Part 8
Chi-square and ANOVA tests

© Maurice Geraghty 2015

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.
- $H_1$: 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.

<table>
<thead>
<tr>
<th>Day</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>95</td>
</tr>
<tr>
<td>Tuesday</td>
<td>65</td>
</tr>
<tr>
<td>Wednesday</td>
<td>60</td>
</tr>
<tr>
<td>Thursday</td>
<td>80</td>
</tr>
<tr>
<td>Friday</td>
<td>100</td>
</tr>
</tbody>
</table>

**EXAMPLE 1 continued**

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

<table>
<thead>
<tr>
<th>Day</th>
<th>O</th>
<th>E</th>
<th>$(O-E)^2/E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>95</td>
<td>80</td>
<td>2.8125</td>
</tr>
<tr>
<td>Tues</td>
<td>65</td>
<td>80</td>
<td>2.8125</td>
</tr>
<tr>
<td>Wed</td>
<td>60</td>
<td>80</td>
<td>5.0000</td>
</tr>
<tr>
<td>Thur</td>
<td>80</td>
<td>80</td>
<td>0.0000</td>
</tr>
<tr>
<td>Fri</td>
<td>100</td>
<td>80</td>
<td>5.0000</td>
</tr>
<tr>
<td>Total</td>
<td>400</td>
<td>400</td>
<td>15.625</td>
</tr>
</tbody>
</table>
EXAMPLE 1

H₀: there is no difference between the observed and the expected frequencies of absences.
- H₁: there is a difference between the observed and the expected frequencies of absences.

Test statistic: \( \chi^2 = \sum \frac{(O - E)^2}{E} = 15.625 \)
Decision Rule: reject H₀ if test statistic is greater than the critical value of 13.277. (4 df, \( \alpha = .01 \))

Conclusion: reject H₀ and conclude that there is a difference between the observed and expected frequencies of absences.

EXAMPLE 2

The U.S. Bureau of the Census (2000) indicated that 54.4% of the population is married, 6.6% widowed, 9.7% divorced (and not re-married), 2.2% separated, and 27.1% single (never been married).

A sample of 500 adults from the San Jose area showed that 270 were married, 22 widowed, 42 divorced, 10 separated, and 156 single.

At the .05 significance level can we conclude that the San Jose area is different from the U.S. as a whole?

<table>
<thead>
<tr>
<th>Status</th>
<th>O</th>
<th>E</th>
<th>( \frac{(O - E)^2}{E} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>270</td>
<td>272</td>
<td>0.015</td>
</tr>
<tr>
<td>Widowed</td>
<td>22</td>
<td>33</td>
<td>3.667</td>
</tr>
<tr>
<td>Divorced</td>
<td>42</td>
<td>48.5</td>
<td>0.871</td>
</tr>
<tr>
<td>Separated</td>
<td>10</td>
<td>11</td>
<td>0.091</td>
</tr>
<tr>
<td>Single</td>
<td>156</td>
<td>135.5</td>
<td>3.101</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>500</td>
<td>7.745</td>
</tr>
</tbody>
</table>
EXAMPLE 2  

- **Design:**  
  - \( H_0: p_1 = .544 \ p_2 = .066 \ p_3 = .097 \ p_4 = .022 \ p_5 = .271 \)  
  - \( H_a: \) at least one \( p_i \) is different  
  - \( \alpha = .05 \)  
- **Model:** Chi-Square Goodness of Fit, df=4  
- **\( H_0 \) is rejected if \( \chi^2 > 9.488 \)\)  
- **Data:** \( \chi^2 = 7.745 \), Fail to Reject Ho  
- **Conclusion:** Insufficient evidence to conclude San Jose is different than the US Average

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:  \( \text{df} = (\text{number of rows}-1)(\text{number of columns}-1) \).  
- The expected frequency is computed as:  \( \text{Expected Frequency} = \frac{\text{(row total)(column total)}}{\text{grand total}} \)

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?

<table>
<thead>
<tr>
<th>Marijuana should be</th>
<th>Men</th>
<th>Women</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal</td>
<td>270</td>
<td>230</td>
<td>500</td>
</tr>
<tr>
<td>Not Legal</td>
<td>205</td>
<td>245</td>
<td>450</td>
</tr>
<tr>
<td>Unsure</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>
EXAMPLE 3

• **Design:** Ho: Gender and Opinion are independent. 
  Ha: Gender and Opinion are dependent. 
  \( \alpha = .05 \) 
  Model: Chi-Square Test for Independence, df=2 
  Ho 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.

Characteristics of F-Distribution

- There is a “family” of F Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- \( F \) cannot be negative, and it is a continuous distribution.
- The \( F \) distribution is positively skewed.
- Its values range from 0 to \( \infty \). As \( F \to \infty \) the curve approaches the Y-axis.
Underlying Assumptions for ANOVA

- The F distribution is also used for testing the equality of more than two means using a technique called analysis of variance (ANOVA). ANOVA requires the following conditions:
  - The populations being sampled are normally distributed.
  - The populations have equal standard deviations.
  - The samples are randomly selected and are independent.

Analysis of Variance Procedure

- The Null Hypothesis: the population means are the same.
- The Alternative Hypothesis: at least one of the means is different.
- The Test Statistic: \( F = \frac{\text{between sample variance}}{\text{within sample variance}} \).
- Decision rule: For a given significance level \( \alpha \), reject the null hypothesis if \( F \) (computed) is greater than \( F \) (table) with numerator and denominator degrees of freedom.

ANOVA - Null Hypothesis

- Ho is true - all means the same
- Ho is false - not all means the same
ANOVA NOTES

If there are \( k \) populations being sampled, then the \( \text{df (numerator)} = k-1 \)
- If there are a total of \( n \) sample points, then \( \text{df (denominator)} = n-k \)
- The test statistic is computed by: \( F = \frac{(SS_F)/(k-1))}{(SS_E)/(n-k)} \).
- \( SS_F \) represents the factor (between) sum of squares.
- \( SS_E \) represents the error (within) sum of squares.
- Let \( T_c \) represent the column totals, \( n_c \) represent the number of observations in each column, and \( \Sigma X \) represent the sum of all the observations.
- These calculations are tedious, so technology is used to generate the ANOVA table.

Formulas for ANOVA

\[
SS_{Total} = \Sigma(X^2) - \left(\frac{\Sigma X}{n}\right)^2
\]
\[
SS_{Factor} = \Sigma \left(\frac{T_c^2}{n_c}\right) - \left(\frac{\Sigma X}{n}\right)^2
\]
\[
SS_{Error} = SS_{Total} - SS_{Factor}
\]

ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>SS_{Factor}</td>
<td>k-1</td>
<td>SS_{Factor}/df_f</td>
<td>MS_F/MS_E</td>
</tr>
<tr>
<td>Error</td>
<td>SS_{Error}</td>
<td>n-k</td>
<td>SS_{Error}/df_E</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>SS_{Total}</td>
<td>n-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EXAMPLE 4

Party Pizza specializes in meals for students. Hsieh Li, President, recently developed a new tofu pizza.

- Before making it a part of the regular menu she decides to test it in several of her restaurants. She would like to know if there is a difference in the mean number of tofu pizzas sold per day at the Cupertino, San Jose, and Santa Clara pizzerias for sample of five days.

- At the .05 significance level can Hsieh Li conclude that there is a difference in the mean number of tofu pizzas sold per day at the three pizzerias?

<table>
<thead>
<tr>
<th></th>
<th>Cupertino</th>
<th>San Jose</th>
<th>Santa Clara</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>13</td>
<td>10</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>12</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>13</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>11</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>51</td>
<td>46</td>
<td>85</td>
<td>182</td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Means</td>
<td>12.75</td>
<td>11.5</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>$\sum^2$</td>
<td>653</td>
<td>534</td>
<td>1447</td>
<td>2634</td>
</tr>
</tbody>
</table>

Example 4 continued

$$SS_{Total} = 2634 - \frac{182^2}{13} = 86$$

$$SS_{Factor} = 2624.25 - \frac{182^2}{13} = 76.25$$

$$SS_{Error} = 86 - 76.25 = 9.75$$
**Example 4 continued**

### ANOVA TABLE

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>76.25</td>
<td>2</td>
<td>38.125</td>
<td>39.10</td>
</tr>
<tr>
<td>Error</td>
<td>9.75</td>
<td>10</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>86.00</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**EXAMPLE 4 continued**

- **Design:** $H_0: \mu_1 = \mu_2 = \mu_3$
- $H_1$: Not all the means are the same
- $\alpha = .05$
- **Model:** One Factor ANOVA
- $H_0$ is rejected if $F > 4.10$
- **Data:** Test statistic: $F = \frac{76.25/2}{9.75/10} = 39.1026$
- $H_0$ is rejected.
- **Conclusion:** There is a difference in the mean number of pizzas sold at each pizzeria.

**One-way ANOVA: Cupertino, San Jose, Santa Clara**

### ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>76.25</td>
<td>2</td>
<td>38.125</td>
<td>39.10</td>
</tr>
<tr>
<td>Error</td>
<td>9.75</td>
<td>10</td>
<td>0.975</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>86.00</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CI for Mean Based on Pooled SDs

- **Level**
  - Cupertino: $4.125 \pm 1.967$ (------)
  - San Jose: $11.500 \pm 1.967$ (-----)
  - Santa Clara: $17.000 \pm 1.967$ (-----)

<table>
<thead>
<tr>
<th>Level</th>
<th>Mean</th>
<th>Stdv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cupertino</td>
<td>4.125</td>
<td>1.967</td>
</tr>
<tr>
<td>San Jose</td>
<td>11.500</td>
<td>1.967</td>
</tr>
<tr>
<td>Santa Clara</td>
<td>17.000</td>
<td>1.967</td>
</tr>
</tbody>
</table>

12.0 14.0 16.0 18.0
Post Hoc Comparison Test

- Used for pairwise comparison
- Designed so the overall significance level is 5%.
- Use technology.
- Refer to Tukey Test Material in Supplemental Material.

---

Grouping Information Using Tukey Method

<table>
<thead>
<tr>
<th>City</th>
<th>Mean</th>
<th>Grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santa Clara</td>
<td>21.000</td>
<td>A</td>
</tr>
<tr>
<td>Cupertino</td>
<td>12.750</td>
<td>B</td>
</tr>
<tr>
<td>San Jose</td>
<td>11.500</td>
<td>B</td>
</tr>
</tbody>
</table>

Means that do not share a letter are significantly different.

---

Individual Value Plot of Cupertino, San Jose, Santa Clara