POLSCI 9590: Methods I

Nominal/Ordinal Measures of Association

Dave Armstrong

Videos

In the videos for today, we learned about:

1. Cross-tabulations
    - Make sure DV is in the rows and IV is in the columns.
    - Take column percentages.
    - Compare those column percentages within each row.
2. Statistical Tests for Independence
3. Measures of Association (i.e., Substantive Significance).
    - Slides didn't cover these much.

More about the Test.

In the test, we can set up the following set of hypotheses:

- H_0 : Variables X and Y are independent in the population.
- H_A : Variables X and Y are not independent in the population.

Notice that there is no direction for the test. Similarly, the results are not generally amenable to directional interpretation (though they could be).

- Rejecting the H_0 because of a large χ^2 statistic means that the variables in the sample are likely related in the population.
- You still have to figure out exactly *how*. More on this in a bit.

Measures of Association: Nominal Data

- ϕ (phi) is a symmetric measure for 2×2 tables based on χ^2. $\phi = \sqrt{\frac{\chi^2}{n}}$. ϕ lives in the range: $0 \leq \phi \leq 1$.
- Cramer's V is a generalization of ϕ to larger tables: $V = \sqrt{\frac{\chi^2}{n \times (min(\#\text{rows}, \#\text{columns}) - 1)}}$
- λ is the proportional reduction in error. Based on classification success.

λ

λ uses two values in its calculation. For the sake of ease, let's call the independent variable x and the dependent variable y. Further let's define $\text{mode}(.)$ as a function that finds the number of observations in the modal category of its argument.

$$E_1 = n - \text{mode}(y)$$
$$E_2 = n - \sum_{m \in \text{ x categories}} \text{mode}(y|x = m)$$
$$\lambda = \frac{E_1 - E_2}{E_1}$$

See example on the next slide...

Example: λ

	Male	Female	
Employed FT	323	228	551
Employed PT	36	83	119
Other	235	300	535
Total	594	611	1205

$$E_1 = 1205 - 551 = 654$$
$$E_2 = (594 - 323) + (611 - 300) = 582$$
$$\lambda = \frac{654 - 582}{654} = 0.11$$

Measures of Association: Ordinal Data

There are some directional measures of association, too. Some of them are functions of *concordant* and *discordant* pairs.

- **concordant** case pairs that are ranked in the same order on both variables.
- **discordant** case pairs that are ranked in a different order on both variables.

Here, generally, we would have directional hypotheses about the relationship between two variables in the population.

- There are no parameters explicitly in the hypotheses because we are not talking about something like a mean.

Calculating Concordant and Discordant Pairs

Concordant pairs are found by taking a cell and finding all other cells that have a higher ranking on both variables. The product of these two numbers is taken.

- The procedure above is done for every cell in the table and then all of those values are summed.

Discordant pairs are found by taking a cell and finding all other cells that have a smaller ranking on one variable and a larger ranking on the other. The product of these two numbers is taken.

- The procedure above is done for every cell in the table and then all of those values are summed.

TABLE 13.8 | Support for Smoking Ban in the Workplace, by Level of Education (in a Table Larger than Two by Two)

	Less than high school diploma	High school diploma, no university	At least some university training	Total
No support for ban	18(a)	10(b)	10(c)	38
Some support for ban	14(d)	12(e)	11(f)	37
Full support for ban	10(g)	14(h)	14(i)	38
Total	42	36	35	113

TABLE 13.9 | Concordant Cells of Table 13.8

Cell	# of concordant cells	# of concordant observations	Contribution to N_s
A	4 (e, f, h, i)	12 + 11 + 14 + 14 = 51	18 * 51 = 918
B	2 (f,i)	11 + 14 = 25	10 * 25 = 250
C	0		
D	2 (h,i)	14 + 14 = 28	14 * 28 = 392
E	1 (i)	14	12 * 14 = 168
F	0		
G	0		
H	0		
I	0		
			$N_s = 1{,}728$

TABLE 13.8 | Support for Smoking Ban in the Workplace, by Level of Education (in a Table Larger than Two by Two)

	Less than high school diploma	High school diploma, no university	At least some university training	Total
No support for ban	18(a)	10(b)	10(c)	38
Some support for ban	14(d)	12(e)	11(f)	37
Full support for ban	10(g)	14(h)	14(i)	38
Total	42	36	35	113

TABLE 13.10 | Discordant Cells of Table 13.8

Cell	# of discordant cells	# of discordant observations	Contribution to N_d
a	0		
b	2 (d, g)	14 + 10 = 24	10 * 24 = 240
c	4 (d, e, g, h)	14 + 12 + 10 + 14 = 50	10 * 50 = 500
d	0		
e	1 (g)	10	12 * 10 = 120
f	2 (g, h)	10 + 14 = 24	11 * 24 = 264
g	0		
h	0		
i	0		
			N_d = 1,124

Calculating γ

If we define ...

- N_{same} as the total number of concordant pairs, and
- $N_{\text{different}}$ as the total number of discordant pairs,

then

$$\gamma = \frac{N_{\text{same}} - N_{\text{different}}}{N_{\text{same}} + N_{\text{different}}}$$

Somer's d

Somer's d is an extension of γ that accounts for the number of ties_y.

- Ties are the number of observations that take the same rank on the dependent variable, but a bigger number on the independent variable.
- For each cell in the matrix, we calculate ties_{yj} as the product of the number of observations in cell j and the sum of the values of cells that have larger values on x, but the same value of y as cell j
- Then, we sum these products across all of the j cells.

$$d = \frac{N_{\text{same}} - N_{\text{different}}}{N_{\text{same}} + N_{\text{different}} + \text{ties}_y}$$

Discordant Pairs and Ties

TABLE 13.8 | Support for Smoking Ban in the Workplace, by Level of Education (in a Table Larger than Two by Two)

	Less than high school diploma	High school diploma, no university	At least some university training	Total
No support for ban	18(a)	10(b)	10(c)	38
Some support for ban	14(d)	12(e)	11(f)	37
Full support for ban	10(g)	14(h)	14(i)	38
Total	42	36	35	113

Cell	Tied Cells	ties_y
18(a)	(b,c)	$18 \times (10 + 10) = 360$
10(b)	c	$10 \times 10 = 100$
10(c)	None	0
14(d)	(e,f)	$14 \times (12 + 11) = 322$
12(e)	f	$12 \times 11 = 132$
11(f)	None	0
10(g)	(h,i)	$10 \times (14 + 14) = 280$
14(h)	i	$14 \times 14 = 196$
Total		1390

Kendall's τ_b (Tau-b)

Kendall's τ_b is an extension of Somer's d that uses not only ties on y, but ties on x as well (calculated the same way).

$$\tau_b = \frac{N_{\text{same}} - N_{\text{different}}}{\sqrt{(N_{\text{same}} + N_{\text{different}} + \text{ties}_y)(N_{\text{same}} + N_{\text{different}} + \text{ties}_x)}}$$

Spearman's ρ_s (Rho)

Spearman's ρ_s is based on the difference in ranks of observations on two variables.

$$\rho_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

TABLE 13.12 | Spearman's *rho* and the Relationship between Races Won and Endorsement Participation

	Japan	US	Canada	Russia	Venezuela	Totals
# of races won	5	2	3	6	7	
Rank	3	5	4	2	1	
# of endorsements	4	3	5	7	6	
Rank	4	5	3	1	2	
D (Rank for wins − rank for endorsements)	−1	0	1	1	−1	0
D^2	1	0	1	1	1	4

Statistical Significance

- For many of these tools, there are formulas for normal approximations (i.e., that make z or t statistics for these things).
- Permutation tests are another way to figure out whether a result is significant.
  - Mimic the process under the null hypothesis lots of different times and calculate the statistic.
  - Do this by randomly re-arranging one of the two variables involved
  - Compare your observed result to the results you get through the simulation above.

Religion and Vote

R | Python | Stata

```
library(DAMisc)
library(dplyr)
library(rio)
ces <- import("ces19.dta")
ces <- factorize(ces)
levels(ces$relig) <- c("Atheist", "Protestant", "Catholic", "Othe
tab <- xt(ces, "vote", "relig")
```

```
tab$tab
```

```
## [[1]]
##    vote/relig    Atheist Protestant   Catholic      Other        Total
##       Liberal  49% (412)  44% (312)  52% (329)  49%  (82)  48% (1,135)
##  Conservative  19% (160)  41% (287)  29% (180)  27%  (45)  29%   (672)
##           NDP  22% (190)  10%  (72)  10%  (64)  19%  (31)  15%   (357)
##         Other  10%  (85)   5%  (33)   9%  (57)   5%   (8)   8%   (183)
##         Total 100% (847) 100% (704) 100% (630) 100% (166) 100% (2,347)
```

```
tab$stats
```

```
## [[1]]
##                statistic p-value
## Chi-squared 138.526489       0
## Cramers V     0.140265       0
## Lambda        0.000000       1
```

Age Group and Relative Economic Position

R Python Stata

```
tab <- xt(ces, "educ", "agegrp", ordinal=TRUE)
tab$tab
```

```
## [[1]]
##   educ/agegrp      18-34      35-54          55+        Total
##           <HS  17%  (79)  14% (134)  20%   (278)  18%   (491)
##    HS/College  35% (160)  35% (339)  39%   (530)  37% (1,029)
##  College Grad  48% (218)  51% (489)  41%   (565)  46% (1,272)
##         Total 100% (457) 100% (962) 100% (1,373) 100% (2,792)
```

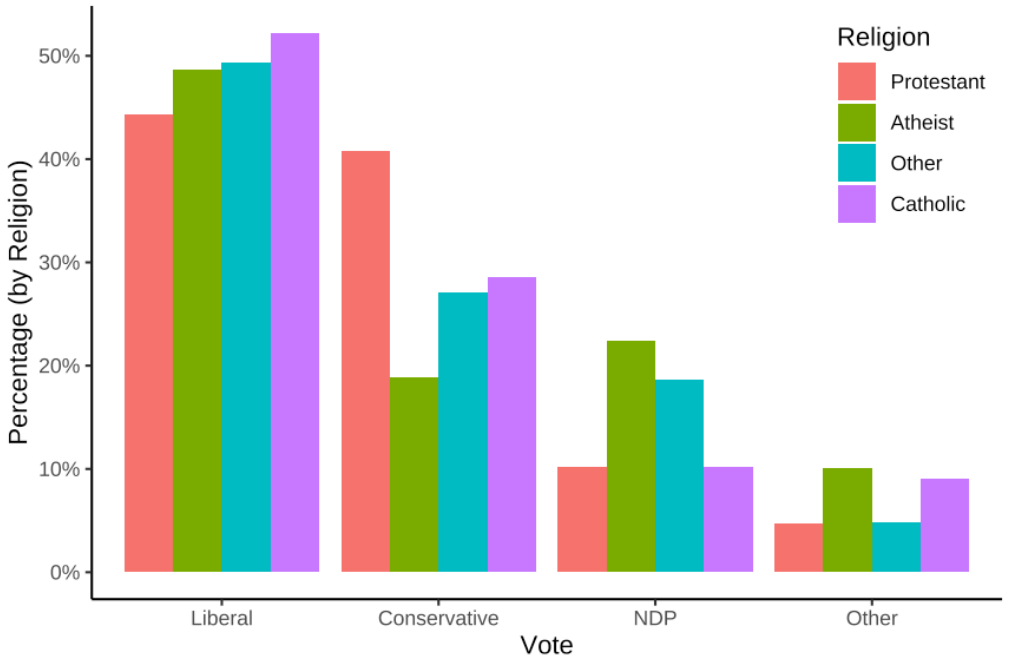
```
tab$stats
```

```
## [[1]]
##                          statistic p-value
## Chi-squared           27.29586773   0.000
## Cramers V              0.06991587   0.000
## Lambda                 0.00000000   0.005
## Kruskal-Goodman Gamma -0.11766132   0.000
## Somers D              -0.07280902   0.000
## Tau-b                 -0.07281304   0.000
```

Substantive Significance

R | Python | Stata

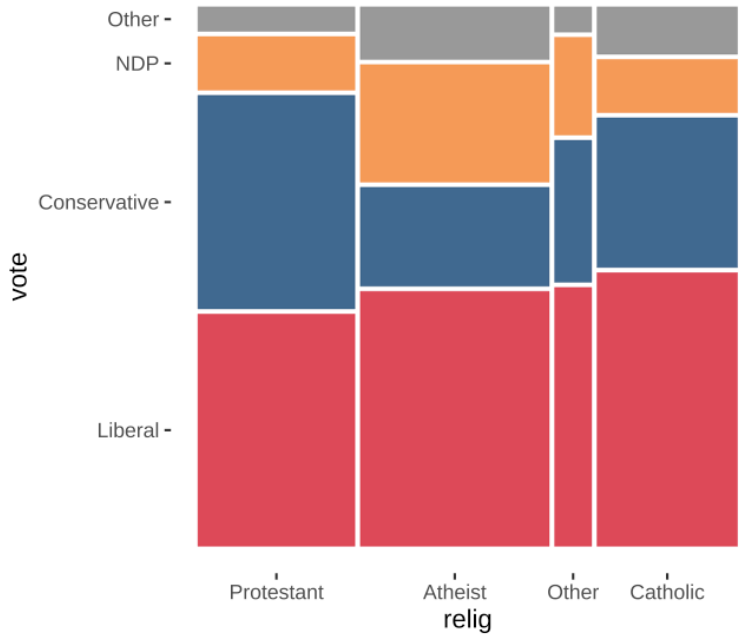
```
ces <- ces %>%
  mutate(relig = factor(as.character(relig),
    levels=c("Protestant", "Atheist",
             "Other", "Catholic")))
d <- ces %>%
  group_by(vote, relig) %>%
  tally() %>%
  na.omit() %>%
  group_by(relig) %>%
  mutate(pct = n/sum(n))

ggplot(d, aes(x=vote, y=pct, fill=relig)) +
  geom_bar(stat="identity", position="dodge") +
  theme_classic() +
  theme(legend.position = c(.9, .8)) +
  labs(x="Vote", fill="Religion",
       y="Percentage (by Religion)") +
  scale_y_continuous(labels = scales::label_percent())
```



Mosaic Plots

R Python Stata

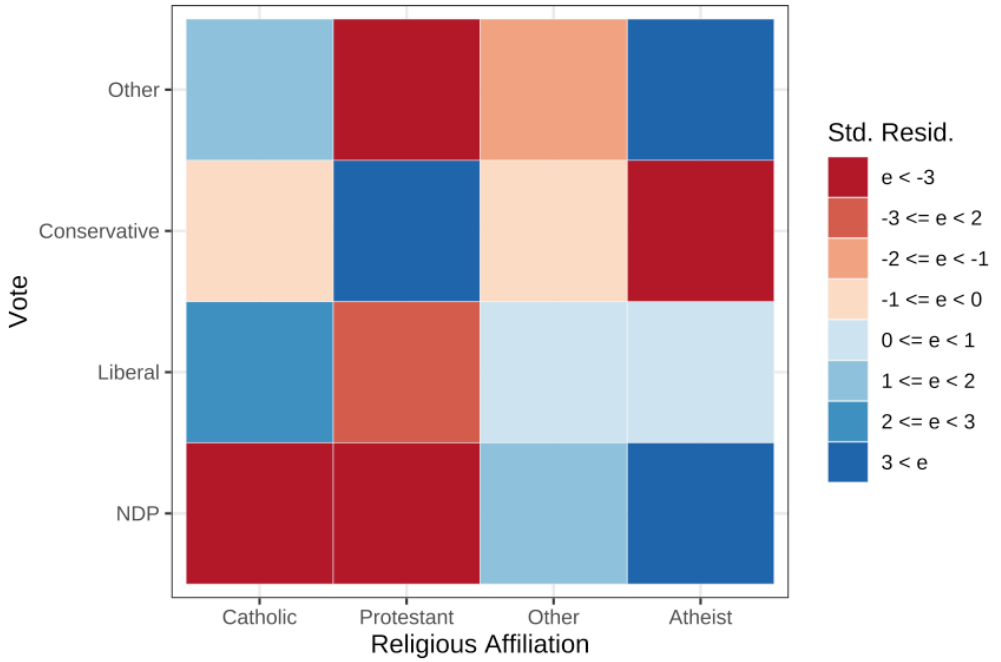
```
library(ggmosaic)
 ggplot(ces %>% filter(!is.na(relig) & !is.na(vote))) +
  geom_mosaic(aes(x = product(relig), fill=vote),
              show.legend = FALSE) +
  theme_mosaic() +
  scale_fill_manual(values = c("#d71920", "#003F72", "#F58220", '
```



Standardized residuals

R Python Stata

```
tab <- table(ces$vote, ces$relig)
tab %>% as.data.frame() %>%
  rename(vote = Var1, relig = Var2) %>%
  mutate(stdres = c(chisq.test(tab)$stdres),
    stdres2 = case_when(stdres < -3 ~ "e < -3",
      stdres >= -3 & stdres < -2 ~ "-3 <= e < 2",
      stdres >= -2 & stdres < -1 ~ "-2 <= e < -1",
      stdres >= -1 & stdres < 0 ~ "-1 <= e < 0",
      stdres >= 0 & stdres < 1 ~ "0 <= e < 1",
      stdres >= 1 & stdres < 2 ~ "1 <= e < 2",
      stdres >= 2 & stdres <= 3 ~ "2 <= e < 3",
      stdres > 3 ~ "3 < e"),
   stdres2 = factor(stdres2,
     levels=c("e < -3", "-3 <= e < 2", "-2 <= e < -1",
        "-1 <= e < 0", "0 <= e < 1", "1 <= e < 2", "2 <= e < 3",
   vote = factor(as.character(vote),
     levels=c("NDP", "Liberal", "Conservative", "Other")),
   relig = factor(as.character(relig),
     levels=c("Catholic", "Protestant", "Other", "Atheist"))) %>%
  ggplot(aes(x=relig, y=vote, fill=stdres2)) +
    geom_tile(col="white") +
    theme_bw() +
    scale_fill_brewer(palette="RdBu") +
    labs(y="Vote", x="Religious Affiliation",
         fill="Std. Resid.")
```



Exercises

1. Using the CES data, find the relationship between `region` and `relig`. What are the appropriate tests for this relationship? Is it statistically and substantively significant.

2. Using the GSS data, find the relationship between `SRH_110` and `SRH_115`. What are the appropriate tests for this relationship? Is it statistically and substantively significant.