Categorical Variables and Contingency Tables: Description and Inference

STAT 526
Professor Olga Vitek

March 3, 2011

Reading:
Agresti Ch. 1, 2 and 3
Faraway Ch. 4
Univariate Binomial and Multinomial Measurements
**Binomial Distribution**

- **Probability distribution:**
  
  - \( Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} \text{Bernouilli}(\pi) \)
  
  - \( \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \pi) \)
  
  - \( p(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \)
  
  - \( \mu = E(Y) = n\pi, \sigma^2 = \text{var}(Y) = n\pi(1 - \pi) \)

- **Log-likelihood:**
  
  - \( L(\pi) = y\log(\pi) + (n - y)\log(1 - \pi) \)

- **Maximum Likelihood Estimator:**
  
  - \( \hat{\pi} = y/n \)
  
  - \( E(\hat{\pi}) = \pi, \text{SE}(\hat{\pi}) = \sqrt{\frac{\pi(1 - \pi)}{n}} \)
Large-sample tests for $\pi$

- For a known $\pi_0$, test
  
  $H_0 : \pi = \pi_0 \text{ vs } H_0 : \pi \neq \pi_0$

- Wald test:
  
  $z_W = \frac{\hat{\pi} - \pi_0}{SE} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})}/n}$
  
  $H_0,\text{approx} \sim N(0, 1)$

- Likelihood ratio Test:
  
  $z_L = 2(L_1 - L_0) = 2\left(y\log\frac{\hat{\pi}}{\pi_0} + (n - y)\log\frac{1 - \hat{\pi}}{1 - \pi_0}\right)$
  
  $H_0,\text{approx} \sim \chi^2_1$

- Score Test:
  
  $z_S = \frac{\hat{\pi} - \pi_0}{SE_0} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)}/n}$
  
  $H_0,\text{approx} \sim N(0, 1)$

Closer to $N(0, 1)$ than Wald
Large-sample CI for $\pi$

- Based on the Wald test statistic:

$$\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Performs poorly unless large $n$

- Based on the Score Test statistic:

$$\hat{\pi} \left( \frac{n}{n + z^2_{\alpha/2}} \right) + \frac{1}{2} \left( \frac{z^2_{\alpha/2}}{n + z^2_{\alpha/2}} \right)$$

$$\pm z_{\alpha/2} \sqrt{\frac{1}{n + z^2_{\alpha/2}} \left[ \hat{\pi}(1-\hat{\pi}) \left( \frac{n}{n + z^2_{\alpha/2}} \right) + \frac{1}{2} \cdot \frac{1}{2} \left( \frac{z^2_{\alpha/2}}{n + z^2_{\alpha/2}} \right) \right]}$$

Performs better than Wald
Multinomial Distribution

• Probability distribution:
  
  - \( (Y_{i1}, \ldots, Y_{ic}) \sim \{Y_{ij} = 1 \text{ if in category } j, \text{ and } 0 \text{ otherwise } \} \)
  
  - \( \sum_{i=1}^{n} Y_{ij} \sim \text{Multinomial}(\pi_1, \ldots, \pi_c), \ n = \sum_{j=1}^{c} n_j \)
  
  - \( p(n_1, n_2, \ldots, n_{c-1}) = \left( \frac{n!}{n_1!n_2! \ldots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c} \)
  
  - \( E(n_j) = n\pi_j \)
    \( \text{var}(n_j) = n\pi_j(1 - \pi_j), \ \text{cov}(n_j, n_k) = -n\pi_j\pi_k \)

• Log-likelihood:
  
  - \( L(\pi) = \sum_{j=1}^{c} n_j \cdot \log\pi_j \)

• Maximum Likelihood Estimator:
  
  - \( \hat{\pi}_j = n_j/n \)
Large-Sample Test for 
\((\pi_1, \ldots, \pi_c)\)

- For known \((\pi_{10}, \pi_{20}, \ldots, \pi_{c0})\), test 
  \(H_0: \pi_j = \pi_{j0} vs H_0: \pi_j \neq \pi_{j0}\)

- **Pearson test:**
  \[
  X^2 = \sum_{j=1}^{c} \frac{(O_j - E_{j0})^2}{E_{j0}} = \sum_{j=1}^{c} \frac{(n_j - n\pi_{j0})^2}{n\pi_{j0}} \sim \chi_{c-1}^2
  \]

  E.g. in genetics: test theories of trait inheritance

- **Likelihood Ratio test:**
  \[
  G^2 = 2(L_1 - L_0) = 2\sum_{j=1}^{n} \log\left(\frac{n_j}{n\pi_{j0}}\right) \sim \chi_{c-1}^2
  \]

  Asymptotically equivalent when \(H_0\) is true.

- For \(n/c < 5\), \(X^2\) converges faster
Poisson Distribution

• Probability distribution:
  - $Y$ - number of events in a fixed interval of space/time
  - $Y \sim \text{Poisson}(\mu)$
  - $p(y) = \frac{e^{-\mu} \mu^y}{y!}$, $y = 0, 1, \ldots$; $E(Y) = \text{var}(Y) = \mu$
  - $Y_1, Y_2, \ldots, Y_c \overset{\text{ind}}{\sim} \text{Poisson}(\mu_i), \sum_{i=1}^c Y_i \sim \text{Poisson}(\sum_{i=1}^c \mu_i)$

• $c$ indep. Poisson r.v. | total $\sim$ Multinomial

\[
P(Y_1 = n_1, \ldots, Y_c = n_c \mid \sum_i Y_i = n) = \frac{P(Y_1 = n_1, \ldots, Y_c = n_c)}{P(\sum_i Y_i = n)}
= \frac{\prod_i \left[ \exp(-\mu_i) \mu_i^{n_i}/n_i! \right]}{\exp(-\sum_i \mu_i) (\sum_i \mu_i)^n/n!} = \frac{n!}{\prod_i n_i! \prod_i \pi_i^{n_i}} \prod_i \pi_i^{n_i}, \quad \pi_i = \frac{\mu_i}{\sum_i \mu_i}
\]
2-Way Contingency Tables
Contingency Tables

- Contingency Table = Classification Table: frequency of outcomes

- Two-Way Table: frequency outcomes of two categorical variables

- $I \times J$ table: a table with $I$ rows and $J$ columns.

- Contingency tables can arise from several sampling schemes
  - Inference depends on the sampling scheme

- Example:

<table>
<thead>
<tr>
<th>Smoking</th>
<th>Lung Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
<td>650</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>709</td>
<td>709</td>
</tr>
</tbody>
</table>
Joint Distribution and Independence

- Underlying probability distribution of $X$ (smoking) and $Y$ (cancer)

- Joint distribution:
  - $\pi_{ij}$, probability of cell $(i, j)$

- Marginal distribution:
  - $\pi_{i+} = \sum_{j=1}^{J} \pi_{ij}$, probability of row $i$
  - $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$, probability of column $j$

- Conditional distribution:
  - $\pi_{j|i} = \pi_{ij}/\pi_{i+}$, distribution of $j$ given $i$

- Independence:
  - $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all $i$ and $j$
Multinomial Sampling

• The total sample size $n$ is fixed, but the row and column totals are not

• $X$ and $Y$ are treated equally
  
  $P(X = i, Y = j) = \pi_{ij}, \ i = 1, \ldots, I; \ j = 1, \ldots, J$

  – describe associations with joint distributions.

  – back to the case of the Multinomial distribution

• Likelihood and log-likelihood:

  $Likelihood = \frac{n!}{n_{11}! \cdots n_{IJ}!} \prod_{i} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}}$

  $L = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log(\pi_{ij}) + constant$
Multinomial Sampling: Testing for Independence

- Hypotheses:
  - $H_0$: reduced model $\pi_{ij} = \pi_i + \pi_j$, for all $i$ and $j$
  - $H_a$: full model $\pi_{ij} \neq \pi_i + \pi_j$, for some $i$ and $j$

- Pearson $\chi^2$ test:
  - $X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} H_0$, approx. $\sim \chi^2_{(I-1)(J-1)}$
  - $O_{ij} = n_{ij}$, $E_{ij} = n\hat{\pi}_i \hat{\pi}_j = n_i n_j / n$
  - Df = $(I - 1)(J - 1) = (IJ - 1) - (I - 1) - (J - 1)$

- Likelihood Ratio test:
  - Full model: $\hat{\pi}_{ij} = n_{ij} / n_{++}$
  - Reduced model: $\hat{\pi}_{i+} = n_{i+} / n_{++}$, $\hat{\pi}_{+j} = n_{+j} / n_{++}$.
  - $G^2 = 2(L_1 - L_0)$
  - $G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \log \frac{n_{ij} n_{++}}{n_i n_j + n_{++}} H_0$, approx. $\sim \chi^2_{(I-1)(J-1)}$
Independent (or Product) Multinomial Sampling

• The row totals $n_{i+}$, $i = 1, \ldots, I$, are fixed
  
  – E.g., $X$ is an explanatory variable, and response $Y$ occurs separately at each setting of $X$.
  
  – View categorical response as function of categorical predictor
  
  – Describe associations in terms of conditional distributions
  
  $P(Y = j|X = i) = \pi_{j|i}$, $i = 1, \ldots, I$; $j = 1, \ldots, J$

• For a fixed $i$, $\{n_{ij}, j = 1, \ldots, J\}$ follow a multinomial distribution

  $$f(n_{i1}, \ldots, n_{iJ}) = \frac{n_{i+}!}{n_{i1}! \cdots n_{iJ}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}$$
Compare Proportions

• Independent Multinomial Sampling

• \( H_0 : \pi_1 = \pi_2 \) vs \( H_a : \pi_1 \neq \pi_2 \)

• ML estimate of the difference:
  \[
  \hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}
  \]

  \[
  SE(\hat{\pi}_1 - \hat{\pi}_2) = \left[ \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} \right]^{1/2}
  \]

• Wald Confidence Interval:
  \[
  \hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \hat{SE}(\hat{\pi}_1 - \hat{\pi}_2)
  \]
  
  - Replace \( \pi \) with \( \hat{\pi} \) to estimate SE

• Usually too narrow

• Better methods (e.g. delta method) exist
Testing for Independence of Rows and Columns

• Independent Multinomial Sampling

• Independence in this context is often called *homogeneity* of the conditional distributions

• $X$ and $Y$ are independent
  \[\iff \pi_j|1 = \cdots = \pi_j|I, \text{ for all } j\]

• Can interpret the independence in terms of product of marginal probabilities

• $\pi_{ij} = \pi_i + \pi_j$ for all $i$ and $j$
  \[\iff \pi_j|1 = \cdots = \pi_j|I\text{ for all } j\]

\[
\begin{align*}
\Rightarrow & \quad \pi_j|i = \pi_{ij}/\pi_i+ = (\pi_i+\pi_j)/\pi_i+ = \pi_j+ \\
\Leftarrow & \quad \text{Let } \pi_j|i = a_j, \text{ then } \pi_j+ = \sum_{i=1}^{I} \pi_{ij} = \sum_{i=1}^{I} \pi_i+a_j = a_j \Rightarrow \pi_{ij} = \pi_i+\pi_j
\end{align*}
\]
Testing for Independence of Rows and Columns

- Test the homogeneity of conditional distributions

\[
\begin{array}{c|ccc|c}
\text{Row} & 1 & \cdots & J & \text{Total} \\
\hline
1 & \pi_{11} & \cdots & \pi_{1J} & \pi_{1+} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
I & \pi_{I1} & \cdots & \pi_{IJ} & \pi_{I+} \\
\hline
\text{Total} & \pi_{+1} & \cdots & \pi_{+J} & \pi_{++}
\end{array}
\]

- Consider the new notation:
  \[\pi_j(x) = P(Y = j | X = x)\]

- Although the interpretation is different, use the same Pearson \(X^2\) test and the LR test
Test for Independence: Odds Ratio

• Odds Ratio:

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

$$= \frac{P(Y = 1|X = 1)/P(Y = 2|X = 1)}{P(Y = 1|X = 2)/P(Y = 2|X = 2)}$$

$$= \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)}$$

• Equally valid for prospective (conditional on $X$), retrospective (conditional on $Y$) and cross-sectional (multinomial) sampling designs

• MLE: $\hat{\theta} = \frac{n_{11}/n_{12}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

  – When some $n_{ij} = 0$, $\hat{\theta}$ is not a good estimator. Is improved by adding 0.5 to each cell count:

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$
Test for Independence:
Odds Ratio

- $X$ and $Y$ are independent

$$\iff \theta = \frac{\frac{\pi_{11}}{\pi_{12}}}{\frac{\pi_{21}}{\pi_{22}}} = \frac{\frac{\pi_{11}}{\frac{\pi_{12}}{\pi_{22}}} = \frac{\pi_{11} \pi_{22}}{\pi_{12} \pi_{21}} = 1$$

- to check, substitute $\pi_{ij} = \pi_{i+} \pi_{+j}$ in the formula above

- Asymptotically, $\log \hat{\theta} \sim N(\log(\theta), \hat{\sigma}^2)$, where

$$\hat{\sigma}^2 = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

- Large-sample CI for $\log \theta$:

$$\log \theta \pm z_{\alpha/2} \widehat{SE}(\log \hat{\theta}) = [L, U]$$

- Large-sample CI for $\theta$ : $[e^L, e^U]$.

  - Usually too wide
Poisson Sampling

• Observe a process over a period of time, and observe the number of occurrences
  – No fixed quantities
  – Poisson sampling assumes each $Y_{ij} \sim \text{Poisson}(\pi_{ij})$

• Denote $Y_{ij}$ the count of cell $(i, j)$

\[
\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij} \sim \text{Poisson} \left( \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} \right)
\]

• Hypothesis of independence of $X$ and $Y$ has the form $\log(\pi_{ij}) = \lambda + \alpha_i + \beta_j$
  – This is the log-linear model of independence for two-way contingency tables
  – Under independence, $\log(\mu_{ij})$ is an additive function of a row effect $\alpha_i$ and a column effect $\beta_j$.
  – Since we don’t have a replicate table, the model with the interaction is saturated
Poisson Sampling

• An additive model

\[ \log \pi_{ij} = \mu + \alpha_i + \beta_j \]

implies the independence of the margins

\[ \pi_{ij} = \frac{E(\text{count})}{\text{sum of all } E(\text{count})} = \frac{e^{\mu+\alpha_i+\beta_j}}{e^{\mu} (\sum_i e^{\alpha_i}) (\sum_j e^{\beta_j})} = \pi_{i+} \pi_{+j}, \]

where

\[ \pi_{i+} = e^{\alpha_i} / \sum_i e^{\alpha_i} = \sum_j \pi_{ij}, \]
\[ \pi_{+j} = e^{\beta_j} / \sum_j e^{\beta_j} = \sum_i \pi_{ij}. \]

• Test for independence: Pearson \( X^2 \) or LR test as before (more on this later)
Hypergeometric Sampling

- Both row and column margins are fixed.

- When $X$ and $Y$ are independent, given the row and column margins, follows hypergeometric distribution

\[
\frac{\left(\prod_{i=1}^{I} n_{i+!}\right) \left(\prod_{j=1}^{J} n_{+j!}\right)}{n_{++}! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!} \]

— the distribution is parameter free

- For a $2 \times 2$ table

\[
P(n_{11} = k) = \frac{\binom{n_{1+}}{k} \binom{n_{2+}}{n_{+1} - k}}{\binom{n_{++}}{n_{+1}}},
\]

\[
max(0, n_{1+} + n_{+1} - n) \leq k \leq min(n_{1+}, n_{+1})
\]

— Fisher’s exact test: $p$-value = total probability of all outcomes more extreme than the one observed.

— Takes discrete values for small samples
Case study: Agresti p.80

#--------------------read the data----------------------
X <- data.frame(y=c(178, 138, 108, 570, 648, 442, 138, 252, 252), belief=rep(c("1-Fundam", "2-Moder", "3-Liber"), 3), degree=rep(c("1-<HS", "2-HS", "3-BS/grad"), 1, each=3))

#------------ a table of observed values (ov)-----------
ov <- xtabs(y ~ degree+belief, data=X)
> ov

   belief
degree   1-Fundam 2-Moder 3-Liber
1-<HS 178.00 138.00 108.00
2-HS  570.00 648.00 442.00
3-BS/grad 138.00 252.00 252.00

#--------------export the table into latex---------------
# export the table into latex
library(xtable)
xtable(ov)

\begin{table}[ht]
\begin{center}
\begin{tabular}{rrrr}
\hline
& 1-Fundam & 2-Moder & 3-Liber \\
1-<HS & 178.00 & 138.00 & 108.00 \\
2-HS & 570.00 & 648.00 & 442.00 \\
3-BS/grad & 138.00 & 252.00 & 252.00 \\
\hline
\end{tabular}
\end{center}
\end{table}
Data visualization

```r
#------------------------dotchart------------------------
dotchart(t(ov), xlab="Observed counts")
```

![Data visualization diagram](image)

1−<HS
2−HS
3−BS/grad

1−Fundam
2−Moder
3−Liber

Observed counts
Data visualization

#-----------------mosaic plot-------------------
mosaicplot(ov, color=TRUE)
2 x 2 table: Compare Proportions

- Independent multinomial sampling: restrictions on the rows
  - compare proportions of columns, given rows
  - also implements the Pearson $X^2$ test with Yates correction for small samples (from each O-E, subtract 0.5 if positive, and add 0.5 if negative)

```r
> prop.test(ov[1:2,1:2])
2-sample test for equality of proportions with continuity correction
data: ov[1:2, 1:2]
X-squared = 8.7451, df = 1, p-value = 0.003104
alternative hypothesis: two.sided
95 percent confidence interval:
  0.03187153 0.15875016
sample estimates:
  prop 1  prop 2
0.5632911 0.4679803

#-----double-check the proportions----------
> 178/(178+138)
[1] 0.5632911
> 570/(570+648)
[1] 0.4679803
```
2 x 2 table: Hypergeometric

- Sampling conditional on both margins
  - Hypergeometric test
  - compare distributions of counts within the 4 cells
  - $H_0$ is specified in terms of OR=1
  - produces CI for the OR

> fisher.test(ov[1:2,1:2])

Fisher’s Exact Test for Count Data

data:  ov[1:2, 1:2]
p-value = 0.002961
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.134415 1.897137
sample estimates:
  odds ratio
    1.465974
I x J table: Pearson $X^2$

- (Independent) multinomial sampling restrictions on a margin, or on the total
  - $H_0$ in terms of independence of rows and columns

```r
> summary(ov)
```

Call: xtabs(formula = y ~ degree + belief, data = X)
Number of cases in table: 2726
Number of factors: 2
Test for independence of all factors:
Chisq = 69.16, df = 4, p-value = 3.42e-14

- **Pearson residuals**
  - $e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\hat{\mu}_{ij}^{1/2}}$
    - divide residual by $\hat{SE}(n_{ij})$ in Poisson sampling

- **Standardized Pearson residuals**
  - $e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij} \cdot (1-p_i) \cdot (1-p_j)}}$
    - divide residual by $\hat{SE}(\text{residual})$ in Poisson sampling
Visualizing the association

# --Compute Pearson and standardized Pearson residuals ---
e <- apply(ov, 1, sum) %*% t(apply(ov, 2, sum)) / sum(ov)
pearsonResid <- (ov - e)/sqrt(e)

pRow <- 1-apply(ov, 1, sum) / sum(ov)
pCol <- 1-apply(ov, 2, sum) / sum(ov)
standPearsonResid <- pearsonResid/ sqrt(pRow %*% t(pCol) )

dotchart( t(standPearsonResid) )
abline(v=c(-2,2))
Ordered Categories

• Ordered categories have more info

• Assign scores to categories
  – Rows: \((u_1 \leq \ldots \leq u_I)\), e.g. \((1, \ldots, I)\)
  – Cols: \((v_1 \leq \ldots \leq v_J)\), e.g. \((1, \ldots, J)\)
  – \(H_0: \operatorname{cor}(u, v) = 0\) vs \(H_a: \operatorname{cor}(u, v) \neq 0\)
  – produces CI for the OR

• Study the linear trend

\[
r = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} (u_i - \bar{u})(v_j - \bar{v})n_{ij}}{\sqrt{\left[\sum_{i=1}^{I} \sum_{j=1}^{J} (u_i - \bar{u})^2 n_{ij}\right] \cdot \left[\sum_{i=1}^{I} \sum_{j=1}^{J} (v_i - \bar{v})^2 n_{ij}\right]}}
\]

\[
\bar{u} = \sum_{i=1}^{I} \sum_{j=1}^{J} u_i n_{ij} / n; \quad \bar{v} = \sum_{i=1}^{I} \sum_{j=1}^{J} v_i n_{ij} / n;
\]

\[
M^2 = (n - 1)r^2 \overset{H_0}{\sim} \chi_1^2
\]
Case Study: Ordered Categories

#-------------------existing implementation-----------------
> library(coin)
> lbl_test(as.table(ov))

Asymptotic Linear-by-Linear Association Test

data: belief (ordered) by degree (1-<HS < 2-HS < 3-BS/grad)
chi-squared = 56.0849, df = 1, p-value = 6.939e-14

#-------------------manually-----------------

u <- as.vector(scale(1:3, center=sum(c(1:3)*ov)/sum(ov),
scale=FALSE))
v <- as.vector(scale(1:3, center=sum(t(ov)*c(1:3))/sum(ov),
scale=FALSE))

r <- sum(u%*%t(v)*ov) / sqrt(sum(u^2*ov) * sum(t(ov) * v^2))

M2 <- (sum(ov) - 1) * r^2
> 1-pchisq(M2, 1, lower=TRUE)
[1] 6.938894e-14
2x2 pairs: Matched Pairs

- Repeated measurements on same subjects
  - ask the same people the same question twice
  - goal: compare proportions
  - absence of association cannot be interpreted as independence

- Example (Agresti Ch. 10.1)
  - Approval of the President’s performance, one month apart, for a same sample of Americans.

<table>
<thead>
<tr>
<th></th>
<th>Approve</th>
<th>Disapprove</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approve</td>
<td>794.00</td>
<td>150.00</td>
</tr>
<tr>
<td>Disapprove</td>
<td>86.00</td>
<td>570.00</td>
</tr>
</tbody>
</table>

- $H_0$ : Marginal homogeneity. $\pi_{1+} = \pi_{+1}$
  - $\delta = \pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12} - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$
  - Equivalent to testing table symmetry
Large-sample test and CI

• CI

\[ \hat{\delta} = p_+^1 - p_+^2 = p^2_+ - p^2_+ \]

\[ \text{var}(\hat{\delta}) = \left[ \pi^1_+ (1 - \pi^1_+) + \pi^1_+ (1 - \pi^1_+) - 2(\pi^1_1 \pi^2_2 - \pi^1_2 \pi^2_1) \right] / n \]

• smaller variance than in independent samples, therefore a more efficient design

\[ \text{var}(\hat{\delta}) = \left[ (p_{12} + p_{21}) - (p_{12} - p_{21})^2 \right] / n \]

• CI: \[ \hat{\delta} \pm z_{\alpha/2} \hat{SE}(\delta) \]

• Wald Test

\[ z = \frac{\hat{\delta}}{\hat{SE}(\delta)} = \frac{n_{21} - n_{12}}{(n_{21} + n_{12})^{1/2}} \]

\[ z^2 \overset{H_0}{\sim} \chi^2_1 \text{ (called McNemar test)} \]

• Only depends on counts outside of the diagonal
President Approval Example

#-----------------Read the data-------------------
Performance <-
matrix(c(794, 86, 150, 570),
nrow = 2, dimnames =
  list("1st Survey" = c("Approve", "Disapprove"),
       "2nd Survey" = c("Approve", "Disapprove"))
)

> Performance
   2nd Survey
1st Survey  Approve Disapprove
   Approve   794    150
   Disapprove   86    570

#-----------------Test---------------------
> mcnemar.test(Performance)

McNemar’s Chi-squared test with continuity correction
data:  Performance
McNemar’s chi-squared = 16.8178, df = 1, p-value = 4.115e-05

• significant change (in fact, drop) in approval ratings