Physiotherapy Statistics Exams

Alfredo Sánchez Alberca
asalber@ceu.es
https://aprendeconalf.es

Table of contents

Preface

Statistics exam collection of the Physiotherapy degree.

1 Descriptive Statistics and Regression exam (2018/05/31)

Exercise 1.1.

The ages of a sample of patients of a physical therapy clinic are:

a. Compute the quartiles.

b. Draw the box plot and identify outliers (do not group data into intervals).

c. Split the sample into two groups, patients younger and older than 65. In which group is the mean more representative. Justify the answer.

d. Which distribution is less symmetric, the one of patients younger than 65 or the one of patients older?

e. Which age is relatively smaller with respect to its group, 50 years in the group of patients younger than 65 or 72 years in the group of patients older than 65?

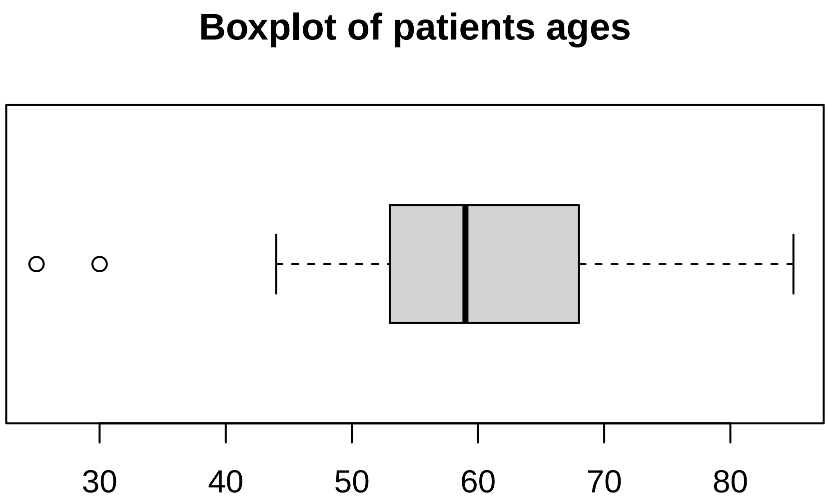
Use the following sums for the computations.
Younger than 65: $\sum x_i = 953$ years, $\sum x_i^2 = 52475$ years^2, $\sum(x_i - \bar{x})^3 = -30846.51$ years^3 and $\sum(x_i - \bar{x})^4 = 939658.83$ years^4.
Older than 65: $\sum x_i = 588$ years, $\sum x_i^2 = 43530$ years^2, $\sum(x_i - \bar{x})^3 = 1485$ years^3 and $\sum(x_i - \bar{x})^4 = 26983.5$ years^4.

Tip

1. $Q_1 = 53$ years, $Q_2 = 59$ years and $Q_3 = 68$ years.
2. There are 2 outliers: 25, 30.

25	30	44	44	51	51	53	56	57	58	58	58	59
59	61	63	63	63	66	68	70	71	72	74	82	85

Warm-up time	15	35	22	28	21	18	25	30	23	20
Injuries	42	2	16	6	17	29	10	3	12	20



2. Let x be the age in patients younger than 65 and y the age in patients older than 65.
   $\bar{x} = 52.9444$ years, $s_x^2 = 112.1636$ years^2, $s_x = 10.5907$ years and $cv_x = 0.2$.
   $\bar{y} = 73.5$ years, $s_y^2 = 39$ years^2, $s_y = 6.245$ years and $cv_y = 0.085$.
   The mean is more representative in patients older than 65 since the coefficient of variation is smaller.
3. $g_{1x} = -1.4426$ and $g_{1y} = 0.7621$, thus the distribution of ages of people younger than 65 is less symmetric.
4. The standard scores are $z_x(50) = -0.278$ and $z_y(72) = -0.2402$, thus 50 years is relative smaller in the group of people younger than 65.

Exercise 1.2.

The table below shows the number of injuries of several teams during a league and the average warm-up time of its players.

a. Draw the scatter plot.

b. Which regression model is more suitable to predict the number of injuries as a function of the warm-up time, the logarithmic or the exponential? Use that regression model to predict the expected number of injuries for a team whose players warm-up 20 minutes a day.

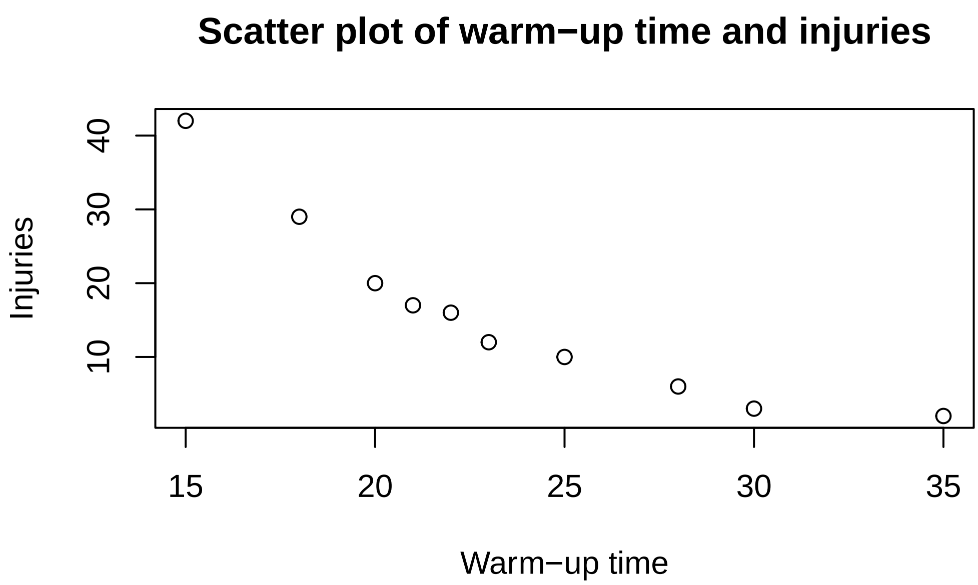
c. Which regression model is more suitable to predict the warm-up time as a function of the number of injuries, the logarithmic or the exponential? Use that regression model to predict the warm-up time required to have no more than 10 injuries in a league.

d. Are these predictions reliable? Which one is more reliable?

Use the following sums for the computations (X warm-up time and Y number of injuries):
$\sum x_i = 237$, $\sum \log(x_i) = 31.3728$, $\sum y_j = 157$, $\sum \log(y_j) = 24.0775$,
$\sum x_i^2 = 5937$, $\sum \log(x_i)^2 = 98.9906$, $\sum y_j^2 = 3843$, $\sum \log(y_j)^2 = 66.3721$,
$\sum x_i y_j = 3115$, $\sum x_i \log(y_j) = 519.1907$, $\sum \log(x_i) y_j = 465.8093$, $\sum \log(x_i)\log(y_j) = 73.3995$.

Solution

a.



a. $\bar{x} = 23.7$ min, $s_x^2 = 32.01$ min^2.
$\overline{\log(x)} = 3.1373$ log(min), $s_{\log(x)}^2 = 0.0565$ log(min)^2.
$\bar{y} = 15.7$ injuries, $s_y^2 = 137.81$ injuries^2.
$\overline{\log(y)} = 2.4078$ log(injuries), $s_{\log(y)}^2 = 0.8399$ log(injuries)^2.
$s_{x\log(y)} = -5.1446$, $s_{\log(x)y} = -2.6744$
Exponential determination coefficient: $r^2 = 0.9844$
Logarithmic determination coefficient: $r^2 = 0.9185$
So the exponential regression model es better to predict the number of injuries as a function of the warm-up time.
Exponential regression model: $y = e^{6.2168+-0.1607x}$.
Prediction: $y(20) = 20.1341$ injuries.

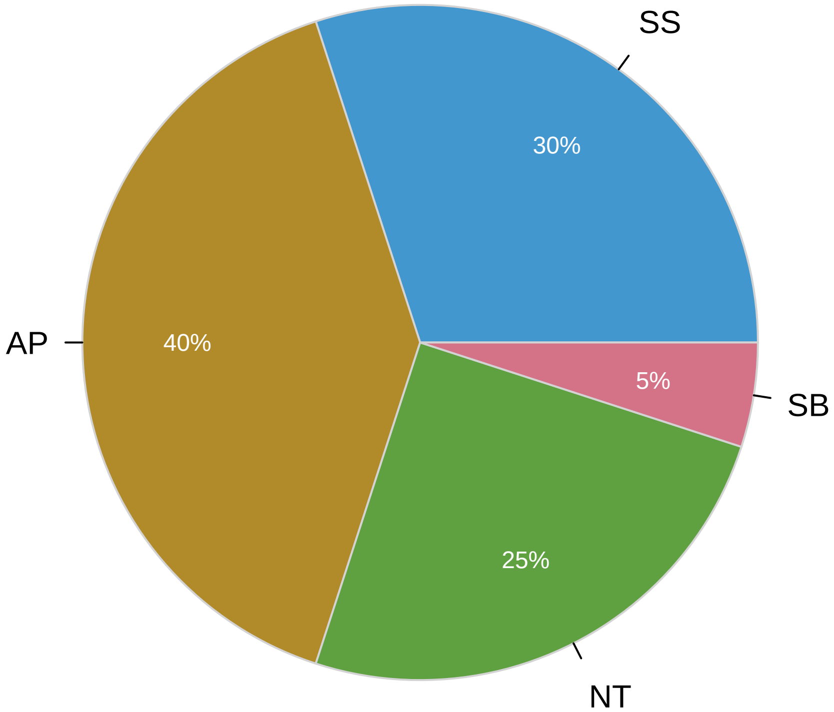
b. The logarithmic model is better to predict the warm-up time as a function of the number or injuries.
   Logarithmic regression model: $x = 164.1851 + -47.3292 \log(y)$.
   Prediction: $x(10) = 55.2056112$ min.

c. Both predictions are very reliable since de deternation coefficient is very high but the last one is a little less reliable as it is for a value further from the data range.

2 Descriptive Statistics and Regression exam (2023/03/23)

Exercise 2.1.

The chart below shows the percentage of grades in a Statistic course with 60 students.



a. Plot the ogive of the score, assuming the following correspondence between grades and scores

Grade	Score
SS	$[0, 5)$
AP	$[5, 7)$
NT	$[7, 9)$
SB	$[9, 10]$

a. Compute the median and interpret it.

b. How many students got a score greater than 8?

c. Study the dispersion of the distribution.

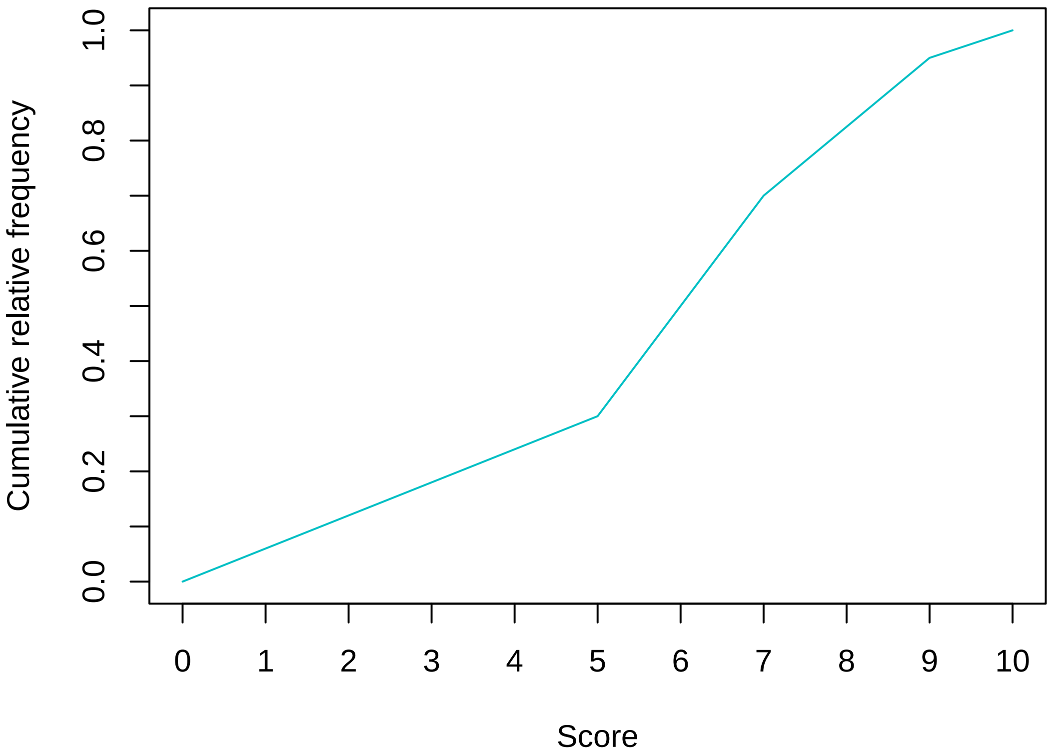
d. Study the skewness of the distribution. Is it normal?

e. If we apply the transformation $y = 10x + 5$ to the scores, how changes the representativeness of the mean. And the skewness?

Use the following sums for the computations (X = Score):
$\sum x_i n_i = 337.5$, $\sum x_i^2 n_i = 2207.25$, $\sum (x_i - \bar{x})^3 n_i = -172.55$ and $\sum (x_i - \bar{x})^4 n_i = 2870.75$.

Tip

a. Ogive



b. $Me = 6$ points.

c. $N(8) = 49.5$ students.

d. $\bar{x} = 5.625$ points, $s_x^2 = 5.1469$ points^2, $s_x = 2.2687$ points and $cv_x = 0.4033$. Thus, there is a moderate dispersion with respect to the mean.

e. $g_1 = -0.2463$ and therefore the distribution is a little bit left skewed.

f. $\bar{y} = 61.25$ points, $s_y^2 = 514.6875$ points^2, $s_y = 22.6867$ points and $cv_y = 0.3704$. As $cv_y < cv_x$ the representativeness of the mean increases. As the slope of the linear transformation is positive, the skewness does not change.

Exercise 2.2.

A study tries to determine if there is a relation between the gestation time (in weeks) and the age of the mother (in years). A sample of 40 mothers was taken and the sums below summarize the results (X=Age and Y=Gestation time):

$\sum x_i = 1262$ years, $\sum \log(x_i) = 137.0078$ log(years), $\sum y_j = 1583.6$ weeks, $\sum \log(y_j) = 147.1305$ log(weeks),
$\sum x_i^2 = 41862$ years^2, $\sum \log(x_i)^2 = 471.4222$ log(years)^2, $\sum y_j^2 = 62734.685$ weeks^2, $\sum \log(y_j)^2 = 541.2096$ log(weeks)^2,
$\sum x_i y_j = 50116.7$ years·weeks, $\sum x_i \log(y_j) = 4645.8$ years·log(weeks), $\sum \log(x_i) y_j = 5428.9192$ log(years)·weeks, $\sum \log(x_i)\log(y_j) = 504.0696$ log(years)·log(weeks).

a. Which regression models, linear, exponential or logarithmic, explains better the relation between the age and the gestation time?

b. Use the best model to predict the gestation time for a mother 45 years old. Is this prediction reliable?

c. According to the linear model, how much increases or decreases the gestation time for every year of the mother?

Solution

a. Linear model: $\overline{x} = 31.55$ years, $s_x^2 = 51.1475$ years^2.
$\bar{y} = 39.59$ weeks, $s_y^2 = 0.999$ weeks^2.
$s_{xy} = 3.853$ years·weeks.
$r^2 = 0.2905$.

Exponential model: $\overline{\ln(y)} = 3.6783$ ln(weeks), $s^2_{\ln(y)} = 0.0006$ ln(weeks)^2
$s_{x\ln(y)} = 0.0958$ years· ln(weeks).
$r^2 = 0.2882$.

Logarithmic model: $\overline{\ln(x)} = 3.4252$ ln(years), $s^2_{\ln(x)} = 0.0536$ ln(years)^2
$s_{\ln(x)y} = 0.1195$ ln(years)weeks.
$r^2 = 0.2668$.

As the linear coefficient of determination is greater, the linear model explains better the relation between de gestation time and the age of the mother.

b. Linear regression model of Y on X: $y = 37.2133 + 0.0753x$.
Predictions: $y(45) = 40.6032$ weeks.
The predictions are not reliable because the coefficient of determination is pretty low.

c. Regression coefficient of Y on X: $b_{yx} = 0.0753$ weeks/year. The gestation time increases 0.0753 weeks per year.