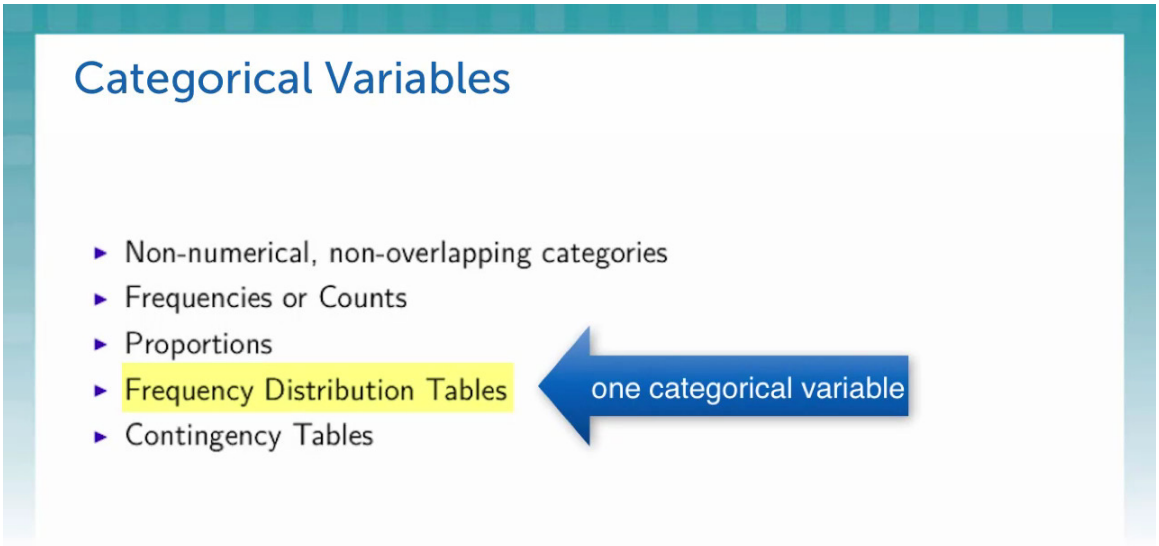
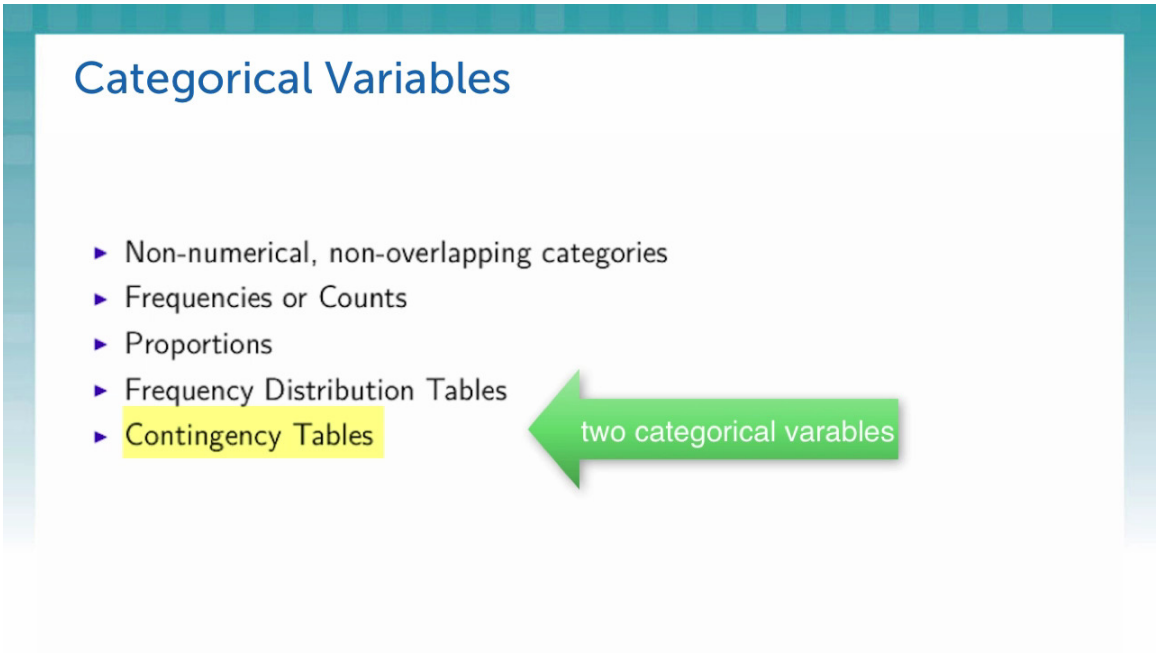

DS 705
Statistical Methods
Inference for Categorical Data

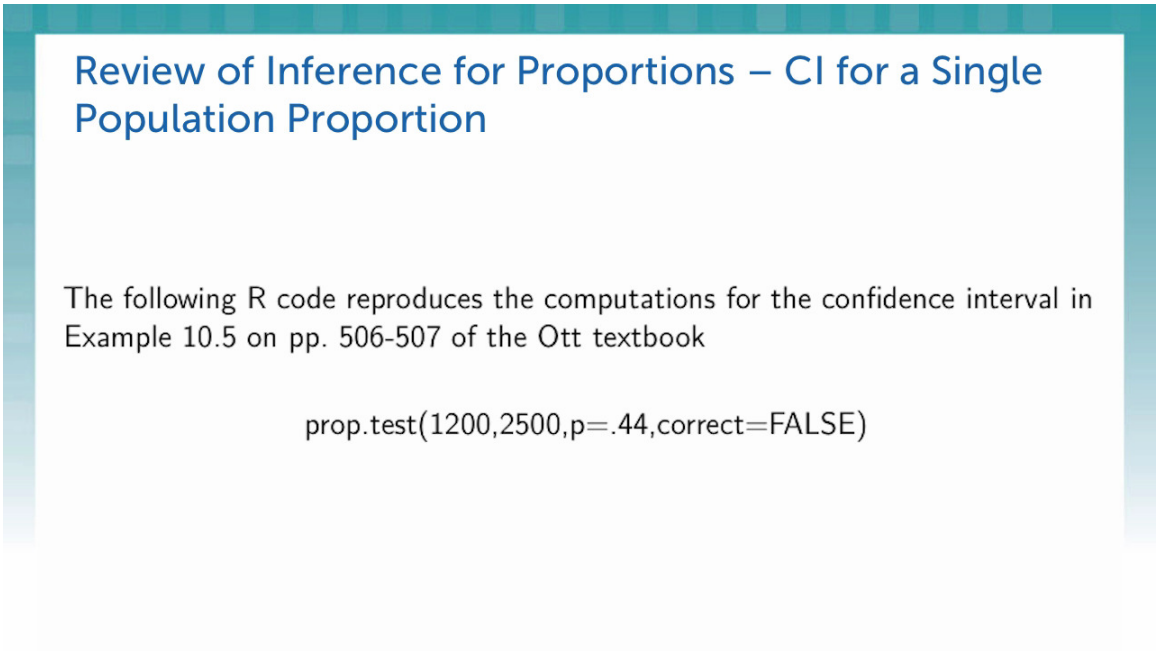




A categorical variable is a variable which takes on values from non-numerical, non-overlapping categories. These are also called qualitative variables. Rather than finding means and standard deviations, we tally up the number of observations in a sample or population that fall within each category. These are called frequencies or counts. From these, we can compute relative

frequencies, which we also call proportions. And we can also find percentages.

When summarizing just one categorical variable, the counts are placed in a frequency distribution table. The frequencies for the cross-classification of two categorical variables are placed in a contingency table.



You may want to grab your textbook to follow along with the next few slides as we review hypothesis tests and confidence intervals for one and two population proportions. The R function prob.test is used in both cases. The option Correct Equals False is turning off the Yates continuity correction, which can overcompensate with larger sample sizes. The default in R is to apply the Yates continuity correction in prob.test.

Recall also for confidence intervals that 95% confidence is the default value. So oftentimes, if that's what's requested in the textbook, you won't see that typed out in the options.

R Output for the Confidence Interval in Example 10.5

```
##
##  1-sample proportions test without continuity correction
##
## data:  1200 out of 2500
## X-squared = 16.234, df = 1, p-value = 5.599e-05
## alternative hypothesis: true p is not equal to 0.44
## 95 percent confidence interval:
##  0.4604617 0.4995996
## sample estimates:
##    p
## 0.48
```

Looking on page 507 of Ott's textbook, we can see that the confidence interval produced by R with a lower bound of 0.46 and an upper bound of 0.499, which would round the 0.50, matches exactly the confidence interval for a single population proportion in the textbook example.

Review of Inference for Proportions – HT for a Single Population Proportion

The following R code reproduces the computations for the hypothesis test in Example 10.5 on pp. 506-507 of the Ott textbook

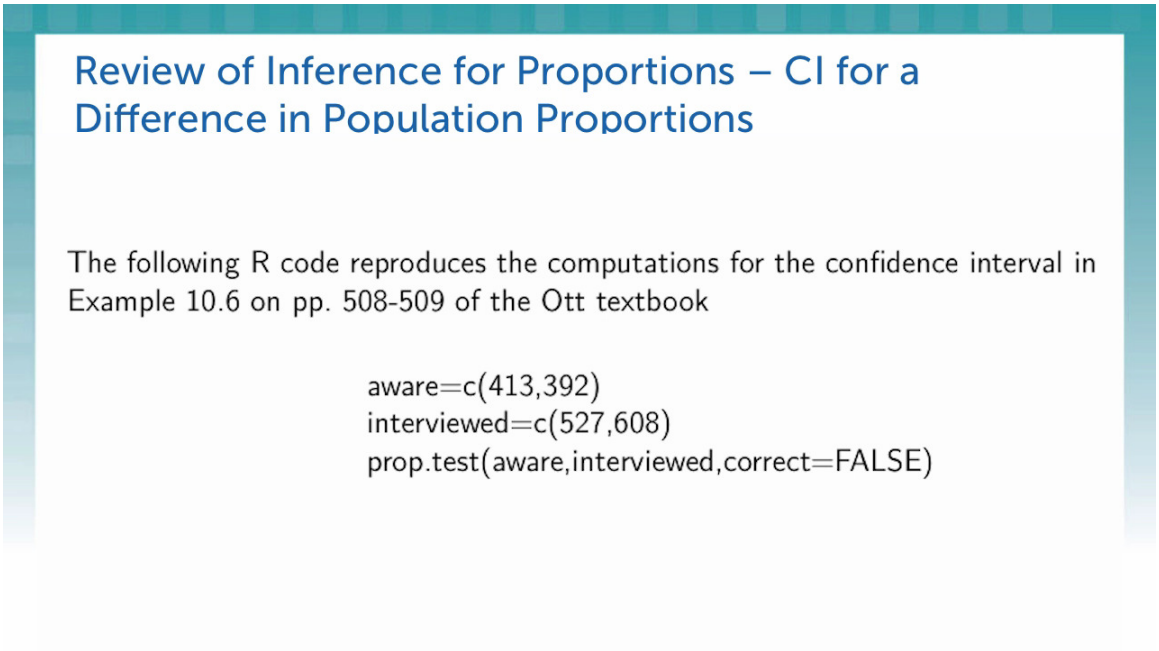
```
prop.test(1200,2500,p=.44,alternative="greater",correct=FALSE)
```

Since the hypothesis test of example 10.5 is one-sided, with the alternative hypothesis of the population proportion Pi being greater than 0.44, we specify the alternative greater in R. Note that we can simply enter the number of successes, 1,200, in the sample size, 2,500, directly into the prob.text function.

R Output for the Hypothesis Test in Example 10.5

```
##
##  1-sample proportions test without continuity correction
##
## data:  1200 out of 2500
## X-squared = 16.234, df = 1, p-value = 2.799e-05
## alternative hypothesis: true p is greater than 0.44
## 95 percent confidence interval:
##  0.4635951 1.0000000
## sample estimates:
##    p
## 0.48
```

Displayed here is the R output for the one sample test for population proportion without the Yates continuity correction. A Chi Square with one degree of freedom is a z-statistic squared. Here, R reports 16.234, the square root of is 4.03, with the difference from the textbooks z of 4.002 due only to rounding. The textbook also states the p-value is 0.00003. Here we see the p-value in scientific notation is 2.799 times 10 to the negative fifth, which when rounded is 0.00003.



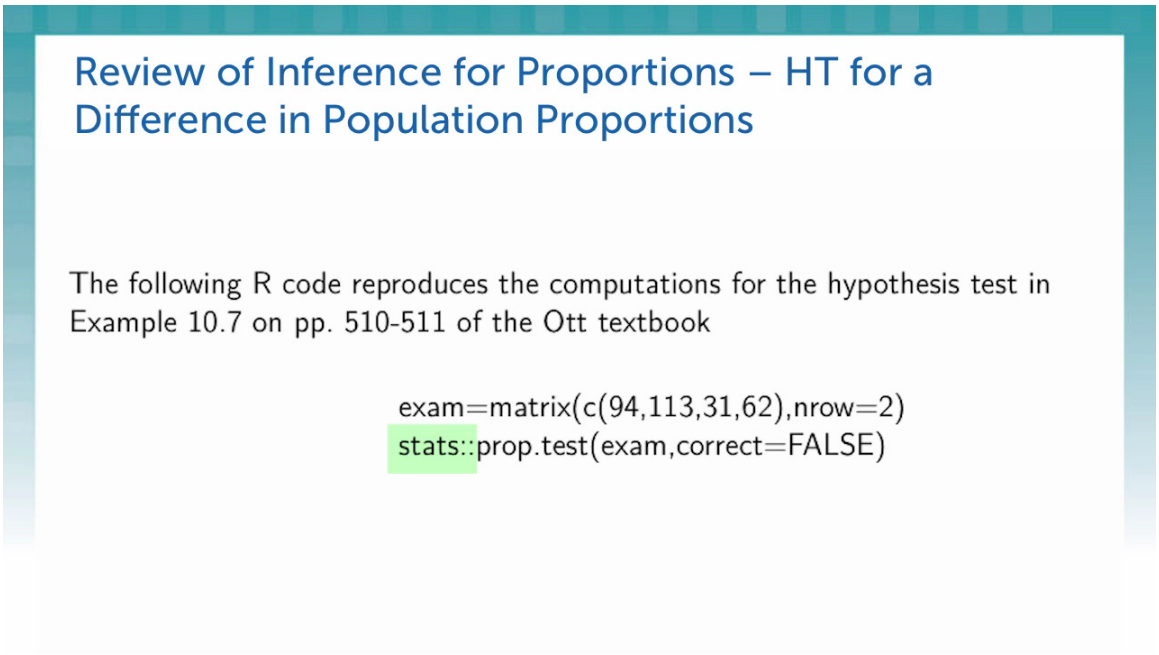
One way to enter data for either a confidence interval or a hypothesis test concerning a difference in population proportions in prob.test is as a vector of the number of successes and another vector of the corresponding sample sizes. Here, for example, 10.6 from table 10.1 on page 509 in the textbook, the number in the sample who are aware of the product are in the vector called Aware. And the sample sizes are in the vector called Interviewed.

Table 10.1 (for Example 10.6, p. 509 in Ott)

	Grand Rapids	Wichita
Number interviewed	608	527
Number aware	392	413

R Output for the Confidence in Example 10.6

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  aware out of interviewed
## X-squared = 26.429, df = 1, p-value = 2.734e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.08714759 0.19074115
## sample estimates:
##    prop 1    prop 2
## 0.7836812 0.6447368
```



In this example, I wanted to show you a different way that R can take the data. It can be entered as a 2 x 2 matrix with the two columns giving counts of successes and failures respectively. The successes go in column one. And the failures go in column two.

It was necessary to specify that we wanted the stats package here because another package that has been installed for this lesson called Mosaic also has the function called prob.test that acts a little differently. So here, we must specify which package we want to call the prob.test test from, which is the package called Stats. So the prob.test function is preceded by Stats followed by two colons.

R Output for the Contingency Table for Example 10.7

Exam Results	Computer Instruction	Traditional Instruction
Pass	94	113
Fail	31	62
Total	125	175

```
exam=matrix(c(94,113,31,62),nrow=2)
exam
```

```
##      [,1] [,2]
## [1,]   94   31
## [2,]  113   62
```

Note the matrix is entered in R so that the counts of successes are in column one. These are the ones who passed the exam an example 10.7. See table 10.2 on page 510.

R Output for the Contingency Table for Example 10.7

Exam Results	Computer Instruction	Traditional Instruction
Pass	94	113
Fail	31	62
Total	125	175

```
exam=matrix(c(94,113,31,62),nrow=2)
exam
```

```
##      [,1] [,2]
## [1,]   94   31
## [2,]  113   62
```

And accounts for the failures are in the second column. These are the ones who didn't pass the English language exam in example 10.7.

R Output for the Hypothesis Test in Example 10.7

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  exam
## X-squared = 3.8509, df = 1, p-value = 0.04972
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.002588801 0.209982628
## sample estimates:
##    prop 1    prop 2
## 0.7520000 0.6457143
```

Fisher Exact Test

The following R code reproduces the computations for the hypothesis test in Example 10.8 on pp. 512-513 of the Ott textbook.

```
count=matrix(c(38,14,4,7),nrow=2)
fisher.test(count,alternative="greater")
```

Table 10.4 for Example 10.8, p. 512 in Ott Text

	Outcome		
Drug	**Success**	**Failure**	**Total**
PV	38	4	42
P	14	7	21
Total	52	11	63

R Setup of the Contingency Table for Example 10.8

```
count=matrix(c(38,14,4,7),nrow=2)
count
```

```
##      [,1] [,2]
## [1,]   38    4
## [2,]   14    7
```

$$H_0: \pi_P \geq \pi_{PV}$$

$$H_a: \pi_P < \pi_{PV}$$

Notice that we only need to enter the inner cells of the 2 x 2 table, not the row and column totals in the margins of the table. R will compute them internally and use them as needed to compute the p-value for the Fisher exact test.

R Setup of the Contingency Table for Example 10.8

```
count=matrix(c(38,14,4,7),nrow=2)
count
```

```
##      [,1] [,2]
## [1,]   38    4
## [2,]   14    7
```

H_0: $\pi_P \geq \pi_{PV}$

H_a: $\pi_P < \pi_{PV}$

If you're looking at the hypotheses on the bottom page 512, you'll notice that the alternative says that the proportion for drug P indicated by Pi sub p, is less than the proportion for drug PV. But you see here, we have specified the alternative greater in the R code. This is because the drug PV outcomes are listed in the first row of the 2 x 2 table. Be careful with one-sided tests to code them in the right direction.

R Output for the Fisher Exact Test in Example 10.8

```
##
##  Fisher's Exact Test for Count Data
##
## data:  count
## p-value = 0.02537
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  1.22629     Inf
## sample estimates:
## odds ratio
##   4.615064
```

As the textbook states, Fisher's exact test computes the p-value as the sum of the probabilities for all the tables having 38 or more successes for the drug PV. Also, testing that the proportion of successes for PV is greater than for drug P is equivalent to saying that the odds ratio is greater than 1.

We'll get to odds ratios a bit later in these slides.