PLSC 502 – Autumn 2024
Measures of Association

November 11, 2024

Our "to do" list:

- *Nominal* Variables: Frequency Tables / Crosstabs, Chi-Square, etc.
- *Binary* Variables: Odds Ratios, ϕ / MCC, and Tetrachoric Correlation
- *Ordinal* Variables: γ and the τs
- *Interval/Ratio* Variables: Linearity, r, and ρ

From a 1997 CBS/*NYT* poll of ≈ 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
```

Two-Way Crosstabs

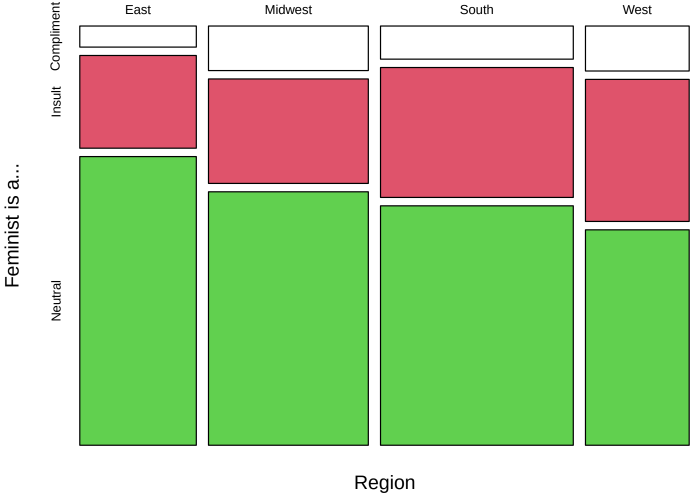
For an *outcome* variable Y and a *predictor* variable X:

- 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 <u>across columns</u>.
- *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 <u>down rows</u>.
- *Cell proportions* (or percentages) are the proportion of the total number of observations in that cell of the table. They sum to 1.0 over <u>all columns and rows</u> (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: 3×4 table

	$X =$				
$Y =$	**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

For a one-way table, we would expect the number of observations in the cell defined by $Y = y$ – that is, the *cell frequency* – 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 under $Y \perp X$:

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

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
```

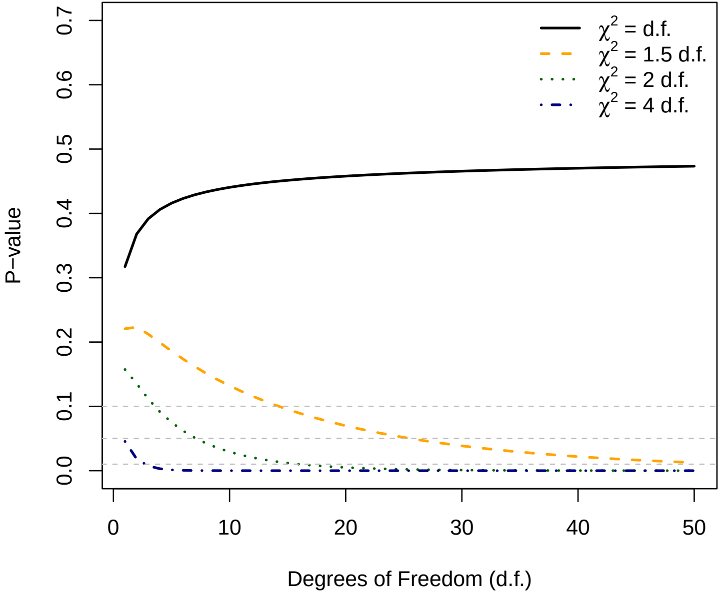
```
> D
     [,1] [,2] [,3]
[1,]    9   12    9
[2,]   12    9    9
[3,]    9    9   12

> chisq.test(D)

 Pearson's Chi-squared test

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

Chi-Square Pointers

Things to remember:

- Large values of W are evidence against the (null / independence) hypothesis.
- In general, if $W \geq 2 \times d.f.$, then P is small (see below).
- Can test vs. *any* expectation (e.g., that $E_{yx} = \frac{N}{k_Y k_X \forall x,y}$)
- Not recommended when $E_{yx} < 5$...

Heuristic χ^2 Values by d.f.



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
```

Three-Way Crosstabs

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
```

```
, , 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
```

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:

$$W \sim \chi^2_1$$

when both X and Y are binary.

z and χ^2

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
```

χ^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 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
```

Association measure: ϕ

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 ϕ

Fun ϕ facts:

- A/K/A the "mean square contingency coefficient" or Matthews' Correlation Coefficient (MCC)
- $\phi \in [-1, 1]$ (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
- ϕ 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 $H_0 : \phi = 0$).

ϕ 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
```

ϕ 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
```

ϕ: 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

Tetachoric Correlation (r_{tet})

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:

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

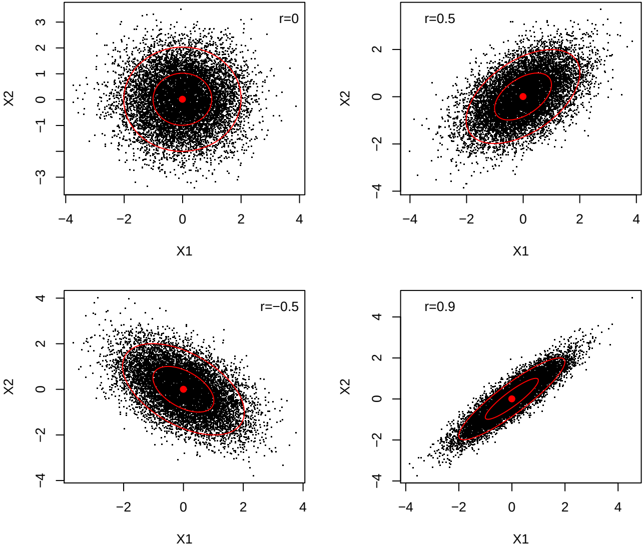
where

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

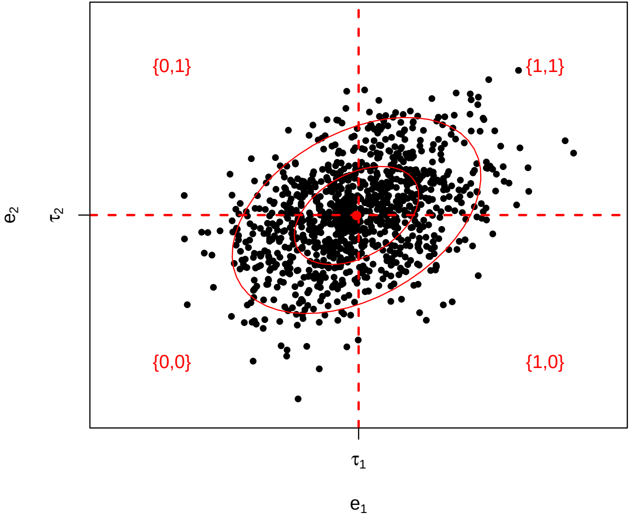
and

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

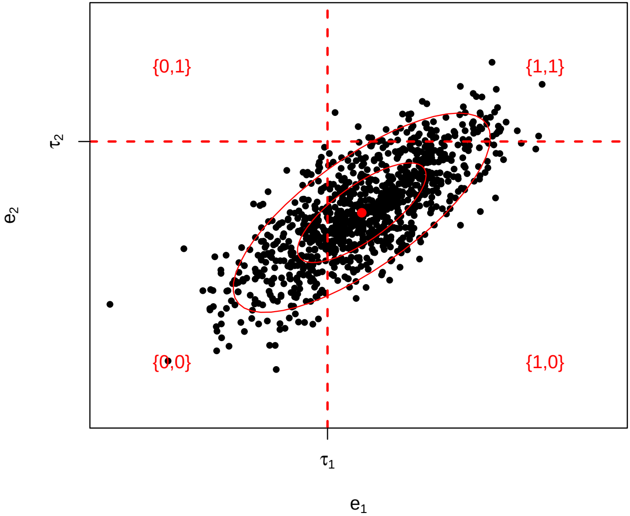
Bivariate Normals Illustrated



Back to Tetrachoric Correlation



Tetrachoric Correlation (continued)



More Tetrachoric Correlation

Idea: Get as close to:

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	π_{00}	π_{10}
$X_2 = 1$	π_{01}	π_{11}

...using three parameters: τ_1, τ_2, and ρ.

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

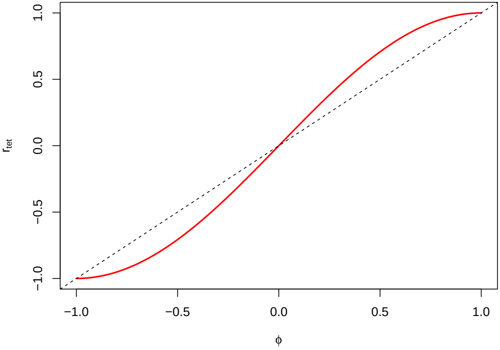
$$\alpha = (OR)^{\frac{\pi}{4}}$$

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
```

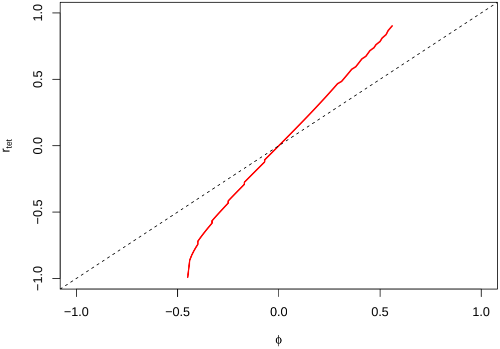
r_{tet} vs. ϕ: Symmetrical Marginals

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



r_{tet} vs. ϕ: 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 2×2 associations, *but*
- In general, we like ϕ / MCC as a single measure of binary association, *provided that the marginals are not badly skewed*
- For more skewed marginals, r_{tet} is probably better (read about this, and the famous Pearson-Yule debate, here)
- 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...

Ordinal Variables

Ordinal variables:

- Key issue: *how to retain the information present in the ordering of the categories without giving the numerical values assigned to them cardinal content*.

- Key concept: **Concordance**

For a pair of values on two observations $i = \{1, 2\}$ and two variables X and Y, a *concordant pair* has:

$$\text{sign}(X_2 - X_1) = \text{sign}(Y_2 - Y_1)$$

and a *discordant* pair has:

$$\text{sign}(X_2 - X_1) = -\text{sign}(Y_2 - Y_1).$$

A(nother) Contingency Table

Consider two ordinal variables X and Y:

		X			
		1	2	3	
	1	n_{11}	n_{12}	n_{13}	n_{1X}
Y	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

Concordance with $\{1, 1\}$ observations:

		X 1	2	3	
	1	n_{11}	n_{12}	n_{13}	n_{1X}
Y	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

Concordance with $\{1, 2\}$ observations:

		X 1	2	3	
	1	n_{11}	n_{12}	n_{13}	n_{1X}
Y	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

Concordant and Discordant Pairs

Discordance with $\{1, 2\}$ observations:

		X 1	2	3	
Y	1	n_{11}	n_{12}	n_{13}	n_{1X}
	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

Discordance with $\{1, 3\}$ observations:

		X 1	2	3	
Y	1	n_{11}	n_{12}	n_{13}	n_{1X}
	2	n_{21}	n_{22}	n_{23}	n_{2X}
	3	n_{31}	n_{32}	n_{33}	n_{3X}
		n_{Y1}	n_{Y2}	n_{Y3}	N

Concordant and Discordant Pairs

For a 3×3 table, the total number of *concordant pairs* is:

$$N_c = n_{11}(n_{22} + n_{23} + n_{32} + n_{33}) + n_{12}(n_{23} + n_{33}) + n_{21}(n_{32} + n_{33}) + n_{22}(n_{33})$$

and the total number of *discordant pairs* is:

$$N_d = n_{13}(n_{21} + n_{22} + n_{31} + n_{32}) + n_{12}(n_{21} + n_{31}) + n_{23}(n_{31} + n_{32}) + n_{22}(n_{31}).$$

This extends to a table of arbitrary size $M \times N$ straightforwardly...

Gamma (γ)

Gamma (γ) is the normed difference between the number of concordant and discordant pairs in the data:

$$\gamma = \frac{N_c - N_d}{N_c + N_d}$$

Equivalently:

$$\gamma = \frac{N_c}{N_c + N_d} - \frac{N_d}{N_c + N_d}$$

About γ

Gamma:

- does not count "ties"
- $\gamma \in [-1, 1]$
- $\gamma = 0 \leftrightarrow$ no association between X and Y, though it can also happen whenever $N_c = N_d$. That is, $\gamma = 0$ is necessary but not sufficient for statistical independence
- Higher absolute values of γ correspond to stronger associations between X and Y
- $\gamma = \pm 1.0$ under conditions of (at least) *weak monotonicity* (e.g., γ will equal 1.0 whenever, as X increases, Y only increases or stays the same)

For a 2×2 table:

		X		(Total)
		0	1	
Y	0	n_{00}	n_{01}	$(n_{00} + n_{01})$
	1	n_{10}	n_{11}	$(n_{10} + n_{11})$
(Total)		$(n_{00} + n_{10})$	$(n_{01} + n_{11})$	(N)

we have:

$$\begin{aligned}
\hat{\gamma} &= \text{“Yule's Q”} \\
&= \frac{n_{00}n_{11} - n_{01}n_{10}}{n_{00}n_{11} + n_{01}n_{10}} \\
&= \frac{OR - 1}{OR + 1}
\end{aligned}$$

Inference on γ

It can be shown that:

$$\hat{\gamma} \sim \mathcal{N}(\gamma, \sigma_\gamma^2)$$

where

$$\sigma_\gamma^2 = \frac{N(1-\hat{\gamma}^2)}{N_c + N_d}$$

So we can approximate:

$$t \approx (\hat{\gamma} - \gamma)\sqrt{\frac{N_c + N_d}{N(1-\hat{\gamma}^2)}}.$$

(Goodman-Kruskal's) "Tau-a":

$$\tau_a = \frac{N_c - N_d}{\frac{1}{2}N(N-1)}$$

(Kendall's) "Tau-b":

$$\tau_b = \frac{N_c - N_d}{\sqrt{[(N_c + N_d + N_{Y^*})(N_c + N_d + N_{X^*})]}}$$

where N_{Y^*} and N_{X^*} are the number of pairs *not tied* on Y and X, respectively.

(Stuart's) "Tau-c":

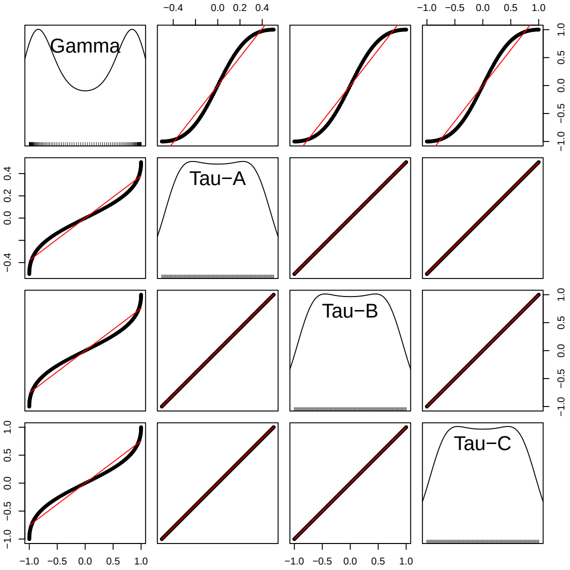
$$\tau_c = (N_c - N_d) \times \left\{ \frac{2m}{[N^2 2(m-1)]} \right\}$$

where m is the number of rows or columns, whichever is smaller.

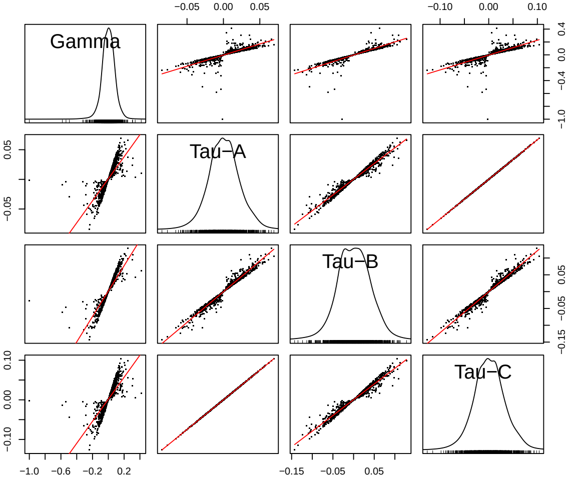
Tau tips:

- All except τ_a have $\tau_{(\cdot)} \in [-1, 1]$
- For all τs, the numerator signs the statistic.
- Like γ, τ_a doesn't do "ties" $\rightarrow$ attenuated range
- $|\tau_b| = 1.0$ only under *strict monotonicity*
- $\tau_b \rightarrow$ "square" tables
- $\tau_c \rightarrow$ "rectangular" (asymmetrical) tables
- $\gamma \geq \tau \;\forall\; \tau_{(\cdot)}$

γ and the τs Compared (2×2 Tables)



γ / τs Comparison (Random 3×3 Tables)



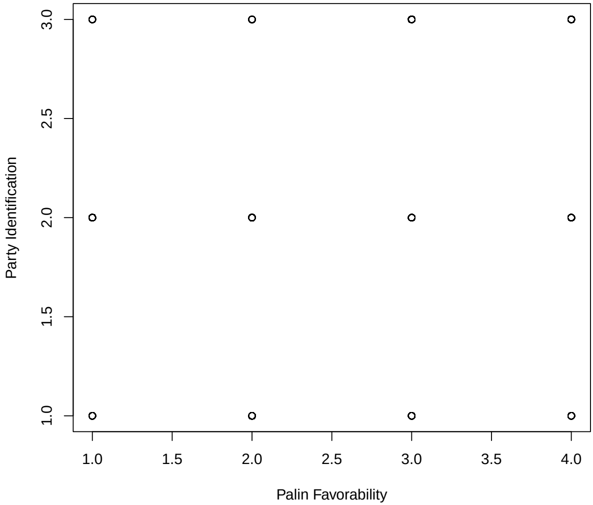
Example: Sarah Palin Support...

September 2008 "Battleground" Poll in PA:

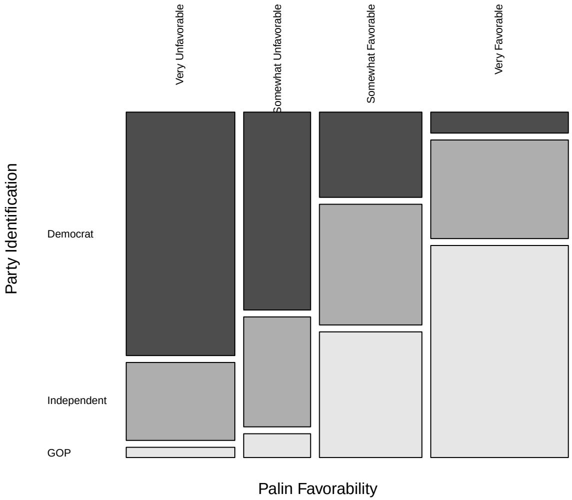
```
> summary(MamaGriz)
    caseid          female                          palin                   pid
 Min.   :    2   Male  :2221   Very Unfavorable     :1200   Democrat   :1709
 1st Qu.:30034   Female:2370   Somewhat Unfavorable: 739   Independent:1391
 Median :31831                 Somewhat Favorable   :1132   GOP        :1491
 Mean   :36776                 Very Favorable       :1520
 3rd Qu.:60674
 Max.   :62125
```

```
> palinpid<-with(MamaGriz, xtabs(~palin+pid))
> addmargins(palinpid)
                      pid
palin                  Democrat Independent  GOP  Sum
  Very Unfavorable          881         282   37 1200
  Somewhat Unfavorable      441         245   53  739
  Somewhat Favorable        291         412  429 1132
  Very Favorable             96         452  972 1520
  Sum                      1709        1391 1491 4591
```

Plotting: Don't



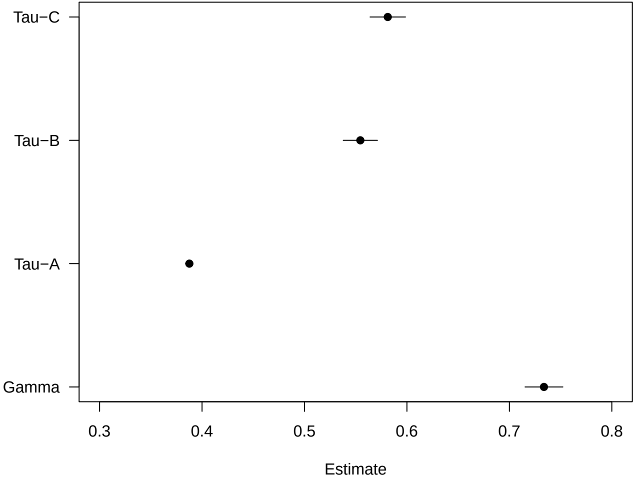
Plotting: Do


Very Unfavorable
Somewhat Unfavorable
Somewhat Favorable
Very Favorable
Party Identification
Democrat
Independent
GOP
Palin Favorability

Estimating γ and the τs

```
> # Gamma:
>
> GoodmanKruskalGamma(palinpid,conf.level=0.95)
  gamma  lwr.ci  ups.ci
0.73376 0.71529 0.75223

> #Tau-A:
>
> KendallTauA(palinpid,conf.level=0.95)
  tau_a  lwr.ci  ups.ci
0.38762 0.38639 0.38884

> # Tau-B:
>
> KendallTauB(palinpid,conf.level=0.95)
  tau_b  lwr.ci  ups.ci
0.55453 0.53784 0.57121

> # Tau-C:
>
> StuartTauC(palinpid,conf.level=0.95)
   tauc  lwr.ci  ups.ci
0.58130 0.56401 0.59859
```

γ and the τs: Party Identification



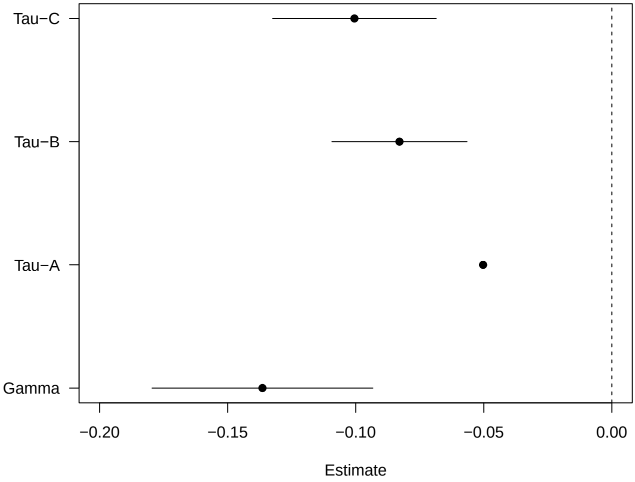
Men vs. Women on Palin

```
> palinfemale<-with(MamaGriz, xtabs(~palin+female))

> addmargins(palinfemale)
                         female
palin                     Male Female  Sum
  Very Unfavorable         508    692 1200
  Somewhat Unfavorable     328    411  739
  Somewhat Favorable       575    557 1132
  Very Favorable           810    710 1520
  Sum                     2221   2370 4591
```

Men vs. Women on Palin

```
> GoodmanKruskalGamma(palinfemale,conf.level=0.95)
    gamma    lwr.ci    ups.ci
-0.136410 -0.179514 -0.093306


> KendallTauA(palinfemale,conf.level=0.95)
    tau_a    lwr.ci    ups.ci
-0.050259 -0.051137 -0.049382


> KendallTauB(palinfemale,conf.level=0.95)
    tau_b    lwr.ci    ups.ci
-0.082912 -0.109268 -0.056556


> StuartTauC(palinfemale,conf.level=0.95)
     tauc    lwr.ci    ups.ci
-0.100497 -0.132442 -0.068552
```

γ and the τs: Men vs. Women



Interval + Ratio-Level Variables

Linearity means:

$$\frac{\partial Y}{\partial X} = m;$$

$$Y = mX + b$$

Other monotonic + "smooth" alternatives:

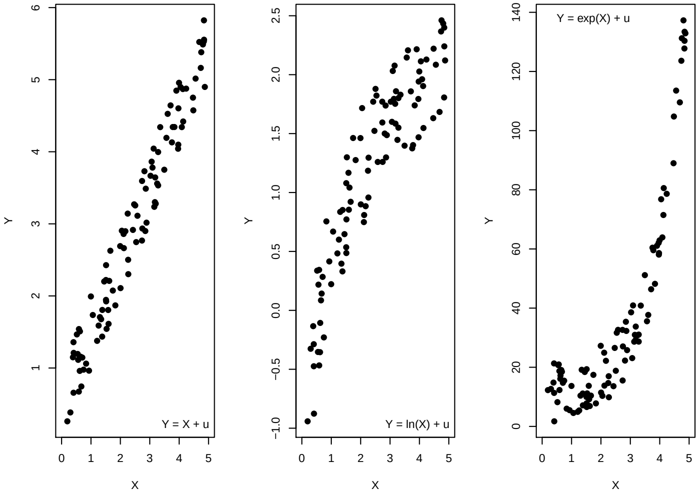
- *Logarithmic*:

$$\frac{\partial^2 Y}{\partial X \partial X} < 0$$

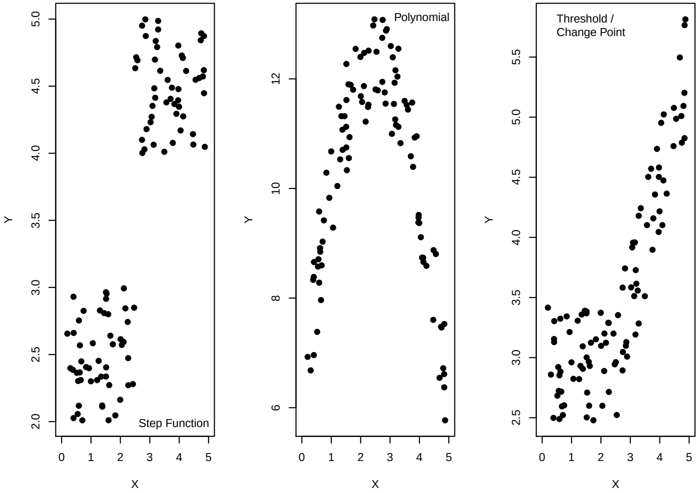
- *Exponential*:

$$\frac{\partial^2 Y}{\partial X \partial X} > 0$$

Linear, Logarithmic, Exponential



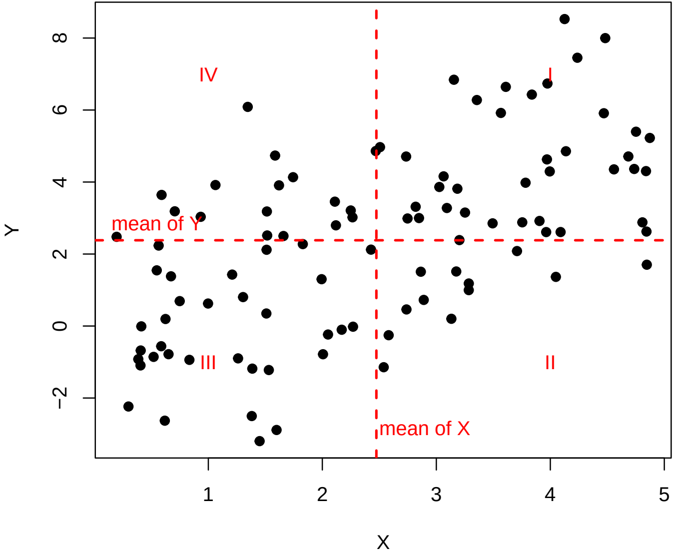
Other Possibilities



"Pearson's product-moment correlation coefficient":

$$
\begin{aligned}
r &= \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \\
&= \frac{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)}{N-1}
\end{aligned}
$$

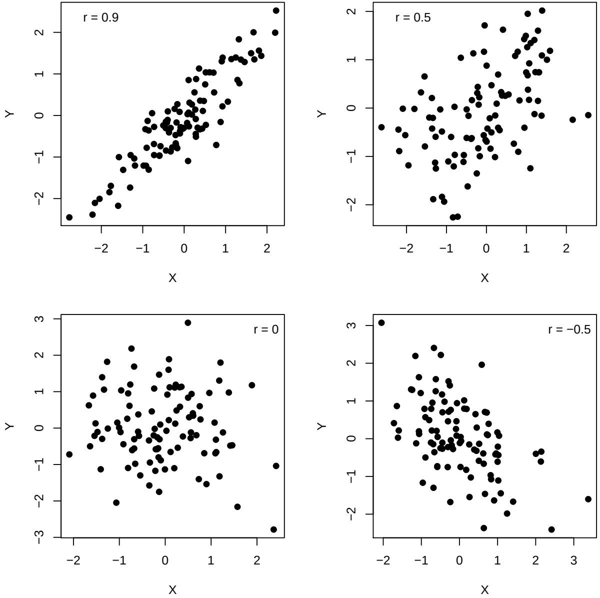
Pearson's r: Intuition



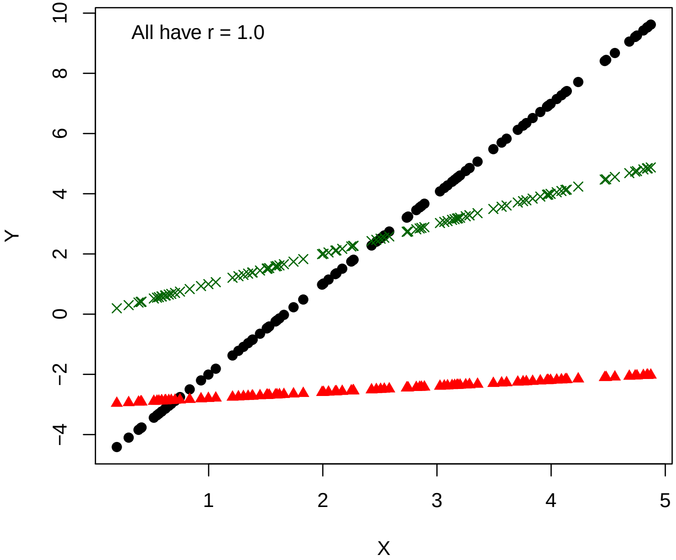
Pearson's r: Characteristics

- $r \in [-1, 1]$
- $r = 0 \leftrightarrow$ no *linear* association between Y and X.
- Sign(r) $\rightarrow$ "direction" of the *linear* association
- $|r| \rightarrow$ "strength" of the *linear* association
- In general:
    - $|r| < 0.3 \rightarrow$ "weak" linear association
    - $0.3 < |r| < 0.7 \rightarrow$ "moderate" linear association
    - $|r| > 0.7 \rightarrow$ "strong" linear association

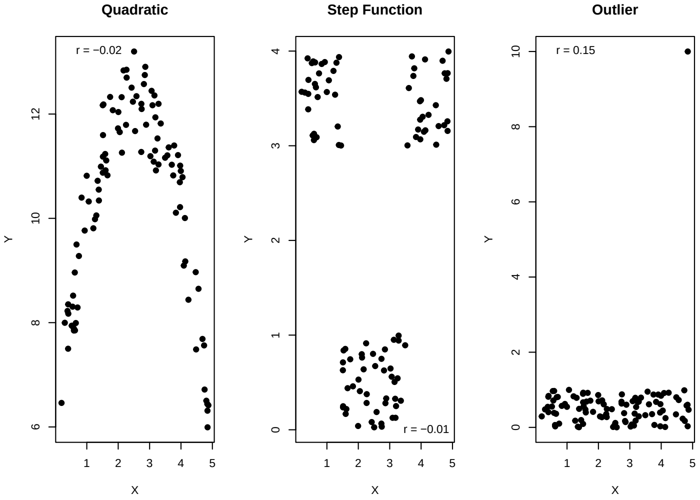
Examples



$r = \pm 1.0 \rightarrow ?$



Nonlinearity, etc.



The sampling distribution of r is:

- complex, and
- skewed as $|r| \to 1.0$.

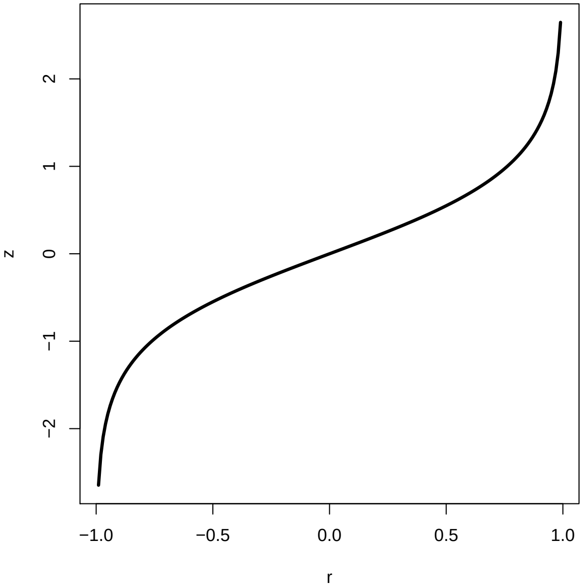
Fisher:

$$\hat{w} \equiv \frac{1}{2}\ln\left(\frac{1+\hat{r}}{1-\hat{r}}\right) \sim \mathcal{N}\left[\frac{1}{2}\ln\left(\frac{1+\hat{r}}{1-\hat{r}}\right), \frac{1}{\sqrt{N-3}}\right]$$

implying:

$$z_r = \frac{\frac{1}{2}\ln\left(\frac{1+\hat{r}}{1-\hat{r}}\right) - \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)}{\sqrt{\frac{1}{N-3}}} \sim \mathcal{N}(0,1)$$

Fisher's z Transformation of r



Alternative Approach (t)

Under $r = 0$, the standard error of $\hat{r}$ is:

$$\sigma_r = \sqrt{\frac{1 - r^2}{N - 2}}$$

This means that we can construct confidence intervals using a t distribution, as:

$$\frac{\hat{r}}{\sigma_r} = \frac{\hat{r}\sqrt{N - 2}}{\sqrt{1 - \hat{r}^2}} \quad \sim \quad t_{N-2}.$$

Note that this converges to z as $N \to \infty$.

Alternative Measure: Spearman's ρ

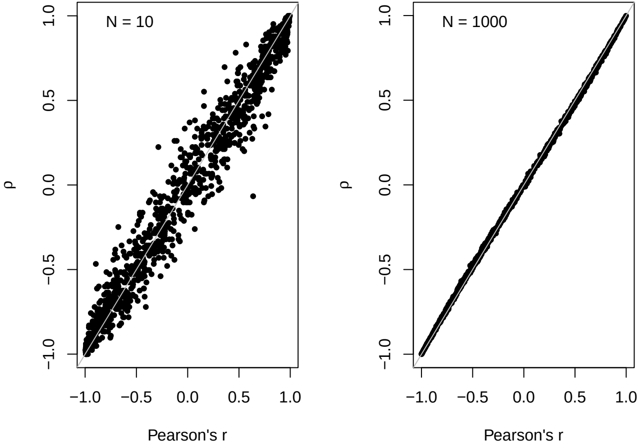
For sorted data on X and Y, where R_{Y_i} and R_{X_i} are the respective ranks,

$$\rho = 1 - \frac{6\sum_{i=1}^{N}(R_{Y_i} - R_{X_i})^2}{N(N^2-1)}$$

Characteristics:

- $\rho \in [-1, 1]$
- Same interpretation as r.
- Also appropriate for use with ordinal data; but
- When many "ties" occur, calculate Pearson's r on the ranks R_{Y_i} and R_{X_i}, and assign "partial" (or "half") ranks to tied individuals.

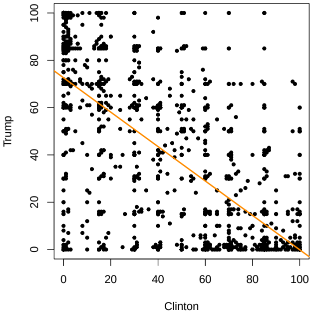
r vs. ρ Comparison (Simulation)



Real Data: ANES 2016 Feeling Thermometers

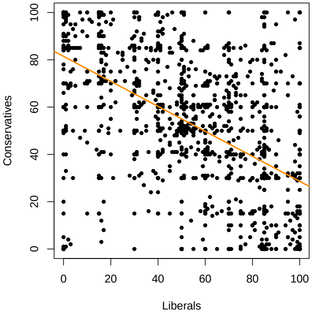
```
> describe(Therms,range=FALSE)
                            vars    n  mean    sd  skew kurtosis   se
Asian-Americans                1 2387 70.17 20.20 -0.38     0.02 0.41
Hispanics                      2 2387 69.35 20.91 -0.41     0.01 0.43
Blacks                         3 2387 69.00 21.19 -0.35    -0.24 0.43
Illegal Immigrants             4 2387 42.54 27.31  0.13    -0.71 0.56
Whites                         5 2387 71.63 19.40 -0.46     0.08 0.40
Dem. Pres. Candidate           6 2387 44.12 34.91  0.12    -1.42 0.71
GOP Pres. Candidate            7 2387 40.53 35.65  0.23    -1.43 0.73
Libertarian Pres. Candidate    8 2387 43.61 19.92 -0.58     0.25 0.41
Green Pres. Candidate          9 2387 43.20 20.87 -0.54     0.22 0.43
Dem. VP                       10 2387 48.24 25.91 -0.22    -0.44 0.53
GOP VP                        11 2387 49.59 33.42 -0.10    -1.21 0.68
John Roberts                  12 2387 53.75 18.39 -0.41     1.44 0.38
Pope Francis                  13 2387 69.55 25.17 -0.73     0.14 0.52
Christian Fundamentalists     14 2387 48.59 28.48 -0.07    -0.72 0.58
Feminists                     15 2387 56.94 26.65 -0.24    -0.47 0.55
Liberals                      16 2387 52.27 27.35 -0.24    -0.67 0.56
Labor Unions                  17 2387 56.70 24.74 -0.27    -0.29 0.51
Poor People                   18 2387 72.20 19.63 -0.36    -0.06 0.40
Big Business                  19 2387 49.34 22.52 -0.15    -0.18 0.46
Conservatives                 20 2387 55.22 25.91 -0.24    -0.45 0.53
SCOTUS                        21 2387 59.34 19.38 -0.32     0.54 0.40
Gays & Lesbians               22 2387 62.83 26.86 -0.46    -0.20 0.55
Congress                      23 2387 41.17 22.32  0.02    -0.34 0.46
Rich People                   24 2387 53.53 20.69 -0.13     0.52 0.42
Muslims                       25 2387 55.80 25.64 -0.29    -0.23 0.52
Christians                    26 2387 74.40 23.80 -0.87     0.35 0.49
Jews                          27 2387 72.20 21.19 -0.45    -0.14 0.43
Tea Party                     28 2387 42.97 27.08 -0.06    -0.70 0.55
Police                        29 2387 75.57 22.50 -1.15     1.13 0.46
Transgender People            30 2387 57.29 26.88 -0.28    -0.31 0.55
Scientists                    31 2387 77.74 19.23 -0.77     0.39 0.39
BLM                           32 2387 48.26 32.66 -0.06    -1.15 0.67
```

Feeling Thermometers: Clinton vs. Trump



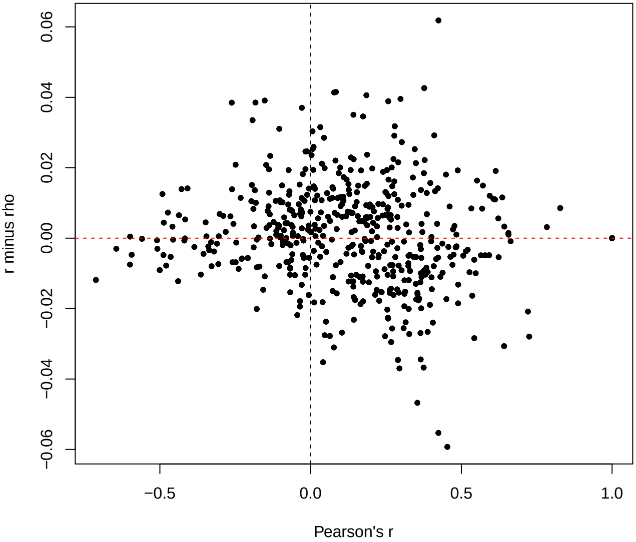

```
> rCT<-with(Therms, cor(‘Dem. Pres. Candidate‘,
           ‘GOP Pres. Candidate‘))
> rCT
[1] -0.71227

> rCT2<-with(Therms, cor.test(‘Dem. Pres. Candidate‘,
            ‘GOP Pres. Candidate‘))
> rCT2

 Pearson’s product-moment correlation

data:  Dem. Pres. Candidate and GOP Pres. Candidate
t = -49.6, df = 2385, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.73148 -0.69192
sample estimates:
     cor
-0.71227


> # Identical:
>
> (rCT*sqrt(nrow(Therms)-2)) / sqrt(1-(rCT^2))
[1] -49.557
```

Liberals and Conservatives

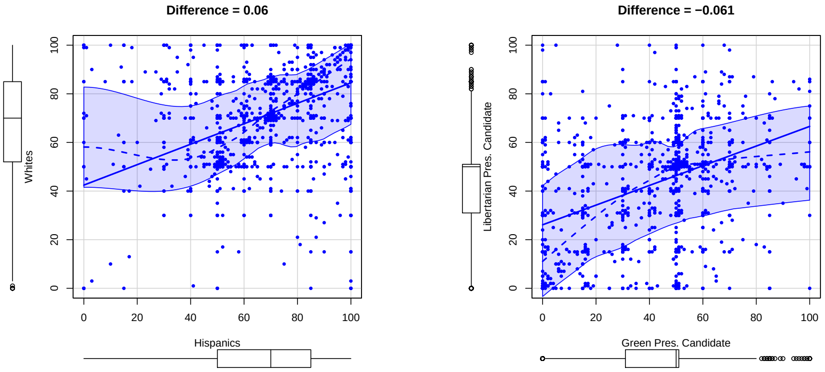



```
> rLC<-with(Therms, cor.test(Liberals,Conservatives))
> rLC

 Pearson's product-moment correlation

data:  Liberals and Conservatives
t = -28.2, df = 2385, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.52983 -0.46966
sample estimates:
     cor
-0.50035


> rhoLC<-with(Therms, SpearmanRho(Liberals,Conservatives))
> rhoLC
[1] -0.49128
```

Pairwise FT Differences between r and ρ



Biggest Differences Between r and ρ



Summary: Measures of Association

Which bivariate measure of association should I use?

		X			
		Nominal	Binary	Ordinal	Interval/Ratio
Y	Nominal	χ^2	χ^2	χ^2	t-test (and η)
	Binary	χ^2	ϕ, Q	γ, τ_c	t-test
	Ordinal	χ^2	γ, τ_c	γ, τ_a, τ_b	Spearman's ρ
	Interval / Ratio	t-test (and η)	t-test	Spearman's ρ	r