# EXAM OF STATISTICS (DESCRIPTIVE STATISTICS AND REGRESSION)

**2nd Physiotherapy** Version A **March, 26 2019**

**Duration**: 1 hour and 15 minutes.

(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: $\sum x_i = 3105$ min, $\sum x_i^2 = 83650$ min$^2$, $\sum(x_i-\bar{x})^3 = 206851.65$ min$^3$ y $\sum(x_i-\bar{x})^4 = 8140374.96$ 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.

---

(1.5 pts.) 2. In a regression study about the relation between two variables $X$ and $Y$ we got $\bar{x} = 7$ and $r^2 = 0.9$. If the equation of the regression line of $Y$ on $X$ is $y - x = 1$, compute

(a) The mean of $Y$.

(b) The equation of the regression line of $X$ on $Y$.

(c) What value does this regression model predict for $x = 6$? And for $y = 10$?

---

**Solution**

(a) $\bar{y} = 8$.

(b) Regression line of $X$ on $Y$: $x = 0.9y - 0.2$.

(c) $y(6) = 7$ and $x(10) = 8.8$.

---

(3.5 pts.) 3. In a tennis club the age $(X)$ and the height $(Y)$ of the ten players conforming the female youth team has been measured.

| Age (years) | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|
| Height (cm) | 128 | 144 | 148 | 154 | 158 | 161 | 165 | 164 | 166 | 167 |

(a) Plot the scatter plot (Height on Age).

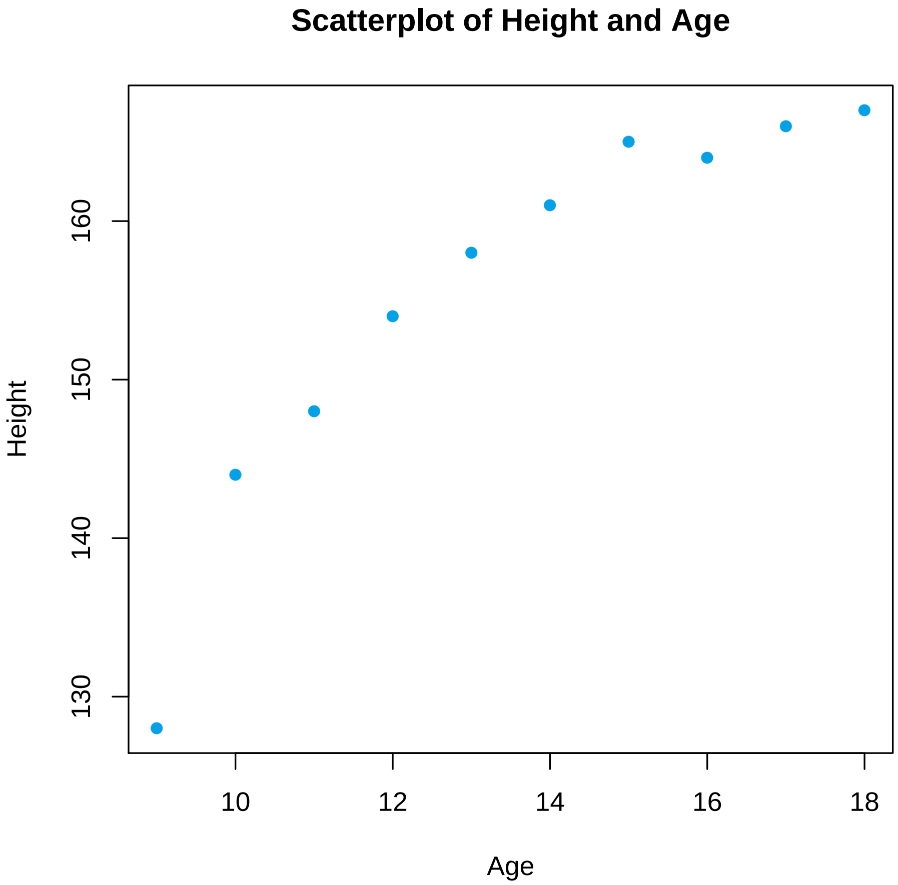
(b) Which regression model bests fits these data, the linear or the logarithmic?

(c) What is the expected height of a player 12.5 years old according to the best of two previous models?

Use the following sums for the computations:
$\sum x_i = 135$ years, $\sum \log(x_i) = 25.7908$ log(years), $\sum y_j = 1555$ cm, $\sum \log(y_j) = 50.4358$ log(cm),
$\sum x_i^2 = 1905$ years$^2$, $\sum \log(x_i)^2 = 67.0001$ log(years)$^2$, $\sum y_j^2 = 243191$ cm$^2$, $\sum \log(y_j)^2 = 254.4404$ log(cm)$^2$,
$\sum x_i y_j = 21303$ years·cm, $\sum x_i \log(y_j) = 682.9473$ years· log(cm), $\sum \log(x_i) y_j = 4035.0697$ log(years)cm,
$\sum \log(x_i)\log(y_j) = 130.2422$ log(years) log(cm).

---

**Solution**




(a)

(b) $\bar{x} = 13.5$ years, $s_x^2 = 8.25$ years$^2$.
$\overline{\log(x)} = 2.5791$ log(years), $s_{\log(x)}^2 = 0.0483$ log(years)$^2$.
$\bar{y} = 155.5$ cm, $s_y^2 = 138.85$ cm$^2$.
$s_{xy} = 31.05$ years·cm, $s_{\log(x)y} = 2.4594$ log(years)cm
Linear coef. determination: $r^2 = 0.8416$
Logarithmic coef. determination: $r^2 = 0.9013$
Therefore, both models fit pretty well, but the logarithmic model fits a little bit better.

(c) Logarithmic regression model: $y = 24.2639 + 50.8848 \log(x)$.
Prediction: $x(12.5) = 152.785$ cm.