QM353: Business Statistics

Chapter 6
Goodness-of-Fit Tests and Contingency Analysis

Chapter Goals

After completing this chapter, you should be able to:

- Use the chi-square goodness-of-fit test to determine whether data fits a specified distribution
- Set up a contingency analysis table and perform a chi-square test of independence

Chi-Square Goodness-of-Fit Test

- Does sample data conform to a hypothesized distribution?
  - Examples:
    - Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)
    - Do measurements from a production process follow a normal distribution?

Chi-Square Goodness-of-Fit Test *(continued)*

- Are technical support calls equal across all days of the week? (i.e., do calls follow a uniform distribution?)
  - Sample data for 10 days per day of week:

	Sum of calls for this day:
Monday	290
Tuesday	250
Wednesday	238
Thursday	257
Friday	265
Saturday	230
Sunday	192
	Σ = 1722

Logic of Goodness-of-Fit Test

- If calls **are** uniformly distributed, the 1722 calls would be expected to be equally divided across the 7 days:

$$\frac{1722}{7} = 246 \quad \text{expected calls per day if uniform}$$

- Chi-Square Goodness-of-Fit Test: test to see if the sample results are consistent with the expected results

  (i.e., actual (observed) data = expected data)

Observed vs. Expected Frequencies

	Observed o_i	Expected e_i
Monday	290	246
Tuesday	250	246
Wednesday	238	246
Thursday	257	246
Friday	265	246
Saturday	230	246
Sunday	192	246
TOTAL	1722	1722

Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week (observed = expected)

H_A: The distribution of calls is not uniform

- The test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} \quad (\text{where } df = k - 1)$$

where:

k = number of categories

o_i = observed cell frequency for category i

e_i = expected cell frequency for category i

The Rejection Region

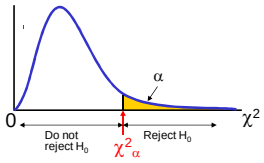
H_0: The distribution of calls is uniform over days of the week

H_A: The distribution of calls is not uniform

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

- Reject H_0 if $\chi^2 > \chi^2_\alpha$

(with k – 1 degrees of freedom)



Chi-Square Test Statistic

H_0: The distribution of calls is uniform over days of the week

H_A: The distribution of calls is not uniform

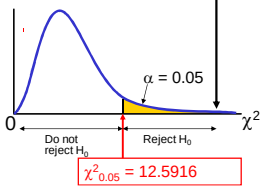
$$\chi^2 = \frac{(290-246)^2}{246} + \frac{(250-246)^2}{246} + \ldots + \frac{(192-246)^2}{246} = 23.05$$

k – 1 = 6 (7 days of the week) so use 6 degrees of freedom **(Appendix G):**

$\chi^2_{0.05} = 12.5916$

Conclusion:

$\chi^2 = 23.05 > \chi^2_\alpha = 12.5916$ so **reject H_0** and conclude that the distribution is not uniform



Goodness-of-Fit Test

1. State the appropriate hypotheses
2. Specify significance level
3. Determine the critical value (Appendix G)
4. Compute the test statistics, χ^2
5. Reach a decision
6. Draw a conclusion

Normal Distribution Example

- Do measurements from a production process follow a normal distribution with μ = 50 and σ = 15?
- Process:
  - Get sample data
  - Group sample results into classes (cells) (Expected cell frequency must be at least 5 for each cell)
  - Compare actual cell frequencies with expected cell frequencies

Normal Distribution Example

(continued)

- Sample data and values grouped into classes:

150 Sample Measurements
80
65
36
66
50
38
57
77
59
…etc…

Class	Frequency
less than 30	10
30 but < 40	21
40 but < 50	33
50 but < 60	41
60 but < 70	26
70 but < 80	10
80 but < 90	7
90 or over	2
TOTAL	150

Normal Distribution Example

(continued)

- What are the expected frequencies for these classes for a normal distribution with μ = 50 and σ = 15 ?

Class	Observed Frequency	Expected Frequency
less than 30	10	?
30 but < 40	21	
40 but < 50	33	
50 but < 60	41	
60 but < 70	26	
70 but < 80	10	
80 but < 90	7	
90 or over	2	
TOTAL	150	

Expected Frequencies

Value	P(X < value)	Expected frequency
less than 30	0.09121	13.68
30 but < 40	0.16128	24.19
40 but < 50	0.24751	37.13
50 but < 60	0.24751	37.13
60 but < 70	0.16128	24.19
70 but < 80	0.06846	10.27
80 but < 90	0.01892	2.84
90 or over	0.00383	0.57
TOTAL	1.00000	150.00

Expected frequencies in a sample of size n=150, from a normal distribution with μ=50, σ=15

Example:

$$P(x<30) = P\left(z < \frac{30-50}{15}\right) = P(z<-1.3333) = 0.0912$$

$$(0.0912)(150) = 13.68$$

The Test Statistic

Class	Frequency (observed, o_i)	Expected Frequency, e_i
less than 30	10	13.68
30 but < 40	21	24.19
40 but < 50	33	37.13
50 but < 60	41	37.13
60 but < 70	26	24.19
70 but < 80	10	10.27
80 but < 90	7	2.84
90 or over	2	0.57
TOTAL	150	150.00

The test statistic is

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

- Reject H_0 if $\chi^2 > \chi^2_\alpha$

(with k – 1 degrees of freedom)

The Rejection Region

H_0: The distribution of values is normal with μ = 50 and σ = 15

H_A: The distribution of calls does not have this distribution

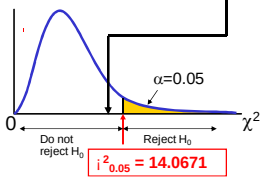
$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(10-13.68)^2}{13.68} + \ldots + \frac{(2-0.57)^2}{0.57} = 12.097$$

8 classes so use 7 d.f.:

$\chi^2_{0.05} = 14.0671$

Conclusion:

$\chi^2 = 12.097 < \chi^2_\alpha = 14.0671$ so **do not reject H_0**



Contingency Analysis

- Situations involving multiple population proportions
- Categorical data
- Used to classify sample observations according to two or more characteristics
- Use Chi-Square to determine independence of the characteristics of interest
- Data summarized in a contingency table
  - Also called a crosstabulation table

Contingency Analysis Example

Left-Handed vs. Gender (two variables)

- Dominant Hand: Left vs. Right
- Gender: Male vs. Female

H_0: Hand preference is independent of gender

H_A: Hand preference is not independent of gender

Contingency Analysis Example

(continued)

Sample results organized in a contingency table:

sample size = n = 300:

120 Females, 12 were left handed

180 Males, 24 were left handed

Gender	Hand Preference: Left	Hand Preference: Right	
Female	12	108	120
Male	24	156	180
	36	264	300

Logic of the Test

H_0: Hand preference is independent of gender

H_A: Hand preference is not independent of gender

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall

Finding Expected Frequencies

120 Females, 12 were left handed

180 Males, 24 were left handed

Overall:

P(Left Handed) = 36/300 = 0.12

If independent, then

P(Left Handed | Female) = P(Left Handed | Male) = 0.12

So we would expect 12% of the 120 females and 12% of the 180 males to be left handed…

i.e., we would expect (120)(0.12) = 14.4 females to be left handed
(180)(0.12) = 21.6 males to be left handed

Expected Cell Frequencies

(continued)

- Expected cell frequencies:

$$e_{ij} = \frac{(i^{th}\ \text{Row total})(j^{th}\ \text{Column total})}{\text{Total sample size}}$$

Example:

$$e_{11} = \frac{(120)(36)}{300} = 14.4$$

Observed vs. Expected Frequencies

Observed frequencies vs. expected frequencies:

Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

Marginal Frequencies

- A marginal frequency is the sum of the row or column
  - e.g., The marginal frequency for females in the study was 120
- The expected marginal frequency for a variable MUST match the observed marginal frequency for that same variable
  - i.e., The expected marginal frequency for females in the study must also be 120

The Chi-Square Test Statistic

The Chi-square contingency test statistic is:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \quad \text{with} \quad \text{d.f.} = (r-1)(c-1)$$

- where:
  o_{ij} = observed frequency in cell (i, j)
  e_{ij} = expected frequency in cell (i, j)
  r = number of rows
  c = number of columns

NOTE: All rows and columns must be used

Observed vs. Expected Frequencies

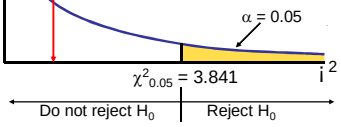
Gender	Hand Preference: Left	Hand Preference: Right	
Female	Observed = 12 Expected = 14.4	Observed = 108 Expected = 105.6	120
Male	Observed = 24 Expected = 21.6	Observed = 156 Expected = 158.4	180
	36	264	300

$$\chi^2 = \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.6848$$

Contingency Analysis

$\chi^2 = 0.6848$ with $\text{d.f.} = (r-1)(c-1) = (1)(1) = 1$

Decision Rule:
If $\chi^2 > 3.841$, reject H_0, otherwise, do not reject H_0



Here, $\chi^2 = 0.6848 < 3.841$, so we do not reject H_0 and conclude that gender and hand preference are independent

Chi-Square Test Cautions

- The chi-square distribution is only an approximation for the true distribution
  - But it is quite good when all expected cell frequencies are > 5
  - When frequencies are >5, the chi-square value may inflate the true probability of a Type I error
- If frequencies are small:
  - Increase sample size first
  - If needed combine the categories of the variables

Chapter Summary

- Used the chi-square goodness-of-fit test to determine whether data fits a specified distribution
  - Example of a discrete distribution (uniform)
  - Example of a continuous distribution (normal)
- Used contingency tables to perform a chi-square test of independence (contingency analysis)
  - Compared observed cell frequencies to expected cell frequencies