Inferential Statistics and Probability
a Holistic Approach

Chapter 11
Chi-square Tests for Categorical Data

1

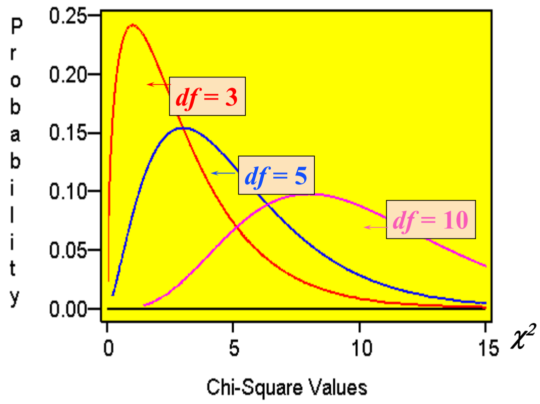
1

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - It is based on degrees of freedom
  - When the degrees of freedom change a new distribution is created

2

2



3

Goodness-of-Fit Test: Equal Expected Frequencies

- Let O_i and E_i be the observed and expected frequencies respectively for each category.
- H_0: there is no difference between Observed and Expected Frequencies
- H_a: there is a difference between Observed and Expected Frequencies
- The test statistic is: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- The critical value is a chi-square value with (k-1) degrees of freedom, where k is the number of categories

4

4

EXAMPLE 1

- The following data on absenteeism was collected from a manufacturing plant. At the .01 level of significance, Can you support the claim that there is a difference in the absence rate by day of the week?

Day	Frequency
Monday	95
Tuesday	65
Wednesday	60
Thursday	80
Friday	100

5

5

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i
Mon	95	0.20
Tues	65	0.20
Wed	60	0.20
Thur	80	0.20
Fri	100	0.20
Total	**400**	**1**

6

6

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i	E_i
Mon	95	0.20	80
Tues	65	0.20	80
Wed	60	0.20	80
Thur	80	0.20	80
Fri	100	0.20	80
Total	**400**	**1**	**400**

7

7

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i	E_i	(O-E)^2/E
Mon	95	0.20	80	2.8125
Tues	65	0.20	80	2.8125
Wed	60	0.20	80	5.0000
Thur	80	0.20	80	0.0000
Fri	100	0.20	80	5.0000
Total	**400**	**1**	**400**	**15.625**

8

8

EXAMPLE 1 *continued*

- H_o: There is no difference absenteeism due to day of the week.
- H_a: There is a difference absenteeism due to day of the week.
- H_o: $p_1=p_2=p_3=p_4=p_5$
- H_a: At least one proportion is different

- Test statistic: chi-square=$\Sigma(O-E)^2/E$=15.625
- Decision Rule: reject H_o if test statistic is greater than the critical value of 13.277. (4 df, α=.01)

- Conclusion: reject H_o and conclude that there is a difference absenteeism due to day of the week.

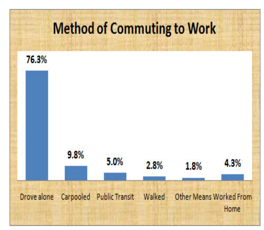
9

9

Goodness-of-Fit Test: Unequal Expected Frequencies

EXAMPLE 2

- In the 2010 United States census, data was collected on how people get to work -- their method of commuting.
- Suppose you wanted to know if people who live in the San Jose metropolitan area (Santa Clara County) commute with similar proportions as the United States.
- Design and conduct a hypothesis test at the 5% significance level.



10

10

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764			
Carpooled	105			
Public Transit	34			
Walked	20			
Other Means	30			
Worked from Home	47			
TOTAL	**1000**			

11

11

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763		
Carpooled	105	0.098		
Public Transit	34	0.050		
Walked	20	0.028		
Other Means	30	0.018		
Worked from Home	47	0.043		
TOTAL	**1000**	**1.000**		

12

12

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763	763	
Carpooled	105	0.098	98	
Public Transit	34	0.050	50	
Walked	20	0.028	28	
Other Means	30	0.018	18	
Worked from Home	47	0.043	43	
TOTAL	**1000**	**1.000**	**1000**	

13

13

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763	763	0.0013
Carpooled	105	0.098	98	0.5000
Public Transit	34	0.050	50	5.1200
Walked	20	0.028	28	2.2857
Other Means	30	0.018	18	8.0000
Worked from Home	47	0.043	43	0.3721
TOTAL	**1000**	**1.000**	**1000**	**16.2791**

14

14

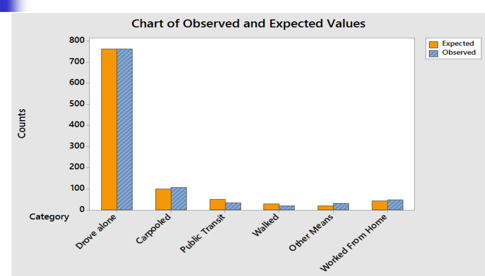
EXAMPLE 2 *continued*

- **Design:**
  Ho: $p_1 = .763$ $p_2 = .098$ $p_3 = .050$ $p_4 = .028$ $p_5 = .018$ $p_6 = .043$
  Ha: At least one p_i is different than what was stated in Ho
- $\alpha=.05$
- Model: Chi-Square Goodness of Fit, df=5
- H_o is rejected if $\chi^2 > 11.071$
- **Data:**
- $\chi^2 = 16.2791$, Reject Ho
- **Conclusion:**
  Workers in Santa Clara County do not have the same frequencies of method of commuting as workers in the entire United States.

15

15

EXAMPLE 2 *continued*



16

16

Explanatory/Response Models

- The remaining models covered in the course can be used for testing claims of the following form:
- Ho: There is no difference in the **Response Variable** due to the **Explanatory Variable**
- Ha: There is a difference in the **Response Variable** due to the **Explanatory Variable**
- If both the explanatory and categorical variables are categorical, then use the **Chi-square Test of Independence Model**

17

17

Chi-square Test of Independence

- Contingency table analysis is used to test whether two traits or variables are related.
- Each observation is classified according to two categorical variables (Explanatory and Response).
- Ha: The variables are dependent
- The *degrees of freedom* is equal to: (number of rows-1)(number of columns-1).
- The expected frequency is computed as: Expected Frequency = (row total)(column total)/grand total

18

18

EXAMPLE 3

- In May 2014, Colorado became the first state to legalize the recreational use of marijuana.
- A poll of 1000 adults were classified by gender and their opinion about legalizing marijuana
- At the .05 level of significance, can we conclude that gender and the opinion about legalizing marijuana for recreational use are dependent events?

Marijuana should be	Men	Women	Total
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

19

19

Example 3 (continued)

- The **observed** is the reported data.

Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
TOTAL	500	500	1000

20

20

Example 3 (continued)

- The **observed** is the reported data.
- The **expected** is

$$\frac{Row\ Total\ \times Column\ Total}{Grand\ Total\ (n)}$$

Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270 250	230 250	500
Not Legal	205 225	245 225	450
Unsure	25 25	25 25	50
TOTAL	500	500	1000

21

21

Example 3 (continued)

- The **observed** is the reported data.
- The **expected** is

$$\frac{Row\ Total\ \times Column\ Total}{Grand\ Total\ (n)}$$

- **Chi-square** is

$$\frac{(Observed - Expected)^2}{Expected}$$

- Sum to get test statistic: χ^2 **= 6.756**

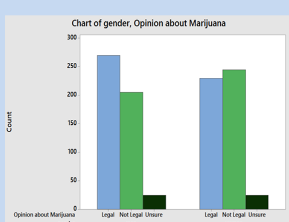
Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270 250 1.600	230 250 1.600	500
Not Legal	205 225 1.778	245 225 1.778	450
Unsure	25 25 0.000	25 25 0.000	50
TOTAL	500	500	1000

22

22

EXAMPLE 3 *continued*

```
Rows: Opinion about Marijuana
Columns: gender

1st value = Observed
2nd Value = Expected
3rd Value = Contribution to Chi-square

              men   women   All

Legal         270     230   500
              250     250
            1.600   1.600

Not Legal     205     245   450
              225     225
            1.778   1.778

Unsure         25      25    50
               25      25
            0.000   0.000

All           500     500  1000
```



23

23

EXAMPLE 3 *continued*

- Explanatory Variable: Gender Response Variable: Opinion
- H_o: There is no difference in Opinion due to gender.
  H_a: There is a difference in Opinion due to gender.
  H_o: Gender and Opinion are independent.
  H_a: Gender and Opinion are dependent.
- α=.05
- Model: Chi-Square Test for Independence, df=2
- H_o is rejected if χ^2 > 5.99
- **Data:** χ^2 = 6.756, Reject Ho
- **Conclusion:** Gender and opinion are dependent variables. Men are more likely to support legalizing marijuana for recreational use.

24

24