PROBABILITY AND MATHEMATICAL STATISTICS

Prasanna Sahoo
Department of Mathematics
University of Louisville
Louisville, KY 40292 USA

THIS BOOK IS DEDICATED TO
AMIT
SADHNA
MY PARENTS, TEACHERS
AND
STUDENTS

PREFACE

This book is both a tutorial and a textbook. In this book we present an introduction to probability and mathematical statistics and it is intended for students already having some elementary mathematical background. It is intended for a one-year senior level undergraduate and beginning graduate level course in probability theory and mathematical statistics. The book contains more material than normally would be taught in a one-year course. This should give the teacher flexibility with respect to the selection of the content and level at which the book is to be used. It has arisen from over 15 years of lectures in senior level calculus based courses in probability theory and mathematical statistics at the University of Louisville.

Probability theory and mathematical statistics are difficult subjects both for students to comprehend and teachers to explain. A good set of examples makes these subjects easy to understand. For this reason alone we have included more than 350 completely worked out examples and over 165 illustrations. We give a rigorous treatment of the fundamentals of probability and statistics using mostly calculus. We have given great attention to the clarity of the presentation of the materials. In the text theoretical results are presented as theorems, proposition or lemma, of which as a rule rigorous proofs are given. In the few exceptions to this rule references are given to indicate where details can be found. This book contains over 450 problems of varying degrees of difficulty to help students master their problem solving skill. To make this less wordy we have

There are several good books on these subjects and perhaps there is no need to bring a new one to the market. So for several years, this was circulated as a series of typeset lecture notes among my students who were preparing for the examination 110 of the Actuarial Society of America. Many of my students encouraged me to formally write it as a book. Actuarial students will benefit greatly from this book. The book is written in simple English; this might be an advantage to students whose native language is not English.

I cannot claim that all the materials I have written in this book are mine. I have learned the subject from many excellent books, such as *Introduction to Mathematical Statistics* by Hogg and Craig, and *An Introduction to Probability Theory and Its Applications* by Feller. In fact, these books have had a profound impact on me, and my explanations are influenced greatly by these textbooks. If there are some resemblances, then it is perhaps due to the fact that I could not improve the original explanations I have learned from these books. I am very thankful to the authors of these great textbooks. I am also thankful to the Actuarial Society of America for letting me use their test problems. I thank all my students in my probability theory and mathematical statistics courses from 1988 to 2003 who helped me in many ways to make this book possible in the present form. Lastly, if it wasn't for the infinite patience of my wife, Sadhna, for last several years, this book would never gotten out of the hard drive of my computer.

The entire book was typeset by the author on a Macintosh computer using TEX, the typesetting system designed by Donald Knuth. The figures were generated by the author using MATHEMATICA, a system for doing mathematics designed by Wolfram Research, and MAPLE, a system for doing mathematics designed by Maplesoft. The author is very thankful to the University of Louisville for providing many internal financial grants while this book was under preparation.

Prasanna Sahoo, *Louisville*

TABLE OF CONTENTS

Chapter 13
SEQUENCES OF RANDOM VARIABLES AND ORDER STASTISTICS

In this chapter, we generalize some of the results we have studied in the previous chapters. We do these generalizations because the generalizations are needed in the subsequent chapters relating mathematical statistics. In this chapter, we also examine the weak law of large numbers, the Bernoulli's law of large numbers, the strong law of large numbers, and the central limit theorem. Further, in this chapter, we treat the order statistics and percentiles.

13.1. Distribution of sample mean and variance

Consider a random experiment. Let X be the random variable associated with this experiment. Let $f(x)$ be the probability density function of X. Let us repeat this experiment n times. Let X_k be the random variable associated with the k^{th} repetition. Then the collection of the random variables $\{ X_1, X_2, ..., X_n \}$ is a random sample of size n. From here after, we simply denote $X_1, X_2, ..., X_n$ as a random sample of size n. The random variables $X_1, X_2, ..., X_n$ are independent and identically distributed with the common probability density function $f(x)$.

For a random sample, functions such as the sample mean $\overline{X}$, the sample variance S^2 are called *statistics*. In a particular sample, say $x_1, x_2, ..., x_n$, we observed $\overline{x}$ and s^2. We may consider

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

as random variables and $\overline{x}$ and s^2 are the realizations from a particular sample.

In this section, we are mainly interested in finding the probability distributions of the sample mean $\overline{X}$ and sample variance S^2, that is the distribution of the statistics of samples.

Example 13.1. Let X_1 and X_2 be a random sample of size 2 from a distribution with probability density function

$$f(x) = \begin{cases} 6x(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

What are the mean and variance of sample sum $Y = X_1 + X_2$?

Answer: The population mean

$$\begin{aligned}
\mu_X &= E\,(X) \\
&= \int_0^1 x\, 6x(1-x)\, dx \\
&= 6 \int_0^1 x^2\,(1-x)\, dx \\
&= 6\, B(3,2) \qquad \text{(here } B \text{ denotes the beta function)} \\
&= 6\, \frac{\Gamma(3)\,\Gamma(2)}{\Gamma(5)} \\
&= 6 \left(\frac{1}{12}\right) \\
&= \frac{1}{2}.
\end{aligned}$$

Since X_1 and X_2 have the same distribution, we obtain $\mu_{X_1} = \frac{1}{2} = \mu_{X_2}$. Hence the mean of Y is given by

$$\begin{aligned}
E(Y) &= E(X_1 + X_2) \\
&= E(X_1) + E(X_2) \\
&= \frac{1}{2} + \frac{1}{2} \\
&= 1.
\end{aligned}$$

Next, we compute the variance of the population X. The variance of X is given by

$$
\begin{aligned}
Var(X) &= E\left(X^2\right) - E(X)^2 \\
&= \int_0^1 6x^3(1-x)\,dx - \left(\frac{1}{2}\right)^2 \\
&= 6\int_0^1 x^3\,(1-x)\,dx - \left(\frac{1}{4}\right) \\
&= 6\,B(4,2) - \left(\frac{1}{4}\right) \\
&= 6\,\frac{\Gamma(4)\,\Gamma(2)}{\Gamma(6)} - \left(\frac{1}{4}\right) \\
&= 6\,\left(\frac{1}{20}\right) - \left(\frac{1}{4}\right) \\
&= \frac{6}{20} - \frac{5}{20} \\
&= \frac{1}{20}.
\end{aligned}
$$

Since X_1 and X_2 have the same distribution as the population X, we get

$$Var(X_1) = \frac{1}{20} = Var(X_2).$$

Hence, the variance of the sample sum Y is given by

$$
\begin{aligned}
Var(Y) &= Var\,(X_1 + X_2) \\
&= Var\,(X_1) + Var\,(X_2) + 2\,Cov\,(X_1, X_2) \\
&= Var\,(X_1) + Var\,(X_2) \\
&= \frac{1}{20} + \frac{1}{20} \\
&= \frac{1}{10}.
\end{aligned}
$$

Example 13.2. Let X_1 and X_2 be a random sample of size 2 from a distribution with density

$$
f(x) = \begin{cases} \frac{1}{4} & \text{for } x = 1, 2, 3, 4 \\ \\ 0 & \text{otherwise.} \end{cases}
$$

What is the distribution of the sample sum $Y = X_1 + X_2$?

Answer: Since the range space of X_1 as well as X_2 is $\{1, 2, 3, 4\}$, the range space of $Y = X_1 + X_2$ is

$$R_Y = \{2, 3, 4, 5, 6, 7, 8\}.$$

Let $g(y)$ be the density function of Y. We want to find this density function. First, we find $g(2)$, $g(3)$ and so on.

$$\begin{aligned}
g(2) &= P(Y = 2) \\
&= P(X_1 + X_2 = 2) \\
&= P(X_1 = 1 \text{ and } X_2 = 1) \\
&= P(X_1 = 1)\ P(X_2 = 1) \qquad \text{(by independence of } X_1 \text{ and } X_2\text{)} \\
&= f(1)\, f(1) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}.
\end{aligned}$$

$$\begin{aligned}
g(3) &= P(Y = 3) \\
&= P(X_1 + X_2 = 3) \\
&= P(X_1 = 1 \text{ and } X_2 = 2) + P(X_1 = 2 \text{ and } X_2 = 1) \\
&= P(X_1 = 1)\ P(X_2 = 2) \\
&\quad + P(X_1 = 2)\ P(X_2 = 1) \qquad \text{(by independence of } X_1 \text{ and } X_2\text{)} \\
&= f(1)\, f(2) + f(2)\, f(1) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{2}{16}.
\end{aligned}$$

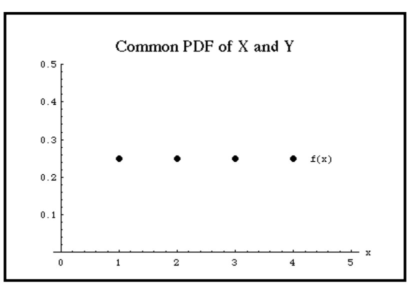
$$
\begin{aligned}
g(4) &= P\,(Y = 4) \\
&= P\,(X_1 + X_2 = 4) \\
&= P\,(X_1 = 1 \text{ and } X_2 = 3) + P\,(X_1 = 3 \text{ and } X_2 = 1) \\
&\quad + P\,(X_1 = 2 \text{ and } X_2 = 2) \\
&= P\,(X_1 = 3)\; P\,(X_2 = 1) + P\,(X_1 = 1)\; P\,(X_2 = 3) \\
&\quad + P\,(X_1 = 2)\; P\,(X_2 = 2) \qquad \text{(by independence of } X_1 \text{ and } X_2\text{)} \\
&= f(1)\, f(3) + f(3)\, f(1) + f(2)\, f(2) \\
&= \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) \\
&= \frac{3}{16}.
\end{aligned}
$$

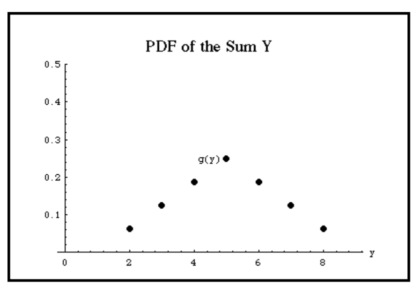
Similarly, we get

$$
g(5) = \frac{4}{16}, \qquad g(6) = \frac{3}{16}, \qquad g(7) = \frac{2}{16}, \qquad g(8) = \frac{1}{16}.
$$

Thus, putting these into one expression, we get

$$
\begin{aligned}
g(y) &= P(Y = y) \\
&= \sum_{k=1}^{y-1} f(k)\, f(y-k) \\
&= \frac{4 - |y-5|}{16}, \qquad y = 2, 3, 4, ..., 8.
\end{aligned}
$$





Remark 13.1. Note that $g(y) = \sum_{k=1}^{y-1} f(k)\, f(y-k)$ is the discrete convolution of f with itself. The concept of convolution was introduced in chapter 10.

The above example can also be done using the moment generating function method as follows:

$$\begin{aligned}
M_Y(t) &= M_{X_1+X_2}(t) \\
&= M_{X_1}(t)\, M_{X_2}(t) \\
&= \left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right) \left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right) \\
&= \left(\frac{e^t + e^{2t} + e^{3t} + e^{4t}}{4}\right)^2 \\
&= \frac{e^{2t} + 2e^{3t} + 3e^{4t} + 4e^{5t} + 3e^{6t} + 2e^{7t} + e^{8t}}{16}.
\end{aligned}$$

Hence, the density of Y is given by

$$g(y) = \frac{4 - |y-5|}{16}, \qquad y = 2, 3, 4, ..., 8.$$

Theorem 13.1. If $X_1, X_2, ..., X_n$ are mutually independent random variables with densities $f_1(x_1)$, $f_2(x_2), ..., f_n(x_n)$ and $E[u_i(X_i)]$, $i = 1, 2, ..., n$ exist, then

$$E\left[\prod_{i=1}^{n} u_i(X_i)\right] = \prod_{i=1}^{n} E[u_i(X_i)],$$

where u_i $(i = 1, 2, ..., n)$ are arbitrary functions.

Proof: We prove the theorem assuming that the random variables $X_1, X_2, ..., X_n$ are continuous. If the random variables are not continuous, then the proof follows exactly in the same manner if one replaces the integrals by summations. Since

$$\begin{aligned}
&E\left(\prod_{i=1}^{n} u_i(X_i)\right) \\
&= E(u_1(X_1) \cdots u_n(X_n)) \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_1(x_1) \cdots u_n(x_n) f(x_1, ..., x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_1(x_1) \cdots u_n(x_n) f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} u_1(x_1) f_1(x_1) dx_1 \cdots \int_{-\infty}^{\infty} u_n(x_1) f_n(x_n) dx_n \\
&= E\left(u_1(X_1)\right) \cdots E\left(u_n(X_n)\right) \\
&= \prod_{i=1}^{n} E\left(u_i(X_i)\right),
\end{aligned}$$

the proof of the theorem is now complete.

Example 13.3. Let X and Y be two random variables with the joint density

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{for } 0 < x,\, y < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of the continuous random variable $Z = X^2Y^2 + XY^2 + X^2 + X$?

Answer: Since

$$\begin{aligned} f(x,y) &= e^{-(x+y)} \\ &= e^{-x}\, e^{-y} \\ &= f_1(x)\, f_2(y), \end{aligned}$$

the random variables X and Y are mutually independent. Hence, the expected value of X is

$$\begin{aligned} E(X) &= \int_0^\infty x\, f_1(x)\, dx \\ &= \int_0^\infty x e^{-x}\, dx \\ &= \Gamma(2) \\ &= 1. \end{aligned}$$

Similarly, the expected value of X^2 is given by

$$\begin{aligned} E\left(X^2\right) &= \int_0^\infty x^2\, f_1(x)\, dx \\ &= \int_0^\infty x^2 e^{-x}\, dx \\ &= \Gamma(3) \\ &= 2. \end{aligned}$$

Since the marginals of X and Y are same, we also get $E(Y) = 1$ and $E(Y^2) = 2$. Further, by Theorem 13.1, we get

$$\begin{aligned} E[Z] &= E\left[X^2Y^2 + XY^2 + X^2 + X\right] \\ &= E\left[\left(X^2 + X\right)\,\left(Y^2 + 1\right)\right] \\ &= E\left[X^2 + X\right]\, E\left[Y^2 + 1\right] \qquad \text{(by Theorem 13.1)} \\ &= \left(E\left[X^2\right] + E\left[X\right]\right)\,\left(E\left[Y^2\right] + 1\right) \\ &= (2+1)\,(2+1) \\ &= 9. \end{aligned}$$

Theorem 13.2. If $X_1, X_2, ..., X_n$ are mutually independent random variables with respective means $\mu_1,\ \mu_2,\ ..., \mu_n$ and variances $\sigma_1^2,\ \sigma_2^2,\ ..., \sigma_n^2$, then the mean and variance of $Y = \sum_{i=1}^{n} a_i\, X_i$, where $a_1,\ a_2,\ ..., a_n$ are real constants, are given by

$$\mu_Y = \sum_{i=1}^{n} a_i\, \mu_i \qquad \text{and} \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2\, \sigma_i^2.$$

Proof: First we show that $\mu_Y = \sum_{i=1}^{n} a_i\, \mu_i$. Since

$$\begin{aligned} \mu_Y &= E(Y) \\ &= E\left(\sum_{i=1}^{n} a_i\, X_i\right) \\ &= \sum_{i=1}^{n} a_i E(X_i) \\ &= \sum_{i=1}^{n} a_i \mu_i \end{aligned}$$

we have asserted result. Next we show $\sigma_Y^2 = \sum_{i=1}^{n} a_i^2\, \sigma_i^2$. Consider

$$\begin{aligned} \sigma_Y^2 &= Var(Y) \\ &= Var\,(a_i\, X_i) \\ &= \sum_{i=1}^{n} a_i^2 Var\,(X_i) \\ &= \sum_{i=1}^{n} a_i^2\, \sigma_i^2. \end{aligned}$$

This completes the proof of the theorem.

Example 13.4. Let the independent random variables X_1 and X_2 have means $\mu_1 = -4$ and $\mu_2 = 3$, respectively and variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 9$. What are the mean and variance of $Y = 3X_1 - 2X_2$?

Answer: The mean of Y is

$$\begin{aligned} \mu_Y &= 3\mu_1 - 2\mu_2 \\ &= 3(-4) - 2(3) \\ &= -18. \end{aligned}$$

Similarly, the variance of Y is

$$\begin{aligned}\sigma_Y^2 &= (3)^2\,\sigma_1^2 + (-2)^2\,\sigma_2^2 \\ &= 9\,\sigma_1^2 + 4\,\sigma_2^2 \\ &= 9(4) + 4(9) \\ &= 72.\end{aligned}$$

Example 13.5. Let $X_1, X_2, ..., X_{50}$ be a random sample of size 50 from a distribution with density

$$f(x) = \begin{cases} \frac{1}{\theta}\,e^{-\frac{x}{\theta}} & \text{for } 0 \leq x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What are the mean and variance of the sample mean $\overline{X}$?

Answer: Since the distribution of the population X is exponential, the mean and variance of X are given by

$$\mu_X = \theta, \qquad \text{and} \qquad \sigma_X^2 = \theta^2.$$

Thus, the mean of the sample mean is

$$\begin{aligned}E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + \cdots + X_{50}}{50}\right) \\ &= \frac{1}{50}\,\sum_{i=1}^{50} E\left(X_i\right) \\ &= \frac{1}{50}\,\sum_{i=1}^{50} \theta \\ &= \frac{1}{50}\,50\,\theta = \theta.\end{aligned}$$

The variance of the sample mean is given by

$$\begin{aligned}Var\left(\overline{X}\right) &= Var\left(\sum_{i=1}^{50} \frac{1}{50}\,X_i\right) \\ &= \sum_{i=1}^{50} \left(\frac{1}{50}\right)^2\,\sigma_{X_i}^2 \\ &= \sum_{i=1}^{50} \left(\frac{1}{50}\right)^2\,\theta^2 \\ &= 50\,\left(\frac{1}{50}\right)^2\,\theta^2 \\ &= \frac{\theta^2}{50}.\end{aligned}$$

Theorem 13.3. If $X_1, X_2, ..., X_n$ are independent random variables with respective moment generating functions $M_{X_i}(t)$, $i = 1, 2, ..., n$, then the moment generating function of $Y = \sum_{i=1}^n a_i X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t).$$

Proof: Since

$$\begin{aligned} M_Y(t) &= M_{\sum_{i=1}^n a_i X_i}(t) \\ &= \prod_{i=1}^{n} M_{a_i X_i}(t) \\ &= \prod_{i=1}^{n} M_{X_i}(a_i t) \end{aligned}$$

we have the asserted result and the proof of the theorem is now complete.

Example 13.6. Let $X_1, X_2, ..., X_{10}$ be the observations from a random sample of size 10 from a distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}, \qquad -\infty < x < \infty.$$

What is the moment generating function of the sample mean?

Answer: The density of the population X is a standard normal. Hence, the moment generating function of each X_i is

$$M_{X_i}(t) = e^{\frac{1}{2}t^2}, \qquad i = 1, 2, ..., 10.$$

The moment generating function of the sample mean is

$$\begin{aligned} M_{\overline{X}}(t) &= M_{\sum_{i=1}^{10} \frac{1}{10} X_i}(t) \\ &= \prod_{i=1}^{10} M_{X_i}\left(\frac{1}{10}\, t\right) \\ &= \prod_{i=1}^{10} e^{\frac{t^2}{200}} \\ &= \left[e^{\frac{t^2}{200}} \right]^{10} = e^{\left(\frac{1}{10}\,\frac{t^2}{2}\right)}. \end{aligned}$$

Hence $\overline{X} \sim N\left(0,\, \frac{1}{10}\right)$.

The last example tells us that if we take a sample of any size from a normal population, then the sample mean also has a normal distribution.

The following theorem says that a linear combination of random variables with normal distributions is again normal.

Theorem 13.4. If $X_1, X_2, ..., X_n$ are mutually independent random variables such that

$$X_i \sim N\left(\mu_i,\, \sigma_i^2\right), \qquad i = 1, 2, ..., n.$$

Then the random variable $Y = \sum_{i=1}^{n} a_i X_i$ is a normal random variable with mean

$$\mu_Y = \sum_{i=1}^{n} a_i\, \mu_i \qquad \text{and} \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2\, \sigma_i^2,$$

that is $Y \sim N\left(\sum_{i=1}^{n} a_i \mu_i,\, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$.

Proof: Since each $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, the moment generating function of each X_i is given by

$$M_{X_i}(t) = e^{\mu_i t + \frac{1}{2}\sigma_i^2 t^2}.$$

Hence using Theorem 13.3, we have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(a_i t) \\ &= \prod_{i=1}^{n} e^{\mu_i t + \frac{1}{2}\sigma_i^2 t^2} \\ &= e^{\sum_{i=1}^{n} \mu_i t + \frac{1}{2}\sum_{i=1}^{n} \sigma_i^2 t^2}. \end{aligned}$$

Thus the random variable $Y \sim N\left(\sum_{i=1}^{n} a_i \mu_i,\, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$. The proof of the theorem is now complete.

Example 13.7. Let $X_1, X_2, ..., X_n$ be the observations from a random sample of size n from a normal distribution with mean μ and variance $\sigma^2 > 0$. What are the mean and variance of the sample mean $\overline{X}$?

Answer: The expected value (or mean) of the sample mean is given by

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{1}{n} \sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mu \\ &= \mu. \end{aligned}$$

Similarly, the variance of the sample mean is

$$Var\left(\overline{X}\right) = \sum_{i=1}^{n} Var\left(\frac{X_i}{n}\right) = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

This example along with the previous theorem says that if we take a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is also normal with mean μ and variance $\frac{\sigma^2}{n}$, that is $\overline{X} \sim N\left(\mu,\, \frac{\sigma^2}{n}\right)$.

Example 13.8. Let $X_1, X_2, ..., X_{64}$ be a random sample of size 64 from a normal distribution with $\mu = 50$ and $\sigma^2 = 16$. What are $P\left(49 < X_8 < 51\right)$ and $P\left(49 < \overline{X} < 51\right)$?

Answer: Since $X_8 \sim N(50, 16)$, we get

$$\begin{aligned}
P\left(49 < X_8 < 51\right) &= P\left(49 - 50 < X_8 - 50 < 51 - 50\right) \\
&= P\left(\frac{49-50}{4} < \frac{X_8 - 50}{4} < \frac{51-50}{4}\right) \\
&= P\left(-\frac{1}{4} < \frac{X_8 - 50}{4} < \frac{1}{4}\right) \\
&= P\left(-\frac{1}{4} < Z < \frac{1}{4}\right) \\
&= 2P\left(Z < \frac{1}{4}\right) - 1 \\
&= 0.1974 \qquad \text{(from normal table).}
\end{aligned}$$

By the previous theorem, we see that $\overline{X} \sim N\left(50,\, \frac{16}{64}\right)$. Hence

$$\begin{aligned}
P\left(49 < \overline{X} < 51\right) &= P\left(49 - 50 < \overline{X} - 50 < 51 - 50\right) \\
&= P\left(\frac{49-50}{\sqrt{\frac{16}{64}}} < \frac{\overline{X} - 50}{\sqrt{\frac{16}{64}}} < \frac{51-50}{\sqrt{\frac{16}{64}}}\right) \\
&= P\left(-2 < \frac{\overline{X} - 50}{\sqrt{\frac{16}{64}}} < 2\right) \\
&= P\left(-2 < Z < 2\right) \\
&= 2P\left(Z < 2\right) - 1 \\
&= 0.9544 \qquad \text{(from normal table).}
\end{aligned}$$

This example tells us that $\overline{X}$ has a greater probability of falling in an interval containing μ, than a single observation, say X_8 (or in general any X_i).

Theorem 13.5. Let the distributions of the random variables $X_1, X_2, ..., X_n$ be $\chi^2(r_1)$, $\chi^2(r_2)$, ..., $\chi^2(r_n)$, respectively. If $X_1, X_2, ..., X_n$ are mutually independent, then $Y = X_1 + X_2 + \cdots + X_n \sim \chi^2\left(\sum_{i=1}^n r_i\right)$.

Proof: Since each $X_i \sim \chi^2(r_i)$, the moment generating function of each X_i is given by

$$M_{X_i}(t) = (1 - 2t)^{-\frac{r_i}{2}}.$$

By Theorem 13.3, we have

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (1 - 2t)^{-\frac{r_i}{2}} = (1 - 2t)^{-\frac{1}{2}\sum_{i=1}^n r_i}.$$

Hence $Y \sim \chi^2\left(\sum_{i=1}^n r_i\right)$ and the proof of the theorem is now complete.

The proof of the following theorem is an easy consequence of Theorem 13.5 and we leave the proof to the reader.

Theorem 13.6. If $Z_1, Z_2, ..., Z_n$ are mutually independent and each one is standard normal, then $Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$, that is the sum is chi-square with n degrees of freedom.

The following theorem is very useful in mathematical statistics and its proof is beyond the scope of this introductory book.

Theorem 13.7. If $X_1, X_2, ..., X_n$ are observations of a random sample of size n from the normal distribution $N\left(\mu, \sigma^2\right)$, then the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ have the following properties:

(A) $\overline{X}$ and S^2 are independent, and
(B) $(n - 1)\,\frac{S^2}{\sigma^2} \sim \chi^2(n - 1)$.

Remark 13.2. At first sight the statement (A) might seem odd since the sample mean $\overline{X}$ occurs explicitly in the definition of the sample variance S^2. This remarkable independence of $\overline{X}$ and S^2 is a unique property that distinguishes normal distribution from all other probability distributions.

Example 13.9. Let $X_1, X_2, ..., X_n$ denote a random sample from a normal distribution with variance $\sigma^2 > 0$. If the first percentile of the statistics $W = \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{\sigma^2}$ is 1.24, where $\overline{X}$ denotes the sample mean, what is the sample size n?

Answer:

$$\begin{aligned}
\frac{1}{100} &= P\,(W \leq 1.24) \\
&= P\left(\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \leq 1.24\right) \\
&= P\left((n-1)\,\frac{S^2}{\sigma^2} \leq 1.24\right) \\
&= P\left(\chi^2(n-1) \leq 1.24\right).
\end{aligned}$$

Thus from χ^2-table, we get

$$n - 1 = 7$$

and hence the sample size n is 8.

Example 13.10. Let $X_1, X_2, ..., X_4$ be a random sample from a normal distribution with unknown mean and variance equal to 9. Let $S^2 = \frac{1}{3}\sum_{i=1}^{4}\left(X_i - \overline{X}\right)$. If $P\left(S^2 \leq k\right) = 0.05$, then what is k?

Answer:

$$\begin{aligned}
0.05 &= P\left(S^2 \leq k\right) \\
&= P\left(\frac{3S^2}{9} \leq \frac{3}{9}\,k\right) \\
&= P\left(\chi^2(3) \leq \frac{3}{9}\,k\right).
\end{aligned}$$

From χ^2-table with 3 degrees of freedom, we get

$$\frac{3}{9}\,k = 0.35$$

and thus the constant k is given by

$$k = 3(0.35) = 1.05.$$

13.2. Laws of Large Numbers

In this section, we mainly examine the weak law of large numbers. The weak law of large numbers states that if $X_1, X_2, ..., X_n$ is a random sample of size n from a population X with mean μ, then the sample mean $\overline{X}$ rarely deviates from the population mean μ when the sample size n is very large. In other words, the sample mean $\overline{X}$ converges in probability to the population mean μ. We begin this section with a result known as Markov inequality which is need to establish the weak law of large numbers.

Theorem 13.8 (Markov Inequality). Suppose X is a nonnegative random variable with mean $E(X)$. Then

$$P(X \geq t) \leq \frac{E(X)}{t}$$

for all $t > 0$.

Proof: We assume the random variable X is continuous. If X is not continuous, then a proof can be obtained for this case by replacing the integrals with summations in the following proof. Since

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_{-\infty}^{t} x f(x) dx + \int_{t}^{\infty} x f(x) dx \\ &\geq \int_{t}^{\infty} x f(x) dx \\ &\geq \int_{t}^{\infty} t f(x) dx \qquad \text{because } x \in [t, \infty) \\ &= t \int_{t}^{\infty} f(x) dx \\ &= t\, P(X \geq t), \end{aligned}$$

we see that

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

This completes the proof of the theorem.

In Theorem 4.4 of the chapter 4, Chebychev inequality was treated. Let X be a random variable with mean μ and standard deviation σ. Then Chebychev inequality says that

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

for any nonzero positive constant k. This result can be obtained easily using Theorem 13.8 as follows. By Markov inequality, we have

$$P((X - \mu)^2 \geq t^2) \leq \frac{E((X - \mu)^2)}{t^2}$$

for all $t > 0$. Since the events $(X - \mu)^2 \geq t^2$ and $|X - \mu| \geq t$ are same, we get

$$P((X - \mu)^2 \geq t^2) = P(|X - \mu| \geq t) \leq \frac{E((X - \mu)^2)}{t^2}$$

for all $t > 0$. Hence

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Letting $k = \frac{\sigma}{t}$ in the above equality, we see that

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Hence

$$1 - P(|X - \mu| < k\sigma) \leq \frac{1}{k^2}.$$

The last inequality yields the Chebychev inequality

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Now we are ready to treat the weak law of large numbers.

Theorem 13.9. Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $i = 1, 2, ..., \infty$. Then

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) = 0$$

for every ε. Here $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

Proof: By Theorem 13.2 (or Example 13.7) we have

$$E(\overline{S}_n) = \mu \qquad \text{and} \qquad Var(\overline{S}_n) = \frac{\sigma^2}{n}.$$

By Chebychev's inequality

$$P(|\overline{S}_n - E(\overline{S}_n)| \geq \varepsilon) \leq \frac{Var(\overline{S}_n)}{\varepsilon^2}$$

for $\varepsilon > 0$. Hence

$$P(|\overline{S}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\,\varepsilon^2}.$$

Taking the limit as n tends to infinity, we get

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\,\varepsilon^2}$$

which yields

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| \geq \varepsilon) = 0$$

and the proof of the theorem is now complete.

It is possible to prove the weak law of large numbers assuming only $E(X)$ to exist and finite but the proof is more involved.

The weak law of large numbers says that the sequence of sample means $\{\overline{S}_n\}_{n=1}^{\infty}$ from a population X stays close to population mean $E(X)$ most of the times. Let us consider an experiment that consists of tossing a coin infinitely many times. Let X_i be 1 if the i^{th} toss results in a Head, and 0 otherwise. The weak law of large numbers says that

$$\overline{S}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \frac{1}{2} \qquad \text{as} \quad n \to \infty \tag{13.0}$$

but it is easy to come up with sequences of tosses for which (13.0) is false:

H H H H H H H H H H H H ··· ···
H H T H H T H H T H H T ··· ···

The strong law of large numbers (Theorem 13.11) states that the set of "bad sequences" like the ones given above has probability zero.

Note that the assertion of Theorem 13.9 for any $\varepsilon > 0$ can also be written as

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu| < \varepsilon) = 1.$$

The type of convergence we saw in the weak law of large numbers is not the type of convergence discussed in calculus. This type of convergence is called convergence in probability and defined as follows.

Definition 13.1. Suppose $X_1, X_2, ...$ is a sequence of random variables defined on a sample space S. The sequence *converges in probability* to the random variable X if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

In view of the above definition, the weak law of large numbers states that the sample mean $\overline{X}$ converges in probability to the population mean μ.

The following theorem is known as the Bernoulli law of large numbers and is a special case of the weak law of large numbers.

Theorem 13.10. Let $X_1, X_2, ...$ be a sequence of independent and identically distributed Bernoulli random variables with probability of success p. Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\overline{S}_n - p| < \varepsilon) = 1$$

where $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

The fact that the relative frequency of occurrence of an event E is very likely to be close to its probability $P(E)$ for large n can be derived from the weak law of large numbers. Consider a repeatable random experiment repeated large number of time independently. Let $X_i = 1$ if E occurs on the i^{th} repetition and $X_i = 0$ if E does not occur on i^{th} repetition. Then

$$\mu = E(X_i) = 1 \cdot P(E) + 0 \cdot P(E) = P(E) \qquad \text{for } i = 1, 2, 3, ...$$

and

$$X_1 + X_2 + \cdots + X_n = n\, N(E)$$

where $N(E)$ denotes the number of times E occurs. Hence by the weak law of large numbers, we have

$$\begin{aligned}\lim_{n\to\infty} P\left(\left|\frac{N(E)}{n} - P(E)\right| > \varepsilon\right) &= \lim_{n\to\infty} P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \varepsilon\right) \\ &= \lim_{n\to\infty} P\left(\left|\overline{S}_n - \mu\right| > \varepsilon\right) \\ &= 0.\end{aligned}$$

Hence, for large n, the relative frequency of occurrence of the event E is very likely to be close to its probability $P(E)$.

Now we present the strong law of large numbers without a proof.

Theorem 13.11. Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $\mu = E(X_i)$ and $\sigma^2 = Var(X_i) < \infty$ for $i = 1, 2, ..., \infty$. Then

$$\lim_{n\to\infty} P(|\overline{S}_n - \mu) \geq \varepsilon) = 0$$

for every ε. Here $\overline{S}_n$ denotes $\frac{X_1+X_2+\cdots+X_n}{n}$.

The type convergence in Theorem 13.11 is called almost sure convergence. The notion of almost sure convergence is defined as follows.

Definition 13.2 Suppose the random variable X and the sequence $X_1, X_2, ...,$ of random variables are defined on a sample space S. The sequence $X_n(w)$ *converges almost surely* to $X(w)$ if

$$P\left(\left\{w \in S \;\Big|\; \lim_{n\to\infty} X_n(w) = X(w)\right\}\right) = 1.$$

It can be shown that the convergence in probability implies the almost sure convergence but not the converse.

13.3. The Central Limit Theorem

Consider a random sample of measurement $\{X_i\}_{i=1}^n$. The X_i's are identically distributed and their common distribution is the distribution of the population. We have seen that if the population distribution is normal, then the sample mean $\overline{X}$ is also normal. More precisely, if $X_1, X_2, ..., X_n$ is a random sample from a normal distribution with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

then

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The central limit theorem (also known as Lindeberg-Levy Theorem) states that even though the population distribution may be far from being normal, still for large sample size n, the distribution of the standardized sample mean is approximately standard normal with better approximations obtained with the larger sample size. Mathematically this can be stated as follows.

Theorem 13.12 (Central Limit Theorem). Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with mean μ and variance $\sigma^2 < \infty$, then the limiting distribution of

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is standard normal, that is Z_n converges in distribution to Z where Z denotes a standard normal random variable.

The type of convergence used in the central limit theorem is called the convergence in distribution and is defined as follows.

Definition 13.3. Suppose X is a random variable with cumulative density function $F(x)$ and the sequence $X_1, X_2, ...$ of random variables with cumulative density functions $F_1(x), F_2(x), ...$, respectively. The sequence X_n *converges in distribution* to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all values x at which $F(x)$ is continuous. The distribution of X is called the *limiting distribution* of X_n.

Whenever a sequence of random variables $X_1, X_2, ...$ converges in distribution to the random variable X, it will be denoted by $X_n \xrightarrow{d} X$.

Example 13.11. Let $Y = X_1 + X_2 + \cdots + X_{15}$ be the sum of a random sample of size 15 from the distribution whose density function is

$$f(x) = \begin{cases} \frac{3}{2}\,x^2 & \text{if } -1 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the approximate value of $P(-0.3 \leq Y \leq 1.5)$ when one uses the central limit theorem?

Answer: First, we find the mean μ and variance σ^2 for the density function $f(x)$. The mean for this distribution is given by

$$\begin{aligned} \mu &= \int_{-1}^{1} \frac{3}{2} x^3 \, dx \\ &= \frac{3}{2} \left[\frac{x^4}{4} \right]_{-1}^{1} \\ &= 0. \end{aligned}$$

Hence the variance of this distribution is given by

$$\begin{aligned} Var(X) &= E(X^2) - [\,E(X)\,]^2 \\ &= \int_{-1}^{1} \frac{3}{2} x^4 \, dx \\ &= \frac{3}{2} \left[\frac{x^5}{5} \right]_{-1}^{1} \\ &= \frac{3}{5} \\ &= 0.6. \end{aligned}$$

$$\begin{aligned} P(-0.3 \leq Y \leq 1.5) &= P(-0.3 - 0 \leq Y - 0 \leq 1.5 - 0) \\ &= P\left(\frac{-0.3}{\sqrt{15(0.6)}} \leq \frac{Y - 0}{\sqrt{15(0.6)}} \leq \frac{1.5}{\sqrt{15(0.6)}} \right) \\ &= P(-0.10 \leq Z \leq 0.50) \\ &= P(Z \leq 0.50) + P(Z \leq 0.10) - 1 \\ &= 0.6915 + 0.5398 - 1 \\ &= 0.2313. \end{aligned}$$

Example 13.12. Let $X_1, X_2, ..., X_n$ be a random sample of size $n = 25$ from a population that has a mean $\mu = 71.43$ and variance $\sigma^2 = 56.25$. Let $\overline{X}$ be the sample mean. What is the probability that the sample mean is between 68.91 and 71.97?

Answer: The mean of $\overline{X}$ is given by $E\left(\overline{X}\right) = 71.43$. The variance of $\overline{X}$ is given by

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{56.25}{25} = 2.25.$$

In order to find the probability that the sample mean is between 68.91 and 71.97, we need the distribution of the population. However, the population distribution is unknown. Therefore, we use the central limit theorem. The central limit theorem says that $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ as n approaches infinity. Therefore

$$\begin{aligned}
& P\left(68.91 \leq \overline{X} \leq 71.97\right) \\
&= \left(\frac{68.91 - 71.43}{\sqrt{2.25}} \leq \frac{\overline{X} - 71.43}{\sqrt{2.25}} \leq \frac{71.97 - 71.43}{\sqrt{2.25}}\right) \\
&= P\left(-0.68 \leq W \leq 0.36\right) \\
&= P\left(W \leq 0.36\right) + P\left(W \leq 0.68\right) - 1 \\
&= 0.5941.
\end{aligned}$$

Example 13.13. Light bulbs are installed successively into a socket. If we assume that each light bulb has a mean life of 2 months with a standard deviation of 0.25 months, what is the probability that 40 bulbs last at least 7 years?

Answer: Let X_i denote the life time of the i^{th} bulb installed. The 40 light bulbs last a total time of

$$S_{40} = X_1 + X_2 + \cdots + X_{40}.$$

By the central limit theorem

$$\frac{\sum_{i=1}^{40} X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1) \qquad \text{as} \qquad n \to \infty.$$

Thus

$$\frac{S_{40} - (40)(2)}{\sqrt{(40)(0.25)^2}} \sim N(0,1).$$

That is

$$\frac{S_{40} - 80}{1.581} \sim N(0,1).$$

Therefore

$$\begin{aligned} P\left(S_{40} \geq 7(12)\right) \\ &= P\left(\frac{S_{40} - 80}{1.581} \geq \frac{84 - 80}{1.581}\right) \\ &= P\left(Z \geq 2.530\right) \\ &= 0.0057. \end{aligned}$$

Example 13.14. Light bulbs are installed into a socket. Assume that each has a mean life of 2 months with standard deviation of 0.25 month. How many bulbs n should be bought so that one can be 95% sure that the supply of n bulbs will last 5 years?

Answer: Let X_i denote the life time of the i^{th} bulb installed. The n light bulbs last a total time of

$$S_n = X_1 + X_2 + \cdots + X_n.$$

The total average life span S_n has

$$E\left(S_n\right) = 2n \qquad \text{and} \qquad Var(S_n) = \frac{n}{16}.$$

By the central limit theorem, we get

$$\frac{S_n - E\left(S_n\right)}{\frac{\sqrt{n}}{4}} \sim N(0,1).$$

Thus, we seek n such that

$$\begin{aligned} 0.95 &= P\left(S_n \geq 60\right) \\ &= P\left(\frac{S_n - 2n}{\frac{\sqrt{n}}{4}} \geq \frac{60 - 2n}{\frac{\sqrt{n}}{4}}\right) \\ &= P\left(Z \geq \frac{240 - 8n}{\sqrt{n}}\right) \\ &= 1 - P\left(Z \leq \frac{240 - 8n}{\sqrt{n}}\right). \end{aligned}$$

From the standard normal table, we get

$$\frac{240 - 8n}{\sqrt{n}} = -1.645$$

which implies

$$1.645\sqrt{n} + 8n - 240 = 0.$$

Solving this quadratic equation for $\sqrt{n}$, we get

$$\sqrt{n} = -5.375 \quad \text{or} \quad 5.581.$$

Thus $n = 31.15$. So we should buy 32 bulbs.

Example 13.15. American Airlines claims that the average number of people who pay for in-flight movies, when the plane is fully loaded, is 42 with a standard deviation of 8. A sample of 36 fully loaded planes is taken. What is the probability that fewer than 38 people paid for the in-flight movies?

Answer: Here, we like to find $P(\overline{X} < 38)$. Since, we do not know the distribution of $\overline{X}$, we will use the central limit theorem. We are given that the population mean is $\mu = 42$ and population standard deviation is $\sigma = 8$. Moreover, we are dealing with sample of size $n = 36$. Thus

$$\begin{aligned} P(\overline{X} < 38) &= P\left(\frac{\overline{X} - 42}{\frac{8}{6}} < \frac{38 - 42}{\frac{8}{6}}\right) \\ &= P\left(Z < -3\right) \\ &= 1 - P(Z < 3) \\ &= 1 - 0.9987 \\ &= 0.0013. \end{aligned}$$

Since we have not yet seen the proof of the central limit theorem, first let us go through some examples to see the main idea behind the proof of the central limit theorem. Later, at the end of this section a proof of the central limit theorem will be given. We know from the central limit theorem that if $X_1, X_2, ..., X_n$ is a random sample of size n from a distribution with mean μ and variance σ^2, then

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} Z \sim N(0, 1) \qquad \text{as} \qquad n \to \infty.$$

However, the above expression is *not equivalent* to

$$\overline{X} \xrightarrow{d} Z \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{as} \qquad n \to \infty$$

as the following example shows.

Example 13.16. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a gamma distribution with parameters $\theta = 1$ and $\alpha = 1$. What is the distribution of the sample mean $\overline{X}$? Also, what is the limiting distribution of $\overline{X}$ as $n \to \infty$?

Answer: Since, each $X_i \sim GAM(1,1)$, the probability density function of each X_i is given by

$$f(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ \\ 0 & \text{otherwise} \end{cases}$$

and hence the moment generating function of each X_i is

$$M_{X_i}(t) = \frac{1}{1-t}.$$

First we determine the moment generating function of the sample mean $\overline{X}$, and then examine this moment generating function to find the probability distribution of $\overline{X}$. Since

$$\begin{aligned} M_{\overline{X}}(t) &= M_{\frac{1}{n}\sum_{i=1}^{n} X_i}(t) \\ &= \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{n}\right) \\ &= \prod_{i=1}^{n} \frac{1}{\left(1-\frac{t}{n}\right)} \\ &= \frac{1}{\left(1-\frac{t}{n}\right)^n}, \end{aligned}$$

therefore $\overline{X} \sim GAM\left(\frac{1}{n}, n\right)$.

Next, we find the limiting distribution of $\overline{X}$ as $n \to \infty$. This can be done again by finding the limiting moment generating function of $\overline{X}$ and identifying the distribution of $\overline{X}$. Consider

$$\begin{aligned} \lim_{n\to\infty} M_{\overline{X}}(t) &= \lim_{n\to\infty} \frac{1}{\left(1-\frac{t}{n}\right)^n} \\ &= \frac{1}{\lim_{n\to\infty}\left(1-\frac{t}{n}\right)^n} \\ &= \frac{1}{e^{-t}} \\ &= e^t. \end{aligned}$$

Thus, the sample mean $\overline{X}$ has a degenerate distribution, that is all the probability mass is concentrated at one point of the space of $\overline{X}$.

Example 13.17. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a gamma distribution with parameters $\theta = 1$ and $\alpha = 1$. What is the distribution of

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad \text{as} \qquad n \to \infty$$

where μ and σ are the population mean and variance, respectively?

Answer: From Example 13.7, we know that

$$M_{\overline{X}}(t) = \frac{1}{\left(1 - \frac{t}{n}\right)^n}.$$

Since the population distribution is gamma with $\theta = 1$ and $\alpha = 1$, the population mean μ is 1 and population variance σ^2 is also 1. Therefore

$$\begin{aligned} M_{\frac{\overline{X}-1}{\frac{1}{\sqrt{n}}}}(t) &= M_{\sqrt{n}\,\overline{X}-\sqrt{n}}(t) \\ &= e^{-\sqrt{n}t}\, M_{\overline{X}}\left(\sqrt{n}\,t\right) \\ &= e^{-\sqrt{n}t}\, \frac{1}{\left(1 - \frac{\sqrt{n}t}{n}\right)^n} \\ &= \frac{1}{e^{\sqrt{n}t}\, \left(1 - \frac{t}{\sqrt{n}}\right)^n}. \end{aligned}$$

The limiting moment generating function can be obtained by taking the limit of the above expression as n tends to infinity. That is,

$$\begin{aligned} \lim_{n\to\infty} M_{\frac{\overline{X}-1}{\frac{1}{\sqrt{n}}}}(t) &= \lim_{n\to\infty} \frac{1}{e^{\sqrt{n}t}\, \left(1 - \frac{t}{\sqrt{n}}\right)^n} \\ &= e^{\frac{1}{2}t^2} \qquad \text{(using MAPLE)} \\ &= \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1). \end{aligned}$$

The following theorem is used to prove the central limit theorem.

Theorem 13.13 (Lévy Continuity Theorem). Let $X_1, X_2, ...$ be a sequence of random variables with distribution functions $F_1(x), F_2(x), ...$ and moment generating functions $M_{X_1}(t), M_{X_2}(t), ...$, respectively. Let X be a random variable with distribution function $F(x)$ and moment generating function $M_X(t)$. If for all t in the open interval $(-h, h)$ for some $h > 0$

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t),$$

then at the points of continuity of $F(x)$

$$\lim_{n\to\infty} F_n(x) = F(x).$$

The proof of this theorem is beyond the scope of this book.

The following limit

$$\lim_{n\to\infty} \left[1 + \frac{t}{n} + \frac{d(n)}{n}\right]^n = e^t, \qquad \text{if } \lim_{n\to\infty} d(n) = 0, \tag{13.1}$$

whose proof we leave it to the reader, can be established using advanced calculus. Here t is independent of n.

Now we proceed to prove the central limit theorem assuming that the moment generating function of the population X exists. Let $M_{X-\mu}(t)$ be the moment generating function of the random variable $X - \mu$. We denote $M_{X-\mu}(t)$ as $M(t)$ when there is no danger of confusion. Then

$$\left.\begin{aligned} M(0) &= 1, \\ M'(0) &= E(X-\mu) = E(X) - \mu = \mu - \mu = 0, \\ M''(0) &= E\left((X-\mu)^2\right) = \sigma^2. \end{aligned}\right\} \tag{13.2}$$

By Taylor series expansion of $M(t)$ about 0, we get

$$M(t) = M(0) + M'(0)\,t + \frac{1}{2}\,M''(\eta)\,t^2$$

where $\eta \in (0, t)$. Hence using (13.2), we have

$$\begin{aligned} M(t) &= 1 + \frac{1}{2}\,M''(\eta)\,t^2 \\ &= 1 + \frac{1}{2}\,\sigma^2\,t^2 + \frac{1}{2}\,M''(\eta)\,t^2 - \frac{1}{2}\,\sigma^2\,t^2 \\ &= 1 + \frac{1}{2}\,\sigma^2\,t^2 + \frac{1}{2}\,\left[M''(\eta) - \sigma^2\right]\,t^2. \end{aligned}$$

Now using $M(t)$ we compute the moment generating function of Z_n. Note that

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu).$$

Hence

$$\begin{aligned} M_{Z_n}(t) &= \prod_{i=1}^{n} M_{X_i-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \prod_{i=1}^{n} M_{X-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= \left[M\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left[1 + \frac{t^2}{2\,n} + \frac{(M''(\eta)-\sigma^2)\,t^2}{2\,n\,\sigma^2}\right]^n \end{aligned}$$

for $0 < |\eta| < \frac{1}{\sigma\sqrt{n}}\,|t|$. Note that since $0 < |\eta| < \frac{1}{\sigma\sqrt{n}}\,|t|$, we have

$$\lim_{n\to\infty} \frac{t}{\sigma\sqrt{n}} = 0, \quad \lim_{n\to\infty} \eta = 0, \quad \text{and} \quad \lim_{n\to\infty} M''(\eta) - \sigma^2 = 0. \tag{13.3}$$

Letting

$$d(n) = \frac{(M''(\eta)-\sigma^2)\,t^2}{2\,\sigma^2}$$

and using (13.3), we see that $\lim\limits_{n\to\infty} d(n) = 0$, and

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2\,n} + \frac{d(n)}{n}\right]^n. \tag{13.4}$$

Using (13.1) we have

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left[1 + \frac{t^2}{2\,n} + \frac{d(n)}{n}\right]^n = e^{\frac{1}{2}\,t^2}.$$

Hence by the Lévy continuity theorem, we obtain

$$\lim_{n\to\infty} F_n(x) = \Phi(x)$$

where $\Phi(x)$ is the cumulative density function of the standard normal distribution. Thus $Z_n \xrightarrow{d} Z$ and the proof of the theorem is now complete.

Remark 13.3. In contrast to the moment generating function, since the characteristic function of a random variable always exists, the original proof of the central limit theorem involved the characteristic function (see for example *An Introduction to Probability Theory and Its Applications, Volume II* by Feller). In 1988, Brown gave an elementary proof using very clever Taylor series expansions, where the use characteristic function has been avoided.

13.4. Order Statistics

Often, sample values such as the smallest, largest, or middle observation from a random sample provide important information. For example, the highest flood water or lowest winter temperature recorded during the last 50 years might be useful when planning for future emergencies. The median price of houses sold during the previous month might be useful for estimating the cost of living. The statistics highest, lowest or median are examples of order statistics.

Definition 13.4. Let $X_1, X_2, ..., X_n$ be observations from a random sample of size n from a distribution $f(x)$. Let $X_{(1)}$ denote the smallest of $\{X_1, X_2, ..., X_n\}$, $X_{(2)}$ denote the second smallest of $\{X_1, X_2, ..., X_n\}$, and similarly $X_{(r)}$ denote the r^{th} smallest of $\{X_1, X_2, ..., X_n\}$. Then the random variables $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are called the order statistics of the sample $X_1, X_2, ..., X_n$. In particular, $X_{(r)}$ is called the r^{th}-order statistic of $X_1, X_2, ..., X_n$.

The sample range, R, is the distance between the smallest and the largest observation. That is,

$$R = X_{(n)} - X_{(1)}.$$

This is an important statistic which is defined using order statistics.

The distribution of the order statistics are very important when one uses these in any statistical investigation. The next theorem gives the distribution of an order statistic.

Theorem 13.14. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with density function $f(x)$. Then the probability density function of the r^{th} order statistic, $X_{(r)}$, is

$$g(x) = \frac{n!}{(r-1)!\,(n-r)!}\,[F(x)]^{r-1}\,f(x)\,[1-F(x)]^{n-r}\,,$$

where $F(x)$ denotes the cdf of $f(x)$.

Proof: Let h be a positive real number. Let us divide the real line into three segments, namely

$$\mathbb{R} = (-\infty,\, x) \bigcup [x,\, x+h] \bigcup (x+h,\, \infty).$$

The probability, say p_1, of a sample value falls into the first interval $(-\infty,\, x]$ and is given by

$$p_1 = \int_{-\infty}^{x} f(t)\, dt.$$

Similarly, the probability p_2 of a sample value falls into the second interval and is

$$p_2 = \int_{x}^{x+h} f(t)\, dt.$$

In the same token, we can compute the probability of a sample value which falls into the third interval

$$p_3 = \int_{x+h}^{\infty} f(t)\, dt.$$

Then the probability that $(r-1)$ sample values fall in the first interval, one falls in the second interval, and $(n-r)$ fall in the third interval is

$$\binom{n}{r-1,\, 1,\, n-r} p_1^{r-1}\, p_2^1\, p_3^{n-r} = \frac{n!}{(r-1)!\,(n-r)!}\, p_1^{r-1}\, p_2\, p_3^{n-r} =: P_h(x).$$

Since

$$g(x) = \lim_{h\to 0} \frac{P_h(x)}{h},$$

the probability density function of the r^{th} statistics is given by

$$\begin{aligned}
g(x) &= \lim_{h\to 0} \frac{P_h(x)}{h} \\
&= \lim_{h\to 0} \left[\frac{n!}{(r-1)!\,(n-r)!}\, p_1^{r-1}\, \frac{p_2}{h}\, p_3^{n-r} \right] \\
&= \frac{n!}{(r-1)!\,(n-r)!} \left[\left(\lim_{h\to 0} \int_{-\infty}^{x} f(t)\, dt \right)^{r-1} \right] \\
&\qquad \left[\lim_{h\to 0} \frac{1}{h} \int_{x}^{x+h} f(t)\, dt \right] \left[\left(\lim_{h\to 0} \int_{x+h}^{\infty} f(t)\, dt \right)^{n-r} \right] \\
&= \frac{n!}{(r-1)!\,(n-r)!}\, \left[F(x)\right]^{r-1}\, f(x)\, \left[1-F(x)\right]^{n-r}.
\end{aligned}$$

The second limit is obtained as $f(x)$ due to the fact that

$$\begin{aligned}
&\lim_{h\to 0} \frac{1}{h} \int_{x}^{x+h} f(t)\, dt \\
&\quad = \lim_{h\to 0} \frac{\int_a^{x+h} f(t)\, dt - \int_a^x f(t)\, dt}{h}, \quad \text{where } x < a < x+h \\
&\quad = \frac{d}{dx} \int_a^x f(t)\, dt \\
&\quad = f(x)
\end{aligned}$$

Example 13.18. Let X_1, X_2 be a random sample from a distribution with density function

$$f(x) = \begin{cases} e^{-x} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the density function of $Y = \min\{X_1, X_2\}$ where nonzero?

Answer: The cumulative distribution function of $f(x)$ is

$$\begin{aligned} F(x) &= \int_0^x e^{-t}\, dt \\ &= 1 - e^{-x} \end{aligned}$$

In this example, $n = 2$ and $r = 1$. Hence, the density of Y is

$$\begin{aligned} g(y) &= \frac{2!}{0!\,1!}\, [F(y)]^0\, f(y)\, [1 - F(y)] \\ &= 2f(y)\, [1 - F(y)] \\ &= 2\, e^{-y} \left(1 - 1 + e^{-y}\right) \\ &= 2\, e^{-2y}. \end{aligned}$$

Example 13.19. Let $Y_1 < Y_2 < \cdots < Y_6$ be the order statistics from a random sample of size 6 from a distribution with density function

$$f(x) = \begin{cases} 2x & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of Y_6?

Answer:

$$\begin{aligned} f(x) &= 2x \\ F(x) &= \int_0^x 2t\, dt \\ &= x^2. \end{aligned}$$

The density function of Y_6 is given by

$$\begin{aligned} g(y) &= \frac{6!}{5!\,0!}\, [F(y)]^5\, f(y) \\ &= 6\left(y^2\right)^5\, 2y \\ &= 12y^{11}. \end{aligned}$$

Hence, the expected value of Y_6 is

$$\begin{aligned} E\left(Y_6\right) &= \int_0^1 y\, g(y)\, dy \\ &= \int_0^1 y\, 12 y^{11}\, dy \\ &= \frac{12}{13}\, \left[y^{13}\right]_0^1 \\ &= \frac{12}{13}. \end{aligned}$$

Example 13.20. Let X, Y and Z be independent uniform random variables on the interval $(0, a)$. Let $W = \min\{X, Y, Z\}$. What is the expected value of $\left(1 - \frac{W}{a}\right)^2$?

Answer: The probability distribution of X (or Y or Z) is

$$f(x) = \begin{cases} \frac{1}{a} & \text{if } 0 < x < a \\ \\ 0 & \text{otherwise.} \end{cases}$$

Thus the cumulative distribution of function of $f(x)$ is given by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \\ \frac{x}{a} & \text{if } 0 < x < a \\ \\ 1 & \text{if } x \geq a. \end{cases}$$

Since $W = \min\{X, Y, Z\}$, W is the first order statistic of the random sample X, Y, Z. Thus, the density function of W is given by

$$\begin{aligned} g(w) &= \frac{3!}{0!\, 1!\, 2!}\, \left[F(w)\right]^0\, f(w)\, \left[1 - F(w)\right]^2 \\ &= 3 f(w)\, \left[1 - F(w)\right]^2 \\ &= 3\, \left(1 - \frac{w}{a}\right)^2\, \left(\frac{1}{a}\right) \\ &= \frac{3}{a}\, \left(1 - \frac{w}{a}\right)^2. \end{aligned}$$

Thus, the pdf of W is given by

$$g(w) = \begin{cases} \frac{3}{a}\, \left(1 - \frac{w}{a}\right)^2 & \text{if } 0 < w < a \\ \\ 0 & \text{otherwise.} \end{cases}$$

The expected value of W is

$$
\begin{aligned}
E&\left[\left(1-\frac{W}{a}\right)^2\right] \\
&= \int_0^a \left(1-\frac{w}{a}\right)^2 g(w)\,dw \\
&= \int_0^a \left(1-\frac{w}{a}\right)^2 \frac{3}{a}\left(1-\frac{w}{a}\right)^2 dw \\
&= \int_0^a \frac{3}{a}\left(1-\frac{w}{a}\right)^4 dx \\
&= -\frac{3}{5}\left[\left(1-\frac{w}{a}\right)^5\right]_0^a \\
&= \frac{3}{5}.
\end{aligned}
$$

Example 13.21. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with uniform distribution on the interval $[0,1]$. What is the probability distribution of the sample range $W := X_{(n)} - X_{(1)}$?

Answer: To find the distribution of W, we need the joint distribution of the random variable $\left(X_{(n)}, X_{(1)}\right)$. The joint distribution of $\left(X_{(n)}, X_{(1)}\right)$ is given by

$$h(x_1, x_n) = n(n-1) f(x_1) f(x_n) \left[F(x_n) - F(x_1)\right]^{n-2},$$

where $x_n \geq x_1$ and $f(x)$ is the probability density function of X. To determine the probability distribution of the sample range W, we consider the transformation

$$\left.\begin{aligned} U &= X_{(1)} \\ W &= X_{(n)} - X_{(1)} \end{aligned}\right\}$$

which has an inverse

$$\left.\begin{aligned} X_{(1)} &= U \\ X_{(n)} &= U + W. \end{aligned}\right\}$$

The Jacobian of this transformation is

$$J = det \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = 1.$$

Hence the joint density of (U, W) is given by

$$
\begin{aligned}
g(u, w) &= |J|\, h(x_1, x_n) \\
&= n(n-1) f(u) f(u+w) [F(u+w) - F(u)]^{n-2}
\end{aligned}
$$

where $w \geq 0$. Since $f(u)$ and $f(u+w)$ are simultaneously nonzero if $0 \leq u \leq 1$ and $0 \leq u + w \leq 1$. Hence $f(u)$ and $f(u + w)$ are simultaneously nonzero if $0 \leq u \leq 1 - w$. Thus, the probability of W is given by

$$\begin{aligned} j(w) &= \int_{-\infty}^{\infty} g(u, w)\, du \\ &= \int_{-\infty}^{\infty} n(n-1) f(u) f(u+w) [F(u+w) - F(u)]^{n-2}\, du \\ &= \int_{0}^{1-w} w^{n-2} du \\ &= n(n-1)(1-w)\, w^{n-2} \end{aligned}$$

where $0 \leq w \leq 1$.

13.5. Sample Percentiles

The sample median, M, is a number such that approximately one-half of the observations are less than M and one-half are greater than M.

Definition 13.5. Let $X_1, X_2, ..., X_n$ be a random sample. The sample median M is defined as

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}\right] & \text{if } n \text{ is even.} \end{cases}$$

The median is a measure of location like sample mean.

Recall that for continuous distribution, $100p^{\text{th}}$ percentile, π_p, is a number such that

$$p = \int_{-\infty}^{\pi_p} f(x)\, dx.$$

Definition 13.6. The $100p^{\text{th}}$ sample percentile is defined as

$$\pi_p = \begin{cases} X_{([np])} & \text{if } p < 0.5 \\ X_{(n+1-[n(1-p)])} & \text{if } p > 0.5. \end{cases}$$

where $[b]$ denote the number b rounded to the nearest integer.

Example 13.22. Let $X_1, X_2, ..., X_{12}$ be a random sample of size 12. What is the 65^{th} percentile of this sample?

Answer:

$$
\begin{aligned}
100p &= 65 \\
p &= 0.65 \\
n(1-p) &= (12)(1-0.65) = 4.2 \\
[n(1-p)] &= [4.2] = 4
\end{aligned}
$$

Hence by definition of 65^{th} percentile is

$$
\begin{aligned}
\pi_{0.65} &= X_{(n+1-[n(1-p)])} \\
&= X_{(13-4)} \\
&= X_{(9)}.
\end{aligned}
$$

Thus, the 65^{th} percentile of the random sample $X_1, X_2, ..., X_{12}$ is the 9^{th}-order statistic.

For any number p between 0 and 1, the $100p^{\text{th}}$ sample percentile is an observation such that approximately np observations are less than this observation and $n(1-p)$ observations are greater than this.

Definition 13.7. The 25^{th} percentile is called the lower quartile while the 75^{th} percentile is called the upper quartile. The distance between these two quartiles is called the interquartile range.

Example 13.23. If a sample of size 3 from a uniform distribution over $[0, 1]$ is observed, what is the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$?

Answer: When a sample of $(2n+1)$ random variables are observed, the $(n+1)^{\text{th}}$ smallest random variable is called the sample median. For our problem, the sample median is given by

$$X_{(2)} = 2^{\text{nd}} \text{ smallest } \{X_1, X_2, X_3\}.$$

Let $Y = X_{(2)}$. The density function of each X_i is given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the cumulative density function of $f(x)$ is

$$F(x) = x.$$

Thus the density function of Y is given by

$$\begin{aligned} g(y) &= \frac{3!}{1!\,1!}\ [F(y)]^{2-1}\ f(y)\ [1-F(y)]^{3-2} \\ &= 6\,F(y)\,f(y)\ [1-F(y)] \\ &= 6y\,(1-y). \end{aligned}$$

Therefore

$$\begin{aligned} P\left(\frac{1}{4} < Y < \frac{3}{4}\right) &= \int_{\frac{1}{4}}^{\frac{3}{4}} g(y)\,dy \\ &= \int_{\frac{1}{4}}^{\frac{3}{4}} 6\,y\,(1-y)\,dy \\ &= 6\left[\frac{y^2}{2} - \frac{y^3}{3}\right]_{\frac{1}{4}}^{\frac{3}{4}} \\ &= \frac{11}{16}. \end{aligned}$$

13.6. Review Exercises

1. Suppose we roll a die 1000 times. What is the probability that the sum of the numbers obtained lies between 3000 and 4000?

2. Suppose Kathy flip a coin 1000 times. What is the probability she will get at least 600 heads?

3. At a certain large university the weight of the male students and female students are approximately normally distributed with means and standard deviations of 180, and 20, and 130 and 15, respectively. If a male and female are selected at random, what is the probability that the sum of their weights is less than 280?

4. Seven observations are drawn from a population with an unknown continuous distribution. What is the probability that the least and the greatest observations bracket the median?

5. If the random variable X has the density function

$$f(x) = \begin{cases} 2\,(1-x) & \text{for } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability that the larger of 2 independent observations of X will exceed $\frac{1}{2}$?

6. Let X_1, X_2, X_3 be a random sample from the uniform distribution on the interval (0, 1). What is the probability that the sample median is less than 0.4?

7. Let X_1, X_2, X_3, X_4, X_5 be a random sample from the uniform distribution on the interval $(0,\, \theta)$, where θ is unknown, and let X_{max} denote the largest observation. For what value of the constant k, the expected value of the random variable kX_{max} is equal to θ?

8. A random sample of size 16 is to be taken from a normal population having mean 100 and variance 4. What is the 90^{th} percentile of the distribution of the sample mean?

9. If the density function of a random variable X is given by

$$f(x) = \begin{cases} \frac{1}{2x} & \text{for } \frac{1}{e} < x < e \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the probability that one of the two independent observations of X is less than 2 and the other is greater than 1?

10. Five observations have been drawn independently and at random from a continuous distribution. What is the probability that the next observation will be less than all of the first 5?

11. Let the random variable X denote the length of time it takes to complete a mathematics assignment. Suppose the density function of X is given by

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is a positive constant that represents the minimum time to complete a mathematics assignment. If $X_1, X_2, ..., X_5$ is a random sample from this distribution. What is the expected value of $X_{(1)}$?

12. Let X and Y be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the probability density function of $W = \max\{X, Y\}$?

13. Let X and Y be two independent random variables with identical probability density function given by

$$f(x) = \begin{cases} \frac{3\,x^2}{\theta^3} & \text{for } 0 \leq x \leq \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

for some $\theta > 0$. What is the probability density function of $W = \min\{X, Y\}$?

14. Let $X_1, X_2, ..., X_n$ be a random sample from a uniform distribution on the interval from 0 to 5. What is the limiting moment generating function of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ as $n \to \infty$?

15. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal distribution with mean μ and variance 1. If the 75^{th} percentile of the statistic $W = \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$ is 28.24, what is the sample size n ?

16. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a Bernoulli distribution with probability of success $p = \frac{1}{2}$. What is the limiting distribution the sample mean $\overline{X}$?

17. Let $X_1, X_2, ..., X_{1995}$ be a random sample of size 1995 from a distribution with probability density function

$$f(x) = \frac{e^{-\lambda}\, \lambda^x}{x!} \qquad\qquad x = 0, 1, 2, 3, ..., \infty.$$

What is the distribution of $1995\overline{X}$?

18. Suppose $X_1, X_2, ..., X_n$ is a random sample from the uniform distribution on $(0, 1)$ and Z be the sample range. What is the probability that Z is less than or equal to 0.5?

19. Let $X_1, X_2, ..., X_9$ be a random sample from a uniform distribution on the interval $[1, 12]$. Find the probability that the next to smallest is greater than or equal to 4?

20. A machine needs 4 out of its 6 independent components to operate. Let $X_1, X_2, ..., X_6$ be the lifetime of the respective components. Suppose each is exponentially distributed with parameter θ. What is the probability density function of the machine lifetime?

21. Suppose $X_1, X_2, ..., X_{2n+1}$ is a random sample from the uniform distribution on $(0, 1)$. What is the probability density function of the sample median $X_{(n+1)}$?

22. Let X and Y be two random variables with joint density

$$f(x,\, y) = \begin{cases} 12x & \text{if } 0 < y < 2x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the expected value of the random variable $Z = X^2 Y^3 + X^2 - X - Y^3$?

23. Let $X_1, X_2, ..., X_{50}$ be a random sample of size 50 from a distribution with density

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\,\theta^\alpha}\, x^{\alpha-1} e^{-\frac{x}{\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What are the mean and variance of the sample mean $\overline{X}$?

24. Let $X_1, X_2, ..., X_{100}$ be a random sample of size 100 from a distribution with density

$$f(x) = \begin{cases} \frac{e^{-\lambda}\,\lambda^x}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that $\overline{X}$ greater than or equal to 1?

Chapter 14

SAMPLING DISTRIBUTIONS ASSOCIATED WITH THE NORMAL POPULATIONS

Given a random sample $X_1, X_2, ..., X_n$ from a population X with probability distribution $f(x; \theta)$, where θ is a parameter, a *statistic* is a function T of $X_1, X_2, ..., X_n$, that is

$$T = T\left(X_1, X_2, ..., X_n\right)$$

which is free of the parameter θ. If the distribution of the population is known, then sometimes it is possible to find the probability distribution of the statistic T. The probability distribution of the statistic T is called the sampling distribution of T. The joint distribution of the random variables $X_1, X_2, ..., X_n$ is called the distribution of the sample. The distribution of the sample is the joint density

$$f(x_1, x_2, ..., x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

since the random variables $X_1, X_2, ..., X_n$ are independent and identically distributed.

Since the normal population is very important in statistics, the sampling distributions associated with the normal population are very important. The most important sampling distributions which are associated with the normal population are the followings: the chi-square distribution, the student's t-distribution, the F-distribution, and the beta distribution. In this chapter, we only consider the first three distributions, since the last distribution was considered earlier.

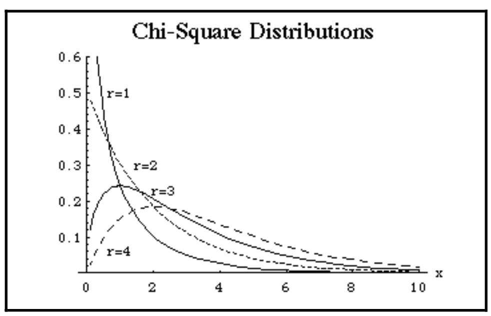
14.1. Chi-square distribution

In this section, we treat the Chi-square distribution, which is one of the very useful sampling distributions.

Definition 14.1. A continuous random variable X is said to have a chi-square distribution with r degrees of freedom if its probability density function is of the form

$$f(x;\, r) = \begin{cases} \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}}\, x^{\frac{r}{2}-1}\, e^{-\frac{x}{2}} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $r > 0$. If X has chi-square distribution, then we denote it by writing $X \sim \chi^2(r)$. Recall that a gamma distribution reduces to chi-square distribution if $\alpha = \frac{r}{2}$ and $\theta = 2$. The mean and variance of X are r and $2r$, respectively.



Thus, chi-square distribution is also a special case of gamma distribution. Further, if $r \to \infty$, then chi-square distribution tends to normal distribution.

Example 14.1. If $X \sim GAM(1, 1)$, then what is the probability density function of the random variable $2X$?

Answer: We will use the moment generating method to find the distribution of $2X$. The moment generating function of a gamma random variable is given by

$$M(t) = (1 - \theta\, t)^{-\alpha}, \qquad \text{if} \qquad t < \frac{1}{\theta}.$$

Since $X \sim GAM(1,1)$, the moment generating function of X is given by

$$M_X(t) = \frac{1}{1-t}, \qquad t < 1.$$

Hence, the moment generating function of $2X$ is

$$\begin{aligned} M_{2X}(t) &= M_X(2t) \\ &= \frac{1}{1-2t} \\ &= \frac{1}{(1-2t)^{\frac{2}{2}}} \\ &= \text{MGF of } \chi^2(2). \end{aligned}$$

Hence, if X is $GAM(1,1)$ or is an exponential with parameter 1, then $2X$ is chi-square with 2 degrees of freedom.

Example 14.2. If $X \sim \chi^2(5)$, then what is the probability that X is between 1.145 and 12.83?

Answer: The probability of X between 1.145 and 12.83 can be calculated from the following:

$$\begin{aligned} &P(1.145 \leq X \leq 12.83) \\ &\quad = P(X \leq 12.83) - P(X \leq 1.145) \\ &\quad = \int_0^{12.83} f(x)\,dx - \int_0^{1.145} f(x)\,dx \\ &\quad = \int_0^{12.83} \frac{1}{\Gamma\left(\frac{5}{2}\right)\, 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\, dx - \int_0^{1.145} \frac{1}{\Gamma\left(\frac{5}{2}\right)\, 2^{\frac{5}{2}}}\, x^{\frac{5}{2}-1}\, e^{-\frac{x}{2}}\, dx \\ &\quad = 0.975 - 0.050 \qquad (\text{from } \chi^2 \text{ table}) \\ &\quad = 0.925. \end{aligned}$$

This above integrals are hard to evaluate and thus their values are taken from the chi-square table.

Example 14.3. If $X \sim \chi^2(7)$, then what are values of the constants a and b such that $P(a < X < b) = 0.95$?

Answer: Since

$$0.95 = P(a < X < b) = P(X < b) - P(X < a),$$

we get

$$P(X < b) = 0.95 + P(X < a).$$

We choose $a = 1.690$, so that

$$P(X < 1.690) = 0.025.$$

From this, we get

$$P(X < b) = 0.95 + 0.025 = 0.975$$

Thus, from chi-square table, we get $b = 16.01$.

The following theorems were studied earlier in Chapters 6 and 13 and they are very useful in finding the sampling distributions of many statistics. We state these theorems here for the convenience of the reader.

Theorem 14.1. If $X \sim N(\mu, \sigma^2)$, then $\left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$.

Theorem 14.2. If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ is a random sample from the population X, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).$$

Theorem 14.3. If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ is a random sample from the population X, then

$$\frac{(n-1)\, S^2}{\sigma^2} \sim \chi^2(n-1).$$

Theorem 14.4. If $X \sim GAM(\theta, \alpha)$, then

$$\frac{2}{\theta}\, X \sim \chi^2(2\alpha).$$

Example 14.4. A new component is placed in service and n spares are available. If the times to failure in days are independent exponential variable, that is $X_i \sim EXP(100)$, how many spares would be needed to be 95% sure of successful operation for at least two years ?

Answer: Since $X_i \sim EXP(100)$,

$$\sum_{i=1}^{n} X_i \sim GAM(100, n).$$

Hence, by Theorem 14.4, the random variable

$$Y = \frac{2}{100} \sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

We have to find the number of spares n such that

$$\begin{aligned}
0.95 &= P\left(\sum_{i=1}^{n} X_i \geq 2 \text{ years}\right) \\
&= P\left(\sum_{i=1}^{n} X_i \geq 730 \text{ days}\right) \\
&= P\left(\frac{2}{100} \sum_{i=1}^{n} X_i \geq \frac{2}{100} 730 \text{ days}\right) \\
&= P\left(\frac{2}{100} \sum_{i=1}^{n} X_i \geq \frac{730}{50}\right) \\
&= P\left(\chi^2(2n) \geq 14.6\right). \\
2n &= 25 \qquad\qquad (\text{from } \chi^2 \text{ table})
\end{aligned}$$

Hence $n = 13$ (after rounding up to the next integer). Thus, 13 spares are needed to be 95% sure of successful operation for at least two years.

Example 14.5. If $X \sim N(10, 25)$ and $X_1, X_2, ..., X_{501}$ is a random sample of size 501 from the population X, then what is the expected value of the sample variance S^2 ?

Answer: We will use the Theorem 14.3, to do this problem. By Theorem 14.3, we see that

$$\frac{(501-1)\, S^2}{\sigma^2} \sim \chi^2(500).$$

Hence, the expected value of S^2 is given by

$$\begin{aligned}
E\left[S^2\right] &= E\left[\left(\frac{25}{500}\right)\left(\frac{500}{25}\right) S^2\right] \\
&= \left(\frac{25}{500}\right) E\left[\left(\frac{500}{25}\right) S^2\right] \\
&= \left(\frac{1}{20}\right) E\left[\chi^2(500)\right] \\
&= \left(\frac{1}{20}\right) 500 \\
&= 25.
\end{aligned}$$

14.2. Student's t-distribution

Here we treat the Student's t-distribution, which is also one of the very useful sampling distributions.

Definition 14.2. A continuous random variable X is said to have a t-distribution with ν degrees of freedom if its probability density function is of the form

$$f(x;\,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\,\nu}\;\Gamma\left(\frac{\nu}{2}\right)\,\left(1+\frac{x^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, \qquad -\infty < x < \infty$$

where $\nu > 0$. If X has a t-distribution with ν degrees of freedom, then we denote it by writing $X \sim t(\nu)$.

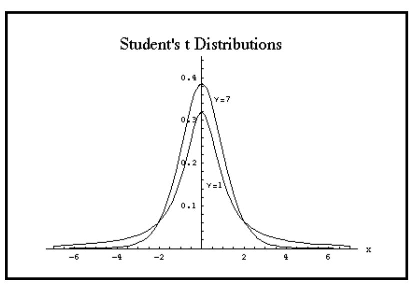
The t-distribution was discovered by W.S. Gosset (1876-1936) of England who published his work under the pseudonym of student. Therefore, this distribution is known as Student's t-distribution. This distribution is a generalization of the Cauchy distribution and the normal distribution. That is, if $\nu = 1$, then the probability density function of X becomes

$$f(x;\,1) = \frac{1}{\pi\,(1+x^2)} \qquad -\infty < x < \infty,$$

which is the Cauchy distribution. Further, if $\nu \to \infty$, then

$$\lim_{\nu\to\infty} f(x;\,\nu) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}x^2} \qquad -\infty < x < \infty,$$

which is the probability density function of the standard normal distribution. The following figure shows the graph of t-distributions with various degrees of freedom.



Example 14.6. If $T \sim t(10)$, then what is the probability that T is at least 2.228 ?

Answer: The probability that T is at least 2.228 is given by

$$\begin{aligned} P(T \geq 2.228) &= 1 - P(T < 2.228) \\ &= 1 - 0.975 \qquad \text{(from t} - \text{table)} \\ &= 0.025. \end{aligned}$$

Example 14.7. If $T \sim t(19)$, then what is the value of the constant c such that $P(|T| \leq c) = 0.95$?

Answer:

$$\begin{aligned} 0.95 &= P(|T| \leq c) \\ &= P(-c \leq T \leq c) \\ &= P(T \leq c) - 1 + P(T \leq c) \\ &= 2\, P(T \leq c) - 1. \end{aligned}$$

Hence

$$P(T \leq c) = 0.975.$$

Thus, using the t-table, we get for 19 degrees of freedom

$$c = 2.093.$$

Theorem 14.5. If the random variable X has a t-distribution with ν degrees of freedom, then

$$E[X] = \begin{cases} 0 & \text{if} \quad \nu \geq 2 \\ DNE & \text{if} \quad \nu = 1 \end{cases}$$

and

$$Var[X] = \begin{cases} \frac{\nu}{\nu-2} & \text{if} \quad \nu \geq 3 \\ DNE & \text{if} \quad \nu = 1, 2 \end{cases}$$

where DNE means does not exist.

Theorem 14.6. If $Z \sim N(0,1)$ and $U \sim \chi^2(\nu)$ and in addition, Z and U are independent, then the random variable W defined by

$$W = \frac{Z}{\sqrt{\frac{U}{\nu}}}$$

has a t-distribution with ν degrees of freedom.

Theorem 14.7. If $X \sim N(\mu, \sigma^2)$ and $X_1, X_2, ..., X_n$ be a random sample from the population X, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1).$$

Proof: Since each $X_i \sim N(\mu, \sigma^2)$,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Thus,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Further, from Theorem 14.3 we know that

$$(n-1)\,\frac{S^2}{\sigma^2} \sim \chi^2(n-1).$$

Hence

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma^2}{\sqrt{n}}}}{\sqrt{\frac{(n-1)\,S^2}{(n-1)\,\sigma^2}}} \sim t(n-1) \qquad \text{(by Theorem 14.6).}$$

This completes the proof of the theorem.

Example 14.8. Let X_1, X_2, X_3, X_4 be a random sample of size 4 from a standard normal distribution. If the statistic W is given by

$$W = \frac{X_1 - X_2 + X_3}{\sqrt{X_1^2 + X_2^2 + X_3^2 + X_4^2}},$$

then what is the expected value of W ?

Answer: Since $X_i \sim N(0,1)$, we get

$$X_1 - X_2 + X_3 \sim N(0,3)$$

and

$$\frac{X_1 - X_2 + X_3}{\sqrt{3}} \sim N(0,1).$$

Further, since $X_i \sim N(0,1)$, we have

$$X_i^2 \sim \chi^2(1)$$

and hence

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4)$$

Thus,

$$\frac{\frac{X_1 - X_2 + X_3}{\sqrt{3}}}{\sqrt{\frac{X_1^2+X_2^2+X_3^2+X_4^2}{4}}} = \left(\frac{2}{\sqrt{3}}\right) W \sim t(4).$$

Now using the distribution of W, we find the expected value of W.

$$\begin{aligned} E\left[W\right] &= \left(\frac{\sqrt{3}}{2}\right) E\left[\frac{2}{\sqrt{3}} W\right] \\ &= \left(\frac{\sqrt{3}}{2}\right) E\left[t(4)\right] \\ &= \left(\frac{\sqrt{3}}{2}\right) 0 \\ &= 0. \end{aligned}$$

Example 14.9. If $X \sim N(0,1)$ and X_1, X_2 is random sample of size two from the population X, then what is the 75$^{\text{th}}$ percentile of the statistic $W = \frac{X_1}{\sqrt{X_2^2}}$?

Answer: Since each $X_i \sim N(0,1)$, we have

$$\begin{aligned} X_1 &\sim N(0,1) \\ X_2^2 &\sim \chi^2(1). \end{aligned}$$

Hence

$$W = \frac{X_1}{\sqrt{X_2^2}} \sim t(1).$$

The 75$^{\text{th}}$ percentile a of W is then given by

$$0.75 = P(W \leq a)$$

Hence, from the t-table, we get

$$a = 1.0$$

Hence the 75$^{\text{th}}$ percentile of W is 1.0.

Example 14.10. Suppose $X_1, X_2, ..., X_n$ is a random sample from a normal distribution with mean μ and variance σ^2. If $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $V^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, and X_{n+1} is an additional observation, what is the value of m so that the statistics $\frac{m(\overline{X}-X_{n+1})}{V}$ has a t-distribution.

Answer: Since

$$\begin{aligned}
& X_i \sim N(\mu, \sigma^2) \\
\Rightarrow\ & \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \\
\Rightarrow\ & \overline{X} - X_{n+1} \sim N\left(\mu - \mu, \frac{\sigma^2}{n} + \sigma^2\right) \\
\Rightarrow\ & \overline{X} - X_{n+1} \sim N\left(0, \left(\frac{n+1}{n}\right)\sigma^2\right) \\
\Rightarrow\ & \frac{\overline{X} - X_{n+1}}{\sigma\sqrt{\frac{n+1}{n}}} \sim N(0,1)
\end{aligned}$$

Now, we establish a relationship between V^2 and S^2. We know that

$$\begin{aligned}
(n-1)\,S^2 &= (n-1)\,\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= n\,\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= n\,V^2.
\end{aligned}$$

Hence, by Theorem 14.3

$$\frac{n\,V^2}{\sigma^2} = \frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2(n-1).$$

Thus

$$\left(\sqrt{\frac{n-1}{n+1}}\right)\frac{\overline{X} - X_{n+1}}{V} = \frac{\frac{\overline{X}-X_{n+1}}{\sigma\sqrt{\frac{n+1}{n}}}}{\sqrt{\frac{\frac{n\,V^2}{\sigma^2}}{(n-1)}}} \sim t(n-1).$$

Thus by comparison, we get

$$m = \sqrt{\frac{n-1}{n+1}}.$$

14.3. Snedecor's F-distribution

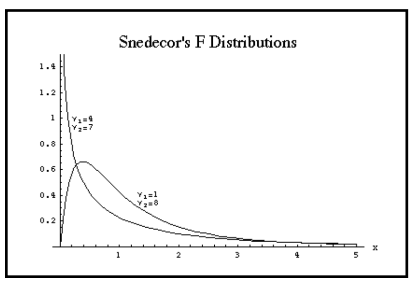
The next sampling distribution to be discussed in this chapter is Snedecor's F-distribution. This distribution has many applications in mathematical statistics. In the analysis of variance, this distribution is used to develop the technique for testing the equalities of sample means.

Definition 14.3. A continuous random variable X is said to have a F-distribution with ν_1 and ν_2 degrees of freedom if its probability density function is of the form

$$f(x;\, \nu_1, \nu_2) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}\,x\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}} & \text{if } 0 \leq x < \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $\nu_1,\ \nu_2 > 0$. If X has a F-distribution with ν_1 and ν_2 degrees of freedom, then we denote it by writing $X \sim F\,(\nu_1, \nu_2)$.

The F-distribution was named in honor of Sir Ronald Fisher by George Snedecor. F-distribution arises as the distribution of a ratio of variances. Like, the other two distributions this distribution also tends to normal distribution as ν_1 and ν_2 become very large. The following figure illustrates the shape of the graph of this distribution for various degrees of freedom.



The following theorem gives us the mean and variance of Snedecor's F-distribution.

Theorem 14.8. If the random variable $X \sim F\,(\nu_1, \nu_2)$, then

$$E[X] = \begin{cases} \frac{\nu_2}{\nu_2-2} & \text{if} \quad \nu_2 \geq 3 \\ \\ DNE & \text{if} \quad \nu_2 = 1, 2 \end{cases}$$

and

$$Var[X] = \begin{cases} \frac{2\,\nu_2^2\,(\nu_1+\nu_2-2)}{\nu_1\,(\nu_2-2)^2\,(\nu_2-4)} & \text{if} \quad \nu_2 \geq 5 \\ \\ DNE & \text{if} \quad \nu_2 = 1, 2, 3, 4. \end{cases}$$

Here DNE means does not exist.

Example 14.11. If $X \sim F(9, 10)$, what $P(X \geq 3.02)$? Also, find the mean and variance of X.

Answer:

$$\begin{aligned} P(X \geq 3.02) &= 1 - P(X \leq 3.02) \\ &= 1 - P(F(9,10) \leq 3.02) \\ &= 1 - 0.95 \qquad (\text{from F} - \text{table}) \\ &= 0.05. \end{aligned}$$

Next, we determine the mean and variance of X using the Theorem 14.8. Hence,

$$E(X) = \frac{\nu_2}{\nu_2 - 2} = \frac{10}{10-2} = \frac{10}{8} = 1.25$$

and

$$\begin{aligned} Var(X) &= \frac{2\,\nu_2^2\,(\nu_1 + \nu_2 - 2)}{\nu_1\,(\nu_2 - 2)^2\,(\nu_2 - 4)} \\ &= \frac{2\,(10)^2\,(19-2)}{9\,(8)^2\,(6)} \\ &= \frac{(25)\,(17)}{(27)\,(16)} \\ &= \frac{425}{432} = 0.9838. \end{aligned}$$

Theorem 14.9. If $X \sim F(\nu_1, \nu_2)$, then the random variable $\frac{1}{X} \sim F(\nu_2, \nu_1)$.

This theorem is very useful for computing probabilities like $P(X \leq 0.2439)$. If you look at a F-table, you will notice that the table start with values bigger than 1. Our next example illustrates how to find such probabilities using Theorem 14.9.

Example 14.12. If $X \sim F(6, 9)$, what is the probability that X is less than or equal to 0.2439 ?

Answer: We use the above theorem to compute

$$
\begin{aligned}
P\left(X \leq 0.2439\right) &= P\left(\frac{1}{X} \geq \frac{1}{0.2439}\right) \\
&= P\left(F(9,6) \geq \frac{1}{0.2439}\right) \qquad \text{(by Theorem 14.9)} \\
&= 1 - P\left(F(9,6) \leq \frac{1}{0.2439}\right) \\
&= 1 - P\left(F(9,6) \leq 4.10\right) \\
&= 1 - 0.95 \\
&= 0.05.
\end{aligned}
$$

The following theorem says that F-distribution arises as the distribution of a random variable which is the quotient of two independently distributed chi-square random variables, each of which is divided by its degrees of freedom.

Theorem 14.10. If $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, and the random variables U and V are independent, then the random variable

$$\frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F\left(\nu_1, \nu_2\right).$$

Example 14.13. Let $X_1, X_2, ..., X_4$ and $Y_1, Y_2, ..., Y_5$ be two random samples of size 4 and 5 respectively, from a standard normal population. What is the variance of the statistic $T = \left(\frac{5}{4}\right) \frac{X_1^2+X_2^2+X_3^2+X_4^2}{Y_1^2+Y_2^2+Y_3^2+Y_4^2+Y_5^2}$?

Answer: Since the population is standard normal, we get

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4).$$

Similarly,

$$Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2 \sim \chi^2(5).$$

Thus

$$
\begin{aligned}
T &= \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2} \\
&= \frac{\frac{X_1^2+X_2^2+X_3^2+X_4^2}{4}}{\frac{Y_1^2+Y_2^2+Y_3^2+Y_4^2+Y_5^2}{5}} \\
&= T \sim F(4,5).
\end{aligned}
$$

Therefore

$$\begin{aligned} Var(T) &= Var[\,F(4,5)\,] \\ &= \frac{2\,(5)^2\,(7)}{4\,(3)^2\,(1)} \\ &= \frac{350}{36} \\ &= 9.72. \end{aligned}$$

Theorem 14.11. Let $X \sim N(\mu_1,\,\sigma_1^2)$ and $X_1, X_2, ..., X_n$ be a random sample of size n from the population X. Let $Y \sim N(\mu_2,\,\sigma_2^2)$ and $Y_1, Y_2, ..., Y_m$ be a random sample of size m from the population Y. Then the statistic

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n-1,\, m-1),$$

where S_1^2 and S_2^2 denote the sample variances of the first and the second sample, respectively.

Proof: Since,

$$X_i \sim N(\mu_1, \sigma_1^2)$$

we have by Theorem 14.3, we get

$$(n-1)\,\frac{S_1^2}{\sigma_1^2} \sim \chi^2(n-1).$$

Similarly, since

$$Y_i \sim N(\mu_2, \sigma_2^2)$$

we have by Theorem 14.3, we get

$$(m-1)\,\frac{S_2^2}{\sigma_2^2} \sim \chi^2(m-1).$$

Therefore

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{(n-1)\,S_1^2}{(n-1)\,\sigma_1^2}}{\frac{(m-1)\,S_2^2}{(m-1)\,\sigma_2^2}} \sim F(n-1,\, m-1).$$

This completes the proof of the theorem.

Because of this theorem, the F-distribution is also known as the variance-ratio distribution.

14.4. Review Exercises

1. Let $X_1, X_2, ..., X_5$ be a random sample of size 5 from a normal distribution with mean zero and standard deviation 2. Find the sampling distribution of the statistic $X_1 + 2X_2 - X_3 + X_4 + X_5$.

2. Let X_1, X_2, X_3 be a random sample of size 3 from a standard normal distribution. Find the distribution of $X_1^2 + X_2^2 + X_3^2$. If possible, find the sampling distribution of $X_1^2 - X_2^2$. If not, justify why you can not determine it's distribution.

3. Let X_1, X_2, X_3 be a random sample of size 3 from a standard normal distribution. Find the sampling distribution of the statistics $\frac{X_1+X_2+X_3}{\sqrt{X_1^2+X_2^2+X_3^2}}$ and $\frac{X_1-X_2-X_3}{\sqrt{X_1^2+X_2^2+X_3^2}}$.

4. Let X_1, X_2, X_3 be a random sample of size 3 from an exponential distribution with a parameter $\theta > 0$. Find the distribution of the sample (that is the joint distribution of the random variables X_1, X_2, X_3).

5. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal population with mean μ and variance $\sigma^2 > 0$. What is the expected value of the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$?

6. Let X_1, X_2, X_3, X_4 be a random sample of size 4 from a standard normal population. Find the distribution of the statistic $\frac{X_1+X_4}{\sqrt{X_2^2+X_3^2}}$.

7. Let X_1, X_2, X_3, X_4 be a random sample of size 4 from a standard normal population. Find the sampling distribution (if possible) and moment generating function of the statistic $2X_1^2+3X_2^2+X_3^2+4X_4^2$. What is the probability distribution of the sample?

8. Let X equal the maximal oxygen intake of a human on a treadmill, where the measurement are in milliliters of oxygen per minute per kilogram of weight. Assume that for a particular population the mean of X is $\mu = 54.03$ and the standard deviation is $\sigma = 5.8$. Let $\bar{X}$ be the sample mean of a random sample $X_1, X_2, ..., X_{47}$ of size 47 drawn from X. Find the probability that the sample mean is between 52.761 and 54.453.

9. Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. What is the variance of $V^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$?

10. If X is a random variable with mean μ and variance σ^2, then $\mu - 2\sigma$ is called the lower 2σ point of X. Suppose a random sample X_1, X_2, X_3, X_4 is drawn from a chi-square distribution with two degrees of freedom. What is the lower 2σ point of $X_1 + X_2 + X_3 + X_4$?

11. Let X and Y be independent normal random variables such that the mean and variance of X are 2 and 4, respectively, while the mean and variance of Y are 6 and k, respectively. A sample of size 4 is taken from the X-distribution and a sample of size 9 is taken from the Y-distribution. If $P\left(\overline{Y} - \overline{X} > 8\right) = 0.0228$, then what is the value of the constant k ?

12. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with density function

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the distribution of the statistic $Y = 2\lambda \sum_{i=1}^{n} X_i$?

13. Suppose X has a normal distribution with mean 0 and variance 1, Y has a chi-square distribution with n degrees of freedom, W has a chi-square distribution with p degrees of freedom, and W, X, and Y are independent. What is the sampling distribution of the statistic $V = \frac{X}{\sqrt{\frac{W+Y}{p+n}}}$?

14. A random sample $X_1, X_2, ..., X_n$ of size n is selected from a normal population with mean μ and standard deviation 1. Later an additional independent observation X_{n+1} is obtained from the same population. What is the distribution of the statistic $(X_{n+1} - \mu)^2 + \sum_{i=1}^{n}(X_i - \overline{X})^2$, where $\overline{X}$ denote the sample mean?

15. Let $T = \frac{k(X+Y)}{\sqrt{Z^2+W^2}}$, where X, Y, Z, and W are independent normal random variables with mean 0 and variance $\sigma^2 > 0$. For exactly one value of k, T has a t-distribution. If r denotes the degrees of freedom of that distribution, then what is the value of the pair (k, r)?

16. Let X and Y be joint normal random variables with common mean 0, common variance 1, and covariance $\frac{1}{2}$. What is the probability of the event $\left(X + Y \leq \sqrt{3}\right)$, that is $P\left(X + Y \leq \sqrt{3}\right)$?

17. Suppose $X_j = Z_j - Z_{j-1}$, where $j = 1, 2, ..., n$ and $Z_0, Z_1, ..., Z_n$ are independent and identically distributed with common variance σ^2. What is the variance of the random variable $\frac{1}{n}\sum_{j=1}^{n} X_j$?

18. A random sample of size 5 is taken from a normal distribution with mean 0 and standard deviation 2. Find the constant k such that 0.05 is equal to the probability that the sum of the squares of the sample observations exceeds the constant k.

19. Let $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_n$ be two random sample from the independent normal distributions with $Var[X_i] = \sigma^2$ and $Var[Y_i] = 2\sigma^2$, for $i = 1, 2, ..., n$ and $\sigma^2 > 0$. If $U = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ and $V = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$, then what is the sampling distribution of the statistic $\frac{2U+V}{2\sigma^2}$?

20. Suppose $X_1, X_2, ..., X_6$ and $Y_1, Y_2, ..., Y_9$ are independent, identically distributed normal random variables, each with mean zero and variance $\sigma^2 > 0$. What is the 95$^{\text{th}}$ percentile of the statistics $W = \left[\sum_{i=1}^{6} X_i^2\right] / \left[\sum_{j=1}^{9} Y_j^2\right]$?

21. Let $X_1, X_2, ..., X_6$ and $Y_1, Y_2, ..., Y_8$ be independent random samples from a normal distribution with mean 0 and variance 1, and $Z = \left[4\sum_{i=1}^{6} X_i^2\right] / \left[3\sum_{j=1}^{8} Y_j^2\right]$?

Chapter 15

SOME TECHNIQUES FOR FINDING POINT ESTIMATORS OF PARAMETERS

A statistical population consists of all the measurements of interest in a statistical investigation. Usually a population is described by a random variable X. If we can gain some knowledge about the probability density function $f(x;\theta)$ of X, then we also gain some knowledge about the population under investigation.

A sample is a portion of the population usually chosen by method of random sampling and as such it is a set of random variables $X_1, X_2, ..., X_n$ with the same probability density function $f(x;\theta)$ as the population. Once the sampling is done, we get

$$X_1 = x_1, \; X_2 = x_2, \; \cdots, \; X_n = x_n$$

where $x_1, x_2, ..., x_n$ are the *sample data*.

Every statistical method employs a random sample to gain information about the population. Since the population is characterized by the probability density function $f(x;\theta)$, in statistics one makes statistical inferences about the population distribution $f(x;\theta)$ based on sample information. A statistical inference is a statement based on sample information about the population. There three types of statistical inferences such as: (1) the estimation (2) the hypothesis testing and (3) the prediction. The goal of this chapter is to examine some well known point estimation methods.

In point estimation, we try to find the parameter θ of the population distribution $f(x;\theta)$ from the sample information. Thus, in the parametric point estimation one assumes the functional form of the pdf $f(x;\theta)$ to be known and only estimate the unknown parameter θ of the population using information available from the sample.

Definition 15.1. Let X be a population with the density function $f(x;\theta)$, where θ is an unknown parameter. The set of all admissible values of θ is called a parameter space and it is denoted by Ω, that is

$$\Omega = \{\theta \in \mathbb{R}^n \mid f(x;\theta) \text{ is a pdf } \}$$

for some natural number m.

Example 15.1. If $X \sim EXP(\theta)$, then what is the parameter space of θ ?

Answer: Since $X \sim EXP(\theta)$, the density function of X is given by

$$f(x;\theta) = \frac{1}{\theta}\, e^{-\frac{x}{\theta}}.$$

If θ is zero or negative then $f(x;\theta)$ is not a density function. Thus, the admissible values of θ are all the positive real numbers. Hence

$$\begin{aligned} \Omega &= \{\theta \in \mathbb{R} \mid 0 < \theta < \infty\} \\ &= \mathbb{R}_+. \end{aligned}$$

Example 15.2. If $X \sim N\left(\mu, \sigma^2\right)$, what is the parameter space?

Answer: The parameter space Ω is given by

$$\begin{aligned} \Omega &= \left\{\theta \in \mathbb{R}^2 \mid f(x;\theta) \sim N\left(\mu, \sigma^2\right)\right\} \\ &= \left\{(\mu, \sigma) \in \mathbb{R}^2 \mid -\infty < \mu < \infty,\ 0 < \sigma < \infty \right\} \\ &= \mathbb{R} \times \mathbb{R}_+ \\ &= \text{upper half plane.} \end{aligned}$$

In general, a parameter space is a subset of $\mathbb{R}^m$. Statistics concerns with the estimation of the unknown parameter θ from a random sample $X_1, X_2, ..., X_n$. Recall that a statistic is a function of $X_1, X_2, ..., X_n$ and free of the population parameter θ.

Definition 15.2. Let $X \sim f(x;\theta)$ and $X_1, X_2, ..., X_n$ be a random sample from the population X. Any statistic that can be used to guess the parameter

θ is called an estimator of θ. The numerical value of this statistic is called an estimate of θ. The estimator of the parameter θ is denoted by $\widehat{\theta}$.

One of the basic problems is how to find an estimator of population parameter θ. There are several methods for finding an estimator of θ. Some of these methods are:

(1) Moment Method
(2) Maximum Likelihood Method
(3) Bayes Method
(4) Least Squares Method
(5) Minimum Chi-Squares Method
(6) Minimum Distance Method

In this chapter, we only discuss the first three methods of estimating a population parameter.

15.1. Moment Method

Let $X_1, X_2, ..., X_n$ be a random sample from a population X with probability density function $f(x; \theta_1, \theta_2, ..., \theta_m)$, where $\theta_1, \theta_2, ..., \theta_m$ are m unknown parameters. Let

$$E\left(X^k\right) = \int_{-\infty}^{\infty} x^k \, f(x; \theta_1, \theta_2, ..., \theta_m) \, dx$$

be the k^{th} population moment about 0. Further, let

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

be the k^{th} sample moment about 0.

In moment method, we find the estimator for the parameters $\theta_1, \theta_2, ..., \theta_m$ by equating the first m population moments (if they exist) to the first m sample moments, that is

$$\begin{aligned}
E\left(X\right) &= M_1 \\
E\left(X^2\right) &= M_2 \\
E\left(X^3\right) &= M_3 \\
&\vdots \\
E\left(X^m\right) &= M_m
\end{aligned}$$

The moment method is one of the classical methods for estimating parameters and motivation comes from the fact that the sample moments are in some sense estimates for the population moments. The moment method was first discovered by British statistician Karl Pearson in 1902. Now we provide some examples to illustrate this method.

Example 15.3. Let $X \sim N\left(\mu, \sigma^2\right)$ and $X_1, X_2, ..., X_n$ be a random sample of size n from the population X. What are the estimators of the population parameters μ and σ^2 if we use the moment method?

Answer: Since the population is normal, that is

$$X \sim N\left(\mu, \sigma^2\right)$$

we know that

$$\begin{aligned} E(X) &= \mu \\ E\left(X^2\right) &= \sigma^2 + \mu^2. \end{aligned}$$

Hence

$$\begin{aligned} \mu &= E(X) \\ &= M_1 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \overline{X}. \end{aligned}$$

Therefore, the estimator of the parameter μ is $\overline{X}$, that is

$$\widehat{\mu} = \overline{X}.$$

Next, we find the estimator of σ^2 equating $E(X^2)$ to M_2. Note that

$$\begin{aligned} \sigma^2 &= \sigma^2 + \mu^2 - \mu^2 \\ &= E\left(X^2\right) - \mu^2 \\ &= M_2 - \mu^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2. \end{aligned}$$

The last line follows from the fact that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2-2\,X_i\,\overline{X}+\overline{X}^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}2\,X_i\,\overline{X}+\frac{1}{n}\sum_{i=1}^{n}\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-2\,\overline{X}\,\frac{1}{n}\sum_{i=1}^{n}X_i+\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-2\,\overline{X}\,\overline{X}+\overline{X}^2\\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2.\end{aligned}$$

Thus, the estimator of σ^2 is $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, that is

$$\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2.$$

Example 15.4. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with probability density function

$$f(x;\theta)=\begin{cases}\theta\,x^{\theta-1} & \text{if } 0<x<1\\ \\ 0 & \text{otherwise,}\end{cases}$$

where $0<\theta<\infty$ is an unknown parameter. Using the method of moment find an estimator of θ ? If $x_1=0.2, x_2=0.6, x_3=0.5, x_4=0.3$ is a random sample of size 4, then what is the estimate of θ ?

Answer: To find an estimator, we shall equate the population moment to the sample moment. The population moment $E(X)$ is given by

$$\begin{aligned}E(X) &= \int_0^1 x\,f(x;\theta)\,dx\\ &= \int_0^1 x\,\theta\,x^{\theta-1}\,dx\\ &= \theta\int_0^1 x^{\theta}\,dx\\ &= \frac{\theta}{\theta+1}\,\left[x^{\theta+1}\right]_0^1\\ &= \frac{\theta}{\theta+1}.\end{aligned}$$

We know that $M_1 = \overline{X}$. Now setting M_1 equal to $E(X)$ and solving for θ, we get

$$\overline{X} = \frac{\theta}{\theta + 1}$$

that is

$$\theta = \frac{\overline{X}}{1 - \overline{X}},$$

where $\overline{X}$ is the sample mean. Thus, the statistic $\frac{\overline{X}}{1-\overline{X}}$ is an estimator of the parameter θ. Hence

$$\widehat{\theta} = \frac{\overline{X}}{1 - \overline{X}}.$$

Since $x_1 = 0.2, x_2 = 0.6, x_3 = 0.5, x_4 = 0.3$, we have $\overline{X} = 0.4$ and

$$\widehat{\theta} = \frac{0.4}{1 - 0.4} = \frac{2}{3}$$

is an estimate of the θ.

Example 15.5. What is the basic principle of the moment method?

Answer: To choose a value for the unknown population parameter for which the observed data have the same moments as the population.

Example 15.6. Suppose $X_1, X_2, ..., X_7$ is a random sample from a population X with density function

$$f(x;\beta) = \begin{cases} \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find an estimator of β by the moment method.

Answer: Since, we have only one parameter, we need to compute only the first population moment $E(X)$ about 0. Thus,

$$\begin{aligned} E(X) &= \int_0^\infty x\, f(x;\beta)\, dx \\ &= \int_0^\infty x\, \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7}\, dx \\ &= \frac{1}{\Gamma(7)} \int_0^\infty \left(\frac{x}{\beta}\right)^7 e^{-\frac{x}{\beta}}\, dx \\ &= \beta\, \frac{1}{\Gamma(7)} \int_0^\infty y^7\, e^{-y}\, dy \\ &= \beta\, \frac{1}{\Gamma(7)}\, \Gamma(8) \\ &= 7\,\beta. \end{aligned}$$

Since $M_1 = \overline{X}$, equating $E(X)$ to M_1, we get

$$7\,\beta = \overline{X}$$

that is

$$\beta = \frac{1}{7}\,\overline{X}.$$

Therefore, the estimator of β by the moment method is given by

$$\widehat{\beta} = \frac{1}{7}\,\overline{X}.$$

Example 15.7. Suppose $X_1, X_2, ..., X_n$ is a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find an estimator of θ by the moment method.

Answer: Examining the density function of the population X, we see that $X \sim UNIF(0,\theta)$. Therefore

$$E(X) = \frac{\theta}{2}.$$

Now, equating this population moment to the sample moment, we obtain

$$\frac{\theta}{2} = E(X) = M_1 = \overline{X}.$$

Therefore, the estimator of θ is

$$\widehat{\theta} = 2\,\overline{X}.$$

Example 15.8. Suppose $X_1, X_2, ..., X_n$ is a random sample from a population X with density function

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimators of α and β by the moment method.

Answer: Examining the density function of the population X, we see that $X \sim UNIF(\alpha, \beta)$. Since, the distribution has two unknown parameters, we need the first two population moments. Therefore

$$E(X) = \frac{\alpha + \beta}{2} \qquad \text{and} \qquad E(X^2) = \frac{(\beta - \alpha)^2}{12} + E(X)^2.$$

Equating these moments to the corresponding sample moments, we obtain

$$\frac{\alpha + \beta}{2} = E(X) = M_1 = \overline{X}$$

that is

$$\alpha + \beta = 2\overline{X} \tag{1}$$

and

$$\frac{(\beta - \alpha)^2}{12} + E(X)^2 = E(X^2) = M_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

which is

$$\begin{aligned}(\beta - \alpha)^2 &= 12 \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - E(X)^2 \right] \\ &= 12 \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \right] \\ &= 12 \left[\frac{1}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X} \right)^2 \right].\end{aligned}$$

Hence, we get

$$\beta - \alpha = \sqrt{ \frac{12}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X} \right)^2 }. \tag{2}$$

Adding equation (1) to equation (2), we obtain

$$2\beta = 2\overline{X} \pm 2 \sqrt{ \frac{3}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X} \right)^2 }$$

that is

$$\beta = \overline{X} \pm \sqrt{ \frac{3}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X} \right)^2 }.$$

Similarly, subtracting (2) from (1), we get

$$\alpha = \overline{X} \mp \sqrt{ \frac{3}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X} \right)^2 }.$$

Since, $\alpha < \beta$, we see that the estimators of α and β are

$$\widehat{\alpha} = \overline{X} - \sqrt{\frac{3}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2} \qquad \text{and} \qquad \widehat{\beta} = \overline{X} + \sqrt{\frac{3}{n} \sum_{i=1}^{n} \left(X_i^2 - \overline{X}\right)^2}.$$

15.2. Maximum Likelihood Method

The maximum likelihood method was first time used by Sir Ronald Fisher in 1912 for finding estimator of a unknown parameter. However, the method originated in the works of Gauss and Bernoulli. Next, we describe the method in details.

Definition 15.3. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with probability density function $f(x;\theta)$, where θ is an unknown parameter. The likelihood function, $L(\theta)$, is the distribution of the sample. That is

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

This definition says that the likelihood function of a random sample $X_1, X_2, ..., X_n$ is the joint density of the random variables $X_1, X_2, ..., X_n$.

The θ that maximizes the likelihood function $L(\theta)$ is called the maximum likelihood estimator of θ, and it is denoted by $\widehat{\theta}$. Hence

$$\widehat{\theta} = Arg \sup_{\theta \in \Omega} L(\theta),$$

where Ω is the parameter space of θ so that $L(\theta)$ is the joint density of the sample.

The method of maximum likelihood in a sense picks out of all the possible values of θ the one most likely to have produced the given observations $x_1, x_2, ..., x_n$. The method is summarized below: (1) Obtain a random sample $x_1, x_2, ..., x_n$ from the distribution of a population X with probability density function $f(x;\theta)$; (2) define the likelihood function for the sample $x_1, x_2, ..., x_n$ by $L(\theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$; (3) find the expression for θ that maximizes $L(\theta)$. This can be done directly or by maximizing $\ln L(\theta)$; (4) replace θ by $\widehat{\theta}$ to obtain an expression for the maximum likelihood estimator for θ; (5) find the observed value of this estimator for a given sample.

Example 15.9. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} (1-\theta)\, x^{-\theta} & \text{if } 0 < x < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

what is the maximum likelihood estimator of θ ?

Answer: The likelihood function of the sample is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Therefore

$$\begin{aligned} \ln L(\theta) &= \ln \left(\prod_{i=1}^{n} f(x_i;\theta) \right) \\ &= \sum_{i=1}^{n} \ln f(x_i;\theta) \\ &= \sum_{i=1}^{n} \ln \left[(1-\theta)\, {x_i}^{-\theta} \right] \\ &= n\, \ln(1-\theta) - \theta \sum_{i=1}^{n} \ln x_i. \end{aligned}$$

Now we maximize $\ln L(\theta)$ with respect to θ.

$$\begin{aligned} \frac{d \ln L(\theta)}{d\theta} &= \frac{d}{d\theta} \left(n\, \ln(1-\theta) - \theta \sum_{i=1}^{n} \ln x_i \right) \\ &= -\frac{n}{1-\theta} - \sum_{i=1}^{n} \ln x_i. \end{aligned}$$

Setting this derivative $\frac{d \ln L(\theta)}{d\theta}$ to 0, we get

$$\frac{d \ln L(\theta)}{d\theta} = -\frac{n}{1-\theta} - \sum_{i=1}^{n} \ln x_i = 0$$

that is

$$\frac{1}{1-\theta} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

or

$$\frac{1}{1-\theta} = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i = \overline{\ln x}.$$

or

$$\theta = 1 - \frac{1}{\overline{\ln x}}.$$

This θ can be shown to be maximum by the second derivative test and we leave this verification to the reader. Therefore, the estimator of θ is

$$\widehat{\theta} = 1 - \frac{1}{\overline{\ln X}}.$$

Example 15.10. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\beta) = \begin{cases} \frac{x^6\, e^{-\frac{x}{\beta}}}{\Gamma(7)\,\beta^7} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of β ?

Answer: The likelihood function of the sample is given by

$$L(\beta) = \prod_{i=1}^{n} f(x_i;\beta).$$

Thus,

$$\begin{aligned} \ln L(\beta) &= \sum_{i=1}^{n} \ln f(x_i, \beta) \\ &= 6\sum_{i=1}^{n} \ln x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i - n\ln(6!) - 7n\ln(\beta). \end{aligned}$$

Therefore

$$\frac{d}{d\beta}\ln L(\beta) = \frac{1}{\beta^2}\sum_{i=1}^{n} x_i - \frac{7n}{\beta}.$$

Setting this derivative $\frac{d}{d\beta}\ln L(\beta)$ to zero, we get

$$\frac{1}{\beta^2}\sum_{i=1}^{n} x_i - \frac{7n}{\beta} = 0$$

which yields

$$\beta = \frac{1}{7n}\sum_{i=1}^{n} x_i.$$

This β can be shown to be maximum by the second derivative test and again we leave this verification to the reader. Hence, the estimator of β is given by

$$\widehat{\beta} = \frac{1}{7}\,\overline{X}.$$

Remark 15.1. Note that this maximum likelihood estimator of β is same as the one found for β using the moment method in Example 15.6. However, in general the estimators by different methods are different as the following example illustrates.

Example 15.11. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of θ ?

Answer: The likelihood function of the sample is given by

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} \left(\frac{1}{\theta}\right) \qquad\qquad \theta > x_i \quad (i = 1, 2, 3, ..., n) \\ &= \left(\frac{1}{\theta}\right)^n \qquad\qquad \theta > \max\{x_1, x_2, ..., x_n\}. \end{aligned}$$

Hence the parameter space of θ with respect to $L(\theta)$ is given by

$$\Omega = \{\theta \in \mathbb{R} \mid x_{\max} < \theta < \infty\} = (x_{\max},\, \infty).$$

Now we maximize $L(\theta)$ on Ω. First, we compute $\ln L(\theta)$ and then differentiate it to get

$$\ln L(\theta) = -n \ln(\theta)$$

and

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{n}{\theta} < 0.$$

Therefore $\ln L(\theta)$ is a decreasing function of θ and as such the maximum of $\ln L(\theta)$ occurs at the left end point of the interval $(x_{\max},\, \infty)$. Therefore, at

$\theta = x_{\max}$ the likelihood function achieve maximum. Hence the likelihood estimator of θ is given by

$$\widehat{\theta} = X_{(n)}$$

where $X_{(n)}$ denotes the n^{th} order statistic of the given sample.

Thus, Example 15.7 and Example 15.11 say that the if we estimate the parameter θ of a distribution with uniform density on the interval $(0, \theta)$, then the maximum likelihood estimator is given by

$$\widehat{\theta} = X_{(n)}$$

where as

$$\widehat{\theta} = 2\,\overline{X}$$

is the estimator obtained by the method of moment. Hence, in general these two methods do not provide the same estimator of an unknown parameter.

Example 15.12. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\; e^{-\frac{1}{2}(x-\theta)^2} & \text{if } x \geq \theta \\ \\ 0 & \text{elsewhere.} \end{cases}$$

What is the maximum likelihood estimator of θ ?

Answer: The likelihood function $L(\theta)$ is given by

$$L(\theta) = \left(\sqrt{\frac{2}{\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{1}{2}(x_i-\theta)^2} \qquad\qquad x_i \geq \theta \;\; (i = 1, 2, 3, ..., n).$$

Hence the parameter space of θ is given by

$$\Omega = \{\theta \in \mathbb{R} \mid 0 \leq \theta \leq x_{\min}\} = [0, x_{\min}],,$$

where $x_{\min} = \min\{x_1, x_2, ..., x_n\}$. Now we evaluate the logarithm of the likelihood function.

$$\ln L(\theta) = \frac{n}{2}\, \ln\left(\frac{2}{\pi}\right) - \frac{1}{2}\, \sum_{i=1}^{n} (x_i - \theta)^2,$$

where θ is on the interval $[0,\, x_{\min}\,]$. Now we maximize $\ln L(\theta)$ subject to the condition $0 \leq \theta \leq x_{\min}$. Taking the derivative, we get

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)\, 2(-1) = \sum_{i=1}^{n} (x_i - \theta).$$

In this example, if we equate the derivative to zero, then we get $\theta = \overline{x}$. But this value of θ is not on the parameter space Ω. Thus, $\theta = \overline{x}$ is not the solution. Hence to find the solution of this optimization process, we examine the behavior of the $\ln L(\theta)$ on the interval $[0,\, x_{\min}\,]$. Note that

$$\frac{d}{d\theta} \ln L(\theta) = -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)\, 2(-1) = \sum_{i=1}^{n} (x_i - \theta) > 0$$

since each x_i is bigger than θ. Therefore, the function $\ln L(\theta)$ is an increasing function on the interval $[0,\, x_{\min}\,]$ and as such it will achieve maximum at the right end point of the interval $[0,\, x_{\min}\,]$. Therefore, the maximum likelihood estimator of θ is given by

$$\widehat{X} = X_{(1)}$$

where $X_{(1)}$ denotes the smallest observation in the random sample $X_1, X_2, ..., X_n$.

Example 15.13. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance σ^2. What are the maximum likelihood estimators of μ and σ^2 ?

Answer: Since $X \sim N(\mu, \sigma^2)$, the probability density function of X is given by

$$f(x; \mu, \sigma) = \frac{1}{\sigma\, \sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

The likelihood function of the sample is given by

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\, \sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}.$$

Hence, the logarithm of this likelihood function is given by

$$\ln L(\mu, \sigma) = -\frac{n}{2}\, \ln(2\pi) - n \ln(\sigma) - \frac{1}{2\sigma^2}\, \sum_{i=1}^{n} (x_i - \mu)^2.$$

Taking the partial derivatives of $\ln L(\mu, \sigma)$ with respect to μ and σ, we get

$$\frac{\partial}{\partial\mu} \ln L(\mu, \sigma) = -\frac{1}{2\sigma^2}\, \sum_{i=1}^{n} (x_i - \mu)\, (-2) = \frac{1}{\sigma^2}\, \sum_{i=1}^{n} (x_i - \mu).$$

and

$$\frac{\partial}{\partial\sigma} \ln L(\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\, \sum_{i=1}^{n} (x_i - \mu)^2.$$

Setting $\frac{\partial}{\partial \mu} \ln L(\mu, \sigma) = 0$ and $\frac{\partial}{\partial \sigma} \ln L(\mu, \sigma) = 0$, and solving for the unknown μ and σ, we get

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

Thus the maximum likelihood estimator of μ is

$$\widehat{\mu} = \overline{X}.$$

Similarly, we get

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

implies

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Again μ and σ^2 found by the first derivative test can be shown to be maximum using the second derivative test for the functions of two variables. Hence, using the estimator of μ in the above expression, we get the estimator of σ^2 to be

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Example 15.14. Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimators of α and β by the method of maximum likelihood.

Answer: The likelihood function of the sample is given by

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\beta - \alpha} = \left(\frac{1}{\beta - \alpha} \right)^n$$

for all $\alpha \leq x_i$ for $(i = 1, 2, ..., n)$ and for all $\beta \geq x_i$ for $(i = 1, 2, ..., n)$. Hence, the domain of the likelihood function is

$$\Omega = \{(\alpha, \beta) \mid 0 < \alpha \leq x_{(1)} \quad \text{and} \quad x_{(n)} \leq \beta < \infty\}.$$

It is easy to see that $L(\alpha, \beta)$ is maximum if $\alpha = x_{(1)}$ and $\beta = x_{(n)}$. Therefore, the maximum likelihood estimator of α and β are

$$\widehat{\alpha} = X_{(1)} \qquad \text{and} \qquad \widehat{\beta} = X_{(n)}.$$

The maximum likelihood estimator $\widehat{\theta}$ of a parameter θ has a remarkable property known as the invariance property. This invariance property says that if $\widehat{\theta}$ is a maximum likelihood estimator of θ, then $g(\widehat{\theta})$ is the maximum likelihood estimator of $g(\theta)$, where g is a function from $\mathbb{R}^k$ to a subset of $\mathbb{R}^m$. This result was proved by Zehna in 1966. We state this result as a theorem without a proof.

Theorem 15.1. Let $\widehat{\theta}$ be a maximum likelihood estimator of a parameter θ and let $g(\theta)$ be a function of θ. Then the maximum likelihood estimator of $g(\theta)$ is given by $g\left(\widehat{\theta}\right)$.

Now we give two examples to illustrate the importance of this theorem.

Example 15.15. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance σ^2. What are the maximum likelihood estimators of σ and $\mu - \sigma$?

Answer: From Example 15.13, we have the maximum likelihood estimator of μ and σ^2 to be

$$\widehat{\mu} = \overline{X}$$

and

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 =: \Sigma^2 \text{ (say)}.$$

Now using the invariance property of the maximum likelihood estimator we have

$$\widehat{\sigma} = \Sigma$$

and

$$\widehat{\mu - \sigma} = \overline{X} - \Sigma.$$

Example 15.16. Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha < x < \beta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the estimator of $\sqrt{\alpha^2 + \beta^2}$ by the method of maximum likelihood.

Answer: From Example 15.14, we have the maximum likelihood estimator of α and β to be

$$\widehat{\alpha} = X_{(1)} \qquad \text{and} \qquad \widehat{\beta} = X_{(n)},$$

respectively. Now using the invariance property of the maximum likelihood estimator we see that the maximum likelihood estimator of $\sqrt{\alpha^2 + \beta^2}$ is $\sqrt{X_{(1)}^2 + X_{(n)}^2}$.

First time, the concept of information in statistics was introduced by Sir Ronald Fisher and now a day it is known as Fisher information.

Definition 15.4. Let X be an observation from a population with probability density function $f(x;\theta)$. Suppose $f(x;\theta)$ is continuous, twice differentiable and it's support of does not depend on θ. Then the Fisher information, $I(\theta)$, in a single observation X about θ is given by

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{d \, \ln f(x;\theta)}{d\theta} \right]^2 f(x;\theta)\, dx.$$

Thus $I(\theta)$ is the expected value of the random variable $\frac{d \, \ln f(X;\theta)}{d\theta}$. Hence

$$I(\theta) = E\left(\left[\frac{d \, \ln f(X;\theta)}{d\theta} \right]^2 \right).$$

In the following lemma, we give an alternative formula for the Fisher information.

Lemma 15.1. The Fisher information contained in a single observation about the unknown parameter θ can be given alternatively as

$$I(\theta) = -\int_{-\infty}^{\infty} \left[\frac{d^2 \, \ln f(x;\theta)}{d\theta^2} \right] f(x;\theta)\, dx.$$

Proof: Since $f(x;\theta)$ is a probability density function,

$$\int_{-\infty}^{\infty} f(x;\theta)\, dx = 1. \tag{3}$$

Differentiating (3) with respect to θ, we get

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x;\theta)\, dx = 0.$$

Rewriting the last equality, we obtain

$$\int_{-\infty}^{\infty} \frac{df(x;\theta)}{d\theta} \frac{1}{f(x;\theta)} f(x;\theta)\, dx = 0$$

which is

$$\int_{-\infty}^{\infty} \frac{d \ln f(x;\theta)}{d\theta} f(x;\theta)\, dx = 0. \tag{4}$$

Now differentiating (4) with respect to θ, we see that

$$\int_{-\infty}^{\infty} \left[\frac{d^2 \,\ln f(x;\theta)}{d\theta^2} f(x;\theta) + \frac{d \ln f(x;\theta)}{d\theta} \frac{df(x;\theta)}{d\theta} \right] dx = 0.$$

Rewriting the last equality, we have

$$\int_{-\infty}^{\infty} \left[\frac{d^2 \,\ln f(x;\theta)}{d\theta^2} f(x;\theta) + \frac{d \ln f(x;\theta)}{d\theta} \frac{df(x;\theta)}{d\theta} \frac{1}{f(x;\theta)} f(x;\theta) \right] dx = 0$$

which is

$$\int_{-\infty}^{\infty} \left(\frac{d^2 \,\ln f(x;\theta)}{d\theta^2} + \left[\frac{d \ln f(x;\theta)}{d\theta} \right]^2 \right) f(x;\theta)\, dx = 0.$$

The last equality implies that

$$\int_{-\infty}^{\infty} \left[\frac{d \,\ln f(x;\theta)}{d\theta} \right]^2 f(x;\theta)\, dx = - \int_{-\infty}^{\infty} \left[\frac{d^2 \,\ln f(x;\theta)}{d\theta^2} \right] f(x;\theta)\, dx.$$

Hence using the definition of Fisher information, we have

$$I(\theta) = - \int_{-\infty}^{\infty} \left[\frac{d^2 \,\ln f(x;\theta)}{d\theta^2} \right] f(x;\theta)\, dx$$

and the proof of the lemma is now complete.

The following two examples illustrate how one can determine Fisher information.

Example 15.17. Let X be a single observation taken from a normal population with unknown mean μ and known variance σ^2. Find the Fisher information in a single observation X about μ.

Answer: Since $X \sim N(\mu, \sigma^2)$, the probability density of X is given by

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Hence

$$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}.$$

Therefore

$$\frac{d\ln f(x;\mu)}{d\mu} = \frac{x-\mu}{\sigma^2}$$

and

$$\frac{d^2\ln f(x;\mu)}{d\mu^2} = -\frac{1}{\sigma^2}.$$

Hence

$$I(\mu) = -\int_{-\infty}^{\infty}\left(-\frac{1}{\sigma^2}\right) f(x;\mu)\,dx = \frac{1}{\sigma^2}.$$

Example 15.18. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with unknown mean μ and known variance σ^2. Find the Fisher information in this sample of size n about μ.

Answer: Let $I_n(\mu)$ be the required Fisher information. Then from the definition, we have

$$\begin{aligned}
I_n(\mu) &= -E\left(\frac{d^2\ln f(X_1, X_2, ..., X_n;\mu)}{d\mu^2}\right) \\
&= -E\left(\frac{d}{d\mu^2}\left\{\ln f(X_1;\mu) + \cdots + \ln f(X_n;\mu)\right\}\right) \\
&= -E\left(\frac{d^2\ln f(X_1;\mu)}{d\mu^2}\right) - \cdots - E\left(\frac{d^2\ln f(X_n;\mu)}{d\mu^2}\right) \\
&= I(\theta) + \cdots + I(\theta) \\
&= n\,I(\theta) \\
&= n\,\frac{1}{\sigma^2} \qquad\qquad \text{(using Example 15.17).}
\end{aligned}$$

This example shows that if $X_1, X_2, ..., X_n$ is a random sample from a population $X \sim f(x;\theta)$, then the Fisher information, $I_n(\theta)$, in a sample of size n about the parameter θ is equal to n times the Fisher information in X about θ. Thus

$$I_n(\theta) = n\,I(\theta).$$

If X is a random variable with probability density function $f(x;\theta)$, where $\theta = (\theta_1, ..., \theta_n)$ is an unknown parameter vector then the Fisher information,

$I(\theta)$, is a $n \times n$ matrix given by

$$\begin{aligned} I(\theta) &= (I_{ij}(\theta)) \\ &= \left(-E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial \theta_i \, \partial \theta_j} \right) \right). \end{aligned}$$

Example 15.19. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance σ^2. What is the Fisher information matrix, $I_n(\mu, \sigma^2)$, of the sample of size n about the parameters μ and σ^2?

Answer: Let us write $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. The Fisher information, $I_n(\theta)$, in a sample of size n about the parameter (θ_1, θ_2) is equal to n times the Fisher information in the population about (θ_1, θ_2), that is

$$I_n(\theta_1, \theta_2) = n\, I(\theta_1, \theta_2). \tag{5}$$

Since there are two parameters θ_1 and θ_2, the Fisher information matrix $I(\theta_1, \theta_2)$ is a 2×2 matrix given by

$$I(\theta_1, \theta_2) = \begin{pmatrix} I_{11}(\theta_1, \theta_2) & I_{12}(\theta_1, \theta_2) \\ \\ I_{21}(\theta_1, \theta_2) & I_{22}(\theta_1, \theta_2) \end{pmatrix} \tag{6}$$

where

$$I_{ij}(\theta_1, \theta_2) = -E\left(\frac{\partial^2 \ln f(X;\theta_1, \theta_2)}{\partial \theta_i \, \partial \theta_j} \right)$$

for $i = 1, 2$ and $j = 1, 2$. Now we proceed to compute I_{ij}. Since

$$f(x;\theta_1, \theta_2) = \frac{1}{\sqrt{2\,\pi\,\theta_2}}\, e^{-\frac{(x-\theta_1)^2}{2\,\theta_2}}$$

we have

$$\ln f(x;\theta_1, \theta_2) = -\frac{1}{2}\, \ln(2\,\pi\,\theta_2) - \frac{(x-\theta_1)^2}{2\,\theta_2}.$$

Taking partials of $\ln f(x;\theta_1, \theta_2)$, we have

$$\begin{aligned} \frac{\partial \ln f(x;\theta_1, \theta_2)}{\partial \theta_1} &= \frac{x-\theta_1}{\theta_2}, \\ \frac{\partial \ln f(x;\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{1}{2\,\theta_2} + \frac{(x-\theta_1)^2}{2\,\theta_2^2}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial \theta_1^2} &= -\frac{1}{\theta_2}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial \theta_2^2} &= \frac{1}{2\,\theta_2^2} + \frac{(x-\theta_1)^2}{\theta_2^3}, \\ \frac{\partial^2 \ln f(x;\theta_1, \theta_2)}{\partial \theta_1 \, \partial \theta_2} &= -\frac{x-\theta_1}{\theta_2^2}. \end{aligned}$$

Hence

$$I_{11}(\theta_1, \theta_2) = -E\left(-\frac{1}{\theta_2}\right) = \frac{1}{\theta_2} = \frac{1}{\sigma^2}.$$

Similarly,

$$I_{12}(\theta_1, \theta_2) = -E\left(\frac{X - \theta_1}{\theta_2^2}\right) = \frac{E(X)}{\theta_2^2} - \frac{\theta_1}{\theta_2^2} = \frac{\theta_1}{\theta_2^2} - \frac{\theta_1}{\theta_2^2} = 0$$

and

$$\begin{aligned} I_{12}(\theta_1, \theta_2) &= -E\left(-\frac{(X - \theta_1)^2}{\theta_2^3} + \frac{1}{2\theta_2^2}\right) \\ &= \frac{E\left((X - \theta_1)^2\right)}{\theta_2^3} - \frac{1}{2\theta_2^2} = \frac{\theta_2}{\theta_2^3} - \frac{1}{2\theta_2^2} = \frac{1}{2\theta_2^2} = \frac{1}{2\sigma^4}. \end{aligned}$$

Thus from (5), (6) and the above calculations, the Fisher information matrix is given by

$$I_n(\theta_1, \theta_2) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Now we present an important theorem about the maximum likelihood estimator without a proof.

Theorem 15.2. Under certain regularity conditions on the $f(x; \theta)$ the maximum likelihood estimator $\widehat{\theta}$ of θ based on a random sample of size n from a population X with probability density $f(x; \theta)$ is asymptotically normally distributed with mean θ and variance $\frac{1}{n\, I(\theta)}$. That is

$$\widehat{\theta}_{ML} \sim N\left(\theta,\, \frac{1}{n\, I(\theta)}\right) \qquad \text{as } n \to \infty.$$

The following example shows that the maximum likelihood estimator of a parameter is not necessarily unique.

Example 15.20. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x; \theta) = \begin{cases} \frac{1}{2} & \text{if } \theta - 1 \leq x \leq \theta + 1 \\ 0 & \text{otherwise,} \end{cases}$$

then what is the maximum likelihood estimator of θ?

Answer: The likelihood function of this sample is given by

$$L(\theta) = \begin{cases} \left(\frac{1}{2}\right)^n & \text{if } \max\{x_1, ..., x_n\} - 1 \leq \theta \leq \min\{x_1, ..., x_n\} + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Since the likelihood function is a constant, any value in the interval $[\max\{x_1, ..., x_n\} - 1,\ \min\{x_1, ..., x_n\} + 1]$ is a maximum likelihood estimate of θ.

Example 15.21. What is the basic principle of maximum likelihood estimation?

Answer: To choose a value of the parameter for which the observed data have as high a probability or density as possible. In other words a maximum likelihood estimate is a parameter value under which the sample data have the highest probability.

15.3. Bayesian Method

In the classical approach, the parameter θ is assumed to be an unknown, but fixed quantity. A random sample $X_1, X_2, ..., X_n$ is drawn from a population with probability density function $f(x;\theta)$ and based on the observed values in the sample, knowledge about the value of θ is obtained.

In Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (known as the prior distribution). This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen (and hence the name prior distribution). A sample is then taken from a population where θ is a parameter and the prior distribution is updated with this sample information. This updated prior is called the posterior distribution. The updating is done with the help of Bayes' theorem and hence the name Bayesian method.

In this section, we shall denote the population density $f(x;\theta)$ as $f(x/\theta)$, that is the density of the population X given the parameter θ.

Definition 15.5. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. The probability density function of the random variable θ is called the prior distribution of θ and usually denoted by $h(\theta)$.

Definition 15.6. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. The conditional density, $k(\theta/x_1, x_2, ..., x_n)$, of θ given the sample $x_1, x_2, ..., x_n$ is called the posterior distribution of θ.

Example 15.22. Let $X_1 = 1$, $X_2 = 2$ be a random sample of size 2 from a distribution with probability density function

$$f(x/\theta) = \binom{3}{x} \theta^x (1-\theta)^{3-x}, \qquad x = 0, 1, 2, 3.$$

If the prior density of θ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the posterior distribution of θ ?

Answer: Since $h(\theta)$ is the probability density of θ, we should get

$$\int_{\frac{1}{2}}^{1} h(\theta)\, d\theta = 1$$

which implies

$$\int_{\frac{1}{2}}^{1} k\, d\theta = 1.$$

Therefore $k = 2$. The joint density of the sample and the parameter is given by

$$\begin{aligned} u(x_1, x_2, \theta) &= f(x_1/\theta) f(x_2/\theta) h(\theta) \\ &= \binom{3}{x_1} \theta^{x_1} (1-\theta)^{3-x_1} \binom{3}{x_2} \theta^{x_2} (1-\theta)^{3-x_2}\, 2 \\ &= 2 \binom{3}{x_1} \binom{3}{x_2} \theta^{x_1+x_2} (1-\theta)^{6-x_1-x_2}. \end{aligned}$$

Hence,

$$\begin{aligned} u(1, 2, \theta) &= 2 \binom{3}{1} \binom{3}{2} \theta^3 (1-\theta)^3 \\ &= 18\, \theta^3 (1-\theta)^3. \end{aligned}$$

The marginal distribution of the sample

$$
\begin{aligned}
g(1,2) &= \int_{\frac{1}{2}}^{1} u(1,2,\theta)\, d\theta \\
&= \int_{\frac{1}{2}}^{1} 18\,\theta^3 (1-\theta)^3\, d\theta \\
&= 18 \int_{\frac{1}{2}}^{1} \theta^3 \left(1 + 3\theta^2 - 3\theta - \theta^3\right)\, d\theta \\
&= 18 \int_{\frac{1}{2}}^{1} \left(\theta^3 + 3\theta^5 - 3\theta^4 - \theta^6\right)\, d\theta \\
&= \frac{9}{140}.
\end{aligned}
$$

The conditional distribution of the parameter θ given the sample $X_1 = 1$ and $X_2 = 2$ is given by

$$
\begin{aligned}
k(\theta/x_1 = 1, x_2 = 2) &= \frac{u(1,2,\theta)}{g(1,2)} \\
&= \frac{18\,\theta^3\,(1-\theta)^3}{\frac{9}{140}} \\
&= 280\,\theta^3\,(1-\theta)^3.
\end{aligned}
$$

Therefore, the posterior distribution of θ is

$$
k(\theta/x_1 = 1, x_2 = 2) = \begin{cases} 280\,\theta^3\,(1-\theta)^3 & \text{if } \frac{1}{2} < \theta < 1 \\ 0 & \text{otherwise.} \end{cases}
$$

Remark 15.2. If $X_1, X_2, ..., X_n$ is a random sample from a population with density $f(x/\theta)$, then the joint density of the sample and the parameter is given by

$$
u(x_1, x_2, ..., x_n, \theta) = h(\theta) \prod_{i=1}^{n} f(x_i/\theta).
$$

Given this joint density, the marginal density of the sample can be computed using the formula

$$
g(x_1, x_2, ..., x_n) = \int_{-\infty}^{\infty} h(\theta) \prod_{i=1}^{n} f(x_i/\theta)\, d\theta.
$$

Now using the Bayes rule, the posterior distribution of θ can be computed as follows:

$$k(\theta/x_1, x_2, ..., x_n) = \frac{h(\theta) \prod_{i=1}^{n} f(x_i/\theta)}{\int_{-\infty}^{\infty} h(\theta) \prod_{i=1}^{n} f(x_i/\theta)\, d\theta}.$$

In Bayesian method, we use two types of loss functions.

Definition 15.7. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. Let $\widehat{\theta}$ be an estimator of θ. The function

$$\mathcal{L}_2\left(\widehat{\theta}, \theta\right) = \left(\widehat{\theta} - \theta\right)^2$$

is called the squared error loss. The function

$$\mathcal{L}_1\left(\widehat{\theta}, \theta\right) = \left|\widehat{\theta} - \theta\right|$$

is called the absolute error loss.

The loss function $\mathcal{L}$ represents the 'loss' incurred when $\widehat{\theta}$ is used in place of the parameter θ.

Definition 15.8. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. Let $\widehat{\theta}$ be an estimator of θ and let $\mathcal{L}\left(\widehat{\theta}, \theta\right)$ be a given loss function. The expected value of this loss function with respect to the population distribution $f(x/\theta)$, that is

$$R_{\mathcal{L}}(\theta) = \int \mathcal{L}\left(\widehat{\theta}, \theta\right)\, f(x/\theta)\, dx$$

is called the risk.

The posterior density of the parameter θ given the sample $x_1, x_2, ..., x_n$, that is

$$k(\theta/x_1, x_2, ..., x_n)$$

contains all information about θ. In Bayesian estimation of parameter one chooses an estimate $\widehat{\theta}$ for θ such that

$$k(\widehat{\theta}/x_1, x_2, ..., x_n)$$

is maximum subject to a loss function. Mathematically, this is equivalent to minimizing the integral

$$\int_{\Omega} \mathcal{L}\left(\widehat{\theta}, \theta\right)\, k(\theta/x_1, x_2, ..., x_n)\, d\theta$$

with respect to $\widehat{\theta}$, where Ω denotes the support of the prior density $h(\theta)$ of the parameter θ.

Example 15.23. Suppose one observation was taken of a random variable X which yielded the value 2. The density function for X is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

and prior distribution for parameter θ is

$$h(\theta) = \begin{cases} \frac{3}{\theta^4} & \text{if } 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

If the loss function is $\mathcal{L}(z, \theta) = (z - \theta)^2$, then what is the Bayes' estimate for θ ?

Answer: The prior density of the random variable θ is

$$h(\theta) = \begin{cases} \frac{3}{\theta^4} & \text{if } 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

The probability density function of the population is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the joint probability density function of the sample and the parameter is given by

$$\begin{aligned} u(x, \theta) &= h(\theta)\, f(x/\theta) \\ &= \frac{3}{\theta^4}\, \frac{1}{\theta} \\ &= \begin{cases} 3\,\theta^{-5} & \text{if } 0 < x < \theta \quad \text{and} \quad 1 < \theta < \infty \\ \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The marginal density of the sample is given by

$$\begin{aligned} g(x) &= \int_x^\infty u(x, \theta)\, d\theta \\ &= \int_x^\infty 3\,\theta^{-5}\, d\theta \\ &= \frac{3}{4}\, x^{-4} \\ &= \frac{3}{4\, x^4}. \end{aligned}$$

Thus, if $x = 2$, then $g(2) = \frac{3}{64}$. The posterior density of θ when $x = 2$ is given by

$$\begin{aligned} k(\theta/x=2) &= \frac{u(2,\theta)}{g(2)} \\ &= \frac{64}{3}\, 3\,\theta^{-5} \\ &= \begin{cases} 64\,\theta^{-5} & \text{if } 2 < \theta < \infty \\ \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

Now, we find the Bayes estimator by minimizing the expression $E\left[\mathcal{L}(\theta, z)/x = 2\right]$. That is

$$\widehat{\theta} = Arg \max_{z\in\Omega} \int_\Omega \mathcal{L}(\theta, z)\, k(\theta/x = 2)\, d\theta.$$

Let us call this integral $\psi(z)$. Then

$$\begin{aligned} \psi(z) &= \int_\Omega \mathcal{L}(\theta, z)\, k(\theta/x = 2)\, d\theta \\ &= \int_2^\infty (z-\theta)^2\, k(\theta/x=2)\, d\theta \\ &= \int_2^\infty (z-\theta)^2\, 64\theta^{-5}\, d\theta. \end{aligned}$$

We want to find the value of z which yields a minimum of $\psi(z)$. This can be done by taking the derivative of $\psi(z)$ and evaluating where the derivative is zero.

$$\begin{aligned} \frac{d}{dz}\psi(z) &= \frac{d}{dz} \int_2^\infty (z-\theta)^2\, 64\theta^{-5}\, d\theta \\ &= 2\int_2^\infty (z-\theta)\, 64\theta^{-5}\, d\theta \\ &= 2\int_2^\infty z\, 64\theta^{-5}\, d\theta - 2\int_2^\infty \theta\, 64\theta^{-5}\, d\theta \\ &= 2\,z - \frac{16}{3}. \end{aligned}$$

Setting this derivative of $\psi(z)$ to zero and solving for z, we get

$$\begin{aligned} 2z - \frac{16}{3} &= 0 \\ \Rightarrow z &= \frac{8}{3}. \end{aligned}$$

Since $\frac{d^2\psi(z)}{dz^2} = 2$, the function $\psi(z)$ has a minimum at $z = \frac{8}{3}$. Hence, the Bayes' estimate of θ is $\frac{8}{3}$.

In Example 15.23, we have found the Bayes' estimate of θ by directly minimizing the $\int_\Omega \mathcal{L}\left(\widehat{\theta},\theta\right)\, k(\theta/x_1, x_2, ..., x_n)\, d\theta$ with respect to $\widehat{\theta}$. The next result is very useful while finding the Bayes' estimate using a quadratic loss function. Notice that if $\mathcal{L}(\widehat{\theta},\theta) = (\theta - \widehat{\theta})^2$, then $\int_\Omega \mathcal{L}\left(\widehat{\theta},\theta\right)\, k(\theta/x_1, x_2, ..., x_n)\, d\theta$ is $E\left((\theta - \widehat{\theta})^2\,/x_1, x_2, ..., x_n\right)$. The following theorem is based on the fact that the function ϕ defined by $\phi(c) = E\left[(X-c)^2\right]$ attains minimum if $c = E[X]$.

Theorem 15.3. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. If the loss function is squared error, then the Bayes' estimator $\widehat{\theta}$ of parameter θ is given by

$$\widehat{\theta} = E(\theta/x_1, x_2, ..., x_n),$$

where the expectation is taken with respect to density $k(\theta/x_1, x_2, ..., x_n)$.

Now we give several examples to illustrate the use of this theorem.

Example 15.24. Suppose the prior distribution of θ is uniform over the interval $(0, 1)$. Given θ, the population X is uniform over the interval $(0, \theta)$. If the squared error loss function is used, find the Bayes' estimator of θ based on a sample of size one.

Answer: The prior density of θ is given by

$$h(\theta) = \begin{cases} 1 & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise .} \end{cases}$$

The density of population is given by

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

The joint density of the sample and the parameter is given by

$$\begin{aligned} u(x,\theta) &= h(\theta)\, f(x/\theta) \\ &= 1 \left(\frac{1}{\theta}\right) \\ &= \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta < 1 \\ 0 & \text{otherwise .} \end{cases} \end{aligned}$$

The marginal density of the sample is

$$\begin{aligned} g(x) &= \int_x^1 u(x,\theta)\, d\theta \\ &= \int_x^1 \frac{1}{\theta}\, d\theta \\ &= \begin{cases} -\ln x & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The conditional density of θ given the sample is

$$k(\theta/x) = \frac{u(x,\theta)}{g(x)} = \begin{cases} -\frac{1}{\theta \ln x} & \text{if } 0 < x < \theta < 1 \\ \\ 0 & \text{elsewhere .} \end{cases}$$

Since the loss function is quadratic error, therefore the Bayes' estimator of θ is

$$\begin{aligned} \widehat{\theta} &= E[\theta/x] \\ &= \int_x^1 \theta\, k(\theta/x)\, d\theta \\ &= \int_x^1 \theta\, \frac{-1}{\theta\, \ln x}\, d\theta \\ &= -\frac{1}{\ln x} \int_x^1 d\theta \\ &= \frac{x-1}{\ln x}. \end{aligned}$$

Thus, the Bayes' estimator of θ based on one observation X is

$$\widehat{\theta} = \frac{X-1}{\ln X}.$$

Example 15.25. Given θ, the random variable X has a binomial distribution with $n = 2$ and probability of success θ. If the prior density of θ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of θ for a squared error loss if $X = 1$?

Answer: Note that θ is uniform on the interval $\left(\frac{1}{2}, 1\right)$, hence $k = 2$. Therefore, the prior density of θ is

$$h(\theta) = \begin{cases} 2 & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The population density is given by

$$f(x/\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \binom{2}{x} \theta^x (1-\theta)^{2-x}, \qquad x = 0, 1, 2.$$

The joint density of the sample and the parameter θ is

$$\begin{aligned} u(x, \theta) &= h(\theta)\, f(x/\theta) \\ &= 2 \binom{2}{x} \theta^x (1-\theta)^{2-x} \end{aligned}$$

where $\frac{1}{2} < \theta < 1$ and $x = 0, 1, 2$. The marginal density of the sample is given by

$$g(x) = \int_{\frac{1}{2}}^{1} u(x, \theta)\, d\theta.$$

This integral is easy to evaluate if we substitute $X = 1$ now. Hence

$$\begin{aligned} g(1) &= \int_{\frac{1}{2}}^{1} 2 \binom{2}{1} \theta\, (1-\theta)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} \left(4\theta - 4\theta^2\right)\, d\theta \\ &= 4 \left[\frac{\theta^2}{2} - \frac{\theta^3}{3} \right]_{\frac{1}{2}}^{1} \\ &= \frac{2}{3} \left[3\theta^2 - 2\theta^3 \right]_{\frac{1}{2}}^{1} \\ &= \frac{2}{3} \left[(3-2) - \left(\frac{3}{4} - \frac{2}{8} \right) \right] \\ &= \frac{1}{3}. \end{aligned}$$

Therefore, the posterior density of θ given $x = 1$, is

$$k(\theta/x = 1) = \frac{u(1, \theta)}{g(1)} = 12\, (\theta - \theta^2),$$

where $\frac{1}{2} < \theta < 1$. Since the loss function is quadratic error, therefore the

Bayes' estimate of θ is

$$\begin{aligned}\widehat{\theta} &= E[\theta/x=1] \\ &= \int_{\frac{1}{2}}^{1} \theta\, k(\theta/x=1)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} 12\,\theta\,(\theta-\theta^2)\, d\theta \\ &= \left[4\,\theta^3 - 3\,\theta^4\right]_{\frac{1}{2}}^{1} \\ &= 1 - \frac{5}{16} \\ &= \frac{11}{16}.\end{aligned}$$

Hence, based on the sample of size one with $X = 1$, the Bayes' estimate of θ is $\frac{11}{16}$, that is

$$\widehat{\theta} = \frac{11}{16}.$$

The following theorem help us to evaluate the Bayes estimate of a sample if the loss function is absolute error loss. This theorem is based the fact that a function $\phi(c) = E\left[\,|X-c|\,\right]$ is minimum if c is the median of X.

Theorem 15.4. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density $f(x/\theta)$, where θ is the unknown parameter to be estimated. If the loss function is absolute error, then the Bayes estimator $\widehat{\theta}$ of the parameter θ is given by

$$\widehat{\theta} = \text{median of } k(\theta/x_1, x_2, ..., x_n)$$

where $k(\theta/x_1, x_2, ..., x_n)$ is the posterior distribution of θ.

The followings are some examples to illustrate the above theorem.

Example 15.26. Given θ, the random variable X has a binomial distribution with $n = 3$ and probability of success θ. If the prior density of θ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of θ for an *absolute difference error loss* if the sample consists of one observation $x = 3$?

Answer: Since, the prior density of θ is

$$h(\theta) = \begin{cases} 2 & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise ,} \end{cases}$$

and the population density is

$$f(x/\theta) = \binom{3}{x} \theta^x (1-\theta)^{3-x},$$

the joint density of the sample and the parameter is given by

$$u(3, \theta) = h(\theta)\, f(3/\theta) = 2\,\theta^3,$$

where $\frac{1}{2} < \theta < 1$. The marginal density of the sample (at $x = 3$) is given by

$$\begin{aligned} g(3) &= \int_{\frac{1}{2}}^{1} u(3, \theta)\, d\theta \\ &= \int_{\frac{1}{2}}^{1} 2\,\theta^3\, d\theta \\ &= \left[\frac{\theta^4}{2}\right]_{\frac{1}{2}}^{1} \\ &= \frac{15}{32}. \end{aligned}$$

Therefore, the conditional density of θ given $X = 3$ is

$$k(\theta/x = 3) = \frac{u(3, \theta)}{g(3)} = \begin{cases} \frac{64}{15}\,\theta^3 & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{elsewhere.} \end{cases}$$

Since, the loss function is absolute error, the Bayes' estimator is the median of the probability density function $k(\theta/x = 3)$. That is

$$\begin{aligned} \frac{1}{2} &= \int_{\frac{1}{2}}^{\widehat{\theta}} \frac{64}{15}\,\theta^3\, d\theta \\ &= \frac{64}{60}\, \left[\theta^4\right]_{\frac{1}{2}}^{\widehat{\theta}} \\ &= \frac{64}{60}\, \left[\left(\widehat{\theta}\right)^4 - \frac{1}{16}\right]. \end{aligned}$$

Solving the above equation for $\widehat{\theta}$, we get

$$\widehat{\theta} = \sqrt[4]{\frac{17}{32}} = 0.8537.$$

Example 15.27. Suppose the prior distribution of θ is uniform over the interval $(2, 5)$. Given θ, X is uniform over the interval $(0, \theta)$. What is the Bayes' estimator of θ for *absolute error loss* if $X = 1$?

Answer: Since, the prior density of θ is

$$h(\theta) = \begin{cases} \frac{1}{3} & \text{if } 2 < \theta < 5 \\ \\ 0 & \text{otherwise ,} \end{cases}$$

and the population density is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

the joint density of the sample and the parameter is given by

$$u(x, \theta) = h(\theta)\, f(x/\theta) = \frac{1}{3\theta},$$

where $2 < \theta < 5$ and $0 < x < \theta$. The marginal density of the sample (at $x = 1$) is given by

$$\begin{aligned} g(1) &= \int_1^5 u(1, \theta)\, d\theta \\ &= \int_1^2 u(1, \theta)\, d\theta + \int_2^5 u(1, \theta)\, d\theta \\ &= \int_2^5 \frac{1}{3\theta}\, d\theta \\ &= \frac{1}{3} \ln\left(\frac{5}{2}\right). \end{aligned}$$

Therefore, the conditional density of θ given the sample $x = 1$, is

$$\begin{aligned} k(\theta/x = 1) &= \frac{u(1, \theta)}{g(1)} \\ &= \frac{1}{\theta \ln\left(\frac{5}{2}\right)}. \end{aligned}$$

Since, the loss function is absolute error, the Bayes estimate of θ is the median of $k(\theta/x=1)$. Hence

$$\frac{1}{2} = \int_2^{\widehat{\theta}} \frac{1}{\theta \, \ln\left(\frac{5}{2}\right)} \, d\theta$$

$$= \frac{1}{\ln\left(\frac{5}{2}\right)} \, \ln\left(\frac{\widehat{\theta}}{2}\right).$$

Solving for $\widehat{\theta}$, we get

$$\widehat{\theta} = \sqrt{10} = 3.16.$$

Example 15.28. What is the basic principle of Bayesian estimation?

Answer: The basic principle behind the Bayesian estimation method consists of choosing a value of the parameter θ for which the observed data have as high a posterior probability $k(\theta/x_1, x_2, ..., x_n)$ of θ as possible subject to a loss function.

15.4. Review Exercises

1. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{2\theta} & \text{if} \quad -\theta < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter θ.

2. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter θ.

3. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2 \, x \, e^{-\theta \, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the moment method find an estimator for the parameter θ.

4. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if} \quad 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter θ.

5. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter θ.

6. Let $X_1, X_2, ..., X_n$ be a random sample of size$n$ from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2\, x\, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter θ.

7. Let X_1, X_2, X_3, X_4 be a random sample from a distribution with density function

$$f(x) = \begin{cases} \frac{1}{\beta}\, e^{\frac{-(x-4)}{\beta}} & \text{for} \quad x > 4 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$. If the data from this random sample are 8.2, 9.1, 10.6 and 4.9, respectively, what is the maximum likelihood estimate of β?

8. Given θ, the random variable X has a binomial distribution with $n = 2$ and probability of success θ. If the prior density of θ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of θ for a squared error loss if the sample consists of $x_1 = 1$ and $x_2 = 2$.

9. Suppose two observations were taken of a random variable X which yielded the values 2 and 3. The density function for X is

$$f(x/\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

and prior distribution for the parameter θ is

$$h(\theta) = \begin{cases} 3\,\theta^{-4} & \text{if } \theta > 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

If the loss function is quadratic, then what is the Bayes' estimate for θ?

10. The Pareto distribution is often used in study of incomes and has the *cumulative density function*

$$F(x;\alpha,\theta) = \begin{cases} 1 - \left(\frac{\alpha}{x}\right)^{\theta} & \text{if } \alpha \leq x \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \alpha < \infty$ and $1 < \theta < \infty$ are parameters. Find the maximum likelihood estimates of α and θ based on a sample of size 5 for value 3, 5, 2, 7, 8.

11. The Pareto distribution is often used in study of incomes and has the *cumulative density function*

$$F(x;\alpha,\theta) = \begin{cases} 1 - \left(\frac{\alpha}{x}\right)^{\theta} & \text{if } \alpha \leq x \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \alpha < \infty$ and $1 < \theta < \infty$ are parameters. Using moment methods find estimates of α and θ based on a sample of size 5 for value 3, 5, 2, 7, 8.

12. Suppose one observation was taken of a random variable X which yielded the value 2. The density function for X is

$$f(x/\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

and prior distribution of μ is

$$h(\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\mu^2} \qquad -\infty < \mu < \infty.$$

If the loss function is quadratic, then what is the Bayes' estimate for μ?

13. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with probability density

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{if } 2\theta \leq x \leq 3\theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. What is the maximum likelihood estimator of θ?

14. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with probability density

$$f(x) = \begin{cases} 1 - \theta^2 & \text{if } 0 \leq x \leq \frac{1}{1-\theta^2} \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. What is the maximum likelihood estimator of θ?

15. Given θ, the random variable X has a binomial distribution with $n = 3$ and probability of success θ. If the prior density of θ is

$$h(\theta) = \begin{cases} k & \text{if } \frac{1}{2} < \theta < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

what is the Bayes' estimate of θ for a *absolute difference error loss* if the sample consists of one observation $x = 1$?

16. Suppose the random variable X has the cumulative density function $F(x)$. Show that the expected value of the random variable $(X - c)^2$ is minimum if c equals the expected value of X.

17. Suppose the continuous random variable X has the cumulative density function $F(x)$. Show that the expected value of the random variable $|X - c|$ is minimum if c equals the median of X (that is, $F(c) = 0.5$).

18. Eight independent trials are conducted of a given system with the following results: S, F, S, F, S, S, S, S where S denotes the success and F denotes the failure. What is the maximum likelihood estimate of the probability of successful operation p ?

19. What is the maximum likelihood estimate of β if the 5 values $\frac{4}{5}$, $\frac{2}{3}$, 1, $\frac{3}{2}$, $\frac{5}{4}$ were drawn from the population for which $f(x;\beta) = \frac{1}{2}(1+\beta)^5 \left(\frac{x}{2}\right)^{\beta}$?

20. If a sample of five values of X is taken from the population for which $f(x;t) = 2(t-1)t^x$, what is the maximum likelihood estimator of t ?

21. A sample of size n is drawn from a gamma distribution

$$f(x;\beta) = \begin{cases} \frac{x^3\, e^{-\frac{x}{\beta}}}{6\beta^4} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimator of β ?

22. The probability density function of the random variable X is defined by

$$f(x;\lambda) = \begin{cases} 1 - \frac{2}{3}\lambda + \lambda\sqrt{x} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimate of the parameter λ based on two independent observations $x_1 = \frac{1}{4}$ and $x_2 = \frac{9}{16}$?

23. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function $f(x;\sigma) = \frac{\sigma}{2}\, e^{-\sigma|x-\mu|}$. Let $Y_1, Y_2, ..., Y_n$ be the order statistics of this sample and assume n is odd and μ is known. What is the maximum likelihood estimator of σ ?

24. Suppose $X_1, X_2, ...$ are independent random variables, each with probability of success p and probability of failure $1-p$, where $0 \leq p \leq 1$. Let N be the number of observation needed to obtain the first success. What is the maximum likelihood estimator of p in term of N ?

25. Let X_1, X_2, X_3 and X_4 be a random sample from the discrete distribution X such that

$$P(X = x) = \begin{cases} \frac{\theta^{2x}\; e^{-\theta^2}}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$. If the data are 17, 10, 32, 5, what is the maximum likelihood estimate of θ ?

26. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population with a probability density function

$$f(x;\alpha,\lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\lambda\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where α and λ are parameters. Using the moment method find the estimators for the parameters α and λ.

27. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;p) = \binom{10}{x} p^x \, (1-p)^{10-x}$$

for $x = 0, 1, ..., 10$, where p is a parameter. Find the Fisher information in the sample about the parameter p.

28. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;\theta) = \begin{cases} \theta^2 \, x \, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. Find the Fisher information in the sample about the parameter θ.

29. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;\mu,\sigma^2) = \begin{cases} \frac{1}{x\,\sigma\,\sqrt{2\,\pi}} \, e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2}, & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$ are unknown parameters. Find the Fisher information matrix in the sample about the parameters μ and σ^2.

30. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;\mu,\lambda) = \begin{cases} \sqrt{\frac{\lambda}{2\pi}} \, x^{-\frac{3}{2}} \; e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $0 < \mu < \infty$ and $0 < \lambda < \infty$ are unknown parameters. Find the Fisher information matrix in the sample about the parameters μ and λ.

31. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\,\theta^\alpha} \, x^{\alpha-1} \, e^{-\frac{x}{\theta}} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\theta > 0$ are parameters. Using the moment method find estimators for parameters α and β.

32. Let $X_1, X_2, ..., X_n$ be a random sample of size$n$ from a distribution with a probability density function

$$f(x;\theta) = \frac{1}{\pi\,[1+(x-\theta)^2]}, \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter θ.

33. Let $X_1, X_2, ..., X_n$ be a random sample of size$n$ from a distribution with a probability density function

$$f(x;\theta) = \frac{1}{2}\,e^{-|x-\theta|}, \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Using the maximum likelihood method find an estimator for the parameter θ.

34. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x\,e^{-\lambda}}{x!} & \text{if } x = 0, 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ is an unknown parameter. Find the Fisher information matrix in the sample about the parameter λ.

35. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population distribution with the probability density function

$$f(x;p) = \begin{cases} (1-p)^{x-1}p & \text{if } x = 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p < 1$ is an unknown parameter. Find the Fisher information matrix in the sample about the parameter p.

Chapter 16

CRITERIA FOR EVALUATING THE GOODNESS OF ESTIMATORS

We have seen in Chapter 15 that, in general, different parameter estimation methods yield different estimators. For example, if $X \sim UNIF(0, \theta)$ and $X_1, X_2, ..., X_n$ is a random sample from the population X, then the estimator of θ obtained by moment method is

$$\widehat{\theta}_{MM} = 2\overline{X}$$

where as the estimator obtained by the maximum likelihood method is

$$\widehat{\theta}_{ML} = X_{(n)}$$

where $\overline{X}$ and $X_{(n)}$ are the sample average and the n^{th} order statistic, respectively. Now the question arises: which of the two estimators is better? Thus, we need some criteria to evaluate the goodness of an estimator. Some well known criteria for evaluating the goodness of an estimator are: (1) Unbiasedness, (2) Efficiency and Relative Efficiency, (3) Uniform Minimum Variance Unbiasedness, (4) Sufficiency, and (5) Consistency.

In this chapter, we shall examine only the first four criteria in details. The concepts of unbiasedness, efficiency and sufficiency were introduced by Sir Ronald Fisher.

16.1. The Unbiased Estimator

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population with probability density function $f(x;\theta)$. An estimator $\widehat{\theta}$ of θ is a function of the random variables $X_1, X_2, ..., X_n$ which is free of the parameter θ. An estimate is a realized value of an estimator that is obtained when a sample is actually taken.

Definition 16.1. An estimator $\widehat{\theta}$ of θ is said to be an unbiased estimator of θ if and only if

$$E\left(\widehat{\theta}\right) = \theta.$$

If $\widehat{\theta}$ is not unbiased, then it is called a biased estimator of θ.

An estimator of a parameter may not equal to the actual value of the parameter for every realization of the sample $X_1, X_2, ..., X_n$, but if it is unbiased then on an average it will equal to the parameter.

Example 16.1. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance $\sigma^2 > 0$. Is the sample mean $\overline{X}$ an unbiased estimator of the parameter μ ?

Answer: Since, each $X_i \sim N(\mu, \sigma^2)$, we have

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

That is, the sample mean is normal with mean μ and variance $\frac{\sigma^2}{n}$. Thus

$$E\left(\overline{X}\right) = \mu.$$

Therefore, the sample mean $\overline{X}$ is an unbiased estimator of μ.

Example 16.2. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance $\sigma^2 > 0$. What is the maximum likelihood estimator of σ^2 ? Is this maximum likelihood estimator an unbiased estimator of the parameter σ^2 ?

Answer: In Example 15.13, we have shown that the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Now, we examine the unbiasedness of this estimator

$$\begin{aligned}
E\left[\widehat{\sigma^2}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= E\left[\frac{n-1}{n}\,\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{n-1}{n}\,E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{n-1}{n}\,E\left[S^2\right] \\
&= \frac{\sigma^2}{n}\,E\left[\frac{n-1}{\sigma^2}\,S^2\right] \qquad \left(\text{since } \frac{n-1}{\sigma^2}\,S^2 \sim \chi^2(n-1)\right) \\
&= \frac{\sigma^2}{n}E\left[\chi^2(n-1)\right] \\
&= \frac{\sigma^2}{n}\,(n-1) \\
&= \frac{n-1}{n}\,\sigma^2 \\
&\neq \sigma^2.
\end{aligned}$$

Therefore, the maximum likelihood estimator of σ^2 is a biased estimator.

Next, in the following example, we show that the sample variance S^2 given by the expression

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

is an unbiased estimator of the population variance σ^2 irrespective of the population distribution.

Example 16.3. Let $X_1, X_2, ..., X_n$ be a random sample from a population with mean μ and variance $\sigma^2 > 0$. Is the sample variance S^2 an unbiased estimator of the population variance σ^2 ?

Answer: Note that the distribution of the population is not given. However, we are given $E(X_i) = \mu$ and $E[(X_i-\mu)^2] = \sigma^2$. In order to find $E\left(S^2\right)$, we need $E\left(\overline{X}\right)$ and $E\left(\overline{X}^2\right)$. Thus we proceed to find these two expected values. Consider

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \end{aligned}$$

Similarly,

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

Therefore

$$E\left(\overline{X}^2\right) = Var\left(\overline{X}\right) + E\left(\overline{X}\right)^2 = \frac{\sigma^2}{n} + \mu^2.$$

Consider

$$\begin{aligned} E\left(S^2\right) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\ &= \frac{1}{n-1}\,E\left[\sum_{i=1}^{n}\left(X_i^2 - 2\overline{X}X_i + \overline{X}^2\right)\right] \\ &= \frac{1}{n-1}\,E\left[\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2\right] \\ &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E\left[X_i^2\right] - E\left[n\,\overline{X}^2\right]\right\} \\ &= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] \\ &= \frac{1}{n-1}\left[(n-1)\,\sigma^2\right] \\ &= \sigma^2. \end{aligned}$$

Therefore, the sample variance S^2 is an unbiased estimator of the population variance σ^2.

Example 16.4. Let X be a random variable with mean 2. Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be unbiased estimators of the second and third moments, respectively, of X about the origin. Find an unbiased estimator of the third moment of X about its mean in terms of $\widehat{\theta_1}$ and $\widehat{\theta_2}$.

Answer: Since, $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the unbiased estimators of the second and third moments of X about origin, we get

$$E\left(\widehat{\theta_1}\right) = E(X) \qquad \text{and} \qquad E\left(\widehat{\theta_2}\right) = E\left(X^2\right).$$

The unbiased estimator of the third moment of X about its mean is

$$\begin{aligned} E\left[(X-2)^3\right] &= E\left[X^3 - 6X^2 + 12X - 8\right] \\ &= E\left[X^3\right] - 6E\left[X^2\right] + 12E\left[X\right] - 8 \\ &= \widehat{\theta_2} - 6\widehat{\theta_1} + 24 - 8 \\ &= \widehat{\theta_2} - 6\widehat{\theta_1} + 16. \end{aligned}$$

Thus, the unbiased estimator of the third moment of X about its mean is $\widehat{\theta_2} - 6\widehat{\theta_1} + 16$.

Example 16.5. Let $X_1, X_2, ..., X_5$ be a sample of size 5 from the uniform distribution on the interval $(0, \theta)$, where θ is unknown. Let the estimator of θ be $k\,X_{\max}$, where k is some constant and $X_{\max}$ is the largest observation. In order $k\,X_{\max}$ to be an unbiased estimator, what should be the value of the constant k ?

Answer: The probability density function of $X_{\max}$ is given by

$$\begin{aligned} g(x) &= \frac{5!}{4!\,0!}\,[F(x)]^4\; f(x) \\ &= 5\left(\frac{x}{\theta}\right)^4 \frac{1}{\theta} \\ &= \frac{5}{\theta^5}\,x^4. \end{aligned}$$

If $k\,X_{\max}$ is an unbiased estimator of θ, then

$$\begin{aligned} \theta &= E\left(k\,X_{\max}\right) \\ &= k\,E\left(X_{\max}\right) \\ &= k\int_0^\theta x\,g(x)\,dx \\ &= k\int_0^\theta \frac{5}{\theta^5}\,x^5\,dx \\ &= \frac{5}{6}\,k\,\theta. \end{aligned}$$

Hence,

$$k = \frac{6}{5}.$$

Example 16.6. Let $X_1, X_2, ..., X_n$ be a sample of size n from a distribution with unknown mean $-\infty < \mu < \infty$, and unknown variance $\sigma^2 > 0$. Show that the statistic $\overline{X}$ and $Y = \frac{X_1+2X_2+\cdots+nX_n}{\frac{n\,(n+1)}{2}}$ are both unbiased estimators of μ. Further, show that $Var\left(\overline{X}\right) < Var(Y)$.

Answer: First, we show that $\overline{X}$ is an unbiased estimator of μ

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \mu = \mu. \end{aligned}$$

Hence, the sample mean $\overline{X}$ is an unbiased estimator of the population mean irrespective of the distribution of X. Next, we show that Y is also an unbiased estimator of μ.

$$\begin{aligned} E\left(Y\right) &= E\left(\frac{X_1 + 2X_2 + \cdots + nX_n}{\frac{n\,(n+1)}{2}}\right) \\ &= \frac{2}{n\,(n+1)} \sum_{i=1}^{n} i\, E\left(X_i\right) \\ &= \frac{2}{n\,(n+1)} \sum_{i=1}^{n} i\, \mu \\ &= \frac{2}{n\,(n+1)}\; \mu\; \frac{n\,(n+1)}{2} \\ &= \mu. \end{aligned}$$

Hence, $\overline{X}$ and Y are both unbiased estimator of the population mean irrespective of the distribution of the population. The variance of $\overline{X}$ is given by

$$\begin{aligned} Var\left[\overline{X}\right] &= Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n^2}\, Var\left[X_1 + X_2 + \cdots + X_n\right] \\ &= \frac{1}{n^2} \sum_{i=1}^{n} Var\left[X_i\right] \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Similarly, the variance of Y can be calculated as follows:

$$\begin{aligned}
Var\,[Y] &= Var\left[\frac{X_1 + 2X_2 + \cdots + nX_n}{\frac{n\,(n+1)}{2}}\right] \\
&= \frac{4}{n^2\,(n+1)^2}\,Var\,[1\,X_1 + 2\,X_2 + \cdots + n\,X_n] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sum_{i=1}^{n} Var\,[i\,X_i] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sum_{i=1}^{n} i^2\,Var\,[X_i] \\
&= \frac{4}{n^2\,(n+1)^2}\,\sigma^2\,\sum_{i=1}^{n} i^2 \\
&= \sigma^2\,\frac{4}{n^2\,(n+1)^2}\,\frac{n\,(n+1)\,(2n+1)}{6} \\
&= \frac{2}{3}\,\frac{2n+1}{(n+1)}\,\frac{\sigma^2}{n} \\
&= \frac{2}{3}\,\frac{2n+1}{(n+1)}\,Var\left[\,\overline{X}\,\right].
\end{aligned}$$

Since $\frac{2}{3}\,\frac{2n+1}{(n+1)} > 1$ for $n \geq 2$, we see that $Var\left[\,\overline{X}\,\right] < Var[\,Y\,]$. This shows that although the estimators $\overline{X}$ and Y are both unbiased estimator of μ, yet the variance of the sample mean $\overline{X}$ is smaller than the variance of Y.

In statistics, between two unbiased estimators one prefers the estimator which has the minimum variance. This leads to our next topic. However, before we move to the next topic we complete this section with some known disadvantages with the notion of unbiasedness. The first disadvantage is that an unbiased estimator for a parameter may not exist. The second disadvantage is that the property of unbiasedness is not invariant under functional transformation, that is, if $\widehat{\theta}$ is an unbiased estimator of θ and g is a function, then $g(\widehat{\theta})$ may not be an unbiased estimator of $g(\theta)$.

16.2. The Relatively Efficient Estimator

We have seen that in Example 16.6 that the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

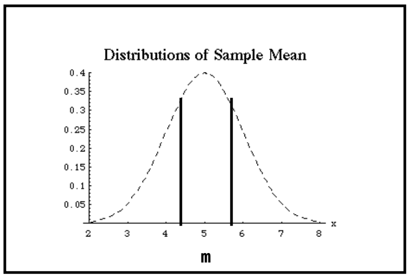
and the statistic

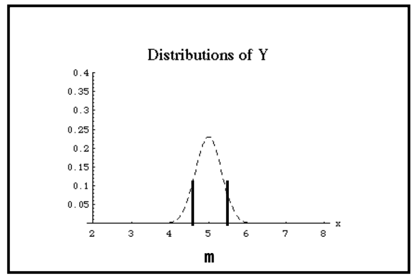
$$Y = \frac{X_1 + 2X_2 + \cdots + nX_n}{1 + 2 + \cdots + n}$$

are both unbiased estimators of the population mean. However, we also seen that

$$Var\left(\overline{X}\right) < Var(Y).$$

The following figure graphically illustrates the shape of the distributions of both the unbiased estimators.





If an unbiased estimator has a smaller variance or dispersion, then it has a greater chance of being close to true parameter θ. Therefore when two estimators of θ are both unbiased, then one should pick the one with the smaller variance.

Definition 16.2. Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of θ. The estimator $\widehat{\theta}_1$ is said to be more efficient than $\widehat{\theta}_2$ if

$$Var\left(\widehat{\theta}_1\right) < Var\left(\widehat{\theta}_2\right).$$

The ratio η given by

$$\eta\left(\widehat{\theta}_1, \widehat{\theta}_2\right) = \frac{Var\left(\widehat{\theta}_2\right)}{Var\left(\widehat{\theta}_1\right)}$$

is called the relative efficiency of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$.

Example 16.7. Let X_1, X_2, X_3 be a random sample of size 3 from a population with mean μ and variance $\sigma^2 > 0$. If the statistics $\overline{X}$ and Y given by

$$Y = \frac{X_1 + 2X_2 + 3X_3}{6}$$

are two unbiased estimators of the population mean μ, then which one is more efficient?

Answer: Since $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, we get

$$\begin{aligned} E\left(\overline{X}\right) &= E\left(\frac{X_1 + X_2 + X_3}{3}\right) \\ &= \frac{1}{3}\,\left(E(X_1) + E(X_2) + E(X_3)\right) \\ &= \frac{1}{3}\,3\mu \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} E(Y) &= E\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right) \\ &= \frac{1}{6}\,\left(E(X_1) + 2E(X_2) + 3E(X_3)\right) \\ &= \frac{1}{6}\,6\mu \\ &= \mu. \end{aligned}$$

Therefore both $\overline{X}$ and Y are unbiased. Next we determine the variance of both the estimators. The variances of these estimators are given by

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1 + X_2 + X_3}{3}\right) \\ &= \frac{1}{9}\,\left[Var(X_1) + Var(X_2) + Var(X_3)\right] \\ &= \frac{1}{9}\,3\sigma^2 \\ &= \frac{12}{36}\,\sigma^2 \end{aligned}$$

and

$$\begin{aligned} Var(Y) &= Var\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right) \\ &= \frac{1}{36}\,\left[Var(X_1) + 4Var(X_2) + 9Var(X_3)\right] \\ &= \frac{1}{36}\,14\sigma^2 \\ &= \frac{14}{36}\,\sigma^2. \end{aligned}$$

Therefore

$$\frac{12}{36}\,\sigma^2 = Var\left(\overline{X}\right) < Var(Y) = \frac{14}{36}\,\sigma^2.$$

Hence, $\overline{X}$ is more efficient than the estimator Y. Further, the relative efficiency of $\overline{X}$ with respect to Y is given by

$$\eta\left(\overline{X}, Y\right) = \frac{14}{12} = \frac{7}{6}.$$

Example 16.8. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population with density

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Are the estimators X_1 and $\overline{X}$ unbiased? Given, X_1 and $\overline{X}$, which one is more efficient estimator of θ ?

Answer: Since the population X is exponential with parameter θ, that is $X \sim EXP(\theta)$, the mean and variance of it are given by

$$E(X) = \theta \qquad \text{and} \qquad Var(X) = \theta^2.$$

Since $X_1, X_2, ..., X_n$ is a random sample from X, we see that the statistic $X_1 \sim EXP(\theta)$. Hence, the expected value of X_1 is θ and thus it is an unbiased estimator of the parameter θ. Also, the sample mean is an unbiased estimator of θ since

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\, n\theta \\ &= \theta. \end{aligned}$$

Next, we compute the variances of the unbiased estimators X_1 and $\overline{X}$. It is easy to see that

$$Var\left(X_1\right) = \theta^2$$

and

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2}\, n\theta^2 \\ &= \frac{\theta^2}{n}. \end{aligned}$$

Hence

$$\frac{\theta^2}{n} = Var\left(\overline{X}\right) < Var\left(X_1\right) = \theta^2.$$

Thus $\overline{X}$ is more efficient than X_1 and the relative efficiency of $\overline{X}$ with respect to X_1 is

$$\eta(\overline{X},\, X_1) = \frac{\theta^2}{\frac{\theta^2}{n}} = n.$$

Example 16.9. Let X_1, X_2, X_3 be a random sample of size 3 from a population with density

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x\, e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where λ is a parameter. Are the estimators given by

$$\widehat{\lambda_1} = \frac{1}{4}\,\left(X_1 + 2X_2 + X_3\right) \qquad \text{and} \qquad \widehat{\lambda_2} = \frac{1}{9}\,\left(4X_1 + 3X_2 + 2X_3\right)$$

unbiased? Given, $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$, which one is more efficient estimator of λ ? Find an unbiased estimator of λ whose variance is smaller than the variances of $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$.

Answer: Since each $X_i \sim POI(\lambda)$, we get

$$E\left(X_i\right) = \lambda \qquad \text{and} \qquad Var\left(X_i\right) = \lambda.$$

It is easy to see that

$$\begin{aligned} E\left(\widehat{\lambda_1}\right) &= \frac{1}{4}\left(E\left(X_1\right) + 2E\left(X_2\right) + E\left(X_3\right)\right) \\ &= \frac{1}{4}\;4\lambda \\ &= \lambda, \end{aligned}$$

and

$$\begin{aligned} E\left(\widehat{\lambda_2}\right) &= \frac{1}{9}\left(4E\left(X_1\right) + 3E\left(X_2\right) + 2E\left(X_3\right)\right) \\ &= \frac{1}{9}\;9\lambda \\ &= \lambda. \end{aligned}$$

Thus, both $\widehat{\lambda_1}$ and $\widehat{\lambda_2}$ are unbiased estimators of λ. Now we compute their variances to find out which one is more efficient. It is easy to note that

$$\begin{aligned} Var\left(\widehat{\lambda_1}\right) &= \frac{1}{16}\left(Var\left(X_1\right) + 4Var\left(X_2\right) + Var\left(X_3\right)\right) \\ &= \frac{1}{16}\,6\lambda \\ &= \frac{6}{16}\lambda \\ &= \frac{486}{1296}\lambda, \end{aligned}$$

and

$$\begin{aligned} Var\left(\widehat{\lambda_2}\right) &= \frac{1}{81}\left(16Var\left(X_1\right) + 9Var\left(X_2\right) + 4Var\left(X_3\right)\right) \\ &= \frac{1}{81}\,29\lambda \\ &= \frac{29}{81}\lambda \\ &= \frac{464}{1296}\lambda, \end{aligned}$$

Since,

$$Var\left(\widehat{\lambda_2}\right) < Var\left(\widehat{\lambda_1}\right),$$

the estimator $\widehat{\lambda_2}$ is efficient than the estimator $\widehat{\lambda_1}$. We have seen in section 16.1 that the sample mean is always an unbiased estimator of the population mean irrespective of the population distribution. The variance of the sample mean is always equals to $\frac{1}{n}$ times the population variance, where n denotes the sample size. Hence, we get

$$Var\left(\overline{X}\right) = \frac{\lambda}{3} = \frac{432}{1296}\,\lambda.$$

Therefore, we get

$$Var\left(\overline{X}\right) < Var\left(\widehat{\lambda_2}\right) < Var\left(\widehat{\lambda_1}\right).$$

Thus, the sample mean has even smaller variance than the two unbiased estimators given in this example.

In view of this example, now we have encountered a new problem. That is how to find an unbiased estimator which has the smallest variance among all unbiased estimators of a given parameter. We resolve this issue in the next section.

16.3. The Uniform Minimum Variance Unbiased Estimator

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population with probability density function $f(x;\theta)$. Recall that an estimator $\widehat{\theta}$ of θ is a function of the random variables $X_1, X_2, ..., X_n$ which does depend on θ.

Definition 16.3. An unbiased estimator $\widehat{\theta}$ of θ is said to be a uniform minimum variance unbiased estimator of θ if and only if

$$Var\left(\widehat{\theta}\right) \leq Var\left(\widehat{T}\right)$$

for any unbiased estimator $\widehat{T}$ of θ.

If an estimator $\widehat{\theta}$ is unbiased then the mean of this estimator is equal to the parameter θ, that is

$$E\left(\widehat{\theta}\right) = \theta$$

and the variance of $\widehat{\theta}$ is

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= E\left[\left(\widehat{\theta} - E\left(\widehat{\theta}\right)\right)^2\right] \\ &= E\left[\left(\widehat{\theta} - \theta\right)^2\right]. \end{aligned}$$

This variance, if exists, is a function of the unbiased estimator $\widehat{\theta}$ and it has a minimum in the class of all unbiased estimators of θ. Therefore we have an alternative definition of the uniform minimum variance unbiased estimator.

Definition 16.4. An unbiased estimator $\widehat{\theta}$ of θ is said to be a uniform minimum variance unbiased estimator of θ if it minimizes the variance $E\left[\left(\widehat{\theta} - \theta\right)^2\right]$.

Example 16.10. Let $\widehat{\theta_1}$ and $\widehat{\theta_2}$ be unbiased estimators of θ. Suppose $Var\left(\widehat{\theta_1}\right) = 1$, $Var\left(\widehat{\theta_2}\right) = 2$ and $Cov\left(\widehat{\theta_1}, \widehat{\theta_2}\right) = \frac{1}{2}$. What are the values of c_1 and c_2 for which $c_1\widehat{\theta_1} + c_2\widehat{\theta_2}$ is an unbiased estimator of θ with minimum variance among unbiased estimators of this type?

Answer: We want $c_1\widehat{\theta_1} + c_2\widehat{\theta_2}$ to be a minimum variance unbiased estimator of θ. Then

$$\begin{aligned} & E\left[c_1\widehat{\theta_1} + c_2\widehat{\theta_2}\right] = \theta \\ \Rightarrow \quad & c_1\, E\left[\widehat{\theta_1}\right] + c_2\, E\left[\widehat{\theta_2}\right] = \theta \\ \Rightarrow \quad & c_1\,\theta + c_2\,\theta = \theta \\ \Rightarrow \quad & c_1 + c_2 = 1 \\ \Rightarrow \quad & c_2 = 1 - c_1. \end{aligned}$$

Therefore

$$\begin{aligned} Var\left[c_1\widehat{\theta}_1 + c_2\widehat{\theta}_2\right] &= c_1^2\, Var\left[\widehat{\theta}_1\right] + c_2^2\, Var\left[\widehat{\theta}_2\right] + 2\, c_1\, c_2\, Cov\left(\widehat{\theta}_1,\, \widehat{\theta}_1\right) \\ &= c_1^2 + 2c_2^2 + c_1c_2 \\ &= c_1^2 + 2(1-c_1)^2 + c_1(1-c_1) \\ &= 2(1-c_1)^2 + c_1 \\ &= 2 + 2c_1^2 - 3c_1. \end{aligned}$$

Hence, the variance $Var\left[c_1\widehat{\theta}_1 + c_2\widehat{\theta}_2\right]$ is a function of c_1. Let us denote this function by $\phi(c_1)$, that is

$$\phi(c_1) := Var\left[c_1\widehat{\theta}_1 + c_2\widehat{\theta}_2\right] = 2 + 2c_1^2 - 3c_1.$$

Taking the derivative of $\phi(c_1)$ with respect to c_1, we get

$$\frac{d}{dc_1}\phi(c_1) = 4c_1 - 3.$$

Setting this derivative to zero and solving for c_1, we obtain

$$4c_1 - 3 = 0 \qquad \Rightarrow \quad c_1 = \frac{3}{4}.$$

Therefore

$$c_2 = 1 - c_1 = 1 - \frac{3}{4} = \frac{1}{4}.$$

In Example 16.10, we saw that if $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are any two unbiased estimators of θ, then $c\,\widehat{\theta}_1 + (1-c)\,\widehat{\theta}_2$ is also an unbiased estimator of θ for any $c \in \mathbb{R}$. Hence given two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$,

$$\mathcal{C} = \left\{\widehat{\theta} \mid \widehat{\theta} = c\,\widehat{\theta}_1 + (1-c)\,\widehat{\theta}_2, \;\; c \in \mathbb{R}\right\}$$

forms an uncountable class of unbiased estimators of θ. When the variances of $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are known along with the their covariance, then in Example 16.10 we were able to determine the minimum variance unbiased estimator in the class $\mathcal{C}$. If the variances of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are not known, then it is very difficult to find the minimum variance estimator even in the class of estimators $\mathcal{C}$. Notice that $\mathcal{C}$ is a subset of the class of all unbiased estimators and finding a minimum variance unbiased estimator in this class is a difficult task.

One way to find a uniform minimum variance unbiased estimator for a parameter is to use the Cramér-Rao lower bound or the Fisher information inequality.

Theorem 16.1. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with probability density $f(x;\theta)$, where θ is a scalar parameter. Let $\widehat{\theta}$ be any unbiased estimator of θ. Suppose the likelihood function $L(\theta)$ is a differentiable function of θ and satisfies

$$\begin{aligned} \frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,...,x_n)\, L(\theta)\, dx_1\cdots dx_n \\ = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,...,x_n)\, \frac{d}{d\theta}L(\theta)\, dx_1\cdots dx_n \end{aligned} \tag{1}$$

for any $h(x_1,...,x_n)$ with $E(h(X_1,...,X_n)) < \infty$. Then

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right]}. \tag{CR1}$$

Proof: Since $L(\theta)$ is the joint probability density function of the sample $X_1, X_2, ..., X_n$,

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(\theta)\, dx_1\cdots dx_n = 1. \tag{2}$$

Differentiating (2) with respect to θ we have

$$\frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} L(\theta)\, dx_1\cdots dx_n = 0$$

and use of (1) with $h(x_1,...,x_n) = 1$ yields

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{d}{d\theta}L(\theta)\ dx_1\cdots dx_n = 0. \tag{3}$$

Rewriting (3) as

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{dL(\theta)}{d\theta}\,\frac{1}{L(\theta)}\ L(\theta)\ dx_1\cdots dx_n = 0$$

we see that

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \frac{d\ln L(\theta)}{d\theta}\ L(\theta)\ dx_1\cdots dx_n = 0.$$

Hence

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta \, \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 0. \tag{4}$$

Since $\widehat{\theta}$ is an unbiased estimator of θ, we see that

$$E\left(\widehat{\theta}\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, L(\theta) \, dx_1 \cdots dx_n = \theta. \tag{5}$$

Differentiating (5) with respect to θ, we have

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, L(\theta) \, dx_1 \cdots dx_n = 1.$$

Again using (1) with $h(X_1, ..., X_n) = \widehat{\theta}$, we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, \frac{d}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 1. \tag{6}$$

Rewriting (6) as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, \frac{dL(\theta)}{d\theta} \, \frac{1}{L(\theta)} \, L(\theta) \; dx_1 \cdots dx_n = 1$$

we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \, \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 1. \tag{7}$$

From (4) and (7), we obtain

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \; \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n = 1. \tag{8}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
1 &= \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right) \; \frac{d \ln L(\theta)}{d\theta} \, L(\theta) \; dx_1 \cdots dx_n \right)^2 \\
&\leq \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\widehat{\theta} - \theta\right)^2 \; L(\theta) \; dx_1 \cdots dx_n \right) \\
&\quad \cdot \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{d \ln L(\theta)}{d\theta} \right)^2 \; L(\theta) \; dx_1 \cdots dx_n \right) \\
&= Var\left(\widehat{\theta}\right) \; E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta} \right)^2 \right].
\end{aligned}$$

Therefore

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}$$

and the proof of theorem is now complete.

If $L(\theta)$ is twice differentiable with respect to θ, the inequality (CR1) can be stated equivalently as

$$Var\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]}. \qquad \text{(CR2)}$$

The inequalities (CR1) and (CR2) are known as Cramér-Rao lower bound for the variance of $\widehat{\theta}$ or the Fisher information inequality. The condition (1) interchanges the order on integration and differentiation. Therefore any distribution whose range depend on the value of the parameter is not covered by this theorem. Hence distribution like the uniform distribution may not be analyzed using the Cramér-Rao lower bound.

If the estimator $\widehat{\theta}$ is minimum variance in addition to being unbiased, then equality holds. We state this as a theorem without giving a proof.

Theorem 16.2. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with probability density $f(x;\theta)$, where θ is a parameter. If $\widehat{\theta}$ is an unbiased estimator of θ and

$$Var\left(\widehat{\theta}\right) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]},$$

then $\widehat{\theta}$ is a uniform minimum variance unbiased estimator of θ. The converse of this is not true.

Definition 16.5. An unbiased estimator $\widehat{\theta}$ is called an efficient estimator if it satisfies Cramér-Rao lower bound, that is

$$Var\left(\widehat{\theta}\right) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}.$$

In view of the above theorem it is easy to note that an efficient estimator of a parameter is always a uniform minimum variance unbiased estimator of a parameter. However, not every uniform minimum variance unbiased estimator of a parameter is efficient. In other words not every uniform minimum variance unbiased estimators of a parameter satisfy the Cramér-Rao lower bound

$$Var\left(\widehat{\theta}\right) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right]}.$$

Example 16.11. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with density function

$$f(x;\theta) = \begin{cases} 3\theta\, x^2 e^{-\theta x^3} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the Cramér-Rao lower bound for the variance of unbiased estimator of the parameter θ ?

Answer: Let $\widehat{\theta}$ be an unbiased estimator of θ. Cramér-Rao lower bound for the variance of $\widehat{\theta}$ is given by

$$Var\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta^2}\right]},$$

where $L(\theta)$ denotes the likelihood function of the given random sample $X_1, X_2, ..., X_n$. Since, the likelihood function of the sample is

$$L(\theta) = \prod_{i=1}^{n} 3\theta\, x_i^2 e^{-\theta x_i^3}$$

we get

$$\ln L(\theta) = n\, \ln\theta + \sum_{i=1}^{n} \ln\left(3x_i^2\right) - \theta \sum_{i=1}^{n} x_i^3.$$

$$\frac{\partial \ln L(\theta)}{\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i^3,$$

and

$$\frac{\partial^2 \ln L(\theta)}{\partial\theta^2} = -\frac{n}{\theta^2}.$$

Hence, using this in the Cramér-Rao inequality, we get

$$Var\left(\widehat{\theta}\right) \geq \frac{\theta^2}{n}.$$

Thus the Cramér-Rao lower bound for the variance of the unbiased estimator of θ is $\frac{\theta^2}{n}$.

Example 16.12. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with unknown mean μ and known variance $\sigma^2 > 0$. What is the maximum likelihood estimator of μ? Is this maximum likelihood estimator an efficient estimator of μ?

Answer: The probability density function of the population is

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\,\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\,(x-\mu)^2}.$$

Thus

$$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\,(x-\mu)^2$$

and hence

$$\ln L(\mu) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the derivative of $\ln L(\mu)$ with respect to μ, we get

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu).$$

Setting this derivative to zero and solving for μ, we see that $\widehat{\mu} = \overline{X}$.

The variance of $\overline{X}$ is given by

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Next we determine the Cramér-Rao lower bound for the estimator $\overline{X}$. We already know that

$$\frac{d\ln L(\mu)}{d\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

and hence

$$\frac{d^2\ln L(\mu)}{d\mu^2} = -\frac{n}{\sigma^2}.$$

Therefore

$$E\left(\frac{d^2\ln L(\mu)}{d\mu^2}\right) = -\frac{n}{\sigma^2}$$

and

$$-\frac{1}{E\left(\frac{d^2 \ln L(\mu)}{d\mu^2}\right)} = \frac{\sigma^2}{n}.$$

Thus

$$Var\left(\overline{X}\right) = -\frac{1}{E\left(\frac{d^2 \ln L(\mu)}{d\mu^2}\right)}$$

and $\overline{X}$ is an efficient estimator of μ. Since every efficient estimator is a uniform minimum variance unbiased estimator, therefore $\overline{X}$ is a uniform minimum variance unbiased estimator of μ.

Example 16.13. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with known mean μ and unknown variance $\sigma^2 > 0$. What is the maximum likelihood estimator of σ^2? Is this maximum likelihood estimator a uniform minimum variance unbiased estimator of σ^2?

Answer: Let us write $\theta = \sigma^2$. Then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\, e^{-\frac{1}{2\theta}(x-\mu)^2}$$

and

$$\ln L(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Differentiating $\ln L(\theta)$ with respect to θ, we have

$$\frac{d}{d\theta}\ln L(\theta) = -\frac{n}{2}\,\frac{1}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Setting this derivative to zero and solving for θ, we see that

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2.$$

Next we show that this estimator is unbiased. For this we consider

$$\begin{aligned}
E\left(\widehat{\theta}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\
&= \frac{\sigma^2}{n}\,E\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right) \\
&= \frac{\theta}{n}\,E(\,\chi^2(n)\,) \\
&= \frac{\theta}{n}\,n = \theta.
\end{aligned}$$

Hence $\widehat{\theta}$ is an unbiased estimator of θ. The variance of $\widehat{\theta}$ can be obtained as follows:

$$\begin{aligned} Var\left(\widehat{\theta}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= \frac{\sigma^4}{n}\,Var\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\right) \\ &= \frac{\theta^2}{n^2}\,Var(\,\chi^2(n)\,) \\ &= \frac{\theta^2}{n^2}\;4\;\frac{n}{2} \\ &= \frac{2\theta^2}{n} = \frac{2\sigma^4}{n}. \end{aligned}$$

Finally we determine the Cramér-Rao lower bound for the variance of $\widehat{\theta}$. The second derivative of $\ln L(\theta)$ with respect to θ is

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = \frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Hence

$$\begin{aligned} E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right) &= \frac{n}{2\theta^2} - \frac{1}{\theta^3}E\left(\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= \frac{n}{2\theta^2} - \frac{\theta}{\theta^3}E\left(\,\chi^2(n)\,\right) \\ &= \frac{n}{2\theta^2} - \frac{n}{\theta^2} \\ &= -\frac{n}{2\theta^2} \end{aligned}$$

Thus

$$-\frac{1}{E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)} = \frac{\theta^2}{n} = \frac{2\sigma^4}{n}.$$

Therefore

$$Var\left(\widehat{\theta}\right) = -\frac{1}{E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)}.$$

Hence $\widehat{\theta}$ is an efficient estimator of θ. Since every efficient estimator is a uniform minimum variance unbiased estimator, therefore $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ is a uniform minimum variance unbiased estimator of σ^2.

Example 16.14. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a normal population known mean μ and variance $\sigma^2 > 0$. Show that $S^2 =$

$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator of σ^2. Further, show that S^2 can not attain the Cramér-Rao lower bound.

Answer: From Example 16.2, we know that S^2 is an unbiased estimator of σ^2. The variance of S^2 can be computed as follows:

$$\begin{aligned}
Var\left(S^2\right) &= Var\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= \frac{\sigma^4}{(n-1)^2}\, Var\left(\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2\right) \\
&= \frac{\sigma^4}{(n-1)^2}\, Var(\,\chi^2(n-1)\,) \\
&= \frac{\sigma^4}{(n-1)^2}\; 2\;(n-1) \\
&= \frac{2\sigma^4}{n-1}.
\end{aligned}$$

Next we let $\theta = \sigma^2$ and determine the Cramér-Rao lower bound for the variance of S^2. The second derivative of $\ln L(\theta)$ with respect to θ is

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = \frac{n}{2\theta^2} - \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Hence

$$\begin{aligned}
E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right) &= \frac{n}{2\theta^2} - \frac{1}{\theta^3} E\left(\sum_{i=1}^{n}(X_i - \mu)^2\right) \\
&= \frac{n}{2\theta^2} - \frac{\theta}{\theta^3} E\left(\,\chi^2(n)\,\right) \\
&= \frac{n}{2\theta^2} - \frac{n}{\theta^2} \\
&= -\frac{n}{2\theta^2}
\end{aligned}$$

Thus

$$-\frac{1}{E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)} = \frac{\theta^2}{n} = \frac{2\sigma^4}{n}.$$

Hence

$$\frac{2\sigma^4}{n-1} = Var\left(S^2\right) > -\frac{1}{E\left(\frac{d^2 \ln L(\theta)}{d\theta^2}\right)} = \frac{2\sigma^4}{n}.$$

This shows that S^2 can not attain the Cramér-Rao lower bound.

The disadvantages of Cramér-Rao lower bound approach are the followings: (1) Not every density function $f(x;\theta)$ satisfies the assumptions of Cramér-Rao theorem and (2) not every allowable estimator attains the Cramér-Rao lower bound. Hence in any one of these situations, one does not know whether an estimator is a uniform minimum variance unbiased estimator or not.

16.4. Sufficient Estimator

In many situations, we can not easily find the distribution of the estimator $\widehat{\theta}$ of a parameter θ even though we know the distribution of the population. Therefore, we have no way to know whether our estimator $\widehat{\theta}$ is unbiased or biased. Hence, we need some other criteria to judge the quality of an estimator. Sufficiency is one such criteria for judging the quality of an estimator.

Recall that an estimator of a population parameter is a function of the sample values that does not contain the parameter. An estimator summarizes the information found in the sample about the parameter. If an estimator summarizes just as much information about the parameter being estimated as the sample does, then the estimator is called a sufficient estimator.

Definition 16.6. Let $X \sim f(x;\theta)$ be a population and let $X_1, X_2, ..., X_n$ be a random sample of size n from this population X. An estimator $\widehat{\theta}$ of the parameter θ is said to be a sufficient estimator of θ if the conditional distribution of the sample given the estimator $\widehat{\theta}$ does not depend on the parameter θ.

Example 16.15. If $X_1, X_2, ..., X_n$ is a random sample from the distribution with probability density function

$$f(x;\theta) = \begin{cases} \theta^x\,(1-\theta)^{1-x} & \text{if } x = 0, 1 \\ \\ 0 & \text{elsewhere ,} \end{cases}$$

where $0 < \theta < 1$. Show that $Y = \sum_{i=1}^n X_i$ is a sufficient statistic of θ.

Answer: First, we find the distribution of the sample. This is given by

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y(1-\theta)^{n-y}.$$

Since, each $X_i \sim BER(\theta)$, we have

$$Y = \sum_{i=1}^n X_i \sim BIN(n, \theta).$$

Therefore, the probability density function of Y is given by

$$g(y) = \binom{n}{y} \theta^y \, (1-\theta)^{n-y}.$$

Further, since each $X_i \sim BER(\theta)$, the space of each X_i is given by

$$R_{X_i} = \{\, 0, 1 \,\}.$$

Therefore, the space of the random variable $Y = \sum_{i=1}^{n} X_i$ is given by

$$R_Y = \{\, 0, 1, 2, 3, 4, ..., n \,\}.$$

Let A be the event $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ and B denotes the event $(Y = y)$. Then $A \subset B$ and therefore $A \bigcap B = A$. Now, we find the conditional density of the sample given the estimator Y, that is

$$\begin{aligned}
f(x_1, x_2, ..., x_n / Y = y) &= P\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \,/ Y = y\right) \\
&= P(A/B) \\
&= \frac{P(A \bigcap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{f(x_1, x_2, ..., x_n)}{g(y)} \\
&= \frac{\theta^y (1-\theta)^{n-y}}{\binom{n}{y} \theta^y \, (1-\theta)^{n-y}} \\
&= \frac{1}{\binom{n}{y}}.
\end{aligned}$$

Hence, the conditional density of the sample given the statistic Y is independent of the parameter θ. Therefore, by definition Y is a sufficient statistic.

Example 16.16. If $X_1, X_2, ..., X_n$ is a random sample from the distribution with probability density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{elsewhere ,} \end{cases}$$

where $-\infty < \theta < \infty$. What is the maximum likelihood estimator of θ ? Is this maximum likelihood estimator sufficient estimator of θ ?

Answer: We have seen in Chapter 15 that the maximum likelihood estimator of θ is $Y = X_{(1)}$, that is the first order statistic of the sample. Let us find the probability density of this statistic, which is given by

$$\begin{aligned} g(y) &= \frac{n!}{(n-1)!} \, [F(y)]^0 \; f(y) \; [1 - F(y)]^{n-1} \\ &= n \, f(y) \; [1 - F(y)]^{n-1} \\ &= n \, e^{-(y-\theta)} \; \left[1 - \left\{ 1 - e^{-(y-\theta)} \right\} \right]^{n-1} \\ &= n \, e^{n\theta} \, e^{-ny}. \end{aligned}$$

The probability density of the random sample is

$$\begin{aligned} f(x_1, x_2, ..., x_n) &= \prod_{i=1}^{n} e^{-(x_i - \theta)} \\ &= e^{n\theta} \, e^{-n\,\overline{x}}, \end{aligned}$$

where $n\overline{x} = \sum_{i=1}^{n} x_i$. Let A be the event $(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$ and B denotes the event $(Y = y)$. Then $A \subset B$ and therefore $A \bigcap B = A$. Now, we find the conditional density of the sample given the estimator Y, that is

$$\begin{aligned} f(x_1, x_2, ..., x_n / Y = y) &= P\left(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \, / Y = y\right) \\ &= P(A/B) \\ &= \frac{P(A \bigcap B)}{P(B)} \\ &= \frac{P(A)}{P(B)} \\ &= \frac{f(x_1, x_2, ..., x_n)}{g(y)} \\ &= \frac{e^{n\theta} \, e^{-n\,\overline{x}}}{n \, e^{n\theta} \, e^{-n\,y}} \\ &= \frac{e^{-n\,\overline{x}}}{n \, e^{-n\,y}}. \end{aligned}$$

Hence, the conditional density of the sample given the statistic Y is independent of the parameter θ. Therefore, by definition Y is a sufficient statistic.

We have seen that to verify whether an estimator is sufficient or not one has to examine the conditional density of the sample given the estimator. To compute this conditional density one has to use the density of the estimator. The density of the estimator is not always easy to find. Therefore, verifying the sufficiency of an estimator using this definition is not always easy. The following *factorization theorem* of Fisher and Neyman helps to decide when an estimator is sufficient.

Theorem 16.3. Let $X_1, X_2, ..., X_n$ denote a random sample with probability density function $f(x_1, x_2, ..., x_n; \theta)$, which depends on the population parameter θ. The estimator $\widehat{\theta}$ is sufficient for θ if and only if

$$f(x_1, x_2, ..., x_n; \theta) = \phi(\widehat{\theta}, \theta)\, h(x_1, x_2, ..., x_n)$$

where ϕ depends on $x_1, x_2, ..., x_n$ only through $\widehat{\theta}$ and $h(x_1, x_2, ..., x_n)$ does not depend on θ.

Now we give two examples to illustrate the factorization theorem.

Example 16.17. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x; \lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\lambda > 0$ is a parameter. Find the maximum likelihood estimator of λ and show that the maximum likelihood estimator of λ is sufficient estimator of the parameter λ.

Answer: First, we find the density of the sample or the likelihood function of the sample. The likelihood function of the sample is given by

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{n} f(x_i; \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^{x_i}\, e^{-\lambda}}{x_i!} \\ &= \frac{\lambda^{n\overline{X}} e^{-n\lambda}}{\prod_{i=1}^{n} (x_i!)}. \end{aligned}$$

Taking the logarithm of the likelihood function, we get

$$\ln L(\lambda) = n\overline{x}\ \ln\lambda - n\lambda - \ln\prod_{i=1}^{n}(x_i!).$$

Therefore

$$\frac{d}{d\lambda}\ln L(\lambda) = \frac{1}{\lambda}\,n\overline{x} - n.$$

Setting this derivative to zero and solving for λ, we get

$$\lambda = \overline{x}.$$

The second derivative test assures us that the above λ is a maximum. Hence, the maximum likelihood estimator of λ is the sample mean $\overline{X}$. Next, we show that $\overline{X}$ is sufficient, by using the Factorization Theorem of Fisher and Neyman. We factor the joint density of the sample as

$$\begin{aligned} L(\lambda) &= \frac{\lambda^{n\overline{x}}e^{-n\lambda}}{\prod_{i=1}^{n}(x_i!)} \\ &= \left[\lambda^{n\overline{x}}\,e^{-n\lambda}\right]\ \frac{1}{\prod_{i=1}^{n}(x_i!)} \\ &= \phi(\overline{X},\lambda)\,h\,(x_1,x_2,...,x_n)\,. \end{aligned}$$

Therefore, the estimator $\overline{X}$ is a sufficient estimator of λ.

Example 16.18. Let $X_1, X_2, ..., X_n$ be a random sample from a normal distribution with density function

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(x-\mu)^2},$$

where $-\infty < \mu < \infty$ is a parameter. Find the maximum likelihood estimator of μ and show that the maximum likelihood estimator of μ is a sufficient estimator.

Answer: We know that the maximum likelihood estimator of μ is the sample mean $\overline{X}$. Next, we show that this maximum likelihood estimator $\overline{X}$ is a sufficient estimator of μ. The joint density of the sample is given by

$$
\begin{aligned}
f(x_1, x_2, ..., x_n; \mu) \\
&= \prod_{i=1}^{n} f(x_i; \mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\overline{x})+(\overline{x}-\mu)\right]^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\overline{x})^2+2(x_i-\overline{x})(\overline{x}-\mu)+(\overline{x}-\mu)^2\right]} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\overline{x})^2+(\overline{x}-\mu)^2\right]} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{n}{2}(\overline{x}-\mu)^2}\, e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^2}
\end{aligned}
$$

Hence, by the Factorization Theorem, $\overline{X}$ is a sufficient estimator of the population mean.

Note that the probability density function of the Example 16.17 which is

$$
f(x; \lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{elsewhere ,} \end{cases}
$$

can be written as

$$
f(x; \lambda) = e^{\{x \ln\lambda - \ln x! - \lambda\}}
$$

for $x = 0, 1, 2, ...$ This density function is of the form

$$
f(x; \lambda) = e^{\{K(x)A(\lambda)+S(x)+B(\lambda)\}}.
$$

Similarly, the probability density function of the Example 16.12, which is

$$
f(x; \mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2}
$$

can also be written as

$$f(x;\mu) = e^{\{x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \frac{1}{2}\ln(2\pi)\}}.$$

This probability density function is of the form

$$f(x;\mu) = e^{\{K(x)A(\mu)+S(x)+B(\mu)\}}.$$

We have also seen that in both the examples, the sufficient estimators were the sample mean $\overline{X}$, which can be written as $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Our next theorem gives a general result in this direction. The following theorem is known as the Pitman-Koopman theorem.

Theorem 16.4. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density function of the exponential form

$$f(x;\theta) = e^{\{K(x)A(\theta)+S(x)+B(\theta)\}}$$

on a support free of θ. Then the statistic $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter θ.

Proof: The joint density of the sample is

$$\begin{aligned}
f(x_1, x_2, ..., x_n; \theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\
&= \prod_{i=1}^{n} e^{\{K(x_i)A(\theta)+S(x_i)+B(\theta)\}} \\
&= e^{\left\{\sum_{i=1}^{n} K(x_i)A(\theta) + \sum_{i=1}^{n} S(x_i) + n\,B(\theta)\right\}} \\
&= e^{\left\{\sum_{i=1}^{n} K(x_i)A(\theta) + n\,B(\theta)\right\}}\, e^{\left[\sum_{i=1}^{n} S(x_i)\right]}.
\end{aligned}$$

Hence by the Factorization Theorem the estimator $\sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter θ. This completes the proof.

Example 16.19. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Using the Pitman-Koopman Theorem find a sufficient estimator of θ.

Answer: The Pitman-Koopman Theorem says that if the probability density function can be expressed in the form of

$$f(x;\theta) = e^{\{K(x)A(\theta)+S(x)+B(\theta)\}}$$

then $\sum_{i=1}^n K(X_i)$ is a sufficient statistic for θ. The given population density can be written as

$$\begin{aligned} f(x;\theta) &= \theta\, x^{\theta-1} \\ &= e^{\{\ln\left[\theta\, x^{\theta-1}\right]} \\ &= e^{\{\ln\theta+(\theta-1)\ln x\}}. \end{aligned}$$

Thus,

$$K(x) = \ln x \qquad\qquad A(\theta) = \theta$$
$$S(x) = -\ln x \qquad\qquad B(\theta) = \ln\theta.$$

Hence by Pitman-Koopman Theorem,

$$\begin{aligned} \sum_{i=1}^n K(X_i) &= \sum_{i=1}^n \ln X_i \\ &= \ln \prod_{i=1}^n X_i. \end{aligned}$$

Thus $\ln \prod_{i=1}^n X_i$ is a sufficient statistic for θ.

Remark 16.1. Notice that $\displaystyle\prod_{i=1}^n X_i$ is also a sufficient statistic of θ, since knowing $\ln\left(\displaystyle\prod_{i=1}^n X_i\right)$, we also know $\displaystyle\prod_{i=1}^n X_i$.

Example 16.20. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. Find a sufficient estimator of θ.

Answer: First, we rewrite the population density in the exponential form. That is

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta}\, e^{-\frac{x}{\theta}} \\ &= e^{\ln\left[\frac{1}{\theta}\, e^{-\frac{x}{\theta}}\right]} \\ &= e^{-\ln\theta - \frac{x}{\theta}}. \end{aligned}$$

Hence

$$\begin{aligned} K(x) &= x \qquad & A(\theta) &= -\frac{1}{\theta} \\ S(x) &= 0 \qquad & B(\theta) &= -\ln\theta. \end{aligned}$$

Hence by Pitman-Koopman Theorem,

$$\sum_{i=1}^{n} K(X_i) = \sum_{i=1}^{n} X_i = n\,\overline{X}.$$

Thus, $n\overline{X}$ is a sufficient statistic for θ. Since knowing $n\overline{X}$, we also know $\overline{X}$, the estimator $\overline{X}$ is also a sufficient estimator of θ.

Example 16.21. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. Can Pitman-Koopman Theorem be used to find a sufficient statistic for θ?

Answer: No. We can not use Pitman-Koopman Theorem to find a sufficient statistic for θ since the domain where the population density is nonzero is not free of θ.

Next, we present the connection between the maximum likelihood estimator and the sufficient estimator. If there is a sufficient estimator for the parameter θ and if the maximum likelihood estimator of this θ is unique, then the maximum likelihood estimator is a function of the sufficient estimator. That is

$$\widehat{\theta}_{\text{ML}} = \psi(\widehat{\theta}_{\text{S}}),$$

where ψ is a real valued function, $\widehat{\theta}_{\text{ML}}$ is the maximum likelihood estimator of θ, and $\widehat{\theta}_{\text{S}}$ is the sufficient estimator of θ.

Similarly, a connection can be established between the uniform minimum variance unbiased estimator and the sufficient estimator of a parameter θ. If there is a sufficient estimator for the parameter θ and if the uniform minimum variance unbiased estimator of this θ is unique, then the uniform minimum variance unbiased estimator is a function of the sufficient estimator. That is

$$\widehat{\theta}_{\text{MVUE}} = \eta(\widehat{\theta}_{\text{S}}),$$

where η is a real valued function, $\widehat{\theta}_{\text{MVUE}}$ is the uniform minimum variance unbiased estimator of θ, and $\widehat{\theta}_{\text{S}}$ is the sufficient estimator of θ.

Finally, we may ask "If there are sufficient estimators, why are not there necessary estimators?" In fact, there are. Dynkin (1951) gave the following definition.

Definition 16.7. An estimator is said to be a necessary estimator if it can be written as a function of every sufficient estimators.

16.5. Consistent Estimator

Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density $f(x;\theta)$. Let $\widehat{\theta}$ be an estimator of θ based on the sample of size n. Obviously the estimator depends on the sample size n. In order to reflect the dependency of $\widehat{\theta}$ on n, we denote $\widehat{\theta}$ as $\widehat{\theta}_n$.

Definition 16.7. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density $f(x;\theta)$. A sequence of estimators $\{\widehat{\theta}_n\}$ of θ is said to be consistent for θ if and only if the sequence $\{\widehat{\theta}_n\}$ converges in probability to θ, that is, for any $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

Note that consistency is actually a concept relating to a sequence of estimators $\{\widehat{\theta}_n\}_{n=n_o}^{\infty}$ but we usually say "consistency of $\widehat{\theta}_n$" for simplicity. Further, consistency is a large sample property of an estimator.

The following theorem states that if the mean squared error goes to zero as n goes to infinity, then $\{\widehat{\theta}_n\}$ converges in probability to θ.

Theorem 16.5. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density $f(x;\theta)$ and $\{\widehat{\theta}_n\}$ be a sequence of estimators of θ based on the sample. If the variance of $\widehat{\theta}_n$ exists for each n and is finite and

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0$$

then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

Proof: By Markov Inequality (see Theorem 13.8) we have

$$P\left(\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2\right) \leq \frac{E\left(\left(\widehat{\theta_n} - \theta\right)^2\right)}{\epsilon^2}$$

for all $\epsilon > 0$. Since the events

$$\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2 \qquad \text{and} \qquad |\widehat{\theta_n} - \theta| \geq \epsilon$$

are same, we see that

$$P\left(\left(\widehat{\theta_n} - \theta\right)^2 \geq \epsilon^2\right) = P\left(|\widehat{\theta_n} - \theta| \geq \epsilon\right) \leq \frac{E\left(\left(\widehat{\theta_n} - \theta\right)^2\right)}{\epsilon^2}$$

for all $n \in \mathbb{N}$. Hence if

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta_n} - \theta\right)^2\right) = 0$$

then

$$\lim_{n\to\infty} P\left(|\widehat{\theta_n} - \theta| \geq \epsilon\right) = 0$$

and the proof of the theorem is complete.

Let

$$B\left(\widehat{\theta}, \theta\right) = E\left(\widehat{\theta}\right) - \theta$$

be the biased. If an estimator is unbiased, then $B\left(\widehat{\theta}, \theta\right) = 0$. Next we show that

$$E\left(\left(\widehat{\theta} - \theta\right)^2\right) = Var\left(\widehat{\theta}\right) + \left[B\left(\widehat{\theta}, \theta\right)\right]^2. \tag{1}$$

To see this consider

$$\begin{aligned}
E\left(\left(\widehat{\theta} - \theta\right)^2\right) &= E\left(\left(\widehat{\theta}^2 - 2\,\widehat{\theta}\,\theta + \theta^2\right)^2\right) \\
&= E\left(\widehat{\theta}^2\right) - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= E\left(\widehat{\theta}^2\right) - E\left(\widehat{\theta}\right)^2 + E\left(\widehat{\theta}\right)^2 - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= Var\left(\widehat{\theta}\right) + E\left(\widehat{\theta}\right)^2 - 2E\left(\widehat{\theta}\right)\,\theta + \theta^2 \\
&= Var\left(\widehat{\theta}\right) + \left[E\left(\widehat{\theta}\right) - \theta\right]^2 \\
&= Var\left(\widehat{\theta}\right) + \left[B\left(\widehat{\theta}, \theta\right)\right]^2.
\end{aligned}$$

In view of (1), we can say that if

$$\lim_{n\to\infty} Var\left(\widehat{\theta}_n\right) = 0 \tag{2}$$

and

$$\lim_{n\to\infty} B\left(\widehat{\theta}_n, \theta\right) = 0 \tag{3}$$

then

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0.$$

In other words, to show a sequence of estimators is consistent we have to verify the limits (2) and (3).

Example 16.22. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population X with mean μ and variance $\sigma^2 > 0$. Is the likelihood estimator

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

of σ^2 a consistent estimator of σ^2?

Answer: Since $\widehat{\sigma^2}$ depends on the sample size n, we denote $\widehat{\sigma^2}$ as $\widehat{\sigma^2}_n$. Hence

$$\widehat{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

The variance of $\widehat{\sigma^2}_n$ is given by

$$\begin{aligned}
Var\left(\widehat{\sigma^2}_n\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{n^2}\, Var\left(\sigma^2\, \frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{n^2}\, Var\left(\frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{n^2}\, Var\left(\chi^2(n-1)\right) \\
&= \frac{2(n-1)\sigma^4}{n^2} \\
&= \left[\frac{1}{n} - \frac{1}{n^2}\right] 2\,\sigma^4.
\end{aligned}$$

Hence

$$\lim_{n\to\infty} Var\left(\widehat{\theta}_n\right) = \lim_{n\to\infty} \left[\frac{1}{n} - \frac{1}{n^2}\right] 2\,\sigma^4 = 0.$$

The biased $B\left(\widehat{\theta}_n, \theta\right)$ is given by

$$\begin{aligned}
B\left(\widehat{\theta}_n, \theta\right) &= E\left(\widehat{\theta}_n\right) - \sigma^2 \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) - \sigma^2 \\
&= \frac{1}{n}\,E\left(\sigma^2\,\frac{(n-1)S^2}{\sigma^2}\right) - \sigma^2 \\
&= \frac{\sigma^2}{n} E\left(\chi^2(n-1)\right) - \sigma^2 \\
&= \frac{(n-1)\sigma^2}{n} - \sigma^2 \\
&= -\frac{\sigma^2}{n}.
\end{aligned}$$

Thus

$$\lim_{n\to\infty} B\left(\widehat{\theta}_n, \theta\right) = -\lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Hence $\frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is a consistent estimator of σ^2.

In the last example we saw that the likelihood estimator of variance is a consistent estimator. In general, if the density function $f(x;\theta)$ of a population satisfies some mild conditions, then the maximum likelihood estimator of θ is consistent. Similarly, if the density function $f(x;\theta)$ of a population satisfies some mild conditions, then the estimator obtained by moment method is also consistent.

Since consistency is a large sample property of an estimator, some statisticians suggest that consistency should not be used alone for judging the goodness of an estimator; rather it should be used along with other criteria.

16.6. Review Exercises

1. Let T_1 and T_2 be estimators of a population parameter θ based upon the same random sample. If $T_i \sim N\left(\theta, \sigma_i^2\right)$ $i = 1, 2$ and if $T = bT_1 + (1-b)T_2$, then for what value of b, T is a minimum variance unbiased estimator of θ ?

2. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}} \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. What is the expected value of the maximum likelihood estimator of θ ? Is this estimator unbiased?

3. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}} \qquad -\infty < x < \infty,$$

where $0 < \theta$ is a parameter. Is the maximum likelihood estimator an efficient estimator of θ?

4. A random sample $X_1, X_2, ..., X_n$ of size n is selected from a normal distribution with variance σ^2. Let S^2 be the unbiased estimator of σ^2, and T be the maximum likelihood estimator of σ^2. If $20T - 19S^2 = 0$, then what is the sample size?

5. Suppose X and Y are independent random variables each with density function

$$f(x) = \begin{cases} 2\,x\,\theta^2 & \text{for } 0 < x < \frac{1}{\theta} \\ \\ 0 & \text{otherwise.} \end{cases}$$

If $k\,(X + 2Y)$ is an unbiased estimator of θ^{-1}, then what is the value of k?

6. An object of length c is measured by two persons using the same instrument. The instrument error has a normal distribution with mean 0 and variance 1. The first person measures the object 25 times, and the average of the measurements in $\bar{X} = 12$. The second person measures the objects 36 times, and the average of the measurements is $\bar{Y} = 12.8$. To estimate c we use the weighted average $a\,\bar{X} + b\,\bar{Y}$ as an estimator. Determine the constants a and b such that $a\,\bar{X} + b\,\bar{Y}$ is the minimum variance unbiased estimator of c and then calculate the minimum variance unbiased estimate of c.

7. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density function

$$f(x) = \begin{cases} 3\,\theta\,x^2\,e^{-\theta\,x^3} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Find a sufficient statistics for θ.

8. Let $X_1, X_2, ..., X_n$ be a random sample from a Weibull distribution with probability density function

$$f(x) = \begin{cases} \frac{\beta}{\theta^\beta}\, x^{\beta-1}\, e^{-\left(\frac{x}{\theta}\right)^\beta} & \text{if } x > 0 \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $\theta > 0$ and $\beta > 0$ are parameters. Find a sufficient statistics for θ with β known, say $\beta = 2$. If β is unknown, can you find a single sufficient statistics for θ?

9. Let X_1, X_2 be a random sample of size 2 from population with probability density

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. If $Y = \sqrt{X_1 X_2}$, then what should be the value of the constant k such that kY is an unbiased estimator of the parameter θ ?

10. Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. If $\overline{X}$ denotes the sample mean, then what should be value of the constant k such that $k\overline{X}$ is an unbiased estimator of θ ?

11. Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise ,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. If X_{med} denotes the sample median, then what should be value of the constant k such that kX_{med} is an unbiased estimator of θ ?

12. What do you understand by an unbiased estimator of a parameter θ? What is the basic principle of the maximum likelihood estimation of a parameter θ? What is the basic principle of the Bayesian estimation of a parameter θ? What is the main difference between Bayesian method and likelihood method.

13. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \frac{\theta}{(1+x)^{\theta+1}} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. What is a sufficient statistic for the parameter θ?

14. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \frac{x}{\theta^2}\, e^{-\frac{x^2}{2\theta^2}} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter. What is a sufficient statistic for the parameter θ?

15. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. What is the maximum likelihood estimator of θ? Find a sufficient statistics of the parameter θ.

16. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $-\infty < \theta < \infty$ is a parameter. Are the estimators $X_{(1)}$ and $\overline{X} - 1$ are unbiased estimators of θ? Which one is more efficient than the other?

17. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{for } 0 \leq x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 1$ is an unknown parameter. What is a sufficient statistic for the parameter θ?

18. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \theta\,\alpha\,x^{\alpha-1} e^{-\theta x^{\alpha}} & \text{for } 0 \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ and $\alpha > 0$ are parameters. What is a sufficient statistic for the parameter θ for a fixed α?

19. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \frac{\theta\,\alpha^{\theta}}{x^{(\theta+1)}} & \text{for } \alpha < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ and $\alpha > 0$ are parameters. What is a sufficient statistic for the parameter θ for a fixed α?

20. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \binom{m}{x}\theta^{x}(1-\theta)^{m-x} & \text{for } x = 0, 1, 2, ..., m \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is parameter. Show that $\frac{\overline{X}}{m}$ is a uniform minimum variance unbiased estimator of θ for a fixed m.

21. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density function

$$f(x;\theta) = \begin{cases} \theta\,x^{\theta-1} & \text{for } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 1$ is parameter. Show that $-\frac{1}{n}\sum_{i=1}^{n}\ln(X_i)$ is a uniform minimum variance unbiased estimator of $\frac{1}{\theta}$.

22. Let $X_1, X_2, ..., X_n$ be a random sample from a uniform population X on the interval $[0, \theta]$, where $\theta > 0$ is a parameter. Is the likelihood estimator $\widehat{\theta} = X_{(n)}$ of θ a consistent estimator of θ?

23. Let $X_1, X_2, ..., X_n$ be a random sample from a population $X \sim POI(\lambda)$, where $\lambda > 0$ is a parameter. Is the estimator $\overline{X}$ of λ a consistent estimator of λ?

24. Let $X_1, X_2, ..., X_n$ be a random sample from a population X having the probability density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1}, & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Is the estimator $\widehat{\theta} = \frac{\overline{X}}{1-\overline{X}}$ of θ, obtained by the moment method, a consistent estimator of θ?

25. Let $X_1, X_2, ..., X_n$ be a random sample from a population X having the probability density function

$$f(x;p) = \begin{cases} \binom{n}{x}\, p^x\, (1-p)^{n-x}, & \text{if } x = 0, 1, 2, ..., n \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p < 1$ is a parameter and m is a fixed positive integer. What is the maximum likelihood estimator for p. Is this maximum likelihood estimator for p is an efficient estimator?

Chapter 17

SOME TECHNIQUES FOR FINDING INTERVAL ESTIMATORS FOR PARAMETERS

In point estimation we find a value for the parameter θ given a sample data. For example, if $X_1, X_2, ..., X_n$ is a random sample of size n from a population with probability density function

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x-\theta)^2} & \text{for } x \geq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

then the likelihood function of θ is

$$L(\theta) = \prod_{i=1}^{n} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x_i-\theta)^2},$$

where $x_1 \geq \theta$, $x_2 \geq \theta, ..., x_n \geq \theta$. This likelihood function simplifies to

$$L(\theta) = \left[\frac{2}{\pi}\right]^{\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2},$$

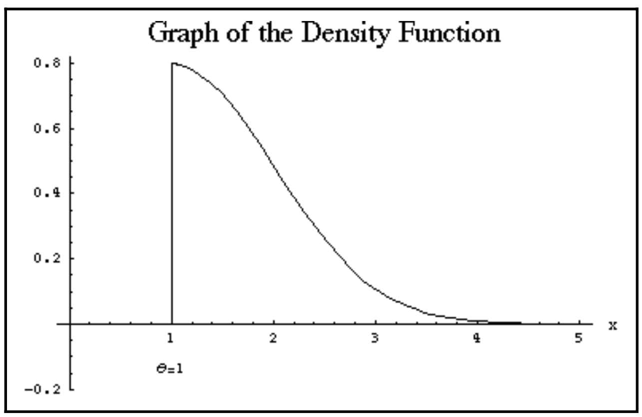
where $\min\{x_1, x_2, ..., x_n\} \geq \theta$. Taking the natural logarithm of $L(\theta)$ and maximizing, we obtain the maximum likelihood estimator of θ as the first order statistic of the sample $X_1, X_2, ..., X_n$, that is

$$\widehat{\theta} = X_{(1)},$$

where $X_{(1)} = \min\{X_1, X_2, ..., X_n\}$. Suppose the true value of $\theta = 1$. Using the maximum likelihood estimator of θ, we are trying to guess this value of θ based on a random sample. Suppose $X_1 = 1.5, X_2 = 1.1, X_3 = 1.7, X_4 = 2.1, X_5 = 3.1$ is a set of sample data from the above population. Then based on this random sample, we will get

$$\widehat{\theta}_{\rm ML} = X_{(1)} = \min\{1.5, 1.1, 1.7, 2.1, 3.1\} = 1.1.$$

If we take another random sample, say $X_1 = 1.8, X_2 = 2.1, X_3 = 2.5, X_4 = 3.1, X_5 = 2.6$ then the maximum likelihood estimator of this θ will be $\widehat{\theta} = 1.8$ based on this sample. The graph of the density function $f(x; \theta)$ for $\theta = 1$ is shown below.



From the graph, it is clear that a number close to 1 has higher chance of getting randomly picked by the sampling process, then the numbers that are substantially bigger than 1. Hence, it makes sense that θ should be estimated by the smallest sample value. However, from this example we see that the point estimate of θ is not equal to the true value of θ. Even if we take many random samples, yet the estimate of θ will rarely equal the actual value of the parameter. Hence, instead of finding a single value for θ, we should report a range of probable values for the parameter θ with certain degree of confidence. This brings us to the notion of confidence interval of a parameter.

17.1. Interval Estimators and Confidence Intervals for Parameters

The interval estimation problem can be stated as follow: Given a random sample $X_1, X_2, ..., X_n$ and a probability value $1 - \alpha$, find a pair of statistics $L = L(X_1, X_2, ..., X_n)$ and $U = U(X_1, X_2, ..., X_n)$ with $L \leq U$ such that the probability of θ being on the random interval $[L, U]$ is $1-\alpha$. That is

$$P\left(L \leq \theta \leq U\right) = 1-\alpha.$$

Recall that a sample is a portion of the population usually chosen by method of random sampling and as such it is a set of random variables $X_1, X_2, ..., X_n$ with the same probability density function $f(x;\theta)$ as the population. Once the sampling is done, we get

$$X_1 = x_1,\ X_2 = x_2,\ \cdots,\ X_n = x_n$$

where $x_1, x_2, ..., x_n$ are the *sample data*.

Definition 17.1. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with density $f(x;\theta)$, where θ is an unknown parameter. The *interval estimator* of θ is a pair of statistics $L = L(X_1, X_2, ..., X_n)$ and $U = U(X_1, X_2, ..., X_n)$ with $L \leq U$ such that if $x_1, x_2, ..., x_n$ is a set of sample data, then θ belongs to the interval $[L(x_1, x_2, ...x_n),\, U(x_1, x_2, ...x_n)]$.

The interval $[l, u]$ will be denoted as an interval estimate of θ whereas the random interval $[L, U]$ will denote the interval estimator of θ. Notice that the interval estimator of θ is the random interval $[L,\, U]$. Next, we define the $100(1-\alpha)\%$ confidence interval for the unknown parameter θ.

Definition 17.2. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with density $f(x;\theta)$, where θ is an unknown parameter. The interval estimator of θ is called a $100(1-\alpha)\%$ *confidence interval* for θ if

$$P(L \leq \theta \leq U) = 1-\alpha.$$

The random variable L is called the *lower confidence limit* and U is called the *upper confidence limit*. The number $(1-\alpha)$ is called the *confidence coefficient* or *degree of confidence*.

There are several methods for constructing confidence intervals for an unknown parameter θ. Some well known methods are: (1) Pivotal Quantity Method, (2) Maximum Likelihood Estimator (MLE) Method, (3) Bayesian Method, (4) Invariant Methods, (5) Inversion of Test Statistic Method, and (6) The Statistical or General Method.

In this chapter, we only focus on the pivotal quantity method and the MLE method. We also briefly examine the the statistical or general method. The pivotal quantity method is mainly due to George Bernard and David Fraser of the University of Waterloo, and this method is perhaps one of the most elegant methods of constructing confidence intervals for unknown parameters.

17.2. Pivotal Quantity Method

In this section, we explain how the notion of pivotal quantity can be used to construct confidence interval for a unknown parameter. We will also examine how to find pivotal quantities for parameters associated with certain probability density functions. We begin with the formal definition of the pivotal quantity.

Definition 17.3. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population X with probability density function $f(x;\theta)$, where θ is an unknown parameter. A *pivotal quantity* Q is a function of $X_1, X_2, ..., X_n$ and θ whose probability distribution is independent of the parameter θ.

Notice that the pivotal quantity $Q(X_1, X_2, ..., X_n, \theta)$ will usually contain both the parameter θ and an estimator (that is, a statistic) of θ. Now we give an example of a pivotal quantity.

Example 17.1. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population X with mean μ and a known variance σ^2. Find a pivotal quantity for the unknown parameter μ.

Answer: Since each $X_i \sim N(\mu, \sigma^2)$,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Standardizing $\overline{X}$, we see that

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

The statistics Q given by

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is a pivotal quantity since it is a function of $X_1, X_2, ..., X_n$ and μ and its probability density function is free of the parameter μ.

There is no general rule for finding a pivotal quantity (or pivot) for a parameter θ of an arbitrarily given density function $f(x;\theta)$. Hence to some extents, finding pivots relies on guesswork. However, if the probability density function $f(x;\theta)$ belongs to the location-scale family, then there is a systematic way to find pivots.

Definition 17.4. Let $g : \mathbb{R} \to \mathbb{R}$ be a probability density function. Then for any μ and any $\sigma > 0$, the family of functions

$$\mathcal{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \mid \mu \in (-\infty,\infty),\, \sigma \in (0,\infty) \right\}$$

is called the *location-scale family* with standard probability density $f(x;\theta)$. The parameter μ is called the *location parameter* and the parameter σ is called the *scale parameter*. If $\sigma = 1$, then $\mathcal{F}$ is called the *location family*. If $\mu = 0$, then $\mathcal{F}$ is called the *scale family*

It should be noted that each member $f(x;\mu,\sigma)$ of the location-scale family is a probability density function. If we take $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, then the normal density function

$$f(x;\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

belongs to the location-scale family. The density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

belongs to the scale family. However, the density function

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

does not belong to the location-scale family.

It is relatively easy to find pivotal quantities for location or scale parameter when the density function of the population belongs to the location-scale family $\mathcal{F}$. When the density function belongs to location family, the pivot for the location parameter μ is $\widehat{\mu} - \mu$, where $\widehat{\mu}$ is the maximum likelihood estimator of μ. If $\widehat{\sigma}$ is the maximum likelihood estimator of σ, then the pivot for the scale parameter σ is $\frac{\widehat{\sigma}}{\sigma}$ when the density function belongs to the scale family. The pivot for location parameter μ is $\frac{\widehat{\mu}-\mu}{\widehat{\sigma}}$ and the pivot for the scale parameter σ is $\frac{\widehat{\sigma}}{\sigma}$ when the density function belongs to location-scale family. Sometime it is appropriate to make a minor modification to the pivot obtained in this way, such as multiplying by a constant, so that the modified pivot will have a known distribution.

Remark 17.1. Pivotal quantity can also be constructed using a sufficient statistic for the parameter. Suppose $T = T(X_1, X_2, ..., X_n)$ is a sufficient statistic based on a random sample $X_1, X_2, ..., X_n$ from a population X with probability density function $f(x; \theta)$. Let the probability density function of T be $g(t; \theta)$. If $g(t; \theta)$ belongs to the location family, then an appropriate constant multiple of $T - a(\theta)$ is a pivotal quantity for the location parameter θ for some suitable expression $a(\theta)$. If $g(t; \theta)$ belongs to the scale family, then an appropriate constant multiple of $\frac{T}{b(\theta)}$ is a pivotal quantity for the scale parameter θ for some suitable expression $b(\theta)$. Similarly, if $g(t; \theta)$ belongs to the location-scale family, then an appropriate constant multiple of $\frac{T-a(\theta)}{b(\theta)}$ is a pivotal quantity for the location parameter θ for some suitable expressions $a(\theta)$ and $b(\theta)$.

Algebraic manipulations of pivots are key factors in finding confidence intervals. If $Q = Q(X_1, X_2, ..., X_n, \theta)$ is a pivot, then a $100(1-\alpha)\%$ confidence interval for θ may be constructed as follows: First, find two values a and b such that

$$P(a \leq Q \leq b) = 1 - \alpha,$$

then convert the inequality $a \leq Q \leq b$ into the form $L \leq \theta \leq U$.

For example, if X is normal population with unknown mean μ and known variance σ^2, then its pdf belongs to the location-scale family. A pivot for μ is $\frac{\overline{X}-\mu}{S}$. However, since the variance σ^2 is known, there is no need to take S. So we consider the pivot $\frac{\overline{X}-\mu}{\sigma}$ to construct the $100(1-2\alpha)\%$ confidence interval for μ. Since our population $X \sim N(\mu, \sigma^2)$, the sample mean $\overline{X}$ is also a normal with the same mean μ and the variance equals to $\frac{\sigma}{\sqrt{n}}$. Hence

$$\begin{aligned}
1 - 2\alpha &= P\left(-z_\alpha \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha\right) \\
&= P\left(\mu - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right).
\end{aligned}$$

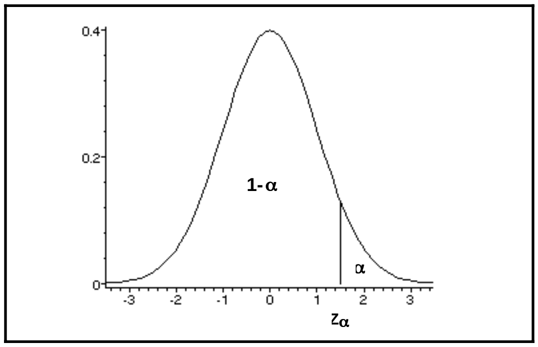
Therefore, the $100(1-2\alpha)\%$ confidence interval for μ is

$$\left[\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}},\ \overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}}\right].$$

Here z_α denotes the $100(1-\alpha)$-percentile (or $(1-\alpha)$-quartile) of a standard normal random variable Z, that is

$$P(Z \leq z_\alpha) = 1 - \alpha,$$

where $\alpha \leq 0.5$ (see figure below). Note that $\alpha = P(Z \leq -z_\alpha)$ if $\alpha \leq 0.5$.



A $100(1-\alpha)\%$ confidence interval for a parameter θ has the following interpretation. If $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ is a sample of size n, then based on this sample we construct a $100(1-\alpha)\%$ confidence interval $[l, u]$ which is a subinterval of the real line $\mathbb{R}$. Suppose we take large number of samples from the underlying population and construct all the corresponding $100(1-\alpha)\%$ confidence intervals, then approximately $100(1-\alpha)\%$ of these intervals would include the unknown value of the parameter θ.

In the next several sections, we illustrate how pivotal quantity method can be used to determine confidence intervals for various parameters.

17.3. Confidence Interval for Population Mean

At the outset, we use the pivotal quantity method to construct a confidence interval for the mean of a normal population. Here we assume first the population variance is known and then variance is unknown. Next, we construct the confidence interval for the mean of a population with continuous, symmetric and unimodal probability distribution by applying the central limit theorem.

Let $X_1, X_2, ..., X_n$ be a random sample from a population $X \sim N(\mu, \sigma^2)$, where μ is an unknown parameter and σ^2 is a known parameter. First of all, we need a pivotal quantity $Q(X_1, X_2, ..., X_n, \mu)$. To construct this pivotal quantity, we find the likelihood estimator of the parameter μ. We know that $\widehat{\mu} = \overline{X}$. Since, each $X_i \sim N(\mu, \sigma^2)$, the distribution of the sample mean is given by

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

It is easy to see that the distribution of the estimator of μ is not independent of the parameter μ. If we standardize $\overline{X}$, then we get

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

The distribution of the standardized $\overline{X}$ is independent of the parameter μ. This standardized $\overline{X}$ is the pivotal quantity since it is a function of the sample $X_1, X_2, ..., X_n$ and the parameter μ, and its probability distribution is independent of the parameter μ. Using this pivotal quantity, we construct the confidence interval as follows:

$$\begin{aligned} 1 - \alpha &= P\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) \\ &= P\left(\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right) \end{aligned}$$

Hence, the $(1 - \alpha)\%$ confidence interval for μ when the population X is normal with the known variance σ^2 is given by

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\, ,\;\; \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

This says that if samples of size n are taken from a normal population with mean μ and known variance σ^2 and if the interval

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\, ,\;\; \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right]$$

is constructed for every sample, then in the long-run $100(1 - \alpha)\%$ of the intervals will cover the unknown parameter μ and hence with a confidence of $(1 - \alpha)100\%$ we can say that μ lies on the interval

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\, ,\;\; \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

The interval estimate of μ is found by taking a good (here maximum likelihood) estimator $\overline{X}$ of μ and adding and subtracting $z_{\frac{\alpha}{2}}$ times the standard deviation of $\overline{X}$.

Remark 17.2. By definition a $100(1-\alpha)\%$ confidence interval for a parameter θ is an interval $[L, U]$ such that the probability of θ being in the interval $[L, U]$ is $1-\alpha$. That is

$$1 - \alpha = P(L \leq \theta \leq U).$$

One can find infinitely many pairs L, U such that

$$1 - \alpha = P(L \leq \theta \leq U).$$

Hence, there are infinitely many confidence intervals for a given parameter. However, we only consider the confidence interval of shortest length. If a confidence interval is constructed by omitting equal tail areas then we obtain what is known as the central confidence interval. In a symmetric distribution, it can be shown that the central confidence interval are the shortest.

Example 17.2. Let $X_1, X_2, ..., X_{11}$ be a random sample of size 11 from a normal distribution with unknown mean μ and variance $\sigma^2 = 9.9$. If $\sum_{i=1}^{11} x_i = 132$, then what is the 95% confidence interval for μ ?

Answer: Since each $X_i \sim N(\mu, 9.9)$, the confidence interval for μ is given by

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}} \,, \quad \overline{X} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}} \right].$$

Since $\sum_{i=1}^{11} x_i = 132$, the sample mean $\overline{x} = \frac{132}{11} = 12$. Also, we see that

$$\sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{9.9}{11}} = \sqrt{0.9}.$$

Further, since $1 - \alpha = 0.95$, $\alpha = 0.05$. Thus

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96 \qquad \text{(from normal table)}.$$

Using these information in the expression of the confidence interval for μ, we get

$$\left[12 - 1.96\sqrt{0.9}, \;\; 12 + 1.96\sqrt{0.9}\right]$$

that is

$$[10.141, \;\; 13.859].$$

Example 17.3. Let $X_1, X_2, ..., X_{11}$ be a random sample of size 11 from a normal distribution with unknown mean μ and variance $\sigma^2 = 9.9$. If $\sum_{i=1}^{11} x_i = 132$, then for what value of the constant k is

$$\left[12 - k\sqrt{0.9},\ 12 + k\sqrt{0.9}\right]$$

a 90% confidence interval for μ ?

Answer: The 90% confidence interval for μ when the variance is given is

$$\left[\overline{x} - \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}},\quad \overline{x} + \left(\frac{\sigma}{\sqrt{n}}\right) z_{\frac{\alpha}{2}}\right].$$

Thus we need to find $\overline{x}$, $\sqrt{\frac{\sigma^2}{n}}$ and $z_{\frac{\alpha}{2}}$ corresponding to $1 - \alpha = 0.9$. Hence

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{11} x_i}{11} \\
&= \frac{132}{11} \\
&= 12. \\
\sqrt{\frac{\sigma^2}{n}} &= \sqrt{\frac{9.9}{11}} \\
&= \sqrt{0.9}. \\
z_{0.05} &= 1.64 \qquad \text{(from normal table).}
\end{aligned}$$

Hence, the confidence interval for μ at 90% confidence level is

$$\left[12 - (1.64)\sqrt{0.9},\ 12 + (1.64)\ \sqrt{0.9}\right].$$

Comparing this interval with the given interval, we get

$$k = 1.64.$$

and the corresponding 90% confidence interval is [10.444, 13.556].

Remark 17.3. Notice that the length of the 90% confidence interval for μ is 3.112. However, the length of the 95% confidence interval is 3.718. Thus higher the confidence level bigger is the length of the confidence interval. Hence, the confidence level is directly proportional to the length of the confidence interval. In view of this fact, we see that if the confidence level is zero, then the length is also zero. That is when the confidence level is zero, the confidence interval of μ degenerates into a point $\overline{X}$.

Until now we have considered the case when the population is normal with unknown mean μ and known variance σ^2. Now we consider the case when the population is non-normal but it probability density function is continuous, symmetric and unimodal. If the sample size is large, then by the central limit theorem

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

Thus, in this case we can take the pivotal quantity to be

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

if the sample size is large (generally $n \geq 32$). Since the pivotal quantity is same as before, we get the sample expression for the $(1 - \alpha)100\%$ confidence interval, that is

$$\left[\overline{X} - \left(\frac{\sigma}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \, , \quad \overline{X} + \left(\frac{\sigma}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \right].$$

Example 17.4. Let $X_1, X_2, ..., X_{40}$ be a random sample of size 40 from a distribution with known variance and unknown mean μ. If $\sum_{i=1}^{40} x_i = 286.56$ and $\sigma^2 = 10$, then what is the 90 percent confidence interval for the population mean μ ?

Answer: Since $1 - \alpha = 0.90$, we get $\frac{\alpha}{2} = 0.05$. Hence, $z_{0.05} = 1.64$ (from the standard normal table). Next, we find the sample mean

$$\overline{x} = \frac{286.56}{40} = 7.164.$$

Hence, the confidence interval for μ is given by

$$\left[7.164 - (1.64) \left(\sqrt{\frac{10}{40}} \right), \; 7.164 + (1.64) \left(\sqrt{\frac{10}{40}} \right) \right]$$

that is

$$[6.344, \;\; 7.984].$$

Example 17.5. In sampling from a nonnormal distribution with a variance of 25, how large must the sample size be so that the length of a 95% confidence interval for the mean is 1.96 ?

Answer: The confidence interval when the sample is taken from a normal population with a variance of 25 is

$$\left[\overline{x} - \left(\frac{\sigma}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \, , \quad \overline{x} + \left(\frac{\sigma}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \right].$$

Thus the length of the confidence interval is

$$\begin{aligned} \ell &= 2 \, z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \\ &= 2 \, z_{0.025} \sqrt{\frac{25}{n}} \\ &= 2 \, (1.96) \sqrt{\frac{25}{n}}. \end{aligned}$$

But we are given that the length of the confidence interval is $\ell = 1.96$. Thus

$$\begin{aligned} 1.96 &= 2 \, (1.96) \sqrt{\frac{25}{n}} \\ \sqrt{n} &= 10 \\ n &= 100. \end{aligned}$$

Hence, the sample size must be 100 so that the length of the 95% confidence interval will be 1.96.

So far, we have discussed the method of construction of confidence interval for the parameter population mean when the variance is known. It is very unlikely that one will know the variance without knowing the population mean, and thus topic of the previous section is not very realistic. Now we treat case of constructing the confidence interval for population mean when the population variance is also unknown. First of all, we begin with the construction of confidence interval assuming the population X is normal.

Suppose $X_1, X_2, ..., X_n$ is random sample from a normal population X with mean μ and variance $\sigma^2 > 0$. Let the sample mean and sample variances be $\overline{X}$ and S^2 respectively. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

and

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

Therefore, the random variable defined by the ratio of $\frac{(n-1)S^2}{\sigma^2}$ to $\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$ has a t-distribution with $(n-1)$ degrees of freedom, that is

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1),$$

where Q is the pivotal quantity to be used for the construction of the confidence interval for μ. Using this pivotal quantity, we construct the confidence interval as follows:

$$\begin{aligned} 1 - \alpha &= P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1) \right) \\ &= P\left(\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \leq \mu \leq \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \right) \end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for μ when the population X is normal with the unknown variance σ^2 is given by

$$\left[\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\,, \;\; \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \right].$$

Example 17.6. A random sample of 9 observations from a normal population yields the observed statistics $\overline{x} = 5$ and $\frac{1}{8}\sum_{i=1}^{9}(x_i - \overline{x})^2 = 36$. What is the 95% confidence interval for μ ?

Answer: Since

$$n = 9 \qquad \overline{x} = 5$$
$$s^2 = 36 \qquad \text{and} \qquad 1 - \alpha = 0.95,$$

the 95% confidence interval for μ is given by

$$\left[\overline{x} - \left(\frac{s}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\,, \;\; \overline{x} + \left(\frac{s}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \right],$$

that is

$$\left[5 - \left(\frac{6}{\sqrt{9}}\right) t_{0.025}(8)\,, \;\; 5 + \left(\frac{6}{\sqrt{9}}\right) t_{0.025}(8) \right],$$

which is

$$\left[5 - \left(\frac{6}{\sqrt{9}} \right) (2.306) \,, \quad 5 + \left(\frac{6}{\sqrt{9}} \right) (2.306) \right].$$

Hence, the 95% confidence interval for μ is given by $[0.388, \ 9.612]$.

Example 17.7. Which of the following is true of a 95% confidence interval for the mean of a population?

(a) The interval includes 95% of the population values on the average.
(b) The interval includes 95% of the sample values on the average.
(c) The interval has 95% chance of including the sample mean.

Answer: None of the statements is correct since the 95% confidence interval for the population mean μ means that the interval has 95% chance of including the population mean μ.

Finally, we consider the case when the population is non-normal but it probability density function is continuous, symmetric and unimodal. If some weak conditions are satisfied, then the sample variance S^2 of a random sample of size $n \geq 2$, converges stochastically to σ^2. Therefore, in

$$\frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$$

the numerator of the left-hand member converges to $N(0,1)$ and the denominator of that member converges to 1. Hence

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

This fact can be used for the construction of a confidence interval for population mean when variance is unknown and the population distribution is nonnormal. We let the pivotal quantity to be

$$Q(X_1, X_2, ..., X_n, \mu) = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$$

and obtain the following confidence interval

$$\left[\overline{X} - \left(\frac{S}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \,, \quad \overline{X} + \left(\frac{S}{\sqrt{n}} \right) z_{\frac{\alpha}{2}} \right].$$

We summarize the results of this section by the following table.

Population	Variance σ^2	Sample Size n	Confidence Limits
normal	known	$n \geq 2$	$\overline{x} \mp z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
normal	not known	$n \geq 2$	$\overline{x} \mp t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}$
not normal	known	$n \geq 32$	$\overline{x} \mp z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
not normal	known	$n < 32$	no formula exists
not normal	not known	$n \geq 32$	$\overline{x} \mp t_{\frac{\alpha}{2}}(n-1) \frac{s}{\sqrt{n}}$
not normal	not known	$n < 32$	no formula exists

17.4. Confidence Interval for Population Variance

In this section, we will first describe the method for constructing the confidence interval for variance when the population is normal with a known population mean μ. Then we treat the case when the population mean is also unknown.

Let $X_1, X_2, ..., X_n$ be a random sample from a normal population X with known mean μ and unknown variance σ^2. We would like to construct a $100(1-\alpha)\%$ confidence interval for the variance σ^2, that is, we would like to find the estimate of L and U such that

$$P\left(L \leq \sigma^2 \leq U\right) = 1 - \alpha.$$

To find these estimate of L and U, we first construct a pivotal quantity. Thus

$$\begin{aligned}
&X_i \sim N\left(\mu, \sigma^2\right), \\
&\left(\frac{X_i - \mu}{\sigma}\right) \sim N(0,1), \\
&\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1). \\
&\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).
\end{aligned}$$

We define the pivotal quantity $Q(X_1, X_2, ..., X_n, \sigma^2)$ as

$$Q(X_1, X_2, ..., X_n, \sigma^2) = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

which has a chi-square distribution with n degrees of freedom. Hence

$$\begin{aligned}
1-\alpha &= P\left(a \leq Q \leq b\right) \\
&= P\left(a \leq \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \leq b\right) \\
&= P\left(\frac{1}{a} \geq \sum_{i=1}^{n}\frac{\sigma^2}{(X_i-\mu)^2} \geq \frac{1}{b}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a} \geq \sigma^2 \geq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{b} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{a}\right) \\
&= P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right)
\end{aligned}$$

Therefore, the $(1-\alpha)\%$ confidence interval for σ^2 when mean is known is given by

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)},\ \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}\right].$$

Example 17.8. A random sample of 9 observations from a normal population with $\mu = 5$ yields the observed statistics $\frac{1}{8}\sum_{i=1}^{9} x_i^2 = 39.125$ and $\sum_{i=1}^{9} x_i = 45$. What is the 95% confidence interval for σ^2 ?

Answer: We have been given that

$$n = 9 \qquad \text{and} \qquad \mu = 5.$$

Further we know that

$$\sum_{i=1}^{9} x_i = 45 \qquad \text{and} \qquad \frac{1}{8}\sum_{i=1}^{9} x_i^2 = 39.125.$$

Hence

$$\sum_{i=1}^{9} x_i^2 = 313,$$

and

$$\begin{aligned}
\sum_{i=1}^{9}(x_i-\mu)^2 &= \sum_{i=1}^{9} x_i^2 - 2\mu\sum_{i=1}^{9} x_i + 9\mu^2 \\
&= 313 - 450 + 225 \\
&= 88.
\end{aligned}$$

Since $1-\alpha = 0.95$, we get $\frac{\alpha}{2} = 0.025$ and $1-\frac{\alpha}{2} = 0.975$. Using chi-square table we have

$$\chi^2_{0.025}(9) = 2.700 \qquad \text{and} \qquad \chi^2_{0.975}(9) = 19.02.$$

Hence, the 95% confidence interval for σ^2 is given by

$$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}, \ \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)} \right],$$

that is

$$\left[\frac{88}{19.02}, \ \frac{88}{2.7} \right]$$

which is

$$[4.583, \ 32.59].$$

Remark 17.4. Since the χ^2 distribution is not symmetric, the above confidence interval is not necessarily the shortest. Later, in the next section, we describe how one construct a confidence interval of shortest length.

Consider a random sample $X_1, X_2, ..., X_n$ from a normal population $X \sim N(\mu, \sigma^2)$, where the population mean μ and population variance σ^2 are unknown. We want to construct a $100(1-\alpha)\%$ confidence interval for the population variance. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\Rightarrow \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n-1).$$

We take $\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2}$ as the pivotal quantity Q to construct the confidence interval for σ^2. Hence, we have

$$\begin{aligned} 1-\alpha &= P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leq Q \leq \frac{1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \right) \\ &= P\left(\frac{1}{\chi^2_{\frac{\alpha}{2}}(n-1)} \leq \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\sigma^2} \leq \frac{1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \right) \\ &= P\left(\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right). \end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for variance σ^2 when the population mean is unknown is given by

$$\left[\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \;\; \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right]$$

Example 17.9. Let $X_1, X_2, ..., X_n$ be a random sample of size 13 from a normal distribution $N(\mu, \sigma^2)$. If $\sum_{i=1}^{13} x_i = 246.61$ and $\sum_{i=1}^{13} 4806.61$. Find the 90% confidence interval for σ^2 ?

Answer:

$$\begin{aligned}
\overline{x} &= 18.97 \\
s^2 &= \frac{1}{n-1} \sum_{i=1}^{13} (x_i - \overline{x})^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{13} \left[x_i^2 - n\overline{x}^2 \right]^2 \\
&= \frac{1}{12} [4806.61 - 4678.2] \\
&= \frac{1}{12}\ 128.41.
\end{aligned}$$

Hence, $12s^2 = 128.41$. Further, since $1 - \alpha = 0.90$, we get $\frac{\alpha}{2} = 0.05$ and $1 - \frac{\alpha}{2} = 0.95$. Therefore, from chi-square table, we get

$$\chi^2_{0.95}(12) = 21.03, \qquad \chi^2_{0.05}(12) = 5.23.$$

Hence, the 95% confidence interval for σ^2 is

$$\left[\frac{128.41}{5.23}, \;\; \frac{128.41}{21.03} \right],$$

that is

$$[6.11, \;\; 24.57].$$

Example 17.10. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution $N\left(\mu, \sigma^2\right)$, where μ and σ^2 are unknown parameters. What is the shortest 90% confidence interval for the standard deviation σ ?

Answer: Let S^2 be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Using this random variable as a pivot, we can construct a $100(1-\alpha)\%$ confidence interval for σ from

$$1-\alpha = P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right)$$

by suitably choosing the constants a and b. Hence, the confidence interval for σ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b}},\ \sqrt{\frac{(n-1)S^2}{a}}\ \right].$$

The length of this confidence interval is given by

$$L(a,b) = \left(\sqrt{n-1}\right)\ S\ \left[\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}}\right].$$

In order to find the shortest confidence interval, we should find a pair of constants a and b such that $L(a,b)$ is minimum. Thus, we have a constraint minimization problem. That is

$$\left.\begin{array}{l} \text{Minimize } \ L(a,b) \\ \text{Subject to the condition} \\ \displaystyle\int_a^b f(u)du = 1-\alpha, \end{array}\right\} \qquad \text{(MP)}$$

where

$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\ 2^{\frac{n-1}{2}}}\ x^{\frac{n-1}{2}-1} e^{-\frac{x}{2}}.$$

Differentiating L with respect to a, we get

$$\frac{dL}{da} = S\sqrt{n-1}\ \left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\ \frac{db}{da}\right).$$

From

$$\int_a^b f(u)\,du = 1-\alpha,$$

we find the derivative of b with respect to a as follows:

$$\frac{d}{da}\int_a^b f(u)\,du = \frac{d}{da}(1-\alpha)$$

that is

$$f(b)\,\frac{db}{da} - f(a) = 0.$$

Thus, we have

$$\frac{db}{da} = \frac{f(a)}{f(b)}.$$

Letting this into the expression for the derivative of L, we get

$$\frac{dL}{da} = S\sqrt{n-1}\,\left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\,\frac{f(a)}{f(b)}\right).$$

Setting this derivative to zero, we get

$$S\sqrt{n-1}\,\left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\,\frac{f(a)}{f(b)}\right) = 0$$

which yields

$$a^{\frac{3}{2}}\,f(a) = b^{\frac{3}{2}}\,f(b).$$

Using the form of f, we get from the above expression

$$a^{\frac{3}{2}}a^{\frac{n-3}{2}}e^{-\frac{a}{2}} = b^{\frac{3}{2}}b^{\frac{n-3}{2}}e^{-\frac{b}{2}}$$

that is

$$a^{\frac{n}{2}}\,e^{-\frac{a}{2}} = b^{\frac{n}{2}}\,e^{-\frac{b}{2}}.$$

From this we get

$$\ln\left(\frac{a}{b}\right) = \left(\frac{a-b}{n}\right).$$

Hence to obtain the pair of constants a and b that will produce the shortest confidence interval for σ, we have to solve the following system of nonlinear equations

$$\left.\begin{aligned} \int_a^b f(u)\,du &= 1-\alpha \\ \ln\left(\frac{a}{b}\right) &= \frac{a-b}{n}. \end{aligned}\right\} \qquad (\star)$$

If a_o and b_o are solutions of $(\star)$, then the shortest confidence interval for σ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b_o}},\;\; \sqrt{\frac{(n-1)S^2}{a_o}}\;\right].$$

Since this system of nonlinear equations is hard to solve analytically, numerical solutions are given in statistical literature in the form of a table for finding the shortest interval for the variance.

17.5. Confidence Interval for Parameter of some Distributions not belonging to the Location-Scale Family

In this section, we illustrate the pivotal quantity method for finding confidence intervals for a parameter θ when the density function does not belong to the location-scale family. The following density functions does not belong to the location-scale family:

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

or

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

We will construct interval estimators for the parameters in these density functions. The same idea for finding the interval estimators can be used to find interval estimators for parameters of density functions that belong to the location-scale family such as

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

To find the pivotal quantities for the above mentioned distributions and others we need the following three results. The first result is Theorem 6.2 while the proof of the second result is easy and we leave it to the reader.

Theorem 17.1. Let $F(x;\theta)$ be the cumulative distribution function of a continuous random variable X. Then

$$F(X;\theta) \sim UNIF(0,1).$$

Theorem 17.2. If $X \sim UNIF(0,1)$, then

$$-\ln X \sim EXP(1).$$

Theorem 17.3. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Then the random variable

$$\frac{2}{\theta} \sum_{i=1}^{n} X_i \sim \chi^2(2n)$$

Proof: Let $Y = \frac{2}{\theta} \sum_{i=1}^{n} X_i$. Now we show that the sampling distribution of Y is chi-square with $2n$ degrees of freedom. We use the moment generating method to show this. The moment generating function of Y is given by

$$\begin{aligned}
M_Y(t) &= M_{\frac{2}{\theta}\sum\limits_{i=1}^{n} X_i}(t) \\
&= \prod_{i=1}^{n} M_{X_i}\left(\frac{2}{\theta} t\right) \\
&= \prod_{i=1}^{n} \left(1 - \theta \, \frac{2}{\theta} t\right)^{-1} \\
&= (1-2t)^{-n} \\
&= (1-2t)^{-\frac{2n}{2}} .
\end{aligned}$$

Since $(1-2t)^{-\frac{2n}{2}}$ corresponds to the moment generating function of a chi-square random variable with $2n$ degrees of freedom, we conclude that

$$\frac{2}{\theta} \sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

Theorem 17.4. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter. Then the random variable $-2\theta \sum_{i=1}^{n} \ln X_i$ has a chi-square distribution with $2n$ degree of freedoms.

Proof: We are given that

$$X_i \sim \theta \, x^{\theta-1}, \qquad 0 < x < 1.$$

Hence, the cdf of f is

$$F(x;\theta) = \int_0^x \theta\, x^{\theta-1} dx = x^{\theta}.$$

Thus by Theorem 17.1, each

$$F\left(X_i;\theta\right) \sim UNIF(0,1),$$

that is

$$X_i^{\theta} \sim UNIF(0,1).$$

By Theorem 17.2, each

$$-\ln X_i^{\theta} \sim EXP(1),$$

that is

$$-\theta\, \ln X_i \sim EXP(1).$$

By Theorem 17.3 (with $\theta = 1$), we obtain

$$-2\,\theta \sum_{i=1}^{n} \ln X_i \sim \chi^2(2n).$$

Hence, the sampling distribution of $-2\,\theta \sum_{i=1}^{n} \ln X_i$ is chi-square with $2n$ degree of freedoms.

The following theorem whose proof follows from Theorems 17.1, 17.2 and 17.3 is the key to finding pivotal quantity of many distributions that do not belong to the location-scale family. Further, this theorem can also be used for finding the pivotal quantities for parameters of some distributions that belong the location-scale family.

Theorem 17.5. Let $X_1, X_2, ..., X_n$ be a random sample from a continuous population X with a distribution function $F(x;\theta)$. If $F(x;\theta)$ is monotone in θ, then the statistics $Q = -2\sum_{i=1}^{n} \ln F\left(X_i;\theta\right)$ is a pivotal quantity and has a chi-square distribution with $2n$ degrees of freedom (that is, $Q \sim \chi^2(2n)$).

It should be noted that the condition $F(x;\theta)$ is monotone in θ is needed to ensure an interval. Otherwise we may get a confidence region instead of a confidence interval. Also note that if $-2\sum_{i=1}^{n} \ln F\left(X_i;\theta\right) \sim \chi^2(2n)$, then

$$-2\sum_{i=1}^{n} \ln\left(1 - F\left(X_i;\theta\right)\right) \sim \chi^2(2n).$$

Example 17.11. If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what is a $100(1-\alpha)\%$ confidence interval for θ?

Answer: To construct a confidence interval for θ, we need a pivotal quantity. That is, we need a random variable which is a function of the sample and the parameter, and whose probability distribution is known but does not involve θ. We use the random variable

$$Q = -2\,\theta \sum_{i=1}^{n} \ln X_i \sim \chi^2(2n)$$

as the pivotal quantity. The $100(1-\alpha)\%$ confidence interval for θ can be constructed from

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq Q \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq -2\,\theta \sum_{i=1}^{n} \ln X_i \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i} \leq \theta \leq \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i}\right). \end{aligned}$$

Hence, $100(1-\alpha)\%$ confidence interval for θ is given by

$$\left[\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i},\ \ \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{-2\sum_{i=1}^{n} \ln X_i}\right].$$

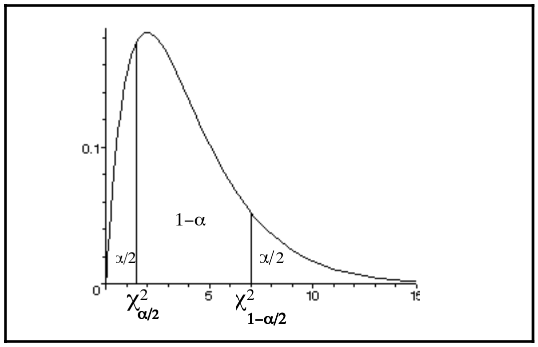
Here $\chi^2_{1-\frac{\alpha}{2}}(2n)$ denotes the $\left(1-\frac{\alpha}{2}\right)$-quantile of a chi-square random variable Y, that is

$$P(Y \leq \chi^2_{1-\frac{\alpha}{2}}(2n)) = 1 - \frac{\alpha}{2}$$

and $\chi^2_{\frac{\alpha}{2}}(2n)$ similarly denotes $\frac{\alpha}{2}$-quantile of Y, that is

$$P\left(Y \leq \chi^2_{\frac{\alpha}{2}}(2n)\right) = \frac{\alpha}{2}$$

for $\alpha \leq 0.5$ (see figure below).



Example 17.12. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a parameter, then what is the $100(1-\alpha)\%$ confidence interval for θ?

Answer: The cumulation density function of $f(x;\theta)$ is

$$F(x;\theta) = \begin{cases} \frac{x}{\theta} & \text{if } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

Since

$$\begin{aligned} -2\sum_{i=1}^{n} \ln F(X_i;\theta) &= -2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\theta}\right) \\ &= 2n\,\ln\theta - 2\sum_{i=1}^{n} \ln X_i \end{aligned}$$

by Theorem 17.5, the quantity $2n\,\ln\theta - 2\sum_{i=1}^{n}\ln X_i \sim \chi^2(2n)$. Since $2n\,\ln\theta - 2\sum_{i=1}^{n}\ln X_i$ is a function of the sample and the parameter and its distribution is independent of θ, it is a pivot for θ. Hence, we take

$$Q(X_1, X_2, ..., X_n, \theta) = 2n\,\ln\theta - 2\sum_{i=1}^{n}\ln X_i.$$

The $100(1-\alpha)\%$ confidence interval for θ can be constructed from

$$
\begin{aligned}
1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq Q \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq 2n\,\ln\theta - 2\sum_{i=1}^{n}\ln X_i \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\
&= P\left(\chi^2_{\frac{\alpha}{2}}(2n) + 2\sum_{i=1}^{n}\ln X_i \leq 2n\,\ln\theta \leq \chi^2_{1-\frac{\alpha}{2}}(2n) + 2\sum_{i=1}^{n}\ln X_i\right) \\
&= P\left(e^{\frac{1}{2n}\left\{\chi^2_{\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}} \leq \theta \leq e^{\frac{1}{2n}\left\{\chi^2_{1-\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}}\right).
\end{aligned}
$$

Hence, $100(1-\alpha)\%$ confidence interval for θ is given by

$$
\left[\, e^{\frac{1}{2n}\left\{\chi^2_{\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}},\;\; e^{\frac{1}{2n}\left\{\chi^2_{1-\frac{\alpha}{2}}(2n)+2\sum_{i=1}^{n}\ln X_i\right\}} \,\right].
$$

The density function of the following example belongs to the scale family. However, one can use Theorem 17.5 to find a pivot for the parameter and determine the interval estimators for the parameter.

Example 17.13. If $X_1, X_2, ..., X_n$ is a random sample from a distribution with density function

$$
f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}
$$

where $\theta > 0$ is a parameter, then what is the $100(1-\alpha)\%$ confidence interval for θ?

Answer: The cumulative density function $F(x;\theta)$ of the density function

$$
f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise} \end{cases}
$$

is given by

$$
F(x;\theta) = 1 - e^{-\frac{x}{\theta}}.
$$

Hence

$$
-2\sum_{i=1}^{n}\ln\left(1 - F\left(X_i;\theta\right)\right) = \frac{2}{\theta}\sum_{i=1}^{n} X_i.
$$

Thus

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2(2n).$$

We take $Q = \frac{2}{\theta}\sum_{i=1}^{n} X_i$ as the pivotal quantity. The $100(1-\alpha)\%$ confidence interval for θ can be constructed using

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq Q \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\chi^2_{\frac{\alpha}{2}}(2n) \leq \frac{2}{\theta}\sum_{i=1}^{n} X_i \leq \chi^2_{1-\frac{\alpha}{2}}(2n)\right) \\ &= P\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n)} \leq \theta \leq \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\frac{\alpha}{2}}(2n)}\right). \end{aligned}$$

Hence, $100(1-\alpha)\%$ confidence interval for θ is given by

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n)},\; \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\frac{\alpha}{2}}(2n)}\right].$$

In this section, we have seen that $100(1-\alpha)\%$ confidence interval for the parameter θ can be constructed by taking the pivotal quantity Q to be either

$$Q = -2\sum_{i=1}^{n} \ln F(X_i;\theta)$$

or

$$Q = -2\sum_{i=1}^{n} \ln\left(1 - F(X_i;\theta)\right).$$

In either case, the distribution of Q is chi-squared with $2n$ degrees of freedom, that is $Q \sim \chi^2(2n)$. Since chi-squared distribution is not symmetric about the y-axis, the confidence intervals constructed in this section do not have the shortest length. In order to have a shortest confidence interval one has to solve the following minimization problem:

$$\left.\begin{aligned} &\text{Minimize } \; L(a,b) \\ &\text{Subject to the condition } \int_a^b f(u)\,du = 1-\alpha, \end{aligned}\right\} \qquad \text{(MP)}$$

where
$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)\ 2^{\frac{n-1}{2}}}\ x^{\frac{n-1}{2}-1} e^{-\frac{x}{2}}.$$

In the case of Example 17.13, the minimization process leads to the following system of nonlinear equations

$$\left.\begin{aligned} \int_a^b f(u)\,du &= 1-\alpha \\ \ln\left(\frac{a}{b}\right) &= \frac{a-b}{2(n+1)}. \end{aligned}\right\} \qquad \text{(NE)}$$

If a_o and b_o are solutions of (NE), then the shortest confidence interval for θ is given by
$$\left[\frac{2\sum_{i=1}^n X_i}{b_o},\ \ \frac{2\sum_{i=1}^n X_i}{a_o}\right].$$

17.6. Approximate Confidence Interval for Parameter with MLE

In this section, we discuss how to construct an approximate $(1-\alpha)100\%$ confidence interval for a population parameter θ using its maximum likelihood estimator $\widehat{\theta}$. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with density $f(x;\theta)$. Let $\widehat{\theta}$ be the maximum likelihood estimator of θ. If the sample size n is large, then using asymptotic property of the maximum likelihood estimator, we have

$$\frac{\widehat{\theta} - E\left(\widehat{\theta}\right)}{\sqrt{V\left(\widehat{\theta}\right)}} \sim N(0,1) \qquad \text{as} \quad n \to \infty,$$

where $V\left(\widehat{\theta}\right)$ denotes the variance of the estimator $\widehat{\theta}$. Since, for large n, the maximum likelihood estimator of θ is unbiased, we get

$$\frac{\widehat{\theta} - \theta}{\sqrt{V\left(\widehat{\theta}\right)}} \sim N(0,1) \qquad \text{as} \quad n \to \infty.$$

The variance $V\left(\widehat{\theta}\right)$ can be computed directly whenever possible or using the Cramér-Rao lower bound

$$V\left(\widehat{\theta}\right) \geq \frac{-1}{E\left[\frac{d^2 \ln L(\theta)}{d\theta^2}\right]}.$$

Now using $Q = \frac{\widehat{\theta}-\theta}{\sqrt{V\left(\widehat{\theta}\right)}}$ as the pivotal quantity, we construct an approximate $(1-\alpha)100\%$ confidence interval for θ as

$$\begin{aligned} 1-\alpha &= P\left(-z_{\frac{\alpha}{2}} \leq Q \leq z_{\frac{\alpha}{2}}\right) \\ &= P\left(-z_{\frac{\alpha}{2}} \leq \frac{\widehat{\theta}-\theta}{\sqrt{V\left(\widehat{\theta}\right)}} \leq z_{\frac{\alpha}{2}}\right). \end{aligned}$$

If $V\left(\widehat{\theta}\right)$ is free of θ, then have

$$1-\alpha = P\left(\widehat{\theta} - z_{\frac{\alpha}{2}}\sqrt{V\left(\widehat{\theta}\right)} \leq \theta \leq \widehat{\theta} + z_{\frac{\alpha}{2}}\sqrt{V\left(\widehat{\theta}\right)}\right).$$

Thus $100(1-\alpha)\%$ approximate confidence interval for θ is

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\sqrt{V\left(\widehat{\theta}\right)},\ \ \widehat{\theta} + z_{\frac{\alpha}{2}}\sqrt{V\left(\widehat{\theta}\right)}\right]$$

provided $V\left(\widehat{\theta}\right)$ is free of θ.

Remark 17.5. In many situations $V\left(\widehat{\theta}\right)$ is not free of the parameter θ. In those situations we still use the above form of the confidence interval by replacing the parameter θ by $\widehat{\theta}$ in the expression of $V\left(\widehat{\theta}\right)$.

Next, we give some examples to illustrate this method.

Example 17.14. Let $X_1, X_2, ..., X_n$ be a random sample from a population X with probability density function

$$f(x;p) = \begin{cases} p^x\,(1-p)^{(1-x)} & \text{if } x = 0, 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is a $100(1-\alpha)\%$ approximate confidence interval for the parameter p?

Answer: The likelihood function of the sample is given by

$$L(p) = \prod_{i=1}^{n} p^{x_i}\,(1-p)^{(1-x_i)}.$$

Taking the logarithm of the likelihood function, we get

$$\ln L(p) = \sum_{i=1}^{n} \left[x_i \ln p + (1 - x_i) \ln(1 - p) \right].$$

Differentiating, the above expression, we get

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - x_i).$$

Setting this equals to zero and solving for p, we get

$$\frac{n\overline{x}}{p} - \frac{n - n\overline{x}}{1-p} = 0,$$

that is

$$(1 - p)\, n\, \overline{x} = p\, (n - n\, \overline{x}),$$

which is

$$n\, \overline{x} - p\, n\, \overline{x} = p\, n - p\, n\, \overline{x}.$$

Hence

$$p = \overline{x}.$$

Therefore, the maximum likelihood estimator of p is given by

$$\widehat{p} = \overline{X}.$$

The variance of $\overline{X}$ is

$$V\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Since $X \sim Ber(p)$, the variance $\sigma^2 = p(1 - p)$, and

$$V\left(\widehat{p}\right) = V\left(\overline{X}\right) = \frac{p(1-p)}{n}.$$

Since $V\left(\widehat{p}\right)$ is not free of the parameter p, we replave p by $\widehat{p}$ in the expression of $V\left(\widehat{p}\right)$ to get

$$V\left(\widehat{p}\right) \simeq \frac{\widehat{p}\,(1 - \widehat{p})}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for the parameter p is given by

$$\left[\widehat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}, \;\; \widehat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} \right]$$

which is

$$\left[\overline{X} - z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}},\;\; \overline{X} + z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{X}\,(1-\overline{X})}{n}}\right].$$

The above confidence interval is a $100(1-\alpha)\%$ approximate confidence interval for proportion.

Example 17.15. A poll was taken of university students before a student election. Of 78 students contacted, 33 said they would vote for Mr. Smith. The population may be taken as 2200. Obtain 95% confidence limits for the proportion of voters in the population in favor of Mr. Smith.

Answer: The sample proportion $\widehat{p}$ is given by

$$\widehat{p} = \frac{33}{78} = 0.4231.$$

Hence

$$\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} = \sqrt{\frac{(0.4231)\,(0.5769)}{78}} = 0.0559.$$

The 2.5^{th} percentile of normal distribution is given by

$$z_{0.025} = 1.96 \qquad\qquad \text{(From table)}.$$

Hence, the lower confidence limit of 95% confidence interval is

$$\begin{aligned}
&\widehat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\\
&\quad = 0.4231 - (1.96)\,(0.0559)\\
&\quad = 0.4231 - 0.1096\\
&\quad = 0.3135.
\end{aligned}$$

Similarly, the upper confidence limit of 95% confidence interval is

$$\begin{aligned}
&\widehat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\\
&\quad = 0.4231 + (1.96)\,(0.0559)\\
&\quad = 0.4231 + 0.1096\\
&\quad = 0.5327.
\end{aligned}$$

Hence, the 95% confidence limits for the proportion of voters in the population in favor of Smith are 0.3135 and 0.5327.

Remark 17.6. In Example 17.15, the 95% percent approximate confidence interval for the parameter p was $[0.3135, 0.5327]$. This confidence interval can be improved to a shorter interval by means of a quadratic inequality. Now we explain how the interval can be improved. First note that in Example 17.14, which we are using for Example 17.15, the approximate value of the variance of the ML estimator $\widehat{p}$ was obtained to be $\sqrt{\frac{p(1-p)}{n}}$. However, this is the exact variance of $\widehat{p}$. Now the pivotal quantity $Q = \frac{\widehat{p}-p}{\sqrt{Var(\widehat{p})}}$ becomes

$$Q = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Using this pivotal quantity, we can construct a 95% confidence interval as

$$0.05 = P\left(- z_{0.025} \leq \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.025} \right)$$

$$= P\left(\left| \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| \leq 1.96 \right).$$

Using $\widehat{p} = 0.4231$ and $n = 78$, we solve the inequality

$$\left| \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| \leq 1.96$$

which is

$$\left| \frac{0.4231 - p}{\sqrt{\frac{p(1-p)}{78}}} \right| \leq 1.96.$$

Squaring both sides of the above inequality and simplifying, we get

$$78\,(0.4231 - p)^2 \leq (1.96)^2\,(p - p^2).$$

The last inequality is equivalent to

$$13.96306158 - 69.84520000\,p + 81.84160000\,p^2 \leq 0.$$

Solving this quadratic inequality, we obtain [0.3196, 0.5338] as a 95% confidence interval for p. This interval is an improvement since its length is 0.2142 where as the length of the interval [0.3135, 0.5327] is 0.2192.

Example 17.16. If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

Answer: The likelihood function $L(\theta)$ of the sample is

$$L(\theta) = \prod_{i=1}^{n} \theta\, x_i^{\theta-1}.$$

Hence

$$\ln L(\theta) = n\, \ln\theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

The first derivative of the logarithm of the likelihood function is

$$\frac{d}{d\theta} \ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i.$$

Setting this derivative to zero and solving for θ, we obtain

$$\theta = -\frac{n}{\sum_{i=1}^{n} \ln x_i}.$$

Hence, the maximum likelihood estimator of θ is given by

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln X_i}.$$

Finding the variance of this estimator is difficult. We compute its variance by computing the Cramér-Rao bound for this estimator. The second derivative of the logarithm of the likelihood function is given by

$$\begin{aligned} \frac{d^2}{d\theta^2} \ln L(\theta) &= \frac{d}{d\theta} \left(\frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i \right) \\ &= -\frac{n}{\theta^2}. \end{aligned}$$

Hence

$$E\left(\frac{d^2}{d\theta^2} \ln L(\theta) \right) = -\frac{n}{\theta^2}.$$

Therefore

$$V\left(\widehat{\theta}\right) \geq \frac{\theta}{n}.$$

Thus we take

$$V\left(\widehat{\theta}\right) \simeq \frac{\theta}{n}.$$

Since $V\left(\widehat{\theta}\right)$ has θ in its expression, we replace the unknown θ by its estimate $\widehat{\theta}$ so that

$$V\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for θ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\,\frac{\widehat{\theta}}{\sqrt{n}},\quad \widehat{\theta} + z_{\frac{\alpha}{2}}\,\frac{\widehat{\theta}}{\sqrt{n}}\right],$$

which is

$$\left[-\frac{n}{\sum_{i=1}^{n}\ln X_i} + z_{\frac{\alpha}{2}}\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right),\quad -\frac{n}{\sum_{i=1}^{n}\ln X_i} - z_{\frac{\alpha}{2}}\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right)\right].$$

Remark 17.7. In the next section 17.2, we derived the exact confidence interval for θ when the population distribution in exponential. The exact $100(1-\alpha)\%$ confidence interval for θ was given by

$$\left[-\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i},\quad -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}\right].$$

Note that this exact confidence interval is not the shortest confidence interval for the parameter θ.

Example 17.17. If $X_1, X_2, ..., X_{49}$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \theta\, x^{\theta-1} & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter, what are 90% *approximate* and *exact* confidence intervals for θ if $\sum_{i=1}^{49}\ln X_i = -0.7567$?

Answer: We are given the followings:

$$n = 49$$

$$\sum_{i=1}^{49}\ln X_i = -0.7576$$

$$1-\alpha = 0.90.$$

Hence, we get

$$z_{0.05} = 1.64,$$

$$\frac{n}{\sum_{i=1}^{n} \ln X_i} = \frac{49}{-0.7567} = -64.75$$

and

$$\frac{\sqrt{n}}{\sum_{i=1}^{n} \ln X_i} = \frac{7}{-0.7567} = -9.25.$$

Hence, the approximate confidence interval is given by

$$[64.75 - (1.64)(9.25), \quad 64.75 + (1.64)(9.25)]$$

that is [49.58, 79.92].

Next, we compute the exact 90% confidence interval for θ using the formula

$$\left[-\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln X_i}, \quad -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln X_i} \right].$$

From chi-square table, we get

$$\chi^2_{0.05}(98) = 77.93 \qquad \text{and} \qquad \chi^2_{0.95}(98) = 124.34.$$

Hence, the exact 90% confidence interval is

$$\left[\frac{77.93}{(2)(0.7567)}, \quad \frac{124.34}{(2)(0.7567)} \right]$$

that is [51.49, 82.16].

Example 17.18. If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} (1-\theta)\,\theta^x & \text{if } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

Answer: The logarithm of the likelihood function of the sample is

$$\ln L(\theta) = \ln\theta \sum_{i=1}^{n} x_i + n\ln(1-\theta).$$

Differentiating we see obtain

$$\frac{d}{d\theta} \ln L(\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n}{1-\theta}.$$

Equating this derivative to zero and solving for θ, we get $\theta = \frac{\overline{x}}{1+\overline{x}}$. Thus, the maximum likelihood estimator of θ is given by

$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

Next, we find the variance of this estimator using the Cramér-Rao lower bound. For this, we need the second derivative of $\ln L(\theta)$. Hence

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{n\overline{x}}{\theta^2} - \frac{n}{(1-\theta)^2}.$$

Therefore

$$\begin{aligned}
&E\left(\frac{d^2}{d\theta^2} \ln L(\theta)\right) \\
&= E\left(-\frac{n\overline{X}}{\theta^2} - \frac{n}{(1-\theta)^2}\right) \\
&= \frac{n}{\theta^2}\, E\left(\overline{X}\right) - \frac{n}{(1-\theta)^2} \\
&= \frac{n}{\theta^2}\, \frac{1}{(1-\theta)} - \frac{n}{(1-\theta)^2} \qquad (\text{since each } \mathrm{X_i} \sim \mathrm{GEO}(1-\theta)) \\
&= -\frac{n}{\theta(1-\theta)} \left[\frac{1}{\theta} + \frac{\theta}{1-\theta}\right] \\
&= -\frac{n\,(1-\theta+\theta^2)}{\theta^2\,(1-\theta)^2}.
\end{aligned}$$

Therefore

$$V\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2 \left(1-\widehat{\theta}\right)^2}{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}.$$

The $100(1-\alpha)\%$ approximate confidence interval for θ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\, \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}},\;\; \widehat{\theta} + z_{\frac{\alpha}{2}}\, \frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}}\right],$$

where
$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

17.7. The Statistical or General Method

Now we briefly describe the statistical or general method for constructing a confidence interval. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density $f(x;\theta)$, where θ is a unknown parameter. We want to determine an interval estimator for θ. Let $T(X_1, X_2, ..., X_n)$ be some statistics having the density function $g(t;\theta)$. Let p_1 and p_2 be two fixed positive number in the open interval $(0,1)$ with $p_1 + p_2 < 1$. Now we define two functions $h_1(\theta)$ and $h_2(\theta)$ as follows:

$$p_1 = \int_{-\infty}^{h_1(\theta)} g(t;\theta)\,dt \qquad \text{and} \qquad p_2 = \int_{-\infty}^{h_2(\theta)} g(t;\theta)\,dt$$

such that

$$P\left(h_1(\theta) < T(X_1, X_2, ..., X_n) < h_2(\theta)\right) = 1 - p_1 - p_2.$$

If $h_1(\theta)$ and $h_2(\theta)$ are monotone functions in θ, then we can find a confidence interval

$$P\left(u_1 < \theta < u_2\right) = 1 - p_1 - p_2$$

where $u_1 = u_1(t)$ and $u_2 = u_2(t)$. The statistics $T(X_1, X_2, ..., X_n)$ may be a sufficient statistics, or a maximum likelihood estimator. If we minimize the length $u_2 - u_1$ of the confidence interval, subject to the condition $1 - p_1 - p_2 = 1 - \alpha$ for $0 < \alpha < 1$, we obtain the shortest confidence interval based on the statistics T.

17.8. Criteria for Evaluating Confidence Intervals

In many situations, one can have more than one confidence intervals for the same parameter θ. Thus it necessary to have a set of criteria to decide whether a particular interval is better than the other intervals. Some well known criteria are: (1) Shortest Length and (2) Unbiasedness. Now we only briefly describe these criteria.

The criterion of shortest length demands that a good $100(1-\alpha)\%$ confidence interval $[L, U]$ of a parameter θ should have the shortest length $\ell = U - L$. In the pivotal quantity method one finds a pivot Q for a parameter θ and then converting the probability statement

$$P(a < Q < b) = 1 - \alpha$$

to

$$P(L < \theta < U) = 1 - \alpha$$

obtains a $100(1-\alpha)\%$ confidence interval for θ. If the constants a and b can be found such that the difference $U - L$ depending on the sample $X_1, X_2, ..., X_n$ is minimum for every realization of the sample, then the random interval $[L, U]$ is said to be the shortest confidence interval based on Q.

If the pivotal quantity Q has certain type of density functions, then one can easily construct confidence interval of shortest length. The following result is important in this regard.

Theorem 17.6. Let the density function of the pivot $Q \sim h(q; \theta)$ be continuous and unimodal. If in some interval $[a, b]$ the density function h has a mode, and satisfies conditions (i) $\int_a^b h(q; \theta) dq = 1 - \alpha$ and (ii) $h(a) = h(b) > 0$, then the interval $[a, b]$ is of the shortest length among all intervals that satisfy condition (i).

If the density function is not unimodal, then minimization of ℓ is necessary to construct a shortest confidence interval. One of the weakness of this shortest length criterion is that in some cases, ℓ could be a random variable. Often, the expected length of the interval $E(\ell) = E(U - L)$ is also used as a criterion for evaluating the goodness of an interval. However, this too has weaknesses. A weakness of this criterion is that minimization of $E(\ell)$ depends on the unknown true value of the parameter θ. If the sample size is very large, then every approximate confidence interval constructed using MLE method has minimum expected length.

A confidence interval is only shortest based on a particular pivot Q. It is possible to find another pivot $Q^\star$ which may yield even a shorter interval than the shortest interval found based on Q. The question naturally arises is how to find the pivot that gives the shortest confidence interval among all other pivots. It has been pointed out that a pivotal quantity Q which is a some function of the complete and sufficient statistics gives shortest confidence interval.

Unbiasedness, is yet another criterion for judging the goodness of an interval estimator. The unbiasedness is defined as follow. A $100(1-\alpha)\%$ confidence interval $[L, U]$ of the parameter θ is said to be unbiased if

$$P(L \leq \theta^\star \leq U) \begin{cases} \geq 1 - \alpha & \text{if } \theta^\star = \theta \\ \\ \leq 1 - \alpha & \text{if } \theta^\star \neq \theta. \end{cases}$$

17.9. Review Exercises

1. Let $X_1, X_2, ..., X_n$ be a random sample from a population with gamma density function

$$f(x;\theta,\beta) = \begin{cases} \frac{1}{\Gamma(\beta)\,\theta^\beta}\, x^{\beta-1}\, e^{-\frac{x}{\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[\frac{2\sum_{i=1}^n X_i}{\chi^2_{1-\frac{\alpha}{2}}(2n\beta)},\quad \frac{2\sum_{i=1}^n X_i}{\chi^2_{\frac{\alpha}{2}}(2n\beta)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for the parameter θ.

2. Let $X_1, X_2, ..., X_n$ be a random sample from a population with Weibull density function

$$f(x;\theta,\beta) = \begin{cases} \frac{\beta}{\theta}\, x^{\beta-1} e^{-x^\beta} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[\frac{2\sum_{i=1}^n X_i^\beta}{\chi^2_{1-\frac{\alpha}{2}}(2n)},\quad \frac{2\sum_{i=1}^n X_i^\beta}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for the parameter θ.

3. Let $X_1, X_2, ..., X_n$ be a random sample from a population with Pareto density function

$$f(x;\theta,\beta) = \begin{cases} \theta\,\beta^\theta\, x^{-(\theta+1)} & \text{for } \beta \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[\frac{2\sum_{i=1}^n \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{1-\frac{\alpha}{2}}(2n)},\quad \frac{2\sum_{i=1}^n \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for $\frac{1}{\theta}$.

4. Let $X_1, X_2, ..., X_n$ be a random sample from a population with Laplace density function

$$f(x;\theta) = \frac{1}{2\theta}\, e^{-\frac{|x|}{\theta}}, \qquad -\infty < x < \infty$$

where θ is an unknown parameter. Show that

$$\left[\frac{2\sum_{i=1}^{n} |X_i|}{\chi^2_{1-\frac{\alpha}{2}}(2n)}, \quad \frac{2\sum_{i=1}^{n} |X_i|}{\chi^2_{\frac{\alpha}{2}}(2n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for θ.

5. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} \frac{1}{2\theta^2}\, x^3\, e^{-\frac{x^2}{2\theta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter. Show that

$$\left[\frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{1-\frac{\alpha}{2}}(4n)}, \quad \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{\frac{\alpha}{2}}(4n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for θ.

6. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta,\beta) = \begin{cases} \beta\,\theta \frac{x^{\beta-1}}{(1+x^{\beta})^{\theta+1}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter and $\beta > 0$ is a known parameter. Show that

$$\left[\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln\left(1 + X_i^{\beta}\right)}, \quad \frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n} \ln\left(1 + X_i^{\beta}\right)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for θ.

7. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = X_{(1)} - \theta$ is a pivotal quantity. Using this pivotal quantity find a a $100(1-\alpha)\%$ confidence interval for θ.

8. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = 2n\left(X_{(1)} - \theta\right)$ is a pivotal quantity. Using this pivotal quantity find a a $100(1-\alpha)\%$ confidence interval for θ.

9. Let $X_1, X_2, ..., X_n$ be a random sample from a population with density function

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{if } \theta < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \mathbb{R}$ is an unknown parameter. Then show that $Q = e^{-\left(X_{(1)} - \theta\right)}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for θ.

10. Let $X_1, X_2, ..., X_n$ be a random sample from a population with uniform density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is an unknown parameter. Then show that $Q = \frac{X_{(n)}}{\theta}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for θ.

11. Let $X_1, X_2, ..., X_n$ be a random sample from a population with uniform density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is an unknown parameter. Then show that $Q = \frac{X_{(n)} - X_{(1)}}{\theta}$ is a pivotal quantity. Using this pivotal quantity find a $100(1-\alpha)\%$ confidence interval for θ.

12. If $X_1, X_2, ..., X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(x-\theta)^2} & \text{if } \theta \leq x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

13. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} (\theta+1)\, x^{-\theta-2} & \text{if} \quad 1 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. What is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

14. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with a probability density function

$$f(x;\theta) = \begin{cases} \theta^2\, x\, e^{-\theta\, x} & \text{if} \quad 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta$ is a parameter. What is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

15. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} \frac{1}{\beta}\, e^{\frac{-(x-4)}{\beta}} & \text{for} \quad x > 4 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$. What is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

16. Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with density function

$$f(x;\theta) = \begin{cases} (\theta-1)\theta^x & \text{for } x = 0, 1, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$. What is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

17. A sample $X_1, X_2, ..., X_n$ of size n is drawn from a gamma distribution

$$f(x;\beta) = \begin{cases} \frac{x^3\, e^{-\frac{x}{\beta}}}{6\beta^4} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

Chapter 18

TEST OF STATISTICAL HYPOTHESES FOR PARAMETERS

18.1. Introduction

Inferential statistics consists of estimation and hypothesis testing. We have already discussed various methods of finding point and interval estimators of parameters. We have also examined the goodness of an estimator.

Suppose $X_1, X_2, ..., X_n$ is a random sample from a population with probability density function given by

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Further, let $n = 4$ and suppose $x_1 = 0.92$, $x_2 = 0.75$, $x_3 = 0.85$, $x_4 = 0.8$ is a set of random sample data from the above distribution. If we apply the maximum likelihood method, then we will find that the estimator $\widehat{\theta}$ of θ is

$$\widehat{\theta} = -1 - \frac{4}{\ln(X_1) + \ln(X_2) + \ln(X_3) + \ln(X_2)}.$$

Hence, the maximum likelihood estimate of θ is

$$\begin{aligned} \widehat{\theta} &= -1 - \frac{4}{\ln(0.92) + \ln(0.75) + \ln(0.85) + \ln(0.80)} \\ &= -1 + \frac{4}{0.7567} = 4.2861 \end{aligned}$$

Therefore, the corresponding probability density function of the population is given by

$$f(x) = \begin{cases} 5.2861\, x^{4.2861} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$$

Since, the point estimate will rarely equal to the true value of θ, we would like to report a range of values with some degree of confidence. If we want to report an interval of values for θ with a confidence level of 90%, then we need a 90% confidence interval for θ. If we use the pivotal quantity method, then we will find that the confidence interval for θ is

$$\left[-1 - \frac{\chi^2_{\frac{\alpha}{2}}(8)}{2\sum_{i=1}^{4} \ln X_i}, \quad -1 - \frac{\chi^2_{1-\frac{\alpha}{2}}(8)}{2\sum_{i=1}^{4} \ln X_i} \right].$$

Since $\chi^2_{0.05}(8) = 2.73$, $\chi^2_{0.95}(8) = 15.51$, and $\sum_{i=1}^{4} \ln(x_i) = -0.7567$, we obtain

$$\left[-1 + \frac{2.73}{2(0.7567)}, \quad -1 + \frac{15.51}{2(0.7567)} \right]$$

which is

$$[\, 0.803, \;\; 9.249\,]\,.$$

Thus we may draw inference, at a 90% confidence level, that the population X has the distribution

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases} \tag{$\star$}$$

where $\theta \in [0.803, 9.249]$. If we think carefully, we will notice that we have made one assumption. The assumption is that the observable quantity X can be modeled by a density function as shown in ($\star$). Since, we are concerned with the parametric statistics, our assumption is in fact about θ.

Based on the sample data, we found that an interval estimate of θ at a 90% confidence level is $[0.803, 9.249]$. But, we assumed that $\theta \in [0.803, 9.249]$. However, we can not be sure that our assumption regarding the parameter is real and is not due to the chance in the random sampling process. The validation of this assumption can be done by the hypothesis test. In this chapter, we discuss testing of statistical hypotheses. Most of the ideas regarding the hypothesis test came from Jerry Neyman and Karl Pearson during 1928-1938.

Definition 18.1. A *statistical hypothesis* H is a conjecture about the distribution $f(x;\theta)$ of a population X. This conjecture is usually about the parameter θ if one is dealing with a parametric statistics; otherwise it is about the form of the distribution of X.

Definition 18.2. A hypothesis H is said to be a *simple hypothesis* if H completely specifies the density $f(x;\theta)$ of the population; otherwise it is called a *composite hypothesis*.

Definition 18.3. The hypothesis to be tested is called the null hypothesis. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by H_o and H_a, respectively.

If θ denotes a population parameter, then the general format of the null hypothesis and alternative hypothesis is

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \in \Omega_a \tag{$\star$}$$

where Ω_o and Ω_a are subsets of the parameter space Ω with

$$\Omega_o \cap \Omega_a = \emptyset \qquad \text{and} \qquad \Omega_o \cup \Omega_a \subseteq \Omega.$$

Remark 18.1. If $\Omega_o \cup \Omega_a = \Omega$, then $(\star)$ becomes

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o.$$

If Ω_o is a singleton set, then H_o reduces to a simple hypothesis. For example, $\Omega_o = \{4.2861\}$, the null hypothesis becomes $H_o : \theta = 4.2861$ and the alternative hypothesis becomes $H_a : \theta \neq 4.2861$. Hence, the null hypothesis $H_o : \theta = 4.2861$ is a simple hypothesis and the alternative $H_a : \theta \neq 4.2861$ is a composite hypothesis.

Definition 18.4. A *hypothesis test* is an ordered sequence

$$(X_1, X_2, ..., X_n; H_o, H_a; C)$$

where $X_1, X_2, ..., X_n$ is a random sample from a population X with the probability density function $f(x;\theta)$, H_o and H_a are hypotheses concerning the parameter θ in $f(x;\theta)$, and C is a Borel set in $\mathbb{R}^n$.

Remark 18.2. Borel sets are defined using the notion of σ-algebra. A collection of subsets $\mathcal{A}$ of a set S is called a σ-algebra if (i) $S \in \mathcal{A}$, (ii) $A^c \in \mathcal{A}$, whenever $A \in \mathcal{A}$, and (iii) $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$, whenever $A_1, A_2, ..., A_n, ... \in \mathcal{A}$. The Borel sets are the member of the smallest σ-algebra containing all open sets of $\mathbb{R}^n$. Two examples of Borel sets in $\mathbb{R}^n$ are the sets that arise by countable union of closed intervals in $\mathbb{R}^n$, and countable intersection of open sets in $\mathbb{R}^n$.

The set C is called the critical region in the hypothesis test. The critical region is obtained using a *test statistics* $W(X_1, X_2, ..., X_n)$. If the outcome of $(X_1, X_2, ..., X_n)$ turns out to be an element of C, then we decide to accept H_a; otherwise we accept H_o.

Broadly speaking, a hypothesis test is a rule that tells us for which sample values we should decide to accept H_o as true and for which sample values we should decide to reject H_o and accept H_a as true. Typically, a hypothesis test is specified in terms of a test statistics W. For example, a test might specify that H_o is to be rejected if the sample total $\sum_{k=1}^{n} X_k$ is less than 8. In this case the critical region C is the set $\{(x_1, x_2, ..., x_n) \,|\, x_1 + x_2 + \cdots + x_n < 8\}$.

18.2. A Method of Finding Tests

There are several methods to find test procedures and they are: (1) Likelihood Ratio Tests, (2) Invariant Tests, (3) Bayesian Tests, and (4) Union-Intersection and Intersection-Union Tests. In this section, we only examine likelihood ratio tests.

Definition 18.5. The *likelihood ratio test statistic* for testing the simple null hypothesis $H_o : \theta \in \Omega_o$ against the composite alternative hypothesis $H_a : \theta \notin \Omega_o$ based on a set of random sample data $x_1, x_2, ..., x_n$ is defined as

$$W(x_1, x_2, ..., x_n) = \frac{\max\limits_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)}{\max\limits_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)},$$

where Ω denotes the parameter space, and $L(\theta, x_1, x_2, ..., x_n)$ denotes the likelihood function of the random sample, that is

$$L(\theta, x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

A *likelihood ratio test* (LRT) is any test that has a critical region C (that is, rejection region) of the form

$$C = \{(x_1, x_2, ..., x_n) \mid W(x_1, x_2, ..., x_n) \le k\},$$

where k is a number in the unit interval $[0, 1]$.

If $H_o : \theta = \theta_0$ and $H_a : \theta = \theta_a$ are both simple hypotheses, then the likelihood ratio test statistic is defined as

$$W(x_1, x_2, ..., x_n) = \frac{L(\theta_o, x_1, x_2, ..., x_n)}{L(\theta_a, x_1, x_2, ..., x_n)}.$$

Now we give some examples to illustrate this definition.

Example 18.1. Let X_1, X_2, X_3 denote three independent observations from a distribution with density

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the LRT critical region for testing $H_o : \theta = 1$ versus $H_a : \theta = 2$?

Answer: In this example, $\theta_o = 1$ and $\theta_a = 2$. By the above definition, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{L(\theta_o, x_1, x_2, x_3)}{L(\theta_a, x_1, x_2, x_3)} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{(1+\theta_o)^3 \prod_{i=1}^{3} x_i^{\theta_o}}{(1+\theta_a)^3 \prod_{i=1}^{3} x_i^{\theta_a}} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{8x_1x_2x_3}{27x_1^2x_2^2x_3^2} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\middle|\; \frac{1}{x_1x_2x_3} \leq \frac{27}{8}k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \;\mid\; x_1x_2x_3 \geq a, \right\} \end{aligned}$$

where a is some constant. Hence the likelihood ratio test is of the form: "Reject H_o if $\displaystyle\prod_{i=1}^{3} X_i \geq a$."

Example 18.2. Let $X_1, X_2, ..., X_{12}$ be a random sample from a normal population with mean zero and variance σ^2. What is the form of the LRT critical region for testing the null hypothesis $H_o : \sigma^2 = 10$ versus $H_a : \sigma^2 = 5$?

Answer: Here $\sigma_o^2 = 10$ and $\sigma_a^2 = 5$. By the above definition, the form of the critical region is given by (with ${\sigma_o}^2 = 10$ and ${\sigma_a}^2 = 5$)

$$\begin{aligned} C &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \frac{L\left({\sigma_o}^2, x_1, x_2, ..., x_{12}\right)}{L\left({\sigma_a}^2, x_1, x_2, ..., x_{12}\right)} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \prod_{i=1}^{12} \frac{\frac{1}{\sqrt{2\pi\sigma_o^2}}\, e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_o}\right)^2}}{\frac{1}{\sqrt{2\pi\sigma_a^2}}\, e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_a}\right)^2}} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \left(\frac{1}{2}\right)^6 e^{\frac{1}{20}\sum_{i=1}^{12} x_i^2} \leq k \right\} \\ &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \sum_{i=1}^{12} x_i^2 \leq a \right\}, \end{aligned}$$

where a is some constant. Hence the likelihood ratio test is of the form: "Reject H_o if $\displaystyle\sum_{i=1}^{12} X_i^2 \leq a$."

Example 18.3. Suppose that X is a random variable about which the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \sim N(0,1)$ is to be tested. What is the form of the LRT critical region based on one observation of X?

Answer: In this example, $L_o(x) = 1$ and $L_a(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. By the above definition, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \;\middle|\; \frac{L_o(x)}{L_a(x)} \leq k \right\}, \qquad \text{where } k \in [0, \infty) \\ &= \left\{ x \in \mathbb{R} \;\middle|\; \sqrt{2\pi}\, e^{\frac{1}{2}x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \;\middle|\; x^2 \leq 2 \ln\left(\frac{k}{\sqrt{2\pi}}\right) \right\} \\ &= \{ x \in \mathbb{R} \mid x \leq a, \} \end{aligned}$$

where a is some constant. Hence the likelihood ratio test is of the form: "Reject H_o if $X \leq a$."

In the above three examples, we have dealt with the case when null as well as alternative were simple. If the null hypothesis is simple (for example, $H_o : \theta = \theta_o$) and the alternative is a composite hypothesis (for example, $H_a : \theta \neq \theta_o$), then the following algorithm can be used to construct the likelihood ratio critical region:

(1) Find the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ for the given sample.

(2) Find $L(\theta_o, x_1, x_2, ..., x_n)$.
(3) Find $\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)$.
(4) Rewrite $\frac{L(\theta_o, x_1, x_2, ..., x_n)}{\max_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)}$ in a "suitable form".
(5) Use step (4) to construct the critical region.

Now we give an example to illustrate these steps.

Example 18.4. Let X be a single observation from a population with probability density

$$f(x;\theta) = \begin{cases} \frac{\theta^x \, e^{-\theta}}{x!} & \text{for } x = 0, 1, 2, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. Find the likelihood ratio critical region for testing the null hypothesis $H_o : \theta = 2$ against the composite alternative $H_a : \theta \neq 2$.

Answer: The likelihood function based on one observation x is

$$L(\theta, x) = \frac{\theta^x \, e^{-\theta}}{x!}.$$

Next, we find $L(\theta_o, x)$ which is given by

$$L(2, x) = \frac{2^x \, e^{-2}}{x!}.$$

Our next step is to evaluate $\max_{\theta \geq 0} L(\theta, x)$. For this we differentiate $L(\theta, x)$ with respect to θ, and then set the derivative to 0 and solve for θ. Hence

$$\frac{dL(\theta, x)}{d\theta} = \frac{1}{x!} \left[e^{-\theta} \, x \, \theta^{x-1} - \theta^x \, e^{-\theta} \right]$$

and $\frac{dL(\theta,x)}{d\theta} = 0$ gives $\theta = x$. Hence

$$\max_{\theta \geq 0} L(\theta, x) = \frac{x^x \, e^{-x}}{x!}.$$

To do the step (4), we consider

$$\frac{L(2, x)}{\max_{\theta \in \Omega} L(\theta, x)} = \frac{\frac{2^x \, e^{-2}}{x!}}{\frac{x^x \, e^{-x}}{x!}}$$

which simplifies to

$$\frac{L(2,x)}{\max\limits_{\theta\in\Omega} L(\theta,x)} = \left(\frac{2e}{x}\right)^x e^{-2}.$$

Thus, the likelihood ratio critical region is given by

$$C = \left\{ x \in \mathbb{R} \;\middle|\; \left(\frac{2e}{x}\right)^x e^{-2} \leq k \right\} = \left\{ x \in \mathbb{R} \;\middle|\; \left(\frac{2e}{x}\right)^x \leq a \right\}$$

where a is some constant. The likelihood ratio test is of the form: "Reject H_o if $\left(\frac{2e}{X}\right)^X \leq a$."

So far, we have learned how to find tests for testing the null hypothesis against the alternative hypothesis. However, we have not considered the goodness of these tests. In the next section, we consider various criteria for evaluating the goodness of an hypothesis test.

18.3. Methods of Evaluating Tests

There are several criteria to evaluate the goodness of a test procedure. Some well known criteria are: (1) Powerfulness, (2) Unbiasedness and Invariancy, and (3) Local Powerfulness. In order to examine some of these criteria, we need some terminologies such as error probabilities, power functions, type I error, and type II error. First, we develop these terminologies.

A statistical hypothesis is a conjecture about the distribution $f(x;\theta)$ of the population X. This conjecture is usually about the parameter θ if one is dealing with a parametric statistics; otherwise it is about the form of the distribution of X. If the hypothesis completely specifies the density $f(x;\theta)$ of the population, then it is said to be a simple hypothesis; otherwise it is called a composite hypothesis. The hypothesis to be tested is called the null hypothesis. We often hope to reject the null hypothesis based on the sample information. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by H_o and H_a, respectively.

In hypothesis test, the basic problem is to decide, based on the sample information, whether the null hypothesis is true. There are four possible situations that determines our decision is correct or in error. These four situations are summarized below:

	H_o is true	H_o is false
Accept H_o	Correct Decision	Type II Error
Reject H_o	Type I Error	Correct Decision

Definition 18.6. Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population X with density $f(x; \theta)$, where θ is a parameter. The *significance level* of the hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o,$$

denoted by α, is defined as

$$\alpha = P\left(\text{Type I Error}\right).$$

Thus, the significance level of a hypothesis test we mean the probability of rejecting a true null hypothesis, that is

$$\alpha = P\left(\text{Reject } \mathrm{H_o} \ / \ \mathrm{H_o} \text{ is true}\right).$$

This is also equivalent to

$$\alpha = P\left(\text{Accept } \mathrm{H_a} \ / \ \mathrm{H_o} \text{ is true}\right).$$

Definition 18.7. Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population X with density $f(x; \theta)$, where θ is a parameter. The *probability of type II error* of the hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{and} \qquad H_a : \theta \notin \Omega_o,$$

denoted by β, is defined as

$$\beta = P\left(\text{Accept } \mathrm{H_o} \ / \ \mathrm{H_o} \text{ is false}\right).$$

Similarly, this is also equivalent to

$$\beta = P\left(\text{Accept } \mathrm{H_o} \ / \ \mathrm{H_a} \text{ is true}\right).$$

Remark 18.3. Note that α can be numerically evaluated if the null hypothesis is a simple hypothesis and rejection rule is given. Similarly, β can be evaluated if the alternative hypothesis is simple and rejection rule is known. If null and the alternatives are composite hypotheses, then α and β become functions of θ.

Example 18.5. Let $X_1, X_2, ..., X_{20}$ be a random sample from a distribution with probability density function

$$f(x;p) = \begin{cases} p^x (1-p)^{1-x} & \text{if } x = 0,1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p \leq \frac{1}{2}$ is a parameter. The hypothesis $H_o : p = \frac{1}{2}$ to be tested against $H_a : p < \frac{1}{2}$. If H_o is rejected when $\sum_{i=1}^{20} X_i \leq 6$, then what is the probability of type I error?

Answer: Since each observation $X_i \sim BER(p)$, the sum the observations $\sum_{i=1}^{20} X_i \sim BIN(20, p)$. The probability of type I error is given by

$$\begin{aligned} \alpha &= P\,(\text{Type I Error}) \\ &= P\,(\text{Reject H}_\text{o} \text{ / H}_\text{o} \text{ is true}) \\ &= P\left(\sum_{i=1}^{20} X_i \leq 6 \;\Big/\; \text{H}_\text{o} \text{ is true} \right) \\ &= P\left(\sum_{i=1}^{20} X_i \leq 6 \;\Big/\; H_o : p = \frac{1}{2} \right) \\ &= \sum_{k=0}^{6} \binom{20}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{20-k} \\ &= 0.0577 \qquad \text{(from binomial table).} \end{aligned}$$

Hence the probability of type I error is 0.0577.

Example 18.6. Let p represent the proportion of defectives in a manufacturing process. To test $H_o : p \leq \frac{1}{4}$ versus $H_a : p > \frac{1}{4}$, a random sample of size 5 is taken from the process. If the number of defectives is 4 or more, the null hypothesis is rejected. What is the probability of rejecting H_o if $p = \frac{1}{5}$?

Answer: Let X denote the number of defectives out of a random sample of size 5. Then X is a binomial random variable with $n = 5$ and $p = \frac{1}{5}$. Hence, the probability of rejecting H_o is given by

$$\begin{aligned}
\alpha &= P\,(\text{Reject H}_\text{o} \;/\; \text{H}_\text{o} \text{ is true}) \\
&= P\,(X \geq 4 \;/\; \text{H}_\text{o} \text{ is true}) \\
&= P\left(X \geq 4 \;\Big/\; p = \frac{1}{5}\right) \\
&= P\left(X = 4 \;\Big/ p = \frac{1}{5}\right) + P\left(X = 5 \;\Big/ p = \frac{1}{5}\right) \\
&= \binom{5}{4} p^4 (1-p)^1 + \binom{5}{5} p^5 (1-p)^0 \\
&= 5\left(\frac{1}{5}\right)^4 \left(\frac{4}{5}\right) + \left(\frac{1}{5}\right)^5 \\
&= \left(\frac{1}{5}\right)^5 \; [20+1] \\
&= \frac{21}{3125}.
\end{aligned}$$

Hence the probability of rejecting the null hypothesis H_o is $\frac{21}{3125}$.

Example 18.7. A random sample of size 4 is taken from a normal distribution with unknown mean μ and variance $\sigma^2 > 0$. To test $H_o : \mu = 0$ against $H_a : \mu < 0$ the following test is used: "Reject H_o if and only if $X_1 + X_2 + X_3 + X_4 < -20$." Find the value of σ so that the significance level of this test will be closed to 0.14.

Answer: Since

$$\begin{aligned}
0.14 &= \alpha \qquad\qquad\qquad\qquad \text{(significance level)} \\
&= P\,(\text{Type I Error}) \\
&= P\,(\text{Reject H}_\text{o} \;/\; \text{H}_\text{o} \text{ is true}) \\
&= P\,(X_1 + X_2 + X_3 + X_4 < -20 \;/ H_o : \mu = 0) \\
&= P\left(\overline{X} < -5 \;/ H_o : \mu = 0\right) \\
&= P\left(\frac{\overline{X} - 0}{\frac{\sigma}{2}} < \frac{-5 - 0}{\frac{\sigma}{2}}\right) \\
&= P\left(Z < -\frac{10}{\sigma}\right),
\end{aligned}$$

we get from the standard normal table

$$1.08 = \frac{10}{\sigma}.$$

Therefore

$$\sigma = \frac{10}{1.08} = 9.26.$$

Hence, the standard deviation has to be 9.26 so that the significance level will be closed to 0.14.

Example 18.8. A normal population has a standard deviation of 16. The critical region for testing $H_o : \mu = 5$ versus the alternative $H_a : \mu = k$ is $\bar{X} > k - 2$. What would be the value of the constant k and the sample size n which would allow the probability of Type I error to be 0.0228 and the probability of Type II error to be 0.1587.

Answer: It is given that the population $X \sim N\left(\mu,\, 16^2\right)$. Since

$$\begin{aligned}
0.0228 &= \alpha \\
&= P\,(\text{Type I Error}) \\
&= P\,(\text{Reject H}_\text{o} \;/\; \text{H}_\text{o} \text{ is true}) \\
&= P\left(\overline{X} > k - 2 \;/ H_o : \mu = 5\right) \\
&= P\left(\frac{\overline{X} - 5}{\sqrt{\frac{256}{n}}} > \frac{k - 7}{\sqrt{\frac{256}{n}}}\right) \\
&= P\left(Z > \frac{k - 7}{\sqrt{\frac{256}{n}}}\right) \\
&= 1 - P\left(Z \leq \frac{k - 7}{\sqrt{\frac{256}{n}}}\right)
\end{aligned}$$

Hence, from standard normal table, we have

$$\frac{(k-7)\sqrt{n}}{16} = 2$$

which gives

$$(k-7)\sqrt{n} = 32.$$

Similarly

$$\begin{aligned}
0.1587 &= P\,(\text{Type II Error}) \\
&= P\,(\text{Accept H}_\text{o} \;/\; \text{H}_\text{a} \text{ is true}) \\
&= P\left(\overline{X} \leq k-2 \;/ H_a : \mu = k\right) \\
&= P\left(\frac{\overline{X}-\mu}{\sqrt{\frac{256}{n}}} \leq \frac{k-2-\mu}{\sqrt{\frac{256}{n}}} \;\Big/\; H_a : \mu = k \right) \\
&= P\left(\frac{\overline{X}-k}{\sqrt{\frac{256}{n}}} \leq \frac{k-2-k}{\sqrt{\frac{256}{n}}} \right) \\
&= P\left(Z \leq -\frac{2}{\sqrt{\frac{256}{n}}} \right) \\
&= 1 - P\left(Z \leq \frac{2\sqrt{n}}{16} \right).
\end{aligned}$$

Hence $0.1587 = 1 - P\left(Z \leq \frac{2\sqrt{n}}{16}\right)$ or $P\left(Z \leq \frac{2\sqrt{n}}{16}\right) = 0.8413$. Thus, from the standard normal table, we have

$$\frac{2\sqrt{n}}{16} = 1$$

which yields

$$n = 64.$$

Letting this value of n in

$$(k-7)\sqrt{n} = 32,$$

we see that $k = 11$.

While deciding to accept H_o or H_a, we may make a wrong decision. The probability γ of a wrong decision can be computed as follows:

$$\begin{aligned}
\gamma &= P\,(\text{H}_\text{a} \text{ accepted and } \text{H}_\text{o} \text{ is true}) + P\,(\text{H}_\text{o} \text{ accepted and } \text{H}_\text{a} \text{ is true}) \\
&= P\,(\text{H}_\text{a} \text{ accepted / H}_\text{o} \text{ is true})\, P\,(\text{H}_\text{o} \text{ is true}) \\
&\qquad + P\,(\text{H}_\text{o} \text{ accepted / H}_\text{a} \text{ is true})\, P\,(\text{H}_\text{a} \text{ is true}) \\
&= \alpha\, P\,(\text{H}_\text{o} \text{ is true}) + \beta\, P\,(\text{H}_\text{a} \text{ is true}).
\end{aligned}$$

In most cases, the probabilities $P\,(\text{H}_\text{o} \text{ is true})$ and $P\,(\text{H}_\text{a} \text{ is true})$ are not known. Therefore, it is, in general, not possible to determine the exact numerical value of the probability γ of making a wrong decision. However, since γ is a weighted sum of α and β, and $P\,(\mathrm{H_o} \text{ is true}) + P\,(\mathrm{H_a} \text{ is true}) = 1$, we have

$$\gamma \leq \max\{\alpha, \beta\}.$$

A good decision rule (or a good test) is the one which yields the smallest γ. In view of the above inequality, one will have a small γ if the probability of type I error as well as probability of type II error are small.

The alternative hypothesis is mostly a composite hypothesis. Thus, it is not possible to find a value for the probability of type II error, β. For composite alternative, β is a function of θ. That is, $\beta : \Omega_o^c :\to [0, 1]$. Here Ω_o^c denotes the complement of the set Ω_o in the parameter space Ω. In hypothesis test, instead of β, one usually considers the *power of the test* $1 - \beta(\theta)$, and a small probability of type II error is equivalent to large power of the test.

Definition 18.8. Let $H_o : \theta \in \Omega_o$ and $H_a : \theta \notin \Omega_o$ be the null and alternative hypothesis to be tested based on a random sample $X_1, X_2, ..., X_n$ from a population X with density $f(x; \theta)$, where θ is a parameter. The *power function* of a hypothesis test

$$H_o : \theta \in \Omega_o \qquad \text{versus} \qquad H_a : \theta \notin \Omega_o$$

is a function $\pi : \Omega \to [0, 1]$ defined by

$$\pi(\theta) = \begin{cases} P\,(\text{Type I Error}) & \text{if } \mathrm{H_o} \text{ is true} \\ \\ 1 - P\,(\text{Type II Error}) & \text{if } \mathrm{H_a} \text{ is true.} \end{cases}$$

Example 18.9. A manufacturing firm needs to test the null hypothesis H_o that the probability p of a defective item is 0.1 or less, against the alternative hypothesis $H_a : p > 0.1$. The procedure is to select two items at random. If both are defective, H_o is rejected; otherwise, a third is selected. If the third item is defective H_o is rejected. If all other cases, H_o is accepted, what is the power of the test in terms of p (if H_o is true)?

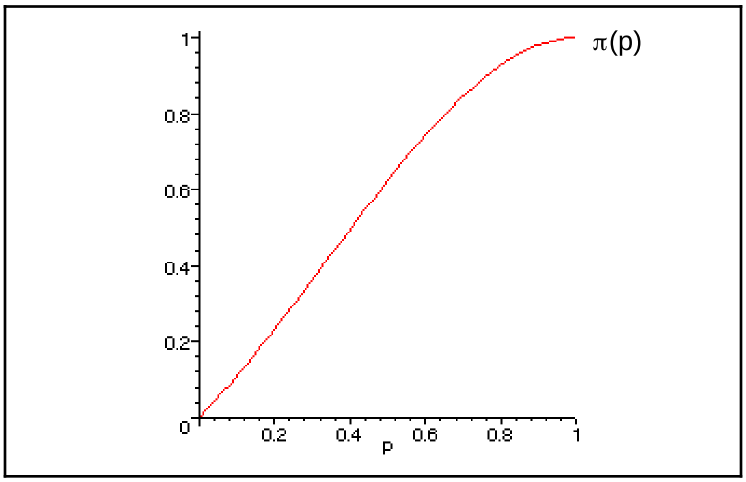
Answer: Let p be the probability of a defective item. We want to calculate the power of the test at the null hypothesis. The power function of the test is given by

$$\pi(p) = \begin{cases} P\,(\text{Type I Error}) & \text{if } p \leq 0.1 \\ \\ 1 - P\,(\text{Type II Error}) & \text{if } p > 0.1. \end{cases}$$

Hence, we have

$$
\begin{aligned}
&\pi(p) \\
&= P\,(\text{Reject H}_o \;/\; \text{H}_o \text{ is true}) \\
&= P\,(\text{Reject H}_o \;/\; \text{H}_o : \text{p} = \text{p}) \\
&= P\,(\text{first two items are both defective /p}) + \\
&\;\; + P\,(\text{at least one of the first two items is not defective and third is/p}) \\
&= p^2 + (1-p)^2\, p + \binom{2}{1}\, p(1-p)p \\
&= p + p^2 - p^3.
\end{aligned}
$$

The graph of this power function is shown below.



Remark 18.4. If X denotes the number of independent trials needed to obtain the first success, then $X \sim GEO(p)$, and

$$P(X = k) = (1-p)^{k-1}\, p,$$

where $k = 1, 2, 3, ..., \infty$. Further

$$P(X \leq n) = 1 - (1-p)^n$$

since

$$
\begin{aligned}
\sum_{k=1}^{n} (1-p)^{k-1}\, p &= p\, \sum_{k=1}^{n} (1-p)^{k-1} \\
&= p\, \frac{1-(1-p)^n}{1-(1-p)} \\
&= 1 - (1-p)^n.
\end{aligned}
$$

Example 18.10. Let X be the number of independent trails required to obtain a success where p is the probability of success on each trial. The hypothesis $H_o : p = 0.1$ is to be tested against the alternative $H_a : p = 0.3$. The hypothesis is rejected if $X \leq 4$. What is the power of the test if H_a is true?

Answer: The power function is given by

$$\pi(p) = \begin{cases} P\,(\text{Type I Error}) & \text{if } p = 0.1 \\ \\ 1 - P\,(\text{Type II Error}) & \text{if p} = 0.3. \end{cases}$$

Hence, we have

$$\begin{aligned} \alpha &= 1 - P\,(\text{Accept H}_\text{o} \;/\; \text{H}_\text{o} \text{ is false}) \\ &= P\,(\text{Reject H}_\text{o} \;/\; \text{H}_\text{a} \text{ is true}) \\ &= P\,(X \leq 4 \;/\; \text{H}_\text{a} \text{ is true}) \\ &= P\,(X \leq 4 \;/\; p = 0.3) \\ &= \sum_{k=1}^{4} P\,(X = k \;/p = 0.3) \\ &= \sum_{k=1}^{4} (1-p)^{k-1}\, p \qquad (\text{where p} = 0.3) \\ &= \sum_{k=1}^{4} (0.7)^{k-1}\,(0.3) \\ &= 0.3\; \sum_{k=1}^{4} (0.7)^{k-1} \\ &= 1 - (0.7)^4 \\ &= 0.7599. \end{aligned}$$

Hence, the power of the test at the alternative is 0.7599.

Example 18.11. Let $X_1, X_2, ..., X_{25}$ be a random sample of size 25 drawn from a normal distribution with unknown mean μ and variance $\sigma^2 = 100$. It is desired to test the null hypothesis $\mu = 4$ against the alternative $\mu = 6$. What is the power at $\mu = 6$ of the test with rejection rule: reject $\mu = 4$ if $\sum_{i=1}^{25} X_i \geq 125$?

Answer: The power of the test at the alternative is

$$\begin{aligned}
\pi(6) &= 1 - P\,(\text{Type II Error}) \\
&= 1 - P\,(\text{Accept H}_\text{o} \ / \ \text{H}_\text{o} \text{ is false}) \\
&= P\,(\text{Reject H}_\text{o} \ / \ \text{H}_\text{a} \text{ is true}) \\
&= P\left(\sum_{i=1}^{25} X_i \geq 125 \ / \ H_a : \mu = 6\right) \\
&= P\left(\overline{X} \geq 5 \ / \ H_a \mu = 6\right) \\
&= P\left(\frac{\overline{X} - 6}{\frac{10}{\sqrt{25}}} \geq \frac{5-6}{\frac{10}{\sqrt{25}}}\right) \\
&= P\left(Z \geq -\frac{1}{2}\right) \\
&= 0.6915.
\end{aligned}$$

Example 18.12. A urn contains 7 balls, θ of which are red. A sample of size 2 is drawn without replacement to test $H_o : \theta \leq 1$ against $H_a : \theta > 1$. If the null hypothesis is rejected if one or more red balls are drawn, find the power of the test when $\theta = 2$.

Answer: The power of the test at $\theta = 2$ is given by

$$\begin{aligned}
\pi(2) &= 1 - P\,(\text{Type II Error}) \\
&= 1 - P\,(\text{Accept H}_\text{o} \ / \ \text{H}_\text{o} \text{ is false}) \\
&= 1 - P\,(\text{zero red balls are drawn } /2 \text{ balls were red}) \\
&= 1 - \frac{\binom{5}{2}}{\binom{7}{2}} \\
&= 1 - \frac{10}{21} \\
&= \frac{11}{21} \\
&= 0.524.
\end{aligned}$$

In all of these examples, we have seen that if the rule for rejection of the null hypothesis H_o is given, then one can compute the significance level or power function of the hypothesis test. The rejection rule is given in terms of a statistic $W\,(X_1, X_2, ..., X_n)$ of the sample $X_1, X_2, ..., X_n$. For instance, in Example 18.5, the rejection rule was: "Reject the null hypothesis H_o if $\sum_{i=1}^{20} X_i \leq 6$." Similarly, in Example 18.7, the rejection rule was: "Reject H_o

if and only if $X_1+X_2+X_3+X_4 < -20$", and so on. The statistic W, used in the statement of the rejection rule, partitioned the set S^n into two subsets, where S denotes the support of the density function of the population X. One subset is called the rejection or critical region and other subset is called the acceptance region. The rejection rule is obtained in such a way that the probability of the type I error is as small as possible and the power of the test at the alternative is as large as possible.

Next, we give two definitions that will lead us to the definition of uniformly most powerful test.

Definition 18.9. Given $0 \leq \delta \leq 1$, a test (or test procedure) T for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$ is said to be a *test of level* δ if

$$\max_{\theta \in \Omega_o} \pi(\theta) \leq \delta,$$

where $\pi(\theta)$ denotes the power function of the test T.

Definition 18.10. Given $0 \leq \delta \leq 1$, a test (or test procedure) for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$ is said to be a *test of size* δ if

$$\max_{\theta \in \Omega_o} \pi(\theta) = \delta.$$

Definition 18.11. Let T be a test procedure for testing the null hypothesis $H_o : \theta \in \Omega_o$ against the alternative $H_a : \theta \in \Omega_a$. The test (or test procedure) T is said to be the *uniformly most powerful* (UMP) test of level δ if T is of level δ and for any other test W of level δ,

$$\pi_T(\theta) \geq \pi_W(\theta)$$

for all $\theta \in \Omega_a$. Here $\pi_T(\theta)$ and $\pi_W(\theta)$ denote the power functions of tests T and W, respectively.

Remark 18.5. If T is a test procedure for testing $H_o : \theta = \theta_o$ against $H_a : \theta = \theta_a$ based on a sample data $x_1, ..., x_n$ from a population X with a continuous probability density function $f(x; \theta)$, then there is a critical region C associated with the the test procedure T, and power function of T can be computed as

$$\pi_T = \int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

Similarly, the size of a critical region C, say α, can be given by

$$\alpha = \int_C L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

The following famous result tells us which tests are uniformly most powerful if the null hypothesis and the alternative hypothesis are both simple.

Theorem 18.1 (Neyman-Pearson). Let $X_1, X_2, ..., X_n$ be a random sample from a population with probability density function $f(x;\theta)$. Let

$$L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

be the likelihood function of the sample. Then any critical region C of the form

$$C = \left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{L(\theta_o, x_1, ..., x_n)}{L(\theta_a, x_1, ..., x_n)} \leq k \right\}$$

for some constant $0 \leq k < \infty$ is best (or uniformly most powerful) of its size for testing $H_o : \theta = \theta_o$ against $H_a : \theta = \theta_a$.

Proof: We assume that the population has a continuous probability density function. If the population has a discrete distribution, the proof can be appropriately modified by replacing integration by summation.

Let C be the critical region of size α as described in the statement of the theorem. Let B be any other critical region of size α. We want to show that the power of C is greater than or equal to that of B. In view of Remark 18.5, we would like to show that

$$\int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \geq \int_B L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n. \tag{1}$$

Since C and B are both critical regions of size α, we have

$$\int_C L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n = \int_B L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n. \tag{2}$$

The last equality (2) can be written as

$$\int_{C\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C\cap B^c} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n$$
$$= \int_{C\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C^c\cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n$$

since

$$C = (C \cap B) \cup (C \cap B^c) \quad \text{and} \quad B = (C \cap B) \cup (C^c \cap B). \tag{3}$$

Therefore from the last equality, we have

$$\int_{C \cap B^c} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n = \int_{C^c \cap B} L(\theta_o, x_1, ..., x_n)\, dx_1 \cdots dx_n. \tag{4}$$

Since

$$C = \left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{L(\theta_o, x_1, ..., x_n)}{L(\theta_a, x_1, ..., x_n)} \leq k \right\} \tag{5}$$

we have

$$L(\theta_a, x_1, ..., x_n) \geq \frac{L(\theta_o, x_1, ..., x_n)}{k} \tag{6}$$

on C, and

$$L(\theta_a, x_1, ..., x_n) < \frac{L(\theta_o, x_1, ..., x_n)}{k} \tag{7}$$

on C^c. Therefore from (4), (6) and (7), we have

$$\begin{aligned}
\int_{C \cap B^c} & L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
& \geq \int_{C \cap B^c} \frac{L(\theta_o, x_1, ..., x_n)}{k}\, dx_1 \cdots dx_n \\
& = \int_{C^c \cap B} \frac{L(\theta_o, x_1, ..., x_n)}{k}\, dx_1 \cdots dx_n \\
& \geq \int_{C^c \cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.
\end{aligned}$$

Thus, we obtain

$$\int_{C \cap B^c} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \geq \int_{C^c \cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n.$$

From (3) and the last inequality, we see that

$$\begin{aligned}
& \int_C L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
& \quad = \int_{C \cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C \cap B^c} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
& \quad \geq \int_{C \cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n + \int_{C^c \cap B} L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n \\
& \quad \geq \int_B L(\theta_a, x_1, ..., x_n)\, dx_1 \cdots dx_n
\end{aligned}$$

and hence the theorem is proved.

Now we give several examples to illustrate the use of this theorem.

Example 18.13. Let X be a random variable with a density function $f(x)$. What is the critical region for the best test of

$$H_o : f(x) = \begin{cases} \frac{1}{2} & \text{if } -1 < x < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

against

$$H_a : f(x) = \begin{cases} 1 - |x| & \text{if } -1 < x < 1 \\ \\ 0 & \text{elsewhere,} \end{cases}$$

at the significance size $\alpha = 0.10$?

Answer: We assume that the test is performed with a sample of size 1. Using Neyman-Pearson Theorem, the best critical region for the best test at the significance size α is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L_o(x)}{L_a(x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{\frac{1}{2}}{1 - |x|} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid |x| \leq 1 - \frac{1}{2k} \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{1}{2k} - 1 \leq x \leq 1 - \frac{1}{2k} \right\}. \end{aligned}$$

Since

$$\begin{aligned} 0.1 &= P(\,C\,) \\ &= P\left(\frac{L_o(x)}{L_a(x)} \leq k \;/\; H_o \text{ is true} \right) \\ &= P\left(\frac{\frac{1}{2}}{1 - |x|} \leq k \;/\; H_o \text{ is true} \right) \\ &= P\left(\frac{1}{2k} - 1 \leq x \leq 1 - \frac{1}{2k} \;/\; H_o \text{ is true} \right), \\ &= \int_{\frac{1}{2k} - 1}^{1 - \frac{1}{2k}} \frac{1}{2}\, dx \\ &= 1 - \frac{1}{2k}, \end{aligned}$$

we get the critical region C to be

$$C = \{ x \in \mathbb{R} \mid -0.1 \leq x \leq 0.1 \}.$$

Thus the best critical region is $C = [-0.1,\ 0.1]$ and the best test is: "Reject H_o if $-0.1 \leq X \leq 0.1$".

Example 18.14. Suppose X has the density function

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on a single observed value of X, find the most powerful critical region of size $\alpha = 0.1$ for testing $H_o : \theta = 1$ against $H_a : \theta = 2$.

Answer: By Neyman-Pearson Theorem, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L(\theta_o, x)}{L(\theta_a, x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{(1+\theta_o)\, x^{\theta_o}}{(1+\theta_a)\, x^{\theta_a}} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{2x}{3x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \frac{1}{x} \leq \frac{3}{2} k \right\} \\ &= \{ x \in \mathbb{R} \mid x \geq a, \} \end{aligned}$$

where a is some constant. Hence the most powerful or best test is of the form: "Reject H_o if $X \geq a$."

Since, the significance level of the test is given to be $\alpha = 0.1$, the constant a can be determined. Now we proceed to find a. Since

$$\begin{aligned} 0.1 &= \alpha \\ &= P\,(\text{Reject } H_o \ / \ H_o \text{ is true}\} \\ &= P\,(X \geq a \ / \ \theta = 1) \\ &= \int_a^1 2x\, dx \\ &= 1 - a^2, \end{aligned}$$

hence

$$a^2 = 1 - 0.1 = 0.9.$$

Therefore

$$a = \sqrt{0.9},$$

since k in Neyman-Pearson Theorem is positive. Hence, the most powerful test is given by "Reject H_o if $X \geq \sqrt{0.9}$".

Example 18.15. Suppose that X is a random variable about which the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \sim N(0,1)$ is to be tested. What is the most powerful test with a significance level $\alpha = 0.05$ based on one observation of X?

Answer: By Neyman-Pearson Theorem, the form of the critical region is given by

$$\begin{aligned} C &= \left\{ x \in \mathbb{R} \mid \frac{L_o(x)}{L_a(x)} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid \sqrt{2\pi}\, e^{\frac{1}{2}x^2} \leq k \right\} \\ &= \left\{ x \in \mathbb{R} \mid x^2 \leq 2 \ln \left(\frac{k}{\sqrt{2\pi}} \right) \right\} \\ &= \{ x \in \mathbb{R} \mid x \leq a, \} \end{aligned}$$

where a is some constant. Hence the most powerful or best test is of the form: "Reject H_o if $X \leq a$."

Since, the significance level of the test is given to be $\alpha = 0.05$, the constant a can be determined. Now we proceed to find a. Since

$$\begin{aligned} 0.05 &= \alpha \\ &= P\left(\text{Reject } H_o \ / \ H_o \text{ is true}\right\} \\ &= P\left(X \leq a \ / \ X \sim UNIF(0,1)\right) \\ &= \int_0^a dx \\ &= a, \end{aligned}$$

hence $a = 0.05$. Thus, the most powerful critical region is given by

$$C = \{ x \in \mathbb{R} \mid 0 < x \leq 0.05 \}$$

based on the support of the uniform distribution on the open interval $(0,1)$. Since the support of this uniform distribution is the interval $(0,1)$, the acceptance region (or the complement of C in $(0,1)$) is

$$C^c = \{ x \in \mathbb{R} \mid 0.05 < x < 1 \}.$$

However, since the support of the standard normal distribution is $\mathbb{R}$, the actual critical region should be the complement of C^c in $\mathbb{R}$. Therefore, the critical region of this hypothesis test is the set

$$\{x \in \mathbb{R} \,|\, x \leq 0.05 \text{ or } x \geq 1\}.$$

The most powerful test for $\alpha = 0.05$ is: "Reject H_o if $X \leq 0.05$ or $X \geq 1$."

Example 18.16. Let X_1, X_2, X_3 denote three independent observations from a distribution with density

$$f(x;\theta) = \begin{cases} (1+\theta)\, x^{\theta} & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the best critical region of size 0.034 for testing $H_o : \theta = 1$ versus $H_a : \theta = 2$?

Answer: By Neyman-Pearson Theorem, the form of the critical region is given by (with $\theta_o = 1$ and $\theta_a = 2$)

$$\begin{aligned} C &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{L(\theta_o, x_1, x_2, x_3)}{L(\theta_a, x_1, x_2, x_3)} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{(1+\theta_o)^3 \prod_{i=1}^{3} x_i^{\theta_o}}{(1+\theta_a)^3 \prod_{i=1}^{3} x_i^{\theta_a}} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{8 x_1 x_2 x_3}{27 x_1^2 x_2^2 x_3^2} \leq k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid \frac{1}{x_1 x_2 x_3} \leq \frac{27}{8} k \right\} \\ &= \left\{ (x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1 x_2 x_3 \geq a, \right\} \end{aligned}$$

where a is some constant. Hence the most powerful or best test is of the form: "Reject H_o if $\displaystyle\prod_{i=1}^{3} X_i \geq a$."

Since, the significance level of the test is given to be $\alpha = 0.034$, the constant a can be determined. To evaluate the constant a, we need the probability distribution of $X_1 X_2 X_3$. The distribution of $X_1 X_2 X_3$ is not easy to get. Hence, we will use Theorem 17.5. There, we have shown that

$-2(1+\theta)\sum_{i=1}^{3} \ln X_i \sim \chi^2(6)$. Now we proceed to find a. Since

$$
\begin{aligned}
0.034 &= \alpha \\
&= P\left(\text{Reject } H_o \;/\; H_o \text{ is true}\right\} \\
&= P\left(X_1X_2X_3 \geq a \;/\; \theta = 1\right) \\
&= P\left(\ln(X_1X_2X_3) \geq \ln a \;/\; \theta = 1\right) \\
&= P\left(-2(1+\theta)\ln(X_1X_2X_3) \leq -2(1+\theta)\ln a \;/\; \theta = 1\right) \\
&= P\left(-4\ln(X_1X_2X_3) \leq -4\ln a\right) \\
&= P\left(\chi^2(6) \leq -4\ln a\right)
\end{aligned}
$$

hence from chi-square table, we get

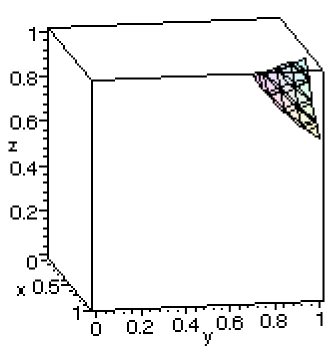
$$-4\ln a = 1.4.$$

Therefore

$$a = e^{-0.35} = 0.7047.$$

Hence, the most powerful test is given by "Reject H_o if $X_1X_2X_3 \geq 0.7047$".

The critical region C is the region above the surface $x_1x_2x_3 = 0.7047$ of the unit cube $[0,1]^3$. The following figure illustrates this region.



Critical region is to the right of the shaded surface

Example 18.17. Let $X_1, X_2, ..., X_{12}$ be a random sample from a normal population with mean zero and variance σ^2. What is the most powerful test of size 0.025 for testing the null hypothesis $H_o : \sigma^2 = 10$ versus $H_a : \sigma^2 = 5$?

Answer: By Neyman-Pearson Theorem, the form of the critical region is given by (with ${\sigma_o}^2 = 10$ and ${\sigma_a}^2 = 5$)

$$
\begin{aligned}
C &= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \frac{L\left({\sigma_o}^2, x_1, x_2, ..., x_{12}\right)}{L\left({\sigma_a}^2, x_1, x_2, ..., x_{12}\right)} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \prod_{i=1}^{12} \frac{\frac{1}{\sqrt{2\pi\sigma_o^2}}\, e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_o}\right)^2}}{\frac{1}{\sqrt{2\pi\sigma_a^2}}\, e^{-\frac{1}{2}\left(\frac{x_i}{\sigma_a}\right)^2}} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \left(\frac{1}{2}\right)^6 e^{\frac{1}{20}\sum_{i=1}^{12} x_i^2} \leq k \right\} \\
&= \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \;\middle|\; \sum_{i=1}^{12} x_i^2 \leq a \right\},
\end{aligned}
$$

where a is some constant. Hence the most powerful or best test is of the form: "Reject H_o if $\displaystyle\sum_{i=1}^{12} X_i^2 \leq a$."

Since, the significance level of the test is given to be $\alpha = 0.025$, the constant a can be determined. To evaluate the constant a, we need the probability distribution of $X_1^2 + X_2^2 + \cdots + X_{12}^2$. It can be shown that the distribution of $\sum_{i=1}^{12} \left(\frac{X_i}{\sigma}\right)^2 \sim \chi^2(12)$. Now we proceed to find a. Since

$$
\begin{aligned}
0.034 &= \alpha \\
&= P\left(\text{Reject } H_o \;/\; H_o \text{ is true}\right\} \\
&= P\left(\sum_{i=1}^{12} \left(\frac{X_i}{\sigma}\right)^2 \leq a \;/\; \sigma^2 = 10 \right) \\
&= P\left(\sum_{i=1}^{12} \left(\frac{X_i}{\sqrt{10}}\right)^2 \leq a \;/\; \sigma^2 = 10 \right) \\
&= P\left(\chi^2(12) \leq \frac{a}{10} \right),
\end{aligned}
$$

hence from chi-square table, we get

$$\frac{a}{10} = 4.4.$$

Therefore

$$a = 44.$$

Hence, the most powerful test is given by "Reject H_o if $\sum_{i=1}^{12} X_i^2 \leq 44$." The best critical region of size 0.025 is given by

$$C = \left\{ (x_1, x_2, ..., x_{12}) \in \mathbb{R}^{12} \mid \sum_{i=1}^{12} x_i^2 \leq 44 \right\}.$$

In last five examples, we have found the most powerful tests and corresponding critical regions when the both H_o and H_a are simple hypotheses. If either H_o or H_a is not simple, then it is not always possible to find the most powerful test and corresponding critical region. In this situation, hypothesis test is found by using the likelihood ratio. A test obtained by using likelihood ratio is called the *likelihood ratio test* and the corresponding critical region is called the *likelihood ratio critical region*.

18.4. Some Examples of Likelihood Ratio Tests

In this section, we illustrate, using likelihood ratio, how one can construct hypothesis test when one of the hypotheses is not simple. As pointed out earlier, the test we will construct using the likelihood ratio is not the most powerful test. However, such a test has all the desirable properties of a hypothesis test. To construct the test one has to follow a sequence of steps. These steps are outlined below:

(1) Find the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ for the given sample.

(2) Evaluate $\max\limits_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)$.

(3) Find the maximum likelihood estimator $\widehat{\theta}$ of θ.

(4) Compute $\max\limits_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)$ using $L\left(\widehat{\theta}, x_1, x_2, ..., x_n\right)$.

(5) Using steps (2) and (4), find $W(x_1, ..., x_n) = \dfrac{\max\limits_{\theta \in \Omega_o} L(\theta, x_1, x_2, ..., x_n)}{\max\limits_{\theta \in \Omega} L(\theta, x_1, x_2, ..., x_n)}$.

(6) Using step (5) determine $C = \{(x_1, x_2, ..., x_n) \,|\, W(x_1, ..., x_n) \leq k\}$, where $k \in [0, 1]$.

(7) Reduce $W(x_1, ..., x_n) \leq k$ to an equivalent inequality $\widehat{W}(x_1, ..., x_n) \leq A$.

(8) Determine the distribution of $\widehat{W}(x_1, ..., x_n)$.

(9) Find A such that given α equals $P\left(\widehat{W}(x_1, ..., x_n) \leq A \,|\, H_o \text{ is true}\right)$.

In the remaining examples, for notational simplicity, we will denote the likelihood function $L(\theta, x_1, x_2, ..., x_n)$ simply as $L(\theta)$.

Example 18.19. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and `known variance` σ^2. What is the likelihood ratio test of size α for testing the null hypothesis $H_o : \mu = \mu_o$ versus the alternative hypothesis $H_a : \mu \neq \mu_o$?

Answer: The likelihood function of the sample is given by

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \right) e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Since $\Omega_o = \{\mu_o\}$, we obtain

$$\begin{aligned} \max_{\mu\in\Omega_o} L(\mu) &= L(\mu_o) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2}. \end{aligned}$$

We have seen in Example 15.13 that if $X \sim N(\mu, \sigma^2)$, then the maximum likelihood estimator of μ is $\overline{X}$, that is

$$\widehat{\mu} = \overline{X}.$$

Hence

$$\max_{\mu\in\Omega} L(\mu) = L(\widehat{\mu}) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Now the likelihood ratio statistics $W(x_1, x_2, ..., x_n)$ is given by

$$W(x_1, x_2, ..., x_n) = \frac{\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2}}{\left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

which simplifies to

$$W(x_1, x_2, ..., x_n) = e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_o)^2}.$$

Now the inequality $W(x_1, x_2, ..., x_n) \leq k$ becomes

$$e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_o)^2} \leq k$$

and which can be rewritten as

$$(\overline{x} - \mu_o)^2 \geq -\frac{2\sigma^2}{n} \ln(k)$$

or

$$|\overline{x} - \mu_o| \geq K$$

where $K = \sqrt{-\frac{2\sigma^2}{n} \ln(k)}$. In view of the above inequality, the critical region can be described as

$$C = \{(x_1, x_2, ..., x_n) \mid |\overline{x} - \mu_o| \geq K\}.$$

Since we are given the size of the critical region to be α, we can determine the constant K. Since the size of the critical region is α, we have

$$\alpha = P\left(\left|\overline{X} - \mu_o\right| \geq K\right).$$

For finding K, we need the probability density function of the statistic $\overline{X} - \mu_o$ when the population X is $N(\mu, \sigma^2)$ and the null hypothesis $H_o : \mu = \mu_o$ is true. Since σ^2 is known and $X_i \sim N(\mu, \sigma^2)$,

$$\frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

and

$$\begin{aligned}
\alpha &= P\left(\left|\overline{X} - \mu_o\right| \geq K\right) \\
&= P\left(\left|\frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}\right| \geq K\frac{\sqrt{n}}{\sigma}\right) \\
&= P\left(|Z| \geq K\frac{\sqrt{n}}{\sigma}\right) \qquad \text{where} \qquad Z = \frac{\overline{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}} \\
&= 1 - P\left(-K\frac{\sqrt{n}}{\sigma} \leq Z \leq K\frac{\sqrt{n}}{\sigma}\right)
\end{aligned}$$

we get

$$z_{\frac{\alpha}{2}} = K \frac{\sqrt{n}}{\sigma}$$

which is

$$K = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

where $z_{\frac{\alpha}{2}}$ is a real number such that the integral of the standard normal density from $z_{\frac{\alpha}{2}}$ to ∞ equals $\frac{\alpha}{2}$.

Hence, the likelihood ratio test is given by "Reject H_o if

$$\left|\overline{X} - \mu_o\right| \geq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}."$$

If we denote

$$z = \frac{\overline{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$

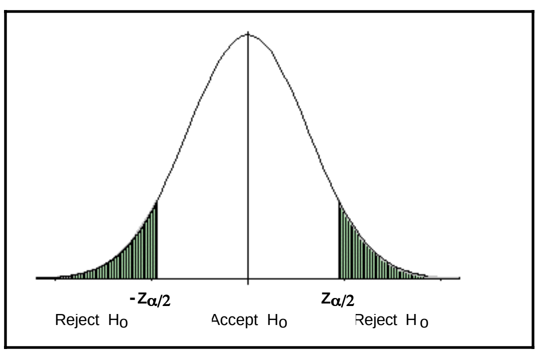
then the above inequality becomes

$$|Z| \geq z_{\frac{\alpha}{2}}.$$

Thus critical region is given by

$$C = \left\{(x_1, x_2, ..., x_n) \mid \; |z| \geq z_{\frac{\alpha}{2}} \right\}.$$

This tells us that the null hypothesis must be rejected when the absolute value of z takes on a value greater than or equal to $z_{\frac{\alpha}{2}}$.



Remark 18.6. The hypothesis $H_a : \mu \neq \mu_o$ is called a two-sided alternative hypothesis. An alternative hypothesis of the form $H_a : \mu > \mu_o$ is called a right-sided alternative. Similarly, $H_a : \mu < \mu_o$ is called the a left-sided

alternative. In the above example, if we had a right-sided alternative, that is $H_a : \mu > \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \; z \geq z_\alpha \}.$$

Similarly, if the alternative would have been left-sided, that is $H_a : \mu < \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \; z \leq -z_\alpha \}.$$

We summarize the three cases of hypotheses test of the mean (of the normal population with variance) in the following table.

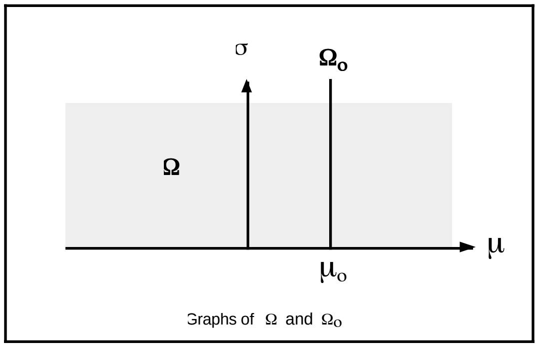
H_o	H_a	Critical Region (or Test)
$\mu = \mu_o$	$\mu > \mu_o$	$z = \frac{\overline{x}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \geq z_\alpha$
$\mu = \mu_o$	$\mu < \mu_o$	$z = \frac{\overline{x}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \leq -z_\alpha$
$\mu = \mu_o$	$\mu \neq \mu_o$	$\lvert z \rvert = \left\lvert \frac{\overline{x}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \right\rvert \geq z_{\frac{\alpha}{2}}$

Example 18.20. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and `unknown variance` σ^2. What is the likelihood ratio test of size α for testing the null hypothesis $H_o : \mu = \mu_o$ versus the alternative hypothesis $H_a : \mu \neq \mu_o$?

Answer: In this example,

$$\begin{aligned}
\Omega &= \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \; -\infty < \mu < \infty, \; \sigma^2 > 0 \right\}, \\
\Omega_o &= \left\{ \left(\mu_o, \sigma^2\right) \in \mathbb{R}^2 \mid \sigma^2 > 0 \right\}, \\
\Omega_a &= \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \mu \neq \mu_o, \; \sigma^2 > 0 \right\}.
\end{aligned}$$

These sets are illustrated below.



The likelihood function is given by

$$L\left(\mu, \sigma^2\right) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Next, we find the maximum of $L\left(\mu, \sigma^2\right)$ on the set Ω_o. Since the set Ω_o is equal to $\left\{\left(\mu_o, \sigma^2\right) \in \mathbb{R}^2 \mid 0 < \sigma < \infty\right\}$, we have

$$\max_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu, \sigma^2\right) = \max_{\sigma^2>0} L\left(\mu_o, \sigma^2\right).$$

Since $L\left(\mu_o, \sigma^2\right)$ and $\ln L\left(\mu_o, \sigma^2\right)$ achieve the maximum at the same σ value, we determine the value of σ where $\ln L\left(\mu_o, \sigma^2\right)$ achieves the maximum. Taking the natural logarithm of the likelihood function, we get

$$\ln\left(L\left(\mu, \sigma^2\right)\right) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_o)^2.$$

Differentiating $\ln L\left(\mu_o, \sigma^2\right)$ with respect to σ^2, we get from the last equality

$$\frac{d}{d\sigma^2}\ln\left(L\left(\mu, \sigma^2\right)\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu_o)^2.$$

Setting this derivative to zero and solving for σ, we obtain

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2}.$$

Thus $\ln\left(L\left(\mu,\sigma^2\right)\right)$ attain maximum at $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2}$. Since this value of σ is also yield maximum value of $L\left(\mu,\sigma^2\right)$, we have

$$\max_{\sigma^2>0} L\left(\mu_o,\sigma^2\right) = \left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}.$$

Next, we determine the maximum of $L\left(\mu,\sigma^2\right)$ on the set Ω. As before, we consider $\ln L\left(\mu,\sigma^2\right)$ to determine where $L\left(\mu,\sigma^2\right)$ achieves maximum. Taking the natural logarithm of $L\left(\mu,\sigma^2\right)$, we obtain

$$\ln\left(L\left(\mu,\sigma^2\right)\right) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the partial derivatives of $\ln L\left(\mu,\sigma^2\right)$ first with respect to μ and then with respect to σ^2, we get

$$\frac{\partial}{\partial\mu}\ln L\left(\mu,\sigma^2\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu),$$

and

$$\frac{\partial}{\partial\sigma^2}\ln L\left(\mu,\sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,$$

respectively. Setting these partial derivatives to zero and solving for μ and σ, we obtain

$$\mu = \overline{x} \qquad \text{and} \qquad \sigma^2 = \frac{n-1}{n}\, s^2,$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$ is the sample variance.

Letting these optimal values of μ and σ into $L\left(\mu,\sigma^2\right)$, we obtain

$$\max_{(\mu,\sigma^2)\in\Omega} L\left(\mu,\sigma^2\right) = \left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}.$$

Hence

$$\frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu,\sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega} L\left(\mu,\sigma^2\right)} = \frac{\left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_o)^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}}{\left(2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{-\frac{n}{2}} e^{-\frac{n}{2}}} = \left(\frac{\sum_{i=1}^{n}(x_i-\mu_o)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{-\frac{n}{2}}.$$

Since

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=(n-1)\,s^2$$

and

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n\,(\overline{x}-\mu_o)^2\,,$$

we get

$$W(x_1,x_2,...,x_n)=\frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o}L\left(\mu,\sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega}L\left(\mu,\sigma^2\right)}=\left(1+\frac{n}{n-1}\frac{(\overline{x}-\mu_o)^2}{s^2}\right)^{-\frac{n}{2}}.$$

Now the inequality $W(x_1,x_2,...,x_n)\leq k$ becomes

$$\left(1+\frac{n}{n-1}\frac{(\overline{x}-\mu_o)^2}{s^2}\right)^{-\frac{n}{2}}\leq k$$

and which can be rewritten as

$$\left(\frac{\overline{x}-\mu_o}{s}\right)^2\geq\frac{n-1}{n}\left(k^{-\frac{2}{n}}-1\right)$$

or

$$\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K$$

where $K=\sqrt{(n-1)\left[k^{-\frac{2}{n}}-1\right]}$. In view of the above inequality, the critical region can be described as

$$C=\left\{(x_1,x_2,...,x_n)\;\middle|\;\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K\right\}$$

and the best likelihood ratio test is: "Reject H_o if $\left|\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}\right|\geq K$". Since we are given the size of the critical region to be α, we can find the constant K. For finding K, we need the probability density function of the statistic $\frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}}$ when the population X is $N(\mu,\sigma^2)$ and the null hypothesis $H_o:\mu=\mu_o$ is true.

Since the population is normal with mean μ and variance σ^2,

$$\frac{\overline{X}-\mu_o}{\frac{S}{\sqrt{n}}}\sim t(n-1),$$

where S^2 is the sample variance and equals to $\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Hence

$$K = t_{\frac{\alpha}{2}}(n-1)\,\frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2}}(n-1)$ is a real number such that the integral of the t-distribution with $n-1$ degrees of freedom from $t_{\frac{\alpha}{2}}(n-1)$ to ∞ equals $\frac{\alpha}{2}$.

Therefore, the likelihood ratio test is given by "Reject $H_o : \mu = \mu_o$ if

$$\left|\overline{X} - \mu_o\right| \geq t_{\frac{\alpha}{2}}(n-1)\,\frac{S}{\sqrt{n}}."$$

If we denote

$$t = \frac{\overline{x} - \mu_o}{\frac{s}{\sqrt{n}}}$$

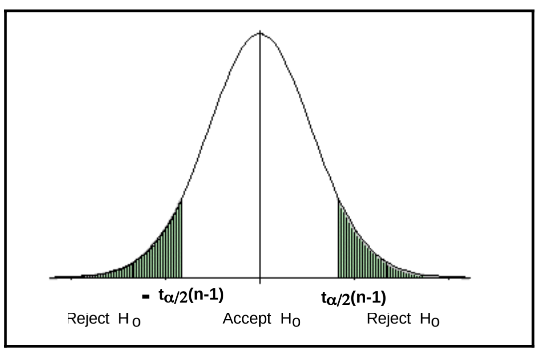
then the above inequality becomes

$$|T| \geq t_{\frac{\alpha}{2}}(n-1).$$

Thus critical region is given by

$$C = \left\{(x_1, x_2, ..., x_n) \mid \; |t| \geq t_{\frac{\alpha}{2}}(n-1)\right\}.$$

This tells us that the null hypothesis must be rejected when the absolute value of t takes on a value greater than or equal to $t_{\frac{\alpha}{2}}(n-1)$.



Remark 18.7. In the above example, if we had a right-sided alternative, that is $H_a : \mu > \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \; t \geq t_{\alpha}(n-1)\}.$$

Similarly, if the alternative would have been left-sided, that is $H_a : \mu < \mu_o$, then the critical region would have been

$$C = \{(x_1, x_2, ..., x_n) \mid \; t \leq -t_\alpha(n-1)\}.$$

We summarize the three cases of hypotheses test of the mean (of the normal population with variance) in the following table.

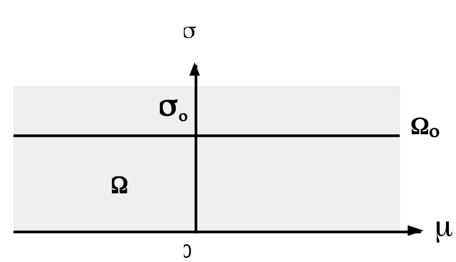
H_o	H_a	Critical Region (or Test)
$\mu = \mu_o$	$\mu > \mu_o$	$t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \geq t_\alpha(n-1)$
$\mu = \mu_o$	$\mu < \mu_o$	$t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \leq -t_\alpha(n-1)$
$\mu = \mu_o$	$\mu \neq \mu_o$	$\lvert t \rvert = \left\lvert \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \right\rvert \geq t_{\frac{\alpha}{2}}(n-1)$

Example 18.21. Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with mean μ and variance σ^2. What is the likelihood ratio test of significance of size α for testing the null hypothesis $H_o : \sigma^2 = \sigma_o^2$ versus $H_a : \sigma^2 \neq \sigma_o^2$?

Answer: In this example,

$$\Omega = \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \; -\infty < \mu < \infty, \; \sigma^2 > 0 \right\},$$
$$\Omega_o = \left\{ \left(\mu, \sigma_o^2\right) \in \mathbb{R}^2 \mid \; -\infty < \mu < \infty \right\},$$
$$\Omega_a = \left\{ \left(\mu, \sigma^2\right) \in \mathbb{R}^2 \mid \; -\infty < \mu < \infty, \; \sigma \neq \sigma_o \right\}.$$

These sets are illustrated below.



Graphs of Ω and Ωo

The likelihood function is given by

$$\begin{aligned} L\left(\mu,\sigma^2\right) &= \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Next, we find the maximum of $L\left(\mu,\sigma^2\right)$ on the set Ω_o. Since the set Ω_o is equal to $\left\{\left(\mu,\sigma_o^2\right) \in \mathbb{R}^2 \mid -\infty < \mu < \infty\right\}$, we have

$$\max_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu,\sigma^2\right) = \max_{-\infty<\mu<\infty} L\left(\mu,\sigma_o^2\right).$$

Since $L\left(\mu,\sigma_o^2\right)$ and $\ln L\left(\mu,\sigma_o^2\right)$ achieve the maximum at the same μ value, we determine the value of μ where $\ln L\left(\mu,\sigma_o^2\right)$ achieves the maximum. Taking the natural logarithm of the likelihood function, we get

$$\ln\left(L\left(\mu,\sigma_o^2\right)\right) = -\frac{n}{2}\ln(\sigma_o^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Differentiating $\ln L\left(\mu,\sigma_o^2\right)$ with respect to μ, we get from the last equality

$$\frac{d}{d\mu}\ln\left(L\left(\mu,\sigma^2\right)\right) = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(x_i-\mu).$$

Setting this derivative to zero and solving for μ, we obtain

$$\mu = \overline{x}.$$

Hence, we obtain

$$\max_{-\infty<\mu<\infty} L\left(\mu,\sigma^2\right) = \left(\frac{1}{2\pi\sigma_o^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

Next, we determine the maximum of $L\left(\mu,\sigma^2\right)$ on the set Ω. As before, we consider $\ln L\left(\mu,\sigma^2\right)$ to determine where $L\left(\mu,\sigma^2\right)$ achieves maximum. Taking the natural logarithm of $L\left(\mu,\sigma^2\right)$, we obtain

$$\ln\left(L\left(\mu,\sigma^2\right)\right) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Taking the partial derivatives of $\ln L\left(\mu, \sigma^2\right)$ first with respect to μ and then with respect to σ^2, we get

$$\frac{\partial}{\partial \mu} \ln L\left(\mu, \sigma^2\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu),$$

and

$$\frac{\partial}{\partial \sigma^2} \ln L\left(\mu, \sigma^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2,$$

respectively. Setting these partial derivatives to zero and solving for μ and σ, we obtain

$$\mu = \overline{x} \qquad \text{and} \qquad \sigma^2 = \frac{n-1}{n}\, s^2,$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is the sample variance.

Letting these optimal values of μ and σ into $L\left(\mu, \sigma^2\right)$, we obtain

$$\max_{(\mu,\sigma^2)\in\Omega} L\left(\mu, \sigma^2\right) = \left(\frac{n}{2\pi(n-1)s^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Therefore

$$\begin{aligned} W(x_1, x_2, ..., x_n) &= \frac{\max\limits_{(\mu,\sigma^2)\in\Omega_o} L\left(\mu, \sigma^2\right)}{\max\limits_{(\mu,\sigma^2)\in\Omega} L\left(\mu, \sigma^2\right)} \\ &= \frac{\left(\frac{1}{2\pi\sigma_o^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma_o^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\left(\frac{n}{2\pi(n-1)s^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i - \overline{x})^2}} \\ &= n^{-\frac{n}{2}}\, e^{\frac{n}{2}} \left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma_o^2}}. \end{aligned}$$

Now the inequality $W(x_1, x_2, ..., x_n) \leq k$ becomes

$$n^{-\frac{n}{2}}\, e^{\frac{n}{2}} \left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma_o^2}} \leq k$$

which is equivalent to

$$\left(\frac{(n-1)s^2}{\sigma_o^2}\right)^{n} e^{-\frac{(n-1)s^2}{\sigma_o^2}} \leq \left(k\left(\frac{n}{e}\right)^{\frac{n}{2}}\right)^2 := K_o,$$

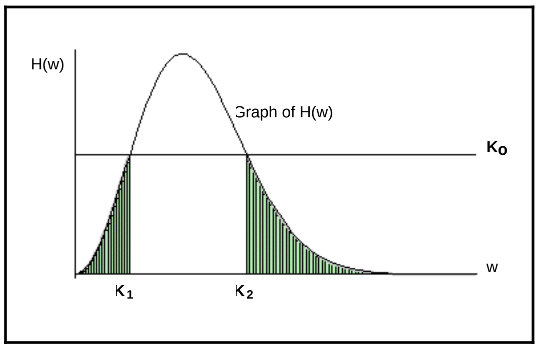
where K_o is a constant. Let H be a function defined by

$$H(w) = w^n\, e^{-w}.$$

Using this, we see that the above inequality becomes

$$H\left(\frac{(n-1)s^2}{\sigma_o^2}\right) \leq K_o.$$

The figure below illustrates this inequality.



From this it follows that

$$\frac{(n-1)s^2}{\sigma_o^2} \leq K_1 \qquad \text{or} \qquad \frac{(n-1)s^2}{\sigma_o^2} \geq K_2.$$

In view of these inequalities, the critical region can be described as

$$C = \left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{(n-1)s^2}{\sigma_o^2} \leq K_1 \;\text{ or }\; \frac{(n-1)s^2}{\sigma_o^2} \geq K_2 \right\},$$

and the best likelihood ratio test is: "Reject H_o if

$$\frac{(n-1)S^2}{\sigma_o^2} \leq K_1 \;\text{ or }\; \frac{(n-1)S^2}{\sigma_o^2} \geq K_2."$$

Since we are given the size of the critical region to be α, we can determine the constants K_1 and K_2. As the sample $X_1, X_2, ..., X_n$ is taken from a normal distribution with mean μ and variance σ^2, we get

$$\frac{(n-1)S^2}{\sigma_o^2} \sim \chi^2(n-1)$$

when the null hypothesis $H_o : \sigma^2 = \sigma_o^2$ is true.

Therefore, the likelihood ratio critical region C becomes

$$\left\{ (x_1, x_2, ..., x_n) \;\middle|\; \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\frac{\alpha}{2}}(n-1) \text{ or } \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\frac{\alpha}{2}}(n-1) \right\}$$

and the likelihood ratio test is: "Reject $H_o : \sigma^2 = \sigma_o^2$ if

$$\frac{(n-1)S^2}{\sigma_o^2} \leq \chi^2_{\frac{\alpha}{2}}(n-1) \text{ or } \frac{(n-1)S^2}{\sigma_o^2} \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)\text{"}$$

where $\chi^2_{\frac{\alpha}{2}}(n-1)$ is a real number such that the integral of the chi-square density function with $(n-1)$ degrees of freedom from 0 to $\chi^2_{\frac{\alpha}{2}}(n-1)$ is $\frac{\alpha}{2}$. Further, $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ denotes the real number such that the integral of the chi-square density function with $(n-1)$ degrees of freedom from $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ to ∞ is $\frac{\alpha}{2}$.

Remark 18.8. We summarize the three cases of hypotheses test of the variance (of the normal population with unknown mean) in the following table.

H_o	H_a	Critical Region (or Test)
$\sigma^2 = \sigma_o^2$	$\sigma^2 > \sigma_o^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\alpha}(n-1)$
$\sigma^2 = \sigma_o^2$	$\sigma^2 < \sigma_o^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\alpha}(n-1)$
$\sigma^2 = \sigma_o^2$	$\sigma^2 \neq \sigma_o^2$	$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \geq \chi^2_{1-\alpha/2}(n-1)$ or $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} \leq \chi^2_{\alpha/2}(n-1)$

18.5. Review Exercises

1. Five trials $X_1, X_2, ..., X_5$ of a Bernoulli experiment were conducted to test $H_o : p = \frac{1}{2}$ against $H_a : p = \frac{3}{4}$. The null hypothesis H_o will be rejected if $\sum_{i=1}^{5} X_i = 5$. Find the probability of Type I and Type II errors.

2. A manufacturer of car batteries claims that the life of his batteries is normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

3. Let $X_1, X_2, ..., X_8$ be a random sample of size 8 from a Poisson distribution with parameter λ. Reject the null hypothesis $H_o : \lambda = 0.5$ if the observed sum $\sum_{i=1}^{8} x_i \geq 8$. First, compute the significance level α of the test. Second, find the power function $\beta(\lambda)$ of the test as a sum of Poisson probabilities when H_a is true.

4. Suppose X has the density function

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{for } 0 < x < \theta \\ \\ 0 & \text{otherwise.} \end{cases}$$

If one observation of X is taken, what are the probabilities of Type I and Type II errors in testing the null hypothesis $H_o : \theta = 1$ against the alternative hypothesis $H_a : \theta = 2$, if H_o is rejected for $X > 0.92$.

5. Let X have the density function

$$f(x) = \begin{cases} (\theta + 1)\, x^{\theta} & \text{for } 0 < x < 1 \text{ where } \theta > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The hypothesis $H_o : \theta = 1$ is to be rejected in favor of $H_1 : \theta = 2$ if $X > 0.90$. What is the probability of Type I error?

6. Let $X_1, X_2, ..., X_6$ be a random sample from a distribution with density function

$$f(x) = \begin{cases} \theta\, x^{\theta - 1} & \text{for } 0 < x < 1 \text{ where } \theta > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

The null hypothesis $H_o : \theta = 1$ is to be rejected in favor of the alternative $H_a : \theta > 1$ if and only if *at least* 5 of the sample observations are larger than 0.7. What is the significance level of the test?

7. A researcher wants to test $H_o : \theta = 0$ versus $H_a : \theta = 1$, where θ is a parameter of a population of interest. The statistic W, based on a random sample of the population, is used to test the hypothesis. Suppose that under H_o, W has a normal distribution with mean 0 and variance 1, and under H_a, W has a normal distribution with mean 4 and variance 1. If H_o is rejected when $W > 1.50$, then what are the probabilities of a Type I or Type II error respectively?

8. Let X_1 and X_2 be a random sample of size 2 from a normal distribution $N(\mu,\ 1)$. Find the *likelihood ratio critical region* of size 0.005 for testing the null hypothesis $H_o : \mu = 0$ against the composite alternative $H_a : \mu \neq 0$?

9. Let $X_1, X_2, ..., X_{10}$ be a random sample from a Poisson distribution with mean θ. What is the most powerful (or best) critical region of size 0.08 for testing the null hypothesis $H_0 : \theta = 0.1$ against $H_a : \theta = 0.5$?

10. Let X be a random sample of size 1 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} (1 - \frac{\theta}{2}) + \theta\, x & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

For a significance level $\alpha = 0.1$, what is the *best (or uniformly most powerful) critical region* for testing the null hypothesis $H_o : \theta = -1$ against $H_a : \theta = 1$?

11. Let X_1, X_2 be a random sample of size 2 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{\theta^x\, e^{-\theta}}{x!} & \text{if } x = 0, 1, 2, 3, ... \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. For a significance level $\alpha = 0.053$, what is the *best critical region* for testing the null hypothesis $H_o : \theta = 1$ against $H_a : \theta = 2$? Sketch the graph of the best critical region.

12. Let $X_1, X_2, ..., X_8$ be a random sample of size 8 from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{\theta^x\, e^{-\theta}}{x!} & \text{if } x = 0, 1, 2, 3, ... \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. What is the *likelihood ratio critical region* for testing the null hypothesis $H_o : \theta = 1$ against $H_a : \theta \neq 1$? If $\alpha = 0.1$ can you determine the best likelihood ratio critical region?

13. Let $X_1, X_2, ..., X_n$ be a random sample of size n from a distribution with probability density function

$$f(x,\theta) = \begin{cases} \frac{x^6\, e^{\frac{x}{\beta}}}{\Gamma(7)\beta^7}, & \text{if } x > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \geq 0$. What is the *likelihood ratio critical region* for testing the null hypothesis $H_o : \beta = 5$ against $H_a : \beta \neq 5$? What is the *most powerful test ?*

14. Let $X_1, X_2, ..., X_5$ denote a random sample of size 5 from a population X with probability density function

$$f(x;\theta) = \begin{cases} (1-\theta)^{x-1}\,\theta & \text{if } x = 1, 2, 3, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is a parameter. What is the *likelihood ratio critical region* of size 0.05 for testing $H_o : \theta = 0.5$ versus $H_a : \theta \neq 0.5$?

15. Let X_1, X_2, X_3 denote a random sample of size 3 from a population X with probability density function

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2}} \qquad -\infty < x < \infty,$$

where $-\infty < \mu < \infty$ is a parameter. What is the *likelihood ratio critical region* of size 0.05 for testing $H_o : \mu = 3$ versus $H_a : \mu \neq 3$?

16. Let X_1, X_2, X_3 denote a random sample of size 3 from a population X with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* for testing $H_o : \theta = 3$ versus $H_a : \theta \neq 3$?

17. Let X_1, X_2, X_3 denote a random sample of size 3 from a population X with probability density function

$$f(x;\theta) = \begin{cases} \frac{e^{-\theta}\,\theta^x}{x!} & \text{if } x = 0, 1, 2, 3, ..., \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* for testing $H_o : \theta = 0.1$ versus $H_a : \theta \neq 0.1$?

18. A box contains 4 marbles, θ of which are white and the rest are black. A sample of size 2 is drawn to test $H_o : \theta = 2$ versus $H_a : \theta \neq 2$. If the null hypothesis is rejected if both marbles are the same color, find the significance level of the test.

19. Let X_1, X_2, X_3 denote a random sample of size 3 from a population X with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq x \leq \theta \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$ is a parameter. What is the *likelihood ratio critical region* of size $\frac{117}{125}$ for testing $H_o : \theta = 5$ versus $H_a : \theta \neq 5$?

20. Let X_1, X_2 and X_3 denote three independent observations from a distribution with density

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{for } 0 < x < \infty \\ \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \beta < \infty$ is a parameter. What is the *best (or uniformly most powerful* critical region for testing $H_o : \beta = 5$ versus $H_a : \beta = 10$?

Chapter 19

SIMPLE LINEAR REGRESSION AND CORRELATION ANALYSIS

Let X and Y be two random variables with joint probability density function $f(x, y)$. Then the conditional density of Y given that $X = x$ is

$$f(y/x) = \frac{f(x, y)}{g(x)}$$

where

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

is the marginal density of X. The conditional mean of Y

$$E\left(Y|X = x\right) = \int_{-\infty}^{\infty} y f(y/x)\, dy$$

is called the regression equation of Y on X.

Example 19.1. Let X and Y be two random variables with the joint probability density function

$$f(x, y) = \begin{cases} x e^{-x(1+y)} & \text{if } x > 0,\ y > 0 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Find the regression equation of Y on X and then sketch the regression curve.

Answer: The marginal density of X is given by

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} x e^{-x(1+y)}\, dy \\ &= \int_{-\infty}^{\infty} x e^{-x}\, e^{-xy}\, dy \\ &= x e^{-x} \int_{-\infty}^{\infty} e^{-xy}\, dy \\ &= x e^{-x} \left[-\frac{1}{x}\, e^{-xy} \right]_0^{\infty} \\ &= e^{-x}. \end{aligned}$$

The conditional density of Y given $X = x$ is

$$f(y/x) = \frac{f(x,y)}{g(x)} = \frac{x e^{-x(1+y)}}{e^{-x}} = x e^{-xy}, \qquad y > 0.$$

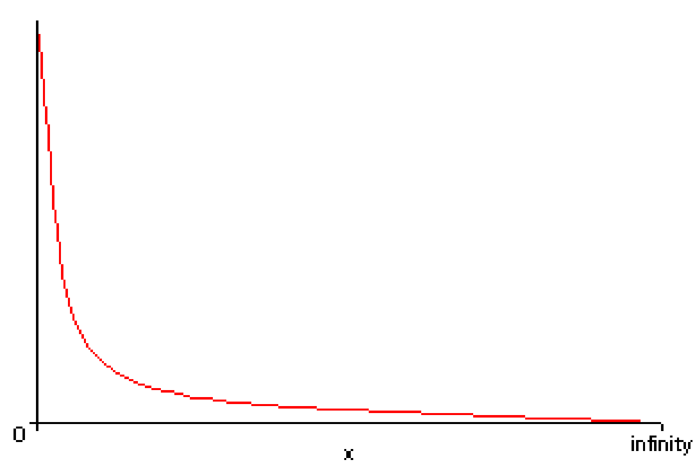
The conditional mean of Y given $X = x$ is given by

$$E(Y/x) = \int_{-\infty}^{\infty} y f(y/x)\, dy = \int_{-\infty}^{\infty} y\, x\, e^{-xy}\, dy = \frac{1}{x}.$$

Thus the regression equation of Y on X is

$$E(Y/x) = \frac{1}{x}, \quad x > 0.$$

The graph of this equation of Y on X is shown below.



Graph of the regression equation E(Y/x) = 1/ x

From this example it is clear that the conditional mean $E(Y/x)$ is a function of x. If this function is of the form $\alpha + \beta x$, then the corresponding regression equation is called a linear regression equation; otherwise it is called a nonlinear regression equation. The term linear regression refers to a specification that is linear in the parameters. Thus $E(Y/x) = \alpha + \beta x^2$ is also a linear regression equation. The regression equation $E(Y/x) = \alpha x^\beta$ is an example of a nonlinear regression equation.

The main purpose of regression analysis is to predict Y_i from the knowledge of x_i using the relationship like

$$E(Y_i/x_i) = \alpha + \beta x_i.$$

The Y_i is called the response or dependent variable where as x_i is called the predictor or independent variable. The term regression has an interesting history, dating back to Francis Galton (1822-1911). Galton studied the heights of fathers and sons, in which he observed a regression (a "turning back") from the heights of sons to the heights of their fathers. That is tall fathers tend to have tall sons and short fathers tend to have short sons. However, he also found that very tall fathers tend to have shorter sons and very short fathers tend to have taller sons. Galton called this phenomenon regression towards the mean.

In regression analysis, that is when investigating the relationship between a predictor and response variable, there are two steps to the analysis. The first step is totally data oriented. This step is always performed. The second step is the statistical one, in which we draw conclusions about the (population) regression equation $E(Y_i/x_i)$. Normally the regression equation contains several parameters. There are two well known methods for finding the estimates of the parameters of the regression equation. These two methods are: (1) The least square method and (2) the normal regression method.

19.1. The Least Squares Method

Let $\{(x_i, y_i) \,|\, i = 1, 2, ..., n\}$ be a set of data. Assume that

$$E(Y_i/x_i) = \alpha + \beta x_i, \tag{1}$$

that is

$$y_i = \alpha + \beta x_i, \qquad i = 1, 2, ..., n.$$

Then the sum of the squares of the error is given by

$$\mathcal{E}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 . \tag{2}$$

The least squares estimates of α and β are defined to be those values which minimize $\mathcal{E}(\alpha, \beta)$. That is,

$$\left(\widehat{\alpha}, \widehat{\beta}\right) = \arg \min_{(\alpha,\beta)} \mathcal{E}(\alpha, \beta).$$

This least squares method is due to Adrien M. Legendre (1752-1833). Note that the least squares method also works even if the regression equation is nonlinear (that is, not of the form (1)).

Next, we give several examples to illustrate the method of least squares.

Example 19.2. Given the five pairs of points (x, y) shown in table below

x	4	0	-2	3	1
y	5	0	0	6	3

what is the line of the form $y = x + b$ best fits the data by method of least squares?

Answer: Suppose the best fit line is $y = x + b$. Then for each x_i, $x_i + b$ is the estimated value of y_i. The difference between y_i and the estimated value of y_i is the error or the residual corresponding to the i^{th} measurement. That is, the error corresponding to the i^{th} measurement is given by

$$\epsilon_i = y_i - x_i - b.$$

Hence the sum of the squares of the errors is

$$\begin{aligned} \mathcal{E}(b) &= \sum_{i=1}^{5} \epsilon_i^2 \\ &= \sum_{i=1}^{5} (y_i - x_i - b)^2 . \end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to b, we get

$$\frac{d}{db}\mathcal{E}(b) = 2 \sum_{i=1}^{5} (y_i - x_i - b)\,(-1).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{5} (y_i - x_i - b) = 0$$

which is

$$5b = \sum_{i=1}^{5} y_i - \sum_{i=1}^{5} x_i.$$

Using the data, we see that

$$5b = 14 - 6$$

which yields $b = \frac{8}{5}$. Hence the best fitted line is

$$y = x + \frac{8}{5}.$$

Example 19.3. Suppose the line $y = bx + 1$ is fit by the method of least squares to the 3 data points

x	1	2	4
y	2	2	0

What is the value of the constant b?

Answer: The error corresponding to the i^{th} measurement is given by

$$\epsilon_i = y_i - bx_i - 1.$$

Hence the sum of the squares of the errors is

$$\begin{aligned} \mathcal{E}(b) &= \sum_{i=1}^{3} \epsilon_i^2 \\ &= \sum_{i=1}^{3} (y_i - bx_i - 1)^2 . \end{aligned}$$

Differentiating $\mathcal{E}(b)$ with respect to b, we get

$$\frac{d}{db}\mathcal{E}(b) = 2 \sum_{i=1}^{3} (y_i - bx_i - 1)\,(-x_i).$$

Setting $\frac{d}{db}\mathcal{E}(b)$ equal to 0, we get

$$\sum_{i=1}^{3} (y_i - bx_i - 1)\, x_i = 0$$

which in turn yields

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

Using the given data we see that

$$b = \frac{6-7}{21} = -\frac{1}{21},$$

and the best fitted line is

$$y = -\frac{1}{21}x + 1.$$

Example 19.4. Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with

$$E(Y_i/x_i) = \theta + 2\ln x_i$$

where θ is an unknown parameter and $x_1, x_2, ..., x_n$ are given constants. What is the least square estimate of the parameter θ?

Answer: The sum of the squares of errors is

$$\mathcal{E}(\theta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \theta - 2\ln x_i)^2 .$$

Differentiating $\mathcal{E}(\theta)$ with respect to θ, we get

$$\frac{d}{d\theta}\mathcal{E}(\theta) = 2\sum_{i=1}^{n} (y_i - \theta - 2\ln x_i)\,(-1).$$

Setting $\frac{d}{d\theta}\mathcal{E}(\theta)$ equal to 0, we get

$$\sum_{i=1}^{n} (y_i - \theta - 2\ln x_i) = 0$$

which is

$$\theta = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} \ln x_i\right).$$

Hence the least squares estimate of θ is $\widehat{\theta} = \overline{y} - \frac{2}{n}\sum_{i=1}^{n} \ln x_i$.

Example 19.5. Given the three pairs of points (x, y) shown below:

x	4	1	2
y	2	1	0

What is the curve of the form $y = x^{\beta}$ best fits the data by method of least squares?

Answer: The sum of the squares of the errors is given by

$$\begin{aligned}\mathcal{E}(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - x_i^{\beta} \right)^2.\end{aligned}$$

Differentiating $\mathcal{E}(\beta)$ with respect to β, we get

$$\frac{d}{d\beta}\mathcal{E}(\beta) = 2\sum_{i=1}^{n} \left(y_i - x_i^{\beta} \right) \left(- x_i^{\beta} \, \ln x_i \right)$$

Setting this derivative $\frac{d}{d\beta}\mathcal{E}(\beta)$ to 0, we get

$$\sum_{i=1}^{n} y_i x_i^{\beta} \, \ln x_i = \sum_{i=1}^{n} x_i^{\beta} \, x_i^{\beta} \, \ln x_i.$$

Using the given data we obtain

$$2\, 4^{\beta} \ln 4 = 4^{2\beta} \ln 4 + 2^{2\beta} \ln 2$$

which simplifies to

$$4 = 2\, 4^{\beta} + 1$$

or

$$4^{\beta} = \frac{3}{2}.$$

Taking the natural logarithm of both sides of the above expression, we get

$$\beta = \frac{\ln 3 - \ln 2}{\ln 4} = 0.2925$$

Thus the least squares best fit model is $y = x^{0.2925}$.

Example 19.6. Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with $E(Y_i/x_i) = \alpha + \beta x_i$, where α and β are unknown parameters, and $x_1, x_2, ..., x_n$ are given constants. What are the least squares estimate of the parameters α and β?

Answer: The sum of the squares of the errors is given by

$$\begin{aligned}\mathcal{E}(\alpha, \beta) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - -\alpha - \beta x_i\right)^2.\end{aligned}$$

Differentiating $\mathcal{E}(\alpha, \beta)$ with respect to α and β respectively, we get

$$\frac{\partial}{\partial \alpha}\mathcal{E}(\alpha, \beta) = 2 \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)(-1)$$

and

$$\frac{\partial}{\partial \beta}\mathcal{E}(\alpha, \beta) = 2 \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)(-x_i).$$

Setting these partial derivatives $\frac{\partial}{\partial \alpha}\mathcal{E}(\alpha, \beta)$ and $\frac{\partial}{\partial \beta}\mathcal{E}(\alpha, \beta)$ to 0, we get

$$\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right) = 0 \tag{3}$$

and

$$\sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right) x_i = 0. \tag{4}$$

From (3), we obtain

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

which is

$$\overline{y} = \alpha + \beta\, \overline{x}. \tag{5}$$

Similarly, from (4), we have

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

which can be rewritten as follows

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})+n\overline{x}\,\overline{y}=n\,\alpha\,\overline{x}+\beta\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})+n\beta\,\overline{x}^2 \qquad (6)$$

Defining

$$S_{xy}:=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

we see that (6) reduces to

$$S_{xy}+n\overline{x}\,\overline{y}=\alpha\,n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right] \qquad (7)$$

Substituting (5) into (7), we have

$$S_{xy}+n\overline{x}\,\overline{y}=\left[\overline{y}-\beta\,\overline{x}\right]n\,\overline{x}+\beta\,\left[S_{xx}+n\overline{x}^2\right].$$

Simplifying the last equation, we get

$$S_{xy}=\beta\,S_{xx}$$

which is

$$\beta=\frac{S_{xy}}{S_{xx}}. \qquad (8)$$

In view of (8) and (5), we get

$$\alpha=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x}. \qquad (9)$$

Thus the least squares estimates of α and β are

$$\widehat{\alpha}=\overline{y}-\frac{S_{xy}}{S_{xx}}\,\overline{x} \quad \text{and} \quad \widehat{\beta}=\frac{S_{xy}}{S_{xx}},$$

respectively.

We need some notations. The random variable Y given $X=x$ will be denoted by Y_x. Note that this is the variable appears in the model $E(Y/x)=\alpha+\beta x$. When one chooses in succession values $x_1,x_2,...,x_n$ for x, a sequence $Y_{x_1},Y_{x_2},...,Y_{x_n}$ of random variable is obtained. For the sake of convenience, we denote the random variables $Y_{x_1},Y_{x_2},...,Y_{x_n}$ simply as $Y_1,Y_2,...,Y_n$. To do some statistical analysis, we make following three assumptions:

(1) $E(Y_x)=\alpha+\beta x$ so that $\mu_i=E(Y_i)=\alpha+\beta x_i$;

(2) $Y_1, Y_2, ..., Y_n$ are independent;

(3) Each of the random variables $Y_1, Y_2, ..., Y_n$ has the same variance σ^2.

Theorem 19.1. Under the above three assumptions, the least squares estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of a linear model $E(Y/x) = \alpha + \beta x$ are unbiased.

Proof: From the previous example, we know that the least squares estimators of α and β are

$$\widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\ \overline{X} \quad \text{and} \quad \widehat{\beta} = \frac{S_{xY}}{S_{xx}},$$

where

$$S_{xY} := \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}).$$

First, we show $\widehat{\beta}$ is unbiased. Consider

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\frac{S_{xY}}{S_{xx}}\right) = \frac{1}{S_{xx}}\ E\left(S_{xY}\right) \\
&= \frac{1}{S_{xx}}\ E\left(\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ E\left(Y_i - \overline{Y}\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ E\left(Y_i\right) - \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ E\left(\overline{Y}\right) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ E\left(Y_i\right) - \frac{1}{S_{xx}}\ E\left(\overline{Y}\right)\ \sum_{i=1}^{n} (x_i - \overline{x}) \\
&= \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ E\left(Y_i\right) = \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ \ (\alpha + \beta x_i) \\
&= \alpha\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x}) + \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ x_i \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ x_i \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ x_i - \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ \overline{x} \\
&= \beta\ \frac{1}{S_{xx}}\ \sum_{i=1}^{n} (x_i - \overline{x})\ (x_i - \overline{x}) \\
&= \beta\ \frac{1}{S_{xx}}\ S_{xx} = \beta.
\end{aligned}$$

Thus the estimator $\widehat{\beta}$ is unbiased estimator of the parameter β.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of α. Consider

$$\begin{aligned} E\left(\widehat{\alpha}\right) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\,\overline{x}\right) = E\left(\overline{Y}\right) - \overline{x}\;E\left(\frac{S_{xY}}{S_{xx}}\right) \\ &= E\left(\overline{Y}\right) - \overline{x}\,E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\;\beta \\ &= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(Y_i\right)\right) - \overline{x}\;\beta \\ &= \frac{1}{n}\;\left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right)\right) - \overline{x}\;\beta \\ &= \frac{1}{n}\;\left(n\alpha + \beta \sum_{i=1}^{n} x_i\right) - \overline{x}\;\beta \\ &= \alpha + \beta\,\overline{x} - \overline{x}\;\beta = \alpha \end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of α and the proof of the theorem is now complete.

19.2. The Normal Regression Analysis

In a regression analysis, we assume that the x_i's are constants while y_i's are values of the random variables Y_i's. A regression analysis is called a normal regression analysis if the conditional density of Y_i given $X_i = x_i$ is of the form

$$f(y_i/x_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{y_i-\alpha-\beta x_i}{\sigma}\right)^2},$$

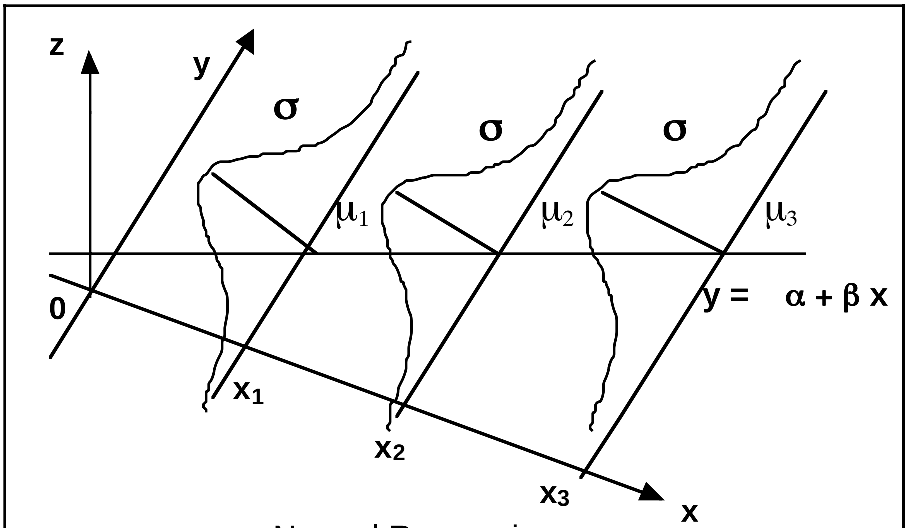
where σ^2 denotes the variance, and α and β are the regression coefficients. That is $Y|_{x_i} \sim N(\alpha + \beta x, \sigma^2)$. If there is no danger of confusion, then we will write Y_i for $Y|_{x_i}$. The following figure shows the regression model of Y populations with equal variances, and with means falling on the straight line $\mu_y = \alpha + \beta\, x$.

Normal regression analysis concerns with the estimation of σ, α, and β. We use maximum likelihood method to estimate these parameters. The maximum likelihood function of the sample is given by

$$L(\sigma, \alpha, \beta) = \prod_{i=1}^{n} f(y_i/x_i)$$

and

$$\ln L(\sigma, \alpha, \beta) = \sum_{i=1}^{n} \ln f(y_i/x_i)$$
$$= -n \ln \sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 .$$



Normal Regression

Taking the partial derivatives of $\ln L(\sigma, \alpha, \beta)$ with respect to α, β and σ respectively, we get

$$\frac{\partial}{\partial \alpha} \ln L(\sigma, \alpha, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)$$
$$\frac{\partial}{\partial \beta} \ln L(\sigma, \alpha, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)\, x_i$$
$$\frac{\partial}{\partial \sigma} \ln L(\sigma, \alpha, \beta) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 .$$

Equating each of these partial derivatives to zero and solving the system of three equations, we obtain the maximum likelihood estimator of β, α, σ as

$$\widehat{\beta} = \frac{S_{xY}}{S_{xx}}, \qquad \widehat{\alpha} = \overline{Y} - \frac{S_{xY}}{S_{xx}}\, \overline{x}, \quad \text{and} \quad \widehat{\sigma} = \sqrt{\frac{1}{n} \left[S_{YY} - \frac{S_{xY}}{S_{xx}}\, S_{xY} \right]},$$

where

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y} \right).$$

Theorem 19.2. In the normal regression analysis, the likelihood estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are unbiased estimators of β and α, respectively.

Proof: Recall that

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y} \right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) Y_i,
\end{aligned}$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. Thus $\widehat{\beta}$ is a linear combination of Y_i's. Since $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable.

First we show $\widehat{\beta}$ is an unbiased estimator of β. Since

$$\begin{aligned}
E\left(\widehat{\beta}\right) &= E\left(\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) Y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) E\left(Y_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) (\alpha + \beta x_i) = \beta,
\end{aligned}$$

the maximum likelihood estimator of β is unbiased.

Next, we show that $\widehat{\alpha}$ is also an unbiased estimator of α. Consider

$$\begin{aligned}
E\left(\widehat{\alpha}\right) &= E\left(\overline{Y} - \frac{S_{xY}}{S_{xx}}\, \overline{x} \right) = E\left(\overline{Y}\right) - \overline{x}\; E\left(\frac{S_{xY}}{S_{xx}} \right) \\
&= E\left(\overline{Y}\right) - \overline{x}\, E\left(\widehat{\beta}\right) = E\left(\overline{Y}\right) - \overline{x}\; \beta \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} E\left(Y_i\right) \right) - \overline{x}\; \beta \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} E\left(\alpha + \beta x_i\right) \right) - \overline{x}\; \beta \\
&= \frac{1}{n} \left(n\alpha + \beta \sum_{i=1}^{n} x_i \right) - \overline{x}\; \beta \\
&= \alpha + \beta\, \overline{x} - \overline{x}\; \beta = \alpha.
\end{aligned}$$

This proves that $\widehat{\alpha}$ is an unbiased estimator of α and the proof of the theorem is now complete.

Theorem 19.3. In normal regression analysis, the distributions of the estimators $\widehat{\beta}$ and $\widehat{\alpha}$ are given by

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 .$$

Proof: Since

$$\begin{aligned}
\widehat{\beta} &= \frac{S_{xY}}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \left(Y_i - \overline{Y}\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right) Y_i,
\end{aligned}$$

the $\widehat{\beta}$ is a linear combination of Y_i's. As $Y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$, we see that $\widehat{\beta}$ is also a normal random variable. By Theorem 19.2, $\widehat{\beta}$ is an unbiased estimator of β.

The variance of $\widehat{\beta}$ is given by

$$\begin{aligned}
Var\left(\widehat{\beta}\right) &= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 Var\left(Y_i/x_i\right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}}\right)^2 \sigma^2 \\
&= \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2\ \sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}$$

Hence $\widehat{\beta}$ is a normal random variable with mean (or expected value) β and variance $\frac{\sigma^2}{S_{xx}}$. That is $\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$.

Now determine the distribution of $\widehat{\alpha}$. Since each $Y_i \sim N(\alpha + \beta x_i,\, \sigma^2)$, the distribution of $\overline{Y}$ is given by

$$\overline{Y} \sim N\left(\alpha + \beta\overline{x},\ \frac{\sigma^2}{n}\right).$$

Since

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right)$$

the distribution of $\overline{x}\,\widehat{\beta}$ is given by

$$\overline{x}\,\widehat{\beta} \sim N\left(\overline{x}\,\beta,\ \overline{x}^2\,\frac{\sigma^2}{S_{xx}}\right).$$

Since $\widehat{\alpha} = \overline{Y} - \overline{x}\,\widehat{\beta}$ and $\overline{Y}$ and $\overline{x}\,\widehat{\beta}$ being two normal random variables, $\widehat{\alpha}$ is also a normal random variable with mean equal to $\alpha + \beta\,\overline{x} - \beta\,\overline{x} = \alpha$ and variance variance equal to $\frac{\sigma^2}{n} + \frac{\overline{x}^2\,\sigma^2}{S_{xx}}$. That is

$$\widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

and the proof of the theorem is now complete.

In the next theorem, we give an unbiased estimator of the variance σ^2. For this we need the distribution of the statistic U given by

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}.$$

It can be shown (we will omit the proof, for a proof see Graybill (1961)) that the distribution of the statistic

$$U = \frac{n\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2).$$

Theorem 19.4. An unbiased estimator S^2 of σ^2 is given by

$$S^2 = \frac{n\,\widehat{\sigma}^2}{n-2},$$

where $\widehat{\sigma} = \sqrt{\frac{1}{n}\,\left[S_{YY} - \frac{S_{xY}}{S_{xx}}\,S_{xY}\right]}$.

Proof: Since

$$\begin{aligned}
E(S^2) &= E\left(\frac{n\,\widehat{\sigma}^2}{n-2}\right) \\
&= \frac{\sigma^2}{n-2}\,E\left(\frac{n\,\widehat{\sigma}^2}{\sigma^2}\right) \\
&= \frac{\sigma^2}{n-2}\,E(\chi^2(n-2)) \\
&= \frac{\sigma^2}{n-2}\,(n-2) = \sigma^2.
\end{aligned}$$

The proof of the theorem is now complete.

Note that the estimator S^2 can be written as $S^2 = \frac{SSE}{n-2}$, where

$$SSE = S_{YY} = \widehat{\beta}\, S_{xY} = \sum_{i=1}^{2} [y_i - \widehat{\alpha} - \widehat{\beta}\, x_i]$$

the estimator S^2 is unbiased estimator of σ^2. The proof of the theorem is now complete.

In the next theorem we give the distribution of two statistics that can be used for testing hypothesis and constructing confidence interval for the regression parameters α and β.

Theorem 19.5. The statistics

$$Q_\beta = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}}$$

and

$$Q_\alpha = \frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n\, (\overline{x})^2 + S_{xx}}}$$

have both a t-distribution with $n-2$ degrees of freedom.

Proof: From Theorem 19.3, we know that

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right).$$

Hence by standardizing, we get

$$Z = \frac{\widehat{\beta} - \beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

Further, we know that the likelihood estimator of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n} \left[S_{YY} - \frac{S_{xY}}{S_{xx}}\, S_{xY} \right]}$$

and the distribution of the statistic $U = \frac{n\, \widehat{\sigma}^2}{\sigma^2}$ is chi-square with $n-2$ degrees of freedom.

Since $Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$ and $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$, by Theorem 14.6, the statistic $\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}} = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,S_{xx}}}} = \frac{\frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \sim t(n-2).$$

Similarly, it can be shown that

$$Q_\alpha = \frac{\widehat{\alpha}-\alpha}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n\,\left(\overline{x}\right)^2 + S_{xx}}} \sim t(n-2).$$

This completes the proof of the theorem.

In the normal regression model, if $\beta = 0$, then $E(Y_x) = \alpha$. This implies that $E(Y_x)$ does not depend on x. Therefore if $\beta \neq 0$, then $E(Y_x)$ is dependent on x. Thus the null hypothesis $H_o : \beta = 0$ should be tested against $H_a : \beta \neq 0$. To devise a test we need the distribution of $\widehat{\beta}$. Theorem 19.3 says that $\widehat{\beta}$ is normally distributed with mean β and variance $\frac{\sigma^2}{S_x x}$. Therefore, we have

$$Z = \frac{\widehat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1).$$

In practice the variance $Var(Y_i/x_i)$ which is σ^2 is usually unknown. Hence the above statistic Z is not very useful. However, using the statistic Q_β, we can devise a hypothesis test to test the hypothesis $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$ at a significance level α. For this one has to evaluate the quantity

$$|t| = \left| \frac{\widehat{\beta}-\beta}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,S_{xx}}}} \right|$$

$$= \left| \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}} \right|$$

and compare it to quantile $t_{\alpha/2}(n-2)$. The hypothesis test, at significance level α, is then "Reject $H_o : \beta = \beta_o$ if $|t| > t_{\alpha/2}(n-2)$".

The statistic

$$Q_\beta = \frac{\widehat{\beta}-\beta}{\widehat{\sigma}} \sqrt{\frac{(n-2)\,S_{xx}}{n}}$$

is a pivotal quantity for the parameter β since the distribution of this quantity Q_β is a t-distribution with $n-2$ degrees of freedom. Thus it can be used for the construction of a $(1-\gamma)100\%$ confidence interval for the parameter β as follows:

$$\begin{aligned} &1-\gamma \\ &= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq \frac{\widehat{\beta}-\beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} \leq t_{\frac{\gamma}{2}}(n-2)\right) \\ &= P\left(\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} \leq \beta \leq \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\right). \end{aligned}$$

Hence, the $(1-\gamma)\%$ confidence interval for β is given by

$$\left[\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}},\;\; \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\right].$$

In a similar manner one can devise hypothesis test for α and construct confidence interval for α using the statistic Q_α. We leave these to the reader.

Now we give two examples to illustrate how to find the normal regression line and related things.

Example 19.7. Let the following data on the number of hours, x which ten persons studied for a French test and their scores, y on the test is shown below:

x	4	9	10	14	4	7	12	22	1	17
y	31	58	65	73	37	44	60	91	21	84

Find the normal regression line that approximates the regression of test scores on the number of hours studied. Further test the hypothesis $H_o : \beta = 3$ versus $H_a : \beta \neq 3$ at the significance level 0.02.

Answer: From the above data, we have

$$\sum_{i=1}^{10} x_i = 100, \qquad \sum_{i=1}^{10} x_i^2 = 1376$$

$$\sum_{i=1}^{10} y_i = 564, \qquad \sum_{i=1}^{10} y_i^2 =$$

$$\sum_{i=1}^{10} x_i y_i = 6945$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

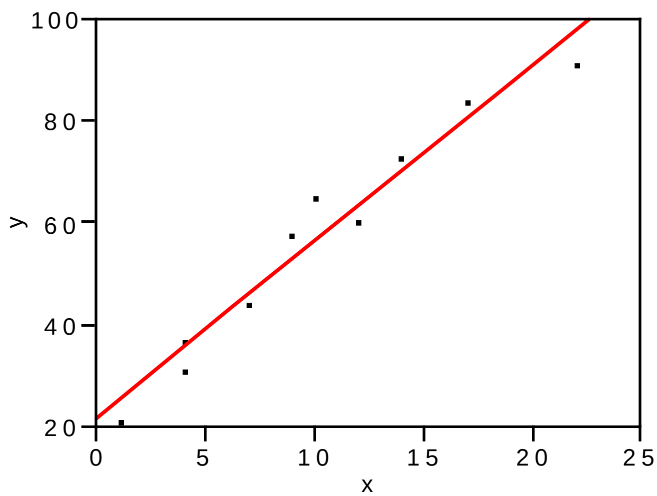
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 3.471 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 21.690.$$

Thus the normal regression line is

$$y = 21.690 + 3.471x.$$

This regression line is shown below.



Regression line y = 21.690 + 3.471 x

Now we test the hypothesis $H_o : \beta = 3$ against $H_a : \beta \neq 3$ at 0.02 level of significance. From the data, the maximum likelihood estimate of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\,S_{xy}\right]}$$

$$= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\,S_{xy}\right]}$$

$$= \sqrt{\frac{1}{10}\,[4752.4 - (3.471)(1305)]}$$

$$= 4.720$$

and

$$|t| = \left| \frac{3.471 - 3}{4.720} \sqrt{\frac{(8)\,(376)}{10}} \right| = 1.73.$$

Hence

$$1.73 = |t| < t_{0.01}(8) = 2.896.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 3$ at the significance level 0.02.

This means that we can not conclude that on the average an extra hour of study will increase the score by more than 3 points.

Example 19.8. The frequency of chirping of a cricket is thought to be related to temperature. This suggests the possibility that temperature can be estimated from the chirp frequency. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

Find the normal regression line that approximates the regression of temperature on the number chirps per second by the striped ground cricket. Further test the hypothesis $H_o : \beta = 4$ versus $H_a : \beta \neq 4$ at the significance level 0.1.

Answer: From the above data, we have

$$\sum_{i=1}^{10} x_i = 170, \qquad \sum_{i=1}^{10} x_i^2 = 2920$$

$$\sum_{i=1}^{10} y_i = 789, \qquad \sum_{i=1}^{10} y_i^2 = 64270$$

$$\sum_{i=1}^{10} x_i y_i = 13688$$

$$S_{xx} = 376, \qquad S_{xy} = 1305, \qquad S_{yy} = 4752.4.$$

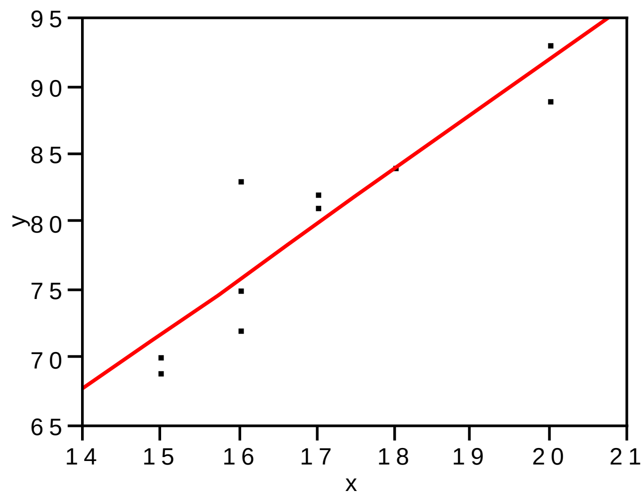
Hence

$$\widehat{\beta} = \frac{s_{xy}}{s_{xx}} = 4.067 \quad \text{and} \quad \widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x} = 9.761.$$

Thus the normal regression line is

$$y = 9.761 + 4.067x.$$

This regression line is shown below.



Regression line y = 9.761 + 4.067x

Now we test the hypothesis $H_o : \beta = 4$ against $H_a : \beta \neq 4$ at 0.1 level of significance. From the data, the maximum likelihood estimate of σ is

$$\begin{aligned}
\widehat{\sigma} &= \sqrt{\frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\,S_{xy}\right]} \\
&= \sqrt{\frac{1}{n}\left[S_{yy} - \widehat{\beta}\,S_{xy}\right]} \\
&= \sqrt{\frac{1}{10}\,[589 - (4.067)(122)]} \\
&= 3.047
\end{aligned}$$

and

$$|t| = \left|\frac{4.067 - 4}{3.047}\sqrt{\frac{(8)\,(30)}{10}}\right| = 0.528.$$

Hence

$$0.528 = |t| < t_{0.05}(8) = 1.860.$$

Thus we do not reject the null hypothesis that $H_o : \beta = 4$ at a significance level 0.1.

Let $\mu_x = \alpha + \beta\, x$ and write $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x$ for an arbitrary but fixed x. Then $\widehat{Y}_x$ is an estimator of μ_x. The following theorem gives various properties of this estimator.

Theorem 19.6. Let x be an arbitrary but fixed real number. Then

(i) $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$,

(ii) $\widehat{Y}_x$ is an unbiased estimator of μ_x, and

(iii) $Var\left(\widehat{Y}_x\right) = \left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right\}\sigma^2$.

Proof: First we show that $\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$. Since

$$\begin{aligned}
\widehat{Y}_x &= \widehat{\alpha} + \widehat{\beta}\, x \\
&= \overline{Y} - \widehat{\beta}\overline{x} + \widehat{\beta}\, x \\
&= \overline{Y} + \widehat{\beta}\,(x - \overline{x}) \\
&= \overline{Y} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \frac{Y_k}{n} + \sum_{k=1}^{n} \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\, Y_k \\
&= \sum_{k=1}^{n} \left(\frac{1}{n} + \frac{(x_k - \overline{x})\,(x - \overline{x})}{S_{xx}}\right) Y_k
\end{aligned}$$

$\widehat{Y}_x$ is a linear estimator of $Y_1, Y_2, ..., Y_n$.

Next, we show that $\widehat{Y}_x$ is an unbiased estimator of μ_x. Since

$$\begin{aligned}
E\left(\widehat{Y}_x\right) &= E\left(\widehat{\alpha} + \widehat{\beta}\, x\right) \\
&= E\left(\widehat{\alpha}\right) + E\left(\widehat{\beta}\, x\right) \\
&= \alpha + \beta\, x \\
&= \mu_x
\end{aligned}$$

$\widehat{Y}_x$ is an unbiased estimator of μ_x.

Finally, we calculate the variance of $\widehat{Y}_x$ using Theorem 19.3. The variance of $\widehat{Y}_x$ is given by

$$\begin{aligned} Var\left(\widehat{Y}_x\right) &= Var\left(\widehat{\alpha} + \widehat{\beta}\, x\right) \\ &= Var\left(\widehat{\alpha}\right) + Var\left(\widehat{\beta}\, x\right) + 2Cov\left(\widehat{\alpha},\, \widehat{\beta}\, x\right) \\ &= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) + x^2\,\frac{\sigma^2}{S_{xx}} + 2\,x\,Cov\left(\widehat{\alpha},\, \widehat{\beta}\right) \\ &= \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right) - 2\,x\,\frac{\overline{x}\,\sigma^2}{S_{xx}} \\ &= \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2. \end{aligned}$$

In this computation we have used the fact that

$$Cov\left(\widehat{\alpha},\, \widehat{\beta}\right) = -\frac{\overline{x}\,\sigma^2}{S_{xx}}$$

whose proof is left to the reader as an exercise. The proof of the theorem is now complete.

By Theorem 19.3, we see that

$$\widehat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{S_{xx}}\right) \qquad \text{and} \qquad \widehat{\alpha} \sim N\left(\alpha,\ \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right).$$

Since $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\, x$, the random variable $\widehat{Y}_x$ is also a normal random variable with mean μ_x and variance

$$Var\left(\widehat{Y}_x\right) = \left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2.$$

Hence standardizing $\widehat{Y}_x$, we have

$$\frac{\widehat{Y}_x - \mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}} \sim N(0,1).$$

If σ^2 is known, then one can take the statistic $Q = \frac{\widehat{Y}_x - \mu_x}{\sqrt{Var\left(\widehat{Y}_x\right)}}$ as a pivotal quantity to construct a confidence interval for μ_x. The $(1-\gamma)100\%$ confidence interval for μ_x when σ^2 is known is given by

$$\left[\widehat{Y}_x - z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)},\quad \widehat{Y}_x + z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)}\right].$$

Example 19.9. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

What is the 95% confidence interval for β? What is the 95% confidence interval for μ_x when $x = 14$ and $\sigma = 3.047$?

Answer: From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

The $(1-\gamma)\%$ confidence interval for β is given by

$$\left[\widehat{\beta} - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}}\,, \;\; \widehat{\beta} + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{n}{(n-2)\,S_{xx}}} \right].$$

Therefore the 90% confidence interval for β is

$$\left[4.067 - t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}}\,, \;\; 4.067 + t_{0.025}(8)\,(3.047)\sqrt{\frac{10}{(8)\,(376)}} \right]$$

which is

$$[\,4.067 - t_{0.025}(8)\,(0.1755)\,, \;\; 4.067 + t_{0.025}(8)\,(0.1755)]\,.$$

Since from the t-table, we have $t_{0.025}(8) = 2.306$, the 90% confidence interval for β becomes

$$[\,4.067 - (2.306)\,(0.1755)\,, \;\; 4.067 + (2.306)\,(0.1755)]$$

which is [3.6623, 4.4717].

If variance σ^2 is not known, then we can use the fact that the statistic $U = \frac{n\,\widehat{\sigma}^2}{\sigma^2}$ is chi-squares with $n-2$ degrees of freedom to obtain a pivotal quantity for μ_x. This can be done as follows:

$$\begin{aligned} Q &= \frac{\widehat{Y}_x - \mu_x}{\widehat{\sigma}}\sqrt{\frac{(n-2)\,S_{xx}}{S_{xx} + n\,(x-\overline{x})^2}} \\ &= \frac{\frac{\widehat{Y}_x - \mu_x}{\sqrt{\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right)\sigma^2}}}{\sqrt{\frac{n\,\widehat{\sigma}^2}{(n-2)\,\sigma^2}}} \sim t(n-2). \end{aligned}$$

Using this pivotal quantity one can construct a $(1-\gamma)100\%$ confidence interval for mean μ as

$$\left[\widehat{Y}_x - t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{S_{xx}+n(x-\overline{x})^2}{(n-2)\,S_{xx}}},\;\; \widehat{Y}_x + t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{S_{xx}+n(x-\overline{x})^2}{(n-2)\,S_{xx}}}\right].$$

Next we determine the 90% confidence interval for μ_x when $x = 14$ and $\sigma = 3.047$. The $(1-\gamma)100\%$ confidence interval for μ_x when σ^2 is known is given by

$$\left[\widehat{Y}_x - z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)},\;\; \widehat{Y}_x + z_{\frac{\gamma}{2}}\sqrt{Var(\widehat{Y}_x)}\right].$$

From the data, we have

$$\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}\,x = 9.761 + (4.067)\,(14) = 66.699$$

and

$$Var\left(\widehat{Y}_x\right) = \left(\frac{1}{10} + \frac{(14-17)^2}{376}\right)\sigma^2 = (0.124)\,(3.047)^2 = 1.1512.$$

The 90% confidence interval for μ_x is given by

$$\left[66.699 - z_{0.025}\,\sqrt{1.1512},\; 66.699 + z_{0.025}\,\sqrt{1.1512}\right]$$

and since $z_{0.025} = 1.96$ (from the normal table), we have

$$[66.699 - (1.96)\,(1.073),\; 66.699 + (1.96)\,(1.073)]$$

which is $[64.596,\; 68.802]$.

We now consider the predictions made by the normal regression equation $\widehat{Y}_x = \widehat{\alpha} + \widehat{\beta}x$. The quantity $\widehat{Y}_x$ gives an estimate of $\mu_x = \alpha + \beta x$. Each time we compute a regression line from a random sample we are observing one possible linear equation in a population consisting all possible linear equations. Further, the actual value of Y_x that will be observed for given value of x is normal with mean $\alpha + \beta x$ and variance σ^2. So the actual observed value will be different from μ_x. Thus, the predicted value for $\widehat{Y}_x$ will be in error from two different sources, namely (1) $\widehat{\alpha}$ and $\widehat{\beta}$ are randomly distributed about α and β, and (2) Y_x is randomly distributed about μ_x.

Let y_x denote the actual value of Y_x that will be observed for the value x and consider the random variable

$$\mathcal{D} = Y_x - \widehat{\alpha} - \widehat{\beta}\, x.$$

Since $\mathcal{D}$ is a linear combination of normal random variables, $\mathcal{D}$ is also a normal random variable.

The mean of $\mathcal{D}$ is given by

$$\begin{aligned} E(\mathcal{D}) &= E(Y_x) - E(\widehat{\alpha}) - x\, E(\widehat{\beta}) \\ &= \alpha + \beta\, x - \alpha - x\, \beta \\ &= 0. \end{aligned}$$

The variance of $\mathcal{D}$ is given by

$$\begin{aligned} Var(\mathcal{D}) &= Var(Y_x - \widehat{\alpha} - \widehat{\beta}\, x) \\ &= Var(Y_x) + Var(\widehat{\alpha}) + x^2\, Var(\widehat{\beta}) + 2\, x\, Cov(\widehat{\alpha}, \widehat{\beta}) \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\overline{x}^2\, \sigma^2}{S_{xx}} + x^2\, \frac{\sigma^2}{S_{xx}} - 2\, x\, \frac{\overline{x}}{S_{xx}} \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{(x - \overline{x})^2\, \sigma^2}{S_{xx}} \\ &= \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2. \end{aligned}$$

Therefore

$$\mathcal{D} \sim N\left(0,\ \frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2\right).$$

We standardize $\mathcal{D}$ to get

$$Z = \frac{\mathcal{D} - 0}{\sqrt{\frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2}} \sim N(0, 1).$$

Since in practice the variance of Y_x which is σ^2 is unknown, we can not use Z to construct a confidence interval for a predicted value y_x.

We know that $U = \frac{n\, \widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-2)$. By Theorem 14.6, the statistic

$\frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2)$. Hence

$$\begin{aligned} Q &= \frac{y_x - \widehat{\alpha} - \widehat{\beta}\, x}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{(n+1)\, S_{xx} + n}} \\ &= \frac{\frac{y_x - \widehat{\alpha} - \widehat{\beta}\, x}{\sqrt{\frac{(n+1)\, S_{xx} + n}{n\, S_{xx}}\, \sigma^2}}}{\sqrt{\frac{n\, \widehat{\sigma}^2}{(n-2)\, \sigma^2}}} \\ &= \frac{\frac{\mathcal{D} - 0}{\sqrt{Var(\mathcal{D})}}}{\sqrt{\frac{n\, \widehat{\sigma}^2}{(n-2)\, \sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{U}{n-2}}} \sim t(n-2). \end{aligned}$$

The statistic Q is a pivotal quantity for the predicted value y_x and one can use it to construct a $(1-\gamma)100\%$ confidence interval for y_x. The $(1-\gamma)100\%$ confidence interval, $[a, b]$, for y_x is given by

$$\begin{aligned} 1 - \gamma &= P\left(-t_{\frac{\gamma}{2}}(n-2) \leq Q \leq t_{\frac{\gamma}{2}}(n-2) \right) \\ &= P(a \leq y_x \leq b), \end{aligned}$$

where

$$a = \widehat{\alpha} + \widehat{\beta}\, x - t_{\frac{\gamma}{2}}(n-2)\, \widehat{\sigma} \sqrt{\frac{(n+1)\, S_{xx} + n}{(n-2)\, S_{xx}}}$$

and

$$b = \widehat{\alpha} + \widehat{\beta}\, x + t_{\frac{\gamma}{2}}(n-2)\, \widehat{\sigma} \sqrt{\frac{(n+1)\, S_{xx} + n}{(n-2)\, S_{xx}}}.$$

This confidence interval for y_x is usually known as the *prediction interval* for predicted value y_x based on the given x. The prediction interval represents an interval that has a probability equal to $1-\gamma$ of containing not a parameter but a future value y_x of the random variable Y_x. In many instances the prediction interval is more relevant to a scientist or engineer than the confidence interval on the mean μ_x.

Example 19.10. Let the following data on the number chirps per second, x by the striped ground cricket and the temperature, y in Fahrenheit is shown below:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

What is the 95% prediction interval for y_x when $x = 14$?

Answer: From Example 19.8, we have

$$n = 10, \quad \widehat{\beta} = 4.067, \quad \widehat{\alpha} = 9.761, \quad \widehat{\sigma} = 3.047 \quad \text{and} \quad S_{xx} = 376.$$

Thus the normal regression line is

$$y_x = 9.761 + 4.067x.$$

Since $x = 14$, the corresponding predicted value y_x is given by

$$y_x = 9.761 + (4.067)\,(14) = 66.699.$$

Therefore

$$\begin{aligned} a &= \widehat{\alpha} + \widehat{\beta}\,x - t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 - t_{0.025}(8)\,(3.047)\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 - (2.306)\,(3.047)\,(1.1740) \\ &= 58.4501. \end{aligned}$$

Similarly

$$\begin{aligned} b &= \widehat{\alpha} + \widehat{\beta}\,x + t_{\frac{\gamma}{2}}(n-2)\,\widehat{\sigma}\sqrt{\frac{(n+1)\,S_{xx} + n}{(n-2)\,S_{xx}}} \\ &= 66.699 + t_{0.025}(8)\,(3.047)\sqrt{\frac{(11)\,(376) + 10}{(8)\,(376)}} \\ &= 66.699 + (2.306)\,(3.047)\,(1.1740) \\ &= 74.9479. \end{aligned}$$

Hence the 95% prediction interval for y_x when $x = 14$ is [58.4501, 74.9479].

19.3. The Correlation Analysis

In the first two sections of this chapter, we examine the regression problem and have done an in-depth study of the least squares and the normal regression analysis. In the regression analysis, we assumed that the values of X are not random variables, but are fixed. However, the values of Y_x for a given value of x are randomly distributed about $E(Y_x) = \mu_x = \alpha + \beta x$. Further, letting ε to be a random variable with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$, one can model the so called regression problem by

$$Y_x = \alpha + \beta\, x + \varepsilon.$$

In this section, we examine the correlation problem. Unlike the regression problem, here both X and Y are random variables and the correlation problem can be modeled by

$$E(Y) = \alpha + \beta\, E(X).$$

From an experimental point of view this means that we are observing random vector (X, Y) drawn from some bivariate population.

Recall that if (X, Y) is a bivariate random variable then the correlation coefficient ρ is defined as

$$\rho = \frac{E\left((X - \mu_X)\,(Y - \mu_Y)\right)}{\sqrt{E\left((X - \mu_X)^2\right)\, E\left((Y - \mu_Y)^2\right)}}$$

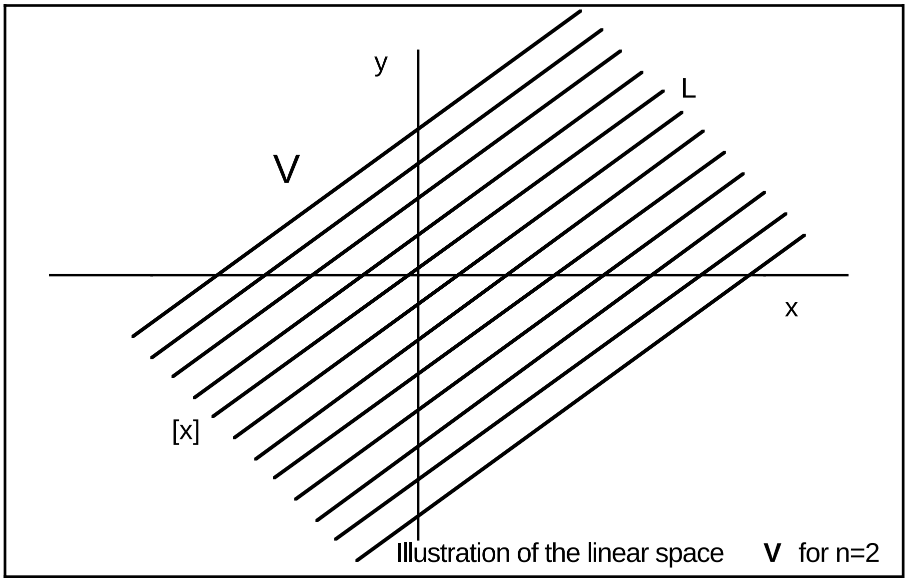
where μ_X and μ_Y are the mean of the random variables X and Y, respectively.

Definition 19.1. If $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is a random sample from a bivariate population, then the sample correlation coefficient is defined as

$$R = \frac{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2}\,\sqrt{\displaystyle\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

The corresponding quantity computed from data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ will be denoted by r and it is an estimate of the correlation coefficient ρ.

Now we give a geometrical interpretation of the sample correlation coefficient based on a paired data set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. We can associate this data set with two vectors $\vec{x} = (x_1, x_2, ..., x_n)$ and $\vec{y} = (y_1, y_2, ..., y_n)$ in $\mathbb{R}^n$. Let $\mathcal{L}$ be the subset $\{\lambda\, \vec{e} \,|\, \lambda \in \mathbb{R}\}$ of $\mathbb{R}^n$, where $\vec{e} = (1, 1, ..., 1) \in \mathbb{R}^n$. Consider the linear space V given by $\mathbb{R}^n$ modulo $\mathcal{L}$, that is $V = \mathbb{R}^n / \mathcal{L}$. The linear space V is illustrated in figure below when $n = 2$.



We denote the equivalence class associated with the vector $\vec{x}$ by $[\vec{x}]$. In the linear space V it can be shown that the points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ are collinear if and only if the the vectors $[\vec{x}]$ and $[\vec{y}]$ in V are proportional.

We define an inner product on this linear space V by

$$\langle [\vec{x}], [\vec{y}] \rangle = \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}).$$

Then the angle θ between the vectors $[\vec{x}]$ and $[\vec{y}]$ is given by

$$\cos(\theta) = \frac{\langle [\vec{x}], [\vec{y}] \rangle}{\sqrt{\langle [\vec{x}], [\vec{x}] \rangle}\,\sqrt{\langle [\vec{y}], [\vec{y}] \rangle}}$$

which is

$$\cos(\theta) = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}\,\sqrt{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})^2}} = r.$$

Thus the sample correlation coefficient r can be interpreted geometrically as the cosine of the angle between the vectors $[\vec{x}]$ and $[\vec{y}]$. From this view point the following theorem is obvious.

Theorem 19.7. The sample correlation coefficient r satisfies the inequality

$$-1 \leq r \leq 1.$$

The sample correlation coefficient $r = \pm 1$ if and only if the set of points $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ for $n \geq 3$ are collinear.

To do some statistical analysis, we assume that the paired data is a random sample of size n from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Then the conditional distribution of the random variable Y given $X = x$ is normal, that is

$$Y|_x \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1),\, \sigma_2^2(1 - \rho^2)\right).$$

This can be viewed as a normal regression model $E(Y|_x) = \alpha + \beta\, x$ where $\alpha = \mu - \rho\frac{\sigma_2}{\sigma_1}\mu_1$, $\beta = \rho\frac{\sigma_2}{\sigma_1}$, and $Var(Y|_x) = \sigma_2^2(1 - \rho^2)$.

Since $\beta = \rho\frac{\sigma_2}{\sigma_1}$, if $\rho = 0$, then $\beta = 0$. Hence the null hypothesis $H_o : \rho = 0$ is equivalent to $H_o : \beta = 0$. In the previous section, we devised a hypothesis test for testing $H_o : \beta = \beta_o$ against $H_a : \beta \neq \beta_o$. This hypothesis test, at significance level γ, is "Reject $H_o : \beta = \beta_o$ if $|t| \geq t_{\frac{\gamma}{2}}(n - 2)$", where

$$t = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}.$$

If $\beta = 0$, then we have

$$t = \frac{\widehat{\beta}}{\widehat{\sigma}}\sqrt{\frac{(n-2)\, S_{xx}}{n}}. \tag{10}$$

Now we express t in term of the sample correlation coefficient r. Recall that

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}, \tag{11}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right], \tag{12}$$

and

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}. \tag{13}$$

Now using (11), (12), and (13), we compute

$$
\begin{aligned}
t &= \frac{\widehat{\beta}}{\widehat{\sigma}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\
&= \frac{S_{xy}}{S_{xx}} \frac{\sqrt{n}}{\sqrt{\left[S_{yy} - \frac{S_{xy}}{S_{xx}}\, S_{xy}\right]}} \sqrt{\frac{(n-2)\, S_{xx}}{n}} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}} \frac{1}{\sqrt{\left[1 - \frac{S_{xy}}{S_{xx}} \frac{S_{xy}}{S_{yy}}\right]}} \sqrt{n-2} \\
&= \sqrt{n-2}\, \frac{r}{\sqrt{1-r^2}}.
\end{aligned}
$$

Hence to test the null hypothesis $H_o : \rho = 0$ against $H_a : \rho \neq 0$, at significance level γ, is "Reject $H_o : \rho = 0$ if $|t| \geq t_{\frac{\gamma}{2}}(n-2)$", where $t = \sqrt{n-2}\, \frac{r}{1-r^2}$.

This above test does not extend to test other values of ρ except $\rho = 0$. However, tests for the nonzero values of ρ can be achieved by the following result.

Theorem 19.8. Let $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ be a random sample from a bivariate normal population $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. If

$$
V = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right) \quad \text{and} \quad m = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right),
$$

then

$$
Z = \sqrt{n-3}\,(V - m) \to N(0,1) \quad \text{as } n \to \infty.
$$

This theorem says that the statistic V is approximately normal with mean m and variance $\frac{1}{n-3}$ when n is large. This statistic can be used to devise a hypothesis test for the nonzero values of ρ. Hence to test the null hypothesis $H_o : \rho = \rho_o$ against $H_a : \rho \neq \rho_o$, at significance level γ, is "Reject $H_o : \rho = \rho_o$ if $|z| \geq z_{\frac{\gamma}{2}}$", where $z = \sqrt{n-3}\,(V - m_o)$ and $m_o = \frac{1}{2} \ln\left(\frac{1+\rho_o}{1-\rho_o}\right)$.

Example 19.11. The following data were obtained in a study of the relationship between the weight and chest size of infants at birth:

x, weight in kg	2.76	2.17	5.53	4.31	2.30	3.70
y, chest size in cm	29.5	26.3	36.6	27.8	28.3	28.6

Determine the sample correlation coefficient r and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

Answer: From the above data we find that

$$\overline{x} = 3.46 \qquad \text{and} \qquad \overline{y} = 29.51.$$

Next, we compute S_{xx}, S_{yy} and S_{xy} using a tabular representation.

$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$	$(x - \overline{x})^2$	$(y - \overline{y})^2$
-0.70	-0.01	0.007	0.490	0.000
-1.29	-3.21	4.141	1.664	10.304
2.07	7.09	14.676	4.285	50.268
0.85	-1.71	-1.453	0.722	2.924
-1.16	-1.21	1.404	1.346	1.464
0.24	-0.91	-0.218	0.058	0.828
		$S_{xy} = 18.557$	$S_{xx} = 8.565$	$S_{yy} = 65.788$

Hence, the correlation coefficient r is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{18.557}{\sqrt{(8.565)\,(65.788)}} = 0.782.$$

The computed t value is give by

$$t = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} = \sqrt{(6-2)}\,\frac{0.782}{\sqrt{1-(0.782)^2}} = 2.509.$$

From the t-table we have $t_{0.005}(4) = 4.604$. Since

$$2.509 = |t| \not\geq t_{0.005}(4) = 4.604$$

we do not reject the null hypothesis $H_o : \rho = 0$.

19.4. Review Exercises

1. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimators of $\widehat{\beta}$ and $\widehat{\sigma}^2$ of β and σ^2.

2. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then show that the maximum likelihood estimator of $\widehat{\beta}$ is normally distributed. What are the mean and variance of $\widehat{\beta}$?

3. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find an unbiased estimator $\widehat{\sigma}^2$ of σ^2 and then find a constant k such that $k\,\widehat{\sigma}^2 \sim \chi^2(2n)$.

4. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for β and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for β.

5. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim N(\beta x_i, \sigma^2)$, where both β and σ^2 are unknown parameters. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find a pivotal quantity for σ^2 and using this pivotal quantity construct a $(1-\gamma)100\%$ confidence interval for σ^2.

6. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of β.

7. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim EXP(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of β.

8. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the maximum likelihood estimator of $\widehat{\beta}$ of β.

9. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, then find the least squares estimator of $\widehat{\beta}$ of β.

10. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, show that the least squares estimator and the maximum likelihood estimator of β are both unbiased estimator of β.

11. Let $Y_1, Y_2, ..., Y_n$ be n independent random variables such that each $Y_i \sim POI(\beta x_i)$, where β is an unknown parameter. If $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ is a data set where $y_1, y_2, ..., y_n$ are the observed values based on $x_1, x_2, ..., x_n$, the find the variances of both the least squares estimator and the maximum likelihood estimator of β.

12. Given the five pairs of points (x, y) shown below:

x	10	20	30	40	50
y	50.071	0.078	0.112	0.120	0.131

What is the curve of the form $y = a + bx + cx^2$ best fits the data by method of least squares?

13. Given the five pairs of points (x, y) shown below:

x	4	7	9	10	11
y	10	16	22	20	25

What is the curve of the form $y = a + b\,x$ best fits the data by method of least squares?

14. The following data were obtained from the grades of six students selected at random:

Mathematics Grade, x	72	94	82	74	65	85
English Grade, y	76	86	65	89	80	92

Find the sample correlation coefficient r and then test the null hypothesis $H_o : \rho = 0$ against the alternative hypothesis $H_a : \rho \neq 0$ at a significance level 0.01.

Chapter 20

ANALYSIS OF VARIANCE

In Chapter 19, we examine how a quantitative independent variable x can be used for predicting the value of a quantitative dependent variable y. In this chapter we would like to examine whether one or more independent (or predictor) variable affects a dependent (or response) variable y. This chapter differs from the last chapter because the independent variable may now be either quantitative or qualitative. It also differs from the last chapter in assuming that the response measurements were obtained for specific settings of the independent variables. Selecting the settings of the independent variables is another aspect of experimental design. It enable us to tell whether changes in the independent variables cause changes in the mean response and it permits us to analyze the data using a method known as analysis of variance (or ANOVA). Sir Ronald Aylmer Fisher (1890-1962) developed the analysis of variance in 1920's and used it to analyze data from agricultural experiments.

The ANOVA investigates independent measurements from several treatments or levels of one or more than one factors (that is, the predictor variables). The technique of ANOVA consists of partitioning the total sum of squares into component sum of squares due to different factors and the error. For instance, suppose there are Q factors. Then the total sum of squares (SS_T) is partitioned as

$$SS_T = SS_A + SS_B + \cdots + SS_Q + SS_{Error},$$

where SS_A, SS_B, ..., and SS_Q represent the sum of squares associated with the factors A, B, ..., and Q, respectively. If the ANOVA involves only one factor, then it is called one-way analysis of variance. Similarly if it involves two factors, then it is called the two-way analysis of variance. If it involves more then two factors, then the corresponding ANOVA is called the higher order analysis of variance. In this chapter we only treat the one-way analysis of variance.

The analysis of variance is a special case of the linear models that represent the relationship between a continuous response variable y and one or more predictor variables (either continuous or categorical) in the form

$$y = X\,\beta + \epsilon \tag{1}$$

where y is an $m \times 1$ vector of observations of response variable, X is the $m \times n$ design matrix determined by the predictor variables, β is $n \times 1$ vector of parameters, and ϵ is an $m \times 1$ vector of random error (or disturbances) independent of each other and having distribution.

20.1. One-Way Analysis of Variance with Equal Sample Sizes

The standard model of one-way ANOVA is given by

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n, \tag{2}$$

where $m \geq 2$ and $n \geq 2$. In this model, we assume that each random variable

$$Y_{ij} \sim N(\mu_i, \sigma^2) \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n. \tag{3}$$

Note that because of (3), each ϵ_{ij} in model (2) is normally distributed with mean zero and variance σ^2.

Given m independent samples, each of size n, where the members of the i^{th} sample, $Y_{i1}, Y_{i2}, ..., Y_{in}$, are normal random variables with mean μ_i and unknown variance σ^2. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n.$$

We will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a} : \text{not all the means are equal.}$$

In the following theorem we present the maximum likelihood estimators of the parameters $\mu_1, \mu_1, ..., \mu_m$ and σ^2.

Theorem 20.1. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Then the MLE's of the parameters μ_i $(i = 1, 2, ..., m)$ and σ^2 of the model are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{nm} \mathrm{SS_W},$$

where $\overline{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}$ and $\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ is the within samples sum of squares.

Proof: The likelihood function is given by

$$L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}}\right\}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu_i)^2}.$$

Taking the natural logarithm of the likelihood function L, we obtain

$$\ln L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = -\frac{nm}{2}\ln(2\,\pi\,\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \mu_i)^2. \qquad (4)$$

Now taking the partial derivative of (4) with respect to μ_1, $\mu_2, ..., \mu_m$ and σ^2, we get

$$\frac{\partial lnL}{\partial \mu_i} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(Y_{ij} - \mu_i) \qquad (5)$$

and

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{nm}{2\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \mu_i)^2. \qquad (6)$$

Equating these partial derivatives to zero and solving for μ_i and σ^2, respectively, we have

$$\mu_i = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\sigma^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2,$$

where

$$\overline{Y}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} Y_{ij}.$$

It can be checked that these solutions yield the maximum of the likelihood function and we leave this verification to the reader. Thus the maximum likelihood estimators of the model parameters are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{nm}\mathrm{SS_W},$$

where $\mathrm{SS_W} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The proof of the theorem is now complete.

Define

$$\overline{Y}_{\bullet\bullet} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{ij}. \tag{7}$$

Further, define

$$\mathrm{SS_T} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{8}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \tag{9}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \tag{10}$$

Here $\mathrm{SS_T}$ is the total sum of square, $\mathrm{SS_W}$ is the within sum of square, and $\mathrm{SS_B}$ is the between sum of square.

Next we consider the partitioning of the total sum of squares. The following lemma gives us such a partition.

Lemma 20.1. The total sum of squares is equal to the sum of within and between sum of squares, that is

$$\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}. \tag{11}$$

Proof: Rewriting (8) we have

$$
\begin{aligned}
\mathrm{SS_T} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[(Y_{ij} - \overline{Y}_{i\bullet}) + (Y_{i\bullet} - \overline{Y}_{\bullet\bullet})\right]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \\
&\qquad\qquad + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) \\
&= \mathrm{SS_W} + \mathrm{SS_B} + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}).
\end{aligned}
$$

The cross-product term vanishes, that is

$$\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})\,(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}) = \sum_{i=1}^{m}(Y_{i\bullet} - Y_{\bullet\bullet})\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet}) = 0.$$

Hence we obtain the asserted result $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$ and the proof of the lemma is complete.

The following theorem is a technical result and is needed for testing the null hypothesis against the alternative hypothesis.

Theorem 20.2. Consider the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where $Y_{ij} \sim N\left(\nu_i, \sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\left(m(n-1)\right)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\, m(n-1)}{\mathrm{SS_W}(m-1)} \sim F(m-1, m(n-1))$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(nm-1)$.

Proof: In Chapter 13, we have seen in Theorem 13.7 that if $X_1, X_2, ..., X_n$ are independent random variables each one having the distribution $N(\mu, \sigma^2)$, then their mean $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ have the following properties:

(i) $\overline{X}$ and $\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independent, and

(ii) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$.

Now using (i) and (ii), we establish this theorem.

(a) Using (ii), we see that

$$\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(n-1)$$

for each $i = 1, 2, ..., m$. Since

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{j=1}^{n}\left(Y_{i'j} - \overline{Y}_{i'\bullet}\right)^2$$

are independent for $i' \neq i$, we obtain

$$\sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).$$

Hence

$$\begin{aligned}\frac{SS_W}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \\ &= \sum_{i=1}^{m}\frac{1}{\sigma^2}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \sim \chi^2(m(n-1)).\end{aligned}$$

(b) Since for each $i = 1, 2, ..., m$, the random variables $Y_{i1}, Y_{i2}, ..., Y_{in}$ are independent and

$$Y_{i1}, Y_{i2}, ..., Y_{in} \sim N\left(\mu_i, \sigma^2\right)$$

we conclude by (i) that

$$\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i\bullet}$$

are independent. Further

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \overline{Y}_{i'\bullet}$$

are independent for $i' \neq i$. Therefore, each of the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

is independent of the statistics $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$, and the statistics

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m$$

are independent. Thus it follows that the sets

$$\sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad i = 1, 2, ..., m \qquad \text{and} \qquad \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m$$

are independent. Thus

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2 \qquad \text{and} \qquad \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2$$

are independent. Hence by definition, the statistics SS_W and SS_B are independent.

Suppose the null hypothesis $H_o : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true.

(c) Under H_o, the random variables $\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet}, ..., \overline{Y}_{m\bullet}$ are independent and identically distributed with $N\left(\mu, \frac{\sigma^2}{n}\right)$. Therefore by (ii)

$$\frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1).$$

Therefore

$$\begin{aligned} \frac{SS_B}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \\ &= \frac{n}{\sigma^2} \sum_{i=1}^{m} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(m-1). \end{aligned}$$

(d) Since

$$\frac{SS_W}{\sigma^2} \sim \chi^2\left(m(n-1)\right)$$

and

$$\frac{SS_B}{\sigma^2} \sim \chi^2(m-1)$$

therefore

$$\frac{\frac{SS_B}{(m-1)\,\sigma^2}}{\frac{SS_W}{(n(m-1)\,\sigma^2}} \sim F(m-1, m(n-1)).$$

That is

$$\frac{\frac{SS_B}{(m-1)}}{\frac{SS_W}{(n(m-1)}} \sim F(m-1, m(n-1)).$$

(e) Under H_o, the random variables Y_{ij}, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$ are independent and each has the distribution $N(\mu, \sigma^2)$. By (ii) we see that

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \sim \chi^2(nm-1).$$

Hence we have

$$\frac{SS_T}{\sigma^2} \sim \chi^2(nm-1)$$

and the proof of the theorem is now complete.

From Theorem 20.1, we see that the maximum likelihood estimator of each μ_i $(i = 1, 2, ..., m)$ is given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet},$$

and since $\overline{Y}_{i\bullet} \sim N\left(\mu_i, \frac{\sigma^2}{n}\right)$,

$$E\left(\widehat{\mu_i}\right) = E\left(\overline{Y}_{i\bullet}\right) = \mu_i.$$

Thus the maximum likelihood estimators are unbiased estimator of μ_i for $i = 1, 2, ..., m$.

Since

$$\widehat{\sigma^2} = \frac{SS_W}{mn}$$

and by Theorem 2, $\frac{1}{\sigma^2}SS_W \sim \chi^2(m(n-1))$, we have

$$E\left(\widehat{\sigma^2}\right) = E\left(\frac{SS_W}{mn}\right) = \frac{1}{mn}\,\sigma^2\,E\left(\frac{1}{\sigma^2}SS_W\right) = \frac{1}{mn}\,\sigma^2\,m(n-1) \neq \sigma^2.$$

Thus the maximum likelihood estimator $\widehat{\sigma^2}$ of σ^2 is biased. However, the estimator $\frac{SS_W}{m(n-1)}$ is an unbiased estimator. Similarly, the estimator $\frac{SS_T}{mn-1}$ is an unbiased estimator where as $\frac{SS_T}{mn}$ is a biased estimator of σ^2.

Theorem 20.3. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. The null hypothesis $H_o : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/(m(n-1))} > F_\alpha(m-1,\, m(n-1)), \tag{12}$$

where α is the significance level of the hypothesis test and $F_\alpha(m-1,\, m(n-1))$ denotes the $100(1-\alpha)$ percentile of the F-distribution with $m-1$ numerator and $nm-m$ denominator degrees of freedom.

Proof: Under the null hypothesis $H_o : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$, the likelihood function takes the form

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(Y_{ij}-\mu)^2}{2\sigma^2}}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{nm} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu)^2}. \end{aligned}$$

Taking the natural logarithm of the likelihood function and then maximizing it, we obtain

$$\widehat{\mu} = \overline{Y}_{\bullet\bullet} \qquad \text{and} \qquad \widehat{\sigma_{H_o}} = \frac{1}{mn} SS_T$$

as the maximum likelihood estimators of μ and σ^2, respectively. Inserting these estimators into the likelihood function, we have the maximum of the likelihood function, that is

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{H_o}}}}\right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2_{H_o}}}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\widehat{\sigma^2_{H_o}}}}\right)^{nm} e^{-\frac{mn}{2\,SS_T}\,SS_T}$$

which is

$$\max L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2_{\mathrm{H_o}}}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{13}$$

When no restrictions imposed, we get the maximum of the likelihood function from Theorem 20.1 as

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{1}{2\widehat{\sigma^2}} \sum_{i=1}^{m}\sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\bullet})^2}.$$

Simplifying the above expression, we see that

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2\,\mathrm{SS_W}}\,\mathrm{SS_W}}$$

which is

$$\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi \widehat{\sigma^2}}} \right)^{nm} e^{-\frac{mn}{2}}. \tag{14}$$

Next we find the likelihood ratio statistic W for testing the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$. Recall that the likelihood ratio statistic W can be found by evaluating

$$W = \frac{\max L(\mu, \sigma^2)}{\max L(\mu_1, \mu_2, ..., \mu_m, \sigma^2)}.$$

Using (13) and (14), we see that

$$W = \left(\frac{\widehat{\sigma^2}}{\widehat{\sigma^2_{\mathrm{H_o}}}} \right)^{\frac{mn}{2}}. \tag{15}$$

Hence the likelihood ratio test to reject the null hypothesis $\mathrm{H_o}$ is given by the inequality

$$W < k_0$$

where k_0 is a constant. Using (15) and simplifying, we get

$$\frac{\widehat{\sigma^2_{\mathrm{H_o}}}}{\widehat{\sigma^2}} > k_1$$

where $k_1 = \left(\frac{1}{k_0}\right)^{\frac{2}{mn}}$. Hence

$$\frac{SS_T/mn}{SS_W/mn} = \frac{\widehat{\sigma^2_{H_o}}}{\widehat{\sigma^2}} > k_1.$$

Using Lemma 20.1 we have

$$\frac{SS_W + SS_B}{SS_W} > k_1.$$

Therefore

$$\frac{SS_B}{SS_W} > k \tag{16}$$

where $k = k_1 - 1$. In order to find the cutoff point k in (16), we use Theorem 20.2 (d). Therefore

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/(m(n-1))} > \frac{m(n-1)}{m-1}k$$

Since $\mathcal{F}$ has F distribution, we obtain

$$\frac{m(n-1)}{m-1}k = F_\alpha(m-1,\, m(n-1)).$$

Thus, at a significance level α, reject the null hypothesis H_o if

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/(m(n-1))} > F_\alpha(m-1,\, m(n-1))$$

and the proof of the theorem is complete.

The various quantities used in carrying out the test described in Theorem 20.3 are presented in a tabular form known as the ANOVA table.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	SS_B	$m-1$	$MS_B = \frac{SS_B}{m-1}$	$\mathcal{F} = \frac{MS_B}{MS_W}$
Within	SS_W	$m(n-1)$	$MS_W = \frac{SS_W}{m(n-1)}$	
Total	SS_T	$mn-1$		

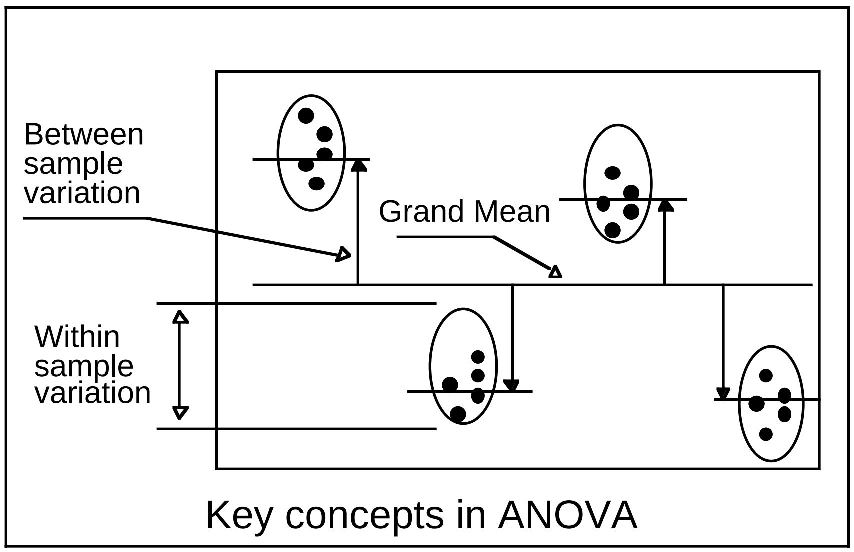
Table 20.1. One-Way ANOVA Table

At a significance level α, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $\mathcal{F} > F_\alpha(m-1,\, m(n-1))$." One can also use the notion of $p-$value to perform this hypothesis test. If the value of the test statistics is $\mathcal{F} = \gamma$, then the p-value is defined as

$$p - \text{value} = P(F(m-1, m(n-1)) \geq \gamma).$$

Alternatively, at a significance level α, the likelihood ratio test is: "Reject the null hypothesis $\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ if $p - \text{value} < \alpha$."

The following figure illustrates the notions of between sample variation and within sample variation.



The ANOVA model described in (2), that is

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

can be rewritten as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where μ is the mean of the m values of μ_i, and $\sum_{i=1}^{m} \alpha_i = 0$. The quantity α_i is called the effect of the i^{th} treatment. Thus any observed value is the sum of an overall mean μ, a treatment or class deviation α_i, and a random element from a normally distributed random variable ϵ_{ij} with mean zero and variance σ^2. This model is called model I, the fixed effects model. The effects of the treatments or classes, measured by the parameters α_i, are regarded as fixed but unknown quantities to be estimated. In this fixed effect model the null hypothesis H_0 is now

$$\mathrm{H_o} : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

and the alternative hypothesis is

$$\mathrm{H_a} : \text{not all the } \alpha_{\mathrm{i}} \text{ are zero.}$$

The random effects model, also known as model II, is given by

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \qquad \text{for } \; i = 1, 2, ..., m, \quad j = 1, 2, ..., n,$$

where μ is the overall mean and

$$A_i \sim N(0, \sigma_A^2) \qquad \text{and} \qquad \epsilon_{ij} \sim N(0, \sigma^2).$$

In this model, the variances σ_A^2 and σ^2 are unknown quantities to be estimated. The null hypothesis of the random effect model is $H_o : \sigma_A^2 = 0$ and the alternative hypothesis is $H_a : \sigma_A^2 > 0$. In this chapter we do not consider the random effect model.

Before we present some examples, we point out the assumptions on which the ANOVA is based on. The ANOVA is based on the following three assumptions:

(1) *Independent Samples:* The samples taken from the population under consideration should be independent of one another.

(2) *Normal Population:* For each population, the variable under consideration should be normally distributed.

(3) *Equal Variance:* The variances of the variables under consideration should be the same for all the populations.

Example 20.1. The data in the following table gives the number of hours of relief provided by 5 different brands of headache tablets administered to 25 subjects experiencing fevers of 38^oC or more. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is same for all 5 brands.

Tablets				
A	B	C	D	F
5	9	3	2	7
4	7	5	3	6
8	8	2	4	9
6	6	3	1	4
3	9	7	4	7

Answer: Using the formulas (8), (9) and (10), we compute the sum of squares SS_W, SS_B and SS_T as

$$SS_W = 57.60, \qquad SS_B = 79.94, \qquad \text{and} \qquad SS_T = 137.04.$$

The ANOVA table for this problem is shown below.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	79.94	4	19.86	6.90
Within	57.60	20	2.88	
Total	137.04	24		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(4,\,20) = 2.8661$. Since

$$6.90 = \mathcal{F} > F_{0.05}(4,\,20) = 2.8661$$

we reject the null hypothesis that the mean number of hours of relief provided by the tablets is same for all 5 brands.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the p-value to be

$$p - \text{value} = P(F(4,20) \geq 6.90) = 0.001.$$

Hence again we reach the same conclusion since p-value is less then the given α for this problem.

Example 20.2. Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following two data sets.

Data Set 1			Data Set 2		
Sample			Sample		
A	B	C	A	B	C
8.1	8.0	14.8	9.2	9.5	9.4
4.2	15.1	5.3	9.1	9.5	9.3
14.7	4.7	11.1	9.2	9.5	9.3
9.9	10.4	7.9	9.2	9.6	9.3
12.1	9.0	9.3	9.3	9.5	9.2
6.2	9.8	7.4	9.2	9.4	9.3

Answer: Computing the sum of squares SS_W, SS_B and SS_T, we have the following two ANOVA tables:

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.3	2	0.1	0.01
Within	187.2	15	12.5	
Total	187.5	17		

and

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find from the F-table that $F_{0.05}(2,\, 15) = 3.68$. For the first data set, since

$$0.01 = \mathcal{F} < F_{0.05}(2,\, 15) = 3.68$$

we do not reject the null hypothesis whereas for the second data set,

$$35.0 = \mathcal{F} > F_{0.05}(2,\, 15) = 3.68$$

we reject the null hypothesis.

Remark 20.1. Note that the sample means are same in both the data sets. However, there is a less variation among the sample points in samples of the second data set. The ANOVA finds a more significant differences among the means in the second data set. This example suggests that the larger the variation among sample means compared with the variation of the measurements within samples, the greater is the evidence to indicate a difference among population means.

20.2. One-Way Analysis of Variance with Unequal Sample Sizes

In the previous section, we examined the theory of ANOVA when samples are same sizes. When the samples are same sizes we say that the ANOVA is in the balanced case. In this section we examine the theory of ANOVA for unbalanced case, that is when the samples are of different sizes. In experimental work, one often encounters unbalance case due to the death of experimental animals in a study or drop out of the human subjects from a study or due to damage of experimental materials used in a study. Our analysis of the last section for the equal sample size will be valid but have to be modified to accommodate the different sample size.

Consider m independent samples of respective sizes $n_1, n_2, ..., n_m$, where the members of the i^{th} sample, $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, are normal random variables with mean μ_i and unknown variance σ^2. That is,

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right), \qquad i = 1, 2, ..., m, \qquad j = 1, 2, ..., n_i.$$

Let us denote $N = n_1 + n_2 + \cdots n_m$. Again, we will be interested in testing the null hypothesis

$$\mathrm{H_o} : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

against the alternative hypothesis

$$\mathrm{H_a} : \text{not all the means are equal.}$$

Now we defining

$$\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n} Y_{ij}, \tag{17}$$

$$\overline{Y}_{\bullet\bullet} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}, \tag{18}$$

$$\mathrm{SS_T} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\bullet\bullet} \right)^2, \tag{19}$$

$$\mathrm{SS_W} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2, \tag{20}$$

and

$$\mathrm{SS_B} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet} \right)^2 \tag{21}$$

we have the following results analogous to the results in the previous section.

Theorem 20.4. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. Then the MLE's of the parameters μ_i $(i = 1, 2, ..., m)$ and σ^2 of the model are given by

$$\widehat{\mu_i} = \overline{Y}_{i\bullet} \qquad i = 1, 2, ..., m,$$
$$\widehat{\sigma^2} = \frac{1}{N} \mathrm{SS_W},$$

where $\overline{Y}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ and $\mathrm{SS_W} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\bullet} \right)^2$ is the within samples sum of squares.

Lemma 20.2. The total sum of squares is equal to the sum of within and between sum of squares, that is $\mathrm{SS_T} = \mathrm{SS_W} + \mathrm{SS_B}$.

Theorem 20.5. Consider the ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad i = 1, 2, ..., m, \quad j = 1, 2, ..., n_i,$$

where $Y_{ij} \sim N\left(\nu_i, \sigma^2\right)$. Then

(a) the random variable $\frac{\mathrm{SS_W}}{\sigma^2} \sim \chi^2\,(N-m)$, and

(b) the statistics $\mathrm{SS_W}$ and $\mathrm{SS_B}$ are independent.

Further, if the null hypothesis $\mathrm{H_o}: \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is true, then

(c) the random variable $\frac{\mathrm{SS_B}}{\sigma^2} \sim \chi^2(m-1)$,

(d) the statistics $\frac{\mathrm{SS_B}\, m(n-1)}{\mathrm{SS_W}\,(m-1)} \sim F(m-1, N-m)$, and

(e) the random variable $\frac{\mathrm{SS_T}}{\sigma^2} \sim \chi^2(N-1)$.

Theorem 20.6. Suppose the one-way ANOVA model is given by the equation (2) where the ϵ_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n_i$. The null hypothesis $\mathrm{H_o}: \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ is rejected whenever the test statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{\mathrm{SS_B}/(m-1)}{\mathrm{SS_W}/(N-m)} > F_\alpha(m-1,\ N-m),$$

where α is the significance level of the hypothesis test and $F_\alpha(m-1,\ N-m)$ denotes the $100(1-\alpha)$ percentile of the F-distribution with $m-1$ numerator and $N-m$ denominator degrees of freedom.

The corresponding ANOVA table for this case is

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	$\mathrm{SS_B}$	$m-1$	$\mathrm{MS_B} = \frac{\mathrm{SS_B}}{m-1}$	$\mathcal{F} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}}$
Within	$\mathrm{SS_W}$	$N-m$	$\mathrm{MS_W} = \frac{\mathrm{SS_W}}{N-m}$	
Total	$\mathrm{SS_T}$	$N-1$		

Table 20.2. One-Way ANOVA Table with unequal sample size

Example 20.3. Three sections of elementary statistics were taught by different instructors. A common final examination was given. The test scores are given in the table below. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that there is a difference in the average grades given by the three instructors.

Elementary Statistics		
Instructor A	Instructor B	Instructor C
75	90	17
91	80	81
83	50	55
45	93	70
82	53	61
75	87	43
68	76	89
47	82	73
38	78	58
	80	70
	33	
	79	

Answer: Using the formulas (17) - (21), we compute the sum of squares SS_W, SS_B and SS_T as

$$SS_W = 10362, \qquad SS_B = 755, \qquad \text{and} \qquad SS_T = 11117.$$

The ANOVA table for this problem is shown below.

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	755	2	377	1.02
Within	10362	28	370	
Total	11117	30		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\, 28) = 3.34$. Since

$$1.02 = \mathcal{F} < F_{0.05}(2,\, 28) = 3.34$$

we accept the null hypothesis that there is no difference in the average grades given by the three instructors.

Note that using a statistical package like MINITAB, SAS or SPSS we can compute the p-value to be

$$p-\text{value} = P(F(2, 28) \geq 1.02) = 0.374.$$

Hence again we reach the same conclusion since p-value is less then the given α for this problem.

We conclude this section pointing out the advantages of choosing equal sample sizes (balance case) over the choice of unequal sample sizes (unbalance case). The first advantage is that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. The second advantage is that the choice of equal sample size minimizes the probability of committing a type II error.

20.3. Pair wise Comparisons

When the null hypothesis is rejected using the F-test in ANOVA, one may still wants to know where the difference among the means is. There are several methods to find out where the significant differences in the means lie after the ANOVA procedure is performed. Among the most commonly used tests are Scheffé test and Tuckey test. In this section, we give a brief description of these tests.

In order to perform the Scheffé test, we have to compare the means two at a time using all possible combinations of means. Since we have m means, we need $\binom{m}{2}$ pair wise comparisons. A pair wise comparison can be viewed as a test of the null hypothesis $H_0 : \mu_i = \mu_k$ against the alternative $H_a : \mu_i \neq \mu_k$ for all $i \neq k$.

To conduct this test we compute the statistics

$$F_s = \frac{\left(\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}\right)^2}{MS_W \left(\frac{1}{n_i} + \frac{1}{n_k}\right)},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, n_i and n_k are the respective sample sizes, and MS_W is the mean sum of squared of within group. We reject the null hypothesis at a significance level of α if

$$F_s > (m-1)F_\alpha(m-1, N-m)$$

where $N = n_1 + n_2 + \cdots + n_m$.

Example 20.4. Perform the analysis of variance and test the null hypothesis at the 0.05 level of significance for the following data given in the table below. Further perform a Scheffé test to determine where the significant differences in the means lie.

Sample		
1	2	3
9.2	9.5	9.4
9.1	9.5	9.3
9.2	9.5	9.3
9.2	9.6	9.3
9.3	9.5	9.2
9.2	9.4	9.3

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2,\, 15) = 3.68$. Since

$$35.0 = \mathcal{F} > F_{0.05}(2,\, 15) = 3.68$$

we reject the null hypothesis. Now we perform the Scheffé test to determine where the significant differences in the means lie. From given data, we obtain $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$. Since $m = 3$, we have to make 3 pair wise comparisons, namely μ_1 with μ_2, μ_1 with μ_3, and μ_2 with μ_3. First we consider the comparison of μ_1 with μ_2. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}\right)^2}{MS_W \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{(9.2 - 9.5)^2}{0.004\,\left(\frac{1}{6} + \frac{1}{6}\right)} = 67.5.$$

Since

$$67.5 = F_s > 2\, F_{0.05}(2,\, 15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we consider the comparison of μ_1 with μ_3. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W\,\left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = \frac{(9.2 - 9.3)^2}{0.004\,\left(\frac{1}{6} + \frac{1}{6}\right)} = 7.5.$$

Since

$$7.5 = F_s > 2\,F_{0.05}(2,\,15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we consider the comparison of μ_2 with μ_3. For this case, we find

$$F_s = \frac{\left(\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}\right)^2}{MS_W\,\left(\frac{1}{n_2} + \frac{1}{n_3}\right)} = \frac{(9.5 - 9.3)^2}{0.004\,\left(\frac{1}{6} + \frac{1}{6}\right)} = 30.0.$$

Since

$$30.0 = F_s > 2\,F_{0.05}(2,\,15) = 7.36$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Next consider the Tukey test. Tuckey test is applicable when we have a balanced case, that is when the sample sizes are equal. For Tukey test we compute the statistics

$$Q = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{k\bullet}}{\sqrt{\frac{MS_W}{n}}},$$

where $\overline{Y}_{i\bullet}$ and $\overline{Y}_{k\bullet}$ are the means of the samples being compared, n is the size of the samples, and MS_W is the mean sum of squared of within group. At a significance level α, we reject the null hypothesis H_0 if

$$|Q| > Q_\alpha(m, \nu)$$

where ν represents the degrees of freedom for the error mean square.

Example 20.5. For the data given in Example 20.4 perform a Tukey test to determine where the significant differences in the means lie.

Answer: We have seen that $\overline{Y}_{1\bullet} = 9.2$, $\overline{Y}_{2\bullet} = 9.5$ and $\overline{Y}_{3\bullet} = 9.3$.

First we compare μ_1 with μ_2. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.3}{\sqrt{\frac{0.004}{6}}} = -11.6189.$$

Since

$$11.6189 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_2$ in favor of the alternative $H_a : \mu_1 \neq \mu_2$.

Next we compare μ_1 with μ_3. For this we compute

$$Q = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.2 - 9.5}{\sqrt{\frac{0.004}{6}}} = -3.8729.$$

Since

$$3.8729 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_3$ in favor of the alternative $H_a : \mu_1 \neq \mu_3$.

Finally we compare μ_2 with μ_3. For this we compute

$$Q = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{3\bullet}}{\sqrt{\frac{MS_W}{n}}} = \frac{9.5 - 9.3}{\sqrt{\frac{0.004}{6}}} = 7.7459.$$

Since

$$7.7459 = |Q| > Q_{0.05}(2,\, 15) = 3.01$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_3$ in favor of the alternative $H_a : \mu_2 \neq \mu_3$.

Often in scientific and engineering problems, the experiment dictates the need for comparing simultaneously each treatment with a control. Now we describe a test developed by C. W. Dunnett for determining significant differences between each treatment mean and the control. Suppose we wish to test the m hypotheses

$$H_0 : \mu_0 = \mu_i \quad \text{versus} \quad H_a : \mu_0 \neq \mu_i \qquad \text{for } \; i = 1, 2, ..., m,$$

where μ_0 represents the mean yield for the population of measurements in which the control is used. To test the null hypotheses specified by H_0 against two-sided alternatives for an experimental situation in which there are m treatments, excluding the control, and n observation per treatment, we first calculate

$$D_i = \frac{\overline{Y}_{i\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}}, \quad i = 1, 2, ..., m.$$

At a significance level α, we reject the null hypothesis H_0 if

$$|D_i| > D_{\frac{\alpha}{2}}(m, \nu)$$

where ν represents the degrees of freedom for the error mean square. The values of the quantity $D_{\frac{\alpha}{2}}(m, \nu)$ are tabulated for various α, m and ν.

Example 20.6. For the data given in the table below perform a Dunnett test to determine any significant differences between each treatment mean and the control.

Control	Sample 1	Sample 2
9.2	9.5	9.4
9.1	9.5	9.3
9.2	9.5	9.3
9.2	9.6	9.3
9.3	9.5	9.2
9.2	9.4	9.3

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	0.280	2	0.140	35.0
Within	0.600	15	0.004	
Total	0.340	17		

At the significance level $\alpha = 0.05$, we find that $D_{0.025}(2, 15) = 2.44$. Since

$$35.0 = D > D_{0.025}(2,\ 15) = 2.44$$

we reject the null hypothesis. Now we perform the Dunnett test to determine if there is any significant differences between each treatment mean and the control. From given data, we obtain $\overline{Y}_{0\bullet} = 9.2$, $\overline{Y}_{1\bullet} = 9.5$ and $\overline{Y}_{2\bullet} = 9.3$. Since $m = 2$, we have to make 2 pair wise comparisons, namely μ_0 with μ_1, and μ_0 with μ_2. First we consider the comparison of μ_0 with μ_1. For this case, we find

$$D_1 = \frac{\overline{Y}_{1\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}} = \frac{9.5 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 8.2158.$$

Since

$$8.2158 = D_1 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_1 = \mu_0$ in favor of the alternative $H_a : \mu_1 \neq \mu_0$.

Next we find

$$D_2 = \frac{\overline{Y}_{2\bullet} - \overline{Y}_{0\bullet}}{\sqrt{\frac{2\,MS_W}{n}}} = \frac{9.3 - 9.2}{\sqrt{\frac{2\,(0.004)}{6}}} = 2.7386.$$

Since

$$2.7386 = D_2 > D_{0.025}(2,\, 15) = 2.44$$

we reject the null hypothesis $H_0 : \mu_2 = \mu_0$ in favor of the alternative $H_a : \mu_2 \neq \mu_0$.

20.4. Tests for the Homogeneity of Variances

One of the assumptions behind the ANOVA is the equal variance, that is the variances of the variables under consideration should be the same for all population. Earlier we have pointed out that the $\mathcal{F}$-statistics is insensitive to slight departures from the assumption of equal variances when the sample sizes are equal. Nevertheless it is advisable to run a preliminary test for homogeneity of variances. Such a test would certainly be advisable in the case of unequal sample sizes if there is a doubt concerning the homogeneity of population variances.

Suppose we want to test the null hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$$

versus the alternative hypothesis

$$H_a : \text{not all variances are equal.}$$

A frequently used test for the homogeneity of population variances is the Bartlett test. Bartlett (1937) proposed a test for equal variances that was modification of the normal-theory likelihood ratio test.

We will use this test to test the above null hypothesis H_0 against H_a. First, we compute the m sample variances $S_1^2, S_2^2, ..., S_m^2$ from the samples of size $n_1, n_2, ..., n_m$, with $n_1 + n_2 + \cdots + n_m = N$. The test statistics B_c is given by

$$B_c = \frac{(N-m)\,\ln S_p^2 - \sum_{i=1}^{m}(n_i - 1)\,\ln S_i^2}{1 + \frac{1}{3\,(m-1)}\left(\sum_{i=1}^{m}\frac{1}{n_i - 1} - \frac{1}{N-m}\right)}$$

where the pooled variance S_p^2 is given by

$$S_p^2 = \frac{\sum_{i=1}^{m}(n_i - 1)\,S_i^2}{N-m} = \text{MS}_\text{W}.$$

It is known that the sampling distribution of B_c is approximately chi-square with $m-1$ degrees of freedom, that is

$$B_c \sim \chi^2(m-1)$$

when $(n_i - 1) \geq 3$. Thus the Bartlett test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level α if

$$B_c > \chi^2_{1-\alpha}(m-1),$$

where $\chi^2_{1-\alpha}(m-1)$ denotes the upper $(1-\alpha)100$ percentile of the chi-square distribution with $m-1$ degrees of freedom.

Example 20.7. For the following data perform an ANOVA and then apply Bartlett test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Data			
Sample 1	Sample 2	Sample 3	Sample 4
34	29	32	34
28	32	34	29
29	31	30	32
37	43	42	28
42	31	32	32
27	29	33	34
29	28	29	29
35	30	27	31
25	37	37	30
29	44	26	37
41	29	29	43
40	31	31	42

Answer: The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	16.2	3	5.4	0.20
Within	1202.2	44	27.3	
Total	1218.5	47		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(2, 44) = 3.23$. Since

$$0.20 = \mathcal{F} < F_{0.05}(2, 44) = 3.23$$

we do not reject the null hypothesis.

Now we compute Bartlett test statistic B_c. From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Further, the pooled variance is

$$S_p^2 = \text{MS}_W = 27.3.$$

The statistics B_c is

$$\begin{aligned} B_c &= \frac{(N-m)\ln S_p^2 - \displaystyle\sum_{i=1}^{m}(n_i-1)\ln S_i^2}{1 + \frac{1}{3\,(m-1)}\left(\displaystyle\sum_{i=1}^{m}\frac{1}{n_i-1} - \frac{1}{N-m}\right)} \\ &= \frac{44\,\ln 27.3 - 11\,[\ln 35.2836 - \ln 30.1401 - \ln 19.4481 - \ln 24.4036\,]}{1 + \frac{1}{3\,(4-1)}\left(\frac{4}{12-1} - \frac{1}{48-4}\right)} \\ &= \frac{1.0537}{1.0378} = 1.0153. \end{aligned}$$

From chi-square table we find that $\chi^2_{0.95}(3) = 7.815$. Hence, since

$$1.0153 = B_c < \chi^2_{0.95}(3) = 7.815,$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

The Bartlett test assumes that the m samples should be taken from m normal populations. Thus Bartlett test is sensitive to departures from normality. The Levene test is an alternative to the Bartlett test that is less sensitive to departures from normality. Levene (1960) proposed a test for the homogeneity of population variances that considers the random variables

$$W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$$

and apply a one-way analysis of variance to these variables. If the F-test is significant, the homogeneity of variances is rejected.

Levene (1960) also proposed using F-tests based on the variables

$$W_{ij} = |Y_{ij} - \overline{Y}_{i\bullet}|, \qquad W_{ij} = \ln|Y_{ij} - \overline{Y}_{i\bullet}|, \qquad \text{and} \quad W_{ij} = \sqrt{|Y_{ij} - \overline{Y}_{i\bullet}|}.$$

Brown and Forsythe (1974c) proposed using the transformed variables based on the absolute deviations from the median, that is $W_{ij} = |Y_{ij} - Med(Y_{i\bullet})|$, where $Med(Y_{i\bullet})$ denotes the median of group i. Again if the F-test is significant, the homogeneity of variances is rejected.

Example 20.8. For the data in Example 20.7 do a Levene test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Answer: From data we find that $\overline{Y}_{1\bullet} = 33.00$, $\overline{Y}_{2\bullet} = 32.83$, $\overline{Y}_{3\bullet} = 31.83$, and $\overline{Y}_{4\bullet} = 33.42$. Next we compute $W_{ij} = \left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$. The resulting values are given in the table below.

Transformed Data			
Sample 1	Sample 2	Sample 3	Sample 4
1	14.7	0.0	0.3
25	0.7	4.7	19.5
16	3.4	3.4	2.0
16	103.4	103.4	29.3
81	3.4	0.0	2.0
36	14.7	1.4	0.3
16	23.4	8.0	19.5
4	8.0	23.4	5.8
64	17.4	26.7	11.7
16	124.7	34.0	12.8
64	14.7	0.0	91.8
49	3.4	0.7	73.7

Now we perform an ANOVA to the data given in the table above. The ANOVA table for this data is given by

Source of variation	Sums of squares	Degree of freedom	Mean squares	F-statistics $\mathcal{F}$
Between	1430	3	477	0.46
Within	45491	44	1034	
Total	46922	47		

At the significance level $\alpha = 0.05$, we find the F-table that $F_{0.05}(3,\, 44) = 2.84$. Since

$$0.46 = \mathcal{F} < F_{0.05}(3,\, 44) = 2.84$$

we do not reject the null hypothesis that the variances are equal. Hence Bartlett test suggests that the homogeneity of variances condition is met.

Although Bartlet test is most widely used test for homogeneity of variances a test due to Cochran provides a computationally simple procedure. Cochran test is one of the best method for detecting cases where the variance of one of the groups is much larger than that of the other groups. The test statistics of Cochran test is give by

$$C = \frac{\max\limits_{1 \leq i \leq m} S_i^2}{\sum\limits_{i=1}^{m} S_i^2}.$$

The Cochran test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots \sigma_m^2$ at a significance level α if

$$C > C_\alpha.$$

The critical values of C_α were originally published by Eisenhart *et al* (1947) for some combinations of degrees of freedom ν and the number of groups m. Here the degrees of freedom ν are

$$\nu = \max_{1 \leq i \leq m} (n_i - 1).$$

Example 20.9. For the data in Example 20.7 perform a Cochran test to examine if the homogeneity of variances condition is met for a significance level 0.05.

Answer: From the data the variances of each group can be found to be

$$S_1^2 = 35.2836, \qquad S_2^2 = 30.1401, \qquad S_3^2 = 19.4481, \qquad S_4^2 = 24.4036.$$

Hence the test statistic for Cochran test is

$$C = \frac{35.2836}{35.2836 + 30.1401 + 19.4481 + 24.4036} = \frac{35.2836}{109.2754} = 0.3328.$$

The critical value $C_{0.5}(3, 11)$ is given by 0.4884. Since

$$0.3328 = C < C_{0.5}(3, 11) = 0.4884.$$

At a significance level $\alpha = 0.05$, we do not reject the null hypothesis that the variances are equal. Hence Cochran test suggests that the homogeneity of variances condition is met.

20.5. Exercises

1. A consumer organization wants to compare the prices charged for a particular brand of refrigerator in three types of stores in Louisville: discount stores, department stores and appliance stores. Random samples of 6 stores of each type were selected. The results were shown below.

Discount	Department	Appliance
1200	1700	1600
1300	1500	1500
1100	1450	1300
1400	1300	1500
1250	1300	1700
1150	1500	1400

At the 0.05 level of significance, is there any evidence of a difference in the average price between the types of stores?

2. It is conjectured that a certain gene might be linked to ovarian cancer. The ovarian cancer is sub-classified into three categories: stage I, stage II and stage III-IV. There are three random samples available; one from each stage. The samples are labelled with three colors dyes and hybridized on a four channel cDNA microarray (one channel remains unused). The experiment is repeated 5 times and the following data were obtained.

Microarray Data			
Array	mRNA 1	mRNA 2	mRNA 3
1	100	95	70
2	90	93	72
3	105	79	81
4	83	85	74
5	78	90	75

Is there any difference between the three mRNA samples at 0.05 significance level?

Chapter 21

GOODNESS OF FITS TESTS

In point estimation, interval estimation or hypothesis test we always started with a random sample $X_1, X_2, ..., X_n$ of size n from a known distribution. In order to apply the theory to data analysis one has to know the distribution of the sample. Quite often the experimenter (or data analyst) assumes the nature of the sample distribution based on his subjective knowledge.

Goodness of fit tests are performed to validate experimenter opinion about the distribution of the population from where the sample is drawn. The most commonly known and most frequently used goodness of fit tests are the Kolmogorov-Smirnov (KS) test and the Pearson chi-square (χ^2) test. There is a controversy over which test is the most powerful, but the general feeling seems to be that the Kolmogorov-Smirnov test is probably more powerful than the chi-square test in most situations. The KS test measures the distance between distribution functions, while the χ^2 test measures the distance between density functions. Usually, if the population distribution is continuous, then one uses the Kolmogorov-Smirnov where as if the population distribution is discrete, then one performs the Pearson's chi-square goodness of fit test.

21.1. Kolmogorov-Smirnov Test

Let $X_1, X_2, ..., X_n$ be a random sample from a population X. We hypothesized that the distribution of X if $F(x)$. Further, we wish to test our hypothesis. Thus our null hypothesis is

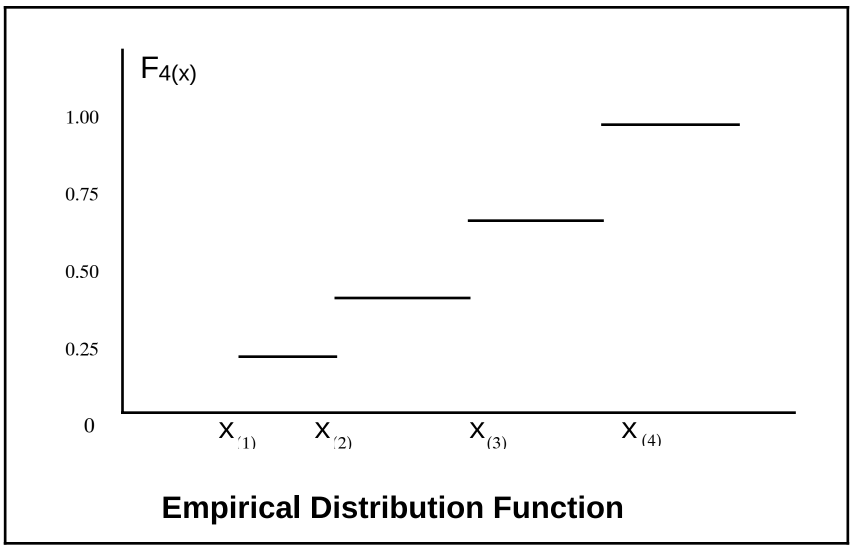
$$H_o : X \sim F(x).$$

We would like to design a test of this null hypothesis against the alternative $H_a : X \not\sim F(x)$.

In order to design a test, first of all we need a statistic which will unbiasedly estimate the unknown distribution $F(x)$ of the population X using the random sample $X_1, X_2, ..., X_n$. Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ be the observed values of the ordered statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$. The empirical distribution of the random sample is defined as

$$F_n(x) = \begin{cases} 0 & \text{if} \quad x < x_{(1)}, \\ \frac{k}{n} & \text{if} \quad x_{(k)} \le x < x_{(k+1)}, \qquad \text{for } k = 1, 2, ..., n-1, \\ 1 & \text{if} \quad x_{(n)} \le x. \end{cases}$$

The graph of the empirical distribution function $F_4(x)$ is shown below.



For a fixed value of x, the empirical distribution function can be considered as a random variable that takes on the values

$$0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, \frac{n}{n}.$$

First we show that $F_n(x)$ is an unbiased estimator of the population distribution $F(x)$. That is,

$$E(F_n(x)) = F(x) \tag{1}$$

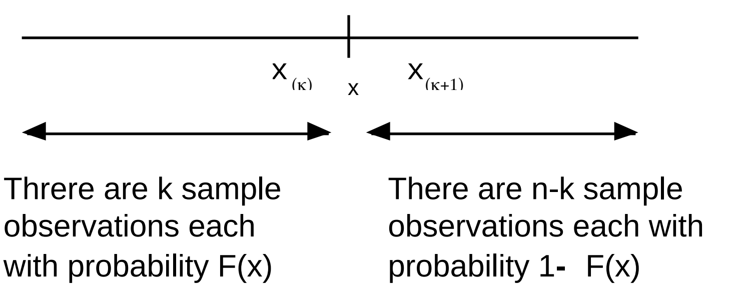
for a fixed value of x. To establish (1), we need the probability density function of the random variable $F_n(x)$. From the definition of the empirical distribution we see that if exactly k observations are less than or equal to x, then

$$F_n(x) = \frac{k}{n}$$

which is

$$n\, F_n(x) = k.$$

The probability that an observation is less than or equal to x is given by $F(x)$.



Distribution of the Empirical Distribution Function

Hence (see figure above)

$$\begin{aligned} P(n\, F_n(x) = k) &= P\left(F_n(x) = \frac{k}{n}\right) \\ &= \binom{n}{k} [F(x)]^k \, [1 - F(x)]^{n-k} \end{aligned}$$

for $k = 0, 1, ..., n$. Thus

$$n\, F_n(x) \sim BIN(n, F(x)).$$

Thus the expected value of the random variable $n\,F_n(x)$ is given by

$$\begin{aligned} E(n\,F_n(x)) &= n\,F(x) \\ n\,E(F_n(x)) &= n\,F(x) \\ E(F_n(x)) &= F(x). \end{aligned}$$

This shows that, for a fixed x, $F_n(x)$, on an average, equals to the population distribution function $F(x)$. Hence the empirical distribution function $F_n(x)$ is an unbiased estimator of $F(x)$.

Since $n\,F_n(x) \sim BIN(n, F(x))$, the variance of $n\,F_n(x)$ is given by

$$Var(n\,F_n(x)) = n\,F(x)\,[1 - F(x)].$$

Hence the variance of $F_n(x)$ is

$$Var(F_n(x)) = \frac{F(x)\,[1 - F(x)]}{n}.$$

It is easy to see that $Var(F_n(x)) \to 0$ as $n \to \infty$ for all values of x. Thus the empirical distribution function $F_n(x)$ and $F(x)$ tend to be closer to each other with large n. As a matter of fact, Glivenkno, a Russian mathematician, proved that $F_n(x)$ converges to $F(x)$ uniformly in x as $n \to \infty$ with probability one.

Because of the convergence of the empirical distribution function to the theoretical distribution function, it makes sense to construct a goodness of fit test based on the closeness of $F_n(x)$ and hypothesized distribution $F(x)$.

Let

$$D_n = \max_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

That is D_n is the maximum of all pointwise differences $|F_n(x) - F(x)|$. The distribution of the Kolmogorov-Smirnov statistic, D_n can be derived. However, we shall not do that here as the derivation is quite involved. In stead, we give a closed form formula for $P(D_n \leq d)$. If $X_1, X_2, ..., X_n$ is a sample from a population with continuous distribution function $F(x)$, then

$$P(D_n \leq d) = \begin{cases} 0 & \text{if } d \leq \frac{1}{2\,n} \\ n! \displaystyle\prod_{i=1}^{n} \int_{2\,i-d}^{2\,i-\frac{1}{n}+d} du & \text{if } \frac{1}{2n} < d < 1 \\ 1 & \text{if } d \geq 1 \end{cases}$$

where $du = du_1 du_2 \cdots du_n$ with $0 < u_1 < u_2 < \cdots < u_n < 1$. Further,

$$\lim_{n\to\infty} P(\sqrt{n}\, D_n \leq d) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2\,k^2\,d^2}.$$

These formulas show that the distribution of the Kolmogorov-Smirnov statistic D_n is distribution free, that is, it does not depend on the distribution F of the population.

For most situations, it is sufficient to use the following approximations due to Kolmogorov:

$$P(\sqrt{n}\, D_n \leq d) \approx 1 - 2e^{-2nd^2} \qquad \text{for } d > \frac{1}{\sqrt{n}}.$$

If the null hypothesis $H_o : X \sim F(x)$ is true, the statistic D_n is small. It is therefore reasonable to reject H_o if and only if the observed value of D_n is larger than some constant d_n. If the level of significance is given to be α, then the constant d_n can be found from

$$\alpha = P(D_n > d_n \,/\, H_o \text{ is true}) \approx 2e^{-2nd_n^2}.$$

This yields the following hypothesis test: Reject H_o if $D_n \geq d_n$ where

$$d_n = \sqrt{-\frac{1}{2n} \ln\left(\frac{\alpha}{2}\right)}$$

is obtained from the above Kolmogorov's approximation. Note that the approximate value of d_{12} obtained by the above formula is equal to 0.3533 when $\alpha = 0.1$, however more accurate value of d_{12} is 0.34.

Next we address the issue of the computation of the statistics D_n. Let us define

$$D_n^+ = \max_{x\in\mathbb{R}} \{F_n(x) - F(x)\}$$

and

$$D_n^- = \max_{x\in\mathbb{R}} \{F(x) - F_n(x)\}.$$

Then it is easy to see that

$$D_n = \max\{D_n^+,\, D_N^-\}.$$

Further it can be shown that

$$D_n^+ = \max\left\{ \max_{1\leq i\leq n} \left[\frac{i}{n} - F(x_{(i)}) \right],\, 0 \right\}$$

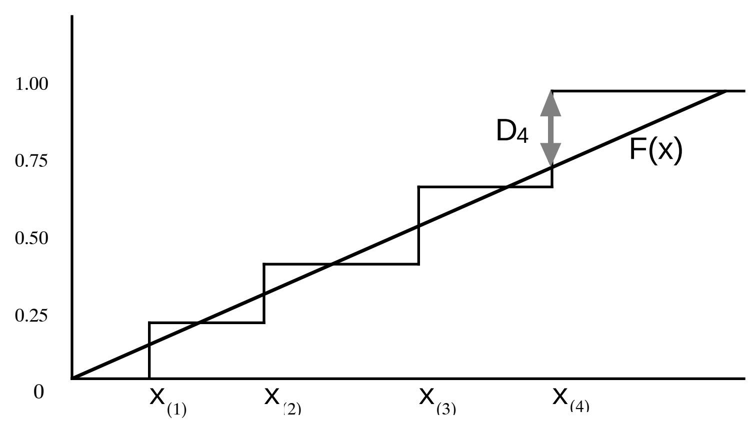
and

$$D_n^- = \max \left\{ \max_{1 \le i \le n} \left[F(x_{(i)}) - \frac{i-1}{n} \right], \, 0 \right\}.$$

Therefore it can also be shown that

$$D_n = \max_{1 \le i \le n} \left\{ \max \left[\frac{i}{n} - F(x_{(i)}), \, F(x_{(i)}) - \frac{i-1}{n} \right] \right\},$$

where $F_n(x_{(i)}) = \frac{i}{n}$. The following figure illustrates the Kolmogorov-Smirnov statistics D_n when $n = 4$.



Kolmogorov-Smirnov Statistic

Example 21.1. The data on the heights of 12 infants are given below: 18.2, 21.4, 22.6, 17.4, 17.6, 16.7, 17.1, 21.4, 20.1, 17.9, 16.8, 23.1. Test the hypothesis that the data came from some normal population at a significance level $\alpha = 0.1$.

Answer: Here, the null hypothesis is

$$H_o : X \sim N(\mu, \sigma^2).$$

First we estimate μ and σ^2 from the data. Thus, we get

$$\overline{x} = \frac{230.3}{12} = 19.2.$$

and

$$s^2 = \frac{4482.01 - \frac{1}{12}(230.3)^2}{12-1} = \frac{62.17}{11} = 5.65.$$

Hence $s = 2.38$. Then by the null hypothesis

$$F(x_{(i)}) = P\left(Z \leq \frac{x_{(i)} - 19.2}{2.38}\right)$$

where $Z \sim N(0,1)$ and $i = 1, 2, ..., n$. Next we compute the Kolmogorov-Smirnov statistic D_n the given sample of size 12 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{12} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{12}$
1	16.7	0.1469	−0.0636	0.1469
2	16.8	0.1562	0.0105	0.0729
3	17.1	0.1894	0.0606	0.0227
4	17.4	0.2236	0.1097	−0.0264
5	17.6	0.2514	0.1653	−0.0819
6	17.9	0.2912	0.2088	−0.1255
7	18.2	0.3372	0.2461	−0.1628
8	20.1	0.6480	0.0187	0.0647
9	21.4	0.8212	0.0121	0.0712
10	21.4			
11	22.6	0.9236	−0.0069	0.0903
12	23.1	0.9495	0.0505	0.0328

Thus

$$D_{12} = 0.2461.$$

From the tabulated value, we see that $d_{12} = 0.34$ for significance level $\alpha = 0.1$. Since D_{12} is smaller than d_{12} we accept the null hypothesis $H_o : X \sim N(\mu, \sigma^2)$. Hence the data came from a normal population.

Example 21.2. Let $X_1, X_2, ..., X_{10}$ be a random sample from a distribution whose probability density function is

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ \\ 0 & \text{otherwise.} \end{cases}$$

Based on the observed values 0.62, 0.36, 0.23, 0.76, 0.65, 0.09, 0.55, 0.26, 0.38, 0.24, test the hypothesis $H_o : X \sim UNIF(0,1)$ against $H_a : X \not\sim UNIF(0,1)$ at a significance level $\alpha = 0.1$.

Answer: The null hypothesis is $H_o : X \sim UNIF(0,1)$. Thus

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

Hence

$$F(x_{(i)}) = x_{(i)} \qquad \text{for } i = 1, 2, ..., n.$$

Next we compute the Kolmogorov-Smirnov statistic D_n the given sample of size 10 using the following tabular form.

i	$x_{(i)}$	$F(x_{(i)})$	$\frac{i}{10} - F(x_{(i)})$	$F(x_{(i)}) - \frac{i-1}{10}$
1	0.09	0.09	0.01	0.09
2	0.23	0.23	−0.03	0.13
3	0.24	0.24	0.06	0.04
4	0.26	0.26	0.14	−0.04
5	0.36	0.36	0.14	−0.04
6	0.38	0.38	0.22	−0.12
7	0.55	0.55	0.15	−0.05
8	0.62	0.62	0.18	−0.08
9	0.65	0.65	0.25	−0.15
10	0.76	0.76	0.24	−0.14

Thus

$$D_{10} = 0.25.$$

From the tabulated value, we see that $d_{10} = 0.37$ for significance level $\alpha = 0.1$. Since D_{10} is smaller than d_{10} we accept the null hypothesis

$$H_o : X \sim UNIF(0,1).$$

21.2 Chi-square Test

The chi-square goodness of fit test was introduced by Karl Pearson in 1900. Recall that the Kolmogorov-Smirnov test is only for testing a specific continuous distribution. Thus if we wish to test the null hypothesis

$$H_o : X \sim BIN(n,p)$$

against the alternative $H_a : X \not\sim BIN(n,p)$, then we can not use the Kolmogorov-Smirnov test. Pearson chi-square goodness of fit test can be used for testing of null hypothesis involving discrete as well as continuous distribution. Unlike Kolmogorov-Smirnov test, the Pearson chi-square test uses the density function the population X.

Let $X_1, X_2, ..., X_n$ be a random sample from a population X with probability density function $f(x)$. We wish to test the null hypothesis

$$H_o : X \sim f(x)$$

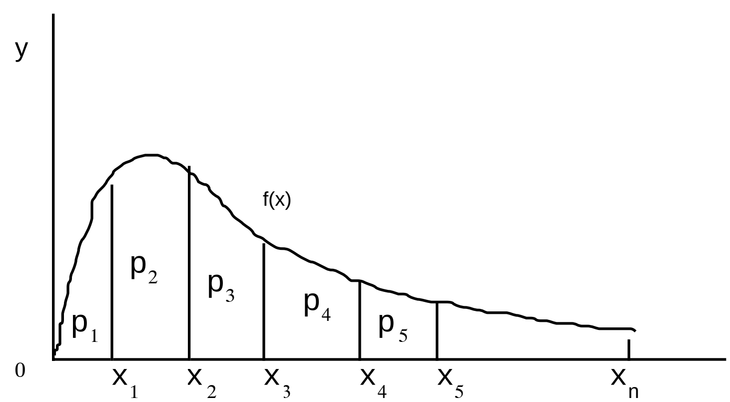
against

$$H_a : X \not\sim f(x).$$

If the probability density function $f(x)$ is continuous, then we divide up the abscissa of the probability density function $f(x)$ and calculate the probability p_i for each of the interval by using

$$p_i = \int_{x_{i-1}}^{x_i} f(x)\, dx,$$

where $\{x_0, x_1, ..., x_n\}$ is a partition of the domain of the $f(x)$.



Discretization of continuous density function

Let $Y_1, Y_2, ..., Y_m$ denote the number of observations (from the random sample $X_1, X_2, ..., X_n$) is $1^{\text{st}}, 2^{\text{nd}}, 3^{\text{rd}}, ..., m^{\text{th}}$ interval, respectively.

Since the sample size is n, the number of observations expected to fall in the i^{th} interval is equal to np_i. Then

$$Q = \sum_{i=1}^{m} \frac{(Y_i - np_i)^2}{np_i}$$

measures the closeness of observed Y_i to expected number np_i. The distribution of Q is chi-square with $m-1$ degrees of freedom. The derivation of this fact is quite involved and beyond the scope of this introductory level book.

Although the distribution of Q for $m > 2$ is hard to derive, yet for $m = 2$ it not very difficult. Thus we give a derivation to convince the reader that Q has χ^2 distribution. Notice that $Y_1 \sim BIN(n, p_1)$. Hence for large n by the central limit theorem, we have

$$\frac{Y_1 - n\,p_1}{\sqrt{n\,p_1\,(1-p_1)}} \sim N(0,1).$$

Thus

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} \sim \chi^2(1).$$

Since

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1\,(1-p_1)} = \frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)},$$

we have This implies that

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(Y_1 - n\,p_1)^2}{n\,(1-p_1)} \sim \chi^2(1)$$

which is

$$\frac{(Y_1 - n\,p_1)^2}{n\,p_1} + \frac{(n - Y_2 - n + n\,p_2)^2}{n\,p_2} \sim \chi^2(1)$$

due to the facts that $Y_1 + Y_2 = n$ and $p_1 + p_2 = 1$. Hence

$$\sum_{i=1}^{2} \frac{(Y_i - n\,p_i)^2}{n\,p_i} \sim \chi^2(1),$$

that is, the chi-square statistic Q has approximate chi-square distribution.

Now the simple null hypothesis

$$H_0 : p_1 = p_{10}, \;\; p_2 = p_{20}, \;\; \cdots \;\; p_m = p_{m0}$$

is to be tested against the composite alternative

$$H_a : \text{at least one } p_i \text{ is not equal to } p_{i0} \text{ for some } i.$$

Here $p_{10}, p_{20}, ..., p_{m0}$ are fixed probability values. If the null hypothesis is true, then the statistic

$$Q = \sum_{i=1}^{m} \frac{(Y_i - n\,p_{i0})^2}{n\,p_{i0}}$$

has an approximate chi-square distribution with $m-1$ degrees of freedom. If the significance level α of the hypothesis test is given, then

$$\alpha = P\left(Q \geq \chi^2_{1-\alpha}(m-1)\right)$$

and the test is "Reject H_o if $Q \geq \chi^2_{1-\alpha}(m-1)$." Here $\chi^2_{1-\alpha}(m-1)$ denotes a real number such that the integral of the chi-square density function with $m-1$ degrees of freedom from zero to this real number $\chi^2_{1-\alpha}(m-1)$ is $1-\alpha$. Now we give several examples to illustrate the chi-square goodness-of-fit test.

Example 21.3. A die was rolled 30 times with the results shown below:

Number of spots	1	2	3	4	5	6
Frequency (x_i)	1	4	9	9	2	5

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic and decision reached?

Answer: In this problem, the null hypothesis is

$$H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}.$$

The alternative hypothesis is that not all p_i's are equal to $\frac{1}{6}$. The test will be based on 30 trials, so $n = 30$. The test statistic

$$Q = \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i},$$

where $p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$. Thus

$$n\,p_i = (30)\,\frac{1}{6} = 5$$

and

$$\begin{aligned} Q &= \sum_{i=1}^{6} \frac{(x_i - n\,p_i)^2}{n\,p_i} \\ &= \sum_{i=1}^{6} \frac{(x_i - 5)^2}{5} \\ &= \frac{1}{5}\,[16 + 1 + 16 + 16 + 9] \\ &= \frac{58}{5} = 11.6. \end{aligned}$$

The tabulated χ^2 value for $\chi^2_{0.95}(5)$ is given by

$$\chi^2_{0.95}(5) = 11.07.$$

Since

$$11.6 = Q > \chi^2_{0.95}(5) = 11.07$$

the null hypothesis $H_o : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ should be rejected.

Example 21.4. It is hypothesized that an experiment results in outcomes K, L, M and N with probabilities $\frac{1}{5}$, $\frac{3}{10}$, $\frac{1}{10}$ and $\frac{2}{5}$, respectively. Forty independent repetitions of the experiment have results as follows:

Outcome	K	L	M	N
Frequency	11	14	5	10

If a chi-square goodness of fit test is used to test the above hypothesis at the significance level $\alpha = 0.01$, then what is the value of the chi-square statistic and the decision reached?

Answer: Here the null hypothesis to be tested is

$$H_o : p(K) = \frac{1}{5},\ p(L) = \frac{3}{10},\ p(M) = \frac{1}{10},\ p(N) = \frac{2}{5}.$$

The test will be based on $n = 40$ trials. The test statistic

$$\begin{aligned}
Q &= \sum_{k=1}^{4} \frac{(x_k - np_k)^2}{n\,p_k} \\
&= \frac{(x_1 - 8)^2}{8} + \frac{(x_2 - 12)^2}{12} + \frac{(x_3 - 4)^2}{4} + \frac{(x_4 - 16)^2}{16} \\
&= \frac{(11 - 8)^2}{8} + \frac{(14 - 12)^2}{12} + \frac{(5 - 4)^2}{4} + \frac{(10 - 16)^2}{16} \\
&= \frac{9}{8} + \frac{4}{12} + \frac{1}{4} + \frac{36}{16} \\
&= \frac{95}{24} = 3.958.
\end{aligned}$$

From chi-square table, we have

$$\chi^2_{0.99}(3) = 11.35.$$

Thus

$$3.958 = Q < \chi^2_{0.99}(3) = 11.35.$$

Therefore we accept the null hypothesis.

Example 21.5. Test at the 10% significance level the hypothesis that the following data

06.88	06.92	04.80	09.85	07.05	19.06	06.54	03.67	02.94	04.89
69.82	06.97	04.34	13.45	05.74	10.07	16.91	07.47	05.04	07.97
15.74	00.32	04.14	05.19	18.69	02.45	23.69	44.10	01.70	02.14
05.79	03.02	09.87	02.44	18.99	18.90	05.42	01.54	01.55	20.99
07.99	05.38	02.36	09.66	00.97	04.82	10.43	15.06	00.49	02.81

give the values of a random sample of size 50 from an exponential distribution with probability density function

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta > 0$.

Answer: From the data $\overline{x} = 9.74$ and $s = 11.71$. Notice that

$$H_o : X \sim EXP(\theta).$$

Hence we have to partition the domain of the experimental distribution into m parts. There is no rule to determine what should be the value of m. We assume $m = 10$ (an arbitrary choice for the sake of convenience). We partition the domain of the given probability density function into 10 mutually disjoint sets of equal probability. This partition can be found as follow.

Note that $\overline{x}$ estimate θ. Thus

$$\widehat{\theta} = \overline{x} = 9.74.$$

Now we compute the points $x_1, x_2, ..., x_{10}$ which will be used to partition the domain of $f(x)$

$$\begin{aligned} \frac{1}{10} &= \int_{x_o}^{x_1} \frac{1}{\theta}\, e^{-\frac{x}{\theta}} \\ &= -\left[e^{-\frac{x}{\theta}} \right]_0^{x_1} \\ &= 1 - e^{-\frac{x_1}{\theta}}. \end{aligned}$$

Hence

$$\begin{aligned} x_1 &= \theta \, \ln \left(\frac{10}{9} \right) \\ &= 9.74 \, \ln \left(\frac{10}{9} \right) \\ &= 1.026. \end{aligned}$$

Using the value of x_1, we can find the value of x_2. That is

$$\frac{1}{10} = \int_{x_1}^{x_2} \frac{1}{\theta}\, e^{-\frac{x}{\theta}}$$

$$= e^{-\frac{x_1}{\theta}} - e^{-\frac{x_2}{\theta}}.$$

Hence

$$x_2 = -\theta\, \ln\left(e^{-\frac{x_1}{\theta}} - \frac{1}{10} \right).$$

In general

$$x_k = -\theta\, \ln\left(e^{-\frac{x_{k-1}}{\theta}} - \frac{1}{10} \right)$$

for $k = 1, 2, ..., 9$, and $x_{10} = \infty$. Using these x_k's we find the intervals $A_k = [x_k,\, x_{k+1})$ which are tabulates in the table below along with the number of data points in each each interval.

Interval A_i	Frequency (o_i)	Expected value (e_i)
$[0,\ 1.026)$	3	5
$[1.026,\ 2.173)$	4	5
$[2.173,\ 3.474)$	6	5
$[3.474,\ 4.975)$	6	5
$[4.975,\ 6.751)$	7	5
$[6.751,\ 8.925)$	7	5
$[8.925,\ 11.727)$	5	5
$[11.727,\ 15.676)$	2	5
$[15.676,\ 22.437)$	7	5
$[22.437,\ \infty)$	3	5
Total	50	50

From this table, we compute the statistics

$$Q = \sum_{i=1}^{10} \frac{(o_i - e_i)^2}{e_i} = 6.4.$$

and from the chi-square table, we obtain

$$\chi^2_{0.9}(9) = 14.68.$$

Since

$$6.4 = Q < \chi^2_{0.9}(9) = 14.68$$

we accept the null hypothesis that the sample was taken from a population with exponential distribution.

21.3. Review Exercises

1. The data on the heights of 4 infants are: $18.2, 21.4, 16.7$ and 23.1. For a significance level $\alpha = 0.1$, use Kolmogorov-Smirnov Test to test the hypothesis that the data came from some uniform population on the interval $(15, 25)$. (Use $d_4 = 0.56$ at $\alpha = 0.1$.)

2. A four-sided die was rolled 40 times with the following results

Number of spots	1	2	3	4
Frequency	5	9	10	16

If a chi-square goodness of fit test is used to test the hypothesis that the die is fair at a significance level $\alpha = 0.05$, then what is the value of the chi-square statistic?

3. A coin is tossed 500 times and k heads are observed. If the chi-squares distribution is used to test the hypothesis that the coin is unbiased, this hypothesis will be accepted at 5 percents level of significance if and only if k lies between what values? (Use $\chi^2_{0.05}(1) = 3.84$.)

REFERENCES

[1] Aitken, A. C. (1944). *Statistical Mathematics.* 3rd edn. Edinburgh and London: Oliver and Boyd,

[2] Arbous, A. G. and Kerrich, J. E. (1951). Accident statistics and the concept of accident-proneness. *Biometrics*, 7, 340-432.

[3] Arnold, S. (1984). Pivotal quantities and invariant confidence regions. *Statistics and Decisions* 2, 257-280.

[4] Bain, L. J. and Engelhardt. M. (1992). *Introduction to Probability and Mathematical Statistics.* Belmont: Duxbury Press.

[5] Brown, L. D. (1988). *Lecture Notes*, Department of Mathematics, Cornell University. Ithaca, New York.

[6] Campbell, J. T. (1934). THe Poisson correlation function. *Proc. Edin. Math. Soc.*, Series 2, 4, 18-26.

[7] Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Belmont: Wadsworth.

[8] Castillo, E. (1988). *Extreme Value Theory in Engineering.* San Diego: Academic Press.

[9] Cherian, K. C. (1941). A bivariate correlated gamma-type distribution function. *J. Indian Math. Soc.*, 5, 133-144.

[10] Dahiya, R., and Guttman, I. (1982). Shortest confidence and prediction intervals for the log-normal. *The canadian Journal of Statistics* 10, 777-891.

[11] David, F.N. and Fix, E. (1961). Rank correlation and regression in a non-normal surface. *Proc. 4th Berkeley Symp. Math. Statist. & Prob.*, 1, 177-197.

[12] Desu, M. (1971). Optimal confidence intervals of fixed width. *The American Statistician* 25, 27-29.

[13] Dynkin, E. B. (1951). Necessary and sufficient statistics for a family of probability distributions. English translation in *Selected Translations in Mathematical Statistics and Probability*, 1 (1961), 23-41.

[14] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I.* New York: Wiley.

[15] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Volume II.* New York: Wiley.

[16] Ferentinos, K. K. (1988). On shortest confidence intervals and their relation uniformly minimum variance unbiased estimators. *Statistical Papers* 29, 59-75.

[17] Freund, J. E. and Walpole, R. E. (1987). *Mathematical Statistics.* Englewood Cliffs: Prantice-Hall.

[18] Galton, F. (1879). The geometric mean in vital and social statistics. *Proc. Roy. Soc.*, 29, 365-367.

[19] Galton, F. (1886). Family likeness in stature. With an appendix by J.D.H. Dickson. *Proc. Roy. Soc.*, 40, 42-73.

[20] Ghahramani, S. (2000). *Fundamentals of Probability.* Upper Saddle River, New Jersey: Prentice Hall.

[21] Graybill, F. A. (1961). *An Introduction to Linear Statistical Models*, Vol. 1. New YorK: McGraw-Hill.

[22] Guenther, W. (1969). Shortest confidence intervals. *The American Statistician* 23, 51-53.

[23] Guldberg, A. (1934). On discontinuous frequency functions of two variables. *Skand. Aktuar.*, 17, 89-117.

[24] Gumbel, E. J. (1960). Bivariate exponetial distributions. *J. Amer. Statist. Ass.*, 55, 698-707.

[25] Hamedani, G. G. (1992). Bivariate and multivariate normal characterizations: a brief survey. *Comm. Statist. Theory Methods*, 21, 2665-2688.

[26] Hamming, R. W. (1991). *The Art of Probability for Scientists and Engineers* New York: Addison-Wesley.

[27] Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* New York: Macmillan.

[28] Hogg, R. V. and Tanis, E. A. (1993). *Probability and Statistical Inference.* New York: Macmillan.

[29] Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, 51, 241-245.

[30] Kapteyn, J. C. (1903). *Skew Frequency Curves in Biology and Statistics.* Astronomical Laboratory, Noordhoff, Groningen.

[31] Kibble, W. F. (1941). A two-variate gamma type distribution. *Sankhya*, 5, 137-150.

[32] Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Erg. Math., Vol 2, Berlin: Springer-Verlag.

[33] Kolmogorov, A. N. (1956). *Foundations of the Theory of Probability.* New York: Chelsea Publishing Company.

[34] Kotlarski, I. I. (1960). On random variables whose quotient follows the Cauchy law. *Colloquium Mathematicum.* 7, 277-284.

[35] Isserlis, L. (1914). The application of solid hypergeometrical series to frequency distributions in space. *Phil. Mag.*, 28, 379-403.

[36] Laha, G. (1959). On a class of distribution functions where the quotient follows the Cauchy law. *Trans. Amer. Math. Soc.* 93, 205-215.

[37] Lundberg, O. (1934). *On Random Processes and their Applications to Sickness and Accident Statistics.* Uppsala: Almqvist and Wiksell.

[38] Mardia, K. V. (1970). *Families of Bivariate Distributions.* London: Charles Griffin & Co Ltd.

[39] Marshall, A. W. and Olkin, I. (1967). A multivariate exponential distribution. *J. Amer. Statist. Ass.*, 62. 30-44.

[40] McAlister, D. (1879). The law of the geometric mean. *Proc. Roy. Soc.*, 29, 367-375.

[41] McKay, A. T. (1934). Sampling from batches. *J. Roy. Statist. Soc.*, Suppliment, 1, 207-216.

[42] Meyer, P. L. (1970). *Introductory Probability and Statistical Applications.* Reading: Addison-Wesley.

[43] Mood, A., Graybill, G. and Boes, D. (1974). *Introduction to the Theory of Statistics* (3rd Ed.). New York: McGraw-Hill.

[44] Moran, P. A. P. (1967). Testing for correlation between non-negative variates. *Biometrika*, 54, 385-394.

[45] Morgenstern, D. (1956). Einfache Beispiele zweidimensionaler Verteilungen. *Mitt. Math. Statist.*, 8, 234-235.

[46] Papoulis, A. (1990). *Probability and Statistics.* Englewood Cliffs: Prantice-Hall.

[47] Pearson, K. (1924). On the moments of the hypergeometrical series. *Biometrika*, 16, 157-160.

[48] Pestman, W. R. (1998). *Mathematical Statistics: An Introduction* New York: Walter de Gruyter.

[49] Pitman, J. (1993). *Probability.* New York: Springer-Verlag.

[50] Plackett, R. L. (1965). A class of bivariate distributions. *J. Amer. Statist. Ass.*, 60, 516-522.

[51] Rice, S. O. (1944). Mathematical analysis of random noise. *Bell. Syst. Tech. J.*, 23, 282-332.

[52] Rice, S. O. (1945). Mathematical analysis of random noise. *Bell. Syst. Tech. J.*, 24, 46-156.

[53] Rinaman, W. C. (1993). *Foundations of Probability and Statistics.* New York: Saunders College Publishing.

[54] Rosenthal, J. S. (2000). *A First Look at Rigorous Probability Theory.* Singapore: World Scientific.

[55] Ross, S. (1988). *A First Course in Probability.* New York: Macmillan.

[56] Seshadri, V. and Patil, G. P. (1964). A characterization of a bivariate distribution by the marginal and the conditional distributions of the same component. *Ann. Inst. Statist. Math.*, 15, 215-221.

[57] Sveshnikov, A. A. (1978). *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions.* New York: Dover.

[58] Taylor, L. D. (1974). *Probability and Mathematical Statistics.* New York: Harper & Row.

[59] Tweedie, M. C. K. (1945). Inverse statistical variates. *Nature*, 155, 453.

[60] Waissi, G. R. (1993). A unifying probability density function. Appl. Math. Lett. 6, 25-26.

[61] Waissi, G. R. (1994). An improved unifying density function. Appl. Math. Lett. 7, 71-73.

[62] Waissi, G. R. (1998). Transformation of the unifying density to the normal distribution. Appl. Math. Lett. 11, 45-28.

[63] Wicksell, S. D. (1933). On correlation functions of Type III. *Biometrika*, 25, 121-133.

[64] Zehna, P. W. (1966). Invariance of maximum likelihood estimators. *Annals of Mathematical Statistics*, 37, 744.

[65] H. Scheffé (1959). *The Analysis of Variance.* New York: Wiley.

[66] Snedecor, G. W. and Cochran, W. G. (1983). *Statistical Methods.* 6th eds. Iowa State University Press, Ames, Iowa.

[67] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society*, London, Ser. A, 160, 268-282.

[68] Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *J. R. Stat. Soc., Suppli.*, 4, 137-183.

[69] Sahai, H. and Ageel, M. I. (2000). *The Analysis of Variance.* Boston: Birkhauser.

[70] Roussas, G. (2003). *An Introduction to Probability and Statistical Inference.* San Diego: Academic Press.

[71] Ross, S. M. (2000). *Introduction to Probability and Statistics for Engineers and Scientists.* San Diego: Harcourt Academic Press.

ANSWERES TO SELECTED REVIEW EXERCISES

CHAPTER 1

1. $\frac{7}{1912}$.
2. 244.
3. 7488.
4. (a) $\frac{4}{24}$, (b) $\frac{6}{24}$ and (c) $\frac{4}{24}$.
5. 0.95.
6. $\frac{4}{7}$.
7. $\frac{2}{3}$.
8. 7560.
10. 4^3.
11. 2.
12. 0.3238.
13. S has countable number of elements.
14. S has uncountable number of elements.
15. $\frac{25}{648}$.
16. $(n-1)(n-2)\left(\frac{1}{2}\right)^{n+1}$.
17. $(5!)^2$.
18. $\frac{7}{10}$.
19. $\frac{1}{3}$.
20. $\frac{n+1}{3n-1}$.
21. $\frac{6}{11}$.
22. $\frac{1}{5}$.

CHAPTER 2

1. $\frac{1}{3}$.

2. $\frac{(6!)^2}{(21)^6}$.

3. 0.941.

4. $\frac{4}{5}$.

5. $\frac{6}{11}$.

6. $\frac{255}{256}$.

7. 0.2929.

8. $\frac{10}{17}$.

9. $\frac{30}{31}$.

10. $\frac{7}{24}$.

11. $\frac{\left(\frac{4}{10}\right)\left(\frac{3}{6}\right)}{\left(\frac{4}{10}\right)\left(\frac{3}{6}\right)+\left(\frac{6}{10}\right)\left(\frac{2}{5}\right)}$.

12. $\frac{(0.01)\,(0.9)}{(0.01)\,(0.9)+(0.99)\,(0.1)}$.

13. $\frac{1}{5}$.

14. $\frac{2}{9}$.

15. (a) $\left(\frac{2}{5}\right)\left(\frac{4}{52}\right)+\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)$ and (b) $\frac{\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)}{\left(\frac{2}{5}\right)\left(\frac{4}{52}\right)+\left(\frac{3}{5}\right)\left(\frac{4}{16}\right)}$.

16. $\frac{1}{4}$.

17. $\frac{3}{8}$.

18. 5.

19. $\frac{5}{42}$.

20. $\frac{1}{4}$.

CHAPTER 3

1. $\frac{1}{4}$.

2. $\frac{k+1}{2k+1}$.

3. $\frac{1}{\sqrt[3]{2}}$.

4. Mode of $X = 0$ and median of $X = 0$.

5. $\theta \ln \left(\frac{10}{9}\right)$.

6. $2 \ln 2$.

7. 0.25.

8. $f(2) = 0.5,\ f(3) = 0.2,\ f(\pi) = 0.3$.

9. $f(x) = \frac{1}{6} x^3 e^{-x}$.

10. $\frac{3}{4}$.

11. $a = 500$, mode $= 0.2$, and $P(X \geq 0.2) = 0.6766$.

12. 0.5.

13. 0.5.

14. $1 - F(-y)$.

15. $\frac{1}{4}$.

16. $R_X = \{3, 4, 5, 6, 7, 8, 9\}$;
$f(3) = f(4) = \frac{2}{20},\ f(5) = f(6) = f(7) = \frac{4}{20},\ f(8) = f(9) = \frac{2}{20}$.

17. $R_X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$;
$f(2) = \frac{1}{36},\ f(3) = \frac{2}{36},\ f(4) = \frac{3}{36},\ f(5) = \frac{4}{36},\ f(6) = \frac{5}{36},\ f(7) = \frac{6}{36},\ f(8) = \frac{5}{36},\ f(9) = \frac{4}{36},\ f(10) = \frac{3}{36},\ f(11) = \frac{2}{36},\ f(12) = \frac{1}{36}$.

18. $R_X = \{0, 1, 2, 3, 4, 5\}$;
$f(0) = \frac{59049}{10^5},\ f(1) = \frac{32805}{10^5},\ f(2) = \frac{7290}{10^5},\ f(3) = \frac{810}{10^5},\ f(4) = \frac{45}{10^5},\ f(5) = \frac{1}{10^5}$.

19. $R_X = \{1, 2, 3, 4, 5, 6, 7\}$;
$f(1) = 0.4,\ f(2) = 0.2666,\ f(3) = 0.1666,\ f(4) = 0.0952,\ f(5) = 0.0476,\ f(6) = 0.0190,\ f(7) = 0.0048$.

20. $c = 1$ and $P(X = even) = \frac{1}{4}$.

21. $c = \frac{1}{2},\ P(1 \leq X \leq 2) = \frac{3}{4}$.

22. $c = \frac{3}{2}$ and $P\left(X \leq \frac{1}{2}\right) = \frac{3}{16}$.

CHAPTER 4

1. -0.995.

2. (a) $\frac{1}{33}$, (b) $\frac{12}{33}$, (c) $\frac{65}{33}$.

3. (c) 0.25, (d) 0.75, (e) 0.75, (f) 0.

4. (a) 3.75, (b) 2.6875, (c) 10.5, (d) 10.75, (e) -71.5.

5. (a) 0.5, (b) π, (c) $\frac{3}{10}\pi$.

6. $\frac{17}{24}\frac{1}{\sqrt{\theta}}$.

7. $\sqrt[4]{\frac{1}{E(x^2)}}$.

8. $\frac{8}{3}$.

9. 280.

10. $\frac{9}{20}$.

11. 5.25.

12. $a = \frac{4h^3}{\sqrt{\pi}}$, $E(X) = \frac{2}{h\sqrt{\pi}}$, $Var(X) = \frac{1}{h^2}\left[\frac{3}{2} - \frac{4}{\pi}\right]$.

13. $E(X) = \frac{7}{4}$, $E(Y) = \frac{7}{8}$.

14. $-\frac{2}{38}$.

15. -38.

16. $M(t) = 1 + 2t + 6t^2 + \cdots$.

17. $\frac{1}{4}$.

18. $\beta^n \prod_{i=1}^{n-1}(k+i)$.

19. $\frac{1}{4}\left[3e^{2t} + e^{3t}\right]$.

20. 120.

21. $E(X) = 3,\ Var(X) = 2$.

22. 11.

23. $c = E(X)$.

24. $F(c) = 0.5$.

25. $E(X) = 0,\ Var(X) = 2$.

26. $\frac{1}{625}$.

27. 38.

28. $a = 5$ and $b = -34$ or $a = -5$ and $b = 36$.

29. -0.25.

30. 10.

31. $-\frac{1}{1-p}\, p \ln p$.

CHAPTER 5

1. $\frac{5}{16}$.

2. $\frac{5}{16}$.

3. $\frac{25}{72}$.

4. $\frac{4375}{279936}$.

5. $\frac{3}{8}$.

6. $\frac{11}{16}$.

7. 0.008304.

8. $\frac{3}{8}$.

9. $\frac{1}{4}$.

10. 0.671.

11. $\frac{1}{16}$.

12. 0.0000399994.

13. $\frac{n^2-3n+2}{2^{n+1}}$.

14. 0.2668.

15. $\frac{\binom{6}{3-k}\binom{4}{k}}{\binom{10}{3}}, \quad 0 \leq k \leq 3$.

16. 0.4019.

17. $1 - \frac{1}{e^2}$.

18. $\binom{x-1}{2} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^{x-3}$.

19. $\frac{5}{16}$.

20. 0.22345.

21. 1.43.

22. 24.

23. 26.25.

24. 2.

25. 0.3005.

26. $\frac{4}{e^4-1}$.

27. 0.9130.

28. 0.1239.

CHAPTER 6

1. $f(x) = e^{-x} \qquad 0 < x < \infty.$

2. $Y \sim UNIF(0, 1).$

3. $f(w) = \frac{1}{w\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{\ln w - \mu}{\sigma}\right)^2}.$

4. 0.2313.

5. $3 \ln 4.$

6. $20.1\sigma.$

7. $\frac{3}{4}.$

8. 2.0.

9. 53.04.

10. 44.5314.

11. 75.

12. 0.4649.

13. $\frac{n!}{\theta^n}.$

14. 0.8664.

15. $e^{\frac{1}{2}k^2}.$

16. $\frac{1}{a}.$

17. 64.3441.

18. $g(y) = \begin{cases} \frac{4}{y^3} & \text{if } 0 < y < \sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$

19. 0.5.

20. 0.7745.

21. 0.4.

22. $\frac{2}{3\theta}\, y^{-\frac{1}{3}} e^{-\frac{y^{\frac{2}{3}}}{\theta}}.$

23. $\frac{4}{\sqrt{2\pi}}\, y\, e^{-\frac{y^4}{2}}.$

24. $\ln(X) \sim \bigwedge\!\!\!\backslash(\mu, \sigma^2).$

25. $e^{\mu - \sigma^2}.$

26. $e^{\mu}.$

27. 0.3669.

29. $Y \sim GBETA(\alpha, \beta, a, b).$

32. (i) $\frac{1}{2}\sqrt{\pi}$, (ii) $\frac{1}{2}$, (iii) $\frac{1}{4}\sqrt{\pi}$, (iv) $\frac{1}{2}$.

33. (i) $\frac{1}{180}$, (ii) $(100)^{13}\, \frac{5!\,7!}{13!}$, (iii) $\frac{1}{360}$.

35. $\left(1 - \frac{\alpha}{\beta}\right)^2.$

36. $E(X^n) = \theta^n\, \frac{\Gamma(n+\alpha)}{\Gamma(\alpha)}.$

CHAPTER 7

1. $f_1(x) = \frac{2x+3}{21}$, and $f_2(y) = \frac{3y+6}{21}$.

2. $f(x,y) = \begin{cases} \frac{1}{36} & \text{if } 1 < x < y = 2x < 12 \\ \frac{2}{36} & \text{if } 1 < x < y < 2x < 12 \\ 0 & \text{otherwise.} \end{cases}$

3. $\frac{1}{18}$.

4. $\frac{1}{2e^4}$.

5. $\frac{1}{3}$.

6. $f_1(x) = \begin{cases} 2(1-x) & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$

7. $\frac{(e^2-1)(e-1)}{e^5}$.

8. 0.2922.

9. $\frac{5}{7}$.

10. $f_1(x) = \begin{cases} \frac{5}{48}x(8-x^3) & \text{if } 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$

11. $f_2(y) = \begin{cases} 2y & \text{if } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$

12. $f(y/x) = \begin{cases} \frac{1}{1+\sqrt{1-(x-1)^2}} & \text{if } (x-1)^2 + (y-1)^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases}$

13. $\frac{6}{7}$.

14. $f(y/x) = \begin{cases} \frac{1}{2x} & \text{if } 0 < y < 2x < 1 \\ 0 & \text{otherwise.} \end{cases}$

15. $\frac{4}{9}$.

16. $g(w) = 2e^{-w} - 2e^{-2w}$.

17. $g(w) = \left(1 - \frac{w^3}{\theta^3}\right) \frac{6w^2}{\theta^3}$.

18. $\frac{11}{36}$.

19. $\frac{7}{12}$.

20. $\frac{5}{6}$.

21. No.

22. Yes.

23. $\frac{7}{32}$.

24. $\frac{1}{4}$.

25. $\frac{1}{2}$.

26. $x\,e^{-x}$.

CHAPTER 8

1. 13.

2. $Cov(X,Y) = 0$. Since $0 = f(0,0) \neq f_1(0) f_2(0) = \frac{1}{4}$, X and Y are not independent.

3. $\frac{1}{\sqrt{8}}$.

4. $\frac{1}{(1-4t)(1-6t)}$.

5. $X + Y \sim BIN(n+m, p)$.

6. $\frac{1}{2}\left(X^2 - Y^2\right) \sim EXP(1)$.

7. $M(s,t) = \frac{e^s-1}{s} + \frac{e^t-1}{t}$.

8.

9. $-\frac{15}{16}$.

10. $Cov(X,Y) = 0$. No.

11. $a = \frac{6}{8}$ and $b = \frac{9}{8}$.

12. $Cov = -\frac{45}{112}$.

13. $Corr(X,Y) = -\frac{1}{5}$.

14. 136.

15. $\frac{1}{2}\sqrt{1+\rho}$.

16. $(1 - p + pe^t)(1 - p + pe^{-t})$.

17. $\frac{\sigma^2}{n}\,[1 + (n-1)\rho]$.

18. 2.

19. $\frac{4}{3}$.

20. 1.

21. $\frac{1}{2}$.

CHAPTER 9

1. 6.

2. $\frac{1}{2}(1+x^2)$.

3. $\frac{1}{2}\,y^2$.

4. $\frac{1}{2}+x$.

5. $2x$.

6. $\mu_X = -\frac{22}{3}$ and $\mu_Y = \frac{112}{9}$.

7. $\frac{1}{3}\,\frac{2+3y-28y^3}{1+2y-8y^2}$.

8. $\frac{3}{2}\,x$.

9. $\frac{1}{2}\,y$.

10. $\frac{4}{3}\,x$.

11. 203.

12. $15-\frac{1}{\pi}$.

13. $\frac{1}{12}\,(1-x)^2$.

14. $\frac{1}{12}\,\left(1-x^2\right)^2$.

15. $\frac{5}{192}$.

16. $\frac{1}{12}$.

17. 180.

19. $\frac{x}{6}+\frac{5}{12}$.

20. $\frac{x}{2}+1$.

CHAPTER 10

1. $g(y) = \begin{cases} \frac{1}{2} + \frac{1}{4\sqrt{y}} & \text{for } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$

2. $g(y) = \begin{cases} \frac{3}{16}\,\frac{\sqrt{y}}{m\sqrt{m}} & \text{for } 0 \leq y \leq 4m \\ 0 & \text{otherwise.} \end{cases}$

3. $g(y) = \begin{cases} 2y & \text{for } 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$

4. $g(z) = \begin{cases} \frac{1}{16}\,(z+4) & \text{for } -4 \leq z \leq 0 \\ \frac{1}{16}\,(4-z) & \text{for } 0 \leq z \leq 4 \\ 0 & \text{otherwise.} \end{cases}$

5. $g(z,x) = \begin{cases} \frac{1}{2}\,e^{-x} & \text{for } 0 < x < z < 2 + x < \infty \\ 0 & \text{otherwise.} \end{cases}$

6. $g(y) = \begin{cases} \frac{4}{y^3} & \text{for } 0 < y < \sqrt{2} \\ 0 & \text{otherwise.} \end{cases}$

7. $g(z) = \begin{cases} \frac{z^3}{15000} - \frac{z^2}{250} + \frac{z}{25} & \text{for } 0 \leq z \leq 10 \\ \frac{8}{15} - \frac{2z}{25} - \frac{z^2}{250} - \frac{z^3}{15000} & \text{for } 10 \leq z \leq 20 \\ 0 & \text{otherwise.} \end{cases}$

8. $g(u) = \begin{cases} \frac{4a^2}{u^3}\,\ln\left(\frac{u-a}{a}\right) + \frac{2a\,(u-2a)}{u^2\,(u-a)} & \text{for } 2a \leq u < \infty \\ 0 & \text{otherwise.} \end{cases}$

9. $h(y) = \frac{3z^2-2z+1}{216}, \;\; z = 1, 2, 3, 4, 5, 6.$

10. $g(z) = \begin{cases} \frac{4h^3}{m\sqrt{\pi}}\,\sqrt{\frac{2z}{m}}\,e^{-\frac{2h^2 z}{m}} & \text{for } 0 \leq z < \infty \\ 0 & \text{otherwise.} \end{cases}$

11. $g(u,v) = \begin{cases} -\frac{3u}{350} + \frac{9v}{350} & \text{for } 10 \leq 3u + v \leq 20,\; u \geq 0,\; v \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

12. $g_1(u) = \begin{cases} \frac{2u}{(1+u)^3} & \text{if } 0 \leq u < \infty \\ 0 & \text{otherwise.} \end{cases}$

13. $g(u,v) = \begin{cases} \frac{5\left[9v^3-5u^2v+3uv^2+u^3\right]}{32768} & \text{for } 0 < 2v+2u < 3v-u < 16 \\ \\ 0 & \text{otherwise.} \end{cases}$

14. $g(u,v) = \begin{cases} \frac{u+v}{32} & \text{for } 0 < u+v < 2\sqrt{5v-3u} < 8 \\ \\ 0 & \text{otherwise.} \end{cases}$

15. $g_1(u) = \begin{cases} 2+4u+2u^2 & \text{if } -1 \leq u \leq 0 \\ \\ 2\sqrt{1-4u} & \text{if } 0 \leq u \leq \frac{1}{4} \\ \\ 0 & \text{otherwise.} \end{cases}$

16. $g_1(u) = \begin{cases} \frac{4}{3}\,u & \text{if } 0 \leq u \leq 1 \\ \\ \frac{4}{3}\,u^{-5} & \text{if } 1 \leq u < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$

17. $g_1(u) = \begin{cases} 4u^{\frac{1}{3}} - 4u & \text{if } 0 \leq u \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$

18. $g_1(u) = \begin{cases} 2\,u^{-3} & \text{if } 1 \leq u < \infty \\ \\ 0 & \text{otherwise.} \end{cases}$

19. $f(w) = \begin{cases} \frac{w}{6} & \text{if } 0 \leq w \leq 2 \\ \\ \frac{2}{6} & \text{if } 2 \leq w \leq 3 \\ \\ \frac{5-w}{6} & \text{if } 3 \leq w \leq 5 \\ \\ 0 & \text{otherwise.} \end{cases}$

20. $BIN(2n, p)$

21. $GAM(\theta, 2)$

22. $CAU(0)$

23. $N(2\mu, 2\sigma^2)$

24. $f_1(\alpha) = \begin{cases} \frac{1}{4}\,(2-|\alpha|) & \text{if } |\alpha| \leq 2 \\ \\ 0 & \text{otherwise,} \end{cases} \qquad f_2(\beta) = \begin{cases} -\frac{1}{2}\,\ln(|\beta|) & \text{if } |\beta| \leq 1 \\ \\ 0 & \text{otherwise.} \end{cases}$

CHAPTER 11

2. $\frac{7}{10}$.

3. $\frac{960}{7^5}$.

6. 0.7627.

CHAPTER 12

CHAPTER 13

3. 0.115.

4. 1.0.

5. $\frac{7}{16}$.

6. 0.352.

7. $\frac{6}{5}$.

8. 101.282.

9. $\frac{1+\ln(2)}{2}$.

10. $[1 - F(x_6)]^5$.

11. $\theta + \frac{1}{5}$.

12. $2\,e^{-2w}$.

13. $6\,\frac{w^2}{\theta^3}\left(1 - \frac{w^3}{\theta^3}\right)$.

14. $N(0,1)$.

15. 25.

16. X has a degenerate distribution with MGF $M(t) = e^{\frac{1}{2}t}$.

17. $POI(1995\lambda)$.

18. $\left(\frac{1}{2}\right)^n (n+1)$.

19. $\frac{8^8}{11^9}\,35$.

20. $f(x) = \frac{60}{\theta}\left(1 - e^{-\frac{x}{\theta}}\right)^3 e^{-\frac{3x}{\theta}}$ for $0 < x < \infty$.

21. $X_{(n+1)} \sim Beta(n+1, n+1)$.

CHAPTER 14

1. $N(0, 32)$.

2. $\chi^2(3)$; the MGF of $X_1^2 - X_2^2$ is $M(t) = \frac{1}{\sqrt{1-4t^2}}$.

3. $t(3)$.

4. $f(x_1, x_2, x_3) = \frac{1}{\theta^3} e^{-\frac{(x_1+x_2+x_3)}{\theta}}$.

5. σ^2

6. $t(2)$.

7. $M(t) = \frac{1}{\sqrt{(1-2t)(1-4t)(1-6t)(1-8t)}}$.

8. 0.625.

9. $\frac{\sigma^4}{n^2} 2(n-1)$.

10. 0.

11. 27.

12. $\chi^2(2n)$.

13. $t(n+p)$.

14. $\chi^2(n)$.

15. $(1, 2)$.

16. 0.84.

17. $\frac{2\sigma^2}{n^2}$.

18. 11.07.

19. $\chi^2(2n-2)$.

20. 2.25.

21. 6.37.

CHAPTER 15

1. $\sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}$.

2. $\frac{1}{1-\bar{X}}$.

3. $\frac{2}{\bar{X}}$.

4. $-\frac{n}{\sum_{i=1}^{n} \ln X_i}$.

5. $\frac{n}{\sum_{i=1}^{n} \ln X_i} - 1$.

6. $\frac{2}{\bar{X}}$.

7. 4.2

8. $\frac{19}{26}$.

9. $\frac{15}{4}$.

10. 2.

11. $\hat{\alpha} = 3.534$ and $\hat{\beta} = 3.409$.

12. 1.

13. $\frac{1}{3}\max\{x_1, x_2, ..., x_n\}$.

14. $\sqrt{1 - \frac{1}{\max\{x_1, x_2, ..., x_n\}}}$.

15. 0.6207.

18. 0.75.

19. $-1 + \frac{5}{\ln(2)}$.

20. $\frac{\bar{X}}{1+\bar{X}}$.

21. $\frac{\bar{X}}{4}$.

22. 8.

23. $\frac{n}{\sum_{i=1}^{n} |X_i - \mu|}$.

24. $\frac{1}{N}$.

25. $\sqrt{\overline{X}}$.

26. $\hat{\lambda} = \frac{n\bar{X}}{(n-1)S^2}$ and $\hat{\alpha} = \frac{n\bar{X}^2}{(n-1)S^2}$.

27. $\frac{10\,n}{p\,(1-p)}$.

28. $\frac{2n}{\theta^2}$.

29. $\begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$.

30. $\begin{pmatrix} \frac{n\lambda}{\mu^3} & 0 \\ 0 & \frac{n}{2\lambda^2} \end{pmatrix}$.

31. $\widehat{\alpha} = \frac{\overline{X}}{\widehat{\beta}}, \quad \widehat{\beta} = \frac{1}{\overline{X}} \left[\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X} \right]$.

32. $\widehat{\theta}$ is obtained by solving numerically the equation $\sum_{i=1}^{n} \frac{2(x_i-\theta)}{1+(x_i-\theta)^2} = 0$.

33. $\widehat{\theta}$ is the median of the sample.

34. $\frac{n}{\lambda}$.

35. $\frac{n}{(1-p)\,p^2}$.

CHAPTER 16

1. $b = \frac{\sigma_2^2 - cov(T_1,T_2)}{\sigma_1^2+\sigma_2^2-2cov(T_1,T_2}$.

2. $\widehat{\theta} = \overline{|X|}$, $E(\overline{|X|}) = \theta$, unbiased.

4. $n = 20$.

5. $k = 2$.

6. $a = \frac{25}{61}$, $b = \frac{36}{61}$, $\widehat{c} = 12.47$.

7. $\displaystyle\sum_{i=1}^{n} X_i^3$.

8. $\displaystyle\sum_{i=1}^{n} X_i^2$, no.

10. $k = 2$.

11. $k = 2$.

13. $\displaystyle\ln \prod_{i=1}^{n} (1 + X_i)$.

14. $\displaystyle\sum_{i=1}^{n} X_i^2$.

15. $X_{(1)}$, and sufficient.

16. $X_{(1)}$ is biased and $\overline{X} - 1$ is unbiased. $X_{(1)}$ is efficient then $\overline{X} - 1$.

17. $\displaystyle\sum_{i=1}^{n} \ln X_i$.

18. $\displaystyle\sum_{i=1}^{n} X_i$.

19. $\displaystyle\sum_{i=1}^{n} \ln X_i$.

22. Yes.

23. Yes.

24. Yes.

25. Yes.

CHAPTER 17

CHAPTER 18

1. $\alpha = 0.03125$ and $\beta = 0.76$.

2. Do not reject H_o.

3. $\alpha = 0.0511$ and $\beta(\lambda) = 1 - \sum_{x=0}^{7} \frac{(8\lambda)^x e^{-8\lambda}}{x!}, \; \lambda \neq 0.5$.

4. $\alpha = 0.08$ and $\beta = 0.46$.

5. $\alpha = 0.19$.

6. $\alpha = 0.0109$.

7. $\alpha = 0.0668$ and $\beta = 0.0062$.

8. $C = \{(x_1, x_2) \,|\, \overline{x}^2 \geq 3.9395\}$.

9. $C = \{(x_1, ..., x_{10}) \,|\, \overline{x} \geq 0.3\}$.

10. $C = \{x \in [0, 1] \,|\, x \geq 0.829\}$.

11. $C = \{(x_1, x_2) \,|\, x_1 + x_2 \geq 5\}$.

12. $C = \{(x_1, ..., x_8) \,|\, \overline{x} - \overline{x} \ln \overline{x} \leq a\}$.

13. $C = \{(x_1, ..., x_n) \,|\, 35 \ln \overline{x} - \overline{x} \leq a\}$.

14.

15.

16.

17.

18.

19.

20.