Inference for Contingency Tables
Large-samples

1. CI for Odds Ratio
   1a. Example 3.1.2

2. CI for Difference of Proportions

3. CI for Relative Risk

4. Delta Method: Applications
   4a. Logit model
   4b. Delta method for Log Odds Ratio

5. Testing Independence in Two-way Contingency Tables
   5a. Pearson Chi-squared Test
   5b. LR Chi-squared Test

6. Pearson and Standardized Residuals

Homework: Ch.3, Sec. 1-2
Problems: 3.1, 3.2, 3.5
1. CI for odds ratio

Consider 2x2 Table

<table>
<thead>
<tr>
<th></th>
<th>n_{11}</th>
<th>n_{12}</th>
<th>n_{1+}</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_{11}</td>
<td>n_{21}</td>
<td>n_{12}</td>
<td>n_{2+}</td>
</tr>
</tbody>
</table>

The sample Odds Ratio: \( \hat{\theta} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} \)

Remark 1. If any \( n_{ij} = 0 \), then \( \hat{\theta} = 0 \) or \( \infty \)

Another estimator

\[
\hat{\theta} = \frac{(n_{11} + .5) (n_{22} + .5)}{(n_{12} + .5) (n_{21} + .5)}
\]

is more appropriate.

\( \hat{\theta} \) and \( \hat{\theta} \) have the same asymptotic normality as \( n \to \infty \),

Transformation: \( \ln \hat{\theta} \)

\( \hat{\theta} \) is asymptotically closer to normal estimated sd of \( \hat{\theta} \):

\[
\hat{\sigma} (\ln \hat{\theta}) = \left( \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2}
\]
CI for $\ln \theta$:

$$\ln \hat{\theta} \pm \frac{z}{2} \hat{\sigma}(\ln \hat{\theta})$$

Since

$$\frac{\ln \theta - \ln \hat{\theta}}{\hat{\sigma}(\ln \hat{\theta})} \approx N(0,1)$$

Taking antilog of the endpoints of CI $\ln \theta$ provides a CI for $\theta$.

La. Example 3.1.2, Table 3.1: Myocardial Infarction

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>28</td>
<td>656</td>
<td>684</td>
</tr>
<tr>
<td>Aspirin</td>
<td>18</td>
<td>676</td>
<td>676</td>
</tr>
</tbody>
</table>

Construct CI for $\theta$:

The sample Odds Ratio $\hat{\theta} = \frac{28 \times 676}{18 \times 656} = 1.56$

$$\ln \hat{\theta} = \ln 1.56 = 0.445$$

$$\hat{\sigma}(\ln \hat{\theta}) = \left( \frac{1}{28} + \frac{1}{656} + \frac{1}{18} + \frac{1}{676} \right)^{1/2} = 0.307$$

95% CI for $\ln \theta$:

$$0.445 \pm 1.96 \times 0.307 \text{ or } (-0.157, 1.047)$$

95% CI for $\theta$:

$$e^{(-0.157, 1.047)} \text{ or } (0.85, 2.85)$$

$H_0: \theta = 1$. Since $1 \in (0.85, 2.85)$ it is plausible.
2. C.I. for \( \pi_1 - \pi_2 \)

Here \( \pi_1 = \pi_{11} \) and \( \pi_2 = \pi_{12} \)

\[
\begin{array}{ccc}
  n_{11} & n_{12} & n_1 \; \text{fixed} \\
  n_{21} & n_{22} & n_2 \; \text{fixed} \\
\end{array}
\]

Totals

Consider \( Y_1 = n_{11} \sim \text{Bin}(n_{11}, \pi_1) \) and \( Y_2 = n_{21} \sim \text{Bin}(n_{21}, \pi_2) \) are independent

\[
\begin{align*}
\hat{\pi}_1 &= \frac{Y_1}{n_1} \\
\hat{\pi}_2 &= \frac{Y_2}{n_2} \\
\end{align*}
\]

\[
E(\hat{\pi}_i) = \pi_i \\
\text{var}(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n_i}, \; i = 1, 2
\]

Indeed: \( Y_i \sim \text{Bin}(n_{ii}, \pi_i) \)

\[
\begin{align*}
E(Y_i) &= n_{ii}\pi_i \\
\text{var}(Y_i) &= n_{ii}\pi_i(1-\pi_i) \\
\text{var}(\hat{\pi}_i) &= \text{var}\left(\frac{Y_i}{n_{ii}}\right) = \frac{1}{n_{ii}^2} \text{var}(Y_i) = \frac{\pi_i(1-\pi_i)}{n_{ii}} \\
\end{align*}
\]

Now, since \( Y_1 \) and \( Y_2 \) are independent:

\[
\text{var}(\hat{\pi}_1 - \hat{\pi}_2) = \text{var}(\hat{\pi}_1) + \text{var}(\hat{\pi}_2)
\]
So that
\[ \hat{\sigma}^2 (\hat{\pi}_1 - \hat{\pi}_2) = \left( \frac{\hat{\pi}_1 (1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2 (1 - \hat{\pi}_2)}{n_2} \right)^{1/2}. \]

Also, we know that
\[ E[\hat{\pi}_1 - \hat{\pi}_2] = \pi_1 - \pi_2 \quad \text{and} \quad \sqrt{\text{Var}(\hat{\pi}_1 - \hat{\pi}_2)} \sim N(0, 1). \]

100\% (1-\alpha) CI for \( \pi_1 - \pi_2 \):
\[ \hat{\pi}_1 - \hat{\pi}_2 \pm Z_{\alpha/2} \cdot \hat{\sigma} (\hat{\pi}_1 - \hat{\pi}_2) \quad \text{(Wald's CI)} \]

3. CI for relative risk \( r = \pi_1 / \pi_2 \).

One can show that \( \hat{r} = \hat{\pi}_1 / \hat{\pi}_2 \) converges to the normal \( N(0, \infty) \) as \( n \to \infty \).

Transformation: \( \ln r \)

CI for \( \ln r \):
\[ \ln \hat{r} \pm Z_{\alpha/2} \cdot \hat{\sigma} (\ln \hat{r}) \]
where
\[ \hat{\sigma} (\ln \hat{r}) = \left( \frac{\hat{\pi}_1 (1 - \hat{\pi}_1)}{\hat{\pi}_1 n_1} + \frac{\hat{\pi}_2 (1 - \hat{\pi}_2)}{\hat{\pi}_2 n_2} \right)^{1/2}. \]
Consider again Table 3.1, \( n_1 = 684 \) \( n_2 = 676 \)

\[ \hat{r} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{.0409}{.0266} = 1.54 \]

\[ \ln \hat{r} = \ln 1.54 = .432 \]

\[ \hat{\sigma}(\ln \hat{r}) = \left( \frac{1-.0409}{.0409} + \frac{1-.0266}{.0266} \right)^{\frac{1}{2}} \]

\[ = .297 \]

95% CI for \( \ln r \):

\[ 0.432 \pm 1.96 \times (.297) \] or \( \ln r \in (0.150, 1.014) \)

\[ r \in (0.86, 2.75) \text{ with } 95\% \]
4. Delta Method

Let $T_n$ be an estimate of a parameter $\Theta$:

$$
\frac{T_n - \Theta}{\sigma \sqrt{n}} \xrightarrow{d} N(0,1), \quad n \to \infty.
$$

where $\sigma = \text{sd}(T_n)$, $(T_n(T_n - \Theta) \to N(0, \sigma^2))$

Consider the smooth function $g(\Theta)$ to be estimated:

Using the Taylor expansion for $g(t)$ in the neighborhood of $t = \Theta$:

$$
\ln \left[ g(T_n) - g(\Theta) \right] \approx \sqrt{n} \left[ T_n - \Theta \right] \cdot g'(\Theta)
$$

Since the variance of the right side is

$$
\sigma^2 \left[ g'(\Theta) \right]^2
$$

we can say that

$$
\frac{g(T_n) - g(\Theta)}{\sigma \cdot g'(\Theta) \sqrt{n}} \xrightarrow{d} N(0,1)
$$

or

$$
\ln \left[ g(T_n) - g(\Theta) \right] \rightarrow N(0; \left[ \frac{g'(\Theta)}{\sigma} \right]^2 \sigma^2)
$$

as $n \to \infty$. 
so one can say that

\[ g(T_n) \sim \text{Normal}(g(\theta), \frac{\sigma^2}{n}) \]

(1-\alpha)100\% CI for \( g(\theta) \):

\[ g(T_n) \pm z_{\alpha/2} \cdot \frac{g(T_n)}{\sqrt{n}} \]

4a Applications:

Logit model: \( \ln \left( \frac{\pi}{1-\pi} \right) \)

Assume that in Binomial sampling the parameter \( \pi \) is unknown.

Data: \( Y = \# \) of successes out of \( n \) trials.

MLE of \( \pi \): \( \hat{\pi} = \frac{Y}{n} \):

\[ \text{E}(\hat{\pi}) = \pi \]

\[ \text{VAR}(\hat{\pi}) = \frac{\pi(1-\pi)}{n} \]

CLT:

\[ \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1), \quad \text{i.e.} \quad T_n = \frac{\hat{\pi}}{\sqrt{\frac{\pi(1-\pi)}{n}}} \]

Consider the log odds function of \( \pi \)

\[ g(\pi) = \ln \left( \frac{\pi}{1-\pi} \right) \]

is called logit
Using the delta method with

\[ g(\hat{\pi}) = \ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) - \text{the sample logit} \]

one can show

\[ \sqrt{n} \left[ \ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) - \ln \left( \frac{\pi}{1 - \pi} \right) \right] \xrightarrow{d} N(0, \frac{1}{\pi(1 - \pi)}) \]

Indeed, let us calculate

\[ g'(\hat{\pi}) = \left( \ln \left( \frac{\pi}{1 - \pi} \right) \right)' = \frac{1 - \pi}{\pi} \cdot \left( \frac{\pi}{1 - \pi} \right)' = \frac{1 - \pi}{\pi} \cdot \frac{1}{(1 - \pi) - \pi (1 - \pi)^2} = \frac{1}{\pi(1 - \pi)} \]

\[ = \frac{1 - \pi}{\pi} \cdot \left( \frac{1 - \pi + \pi}{(1 - \pi)^2} \right) = \frac{1 - \pi}{\pi(1 - \pi)^2} = \frac{1}{\pi(1 - \pi)} \]

and

\[ \text{var} \left( \ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) \right) = \left( g'(\hat{\pi}) \right)^2 \cdot \text{var} \left( \hat{\pi} \right) = \frac{1}{\pi^2(1 - \pi)^2} \cdot \frac{1}{\pi(1 - \pi)^2} = \frac{1}{\pi^2(1 - \pi)^2} \cdot \frac{1}{n(1 - \pi)^2} \]

\[ \Rightarrow \sqrt{n} \left[ \ln \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) - \ln \left( \frac{\pi}{1 - \pi} \right) \right] \xrightarrow{d} N\left(0, \frac{1}{\pi(1 - \pi)} \right) \]
4.8. Delta Method for Log Odds Ratio

Let us consider \( \log \Theta \), where \( \Theta = \frac{\pi_{11} \cdot \pi_{22}}{\pi_{12} \cdot \pi_{21}} \).

Here \( g(\pi) = \log \Theta = \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21} \).

\[ \pi_{\Theta} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) \]

**General formula.** Let \( (\pi_1, \pi_2, \ldots, \pi_{c-1}) = \pi_{\Theta} \)

\[ \hat{\pi}_i = \frac{\hat{y}_i}{n} \]

\[ \frac{\pi_i}{\pi_{c-1}} = \frac{\hat{y}_i}{\hat{y}_{c-1}} \]

\[ \mathbb{E} \hat{\pi}_i = \pi_i \]

\[ \text{Var}(\hat{\pi}_i) = \frac{\pi_i (\pi_i - 1)}{n} \]

\[ \text{cov}(\hat{\pi}_i, \hat{\pi}_j) = -\frac{\pi_i \pi_j}{n} \quad i \neq j \]

\[ \sqrt{n} (\hat{\pi}_1 - \pi_1) \approx \text{Multivariate Normal} (0, \Sigma) \]

as \( n \to \infty \).
Denote

\[ \phi_i = \frac{d g(\pi_i)}{\partial \pi_i}, \quad i = 1, 2, \ldots, c \]

**Delta Method:**

\[
\sqrt{n} \left[ g\left( \frac{\hat{\pi}}{n} \right) - g(\bar{\pi}) \right] \sim \mathcal{N}(0, \sigma^2)
\]

with

\[
\sigma^2 = \frac{1}{c} \sum_{i=1}^{c} \phi_i^2 \cdot \pi_i - \left( \frac{1}{c} \sum_{i=1}^{c} \phi_i \cdot \pi_i \right)^2
\]

If we denote by \( \hat{\sigma}^2 \) the estimator of \( \sigma^2 \), then the estimated standard error of

\[ g\left( \frac{\hat{\pi}}{n} \right) \]

is

\[ \frac{\hat{\sigma}}{\sqrt{n}} \]


(1 - \( \alpha \))100\% Wald CI for \( g(\bar{\pi}) \):

\[ g\left( \frac{\hat{\pi}}{n} \right) \pm z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \]
Log Odds Ratios: \[ g(\pi) = \log \pi_1 + \log \pi_{21} - \log \pi_2 - \log \pi_{21} \]

\[ \phi_{ii} = \frac{\partial g}{\partial \pi_i} = -1/\pi_i \quad \phi_{21} = \frac{\partial g}{\partial \pi_{21}} = -1/\pi_{21} \]

\[ \phi_{12} = \frac{\partial g}{\partial \pi_{12}} = -1/\pi_{12} \quad \phi_{22} = \frac{\partial g}{\partial \pi_{22}} = 1/\pi_{22} \]

\[ \sum_{ij} \phi_{ij} \cdot \pi_{ij} = 1 + 1 - 1 - 1 = 0 \]

\[ \hat{\theta}^2 = \sum_{ij} \phi_{ij}^2 \cdot \pi_{ij} = \frac{1}{\pi_{11}} \cdot \pi_{11} + \frac{1}{\pi_{12}} \cdot \pi_{12} + \frac{1}{\pi_{21}} \cdot \pi_{21} + \frac{1}{\pi_{22}} \cdot \pi_{22} \]

\[ \frac{\hat{\theta}}{\sqrt{n}} = \left( \frac{1}{\pi_{11}^{1/2}} + \frac{1}{\pi_{12}^{1/2}} + \frac{1}{\pi_{21}^{1/2}} + \frac{1}{\pi_{22}^{1/2}} \right)^{1/2} \]

i.e.

\[ \frac{\hat{\theta}}{\sqrt{n}} = \left( \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2} \]

since \[ n_i \cdot \pi_{ij} = n \cdot \frac{\pi_{ij}}{n} = n \cdot \pi_{ij} \quad \text{and} \quad n \]

Remark 2: \( \hat{\theta} \) converge more slowly to Normal r.v. than \( \log \hat{\theta} \).
5. Testing independence in Two-way Contingency Tables

5a. Chi-squared Test

Data: Contingency Table: \( n_{ij}, i = 1, \ldots, I \), \( j = 1, \ldots, J \) with

Null Hypothesis:
\[
E(n_{ij}) = n_{ij} \rightarrow n_{ij} = n \frac{\pi_i \pi_j}{\pi} \]

\( H_0: \pi_{ij} = \pi_i \pi_j \) for all \( i, j \)

Remark 1: Usually, \( \pi_i \) and \( \pi_j \) are unknown and have to be estimated.

ML estimators:
\[
\hat{\pi}_i = \frac{n_i}{n} \quad \text{and} \quad \hat{\pi}_j = \frac{n_j}{n}
\]

where \( n_i = \sum_j n_{ij} \) and \( n_j = \sum_i n_{ij} \)

Denote the estimated expected frequencies (under \( H_0: \pi_{ij} = n \frac{\pi_i \pi_j}{\pi} \)) by
\[
\hat{\pi}_{ij} = n \frac{\hat{\pi}_i \hat{\pi}_j}{\hat{\pi}} = n \frac{n_i \cdot n_j}{n} = \frac{n_{ij}}{n}
\]

Test Statistic (Pearson (1900)):
\[
X^2 = \sum_{i} \sum_{j} \frac{(n_{ij} - \hat{\pi}_{ij})^2}{\hat{\pi}_{ij}}
\]
Null Distribution: Under Hypothesis $H_0$:

$X^2$ has $\chi^2$ distribution with
d.f. $= (I-1) \times (J-1)$. 

Since, the total number of classes (cells) $= I \times J$
d.f. of $X^2$ with given $\{ \pi_{i,j} \}$ and $\{ \hat{\pi}_{i,j} \}$ is

$$I \times J - 1$$

But with estimated $\{ \hat{\pi}_{i,j} \}$ and $\{ \pi_{i,j} \}$:

$$\text{d.f.} = I \times J - 1 - (I-1) - (J-1) =$$

$$= I \times J - I - (J-1) = I(J-1) - (J-1) =$$

$$= (J-1)(I-1).$$

Remark 2: dimension of $\sum_{i=1}^{I} \pi_{i,j}$ is $I-1$

since there is one tie: $\sum_{i=1}^{I} \pi_{i,j} = 1$

Similarly: $\dim \sum_{j=1}^{J} \pi_{i,j} = J-1$
58. LR Chi-squared Test

Consider the multinomial sampling again.
The kernel of likelihood is equal to
\[ \prod_{i,j} x_{ij} \quad x_{ij} > 0 \quad \sum_{i,j} x_{ij} = 1 \]

Under Hypothesis \( H_0: \pi_{ij} = \pi_i \cdot \pi_j \) for all \( i,j \),
the MLE of \( \pi_{ij} \) are
\[ \hat{\pi}_{ij} = \frac{n_i}{n} \cdot \frac{n_j}{n} \]

In the general case the
MLE of \( \pi_{ij} \): \[ \hat{\pi}_{ij} = \frac{x_{ij}}{n} \]

So that the ratio of the likelihoods equals
\[ L = \left( \frac{\pi_i}{\pi_i + \pi_j} \right)^{n_{ij}} \]
\[ = \frac{\prod_{i,j} x_{ij}^{n_{ij}}}{\prod_{i,j} \left( \frac{\pi_i}{\pi_i + \pi_j} \right)^{n_{ij}}} \]
\[ = \frac{1}{\prod_{i,j} \left( \frac{n_i}{n} \cdot \frac{n_j}{n} \right)^{n_{ij}}} \cdot n \]
\[ = \frac{\prod_{i,j} x_{ij}^{n_{ij}}}{\prod_{i,j} \left( \frac{n_i}{n} \cdot \frac{n_j}{n} \right)^{n_{ij}}} \cdot n \]
The LR Chi-squared Test Statistic:

\[ G^2 = -2 \log \Lambda = 2 \sum_{ij} n_{ij} \log \left( \frac{n_{ij}}{\hat{\mu}_{ij}} \right) \]

with \[ \hat{\mu}_{ij} = \frac{n_i + n_j}{n} \]

Null Distribution: Under \( H_0 \)

\[ G^2 \sim \text{chi-squared with d.f.} = (I-1)(J-1) \]

as \( n \to \infty \) and the number of cells is fixed.

Remark. The convergence is not good for \( \frac{n}{IJ} < 5 \).

Example: 3.2.2, Table 3.2

\( H_0: \text{Education \& Religious Fundamentalism are independent categories} \)

Let us calculate: \[ X^2 = 69.2 \quad \text{d.f.} = (3-1)(5-1) \]

\[ \hat{\mu}_{ij} = \frac{ni + nj}{n} = \frac{424 \times 886}{2746} = 137.8 \]

\[ p-value = P \left( \chi^2_{df} > 69.2 \right) < .0001 \]

Strong evidence of association.
6. Pearson and Standardized Residuals

Cell-by-cell comparison of observed \( n_{ij} \) and estimated expected frequencies helps show the nature of the association:

The Pearson residual:

\[
e_{ij} = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}} \sim N(0, 1) \text{ under } H_0
\]

\[
\sum_{ij} e_{ij}^2 = \chi^2 \text{ is a Pearson Statistic}
\]

The standardized Pearson residual:

\[
e_{ij}^* = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{p}_{ij} (1 - \hat{p}_{ij}) (1 - \hat{p}_{ij})}} \cdot \frac{\hat{n}_{ij}}{\hat{p}_{ij} + \hat{n}_{ij}} \cdot \frac{\hat{p}_{ij}}{\hat{n}_{ij}}
\]

Note that if \( |e_{ij}^*| > 2 \) or 3

this indicates lack of fit of \( H_0 \) in that cell.
Example 3.2.2, Table 3.2

Education and Religious Fundamentalism Revisited

\[ \chi^2_{11} = 178 \quad \mu_{11} = 137.8 \quad = \frac{424 	imes 886}{2726} \]

\[ p_{1+} = \frac{424}{2726} = 0.156 \]

\[ p_{+1} = \frac{886}{2726} = 0.325 \]

\[ e_{11}^* = \frac{178 - 137.8}{\sqrt{137.8 (1 - 0.156)(1 - 0.325)^{1/2}}} = 4.5 > 3 \]

i.e. the Table 3.2 shows large positive residuals with "les than 1/2 high school education & fundamentalist views"