CEU San Pablo University — *Department of Mathematics and Data Science*

EXAM OF STATISTICS (Descriptive Statistics and Regression)

2nd Physiotherapy — Modelo A — **May, 30 2023**

1. To see if the confinement due to COVID-19 influenced the performance of a course, the number of failed subjects of each student in the current course and in the previous year course has been counted, obtaining the table below.

   Failed subjects

   Previous year course

   Current course

   0

   7

   8

   1

   15

   12

   2

   11

   8

   3

   5

   7

   4

   4

   3

   5

   2

   2

   6

   1

   2

   8

   0

   1

   a) Draw the box plots of the failed subjects in the current and the previous year courses and compare them.

   b) Can we assume that both samples come from a normal population?

   c) In which sample is the mean more representative?

   d) Which number of failed subjects is relatively greater, 7 in the current course or 6 in the previous year course?

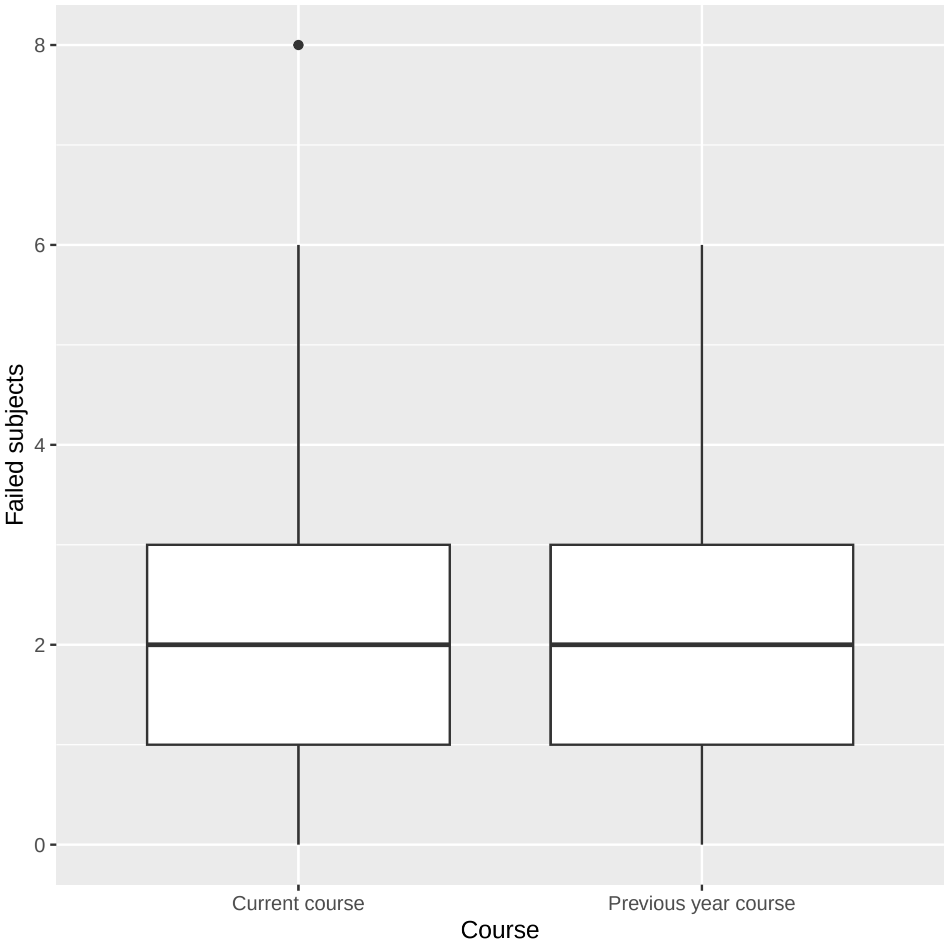

Figura 1: plot of chunk box-plot

Use the following sums for the computations:
Previous year course: $\sum x_i = 84$ subjects, $\sum x_i^2 = 254$ subjects^2, $\sum(x_i - \bar{x})^3 = 122{,}99$ subjects^3 y $\sum(x_i - \bar{x})^4 = 669{,}21$ subjects^4.
Current course: $\sum x_i = 91$ subjects, $\sum x_i^2 = 341$ subjects^2, $\sum(x_i - \bar{x})^3 = 301{,}16$ subjects^3 y $\sum(x_i - \bar{x})^4 = 2012{,}88$ subjects^4.

Solución

a) Box plot

a) Previous year course: $\bar{x} = 1{,}8667$ subjects, $s_x^2 = 2{,}16$ subjects^2, $s_x = 1{,}4697$ subjects, $g_1 = 0{,}8609$ and $g_2 = 0{,}1874$.
Current course: $\bar{x} = 2{,}1163$ subjects, $s_x^2 = 3{,}4516$ subjects^2, $s_x = 1{,}8578$ subjects, $g_1 = 1{,}0922$ and $g_2 = 0{,}9292$.
In both courses the coefficients of skewness and kurtosis are between -2 and 2, so we can assume that both samples come from a normal population.

b) Previous year course: $cv = 0{,}7873$.
Current year course: $cv = 0{,}8779$.
As the coefficient of variation of the previous year course is smaller, its mean is a little bit more representative.

c) Previous year course: $z(6) = 2{,}8124$.
Current course: $z(7) = 2{,}6287$.

Thus, 6 failed subjects is relative greater in the previous year course than 7 failed subjects in the current course.

---

2. The following table shows the reduction of inflammation in trauma (in percentage) for different doses of dexketoprofen given for 4 days (in mg).

Dose (mg)	50	62	71	75	82	90	96	102
Inflammation reduction (%)	38	45	60	68	70	86	88	95

a) Draw the scatter diagram of inflammation reduction versus dose of dexketoprofen.

b) What percentage of inflammation reduction variability does the linear model explain? And the logarithmic model?

c) According to the best of the two previous models, what is the expected percentage of inflammation reduction if we use 75 mg of dexketoprofen? Which dose should be administered to attain an inflammation reduction of 90 %? Are these predictions reliable?

Use the following sums for the computation (X=Dose, Y=inflammation reduction):
$\sum x_i = 628$ mg, $\sum \log(x_i) = 34{,}7152$ log(mg), $\sum y_j = 550\,\%$, $\sum \log(y_j) = 33{,}4922$ log(%),
$\sum x_i^2 = 51454$ mg^2, $\sum \log(x_i)^2 = 151{,}0394$ log(mg)^2, $\sum y_j^2 = 40758\,\%^2$, $\sum \log(y_j)^2 = 140{,}9659$ log(%)^2,
$\sum x_i y_j = 45668$ mg$\cdot$%, $\sum x_i \log(y_j) = 2668{,}6416$ mg$\cdot$log(%), $\sum \log(x_i) y_j = 2420{,}2169$ log(mg) %,
$\sum \log(x_i)\log(y_j) = 145{,}8748$ log(mg) log(%).

Solución

a)

a) **Linear model**
$\bar{x} = 78{,}5$ mg, $s_x^2 = 269{,}5$ mg^2.
$\bar{y} = 68{,}75\,\%$, $s_y^2 = 368{,}1875\,\%^2$.
$s_{xy} = 311{,}625$ mg$\cdot$%.
$r^2 = 0{,}9787$, so the linear model explains 97,87 % of the variability of the inflammation reduction.
Logarithmic model
$\overline{\log(x)} = 4{,}3394$ log(mg), $s^2_{\log(x)} = 0{,}0496$ log(mg)^2.
$s_{\log(x)y} = 4{,}1936$ log(mg)$\cdot$%.
$r^2 = 0{,}9637$, so the logarithmic model explains 96,37 % of the variability of the inflammation reduction.

b) The linear model fits better than the logarithmic one as its coefficient of determination is greater.
Regression line of Y on X: $y = -22{,}0202 + 1{,}1563x$.
Prediction: $y(75) = 64{,}7029\,\%$.
Regression line of X on Y: $x = 20{,}3117 + 0{,}8464y$.
Prediction: $x(90) = 96{,}4855$ mg.
Although the linear coefficient of determination is close to 1, the sample size is too small to consider these predictions reliable.

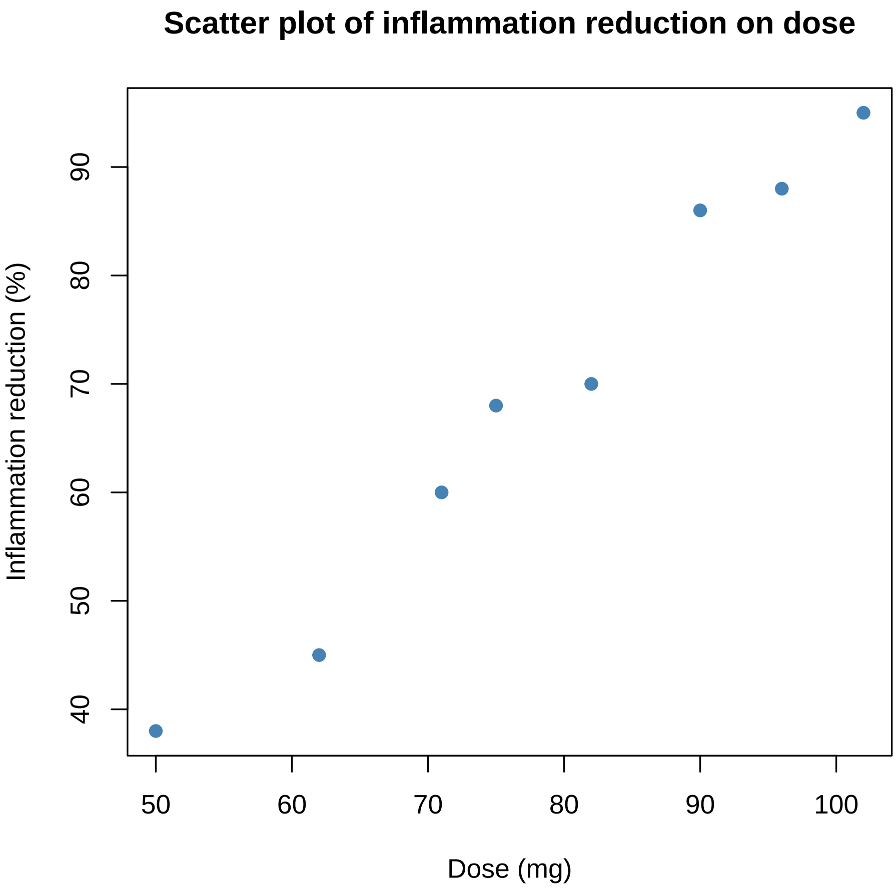

Figura 2: plot of chunk scatter plot