Contingency Table Analysis

Bruce A Craig

Department of Statistics
Purdue University

Reading: Faraway Ch. 6, Agresti Ch. 1-3

Contingency Tables

- Used to summarize data on two or more categorical variables
- Two types of categorical variables:
  - Nominal: no obvious ordering of categories
  - Ordinal: has a natural default ordering of categories
- Common scales for ordinal data are:
  - Rating data
    - Likert-scale: Strongly agree to Strongly disagree
  - Interval-scale data
    - Discretize a continuous variable: Age ranges
- Analysis will focus on models that predict the counts in each cell of the table

Contingency Table Structure

- Assume X has I categories and Y has J categories
- Table has IJ cells

				Y	
X	1	2	$\cdots$	J	Total
1	y_{11}	y_{12}	$\cdots$	y_{1J}	$y_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2J}	$y_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$y_{I.}$
Total	$y_{.1}$	$y_{.2}$	$\cdots$	$y_{.J}$	n

Contingency Table Distributions

- Binomial
- Multinomial
- Poisson
- Hypergeometric

- We'll now discuss each distribution as it applies to a 2×2 table of counts in terms of ML estimators and inference

Binomial Distribution

- Used to model number of successes in m trials
- Probability distribution:
  - $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} Bernoulli(p)$
  - $\sum_{i=1}^{m} Y_i \sim B(m, p)$
  - $p(y) = \binom{m}{y} p^y (1-p)^{m-y}$
  - $E(Y) = \mu = mp$ and $\text{Var}(Y) = mp(1-p)$
- Log-likelihood:
  - $l(p) = y \log(p) + (m-y) \log(1-p) + C(m, y)$
- Maximum Likelihood Estimator:
  - $\hat{p} = y/m$
  - $E(\hat{p}) = p$ and $SD(\hat{p}) = \sqrt{p(1-p)/m}$

Large-Sample Tests for p

- Wald test:

$$z_W = \frac{\hat{p} - p_0}{SE\{\hat{p}\}} = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/m}} \overset{H_0, approx}{\sim} \mathcal{N}(0,1)$$

- Likelihood ratio Test:

$$2(l(p) - l(p_0) = 2\left(y \log \frac{\hat{p}}{p_0} + (m-y)\log \frac{1-\hat{p}}{1-p_0}\right) \overset{H_0, approx}{\sim} \chi^2_1$$

- Score Test:

$$z_S = \frac{\hat{p} - p_0}{SE\{p_0\}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/m}} \overset{H_0, approx}{\sim} \mathcal{N}(0,1)$$

Large-sample CI for p

- Using Wald test statistic:

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

  - Performs poorly unless large m
- Using score test statistic:

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2m}}{1 + \frac{z^2_{\alpha/2}}{m}} \pm \frac{z_{\alpha/2}}{1 + \frac{z^2_{\alpha/2}}{m}}\sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \frac{z^2_{\alpha/2}}{4m^2}}$$

  - Performs better than Wald
  - For 95%, can approximate using $m^* = m + 4$ and $\hat{p}^* = (y+2)/(m+4)$ in Wald formula
- Can also use likelihood ratio approach (Topic 4)

Multinomial Distribution

- Used to model variables with **two or more** categories
- Probability distribution:
  - $\mathbf{Y} = (Y_1, \ldots, Y_c) \sim Multinomial(n, p_1, \ldots, p_{c-1})$
    - n is considered known (number of trials)
    - $p_c = 1 - \sum_{j=1}^{c-1} p_j$
  - $p(y_1, y_2, \ldots, y_c) = \left(\frac{n!}{y_1! y_2! \ldots y_c!}\right) p_1^{y_1} p_2^{y_2} \ldots p_c^{y_c}$
  - $E(Y_j) = np_j$, $\mathrm{Var}(Y_j) = np_j(1-p_j)$, and $\mathrm{Cov}(Y_j, Y_k) = -np_j p_k$
  - Marginal dist for each Y_j is $\mathrm{B}(n, p_j)$
- Log-likelihood:
  - $l(\mathbf{p}) = \sum_{j=1}^{c} y_j \log p_j$
- Maximum Likelihood Estimator:
  - $\hat{p}_j = y_j/n$

Large-Sample Test for $(p_1, \ldots, p_{c-1})$

- Testing $H_0 : \mathbf{p} = \mathbf{p_0} = (p_{10}, p_{20}, \ldots p_{c0})$
- Pearson test:

$$X^2 = \sum_{j=1}^{c} \frac{(O_j - E_{j0})^2}{E_{j0}} = \sum_{j=1}^{c} \frac{(y_j - np_{j0})^2}{np_{j0}} \overset{H_0,\ approx}{\sim} \chi^2_{c-1}$$

- Likelihood Ratio test:

$$G^2 = 2(l(\hat{\boldsymbol{p}}) - l(\mathbf{p_0})) = 2\sum_{j=1}^{c} y_j \log\left(\frac{y_j}{np_{j0}}\right) \overset{H_0,\ approx}{\sim} \chi^2_{c-1}$$

- 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
  - $p(y) = \frac{\exp\{-\lambda\}\lambda^y}{y!},\ y = 0, 1, \ldots$
  - $E(Y) = \mathrm{Var}(Y) = \lambda$
  - $Y_1, Y_2, \ldots, Y_k \overset{indep}{\sim} Pois(\lambda_i)$ then $\sum_{i=1}^{k} Y_i \sim Pois(\sum_{i=1}^{k} \lambda_i)$
- If c indep. Poisson random variables

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_c = y_c \mid \sum_i Y_i = n) &= \frac{P(Y_1 = y_1, \ldots, Y_c = y_c)}{P(\sum_i Y_i = n)} \\
&= \frac{\prod_i \left[\exp\{-\lambda_i\}\lambda_i^{y_i}/y_i!\ \right]}{\exp\{-\sum_i \lambda_i\}(\sum_i \lambda_i)^n/n!} \\
\text{(link to multinomial)} &= \frac{n!}{\prod_i y_i!} \prod_i \left(\lambda_i / \sum_i \lambda_i\right)^{y_i}
\end{aligned}
$$

Hypergeometric Distribution

- Probability distribution:
  - Y - number of successes in n draws from population of size N with K successes
  - Diff from binomial: draws **without replacement**

$$p(y) = \frac{\binom{K}{y}\binom{N-K}{n-y}}{\binom{N}{n}}$$

  - $E(Y) = nK/N = np$, $\text{Var}(Y) = np(1-p)(N-n)/(N-1)$
  - $(N-n)/(N-1)$ is the finite population correction factor
  - When $n = 1$, Y had Bernoulli distribution
  - When N and K are large compared to n, $Y \approx B(n, p)$

Two-way Contingency Tables

- Joint summary of outcomes for two categorical variables
- Tables can arise from several sampling schemes
  - Inference method depends on the sampling scheme
- Example: semiconductor wafers cross classified

	Particles Found		
Quality	No	Yes	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Distributions for Example

- Poisson: Observed manufacturing process for fixed period of time: 450 wafers were produced and cross-classified
- Multinomial: Sampled 450 wafers and cross-classified each of them
- Binomial: Select 400 wafers without particles and 50 with particles and then recorded the good or bad outcome

- All sampling schemes are plausible
- All lead to exactly the same conclusion

Poisson Model

- Model counts as function of quality and particles

```
> y = c(320,14,80,36)
> particle = c("No","Yes","No","Yes")
> quality = c("Good","Good","Bad","Bad")
> modp = glm(y~particle+quality)
> summary(modp)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.63581    0.09433  49.144   <2e-16 ***
particleYes -2.07944    0.15000 -13.863   <2e-16 ***
qualityGood  1.05755    0.10777   9.813   <2e-16 ***

    Null deviance: 474.10  on 3  degrees of freedom
Residual deviance:  54.03  on 1  degrees of freedom
AIC: 83.774
```

Summary: Poisson Model

- Null model assumes rate the same for all combinations of particle and quality
- Can be rejected given deviance 474.1 on 3 df

- Additive model is improvement but not great fit
  - $\hat{\mu}_{GN} = 296.89, \hat{\mu}_{GY} = 37.11, \hat{\mu}_{BN} = 103.11, \hat{\mu}_{BY} = 12.89$
- Addition of interaction term would saturate model
- Deviance would drop from 54.3 on 1 df $\rightarrow$ 0 on 0 df
- Thus, likelihood ratio/deviance test supports interaction
- Conclusion: presence of particles related to the quality

Binomial Model

- Probability of good outcome same in both particle groups

```
> y1 = matrix(y,ncol=2,byrow=F)
> modb = glm(y1~1, family=binomial)
> summary(modb)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0576     0.1078   9.813   <2e-16 ***

    Null deviance: 54.03  on 1  degrees of freedom
Residual deviance: 54.03  on 1  degrees of freedom
AIC: 66.191
```

Summary: Binomial Model

- Null model assumes probability of good the same in both particles and no-particles groups
  - Estimated probability: $\hat{p} = 0.742 = 334/450$
- Null can be rejected given deviance 54.03 on 1 degree of freedom (allowing probability to vary $\rightarrow$ saturated model)

- Conclusion: Homogeneity of probabilities test corresponds with test of independence in Poisson model

Distributions of X and Y

- Joint distribution:
  - Defined by the p_{ij}, probability of cell (i, j)
- Marginal distributions:
  - For X: Defined by $p_{i.} = \sum_{j=1}^{J} p_{ij}$, probability of row i
  - For Y: Defined by $p_{.j} = \sum_{i=1}^{I} p_{ij}$, probability of column j
- Conditional distributions:
  - For $Y|X$: Defined by $p_{j|i} = p_{ij}/p_{i.}$, dist of j given i
  - For $X|Y$: Defined by $p_{i|j} = p_{ij}/p_{.j}$, dist of i given j
- Independence of X and Y:
  - $p_{ij} = p_{i.}p_{.j}$ for all i and $j \rightarrow p_{i|j} = p_{i.}$

Multinomial Model: Testing Independence

- Hypotheses:
  - H_0 : reduced model $p_{ij} = p_{i.}p_{.j}$, for all i and j
  - H_a : full model $p_{ij} \neq p_{i.}p_{.j}$, for some i and j
- Pearson $\chi 2$ test:
  - $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \overset{H_0,\ approx.}{\sim} \chi^2_{(I-1)(J-1)}$
  - $O_{ij} = y_{ij}$, $E_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = y_{i.}y_{.j}/n$
- Likelihood Ratio test:
  - Full model: $\hat{p}_{ij} = y_{ij}/n$
  - Reduced model: $\hat{p}_{i} = y_{i.}/n$; $\hat{p}_{.j} = y_{.j}/n$

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log \frac{y_{ij}n}{y_{i.}y_{.j}} \overset{H_0,\ approx.}{\sim} \chi^2_{(I-1)(J-1)}$$

Multinomial Model

- Assume total sample size n is fixed
- Will assess if particles and quality are independent

```
> obsval = xtabs(y~quality*particle)
> probp = prop.table(xtabs(y~particle))
> probq = prop.table(xtabs(y~quality))
> fitval = outer(probq,probp)*450
```

```
       particle                          particle
quality       No      Yes          quality  No Yes
   Bad  103.1111 12.88889             Bad   80  36
   Good 296.8889 37.11111             Good 320  14
```

```
##Deviance
> 2*sum(obsval*log(obsval/fitval))
[1] 54.03045
> ##Pearson
> sum((obsval-fitval)^2/fitval)
[1] 62.81231
```

Summary: Multinomial

- Estimated means under multinomial same as Poisson
- This is expected given described relationship between multinomial and Poisson distributions
- Deviance of 54.03 on 1 degree of freedom or the Pearson $X^2 = 62.8$ does not suggest a good fit
- Conclusion: Multinomial test of independence same as homogeneity of probabilities test (binomial) and test of independence (Poisson)

Hypergeometric Model

- Could consider example table created in which marginal total are fixed and the cell counts were observed
- Less common in practice but provides more accurate test of independence if possible to consider this way
- Famous example: Fisher's Lady Tasting Tea...told there were 4 cups of each type and asked to identify them
- With margins fixed, only one cell count can vary
- Can compute probability of all outcomes/tables as extreme as the one observed table
- Know as Fisher's exact test
- Does not require need to make χ^2 distribution assumption

Fisher's Exact Test

```
> obsval = xtabs(y~quality+particle)
> obsval
         particle
quality   No Yes
   Bad    80  36
   Good 320  14

> fisher.test(obsval)

Fisher's Exact Test for Count Data
data:  obsval
p-value = 2.955e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.04641648 0.19643940
sample estimates:
odds ratio
0.09791147
```

***Odds ratio is a measure of association $\frac{\text{Odds(No|Bad)}}{\text{Odds(No|Good)}}$

Homogeneity of Conditional Distributions

- Alternative view of multinomial resulting in same test
- Extension of binomial test to Y having $J > 2$ outcomes
- Consider row totals $y_{i.}$ are fixed
  - X is an explanatory variable and response Y occurs separately at each setting of X
  - Categorical response a function of categorical predictor
  - Describe association using conditional distributions

$$P(Y = j|X = i) = p_{j|i}, \quad i = 1, \cdots, I; \quad j = 1, \cdots, J$$

- For a fixed i, $\{y_{ij}, j = 1, \cdots, J\}$ follow a multinomial dist

$$f(y_{i1} \ldots, y_{iJ}\} = \frac{y_{i.}!}{y_{i1}! \cdots y_{iJ}!} \prod_{j=1}^{J} p_{j|i}^{y_{ij}}$$

Testing for Homogeneity

- Interpret contingency table in terms of cond dists

		Y		
X	1	$\cdots$	J	Total
1	p_{11} $(p_{1\|1})$	$\cdots$	p_{1J} $(p_{J\|1})$	$p_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	p_{I1} $(p_{1\|I})$	$\cdots$	p_{IJ} $(p_{J\|I})$	$p_{I\cdot}$
Total	$p_{\cdot 1}$	$\cdots$	$p_{\cdot J}$	p

- Consider the new notation: $p_j(x) = P(Y = j|X = x)$
- Although the interpretation is different, use the same Pearson X^2 test and the LR test

Why Same Test?

- Want to show:

  X and Y are independent $\iff p_{j|1} = \cdots = p_{j|I}$, for all j
- Same as interpreting independence in terms of the product of marginal probabilities:

  $p_{ij} = p_{i.}p_{.j}$ for all i and $j \iff p_{j|1} = \cdots = p_{j|I}$ for all j

$$\text{“}\Rightarrow\text{”} \quad p_{j|i} = p_{ij}/p_{i.} = (p_{i.}p_{.j})/p_{i.} = p_{.j} \text{ for all } i$$

$$\text{“}\Leftarrow\text{”} \quad \text{Let } p_{j|i} = a_j, \text{ then } p_{.j} = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} p_{i.}a_j = a_j$$

$$\Longrightarrow p_{ij} = p_i p_{j|i} = p_{i.}p_{.j}$$

Independence of Rows and Columns

- Can also assess association through the odds ratio:

$$\begin{aligned}\theta &= \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}} \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)} \\ &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)}\end{aligned}$$

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

Test for Independence: Odds Ratio

- X and Y are independent $\iff \theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1$
- MLE: $\hat{\theta} = \frac{y_{11}/y_{12}}{y_{12}/y_{22}} = \frac{y_{11}y_{22}}{y_{12}y_{21}}$
  - When some $y_{ij} = 0$, $\hat{\theta}$ is not a good estimator. Improved by adding 0.5 to each cell count:

$$\tilde{\theta} = \frac{(y_{11} + 0.5)(y_{22} + 0.5)}{(y_{12} + 0.5)(y_{21} + 0.5)}$$

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

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

Test for Independence: Odds Ratio

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

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

- Large-sample CI for θ : $[e^L, e^U]$
  - Usually too wide
  - Could consider delta method or bootstrapping

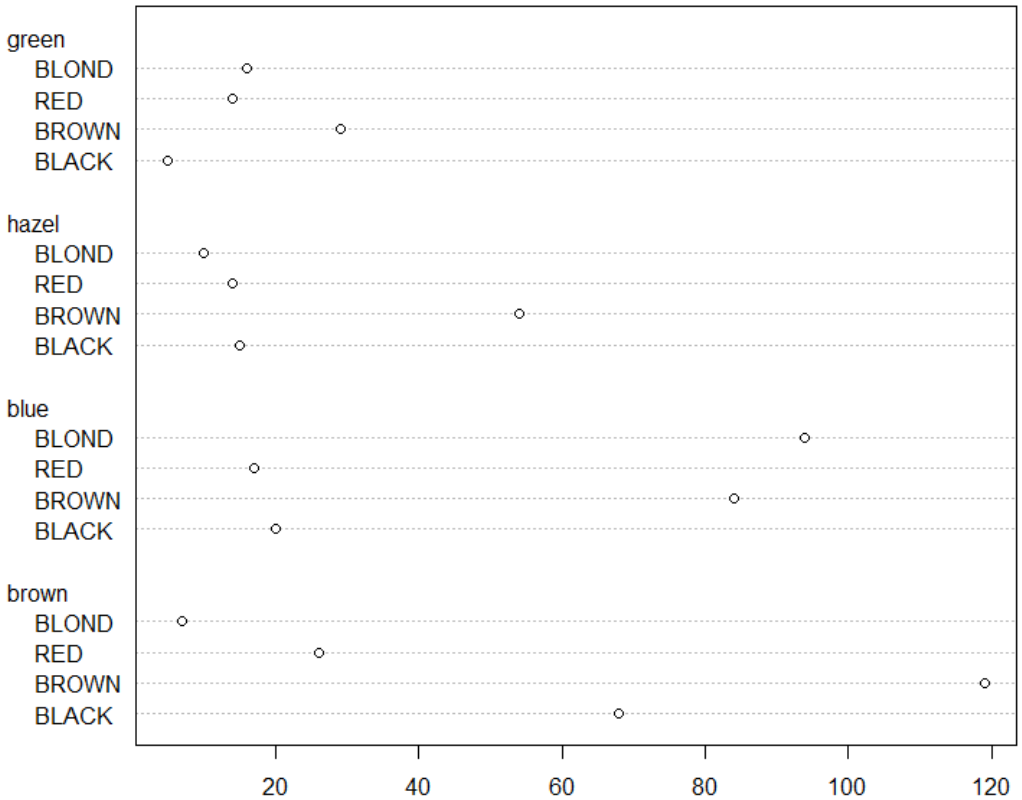
Example #2

- Studying the relationship between eye and hair color
  - 4×4 contingency table

```
> ct = xtabs(y ~ hair+eye,haireye)
> ct
       eye
hair    green hazel blue brown
  BLACK     5    15   20    68
  BROWN    29    54   84   119
  RED      14    14   17    26
  BLOND    16    10   94     7
> summary(ct)
Call: xtabs(formula = y ~ hair + eye, data = haireye)
Number of cases in table: 592
Number of factors: 2
Test for independence of all factors:
Chisq = 138.29, df = 9, p-value = 2.325e-25
```

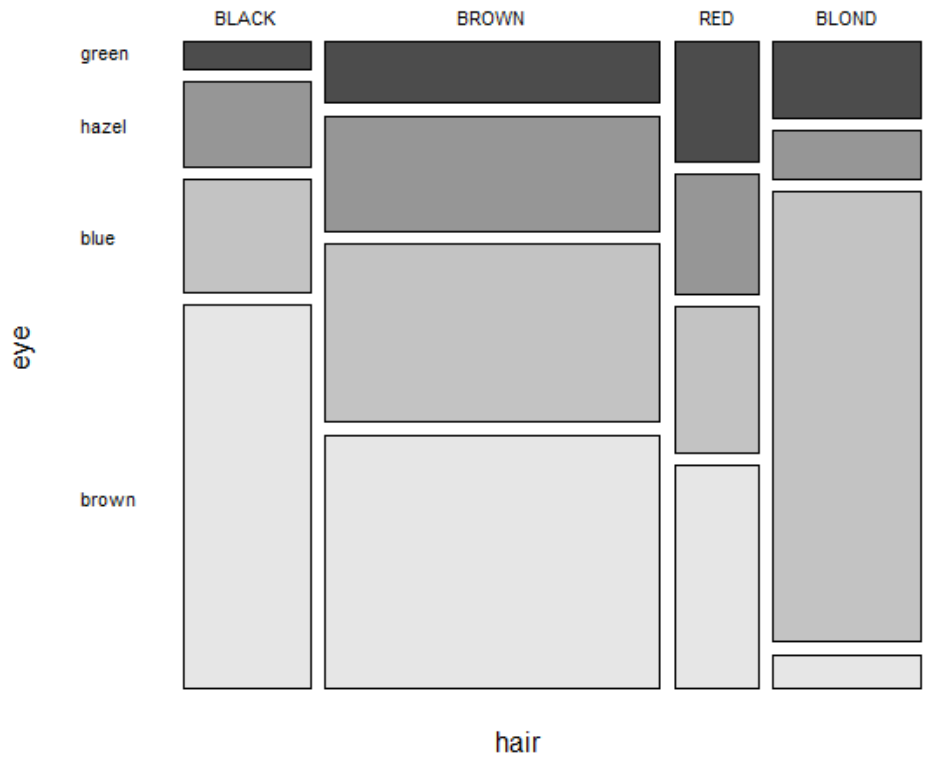
Data visualization #1

```
> dotchart(ct)
```



Data visualization #2

```
> mosaicplot(ct,color=TRUE,main=NULL,las=1)
```



Fitting Poisson Model

```
> modp = glm(y ~ hair+eye, family=poisson,haireye)
> summary(modp)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4575     0.1523  16.136  < 2e-16 ***
hairBROWN     0.9739     0.1129   8.623  < 2e-16 ***
hairRED      -0.4195     0.1528  -2.745  0.00604 **
hairBLOND     0.1621     0.1309   1.238  0.21569
eyehazel      0.3737     0.1624   2.301  0.02139 *
eyeblue       1.2118     0.1424   8.510  < 2e-16 ***
eyebrown      1.2347     0.1420   8.694  < 2e-16 ***
---
Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 453.31  on 15  degrees of freedom
Residual deviance: 146.44  on  9  degrees of freedom   ****Clearly poor fit
AIC: 241.04
```

Further Comparisons

- Both Pearson and deviance tests suggest lack of fit assuming the independence model
- Adding interaction to Poisson model results in saturated model
- Can look at subsets of columns/rows to assess differences
  - Based on mosaic plot, are conditional distributions of eye color different in those with black or brown hair? If yes, is this primarily due to different proportions of brown eyes?
- Faraway discusses use of correspondence analysis to assess relationship

Comparing Multinomial Proportions

- Might also compare proportions across conditional dists
  - Compare prop brown eyes in black and brown hair groups
- $H_0: \ p_{j|1} = p_{j|2}$ vs $H_a: \ p_{j|1} \neq p_{j|2}$
- ML estimate of the difference:
  - $\hat{p}_{j|1} - \hat{p}_{j|2} = \frac{y_{1j}}{y_{1.}} - \frac{y_{2j}}{y_{2.}}$
  - $SD(\hat{p}_{j|1} - \hat{p}_{j|2}) = \left[\frac{p_{j|1}(1-p_{j|1})}{y_{1.}} + \frac{p_{j|2}(1-p_{j|2})}{y_{2.}}\right]^{1/2}$
- Wald Confidence Interval:
  - Replace p with $\hat{p}$ to obtain the SE
  - $\hat{p}_{j|1} - \hat{p}_{j|2} \pm z_{\alpha/2}\mathrm{SE}(\hat{p}_{j|1} - \hat{p}_{j|2})$
- Usually too narrow
- Better methods (e.g., delta method) exist

Matched Pairs

- So far, we've compared cases where two categorical variables were measured on the same object
- Can be cases where one categorical variable is measured on matched objects
  - Repeat measurements on same subject (pre / post)
  - Activity scored by two reviewers (interrater reliability)
- Similar in design to a paired t test
- Expect there to be strong evidence of association
- Interested more in symmetry ($H_0 : p_{ij} = p_{ji}$)

Example: Agresti 10.1

- Approval of the President's performance, one month apart, for the same sample of Americans
- X =original view and Y = view one month later
- Did events in the month alter approval rating?

	Approve	Disapprove
Approve	794	150
Disapprove	86	570

- Want to test marginal prob homogeneity ($H_0 : p_{1.} = p_{.1}$)
- Equivalent to testing table symmetry

$$\begin{aligned} p_{1.} - p_{.1} &= (p_{11} + p_{12}) - (p_{11} + p_{21}) \\ &= p_{12} - p_{21} \end{aligned}$$

Example: Agresti 10.1

- Observed approval ratings:

$$\hat{p}_{1.} = \frac{794 + 150}{1600} = 0.59 \qquad \hat{p}_{.1} = \frac{794 + 86}{1600} = 0.55$$

- Is there a drop in approval?
- Estimates use some of the same data $\rightarrow$ not independent
- In 2×2 case, test simplifies to binomial test $H_0 : p = 0.5$
  - Observe $y_{12} = 150$ and $y_{21} = 86$
  - Under H_0, for those who differ across months, should be equal chance to fall in (1,2) or (2,1) cell

Large-sample test and CI

- Confidence interval
  - Let $\hat{\delta} = \hat{p}_{1.} - \hat{p}_{.1} = (y_{12} - y_{21})/1600$

$$\begin{aligned} \mathrm{Var}(\hat{\delta}) &= \left[p_{1.}(1-p_{1.}) + p_{.1}(1-p_{.1}) - 2(p_{11}p_{22} - p_{12}p_{21})\right]/n \\ &= \left[p_{12}(1-p_{12}) + p_{21}(1-p_{21}) + 2p_{12}p_{21}\right]/n \\ &= \left[(p_{12}+p_{21}) - (p_{12}-p_{21})^2\right]/n \end{aligned}$$

  - Smaller variance than in indep samples (more efficient design)
  - Plug in estimates of $\hat{p}_{12}$ and $\hat{p}_{21}$ to get $\mathrm{SE}(\hat{\delta})$
  - CI: $\hat{\delta} \pm z_{\alpha/2}\mathrm{SE}(\hat{\delta})$
- Wald Test : $z = \frac{\hat{\delta}}{\mathrm{SE}(\delta)}$
- Score Test : $z = \frac{y_{21}-y_{12}}{(y_{21}+y_{12})^{1/2}}$
  - Also known as McNemar's test
- Inference depends on the off-diagonal counts

Using R - Binomial Test

- Need to calculate $2P(Y_{12} \geq 150 | m = 236, p = 0.5)$
- Will use Normal approximation and binomial dist

```
#### Normal approx
> zstat = (150 - 236*.5)/sqrt(236*.25)
> pz = 2*pnorm(zstat,lower=F)
> pz
[1] 3.099293e-05
#### Using binomial distribution
> pb = 2*pbinom(149,236,p=0.5,lower=FALSE)
> pb
[1] 3.715936e-05
#### Using McNemar's function
> mcnemar.test(ct)
McNemar's Chi-squared test with continuity correction
data:  ct
McNemar's chi-squared = 16.818, df = 1, p-value = 4.115e-05
```

- Strong evidence that the approval rating has dropped

Using R - Fitting a GLM

- Need to create a factor that denotes different probabilities under the symmetric table hypothesis

```
> symfac = c("AA","AB","AB","BB")
> mods = glm(y ~ symfac,family=poisson)
> summary(mods)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.67708    0.03549 188.147  < 2e-16 ***
symfacAB    -1.90640    0.07414 -25.714  < 2e-16 ***
symfacBB    -0.33145    0.05490  -6.037 1.57e-09 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 933.886  on 3  degrees of freedom
Residual deviance:  17.575  on 1  degrees of freedom
AIC: 53.418
```

- A deviance of 17.575 on 1 df strongly suggests lack of fit
- Pearson $X^2 = 17.36$ (equals zstat^2 from previous slide)

Cochran Mantel Haenszel Test

- The data can be presented in format where each subject contributes a 2×2 table

		Response	
Subject	Month	Approve	Disapprove
1	1	1	0
1	2	1	0
2	1	0	1
2	2	0	1
3	1	0	1
3	2	1	0

- Test of conditional independence in $2 \times 2 \times n$ table is a test of marginal homogeneity
- Cochran Mantel Haenszel test equivalent to McNemar's

Matched Pairs II

- What about a matched pairs design when the categorical response has more than two levels?
- Homogeneity of marginals and symmetry not equivalent
- Could apply generalized McNemar's test
- More easily assess using GLMs
- Example: Unaided distance vision
  - A total of 7477
  - Each eye graded for distance vision
  - Is the joint distribution of grades symmetric?

Eye Grade Study Data

```
> ct = xtabs(y~right+left, eyegrade)
> ct
        left
right    best second third worst
  best   1520    266   124    66
  second  234   1512   432    78
  third   117    362  1772   205
  worst    36     82   179   492

###Create a symmetric table factor (10 levels)
> symfac = c("BB","BS","BT","BW",
+            "BS","SS","ST","SW",
+            "BT","ST","TT","TW",
+            "BW","SW","TW","WW")
```

Eye Grade Study GLM

```
> mods = glm(y~symfac,family=poisson,eyegrade)
> summary(mods)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.326466   0.025649 285.638  < 2e-16 ***
symfacBS    -1.805005   0.051555 -35.011  < 2e-16 ***
symfacBT    -2.534816   0.069334 -36.559  < 2e-16 ***
symfacBW    -3.394640   0.102283 -33.189  < 2e-16 ***
symfacSS    -0.005277   0.036322  -0.145    0.884
symfacST    -1.342529   0.043787 -30.660  < 2e-16 ***
symfacSW    -2.944439   0.083114 -35.427  < 2e-16 ***
symfacTT     0.153399   0.034960   4.388 1.15e-05 ***
symfacTW    -2.068970   0.057114 -36.225  < 2e-16 ***
symfacWW    -1.127987   0.051869 -21.747  < 2e-16 ***
---

    Null deviance: 8692.334  on 15  degrees of freedom
Residual deviance:   19.249  on  6  degrees of freedom
AIC: 156.63
```

GLM Results

- Deviance is 19.249 on 6 degrees of freedom

```
> pchisq(19.249,6,lower=F)
[1] 0.003763139
```

- Given $P = 0.0038$, we conclude the table is not symmetric
- Are marginal dists of right and left eyes different?

```
 > margin.table(ct,1)
right
  best second  third  worst
  1976   2256   2456    789
> margin.table(ct,2)
left
  best second  third  worst
  1907   2222   2507    841
```

- Is this the reason for our lack of symmetry?

Quasi-Symmetry Model

- Suppose we set $p_{ij} = \alpha_i \beta_j \gamma_{ij}$ with $\gamma_{ij} = \gamma_{ji}$
- The parameters α and β describe marginal dists
- Under this specification

$$\eta_{ij} = \log np_{ij} = \log n + \log \alpha_i + \log \beta_j + \log \gamma_{ij}$$

```
> mods1 = glm(y~right+left+symfac,family=poisson,eyegrade)
> summary(mods1)
```

- Overparameterized as specified in R
- Let R handle this but be wary that this has a bearing on parameter interpretation

Quasi-Symmetry Model Output

```
Coefficients: (3 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.32647    0.02565 285.638  < 2e-16 ***
rightsecond -2.43955    0.09055 -26.942  < 2e-16 ***
rightthird  -1.61523    0.06955 -23.223  < 2e-16 ***
rightworst  -0.72288    0.05641 -12.816  < 2e-16 ***
leftsecond  -2.33241    0.09149 -25.493  < 2e-16 ***
leftthird   -1.39721    0.07012 -19.927  < 2e-16 ***
leftworst   -0.40510    0.05641  -7.182 6.88e-13 ***
symfacBS     0.57954    0.09462   6.125 9.07e-10 ***
symfacBT    -1.03453    0.08633 -11.983  < 2e-16 ***
symfacBW    -2.84322    0.10266 -27.696  < 2e-16 ***
symfacSS     4.76669    0.16668  28.598  < 2e-16 ***
symfacST     2.54814    0.11038  23.085  < 2e-16 ***
symfacSW          NA         NA      NA       NA
symfacTT     3.16584    0.11415  27.734  < 2e-16 ***
symfacTW          NA         NA      NA       NA
symfacWW          NA         NA      NA       NA

    Null deviance: 8692.3336  on 15  degrees of freedom
Residual deviance:    7.2708  on  3  degrees of freedom
AIC: 150.65
```

GLM Results

- Deviance is 7.2708 on 3 degrees of freedom

```
> pchisq(7.2708,3,lower=F)
[1] 0.06374946
```

- Given $P = 0.0637$, we conclude the model fits
- Better to compare the models

```
> anova(mods,mods1,test="Chi")
Analysis of Deviance Table

Model 1: y ~ symfac
Model 2: y ~ right + left + symfac
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         6    19.2492
2         3     7.2708  3   11.978 0.007457 **
```

- We find a lack of marginal homogeneity provided quasi-symmetry holds

Quasi-Independence Model

- Many women had similar results for left and right eye
- Are responses independent for the off-diagonal counts?
- Focus now is just on off-diagonal counts (like McNemar's)
- Known as the quasi-independence model

```
> mods2 = glm(y~right+left,family=poisson,
+             subset= -c(1,6,11,16), eyegrade)
> summary(mods2)
```

Quasi-Independence Model Output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.38841    0.07454  58.870  < 2e-16 ***
rightsecond  0.78176    0.06409  12.197  < 2e-16 ***
rightthird   0.68723    0.06470  10.622  < 2e-16 ***
rightworst  -0.45698    0.07552  -6.051 1.44e-09 ***
leftsecond   0.88999    0.06789  13.110  < 2e-16 ***
leftthird    0.87160    0.06689  13.030  < 2e-16 ***
leftworst   -0.17689    0.07481  -2.364   0.0181 *

    Null deviance: 900.99  on 11  degrees of freedom
Residual deviance: 199.11  on  5  degrees of freedom
AIC: 294.81
```

- Clearly a poor fit
- We can see a majority of diffs between left and right are mostly one grade

Interrater Reliability

- Often pairing the result of two raters each scoring a set of objects, such as images or papers
- Scoring may involve two or more levels
- Observed agreements, $OA = \sum_i y_{ii}$, includes possible chance agreements
- Cohen's Kappa introduced to adjust for this possibility
- Expected agreements based on indep, $EA = \sum_i y_{.i} y_{.i}/n$

$$\kappa = \frac{OA - EA}{n - EA}$$

- Is a form of correlation $\rightarrow -1 \leq \kappa \leq 1$
- When more than two ordered categories, can "score" disagreements (i.e., -1 and -2 more similar than -1 and 2)

Ordered Categories

- Ordered categories provide more info
- Often will treat one variable as response
- Will discuss models with a response later
- Here, will assign scores to the categories
  - Rows: $(u_1 \leq, \ldots \leq u_I)$
  - Cols: $(v_1 \leq, \ldots \leq v_j)$
  - $H_0: \ r = cor(u, v) = 0$ vs $H_a: \ r = cor(u, v) \neq 0$
- Choice of scores requires some judgement
- May consider different sets to assess robustness

Linear-by-linear Association Model

- Study the linear association in u and v

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

  - $\bar{u} = \sum_{i=1}^{I}\sum_{j=1}^{J} u_i \; y_{ij}/n = \sum_{i=1}^{I} u_i \; y_{i.}/n$

    $\bar{v} = \sum_{i=1}^{I}\sum_{j=1}^{J} v_j \; y_{ij}/n = \sum_{j=1}^{J} v_j \; y_{.j}/n$
  - $M^2 = (n-1)r^2 \overset{H_0}{\sim} \chi^2_1$
  - In R: lbl_test function in coin package

Case Study: Ordered Categories

```
> ct = xtabs( ~ PID+educ,nes96)
> ct
          educ
PID        MS HSdrop HS Coll CCdeg BAdeg MAdeg
  strDem    5     19 59   38    17    40    22
  weakDem   4     10 49   36    17    41    23
  indDem    1      4 28   15    13    27    20
  indind    0      3 12    9     3     6     4
  indRep    2      7 23   16     8    22    16
  weakRep   0      5 35   40    15    38    17
  strRep    1      4 42   33    17    53    25
```

- Will consider scores using 1 through 7 for each variable

Case Study: Independence : Nominal

```
> summary(ct)
Call: xtabs(formula = ~PID + educ, data = nes96)
Number of cases in table: 944
Number of factors: 2
Test for independence of all factors:
Chisq = 38.8, df = 36, p-value = 0.3446
Chi-squared approximation may be incorrect

> groupct = as.data.frame.table(ct)
> modi = glm(Freq~PID+educ,family=poisson,groupct)
> summary(modi)
> pchisq(deviance(modi),df.residual(modi),lower=F)
[1] 0.2696086
```

- Nominal model: No evidence against independence

Case Study: Independence : Scored

```
> library(coin)
> lbl_test(as.table(ct))

Asymptotic Linear-by-Linear Association Test

data:  educ (ordered) by
 PID (strDem < weakDem < indDem < indind < indRep < weakRep < strRep)
Z = 3.1822, p-value = 0.001461
alternative hypothesis: two.sided

> #####-----manually------#####
> u <- as.vector(scale(1:7, center=sum(c(1:7)*ct)/sum(ct),scale=FALSE))
> v <- as.vector(scale(1:7, center=sum(t(ct)*c(1:7))/sum(ct),scale=FALSE))
> r <- sum(u%*%t(v)*ct) / sqrt(sum(u^2*ct) * sum(t(ct) * v^2))
> M2 <- (sum(ct) - 1) * r^2
> 1-pchisq(M2, 1, lower=TRUE)
[1] 0.001461437
```

- Scored model: Strong evidence against independence

Linear-by-linear Association Model

- Consider modeling the p_{ij} as

$$\eta_{ij} = \log np_{ij} = \log n + \alpha_i + \beta_j + \gamma u_i v_j$$

- When $\gamma = 0$, we have independence
- Similar test to previous one

```
groupct$uscore = unclass(groupct$PID)
groupct$vscore = unclass(groupct$educ)

modis = glm(Freq~PID+educ+I(uscore*vscore),family=poisson,groupct)
anova(modi,modis,test="Chi")
```

GLM Results

```
> summary(modis)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.245874   0.290541   4.288 1.80e-05 ***
PIDweakDem         -0.231690   0.109657  -2.113 0.034613 *
PIDindDem          -0.870935   0.142609  -6.107 1.01e-09 ***
PIDindind          -2.072665   0.214863  -9.646  < 2e-16 ***
PIDindRep          -1.272907   0.203997  -6.240 4.38e-10 ***
PIDweakRep         -0.940284   0.231496  -4.062 4.87e-05 ***
PIDstrRep          -0.922944   0.270247  -3.415 0.000637 ***
educHSdrop          1.288644   0.311257   4.140 3.47e-05 ***
educHS              2.749103   0.290065   9.478  < 2e-16 ***
educColl            2.360892   0.300152   7.866 3.67e-15 ***
educCCdeg           1.519473   0.321228   4.730 2.24e-06 ***
educBAdeg           2.330228   0.327217   7.121 1.07e-12 ***
educMAdeg           1.630806   0.353532   4.613 3.97e-06 ***
I(uscore * vscore)  0.028745   0.009062   3.172 0.001513 **

    Null deviance: 626.798  on 48  degrees of freedom
Residual deviance:  30.568  on 35  degrees of freedom
AIC: 268.26
```

GLM Results

```
> anova(modi,modis,test="Chi")
Analysis of Deviance Table

Model 1: Freq ~ PID + educ
Model 2: Freq ~ PID + educ + I(uscore * vscore)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        36     40.743
2        35     30.568  1   10.175 0.001424 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Similar result as presented earlier
- $\hat{\gamma}$ represents log-odds ratio when scores equally-spaced
- Can apply same ideas to nominal by ordinal table

Column-Effect Model

- This model also called an 'ordinal-by-nominal' model
  - Only one of the variables is scored
- For this example, residual analysis suggests the linear-by-linear association model does not explain all structure in the data
  - Perhaps relationship between party affiliation and education level is not monotone
- Can assess this by treating one of the two variables as nominal
- Let's treat education as nominal, but party preference as ordinal with equally-space scores

Column-Effect Model

- Poisson GLM model:

$$\begin{aligned} \log E(Y_{ij}) &= \log \lambda_{ij} \\ &= \log n + \alpha_i + \beta_j + \gamma_j u_i \end{aligned}$$

- γ_j - separate parameter for each column
- Can assess $\hat{\gamma}_j$ to see if trend is monotone
- Examination of output on following slides reveal this is not the case
- Faraway discusses altering the scores to better fit the data

GLM Model Fit

```
> fit7<-glm(Freq~PID+educ+educ:uscore,groupct,family=poisson)

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.945679   0.471322   4.128 3.66e-05 ***
PIDweakDem        -0.075007   0.109326  -0.686 0.492661
PIDindDem         -0.560399   0.140521  -3.988 6.66e-05 ***
PIDindind         -1.610437   0.210250  -7.660 1.86e-14 ***
PIDindRep         -0.660584   0.192522  -3.431 0.000601 ***
PIDweakRep        -0.178992   0.211862  -0.845 0.398193
PIDstrRep         -0.013421   0.241404  -0.056 0.955663
educHSdrop         1.061194   0.527977   2.010 0.044439 *
educHS             2.161235   0.484489   4.461 8.16e-06 ***
educColl           1.650803   0.491805   3.357 0.000789 ***
educCCdeg          0.971275   0.513874   1.890 0.058744 .
educBAdeg          1.722897   0.489151   3.522 0.000428 ***
educMAdeg          1.281529   0.501813   2.554 0.010655 *
```

GLM Model Fit

```
educMS:uscore      -0.312217   0.154051  -2.027 0.042692 *
educHSdrop:uscore -0.194451   0.077228  -2.518 0.011806 *
educHS:uscore      -0.055347   0.048196  -1.148 0.250810
educColl:uscore     0.004460   0.050603   0.088 0.929760
educCCdeg:uscore   -0.008699   0.060667  -0.143 0.885978
educBAdeg:uscore    0.034554   0.048782   0.708 0.478740
educMAdeg:uscore          NA         NA      NA       NA

    Null deviance: 626.798  on 48  degrees of freedom
Residual deviance:  22.761  on 30  degrees of freedom
AIC: 270.46
```

- Not a monotone pattern
- First two are the only two signficantly different from 0