Regression III
Bios 662

Michael G. Hudgens, Ph.D.

mhudgens@bios.unc.edu

http://www.bios.unc.edu/∼mhudgens

2008-10-16 15:51

Outline

- Measures of association
- Parametric/large N
  - Pearson correlation coefficient
- Nonparametric (i.e., rank based)
  - Spearman rank correlation coefficient
  - Kendall's τ

Correlation

- The *correlation* between RVs X and Y is

$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

- Note:

$$\rho = \frac{\beta_{Y \cdot X}\sigma_X}{\sigma_Y} = \frac{\beta_{X \cdot Y}\sigma_Y}{\sigma_X}$$

Correlation

- Estimate ρ by

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (X_i - \bar{X})^2}} = \frac{[XY]}{\sqrt{[X^2][Y^2]}}$$

  sample Pearson product moment correlation coefficient

- Can show

$$r = \hat{\beta}_{Y \cdot X} \frac{s_X}{s_Y} = sign(\hat{\beta}_{Y \cdot X}) \sqrt{r^2}$$

  where r^2 is as in previous set of slides, i.e., the proportion of total variation attributable to regression

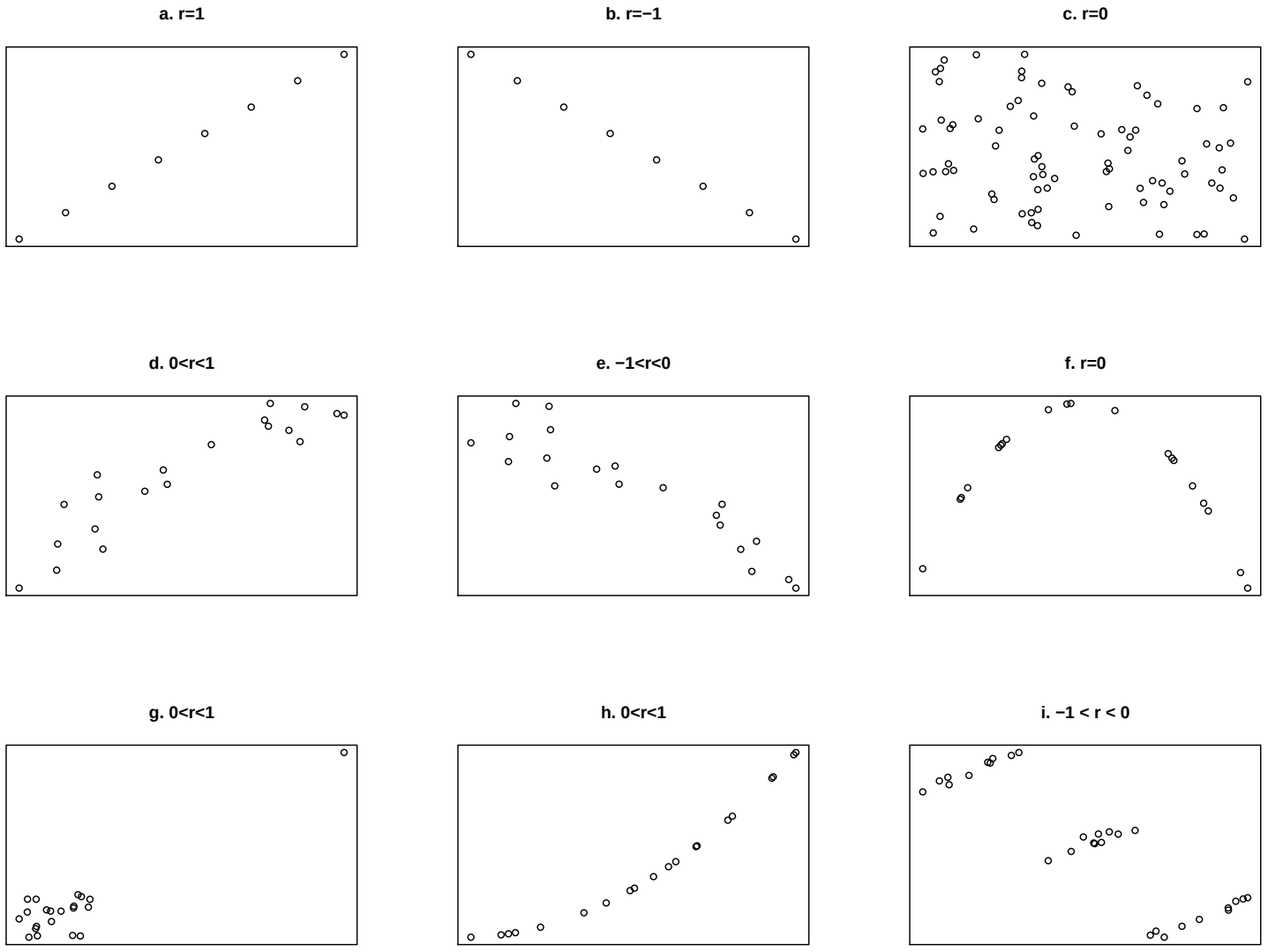
Correlation

- The correlation coefficient r has the following properties
- $r \in [-1, 1]$
- $r = 1$ iff all obs on a straight line with pos slope
- $r = -1$ iff all obs on a straight line with neg slope
- invariant under multiplication and addition of constants to X or Y
- measures linear association between two variables
- tends to be close to zero if no linear association

Correlation in R

```
> x <- 1:10
> y <- x

> cor(y,x)
[1] 1

> cor(y,3*x)
[1] 1

> cor(y,3*x+10)
[1] 1

> x <- x^2
> cor(y,x)
[1] 0.9745586

> cor(y/100,3*x+10)
[1] 0.9745586
```

Correlation: Figure 9.11



Correlation

- The test statistic

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} \sim t_{N-2}$$

  can be used to test $H_0 : \rho = 0$

- Claim: this test is equivalent to testing $H_0 : \beta_{Y.X} = 0$
- Proof of claim on next slides

Correlation

- First note

$$(N-2)s^2_{Y\cdot X} = SSE = SST - SSR$$

$$= [Y^2] - \frac{[XY]^2}{[X^2]}$$

$$= [Y^2]\left(1 - \frac{[XY]^2}{[Y^2][X^2]}\right)$$

$$= (N-1)s^2_Y(1-r^2)$$

- Next recall

$$\hat{\beta}_{Y\cdot X} = \frac{[XY]}{[X^2]}$$

Correlation

- Then

$$t = \frac{\hat{\beta}_{Y\cdot X}}{s_{Y\cdot X}/\sqrt{[X^2]}} = \frac{[XY]/[X^2]}{s_{Y\cdot X}/\sqrt{[X^2]}}$$

$$= \frac{[XY]/\sqrt{[X^2]}}{s_{Y\cdot X}} = \frac{r\sqrt{[Y^2]}}{s_{Y\cdot X}}$$

$$= \frac{r s_Y \sqrt{N-1}}{\sqrt{(1-r^2)s_Y^2(N-1)/(N-2)}}$$

$$= \frac{r}{\sqrt{(1-r^2)/(N-2)}}$$

Correlation

- In general,

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} \sim t_{N-2} \qquad (1)$$

  if

  1. (X, Y) bivariate normal (Sec 9.3.3 of text), or
  2. $Y|X$ normal with constant variance (i.e., the usual regression model holds)

- (1) holds approx for large N (cf Graybill 1976, Section 6.10)

Correlation

- Example: Cholesterol was measured in 100 spouse pairs
- If there is no environmental effect (e.g., diet) on cholesterol we would expect $\rho = 0$
- $H_0 : \rho = 0$ vs $H_A : \rho \neq 0$
- $t_{.975,98} = 1.98$, so $C_{0.05} = \{t : |t| > 1.98\}$
- Observe $r = .25$ such that

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} = \frac{.25}{\sqrt{(1 - .25^2)/98}} = 2.556$$

- $p = 2 * \{1 - F_{t,98}(2.556)\} = 0.0121$

Correlation: SAS

```
proc corr;  var x y;
```

Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.25000 0.0121
y	0.25000 0.0121	1.00000

Correlation using Fisher's Transformation

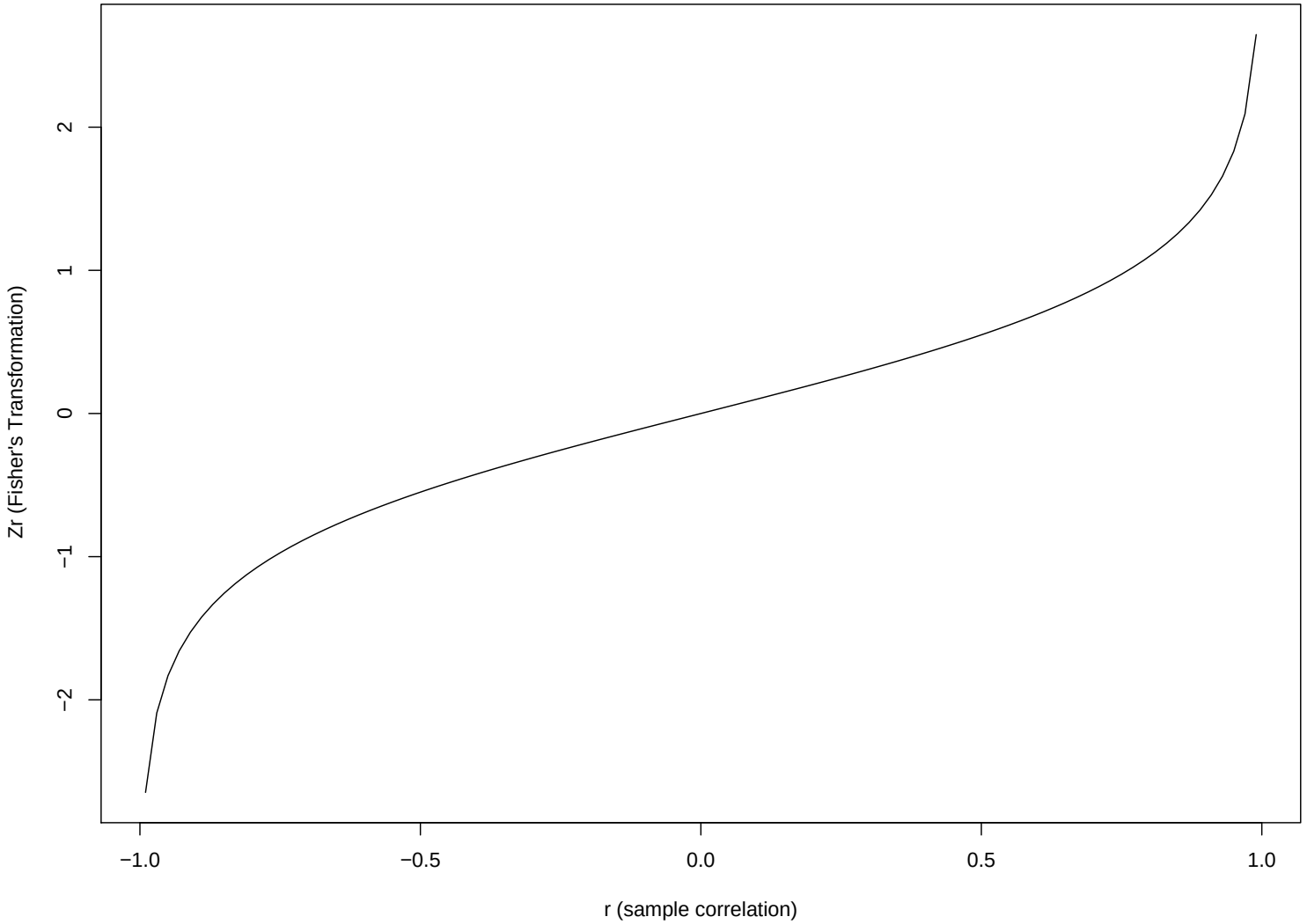
- R. A. Fisher developed a test of $H_0 : \rho = \rho_0$
- He showed

$$z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{N-3}\right)$$

- Under $H_0 : \rho = \rho_0$

$$z = \frac{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{1/(N-3)}} \sim N(0,1)$$

Correlation: Fisher's Transformation



Correlation using Fisher's Transformation: Example

- Cholesterol example
- $N = 100$, $r = .25$
- $H_0 : \rho = 0$

$$z_r = \frac{1}{2}\log\left(\frac{1.25}{.75}\right) = .2554$$

$$z = \frac{.2554 - 0}{\sqrt{1/97}} = 2.515$$

$$p = 2 * \{1 - \Phi(2.515)\} = 0.0119$$

Correlation using Fisher's Transformation

- The Fisher transformation can be used for a CI for ρ

$$z_L = z_r - z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$z_U = z_r + z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$r_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}; r_U = \frac{e^{2z_U}-1}{e^{2z_U}+1}$$

Correlation using Fisher's Transformation: Example

- 95% CI when $r = .25$ and $n = 100$

$$(z_L, z_U) = .2554 \pm 1.96/\sqrt{97} = (0.0564, 0.4544)$$

$$r_L = \frac{e^{2*0.0564} - 1}{e^{2*0.0564} + 1} = 0.05635$$

$$r_U = \frac{e^{2*0.4544} - 1}{e^{2*0.4544} + 1} = 0.4255$$

Correlation using Fisher's Transformation: SAS

```
proc corr fisher(biasadj=no);
  var x y;
run;
```

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z
x	y	100	0.25000	0.25541

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	95% Confidence Limits		p Value for H0:Rho=0
x	y	0.056350	0.425524	0.0119

Correlation using Fisher's Transformation: R

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 2.556, df = 98, p-value = 0.01212
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.05634962 0.42552363
sample estimates:
      cor
0.2500007
```

Correlation using Fisher's Transformation

- Comparing 2 correlations: 2 independent samples

$$H_0 : \rho_1 = \rho_2 \text{ vs } H_A : \rho_1 \neq \rho_2$$

- Compute z_{r_1} and z_{r_2}

$$V(z_{r_1} - z_{r_2}) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

- Thus under H_0

$$z = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \sim N(0, 1)$$

Correlation using FT: Example

- Example: If blood pressure level is inherited, one would expect the correlation between blood pressure of mothers and their natural children to be greater than between mothers and their adopted children
- In a study, 1000 mothers and one of their randomly choosen natural children had their blood pressure measured.
- In a separate sample, 100 mothers and their adopted children also had their BP measured

Correlation using FT: Example

- Let

  ρ_1 = popn correlation for natural pairs

  ρ_2 = popn correlation for adopted pairs

- Hypotheses

$$H_0 : \rho_1 = \rho_2 \text{ vs } H_A : \rho_1 > \rho_2$$

- Critical region

$$C_{.05} = \{z : z > 1.645\}$$

Correlation using FT: Example

- $r_1 = 0.32$; $r_2 = 0.06$
- $z_{r_1} = 0.3316$; $z_{r_2} = 0.0601$
- Thus

$$z = \frac{.3316 - .0601}{\sqrt{\frac{1}{997} + \frac{1}{97}}} = 2.55$$

Correlation Homogeneity

- Testing the homogeneity of k correlations
- Fisher's transformation can be used to test the hyp that several correlations are equal

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$$

vs

$$H_A : \text{ at least one } \neq$$

Correlation Homogeneity

- Let

$$T_1 = \sum_{i=1}^{k} (n_i - 3) z_{r_i}$$

and

$$T_2 = \sum_{i=1}^{k} (n_i - 3) z_{r_i}^2$$

- Under H_0

$$H = T_2 - \frac{T_1^2}{\sum (n_i - 3)} \sim \chi^2_{k-1}$$

- Cf. Graybill (1976, p. 405)

Correlation Homogeneity: Example

- Does the correlation between LDL-cholesterol and HDL-cholesterol change with age in women not taking hormones?

Age	n	r	z_r
20-29	277	-.08	-.0802
30-39	479	-.25	-.2554
40-49	508	-.19	-.1923
50-59	373	-.18	-.1820
60-69	216	-.15	-.1511

Correlation Homogeneity: Example

- Null hyp $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$
- Critical region $C_{.05} = \{H : H > 9.49\}$
- Compute test statistic

$$T_1 = 274(-.0802) + \cdots + 213(-.1511) = -340.181$$

$$T_2 = 274(-.0802)^2 + \cdots + 213(-.1511)^2 = 68.605$$

$$H = 68.605 - \frac{(-340.181)^2}{1853} = 6.15$$

Rank Correlation Coefficients

- Using ranks makes robust to outliers
- Spearman Rank Correlation, Kendall's τ
- Nonparametric measure of association

Spearman Rank Correlation

1. Y's and X's are ranked from 1 to N separately
2. The correlation of the ranks is then computed

Spearman Correlation

- Example: 10 children are ranked according to their mathematics and muscial abilities

Child	Math	Music
A	7	5
B	4	7
C	3	3
D	10	10
E	6	1
F	2	9
G	9	6
H	8	2
I	1	8
J	5	4

Spearman Correlation

- Let R_{1i} and R_{2i} be the ranks of Y_i and X_i
- SC test statistic

$$r_s = \frac{\sum(R_{1i}-\bar{R}_1)(R_{2i}-\bar{R}_2)}{\sqrt{\sum_i(R_{1i}-\bar{R}_1)^2 \sum_i(R_{2i}-\bar{R}_2)^2}}$$

$$= 1 - \frac{6\sum d_i^2}{N^3-N}$$

  where $d_i = R_{1i} - R_{2i}$

- Form of r_s with $\sum d_i^2$ not correct if ties are present
- Note $R_{1i} = R_{2i}$ for all $i \Rightarrow d_i = 0$ for all $i \Rightarrow r_s = 1$

Spearman Correlation

- Suppose N is odd and $N = 2m + 1$
- Then the most extreme discordant rankings are

i	1	2	$\cdots$						N
R_{1i}	1	2	$\cdots$	m	$m+1$	$m+2$	$\cdots$	$2m$	$2m+1$
R_{2i}	$2m+1$	$2m$	$\cdots$	$m+2$	$m+1$	m	$\cdots$	2	1
d_i	$-2m$	$2-2m$	$\cdots$	-2	0	2	$\cdots$	$2m-2$	$2m$

Spearman Correlation

- Under this configuration

$$\begin{aligned}\sum_{i=1}^{N} d_i^2 &= 4m^2 + 4(m-1)^2 + \cdots 2^2 \\ &\quad +2^2 + \cdots + 4(m-1)^2 + 4m^2 \\ &= 8\sum_{j=1}^{m} j^2 \\ &= 8m(m+1)(2m+1)/6 \\ &= \{4 * (N-1)/2 * (N+1)/2 * N\}/3 \\ &= (N^3 - N)/3\end{aligned}$$

Spearman Correlation

- Thus

$$r_s = 1 - \frac{6(N^3 - N)}{3(N^3 - N)} = 1 - 2 = -1$$

- In a similar way, it can be shown that if N is even, the most extreme rankings give $r_S = -1$
- So $r_s = 1$ if perfect agreement in the ranks

  $r_s = -1$ if perfect disagreement in the ranks

Spearman Correlation

Child	Math	Music	d
A	7	5	2
B	4	7	-3
C	3	3	0
D	10	10	0
E	6	1	5
F	2	9	-7
G	9	6	3
H	8	2	6
I	1	8	-7
J	5	4	1

- Spearman correlation

$$r_s = 1 - \frac{6(2^2 + (-3)^2 + \cdots + 1^2)}{10^3 - 10} = \frac{6(182)}{990} = -.103$$

Spearman Correlation: SAS and R

```
proc corr spearman;  var math music; run;
```

Spearman Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	math	music
math	1.00000	-0.10303 0.7770
music	-0.10303 0.7770	1.00000

```
> cor(math,music,method="spearman")
[1] -0.1030303
```

Spearman Correlation

- SC used to test null hyp of independence

$$H_0 : X \perp Y \text{ vs } H_A : X \not\perp Y$$

  i.e. $H_A : X$ and Y not independent

- Distribution of r_s under H_0 derived using permutation-based argument
- We can list the R_{1i} in ascending order
- There are $N!$ possible orderings of R_{2i}
- Under H_0, each of these orderings is equally likely

Spearman Correlation

- Example $N = 3$

R_{1i}	1	2	3	$\sum d_i^2$	r_s
R_{2i}	1	2	3	0	1
R_{2i}	1	3	2	2	.5
R_{2i}	2	1	3	2	.5
R_{2i}	2	3	1	6	-.5
R_{2i}	3	1	2	6	-.5
R_{2i}	3	2	1	8	-1

Spearman Correlation

- CDF of r_s

k	$\Pr[r_s \leq k]$
-1	1/6
-.5	1/2
.5	5/6
1	1

Spearman Correlation

- Text Table A.12, p. 838, gives the two sided critical values for testing $H_0 : X \perp Y$
- If N is large (> 10 Neter et al. 1996, page 652),

$$t_s = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t_{N-2}$$

Spearman Correlation: Example

- Example: math (X) and music (Y)
- $N = 10$; $r_s = -0.1030$
- From Table A.12

$$C_{0.05} = \{r_s : |r_s| > .648\}$$

Spearman Correlation: Example

- Assume $N = 10$ is large enough to use t approximation
- $C_{0.05} = \{t_s : |t_s| > t_{8,.975} = 2.306\}$
- $t_s = \frac{-.1030\sqrt{8}}{\sqrt{1-(-.1030)^2}} = -.2929$
- $p = 2 * \Pr[t_8 < -.2929] = 0.777$

Spearman Correlation: Ties

- In the presence of ties, ranks are replaced by midranks
- However, critical values in Table A.12 are only approximate
- If N is large, use t_s as before; i.e.,

$$t_s = \frac{r_s\sqrt{N-2}}{\sqrt{1-r_s^2}} \sim t_{N-2}$$

Kendall's τ

- Kendall's τ: another rank correlation statistic
- Data: (X_i, Y_i) for $i = 1, 2, \ldots, N$
- Definitions: Two pairs of observations are

  concordant if $(X_i - X_j)(Y_i - Y_j) > 0$

  discordant if $(X_i - X_j)(Y_i - Y_j) < 0$

Kendall's τ

- Let p_c = the probability that a randomly chosen pair of observations is concordant; and p_d = the prob discordant

$$\tau = p_c - p_d$$

- Note: $-1 \leq \tau \leq 1$; if X and Y are independent, $\tau = 0$

Kendall's τ

- Note that there are $\binom{N}{2}$ pairs of observations
- Let P be the number of concordant pairs
- Let Q be the number of discordant pairs
- The estimate of τ is

$$r_k = \frac{P-Q}{\binom{N}{2}} = 1 - \frac{2Q}{\binom{N}{2}} = \frac{2P}{\binom{N}{2}} - 1$$

- Replacing X and Y's with ranks does not change τ

Kendall's τ

- $H_0 : \tau = 0$ vs. $H_A : \tau \neq 0$
- The distribution of r_k under H_0 is computed using permutation principles
- As with r_s, there are $N!$ equally likely outcomes
- Kendall, *Rank Correlation Methods*, Hafner Publishing, 1962 gives a table of the distribution of $P - Q$ for $4 \leq N \leq 10$

Kendall's τ

- Upper one-sided critical values of r_k
- Note the distn of r_k is symmetric about 0

N	.05	.025
5	.8	1.00
6	.73	.87
7	.62	.71
8	.57	.64
9	.50	.56
10	.42	.51

Kendall's τ

- Cigarette consumption and lung cancer mortality in England and Wales, 1930-1969

Period	$\log_{10}$ mortality	$\log_{10}$ tobacco (lb / person)
1930-34	-2.35	-.26
1935-39	-2.20	-.03
1940-44	-2.12	.30
1945-49	-1.95	.37
1950-54	-1.85	.40
1955-59	-1.80	.50
1960-64	-1.70	.55
1965-69	-1.58	.55

Kendall's τ

- $C_{0.05} = \{r_k : |r_k| \geq 0.64\}$
- Obs 1: (-2.35, -.26), Obs 2: (-2.20, -.03)

  $\{-2.35 - (-2.2)\}\{-2.6 - (-.03)\} > 0 \Rightarrow$ concordant
- Obs 1 and Obs 3:

  $\{-2.35 - (-2.12)\}(-2.6 - .3) > 0 \Rightarrow$ concordant
- $P - Q = 27 \Rightarrow$

$$r_k = \frac{27}{\binom{8}{2}} = \frac{27}{28} = 0.96$$

Kendall's τ

- If N is sufficiently large (≥ 10), under $H_0 : \tau = 0$

$$r_k \sim N\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$$

$$P - Q \sim N\left(0, \frac{N(N-1)(2N+5)}{18}\right)$$

or

$$Z = \frac{P-Q}{\sqrt{\frac{N(N-1)(2N+5)}{18}}} \sim N(0,1)$$

Kendall's τ

- If there are tied observations, r_k cannot be 1 or -1.
- Let

$$t_x = \frac{1}{2}\sum_i t_{xi}(t_{xi}-1) \text{ and } t_y = \frac{1}{2}\sum_i t_{yi}(t_{yi}-1)$$

where t_{zi} denotes the number of observations in the i^{th} set of ties for $z = x, y$

Kendall's τ

- Let

$$W = \sqrt{\left\{\frac{1}{2}N(N-1) - t_x\right\}\left\{\frac{1}{2}N(N-1) - t_y\right\}}$$

- Then

$$r_{k_b} = \frac{P - Q}{W}$$

This statistic is known as *Kendall's* τ_b

Kendall's τ: Revisit tobacco example

- Recall $N = 8$ and there was one set of ties (of size 2) for the tobacco variable
- Thus

$$W = \sqrt{\left\{\frac{1}{2}8(8-1)\right\}\left\{\frac{1}{2}8(8-1)-1\right\}}$$

- Yielding

$$r_{k_b} = \frac{27}{\sqrt{27 * 28}} = 0.98198$$

This statistic is known as *Kendall's* τ_b

Kendall's τ: Revisit tobacco example

- SAS

```
proc corr kendall;
  var mortality tobacco;
run;
```

Kendall Tau b Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	mortality	tobacco
mortality	1.00000	0.98198 0.0008
tobacco	0.98198 0.0008	1.00000

- R

```
cor(mortality, tobacco, method="kendall")
cor.test(mortality, tobacco, method="kendall")
```

Kendall's τ

- Kendall's score $P - Q$

$$r_{k_a} = \frac{P - Q}{\binom{N}{2}}$$

and

$$r_{k_b} = \frac{P - Q}{W}$$

- Tests based on r_{k_a} and r_{k_b} are equivalent
- Asymptotic variance of $P - Q$ under H_0 is given on page 336 of text

$$Z = \frac{P - Q}{\sqrt{V(P - Q)}} \sim N(0, 1)$$

Kendall's τ: Example

- In general, $V(P-Q)$ equals

$$\frac{N(N-1)(2N+5)}{18} - \sum_i \frac{t_{xi}(t_{xi}-1)(2t_{xi}+5)}{18} - \ldots$$

- For tobacco example, $V(P-Q)$ equals

$$\frac{8(8-1)(2*8+5)}{18} - \frac{2(2-1)(2*2+5)}{18} - 0 = 64.333$$

- Thus

$$z = \frac{27}{\sqrt{64.333}} = 3.366$$

yielding $p = 2 * \{1 - \Phi(3.366)\} = 0.0008$

Correlation: Summary/Remarks

- r appropriate if (X, Y) bivariate normal; sensitive to outliers, major(?) departures from normality
- Nonparametric alternatives: r_s and r_k
- If (X, Y) bivariate normal with correlation ρ,

$$r \rightarrow^p \rho$$

$$r_s \rightarrow^p \frac{6}{\pi} sin^{-1}(\rho/2)$$

$$r_k \rightarrow^p \frac{2}{\pi} sin^{-1}(\rho)$$

  (Kraemer 1998 "Rank Correlation" *Encyc of Bios*)
- ARE of r_s and r_k compared to r: $9/\pi^2 = 0.912$ (Conover 1980 *Practical Nonparametric Statistics*)