AP Statistics Notes

anastasia

Fall 2024 & Spring 2025

Contents

1 Exploring One-Variable Data

1.1 Representing Categorical and Quantitative Variables with Graphs

Data contains information about a group of individuals. The information is organized using variables.

Individuals are objects described by a set of data. Individuals may be people but may be animals or inanimate objects.

Variables are characteristics of individuals. A variable may take on different values of different variables. Variables can be split into two types: categorical or quantitative.

Categorical variables place individuals into specific groups.

Quantitative variables takes on numerical values for which it makes sense to do arithmetic operations like adding and averaging. Quantitative variables fall into two categories: discrete and continuous.

Be careful - just because it is a number doesn't make it quantitative.

Discrete variables are numerical values where counting makes sense; in other words, decimals would not be an appropriate way to record the data.

Continuous variables are numerical values where decimals are appropriate; it usually involves some form of measuring.

The difference between discrete and continuous isn't always clear. An example of this would be age.

One of the easiest ways to display categorical data is with a table.

Count is the amount of that category in a table and relative count is

$$\frac{\text{count}}{\text{total}}$$

If you wanted to display two categorical variables at a time, we could make a two way table.

To better visualize the data, there are graphs that we can make from the data. We want to visualize the graphs to get a better idea of the distribution.

Distribution of a variable tells us what values the variable takes and how often it takes these values.

Bar Graphs have the following characteristics:

- Label each axis clearly
- The x-axis will contain the categorical variable and the y-axis will display teh counts
- Each category has its own bar and the bars cannot touch
- Order is not important when creating the x-axis

To make a histogram, we need to put the data into even intervals that capture our data. We will do this first by hand by counting how many data scores are in each bin.

To find the interval width, we can use the formula

$$\frac{\text{max-min}}{\text{\# of wanted intervals}}$$

To make the histogram:

- Draw rectanges for each interval with height representing the count
- Bars must touch
- Label the x-axis with the lower bound values of each interval

1.2 Representing Quantitative Variables with Graphs

Stemplots (or stem and leaf plots) are an alternate way to illustrate data using a semi-graph. It is similar to a histogram, but the data isn't lost. If the data has two digits, the stem is the first digit and the leaf is the second. If the data has 3 digits, the stem is the first two digits. You must always add a key to the graph.

Back-to-Back Stemplots are created when you can separate the data into two categories. The stems are the same, but the data can be split into different categories. You still must have a key for both sides.

Split Stemplots are the last type of stem and leaf plots. You can split the stem; in a similar way to creating more bins on a histogram if the bin width resulted in a skyscraper.

Dot plots are a very simple type of graph that involves plotting data values with dots above the values on a number line.

To construct a dot plot:

- Label your axis and title your graph. Draw a horizontal line and label it with the variable.
- Scale the axis based on the values of the variable.
- Mark a dot above the number on the horizontal axis corresponding to each data value.

Cumulative relative frequency graphs display percentiles. A percentile will tell you what percent of data falls above a value.

In order to create one of these graphs, you must make a table of the cumulative relative frequencies in order to graph it. This can be done by first finding the relative frequencies and then you add them together to get the cumulative relative frequency. This graph is also called an ogive.

1.3 Describing Distributions of Quantitative Variables

We can describe the distribution of a quantitative variable by its shape, outliers, center, and spread.

After graphing the distribution, the first thing we identify is the shape. A unimodal shape has one peak, a bimodal shape has two peaks, a uniform shape has no peak. If a graph is skewed it will be skewed in the direction in which the tail is located.

There are three measures of center: mean, median, and mode. Mean is defined as

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

, where $\bar{x}$ is the mean of sample, x_i is each individual observation, and n is the number of observations. Mean is called non-resistant because the mean is strongly influenced by outliers.

Median is a different measure of center that is resistant. The median is found by ordering the data and finding the middle value in the list. The location of the median is

$$\frac{n+1}{2}$$

Mode is the most occurring value in a data set.

To summarize, a unimodal symmetric graph will have the median, mean, and mode similar to each other. A unimodal, left-skewed graph will have mean $<$ median $<$ mode, and a unimodal, right-skewed graph will have mode $<$ median $<$ mean.

There are three measures of spread: range, standard deviation, and IQR.

Range is the difference between the maximum number and the minimum number.

The standard deviation is the average deviation of an observation from the mean of the data set. It is calculated by

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

, where s_x is the standard deviation, and s_x^2 is the variance of sample.

The variance is the average squared deviation. The quantitative variable typically varied by the mean by "standard deviation units".

Standard deviation has a few properties:

- The standard deviation is always positive.
- The standard deviation is always 0 when all observations are equal.
- The standard deviation has the same units of measure as the original variable measured.
- The standard deviation is non-resistant.
- The greater the standard deviation, the greater the distribution.

IQR is inter-quartile range and it uses percentiles to describe the spread of distribution. The 0th percentile is the minimum, the 25th percentile is Quartile 1, the 50th percentile is the median, the 75th percentile is Quartile 3, and the 100th percentile is the maximum.

An outlier is an individual piece of data that falls outside the overall pattern of the distribution. We can determine if a point is an outlier:

- First, find the five-number summary (the percentiles mentioned above)
- Find the IQR by subtracting the value of Quartile 1 from the value of Quartile 3
- Compute Quartile 1 - (1.5 * IQR). Any data above this number is an outlier.
- Compute Quartile 3 - (1.5 * IQR). Any data above that number is an outlier.

Using this data, we can make box plots or modified box plots depending on if the data set as an outlier determined.

1.4 Comparing Distributions of Quantitative Variables

In the world of statistics, it is not enough to just report the graph or just report the summary statistics.

Given a set of data, first you must create a graph for the data. If the data is categorical, use a bar graph. If the data is quantitative, if it's discrete, use a dot plot, stem plot, or boxplot. If it is continuous, use a histogram or box plot.

You want to summarize your findings after this. First, describe the shape. Next, describe if there are any outliers. Then, use the mean or median for the center of data. Lastly, describe the spread using range or IQR if you described the center with the median, or standard deviation if you described the center with mean.

1.5 Z-Scores and the Empirical Rule

Normal distributions are appropriate for many distributions whoes shapes are unimodal and approximately symmetric.

In the normal distribution, Mean $= \mu$, Standard deviation $= \sigma$, and this is written as

$$N(\mu, \sigma)$$

A normal distribution curve has these properties:

- Symmetric around the mean
- mean = median = mode
- 50% of observations are greater than the mean and 50% are less than the mean
- The standard deviation tells us how measurements for a group of observations are spread out from the mean
- In a normal distribution, approximately all the data lies 3 standard deviations above and below the mean

In a normal distribution, 68% of the data is likely to be found within 1 standard deivation from the mean, 95% found 2 standard deviations away, and 99.7% 3 standard deviations away.

To calculate how many standard deviations away from the mean an observation is, we use a z-score.

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

Observations larger than the mean have positive z-scores, and observations smaller than the mean have negative z-scores.

1.6 The Standard Normal Curve

Using a calculator, we can determine what percentage of data falls above, below, or between specific z-scores.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 2:normalcdf(), we can find the percentage of data in between two values.

We can also find the z-score that corresponds to a percentile.

Using the calculator commands: 2ND $\rightarrow$ Vars $\rightarrow$ 3:invNorm(), we can find this.

2 Exploring Two-Variable Data

2.1 Two Categorical Variables

A side by side bar graph merges two bar graphs into one, in an attempt to compare the distributions of the two categorical variables.

A segmented bar graph is another way to display data, where each group is split by its relative frequency.

A mosaic plot is similar to a segmented bar graph, but it draws attention to the sizes of each group.

Joint relative frequencies are the ratio of the frequency in a cell and the total number of data values.

Marginal relative frequencies is the ratio of the sum in a row or column and the total number of data values.

Conditional relative frequencies are the ratio of a joint relative frequency and related marginal relative frequency.

Basically - joint relative frequency is the cell count divided by the table total, marginal is the row/column total divided by the table total and the conditional relative frequency is the intersection divided by the row/column total.

2.2 Scatterplots and Correlation

We have worked with bar graphs, box plots, and histograms. This is called variate data.

When we compare two variables or bivariate data, we are exploring the relationship between the two.

You should start with a scatter plot. A scatterplot shows the representation between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the plot fixed by the values of both variables for that individual.

A response variable measures the outcome of a study or an observation.

An explanatory variable helps explain or influences change in a response variable.

Commonly, explanatory variables are called independent and respondent are called dependent.

You can describe the overall pattern of a scatterplot by the direction, form, and strength of the relationship. An important kind of deviation is an outlier.

Form is the overall pattern or deviations from the pattern. Direction is whether the graph has a positive or negative slope. Strenght is how close the points lie to a simple form (such as a line).

Here are the steps to creating a scatterplot on a calculator:

- Load the x-values into list 1 and y-values into list 2.
- Using StatPlot - highlight the mini-scatterplot: XList:L1 and YList:L2
- Zoom:9 to fit the scatterplot and then graph

In order to strengthen the analysis when comparing two variables, we can attach a number, called the correlation coefficient (r), to describe the linear relationship between two variables.

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

The formula to find the correlation is:

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_1 - \overline{y}}{s_y} \right)$$

Correlation is a number between -1 and 1. The strongest correlations are closer to 1 or -1.

Correlation describes only linear relationship between two variables.

Correlation does not have units and changing units on either axis will not affect correlation.

Switching the x and y axes will not change the correlation.

Correlation is very strongly affected by outliers.

2.3 Linear Regression

Linear regression or least squares regression allows you to fit a line to a scatterplot in order to be able to better interpret the relationship between two variables as well as to make predictions about the response variable.

The fitted line is called the line of best fit and has an equation:

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

The way the line is fitted to the data is through a process called the method of least squares. The main idea is that the square of the vertical distance between each data point and the line is minimized.

Using a calculator we can find the slope of the least squares regression line by doing:

$$\text{Slope} = r\left(\frac{S_y}{S_x}\right)$$

Using this, we can find the y-intercept as well:

$$\text{y-intercept} = \overline{y} - \text{slope} \cdot \overline{x}$$

We can get the line by using Stat $\rightarrow$ Calc $\rightarrow$ 8:LinReg(a+bx)

Correlation does not imply causation.

To describe the strength of a prediction, we use the coefficient of determination. Basically we just use r^2, and this gives the proportion of variation in the values of y that is explained by least-squares regression on y on x.

A residual is a vertical distance between an observed value of the response variable and the value predicted by the regression line.

Residual value is the Actual value minus the Predicted value.

A residual plot is plotting residuals against the explanatory variable. Essentially, it turns the regression line horizontal.

In order to draw a residual plot, you must first perform a LinReg. Next, create a StatPlot where XList is L1 and YList is RESID (From 2nd $\rightarrow$ Stat -¿ :(this one depends on the calculator))

2.4 Influential Points and Departure from Linearity

The standard deviation of the residuals gives the approximate size of a "typical" prediction error. Large values means our line is expected to give larger residuals.

An influential point in a data set, is a point that has leverage on the correlation and regression line.

If your scatterplot when graphing data does not show a linear pattern, or the residual plot pattern is not random, consider transforming the graph.

Some of the most common patterns involve transforming the x or y variables by the natural log or a square root.

3 Collecting Data

3.1 Planning a Study

In order to better understand the characteristics of a population, statisticians and researchers often use a sample from that population and make inferences based on the summary results from the sample.

A population is the entire group we want information from.

A sample is a part of the population we actually examine.

A census collects data from every individual in the population.

An observational study observes individuals and measures variables of interest but does not attempt to influence the responses.

An experiment deliberately imposes some treatment on individuals to measure their responses.

It is only appropriate to make generalizations about a population based on samples that are renadomly selected or otherwise representative of that population.

A convenience sample uses subjects that are readily avaliable.

A voluntary response sample is a sample obtained by allowing subjects to decide whether or not to respond.

A simple random sample consists of n individuals from the popultion chosen in such a way that every set of n individuals has an equal chance in the sample selected.

A stratified random sampling is when you divide the population into groups of similar individuals then select a SRS within each strata. Combine the SRSs from each strata to form your full sample.

Cluster sampling is dividing the population into sections (clusters) then randomly choose a few of these clusters. Every member of the cluster becomes your sample.

Systematic random samplling is one randomly selects an arbitrary starting point and then select every kth member of the population.

When an item from a population can only be selected once, this is called without replacement. When it can be selected more than once, it is called with replacement.

Samples are biased if they are systematically not representative of the desired population.

Voluntary response is when a sample is comprised entirely of volunteers or people who choose to participate, the sample will typically not be representative of the population.

Undercoverage occurs when some groups in the population are left out of the process of choosing a sample.

Non-response occurs when an individual chosen for a sample can't be contacted or refuses to respond.

Response bias is bias cuased by the behavior of the respondent or interviewer.

Untruthful answers occur when people give untruthful answers for several reasons.

Ignorance is when people will give silly answers just so that they appear to know something about the subject.

Lack of Memory is giving a wrong answer simply because the respondent doesn't remember the correct answer.

Timing is when a survey is taken can have an impact on answers.

Phrasing is subtle differences that can make a large difference in results.

When drawing a sample, two types of errors may occur:

Sampling Error: The difference between a sample result and the true population result. This error results from chance variation.

Non-sampling Error: Occurs when the sample data are incorrectly collected, recorded, or analyzed. Usually occurs when the sample is selected in a non-random fashion.

3.2 Selecting a Random Sample

The Hat Method:

- Write down all the names or numbers on their own slip of paper. Then put all the pieces of paper into a hat, mix well in-between selections, and pull out the desired number of slips.

Calculator Random Number Generator:

- MATH - PROB - 5: randInt(lower, upper, n)

Random Digit Table:

To choose SRS with a random digit table, you must label, identify how any digits you will take at a time, indicate when you want to stop sampling, and use the random numbers to identify subjects to be selected from your population.

An observational study observes individuals and measures variables of interest but does not attempt to influence the reponse.

An experiment deliberately imposes some treatment on individulas to measure their responses.

Experimental Unit: the things on which the experiment is done

Subjects: experimental units that are human beings

Treatment: a specific experimental condition applied to the units

3.3 Experimental Design

Factor: The explatory variables in an experiment

Level: the various groups the factors take

Principles of Experimental Design

1. Comparison - we need to make sure we are using a design that compares two or more treatments

2. Randomization - Randomization produces groups of experimental units we expect to be similar in all respects before treatment is applied.

3. Control - Control group is treated identically in all respects to the group receiving the treatment except that the members of the control group do not receive the treatment.

4. Replication - Use enough experimental units in each group so that any difference in the effects of the treatment can be distinguished from change differences between groups.

Experimental terms

- Placebo: a "dummy" treatment
- Placebo Effect: subjects receiving the placebo have a response that is similar to what we would expect if they received the treatment
- Single Blind: when the subject does not know what treatment they are receiving to remove the power of suggestion
- Double Blind: experiments in which the subject and administrator do not know who receives the treatment

In a poorly designed experiment, it might be difficult to tell if the explatory variable causes a change or if it was another variable that wasn't measured.

Confounding variables are variables that might affect the outcome, but we did not control or account for them in our experiment.

One way that we have seen already of removing the effect of any confounding variables is to randomly assign subjects to the treatment or control group. This way any possible bias in population should be evenly spread among the treatment and control groups. Sometimes instead of relying on randomization to make the groups as even as possible we actually force the groups to be similar.

An extraneous variable is one that is not an explanatory variable in the study but is thought to affect the response variable.

Confounding variable refers to another variable that may affect the response and is in some way tied together with the factor under investigation. It leaves us unable to tell which of the two variables caused the observed response.

Lurking variable refers to a variable that drives each of the two variables under investigation, making it appear there's some association between them.

Inference is drawing conclusions beyond the data at hand.

Random selections of individuals allows inference for the population and random assignment in an experiment allows inference for cause and effect.

Both random sampling and random assignment introduce chance variation into a statistical study.

To write an experiment:

1. Determine what type of design is best for your experiment.
2. Diagram your experiment.
3. Tell exactly how you will randomly assign variables.
4. Explain exactly what you are comparing once you gather the data.

Completely Randomized Designed:

- The experimental units are assigned to treatments completely by chance.
- Treatment groups and control groups will be about equal in size in a completely randomized design.
- There are mathematical reasons for having groups of equal sizes.

Randomized Block Design:

- When groups of experimental units are similar, it's often a good idea to gather them together into blocks.
- Blocking isolates the variability due to the differences between the blocks so that we can see the differences due to the treatments more clearly.
- When randomization occurs only within the blocks, we call the design a randomized block design.
- Control what you can, block on what you can't control, and randomize to create compatible groups.

Matched Pairs Design:

- These are experimental designs in which either the same individual or two matched individuals are assigned to receive the treatment and the control.
- Often the "pair" in a matched pairs design is just one experimental unit which serves as its own control.
- In the case where an individual receives both the treatment and the control, the order in which this happens should be random.

4 Probability, Random Variables, and Probability Distributions

4.1 Basic Probability and Simulations

- Outcomes are governed by change, but in many repetitions, a pattern emerges.
- Sample space: all of the possible outcomes.
- Event: a specific desired outcome or set of outcomes.
- Notation: probability of Event A $\rightarrow$ P(A).
- Range of probabilities: probability of event is between 0 and 1.
- Sum of probabilities: probability of whole sample is 1.

Theoretical probability is what should happen given the sample space: # of desired outcomes/total # outcomes.

Empirical probability is when you are performing a simulation or experiment, it is what does happen given the trials: # successes/# trials.

The law of large numbers states that simulated probabilities tend to get closer to the true probability as the number of trials increases.

The notation for the probability that an event does not occur is $P(A^C)$. The probabiliy an event does not occur is $P(A^C) = 1 - P(A)$.

- The idea of probability is that randomness is predictable in the long run.
- Probability does not allow us to make short run predictions.
- Probability tells us random behavior events out in the long run.
- Future outcomes are not affected by past behavior.

The imitation of chance behavior, based on a model that accurately reflects the situation, is called a simulation. Simulations are usually done with a table of random digits, random number generator, dice, deck of cards, spinner, etc.

Four principles of simulation:

- State: identify the probability calculation.
- Plan: Describe how to use your chance process.
- Do: Perform the simulation
- Conclude: use the results of your simulation to answer the question.

4.2 The Addition Rule

- When two events have no outcomes in common, we refer to them as mutually exclusive events.
- P(A and B) = 0
- P(A or B) = P(A)+P(B)

If A and B are any two events resulting from some chance process, then

P(A or B) = P(A) + P(B) - P(A and B)

Notice how if A and B are mutually exclusive, P(A and B) = 0 and P(A or B) = P(A) + P(B).

4.3 Venn Diagrams and the Multiplication Rule

- We can represent events with a Venn diagram - a display of potential probabilities.
- The box around the Venn diagram represents the total sample space where P(S) = 1.
- The circles themselves represent the probability of each event.

Union: $\cup$ is the probability that either occurs.

Intersection: $\cap$ is the probability both will occur.

- Two events are independent if they do not influence one another.
- The occurrence of one has no effect on the occurrence of the other.
- If A and B are independent events, then the probability that both A and B occur is found using the multiplication rule:

$$P(A \text{ and } B) = P(A \cup B) = P(A) \cdot P(B)$$

- Two events are dependent if they influence one another.
- The occurrence of one affects the occurrence of the other.
- If A and B are dependent events, then the probability that both A and B occur is found using:

$$P(A \text{ and } B) = P(A \cup B) = P(A) \cdot P(B|A)$$

- Where P(B|A) is the probability B "given" that A has occurred.
- The conditional probability of an event is the probability that one event will happen, if it is known that another event has happened.

4.4 Conditional Probability and Tree Diagrams

- If A and B are dependent events, then the probability that both A and B occur is found using $P(A \cup B) = P(A) \cdot P(B|A)$ where P(B|A) is the probability of B "given" that A has already occurred.
- The conditional probability of an event is the probability that one event will happen, if it is known that another event has happened.
- We can rearrange the multiplication rule a little and get a formula for conditional probability.

$$P(B—A) = \frac{P(A \cup B)}{P(A)}$$

- If we draw two cards with replacement, we know that those events are independent.
- If we draw two cards without replacement, we know that those events are dependent. Using the following formula we can prove independence:

$$P(A \cup B) = P(A) \cdot P(B) \text{ or } P(A) = P(A|B)$$

4.5 Discrete and Continuous Random Variables

A random variable takes numerical values that describe the outcomes of some chance process. Random variables involve the same probability rules we have already learned, but we will extend those rules to be able to model more events.

A probability distribution tells us the value that our random variable can take and the probability associated with each value.

Every probability distribution must satisfy each of the following requirements.

1. There is a numerical random variable X, and each of the values of X has an associated probability with it.
2. The sum of all the probabilities in the distribution $\sum P(x)$ must equal 1.
3. $0 \leq P(x) \leq 1$ for all probabilities in the distribution.

A discrete random variable takes on a fixed set of possible values with whole number outcomes.

- Random variables are usually capital letters.
- Random variables can be discrete or continuous
- Random variables must be numeric in value

The mean of a discrete random variable X, is the mean outcome for infinitely many trials.

We think of this as the "expected" value because it is the average value we would expect to get if the trials could continue indefinitelyindefinitely.

To find the mean, μ_x, or expected value, $E(x)$, of a discrete random variable X, multiply each possible value by its probability, then add all the products.

$$E(x) = \sum x_i \cdot p_i$$

The standard deviation of a random variable X is a measure of how much the values of the variable typically vary from the expected value.

$$\sigma_x^2 = Var(X) = \sum (x_i - \mu_x)^2 \cdot p_i$$

$$\sigma_x = \sqrt{\sum (x_i - \mu_x)^2 \cdot p_i}$$

- Continuous Random Variables take on all possible values in an interval of numbers. The probability distribution of X is described by a density curve.
- The probability of any event is the area under the density curve and above, below, or between the values x that define the event.
- Area under a density curve is always equal to 1.
- The probability is on the y-axis in a distribution, so finding the area equates to finding the probability of getting an interval of numbers.
- The probability at an event is 0 for a continuous random variable because the area under a point is 0.

Calculator: Mean and Standard Deviation from a Probability Distribution Table

- Enter outcomes into L1 and probabilities in L2
- Run 2nd - Vars - 1VarStats with List:L1 and FreqList:L2
- Mean will be represented by $\overline{x}$
- Standard deviation will be represented by σ_x

4.6 Combining Random Variables

If X and Y are two independent random variables:

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

$$E(X-Y) = E(X) - E(Y)$$

$$\sigma_{X+Y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

$$\sigma_{X_Y} = \sqrt{\sigma_x^2 + \sigma_y^2}$$

If X is a random variable and a and b are both constants:

$$E(a + bX) = a + bE(x)$$

$$\sigma_{a+bX} = |b|E(x)$$

4.7 The Binomial Distribution

Requirements for a Bernoulli Trial:

1. Two Possible Outcomes
2. Probability of Success is the Same for Each Trial
3. Trials are Independent

If we take a Bernoulli trial and then we are interested in the number of successes in a specific number of trials, we create a Binomial Probability Distribution.

There are four requirements for a setting to follow the binomial probability distribution. As long as these four things are true, we can use the binomial probability model for calculating the probabilities.

- BINARY: Each trial falls into one of two categories - we call them "success" or "failure". Success does not necessarily mean a positive outcome, but instead, the outcome we are looking for.
- Independent Trials
  - Each trial is independent of the next
  - Reasonable assumption for coins, cards with replacement, spinners, rolling a die
  - What about sampling without replacement, when it can't be avoided? This is where the "10% condition" comes in. The 10% condition says that if you are sampling from a large enough population, you can proceed with sampling without replacement as though the trials were independent.
- Number of Trials is fixed: we are counting the number of successes from a set number of trials (called n)
- Same probability: The probability of successes (called p) is the same for each trial

If all four points are satisfied, you can compute the probability you get a certain number of successes in a specified number of trials:

n is the number of independent trials, p is the probability of successes, and x is the number of successes.

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

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

The mean is $\mu_x = np$ and the standard deviation is $\sigma_x = \sqrt{np(1-p)}$.

Distributions in your calculator often have "pdf" and "cdf".

pdf is the probability distribution function and is used when trying to calculate $P(X = k)$ and cdf is the cumulative distribution function and is used when you are trying to calculate $P(X \leq k)$.

Specifically for Binomial distributions we want

- 2nd - VARS - A: binompdf(trials, probability of success, k) for $P(X = k)$
- 2nd - VARS - B: binomcdf(trials, probability of success, k) for $P(X \leq k)$

4.8 The Geometric Distribution

WHen we perform many independent trials of the same chance process and are interested in the occurrence of the first success, a geometric setting arises.

If X has a geometric distribution with a probability of success p on each trial and x represents the trial that you get your first success, the probability that X equals x is given by

$$P(X = x) = (1 - p)^{x-1}(p)$$

Specifically for geometric distributions we want...

- 2nd - Vars - E: geometpdf(probability of success, k) for $P(X = x)$
- 2nd - Vars - F: geometcdf(probability of success, k) for $P(X \leq x)$

For this distribution, $\mu_x = \frac{1}{p}$ and $\sigma_x = \frac{\sqrt{1-p}}{p}$

5 Sampling Distributions

5.1 Sampling Distributions of Sample Proportions

As we begin to use sample data to draw conclusions about a larger population, we must be clear about whether a number describes a sample or a population.

Parameter: is a number that describes some characteristic of a population

Statistic: is a number that describes some characteristic of a sample

Unbiased Estimator: a sample proportion or sample mean that is equal to the population proportion or population mean

$\hat{p}$ (sample proportion) estimates p (population proportion)

$\overline{x}$ (sample mean) estimates μ (population mean)

s_x (sample standard deviation) estimates σ (population standard deviation)

Rather than showing real repeated samples, imagine what would happen if we were to actually draw many samples. Now imagine what would happen if we looked at the sample proportions for these samples. The histogram we'd get if we could see all the proportions from all possible samples is called the sampling distribution of the sample proportions.

- We would expect the histogram of the sample proportions to center at the true proportion, p, in the population.
- The spread is calculated as standard deviation based on the true proportion, p, and the sample size, n. As the sample size gets larger the standard deviation will get smaller.
- The shape of the histogram would be unimodal and symmetric.
- More specifically, a normal model is just the right one for the histogram of sample proportions.

Assumptions and Conditions

- Randomness: The sample should be a simple random sample of the population
- Independence (10% Condition): The sample size, n, must be no larger than 10$ of the population
- Normality (Large Counts Condition): The sample size has to be big enough so that both number of successes and number of failures are at least 10. We also refer to this as the Success/Fail Condition.

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of p is modeled by a normal model with:

Sample proportions: $\hat{p} = \frac{\#\text{successes}}{\text{sample size}}$

Mean of sample proportions: $\mu_{\hat{p}} = p$

Standard deviation of sample proportions: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

5.2 Sampling Distributions of Sample Means

The sampling mean distribution notation:

- Parameters: μ and σ
- Statistics: $\bar{x}$ and s
- Sampling Distribution Mean: $\mu_{\bar{x}}$

- Sampling Distribution Standard Deviation: $\sigma_{\bar{x}}$

Conditions for sample means:

- Random - As long as the sampling method is random, our mean is an unbiased estimator.
- Independent - When sampling, we have to make sure the 10% Condition is satisfied.
- Normal - No longer checking Large Counts, instead we have the Central Limit Theorem.

Central Limit Theorem: The central limit theorem (CLT) states that when the sample size is sufficiently large, a sampling distribution of the mean of random variable will be approximately normally distributed.

The central limit theorem requires that the sample values are independent of each other and that n is sufficiently large.

Therefore, if the population distribution is normal, then so is the sampling distribution of $\bar{x}$. This is true no matter what the sample size of n is.

If the population distribution is not normal, the central limit theorem tells us that the sampling distribution of $\bar{x}$ will be approximately normal in most cases of $n \geq 30$.

Summary:

- Shape: approximately normal
- Center: $\mu_{\bar{x}} = \mu$
- Spread: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

5.3 Combining Sample Proportions and Sample Means

Sampling distribution of $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1$ is the sample proportion from the first group and $\hat{p}_2$ is the sample proportion from the second group.

If both samples were randomly selected from the population, then $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$.

If both samples satisfy the 10% condition, then $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}$

If both samples satisfy the success/fail condition, then the shape is approximately normal.

Samplying distribution of $\bar{x}_1 - \bar{x}_2$ where $\bar{x}_1$ is the sample mean from the first group and $\bar{x}_2$ is the sample mean from the second group:

Same conditions apply for the center and spread, and remember that the central limit theorem rather than success/fail is used for shape.

The center is $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and the spread is $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

6 Inference for Categorical Data: Proportions

6.1 Constructing a One Proportion z-Interval

Definition: A confidence interval for a population parameter is an interval of plausible values for that unknown parameter.

It is constructed in such a way so that, with a chosen degree of confidence, the value of the parameter will be captured inside the interval.

The chosen degree of confidence is called the confidence level. The confidence level gives information about how much "confidence" we will have in the method used to construct the interval.

To create an interval of plausible values for a parameter, we need two components:

- A point estimate is a single value used to estimate the population parameter such as a sample population.
- A margin of error represents the maximum expected difference between the true population parameter and the sample estimate.

To calculate the interval you use the formula

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the point estimate, z^* is the critical value and $z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the margin of error.

To calculate the critical value, you can use invNorm. Generally for a 90% confidence interval, $z^* = 1.64$, for a 95% confidence interval, $z^* = 1.96$ and for a 99% confidence interval, $z^* = 2.58$.

6.2 Constructing a One Proportion z-Test

A significance test is another inference method that assesses evidence provided by data about a claim. Significance tests tell us if sample data gives us convincing evidence against a null hypothesis.

- A null hypothesis (H_0) is the claim being assess in a significance test. Usually, the null hypothesis is a statement of "no change from the expected value."
- An alternative hypothesis (H_A) proposes what we should conclude if we find the null hypothesis to be unlikely.

Hypotheses always refer to the population not the sample.

A p-value is the probability of getting results as extreme or more extreme in the direction of the null hypothesis by random chance alone assuming the claim of the null hypothesis is true.

- Small p-values give convincing evidence against the null hypothesis since the result we got is unlikely to occur.
- Large p-values fail to give convincing evidence against the null hypothesis since the result we got is likely to occur.

The significance level (α) is a fixed value that we will regard as the decisive value that determines if the p-value is small or large.

- Typically we choose $\alpha = 0.05$ which says we need data so strong that it would happen by chance less than 5% of the time.

To construct a test follow the steps:

- Define the parameter: $p =$ true proportion of {parameter in context}
- State the hypotheses. (If you are not given a claimed proportion, we use a conservative estimate which is 0.50)
- Check the Assumptions and Conditions
- Name the Inference method
- Calculate the test statistic
- Obtain the p-value
- Make a decision
- Write your conclusion in context

6.3 Relating Confidence Intervals and Significance Tests

Margin of error is point estimate $\pm$ margin of error.

Increasing confidence level increases critical value and margin of error and gives a wider interval.

Decreasing confidence level decreases critical value and margin of error and gives a narrower interval.

Increasing sample size decreases standard error and decreases margin of error and gives a narrower interval.

Decreasing sample size increased standard error, margin of error and gives a wider interval.

Keep in mind that the margin of error in a confidence covers only accounts for sampling variability and does not account for bias in the sampling methods.

6.4 Inference for Comparing Two Population Proportions

To construct a two proportion z-interval,

- Define the parameter. p_1 and p_2 are the true proportion of parameters in context for a population
- Check the assumptions and conditions (randomness, independence, and normality)
- Name the inference method
- Calculate the interval: $(\hat{p}_1 - \hat{p}_2 \pm z * \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_1(1-\hat{p}_2)}{n_2}})$
- Write the conclusion in context

To construct a two proportion z-test,

- Define the parameter
- State the hypothesis - The null is $H_0 : p_1 = p_2$ and the altternate is $p_1 < p_2, p_1 > p_2$ or $p_1 \neq p_2$
- Check the assumptions and conditions
- Name the Inference Method
- Calculate the Test Statistic (z-score). The null hypothesis states that there is no difference between the two population proportions. If this is true, the observations really come from a single population.

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

  where $\hat{p}_c = \frac{x_1+x_2}{n_1+n_2}$, where x is the number of successes and n is sample size.
- Obtain the p-value (using normalcdf)
- Make a decision, if the p-value is less than α you reject the null otherwise you fail to reject

- State the conclusion in context

6.5 Errors & Power

A Type I error occurs when

- Reject H_0 incorrectly
- The probability of a Type I error is equal to the significance level
- P(Type I error) $= \alpha$
- Our significance level tells us what p-value is "low enough"

A type II error occurs when

- Fail to reject H_0 incorrectly
- P(Type II Error) $= \beta$

Power

- Reject H_0 correctly
- The probability we reject the null correctly is 1 minus the probability we reject the null incorrectly
- Power $= 1 - \beta$

Errors and their relationships

- Type I and Type II errors have an indirect relationship: as the probability of one increases, the probability of the other decreases
- Our Type I error is set with our significance level
- The higher our significance level is, the lower our probability of failing to reject becomes, which is why α and β have an indirect relationship
- The higher our significance level, the more likely we will reject the null, which increases the likelihood we do that incorrectly (as well as correctly)
- Therefore, Type I error and Power have a direct relationship: as the probability of Type I Error increases, the higher the Power of the test

7 Inference for Quantitative Data: Means

7.1 Constructing a One Sample t-Interval

Sampling Distribution for Means:

- Shape - normal population indicates a normal sampling distribution. $n < 30$ then data needs to be stated as approximately normal. $n \geq 30$ satisfies Central Limit Theorem (CLT) which guarantees approximately normal
- Center - as long as you are taking a random sample or using random assignment, $\mu_{\overline{x}} = \mu$
- When sampling, as long as 10% condition is met, $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{x}}$

As long as the above conditions are met, the sampling distribution of $\overline{x} : \overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

When we do not know the population standard deivation (which we usually don't), we must estimate it from our sample using the sample standard deviation, s. However, when we do so, the test statistic (z) that we previously used changes. The new test statistic is now called the t-statistic and has a new distribution associated with it.

The new t-distribution is not exactly like the standard normal curve, but it is very close:

- It it still centered at 0
- It is bell shaped
- Its spread is slightly greater than the standard normal distribution
- It has more area in the tails

The special thing about t-distributions is the fact that it is a family of distributions. There is a unique density curve for each sample size (and the dependence on sample size is taken care of by degrees of freedom: $n-1$.)

The tails on a t-distribution are "fatter" than a standard normal distribution. This is true because the small our sample size, the more variation we have, hence the fatter tails. The larger our sample size gets, the closer the t-distribution will have towards the standard normal distribution.

Constructing a confidence interval

- Define the parameter. μ = true mean of {population parameter in context}
- Check the assumptions and conditions:
  - Random: The sample should be a random sample of the pouplation or random assignment in an experiment.
  - Independence (10% Condition): The sample size, n, must be no longer than 10% of the population.
  - Normality: THere are multiple ways to verify this condition.
    - Stated in problem: It may be stated in the problem that the sampling distribution is approx. normal
    - Central Limit Theorem: When n is large, the sampling distribution of the sample means is approx. normal.
    - Visual representation: You are given a graphical representation (histogram or boxplot) that depicts a shape that is approximately normal. You may also be given data that you have to graph yourself (include a sketch). You are looking for no strong skew or outliers.
- Name the inference method: One Sample t-Interval

- Calculate the interval: point estimate $\pm$ margin of error: $\overline{x} \pm t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right)$. Use 2nd-Vars-4:invT() in the calculator to calculate the critical value.
- Write your conclusion in context

You can determine the sample size from the formula $ME = t^*_{n-1}\left(\frac{s}{\sqrt{n}}\right)$.

7.2 Constructing a One Sample t-Test

Constructing a significance test:

1. Define the parameter: $\mu =$ true mean of {population parameter in context}.
2. State the hypothesis: the null and alternative.
3. Check the assumptions and conditions (same as last topic)
4. Method: One Sample t-Test
5. Test statistic is statistic-parameter divided by standard deviation of statistic and $t = \frac{\overline{x}-\mu_0}{\left(\frac{s}{\sqrt{n}}\right)}$.
6. Obtain the p-value from tcdf.
7. Make a decision.
8. State your conclusion on context.

7.3 Inference for Paired Data

Comparative studies are more convincing than single-sample investigations. For that reason, one-sample inference is less common than comparative inference. Study designs that involve making two observations on the same individual or one observation on each of two similar individuals, result in paired data.

When paired data result from measuring the same quantitative variable twice, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample t-procedures to perform inference about the mean difference: μ_D. These methods are called matched pairs procedures.

7.4 Inference for Comparing Two Sample Means

Same PANIC and PHANTOM as before.

Note that the degrees of free dom $n-1$ is where n is the smaller sample size (this is known as the conservative df).

The confidence interval formula is $(\overline{x}_1 - \overline{x}_2) \pm t^*_{n-1}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

The test statistic in a two sample t-test can be calculated by $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

8 Inference for Categorical Data: Chi-Square

8.1 Chi Square Test for Goodness of Fit

A goodness-of-fit test is used to test the hypothesis that an observed frequency distribution fits to some claimed distribution.

To measure the difference betwen the observed and expected counts, and to determine if the difference is significant, we will introduce a new test statistic, called the chi-square statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O represents each observed count in the distribution and E represents each corresponding expected cout.

- The sampling distribution of the chi-square statistic is not a normal distribution.
- It is a right-skewed distribution that allows only for positive values because the statistic cannot be negative.

When all the counts are at least 5, the sampling distribution of the χ^2 statistic is close to a chi-square distribution with degrees of freedom (df) equal to the number of categories minus 1.

- The chi-square distributions are a family of distributions that take only positive values and are skewed to the right.
- A particular chi-square distribution is specified by giving its degrees of freedom.
- The chi-square goodness-of-fit test uses the chi-square distribution with df = number of categories - 1

The null hypothesis in a chi-square goodness-of-fit test should stake a claim about the distribution of a single categorical variable in the population of interest.

The alternative hypothesis in a chi-square goodness-of-fit test is that the categorical variable does not have the specified distribution, and is easily given in words.

Conditions:

- Random - the data came from a well-designed random sample or randomized experiment
- Independent - When sampling without replacement, the 10% condition is met
- Large counts - All expected counts are at least 5 lets us say that the sampling distribution will follow a chi-squared distributions

When all conditions are met, the chi-squared goodness of fit test can be performed with the hypotheses:

- H_0: the claimed distribution is correct
- H_a: at least one proportion in the claimed distribution is incorrect

8.2 Chi Square Test for Homogeneity

The Chi Square Test for Homogeneity compares the distributions of one categorical variable across two or more populations to see if they are the same or different.

To construct a chi-square test for Homogeneity:

State the hypothesis:

- H_0: There is no difference in the distribution of the categorical variable among two or more groups
- H_A: There is a difference in the distribution of the categorical variable among two or more groups

Check the assumptions and conditions

- Randomness: The individuals whose counts are available for analysis should be a random sample of the population
- 10% condition: The sample size, n, must be no larger than 10% of the population
- Large Counts: We should expect to see at least 5 counts in each category of the categorical variable.

Name the inference method: Chi Square Test for Homogeneity

Calculate the Test Statistics:

- Observed Counts - Actual Frequencies of the variable from your sample
- Expected counts - projected frequencies of the variable if the null hypothesis is true. This is the row total times the column total divided by the table total.

Obtain the P-Value:

2nd-Vars-8:χ^2cdfwith lower being χ^2 and upper being ∞ and df is the number of rows - 1 times the number of columns - 1

In a calculator, enter the observed values into matrix A, and enter the expected values into matrix B, which will population when you run the χ^2 test. Then do stat-tests-c with observed in A and Expected in B.

Make a decision then write the conclusion in context.

8.3 Chi Square Test for Independence

9 Inference for Quantitative Data: Slopes