EXAM OF STATISTICS (DESCRIPTIVE STATISTICS AND REGRESSION)

2nd Physiotherapy Version A **March, 25 2022**

Duration: 1 hour.

(5 pts.) 1. The time required by a drug A to be effective has been measured (in minutes) in a sample of 150 patients. The table below summarize the results.

Response time	Patients
$(0, 5]$	5
$(5, 10]$	15
$(10, 15]$	32
$(15, 20]$	36
$(20, 30]$	42
$(30, 60]$	20

(a) Are there outliers in the sample? Justify the answer.

(b) What is the minimum time for the 20% of patients with highest response time?

(c) What is the average response time? Is the mean representative?

(d) Can we assume that the sample comes from a normal population?

(e) If we take another sample of patients with mean 18 min and standard deviation 15 min, in which group is greater a response time of 25 min?

Use the following sums for the computations: $\sum x_i n_i = 3105$ min, $\sum x_i^2 n_i = 83650$ min^2, $\sum(x_i - \bar{x})^3 n_i = 206851.65$ min^3 y $\sum(x_i - \bar{x})^4 n_i = 8140374.96$ min^4.

Solution

(a) $Q_1 = 12.7344$ min, $Q_3 = 25.8333$ min, $IQR = 13.099$ min, $f_1 = -6.9141$ min and $f_2 = 45.4818$ min. Therefore there are outliers in the sample since the upper limit of the last interval is above the upper fence.

(b) $P_{80} = 27.619$ min.

(c) $\bar{x} = 20.7$ min, $s^2 = 129.1767$ min^2, $s = 11.3656$ min and $cv = 0.5491$. The mean is not very representative since the $cv > 0.5$.

(d) $g_1 = 0.9393$ and $g_2 = 0.2523$. Since g_1 and g_2 are between -2 and 2, we can assumme that the sample comes from a normal (bell-shaped) population.

(e) The standard score of the first sample is $z(25) = 0.3783$ and the standard score of the second one $z(25) = 0.4667$, thus a time of 25 min is relatively greater in the second sample.

(5 pts.) 2. We have measured the average number of weekly hours of study X and the score in a clinic entrance test Y of 8 candidates, getting the following results:

$\sum x_i = 15.9$, $\sum \log(x_i) = 3.852$, $\sum y_j = 41.5$, $\sum \log(y_j) = 11.511$,
$\sum x_i^2 = 42.23$, $\sum \log(x_i)^2 = 5.559$, $\sum y_j^2 = 274.25$, $\sum \log(y_j)^2 = 20.742$,
$\sum x_i y_j = 103.3$, $\sum x_i \log(y_j) = 28.047$, $\sum \log(x_i) y_j = 32.616$.

(a) Compute the equations of the logarithmic and exponential regression models of the score as a function of the average number of hours of study.

(b) Use the best of the previous models to predict the score for somebody that study 3.2 hours a week. Is this prediction reliable?

Solution

(a) $\bar{x} = 1.9875$ hours, $s_x^2 = 1.3286$ hours^2.
$\overline{\log(x)} = 0.4815$ log(hours), $s_{\log(x)}^2 = 0.463$ log(hours)^2.
$\bar{y} = 5.1875$ points, $s_y^2 = 7.3711$ points^2.
$\overline{\log(y)} = 1.4389$ log(points), $s_{\log(y)}^2 = 0.5224$ log(points)^2.
$s_{x\log(y)} = 0.6461$, $s_{\log(x)y} = 1.5792$
Logarithmic regression model: $y = 3.5453 + 3.4106\log(x)$
Exponential regression model: $y = e^{0.4723+0.4863x}$

(b) Logarithmic coefficient of determination: $r^2 = 0.7307$
Exponential coefficient of determination: $r^2 = 0.6015$
Therefore, the best regression model to predict is the exponential.
Prediction: $y(3.2) = 14.4592$.
