STAT 518 --- Section 4.2 --- Tests for $r \times c$ Tables

- We now consider more general two-way tables:

- In Sec. 4.1 we had two samples in which a two-category variable is measured on each individual in each sample.

- Now suppose we have $r$ samples in which the same $c$-category variable is measured on each individual in each sample.

Comparing Multinomial Probabilities Across Several Independent Samples

- Suppose we have $r$ independent samples, with respective sizes $n_1, n_2, \ldots, n_r$. We classify each individual in each sample into class 1, 2, ..., $c$.

- Our data (which could be nominal or ordinal) could be arranged in an $r \times c$ table as follows:

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th>Class 2</th>
<th>\ldots</th>
<th>Class $c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>$O_{11}$</td>
<td>$O_{12}$</td>
<td>\ldots</td>
<td>$O_{1c}$</td>
</tr>
<tr>
<td>Sample 2</td>
<td>$O_{21}$</td>
<td>$O_{22}$</td>
<td>\ldots</td>
<td>$O_{2c}$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
</tr>
<tr>
<td>Sample $r$</td>
<td>$O_{r1}$</td>
<td>$O_{r2}$</td>
<td>\ldots</td>
<td>$O_{rc}$</td>
</tr>
<tr>
<td>\hline</td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>\ldots</td>
<td>$c_c$</td>
</tr>
<tr>
<td>\hline</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>\ldots</td>
<td>$n_r$</td>
</tr>
<tr>
<td>\hline</td>
<td>$N$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chi-Square Test for Homogeneity in a Two-Way Table

• This is a basic extension of the two-tailed z-test comparing \( p_1 \) and \( p_2 \).

Hypotheses: \( H_0: p_{ij} = p_{2j} = \cdots = p_{kj} \) for all \( j \).

\( H_1: p_{ij} \neq p_{kj} \) for some \( j \) and for some \( i, k \).

Test Statistic

\[
T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}} - N, \quad \text{where} \quad E_{ij} = \frac{n_i c_j}{N}
\]

which has an asymptotic \( \chi^2 \) distribution with \( (r-1)(c-1) \) degrees of freedom when \( H_0 \) is true.

• Note if \( H_0 \) is true and all the populations have the same set of class probabilities, the expected count in cell \((i, j)\) is the size of the i-th sample times the proportion of observations (of all \( N \)) falling in category \( j \).

• If \( r = c = 2 \), this \( T = T_1^2 \) from Section 4.1.

• If \( T \) is far from zero, this indicates that \( H_0 \) is false and that the probability distribution differs among the \( r \) populations.

Decision Rule:

Reject \( H_0 \) if \( T > \chi^2_{1-\alpha, (r-1)(c-1)} \) found in Table A2.
• The P-value is found through interpolation in Table A2 or using R.

• Note: The $\chi^2$ approximation for $T$ is valid for large samples, say, if all Eij's are greater than 0.5 and at least half of the Eij's are greater than 1.0.

• If some expected cell counts are too small, two or more categories could be combined, as long as this is sensible.

Example 1: Page 202 gives test score category counts from a sample of public school students and from a sample of private school students. Is the probability distribution of scores equal for public and private school students? Use $\alpha = 0.05$.

$$E_{11} = \frac{(46)(36)}{128} = 12.94, \quad E_{12} = \frac{(46)(46)}{128} = 16.53, \quad E_{13} = \frac{(46)(34)}{128} = 12.22, \quad \text{etc.}$$

Data:

<table>
<thead>
<tr>
<th>Score</th>
<th>Low</th>
<th>Marginal</th>
<th>Good</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private</td>
<td>6</td>
<td>14</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Public</td>
<td>30</td>
<td>32</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>46</td>
<td>34</td>
<td>12</td>
</tr>
</tbody>
</table>

$H_0$: $p_{ij} = p_{2j}$ for all $j = 1, 2, 3, 4$

$H_1$: $p_{ij} \neq p_{2j}$ for some $j$

Test statistic:

$$T = \frac{6^2}{12.94} + \frac{14^2}{16.53} + \frac{17^2}{12.22} + \frac{9^2}{4.31} + \frac{30^2}{23.06} + \frac{32^2}{29.47} + \frac{17^2}{21.78} + \frac{3^2}{7.69} - 128 = 17.29$$

Decision rule and conclusion:

Reject $H_0$ if $T > \chi^2_{.95, 3} = 7.815$ (Table A2). Since $17.29 > 7.815$, we reject $H_0$ and conclude the probability distribution differs for public and private school students.

P-value $\approx .0006$ from R
Chi-Square Test for Independence

- Now we consider observations in a single sample of size $N$ that are classified according to two categorical variables.

- Such data can also be presented in a two-way table.

Example: Suppose the people in the “favorite-sport” survey had been further classified by gender:

<table>
<thead>
<tr>
<th>Gender</th>
<th>Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
</tr>
</tbody>
</table>

- Two categorical variables: Gender and Sport

Question: Are the two classifications independent or dependent?

- For instance, does people’s favorite sport depend on their gender? Or does gender have no association with favorite sport?

- Unlike the $r$-sample problem, in this situation both column totals and row totals are random (only $N$ is fixed).
Observed Counts for a $r \times c$ Contingency Table
($r =$ # of rows, $c =$ # of columns)

<table>
<thead>
<tr>
<th>Column Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Row</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Variable</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Col. Totals</td>
</tr>
</tbody>
</table>

Probabilities for a $r \times c$ Contingency Table:

<table>
<thead>
<tr>
<th>Column Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>Row</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Variable</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Note: If the two classifications are independent, then:
$p_{11} = (p_{\text{row } 1})(p_{\text{col } 1})$ and $p_{12} = (p_{\text{row } 1})(p_{\text{col } 2})$, etc.

So under the hypothesis of independence, we expect the cell probabilities to be the product of the corresponding marginal probabilities:

$$p_{ij} = (p_{\text{row } i})(p_{\text{col } j})$$
Hence if $H_0$ is true, the (estimated) expected count in cell $(i, j)$ is simply:

$$N_{ij} = N(p_{rowi})(p_{coli}) \approx N\left(\frac{R_i}{N}\right)\left(\frac{C_j}{N}\right) = \frac{R_i C_j}{N}$$

$\chi^2$ test for independence

$H_0$: The classifications are independent
$H_a$: The classifications are dependent

Test statistic:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{O_{ij}^2}{E_{ij}}\right) - N$$

where the expected count in cell $(i, j)$ is

$$E_{ij} = \frac{R_i C_j}{N}$$

Decision Rule:

Reject $H_0$ if $T > \chi^2_{1-\alpha, (r-1)(c-1)}$

\[ \uparrow \text{Table A2} \]

- The P-value is found through interpolation in Table A2 or using R.

Note: The same large-sample rule of thumb applies as in the previous $\chi^2$ test.
Example: Does the incidence of heart disease depend on snoring pattern? (Test using $\alpha = .05$.) Random sample of 2484 adults taken; results given in a contingency table:

<table>
<thead>
<tr>
<th>Snoring Pattern</th>
<th>Never</th>
<th>Occasionally</th>
<th>≈Every Night</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Disease</td>
<td>Yes</td>
<td>24</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1355</td>
<td>603</td>
</tr>
<tr>
<td>Total</td>
<td>1379</td>
<td>638</td>
<td>467</td>
</tr>
</tbody>
</table>

Expected Cell Counts:

$E_{11} = \frac{(110)(1379)}{2484} = 61.07$, $E_{12} = \frac{(110)(638)}{2484} = 28.25$,

$\cdots$, $E_{23} = \frac{(2374)(467)}{2484} = 446.32$

Test statistic:

$T = \frac{24^2}{61.07} + \frac{35^2}{28.25} + \frac{51^2}{20.68} + \frac{1355^2}{1317.93} + \frac{603^2}{609.75} + \frac{416^2}{446.32} - 2484 = 71.75$

Decision rule and conclusion:

$(r-1)(c-1) = 2$, so

Reject $H_0$ if $T > \chi^2_{.95, 2} = 5.991$. Since $71.75 > 5.99$, we reject $H_0$ and conclude the incidence of heart disease is associated with snoring pattern.

P-value $\approx 0$ from R.
Tests for $r \times c$ Tables with Fixed Marginal Totals

- If the table has $r$ rows and $c$ columns and both the row totals and column totals are fixed, an extended version of the Exact Test is available.

- In this case, there are no one-tailed alternatives possible – the hypotheses are simply the same as for the $\chi^2$ test for homogeneity or the $\chi^2$ test for independence, depending on the sampling setup.

- The P-value are obtained using `fisher.test` in R, as the exact null distribution is cumbersome.

- The exact P-value is obtained by considering all possible tables resulting in the given margins, and sorting these by how favorable to $H_1$ they are.

- The exact P-value is the proportion of possible tables that are as or more favorable to $H_1$ as the table we observed.

Example Data (alteration of bank data to a $3 \times 3$ table):

<table>
<thead>
<tr>
<th>Race</th>
<th>Position</th>
<th>Acct. Rep</th>
<th>Teller</th>
<th>Data Analyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td></td>
<td>0</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Asian</td>
<td></td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

P-value and conclusion: $P$-value $= 0.0566$ from R. At $\alpha = 0.05$, cannot conclude the probabilities of the various jobs differ across the races.
Section 4.3 --- Median Test

• We return to the situation in which we want to know whether several (c) populations have the same median.

• For c > 2, this is similar to the setup of the Kruskal-Wallis test.
• For c = 2, this is similar to the setup of the Mann-Whitney test.

• The difference is in the conditions of the tests:
The M-W and K-W tests assume that under $H_0$, the c populations have identical distributions. While the Median Test assumes only that under $H_0$, the c populations have the same median.
• So the Median Test can be applied more generally.

• Suppose from each of c populations, we have a random sample, with sizes $n_1, n_2, \ldots, n_c$.

• We assume that the c samples are independent and that the data are at least ordinal, so that the “median” is a meaningful measure.

• Calculate the grand median of all $N = n_1 + n_2 + \ldots + n_c$ observations, and arrange the data into a $2 \times c$ table:

\[
\begin{array}{cccc}
 & 1 & 2 & \cdots & c \\
\text{Sample} & 0_{11} & 0_{12} & \cdots & 0_{1c} \\
\gt \text{Grand Median} & 0_{21} & 0_{22} & \cdots & 0_{2c} \\
\lt \text{Grand Median} & n_1 & n_2 & \cdots & n_c \\
\end{array}
\]
Hypotheses:

$H_0$: All $c$ populations have the same median
$H_1$: At least 2 populations have different medians

- The null hypothesis implies that being in the top row or bottom row is independent of which column (population) an observation is in.

- Note that the expected cell count under $H_0$ is

$$E_{1i} = \frac{n_i a}{N}$$

for the top-row cells, and

$$E_{2i} = \frac{n_i b}{N}$$

for the bottom-row cells.

So the test statistic, as in the $\chi^2$ test for independence, is

$$T = \sum_{i=1}^{c} \frac{(O_{1i} - \frac{n_i a}{N})^2}{\frac{n_i a}{N}} + \sum_{i=1}^{c} \frac{(O_{2i} - \frac{n_i b}{N})^2}{\frac{n_i b}{N}}$$

which can be simplified into

$$T = \frac{N^2}{ab} \sum_{i=1}^{c} \frac{(O_{1i} - \frac{n_i a}{N})^2}{n_i} = \left( \frac{N^2}{ab} \sum_{i=1}^{c} \frac{O_{1i}^2}{n_i} \right) - \frac{N a}{b}$$

since

$$O_{2i} = n_i - O_{1i}$$

- The asymptotic null distribution of $T$ is $\chi^2_{c-1}$

Decision rule:

Reject $H_0$ if $T > \chi^2_{1-\alpha, c-1}$
• The P-value is found through interpolation in Table A2 or using R.

Note: The same large-sample rule of thumb applies as in the previous $\chi^2$ test.

• The median test may be generalized to test about any particular quantile – in that case, the appropriate “grand quantile” is used instead of the “grand median”.

Example 1: Bidding/Buy-It-Now Data from Section 5.1 notes. At $\alpha = .05$, are the median selling prices significantly different for the two groups?
Data:
Bidding: 199, 210, 228, 232, 245, 246, 246, 249, 255
BIN: 210, 225, 225, 235, 240, 250, 251

Grand Median: 237.5 \quad c = \underline{2}. \quad 2 \times c \text{ table:}

<table>
<thead>
<tr>
<th></th>
<th>Bidding</th>
<th>BIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grand Median</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>$\geq$ Grand Median</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$\leq$ Grand Median</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

16

Test statistic $T = \frac{16^2}{(8)(8)} \left( \frac{5^2}{9} + \frac{3^2}{7} \right) - \frac{(16)(8)}{(8)} = 0.254$

Decision Rule and Conclusion: \text{Reject } H_0 \text{ if } T > \chi^2_{.95,1} = 3.84 \text{ Table}

Since 0.254 $\leq$ 3.84, we fail to reject $H_0$. The two $A_2$ methods may have the same median price.

P-value \approx 0.614 from R.
Example 2: Data on page 221 gives corn yields for four different growing methods. At $\alpha = .05$, are the median yields significantly different for the four methods?

Grand Median: $89$  \hspace{0.5cm}  $c = \frac{4}{4}$  \hspace{0.5cm}  $2 \times c$ table:

\begin{center}
\begin{tabular}{cccc}
Method & 1 & 2 & 3 & 4 \\
$\geq$ Grand Median & 6 & 3 & 7 & 0 & 16 \\
$\leq$ Grand Median & 3 & 7 & 10 & 8 & 18 \\
 & 9 & 10 & 7 & 8 & 34 \\
\end{tabular}
\end{center}

Test statistic

\[ T = \frac{34^2}{(16)(18)} \left( \frac{6^2}{9} + \frac{3^2}{10} + \frac{7^2}{7} + \frac{0^2}{8} \right) - \frac{(34)(16)}{18} = 17.54 \]

Decision Rule and Conclusion:

Reject $H_0$ if $T > \chi^2_{0.05,3} = 7.815$. Since $17.54 > 7.815$ we reject $H_0$ and conclude the median yields differ among the 4 methods.

P-value $\approx .0005$ from R.

Comparison of Median Test to Competing Tests

- The classical parametric approach for comparing the centers of several populations is the \textit{ANOVA F-test}.

- In Sec. 5.1 we examined the efficiency of the Mann-Whitney test relative to the median test when $c = 2$.

- Of these options, the median test is the most flexible since it makes the fewest assumptions about the data.

- The A.R.E. of the median test relative to the F-test is 0.64 with normal populations and 2.00 with double exponential (heavy-tailed) populations.