AP Statistics

Arnav Patri

March 13, 2022

Contents

I One-Variable Data **2**

1 One-Variable Quantitative Data **3**
1.1 Graphically Displaying Distributions . . . 3
1.2 Numerically Describing Distributions . . . 4
1.3 Density Curves . . . 5
1.4 The Normal Distribution . . . 6

2 One-Variable Categorical Data **9**

II Two-Variable Data **10**

3 Two-Variable Quantitative Data **11**
3.1 Correlation . . . 11
3.2 Regression and Residuals . . . 12
3.3 Transformations to Achieve Linearity . . . 13

4 Two-Variable Categorical Data **15**

III Collecting Data **16**

5 Sampling **17**

6 Experimentation **19**

IV Probability, Random Variables, and Probability Distributions **20**

7 Probability **21**

8 Random Variables and Probability Distributions **23**

V Sampling Distributions **24**

VI Inference **25**

9 Confidence Intervals **26**

10 Significance Tests **27**

11 Chi-Square Tests **28**

12 Slopes **29**

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. 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 uncountably infinite set of possible values.

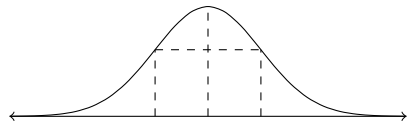
The quantitative variable being described must always be defined, typically as a capital letter. An arbitrary particular value is denoted with a lowercase letter and a superscript is added to denote a defined value.

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 (inclusive). Rather than each class being represented as a range, they are labeled by their medians.

Graphically Describing Distributions

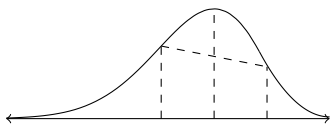
A distribution can be described by its shape, outliers, center, and spread (SOCS).

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

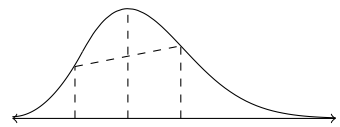


Symmetric

Skew is dependent on whether the tail is 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. 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, 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 (not necessarily unique) 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 $\bar{x}$).

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

The first, second, and third **quartiles**, denoted Q_1, Q_2, and Q_3, are the values with 25%, 50%, and 75% of the data below them respectively.[1] The second quartile is the **median**, and is, along with the mean, a measure of center.
The **range** is the difference between the highest and lowest values of a data set.
The **interquartile range**, denoted IQR, is the difference between the values of the third and first quartiles.[2] It therefore shows the "middle half" of the distribution.

$$IQR = Q_3 - Q_1$$

The **standard deviation** is the average distance from the mean. It is denoted by 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.

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

[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.

It is crucial to note that the mean and standard deviation are only to be used when there are no outliers[a] due to their sensitivity.

[a]An **outlier** is any value that varies from the middle 50% by over 1.5 multiplied by the interquartile range.

$$(|Q_1 - x_i| \vee |x_i - Q_3|) > 1.5IQR \implies x_i \text{ is an outlier}$$

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

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

$$y_i = ax_i + b$$

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 curve**, typically denoted f, displays the relative frequencies of every possible value of its variable. As such, the total area under the curve must be exactly 1 and its values must always be positive.

A density curve is a **continuous** distribution, so the probability of X being any particular value is zero. The function's actual values can be disregarded.

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

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

The area under the density curve from its leftmost value to any given value is that value's **percentile**. It is the percentage of the data that that particular value exceeds, which means that it is also equal to the probability of the value being below it, and can be found using a cumulative distribution function, generally notated F.[4]

$$\text{Percentile} = P(X < x)$$

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 $\int_{x_{\min}}^{x} f(t)\,dt$. As dt is a differential, it is infinitesimal, so adding or subtracting $f(x)\,dt$ 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.

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

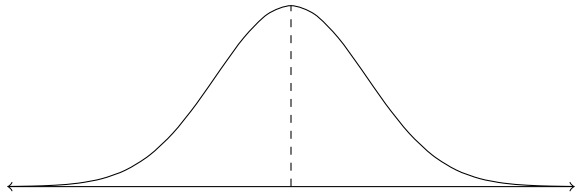
[3]The values of n, $\sum x_i$, $\bar{x}$, Q_1, Q_2, Q_3, range, IQR, s_x, and s_x^2 are all calculable automatically given a data set. The data can be entered into `L`n (`Stat/EDIT/1`), and `1-Var Stats` (`Stat/CALC/1`) can be performed with `L`n (`ALPHA/`n) as its parameter.

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

$$F(x) = \int_{x_{\min}}^{x} f(t)\,dt$$

The median of a density curve is the point that splits the curve into two regions of equal area[5] while its mean is the point at which the entire curve would be balanced[6].

1.4 The Normal Distribution



The Normal Distribution

The **normal** distribution (typically denoted φ) is a density curve that is bell-shaped and symmetrical with inflection points one standard deviation from the mean.[7] Its mean and standard deviation are 0 and 1 respectively.
For a curve to be normal, it must be some transformation[8] 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 $X \sim \mathcal{N}(\mu, \sigma)$.[9]
To justify the normality of a data set, a modified box plot can be created and symmetry and a lack of variance shown.
Percentile can be found using the **cumulative normal distribution function** Φ.[10]

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

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

[6] The mean of the curve is the x-value of the centroid of $f(x)$ over its entire domain.

$$\mu = \frac{\int_{x_{\min}}^{x_{\max}} x f(x)\, dx}{\int_{x_{\min}}^{x_{\max}} f(x)\, dx}$$

[7] The normal distribution is defined as such:

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

[8] 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").

[9] A normal distribution of X with parameters μ and σcan also be denoted as such:

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

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

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

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:

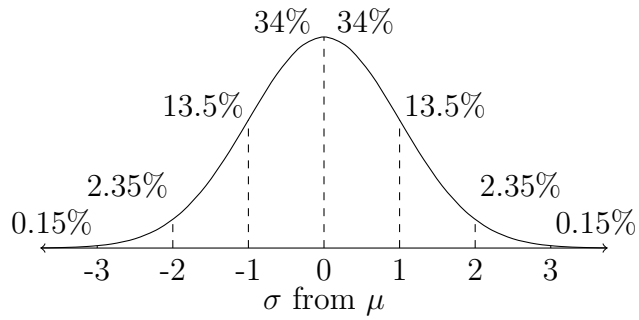
$$P(x_1 < X < x_2) = \texttt{normalcdf}\,(\text{lower} : x_1, \text{upper} : x_2, \mu : \mu, \sigma : \sigma)$$

The **inverse normal function** Φ^{-1} takes a percentile as an input and returns a value.[11]

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

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.[12]

Standardization and z-Scores

The ***z*-score** is the number of standard deviations away from the mean.[13]

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

z-scores allow the standard normal distribution to always be used[14], 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.

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

1. Define the variable its distribution, using the given parameters (and justifying normality).

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

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

[12] 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 `sto` $\rightarrow \texttt{L}_n$

[13] 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 which case the latter would be uppercase, as in Z_X, z_X, and $z_{x,i}$

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

(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.

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

Chapter 2

One-Variable Categorical Data

Categorical variables assign labels that place each individual into one of several groups, while **quantitative variables** provide values that describe or measure some characteristic.
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 fraction 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, from can be described as linear or nonlinear. **Strength** describes how strongly correlated the variables are.
Association direction describes how the explanatory

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 agains x (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 nor 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 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:

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

Outliers are points with large residuals.

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**.

$$\boxed{\text{Extrapolation should be avoided, as it is more likely not to be accurate than interpolation.}}$$

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

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

The definition of a makes it clear 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.

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

[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.

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

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

[6] Give 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 $\hat{y} = \bar{y}$.

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

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

$$\boxed{\text{Exponentiating one variable and taking the logarithm of the other are equivalent operations.}}$$

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.
A **marginal relative frequency** is the proportion of individuals with a specific value for one categorical variable. A **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.
A **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.
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 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 **variables** 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.

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 non-representative

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.

The complements of binary operators should be noted:

$$(A < B)^C = A \geq B \qquad (A > B)^C = A \leq B \qquad (A = B)^C = a \neq B$$

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

In order for two events to be **mutually exclusive** (**disjoint**), it must be impossible for both of them to occur. 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 apply to the complements of 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 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 is disguised by the fact that "union" in statistics means "symmetric difference" in logic.

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

Chapter 8

Random Variables and Probability Distributions

Part V

Sampling Distributions

Part VI

Inference

Chapter 9

Confidence Intervals

Chapter 10

Significance Tests

Chapter 11

Chi-Square Tests

Chapter 12

Slopes