PLSC 502 – Autumn 2022
Measures of Association for Ordinal + Interval/Ratio Variables

November 10, 2022

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

Concordant and Discordant Pairs

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	
	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

Discordance with $\{1, 3\}$ 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

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)

γ and 2×2 Tables

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:

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

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":

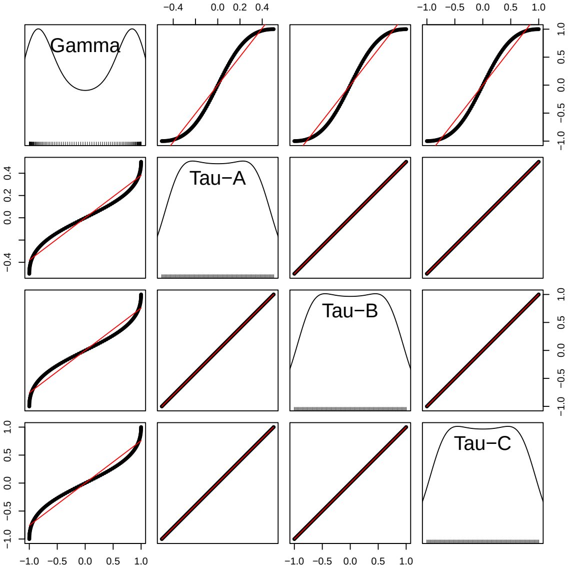
$$\tau_c = (N_c - N_d) \times \left\{ \frac{2m}{[N^2 2(m-1)]} \right\}$$

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

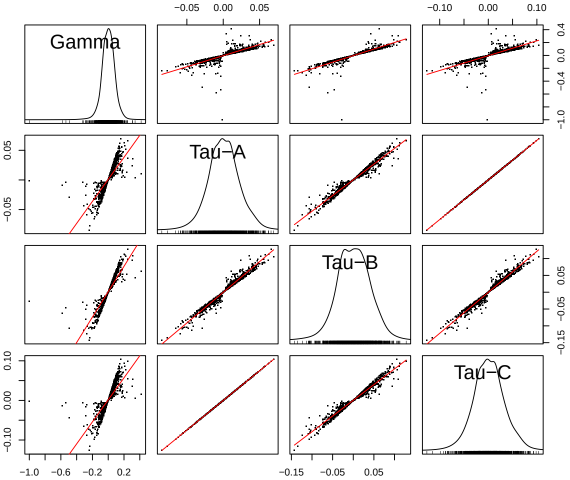
Tau tips:

- 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)



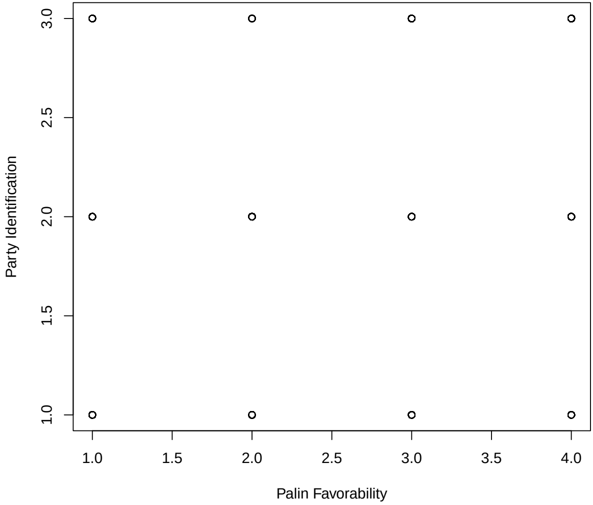
Example: Sarah Palin Support...

September 2008 "Battleground" Poll in PA:

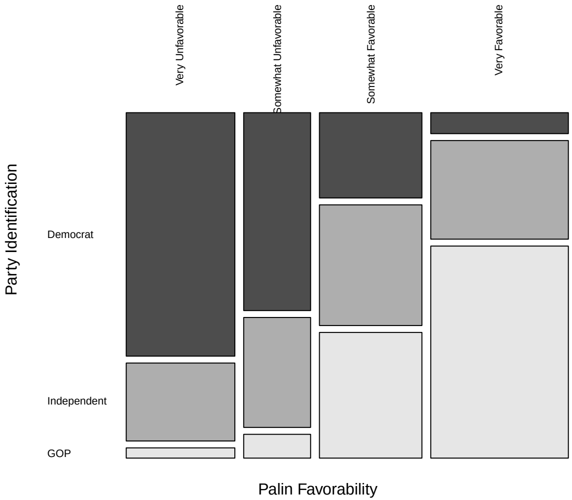
```
> summary(MamaGriz)
    caseid          female                          palin                    pid
 Min.   :    2   Male  :2221   Very Unfavorable    :1200   Democrat   :1709
 1st Qu.:30034   Female:2370   Somewhat Unfavorable: 739   Independent:1391
 Median :31831                 Somewhat Favorable  :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



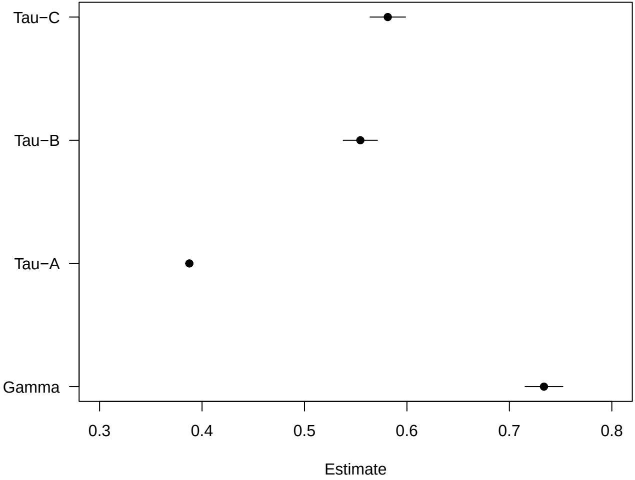
Plotting: Do


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

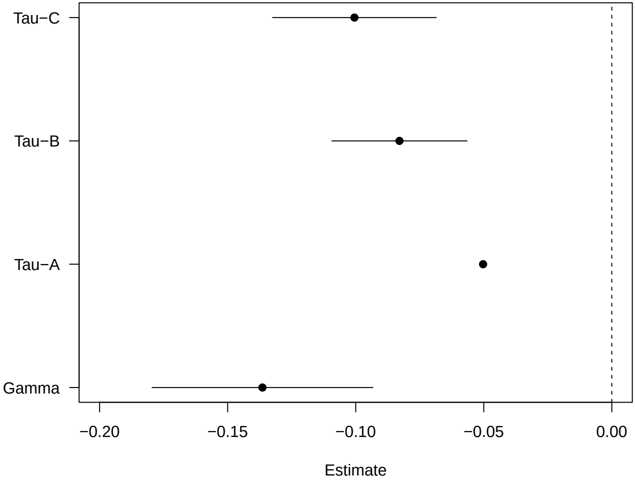
```
> GoodmanKruskalGamma(palinfemale,conf.level=0.95)
    gamma    lwr.ci    ups.ci
-0.136410 -0.179514 -0.093306
```

```
> 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 Data

Linearity

Linearity means:

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

$$Y = mX + b$$

Other monotonic + "smooth" alternatives:

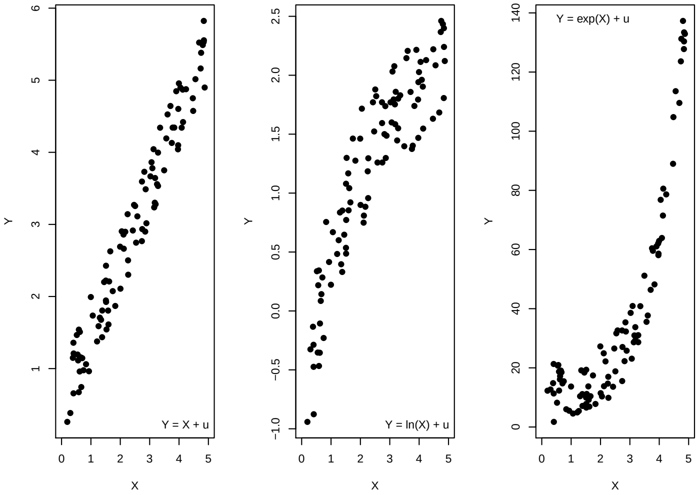
- *Logarithmic*:

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

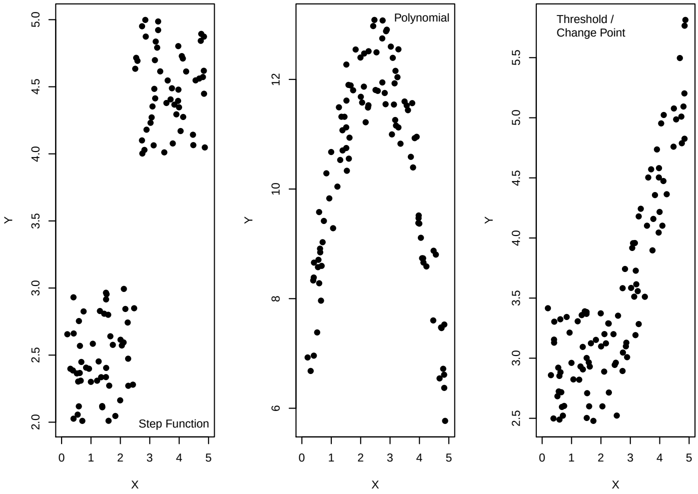
- *Exponential*:

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

Linear, Logarithmic, Exponential



Other Possibilities

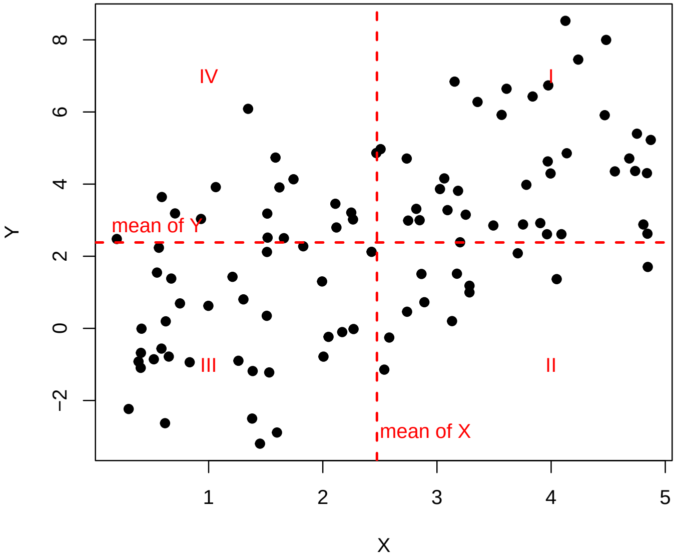


Linear Association: Pearson's r

"Pearson's product-moment correlation coefficient":

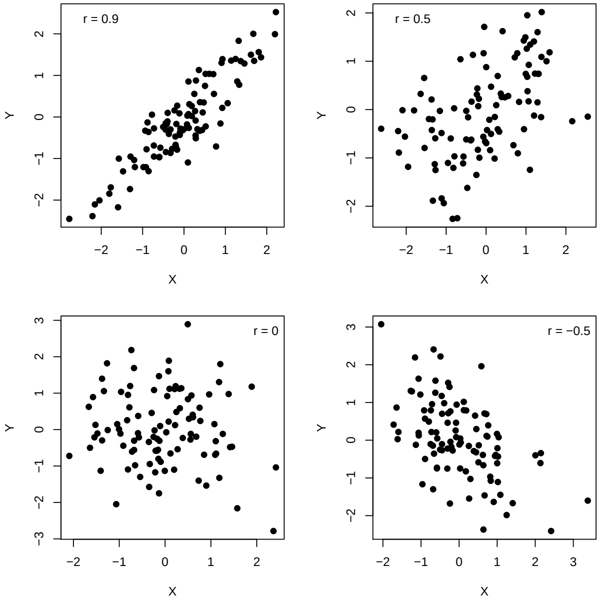
$$
\begin{aligned}
r &= \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \\
&= \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \\
&= \frac{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)}{N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

Pearson's r: Intuition

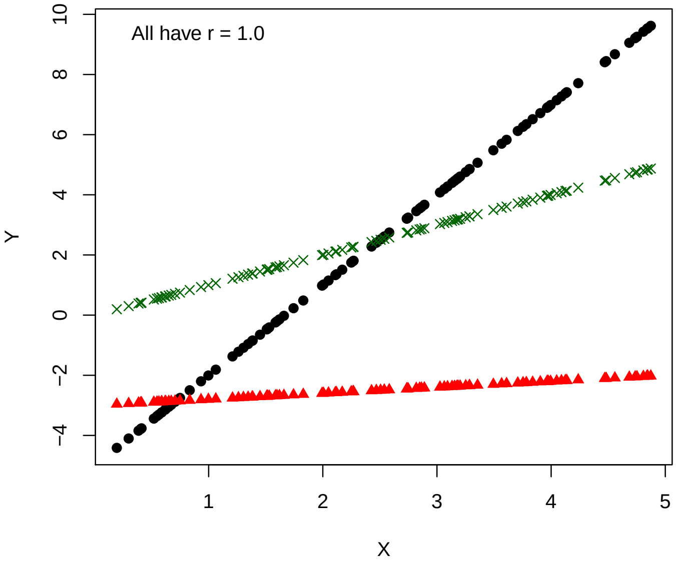


Pearson's r: Characteristics

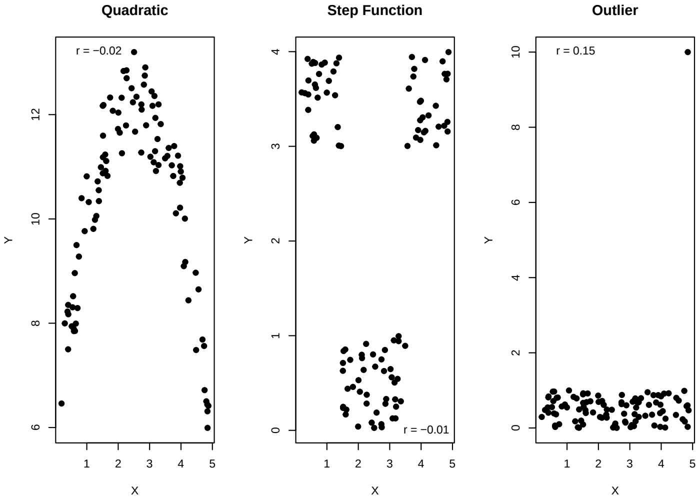
- $r \in [-1, 1]$
- $r = 0 \leftrightarrow$ no *linear* association between Y and X.
- Sign(r) $\rightarrow$ "direction" of the *linear* association
- $|r| \rightarrow$ "strength" of the *linear* association
- In general:
    - $|r| < 0.3 \rightarrow$ "weak" linear association
    - $0.3 < |r| < 0.7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


r = 0.9
r = 0.5
r = 0
r = −0.5
X
Y

$r = \pm 1.0 \rightarrow ?$



Nonlinearity, etc.



The sampling distribution of r is:

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

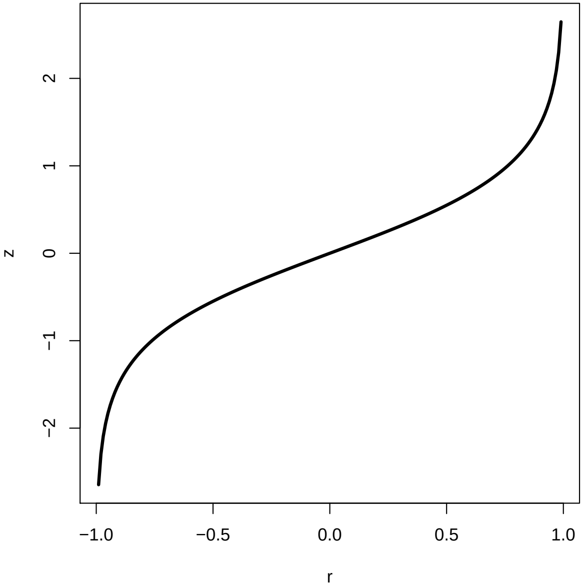
Fisher:

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

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

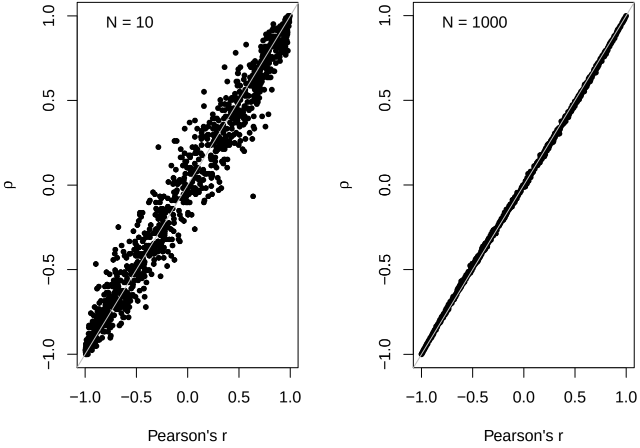
For sorted data on X and Y, where R_{Y_i} and R_{X_i} are the respective ranks,

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

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.

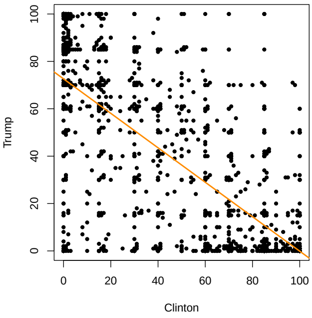
r vs. ρ Comparison (Simulation)



Real Data: ANES 2016 Feeling Thermometers

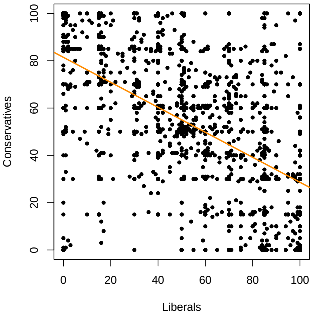
```
> describe(Therms,range=FALSE)
                             vars    n  mean    sd  skew kurtosis   se
Asian-Americans                 1 2387 70.17 20.20 -0.38     0.02 0.41
Hispanics                       2 2387 69.35 20.91 -0.41     0.01 0.43
Blacks                          3 2387 69.00 21.19 -0.35    -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



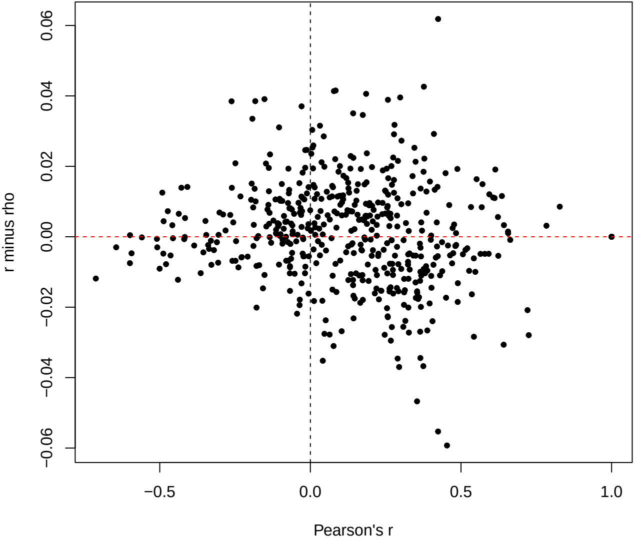

```
> rCT<-with(Therms, cor(‘Dem. Pres. Candidate‘,‘GOP Pres. Candidate‘))
> rCT
[1] -0.71227

> 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

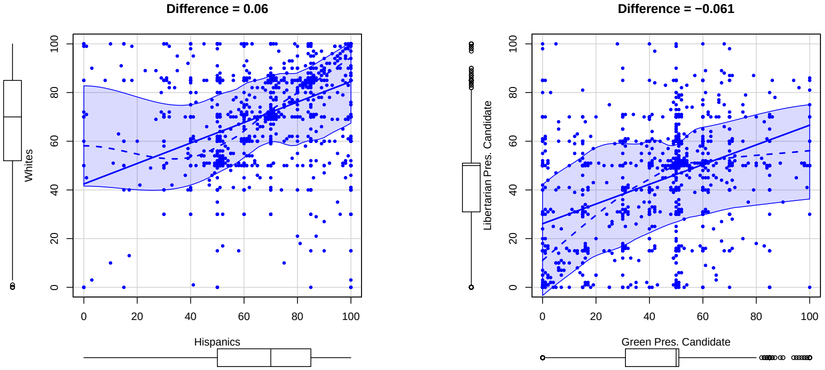



```
> rLC<-with(Therms, cor.test(Liberals,Conservatives))
> rLC

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