# PLSC 502 – Autumn 2022
# Measures of Association, Part I: Nominal + Binary Variables

November 3, 2022

# Some Data

From a 1997 CBS/*NYT* poll of $\approx 1000$ Americans:

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

```
> summary(Fem)

    respon        intrace          relgpref         cenreg         timezone
 Min.   :   1   Asian: 58   Catholic  :224   East   :191   Bering  :  1
 1st Qu.: 264   Black:217   Jewish    : 15   Midwest:262   Central :275
 Median : 523   White:664   None      :147   South  :316   Eastern :492
 Mean   : 527               Other     : 39   West   :170   Hawaii  :  2
 3rd Qu.: 788               Protestant:514                 Mountain: 52
 Max.   :1050                                              Pacific :117
   race          feminsult
 Asian: 11   Compliment: 84
 Black: 93   Insult    :274
 Other: 36   Neutral   :581
 White:799
```

For each category of a nominal $Y$, the proportion of observations that have $Y = y$ is:

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

Frequency table:

```
> table(Fem$feminsult)

Compliment     Insult    Neutral
        84        274        581

> tab1(Fem$feminsult) # from -epiDisplay-

Fem$feminsult :
           Frequency Percent Cum. percent
Compliment        84     8.9          8.9
Insult           274    29.2         38.1
Neutral          581    61.9        100.0
  Total          939   100.0        100.0
```

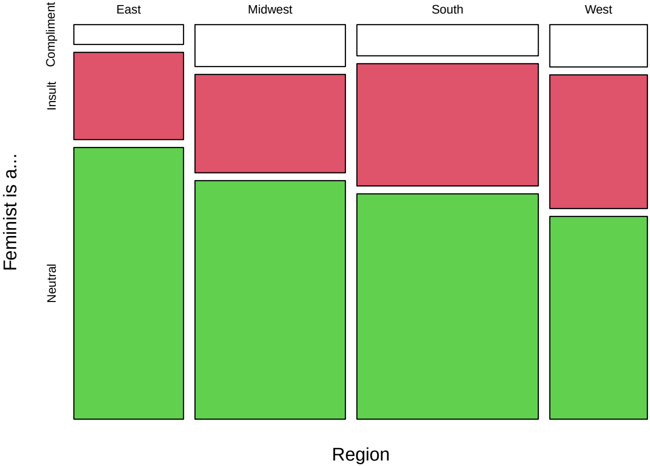
For an *outcome* variable $Y$ and a *predictor* variable $X$:

- Conventionally, we place the $Y$ variable on the "vertical" axis of the table (that is, values of $Y$ define *rows* of the cross-table) and the $X$ variable on the "horizontal" axis (values of $X$ define *columns* of the crosstab).
- *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).

## Feminist as a compliment/insult, by region:

```
> tabpct(Fem$feminsult, Fem$cenreg)

Original table
             Fem$cenreg
Fem$feminsult  East  Midwest  South  West  Total
   Compliment    10       29     26    19     84
   Insult        44       68    102    60    274
   Neutral      137      165    188    91    581
   Total        191      262    316   170    939

Row percent
             Fem$cenreg
Fem$feminsult    East  Midwest   South    West  Total
   Compliment      10       29      26      19     84
               (11.9)   (34.5)    (31)  (22.6)  (100)
   Insult          44       68     102      60    274
               (16.1)   (24.8)  (37.2)  (21.9)  (100)
   Neutral        137      165     188      91    581
               (23.6)   (28.4)  (32.4)  (15.7)  (100)

Column percent
             Fem$cenreg
Fem$feminsult  East       %  Midwest       %  South       %  West       %
   Compliment    10   (5.2)       29  (11.1)     26   (8.2)    19  (11.2)
   Insult        44  (23.0)       68  (26.0)    102  (32.3)    60  (35.3)
   Neutral      137  (71.7)      165  (63.0)    188  (59.5)    91  (53.5)
   Total        191   (100)      262   (100)    316   (100)   170   (100)
```

# Mosaic Plot



East
Midwest
South
West
Compliment
Insult
Neutral
Feminist is a...
Region

Preliminaries:

- $N$ 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 cell $\{x, y\}$
- $R_y = \sum_{k_X} n_{yx}$ = "marginals" of $Y$
- $C_x = \sum_{k_Y} n_{yx}$ = "marginals" of $X$

# Example: $3 \times 4$ table

| | $Y =$ | | | | |
|---|---|---|---|---|---|
| $X =$ | **East** | **Midwest** | **South** | **West** | **Total** |
| **Compliment** | $n_{CE}$ | $n_{CM}$ | $n_{CS}$ | $n_{CW}$ | $R_C$ |
| **Insult** | $n_{IE}$ | $n_{IM}$ | $n_{IS}$ | $n_{IW}$ | $R_I$ |
| **Neutral** | $n_{NE}$ | $n_{NM}$ | $n_{NS}$ | $n_{NW}$ | $R_N$ |
| **Total** | $C_E$ | $C_M$ | $C_S$ | $C_W$ | $N$ |

# Independence

For a one-way table, we would expect the cell defined by $Y = y$ to be:

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

For a two-way table, the expected cell frequency is:

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

*Statistical independence* implies:

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

This suggests that if $Y \perp X$, then

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

Chi-square statistic:

$$W = \sum \frac{(n_{yx} - E_{yx})^2}{E_{yx}}$$

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 Examples: Independence ($N = 90$)

```
> I
     [,1] [,2] [,3]
[1,]   10   10   10
[2,]   10   10   10
[3,]   10   10   10
> chisq.test(I)

Pearson's Chi-squared test

data:  I
X-squared = 0, df = 4, p-value = 1

> I
     [,1] [,2] [,3]
[1,]    5    5    5
[2,]   20   20   20
[3,]    5    5    5
> chisq.test(I)

Pearson's Chi-squared test

data:  I
X-squared = 0, df = 4, p-value = 1

> I
     [,1] [,2] [,3]
[1,]   20    5    5
[2,]   20    5    5
[3,]   20    5    5
> chisq.test(I)

Pearson's Chi-squared test

data:  I
X-squared = 0, df = 4, p-value = 1
```

# Chi-Square Examples: Dependence ($N = 90$)

```
> D
     [,1] [,2] [,3]
[1,]   20    5    5
[2,]    5   20    5
[3,]    5    5   20

> chisq.test(D)

Pearson's Chi-squared test

data:  D
X-squared = 45, df = 4, p-value = 0.000000004
```

```
> D
     [,1] [,2] [,3]
[1,]    9   12    9
[2,]   12    9    9
[3,]    9    9   12

> chisq.test(D)

Pearson's Chi-squared test

data:  D
X-squared = 1.8, df = 4, p-value = 0.8
```

Things to remember:

- 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 $E_{yx} < 5$...

Alternative: "Fisher's Exact Test" for independence:

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

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

# One-Way Chi-Square

```
> oneway<-with(Fem, table(feminsult))
> oneway
feminsult
Compliment     Insult    Neutral
        84        274        581
```

```
> X1<-chisq.test(table(Fem$feminsult))
> X1

Chi-squared test for given probabilities

data:  table(Fem$feminsult)
X-squared = 402, df = 2, p-value <0.0000000000000002
```

# Two-Way Chi-Square

```
> region<-with(Fem, table(feminsult,cenreg))

> region

            cenreg
feminsult     East Midwest South West
  Compliment    10      29    26   19
  Insult        44      68   102   60
  Neutral      137     165   188   91


> chisq.test(region)

Pearson's Chi-squared test

data:  region
X-squared = 17, df = 6, p-value = 0.008
```

# An Alternative: CrossTable

```
> region2<-with(Fem,
+             CrossTable(feminsult,cenreg,prop.chisq=FALSE,chisq=TRUE))


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


Total Observations in Table:  939

.
.
.
```

# CrossTable (continued)

```
.
.
.
             | cenreg
  feminsult |      East |   Midwest |     South |      West | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|
  Compliment |        10 |        29 |        26 |        19 |        84 |
             |     0.119 |     0.345 |     0.310 |     0.226 |     0.089 |
             |     0.052 |     0.111 |     0.082 |     0.112 |           |
             |     0.011 |     0.031 |     0.028 |     0.020 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
      Insult |        44 |        68 |       102 |        60 |       274 |
             |     0.161 |     0.248 |     0.372 |     0.219 |     0.292 |
             |     0.230 |     0.260 |     0.323 |     0.353 |           |
             |     0.047 |     0.072 |     0.109 |     0.064 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
     Neutral |       137 |       165 |       188 |        91 |       581 |
             |     0.236 |     0.284 |     0.324 |     0.157 |     0.619 |
             |     0.717 |     0.630 |     0.595 |     0.535 |           |
             |     0.146 |     0.176 |     0.200 |     0.097 |           |
-------------|-----------|-----------|-----------|-----------|-----------|
Column Total |       191 |       262 |       316 |       170 |       939 |
             |     0.203 |     0.279 |     0.337 |     0.181 |           |
-------------|-----------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  17.26     d.f. =  6     p =  0.008373
```

Conditioning $Y$ on two variables (say, $X_1$ and $X_2$)...

- Typically can't *show* the table(s)
- Independence:
  - Marginal independence: Variables $Y$ and (say) $X_1$ are independent *irrespective of the values of* $X_2$
  - Conditional independence: Variables $Y$ and (say) $X_1$ are independent *for a particular value of* $X_2$
  - Marginal independence can also be three-way...
  - Testing: the Cochran-Mantel-Haenszel test (see the link for details; also here)

# Three-Way Crosstabs: Example

```
> 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

> mantelhaen.test(threeway)

Cochran-Mantel-Haenszel test

data:  threeway
Cochran-Mantel-Haenszel M^2 = 17, df = 6, p-value = 0.01
```

```
> 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
```

# Measures of Association: Binary Variables

Binary variables are a bit weird...

- Ambiguous level of measurement...
- Related to proportions... For $Y \in \{0,1\}$:
  - $E(Y) \equiv \sum Y/N = \hat{\pi}$
  - Same as $\widehat{\Pr(Y_i = 1)}$
  - Variance is $\hat{\pi}(1-\hat{\pi})$
- Also potentially interval / ratio (as a "count")

We know that for two estimates $\hat{\pi}_1$ and $\hat{\pi}_2$, based on samples of size $N_1$ and $N_2$,

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\pi_1 - \pi_2}}$$

where

$$\hat{\sigma}_{\pi_1 - \pi_2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{N_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{N_2}}$$

We can think about this as samples of $Y$ drawn from (say) $X = 0$ and $X = 1$:

$$\hat{\sigma}_{\pi_{Y|X=0} - \pi_{Y|X=1}} = \sqrt{\frac{\hat{\pi}_{Y|X=0}(1 - \hat{\pi}_{Y|X=0})}{N_{X=0}} + \frac{\hat{\pi}_{Y|X=1}(1 - \hat{\pi}_{Y|X=1})}{N_{X=1}}}$$

We also know that:

$$W = \sum_{k_X k_Y} \frac{(N_{XY} - E_{XY})^2}{E_{XY}}$$

and that:

$$W \sim \chi_1^2$$

when both $X$ and $Y$ are binary.

In fact, $z^2 = W$...

```
> T <- table(Y,X)
> T
   X
Y   0 1
  0 5 3
  1 4 8

> chisq.test(T,correct=FALSE)

Pearson's Chi-squared test

data:  T
X-squared = 1.65, df = 1, p-value = 0.2

> p1<-4/9
> p2<-8/11
> p <- 12/20
> se <- sqrt(((p*(1-p)*(1/9+1/11))))
> Z <- (p1-p2) / se
> Z
[1] -1.2845

> Z^2
[1] 1.6498
```

# $\chi^2$ Is *Not* A Measure Of Association

```
> chisq.test(T, correct=FALSE)

Pearson's Chi-squared test

data:  T
X-squared = 1.65, df = 1, p-value = 0.199

> X <- rep(X,times=10)
> Y <- rep(Y,times=10)
> T10 <- table(Y,X)
> T10
   X
Y    0  1
  0 50 30
  1 40 80
> chisq.test(T10,correct=FALSE)

Pearson's Chi-squared test

data:  T10
X-squared = 16.5, df = 1, p-value = 0.0000487
```

# "Contingency Tables"

*Contingency table*:

| | $X = 0$ | $X = 1$ | |
|---|---|---|---|
| $Y = 0$ | $N_{00}$ | $N_{10}$ | $N_{\bullet 0}$ |
| $Y = 1$ | $N_{01}$ | $N_{11}$ | $N_{\bullet 1}$ |
| | $N_{0\bullet}$ | $N_{1\bullet}$ | $N$ |

**Q: How much more or less likely is $Y = 1|X = 1$ than $Y = 1|X = 0$?**

Recall that the *odds* of $Y = 1|X = 1$ are:

$$\begin{aligned} O_{Y=1|X=1} &= \frac{\Pr(Y = 1|X = 1)}{\Pr(Y = 0|X = 1)} \\ &= \frac{\hat{\pi}_{Y=1|X=1}}{\hat{\pi}_{Y=0|X=1}} \\ &= \frac{N_{11}/N_{1\bullet}}{N_{10}/N_{1\bullet}} \\ &= \frac{N_{11}}{N_{10}} \end{aligned}$$

And similarly:

$$O_{Y=1|X=0} = \frac{N_{01}}{N_{00}}$$

The *odds ratio* is then:

$$
\begin{aligned}
OR &= \frac{O_{Y=1|X=1}}{O_{Y=1|X=0}} \\
&= \frac{N_{11}/N_{10}}{N_{01}/N_{00}}
\end{aligned}
$$

# Odds Ratio Facts...

Odds ratios (OR):

- $OR$ expresses the *relative* odds of an event ($Y = 1$) under one condition ($X = 1$) versus another ($X = 0$).
- $OR \in [0, \infty)$
- Interpretation:
  - $OR = 1 \;\leftrightarrow\;$ no association
  - $OR > 1 \;\leftrightarrow\;$ positive association
  - $OR < 1 \;\leftrightarrow\;$ negative association
- The "inverse odds ratio" ($O_{Y=0|X=1}/O_{Y=0|X=0}$) is simply the reciprocal of $OR$.

```
> T
   X
Y   0 1
  0 5 3
  1 4 8

> OR <- (T[1,1])*T[2,2] / (T[1,2]*T[2,1])
> OR
[1] 3.33333

> require(DescTools)
> OddsRatio(T)
[1] 3.33333
```

For the contingency table above,

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{1\bullet}N_{0\bullet}N_{\bullet 0}N_{\bullet 1}}}$$

Also,

$$\phi^2 = \frac{\chi^2}{N} \quad \text{so} \quad |\phi| = \sqrt{\frac{\chi^2}{N}}$$

# A Few Things About $\phi$

- A/K/A the "mean square contingency coefficient" or Matthews' Correlation Coefficient (MCC)

- $\phi \in [0, 1]$ (but see below...)

- In general:
  - $\phi \in [0.7, 1.0]$ = a strong positive association
  - $\phi \in [0.4, 0.7]$ = a moderate positive association
  - $\phi \in [0.1, 0.4]$ = a weak positive association
  - $\phi \in [-0.1, 0.1]$ = no association
  - $\phi \in [-0.1, -0.4]$ = a weak negative association
  - $\phi \in [-0.4, -0.7]$ = a moderate negatie association
  - $\phi \in [-0.7, -1.0]$ = a strong negative association

- $\phi$ equals Pearson's correlation coefficient ($r$) applied to two binary variables.

- The equation above means that $\phi^2 \times N \sim \chi^2_1$, which can be used for hypothesis testing (e.g., for $H_0 : \phi = 0$).

# $\phi$ Examples...

```
> T
   X
Y   0 1
  0 5 3
  1 4 8

> require(psych)
> phi(T)
[1] 0.29

> cor(X,Y)
[1] 0.287213
```

# $\phi$ Examples (continued)

```
> Tpos<-as.table(rbind(c(10,0),c(0,10)))
> Tpos
   A  B
A 10  0
B  0 10
> phi(Tpos)
[1] 1

> Tneg<-as.table(rbind(c(0,10),c(10,0)))
> Tneg
   A  B
A  0 10
B 10  0
> phi(Tneg)
[1] -1

> T0<-as.table(rbind(c(5,5),c(5,5)))
> T0
  A B
A 5 5
B 5 5
> phi(T0)
[1] 0
```

# $\phi$: Restricted Range

From the Stata manual (entry for `tetrachoric`):

from −1 to 1. To illustrate, consider the following set of tables for two binary variables, X and Z:

| | Z = 0 | Z = 1 | |
|---|---|---|---|
| X = 0 | $20-a$ | $10+a$ | 30 |
| X = 1 | $a$ | $10-a$ | 10 |
| | 20 | 20 | 40 |

For $a$ equal to 0, 1, 2, 5, 8, 9, and 10, the Pearson and tetrachoric correlations for the above table are

| $a$ | 0 | 1 | 2 | 5 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|
| Pearson | 0.577 | 0.462 | 0.346 | 0 | −0.346 | −0.462 | −0.577 |
| Tetrachoric | 1.000 | 0.792 | 0.607 | 0 | −0.607 | −0.792 | −1.000 |

Setup:

- $N$ observations, with
- $T_i$ a *latent* trait for each observation;
- two *raters*, $\{1, 2\}$, each of which
  - observes a "noisy" version of $T_i$:

$$\begin{aligned} T_i^{*1} &= T_i + e_{1i} \\ T_i^{*2} &= T_i + e_{2i} \end{aligned}$$

  - and gives a binary rating to $i$; equals 0 if $T_i < \tau$, 1 if $T_i > \tau$. Call these $X_{1i}$ and $X_{2i}$.
- Assume that $\{e_{1i}, e_{2i}\} \sim \Phi_2(0, 0, 1, 1, \rho)$ (*bivariate normal*)

The Bivariate Normal is:

$$\Pr(X_1, X_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left[\frac{-z}{2(1-\rho^2)}\right]$$

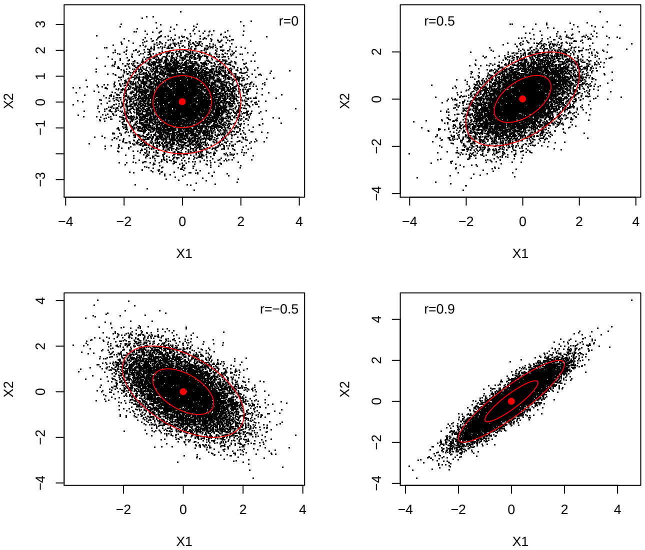
where

$$z = \left[\frac{(X_1 - \mu_{X_1})^2}{\sigma_{X_1}^2} + \frac{(X_2 - \mu_{X_2})^2}{\sigma_{X_2}^2} - \frac{2\rho(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})}{\sigma_{X_1}\sigma_{X_2}}\right]$$

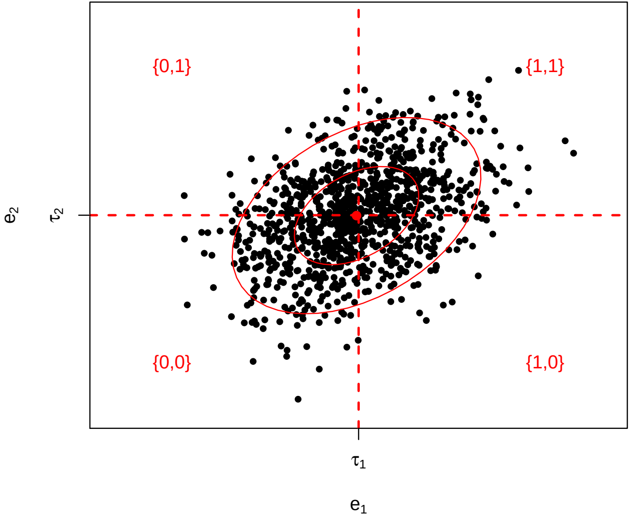
and

$$\rho \ = \ \operatorname{corr}(X_1, X_2)$$

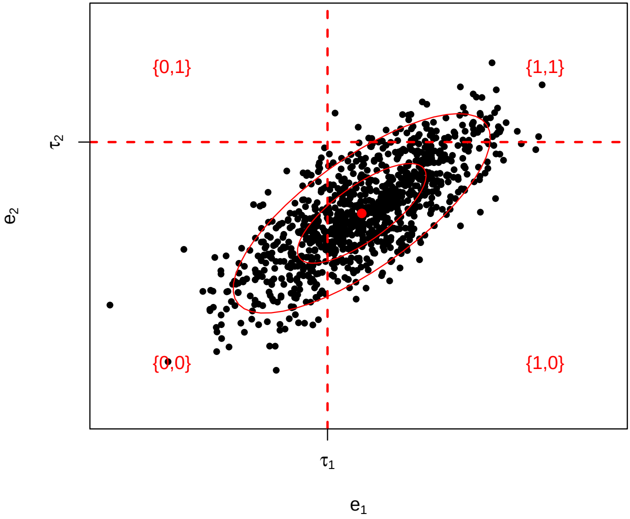
# Bivariate Normals Illustrated



# Back to Tetrachoric Correlation



# Tetrachoric Correlation (continued)



# More Tetrachoric Correlation

Idea: Get as close to:

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 0$ | $\pi_{00}$ | $\pi_{10}$ |
| $X_2 = 1$ | $\pi_{01}$ | $\pi_{11}$ |

...using three parameters: $\tau_1$, $\tau_2$, and $\rho$.

Tetrachoric correlation $r_{tet}$:

- $r_{tet} \in [-1, 1]$
- Assumes two continuous, *Normal* underlying (latent) variables...
- Fitted via ML, etc. but also has a simple approximate formula:

$$r_{tet} \approx \frac{\alpha - 1}{\alpha + 1}$$

where

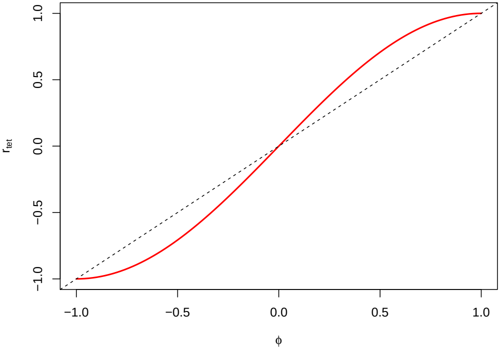
$$\alpha = (OR)^{\frac{\pi}{4}}$$

# $r_{tet}$: An Example

```
> require(polycor)
> T
   X
Y   0 1
  0 5 3
  1 4 8

> polychor(T)
[1] 0.4399

> # Compare:
>
> phi(T)
[1] 0.29

> # Approximate formula:
>
> alpha <- (OR)^(pi/4)
> rtet <- (alpha - 1) / (alpha + 1)
> rtet
[1] 0.440458
```

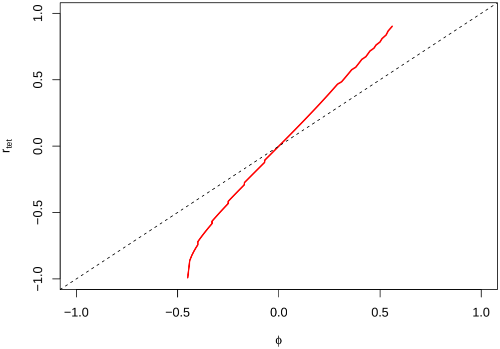
# $r_{tet}$ vs. $\phi$: Symmetrical Marginals

```
> addmargins(ST)
      A   B Sum
A     0 100 100
B   100   0 100
Sum 100 100 200
```



# $r_{tet}$ vs. $\phi$: Asymmetrical Marginals

```
> addmargins(AT)
      A   B Sum
A     0 150 150
B   100 150 250
Sum 100 300 400
```



# Binary Association Summary

Some general thoughts:

- Odds ratios are natural for describing $2 \times 2$ associations, *but*
- In general, we like $\phi$ / MCC as a single measure of binary association
- Some of the other things we'll discuss next week are also useful for binary responses (e.g., Spearman's $r$)
- We'll also discuss binary variables a bit later, in the context of classification...