EXAM OF STATISTICS
(DESCRIPTIVE STATISTICS AND REGRESSION)

Pharmacy/Biotechnology 1st year — Version A — **October, 26 2020**

Duration: 1 hour.

(4 pts.) 1. The table below shows the daily number of patients hospitalized in a hospital during the month of September.

patients	Frecuencia
$(10, 14]$	6
$(14, 18]$	10
$(18, 22]$	7
$(22, 26]$	6
$(26, 30]$	1

(a) Study the spread of the 50% of central data.

(b) Compute the mean and study the dispersion with respect to it.

(c) Study the normality of the patients distribution.

(d) If the mean was 35 patients and the variance 40 patients^2 during the month of April, which month had a higher relative variability?

(e) Which number of people hospitalized was greater, 20 persons in September or 40 in April?

Use the following sums for the computations:
$\sum x_i n_i = 544$ patients, $\sum x_i^2 n_i = 10464$ patients^2, $\sum (x_i - \bar{x})^3 n_i = 736.14$ patients^3 and $\sum (x_i - \bar{x})^4 n_i = 25367.44$ patients^4.

Solution

(a) $Q_1 = 16$, $Q_3 = 20$ and $IQR = 4$. Thus the central dispersion is moderate.

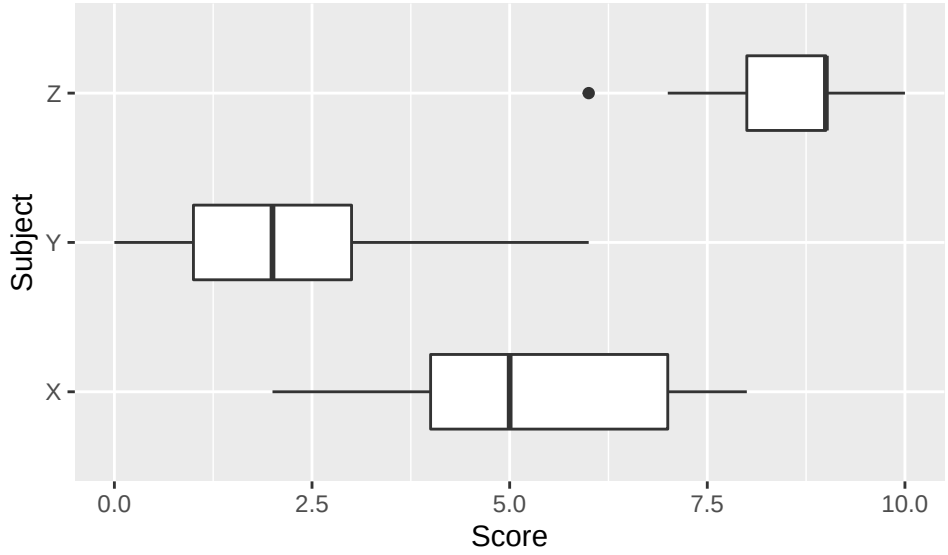
(b) $\bar{x} = 18.1333$, $s^2 = 19.9822$, $s = 4.4701$ and $cv = 0.2465$. Thus, the dispersion with respect to the mean is small and the mean represents well.

(c) $g_1 = 0.2747$ and $g_2 = -0.8823$. As the coefficient of skewness and the coefficient of kurtosis fall between -2 and 2, we can assume that the sample comes from a normal population.

(d) Let Y be the daily number of patients hospitalized during April. Then, $cv_y = 0.1807$. Since the coefficient of variation in April is greater than the one in September, there is a relative higher variability in April.

(e) September: $z(20) = 0.4176$.
April: $z(40) = 0.7906$.
Thus, 40 patients hospitalized in April is relatively higher than 20 in September.

(1 pts.) 2. The chart below shows the distribution of scores in three subjects.



(a) Which subject is more difficult?

(b) Which subject has more central dispersion?

(c) Which subjects have outliers?

(d) Which subject is more asymmetric?

Solution

(a) Subject Y because its scores are smaller.

(b) Subject X because the box is wider.

(c) Subject Z because there is a score out of the whiskers.

(d) Subject Z because the distance from the first quartile to the median (left side of the box) is greater than de distance from the third quartile to the median (right side of the box).

(5 pts.) 3. In a sample of 10 families with a child older than 20 it has been measured the height of the father (X), the mother (Y) and the children (Z) in centimeters, getting the following results:

$\sum x_i = 1774$ cm, $\sum y_i = 1630$ cm, $\sum z_i = 1795$ cm,
$\sum x_i^2 = 315300$ cm^2, $\sum y_i^2 = 266150$ cm^2, $\sum z_i^2 = 322737$ cm^2,
$\sum x_i y_j = 289364$ cm^2, $\sum x_i z_j = 318958$ cm^2, $\sum y_i z_j = 292757$ cm^2.

(a) On which height does the height of the child depend more linearly, the height of the father or the mother?

(b) Using the best linear regression model, predict the height of a child with a father 181 cm tall and a mother 163 cm tall?

(c) How much will increase the height of the child for each centimeter that increases the height of the father? And for each centimeter that increases the height of the mother?

(d) How would the reliability of the prediction be if the heights were measured in inches? (An inch is 2.54 cm).

Solution

(a) $\bar{x} = 177.4$ cm, $s_x^2 = 59.24$ cm^2,
$\bar{y} = 163$ cm, $s_y^2 = 46$ cm^2,
$\bar{z} = 179.5$ cm, $s_z^2 = 53.45$ cm^2,
$s_{xz} = 52.5$ cm^2 and $s_{yz} = 17.2$ cm^2.
$r_{xz}^2 = 0.8705$ and $r_{yz}^2 = 0.1203$, thus the height of the child depends linearly more on the height of the father since the $r_{xz}^2 > r_{yz}^2$.

(b) Regression line of Z on X: $z = 22.2836 + 0.8862z$ and $z(181) = 182.6904$.

(c) The height of the child will increase 0.8862 cm per cm of the height of the father and 0.3739 cm per cm of the height of the mother.

(d) The same, as after applying the same linear transformation to X and Z, the variances are multiplied by the square of the slopes and the covariance is multiplied by the product of slopes.
