## Contents

5 Tests for Trends and Association 2

5.1 Association between Quantitative Variables ............... 3
  5.1.1 Pearson’s Correlation Coefficient .................. 4
  5.1.2 Slope of Least Squares Line .................... 6
  5.1.3 Permutation Test for $\rho$ or $\beta_1$ ................. 7
  5.1.4 Spearman Rank Correlation ....................... 12
  5.1.5 Kendall’s $\tau$ .................................. 16

5.2 Qualitative Variables .................................. 22
  5.2.1 Contingency Tables ................................ 22
  5.2.2 Chi-square Test for Association .................... 24
  5.2.3 Fisher’s Exact Test for a $2 \times 2$ Contingency Table 32
Example 1: homework and final grade The following data shows a subset of the homework and final exam grade of ST372 students in Fall 2007.

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>0</td>
<td>96</td>
<td>65</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>Final exam</td>
<td>0</td>
<td>166</td>
<td>130</td>
<td>118</td>
<td>130</td>
</tr>
</tbody>
</table>

Question: How to assess the relationship between homework and final exam?
Example 2: reading ability. Is there a significant positive correlation between the rankings of 10 children on a reading test $X$ and their teacher’s ranking of their reading ability $Y$?

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Y</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

We first consider assessing the relationship between two continuous variables $X$ and $Y$ by using

- correlation coefficient $\rho$
- slope of least squares regression line $\beta_1$
5.1 Association between Quantitative Variables

5.1.1 Pearson’s Correlation Coefficient

Pearson’s Correlation Coefficient:

\[ \rho = \frac{E\{(X - \mu_x)(Y - \mu_y)\}}{\sigma_x \sigma_y}, \]

where

- \( \mu_x \) and \( \mu_y \) are the means of \( X \) and \( Y \);
- \( \sigma_x \) and \( \sigma_y \) are the standard deviations of \( X \) and \( Y \);
- the numerator is \( Cov(X, Y) \).

Some properties of \( \rho \):

- measures the linear relationship between \( X \) and \( Y \);
- \( -1 \leq \rho \leq 1 \);
• $\rho = 0 \Rightarrow$ no linear relationship;

• $\rho > 0$: positive linear association, $Y$ tends to increase as $X$ increases, and vice versa.

Suppose we observe $(X_i, Y_i), i = 1, \cdots, n$. We can estimate $\rho$ by the Sample correlation coefficient:

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{S_{xy}}{S_x S_y},$$

where

• $S_{xy} = 1/(n - 1) \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$

• $S_x^2 = 1/(n - 1) \sum_{i=1}^{n} (X_i - \bar{X})^2$

• $S_y^2 = 1/(n - 1) \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

**Hypothesis test:**

• If $(X_i, Y_i), i = 1, \cdots, n$ is a random sample from a bivariate
normal distribution, then we can test $H_0: \rho = 0$ with the test statistic

$$t_{corr} = \sqrt{\frac{n - 2}{1 - r^2}} \sim t_{n-2} \text{ under } H_0.$$  

That is,

- $H_a: \rho \neq 0$, reject $H_0$ if $|t_{corr}| > t_{\alpha/2,n-2}$$
- $H_a: \rho > 0$, reject $H_0$ if $t_{corr} > t_{\alpha,n-2}$
- $H_a: \rho < 0$, reject $H_0$ if $t_{corr} < -t_{\alpha,n-2}$

### 5.1.2 Slope of Least Squares Line

Simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

- $\beta_0$: intercept;
- $\beta_1$: slope, $\beta_1 = 0 \Rightarrow$ no linear relationship between $Y_i$ and $X_i$;
• \( \epsilon_i \): random error with mean 0 and finite variance.

Least squares estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are the minimizers of

\[
SSE = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i,
\]

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = r \frac{S_y}{S_x},
\]

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.
\]

If \( \epsilon_i \) i.i.d. are normally distributed, we can test \( H_0 : \beta_1 = \beta_{10} \) with

\[
t_{slope} = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}} \sim t_{n-2} \text{ under } H_0,
\]

where \( MSE = SSE/(n - 2) \).
5 Tests for Trends and Association

5.1.3 Permutation Test for $\rho$ or $\beta_1$

The validity of the $t$-tests require the normal distribution assumption. When we are not willing to make distributional assumptions, we can perform permutation test to obtain the reference null distribution of $\hat{\rho}$ or $\hat{\beta}_1$.

Under $H_0: \rho = 0$ or $H_0: \beta_1 = 0$,

- $Y_i$ is likely to appear with $X_j$ as it is to appear with $X_i$ for $j \neq i$;
- i.e. all $n!$ ways of arranging the $Y_i$’s with the $X_i$’s are equally likely under $H_0$.

Steps:

- Calculate $\hat{\beta}_{1,obs}$ (or $r_{obs}$).
- Permute the $Y$’s among the $X$’s in $n!$ ways (or a sample $R$ of the permutations). That is, keep the order of $X$ unchanged and permute $Y$. 
• For each permutation, calculate $\hat{\beta}_1^*$ or $r^*$.

• Upper-tailed test $H_0 : \beta_1 > 0$:

$$p\text{-value} = \frac{\#\hat{\beta}_1^*'s \geq \hat{\beta}_{1,\text{obs}}}{R}.$$ 

• Lower-tailed test $H_0 : \beta_1 < 0$:

$$p\text{-value} = \frac{\#\hat{\beta}_1^*'s \leq \hat{\beta}_{1,\text{obs}}}{R}.$$ 

• Two-tailed test $H_0 : \beta_1 \neq 0$:

$$p\text{-value} = \frac{\#|\hat{\beta}_1^*|'s \geq |\hat{\beta}_{1,\text{obs}}|}{R}.$$ 

**Remark.** Recall that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r \frac{S_y}{S_x},$$

and $S_y$ and $S_x$ are unchanged with permutations. Therefore, it's
equivalent to base the test on

- \( r = \frac{S_{xy}}{S_x S_y} \)
- or \((n - 1)S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y} \)
- or \( \sum_{i=1}^{n} X_i Y_i \)

**Large sample approximation for** \( r \): \( Var(r) = \frac{1}{n-1} \), so for large \( n \),

\[
Z_r = \frac{r}{\sqrt{1/(n-1)}} = r \sqrt{n-1} \sim N(0, 1) \text{ approximately.}
\]

**Example 5.1.1** *ST745* grades. \( X_i \): *middle term exam score*, \( Y_i \): *final exam score*, \( i = 1, \cdots, 21 \). We have the following summary:

\[
\sum_{i=1}^{n} X_i = 1956, \sum_{i=1}^{n} Y_i = 1917, \sum_{i=1}^{n} X_i Y_i = 179203, \sum_{i=1}^{n} X_i^2 = 182738, \sum_{i=1}^{n} Y_i^2 = 176499.
\]

*Calculate* \( r \) *and* \( \hat{\beta}_1 \), *and test* \( H_0 : \beta_1 = 0 \).
See R code for the hypothesis testing results.
5.1.4 Spearman Rank Correlation

Consider the following data:

<table>
<thead>
<tr>
<th>$X_i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_i$</td>
<td>1</td>
<td>16</td>
<td>81</td>
<td>256</td>
<td>625</td>
<td>1296</td>
<td>2401</td>
</tr>
</tbody>
</table>

- The sample correlation coefficient: $r = .89$.
- However, there is a perfect relationship: $Y_i = X_i^4$.
- Pearson’s CC can only capture the linear relationship.

Notice that for the above data: when the rank of $X_i$ increases, the rank of $Y_i$ also increases.

Instead of limiting our definition of association to linear relationship, we consider measuring the extent to which $Y$ increases with $X$ by comparing the ranks of $X_i$’s with those of $Y_i$’s.
• **Spearman’s rank correlation** \((r_s)\) is the standard Pearson correlation applied to the ranks of \(X_i\)'s and the ranks of \(Y_i\)'s.

• \(r_s\) measures how well one variable is monotonically dependent on the other variable. When there are no ties, \(r_s = 1\) (or \(-1\)) means one variable is a perfect monotone increasing (or decreasing) function of the other.

• Table A12 gives the limited critical values for the distribution of \(r_s\) under \(H_0\).

• Large sample approximation: for large \(n\),

\[
Var(r_s) = \frac{1}{n - 1}, \quad Z = \frac{r_s}{\sqrt{Var(r_s)}} = r_s \sqrt{n - 1} \sim N(0, 1)
\]

approximately under \(H_0\) : no association between \(X\) and \(Y\).

• One treatment of ties
  – use midranks to ties among \(X_i\)'s (or \(Y_i\)'s)
– then apply Pearson correlation to the ranks adjusted for ties
– use permutation to obtain an exact test
– or use the large sample approximation

• There exists some other complicated adjustments for ties but we recommend apply the permutation to ranks.

For the above artificial data:

R code and outputs:

```
x  1  2  3  4  5  6  7
y  1 16  81 256 625 1296 2401
> cor(x,y)
[1] 0.8903055
> rank(x)
[1] 1 2 3 4 5 6 7
> rank(y)
[1] 1 2 3 4 5 6 7
> cor(rank(x), rank(y))
[1] 1
```
**Example 5.1.2** Calculate the Spearman coefficient for the “Reading ability” data set, and test $H_0 : r_s = 0$ versus $H_a : r_s > 0$

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>9</td>
</tr>
</tbody>
</table>

```r
> # Spearman correlation
> (rs.obs = cor(x, y))
[1] 0.9272727

> # permutation test for the Spearman correlation
> permr <- perm.approx.r(x, y, 1000)
> mean(permr >= rs.obs)
[1] 0
```

From Table A12, $p$-value $= P(r_s \geq 0.927) > P(r_s > 0.78) = 0.005$. Q: Carry out the test based on the large sample approximation.
Example 5.1.3 Scores (with ties) of ten projects at a science fair:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>JudgeA x</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>JudgeB y</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

> (x = rank(x))
[1]  7.5  7.5  4.5  7.5  1.0  2.5  2.5 10.0  7.5  4.5
> (y = rank(y))
[1]  6.0  8.0  8.0  2.5  4.5  1.0  2.5  8.0  4.5 10.0
>
> ## Spearman’s rank correlation
> (rs.obs = cor(x, y))
[1] 0.3750694
>
> ## permutation test for the Spearman correlation
> permr <- perm.approx.r(x, y, 1000)
> mean(permr >= rs.obs)
[1] 0.131
5.1.5 Kendall’s $\tau$

For this measure of association, we consider whether pairs are concordant or discordant.

Consider the exam1 score $X_i$ and exam2 scores $Y_i$ of two students.

- A: $X_1 = 43$, $Y_1 = 64$
- B: $X_2 = 89$, $Y_2 = 72$

Note that as exam1 score increases from subject A to subject B, the exam2 score also increases, i.e. B performs better than A consistently in two exams.

We say a pair of points $(X_i, Y_i)$ and $(X_j, Y_j)$ are

- **concordant** if

\[ X_i < X_j \Rightarrow Y_i < Y_j, \text{ or } (X_i - X_j)(Y_i - Y_j) > 0 \]
• discordant if

\[ X_i < X_j \Rightarrow Y_i > Y_j, \text{ or } (X_i - X_j)(Y_i - Y_j) < 0. \]

We say that \(X\)'s and \(Y\)'s have

• a positive association if pairs are more likely to be concordant than discordant;

• a negative association if pairs are more likely to be discordant than concordant;

• no association if pairs are equally likely to be discordant or concordant.

Assuming no ties, Kendall’s \(\tau\) is defined as

\[ \tau = 2P\{(X_i - X_j)(Y_i - Y_j) > 0\} - 1, \]

so that \(\tau \in [-1, 1]\).

**Estimation of \(\tau\) (standard approach)**
For $i = 1, \cdots, n, j = 1, \cdots, n$, let

$$U_{ij} = \begin{cases} 
1, & (X_i - X_j)(Y_i - Y_j) > 0 \\
0, & (X_i - X_j)(Y_i - Y_j) < 0 \\
1/2, & (X_i - X_j)(Y_i - Y_j) = 0,
\end{cases}$$

and

$$V_i = \sum_{j=i+1}^{n} U_{ij},$$

that is, the number of pairs $(X_j, Y_j)$ concordant with $(X_i, Y_i)$ for $j \geq i + 1$. Then

$$\sum_{i=1}^{n-1} V_i / \binom{n}{2}$$

is the fraction of concordant pairs. One estimation of $\tau$:

$$r_\tau = 2 \left( \frac{\sum_{i=1}^{n-1} V_i}{\binom{n}{2}} \right) - 1$$
Estimation of $\tau$ (simpler approach, equivalent when no ties)

1. Order the paired data $(X_i, Y_i)$ so that $X$’s are in the increasing order $X_1 < X_2 < \cdots < X_n$.

2. Count the number of pairs $(Y_i, Y_j)$ such that $Y_i < Y_j$. If $Y_i = Y_j$, add $1/2$ to the sum.

3. $r_\tau = 2 \left( \sum \text{in step 2} \right) \left( \begin{array}{c} n \\ hline 2 \end{array} \right) - 1$

Large sample approximation for distribution of $r_\tau$: under $H_0$: no association, $r_\tau$ is approximately normal with $E(r_\tau) = 0$ and

$$\text{Var}(r_\tau) = \frac{4n + 10}{9(n^2 - n)}.$$

Note: the variance needs be adjusted when there are ties (see Higgins).
Example 5.1.4 A subset of the ST372 grades:

<table>
<thead>
<tr>
<th>student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>homework $X_i$</td>
<td>0</td>
<td>96</td>
<td>65</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>final exam $Y_i$</td>
<td>0</td>
<td>166</td>
<td>130</td>
<td>118</td>
<td>130</td>
</tr>
</tbody>
</table>

Step1: order the pairs

| homework $X_i$ | 0  | 56 | 58 | 65 | 96 |
| final exam $Y_i$ | 0  | 130| 118| 130| 166|

Step2: total $4 + 1.5 + 2 + 1 = 8.5$ pairs $(Y_i, Y_j)$ such that $Y_i < Y_j$, $\binom{5}{2} = 10$.

Step3: $r_\tau = \frac{2 \cdot 8.5}{10} - 1 = 0.7$. See R code for permutation test.
5.2 Qualitative Variables

5.2.1 Contingency Tables

Suppose individuals are placed into categories according to two characteristics. The two-way contingency table displays the counts of individuals falling into each of the categories. For example:

- Simple Random Sample (SRS) with questions: “favorite member of Beatles” and “favorite member of U2”

<table>
<thead>
<tr>
<th>Beatles</th>
<th>Bone</th>
<th>Edge</th>
<th>Larry</th>
<th>Adam</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>Paul</td>
<td>14</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>George</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Ringo</td>
<td>5</td>
<td>9</td>
<td>10</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>33</td>
<td>14</td>
<td>9</td>
<td>98</td>
</tr>
</tbody>
</table>
• Stratified sample: attitudes about Jell-O

<table>
<thead>
<tr>
<th></th>
<th>Hate</th>
<th>Neutral</th>
<th>Love</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utahns</td>
<td>10</td>
<td>20</td>
<td>70</td>
<td>100</td>
</tr>
<tr>
<td>Californians</td>
<td>50</td>
<td>40</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Alaskans</td>
<td>20</td>
<td>60</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

• Designed experiment (CRD):

<table>
<thead>
<tr>
<th></th>
<th>No benefit</th>
<th>Mild benefit</th>
<th>Strong benefit</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug A</td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>Drug B</td>
<td>15</td>
<td>15</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>
5.2.2 Chi-square Test for Association

Observations in an $r \times c$ contingency table:

<table>
<thead>
<tr>
<th></th>
<th>Col 1</th>
<th>Col 2</th>
<th>⋮</th>
<th>Col c</th>
<th>Row Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 1</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>⋮</td>
<td>$n_{1c}$</td>
<td>$n_1.$</td>
</tr>
<tr>
<td>Row 2</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
<td>⋮</td>
<td>$n_{2c}$</td>
<td>$n_2.$</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>Row $r$</td>
<td>$n_{r1}$</td>
<td>$n_{r2}$</td>
<td>⋮</td>
<td>$n_{rc}$</td>
<td>$n_r.$</td>
</tr>
<tr>
<td>Col Totals</td>
<td>$n.1$</td>
<td>$n.2$</td>
<td>⋮</td>
<td>$n.c$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Two different cases:

- Case 1 (SRS): all $n$ individuals are randomly selected and classified according to row/column characteristics.
- Case 2 (stratified or CRD):
  - a fixed number $n_i.$ is selected according to row characteristics,
\(i = 1, \cdots, r\)

– then classified according to column characteristics

**Hypotheses:**

- **Case 1:** \(H_0 : p_{ij} = p_i \cdot p_j\) (independence), where
  \[
  p_{ij} = \frac{E(n_{ij})}{n}, \quad p_i = \sum_{j=1}^{c} p_{ij}, \quad p_j = \sum_{i=1}^{r} p_{ij},
  \]
  and \(p_{ij}\) is the expected proportion of the cell \((i, j)\), \(p_i\) is the expected proportion of row \(i\), and \(p_j\) is the expected proportion of the column \(j\).

- **Case 2:** \(H_0 : p_{j|i} = p_{j|i'}\) for all \(i, i'\) and \(j\) (homogeneity), where
  \[
  p_{j|i} = \frac{p_{ij}}{p_i}.
  \]
  is the conditional probability of column \(j\) given row \(i\). E.g. for the example “attitudes about Jell-O”, \(p_{1|2}\): the expected proportion of
Californians who hate Jell-O.

- The two null hypotheses are equivalent: test if there is any association between the row and the column factors.

**Chi-square test statistic:**

- Observed counts in each cell: $n_{ij}$
- Expected counts under $H_0$:
  \[
  e_{ij} = \frac{n_i n_j}{n}
  \]
- Chi-square test statistic
  \[
  \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.
  \]
- If $e_{ij} \geq 5$ for all $i, j$, then $\chi^2$ is distributed $\chi^2_{(r-1)(c-1)}$ under $H_0$. 

Permutation $\chi^2$ test: when some $e_{ij} < 5$, the chi-square distribution may not be valid. But we can still create permutation distribution of the $\chi^2$ statistic under $H_0$.

**Example 5.2.1** satisfaction with pain-relief treatment versus gender

<table>
<thead>
<tr>
<th></th>
<th>Not satisfied</th>
<th>Somewhat satisfied</th>
<th>Very satisfied</th>
<th>row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Male</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>col total</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

The expected counts: $e_{11} = \frac{4 \times 2}{7} = 8/7$, $e_{12} = 12/7$, $e_{13} = 8/7$, $e_{21} = 6/7$, $e_{22} = 9/7$, $e_{23} = 6/7$. So

$$\chi^2_{obs} = \frac{(8/7 - 2)^2}{8/7} + \cdots + \frac{(6/7 - 2)^2}{6/7} = 4.28.$$
The permutation null distribution of $\chi^2$. Under $H_0$,

- all assignments of the 4 females and 3 males to the 3 column groups are equally likely;
- or equivalently, all assignments of 2 non-, 3 somewhat-, and 2 very-satisfied to the genders are equally likely

**Steps for the permutation chi-square test:**

1. Calculate $\chi^2_{obs}$

2. For each permutation, randomly choose $n_i$ of the column labels to be placed in row $i$ and calculate $\chi^2_*$ for each permutation. For the pain-relief treatment example,
   - there are total 7 subjects, having 7 labels $N_1, N_2, S_3, S_4, S_5, V_6, V_7$
   - assigning 7 labels: 4 to one treatment, and 3 to the other treatment has $\frac{7!}{4!3!} = 35$ ways
3. Calculate \( p\)-value = \#\{\chi^2* > \chi^2_{obs}\}/R \), where \( R \) is the number of permutations (or a sample of all permutations).

**Simple implementation of Step 2:**

- Create vectors \( x \) and \( y \) (length \( n \)) with row labels \( (1, \cdots, r) \) and column labels \( (1, \cdots, c) \) as elements. For the pain-relief treatment example

  \[
  \begin{array}{ccccccc}
  x & 1 & 1 & 1 & 1 & 2 & 2 & 2 \\
  y & 1 & 1 & 2 & 2 & 2 & 3 & 3 \\
  \end{array}
  \]

- Randomly permute values in \( x \) (or in \( y \)) while keeping the other vector unchanged to get table and \( \chi^2* \) statistic. Example tables for pain-relief example under \( H_0 \) (note that each table has the same row totals and column totals)

  Permutated \( x \): 1, 1, 2, 2, 1, 1, 2 gives the permuted frequency table:
Permutated $x$: 1, 2, 1, 2, 2, 1, 1 gives the permuted frequency table:

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>S</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Example 5.2.2 See the R code for the permutation test of the following examples

- Satisfaction v.s. Gender
- Gender v.s. Party
- Presidential preference v.s. Region
5.2.3 Fisher’s Exact Test for a $2 \times 2$ Contingency Table

Permutation test applied to a $2 \times 2$ contingency table is Fisher’s exact test. Fisher’s exact test is used for small sample size, as p-value can be calculated exactly under the null hypothesis rather than based on large sample approximation. Fisher’s exact test is named after its inventor R. A. Fisher.

The lady tasting tea experiment: The lady, Muriel Bristol, claimed that she was able to tell whether the tea or the milk was added first to a cup. Fisher prepared 8 cups of tea, 4 with tea added first and 4 with milk added first. The lady was informed of the design (4 tea first, 4 milk first). Then Fisher presented the 8 cups to her in random order. She was asked to identify the 4 cups with milk first. Below is the result:
The question is: does the lady have the discriminating skill? What’s the probability that she got such answers when everything is due to chance (p-value)?

The null and alternative hypotheses:

\[ H_0: \text{there is no association between the true order of pouring and the lady’s guess} \]

versus \[ H_a: \text{there is a positive association.} \]

We can generalize the table as follows:
- Since the design fixes the row and column totals to 4 each, the entire table is fixed after $X$ is chosen ($X = 3$ in the lady tea example).

- Rephrase the problem. There are total 8 cups, among which 4 have milk added first ("success"). The lady is asked to choose 4 cups (that she believes has milk added first). $X$ is the number of milk-first cups among the 4 that the lady choose, that is, the number of success among 4 randomly chosen from the population.

- There are total $\binom{8}{4}$ of ways of choosing 4 cups among 8.
5 Tests for Trends and Association

• Suppose $H_0$ is true, i.e. the lady has no discriminating skill. Then all $\binom{8}{4}$ are equally likely. Under $H_0$, $X$ follows the hypergeometric distribution $Hyper(m = 4, n = 4, k = 4)$.

• $Hyper(m, n, k)$ is a discrete distribution. The hypergeometric distribution can be understood by using the urn model. Suppose a urn has total $n$ black marbles, and $m$ white marbles (“successes”). Suppose you are asked to draw $k$ marbles from the urn without replacement, and denote $X$ as the number of white marbles you get among $k$. Then $X \sim Hyper(m, n, k)$ and

$$P(X = x) = \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{m+n}{k}},$$

for $\max\{0, k - \min(n, k)\} \leq x \leq \min(m, k)$.

• Under $H_0$:

$$P(X = 3) = \frac{\binom{4}{3} \binom{4}{1}}{\binom{8}{4}} = 0.229$$
\[ P(X = 4) = \frac{\binom{4}{4} \binom{4}{0}}{\binom{8}{4}} = 0.014 \]

- In R, function `dhyper(x, m, n, k)` gives \( P(X = x) \) for hypergeometric distribution.

```r
dhyper(0:4, m, n, k)
# 0.01428571 0.22857143 0.51428571 0.22857143 0.01428571
```

- The probability that \( X = 3 \) is equivalent to the probability that we get exactly 3 white marbles among 4 draws in the urn containing total 4 white marbles and 4 black marbles.

- For testing \( H_0 \): no association between the true order and the lady’s guess  
  versus  
  \( H_a \): there is a positive association (i.e. the lady has the discriminating skill).
Then $H_a$ implies that $X$ is large and

\[ p\text{-value} = P(X \geq 3) = P(X = 3) + P(X = 4) = 0.243. \]

So there is no significant positive association.

**Example 5.2.3** Cross-classification of 13 states by presidential preference and region

<table>
<thead>
<tr>
<th></th>
<th>Bush</th>
<th>Kerry</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>West</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>East</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
</tbody>
</table>

Use Fisher’s exact test to test $H_0$: no association between region and preference versus $H_a$: western states prefer Bush more.