LECTURE NOTES

ON

BUSINESS STATISTICS

MBA I SEMESTER
(JNTUA-R14)

G. RAJASEKHAR
ASSOCIATE PROFESSOR

DEPARTMENT OF MANAGEMENT STUDIES
CHADALAWADA RAMANAMMA ENGINEERING COLLEGE
CHADALAWADA NAGAR, RENIGUNTA ROAD, TIRUPATI (A.P) - 517506
JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR
MBA Semester – I Th C

(14E00105) BUSINESS STATISTICS

The objective of this course is to familiarize the students with the statistical
techniques popularly used in managerial decision making. It also aims at developing the
computational skill of the students relevant for statistical analysis.

1. Introduction of statistics - Measures of Central Tendency - Arithmetic – Weighted
   mean – Median, Mode – Geometric mean and Harmonic mean – Measures of Dispersion,
   range, quartile deviation, mean deviation, standard deviation, coefficient of variation –
   Application of measures of central tendency and dispersion for business decision making.

2. Correlation: Introduction, Significance and types of correlation – Methods of
   correlation – Co-efficiency of correlation. Regression analysis – Meaning and utility of
   regression analysis – Comparison between correlation and regression – Interpretation of
   regression coefficients.

3. Probability – Meaning and definition of probability – Significance of probability in
   business application – Theories of probability – Addition and multiplication – Conditional
   laws of probability – Binominal – Poisson – Uniform – Normal and exponential
   distribution.

4. Testing of Hypothesis - Hypothesis testing: one sample and two sample tests for means
   and proportions of large samples (z-test), one sample and two sample tests for means of
   small samples (t-test), F-test for two sample standard deviations. ANOVA one and two
   way – Design of experiments.

5. Non-Parametric Methods: Chi-square test for single sample standard deviation. Chisquare
   tests for independence of attributes and goodness of fit - Sign test for paired data.

References:
- Statistics for Management, Richard I Levin, David S.Rubin, Pearson,
- Business Statistics, J.K.Sharma, Vikas house publications house Pvt Ltd
- Complete Business Statistics, Amir D. Aezel, Jayavel, TMH,
- Statistics for Management, P.N.Arora, S.Arora, S.Chand
- Business Statistics for Contemporary decision making, Black Ken, New age
  publishers.
- Statistical Methods, Gupta S.P., S.Chand.

Textbooks:

References:
3. Anderson D.R., Sweeney D.J. and Williams T.A., Statistics for business and economics,
UNIT I
FUNDAMENTALS OF PROBABILITY

Basic Definitions and Rules for Probability - Conditional Probability Independence of Events - Baye’s Theorem and Random Variables - Probability Distributions – Binomial - Poisson and Normal Distributions

Statistics

Statistics is the science of the collection, organization, and interpretation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments and closely related to probability theory, with which it is often grouped.

In statistics, a variable has two defining characteristics:

A variable is an attribute that describes a person, place, thing, or idea. The value of the variable can "vary" from one entity to another.

For example, a person’s hair color is a potential variable, which could have the value of "blond" for one person and "brunette" for another.

Categorical vs. Quantitative Variables

Variables can be classified as categorical (aka, qualitative) or quantitative (aka, numerical).

Categorical.

Categorical variables take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of categorical variables.

Quantitative. Quantitative variables are numerical. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

In algebraic equations, quantitative variables are represented by symbols (e.g., $x$, $y$, or $z$).

Discrete vs. Continuous Variables

Quantitative variables can be further classified as discrete or continuous. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete variable.
Some examples will clarify the difference between discrete and continuous variables.

Suppose the fire department mandates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.

Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.3 heads. Therefore, the number of heads must be a discrete variable.

Univariate vs. Bivariate Data

Statistical data is often classified according to the number of variables being studied.

Univariate data. When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

Bivariate data. When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

Statisticians use summary measures to describe patterns of data.

Measures of central tendency refer to the summary measures used to describe the most "typical" value in a set of values.

The Mean and the Median

The two most common measures of central tendency are the median and the mean, which can be illustrated with an example. Suppose we draw a sample of five women and measure their weights. They weigh 100 pounds, 100 pounds, 130 pounds, 140 pounds, and 150 pounds.
To find the median, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values. Thus, in the sample of five women, the median value would be 130 pounds; since 130 pounds is the middle weight.

The mean of a sample or a population is computed by adding all of the observations and dividing by the number of observations. Returning to the example of the five women, the mean weight would equal 

\[(100 + 100 + 130 + 140 + 150)/5 = 620/5 = 124 \text{ pounds.}\]

In the general case, the mean can be calculated, using one of the following equations:

- Population mean = \( \mu = \frac{\Sigma X}{N} \) OR \( \text{Sample mean} = \bar{x} = \frac{\Sigma x}{n} \)

where \( \Sigma X \) is the sum of all the population observations, \( N \) is the number of population observations, \( \Sigma x \) is the sum of all the sample observations, and \( n \) is the number of sample observations.

When statisticians talk about the mean of a population, they use the Greek letter \( \mu \) to refer to the mean score. When they talk about the mean of a sample, statisticians use the symbol \( x \) to refer to the mean score.

The Mean vs. the Median

As measures of central tendency, the mean and the median each have advantages and disadvantages. Some pros and cons of each measure are summarized below.

**The median may be a better indicator of the most typical value if a set of scores has an outlier. An outlier is an extreme value that differs greatly from other values.**

However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

To illustrate these points, consider the following example. Suppose we examine a sample of 10 households to estimate the typical family income. Nine of the households have incomes between $20,000 and $100,000; but the tenth household has an annual income of $1,000,000,000. That tenth household is an outlier. If we choose a measure to estimate the income of a typical household, the mean will greatly over-estimate the income of a typical family (because of the outlier); while the median will not.

**Effect of Changing Units**

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of central tendency are affected when we change units. If you add a constant to every value, the mean and median increase by the same constant. For example, suppose you have a set of scores with...
a mean equal to 5 and a median equal to 6. If you add 10 to every score, the new mean will be $5 + 10 = 15$; and the new median will be $6 + 10 = 16$.

Suppose you multiply every value by a constant. Then, the mean and the median will also be multiplied by that constant. For example, assume that a set of scores has a mean of 5 and a median of 6. If you multiply each of these scores by 10, the new mean will be $5 \times 10 = 50$; and the new median will be $6 \times 10 = 60$.

The Range

The range is the difference between the largest and smallest values in a set of values.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. For this set of numbers, the range would be 11 - 1 or 10.

The Interquartile Range (IQR)

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles.

Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

Q1 is the "middle" value in the first half of the rank-ordered data set. Q2 is the median value in the set.

Q3 is the "middle" value in the second half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. Q1 is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, Q1 = (3 + 4)/2 or Q1 = 3.5. Q3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, Q3 = (6 + 7)/2 or Q3 = 6.5. The interquartile range is Q3 minus Q1, so IQR = 6.5 - 3.5 = 3.

An Alternative Definition for IQR

In some texts, the interquartile range is defined differently. It is defined as the difference between the largest and smallest values in the middle 50% of a set of data.

To compute an interquartile range using this definition, first remove observations from the lower quartile. Then, remove observations from the upper quartile. Then, from the remaining observations, compute the difference between the largest and smallest values.
For example, consider the following numbers: 1, 3, 4, 3, 5, 6, 7, 11. After we remove observations from the lower and upper quartiles, we are left with: 4, 5, 5, 6. The interquartile range (IQR) would be 6 - 4 = 2.

When the data set is large, the two definitions usually produce the same (or very close) results. However, when the data set is small, the definitions can produce different results.

The Variance

In a population, variance is the average squared deviation from the population mean, as defined by the following formula:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

where $\sigma^2$ is the population variance, $\mu$ is the population mean, $X_i$ is the $i$th element from the population, and $N$ is the number of elements in the population. The variance of a sample, is defined by slightly different formula, and uses a slightly different notation:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

where $s^2$ is the sample variance, $\bar{x}$ is the sample mean, $x_i$ is the $i$th element from the sample, and $n$ is the number of elements in the sample. Using this formula, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate an unknown population variance, based on data from a sample, this is the formula to use.

The Standard Deviation

The standard deviation is the square root of the variance. Thus, the standard deviation of a population is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

where $\sigma$ is the population standard deviation, $\sigma^2$ is the population variance, $\mu$ is the population mean, $X_i$ is the $i$th element from the population, and $N$ is the number of elements in the population.

And the standard deviation of a sample is:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

where $s$ is the sample standard deviation, $s^2$ is the sample variance, $\bar{x}$ is the sample mean, $x_i$ is the $i$th element from the sample, and $n$ is the number of elements in the sample.
Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of variability are affected when we change units.

If you add a constant to every value, the distance between values does not change. As a result, all of the measures of variability (range, interquartile range, standard deviation, and variance) remain the same.

On the other hand, suppose you multiply every value by a constant. This has the effect of multiplying the range, interquartile range (IQR), and standard deviation by that constant. It has an even greater effect on the variance. It multiplies the variance by the square of the constant.

<table>
<thead>
<tr>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>An experiment is a situation involving chance or probability that leads to results called outcomes.</td>
<td>In the problem above, the experiment is spinning the spinner.</td>
</tr>
<tr>
<td>An outcome is the result of a single trial of an experiment.</td>
<td>The possible outcomes are landing on yellow, blue, green or red.</td>
</tr>
<tr>
<td>An event is one or more outcomes of an experiment.</td>
<td>One event of this experiment is landing on blue.</td>
</tr>
<tr>
<td>Probability is the measure of how likely an event is.</td>
<td>The probability of landing on blue is one fourth.</td>
</tr>
</tbody>
</table>

In order to measure probabilities, mathematicians have devised the following formula for finding the probability of an event.
Probability of an Event

The number of ways event A can occur

\[ P(A) = \frac{\text{The number of ways event A can occur}}{\text{The total number of possible outcomes}} \]

Rules for probability

Two events are mutually exclusive if they have no sample points in common.

The probability that event A occurs, given that event B has occurred, is called a conditional probability.

The conditional probability of A, given B, is denoted by the symbol \( P(A|B) \).

The probability that event A will not occur is denoted by \( P(A') \).

Bayes' theorem (also known as Bayes' rule) is a useful tool for calculating conditional probabilities. Bayes' theorem can be stated as follows:

Bayes' theorem.

Let \( A_1, A_2, \ldots, A_n \) be a set of mutually exclusive events that together form the sample space \( S \). Let \( B \) be any event from the same sample space, such that \( P(B) > 0 \). Then,

\[
P( A_k | B ) = \frac{P( A_k ) P( B | A_k )}{\sum_{i=1}^{n} P( A_i ) P( B | A_i )}
\]

Note: Invoking the fact that \( P( A_k \cap B ) = P( A_k ) P( B | A_k ) \), Baye's theorem can also be expressed as

\[
P( A_k ) P( B | A_k )
\]

\[
P( A_k | B ) = P( A_1 ) P( B | A_1 ) + P( A_2 ) P( B | A_2 ) + \ldots
\]

\[
+ P( A_n ) P( B | A_n )
\]
Unless you are a world-class statistician, Bayes' theorem (as expressed above) can be intimidating. However, it really is easy to use. The remainder of this lesson covers material that can help you understand when and how to apply Bayes' theorem effectively.

When to Apply Bayes' Theorem

Part of the challenge in applying Bayes' theorem involves recognizing the types of problems that warrant its use. You should consider Bayes' theorem when the following conditions exist.

- The sample space is partitioned into a set of mutually exclusive events \( \{ A_1, A_2, \ldots, A_n \} \).
- Within the sample space, there exists an event \( B \), for which \( P(B) > 0 \).
- The analytical goal is to compute a conditional probability of the form: \( P( A_k | B ) \).
- You know at least one of the two sets of probabilities described below.
  - \( P( A_k ) \) and \( P( B | A_k ) \) for each \( A_k \)

Sample Problem

Bayes' theorem can be best understood through an example. This section presents an example that demonstrates how Bayes' theorem can be applied effectively to solve statistical problems.

Example 1

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

Solution:

The sample space is defined by two mutually-exclusive events - it rains or it does not rain. Additionally, a third event occurs when the weatherman predicts rain. Notation for these events appears below.

- Event \( A_1 \). It rains on Marie's wedding.
Event A: It does not rain on Marie's wedding

Event B: The weatherman predicts rain.

In terms of probabilities, we know the following:

\[ P( A_1 ) = \frac{5}{365} = 0.0136985 \]  [It rains 5 days out of the year.]

\[ P( A_2 ) = \frac{360}{365} = 0.9863014 \]  [It does not rain 360 days out of the year.]

\[ P( B | A_1 ) = 0.9 \]  [When it rains, the weatherman predicts rain 90% of the time.]

\[ P( B | A_2 ) = 0.1 \]  [When it does not rain, the weatherman predicts rain 10% of the time.]

We want to know \( P( A_1 | B ) \), the probability it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes' theorem, as shown below.

\[
P( A_1 | B ) = \frac{P( A_1 ) P( B | A_1 )}{P( A_1 ) P( B | A_1 ) + P( A_2 ) P( B | A_2 )}
\]

\[
P( A_1 | B ) = \frac{(0.014)(0.9)}{(0.014)(0.9) + (0.986)(0.1)}
\]

\[ P( A_1 | B ) = 0.111 \]

Note the somewhat unintuitive result. Even when the weatherman predicts rain, it only rains only about 11% of the time. Despite the weatherman's gloomy prediction, there is a good chance that Marie will not get rained on at her wedding.

When the numerical value of a variable is determined by a chance event, that variable is called a random variable.

Discrete vs. Continuous Random Variables

Random variables can be discrete or continuous.

Discrete.

Discrete random variables take on integer values, usually the result of counting. Suppose, for example, that we flip a coin and count the number of heads. The number of heads results from a random process - flipping a coin. And the number of heads is represented by an integer value - a number between 0 and plus infinity. Therefore, the number of heads is a discrete random variable.
Continuous random variables, in contrast, can take on any value within a range of values. For example, suppose we flip a coin many times and compute the average number of heads per flip. The average number of heads per flip results from a random process - flipping a coin. And the average number of heads per flip can take on any value between 0 and 1, even a non-integer value. Therefore, the average number of heads per flip is a continuous random variable.

Binomial distribution

A binomial experiment (also known as a Bernoulli trial) is a statistical experiment that has the following properties:

The experiment consists of \( n \) repeated trials.
Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
The probability of success, denoted by \( P \), is the same on every trial. The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Consider the following statistical experiment. You flip a coin 2 times and count the number of times the coin lands on heads. This is a binomial experiment because:

The experiment consists of repeated trials. We flip a coin 2 times.
Each trial can result in just two possible outcomes - heads or tails.
The probability of success is constant - 0.5 on every trial. The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

Notation

The following notation is helpful, when we talk about binomial probability.

- \( x \): The number of successes that result from the binomial experiment.
- \( n \): The number of trials in the binomial experiment.
- \( P \): The probability of success on an individual trial.
- \( Q \): The probability of failure on an individual trial. (This is equal to 1 - \( P \).)

\( b(x; n, P) \): Binomial probability - the probability that an \( n \)-trial binomial experiment results in exactly \( x \) successes, when the probability of success on an individual trial is \( P \).
\( nCr \): The number of combinations of \( n \) things, taken at a time.

Binomial Distribution

A binomial random variable is the number of successes \( x \) in \( n \) repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution (also known as a Bernoulli distribution).

Suppose we flip a coin two times and count the number of heads (successes). The binomial random variable is the number of heads, which can take on values of 0, 1, or 2. The binomial distribution is presented below.

<table>
<thead>
<tr>
<th>Number of heads</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
</tr>
</tbody>
</table>

The binomial distribution has the following properties:

The mean of the distribution \( (\mu_x) \) is equal to \( n \times P \).

The variance \( (\sigma_x^2) \) is \( n \times P \times (1 - P) \).

The standard deviation \( (\sigma_x) \) is \( \sqrt{n \times P \times (1 - P)} \).

Binomial Probability

The binomial probability refers to the probability that a binomial experiment results in exactly \( x \) successes. For example, in the above table, we see that the binomial probability of getting exactly one head in two coin flips is 0.50. Given \( x \), \( n \), and \( P \), we can compute the binomial probability based on the following formula:

Binomial Formula. Suppose a binomial experiment consists of \( n \) trials and results in \( x \) successes. If the probability of success on an individual trial is \( P \), then the binomial probability is:

\[
b(x; n, P) = \binom{n}{x} \times P^x \times (1 - P)^{n-x}
\]
Example 1

Suppose a die is tossed 5 times. What is the probability of getting exactly 2 fours?

Solution:

This is a binomial experiment in which the number of trials is equal to 5, the number of successes is equal to 2, and the probability of success on a single trial is 1/6 or about 0.167. Therefore, the binomial probability is:

\[ b(2; 5, 0.167) = \binom{5}{2} \times (0.167)^2 \times (0.833)^3 \]

\[ b(2; 5, 0.167) = 0.161 \]

Cumulative Binomial Probability

A cumulative binomial probability refers to the probability that the binomial random variable falls within a specified range (e.g., is greater than or equal to a stated lower limit and less than or equal to a stated upper limit).

For example, we might be interested in the cumulative binomial probability of obtaining 45 or fewer heads in 100 tosses of a coin (see Example 1 below). This would be the sum of all these individual binomial probabilities.

\[ b(x \leq 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \ldots + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5) \]

Example 1

What is the probability of obtaining 45 or fewer heads in 100 tosses of a coin?

Solution:

To solve this problem, we compute 46 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

\[ b(x \leq 45; 100, 0.5) = b(x = 0; 100, 0.5) + b(x = 1; 100, 0.5) + \ldots + b(x = 44; 100, 0.5) + b(x = 45; 100, 0.5) = 0.184 \]

Example 2

The probability that a student is accepted to a prestigious college is 0.3. If 5 students from the same school apply, what is the probability that at most 2 are accepted?

Solution:
To solve this problem, we compute 3 individual probabilities, using the binomial formula. The sum of all these probabilities is the answer we seek. Thus,

\[
b(x \leq 2; 5, 0.3) = b(x = 0; 5, 0.3) + b(x = 1; 5, 0.3) + b(x = 2; 5, 0.3) = 0.1681 + 0.3601 + 0.3087
\]

\[
b(x \leq 2; 5, 0.3) = 0.8369
\]

Example 3

What is the probability that the world series will last 4 games? 5 games? 6 games? 7 games? Assume that the teams are evenly matched.

Solution:

This is a very tricky application of the binomial distribution. If you can follow the logic of this solution, you have a good understanding of the material covered in the tutorial, to this point.

In the world series, there are two baseball teams. The series ends when the winning team wins 4 games. Therefore, we define a success as a win by the team that ultimately becomes the world series champion.

For the purpose of this analysis, we assume that the teams are evenly matched. Therefore, the probability that a particular team wins a particular game is 0.5.

Let's look first at the simplest case. What is the probability that the series lasts only 4 games. This can occur if one team wins the first 4 games. The probability of the National League team winning 4 games in a row is:

\[
b(4; 4, 0.5) = \binom{4}{4} * (0.5)^4 * (0.5)^0 = 0.0625
\]

Similarly, when we compute the probability of the American League team winning 4 games in a row, we find that it is also 0.0625. Therefore, probability that the series ends in four games would be 0.0625 + 0.0625 = 0.125; since the series would end if either the American or National League team won 4 games in a row.

Now let's tackle the question of finding probability that the world series ends in 5 games. The trick in finding this solution is to recognize that the series can only end in 5 games, if one team has won 3 out of the first 4 games. So let's first find the probability that the American League team wins exactly 3 of the first 4 games.

\[
b(4; 4, 0.5) = \binom{4}{3} * (0.5)^3 * (0.5)^1 = 0.0625
\]
Given that the American League team has won 3 of the first 4 games, the American League team has a 50/50 chance of winning the fifth game to end the series. Therefore, the probability of the American League team winning the series in 5 games is 0.25 * 0.50 = 0.125. Since the National League team could also win the series in 5 games, the probability that the series ends in 5 games would be 0.125 + 0.125 = 0.25.

The rest of the problem would be solved in the same way. You should find that the probability of the series ending in 6 games is 0.3125; and the probability of the series ending in 7 games is also 0.3125.

While this is statistically correct in theory, over the years the actual world series has turned out differently, with more series than expected lasting 7 games. For an interesting discussion of why world series reality differs from theory,

Negative Binomial Experiment

A negative binomial experiment is a statistical experiment that has the following properties:

- The experiment consists of x repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by $p$, is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.
- The experiment continues until r successes are observed, where r is specified in advance.
- Consider the following statistical experiment. You flip a coin repeatedly and count the number of times the coin lands on heads. You continue flipping the coin until it has landed 5 times on heads. This is a negative binomial experiment because:
  - The experiment consists of repeated trials. We flip a coin repeatedly until it has landed 5 times on heads.
  - Each trial can result in just two possible outcomes - heads or tails.
  - The probability of success is constant - 0.5 on every trial.
The trials are independent; that is, getting heads on one trial does not affect whether we get heads on other trials.

The experiment continues until a fixed number of successes have occurred; in this case, 5 heads.

Notation

The following notation is helpful, when we talk about negative binomial probability.

\[ x \]: The number of trials required to produce \( r \) successes in a negative binomial experiment.

\[ r \]: The number of successes in the negative binomial experiment.

\[ p \]: The probability of success on an individual trial.

\[ q \]: The probability of failure on an individual trial. (This is equal to \( 1 - p \).)

\[ b^*(x; r, p) \]: Negative binomial probability - the probability that an \( x \)-trial negative binomial experiment results in the \( r \)th success on the \( x \)th trial, when the probability of success on an individual trial is \( p \).

\[ nC_r \]: The number of combinations of \( n \) things, taken \( r \) at a time.

Negative Binomial Distribution

A negative binomial random variable is the number \( X \) of repeated trials to produce \( r \) successes in a negative binomial experiment. The probability distribution of a negative binomial random variable is called a negative binomial distribution. The negative binomial distribution is also known as the Pascal distribution.

Suppose we flip a coin repeatedly and count the number of heads (successes). If we continue flipping the coin until it has landed 2 times on heads, we are conducting a negative binomial experiment. The negative binomial random variable is the number of coin flips required to achieve 2 heads. In this example, the number of coin flips is a random variable that can take on any integer value between 2 and plus infinity. The negative binomial probability distribution for this example is presented below.

<table>
<thead>
<tr>
<th>Number of coin flips</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Negative Binomial Probability

The negative binomial probability refers to the probability that a negative binomial experiment results in \( r - 1 \) successes after trial \( x - 1 \) and \( r \) successes after trial \( x \). For example, in the above table, we see that the negative binomial probability of getting the second head on the sixth flip of the coin is 0.078125.

Given \( x \), \( r \), and \( P \), we can compute the negative binomial probability based on the following formula:

\[
b(x; r, P) = \binom{x-1}{r-1} P^r (1 - P)^{x - r}
\]

The Mean of the Negative Binomial Distribution

If we define the mean of the negative binomial distribution as the average number of trials required to produce \( r \) successes, then the mean is equal to:

\[
\mu = \frac{r}{P}
\]

where \( \mu \) is the mean number of trials, \( r \) is the number of successes, and \( P \) is the probability of a success on any given trial.

Alternative Views of the Negative Binomial Distribution
As if statistics weren’t challenging enough, the above definition is not the only definition for the negative binomial distribution. Two common alternative definitions are:

The negative binomial random variable is $R$, the number of successes before the binomial experiment results in $k$ failures. The mean of $R$ is:

$$\mu_R = \frac{kP}{Q}$$

The negative binomial random variable is $K$, the number of failures before the binomial experiment results in $r$ successes. The mean of $K$ is:

$$\mu_K = \frac{rQ}{P}$$

The moral: If someone talks about a negative binomial distribution, find out how they are defining the negative binomial random variable. On this web site, when we refer to the negative binomial distribution, we are talking about the definition presented earlier. That is, we are defining the negative binomial random variable as $X$, the total number of trials required for the binomial experiment to produce $r$ successes.

Attributes of a Poisson Experiment

A Poisson experiment is a statistical experiment that has the following properties:

The experiment results in outcomes that can be classified as successes or failures.

The average number of successes ($\mu$) that occurs in a specified region is known.

The probability that a success will occur is proportional to the size of the region.

The probability that a success will occur in an extremely small region is virtually zero.

Note that the specified region could take many forms. For instance, it could be a length, an area, a volume, a period of time, etc.

Notation

The following notation is helpful, when we talk about the Poisson distribution.

- $\mu$: A constant equal to approximately 2.71828. (Actually, $\mu$ is the base of the natural logarithm system.)
- $\mu$: The mean number of successes that occur in a specified region.
**Poisson Distribution**

A Poisson random variable is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution. Given the mean number of successes ($\mu$) that occur in a specified region, we can compute the Poisson probability based on the following formula:

**Poisson Formula.** Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is $\mu$. Then, the Poisson probability is:

$$P(x; \mu) = \frac{(e^{-\mu})(\mu^x)}{x!}$$

where $x$ is the actual number of successes that result from the experiment, and $e$ is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to $\mu$.
- The variance is also equal to $\mu$.

**Example 1**

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

*Solution:* This is a Poisson experiment in which we know the following:

- $\mu = 2$; since 2 homes are sold per day, on average.
- $x = 3$; since we want to find the likelihood that 3 homes will be sold tomorrow.
- $e = 2.71828$; since $e$ is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(3; 2) = \frac{(2.71828^{-2})(2^3)}{3!}$$
Thus, the probability of selling 3 homes tomorrow is 0.180.

**Example 1**

Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

**Solution:**

This is a Poisson experiment in which we know the following:

- \( \mu = 5 \); since 5 lions are seen per safari, on average.
- \( x = 0, 1, 2, \) or 3; since we want to find the likelihood that tourists will see fewer than 4 lions; that is, we want the probability that they will see 0, 1, 2, or 3 lions.
- \( e = 2.71828 \); since \( e \) is a constant equal to approximately 2.71828.

To solve this problem, we need to find the probability that tourists will see 0, 1, 2, or 3 lions. Thus, we need to calculate the sum of four probabilities: \( P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5) \). To compute this sum, we use the Poisson formula:

\[
P(x < 3, 5) = P(0; 5) + P(1; 5) + P(2; 5) + P(3; 5)
\]

\[
P(x < 3, 5) = \frac{(-5)^0 e^{-5}}{0!} + \frac{(-5)^1 e^{-5}}{1!} + \frac{(-5)^2 e^{-5}}{2!} + \frac{(-5)^3 e^{-5}}{3!}
\]

\[
P(x < 3, 5) = \frac{0.006738}{1} + \frac{0.006738 	imes 5}{1} + \frac{0.006738 	imes 25}{2} + \frac{0.006738 	imes 125}{6}
\]

\[
P(x < 3, 5) = 0.0067 + 0.03369 + 0.084224 + 0.140375
\]

\[
P(x < 3, 5) = 0.2650
\]

**Department of Management Studies**
The normal distribution refers to a family of continuous probability distributions described by the normal equation.

The Normal Equation

The normal distribution is defined by the following equation:

Normal equation. The value of the random variable $Y$ is:

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

where $X$ is a normal random variable, $\mu$ is the mean, $\sigma$ is the standard deviation, $\pi$ is approximately 3.14159, and $e$ is approximately 2.71828.

The random variable $X$ in the normal equation is called the normal random variable. The normal equation is the probability density function for the normal distribution.

The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.
Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

The total area under the normal curve is equal to 1.

The probability that a normal random variable $X$ equals any particular value is 0.

The probability that $X$ is greater than $a$ equals the area under the normal curve bounded by $a$ and plus infinity (as indicated by the non-shaded area in the figure below).

The probability that $X$ is less than $a$ equals the area under the normal curve bounded by $a$ and minus infinity (as indicated by the shaded area in the figure below).

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

Collectively, these points are known as the empirical rule or the 68-95-99.7 rule. Clearly, given a normal distribution, most outcomes will be within 3 standard deviations of the mean.

To find the probability associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table. In the examples below, we illustrate the use of Stat Trek's Normal Distribution Calculator, a free tool available on this site. In the next lesson, we demonstrate the use of normal distribution tables.

Example 1
An average light bulb manufactured by the Acme Corporation lasts 300 days with a standard deviation of 50 days. Assuming that bulb life is normally distributed, what is the probability that an Acme light bulb will last at most 365 days?

Solution:

Given a mean score of 300 days and a standard deviation of 50 days, we want to find the cumulative probability that bulb life is less than or equal to 365 days. Thus, we know the following:

- The value of the normal random variable is 365 days.
- The mean is equal to 300 days.
- The standard deviation is equal to 50 days.

We enter these values into the Normal Distribution Calculator and compute the cumulative probability. The answer is: \( P( X \leq 365) = 0.90 \). Hence, there is a 90% chance that a light bulb will burn out within 365 days.

Example 2

Suppose scores on an IQ test are normally distributed. If the test has a mean of 100 and a standard deviation of 10, what is the probability that a person who takes the test will score between 90 and 110?

Solution:

Here, we want to know the probability that the test score falls between 90 and 110. The "trick" to solving this problem is to realize the following:

\[ P( 90 < X < 110 ) = P( X < 110 ) - P( X < 90 ) \]

We use the Normal Distribution Calculator to compute both probabilities on the right side of the above equation.

To compute \( P( X < 110 ) \), we enter the following inputs into the calculator: The value of the normal random variable is 110, the mean is 100, and the standard deviation is 10. We find that \( P( X < 110 ) \) is 0.84.

To compute \( P( X < 90 ) \), we enter the following inputs into the calculator: The value of the normal random variable is 90, the mean is 100, and the standard deviation is 10. We find that \( P( X < 90 ) \) is 0.16.

We use these findings to compute our final answer as follows:

\[ P( 90 < X < 110 ) = P( X < 110 ) - P( X < 90 ) \]

Department of Management Studies
P(90 < X < 110) = 0.84 - 0.16

Thus, about 68% of the test scores will fall between 90 and 110.

Standard Normal Distribution

The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

The normal random variable of a standard normal distribution is called a standard score or a z-score. Every normal random variable $X$ can be transformed into a z-score via the following equation:

$$z = \frac{X - \mu}{\sigma}$$

where $X$ is a normal random variable, $\mu$ is the mean of $X$, and $\sigma$ is the standard deviation of $X$.

Standard Normal Distribution Table

A standard normal distribution table shows a cumulative probability associated with a particular z-score. Table rows show the whole number and tenths place of the z-score. Table columns show the hundredths place. The cumulative probability (often from minus infinity to the z-score) appears in the cell of the table.

For example, a section of the standard normal table is reproduced below. To find the cumulative probability of a z-score equal to -1.31, cross-reference the row of the table containing -1.3 with the column containing 0.01. The table shows that the probability that a standard normal random variable will be less than -1.31 is 0.0951; that is, $P(Z < -1.31) = 0.0951$.

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Of course, you may not be interested in the probability that a standard normal random variable falls between minus infinity and a given value. You may want to know the probability that it lies between a given value and plus infinity. Or you may want to know the probability that a standard normal random variable lies between two given values. These probabilities are easy to compute from a normal distribution table. Here's how.

Find $P(Z > a)$. The probability that a standard normal random variable $Z$ is greater than a given value $a$ is easy to find. The table shows the $P(Z < a)$. The $P(Z > a) = 1 - P(Z < a)$. 

<table>
<thead>
<tr>
<th>Z</th>
<th>0.080</th>
<th>0.079</th>
<th>0.077</th>
<th>0.076</th>
<th>0.074</th>
<th>0.073</th>
<th>0.072</th>
<th>0.070</th>
<th>0.069</th>
<th>0.068</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.4</td>
<td>8</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Z</th>
<th>0.096</th>
<th>0.095</th>
<th>0.093</th>
<th>0.091</th>
<th>0.090</th>
<th>0.088</th>
<th>0.086</th>
<th>0.085</th>
<th>0.083</th>
<th>0.082</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.3</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Z</th>
<th>0.115</th>
<th>0.113</th>
<th>0.111</th>
<th>0.109</th>
<th>0.107</th>
<th>0.105</th>
<th>0.103</th>
<th>0.102</th>
<th>0.100</th>
<th>0.098</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Z</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.998</th>
<th>0.999</th>
<th>0.999</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Suppose, for example, that we want to know the probability that a z-score will be greater than 3.00. From the table (see above), we find that \( P(Z < 3.00) = 0.9987 \). Therefore, \( P(Z > 3.00) = 1 - P(Z < 3.00) = 1 - 0.9987 = 0.0013 \).

Find \( P(a < Z < b) \). The probability that a standard normal random variable lies between two values is also easy to find. The \( P(a < Z < b) = P(Z < b) - P(Z < a) \).

For example, suppose we want to know the probability that a z-score will be greater than -1.40 and less than -1.20. From the table (see above), we find that \( P(Z < -1.20) = 0.1151 \); and \( P(Z < -1.40) = 0.0808 \).

Therefore, \( P(-1.40 < Z < -1.20) = P(Z < -1.20) - P(Z < -1.40) = 0.1151 - 0.0808 = 0.0343 \).

In school or on the Advanced Placement Statistics Exam, you may be called upon to use or interpret standard normal distribution tables. Standard normal tables are commonly found in appendices of most statistics texts.

The Normal Distribution as a Model for Measurements

Often, phenomena in the real world follow a normal (or near-normal) distribution. This allows researchers to use the normal distribution as a model for assessing probabilities associated with real-world phenomena. Typically, the analysis involves two steps.

Transform raw data. Usually, the raw data are not in the form of z-scores. They need to be transformed into z-scores, using the transformation equation presented earlier: \( z = (X - \mu) / \sigma \).

Find probability. Once the data have been transformed into z-scores, you can use standard normal distribution tables, online calculators (e.g., Stat Trek’s free normal distribution calculator), or handheld graphing calculatorsto find probabilities associated with the z-scores.

Important two mark questions: What is meant by probability? What are the properties of normal distribution? What are the applications of poisson? Define bayes theorem. Important 16 mark question:

Explain the properties of normal distribution and explain with examples.
Introduction to Sampling Distributions - Sampling Distribution of Sample Mean and Sample Proportion - Application of Central Limit Theorem - Sampling Techniques - Estimation and Confidence Intervals - Point and Confidence Interval Estimates for Population Parameters of Large Sample and Small Samples - Determining the Sample Size.

Sampling distribution:

A sampling distribution or finite-sample distribution is the distribution of a given statistic based on a random sample of size n. It may be considered as the distribution of the statistic for all possible samples of a given size. The sampling distribution depends on the underlying distribution of the population, the statistic being considered, and the sample size used. The sampling distribution is frequently opposed to the asymptotic distribution, which corresponds to the limit case $n \to \infty$.

For example, consider a normal population with mean $\mu$ and variance $\sigma^2$. Assume we repeatedly take samples of a given size from this population and calculate the arithmetic mean for each sample—this statistic is called the sample mean. Each sample has its own distribution of the sample mean.

This is an example of a simple statistic taken from one of the simplest statistical populations. For other statistics and other populations, the formulas are frequently more complex. In such cases, the sampling distributions may be approximated through Monte-Carlo simulations, bootstrap method, or asymptotic distribution theory.

The standard deviation of the sampling distribution of the statistic is referred to as the standard error of that quantity. For the case where the statistic is the sample mean, the standard error is:

$$\sigma \sqrt{\frac{1}{n}}$$
A very important implication of this formula is that you must quadruple the sample size (4x) to achieve half (1/2) the measurement error. When designing statistical studies where cost is a factor, this may have a factor in understanding cost-benefit tradeoffs.

Alternatively, consider the sample median from the same population. It has a different sampling distribution which is generally not normal (but may be close under certain circumstances).

<table>
<thead>
<tr>
<th>Examples</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Sampling</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| distribution |
| --- | --- |
| | |
Bernoulli:

Two independent normal populations: and

Any absolutely continuous distribution $F$ with density $f$

Any distribution with distribution function $F$

Sampling Distribution of the Mean

Sample mean from samples of size $n$

Sample proportion of successful trials

Difference between sample means,

\[ \text{Median} = X(k - \frac{\sum_{i=1}^{k} x_i}{k}) \quad \text{Size } n = 2k \]

Maximum from a random sample of size $n$

The sampling distribution of the mean is a very important distribution. In later chapters you will see that it is used to construct confidence intervals for the mean and for significance testing.
Given a population with a mean of \( \mu \) and a standard deviation of \( \sigma \), the sampling distribution of the mean has a mean of \( \mu \) and a standard deviation of

\[
\sigma_M = \frac{\sigma}{\sqrt{n}}
\]

where \( n \) is the sample size. The standard deviation of the sampling distribution of the mean is called the standard error of the mean. It is designated by the symbol: \( \sigma_M \)

Note that the spread of the sampling distribution of the mean decreases as the sample size increases.

An example of the effect of sample size is shown above. Notice that the mean of the distribution is not affected by sample size.

**Sampling Distribution of a Proportion**

Assume that 0.80 of all third grade students can pass a test of physical fitness. A random sample of 20 students is chosen: 13 passed and 7 failed. The parameter \( \pi \) is used to designate the proportion of subjects in the population that pass (.80 in this case) and the statistic \( p \) is used to designate the proportion who pass in a sample (13/20 = .65 in this case). The sample size (\( N \)) in this example is 20. If repeated samples of size \( N \) where taken from the population and the proportion passing (\( p \)) were determined for each sample, a distribution of values of \( p \) would be formed. If the sampling went on forever,
the distribution would be the sampling distribution of a proportion. The sampling distribution of a proportion is equal to the binomial distribution. The mean and standard deviation of the binomial distribution are: \( \mu = \pi \)

\[
\sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}}
\]

and For the present example, \( N = 20, \pi = 0.80 \), the mean of the sampling distribution of \( p \) (\( \mu \)) is .8 and the standard error of \( p \) (\( \sigma_p \)) is 0.089. The shape of the binomial distribution depends on both \( N \) and \( \pi \). With large values of \( N \) and values of \( \pi \) in the neighborhood of .5, the sampling distribution is very close to a normal distribution.

Applications of the Central Limit Theorem

Theorem 1 Suppose that \( X_1; X_2; \ldots \) is an infinite sequence of independent, identically distributed random variables with common mean \( \mu = E(X_1) \) and finite variance \( \sigma^2 = V(X_1) \). Then, if we let \( S_n = X_1 + X_2 + \ldots + X_n \) we have that lim

\[
\lim_{n \to \infty} P \left( \frac{S_n - n\mu}{\sigma} \right) = \Phi(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-x^2/2} dx
\]

There are many applications of this theorem to real-world problems, and in these notes we will give two: An application to hypothesis testing, and an application to noise cancellation.

Hypothesis Testing

Here we will give an example of how to use the CLT to test hypotheses. We have already seen how to do this using a chi-square test to determine whether to reject a hypothesized population distribution (with finitely many classes) as being false. Here we will do this for when the population breaks down into two classes, smokers and non-smokers. Rejecting and Accepting Hypotheses

Before we embark on the stated goal of this section, it is worth mentioning the difference between accepting and rejecting hypotheses, and what sorts of conclusions we can draw from statistical sampling.
A primary misconception of certain statistical tests is that if a certain set of data supporting a hypothesis passes the test, then the hypothesis is likely true. This is the case of some statistical tests, but not of others.

The Main Problem

Problem. You read in a newspaper that 20% of Georgians smoke, and you decide to test this hypothesis by doing a poll on 10,000 randomly selected Georgians with replacement (if the population you are testing is very large, then you would not need to test with replacement). Suppose that 2050 of the responses are "smoker", while 7950 are "non-smoker". Is the claim "20% of Georgians smoke" unreasonable?

Well, in order to answer this question we would need more information; we would need to know what we mean by "unreasonable". Here we will mean "unreasonable" with respect to a certain statistical test which we presently describe:

Let \( X_i = 1 \) if respondent i says he/she is a smoker, and let \( X_i = 0 \) if he/she is not a smoker. These \( X_i \)'s as independent Bernoulli random variables. Let

\[
S_n = X_1 + \ldots + X_{10000}.
\]

If our hypothesis that 20% of Georgians smoke were correct, then \( \mu = E(X_i) = 0.2 \); and so, the Central Limit Theorem would tell us that \( S_{10000} \) is approximately \( N(0; 1) \); (1) in the sense that

\[
P \left( \frac{S_{10000} - 2000}{40} \right) \leq c
\]

That is, we observe \( S_{10000} \) smokers. Then, on the basis of the central limit theorem and (1) we wouldn't expect that is an atypical value for \( N(0; 1) \). In particular, we wouldn't expect that \( P \left( N(0; 1) \geq \frac{S_{10000} - 2000}{40} \right) < 0.05 \)

if \( \mu = 0.2 \) is the true mean.

Thus, we have the following basic statistical test

Statistical Test. Fix an \( \alpha > 0 \), typically \( \alpha = 0.05 \) or 0.01. Compute as in

\[
(2) \quad IIFP(\mu; 0) 2 ( \mu ) = 1 \quad \mu ( ) + \mu ( ) < \mu
\]

then we reject the hypothesis that the mean value of \( X_1 \) is \( \mu \); and, if this inequality is not satisfied, we do not reject it, which is not the same as saying we accept it (our world, and in fact natural human language, is built on multi-valued logic, where we deem hypotheses to be True and False, as well as undecided. 2)
and one can readily compute that $P(\mathcal{N}(0, 1) \geq 1.25) = 0.09$ (in fact, this probability is around 0.75 or so). Thus, we do not reject the hypothesis that 20% of Georgians smoke.

Noise Cancellation

Suppose that a man is driving through the desert, and runs out of gas. He grabs his cellphone to make a call for help, dialing 911, but he is just at the edge of the broadcast range for his cellphone, and so his call to the 911 dispatcher is somewhat noisy and garbled. Suppose that the 911 dispatcher has the ability to use several cellphone towers to clean up the signal. Suppose that there are about 100 towers near to the stranded driver, and suppose that the signals they each receive at a particular instant in time is given by

$$X_1, \ldots, X_{100};$$

where

$$X_i = S + Y_i;$$

where $S$ is the true signal being sent to the towers, and where $Y_i$ is the noise. Suppose that all the noises $Y_1, \ldots, Y_{100}$ are independent and identically distributed. In fact, it is not unreasonable to assume that they are all normal with mean 0 and variance $\sigma^2$, though we will not need this assumption for what follows. The 911 dispatcher cleans up the signal by computing the average

$$X = \frac{X_1 + \cdots + X_{100}}{100} = S + Y_1 + \cdots + Y_{100}$$

Now, by the Central Limit Theorem, we would expect that $Y_1 + \cdots + Y_{100}$ is approximately $\mathcal{N}(0; \sigma^2)$. (3)

Of course we need to be careful here; the central limit theorem says that random variables such as $(X_1 + \cdots + X_n)/\sqrt{n}$ are approximately normal provided that $n$ is sufficiently large, but does not say that in our particular case we can take $n$ around 100 and get (3) to hold. There are more powerful versions of the central limit theorem, which give conditions on $n$ under which (3) holds under a precise notion of "is approximately". At any rate, if we assume that the $Y_i$s are all independent normal random variables, then we don't even need the central limit theorem, because in that case we have that $X - S$ is exactly $\mathcal{N}(0; \sigma^2)$. (3)

Now, suppose that, in fact, all the noises $Y_i$s have variance $\sigma^2 = 1$. Then, the central limit theorem in the guise (3) would be telling us that the new noise $X - S$ is approximately normal with variance 1=100, a 100-fold improvement in the noise variance! Just How Good is the Averaging Method for Noise Cancellation Can we Do Better?

It turns out that not only does averaging give us a pretty good way to cancel noise, but it is, in some sense, the best thing we could possibly try. The proper language is that taking the average $X$ gives us a
maximum likelihood estimate for the signal S, which is the same as the expected value of Xi for all i = 1; 100. Let us make this more precise:

Maximum Likelihood Estimates. Suppose X is some random variable having a distribution that depends on a list of unknown, underlying parameters \( \beta_1; \ldots; \beta_k \). Let \( f(x; \beta_1; \ldots; \beta_k) \) denote the pdf for X, given the parameters \( \beta_1; \ldots; \beta_k \). Suppose we make n independent observations of our random source X, and suppose these observations are the values \( x_1; \ldots; x_n \). Then, the likelihood value of this observation is

\[
L(x_1; \ldots; x_n; \beta_1; \ldots; \beta_k) = n \prod_{i=1}^{n} f(x_i; \beta_1; \ldots; \beta_k)
\]

Then, we say that \( \hat{\beta}_1; \ldots; \hat{\beta}_k \) are maximum likelihood estimates for \( \beta_1; \ldots; \beta_k \), given the observations \( x_1; \ldots; x_n \), if these values \( \hat{\beta}_i \) maximize \( L(x_1; \ldots; x_n; \beta_1; \ldots; \beta_k) \).

Note that the \( \hat{\beta}_i \)s which maximize L may not be unique (there may be more than one global max). In the discrete setting it is easy to describe what this means, because in that setting \( L(x_1; \ldots; x_n; \beta_1; \ldots; \beta_k) \) is just the probability that if we make n independent observations of X, that those observations are \( x_1; \ldots; x_n \), given the distribution parameters \( \beta_1; \ldots; \beta_k \). So, finding \( \hat{\beta}_i \) that maximize this probability would maximize the likelihood (probability) that this sequence \( x_1; \ldots; x_n \) is observed.

In our case, let us suppose that the received signals \( X_i \) are, in fact, normal, with mean S, and variance \( \sigma^2 \); that is to say, the noises \( Y_i \) are \( N(0; \sigma^2) \). Now suppose that we have definite values for these observations (that is, our observed signals are `instantiated'), and suppose that those values are \( x_1; \ldots; x_{100} \). The likelihood function here is

\[
L(x_1; \ldots; x_{100}; S; \sigma^2) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - S)^2}{2\sigma^2}\right)
\]

For example, perhaps X is normal with mean \( \mu \) and variance \( \sigma^2 \), where neither \( \mu \) nor \( \sigma^2 \) are known, in which case \( \hat{\mu} = \bar{x} \) and \( \hat{\sigma}^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \).

If we seek S which maximizes this (for any given value for \( \sigma^2 \)), we can ignore the factor \( (2\pi)^{50} \sigma^{100} \), and we maximize the log of the remaining exponential factor; thus, we just need to maximize

\[
(x_1 - S)^2 + \ldots + (x_{100} - S)^2\sigma^2
\]

We can ignore the \( \sigma^2 \); so, we maximize

\[
(x_1 - S)^2 + \ldots + (x_{100} - S)^2
\]

Taking a derivative with respect to S and setting equal to 0 we have \( (x_1 + \ldots + x_{100}) 100S = 0 \):

So,

\[
S = \frac{1}{100} (x_1 + \ldots + x_{100})
\]

which is our sample mean. The fact that the expression (4) is a down-turning parabola means that, indeed, this is a maximum.
Thus, we see that by averaging we obtain a maximum likelihood estimate for $S$, and therefore, in some sense, this is the best we could hope to do to recover $S$.

Sampling methods

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

Probability and nonprobability sampling

A probability sampling scheme is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (In effect, the person who is selected from that household is taken as representing the person who isn't selected.)

In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population does have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

- Every element has a known nonzero probability of being sampled and involves random selection at some point.
Nonprobability sampling is any sampling method where some elements of the population have no chance of selection (these are sometimes referred to as 'out of coverage'/'undercovered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, nonprobability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example:

We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a nonprobability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling. In addition, nonresponse effects may turn any probability design into a nonprobability design if the characteristics of nonresponse are not well understood, since nonresponse effectively modifies each element's probability of being sampled.

Simple random sampling

In a simple random sample (‘SRS’) of a given size, all such subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection: the frame is not subdivided or partitioned. Furthermore, any given pair of elements has the same chance of selection as any other such pair (and similarly for triples, and so on). This minimises bias and simplifies analysis of results. In particular, the variance between individual results within the sample is a good indicator of variance in the overall population, which makes it relatively easy to estimate the accuracy of results.

However, SRS can be vulnerable to sampling error because the randomness of the selection may result in a sample that doesn’t reflect the makeup of the population. For instance, a simple random sample of ten people from a given country will on average produce five men and five women, but any given trial is likely to overrepresent one sex and underrepresent the other. Systematic and stratified techniques, discussed below, attempt to overcome this problem by using information about the population to choose a more representative sample.
SRS may also be cumbersome and tedious when sampling from an unusually large target population. In some cases, investigators are interested in research questions specific to subgroups of the population. For example, researchers might be interested in examining whether cognitive ability as a predictor of job performance is equally applicable across racial groups. SRS cannot accommodate the needs of researchers in this situation because it does not provide subsamples of the population. Stratified sampling, which is discussed below, addresses this weakness of SRS.

Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

Systematic sampling

Systematic sampling relies on arranging the target population according to some orderingscheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every kth element from then onwards. In this case, \( k = \frac{\text{population size}}{\text{sample size}} \). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the kth element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

As long as the starting point is randomized, systematic sampling is a type of probability sampling. It is easy to implement and the stratification induced can make it efficient. If the variable by which the list is ordered is correlated with the variable of interest. 'Every 10th' sampling is especially useful for efficient sampling from databases.

Example: Suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting (e.g.) every 10th street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts. (Note that if we always start at house #1 and end at #991, the sample is slightly biased towards the low end; by randomly selecting the start between #1 and #10, this bias is eliminated.)
However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate than simple random sampling.

Example: Consider a street where the odd-numbered houses are all on the north (expensive) side of the road, and the even-numbered houses are all on the south (cheap) side. Under the sampling scheme given above, it is impossible to get a representative sample; either the houses sampled will all be from the odd-numbered, expensive side, or they will all be from the even-numbered, cheap side.

Another drawback of systematic sampling is that even in scenarios where it is more accurate than SRS, its theoretical properties make it difficult to quantify that accuracy. (In the two examples of systematic sampling that are given above, much of the potential sampling error is due to variation between neighbouring houses - but because this method never selects two neighbouring houses, the sample will not give us any information on that variation.)

As described above, systematic sampling is an EPS method, because all elements have the same probability of selection (in the example given, one in ten). It is not 'simple random sampling' because different subsets of the same size have different selection probabilities - e.g. the set \{4,14,24,...,994\} has a one-in-ten probability of selection, but the set \{4,13,24,34,...\} has zero probability of selection.

Systematic sampling can also be adapted to a non-EPS approach; for an example, see discussion of PPS samples below.

Stratified sampling

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.
Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group’s size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling). A stratified sampling approach is most effective when three conditions are met: Variability within strata are minimized, Variability between strata are maximized, and The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages over other sampling methods
Focuses on important subpopulations and ignores irrelevant ones. Allows use of different sampling techniques for different subpopulations. Improves the accuracy/efficiency of estimation. Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size.

Disadvantages
Requires selection of relevant stratification variables which can be difficult.
Is not useful when there are no homogeneous subgroups.

Can be expensive to implement.

Poststratification

Stratification is sometimes introduced after the sampling phase in a process called "poststratification". This approach is typically implemented due to a lack of prior knowledge of an appropriate stratifying variable or when the experimenter lacks the necessary information to create a stratifying variable during the sampling phase. Although the method is susceptible to the pitfalls of post hoc approaches, it can provide several benefits in the right situation. Implementation usually follows a simple random sample. In addition to allowing for stratification on an ancillary variable, poststratification can be used to implement weighting, which can improve the precision of a sample's estimates.[3]

Oversampling

Choice-based sampling is one of the stratified sampling strategies. In choice-based sampling,[4] the data are stratified on the target and a sample is taken from each strata so that the rare target class will be more represented in the sample. The model is then built on this biased sample. The effects of the input variables on the target are often estimated with more precision with the choice-based sample even when a smaller overall sample size is taken, compared to a random sample. The results usually must be adjusted to correct for the oversampling.

Probability proportional to size sampling

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1. In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490 students respectively (total 1500 students), and we want to use student population as the basis for a PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school (1011 to 1500). We then generate a
Random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

The PPS approach can improve accuracy for a given sample size by concentrating sample on large elements that have the greatest impact on population estimates. PPS sampling is commonly used for surveys of businesses, where element size varies greatly and auxiliary information is often available - for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel's number of rooms as an auxiliary variable. In some cases, an older measurement of the variable of interest can be used as an auxiliary variable when attempting to produce more current estimates.

Cluster sampling

Sometimes it is cheaper to 'cluster' the sample in some way e.g. by selecting respondents from certain areas only, or certain time-periods only. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.)

Cluster sampling is an example of 'two-stage sampling' or 'multistage sampling': in the first stage a sample of areas is chosen; in the second stage a sample of respondents within those areas is selected.

This can reduce travel and other administrative costs. It also means that one does not need a sampling frame listing all elements in the target population. Instead, clusters can be chosen from a cluster-level frame, with an element-level frame created only for the selected clusters. Cluster sampling generally increases the variability of sample estimates above that of simple random sampling, depending on how the clusters differ between themselves, as compared with the within-cluster variation.

Nevertheless, some of the disadvantages of cluster sampling are the reliance of sample estimate precision on the actual clusters chosen. If clusters chosen are biased in a certain way, inferences drawn about population parameters from these sample estimates will be far off from being accurate.

Multistage sampling Multistage sampling is a complex form of cluster sampling in which two or more levels of units are embedded one in the other. The first stage consists of constructing the clusters that
will be used to sample from. In the second stage, a sample of primary units is randomly selected from each cluster (rather than using all units contained in all selected clusters). In following stages, in each of those selected clusters, additional samples of units are selected, and so on. All ultimate units (individuals, for instance) selected at the last step of this procedure are then surveyed.

This technique, thus, is essentially the process of taking random samples of preceding random samples. It is not as effective as true random sampling, but it probably solves more of the problems inherent to random sampling. Moreover, it is an effective strategy because it banks on multiple randomizations. As such, it is extremely useful.

Multistage sampling is used frequently when a complete list of all members of the population does not exist and is inappropriate. Moreover, by avoiding the use of all sample units in all selected clusters, multistage sampling avoids the large, and perhaps unnecessary, costs associated with traditional cluster sampling.

Matched random sampling

A method of assigning participants to groups in which pairs of participants are first matched on some characteristic and then individually assigned randomly to groups.[5]

The procedure for matched random sampling can be briefed with the following contexts,

Two samples in which the members are clearly paired, or are matched explicitly by the researcher. For example, IQ measurements or pairs of identical twins.

Those samples in which the same attribute, or variable, is measured twice on each subject, under different circumstances. Commonly called repeated measures. Examples include the times of a group of athletes for 1500m before and after a week of special training; the milk yields of cows before and after being fed a particular diet.

Quota sampling

In quota sampling, the population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.
It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for many years.

Convenience sampling or Accidental Sampling

Convenience sampling (sometimes known as grab or opportunity sampling) is a type of nonprobability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a sample population selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer was to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey was to be conducted at different times of day and several times per week. This type of sampling is most useful for pilot testing. Several important considerations for researchers using convenience samples include:

Are there controls within the research design or experiment which can serve to lessen the impact of a non-random convenience sample, thereby ensuring the results will be more representative of the population?

Is there good reason to believe that a particular convenience sample would or should respond or behave differently than a random sample from the same population?

Is the question being asked by the research one that can adequately be answered using a convenience sample?

In social science research, snowball sampling is a similar technique, where existing study subjects are used to recruit more subjects into the sample.

Line-intercept sampling
Line-intercept sampling is a method of sampling elements in a region whereby an
FMFNFOUJF
TBNQMFEJGBDIPTFOMJOFTFHNFOU
DBMMFEBʕUSBOTFDUʚ
JOUFSTFDUTUIFFMFNFOU
Panel sampling

Panel sampling is the method of first selecting a group of participants through a random sampling method and then asking that group for the same information again several times over a period of time. Therefore, each participant is given the same survey or interview at two or more time points; each period of data collection is called a "wave". This sampling methodology is often chosen for large scale or nationwide studies in order to gauge changes in the population with regard to any number of variables from chronic illness to job stress to weekly food expenditures. Panel sampling can also be used to inform researchers about within-person health changes due to age or help explain changes in continuous dependent variables such as spousal interaction. There have been several proposed methods of analyzing panel sample data, including MANOVA, growth curves, and structural equation modeling with lagged effects. For a more thorough look at analytical techniques for panel data, see Johnson (1995).

Event sampling methodology

Event sampling methodology (ESM) is a new form of sampling method that allows researchers to study ongoing experiences and events that vary across and within days in its naturally-occurring environment. Because of the frequent sampling of events inherent in ESM, it enables researchers to measure the typology of activity and detect the temporal and dynamic fluctuations of work experiences. Popularity of ESM as a new form of research design increased over the recent years because it addresses the shortcomings of cross-sectional research, where once unable to, researchers can now detect intra-individual variances across time. In ESM, participants are asked to record their experiences and perceptions in a paper or electronic diary.

There are three types of ESM:

Signal contingent – random beeping notifies participants to record data. The advantage of this type of ESM is minimization of recall bias.

Event contingent – records data when certain events occur

Interval contingent – records data according to the passing of a certain period of time ESM has several disadvantages. One of the disadvantages of ESM is it can sometimes be perceived as invasive and intrusive by participants. ESM also leads to possible self-selection bias. It may be that only certain types of individuals are willing to participate in this type of study creating a non-random sample. Another concern is related to participant cooperation. Participants may not be actually fill out their diaries at the specified time.
times. Furthermore, ESM may substantively change the phenomenon being studied. Reactivity or priming effects may occur, such that repeated measurement may cause changes in the participants' experiences. This method of sampling data is also highly vulnerable to common method variance.

Further, it is important to think about whether or not an appropriate dependent variable is being used in an ESM design. For example, it might be logical to use ESM in order to answer research questions which involve dependent variables with a great deal of variation throughout the day. Thus, variables such as change in mood, change in stress level, or the immediate impact of particular events may be best studied using ESM methodology. However, it is not likely that utilizing ESM will yield meaningful predictions when measuring someone performing a repetitive task throughout the day or when dependent variables are long-term in nature (coronary heart problems).

Estimation and Confidence Intervals

Point and Interval Estimates

Confidence Intervals/Margin of Error
Finite Population Correction Factor
Known vs. Unknown Population Variance
Statistical Precision
Testing $\rho = a$ (Correlation Coefficient): Fisher $z$
Testing $\rho = 0$ (Correlation Coefficient)
Testing $P = a$ (Population Proportion)

Point and Interval Estimates

Recall how the critical value(s) delineated our region of rejection. For a two-tailed test the distance to these critical values is also called the margin of error and the region between critical values is called the confidence interval. Such a confidence interval is commonly formed when we want to estimate a population parameter, rather than test a hypothesis. This process of estimating a population parameter from a sample statistic (or observed statistic) is called statistical estimation. We can either form a point estimate or an interval estimate, where the interval estimate contains a range of reasonable or tenable values with the point estimate our "best guess." When a null hypothesis is rejected, this procedure can give us more information about the variable under investigation. It can also test many hypotheses simultaneously. Although common in science, this use of statistics may be underutilized in the behavioral sciences extensively to calculate the margin of error which in turn is used to calculate confidence intervals.

Remember, if we sample enough times, we will obtain a very reasonable estimate of both the population mean and population standard deviation. This is true whether or not the population is normally distributed. However, normally distributed populations are very common. Populations which are not normal
are often "heap-shaped" or "mound-shaped". Some skewness might be involved (mean left or right of median due to a "tail") or those dreaded outliers may be present. It is good practice to check these concerns before trying to infer anything about your population from your sample.

Since 95.0% of a normally distributed population is within 1.96 (95% is within about 2) standard deviations of the mean, we can often calculate an interval around the statistic of interest which for 95% of all possible samples would contain the population parameter of interest. We will assume for the sake of this discussion that this statistic/parameter is the mean.

We can say we are 95% confident that the unknown population parameter lies within our given range. This is to say, the method we use will generate an interval containing the parameter of interest 95% of the time. For life-and-death situations, 99% or higher confidence intervals may quite appropriately be chosen. The population parameter either is or is not within the confidence interval so we must be careful to say we have 95% confidence that it is within, not that there is a 95% probability that it is. Since we expect it to 95% of the time, this can be a point of confusion.

Example:

Assume the population is the U.S. population with a mean IQ of 100 and standard deviation of 15. Assume further that we draw a sample of \( n = 5 \) with the following values: 100, 100, 100, 100, 150. The sample mean is then 110 and the sample standard deviation is easily calculated as \( \sqrt{(102+102+102+102+402)/(5-1)} = \sqrt{500} \) or approximately 22.4. The standard error of the mean is \( \sqrt{500}/\sqrt{5} = \sqrt{100} = 10 \). Our 95% confidence intervals are then formed with \( z = \pm 1.96 \). Thus based on this sample we can be 95% confident that the population mean lies between 110-19.6 and 110+19.6 or in (90.4, 129.6). Suppose, however, that you did not know the population standard deviation. Then since this is also a small sample you would use the \( t \)-statistics. The \( t \)-value of 2.776 would give you a margin of error of 27.8 and a corresponding confidence interval of (82.2, 137.8).

If you are sampling without replacement and your sample size is more than, say, 5% of the finite population \( N \), you need to adjust (reduce) the standard error of the mean when used to estimate the mean by multiplying it by the finite population correction factor as specified above. If we can assume that the population is infinite or that our sample size does not exceed 5% of the population size (or we are sampling with replacement), then there is no need to apply this correction factor.

Known vs. Unknown Population Variance

There is a pervasive joke in inferential statistics about knowing the population variance or population standard deviation. If such a value were known, then we have a big handle on how the population is distributed and would seem to have little reason to do inferential statistics on a sample. However, there are times when a test, like an IQ test, might be designed with a given variance in mind and such an assumption is meaningful.
The only practical difference is that unless our sample size is large enough (n > 30) we should use the more conservative \( t \) distribution rather than the normal distribution to obtain the critical value of our test statistic when the population variance (standard deviation) is unknown and also use \( s^2 \) (or \( s \)) as an estimate of the population variance (or standard deviation) when calculating the standard error of the mean or \( t \) value.

Statistical Precision

Statistical Precision can be thought of as how narrow our margin of error is. For increased precision a larger sample size is required. However, the precision increases slowly due to the square root of \( n \) in the denominator of the formula. Thus to cut a margin of error in half would require one to increase the sample size by a factor of four. Of course, the margin of error is also influenced by our level of significance or confidence level, but that tends to stay fixed within a field of study. A 99% confidence interval will be wider than a 95% confidence interval or less precise. The same basic situation applies for the correlation coefficient and population proportion tests described below even though different formulae determine our test statistic.

Testing rho=\( a \) (Correlation Coefficient): Fisher \( z \)

Sociologists might commonly test hypotheses regarding the correlation between two variables or construct an interval estimate of such a correlation. However, a transformation of variable is necessary since the sampling distribution is skewed when there is a correlation. The special case of testing for no correlation will be handled with a normal distribution in the next section.

The Fisher \( z \) transformation transforms the correlation coefficient \( r \) into the variable \( z \) which is approximately normal for any value of \( r \), as long as the sample size is large enough. However, the transformation goes beyond simple algebra so a conversion table is included in the Hinkle text. We don’t expect to test over this material so this is included here only for reference.

The transformation involves the logarithm function which relates a given number (\( y \)) to another number (\( x \)) which is the exponent required to raise some base (\( b \)) to, to obtain the given number (\( y = bx \)). The usual base used is that of the natural logarithm or base \( e = 2.71828... \) (It can also be described as the hyperbolic cotangent function.)

\[ z_r = \frac{1}{2} \log \left( \frac{1+r}{1-r} \right) \]

The standard error of \( z_r \) is given by \( z_{sr} = \sqrt{\frac{1}{n-3}} \).

\( n \), of course, is the sample size. We then proceed with hypothesis testing or confidence interval construction by forming the test statistic in the usual manner of \( \frac{\text{statistic}-\text{parameter}}{\text{standard error of the statistic}} \). It is usual to call the population correlation coefficient rho.

Testing rho=0 (Correlation Coefficient)
A common test in the behavioral sciences is that of whether or not a relationship exists between two variables. Zero correlation in a population is a special case where the $t$ distribution can be used after a slightly different transformation.

\[ t = r \cdot \sqrt{\frac{(n-2)}{(1-r^2)}}. \]

$n$ is our usual sample size and $n-2$ the degrees of freedom (with one lost for the variance of each variable).

Example: Consider a two-tailed test to check $H_0: \rho = 0$ at $\alpha = 0.05$ for a sample of 22 ordered pairs when $r = 0.45$.

Solution: The critical $t$ values are +/-2.086. $t = 0.45 \cdot \sqrt{\frac{(22-2)}{(1-0.45^2)}} = 2.254$. We can thus reject the null hypothesis or as commonly stated find the relationship to be statistically significant. We can construct a corresponding confidence interval, which should be done using the Fisher $z$ transformation of the previous section (since $\rho \neq 0$): $r = 0.45$ transforms to $z_r = 0.485$ and $s_zr = 0.229$. The 95% confidence interval is then $z_r +/- 1.96 \cdot 0.229 = 0.485 +/- 0.450$ or $(0.035, 0.935)$. Note that these are still $z$ scores which transform back to $(0.035, 0.733)$ as $r$ values. (The inverse transformation might easiest be done with a table of values or via the time honored guess and check method, instead of using the inverse hyperbolic cotangent.) The transformation equation above is commonly avoided by the use of tables compiled to give critical $r$ values for various sample sizes and alphas.

Testing $P = a$ (Population Proportion)

An uppercase $P$ is used for population proportion since the Greek letter pi almost always refers to the ratio of a circle’s circumference to its diameter (3.1415...).

When two possibilities exist for a particular variable in a population, the binomial distribution provides an easily identifiable standard error of the proportion in terms of $p$, the hypothetical proportion value, $q = 1 - p$, and the sample size $n$. Since the binomial tends toward the normal distribution quickly we can use the normal distribution when $np$ AND $nq$ both exceed some magic number, say 10. Some less conservative disciplines might even push that magic number down to 5, whereas more conservative disciplines push it up to 15. This magic number check helps ensure adequately sized samples when $p$ takes on values far away from $\frac{1}{2}$, i.e. near either 0 or 1. The formula for the standard error of the proportion is: $sp = \sqrt{pq/n}$.

(Take care here not to assume you can find this by dividing the standard deviation for a binomial distribution by the square root of the sample size!)

Example: Suppose 35% of first-year college students applied to some other college than where they are actually attending. Suppose further that you will be asking a simple random sample of size $n = 1000$ from the population of about $N = 1,600,000$ and desire a result within 3% of the true value. How probable is this?
Solution: We expect a mean sample proportion of $p = 0.35$ distributed normally with a standard deviation of $\sqrt{pq/n} = 0.0151$. We can calculate $P(0.32 < p < 0.38) = P(-1.989 < z < 1.989) = 0.953$ or slightly more than 95% of all samples will give such a result.

We should note that the confidence interval constructed about our test statistic using the hypothesized population parameter and the confidence interval constructed using the actual sample statistic will differ.

TESTING OF HYPOTHESIS 9

Hypothesis Testing - General Procedure for Hypothesis Testing - Errors in Hypothesis Testing - OneSample and Two Sample Tests for Means and Proportions of Large Samples (Z-Test) - One Sample and Two Sample Tests for Means of Small Samples (T-Test) - F-Test for Two Sample Standard deviations

<table>
<thead>
<tr>
<th>Name</th>
<th>Assumptions or notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>For normal population or $n &gt; 30$ and $\sigma$ known.</td>
<td>($z$ is the distance from the mean in relation to the standard deviation of the mean). For non-normal distributions it is possible to calculate a minimum proportion of a population that falls within $k$ standard deviations for any $k$.</td>
</tr>
<tr>
<td>One-sample $z$-test</td>
<td>Normal population and independent observations and $\sigma_1$</td>
</tr>
<tr>
<td>Two-sample $z$-test</td>
<td>$\sigma_1$ and $\sigma_2$ are known</td>
</tr>
<tr>
<td>Two-sample pooled $t$-test</td>
<td>(Normal populations or $n_1 + n_2 &gt; 40$) and independent</td>
</tr>
</tbody>
</table>
equal variances* observations and $\sigma_1 = \sigma_2$ and $\sigma_1$ and $\sigma_2$ unknown

Two-sample unpoled $t$-test, (Normal populations or $n_1 + n_2 > 40$) and independent

unequal variances* observations and $\sigma_1 \neq \sigma_2$ and $\sigma_1$ and $\sigma_2$ unknown

$n \cdot \hat{p}_0 > 10$ and $n \cdot (1 - \hat{p}_0) > 10$ and it is a SRS (Simple Random Sample).

One-proportion $z$-test

Two-proportion $z$-test, pooled

$n_1 \cdot \hat{p}_1 > 5$ and $n_1 \cdot (1 - \hat{p}_1) > 5$ and $n_2 \cdot \hat{p}_2 > 5$ and $n_2 \cdot (1 - \hat{p}_2) > 5$

for $d_0 = 0$

and independent observations.

Two-proportion $z$-test, unpooled for $|d_0| > 0$

and independent observations.

The sampling method for each population is simple random sampling.

The samples are independent.

Each sample includes at least 10 successes and 10 failures. (Some texts say that 5 successes and 5 failures are enough.)

$s = \text{sample standard deviation} \quad s^2 = \text{sample variance}$

$s_1 = \text{sample 1 standard deviation} \quad s_2 = \text{sample 2 standard deviation} \quad t = \text{t statistic}$

$df = \text{degrees of freedom}$

= sample mean of differences

$d_0 = \text{hypothesized population mean difference} \quad s_d = \text{standard deviation of differences}$
One of the following

One-sample chi-square test
- All expected counts are at least 5
- All expected counts are > 1 and no more that 20% of expected counts are less than 5

*Two-sample F test for

\[ F > F(\alpha / 2, n_1 - 1, n_2 - 1) \]

equality of variances

In general, the subscript 0 indicates a value taken from the null hypothesis, \( H_0 \), which should be used as much as possible in constructing its test statistic. ... Definitions of other symbols:

often stated in the following form.

\[ H_0: P_1 = P_2 \]

\[ H_a: P_1 \neq P_2 \]

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
Test method. Use the two-proportion z-test (described in the next section) to determine whether the hypothesized difference between population proportions differs significantly from the observed sample difference.

Analyze Sample Data

Using sample data, complete the following computations to find the test statistic and its associated P-Value.

Pooled sample proportion. Since the null hypothesis states that $P_1 = P_2$, we use a pooled sample proportion $(p)$ to compute the standard error of the sampling distribution.

$$p = (p_1 \cdot n_1 + p_2 \cdot n_2) / (n_1 + n_2)$$

where $p_1$ is the sample proportion from population 1, $p_2$ is the sample proportion from population 2, $n_1$ is the size of sample 1, and $n_2$ is the size of sample 2.

Standard error. Compute the standard error (SE) of the sampling distribution difference between two proportions.

$$SE = \sqrt{p \cdot (1 - p) \cdot \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}$$

where $p$ is the pooled sample proportion, $n_1$ is the size of sample 1, and $n_2$ is the size of sample 2.

Test statistic. The test statistic is a z-score $(z)$ defined by the following equation.
\[ z = \frac{(p_1 - p_2)}{SE} \]

where \( p_1 \) is the proportion from sample 1, \( p_2 \) is the proportion from sample 2, and SE is the standard error of the sampling distribution.

**P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the Normal Distribution Calculator to assess the probability associated with the z-score. (See sample problems at the end of this lesson for examples of how this is done.)

The analysis described above is a two-proportion z-test.

**Interpret Results**

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level and rejecting the null hypothesis when the P-value is less than the significance level.

**Test Your Understanding of This Lesson**

In this section, two sample problems illustrate how to conduct a hypothesis test for the difference between two proportions. The first problem involves a two-tailed test; the second problem, a one-tailed test.

**Problem 1: Two-Tailed Test**
Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

**Solution:** The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

**State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: \( P_1 = P_2 \)

Alternative hypothesis: \( P_1 \neq P_2 \)

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the proportion from population 1 is too big or if it is too small.

**Formulate an analysis plan.** For this analysis, the significance level is 0.05. The test method is a two-proportion z-test.

**Analyze sample data.** Using sample data, we calculate the pooled sample proportion (\( p \)) and the standard error (SE). Using those measures, we compute the z-score test statistic (\( z \)).

\[
p = \frac{(p_1 \times n_1 + p_2 \times n_2)}{(n_1 + n_2)} = \frac{[(0.38 \times 100) + (0.51 \times 200)]}{(100 + 200)} = \frac{140}{300} = 0.467
\]
SE = sqrt{ p * (1 - p) * \left[(1/n_1) + (1/n_2)\right] } 

SE = \sqrt{0.467 \times 0.533 \times (\frac{1}{100} + \frac{1}{200})} = \sqrt{0.003733} = 0.061

z = \frac{p_1 - p_2}{SE} = \frac{0.51 - 0.38}{0.061} = 2.13

where \( p_1 \) is the sample proportion in sample 1, where \( p_2 \) is the sample proportion in sample 2, \( n_1 \) is the size of sample 2, and \( n_2 \) is the size of sample 2.

Since we have a two-tailed test, the P-value is the probability that the z-score is less than -2.13 or greater than 2.13.

We use the Normal Distribution Calculator to find \( P(z < -2.13) = 0.017 \), and \( P(z > 2.13) = 0.017 \). Thus, the P-value = 0.017 + 0.017 = 0.034.

Interpret results. Since the P-value (0.034) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the...
sampling method was simple random sampling, the samples were independent, each population was at least 10 times larger than its sample, and each sample included at least 10 successes and 10 failures.

Problem 2: One-Tailed Test

Suppose the previous example is stated a little bit differently. Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is more effective for women than for men. To test this claim, they choose a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we conclude that the drug is more effective for women than for men? Use a 0.01 level of significance.

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: \( P_1 \geq P_2 \)

Alternative hypothesis: \( P_1 < P_2 \)

Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the proportion of women catching cold (\( p_1 \)) is sufficiently smaller than the proportion of men...
Formulate an analysis plan. For this analysis, the significance level is 0.01. The test method is a two-proportion z-test.

Analyze sample data. Using sample data, we calculate the pooled sample proportion (p) and the standard error (SE). Using those measures, we compute the z-score test statistic (z).

\[
p = \frac{(p_1 \times n_1 + p_2 \times n_2)}{(n_1 + n_2)} = \frac{(0.38 \times 100) + (0.51 \times 200)}{100 + 200} = \frac{140}{300} = 0.467
\]

\[
SE = \sqrt{p \times (1 - p) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}
\]

\[
SE = \sqrt{0.467 \times 0.533 \times (0.001 + 0.0005)} = \sqrt{0.003733} = 0.061
\]

\[
z = \frac{(p_1 - p_2)}{SE} = \frac{(0.38 - 0.51)}{0.061} = -2.13
\]

where \(p_1\) is the sample proportion in sample 1, where \(p_2\) is the sample proportion in sample 2, \(n_1\) is the size of sample 2, and \(n_2\) is the size of sample 2.

Since we have a one-tailed test, the P-value is the probability that the z-score is less than -2.13. We use the Normal Distribution Calculator to find \(P(z < -2.13) = 0.017\). Thus, the P-value = 0.017.

Interpret results. Since the P-value (0.017) is greater than the significance level (0.01), we cannot reject the null hypothesis.
Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, each population was at least 10 times larger than its sample, and each sample included at least 10 successes and 10 failures.

Hypothesis Test of the Mean

This lesson explains how to conduct a hypothesis test of a mean, when the following conditions are met:

- The sampling method is simple random sampling.
- The sample is drawn from a normal or near-normal population.

Generally, the sampling distribution will be approximately normally distributed if any of the following conditions apply.

- The population distribution is normal.
- The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
- The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
- The sample size is greater than 40, without outliers.
This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of hypotheses. Each makes a statement about how the population mean \( \mu \) is related to a specified value \( M \). (In the table, the symbol \( \neq \) means "not equal to").

<table>
<thead>
<tr>
<th>Set</th>
<th>Null Hypothesis</th>
<th>Alternative Hypothesis</th>
<th>Number of Tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu = M )</td>
<td>( \mu \neq M )</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \mu \geq M )</td>
<td>( \mu &lt; M )</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( \mu \leq M )</td>
<td>( \mu &gt; M )</td>
<td></td>
</tr>
</tbody>
</table>

The first set of hypotheses (Set 1) is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.
Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- **Test method.** Use the one-sample t-test to determine whether the hypothesized mean differs significantly from the observed sample mean.

Analyze Sample Data

Using sample data, conduct a one-sample t-test. This involves finding the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

- **Standard error.** Compute the standard error (SE) of the sampling distribution.

\[
SE = s \times \sqrt{\left(\frac{1}{n}\right) \times \left(1 - \frac{n}{N}\right) \times \left[\frac{N}{N - 1}\right]}
\]

where \(s\) is the standard deviation of the sample, \(N\) is the population size, and \(n\) is the sample size.

When the population size is much larger (at least 10 times larger) than the sample size, the standard error can be approximated by:

\[
SE = \frac{s}{\sqrt{n}}
\]

- **Degrees of freedom.** The degrees of freedom (DF) is equal to the sample size \((n)\) minus one. Thus, \(DF = n - 1\).
Test statistic. The test statistic is a t-score (t) defined by the following equation.

\[ t = \frac{(x - \mu)}{SE} \]

where \( x \) is the sample mean, \( \mu \) is the hypothesized population mean in the null hypothesis, and \( SE \) is the standard error.

P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the t-score, given the degrees of freedom computed above. (See sample problems at the end of this lesson for examples of how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding of This Lesson

In this section, two sample problems illustrate how to conduct a hypothesis test of a mean score. The first problem involves a two-tailed test; the second problem, a one-tailed test.

Sample Planning Wizard

As you probably noticed, the process of testing a hypothesis about a mean score can be complex and time-consuming. Stat Trek's Sample Planning Wizard can do the same job quickly, easily, and error-free. In
addition to conducting the hypothesis test, the Wizard creates a summary report that lists key findings and documents analytical techniques. Whenever you need to test a hypothesis, consider using the Sample Planning Wizard. The Sample Planning Wizard is a premium tool available only to registered users. 

Problem 1: Two-Tailed Test

An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for 5 hours (300 minutes) on a single gallon of regular gasoline. Suppose a simple random sample of 50 engines is tested. The engines run for an average of 295 minutes, with a standard deviation of 20 minutes. Test the null hypothesis that the mean run time is 300 minutes against the alternative hypothesis that the mean run time is not 300 minutes. Use a 0.05 level of significance. (Assume that run times for the population of engines are normally distributed.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  *Null hypothesis:* \( \mu = 300 \)

  *Alternative hypothesis:* \( \mu \neq 300 \)

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the sample mean is too big or if it is too small.
Formulate an analysis plan. For this analysis, the significance level is 0.05. The test method is a one-sample t-test.

Analyze sample data. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t-score test statistic (t).

\[ SE = \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{50}} = \frac{20}{7.07} = 2.83 \]

\[ DF = n - 1 = 50 - 1 = 49 \]

\[ t = \frac{(x - \mu)}{SE} = \frac{(295 - 300)}{2.83} = 1.77 \]

where \( s \) is the standard deviation of the sample, \( x \) is the sample mean, \( \mu \) is the hypothesized population mean, and \( n \) is the sample size.

Since we have a two-tailed test, the P-value is the probability that the t-score having 49 degrees of freedom is less than -1.77 or greater than 1.77.

We use the t Distribution Calculator to find \( P(t < -1.77) = 0.04 \), and \( P(t > 1.75) = 0.04 \). Thus, the P-value = 0.04 + 0.04 = 0.08.

Interpret results. Since the P-value (0.08) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, and the population was normally distributed.
Problem 2: One-Tailed Test

Bon Air Elementary School has 300 students. The principal of the school thinks that the average IQ of students at Bon Air is at least 110. To prove her point, she administers an IQ test to 20 randomly selected students. Among the sampled students, the average IQ is 108 with a standard deviation of 10. Based on these results, should the principal accept or reject her original hypothesis? Assume a significance level of 0.01.

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: \( \mu \geq 110 \)

  Alternative hypothesis: \( \mu < 110 \)

  Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the sample mean is too small.

- Formulate an analysis plan. For this analysis, the significance level is 0.01. The test method is a one-sample t-test.

- Analyze sample data. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t-score test statistic (t).
where $s$ is the standard deviation of the sample, $x$ is the sample mean, $\mu$ is the hypothesized population mean, and $n$ is the sample size.

Since we have a one-tailed test, the P-value is the probability that the t-score having 19 degrees of freedom is less than -0.894.

We use the t Distribution Calculator to find $P(t < -0.894) = 0.19$. Thus, the P-value is 0.19.

Interpret results. Since the P-value (0.19) is greater than the significance level (0.01), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, and the population was normally distributed.

Hypothesis Test for the Difference Between Two Means

This lesson explains how to conduct a hypothesis test for the difference between two means. The test procedure, called the two-sample t-test, is appropriate when the following conditions are met:

- The sampling method for each sample is simple random sampling.
The samples are independent.

- Each population is at least 10 times larger than its respective sample.

- Each sample is drawn from a normal or near-normal population. Generally, the sampling distribution will be approximately normal if any of the following conditions apply.

  - The population distribution is normal.

  - The sample data are symmetric, unimodal, without outliers, and the sample size is 15 or
The sample data are slightly skewed, unimodal, without outliers, and the sample size is 16 to 40.

The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The table below shows three sets of null and alternative hypotheses. Each makes a statement about the difference \( d \) between the mean of one population \( \mu_1 \) and the mean of another population \( \mu_2 \). (In the table, the symbol \( \neq \) means "not equal to ".)

<table>
<thead>
<tr>
<th>Set</th>
<th>Null hypothesis</th>
<th>Alternative hypothesis</th>
<th>Number of tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu_1 - \mu_2 = d )</td>
<td>( \mu_1 - \mu_2 \neq d )</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>( \mu_1 - \mu_2 \geq d )</td>
<td>( \mu_1 - \mu_2 &lt; d )</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>( \mu_1 - \mu_2 \leq d )</td>
<td>( \mu_1 - \mu_2 &gt; d )</td>
<td>1</td>
</tr>
</tbody>
</table>
The first set of hypotheses (Set 1) is an example of a **two-tailed test**, since an extreme value on either side of the **sampling distribution** would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are **one-tailed tests**, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population means (i.e., \( d = 0 \)), the null and alternative hypothesis are often stated in the following form.

\[
H_0: \mu_1 = \mu_2 \\
H_a: \mu_1 \neq \mu_2
\]

**Formulate an Analysis Plan**

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- **Test method.** Use the **two-sample t-test** to determine whether the difference between means found in the sample is significantly different from the hypothesized difference between means.

**Analyze Sample Data**

Using sample data, find the standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

- **Standard error.** Compute the **standard error** \((SE)\) of the sampling distribution.

\[
SE = \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}
\]
where \( s_1 \) is the standard deviation of sample 1, \( s_2 \) is the standard deviation of sample 2, \( n_1 \) is the size of sample 1, and \( n_2 \) is the size of sample 2.

**Degrees of freedom.** The degrees of freedom (DF) is:

\[
DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left\{ \left(\frac{s_1}{n_1}\right)^2/(n_1 - 1) \right\} + \left\{ \left(\frac{s_2}{n_2}\right)^2/(n_2 - 1) \right\}}
\]

If DF does not compute to an integer, round it off to the nearest whole number. Some texts suggest that the degrees of freedom can be approximated by the smaller of \( n_1 - 1 \) and \( n_2 - 1 \); but the above formula gives better results.

**Test statistic.** The test statistic is a t-score (t) defined by the following equation.

\[
t = \frac{(x_1 - x_2) - d}{SE}
\]

where \( x_1 \) is the mean of sample 1, \( x_2 \) is the mean of sample 2, \( d \) is the hypothesized difference between population means, and \( SE \) is the standard error.

**P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the t-score, having the degrees of freedom computed above. (See sample problems at the end of this lesson for examples of how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis.

Typically, this involves comparing the P-value to the significance level and rejecting the null hypothesis when the P-value is less than the significance level.
Test Your Understanding of This Lesson

In this section, two sample problems illustrate how to conduct a hypothesis test of a difference between mean scores. The first problem involves a two-tailed test; the second problem, a one-tailed test.

Problem 1: Two-Tailed Test

Within a school district, students were randomly assigned to one of two Math teachers - Mrs. Smith and Mrs. Jones. After the assignment, Mrs. Smith had 30 students, and Mrs. Jones had 25 students.

At the end of the year, each class took the same standardized test. Mrs. Smith's students had an average test score of 78, with a standard deviation of 10; and Mrs. Jones' students had an average test score of 85, with a standard deviation of 15.

Test the hypothesis that Mrs. Smith and Mrs. Jones are equally effective teachers. Use a 0.10 level of significance. (Assume that student performance is approximately normal.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: \( \mu_1 - \mu_2 = 0 \)

  Alternative hypothesis: \( \mu_1 - \mu_2 \neq 0 \)

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.10. Using sample data, we will conduct a two-sample t-test of the null hypothesis.
Analyze sample data. Using sample data, we compute the standard error \( (SE) \), degrees of freedom \( (DF) \), and the t-score test statistic \( (t) \).

\[
SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

\[
SE = \sqrt{\left(\frac{10^2}{30}\right) + \left(\frac{15^2}{25}\right)} = \sqrt{3.33 + 9} = \sqrt{12.33} = 3.51
\]

\[
DF = \frac{s_1^2/n_1 + s_2^2/n_2}{\left[\left(s_1^2/n_1\right)/(n_1 - 1)\right] + \left[s_2^2/n_2\right]/(n_2 - 1)}
\]

\[
DF = \frac{10^2/30 + 15^2/25}{\left[\left(10^2/30\right)/29\right] + \left[15^2/25\right]/24} = 40.47
\]

\[
t = \frac{(x_1 - x_2) - d}{SE} = \frac{(78 - 85) - 0}{3.51} = -1.99
\]

where \( s_1 \) is the standard deviation of sample 1, \( s_2 \) is the standard deviation of sample 2, \( n_1 \) is the size of sample 1, \( n_2 \) is the size of sample 2, \( x_1 \) is the mean of sample 1, \( x_2 \) is the mean of sample 2, \( d \) is the hypothesized difference between the population means, and \( SE \) is the standard error.

Since we have a two-tailed test, the P-value is the probability that a t-score having 40 degrees of freedom is more extreme than -1.99; that is, less than -1.99 or greater than 1.99.

We use the t Distribution Calculator to find \( P(t < -1.99) = 0.027 \), and \( P(t > 1.99) = 0.027 \). Thus, the P-value = \( 0.027 + 0.027 = 0.054 \).

Interpret results. Since the P-value (0.054) is less than the significance level (0.10), we cannot accept the null hypothesis.
Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the samples were drawn from a normal population.

Problem 2: One-Tailed Test

The Acme Company has developed a new battery. The engineer in charge claims that the new battery will operate continuously for at least 7 minutes longer than the old battery.

To test the claim, the company selects a simple random sample of 100 new batteries and 100 old batteries. The old batteries run continuously for 190 minutes with a standard deviation of 20 minutes; the new batteries, 200 minutes with a standard deviation of 40 minutes.

Test the engineer's claim that the new batteries run at least 7 minutes longer than the old. Use a 0.05 level of significance. (Assume that there are no outliers in either sample.)

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: \( \mu_1 - \mu_2 \geq 7 \)

  Alternative hypothesis: \( \mu_1 - \mu_2 < 7 \)

  Note that these hypotheses constitute a one-tailed test. The null hypothesis will be rejected if the mean difference between sample means is too small.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a two-sample t-test of the null hypothesis.
Analyze sample data. Using sample data, we compute the standard error (SE), degrees of freedom (DF), and the t-score test statistic (t).

\[
SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}
\]

\[
SE = \sqrt{\left(\frac{40^2}{100} + \frac{20^2}{100}\right)} = \sqrt{16 + 4} = 4.472
\]

\[
DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{(s_1^2)}{n_1 - 1}\right] + \left[\frac{(s_2^2)}{n_2 - 1}\right]}
\]

\[
DF = \frac{\left(\frac{40^2}{100} + \frac{20^2}{100}\right)^2}{\left[\frac{(40^2)}{99}\right] + \left[\frac{(20^2)}{99}\right]} = 400 / (2.586 + 0.162) = 145.56
\]

\[
t = \frac{(x_1 - x_2) - d}{SE} = \frac{(200 - 190) - 7}{4.472} = \frac{3}{4.472} = 0.67
\]

where \(s_1\) is the standard deviation of sample 1, \(s_2\) is the standard deviation of sample 2, \(n_1\) is the size of sample 1, \(n_2\) is the size of sample 2, \(x_1\) is the mean of sample 1, \(x_2\) is the mean of sample 2, \(d\) is the hypothesized difference between population means, and SE is the standard error.

For this one-tailed test, the P-value is the probability of obtaining a t-score test statistic that is more extreme than 0.67 (i.e., greater than 0.67), assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.

To find the probability that the t-score test statistic is greater than 0.67, we use the t Distribution Calculator. The calculator tells us that \(P(t \leq 0.67) = 0.75\). Thus, the P-value is:

\[
P(t \geq 0.67) = 1 - P(t \leq 0.67) = 1 - 0.75 = 0.25
\]
Interpret results. Since the P-value (0.25) is greater than the significance level (0.05), we cannot reject the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples were independent, the sample size was much smaller than the population size, and the sample size was large without outliers.

Hypothesis Test for Difference Between Matched Pairs

This lesson explains how to conduct a hypothesis test for the difference between paired means. The test procedure, called the matched-pairs t-test, is appropriate when the following conditions are met:

- The sampling method for each sample is simple random sampling.
- The test is conducted on paired data. (As a result, the data sets are not independent.)
- Each sample is drawn from a normal or near-normal population. Generally, the sampling distribution will be approximately normal if any of the following conditions apply.
  - The population distribution is normal.
  - The sample data are symmetric, unimodal, without outliers, and the sample size is 15 or less.
  - The sample data are slightly skewed, unimodal, without outliers, and the sample size is 16 to 40.
  - The sample size is greater than 40, without outliers.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze
sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

The hypotheses concern a new variable \( d \), which is based on the difference between paired values from two data sets.

\[
d = x_1 - x_2
\]

where \( x_1 \) is the value of variable \( x \) in the first data set, and \( x_2 \) is the value of the variable from the second data set that is paired with \( x_1 \).

The table below shows three sets of null and alternative hypotheses. Each makes a statement about how the true difference in population values \( \mu_d \) is related to some hypothesized value \( D \). (In the table, the symbol \( \neq \) means "not equal to ".)

<table>
<thead>
<tr>
<th>Set</th>
<th>Null hypothesis</th>
<th>Alternative hypothesis</th>
<th>Number of tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \mu_d = D )</td>
<td>( \mu_d \neq D )</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \mu_d \geq D )</td>
<td>( \mu_d &lt; D )</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( \mu_d \leq D )</td>
<td>( \mu_d &gt; D )</td>
<td></td>
</tr>
</tbody>
</table>
The first set of hypotheses (Set 1) is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- Test method. Use the matched-pairs t-test to determine whether the difference between sample means for paired data is significantly different from the hypothesized difference between population means.

Analyze Sample Data

Using sample data, find the standard deviation, standard error, degrees of freedom, test statistic, and the P-value associated with the test statistic.

- Standard deviation. Compute the standard deviation ($s_d$) of the differences computed from $n$ matched pairs.

$$s_d = \sqrt{\left( \sum (d_i - d)^2 \right) / (n - 1)}$$

where $d_i$ is the difference for pair $i$, $d$ is the sample mean of the differences, and $n$ is the number of paired values.
Standard error. Compute the standard error (SE) of the sampling distribution of d.

\[ SE = s_d \times \sqrt{ \left( \frac{1}{n} \right) \times \left( 1 - \frac{n}{N} \right) \times \left[ \frac{N}{(N-1)} \right] } \]

where \( s_d \) is the standard deviation of the sample difference, \( N \) is the population size, and \( n \) is the sample size. When the population size is much larger (at least 10 times larger) than the sample size, the standard error can be approximated by:

\[ SE = \frac{s_d}{\sqrt{n}} \]

Degrees of freedom. The degrees of freedom (DF) is: \( DF = n - 1 \).

Test statistic. The test statistic is a t-score (t) defined by the following equation.

\[ t = \frac{(x_1 - x_2) - D}{SE} = \frac{(d - D)}{SE} \]

where \( x_1 \) is the mean of sample 1, \( x_2 \) is the mean of sample 2, \( d \) is the mean difference between paired values in the sample, \( D \) is the hypothesized difference between population means, and \( SE \) is the standard error.

P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t-score, use the t Distribution Calculator to assess the probability associated with the t-score, having the degrees of freedom computed above. (See the sample problem at the end of this lesson for guidance on how this is done.)

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.
Test Your Understanding of This Lesson

Problem

Forty-four sixth graders were randomly selected from a school district. Then, they were divided into 22 matched pairs, each pair having equal IQ's. One member of each pair was randomly selected to receive special training. Then, all of the students were given an IQ test. Test results are summarized below.

<table>
<thead>
<tr>
<th>Pair</th>
<th>No Traini</th>
<th>Traini</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95</td>
<td>90</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>89</td>
<td>85</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>76</td>
<td>73</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>92</td>
<td>90</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>91</td>
<td>90</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>67</td>
<td>68</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>85</td>
<td>83</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>87</td>
<td>83</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>85</td>
<td>82</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>68</td>
<td>65</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>84</td>
<td>83</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>68</td>
<td>65</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>79</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>84</td>
<td>83</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>83</td>
<td>83</td>
<td>0</td>
</tr>
</tbody>
</table>

Department of Management Studies
Do these results provide evidence that the special training helped or hurt student performance? Use an 0.05 level of significance. Assume that the mean differences are approximately normally distributed.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

  Null hypothesis: \( \mu_d = 0 \)

  Alternative hypothesis: \( \mu_d \neq 0 \)

  Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the difference between sample means is too big or if it is too small.
Formulate an analysis plan. For this analysis, the significance level is 0.05. Using sample data, we will conduct a matched-pairs t-test of the null hypothesis.

Analyze sample data. Using sample data, we compute the standard deviation of the differences (s), the standard error (SE) of the mean difference, the degrees of freedom (DF), and the t-score test statistic (t).

\[ s = \sqrt{\frac{\sum (d_i - d)^2}{n - 1}} = \sqrt{\frac{270}{22-1}} = \sqrt{12.857} = 3.586 \]

\[ SE = \frac{s}{\sqrt{n}} = \frac{3.586}{\sqrt{22}} = 0.765 \]

\[ DF = n - 1 = 22 - 1 = 21 \]

\[ t = \frac{(x_1 - x_2) - D}{SE} = \frac{(d - D)}{SE} = \frac{(1 - 0)}{0.765} = 1.307 \]

where \(d_i\) is the observed difference for pair \(i\), \(d\) is mean difference between sample pairs, \(D\) is the hypothesized mean difference between population pairs, and \(n\) is the number of pairs.

Since we have a two-tailed test, the P-value is the probability that a t-score having 21 degrees of freedom is more extreme than 1.307; that is, less than -1.307 or greater than 1.307.

We use the t Distribution Calculator to find \(P(t < -1.307) = 0.103\), and \(P(t > 1.307) = 0.103\). Thus, the P-value = 0.103 + 0.103 = 0.206.

Interpret results. Since the P-value (0.206) is greater than the significance level (0.05), we cannot reject the null hypothesis.
Specifically, the approach is appropriate because the sampling method was simple random sampling, the samples consisted of paired data, and the mean differences were normally distributed. In addition, we used the approximation formula to compute the standard error, since the sample size was small relative to the population size.

One-Way Analysis of Variance
(A worked Problem: Procedure explained in more in Chapter 13)

One-Way Analysis of Variance (ANOVA) Example Problem

Introduction

Analysis of Variance (ANOVA) is a hypothesis-testing technique used to test the equality of two or more population (or treatment) means by examining the variances of samples that are taken. ANOVA allows one to determine whether the differences between the samples are simply due to random error (sampling errors) or whether there are systematic treatment effects that causes the mean in one group to differ from the mean in another. Most of the time ANOVA is used to compare the equality of three or more means, however when the means from two samples are compared using ANOVA it is equivalent to using a t-test to compare the means of independent samples. ANOVA is based on comparing the variance (or variation) between the data samples to variation within each particular sample. If the between variation is much larger than the within variation, the means of different samples will not be equal. If the between and within variations are approximately the same size, then there will be no significant difference between sample means.

Assumptions of ANOVA:

(i) All populations involved follow a normal distribution.
(ii) All populations have the same variance (or standard deviation).
(iii) The samples are randomly selected and independent of one another.

Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.

As a psychologist who works with people who have Down's syndrome, you design a study intended to determine which rewards are most effective for use in training your patients. You select four different, independent, groups of six patients and record the number of days it takes to teach them a particular task, with each group receiving one of four types of rewards: Reward 1, Reward 2, Reward 3, and Reward 4. the number of days are given in the following table.
Use the data above to conduct a one-way analysis of variance.

I. State your hypotheses

Null hypothesis: The type of reward does not make a difference in the number of days required for Down's Syndrome patients to learn a task.

Research hypothesis: The type of reward makes a difference in the number of days required for Down's syndrome patients to learn a task

II. After stating the hypotheses, always begin an analysis of variance problem by computing all required sums.

<table>
<thead>
<tr>
<th>Reward 1</th>
<th>Reward 2</th>
<th>Reward 3</th>
<th>Reward 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

III. Compute SS$_{total}$

$$SS_{total} = \sum \sum X^2 - \frac{(\sum \sum X)^2}{N_{total}} = 2478 - \frac{218^2}{24} = 2478 - \frac{47524}{24}$$

$$SS_{total} = 2478 - 1980.167 = 497.833$$

IV. Compute SS$_{bg}$
V. Compute SS_{wg}

\[ SS_{wg} = SS_{total} - SS_{bg} = 497.833 - 413.5 = 84.333 \]

VI. Compute df_{total}

\[ df_{total} = N_{total} - 1 = 24 - 1 = 23 \]

VII. Compute df_{bg}

\[ df_{bg} = k - 1 = 4 - 1 = 3 \]

VIII. Compute df_{wg}

\[ df_{wg} = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) = (6 - 1) + (6 - 1) + (6 - 1) + (6 - 1) \]
\[ df_{wg} = 5 + 5 + 5 + 5 = 20 \]

IX. Compute MS_{bg}

\[ MS_{bg} = \frac{SS_{bg}}{df_{bg}} = \frac{413.5}{3} = 137.833 \]

X. Compute MS_{wg}

\[ MS_{wg} = \frac{SS_{wg}}{df_{wg}} = \frac{84.333}{20} = 4.217 \]

XI. Compute F

\[ F = \frac{MS_{bg}}{MS_{wg}} = \frac{137.833}{4.217} = 32.685 \]

XII. Find the critical value of the F ratio in Table F and determine the significance of F

A. df = (3, 20)
B. Critical Value at alpha = .05 is 3.10
C. Because $F >$ Critical Value we can reject the Null Hypothesis and accept the Research Hypothesis.

D. The probability of an $F$ ratio this large happening just by chance is less than .05 ($p < .05$).

XIII. Create the Source Table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between</td>
<td>413.500</td>
<td>3</td>
<td>137.833</td>
<td>32.685</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Within</td>
<td>84.333</td>
<td>20</td>
<td>4.217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>497.833</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

XIV. Because $F$ is greater than the critical value we must compute the HSD

A. $MS_{wg} = 4.217$

B. $n = 6$ (Hint: The number of scores in each group)

C. $q = 3.96$ (Hint: To enter table $Q$ use $k = 4 \ldots k$ is the number of groups...and $df_{wg} = 20$)

D. Compute HSD

Two-Way Analysis of Variance

(A worked Problem: Procedure explained in more detail in Chapter 14)

An understanding of the one-way ANOVA is crucial to understanding the two-way ANOVA, so be sure that the concepts involved in the one-way ANOVA are clear. Important background information and review of concepts in ANOVA can be found in Ray so be sure to read that chapter carefully. The sort of experiment that produces data for analysis by a two-factor ANOVA is one in which there are two factors (independent variables). In Ray”s example (p. 182 ff.), an experimenter is interested in assessing the impact of housing (the first factor) and feeding schedule (the second factor) on errors made in running a maze (the dependent variable). In this experiment, the housing factor can take on two levels (enriched or standard) and the feeding schedule can take on two levels (ad lib or once a day). Thus, this experiment is a $2 \times 2$ independent groups design, which means that there are 4 unique conditions to the experiment. Of the 40 mice in the experiment, 20 are randomly assigned to the enriched housing and 20 are assigned to the standard housing. Of the 20 mice assigned to the enriched housing, 10 are fed ad lib and 10 are fed once a day. Likewise, of the 20 mice in the standard housing, 10 are fed ad lib and 10 are fed once a day.

As a budding psychologist, you wonder whether you can teach old dogs new tricks. So you go to the pound and adopt 15 old dogs and 15 puppies. Then you attempt to teach each of the 30 dogs one of the standard dog tricks, "sit", "stay", and "roll over." Teaching only one trick to each dog, you keep a record of how many days it takes before they learn the tricks. The results of your experiment are listed in the table below. Use that data to conduct a two-way analysis of variance to determine if old dogs can learn new tricks.
Because there are 2 rows and 3 columns this makes this problem a 2 by 3 two-way analysis of variance. To begin any two-way analysis of variance we must first compute all the means and all the sums for the cells, rows, and columns. To make the table easier to read and use I have replaced the data in each cell with the appropriate sums and means. (Refer to the table at the top of the problem if you want to see the actual data rather than the sums.)
I. We always begin by stating our Null and Research Hypotheses for all three F ratios

A. Rows
- \( H_0 \): There is no difference between the time it takes an old dog to learn a trick and a puppy to learn a trick.
- \( H_1 \): There is a difference between the time it takes an old dog to learn a trick and a puppy to learn a trick.

B. Columns
- \( H_0 \): There will be no difference between the time it takes to learn the different types of tricks.
- \( H_1 \): There will be a difference between the time it takes to learn the different types of tricks.

C. Interaction
- \( H_0 \): There is no interaction
- \( H_1 \): There is an interaction

II. The Source Table.....To complete the two-way analysis of variance we will fill out the following source table:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

III. Compute Sums of Squares

A. Sums of Squares Total (\( SS_{\text{total}} \))

\[
SS_{\text{total}} = \sum \sum X^2 - \frac{(\sum X)^2}{N_{\text{total}}} = 2065 - \frac{(211)^2}{30} = 2065 - \frac{44521}{30} = 2065 - 1484.033 = 580.967
\]
B. Sums of Squares Rows (SS_r)

\[
SS_{\text{rows}} = \left[ \frac{\left( \sum_{i=1}^{n_1} X_{i1} \right)^2}{n_1} + \frac{\left( \sum_{i=2}^{n_2} X_{i2} \right)^2}{n_2} \right] - \frac{\left( \sum \sum X^2 \right)}{N_{\text{total}}} = -1484.033
\]

\[
SS_{\text{rows}} = \frac{5625 + 18496}{15} - 1484.033 = 24121 - 1608.067 - 1484.033
\]

\[
SS_{\text{rows}} = 124.034
\]

C. Sums of Squares Columns (SS_c)

\[
SS_{\text{cols}} = \left[ \frac{\left( \sum_{i=1}^{n_1} X_{i1} \right)^2}{n_1} + \frac{\left( \sum_{i=2}^{n_2} X_{i2} \right)^2}{n_2} + \frac{\left( \sum_{i_3} X_{i3} \right)^2}{n_{i_3}} \right] - \frac{\left( \sum \sum X^2 \right)}{N_{\text{total}}} = -1484.033
\]

\[
SS_{\text{cols}} = \frac{25^2 + 76^2 + 110^2}{10} - 1484.033 = 625 + 5776 + 12100 - 1484.033
\]

\[
SS_{\text{cols}} = \frac{18501}{10} - 1484.033 = 1850.1 - 1484.033 = 366.067
\]

D. Sums of Squares Within Groups (SS_{wg})

\[
SS_{\text{wg}} = \sum \sum X^2 - \sum \left[ \frac{\left( \sum_{i=1}^{n_{cell}} X_{i} \right)^2}{n_{cell}} \right] = 2065 - \frac{9^2 + 26^2 + 40^2 + 16^2 + 50^2 + 70^2}{5}
\]

\[
SS_{\text{wg}} = 2065 - \frac{81 + 676 + 1600 + 256 + 2500 + 4900}{5} = 2065 - \frac{10013}{5} = 2065 - 2002.6
\]

\[
SS_{\text{wg}} = 62.4
\]

E. Sums of Squares Interaction (SS_{rxc})

\[
SS_{\text{rxc}} = SS_{\text{total}} - SS_{\text{rows}} - SS_{\text{cols}} - SS_{\text{wg}} = 580.967 - [124.034 + 366.067 + 62.4]
\]

\[
SS_{\text{rxc}} = 580.967 - 552.501 = 28.466
\]

F. Copy the Sums of Squares to the source table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>124.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columns</td>
<td>366.067</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>28.466</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>62.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IV. Compute degrees of freedom

A. Degrees of Freedom Total (df\text{total})
\[ df_{\text{total}} = N_{\text{total}} - 1 = 30 - 1 = 29 \]

B. Degrees of Freedom Rows
\[ df_{r} = n_{r} - 1 = 2 - 1 = 1 \]

C. Degrees of Freedom Columns
\[ df_{c} = n_{c} - 1 = 3 - 1 = 2 \]

D. Degrees of Freedom Interaction
\[ df_{rxc} = df_{r} \cdot df_{c} = 1 \cdot 2 = 2 \]

E. Degrees of Freedom Within (df\text{wg})
\[ df_{\text{wg}} = N_{\text{total}} - \text{Number of Cells} = 30 - 6 = 24 \]

F. Copy Sums of Squares to the Source Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>124.034</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columns</td>
<td>366.067</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>28.466</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>62.4</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>580.967</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. Compute Mean Squares

A. Mean Square Rows (MS\text{r})
\[ MS_{r} = \frac{SS_{r}}{df_{r}} = \frac{124.034}{1} = 124.034 \]

B. Mean Square Columns (MS\text{c})
\[ MS_{c} = \frac{SS_{c}}{df_{c}} = \frac{366.067}{2} = 183.034 \]

C. Mean Square Interaction (MS\text{rxc})
D. Mean Square Within (MS_{wg})

\[ MS_{wg} = \frac{SS_{wg}}{d_{wg}} = \frac{62.400}{24} = 2.6 \]

E. Copy Mean Squares to the Source Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>124.034</td>
<td>1</td>
<td>124.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columns</td>
<td>366.067</td>
<td>2</td>
<td>183.034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction</td>
<td>28.466</td>
<td>2</td>
<td>14.233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>62.4</td>
<td>24</td>
<td>2.600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>580.967</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

VI. Compute the F ratios

A. F ratio for Rows (F_r)

\[ F_{rows} = \frac{MS_{rows}}{MS_{wg}} = \frac{124.034}{2.6} = 47.705 \]

B. F ratio for Columns (F_c)

\[ F_{cols} = \frac{MS_{cols}}{MS_{wg}} = \frac{183.034}{2.6} = 70.398 \]

C. F ratio for Interaction (F_{rxc})

\[ F_{rxc} = \frac{MS_{rxc}}{MS_{wg}} = \frac{14.233}{2.6} = 5.474 \]

D. Copy F ratios to the Source Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
</table>
VII. Conclusions and Significance of the F ratios

A. \( F_r \)

1. Critical value with \((1,24)\) degrees of freedom = 4.26
2. \( F_r = 47.705 \) this is greater than the critical value. Therefore we reject \( H_0 \) for Rows. The probability of this F ratio happening just by chance is <.05.
3. The puppies learned significantly faster than the old dogs.

B. \( F_c \)

1. Critical value with \((2,24)\) degrees of freedom = 3.40.
2. \( F_c = 70.398 \) this is greater than the critical value. Therefore we reject \( H_0 \) for Columns. The probability of this F ratio happening just by chance is <.05.
3. Because there are three different columns we must now compare the means from each column with each of the other column means using the HSD.
   a) Find the value of \( q \) in table Q with \( k \), the number of groups being compared, equal to 3 and the degrees of freedom within groups equal to 24. \( q = 3.53 \)
   b) Compute the HSD

\[
HSD = q \cdot \sqrt{ \frac{MS_{W}}{n} } = 3.53 \cdot \sqrt{ \frac{2.6}{10} } = 3.53 \cdot \sqrt{0.26} = 3.53 \cdot 0.510 = 1.800
\]

c) Compare the column means

\[
\begin{align*}
|\bar{X}_{a} - \bar{X}_{c1}| &= |2.5 - 7.6| = 5.1 \\
|\bar{X}_{a} - \bar{X}_{c2}| &= |2.5 - 11| = 8.5 \\
|\bar{X}_{a} - \bar{X}_{c3}| &= |7.6 - 11| = 3.4
\end{align*}
\]

4. Because all the comparisons are greater than the HSD all the different types of tricks are significantly different from one another.

C. \( F_{rc} \)

1. Critical value with \((2,24)\) degrees of freedom = 3.40.
2. \( F_{yx} = 5.474 \) this is greater than the critical value. Therefore we reject \( H_0 \) for Interaction. The probability of this \( F \) ratio happening just by chance is <.05.

3. Because the interaction is significant we must now graph the cell means. Remember that when creating this graph use the dependent variable as the Y axis label and either rows or columns, which ever has the most groups, as the X axis.

![Graph showing number of days to learn different tricks for puppies and old dogs.](image)

You can see from the graph that the puppies learn at a much faster rate overall than the old dogs. And you can also see that the sit, shake, and roll over tricks are progressively more difficult for the old dogs, but there seems to be very little difference between shake and roll over for the younger dogs.

VII. The final completed Source Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sums of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>124.034</td>
<td>1</td>
<td>124.034</td>
<td>47.705</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Columns</td>
<td>366.067</td>
<td>2</td>
<td>183.034</td>
<td>70.398</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Interaction</td>
<td>28.466</td>
<td>2</td>
<td>14.233</td>
<td>5.474</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>Within</td>
<td>62.4</td>
<td>24</td>
<td>2.600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>580.967</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NON-PARAMETRIC METHODS

The Chi-Square Test - Statistic - Applications of Chi-Square Tests - Test of Independence of Attributes - Goodness of Fit - Yates Correction for Continuity Sign Test for Paired Data, Rank Sum Test: Mann-Whitney U Test and Kruskal Wallis Test, One Sample Run Test, Rank Correlation
This lesson explains how to conduct a chi-square goodness of fit test. The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company. The sample problem at the end of the lesson considers this example.

The test procedure described in this lesson is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The population is at least 10 times as large as the sample.
- The variable under study is categorical.
- The expected value for each level of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

For a chi-square goodness of fit test, the hypotheses take the following form.
H₀: The data are consistent with a specified distribution. Hₐ: The data are *not* consistent with a specified distribution. Typically, the null hypothesis specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis is that *at least* one of the specified proportions is not true.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- **Test method.** Use the chi-square goodness of fit test to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square goodness of fit test is described in the next section, and demonstrated in the sample problem at the end of this lesson.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to the number of levels (k) of the categorical variable minus 1: $DF = k - 1$.

- **Expected frequency counts.** The expected frequency counts at each level of the categorical variable are equal to the sample size times the hypothesized proportion from the null hypothesis:
  \[ E_i = np_i \]

  where $E_i$ is the expected frequency count for the $i$th level of the categorical variable, $n$ is the total sample size, and $p_i$ is the hypothesized proportion of observations in level $i$. 
Test statistic. The test statistic is a chi-square random variable \( \chi^2 \) defined by the following equation.

\[
\chi^2 = \sum (O_i - E_i)^2 / E_i
\]

where \( O_i \) is the observed frequency count for the \( i \)th level of the categorical variable, and \( E_i \) is the expected frequency count for the \( i \)th level of the categorical variable.

P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding of This Lesson

Problem

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars. The cards are sold in packages of 100.

Suppose a randomly-selected package of cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.
The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

- **Null hypothesis:** The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
- **Alternative hypothesis:** At least one of the proportions in the null hypothesis is false.

Formulate an analysis plan. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.

Analyze sample data. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

\[ DF = k - 1 = 3 - 1 = 2 \]

\[ (E_i) = n \cdot p_i \]

\[ (E_1) = 100 \cdot 0.30 = 30 \quad (E_2) = 100 \cdot 0.60 = 60 \quad (E_3) = 100 \cdot 0.10 = 10 \]

\[ X^2 = \sum \left( \frac{(O_i - E_i)^2}{E_i} \right) \]

\[ X^2 = \left[ \frac{(50 - 30)^2}{30} \right] + \left[ \frac{(45 - 60)^2}{60} \right] + \left[ \frac{(5 - 10)^2}{10} \right] \]

\[ X^2 = \frac{400}{30} + \frac{225}{60} + \frac{25}{10} = 13.33 + 3.75 + 2.50 = 19.58 \]
where $DF$ is the degrees of freedom, $k$ is the number of levels of the categorical variable, $n$ is the number of observations in the sample, $E_i$ is the expected frequency count for level $i$, $O_i$ is the observed frequency count for level $i$, and $X^2$ is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.

We use the Chi-Square Distribution Calculator to find

$$P(X^2 > 19.58) = 0.0001.$$  

Interpret results. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was more than 10 times larger than the sample, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.

This lesson explains how to conduct a chi-square goodness of fit test. The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company. The sample problem at the end of the lesson considers this example.

The test procedure described in this lesson is appropriate when the following conditions are met:
The sampling method is simple random sampling.

- The population is at least 10 times as large as the sample.
- The variable under study is categorical.
- The expected value for each level of the variable is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa.

For a chi-square goodness of fit test, the hypotheses take the following form.

\( H_0 \): The data are consistent with a specified distribution. \( H_A \): The data are not consistent with a specified distribution.

Typically, the null hypothesis specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis is that at least one of the specified proportions is not true.

Formulate an Analysis Plan
The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- **Significance level.** Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- **Test method.** Use the chi-square goodness of fit test to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis. The chi-square goodness of fit test is described in the next section, and demonstrated in the sample problem at the end of this lesson.

### Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequency counts, test statistic, and the P-value associated with the test statistic.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to the number of levels (k) of the categorical variable minus 1: $DF = k - 1$.

- **Expected frequency counts.** The expected frequency counts at each level of the categorical variable are equal to the sample size times the hypothesized proportion from the null hypothesis

$$E_i = np_i$$

where $E_i$ is the expected frequency count for the $i$th level of the categorical variable, $n$ is the total sample size, and $p_i$ is the hypothesized proportion of observations in level $i$.

- **Test statistic.** The test statistic is a chi-square random variable ($\chi^2$) defined by the following equation.

$$\chi^2 = \sum (O_i - E_i)^2 / E_i$$
where $O_i$ is the observed frequency count for the $i$th level of the categorical variable, and $E_i$ is the expected frequency count for the $i$th level of the categorical variable.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

**Interpret Results**

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

**Test Your Understanding of This Lesson**

**Problem**

Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars. The cards are sold in packages of 100.

Suppose a randomly-selected package of cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

**Solution**

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
Null hypothesis: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.

Alternative hypothesis: At least one of the proportions in the null hypothesis is false.

Formulate an analysis plan. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.

Analyze sample data. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

\[ \text{DF} = k - 1 = 3 - 1 = 2 \]

\[ (E_i) = n \times p_i \]

\[ (E_1) = 100 \times 0.30 = 30 \quad (E_2) = 100 \times 0.60 = 60 \quad (E_3) = 100 \times 0.10 = 10 \]

\[ X^2 = \sum \left( \frac{(O_i - E_i)^2}{E_i} \right) \]

\[ X^2 = \left( \frac{(50 - 30)^2}{30} \right) + \left( \frac{(45 - 60)^2}{60} \right) + \left( \frac{(5 - 10)^2}{10} \right) \]

\[ X^2 = \left( \frac{400}{30} \right) + \left( \frac{225}{60} \right) + \left( \frac{25}{10} \right) = 13.33 + 3.75 + 2.50 = 19.58 \]

where DF is the degrees of freedom, \( k \) is the number of levels of the categorical variable, \( n \) is the number of observations in the sample, \( E_i \) is the expected frequency count for level \( i \), \( O_i \) is the observed frequency count for level \( i \), and \( X^2 \) is the chi-square test statistic.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 19.58.
We use the Chi-Square Distribution Calculator to find

\[ P(X^2 > 19.58) = 0.0001. \]

Interpret results. Since the P-value (0.0001) is less than the significance level (0.05), we cannot accept the null hypothesis.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, the population was more than 10 times larger than the sample,

the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5.

Chi-Square Test for Independence

This lesson explains how to conduct a chi-square test for independence. The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The sample problem at the end of the lesson considers this example.

The test procedure described in this lesson is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- Each population is at least 10 times as large as its respective sample.
The variables under study are each categorical.

If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Suppose that Variable A has \( r \) levels, and Variable B has \( c \) levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

\[ H_0: \text{Variable A and Variable B are independent}. \]
\[ H_A: \text{Variable A and Variable B are not independent}. \]

The alternative hypothesis is that knowing the level of Variable A can help you predict the level of Variable B.

Note: Support for the alternative hypothesis suggests that the variables are related; but the relationship is not necessarily causal, in the sense that one variable "causes" the other.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

- Test method. Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.
Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- Degrees of freedom. The degrees of freedom (DF) is equal to:

\[
DF = (r - 1) \times (c - 1)
\]

where \( r \) is the number of levels for one categorical variable, and \( c \) is the number of levels for the other categorical variable.

- Expected frequencies. The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute \( r \times c \) expected frequencies, according to the following formula.

\[
E_{r,c} = \frac{n_r \times n_c}{n}
\]

where \( E_{r,c} \) is the expected frequency count for level \( r \) of Variable A and level \( c \) of Variable B, \( n_r \) is the total number of sample observations at level \( r \) of Variable A, \( n_c \) is the total number of sample observations at level \( c \) of Variable B, and \( n \) is the total sample size.

- Test statistic. The test statistic is a chi-square random variable (\( \chi^2 \)) defined by the following equation.

\[
\chi^2 = \sum (O_{r,c} - E_{r,c})^2 / E_{r,c}
\]

where \( O_{r,c} \) is the observed frequency count at level \( r \) of Variable A and level \( c \) of Variable B, and \( E_{r,c} \) is the expected frequency count at level \( r \) of Variable A and level \( c \) of Variable B.
P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding of This Lesson

Problem

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

<table>
<thead>
<tr>
<th>Voting Preferences</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Republican</td>
</tr>
<tr>
<td>Male</td>
<td>200</td>
</tr>
<tr>
<td>Female</td>
<td>250</td>
</tr>
<tr>
<td>Column total</td>
<td>450</td>
</tr>
</tbody>
</table>

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.
Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the **null hypothesis** and an alternative hypothesis.

  \[ H_0: \text{Gender and voting preferences are independent.} \quad H_a: \text{Gender and voting preferences are not independent.} \]

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a **chi-square test for independence**.

- **Analyze sample data.** Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the **degrees of freedom**, we determine the P-value.

  \[ \text{DF} = (r - 1) \times (c - 1) = (2 - 1) \times (3 - 1) = 2 \]

  \[ E_{r,c} = \frac{n_r \times n_c}{n} \]

  \[ E_{1,1} = \frac{400 \times 450}{1000} = 180000/1000 = 180 \]

  \[ E_{1,2} = \frac{400 \times 450}{1000} = 180000/1000 = 180 \]

  \[ E_{1,3} = \frac{400 \times 100}{1000} = 40000/1000 = 40 \]

  \[ E_{2,1} = \frac{600 \times 450}{1000} = 270000/1000 = 270 \]

  \[ E_{2,2} = \frac{600 \times 450}{1000} = 270000/1000 = 270 \]

  \[ E_{2,3} = \frac{600 \times 100}{1000} = 60000/1000 = 60 \]

  \[ \chi^2 = \sum \left( \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \right) \]

  \[ \chi^2 = \frac{(200 - 180)^2}{180} + \frac{(150 - 180)^2}{180} + \frac{(50 - 40)^2}{40} \]

  \[ + \frac{(250 - 270)^2}{270} + \frac{(300 - 270)^2}{270} + \frac{(50 - 60)^2}{40} \]

  \[ \chi^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60 \]
\[ \chi^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2 \]

where DF is the degrees of freedom, \( r \) is the number of levels of gender, \( c \) is the number of levels of the voting preference, \( n_r \) is the number of observations from level \( r \) of gender, \( n_c \) is the number of observations from level \( c \) of voting preference, \( n \) is the number of observations in the sample, \( E_{r,c} \) is the expected frequency count when gender is level \( r \) and voting preference is level \( c \), and \( O_{r,c} \) is the observed frequency count when gender is level \( r \) and voting preference is level \( c \).

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

We use the Chi-Square Distribution Calculator to find \( P(\chi^2 > 16.2) = 0.0003 \).

Interpret results. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.

Note: If you use this approach on an exam, you may also want to mention why this approach is appropriate. Specifically, the approach is appropriate because the sampling method was simple random sampling, each population was more than 10 times larger than its respective sample, the variables under study were categorical, and the expected frequency count was at least 5 in each cell of the contingency table.

The Mann-Whitney U-Test

Sometimes distributions of variables do not show a normal distribution, or the samples taken are so small that one cannot tell if they are part of a normal distribution or not. Using the t-test to tell if there is a significant difference between samples is not appropriate here. The Mann-Whitney U-test can be used in these situations. This test can be used for very small samples (between 5 and 20). It can also be used when the variable being recorded is measured using an arbitrary scale which cannot be measured accurately (e.g., a colour scale measured by eye or a behavioural trait such as aggression). The following example will illustrate the method. The size of leaves taken from bramble bushes were measured to see if there is a difference between the size of the leaves growing in full sunlight and those growing in the shade.
Width of leaf / cm
Sunlight 6.0 4.8 5.1 5.5 4.1 5.3 4.5 5.1
Shade 6.5 5.5 6.3 7.2 6.8 5.5 5.9 5.5

The Mann-Whitney U-test is chosen because the sample size is so small it is not clear if these are samples taken from normally distributed data.

1. Set up the Null Hypothesis: There is no difference between the leaves taken from the sunlit bramble and the shaded bramble. Alternative Hypothesis: There is a difference between the leaves taken from the sunlit bramble and the shade bramble.

2. Let n1 be the size of the smallest sample and n2 be the size of the biggest sample. In this example both are of the same size so it does not matter which you choose. n1 = 8 and n2 = 8

3. Rank all the values for both samples from the smallest (=1) to the largest. Set them up as shown in the table below.

<table>
<thead>
<tr>
<th>Sunlight Rank of X</th>
<th>Shade Rank of Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.0 12</td>
<td>6.5 14</td>
</tr>
<tr>
<td>4.8 3</td>
<td>5.5 8.5</td>
</tr>
<tr>
<td>5.1 4.5</td>
<td>6.3 13</td>
</tr>
<tr>
<td>5.5 8.5</td>
<td>7.2 16</td>
</tr>
<tr>
<td>4.1 1</td>
<td>6.8 15</td>
</tr>
<tr>
<td>5.3 6</td>
<td>5.5 8.5</td>
</tr>
<tr>
<td>4.5 2</td>
<td>5.9 10</td>
</tr>
<tr>
<td>5.1 4.5</td>
<td>5.5 8.5</td>
</tr>
<tr>
<td>41.5</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Note where the values are the same and share the same rank, take an average of the rank values.

4. Total the ranks of each sample R1 and R2 (see the bottom of the table above).

5. Calculate the U values for both samples:

$$U = n_1 n_2 + n_1 R_1 - n_1 n_2$$
6. Use the table to find the critical value for the U statistic at the 5% level for samples of this size (n1 = 8 and n2 = 8).

\[ U_{crit} = 13 \]

7. Reject the Null Hypothesis if the smallest value of U1 or U2 is below Ucrit. In this case U2 is below 13 we can reject the Null Hypothesis and accept the Alternative Hypothesis. The difference between the size of the bramble leaves in the light and the dark is significant for P>0.05. Bramble leaves in the dark seem to be significantly bigger.

Kruskal-Wallis Test

The Kruskal-Wallis Test was developed by Kruskal and Wallis jointly and is named after them. The Kruskal-Wallis test is a nonparametric (distribution free) test, which is used to compare three or more groups of sample data. Kruskal-Wallis Test is used when assumptions of ANOVA are not met. ANOVA is a statistical data analysis technique that is used when the independent variable groups are more than two. In ANOVA, we assume that distribution of each group should be normally distributed. In Kruskal-Wallis Test, we do not assume any assumption about the distribution. So Kruskal-Wallis Test is a distribution free test. If normality assumptions are met, then the Kruskal-Wallis Test is not as powerful as ANOVA. Kruskal-Wallis Test is also an improvement over the Sign test and Wilcoxon’s sign rank test, which ignores the actual magnitude of the paired magnitude.

Hypothesis in Kruskal-Wallis Test:

Null hypothesis: In Kruskal-Wallis Test, null hypothesis assumes that the samples are from identical populations.

Alternative hypothesis: In Kruskal-Wallis Test, alternative hypothesis assumes that the sample comes from different populations.

Hypothesis in Kruskal-Wallis Test:

1. In Kruskal-Wallis Test, we assume that the samples drawn from the population are random.

2. In Kruskal-Wallis Test, we also assume that the cases of each group are independent.

3. The measurement scale for Kruskal-Wallis Test should be at least ordinal.

Procedure for Kruskal-Wallis Test:
1. Arrange the data of both samples in a single series in ascending order.

2. Assign rank to them in ascending order. In the case of a repeated value, assign ranks to them by averaging their rank position.

3. Once this is complete, ranks or the different samples are separated and summed up as $R_1, R_2, R_3$, etc.

4. To calculate the value of Kruskal-Wallis Test, apply the following formula:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

Where, $H = $ Kruskal-Wallis Test

$n = $ total number of observations in all samples

$R_i = $ Rank of the sample

Kruskal-Wallis Test statistics is approximately a chi-square distribution, with $k-1$ degree of freedom where $n_i$ should be greater than 5. If the calculated value of Kruskal-Wallis Test is less than the chi-square table value, then the null hypothesis will be accepted. If the calculated value of Kruskal-Wallis Test $H$ is greater than the chi-square table value, then we will reject the null hypothesis and say that the sample comes from a different population.

Spearman’s Rank correlation coefficient

A correlation can easily be drawn as a scatter graph, but the most precise way to compare several pairs of data is to use a statistical test - this establishes whether the correlation is really significant or if it could have been the result of chance alone. Spearman’s Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables.

The result will always be between 1 and minus 1. Method - calculating the coefficient

Create a table from your data.

Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest number in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking. This should be done for both sets of measurements.
Tied scores are given the mean (average) rank. For example, the three tied scores of 1 euro in the example below are ranked fifth in order of price, but occupy three positions (fifth, sixth and seventh) in a ranking hierarchy of ten. The mean rank in this case is calculated as $(5+6+7) ÷ 3 = 6$.

Find the difference in the ranks ($d$): This is the difference between the ranks of the two values on each row of the table. The rank of the second value (price) is subtracted from the rank of the first (distance from the museum).

Square the differences ($d^2$) To remove negative values and then sum them ($d^2$).

<table>
<thead>
<tr>
<th>Convenience Store</th>
<th>Distance from CAM (m)</th>
<th>Rank of bottle price (€)</th>
<th>Rank of 50cl bottle price (€)</th>
<th>Difference in ranks (d)</th>
<th>$d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>10</td>
<td>1.80</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>9</td>
<td>1.20</td>
<td>3.5</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30.25</td>
</tr>
<tr>
<td>3</td>
<td>270</td>
<td>8</td>
<td>2.00</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>49</td>
</tr>
<tr>
<td>4</td>
<td>375</td>
<td>7</td>
<td>1.00</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>425</td>
<td>6</td>
<td>1.00</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Data Table: Spearman's Rank Correlation

Calculate the coefficient \( R \) using the formula below. The answer will always be between 1.0 (a perfect positive correlation) and -1.0 (a perfect negative correlation).

When written in mathematical notation the Spearman Rank formula looks like this:

\[
R = 1 - \frac{6 \sum d^2}{n^3 - n}
\]

Now to put all these values into the formula.

Find the value of all the \( d^2 \) values by adding up all the values in the Difference² column. In our example this is 285.5. Multiplying this by 6 gives 1713.

Now for the bottom line of the equation. The value \( n \) is the number of sites at which you took measurements. This, in our example is 10. Substituting these values into \( n^3 - n \) we get 1000 - 10.
We now have the formula: \( R = 1 - \frac{1713}{990} \) which gives a value for \( R \):

\[
1 - 1.73 = -0.73
\]

What does this \( R \) value of -0.73 mean?

The closer \( R \) is to +1 or -1, the stronger the likely correlation. A perfect positive correlation is +1 and a perfect negative correlation is -1. The \( R \) value of -0.73 suggests a fairly strong negative relationship.

A further technique is now required to test the significance of the relationship.

The \( R \) value of -0.73 must be looked up on the Spearman Rank significance table below as follows:

- Work out the 'degrees of freedom' you need to use. This is the number of pairs in your sample minus 2 (\( n-2 \)). In the example it is 8 (10 - 2).
- Now plot your result on the table.
If it is below the line marked 5%, then it is possible your result was the product of chance and you must reject the hypothesis.

- If it is above the 0.1% significance level, then we can be 99.9% confident the correlation has not occurred by chance.

- If it is above 1%, but below 0.1%, you can say you are 99% confident.

- If it is above 5%, but below 1%, you can say you are 95% confident (i.e. statistically there is a 5% likelihood the result occurred by chance).

In the example, the value 0.73 gives a significance level of slightly less than 5%. That means that the probability of the relationship you have found being a chance event is about 5 in a 100. You are 95% certain that your hypothesis is correct. The reliability of your sample can be stated in terms of how many researchers completing the same study as yours would obtain the same results: 95 out of 100.

The fact two variables correlate cannot prove anything - only further research can actually prove that one thing affects the other.

- Data reliability is related to the size of the sample. The more data you collect, the more reliable your result.

CORRELATION, REGRESSION AND TIME SERIES ANALYSIS


**Correlation analysis**

*Use the Correlation transformer to determine the extent to which changes in the value of an attribute (such as length of employment) are associated with changes in another attribute (such as salary). The data for a correlation analysis consists of two input columns. Each column contains values for one of the attributes of interest.*
Correlation transformer can calculate various measures of association between the two input columns. You can select more than one statistic to calculate for a given pair of input columns.

The data in the input columns also can be treated as a sample obtained from a larger population, and the Correlation transformer can be used to test whether the attributes are correlated in the population. In this context, the null hypothesis asserts that the two attributes are not correlated, and the alternative hypothesis asserts that the attributes are correlated.

The Correlation transformer calculates any of the following correlation-related statistics on one or more pairs of columns:

Correlation coefficient \( r \)

The correlation coefficient \( r \) is a measure of the linear relationship between two attributes or columns of data. The correlation coefficient is also known as the Pearson product-moment correlation coefficient. The value of \( r \) can range from -1 to +1 and is independent of the units of measurement. A value of \( r \) near 0 indicates little correlation between attributes; a value near +1 or -1 indicates a high level of correlation.

When two attributes have a positive correlation coefficient, an increase in the value of one attribute indicates a likely increase in the value of the second attribute. A correlation coefficient of less than 0 indicates a negative correlation. That is, when one attribute shows an increase in value, the other attribute tends to show a decrease.

Consider two variables \( x \) and \( y \):

- If \( r = 1 \), then \( x \) and \( y \) are perfectly positively correlated. The possible values of \( x \) and \( y \) all lie on a straight line with a positive slope in the \((x,y)\) plane.

- If \( r = 0 \), then \( x \) and \( y \) are not correlated. They do not have an apparent linear
relationship. However, this does not mean that $x$ and $y$ are statistically independent.

- If $r = -1$, then $x$ and $y$ are perfectly negatively correlated. The possible values of $x$ and $y$ all lie on a straight line with a negative slope in the $(x,y)$ plane.

### Covariance

- **Covariance** is a measure of the linear relationship between two attributes or columns of data. The value of the covariance can range from $-\infty$ to $+\infty$. However, if the value of the covariance is too small or too large to be represented by a number, the value is represented by **NULL**.

- Unlike the correlation coefficient, the covariance is dependent on the units of measurement. For example, measuring values of two attributes in inches rather than feet increases the covariance by a factor of 144.

### Finding Regression Estimations

To better understand what regression estimation is, one must become familiar with the process of finding the regression estimation.

The first steps for finding regression estimations are to collect **bivariate data** and plot it on a scatter plot. The scatter plot should have a linear correlation, in order to have a regression estimation. By having a linear correlation, one can then draw a line of best fit or regression line. Once these steps are complete, one can predict missing values (regression estimations) by using the regression equation. The regression equation describes the line of best fit and is defined as $Y' = a + bX$, where $Y'$ is the value that one is trying to predict, $X$ is the value that one is given, $a$ is the point where the regression line crosses the **y-axis** of the scatter plot, and $b$ represents the slope of the regression line. Most may better recognize this equation as slope-intercept form. Finally, to find the regression estimation, plug $a$, $b$, and $X$ into the regression equation and solve for $Y'$.

### Introduction

A time series is a set of numerical values of a given variable listed at successive intervals of time. That is, the data regarding the variable is listed in chronological order. Usually the interval of time is taken as uniform.
Example: Yearly production of wheat in the country, hourly temperature of a city, bimonthly electricity bills etc. Almost all the data like industrial production, agricultural production, exports, imports, diary products can be arranged in chronological order.

Learning Objective 1

Know the Concept of Time-series

A time-series is a set of numerical values of a given variable listed at successive intervals of time. That is, the data regarding the variable is listed in chronological order. Usually the interval of time is taken as uniform.

Example: Yearly production of wheat in the country, hourly temperature of a city, bimonthly electricity bills etc. Almost all the data like industrial production, agricultural production, exports, imports, diary products can be arranged in chronological order.

Time Series Analysis

Given a time series, we wish to

i. Study the forces that influence the variations in time series, and

ii. Study the behaviour of phenomenon over the given period of time.

For example, consider the sale of T.V sets (in thousands) by a producing company

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number sold (in thousands)</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>10</td>
<td>18</td>
</tr>
</tbody>
</table>

We would like to analyse the above data and give some trends about the sales. For example, the company would like to know as to why the sales dropped in 1998 and 1999, and then why the sales increased. That is, the company would like to analyse the various forces that affect the sales.

There can be changes in the values of the variable recorded over different points of time due to various forces. Analysing the effect of all such forces on the values of the variable is generally known as the analysis of time series. Broadly there can be four types of changes in the values of the variable as discussed below:

i. Changes which generally occur due to general tendency of the data to increase or decrease.

ii. Changes which occur due to change in climate, weather conditions, festivals etc.
iii. Changes which occur due to booms and depressions.

iv. Changes which occur due to some unpredictable forces like floods, famines, earthquakes etc.

Learning Objective 2

Understand Different Components of a Time-series Components of Time Series

The behaviour of a time series over periods of time is called the movement of the time series. The time series is classified into the following four components:

- Long Term Trend or Secular Trend: This refers to the smooth or regular long term growth or decline of the series. This movement can be characterised by a trend curve. If this curve is a straight line, that is called a trend line. If the variable is increasing over a long period of time, then it is called an upward trend. If the variable is decreasing over a long period of time, then it is called downward trend. If the variable moves upward or downwards along a straight line then the trend is called a linear trend, otherwise it is called a non-linear trend.

- Seasonal Variations: Variations in a time series that are periodic in nature and occur regularly over short periods of time during an year are called seasonal variations. By definition, these variations are precise and can be forecasted.

Examples:

i. The prices of vegetables drop down after rainy season or in winter months and they go up during summer, every year.

ii. The prices of cooking oils reduce after the harvesting of oil seeds and go up after some time.

- Cyclic Variations: The long-term oscillations that represent consistent rises and declines in the values of the variable are called cyclic variations. Since these are long-term oscillations in the time series, the period of oscillation is usually greater than one year. The oscillations are about a trend curve or a trend line. The period of one cycle is the time-distance between two successive peaks or two successive troughs.

- Random Variations: These are called irregular movements. Movements that occur usually in brief periods of time, without any pattern and are unpredictable in nature are called irregular movements. These movements do not have any regular period or time of occurrences. Example: The effects of national strikes, floods, earthquakes etc. It is very difficult to study the behaviour of such a time series.
Learning Objective 3

Know Different Methods for Time-series Analysis

Methods of Measuring Trend:

We shall be studying the following methods of measuring the trend of a time series:

1. Free hand or graphic method
2. Semi averages method
3. Moving average method
4. Method of least squares

Free hand or graphic method: This is the simplest method to drawing a trend curve. We plot the values of the variable against time on a graph paper and join these points. The trend line is then fitted by inspecting the graph of the time series. Fitting a trend line by this method is arbitrary. The trend line is usually drawn such that the numbers of fluctuations on either side are approximately the same. The trend line should be a smooth curve. The free hand method has some disadvantages. They are:

i. it depends on individual judgement

ii. It cannot be used for any predictions of trends, as drawing the trend curve is arbitrary.

Example: Find trend with the help of freehand curve method for the data given below:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Production (in lakh ton)</td>
<td>15</td>
<td>18</td>
<td>16</td>
<td>22</td>
<td>19</td>
<td>24</td>
<td>20</td>
<td>28</td>
<td>22</td>
<td>30</td>
<td>26</td>
</tr>
</tbody>
</table>

Solution:
The methods of fitting a linear trend with the help of semi average method are as follows:

i. *The number of years in even:* The data of the time series are divided into two equal parts. The total of the items in each of the part is done and it is then divided by the number of items to obtain arithmetic means of the two parts. Each average is then centred in the period of time from which it has been computed and plotted on the graph paper. A straight line is drawn passing through these points. This is the required trend line.

ii. *The number of years is odd:* When the number of years is odd, the value of the middle year is omitted to divide the time series into two equal parts. Then the procedure (i) is followed.

A trend value of any future year may be predicted by multiplying the periodic increment by the number of years into the future that is desired and adding the result to the best trend value listed in the series.
Merits:

i. The method is simple

ii. The trend line can be extended on either side in order to obtain past or future estimates.

iii. This is an objective method, as any one applying this method get the same trend line.

Demerits:

i. The method of semi average assumes a straight line relationship between the plotted points, regardless of the fact whether such relationship exists or not.

ii. This method has an in built limitation of arithmetic mean. This method is not suitable is case of very low or very large extreme values.

iii. There is no assurance that the influence of cycle is eliminated.

Method of Moving Averages

This method is used for smoothing the time series. That is, it smoothen the fluctuations of the data by the method of moving averages.

a. When Period of moving average is odd: To determine the trend by this method, we use the following method:

i. Obtain the time series

ii. Select a period of moving average such as 3 years, 5 years etc.

iii. Compute moving totals according to the length of the period of moving average.

If the length of the period of moving average is 3 i.e., 3-yearly moving average is to be calculated, compute moving totals as follows:

\[ a + b + c, b + c + d, c + d + e, d + e + f, \ldots \]

for 5-years moving average, moving totals are computed as follows: \[ a + b + c + d + e, b + c + d + e + f, c + d + e + f + g, \ldots \]

Placing the moving totals at the centre of the time span from which they are computed.

iv. Compute moving averages by moving totals in step (3) by the length of the period of moving average and place them at the centre of the time span from which the moving totals are computed. These moving averages are also called the trend values.
By plotting these trend values (if desired) one can obtain the trend curve with the help of which we can determine the trend whether it is increasing or decreasing.

History of early price indices

No clear consensus has emerged on who created the first price index. The earliest reported research in this area came from Welsh poet Henry Rice Vaughan who examined price level change in his 1675 book *A Discourse of Coin and Coinage*. Vaughan wanted to separate the inflationary impact of the influx of precious metals brought by Spain from the New World from the effect due to currency debasement. Vaughan compared labor statutes from his own time to similar statutes dating back to Edward III. These statutes set wages for certain tasks and provided a good record of the change in wage levels. Vaughan reasoned that the market for basic labor did not fluctuate much with time and that a basic laborer's salary would probably buy the same amount of goods in different time periods, so that a laborer's salary acted as a basket of goods. Vaughan's analysis indicated that price levels in England had risen six to eightfold over the preceding century. [1]

While Vaughan can be considered a forerunner of price index research, his analysis did not actually involve calculating an index. [1] In 1707 Englishman William Fleetwood created perhaps the first true price index. An Oxford student asked Fleetwood to help show how prices had changed. The student stood to lose his fellowship since a fifteenth century stipulation barred students with annual incomes over five pounds from receiving a fellowship. Fleetwood, who already had an interest in price change, had collected a large amount of price data going back hundreds of years. Fleetwood proposed an index consisting of averaged price relatives and used his methods to show that the value of five pounds had changed greatly over the course of 260 years. He argued on behalf of the Oxford students and published his findings anonymously in a volume entitled *Chronicon Preciosum*. [2]

Formal calculation

Further information: [List of price index formulas](#)

Given a set $C$ of goods and services, the total market value of transactions in $C$ in some period $t$ would be

$$\sum_{c \in C} (p_{c,t} \cdot q_{c,t})$$

where
If, across two periods \( t_0 \) and \( t_n \), the same quantities of each good or service were sold, but under different prices, then

\[
q_{c,t_n} = q_c = q_{c,t_0} \quad \forall c
\]

and

\[
P = \frac{\sum(p_{c,t_n} \cdot q_c)}{\sum(p_{c,t_0} \cdot q_c)}
\]

would be a reasonable measure of the price of the set in one period relative to that in the other, and would provide an index measuring relative prices overall, weighted by quantities sold.

Of course, for any practical purpose, quantities purchased are rarely if ever identical across any two periods. As such, this is not a very practical index formula.

One might be tempted to modify the formula slightly to

\[
P = \frac{\sum(p_{c,t_n} \cdot q_{c,t_n})}{\sum(p_{c,t_0} \cdot q_{c,t_0})}
\]

This new index, however, doesn't do anything to distinguish growth or reduction in quantities sold from price changes. To see that this is so, consider what happens if all the prices double between \( t_0 \) and \( t_n \) while quantities stay the same: \( P \) will double. Now consider what happens if all the quantities double between \( t_0 \) and \( t_n \) while all the prices stay the same: \( P \) will double. In either case the change in \( P \) is identical. As such, \( P \) is as much a quantity index as it is a price index.

Various indices have been constructed in an attempt to compensate for this difficulty.

### Paasche and Laspeyres price indices

The two most basic formulae used to calculate price indices are the Paasche index (after the economist Hermann Paasche[ˈpaːʃə]) and the Laspeyres index (after the economist Etienne Laspeyres[ˈlaspejə]).

The Paasche index is computed as

\[
P_P = \frac{\sum(p_{c,t_n} \cdot q_{c,t_n})}{\sum(p_{c,t_0} \cdot q_{c,t_n})}
\]

while the Laspeyres index is computed as

\[
P_L = \frac{\sum(p_{c,t_n} \cdot q_{c,t_0})}{\sum(p_{c,t_0} \cdot q_{c,t_0})}
\]
where $P_t$ is the relative index of the price levels in two periods, $t$ is the base period (usually the first year), and $t_n$, the period for which the index is computed.

Note that the only difference in the formulas is that the former uses period $n$ quantities, whereas the latter uses base period (period 0) quantities.

When applied to bundles of individual consumers, a Laspeyres index of 1 would state that an agent in the current period can afford to buy the same bundle as she consumed in the previous period, given that income has not changed; a Paasche index of 1 would state that an agent could have consumed the same bundle in the base period as she is consuming in the current period, given that income has not changed.

Hence, one may think of the Laspeyres index as one where the **numeraire** is the bundle of goods using current year prices but base year quantities. Similarly, the Paasche index can be thought of as a price index taking the bundle of goods using current prices and current quantities as the numeraire.

The Laspeyres index tends to overstate inflation (in a cost of living framework), while the Paasche index tends to understate it, because the indices do not account for the fact that consumers typically react to price changes by changing the quantities that they buy. For example, if prices go up for good $c$, *ceteris paribus*, quantities of that good should go down.

**Fisher index and Marshall–Edgeworth index**

A third index, the **Marshall–Edgeworth index** (named for economists Alfred Marshall and Francis Ysidro Edgeworth), tries to overcome these problems of under- and overstatement by using the arithmetic means of the quantities:

$$P_{ME} = \frac{\sum [p_{c,t_n} \cdot \frac{1}{2} \cdot (q_{c,t_0} + q_{c,t_n})]}{\sum [p_{c,t_0} \cdot \frac{1}{2} \cdot (q_{c,t_0} + q_{c,t_n})]}$$

A fourth, the Fisher index (after the American economist Irving Fisher), is calculated as the geometric mean of $P_P$ and $P_L$:

$$P_F = \sqrt{P_P \cdot P_L}$$

Fisher’s index is also known as the “ideal” price index.

However, there is no guarantee with either the Marshall–Edgeworth index or the Fisher index that the overstatement and understatement will exactly cancel the other.

While these indices were introduced to provide overall measurement of relative prices, there is ultimately no way of measuring the imperfections of any of these indices (Paasche, Laspeyres, Fisher, or Marshall–Edgeworth) against reality.

**Practical measurement considerations**

**Normalizing index numbers**

Price indices are represented as **index numbers**, number values that indicate relative change but not absolute values (i.e. one price index value can be compared to another or a base, but the number alone has no meaning).
indices generally select a base year and make that index value equal to 100. You then express every other year as a percentage of that base year. In our example above, let's take 2000 as our base year. The value of our index will be 100. The price

- 2000: original index value was $2.50; $2.50/$2.50 = 100%, so our new index value is 100
- 2001: original index value was $2.60; $2.60/$2.50 = 104%, so our new index value is 104
- 2002: original index value was $2.70; $2.70/$2.50 = 108%, so our new index value is 108
- 2003: original index value was $2.80; $2.80/$2.50 = 112%, so our new index value is 112

When an index has been normalized in this manner, the meaning of the number 108, for instance, is that the total cost for the basket of goods is 4% more in 2001, 8% more in 2002 and 12% more in 2003 than in the base year (in this case, year 2000).

Relative ease of calculating the Laspeyres index

As can be seen from the definitions above, if one already has price and quantity data (or, alternatively, price and expenditure data) for the base period, then calculating the Laspeyres index for a new period requires only new price data. In contrast, calculating many other indices (e.g., the Paasche index) for a new period requires both new price data and new quantity data (or, alternatively, both new price data and new expenditure data) for each new period. Collecting only new price data is often easier than collecting both new price data and new quantity data, so calculating the Laspeyres index for a new period tends to require less time and effort than calculating these other indices for a new period.

Calculating indices from expenditure data

Sometimes, especially for aggregate data, expenditure data is more readily available than quantity data. For these cases, we can formulate the indices in terms of relative prices and base year expenditures, rather than quantities.

Here is a reformulation for the Laspeyres index:

Let $E_{c,t0}$ be the total expenditure on good $c$ in the base period, then (by definition) we have $E_{c,t0} = p_{c,t0} \cdot q_{c,t0}$

and therefore also $p_{c,t0} = \frac{E_{c,t0}}{q_{c,t0}}$. We can substitute these values into our Laspeyres formula as follows:

$$P_L = \frac{\sum (p_{c,t_n} \cdot q_{c,t_n})}{\sum (p_{c,t_0} \cdot q_{c,t_0})} = \frac{\sum (\frac{p_{c,t_n}}{p_{c,t_0}} \cdot E_{c,t_0})}{\sum E_{c,t_0}} = \frac{\sum (p_{c,t_n} \cdot E_{c,t_0})}{\sum E_{c,t_0}}$$

A similar transformation can be made for any index.

Chained vs non-chained calculations

So far, in our discussion, we have always had our price indices relative to some fixed base period. An alternative is to take the base period for each time period to be the immediately preceding time period. This can be done with any of the above indices. Here's an example with the Laspeyres index, where $t_n$ is the period for which we wish to calculate the index and $t_0$ is a reference period that anchors the value of the series:

$$P_{t_n} = \frac{\sum (p_{c,t_1} \cdot q_{c,t_0})}{\sum (p_{c,t_0} \cdot q_{c,t_0})} \times \frac{\sum (p_{c,t_2} \cdot q_{c,t_1})}{\sum (p_{c,t_1} \cdot q_{c,t_1})} \times \cdots \times \frac{\sum (p_{c,t_n} \cdot q_{c,t_{n-1}})}{\sum (p_{c,t_{n-1}} \cdot q_{c,t_{n-1}})}$$

Each term
answers the question "by what factor have prices increased between period $t_{n-1}$ and period $t_n$". When you multiply these all together, you get the answer to the question "by what factor have prices increased since period $t_{t1}$".

Nonetheless, note that, when chain indices are in use, the numbers cannot be said to be "in period $t_{t1}$" prices.

Index number theory

Price index formulas can be evaluated based on their relation to economic concepts (like cost of living) or on their mathematical properties. Several different tests of such properties have been proposed in index number theory literature. W.E. Diewert summarized past research in a list of nine such tests for a price index $I(P_{t0}, P_{tm}, Q_{t0}, Q_{tm})$, where $P_{t0}$ and $P_{tm}$ are vectors giving prices for a base period and a reference period while $Q_{t0}$ and $Q_{tm}$ give quantities for these periods.[5]

1. Identity test:

$$I(p_{tm}, p_{tm}, \alpha \cdot q_{tm}, \beta \cdot q_{tn}) = 1 \text{ for all } \alpha, \beta \in (0, \infty)$$

The identity test basically means that if prices remain the same and quantities remain in the same proportion to each other (each quantity of an item is multiplied by the same factor of either $\alpha$, for the first period, or $\beta$, for the later period) then the index value will be one.

2. Proportionality test:

$$I(p_{tm}, \alpha \cdot p_{tm}, q_{tm}, q_{tn}) = \alpha \cdot I(p_{tm}, p_{tn}, q_{tm}, q_{tn})$$

If each price in the original period increases by a factor $\alpha$ then the index should increase by the factor $\alpha$.

3. Invariance to changes in scale test:

$$I(\alpha \cdot p_{tm}, \alpha \cdot p_{tn}, \beta \cdot q_{tm}, \gamma \cdot q_{tn}) = I(p_{tm}, p_{tn}, q_{tm}, q_{tn}) \text{ for all } \alpha, \beta, \gamma \in (0, \infty)$$

The price index should not change if the prices in both periods are increased by a factor and the quantities in both periods are increased by another factor. In other words, the magnitude of the values of quantities and prices should not affect the price index.

4. Commensurability test:

The index should not be affected by the choice of units used to measure prices and quantities.

5. Symmetric treatment of time (or, in parity measures, symmetric treatment of place):

$$I(p_{tn}, p_{tm}, q_{tn}, q_{tm}) = \frac{1}{I(p_{tm}, p_{tn}, q_{tm}, q_{tn})}$$

Reversing the order of the time periods should produce a reciprocal index value. If the index is calculated from the most recent time period to the earlier time period, it should be the reciprocal of the index found going from the earlier period to the more recent.
6. Symmetric treatment of commodities:

All commodities should have a symmetric effect on the index. Different permutations of the same set of vectors should not change the index.

7. Monotonicity test:

\[ I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}) \leq I(p_{t_m}, p_{t_r}, q_{t_m}, q_{t_r}) \iff p_{t_n} \leq p_{t_r} \]

A price index for lower later prices should be lower than a price index with higher later period prices.

8. Mean value test:

The overall price relative implied by the price index should be between the smallest and largest price relatives for all commodities.

9. Circularity test:

\[ I(p_{t_m}, p_{t_n}, q_{t_m}, q_{t_n}) \cdot I(p_{t_n}, p_{t_r}, q_{t_n}, q_{t_r}) = I(p_{t_m}, p_{t_r}, q_{t_m}, q_{t_r}) \iff t_m \leq t_n \leq t_r \]

Given three ordered periods \( t_{m}, t_{n}, t_{r} \), the price index for periods \( t_{m} \) and \( t_{r} \) times the price index for periods \( t_{n} \) and \( t_{r} \) should be equivalent to the price index for periods \( t_{m} \) and \( t_{r} \).

Quality change

Price indices often capture changes in price and quantities for goods and services, but they often fail to account for variation in the quality of goods and services. Statistical agencies generally use matched-model price indices, where one model of a particular good is priced at the same store at regular time intervals. The matched-model method becomes problematic when statistical agencies try to use this method on goods and services with rapid turnover in quality features. For instance, computers rapidly improve and a specific model may quickly become obsolete. Statisticians constructing matched-model price indices must decide how to compare the price of the obsolete item originally used in the index with the new and improved item that replaces it. Statistical agencies use several different methods to make such price comparisons. \[6\]

The problem discussed above can be represented as attempting to bridge the gap between the price for the old item at time \( t \), \( P(M)_{t} \), with the price of the new item at the later time period, \( P(N)_{t+1} \). \[7\]

- The overlap method uses prices collected for both items in both time periods, \( t \) and \( t+1 \). The price relative \( P(N)_{t+1}/P(N)_{t} \) is used.
- The direct comparison method assumes that the difference in the price of the two items is not due to quality change, so the entire price difference is used in the index. \( P(N)_{t+1}/P(M)_{t} \) is used as the price relative.
- The link-to-show-no-change assumes the opposite of the direct comparison method; it assumes that the entire difference between the two items is due to the change in quality. The price relative based on link-to-show-no-change is \( 1 \). \[8\]
- The deletion method simply leaves the price relative for the changing item out of the price index. This is equivalent to using the average of other price relatives in the index as the price relative for the changing item. Similarly, class mean imputation uses the average price relative for items with similar characteristics (physical, geographic, economic, etc.) to \( M \) and \( N \).