Multivariate Data Analysis

Course Taught at SUAD

Session 3: Probability & Sampling Distributions

Dr. Tanujit Chakraborty

Assistant Professor of Statistics

Sorbonne University

Today's Topics…

- Probability vs. Statistics
- Concept of random variable
- Probability distribution concept
- Discrete probability distribution
- Continuous probability distribution
- Central Limit Theorem
- Standard Sampling Distributions

Probability and Statistics

Probability is the chance of an outcome in an experiment (also called event).

Event: Tossing a fair coin
Outcome: Head, Tail

Probability deals with **predicting** the likelihood of **future** events.	Statistics involves the **analysis** of the **frequency** of **past** events

Example: Consider there is a drawer containing 100 socks: 30 red, 20 blue and 50 black socks.

We can use probability to answer questions about the selection of a random sample of these socks.

- **PQ1.** What is the probability that we draw two blue socks or two red socks from the drawer?
- **PQ2.** What is the probability that we pull out three socks or have matching pair?
- **PQ3.** What is the probability that we draw five socks and they are all black?

Statistics

Instead, if we have no knowledge about the type of socks in the drawers, then we enter into the realm of statistics. Statistics helps us to infer properties about the population on the basis of the random sample.

Questions that would be statistical in nature are:

- **Q1**: A random sample of 10 socks from the drawer produced one blue, four red, five black socks. What is the total population of black, blue or red socks in the drawer?
- **Q2**: We randomly sample 10 socks, and write down the number of black socks and then return the socks to the drawer. The process is done for five times. The mean number of socks for each of these trial is 7. What is the true number of black socks in the drawer?
- etc.

Probability vs. Statistics

In other words:

- In probability, we are given a model and asked what kind of data we are likely to see.
- In statistics, we are given data and asked what kind of model is likely to have generated it.

Example: Measles Study

- A study on health is concerned with the incidence of childhood measles in parents of childbearing age in a city. For each couple, we would like to know how likely, it is that either the mother or father or both have had childhood measles.
- The current census data indicates that 20% adults between the ages 17 and 35 (regardless of sex) have had childhood measles.
  - This give us the probability that an individual in the city has had childhood measles.

Defining Random Variable

Definition: **Random Variable**

A random variable is a rule that assigns a numerical value to an outcome of interest.

Example : In "measles Study", we define a random variable X as the number of parents in a married couple who have had childhood measles.

This random variable can take values of $0, 1$ *and* 2.

Note:

- Random variable is not exactly the same as the variable defining a data.
- The probability that the random variable takes a given value can be computed using the rules governing probability.
  - For example, the probability that $X = 1$ means either mother or father but not both has had measles is 0.32. Symbolically, it is denoted as **P(X=1) = 0.32.**

Probability Distribution

Definition : **Probability distribution**

A probability distribution is a definition of probabilities of the values of random variable.

Example : Given that 0.2 is the probability that a person (in the ages between 17 and 35) has had childhood measles. Then the probability distribution is given by

X	Probability
0	0.64
1	0.32
2	0.04

?

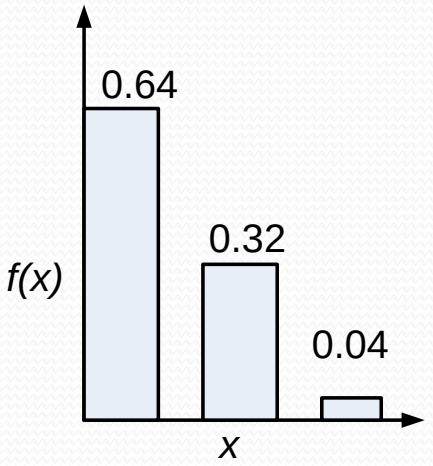
Probability Distribution

- In data analytics, the probability distribution is important with which many statistics making inferences about population can be derived .
  - In general, a probability distribution function takes the following form

x	x_1	$x_2 \dots \dots \dots \dots \dots x_n$
$f(x) = P(X = x)$	$f(x_1)$	$f(x_2) \dots \dots \dots f(x_n)$

Example: Measles Study

x	0	1	2
$f(x)$	0.64	0.32	0.04



Usage of Probability Distribution

- Distribution (discrete/continuous) function is widely used in simulation studies.
  - A simulation study uses a computer to simulate a real phenomenon or process as closely as possible.
  - The use of simulation studies can often eliminate the need of costly experiments and is also often used to study problems where actual experimentation is impossible.

Examples :

1) A study involving testing the effectiveness of a new drug, the number of cured patients among all the patients who use such a drug approximately follows a binomial distribution.

2) Operation of ticketing system in a busy public establishment (e.g., airport), the arrival of passengers can be simulated using Poisson distribution.

Binomial Distribution

- In many situations, an outcome has only two outcomes: success and failure.
  - Such outcome is called dichotomous outcome.
- An experiment when consists of repeated trials, each with dichotomous outcome is called Bernoulli process. Each trial in it is called a Bernoulli trial.

Example : Firing bullets to hit a target.

- Suppose, in a Bernoulli process, we define a random variable $X \equiv$ the number of successes in trials.
- Such a random variable obeys the binomial probability distribution, if the experiment satisfies the following conditions:

1) The experiment consists of n trials.
2) Each trial results in one of two mutually exclusive outcomes, one labelled a *"success"* and the other a *"failure"*.
3) The probability of a success on a single trial is equal to $\boldsymbol{p}$. The value of p remains constant throughout the experiment.
4) The trials are independent.

Defining Binomial Distribution

Definition: **Binomial distribution**

The function for computing the probability for the binomial probability distribution is given by

$$f(x) = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x}$$

$$for\ x = 0, 1, 2, \ldots, n$$

Here, $f(x) = P(X = x)$, where X denotes "the number of success" and $X = x$ denotes the number of success in n trials.

Binomial Distribution

Example : Measles study

X = having had childhood measles a success

p = 0.2, the probability that a parent had childhood measles

n = 2, here a couple is an experiment and an individual in a trial, and the number of trials is two.

Thus,

$$P(x = 0) = \frac{2!}{0!(2-0)!}(0.2)^0(0.8)^{2-0} \quad = \mathbf{0.64}$$

$$P(x = 1) = \frac{2!}{1!\,(2-1)!}(0.2)^1(0.8)^{2-1} = \mathbf{0.32}$$

$$P(x = 2) = \frac{2!}{2!\,(2-2)!}(0.2)^2(0.8)^{2-2} = \mathbf{0.04}$$

Binomial Distribution

Example : Verify with real-life experiment

Suppose, 10 pairs of random numbers are generated by a computer (Monte-Carlo method)

15 38 68 39 49 54 19 79 38 14

If the value of the digit is 0 or 1, the outcome is “had childhood measles”, otherwise, (digits 2 to 9), the outcome is “did not”.

For example, in the first pair (i.e., 15), representing a couple and for this couple, $x = 1$. The frequency distribution, for this sample is

x	0	1	2
$f(x)=P(X=x)$	0.7	0.3	0.0

Note: This has close similarity with binomial probability distribution!

Exercise:

- The ***Los Angeles Times*** **(December 13, 1992)** reported that what airline passengers like to do most on long flights is rest or sleep; in a survey of 3697 passengers, almost 80% did so. Suppose that for a particular route the actual percentage is exactly 80%, and consider randomly selecting six passengers.

a. Calculate $p(4)$, and interpret this probability.

b. Calculate $p(6)$, the probability that all six selected passengers rested or slept.

c. Determine $P(X \geq 4)$.

The Multinomial Distribution

The binomial experiment becomes a multinomial experiment, if we let each trial has more than two possible outcome.

Definition: **Multinomial distribution**

If a given trial can result in the ***k*** outcomes $E_1, E_2, \dots\dots, E_k$ with probabilities $p_1, p_2, \dots\dots, p_k$, then the probability distribution of the random variables $X_1, X_2, \dots\dots, X_k$ representing the number of occurrences for $E_1, E_2, \dots\dots, E_k$ in n independent trials is

$$f(x_1, x_2, \dots\dots, x_k) = \binom{n}{x_1, x_2, \dots\dots, x_k} {p_1}^{x_1} {p_2}^{x_2} \dots\dots {p_k}^{x_k}$$

where $\binom{n}{x_1, x_2, \dots\dots, x_k} = \frac{n!}{x_1! x_2! \dots\dots x_k!}$

$$\sum_{i=1}^{k} x_i = n \text{ and } \sum_{i=1}^{k} p_i = 1$$

The Poisson Distribution

There are some experiments, which involve the occurring of the number of outcomes during a given time interval (or in a region of space).

Such a process is called Poisson process.

Example :

Number of clients visiting a ticket selling counter in a metro station.

The Poisson Distribution

Properties of Poisson process

- The number of outcomes in one time interval is independent of the number that occurs in any other disjoint interval [Poisson process has no memory]
- The probability that a single outcome will occur during a very short interval is proportional to the length of the time interval and does not depend on the number of outcomes occurring outside this time interval.
- The probability that more than one outcome will occur in such a short time interval is negligible.

Definition : **Poisson distribution**

The probability distribution of the Poisson random variable X, representing the number of outcomes occurring in a given time interval t, is

$$f(x,\lambda t) = P(X = x) = \frac{e^{-\lambda t}.(\lambda t)^x}{x!}, x = 0, 1, \dots \dots$$

where λ is the average number of outcomes per unit time and $e = 2.71828 \dots$

Examples

Suppose that the number of telephone calls coming into a telephone exchange between 10 A.M and 11 A.M. say, X_1 is a random variable with Poisson distribution with parameter 2. Similarly the number of calls arriving between 11 A.M. to 12 noon, say, X_2 has a Poisson distribution with parameter 6. If X_1 and X_2 are independent, what is the probability that more than 5 calls come in-between 10 A.M. and 12 noon?

Descriptive measures

Given a random variable X in an experiment, we have denoted $f(x) = P(X = x)$, the probability that $X = x$. For discrete events $f(x) = 0$ for all values of x except $x = 0, 1, 2, \ldots\ldots$

Properties of discrete probability distribution

1. $0 \leq f(x) \leq 1$
2. $\sum f(x) = 1$
3. $\mu = \sum x.f(x)$ [is the mean]
4. $\sigma^2 = \sum(x - \mu)^2 . f(x)$ [is the variance]

In $2, 3$ *and* 4, summation is extended for all possible discrete values of x.

Note: For discrete uniform distribution, $f(x) = \frac{1}{n}$ with $x = 1, 2, \ldots\ldots, n$

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

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

Descriptive measures

Binomial distribution

The binomial probability distribution is characterized with p (the probability of success) and n (is the number of trials). Then

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

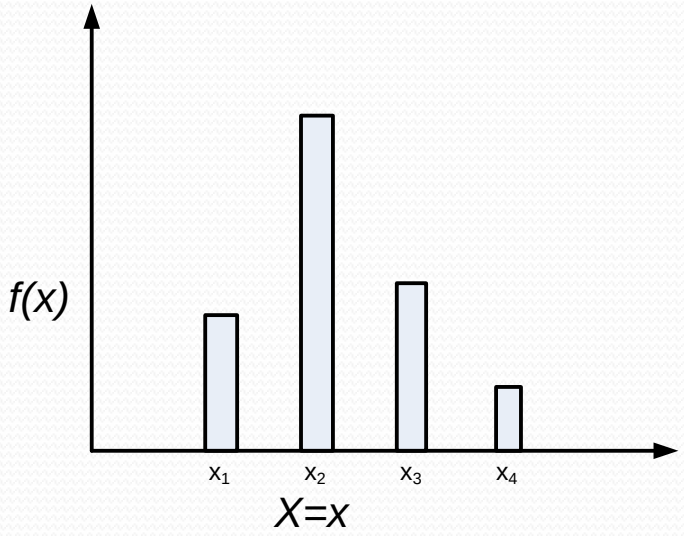
Poisson Distribution

The Poisson distribution is characterized with λ where $\lambda = the\ mean\ of\ outcomes$ and $t = time\ interval$.

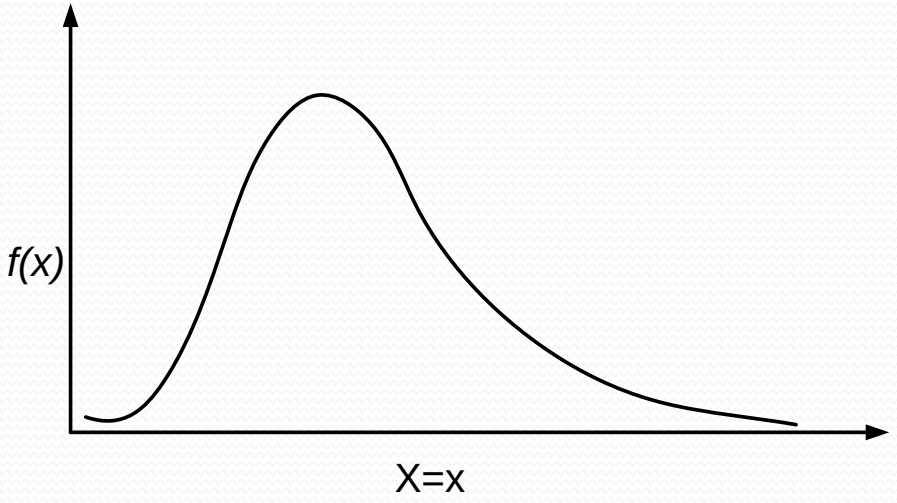
$$\mu = \lambda t$$

$$\sigma^2 = \lambda t$$

Discrete Vs. Continuous Probability Distributions



Discrete Probability distribution



Continuous Probability Distribution

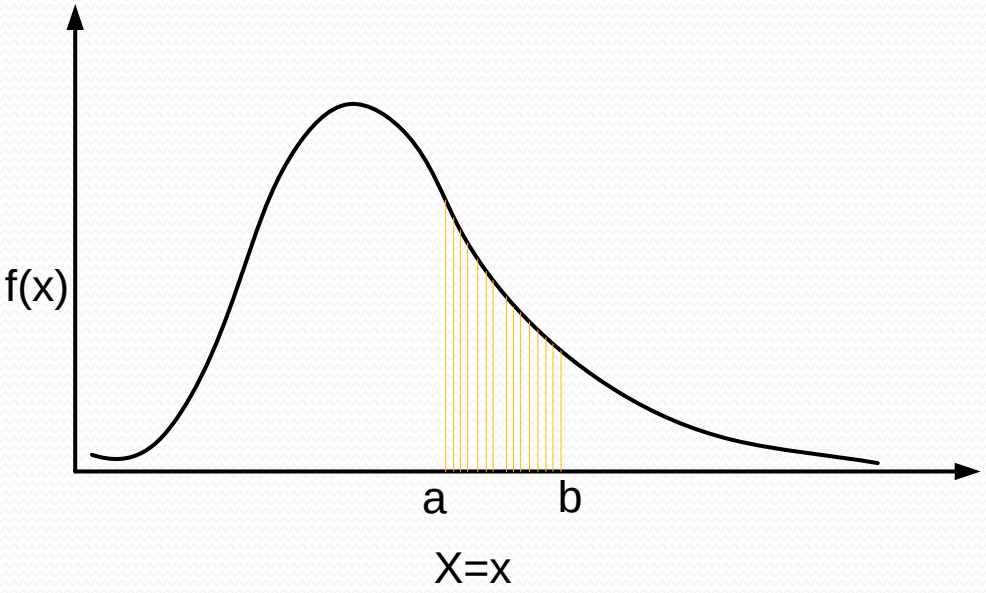
Continuous Probability Distributions

- When the random variable of interest can take any value in an interval, it is called continuous random variable.
  - Every continuous random variable has an infinite, uncountable number of possible values (i.e., any value in an interval)
- Consequently, continuous random variable differs from discrete random variable.

Properties of Probability Density Function

The function $f(x)$ is a probability density function for the continuous random variable X, defined over the set of real numbers R, if

1. $f(x) \geq 0$, for all $x \in R$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$
4. $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
5. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$



Continuous Uniform Distribution

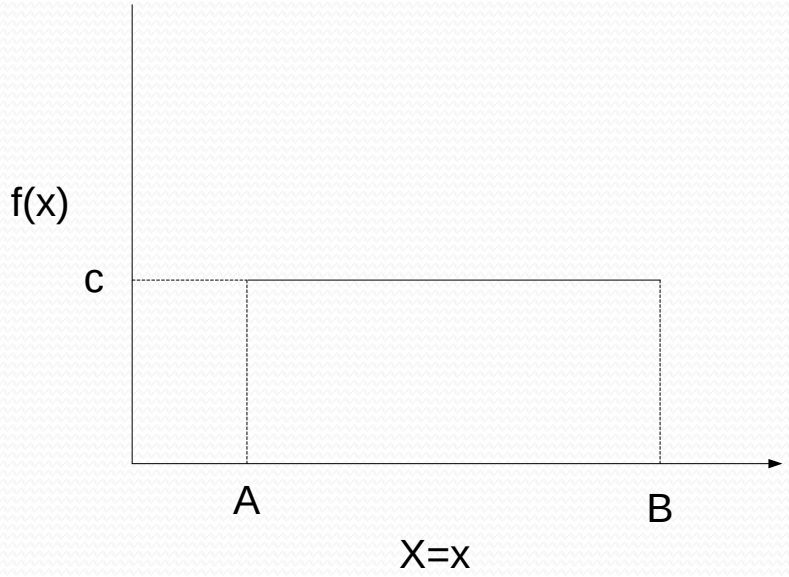
- One of the simplest continuous distribution in all of statistics is the continuous uniform distribution.

Definition : **Continuous Uniform Distribution**

The density function of the continuous uniform random variable X on the interval $[A, B]$ is:

$$f(x: A, B) = \begin{cases} \dfrac{1}{B - A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$$

Continuous Uniform Distribution



Note:

a) $\int f(x)dx = \frac{1}{B-A} \times (B - A) = 1$

b) $P(c < x < d) = \frac{d-c}{B-A}$ where both c and d are in the interval (A,B)

c) $\mu = \frac{A+B}{2}$

d) $\sigma^2 = \frac{(B-A)^2}{12}$

Example

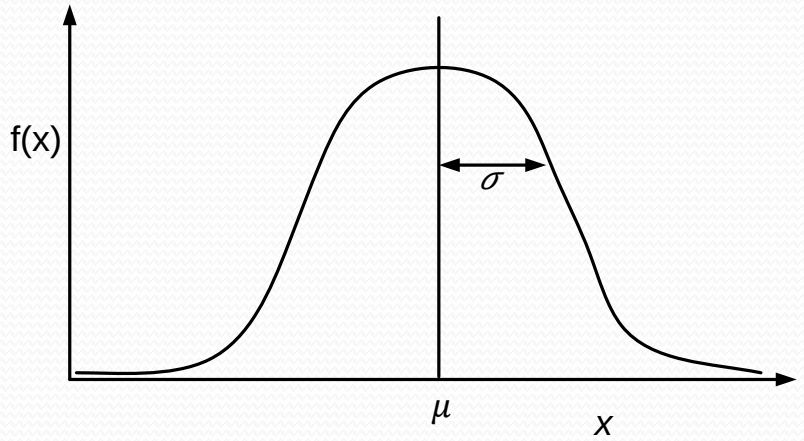
Suppose a train arrives at a subway station regularly every 20 min. If a passenger arrives at the station without knowing the timetable, then find the probability that the man will have to wait at least 10 min? What is the average waiting time ?

Normal Distribution

- The most often used continuous probability distribution is the normal distribution; it is also known as **Gaussian** distribution.
- It's graph called the normal curve is the bell-shaped curve.
- Such a curve approximately describes many phenomenon occur in nature, industry and research.
  - Physical measurement in areas such as meteorological experiments, rainfall studies and measurement of manufacturing parts are often more than adequately explained with normal distribution.
- A continuous random variable X having the bell-shaped distribution is called a normal random variable.

Normal Distribution

- The mathematical equation for the probability distribution of the normal variable depends upon the two parameters μ and σ, its mean and standard deviation.



Definition : **Normal distribution**

The density of the normal variable x with mean μ and variance σ^2 is

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

where $\pi = 3.14159$ ... and $e = 2.71828$ ….., the Naperian constant

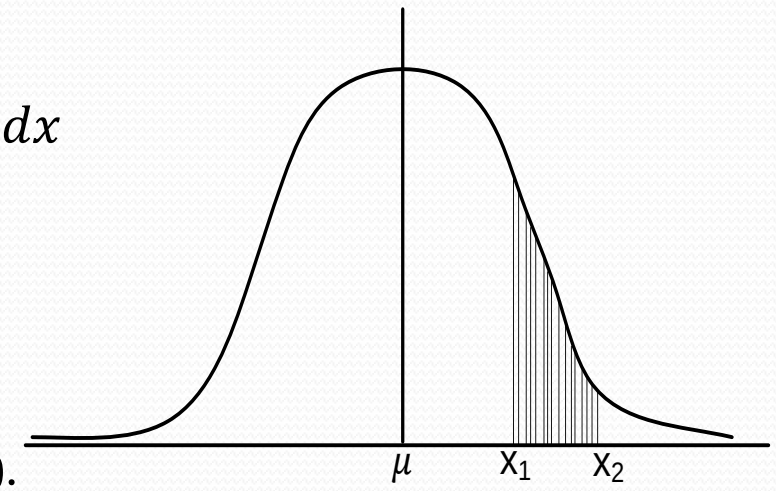
Properties of Normal Distribution

- The curve is symmetric about a vertical axis through the mean μ.
- The random variable x can take any value from $-\infty \; to \; \infty$.
- The most frequently used descriptive parameter s define the curve itself.
- The mode, which is the point on the horizontal axis where the curve is a maximum occurs at $x = \mu$.
- The total area under the curve and above the horizontal axis is equal to 1.

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = 1$$

- $\mu = \int_{-\infty}^{\infty} x.f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x.e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$
- $\sigma^2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2.e^{-\frac{1}{2}[(x-\mu)/\sigma^2]} dx$
- $P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$

denotes the probability of x in the interval (x_1, x_2).

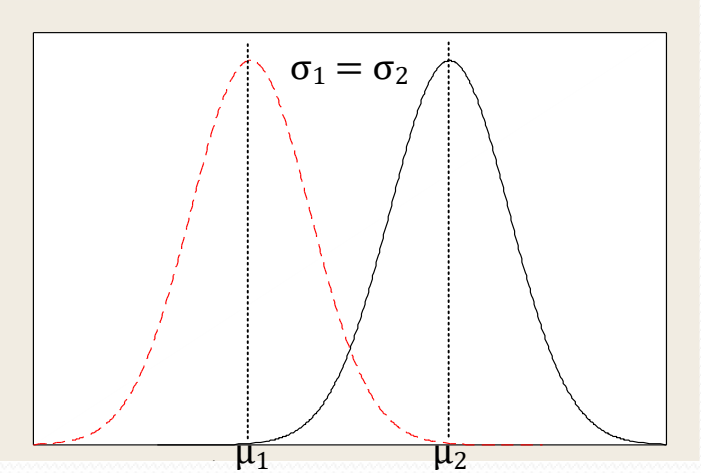


Examples

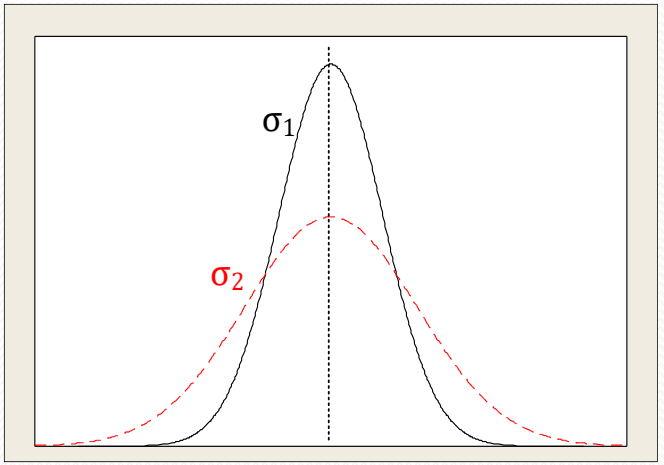
- There are 600 data science students in the under graduate classes of a university, and the probability for any student to need a copy of a particular book from the university library on any day is 0.05. How many copies of the book should be kept in the university library so that the probability may be grater than 0.90 that none of the students needing a copy from the library has to come back disappointed?

- Emissions of nitrogen oxides, which are major constituents of smog, can be modelled using a normal distribution. Let X denote the amount of this pollutant emitted by a randomly selected vehicle (in parts per billion). The distribution of X can be described by a normal distribution with mean 1.6 and standard deviation 0.4. Suppose that the EPA wants to offer some sort of incentive to get the worst polluters off the road. What emission levels constitute the worst 10% of the vehicles?

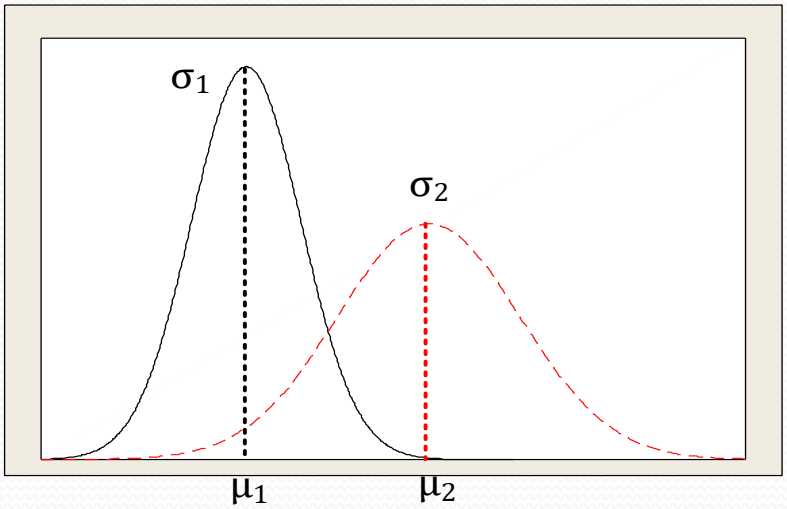
Normal Distribution



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$



Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$



Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$

Chebyshev's Rule

- The mean and standard deviation can be combined to make informative statements about how the values in a data set are distributed and about the relative position of a particular value in a data set.
- To do this, it is useful to be able to describe how far away a particular observation is from the mean in terms of the standard deviation.
- For example, we might say that an observation is 2 standard deviations above the mean or that an observation is 1.3 standard deviations below the mean.
- Sometimes in published articles, the mean and standard deviation are reported, but a graphical display of the data is not given.
- However, using a result called Chebyshev's Rule, it is possible to get a sense of the distribution of data values based on our knowledge of only the mean and standard deviation.

Chebyshev's Inequality

If X is a r.v with mean μ and variance σ^2, then for any positive number k, we have

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

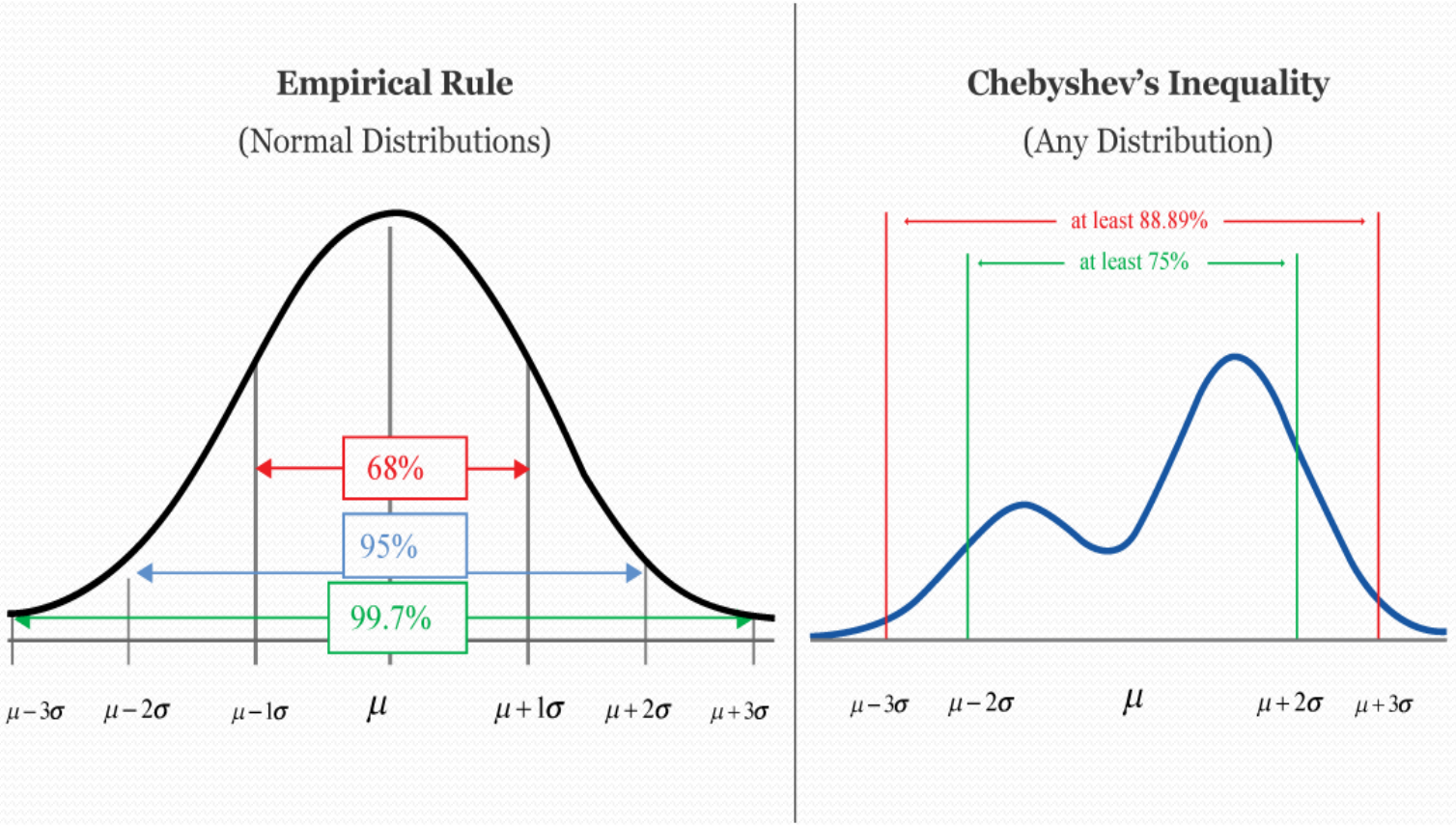
$$P(|X-\mu| < c) \geq 1-\frac{\sigma^2}{c^2}, or, P(|X-\mu| < k\sigma) \geq 1-\frac{1}{k^2}.$$

No. of S.D, k	$1-\frac{1}{k^2}$	% within k S.D of the Mean
2	$1-\frac{1}{4}$ =0.75	At least 75%
3	0.89	At least 89%
4	0.94	At least 94%
4.472	0.95	At least 95%
5	0.96	At least 96%
10	0.99	At least 99%

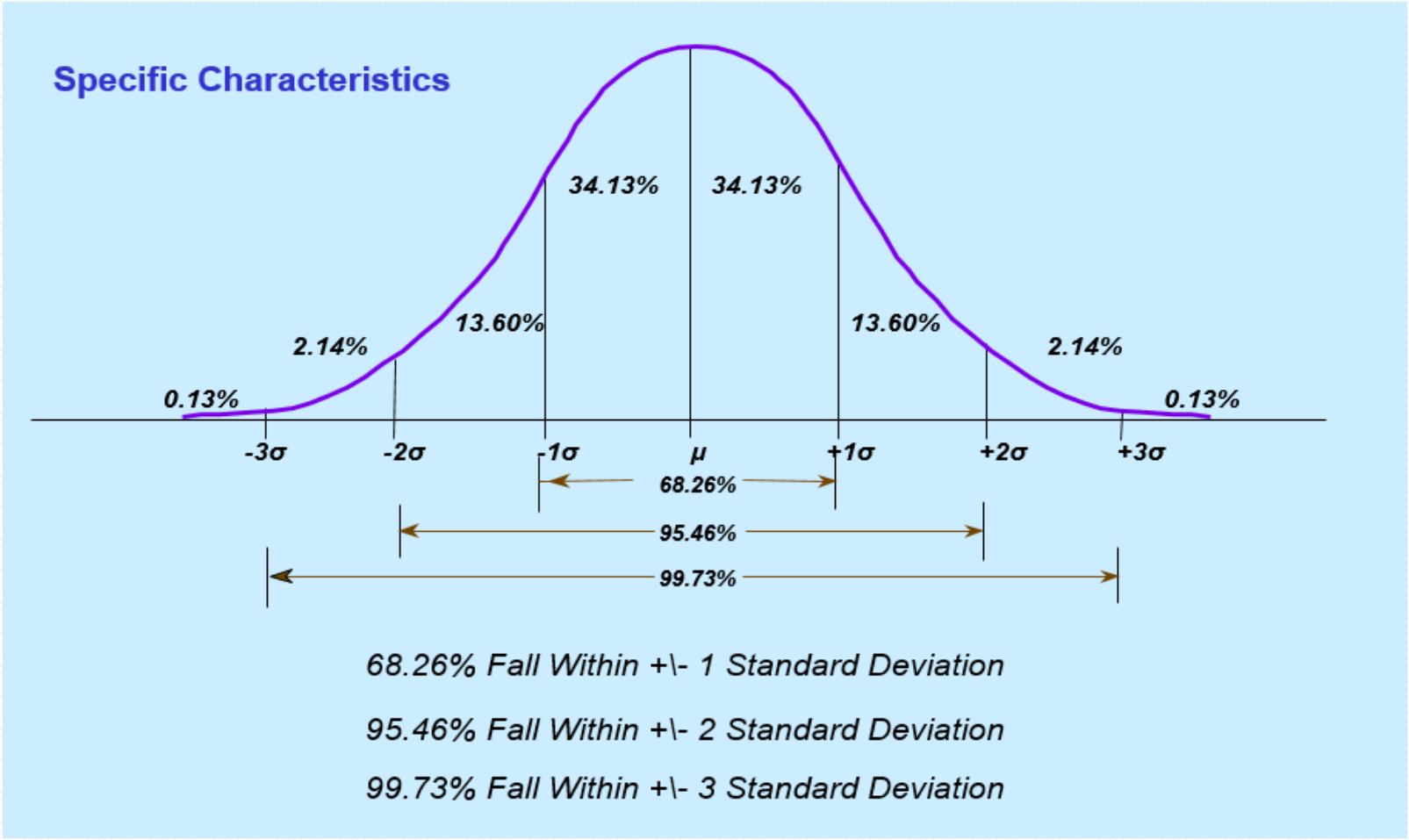
Empirical Rule

- The fact that statements based on Chebyshev's Rule are frequently conservative suggests that we should look for rules that are less conservative and more precise.
- One useful rule is the **Empirical Rule**, which can be applied whenever the distribution of data values can be reasonably well described by a normal curve.
- The Empirical Rule : If the histogram of values in a data set can be reasonably well approximated by a normal curve, then
  - Approximately 68% of the observations are within 1 standard deviation of the mean.
  - Approximately 95% of the observations are within 2 standard deviations of the mean.
  - Approximately 99.7% of the observations are within 3 standard deviations of the mean.

Chebyshev's Inequality VS. Empirical Rule



Normal Curve (6-sigma)



What happens when X follows any continuous distribution? (Chebyshev's Inequality)

Z score

- The ***z* score** corresponding to a particular value is

$$z\ score = \frac{value - mean}{standard\ deviation}$$

- The *z* score tells us how many standard deviations the value is from the mean.
- It is positive or negative according to whether the value lies above or below the mean.
- The process of subtracting the mean and then dividing by the standard deviation is sometimes referred to as *standardization*, and a *z* score is one example of what is called a *standardized score*.

Examples

- A symmetric die is thrown 600 times. Find the lower bound for the probability of getting 80 to 120 sixes.

- Let $f(x) = \begin{matrix} \frac{2}{3}x, 1 \leq x \leq 2 \\ 0, \quad elsewhere \end{matrix}$

Give a bound using Chebyshev's for $P\left(\frac{10}{9} \leq X \leq 2\right)$. Calculate the actual probability. How do they compare?

Example

- A student took two national aptitude tests. The national average and standard deviation were 475 and 100, respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use ***z* scores** to determine on which exam the student performed better relative to the other test takers.
- A sample of concrete specimens of a certain type is selected, and the compressive strength of each specimen is determined. The mean and standard deviation are calculated as $\bar{x} = 3000$ and $s = 500$, and the sample histogram is found to be well approximated by a normal curve.
  - Approximately what percentage of the sample observations are between 2500 and 3500?
  - Approximately what percentage of sample observations are outside the interval from 2000 to 4000?
  - What can be said about the approximate percentage of observations between 2000 and 2500?
  - Why would you not use Chebyshev's Rule to answer the questions posed in Parts (a)–(c)?

Standard Normal Distribution

- The normal distribution has computational complexity to calculate $P(x_1 < x < x_2)$ for any two (x_1, x_2) and given μ and σ
- To avoid this difficulty, the concept of z-transformation is followed.

$$z = \frac{x-\mu}{\sigma} \qquad \text{[Z-transformation]}$$

- X: Normal distribution with mean μ and variance σ^2.
- Z: Standard normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.
- Therefore, if *f(x)* assumes a value, then the corresponding value of $f(z)$ is given by

$$f(x:\mu,\sigma): \; P(x_1 < x < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

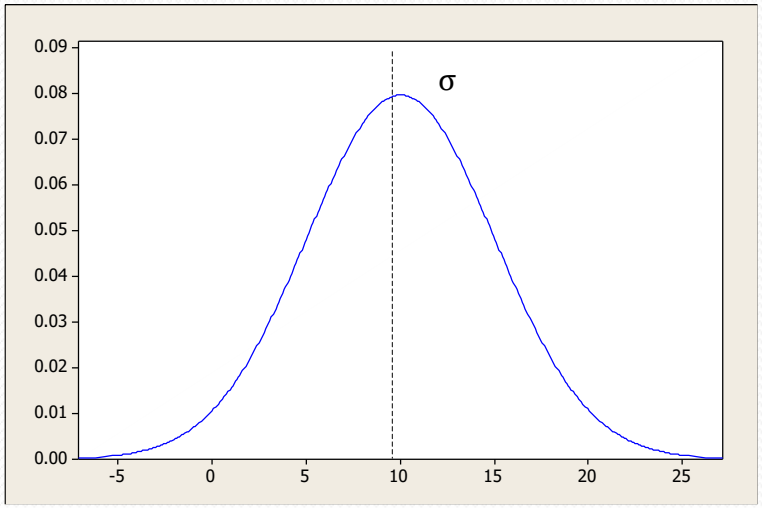
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} dz$$

$$= f(z:0,\sigma)$$

Standard Normal Distribution

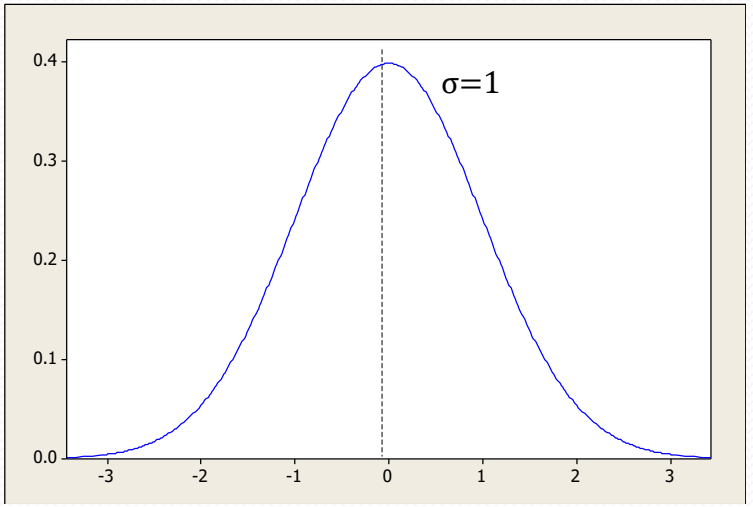
Definition : **Standard normal distribution**

The distribution of a normal random variable with mean 0 and variance 1 is called a standard normal distribution.



$x=\mu$

$f(x: \mu, \sigma)$



$\mu=0$

$f(z: 0, 1)$

Question: Using Standard Normal Distribution, show that $\Gamma(1/2)=\sqrt{1/2}$

Sampling Distributions

Random Sampling:

- The outcome of a statistical experiment may be recorded either as a numerical value or as a descriptive representation.

- Here we focus on sampling from distributions or populations and study such important quantities as the *sample mean* and *sample variance*.

Population

- A **population** consists of the totality of the observations with which we are concerned.
- The number of observations in the population is defined to be the size of the population.
- Each observation in a population is a value of a random variable X having some probability distribution $f(x)$.
- Hence, the mean and variance of a random variable or probability distribution are also referred to as the mean and variance of the corresponding population.

Sample

- In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible to observe the entire set of observations that make up the population.

- This brings us to consider the notion of sampling.

- A **sample** is a subset of a population.

- All too often we are tempted to choose a sample by selecting the most convenient members of the population.

- Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be **biased**.

Random Sampling

- To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.
- In selecting a random sample of size n from a population $f(x)$, let us define the random variable $X_i, i = 1, 2, \ldots, n$, to represent the ith measurement or sample value that we observe.
- The random variables $X_1, X_2, \ldots, X_n$ will then constitute a random sample from the population $f(x)$ with numerical values $x_1, x_2, \ldots, x_n$ if the measurements are obtained by repeating the experiment n independent times under essentially the same conditions.
- Because of the identical conditions under which the elements of the sample are selected, it is reasonable to assume that the n random variables $X_1, X_2, \ldots, X_n$ are independent and that each has the same probability distribution $f(x)$.
- The probability distributions of $X_1, X_2, \ldots, X_n$ are, respectively, $f(x_1), f(x_2), \ldots, f(x_n)$, and their joint probability distribution is $f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n)$.

Some important Statistics

- Our main purpose in selecting random samples is to elicit information about the unknown population parameters.

- Any function of the random variables constituting a random sample is called a **statistic**.

- Let $X_1, X_2, \ldots, X_n$ represent *n* random variables.
  - Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.
  - Sample Variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. The computed value of S^2 for a given sample is denoted by s^2.

Sampling Distribution

More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistics.

Definition : Sampling distribution

The sampling distribution of a statistic is the probability distribution of that statistic.

- The probability distribution of sample mean (hereafter, will be denoted as $\bar{X}$) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like $\bar{X}$, we call sampling distribution of variance (denoted as S^2).
- Using the values of $\bar{X}$ and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution

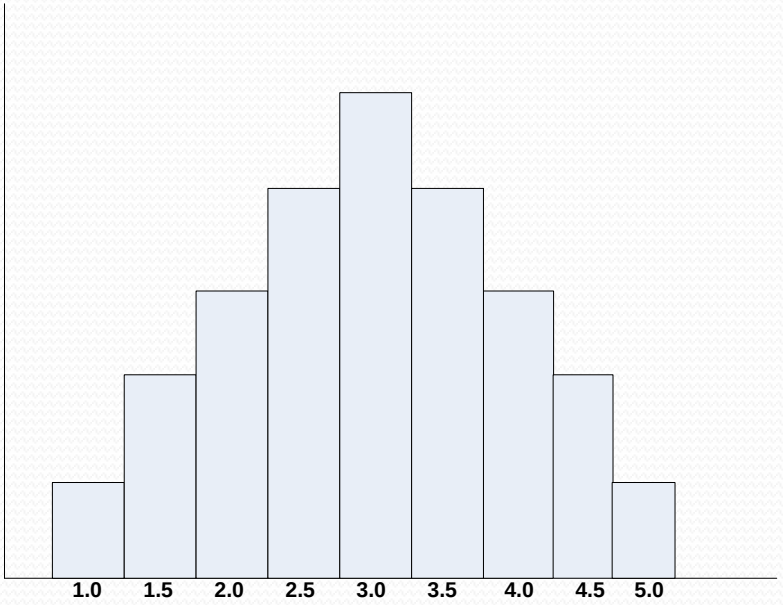
Example 1: Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls. Following table lists all possible samples and their mean.

Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence probability distribution of sample statistics.

2. The sampling distribution of a statistic depends on
   - the size of the population
   - the size of the samples and
   - the method of choosing the samples.

?

Theorem on Sampling Distribution

Theorem 1: **Sampling distribution of mean and variance**

The sampling distribution of a random sample of size n drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $\frac{\sigma^2}{n} = V(\bar{X})$.

Example 2: With reference to data in Example 1

For the population, $\mu = \frac{1+2+3+4+5}{5} = 3$

$$\sigma^2 = \frac{(25-1)}{12} = 2$$

Applying the theorem, we have $\bar{X} = 3$ *and* $\mathrm{V}(\bar{X}) = 1$.

Hence, the theorem is verified!

Central Limit Theorem

Theorem 1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

This further, can be established with the famous "central limit theorem", which is stated below.

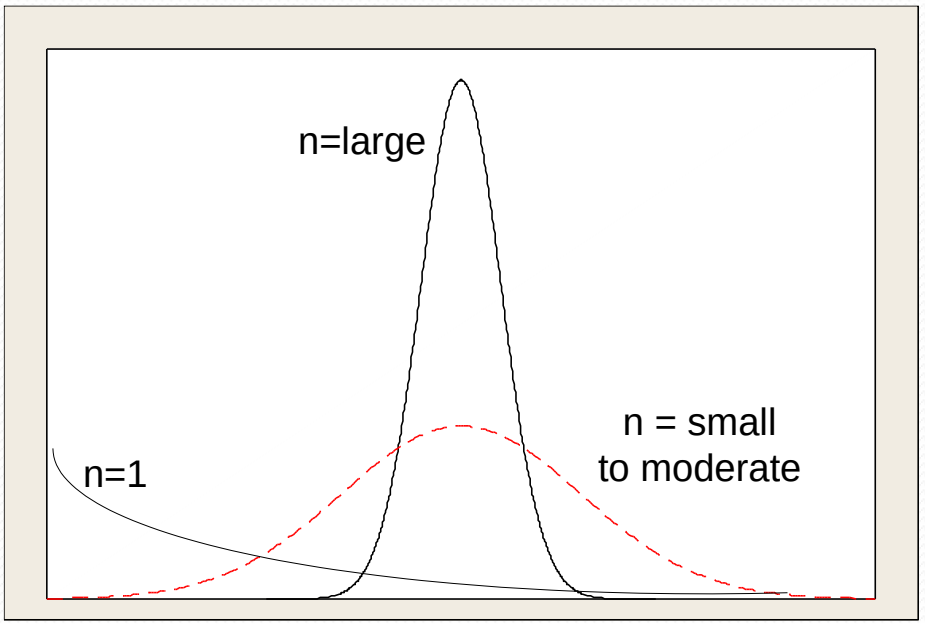
Theorem 2: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$; ***i.e.*, $E(\bar{X}) = \mu$ and $V(\bar{X}) = \frac{\sigma^2}{n}$.**

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of $\overline{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



- One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2.

- For standard normal distribution, we have the z-transformation

$$Z = \frac{\overline{X} - \mu}{S} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Exercise:

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Sampling Distribution of the Difference between Two Means

- Suppose that we have two populations, the first with mean μ_1 and variance σ_1^2, and the second with mean μ_2 and variance σ_2^2.

- Let the statistic $\bar{X}_1$ represent the mean of a random sample of size n_1 selected from the first population, and the statistic $\bar{X}_2$ represent the mean of a random sample of size n_2 selected from the second population, independent of the sample from the first population.

Sampling Distribution of the Difference between Two Means

Distribution of $\bar{X}_1 - \bar{X}_2$:

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normal distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

is approximately a standard normal variable.

Sampling Distribution of the Difference between Two Means

- If both n_1 and n_2 are greater than or equal to 30, the normal approximation for the distribution of $\bar{X}_1 - \bar{X}_2$ is very good when the underlying distributions are not too far away from normal.

- When n_1 and n_2 are less than 30, the normal approximation is reasonably good except when the populations are decidedly nonnormal.

- If both populations are normal, then $\bar{X}_1 - \bar{X}_2$ has a normal distribution no matter what the sizes of n_1 and n_2 are.

Exercise

Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type *A*, and the drying time, in hours, is recorded for each. The same is done with type *B*. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{X}_A - \bar{X}_B > 1.0)$, where $\bar{X}_A \; and \; \bar{X}_B$ are average drying times for samples of size $n_A = n_B = 18$.

Sampling Distribution of S^2

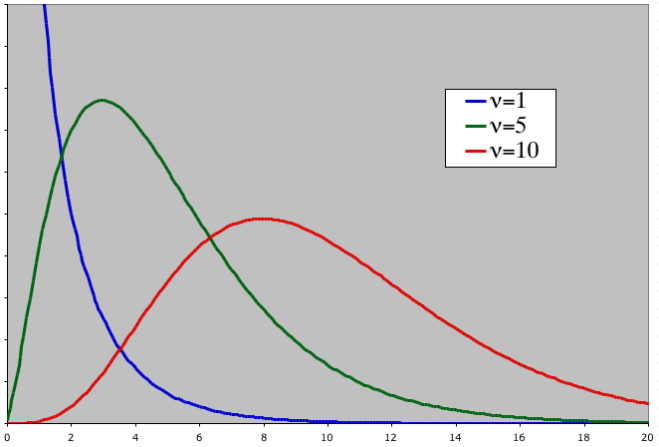
- If an engineer is interested in the population mean resistance of a certain type of resistor, the sampling distribution of $\bar{X}$ will be exploited once the sample information is gathered.

- On the other hand, if the variability in resistance is to be studied, clearly the sampling distribution of S^2 will be used in learning about the parametric counterpart, the population variance σ^2.

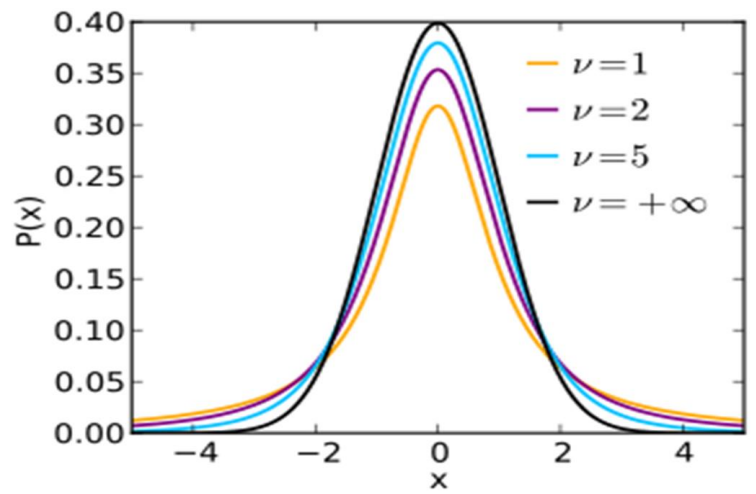
Standard Sampling Distributions

- Apart from the standard normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.
    - χ^2: Describes the distribution of variance.
    - t: Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
    - F: Describes the distribution of the ratio of two variables.

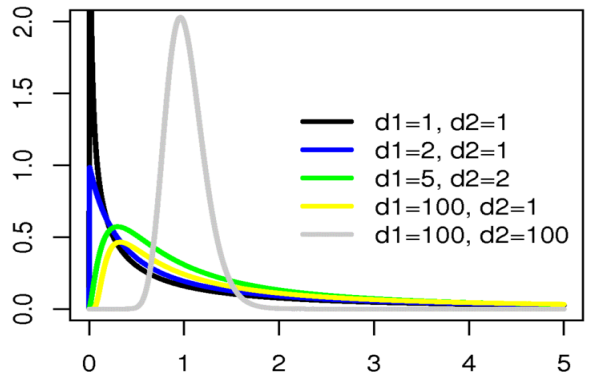
Standard Sampling Distributions



χ^2 - (Chi-Square) distribution curve



t- distribution curve



F – distribution curve

The χ^2 Distribution

A common use of the χ^2 distribution is to describe the distribution of the sample variance.

Definition 1: χ^2 distribution

If $x_1, x_2 \ldots\ldots\ldots x_n$ are independent random variables having identical normal distribution with mean μ and variance σ^2, then the random variable

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

has a Chi squared distribution with *n* degrees of freedom. (How?)

The χ^2 Distribution

Note: Each of the n independent random variable $\left(\frac{x_i-\mu}{\sigma}\right)^2, i = 1, 2, 3, \ldots\ldots n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2- distribution for sample variance.
We can write

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}[(x_i - \overline{x}) + (\overline{x} - \mu)]^2$$

$$= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n.(\overline{x} - \mu)^2$$

or, $$\frac{1}{\sigma^2}\sum(x_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\overline{x}-\mu)^2}{\sigma^2/n}$$

Chi-square distribution with n-degree

Chi-square distribution with (n-1) degree of freedom

Chi-square distribution with 1 degree of freedom $[= Z^2]$

Note: To **calculate degrees of freedom**, subtract the number of relations from the number of observations.

The χ^2 Distribution

Definition 2: χ^2-distribution for Sampling Variance

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has a chi-squared distribution with $v = n - 1$ degrees of freedom.

This way χ^2- distribution is used to describe the sampling distribution of S^2.

Chi-Squared Distribution

Definition 3: Chi-squared distribution

The continuous random variable x has a Chi-squared distribution with v degrees of freedom, is given by

$$f(x:v) = \begin{cases} \dfrac{1}{2^{\frac{v}{2}}\,\Gamma(v/2)} x^{v/2-1} e^{-\frac{x}{2}}, & x > 0 \\ 0 & Otherwise \end{cases}$$

where v is a positive integer and

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1}\, dt.$$

- The Chi-squared distribution plays an important role in statistical inference .
- The mean and variance of Chi-squared distribution are:

$$\mu = v \text{ and } \sigma^2 = 2v \text{ (Prove !)}$$

Exercise

A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced that the batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

The t Distribution

- **The t Distribution**

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
2. This require the known value of σ a priori.
3. However, in many situation, σ is certainly no more reasonable than the knowledge of the population mean μ.
4. In such situation, only measure of the standard deviation available may be the sample standard deviation S.
5. It is natural then to substitute S for σ. The problem is that the resulting statistics is not normally distributed!
6. The t distribution is to alleviate this problem. This distribution is called $student's\ t$ or simply $t - distribution$.

The t Distribution

Definition: t −distribution

The t −distribution with v degrees of freedom actually takes the form

$$t(v) = \frac{Z}{\sqrt{\frac{\chi^2(v)}{v}}}$$

where Z is a standard normal random variable, and $\chi^2(v)$ is χ^2 random variable with v degrees of freedom.

The probability density function :

$$f(t) = \frac{\Gamma[(\vartheta+1)/2]}{\Gamma(\vartheta/2)\sqrt{\pi\vartheta}}\left(1+\frac{t^2}{\vartheta}\right)^{-(\vartheta+1)/2}, \qquad -\infty < t < \infty$$

This is known as t distribution with $\vartheta = n - 1$ degrees of freedom.

The *t* Distribution

Corollary: Let $X_1, X_2, \ldots\ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ.

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Using this definition, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has the standard normal distribution.

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has the χ^2 distribution with $(n-1)$ degrees of freedom.

Thus, $T = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$ or $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

This is the $t - distribution$ with $(n-1)$ degrees of freedom.

t Distribution

- If the sample size is small, the values of S^2 fluctuate considerably from sample to sample.
- The distribution of T deviates appreciably from that of a standard normal distribution.
- If the sample size is large enough, say $n \geq 30$, the distribution of T does not differ considerably from the standard normal.
- For $n < 30$, it is useful to deal with the exact distribution of T.
- In developing the sampling distribution of T, we shall assume that our random sample was selected from a normal population.

Exercise

A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he samples 25 batches each month. If the computed *t*-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{x} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

F Distribution

- While it is of interest to let sample information shed light on two population means, it is often the case that a comparison of variability is equally important, if not more so.
- The *F*-distribution finds enormous application in comparing sample variances.
- Applications of the *F*-distribution are found in problems involving two or more samples.
- The statistic *F* is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom.
- Hence, we can write

$$F = \frac{\chi_1^2/\vartheta_1}{\chi_2^2/\vartheta_2}$$

where χ_1^2 and χ_2^2 are independent random variables having chi-squared distributions with $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ degrees of freedom, respectively.

F Distribution

- The curve of the F-distribution depends not only on the two parameters ϑ_1 and ϑ_2 but also on the order in which we state them.
- Let f_α be the f-value above which we find an area equal to α.
- Writing $f_\alpha(\vartheta_1, \vartheta_2)$ for f_α with ϑ_1 and ϑ_2 degrees of freedom, then

$$f_{1-\alpha}(\vartheta_1, \vartheta_2) = \frac{1}{f_\alpha(\vartheta_2, \vartheta_1)}.$$

- Thus the f-value with 6 and 10 degrees of freedom, leaving an area of 0.95 to the right is

$$f_{0.95}(6,10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246$$

F Distribution

- Probability density function:

$$h(x) = \frac{\Gamma[(\vartheta_1 + \vartheta_2)/2]}{\Gamma(\vartheta_1/2)\Gamma(\vartheta_2/2)} \left(\frac{\vartheta_1}{\vartheta_2}\right)^{\vartheta_1/2} \frac{x^{(\vartheta_1/2)-1}}{\left[1 + \left(\frac{\vartheta_1}{\vartheta_2}\right)x\right]^{(\vartheta_1+\vartheta_2)/2}}, 0 < x < \infty$$

with ϑ_1 and ϑ_2 degrees of freedom.

- If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2 respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an *F*-distribution with $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ degrees of freedom.

The F Distribution

Definition: F distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom. Hence,

$$\mathrm{F}(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Corollary : Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n-1)$ degrees of freedom.

Therefore, if we assume that we have sample of size n_1 from a population with variance ${\sigma_1}^2$ and an independent sample of size n_2 from another population with variance ${\sigma_2}^2$, then the statistics

$$F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$$

Note: The F distribution finds enormous applications in comparing sample variances.

Exercise

Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0.8 and 0.5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level.

Reference Book

Foundations of Statistics for Data Scientists With R and Python

By Alan Agresti, Maria Kateri (2022)