AP Stats Notes

Franklin Chen

11 November 2024 - idk some time in may

Contents

using *The Practice of Statistics for the AP Exam: 6th Edition* by Starnes and Tabor

1 Data Analysis

1.1 What is Statistics?

Definition 1 (Statistics)**.** The science of collecting, analyzing, and drawing conclusions from data.

Data is collected from *individuals* about certain *variables*.

Definition 2 (Individual, Variable)**. Individuals** are objects described in a dataset. Typically people, but not always.

Variables are attributes that can take different values for different individuals.

For example, *individuals* may be households, and *variables* may be region, nnumber of people, household income, etc. It's important to distinguish between *categorial* and *quantative* variables:

Definition 3 (Categorical and Quantative Variables)**. Categorical Variables** are variables whose values can be placed into distinct categories.

Quantative Variables are variables whose values are quantities, typically counts or measurements.

For example, region would be categorical, while household income would be quantative. *Not all numbers are quantative*; eg. zip code.

1.2 Analyzing Categorical Data

1.2.1 One-Variable Categorical Data

Definition 4 (Frequency and Relative Frequency Tables)**. Frequency Tables**shows the number of individuals that have values of a certain category. **Relative Frequency Tables** shows the proportion or percent of individuals in each category.

Note (relative) frequencies are not data; they summarize *data.* Bar graphs and Pie Graphs summarize relative frequency tables.

Beware of misleading graphs; we mainly react to the area of each bar, not the actual height.

	Like Skateboards	Do Not Like Skateboards	Totals
Like Snowmobiles	80	25	105
Do not like Snowmobiles	45	10	55
Totals	125	35	160

MathBits.com

Figure 1: An example two-way table with additional summary information.

1.2.2 Two-Variable Categorical Data

Use a two-way table to summarize data about two categorical variables. These tables can be used to answer questions about *marginal, joint, and conditional relative frequencies.*

Margial relative frequencies give the percent or proportion of individuals that have a given value for one categorical variable. For example, the marginal relative frequency of liking skateboards is $\frac{125}{160} \approx 78.125\%$.

Joint relative frequencies give the percent or proportion of individuals that have a specific value for both categorical variables. For example, the joint frequency of liking both skateboards and snowmobiles is $\frac{80}{160} = 50\%$.

Conditional relative frequencies give the percent or proportion of individuals that have a specific value for one categorical variable relative to other individuals with the same other categorical variable. For example, the conditional relative frequency of those who like snowmobiles out of all individuals that like skateboards is $\frac{80}{125} = 64\%$.

These frequencies can be summarized in *side by side bar graphs, segemented bar graphs, or mosaic plots.*

Graphs and these tables can be used to show **assocation** between two variables. There is association between two variables if knowing the value of one helps to predict the other. For example, knowing that an individual likes skateboards helps predict whether they like snowmobiles ($\frac{80}{125} = 64\%$ vs $\frac{25}{35} \approx 71.4\%$). **ASSOCIATION DOES NOT IMPLY CAUSATION!**

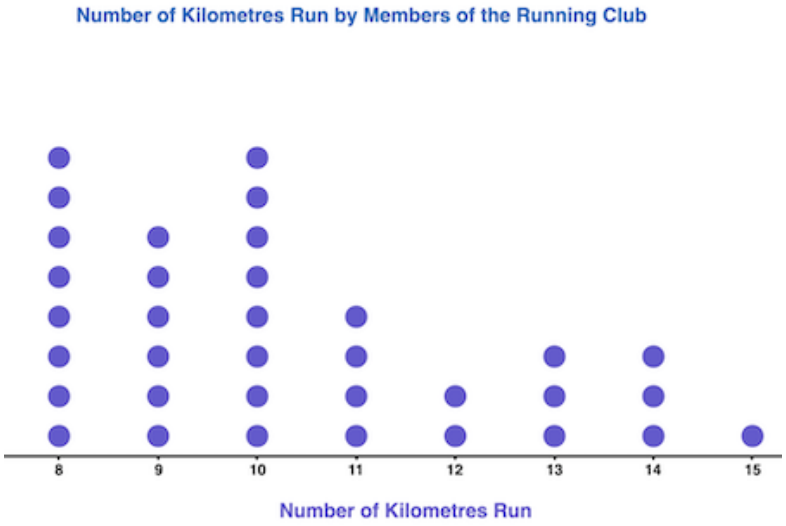

Figure 2: A dotplot showing the distribution of kilometers run by members of the running club.

1.3 Analyzing Quantative Data with Graphs

Dotplots (as shown above) show each individual as a dot above their quantative data value.

When describing the shape of a dotplot (or other quantative graphs), *focus on main features*: major peaks, clusters, or gaps. Especially note whether the distribution is roughly symmetric or skewed:

Definition 5 (Symmetric, Skewed)**.** A distribution is rougly **symmetric** if the right side of the graph has rougly the same shape as the left side.

A distribution is **skewed to the right** if the right 'tail' has less values than the left; typically, the left has a peak whereas the right does not. **Left-skewed** definition are defined similarly to right-skewed distributions.

For example, the distribution of the number of kilometers run is right-skewed because the right 'tail' has less values.

Graphs with a single peak are considered *unimodal*, like the dotplot. Distributions with two peaks are considered *bimodal*, and beyond that is considered *multimodal.*

When describing a distribution of quantative data, use the acronym ROCS: **R**ange (max - min), **O**utliers (clear departures from the data), **C**enter (mean or median), and **S**hape (symmetry, skew, gaps, peaks).

Leaf plots exist. Stem represents first few digits, leaf represents final digit.

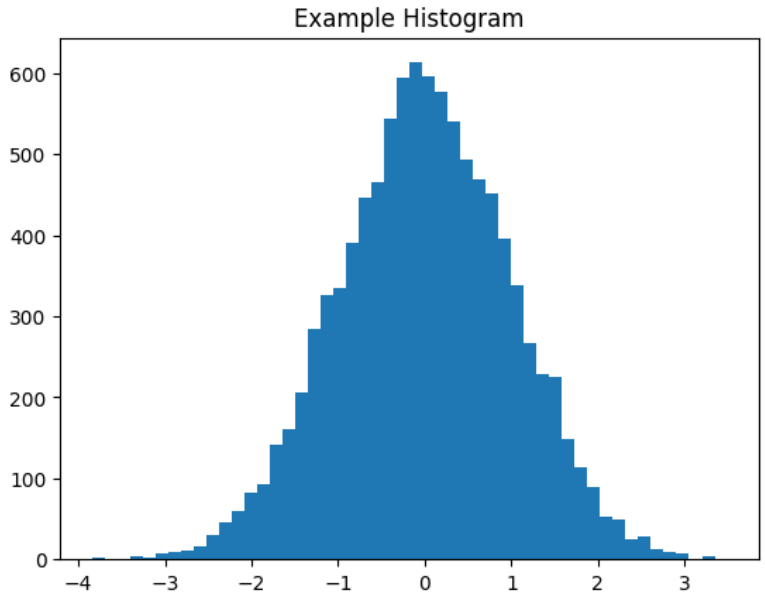

Figure 3: An example histogram with a normal distribution.

Histograms are a notable way of displaying quantative data, as they avoid showing individual data points. Histograms divide the variable into many 'bins' (bars), with the height representing the frequency. Smaller bins show more detail at the cost of a less clear pattern.

Don't confuse histograms and bar graphs. Histograms are used for quantative data, while bar graphs are used for qualatative data.

Use percentages when comapring to distributions in order to remove the effect of a larger sample.

1.4 Describing Quantative Data with Numbers

Definition 6 (Mean: $\bar{x}, \mu$)**.** The average of all individual data values. If there are n observations $x_1, x_2, ..., x_n$, the sample mean is calculated by

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

The mean of a **sample** is referred to using $\bar{x}$, while the mean of a **population** is referred to using μ.

Statistics come from **samples** (small subset of population) and **parameters** come from **populations** (all possible samples of what's being tested).

The mean is not **resistant** as it is sensitive to strong outliers in a distribution. The median *is* a resistant measure of the center of the distribution.

Definition 7 (Median)**.** The 'midpoint' of a distribution. Either the middle element (n is odd) or the average of the two middle elements (n is even) in a **SORTED** distribution.

Using both the mean and the median, one can predict the skew of the data. If a distribution is rougly symmetric without outliers, **the mean and median will be similar**. If a distribution is strongly skewed, **the mean will be pulled in the direction of the skew**. (Mean < Median for left-skewed, Mean > Median for right-skewed)

The **range** (max - min) is one way to show the variability of a distribution. Note that the range is *not* resistant.

Definition 8 (Standard Deviation)**.** The **standard deviation** (s_x, σ) measures the 'average' distance of the values in a distribution from the mean. Standard deviation is calculated by

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

The squared stdev is known as **variance** (s_x^2, σ^2). Remember, s_x refers to a sample while σ refers to a population. Larger stdev indicates greater variation, but is not a resistant measure of variability. **Stdev measures variance around the mean; if the mean is skewed, so will stdev!**

The **Interquartile Range (IQR)** is another way to measure variance, using $IQR = Q_3 - Q_1$ where Q represents the quartiles. IQR can be thought of as the range of the 'middle half' of the distribution. *IQR is a resistant measure.*

$$\text{Lower Outliers} < Q_1 - 1.5 \times IQR \text{ or High Outliers} > Q_3 + 1.5 \times IQR$$

boxplots and the five-number summary = min, Q1, median, Q3, and max exist. **Boxplots don't show gaps, clusters, or multiple peaks.**

be careful with language- 'skews' is a shape, IQR and range are single numbers (no 'in the middle of the IQR')

2 Modelling Distributions

2.1 Describing Location in a Distribution

Definition 9 (Percentile)**.** The pth percentile of a distribution is the value with p% of observations less than (*or equal, depending on who you ask*) than it. Note that this distinction of whether or not to include "or equal" only matters for discrete variables.

For example, in a class of 25 students, if a student gets a score greater than or equal to 21 other students, then they would be at the 84th percentile in the class's test score distribution ($\frac{21}{25} \approx 84\%$). **An obervation is never "in" a percentile; an observation is AT a percentile** (percentile is just a number; like IQR and range). Also, definition, $Q_1 \approx 25$th percentile, Q_2(median) ≈ 50th percentile, and $Q_3 \approx 75$th percentile. Percentiles can be graphically shown in a **cumulative relative frequency graph**, where the y-coordinates of points are graphed based on their percentile.

Definition 10 (Standardized Score (z-score))**.** How many standard deviations an individual value is from the mean. For a value p, mean μ, and stdev σ:

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

z-scores provide a way to objectively compare measurements from two distributions while still considering means and variability.

2.1.1 Transforming Data

Effect of adding a constant a:

- Adds a to measures of center and location (mean, median, quartiles, min, max)
- Does not change measures of variability (range, IQR, stdev)
- Does not change the shape of the distribution (percentile; translation along axis)

Effect of multiplying by constant b:

- Multiplies measures of center and location by b (mean, median, quartiles, min, max)
- Multiplies measures of variability by $|b|$ (range, IQR, stdev)

- Does not change the *overall* shape of the distribution (percentile; like squish or squash)

z-scores transform any distribution into one with mean 0 and stdev 1, but with the same original shape.

2.2 Density Curves

Definition 11 (Density Curves)**.** A curve that models the distribution of a quantative variable with a curve that

- Is always on (or above) the horizontal axis
- Has an area of exactly 1 underneath

The area under the curve within any interval of values estimates the proportion of all observations that fall into this interval.

No set of quantative data is fully described by a density curve. Density curves are approximations that are easy to use by smoothing out small irregularities in the data. Similarly to finite distributions, distributions will have shapes, often with skews and peaks.

The median of the density curve splits the curve into two equal-area halves of area = 0.5. The mean of the density curve is harder to define. For a density curve described by a function $p(x)$, the mean is given as the below. (Intuitively, this is the idea of a 'weighted average'; weight being its relative density- $p(x)$, in this case- being extended to a continous distribution.)

$$\text{mean of } p(x) = \int_{-\infty}^{\infty} xp(x)dx$$

While the mean-median location principles still apply, they aren't too useful as it's relatively hard to locate the mean of a curve by simply looking at it.

2.2.1 Normal Distributions

As $\lim_{n\to\infty} n$, binomial distributions will approach a **Normal distribution** (see Chapter 6 or something), which are described using Normal curves.

Definition 12 (Normal Curve)**.** Any normal distribution is described by a symmetric, single-peaked, bell-shaped density curve. Its center is equal to the mean and the median, and the stdev represents the 'width' of the curve. Any normal distribution is fully described by its mean and stdev. A normal

curve's mean is its peak, while the stdev are its inflection points (symmetric around the mean).

Normal distributions with $\mu = 0$ and $\sigma = 1$ is known as a **standard Normal distribution** (which is the same as any other Normal distribution normalized into z-scores). *Remember that mean and stdev only fully describe Normal distributions!*

All normal distributions follow the **empirical rule**: 68%, 95%, and 99.7% of all observations will fall within 1, 2, and 3 stdev around the mean respectively. `normalcdf(upper, lower, mean, stdev)` can be used to find the area under the normal curve (with a given mean and stdev) between upper and lower. Similarly, `invNorm(area to the left, mean, stdev)` calculates the area's percentile value given the mean and stdev. **When using calculator functions like normalcdf, make sure to 1) label what the inputs mean and 2) answer the question asked AS A SENTENCE!**

2.2.2 Assessing Normality

uhh i'll do it later

3 Two-Variable Data

3.1 Scatterplots and Correlation

When analyzing two sets of data, we usually hope to establish a cause-effect relationship between the **independent / explanatory variable** and the **dependent / response variable**. We want to show that knowing the value of the independent value will help predict the dependent variable's value. **Just because knowing the independent value helps us predict the dependent value does not imply direct causation; there may be some other variable influencing both!**

We can analyze two-variable using a scatterplot. To describe a scatterplot, consider the following:

- **DIRECTION**: Does the dependent variable increase (*positive assocation*) or decrease (*negative association*) with the independent variable? Or does it look like there's no relation between the two? (*no assocation*)
- **FORM**: Is the data linear or nonlinear? Nonlinear data still shows a pattern, although it's harder to describe.
- **STRENGTH**: How closely does data follow the predicted pattern? Not very much (*strong association*) or a lot (*weak assocation*)? Or somewhere between? (*moderate association*)
- **OUTLIERS**: Look at distinct departures from the described pattern and distinct clusters of points.

We can directly quantify correlation for linear relationships using the **(linear) correlation r**.

Definition 13 (Correlation r)**.** A measure of the direction and strength of a linear association. Always between -1 and 1, with values closer to each extreme indicating stronger relationships ($r = \pm 1 \rightarrow$ perfectly linear relationship, $r = 0 \rightarrow$ no linear relationship). Positive r suggests positive association, while negative r suggests negative association.

$$r = \frac{1}{n-1}\sum(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}) = \frac{1}{n-1}\sum z_x z_y$$

On a surface level, the formula for r standardizes each variable to avoid the issue of different linear slopes. Other than that, don't ask why this formula works; it's wayyyy too much of a hassle to reasonably prove.

Obviously, r only works if both variables are quantative; use the word *association* if either variables are categorical. r doesn't differentiate between explanitory and response variables; r will be the same even if you swap the two sets. r also has no units; it's just a scalar.

The linear correlation should only be used when talking about LINEAR relationships: a perfectly quadratic relationship will not give $r = 1$. Linear correlation does not care about form; even if something is curved, linear correlation only measures its 'closeness' to a linear line. **Linear correlation is not resistant; it is GREATLY influenced by extreme points.**

3.2 Least-Squares Regression

Definition 14 (Regression Line)**.** A line of form $\hat{y} = a + bx$ which models how the dependent variable y changes as the independent x changes. ($\hat{y}$ = predicted value of y for given x) **Regression lines require that we consider the x as the independent variable.**

Alternatively, we can think of $\hat{y}$ as the average price for a sample of individuals with the same x value (assuming that our pattern is true). Beware of **extrapolation**: using a regression line to predict outside of the interval of x-values used to obtain the line. The further the values are from the data used to generate the line, the less reliable the prediction is. *Just because regression lines can predict extreme values doesn't mean those values will ever happen!*

3.2.1 Residuals

To determine if a regression line is a good prediction, plot the **residuals** ($y - \hat{y}$ for all data points in the set). No pattern suggests that the regresison model is accurate, while a leftover non-zero pattern suggests that this model isn't appropriate. (Close to the x-axis = strength)

Intuitively, we look at the remaining pattern because **form of residual plot = form of association - form of model**. As ssuch, no residual plot form suggests that form of association = form of model (good), while a residual plot form suggests that form of assocation $\neq$ form of model (bad!!!).

3.2.2 Least-Square Regression Lines

A good regression line minimizes the **sum of the squared residuals.** (Why not the residuals themselves? The sum of all the residuals = 0, so we square it to avoid sign- similar to stdev.) The line that minimizes the squared residuals is known as the **least squares regression line**.

The least-square regression line $\hat{y} = a + bx$ can be calculated as follows:

$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

The formula for the slope can be thought of as "normalizing" the correlation to fit with the specific scenario using the coefficent $\frac{s_y}{s_x}$. (as far as i can tell there's no super intuitive way to think about this. sorry)

The formula for the y-intercept comes from the fact the least-squares regression line will always pass through $(\bar{x}, \bar{y})$. Using simple algebra we rearrange for a.

This is called *regression to the mean* because something something predicted values "regress" to the mean somehow.

3.2.3 Problems with Least-Square Regression Lines

Similar to correlation, least-square regression lines are 1) only used to describe linear relationships, 2) not resistant, and 3) only describe patterns in the data, not direct causation.

2 suggests that points with relatively high or low x-values have **high leverage** over the regression line versus other points in the data set. These **outliers** (original def + high residual) are **influential points**; if removed, they substantially change the regression line or measures of strength (see 3.2.4).

3.2.4 Strength of regression lines

To determine the strength of a regression line, we have two methods: the **standard deviation of the residuals** or the **coefficent of determination** r^2.

The standard deviation of the residuals (s) is pretty obvious: calculate the stdev of the data relative to the prediction. (why n - 2? idk it looks rlly complicated)

$$s = \sqrt{\frac{\sum(y_i - \hat{y}_i)}{n-2}}$$

On the other hand, the coefficent of determination measures the percent reduction in the sum of squared residuals when using the regression line vs. using $\hat{y} = \bar{y}$. In other words, it represents the percent of variability in the response variable that is accounted for by the least squares regression line. ($1-r^2$ can be thought of as the % of variability that's caused by other factors

than the independent variable.) (It's also equal to the square of the linear correlation, hence the symbol r^2. why? idk)

$$r^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

3.3 Transforming to Achieve Linearity

The techniques discussed in 3.1 and 3.2 (obviously) only apply to purely linear relationships. In some special cases, we may be able to **tranform the data** from a non-linear pattern into a linear pattern to apply our techniques.

3.3.1 Power Models

Definition 15 (Power Model)**.** Any regression model that takes the form $\hat{y} = ax^p$. Scenarios modeled by these regression models should be expected to do so by geometry (eg. area proportional to square, volume proportional to cube) or some other factor like physics.

While power models (by definition) describe a non-linear relationship between x and y, there will always be a linear relationship between x^p and y. As such, if we raise x to the pth power or take the pth root of y there should be a linear relationship between x^p and y (or x and $\sqrt[p]{y}$).

However, this all assumes p is known. If p isn't known, you could either guess and check (cringe) or use logarithms.

3.3.2 Transforming with Logarithms

If a scenario is modeled by $y = ax^p$, then $\ln(y) = \ln(ax^p) \to \ln(y) = \ln(a) + p\ln(x)$. Thus, there is a linear relationship between $\ln(x)$ and $\ln(y)$. (*Note that the power p now becomes the slope of the line!*) If the resulting graph turns out ot be linear, we can fit a linear regression line to the graph and then use the relation $\hat{y} = e^{a+b\ln(x)}$ to predict values. (This relationship is the same as the power model: $\hat{y} = e^a \times e^{b\ln(x)} = e^a x^b$; e^a is a constant.)

If a scenario is modeled by $y = ab^x$ (an **exponential model**), then $\ln(y) = \ln(ab^x) \to \ln(y) = \ln(a) + \ln(b)x$. Thus, there is a linear relationship between x and $\ln(y)$. IF the resulting graph turns out to be linear, we can fit a linear regression line to the graph and use the relation $\hat{y} = e^{a+bx}$. (This relationship is the same as the exponential model: $\hat{y} = e^a \times e^{bx} = e^a \times (e^b)^x$: e^a and e^b are constants.)

If unsure if an exponential or power model is better, consider their residual plots. If both models have similarly random residual plots, choose the model with the largest r^2.

4 Collecting Data

5 Probability

6 Random Variables and Distributions

7 Sampling Distributions

8 Confidence Intervals

9 Significance Tests

Significance Tests are like the opposite of confidence intervals. Instead of using a statistic to find a parameter, significance tests use statistics to test claims about a parameter.

Definition 16 (Significance Test)**.** A formal procedure for using observed statistics in order to decide between to competing claims (*hypotheses*) about parameters.

9.1 Basics of Significance Tests

9.1.1 Hypotheses

Definition 17 (Null Hypothesis, H_0)**.** A claim about a parameter that we weigh evidence **against** in a significance test. Usually a statement of 'no difference' (as claimed).

Definition 18 (Alternative Hypothesis, H_a)**.** The claim that we are trying to find evidence for. Directly contradicts the null hypothesis.

This "Null Hypothesis" and "Alternative Hypothesis" can be thought as trying to prove someone "guilty" or "not guilty."

For example, if a player claims they're a 80% free throw player, the null hypothesis would be $H_0 : p = 0.80$, and the alternative hypothesis would be $H_a : p < 0.80$.

The alternative hypothesis is **one-sided** because we suspect that the player makes less than 80% of his free throws. If we believe that it's equally plausible that they make more than 80% of their free throws, then we would use a **two-sided hypothesis**- $H_a : p \neq 0.80$.

Hypotheses express our beliefs before looking at the data. Molding a hypothesis around data shows nothing.

9.1.2 P-values

Definition 19 (P-value)**.** The probability of getting the values observed in the data under the assumption that the null hypothesis H_0 is true.

Small P-values provide convincing evidence for the alternative hypothesis because small values suggest the observed result is unlikely to happen when the null hypothesis is true. Similarly, large P-values provide convincing evidence for the null hypothesis because large values suggest the observed result is likely to happen due to chance if the null hypothesis is true.

In terms of probability notation: P-value = P(observed data | null hypothesis is true).

For two-sided tests, we look at the distance between the null hypothesis and the observed data. For example, if $H_0 : p = 0.5$, $H_a : p \neq 0.5$ and $\hat{p} = 0.65$ (observed proportion), then the P-value = P($\hat{p} \leq 0.35$ or $\hat{p} \geq 0.6$ | $p = 0.5$). We look at $\hat{p} \leq 0.35$ because $|p - 0.35| = |p - 0.65|$.

Based on the P-value, we make a conclusion about data:

- If the P-value is small (unlikely to happen by chance), then we "reject H_0" and conclude that there is convincing evidence for H_a (in context).
- If the P-value is large (likely to happen by chance), then we "fail to reject H_0" and conclude that there is not convincing evidence for H_a (in context).

How small does a P-value have to be in order to reject H_0? We use a given **significance value** for this boundary.

Definition 20 (Significance Level, α)**.** The value that we use as a boundary for deciding whether a P-value is significant enough to disqualify the null hypothesis. *α should be stated before data is produced (cherrypicking).*

P-value $< \alpha \Rightarrow$ reject H_0 $\Rightarrow$ convincing evidence for H_a in context

P-value $> \alpha \Rightarrow$ fail to reject H_0 $\Rightarrow$ not convincing evidence for H_a in context

If P is less than the significance level, we say that the result is "statistically significant at the $\alpha =$ ____ level." Alternatively, "the results were significant ($P = 0.03 < \alpha = 0.05$)." *Keep in mind a P-value is more informative than a statement of significance!*

NEVER "accept H_0" or conclude that H_0 is true! Always use 'reject' or 'fail to reject!'

9.1.3 Type I and Type II Errors

When drawing a conclusion from a significance test, our conflusion may be wrong. There are two types we can make with the conclusion process, helpfully named "Type I" and "Type II" errors. Only one type of error is possible at once.

Definition 21 (Type I and II errors)**.** A **Type I error** occurs when we reject H_0 when H_0 is true; the data gives convincing evidence that H_a is true despite being false.

A **Type II error** occurs when we fail to reject H_0 when H_a is true; the data fails to give convincing evidence that H_a is true despite it being true.

Note $P(\text{Type I error}) = \alpha$. However, significance is inversely proportional to Type II error: as significance decreases, $P(\text{Type II error})$ increases.

9.1.4 Power

Definition 22 (Power)**.** The **power** of a test is the probability that the test finds convincing evidence for H_a given that the parameter is a specific value that does follow H_a. In other words:

$$\text{power} = P(\text{reject } H_0 \mid H_0 \text{ is false})$$

Power depends on a specific value. For example, if the power of a test to detect p = 0.08 is 0.29 given $H_0 : p = 0.10$, if the true proportion in the population is p = 0.08, there is a 0.29 probability that researchers will find convincing evidence for H_a.

Note Power = 1 - P(Type II Error), or written alternatively, P(Type II Error) = 1 - Power. Also note that the power of a test increases with higher sample number, higher signifiance level, further null and alternative parameter values, and 'wise choices when collecting data' (reducing variability). *Larger signifiance values reduce the chance for Type II error, but increase the chance for Type I error.*

9.1.5 Steps for Signifiance Tests

Overall, significance tests can be summarized in four steps:

1. **STATE** the hypotheses, signifiance level, and parameters.
2. **PLAN** the appropriate inference method and check conditions.
3. **DO** the signifiance test (see Section 7.2):
   - Give the sample statistic(s)
   - Calculate the standardized test statistic
   - Find the P-value.
4. **CONCLUDE** whether or not you believe in the null or alternate hypotheses, with reference to P-values, in the context of the problem. Make sure to reference the parameter, not the sample statistic!

9.2 Tests About a Population Proportion

Like confidence intervals, signfiicance tests should satisfy several conditions; random sampling, no bias, 10% if applicable ($n < 0.10N$), and large counts ($np_0 \geq 10$ and $n(1-p_0) \geq 10$). Note that the large counts condition uses p_0: the number of successes and failures *assuming the null hypothesis is true* are both greater than or equal to 10.

To conduct a **one-sample z test for a proportion** (CITE THIS WHEN FOLLOWING THE SIGNIFIANCE TEST STEPS) about a proportion, look at the normal distribution *assuming the null hypothesis is true.*

From Section 7.2 (add ref later), $\mu_{\hat{p}} = p_0$ and $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$.

Using this, we standardize the statistic with respect to the null value to get the **standardized test statistic**; how many stdev units away the sample statistic is from what we would expect, assuming the null hypothesis is true. **YOU HAVE TO PUT THE STANDARDIZED TEST STATISTIC ON THE TEST!**

$$\text{standardized test statistic} = z_{sts} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$$

Using this, we can directly find the P-value using `normalcdf` with the appropriate values:

- $P(z > z_{sts})$ if $H_a : p > p_0$
- $P(z < z_{sts})$ if $H_a : p < p_0$
- $P(z > |z_{sts}|) + P(z < -|z_{sts}|)$ if $H_a : p \neq p_0$

Keep in mind that just because the P-value is low *does not prove that* H_0 *is false*; sample proportions may be small due to sampling variability, or we have made a Type I error somehow. **AVOID JUMPING TO CONCLUSIONS**: just because a sample statistic looks unlikely, doesn't mean its P-value is!

It should be noted there is a link between *two-sided tests* and confidence intervals: a K% confidence interval gives an approximate set of p_0's that would not be rejected by a two-sided test at the $\alpha = \frac{K}{100}$ level, with all other values being rejected.

Intuitively, if a confidence interval for p using *the same* $\hat{p}$ does not include p_0, as the distribution (test) about p_0 has a similar stdev (due to the same α) the distribution is unlikely to include the actual value of p.

9.3 Tests About a Difference in Proportions

Tests about difference in proportions work very similarly to tests about single proportions. For null hypothesis $H_0 : p_1 - p_2 = 0$, large counts can't directly be applied; we need estimate the true difference in proportions p using a weighted average of the two proportions:

$$\hat{p}_C = \frac{\text{number of successes in both samples combined}}{\text{number of individuals in both samples combined}} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

We use this pooled proportion (*even for non-zero differences*, although the given formulas have to be changed) because the combined proportion provides *the best estimation of the true difference in proportions.* In the one-sample z test for a proportion, we used p_0 because we're looking at the distribution relative to it. For **two-sample z tests for the difference between two proportions**, we look at the distribution 'relative' to the true proportion; which our best guess is $\hat{p}_C$. (something something using each individual $\hat{p}_i$ causes too much error + causes some error in the underlying math when we're testing for 0 diff)

In this case, we use the combined proportion for large counts *for both samples*: that is, $n_1\hat{p}_C, n_1(1-\hat{p}_C), n_2\hat{p}_C, n_2(1-\hat{p}_C)$ all need to be greater or equal than 10. Using the combined proportion works for a non-zero difference in the null hypothesis, although is not strictly neccessary (large counts for each individual distribution can be used).

The standardized test statistic also uses the combined estimate (for both proportions are equal; otherwise, each individual statistic can be used):

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

10 Estimating Means

10.1 One-Sample *t* intervals

To create a confidence interval for a mean, we can use the same point estimate $\pm$ margin of error = (critical value)(SE of statistic) idea. However, this is slightly flipped, as (from Section 7) the stdev of a sample distribution relies on the stdev of the population itself. If we know the stdev of the population, we can the familiar *one-sample z interval for a population interval*:

$$\hat{\mu} = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Using the sample stdev alone is not a good enough approximation. For an example, using the sample stdev to estimate a 99% confidence interval would result in a $\approx 92.5\%$ confidence interval. This is because there is significant variation in the sample stdev itself.

To account for this, we use the larger *t-distribution* instead. The t-distribution is like the z-distribution (measures how many stdevs needed for desired confidence), but has higher area in the tails to compensate for the errors in the sample stdev. Notably, t-distributions vary depending on the sample size; that is, they have a specific *degree of freedom df.* (Use `invT` to calculate the specific t^* value needed.) When using t-distributions, make sure to set $df = n - 1$. Thus, for the *the one-sample t interval for a population mean,*

$$\hat{\mu} = \bar{x} \pm t^* \frac{s_x}{\sqrt{n}}$$

The random (and 10%) clauses still apply, although the Large Counts condition is slightly modified due to the central limit theorem:

- If the sample size is small ($n < 30$), then graph the sample data (dotplot is preferred, although boxplot works it's kinda funky) and consider whether the data plausibly comes from a normal population. No strong outliers or skew indicates yes. *You will have to show your graph when assesing the normality of the data.*
- If the sample size is big ($n \geq 30$), then you can still use a one-sample t interval for a population mean (although the true confidence level will be lower than presented).

The `TInterval` function on the calculator can be used to simulate a one-sample t interval for a population mean if the above are satisfied.

As df $\to \infty$, the t-distribution approaches the normal distribution. Intuitively, as sample size increases, the sample stdev will get closer to the population stdev. The quantity $\frac{s_x}{\sqrt{n}}$ is known as the *standard error of the sample mean* ($s_{\bar{x}}$).

10.2 Two-Sample *t* intvervals for a Difference between Two Means

Like two-sample z intervals, these require the one-sample requirements to be fufilled for the populations, and follow the same idea (sqrt the variance):

$$\hat{\mu}_{x_1-x_2} = (\bar{x_1} - \bar{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

However, the degrees of freedom are significantly more confusing to decide. There are two practical options to decide the df to use:

- **Technology**: Use the formula below to calculate df. Note that in this case, df is usually not a whole number. Results in confidence intervals that are approximately the stated confidence levels.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

- **Conservative**: $df = \min(n_1 - 1, n_2 - 1)$. Note this method causes a margin of error as large if not larger than is needed for the desired confidence level.

10.3 Paired *t* interval For a Mean Difference

(also known as a one-sample t inerval for a mean difference)

For comparing data which interviews the "same" (or highly related) individuals (known as **paired data**), "combine" the data into one dataset by subtracting each individuals' values, and analyze that using the tactics from section 10.1. (n = the number of pairs)

The requirements are the same as the 1-sample t interval, although the meaning of 10% ($n_{diff} < 0.10N_{diff}$) is different: we suspect that the data is not independent, so 10% guarantees that the differences are independent.

11 Confidence with Means

12 Chi-Square Tests

13 Inference for Slopes and Tables

$$\prod_{k=1}^{4}(x_k^2+1)=\prod_{k=1}^{4}(x_k-i)(x_k+i)=\prod_{k=1}^{4}(x_k+i)\times\prod_{k=1}^{4}(x_k-i)$$

$$=P(i)P(-i)=(b-d+1)^2+(a-c)^2=(5-1)^2+0=\boxed{16}$$