Math 373 Lecture 38

Significant positive and negative correlations

Given: a n-element sample of quantitative bivariate data \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}. Let \( \rho \) be the correlation coefficient for the entire population (this value is set by the null hypothesis). Let \( r \) be the correlation coefficient of the sample. Since \( r \) involves two estimated parameters, \( \bar{x} \) and \( \bar{y} \), there are \( df = n-2 \) degrees of freedom.

In Lecture 29 we tested for linear dependency using
\[ F = \frac{MSR/MSE} {1 - r^2} \]

Since \( F \) is a ratio of variances which are sums of squares, it can’t distinguish between a positive correlation and a negative correlation. To do this we use the square root of \( F \), \( r \), which has the \( t \)-distribution.

Quantitative test for significant linear correlation.

Given an \( n \)-element sample of bivariate data with correlation coefficient \( r \).

\( H_a: \rho \neq 0 \) or \( \rho > 0 \), or \( \rho < 0 \) \( H_0: \rho = 0, \rho \leq 0, \rho \geq 0 \)

Test statistic: \( t = \frac{r \sqrt{n-2}} {\sqrt{1-r^2}} \) which has the \( t \)-distribution.

\( df = n-2 \). As before, Table 4 gives the critical values \( t_a \).

Example from Lecture 29. You wish to determine if there is a significant correlation (positive correlation respectively) between the inches of water provided to a tomato plant and pounds of tomatoes it produces. You give 8 plants 4 different amounts of water: 0, 1, 1.5, 2". Each amount is applied to two plants with the following results.

<table>
<thead>
<tr>
<th>x in inches</th>
<th>0.5</th>
<th>1</th>
<th>1.5</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>y in pounds</td>
<td>0, 1</td>
<td>2, 3</td>
<td>3, 3</td>
<td>3, 4</td>
</tr>
</tbody>
</table>

\( n = 8 \), \( \alpha = .05 \)

\( r = .8718 \)

\( df = n-2 = 6 \)

\( t = \frac{.8718 \sqrt{6}} {\sqrt{1-.8718^2}} = 4.359 \)

(a) Is there a significant correlation between \( x \) and \( y \)?

\( H_a: \rho \neq 0 \), \( H_0: \rho = 0 \) (no correlation). Null region: \( \rho \in [0] \)

Acceptance region for \( t \): \( t \in [-t_{a/2}, t_{a/2}] = [-2.447, 2.447] \)

Yes, there is a significant correlation between \( x \) and \( y \).


(b) Is there a positive correlation between \( x \) and \( y \)?

\( H_a: \rho > 0, H_0: \rho \leq 0 \) (no positive correlation). Null: \([-1, 0]\)

Null region: \( \rho \in (-\infty, 0) \)

Acceptance region: \( t \in (-\infty, t_a] = (-\infty, 1.943] \)

Yes, there is a significant positive correlation.

For a two-sided test you can use either this \( t \) test or the F test of Lecture 30. The F test extends to cases with more than two variables. This \( t \) test handles the one-sided cases (positive or negative correlations).

Significant correlations between rankings

The above test dealt with quantitative data. We assumed that the residual error (the deviations of the data from the regression line) were independent and normal. Now suppose both variables are rankings. When is there is a significant correlation? To rank items: order them in a line (ascending or descending order), assign 1, 2, 3, ... to the first, second, third, ..., average items of equal rank. Ranks are neither independent (when one person moves up, someone else must move down) nor normal.

Given: an \( n \)-element sample of bivariate rank data \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}. Let \( \rho \) be the correlation coefficient for the entire population (the value is set by the null hypothesis). Let \( r \) be the correlation coefficient of the sample. The range of \( r \) is \([-1, 1]\). The distribution is symmetric around 0.

Spearman’s rank correlation test. Given an \( n \)-element bivariate rank data set with correlation coefficient \( r \),

\( H_a: \rho \neq 0 \), or \( \rho > 0 \), or \( \rho < 0 \) \( H_0: \rho = 0, \rho \leq 0, \rho \geq 0 \)

Test statistic: \( r \). Table 9 gives the critical values \( r_a \).

Table 9 gives the endpoint of the right-tailed rejection region \( (r_a, 1] \). The corresponding acceptance region is \([-1, r_a] \). The left-tailed acceptance region is \([-r_a, 1] \). The two-tailed acceptance region is \([-r_a/2, r_a/2] \).

In the quantitative correlation test, we used the statistic
\( t = \frac{r \sqrt{n-2}} {\sqrt{1-r^2}} \) which had Student’s \( t \) distribution. In this rank correlation test, there is no formula which converts \( r \) to a known distribution. Hence it has a separate table (table 9).

A teacher is asked to rank his students according to IQ. His ranking is then compared with the student’s IQ scores. Test for a positive correlation.

<table>
<thead>
<tr>
<th>Student</th>
<th>Teacher’s ranking</th>
<th>IQ score</th>
<th>IQ pre-ranking</th>
<th>IQ ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>110</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>90</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>120</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>100</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>100</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>100</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>90</td>
<td>7</td>
<td>6.5</td>
</tr>
</tbody>
</table>

IQ scores are not rankings but we can convert them to rankings by giving 1 to the highest IQ, 2 to the second, ..., etc. As before ranking of items with equal IQ are assigned the average ranking. Hence the three students ranked 3, 4, 5 in IQ all have IQ 100 and hence are given revised ranks of 4, 4, and 4.

Calculating the correlation coefficient for the pairs
\( (3, 2), (6, 6.5), (1, 1), (5, 4), (2, 4), (4, 4), (7, 6.5) \) gives \( r = .8795 \)

\( H_a: \rho > 0 \), \( H_0: \rho \leq 0 \) (no positive correlation).

Null region: \( \rho \in [-1, 0], n = 7 \)

Acceptance region: \( r \in [-1, r_a] = [-1, 7.14] \)

\( r = .8795 \notin [-1, r_a] = [-1, 7.14] \) = accept. reg. for no pos. correlation.

Yes, there is a significant positive correlation.