Harvard Department of Government 2003
Faraway Chapter 6, Contingency Tables

JEFF GILL
Visiting Professor, Fall 2024

Contingency Tables

- Used to show cross-classified categorical data on multiple variables.
- Recall the definitions of *nominal* and *ordinal*.
- Also applies to *interval* data that has been discretized.
- Simple approach: χ^2 testing from basic stats.
- As we will see some of the variants can get more mathematically and conceptually involved.

What Are Degrees of Freedom?

- The 1990 Election to the Ohio State House, Precinct 1 of 131, District 42.

	Vote	Abstain	
Black	?	?	221
White	?	?	484
	222	483	705

Cross Tabulation

- The table below shows occupational background and legal history for current United States Senators.
- Test for a relationship between attorney as a previous occupation and whether or not the Senator has been indicted for accepting illegal campaign contributions.

	Attorney	Not Attorney
Been Indicted	10	1
Never Indicted	48	41

- Using $\alpha = 0.05$, $df = 1$, the χ^2 critical value is 3.84:

```
qchisq(p=0.05,df=1,lower.tail=FALSE)
[1] 3.8415
```

Illustration of NHST

- Senate example. . .

```
senate <- data.frame("freq"=c(10,1,48,41),
                     "indicted"=c("yes","yes","no","no"),
                     "attorney"=c("yes","no","yes","no"))
summary(xtabs(freq ~ indicted + attorney, data=senate))

Call: xtabs(formula = freq ~ indicted + attorney, data = senate)
Number of cases in table: 100
Number of factors: 2
Test for independence of all factors:
        Chisq = 5.495, df = 1, p-value = 0.01907
        Chi-squared approximation may be incorrect
```

Example from Johnson and Reynolds (page 350)

	Dependent Variable: Race	
Opinion	White	Nonwhite
Keep troups in Iraq	367	56
Bring troops home	287	182

1. The direction of the test is along the "a-d" or main diagonal:
   - H_1: Whites are more likely to support continuation than Nonwhites.
   - H_0: There is no difference by race.

Example from Johnson and Reynolds (page 350)

2. The test statistic is:

$$X^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$= \frac{892(66794 - 16072)}{(423)(469)(654)(238)}$$

$$= 74.31726$$

   which is distributed χ^2_1 under H_0 .

3. H_1 is supported for "big" values of X^2 , H_0 for typical values under the χ^2_1 distribution.

4. Choose $\alpha = 0.95$.

5. The critical value is 3.84 (from `R` or tables in the back of some books).

6. Since $74 > 3.84$ we reject H_0 and assert that there is these are not independent (not a proof!).

7. Keep in mind that the NHST is widely misunderstood and mis-applied.

State Legislators Example

```
soule.freq <- c(76,24,90,10)
soule.df <- data.frame(soule.freq,
        expand.grid(Willing.to.Remain=1:2,Future.Ambitions=1:2))
soule.df$Willing.to.Remain <- factor(soule.df$Willing.to.Remain)
levels(soule.df$Willing.to.Remain) <- c("Probably","Not")
soule.df$Future.Ambitions <- factor(soule.df$Future.Ambitions)
levels(soule.df$Future.Ambitions) <- c("Progressive","Not")
soule.df

  soule.freq Willing.to.Remain Future.Ambitions
1         76          Probably      Progressive
2         24               Not      Progressive
3         90          Probably              Not
4         10               Not              Not
```

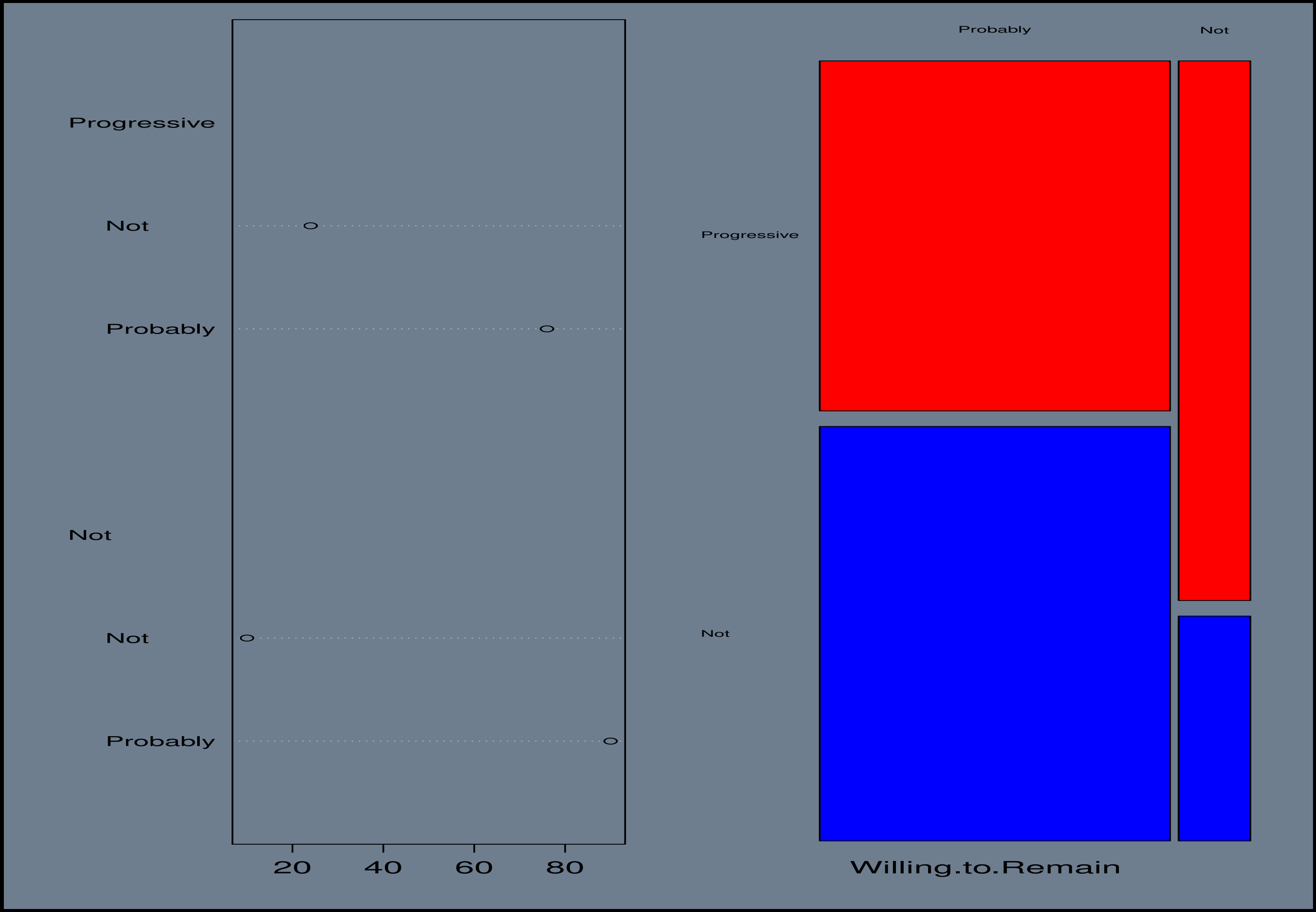
State Legislators Example

```
( leg <- xtabs(soule.freq ~ Willing.to.Remain + Future.Ambitions, soule.df) )

                 Future.Ambitions
Willing.to.Remain Progressive Not
         Probably          76  90
         Not               24  10

summary(xtabs(soule.freq ~ Willing.to.Remain + Future.Ambitions, soule.df))

Call: xtabs(formula = soule.freq ~ Willing.to.Remain + Future.Ambitions,
    data = soule.df)
Number of cases in table: 200
Number of factors: 2
Test for independence of all factors: Chisq = 6.945, df = 1, p-value = 0.008403
```

State Legislators Example

```
postscript("Class.MLE/Images/soule.ps")
par(mfrow=c(1,2),mar=c(4,1,1,1),bg="slategrey",cex.lab=1.25,cex.axis=1.25)
dotchart(leg)
mosaicplot(leg,color=c(498,132),main=NULL,las=1)
dev.off()
```

State Legislators Example



The Death Penalty In Florida

- Data on 326 defendents in homicide indictments in 20 Florida counties during 1976-77.
- Source: Radelet M. (1981) "Racial characteristics and the imposition of the death penalty." *American Sociological Review* 46, 918-927.
- Data were collected in 20 of Florida's 67 counties on all indictments for Murder I, II, and III that occurred in 1976 and 1977.
- The counties were selected with the probability of inclusion of each county in the sample proportional to its population size.
- Format: A data frame with 8 observations on the following 4 variables.

y	a numeric vector
penalty	Did the subject recieve the death penalty? no or yes
victim	Was the victim black or white?
defend	Was the defendent black or white?

The Death Penalty In Florida

Table 1. Relationship and Racial Characteristics of Victims and Defendants for All Homicide Indictments

	Number of Cases	First Degree Indictments	Probability of First Degree Indictment	Sentenced to Death	Probability of Death Penalty (all cases)	Probability of Death Penalty (first degree indictments)
Nonprimary						
White victim						
Black defendant	63	58	.921	11	.175	.190
White defendant	151	124	.821	19	.126	.153
Black victim						
Black defendant	103	56	.544	6	.058	.107
White defendant	9	4	.444	0	.000	.000
Primary						
White victim						
Black defendant	3	1	.333	0	.000	.000
White defendant	134	73	.545	3	.022	.041
Black victim						
Black defendant	166	51	.307	0	.000	.000
White defendant	8	4	.500	0	.000	.000
N	637	371	.582	39	.061	.105
Nonprimary						
Black defendant	166	114	.687	17	.102	.149
White defendant	160	128	.800	19	.119	.148
White victim	214	182	.850	30	.140	.165
Black victim	112	60	.536	6	.054	.100

The Death Penalty In Florida

```
library(faraway);  data(death);  death
    y penalty victim defend
1  19     yes      w      w
2 132      no      w      w
3   0     yes      b      w
4   9      no      b      w
5  11     yes      w      b
6  52      no      w      b
7   6     yes      b      b
8  97      no      b      b

xtabs(y ~ defend + penalty, data=death)
      penalty
defend  no yes
     b 149  17
     w 141  19

summary(xtabs(y ~ defend + penalty, data=death))
```

The Death Penalty In Florida

```
library(faraway);  data(death);  death
    y penalty victim defend
1  19     yes      w      w
2 132      no      w      w
3   0     yes      b      w
4   9      no      b      w
5  11     yes      w      b
6  52      no      w      b
7   6     yes      b      b
8  97      no      b      b

xtabs(y ~ defend + penalty, data=death)
      penalty
defend  no yes
     b 149  17
     w 141  19

summary(xtabs(y ~ defend + penalty, data=death))

        Chisq = 0.22145, df = 1, p-value = 0.638
```

The Death Penalty In Florida

```
xtabs(y ~ victim + penalty, data=death)

victim  no yes
     b 106   6
     w 184  30

summary(xtabs(y ~ victim + penalty, data=death))
```

The Death Penalty In Florida

```
xtabs(y ~ victim + penalty, data=death)

victim  no yes
     b 106   6
     w 184  30

summary(xtabs(y ~ victim + penalty, data=death))

        Chisq = 5.615, df = 1, p-value = 0.01781
```

The Death Penalty In Florida

```
xtabs(y ~ defend + victim, data=death)

      victim
defend   b   w
     b 103  63
     w   9 151

summary(xtabs(y ~ defend + victim, data=death))
```

The Death Penalty In Florida

```
xtabs(y ~ defend + victim, data=death)

      victim
defend   b   w
     b 103  63
     w   9 151

summary(xtabs(y ~ defend + victim, data=death))

        Chisq = 115.01, df = 1, p-value = 7.837e-27
```

The Death Penalty In Florida

```
ftable(xtabs(y ~ victim + defend + penalty, data = death))

              penalty  no yes
victim defend
b      b                97   6
       w                 9   0
w      b                52  11
       w               132  19

summary(xtabs(y ~ victim + defend + penalty, data = death))
```

The Death Penalty In Florida

```
ftable(xtabs(y ~ victim + defend + penalty, data = death))

              penalty  no yes
victim defend
b      b               97   6
       w                9   0
w      b               52  11
       w              132  19

summary(xtabs(y ~ victim + defend + penalty, data = death))

        Chisq = 122.4, df = 4, p-value = 1.642e-25
```

Case Study: Spending in Congress

- Fiscal behavior for House members over three periods: 409 roll call votes in the 103^{rd} Congress from January 1, 1993 to September 31, 1994, 573 in the 104^{th} Congress from January 5, 1995 to April 16, 1996 (until the fiscal year 1996 budget was finally passed, six and a half months late), and 751 in the 104^{th} Congress from January 5, 1995 to December 1, 1996 (the complete session).
- "Spending" vote: in favor of a bill or amendment that increases federal outlays
- "Saving" vote: specifically decreases federal spending (i.e. program cuts).
- The fiscal impact of each House member's vote is cross-indexed (432 in the 103^{rd}, 437 in the 104^{th}) and calculated as the total increase to the budget or the total decrease to the budget.

Partisan Differences and the Exchange of Power

- The 1995 Republican control of the House provided nothing to counter the conventional finding that party affiliation is the best predictor of how a legislator will vote (Collie 1985, Cooper and Young 1997, Rohde 1991, 1992).
- This is observed regardless of unified or divided party control of government.
- Gingrich and the leadership fostered great loyalty among the Republicans in the 104^{th} Congress through a clear agenda (the "Contract with America"), the removal of proxy voting by chairs, appointment of chairs, and the use of the centralized budget process.
- Did the majority change and the new rules alter fiscal behavior as measured by specific spending and saving proposals?

House Spending and Saving Means by Party (in millions)

104^{th} House: Through 4/16/96

	Rep	Dem	S.Dem
Total Spending	5,524	207,353	94,654
	(5,176)	(269,317)	(202,067)
Total Saving	-25,342	-7,253	-9,979
	(32,792)	(9,657)	(14,049)

103^{rd} House

	Rep	Dem	S.Dem
Total Spending	90,652	134,486	130,810
	(18,191)	(8,911)	(12,046)
Total Saving	-78,503	-76,149	-58,567
	(18,415)	(15,271)	(15,190)

104^{th} House: Through 12/1/96

	Rep	Dem	S.Dem
Total Spending	112,093	116,979	121,131
	(4,788)	(10,671)	(5,204)
Total Saving	-102,968	-61,448	-59,892
	(8,419)	(17,410)	(20,384)

Spending Findings from the Partisan Table

- The 103^{rd} House: Democrats are 49% higher on average than the Republicans.
- First two thirds of the 104^{th} House (up until the FY1996 budget agreement) was even more substantial: Democratic spending was *forty times higher* on average.
- But this changes rapidly in the remaining six months so that the mean spending difference closed to 3.4%: in mean adjusted numbers, the Republicans moved from $5.5 billion to $112 billion, whereas the Democrats moved from $207 billion to $117 billion.

Savings Findings from the Partisan Table

- The mean savings for the Democrats and Republicans in the 103^{rd} House is almost identical.
- But the southern Democrats (Southern Democrats are defined consistent with V.O. Key's (1949) classification: AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, TX, VA.) slightly lower.
- The Republicans in the 104^{th} save only 31% more by mean than the they did as the minority party.
- Southern Democrats moved closer towards the non-southern Democrats once in the minority.
- The southern Democrats are statistically distinct from the Republicans except for savings in the 103^{rd} Congress.

Interaction Between Region and Party

- A fundamental shift has moved the southern electorate increasingly toward the Republican party over the last 20 years.
- Southern Democrats voted to save less than the other Democrats in the 103^{rd} House.
- In contrast Southern Democrats were more like the other Democrats in the 104^{th} House for both time periods.
- Voting to cut fewer programs than other Democrats in the 104^{th} was a defensive strategy that enabled Southern Democrats to claim that they were protecting local interests while still being fiscally conservative.
- This suggests cross-pressures with regard to spending versus saving and region.

A House Spending and Saving Means by Region

103^{rd} House

	West	Midwest	South	Northeast
Spending Votes	114,273	114,595	114,506	121,973
	(102,737)	(26,134)	(25,146)	(21,350)
Saving Votes	-84,872	-89,706	-83,501	-76,361
	(24,697)	(22,959)	(25,282)	(25,915)

104^{th} House: Through 4/16/96

	West	Midwest	South	Northeast
Spending Votes	101,666	70,275	49,128	125,428
	(462)	(424)	(383)	(478)
Saving Votes	-16,818	-21,249	-16,916	-14,268
	(151)	(183)	(140)	(170)

104^{th} House: Through 12/1/96

	West	Midwest	South	Northeast
Spending Votes	114,536	113,671	116,250	114,770
	(11,816)	(6,502)	(5,684)	(7,448)
Saving Votes	-84,872	-89,706	-83,501	-76,361
	(24,697)	(22,959)	(25,28288)	(25,915)

A Linear Model of Interaction Effects

- Consider a model for explaining the outcome variable, X_{ij}, which is the fiscal behavior in region i for party j in terms of spending or saving votes:

$$X_{ij} = \mu + f(\alpha_i) + g(\beta_i) + \gamma_{ij,i'j'} + \epsilon_{ij}$$

where: μ is the national mean, α_i are the region effects ($i = 1, 2, 3, 4$), β_i are the party effects ($j = 1, 2$), $\gamma_{ij,i'j'}$ are the 6 possible interaction effects, ϵ_{ij} are the 8 residuals, and each cell is indexed through differing k.

- Diagramatically:

	West	Midwest	South	Northeast
Democrats				
Republicans				

So $a = 4$ and $b = 2$ in this case.

- The party category has been collapsed so that the southern Democrats are now returned to the Democrats as a whole to avoid multicollinearity effects that would arise if a party category were also defined by region.

A Linear Model of Interaction Effects, Assumptions

- If we treat the region effects as row variables and treat the party effects as column variables, then we can employ a fixed effects cell means model (2-way cross classification model) to test for significance of the interaction term.
- A requirement of the standard cell means model is $\epsilon_{ij} \sim i.i.d.\ n(0, \sigma^2)$ with finite σ^2.
- Defining: $f(x) = \log(x)$ (spending is right skewed), and $g(y) = (-y)^{\frac{1}{4}}$ (saving is left skewed) produces residuals that are approximately normally distributed with mean zero and no evidence of heteroscedasticity.
- The model is now restated as:

$$X_{ij} = \mu + \log(\alpha_i) + (-\beta_i^{\frac{1}{4}}) + \gamma_{ij,i'j'} + \epsilon_{ij}.$$

for $i = 1, 2, 3, 4$ and $j = 1, 2$.

A Linear Model of Interaction Effects, Assumptions

- The sensitivity of the cell means model to residual deviances from normality is primarily a function of the degree of inequality of the category variances.

- Since we know:

$$\sum_i \sum_j log(\alpha_{ij}) \cong 0, \qquad \sum_i \sum_j (-\beta_{ij})^{\frac{1}{4}} \cong 0, \qquad \sum_i \sum_j \epsilon_{ij} \xrightarrow[n \to \infty]{} 0$$

  directly from the model assumptions, then the expected value of the k^{th} observation in the i^{th} region for the j^{th} party is:

$$E(x_{ijk}) = \mu_{ij} + \gamma_{ij,i'j'}$$

- Normality of the residuals (achieved by transformation above) leads to the following well known and desirable result: the best linear unbiased estimate of a cell population mean:

$$\text{BLUE}(\mu_{ij}) = \hat{\mu}_{ij} = \overline{X}_{ij},$$

  where: $E(\hat{\mu}_{ij}) = \mu_{ij}$, and: $\text{var}(\hat{\mu}_{ij}) = \frac{\sigma^2}{n_{ij}}$.

A Linear Model of Interaction Effects, Hypotheses

- We are interested in the interaction effects from the $a = 2$ parties and $b = 4$ regions, γ_{ij}, whose sources are identified by:

$$\gamma_{ij,i'j'} = \mu_{ij} + \mu_{ij'} + \mu_{i'j} + \mu_{i'j'} \qquad where: i < i', j < j'$$

- There are $\frac{1}{4}ab(a-1)(b-1) = 6$ possible interaction effects (defined by the number of possible "quadrants" in the table), and there are a maximum of $(a-1)(b-1) = 3$ independent interaction effects:

	West	Midwest	South	Northeast
Democrats	μ_{11}	μ_{12}		
Republicans	μ_{21}	μ_{22}		

	West	Midwest	South	Northeast
Democrats	μ_{11}		μ_{13}	
Republicans	μ_{21}		μ_{23}	

A Linear Model of Interaction Effects, Hypotheses

	West	Midwest	South	Northeast
Democrats	μ_{11}			μ_{14}
Republicans	μ_{21}			μ_{24}

	West	Midwest	South	Northeast
Democrats		μ_{12}	μ_{13}	
Republicans		μ_{22}	μ_{23}	

	West	Midwest	South	Northeast
Democrats		μ_{12}		μ_{14}
Republicans		μ_{22}		μ_{24}

	West	Midwest	South	Northeast
Democrats			μ_{13}	μ_{14}
Republicans			μ_{23}	μ_{24}

A Linear Model of Interaction Effects, Hypotheses

- The hypothesis test for interaction between region and party is therefore:

$$H_0 : \gamma_{ij,i'j'} = 0 \quad \forall i, j \qquad \text{vs.} \qquad H_1 : \gamma_{ij,i'j'} \neq 0 \quad \text{for any } i, j$$

- So under the null hypothesis, $E(x_{ijk}) = \mu_{ij}$.

- The test statistic is derived from Cochran's Theorem and Slutsky's Theorem:

$$Q' = \frac{(a-1)(b-1)Q}{\hat{s}^2} \sim \chi^2_{(a-1)(b-1)}$$

where: $\hat{s}^2$ is the pooled sample variance, and

$$Q = SSE(H_0) - SSE(H_1),$$

which is the sum of squared errors assuming *no* interaction minus the sum of squared errors assuming interaction.

Results: Region and Party Interaction in the House

		Spending Votes	Saving Votes
	Q'	0.01965	0.962387
103^{th} House:	p-value	0.9992717	0.8103521
104^{th} House:	Q'	0.69523	0.11127
Through 4/16/96	p-value	0.9946	0.9905
104^{th} House:	Q'	0.010544	3.3089
Through 12/1/96	p-value	0.999713	0.3464059

Q'-statistic from cell means model, p-values from $Q' \sim \chi^2_3$.

A Linear Model of Interaction Effects, Results

- Surprisingly, the results for each Q_i' reported in the table provide no evidence, given these data, to infer an interaction between party and region for spending or saving in all three periods.

- What makes the failure to reject the null hypothesis interesting in this case is that in doing so, we are failing to find evidence of any of the six mathematically possible interaction effects (three of which are independent) given these data. Rejecting the null hypothesis and concluding that at least one of the interaction effects is significant is in some ways a weaker result.

- The southern Democrats simply do not differ enough from the other Democrats on savings behavior: 27% in the 103^{rd} House, 38% in the first 16 months of the 104^{th} House, and only 2.6% for the entire 104^{th} House.

A Linear Model of Interaction Effects, Results

- There is another contributing factor to this poor interaction term for saving: Republicans are disciplined in terms of spending behavior, but not as disciplined in terms of saving behavior in the 104^{th} House.

- This serves to weaken any potentially observable interaction effects between party and region because it diminishes the effect of fiscally conservative southern districts belonging to Republicans.

- The model reveals a problem with analyzing the southern Democrats as a distinct group: there is little evidence from these data that the southern Democrats are held to a different standard than southern Republicans by their constituents.

- In fact, there is evidence in the figure that southern Republicans are also required to be conservative with regard to spending, but not at the expense of protecting local federal largesse (i.e. fewer votes to cut).

New Running Example

- 1960 American National Election Study, cross classify married/single with think Nixon is working class versus middle or upper class:

```
y <- c(281+496,19+32,89,12)
respondent.status <- gl(n=2, k=1, length=4, labels=c("married","single"),
                        ordered=FALSE)
nixon.ses         <- gl(n=2, k=2, labels=c("upper/middle","working"),
                        ordered=FALSE)
( nixon <- data.frame(y,respondent.status,nixon.ses) )
    y respondent.status    nixon.ses
1 777           married upper/middle
2  51            single upper/middle
3  89           married      working
4  12            single      working

( ov <- xtabs(y ~ respondent.status + nixon.ses) )
                 nixon.ses
respondent.status upper/middle working
          married          777      89
          single            51      12
```

Choice #1: Poisson Model

- Tests differences in counts across the 4 cells vs. the null that all 4 occur from the same rate parameter. *Fixes no marginals or totals.*

- This additive model requires marginals with treatment contrast.

- MODEL SPECIFICATION: $\log(\mu) = \gamma + \alpha_i + \beta_j$ where:
  γ is the intercept,
  α_i is the respondent status effect ($i = 1, 2$),
  and β_j is the Nixon SES effect ($j = 1, 2$).

- And the resulting log-likelihood function is simply: $\ell(\mu) = \sum_{i=1}^{n} y_i \log \mu_i$.

- Effects are assumed to be independent since no interaction is specified (giving an interaction here produces the saturated model with no DF).

Choice #1: Poisson Model

```
model.poisson <- glm(y ~ respondent.status + nixon.ses, family=poisson(),data=ov)
summary(model.poisson)

Deviance Residuals:
      1        2        3        4
 0.1852  -0.6983  -0.5358   1.7768

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              6.64879    0.03586  185.40   <2e-16
respondent.statussingle -2.62075    0.13049  -20.08   <2e-16
nixon.sesworking        -2.10389    0.10540  -19.96   <2e-16
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1480.1764  on 3  degrees of freedom
Residual deviance:    3.9658  on 1  degrees of freedom
AIC: 34.897
```

Choice #1: Poisson Model

```
# A CHI-SQUARE TEST OF EACH PREDICTOR'S INCLUSION VALUE
drop1(model.poisson,test="Chi")

                  Df Deviance    AIC    LRT   Pr(Chi)
<none>                   3.97  34.90
respondent.status  1   831.14 860.07 827.18 < 2.2e-16
nixon.ses          1   653.00 681.93 649.03 < 2.2e-16
```

Choice #2: Multinomial Model

- Now assume that for some reason the size of the survey was capped, and *fix the total sample size* at 929 respondents.

- Then the question that this produces is what is the probability implied by the table placements, leading (of course) to a multinomial model.

- If y_{ij} is the observation in cell row i and column j and p_{ij} is the associated probability, then the likelihood and log-likelihood are given by:

$$L(\mathbf{p}) = \frac{n!}{\prod_i \prod_j y_{ij}} \prod_i \prod_j p_{ij}^{y_{ij}} \qquad \ell(\mathbf{p}) \propto \sum_i \sum_j y_{ij} \log p_{ij}.$$

- The implied test here is independence of the two explanatory variables, defined canonically as $p_{ij} = p_i p_j$.

- Using the property that the probabilities sum to one, we get the MLEs:

$$\hat{p}_i = \sum_j y_{ij}/n \qquad \hat{p}_j = \sum_i y_{ij}/n$$

Choice #2: Multinomial Model

```
attach(nixon)
# p.hat(j=1) AND p.hat(j=2), SUMS OVER ROWS
(pp <- prop.table( xtabs(y ~ respondent.status)))
# SAME AS: (777+89)/929, (51+12)/929

respondent.status
   married     single
0.93218515 0.06781485

# p.hat(i=1) AND p.hat(i=2), SUMS OVER COLUMNS
(qp <- prop.table( xtabs(y ~ nixon.ses)))
# SAME AS: (777+51)/929, (81+12)/929

nixon.ses
upper/middle      working
   0.8912809    0.1087191
```

Choice #2: Multinomial Model

```
# FITTED VALUES FROM mu_ij = n p.hat_i p.hat_j
(fv <- outer(qp,pp)*929)
              respondent.status
nixon.ses        married  single
  upper/middle 771.8493 56.1507
  working       94.1507  6.8493

# COMPARE WITH SATURATED MODEL THAT GIVES mu_ij = y_ij
# DEVIANCE TEST
2*sum(ov*log(ov/fv))
[1] 43.247
```

Choice #2: Multinomial Model

```
# PEARSON's X^2 TEST:
sum( (ov-fv)^2/fv )
[1] 42.90185

# PEARSON'S X^2 TEST WITH YATES CORRECTION (ADDS 0.5 TOWARDS ZERO FOR Y-mu)
prop.test(ov)
        2-sample test for equality of proportions with
        continuity correction

data:  ov
X-squared = 3.8008, df = 1, p-value = 0.05123
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01986031  0.19526996
sample estimates:
   prop 1    prop 2
0.8972286 0.8095238
```

Choice #3: Binomial Model

- Stricter assumption: *fix the row marginals* and run two binomials for comparison (tests for independence of effects).
- married: 777 success out of 866 trials, single: 51 success out of 63 trials.

```
(m <- matrix(y,nrow=2))
     [,1] [,2]
[1,]  777   89
[2,]   51   12

# FIT THE NULL MODEL, CHECK THE DEVIANCE
modb <- glm(m ~ 1, family=binomial)
deviance(modb)
[1] 3.965824
```

Choice #4: Hypergeometric Model

- Strictest assumption yet: *row and column marginals are fixed.*
- The hypergeometric analysis asks, given all marginals fixed, what is the *exact* probability of observing these cell values (also called Fisher's Exact Test):

$$p(\text{table}) = \frac{(y_{11}+y_{12})!(y_{11}+y_{21})!(y_{12}+y_{22})!(y_{21}+y_{22})!}{y_{11}!y_{12}!y_{21}!y_{22}!}$$

```
fisher.test(ov)       Fisher's Exact Test for Count Data
data:  ov
p-value = 0.0553
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9587894 4.0879694
sample estimates:
odds ratio
  2.052193

prod(diag(ov))/prod(diag(ov[,2:1]))   # ODDS RATIO BY HAND:
[1] 2.054197
```

Binary Outcomes In Epidemiology (Review Slide)

- Binary outcomes are often called *events*, meaning they either happened or didn't.
- Usually these are labeled 0 and 1, where the one denotes "happened."
- Sometimes the 1 is called a "success."
- These are only labels and switching the assignment never changes the construction or reliability of the statistical model.
- Tables of events have a very specific construction:

2×2 Contingency Table

	Experimental-Manipulation		
Outcome	Treatment	Control	Row Total
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Column Total	$a+c$	$b+d$	

- Hypothesized relationships are usually down the primary diagonal of the table.

Binary Outcome Terms In Epidemiology

- Associated with these tables are some important, and frequently used, terms.
- The proportion of subjects with positive outcome is:

$$p_T = \frac{a}{a+c} \qquad \text{under treatment}$$

$$p_C = \frac{b}{b+d} \qquad \text{under control.}$$

- The difference in proportions is:

$$d_{prop} = p_T - p_C,$$

  which is known as risk in prospective studies.
- Risk is an important concept in epidemiology.

Binary Outcome Terms In Epidemiology

- The absolute risk difference is: $ARD = |p_T - p_C|,$

  and if the treatment is supposed to improve health is called the absolute risk reduction (ARR).

- In clinical trials $1/ARD$ is "the number needed to treat/harm" (depending on direction of effect).

- The risk ratio, also called the relative risk, is: $RR = \dfrac{p_T}{p_C}.$

  - $RR = 1$: there is no difference in risk between the two groups.
  - $RR < 1$: the event is less likely to occur in the Treatment group than in the Control group.
  - $RR > 1$: the event is more likely to occur in the Treatment group than in the Control group.

Duration of oral contraceptive (OC) use among 195 cases and 979 controls who were never pregnant

Duration of OC use before first pregnancy (years)	Cases	Controls	Relative risk estimate*	95% confidence interval
Never used OCs	147	509	(1.0)†	
<1	16	139	1.2	0.6–2.3
1–4	11	204	0.8	0.4–1.6
≥5	13	102	1.1	0.6–2.3
Unknown	8	25		
Total	48	470	1.1	0.7–1.7

* Relative risk estimates aggregated over decade of age.
† Reference category.

Example: Deep Vein Thrombosis and the Contraceptive Pill

- This is an example of the difference between *relative risk* and *absolute risk difference*.
- Women aged 15–45 *not* on the contraceptive pill have a DVT risk of roughly: 20 per 100,000 women per year.
- Use of the pill increases this to 40 per 100,000 women per year.
- So the relative risk (RR) is 2, however the absolute risk difference (ARD) is:

$$ARD = |p_T - p_C| = \frac{40}{100,000} - \frac{20}{100,000} = \frac{20}{100,000} = 0.00002,$$

  which is an additional 2 cases in $10,000$ person-years of exposure.
- Also, pregnant women have an ARD of about 80 per 100,000 women per year.
- So RR gives a somewhat deceptive view of the change since the risks are low anyways.

Wilcoxon Signed Rank Test Setup

- Consider ordinal data as "ranks."
- Ranks are resistant to outliers since an outlying value will only ever be 1 unit away from the following value.
- Removes information about distributional shape (therefore sometimes called nonparametric).
- Group 1: $x_{11}, x_{12}, \ldots, x_{1n}$.
- Group 2: $x_{21}, x_{22}, \ldots, x_{2n}$.
- Pair these: $(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})$.
- Define the paired difference: $d_1 = x_{11} - x_{21}, d_2 = x_{12} - x_{22}, \ldots, d_n = x_{1n} - x_{2n}$.
- H_0 : $\text{median}(\delta) = 0$.
- H_A : $\text{median}(\delta) \neq 0$, (most common), H_A : $\text{median}(\delta) > 0$, or H_A : $\text{median}(\delta) < 0$.
- Assumptions: the d_i come from a symmetric distribution $\forall i$, d_i independent of $d_j, i \neq j$.

Wilcoxon Signed Rank Test Steps

- From the data obtain:
  - the absolute value of the paired differences: $|d_1|, |d_2|, \ldots, |d_n|$ and
  - the sign of the paired differences: $\text{sign}(d_1), \text{sign}(d_2), \ldots, \text{sign}(d_n)$.
- Rank the absolute differences, *discarding values equal to zero*, and averaging ties, calling these $R_1, R_2, \ldots, R_{n'}$, where the new $n' < n$ from the discards.
- Calculate:

$$T^+ = \left| \sum_{i=1}^{n'} \text{sign}(d_i) \times R_i \right| = \left| \sum_{i=1}^{n'} \text{signed rank of } d_i \right|$$

  (also labeled V and W in some texts).
- The test statistic is:

$$z = \frac{T^+ - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}},$$

  which is asymptotically standard normal.

Wilcoxon Signed Rank Test Steps Example

1	2	3	4	5	6	7
Subj.	X_A	X_B	original $X_A - X_B$	absolute $X_A - X_B$	rank of absolute $X_A - X_B$	signed rank
1	78	78	0	0	---	---
2	24	24	0	0	---	---
3	64	62	+2	2	1	+1
4	45	48	−3	3	2	−2
5	64	68	−4	4	3.5	−3.5
6	52	56	−4	4	3.5	−3.5
7	30	25	+5	5	5	+5
8	50	44	+6	6	6	+6
9	64	56	+8	8	7	+7
10	50	40	+10	10	8.5	+8.5
11	78	68	+10	10	8.5	+8.5
12	22	36	−14	14	10	−10
13	84	68	+16	16	11	+11
14	40	20	+20	20	12	+12
15	90	58	+32	32	13	+13
16	72	32	+40	40	14	+14
						W = 67.0
						N = 14

Two-Sample Tests of Center, Example in R

- Consider comparing two groups measuring average monthly bill cosponsorship:

```
A <- c(5.8, 1.0, 1.1, 2.1, 2.5, 1.1, 1.0, 1.2, 3.2, 2.7)
B <- c(1.5, 2.7, 6.6, 4.6, 1.1, 1.2, 5.7, 3.2, 1.2, 1.3)
wilcox.test(A,B,paired = TRUE,exact=FALSE)

    Wilcoxon signed rank test with continuity correction

data:  A and B
V = 18.5, p-value = 0.386
alternative hypothesis: true location shift is not equal to 0
```

- The "continuity correction" adds 0.5 in the direction of the mean to each value before the sum, to help with small sample size.

- Otherwise stipulate:

```
wilcox.test(A,B,paired = TRUE,exact=FALSE, correct = FALSE)
```

- Note: by default, an exact p-value is computed if the samples contain less than 50 finite values and there are no ties; otherwise, a normal approximation is used.

Comparing Two Paired Groups, Nominal Outcome

- Suppose we survey the exact same respondents at two points in time and ask if they want to continue current laws on sales at "gun shows" in the US.

- The analysis focuses on the discordant pairs in the table:

		Second Survey		
		Yes	No	Total
First Survey	Yes	4	11	15
	No	3	241	244
	Total	7	252	259

  since equality means that there are the same number of switchers.

- The null hypothesis is that the discordant pairs are equal, $H_0 : b = c$.

Comparing Two Paired Groups, Nominal Outcome

- The χ^2 McNemar's test statistic is given by:

$$\chi^2_{\text{McNemar}} = \frac{(b-c)^2}{b+c} = \frac{(11-3)^2}{11+3} = 4.5714,$$

  which has $df = 1$ since it comes from at 2×2 table (number of rows minus one times number of columns minus one).

- So from `pchisq(4.5714,df=1,lower.tail=FALSE)` we get 0.03251 .

- A Yate's-style correction for small samples uses: $\chi^2_{\text{McNemar}} = \dfrac{(|b-c|-1)^2}{b+c}$,

  but this is not necessary here.

- Directly in R:

```
x <- matrix(c(4,3,11,241),2,2)
mcnemar.test(x,correct=FALSE)

    McNemar's Chi-squared test

data:  x
McNemar's chi-squared = 4.5714, df = 1, p-value = 0.03251
```

Paired Samples With Nominal Outcome

- Now let's ask about the *exact* probability of getting this table or one more "extreme."
- This is a form of Fisher's Exact Test, which we will see in more detail with independent data.
- Start with the observed table:

(i)	
4	11
3	241

- We then identify 3 more tables that make the off-diagonal values more discordant:

(ii)		(iii)		(iv)	
4	12	4	13	4	14
2	241	1	241	0	241

- For the original table and the three more discordant tables, calculate the statistic:

$$p = \frac{(b+c)!}{b!c!}\left(\frac{1}{2}\right)^{b+c}.$$

Paired Samples With Nominal Outcome

- For the 4 tables, the probabilities are:

$$p(i) = \frac{14!}{11!3!}\left(\frac{1}{2}\right)^{11+3} = 0.022217$$
$$p(ii) = \frac{14!}{12!2!}\left(\frac{1}{2}\right)^{11+3} = 0.005554$$
$$p(iii) = \frac{14!}{13!1!}\left(\frac{1}{2}\right)^{11+3} = 0.000854$$
$$p(iv) = \frac{14!}{14!0!}\left(\frac{1}{2}\right)^{11+3} = 0.000061$$

- The total is 0.028686, which is 0.057372 when doubled to put all of the rejection region in one tail (rounds to 0.06).

Paired Samples With Nominal Outcome

- An approximate 95% CI is calculated by:

$$p_1 = \frac{a+b}{N}, \qquad p_2 = \frac{a+c}{N}$$

$$(p_1 - p_2) = \frac{(b-c)}{N}$$

$$SE(p_1 - p_2) = \frac{\sqrt{b + c - (b-c)^2/N}}{N}$$

$$CI_{\alpha=0.05} = [(p_1 - p_2) \pm 1.96 \times SE(p_1 - p_2)]$$