Inference for Categorical Data

Categorical Variables

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

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()**

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)
```

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
```

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)
```

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
```

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()**

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)
```

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')
```

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 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")
```

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

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

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

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

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

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
```

Chi-Square Tests

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

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()**

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: Getting the Data into R

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

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
```

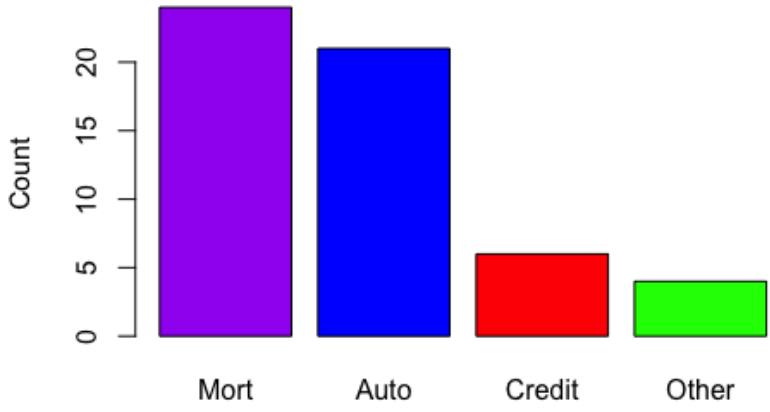
GOF Test: Checking Expected Values in R

```
55*proportions
```

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

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"))
```



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
```

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
```

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()**

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 Contingency Table

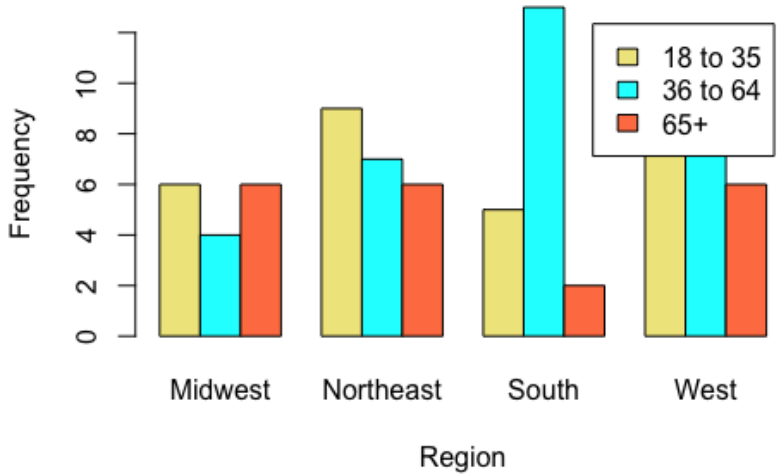
To see the crosstabs, use the 'table' function in R

```
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
```

Making a Bar Graph for Two Categorical Variables

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



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)
```

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 ($P = 0.2247$).

Expected Cell Counts for Health Exam data

```
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
```

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
```

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
```

Column Percents

```
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
```

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 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
```

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

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."

Interpreting the OR Confidence Interval

Recall output from R

95 percent confidence interval: $0.6263 < \text{OR} < 0.8981$

"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."

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
```

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
```

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

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."

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."