Magnet Statistics

Ace Chun

January 22, 2023

Contents

1 Unit 1

1.1 What is data?

Data is described as a collection of numbers, characters, or images that can describe some system. For data to be useful, we must be provided *context* about its surrounding circumstances, details that make the data meaningful for intended use. To analyze the context of data, we can ask the "Five W's": Who, What, When, Where, Why, and in addition, How. Specifically, emphasis is placed on "Who" and "What", as those are the properties that are essential to understanding any dataset.

Data is usually categorized into three different types: Identifiers, Categoricals, and Quantitatives.

Identifiers are not measures of anything in particular; rather, they are useful as distinct labels for different samples. They are not typically used to perform data analysis.

Categorical variables communicate what group or category a specific data point belongs to. With categorical variables, the labels are distinct, usually describing a type of something.

Quantitative variables are those that measure numerical values with units of measurement. More often, a set of data is considered to have quantitative variables if it makes sense to take an average of the data.

1.2 Categorical Data

There are three rules of data analysis:

1. Make a picture

2. Make a picture

3. Make a picture

The reason why this is so widely encouraged is because one cannot begin to analyze data without having some idea of what it looks like, or how it behaves.

Data with categorical variables can be sufficiently described by a frequency table, which lists each category and its corresponding cases.

A relative frequency table displays percentages or proportions of the data belonging to the category, rather than raw counts; in certain cases, this makes it easier for analysts to see and describe the tendencies that may be apparent in data.

These tables display the distribution of a categorical variable across its different labels.

A particular goal of categorical data analysis is to find correlations between datapoints, conditional of certain characteristics. Such an analysis can be conducted by finding a conditional distribution, which shows the distribution of one variable for a specific subgroup of individuals to examine if a certain characteristic influenced the outcome of another; this can be summarized in a contingency table.

If the distribution of a single variable is the same across all categories regardless of its circumstances, then it is independent of the category. Vice versa, if the variable changes according to the category, then it is likely dependent on the status of that category.

Data sets can sometimes display 'quirks', or random phenomenons that are not necessarily indicative of any specific effect. To address this, one may simulate a situation with more or random data; if a pattern observed in the original data is not present in the other data, then it is probable that the pattern was just a quirk of the data set.

1.2.1 Vocabulary

- Area principle — In a display of data, the representation of each data value should take up about the same amount of area.

- Frequency table — A table that lists the labels of a categorical variable as well as the number of observations for each category

- Relative frequency table — A table that lists the contents of each category as a relative percentage as compared to a whole category or dataset

- Distribution — The distribution of a variable tells us the possible values and of the variable, as well as the relative frequency of each value
- Contingency Table — Displays counts and/or percentages of data points that fall into a certain category, contingent on its label in another category
- Marginal distribution — Displayed within a contingency table, the distribution of one of the variables alone (or not contingent on the other variable)
- Conditional distribution — Distribution of a variable about data restricted to certain criteria
- Independence — Variables are independent when the conditional distribution for one variable is the same for each category of the other variable
- Simulation — Random reenactment of data to analyze the likelihood of association between variables
- Simpson's Paradox — When averages are taken across groups, said averages can appear to contradict overall averages.

1.3 Quantitative Data

Generally, quantitative variables are described in three ways: center, spread, and shape.

1.3.1 Center

Also known as a "point estimate", this approach attempts to describe the quantitative dataset in terms of a single number that represents the typical data. Examples of point estimates include the median, mean, and the mode (which is the least common).

1.3.2 Spread

The spread of a dataset describes how similar or varied a set of data is over a specific distance. The simplest measure of spread is the range of a dataset; that is, for a set of data X with n elements, its range is the value $X_n - X_1$. However, this approach can be flawed because of the nature of this calculation; by taking the difference between the two "atypical" datapoints in a set (that is, the two datapoints that would not describe the typical entry in the set of data), we gauge only an extreme idea of how spread the data is.

Another approach is standard deviation, which measures the average distance of a data point from the mean. Inherently, there's a problem with this simple statement: sticking with the dataset X, if we were to measure the average distance to the mean of a dataset, the formula would look something like:

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})$$

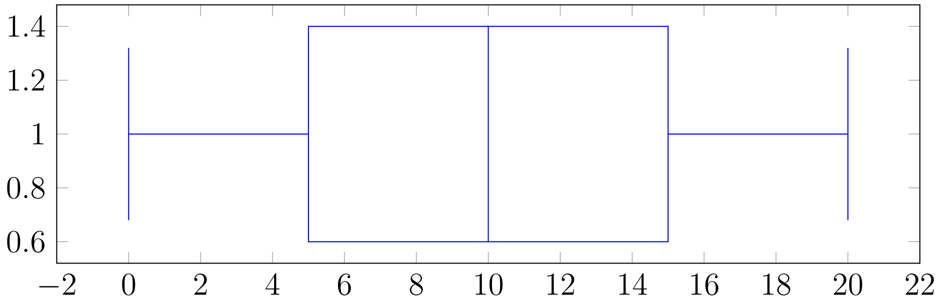
However, this formula is not very helpful; since $X_i - \overline{X}$ gives us *signed* distance from the mean, the result would simply be zero; this is, obviously, not the expected answer. Therefore, the distance should be squared in order to calculate a positive metric of distance without having to deal with the idea of the piecewise absolute value function. The modified formula looks something like

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

This formula is defined to be the variance of a dataset, denoted by S^2. The standard deviation, S, is defined to be the square root of the variance, or:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Yet another metric for describing the spread of a data set is the idea of the interquartile range; that is, the median of the upper half of the data set subtracted by the median of the lower half of the data set. Usually, the median of the upper half of a data set is denoted Q_3, or quartile three; vice versa, the median of the lower half of the data set is denoted Q_1. The Interquartile Range, or IQR, is then $Q_3 - Q_1$. The Box-and-Whiskers method of data visualization was developed by mathematician John Tukey, who wanted to draw attention to exactly the most typical, or the most useful half of a certain data set. Within a Box-and-Whiskers plot, the interquartile range is shown as a single box representing the middle half of the data. Lines are drawn to the lower and upper ends of a data set; this is because the data not encompassed inside of the box is still *important data*, and should be represented in an appropriate graph.



A problem with some Box-and-Whiskers plots is with the inclusion of outliers; if extreme outliers are to be counted, then the whiskers would simply be too distracting to fulfill their original purpose; thus, Tukey was forced to create a "definition" of sorts for outliers. This is known as Tukey's Outlier Algorithm; we set the 'fences' of the data at $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$. However, Tukey maintained that the status of outliers of the data set were heavily dependent on context, and that one could not simply make a hard and fast rule to calculate these outliers.

1.3.3 Shape

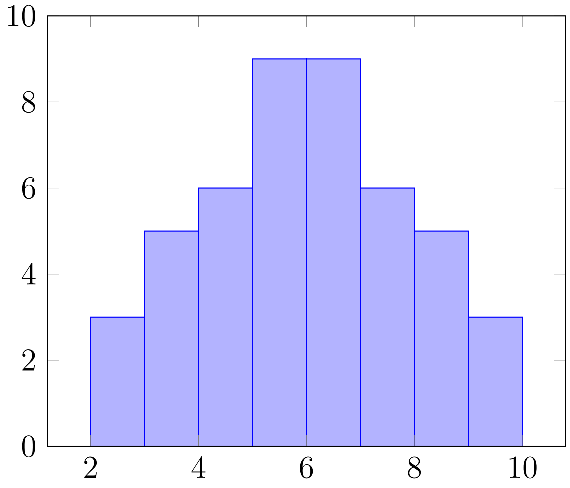
The third method of describing a quantitative data set is the shape of a data set. There are three typical shapes that data sets align into:

1. Symmetrical: the spread on either side of the median is around the same.

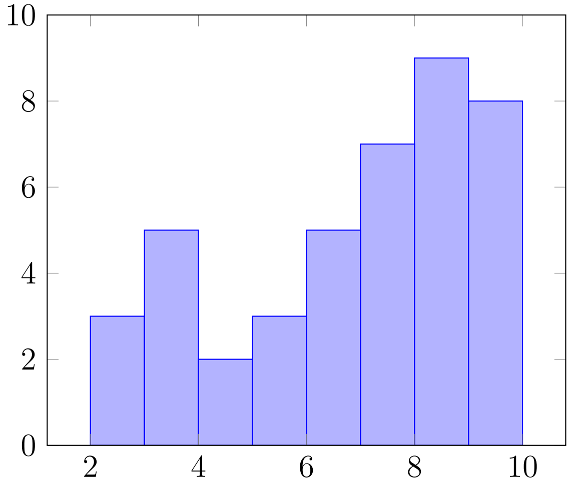
2. Skewed left: the spread on the left side of the median is considerably greater than the spread of the data on the right side.

3. Skewed right: the spread on the right side of the median is greater than the spread on the left side.

A symmetrical data set looks "mound-shaped"; its median and mean are somewhere around the same value.



A data set that is skewed left has a high concentration of its data points on the right side, therefore leaving the left side with more spread:



Vice versa, a data set that is skewed right has a high concentration of data points on the left side of the data.

Usually, a symmetric data set is considered “well-behaved”; the median and mean are mostly similar, so it is valid to use both metrics in describing the data. However, with heavily skewed, “naughty” data sets, the median and mean will be largely different; it is cases like these that render the mean unhelpful. Well-behaved datasets are considred to be unimodal, symmetric, and have a limited number of outliers.

1.3.4 Resistance

There are two kinds of people in life: those who stay with you in times of crisis, but are hard to be around, and those who are easy to work with, but are nowhere to be found when crises (outliers) show up. Statistics is much the same. Statistics act much the same. There are resistant statistics, which resist changing in the presence of outliers, and non-resistant statistics, which change drastically. For example, the mean and the IQR are considered resistant; the average and standard deviation are not.

1.3.5 Five Number Summary

Typically, data is summarized in a five-number summary, which contains

1. Smallest data value
2. First quartile
3. Median
4. Third quartile
5. Largest data value

1.4 Classifying points

1.4.1 Z-scores

Suppose we have a scenario with two tests: if I get a score of 83 on test 1, and a score of 88 on test 2, is it fair to say that I did better on test 2 than test 1? It isn't, because we should look at the overall picture of the data; the tests may have been different in difficulty. If we're provided the information that, for test 1, $\bar{x} = 78$ and for test 2, $\bar{x} = 84$, is it still fair to say that my second score was better? If we have the data $S = 1.6$ for test 1, and $S = 1.2$ for test 2, we can figure out how many standard deviations our score is from the mean. For test 1, this works out to

$$\frac{83 - 78}{1.6} = 3.13$$

and for test 2, this works out to

$$\frac{88 - 84}{1.2} = 3.33$$

This process can be described as "using the standard deviation as a ruler." This measure is called the z-score:

$$z = \frac{x - \bar{x}}{S_x}$$

This calculation can be done on every data set; however, on heavily skewed data sets, the z-score will not mean what is expected from a symmetric data set.

1.4.2 Density Curve

The density curve is a continuous function with two properties:

1. $f(x) \geq 0$
2. $\int_a^b f(x)dx = 1$

Density curves are drawn when we are imagining a distribution. While a density curve is not ostensibly a histogram, we can treat it as a very "dense" histogram.

With the property that the area under the curve is 1, we can use the bounds to analyze the percentiles of a specific data point.

When we talk about actual data sets, we use alphabetical letters. For example, to indicate the mean of an actual data set, we use $\bar{X}$; however, with theoretical distributions from a function, we denote it as μ. Similarly, with the standard deviation; in an actual data set, it is denoted S; in a theoretical distribution, it is denoted σ.

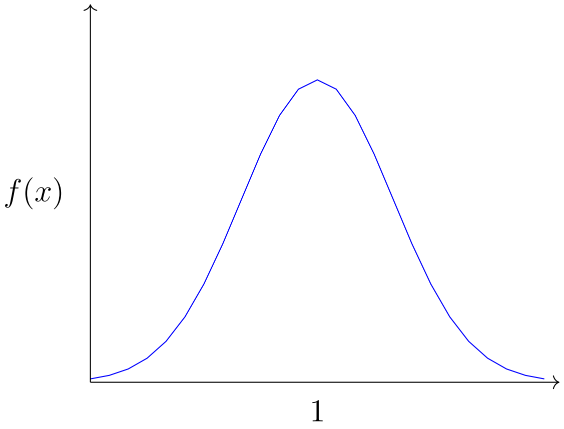
The z-score of such a distribution follows the same rules:

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

1.4.3 Normal Distribution

A special distribution function, often called the Gaussian distribution, or the Normal curve, is expressed as

$$f(x) = e^{\frac{-x^2}{2}}$$



The area of this curve as it is right now is not 1. However, we can take its integral and divide the function as a whole by the area. We have the integral

$$\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx$$

This happens to evaluate to an exact value of $\sqrt{2\pi}$

Thus, our function becomes

$$\frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

This final curve is called the *standard normal distribution*. The distribution has several properties:

1. It is perfectly symmetric.
2. $\mu_x = 0$
3. σ_x is found at the point of inflection: $x = 1$

There is a special notation in terms of an ordered pair; for the curve above, it is denoted $N(0, 1)$, where the first number in the 2-tuple is the mean and the second number is the standard deviation. In a normal distribution, 68% of the distribution is within 1σ of μ. This is found with the integral

$$\int_{-1}^{1} f(x)dx = 0.682$$

Similarly, 95.4% of the distribution is within 2σ. 99.7% of the distribution is within 3σ. These three ideas are called the Empirical rule; they are reference points.

We can shift the distribution in some way.

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

As an example, let's take the distribution of SAT math scores, described by $N(510, 110)$. The percentile of a data point is the percent of data points that did worse than that specific data point. Let's say that one person got a 730 on the SAT math. A score of 730 has a z-score of 2; therefore, it has a percentile of 97.5. A score of 400, meanwhile, is in the 16th percentile.

As another example, we can look at the dewlaps of an iguana. Its distribution is defined by $N(8, 1.3)$, and we want to find $P(x > 6)$. To find this percentile, we could alter the standard normal distribution; however, the better way to do this is to find the z-score of the data point for its specific distribution, and find its corresponding integral on the standard function.

In our example,

$$z = \frac{6-8}{1.3} = -1.54$$

$$\int_{-1.54}^{\infty} e^{\frac{-x^2}{2}} dx = 0.938$$

which indicates that 93.8% of the data fits the restraint of $x > 6$.

Ultimately, if we can convert to a z-score, then it is reducible to the same thing. There exists a table with corresponding z-scores and indexes. Note: this table is affectionately dubbed "Cha-A-a-A-rt" by Mr. Stein.

If we have $z = 1.645$, we can find our original x value given the formula for a z-score:

$$1.645 = \frac{x-8}{1.3}$$

This process is known as an inverse-normal.

Datasets cannot be normal. "Normal" describes a distribution, and so datasets cannot strictly be normal. We can, however, "assess normality"; it asks, is it reasonable to believe that the data comes from a normal distribution?

1.4.4 Normal Probability Plots

Suppose we have a dataset where we are told that certain data points fulfill a certain percentile marking. We can find the z-score of each percentile, and then plot this onto a data by z-score plot: (data, z-score). If the distribution were normal, then the plot would be linear; this is because the z-score is a linear transformation, and thus, if each percentile fit exactly the corresponding z-score of a normal distribution, the would follow a general line.

2 Unit 2

2.1 Two variables

With two quantitative variables, the relation between them is specifically called *correlation.* With any other type of variable, it is simply a *relationship.* In a graph where y is dependent on the x, x is known as the explanatory variable, while y is known as the responsive variable.

When we talk about bivariate data, we consider

1. Direction (as x changes, which direction does y move?)
2. Strength (how is the data spread around the center model?)
3. Linearity (what is the shape of this model?)

When we discuss direction, we are talking about the center model; this may be a line fit to the data, or another curve. Strength is a measure of spread with respect to the center model. Linearity discusses "how straight" a model is; if you can draw a cloud around data, linearity can be determined by gaps within the cloud; if there are large gaps, then it is likely not linear.

These three characteristics are analogous to center, spread, and shape.

A metric for measuring strength is the correlation coefficient, r.

2.1.1 Correlation Coefficient (r)

When we compare data, especially on a scatterplot, it is not accurate to plot the raw data; a difference in units could mess up any sort of analysis. Therefore, we must *standardize* the plane; we convert each of the axes to measures of z-scores and then plot it on a new graph, known as the standardized plane.

The data will be mostly centered around the origin, with data divided into the four quadrants; if we assume linearity, then it is apparent that if the explanatory data has a positive correlation with the responsive data, then the data would lie in quadrants I and III, meaning that both data for a given point are either both positive or both negative. Similarly, for a negative correlation, then the data from one of the axes would be negative, while the other would be positive. To determine the direction of the data, we can then multiply together the z-scores of the x and y data for a given point; if there are more positive numbers than negative numbers, the data is more likely positively correlated. Taking a sum of these products gives us an aggregate number that tells us something about the direction of the data. However, it is not necessarily a fair metric; the more data points are present, the larger this sum will be. Thus, we can divide this sum by $n-1$, which gives us a sort of "average"; this is the r number.

Formulaically, this process is summarized as

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\cdot\left(\frac{y_i - \bar{y}}{S_y}\right)$$

This r number has several properties.

1. r is not resistant; it is based on the mean and standard deviation, which are not resistant statistics.

2. r is unitless.

3. r does not measure linearity; in fact, it assumes linearity and measures the strength around an assumed linear model.

4. $r = \{1, -1\}$ indicates that the data fall in a perfect line.

5. $-1 \leq r \leq 1$

2.1.2 Linear Models

For data that is related linearly, there exists some line

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

where the hat indicates that the function is a prediction of the data. This is just a simple linear model; it is just a simplified version of the equation

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots$$

In this sense, it doesn't make sense to add the constant a after the bx term.

The residual is the difference between y (the actual value) and $\hat{y}$ (the prediction).

$$y - \hat{y}$$

We would like to find the "best" line of fit for a set of data; since "best" is a subjective term, it is defined strictly as the line that minimizes the expression

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

which is known as the Mean Squared Residual (MSR). The line that fulfill this criteria is known as the Least Square Regression Line (LSRL). Notice that this expression looks remarkably close to the equation for variance; in fact, if we substitute $\hat{y}$ for $\overline{y}$, it becomes the equation for variance.

Remarks about z-scores

$$\begin{align} \overline{z} &= 0 \tag{1}\\ \sum z &= 0 \tag{2}\\ S_z &= 1 \tag{3}\\ S_z^2 &= 1 \tag{4}\\ \frac{1}{n-1}\sum(z-\overline{z})^2 &= 1 \text{ (refer to line (1))} \tag{5}\\ \therefore \frac{1}{n-1}\sum z^2 &= 1 \tag{6}\\ \therefore \sum z^2 &= n-1 \tag{7} \end{align}$$

We can plot our data in the standardized plane.

$$\hat{Z}_y = a + bZ_x$$

We want to find $a,\ b$ such that

$$\frac{1}{n-1}\sum(Z_y - \hat{Z}_y)^2$$

is minimized. By revisiting our original model, our expression becomes

$$\frac{1}{n-1}\sum(Z_y-(a+bZ_x))^2$$

$$\frac{1}{n-1}\sum((Z_y-bZ_x)-a)^2$$

$$\frac{1}{n-1}\sum((Z_y-bZ_x)^2-2a(Z_y-bZ_x)+a^2)$$

$$\frac{1}{n-1}\sum(Z_y-bZ_x)^2-\frac{2a}{n-1}\sum Z_y+\frac{2ab}{n-1}\sum Z_x+\frac{1}{n-1}\sum a^2$$

It's apparent that the terms $\frac{2a}{n-1}\sum Z_y$ and $\frac{2ab}{n-1}\sum Z_x$ are zero, since they are sums of z-scores. The expression then becomes

$$\frac{1}{n-1}\sum(Z_y-bZ_x)^2+\frac{1}{n-1}\sum a^2$$

To then minimize this expression, a must equal 0; if it is anything other than 0, then the line is not minimized. The line must, therefore, past through the grand mean, or centroid: the point $(\bar{x},\bar{y})$ We now know that our simple model is much simpler:

$$\hat{Z}_y=bZ_x$$

and now we must minimize

$$\frac{1}{n-1}\sum(Z_y-bZ_x)^2$$

$$\frac{1}{n-1}\sum(Z_y^2-2bZ_xZ_y+b^2Z_x^2)$$

$$\frac{1}{n-1}\sum(Z_y^2)-\frac{2b}{n-1}\sum Z_xZ_y+\frac{b^2}{n-1}\sum b^2Z_x^2$$

Refer back to line (7) of the remarks about z-scores ($\sum z^2=n-1$) and the definition of r ($\frac{1}{n-1}\sum Z_xZ_y$)

$$\frac{n-1}{n-1}-r+\frac{b^2(n-1)}{n-1}$$

$$b^2-2rb+1$$

This is the MSR. We want to minimize this polynomial; we can use the *Not the First Derivative Test*:

$$\frac{d}{db}(b^2 - 2rb + 1) = 2b - 2r = 0$$

$$\therefore b = r$$

This is an incredible result; we have found the solution to the prediction problem; the offset a is equal to 0, and the slope b is equal to r. The correlation problem is, in a way, equivalent to the correlation problem. We have just found a line for data in the standardized $(z - z)$ plane:

$$\hat{Z}_y = rZ_x$$

Recall that the units of z-scores are standard deviations. Thus,

$$b = r\frac{S_y}{S_x}$$

This equation is known as the regression formula. If we know the spread of x and y, and if we know the correlation between the variables, then we have the exact slope of a prediction line for those variables.

However, in the non-standardized plane, the y-intercept is not necessarily 0; however, we can set up an equation to calculate the y-intercept of the line in data space:

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

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

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

In the standardized (z-z) plane, the equation of the minimum line is

$$\hat{Z}_y = bZ_x$$

We also know that the MSR of the line is

$$1 - 2br + b^2$$

minimized when $b = r$. Plugging this into the polynomial, we get that the minimum MSR is $1 - r^2$; this proves, mathematically, that $-1 \leq r \leq 1$. In addition, this means that when $r = \pm 1$, we have a perfect line; this means that the MSR is zero.

Variance is the MSR off of a very basic prediction line: $\hat{y} = \bar{y}$. Using the formula for MSR

$$\frac{1}{n-1}\sum(y-\hat{y})^2$$

and plugging in $\bar{y}$, we get the exact formula for S^2. We can compare this formula to our LSRL line

$$\frac{1}{n-1}\sum(y-\hat{y})$$

After converting both plots to z-scores, notice that S_y^2 in terms of z-scores is simply 1; the MSR of the LSRL is $1 - r^2$ in terms of z-scores. Therefore, the difference between these two quantities is r^2. r^2 tells us how much "better" the resulting line is compared to the most basic model of $\hat{y} = \bar{y}$, which is why many packages provide information of r^2 moreso than r itself.

"r^2% of the variation in y is explained by the model based on x".

r^2 is the percentage of the variation that we are able to explain with our model.

2.2 Residuals

Residuals describe the aspect of data that has not been described by a model. It is defined as the difference between the observed value and the predicted value, or

$$e = y - \hat{y}$$

Intuitively, the emphasis on this makes sense; in order to understand how well a model does, one must first understand the aspects in which the model fails.

After fitting a regression model, we can plot the residuals onto a separate graph (this can be plotted against x or $\hat{y}$, but there is not much of a distinction between the two). The residual plot should not have any direction or shape; if it does, then the regression model is likely incorrect.

To quantify the typical size of a residual, we can refer to the concept of the standard deviation, which tell us how much a set of data typically differs from the mean. In a similar fashion, the formula for standard error is

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

The standard error simply summarizes the typical residual, analogous to how the standard deviation does so for data of a single variable.

The worst pattern to see on a residual plot is fanning, or heteroscedasticity, where the residuals progressively “fan out” from the center; this means that our model is doing even worse for larger numbers.

2.3 r^2

It has been proven, mathematically, what the significance of r^2 is; it is the difference between the variance of the y-values and the mean squared residual of the optimized LSRL line. Recall the equation

$$e = y - \hat{y}$$

This can be rearranged into

$$y = \hat{y} + e$$

We can, then, apply a function of *variance* onto the components of the equation:

$$var(y) = var(\hat{y}) + var(e)$$

Since r^2 attempts to describe the variation accounted for in a model, we can see that

$$r^2 = 1 - \frac{var(\hat{y})}{var(y)}$$

Because variances are positive, it is apparent that $0 \leq r^2 \leq 1$.

Essentially, r^2 quantifies how well a linear model fits a set of data by expressing the fraction of overall variability in the response variable against the explanatory variable. Because of its fractional nature, r^2 is denoted as a percentage.

2.4 Re-expressing Data

When we face data that is not quite ‘straight enough’, we can apply some sort of function to the data to view it in a different frame of reference. If this re-expression produces some linear array of data with a fitted regression line that has an appropriate residual plot, then we can find a fitted function to the data.

We have 4 goals in data re-expression:

1. Make the distribution of a variable more symmetric

2. Make the spread of several groups more alike, even if the centers are different

3. Make the form of the scatterplot more linear (and its residual plot uninteresting)

4. Make the scatterplot spread out evenly

To help us with this, we can use the Ladder of Powers.

Power	Name	Comment
2	Square of data values	This may be useful for unimodal distributions skewed to the left.
1	Raw data, no change	We make no changes in this case.
$\frac{1}{2}$	Square root of data values	Counted data can benefit from this re-expression.
0	Logarithm of data values	Can be helpful with values that cannot be negative, especially those that grow by percentage increases
$-\frac{1}{2}$	Negative reciprocal square root	Taking the negative of this re-expression can preserve direction.
-1	Negative Reciprocal	Ratios of two quantities can benefit from this.

The re-expression becomes stronger as we descend the ladder.

2.5 Summary

Correlation: Denoted r, correlation assumes linearity between two variables and measures the strength of the association between them.

$$r = \frac{1}{N-1}\sum_{i=1}^{N} Z_{xi}Z_{yi}$$

Prediction line: A prediction line for data exists in the form $\hat{y} = a + bx$, where $b = r \cdot \frac{S_y}{S_x}$. The prediction line must pass through the grand mean

$((\bar{x}, \bar{y}))$, so $a = \bar{y} - b \cdot \bar{x}$, Therefore, the equation for a prediction line is

$$\hat{y} = \left(\bar{y} - r\frac{S_y}{S_x}\bar{x}\right) + r\frac{S_y}{S_x}x$$

R-squared: R^2 describes the percentage in the variation of our y variable that can be explained by the model based on our x variable. R^2 is the square of the correlation coefficient r, and is also equivalent to

$$1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

Residuals: The residual plot is x graphed against $y - \hat{y}$. In order for a model to be adequate, the residual plot should not display any patterns or direction; it should look like completely random data.
Re-expression: If data does not seem to be linear, then we can re-express each axis in terms of a different function (for example, $x \to \ln x$ or $y \to y^2$). If the data seems linear enough after re-expression, we can find the prediction line and rearrange the equation for $\hat{y}$. The residual plots should indicate an adequate model.

3 Unit 3

3.1 Understanding "Randomness"

Randomness can be thought of as fulfilling the criteria:

1. Unpredictable in the short run
2. Predictable in the long run

Random numbers and data can be used for simulation, in order to generate an accurate picture of the behavior of some data. For example, we may simulate data with random numbers to detect 'accidental correlation' and determine if a relation actually exists within two arrays.

A *simulation* simply mimics reality through random numbers to represent outcomes of events. Each time we obtain a simulated answer, we have performed a *trial.* A *simulation component* is a basic building block of a simulation, used to model random occurrences. The outcome of a trial is a *response variable.*

To formally simulate trials, we can:

1. Explain how components combine to model a trial.
2. State the response variable and its value in context.
3. Run several trials.
4. Collect and summarize results of all trials.
5. Find and state the conclusion of the simulation.

The intent of a simulation is gain insight about situations not already understood. The conclusion of a simulation depends on the number of trials performed — too few trials will result in inaccurate conclusions, while too many may be resource intensive.

3.2 Survey Design

- Population — Group of individuals of interest
  - Usually, an entire population will be too large to measure wholly.
  - A census polls the entire population, however, this often takes too many resources.
- Sample — Group of individuals who are polled
  - We must be careful with sampling, however, because results may vary across different samples.
- Sampling Frame — Group of individuals who have any chance of being chosen
  - Ideally, the sampling frame would be the same as the population.
  - If the frame is smaller than the entire population, then we would not have a fair sample to work with.
- Parameters — Numbers (measurements) about entire populations
- Statistics — Numbers (measurements) about samples
- Probability Sample — Sample which uses a randomization procedure to choose members of the sample

  - While a random sample may not be more representative of the population, to the extent that it is not, we can quantify the likelihood of its statistics from the corresponding parameters.
  - We can figure out probabilities with Probability Samples, unlike convenience samples.

- Convenience Sample — Sample which uses no randomization techniques

3.3 Types of Random Samples

There are multiple types of random samples.

- Simple Random Sample (SRS) — A sample take of size n in which every group of size n from the population of interest has the same chance of being chosen

- Stratified Random Sample — A population is divided into relevant (not arbitrary) subgroups and a SRS is taken within each subgroup.

- Cluster Sampling — A population is split into representative clusters that make measurements more practical; one or two clusters are chosen at random and censuses are performed within them.

- Systematic Sampling — Individuals from a population are chosen systematically (i.e., every tenth person from a list is chosen)

Sampling schemes that combine these types of random samples are known as multistage samples. For any given question of interest, there are multiple statistics but there is a single parameter. Differences between statistics due to randomness are known as *sampling variability.*

Statistics are likely not equal to the parameter. This can be chalked up to two possible reasons:

- Luck, or sampling variability

- Bias

Our goal is to minimize the bias and quantify the variability.

There are two different types of bias.

- Sampling Bias — Bias in the way samples are collected in a survey
  - Undercoverage — size of sampling frame is smaller than population
  - Nonresponse — randomly selected participants refuse to participate
  - Voluntary Response — population members opt into the survey
  - Convenience Bias — no probability sample at all
- Non Sampling Bias — Question Bias, Response Bias

3.4 Experimental Design

In an experiment, a researcher imposes the treatment. In a study, the relationship between variables is simply observed.

An (observational) study simply records subjects "in the wild" and doesn't interfere in any way with the choices made by the subjects. A retrospective study identifies subjects and then collects data on their past records and experiences. Conversely, a prospective study identifies subjects in advance and collects data as events happen.

There are three principles of experimental design:

1. Comparison
2. Random assignment of treatments
3. Replication

An experiment that fulfills these three principles is known as a Completely Randomized Experiment (CRE). Statistical Significance describes a difference too big to be attributed to chance.

A Blocked Randomized Experiment, meanwhile, divides the volunteer pool into different blocks based on some criteria. A CRE is conducted within each block.

Experiments study the relationship between at least two variables. There must be at least one explanatory variable, known as a factor, that is manipulated in the experiment, as well as one response variable in which changes are measured with respect to the factor.

Individuals on whom an experiment is conducted are generally called experimental units. The different values that an experimenter chooses for a factor are known as levels. The combination of specific levels that a unit receives is known as a treatment.

There are four guidelines for an effective experiment.

1. Control — Sources of variation between treatment groups are controlled by making testing conditions as similar as possible for all groups.

2. Randomize — Randomization allows the experimenters to equalize the effects of unknown or uncontrollable sources of variation. Assigning different units treatments at random allows for the use of statistical methods to draw conclusions.

3. Replicate — Meaningful conclusions can only be drawn if the experiment is repeatable and its results are replicable. The outcome of an experiment on a single subject is an anecdote, not complete data.

4. Block — Attributes of experimental units that are not being observed may affect the results of an experiment. Similar individuals are, then, grouped together and assigned randomized treatment within blocks. It's a compromise between complete randomness and control.

What is the purpose of blocking? It is not to reduce bias; bias exists in sampling, and this is not a sampling question. Blocking is done to reduce variation.

An experiment is considered blind if its subjects don't know which group they are in. "Double Blindness" means that both the researchers and the subjects do not know the groups.

The best experiments are usually randomized, comparative, double-blind, and placebo-controlled.

When levels of one factor are associated with the levels of another factor, the two factors are confounded.

4 Unit 4

4.1 Probability

Suppose we have an urn that contains some number of marbles; we have two blue, one green, and two red marbles. Intuitively, the probability of selecting a red marble is $\frac{2}{5}$. What does this mean, exactly?

We have two views of probability. Frequentism describes probability in terms of the number of successes in the long run; as the number of tries goes to infinity, the ratio of the number of successes to total outcomes is our probability.

The Bayesian view is entirely different. Before anything happens, there exists a prior probability of success. Then, evidence is "inserted", and the end result is a posterior probability.

If we ask the question "What is the probability that the coin is fair?" To the frequentist, there is no direct answer to this question. To the Bayesian, the prior probability that the coin is fair starts out very high (for example, 99%). Then, they flip the coin many many times; based on the evidence, the probability that the coin is either fair or not fair is either increased or decreased. To the Bayesian, every piece of new information somehow changes the probability in a way.

A principle known as the Law of Large Numbers (LLN) says that a random process repeated over and over will eventually settle into a single number; a probability. However, the LLN assumes that the phenomenon do not change with respect to probability, and that events are independent.

4.2 Some Notation

- $P(A \cup B)$ — Union, or Probability of A or B
- $P(A \cap B)$ — Intersection, or Probability of A and B
- $P(A|B)$ — Probability of A given B
- $P(A^c)$ — Complement of the probability of A happening (or $1 - P(A)$)

A and B are disjoint or mutually exclusive events if $P(A \cap B) = 0$ or $P(A|B) = 0$ (or $P(B|A) = 0$). A and B are independent events if $P(B|A) = P(B)$, or vice versa.

We have some rules.

Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Multiplication Rule

$$P(A \cap B) = P(A) \cdot P(B|A)$$

Let's suppose that two events, A and B, are disjoint. Are they independent?

We can answer this question in two ways: intuitively and mathematically. Intuitively, they cannot be independent because the two events that are disjoint still affect each other; they just have an opposite relation. Thus, they cannot be independent because they are not completely separate events.

$$P(A \cap B) = 0$$
$$\frac{P(A \cap B)}{P(A)} = P(B|A) = 0$$
$$P(B|A) \neq P(B)$$

Thus, they are not independent.

4.3 Random Variables

A variable is, like the name suggests, something that varies. For example, in the equation

$$3x + 5 = 20$$

there are zero variables. x is something that has a fixed variable; it may be unknown, but it is not a variable. On the other hand, the equation

$$y = 2x + 3$$

has two variables; there are two things that vary in the equation.

A random variable is a variable whose outcomes are determined randomly.

There are two kinds of random variables.

1. Discrete Random Variable: A random variable whose set of outcomes has either a finite or an infinite but countable number of elements

2. Continuous Random Variable: A random variable whose set of outcomes has a non-countable and infinite set of outcomes

Some continuous random variable problems may seem familiar. For example,

$$\text{Find } P(x \geq 9) \text{ if } \mu_x = 6,\ \sigma_x = 8$$

This is simply a normal distribution problem; we can calculate the z-score and then reference the normal probability chart to find the answer.

For discrete random variable problems, we can start with writing out the possibilities. For example,

X	-1	0	3	19	99
P(X)	0.5	0.4	0.09	0.0099	0.0001

We find the center of our discrete distribution by multiplying by the weights and dividing by the sum of the weights, which will always be one. Therefore,

$$\mu = \sum x \cdot P(x)$$

This number, μ, is the expected value. For the example above, the expected value is -0.032.

We can extend this definition to the variance.

$$\sigma^2 = \sum (x - \mu)^2 \cdot P(x)$$

Similarly, with standard deviation,

$$\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)}$$

There are some transformation rules.

$$\text{If } T = X + a,\ \mu_T = \mu_X + a,\ \sigma_T^2 = \sigma_X^2,\ \sigma_T = \sigma_X$$

$$\text{If } S = bX,\ \mu_S = b\mu_X,\ \sigma_S^2 = b^2\sigma_X^2,\ \sigma_S = |b|\sigma_X$$

The second most important theorem (the first being the correlation equation) discusses the addition or subtraction of two random, independent variables.

$$\mu_{X \pm Y} = \mu_X \pm \mu_Y$$

The pythagorean theorem of statistics says (if X and Y are independent):

$$\sigma_{X \pm Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Therefore,

$$\sigma_{X \pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$$

4.4 Binomial Random Variables

A **Bernoulli Event** is a random variable with exactly two outcomes: success and failure. 'Success' is simply what we choose to count, while 'Failure' is everything else.

A **Binomial Random Variable** is a random variable which counts the number of successes in n identical and independent Bernoulli events. We have a couple of requirements:

1. Exactly two outcomes (success or failure)
2. Fixed number of trials (n)
3. Probability of success (p) and probability of failure ($q = 1 - p$) must be the same for each trial
4. Each trial must be independent

In order to calculate probability density of specific events, we have the Binomial Probability Density Function (also called the binompdf).

$$P(X = x) = \frac{n!}{x!(n-x)!}p^x q^{n-x}$$

where X is our Binomial Random Variable, x is our target number of successes, n is our number of trials, p is our probability of success, and q is our probability of failure.

If X is binomial, then

$$\mu_X = p \cdot n$$

and

$$\sigma_X^2 = \mu_X \cdot (1 - p) = n \cdot p \cdot q$$

$$\sigma_X = \sqrt{npq}$$

These are incredibly important formulas.

4.4.1 Proofs of formulas

Begin with a simple Bernoulli event.

X	1	0
$P(X)$	p	q

$$\mu_X = 1 \cdot p + 0 \cdot q = p$$

$$\begin{aligned}
\sigma_X^2 &= (1-\mu_X)^2 \cdot p + (0-\mu_X)^2 \cdot q \\
&= (1-p)^2 \cdot q + p^2 \cdot q \\
&= q^2 p + p^2 q \\
&= pq(p+q) \\
&= pq
\end{aligned}$$

So we know that the mean of a Bernoulli event is p, and the variance is pq. A Binomial Random Event is simply n independent and identical Bernoulli events. We know that the means of events simply add. Therefore, we have n events with a mean of p; the mean of the binomial must be np.

We can add variances iff the events are independent. The binomial is made of n independent events; therefore, we may just add them up, leading to a variance of npq.

4.4.2 Rules of Thumb

"10% condition"

If the sample size is less than 10% of the population size, we can ignore the effect of non-replacement on the independence assumption.

"Success/Failure" condition:

If $np > 10$ and $nq > 10$, then you can safely approximate a binomial with a normal distribution.

4.5 Geometric Random Variables

Geometric Random Variables count until some success event happens. They are also called "wait-time" events. If we have some event with with a success probability of p, then our distribution would look something like

X	1	2	3	$\cdots$	∞
$P(X)$	p	qp	q^2p	$\cdots$	$\frac{p}{1-q}$

Note that

$$\frac{p}{1-q} = \frac{p}{p} = 1$$

And so, with a GRV,

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

The mean of a geometric variable is

$$\mu_X = \frac{1}{p}$$

$$\sigma^2 = \frac{1-p}{p^2}$$

$$\sigma = \sqrt{\frac{1-p}{p^2}}$$

4.6 Review

- Chapter 13
    - And/Or/Given
    - Multiplication Rule
    - Addition Rule
    - Independent vs. Disjoint
- Chapter 14
    - Conditional Probability
    - Reversing the Condition (tree diagram)
- Chapter 15
    - Random Variables
    - μ, σ^2, σ for a discrete random variable
    - Transforming random variables
    - Pythagorean Theorem (if x and y are independent)
    - If X and Y are not independent, then:
        - If X and Y are positively correlated, then the STD goes up

        - ∗ If they are negatively correlated, the STD goes down

- Chapter 16
    - Binomial RV
    - Calculate binompdf, binomcdf
    - Mean - np, variance - npq
    - When to use a normal instead of a binomial
    - Geometric RV

5 Unit 5

5.1 Sampling Distributions

Random sampling of a population yields different samples each time; they are naturally vulnerable to sampling variability. It is helpful, then, to quantify how much sampling variability we should expect between different polls. By aggregating the results of different random samples taken of a population, we can draw a distribution.

A samping distribution is the set of a particular statistic repeatedly taken from all possible samples of size n from a certain population. There are two different types of sampling distributions: sampling distributions of sample proportions, and sampling distributions of sample means.

The difference between a sample proportion and the population proportion is known as sampling error.

The main point here is that the statistics (not the parameters) of some sample are, in themselves, random variables.

5.1.1 Notation

For sample proportions,

- $\hat{p}$ — Sample proportion
- p — Population proportion
- $\mu_{\hat{p}}$ — Mean of the sampling distribution (average of $\hat{p}$s)
- $\sigma_{\hat{p}}$ — Standard deviation of the sampling distribution (spread of $\hat{p}$s)

For sample means,

- $\overline{x}$ — Sample mean
- s — Sample standard deviation
- μ — Population mean
- σ — Population standard deviation
- $\mu_{\overline{x}}$ — Mean of the sampling distribution (average of $\overline{x}$s)
- $\sigma_{\overline{x}}$ — Standard deviation of the sampling distribution (spread of $\overline{x}$s)

5.1.2 Formulas

	Proportions	Conditions	Means
Center	$\mu_{\hat{p}} = p$	if sample is taken without bias	$\mu_{\overline{x}} = \mu$
Spread	$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$	if picks are independent	$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Shape	Approximately normal	if n is big enough
		if $np > 10$, $nq > 10$
			Approximately normal

Always check the conditions for center, spread, and shape.

The Central Limit Theorem is the fundamental theorem of statistics. It states that the shape of a sampling distribution $\bar{x}$ will be approximately normal if n is sufficiently large. This means that, no matter how messed up the starting distribution is, the distribution of sample means will always tend towards normality, given that there are enough samples taken.

5.2 Confidence Intervals

Recall that sampling distributions of population proportions can be approximated with a normal distribution. A critical value, denoted

$$Z_x^*$$

denotes the number of standard deviations away from the mean with which x% of the values lie. The value of Z_x^* can be found with a normal probability table. In a sampling distribution of $\hat{p}$s, x% of $\hat{p}$s are within Z_x^* $\sigma_{\hat{p}}$ of $\mu_{\hat{p}}$.

If I were to sample repeatedly, some $x\%$ of the samples would produce a $\hat{p}$ such that p is within $Z_x^* \cdot \sqrt{\frac{pq}{n}}$ of the particular $\hat{p}$.

This is called a $x\%$ confidence interval. For example, a 95% confidence interval would mean that p is within

$$\left[\hat{p} - 1.96 \cdot \sqrt{\frac{pq}{n}}, \hat{p} + 1.96 \cdot \sqrt{\frac{pq}{n}}\right]$$

There is a problem, here — we use our unknown values in our own formula. We need to replace the ps and qs with the experimental versions, or $\hat{p}$ and $\hat{q}$, instead:

$$\left[\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 1.96 \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

The formula for the confidence interval is

$$\hat{p} \pm Z_x^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p}$ is our estimate and $\sqrt{\frac{\hat{p}\hat{q}}{n}}$ is our Standard Error (SE). $Z_x^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$ is called the Margin of Error (ME).

Confidence intervals, however, do not distinguish between numbers that are somewhat far from the interval and numbers that are extremely far; all it can tell us is whether or not the number is within the interval. Hypothesis tests exist to solve this problem.

5.3 Hypothesis Tests

We have a couple of key questions: what is the claim? What is the evidence presented?

There are five steps to completing a hypothesis test.

1. Write null and alternative hypotheses
2. Check assumptions
3. Calculate a p-value
4. Make a decision
5. Interpret the results

5.3.1 Hypotheses

The null hypotheses (H_0) must be written about a parameter and must include =. A null hypothesis asks, what if there is no change? What if there was no change to the original claim? For example, if we were to be testing whether or not a certain claim that p is less than 90% is correct. Then,

$$H_0 : p = 0.9$$

There also exists the alternative hypothesis (H_a), which denotes the opposite of H_0. There are three different ways to write this:

- The parameter is less than the number ($H_a : p < 0.9$, left tailed test)
- The parameter is greater than the number ($H_a : p > 0.9$, right tailed test)
- The parameter is not equal to the number ($H_a : p \neq 0.9$, two tailed test)

Choosing any one out of these tests depends entirely upon context. However, the default should be the two tailed test.

5.3.2 Basic logic

First, assume the null hypothesis is true: ask, how likely is it to see these results? In other words, calculate

$$P(\text{results} \mid H_0 \text{ is true})$$

If the p-value is extremely low, then it is extremely unlikely that the null hypothesis is true. If the probability evaluates to a reasonable figure, we don't have the evidence to determine whether or not H_0 is true. Note that this does not say that H_0 is true; there is an important distinction here.

Then, we should check our required prerequisites: the absence of bias, independence, and normality. If these conditions are met, we can assume

$$\mu_{\hat{p}} = p, \ \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Notice the lack of hats on the right hand sides of the equations.

The third step is to calculate the p-value. For example, if we were finding the chance of seeing a result of less than 0.827 given that our null hypothesis is true:

$$P(\hat{p} < 0.827 | p = 0.9), z = -2.12$$

Based on the normal probability table,

$$p = 0.017$$

Then, we must make a decision: is the p-value low enough to reject the null hypothesis? Is it too high that we cannot entirely reject the null hypothesis? We have the significance level, denoted α: it is the level below which we are to consider a p-value to be low. Recall that statistical significance indicates whether or not some statistic is 'odd' enough to not be attributed to chance; in this case, a p-value below α is low enough that it cannot be attributed to chance.

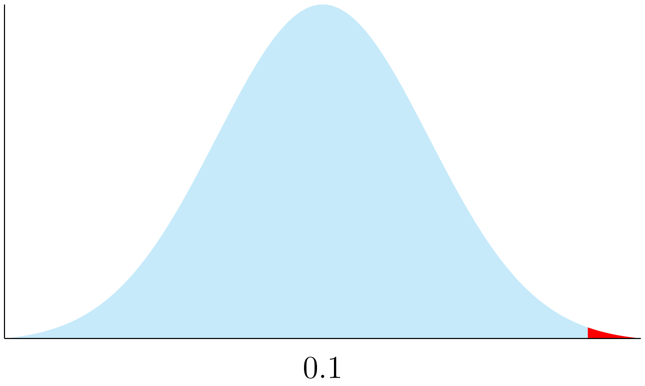
Finally, we need to interpret the results. The flow of logic needs to be reflected in this interpretation, and the words must place the situation into context. The statistical terms must be translated into regular english. It needs to discuss the meaning of these results.

5.4 Power of a test

	Defendant is innocent (Null Hypothesis)	Defendant is guilty (Alternative Hypothesis)
Jury finds defendant guilty (reject the Null)	Type I Error (rejecting a true Null Hypothesis)	:)
Jury finds defendant not guilty (Fail to reject the Null)	:)	Type II Error (failing to reject a false Null)

When making a decision, the Jury is told to make the decision such that it is "beyond a reasonable doubt". In statistics, this logic is what lends itself to the significance level — α.

We can think about probabilities. If we assume that the null is true, then the sampling distribution looks like



If our α level is 0.05, the red region in the distribution above represents our 'reject' region. α is the probability of a Type I error.

When we choose an α, we are choosing the hhighest probability of a Type I error we are willing to accept. If we minimize this α, we do so at the expense of the probability of Type II error.

We define β to be the probability of Type II error.

The power of a test is $1 - \beta$.

5.5 Two Proportion Z-Test

In certain situations with two samples, for example, detecting a difference in proportions before and after some action, we cannot get by with a simple one-proportion test. Instead, we need to perform a two proportion z-test. For example, we can have a situation in which we need to test whether a proportion is the same before and after some course of action is taken.

$$H_0 = p_{\text{before}} = p_{\text{after}},\ p_{\text{before}} - p_{\text{after}} = 0$$

$$H_A = p_{\text{before}} < p_{\text{after}},\ p_{\text{before}} - p_{\text{after}} < 0$$

We start by assuming that the null hypothesis is true, or that the difference between the two proportions is 0. Then, we can calculate our p-value:

$$P(\text{Seeing a difference in proportions} \mid H_0 \text{ is true})$$

We also need to consider our assumptions.

	Two Samples	Two Experimental Groups
Center	We have independent random samples, $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$	Treatments were assigned randomly, $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$
Spread	Sample picks are independent, 10% condition $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	Individuals in group are independent, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$
Shape	$n_1p_1,\ n_1q_1 > 10$, $n_2p_2,\ n_2q_2 > 10$, we can approximate with normal distribution	$n_1p_1,\ n_1q_1 > 10$, $n_2p_2,\ n_2q_2 > 10$, we can approximate with normal distribution

In order to get a z-score, we need our mean, our observed value, and our standard deviation. In this case,

$$\mu = 0,\ \hat{p} = \hat{p}_{\text{before}} - \hat{p}_{\text{after}}$$

To calculate the standard deviation, recall the Pythagorean theorem:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma_{\hat{p}_b-\hat{p}_a} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

We can use this theorem because the two samples from before and after are different from each other. However, we have a problem; using this requires the use of the actual parameters, which we do not have. Instead, we need a statistic known as $\hat{p}_{\text{pooled}}$.

$$\hat{p}_{\text{pooled}} = \frac{x_1 + x_2}{n_1 + n_2}$$

where

$$x_i = \hat{p}_i \cdot n_i$$

We approximate our actual proportion, p, with this $\hat{p}_{\text{pooled}}$.

$$\sigma_{\hat{p}_b-\hat{p}_a} = \sqrt{\frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1 - \hat{p}_{\text{pooled}})}{n_2}}$$

So, we can calculate our z-score with the information given, and we must interpret it in context.

$$z = \frac{\hat{p}_{\text{before}} - \hat{p}_{\text{after}}}{\sigma_{\hat{p}_b - \hat{p}_a}}$$

We can then make a confidence interval with this information:

$$\text{Confidence Interval: Estimate} \pm z_x^* \cdot SE$$

Where our estimate will just be

$$\hat{p}_{\text{before}} - \hat{p}_{\text{after}}$$

and our $z*$ will be our critical value. For a 95% confidence interval, this is 1.96. Now, for our SE, we cannot use $\hat{p}_{\text{pooled}}$ because we assumed that the two proportions were the same in doing so. Instead, we use our actual observed proportions:

$$SE = \sqrt{\frac{\hat{p}_b \hat{q}_b}{n_b} + \frac{\hat{p}_a \hat{q}_a}{n_a}}$$

So our final interval is

$$(\hat{p}_{\text{before}} - \hat{p}_{\text{after}}) + z_x^* \sqrt{\frac{\hat{p}_b \hat{q}_b}{n_b} + \frac{\hat{p}_a \hat{q}_a}{n_a}}$$

6 Unit 6

6.1 One Sample T-Tests

In unit 5, we learned about performing confidence intervals and hypothesis tests on proportions. The theorem that enables us do so is the Central Limit Theorem, which says that we can approximate sampling distributions with normal distributions, given that our sample size is "large enough". Luckily, the CLT applies to means, too, so we can perform the exact same processes with means as we did with proportions.

Recall the formula for confidence intervals:

$$\text{initial estimate} \pm z^* \cdot SE(\hat{p})$$

We would want to extend this definition to means, as well:

$$\text{initial mean} \pm z^* \cdot SE(\bar{x})$$

We know that the standard deviation of a sample size for means will be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

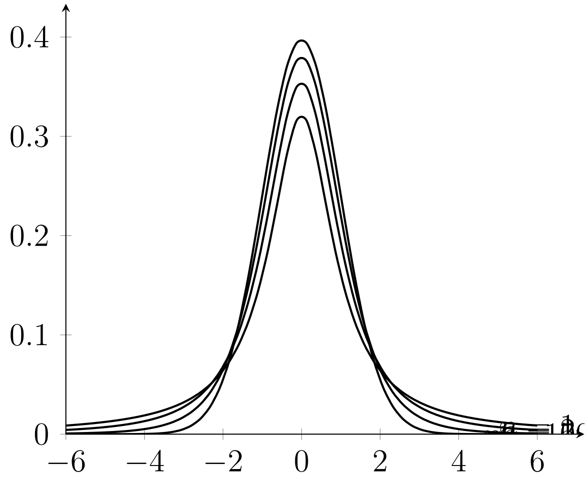
So we can just plug this $\sigma_{\bar{x}}$ in, right?

But how do we figure out the initial σ? Do we use an estimated standard deviation from the sample size that gave us our mean estimate? Would this even be a reliable approximation?

In the 20th century, statistician William S. Gosset investigated the ability to approximate the standard error with the measured standard deviation from the example. Eventually, he developed a distribution known as the Student's t, which is actually a family of distributions that change according to the degree of freedom (sample size - 1).

The Student's t distribution is somewhat fatter than a typical normal distribution, although it has been proven that as the limit of the sample size goes to infinity, the distribution approaches the Normal.

The diagram below displays the Student's t at $n = 100,\ n = 5,\ n = 2,\ n = 1$.



Since we just changed our distribution, then, we can't use z^* as the critical value; instead, we need to use a new critical value, defined as

$$t^*_{n-1} = t^*_{df}$$

Just like we have a normal probability table to find critical values for specific z-scores, we also have t-tables. Therefore, our confidence interval for means is

$$\bar{y} \pm t^*_{n-1} \cdot \frac{s}{\sqrt{n}}$$

Making a confidence interval for means requires similar assumptions and conditions as proportions.

1. Randomization: the data must be random and/or nonbiased for conclusions to be valid and meaningful.
2. Independence: the sample size must be below 10% of the entire population in order to ignore the effects of nonreplacement on the independence of the sample.
3. Nearly Normal: the initial distribution of the data must be somewhat unimodal and symmetric, and shouldn't have skewness or outliers.

We can also perform hypothesis tests for means. Given some null hypothesis

$$H_0 : \mu = \mu_0$$

and some H_A, we can first check our assumptions, and then calculate the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}$$

where

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

We then use a t-chart to find a p-value, and come up with a conclusion.

6.2 Comparing two means

In comparing two sample means, the population parameter of interest becomes the quantity

$$\mu_1 - \mu_2$$

Analogously, the statistic of interest is

$$\bar{x}_1 - \bar{x}_2$$

Much like with parameters, we can conduct hypothesis tests and construct intervals for sampling distributions of means. Now, we have a formula for combining the standard deviations of independent distributions (the Pythagorean theorem). Thus, if we just apply the theorem to our sampling distributions,

$$SD(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2}$$

In cases where we don't know the true standard deviations, we must calculate standard error instead:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}$$

6.2.1 Assumptions

As always, these comparisons require a set of assumptions.

1. Randomization: data must be unbiased, randomized, and representative of the population to assume that the center is $\mu_1 - \mu_2$.

2. Independence: the two samples must be independent from each other, and each sample must constitute less than 10% of the entire sampling population to ignore non-replacement. Once this has been assumed, we can use the formulas for $\sigma_{\bar{x}_1 - \bar{x}_2}$.

3. Nearly Normal: both initial distributions should be well behaved and somewhat normal. Some skewness or outlirs are forgiveable for larger sample sizes due to the CLT. Once this has been assumed, we can use Gosset's t to approximate the sampling distribution.

6.2.2 Two sample t-intervals

Two sample t-intervals work the same way one sample t-intervals do; we need a center $(\bar{x}_1 - \bar{x}_2)$ and a margin of error, which is composed of the standard error and the critical t value.

$$ME = t^* \cdot SE(\bar{x}_1 - \bar{x}_2)$$

The only problem is the t^* value, then; to find such a value, we need to find a combined "degrees of freedom" quantity to use in a t-chart lookup. The formula for this combined *df* quantity is somewhat ugly:

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

6.2.3 Two sample t-tests

Like two proportion z-tests, two sample t-tests work by first setting the central assumed mean as the initial difference between the mean paramters.

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

After deciding our alternative hypothesis and assumptions, we can obtain the t-score:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{SE(\bar{x}_1 - \bar{x}_2)}$$

where

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Using the formula for *df* as seen in section 6.2.2, we can use the t-chart to obtain a p-value.

6.2.4 Pooled t-tests

Pooled t-tests make the assumption that the variances of the two samples are equal.

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Thus, the standard error with this pooled variance statistic becomes

$$SE_{\text{pooled}}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}}$$

In addition, the formula for the degrees of freedom is much simpler with a pooled t-test:

$$df = n_1 + n_2 - 2$$

However, pooled tests are rarely ever used because they rely heavily on the assumption that the variances are equal, which is almost never the case.

6.3 Paired samples

Some studies will conduct research on the same groups of subjects, recording changes based on time or treatment. Because these are the same people, we cannot say that the two sample groups are independent, and we cannot perform two sample t-tests or two sample t-intervals. These types of samples are known as matched pairs. To conduct hypothesis tests and intervals on these types of data, we need to first take the difference in between each metric for each given "row" of data, and then perform a one-sample t-test or t-interval. This is known as a paired t-test.

The degrees of freedom for a paired t-test will simply be the number of pairs minus 1.

6.3.1 Assumptions

1. Paired condition: the data must be paired to perform a paired t-test.
2. Independence: the data must come from random samples, the differences between each subject should be independent from one another, and we should think about the effects of non-replacement.
3. Nearly normal: a histogram of the differences should look well behaved and somewhat normal. This can be mitigated with a larger sample size.

6.3.2 Paired t-tests

With paired t-tests, we are assuming the null hypothesis

$$H_0 : \mu_d = \Delta_0$$

where Δ_0 is our initially assumed difference and μ_d is the mean of the distribution of differences within each pair. Naturally,

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$$

and the t score is calculated in the same way.

7 Unit 7

7.1 Chi-Square Tests

To analyze categorical and count data, we can use χ^2 distribution tests. There are three types of χ^2 tests:

1. χ^2 test of Goodness of Fit — One sample, One variable
2. χ^2 test of Independence — One sample, Two variables
3. χ^2 test of Homogeneity — More than one sample, one variable

7.1.1 Goodness-of-fit tests

We first decide the hypotheses for the data. For example, our null hypothesis can be

"The observed data conforms with Benford's law"

The alternative hypothesis, then, would be the notion that the data does not conform with Benford's law.

We have a few required assumptions:

1. The data must be categorical or count data.
2. The selection must have been random and the data points must be independent.
3. N needs to be big enough.

Then, we need to look at and compare the observed data against the expected data. Intuitively, we need to find a way to measure the "distance" of the observed data points from the expected data points.

$$O - E$$

However, since the signed values would mess up the measure, we can take

$$(O - E)^2$$

The measure must be relative to the size of the sample, and thus the formula becomes

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

These measures are known as chi-square components. We can then add the components:

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

This is the chi-square test statistic. The only thing we need to consider, then, is the distance of this test statistic from the ideal value if the expected and observed were the same, 0.

The degrees of freedom in a goodness-of-fit test is the number of categories in the test minus one.

We can find a p-value from a chi-square probability chart given the quantities of degrees of freedom and the chi-square test statistic.

7.1.2 Independence tests

Our null hypothesis will be something along the lines of determining whether some event happening is independent of the variables present. We have a few assumptions:

1. Data must be categorical.
2. Data must be assigned randomly.
3. N must large enough.

Our expected count for any row-column pair would is equal to

$$\frac{\text{Row count} \cdot \text{Column count}}{\text{Grand total}}$$

We can use this expected count statistic to calcuate the chi-square statistic and conduct the test. The degrees of freedom for this test is equal to

$$(\text{number of rows} - 1)(\text{number of columns} - 1)$$

7.1.3 Homogeneity tests

This test operates essentially the same way as an independence test; the distinction lies in what is being tested for. In this case, we are seeing if two samples are the same for a set of particular samples. The expected count for a row-column pair, again, is

$$\frac{\text{Row count} \cdot \text{Column count}}{\text{Grand total}}$$

With the same formula for degrees of freedom, we find a p-value from a chi-square distribution chart.

7.2 Linear Regression T-Test

Recall that we can make a regression line based on samples:

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

However, this is only *a* regression line, based on a single sample. *The* regression line, on the other hand, is

$$\mu_y = \alpha + \beta x$$

Note the use of greek letters; α and β are true parameters that we attempt to approximate with a and b. The question becomes, what can we infer about β by knowing b? To answer this question, we perform a linear regression t-test.

We first assume that there's no relationship between x and y, or,

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

We have a few assumptions that we need, as always:

1. Random selection

2. We should be able to do a linear regression in the first place — the scatterplot should be "straight enough"

3. Homoscedasticity — the error around the line must be independent of x. The residuals plot should be *boring*, and there should be no trend.

4. The error around the line must be spread normally for every value of x; that is, most residuals should be close to 0.

Our degrees of freedom is equal to

$$df = n - 2$$

This is because, by definition, a line with only two points has no error.

To calculate our t score, we need an estimate and a standard error:

$$t = \frac{b - 0}{SE}$$

We know what b is, the estimate of our slope given the sample. How do we calculate SE? SE is equal to the "SE coefficient" value on a given coefficient table. Formulaically,

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

where s_e is the spread of the data around the line (or, alternatively, the spread of the residuals).