Section 5.4: Measures of Rank Correlation

• **Correlation is used in cases of paired data, to describe the <u>association</u> between the two random variables, say *X* and *Y*.**

For all measures of correlation:

• **The correlation is always between -1 and 1.**
• **Positive correlation => The two variables are <u>positively associated</u> (large values of one variable correspond to large values of the other variable)**
• **Negative correlation => The two variables are <u>negatively associated</u> (large values of one variable correspond to small values of the other variable)**
• **Correlation near 0 => large values of one variable tend to appear randomly with either large or small values of the other variable.**

How far the correlation is from 0 measures the *strength* of the relationship:

• **nearly 1 => Strong positive association between the two variables**
• **nearly -1 => Strong negative association between the two variables**
• **near 0 => Weak association between the two variables**

• **When the correlation is zero, this sometimes (but not always) means that *X* and *Y* are <u>independent</u>.**

• The <u>Pearson (product-moment) correlation coefficient</u> (denoted r) is a numerical measure of the <u>strength</u> and <u>direction</u> of the <u>linear</u> relationship between two variables.

Formula for r (the Pearson correlation coefficient between two paired data sets $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$):

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]^{1/2}} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)^{1/2}\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)^{1/2}}$$

This is the same as:

$$r = \frac{\text{sample covariance of } X_i\text{'s and } Y_i\text{'s}}{(\text{sample std. dev. of } X_i\text{'s})(\text{sample std. dev. of } Y_i\text{'s})}$$

• If the bivariate distribution of (X, Y) is unknown, then the Pearson correlation coefficient cannot be used for hypothesis tests and confidence intervals.

Spearman Correlation Coefficient

• An alternative measure of correlation simply ranks the two samples (<u>separately</u>, not combined) and calculates the Pearson measure on the ranks $R(X_i)$ and $R(Y_i)$ rather than on the actual data values.

• This produces the <u>Spearman Correlation Coefficient</u>.

• **Since the average of the *n* ranks (1, 2, …, *n*) in each sample is:** $\frac{n+1}{2}$

the formula for the Spearman Correlation Coefficient is

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}\left(\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}}$$

• **We can use Spearman's ρ as a test statistic to test whether *X* and *Y* are independent.**

Null Hypothesis:

H_0: The X_i's and Y_i's are mutually independent

3 Possible Alternatives

Two-Tailed	Lower-Tailed	Upper-Tailed
H_1: The X and Y variables are associated (either positively or negatively)	H_1: The X and Y variables are negatively associated	H_1: The X and Y variables are positively associated

• **The exact null distribution of ρ is tabulated (for $n \leq 30$) in Table A10. Note** $w_{1-p} = -w_p$

• **For larger sample sizes (or with many ties), the approximate quantiles may be used:**

$$w_p \approx \frac{z_p}{\sqrt{n-1}}$$

where z_p is a standard normal quantile.

<u>Decision Rules</u>

Two-tailed	**Lower-tailed**	**Upper-tailed**
Reject H_0 if $\lvert\rho\rvert > w_{1-\alpha/2}$	Reject H_0 if $\rho < w_\alpha = -w_{1-\alpha}$	Reject H_0 if $\rho > w_{1-\alpha}$
from Table A10	from Table A10	from Table A10

• **Approximate <u>P-values</u> can be obtained from the normal distribution using one of equations (12)-(14) on pp. 317-318, or by interpolating within Table A10, but we will typically use software to get approximate P-values.**

Example: The GMAT score and GPA for 12 MBA graduates are given on p. 316. Is there evidence of positive correlation between GMAT and GPA?

From R, Spearman's $\rho = 0.59$.

H_0: GMAT and GPA independent

H_1: GMAT and GPA positively associated

Reject H_0 if $\rho > w_{.95} = .4965$ ← Table A10, $n=12$.

Since $0.59 > .4965$, we reject H_0 and conclude GMAT and GPA have positive correlation.

On computer: Use `cor.test` function in R with `method="spearman"` (see code on course web page).

From R, P-value $\approx .0217$

Kendall's Tau

• **Another measure of correlation, Kendall's Tau, is based on the idea of <u>concordant</u> and <u>discordant</u> pairs.**

• **Consider two bivariate observations, say, (X_i, Y_i) and (X_j, Y_j).**

• **The two observations are <u>concordant</u> if both numbers in one observation are larger than the corresponding numbers in the other observation.**

• **The two observations are <u>discordant</u> if the numbers in observation *i* differ in opposite directions as the corresponding numbers in observation *j*.**

Examples:

If $X_i < X_j$ and $Y_i < Y_j$, then the *i*-th and *j*-th observations are: Concordant

If $X_i < X_j$ and $Y_i > Y_j$, then the *i*-th and *j*-th observations are: discordant

If $X_i > X_j$ and $Y_i < Y_j$, then the *i*-th and *j*-th observations are: discordant

If $X_i > X_j$ and $Y_i > Y_j$, then the *i*-th and *j*-th observations are: Concordant

Let N_c = the number of concordant pairs

and N_d = the number of discordant pairs

• **There are** $\binom{n}{2} = \frac{n(n-1)}{2}$ **possible pairs of bivariate observations.**

• **If there are no ties (no cases when $X_i = X_j$ or $Y_i = Y_j$), then**

$$\tau = \frac{N_c - N_d}{n(n-1)/2}$$

• **A general definition of Kendall's tau that allows for ties is**

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

where we compute N_c and N_d by:

If $\frac{Y_j - Y_i}{X_j - X_i} > 0 \Rightarrow$ add 1 to N_c (concordant)

If $\frac{Y_j - Y_i}{X_j - X_i} < 0 \Rightarrow$ add 1 to N_d (discordant)

If $\frac{Y_j - Y_i}{X_j - X_i} = 0 \Rightarrow$ add $\frac{1}{2}$ to N_c and $\frac{1}{2}$ to N_d

If $X_i = X_j \Rightarrow$ ignore this pair

Examples on p. 316 data:

Pair: 1+2 : $\frac{4.0 - 4.0}{610 - 710} = 0 \Rightarrow$ add $\frac{1}{2}$ to N_c and $\frac{1}{2}$ to N_d

1+3 : $\frac{3.9 - 4.0}{640 - 710} > 0 \Rightarrow$ add 1 to N_c

... Continue for all pairs, where $i < j$.

• **We can use** $T = N_c - N_d$

as a test statistic to test for independence of X **and** Y**.**

Null Hypothesis:

H_0: The X_i and Y_i are mutually independent

3 Possible Alternatives

Two-Tailed	Lower-Tailed	Upper-tailed
H_1: The X and Y variables are associated (either positively or negatively)	H_1: Negative association (smaller values of X correspond to larger values of Y and vice versa)	H_1: Positive association (larger values of X correspond to larger values of Y)

• **The exact null distribution of** T **is tabulated (for** $n \leq 60$**) in Table A11. Note** $w_{1-p} = -w_p$

• **For larger sample sizes (or with many ties), the quantile for** T **is approximately:**

$$w_p = z_p \sqrt{n(n-1)(2n+5)/18}$$

where z_p **is a standard normal quantile.**

Decision Rules

Two-tailed	Lower-tailed	Upper-tailed
Reject H_0 if $T < W_{\alpha/2}$ or if $T > W_{1-\alpha/2}$	Reject H_0 if $T < W_\alpha$	Reject H_0 if $T > W_{1-\alpha}$

Table A11

• **Approximate P-values can be obtained from the normal distribution using one of equations (20)-(21) on p. 322, or by interpolating within Table A11, but we will typically use software to get approximate P-values.**

Example: Recall the GMAT score and GPA for 12 MBA graduates on p. 316. Is there evidence of positive correlation between GMAT and GPA? H_1

H_1: Positive association between GMAT + GPA

Reject H_0 if $T > W_{.95} = 24$ ← Table A11 with $n=12$

$T = N_c - N_d = 44.5 - 17.5 = 27$ ← tedious to find by hand

Since $27 > 24$, reject H_0 and conclude positive association between GMAT and GPA.

– Note $\tau = \frac{44.5-17.5}{44.5+17.5} = .4355$ (R gives .439 and gives us an approximate P-value = .0289) (uses method based on ties)

On computer: Use `cor.test` function in R with `method="kendall"` (see code on course web page).

Daniels Test for Trend

• The Daniels Test is a more powerful test for trend than the Cox-Stuart Test from Chapter 3.

• If we have a time-ordered sample $X_1, \ldots, X_n$, we create paired data: $(\text{Time}_1, X_1), \ldots, (\text{Time}_n, X_n)$.

• Then the test of independence based on Spearman's rho or Kendall's tau is performed, with

H_0: no trend

and the possible alternatives being:

H_1: either an increasing or decreasing trend	H_1: decreasing trend	H_1: increasing trend

Example on global temperature data again: Is there evidence of an increasing temperature trend?

H_0: no trend vs. H_1: increasing trend

Time: $(1, 2, \ldots, 13)$

X: $(-.493, -.457, \ldots, .923)$

Spearman's $\rho = .929$ (P-value of test ≈ 0)

Kendall's $\tau = .821$ (P-value ≈ 0)

– Reject H_0 and conclude there is an increasing temperature trend.

Comparison to Competing Tests

• If the distribution of X and Y is bivariate normal, a t-test based on Pearson's correlation coefficient is used to test for independence.

• The A.R.E. of the tests based on Spearman's and Kendall's measures relative to that t-test are each 0.912 when the data are bivariate normal.

• However, the nonparametric tests can have better efficiency than the t-tests for many nonnormal distributions.

• These nonparametric tests only require the data to be continuous, rather than requiring normality.

• As measures of correlation, Spearman's rho and Kendall's tau are appropriate as long as the data are at least ordinal on the measurement scale.

• Kendall's tau is often used as a measure of association when the data are binary and ~~ordered~~ paired (for example, Fail/Pass).

Example: 20 students each took both a Pass-Fail test in Math and a Pass-Fail test in History. Describe the association between the two tests.

$\tau = .492 \Rightarrow$ In this sample, there is moderate positive association between the math and history tests.