Badanie wspolzaleznosci zjawisk

Lucas van der Velde

Task 1: Correlation coefficient We would like to test whether the length of two jumps of professional ski-jumpers during a competence are correlated. To test this, we use data from table below.

Jump 1	97	90.5	89	85	90.5	86	84	83	87	89
Jump 2	101	91.5	91	97.5	88	87.5	88.5	85.5	83	83

1. Compute the correlation coefficient and test its significance
2. Some might believe that the winner is an outlier, as she is the only one who jumped over 95 meters in both attempts. Compute the correlation one more time after excluding the potential outlier. Do results change?

Solution I think this task went well. Most of those who started it were on the right track. BUT, it did require a lot of time for all the computations. Below, you can find my solutions

1. $$\bar{x} = 88.1 \qquad d(x) = 4.088 \qquad \bar{y} = 89.65 \qquad d(y) = 5.87$$

   $$\bar{x}\bar{y} * n = 78981.65 \qquad \sum xy = 79091.75$$

   $$c_{x,y} = \frac{1}{n-1}(\sum xy - \bar{x}\bar{y} * n) = \frac{1}{9}(79091.75 - 78981.65) = 12.23$$

   $$r_{x,y} = \frac{c_{x,y}}{d(x)d(y)} = \frac{12.23}{4.088 * 5.87} = 0.51$$

   $$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} = \frac{0.51}{\sqrt{1-0.51^2}}\sqrt{10-2} = 1.676$$

   For $\alpha = 0.05$, $t_{0.05,8} = 2.30$. We lack evidence to reject the null in the favor of the alternative. The correlation between jumps is not statistically significant.

2. Now, we exclude the winner. This is a little bit less computationally intensive, but you still had to do a fair share of work. Under time pressure, I think the best approach is to consider that we can write the mean of all observations but i as a function of the mean, the value of x_i and the number of observations: $\bar{x}_{\forall j \neq i} = (\bar{x} - x_i/n) * \frac{n}{n-1}$. This should reduce the number of time you use the calculator. Below, I will write just the last step. If there are any doubts concerning previous steps, please consult by email.

   $$r_{x,y} = \frac{c_{x,y}}{d(x)d(y)} = \frac{-0.267}{2.79 * 4.57} = -0.0209$$

   For $\alpha = 0.05$, $t_{0.05,7} = 2.36$. We lack evidence to reject the null in the favor of the alternative. The correlation between jumps is not statistically significant.

Task 2 Table below shows how European countries are ranked with respect to two variables: differences in average wages between men and women, and differences in employment rates between men and women. Lower values of the rankings correspond to lower differences. Using data from the table, compute the correlation coefficient. Provide an interpretation for the results.

	Wages	Employment		Wages	Employment
Belgium	4	15	Lithuania	10	1
Bulgaria	14	11	Luxembourg	2	17
Czechia	27	24	Hungary	17	21
Denmark	19	5	Malta	8	28
Germany	26	9	Netherlands	20	13
Estonia	28	8	Austria	25	12
Ireland	12	19	Poland	6	22
Greece	9	27	Portugal	15	6
Spain	16	20	Romania	1	25
France	18	10	Slovenia	5	7
Croatia	7	18	Slovakia	23	23
Italy	3	26	Finland	22	4
Cyprus	13	16	Sweden	11	3
Latvia	21	2	United Kingdom	24	14

Solution Most of you who started this task completed it with no problems. Only one was not perfect, and that was only because of a small mistake in the number of observations.

$$r_{x,y} = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 * 4842}{28 * (28^2 - 1)} = 1 - \frac{29052}{21924} = -0.33$$

The interpretation of the results requires some more nuanced. Please remember that answer of the form "ujemny umiarkowany zwiazek", though correct, are not enough. A reference to the question is needed. For example, that "countries that present smaller disparities in earning between men and women, are also those that present larger disparities in employment." A candidate explanation for this negative correlation relates to selection into the labor market. If the threshhold for women participation is high, only the most productive women will enter the labor market. These women will receive higher wages, which means that the wage gap (the average difference in earnings with respect to men) will be smaller.