EXAM OF STATISTICS (DESCRIPTIVE STATISTICS AND REGRESSION)

Pharmacy/Biotechnology 1st year — Version A — **October, 14 2019**

Duration: 1 hour and 15 minutes.

(4.5 pts.) 1. It has been measured the systolic blood pressure (in mmHg) in two groups of 100 persons of two populations A and B. The table below summarize the results.

Systolic blood pressure	Num persons A	Num persons B
$(80, 90]$	4	6
$(90, 100]$	10	18
$(100, 110]$	28	30
$(110, 120]$	24	26
$(120, 130]$	16	10
$(130, 140]$	10	7
$(140, 150]$	6	2
$(150, 160]$	2	1

(a) Which of the two systolic blood pressure distributions is less asymmetric? Which one has a higher kurtosis? According to skewness and kurtosis can we assume that populations A and B are normal?

(b) In which group is more representative the mean of the systolic blood pressure?

(c) Compute the value of the systolic blood pressure such that 30% of persons of the group of population A are above it?

(d) Which systolic blood pressure is relatively higher, 132 mmHg in the group of population A, or 130 mmHg in the group of population B?

(e) If we measure the systolic blood pressure of the group of population A with another tensiometer, and the new pressure obtained (Y) is related with the first one (X) according to the equation $y = 0.98x - 1.4$, in which distribution, X or Y, is more representative the mean?

Use the following sums for the computations:
Group A: $\sum x_i n_i = 11520$ mmHg, $\sum x_i^2 n_i = 1351700$ mmHg^2, $\sum (x_i - \bar{x})^3 n_i = 155241.6$ mmHg^3 y $\sum (x_i - \bar{x})^4 n_i = 16729903.52$ mmHg^4.
Group B: $\sum x_i n_i = 11000$ mmHg, $\sum x_i^2 n_i = 1230300$ mmHg^2, $\sum (x_i - \bar{x})^3 n_i = 165000$ mmHg^3 y $\sum (x_i - \bar{x})^4 n_i = 13632500$ mmHg^4.

Solution

(a) Group A: $\bar{x} = 115.2$ mmHg, $s^2 = 245.96$ mmHg^2, $s = 15.6831$ mmHg, $g_{1A} = 0.4024$ and $g_{2A} = -0.2346$.
Group B: $\bar{x} = 110$ mmHg, $s^2 = 203$ mmHg^2, $s = 14.2478$ mmHg, $g_{1B} = 0.5705$ and $g_{2B} = 0.3081$.
Thus the distribution of the population A group is less asymmetric since g_{1A} is closer to 0 than g_{1B} and the populaton B group has a higher kurtosis since $g_{2B} > g_{2A}$. Both populations can be cosidered normal since g_1 and g_2 are between -2 and 2.

(b) $cv_A = 0.1361$ and $cv_B = 0.1295$, thus, the mean of group B is a little bit more representative since its coef. of variation is smaller than the one of group A.

(c) $P_{70} \approx 125$ mmHg.

(d) The standard scores are $z_A(132) = 1.0712$ and $z_B(130) = 1.4037$. Thus, 130 mmHg in group B is relatively higher than 132 mmHg in group A.

(e) $\bar{y} = 111.496$, $s_y = 15.3694$ and $cv_y = 0.1378$. Thus the mean of X is more representative than the mean of Y since $cv_x < cv_y$.

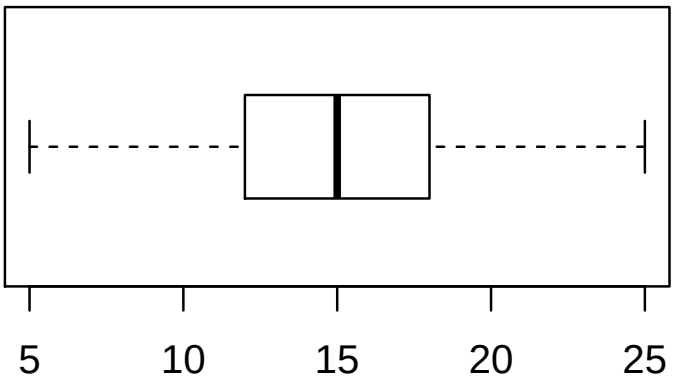
(1 pts.) 2. In a symmetric distribution the mean is 15, the first quartile 12 and the maximum value is 25.

(a) Draw the box and whiskers plot.

(b) Could an hypothetical value of 2 be considered an outlier in this distribution?

Solution

(a) $Q_1 = 12$, $Q_2 = 15$, $Q_3 = 18$, $IQR = 6$, $f_1 = 3$, $f_2 = 27$, $Min = 5$ and $Max = 25$.



(b) Yes, because $2 < f_1$.

(4.5 pts.) 3. A pharmaceutical company is trying three different analgesics to determine if there is a relation among the time required for them to take effect. The three analgesics were administered to a sample of 20 patients and the time it took for them to take effect was recorded. The following sums summarize the results, where X, Y and Z are the times for the three analgesics.

$\sum x_i = 668$ min, $\sum y_i = 855$ min, $\sum z_i = 1466$ min,
$\sum x_i^2 = 25056$ min^2, $\sum y_i^2 = 42161$ min^2, $\sum z_i^2 = 123904$ min^2,
$\sum x_i y_j = 31522$ min^2, $\sum y_j z_j = 54895$ min^2.

(a) Is there a linear relation between the times X and Y? And between Y and Z? How are these linear relationships?

(b) According to the regression line, how much will the time X increase for every minute that time Y increases?

(c) If we want to predict the time Y using a linear regression model, ¿which of the two times X or Z is the most suitable? Why?

(d) Using the chosen linear regression model in the previous question, predict the value of Y for a value of X or Z of 40 minutes.

(e) If the correlation coefficient between the times X and Z is $r = -0.69$, compute the regression line of X on Z.

Solution

(a) $\bar{x} = 33.4$ min, $s_x^2 = 137.24$ min^2,
$\bar{y} = 42.75$ min, $s_y^2 = 280.4875$ min^2,
$\bar{z} = 73.3$ min, $s_z^2 = 822.31$ min^2,
$s_{xy} = 148.25$ min^2 and $s_{yz} = -388.825$ min^2.
Thus, there is a direct linear relation between X and Y and an inverse linear relation between Y and Z.

(b) $b_{xy} = 0.5285$ min.

(c) $r_{xy}^2 = 0.5709$ and $r_{yz}^2 = 0.6555$, thus the regression line of Y on Z explains better Y than the regression line of Y on X since $r_{yz}^2 > r_{xy}^2$.

(d) Regression line of Y on Z: $y = 77.4095 + -0.4728z$ and $y(40) = 58.4957$.

(e) $s_{xz} = -231.7967$ and the regression line of X on Z is $x = 54.0622 + -0.2819z$.
