Section summary

- The **population** is the entire group that the researchers are interested in. Because it is usually too costly to gather the data for the entire population, researchers will collect data from a **sample**, representing a subset of the population.
- A **parameter** is a true quantity for the entire population, while a **statistic** is what is calculated from the sample. A parameter is about a population and a statistic is about a sample. Remember: *p goes with p and s goes with s.*
- Two common summary quantities are **mean** (for numerical variables) and **proportion** (for categorical variables).
- Finding a good estimate for a population parameter requires a random sample; do not generalize from anecdotal evidence.
- There are two primary types of data collection: observational studies and experiments. In an **experiment**, researchers impose a treatment to look for a causal relationship between the treatment and the response. In an **observational study**, researchers simply collect data without imposing any treatment.
- Remember: *Correlation is not causation*! In other words, an association between two variables does not imply that one causes the other. Proving a causal relationship requires a well-designed experiment.

Section summary

- In an **observational study**, one must always consider the existence of **confounding factors**. A confounding factor is a "spoiler variable" that could explain an observed relationship between the explanatory variable and the response. Remember: For a variable to be confounding it must be associated with both the explanatory variable *and* the response variable.
- When taking a sample from a population, avoid **convenience samples** and **volunteer samples**, which likely introduce bias. Instead, use a **random** sampling method.
- Generalizations from a sample can be made to a population only if the sample is random. Furthermore, the generalization can be made only to the population from which the sample was randomly selected, not to a larger or different population.
- Random sampling from the entire population of interest avoids the problem of **undercoverage bias**. However, **response bias** and **non-response** bias can be present in any type of sample, random or not.
- In a **simple random sample**, every *individual* as well as every *group of individuals* has the same probability of being in the sample. A common way to select a simple random sample is to number each individual of the population from 1 to N. Using a random digit table or a random number generator, numbers are randomly selected without replacement and the corresponding individuals become part of the sample.
- A **systematic random sample** involves choosing from of a population using a random starting point, and then selecting members according to a fixed, periodic interval (such as every 10th member).
- A **stratified random sample** involves randomly sampling from *every* **strata**, where the strata should correspond to a variable thought to be associated with the variable of interest. This ensures that the sample will have appropriate representation from each of the different strata and reduces variability in the sample estimates.
- A **cluster random sample** involves randomly selecting a set of **clusters**, or groups, and then collecting data on all individuals in the selected clusters. This can be useful when sampling clusters is more convenient and less expensive than sampling individuals, and it is an effective strategy when each cluster is approximately representative of the population.
- Remember: *Individual strata should be homogeneous (self-similar), while individual clusters should be heterogeneous (diverse).* For example, if smoking is correlated with what is being estimated, let one stratum be all smokers and the other be all non-smokers, then randomly select an appropriate number of *individuals* from *each* strata. Alternately, if age is correlated with the variable being estimated, one could randomly select a *subset* of clusters, where each cluster has mixed age groups.

Section summary

- In an **experiment**, researchers impose a **treatment** to test its effects. In order for observed differences in the response to be attributed to the treatment and not to some other factor, it is important to make the treatment groups and the conditions for the treatment groups as similar as possible.
- Researchers use **direct control**, ensuring that variables that are within their power to modify (such as drug dosage or testing conditions) are made the *same* for each treatment group.
- Researchers **randomly** assign subjects to the treatment groups so that the effects of uncontrolled and potentially confounding variables are *evened out* among the treatment groups.
- **Replication**, or imposing the treatments on many subjects, gives more data and decreases the likelihood that the treatment groups differ on some characteristic due to chance alone (i.e. in spite of the randomization).
- An ideal experiment is **randomized**, **controlled**, and **double-blind**.
- A **completely randomized experiment** involves randomly assigning the subjects to the different treatment groups. To do this, first number the subjects from 1 to N. Then, randomly choose some of those numbers and assign the corresponding subjects to a treatment group. Do this in such a way that the treatment group sizes are balanced, unless there exists a good reason to make one treatment group larger than another.
- In a **blocked experiment**, subjects are first separated by a variable thought to affect the response variable. Then, within *each* block, subjects are randomly assigned to the treatment groups as described above, allowing the researcher to compare like to like within each block.
- When feasible, a **matched-pairs experiment** is ideal, because it allows for the best comparison of like to like. A matched-pairs experiment can be carried out on pairs of subjects that are meaningfully paired, such as twins, or it can involve all subjects receiving both treatments, allowing subjects to be compared to *themselves*.
- A treatment is also called a **factor** or explanatory variable. Each treatment/factor can have multiple **levels**, such as yes/no or low/medium/high. When an experiment includes many factors, multiplying the number of levels of the factors together gives the total number of treatment groups.
- In an experiment, blocking, randomization, and direct control are used to *control for confounding factors.*

Section summary

- A **scatterplot** is a **bivariate** display illustrating the relationship between two numerical variables. The observations must be **paired**, which is to say that they correspond to the same case or individual. The linear association between two variables can be positive or negative, or there can be no association. **Positive association** means that larger values of the first variable are associated with larger values of the second variable. **Negative association** means that larger values of the first variable are associated with smaller values of the second variable. Additionally, the association can follow a linear trend or a curved (nonlinear) trend.
- When looking at a **univariate** display, researchers want to understand the distribution of the variable. The term **distribution** refers to the values that a variable takes and the frequency of those values. When looking at a distribution, note the presence of clusters, gaps, and **outliers**.
- Distributions may be **symmetric** or they may have a long tail. If a distribution has a long left tail (with greater density over the higher numbers), it is **left skewed**. If a distribution has a long right tail (with greater density over the smaller numbers), it is **right skewed**.
- Distributions may be **unimodal**, **bimodal**, or **multimodal**.
- Two graphs that are useful for showing the distribution of a small number of observations are the **stem-and-leaf plot** and **dot plot**. These graphs are ideal for displaying data from small samples because they show the exact values of the observations and how frequently they occur. However, they are impractical for larger data sets.
- For larger data sets it is common to use a **frequency histogram** or a **relative frequency histogram** to display the distribution of a variable. This requires choosing bins of an appropriate width.
- To see cumulative amounts, use a **cumulative frequency histogram**. A **cumulative relative frequency histogram** is ideal for showing **percentiles**.
- **Descriptive statistics** describes or summarizes data, while **inferential statistics** uses samples to generalize or infer something about a larger population.

Section summary

- In this section we looked at univariate summaries, including two measures of **center** and three measures of **spread**.
- When **summarizing** or **comparing distributions**, always comment on center, spread, and shape. Also, mention outliers or gaps if applicable. Put descriptions in *context*, that is, identify the variable(s) being summarized by name and include relevant units. Remember: *Center, Spread, and Shape! In context!*
- **Mean** and **median** are measures of center. (A common mistake is to report **mode** as a measure of center. However, a mode can appear anywhere in a distribution.)
  - The **mean** is the sum of all the observations divided by the number of observations, n. $\bar{x} = \frac{1}{n}\sum x_i = \frac{\sum x_i}{n} = \frac{x_1+x_2+\ldots+x_n}{n}$
  - In an ordered data set, the **median** is the middle number when n is odd. When n is even, the median is the average of the two middle numbers.
- Because large values exert more "pull" on the mean, large values on the high end tend to increase the mean more than they increase the median. In a **right skewed** distribution, therefore, the mean is greater than the median. Analogously, in a **left skewed** distribution, the mean is less than the median. Remember: *The mean follows the tail! The skew is the tail!*
- **Standard deviation (SD)** and **Interquartile range (IQR)** are measures of spread. SD measures the typical spread from the mean, whereas IQR measures the spread of the middle 50% of the data.
  - To calculate the standard deviation, subtract the average from each value, square all those differences, add them up, divide by $n-1$, then take the square root. Note: The standard deviation is the square root of the variance.
    $s_X = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$
  - The IQR is the difference between the third quartile Q_3 and the first quartile Q_1.
    $IQR = Q_3 - Q_1$
- **Range** is also sometimes used as a measure of spread. The range of a data set is defined as the difference between the maximum value and the minimum value, i.e. $max - min$.
- **Outliers** are observations that are extreme relative to the rest of the data. Two rules of thumb for identifying observations as outliers are:
  - more than 2 standard deviations above or below the mean
  - more than $1.5 \times IQR$ below Q_1 or above Q_3

  Note: These rules of thumb generally produce different cutoffs.
- Mean and SD are sensitive to outliers. Median and IQR are more robust and less sensitive to outliers.
- The **empirical rule** states that for normal distributions, about 68% of the data will be within one standard deviation of the mean, about 95% will be within two standard deviations of the mean, and about 99.7% will be within three standard deviations of the mean.
- **Linear transformations of data**. Adding a constant to every value in a data set shifts the mean but does not affect the standard deviation. Multiplying the values in a data set by a constant will multiply the mean and the standard deviation by that constant, except that the standard deviation must always remain positive.
- **Box plots** do not show the *distribution* of a data set in the way that histograms do. Rather, they provide a visual depiction of the **5-number summary**, which consists of: min, Q_1, Q_2, Q_3, max. It is important to be able to identify the median, IQR, and direction of skew from a box plot.

Section summary

- When an outcome depends upon a chance process, we can define the **probability** of the outcome as the proportion of times it would occur if we repeated the process an infinite number of times. Also, even when an outcome is not truly random, modeling it with probability can be useful.
- The **Law of Large Numbers** states that the **relative frequency**, or proportion of times an outcome occurs after n repetitions, stabilizes around the true probability as n gets large.
- The probability of an event is always between 0 and 1, inclusive.
- The probability of an event and the probability of its **complement** add up to 1. Sometime we use $P(A) = 1 - P(\text{not } A)$ when $P(\text{not } A)$ is easier to calculate than $P(A)$.
- A and B are **disjoint**, i.e. **mutually exclusive**, if they cannot happen together. In this case, the events do not overlap and $P(A \text{ and } B) = 0$.
- In the *special case* where A and B are **disjoint** events: $P(A \text{ or } B) = P(A) + P(B)$.
- When A and B are not disjoint, adding $P(A)$ and $P(B)$ will overestimate $P(A \text{ or } B)$ because the overlap of A and B will be added twice. Therefore, when A and B are not disjoint, use the **General Addition Rule**:
  $P(A \text{ or B}) = P(A) + P(B) - P(A \text{ and } B)$.[20]
- To find the probability that *at least one* of several events occurs, use a special case of the rule of **complements**: $P(\text{at least one}) = 1 - P(\text{none})$.
- When only considering two events, the probability that one *or* the other happens is equal to the probability that *at least one* of the two events happens. When dealing with more than two events, the General Addition Rule becomes very complicated. Instead, to find the probability that A or B or C occurs, find the probability that none of them occur and subtract that value from 1.
- Two events are **independent** when the occurrence of one does not change the likelihood of the other.
- In the *special case* where A and B are **independent**: $P(A \text{ and } B) = P(A) \times P(B)$.

[20] Often written: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Section summary

- A **conditional probability** can be written as $P(A|B)$ and is read, "Probability of A given B". $P(A|B)$ is the probability of A, given that B has occurred. In a conditional probability, we are given some information. In an **unconditional probability**, such as $P(A)$, we are not given any information.
- Sometimes $P(A|B)$ can be deduced. For example, when drawing without replacement from a deck of cards, $P(\text{2nd draw is an Ace} \mid \text{1st draw was an Ace}) = \frac{3}{51}$. When this is not the case, as when working with a table or a Venn diagram, one must use the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$.
- In the last section, we saw that two events are **independent** when the outcome of one has no effect on the outcome of the other. When A and B are independent, $P(A|B) = P(A)$.
- When A and B are **dependent**, find the probability of A *and* B using the **General Multiplication Rule**: $P(A \text{ and } B) = P(A|B) \times P(B)$.
- In the *special case* where A and B are **independent**, $P(A \text{ and } B) = P(A) \times P(B)$.
- If A and B are **mutually exclusive**, they must be **dependent**, since the occurrence of one of them changes the probability that the other occurs to 0.
- When sampling **without replacement**, such as drawing cards from a deck, make sure to use **conditional probabilities** when solving *and* problems.
- Sometimes, the conditional probability $P(B|A)$ may be known, but we are interested in the "inverted" probability $P(A|B)$. **Bayes' Theorem** helps us solve such conditional probabilities that cannot be easily answered. However, rather than memorize Bayes' Theorem, one can generally draw a tree diagram and apply the conditional probability rule $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$. The resulting answer often has the form $\frac{w \times x \,+\, y \times z}{w \times x}$, where w, x, y, z are numbers from a tree diagram.

Section summary

- $\binom{n}{x}$, the **binomial coefficient**, describes the number of combinations for arranging x successes among n trials. $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, where $n! = 1 \times 2 \times 3 \times ...n$, and 0!=0.
- The **binomial formula** can be used to find the probability that something happens *exactly x times in n trials*. Suppose the probability of a single trial being a success is p. Then the probability of observing exactly x successes in n independent trials is given by

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

- To apply the binomial formula, the events must be **independent** from trial to trial. Additionally, n, the number of trials must be fixed in advance, and p, the probability of the event occurring in a given trial, must be the same for each trial.
- To use the binomial formula, first confirm that the binomial conditions are met. Next, identify the number of trials n, the number of times the event is to be a "success" x, and the probability that a single trial is a success p. Finally, plug these three numbers into the formula to get the probability of exactly x successes in n trials.
- The $p^x(1-p)^{n-x}$ part of the binomial formula is the probability of just one combination. Since there are $\binom{n}{x}$ combinations, we add $p^x(1-p)^{n-x}$ up $\binom{n}{x}$ times. We can think of the binomial formula as: [# of combinations] × P(a single combination).
- To find a probability involving *at least* or *at most*, first determine if the scenario is binomial. If so, apply the binomial formula as many times as needed and add up the results. e.g. P(at least 3 Heads in 5 tosses of a fair coin) $=$ P(exactly 3 Heads) + P(exactly 4 Heads) + P(exactly 5 Heads), where each probability can be found using the binomial formula.

Section summary

- When a probability is difficult to determine via a formula, one can set up a **simulation** to estimate the probability.
- The **relative frequency** theory of probability and the **Law of Large Numbers** are the mathematical underpinning of simulations. A larger number of trials should tend to produce better estimates.
- The first step to setting up a simulation is to assign digits to represent outcomes. This should be done in such a way as to give the event of interest the correct probability. Then, using a random number table, calculator, or computer, generate random digits (outcomes). Repeat this a specified number of trials or until a given stopping rule. When this is finished, count up how many times the event happened and divide that by the number of trials to get the estimate of the probability.

Section summary

- A **discrete probability distribution** can be summarized in a table that consists of all possible outcomes of a random variable and the probabilities of those outcomes. The outcomes must be disjoint, and the sum of the probabilities must equal 1.
- A probability distribution can be represented with a histogram and, like the distributions of data that we saw in Chapter 2, can be summarized by its **center**, **spread**, and **shape**.
- When given a probability distribution table, we can calculate the **mean** (expected value) and **standard deviation** of a random variable using the following formulas.

  $$\begin{aligned} E(X) = \mu_X &= \sum x_i \cdot P(x_i) \\ &= x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \cdots + x_n \cdot P(x_n) \\ Var(X) = \sigma_x^2 &= \sum (x_i - \mu_X)^2 \cdot P(x_i) \\ SD(X) = \sigma_X &= \sqrt{\sum (x_i - \mu_X)^2 \cdot P(x_i)} \\ &= \sqrt{(x_1 - \mu_X)^2 \cdot P(x_1) + (x_2 - \mu_X)^2 \cdot P(x_2) + \cdots + (x_n - \mu_X)^2 \cdot P(x_n)} \end{aligned}$$

  We can think of $P(x_i)$ as the *weight*, and each term is weighted its appropriate amount.
- The **mean** of a probability distribution does not need to be a value in the distribution. It represents the average of many, many repetitions of a random process. The **standard deviation** represents the typical variation of the outcomes from the mean, when the random process is repeated over and over.
- **Linear transformations**. Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- **Combining random variables**. Let X and Y be random variables and let a and b be constants.
  - The expected value of the sum is the sum of the expected values.

    $E(X + Y) = E(X) + E(Y)$

    $E(aX + bY) = a \times E(X) + b \times E(Y)$
  - When X and Y are **independent**: The standard deviation of a sum or a difference is the square root of the sum of each standard deviation squared.

    $SD(X + Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

    $SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$

    $SD(aX + bY) = \sqrt{(a \times SD(X))^2 + (b \times SD(Y))^2}$

  The SD properties require that X and Y be independent. The expected value properties hold true whether or not X and Y are independent.

Section summary

- A **Z-score** represents the number of standard deviations a value in a data set is above or below the mean. To calculate a Z-score use: $Z = \frac{x - \text{mean}}{SD}$.
- *Z-scores do not depend on units.* When looking at distributions with different units or different standard deviations, Z-scores are useful for comparing how far values are away from the mean (relative to the distribution of the data).
- The **normal distribution** is the most commonly used distribution in Statistics. Many distribution are approximately normal, but none are exactly normal.
- The empirical rule (68-95-99.7 Rule) comes from the normal distribution. The closer a distribution is to normal, the better this rule will hold.
- It is often useful to use the standard normal distribution, which has mean 0 and SD 1, to approximate a discrete histogram. There are two common types of **normal approximation problems**, and for each a key step is to find a Z-score.

  A: *Find the percent or probability of a value greater/less than a given x-value.*
  1. Verify that the distribution of interest is approximately normal.
  2. Calculate the Z-score. Use the provided population mean and SD to standardize the given x-value.
  3. Use a calculator function (e.g. `normcdf` on a TI) or a normal table to find the area under the normal curve to the right/left of this Z-score; this is the *estimate* for the percent/probability.

  B: *Find the x-value that corresponds to a given percentile.*
  1. Verify that the distribution of interest is approximately normal.
  2. Find the Z-score that corresponds to the given percentile (using, for example, `invNorm` on a TI).
  3. Use the Z-score along with the given mean and SD to solve for the x-value.

- Because the sum or difference of two normally distributed variables is itself a normally distributed variable, the normal approximation is also used in the following type of problem.

  Find the probability that a sum $X + Y$ or a difference $X - Y$ is greater/less than some value.
  1. Verify that the distribution of X and the distribution of Y are approximately normal.
  2. Find the mean of the sum or difference. Recall: the mean of a sum is the sum of the means. The mean of a difference is the difference of the means. Find the SD of the sum or difference using: $SD(X + Y) = SD(X - Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$.
  3. Calculate the Z-score. Use the calculated mean and SD to standardize the given sum or difference.
  4. Find the appropriate area under the normal curve.

Section summary

- The symbol $\bar{x}$ denotes the sample average. $\bar{x}$ for any particular sample is a number. However, $\bar{x}$ can vary from sample to sample. The distribution of all possible values of $\bar{x}$ for repeated samples of a fixed size from a certain population is called the **sampling distribution** of $\bar{x}$.
- The standard deviation of $\bar{x}$ describes the typical error or distance of the sample mean from the population mean. It also tells us how much the sample mean is likely to vary from one random sample to another.
- The standard deviation of $\bar{x}$ will be *smaller* than the standard deviation of the population by a factor of $\sqrt{n}$. The larger the sample, the better the estimate tends to be.
- Consider taking a simple random sample from a population with a fixed mean and standard deviation. The **Central Limit Theorem** ensures that regardless of the shape of the original population, as the sample size increases, the distribution of the sample average $\bar{x}$ becomes more normal.
- Three important facts about the sampling distribution of the sample average $\bar{x}$:
  - The mean of a sample mean is denoted by $\mu_{\bar{x}}$, and it is equal to μ. (*center*)
  - The SD of a sample mean is denoted by $\sigma_{\bar{x}}$, and it is equal to $\frac{\sigma}{\sqrt{n}}$. (*spread*)
  - When the population is normal or when $n \geq 30$, the sample mean closely follows a normal distribution. (*shape*)
- These facts are used when solving the following two types of **normal approximation** problems involving a *sample mean* or a *sample sum.*

  A: *Find the probability that a sample average will be greater/less than a certain value.*
  1. Verify that the population is approximately normal or that $n \geq 30$.
  2. Calculate the Z-score. Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to standardize the sample average.
  3. Find the appropriate area under the normal curve.

  B: *Find the probability that a sample sum/total will be greater/less than a certain value.*
  1. Convert the sample sum into a sample average, using $\bar{x} = \frac{sum}{n}$.
  2. Do steps 1-3 from Part A above.

Section summary

In the previous chapter, we introduced the binomial formula to find the probability of exactly x successes in n trials for an event that has probability p of success. Instead of looking at this scenario piecewise, we can describe the entire *distribution* of the number of successes and their corresponding probabilities.

- The distribution of the *number of successes* in n independent trials gives rise to a **binomial distribution**. If X has a binomial distribution with parameters n and p, then $P(X = x) = \binom{n}{x} p^x (1-p)^n - x$, where $x = 0, 1, 2, 3 \ldots, n$.
- To write out a binomial probability **distribution table**, list all possible values for x, the number of successes, then use the binomial formula to find the probability of each of those values.
- Because a binomial distribution can be thought of as the *sum* of a bunch of 0s and 1s, the **Central Limit Theorem** applies. As n gets larger, the shape of the binomial distribution becomes more normal.
- We call the rule of thumb for when the binomial distribution can be well modeled with a normal distribution the **success-failure** condition. The success-failure condition is met when there are at least 10 successes and 10 failures, or when $np \geq 10$ and $n(1-p) \geq 10$.
- If X follows a binomial distribution with parameters n and p, then:
  - The mean is given by $\mu_X = np$. *(center)*
  - The standard deviation is given by $\sigma_X = \sqrt{np(1-p)}$. *(spread)*
  - When $np \geq 10$ and $n(1-p) \geq 10$, the binomial distribution is approximately normal. *(shape)*
- It is often easier to use **normal approximation to the binomial distribution** rather than evaluate the binomial formula many times. These three properties of the binomial distribution are used when solving the following type of problem.

  Find the probability of getting more than / fewer than x yeses in n trials or in a sample of size n.

  1. Identify n and p. Verify than $np \geq 10$ and $n(1-p) \geq 10$, which implies that normal approximation is reasonable.
  2. Calculate the Z-score. Use $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$ to standardize the x value.
  3. Find the appropriate area under the normal curve.

TI-84: CALCULATING SUMMARY STATISTICS

Use the `STAT`, `CALC`, `1-Var Stats` command to find summary statistics such as mean, standard deviation, and quartiles.

1. Enter the data as described previously.
2. Press `STAT`.
3. Right arrow to `CALC`.
4. Choose `1:1-Var Stats`.
5. Enter `L1` (i.e. `2ND 1`) for List. If the data is in a list other than `L1`, type the name of that list.
6. Leave `FreqList` blank.
7. Choose `Calculate` and hit `ENTER`.

TI-83: Do steps 1-4, then type `L1` (i.e. `2nd 1`) or the list's name and hit `ENTER`.

Calculating the summary statistics will return the following information. It will be necessary to hit the down arrow to see all of the summary statistics.

$\bar{x}$	Mean	n	Sample size or # of data points
Σx	Sum of all the data values	minX	Minimum
Σx^2	Sum of all the squared data values	Q_1	First quartile
Sx	Sample standard deviation	Med	Median
σx	Population standard deviation	maxX	Maximum

TI-83/84: DRAWING A BOX PLOT

1. Enter the data to be graphed as described previously.
2. Hit `2ND Y=` (i.e. `STAT PLOT`).
3. Hit `ENTER` (to choose the first plot).
4. Hit `ENTER` to choose `ON`.
5. Down arrow and then right arrow three times to select box plot with outliers.
6. Down arrow again and make `Xlist: L1` and `Freq: 1`.
7. Choose `ZOOM` and then `9:ZoomStat` to get a good viewing window.

3.3.3 Calculator: binomial probabilities

TI-83/84: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

Use MATH, PRB, nCr to evaluate n choose r. Here r and x are different letters for the same quantity.

1. Type the value of n.
2. Select MATH.
3. Right arrow to PRB.
4. Choose 3:nCr.
5. Type the value of x.
6. Hit ENTER.

Example: 5 nCr 3 means 5 choose 3.

CASIO FX-9750GII: COMPUTING THE BINOMIAL COEFFICIENT $\binom{n}{x}$

1. Navigate to the RUN-MAT section (hit MENU, then hit 1).
2. Enter a value for n.
3. Go to CATALOG (hit buttons SHIFT and then 7).
4. Type C (hit the ln button), then navigate down to the bolded C and hit EXE.
5. Enter the value of x. Example of what it should look like: 7C3.
6. Hit EXE.

TI-84: COMPUTING THE BINOMIAL FORMULA, $P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$

Use 2ND VARS, binompdf to evaluate the probability of *exactly* x occurrences out of n independent trials of an event with probability p.

1. Select 2ND VARS (i.e. DISTR)
2. Choose A:binompdf (use the down arrow to scroll down).
3. Let trials be n.
4. Let p be p
5. Let x value be x.
6. Select Paste and hit ENTER.

TI-83: Do step 1, choose 0:binompdf, then enter n, p, and x separated by commas: binompdf(n, p, x). Then hit ENTER.

TI-84: COMPUTING $P(X \le x) = \binom{n}{0}p^0(1-p)^{n-0} + ... + \binom{n}{x}p^x(1-p)^{n-x}$

Use 2ND VARS, binomcdf to evaluate the cumulative probability of *at most* x occurrences out of n independent trials of an event with probability p.

1. Select 2ND VARS (i.e. DISTR)
2. Choose B:binomcdf (use the down arrow).
3. Let trials be n.
4. Let p be p
5. Let x value be x.
6. Select Paste and hit ENTER.

TI-83: Do steps 1-2, then enter the values for n, p, and x separated by commas as follows: binomcdf(n, p, x). Then hit ENTER.

CASIO FX-9750GII: BINOMIAL CALCULATIONS

1. Navigate to STAT (MENU, then hit 2).
2. Select DIST (F5), and then BINM (F5).
3. Choose whether to calculate the binomial distribution for a specific number of successes, $P(X = k)$, or for a range $P(X \le k)$ of values (0 successes, 1 success, ..., x successes).
   - For a specific number of successes, choose Bpd (F1).
   - To consider the range 0, 1, ..., x successes, choose Bcd(F1).
4. If needed, set Data to Variable (Var option, which is F2).
5. Enter the value for x (x), Numtrial (n), and p (probability of a success).
6. Hit EXE.

GUIDED PRACTICE 3.71

Find the number of ways of arranging 3 blue marbles and 2 red marbles.[55]

GUIDED PRACTICE 3.72

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get exactly 3 blue marbles.[56]

GUIDED PRACTICE 3.73

There are 13 marbles in a bag. 4 are blue and 9 are red. Randomly draw 5 marbles *with replacement*. Find the probability you get *at most* 3 blue marbles (i.e. less than or equal to 3 blue marbles).[57]

[55] Here $n = 5$ and $x = 3$. Doing 5 nCr 3 gives the number of combinations as 10.
[56] Here, $n = 5$, $p = 4/13$, and $x = 3$, so set trials $= 5$, p $= 4/13$ and x value $= 3$. The probability is 0.1396.
[57] Similarly, set trials $= 5$, p $= 4/13$ and x value $= 3$. The cumulative probability is 0.9662.

4.1.5 Calculator: finding normal probabilities

TI-84: FINDING AREA UNDER THE NORMAL CURVE

Use `2ND VARS`, `normalcdf` to find an area/proportion/probability between two Z-scores or to the left or right of a Z-score.

1. Choose `2ND VARS` (i.e. `DISTR`).
2. Choose `2:normalcdf`.
3. Enter the `lower` (left) Z-score and the `upper` (right) Z-score.
   - If finding just a lower tail area, set `lower` to `-5`.
   - If finding just an upper tail area, set `upper` to `5`.
4. Leave μ as `0` and σ as `1`.
5. Down arrow, choose `Paste`, and hit `ENTER`.

TI-83: Do steps 1-2, then enter the lower bound and upper bound separated by a comma, e.g. `normalcdf(2, 5)`, and hit `ENTER`.

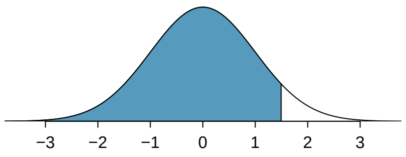
CASIO FX-9750GII: FINDING AREA UNDER THE NORMAL CURVE

1. Navigate to `STAT` (`MENU`, then hit `2`).
2. Select `DIST` (`F5`), then `NORM` (`F1`), and then `Ncd` (`F2`).
3. If needed, set `Data` to `Variable` (`Var` option, which is `F2`).
4. Enter the `Lower` Z-score and the `Upper` Z-score. Set σ to `1` and μ to `0`.
   - If finding just a lower tail area, set `Lower` to `-5`.
   - For an upper tail area, set `Upper` to `5`.
5. Hit `EXE`, which will return the area probability (`p`) along with the Z-scores for the lower and upper bounds.

EXAMPLE 4.11

Use a calculator to determine what percentile corresponds to a Z-score of 1.5.

Always first sketch a graph:[7]



To find an area under the normal curve using a calculator, first identify a lower bound and an upper bound. Theoretically, we want all of the area to the left of 1.5, so the left endpoint should be -∞. However, the area under the curve is nearly negligible when Z is smaller than -4, so we will use -5 as the lower bound when not given a lower bound (any other negative number smaller than -5 will also work). Using a lower bound of -5 and an upper bound of 1.5, we get $P(Z < 1.5) = 0.933$.

[7] normalcdf gives the result without drawing the graph. To draw the graph, do 2nd VARS, DRAW, 1:ShadeNorm. However, beware of errors caused by other plots that might interfere with this plot.

GUIDED PRACTICE 4.12

Find the area under the normal curve to right of $Z = 2$. [8]

GUIDED PRACTICE 4.13

Find the area under the normal curve between -1.5 and 1.5. [9]

TI-84: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

Use `2ND VARS`, `invNorm` to find the Z-score that corresponds to a given percentile.

1. Choose `2ND VARS` (i.e. `DISTR`).
2. Choose `3:invNorm`.
3. Let `Area` be the percentile as a decimal (the area to the left of desired Z-score).
4. Leave μ as `0` and σ as `1`.
5. Down arrow, choose `Paste`, and hit `ENTER`.

TI-83: Do steps 1-2, then enter the percentile as a decimal, e.g. `invNorm(.40)`, then hit `ENTER`.

CASIO FX-9750GII: FIND A Z-SCORE THAT CORRESPONDS TO A PERCENTILE

1. Navigate to `STAT` (`MENU`, then hit `2`).
2. Select `DIST` (`F5`), then `NORM` (`F1`), and then `InvN` (`F3`).
3. If needed, set `Data` to `Variable` (`Var` option, which is `F2`).
4. Decide which tail area to use (`Tail`), the tail area (`Area`), and then enter the σ and μ values.
5. Hit `EXE`.

EXAMPLE 4.14

Use a calculator to find the Z-score that corresponds to the 40th percentile.

Letting Area be 0.40, a calculator gives -0.253. This means that $Z = -0.253$ corresponds to the 40th percentile, that is, $P(Z < -0.253) = 0.40$.

GUIDED PRACTICE 4.15

Find the Z-score such that 20 percent of the area is to the right of that Z-score.[10]

[8] Now we want to shade to the right. Therefore our lower bound will be 2 and the upper bound will be +5 (or a number bigger than 5) to get $P(Z > 2) = 0.023$.

[9] Here we are given both the lower and the upper bound. Lower bound is -1.5 and upper bound is 1.5. The area under the normal curve between -1.5 and 1.5 = $P(-1.5 < Z < 1.5) = 0.866$.

[10] If 20% of the area is the right, then 80% of the area is to the left. Letting area be 0.80, we get $Z = 0.841$.

Chapter highlights

Chapter 1 focused on various ways that researchers collect data. The key concepts are the difference between a sample and an experiment and the role that randomization plays in each.

- Researchers take a **random sample** in order to draw an **inference** to the larger population from which they sampled. When examining observational data, even if the individuals were randomly sampled, a correlation does not imply a causal link.
- In an **experiment**, researchers impose a treatment and use **random assignment** in order to draw **causal conclusions** about the effects of the treatment. While often implied, inferences to a larger population may not be valid if the subjects were not also *randomly sampled* from that population.

Related to this are some important distinctions regarding terminology. The terms stratifying and blocking cannot be used interchangeably. Likewise, taking a simple random sample is different than randomly assigning individuals to treatment groups.

- **Stratifying vs Blocking**. Stratifying is used when sampling, where the purpose is to *sample* a subgroup from each stratum in order to arrive at a better *estimate* for the parameter of interest. Blocking is used in an experiment to *separate* subjects into blocks and then *compare* responses within those blocks. All subjects in a block are used in the experiment, not just a sample of them.
- **Random sampling vs Random assignment**. Random sampling refers to sampling a subset of a population for the purpose of inference to that population. Random assignment is used in an experiment to separate subjects into groups for the purpose of comparison between those groups.

When randomization is not employed, as in an **observational study**, neither inferences nor causal conclusions can be drawn. Always be mindful of possible **confounding factors** when interpreting the results of observation studies.

Chapter highlights

A raw data matrix/table may have thousands of rows. The data need to be summarized in order to makes sense of all the information. In this chapter, we looked at ways to summarize data **graphically**, **numerically**, and **verbally**.

Categorical data

- A single **categorical variable** is summarized with **counts** or **proportions** in a **one-way table**. A **bar chart** is used to show the frequency or relative frequency of the categories that the variable takes on.
- Two categorical variables can be summarized in a **two-way table** and with a **side-by-side bar chart** or a **segmented bar chart**.

Numerical data

- When looking at a single **numerical variable**, we try to understand the **distribution** of the variable. The distribution of a variable can be represented with a frequency table and with a graph, such as a **stem-and-leaf plot** or **dot plot** for small data sets, or a **histogram** for larger data sets. If only a summary is desired, a **box plot** may be used.
- The **distribution** of a variable can be described and summarized with **center** (mean or median), **spread** (SD or IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- **Z-scores** and **percentiles** are useful for identifying a data point's relative position within a data set.
- **Outliers** are values that appear extreme relative to the rest of the data. Investigating outliers can provide insight into properties of the data or may reveal data collection/entry errors.
- When **comparing the distribution** of two variables, use two dot plots, two histograms, a back-to-back stem-and-leaf, or parallel box plots.
- To look at the **association** between two numerical variables, use a **scatterplot**.

Graphs and numbers can summarize data, but they alone are insufficient. It is the role of the researcher or data scientist to ask questions, to use these tools to identify patterns and departure from patterns, and to make sense of this in the context of the data. Strong writing skills are critical for being able to communicate the results to a wider audience.

Chapter highlights

This chapter focused on understanding likelihood and chance variation, first by solving individual probability questions and then by investigating probability distributions.

The main probability techniques covered in this chapter are as follows:

- The **General Multiplication Rule** for **and** probabilities (intersection), along with the special case when events are **independent**.
- The **General Addition Rule** for **or** probabilities (union), along with the special case when events are **mutually exclusive**.
- The **Conditional Probability Rule**.
- Tree diagrams and **Bayes' Theorem** to solve more complex conditional problems.
- The **Binomial Formula** for finding the probability of exactly x successes in n independent trials.
- **Simulations** and the use of random digits to estimate probabilities.

Fundamental to all of these problems is understanding when events are independent and when they are mutually exclusive. Two events are **independent** when the outcome of one does not affect the outcome of the other, i.e. $P(A|B) = P(A)$. Two events are **mutually exclusive** when they cannot both happen together, i.e. $P(A \text{ and } B) = 0$.

Moving from solving individual probability questions to studying probability distributions helps us better understand chance processes and quantify expected chance variation.

- For a **discrete probability distribution**, the **sum** of the probabilities must equal 1. For a **continuous probability distribution**, the **area under the curve** represents a probability and the total area under the curve must equal 1.
- As with any distribution, one can calculate the mean and standard deviation of a probability distribution. In the context of a probability distribution, the **mean** and **standard deviation** describe the average and the typical deviation from the average, respectively, after many, many repetitions of the chance process.
- A probability distribution can be summarized by its **center** (mean, median), **spread** (SD, IQR), and **shape** (right skewed, left skewed, approximately symmetric).
- Adding a constant to every value in a probability distribution adds that value to the mean, but it does not affect the standard deviation. When multiplying every value by a constant, this multiplies the mean by the constant and it multiplies the standard deviation by the absolute value of the constant.
- The mean of the sum of two random variables equals the sum of the means. However, this is not true for standard deviations. Instead, when finding the standard deviation of a sum or difference of random variables, take the square root of the sum of each of the standard deviations squared.

The study of probability is useful for measuring uncertainty and assessing risk. In addition, probability serves as the foundation for inference, providing a framework for evaluating when an outcome falls outside of the range of what would be expected by chance alone.

Chapter highlights

This chapter began by introducing the normal distribution. A common thread that ran through this chapter is the use of the **normal approximation** in various contexts. The key steps are included for each of the normal approximation scenarios below.

1. Normal approximation for **data**:
   - Verify that population is approximately normal.
   - Use the given mean μ and SD σ to find the Z-score for the given x value.
2. Normal approximation for a **sample mean/sum**:
   Verify that population is approximately normal or that $n \geq 30$.
   Use $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ to find the Z-score for the given/calculated sample mean.
3. Normal approximation for the **number of successes** (binomial):
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$ to find the Z-score for the given number of successes.
4. Normal approximation for a **sample proportion**:
   - Verify that $np \geq 10$ and $n(1-p) \geq 10$.
   - Use $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ to find the Z-score for the given sample proportion.
5. Normal approximation for the **sum of two independent random variables**:
   - Verify that each random variable is approximately normal.
   - Use $E(X+Y) = E(X) + E(Y)$ and $SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2}$ to find the Z-score for the given sum.

Cases 1 and 2 apply to **numerical** variables, while cases 3 and 4 are for **categorical** yes/no variables. Case 5 applies to both numerical and categorical variables.

Note that in the binomial case and in the case of proportions, we never look to see if the *population* is normal. That would not make sense because the "population" is simply a bunch of no/yes, 0/1 values and could not possibly be normal.

The **Central Limit Theorem** is the mathematical rule that ensures that when the sample size is sufficiently large, the sample mean/sum and sample proportion/count will be approximately normal.

Project 1 - Due Friday, October 18

The focus of this project is to expose you to some of the features in Microsoft Excel that can be used to create some of the objects and to compute some of the values that we have seen in Chapters 1-3 of our text. In particular, in this project you will be asked to create various charts, like a frequency table, a bar chart, an ogive, and a pie chart. Additionally, you will be asked to compute the mean, median, mode, and standard deviation for a set of data.

Although Microsoft Excel is certainly not the only application available (and perhaps is far from the best), it is one of the more common tools used in the "real-world". So, an exposure to some of its computing and displaying capacity can be valuable.

Note that Microsoft Excel is available in computer labs campus-wide. If you do not have a computer with Excel installed, please plan to complete this project in one of the labs or use Microsoft 365 now available for free for CU Boulder students through OIT.

The project is due by class on Friday, October 18. There are two components: A worksheet which you will complete using your results from Excel, and an Excel workbook containing all required components. The worksheet is to be handed in class. The Excel workbook will need to be submitted via the Canvas Project 1 File Upload.

You must work in groups of 2 or 3, handing in one worksheet and submitting one Excel worksheet for the group. On Wednesday, October 2, you will be asked to specify your group members. You will need to upload your Excel worksheet to Canvas by class time on Friday, October 18. You will hand in the worksheet in class on that date, as well.

For those of you unfamiliar with Excel, useful screencasts will be provided in Canvas. There are also many resources online for help and tutorials on the tasks required of you in this project.

To begin. In the folder for Project 1, you should find an Excel file for your section. Download the file to your computer. The data sheet contains the data that you will analyze. The data highlighted in green should NOT be altered in anyway. Changing values or even sorting the data may result in inaccurate results later.

Additionally, the explicit locations for the placement of each of the items you are instructed to create/compute are labeled and highlighted in yellow. You are welcome to use other space in your data sheet if needed (for intermittent computations or analysis), but to earn credit for the required components, they must be in their specified location. Do NOT relocate any of these cells. Even adding a row or column to the data set can result in a loss of points.

Lastly, all objects and computations completed in the Excel file should be done *robustly*, meaning that when any of the original data values are changed, the objects and values that you designed will update accordingly. Part of your grade for this project will be based on the required components updating when some of the data is changed. Use cell references rather than hard-coding any numeric values into your formulas to achieve this robustness.

The data in the spreadsheet are responses from 650 students collected for MATH 2510.

1. Complete the frequency table for the PREFERRED COLOR responses from the sample of 650 respondents in the data sheet. {Hint: COUNTIF}
2. Create a Bar chart to display the distribution of preferred colors. Make sure to label the Bar chart properly.
3. Complete the frequency table for the CLASS STANDING responses from the sample of 650 respondents in the data sheet. {Hint: COUNTIF}
4. Create a Pie chart to display the distribution of class standings. Make sure to label the Pie chart properly.
5. Compute the mean height of the sample of 650 respondents in the data sheet. {Hint: AVERAGE}
6. Determine the five-number summary of the heights of the sample of 650 respondents in the data sheet. {Hint: QUARTILE}
7. Compute the mean height of only the female respondents in the sample of 650 respondents in the data sheet. {Hint: AVERAGEIF}
8. Compute the mean height of only the male respondents in the sample of 650 respondents in the data sheet. {Hint: AVERAGEIF}
9. Complete the FREQUENCY column in the frequency table for the heights of only the female respondents in the sample of 650 respondents in the data sheet. {Hint: COUNTIFS}
10. Unfortunately, there is not (yet) a function called STDEVIF. So, in order to compute the standard deviations for the heights of only the female respondents in the sample of 650 respondents in the data sheet, please use the defining formula and the following steps.
    (a) Complete the column in the frequency table containing the values of $x - \bar{x}$. You have already computed $\bar{x}$, so simply reference the cell containing that value.
    (b) Now complete the column containing the values $(x - \bar{x})^2$.
    (c) Taking in consideration the frequencies of each height, determine the sample variance.
    (d) Lastly, use the sample variance to determine the sample standard deviation.
11. Compute the mean number of coin flips that landed heads as reported by these 650 students. {Hint: AVERAGE}
12. Compute the sample standard deviation for the number of coin flips that landed heads as reported by this sample of 650 students. {Hint: STDEV}
13. Determine the upper and lower bounds of the 84% Chebyshev interval for the number of heads, and then determine what percentage of the outcomes that actually lie within that range. (If it is not 84% or greater, something is wrong.) {Hint: COUNTIFS}
14. Complete the binned frequency table with 10 bins { '1 to 5', '6 to 10', '11 to 15', ..., '46 to 50'} for the number of heads flipped as reported by this sample. Then add to the table values for Relative Frequency, Cumulative Frequency, and Relative Cumulative Frequency. (Be mindful of your class boundaries and how to correctly communicate this to the Excel function.) {Hint: COUNTIFS}
15. In a **single graph**, create an Ogive for the number of heads flipped in each trial into a **single graph**.

Use the information found in the Excel workbook you just created to answer the questions on the following worksheet. Bring the worksheet to class on Friday, October 18 to turn in. Upload your Excel workbook to Canvas by class time on that same day.

Project 1 Worksheet - Due Friday, October 18

NAME 1:____

NAME 2: ____

NAME 3: ____

1. What was the most popular answer to the “Which of the following colors do you most prefer?” question? How many survey respondents answered with that color?

2. How many students responded that they are Sophomores? Is the sampling method used to collect the data likely a good method to infer the proportion of all students at CU that are Sophomores? Why or why not?

3. What is the mean height of all 650 students? What is the average of the female and male mean heights? You should find that your two values are not exactly equal. Explain why this is the case.

4. What is the standard deviation of the heights of the female respondents in the sample? Explain whether the standard deviation of the heights of all 650 respondents is likely to be greater than or less than the standard deviation for the female respondents alone and why. (You need not compute the standard deviation of the entire 650 students to answer this, but if you do, then an explanation like “It is greater, because the computed value is larger” is not sufficient. Your explanation should relate that you understand what the standard deviation measures.)

5. Although you were not asked to create a histogram for the coin flip data, what does the frequency table indicate about the shape of the distribution? (unimodal, bimodal, symmetric, skewed) How does the ogive also illustrate the shape of the distribution?