Inference for Categorical Data

Categorical Variables

- Non-numerical, non-overlapping categories
- Frequencies or Counts
- Proportions
- Frequency Distribution Tables
- Contingency Tables

Categorical Variables

- Non-numerical, non-overlapping categories
- Frequencies or Counts
- Proportions
- Frequency Distribution Tables
- Contingency Tables

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.

Fast Facts: One-Sample Z Procedures for a Proportion

Why: Hypothesis test - To *compare* an unknown population proportion to some hypothetical value.
Confidence Interval - To *estimate* an unknown population proportion.

When: The following conditions are necessary for these procedures to be accurate and valid.

1. The sample is selected randomly
2. The sample contains at least 10 successes and 10 failures

How: Use R function **prop.test()**

Fast Facts: One-Sample Z Procedures for a Proportion

Why: Hypothesis test - To *compare* an unknown population proportion to some hypothetical value.
Confidence Interval - To *estimate* an unknown population proportion.

When: The following conditions are necessary for these procedures to be accurate and valid.

1. The sample is selected randomly
2. The sample contains at least 10 successes and 10 failures

How: Use R function **prop.test()**

No audio

Review of Inference for Proportions - CI for a Single Population Proportion

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

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

Review of Inference for Proportions - CI for a Single Population Proportion

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

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

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 prop.test is used both cases.

The option correct=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 prop.test.

R output for the confidence interval in Example 10.5, pp. 506-507

```
##
##  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
```

R output for the confidence interval in Example 10.5, pp. 506-507

```
##
##  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 produce by R with a lower bound of 0.46 and an upper bound of .499, which would round to .50, matches exactly the confidence interval for a single population proportion in the textbook example.

bottom panel note: the R function binom.test does the same, only provides the interval or test based on the exact binomial distribution rather than the normal approximation

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)
```

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 .44, we specify the alternative greater in R.

Note that we can simply enter the number of successes, 1200, and the sample size, 2500, directly into the prop.test function.

bottom panel note: Enter ?prop.test in R to see more

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
```

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
```

Here is the R output for the one-sample test for a population proportion without the Yates' continuity correction. Chi-square with 1 df is z squared (here R reports 16.234, the square root of which is 4.03 - with the difference from the textbook's z of 4.00 due to rounding). The textbook states the p-value as .00003 - here we see the p-value in scientific notation as 2.799 times 10 to the negative 5th - which when rounded, is .00003

Fast Facts: Two-Sample Z Procedures for Proportions

Why: Hypothesis test - To *compare* two unknown population proportions.
Confidence Interval - To *estimate* the difference between two unknown population proportions.

When: The following conditions are necessary for these procedures to be accurate and valid.

1. The sample is selected randomly
2. The samples are selected independently
3. Both samples contains at least 10 successes and 10 failures

How: Use R function **prop.test()**

Fast Facts: Two-Sample Z Procedures for Proportions

Why: Hypothesis test - To *compare* two unknown population proportions. Confidence Interval - To *estimate* the difference between two unknown population proportions.

When: The following conditions are necessary for these procedures to be accurate and valid.

1. The sample is selected randomly
2. The samples are selected independently
3. Both samples contains at least 10 successes and 10 failures

How: Use R function **prop.test()**

No audio

Review of Inference for Proportions - CI for a Difference in Population Proportions

The following R code reproduces the computations for the confidence interval in Example 10.6 on pp. 508-509 of the Ott textbook

```
aware=c(413,392)
interviewed=c(527,608)
prop.test(aware,interviewed,correct=FALSE)
```

Review of Inference for Proportions - CI for a Difference in Population Proportions

The following R code reproduces the computations for the confidence interval in Example 10.6 on pp. 508-509 of the Ott textbook

```
aware=c(413,392)
interviewed=c(527,608)
prop.test(aware,interviewed,correct=FALSE)
```

One way to enter data for either a confidence interval or a hypothesis test concerning a difference in population proportions in prop.test is as a vector of the number of successes and a 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

Figure 1:

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
```

Review of Inference for Proportions - HT for a Difference in Population Proportions

The following R code reproduces the computations for the hypothesis test in Example 10.7 on pp. 510-511 of the Ott textbook

```
exam=matrix(c(94,113,31,62),nrow=2)
stats::prop.test(exam,correct=FALSE,alternative='greater')
```

Review of Inference for Proportions - HT for a Difference in Population Proportions

The following R code reproduces the computations for the hypothesis test in Example 10.7 on pp. 510-511 of the Ott textbook

```
exam=matrix(c(94,113,31,62),nrow=2)
stats::prop.test(exam,correct=FALSE,alternative='greater')
```

In this example I wanted to show you a different way that R can take the data. It can be entered as a 2x2 matrix with the two columns giving counts of successes and failures, respectively. The successes go in column 1, the failures go in column 2.

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 a function called prop.test - which behaves a little differently, so here we must specify which package we want to call the prop.test from, which is the package called stats, so the prop.test function is preceded by stats followed by two colons.

bottom panel note: Counts are from Table 10.2 on p. 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

Figure 2:

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

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

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

Figure 2:

```
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 1 - these are the ones who passed the exam in Example 10.7, see Table 10.2 on page 510 - and the counts for failures are in the second column - these are the ones who didn't pass the English language exam in Example 10.7.

bottom panel note: See Table 10.2 on page 510 of the Ott textbook

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 = 0.02486
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.01926052 1.00000000
## 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")
```

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")
```

Bottom panel note: Fisher's Exact test is used when at least one of the expected cell counts in a 2x2 table is under 5.

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

Figure 3:

R setup of the contingency table for Example 10.8

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

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

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

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

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
```

$$H_0\colon \pi_P \geq \pi_{PV}$$
$$H_a\colon \pi_P < \pi_{PV}$$

Notice that we only need to enter the inner cells of the 2x2 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.

If you're looking at the hyotheses on the bottom of page 512, You'll notice that the alternative says the proportion for drug P (indicated by pi_P) is LESS than the proportion for drug PV, but recall that in the R code we had specified the alternative "greater." This is because the drug PV outcomes are listed in the first row of the 2x2 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
```

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 tables having 38 or more successes for the drug PV.

Also, testing the that proportion of successes for PV is greater than for drug P is equivalent to saying the odds ratio is greater than 1. We'll get to odds ratios a bit later in these slides.

Chi-Square Tests

- One categorical variable
  - Goodness-of-fit test
- Two categorical variables
  - Test for independence
  - Test for homogeneity

Chi-Square Tests

- One categorical variable
  - Goodness-of-fit test
- Two categorical variables
  - Test for independence
  - Test for homogeneity

When just one categorical variable is under consideration, the chi-square test for goodness-of-fit can be used to test the hypothesis that the sample was drawn from a specified distribution vs the alternative that is was not. You may recall that the Shapiro-Wilk test for normality is also a goodness-of-fit test.

For two categorical factors, the chi-square statistic can be used to test for the independence of the two factors vs the alternative that the factors are associated. With the test for independence, the sampling scheme must be that a random sample has been drawn from the population of interest, thus making the row and column totals random counts.

The chi-square test for homogeneity has identical computations for

Fast Facts: Chi-Square Test for Goodness-of-Fit

Why: To determine whether or not a sample was drawn from a particular distribution with hypothesized proportions for specified categories.

When: The following conditions are necessary for this procedure to be accurate and valid.

1. The samples are selected randomly
2. The sample is large enough so that the expected cell frequencies are all at least 5

How: Use R function **chisq.test()**

Fast Facts: Chi-Square Test for Goodness-of-Fit

Why: To determine whether or not a sample was drawn from a particular distribution with hypothesized proportions for specified categories.

When: The following conditions are necessary for this procedure to be accurate and valid.

1. The samples are selected randomly
2. The sample is large enough so that the expected cell frequencies are all at least 5

How: Use R function **chisq.test()**

No audio

Example A: Chi-square GOF test

Suppose it is reported in a media release that 24% of all personal loans are for home mortgages, 38% were for automobile purchases, 18% were for credit card loans, and the rest were for other types of loans. Records for a random sample of 55 loans was obtained and each was classified into one of these categories. The results are in the following table.

	Mortgage	Auto	Credit	Other
Number of loans	24	21	6	4

GOF Test: The Request

Conduct the appropriate test to determine if the distribution reported in the media release for the frequency of the types of loans fits the actual distribution of types loans in the population. Use $\alpha = 0.01$.

GOF Test: The Hypotheses

H_0: $\pi_{Mortgage} = 0.24, \pi_{Auto} = 0.38, \pi_{Credit} = 0.18, \pi_{Other} = 0.20$

H_a: At least one π_i differs from its hypothesized value

GOF Test: The Hypotheses

H_0: $\pi_{Mortgage} = 0.24, \pi_{Auto} = 0.38, \pi_{Credit} = 0.18, \pi_{Other} = 0.20$
H_a: At least one π_i differs from its hypothesized value

Verbally, the null hypothesis is claiming that the distribution claimed by the media release is correct. The alternative hypothesis is simply that the distribution is not correct since at least one of the hypothesized probabilities is not right.

Many times, the chi square goodness-of-fit test is used to determine if the categories have equal probabilities - like testing to see if a die is fair, for example. In those cases it isn't necessary to specify the proportions because they are self evident. If a 6-sided die is equally balanced, then each outcome should have a probability of 1 out of 6. If we were testing to see if the proportions of loans were equally likely here, the null hypothesis probabilities would all be one fourth, since there are 4 categories.

GOF Test: Getting the Data into R

```
observed=c(24,21,6,4)
proportions=c(.24,.38,.18,.20)
```

GOF Test: Getting the Data into R

```
observed=c(24,21,6,4)
proportions=c(.24,.38,.18,.20)
```

We simply create a vector that contains the observed cell counts, here I named it "observed" and a vector holding the hypothesized proportions, which I called "proportions."

You have probably noticed by now that we are using the terms proportions and probabilties interchangeably.

In this test our presumption is that the underlying variable has a multinomial probability distribution with the probabilities specified in the null hypothesis. Multinomial distributions are characterized by having n identical, independent trials, each having k possible outcomes, where the probabilities of each of the k outcomes remains constant from trial to trial.

GOF Test: Getting the Test Statistic & P-value in R

```r
chisq.test(x=observed,p=proportions)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 14.828, df = 3, p-value = 0.00197
```

GOF Test: Getting the Test Statistic & P-value in R

```
chisq.test(x=observed,p=proportions)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 14.828, df = 3, p-value = 0.00197
```

One quick check to see that we have coded it right is to look at the degrees of freedom. It should be the number of categories minus 1. Since there were 4 loan categories being tested and we see the degrees of freedom given as 3, we should start to get warm fuzzies about now.

What should we conclude? Was the media report correct? No, according to the sample data resulting in a test statistic of 14.828 and a p-value of .00197, which is less than .01, we should reject the null hypothesis and claim that the actual distribution for the types of personal loans is different from what was reported.

GOF Test: Checking Expected Values in R

```r
55*proportions
```

```
## [1] 13.2 20.9  9.9 11.0
```

GOF Test: Checking Expected Values in R

```
55*proportions
```

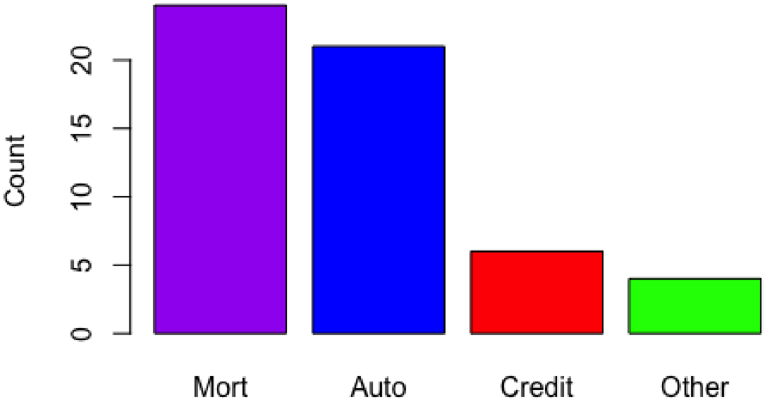
```
## [1] 13.2 20.9  9.9 11.0
```

With a smaller sample like this one, it would behoove us to check the sample size requirement for this chi-square test. You see the expected cell frequencies are easily obtained by multiplying the vector of hypothesized proportions by the sample size.

Notice also that the requirement isn't that the observed counts are all at least 5, but that the expected counts are all at least 5. So even though there was an observed cell frequency of 4 here, our sample was still large enough to trust the chi-square test for goodness-of-fit here, at least we can trust it to the extent that we didn't just make a Type 1 error - which was controlled at the 1% level of significance in this test.

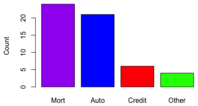
Making a Bar Graph for One Categorical Variable

```r
observed=c(24,21,6,4)
barplot(observed,names.arg=c("Mort","Auto","Credit","Other"),
        ylab="Count",col=c("purple","blue","red","green"))
```



Making a Bar Graph for One Categorical Variable

```
observed=c(24,21,6,4)
barplot(observed,names.arg=c("Mort","Auto","Credit","Other"),
        ylab="Count",col=c("purple","blue","red","green"))
```



No audio.

Bottom Panel Note:

See the .Rmd file for the code.

What About Data in a Larger Data File?

The HealthExam data frame contains a variety of both quantitative and categorical variables for 80 patients. In the file, the variable Region indicates the region of the U.S. for each of the 80 patients.

Consider the following table of counts for patients falling in the four regions of the U.S.

```r
data("HealthExam")
table(HealthExam$Region)
```

```
##
##   Midwest Northeast     South      West
##        16        22        20        22
```

What About Data in a Larger Data File?

The HealthExam data frame contains a variety of both quantitative and categorical variables for 80 patients. In the file, the variable Region indicates the region of the U.S. for each of the 80 patients.

Consider the following table of counts for patients falling in the four regions of the U.S.

```
data("HealthExam")
table(HealthExam$Region)
```

```
##
##   Midwest Northeast     South      West
##        16        22        20        22
```

No audio.

Bottom panel note:
You should load the HealthExam data file into your R session and take a look at it. It is in the DS705Data package.

GOF Test on Variable From Larger Data File

Test that the regions are equally represented in the population. Use $\alpha = 0.05$.

H_0: $\pi_{Midwest} = 0.25, \pi_{Northeast} = 0.25, \pi_{South} = 0.25, \pi_{West} = 0.25$
H_a: *At least one π_i differs from its hypothesized value*

```r
observed=table(HealthExam$Region)  # get observed cell counts
proportions=c(.25,.25,.25,.25)  # specify proportions from H0
chisq.test(x=observed,p=proportions) # test for goodness-of-fit
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 1.2, df = 3, p-value = 0.753
```

GOF Test on Variable From Larger Data File

Test that the regions are equally represented in the population. Use $\alpha = 0.05$.

H_0: $\pi_{Midwest} = 0.25, \pi_{Northeast} = 0.25, \pi_{South} = 0.25, \pi_{West} = 0.25$
H_a: *At least one* π_i *differs from its hypothesized value*

```
observed=table(HealthExam$Region)  # get observed cell counts
proportions=c(.25,.25,.25,.25)  # specify proportions from H0
chisq.test(x=observed,p=proportions) # test for goodness-of-fit
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 1.2, df = 3, p-value = 0.753
```

If you are pulling counts for a categorical variable out of a larger data frame with the table function, you can easily store the counts in an object and place them into the chisq.test function exactly as if you specified them yourself. You will, however, have to provide hypothesized proportions for the goodness-of-fit test. In this case, we are testing to see of the proportions are all equal to each other, so since there are 4 categories, each hypothetical proportion will be 0.25.

With a p-value of 0.753, H0 will not be rejected at a 5% level of significance. There is not enough evidence to conclude that any of the proportions for the geographic regions differ from 0.25 in the population that this sample represents.

Fast Facts: Chi-Square Test for Independence

Why: To determine if two categorical variables (called factors) are associated or independent.

When: The following conditions are necessary for this procedure to be accurate and valid.

1. The samples are selected randomly
2. The sample is large enough so that the expected cell frequencies are all at least 5

How: Use R function **chisq.test()**

Fast Facts: Chi-Square Test for Independence

Why: To determine if two categorical variables (called factors) are associated or independent.

When: The following conditions are necessary for this procedure to be accurate and valid.

1. The samples are selected randomly
2. The sample is large enough so that the expected cell frequencies are all at least 5

How: Use R function **chisq.test()**

No audio

Example B: Health Exam Data

The Age Group and Region for the first 6 out of 80 subjects is as follows

```
##   HealthExam.AgeGroup HealthExam.Region
## 1            36 to 64              West
## 2            36 to 64             South
## 3                 65+           Midwest
## 4            36 to 64              West
## 5            36 to 64         Northeast
## 6                 65+           Midwest
```

Example B: Health Exam Data

The Age Group and Region for the first 6 out of 80 subjects is as follows

```
##   HealthExam.AgeGroup HealthExam.Region
## 1            36 to 64              West
## 2            36 to 64             South
## 3                 65+           Midwest
## 4            36 to 64              West
## 5            36 to 64         Northeast
## 6                 65+           Midwest
```

Instead of having the counts as basic summary statistics for our categorical variables, we may have a large data frame that contains the individual observations. That's OK. R will know just what to do with them and they can be entered into the chisq.test function in the same way as the vectors or matrices containing the frequencies.

Example B: Health Exam Data Contingency Table

To see the crosstabs, use the 'table' function in R

```r
tbl <- with(HealthExam,table(AgeGroup,Region))
addmargins(tbl)
```

```
##           Region
## AgeGroup   Midwest Northeast South West Sum
##   18 to 35       6         9     5    8  28
##   36 to 64       4         7    13    8  32
##   65+            6         6     2    6  20
##   Sum           16        22    20   22  80
```

Example B: Health Exam Data Contingency Table

To see the crosstabs, use the 'table' function in R

```
tbl <- with(HealthExam,table(AgeGroup,Region))
addmargins(tbl)
```

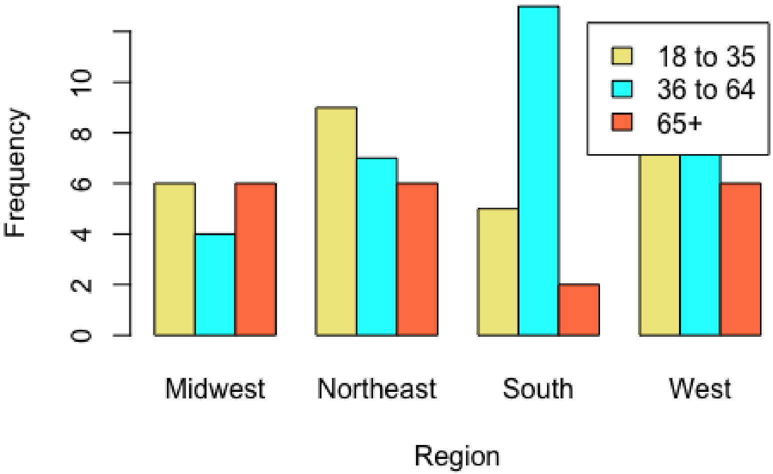
```
##           Region
## AgeGroup   Midwest Northeast South West Sum
##   18 to 35       6         9     5    8  28
##   36 to 64       4         7    13    8  32
##   65+            6         6     2    6  20
##   Sum           16        22    20   22  80
```

When your data comes as individual observations in a data frame, it is a good idea to just look at the counts to get a feel for what relationship might exist between the factors and to make sure that there aren't any unexpected surprises in your data set.

NEW AUDIO: Note that the addmargins function will display the row and column totals.

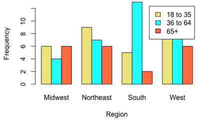
Making a Bar Graph for Two Categorical Variables

```
barplot(tbl,xlab="Region", ylab="Frequency",
        col=c("khaki","cyan","coral"),legend=rownames(tbl),beside=T)
```



Making a Bar Graph for Two Categorical Variables

```
barplot(tbl,xlab="Region", ylab="Frequency",
        col=c("khaki","cyan","coral"),legend=rownames(tbl),beside=T)
```



No audio.

Bottom panel note:

R has many fun colors to choose from! Search the web for the names.

R Code for Tests of Independence or Homogeneity

Whether it is a test for independence or homogeneity, the R code is the same.

```
chisq.test(AgeGroup,Region,data=HealthExam)
```

R Code for Tests of Independence or Homogeneity

Whether it is a test for independence or homogeneity, the R code is the same.

```
chisq.test(AgeGroup,Region,data=HealthExam)
```

The chisq.test function can be used with vectors or matrices containing the contingency table frequencies in the same way that was shown for the prop.test function previously in this presentation.

However, when categorical data is listed out in a data frame, the variables can be loaded directly into the chisq.test function by their names in the data frame.

NEW Bottom Panel Note:

A table can be stored and used directly in the chisq.test() function as chisq.test(tbl).

Example B: Health Exam Output from chisq.test

Since the 80 people selected in this study randomly fell into the age categories and geographic regions, the chi-square test here is for independence (not homogeneity).

```
##
##  Pearson's Chi-squared test
##
## data:  HealthExam$AgeGroup and HealthExam$Region
## X-squared = 8.188, df = 6, p-value = 0.2247
```

Chi-square test for Health Exam data

H_0: Age Group and Region are independent.
H_a: Age Group and Region are associated.

Conclusion: Do not reject H_0 at $\alpha = 0.05$. There is insufficient evidence in this sample to claim that Age Group and Region are associated for the population of U.S. adults ($P = 0.2247$).

Chi-square test for Health Exam data

H_0: Age Group and Region are independent.
H_a: Age Group and Region are associated.

Conclusion: Do not reject H_0 at $\alpha = 0.05$. There is insufficient evidence in this sample to claim that Age Group and Region are associated for the population of U.S. adults ($P = 0.2247$).

You see the conclusion here is to not reject the null hypothesis . . .But wait! some of those cell counts were pretty small - we should check the expected cell counts to see if any are under 5.

Expected Cell Counts for Health Exam data

```r
result=chisq.test(HealthExam$AgeGroup,HealthExam$Region)
result$expected
```

```
##                    HealthExam$Region
## HealthExam$AgeGroup Midwest Northeast South West
##            18 to 35     5.6       7.7     7  7.7
##            36 to 64     6.4       8.8     8  8.8
##            65+          4.0       5.5     5  5.5
```

Expected Cell Counts for Health Exam data

```r
result=chisq.test(HealthExam$AgeGroup,HealthExam$Region)
result$expected
```

```
##                    HealthExam$Region
## HealthExam$AgeGroup Midwest Northeast South West
##            18 to 35     5.6       7.7     7  7.7
##            36 to 64     6.4       8.8     8  8.8
##            65+          4.0       5.5     5  5.5
```

To get the expected cell counts you see that its necessary to assign the chisq.test output to an object in R and then call from that object the expected values using this code here "result dollar sign expected."

Do you see that the expected cell frequency for the 65 and over age group in the Midwest REgion is 4? While it is only one cell count, and it is very close to 5, even so, using the chi-square distribution for the test statistic may not be such a good approximation, even to the extent that we should at least look at another test - one that can handle small expected cell frequencies. Fisher's Exact Test is just the one. It can handle tables larger than 2x2. Let's see what is says about the Health Exam data.

Fishers Exact Test for Health Exam data - more than a 2x2 table

```r
fisher.test(HealthExam$AgeGroup,HealthExam$Region)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  HealthExam$AgeGroup and HealthExam$Region
## p-value = 0.2443
## alternative hypothesis: two.sided
```

Fishers Exact Test for Health Exam data - more than a 2x2 table

```
fisher.test(HealthExam$AgeGroup,HealthExam$Region)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  HealthExam$AgeGroup and HealthExam$Region
## p-value = 0.2443
## alternative hypothesis: two.sided
```

Bottom panel note: Note that in this case the result is nearly identical to the chi-square test.

Row Percents

```r
options(digits=3)
demographics=table(HealthExam$AgeGroup,HealthExam$Region)
prop.table(demographics,1)*100
```

```
##
##             Midwest Northeast South West
##   18 to 35    21.4      32.1  17.9 28.6
##   36 to 64    12.5      21.9  40.6 25.0
##   65+         30.0      30.0  10.0 30.0
```

Row Percents

```
options(digits=3)
demographics=table(HealthExam$AgeGroup,HealthExam$Region)
prop.table(demographics,1)*100
```

```
##
##            Midwest Northeast South West
##   18 to 35    21.4      32.1  17.9 28.6
##   36 to 64    12.5      21.9  40.6 25.0
##   65+         30.0      30.0  10.0 30.0
```

If I was interested in looking at the distribution of people in the 4 geographic Regions for each Age Group. Base on the way the contingency table is arranged, I would need row percents. That is, the rows add up to 100 percent.

Comparisons of percentages among Age Groups can now be made for each Region. So I can say something like "21.4% of the all people in the sample age 18 to 35 live in the Midwest, while only 12.5% of the 36 to 65 year-olds live in the Midwest and 30% of people over 65 live in the Midwest."

These percentages may seem far apart, but they weren't different enough for our chi-square test here to reject the hypothesis of independence. The sample size is big enough to conduct the

Column Percents

```r
options(digits=3)
demographics=table(HealthExam$AgeGroup,HealthExam$Region)
prop.table(demographics,2)*100
```

```
##
##              Midwest Northeast South West
##   18 to 35     37.5      40.9  25.0 36.4
##   36 to 64     25.0      31.8  65.0 36.4
##   65+          37.5      27.3  10.0 27.3
```

Column Percents

```r
options(digits=3)
demographics=table(HealthExam$AgeGroup,HealthExam$Region)
prop.table(demographics,2)*100
```

```
##
##            Midwest Northeast South West
##   18 to 35    37.5      40.9  25.0 36.4
##   36 to 64    25.0      31.8  65.0 36.4
##   65+         37.5      27.3  10.0 27.3
```

If I was interested in looking at the distribution of people in the 3 Age Groups for each Region. Base on the way the contingency table is arranged, I would need row percents. Notice for this one it is the columns that add up to 100 percent.

Comparisons of percentages among Geographic Regions can now be made for each Age Group. So I can say something like "37.5% of the all people in the sample in the Midwest are 18 to 35 years old, 40.9% in the Northeast are 18 to 35, 25% in the South are 18 to 35, 36.4% in the West are 18 to 35."

The number 2 in the prop.table function is what directs R to compute column percents. In matrix notation, the rows get mentioned first and the columns get mentioned second, so a 2

Odds Ratios

Let's go back to the text book for an example of odds ratios. Example 10.16 on pp. 533-535 of the Ott textbook uses the following data.

	Employee Response		
Job Stress	**Favorable**	**Unfavorable**	**Total**
Low	250	750	1,000
High	400	1,600	2,000
Total	650	2,350	3,000

Figure 6:

Odds Ratios

The R code for entering the data in Example 10.16 on pp. 533-535 of the Ott textbook is

```
counts=matrix(c(250,400,750,1600),nrow=2)
rownames(counts) <- c("Low","High")
colnames(counts) <- c("Favorable","Unfavorable")
counts
```

```
##      Favorable Unfavorable
## Low        250         750
## High       400        1600
```

R function for Odds Ratio and Relative Risk (package: mosaic)

```
oddsRatio(counts,verbose=TRUE)
```

R function for Odds Ratio and Relative Risk (package: mosaic)

```
oddsRatio(counts,verbose=TRUE)
```

Running the oddsRatio function will require you to install the package called mosaic first, but it does a nice job of computing the proportions, relative risk, odds, and odd ratio as well as the confidence intervals for the relative risk and odds ratio.

bottom panel note: For Example 10.16 on pp. 533-535 of the Ott textbook

R output for oddsRatio(counts,verbose=TRUE)

```
Proportions
       Prop. 1:  0.25
       Prop. 2:  0.2
     Rel. Risk:  0.8

Odds
        Odds 1:  0.3333
        Odds 2:  0.25
    Odds Ratio:  0.75

95 percent confidence interval:
     0.6965 < RR < 0.9189
     0.6263 < OR < 0.8981
```

R output for oddsRatio(counts,verbose=TRUE)

```
Proportions
       Prop. 1:  0.25
       Prop. 2:  0.2
     Rel. Risk:  0.8

Odds
        Odds 1:  0.3333
        Odds 2:  0.25
    Odds Ratio:  0.75

95 percent confidence interval:
     0.6965 < RR < 0.9189
     0.6263 < OR < 0.8981
```

With the option verbose=TRUE, we get all the output we want here. We get the proportions of a favorable response for both the low and high stress jobs along with their ratio, the relative risk with row 2 in the numerator; .2 divided by .25 equals .8.

We get the odds of a favorable response for the low stress job as 250 divided by 750, which is 0.3333, and the odds of a favorable response for the high stress job, which is 400 divided by 1600, which is 0.25.

And, of course, we get the ratio of those odds, with the odds for row 2 in the numerator as .25 divided by .3333 to get .75.

95% percent confidence intervals for the relative risk and odds ratio are also displayed. The level of confidence can be adjusted in the

Odds Ratio: Interpretation Options when the OR $= 0.75$

1. As a multiple

"The odds of a favorable response for employees in a high stress job are 0.75 times as large as the odds of a favorable response for employees in a low stress job."

or

"The odds of a favorable response for employees in a high stress job are only three-fourths of the odds for employees in a low stress job."

Odds Ratio: Interpretation Options when the OR = 0.75

1. As a multiple

"The odds of a favorable response for employees in a high stress job are 0.75 times as large as the odds of a favorable response for employees in a low stress job."

or

"The odds of a favorable response for employees in a high stress job are only three-fourths of the odds for employees in a low stress job."

Odds ratios can be interpreted in a variety of ways. In any case, one must proceed with caution when interpreting odds ratios, because they can so easily be misrepresented or misunderstood. Take some time to read these interpretations carefully.

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Odds Ratio: Interpretation Options when the OR = 0.75

2. As a percent

"The odds of a favorable response for employees in a high stress job are only 75% of the odds for employees in a low stress job."

or

"The odds of a favorable response for employees in a high stress job are 25% less than the odds of a favorable response for employees in a low stress job."

Odds Ratio: Interpretation Options when the OR = 0.75

2. As a percent

"The odds of a favorable response for employees in a high stress job are only 75% of the odds for employees in a low stress job."

or

"The odds of a favorable response for employees in a high stress job are 25% less than the odds of a favorable response for employees in a low stress job."

I like the second option here and I believe it is more common to express an odds ratio as a percent when its less than 1.

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Interpreting the OR Confidence Interval

Recall output from R

95 percent confidence interval: $0.6263 < OR < 0.8981$

"With 95% confidence, the odds of a favorable response from an employee in a high stress job are 63 to 90 percent as high as for an employee in a low stress job."

Interpreting the OR Confidence Interval

Recall output from R

95 percent confidence interval: $0.6263 < OR < 0.8981$

"With 95% confidence, the odds of a favorable response from an employee in a high stress job are 63 to 90 percent as high as for an employee in a low stress job."

An odds ratio of 1 would tell us that the odds of an event for the first group are identical to the odds for the second group. When we see a confidence interval that does not contain 1, we can conclude that there is a statistically significant relationship between the two categorical factors.

We could have equally said "With 95% confidence, that the odds of a favorable response from an employee in a high stress job are 10 to 37 percent less than for an employee in a low stress job."

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Let's reconstruct the 2x2 table so our output matches the textbook example output

```r
counts=matrix(c(400,250,1600,750),nrow=2)
rownames(counts) <- c("High","Low")
colnames(counts) <- c("Favorable","Unfavorable")
counts
```

```
##      Favorable Unfavorable
## High       400        1600
## Low        250         750
```

Let's reconstruct the 2x2 table so our output matches the textbook example output

```
counts=matrix(c(400,250,1600,750),nrow=2)
rownames(counts) <- c("High","Low")
colnames(counts) <- c("Favorable","Unfavorable")
counts
```

```
##      Favorable Unfavorable
## High       400        1600
## Low        250         750
```

By entering the 2x2 table into R such that the frequencies for the Low Stress Job are in row 2, so that R puts them in the numerator of the odds ratio, we can replicate the output for example 10.16 in the textbook.

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

R output for Example 10.16 (again)

```
Proportions
       Prop. 1:  0.2
       Prop. 2:  0.25
     Rel. Risk:  1.25

Odds
        Odds 1:  0.25
        Odds 2:  0.3333
    Odds Ratio:  1.333

95 percent confidence interval:
     1.088 < RR < 1.436
     1.113 < OR < 1.597
```

R output for Example 10.16 (again)

```
Proportions
       Prop. 1:  0.2
       Prop. 2:  0.25
     Rel. Risk:  1.25

Odds
        Odds 1:  0.25
        Odds 2:  0.3333
    Odds Ratio:  1.333

95 percent confidence interval:
     1.088 < RR < 1.436
     1.113 < OR < 1.597
```

Notice now that the odds ratio and confidence interval bounds for the odds ratio now match the values given on page 534 of Ott's textbook.

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Odds Ratio: Interpretation Options when the OR = 1.333

1. As a multiple

"The odds of a favorable response for employees in a low stress job are 1.33 times the odds of a favorable response for employees in a high stress job."

Odds Ratio: Interpretation Options when the OR = 1.333

1. As a multiple

"The odds of a favorable response for employees in a low stress job are 1.33 times the odds of a favorable response for employees in a high stress job."

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Odds Ratio: Interpretation Options when the OR = 1.333

2. As a percent

"The odds of a favorable response for employees in a low stress job are only 133% of the odds for employees in a high stress job."

or

"The odds of a favorable response for employees in a low stress job are 33% more than the odds of a favorable response for employees in a high stress job."

Odds Ratio: Interpretation Options when the OR = 1.333

2. As a percent

"The odds of a favorable response for employees in a low stress job are only 133% of the odds for employees in a high stress job."

or

"The odds of a favorable response for employees in a low stress job are 33% more than the odds of a favorable response for employees in a high stress job."

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook

Interpreting the OR Confidence Interval

Recall output from R

95 percent confidence interval: $1.113 < \text{OR} < 1.597$

"With 95% confidence, the odds of a favorable response from an employee in a low stress job are 11 to 60 percent higher than for an employee in a high stress job."

Interpreting the OR Confidence Interval

Recall output from R

95 percent confidence interval: $1.113 < \text{OR} < 1.597$

"With 95% confidence, the odds of a favorable response from an employee in a low stress job are 11 to 60 percent higher than for an employee in a high stress job."

bottom panel note: Example 10.16 on pp. 533-535 of the Ott textbook