AP Statistics

Arnav Patri

April 7, 2022

Contents

Part I

One-Variable Data

Chapter 1

One-Variable Quantitative Data

Statistics is the science of collecting and analyzing data.
A data set is a collection of data on several **individuals**. These individuals can be anything.
Data provides values for **variables**, which describe some *characteristic* of an *individual*.
Quantitative variables provide numerical values that describe or measure some characteristic while **categorical variables** assign labels that place each individual into one of several groups.
A variable's **distribution** describes the *frequency* with which a variable takes on its possible values.

1.1 Graphically Displaying Distributions

Quantitative variables can either be **discrete**, having some countable set of possible values, or **continuous**, having an uncountable set of possible values.

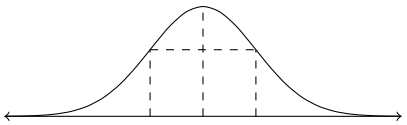
> The quantitative variable being described must always be defined, typically as a capital letter. An arbitrary particular value is denoted with a lowercase letter and a superscript is added to denote a defined value.

Dot plots assign the horizontal axis to the variable and show each value's frequency with a number of dots above, each corresponding to an individual.
Stem (and leaf) plots can only display quantities. The stem (vertical axis) corresponds to the first digit while the leaf (horizontal axis) corresponds to the remaining digits. The stem is always on the left, and the leaves should always be ordered from least to greatest. A key is required to denote the magnitudes displayed.
Histograms show bars that are assigned equal intervals. To find the length of each interval, the range can be divided by the number of classes and the result rounded up. Each bar's height shows the number of individuals within the class.
Time plots measure a variable over time.
A **cumulative frequency curve** (OGIVE) is a line graph that shows the cumulative relative frequency, the sum of all lower classes' **relative frequencies**, the percentages of data contained within each class, (inclusive). Rather than each class being represented as a range, they are labeled by their *medians*.

Graphically Describing Distributions

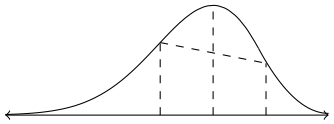
A distribution can be described by its *shape*, *outliers*, *center*, and *spread* (**SOCS**).

Shape In order for a distribution to be *symmetric*, values that are equidistant from the mode must have the same frequencies.

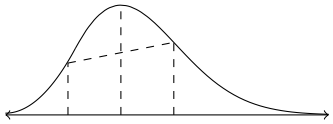


Symmetric

Skew is dependent on the direction of the tail.



Skew Left



Skew Right

Outliers An **outlier** is a point that does not fit the general trend of the data. It is marked as an asterisk on a modified box plot.

Center The *mean* ($\bar{x}$), *median* (Q_2), and *mode* are measures of center, meaning that they measure where the data is centered. The relationship between these quantities is related to the skew.

Skew	Left	Center	Right
Relationship	$\bar{x} < Q_2 <$ mode	$\bar{x} = Q_2 =$ mode	$\bar{x} > Q_2 >$ mode

Spread The *standard deviation*, *range*, and *interquartile range* are measures of **spread**, meaning that they measure how spread apart the data is. The greater they are, the more variability there is in the data.

1.2 Numerically Describing Distributions

The **mean** is the sum of each value in the distribution divided by the total number of values. It is also referred to as the *average* or *expected value* of its variable. For a sample, it is denoted using a bar over the lowercase form of its variable (X becoming $\bar{\boldsymbol{x}}$).

$$\bar{x} = \frac{\sum x_i}{n}$$

The first, second, and third **quartiles**, denoted $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$, and $\boldsymbol{Q_3}$, are the values with 25%, 50%, and 75% of the data below them respectively.[1] The second quartile is the **median**.
The **mode** is the value that appear with the greatest frequency. It is, along with the mean and median, a measure of center.

If multiple values have the same greatest frequency, they are all modes.

The **range** is the difference between the highest and lowest values of a data set.
The **interquartile range**, denoted $\boldsymbol{IQR}$, is the difference between the values of the third and first quartiles.[2] It therefore shows the "middle half" of the distribution.

$$IQR = Q_3 - Q_1$$

[1]The probability of X falling below each quartile is equal to its number multiplied by 25%.

$$P(X < Q_1) = 0.25 \qquad P(X < Q_2) = 0.5 \qquad P(X < Q_3) = 0.75$$

[2]A box plot shows the interquartile range as the box and the median as the line within it.

The **standard deviation** is the average distance from the mean. It is denoted by $\boldsymbol{s}$ with the subscript of its variable's lowercase form (for a sample). Along with the range and interquartile range, it is a measure of spread/variability.

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

The **variance** is the square of the standard deviation and is accordingly denoted by $\boldsymbol{s^2}$ with the appropriate subscript. [3]

An *outlier* is defined as a point that is more than 1.5 times the interquartile range from the "middle half" of the data.

$$x_i \notin (Q_1 - 1.5IQR, Q_2 + 1.5IQR) \implies x_i \text{ is an outlier}$$

> It is crucial to note that the mean, standard deviation, variance, and range are only to be used when there are no outliers, as outliers disproportionately affect them. The median, mode, and interquartile range can always be used, though, as they are **resistant**.

A value's **percentile** is the percentage of all values that are at less than or equal to it. This is simply the probability of the variable being at most the value.

$$\text{percentile} = P(X \leq x)$$

A **linear transformation** shifts all values by the same amount a and/or proportionally to their values (with constant of proportionality b), resulting in a new variable.

$$y_i = a + bx_i$$

It should be noted that the measures of *center* are changed by both a and b, but only the former alters the measures of *spread*.

1.3 Density Curves

A **density function** typically denoted $\boldsymbol{f}$, outputs the "*relative frequencies*" of every possible value of its variable. As such, the total area under it must be exactly 1 and its values must always be positive.

The graph of a density function is a **density curve**. A density curve is a *continuous* distribution, so the probability of X being any particular value is 0.[4]

> Because the probability of X being any particular value is 0, the function's output values do not actually represent anything, though they can be likened to the relative frequencies of distribution curves.

Because a density curve is effectively an idealized distribution curve, representing a population rather than a sample, different variables are used.

Quantity	Sample Data	Population Value
Mean	$\bar{x}$	μ
Standard Deviation	s	σ

[3]The values of n, $\sum x_i$, $\bar{x}$, Q_1, Q_2, Q_3, range, IQR, s_x, and s_x^2 are all calculable automatically given a data set. The data can be entered into Ln (`Stat/EDIT/1`), and `1-Var Stats` (`Stat/CALC/1`) can be performed with Ln (`ALPHA/`n) as its parameter.

[4]Probability is equal to the integral over the desired interval, so when both bounds are the same, it will be equal to zero.

$$\int_x^x f(t)\,\mathrm{d}t = F(x) - F(x) = 0$$

The area under the density curve from its leftmost value to any given value is that value's *percentile*. It can be found using a **cumulative distribution function**, generally notated $\boldsymbol{F}$.[5]

$$P(X \leq x) = F(x)$$

As a density curve is categorically *continuous*, the inclusivity of the binary relation of the probability statement is irrelevant to the value of the percentile.[a]

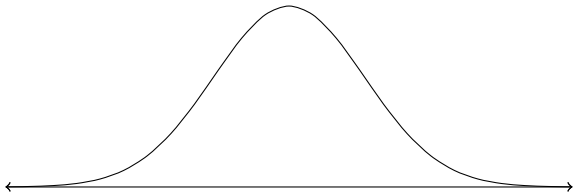
[a] The lack of regard for inclusivity with percentile, as well as other probability statements made regarding continuous distributions, is due to the area being $\int_{-\infty}^{x} f(t)\,\mathrm{d}t$. As $\mathrm{d}t$ is a differential, it is infinitesimal, so adding or subtracting $f(x)\,\mathrm{d}t$ to or from the integral has no tangible impact on its value.

The probability of X falling within two values is equal to the difference of their percentiles and the chance of X falling above a value is 1 minus its percentile.

$$(x_1 < X < x_2) = P(X \leq x_2) - P(X \leq x_1) \qquad P(X > x) = 1 - P(X \leq x)$$

The *median* of a density curve is the point that splits the curve into two regions of equal area[6] while its *mean* is the point at which the entire curve would be balanced[7]. Its *mode* its its absolute maximum.

1.4 The Normal Distribution



The Normal Distribution

The **Normal distribution** $\boldsymbol{\varphi}$ is a *density curve* that is *bell-shaped* and *symmetric* with *inflection points* one standard deviation from the mean.[8] Its *mean*, *median* and *mode* 0 while its *standard deviation* is 1.

[5] The cumulative distribution function is the area under f from its minimum value to the desired x value. It is thusly defined as the following integral:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

(As the density function is equal to 0 outside of the domain of x, an infinity can be used as the lower bound.

[6] $P(X < x) = P(X > x) = 0.5$, as the areas are equal and the total area must be equal to 1. This is also consistent with the its prior definition.

[7] The mean of the curve is the x-value of the centroid of $f(x)$ over its entire domain, though as its total area is always equal to 1 by definition, this can be simplified.

$$\mu = \frac{\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x}{\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x} = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

(The bounds of the integral being infinities does not affect its value, as a distribution curve is 0 outside of its domain.) This is consistent with the definition of the mean for a sample.

$$\mu = \frac{\sum x_i}{n} = \sum \left[\frac{x_i}{n}\right] = \sum x_i P(X = x_i) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

[8] The Normal distribution is defined as such:

$$\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

For a curve to be **Normal**, displaying **Normality**, it must be some transformation[9] of the Normal distribution curve. A Normal curve with parameters μ and σ is denoted $\mathcal{N}(\mu, \sigma)$, and its application to a variable X is denoted $X \sim \mathcal{N}(\mu, \sigma)$.[10]
To justify the Normality of a data set, a modified box plot can be created and symmetry and a lack of variation shown.
Percentile can be found using the **cumulative Normal distribution function** Φ.[11]

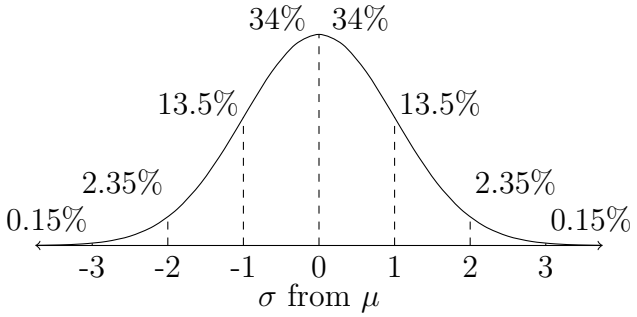
$$\Phi(x) = P(X < x)$$

The **inverse Normal function $\mathbf{\Phi^{-1}}$** takes a percentile as an input and returns a value.[12]

$$\Phi^{-1}(P(X < x)) = x$$

The Normal distribution follows the **empirical rule**, which states the following:

Range	$\mu \pm \sigma$	$\mu \pm 2\sigma$	$\mu \pm 3\sigma$	$(\mu, \mu+\sigma)$	$(\mu+\sigma, \mu+2\sigma)$	$(\mu+2\sigma, \mu+3\sigma)$	$(\mu+3\sigma, \infty)$
Proportion of Data	60%	95%	99.7%	34%	13.5%	2.35%	0.15%



To *simulate* a situation that shows Normality, *random* numbers can be generated following a Normal distribution.[13]

---

[9] Changing the mean shifts the graph of the Normal distribution horizontally while changing the standard deviation stretches or shrinks it vertically (the larger, the flatter, and the more data in the "tails").

[10] A Normal distribution of X with parameters μ and σ can also be denoted as such:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

[11] The cumulative Normal distribution function is defined as such, where erf is the error function:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

Percentile can be evaluated on a calculator (for a standard or nonstandard Normal distribution) using the `normalcdf` (`2nd/distr/2`) function:

$$P(X < x) = \texttt{normalcdf}\,(\text{lower} : -\infty, \text{upper} : x, \mu : \mu, \sigma : \sigma)$$

The probability of X falling within a specific range can be found with this function as well:

$$P(x_1 < X < x_2) = \texttt{normalcdf}\,(\text{lower} : x_1, \text{upper} : x_2, \mu : \mu, \sigma : \sigma)$$

[12] A percentile's corresponding x value can be found using the `invNorm` (`2nd/distr/3`)function:

$$x = \Phi^{-1}(P(X < x)) = \texttt{invNorm}\,(\text{area} : P(X < x), \mu : \mu, \sigma : \sigma, \text{Tail} : \text{LEFT})$$

[13] To generate n random numbers following a Normal distribution, the `randomnorm` (`math/PROB/6`) function can be used as such:

$$\texttt{randomnorm}\,(\mu : \mu, \sigma : \sigma, n : n) = \{x_1, x_2, \ldots, x_n\}$$

This list can then be stored using $\texttt{sto} \rightarrow \texttt{L}_n$

Standardization and z-Scores

The ***z*-score** is the number of standard deviations away from the mean.[14]

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

z-scores allow the **standard Normal distribution** to always be used[15], as it turns x into z, μ into 0 (as the numerator becomes zero), and σ into 1 (as z is measured in units of σ).

$$f(x) = \varphi(z) \qquad\qquad F(x) = \Phi(z)$$

> z-scores can always be calculated, but Normality must be confirmed in order for percentile to be calculable.

The process of converting a value to its z-score is called **standardization**.

When answering a question involving calculating the value with a corresponding percentile, the following should be done:

1. Define the variable and its distribution, using the given parameters (and justifying Normality).
   (a) To show Normality, the situation can be quoted, a modified box plot can be used, or a **Normal probability plot**, which shows x vs z, can be created, and its straightness verified.
   (b) If a graph is not given, draw the distribution, labeling the mean, given/desired particular value(s), and given/desired percentiles.
2. Calculate/denote the z-scores of any particular values.
   (a) Use the definition of z-score as the number of standard deviations from the mean if given particular values.
   (b) Use the inverse Normal function if given percentiles.
3. Use the formula for z-score to calculate the desired variable.

$$x = \mu + z\sigma \qquad\qquad \mu = x - z\sigma \qquad\qquad \sigma = \frac{x - \mu}{z}$$

[14] z-scores follow the same conventions as variables, Z being used for any value, z for any particular value, and z_i for a defined particular value. The subscript may also denote the variable that it is the z-score of, though, in case the cases would correspond, as in Z_X, z_x, and $z_{x,i}$

[15] z should be used in place of x when working with φ and Φ.

Chapter 2

One-Variable Categorical Data

A **frequency table** shows the number of individuals that have a certain value while a **relative frequency table** shows the percentage of all individuals in the data set that have that particular value.
Bar graphs show each category as a bar, the height of which corresponds to its frequency.
Pie charts show each category as some section of a circle that is bounded by two radii. The areas of each slice is proportional to the frequency.

Part II

Two-Variable Data

Chapter 3

Two-Variable Quantitative Data

A **response** (*dependent*) **variable** measures the outcome while a **explanatory** (*independent*) **variable** explains the data.
The term **bivariate data** refers to data with two variables that are recorded for the same set of *individuals*.
A **scatterplot** is, as the name suggests, a set of points, each representing an individual, scattered about a grid with x and y axes that correspond to the response and explanatory variables respectively.[1] They are the most effective way for the relationship between two quantitative variables measured on the same individuals.

When storing data within lists, it is crucial that each element x_n and y_n correspond to the same individual.

A scatterplot can be described by its *form*, *direction*, and *strength*.
Form describes the nature (shape) of the variable's relationships (linear, polynomial, root, exponential, logarithmic, sinusoidal, etc.). More generally, form can be described as linear or nonlinear. **Strength** describes how strongly correlated the variables are.
Association direction describes whether a positive change in the explanatory variable results in an increase (positive) or decrease (negative) in the explanatory variable.

3.1 Correlation

The **correlation coefficient** $\boldsymbol{r}$ is equal to the sum of the products of the z-scores of each variable for each individual divided by one less than the number of individuals.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{\sum z_{xi} z_{yi}}{n-1}$$

The magnitude of r corresponds to strength while its sign indicates the *association direction*.

[1] To plot the data of two lists as a scatterplot, `stat plot` (`2nd/stat plot`)) can be pressed, and then a plot selected and toggled `On`, its first option selected for type, and the `X` and `Ylist` set to their corresponding lists. To cycle through individuals in `Xlist` index order, it `trace` can be selected.

The following should be noted regarding r:

- It is the same for x against y and y against x (making it *commutative*).
- It requires that both variables be quantitative.
- It is independent of the units of the variables.
- Its magnitude cannot exceed 1.
- It only measures the strength of *linear* relationships.

  The correlation may be 0 even when there is a strong nonlinear pattern.

- It is not *resistant*, as its formula includes non-resistant means and standard deviations, meaning that it is prone to being affected by outliers.
- It does not completely describe bivariate data.

  A strong correlation alone is not enough to ensure linearity.

- It is not affected by *linear transformations* of either variable, even if they are not transformed correspondingly.

3.2 Regression and Residuals

The **line of best fit** of a scatterplot is the line that best predicts the data. A method for finding a linear line of best fit is **least-squares regression**.[2] The line generated by this method is referred to as the **least-squares regression line** or **LSRL**.

The LSRL enables predictions to be made regarding how the response variable when the explanatory variable is changed. It is defined as the line the minimizes the sum of the squares of the **residuals** (denoted $\boldsymbol{e}$), the differences between the predicted and actual values of the response variable.[3]

A **residual plot** shows the residuals of each data point. A random pattern suggests linearity.[4]

The regression line of a residual plot is always $\hat{y} = 0$, as the sum of all residuals is equal to 0. This also means that the mean of the residuals is equal to 0.

Due to the fact that it uses a sum of every point's values, the LSRL is not resistant to **influential observations**. Influential observations are often outliers in the y direction, especially in smaller data sets. To identify whether a point is an outlier, a modified box plot can be created. A point is influential if removing it markedly changes r.

The predicted value for the explanatory variable is denoted by putting a "hat" over the variable. The LSRL can be described as a linear transformation that maps x onto y:

$$\hat{y} = a + bx$$

Points with large residuals are *outliers*.

Predictions made within the data are **interpolations** while those made outside the bounds of the data

[2] On a calculator, a linear regression can be carried out by entering the data into two corresponding lists. From there, `LinReg(ax+b)` (`stat/CALC/4`) with the appropriate `X` and `Ylist`s entered. This also calculates the r value. To store the equation, `Store RegEQ` can be set to $\texttt{Y}_n$. Before a regression is run, though, `DiagnosticOn` (`2nd/catalog/DiagnosticOn`) should be selected.

[3] A list of residuals can be created by setting $\texttt{L}_n$ to `RESID` (`2nd/list/7`), though the regression should first be run.

[4] This can be found on a calculator by setting `Ylist` to `RESID` (`2nd/list/7`) when creating a scatterplot. This should be done after running a regression.

are **extrapolations**.

Extrapolation should be avoided, as it is more likely not to be accurate than interpolation.

a and b can be calculated as such:[5]

$$a = \bar{y} - bx \qquad\qquad b = r\frac{s_y}{s_x}$$

The definition of a means that the LSRL always passes through $(\bar{x}, \bar{y})$.
The **coefficient of determination r^2** is the percentage of the change/variation in the response variable that can be explained by the LSRL relating the response variable to the explanatory variable. It is maximized by the LSRL.[6]
Clusters are groups of points that are similar.
To add a categorical variable, the data can be displayed using different symbols for each group.

3.3 Transformations to Achieve Linearity

The form of a relationships is not always linear, but the LSRL calculates a linear line of best fit. To resolve this, x and/or y can be transformed using powers, roots, or logarithms. This process is known as **linearization**.
The axes of the scatterplot are changed in accordance with the linearization.
An **exponential model** takes the form of an exponential function.

$$\hat{y} = ab^x$$

A **logarithmic model** takes the form of a natural logarithmic function.

$$\hat{y} = a + b\ln x$$

[5] a and b can be derived by differentiating the sum of the squares of the residuals.

$$E = \sum(y_i - \hat{y})^2 = \sum(y - a - bx_i)$$

Because the error is being minimized, its derivatives are equal to 0.

$$\begin{aligned}
\partial_a E &= \partial_a \sum(y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i)^2(\partial_a[y_i - a - bx_i]) \\
&= \sum(y_i - a - bx_i)(-1) \\
&= \sum y_i - \sum a - \sum bx_i = \sum y_i - an - b\sum x_i \\
a &= \frac{\sum y_i - b\sum x_i}{n} = \bar{y} - b\bar{x}
\end{aligned}$$

$$\begin{aligned}
\partial_b E &= \partial_b \sum(y_i - a - bx_i)^2 \\
0 &= \sum 2(y_i - a - bx_i)(\partial_b[y_i - a - bx_i]) \\
&= \sum(y_i - a - bx_i)(-x_i) = \sum(x_i y_i - x_i\bar{y} + b\bar{x}x_i - bx_i^2) \\
&= \sum(x_i y_i - x_i\bar{y}) - b\sum(x_i^2 - \bar{x}x_i) \\
b &= \frac{\sum(x_i y_i - \bar{y}x_i)}{\sum(x_i^2 - \bar{x}x_i)} = \sum\left(\frac{y_i - \bar{y}}{x_i - \bar{x}}\right) = \sum\left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}\right) \\
&= \frac{1}{s_x(n-1)}\sum\left(\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x}\right) \\
&= \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \times \frac{s_y}{s_x} = r\frac{s_y}{s_x}
\end{aligned}$$

[6] Given no information regarding x, the predicted value of y is simply $\bar{y}$, so the residual will be the difference between y and $\bar{y}$. The sum of all residuals must therefore be 0:

$$\sum e_i = \sum(y_i - \bar{y}) = \sum y_i - \sum \bar{y} = \sum y_i - n\bar{y} = \sum y_i - \frac{n\sum y_i}{n} = -\sum y_i - \sum y_i = 0$$

Because the sum of the residuals is 0, the sum of the squares of the residuals can instead be used to compare errors between approximations. r^2 is equal to the percentage of error removed by using the LSRL rather than $\hat{y} = \bar{y}$.

Exponentiating one variable and taking the logarithm of the other are equivalent operations.

A **power model** takes the form of a polynomial with direct variation.

$$\hat{y} = ax^b$$

Taking the logarithm of both sides simply results in a power model after it is re-expressed.

$$\begin{aligned} \ln \hat{y} &= a + b \ln x \\ e^{\ln \hat{y}} &= e^{a+b \ln x} \\ \hat{y} &= e^a e^{\ln x^b} = cx^b \end{aligned}$$

The model that bests fits the data is the one that minimizes s, the standard deviation of the residuals.

Chapter 4

Two-Variable Categorical Data

Segmented bar graphs have one bar per variable. Each bar is divided into segments according to their *relative frequency*. The height of each bar is equal to 1.
Side-by-side bar graphs show the distribution of one categorical variable for each value of another. The grouping of the bars is based on one of the categorical variables while the bars themselves show the frequencies of the values of the other.
Mosaic plots are variants of segmented bar graphs that display the number of individuals in a category by making each bar's width proportional to it.
A **two-way table** shows data on the relationship between two categorical variables for some group of individuals.
There are 3 types of *relative frequencies*:

1. **Marginal relative frequency** is the proportion of individuals with a specific value for one categorical variable.
$$\text{marginal relative frequency} = \frac{n}{N}$$
2. **Joint relative frequency** is the proportion of individuals with a specific value for one categorical variable as well as a specific value for another categorical variable.
$$\text{joint relative frequency} = \frac{n_{A \wedge B}}{N}$$
3. **Conditional relative frequency** is the proportion of individuals that have a specific value for one variable that also have a specific value for another variable.
$$\text{conditional relative frequency} = \frac{n_A}{n_B}$$

If knowing the value of one variable helps in predicting that of another, there is an *association* between them. To test for association, the marginal and conditional relative frequencies can be compared. If they vary *significantly*, there is an association.

Part III

Collecting Data

Chapter 5

Sampling

A **population** consists of every **individual** that is in a defined group, while a **sample** is a subset of a population of interest.
A **study** refers to a **sample** or an **experiment**. Studies involving humans must be screened by an **institutional review board** before they can happen. All participants in said studies must give their **informed consent** prior to their participation. Any information regarding specific individuals must be kept **confidential**.
Observational studies observe individuals, measuring *variablesvariable* of interest but not attempting to influence a response.
Sampling attempts to gain information regarding a population by studying subgroups. A **census**, on the other hand, attempts to gain information regarding all individuals within an area of interest.
Causality occurs when one variable's value results in that of another.
Generalizability is the ability for a study's results to be generalized to the population at large.

Sample Designs and Bias

The **sample design**, how the sampling is carried out, is crucial to take into account when attempting to collect data that is **representative** of the population of interest.
Bias is the systematic skewing of results.
Voluntary response bias occurs when data is only collected by those that want to have their data collected. Those with strong opinions are more likely to want to be heard, so they more actively respond. This tends to result in a negative bias.
Under-coverage bias occurs when some portion of the population is left out.
Nonresponse bias occurs when some individuals that are chose elect not to participate.
Response bias occurs when respondents change their results due to the sampling design.[1].
Convenience sampling fails to use randomness. While this makes data easier to obtain, that data will likely be unrepresentative of the population.
The fact that different random samples of the same size from the same population may produce different estimates is called **sampling variability**. This is reduced by increasing the sample size.
Inference (generalization of results) regarding a population requires that all individuals taking part in a sample be randomly selected from the population.
Evidence of *causality* requires a strong association that consistently appears across many studies.
Observed results that are too improbable to be explained by chance alone are **statistically significant**.

[1]If questions being asked in a survey are too complicated, responses will likely be unrepresentative

Random Sampling

To eliminate some potential bias, chance can be utilized in choosing the sample. The simplest way to do this is to use a **simple random sample** (**SRS** or probability design). An SRS of size n consists of n individuals from the population chosen such that each individual has an equal chance of being chosen.
To create an SRS, numerical values can be assigned[2] to each individual in the population such that each number has the same number of digits. The first number should be either 1 or 0 with the appropriate number of leading zeros to account for the total number of individuals. Numbers can then be randomly selected.
Systematic random sampling follows the same rules for numerical assignment as simple random sampling, but rather than randomly selecting n numbers, a single number from the minimum to k (or $k-1$ if the minimum is 0) is selected, where k is the population size divided by n, and each number that is the sum of that number and an integer multiple of k is used.
When the population is large and diversified among many categories, a **stratified random sample** can be used. The population is divided into non-overlapping subgroups called **strata**, each of which is subjected to an SRS before they are recombined. When each stratum is **homogenous**, individuals within the same stratum having similar values, and strata are different from each other, stratified random sampling is preferable to simple random sampling.
Cluster sampling divides the population into groups of people that are geographically close to each other called *clusters*. When each cluster is **heterogenous**, individuals within the same cluster not having similar values, and all clusters are similar, cluster sampling can be used, saving time and money.

[2]A common method for assigning numbers to each member of a population is to assign them alphabetically.

Chapter 6

Experimentation

Experiments are studies that impose a **treatment**, the experimental condition applied, upon some group, called **experimental units** (or **subjects** if human), to observe the results. They often aim to show that a change in one variable, the **explanatory variable** or factors, causes a change in another variable, the **response variable**.
Many experiments combine several factors, so each treatment is made by combining specific values, called **levels**, of each factor.
Factors attempt to explain the results.
Experiments provide evidence for **causality**, which cannot be done by samples. They also control **lurking variables**, external variables that affect the response variable. Additionally, they enable the combination of several factors.
Confounding variables are variables that are tied together such that one's affects on the response variable cannot be distinguished from the other's. Well-designed experiments with random assignment enable **inference** regarding causality.

Experimental Design

A **comparative design** is one that compares two or more treatments. A **control group** is a group who's treatment is set up to be compared to the real treatments. They are given a **placebo**, a dummy treatment. If neither the subjects nor those measuring their treatments are aware of who is receiving what treatment, the experiment is **double-blind**, as is the case with many medical and behavioral experiments. If one group knows, it is **single-blind**.
When **randomization** is used to group subjects, the resultant groups should be similar in all respects prior to the application of treatments.
Control, keeping all variables apart from the treatment the same for all groups, helps to avoid confounding and reduces variation in responses, making it easier to determine a treatment's efficacy.
Each treatment should be imposed on enough experimental units that the effects of the treatments can be distinguished from chance differences between groups, ensuring **replicability**.
A **completely randomized design** assigns treatments to experimental units completely at random.
A **randomized block design** divides the experimental units into groups, referred to as **blocks**, that are similar with respect to a variable that is expected to affect the response. Within each block, responses are compared and combined with those of other blocks after accounting for differences between blocks.
A **matched pairs design** can be used to compare to two treatments. It may involve each subject receiving both in a random order or two similar subjects being paired and the treatments being randomly assigned within each pair.

Part IV

Probability, Random Variables, and Probability Distributions

Chapter 7

Probability

Probability is the long-run relative frequency. It must be between 0 and 1 (inclusive).
The short-term is unpredictable, but the long-term is predictable.
The **law of large numbers** states that as the number of trials approaches infinity, the **experimental** (observed) probability will converge to the **theoretical** (calculated) probability.
A **sample space** S is a list of all possible outcomes. It can be used in calculating theoretical probabilities when each outcome is equally likely. A **probability model** is a description of a random process. It is comprised of a sample space and a list of each outcome's corresponding probability.
An **event** is any collection of outcomes.
For a probability model to be valid, any individual event's probability must be within $[0, 1]$ and the sum of the probabilities of all outcomes must be equal to zero.
The **complement rule** states that the probability of some event not occurring, denoted by the superscript C above the event, is equal to 1 minus its probability of occurring.

$$P\left(A^C\right) = 1 - P(A)$$

It is quite clear that the converse of the complement rule also holds true, which means that a complement's complement is nothing but the original event.

$$P\left(\left(A^C\right)^C\right) = P(A)$$

The complements of binary relations should be noted:

$$(A < B)^C = A \geq B \quad (A > B)^C = A \leq B \quad (A = B)^C = A \neq B \quad (A < B < C)^C = (B \leq A) \vee (B \geq C)$$

The **intersection** of two events, denoted by a $\cap$ between them, occurs when both events occur. It follows the commutative property

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

The **union** of two events occurs when exactly one of the two events occurs. This is also commutative.

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

The **general addition rule** states that the probability of exactly one of two events occurring is equal to their sums of their probabilities of occurring by themselves minus the that of both occurring.

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

Complements can be applied to the unions and intersections of two events.[1]

$$P(A \cup B)^C = P(A \cap B) + P\left(A^C \cap B^C\right) \qquad P(A \cap B)^C = P(A \cup B) + P\left(A^C \cap B^C\right)$$

The probability of one event occurring and another not is equal to the difference between the probabilities of the first event occurring and both occurring.

$$P\left(A \cap B^C\right) = P(A) - P(A \cap B)$$

For two events to be **mutually exclusive** or **disjoint**, it must be impossible for both to occur.

$$P(A \cap B) = 0$$

Mutually exclusive events follow the **addition rule for mutually exclusive events**, which states that the probability of their intersection is equal to their sum.

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

The probability of one event occurring given that another has already occurred is a **conditional probability**. It is equal to the probability of both events occurring divided by that of the given event.[2]

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

This allows probability of the intersection of two events to be calculated.

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

The commutative property is not always followed by conditional probabilities.

$$\neg\Box[P(A|B) = P(B|A)]$$

Two events are **independent** if the one occurring does not affect the probability of the other occurring.

$$P(A|B) = P(A) \wedge P(B|A) = P(B)$$

Mutually exclusive events have no common outcomes while *independent* events are not affected by one of the events occurring.

The probability of the intersection of two independent events is simply the product of their probabilities.

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

If two events are not independent, they are **dependent**.

---

[1]The way that complements affect unions and intersections follows De Morgan's law:

$$\neg(A \vee B) = \neg A \wedge \neg B \qquad \neg(A \wedge B) = \neg A \vee \neg B$$

This truth is disguised by the fact that "union" in statistics means "symmetric difference" in logic.

$a := x \in A \wedge b := x \in B$

$$\begin{aligned}
(A \triangle B)^C &\equiv \neg(x \in A \oplus x \in B) \\
&\equiv \neg(a \oplus b) \\
&\equiv \neg((a \wedge \neg b) \vee (b \wedge \neg a)) \\
&\equiv \neg(a \wedge \neg b) \wedge \neg(b \vee \neg a) \\
&\equiv (\neg a \vee b) \wedge (\neg b \vee a) \\
&\equiv (\neg a \wedge (\neg b \vee a)) \vee (b \wedge (\neg b \vee a)) \\
&\equiv ((\neg a \wedge \neg b) \vee (\neg a \wedge a)) \vee ((b \wedge \neg b) \vee (b \wedge a)) \\
&\equiv (\neg a \wedge \neg b) \vee 0 \vee 0 \vee (b \wedge a) \\
&\equiv (x \notin A \wedge x \notin B) \vee (x \in B \wedge x \in A) \\
&\equiv (A^C \cap B^C) \cup (A \cap B)
\end{aligned}$$

$$\begin{aligned}
(A \cap B)^C &\equiv \neg(x \in A \wedge x \in B) \\
&\equiv \neg(a \wedge b) \\
&\equiv \neg a \vee \neg b \\
&\equiv (\neg a \wedge \neg b) \vee 0 \\
&\equiv (\neg a \vee \neg b) \vee (p \wedge (q \wedge \neg q)) \\
&\equiv (\neg a \vee \neg b) \vee ((a \wedge a) \wedge (a \wedge \neg b)) \\
&\equiv (\neg a \vee \neg b) \vee ((a \wedge b) \wedge (a \wedge \neg b)) \\
&\equiv (\neg a \vee \neg b) \wedge (((a \wedge b) \vee (a \wedge \neg a)) \wedge ((\neg b \wedge b) \vee (\neg b \wedge a))) \\
&\equiv (\neg a \wedge \neg b) \wedge (a \vee (b \wedge \neg a)) \wedge ((\neg b \vee (b \wedge \neg a))) \\
&\equiv (\neg a \wedge \neg b) \vee ((a \wedge \neg b) \vee (b \wedge \neg a)) \\
&\equiv (x \notin A \wedge x \notin B) \vee ((x \in A \wedge x \notin B) \vee (x \in B \wedge x \notin A)) \\
&\equiv (A \triangle B) \cup (A^C \cap B^C)
\end{aligned}$$

[2]On a two-way table, $P(A|B)$ is the intersection of A and B divided by the total of B.

Simulation

Chapter 8

Random Variables

Random variables take numerical values that describe the outcomes of some chance process.
A *probability distribution* describes all possible outcomes and their probabilities.
For a probability distribution to be valid, the sum of all probabilities must be one and all individual probabilities be within $[0, 1]$.
The probability of X being equal to x_i can be denoted as either $P(x_i)$ or as p_i.

$$P(X = x_i) = P(x_i) = p_i$$

The *mean* or **expected value** (denoted $E(X)$) is a weighted average of all possible values of X.

8.1 Discrete Random Variables

A **discrete random variables** have a fixed finite set of possible values with gaps between them.
The *mean* μ of a discrete random variable is equal to the sum of the products of each value and its frequency.[1]

$$\mu_X = \sum x_i p_i$$

The mean of a discrete random variable need not be one of its possible values.

The *variance* σ^2 of a discrete random variable is the sum of the products of the square of the distance from the mean multiplied by frequency.

$$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$

8.2 Continuous Random Variables

A **continuous random variable** can take on any value within an interval (potentially unbounded).
The probability of X falling within an interval (represented as a binary relation) is equal to the area under the variable's *density curve*[2] between the bounds of the region.[3]

[1] A random variable's sample space can be input into two lists, one corresponding to its values and the other to their frequencies. From there, `1-Var Stats` can be run the the former input as the `List` and the latter as the `FreqList`. This provides the values of μ_X, σ_X, and σ_X^2.
[2] See section 1.3 for more on density curves
[3] The probability of a continuous variable is, of course, an integral.

$$P(a < X < b) = \int_a^b f_X(x)\,\mathrm{d}x$$

As the number of values that a continuous variable may adopt is uncountably infinite, the probability of X being any particular value is always equal to 0.

$$P(X = x) = 0$$

As such, the inclusivity of the bounds of the interval is irrelevant.

$$P(x_1 \leq X \leq x_2) = P(X = x_1) + P(x_1 < X < x_2) + P(X = x_2) = P(x_1 < X < x_2)$$

A density curve's *mean* is the value of at which it would be balanced.[4]

8.3 Combinations and Linear Transformations of Random Variables

Two *independent* variables can be combined by using all possibilities for the sum or difference of x_i and y_i and the set of possible values for a new variable Z. The probability of Z being a particular value z_i is simply the product of the corresponding probabilities of x_i and y_i.[5]
When adding or subtracting two independent random variables, the means are added or subtracted from each other and the variances are always added.[6]

$$\mu_{X\pm Y} = \mu_X \pm \mu_Y \qquad\qquad \sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

It should be noted that *standard deviation* must must be derived from variance.

$$\sigma_{X\pm Y} \neq \sigma_X \pm \sigma_Y$$

A **linear transformation** takes the form of a constant added to the product of a constant and a variable.

$$Y = a + bX$$

The mean can simply be plugged in to this formula to find the new mean while the standard deviation is only multiplied.

$$\mu_Y = a + b\mu_X \qquad\qquad \sigma_Y = b\sigma_X$$

[4]The mean and variance of a continuous variable are defined just like that of a discrete one, only using integrals in place of sums and $f_X(x)$ rather than p_i.

$$\mu_X = \sum x_i p_i = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \qquad\qquad \sigma^2_X = \sum (x_i - \mu_X)^2 p_i = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\,\mathrm{d}x$$

[5]A variable Z and its corresponding sample space can be defined as a combination of X and Y as such:

$$Z = \{x \pm y \mid (x, y) \in X \times Y\} \qquad\qquad S_Z = \{(x \pm y, p(x) \times p(y)) \mid (x, y) \in X \times Y\}$$

[6]Means can be added or subtracted correspondingly with a combination of random variables regardless of independence.

Chapter 9

Probability Distributions

9.1 Binomial Distributions

For a *discrete* random variable to be **binomial**, it must meed the following conditions:

1. The possible outcomes of each trial must be **binary**, meaning that each trial can either be a success or a failure.
2. Each trial must be *independent*.
3. The number of trials must be fixed in advance.
4. The probability of success must be constant between trials.

A binomial distribution with parameters n trials and probability p of any given trial being successes can be denoted $B(n, p)$.
The probability of x successes in n trials, each with a probability p of being successful, is defined as such:[1]

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is the product of the probability of x successes, that of $n - x$ failures (necessitated by the definition of a binomial), and the total number of possibilities in which there are x successes.[2]
The **sample space** of a binomial variable can be derived as each term of a binomial expansion[3] of the conjugate of p added to p itself.

$$S_{B(n,p)} = ((1-p)+p)^n = \sum_{x=0}^{n} \binom{n}{x} (a-p)^{n-x} p^x$$

[1]The binomial $\binom{n}{x}$ is defined as such:

$$\binom{n}{x} = \frac{n!}{(n-x)!}$$

[2]The probability of exactly and at most x successes can be found using `binompdf` and `binomcdf` respectively:

$$P(x) = \texttt{binompdf}\,(n : n, p : p, x : x) \qquad P(X < x) = \texttt{binomcdf}\,(n : n, p : p, x : x)$$

[3]The binomial expansion of $a + b$ raised to power n is the following:

$$(a+b)^n = \sum_{n=0}^{n} \binom{n}{k} a^{n-k} b^k$$

A binomial variable's *expected value* (*mean*) is equal to the product of the number of trials and the probability of success while its standard deviation is the product of the probability of success and the number of trials.[4]

$$\mu_X = np$$

The *variance* of a binomial variable is the product of the number of trials and the probability of success and its conjugate.

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

9.2 Geometric Distributions

[4]The probability is simply the long-run relative frequency, so the expected number of successes for a binomial distribution is simply the expected relative frequency of successes multiplied by the total number of trials.

Part V

Sampling Distributions

Chapter 10

Sampling Distributions

10.1 Sampling Distributions of Proportions

Sampling Distributions of Differences in Proportions

10.2 Sampling Distributions of Means

Sampling Distributions of Differences in Means

Part VI

Inference

Chapter 11

Confidence Intervals

A statistic is a **point estimator** used to estimate an unknown population parameter. A **point estimate** is a statistic's value, and is referred to as such because it is a single point. As such, it is almost never accurate.
A point estimate can be made made more reliable by either increasing the sample size or using a more accurate sampling procedure.
The **standard error** $\boldsymbol{s}$ (with the appropriate subscript denoting its statistic) of a statistic is the point estimate of the standard deviation of the sampling distribution.
A **confidence interval** provides an interval of plausible values for an unknown parameter based on sample data. It is equal to the point estimate plus or minus the **margin of error** (**ME**).

$$\text{confidence interval} = \text{point estimate} \pm \text{margin of error}$$

The probability that a confidence interval contains the true parameter value is the confidence interval's **confidence level** ($\boldsymbol{C}$). The margin of error describes the maximum deviation of the estimate from the parameter. It is the product of the critical value and the standard error of the statistic.

$$\text{ME} = \text{critical value} \times \text{standard error}$$

The **critical value** is equal to the the number of standard deviations from the mean within which the probability of a random variable falling is equal to C.

$$P(-\text{critical value} < \text{standardized test statistic} < \text{critical value}) = C$$

11.1 Confidence Intervals about Proportions

Confidence Intervals about Differences in Proportions

11.2 Confidence Intervals about Means

In order for Normality to be verified, the central limit theorem can be used, necessitating that n be at least 30, or a modified box plot can be created with the data and observed to be symmetrical without outliers. When the standard deviation of X (not of the sampling distribution of $\bar{x}$) is known, a confidence interval about $\bar{x}$ can be constructed using the templates for confidence intervals and margin of error, simply using the sampling distribution's standard deviation rather than the statistic's standard error.

$$\text{confidence interval} = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

This is a **one-sample *z*-interval for a population mean**. When σ is not known, the standard deviation can be replaced by the standard error to calculate the **standard error of $\bar{x}$**.

$$s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

The margin of error is appropriately changed:

$$\text{ME} = z^* \frac{s_x}{\sqrt{n}}$$

This results in the intervals being too small, though, and the confidence level decreases from what would be expected given z^*. The critical value can also change, though, becoming t^* rather than z^*, making the intervals longer and making the confidence level representative. The reason that t is used is that a ***t* distribution** is used rather than a Normal one. (t^* can still be interpreted in the same way as z^*, though (the number of standard errors from the point estimate).)
The specific t distribution used is specified by **degrees of freedom** (**df**), which is 1 less than the sample size.[1]

$$\text{df} = n - 1$$

The confidence interval constructed about $\bar{x}$ using t^* and s_x is a **one-sample *t* interval for a population mean**.

$$\text{ME} = t^* \frac{s_x}{\sqrt{n}}$$

Because t^* is dependent on df, which is 1 less than the sample size, and s_x is dependent on the data, which has not been produced, so the sample size cannot be solved for given a confidence level and the margin of error, so z^* and σ, a value of s_x from a previous study are instead used.

$$\text{ME} \geq z^* \frac{\sigma}{\sqrt{n}}$$
$$n \geq \left(\frac{\text{ME}}{z^*} \right)^2$$

Confidence Intervals about Differences in Means

A confidence interval about a difference in means is a **two-sample *t* interval for a mean difference**. In order for it be constructed about, Normality and independence must be justified and both samples must be independent.
The center of the confidence interval is the difference between the sample means, denoted by a subscript **diff**, while its standard error is simply the square root of the sum of the variances of the standard errors of the individual statistics.

$$\bar{x}_{\text{diff}} = \overline{x_1 - x_2} = \bar{x}_1 - \bar{x}_2 \qquad \text{confidence interval} = \bar{x}_{\text{diff}} \pm s_{\bar{x}_1 - \bar{x}_2} = (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

[1] The density curve of a t distribution with degrees of freedom ν is defined (using the gamma function Γ or the beta function B) as such:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} = \frac{1}{\sqrt{\nu}\,\text{B}\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

As df approaches infinity, the t distribution approaches a normal curve (the tails approaching 0 more quickly), so t^* approaches z^*. This is because a greater sample size means that s_x will be closer to σ.

If both standard deviations are known, they are used along with z^*.

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The degrees of freedom of a difference of means is the equal to the estimated variance divided by the sum of the estimated variances of each statistic divided by 1 less than their sample sizes.

$$\text{df} = \frac{s_{\bar{x}_1 - \bar{x}_2}^4}{\frac{s_{\bar{x}_1}^4}{n_1 - 1} + \frac{s_{\bar{x}_2}^4}{n_2 - 1}} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

When the data is *paired*, the response being collected from the same set of individuals independently, the difference can be treated as a single mean, as each sample difference is known, and a **paired *t* interval** can be created. s_{diff} can therefore be used as well.[2]

$$s_{\text{diff}} = s_{1-2} = \sqrt{s_1^2 + s_2^2} \qquad \text{confidence interval} = \bar{x}_{\text{diff}} \pm t^* \frac{s_{\text{diff}}}{\sqrt{n}}$$

[2]When $n_1 = n_2$, it can be verified that the values of $s_{\bar{x}_1 - \bar{x}_2}$ derived by s_{diff} and by s_1 and s_2 individually are the same.

$$\begin{aligned} s_{\text{diff}} &= \sqrt{s_1^2 + s_2^2} \\ s_{\bar{x}_1 - \bar{x}_2} &= \frac{s_{\text{diff}}}{\sqrt{n}} & &= \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}} \\ &= \sqrt{\frac{s_1^2 + s_2^2}{n}} & &= \sqrt{\frac{s_1^2 + s_2^2}{n}} \end{aligned}$$

Chapter 12

Significance Tests

A **significance test** is a procedure that uses observed data to test between two claims, often made regarding parameters, about hypotheses.

> In order for a significance test to be performed, randomness, independence, and Normality must be verified.

The **null hypothesis** ($\boldsymbol{H_0}$) claims that the parameter is equal to a **null value**, what it was previously assumed to be, denoted by a subscript 0 on the parameter. It is often a statement of no change or difference.

$$H_0 : \text{parameter} = \text{null value}$$

The claim that is attempting to be supported is the **alternative hypothesis** ($\boldsymbol{H_a}$). It can either be **one-sided**, claiming that the parameter is greater or less than the null value, or **two-sided**, claiming simply that the parameter is not equal to the null value.

$$H_a : \text{parameter} \gtrless \text{null value} \vee \text{parameter} \neq \text{null value}$$

A test's ***P*-value** is the probability of *significant* evidence being found that supports H_a that is at least as strong as that observed assuming that H_0 is true.

$$P\text{-value} = P(\text{statistic supports } H_a \mid H_0)$$

The smaller the P-value, the lower the chances of receiving evidence of the alternative. A small P-value therefore supports the H_a.
If the P-value is less than the **significance level** $\boldsymbol{\alpha}$, H_0 can be rejected and it can be concluded that there is convincing evidence for H_a. If the P-value is greater than or equal to α, H_0 cannot be rejected, and it can be concluded that there is not convincing evidence for H_a.
The P-value is calculated using the **standardized test statistic**, which is the number of standard deviations from the null value of the parameter.
For a null hypothesis to be significant is for a significance test to provide a P-value less than the significance level.

$$\text{standardized test statistic} = \frac{\text{statistic} - \text{null parameter}}{\text{standard error of statistic from null parameter}} = \text{s}$$

The P-value is equal to the probability of z satisfying the H_a assuming that H_0 is true. It can therefore be calculated as such using a *cumulative distribution function*, so long as Normality and independence are justified.

$$P\text{-value} = \begin{cases} P(\text{parameter} \gtrless \text{null parameter}) = P(\text{S} \gtrless \text{s}) & H_a : \text{parameter} \gtrless \text{null parameter} \\ P(|\text{parameter}| < |\text{null parameter}|) = P(\text{S} < -|\text{s}|) + P(\text{S} > |\text{s}|) & H_a : \text{parameter} \neq \text{null parameter} \end{cases}$$

Conclusions should only ever be made regarding the rejection of H_0 and the convincing support of H_a. H_0 should never be supported and H_a should never be rejected.

When answering a question regarding a significance test, the following process can be followed:

1. State State the hypotheses to be tested and the significance level and define any parameters.

2. Plan Identify the appropriate methods of inference and verify its conditions.

3. Do State the sample statistic(s), calculate the standardized test statistic(s), and calculate the P-value.

4. Conclude Make a conclusion regarding the hypotheses within the problem's context.

When performing significance tests, two types of errors may occur:

- A **Type I error** occurs when H_0 is rejected despite H_a being false; the data provided convincing evidence for H_a despite it being incorrect.
- A **Type II error** occurs when H_0 is not rejected despite H_a being true; the data did not provide convincing evidence for H_a despite it being correct.

	H_a is false	H_a is true
H_0 is rejected	Type I error	Correct conclusion
H_0 is not rejected	Correct conclusion	Type II error

The probability of a Type I error occurring is equal to α.
As α increases, the probability of a Type I error increases but that of a Type II error decreases.
A confidence interval for $\hat{p}$ (using $s_{\hat{p}}$) can be used in tandem with a sample statistic to provide a set of plausible values for the true parameter, should the alternative hypothesis be convincingly supported.
A two-sided test of of H_0 at significance level α usually provides the same conclusion as a confidence level of the complement of α.

$$[P(\mathrm{S} \neq \mathrm{s}) < \alpha] \approxeq [\text{null parameter} \in (\text{statistic} \pm \mathrm{ME})]$$

A test's **power** is the probability of convincing evidence being found that convincingly supports H_a given a value for the parameter being tested. This is also is equal to the probability of avoiding a Type II error.

$$\text{power} = 1 - P(\text{Type II Error})^C = P(\text{statistic convincingly supports } H_a \mid H_a \text{ is true})$$

Power can be increased in three ways:

1. Increasing the sample size
    - A large sample means that more data is collected and more information is given regarding the true population parameter. This also increases n, which decreases the standard error of the statistic, reducing the value of the standardized test statistic and therefore the P-value, making it more likely to fall below α.
2. Increasing the significance level
    - Increasing the significance level increases the probability of H_0 being rejected when H_a is true, as the maximum P-value for H_0 to be rejected increases.
3. Increasing the **effect size**, the minimum difference between the null parameter value and the alternative parameter value for the change to matter
    - Increasing the size of the difference that needs to be detected makes that difference more likely to be detected, as larger differences are easier to detect.

12.1 Significance Tests about Proportions

In order for a significance test of $H_0 : p = p_0$ to be performed, it must be verified that the distribution of $\hat{p}$ is approximately Normal assuming H_0 and that the standard error can be calculated, so Large Counts and the 10% condition (not for experiments) must be satisfied and interpreted for Normality and independence respectively.

To perform **1-proportion *z* test**, a significance test about one proportion, z must be calculated.

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

Significance Tests about Differences in Proportions

A **2-proportion *z* test** can be performed to compare the proportions for two populations is based on the difference between sample proportions.

$$H_0 : p_1 - p_2 = p_0 \qquad H_a : p_1 - p_2 \gtrless p_0 \qquad H_a : p_1 - p_2 \neq p_0$$

Typically, p_0 is 0, so these hypotheses can be rewritten.

$$H_0 : p_1 = p_2 \qquad H_a : p_1 \gtrless p_2 \qquad H_a : p_1 \neq p_2$$

A significance test first assumes that the null hypothesis $H_0 : p_1 = p_2$ is true. This common value is referred to as p.
The **combined sample proportion** is denoted $\hat{p}_C$ and is equal to the total successes divided by the total sample size, making it effectively a weighted average. It is the sample proportion that assumes that the parameter values are equal.

$$\hat{p}_C = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

The Large Counts condition must be met with $\hat{p}_C$.

$$n_1\hat{p}_C, n_1(1-\hat{p}_C), n_2\hat{p}_C, n_2(1-\hat{p}_C) \geq 10$$

For a significance test to be run about a difference of proportions, the randomness, independence (10%) (for each proportion), and Large Counts conditions must be met.

The *standardized test statistic* is the z-score calculated using the difference in proportions and its standard error assuming the mean to be 0 (H_0 to be true).

$$z = \frac{\hat{p}_1 - \hat{p}_2 - \mu_{\hat{p}_1 - \hat{p}_2}}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_C\left(1-\hat{p}_C\right)\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}}$$

12.2 Significance Tests about Means

Significance Tests about Differences in Means

To perform a significance test for a population mean, a **1-sample t test** randomness, independence (10%), and Normality (*CLT* or distribution) must be verified.

The *standardized test statistic* for a significance test about a mean is t.

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{s_x/\sqrt{n}}$$

The *t-distribution* used to calculate the P-value uses *degrees of freedom* 1 less than the sample size.

$$\text{df} = n - 1$$

Minute, practically unimportant changes in in μ_0 can drastically shrink the P-value when the sample size is always large enough. A very large sample size results in the null hypothesis almost always being rejected. ***P*-hacking** takes advantage of this fact.

Chapter 13

Chi-Square Tests

Chapter 14

Slopes

Part VII

Back Matter

Chapter 15

Index