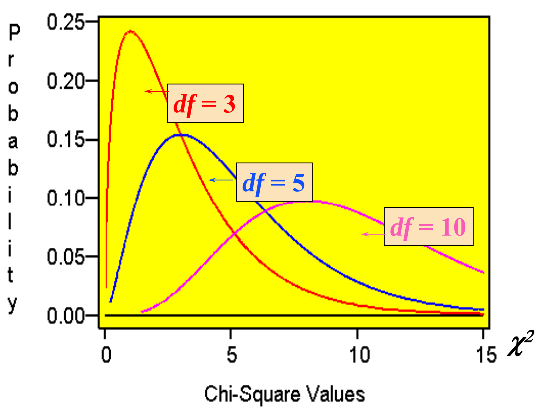
Inferential Statistics and Probability
a Holistic Approach

Chapter 11

Chi-square Tests for Categorical Data

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



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

EXAMPLE 1

- The following data on absenteeism was collected from a manufacturing plant. At the .01 level of significance, test to determine whether 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

EXAMPLE 1 *continued*

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

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

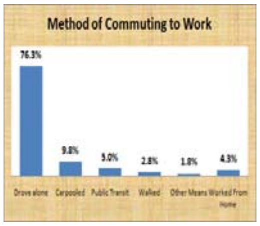
EXAMPLE 1 *continued*

- H_o: there is no difference between the observed and the expected frequencies of absences.
- H_a: there is a difference between the observed and the expected frequencies of absences.
- 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 between the observed and expected frequencies of absences.

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.



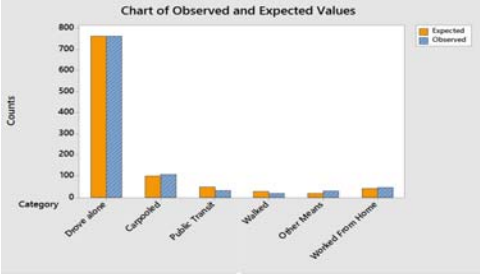
EXAMPLE 2 *continued*

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**

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
- α=.05
- Model: Chi-Square Goodness of Fit, df=5
- H_o is rejected if χ^2 > 11.071
- **Data:**
- χ^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.

EXAMPLE 2 *continued*



Contingency Table Analysis

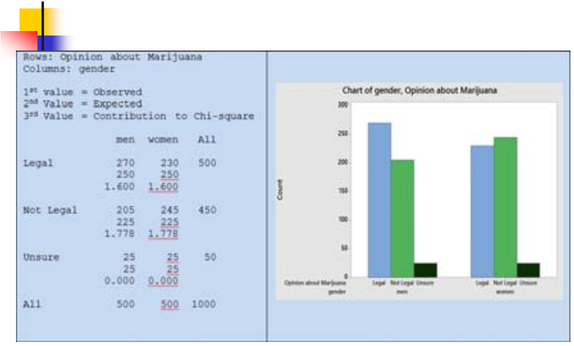
- Contingency table analysis is used to test whether two traits or variables are related.
- Each observation is classified according to two variables.
- The usual hypothesis testing procedure is used.
- The *degrees of freedom* is equal to: (number of rows-1)(number of columns-1).
- The expected frequency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

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**

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
```

EXAMPLE 3 *continued*

- **Design:** 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 $\chi^2 > 5.99$
- **Data:** $\chi^2 = 6.756$, Reject Ho
- **Conclusion:** Gender and opinion are dependent variables. Men are more likely to support legalizing marijuana for recreational use.