1 Contingency Tables

Measure association between categorical variables. Similar to Pearson’s Correlation Coefficient for numerical variables.

1.1 Probability Structure

Distribution of ONE categorical variable \(X\), with possible outcomes \(x_1, \ldots, x_k\).
The distribution is described by the outcomes and their probabilities: \(\pi_i = P(X = x_i)\), with

\[
\sum_i \pi_i = 1
\]

The distribution table of the random variable \(X\):

\[
\begin{array}{c|c}
  x_i & \pi_i \\
  x_1 & \pi_1 \\
  x_2 & \pi_2 \\
  \vdots & \vdots \\
  x_k & \pi_k \\
\end{array}
\]

Example 1
Random variable \(X=\) Use of phone while driving(Yes/No), therefore categorical

<table>
<thead>
<tr>
<th>(x)</th>
<th>(P(X = x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Joint distribution of two categorical random variables lead to probability tables.
Let \(X\) be a random variable with \(I\) categories,
Let \(Y\) be a random variable with \(J\) categories.
Then the joint distribution of \(X\) and \(Y\) is described by the joint probabilities

\[
P(X = i, Y = j) = \pi_{ij}, \quad 1 \leq i \leq I, 1 \leq j \leq J, \quad \text{with} \sum_{ij} \pi_{ij} = 1
\]

These can be organized in a table:

<table>
<thead>
<tr>
<th>(X)</th>
<th>(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  \ldots  (J)</td>
</tr>
<tr>
<td>1</td>
<td>(\pi_{11}) (\pi_{12}) \ldots (\pi_{1J})</td>
</tr>
<tr>
<td>2</td>
<td>(\pi_{21}) (\pi_{22}) \ldots (\pi_{2J})</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>1</td>
<td>(\pi_{I1}) (\pi_{I2}) \ldots (\pi_{IJ})</td>
</tr>
</tbody>
</table>
Example 2
Let $X =$ person uses phone while driving (yes, no), $I = 2$
Let $Y =$ person received speeding ticket or was involved in a car accident, (yes, no), $J = 2.$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>yes</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>yes</td>
</tr>
<tr>
<td>no</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>no</td>
</tr>
</tbody>
</table>

The table gives the joint distribution of $X$ and $Y,$ for example the probability that a person using the phone AND getting in an accident $P(X = \text{yes} AND Y = \text{yes}) = 0.1.$

The *marginal distributions* are the row and the column totals of the joint probabilities and describe the distributions of $X$ and $Y,$ ignoring the other variable. They are denoted by $\pi_{i+}$ and $\pi_{+j},$ the “+” replacing the index totalled.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>Total</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>1</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>$;\ldots$</td>
</tr>
<tr>
<td>I</td>
<td>$\pi_{I1}$</td>
<td>$\pi_{I2}$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>Total</td>
<td>$\pi_{+1}$</td>
<td>$\pi_{+2}$</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>

Example 3
Let $X =$ person uses phone while driving (yes, no), $I = 2$
Let $Y =$ person received speeding ticket or was involved in a car accident, (yes, no), $J = 2.$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.1</td>
<td>0.6</td>
</tr>
<tr>
<td>no</td>
<td>0.01</td>
<td>0.4</td>
</tr>
<tr>
<td>Total</td>
<td>0.11</td>
<td>0.89</td>
</tr>
</tbody>
</table>

With the interpretation $P(X = \text{yes})=0.6,$ $P(Y = \text{yes})=0.11.$

The marginal distributions of $X$ and $Y$ are, respectively:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X = x)$</th>
<th>$Y$</th>
<th>$P(Y = y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>0.6</td>
<td>Yes</td>
<td>0.11</td>
</tr>
<tr>
<td>No</td>
<td>0.4</td>
<td>No</td>
<td>0.89</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

When interpreting one of the variables as a response variable (let’s say $Y$), we are interested in the probabilities for the different outcomes of $Y,$ given a certain outcome of the other variable (this must be then $X$).

$$P(Y = j | X = i) = \frac{\pi_{ij}}{\pi_{i+}}, \quad 1 \leq i \leq I, 1 \leq j \leq J$$

These probabilities make up the *conditional distribution of $Y$ given $X$*.
Example 4

\[ P(Y = \text{yes}|X = \text{yes}) = \frac{0.1}{0.6} = 0.166, \quad P(Y = \text{yes}|X = \text{no}) = \frac{0.01}{0.4} = 0.025 \]

The conditional distribution of \( Y \) given \( X=\text{yes} \) and \( X=\text{no} \) are, respectively:

\[
\begin{array}{l|c}
\text{y} & P(Y = y|X = \text{yes}) \\
\hline
\text{Yes} & 0.166 \\
\text{No} & 0.833 \\
\hline
1 & 1
\end{array}
\]

\[
\begin{array}{l|c}
\text{y} & P(Y = y|X = \text{no}) \\
\hline
\text{Yes} & 0.025 \\
\text{No} & 0.975 \\
\hline
1 & 1
\end{array}
\]

Association between \( X \) and \( Y \)?

The same concepts can be applied to sample data, (replacing \( \pi \) by \( p \)), and the resulting numbers are then interpreted as estimates for the true joint, marginal, and conditional probabilities.

1.1.1 Sensitivity and Specificity

Sensitivity and specificity are certain conditional probabilities when discussing correct classifications using diagnostic tests. Sensitivity = P(Test positive | Diagnosis yes), Specificity = P(Test negative | Diagnosis no)

The higher sensitivity and specificity the better the diagnostic test.

Positive Predictive Value = P(Diagnosis yes | Test positive),
Negative Predictive Value = P(Diagnosis no | Test negative)

Be aware that even when sensitivity and specificity are high the positive and negative predictive values do not have to be high.

Example 5

(From Statistics Notes(http://pubmedcentralcanada.ca/pmcc/articles/PMC2540489/pdf/bmj00444-0038.pdf))

\( Y = \) Pathology (abnormal, normal)
\( X = \) Liver scan (abnormal, normal)

\[
\begin{array}{l|cc|c}
\text{Liver Scan} & \text{Pathology} & & \\
& \text{Abnormal(a)} & \text{Normal(n)} & \text{Total} \\
\hline
\text{Abnormal(a)} & 231 & 32 & 263 \\
\text{Normal(n)} & 27 & 54 & 81 \\
\hline
\text{Total} & 258 & 86 & 344
\end{array}
\]

Estimate for joint probability \( p_{aa} = 231/344 \)
Estimate for marginal probability \( p_{a+} = 263/344 \)
Estimate for marginal probability \( p_{+a} = 258/344 \)
Estimate for conditional probability \( \hat{P}(P = a|LS = a) = 231/263 \)
Estimate for sensitivity \( \hat{P}(LS = a|P = a) = 231/258 \)
Estimate for specificity \( \hat{P}(LS = n|P = n) = 54/86 \)
Estimate for positive predictive value \( \hat{P}(P = a|LS = a) = 231/263 \)
Estimate for negative predictive value \( \hat{P}(P = n|LS = n) = 54/81 \)
Definition 1
Two categorical random variables are statistically independent if the conditional distributions of $Y$ are identical for all levels of $X$.

Example 6
The result of the liver scan would be independent from the pathology, if for abnormal and normal pathology the probability of getting a normal liver scan would be the same.

Remark:
Two variables are statistically independent iff the joint probabilities are the product of the marginal probabilities

$$\pi_{ij} = \pi_{i+} \pi_{+j} \quad \text{for} \quad 1 \leq i \leq I, 1 \leq j \leq J$$

1.1.2 Binomial and Multinomial Sampling

Watch out!!

- If $X$ is a group variable (for example placebo versus treatment) and the sample sizes are fixed for each group (for example 50 for placebo and 50 for treatment), then it does not make sense to look at the joint probabilities.
  
  The conditional distributions of $Y$ are then either binomial or multinomial for each level of $X$, depending on $J$ ($J = 2 \rightarrow$ binomial, $J > 2 \rightarrow$ multinomial).

- The same is true when $X$ is considered an explanatory variable, and $Y$ the response variable. Here again it makes more sense to look at the conditional distributions of $Y$ given $X$. With the same consequence for the distribution as above. (Example above: response = pathology, explanatory = liver scan)

- When both variables can be considered response variables (Example: ask randomly chosen students at MacEwan, which year of study they are in (1st, 2nd, 3rd, 4th), and if they use public transportation (yes/no)), then the joint distribution of $X$ and $Y$ is a multinomial distribution with the cell being the possible outcomes.

1.2 2 × 2 tables

1.2.1 Comparing two Proportions

Use

$$\pi_1 = P(Y = \text{success}|X = 1), \pi_2 = P(Y = \text{success}|X = 2)$$

and $n_1 = n_{1+}, n_2 = n_{2+}$.

Then the Wald-confidence interval for $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Plus four method (Agresti-Cull): For small samples the confidence interval can be improved by adding two imaginary observations to each sample (one success and one failure).
An approximately normally distributed test statistic for comparing $\pi_1$ and $\pi_2$ is

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

with the pooled estimator

$$p_p = \frac{p_1n_1 + p_2n_2}{n_1 + n_2}$$

**Example 7**

A genetic fingerprinting technique called polymerase chain reaction (PCR), can detect as few as 5 cancer cells in every 100,000 cells, which is much more sensitive than other methods for detecting such cells, like using a microscope.

Investigators examined 178 children diagnosed with acute lymphoblastic leukemia who appeared to be in remission using a standard criterion after undergoing chemotherapy (Cave et al, 1998). Using PCR traces of cancer were detected in 75 of the children. During 3 years of follow up 30 of the children suffered a relapse. Of the 103 children who did not show traces of cancer 8 suffered a relapse.

Do these data provide sufficient evidence that children are more likely to suffer a relapse, if the PCR is positive (detect cells)?

**Data:**

<table>
<thead>
<tr>
<th>PCR</th>
<th>Relapse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>positive</td>
<td>30</td>
<td>45</td>
</tr>
<tr>
<td>negative</td>
<td>8</td>
<td>95</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>140</td>
</tr>
</tbody>
</table>

$p_1 = 30/75 = 0.4$, $p_2 = 8/103 = 0.07767$, then the 95% ci for $\pi_1 = \pi_2$

$$(p_1 - p_2) \pm 1.96 \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \to [0.200, 0.445]$$

Since 0 does not fall within the confidence interval at 95% confidence the data do provide sufficient evidence that the proportion of children who suffer a relapse differs for those with positive PCR and a negative PCR. The proportion of children who have a positive PCR is between 0.2 and 0.445 larger than of children with a negative PCR.

**1.2.2 Relative Risk**

The relative risk measures how much more likely a success is in one group than in another. It is a measure of association for the variables defining a $2 \times 2$ table, similar to how Pearson’s correlation coefficient is measuring the association between two numerical variables.

**Definition 2**

$$\text{relative risk} = RR = \frac{\pi_1}{\pi_2}$$
• \( rr = 1 \Leftrightarrow \pi_1 = \pi_2 \) and the variables are independent.

• If \( \pi_1 = 0.5001 \) and \( \pi_2 = 0.5 \) then \( \pi_1 - \pi_2 = 0.001 \) and \( rr = 1.0002 \), but

• if \( \pi_1 = 0.0002 \) and \( \pi_2 = 0.0001 \) then still \( \pi_1 - \pi_2 = 0.0001 \) and now \( rr = 2 \) (the probability in group 1 is 2 times as high as in group 2).

In the second case the difference between the probabilities is more consequential, and the relative risk the better description for the difference.

Confidence interval for the log of the population \( rr \)

\[
\ln(p_1/p_2) \pm z_{\alpha/2} \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}
\]

Using the exponential function for upper and lower bound of this ci, results in a ci for the population \( rr \).

**Example 8**
PCR and relapse:

Estimated \( rr \):

\[
\hat{r}r = \frac{0.4}{0.07767} = 5.15
\]

The relative risk of a relapse for children with positive PCR versus negative PCR is estimated to be 5.15. The probability of a relapse 5.15 times larger for children with positive PCR. The 95% ci for \( \ln(rr) \):

\[
\ln(5.14999) \pm 1.96\sqrt{0.6/30 + 0.92233/8} \rightarrow 1.63900 \pm 0.72092 \rightarrow [0.91807, 2.35992]
\]

Then the 95% for the true relative risk is \( \exp(0.91807), \exp(2.35992) \] = [2.5044, 10.5901].

We are 95% confident that the true \( rr \) falls between 2.50 and 10.59.

Since 1 does not fall within the confidence interval at confidence level of 95% the data do provide sufficient evidence that the \( rr \) for a relapse is different from 1. \( rr = 1 \) would mean that the probabilities for a relapse is the same for children with positive and negative PCR. A relapse is between 2.5 and 10.6 times more likely for children with a positive PCR.

### 1.3 The Odds Ratio

The odds ratio is another measure of association in a two way tables.

Let \( \pi \) be the probability of success

\[
\text{odds} = \frac{\pi}{1 - \pi}
\]

If \( \pi = 0.75 \), then \( \text{odds} = 0.75/0.25 = 3 \), success is three times more likely than failure. For every failure we expect three successes.

If \( \pi = 0.25 \), then \( \text{odds} = 0.25/0.75 = 1/3 \), failure is three times more likely than success.

In reverse, finding the probability when knowing the odds:

\[
\pi = \frac{\text{odds}}{\text{odds} + 1}
\]
Example 9
Assume the odds for a randomly chosen student to get an A in Stat 151 are 1:10, therefore the probability, \( \pi \) of getting an A in Stat 151 is:

\[
\pi = \frac{\text{odds}}{\text{odds} + 1} = \frac{1/10}{1 + 1/10} = \frac{1/10}{(10 + 1)/10} = \frac{1}{10 + 1} = \frac{1}{11}
\]

Definition 3
In a two way table, let \( \text{odds}_1 = \pi_1/(1 - \pi_1) \) the odds of success in row 1, and \( \text{odds}_2 = \pi_2/(1 - \pi_2) \) the odds of success in row 2.
Then

\[
\theta = \frac{\text{odds}_1}{\text{odds}_2}
\]

is the odds ratio.

Properties:

- \( \theta \geq 0 \)
- When \( X \) and \( Y \) are independent, then \( \theta = 1 \) (the probabilities and odds for success are the same in both rows).
- \( \theta > 1 \) means the odds for success are higher in row 1 than in row 2. For example \( \theta = 2 \), means that the odds for success are two times higher in row 1 than in row 2. Individuals from row one are more likely to have a successes than those in row 2.
- \( \theta < 1 \) means the odds are smaller in row 1 than in row 2. Successes are less likely for individuals in row 1 than for those in row 2 (\( \pi_1 < \pi_2 \)).
- \( \theta_1 \) and \( \theta_2 \) indicate the same strength in association if \( \theta_1 = 1/\theta_2 \). (0.25=1/4).
- The odds ratio does not change when the table is transposed, i.e. when the rows become the columns, and the columns the rows. (This is the advantage of the odds ratio over the relative risk.)
- When both variables are response variables then

\[
\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}
\]

and the sample odds ratio \( \hat{\theta} \):

\[
\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}
\]

which is also the ML estimator for \( \theta \) (what does that mean again?).

Same as for the relative risk \( \hat{r} \), the distribution of \( \hat{\theta} \) is very skewed. Therefore instead of finding a confidence interval for \( \theta \), we will find one for \( \ln(\theta) \) and then transform the result to obtain a confidence interval for \( \theta \).
With
\[ SE(\ln(\hat{\theta})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \]

A confidence interval for \( \ln(\theta) \) is
\[ \ln(\hat{\theta}) \pm z_{\alpha/2}SE(\ln(\hat{\theta})) \]

Example 10
Relapse and PCR

\[ \hat{\theta} = \frac{30 \times 95}{8 \times 45} = 7.92 \]

The odds for a relapse are estimated to be almost 8 times higher for children with positive PCR than
for children with a negative PCR.

A 95% confidence interval for \( \ln(\theta) \)
\[ \ln(7.92) \pm 1.96 \sqrt{\frac{1}{30} + \frac{1}{45} + \frac{1}{8} + \frac{1}{95}} \rightarrow 2.0694 \pm 0.8568 \rightarrow [1.213, 2.926] \]

Which gives the following confidence interval for \( \theta \): \([\exp(1.213), \exp(2.926)] = [3.36, 18.65]\). Since
one does not fall within the interval at 95% confidence we do have sufficient evidence that a relapse
is associated with PCR.

Example 11
Hepatitis B Vaccine and Risk of Multiple Sclerosis (Alberto Ascherio et al., N Engl J Med 2001;
344:327-332)

Does the vaccine cause MS?

Data from a large study was used. For each MS patient five matching healthy women were included
for the analysis (relating to age and other factors).

Retrospective study, (was a women vaccinated, after finding if she has or has not MS).

This is a Case-control study!

Because this is a case-control study some limitations to what can be done apply.
Since the proportion of MS/Healthy patients has been chosen by the sampling design, we can not
estimate probabilities based on \( \pi_{MS} \), like \( P(MS|Vaccine) \), but we can estimate \( P(Vaccine|MS) \).
(We can make comparisons between the two samples MS and non-MS, but not vaccine and non-
vaccine)
Therefore we can not use this data for comparing the proportion of MS patients within the group of
vaccinated and non vaccinated individuals. This data can also not be used for estimating the relative
risk of MS in the two groups who were vaccinated or not.
In order to answer the question "Does the vaccine cause MS?" the only measure that is applicable is
the odds ratio, because it is symmetric in \( X \) and \( Y \) (transposing the table has no effect on the odds
ratio).

Data:
### Table

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>MS</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>32</td>
<td>116</td>
</tr>
<tr>
<td>no</td>
<td>158</td>
<td>608</td>
</tr>
<tr>
<td>Total</td>
<td>190</td>
<td>724</td>
</tr>
</tbody>
</table>

Two approaches can be taken to find the estimate for the odds ratio:

1. \( p_1 = \frac{32}{116} = 0.27586 \), \( p_2 = \frac{158}{608} = 0.25987 \) (these numbers have no valuable interpretation as explained above and are only used for calculation purposes)

   (a) \( o_1 \): odds for vaccinated women: \( o_1 = \frac{p_1}{1 - p_1} = \frac{32}{84} = 0.38095 \)

   (b) \( o_2 \): odds for non-vaccinated women: \( o_2 = \frac{p_2}{1 - p_2} = \frac{158}{450} = 0.35111 \)

   (c) \( \hat{\theta} \): odds ratio \( \hat{\theta} = \frac{o_1}{o_2} = 1.0850 \)

2. \[ \hat{\theta} = \frac{32 \times 450}{158 \times 84} = 1.0850 \]

The odds of getting MS is a little higher for women who were vaccinated than for those who were not.

A 95% CI for \( \ln(\theta) \)

\[ \ln(1.0850) \pm 1.96 \sqrt{\frac{1}{32} + \frac{1}{84} + \frac{1}{158} + \frac{1}{450}} \rightarrow 0.081580 \pm 0.10134 \rightarrow [-0.01976, 0.18292] \]

Which gives following CI for \( \theta \): \([\exp(-0.01976), \exp(0.18292)] = [0.98, 1.2] \). Since 1 falls within the interval at 95% confidence we do not have sufficient evidence that MS is associated with the vaccine.

**Remark:**

- An **odds ratio** of \( \theta = 2 \), means that the odds for group 1 are double the odds for group 2.

- A relative risk of \( rr = 2 \), means that the probability for success in group1 is double the success probability in group 2.

\[ \theta = rr \times \frac{1 - \pi_2}{1 - \pi_1} \]

When both probabilities are close to zero the last fraction is close to one, and \( \theta \approx rr \).

**Remark:**

Odds ratio as well as relative risk are measures of association in a 2×2 table and give how much higher the risk or odds are for one group versus another one. Usually the relative risk is easier to interpret because it compares probabilities instead of odds. Beside the difference in interpretation there is a difference when they are applicable. The odds ratio can be used in all types of studies,
but the relative risk can only be used, if the rows can be interpreted as random samples from a populations with and without the risk factor. This means that in particular for case-control studies the relative risk is not applicable. A case-control design involves the selection of research subjects on the basis of the outcome measurement rather than on the basis of the exposure (see above, we chose MS patients and healthy subjects, instead of people who were vaccinated and not vaccinated).

1.3.1 Types of studies

- Case-control, for each case find matching individuals for the control group at a given ratio. The cases are defined to have the outcome to be studied (disease = yes)
- Cohort studies (individuals choose themselves which groups they are in, ex. person smokes (yes/no))
- Clinical studies (individuals are randomly assigned to treatment groups by experimenter)
- Cross-sectional studies (random samples are taken and individuals are assessed which group they fall within and their response, ex.: relapse, PCR study)
- Observational studies: cohort, cross sectional, and case control (more practical)
- Experimental studies: clinical (less pitfalls)

1.4 $\chi^2$ tests for Independence

Imagine two categorical variables $X$ with $I$ levels and $Y$ with $J$ levels. Then the joint distribution of $X$ and $Y$ is given by.

$$P(X = i, Y = j) = \pi_{ij}, \quad 1 \leq i \leq I, 1 \leq j \leq J$$

For samples of size $n$ from this joint distribution the expected cell frequencies are then

$$\mu_{ij} = n\pi_{ij}, \quad 1 \leq i \leq I, 1 \leq j \leq J$$

(similar to the mean of a binomial distribution).

If we want to test the null hypothesis that the joint distribution of $X$ and $Y$ equals $\{\pi_{ij}\}$, then this can be done comparing the expected cell frequencies $\{\mu_{ij}\}$ for the claimed distribution with the observed cell counts $\{n_{ij}\}$. The larger the differences $\{n_{ij} - \mu_{ij}\}$ the more likely the null hypothesis is wrong.

1.4.1 Pearson’s statistic

(compare with Stat 151/141)

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$

(studied by Pearson in 1900) Properties:
• $X^2 \geq 0$
• $X^2 = 0$ when $n_{ij} = \mu_{ij}$
• For large $n$: $X^2$ is approximately $\chi^2$ distributed (diagram)
  The approximation is close enough when $\mu_{ij} \geq 5$ for all cells.
• $\mu_{X^2} = df$, $\sigma_{X^2} = \sqrt{2df}$
• $X^2$ is a score statistic, it is based on the distribution on the standard errors and covariances of
  the $n_{ij}$ when $H_0$ is true.
• Large values of $X^2$ give evidence against $H_0$.

1.4.2 Likelihood Ratio Statistic
Reminder:
• The likelihood function $l$ gives the probability of the data to occur depending on the parameter(s).
• Find the largest value of $l$ if $H_0$ is true $\rightarrow$ call this $l_0$
• Find the largest value of $l$ if $H_0$ is true or not $\rightarrow$ call this $l_1$
• $G^2 = -2\ln(l_0/l_1)$ is the test statistic.

We get the likelihood-ratio $\chi^2$ statistic

$$G^2 = 2 \sum_{ij} n_{ij} \ln \frac{n_{ij}}{\mu_{ij}}$$

Properties:
• $G^2 \geq 0$
• $G^2 = 0$ when $n_{ij} = \mu_{ij}$
• For large $n$: $G^2$ is $\chi^2$ distributed
  The approximation is close enough when $\mu_{ij} \geq 5$ for all cells.
• $\mu_{G^2} = df$, $\sigma_{G^2} = \sqrt{2df}$
• Large values of $G^2$ give evidence against $H_0$. 

11
1.4.3 Tests for independence

Both test statistics can be used for a test for independence. We only have to answer the question, how to find $\mu_{ij}$ when $X$ and $Y$ are independent.

When $X$ and $Y$ are independent, then

$$\mu_{ij} = n\pi_{ij} = n\pi_i \pi_j$$

which can be estimated by

$$\hat{\mu}_{ij} = np_i p_j = \frac{n_{i+} n_{+j}}{n} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

The degree of freedom of the test statistic is the difference between the number of parameters under $H_a$ and $H_0$.

- When $H_a$ is true $\{\pi_{ij}\}$ are $I \times J$ parameters, but one of them can not be chosen freely, since the total has to be one, therefore there are $IJ - 1$ parameters under $H_a$

- When $H_0$ holds the parameters are the marginal distributions $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ with $I - 1$ and $J - 1$ parameters (subtract 1 because the total has to be 1), respectively, for a total of $(I - 1) + (J - 1) = I + J - 2$

- Then $df = (IJ - 1) - (I + J - 2) = (I - 1)(J - 1)$ (some algebra).

This now leads to

Tests for Independence

1. Hyp.: $H_0 : X$ and $Y$ are independent versus $H_a : H_0$ is not true, choose $\alpha$

2. Assumptions: random samples, $\{\hat{\mu}_{ij} \geq 5\}$.

3. Test Statistic: $X^2_0$ or $G^2_0$, $df = (I - 1)(J - 1)$

4. P-value: P-value=$P(\chi^2 > X^2)$ or P-value=$P(\chi^2 > G^2)$

Example 12

A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a small community. It consists of a two-shot sequence in two weeks. A survey of 1000 residents the following spring provided the following information:

<table>
<thead>
<tr>
<th>No vaccine</th>
<th>One Shot</th>
<th>Two Shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>No Flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>109</td>
<td>578</td>
</tr>
</tbody>
</table>

1. Hyp.: $H_0 :$ Getting the flu and the vaccine are independent versus $H_a : H_0$ is not true, $\alpha = 0.05$

2. Assumptions: random samples, $\{\hat{\mu}_{ij} \geq 5\}$ (check later).
3. Test Statistic: $X_0^2$ or $G_0^2$, $df = (I - 1)(J - 1)$  
First find the expected cell counts:

<table>
<thead>
<tr>
<th></th>
<th>No vaccine</th>
<th>One Shot</th>
<th>Two Shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>14.40</td>
<td>5.01</td>
<td>26.59</td>
<td></td>
</tr>
<tr>
<td>No Flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
<td>954</td>
</tr>
<tr>
<td></td>
<td>298.60</td>
<td>103.99</td>
<td>551.41</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>

Then $df = (3 - 1)(2 - 1) = 2$ and

$$X_0^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = 6.404 + 3.169 + 6.944 + 0.309 + 0.153 + 0.335 = 17.31$$

and

$$G_0^2 = 2 \sum_{i,j} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}} = 2 \left( 24 \ln \frac{24}{14.40} + \cdots + 565 \ln \frac{565}{551.41} \right) = 17.26$$

4. P-value: $P(\chi^2 > X_0^2) < 0.005$ or $P(\chi^2 > G_0^2) < 0.005$ (table VII).

5. Reject $H_0$, since the P-value is smaller than $\alpha = 0.05$.

6. At significance level of 5% the data provide sufficient evidence that there is an association between the vaccination status and getting the flu.

To study the nature of the association we could, for each cell, compare the observed and the expected cell count. But since for larger estimated expected cell counts the standard error increases we need to standardize:

$$Z_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_i+)(1 - p_{+j})}}$$

is the standardized cell residual (called adjusted residual by SPSS) for the cell in row $i$ and column $j$.

If $H_0$ would be true these would be approximately standard normally distributed. Therefore standardized cell residuals which do not fall between $-1.96$ and $1.96$ indicate that the particular cell gives evidence that $H_0$ is violated.

Should we do a multiple comparison?

**Example 13**

The standardized cell residuals for the flu example:

<table>
<thead>
<tr>
<th></th>
<th>No vaccine</th>
<th>One Shot</th>
<th>Two Shots</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flu</td>
<td>24</td>
<td>9</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>3.1</td>
<td>1.9</td>
<td>-4.2</td>
<td></td>
</tr>
<tr>
<td>No Flu</td>
<td>289</td>
<td>100</td>
<td>565</td>
<td>954</td>
</tr>
<tr>
<td></td>
<td>-3.1</td>
<td>-1.9</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>109</td>
<td>578</td>
<td>1000</td>
</tr>
</tbody>
</table>
The significant positive residual for people who did not get vaccinated and did get the flu, indicates that these cases occurred more often than we would predict under independence. The significant negative residual for people who were vaccinated twice and did not get the flu, indicates that these cases occurred less often than we would predict under independence. Together this means that non-vaccinated people did get the flu more often than expected and vaccinated did get the flu less often, therefore we can conclude that the vaccine helps preventing people from getting the flu.

Since we have to do 3 tests (one for each column, since we only have two rows and the standardized cell cell residuals add up to zero for each row and column), it might be more appropriate to do a multiple comparison, using a significance level of \( \alpha^* = \alpha / 3 \) for each test.

The sample odds ratio for no vaccine and 2 shots is

\[
\hat{\theta} = \frac{24 \times 565}{289 \times 13} = 3.6
\]

The estimated odds for getting the flu are 3.6 times higher for people who did not get vaccinated compared to those who got 2 shots. Or we can say, that the estimated odds for getting the flu are 260\% higher for people who did not get vaccinated compared to those who got 2 shots. Yep, the vaccine works.

### 1.5 Testing Independence of Ordinal Variables

The trick for studying the association between two categorical variables is to assign numerical scores to the categories which reflect the distance between the categories and then use tools which we would usually use for numerical variables, like Pearson’s correlation coefficient.

#### 1.5.1 Linear relationship between categorical variables

We are going to introduce a test to test if two ordinal variables are linearly related (if there is a linear trend association).

Assume scores have been assigned (how to do this we will look at later). The test statistic based on the covariance (correlation) is then sensitive to a linear trend.

Let \( u_1 \leq u_2 \leq \cdots \leq u_I \) denote the scores for the row categories, and \( v_1 \leq v_2 \leq \cdots \leq v_J \) denote scores for the column categories (the ordering of the scores should reflect the ordering of the categories) with greater distances between categories which are considered farther apart.

The mean score for the rows and columns are then, respectively:

\[
\bar{u} = \frac{\sum_i n_i + u_i}{n} = \frac{1}{n} \sum_i p_i + u_i, \quad \bar{v} = \frac{\sum_j n_j + v_j}{n} = \frac{1}{n} \sum_j p_j + v_j
\]

The sample covariance for \( X \) and \( Y \) is

\[
\hat{\text{Cov}}(X, Y) = \sum_{ij} p_{ij} (u_i - \bar{u})(v_j - \bar{v})
\]

the sample standard deviations for \( X \) and \( Y \), respectively:

\[
s_X = \sqrt{\sum_i p_i (u_i - \bar{u})^2}, \quad s_Y = \sqrt{\sum_j p_j (v_j - \bar{v})^2}
\]
and the sample correlation

\[ r = \frac{\text{Cov}(X,Y)}{s_x s_y} \]

Properties:

- \(-1 \leq r \leq 1\)
- \(r < 1\) indicates a negative relationship (the larger \(X\) the smaller \(Y\)).
- \(r > 1\) indicates a positive relationship (the larger \(X\) the larger \(Y\)).
- \(r \approx 0\) no relationship.

This is an estimate for the population correlation \(\rho\).

If \(\rho = 0\)

\[ M^2 = (n - 1)r^2, df = 1 \]

is approximately \(\chi^2\)-distributed with \(df = 1\). Based on this we can develop a test for linear trend for two ordinal variables.

**A test about the correlation (for linear trend):**

1. Hyp.: \(H_0 : \rho = 0\) versus \(H_a : \rho \neq 0\), choose \(\alpha\)
2. random samples
3. Test Statistic:

\[ M^2_0 = (n - 1)r^2, df = 1 \]

4. P-value: \(P = P(\chi^2 > M^2_0)\)

**Example 14**

**Effect of smoking on the development of tartar.** (Toutenberg: Statistical Analysis of designed experiments (1992))

<table>
<thead>
<tr>
<th>Smoking</th>
<th>Tartar</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>none</td>
<td>middle</td>
<td>heavy</td>
<td>Total</td>
</tr>
<tr>
<td>no</td>
<td>284</td>
<td>236</td>
<td>48</td>
<td>568</td>
</tr>
<tr>
<td>middle</td>
<td>606</td>
<td>983</td>
<td>209</td>
<td>1798</td>
</tr>
<tr>
<td>heavy</td>
<td>1028</td>
<td>1871</td>
<td>425</td>
<td>3324</td>
</tr>
<tr>
<td>Total</td>
<td>1918</td>
<td>3090</td>
<td>682</td>
<td>5690</td>
</tr>
</tbody>
</table>

Use coding 1,2,3 for tartar and smoking, then \(r = 0.095\), which seems to be quite low, this is due to the small number of possible values. Use coding 1,2,3 for tartar, and for smoking 0, 2, 3,( indicate a larger difference between no smoking and some, than between some and heavy) then \(r = 0.103\), which is a little larger.

To test if there is a linear trend between these two variables

1. Hyp.: \(H_0 : \rho = 0\) versus \(H_a : \rho \neq 0\), choose \(\alpha = 0.05\)
2. random samples

3. Test Statistic:

\[ M_0^2 = (n - 1)r^2 = (56890.103^2 = 60.35, df = 1 \]

4. P-value: \[ P = P(\chi^2 > M_0^2) \approx 0 \]

5. Decision: Reject \( H_0 \)

6. Context: At significance level of 5% the data provide sufficient evidence that the more a person smokes the more extreme the tartar build up.

Advantages of test for linear trend over test for independence:

- gain in power (power is the probability of a correct decision, when not rejecting \( H_0 \)). i.e. if the test is not significant then it is more likely that there truly is no (positive or negative) association between the ordinal variables.

- smaller sample size required to detect association (test is more sensitive).

- smaller samples are required for the approximation of the distribution to be close.

1.5.2 Assigning Scores to Ordinal Variables

Agresti "For most data sets, the choice of scores has little effect on the results. Different choices of ordered scores usually give similar results."

The biggest problems occur when the data is unbalanced, i.e. if some categories have many more observations than other. (For example when looking at the evaluation of an excellent instructor, then categories "excellent and "very good" get high frequencies, but "poor" gets none or very little). Usually it is reasonable to use evenly spaced scores for the categories, like 1, 2, 3, \ldots or 10, 20, \ldots for \( X \). Another solution is to use midranks.

For example for the data in the previous table this would give scores

<table>
<thead>
<tr>
<th>categories</th>
<th>frequencies</th>
<th>midrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>568</td>
<td>((\frac{568+1}{2})=284.5)</td>
</tr>
<tr>
<td>middle</td>
<td>1798</td>
<td>((\frac{(568+1) + (568+1798)}/2=1467.5)</td>
</tr>
<tr>
<td>heavy</td>
<td>3324</td>
<td>((\frac{(568+1798+1)+(568+1798+3324)}/2=4028.5)</td>
</tr>
</tbody>
</table>

In the case when using midranks then \( r \) is called Spearman’s \( \rho \) (or should it be called \( r \)) or Spearman’s correlation.

For the example Spearman’s correlation=0.086 \((M^2=42.08, df = 1, P < 0.001)\). This is consistent with the results using Pearson’s correlation, but the better choice because this statistic takes into account that the data is ordinal and not numerical, and the scores are not dropping out of the sky, but are based on the data.
1.5.3 Kendall’s $\tau$

Kendall’s $\tau - b$ (tau) is an alternative measure for the relationship between two ordinal measures with the same number of outcomes (the contingency table is square). It is based on the number of concordant and discordant observations in the sample.

A pair of observations, $(x_i, y_i), (x_j, y_j)$, is called concordant if both $x_i < x_j$ and $y_i < y_j$, or both $x_i > x_j$ and $y_i > y_j$.

A pair of observations, $(x_i, y_i), (x_j, y_j)$, is called discordant if both $x_i < x_j$ but $y_i > y_j$, or both $x_i > x_j$ but $y_i < y_j$.

Let $n_c$ be the number of concordant pairs, and $n_d$ the number of discordant pairs.

Then Kendall’s $\tau$ is

$$\hat{\tau} = 2 \frac{n_c - n_d}{n(n - 1)}$$

**Example 15**

<table>
<thead>
<tr>
<th>Smoking</th>
<th>Tartar</th>
<th>none</th>
<th>middle</th>
<th>heavy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>284</td>
<td>236</td>
<td>48</td>
<td>568</td>
<td></td>
</tr>
<tr>
<td>middle</td>
<td>606</td>
<td>983</td>
<td>209</td>
<td>1798</td>
<td></td>
</tr>
<tr>
<td>heavy</td>
<td>1028</td>
<td>1871</td>
<td>425</td>
<td>3324</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1918</td>
<td>3090</td>
<td>682</td>
<td>5690</td>
<td></td>
</tr>
</tbody>
</table>

Then

$$n_c = 284(983 + 209 + 1871 + 425) + 236(209 + 425) + 606(1871 + 425) + 983(425) = 2,949,367$$

and

$$n_d = 48(606 + 983 + 1028 + 1871) + 236(606 + 1028) + 209(1028 + 1871) + 983(1028) = 2,217,463$$

therefore

$$\hat{\tau} = 0.045$$

The problem with $\tau$ is that it ignores the presence of ties. The measure Kendall’s $\tau - b$ is a modification of Kendall’s $\tau$ accounting for ties applicable in square tables only, and Kendall’s $\tau - c$ generalizes Kendall’s $\tau - b$ for non square tables. The interpretation remains the same.
For this data Kendall’s $\tau - b$ is 0.080 ($t=6.436$, $df=1$, $p<0.001$).

<table>
<thead>
<tr>
<th>Symmetric Measures</th>
<th>Value</th>
<th>Asymp. Std. Error$^a$</th>
<th>Approx. $T^b$</th>
<th>Approx. Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinal by Kendall's tau-b</td>
<td>0.080</td>
<td>0.012</td>
<td>6.437</td>
<td>0.000</td>
</tr>
<tr>
<td>Ordinal Spearman Correlation</td>
<td>0.086</td>
<td>0.013</td>
<td>6.545</td>
<td>0.000$^c$</td>
</tr>
<tr>
<td>Interval by Interval Pearson’s R</td>
<td>0.095</td>
<td>0.013</td>
<td>7.188</td>
<td>0.000$^c$</td>
</tr>
<tr>
<td>N of Valid Cases</td>
<td>5690</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

1.5.4 $2 \times J$ tables

In this case the trend test, tests if the median scores differ for the two groups (given by $X$), when using the midrank scores this test is equivalent to the Wilcoxon Rank Sum Test (also called Mann-Whitney Test) In Stat 252 we said this test is only applicable if there are not too many ties. One has to be careful with the interpretation.

1.5.5 $I \times 2$ tables

In this case the response has only two outcomes (success/failures), and the question is if the ordinal variable has a positive or negative effect on the success rate of the response. In this case the test is called the Cochran-Armitage trend test.

Warning: Do not use this test if one of the variables is nominal and has more than 2 categories. In this case the test for trend is inappropriate.

1.6 Exact Tests (in Lab?)

1.7 Three Way Tables

For three categorical variables $X$, $Y$, and $Z$, where

$Y$ - response variable
$X$ - explanatory variable
$Z$ - confounding variable (to be adjusted for)

Example 16

$Y$ - response variable (cancer(yes, no))
$X$ - explanatory variable (second hand smoking (none, some, lots))
$Z$ - confounding variable (age(age groups))

One way of controlling for $Z$ is by looking at tables for $X$ and $Y$ for each level of $Z$, called partial tables.
Example 17
Data from IMDb (January 11, 2011)

\[ Y = \text{rating for movies} \quad X = \text{type of movie, horror(Alien) or animated movie (Wall-e), controlling for} \]

\[ Z = \text{sex (age group).} \]

Male

<table>
<thead>
<tr>
<th>Movie</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alien</td>
<td>1502</td>
<td>1404</td>
<td>7090</td>
<td>49158</td>
<td>77097</td>
<td>136251</td>
</tr>
<tr>
<td>Wall-e</td>
<td>5654</td>
<td>2261</td>
<td>8199</td>
<td>43116</td>
<td>94079</td>
<td>153309</td>
</tr>
<tr>
<td>Total</td>
<td>7156</td>
<td>3665</td>
<td>15289</td>
<td>92274</td>
<td>171176</td>
<td>289560</td>
</tr>
</tbody>
</table>

Female

<table>
<thead>
<tr>
<th>Movie</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
<th>7-8</th>
<th>9-10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alien</td>
<td>430</td>
<td>356</td>
<td>1313</td>
<td>5075</td>
<td>6777</td>
<td>13951</td>
</tr>
<tr>
<td>Wall-e</td>
<td>1104</td>
<td>441</td>
<td>1274</td>
<td>5998</td>
<td>19106</td>
<td>27923</td>
</tr>
<tr>
<td>Total</td>
<td>1534</td>
<td>797</td>
<td>2587</td>
<td>11073</td>
<td>25883</td>
<td>41874</td>
</tr>
</tbody>
</table>

The partial tables describe the conditional joint distribution of \( X \) and \( Y \) (conditional on the different levels of \( Z \)).

Given that the rater was female the probability that a rater chose to rate Alien and rated it 9 or 10 equals \( \frac{6777}{41874} \).
Ignoring $Z$ entirely results in the marginal table for $X$ and $Y$. To obtain the marginal table just add the equivalent table entries of the partial tables.

**Example 18**

**Rows:** movie  **Columns:** rating

<table>
<thead>
<tr>
<th>sex</th>
<th>movie</th>
<th>Alien</th>
<th>Rating</th>
<th>Count</th>
<th>12</th>
<th>34</th>
<th>56</th>
<th>78</th>
<th>910</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>movie</td>
<td>Alien</td>
<td>Count</td>
<td></td>
<td>430</td>
<td>356</td>
<td>1313</td>
<td>5075</td>
<td>6777</td>
<td>13951</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td>3.1%</td>
<td>2.6%</td>
<td>9.4%</td>
<td>36.4%</td>
<td>48.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td>28.0%</td>
<td>44.7%</td>
<td>50.8%</td>
<td>45.8%</td>
<td>26.2%</td>
<td>33.3%</td>
</tr>
<tr>
<td></td>
<td>Wall-e</td>
<td>Count</td>
<td></td>
<td></td>
<td>1104</td>
<td>441</td>
<td>1274</td>
<td>5998</td>
<td>19106</td>
<td>27923</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td></td>
<td>4.0%</td>
<td>1.6%</td>
<td>4.6%</td>
<td>21.5%</td>
<td>68.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td></td>
<td>72.0%</td>
<td>55.3%</td>
<td>49.2%</td>
<td>54.2%</td>
<td>73.8%</td>
<td>66.7%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td></td>
<td></td>
<td>1534</td>
<td>797</td>
<td>2587</td>
<td>11073</td>
<td>25883</td>
<td>41874</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td></td>
<td>3.7%</td>
<td>1.9%</td>
<td>6.2%</td>
<td>26.4%</td>
<td>61.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td></td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>m</td>
<td>movie</td>
<td>Alien</td>
<td>Count</td>
<td></td>
<td>1502</td>
<td>1404</td>
<td>7090</td>
<td>49158</td>
<td>77097</td>
<td>136251</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td></td>
<td>1.1%</td>
<td>1.0%</td>
<td>5.2%</td>
<td>36.1%</td>
<td>56.6%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td></td>
<td>21.0%</td>
<td>38.3%</td>
<td>46.4%</td>
<td>53.3%</td>
<td>45.0%</td>
<td>47.1%</td>
</tr>
<tr>
<td></td>
<td>Wall-e</td>
<td>Count</td>
<td></td>
<td></td>
<td>5654</td>
<td>2261</td>
<td>8199</td>
<td>43116</td>
<td>94079</td>
<td>153309</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td></td>
<td>3.7%</td>
<td>1.5%</td>
<td>5.3%</td>
<td>28.1%</td>
<td>61.4%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td></td>
<td>79.0%</td>
<td>61.7%</td>
<td>53.6%</td>
<td>46.7%</td>
<td>55.0%</td>
<td>52.9%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>Count</td>
<td></td>
<td></td>
<td>7156</td>
<td>3665</td>
<td>15289</td>
<td>92274</td>
<td>171176</td>
<td>289560</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within movie</td>
<td></td>
<td></td>
<td>2.5%</td>
<td>1.3%</td>
<td>5.3%</td>
<td>31.9%</td>
<td>59.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>% within rating</td>
<td></td>
<td></td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

We can now use these tables to test for conditional independence and marginal independence.

**Example 19**

From SPSS:
At significance level of 5% the data provide sufficient evidence for conditional association for men and women, and marginal association of movies and rating.

We can also do conditional and marginal trend tests (I used scores 1,2,3,4,5 for the rating): Use SPSS (1 = Alien, 2 = Wall-e) to find the correlations, and $M_0^2 = (r^2(n - 1))$:

<table>
<thead>
<tr>
<th>Test</th>
<th>$\chi^2$</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>conditional independence (female)</td>
<td>1793.102</td>
<td>4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>conditional independence (male)</td>
<td>3778.475</td>
<td>4</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>marginal independence</td>
<td>4692.375</td>
<td>4</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>$r$</th>
<th>$M_0^2$</th>
<th>df</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>trend (conditional female)</td>
<td>0.188</td>
<td>1480.0</td>
<td>1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>trend (conditional male)</td>
<td>-0.024</td>
<td>166.79</td>
<td>1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>marginal trend</td>
<td>-0.006</td>
<td>11.93</td>
<td>1</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

At significance level of 5% the data provide sufficient evidence that the median ratings for Alien and Wall-e conditional for women and men and combined are not the same.

The sign of $r$ indicates that the median rating by women for Wall-e is higher than for Alien, and Alien is higher rated by men.

The marginal result indicates that the median rating for Alien is higher overall than for Wall-e. (We are missing an important fact here. This result is due to the fact that men are much more likely to give a rating on IMDb). These findings indicate that we have an interaction effect between sex and movie on rating, an appropriate model should give us more insight.

In the case that the marginal and conditional distributions are 2 $\times$ 2 tables we also can look at conditional and marginal odds ratios etc. (compare text book).

### 1.7.1 Simpson’s Paradox

In a paper by Charig et al. (British Medical Journal, Clinical Research Ed., March 1986, 292(6524): 879-882) two different treatments for kidney stones are compared. Call the treatments A and B. It is reported that treatment A is successful for 78% of patients and treatment B is successful for 83% of patients. So one would assume that treatment B is "better" than treatment A. BUT they also report that for small kidney stones treatment A is successful for 93% of all cases, but B only in 87% of all cases and that for large kidney stones treatment A is successful for 73% of all cases, but B only in 69% of all cases.

So treatment A is "better" than treatment B?

<table>
<thead>
<tr>
<th>Treatment A</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>No Success</td>
</tr>
<tr>
<td>small</td>
<td>81</td>
</tr>
<tr>
<td>large</td>
<td>192</td>
</tr>
<tr>
<td>Total</td>
<td>273</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Treatment B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>No Success</td>
</tr>
<tr>
<td>small</td>
<td>234</td>
</tr>
<tr>
<td>large</td>
<td>55</td>
</tr>
<tr>
<td>Total</td>
<td>289</td>
</tr>
</tbody>
</table>
Treatment A is better than treatment B, because looking at the conditional distributions for treatment and success, given the size, treatment A shows a higher success rate for small as well as large kidney stones, i.e. 93% instead of 87%, and 73% instead of 69%.

How come, that overall treatment B seems to be the better treatment?
Both treatments are much more successful for small kidney stones than for large ones. Find that the sample for treatment B mostly consists of small kidney stones, but the sample for treatment A mostly consists of patients treated for large kidney stones, therefore showing a favourable overall success rate for treatment B.
As a result the overall success rate for treatment B is closer to the success rate for small kidney stone (looks good), but the overall success rate for treatment A is closer to the success rate for large kidney stones (looks not as good).

1.7.2 Homogeneous association
We speak of an homogeneous association between \( X \) and \( Y \) controlling for \( Z \), when the association between \( X \) and \( Y \) is the same for all levels of \( Z \). In ANOVA and multiple regression we said that \( X \) and \( Z \) do not interact in their effect on \( Y \).
For categorical data means that the odds ratios (for each combination of 2 levels from \( X \) and two levels of \( Y \)) are not affected by \( Z \).
How to test for homogeneous association we see later after building models.

1.8 Cochran-Mantel-Haenzel Test
Assume three variables, \( X, Y, Z \), all being binary.
Let \( \vartheta_{XY(k)} \) denote the conditional odds ratio between \( X \) and \( Y \), given \( Z = k \).
Then a homogeneous relationship between \( X, Y \) and \( Z \) implies that
\[
\vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)}
\]
We speak of conditional independence between \( X \) and \( Y \), given \( Z \), iff
\[
\vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)} = 1
\]

Cochran-Mantel-Haenzel Test for conditional independence
1. \( H_0 : \vartheta_{XY(1)} = \vartheta_{XY(2)} = \cdots = \vartheta_{XY(K)} = 1 \)
   \( H_a : \) at least one conditional OR is different from 1
2. large random sample
3. \[
\chi^2_0 = \frac{\left( \sum_k n_{11k} - \hat{\mu}_{11k} \right)^2}{\sum_k \text{Var}(n_{11k})}, \quad df = 1
\]
   with
\[
\text{Var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{1+k}n_{2+k}}{n_{1+k}^2(n_{++k} - 1)}
\]
4. P-value= \( P(\chi^2 > \chi^2_0) \)