EXAM OF STATISTICS (DESCRIPTIVE STATISTICS AND REGRESSION)

2nd Physiotherapy Version A **June, 18 2019**

Duration: 1 hour and 15 minutes.

(5 pts.) 1. A study tries to determine the effectiveness of an occupational risk prevention program in jobs that require to be sit a lot of hours. A sample of individuals between 40 and 50 years that spent more than 5 hours sitting were drawn. It was observed if they followed or not the occupational risk prevention program and the number of spinal injuries after 10 years. The results are shown in the table below.

With prevention program	1	3	2	4	4	0	2	4	2	2	5	2	3	2	0
Wihtout prevention program	6	3	1	3	7	6	5	5	9	5	5	4	4	3	

(a) Plot the polygon of cumulative relative frequencies of the total sample.

(b) According to the interquartile range, which sample has more central spread of the spinal injuries, the sample of people following the prevention program or the sample of people not following the prevention program?

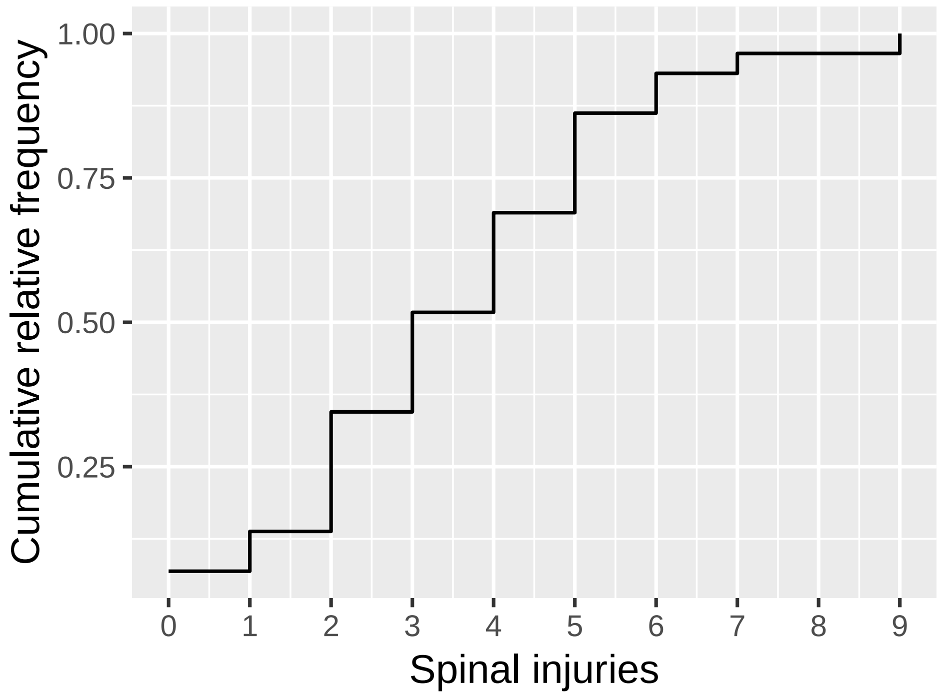
(c) Which sample has a greater relative spread with respect to the mean of the spinal injuries, the sample of people following the prevention program or the sample of people not following the prevention program?

(d) Which sample has a more normal kurtosis of the number of spinal injuries, the sample of people following the prevention program or the sample of people not following the prevention program?

(e) Which number of spinal injuries is relatively greater, 2 injuries of a person following the prevention program or 4 injuries of a person not following the prevention program?

Use the following sums for the computations:
With prevention program: $\sum x_i = 36$ injuries, $\sum x_i^2 = 116$ injuries^2, $\sum(x_i - \bar{x})^3 = -0.48$ injuries^3 and $\sum(x_i - \bar{x})^4 = 135.97$ injuries^4.
Without prevention program: $\sum y_i = 66$ injuries, $\sum y_i^2 = 362$ injuries^2, $\sum(y_i - \bar{y})^3 = 27.92$ injuries^3 and $\sum(y_i - \bar{y})^4 = 586.9$ injuries^4.

Solution



(a)

(b) With prevention program: $Q_1 = 2$ injuries, $Q_3 = 4$ injuries, $IQR = 2$ injuries.
Without prevention program: $Q_1 = 3$ injuries, $Q_3 = 6$ injuries, $IQR = 3$ injuries.
The sample not following the prevention program has more central spread since the interquartile range is greater.

(c) With prevention program: $\bar{x} = 2.4$ injuries, $s^2 = 1.9733$ injuries^2, $s = 1.4048$ injuries and $cv = 0.5853$.
Without prevention program: $\bar{y} = 4.7143$ injuries, $s^2 = 3.6327$ injuries^2, $s = 1.906$ injuries and $cv = 0.4043$.
The sample following the prevention program has a greater relative spread with respect to the mean since the coef. of variation is greater.

(d) With prevention program: $g_2 = -0.6722$.
Without prevention program: $g_2 = 0.1768$.
Thus the sample not following the prevention program has a more normal kurtosis, since g_2 is closer to 0.

(e) With prevention program: $z(2) = -0.2847$.
Without prevention program: $z(4) = -0.3748$.
Thus 4 injuries in the sample not following the prevention program is relatively smaller, since its standard score is smaller.

(4 pts.) 2. The evolution of the price of a muscle relaxant between 2015 and 2019 is shown in the table below.

Year	2015	2016	2017	2018	2019
Price (€)	1.40	1.60	1.92	2.30	2.91

(a) Which regression model is better to predict the price, the linear or the exponential?

(b) Use the best of the two previous models to predict the price in 2020.

Solution

(a) $\bar{x} = 2017$ years, $s_x^2 = 2$ years^2.
$\bar{y} = 2.026$ €, $s_y^2 = 0.2882$ €^{2}.
$\overline{\log(y)} = 0.672 \log(€)$, $s^2_{\log(y)} = 0.0673 \log(€)^2$.
$s_{xy} = 0.744$, $s_{x\log(y)} = 0.3653$
Linear coef. determination: $r^2 = 0.9603$
Exponential coef. determination: $r^2 = 0.9909$
Thus the exponential regression model is better to predict the price since the coef. of determination is greater.

(b) Exponential regression model: $y = e^{-367.6861+0.1826x}$.
Prediction: $y(2020) = 3.3867$ €.

(1 pts.) 3. In a linear regression study between two variables X and Y we know $\bar{x} = 3$, $s_x^2 = 2$, $s_y^2 = 10.8$ and the regression line of Y on X is $y = 90.9 - 2.3x$.

(a) Compute the mean of Y.

(b) Compute and interpret the linear correlation coefficient.

Solution

(a) $\bar{y} = 84$.

(b) $r = -0.9898$.
