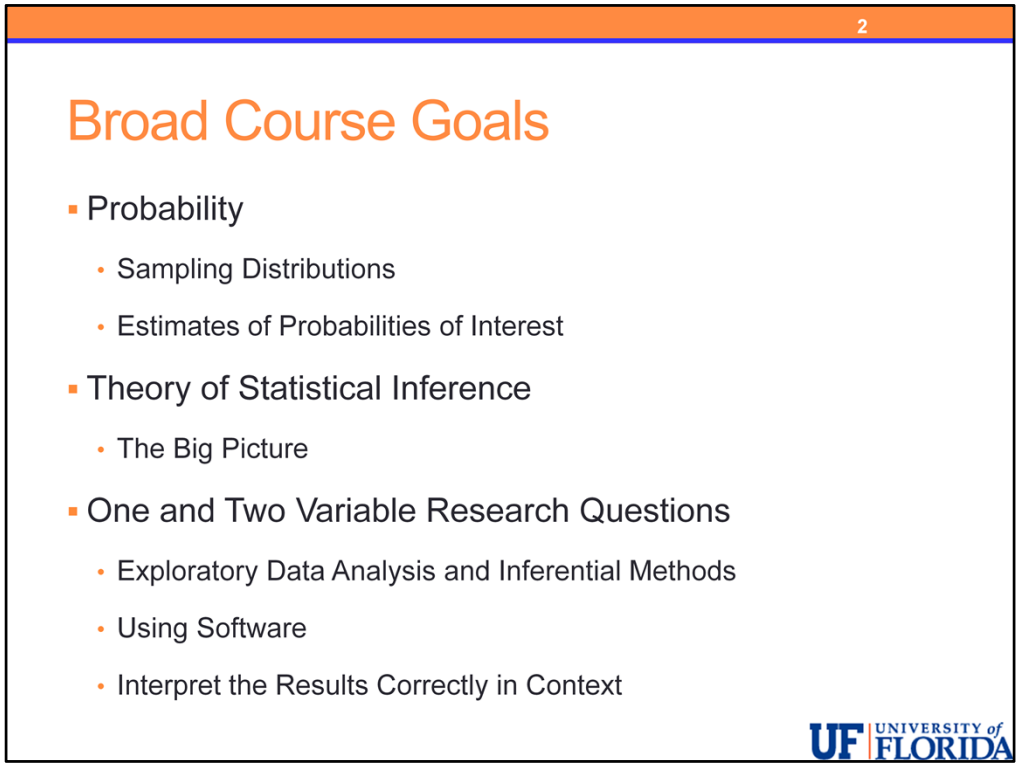
COURSE SUMMARY

Putting Everything Together

UF UNIVERSITY of FLORIDA

Now, we will give an overview of entire course. We will discuss as many details as possible, however, we will not be able to review everything in depth.

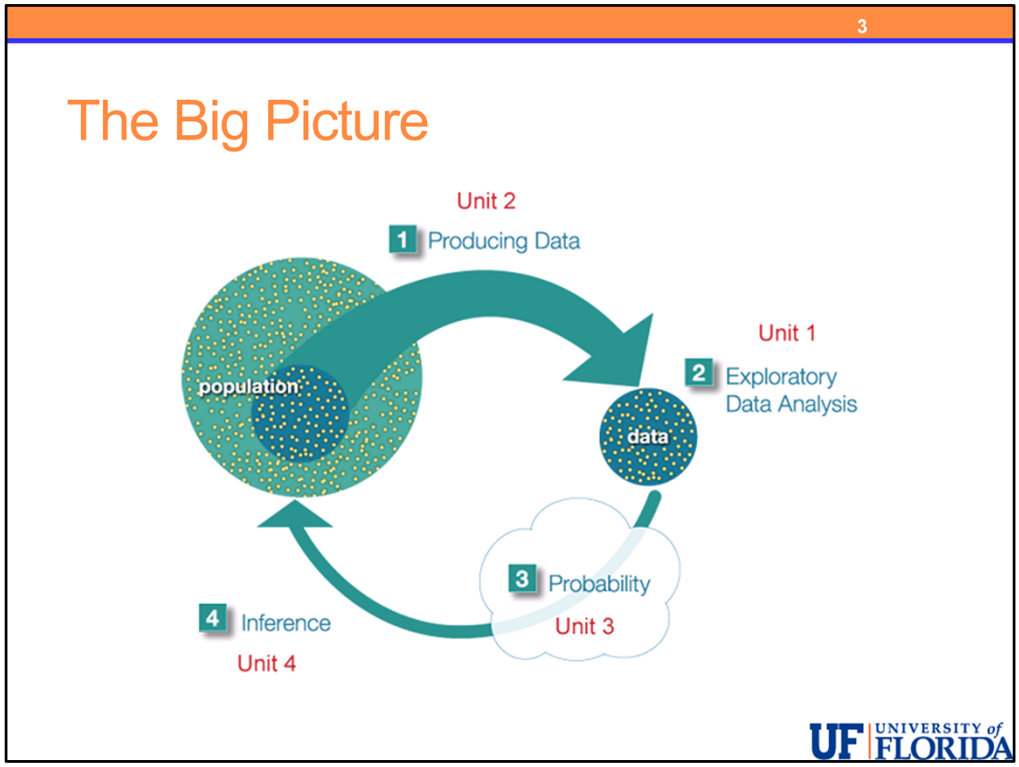


In this course we have a few major goals. We want you to

- Develop your understanding of probability and probability distributions. Including their application to statistics through:
    - The concept of the sampling distribution of our sample statistic.
    - and Real-world problems where we are interested in estimating certain probabilities in our population.

- Develop your understanding of the process of statistical inference using the relatively simplistic examples of one mean or one proportion – this is the big picture we have presented.

- Be able to correctly identify the main cases for research questions involving one or two variables. In each case you:
    - Know which exploratory and inferential methods to apply, including non-parametric alternatives.
    - Apply the appropriate standard method in software (or by-hand for the simplest problems).
    - Interpret the results correctly in the context of the problem.



We have now covered all of the concepts that make up the big picture. Let's review.

We are interested in learning something about a particular population.

In order to learn about the population, we take a random sample from the population.

From that sample, we produce our data (Step 1).

When we have our data, we conduct exploratory data analysis to obtain some STATISTIC from our sample. (Step 2)

In our methods we have seen statistics such as p-hat, x-bar, the difference between two sample means, the sample correlation coefficient (r), and the estimated slope, beta_1-hat.

In each of these cases, the probability "cloud" (Step 3) represents the process of learning about the BEHAVIOR of our statistic, in particular, we want to know the sampling distribution of the statistic and its associated standard deviation, which we call the standard error of the statistic.

Combining the value of the statistic from our data (our estimate) with information about the sampling distribution of the statistic, we can

- Construct, for example, 95% confidence intervals which, in repeated sampling, will

contain the true value in the population (our parameter) 95% of the time.

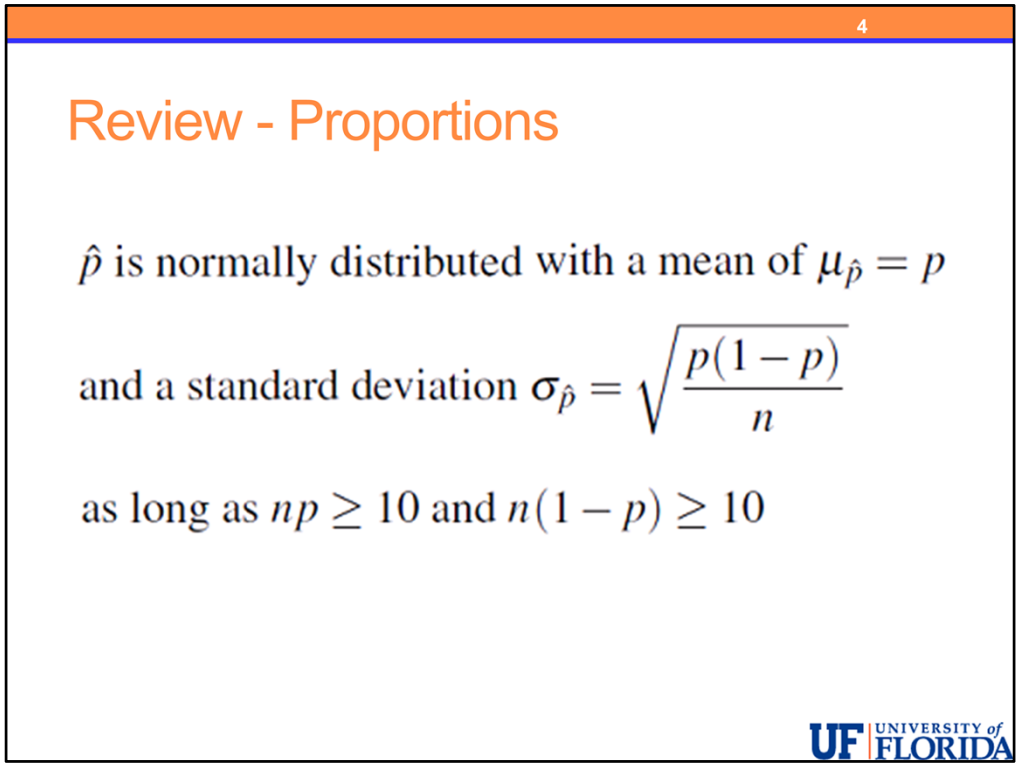
- Conduct hypothesis tests about our parameter using the data from our sample. In this case, we calculate the p-value which tells us the chance we could see a result such as ours or more extreme by random chance alone – in other words, assuming the null hypothesis is true, we find the probability that data such as ours or more extreme could have been produced.

    - When this probability is small, it would be very unlikely to obtain results such as ours or more extreme assuming the null hypothesis is true – and by inductive reasoning we say that there is evidence to reject the null hypothesis and conclude that the alternative hypothesis is actually true.

    - When this probability is large, it would not be unlikely to obtain results such as ours or more extreme assuming the null hypothesis is true – and thus we do not have enough evidence to reject the null hypothesis and we are unable to conclude the alternative is true. In this case, we have not proven the null hypothesis IS true we simply have not found any evidence to reject it.

For both confidence intervals and hypothesis tests, the standard error and hence the sampling distribution are key components.

Without information about the sampling distribution and standard error, we can't make inferences about the population of interest.



Before we begin our data analysis examples, let's review some details and concepts of sampling distributions and inference.

In the case of the sample proportion, we found that the distribution of all possible p-hats – the sampling distribution has a mean equal to the population proportion, p, and a standard deviation of the square root of p times (1-p) over n. This value is called the standard error of p-hat and measures the sampling variability of the estimator p-hat.

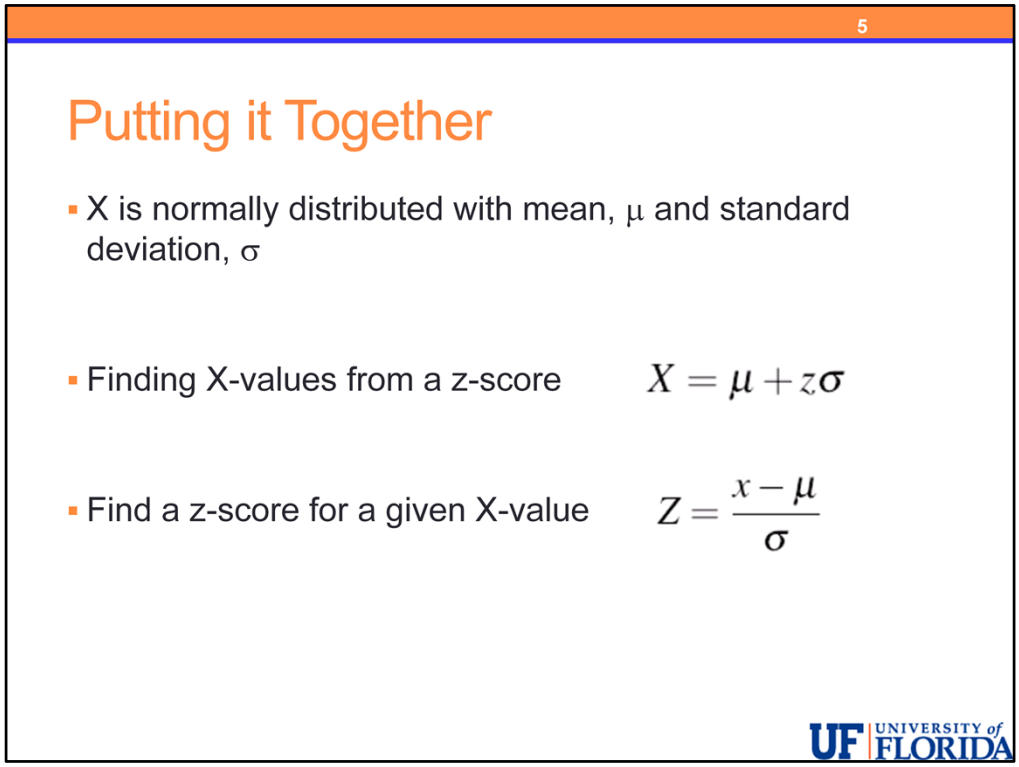
We also found that the sampling distribution of p-hat is approximately normally distributed as long as the sample size is large enough relative to the population proportion, p, specifically we need np and n(1-p) to be at least 10.

Knowledge of the sampling distribution and standard error are the basis of our ability to determine what range of values of p-hat are likely or unlikely which is the basis for constructing confidence intervals and conducting hypothesis tests.

Another very important idea is the difference between the parameter and a statistic. The parameter is the truth in the population whereas the value of our statistic is the estimate of the population parameter based upon our data.

Our inferential methods use the statistic from our single sample to estimate or test hypotheses about THE PARAMETER in the population.

It is crucial to realize that the results of any inferential method DO NOT apply to our sample – we know the EXACT results for our sample so there are no questions to answer about the sample itself, only about the population from which the sample was taken.



When we discussed normal probabilities and their applications, we presented these two equations.

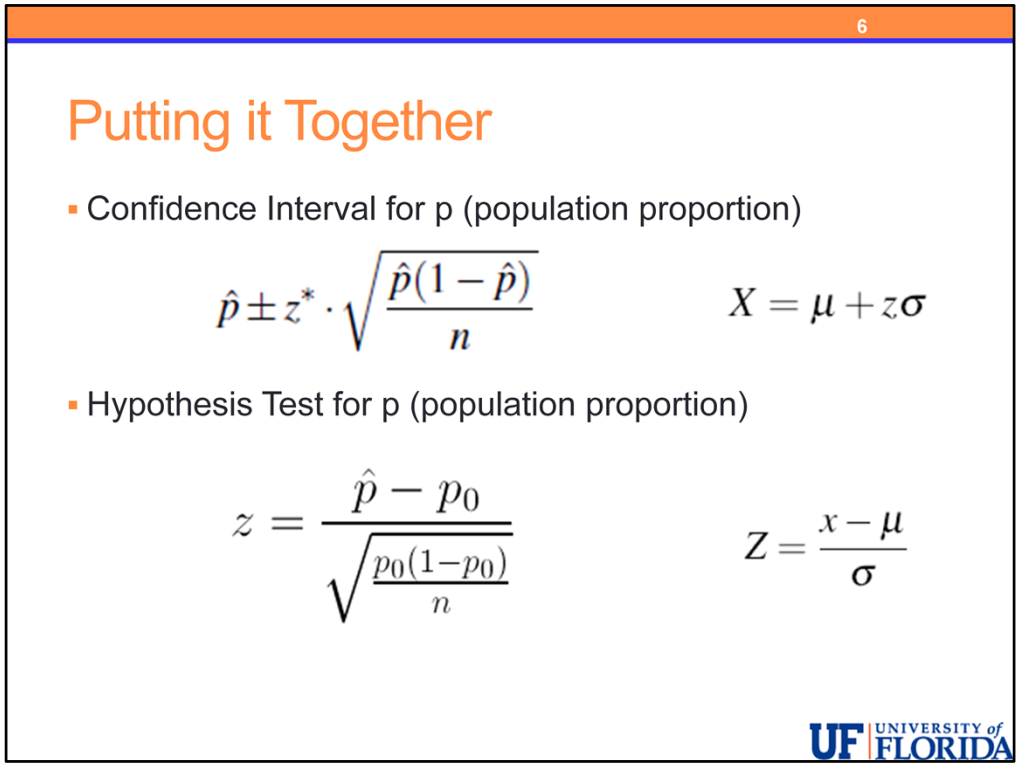
Here X represents a random variable which is normally distributed with mean mu and standard deviation sigma.

The z represents a z-score giving how many standard deviations is the value of X away from the mean mu.

The first equation allows us to convert from a known z-score to find the value of X as long as we know the mean and standard deviation of X.

The second equation calculates the z-score for a particular value of X.

Combining these equations with what we know about the sampling distribution of p-hat produces the equations we learned for confidence intervals and hypothesis tests.



The general equation x = mu + z(sigma) provides the basis for the construction of our confidence intervals, we simply need to substitute the standard deviation of our statistic in place of the generic standard deviation, sigma, in the general equation.

For example, for 95% confidence, the idea is that we know that we need to go 1.96 standard deviations on either side of our estimate from our data to be 95% confident that our resulting interval will capture the true population proportion. Knowing an estimate of the standard deviation of p-hat, we can determine the range of this interval.

If we didn't know the standard error or we did not know the distribution was normal – none of this would work!!

For hypothesis tests, we want to measure how many standard deviations away from the null value is the estimate from our data? We have seen z-scores a few times during the semester and they always have the same form: In the numerator we have the difference between "my value for the random variable" and the mean and in the denominator we have the standard deviation of the random variable under consideration.

In the section on normal random variables we had (x – mu) in the numerator & sigma in the denominator. It is important to understand the mu we subtract in the numerator and the sigma in the denominator are the mean and standard deviation of the random variable X.

Thus, for hypothesis tests about the population proportion we can find the z-score by

substituting

- p-hat in place of X (since our random variable is p-hat).

- p_zero, the null value, in place of mu (this is the assumed true population proportion which would be the mean of our random variable p-hat under our assumption that the null hypothesis is true).

- And the square root of p_zero times (1-p_zero) over n in place of sigma – since this is the standard deviation of our random variable p-hat under our assumption that the null hypothesis is true.

It is possible for someone to apply inferential methods throughout their career and not really understand these connections and for complex methods it becomes difficult to be able to put all of these pieces together without the required mathematical background .

However, hopefully you can see that material presented has been building the foundation for the development and understanding of these equations.

We learned how to summarize our data in exploratory data analysis – including introducing you to some of the needed ideas for normal probabilities with the discussion of the standard deviation rule.

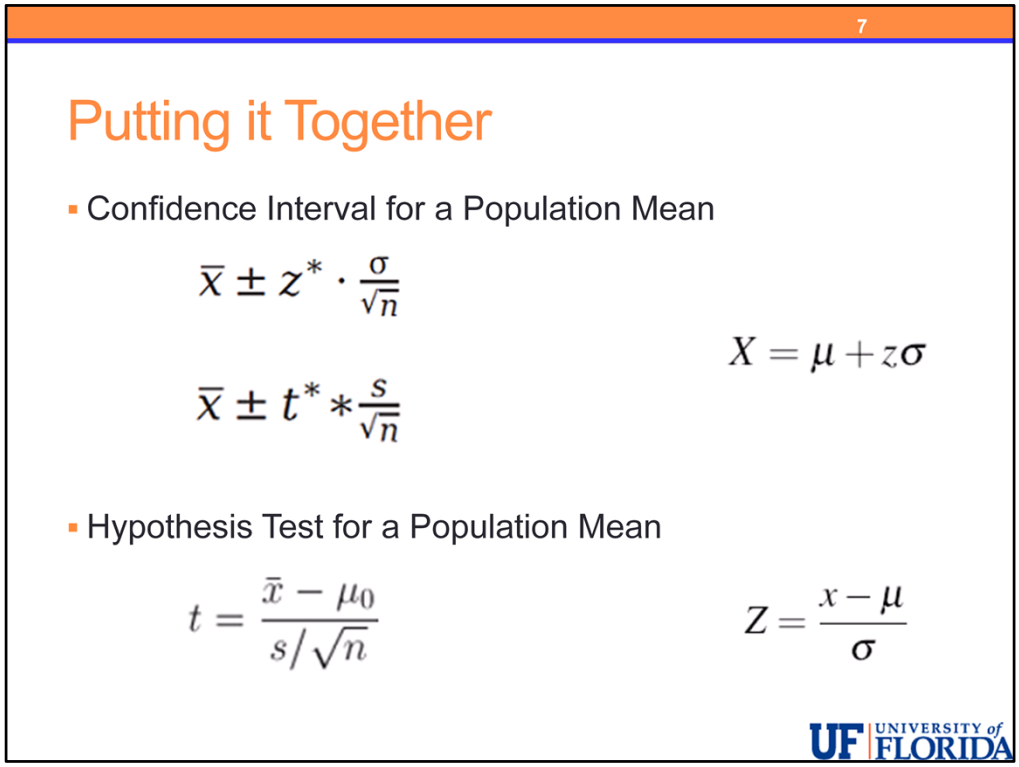
Then, after some discussion on sampling and design, we discussed random variables and normal probabilities so that we could develop the skills needed to find cut-offs for confidence intervals and p-values of hypothesis tests.

Then we discussed sampling distributions for x-bar and p-hat where we defined and verified the mean and standard deviation of these statistics and then tried to convince you that they are approximately normally distributed under certain conditions.

Once we know the sampling distribution is normal and we know the mean and standard deviation of that normal distribution, we can use this to find the cutoffs for confidence intervals and the p-values for hypothesis tests.

Notice that the normal distribution is only used for hypothesis tests once we calculate the p-value using our z-score which is our test statistic. Just because we “name” it z doesn’t make it normally distributed – a z-score will always measure the number of standard deviations away but if the original random variable is normal, we can take it the step further and determine probabilities associated with that z-score.

Not all confidence intervals and hypothesis tests use this “standardized score” approach but many do which makes this idea a fundamental concept in the development of a wide variety of statistical methods.



Similarly, for sample means we developed equations for confidence intervals and hypothesis tests.

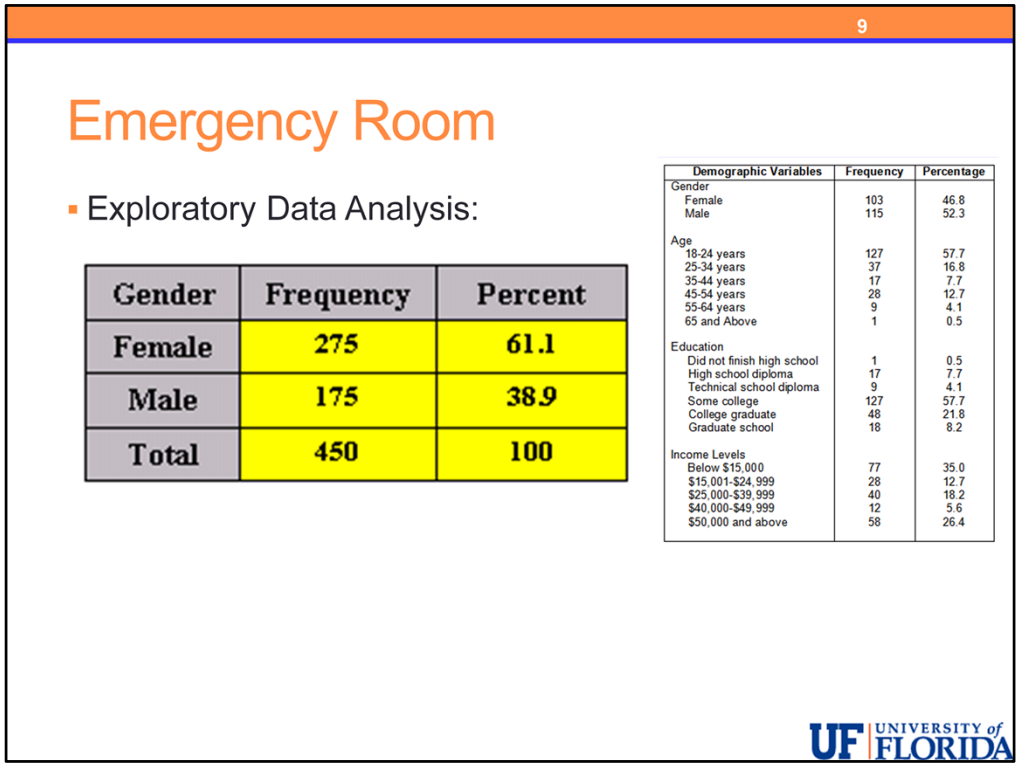
We learned in the section on estimation that if we do not know the true population standard deviation and instead substitute the sample standard deviation then the appropriate sampling distribution will be a t-distribution with n-1 degrees of freedom.

In practice we rarely know the true population standard deviation and thus we focused more on using software to provide the results for confidence intervals and hypothesis tests for one population mean and focused instead on correctly interpreting the results in context.

Although we are using a t-distribution instead of the normal distribution, the fundamental idea behind these methods still relies on the concept of the standard error of the sampling distribution and standardized values. We simply use the t-distribution instead of the normal distribution as the basis for determining what range of values are likely or unlikely.

Now we will give examples of each of the main cases covered in the course and an overview of exploratory and inferential methods.

We begin with a simple binary categorical variable: Gender in emergency room patients.



In a survey of a random sample of 450 emergency room patients at a certain hospital, 275 were female and 175 were male.

Our raw data consist of a list of 450 observations containing the gender of each patient. We can summarize this data numerically using a frequency distribution. This table could also be considered a visual display but we could also create a pie chart or bar chart if desired.

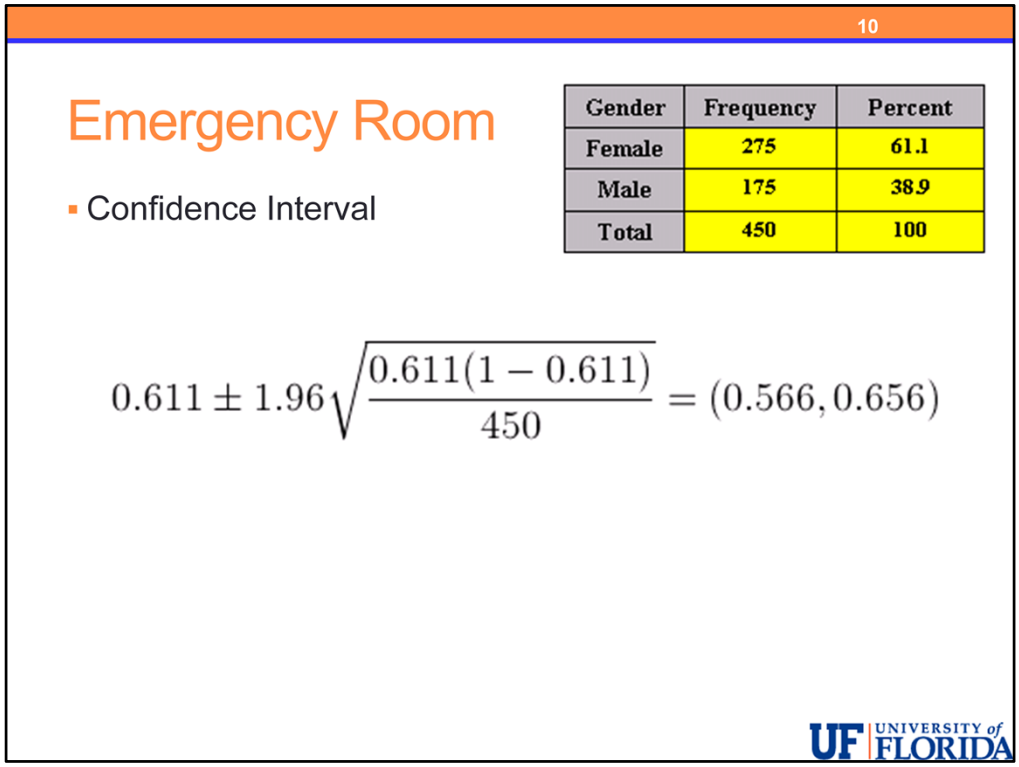
In practice, we would summarize the results presented in the frequency table in a short sentence or possibly in a large table containing this type of information for many variables in the study.

Here is an example of such a table from a different study.

Often, the purpose of this type of one variable analysis will be to give an overall descriptive summary of the patients in the sample. How well does it represent the population to which you want your results to apply. The closer your sample matches the population of interest the less limitations there are in your results.

Here we see that among the 450 patients surveyed 61.1% were female.

Suppose we wanted to determine if there is evidence that the true proportion of female patients in this ER is different from 50%.



When checking the sample size for confidence intervals we check if n(p-hat) and n(1-p-hat) are both at least 10.

Since both 450(0.611) and 450(1-0.611) are at least 10, we can construct a 95% confidence interval for the true proportion of females using the equation we presented.

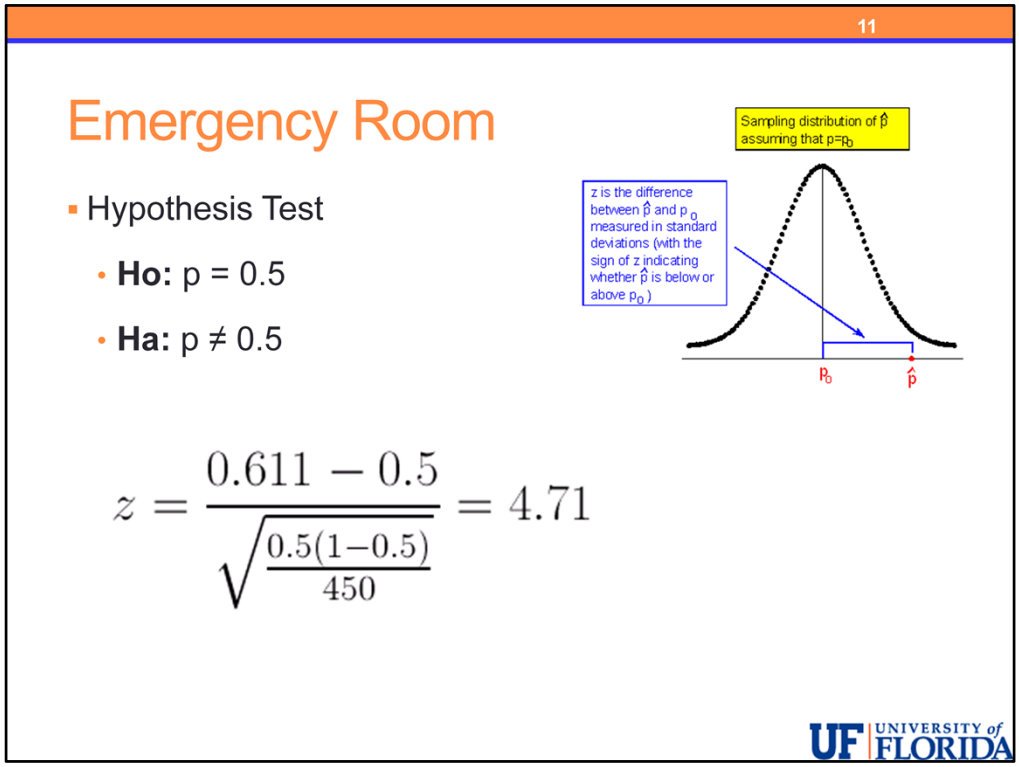
The appropriate confidence multiplier in this case is from a normal distribution due to the fact that for large enough samples, the distribution of all possible p-hats (the sampling distribution) will be approximately normally distributed.

This confidence interval consists of our estimate plus or minus our confidence multiplier, 1.96, times the estimated standard error of our statistic.

The resulting interval is 0.566 to 0.656.

Thus, we are 95% confident that between 56.6% and 65.6% of all ER patients at this hospital are female.

Based upon this confidence interval, since 50% is not a plausible value, we can conclude that the proportion of females in not 50% in this ER population. In fact, the confidence interval estimates the true proportion to be greater than 50%.



When checking the sample size for hypothesis tests we check if n(p-zero) and n(1-p-zero) are both at least 10.

Since 450(0.5) and 450(1-0.5) are at least 10, we can also answer the question using a hypothesis test with

Ho: p = 0.5
Ha: p ≠ 0.5

We calculate the test statistic as illustrated giving z = 4.71.

This test statistic, tells us that our p-hat is 4.71 standard errors above the hypothesized value. This is extremely unlikely for a normally distributed quantity.

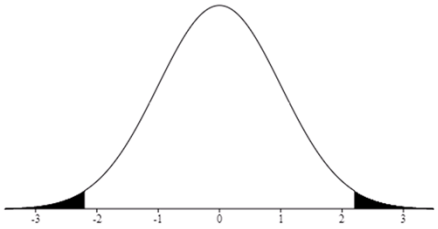
The p-value is basically zero for such a large z-score under the normal curve and thus there is enough evidence to conclude that the true proportion of female patients in this ER is not equal to 50%.

Remember that the p-value is the probability of obtaining a result as or more extreme than our data – in the direction (or directions) of our alternative hypothesis ASSUMING THE NULL HYPOTHESIS IS TRUE.

Emergency Room

- Suppose Sample Size =100, of which 61 are Female
- Hypothesis Test
  - **Ho:** p = 0.5
  - **Ha:** p ≠ 0.5
- P-value
  - = 2(0.0139)
  - = 0.0278

$$z = \frac{0.61 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 2.2$$



UF UNIVERSITY of FLORIDA

For a little better practice, let's suppose we had similar evidence from a smaller sample size.

What if we took a random sample of 100 where 61 patients are female.

Then the test statistic would be Z = 2.2.

Since our alternative hypothesis is two-sided, to find the p-value we need to calculate the area both above 2.2 and below -2.2.

Our p-value is 2(0.0139) = 0.0278 which is less than 0.05 and thus we reject the null hypothesis.

There is enough evidence to conclude that the true proportion of all patients in this ER who are female is not equal to 50%.

Notice that when we conduct hypothesis tests in practice they are usually two-sided. Most definitely the only one-sided tests which should be conducted are ones for which you know BEFORE you collect your data that you wish to prove ONLY one direction.

Sometimes this is the case – we want to prove our drug results in weight loss – or we want to prove our treatment increases red blood cell counts.

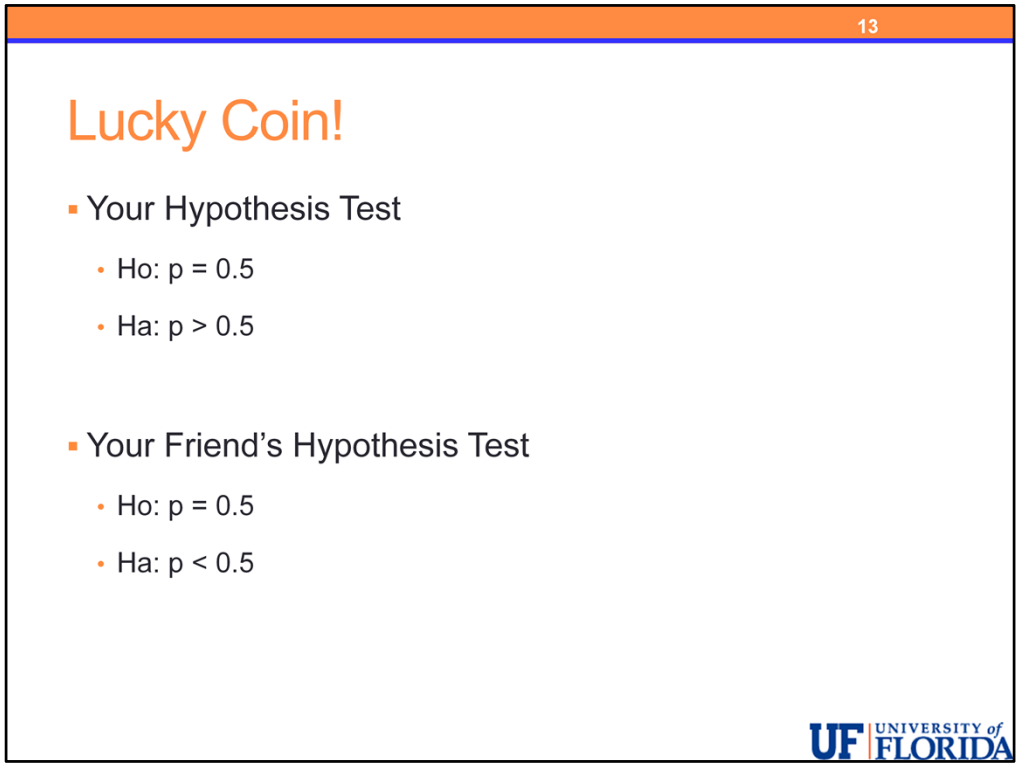
The main point is that you CANNOT change your hypotheses to a one-sided test AFTER you see your data. In this instance we cannot decide after seeing that 61% were female in our sample to change our desired alternative hypothesis to > instead of simply ≠.

Be sure to state your hypotheses based upon what is provided in the scenario NOT based upon the information you are provided about the sample.

Also notice that a disadvantage of conducting a one-sided test is that if it turns ours that the truth is the opposite of what you desire to show, your test will not be designed to discover that information and this is certainly one reason that the standard practice is to conduct two-sided tests followed by confidence intervals for the estimation of any effects of interest.

Before moving on, notice that since we rejected the null hypothesis we could have made a Type I error in this case. To describe this error in context we could say:

- It is possible that we could have concluded the true proportion of females is different from 0.5 when in fact it is equal to 0.5.



To review the p-value calculation for one-sided tests we will consider a simple example of tests about the fairness of a coin.

Suppose you have a lucky coin that you believe lands on heads more often than tails.

Then your hypotheses would be

Ho: p = 0.5
Ha: p > 0.5

Suppose your friend, who has seen you use this coin on numerous occasions, thinks you are crazy and if anything your so-called "lucky" coin lands on tails more often than heads!

Then your friend's hypotheses would be

Ho: p = 0.5
Ha: p < 0.5

Since it is your lucky coin, it is decided to allow you to flip it 100 times.

You do and you get 48 heads on 100 tosses.

Now, it is fairly clear that based upon the results of this sample, there is no evidence that p

> 0.5, in other words, we expect to find a very large p-value for that alternative hypothesis.

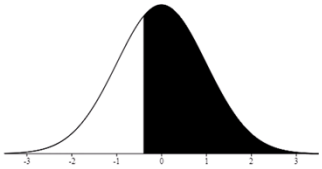
There is some anecdotal evidence to support your friend's claim that $p < 0.5$ but we will need to take into account the sampling variability in 100 tosses of a fair coin to assess if this is enough evidence to support your friend's claim.

Let's calculate the p-value for each test using this sample to illustrate the process.

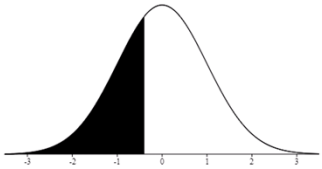
Lucky Coin!

$$z = \frac{0.48 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = -0.4$$

- Your Alternative: Ha: p > 0.5
  - P-value = 0.6554



- Your Friend's Alternative: Ha: p < 0.5
  - P-value = 0.3446



UF UNIVERSITY of FLORIDA

Since your alternative hypothesis was "greater than" – the p-value for your test finds the probability of obtaining a z-score such as that in the data or larger – as these are the values that are "as or more extreme" in the direction of the alternative.

Your p-value is the area to the right of -0.4 which is 0.6554.

Your friend's alternative hypothesis was "less than" – the p-value of this test finds the probability of obtaining a z-score such as that in the data or smaller – as these are the values that are "as or more extreme" in the direct of the alternative.

Your friend's p-value is the area to the left of -0.4 which is 0.3446.

In both cases, there is not enough evidence to reject the null hypothesis. We didn't need to know the p-value to know this was the case for your test but for your friends, we needed to know how rare this value was in order to determine if there was evidence to support the claim that $p < 0.5$.

If you and your friend had decided prior to collecting your data to simply test the two-sided alternative, the p-value for this test would have been 2(0.3446)

It could be that the coin really is fair! But … then again maybe not.

We aren't able to prove the null hypotheses we only know this data does not give us

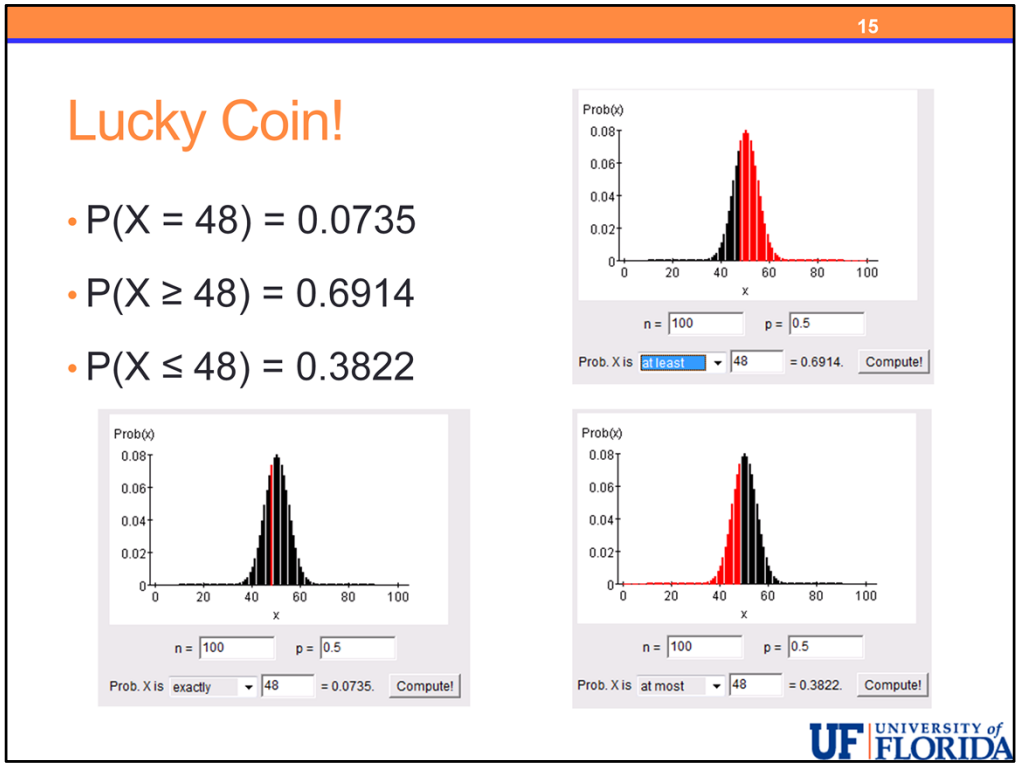
evidence to reject it … in either direction.

In this case, since we failed to reject the null hypothesis, it could be that we have made a Type II error. In context we could say:

For your test: It is possible that we concluded the true proportion of heads on this coin is not greater than 50% when in fact it is! (this would make you happy)

For your friend's test: It is possible that we concluded the true proportion of heads on this coin is not less than 50% when in fact it is.

Two-sided test: It is possible that we concluded the true proportion of heads on this coin is not different from 50% when in fact it is different from 50%.



Let's take this opportunity to review the binomial distribution and use our ability to calculate binomial probabilities to look at the "what if's" of this situation.

We will illustrate the calculations under the null hypothesis visually and then present a full set of results for discussion.

We will go back to using a binomial distribution to calculate probabilities using n = 100 and p = our current guess at the truth.

Under the null hypothesis, we assume p = 0.5. Now we calculate three probabilities

- P(X = 48)
- P(X ≥ 48)
- P(X ≤ 48)

In this particular case where we assume p = 0.5, the last two probabilities are the exact p-values of your test and your friend's test respectively. We applied the exact distribution instead of approximating it by an appropriate normal distribution.

Here we find that if the coin was exactly fair, there is a 7.4% chance we could obtain 48 heads out of 100 tosses. There would be a 38.2% chance of getting 48 heads or less and a 69.1% chance of getting 48 heads or more.

Notice these last two probabilities do not add to 100% as they both contain P(X = 48).

Now we will start assuming other values for the truth which are not the null hypothesis. The question is, how large or small does the true proportion have to be before it becomes very unlikely that this would happen.

This may help you see how all of this fits together and help you understand a little more about how we calculate type II errors and power.

Although the probabilities we will calculate are not directly related to either type II error or power, they will be calculated through a similar process – by assuming a value for the truth and then calculating probabilities based upon that assumption.

Truth	P(X = 48)	P(X ≥ 48)	P(X ≤ 48)
0.45	0.066	0.307	0.760
0.46	0.074	0.381	0.693
0.47	0.078	0.459	0.619
0.48	0.080	0.539	0.540
0.49	0.078	0.618	0.460
0.5	**0.074**	**0.691**	**0.382**
0.51	0.067	0.758	0.308
0.52	0.058	0.816	0.242
0.53	0.048	0.865	0.184
0.54	0.039	0.904	0.135
0.55	0.030	0.934	0.096
0.56	0.022	0.956	0.066
0.57	0.016	0.972	0.044
0.58	0.011	0.983	0.028
0.59	0.007	0.990	0.017
0.6	0.004	0.994	0.010
0.61	0.003	0.997	0.006

UF UNIVERSITY of FLORIDA

I didn't use the applet for these calculations as it would have taken quite a while. I used an EXCEL formula BINOMDIST and the ability to "fill down" to obtain this table very quickly.

The row in bold represents the probabilities we calculated on the previous slide where we assumed the null hypothesis is true.

If we look at the P(X = 48), we see that it is the largest when the coin's true percentage is 0.48, as we would expect. It is fairly likely to happen if the true value was 45% through about 52 or 53% but it does not become extremely rare as an individual outcome until we get to true values approaching 60%.

For example, if the coin were 60% heads, there would only be a 0.4% chance we could ever see 48 heads in 100 tosses of the coin.

Considering it from your perspective – it is your lucky coin after all, in order to really investigate how rare this would be, we should consider the P(X ≤ 48).

Looking in that column, we see that if the true probability of heads is 55%, overall there is still a 13.5% chance of getting 48 heads or lower in 100 tosses. Not very rare!

If the true probability is 57% that probability drops to 0.044 which is somewhat rare.

When the true probability is 60% there is a 1% chance of getting 48 heads or less in 100

tosses.

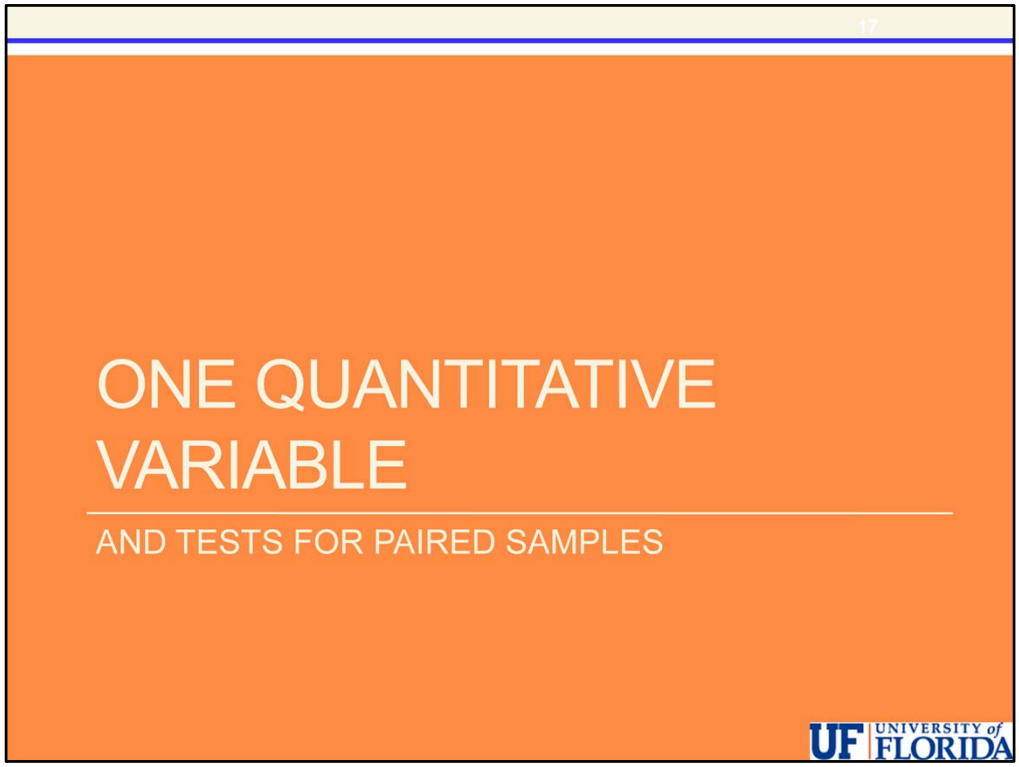
So, since we found a sample with X = 48 for this coin, we can see from this table that there are numerous values of p for which we could easily have seen our result.

Besides being a quick review of the binomial distribution, the final conclusion of this example is:

Just because we fail to reject the null hypothesis, doesn't mean the null hypothesis is true. In this case with 48 heads in 100 tosses, p could easily be 0.55 based upon the results in this table.

In fact the 95% confidence interval would range from 0.382 to 0.578. Which says that any value between 0.382 and 0.578 is a plausible value for the true probability of heads on this coin.

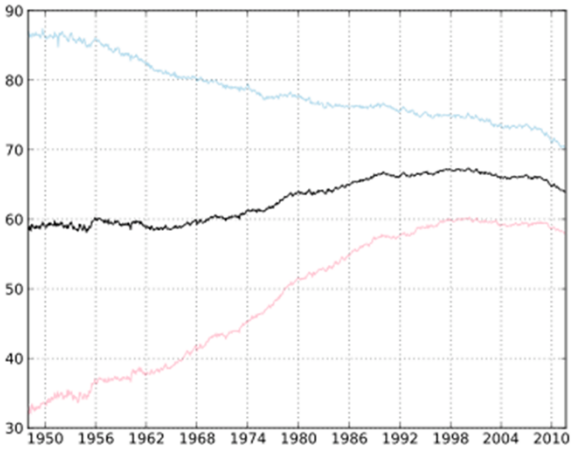
You could still be correct about your lucky coin – but so could your friend! More data would be needed to settle this argument.



Now let's review methods for one quantitative variable.

We will use an example which will result in a paired t-test regarding the labor force participation rate for women in the 60's and 70's.

Labor Force Participation Rate



UF UNIVERSITY of FLORIDA

The labor force participation rate (LFPR) is the proportion of individuals in a particular population who are currently working.

This graph (from http://en.wikipedia.org/wiki/Labor_force) shows:

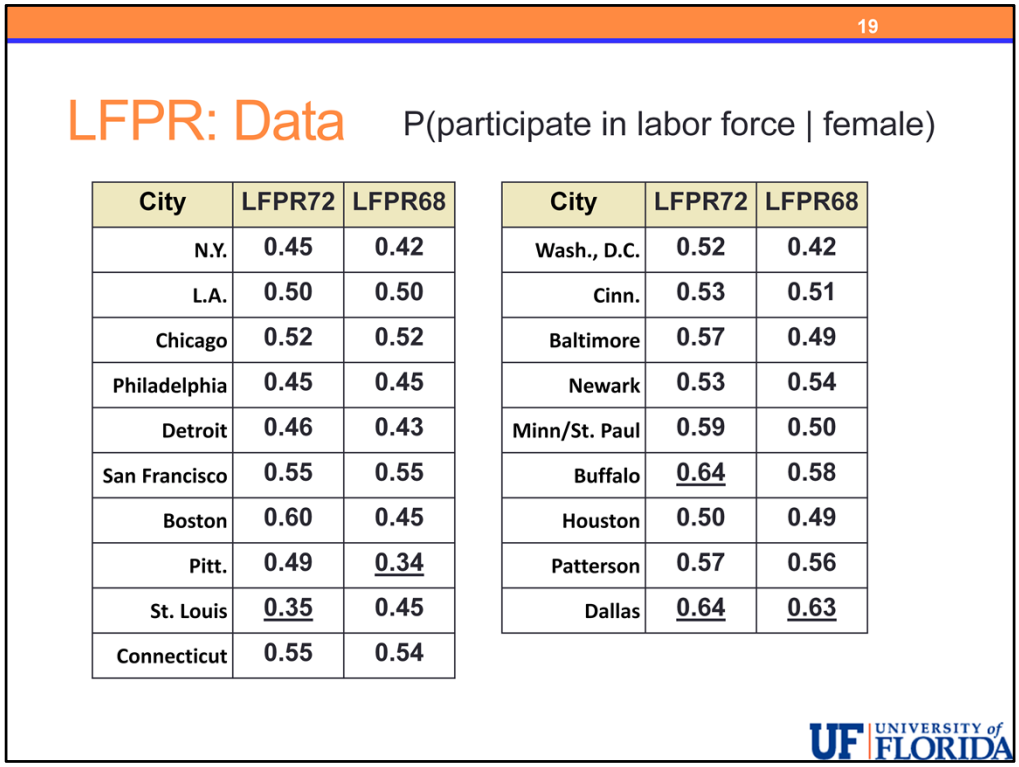
In the center as a black line – the labor force participation rate for all US adults.

On the top – as a light blue line – the labor force participation rate for US men.

On the bottom – as a pink line – the labor force participation rate for US women.

There was a clear increasing trend among women and a decreasing trend among men over most of the period.

We are going to investigate data containing the labor force participation rate for women in 19 cities for two years, 1968 and 1972.



Here is the raw data which comes from the data and story library. (http://lib.stat.cmu.edu/DASL/Stories/WomenintheLaborForce.html)

Each pair represents the results for the listed city with the first value from 1972 and the second from 1968.

We will begin by summarizing the results in 1968 and 1972 individually.

In the data, the values corresponding to the min and max for each year are underlined.

You can see that the minimum in each year came from a different city. The maximum value in 1972 occurred twice and one of these cities – Dallas – was the maximum in both years.

Although not completely crucial to this problem. I will point out that these particular measures – the labor force participation rates – are actually estimates of a probability. In this case, we have a conditional probability. The labor force participation rate among women can be restated as the probability of a person participating in the labor force given the person is female.

We could write this in symbols as P(participate in labor force | female)

I point this out mainly to illustrate that there are many applications of the concepts of probability that we discussed hidden in real-world problems in a wide variety of disciplines.

Labor Force Participation Rate

Variable	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
LFPR72	0.350	0.490	0.530	0.570	0.640
LFPR68	0.340	0.450	0.500	0.540	0.630

Variable	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
LFPR72	0.527	0.071	0.493	0.561
LFPR68	0.493	0.068	0.460	0.526

UF UNIVERSITY of FLORIDA

Here we have the SAS output summarizing the labor force participation rate for these two years.

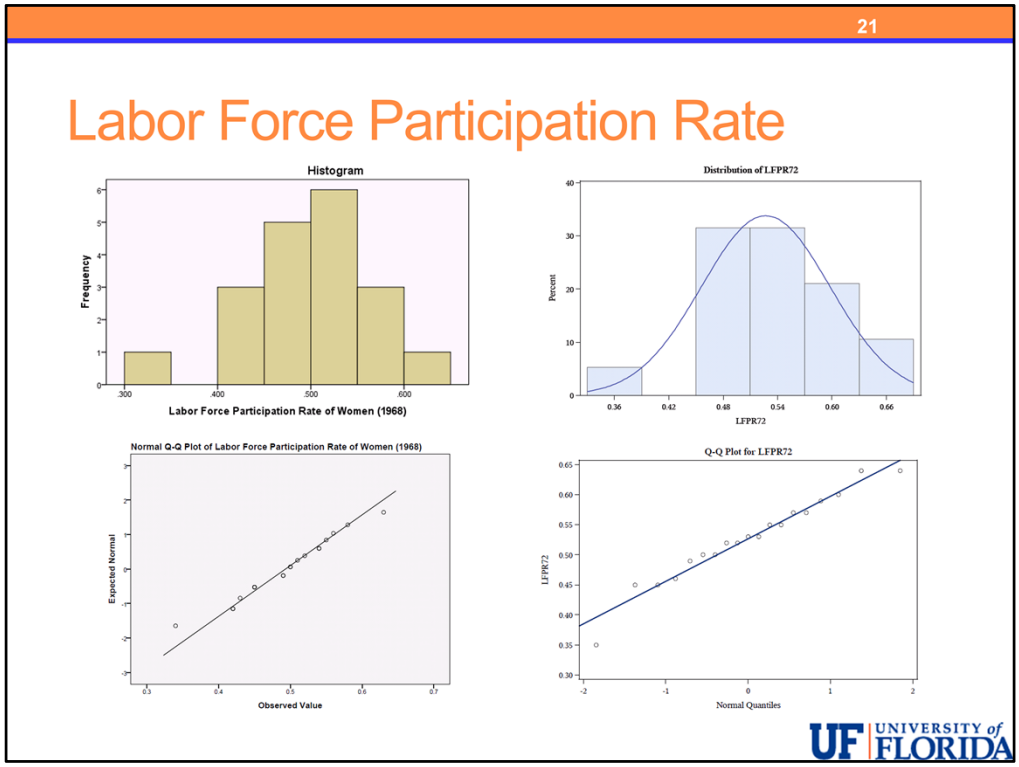
We can see that in 1972 all of the values for the 5-numbary summary are larger in 1972 than 1968 indicating an overall increase in this measure between these two years.

The mean LFPR in 1972 was 0.527 whereas in 1968 it was 0.493. Soon we will determine if this difference is statistically significant.

The variation within each year is similar as the range is exactly the same in both years, the standard deviation is similar (0.071 in 1972 and 0.068 in 1968), and the IQR is also similar (0.57 – 0.49 = 0.08 in 1972 and 0.54 – 0.45 = 0.09 in 1968).

When we calculate 95% confidence intervals for each year, the intervals have significant overlap. For independent samples, unless the overlap is small, this usually indicates that the difference will not be statistically significant in the corresponding t-test.

However, in the case of paired samples, such as in this data, we cannot base our conclusion on results which assume independent samples. We are interested specifically in the trend within each city. Overall, is there a change in the LFPR values?



Before conducting the paired t-test, let's look at a few additional types exploratory data analysis.

On the left we have the SPSS results for a histogram and normal QQ-plot of the LFPR values for 1968.

On the right we have the same results using SAS for 1972.

For both years, there seems to be one low outlier but in general the distributions are reasonably normally distributed.

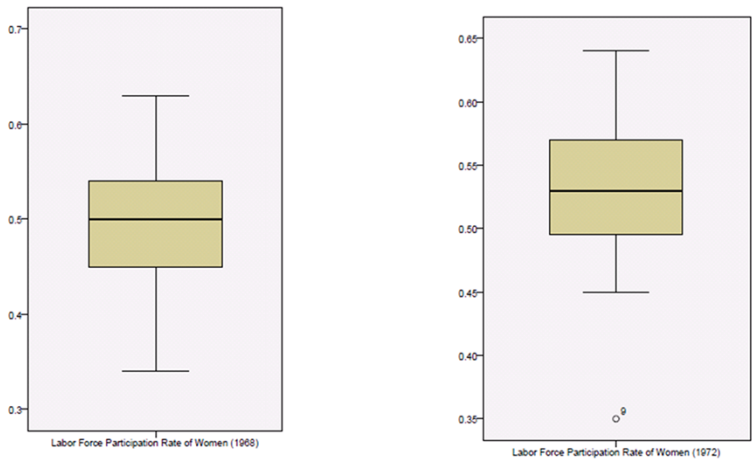
The histograms, give us a good overall picture of how the values of the variables are distributed for this sample of cities.

The normal QQ-plots are mostly used to investigate the validity of normality assumptions required by inferential methods you wish to use. For the moment, we aren't particularly concerned with the normality but we can see that both are approximately normal.

For the moment, it is difficult to make any comparisons based upon the histograms, especially when taken from two different packages.

In addition, in order to answer our question, we need to consider the pairing.

Labor Force Participation Rate



UF UNIVERSITY of FLORIDA

In the boxplots, from SPSS, we can see that the minimum in 1968 (which was Pittsburg with 0.34) is not considered an outlier whereas in 1978, the minimum is an outlier (this was St. Louis with 0.35)

LFPR: Data

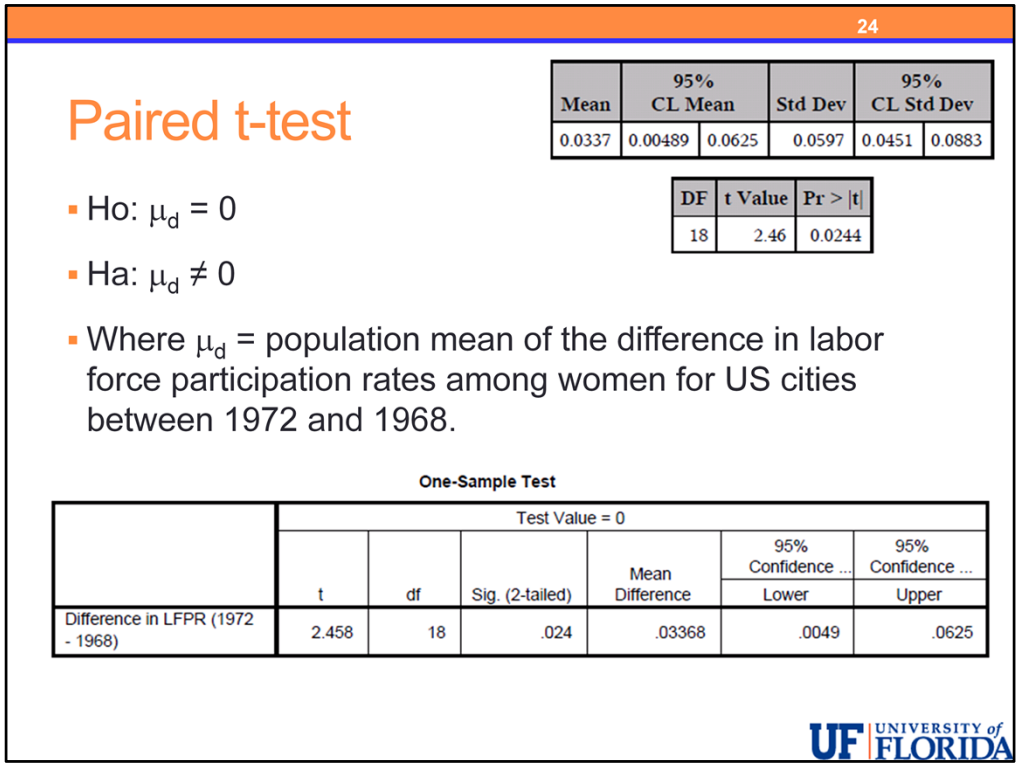
City	LFPR72	LFPR68	Diff
N.Y.	0.45	0.42	**0.03**
L.A.	0.50	0.50	**0**
Chicago	0.52	0.52	**0**
Philadelphia	0.45	0.45	**0**
Detroit	0.46	0.43	**0.03**
San Francisco	0.55	0.55	**0.00**
Boston	0.60	0.45	**0.15**
Pitt.	0.49	0.34	**0.15**
St. Louis	0.35	0.45	**-0.1**
Connecticut	0.55	0.54	**0.01**

City	LFPR72	LFPR68	Diff
Wash., D.C.	0.52	0.42	**0.10**
Cinn.	0.53	0.51	**0.02**
Baltimore	0.57	0.49	**0.08**
Newark	0.53	0.54	**-0.01**
Minn/St. Paul	0.59	0.50	**0.09**
Buffalo	0.64	0.58	**0.06**
Houston	0.50	0.49	**0.01**
Patterson	0.57	0.56	**0.01**
Dallas	0.64	0.63	**0.01**

UF UNIVERSITY of FLORIDA

To begin our paired analysis, we can calculate the differences for each city.

You can see that most values are positive with a few negative and some with no measurable change.



Although researchers may be interested in showing an increase, in keeping with the most common analysis in practice we will conduct a two-sided test.

Our hypotheses will be that mu-sub-d = 0 for the null hypothesis and mu-sub-d ≠ 0 for the alternative.

It is important to understand exactly what the symbolic parameters in your hypotheses mean in context.

In this case, we can define mu-sub-d to be the population mean of the difference in labor force participation rates among women for US cities between 1972 and 1968.

We have a small sample (n = 19) so we will need to investigate the normality assumption.

We will be using the sample standard deviation as we do not know the population standard deviation of the differences under study – thus we will be conducting a t-test.

In particular, we are conducting the paired t-test, however, this is the same process as a one-sample t-test except that usually the mean difference specified in the null hypothesis for a paired t-test is zero whereas for a one-sample t-test the null value is not usually zero.

Partial output for both packages is provided. We find

- A test statistic of t = 2.46 from SAS and 2.458 from SPSS
- A p-value of 0.0244 from SAS and 0.024 from SPSS
- The degrees of freedom are stated to be 18 indicating that n is 19 (as it should be)

Since the p-value is less than 0.05, we can reject the null hypothesis. We can say:

There is enough evidence to conclude that the population mean of the difference in *labor force participation rates among women* for US cities between 1972 and 1968 is not zero.

Although that interpretation is completely accurate, it may be re-worded for easier understanding as:

There was a statistically significant change in the population mean *labor force participation rate among women* for US cities between 1968 and 1972.

Now let's investigate the change:

- The estimated mean difference is 0.0337 from SAS and 0.03368 from SPSS
- The 95% confidence interval rounded to three decimal places is (0.005, 0.0625)

We can interpret this by saying:

Based upon our data, we estimate that the population mean *labor force participation rate among women* for US cities increased by 0.034 between 1968 and 1972. The 95% confidence interval suggests this value could be as low as 0.005 to as high as 0.0625.

Or we could simply say:

We are 95% confident that the population mean *labor force participation rate among women* for US cities increased by between 0.005 to 0.0625 from 1968 to 1972.

Remember that both our confidence intervals and hypothesis tests are about the population NOT our current sample – we know exactly what happened in our sample.

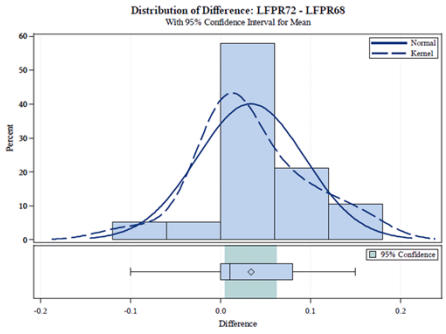
As always, it is possible that we have made an error.

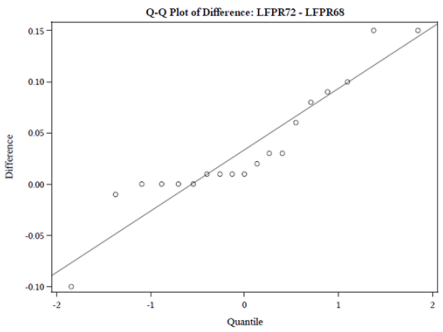
For this hypothesis test, we could have claimed there was a change in the population mean labor force participation rate when in fact there was not which would be a Type I error.

And for the confidence interval, similarly, it is possible that it does not contain the true value.

We know that if we repeated this process, 95% of the time, the interval we obtain from this process would capture the target value but we cannot know if our current interval from 0.005 to 0.065 contains the true mean or not!

Labor Force Participation Rate





Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Labor Force Participation Rate of Women (1972) and Labor Force Participation Rate of Women (1968) equals 0.	Related-Samples Sign Test	.007[1]	Reject the null hypothesis.
2	The median of differences between Labor Force Participation Rate of Women (1972) and Labor Force Participation Rate of Women (1968) equals 0.	Related-Samples Wilcoxon Signed Rank Test	.011	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1]Exact significance is displayed for this test.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.457704	Pr > |t|	0.0244
Sign	M	5.5	Pr >= |M|	0.0074
Signed Rank	S	46	Pr >= |S|	0.0062

UF UNIVERSITY of FLORIDA

SAS does a better job than SPSS of providing details that help to validate the assumptions. In SPSS you would need to analyze the differences yourself where in SAS we obtain these graphs automatically upon conducting a paired t-test.

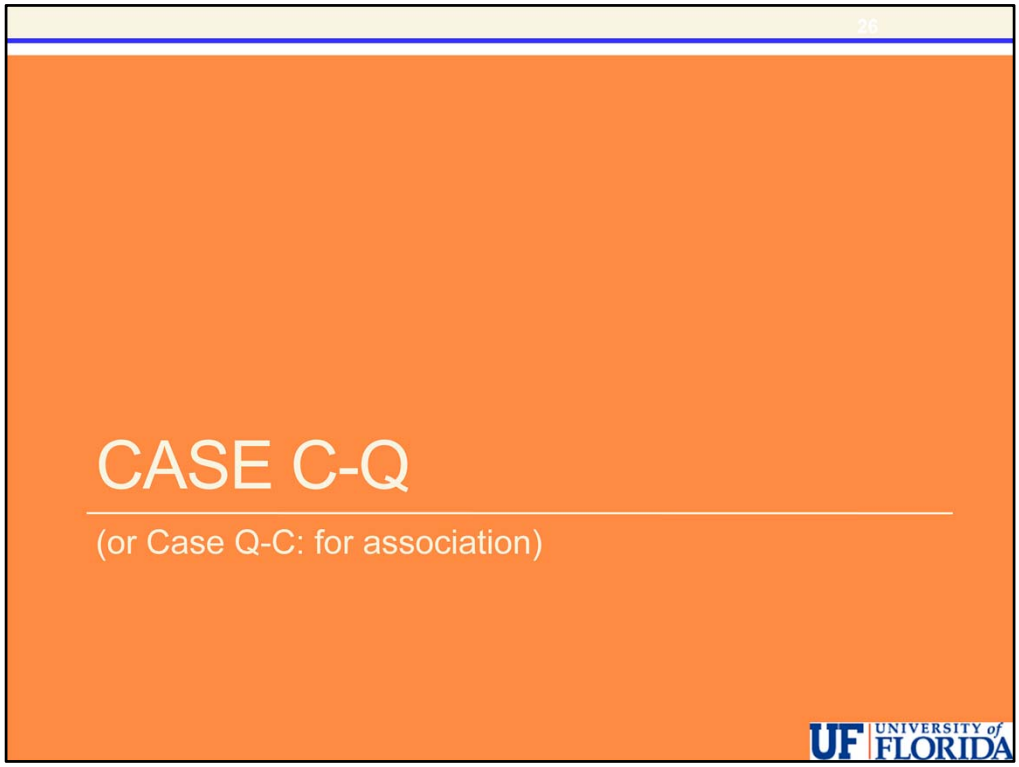
We can see that the distribution of the differences is somewhat not normal, however for such a small sample size, this would not be unexpected coming from a normal population.

It would be reasonable to apply the paired t-test.

If you are concerned, you could also apply the sign test and the Wilcoxon signed-rank test.

Both were applied in SAS and SPSS and both are statistically significant lending support to our conclusion of a statistically significant change.

You might notice that the p-value for the signed-rank test is different in SAS and SPSS. SAS uses the exact p-value where SPSS uses an asymptotic approach – which for small sample sizes may not be very accurate. There may be a way to find the exact p-value in SPSS but it wasn't a direct option that I could find.

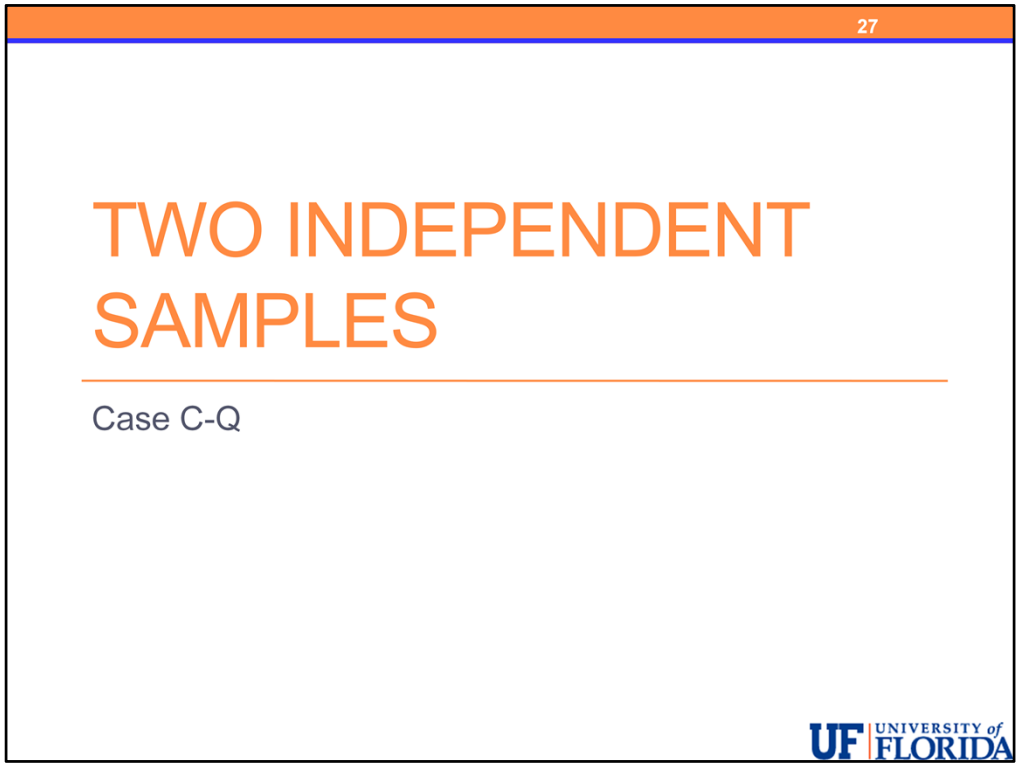


In Case C-Q we covered three main scenarios.

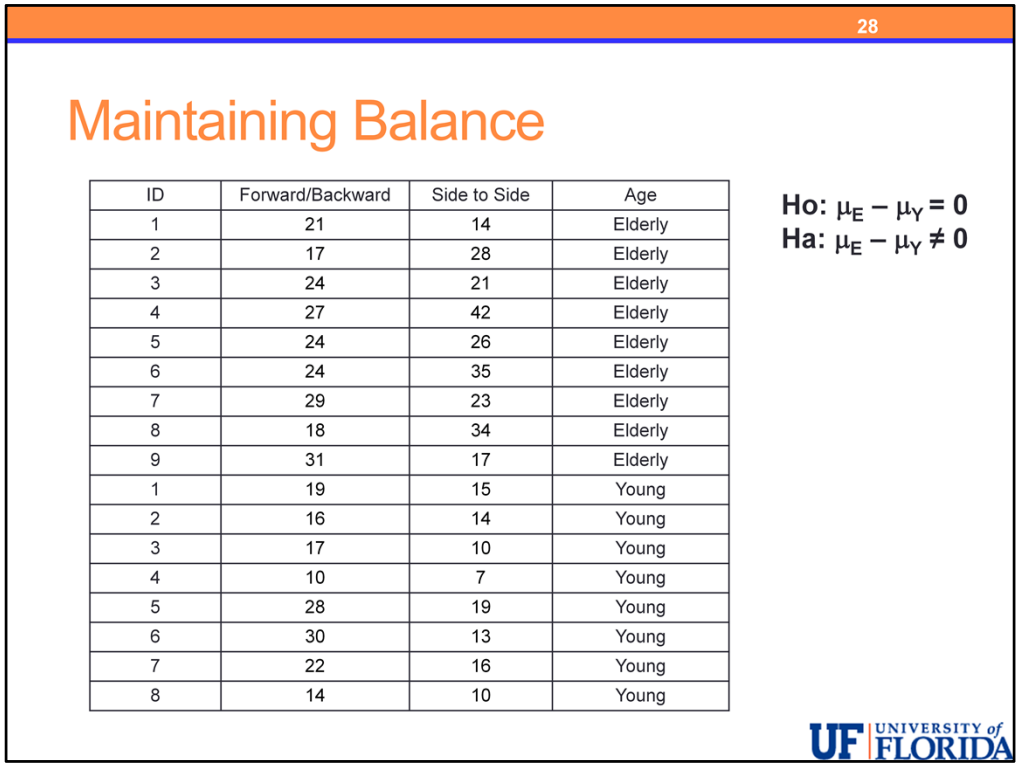
The first is the paired t-test which we have already reviewed.

The remaining methods are for two independent samples or for more than two independent samples.

These methods for independent samples can also be used in Case Q-C to show an association between the two variables but they will not allow us to predict a categorical outcome from a quantitative predictor as may be desired in Case Q-C.



We will be begin with an example comparing two groups.



Data Reference: http://lib.stat.cmu.edu/DASL/Stories/MaintainingBalance.html

Is age related to the ability maintain balance while concentrating? The data comes from the data and story library. The data we will use was simulated to be similar to, but without some of the problems of, the original data.

Nine elderly and eight young subjects participated in this experiment.

Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position and to react as quickly as possible to an unpredictable noise by pressing a hand held button. The noise came randomly and the subject concentrated on reacting as quickly as possible.

The platform automatically measured how much each subject swayed in millimeters in both the forward/backward and the side-to-side directions.

These are two independent samples but we also have two different response variables to analyze:

- Forward to Backward Sway Range and
- Side to Side Sway Range

In each case our null hypotheses will be that the difference in the population mean sway range between elderly and young is zero and our alternative will be that this difference will

not be equal to zero.

SPSS Summaries for Both Sway Ranges

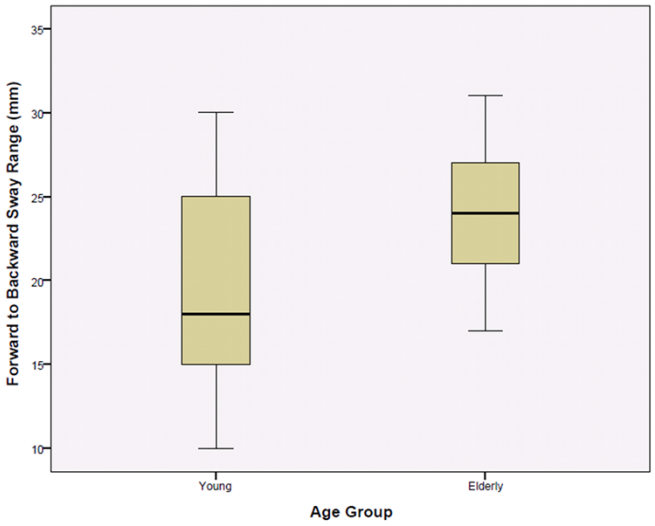
	Age Group	N	Mean	Std. Deviation	Std. Error Mean
Forward to Backward Sway Range (mm)	Elderly	9	23.89	4.702	1.567
	Young	8	19.50	6.845	2.420
Side to Side Sway Range (mm)	Elderly	9	26.67	9.083	3.028
	Young	8	13.00	3.854	1.363

UF UNIVERSITY of FLORIDA

Here are the summaries produced by SPSS for both sway ranges.

Notice in SPSS the output lists elderly and then young. This indicates that the SPSS output that follows will be estimating mu-sub_Elderly minus mu-sub-Young.

Forward to Backward Sway Range

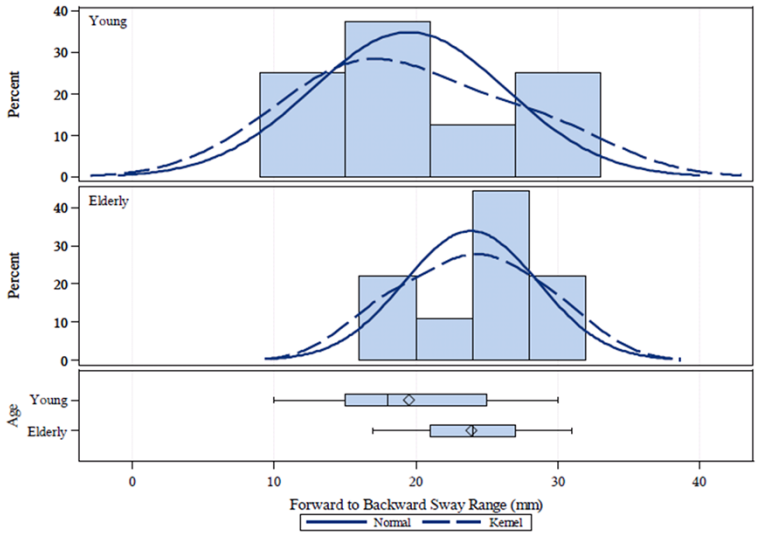


UF UNIVERSITY of FLORIDA

Now, boxplots for forward to backward sway range from SPSS. There seems to be some difference in the variation but, as this is a very small sample size, possibly this could be due to chance.

It does seem that the mean and median forward to backward sway range for elderly individuals is larger than that for young individuals but again, the sample size is small.

Forward to Backward Sway Range

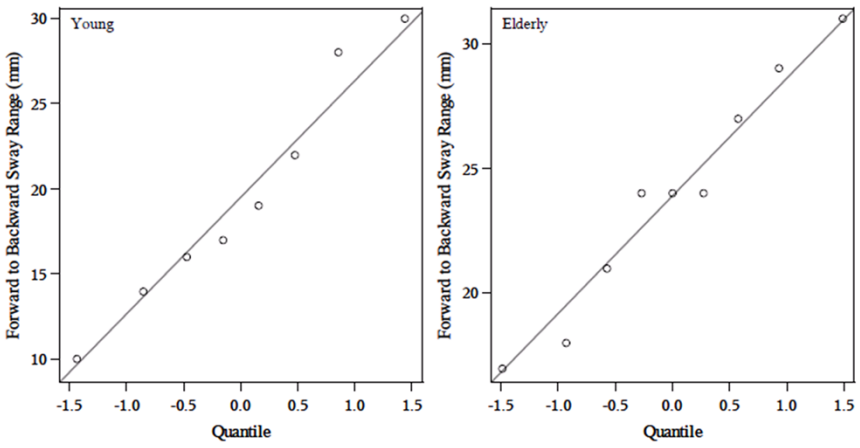


UF UNIVERSITY of FLORIDA

Here is the output from SAS when conducting the two-sample t-test.

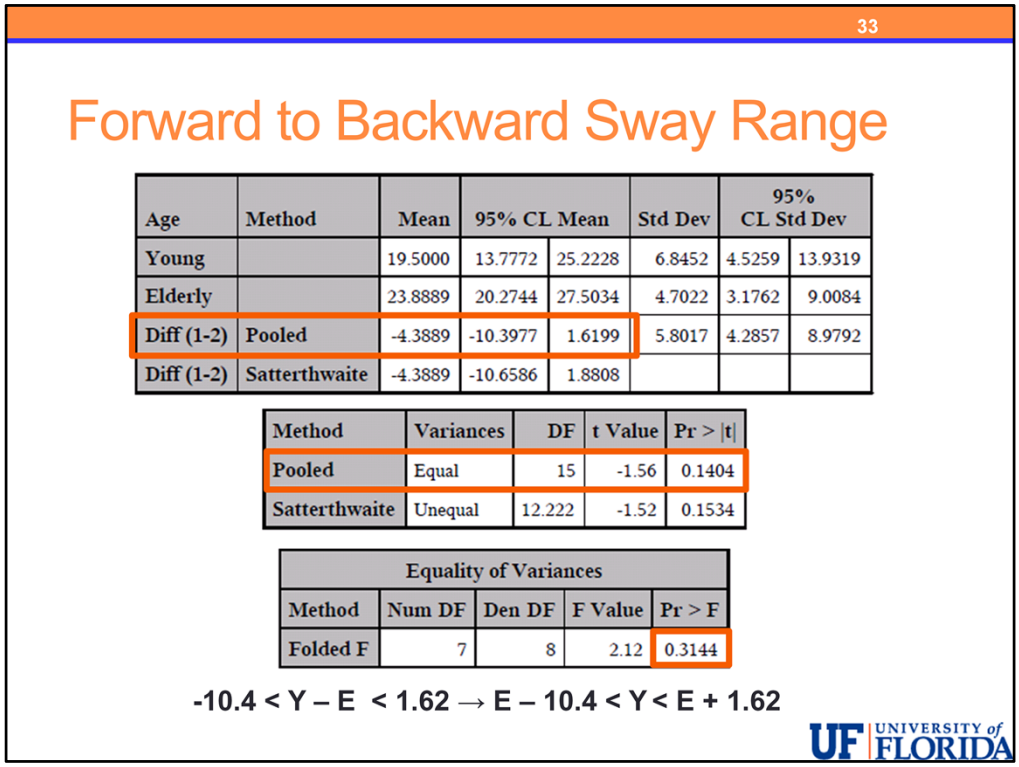
Both distributions seem reasonably normal comparing the densities (solid vs. dotted line) on these histograms. We also see the boxplots again, in this case horizontally, under the histograms. There are no outliers in the data.

Forward to Backward Sway Range



UF UNIVERSITY of FLORIDA

These are the QQ-plots from SAS which also show no reason for concern regarding the normality assumption.



Now we move into the output for the Two-sample T-test on forward to backward sway range between young and elderly patients.

Notice in SAS the output lists young and then elderly. This indicates that the SAS output will be estimating mu-sub_Young minus mu-sub-Elderly.

For SAS we begin by looking for the p-value of the test for equality of variances, which is 0.3144, outlined in the lower right of this output.

Thus we fail to reject the null hypothesis that the variances are equal and so we can use the equal variances row in the tables, also outlined.

We find a p-value for the equal variances two sample t-test of 0.1404 and so there is not enough evidence to conclude that the population mean ***forward to backward sway range*** differs between young and elderly individuals.

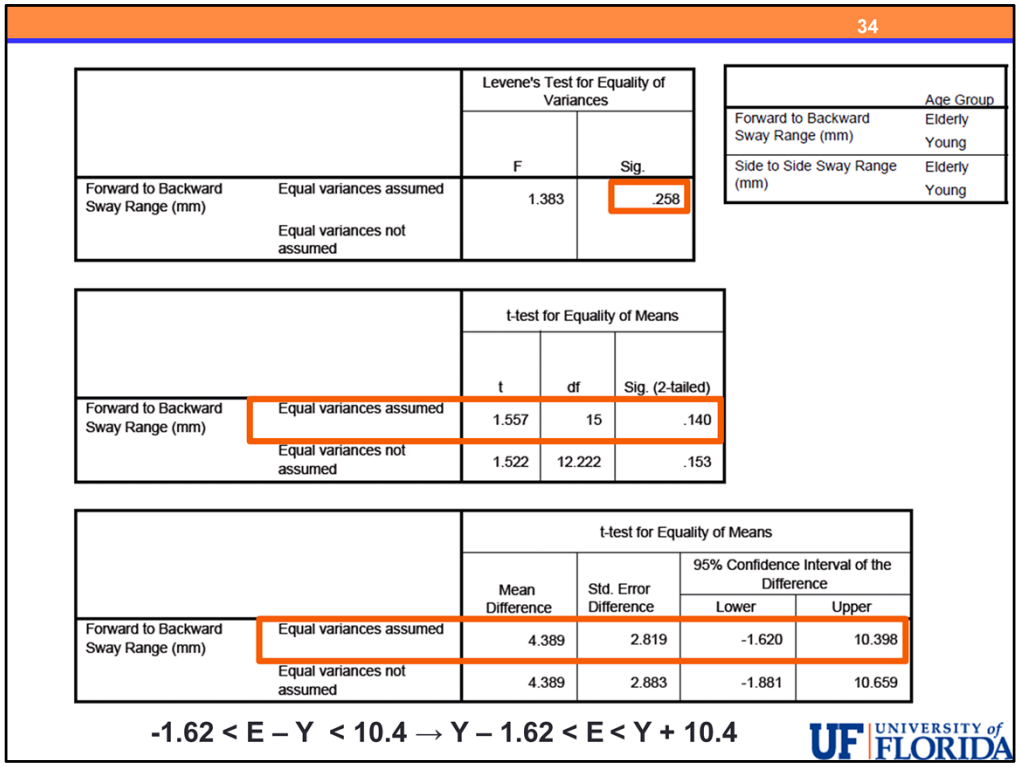
The appropriate 95% confidence interval for the difference between the population mean for young and that for elderly is given as -10.4 to 1.62.

We can interpret our estimate and confidence interval as follows.

Based upon this study, we estimate that the mean forward to backward sway range for young individuals is 4.4 mm less than that for elderly individuals. However, the 95%

confidence interval indicates that the mean for young individuals could be as much as 10.4 mm less to as much as 1.62 mm MORE than that for elderly individuals.

Plausible values for the true mean difference (young – elderly) range from large negative values to small positive values and include the possibility that the true mean difference could be zero.



In SPSS, we have the reverse order for our comparison, elderly – young. So our test statistic, mean difference, and confidence interval values are all reversed. Otherwise, the results are equivalent.

For SPSS we begin by looking for the p-value of the test for equality of variances, which is 0.258, outlined in the right column of the first table of this output. Notice the p-value is different from SAS and indeed the test used by SPSS may be preferred as it is less sensitive to outliers and departures from normality. It is possible that SAS users and SPSS users may get different results for this test and thus choose a different row for their t-test.

In this case we get the same conclusion as for SAS by failing to reject the null hypothesis that the variances are equal and so we would still use the equal variances row in the tables outlined in the output.

We find a p-value for the equal variances two sample t-test of 0.140 and so there is not enough evidence to conclude that the population mean ***forward to backward sway range*** differs between elderly and young and individuals.

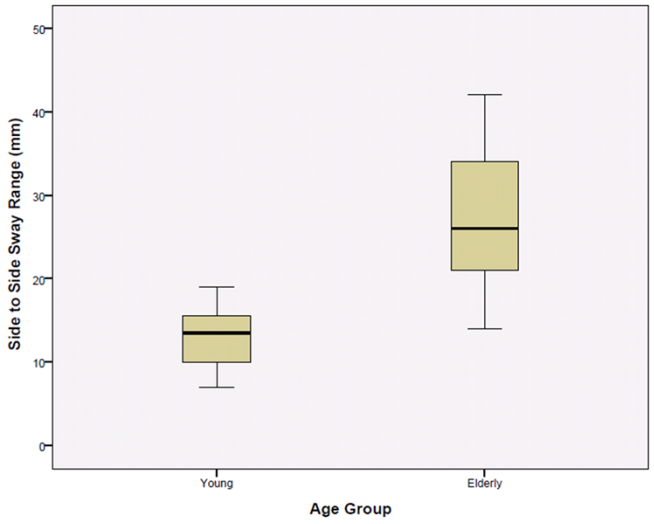
The appropriate 95% confidence interval for the difference between the population mean for elderly and that for young is given as -1.62 to 10.4.

We can interpret our estimate and confidence interval as follows.

Based upon this study, we estimate that the mean forward to backward sway range for elderly individuals is 4.4 mm greater than that for young individuals. However, the 95% confidence interval indicates that the mean for elderly individuals could be as much as 1.62 mm less than to as much as 10.4 mm more than that for young individuals.

Plausible values for the true mean difference (elderly – young) range from small negative values to large positive values and include the possibility that the true mean difference could be zero.

Side to Side Sway Range



UF UNIVERSITY of FLORIDA

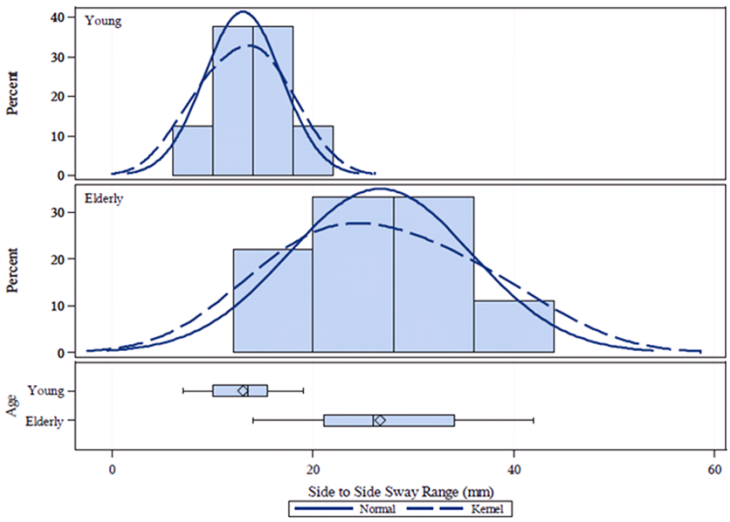
Now for Side to Side Sway Range.

The boxplots show a much larger difference in variation with the distribution of young individuals having a much smaller spread than that for elderly individuals.

It does seem a more obvious difference exists for side-to-side sway.

Elderly individuals tend to have larger side to side sway than young individuals.

Side to Side Sway Range

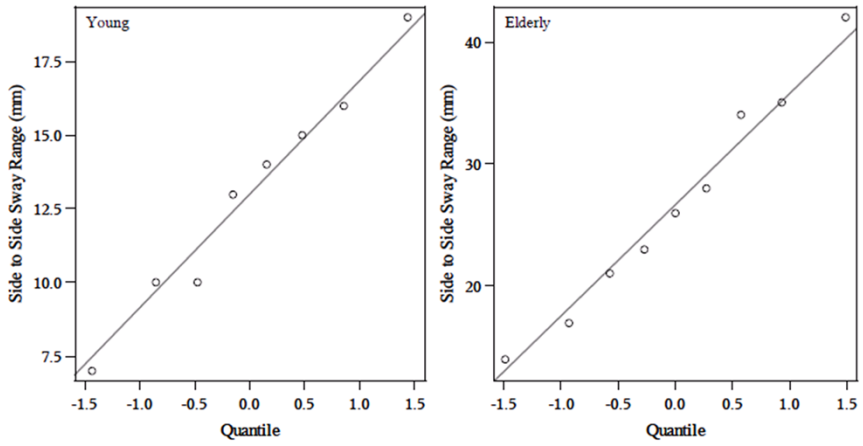


UF UNIVERSITY of FLORIDA

Here is the output from SAS when conducting the two-sample t-test.

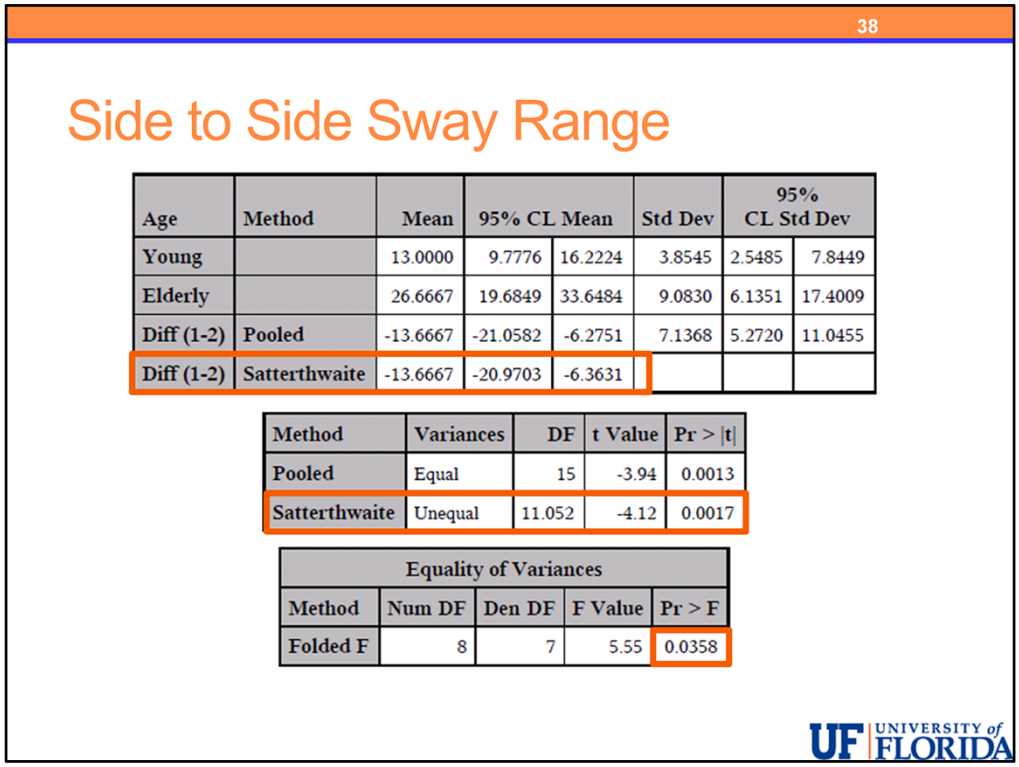
Both distributions seem reasonably normal comparing the densities (solid vs. dotted line) on these histograms. We also see the boxplots and there are no outliers in the data.

Side to Side Sway Range



UF UNIVERSITY of FLORIDA

These are the QQ-plots from SAS which also show no reason for concern regarding the normality assumption.



In the SAS output, we begin by looking for the p-value of the test for equality of variances, which is 0.0358.

Thus here we do reject the null hypothesis that the variances are equal and so we can NOT use the equal variances row in the tables, we should instead use the unequal variances row outlined in the output.

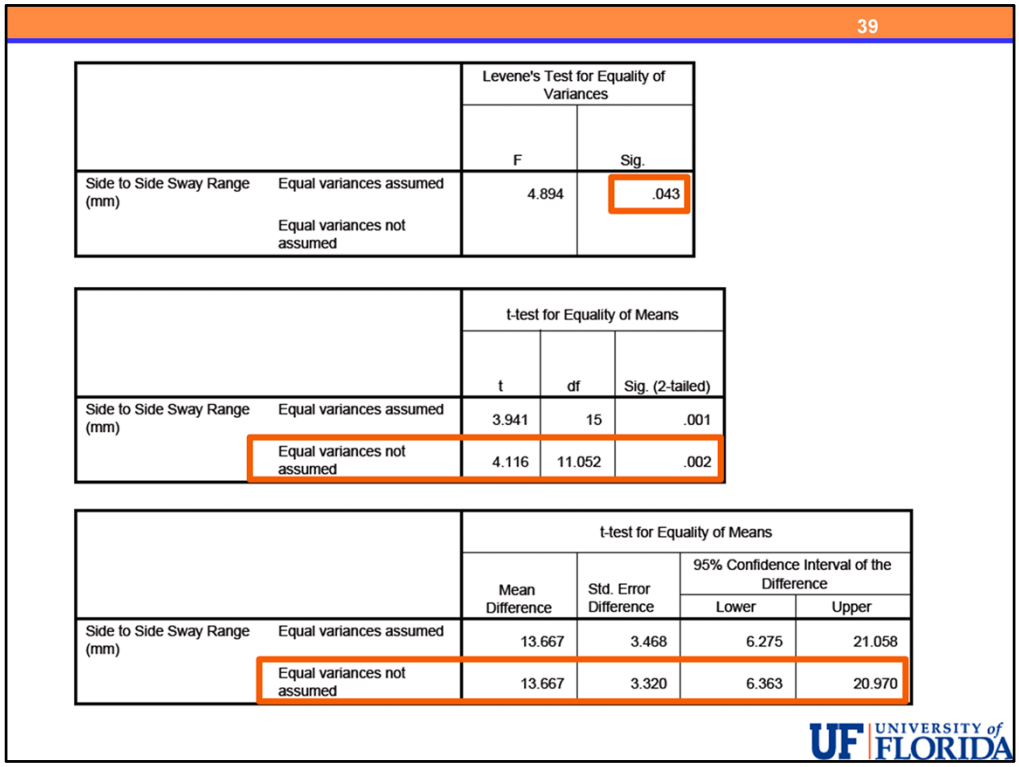
We find a p-value for the unequal variances two sample t-test of 0.0017 and so there is enough evidence to conclude that the population mean ***side to side sway range*** differs between young and elderly individuals.

The appropriate 95% confidence interval for the difference between the population mean for young and that for elderly is given as -20.97 to -6.36.

We can interpret our estimate and confidence interval as follows.

Based upon this study, we estimate that the mean side to side sway range for young individuals is 13.7 mm less than that for elderly individuals. However, the 95% confidence interval indicates that the mean for young individuals could be as little as 6.36 mm to as much as 20.97 mm less than that for elderly individuals.

Plausible values for the true mean difference (young – elderly) are all negative and hence zero is not a plausible value.



Once again the results in SPSS are reversed and yet reveal the same conclusion. We begin by looking for the p-value of the test for equality of variances, which is 0.043.

Thus here we do reject the null hypothesis that the variances are equal and so we can NOT use the equal variances row in the tables, we should use the unequal variances row - outlined in the output.

We find a p-value for the unequal variances two sample t-test of 0.002 and so there is enough evidence to conclude that the population mean ***side to side sway range*** differs between young and elderly individuals.

The appropriate 95% confidence interval for the difference between the population mean for elderly and that for young is given as 6.36 to 20.97.

We can interpret our estimate and confidence interval as follows.

Based upon this study, we estimate that the mean side to side sway range for elderly individuals is 13.7 mm more than that for young individuals. However, the 95% confidence interval indicates that the mean for elderly individuals could be as little as 6.36 mm to as much as 20.97 mm more than that for young individuals.

Plausible values for the true mean difference (elderly – young) are all positive and hence zero is not a plausible value.

Non-Parametric Tests

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Forward to Backward Sway Range (mm) is the same across categories of Age Group.	Independent-Samples Mann-Whitney U Test	.139 [1]	Retain the null hypothesis.
2	The distribution of Side to Side Sway Range (mm) is the same across categories of Age Group.	Independent-Samples Mann-Whitney U Test	.001 [1]	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1] Exact significance is displayed for this test.

UF UNIVERSITY of FLORIDA

We would get the same conclusions from the non-parametric Wilcoxon Rank-Sum test. The SPSS results are shown here with a p-value for forward to backward of 0.139 and one for side to side of 0.001.

Wilcoxon Scores (Rank Sums) for Variable f_b Classified by Variable Age					
Age	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Elderly	9	96.50	81.0	10.360417	10.722222
Young	8	56.50	72.0	10.360417	7.062500
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable s_s Classified by Variable Age					
Age	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Elderly	9	112.50	81.0	10.379561	12.50000
Young	8	40.50	72.0	10.379561	5.06250
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	56.5000
Normal Approximation	
Z	-1.4478
One-Sided Pr < Z	0.0738
Two-Sided Pr > |Z|	0.1477
t Approximation	
One-Sided Pr < Z	0.0835
Two-Sided Pr > |Z|	0.1670
Z includes a continuity correction of 0.5.	

Wilcoxon Two-Sample Test	
Statistic	40.5000
Normal Approximation	
Z	-2.9866
One-Sided Pr < Z	0.0014
Two-Sided Pr > |Z|	0.0028
t Approximation	
One-Sided Pr < Z	0.0044
Two-Sided Pr > |Z|	0.0087
Z includes a continuity correction of 0.5.	

UF UNIVERSITY of FLORIDA

The SAS results are more complex. The two-sided p-values for either the Z or t approximation are acceptable.

For forward to backward on the left, we find a p-value of 0.1477 for the Z or 0.1670 for the t.

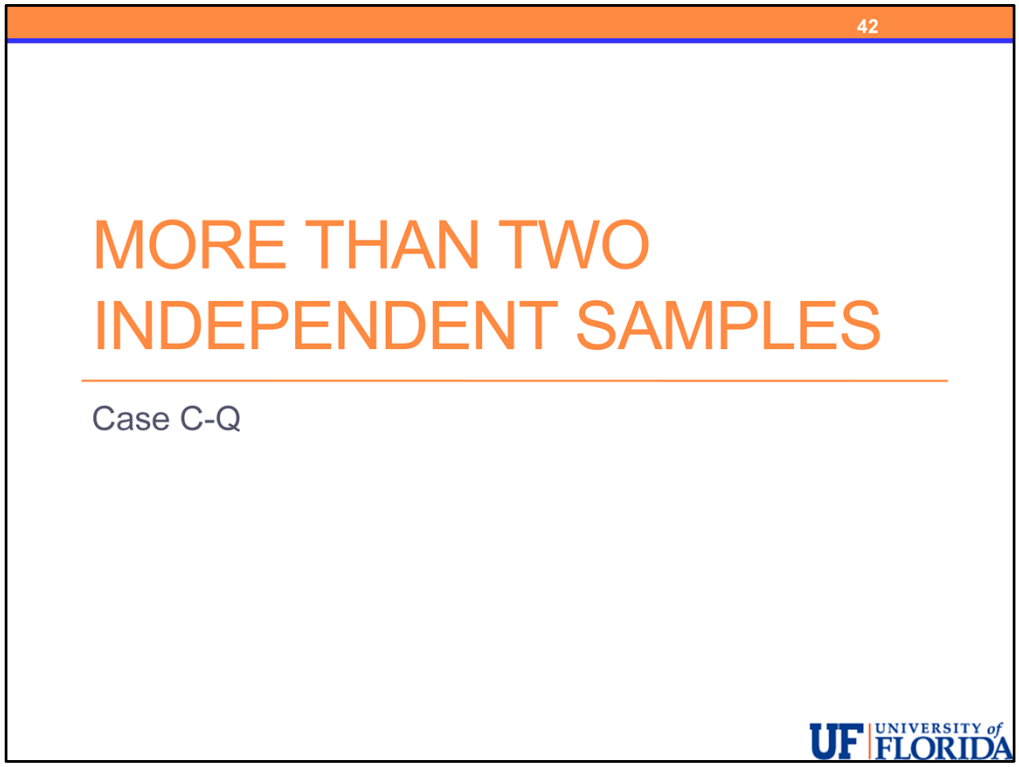
And for side to side on the right, we find a p-value of 0.0028 for the Z or 0.0087 for the t.

Finally, for our test involving forward to backward sway range, since we failed to reject the null hypothesis, it is possible that we could have made a type II error.

In context we would not conclude that there is a difference in the mean forward to backward sway when in fact there is a difference.

And for our test involving side to side sway range, since we rejected the null hypothesis, it is possible that we could have made a type I error.

In context we would conclude that there is a difference in the mean side to side sway when in fact there is NOT a difference.



Now we will look quickly at an example of ANOVA.

Data: Hot Dogs

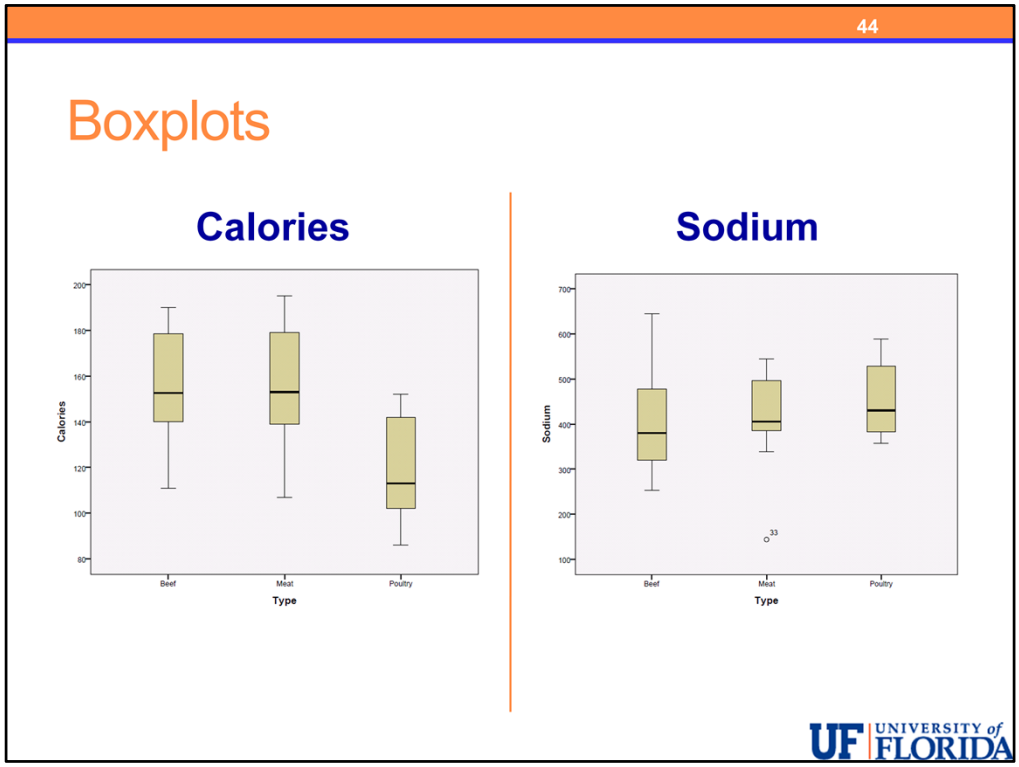
Type	Calories	Sodium
Beef	186	495
Beef	181	477
Meat	190	545
Meat	147	360
Poultry	87	359
Poultry	144	545

UF UNIVERSITY of FLORIDA

Data: http://lib.stat.cmu.edu/DASL/Stories/Hotdogs.html

In our previous section on Case C-Q, we discussed an example regarding the calories and sodium content of hot dogs.

We have two different response variables (calories and sodium) and we wish to compare beef, poultry, and other "meat" hot dogs.



Here are the SPSS boxplots for calories by hot dog type on the left and sodium by hot dog type on the right.

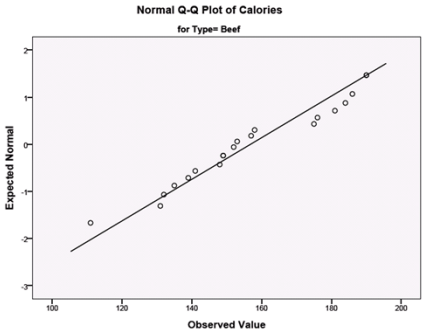
For calories, it seems clear that poultry hot dogs tend to be lower in calories but there is not much difference between Beef and Meat for calories. The variation in calories is similar for all types.

For sodium, there is no clear difference. There is one low outlier for type = meat.

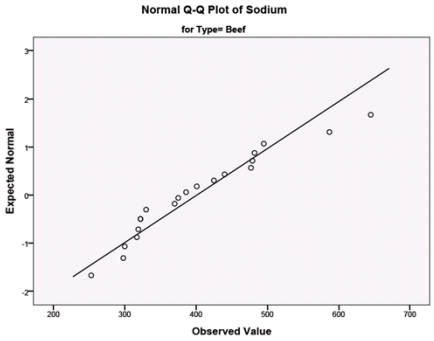
The variation in sodium content is not as consistent between the three types as for calories – however, neither is there a clear indication of a large difference in variation between these groups.

QQ-Plots: Beef

Calories



Sodium

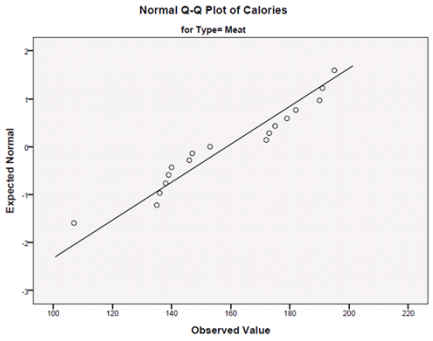


UF UNIVERSITY of FLORIDA

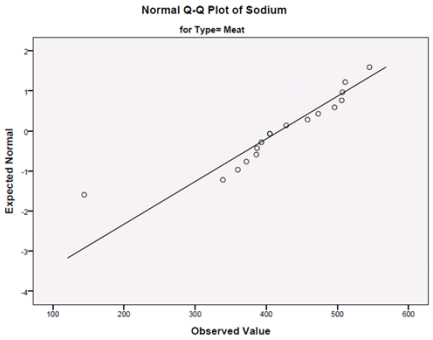
The QQ-plots for Beef show no major problems.

QQ-Plots: Meat

Calories



Sodium

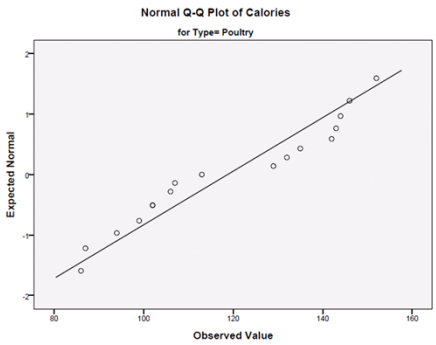


UF UNIVERSITY of FLORIDA

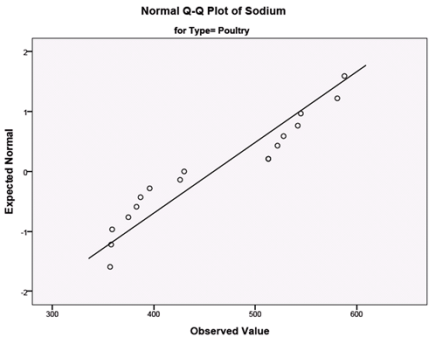
The QQ-plots for meat for both variables are not too bad but the one for sodium does show a fairly unusual outlier – as we saw in the boxplots.

QQ-Plots: Poultry

Calories



Sodium



UF UNIVERSITY of FLORIDA

The QQ-plots for poultry show no major problems.

Overall, the normality assumption seems reasonable for these responses within our hot dog type groups.

ANOVA - SPSS

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Calories	Between Groups	17692.195	2	8846.098	16.074	.000
	Within Groups	28067.138	51	550.336		
	Total	45759.333	53			
Sodium	Between Groups	31738.715	2	15869.357	1.778	.179
	Within Groups	455248.785	51	8926.447		
	Total	486987.500	53			

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Calories is the same across categories of TypeCode.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The distribution of Sodium is the same across categories of TypeCode.	Independent-Samples Kruskal-Wallis Test	.095	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

UF UNIVERSITY of FLORIDA

The results from SPSS for both the standard ANOVA and the Kruskal-Wallis test are provided here for both variables.

For calories both the ANOVA and Kruskal-Wallis test have p-values of 0.000 and thus we reject the null hypothesis.

There are statistically significant differences in mean calorie count between these three hot dog types. Although no formal test was conducted, based upon the boxplots, it seems clear that the mean calories for poultry hot dogs is different from (and in fact lower than) both beef and other meat hot dogs. However, the boxplots for the other two groups are extremely similar and thus are not likely to be found to have different means.

For sodium, however, both the ANOVA and Kruskal-Wallis test have p-values over 0.05. For the ANOVA we have a p-value of 0.179. For the Kruskal-Wallis test, we have a p-value of 0.095. In either case, we fail to reject the null hypothesis and find no evidence of any differences in the population mean sodium content between these three hot dog types.

SAS - ANOVA

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	17692.19510	8846.09755	16.07	<.0001
Error	51	28067.13824	550.33604		
Corrected Total	53	45759.33333			

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	31738.7147	15869.3574	1.78	0.1793
Error	51	455248.7853	8926.4468		
Corrected Total	53	486987.5000			

UF UNIVERSITY of FLORIDA

In SAS we have the same results for the standard ANOVA.

Boxplots

Calories

Wilcoxon Scores (Rank Sums) for Variable Calories Classified by Variable Type					
Type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Beef	20	675.50	550.00	55.809023	33.775000
Meat	17	577.50	467.50	53.675400	33.970588
Poultry	17	232.00	467.50	53.675400	13.647059
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	19.2514
DF	2
Pr > Chi-Square	<.0001

Sodium

Wilcoxon Scores (Rank Sums) for Variable Sodium Classified by Variable Type					
Type	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Beef	20	441.00	550.00	55.819666	22.050000
Meat	17	478.50	467.50	53.685635	28.147059
Poultry	17	565.50	467.50	53.685635	33.264706
Average scores were used for ties.					

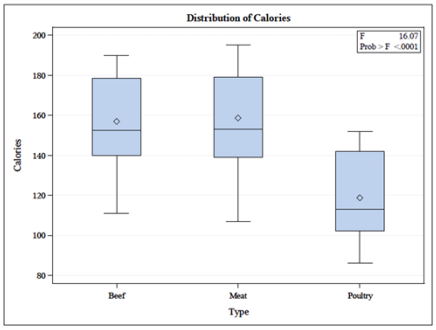
Kruskal-Wallis Test	
Chi-Square	4.7128
DF	2
Pr > Chi-Square	0.0948

UF UNIVERSITY of FLORIDA

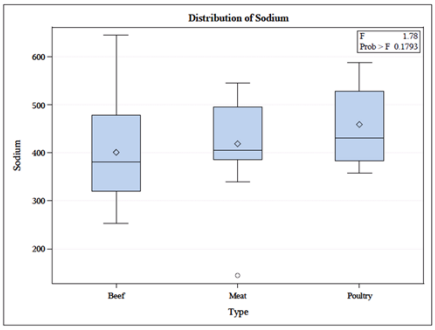
And for the Kruskal-Wallis test.

Boxplots

Calories

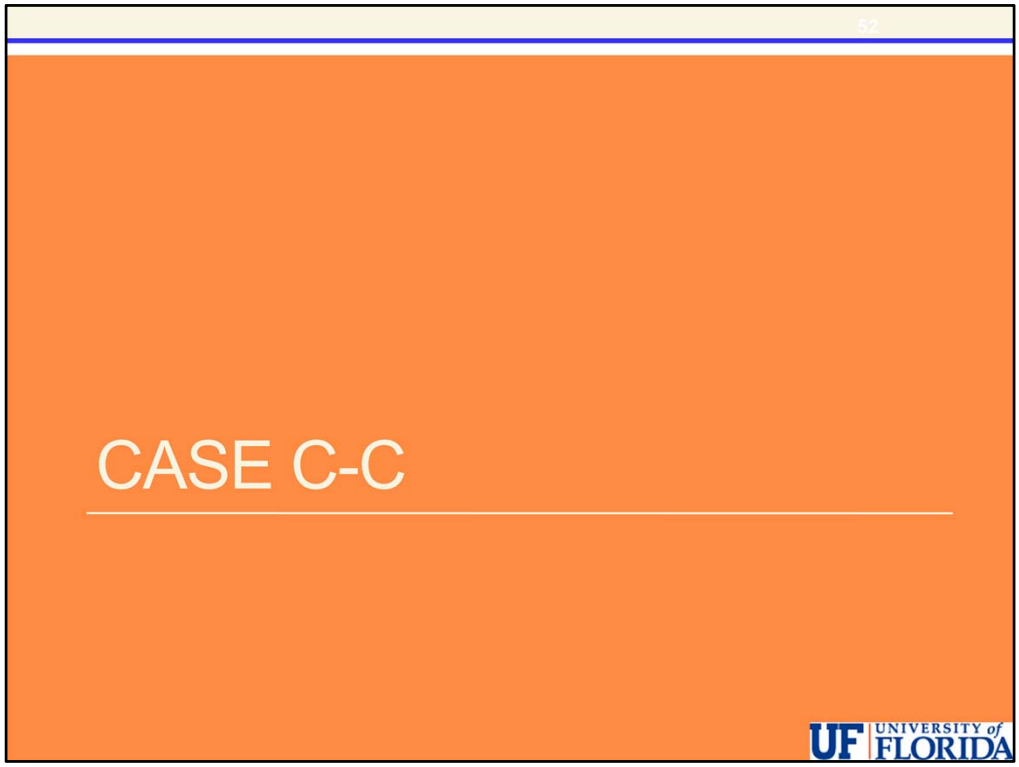


Sodium



UF UNIVERSITY of FLORIDA

These boxplots are provided by SAS with the ANOVA analysis. They provide the p-value for the F-test and illustrate the mean as well as the median for comparison.



Original Data: http://bolt.mph.ufl.edu/2012/12/23/learn-by-doing-case-c-c-software/

When we discussed exploratory data analysis for Case C-C, we used a dataset based on a 1999 study at the University of Pennsylvania and Children's Hospital of Philadelphia, in which parents were surveyed about the lighting conditions under which their children slept between birth and age 2 (lamp, night-light, or no light) and whether or not their children developed nearsightedness (myopia). The purpose of the study was to explore the effect of a young child's nighttime exposure to light on later nearsightedness.

Notice this is an observational study which does not control for any other possible lurking variables.

Data: Nightlight

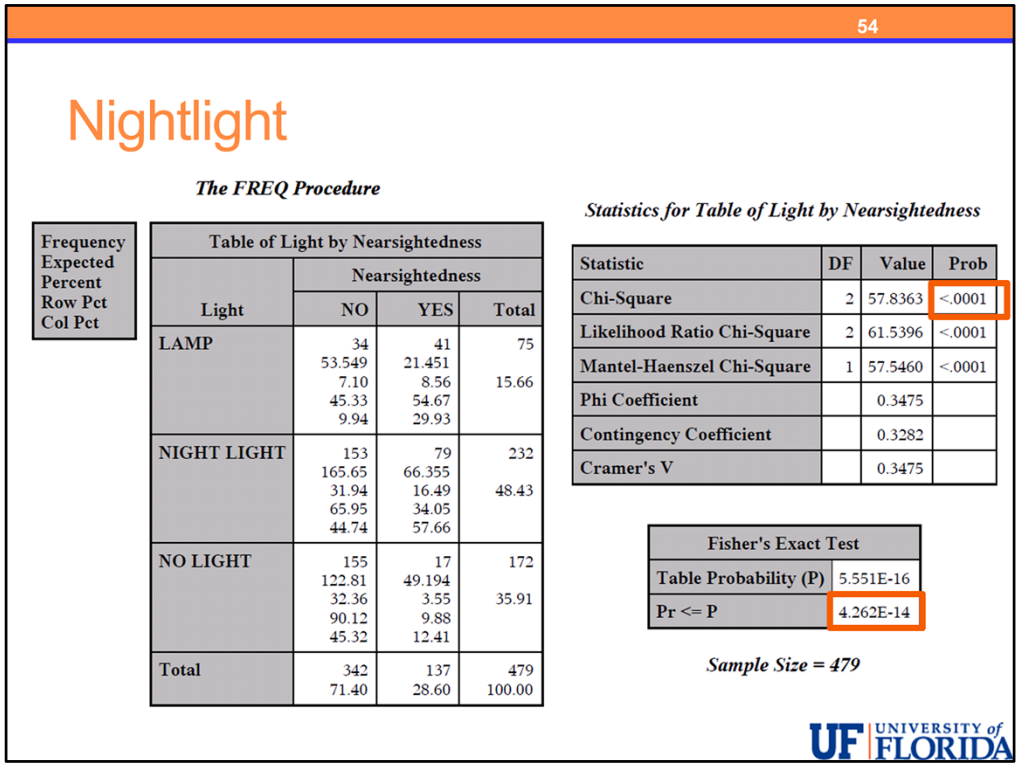
Obs	Anylight	Light	Nearsightedness
1	NO	NO LIGHT	NO
2	YES	NIGHT LIGHT	NO
3	YES	LAMP	NO
4	NO	NO LIGHT	NO
5	NO	NO LIGHT	NO
6	YES	NIGHT LIGHT	YES
7	YES	LAMP	NO
8	YES	NIGHT LIGHT	YES
9	NO	NO LIGHT	NO
10	YES	NIGHT LIGHT	YES

UF UNIVERSITY of FLORIDA

Here is a few lines of the data.

Notice the variable values are not coded.

We have added a new variable called Anylight which is NO for children with no light and YES for children with a lamp or night light.



To investigate the association between type of light and nearsightedness, using the original three level light variable, we can conduct a chi-squared test or fisher's exact test.

The null hypothesis is that there is no relationship between the type of light and future nearsightedness in other words, that type of light and future nearsightedness are independent.

The alternative hypothesis is that there IS a relationship between the type of light and future nearsightedness in other words, that type of light and future nearsightedness are dependent.

In SAS, the values in each cell are in the following order – specified in the "legend" in the upper left corner of the table. Frequency, Expected Count, Overall Percent, Row Percent, Column Percent

Using the row percentages, our contingency table shows that among children with no light, 9.88% developed nearsightedness, among children with a nightlight, 34.05% developed nearsightedness and among children with a lamp, 54.67% developed nearsightedness.

Without using any inferential statistics, this difference seems extreme. And, in fact, the p-value of both the chi-square test (given as < 0.0001) and Fisher's exact test (which gives a tiny probability of 4.3x10^(-14)) show an extremely highly significant result.

Thus we can reject the null hypothesis.

We conclude that there is enough evidence of an association between the type of light at night and the future development of nearsightedness in the population. Type of light used at night and development of nearsightedness are dependent.

Nightlight

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	57.836[a]	2	.000	.000
Likelihood Ratio	61.540	2	.000	.000
Fisher's Exact Test	61.016			.000
N of Valid Cases	479			

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.45.

			Nearsightedness		
			NO	YES	Total
Light	LAMP	Count	34	41	75
		Expected Count	53.5	21.5	75.0
		% within Light	45.3%	54.7%	100.0%
		% within Nearsightedness	9.9%	29.9%	15.7%
		% of Total	7.1%	8.6%	15.7%
	NIGHT LIGHT	Count	153	79	232
		Expected Count	165.6	66.4	232.0
		% within Light	65.9%	34.1%	100.0%
		% within Nearsightedness	44.7%	57.7%	48.4%
		% of Total	31.9%	16.5%	48.4%
	NO LIGHT	Count	155	17	172
		Expected Count	122.8	49.2	172.0
		% within Light	90.1%	9.9%	100.0%
		% within Nearsightedness	45.3%	12.4%	35.9%
		% of Total	32.4%	3.5%	35.9%
Total		Count	342	137	479
		Expected Count	342.0	137.0	479.0
		% within Light	71.4%	28.6%	100.0%
		% within Nearsightedness	100.0%	100.0%	100.0%
		% of Total	71.4%	28.6%	100.0%

UF UNIVERSITY of FLORIDA

The SPSS output gives exactly the same information. The only difference is the order that the cell values are presented.

In SAS the values were Frequency, Expected Count, Overall Percent, Row Percent, Column Percent.

In SPSS they are given as Count – which is the frequency, expected count, then % within light which is the ROW percent, then % within nearsightedness which is the column percent, with the overall percent being provided last.

Understanding the output provided by your software is important now and most definitely in practice.

The p-value of the appropriate chi-square test and Fisher's exact test are outlined in the table and are reported to be 0.000 which doesn't mean the p-value is exactly equal to zero but it is zero rounded to three decimal places.

Again, our conclusion is that there is a highly statistically significant association between type of light and nearsightedness.

Nightlight

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	46.041[a]	1	.000	.000	.000
Continuity Correction[b]	44.622	1	.000		
Likelihood Ratio	51.605	1	.000	.000	.000
Fisher's Exact Test				.000	.000
N of Valid Cases	479				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 49.19.

b. Computed only for a 2x2 table

			Nearsightedness		
			NO	YES	Total
anylight	NO	Count	155	17	172
		Expected Count	122.8	49.2	172.0
		% within anylight	90.1%	9.9%	100.0%
		% within Nearsightedness	45.3%	12.4%	35.9%
		% of Total	32.4%	3.5%	35.9%
	YES	Count	187	120	307
		Expected Count	219.2	87.8	307.0
		% within anylight	60.9%	39.1%	100.0%
		% within Nearsightedness	54.7%	87.6%	64.1%
		% of Total	39.0%	25.1%	64.1%
Total		Count	342	137	479
		Expected Count	342.0	137.0	479.0
		% within anylight	71.4%	28.6%	100.0%
		% within Nearsightedness	100.0%	100.0%	100.0%
		% of Total	71.4%	28.6%	100.0%

UF UNIVERSITY of FLORIDA

To investigate the association between the variable anylight and nearsightedness we can conduct a chi-squared test with a continuity correction or fisher's exact test.

The null hypothesis is that there is no relationship between whether or not the child slept with any light and future nearsightedness in other words, exposure to light during sleep and future nearsightedness are independent.

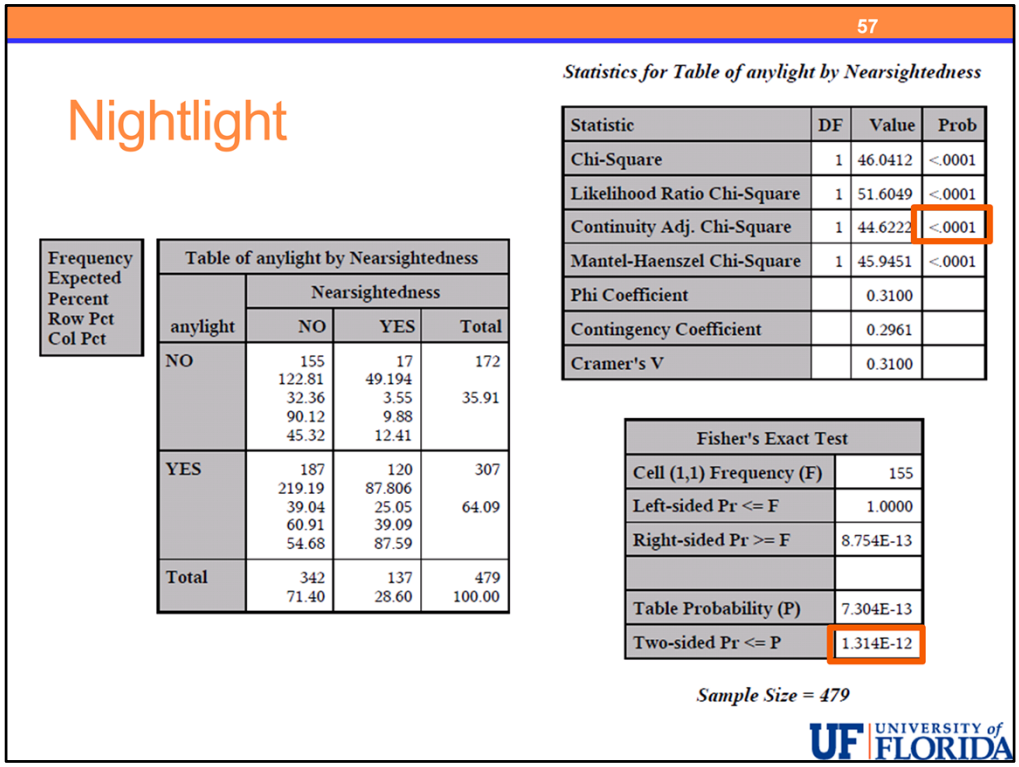
The alternative hypothesis is that there IS a relationship between whether or not the child slept with any light and future nearsightedness in other words, exposure to light during sleep and future nearsightedness are dependent.

Using the row percentages, our contingency table shows that among children with no light, 9.88% developed nearsightedness whereas among children with a nightlight or lamp, 39.09% developed nearsightedness.

In SPSS, the p-value of both the continuity adjusted chi-square test and Fisher's exact test are given as 0.000 giving an extremely highly significant result.

Thus we can reject the null hypothesis.

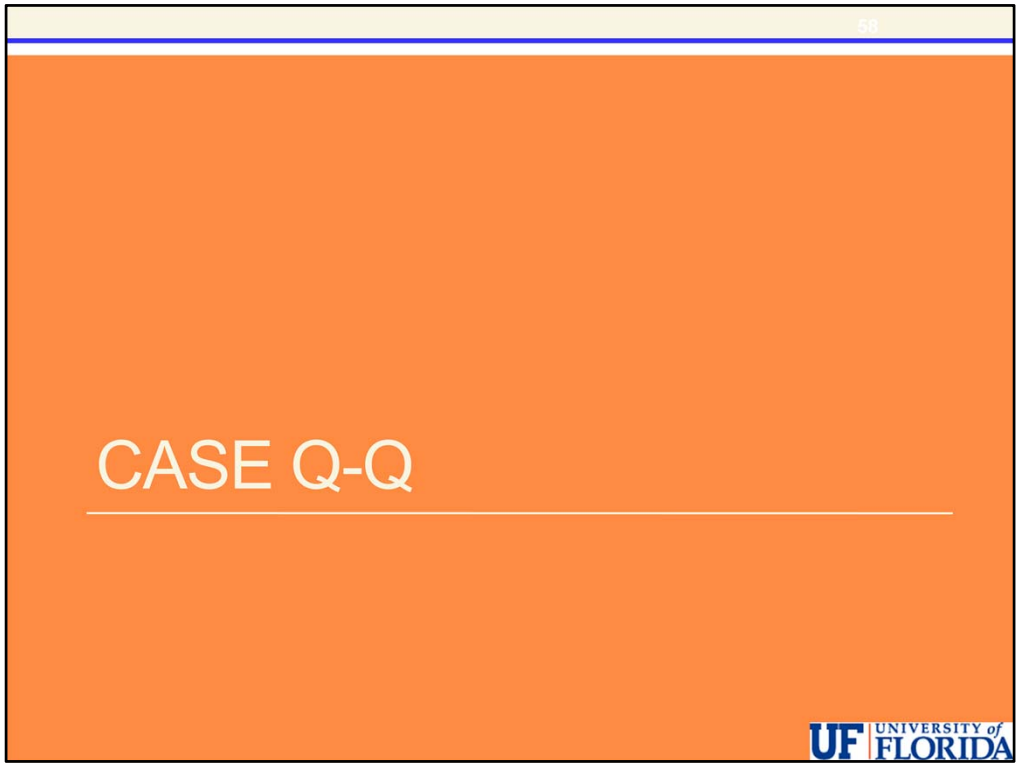
We conclude that there is enough evidence of an association between whether or not the child slept with any light and the future development of nearsightedness in the population. Exposure to light during sleep and future nearsightedness are dependent.



The only difference between the SAS output and SPSS output is in the reporting of the p-values.

In SAS, the p-value of the continuity adjusted chi-square test is given as < 0.0001 and for Fisher's exact test it is given as 1.3x10^(-12).

Both of these are extremely small and so we would again reject the null hypothesis.



Dataset information:

http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html

For case Q-Q, we will use a dataset containing information about U.S. states during the 1970's.

Facts on US States in 1970's

Obs	State	Population	Income	Illiteracy	Life_Exp	Murder	HS_Grad	Frost	Area
1	Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
2	Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
3	Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
4	Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
5	California	21198	5114	1.1	71.71	10.3	62.6	20	156361
6	Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
7	Connecticut	3100	5348	1.1	72.48	3.1	56	139	4862

UF UNIVERSITY of FLORIDA

A few lines of the data are shown here.

The variables are:

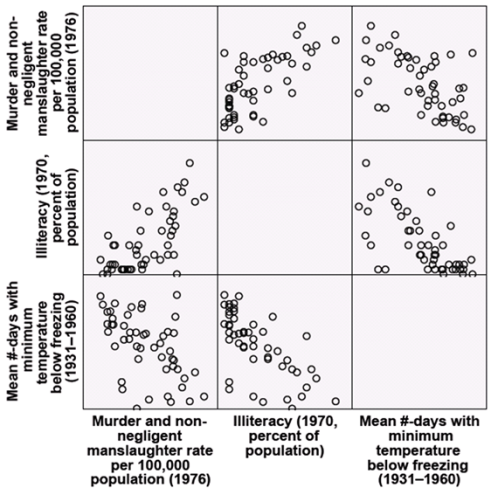
- state name
- Population
- per-capita income
- illiteracy rate
- life expectancy
- Murder and non-negligent manslaughter rate per 100,000 population
- Percent high school graduates
- Mean number of days with the minimum temperature below freezing in capital or large city

And the

- Land area in square miles.

In particular we will investigate the associations between murder, frost, and illiteracy.

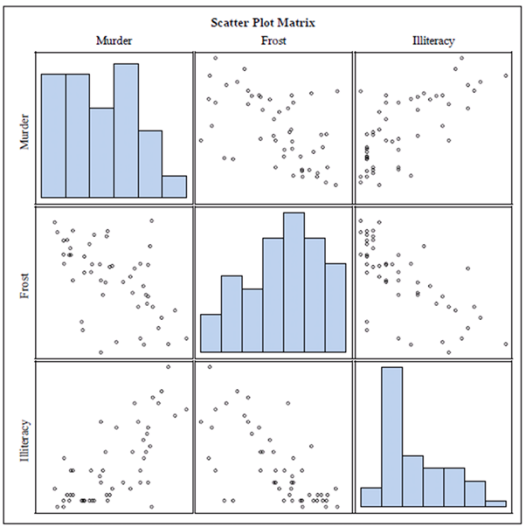
Facts on US States in 1970's



UF UNIVERSITY of FLORIDA

This is a scatterplot matrix from SPSS showing the scatterplots of all possible pairings between the variables murder, frost, and illiteracy.

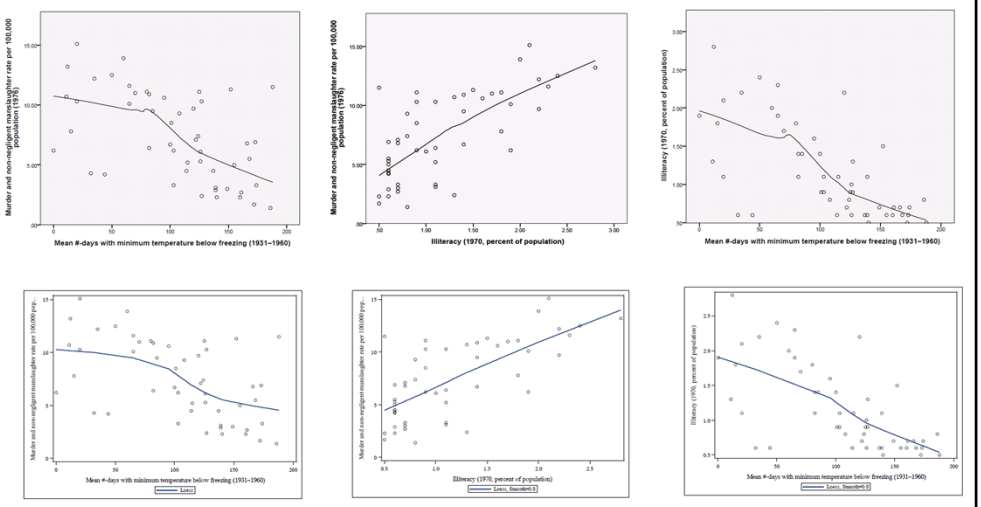
Facts on US States in 1970's



And a similar scatterplot matrix from SAS.

None of these scatterplots show any clear non-linear trends although there may be some outliers.

Facts on US States in 1970's



UF UNIVERSITY of FLORIDA

Here we have individual scatterplots with LOESS curves for
Murder vs. frost (on the left)
Murder vs. illiteracy (center)
Illiteracy vs. frost (right)

Of the three plots, the murder vs. illiteracy scatterplot in the center shows the most linear trend followed by illiteracy vs. frost (on the right) and finally murder vs. frost (on the left).

Although the plot for murder vs. frost (on the left) may be truly non-linear, we will investigate all three of these relationships further using correlation and regression.

From these plots we would expect a negative correlation between murder and frost (on the left) and between illiteracy and frost (on the right)

And a positive correlation between murder and illiteracy.

Facts on US States in 1970's

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0	Murder	Frost	Illiteracy
Murder Murder and non-negligent manslaughter rate per 100,000 population (1976)	1.00000	-0.53888 <.0001	0.70298 <.0001
Frost Mean #-days with minimum temperature below freezing (1931–1960)	-0.53888 <.0001	1.00000	-0.67195 <.0001
Illiteracy Illiteracy (1970, percent of population)	0.70298 <.0001	-0.67195 <.0001	1.00000

Spearman Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0	Murder	Frost	Illiteracy
Murder Murder and non-negligent manslaughter rate per 100,000 population (1976)	1.00000	-0.54384 <.0001	0.67236 <.0001
Frost Mean #-days with minimum temperature below freezing (1931–1960)	-0.54384 <.0001	1.00000	-0.68319 <.0001
Illiteracy Illiteracy (1970, percent of population)	0.67236 <.0001	-0.68319 <.0001	1.00000

UF UNIVERSITY of FLORIDA

First we have the SAS output for both Pearson's and Spearman's correlation between all combinations.

The results are all highly statistically significant.

For Murder vs. Frost, Pearson's correlation is -0.539 and Spearman's is -0.544. Both indicating a moderately strong negative linear association between murder and frost. As the mean number of days below freezing increases, the murder rate tends to decrease.

For Murder vs. Illiteracy, Pearson's correlation is 0.703 and Spearman's is 0.672. Both indicating a somewhat strong positive linear association between murder and illiteracy. As the illiteracy rate increases, the murder rate tends to increase.

For Frost vs. Illiteracy, Pearson's correlation is -0.672 and Spearman's is -0.683. Both indicating a somewhat strong negative linear association between frost and illiteracy. As the mean number of days below freezing increases, the illiteracy rate tends to decrease.

These values confirm what we found in the previous scatterplots.

Facts on US States in 1970's

Correlations

		Illiteracy (1970, percent of population)	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Mean #-days with minimum temperature below freezing (1931–1960)
Illiteracy (1970, percent of population)	Pearson Correlation	1	.703**	-.672**
	Sig. (2-tailed)		.000	.000
	N	50	50	50
Murder and non-negligent manslaughter rate per 100,000 population (1976)	Pearson Correlation	.703**	1	-.539**
	Sig. (2-tailed)	.000		.000
	N	50	50	50
Mean #-days with minimum temperature below freezing (1931–1960)	Pearson Correlation	-.672**	-.539**	1
	Sig. (2-tailed)	.000	.000	
	N	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

UF UNIVERSITY of FLORIDA

We find the same results for Pearson's correlation in SPSS.

Facts on US States in 1970's

Correlations

			Illiteracy (1970, percent of population)	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Mean #-days with minimum temperature below freezing (1931–1960)
Spearman's rho	Illiteracy (1970, percent of population)	Correlation Coefficient	1.000	.672**	-.683**
		Sig. (2-tailed)	.	.000	.000
		N	50	50	50
	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Correlation Coefficient	.672**	1.000	-.544**
		Sig. (2-tailed)	.000	.	.000
		N	50	50	50
	Mean #-days with minimum temperature below freezing (1931–1960)	Correlation Coefficient	-.683**	-.544**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	50	50	50

UF UNIVERSITY of FLORIDA

And for Spearman's correlation. The only difference is in the order the variables are presented.

Murder vs. Frost

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	193.91028	193.91028	19.64	<.0001
Error	48	473.83552	9.87157		
Corrected Total	49	667.74580			

Root MSE	3.14191	R-Square	0.2904
Dependent Mean	7.37800	Adj R-Sq	0.2756
Coeff Var	42.58479		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	11.37569	1.00549	11.31	<.0001	9.35401	13.39737
Frost	Mean #-days with minimum temperature below freezing (1931–1960)	1	-0.03827	0.00863	-4.43	<.0001	-0.05563	-0.02091

UF UNIVERSITY of FLORIDA

Now we can continue with simple linear regression.

Values of particular interest are outlined.

We have an R-squared of 0.2904 indicating that 29% of the variation in murder rate can be explained by the mean number of days below freezing.

The slope is statistically significant with a p-value <0.0001.

The linear regression equation is: Predicted Murder Rate = 11.38 – 0.038(Frost).

The 95% confidence interval for the slope is -0.056 to -0.021.

We can interpret the slope and it's confidence interval by saying: For each 1 day increase in the *mean number of days with minimum temperature below freezing*, the **average** murder rate decreases by 0.038. The 95% confidence interval suggests this decrease could be as little as 0.021 to as much as 0.056.

Murder vs. Frost

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.539[a]	.290	.276	3.14191

ANOVA[a]

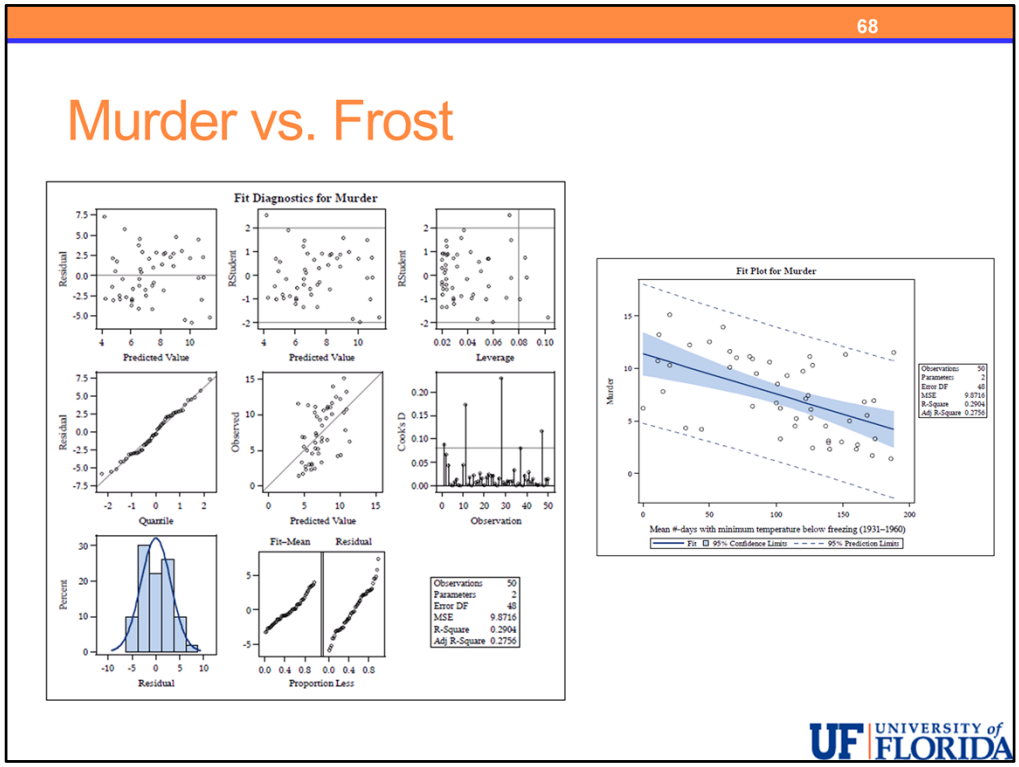
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	193.910	1	193.910	19.643	.000[b]
	Residual	473.836	48	9.872		
	Total	667.746	49			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	11.376	1.005		11.314	.000	9.354	13.397
	Mean #-days with minimum temperature below freezing (1931–1960)	-.038	.009	-.539	-4.432	.000	-.056	-.021

UF UNIVERSITY of FLORIDA

The results from SPSS are exactly the same except for differences in rounding.



In SAS we obtain the following diagnostic plots and a fit plot by default when conducting a regression analysis.

We need to verify that the relationship is reasonably linear, which we have here.

We need to check that the residuals are approximately normally distributed. Looking at the QQ-plot and histogram of the residuals, the normality assumption seems completely reasonable.

We need to check the assumption of constant variance. From the plot of the residuals by the predicted values, there is no clear violation of this assumption. The points are relatively evenly distributed with similar spread around the horizontal line at zero over the range of predicted values. We could also look at the scatterplot of the data to see that the constant variance assumption is reasonable.

We haven't learned about all of the graphs displayed here by SAS but if you go on to a regression course you will learn more about some of these plots and measures.

Murder vs. Illiteracy

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	329.98270	329.98270	46.89	<.0001
Error	48	337.76310	7.03673		
Corrected Total	49	667.74580			

Root MSE	2.65268	R-Square	0.4942
Dependent Mean	7.37800	Adj R-Sq	0.4836
Coeff Var	35.95397		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	2.39678	0.81844	2.93	0.0052	0.75118	4.04237
Illiteracy	Illiteracy (1970, percent of population)	1	4.25746	0.62171	6.85	<.0001	3.00742	5.50750

UF UNIVERSITY of FLORIDA

For murder vs. illiteracy, we have an R-squared of 0.4942 indicating that 49% of the variation in murder rate can be explained by illiteracy.

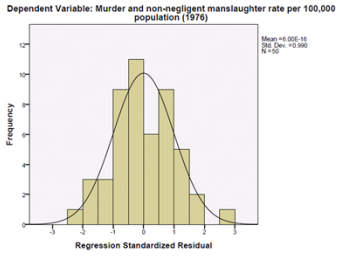
The slope is statistically significant with a p-value <0.0001.

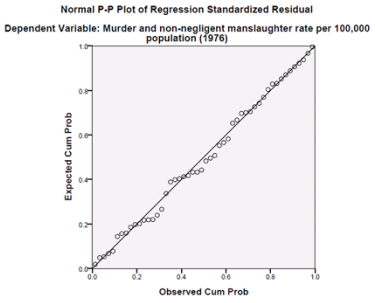
The linear regression equation is: Predicted Murder Rate = 2.40 + 4.26(Illiteracy).

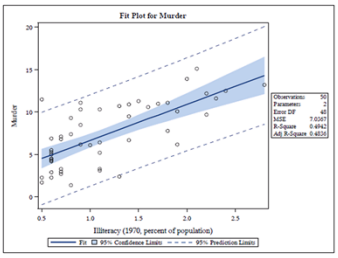
The 95% confidence interval for the slope is 3.007 to 5.508.

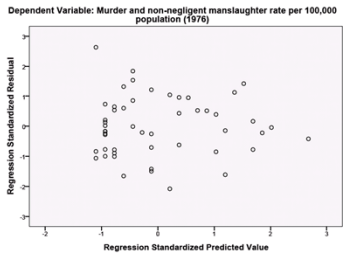
We can interpret the slope and it's confidence interval by saying: For each 1 percentage point increase in the *illiteracy rate*, the **average** murder rate increases by 4.26. The 95% confidence interval suggests this increase could be as little as 3.007 to as much as 5.508.

Murder vs. Illiteracy









UF UNIVERSITY of FLORIDA

Here we use the SPSS versions of the needed graphs to validate assumptions.

Linearity is reasonable from the scatterplot.

The histogram and normal probability plot indicate normality is reasonable. In regression, SPSS gives a PP-plot instead of a QQ-plot but these graphs are identical in what we expect to see and how they are interpreted and can be used interchangeably.

Finally the plot of the residuals by the predicted values shows no major issues although there does seem to be a slight decrease in the spread as the predicted value increases, this could be driven by two odd points in the scatterplot – one high value on the left side which is unusually far from the line and one on the right side corresponding to the largest x-value as if we ignore those two points, what remains seems to better satisfy the constant variance assumption.

Illiteracy vs. Frost

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.672[a]	.452	.440	.45610

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.220	1	8.220	39.513	.000[b]
	Residual	9.985	48	.208		
	Total	18.205	49			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.993	.146		13.655	.000	1.700	2.287
	Mean #-days with minimum temperature below freezing (1931–1960)	-.008	.001	-.672	-6.286	.000	-.010	-.005

UF UNIVERSITY of FLORIDA

Finally for illiteracy vs. frost, we have an R-squared of 0.452 indicating that 45% of the variation in illiteracy can be explained by frost.

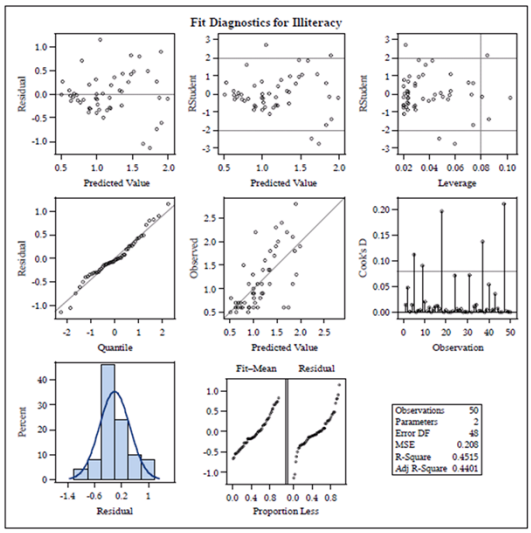
The slope is statistically significant with a p-value reported as 0.000.

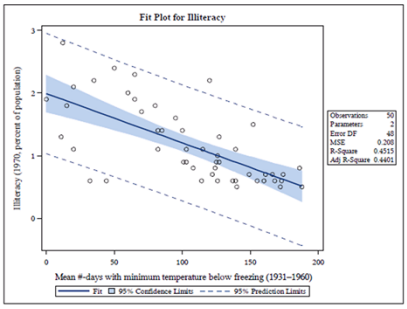
The linear regression equation is: Predicted Percent Illiteracy = 1.993 – 0.008(Frost).

The 95% confidence interval for the slope is -0.010 to -0.005.

We can interpret the slope and it's confidence interval by saying: For each 1 day increase in the *mean number of days with minimum temperature below freezing*, the **average** illiteracy percentage decreases by 0.008. The 95% confidence interval suggests this decrease could be as little as 0.005 to as much as 0.01.

Illiteracy vs. Frost





UF UNIVERSITY of FLORIDA

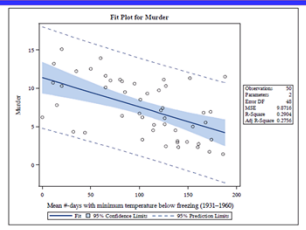
In this case, linearity is reasonably.

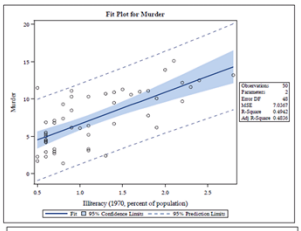
The residuals are reasonably normally distributed based upon the QQ-plot and histogram of the residuals.

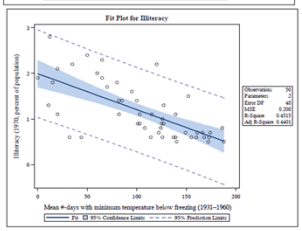
However, in this case, there does seem to be a strange pattern in the residual vs. predicted values plot and the original scatterplot. The residuals vs. predicted values shows an increasing spread as the predicted value increases. The scatterplot shows a similar trend in that as the variable Frost increases, the variation around the regression line seems to be decreasing. Thus there is some concern about the validity of the constant variance assumption.

Lurking Variables

- Cautions about Cause
- Murder is associated with Frost
- Murder is associated with Illiteracy
- Illiteracy is associated with Frost







UF UNIVERSITY of FLORIDA

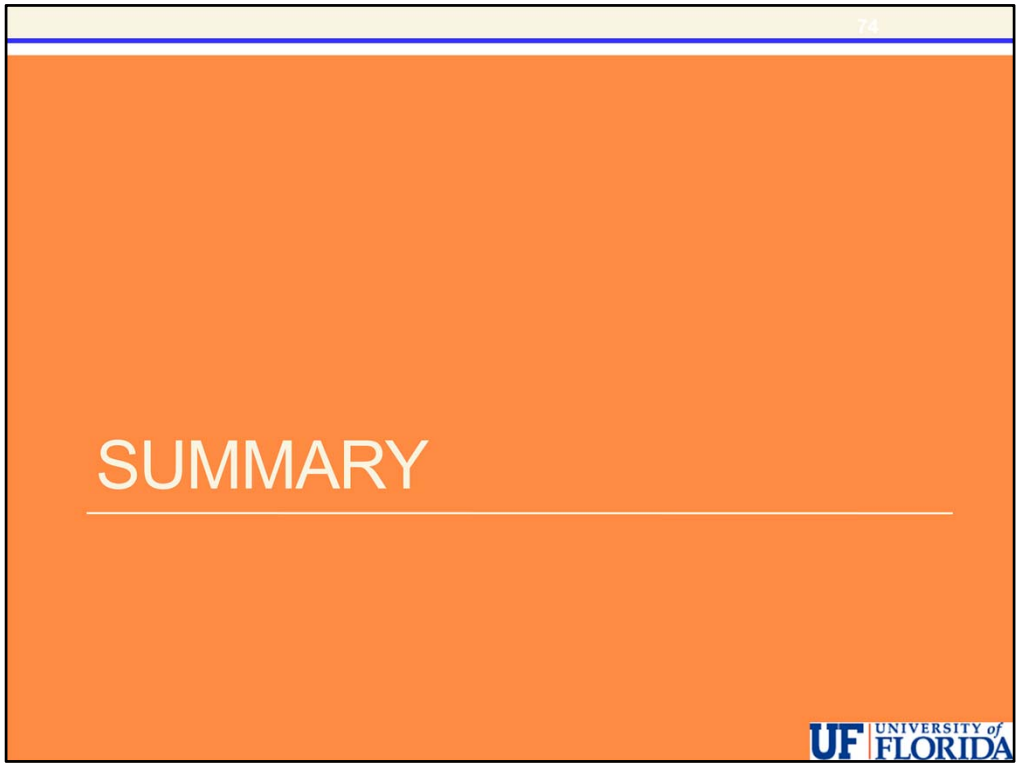
Although we found associations in each of these three regression models, we must be careful about concluding the relationship is causal.

The first relationship found as the mean number of days with minimum temperature below freezing increases, the murder rate decreases but we CANNOT say that more days below freezing CAUSES the murder rate to decrease.

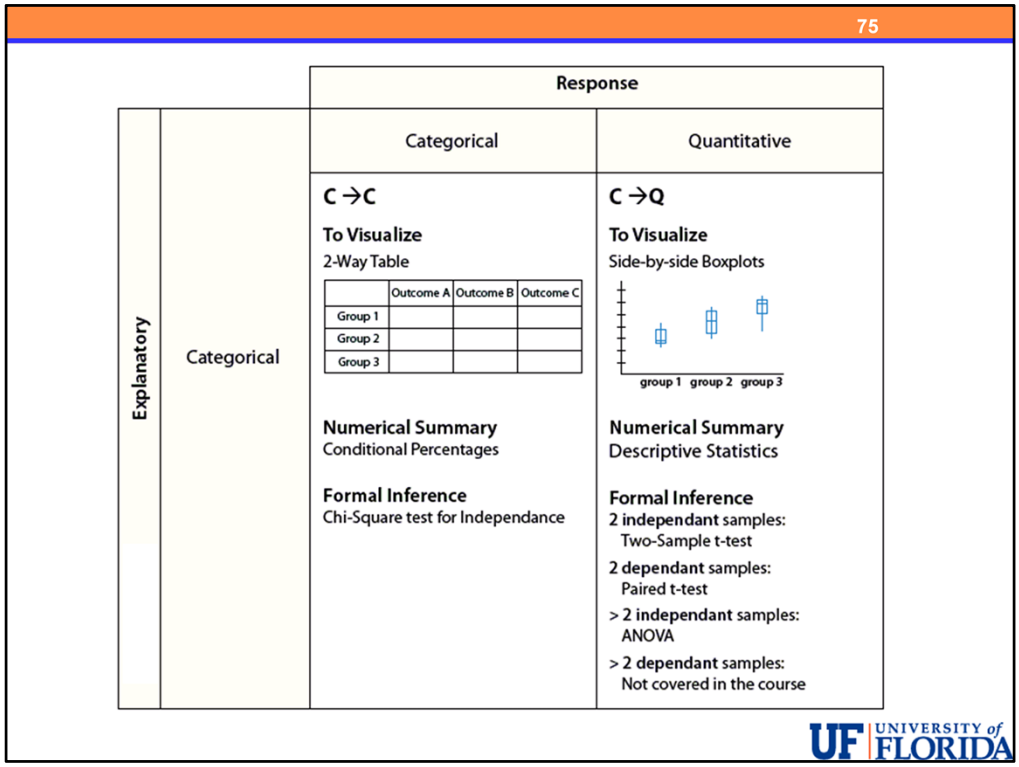
In the second relationship we see that as the illiteracy percentage increases, the murder rate also increases but again we CANNOT say that higher illiteracy percentage CAUSES the murder rate to increase.

The fact that illiteracy and frost are also related in the third relationship shows that when considering the relationship between murder and frost, we must be aware that illiteracy is also related to both frost and murder and thus illiteracy is a potential lurking variable in this relationship between murder and frost.

In general, unless you have performed a randomized controlled experiment, you should always be cautious about claiming a direct causal link between the explanatory and response variables in any analysis!



Now, let's summarize the standard methods presented in the course.



When both our explanatory variable and response variable are categorical, we visualize the results using a two-way table and summarize the data numerically using conditional percentages to compare the distribution of the response variable within the levels of the explanatory variable. We used a chi-square test as the standard method and Fisher's exact test as the non-parametric alternative. Remember to use the continuity adjusted p-value for the special case when we have a 2x2 table – when both our explanatory variable and our response variable have only two levels.

When our explanatory variable is categorical and our response is quantitative, we visualize the data using side-by-side boxplots and numerically summarize using measures such as the sample mean, standard deviation and 5-number summary.

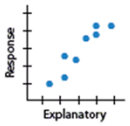
The inferential methods depend upon whether the categorical explanatory variable has two levels or more than two levels and whether the samples are dependent or independent.

For dependent samples, we only considered the case where the explanatory variable has two levels and in this case we can apply the paired t-test and estimate the population mean difference using a confidence interval. The non-parametric alternatives are the sign test and the Wilcoxon signed-rank test.

For independent samples, when the explanatory variable has two levels, we can apply the two-sample t-test as our standard method and estimate the difference in the population means using a confidence interval. The non-parametric alternative is the Wilcoxon rank-

sum test which can also be called the Mann-Whitney U test.

For independent samples where the explanatory variable has more than two levels, the standard method is ANOVA and the non-parametric alternative is the Kruskal-Wallis test.

		Response	
		Categorical	Quantitative
Explanatory	Quantitative	**Q → C** **Logistic Regression** Not covered in this course	**Q → Q** **To Visualize** Scatterplot  **Numerical Summaries** Correlation Coefficient **Formal Inference** Regression line. Significance test for the linear relationship (t-test for the slope).

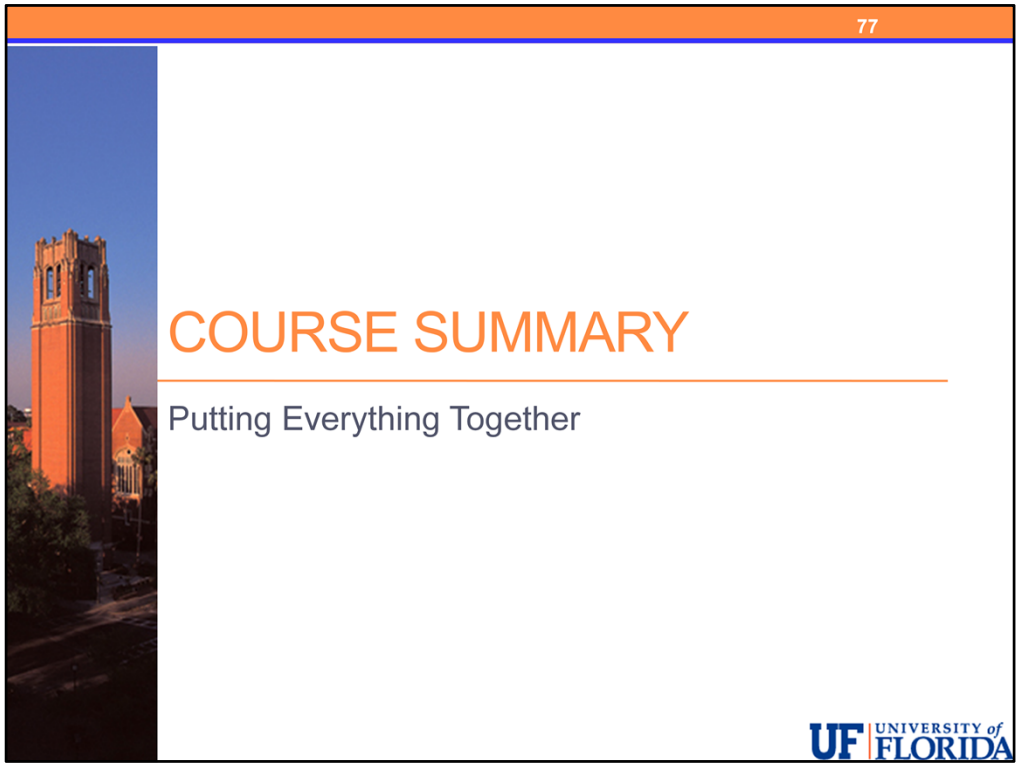
UF UNIVERSITY of FLORIDA

When the explanatory variable is quantitative and the response is categorical, we have not learned any formal methods of inference, however, we have mentioned numerous times that you can still apply the methods in Case C-Q in order to determine if there is an association. Logistic regression is the formal approach to the task of predicting a categorical outcome from a quantitative predictor.

When both the explanatory and response variables are quantitative, we visualize the relationship with a scatterplot and if the relationship is linear we can summarize numerically with the correlation. The inferential methods in this case are tests about the correlation coefficient and the slope of the linear regression equation.

In the case of non-linear relationships which are still either increasing or decreasing, we discussed one non-parametric method – Spearman's rank correlation – which can be used to measure the strength and direction of the relationship and can be tested for significance.

For each of our inferential methods, it is also very important to know the conditions under which the test can be applied. We reviewed that information in the examples but didn't restate those details in this summary.



Now that we have reached the end of the course, we hope you feel you have learned a lot about using statistics in practice in situations involving one or two variables.

We also hope that you have an understanding of and appreciation for the process of statistical inference and how probability plays an important role behind the scenes.