Section 5.4: Measures of Rank Correlation

- Correlation is used in cases of paired data, to describe the association between the two random variables, say $X$ and $Y$.

For all measures of correlation:

- The correlation is always between -1 and 1.
- Positive correlation $\Rightarrow$ The two variables are positively associated (large values of one variable correspond to large values of the other variable)
- Negative correlation $\Rightarrow$ The two variables are negatively associated (large values of one variable correspond to small values of the other variable)
- Correlation near 0 $\Rightarrow$ 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 $\Rightarrow$ Strong positive association between the two variables
- nearly -1 $\Rightarrow$ Strong negative association between the two variables
- near 0 $\Rightarrow$ Weak association between the two variables

- When the correlation is zero, this sometimes (but not always) means that $X$ and $Y$ are independent.
• The Pearson (product-moment) correlation coefficient (denoted \( r \)) is a numerical measure of the strength and direction of the linear relationship between two variables.

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

\[
\begin{align*}
\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} x_i y_i - n \overline{x} \overline{y} \\
\left[ \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}{(\sum_{i=1}^{n} x_i^2 - n \overline{x}^2)(\sum_{i=1}^{n} y_i^2 - n \overline{y}^2)} \right]^{\frac{1}{2}}
\end{align*}
\]

This is the same as:

\[
\frac{\text{sample covariance of } x_i \text{'s and } y_i \text{'s}}{\text{(sample std. dev. of } x_i \text{'s}) \times \text{(sample std. dev. of } y_i \text{'s})}
\]

• 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 (separately, 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 **Spearman Correlation Coefficient**.
• 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 $\rho$ 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

<table>
<thead>
<tr>
<th>Two-Tailed</th>
<th>Lower-Tailed</th>
<th>Upper-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$: The $X$ and $Y$ variables are associated (either positively or negatively)</td>
<td>$H_1$: The $X$ and $Y$ variables are negatively associated</td>
<td>$H_1$: The $X$ and $Y$ variables are positively associated</td>
</tr>
</tbody>
</table>

• The exact null distribution of $\rho$ is tabulated (for $n \leq 30$) in Table A10. Note $w_{1-\rho} = -\frac{1}{\sqrt{30}}$
• 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.

<table>
<thead>
<tr>
<th>Decision Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-tailed</td>
</tr>
<tr>
<td>Lower-tailed</td>
</tr>
<tr>
<td>Upper-tailed</td>
</tr>
<tr>
<td>Reject Ho if</td>
</tr>
<tr>
<td>(</td>
</tr>
<tr>
<td>( p &lt; W_\alpha = -W_{1-\alpha} ) from Table A10</td>
</tr>
<tr>
<td>Reject Ho if</td>
</tr>
<tr>
<td>( p &gt; W_{1-\alpha} )</td>
</tr>
</tbody>
</table>

• Approximate P-values 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 Ho if \( p > W_{0.95} = 0.4965 \) from Table A10, \( n=12 \).

Since \( 0.59 > 0.4965 \) we reject \( H_0 \) and conclude GMAT and GPA have positive correlation.

On computer: Use \texttt{cor.test} function in R with method=“spearman” (see code on course web page).

From R, \( p\)-value \( \approx 0.0217 \).
Kendall’s Tau

- Another measure of correlation, Kendall’s Tau, is based on the idea of concordant and discordant pairs.

- Consider two bivariate observations, say, \((X_i, Y_i)\) and \((X_j, Y_j)\).

- The two observations are concordant if both numbers in one observation are larger than the corresponding numbers in the other observation.

- The two observations are discordant 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 \( N_d \)

Pair: 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

<table>
<thead>
<tr>
<th>Two-Tailed</th>
<th>Lower-Tailed</th>
<th>Upper-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_i: ) The ( X ) and ( Y ) variables are associated (either positively or negatively)</td>
<td>( H_i: ) Negative association (smaller values of ( X ) correspond to larger values of ( Y ) and vice versa)</td>
<td>( H_i: ) Positive association (larger values of ( X ) correspond to larger values of ( Y ))</td>
</tr>
</tbody>
</table>

• 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.
Two-tailed  
Reject \( H_0 \) if \( T < W_{\alpha/2} \)
or if \( T > W_{1-\alpha/2} \)

<table>
<thead>
<tr>
<th>Decision Rules</th>
<th>Lower-tailed</th>
<th>Upper-tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject ( H_0 ) if ( T &lt; W_\alpha )</td>
<td>Reject ( H_0 ) if ( T &gt; W_{1-\alpha} )</td>
<td>Table A11</td>
</tr>
</tbody>
</table>

- 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?

\( H_0 \): 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 \( T = \frac{44.5 - 17.5}{44.5 + 17.5} = .4355 \) (\( R \) 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 $p = .929$ (p-value of test $\approx 0$)

Kendall’s $\tau = .821$ (p-value $\approx 0$)

- Reject $H_0$ and conclude there is an increasing temperature trend.
Comparison to Competing Tests

• If the distribution of $X$ and $Y$ is \underline{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 \underline{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 \underline{ordinal} on the measurement scale.

• Kendall’s tau is often used as a measure of association when the data are binary and \underline{ordered}-(for example, \underline{paired}).

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.