# Inferential Statistics for b and r

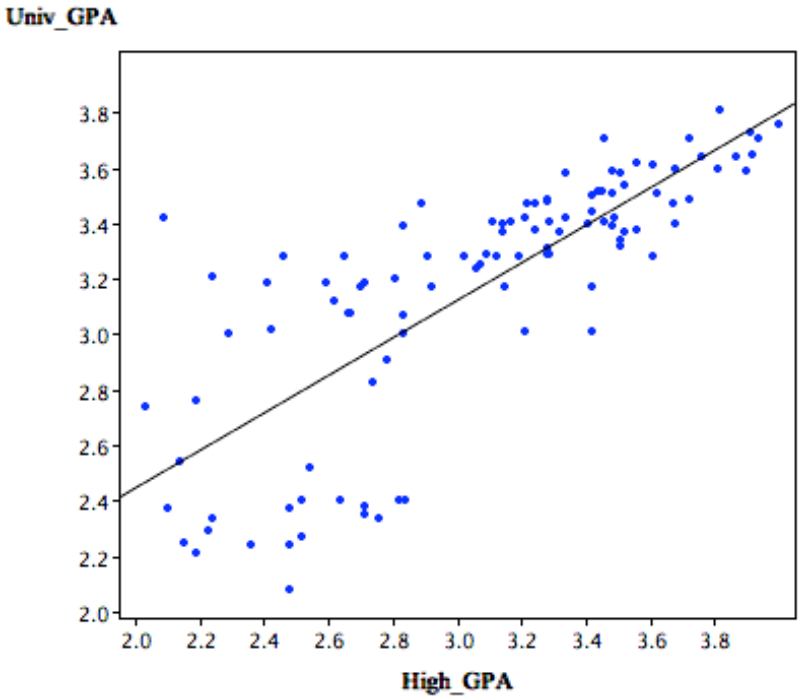
**Prerequisites**
Sampling Distribution of r, Confidence Interval for r

This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation. As you will see, if the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

**Assumptions**
Although no assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics. Naturally, these assumptions refer to the population not the sample.

1. Linearity: The relationship between the two variables is linear.
2. Homoscedasticity: The variance around the regression line is the same for all values of X. A clear violation of this assumption is shown in Figure 1. Notice that the predictions for students with high high-school GPA's is very good whereas the prediction for students with low high-school GPA's is not very good. In other words, the points for students with high high-school GPA's are close to the regression line whereas the points for low high-school GPA students do not.

Figure 1. University GPA as a function of High School GPA.




3. The errors of prediction are distributed normally. This means that the distributions of deviations from the regression line are normally distributed. It does not mean that X or Y is normally distributed.

**Significance Test for the Slope (b)**
Recall the general formula for a t test

As applied here, the statistic is the sample value of the slope (b) and the hypothesized value is 0. The degrees of freedom for this test are:

```
df = N-2
```

where N is the number of pairs of scores.

The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}}$$

where $s_b$ is the estimated standard error of b, $s_{est}$ is the standard error of the estimate. SSX is the sum of squared deviations of X from the the mean of X. It

is calculated as

$$SSX = \sum (X - M_X)^2$$

where $M_X$ is the mean of X. As shown previously, the standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1-r^2)SSY}{N-2}}$$

These formulas are illustrated with the data shown in Table 1. These data are reproduced from the introductory section. The column X, has the values of the *predictor variable* and the column Y has the *criterion variable*. The third column, x, contains the the differences between the column X and the mean of X. The fourth column, $x^2$, is the square of the x column. The fifth column, y, contains the the differences between the column Y and the mean of Y. The last column, $y^2$, is simply the square of the y column.

Table 1. Example data.

| | X | Y | x | $x^2$ | y | $y^2$ |
|---|---|---|---|---|---|---|
| | 1.00 | 1.00 | -2.00 | 4 | -1.06 | -1.1236 |
| | 2.00 | 2.00 | -1.00 | 1 | -0.06 | 0.0036 |
| | 3.00 | 1.30 | 0.00 | 0 | -0.76 | 0.5776 |
| | 4.00 | 3.75 | 1.00 | 1 | 1.69 | 2.8561 |
| | 5.00 | 2.25 | 2.00 | 4 | 0.19 | 0.0361 |
| sum | 15.00 | 10.30 | 0.00 | 10.00 | 0.00 | 4.5970 |

The computations of the standard error of the estimate ($s_e$) for these data is shown on the section on the standard error of the estimate. It is equal to 0.964.

$s_e = 0.964$

SSX is the sum of squared deviations from the mean of X. It is therefore equal to the sum of the x2 column and is equal to 10.

SSX = 10.00

We now have all the information to compute the standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

As shown previously, the slope (b) is 0.425. Therefore,

$$t = \frac{0.425}{0.305} = 1.39$$

df = N-2=5-2=3.

The p value for a two-tailed test is 0.26. Therefore, the slope is not significantly different from 0.

**Confidence Interval for the Slope**

The method for computing a confidence interval for the population slope is very similar to methods for computing other confidence interval. For the 95% confidence interval the formula is:

lower limit: $b - (t_{.95})(s_b)$
upper limit: $b + (t_{.95})(s_b)$

where $t_{.95}$ is the value of t to use for the 95% confidence interval.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table 2. The first column, df, stands for degrees of freedom.

Table 2. Abbreviated t table.

| df | 0.95 | 0.99 |
|---|---|---|
| 2 | 4.303 | 9.925 |
| 3 | 3.182 | 5.841 |
| 4 | 2.776 | 4.604 |
| 5 | 2.571 | 4.032 |
| 8 | 2.306 | 3.554 |
| 10 | 2.228 | 3.169 |
| 20 | 2.086 | 2.845 |
| 50 | 2.009 | 2.678 |
| 100 | 1.984 | 2.626 |

You can also use the "inverse t distribution" calculator to find the t values to use in confidence interval.

Applying these formulas to the example data,

lower limit: 0.425 - (3.182)(0.305)= -0.55
upper limit: 0.425 + (3.182)(0.305)= 1.40

**Significance Test for the Correlation**

The formula for a significance test of Pearson's correlation is shown below:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

where N is the number of paris of scores. For the example data,

$$t = \frac{0.627\sqrt{5-2}}{\sqrt{1-0.627^2}} = 1.39$$

Notice that this is the same t value obtained in the test of t b. As in that test the degrees of freedom is N-2 = 3.

**Confidence Interval for the Correlation**

There are several steps in computing a confidence interval on r (the population value of Pearson's r). Recall from the chapter on sampling distributions that:

1. The sampling distribution of Pearson's r is skewed.
2. Fisher's z' transformation of r is normal.
3. z' = 0.5 ln[(1+r)/(1-r)].
4. z' has a standard error of .

The calculation of the confidence interval involves the following steps:

1. Converting r to z'. For our example data, the r of 0.627 is transformed to a z' 0.736. This can be done using the formula above or the r to z' calculator.
2. Find the standard error of z' ($s_{z'}$). For our example, N = 5 so $s_{z'}$ = 0.707.
3. Compute the confidence interval in terms of z' using the formula

   lower limit = z' - $(z_{.95})(s_{z'})$

   upper limit = z' + $(z_{.95})(s_{z'})$

   For the example,

   lower limit = 0.736 - (1.96)(0.707) = -0.650

```
upper limit = 0.736 + (1.96)(0.707) =  2.122
```

4. Convert the interval for z' back to Pearson's correlation. This can be done with the r to z' calculator.
   For the example,

   ```
   lower limit = -0.57
   upper limit =  0.97
   ```

   The interval is so wide because the sample size is so small.