# PLSC 502 – Autumn 2016 Measures of Association: Nominal Variables

October 27, 2016

# Frequency Tables

$$P_y = \frac{n_y}{N}.$$

| Category | Frequency | Proportion |
|---|---|---|
| No Civil War | 30 | 0.70 |
| Civil War | 13 | 0.30 |
| Total | 43 | 1.00 |

# Two-Way Crosstabs

- *Row proportions* (or percentages) are the proportion of observations in that row of the table (that is, with $Y = y$) falling into the column defined by $X = x$. They sum to 1.0 across columns.
- *Column proportions* (or percentages) are the proportion of observations in that column of the table (that is, with $X = x$) falling into the row defined by $Y = y$. They sum to 1.0 down rows.
- *Cell proportions* (or percentages) are the proportion of the total number of observations in that cell of the table. They sum to 1.0 overall columns and rows (cells).

# Two-Way Table

| | Sub-Saharan? | | |
|---|---|---|---|
| **Civil War?** | **No** | **Yes** | **Total** |
| **No** | 5 | 25 | 30 |
| (Row) | (0.17) | (0.83) | (1.00) |
| [Column] | [0.83] | [0.68] | [0.70] |
| {Cell} | {0.12} | {0.58} | {0.70} |
| **Yes** | 1 | 12 | 13 |
| (Row) | (0.08) | (0.92) | (1.00) |
| [Column] | [0.17] | [0.32] | [0.30] |
| {Cell} | {0.02} | {0.28} | {0.30} |
| **Total** | 6 | 37 | 43 |
| | (0.14) | (0.86) | (1.00) |
| | [1.00] | [1.00] | [1.00] |
| | {0.14} | {0.86} | {1.00} |

# The Setup

- $N$ total observations on nominal-level variables $Y$ and $X$
- $k_Y$ / $k_X$ = the number of different categories of $Y$ and $X$
- $n_{yx}$ = number of observations in the cell corresponding to category $\{x, y\}$
- $R_y = \sum_{k_X} n_{yx} =$ "marginals" of $Y$
- $C_x = \sum_{k_Y} n_{yx} =$ "marginals" of $X$

# Example: $2 \times 2$ table

| | $Y = ?$ | | |
|---|---|---|---|
| $X = ?$ | **0** | **1** | **Total** |
| **0** | $n_{00}$ | $n_{01}$ | $R_0$ |
| **1** | $n_{10}$ | $n_{11}$ | $R_1$ |
| **Total** | $C_0$ | $C_1$ | $N$ |

# Independence

Expectations...

$$E_{yx} = \frac{R_y \times C_x}{N}$$

For a one-way table:

$$E_y = N \times \frac{1}{k_Y}$$

*Statistical independence* implies:

$$H_0 : f(Y|X) = f(Y)$$

Suggests that if $Y \perp X$, then

- On average, $n_{yx} = E_{yx}$
- $n_{yx} - E_{yx}$ should be small

Chi-square statistic:

$$W = \sum_{k_Y k_x} \frac{(n_{yx} - E_{yx})^2}{E_{yx}}$$

Because

$$n_{yx} - E_{yx} \sim \mathcal{N}(0, \sigma_E^2)$$

we can show that:

$$W \sim \chi^2_{(k_Y - 1)(k_X - 1)}.$$

# Chi-Square Pointers

- Large values of $W$ are evidence against the (null / independence) hypothesis.

- In general, if $W \geq d.f.$, then $P$ is small.

- Can test vs. *any* expectation (e.g., that $E_{yx} = \frac{N}{k_Y k_X \, \forall \, x,y}$)

- Not recommended when $E_{yx} < 5$...

# Fisher's Exact Test

$$P = \frac{(R_1!R_2!...R_{k_Y}!)(C_1!C_2!...C_{k_X}!)}{N!\prod_{k_Y,k_X} n_{yx}!}.$$

- Intuition:
  - $N!\prod_{k_Y,k_X} n_{yx}! =$ possible ways in which one could arrange the data on $N$ observations in a $k_y \times k_X$ contingency table
  - $(R_1!R_2!...R_{k_Y}!)(C_1!C_2!...C_{k_X}!)$ reflects the possible orderings with the marginals determined by the values of $R$ and $C$.
- Difficult as tables get large...

# Example: Feminism as an Insult

*"Do you consider calling someone a feminist to be a compliment, an insult, or a neutral description?"*

```
> summary(DH)

     lcard        respon        intrace           feminsult         region        timezone
 Min.   :1   Min.   :   1   White:743   Compliment: 86   East   :206   Eastern :543
 1st Qu.:1   1st Qu.: 264   Black:244   Insult    :276   Midwest:287   Central :302
 Median :1   Median : 526   Asian: 64   Neutral   :595   South  :354   Mountain: 60
 Mean   :1   Mean   : 526               NA's      : 94   West   :204   Pacific :143
 3rd Qu.:1   3rd Qu.: 788                                              Bering  :  1
 Max.   :1   Max.   :1051                                              Hawaii  :  2
    race           religion
 White:885   Protestant:571
 Black:103   Catholic  :236
 Asian: 15   Jewish    : 18
 Other: 41   Other     : 45
 NA's :  7   None      :162
             NA's      : 19
```

# One-Way Tables

```
> oneway<-table(feminsult)

> oneway

feminsult
Compliment     Insult    Neutral
        86        276        595

> prop.table(oneway)

feminsult
Compliment     Insult    Neutral
   0.08986    0.28840    0.62173

> chisq.test(table(feminsult))

Chi-squared test for given probabilities

data:  table(feminsult)
X-squared = 414.8, df = 2, p-value < 2.2e-16
```

```
> region<-table(feminsult,region)

> addmargins(region)
           region
feminsult    East Midwest South West Sum
  Compliment   11      29    26   20  86
  Insult       45      69   102   60 276
  Neutral     137     167   192   99 595
  Sum         193     265   320  179 957

> prop.table(region)
           region
feminsult       East Midwest   South    West
  Compliment 0.01149 0.03030 0.02717 0.02090
  Insult     0.04702 0.07210 0.10658 0.06270
  Neutral    0.14316 0.17450 0.20063 0.10345

> prop.table(region,1)
           region
feminsult      East Midwest  South   West
  Compliment 0.1279  0.3372 0.3023 0.2326
  Insult     0.1630  0.2500 0.3696 0.2174
  Neutral    0.2303  0.2807 0.3227 0.1664
```

# Two-Way Tables (continued)

```
> prop.table(region,2)
           region
feminsult       East Midwest   South    West
  Compliment 0.05699 0.10943 0.08125 0.11173
  Insult     0.23316 0.26038 0.31875 0.33520
  Neutral    0.70984 0.63019 0.60000 0.55307


> chisq.test(region)

Pearson's Chi-squared test

data:  region
X-squared = 13.85, df = 6, p-value = 0.03133
```

# An Alternative: CrossTable

```
> require(gmodels)
> region2<-CrossTable(feminsult,Day18$region, prop.chisq=FALSE, chisq=TRUE)

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  957
.
.
.
```

# CrossTable (continued)

```
.
.
.
             | Day18$region
   feminsult |      East |   Midwest |     South |      West | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|
  Compliment |        11 |        29 |        26 |        20 |        86 |
             |     0.128 |     0.337 |     0.302 |     0.233 |     0.090 |
             |     0.057 |     0.109 |     0.081 |     0.112 |           |
             |     0.011 |     0.030 |     0.027 |     0.021 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
      Insult |        45 |        69 |       102 |        60 |       276 |
             |     0.163 |     0.250 |     0.370 |     0.217 |     0.288 |
             |     0.233 |     0.260 |     0.319 |     0.335 |           |
             |     0.047 |     0.072 |     0.107 |     0.063 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
     Neutral |       137 |       167 |       192 |        99 |       595 |
             |     0.230 |     0.281 |     0.323 |     0.166 |     0.622 |
             |     0.710 |     0.630 |     0.600 |     0.553 |           |
             |     0.143 |     0.175 |     0.201 |     0.103 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
Column Total |       193 |       265 |       320 |       179 |       957 |
             |     0.202 |     0.277 |     0.334 |     0.187 |           |
-------------|-----------|-----------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  13.85     d.f. =  6     p =  0.03133
```

# Three-Way Crosstabs

```
> threeway<-table(feminsult,region,intrace)
> addmargins(threeway)
, , intrace = White

            region
feminsult    East Midwest South West Sum
  Compliment   10      20    18   14  62
  Insult       34      47    71   42 194
  Neutral      98     120   131   75 424
  Sum         142     187   220  131 680

, , intrace = Black

            region
feminsult    East Midwest South West Sum
  Compliment    1       9     7    2  19
  Insult        8      12    26   13  59
  Neutral      33      40    49   19 141
  Sum          42      61    82   34 219
```

# Three-Way Crosstabs (continued)

```
, , intrace = Asian

            region
feminsult    East Midwest South West Sum
  Compliment    0       0     1    4   5
  Insult        3      10     5    5  23
  Neutral       6       7    12    5  30
  Sum           9      17    18   14  58

, , intrace = Sum

            region
feminsult    East Midwest South West Sum
  Compliment   11      29    26   20  86
  Insult       45      69   102   60 276
  Neutral     137     167   192   99 595
  Sum         193     265   320  179 957

> chisq.test(threeway)

Chi-squared test for given probabilities

data:  threeway
X-squared = 1490, df = 35, p-value < 2.2e-16
```

# Small Cell Frequencies

```
> table(feminsult,race)
           race
feminsult    White Black Asian Other
  Compliment    69    13     1     3
  Insult       244    21     2     8
  Neutral      496    61     9    25


> chisq.test(table(feminsult,race))

Pearson's Chi-squared test

data:  table(feminsult, race)
X-squared = 6.453, df = 6, p-value = 0.3744

Warning message:
In chisq.test(table(feminsult, race)) :
  Chi-squared approximation may be incorrect
```

# Small Cell Frequencies (continued)

```
> fisher.test(table(feminsult,race), workspace=20000000)

Fisher's Exact Test for Count Data

data:  table(feminsult, race)
p-value = 0.3681
alternative hypothesis: two.sided
```