An Introduction to Statistics

Keone Hon
<keone.hon@gmail.com>

Contents

1 Descriptive Statistics **2**

1.1 Descriptive vs. Inferential 2

1.2 Means, Medians, and Modes 2

1.3 Variability 4

1.4 Linear Transformations 5

1.5 Position 6

1.6 Dispersion Percentages 7

2 Graphs and Displays **9**

2.1 Histograms 9

2.1.1 Introduction 9

2.1.2 Medians, Modes, and Means Revisited 10

2.1.3 z-Scores and Percentile Ranks Revisited 11

2.2 Stem and Leaf Displays 11

2.3 Five Number Summaries and Box and Whisker Displays 12

3 Probability **13**

3.1 Introduction 13

3.2 Random Variables 16

3.2.1 Definition 16

3.2.2 Expected Value 17

3.2.3 Variance and Standard Deviation 17

3.2.4 "Shortcuts" for Binomial Random Variables 18

4 Probability Distributions **19**

4.1 Binomial Distributions 19

4.2 Poisson Distributions 21

4.2.1 Definition 21

4.2.2 As an Approximation to the Binomial 22

4.3 Normal Distributions 23

4.3.1 Definition and Properties 23

4.3.2 Table of Normal Curve Areas 23

4.3.3 Working Backwards 25

4.3.4 As an Approximation to the Binomial 26

5 The Population Mean **28**

5.1 The Distribution of Sample Means 28

5.2 Confidence Interval Estimatess 29

5.3 Choosing a Sample Size 30

5.4 The Hypothesis Test 31

5.5 More on Errors 32

5.5.1 Type I Errors and Alpha-Risks 32

5.5.2 Type II Errors and Beta-Risks 34

5.6 Comparing Two Means 35

5.6.1 Confidence Interval Estimates 35

Chapter 1

Descriptive Statistics

1.1 Descriptive vs. Inferential

There are two main branches of statistics: descriptive and inferential. Descriptive statistics is used to say something about a set of information that has been collected only. Inferential statistics is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population. Thus, inferential statistics involves generalizing beyond the data, something that descriptive statistics does not do.

Other distinctions are sometimes made between data types.

- Discrete data are whole numbers, and are usually a count of objects. (For instance, one study might count how many pets different families own; it wouldn't make sense to have half a goldfish, would it?)
- Measured data, in contrast to discrete data, are continuous, and thus may take on any real value. (For example, the amount of time a group of children spent watching TV would be measured data, since they could watch any number of hours, even though their watching habits will probably be some multiple of 30 minutes.)
- Numerical data are numbers.
- Categorical data have labels (i.e. words). (For example, a list of the products bought by different families at a grocery store would be categorical data, since it would go something like {milk, eggs, toilet paper, ...}.)

1.2 Means, Medians, and Modes

In everyday life, the word "average" is used in a variety of ways - batting averages, average life expectancies, etc. - but the meaning is similar, usually

the center of a distribution. In the mathematical world, where everything must be precise, we define several ways of finding the center of a set of data:

> **Definition 1: median**.
> The median is the middle number of a set of numbers arranged in numerical order. If the number of values in a set is even, then the median is the sum of the two middle values, divided by 2.

The median is not affected by the magnitude of the extreme (smallest or largest) values. Thus, it is useful because it is not affected by one or two abnormally small or large values, and because it is very simple to calculate. (For example, to obtain a relatively accurate average life of a particular type of lightbulb, you could measure the median life by installing several bulbs and measuring how much time passed before half of them died. Alternatives would probably involve measuring the life of each bulb.)

> **Definition 2: mode**.
> The mode is the most frequent value in a set. A set can have more than one mode; if it has two, it is said to be bimodal.

Example 1:
The mode of {1, 1, 2, 3, 5, 8} is 1.
The modes of {1, 3, 5, 7, 9, 9, 21, 25, 25, 31} are 9 and 25. Thus, the set is bimodal.

The mode is useful when the members of a set are very different - take, for example, the statement "there were more Ds on that test than any other letter grade" (that is, in the set {A, B, C, D, E}, D is the mode). On the other hand, the fact that the mode is absolute (for example, 2.9999 and 3 are considered just as different as 3 and 100 are) can make the mode a poor choice for determing a "center". For example, the mode of the set {1, 2.3, 2.3, 5.14, 5.15, 5.16, 5.17, 5.18, 10.2} is 2.3, even though there are many values that are close to, but not exactly equal to, 5.16.

> **Definition 3: mean**.
> The mean is the sum of all the values in a set, divided by the number of values. The mean of a whole population is usually denoted by μ, while the mean of a sample is usually denoted by $\overline{x}$. (Note that this is the arithmetic mean; there are other means, which will be discussed later.)

Thus, the mean of the set $\{a_1, a_2, \cdots, a_n\}$ is given by

$$\mu = \frac{a_1 + a_2 + \cdots + a_n}{n} \tag{1.1}$$

The mean is sensitive to *any* change in value, unlike the median and mode, where a change to an extreme (in the case of a median) or uncommon (in the case of a mode) value usually has no effect.

One disadvantage of the mean is that a small number of extreme values can distort its value. For example, the mean of the set {1, 1, 1, 2, 2, 3, 3, 3, 200} is 24, even though almost all of the members were very small. A variation called the **trimmed mean**, where the smallest and largest quarters of the values are removed before the mean is taken, can solve this problem.

1.3 Variability

Definition 4: range.
The range is the difference between the largest and smallest values of a set.

The range of a set is simple to calculate, but is not very useful because it depends on the extreme values, which may be distorted. An alternative form, similar to the trimmed mean, is the interquartile range, or IQR, which is the range of the set with the smallest and largest quarters removed. If $Q1$ and $Q3$ are the medians of the lower and upper halves of a data set (the values that split the data into quarters, if you will), then the IQR is simply $Q3 - Q1$.

The IQR is useful for determining outliers, or extreme values, such as the element {200} of the set at the end of section 1.2. An outlier is said to be a number more than 1.5 IQRs below $Q1$ or above $Q3$.

Definition 5: variance.
The variance is a measure of how items are dispersed about their mean. The variance σ^2 of a whole population is given by the equation

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma x^2}{n} - \mu^2 \tag{1.2}$$

The variance s^2 of a sample is calculated differently:

$$s^2 = \frac{\Sigma(x-\overline{x})^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)} \tag{1.3}$$

Definition 6: standard deviation.
The standard deviation σ (or s for a sample) is the square root of the variance. (Thus, for a population, the standard deviation is the

square root of the average of the squared deviations from the mean. For a sample, the standard deviation is the square root of the sum of the squared deviations from the mean, divided by the number of samples minus 1. Try saying that five times fast.)

Definition 7: relative variability.
The relative variability of a set is its standard deviation divided by its mean. The relative variability is useful for comparing several variances.

1.4 Linear Transformations

A linear transformation of a data set is one where each element is increased by or multiplied by a constant. This affects the mean, the standard deviation, the IQR, and other important numbers in different ways.

Addition. If a constant c is added to each member of a set, the mean will be c more than it was before the constant was added; the standard deviation and variance will not be affected; and the IQR will not be affected. We will prove these facts below, letting μ and σ be the mean and standard deviation, respectively, before adding c, and μ_t and σ_t be the mean and standard deviation, respectively, after the transformation. Finally, we let the original set be $\{a_1, a_2, \ldots, a_n\}$, so that the transformed set is $\{a_1 + c, a_2 + c, \ldots, a_n + c\}$.

$$\begin{aligned}
\mu_t &= \frac{(a_1 + c) + (a_2 + c) + \cdots + (a_n + c)}{n} = \frac{a_1 + a_2 + \cdots + a_n + n \cdot c}{n} \\
&= \frac{a_1 + a_2 + \cdots + a_n}{n} + \frac{cn}{n} = \mu + c \\
\sigma_t &= \sqrt{\frac{\sum_{i=1}^{n} ((a_i + c) - (\mu + c))^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (a_i - \mu)^2}{n}} = \sigma \\
IQR_t &= Q3_t - Q1_t = (Q3 + c) - (Q1 + c) = Q3 - Q1 = IQR
\end{aligned}$$

where we use the result of the first equation to replace μ_t with $\mu + c$ in the second equation. Since the variance is just the square of the standard deviation, the fact that the standard deviation is not affected means that the variance won't be, either.

Multiplication.

Another type of transformation is multiplication. If each member of a set is multiplied by a constant c, then the mean will be c times its value before the constant was multiplied; the standard deviation will be $|c|$ times its value before

the constant was multiplied; and the IQR will be $|c|$ times its value. Using the same notation as before, we have

$$\begin{aligned}
\mu_t &= \frac{(a_1c)+(a_2c)+\cdots+(a_nc)}{n} = \frac{(a_1+a_2+\cdots+a_n)\cdot c}{n} \\
&= \frac{a_1+a_2+\cdots+a_n}{n}\cdot c = \mu\cdot c \\
\sigma_t &= \sqrt{\frac{\sum_{i=1}^{n}((a_ic)-(\mu c))^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}c^2(a_i-\mu)^2}{n}} = \sqrt{\frac{c^2\sum_{i=1}^{n}(a_i-\mu)^2}{n}} \\
&= \sqrt{c^2\frac{\sum_{i=1}^{n}(a_i-\mu)^2}{n}} = \sqrt{c^2}\cdot\sigma = |c|\sigma \\
IQR_t &= |Q3_t - Q1_t| = |Q3\cdot c - Q1\cdot c| = |c|(Q3-Q1) = |c|IQR
\end{aligned}$$

1.5 Position

There are several ways of measuring the relative position of a specific member of a set. Three are defined below:

Definition 8: simple ranking.
As the name suggests, the simplest form of ranking, where objects are arranged in some order and the rank of an object is its position in the order.

Definition 9: percentile ranking.
The percentile ranking of a specific value is the percent of scores/values that are below it.

Definition 10: z-score.
The z-score of a specific value is the number of standard deviations it is from the mean. Thus, the z-score of a value x is given by the equation

$$z = \frac{x-\mu}{\sigma} \tag{1.4}$$

where μ is the mean and σ is the standard deviation, as usual.

Example 2:
In the set of grade point averages {1.1, 2.34, 2.9, 3.14, 3.29, 3.57, 4.0}, the value 3.57 has the simple ranking of 2 out of 7 and the percentile ranking of $\frac{5}{7} \approx 71.43\%$. The mean is $\frac{20.34}{7} \approx 2.91$ and the standard deviation is 0.88, so the z-score is $\frac{3.57-2.91}{0.88} = 0.75$.

Conversely, if given a z-score, we can find a corresponding value, using the equation

$$x = \mu + z\sigma \tag{1.5}$$

Example 3:
The citizens of Utopia work an average (mean) of 251 days per year, with a standard deviation of 20 days. How many days correspond to a z-score of 2.3?

Since each z corresponds to one standard deviation, a z-score of 2.3 means that the desired value is 2.3 standard deviations more than the mean, or $251 + 2.3 \cdot 20 = 297$.

1.6 Dispersion Percentages

> **Theorem 1: empirical rule**
> For data with a "bell-shaped" graph, about 68% of the values lie within one standard deviation of the mean, about 95% lie withing two standard deviations, and over 99% lie within three standard deviations of the mean.

Note that since 99% of the data fall within a span of six standard deviations (z-scores of -3 to +3), the standard deviation of a set of values that are somewhat bell-shaped should be about $\frac{1}{6}$ of the range. This can be useful in checking for arithmetic errors.

> **Theorem 2: Chebyshev's Theorem**
> For any set of data, at least $1 - \frac{1}{k^2}$ of the values lie within k standard deviations of the mean (that is, have z-scores between $-k$ and $+k$).

Example 4:
Matt reads at an average (mean) rate of 20.6 pages per hour, with a standard deviation of 3.2. What percent of the time will he read between 15 and 26.2 pages per hour?

15 pages per hour corresponds to a z-score of $\frac{15-20.6}{3.2} = -1.75$, and 26.2 pages per hour corresponds to a z-score of $\frac{26.2-20.6}{3.2} = 1.75$. Chebyshev's Theorem says that $1 - \frac{1}{1.75^2} = .673$ of the values will be within 1.75 standard deviations, so 67.3% of the time, Matt's reading speed will be between 15 and 26.2 pages per hour.

Chapter 2

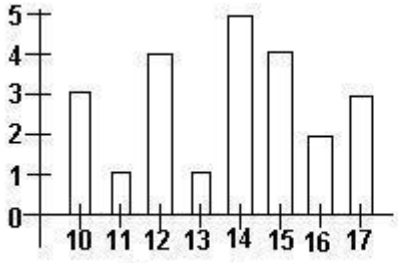
Graphs and Displays

2.1 Histograms

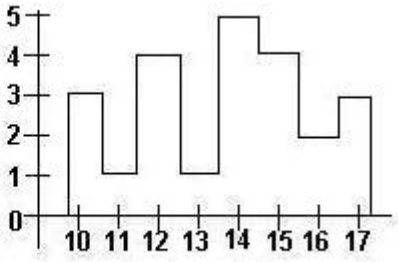
2.1.1 Introduction

Definition 11: histogram.
A histogram is a graphical representation of data, where relative frequencies are represented by relative areas. A histogram's height at a specific point represents the relative frequency of a certain item.

A histogram is similar to a bar graph, except the sides of the bars are widened until there is no space between bars:



a bar graph



the same graph, in histogram format

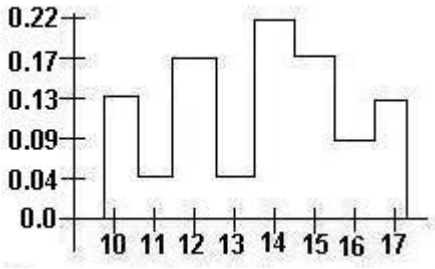
A histogram isn't limited to displaying frequencies. The y-axis (vertical axis) may labeled with relative frequencies. To determine relative frequency, we use the simple formula

$$\text{relative frequency} = \frac{\text{frequency of item}}{\text{total}} \tag{2.1}$$

In the example we used above, the total is $3 + 1 + 4 + 1 + 5 + 4 + 2 + 3 = 23$, so the relative frequencies are as follows:

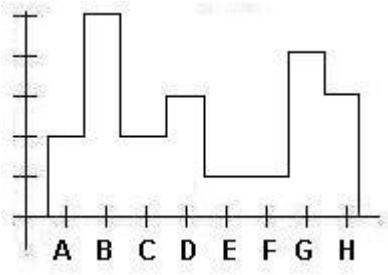
x-value	frequency	relative frequency
10	3	$3/23 \approx .13$
11	1	$1/23 \approx .04$
12	4	$4/23 \approx .17$
13	1	$1/23 \approx .04$
14	5	$5/23 \approx .22$
15	4	$4/23 \approx .17$
16	2	$2/23 \approx .09$
17	3	$3/23 \approx .13$

The resulting histogram is shown below. Note that it has the same shape as the histogram labeled with actual frequencies. This is true in general.

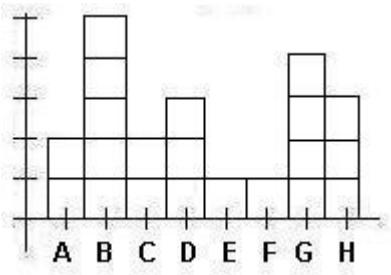


If we were given a histogram with an unlabeled y-axis, we could still determine the relative frequencies of the items because the relative frequency is equal to the fraction of the total area that a certain column covers.

Example 5:
Determine the relative frequency of items A-H using the histogram below:



We cut the histogram up into rectangles of equal area:



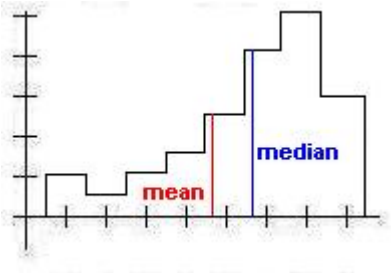
There are 21 rectangles in all. So if an item has an area of x rectangles, it will have relative frequency $\frac{x}{21}$.

item	area	relative frequency
A	2	$2/21 \approx .1$
B	5	$5/21 \approx .24$
C	2	$2/21 \approx .1$
D	3	$3/21 \approx .14$
E	1	$1/21 \approx .05$
F	1	$1/21 \approx .05$
G	4	$4/21 \approx .19$
H	3	$3/21 \approx .14$

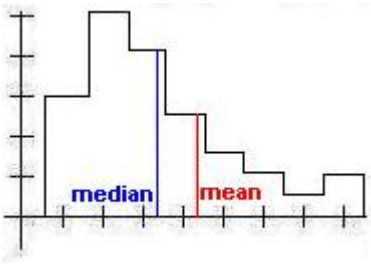
2.1.2 Medians, Modes, and Means Revisited

Histograms can be used to show the median, mode, and mean of a distribution. Since the mode is the most frequent value, it is the point on the histogram where the graph is highest. Since the median is in the middle of a distribution (so it divides the distribution in half), it may be represented by the line that divides the area of the histogram in half. Finally, the mean is a line that passes through the center of gravity of the histogram.

If the mean is less than the median, then the distribution is said to be *skewed to the left*. It will be spread widely to the left. Here is an example of a distribution that is skewed to the left:



Conversely, if the mean is greater than the median, then the distribution is *skewed to the right*, and it will be spread widely to the right. Here is an example of a distribution that is skewed to the right:



2.1.3 z-Scores and Percentile Ranks Revisited

Earlier, we exchanged frequency for relative frequency on the vertical axis of a histogram. In the same spirit, we can label the horizontal axis in terms of z-scores rather than with the names of the items in the set.

One common way that data is given is through percentile rankings. With this information, we can construct a histogram:

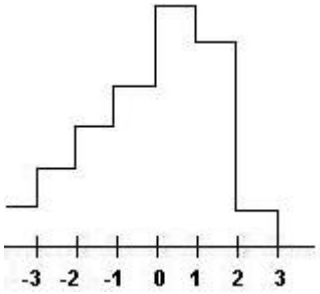
Example 6:
Construct a histogram using the following data:

z-score	percentile ranking
-3	5
-2	15
-1	30
0	50
1	80
2	95
3	100

From $-\infty to -3$, the relative frequency is 0.05. From -3 to -2, the relative frequency increases from 0.05 to 0.15, so the relative frequency at between -3 and -2 is 0.1. From -2 to -1, the relative frequency increases from 0.15 to 0.3, so the relative frequency between the two is 0.15. Using similar reasoning, we obtain the following:

z-score	relative frequency
$-\infty$ to -3	.05
-3 to -2	.10
-2 to -1	.15
-1 to 0	.20
0 to 1	.30
1 to 2	.25
2 to 3	.05

Plotting these values on a histogram, we obtain



2.2 Stem and Leaf Displays

A stem and leaf display is similar to a histogram, since it shows how many values in a set fall under a certain interval. However, it has even more information - it shows the actual values within the interval. The following example demonstrates a stem and leaf display:

Example 7:
Here is a stem and leaf display of the set {10, 14, 19, 22, 25, 25, 28, 31, 33, 39, 39, 40, 40, 40, 41, 44, 45}:

Stems	Leaves
1	0 4 9
2	2 5 5 8
3	1 3 9 9
4	0 0 0 1 4 5

If we draw a line around the stem and leaf display, the result is a histogram, albeit one whose orientation is different from the kind we are used to seeing:

1	0 4 9
2	2 5 5 8
3	1 3 9 9
4	0 0 0 1 4 5

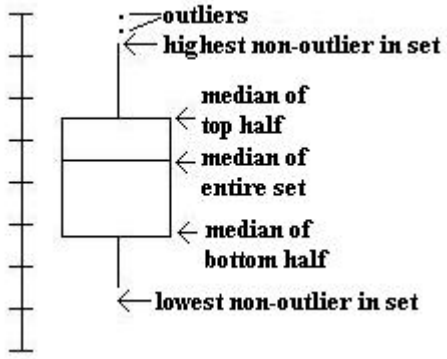
The stems and leaves need not be single digits, although all leaves should be the same size in order to make the display accurate.

2.3 Five Number Summaries and Box and Whisker Displays

The five-number summary is a group of five numbers that help describe the shape and variation of a distribution. These five numbers are $Q2$, the median of the set; $Q1$, the median of the lower half of the set; $Q3$, the median of the upper half of the set, and the maximum and minimum numbers that are not outliers. (See section 1.3 for the definition of an outlier.)

The box and whisker display is a graphical representation of the five-number summary. Horizontal lines are drawn at $Q1$, $Q2$, and $Q3$; a box is then drawn around these lines, forming a double box with a shared side. Two perpendicular lines ("whiskers") are drawn from the top and bottom of the box to the maximum and minimum non-outliers. Any outliers are then plotted as individual points.

The diagram below shows the basic box and whisker format.



Note that because a median splits a set in half, the top whisker represents the top quarter, the top box represents the second quarter, the bottom box represents the third quarter, and the bottom whisker represents the bottom quarter.

Chapter 3

Probability

3.1 Introduction

Definition 12: probability.
The probability of a specific event is a mathematical statement about the likelihood that it will occur. All probabilities are numbers between 0 and 1, inclusive; a probability of 0 means that the event will never occur, and a probability of 1 means that the event will always occur.

The sum of the probabilities of all possible outcomes of any event is 1. (This is because *something* will happen, so the probability of some outcome occurring is 1.)

Definition 13: complimentary event.
With respect to an event E, the complimentary event, denoted E', is the event that E does not occur. For example, consider the event that it will rain tomorrow. The compliment of this event is the event that it will not rain tomorrow.

Since an event must either occur or not occur, from above, it must be the case that

$$P(E) + P(E') = 1 \tag{3.1}$$

Definition 14: mutually exclusive.
Two or more events are mutually exclusive if they cannot occur simultaneously.

Two events A and B are mutually exclusive if $A \cup B = 0$ - that is, if they have no members in common.

Example 8:
Let A = the event that it is Monday, B = the event that it is Tuesday, and C = the event that it is the year 2004.

A and B are mutually exclusive events, since it cannot be both Monday and Tuesday at the same time. A and C are not mutually exclusive events, since it can be a Monday in the year 2004.

Theorem 3: Principle of Inclusion and Exclusion

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y) \tag{3.2}$$

Recall that when two events A and B are mutually exclusive $P(A \cap B) = 0$. Using this fact and the Principle of Inclusion and Exclusion (PIE), we conclude that when two events are mutually exclusive, the probability that both of them will occur ($P(A \cup B)$) is $P(A) + P(B)$.

Definition 15: independent.
Two events are said to be independent if the chance that each one occurs is not affected by the outcome of any of the others.

Example 9:
Suppose that two dice are rolled. The outcome of either die will not affect the outcome of the other die, so the two dice are independent. On the other hand, the event that John Kerry will become president of the U.S. and the event that John Edwards will become vice president are not independent, since the two are (were?) running mates, so if one is elected, so is the other.

Theorem 4: Independence Principle
If two events are independent, then the probability that both will occur is equal to the product of their individual probabilities. In other words, if A and B are independent, then

$$P(A \cap B) = P(A) + P(B) \tag{3.3}$$

Definition 16: factorial.
The factorial of an integer n is defined as the product of all the positive integers from 1 to n, that is

$$n! \quad \text{(read "n factorial")} = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 \tag{3.4}$$

0! is defined to be 1.

Definition 17: combination.
A combination is a set of unordered (i.e. order does not matter) items. If we are to choose k distinct objects from a total of n objects, then there are $\binom{n}{k}$ different combinations, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.5}$$

Formula 1: Binomial Formula
Suppose that an event occurs with probability p. Then the probability that it will occur exactly x times out of a total of n trials is

$$\binom{n}{x} \cdot p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x (1-p)^{n-x} \tag{3.6}$$

Example 10:
Derive a formula for calculating P(x) (the probability of x successes out of n trials, where the probability of each success is p) in terms of x, n, p, *and* $P(x-1)$.

The probability of $x-1$ successes is, using the binomial formula,

$$\begin{aligned} P(x-1) &= \binom{n}{x-1} \cdot p^{x-1}(1-p)^{n-(x-1)} \\ &= \frac{n!}{(x-1)!(n-x+1)!} p^{x-1}(1-p)^{n-x+1} \end{aligned}$$

The probability of x successes is, again using the binomial formula,

$$\begin{aligned} P(x) &= \binom{n}{x} \cdot p^x (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} \cdot p^x (1-p)^{n-x} \end{aligned}$$

We want to find an expression K such that $P(x-1) \cdot K = P(x)$. Thus,

$$\begin{aligned}
\frac{n! \cdot p^{x-1}(1-p)^{n-x+1}}{(x-1)!(n-x+1)!} \cdot K &= \frac{n! \cdot p^x (1-p)^{n-x}}{x!(n-x)!} \cdot \\
\frac{n! \cdot p^{x-1}(1-p)^{n-x}(1-p)}{(x-1)!(n-x)!(n-x+1)} \cdot K &= \frac{n! \cdot p \cdot p^{x-1}(1-p)^{n-x}}{x(x-1)!(n-x)!} \\
\frac{1-p}{(n-x+1)} \cdot K &= \frac{p}{x} \\
K &= \frac{n-x+1}{x} \cdot \frac{p}{1-p}
\end{aligned}$$

Hence, the desired formula is

$$P(x) = \frac{n-x+1}{x} \cdot \frac{p}{1-p} \cdot P(x-1) \tag{3.7}$$

3.2 Random Variables

3.2.1 Definition

Definition 18: random variable.
A variable that may take on different values depending on the outcome of some event.

Example 11:
On the AP Statistics exam, $\frac{1}{5}$ of the class received an 5, $\frac{1}{3}$ received a 4, $\frac{1}{6}$ received a 3, $\frac{1}{20}$ received a 2, and $\frac{1}{3}$ received an 1. If x represents the score of a randomly chosen student in the class, then x is a random variable that takes the values 1, 2, 3, 4, and 5.

A random variable may be discrete (it takes on a finite number of values) or continuous (it can take on any value in a certain interval, that is, an infinite number of values).

Definition 19: probability distribution.
A list or formula that gives the probability for each discrete value of a random variable.

3.2.2 Expected Value

Definition 20: expected value.
The expected value (also called the mean) of a random variable is the sum of the product of each possible value and its corresponding probability. In mathematical terms, if the random variable is X, the possible values are $x_1, x_2, \ldots x_n$, and the corresponding probabilities are $P(x_1), P(x_2), \ldots, P(x_n)$ then

$$E(X) = \Sigma_{i=1}^{n}(x_i \cdot P(x_i)) \tag{3.8}$$

Example 12:
Investing in Sharma Furniture Co. has a 60% *chance of resulting in a $10,000 gain and a* 40% *chance of resulting in a $3,000 loss. What is the expected value of investing in Sharma Furniture Co.?*

$E(X) = .6 \cdot 10000 + .4 \cdot -3000 = 6000 - 1200 = 4800$

Thus, the expected value is $4,800.

3.2.3 Variance and Standard Deviation

Definition 21: variance of a discrete random variable.
The variance σ^2 of a random variable, which takes on discrete values x and has mean μ, is given by the equation

$$\sigma^2 = \Sigma(x-\mu)^2 P(x) = \Sigma(x^2 P(x)) - \mu^2 \tag{3.9}$$

Definition 22: standard deviation of a discrete random variable.
The standard deviation σ of a discrete random variable is equal to the square root of the variance, i.e. $\sigma = \sqrt{\sigma^2}$.

Example 13:
In a contest sponsored by Hansen sodas, you win a prize if the cap on your bottle of sodas says "WINNER"; however, you may only claim one prize. Eager to win, you blow all your savings on sodas; as a result, you have a 0.05% chance of winning $1,000,000, a 1% chance of winning $20,000, and a 90% chance of winning $10. Ignoring the money you spent on sodas, what is your expected

value and standard deviation?

First, we calculate the mean.

$$\mu = 1000000 \cdot 0.0005 + 0.01 \cdot 20000 + 0.9 \cdot 10 = 500 + 200 + 9 = 709$$

Now, the variance.

$$\begin{aligned} \sigma^2 &= 1000000^2 \cdot 0.0005 + 20000^2 \cdot 0.01 + 10^2 \cdot .9 - 709^2 \\ &= 500000000 + 4000000 + 90 - 502681 \\ &= 503497409 \end{aligned}$$

Finally, the standard deviation.

$$\sigma = \sqrt{503497409} \approx 22438$$

The standard deviation is over $22,000! Although the expected value looks nice, there's a good chance that you'll get a not-so-good amount of money.

3.2.4 "Shortcuts" for Binomial Random Variables

The past examples required a fair amount of arithmetic. To save time, there are simpler ways to find expected values, variances, and standard deviations of binomial random variables (that is, random variables with only two outcomes).

Theorem 5: Mean and Standard Deviation of a Binomial
In a situation with two outcomes where the probability of an outcome O is p and there are n trials, the expected number of Os is np. That is,

$$\mu = np \tag{3.10}$$

Furthermore, the variance σ^2 is given by the equation

$$\sigma^2 = np(1-p) \tag{3.11}$$

and the standard deviation σ by the equation

$$\sigma = \sqrt{np(1-p)} \tag{3.12}$$

Chapter 4

Probability Distributions

4.1 Binomial Distributions

Previously, we computed the probability that a binomial event (one with two possible outcomes for each trial) would occur. In the same way, we can compute the probability of every possible combination of outcomes and create a table or histogram from it.

Example 14:
On a five-item true or false test, Simon has an 80% chance of choosing the correct answer for any of the questions. Find the complete probability distribution for the number of correct answers that Simon can get. Then determine the mean and standard deviation of the probability distribution.

$$\begin{aligned}
P(0) &= \binom{5}{0}(0.8)^0(0.2)^5 = 1 \cdot 1 \cdot 0.00032 = 0.00032 \\
P(1) &= \binom{5}{1}(0.8)^1(0.2)^4 = 5 \cdot 0.8 \cdot 0.0016 = 0.0064 \\
P(2) &= \binom{5}{2}(0.8)^2(0.2)^3 = 10 \cdot 0.64 \cdot 0.008 = 0.0512 \\
P(3) &= \binom{5}{3}(0.8)^3(0.2)^2 = 10 \cdot 0.512 \cdot 0.04 = 0.2048 \\
P(4) &= \binom{5}{4}(0.8)^4(0.2)^1 = 5 \cdot 0.4096 \cdot 0.2 = 0.4096 \\
P(5) &= \binom{5}{5}(0.8)^5(0.2)^0 = 1 \cdot 0.32768 \cdot 1 = 0.32768
\end{aligned}$$

Note that $0.00032 + 0.0064 + 0.0512 + 0.2048 + 0.4096 + 0.32768 = 1$.

The expected value may be calculated in two ways: $\mu = np = 5 \cdot 0.8 = 4$ or $\mu = 0 \cdot 0.00032 + 1 \cdot 0.0064 + 2 \cdot 0.0512 + 3 \cdot 0.2048 + 4 \cdot 0.4096 + 5 \cdot 0.32768 = 4$. In either case, the answer is the same.

The variance is $\sigma^2 = np(1-p) = 5 \cdot 0.8 \cdot (1-0.8) = 0.8$, so the standard deviation is $\sigma = \sqrt{0.8} \approx 0.894$.

Binomial distributions are often used in quality control systems, where a small number of samples are tested out of a shipment, and the shipment is only accepted if the number of defects found among the sampled units falls below a specified number. In order to analyze whether a sampling plan effectively screens out shipments containing a great number of defects and accepts shipments with very few defects, one must calculate the probability that a shipment will be accepted given various defect levels.

Example 15:
A sampling plan calls for twenty units out of a shipment to be sampled. The shipment will be accepted if and only if the number of defective units is less than or equal to one. What is the probability that a shipment with defect level n will be accepted for $n = 0.05, 0.10, 0.15, 0.20, 0.25, 0.30$?

$$\begin{aligned} P(\text{acceptance}) &= P(0 \text{ defective units}) + P(1 \text{ defective unit}) \\ &= \binom{20}{20} n^0 \cdot (1-n)^{20} + \binom{20}{19} n^1 \cdot (1-n)^{19} \\ &= (1-n)^{20} + 20 \cdot (1-n)^{19} \cdot n \end{aligned}$$

$$\begin{aligned} \text{If } n = 0.05, P(\text{acceptance}) &= (0.95)^{20} + 20(0.95)^{19}(0.05)^1 = 0.736 \\ \text{If } n = 0.10, P(\text{acceptance}) &= (0.90)^{20} + 20(0.90)^{19}(0.10)^1 = 0.392 \\ \text{If } n = 0.15, P(\text{acceptance}) &= (0.85)^{20} + 20(0.85)^{19}(0.15)^1 = 0.176 \\ \text{If } n = 0.20, P(\text{acceptance}) &= (0.80)^{20} + 20(0.80)^{19}(0.20)^1 = 0.069 \\ \text{If } n = 0.25, P(\text{acceptance}) &= (0.75)^{20} + 20(0.75)^{19}(0.25)^1 = 0.024 \\ \text{If } n = 0.30, P(\text{acceptance}) &= (0.70)^{20} + 20(0.70)^{19}(0.30)^1 = 0.008 \end{aligned}$$

4.2 Poisson Distributions

4.2.1 Definition

Definition 23: Poisson distribution.
In a binomial distribution, when the number of trials n is large and the probability of success p is small, the distribution approaches the Poisson distribution. In the Poisson distribution, the probability of x successes is given by the equation

$$P(x \text{ successes}) = \frac{\mu^x}{x!}e^{-\mu} \tag{4.1}$$

where μ is the mean.

At first, the requirement that n be large, p be small, and the mean (np) be a known, moderate number seems overly restrictive. However, there are many cases where this occurs. For example, a grocer might sell 5 heads of lettuce each day. It's impractical to say how many heads of lettuce he didn't sell, because we do not know how many customers visited his store or how many they could have bought (and there is really no way to determine the latter). However, we can assume that there were many chances for someone to buy a head of lettuce, so n is very large. The chance of someone buying a head of lettuce at any given moment is very small, so p is small. Finally, the mean, 5 heads of lettuce per day, is known. Thus, the Poisson distribution could probably be used to describe this situation.

Here is another application of the Poisson distribution:

Example 16:
The Morgan household gets an average of 3 telephone calls per day. Using the Poisson distribution, find the probability of n phone calls for $0 \leq n \leq 6$ *in one day. Then find the probability of n phone calls in half a day for* $0 \leq n \leq 3$.

$$\begin{aligned}
P(0) &= \frac{3^0}{0!}e^{-3} = e^{-3} \approx 0.0498 \\
P(1) &= \frac{3^1}{1!}e^{-3} = 3e^{-3} \approx 0.149 \\
P(2) &= \frac{3^2}{2!}e^{-3} = \frac{9}{2}e^{-3} \approx 0.224 \\
P(3) &= \frac{3^3}{3!}e^{-3} = \frac{9}{2}e^{-3} \approx 0.224
\end{aligned}$$

$$
\begin{aligned}
P(4) &= \frac{3^4}{4!}e^{-3} = \frac{27}{8}e^{-3} \approx 0.168 \\
P(5) &= \frac{3^5}{5!}e^{-3} = \frac{81}{40}e^{-3} \approx 0.100 \\
P(6) &= \frac{3^6}{6!}e^{-3} = \frac{81}{80}e^{-3} \approx 0.050
\end{aligned}
$$

Notice that $P(0)+P(1)+P(2)+P(3)+P(4)+P(5)+P(6) \approx 0.96$, not 1. This is because there is still a small probability of 7, 8, etc. calls.

In half a day, we can expect an average of 1.5 calls per day. (It's okay to have a non-integral mean, although it is true that the number of successes, x, must be an integer.) Thus,

$$
\begin{aligned}
P(0) &= \frac{1.5^0}{0!}e^{-1.5} = e^{-1.5} \approx 0.223 \\
P(1) &= \frac{1.5^1}{1!}e^{-1.5} = 1.5e^{-1.5} \approx 0.335 \\
P(2) &= \frac{1.5^2}{2!}e^{-1.5} = 1.125e^{-1.5} \approx 0.251 \\
P(3) &= \frac{1.5^3}{3!}e^{-1.5} = 0.5625e^{-1.5} \approx 0.126 \\
P(4) &= \frac{1.5^4}{4!}e^{-1.5} = 0.2109e^{-1.5} \approx 0.047
\end{aligned}
$$

4.2.2 As an Approximation to the Binomial

Earlier we stated that the Poisson distribution was useful because it only required knowing the mean. However, even if we *do* know n and p, we can still use the Poisson distribution as an approximation. In general, if $n \geq 20$ and $p \leq 0.05$, the approximation will be "close" (of course, "close" is a relative term).

Example 17:
A standard die is rolled 120 times. What is the probability of exactly 10 sixes? 15? 20? 25?

In this case, we know n (it's 120) and p (it's $\frac{1}{6}$). However, using the binomial formula would require calculating very large and very small numbers - for the first one, $\binom{120}{10}$. (For the record, it's 116068178638776. Try to remember that!) Instead, we'll use the Poisson approximation, even though we will sacrifice some accuracy as $p \not\leq 0.05$.

The expected value is $np = 120 \cdot \frac{1}{6} = 20$. So

$$\begin{aligned} P(10) &= \frac{20^{10}}{10!}e^{-20} \approx 0.0058 \\ P(15) &= \frac{20^{15}}{15!}e^{-20} \approx 0.0516 \\ P(20) &= \frac{20^{20}}{20!}e^{-20} \approx 0.0888 \\ P(25) &= \frac{20^{25}}{25!}e^{-20} \approx 0.0446 \end{aligned}$$

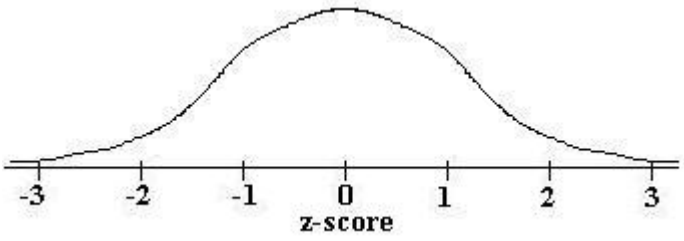
The actual values are

$$\begin{aligned} P(10) &= \binom{120}{10}\left(\frac{1}{6}\right)^{10}\left(\frac{5}{6}\right)^{110} \approx 0.0037 \\ P(15) &= \binom{120}{15}\left(\frac{1}{6}\right)^{15}\left(\frac{5}{6}\right)^{105} \approx 0.0488 \\ P(20) &= \binom{120}{20}\left(\frac{1}{6}\right)^{20}\left(\frac{5}{6}\right)^{100} \approx 0.0973 \\ P(25) &= \binom{120}{25}\left(\frac{1}{6}\right)^{25}\left(\frac{5}{6}\right)^{95} \approx 0.0441 \end{aligned}$$

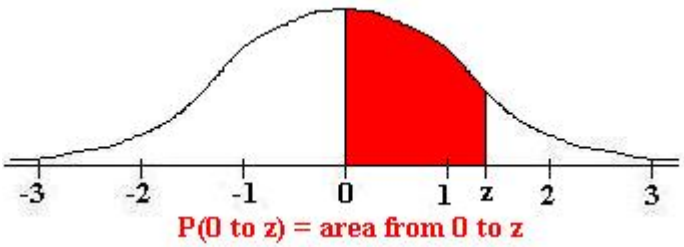
4.3 Normal Distributions

4.3.1 Definition and Properties

The normal curve is a bell-shaped, symmetrical graph with an infinitely long base. The mean, median, and mode are all located at the center.



A value is said to be normally distributed if its histogram is the shape of the normal curve. The probability that a normally distributed value will fall between the mean and some z-score z is the area under the curve from 0 to z:



4.3.2 Table of Normal Curve Areas

The area from the mean to z-score z is given in the table below:

z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4958	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Example 18:
The price of a gallon of gasoline at the gasoline stations in Nevada is normally distributed with a mean of \$2.00 and a standard deviation of \$0.50.

What is the probability that at a randomly chosen gasoline station in Nevada, the price of gasoline will be between \$2.00 and \$2.75?

The z-score of \$2.00 is 0, and the z-score of \$2.75 is $\frac{\$2.75 - \$2.00}{\$0.50} = 1.5$. According to the table above, the area between 0 and 1.5 is 0.4332 of the total area, so the probability is 0.4332.

What is the probability that at a randomly chosen gasoline station in Nevada, the price of gasoline will be between \$1.25 and \$2.00?

The z-score of \$1.25 is $\frac{\$1.75 - \$2.00}{\$0.50} = -1.5$, and the z-score of \$2.00 is 0. But since the normal curve is symmetric, the area between -1.5 and 0 is the same as the area between 0 and 1.5, which is, as before, 0.4332 of the total area. Thus the probability is also 0.4332.

What percent of the gasoline stations have prices between \$1.13 and \$3.20?

The z-score of \$1.13 is $\frac{\$1.13 - \$2.00}{\$0.50} = -1.74$, and the z-score of \$3.20 is $\frac{\$3.20 - \$2.00}{\$0.50} = 2.40$. The area between -1.74 and 2.40 is equal to the area between -1.74 and 0 plus the area between 0 and 2.40, which is $0.4591 + 0.4918 = 0.9509 \approx 95.1\%$.

What is the probability that a randomly chosen gas station will have prices greater than \$3.15 per gallon?

The z score of \$3.15 is $\frac{\$3.15 - \$2.00}{\$0.50} = 2.30$. We want to find the probability P(greater than 2.30). But how can we do this? 0 isn't greater than 2.30, so it appears that we can't use the probabilities from the above table.

Appearances can be deceiving. First, note that, in terms of z-scores, P(between 0 and 2.30) + P(greater than 2.30) = P(greater than 0). Thus, P(greater than 2.30) = P(greater than 0) - P(between 0 and 2.30).

The probability that the gas station will be between 0 and 2.30 is 0.4893, and the probability that the gas station will be greater than 0 is 0.5, so the probability that the gas station will be greater than 2.30 is $0.5 - 0.4893 = 0.0107$

4.3.3 Working Backwards

In the previous examples, we found percentages and probabilities given raw data. We can work in the reverse direction just as easily. For example, suppose we wanted to know what z-score 90% of a normal distribution was greater than. Thus, we want 10% to be less than some score z, or 40% to be between z and 0. Looking at the table, we see that a z-score of 1.28 corresponds to 0.3997, which is very close to 0.4, so z is -1.28. Thus, 90% of the normal distribution has a z score greater than -1.28. Working similarly, we obtain the following facts:

- 90% of the normal distribution is between z-scores of -1.645 and 1.645.
- 95% of the normal distribution is between z-scores of -1.96 and 1.96.
- 99% of the normal distribution is between z-scores of -2.58 and 2.58.
- 90% of the normal distribution is less than the z-score of 1.28, and 90% of the normal distribution is greater than the z-score of -1.28.
- 95% of the normal distribution is less than the z-score of 1.645, and 95% of the normal distribution is greater than the z-score of -1.645.
- 99% of the normal distribution is less than the z-score of 2.33, and 99% of the normal distribution is greater than the z-score of -2.33.
- 68.26% of the normal distribution is between z-scores of -1 and 1.
- 95.44% of the normal distribution is between z-scores of -2 and 2.
- 99.74% of the normal distribution is between z-scores of -3 and 3.

If we do not know the mean or standard deviation, we can also work backward to find it.

Example 19:
A normal distribution has a mean of 36, and 19% of the values are above 50. What is the standard deviation?

Since 0.19 of the distribution is above 50, 0.31 of the distribution is between 36 (the mean) and 50. Looking at the table of normal curve areas, 0.31 corresponds to a z-score of 0.88, so 50 is 0.88 standard deviations above the mean. Thus, $50 = 36 + 0.88\sigma$, or $0.88\sigma = 14$. Therefore $\sigma = 15.91$.

4.3.4 As an Approximation to the Binomial

The normal may also be viewed as a limiting case to the binomial, so we may use it to approximate the value of a binomial for large n. However, because the binomial only takes values at integers, whereas the normal is a continuous curve, we will represent an integer value with a unit-long interval centered at that integer. (For example, 4 would be represented by the interval from 3.5 to 4.5.)

Example 20:
Shaquille O'Neal averages 0.627 on free throws. What is the probability that out of 100 attempts, he will have made exactly 70 of them?

First, we calculate μ and σ. $\mu = np = 100 \cdot 0.627 = 62.7$ and $\sigma = \sqrt{np(1-p)} = \sqrt{62.7 \cdot 0.373} \approx 4.84$.

Then, recalling that we will represent 70 with the interval from 69.5 to 70.5, we calculate some z-scores. The z-score of 69.5 is $\frac{69.5-62.7}{4.84} = 1.405$, and the z-score of 70.5 is $\frac{70.5-62.7}{4.84} = 1.612$. The area from 0 to 1.612 is 0.4463 and the area from 0 to 1.405 is 0.4207, so the final probability is $0.4463 - 0.4207 = 0.0256$.

In general, the normal is considered a "good" approximation when both np and $n(1-p)$ are greater than 5.

Chapter 5

The Population Mean

Oftentimes it is impossible or impractical to survey an entire population. For example, a manufacturer cannot test every battery, or it wouldn't have any to sell. Instead, a sample must be taken and tested. This gives birth to many questions: what size sample should be taken to be accurate? How accurate is accurate? How can we ensure that a sample is representative of a population? What conclusions can we draw about the population using the sample? And so on. In this section, we'll discuss samples and answer some of these questions.

5.1 The Distribution of Sample Means

As mentioned before, we want to estimate a population's mean by surveying a small sample. If the sample is very small, say it contains one member, then the mean of the sample is unlikely to be a good estimate of the population. As we increase the number of members, the estimate will improve. Thus, bigger sample size generally results in a sample mean that is closer to the population mean.

Similarly, if we survey individual members of a population, their values are unlikely to be normally distributed - individuals can easily throw things off with widely varying values. However, if we take several samples, then the sample means are likely to be normally distributed, because the individuals in each sample will generally balance each other out.

Theorem 6: Central Limit Theorem
Start with a population with a given mean μ and standard deviation σ. Take samples of size n, where n is a sufficiently large (generally at least 30) number, and compute the mean of each sample.

- The set of all sample means will be approximately normally distributed.

- The mean of the set of samples will equal μ, the mean of the population.
- The standard deviation, $\sigma_{\overline{x}}$, of the set of sample means will be approximately $\frac{\sigma}{\sqrt{n}}$.

Example 21:
Foodland shoppers have a mean \$60 grocery bill with a standard deviation of \$40. What is the probability that a sample of 100 Foodland shoppers will have a mean grocery bill of over \$70?

Since the sample size is greater than 30, we can apply the Central Limit Theorem. By this theorem, the set of sample means of size 100 has mean \$60 and standard deviation $\frac{\$40}{\sqrt{100}} = \$4$. Thus, \$70 represents a z-score of $\frac{\$70 - \$60}{\$4} = 2.5$. Since the set of sample means of size 100 is normally distributed, we can compare a z-score of 2.5 to the table of normal curve areas. The area between $z = 0$ and $z = 2.5$ is 0.4938, so the probability is 0.5 - 0.4938 = 0.0062.

5.2 Confidence Interval Estimatess

We can find the probability that a sample lies within a certain interval of the population mean by using the central limit theorem and the table or normal curve areas. But this is the same as the probability that the population mean lies within a certain interval of a sample. **Thus, we can determine how confident we are that the population mean lies within a certain interval of a sample mean.**

Example 22:
At a factory, batteries are produced with a standard deviation of 2.4 months. In a sample of 64 batteries, the mean life expectancy is 12.35. Find a 95% confidence interval estimate for the life expectancy of all batteries produced at the plant.

Since the sample has n larger than 30, the central limit theorem applies. Let the standard deviation of the set of sample means of size 64 be $\sigma_{\overline{x}}$. Then by the central limit theorem, $2.4 = \frac{\sigma_{\overline{x}}}{\sqrt{64}}$, so $\sigma_{\overline{x}} = 0.3$ months.

Looking at the table of normal curve areas (or referring to section 4.3.3), 95% of the normal curve area is between the z-scores of -1.96 and 1.96. Since the standard deviation is 0.3, a z-score of -1.96 represents a raw score of -0.588 months, and a z-score of 1.96 represents a raw score of 0.588 months. So we have 95% confidence that the life expectancy will be between $12.35 - 0.588 = 11.762$ months and $12.35 + 0.588 = 12.938$ months.

If we do not know σ (the standard deviation of the entire population), we use s (the standard deviation of the sample) as an estimate for σ. Recall that s is defined as

$$s = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n-1}} = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} \tag{5.1}$$

where x takes on each individual value, $\overline{x}$ is the sample mean, and n is the sample size.

5.3 Choosing a Sample Size

Note that as the degree of confidence increases, the interval must become larger; conversely, as the degree of confidence decreases the interval becomes more precise. This is true in general; if we want to be more sure that we are right, we sacrifice precision, and if we want to be closer to the actual value, we are less likely to be right.

There is a way to improve both the degree of confidence and the precision of the interval: by increasing the sample size. So it seems like greater sample size is always desirable; however, in the real world, increasing the sample size costs time and money.

Generally, we will be asked to find the minimum sample size that will result in a desired confidence level and range.

Example 23:
A machine fills plastic bottles with Mountain Dew brand soda with a standard deviation of 0.04 L. How many filled bottles should be tested to determine the mean volume of soda to an accuracy of 0.01 L with 95% confidence?

Let $\sigma_{\overline{x}}$ be the standard deviation of the sample. Since $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, we have $\sigma_{\overline{x}} = \frac{0.04L}{\sqrt{n}}$. Also, since 95% confidence corresponds to z-scores from -1.96 to

1.96 on the normal curve, we have $0.01L = \sigma_{\overline{x}} \cdot 1.96$. Substituting, we obtain

$$\begin{aligned} 0.01L &= \frac{0.04L}{\sqrt{n}} \cdot 1.96 \\ \sqrt{n} &= \frac{0.04L}{0.01L} \cdot 1.96 \\ \sqrt{n} &= 7.84 \\ n &= 61.4656 \end{aligned}$$

So at least 62 bottles should be tested.

5.4 The Hypothesis Test

Oftentimes we want to determine whether a claim is true or false. Such a claim is called a hypothesis.

> **Definition 24: null hypothesis**.
> A specific hypothesis to be tested in an experiment. The null hypothesis is usually labeled H_0.

> **Definition 25: alternative hypothesis**.
> A hypothesis that is different from the null hypothesis, which we usually want to show is true (thereby showing that the null hypothesis is false). The alternative hypothesis is usually labeled H_a.
>
> If the alternative involves showing that some value is greater than or less than a number, there is some value c that separates the null hypothesis rejection region from the fail to reject region. This value is known as the **critical value**.

The null hypothesis is tested through the following procedure:

1. Determine the null hypothesis and an alternative hypothesis.
2. Pick an appropriate sample.
3. Use measurements from the sample to determine the likelihood of the null hypothesis.

> **Definition 26: Type I error**.
> If the null hypothesis is true but the sample mean is such that the null hypothesis is rejected, a Type I error occurs. The probability that such an error will occur is the α **risk**.

Definition 27: Type II error.
If the null hypothesis is false but the sample mean is such that the null hypothesis cannot be rejected, a Type II error occurs. The probability that such an error will occur is called the β **risk**.

Example 24:
A government report claims that the average temperature on the planet Venus is at least 300° C. You don't believe this - you think the average temperature must be lower - so you carry out an experiment during which you will measure the temperature on Venus at 100 random times, then compute the mean of the measured temperatures. If the mean temperature is over 20° C less than the report's claim, then you will declare the report's claim false.

Thus, the null hypothesis is $H_0 : T = 300$ and the alterative hypothesis is $H_a : T < 300$. The value $c = 280$ separates the rejection region from the fail to reject region; that is, if $T < 280$, the null hypothesis will be rejected, and if $T \geq 280$, then the null hypothesis will not be rejected.

Suppose that the actual temperature on Venus is indeed 300° C (or greater), as the report stated. If the sample mean has $T \geq 280$, then the null hypothesis will correctly be accepted. If the sample mean has $T < 280$ then the null hypothesis will incorrectly be rejected; this is a Type I error. On the other hand, if the actual temperature on Venus is less than 300° C, but the sample mean has $T \geq 280$, then the null hypothesis will incorrectly be accepted; this is a Type II error. If the sample mean has $T < 280$, then the null hypothesis will correctly be rejected.

5.5 More on Errors

5.5.1 Type I Errors and α-Risks

We can calculate the α-risk (that is, the probability of a Type I error) by drawing the normal curve assuming that the null hypothesis is true, then determining the area of the region which corresponds to the probability that the test results in the rejection of the null hypothesis.

Example 25:
A school district claims that an average of $7,650 is being spent on each child

each year. The state budget director, suspicious that some of the money allocated to the school district is not being spent on the schools, suspects that the figure is actually smaller. He will measure the exact amount of money spent on 36 students and wil reject the school district's claim if the average amount spent is less than \$7,200. If the standard deviation in amount spent per pupil is \$1,200, what is the α-risk?

The α-risk is the probability that a Type I error, where the null hypothesis is true but is rejected, so to calculate it, we will look at the case where the null hypothesis *is* true, i.e. $\mu = 7650$. Since the sample size is over 30, we apply the central limit theorem. The standard deviation of a sample of size 36 is $\frac{1200}{\sqrt{36}} = \frac{1200}{6} = 200$. The mean of the sample means is equal to the actual mean, which is 7650. Thus, the z-score that corresponds to 7200 is $\frac{7200-7650}{200} = -\frac{450}{200} = -2.25$. The area under the normal curve from -2.25 to 0 is 0.4878, so the area of the z-scores less than -2.25 is -0.0122. Thus, the α-risk is -0.0122.

In the previous example, we calculated the α-risk given the critical value(s). However, it is often more useful to determine critical values given an acceptable α-risk level (called the **significance level** of the test).

After we determine the critical values, it is a simple task to determine whether the sample mean (or, potentially, some other sample statistic) falls in the rejection range or the fail to reject range. This is how we draw conclusions: by showing that the null hypothesis is improbable.

> **Definition 28: p-value**.
> The p-value of a test is the smallest value of α for which the null hypothesis would be rejected. An alternative definition is the probability of obtaining the experimental result if the null hypothesis is true. Smaller p-values mean more significant differences between the null hypothesis and the sample result.

Example 26:
A drug company claims that their medication will render a patient unconscious in an average of 5.76 minutes. A researcher decides to test this claim with 300 patients. He obtains the following data: $\Sigma x = 2789$ *and* $\Sigma(x - \overline{x})^2 = 3128$. *Should the drug company's claim be rejected at a level of 10% significance?*

We calculate the sample mean and sample standard deviation as follows:

$$\begin{aligned} \overline{x} &= \frac{\Sigma x}{n} = \frac{2789}{300} \approx 9.297 \\ s &= \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}} = \sqrt{\frac{3128}{299}} \approx 3.234 \\ \sigma_{\overline{x}} &\approx \frac{s}{\sqrt{n}} = \frac{3.234}{300} \approx 0.187 \end{aligned}$$

A level of 10% significance means that $\alpha = 0.1$. Recall that a Type I error is one in which the null hypothesis is true but the data causes it to be rejected. So for our purposes, we will consider the case where the null hypothesis is true, and the time it takes to render a patient unconscious is 5.76 minutes. A Type I error will occur either if the sample mean is above or below 5.76, so we must consider both possibilites - this is called a two-sided, or two-tailed, test - the bottom 0.05 and the top 0.05.

According to the table of normal curve areas, between -1.645 and 0, the area under the normal curve is 0.45, so if the z-score is less than -1.645, the area is 0.05, as desired. Similarly, if the z-score is greater than 1.645, the area is also 0.05. So a Type I error will occur if the z-score is less than -1.645 or greater than 1.645. We must now convert to raw score. Since the mean of sample means is 5.76 and the standard deviation of sample means is 0.187, a z-score of -1.645 corresponds to a raw score of $5.76+0.187(-1.645) = 5.45$, and a z-score of 1.645 corresponds to a raw score of $5.76 + 0.187(1.645) = 6.07$. So the critical values of the experiment should be 5.45 and 6.07. Since $9.297 > 6.06$, the claim should be rejected.

5.5.2 Type II Errors and β-Risks

The β-risk (that is, the probability of failing to reject the null hypothesis) differs depending on the actual population statistic. We calculate it in a way similar to the calculation of the α-risk - by considering a possible actual value for the population, using this value as the center of the normal distribution, and calculating the area beyond the critical point that corresponds to the fail to reject zone.

Example 27:
A poll claims that children spend an average of 203 minutes per day watching television. A researcher believes this is too high and conducts a study of 144 patients. The sample standard deviation is 120 minutes. He will reject the

claim if the sample mean is under 180 minutes per day. What is the probability of a Type II error if the actual average is 210? 175? 170?

The standard deviation of the sample means is

$$\sigma_{\overline{x}} \approx \frac{s}{\sqrt{n}} = \frac{120}{\sqrt{144}} = \frac{120}{12} = 10$$

The null hypothesis will not be rejected if the sample mean is above 180. Thus, we look for the probability of the sample mean being above 180 to determine the β-risk. If the actual average is 210 minutes, then the sample means will be normally distributed about 210, so the z-score of the critical value, 180, is $\frac{180-210}{10} = -3$, which corresponds to a probability of 0.4987. Thus, the probability of the sample mean being less than 180 (i.e. rejection) is $0.5 - 0.4987$, and the probability of this *not* being the case (i.e. failure to reject) is $0.5 + 0.4987 = 0.9987$. This is the β-risk.

If the actual average is 175 minutes, then the sample means will be normally distributed about 175, so the z-score of the critical value, 180, is $\frac{180-175}{10} = 0.5$, which corresponds to a probability of 0.1915. Thus, the probability of the sample mean being greater than 180 is $0.5 - 0.1915 = .3085$.

If the actual average is 170 minutes, then the sample means will be normally distributed about 170, so the z-score of the critical value, 170, is $\frac{180-170}{10} = 1$, which corresponds to a probability of 0.3413. This is the probability that the sample mean is between 170 and 180. Thus, the probability of the sample mean being greater than 180 is $0.5 - 0.3413 = .1587$.

In the past example, we were given a critical value. Instead, we may be given an acceptable α-risk, from which we can calculate the critical value. We can then use the critical value to calculate β-risks for possible actual values. In general, we can use the following pathway:

acceptable α-risk (significance level) $\longleftrightarrow$ critical value(s) $\longrightarrow$ β-risk

5.6 Comparing Two Means

5.6.1 Confidence Interval Estimates

We can compare the means of two populations by comparing the means of their samples. More specifically, we can compute a confidence interval estimate for the difference between the two means (i.e. $\mu_1 - \mu_2$). We do this using the same method that we did for finding a confidence interval estimate for *one* value,

except that we use

$$\mu = \mu_1 - \mu_2 \tag{5.2}$$

$$\sigma_{\overline{x}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{5.3}$$

$$\overline{x} = \overline{x}_1 - \overline{x}_2 \tag{5.4}$$

Index

alternative hypothesis, 31

bar graph, 9
bimodal, 3
binomial distribution, 19
    in quality control systems, 20
Binomial Formula, 15
binomial random variable, 18
box and whisker display, 12

Central Limit Theorem, 28
Chebyshev's Theorem, 7
combination, 15
complimentary event, 13
confidence interval
    estimate of, 29

data types
    discrete, 2
    measured, 2
    numerical, 2
definitions
    alternative hypothesis, 31
    combination, 15
    complimentary event, 13
    expected value, 17
    factorial, 15
    histogram, 9
    independent, 14
    mean, 3
    median, 3
    mode, 3
    mutually exclusive, 13
    null hypothesis, 31
    p-value, 33
    percentile ranking, 6
    Poisson distribution, 21
    probability, 13
    probability distribution, 16
    random variable, 16
    range, 4
    relative variability, 5
    simple ranking, 6
    standard deviation, 4
    Type I error, 31
    Type II error, 32
    variance, 4
    z-score, 6
descriptive statistics, 2

empirical rule, 7
examples
    binomial distribution, 19, 20
    binomial formula, 15
    calculating an alpha risk, 32
    calculating beta-risk, 34
    calculating modes, 3
    central limit theorem, 29
    Chebyshev's Theorem, 7
    choosing a sample size, 30
    confidence interval estimates, 29
    constructing a histogram from z-scores, 11
    determining relative frequency from a histogram, 10
    determining statistical significance, 33
    expected value, 17
    independent and non-independent events, 14
    mutually exclusive events, 13
    normal as a binomial approximation, 26
    normal distribution, 24, 26

Poisson distribution, 22
Poisson distributions, 21
random variables, 16
simple, percentile, and z-score ranking, 6
standard deviation of a discrete random variable, 17
stem and leaf display, 11
testing a hypothesis and error types, 32
z-scores, 7
expected value, 17

factorial, 15
formulae
Binomial Formula, 15

histogram, 9, 10
and relative frequencies, 10
as compared to bar graph, 9
representation of mean, median, and mode, 10

Independence Principle, 14
independent, 14
inferential statistics, 2
interquartile range, 4

mean, 3, 10
arithmetic, 3
expected value, 17
on a histogram, 10
properties of, 4
trimmed mean, 4
Mean and Standard Deviation of a Binomial, 18
median, 3, 10
in a box and whisker display, 12
on a histogram, 10
mode, 3, 10
on a histogram, 10
mutually exclusive, 13

normal curve, 23
area of, 23
null hypothesis, 31

outlier, 4

p-value, 33
percentile ranking, 6
Poisson distribution, 21
as an approximation to the binomial, 22
Principle of Inclusion and Exclusion, 14
probability, 13
probability distribution, 16

random variable, 16
standard deviation of, 17
variance of, 17
range, 4
interquartile range, 4
relative frequency, 9
relative variability, 5

sample size, 30
simple ranking, 6
skewed to the left, 10
skewed to the right, 10
standard deviation, 4
approximation of, 7
of a discrete random variable, 17
of a sample, 30

theorems
Central Limit Theorem, 28
Chebyshev's Theorem, 7
empirical rule, 7
Independence Principle, 14
Mean and Standard Deviation of a Binomial, 18
Principle of Inclusion and Exclusion, 14
trimmed mean, 4
Type I error, 31
Type II error, 32

variance, 4
of a discrete random variable, 17
of population, 4

of sample, 4

z-score, 6