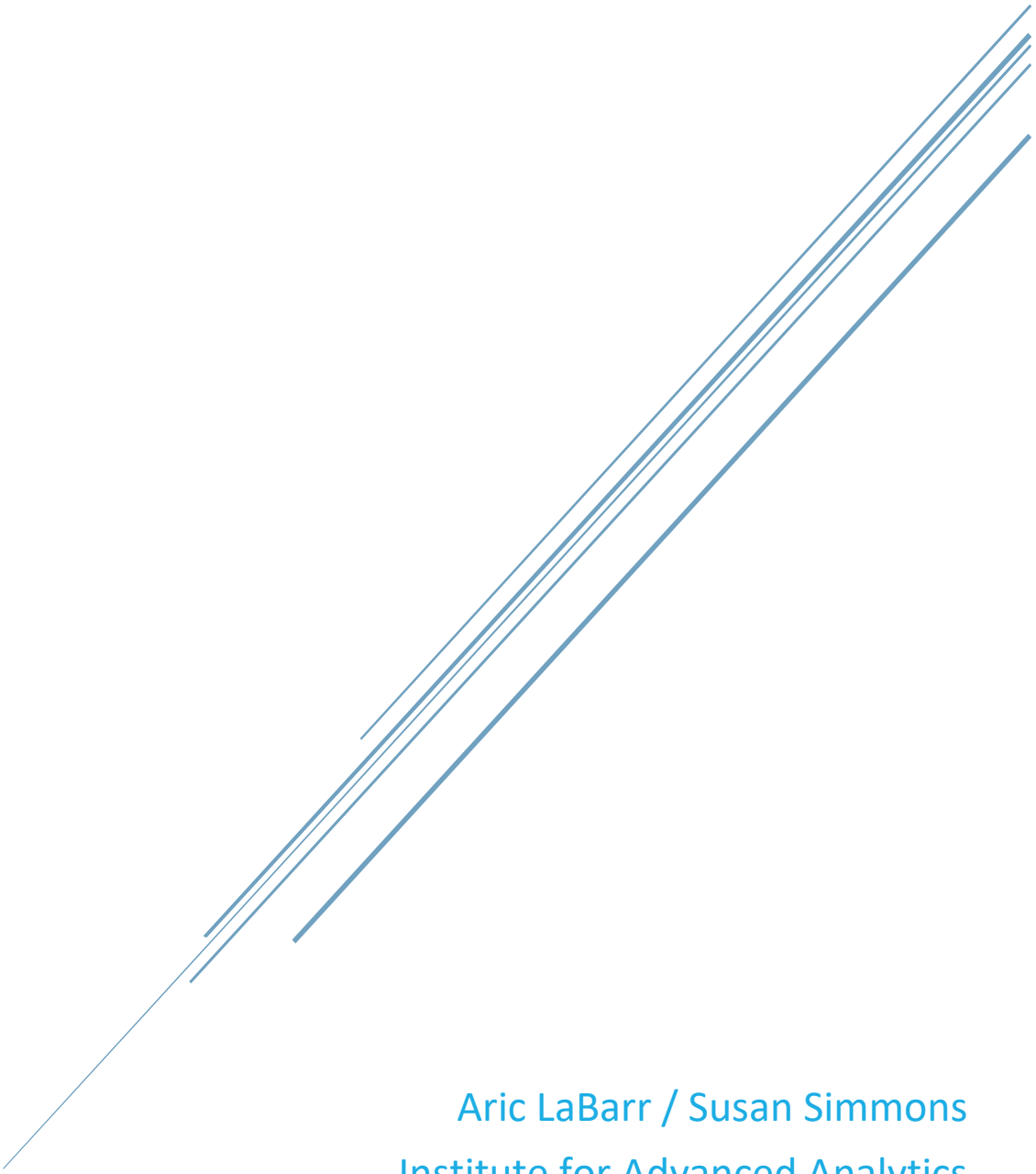
# STATISTICS PRIMER

Course Notes



Aric LaBarr / Susan Simmons

Institute for Advanced Analytics

CONTENTS

## FUNDAMENTAL STATISTICAL CONCEPTS

Statistics boils down to three main pieces:

1. Data Collection
2. Analysis
3. Inference

### DATA COLLECTION

The most important part of statistics is the data collection. Data analysis and inference are incorrect when data is not collected properly. The three terms in statistics most important to data collection are population, frame, and sample.

- Population:

- Sample:

- Sampling Frame:

Most of our focus will be on the population and the sample. Different terminology is used when describing populations and samples. These terms are fundamental to understanding statistics.

- Parameter:

- Statistic:

### EXAMPLE

A retail store chain is trying to determine if a new product they introduced is selling well across their stores. The retail store chain has 2135 stores nation- wide. The analyst in charge of this project is tasked to estimate the average daily sales in dollars of this new product across all of the stores. Even though every store can be contacted, older computing technology and a lack of re- porting efficiency in the company makes obtaining detailed sales information for each store in the chain difficult, expensive, and time consuming. The analyst decides to randomly pick 179 stores spread evenly throughout the nation and calculate their average daily sales in dollars for the new product. The average daily sales from these 179 stores is $129.19. What is the population, sample, parameter, and statistic in this example? Does there appear to be any issues regarding the population and frame in this example?

- Population:
- Sample:
- Parameter:
- Statistic:
- Sampling Frame Issues:

## SAMPLING TECHNIQUES

There are many different ways to sample from a population. Unfortunately, there are many mistakes that can be made when sampling data, which potentially leads to bias in a sample. Bias occurs when certain outcomes are favored over other outcomes in samples. Two common types of bias are selection bias and sampling bias.

1. Selection Bias:
    a. Undercoverage:

b. Nonresponse:

2. Sampling Bias:

    a. Convenience Sampling:

    b. Voluntary Sampling:

Unlike the previous sampling techniques, statistical sampling techniques use selection methods based on chance selection instead of convenience or judgment. Four statistical techniques we will discuss are simple random sampling (SRS), stratified random sampling, cluster sampling, and systematic sampling.

1. **Simple Random Sampling (SRS):** A method of sampling items from a population such that every possible sample of a specified size has an equal chance of being selected.
    a. Advantage:

    b. Disadvantage:

2. **Stratified Random Sampling:** A method of sampling items where the population is divided *a priori* into subgroups called strata so that each member in the population belongs to only one strata. The strata should be formed such that population values of interest within the strata are similar. Sample items from **every** stratum with SRS.
    a. Advantage:

    b. Disadvantage:

3. **Cluster Sampling:** A method of sampling items where the population is divided *a priori* into subgroups called clusters where each cluster is intended to be a mini-population. Sample items from a **sample** of m clusters selected with simple random sampling. In cluster sampling, all items within the selected cluster are observed.
    a. Advantage:

    b. Disadvantage:

4. **Systematic Sampling:** A method of sampling items that involves selecting every $k^{th}$ item in the population after randomly selecting a starting point between 1 and k. The value k is determined as the ratio of the population size over the desired sample size.
    a. Advantage:

    b. Disadvantage:

EXAMPLE

Suppose a large worldwide financial company wants to develop a new retirement plan for the company. They want to survey different managers of branches around the world to find out the most important strategies the new retirement plan should contain. They have 5000 branches worldwide and want to personally interview these branch managers. They have information about the branch size (either small, medium, or large) and the state/province location of the branch. Develop four separate strategies to sample these branch managers based on the four different statistical sampling techniques discussed above.

- Simple Random Sampling:

- Stratified Random Sampling:

- Cluster Sampling:

- Systematic Sampling:

## TYPES OF DATA

Now that we know how to collect data, we will learn about different types of data along with some brief analysis techniques for the data. There are four main types of data collected in statistics. They are split into two groups as follows:

- Quantitative vs. Qualitative
    - Quantitative:

    - Qualitative:

        - Nominal:

        - Ordinal:

- Time-Series vs. Cross-Sectional
    - Time-Series:

    - Cross-Sectional:

# PROBABILITY

## EVENTS

Before working with distributions of data, it is helpful to understand the characteristics and properties of probability. Probability is the chance that a particular event will occur. Most people confuse probabilities and percent- ages. Probabilities are numbers that are between 0 and 1. Percentages are numbers that are between 0 and 100. A probability of 0 means that an event will not occur, while a probability of 1 means the event will occur. When probabilities are between 0 and 1, the event may or may not occur. A higher probability signifies a higher chance of an event occurring. One way of estimating probabilities is to use the relative frequency of the event (assuming the information used to calculate the relative frequency is representative).

The sample space is the collection of all outcomes in a random process (rolling a die, choosing a card, etc). When listing all outcomes in a random process, the sum of all probabilities must be one.

An event, $A$, is a collection of one or more outcomes from a process whose result cannot be predicted with certainty. The probability of an event $A_i$ is denoted $P(A_i)$.

## RULES OF PROBABILITY

Now that probabilities are defined, the rules for these probabilities are established.

- Bounded Rule:

$$0 \leq P(A_i) \leq 1 \text{ for all } i$$

- Complement Rule:

$$P(\textbf{\textit{not}}\ A) = P(\bar{A}) = P(A^c) = 1 - P(A)$$

The **union** of an event *A* and an event *B* is the event containing all sample points that are in *A* **or** *B* (or both). The union of *A* and *B* is denoted by $A \cup B$. This is displayed in Venn Diagrams in Figure 1 below.

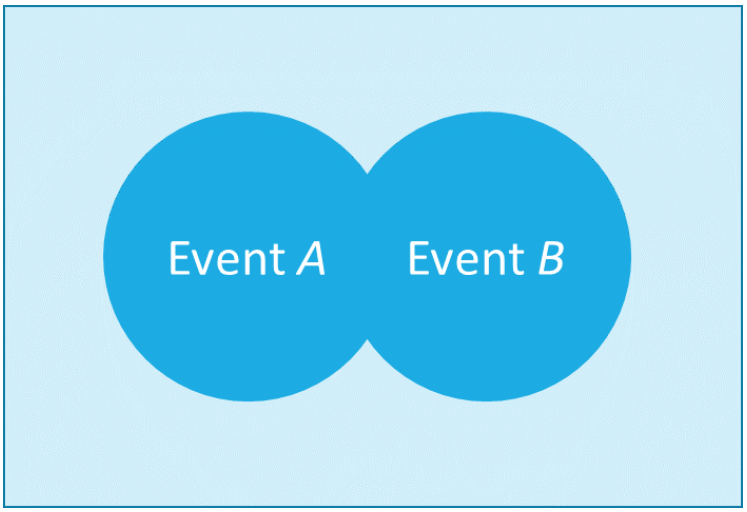



**Figure 1: Union of Events A and B**

The **intersection** of an event *A* and an event *B* is the event containing all sample points that are in **both** *A* and *B*. The intersection of *A* and *B* is denoted by $A \cap B$. This is displayed in Venn Diagrams in Figure 2 below.

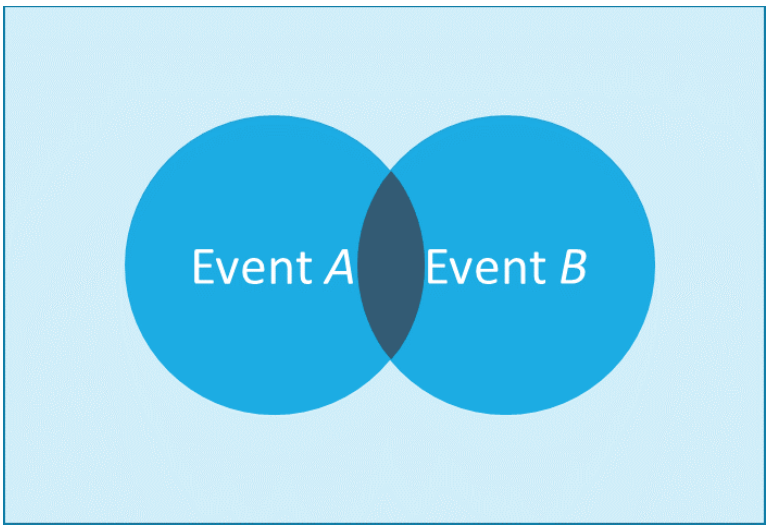



**Figure 2: Intersection of Events A and B**

- Addition Rules:
  - Any Two Events:

$$P(A \textbf{ or } B) = P(A) + P(B) - P(A \textbf{ and } B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  - **Mutually Exclusive Events**: Two events are **mutually exclusive** is the events have no sample points in common – DO NOT INTERSECT. This also means that the events cannot both occur. If one occurs, the other cannot. This changes the addition rule ($P(A \cap B) = 0$) to the following:

$$P(A \cup B) = P(A) + P(B)$$

- **Conditional (Joint) Probabilities**: The probability that an event will occur **given** that some other event has already happened. The conditional probability of an event *A* given that an event *B* has already occurred is denoted by $P(A|B)$.
  - Any Two Events:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplication Rules:
  - Any Two Events:

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

Events can also be either independent or dependent. Two events are **independent** if the occurrence of one event doesn't influence the probability of the occurrence of the other event. Therefore, two events are **dependent** if the occurrence of one event impacts the probability of the occurrence of the other event. Mutually exclusive events are a **special case** of dependent events because the occurrence of one event precludes the occurrence of a second event. If two events are independent, this changes the calculation of conditional probabilities which impacts the multiplication rule:

- Conditional Probabilities (**under independence**):

$$P(A|B) = P(A) \qquad \text{or} \qquad P(B|A) = P(B)$$

- Multiplication Rule (**under independence**):

$$P(A \cap B) = P(A) \times P(B)$$

**Marginal probabilities** can be thought of as unconditional probabilities – probabilities of events without any condition.

EXAMPLE

**Table 1: Number of Credit Cards Per Age Group**

| # Credit Cards | 20-29 Years Old | 30-39 Years Old | 40-49 Years Old | 50+ Years Old | TOTAL |
|---|---|---|---|---|---|
| 0 | 56 | 24 | 33 | 98 | 210 |
| 1-2 | 182 | 273 | 187 | 387 | 1029 |
| 3-4 | 147 | 358 | 413 | 212 | 1130 |
| 5-6 | 65 | 195 | 154 | 157 | 571 |
| 7-8 | 23 | 101 | 98 | 88 | 319 |
| 9+ | 10 | 67 | 123 | 11 | 211 |
| TOTAL | 492 | 1018 | 1008 | 952 | 3470 |

A major credit card company is trying to analyze data on their customers in Raleigh, NC that describes how many credit cards a person owns depending on their age. Table 2.1 displays the results. From these results determine the probability of the following:

- A person is between the age of 20 and 29 **and** owns 3-4 credit cards

- A person is between the age of 20 and 29 **or** owns 3-4 credit cards

- A person owns 5-6 credit cards

- A person owns at least one credit card

- A person owns 1-2 credit cards given they are between the age of 30 and 39

- A person is above the age of 40 given they own 9 or more credit cards

# DISTRIBUTIONS

## PROBABILITY DISTRIBUTION

Distributions are a collection of ordered data values with how often each data value occurs with respect to all the others. When we typically think of data, we consider each of the variables in a data set to be a random variable. Random variables are variables that assign a numerical value to each outcome of a random experiment or trial. Probability distributions describe the distribution of probabilities for all possible outcomes of the random variable of interest.

- Two Types of Random Variables:
  - Discrete:
  - Continuous:

The total area under the probability distribution curve is equal to one. There are many different probability distributions for each type of random variable. Two of the most common discrete variable distributions are the Binomial and the Poisson distributions. Two of the most common continuous variable distributions are the Normal and Exponential Distributions. Before giving details of these distributions, we must learn how distributions are typically described.

## DESCRIBING DISTRIBUTIONS

Three common ways of numerically summarizing distributions are measures of location (central tendency), spread (variation), and shape (skewness and kurtosis). Where the data comes from is important when measuring center/location, spread, and shape. Remember the difference between parameters and statistics. A parameter is a measure computed entirely from the population and is usually assumed that the value is unknown and does not change. A statistic is a measure computed entirely from the sample and therefore changes between samples.

### MEASURES OF CENTER/LOCATION

There are a variety of ways to measure the center of a distribution or different location measures. First we will focus on measures of center:

- **Mean:**
  - Population Mean:

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

- Sample Mean:

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

- **Median:** the value in the middle of the data when arranged in ascending order.
  - For odd numbered observations, it is the observation that has an equal number of observations above and below it.
  - For even numbered observations, it is the average of the two observations in the middle (equal number of observations above and below them).

Location is not only determined by the center of a distribution. Sometimes, you desire to know not only where the center of a distribution is, but also where points along the distribution are. These are measures of location:

- **Percentiles:** the $p^{th}$ percentile in a collection of ordered data is the value such that $p$% of the data falls at or below this value and (100-$p$)% of the data are above this value.
  - **Quantiles** are the same thing as percentiles, except quantiles are between 0 and 1 (so 93$^{rd}$ percentile is the 0.93 quantile).
  - **Quartiles:** special (named) percentiles where the first quartile, $Q_1$, has $p = 25$ and the third quartile, $Q_3$, has $p = 75$. The second quartile, $Q_2$, is also called the median.

## MEASURES OF SPREAD

As set of data has spread (or variation) if all of the data are not the same value. There are a variety of measures of spread. The most common are listed here:

- **Range:** the difference between the maximum and minimum value in the data set.

- **Interquartile Range:** the difference between the third and first quartile.
- **Variance:** the "average" of the squared distances between each data value and the mean.
  - Population Variance:

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

  - Sample Variance:

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

- **Standard Deviation:** the square root of the variance.
    - Population Standard Deviation:

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

    - Sample Standard Deviation:

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

The variance and standard deviation possess two common characteristics. First, if the value of the variance or standard deviation equals zero, then all of the data in the data set has the same value. Second, all measures of spread are positive (or nonnegative if zero spread) in value.

## MEASURES OF SHAPE

When looking at the shape of a distribution, the skewness (non-symmetry) and kurtosis (thickness of tails) of a distribution are typical descriptive terms.

- Symmetric vs. Skewed Data: symmetric data sets have values evenly spread around the center, while skewed data sets are not symmetric.
- Sample Skewness:

$$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

- Mean vs. Median:
    - Right-Skewed Data: Mean > Median

    - Left-Skewed Data: Mean < Median

- Sample Kurtosis: deals with the thickness of the tails of a distribution.

$$g_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

- In Figure 3, three distributions are shown with the three different distinctions of kurtosis – mesokurtic, platykurtic, and leptokurtic.
    - Mesokurtic (ex: Normal distribution): $g_2 = 3$
    - Platykurtic (ex: Raised Cosine distribution): $g_2 < 3$
    - Leptokurtic (ex: Double Exponential distribution): $g_2 > 3$

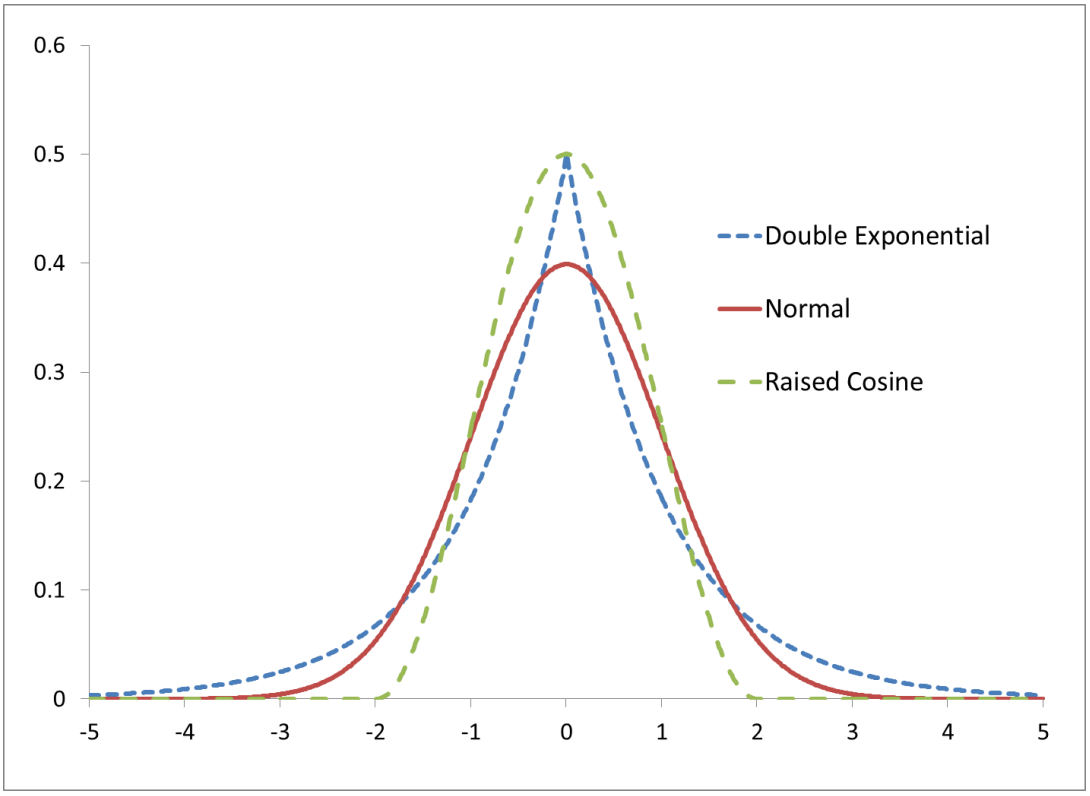



**Figure 3: Comparison of Kurtosis Distinctions**

To get a baseline measure of zero instead of three, sometimes people measure **excess kurtosis** instead of kurtosis. The equation for excess kurtosis is the following:

$$g_2^* = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4\right) - 3$$

## OUTLIER EFFECTS

Outliers are data points that deviate from the common pattern or grouping of the majority of the data. A few of these anomalous observations lead to large changes in the values of certain numerical summaries of data. We are going to focus on the mean, median, and variance (or standard deviation). When outliers are present in a data set, these numerical summaries have the following changes:

- **Mean:** the value of the mean either increases or decreases depending on the whether the outlier is larger than the data or smaller than the center of the data.
- **Median:** since the median is calculated from the center points of the data, outliers do not dramatically change the value if at all.
- **Variance (or Standard Deviation):** all outliers will increase the variance and standard deviation of the data.
- **Skewness:** the value of the skewness coefficient will increase or decrease depending on whether the outlier is larger than or smaller than the center of the data.
- **Kurtosis:** all outliers will increase the kurtosis coefficient (and excess kurtosis) of the data.

EXAMPLE

A marketing firm collected data on annual household incomes for Outland, NC. They surveyed all 182 households in the small town of Outland. Use the file *Income Outlier Data* on the class website to calculate the mean, median, standard deviation, skewness, and kurtosis of the household incomes of Outland, NC. Now imagine that one of the marketing firm's vice presidents has decided to move to a more rural setting and picks Outland, NC. Now the data set will have 183 households. The vice president's household annual salary is $348,000. Add the new household to the data set and calculate the new mean, median, standard deviation, skewness, and kurtosis. Describe your results for the following:

- Mean:
- Median:
- Standard Deviation:
- Skewness:
- Kurtosis:

## DISCRETE DISTRIBUTIONS

Discrete data are data whose possible values are countable, while continuous data has values that are uncountable and may assume any value in an interval. Continuous data may be transformed into discrete data by binning. Discrete data may not be transformed into continuous data, except in extreme situations. Frequency distributions describe discrete data. A frequency distribution is just a summary of data that displays the number of observations that belong to each category in the data.

There is a difference between the frequency of a category and the relative frequency of a category.

- Frequency:

- Relative Frequency:

Frequencies and relative frequencies may also be converted into cumulative frequencies and cumulative relative frequencies.

- Cumulative Frequency:

- Cumulative Relative Frequency:

All of these may be plotted in frequency (or relative frequency) histograms.

EXAMPLE

A financial firm has researched 241 mutual funds they offer to customers to invest in. They calculated the annual return over the last 3 years for each mutual fund. Fill in the columns for relative frequency, cumulative frequency, and cumulative relative frequency. If an investor randomly selected one of these mutual funds to invest in, what is the probability their mutual fund had a negative annual return? How about an annual return above 5%?

**Table 2: Mutual Fund Returns Example**

| Annual Return | Frequency | Relative Frequency | Cumulative Frequency | Cumulative Relative Frequency |
|---|---|---|---|---|
| -2% < -1% | 10 | | | |
| -1% < 0% | 18 | | | |
| 0% < 1% | 23 | | | |
| 1% < 2% | 44 | | | |
| 2% < 3% | 56 | | | |
| 3% < 4% | 39 | | | |
| 4% < 5% | 27 | | | |
| 5% < 6% | 12 | | | |
| 6% < 7% | 9 | | | |
| 7% < 8% | 3 | | | |

## CONTINUOUS DISTRIBUTIONS

For discrete data, probability distributions are composed of the probabilities of each of the possible values of the data. These probabilities are typically displayed in histograms where the sum of the individual rectangles in the histogram is equal to one. In continuous probability distributions, the number of possible values is uncountable and therefore described by a probability curve instead of a histogram. Figure 4 shows this property.

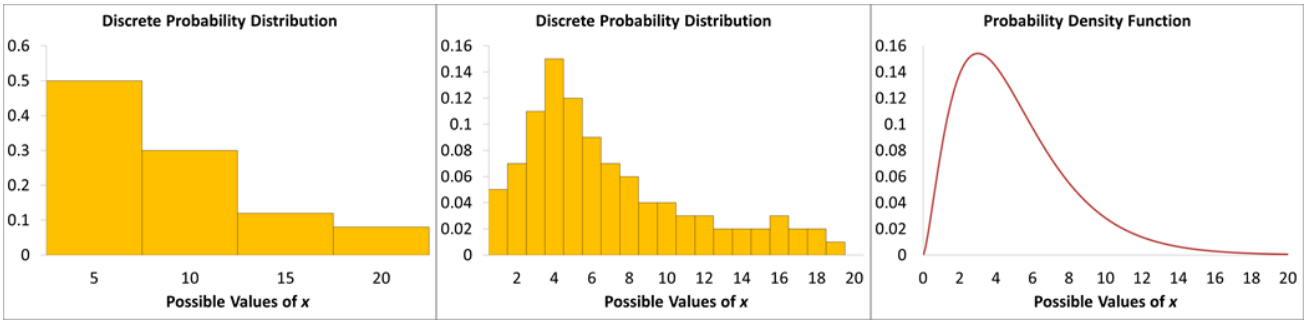



**Figure 4: Comparing Discrete to Continuous Distributions**

One of the most common continuous distributions is the Normal Distribution.

## NORMAL DISTRIBUTION

The Normal distribution is one of the most important distributions in statistics. It is the foundation of statistical theory that makes statistical inference about populations possible. The Normal distribution is a specific continuous probability distribution. The Normal distribution is important because it oc- curs naturally throughout nature. The Normal distribution is a bell-shaped distribution with other important characteristics:

- Unimodal:
- Symmetrical: $g_1 = 0$
- Mean and median are equal
- Asymptotic to the x-axis towards $\pm\infty$
- Completely defined by mean and standard deviation:
- Excess kurtosis of 0: $g_2^* = 0$

Figure 5 displays a typical Normal distribution. The equation for the actual density curve of the Normal distribution is the following:

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

where $\mu$ is the mean of the Normal distribution and $\sigma$ is the standard deviation.

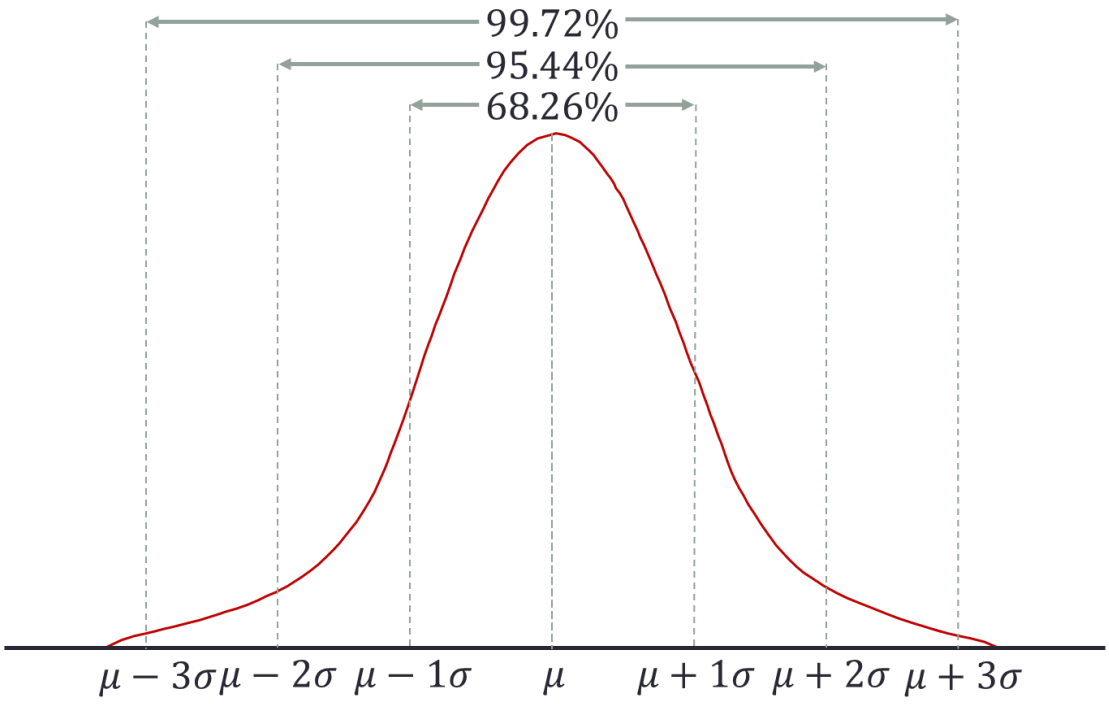



**Figure 5: Normal Distribution and the Empirical Rule**

Since the Normal distribution is perfectly symmetric, half of the observations are above and half of the observations below the mean. This implies that the probability of a Normal random variable taking a value above (or below) the mean is 0.5. This is not the only probability assumption the Normal distribution has. Since the Normal distribution is bell-shaped, the Empirical Rule (68-95-99.7 Rule) describes the Normal distribution as well.

- **Empirical Rule:** if the data is normally distributed, then the following intervals are true:
  - $\mu \pm 1\sigma$ contains approximately 68% of the values in the distribution
  - $\mu \pm 2\sigma$ contains approximately 95% of the values in the distribution
  - $\mu \pm 3\sigma$ contains approximately 99.7% of the values in the distribution

This rule is also displayed in Figure 5.

EXAMPLE

What is the approximate probability a random variable that follows a Normal distribution with $\mu = 0$ and $\sigma = 2$ ($N(0, 2)$) takes a value between -2 and 4?

EXAMPLE

Assume the weekly number of credit card transactions for customers of a major credit card follows a Normal distribution with $\mu = 8.5$ and $\sigma = 1.5$ ($N(8.5, 1.5)$). What is the probability that a randomly selected customer has between 7 and 13 transactions in a week?

### EXAMPLE

Assume the number of customers of a major television provider that churn every month follows a Normal distribution with $\mu = 1325$ and $\sigma = 568$ ($N(1325, 568)$). What is the probability that a randomly selected month had more than 2461 customers that churn?

## STANDARD NORMAL DISTRIBUTION

The Empirical Rule works well if the numbers we are trying to find are exactly a certain number of standard deviations away from the mean. This is not always the case, and most of the time never is the case. What can we do in situations such as these? With integrals, we can determine the area under the Normal density curve at any fraction of standard deviations away from the mean. Another technique is to calculate these integrals for one Normal distribution and standardize all other Normal distributions to this Standard Normal distribution.

The Standard Normal distribution has $\mu = 0$ and $\sigma = 1$. The file *Standard Normal Table* on the class website, contains the probability values for intervals under the Standard Normal distribution up to two decimal places. This table displays the probability of values 3.49 standard deviations above and below the mean. With this table, we can calculate the probability between any fractions of standard deviations away from the mean. By converting any Normal distribution to this Standard Normal distribution, we can use this table for all Normal distributions. This can be seen in Figure 6.

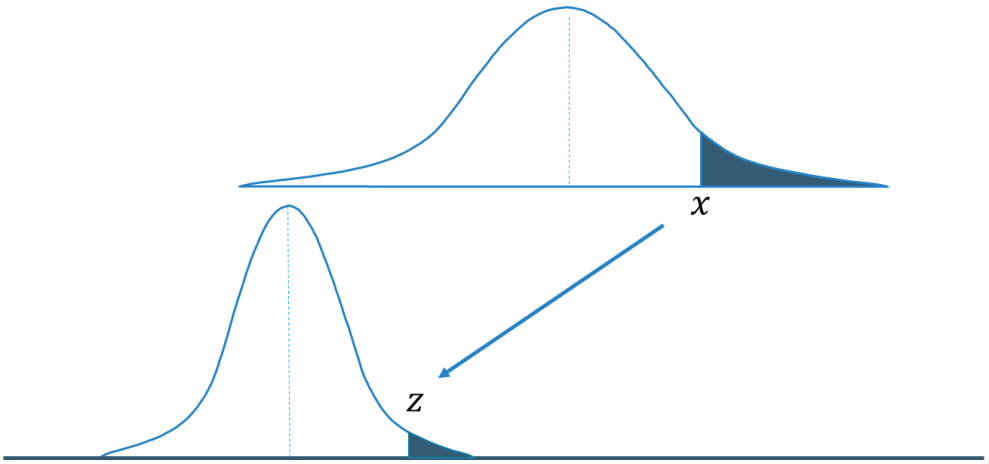



**Figure 6: Converting Normal to Standard Normal**

The horizontal axis of the Standard Normal distribution is scaled in z- values that measure the number of standard deviations a point is from the mean. The standardized Normal z value is defined as:

$$z = \frac{(x - \mu)}{\sigma}$$

Again, these z-values measure the number of standard deviations a point is from the mean. Therefore, positive z-values represent values above the mean, and negative z-values represent values below the mean.

EXAMPLE

Assuming we have a Normal distribution, find the following probabilities if the mean is 0 and standard deviation is 1:

- $P(z > 2)$
- $P(z \leq 1.12)$
- $P(-1.33 \leq z \leq 1.33)$
- $P(z \geq 3.02)$
- $P(z \leq 6.87)$
- $P(z \leq -6.87)$

EXAMPLE

Assume new employees at a company have previous years of professional experience that follow a Normal distribution, find the following probabilities if the mean is 5 and standard deviation is 2.5:

- What is the probability a new employee has more than 5 years of previous experience?
- What is the probability a new employee has less than 2 years of previous experience?
- What is the probability a new employee has between 1 and 7.5 years of previous experience?

EXAMPLE

You manage a credit card company. The weekly number of transactions your customers have follows a Normal distribution with a mean of 20.5 with a standard deviation of 9.

- What is the probability that a random customer uses their credit card more than 28 times in a week (4/day average)?

- What is the number of transactions for the lower 10% of customers?

## OTHER COMMON DISTRIBUTIONS

There are many different types of discrete and continuous distributions used for analysis in statistics. A couple of the common discrete distributions is the Binomial distribution and Poisson distribution. A few of the common continuous distributions is the Uniform distribution, Exponential distribution, *t*-distribution, *F*-distribution, and $\chi^2$-distribution. These distributions will be covered in later chapters as their need arises.

## SAMPLING DISTRIBUTIONS

Sample statistics are just point estimates of population parameters. However, because they are just estimates, they have error. Although parameters are assumed to remain constant, statistics change depending on the sample which leads to this error – called **sampling error**.

- **Sampling Error:**

Most of the time we do not have the ability to collect information from the whole population to see what kind of sampling error we actually have. For that reason, it would be nice to know if there is a common pattern/distribution for the point estimates and therefore the sampling errors of these point estimates. Although statistics calculated from samples have sampling error, statistical inference is possible because statistics have a predictable distribution called a sampling distribution.

- **Sampling Distribution:**

### SAMPLE MEANS

Intuitively, if a population is normally distributed, with a mean $\mu$ and standard deviation $\sigma$, the sampling distribution of sample means from that population are also normally distributed. However, sample means have an even more powerful property – the **Central Limit Theorem**.

- **Central Limit Theorem:** If we use a large sample ($n \geq 50$), the Central Limit Theorem (CLT) states that the sampling distribution of $\bar{x}$ is approximately normally distributed, **regardless of the population distribution.**

In small sample sizes we would still require the population to be normally distributed for the sampling distribution of sample means to be normally distributed.

- Distribution of Sample Mean:

$$N(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Since the sample mean has a Normal distribution, we can standardize this Normal distribution for analysis with the Standard Normal distribution in a similar way as before. The standardized z-value will have a similar form to the previous value where we subtract the mean of the distribution (now $\mu_{\bar{x}}$) and divide by the standard deviation of the distribution (now $\sigma_{\bar{x}}$):

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

Now we can answer similar questions as before, but with sample means instead.

EXAMPLE

Assume the average number of daily web page hits a company gets follows a Normal distribution with mean 2341.36 and standard deviation 516.79.

- What is the probability that a sample of 49 days over the past year has an average web page hit above 2500? How about a sample of 121 days?

- What if the distribution of web page hits was not a Normal distribution?

EXAMPLE

Assume that I own chain of retail stores located at major cities across the country. The daily sales in thousands of dollars at each of my stores has a mean of 17.06 with a standard deviation of 5.12.

- What is the probability that a sample of 64 of my stores averages sales of more than $19K in a day?

I am worried about one of my regional manager's performance in daily sales. He manages 100 of my stores. They only average $14.35K in sales per day.

- What is the probability I randomly select 100 of my stores and get sales numbers this low?

## SAMPLE PROPORTIONS

There are more statistics than the sample mean that are commonly measured. One of the other popular measures of a population or sample is the proportion of individuals/objects that fall within a certain category.

- **Population Proportion:**

$$p = \frac{x}{N}$$

- **Sample Proportion:**

$$\hat{p} = \frac{x}{n}$$

The Central Limit Theorem holds for sample means. This implies that it would also hold for proportions as well. Proportions are just the average number of successes in a sample with successes labeled with a value of 1 and failure labeled as a value of 0. Therefore, the Central Limit Theorem holds for sample proportions with the following:

$$\mu_{\hat{p}} = p \qquad \text{and} \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Since the sample proportion has a Normal distribution, we can standardize this Normal distribution for analysis with the Standard Normal distribution in a similar way as with the sample mean. The standardized z-value will have a similar form to the previous value where we subtract the mean of the distribution (now $\mu_{\hat{p}}$) and divide by the standard deviation of the distribution (now $\sigma_{\hat{p}}$):

$$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Now we can answer similar questions as before, but with sample proportions instead.

EXAMPLE

The NC State Board of Education is interested in gathering information about the drop-out rate of high school students across the state of North Carolina. The proportion of high school students that drop out is 5.24% across the state.

- What is the probability that less than 8 out of 169 random students across the state drop out of high school?

# CONFIDENCE INTERVALS

## POINT AND INTERVAL ESTIMATION

We have previously discussed how to appropriately draw samples from a population and briefly analyze data from these samples. Statistical inference is possible because statistics calculated from samples follow predictable patterns called sampling distributions. These sampling distributions bridge the gap between statistics and parameters. Sample statistics are a point estimation of population parameters.

- **Unbiased Point Estimator:**

Every sample produces a different statistic and therefore a different estimate of the population parameter. Up until this point, the population parameters $\mu$, $\sigma$, and $p$ have been assumed known to calculate probabilities about samples. Now this assumption of knowing population parameters will be relaxed.

Since the true accuracy of a point estimate is unknown without knowledge of the population parameter, providing a margin of error around a point estimate reveals more information about the potential accuracy. Point estimates with margins of error become interval estimates.

- **Confidence Interval:**
    - **Confident:**
    - **Confidence Interval Form:**
        - **Margin of Error:**
        - **Standard Error:**

These confidence intervals depend on the point estimate from which they are calculated. Even with the same sample size and level of confidence, every sample will provide a different confidence interval because of the point estimate. Confidence intervals are random values that depend on the sample. Therefore, confidence intervals may not contain the true population parameter.

Misinterpretation of confidence intervals often occurs. An interpretation of a 95% (typically called the $1-\alpha$ level) confidence interval is not that there is a 0.95 probability that the population parameter is contained in

the specific confidence interval from that sample. The population parameter is assumed to be fixed. Constant values cannot have a randomness or probability associated with them. A confidence interval either does or does not contain the population parameter. A proper interpretation would be if all possible confidence intervals are calculated from the population, 95% of these intervals would contain the population parameter. This implies that the confidence is in the procedure, not the exact interval.

## CONFIDENCE INTERVAL FOR PROPORTION

The sampling distribution of the sample proportion $\hat{p}$ helps determine the margin of error around a specific point estimate.

- Sampling Distribution for Sample Proportion:

- Confidence Interval for Sample Proportion:

This confidence interval is a large sample confidence interval. Remember that the Central Limit Theorem only holds for sufficiently large sample sizes. The Normal distribution approximation is a good approximation if $n \times \hat{p} \geq 5$ and $n \times (1 - \hat{p}) \geq 5$. Figure 7 depicts the visual representation of the confidence interval around the sample proportion.

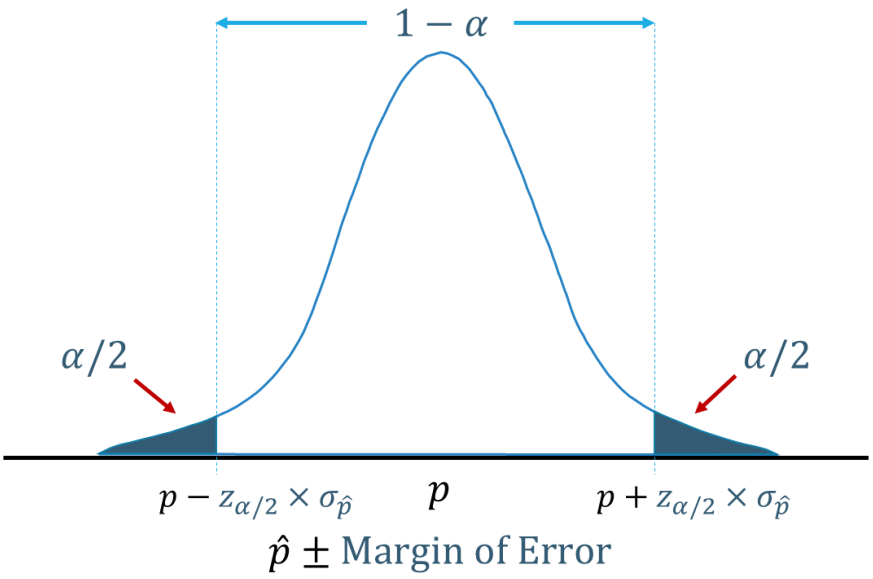



**Figure 7: Confidence Interval for Sample Proportion**

The confidence interval also changes depending on the sample size or the confidence level:

- Change of sample size:

- Change of confidence level:

EXAMPLE

An electronics manufacturer provides a full warranty on a certain type of tele- vision they make. The company will replace the television if any problems occur in the first year of use. The manager in charge of the warranty division wants to determine the proportion of warranties that are claimed. The manager samples 150 customer records and found that 17 of the customers used their warranty.

- Create a 95% confidence interval for the estimate of the proportion of customers who use their warranties.

CONFIDENCE INTERVAL FOR MEAN

The sampling distribution of the sample mean $\bar{x}$ helps determine the margin of error around a specific point estimate. Unlike the sample proportion, the sample mean requires more estimation of population parameters. The standard deviation of the sample mean is calculated from the population standard deviation, which is typically unknown. Therefore, we must estimate this with the sample standard deviation, s. However, we have added extra error into our calculations since estimating two statistics has more error than just estimating one. The normal distribution is no longer a good approximation for the sampling distribution of $\bar{x}$ because it doesn't account for this extra error. Instead we must use the *t*-distribution.

The t distribution is a family of similar probability distributions. They are symmetric, but have thicker tails than the Normal distribution. The t distribution has degrees of freedom as it only parameter as compared to the mean and standard deviation from a Normal distribution. Degrees of freedom are the number of independent pieces of information that go into the computation of s. The more degrees of freedom leads to less dispersion in the distribution. As the degrees of freedom increases, the *t*-distribution becomes approximately more and more like the standard Normal distribution as seen in Figure 8.

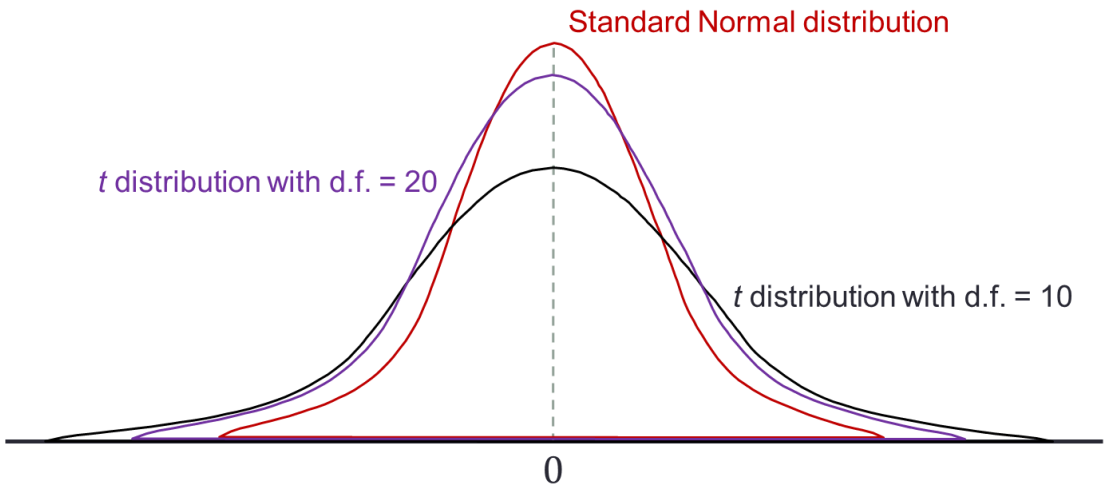



**Figure 8: Comparing Normal to t-Distributions**

For confidence intervals for sample means, the degrees of freedom for the $t$-distribution equals $n - 1$. This will change for other calculations later in the course.

- **Confidence Interval for Sample Mean:**

Figure 9 depicts the visual representation of the confidence interval around the sample proportion.

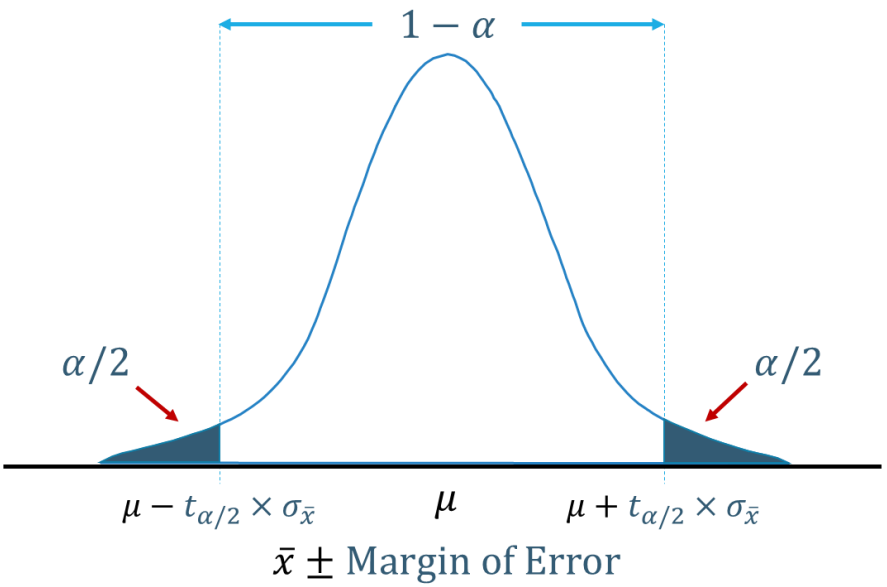



**Figure 9: Confidence Interval for Sample Mean**

This confidence interval is both a large sample and small sample confidence interval. Different size samples may be used depending on the population's distribution. The Central Limit Theorem holds for large sample sizes. However, if the population follows a Normal distribution, any sample size may be used.

The confidence interval also changes depending on the sample size or the confidence level.

- Change of Sample Size:

- Change of Confidence Level:

EXAMPLE

Assume a sample of stores has daily sales in thousands of dollars with a sample mean 17.06 and sample standard deviation 5.12.

- What additional assumption is needed if the analyst who collected this sample wants to create a 90% CI with a sample size of 20 stores? Calculate this interval.

- Would this assumption be needed if the sample size was 100? Calculate this interval.

## SAMPLE SIZE CALCULATION

We have discussed the trade-offs between sample size and margin of error of confidence intervals. Under the same confidence level, reducing margin of error requires increased sample size. The question now is how much bigger of a sample is needed for the desired margin of error. In ideal situations we take as large of a sample as possible. However, analysts may save money if they only take the sample that obtains the desired margin of error instead of the biggest sample possible.

The calculations for the desired sample size depends on the statistic of interest.

- Proportion Sample Size:

$$n = \frac{z^2 \times 0.5 \times (1 - 0.5)}{M.O.E.^2}$$

  - Since the true value of $p$ is not known, we must use an estimate. However, since we are using this calculation to determine the size of the sample we will take, we cannot use $\hat{p}$ because our actual sample has not yet been taken. The most conservative and preferred estimate is 0.5.

- Mean Sample Size:

$$n = \frac{z^2 \hat{\sigma}^2}{M.O.E.^2}$$

  - Since the true value of $\sigma$ is not known, we must use an estimate $\hat{\sigma}$. However, since we are using this calculation to determine the size of the sample we will take, we cannot use s. This estimate may come from a pilot sample (or can get a really rough estimate of standard deviation by using estimated range ÷ 4).

## HYPOTHESIS TESTING

We use confidence intervals to estimate population parameters. Hypothesis testing uses evidence from data to test some claim about the population parameter. This is the same thought process to probabilities and z-values that we previously discussed. However, hypothesis testing relaxes the assumption that we know the population parameters $\mu$, $\sigma$, and $p$.

### STEPS OF HYPOTHESIS TESTING

The main idea behind hypothesis testing is that enough evidence against a claim allows us to consider that claim is not true. Imagine if you had a coin that you believed was fair. You flip the coin 5 times and get heads each time. Is this a possible occurrence? Yes, it is possible to flip a fair coin 5 times and get 5 straight heads. Is this a probable occurrence? The probability of this happening is only 0.03125 or 3.125% chance. Do you still believe the coin is fair? This is the basis for hypothesis testing.

There are five main steps to hypothesis testing:

1. State Hypothesis
2. Test Statistic
3. P-value
4. Decision Rule
5. Conclusion

The following sections addresses each of these steps one at a time.

#### STATE HYPOTHESIS

In hypothesis testing, there are two hypotheses to state. The two hypotheses are the **null hypothesis** and **alternative hypothesis**.

- **Null Hypothesis:**

- **Alternative Hypothesis:**

All hypotheses are stated about the population parameters because they are the unknown piece. Statistics are known values from a sample. The population parameter is typically unknown, which is why we are typically interested in testing it.

There are two different types of alternative hypotheses depending on the tests you are interested in. These are either **one sided** or **two sided** tests. One-sided tests allow for a possibility to only occur on one side of the distribution as you can see in Figure 10.

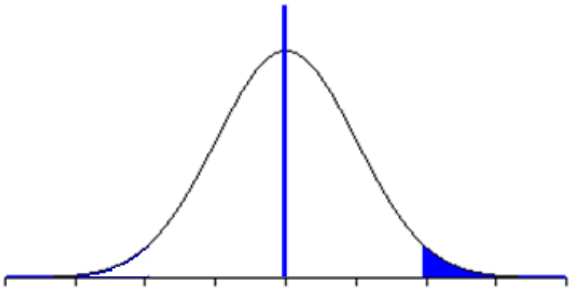


**Figure 10: Example of One-Sided Hypothesis Tests**

However, a two-sided test allows for possibilities in either side of the distribution as you can see in Figure 11.

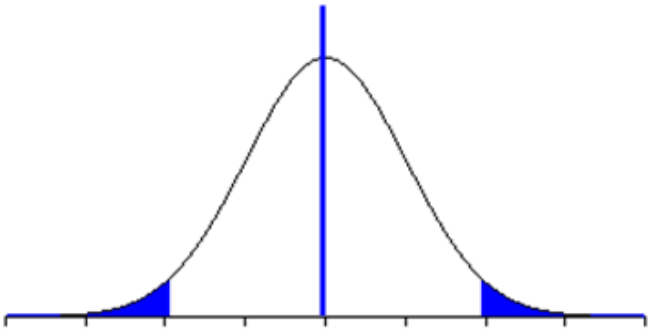


**Figure 11: Example of Two-Sided Hypothesis Test**

## TEST STATISTIC

The test statistic portion of the hypothesis test is similar to evidence in a court case. The test statistic summarizes the amount of information pro- vided in the sample that supports the null hypothesis. Test statistics have a common form they all follow:

- Form of a Test Statistic:

- Test Statistic for Proportion:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- Test Statistic for Mean:

$$t = \frac{\bar{x} - \mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$$

### P-VALUE & DECISION RULE

Once we determine the test statistic, we must calculate the probability that we got the information in our sample **provided that the null hypothesis is true**. Until we obtain information to states that the null hypothesis is no longer true, we must assume that the null hypothesis is true. This probability that we calculate is called the **p-value**.

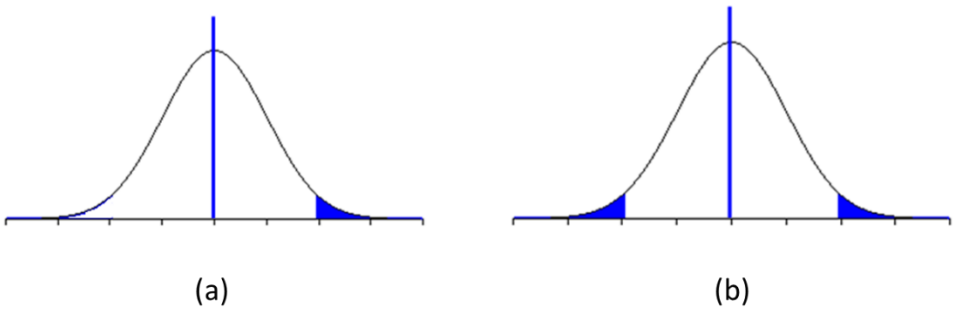



**Figure 12: One-Sided vs. Two-Sided P-value**

If the p-value is low, this implies that the sample we obtained from the population is extremely rare if we assume that the null hypothesis is true. Figure 12(a) displays the p-value for a one-sided test. Figure 12(b) displays the p-value for a two-sided test.

The question now becomes, how low is low enough of a p-value where we no longer believe the null hypothesis. This is defined by the **significance level**, $\alpha$. If the p-value is smaller than $\alpha$, then you **reject the null hypothesis**. If the p-value is larger than $\alpha$, then you **do not reject the null hypothesis**.

Of course, there is a chance that we made the incorrect choice when rejecting the null hypothesis. Just like with the coin example at the beginning of this chapter, just because something is improbable doesn't mean that it is impossible. The significance level is the amount of error you are comfortable with making. In other words, what percentage (or probability) of the time would you reject the null hypothesis incorrectly. This is referred to a **Type I error**. The opposite mistake is a **Type II error**.

- Type I Error:

- Type II Error:

Figure 13 helps to better see where the Type I and Type II errors occur.

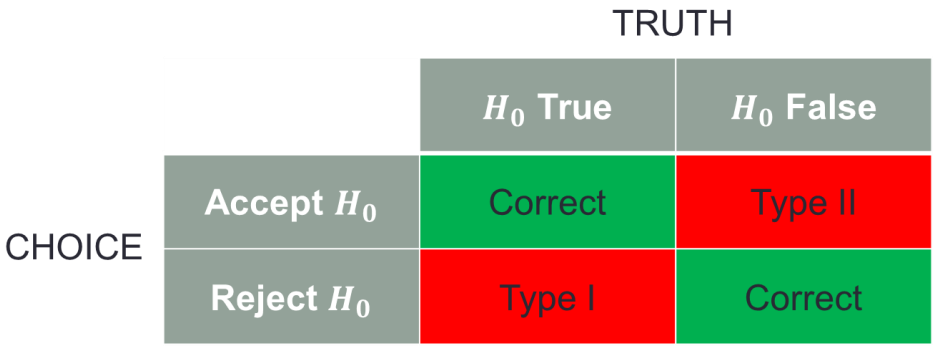



**Figure 13: Type I vs. Type II Errors**

### CONCLUSION

Most people forget this step, but it is the step that ties the entire problem together. It doesn't really make any sense for us to tell someone that the test statistic from the data provided a p-value which was low enough to reject our original null hypothesis. I could say that statement for any problem that rejects the null hypothesis because it really doesn't tell me anything about what I am trying to find out. The important aspect of any conclusion for statistical analysis is to put the problem in real world terminology so anyone without a statistical background can understand it.

### EXAMPLE (MEAN)

You work for a business school as an analyst. The dean of the business school just went on record saying that students who just graduated average at least $3000 per month in salary. You took a sample of 12 people and assume that salaries follow a Normal distribution. The average monthly salary of these individuals was $2,940, with a standard deviation of $165.70.

- With a significance level of 0.05, conduct a hypothesis test on this claim.

- Now assume the dean said they make on average $3000 per month (not necessarily more). With a significance level of 0.05, conduct a hypothesis test on this claim.

### EXAMPLE (PROPORTION)

You are interested in hair color and eye color across 2 different regions of the country. You do not believe that less than 32% of people have blue eyes. You have a sample of 762 people with 222 people in that sample having blue eyes.

- With a significance level of 0.05, conduct a hypothesis test on this claim.

## COMPARING HYPOTHESIS TESTS TO CONFIDENCE INTERVALS

Under certain conditions, hypothesis testing and confidence intervals are con- ducting the same test. Figure 14 and Figure 15 show this concept. The distribution centered with the solid line shows the $\alpha = 0.05$ hypothesis test rejection region. The distribution centered with the dashed line shows the 95% confidence interval around the sample statistic.

In Figure 14, if the null hypothesis (the dark solid line in the first distribution) is in the confidence interval of the sample mean (the lightly shaded region in the second distribution), then the sample mean is not in the rejection region of the hypothesis test (the dark shaded region in the tails of the first distribution). Therefore, you do not reject the null hypothesis in the hypothesis test.

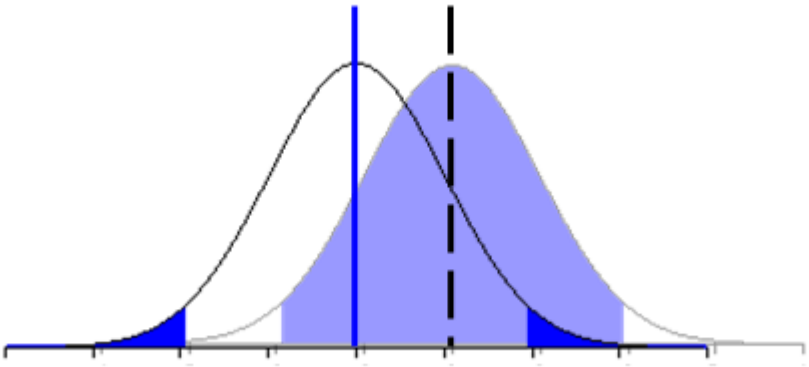


Figure 14: Comparing HT to CI without Rejection

In Figure 15, if the null hypothesis (the dark solid line in the first distribution) is not in the confidence interval of the sample mean (the lightly shaded region in the second distribution), then the sample mean is in the rejection region of the hypothesis test (the dark shaded region in the tails of the first distribution). Therefore, you do reject the null hypothesis in the hypothesis test.

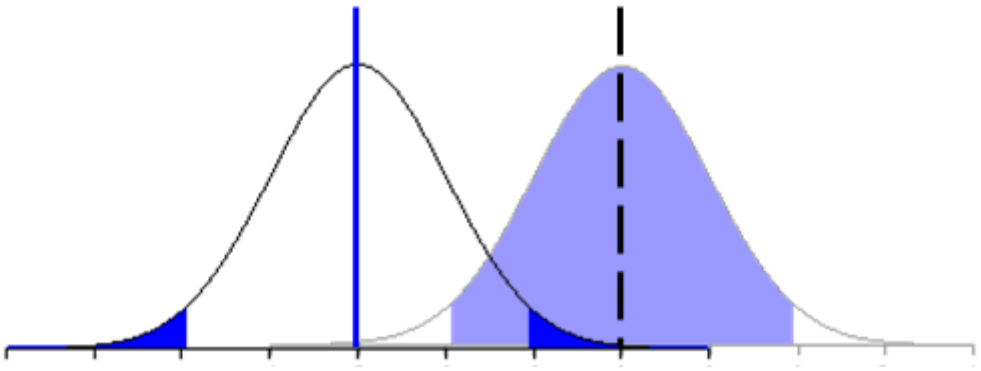


Figure 15: Comparing HT to CI with Rejection

There are two conditions to allow these comparisons.

1. The hypothesis test is a two-sided test.
2. $C = (1 - \alpha)$

## CORRELATION AND LINEAR REGRESSION

Instead of trying to figure out probabilities of events happening or predicting an unknown parameter, more often analysts are tasked to explain possible relationships between two or more variables. Methods exist to try and predict one variable with another set of variables.

When explaining relationships between variables, there are two different types of variables – variables we are trying to explain and variables we are using to explain.

- **Dependent/Target/Response Variable:**

- **Independent/Explanatory/Predictor Variable:**

It is always beneficial to graph pairs of variables in a scatter plot to determine their relationship. Scatter plots reveal the relationship between variables. The possible relationships we will discuss are linear, nonlinear, and no relationship.

- **Linear Relationship:**

- **Nonlinear Relationship:**

Both linear and nonlinear relationships can be positive or negative.

- **Positive Relationship:**

- **Negative Relationship:**

Figure 16 shows an example of a positive linear (a), negative linear (b), positive nonlinear (c), negative nonlinear (d), and no relationship (e and f).

Now we must learn how to quantify these relationships.

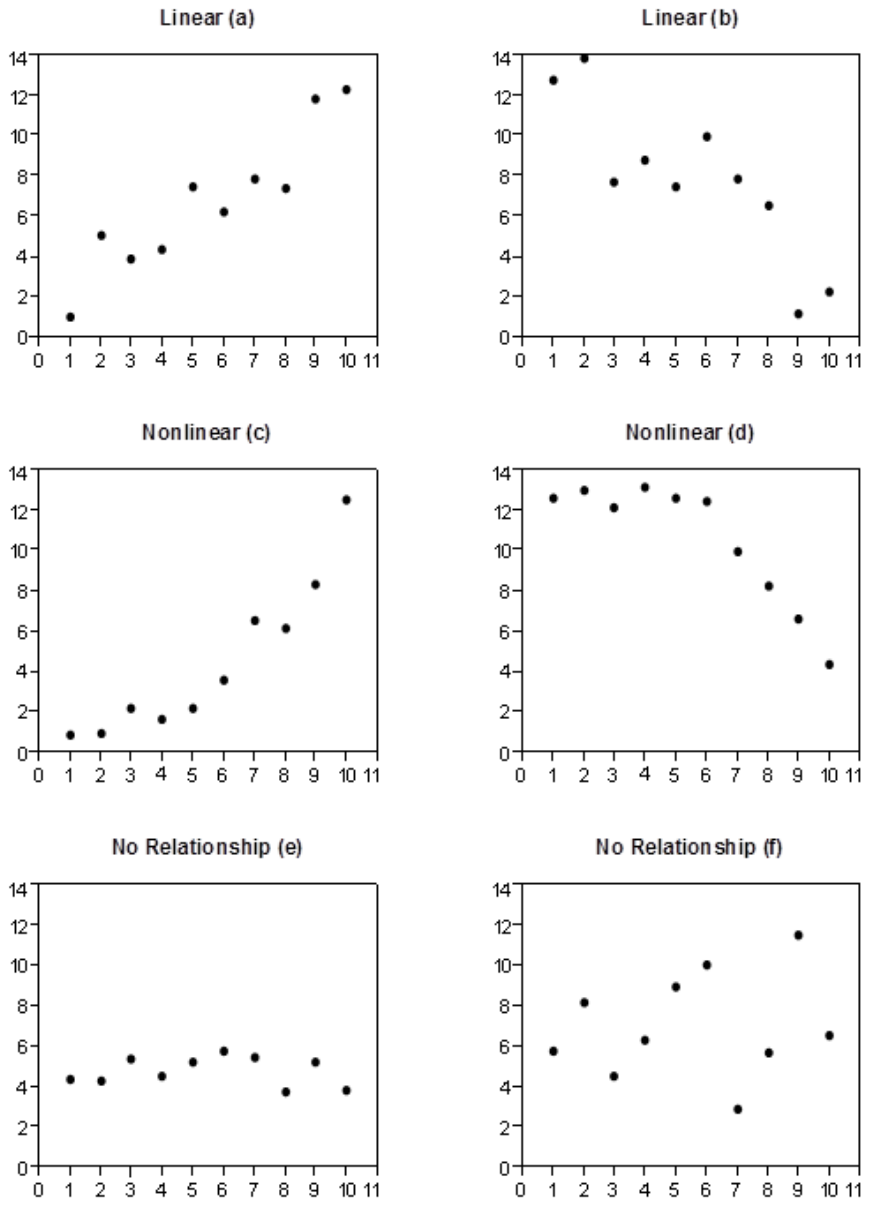



**Figure 16: Visual Examples of Relationships Between Two Variables**

## CORRELATION

We will focus on linear relationships for this course. Not all positive (or negative) relationships have the same strength of association. The **Pearson correlation coefficient**, r, is a quantitative measure that summarizes the direction and strength of the **linear** relationship between two variables.

- (Sample) Pearson Correlation Coefficient:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- 3 Properties:
    1.

    2.

    3.

These correlations are just combinations of sample z-values. This means that both variables are standardized and their units of measurement play no role in the calculation of the correlation. The correlation remains the same if either variable is converted into another unit of measure.

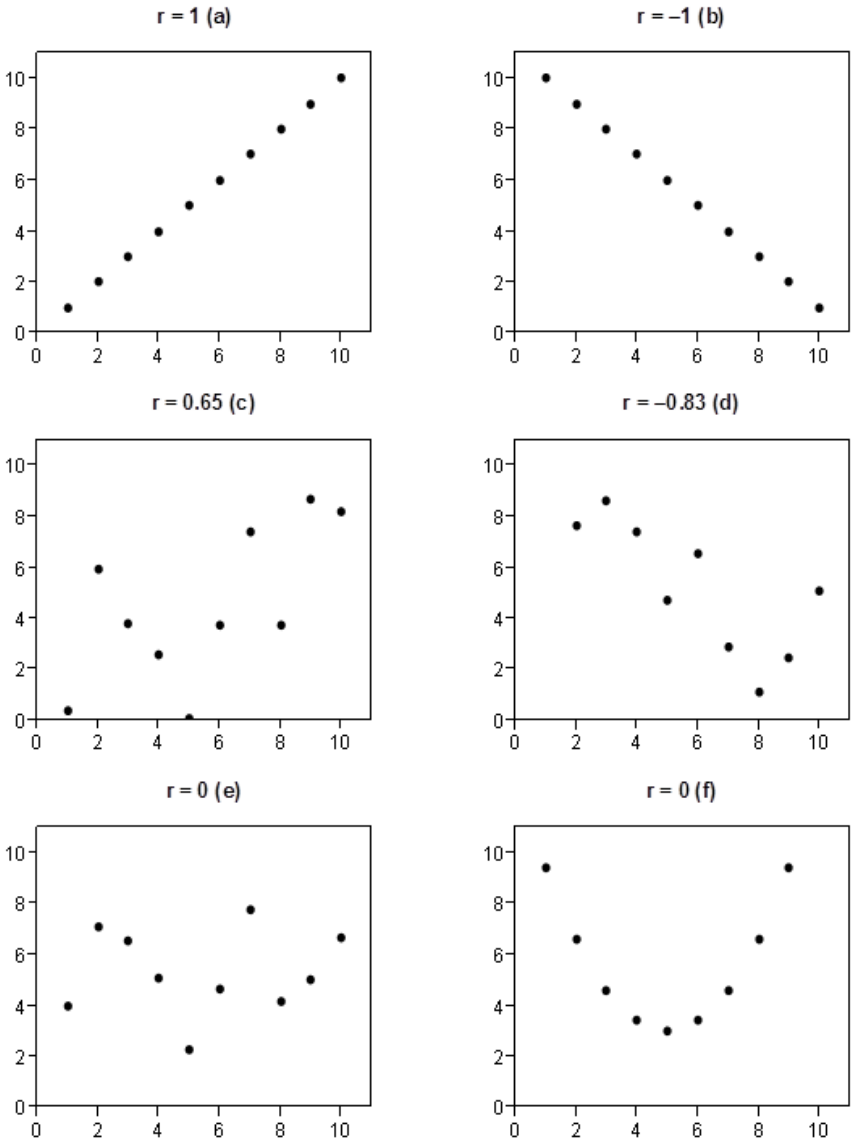



**Figure 17: Visual Examples of Correlation**

Figure 17 shows an example of a perfect positive (a) and negative (b) linear relationship. It also shows a positive (c) and negative (d) linear relationship that is not perfect, but still significant. Figure 17(e) shows the correlation of zero for two variables that have no relationship. However, correlation is only a measure of linear relationships. Figure 17(f) shows another example where two variables have a correlation of zero but with an obvious nonlinear relationship. A Pearson correlation of zero **does not** imply there is no relationship between the variables, just no **linear** relationship. It is good practice to always graph a scatter plot of two variables to understand their relationship because only looking at the correlation value does not reveal everything.

## POTENTIAL ISSUES WITH CORRELATION

Similar to other quantitative values of analysis, there are potential issues in the calculation and interpretation of correlation. Two common mistakes deal with outliers and causation.

Outliers may affect the value of the sample correlation coefficient and lead to false conclusions about correlation if you don't visualize the data to help us see what might be going on. Some outliers may inflate the correlation coefficient and make a false relationship as seen in Figure 18:

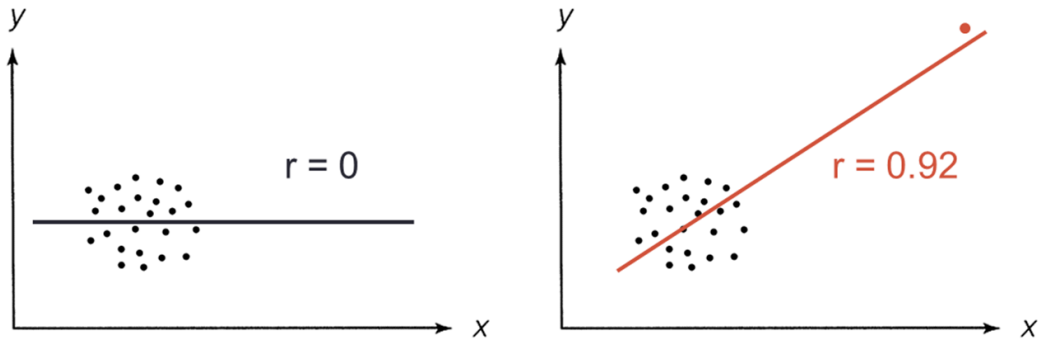



**Figure 18: Outlier Inflating Correlation**

While others outliers may deflate the correlation coefficient and hide a relationship as seen in Figure 19.

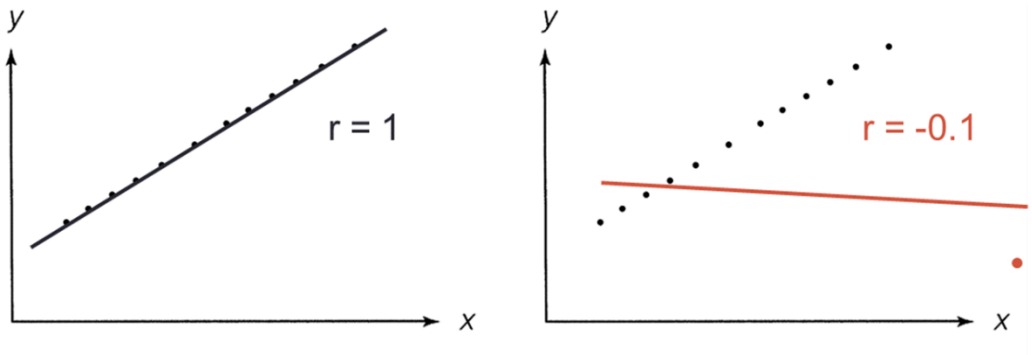



**Figure 19: Outlier Deflating Correlation**

Another common mistake with correlations is spurious correlation. Spurious correlation is a correlation between two otherwise unrelated variables. In other words, correlation does not imply causation. Even

though there might be significant correlation between variables that does not imply that there is a causal relationship between the variables. Here are some classic examples:

- As ice cream sales increase, the rate of sharp attacks increases sharply. Therefore, ice cream causes shark attacks.
- With a decrease in the number of pirates, there has been an increase in global warming over the same period. Therefore, global warming is caused by a lack of pirates.

## SIMPLE LINEAR REGRESSION

In regression we focus on assessing the significance of the independent variables in explaining the variability or behavior of the dependent variable. We try predicting the values of the dependent variable with given values of the independent variables. In the previous section, we learned that sample correlation tries to explain the linear relationship between two variables. Simple linear regression tries to estimate this linear relationship between the two variables. Later we will extend simple linear regression to include more than one independent variable.

- **Population Simple Linear Regression Model:**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

This simple linear regression model can be visualized in Figure 20.

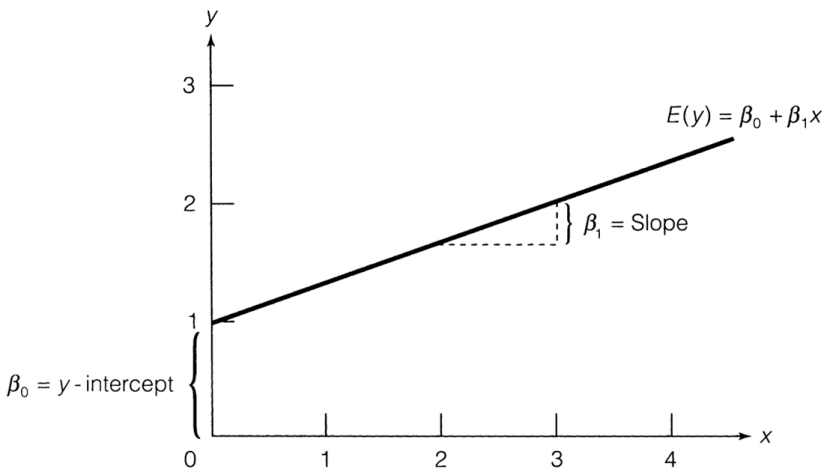



**Figure 20: Simple Linear Regression Model**

The interpretation of the slope and intercept coefficients in the simple regression model are similar to those of any line.

- The intercept coefficient indicates the mean value of $y$ when $x = 0$.
- The slope coefficient measures the change in the average value of $y$ for each single unit change in $x$.

The slope coefficient helps explain how the change in the independent variable affects the dependent variable.

The existence of the population simple linear regression model depends on four main assumptions:

1. Linearity of the Mean:

2. Normality of Errors:

3. Equal Variance of Errors:

4. Independence of Errors:

These assumptions are summarized visually in Figure 21.

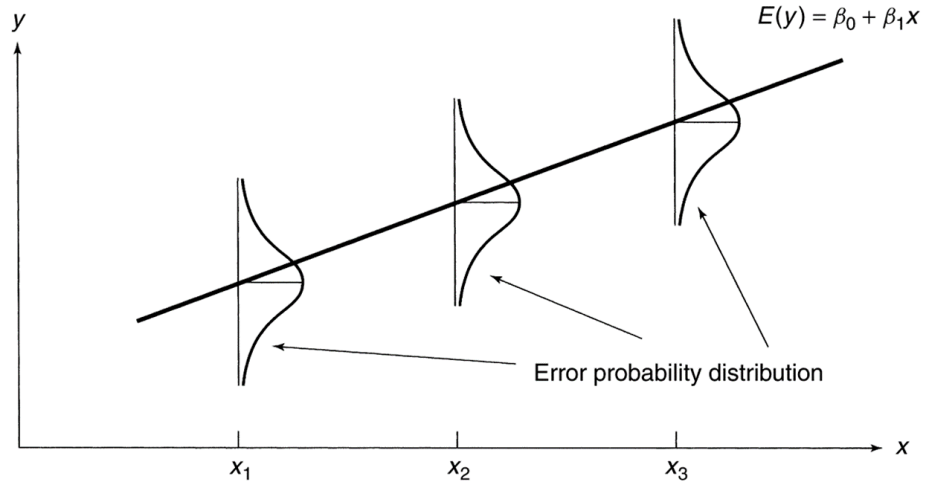



**Figure 21: Assumptions of Linear Regression**

## LEAST SQUARES METHOD

The population regression model is a parameter model. We do not know the true values of the coefficients in this model and can only estimate them. We estimate the true regression model by estimating a regression line through a sample of data from the population. However, we must determine which regression line through the data is the best regression line. A good approach to finding the best regression line is to create a line where the errors in the regression line are minimized. A regression line is a line of predicted values of $y$. Since the values of $y$ are not necessarily in a perfectly straight line, the regression line will have some prediction errors called **residuals**.

- **Residual:**

Figure 22 visually shows the calculation of two residuals from a regression model. The criterion that we will use to decide the best regression line is the **least squares criterion**. The least squares criterion is a criterion for determining the regression line that minimizes the sum of the squared residuals.

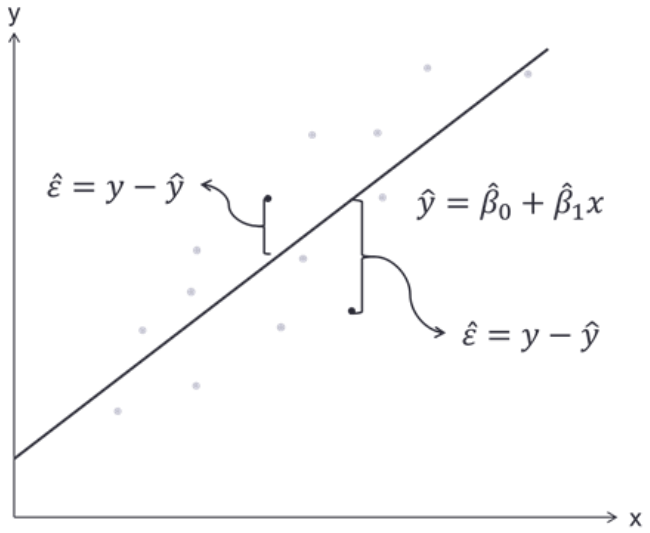



**Figure 22: Visualizing Residuals of Simple Linear Regression**

It can be shown that there is only one line for which the SSE is minimized. This line is called the **sample simple linear regression line** (or **line of best fit**).

- Sample Simple Linear Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  - Sample Slope Calculation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  - Relationship Between Slope and Correlation:

$$\hat{\beta}_1 = r \times \frac{s_y}{s_x}$$

  - Sample Intercept Calculation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Computers are usually used to calculate these coefficients because their calculations are potentially burdensome. The least squares regression line has important properties.

- Sum of the residuals equals zero.
- Sum of squared residuals is minimized.

- Simple linear regression line passes through the point $(\bar{x}, \bar{y})$.
- Coefficient estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates of $\beta_0$ and $\beta_1$ respectively.

## COEFFICIENT OF DETERMINATION

Now that the regression line is calculated, a reasonable question is how much of the variability that occurs in the dependent variable y is explained by the relationship with the independent variable x. This calculation is called the **coefficient of determination**. To determine this, we must first calculate some other values summarizing our regression:

- Total Sum of Squares:

$$TSS = SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Sum of Squares Error (Residuals):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Sum of Squares Regression (Model):

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

These components all add together as $TSS = SSR + SSE$. The difference between the $TSS$ (naïve model's error) and the $SSE$ ("best" model's error) is seen in Figure 23.

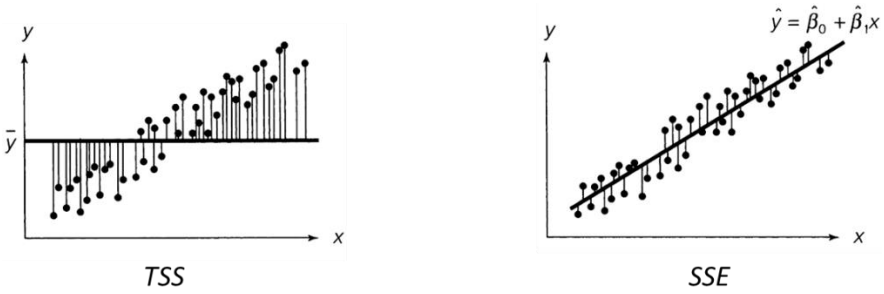



**Figure 23: Difference Between TSS and SSE**

From these equations we calculate the coefficient of determination, $R^2$.

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}$$

In the simple linear regression case where we only have one explanatory variable, $R^2 = r^2$ where $r$ is the sample correlation coefficient.

Again, there is a useful interpretation to $R^2$. About $100(R^2)\%$ of the variation in $y$ can be attributed to using $x$ as a predictor in a linear model.

### EXAMPLE

An analyst for the State of North Carolina has collected data on salaries and years of education with the intention of finding out how education affects the salary an individual makes monthly (dollars). The correlation between the two variables is 0.327. The simple linear regression line between these two variables is the following:

$$\hat{y}_i = 66.271 + 66.054x_i$$

- Interpret the values of the coefficients.

- How much salary per month do you expect someone with only a high school education (13 years) to earn monthly?

- How much more salary can a person earn monthly if they go to school 4 more years after high school?

- Calculate the coefficient of determination and interpret the value.

### REGRESSION INFERENCE

What would we conclude about the relationship between the response variable and explanatory variable if the slope of the population regression line, $\beta_1$, equaled zero? The slope of the regression line explains the average change in $y$ given a one unit increase of $x$. Therefore, if $\beta_1 = 0$, then $x$ and $y$ are unrelated. This is an important calculation in determining the relationship between the two variables. However, we do not know the true value of $\beta_1$ because it is a population parameter.

In simple linear regression, $\hat{\beta}_1$ is a statistic calculated from a sample that we are using to estimate the population parameter $\beta_1$. This is also true for $\hat{\beta}_0$ and $\beta_0$. Since $\hat{\beta}_1$ is our estimate of $\beta_1$, we will have to test to see if $\hat{\beta}_1 = 0$. The statistic $\hat{\beta}_1$ will change for every sample, but in a predictable way that is called the sampling distribution.

- Sampling Distribution for $\hat{\beta}_1$:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)$$

  - Estimate for $\sigma_\varepsilon$:

A hypothesis test for testing the slope of the population regression model is derived from the previous sampling distribution. Regression coefficient hypothesis tests are all two-sided because the sign of $\beta_1$ is not important as long as it is not equal to zero. The only time the sign of $\beta_1$ is important is in the interpretation of the results of the regression model, not the hypothesis test. Therefore, the hypotheses of the test are:

$$H_0: \beta_1 = 0 \qquad vs. \qquad H_A: \beta_1 \neq 0$$

The test statistic for this hypothesis test is of the same form as the previously studied test statistics:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Null Value}}{\text{Standard Error}}$$

$$t = \frac{\hat{\beta} - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\left(\frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)} \qquad \text{d.f.} = n - 2$$

We can also construct a confidence interval for the slope coefficient. The confidence interval has the same form as the other confidence intervals discussed in the course with Point Estimate ± Critical Value × Standard Error.

- Confidence Interval for Slope:

EXAMPLE

A clothing manufacturer wants to understand the relationship between height and weight of American men between the ages of 18 and 24 when designing their next line of clothing. They randomly sampled 101 American males of that age group to collect their data. From this data they determined the simple linear regression using height to predict weight is:

$$\hat{y}_i = -167 + 4.7x_i \qquad \sum_{i=1}^{n}(x_i - \bar{x})^2 = 900 \qquad SSE = 259320$$

- Calculate a hypothesis test to determine if the slope of the true regression line equals zero with $\alpha = 0.05$.

- Create a 90% confidence interval for $\hat{\beta}_1$.

- Can you compare these results?

## COMPLETE EXAMPLE FOR SLR

The director of admissions of a small college in the Mid-west has hired you as an analyst to administer a newly designed entrance test. This test ranges from a score of 1 to 7. You administer the test to 213 students selected randomly from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of their freshman year (y) can be predicted from the entrance test score (x). The sample's average GPA at the end of freshman year was 2.67, with a sample standard deviation of 0.72. The average entrance test score was 4.81, with a sample standard deviation of 0.69. The correlation between these two variables is 0.735. Use this information to answer the following questions.

| Parameter | Estimate | Std. Error | t Ratio | P-Value |
|---|---|---|---|---|
| Intercept | _____ | 0.238997 | _____ | _____ |
| Slope | _____ | 0.049238 | _____ | _____ |

- Create the sample linear regression line for predicting GPA from the entrance test score.

- What would a predicted GPA at the end of freshman year be for a student with scored a 6.1? What about a student who scored a 2.9?

- What would be the expected increase in GPA at the end of freshman year with an increase of 1.5 points on the entrance test?

- State the hypotheses for a test to determine whether the slope of the true regression line is equal to zero.

- Fill in the blank for the above table.

- Summarize the results of the hypothesis test for the slope.

- What is the value and interpretation of $R^2$ for this problem?

## MULTIPLE LINEAR REGRESSION

### MULTIPLE REGRESSION MODEL

The simple linear regression model is characterized with only one independent variable $x$ describing the dependent variable $y$. Although $x$ may reasonably explain some variability in $y$, multiple independent variables in the model may explain some of the unexplained variation.

- **Population Multiple Linear Regression Model (with *k* Variables):**

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

The four main assumptions for the population multiple regression model are similar to those of the simple linear regression model and sometimes an extra assumption is added.

1. Zero Mean of Errors & Linearity:

2. Normality of Errors:

3. Equal Variance of Errors:

4. Independence of Errors:

5. (Assumption Not Always Mentioned) No **Perfect** Collinearity

Just like other population parameters, we do not know the true values of the coefficients to the multiple regression model. These parameter coefficients are estimated with the sample multiple regression model.

- Sample Multiple Linear Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_k x_{k,i}$$

With multiple variables in the model, the interpretation of $\hat{\beta}_j$ changes slightly. The estimate $\hat{\beta}_j$ is the predicted (or expected or average) change in $y$ with a one unit increase in $x_j$ given **all other variables are held constant**.

Just like with linear regression, the multiple regression model is defined by a line through the points such that the **sum of the squared errors (SSE) is minimized**. The mathematics behind the development of the regression model are derived with matrix algebra that is not discussed in the statistics portion of this course and will be covered in the linear algebra portion.

Similar to simple linear regression, the coefficient of determination ($R^2$) is the percentage of variation in the dependent variable explained by its relationship to the independent variables in the model. However, mathematically the $R^2$ value will increase with the addition of any independent variable, whether significant to the prediction or not. The increase of every variable in the model decreases the degrees of freedom in the model. The addition of a low quality variable may not justify the loss of the degree of freedom. The adjusted $R^2$ measure, $R_A^2$, takes this cost into account:

$$R_A^2 = 1 - \left(\frac{n-1}{n-(k+1)}\right)\left(\frac{SSE}{TSS}\right)$$

Unlike $R^2$, the adjusted $R^2$ value may take a negative value because $R_A^2 \leq R^2$.

EXAMPLE

A real estate company is trying to model housing prices (in dollars) of their customers with the variables:

$x_1 = \text{Size of Home (square feet)}$

$x_2 = \text{Age of Home (years)}$

$x_3 = \text{Acreage of Land (acres)}$

$x_4 = \text{Number of Bedrooms}$

Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$
$$SSE = 27695831, \qquad SSR = 45963293, \qquad TSS = 73659124$$

- Interpret the coefficient on $x_2$ in the model.

- Calculate $R^2$ and $R_A^2$.

## INFERENCE FOR MULTIPLE REGRESSION

The inference behind the multiple regression model is an extension to the inference behind the simple regression model. If all of the coefficients in the population regression model ($\beta_1, \ldots, \beta_k$) are equal to zero, then the model with the independent variables does not accurately describe the dependent variable. Instead of testing each variable individually, we can test the significance of the overall model.

- Hypotheses:

- *F*-Test Statistic:

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{n-(k+1)}\right)}$$

- Mean Square Regression:

- Mean Square Error:

The *F*-test statistic come from the F distribution, which is shown in Figure 24.

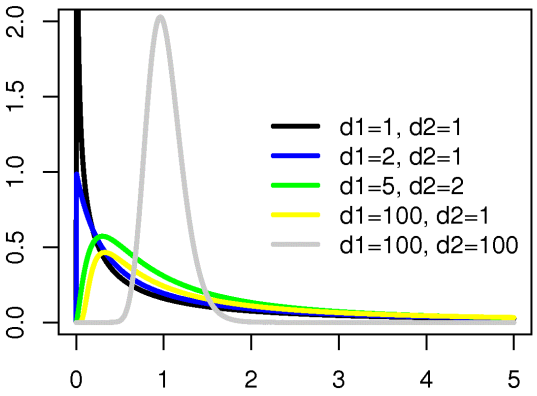



**Figure 24: F Distribution**

Here are some characteristics of the F distribution:

- Bounded below by 0.
- Right-skewed.
- Numerator and denominator degrees of freedom
  - Numerator: $k$
  - Denominator: $n-k-1$

If the test rejects the null hypothesis, then at least one of the coefficients does not statistically equal zero. Once the model proves significant, each variable should individually be tested for significance. These individual tests of significance are the exact same as the test of significance of the simple regression model with a slight adjustment to the standard error.

- Hypotheses:

$$H_0: \beta_j = 0 \quad vs. \quad H_A: \beta_j \neq 0$$

- Test Statistic:

$$t = \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\left(\frac{s_\varepsilon}{\left(1-R_j^2\right)\sqrt{\sum_{i=1}^{n}\left(x_{j,i}-\bar{x}_j\right)^2}}\right)} \qquad \text{d.f.} = n-(k+1)$$

In a similar method as simple linear regression, a confidence interval can be calculated for each of the coefficients in the multiple regression model:

$$\hat{\beta}_j \pm \left(t_{\alpha/2}\right) s_{\hat{\beta}_j}$$

EXAMPLE

A real estate company is trying to model housing prices (in dollars) of their customers with the variables:

$x_1 = \text{Size of Home (square feet)}$

$x_2 = \text{Age of Home (years)}$

$x_3 = \text{Acreage of Land (acres)}$

$x_4 = \text{Number of Bedrooms}$

Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$
$$SSE = 27695831, \qquad SSR = 45963293, \qquad TSS = 73659124$$

- Test the overall significance of the model.

- Test the individual significance of the variable $x_3$ using $s_{\hat{\beta}_3} = 3313$.

## CATEGORICAL PREDICTOR VARIABLES

All of the previous variables discussed have been quantitative in nature. However, many situations will occur where qualitative variables are good choices of explanatory variables. Variables such as gender or race have no numerical value. These variables can be added to a regression model with dummy variables.

- Dummy Variables:
    - Two Categories:

    - More than Two Categories:

    - Dummy Variable Trap:

Categorical variables need to be coded differently because they are not numerical in nature. However, there are a variety of ways to code these variables. Two of the more popular ways are dummy/reference coding and effects coding. Table 3(a) shows an example of a three category variable (A, B, C) coded with reference coding and Table 3(b) shows the same example with effects coding.

**Table 3: Reference(a) vs. Effects(b) Coding for a 3 Level Categorical Variable**

| | $x_1$ | $x_2$ |
|---|---|---|
| A | 1 | 0 |
| B | 0 | 1 |
| C | 0 | 0 |

(a)

| | $x_1$ | $x_2$ |
|---|---|---|
| A | 1 | 0 |
| B | 0 | 1 |
| C | -1 | -1 |

(b)

The only difference between the two coding methods is the interpretation of the coefficients in the model. The coefficient of the dummy variable in dummy/reference coding is that it is a shift from the reference variable. If the variable is not statistically significant, then there is no difference between the two categories. For example, in Table 3(a) a coefficient on the variable $x_1$ is testing if there is a difference between category A and category C, while the coefficient on the variable $x_2$ is testing if there is a difference between category B and category C.

In effects coding the comparison is to the overall average across all categories instead of a specific one. For example, in Table 3(b) a coefficient on the variable $x_1$ is testing if there is a difference between category A and the average of categories A, B, and C, while the coefficient on the variable $x_2$ is testing if there is a difference between category B and the average of categories A, B, and C.

EXAMPLE

Develop both effects coding and dummy/reference coding for a categorical variable with 4 categories.

EXAMPLE

The real estate company in the previous examples has decided to add a variable that determines whether the house is located on a golf course.

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4 + 12{,}550x_5 \qquad s_{\hat{\beta}_5} = 4{,}532$$

- Interpret the coefficient on the variable $x_5$ assuming that reference coding was used with a 1 for on a golf course and 0 for not.

- Calculate the test of significance for this new variable.

## POLYNOMIAL REGRESSION

It is unreasonable to assume that the multiple linear regressions we have previously discussed will adequately explain all possible sets of data. Relationships between an independent variable and a response variable do not always remain constant across all of the levels of the independent variable. An example of this would be the economic idea of diminishing returns. In diminishing returns, as the value of an input increases, the marginal effect of this input on the output diminishes.

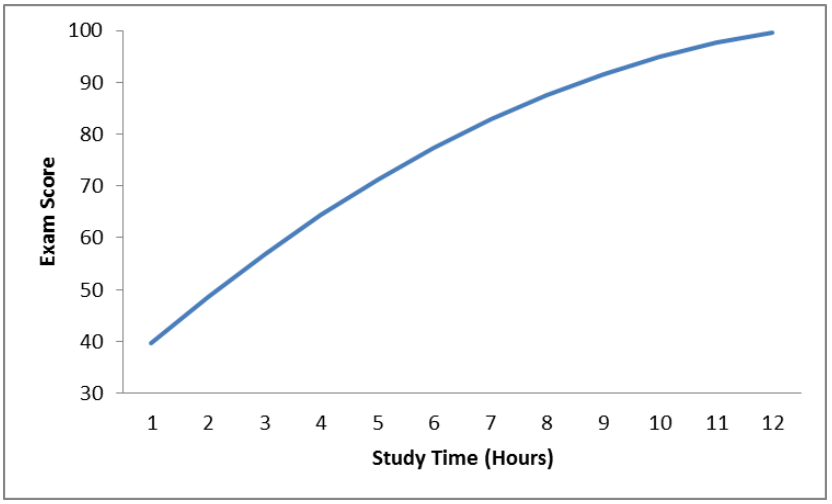



**Figure 25: Diminishing Returns on Studying**

Imagine you are studying for an exam. The more you study, presumably the higher the grade you would receive on the exam. However, the first hour of studying provides more valuable return on your grade than the eleventh hour of studying. This is shown in Figure 25.

These types of relationships are typically modeled with polynomial regression terms, with the most common being the quadratic model.

- Sample Polynomial Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{1,i}^2 + \cdots + \hat{\beta}_k x_{1,i}^k$$

- Sample Quadratic Regression Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{1,i}^2$$

  - $\beta_1$ Parameter:

  - $\beta_2$ Parameter:

The interpretation of the coefficients in a quadratic regression model are slightly different, especially around the second one as seen in Figure 26.

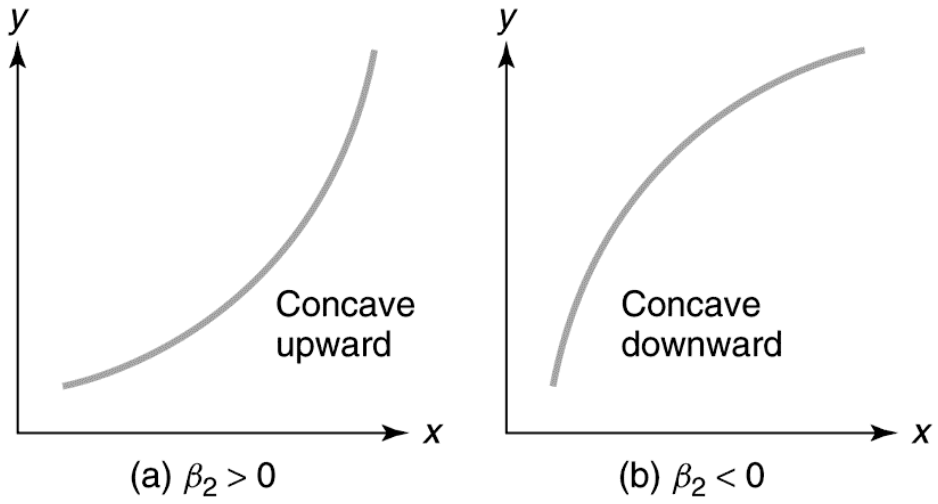



**Figure 26: Differences in Sign Changes for Quadratic Model**

Polynomial regression models are still considered linear regression models because the parameters are linearly related to the response variable – they are a **linear combination**. A linear combination is multiplying each term by constant and adding the terms together.

## INTERACTION TERMS

The previous section discussed how an independent variable may have a different relationship with the response variable across different levels of the independent variable. This was modeled through polynomial terms in a multiple linear regression model. Another scenario is when an independent variable may have a different relationship with the response variable across different levels of **another** independent variable.

- **Interaction:**

- **Independence:**

In a multiple linear regression model, an interaction term is denoted as the multiplication of the two interacting variables.

- Sample Model with Interaction Between Two Variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,1} + \hat{\beta}_3 x_{1,i} x_{2,i}$$

Let's revisit the previous section's example of diminishing returns for hours of studying. Another variable, the number exams studying for, is included as an interaction in the model. Presumably, as the number of exams increases, the returns on the studying time will get lower as shown in Figure 27.

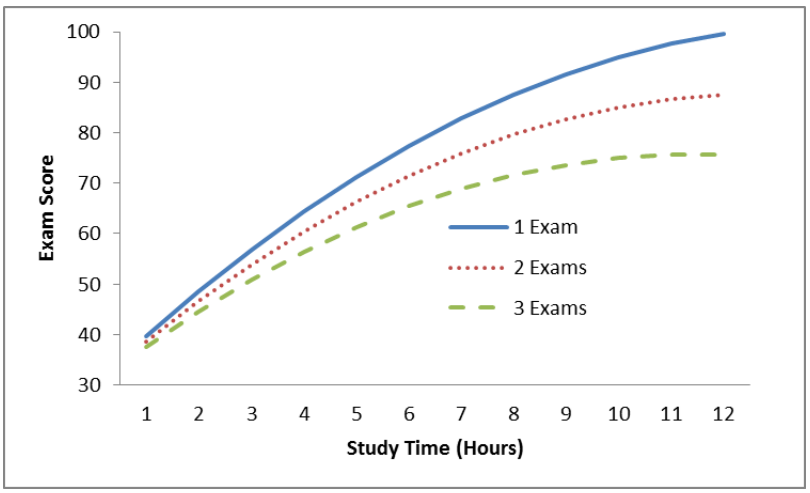



**Figure 27: Polynomial and Interaction Terms Displayed Visually**

## REGRESSION CAUTIONS

Regression models are not perfect and come with their fair share of cautions/concerns. This section is not comprehensive with regards to the cautions of regression, but does deal with three common ones – extrapolation, misspecification, and multicollinearity.

### EXTRAPOLATION

Analysts for a real estate firms try to model and predict what price a house in a certain neighborhood would sell for. Assume a model has been built for a certain neighborhood that typically has homes that range from 2000-3000 square feet in size. A new builder comes into the neighborhood and wants to build a house that is 4000 square feet in size. The real estate analyst should not use the current neighborhood model to predict a selling price for this house. Unfortunately, this problem happens all of the time in the real world – it is called **extrapolation**.

- **Extrapolation:**

The problem with extrapolation comes from not knowing the relationship between certain variables outside of the range of the data. Figure 28 gives a good example of this. The solid line in the graph represents the model that is built from the available data. However, predicting values outside the range of this line (here greater than the data values) poses a potential problem. The relationship between the input and the response variable is unknown outside of the data range, which could lead to predictions that are extremely inaccurate.

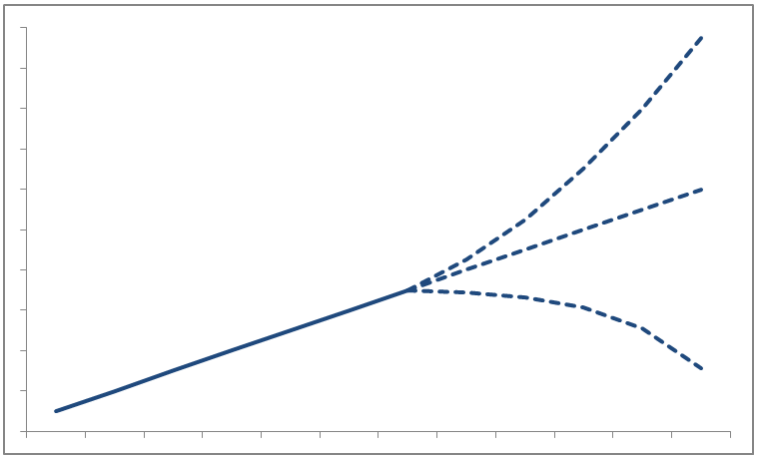


Figure 28: Example of Extrapolation

This problem is highlighted even more in polynomial regressions because the polynomial relationship might not extend past the data range available. Extrapolation in multiple linear regression becomes more complicated.

When more than one variable is in the model, it is not enough to check if the new observation to be predicted is inside each variable's range independently. The new observation must be within the joint range of the variables used to build the model.

- **Hidden Extrapolation:**

Figure 29 gives a visualization of this property. Assume the real estate analyst previously discussed built a model from homes between 2000-3000 square feet in size on acreage between 0.15-0.35 acres. The solid colored dot is a new data point to be predicted. This new observation is within each of the bounds individually, but not inside the joint bounds used to build the model.

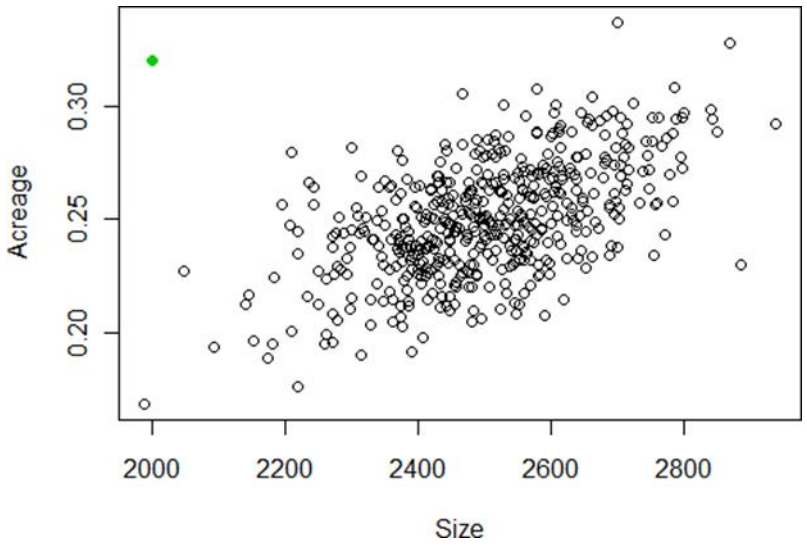



Figure 29: Example of Hidden Extrapolation

## MODEL MISSPECIFICATION

Now that we are assuming that the true population model contains multiple independent variables, we must address the issue of not selecting the proper independent variables. How is the model affected if we include irrelevant variables in the model? How is the model affected if we omit important variables in our model? These questions are important in the analysis of multiple regression models.

We will first address the issue of overspecification of a regression model. An overspecified model is a model that includes irrelevant variables in the multiple regression model.

- True Population Regression Model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$$

- Tested Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$$

In the population model, the true value of $\beta_2 = 0$. In terms of the unbiasedness of the estimate $\hat{\beta}_1$, there is no effect if the model is overspecified. However, assuming $x_1$ is correlated with $x_2$, the variance in the estimate of $\hat{\beta}_1$ is always higher than it would be in the model that is not overspecified. Even if the two variables are uncorrelated, the overspecified model will never have estimates with a lower standard deviation.

The opposite of overspecification is underspecification. This occurs when the model excludes a relevant variable. A simple example of this is defined below.

- True Population Regression Model:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

- Tested Model:

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1,i}$$

Unlike overspecified models, underspecified models potentially have biased estimates of the coefficients in the model. The true bias of the estimate of $\tilde{\beta}_1$ in the above model is the following:.

- Bias of $\tilde{\beta}_1$:

**Table 4: Direction of Omitted Variable Bias**

| | $r_{x_1,x_2} > 0$ | $r_{x_1,x_2} < 0$ |
|---|---|---|
| $\beta_2 > 0$ | POSITIVE | NEGATIVE |
| $\beta_2 < 0$ | NEGATIVE | POSITIVE |

Table 4 summarizes the conclusions from this equation of bias. This calculation of direction of bias works well in this simple case. However, when multiple variables are in the model, multiple correlations between variables may occur, which leads to unpredictable directions of bias. The variability of the estimate in the underspecified model will always be incorrectly lower than the variability of the estimate in the correctly specified model.

## MULTICOLLINEARITY

Although the overall model may appear significant, and each individual variable may appear significant, decision makers should examine the model's regression coefficients to determine if the values appear reasonable. Specifically, the signs of the coefficients should be compared with the sign of the correlation coefficient between the independent variable and the dependent variable. If these two signs are different, the model may have the problem of **multicollinearity**.

- **Multicollinearity:**

- Problems/Signs of Multicollinearity:
    - Incorrect Signs of Coefficients
    - Extreme Differences in Coefficients After Addition of Variable
    - Switches in Significance
    - Standard Deviation of Model Error Increases After Addition of Variable

Another potential problem of multicollinearity is the inaccuracy of the individual t-tests of significance for the coefficients of the multiple regression model. The variance of the model becomes inflated and leads to t-tests that show insignificance when, in fact, the variable is significant. The **variance inflation factor** is the amount of inflation that the standard error of the parameter estimates have due to multicollinearity. This piece comes from the standard error of parameter estimates in multiple linear regression.

- **Variance Inflation Factor (VIF):**

$$VIF = \frac{1}{1 - R_j^2}$$

Generally, if the $VIF < 10$ (some books say 5) multicollinearity is not considered a problem for the respective independent variable. However, if the $VIF > 10$ (some books say 5), multicollinearity is a problem. One solution to multicollinearity is to drop one of the correlated independent variables in the model. Do not drop all the variables in the multicollinearity as they provide important (but similar) information. There are many more solutions to multicollinearity such as Ridge Regression and Principal Components Regression that are beyond the scope of this course.

Multicollinearity naturally arises in a polynomial regression because of the structure of the variables. One possible solution is to standardize the variables first. This will remove any collinearity with the intercept by making the independent variables orthogonal to the intercept column in the design matrix.

### EXAMPLE

A real estate company is trying to model housing prices (in dollars) of their customers with the variables:

$x_1 = \text{Size of Home (square feet)}$

$x_2 = \text{Age of Home (years)}$

$x_3 = \text{Acreage of Land (acres)}$

$x_4 = \text{Number of Bedrooms}$

Using a sample of 105 houses they derive the following model:

$$\hat{y} = 24{,}312 + 86.5x_1 - 324x_2 + 9{,}610x_3 + 3{,}617x_4$$

They added a new variable $x_5$, the number of bathrooms in the home. The model was updated to the following:

$$\hat{y} = 28{,}438 + 62.9x_1 - 336x_2 + 9{,}610x_3 - 2{,}102x_4 + 3{,}498x_5$$

- Without examining the numbers in the models themselves, is there any potential for multicollinearity with the new variable and the older variables?

- After comparing the two models, what signs do you see that might signal multicollinearity?

## RESIDUAL ANALYSIS

Previously we calculated residuals for a regression model. We also introduced the four main assumptions behind regression models:

1. Zero Mean of **Errors** & Linearity
2. Normality of **Errors**
3. Equal Variance of **Errors**
4. Independence of **Errors**

The four main assumptions to regression models all deal with the errors from the model. These concepts are connected because by plotting and testing residuals we can test the validity of the assumptions in the regression model.

### LINEARITY

Residuals estimate the error term $\varepsilon$. Therefore, the residuals should be randomly scattered around zero on a residual plot as shown in Figure 30.

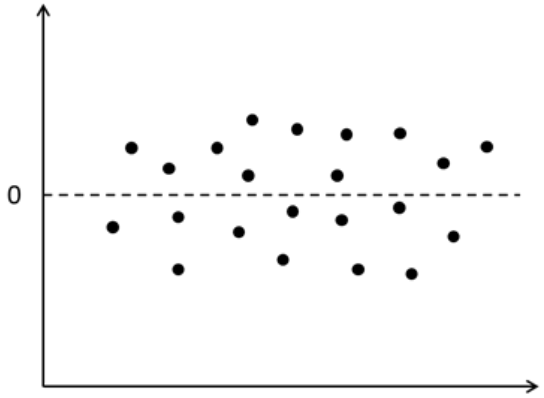



**Figure 30: Example of Randomly Scattered Residuals**

Any pattern in the residual plot is a sign of a problem with the assumptions in the model. Different patterns reveal different attributes about the residuals. Residuals with curved patterns reveal a potential lack of linearity in the data. Transformations, such as polynomial regression terms, are a possible solution to this problem.

### EXAMPLE

A health analyst is trying to establish a relationship between cholesterol levels and a new medication to help lower cholesterol. It is expected that as the drug dosage increases, the cholesterol level should drop. The analyst uses a linear regression model between drug dose and cholesterol and gets the residual plot in Figure 31.

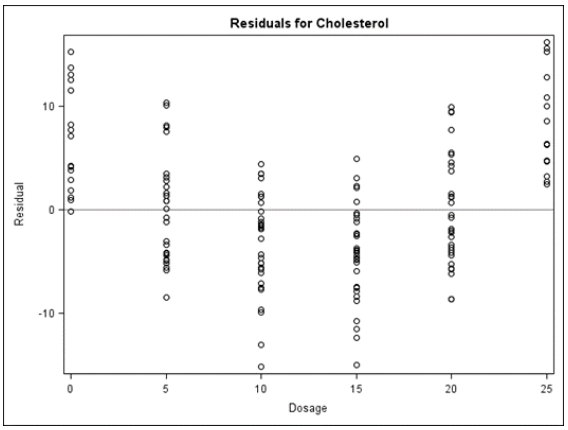



**Figure 31: Residuals from Cholesterol Model**

- From the plot, what potential relationship exists between drug dosage and cholesterol level?

## NORMALITY

The assumption of Normality for the error term is an important one from the stand point of testing variables, but not estimating coefficients. The least squares estimation process does not need the error term to be Normal for the estimates to be calculated. The Normality assumption is the underlying foundation behind the *t*-tests and F-test. If the assumption is not met, the tests of significance (along with any confidence intervals) will no longer be valid.

This assumption is hard to perfectly meet in practice. However, the results of the tests of significance only change slightly for small deviations away from Normality.

There are three common ways to check for Normality through the residuals:

1. Histograms of Residuals: these can be difficult to use since normality is hard to visualize perfectly as seen in Figure 32. Skewness might be easy to visualize, but kurtosis is not.

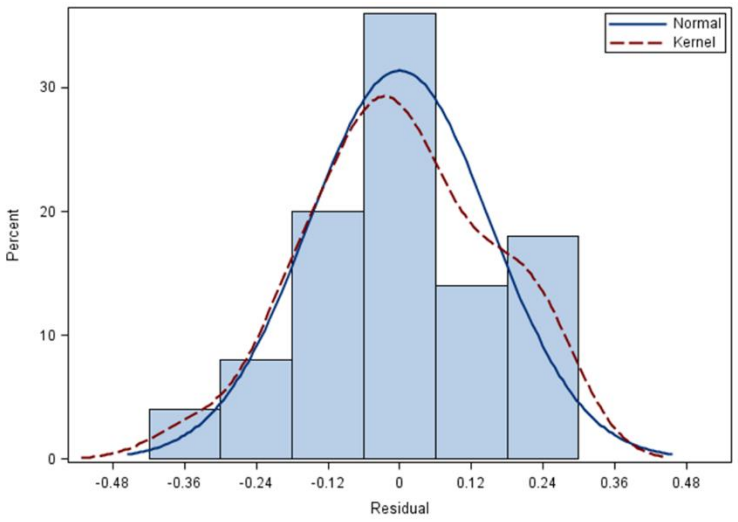



**Figure 32: Histogram of Residuals**

2. Normal Probability Plot (QQ-plot): plots of residuals against expected quantiles from a Normal distribution with the same mean and standard deviation as the residuals. If the residuals are approximately equal to their expected place on the Normal distribution, a straight diagonal line is formed as seen in Figure 33.

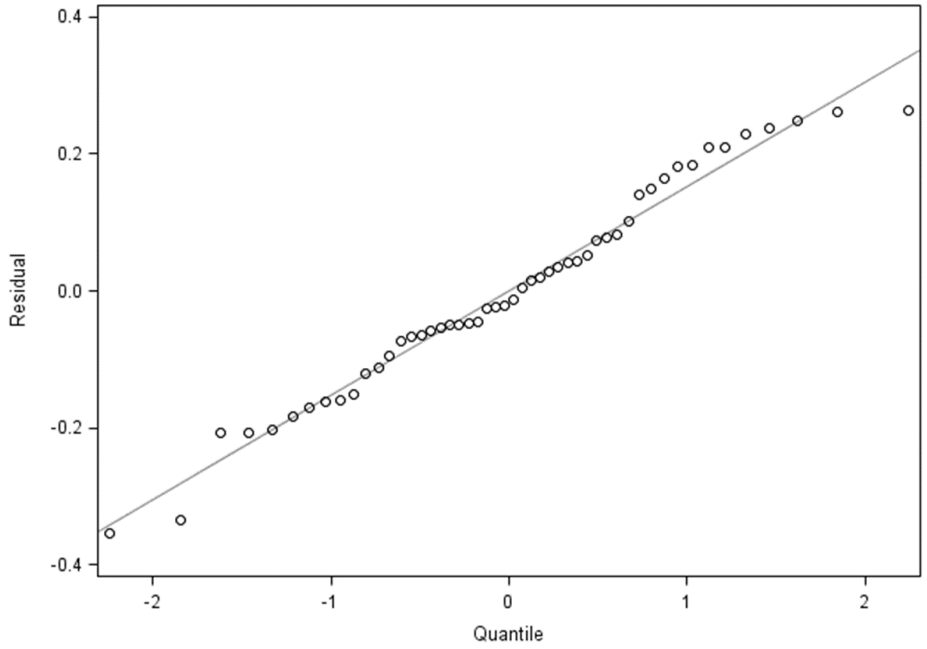



**Figure 33: Normality Probability Plot (QQ-plot)**

- Departures Due to Skewness:

- Departures Due to Kurtosis:

3. Tests of Normality: these are formal statistical tests that allow us to put a hypothesis and p-value to the question of normality. Two of the most popular tests for normality are the Kolmogrov-Smirmov (KS) and Anderson-Darling (AD) tests. Both has the same hypotheses – assume normal until proved otherwise.

One possible solution to having residuals that are not normally distributed is the use of transformations on the dependent variable. A common transformation would be the **Box-Cox transformation**.

- **Box-Cox Transformation:**

$$y` = \begin{cases} \frac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

The value of $\lambda$ is optimized to best fit the data to the needed transformation. This can be tested through trial and error by hand, or the use of computer software.

## HOMOSCEDASTICITY

When a set of random variables has an equal and finite variance, they are said to be homoscedastic. One of the assumptions of the errors in a regression model is that they are homoscedastic. Heteroscedasiticty occurs when the variances are not equal across the set of data. The visual difference between these two can be seen in Figure 34.

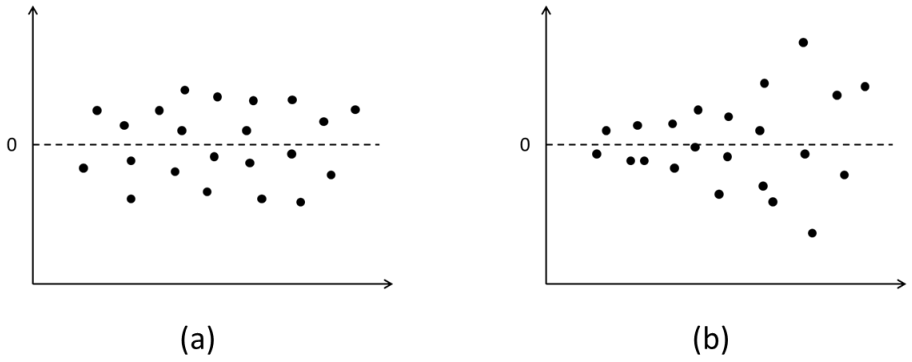



**Figure 34: Comparison of Homoscedastic(a) and Heteroscedastic(b) Errors**

Similar to a lack of Normality, the breaking of the homoscedasticity of errors assumption does not affect the parameter estimates, but the standard errors of the coefficients in the model are compromised. Also similar to a lack of Normality, heteroscedasticity can potentially be solved using a transformation of the response variable as previously described. Another possible solution to heteroscedastic errors is transformation through Weighted Least Squares (WLS), which will not be covered here.

## INDEPENDENCE

Chapter 1 previously discussed the difference between cross-sectional and time series data. Regression models with times series data can lead to problems in the modeling process. The value of a times series at the point in time $t$ is often highly correlated with the value of that same series at the point in time $t + 1$.

The Durbin-Watson $d$ statistic tests for possible correlation between the residuals.

- Hypotheses Statement:

- Durbin-Watson $d$ Statistic:

$$d = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

This statistic has the following 4 properties:

1. $0 \leq d \leq 4$
2. $d \approx 2$: Uncorrelated
3. $d < 2$: Positively Correlated
4. $d > 2$: Negatively Correlated

The sampling distribution for this test statistic is extremely complex and the knowledge of how to calculate a p-value from this test is not covered in this course.

## OUTLIERS & INFLUENTIAL OBSERVATIONS

Observations that fall outside the typical pattern of a majority of the observations in a data set affect the results in a regression model. There are two types of anomalous observations – **outliers** and **influential observations**.

- **Outlier:**

How large of a residual does an observation need to have to be considered an outlier? Typically an outlier is defined as an observation with a residual that is three standard deviations away from zero (the mean). By standardizing the residuals, we can look for observations with values greater than 3.

- **Standardized Residual:**

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{s} = \frac{(y_i - \hat{y}_i)}{s}$$

- **Studentized Residual:**

$$\hat{\varepsilon}_i^{**} = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} = \frac{(y_i - \hat{y}_i)}{s\sqrt{1-h_i}}$$

The **leverage** of an observation, $h_i$, is the influence of that particular observation on the respective predicted value. In other words, how do the respective values of the independent variables for the $i^{th}$ observation affect the prediction $\hat{y}_i$. Equation for calculating leverage is extremely complicated, therefore, computers are needed for calculation.

- **Influential Observation:**

How large of a leverage value does an observation need to have to be considered an influential observation? Typically an influential observation is one that has a leverage value $h_i > \frac{2(k+1)}{n}$, where *k* is the number of variables in the model.

Leverage isn't the only measure of how much a point influences a regression model. Here are three other common statistical measures of influence.

- **Cook's D:** the influence of each observation on the estimated $\beta$ coefficients.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\left(\frac{h_i}{(1-h_i)^2}\right)$$

  - Influential if $D_i > \frac{4}{n}$

- **DFFITS:** measures the difference between the predicted value of y with and without the observation in the regression.
    - Influential if $|DFFITS| > 2 \times \sqrt{\frac{k+1}{n}}$
- **DFBETA:** measures the difference between the estimated coefficient for each variable with and without the observation in the regression.
    - Influential if $|DFBETA| > \frac{2}{\sqrt{n}}$

### COMPREHENSIVE EXAMPLE

The director of a company's human resources department is trying to develop a model for predicting the starting salary (in dollars) of incoming employees based on the following variables:

$x_1 = \text{Years of Professional Work Experience}$

$x_2 = \text{Age in Years}$

$x_3 = \text{Gender } (1 = \text{Male}, 0 = \text{Female})$

$x_4 = \text{Previous Work Salary}$

$x_5 = \text{Communication Test (assigns scores from 0 to 100)}$

The director used a sample of 108 employees to develop the following model.

$$\hat{y} = 15{,}006.2 + 1{,}365.5x_1 + 309.3x_2 + 1{,}822.7x_3 + 0.7x_4 + 46.2x_5$$

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Model | | 89,816.28 | | | |
| Error | | 45,671.59 | | | |
| Total | | 135,487.87 | | | |

| Parameter | Estimate | Std. Error | T-Value | P-Value |
|---|---|---|---|---|
| Intercept | 15,006.2 | 5,983.5 | 2.508 | 0.0138 |
| $x_1$ | 1,365.5 | 204.8 | | |
| $x_2$ | 309.9 | 174.0 | | |
| $x_3$ | 1,822.7 | 1,368.1 | | |
| $x_4$ | 0.7 | 0.2 | | |
| $x_5$ | 46.2 | 14.3 | | |

- Fill in the blanks for the above table.
- Is the overall model a significant model? Explain.

- What would be a predicted starting salary for a 35 year old male with 10 years of previous work experience at a previous salary of $48,000 that scored an 89 on the communication test? How a female with the same qualifications?

- Interpret the coefficients of the variable $x_3$.

- The director of human resources says they are offended that the model predicts the company pays men more than women. Would the director be statistically correct in making this statement? Explain.

- Calculate the $R^2$ and $R_A^2$ from this problem.

- Based solely on the variables in the model, is there any potential for multicollinearity? Explain which variables may be correlated and a potential solution.

- Use the table below to calculate the VIF for each variable. What conclusions can you draw from the results to support your answer to the previous question?

| Variable | $R_j^2$ | VIF |
|---|---|---|
| $x_1$ | 0.54 | |
| $x_2$ | 0.63 | |
| $x_3$ | 0.88 | |
| $x_4$ | 0.98 | |
| $x_5$ | 0.21 | |

- Are there any variables you would suggest dropping from the model to account for any multicollinearity? Explain.

- What potential problems arise from this multicollinearity?

- Now which variables should we focus on dropping?

# ANALYSIS OF VARIANCE

We have studied many different types of analysis. Most of the analysis that we have covered to this point deal with continuous variables. In the previous section on regression, we first introduced the idea of dummy variables to include categorical variables in analysis. Here we will discuss further details on how to analyze categorical variables with hypothesis testing. This is called **analysis of variance (ANOVA)**.

## TWO-SAMPLE HYPOTHESIS TEST

ANOVA deals with comparing statistics across categories, but we must start small and work our way up. This section covers comparing only two categories.

### TWO POPULATION MEANS (EQUAL VARIANCES)

We have already discussed basic hypothesis testing techniques. However, when dealing with two sample hypothesis testing, the hypothesis of interest changes from what we have seen before because we are sampling from multiple populations.

- Traditional Hypotheses Statements:

- Two-Sample Hypotheses Statements:

We also assume that these two samples are independent of each other. Independent samples are samples selected from two different populations in such a way that the occurrence of values in one sample has no influence on the probability of occurrence of values in the other sample.

There are two sets of two sample analysis of means. Either the two samples can have equal variances or unequal variances with the first case being easier to analyze than the second. In the next section we will learn how to test for the equality of variance. First we will assume that the two samples have equal variances.

- Hypotheses Statements:

- Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{d.f.} = n_1 + n_2 - 2$$

- The assumption of equal variances implies that both $s_1^2$ and $s_2^2$ are estimating the same population $\sigma^2$. However, it would be unreasonable to only use one of these estimates since both provide valuable information. The pooled standard deviation $s_p$ is an "average" of these standard deviations.
- **Assumptions of Test Statistic:**
    1.
    2.

The p-value is calculated from a *t*-distribution with the appropriate degrees of freedom as defined above. Both the decision rule and conclusion of the hypothesis test are in similar form to what we have previously seen.

Confidence intervals may also be derived for means from two populations. If the variances between the two populations are assumed equal, then the confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

EXAMPLE

A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of the employees at the company. The manager assumes that the variability of salaries between genders is the same, but wants to run a hypothesis test to see if males are paid more than females. The manager samples 62 males and 77 females from the company. The sample of males had an average salary of $87,547 with a standard deviation of $5,910. The sample of females had an average salary of $78,289 with a standard deviation of $6,276.

- Run the hypothesis test and state your conclusion.

## COMPARING TWO MEANS (UNEQUAL VARIANCES)

If the variance of the two populations are not equal, then a different test statistic must be used in the hypothesis test.

- Hypotheses Statements:

- Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \text{d.f.} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)}$$

  - Under the assumption of unequal variances, we have two population variances, $\sigma_1^2$ and $\sigma_2^2$. Therefore, we need both estimates of each of these treated separately and cannot "pool" them, just use them both in standard error calculation.
  - **Assumptions of Test Statistic:**
    1.
    2.

The p-value is calculated from a *t*-distribution with the appropriate degrees of freedom as defined above. Both the decision rule and conclusion of the hypothesis test are in similar form to what we have previously seen.

Confidence intervals may also be derived for means from two populations. If the variances between the two populations are assumed unequal, then the confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

### EXAMPLE

A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of the employees at the company. The manager does not want to assume that the variability of salaries between genders is the same, but wants to run a hypothesis test to see if males are paid more than females. The manager samples 62 males and 77 females from the company. The sample of males had an average salary of $87,547 with a standard deviation of $5,910. The sample of females had an average salary of $78,289 with a standard deviation of $6,276.

- Run the hypothesis test and state your conclusion.

## TWO POPULATION VARIANCES

Previously we introduced hypothesis testing for means when the variances of the two populations are either equal or not equal. Here we introduce a formal test to tell if the two population variances are equal to each other. Although we commonly work with standard deviations, there is no statistical test for standard deviations, but there are tests for variances. If we want to determine whether we have two populations with the same variance, the hypothesis test is the following:

- Hypotheses Statements:

- Test Statistic:

$$F = \frac{s_i^2}{s_j^2} \qquad \text{numerator d.f.} = n_i - 1, \text{denominator d.f.} = n_j - 1$$

  - For a two-tailed test, place the larger sample variance in the numerator to make the value of the F statistic greater than one. This pushes the test into the upper tail of the F distribution.
  - **Assumptions of Test Statistic:**
    1.
    2.

The p-value is calculated from an F-distribution with the appropriate degrees of freedom as defined above. Both the decision rule and conclusion of the hypothesis test are in similar form to what we have previously seen. When running a two sample test of population means, the test of equal variances should be run before the test statistic is calculated to determine which version of the test statistic to use.

EXAMPLE

A human resources manager of a large business firm is trying to determine if there exists gender bias in the pay scale of the employees at the company. The manager wants to test if the variability of salaries between genders is the same before running a test that compares their means. The manager samples 62 males and 77 females from the company. The sample of males had an average salary of $87,547 with a standard deviation of $5,910. The sample of females had an average salary of $78,289 with a standard deviation of $6,276.

- Conduct the hypothesis test and state your results.

## PAIRED DIFFERENCES

The previous section on testing differences in means from two populations works well in certain situations. However, there are some instances where a paired differences sample is used to control for sources of variation that might distort the conclusions of the study.

In the example from the previous section where we tested whether men make more than women at a certain company, did we account for the fact that there might be more men in seniority positions and that is why their income across the company was higher in value? A more accurate assessment of whether men make more than women would be to sample males and females with similar job titles in the company to see if they had a difference in salary.

- **Paired (Matched) Sampling:**

When paired sampling occurs, the hypothesis test changes because our focus shifts from the values in the populations to the values of the differences in the populations. The population parameter now becomes $\mu_d$ which is the population difference in means. Therefore, we must calculate a statistic that would best represent this parameter. We use the sample paired difference to calculate a mean of paired differences $\bar{x}_d$.

- Hypotheses Statements:

- Test Statistic:

$$t = \frac{\bar{x}_d - D_0}{\frac{s_d}{\sqrt{n_d}}} \qquad \text{d.f.} = n_d - 1$$

  - Under the paired difference set-up, the focus is on the differences. Therefore, the sample size, $n_d$, is on the differences and the standard deviation, $s_d$, is of the differences.
  - **Assumptions are the same for means hypothesis testing only with the focus on the differences instead of the individual samples.**

The p-value is calculated from a *t*-distribution with the appropriate degrees of freedom as defined above. Both the decision rule and conclusion of the hypothesis test are in similar form to what we have previously seen.

Confidence intervals may also be derived for means from paired differences:

$$\bar{x}_d \pm t^*_{\alpha/2} \times \frac{s_d}{\sqrt{n_d}}$$

EXAMPLE

The human resources manager of the previous example is trying to determine if there exists gender bias in the pay scale of the employees at the company in a more meaning method than previously used. The manager samples 51 pairs of male and female employees where the pair has the same job title and experience at the company. The average difference in salaries between the two genders is $2,131 with a sample standard deviation of differences at $7,898.

- Does the human resource manager reach the same conclusion as earlier about male and female salaries?

## TWO POPULATION PROPORTIONS

The hypothesis test for population proportions can also be extended to a comparison of two population proportions. The hypothesis test for comparing if two population proportions are equal is very similar in nature to the comparison of mean hypothesis tests.

- Hypotheses Statements:

- Test Statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\bar{p}(1-\bar{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

  - For the test statistic, the estimates of the proportion are combined together in an "average" calculation.
  - **Assumptions of Test Statistic:**
    1. Independent samples
    2. Large sample size ( $n_1\hat{p}_1 \geq 5,\ n_2\hat{p}_2 \geq 5,\ (1-n_1)\hat{p}_1,\ (1-n_2)\hat{p}_2$ )

The p-value is calculated from a normal distribution. Both the decision rule and conclusion of the hypothesis test are in similar form to what we have previously seen.

Confidence intervals may also be derived for proportions from two populations:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

## EXAMPLE

A researcher at a large university on the west coast is interested in comparing some factors between upperclassmen (juniors and seniors) and underclassmen (freshman and sophomores) in the undergraduate school. The researcher believes that more experience in college may help students perform better in the classroom. The researcher is interested in testing if the average GPA of upperclassmen is greater than the average GPA of underclassmen.

- The researcher does not plan on using a paired sampling techniques. Do you agree? Explain.

The researcher sampled 89 underclassmen who had an average GPA of 2.75 with a standard deviation of 0.91. The researcher sampled 102 upperclassmen who had an average GPA of 3.07 with a standard deviation of 1.02.

- The researcher wants to conduct a two sample hypothesis test of means, but needs to know if he should assume that the standard deviations are equal. Test if the standard deviations are equal with an $\alpha = 0.1$.

- From your results to the previous question, conduct the appropriate hypothesis test to test if upperclassmen have a higher average GPA then underclassmen.

The same researcher from the previous question also believes that a higher proportion of upperclassmen live off campus compared to the proportion of underclassmen. While sampling the students in the previous sample, the researcher also asked whether the student lived off campus. Of the 89 under- classmen sampled, 27 lived off campus. Of the 102 upperclassmen sampled, 65 lived off campus.

- Construct a 95% confidence interval for the difference between the proportion of upperclassmen living off campus to the proportion of under- classmen living off campus.

- Conduct the appropriate hypothesis test to test the researcher's claim.

- Can you compare the confidence interval to the hypothesis test?

## ANALYSIS OF VARIANCE (ANOVA)

Previously, we discussed the idea of testing whether two populations had the same mean in independent samples. What if we want to determine if more than two populations have the same mean or if there is a difference in categories when there are more than two possibilities? These tests are called **analysis of variance (ANOVA)** tests.

### ONE-WAY ANOVA

The simplest form of ANOVA is the one-way analysis of variance.

- **One-Way ANOVA:**

The purpose of a one-way analysis of variance is to test if different levels of a categorical variable are a possible cause of the variation we are seeing in the response variable.

In ANOVA, the categorical variable and possible categories in the variable take on different names.

- **Factors:**

- **Levels:**

One-way ANOVA implies that we are only considering one factor. Another way to think about a one-way ANOVA is that we are testing if more than two populations have different means. In this case, the hypotheses for this kind of problem is similar to the overall significance of the multiple regression model.

- Hypotheses Statements:

- Assumptions:
    - Normally distributed categories
    - Independence
    - Equality of variance across all categories
        - We had previously learned how to test if a pair of variances were equal to each other with the F distribution. Now our hypothesis test changes because we are testing if more than two variances are all equal to each other.
        - Hypothesis Statement:

$$H_0: \sigma_1^2 = \cdots = \sigma_k^2 \qquad \text{vs.} \qquad H_A: \text{At least two are not equal}$$

- Test Statistic:

$$F = \frac{s_{max}^2}{s_{min}^2}$$

- P-value: calculated from Hartley's $F_{\max}$ distribution.

Testing if the population means are different is very similar to the F test we used in testing the overall significance in the multiple regression model. However, variability in the one-way ANOVA can come from two different places. Variation in this estimation may come from the variation within a level of a factor, or across all the levels of a factor. Again this is similar to multiple regression. In multiple regression we had variability within our model with SSR (the difference between our model and the overall average) and variability between our model and the data with SSE (the difference between our model and the actual values in the data). In one-way ANOVA we have variability within each of the populations and variability across all populations.

- **Within-Sample Variability:**

- **Between-Sample Variability:**

In a similar manner to multiple regression's notions of SSR, SSE, and TSS, ANOVA has its own representation of these. They can be visualized in Figure 35.

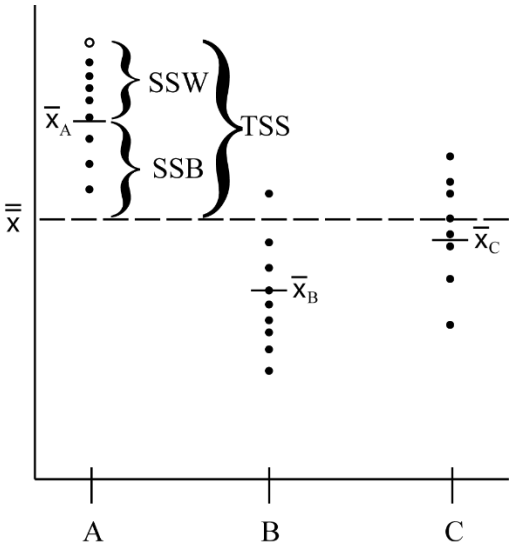



Figure 35: Sum of Squares in ANOVA

- Total Sum of Squares:

$$TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{\bar{x}}\right)^2$$

- Sum of Squares Between:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

- Sum of Squares Within:

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{i,j} - \bar{x}_i\right)^2$$

The Total Sum of Squares has a degree of freedom of $N-1$ where $N$ is the total sample size across all populations. We lose the degree of freedom by calculating the overall mean of the data. The Sum of Squares Between has $k-1$ degrees of freedom because we only need to calculate the mean of $k-1$ means. The Sum of Squares Within has all of the remaining degrees of freedom with $N-k$.

Similar to multiple regression, we can calculate a mean sum of squares for both the between and within variability.

- Mean Squares Between:

- Mean Squares Within:

Finally, we can develop the test of determining if all of the populations are equal to each other with the F test.

- Hypotheses Statements:

- Test Statistic:

$$F = \frac{\left(\frac{SSB}{k-1}\right)}{\left(\frac{SSW}{N-k}\right)} = \frac{MSB}{MSW}$$

The p-value comes from the F distribution with $k-1$ numerator degrees of freedom and $N-k$ denominator degrees of freedom.

Results from analysis of variance are typically displayed in ANOVA tables similar to Table 5.

**Table 5: ANOVA Table**

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Between | $k-1$ | $SSB$ | $\frac{SSB}{k-1}$ | $\frac{MSB}{MSW}$ | … |
| Within | $N-k$ | $SSW$ | $\frac{SSW}{N-k}$ | | |
| Total | $N-1$ | $TSS$ | | | |

EXAMPLE

A marketing analyst is interested in testing the effectiveness of 4 different commercials describing their company's new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The marketing analyst wants to test if there is a difference in sales between the commercials.

- Fill out the remaining pieces of the ANOVA table below and answer the analyst's question.

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Between | | 2.3236 | | | |
| Within | | 0.9587 | | | |
| Total | | 3.2823 | | | |

MULTIPLE COMPARISONS

If the null hypothesis in the ANOVA F-test is rejected, then the next question becomes which of the pairs of population means are different from each other. A one-way ANOVA only reveals if at least two of the population means are different, but doesn't reveal which two (or possibly more) means are different. Similar

to the multiple regression model where we tested each variable independently after testing the overall model, we must now test each pair of population means.

One way to test if each pair of population means equals each other is to construct confidence intervals for each pair's difference in population mean similar to the previous set of notes on two sample hypothesis testing. Therefore, we would compute multiple confidence intervals to make multiple inferences about our population means. However, this poses an inferential problem. We learned that the process of confidence intervals is accurate to the confidence level. For example, a 95% confidence interval process correctly contains the parameter in approximately 95% of all of the possible confidence intervals. If we construct multiple 95% confidence intervals however, then the percentage of experiments in which all of the set of confidence intervals constructed contains the true parameter is below 95%.

- **Comparison-wise Error Rate:**

- **Experiment-wise Error Rate:**

This is best seen in an example. Imagine you were trying to predict the toss of a fair coin. Each toss of the coin affords you a 50% chance of correctly guessing the outcome. You could say that in the long run, your accuracy rate is 50%. However, the probability of you guessing multiple flips of the coin in a row is smaller than 0.5. Although your accuracy and method for guessing each coin flip is correct per flip, when trying to make multiple correct guesses, your overall accuracy diminishes.

Instead of controlling the comparison-wise error rate, we can control the experiment-wise error rate through techniques other than the traditional confidence interval. The Tukey-Kramer procedure for multiple comparisons allows us to simultaneously test all pairs of means from the population without raising the experiment-wise error rate. These comparisons are typical called post-hoc tests because they are calculated after the ANOVA procedure. The Tukey-Kramer procedure develops a more conservative estimate for the confidence intervals.

- **Tukey-Kramer Honest Significant Difference (HSD) / Critical Range:**

- Critical Range (Margin of Error):

$$MOE = q_{\alpha} \times \sqrt{\frac{MSW}{2} \times \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

### EXAMPLE

A marketing analyst is interested in testing the effectiveness of 4 different commercials describing their new product. The marketing analyst randomly assigns a commercial to each of 32 cities across the country and measures the average increase in sales of their new product at their stores. The four commercial average sales were 1.2 million for commercial A, 1.8 million for commercial B, 0.76 million for commercial C, and 1.3 million for commercial D.

- Use the following ANOVA table and the Tukey-Kramer Critical Range to see if any differences exist in the average sales of the four commercials.

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Between | 3 | 2.3236 | 0.775 | 22.79 | < 0.05 |
| Within | 28 | 0.9587 | 0.034 | | |
| Total | 31 | 3.2823 | | | |

### FIXED VS. RANDOM EFFECTS

The inference drawn from an ANOVA procedure depends on whether the factor levels in the procedure are selected purposefully or randomly.

- **Fixed Effects Model:**

- **Random Effects Model:**

This distinction should be made to determine whether the analysis is restricted to the levels of interest or it can be extended beyond these values.

### ANOVA WITH RANDOMIZED BLOCKS

In two sample hypothesis testing we introduced the technique of paired sampling to help control outside sources of variation in the data affecting our results. A similar technique in ANOVA is blocking. We use

blocking when another variable that we can measure has a potential of affecting our results. **Blocking** occurs when an additional factor of at least two levels is accounted for in the analysis.

- Assumptions:
  1.
  2.
  3.

With this new blocking factor, we must perform a **randomized complete block ANOVA** instead of a one-way ANOVA. However, with the addition of the blocks comes a new source of variation that we must account for in the model.

- Sum of Squares Blocks:

$$SSBL = \sum_{j=1}^{b} k(\bar{x}_j - \bar{\bar{x}})^2$$

The block size replaces the sample size in our previous calculations in SSB, SSW, and TSS. Since we have a new source of variability, the SSW gets smaller.

$$TSS = SSB + SSBL + SSW$$

$$SSW = TSS - (SSB + SSBL)$$

Now the ANOVA table looks slightly different as seen in Table 6.

**Table 6: ANOVA Table with Blocking**

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Blocking | $b-1$ | $SSBL$ | $\frac{SSBL}{b-1}$ | $\frac{MSBL}{MSW}$ | … |
| Between | $k-1$ | $SSB$ | $\frac{SSB}{k-1}$ | $\frac{MSB}{MSW}$ | … |
| Within | $(k-1)(b-1)$ | $SSW$ | $\frac{SSW}{(k-1)(b-1)}$ | | |
| Total | $N-1$ | $TSS$ | | | |

Similar to one-way ANOVA, we can calculate a mean squares for blocking as seen in Table 6.

Not only do we have a test of whether the means of the populations of interest are different, but we also have a test to see if the blocks in the model are significantly different from each other.

- Hypotheses Statements:

- Test Statistic:

$$F = \frac{\left(\frac{SSBL}{b-1}\right)}{\left(\frac{SSW}{(k-1)(b-1)}\right)} = \frac{MSB}{MSW}$$

The p-value comes from the F distribution with $b-1$ numerator degrees of freedom and $(k-1)(b-1)$ denominator degrees of freedom.

The Tukey-Kramer post-hoc ANOVA comparisons does not work for the completely randomized block ANOVA. Instead we need to use another technique called the Fisher's Least Significant Difference Test. Similar to the Tukey-Kramer Critical Range, if the difference between the means exceeds the Fisher's Least Significant Difference (LSD), then the two means are different from each other.

- **Fisher's Least Significant Difference:**

$$LSD = t^* \times \sqrt{MSW} \times \sqrt{\frac{2}{b}}$$

EXAMPLE

The same marketing analyst from earlier in the notes is interested in testing the effectiveness of 4 different commercials describing their new product. However, this time the marketing analyst wants to block by region of the country because the location of the cities might play a role in the sales of the product as well as the commercial. The marketing analyst randomly assigns one commercial to each of 4 cities in 8 regions across the country and measures the average increase in sales of their new product at their stores.

- Fill out the remaining pieces of the ANOVA table below and answer the analyst's question. Also determine if sales were different across regions.

| Source | DF | SS | MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Blocking | | 1.587 | | | |
| Between | | 2.923 | | | |
| Within | | 0.470 | | | |
| Total | | 4.980 | | | |

## CATEGORICAL DATA ANALYSIS

Up until now, we have dealt only with categorical variables as independent variables trying to explain variation in continuous response variables. This led us to dummy variables in regression and analysis of variance (ANOVA). Now we will consider the final case where our categorical variables is our response variable. A summary of common ways to analyze different combinations of continuous and categorical variables is in Table 7.

**Table 7: Analysis by Variable Type**

| Response Variable | Predictor Variable(s): Categorical | Continuous | Both |
|---|---|---|---|
| **Continuous** | ANOVA | Linear Reg. | Dummy Variable Reg. |
| **Categorical** | Tests of Association | Logistic Reg. | Logistic Reg. |

This section will only cover the situation where we are using one categorical variable to describe another categorical variable.

### DESCRIBING CATEGORICAL DATA

Before learning how to test for relationships between categorical variables, we must first examine their structure more carefully. Categorical variables come in two types – nominal and ordinal.

- Nominal:

- Ordinal:

By examining the distribution of the categorical variables of interest, we can see potential associations between the variables. An association exists between two variables if the distribution of one of the variables changes when the level of the other categorical variable changes. The distribution will stay the same across different levels if the two variables have no association. Frequency tables display when categorical variables take on different values. Table 8 tries to compare the distributions of hunger and mood to answer the question of whether mood is associated with hunger.

**Table 8: Comparison of Moods vs. Hunger**

| | Happy | Angry |
|---|---|---|
| Not Hungry | 78% | 22% |
| Hungry | 78% | 22% |

(a)

| | Happy | Angry |
|---|---|---|
| Not Hungry | 87% | 13% |
| Hungry | 40% | 60% |

(b)

In Table 8(a), the distribution for happiness does not change for either value of hunger. This would imply that there is no relationship between the two variables. In Table 8(b), the distribution of happiness changes dramatically depending on the value of hunger. This would imply that there is a relationship between the two variables. These tables are similar to the frequency tables we discussed previously in Chapter 3.

## TESTS OF ASSOCIATION

How much of a change in the distribution of mood would result in the conclusion that a relationship exists between mood and hunger? Although we said a change in distribution would result in an association, there are some formal tests of association we can consider.

### GENERAL TESTS

The first test we will examine is the Pearson Chi-Square Test. This test can be used to compare **any two** categorical variables.

- Hypotheses Statements:

- Test Statistic:

$$Q_P = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(Obs_{i,j} - Exp_{i,j}\right)^2}{Exp_{i,j}} \qquad \text{d.f.} = (\#\text{Rows} - 1) \times (\#\text{Columns} - 1)$$

  - Expected Cell Value:

$$Exp_{i,j} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Sample Size}}$$

The p-value from this test comes from the $\chi^2$-distribution with the degrees of freedom specified above. The decision rule and conclusion statements are the same as previous hypothesis tests.

Chi-square distributions are squared standard normal distributions. They are bounded below by zero, right-skewed, and have one set of degrees of freedom. Figure 36 displays five different $\chi^2$-distributions.

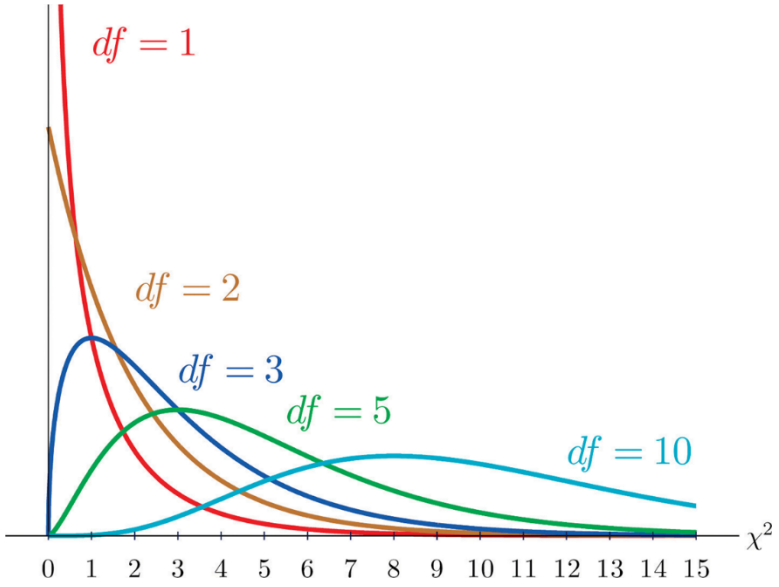



Figure 36: Chi-Square Distribution

The Pearson Chi-Square Test measures the difference between the expected cell count (which equals the observed cell count if no association exists) and the observed cell count. If the difference between the two is large enough, then the null hypothesis of no association between the variables is rejected.

Another similar test that uses observed and expected cell counts is the Likelihood Ratio Chi-Squared Test. Instead of using the differences between expected and observed cell counts, this test uses the ratio between the two.

- Hypotheses Statements:

- Test Statistic:

$$Q_{LR} = 2 \times \sum_{i=1}^{R} \sum_{j=1}^{C} Obs_{i,j} \times log\left(\frac{Obs_{i,j}}{Exp_{i,j}}\right) \qquad \text{d.f.} = (\#\text{Rows} - 1) \times (\#\text{Columns} - 1)$$

  - Expected Cell Value:

$$Exp_{i,j} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Sample Size}}$$

The p-value from this test comes from the $\chi^2$-distribution with the degrees of freedom specified above. The decision rule and conclusion statements are the same as previous hypothesis tests.

EXAMPLE

A manager of a major car dealership wants to determine if the membership of a client in their loyalty program is associated with the color of car that they buy. With this knowledge, it potentially could help the sales staff show different cars to different clients to help improve the likelihood of a sale. The manager pull information from the previous year's sales.

- Fill in the expected value portion of the table below.

Observed

| Color | Yes | No | Total |
|---|---|---|---|
| Black | 149 | 101 | 250 |
| White | 101 | 66 | 167 |
| Blue | 72 | 108 | 180 |
| Red | 96 | 161 | 257 |
| Green | 39 | 65 | 104 |
| Total | 457 | 501 | 958 |

Expected

| Color | Yes | No | Total |
|---|---|---|---|
| Black | | | 250 |
| White | | | 167 |
| Blue | | | 180 |
| Red | | | 257 |
| Green | | | 104 |
| Total | 457 | 501 | 958 |

- Compute $Q_P$ and $Q_{LR}$ and summarize the results.

## ORDINAL TESTS

For a better test of association between ordinal variables, the Mantel-Haenszel Chi-Square Test should be used instead. The Mantel-Haenszel Chi-Square Test takes the ordinality of the variables into account. The Pearson and Likelihood Ratio $\chi^2$ tests can also be used for ordinal variables, but they do not take the ordinality into account in the calculation of the statistic. Ideally, all information about a variable should be used whenever possible.

- Hypotheses Statements:

- Test Statistic:

$$Q_{MH} = (n-1)r^2 \qquad \text{d.f.} = 1$$

  - $r^2$: the Pearson correlation coefficient between the row and column variable.

The p-value from this test comes from the $\chi^2$-distribution with the degrees of freedom specified above. The decision rule and conclusion statements are the same as previous hypothesis tests.

## MEASURES OF ASSOCIATION

The downside of the tests of association mentioned in the previous section is that they are just tests of whether an association between categorical variables exists, not the strength of that association. There are different statistics to measure the strength of an association between two categorical variables. The first measure of association is called the odds ratio. The odds ratio is only used for analyzing $2 \times 2$ contingency tables.

- **Odds Ratio:**

- **Odds:**

$$Odds = \frac{p}{1-p}$$

Although many people mistakenly interchange odds and probabilities, they are two different concepts. This is best highlighted with an example.

### EXAMPLE

Use the following table summarizing data around loyalty program customers and whether they bought our new product.

| | Bought Product - YES | Bought Product - NO | Total |
|---|---|---|---|
| **Loyal** | 20 | 60 | **80** |
| **Non-Loyal** | 10 | 90 | **100** |
| **Total** | **30** | **150** | **180** |

- Calculate the probability of a loyalty program customer buying the product.

- Calculate the probability of a non-loyalty program customer buying the product.

- Calculate the odds of a loyalty program customer buying the product.

- Calculate the odds ratio of loyal vs. non-loyal customers buying the product.

Odds ratios cannot be used for contingency tables greater in size than $2 \times 2$. For these comparisons, the Cramer's V Statistic is used. Again, Cramer's V is only a measure of the amount of association.

- Cramer's V:

$$V = \sqrt{\frac{\left(\frac{Q_P}{n}\right)}{min(\#Rows - 1, \#Columns - 1)}}$$

Cramer's V is bounded between 0 and 1 where closer to 0 implies a weaker relationship. The only exception is a $2 \times 2$ table where Cramer's V is then bounded between -1 and 1. Closer to 0 still implies a weaker relationship, but Cramer's V allows for a direction of relationship in a $2 \times 2$ table.

EXAMPLE

The same manager as the previous example now wants to know the strength of the relationship between the color of car and loyalty program.

- Use the appropriate measure of association to calculate this.