Basics for Advanced Regression

E. C. Hedberg

August 24, 2013

1 The normal distribution

In these notes I will be making many interval-ratio draws from the normal (or Z) distribution. When we simulate these data (note that I didnt say this data since data is plural), I can control the mean (μ) and the standard deviation (σ) or variance (σ^2). Below we define these quantities. For example, a variable can be described as normally distributed with the symbols $N \sim (\mu, \sigma^2)$ for normally distributed with mean μ and variance σ^2. Figure 1 is a picture of the normal distribution for a range of x from -5 to 5, with a mean of 0 and variance of 1.

2 Univariate Statistics

2.1 The mean, statistics, and parameters

To calculate the mean, simply add up the numbers for each of the cases and divide by the total number of cases. For example, the mean of the numbers 1,2,3,4,5,6,7,8,9 is (1+2+3+4+5+6+7+8+9) / 9 = 41/9 = 4.556. Statisticians and mathematicians symbolize the addition of values in a set by using the summation symbol, Σ. The $\sum_{i=1}^{N}$ phrase represents the summation of each i^{th} case, starting with the first one $i = 1$ and ending with the N^{th} case. To calculate the mean, this sum is divided by N, or the total number of cases.

The formula of the mean is:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N},$$

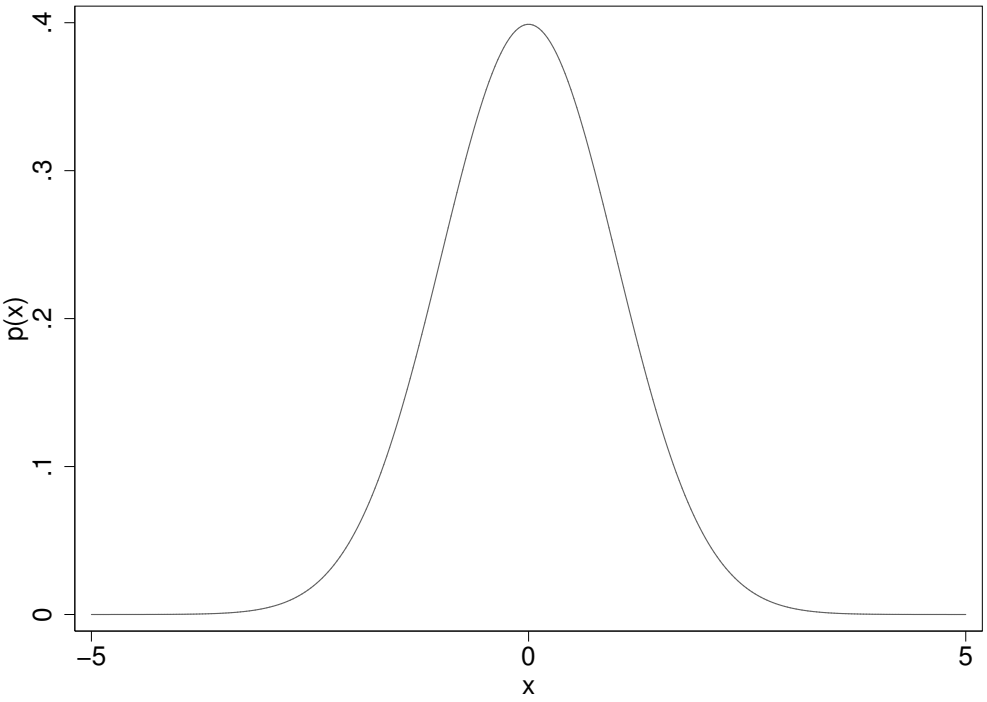

Figure 1: A variable x distributed $N \sim (0, 1)$.

which is estimated by the statistic:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

In the previous paragraph, the term statistic is used. It is important to note that there are two concepts: parameters and statistics, which are often conflated by students. Frequentists generally think that the population at large can be described with parameters. For example, one parameter is the mean, or μ. For example, a statistician might say that the population of objects (e.g., people) has some mean income. The statistician could then use a sample of objects to estimate this parameter. This estimate is the statistic.

For example, the estimate of the population mean, μ, of some variable x is the mean statistic $\bar{x}$. If the variable was w, the statistic would be $\bar{w}$. The estimate of the population variance σ^2 is the statistic s^2. One way to think about this is with the nemonic **PPSS**: Populations have Parameters, Samples have Statistics.

Of course, these statistics are not going to be absolutely correct, and will vary from one sample to the next. Thus, one of the things that researchers fret most about is how well a certain statistic estimates a parameter. This is the sampling variation of that statistic, or the variance of the estimate. The square root of the variance of a statistic is the standard error. More on that later.

2.2 Deviance

The next important idea is that of deviance. This is central to most of what statistics is about since statisticians often wonder how one population differs from another that has experienced some effect or treatment. The most basic deviation is the difference between a cases value of x and the mean of x, $\bar{x}$. Deviance can be symbolized as:

$$d_i = x_i - \bar{x}.$$

Even this basic idea has substantive interest. It is often important to know how some case differs from what is typical. Here, the mean is considered typical, and it is valuable to know whether a case is typical, above typical, or below typical. Negative deviance values are below the mean, positive deviance values are above the mean, and 0 deviance is exactly the mean.

2.3 The sum of squares

It is often important to understand how much deviance there is in our sample. One method might be to add up each deviance score, but because of properties of the mean this will wind up at 0 which makes this a rather useless method. As luck would have it, there is a simple solution to this problem.

By squaring each deviance, the negative values of defiance are effectively made positive (since the square of any negative number is positive) and then by adding up each squared deviance term, a measurement of total deviance is obtained. We call this total deviance the sum of squares, or SS:

$$SS = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

One interesting way to view SS is to consider it a measure of how much information is available in a variable. The larger the SS, the more information that is contained in the variable that can be modeled and correlated with other variables. If the SS is small, then little will be correlated with it.

2.4 The variance

The next quantity is the variance. This is simply the average of SS. In the population, this is

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}.$$

The estimate of the population variance is the sum of squared deviations divided by N-1:

$$s_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}.$$

We divide by N-1 instead of N because we lose a degree of freedom since this statistic employes another statistic (the mean). This is also a correction to remove bias, since this estimate of the population variance without this correction is slightly smaller than the expected parameter.

2.5 The standard deviation

A problem with the variance is that it in units of squared deviations. To get back to simple deviations, we take the square root of the variance to get the

standard deviation. In the population this is

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}},$$

and the statistic is

$$s_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}.$$

2.6 Back to the normal distribution

We now have all the ingredients for the normal curve. The formula using population parameters is

$$\Pr(x_i) = \frac{\exp\left[-\frac{1}{2}\sigma\left(x_i - \mu\right)^2\right]}{\sigma\sqrt{2\pi}},$$

which gives the probability of observing the i^{th} value. You can see the exponent of the squared deviance and standard deviation on top, $-\frac{1}{2}\sigma\left(x_i - \mu\right)^2$, and the standard deviation on the bottom, $\sigma\sqrt{2\pi}$. Figure 1 shows this function for a mean of 0 and a variance of 1.

To make the point, this formula can be expressed as an estimated probability for any i^{th} case using the estimated parameters for the mean and standard deviation:

$$\widehat{\Pr(x_i)} = \frac{\exp\left[-\frac{1}{2}s\left(x_i - \bar{x}\right)^2\right]}{s\sqrt{2\pi}}.$$

Note the use of the hat ($\hat{}$) symbol. It is another way to signal to readers that the number being calculated is an estimate of a population parameter.

2.7 Why the mean is the best estimate of what is typical for continuous variables

These notes are for regression. So now is the time to know that regression produces a conditional mean (a mean that exists under certain specified conditions). Since learning regression is what we are here to do, it makes sense to show how useful means really are. If we use as our criteria for whether a summary statistic about the data is good or not is the deviation

between the observation and our statistic, then we want a summary statistic to minimize these differences.

As mentioned earlier, adding up all of the deviances will produce 0. To circumvent this problem, the deviances were squared and then added up. Therefore, we want our mystery statistic, θ to be a *function* of these squared deviations:

$$S(\theta) = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (x_i - \theta)^2$$

A "good" estimate will have the least amount of differences. Thus, we want our statistic to minimize $\sum_{i=1}^{N} (x_i - \theta)^2$. So, how to we figure out which statistics will minimize the sum of these squared deviations? Let's break it down into known and unknown quantities. Someone smarter than I figured out that if you rearrange the terms, you can express this function as

$$S(\theta) = \sum_{i=1}^{N} x_i^2 - 2\theta \sum_{i=1}^{N} x_i + N\theta^2$$

Since, theoretically, we know our data, and thus know the sum of x squared, $\sum_{i=1}^{N} x_i^2$, and the sum of x, $sum_{i=1}^{N} x_i$, and N, then this function only has one unknown, θ. We can use calculus to find the first derivative of this function with respect to the mean

$$\frac{dS(\theta)}{d(\theta)} = -2 \sum_{i=1}^{N} x_i + 2N\theta$$

Then we set this first derivative to 0 (the first derivative of a function is its slope, and when the slope of a function is zero, the function is at a minimum or maximum) and solve for θ:

$$\frac{dS(\theta)}{d(\theta)} = -2 \sum_{i=1}^{N} x_i + 2N\theta$$

$$0 = -2 \sum_{i=1}^{N} x_i + 2N\theta$$

$$2 \sum_{i=1}^{N} x_i = 2N\theta$$

$$\sum_{i=1}^{N} x_i = N\theta$$

$$\frac{\sum_{i=1}^{N} x_i}{N} = \theta$$

Now that we have it, we name the statistic μ also call it the mean. For example, here are some summary statistics from a random variable in Table 1 Graphing functions so that you can visualize them is extremely important.

Table 1: Summary of x

variable	**mean**	s
x	9.885959	1.074514

The next step is to draw a graph showing how the sum of squares would change for other values of the mean. We can draw the

$$S(\theta) = \sum_{i=1}^{N} x_i^2 - 2\theta \sum_{i=1}^{N} x_i + N\theta^2$$

In Figure 2, we show that the mean we estimate does indeed present the minimum sum of squared errors. Any other value for the mean increases error.

2.8 Why the proportion is the best estimate of what is typical for dichotomous variables

While the least squares principle works well for dichotomous variables (the math doesnt care if the numbers only vary between zero and one), the next step is to introduce the idea of likelihood. This will be important once we start estimating logits and probits and other models with non-normal distributions.

Lets start with an example. A researcher takes random poll of 100 people and asks them if they like sardines. Some people said they liked sardines and were coded as 1, while others didnt like sardines and were coded as 0. The frequency table is in Table 2. The researcher would like to use this data to determine the probability that someone in the population will like sardines. Of course, we would estimate the mean of this variable to get the proportion; see Table 3. Why is this the right thing to do? Lets start with the function

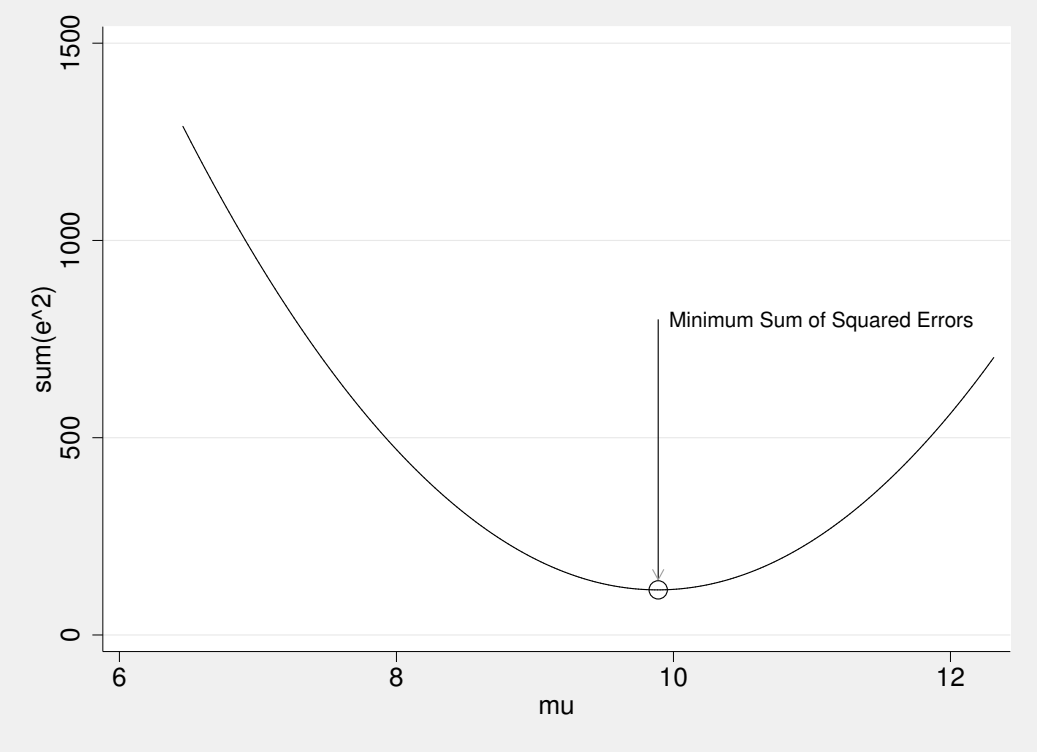

Figure 2: The function $S(\theta) = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (x_i - \theta)^2$ for random variable x.

Table 2:

Item	Number	Percent
No	73	73
Yes	27	27
Total	100	100

Source:

Table 3:

variable	**mean**
x	.27

that determines the chance of observing a 1 or 0 for some variable x, given a probability p

$$f\left(x|p\right) = \Pr\left(x = q\right) = p^{q}\left(1-p\right)^{1-q}, q = 0, 1$$

The next important topic is the idea of a likelihood function. Researchers typically want to know the chance of observing the data they have if they assume some pre-specified model. In other words, if our model says that the chance of observing a 1 for any case is 50 percent, then how likely is it that we observed only 27 out of a 100? The likelihood function serves this purpose. It quantifies how likely we are to observe our data if we assume a specific model. In this case, our model is just some proportion we expect. Later, our models will become much more complicated multivariate functions. The likelihood function is

$$L\left(p|x_1...x_N\right) = \prod_{i=1}^{N} f\left(x_i|p\right)$$

To calculate the likelihood function, take the assumed chance of observing (which is p for $x = 1$ and $(1-p)$ for $x = 0$) for each case and multiply them together. (The symbol Π works like the summation symbol, Σ, except it means multiply everything instead of adding everything). The likelihood function reduces to

$$L\left(p|x_1...x_N\right) = p^{\sum_{i=1}^{N} x_i}\left(1-p\right)^{N-\sum_{i=1}^{N} x_i}$$

Which also looks scarier than it is. Its just the assumed models probability, p, to the power of the total number of observed 1s times $1-p$ to the power of the total number of observed 0s. Taking the natural log of this function gives us

$$\ln\left(L\left(p|x_1...x_N\right)\right) = \left(\sum_{i=1}^{N} x_i\right)\ln\left(p\right) + \left(N - \sum_{i=1}^{N} x_i\right)\ln\left(1-p\right)$$

Remember from our data that there are 27 yes answers and 100 - 27 = 73 no answers, so thus function is

$$\ln\left(L\left(p|x_1...x_N\right)\right) = \left(27\right)\ln\left(p\right) + \left(73\right)\ln\left(1-p\right)$$

Figure 3 tells a simple story: if one assumes a model where the chance of

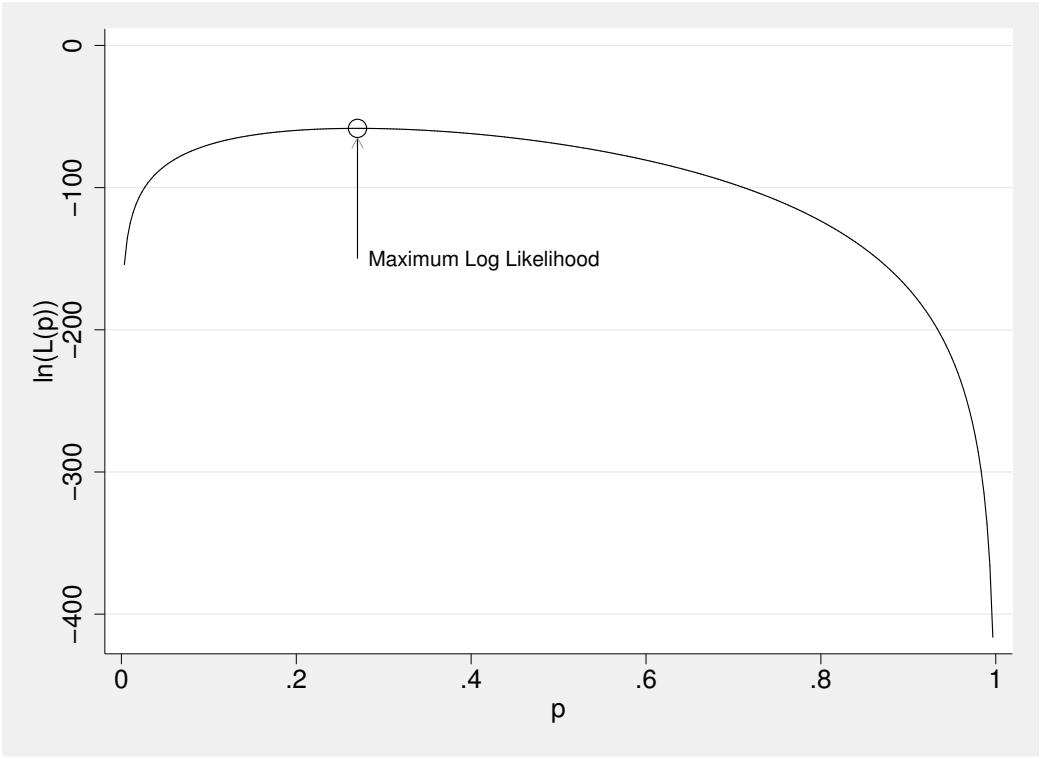

Figure 3: The function $\ln(L(p)) = (27)\ln(p) + (73)\ln(1-p)$ for random variable x.

observing a 1 is the estimated mean, then the likelihood function is maximized. If any other model is assumed, the likelihood function decreases. This suggests that the optimum model for our data is the estimated mean. This may seem like we are going in circles, but when we have more complicated models, the idea of maximizing the likelihood of observing the data we have becomes powerful.

What process can be used to find the estimate that maximizes the likelihood? Again, the first step is to find the first derivative (which gives us an equation for slope) of this function with respect to p:

$$\frac{d\ln\left(L\left(p\right)\right)}{dp} = \frac{\sum_i^N x_i}{p} - \frac{N - \sum_i^N x_i}{1-p}$$

The next step is to set the derivative of this function to 0 (remember, when the slope is 0, the function is either at a local minima or maxima) and solve for p

$$0 = \frac{\sum_i^N x_i}{p} - \frac{N - \sum_i^N x_i}{1-p}$$

$$0 = \left(\sum_i^N x_i\right)(1-p) - \left(N - \sum_i^N x_i\right)p$$

$$-\sum_i^N x_i = -pN$$

$$-\frac{\sum_i^N x_i}{N} = -p$$

$$\frac{\sum_i^N x_i}{N} = p$$

There we are, the formula that maximizes the likelihood is the mean, or proportion of 1s. In this case p = 0.27. It would be a simple matter to plug in the numbers (if they had been provided) to verify that p = 0.27.

3 Bivariate Statistics

4 Covariance

The covariance is a simple idea: instead of squaring the deviance of one variable from its mean, the covariance multiplies the deviances of two variables

from their means:

$$\operatorname{cov}(y, x) = \frac{\sum_i^N (y_i - \bar{y})(x_i - \bar{x})}{N - 1}$$

We can visualize this by using a scatterplot. Figure 4 is a scatterplot of

Table 4:

y	**x**
8.195923	4.514571
9.208269	5.355424
10.22313	6.229963
8.563372	3.842484
10.65503	4.900496
9.934275	4.995431
8.293035	4.799414
9.181551	4.337225
9.221762	5.783111
7.294033	4.062926

the data with the means of x and y marked with vertical and horizontal lines and each point marked with the deviations from the mean. The data appear in Table 4. Now, consider Figure 5. The means of x and y have been graphed into four quadrants. Each quadrant is defined by the sign of the deviation from the means of x and y. Observations whose values of x and y are both larger than the mean of x and y respectively will fall into quadrant one $(+, +)$. Observations whose values have x and y are both lower than the respective means will fall into quadrant three $(-, -)$. The key is that the product of any deviations in either of these quadrants will be positive (a negative times a negative is positive). On the other hand, points that fall into the other quadrants (quadrant two and quadrant four) will have products that are negative (a negative times a positive is negative).

Why do we care about the products of the deviations? Consider the numerator of the covariance formula, $\sum_i^N (y_i - \bar{y})(x_i - \bar{x})$. All this is doing is adding up these products. Therefore, more positive products will produce a large positive number, more negative products will produce a large negative

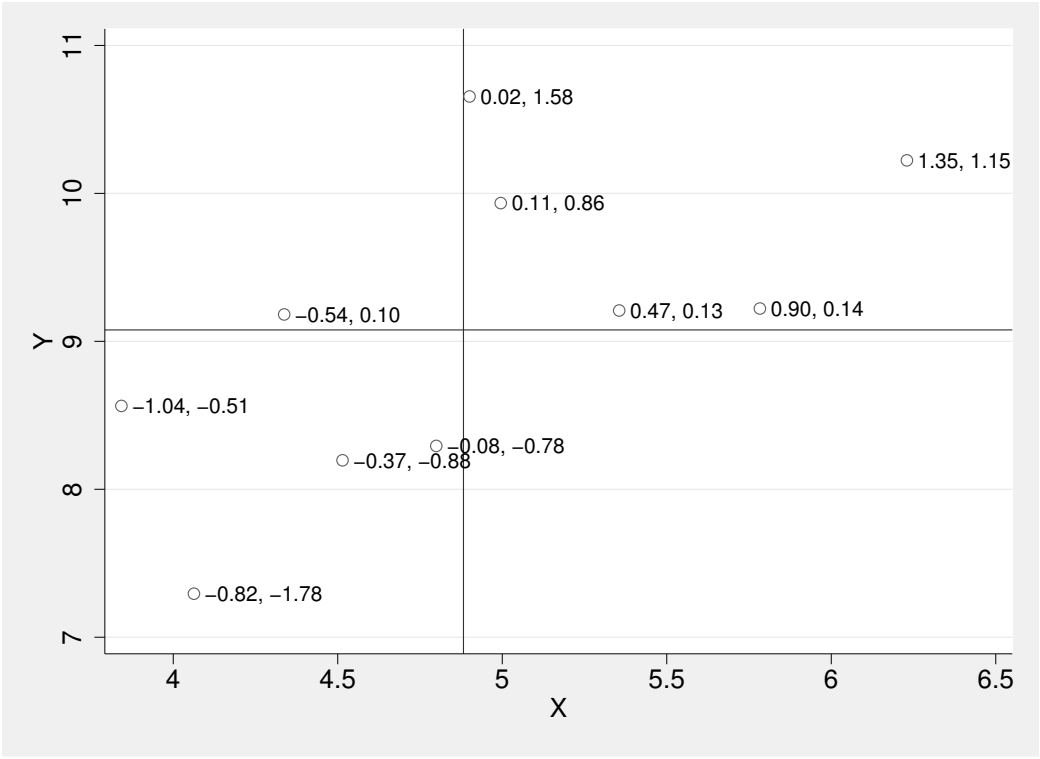

Figure 4: Plot of variables y and x with solid lines for means, labeled with deviations from the means

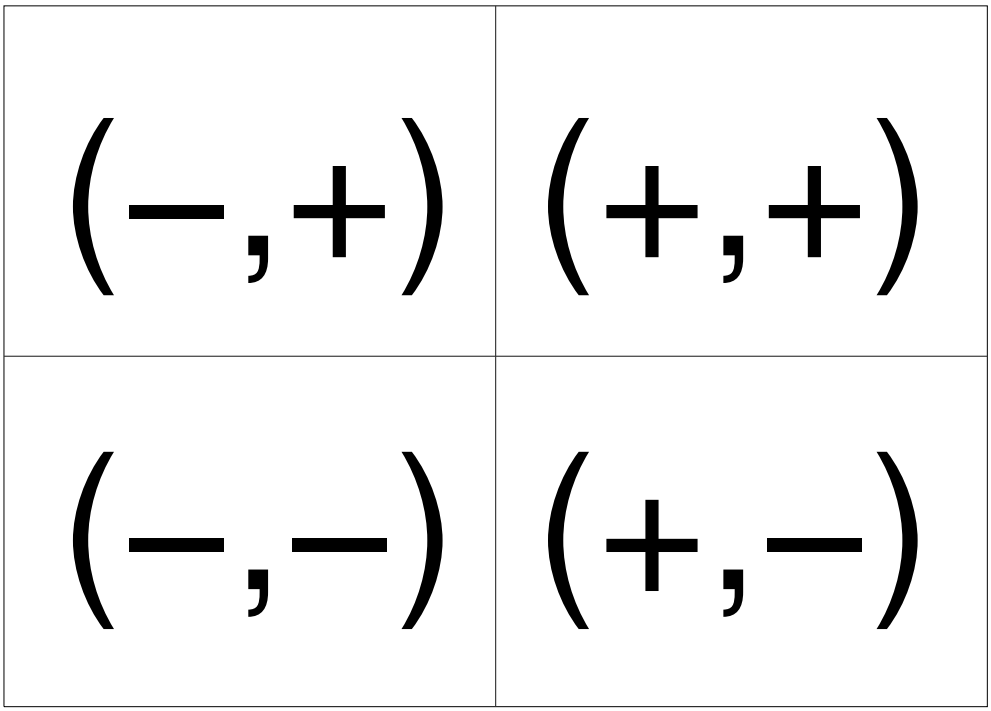

Figure 5: Plot of variables y and x with solid lines for means, labeled with deviations from the means

number. In addition, larger deviations of either x or y will create larger numbers, and large deviations of both x and y will create a really big number.

If more points fall into the positive quadrants $(+,+;-,-)$ the result is a positive number. If more points fall into the negative quadrants $(-,+;+,-)$, the result is negative. If equal numbers fall into each quadrant, then they will balance out and the number will be close to 0.

Figure 4 shows the deviations marked for each point on the scatterplot: what do you think the covariance is?

The denominator of the covariance formula (N-1) just turns it into an average, like the variance. Looking at Figure 4, we can tell that the relationship between x and y is positive. The direction of a relationship is generally thought of as what happens to y as x increases. Thus, the covariance in Figure 4 is 0.47. This tells me that when x increases, so does y. If this number was negative, then as x increases y would decrease.

4.1 Correlation

One of the biggest problems with covariance is that it is really hard to interpret anything meaningful from it. Covariance identifies if the relationship is positive or negative, but it doesnt provide information about the magnitude of the relationship. In addition, the units get all messy. For instance, if the variables were wages and years of education, then the units would be wage-years. What is a wage-year? The solution to this problem is standardization. It would be helpful to standardize this quantity, and one method is to use the sum of squares of the two variables involved. This is basically what the correlation coefficient is about. It is a standardized covariance. The correlation formula is

$$r_{y,x} = \frac{\sum_i^N (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i^N (y_i - \bar{y})^2 \sum_i^N (x_i - \bar{x})^2}}$$

The numerator of this equation is simply the covariance formula while the denominator is the sum of the squares equation. Since the units of the variables are the same for the numerator and denominator, the units cancel out Correlation coefficients range from -1 to 1. If it is 0, it means that there is no covariance between x and y. If it is close to 1, it means that there is a strong positive covariance between x and y. If it close to -1, it means that there is a strong negative covariance between x and y. All information about the units is lost. It does not matter what x and y are measured by.

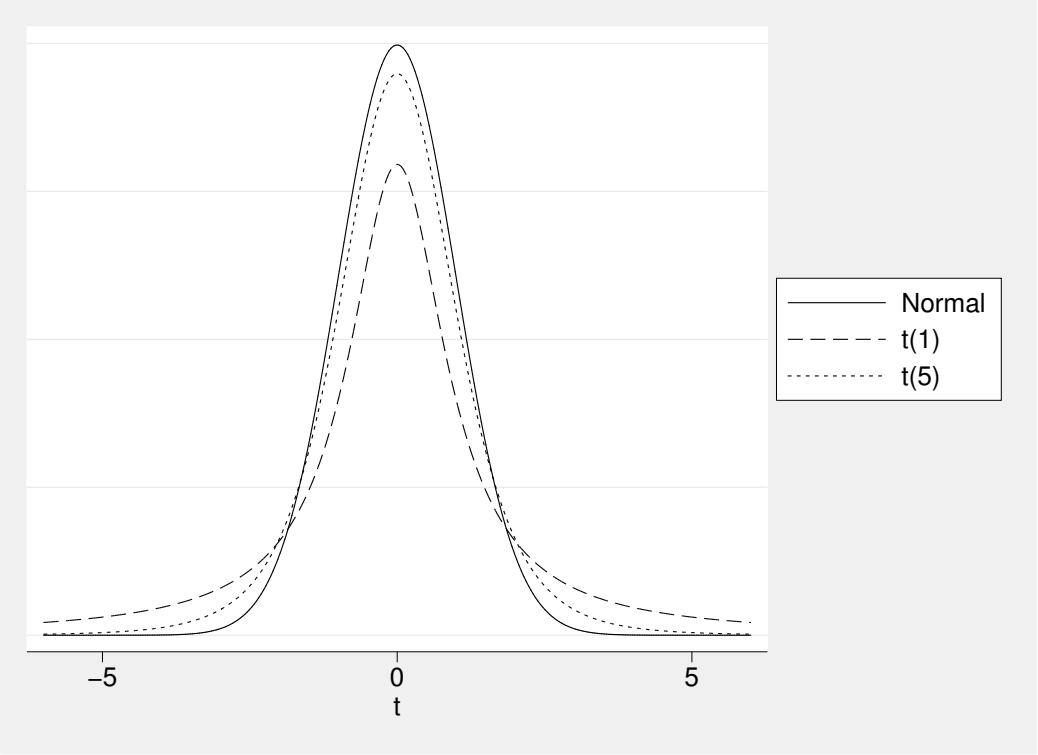

Figure 6: Plot of the normal distribution and two t distributions with 1 and 5 degrees of freedom

4.1.1 Testing the correlation coefficient (and intro to hypothesis testing)

What if we wanted to know if our correlation coefficient was *statistically* different than 0? In this case, we need to consider the sampling distribution of the correlation coefficient. I go into this in more detail below, but for now remember that we have a random *sample*, and if we started our research another day, we would get a different *sample*. Thus, the estimated correlation coefficient from our data is one of many possible estimates. We want to know the distribution of these estimates from different hypothetical samples, so we can know how likely our estimate is, assuming a world where a null hypothesis is true. In this case, our null hypothesis is that there is no correlation. For the data in Figure 4, the correlation is 0.61. The question is, is this statistically different than 0? Since we have a sample, we will rely on the student's t

distribution. As you can see in Figure 6, the t distribution is similar to the Z distribution except that it has slightly different shapes for different degrees of freedom. We can estimate a test statistic for a correlation with

$$t = \frac{|r|\sqrt{N-2}}{\sqrt{1-r^2}}$$

which in this case is

$$t = \frac{0.61\sqrt{10-2}}{\sqrt{1-0.61^2}}$$
$$t = 2.18$$

Now, in our case in Table 4, we have 10 cases, and with 2 means (for two variables) we have 10-2 = 8 degrees of freedom. Figure 7 is the sampling distribution of our statistic if we assume that the null hypothesis, or the hypothesis that the correlation is 0, is true. If you look carefully, you will see that the probability of finding a correlation of 0.61 with 10 cases is about 0.06. The math to find that number is a little tedious, so most stats books tell you what test you need to get in order to have a sufficiently low probability to reject the null hypothesis. In our case, that value is about 2.31. Why 2.31? That is a number large enough to claim that if we assume the null hypothesis, the chance of finding that test in our data is less than 0.05. Another way of saying that is that the Type I error rate is less than 5 percent. We sometimes call the Type I error rate α and so we say that $\alpha = 0.05$.

In this case, with a test statistic of 2.18, which is smaller than the critical value of 2.31, we have to accept the null hypothesis. It was close, but not close enough.

4.2 Chi-square

In the case of two categorical variables, neither covariance or correlation should be used to determine relationships. Instead, another method based on the idea of deviation is appropriate; the Chi-square (χ^2) test. The idea is to state a null hypothesis in terms of the researchers expectation for the results of their analysis. The researcher states what she/he expects, and then offer an alternative hypothesis. If the data deviates in a substantial degree from what the researcher expects, then the researcher must fail to reject the null hypothesis. In common practice, a null hypothesis indicates that there will be no statistically significant relationship between two variables while

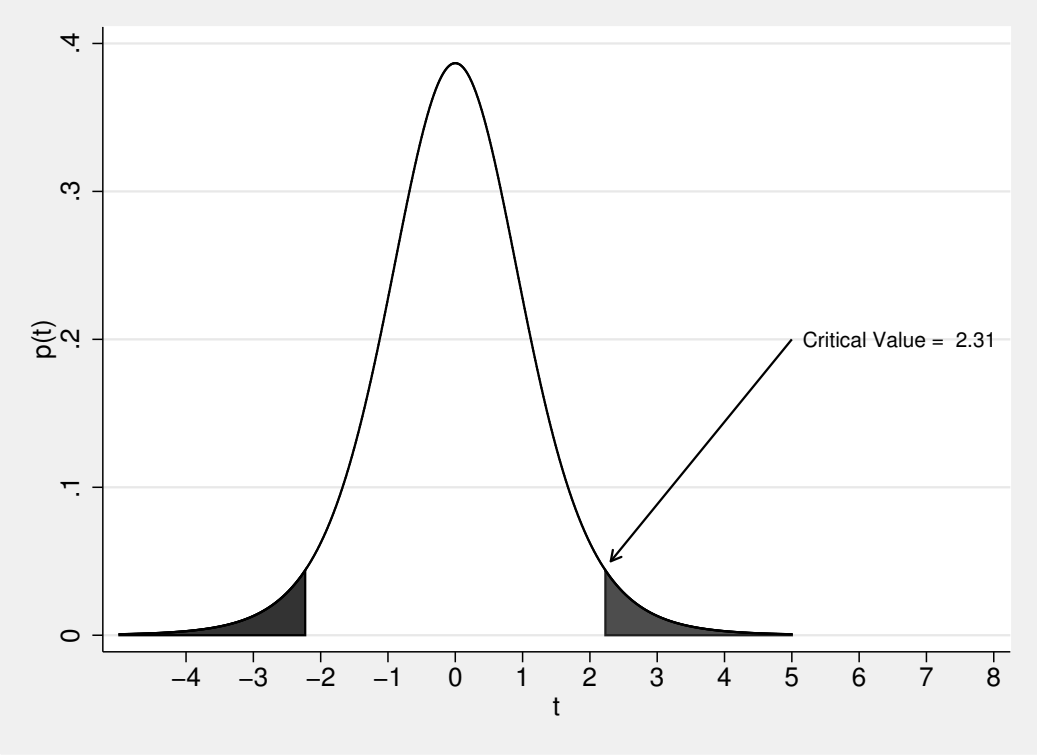

Figure 7: Plot of t distribution with 8 degrees of freedom and $\alpha = 0.05$ critical tails marked.

the alternative hypothesis indicates that there is a statistically significant relationship between the two variables.

The trick to Chi-square tests is that the expected value can be anything. You may remember the Chi-square tests from your research methods or statistics class. Frequently, data is presented in table format. In such a table we have (at least) two categorical variables x and y. The variable x takes on values such as $x \in \{1...i\}$ and variable y takes on values such as $y \in \{1...j\}$. Each cell in the table has a number of observations associated with it, n_{ij}, for a total of $N = \sum_i \sum_j n_{ij}$ observations. In the case of a two variable table, x makes up the rows and y makes up the columns. We also have totals for each row, n_{i+}, and for each column, n_{+j}. Chi-square tests then make a comparison between the observed frequency, n_{ij} and what the researcher expects each cell frequency to be, m_{ij}. This test is

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}.$$

A big value for this number tells us that there is a difference between our data and our expectation. Depending on what we are doing, this expectation is different. If we are testing to see if two variables are independent, when we test against a random expectation, or we expect the cells to simply be a function of the margins. So, in this case, m_{ij} is

$$m_{ij} = \frac{n_{i+} n_{+j}}{N}$$

So, what's a big value? In this case, it depends on how many cells are in the table. Any statistical test is compared to a function, generally based on some degrees of freedom. For example, a Chi-square test is compared against a Chi-square distribution with $(rows - 1)(columns - 1)$ degrees of freedom. If the value of the test is far enough along that distribution to be considered unlikely if the expectation is true, then we reject the expectation.

5 Basic Statistical Inference

5.1 Sampling distributions

In the correlation section we considered the idea of multiple samples. Almost all the methods in these notes assume a simple random sample. **If it is not**

a simple random sample, such as a two-stage cluster randomized sample from a major survey, these notes do not apply. While the population mean, μ , is fixed, the estimation of the mean, $\bar{x}$, is random because the selection of the sample is random. It therefore makes sense that if we were to gather another sample, that new samples estimation of the mean will be different. This is the idea of sampling distributions: given a large number of samples, how will the estimate likely vary? What would the range of our estimate be 95 percent of the time?

5.1.1 Basic principles of the central limit theorem

If repeated random samples of size N are drawn from any population with mean μ and standard deviation σ, as N becomes large the sampling distribution of sample means will approach normality, with a mean of μ and standard deviation $\frac{\sigma}{\sqrt{N}}$.

There are two important parts here: that the mean of the sampling distribution should be close to the actual mean, and the standard deviation of the sampling distribution should be close to the standard deviation of the population divided by the square root of the sample size.

Keep in mind that the population distribution doesnt have to be normal– but the sampling distribution becomes normal with large N.

It is important to note that the central limit theorem suggests that the mean of our sample can be used as our guess of the population mean and we can quantify the uncertainly of this estimate with the variance of the estimates, of which the standard error is the square-root. THis standard error is an estimate of the sampling distributions standard deviation.

5.2 The standard error of the mean

Of course, it is extremely expensive to obtain repeated samples to get a good estimate of the mean. Researchers generally only have the resources for a single sample.

The formula for the standard deviation of samples of the a mean is similar to the standard deviation of the sampling distribution, except we substitute the population standard deviation, σ, with the estimate of the standard deviation from our sample, s. We start with the variance of the statistic, $\bar{x}$:

$$V(\bar{x}) = \frac{\sigma^2}{N}$$

the square-root of which is the standard error

$$SE\left(\bar{x}\right) = \sqrt{\frac{\sigma^2}{N}}$$

or the more familiar formula

$$SE\left(\bar{x}\right) = \frac{\sigma}{\sqrt{N}}$$

and our estimate of this replaces σ with s

$$SE\left(\bar{x}\right) = \frac{s}{\sqrt{N}}$$

This gives us the variance of our mean estimate. All statistics have variances associated with them. Much of the more advanced statistics and sample statistics that are taught in graduate schools focus on getting the right variances, and thus the right standard errors. Much the time, the formulas of the point estimates (the statistics) are pretty much the same.

5.3 The standard error of a proportion

The variance of a mean is not a function of the value of the mean, it just depends on the standard deviation. In the case of proportions, the standard deviation is a function of the estimate itself:

$$\sigma_p = p\left(1 - p\right)$$

therefore, the standard error of the proportion is

$$SE\left(p\right) = \sqrt{\frac{p\left(1 - p\right)}{N}}$$

5.4 Hypothesis testing

So now that we have an estimate of the mean, and a sense of our uncertainty about this estimate, we can now ask a theoretical question: lets say we did draw several samples from the population, what is the range for 95 percent of the means we would get? We can do this by calculating a confidence interval of the estimated statistic. Confidence intervals generally take the form of

$$CI\left(\theta, \left(1 - \alpha\right) \times 100\right) = \hat{\theta} \pm t_{df,\alpha/2} SE\left(\hat{\theta}\right)$$

Where θ is the parameter of interest, $\hat{\theta}$ is the statistic that estimates the parameter, $SE\left(\hat{\theta}\right)$ is the standard error of that statistic, α is the type I error rate, and t is the critical value of the t-distribution associated with the df degrees of freedom of the statistic and the type I error rate.

In this case, the parameter of interest is the population mean, μ. We are estimating that with the sample mean statistic, $\bar{y}$. Note that we are now talking about a variable y instead of x, this is because we are transitioning from any old variable to outcomes, which are typically represented with the variable y. For that statistic, we also have a standard error $SE\left(\bar{y}\right)$.

In the social sciences, the common type I error rate is 5 percent, or $\alpha = 0.05$. What is the type I error rate? Type I error is the chance of falsely rejecting the hypothesis that the true parameter could fall outside the confidence interval. In other words, if we are going to make the assertion that some other parameter, θ_{null}, is different than our population parameter, θ, for which we have a sample estimate, $\hat{\theta}$, we want there to be only a $(1-\alpha)\times 100$ percent chance of being wrong.

Finally, we need to scale our standard error by something to create the confidence interval. Since the sampling distribution is normal with a standard deviation of $SE\left(\hat{\theta}\right)$, we want to form our confidence interval so that it gives us the bounds of percent of the possible sample estimates.

One option is to use the normal Z distribution, in which case, for example, a 95 percent confidence interval (meaning that $alpha = 0.05$) is bounded by -1.96 and 1.96. However, we typically use the t-distribution rather than the normal distribution because it flattens out as the sample size gets small, making the confidence intervals wider (see Figure 6). Thus, in order to get the right number from the t-distribution, we need to know the degrees of freedom, which in a one sample case is $N-1$. Thus, confidence intervals for smaller samples will always be larger. They are larger for two reasons. First, if we look at the formula for a standard error,

$$SE\left(\bar{x}\right) = \frac{s}{\sqrt{N}}$$

we see that the denominator is essentially the sample size. Smaller samples, larger standard errors. The second reason is that we use the sample size to determine the degrees of freedom when selecting the value from the t distribution. Smaller samples, larger values of t for any given α.

Why do we care about this? We care because we may want to make an assertion that the estimated parameter, $\hat{\theta}$, is different from some other

value, θ_{null}. If θ_{null} is outside the confidence interval, then the estimate is statistically different. Another way is to find the difference between $\hat{\theta}$ and θ_{null}, and divide by $SE\left(hat\theta\right)$. This gives you the value of t for a confidence interval, and you can compare it to the t you would use to construct a confidence interval, or the critical t. If

$$\frac{\hat{\theta} - \theta_{null}}{SE\left(\hat{\theta}\right)} > t_{critical,df,\alpha/2}$$

then the estimate is statistically different.

Can we be absolutely sure? No, because our confidence interval only covers $(1 - \alpha) \times 100$ percent of the sampling distribution, another sample could have given us a different answer.

This is all backing into hypothesis testing. We could play this game all day of calculating confidence intervals and seeing if some arbitrary value falls within the bounds. A much simpler approach is to do a hypothesis test.

Most stats text books will tell you that there are 5 steps to doing a hypothesis test that include making an assertion of our type I error we assume, stating the critical value, etc. I think these steps are silly, since no one ever does these discreetly and in that order. This is how it is done in practice. We typically think of θ_{null} as the null hypothesis (or the value of the parameter or model we will assume to be true). We can write the null hypothesis like this (in terms of the population parameter θ:

$$H_0 : \theta = \theta_{null}$$

where we say that the population parameter, θ, is equal to the null value, θ_{null}. The alternative hypothesis is that these values are unequal

$$H_1 : \theta \neq \theta_{null}$$

First, we calculate our test statistic. Test statistics for parameters that use the t-distribution generally take the form of

$$t = \frac{\hat{\theta} - \theta_{null}}{SE\left(\hat{\theta}\right)}$$

for t-tests when we have a specific degrees of freedome, or

$$z = \frac{\hat{\theta} - \theta_{null}}{SE\left(\hat{\theta}\right)}$$

for z tests when there are no real degrees of freedom to consider, or

$$\chi^2 = 2\left(\ln\left(L\left(\theta_a\right)\right) - \ln\left(L\left(\theta_{null}\right)\right)\right)$$

when comparing the results of two likelihood functions (the value is compared to the Chi-square distribution of a single degree of freedom). Basically, what we are doing is calculating the difference between what we estimated and what we assume is true (the null hypothesis) and then we divide that difference by the standard error of our estimate.

We then look up the probability of that test on the appropriate distribution associated with the statistic and degrees of freedom. If that probably is less than 0.05, then we say we can reject the null. I don't talk about one-tail tests.

A probability less than 0.05 means that what we get from our sample is highly unlikely if we assume the null hypotheses, so we can then reject the null hypothesis. There is still a chance that the null is correct and we got a weird sample, thats our type I error rate, but we feel that 5 percent or less is acceptable.

I really dont see many comparisons between a population mean and some other value. What I do see a lot of are comparisons between two groups that are sampled independently. We should be careful when we think about this, because in observational studies like the General Social Survey, we dont sample Republicans separately from Democrats. We randomly sample the population, and some are Republicans and some are Democrats. Even in experiments, we randomly sample a group then randomly assign treatment or control.

Regardless, we always talk about testing two independent means, for example between groups A and B. They are independent because the observations from group A are in no way related to observations from group B. However, we are not really dealing with a test of two statistics, we are estimating a single statistic, *the difference*. In population terms, this is

$$\theta = \mu_A - \mu_B,$$

and from our sample estimates it is

$$\hat{\theta} = \bar{y}_A - \bar{y}_B.$$

We then state the null hypothesis as this difference being 0 (i.e., $\theta_{null} = 0$) as

$$H_0 : \theta = \theta_{null}$$

$$H_0 : \mu_A - \mu_B = 0$$

and the alternative hypothesis as

$$H_1 : \mu_A - \mu_B \neq 0.$$

In the case of independent means, we can estimate the standard error of this difference using a pooled standard deviation

$$SE\left(\bar{y}_A - \bar{y}_B\right) = \frac{s_{pooled}}{\sqrt{N}}$$

where

$$s_{pooled} = \sqrt{\frac{\left(n_A - 1\right) s_A^2 + \left(n_B - 1\right) s_B^2}{n_A + n_B - 2}}.$$

The t test is then

$$t = \frac{\bar{y}_A - \bar{y}_B}{SE\left(\bar{y}_A - \bar{y}_B\right)}$$

We then evaluate the value of t against some level of α using $N - 2$ degrees of freedom.

If there is a relationship between groups A and B, like one group was a set of pre-tests and the second group was a set of post tests, each pre-test and post-test was from the same set of people, then we would be doing a paired t-test. In this case the test is

$$t = \frac{\bar{d}\sqrt{N}}{s_d}$$

where

$$\bar{d} = \frac{\sum_i^N \left(y_{Ai} - y_{Bi}\right)}{N}$$

and

$$s_d = \sqrt{\frac{\sum_i^N \left(y_{Ai} - y_{Bi} - \bar{d}\right)^2}{N - 1}}$$

with N pairs of values. The test is then evaluated against the t distribution using $N - 1$ degrees of freedom.

5.5 The analysis of variance (ANOVA)

When we are considering three or more groups, we ask whether these groups are different from each other. Another way to think about this is to think about how much of the variance occurs between the groups compared to how much variance occurs within the groups. This ratio is the idea behind ANOVA.

The building blocks of ANOVA is that we can think about each i^{th} case as part of the j^{th} group. We can then conceptualize a way to think about deviance that involves groups.

Here, we can think of the total difference between an observation and the overall mean, $y_{ij} - \bar{y}$, as the difference between that observation and its groups mean, $y_{ij} - \bar{y_j}$, plus the difference between that groups mean and the overall mean, $\bar{y_j} - \bar{y}$:

$$y_{ij} - \bar{y} = (y_{ij} - \bar{y_j}) + (\bar{y_j} - \bar{y})$$

This idea has implications for the sum of squares. We can now think of the total sum of squares (SST) as breaking down into the sum of squares within groups (SSW) and the sum of squares between groups (SSB):

$$SST = SSW + SSB$$

or

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y_j})^2 + \sum_{j=1}^{k} n_j (\bar{y_j} - \bar{y})^2$$

There is some new notation here. We now have k groups with the subscript j, and each j^{th} group has cases with the subscript i.

ANOVAs test statistic relies on the F-distribution to evaluate type I error. The F-distribution is used to test ratios and is thus defined by two degrees of freedom, one for the numerator and one for the denominator of the ratio, and there are no negative statistics. As the number of each increases to a large number, the distribution starts to look normal.

As was just mentioned, the test statistic is a ratio. Again, we are concerned with the ratio of the variance that occurs between groups compared to that within groups. Thus, this ratio uses as the numerator the mean square between, which is the sum of squares between divided by the number of groups - 1:

$$MSB = \frac{\sum_{j=1}^{k} n_j (\bar{y_j} - \bar{y})^2}{k - 1}$$

The denominator of the ratio is the mean square within, which is the sum of squares within divided by the number of cases minus k group means

$$MSW = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(y_{ij} - \bar{y_{ij}}\right)^2}{N - k}$$

The ratio then looks like this

$$F = \frac{MSB}{MSW}$$

and we test how much of the F-distribution (with $kandN - k$ degrees of freedom) is left after this value for our type I error rate of our test of the the null hypothesis that all group means are the same. In other words, our null hypothesis is that all the group means are the same:

$$H_0 : \mu_1 = \mu_2 ... = \mu_k.$$