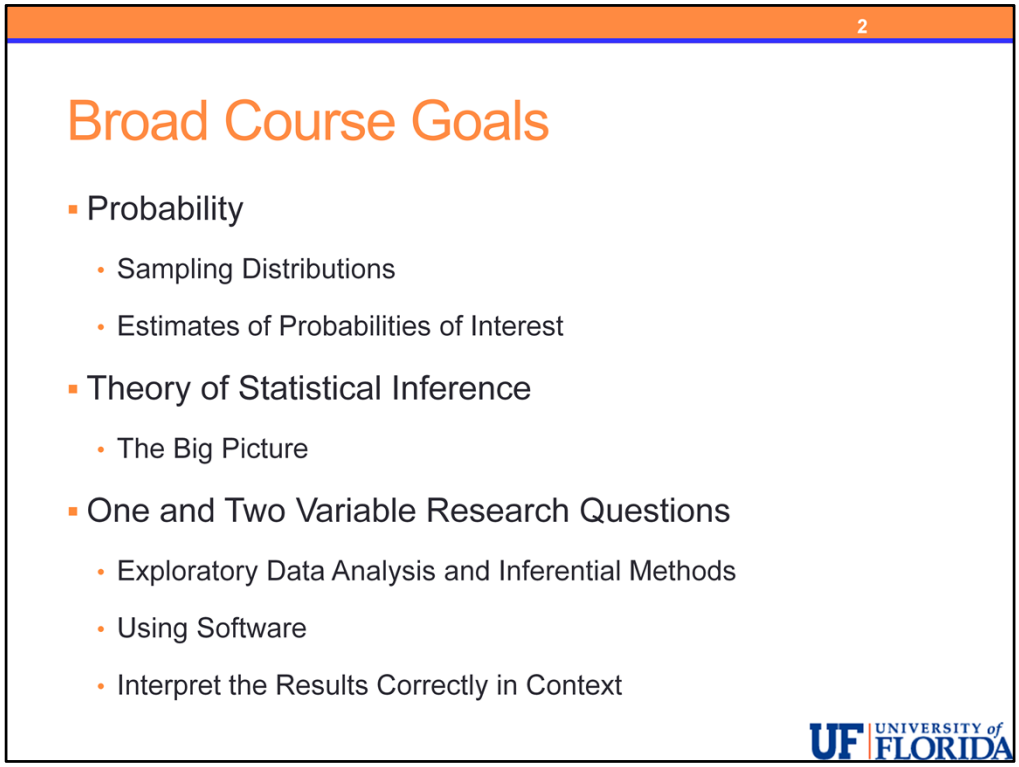
COURSE SUMMARY

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.

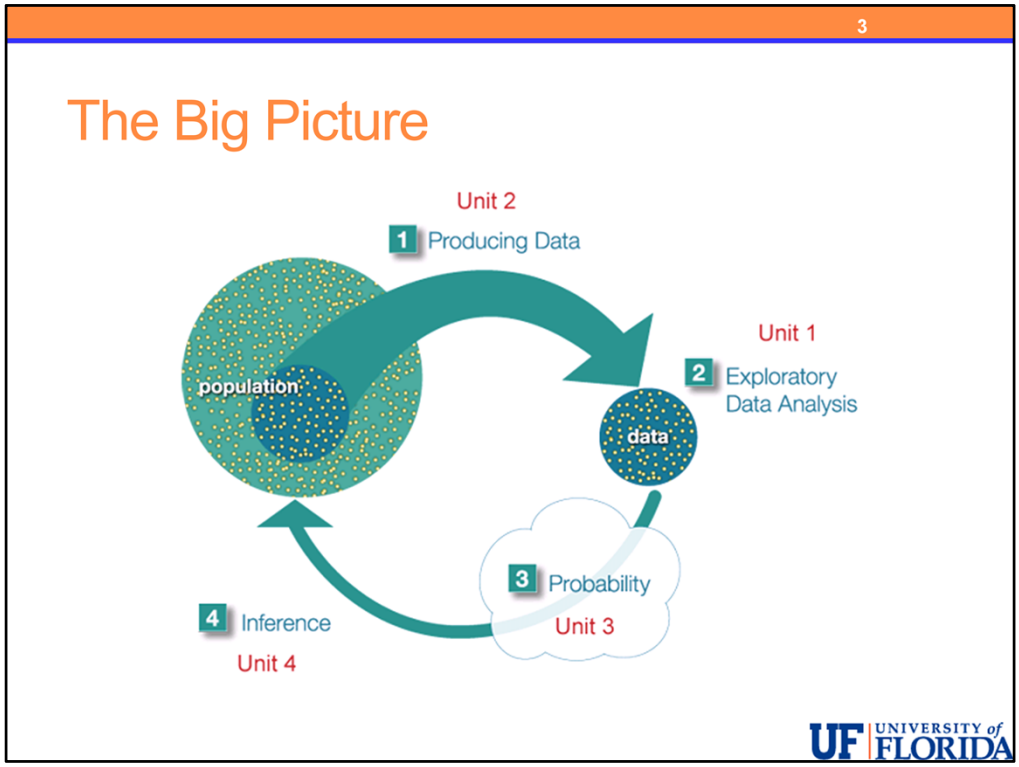


In this course we have a few major goals. We want you to

- 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.

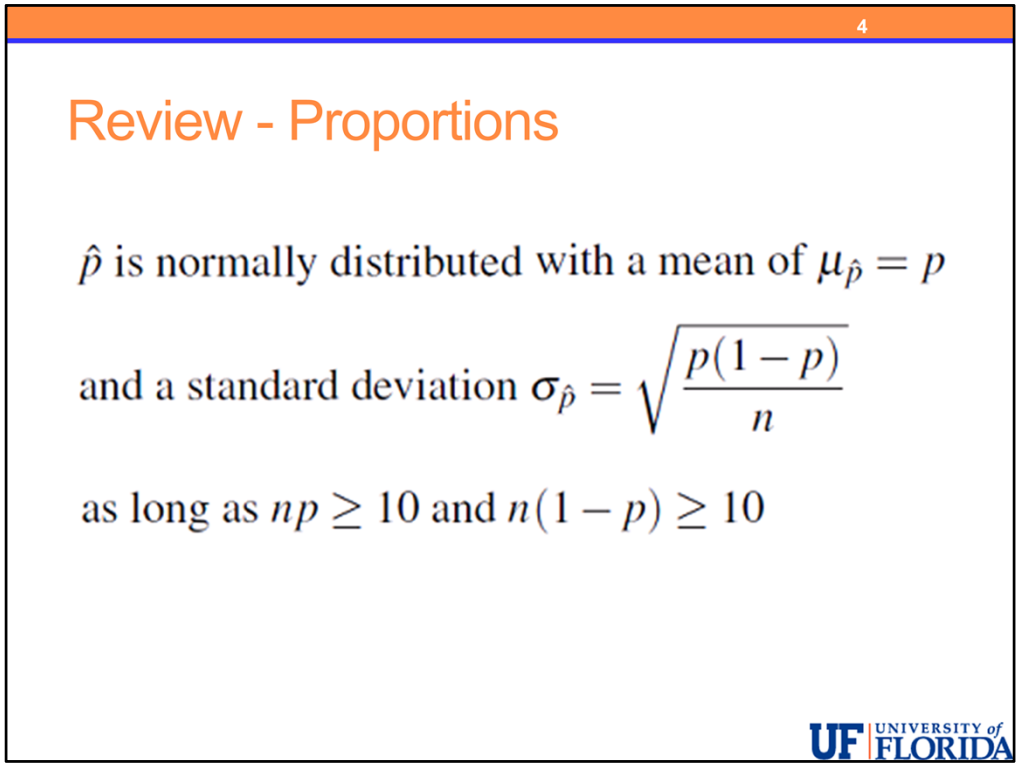
- Conduct hypothesis tests about our parameter using the data from our sample. In this case, we calculate the p-value which tells us the chance we could see a result such as ours or more extreme by random chance alone – in other words, assuming the null hypothesis is true, we find the probability that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

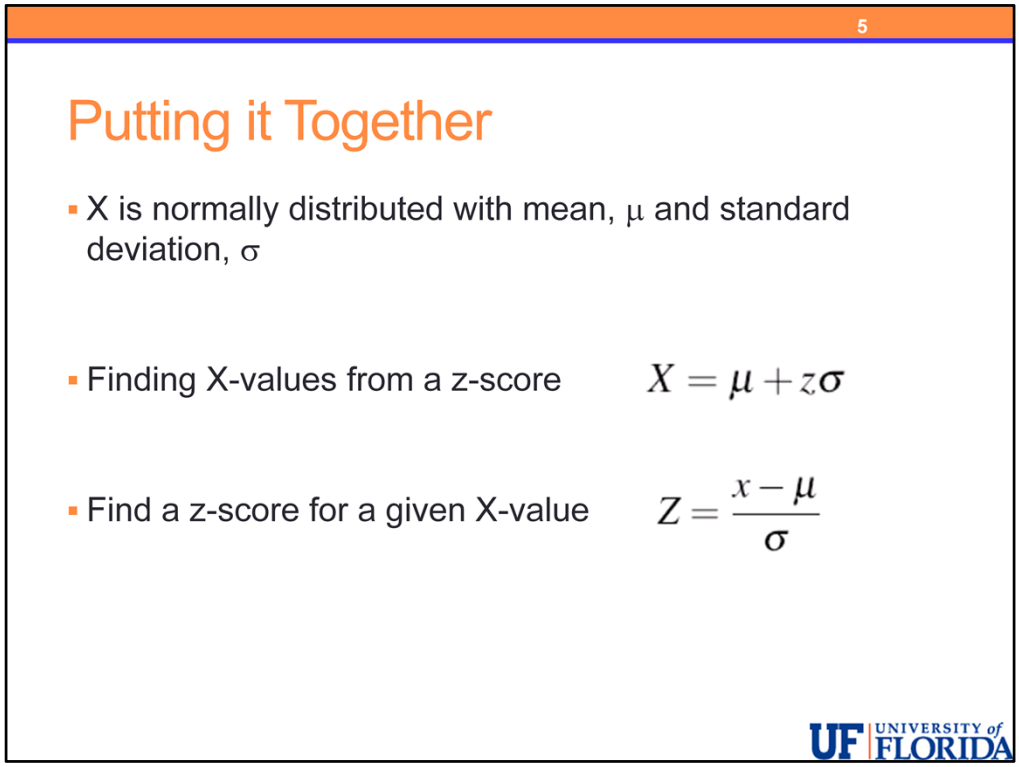
We also found that the sampling distribution of p-hat is approximately normally distributed as long as the sample size is large enough relative to the population proportion, p, specifically we need np and n(1-p) to be at least 10.

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.

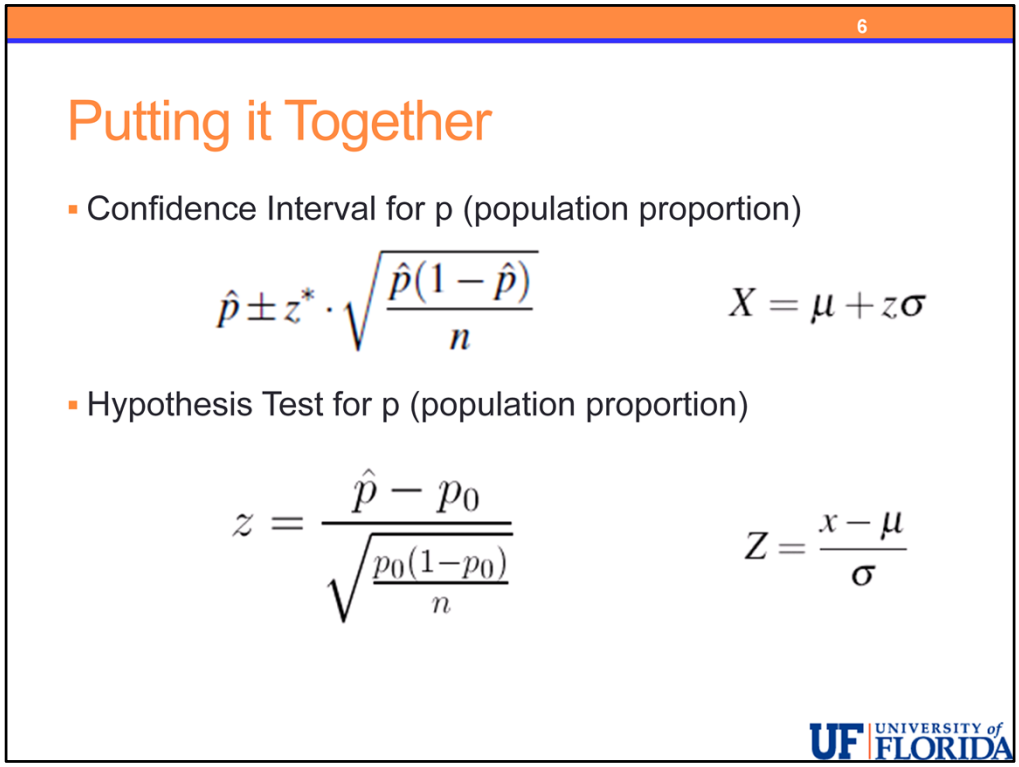
Here X represents a random variable which is normally distributed with mean mu and standard deviation sigma.

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.

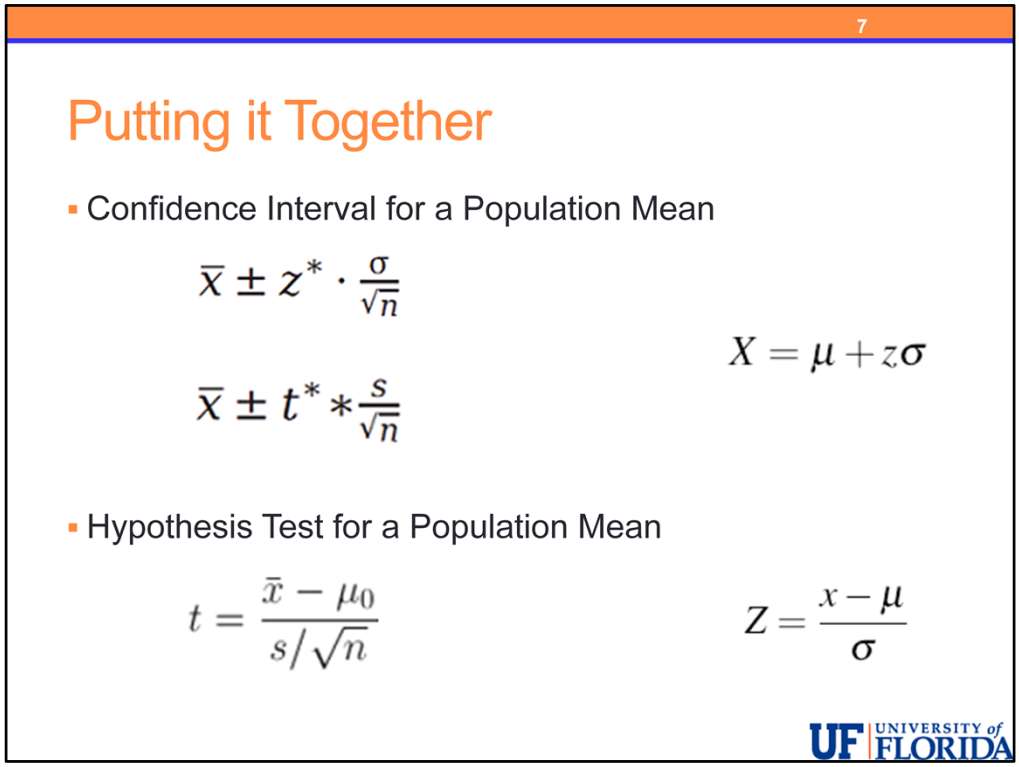
Then, after some discussion on sampling and design, we discussed random variables and normal probabilities so that we could develop the skills needed to find cut-offs for confidence intervals and p-values of hypothesis tests.

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.

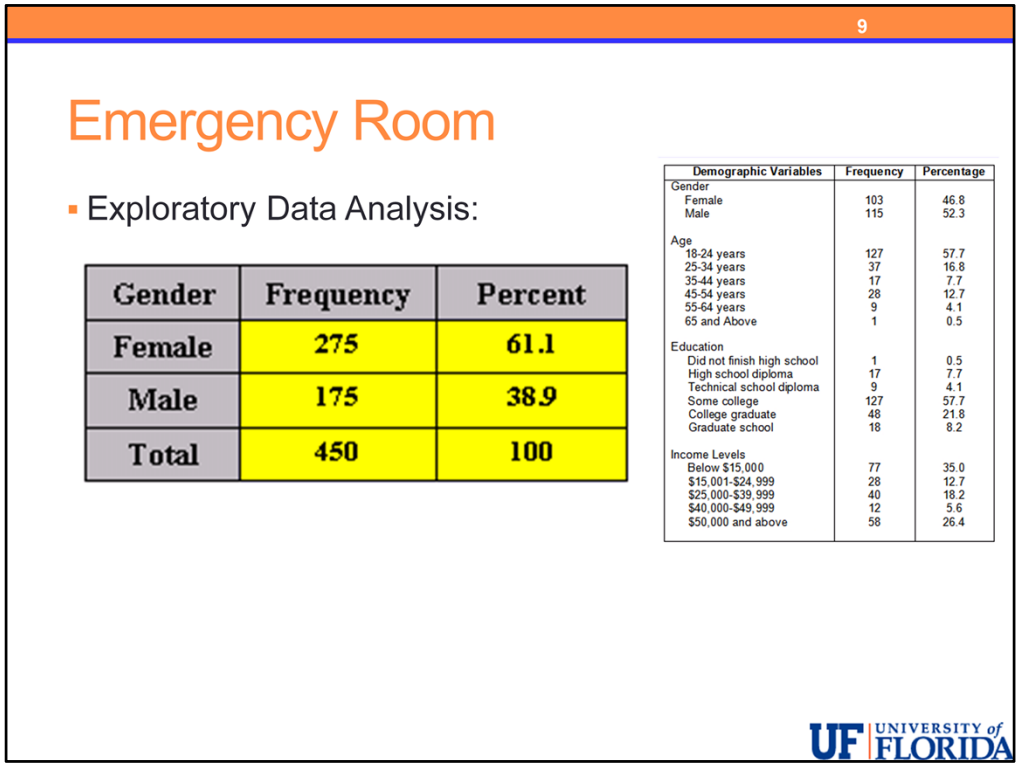
We learned in the section on estimation that if we do not know the true population standard deviation and instead substitute the sample standard deviation then the appropriate sampling distribution will be a t-distribution with n-1 degrees of freedom.

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.

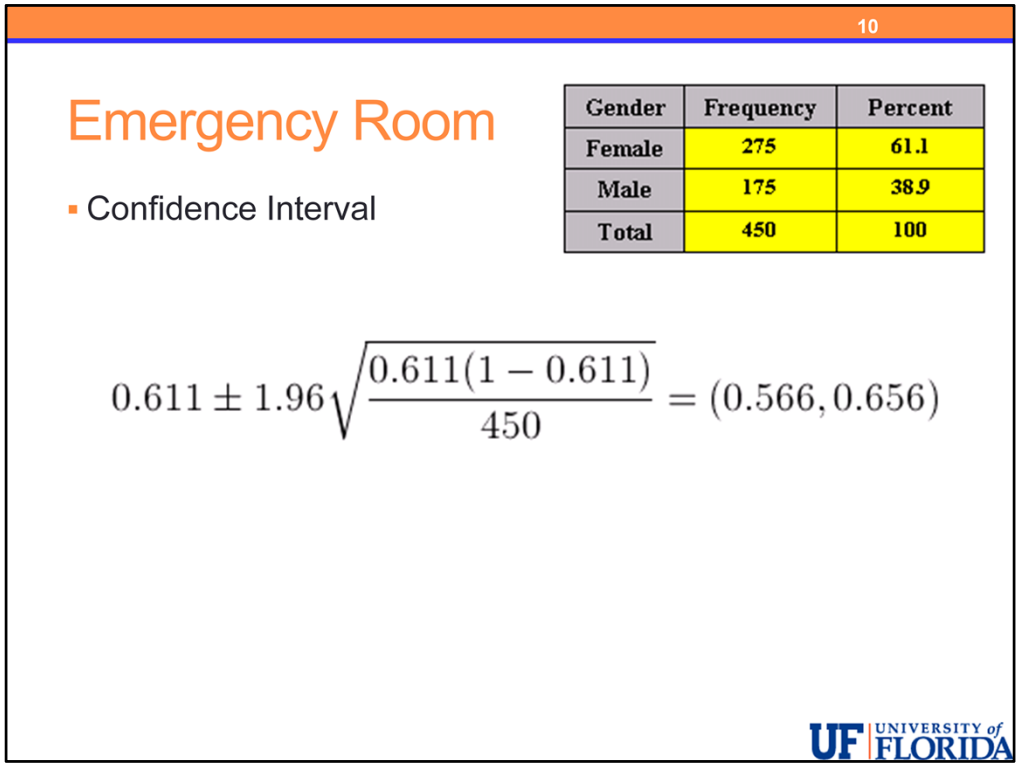
In practice, we would summarize the results presented in the frequency table in a short sentence or possibly in a large table containing this type of information for many variables in the study.

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.

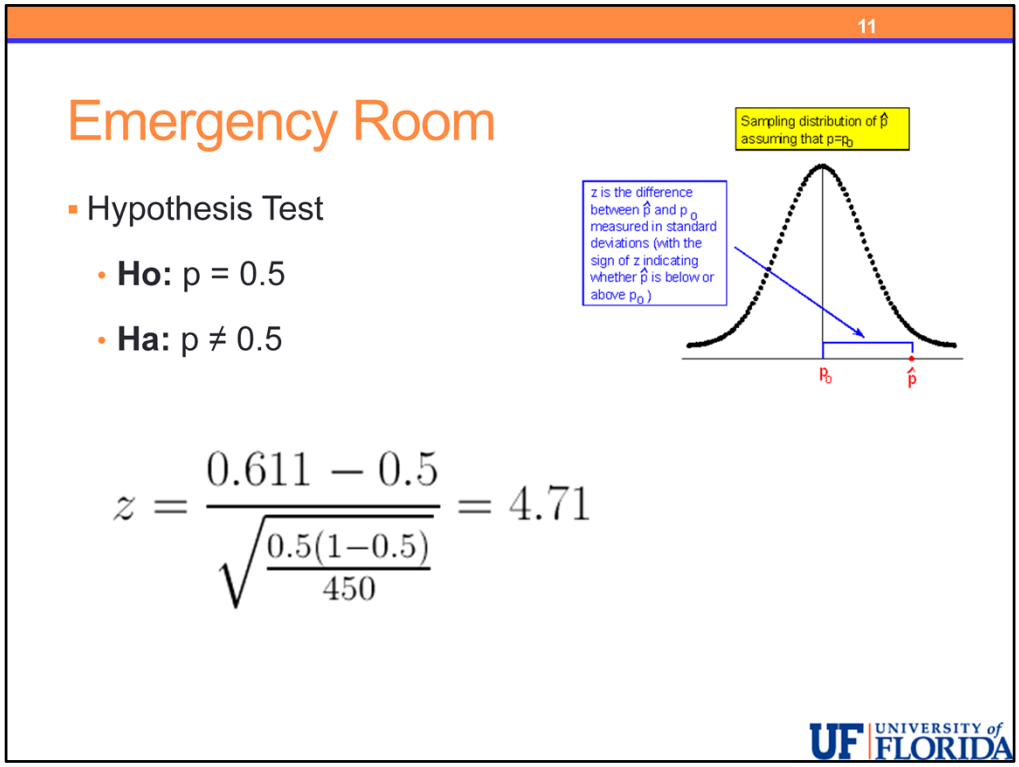
The appropriate confidence multiplier in this case is from a normal distribution due to the fact that for large enough samples, the distribution of all possible p-hats (the sampling distribution) will be approximately normally distributed.

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.

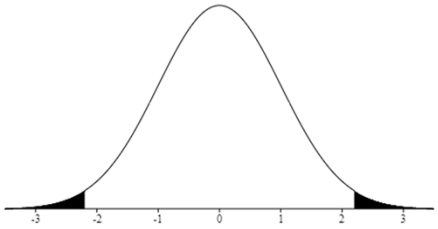
The p-value is basically zero for such a large z-score under the normal curve and thus there is enough evidence to conclude that the true proportion of female patients in this ER is not equal to 50%.

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.

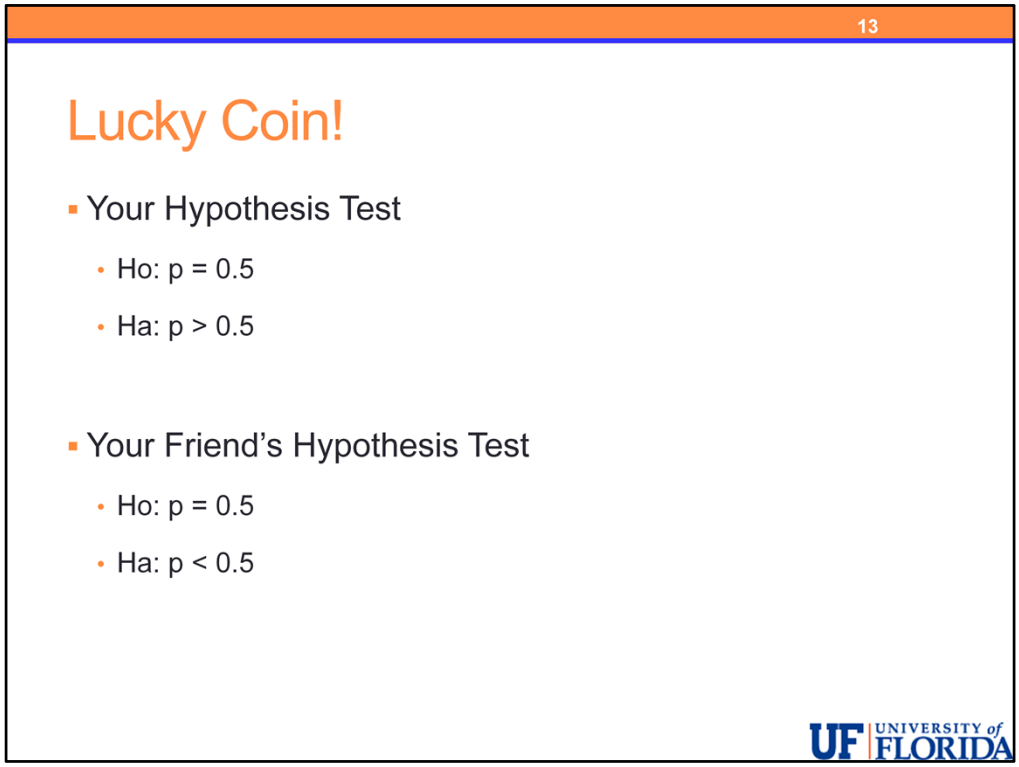
The main point is that you CANNOT change your hypotheses to a one-sided test AFTER you see your data. In this instance we cannot decide after seeing that 61% were female in our sample to change our desired alternative hypothesis to > instead of simply ≠.

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.

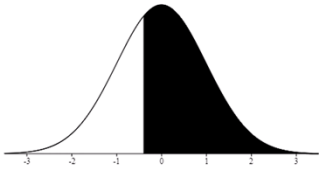
There is some anecdotal evidence to support your friend's claim that $p < 0.5$ but we will need to take into account the sampling variability in 100 tosses of a fair coin to assess if this is enough evidence to support your friend's claim.

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

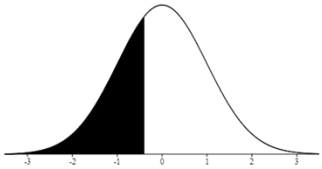
Lucky Coin!

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

- 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

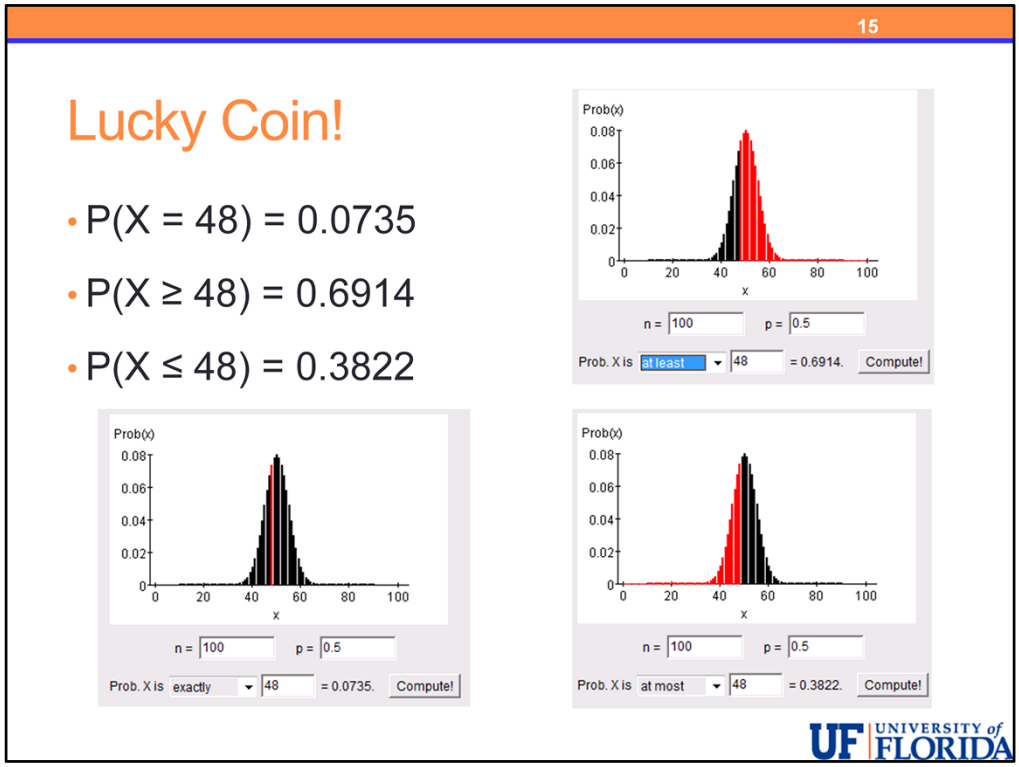
evidence to reject it … in either direction.

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.