Statistical Analysis

Chi-Square Testing

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Today’s Lesson

1. Chi-Square Goodness-of-Fit Tests
    - Uniform Patterns
    - Specific Patterns
    - Normal distribution pattern
2. A Chi-Square Test for Independence

Chi-Square Goodness-of-Fit Tests

- Non-parametric Tests
- Often use categorical data for statistical inference
- No longer assumption of a normal distribution
- Collect count data to study how counts are distributed among cells

Chi-Square Goodness-of-Fit Tests for a Uniform Pattern

- There may be three or more categories
- Count the number of observations in each category.
- Always comparing the expected with the observed.
- Hypothesis testing for a uniform patter:
  - H_0: All categories have the same number of observations
  - H_a: Not all categories have the same number of observations
  - Test is conducted at a significance level
- Collect count data to study how counts are distributed among cells

Example

We would expect that the day of the week that a car is produced has no relationship to the number of defects. However, car buyers are often warned that cars made on a Friday or a Monday have more defects.

Data has been collected on the number of defects by day the car was made. Monday through Friday; 32, 22, 26, 19, and 30. Test for a uniform pattern at significance of 5%.

Step 1: State Hypothesis

- H_0: Defects are the same for each manufacturing day
- H_a: Defects are notthe same for each manufacturing day

Example *continued*

2. Calculate the expected defects per day.

$$\frac{32 + 22 + 26 + 19 + 30}{5} = \frac{129}{5} = 25.8$$ This is the expected if uniform

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

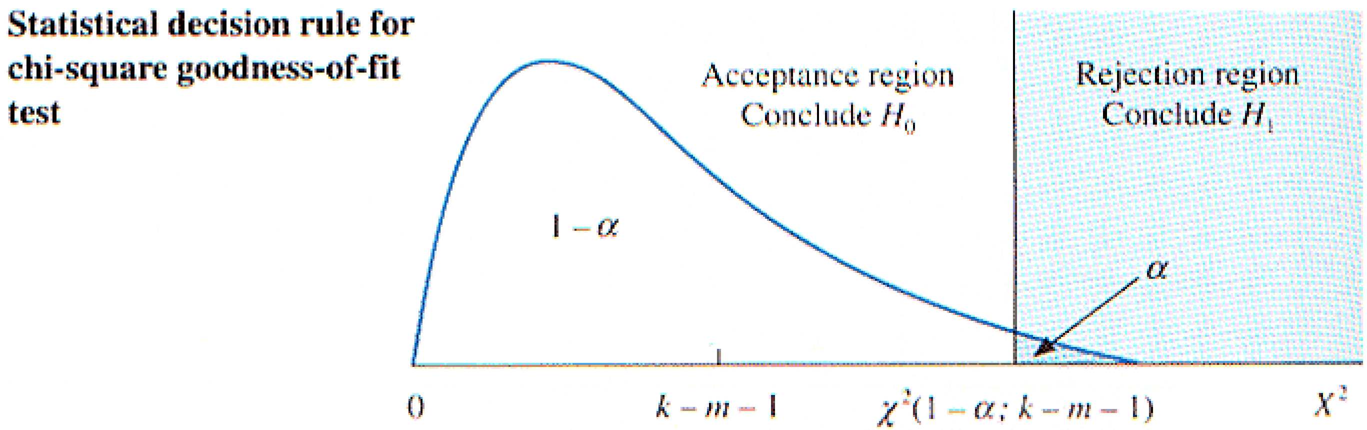
$$= \frac{(32-25.8)^2 + (22-25.8)^2 + (26-25.8)^2 + (19-25.8)^2 + (30-25.8)^2}{25.8}$$

$$\chi^2 = 4.53$$

From the Chi-Square table the value for df 4, and significance of 5%, 9.48.

Interpretation

Since the calculated Chi Square value is less than the value from the table, we do not reject the null. The is not evidence to support that there is a difference in number of defects by day.



Chi-Square for a Specific Pattern

A store wants to test if the preference for Brand of Microwaves is the same in their location of Milwaukee as in Cleveland.

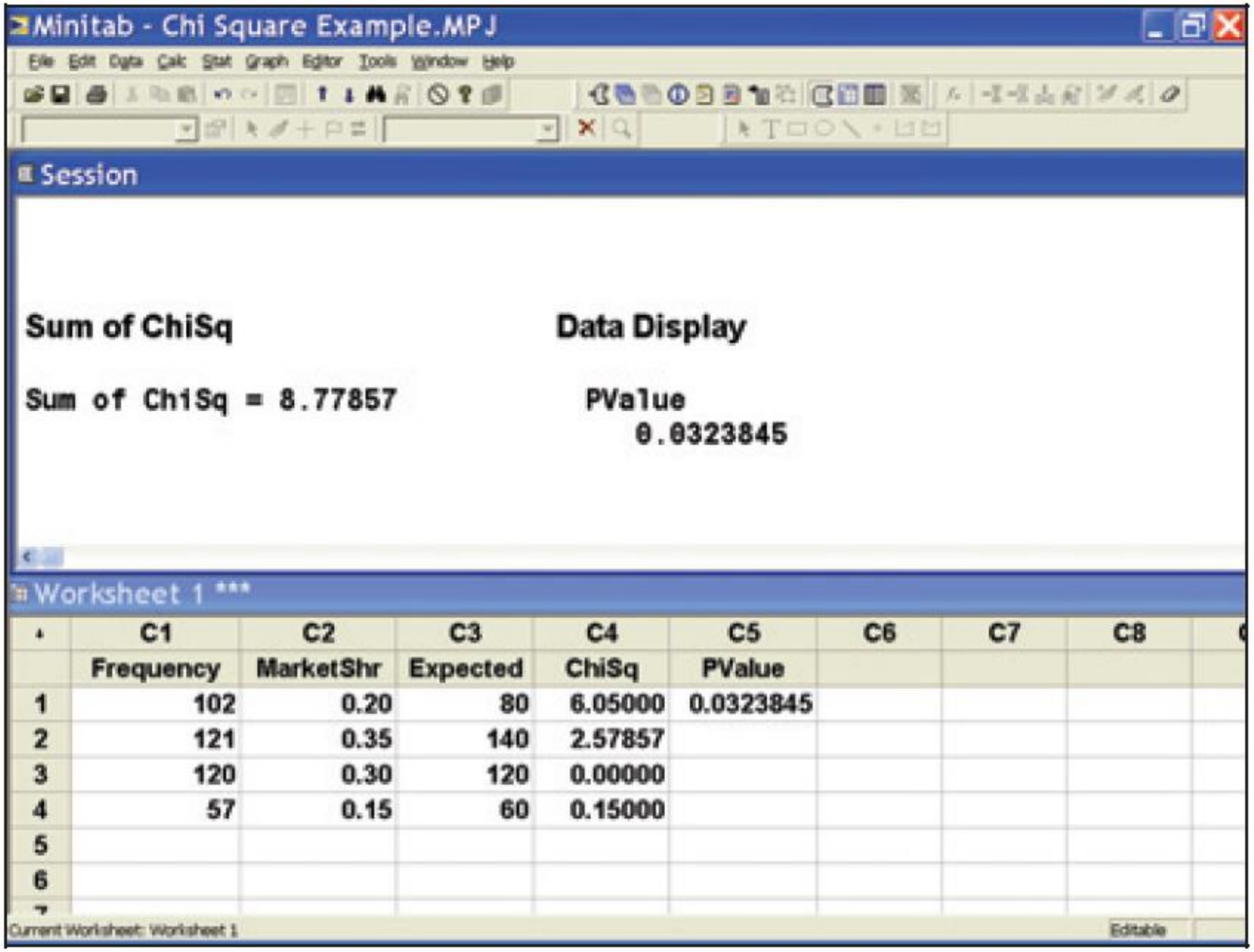
- H_0: The pattern of brand preferences in Milwaukee is the same as Cleveland
- H_a: The pattern of brand preferences in Milwaukee is **not** the same as Cleveland

The Microwave Oven Case: Studying Consumer Preferences

- Market shares in Cleveland
  - Brand 1 20%
  - Brand 2 35%
  - Brand 3 30%
  - Brand 4 15%

- Observed Frequency in Milwaukee
  - Brand 1 102
  - Brand 2 121
  - Brand 3 120
  - Brand 4 57

The Microwave Oven Case *continued*



↓	C1	C2	C3	C4	C5	C6	C7	C8
	Frequency	MarketShr	Expected	ChiSq	PValue			
1	102	0.20	80	6.05000	0.0323845			
2	121	0.35	140	2.57857				
3	120	0.30	120	0.00000				
4	57	0.15	60	0.15000				
5								
6								

The Microwave Oven Case *continued*

- $H_0 : p_1 = .20,\ p_2 = .35,\ p_3 = .30,\ p_4 = .15$
- $H_a : H_0$ fails to hold

$$\chi^2 = \sum_{i=1}^{k=4} \frac{(f_i - E_i)^2}{E_i}$$

$$= \frac{(102 - 80)^2}{80} + \frac{(121 - 140)^2}{140} + \frac{(120 - 120)^2}{120} + \frac{(57 - 60)^2}{60}$$

$$= \frac{484}{80} + \frac{361}{140} + \frac{0}{120} + \frac{9}{60} = 8.7786$$

$$\chi^2 = 8.7786 > \chi^2_{.05} = 7.81473$$

A Goodness of Fit Test for a Normal Distribution

1. Test the following null and alternative hypotheses:
   H_0: the population has a normal distribution
   H_a: population does not have normal distribution
2. Select random sample and compute sample mean and standard deviation
3. Define k intervals for the test
4. Record observed frequency (f_i) for each interval
5. Calculate expected frequency (E_i)
6. Calculate the chi-square statistic
7. Make a decision

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(f_i - E_i\right)^2}{E_i}$$

A Chi-Square Test for Independence

- Each of n randomly selected items is classified on two dimensions into a contingency table with r rows and c columns and let:
  - f_{ij} = observed cell frequency for i^{th} row and j^{th} column
  - r_i = i^{th} row total
    c_j = j^{th} column total
- Expected cell frequency for i^{th} row and j^{th} column under independence

$$\hat{E}_{ij} = \frac{r_i c_j}{n}$$

Example of test for Independence with contingency tables

The American Marketing Association wants to determine the relationship between the importance store owner attach to advertising and the size of the store they own. We will test at α = 0.05

- H_0: There is no relationship between store size and importance placed by store owner.
- H_a: There is a relationship between store size and importance placed by store owner.

Example

Advertising

Size	**Importance**	**Not Important**	**No Opinion**
Small	20	52	32
Medium	53	47	28
Large	67	32	25

Example *continued*

Size		Importance	Not Important	No Opinion	Total
Small	**Obs**	20	52	32	**104**
	Exp	40.899	38.27	24.83	
Medium	**Obs**	53	47	28	**128**
	Exp	50.34	47.1	30.56	
Large	**Obs**	67	32	25	**124**
	Exp	47.76	45.63	29.61	
Total		**140**	**131**	**85**	**356**

A Chi-Square Test for Independence *continued*

- H_0 : the two classifications are statistically independent
- H_a : the two classifications are statistically dependent

- Test statistic

$$c^2 = \sum_{\text{all cells}} \frac{(f_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

- Reject H_0 if $\chi^2 > \chi_\alpha^2$ or if $p\text{-}value < \alpha$
- χ_α^2 and the $p\text{-}value$ are based on $(r\text{-}1)(c\text{-}1)$ degrees of freedom

Interpretation

χ^2 = 30.59

$df = (r\text{-}1)\ (c\text{-}1) = (3\text{-}1)(3\text{-}1) = 4$

Table value = **9.84**

Interpretation: Since the calculated value of 30.59 is greater than 9.84, there is evidence to support a relationship between the size of the store and the importance the owner places on advertising.