# STAT 145 Notes

## Contents

## Fun stuff

## Course outline

### .1 A: Data (Ch1-2)

1. How to collect and describe data.

2. How data collection influences the types of conclusions to be drawn.

3. Ways to sumarize and visualize data.

### .2 B: Understanding inference (Ch3-4)

### .3 C: Inference with normal and t-distributions (Ch5-6)

### .4 D: Inference for multiple parameters (Ch7-10)

## Chapter 1: Collecting Data

### .1 1.1 The structure of data

1. Basics: Data structure, variable types
   (a) *Statistics* is the science of collecting, describing, and analyzing data.
   (b) Here we focus on data collection, chapter 2 describes data, and later chapters analyze data.

(c) Dataset structure:
   - Rows are *cases* or *units*
   - Columns are *variables*
   - Example: Student survey, rows are students, columns are GPA, gender, email, major, age, etc

(d) Variable types:
   - *Categorical variables* divide cases into groups
   - *Quantitative variables* give a numerical quantity which can be applied operations to such as averaging.
   - Example: Classify variables above.

2. Investigating variables and relationships between variables

   (a) Single variable questions: What is the average student age? Majority gender? Number of 4.0 GPAs?

   (b) Variable relationship questions: Who has higher GPA, male or female? Does age determine GPA? (many pairings, quant to cat, quant to quant, cat to cat)

   (c) Explanatory and response variables

      i. Ask if one variable helps to explain another.

      ii. Explanatory variable helps us understand (or predict) another variable. Also called independent variable.

      iii. Response variable is what we aim to understand (or predict). Also called dependent variable.

      iv. Example: Does student age predict GPA?

## .2 1.2 Sampling from a population

1. Data collection: Super important before any analysis can be done and trusted.

   (a) A *population* is the entire group of indviduals or objects of interest.

   (b) A *sample* is a subset of the population which data is collected for.

   (c) *Statistical inference* is the process of using data from a sample to gain information about the entire population.

   (d) *Sampling bias* occurs when the method of selcting a sample causes the sample to differ from the population in some relevant way. Such a sample would not generalize to the entire population.

   (e) Key question: How to tell if the sample is good enough for inference? How much to trust the findings? Choose a random sample.

   (f) A *simple random sample* of $n$ units is such that all groups of size $n$ have the same chance of becoming the sample.

   (g) Still pitfalls exist. How to randomly select? Computer random number generator. Need all selected participants to be willing. May not be completely doable.

2. Types of bias:

   (a) *Voluntary response bias* occurs when volunteers are asked to participate.

   (b) *Convenience sampling* occurs when you use the group which is easiest to access for the sample.

   (c) *Response bias* occurs when people do not answer questions honestly.

   (d) *Non-response bias* occurs when people do not respond and therefore are excluded from the sample.

   (e) *Undercoverage* occurs when a part of the population is not considered for inclusion into the sample.

### .3 1.3 Experiments and observational studies

1. Association vs causation:
   (a) Two variables are *associated* if values of one variable tend to be related to values of another variable.
   (b) Two variables are *causally associated* if changing the value of one variable influences the value of the other variable.
   (c) Example: Sharks and ice cream sales. Heat wave.
   (d) A *confounding variable* is a third variable that is associated with both the explanatory and response variable. These can offer a plausible explanation for the associate between the two variables.
   (e) How to avoid confounding variables? Avoid observational studies and instead perform random experiments.

2. Randomized experiements:
   (a) In a *random experiment* the value of the explanatory variable for each unit is determined randomly before the response variable is measured.
   (b) Allows to establish a causal relationship.
   (c) Two basic types of random experiments:
      - A *randomized comparative experiment* randomly assigns cases to to different treatment groups and then compares results on the response variable. This includes examples such as control groups and placebos.
         - Placebos require blinding to hide the lie.
         - Single-blind experiments don't tell participants which group they are in.
         - Double-blind experiments don't tell participants which group and the data recorders also don't know which group.
      - A *matched pairs experiment* get both treatments in random order and then compare individual differences.
   (d) In all, randomized experiments are the best to determine causality, though they are not always possible or ethical. Observational studies may then be used to determine association.

3. Two fundamental questions about data collection:
   (a) Was the sample randomly selected? If yes, can generalize to the entire population.
   (b) Was the explanatory variable randomly assigned? If yes, conclusions can be made about causality.

## Chapter 2: Describing Data

Goal of chapter:

1. Describe both types of variable (categorical and quantitative) and their relationships between them.

2. Visualize data via graphs

3. Summarize key aspects of data via *summary statistics*.

## .1 2.1 Categorical variables

1. One categorical variable
   (a) A *frequency table* gives counts of each category.
   (b) *Proportions* (AKA relative frequencies) are computed as

   $$\text{Category proportion} = \frac{\text{Category number}}{\text{Total number}}$$

   (c) Visualize as a bar char or a pie chart.
   (d) $\hat{p}$ denotes proportion for a sample. $p$ denotes proportion for a population. This distinction is key with writing.
2. Two categorical variables
   (a) Two way table comparison. Counts vs proportions.
   (b) Difference in proportions.
   (c) Visualize as segmented (stacked) bar chart or a side-by-side bar chart.

## .2 2.2-2.3 One quantitative variable: Shape and center, measures of spread

1. For a quantitative variable, three main questions are key.
   (a) What is the *shape*?
   (b) What is the *center*?
   (c) How does the data vary (known as *spread*)?
   (d) Combined these compose the *distribution* of the data.
2. The *shape* of a distribution:
   (a) Dotplots or histograms can be used to visualize smaller sized data. Histograms require binning of data.
   (b) *Outliers* are observed data which is much larger or much smaller than the rest of the data values.
   (c) Symmetric or skewed distributions:
   - Symmetric distributions are bell shaped. Bimodal (non-bell shaped) symmetric is possible.
   - Skewed distributions have a long tale on right or left side. A tail on the left is called *right-skewed*.
   - Many other shapes are possible.
3. The *center* of a distribution:
   (a) Two main summary statistics describe the center of a distribution, the mean and the median.
   - The *mean* is the average (sum of all values divided by the count). Sample mean is computed as

     $$\overline{x} = \frac{1}{n}\sum x_i$$

     and population mean is denoted with Greek letter $\mu$.
   - The *median* (denoted $m$) is the middle entry if all values are listed in order. For even counts, the two middles are averaged.
   - The mean is the balance point of a distribution. The median splits the data into two equal halfs.
   - The median is more *resistant* to outliers while the mean is now.

(b) Can use the mean and median of a distribution to tell if it is symmetric or skewed.

- $m = \overline{x}$ is symmetric
- $m < \overline{x}$ is right skew
- $m > \overline{x}$ is left skew

4. The *spread* of a distribution:

(a) We need a way to measure how spread out data is. Is it clustered close to the mean or farther out?

(b) The *deviation* of a single data value $x$ from the mean $\overline{x}$ is the difference $x - \overline{x}$. Positive means bigger than the mean, negative means smaller.

(c) Adding up all deviations will always give zero!

(d) Instead we take the square root of the averaged sum of squares. The sample standard deviation is then

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

The population standard deviation is $\sigma$.

(e) In essence, the standard deviation is a rough estimate of the typical distance from the mean. Large $s$ means more spread.

(f) The standard deviation 95% rule:

- For symmetric, bell-shaped data, 95% of the data should fall within two standard deviations of the mean. That is, in the interval $[\overline{x} - 2s, \overline{x} + 2s]$.

(g) For a single sample $x$ we can measure how many standard deviations from the mean as

$$x = \bar{x} + s \cdot z \quad \rightarrow \quad z\text{-score} = \frac{x - \bar{x}}{s}$$

which defines the $z-$score. Then for symmetric bell-shaped data, only 5% of the data will have a $z-$score bigger than 2 in absolute value.

(h) Explanations of $s$ formula.

- `https://en.wikipedia.org/wiki/Standard_deviation#Corrected_sample_standard_deviatio`
- Bessel's correction: `https://en.wikipedia.org/wiki/Bessel%27s_correction`
- `https://www.khanacademy.org/computer-programming/spin-off-of-fishy-statistics-unbia`
  `6742828205522944`

5. Percentiles and IQR:

(a) Again, the mean and therefore standard deviation are sensitive to outliers (though they do have the advantage that all data values are used in calculation). We can replace mean with median and measure spread thru this lens.

(b) The *Pth percentile* is the value of the quantitative variable which is greater than $P$ percent of the data.

(c) Five number summary:

- Minimum, $Q_1$, median, $Q_3$, maximum
- $Q_1$ is the 25th percentile (first quartile).
- $Q_3$ is the 75th percentile (third quartile).
- This divides the data into fourths with an equal count of data in each of the 4 buckets.
- Better than mean and SD to tell distribution shape and skewness.

(d) Range = max - min

(e) Interquartile range (IQR) = $Q_3 - Q_1$.

### .3 2.4 Boxplots and quantitative/categorical relationships

1. Boxplots

   (a) A *boxplot* is a graphical display of the five number summary for a single quantitative variable. The main goal is to show the general shape of a distribution. Components:

   - Data scale
   - Box stretching from $Q_1$ to $Q_3$
   - Line drawn at the median
   - Line from each quartile drawn to the most extreme value not an outlier
   - Individual outliers drawn with an asterisk
   - Symmetry and skew can be seen

   (b) IQR method for detecting *outliers* is:

   - Smaller than $Q_1 - 1.5(IQR)$
   - Bigger than $Q_3 + 1.5(IQR)$

2. One quantitative and one categorical variable:

   (a) Side-by-side boxplots or dotplots

   (b) Comparative summary statistics

### .4 2.5 Two quantitative variables: Scatterplot and correlation

1. SKIP

### .5 2.6 Two quantitative variables: Linear regression

1. SKIP

### .6 2.7 Data visualization and multiple variables

1. SKIP

## Chapter 3: Confindence intervals

1. Develop the key ideas of statistical inference (estimation and testing) using simulation methods to build understanding and carry out analysis.

2. Key goal is to use data in the sample to understand what might be true about the entire population. With these conclusions will come accuracy.

### .1 3.1 Sampling distributions

1. *Statistical inference* is the process of drawing conclusions about the entire population based on the information in the sample.

2. Population parameters and sample statistics:

   (a) A *parameter* is a number that describes some aspect of a population.

   (b) A *statistic* is a number computed from the data in a sample.

   (c) Names for parameters and statistics are the same, but notation differs. For example, mean for a statistic is $\bar{x}$ and for a population is $\mu$. See table in text for full list.

3. Sample statistics as estimates of population parameters
    (a) If we only have one sample and don't know the value of the population parameter, the sample statistic is our *best estimate* of the true parameter value.
    (b) Parameters are fixed though usually unknown. Sample statistics vary from sample to sample, but at least we compute their values.
    (c) A *sampling distribution* is the distribution of sample statistics computed from different samples of the same size from the same population.
    (d) For most parameters, if samples are randomly selected and sample size is large enough, the sampling distribution will be symmetric and bell-shaped with center at the value of the population parameter. This is known as the Central Limit Theorem.
4. Measuring the sampling variability: The standard error
    (a) The spread of the sample statistic is the most important in knowning how accurate the estimate is. Standard deviation captures this.
    (b) The *standard error* of a statistic, denoted $SE$, is the standard deviation of the sample statistic.
    (c) In the next section we use $SE$ to quantify uncertainty via confidence intervals.
    (d) Sample size is important. As sample size increases, $SE$ decreases making the statistic a better estimate of the population parameter.
    (e) Random sampling is important. Bias in sampling can lead to false conclusions about the population parameter.

## .2 3.2 Understanding and interpreting confidence intervals

1. First we aim to identify the parameter we are interested in. To estimate, repeat random sampling and compute the sampling statistic. In the process, measure the variation on sample statistics. Result is an estimate of the parameter with a range of plausible values.
2. Identifying the parameter of interest: 5 options so far
    (a) Single categorical variable, proportion $p$ (parameter) has statistic $\hat{p}$.
    (b) Single quantitative variable, mean $\mu$ (parameter) has statistic $\bar{x}$.
    (c) Two categorical variables, difference in proportions $p_1 - p_2$ (parameter) has statistic $\hat{p_1} - \hat{p_2}$.
    (d) One categorical and one quantitative variable, difference in means $\mu_1 - \mu_2$ (parameter) has statistic $\bar{x}_1 - \bar{x}_2$.
    (e) (Will see later on) Two quantitative variables, correlation $\rho$ (parameter) has statistic $r$.
3. Parameter estimation with an interval (margin of error)
    (a) Aim is a range of possible values for the population parameter as

    $$\text{sample statistic} \pm \text{margin of error}$$

    which is the interval $[SS - MoE, SS + MoE]$
4. Confidence intervals
    (a) A *confidence interval* for a parameter is an interval computed from the sample data by a method that will capture the parameter for a proportion of all samples. The success rate (proportion of all samples whose intervals contain parameter) is the *confidence level*.

(b) If we estimate the standard error $SE$ the sampling distribution is symmetric and bell shaped, a 95% confidence interval can be estimated using

$$\text{Statistic} \pm 2 \cdot SE.$$

The tails of the bell curve more than $2SE$ away from the mean comprise 5% of the sample statistics.

(c) We interpret a CI as being 95% certain that the population parameter lies in the interval.

(d) Note, margin of error (amount added and subtracted in confidence interval), standard error (standard dev of many sample statistics), and standard deviation of a sample (single sample of data values) are different! Stats terminology sux.

5. In practice we only have one sample for our data. The next section shows how bootstrapping can be used to generate artificial samples from a single one.

## .3 3.3 Constructing bootstrap confidence intervals

1. Bootstrapping samples:
   (a) We need many different samples to construct a confidence interval. In practice, we only have one sample to work with. We can use that one sample to create a sampling distribution.
   (b) *Sampling with replacement* randomly samples the sample. Once an item is selected, it is still available to be selected again.
   (c) A *bootstrap sample* is the collection of samples with replacement from the original sample.
   (d) A *bootstrap statistic* is computed from the bootstrap sample.
   (e) Collecting all bootstrap statistics give the *bootstrap distribution.* Assuming the original sample was chosen randomly and is big enough, the bootstrap distribution is a good approximation to the sampling distribution.
2. Estimating standard error from a bootstrap distribution
   (a) The standard deviation of the bootstrap statistics in a bootstrap distribution gives a good approximation of the standard error of the statistic.
   (b) Likewise the confidence interval from the bootstrap distribution approximates the sample statistic confidence interval.

## .4 3.4 Bootstrap confidence intervals using percentiles

1. Confidence intervals based on bootstrap percentiles
   (a) 95% confidence intervals are found from a symmetric, bell-shaped bootstrap distribution via the rough $\mathit{Statistic} \pm 2 \cdot SE$. Instead of using $SE$, we could just remove the upper and lower 2.5% of the bootstrap distribution. This would yield similar results. Generalizing this idea to any percentile gives a more flexible approach.
   (b) A 90% confidence interval would remove 5% from both tails of the bootstrap distribution.
   (c) In general, for a symmetric bootstrap distribution, a $XX\%$ confidence interval
2. Caution on constructing bootstrap confidence intervals
   (a) Bootstrap intervals don't always work. When data looks more discrete (rather than continuous) or non-bell-shaped, results are note reasonable.
   (b) Plotting the bootstrap distribution is always a best practice before discussion of confidence intervals.

# Chapter 4: Hypothesis testing

## .1 4.1 Introducing hypothesis tests

1. Two main areas of statistical inference: estimation and testing
   (a) Chapter 3 gives estimate with a confidence interval.
   (b) Here we use statistical testing to answer questions about our data.
   (c) A *statistical test* is used to determine whether the results from a sample are convincing enough to allow us to conclude something about the population.
2. Null and alternative hypothesis
   (a) The *null hypothesis* $H_0$ claims that there is no effect or no difference.
   (b) The *alternative hypothesis* $H_a$ claims what we seek as significant evidence.
   (c) The *hypothesis test* examines whether sample data provides enough evidence to refute the null hypothesis and support that alternative hypothesis.
   (d) Note, $H_0$ and $H_a$ are claims about population parameters, not sample statistics.
   (e) Note, $H_0$ is a statement about equality while $H_a$ claims less than, more than, or not equal depending on the question asks.

## .2 4.2 Measuring evidence with p-values

1. How to measure evidence?
   (a) A hypothesis test examines whether data from a sample provides enough evidence to refute the null hypothesis and support the alternative hypothesis.
   (b) How to measure evidence? We need to have a sense of what is likely to occur by random chance.
2. Randomization distribution
   (a) We want to understand how statistics randomly vary from sample to sample, if the null hypothesis is true.
   (b) We will bootstrap to simulate samples in a way that his consistent with the null hypothesis.
   (c) This is called *random samples*.
   (d) For each random sample, we calculate the statistic of interest.
   (e) The values of the statistic of interest for all random samples generates a *randomization distribution*.
   (f) How to generate a randomization distribution:
      - Put each case randomly into 2 groups.
      - Compute sample statistic.
      - Repeat.
   (g) We then compare our observed statistic to the randomization distribution.
3. Measuring strength of evidence with a $p$-value
   (a) The *p-value* is the proportion of samples, when the null hypothesis is true, that would gave a statistic as extreme as (or more extreme than) the observed sample.

$$p - value = \frac{\text{number of random samples with statistic bigger than (same sign) observed stat}}{\text{total random samples}}$$

   (b) The $p$-value only considers one side of the randomized distribution.

   (c) The farther to the tail of the randomization, the smaller the $p$-value.

4. $p$-values and the alternative hypothesis.

   (a) The randomize distribution only considers the null hypothesis.

   (b) The $p$-value connects to the alternative hypothesis and considers one or both tails of the distribution. We call these *right-tailed*, *left-tailed*, and *two-tailed* tests.

   (c) Note, for a *two-tailed* test, we find the proportion of the simulated statistic in the smaller tail at or beyond the observed statistic. The $p$-value doubles this proportion to account for the other tail.

## .3 4.3 Determining statistical significance

1. Statistical significance

   (a) The smaller the $p$-value, the stronger the statistical evidence is against the null hypothesis in support of the alternative hypothesis.

   (b) If the $p$-value is small enough, we say the sample results are *statistically significant* and we have convincing evidence against $H_0$ and in favor of $H_a$.

   (c) The *significance level* $\alpha$ for a test of hypothesis is a boundary below which concludes the $p$-value is statistically significant. Common levels are $\alpha = 0.05, 0.01, 0.1$. If not specified, $\alpha = 0.05$ is used.

   (d) If the $p$-value is less than $\alpha$, we reject $H_0$. If the $p$-value $\geq \alpha$, we cannot reject $H_0$ and do not have convincing evidence that $H_a$ is true.

2. Summary of hypothesis tests

   (a) State the null and alternative hypothesis.

   (b) Determine the value of the observed sample statistic.

   (c) Find the $p$-value

   (d) Reject or do not reject $H_0$.

   (e) Write a sentence explaining the conclusion.

3. Less formal rules for statistical decisions.

   (a) Instead of reject / do not reject, can replace with strength levels (little, some, moderate, strong, very strong).

## .4 4.4 A closer look at testing

1. Pitfalls to hypothesis testing: Type 1 and type 2 errors

   (a) Two generic decisions: reject $H_0$, do not reject $H_0$.

   (b) Type 1 error: $H_0$ is in reality true, decide to reject $H_0$ (false positive)

   (c) Type 2 error: $H_0$ is false in reality, decide not to reject $H_0$ (false negative)

   (d) Good to avoid both error types, but sometimes one is more important to avoid. Choosing significant level $\alpha$ is a way to control this balance.

   (e) $\alpha = 0.01$ is decreased making it harder to reject $H_0$ which avoids Type 1 error.

   (f) $\alpha = 0.1$ is increased making it easier to reject $H_0$ which avoids Type 2 error.

2. The problem of multiple testing:

(a) When multiple tests are conducted, if the null hypothesis are all true, the proportion of the tests that will yield a statistically significant result by random chance if about $\alpha$, the significance level.

(b) With publications, only significant results are often published. This could be a result of multiple testing with non-sig results not published.

(c) It is important to replicate or reproduce results with another study as with clinical trials.

3. Effect of sample size

(a) As sample size increases, there will be less spread in the kinds of statistics we will see just by random chance. Statistics in the randomization distribution will be closely concentrated around the null value.

(b) A large sample size makes it easier to reject $H_0$ when $H_a$ is true, decreasing the chance of Type 2 errors.

(c) Larger samples are always better!

## .5 4.5 Making connections

1. Connecting randomization and bootstrap distributions.

(a) Sampling distribution: shows the distribution of sample statistics from a population, and is usually centered at the true value of the population parameter.

(b) Bootstrap distribution: simulates a distribution of sample statistics for the population, but is generally centered at the value of the original sample statistic.

(c) Randomization distribution: Simulates a distribution of sample statistics for a population for which the null hypothesis is true, and is generally centered at the value of the null parameter.

(d) Both simulate many samples then collect values of a sample statistic to form a distribution.

(e) Both have typical values in the middle and unusual values in the tails.

(f) Both use information from the original sample to make inference what might be true about a population, parameter, or relationship.

2. Connecting confidence intervals and hypothesis tests

(a) Confidence intervals: show us plausible values of the population parameter

(b) Hypothesis tests: used to determine whether a given parameter in a null hypothesis is plausible or not

(c) The formal decision of a two-tailed hypothesis test is related to whether or not the null parameter falls within a confidence interval:

- when the parameter value of $H_0$ falls outside a 95% confidence interval, then it is not a plausible value for the parameter and we should reject $H_0$ at 5% level in a two-tailed test
- when the parameter value of $H_0$ falls inside a 95% confidence interval, then it is a plausible value for the parameter and we should not reject $H_0$ at 5% level in a two-tailed test

(d) Combining ideas of confidence intervals and hypothesis enables 2 things:

- Conclude if we should reject $H_0$ or not
- Confidence interval gives scale of desired parameter

3. Creating randomization distributions

(a) Two main criteria to consider when creating randomization samples for a statistical test:

- be consistent with the null hypothesis
- use the data in the original sample (and possibly reflect the way the original data was collected)

(b) 5 examples of tests to create randomization distributions for:

- Difference in means $\mu_1 = \mu 2$
- Difference in proportions: $p_1 = p_2$
- Correlation (later)
- Single proportion $p$
- Single mean $\mu$

# Chapter 5: Approximating with a distribution

## .1 5.1 Hypothesis tests using normal distributions

1. Normal distribution

   (a) The *normal distribution* follows a bell-shaped curve. We use two parameters (mean $\mu$ and standard deviation $\rho$) to distinguish one normal curve from another and abbreviate with notation $N(\mu, \rho)$.

   (b) Mean $\mu$ locates the center of the distribution

   (c) Standard deviation $\rho$ tells how tall and skinny (or short and wide) the distribution is.

   (d) The standard normal distribution is $Z = N(0, 1)$ with mean 0 an standard deviation 1.

   (e) Convert any normal distribution $N(\mu, \rho)$ to $Z$ by

   $$Z = \frac{X - \mu}{\rho}$$

   where $X$ is any value on $N(\mu, \rho)$. We use $Z$ because this is the $Z-$score from before.

   (f) Formula for $N(\mu, \rho)$ at any location $x$ is

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

2. The central limit theorem

   (a) For random samples of sufficiently large sample size, the distribution of many common sample statistics can be approximated with a normal distribution.

   (b) A bit vague. What precisely is random? How large a sample? How good an approximation?

   (c) If true, we can trade randomization distributions for normal distribution to compute $p$-values and confidence intervals.

3. $p-$value from a normal distribution

   (a) The normal distribution that best approximates a bell-shaped randomization distribution has mean equal to the null value of the parameter, with standard deviation equal to the standard error:

   $$N(\text{null parameter}, SE).$$

   A $p-$ value can be found as the proportion of this normal distribution beyond the observed sample statistic in the direction of the alternative (or twice the smaller tail for a two-tailed test).

   (b) Can also standardize the test statistic and compare to the standard normal distribution. For the distribution of the statistic under $H_0$ which is normal, we compute a standardized *test statistic* using

   $$z = \frac{\text{Sample statistic} - \text{Null parameter}}{SE}.$$

   The $p-$ value for the test is the proportion of the standard normal distribution beyond the standardized test statistic, depending on the direction of the alternative hypothesis.

### .2 5.2 Confidence intervals using normal distributions

SKIP

## Chapter 6: Inference for means and proportions

### .1 6.1 Inference for a proportion

1. 6.1D: Distribution of a proportion
   (a) For categorical data, the parameter of interest is a population proportion $p$.
   (b) For large enough sample size, CLT says the sample proportion $\hat{p}$ will follow a normal distribution centered at $p$.
   (c) The normal distribution also has a standard deviation, the standard error (SE) of the sample proportions.
   (d) $SE$ can be estimated by bootstrapping or randomization distribution, but a version of the CLT gives another way
   (e) Theorem: CLT alternate version. For random samples of size $n$ from population with proportion $p$, the distribution of sample proportions is centered at population proportion $p$ with standard error
   $$SE = \sqrt{\frac{p(1-p)}{n}}$$
   and is reasonably normally distributed if $np \geq 10$ and $n(1-p) \geq 10$.
   (f) The conditions $np \geq 10$ and $n(1-p) \geq 10$ prevent the normal distribution center from being too close to 0 and 1 respectively. If so, the $0 \leq p \leq 1$ restriction will cutoff the normal curve making it not normalish.
2. 6.1CI: Confidence interval for a proportion
   (a) In section 5.2 we found for a normal distribution of a sample statistic, confidence intervals are found via
   $$\text{Sample Statistic} \pm z^* \cdot SE$$
   where $z^*$ is a distribution percentile and $SE$ is the standard error of the sample statistic.
   (b) In section 6.1-D, we have for a large sample, the distribution of the sample proportion is normalish with standard error
   $$SE = \sqrt{\frac{p(1-p)}{n}}$$
   for $n$ the sample size and $p$ the population proportion. Replace unknown $p$ with sample proportion $\hat{p}$ and we have
   $$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
   and
   $$\hat{p} \pm z^* SE = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
   (c) How big of a sample do we need to estimate a proportion? If we know what margin of error $ME$ is acceptible,
   $$ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \rightarrow \quad n = \left(\frac{z^*}{ME}\right)^2 \hat{p}(1-\hat{p}),$$
   though if we do not know $\hat{p}$ since we haven't collected the sample, often $\hat{p} = 0.5$ is used because that is where $ME$ is largest.

3. 61HT: Hypothesis testing for a proportion

   (a) We have already seen that when a randomization distribution is normal, we can compute a $p$-value using a standard normal distribution and standardized test statistic

   $$z = \frac{\text{Sample statistic} - \text{Null parameter}}{SE}$$

   The sample statistic is computed from the sample data and the null parameter is specified by the null hypothesis $H_0$.

   (b) For a population proportion,

   $$z = \frac{\hat{p} - p_0}{SE}$$

   for $\hat{p}$ the sample proportion and $p_0$ the proportion for $H_0$. Using our new formula for $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$, we get the following result.

   (c) Hypothesis test for a proportion: To test $H_0 : p = p_0$ vs $H_a : p \neq p_0$ (or a one-tail alternative, we use the standardized test statistic

   $$z = \frac{\text{Sample statistic} - \text{Null parameter}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

   where $\hat{p}$ is the proportion in a random sample of size $n$. Provided the sample is large enough so that $np_0 \geq 10$ and $n(1 - p_0) \geq 10$, the $p$-value of the test is computed using the standard normal distribution.

## .2 6.2 Inference for a mean

1. Distribution of for a mean

   (a) CLT for sample means:

   - For quantitative data, the parameter of interest is usually the population mean $\mu$.
   - The distribution of sample means $\bar{x}$ often follows a normal distribution with center $\mu$.
   - The standard deviation of this normal distribution is important to know, which we view through the standard error ($SE$) of the sample means. The normal distribution curve gives us the formula:

     $$SE = \frac{\sigma}{\sqrt{n}}$$

     for random samples of size $n$ from a population with mean $\mu$ and standard deviation $\sigma$.
   - Two difficulties in practice: We likely don't know $\sigma$ and usually compute the standard deviation $s$ of the sample statistic giving an estimate to the standard error of the sample means:

     $$SE = \frac{s}{\sqrt{n}}.$$

     The second difficulty is the distribution is no longer a standard normal. We address this issue below.

   (b) The $t$-distribution

   - The distribution of sample means using the sample standard deviation: When choosing random samples of size $n$ from a population with mean $\mu$, the distribution of the sample means is centered at the population mean $\mu$ and has standard error estimated by

     $$SE = \frac{s}{\sqrt{n}}$$

for $s$ the standard deviation of the sample. The standardized sample means approximately follow a $t$-distribution with $n-1$ degrees of freedom (df). For small sample sizes ($n < 30$), the $t$- distribution is only a good approximation if the underlying population has a distribution that is approximately normal.

- $t$-distributions and normal distributions are very similar. Main advantage with $t$-distributions is they are better for small sample size when we approximate $\sigma$ with $s$.

(c) Conditions for the $t$-distribution: Need the distribution of the population to be approximately normal. For a small sample, this is hard to assess since it can appear skewed or appear to have outliers.

2. Confidence interval for a mean

(a) A sample mean based on a random sample of size $n$ has sample statistic $\bar{x}$ and standard error

$$SE = \frac{s}{\sqrt{n}}$$

If $t^*$ is an endpoint chosen from a $t$-distribution with $n-1$ df to give the desired level of confidence, and if the distribution of the population is approximately normal or the sample size is large ($n \geq 30$), then the confidence interval for the population mean $\mu$ is

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

(b) Determining the sample size for estimating the mean: The margin of error from the confidence interval is

$$ME = t^* \frac{s}{\sqrt{n}} \quad \rightarrow \quad n = \left(\frac{t^* \cdot s}{ME}\right)^2$$

Issues:

- $t^*$ requires $n$, use $z^*$ instead since it should resemble $t^*$ for $n$ large.
- $s$ is computed from the sample which we haven't collected yet. Use a guess of the population parameter $\sigma$, call it $\tilde{\sigma}$.

(c) Determining the sample size for estimating the mean:

$$n = \left(\frac{z^* \cdot \tilde{\sigma}}{ME}\right)^2$$

3. Hypothesis testing for a mean

(a) $t$-test for a mean: To test $H_0$: $\mu = \mu_0$ vs $H_a$: $\mu \neq \mu_0$ (or a one-tail alternative) use the $t$-statistic

$$t = \frac{\text{Statistic} - \text{Null value}}{SE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x}$ is the mean and $s$ is the standard deviation in a random sample of size $n$. Provided the underlying population is reasonably normal (or the sample size is large), the $p$-value of the test is computed using the appropriate tail(s) of a $t$-distribution with $n-1$ degrees of freedom.

## .3 6.3 Inference for a difference in proportions

1. Distribution of for a difference in proportions

(a)

2. Confidence interval for a difference in proportions

(a)

3. Hypothesis testing for a difference in proportions

(a)

### .4 6.4 Inference for a difference in means

1. Distribution of for a difference in means

   (a)

2. Confidence interval for a difference in means

   (a)

3. Hypothesis testing for a difference in means

   (a)

### .5 6.5 Paired difference in means

## Chapter 7: Chi-square tests for categorical variables

### .1 7.1 Testing goodness of fit for a single categorical variable

### .2 7.2 Testing for an association between two categorical variables

## Chapter 8: ANOVA to compare means

### .1 8.1 Analysis of variance

### .2 8.2 Pairwise comparisons and inference after ANOVA

## Chapter 9: Inference for regression

### .1 9.1 Inference for slope and correlation

### .2 9.2 ANOVA for regression

### .3 9.3 Confidence and prediction intervals

## Chapter 10: Multiple regression

### .1 10.1 Multiple predictors

### .2 10.2 Checking conditions for a regression model

### .3 10.3 Using multiple regression

## Chapter P: Probability basics

### .1 P.1 Probability rules

### .2 P.2 Tree diagrams and Baye's rule

### .3 P.3 Random variables and probability functions

### .4 P.4 Binomial probabilities

### .5 P.5 Density curves and the normal distribution