Sample: Statistics and Probability - Correlation Coefficients

1. What is the relationship, if any, between education and gender? Discuss any differences that may exist and describe the characteristics of the sample.

There can be three possible situations:

1) gender does not affect the education level;
2) persons with gender = 1 are highly educated than persons with gender = 2
3) persons with gender = 2 are highly educated than persons with gender = 1

We want to establish the situation that is suitable for the given sample.

Consider the data characteristics.

There are two samples of the data – one for each sex value. And using this sample we need to compare means of two populations – education level of all males with education level of all females. Data in these samples has the following characteristics:

- numerical;
- assumed to be random;
- the population distribution is not known;
- sample sizes are not equal;
- sample sizes are large (> 30).

Using these characteristics and the purpose of the investigation we can conclude that two – sampled t-test should be used to compare means of two samples. Provide this test.

First test whether the first situation is presented. The appropriate hypotheses look:

H0: Mean1 = Mean2

Ha: Mean1 ≠ Mean2

These hypotheses need to provide two - tiled test.

State the significance level alpha = 0.05. It is most frequently used and enough for the social researches.

First, calculate statistic characteristics of the samples (mean, standard deviation and sample sizes). These values are:

Mean1 = 13.675, SD1 = 4.517, n1 = 622

Mean2 = 13.319, SD2 = 5.981, n2 = 797

Now calculate the standard error. Use formula for t-test for samples with different standard deviations and different sample sizes:

$$SE = \sqrt{\frac{SD1^2}{n1} + \frac{SD2^2}{n2}} \approx 0.279$$

Determine degrees of freedom for the test:

$$df = \frac{\left(\frac{SD1^2}{n1} + \frac{SD2^2}{n2}\right)^2}{\frac{\left(\frac{SD1^2}{n1}\right)^2}{n1-1} + \frac{\left(\frac{SD2^2}{n2}\right)^2}{n2-1}} \approx 1416$$

Calculate the test statistic:

$$t = \frac{Mean1 - Mean2}{SE} \approx 1.279$$

Using t-table, determine the p-value (two-tiled):

$0.1 < P(t = 1.279, df = 1416) < 0.2$

As we can see, alpha < p-value. So, we cannot reject the null hypothesis and we should conclude that there are no significant differences between education levels for different gender. So, the first situation is presented by the given sample.

2. What is the relationship, if any, between parental education and the education of the respondent? If a relationship exists, which parent has the strongest effect on the educational level of the respondent?

The data that represent education level of respondent and his/her mother and father is numerical, matched-pair and assumed to be collected randomly, so we can use correlation coefficient to determine whether there is a relationship, between parental education and the education of the respondent. But first we can use dot-point diagram to determine the relationship roughly, visual only. The scatter plot of the data is shown at the figure 1.

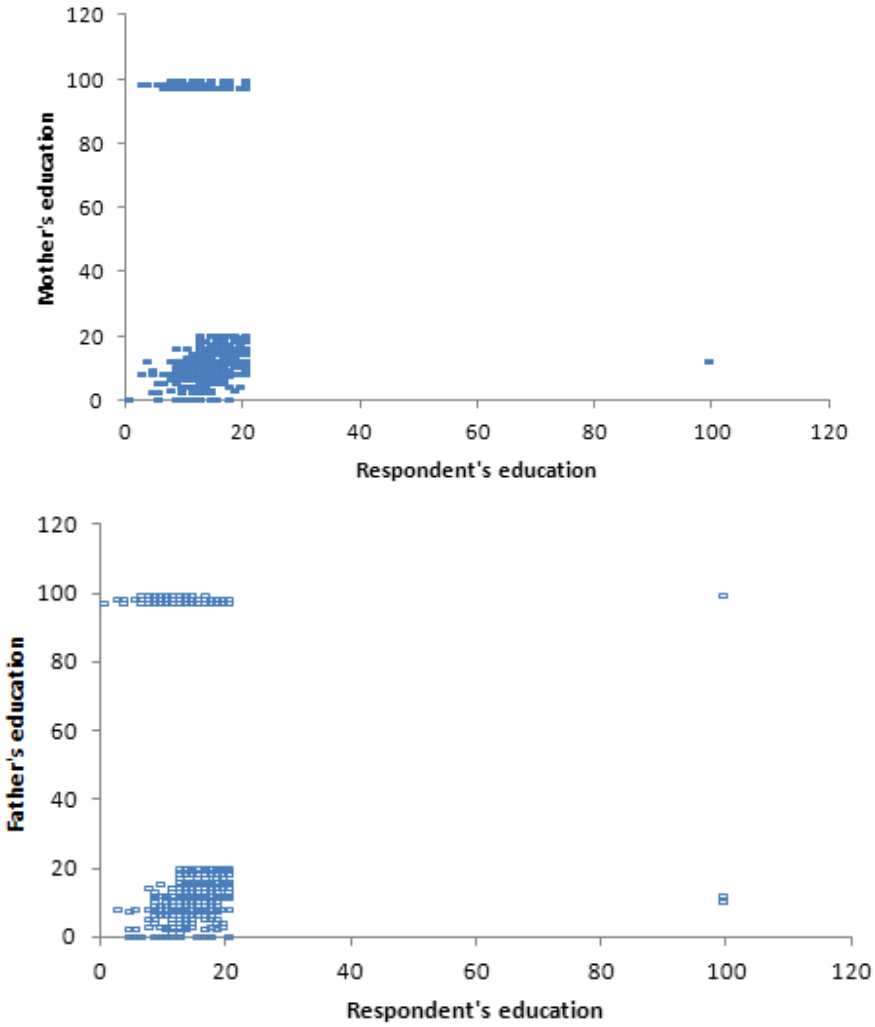

Figure 1 – Relationship between parental education and the education of the respondent.

As we can see, there is no visible relationship between parental education and the education of the respondent. Verify this fact using the correlation coefficient test.

Calculate both correlation coefficients:

Rm = -0.15

Rp = -0.14

So we can observe the very weak negative correlation. Verify if it is significant.

Data in the samples has the following characteristics:

- numerical;
- assumed to be random;
- sample sizes are large (> 30).

Using these characteristics we can conclude that one – sampled t-test can be used to test the significance of correlation coefficient. Provide this test.

The hypotheses look:

H0: R = 0

Ha: R ≠ 0

These hypotheses need to provide two - tiled test.

State the significance level alpha = 0.05. It is most frequently used and enough for the social researches.

The sample sizes are equal:

Nm = Np = 1419.

Determine degrees of freedom for the test:

$$dfm = dfp = N - 2 = 1417$$

Calculate the test statistic:

$$t = \frac{R}{\sqrt{\frac{1 - R^2}{N - 2}}} => tm = -5.92, tp = -5.57$$

Using t-table, determine the p-value (two-tiled):

Pm < 0.001

Pp < 0.001

As we can see, alpha > p-value. So, we should reject the null hypothesis and conclude that there are significant negative correlation between parental education and the education.

Effects of both parents education are approximately equal, but the mother's education level is little more influential.