EXERCISES OF STATISTICS

Subject: Statistics Applied tho the Health Sciences
Course: 2^{nd}
Degree: Physiotherapy
Year: 2015-2016

Authors: Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
José Miguel Cárdenas Rebollo (cardenas@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)

CEU
Universidad San Pablo

Contents

1 Descriptive Statistics

1. Classify the following variables

   (a) Daily hours of exercise.

   (b) Nationality.

   (c) Blood pressure.

   (d) Severity of illness.

   (e) Number of sport injuries in a year.

   (f) Daily calorie intake.

   (g) Size of clothing.

   (h) Subjects passed in a course.

2. The number of injuries suffered by the members of a soccer team in a league were

0	1	2	1	3	0	1	0	1	2	0	1
1	1	2	0	1	3	2	1	2	1	0	1

   (a) Construct the frequency distribution table of the sample.

   (b) Draw the bar chart of the sample and the polygon.

   (c) Draw the cumulative frequency bar chart and the polygon.

3. A survey about the daily number of medicines consumed by people over 70 years, shows the following results:

3	1	2	2	0	1	4	2	3	5	1	3	2	3	1	4	2	4	3	2
3	5	0	1	2	0	2	3	0	1	1	5	3	4	2	3	0	1	2	3

   (a) Construct the frequency distribution table of the sample.

   (b) Draw the bar chart of the sample and the polygon.

   (c) Draw the cumulative relative frequency bar chart and the polygon.

4. In survey about the dependency of older people, 23 persons over 75 years were asked about the help they need in daily life. The answers were

B	D	A	B	C	C	B	C	D	E	A	B	C	E	A	B	C	D	B	B	A	A	B

   where the meanings of letters are:

   A No help.

   B Help climbing stairs.

   C Help climbing stairs and getting up from a chair or bed.

   D Help climbing stairs, getting up and dressing.

   E Help for almost everything.

   Construct the frequency distribution table and the suitable chart.

5. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

   (a) Construct the frequency distribution table of the sample.

   (b) Draw a suitable chart for the frequency distribution.

(c) Draw a suitable chart for the cumulative frequency distribution.

6. The following frequency distribution table represents the distribution of time (in min) required by people attended in a medical dispensary.

Time	n_i	f_i	N_i	F_i
$[0, 5)$	2			
$[5, 10)$			8	
$[10, 15)$				0.7
$[15, 20)$	6			

(a) Complete the table.

(b) Draw the ogive.

7. Use the data of exercise 2 to calculate the following statistics and interpret them.

(a) Mean.

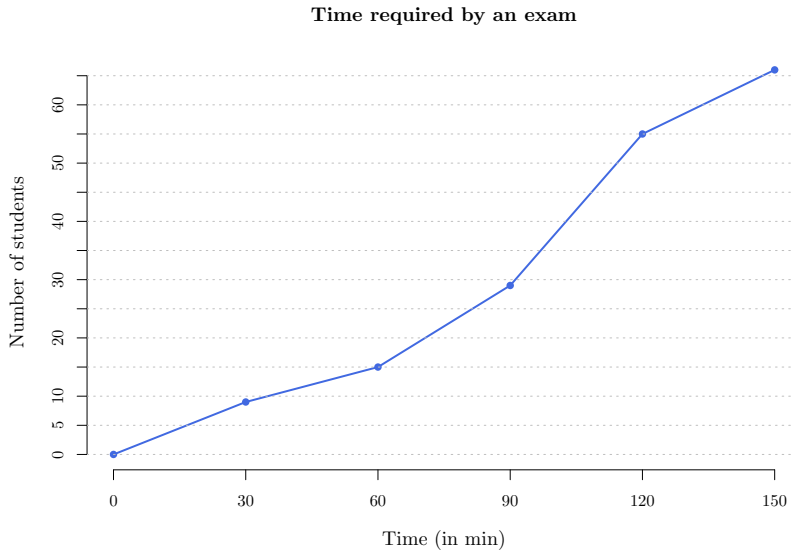
(b) Median.

(c) Mode.

(d) Quartiles.

(e) Percentile 32.

8. The chart below shows the cumulative distribution of the time (in min) required by 20 students to do an exam.



(a) A which time have finished half of the students? And 90% of students?

(b) Which percentage of students have finished after 100 minutes?

(c) Which is the time that best represent the time required by students in the sample to finish the exam? Is this value representative or not?

9. In a study about the children growth two samples where drawn, one for newborns and the other for one year old. The height in cm of children in both samples were

Newborn children: 51, 50, 51, 53, 49, 50, 53, 50, 47, 50
One year old children: 62, 65, 69, 71, 65, 66, 68, 69

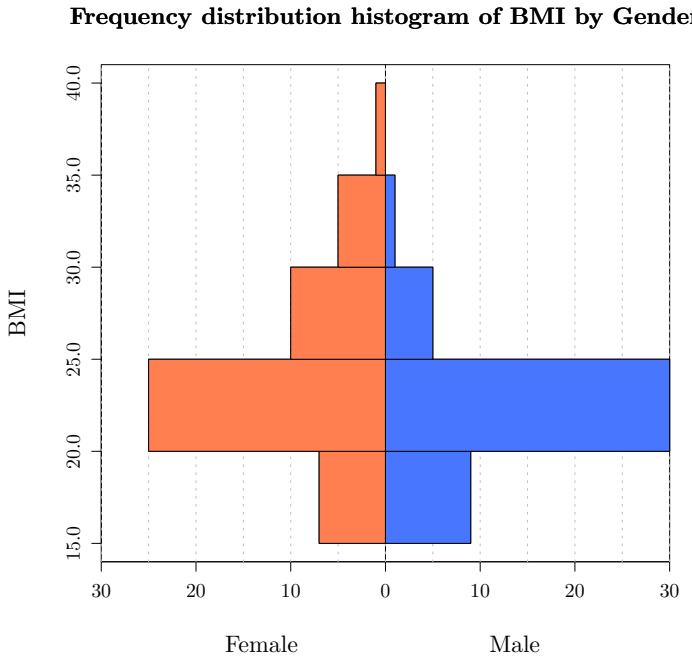
In which group is more representative the mean? Justify the answer.

10. To determine the accuracy of a method for measuring hematocrit in blood, the measurement was repeated 8 times on the same blood sample. The results in percentage of hematocrit in plasma were

    42.2 42.1 41.9 41.8 42 42.1 41.9 42

    What do you think about the accuracy of the method?

11. The histogram below shows the frequency distribution of the body mass index (BMI) of a group of people by gender.



    (a) Draw the pie chart for the gender.

    (b) In which group is more representative the mean of the BMI?

    (c) Calculate the mean for the whole sample.

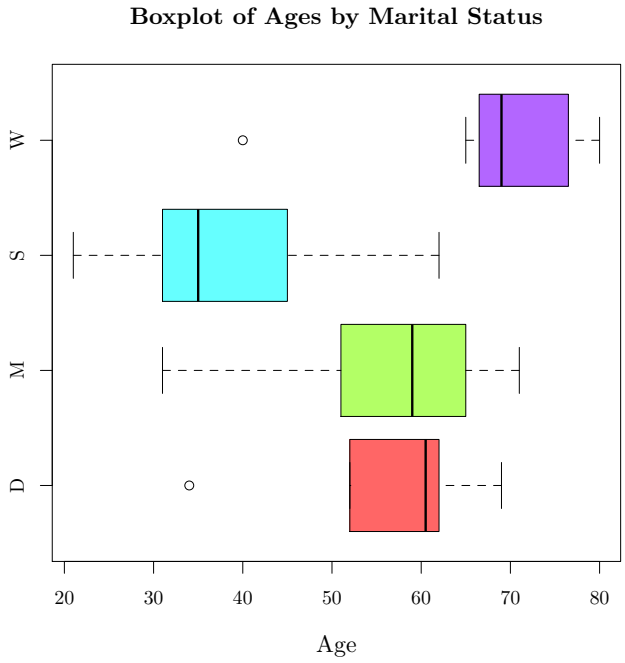
    Use the following sums
    Males: $\sum x_i = 1002$ kg/m^2 $\sum x_i^2 = 22781$ kg^2/m^4
    Females: $\sum x_i = 1160$ kg/m^2 $\sum x_i^2 = 29050$ kg^2/m^4

12. The following table represents the frequency distribution of the yearly uses of a health insurance in a sample of clients of a insurance company.

Uses:	0	1	2	3	4	5	7
Clients:	4	8	6	3	2	1	1

    Draw the box plot. How is the symmetry of the distribution?

13. The box plots below correspond to the age of a sample of people by marital status.



(a) Which group has higher ages?

(b) Which group has lower central dispersion?

(c) Which groups have outliers?

(d) Which group has a distribution of ages more asymmetric?

14. The following table represents the frequency distribution of ages at which a group of people suffered a heart attack.

Age	[40-50)	[50-60)	[60-70)	[70-80)	[80-90)
Persons	6	12	23	19	5

Could we assume that the sample comes from a normal population?

Use the following sums: $\sum x_i = 4275$ years, $\sum (x_i - \bar{x})^2 = 7462$ years^2, $\sum (x_i - \bar{x})^3 = -18249$ years^3, $\sum (x_i - \bar{x})^4 = 2099636$ years^4.

15. To compare two rehabilitation treatments A and B for an injury, every treatment was applied to a different group of people. The number of days required to cure the injury in every group is shown in the following table:

Days	A	B
20-40	5	8
40-60	20	15
60-80	18	20
80-100	7	7

(a) In which treatment is more representative the mean?

(b) In which treatment the distribution of days is more skew?

(c) In which treatment the distribution is more peaked?

Use the following sums:
A: $\sum x_i = 3040$ days, $\sum (x_i - \bar{x})^2 = 14568$ days^2, $\sum (x_i - \bar{x})^3 = 17011.2$ days^3, $\sum (x_i - \bar{x})^4 = 9989603$

days^4
B: $\sum x_i = 3020$ days, $\sum(x_i - \bar{x})^2 = 16992$ days^2, $\sum(x_i - \bar{x})^3 = -42393.6$ days^3, $\sum(x_i - \bar{x})^4 = 12551516$ days^4

16. The systolic blood pressure (in mmHg) of a sample of persons is

    135 128 137 110 154 142 121 127 114 103

    (a) Calculate the central tendency statistics.
    (b) How is the relative dispersion with respect to the mean?
    (c) How is the skewness of the sample distribution?
    (d) How is the kurtosis of the sample distribution?
    (e) If we know that the method used for measuring the blood pressure is biased, and, in order to get the right values, we have to apply the linear transformation $y = 1.2x - 5$, which are values of the statistics required to answer the previous questions for the corrected values of the blood pressure?

    Use the following sums: $\sum x_i = 1271$ mmHg, $\sum(x_i - \bar{x})^2 = 2188.9$ mmHg^2, $\sum(x_i - \bar{x})^3 = 2764.32$ mmHg^3, $\sum(x_i - \bar{x})^4 = 1040080$ mmHg^4.

17. The table below contains the frequency of pregnancies, abortions and births of a sample of 999 women in a city.

Num	Pregnancies	Abortions	Births
0	61	751	67
1	64	183	80
2	328	51	400
3	301	10	300
4	122	2	90
5	81	2	62
6	29	0	0
7	11	0	0
8	2	0	0

    (a) How many birth outliers are in the sample?
    (b) Which variable has lower spread with respect to the mean?
    (c) Which value is relatively higher, 7 pregnancies or 4 abortions? Justify the answer.

    Use the following sums:
    Pregnancies: $\sum x_i = 2783$, $\sum x_i^2 = 9773$.
    Abortions: $\sum x_i = 333$, $\sum x_i^2 = 559$.
    Births: $\sum x_i = 2450$, $\sum x_i^2 = 7370$.

2 Regression and correlation

18. Give some examples of:

    (a) Non related variables.
    (b) Variables that are increasingly related.
    (c) Variables that are decreasingly related.

19. In an study about the effect of different doses of a medicament, 2 patients got 2 mg and took 5 days to cure, 4 patients got 2 mg and took 6 days to cure, 2 patients got 3 mg ant took 3 days to cure, 4 patients got 3 mg and took 5 days to cure, 1 patient got 3 mg and took 6 days to cure, 5 patients got 4 mg and took 3 days to cure and 2 patients got 4 mg and took 5 days to cure.

(a) Construct the joint frequency table.

(b) Get the marginal frequency distributions and compute the main statistics for every variable.

(c) Compute the covariance and interpret it.

20. The table below shows the two-dimensional frequency distribution of a sample of 80 persons in a study about the relation between the blood cholesterol (X) in mg/dl and the high blood pressure (Y).

$X \setminus Y$	$[110, 130)$	$[130, 150)$	$[150, 170)$	n_x
$[170, 190)$		4		12
$[190, 210)$	10	12	4	
$[210, 230)$	7		8	
$[230, 250)$	1			18
n_y		30	24	

(a) Complete the table.

(b) Construct the linear regression model of cholesterol on pressure.

(c) Use the linear model to calculate the expected cholesterol for a person with pressure 160 mmHg.

(d) According to the linear model, what is the expected pressure for a person with cholesterol 270 mg/dl?

Use the following sums: $\sum x_i = 16960$ mg/dl, $\sum y_j = 11160$ mmHg, $\sum x_i^2 = 3627200$ (mg/dl)^2, $\sum y_j^2 = 1576800$ mmHg^2 y $\sum x_i y_j = 2378800$ mg/dl·mmHg.

21. A research study has been conducted to determine the loss of activity of a drug. The table below shows the results of the experiment.

Time (in years)	1	2	3	4	5
Activity (%)	96	84	70	58	52

(a) Construct the linear regression model of activity on time.

(b) According to the linear model, when will the activity be 80%? When will the drug have lost all activity?

22. A basketball team is testing a new stretching program to reduce the injuries during the league. The data below show the daily number of minutes doing stretching exercises and the number of injuries along the league.

Stretching minutes	0	30	10	15	5	25	35	40
Injuries	4	1	2	2	3	1	0	1

(a) Construct the regression line of the number of injuries on the time of stretching.

(b) What is the reduction of injuries for every minute of stretching?

(c) How many minutes of stretching are require for having no injuries? Is reliable this prediction?

Use the following sums (X=Number of minutes stretching, and Y=Number of injuries): $\sum x_i = 160$ min, $\sum y_j = 14$ injuries, $\sum x_i^2 = 4700$ min^2, $\sum y_j^2 = 36$ injuries^2 and $\sum x_i y_j = 160$ min·injuries.

23. For two variables X and Y we have

- The regression line of Y on X is $y - x - 2 = 0$.
- The regression line of X on Y is $y - 4x + 22 = 0$.

Calculate:

(a) The means $\bar{x}$ and $\bar{y}$.

(b) The correlation coefficient.

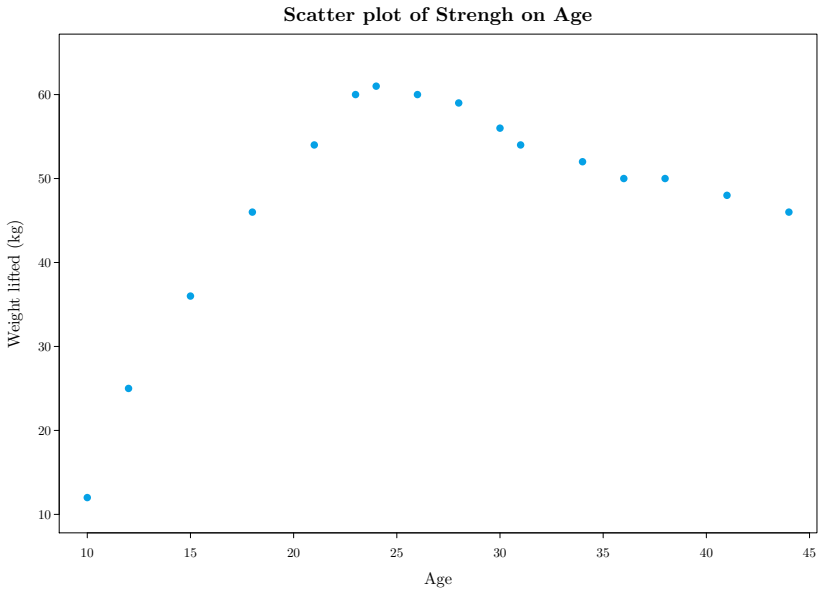
24. The means of two variables X and Y are $\bar{x} = 2$ and $\bar{y} = 1$, and the correlation coefficient is 0.

    (a) Predict the value of Y for $x = 10$.

    (b) Predict the value of X for $y = 5$.

    (c) Plot both regression lines.

25. A study to determine the relation between the age and the physical strength gave the scatter plot below.



    (a) Calculate the linear coefficient of determination for the whole sample.

    (b) Calculate the linear coefficient of determination for the sample of people younger than 25 years old.

    (c) Calculate the linear coefficient of determination for the sample of people older than 25 years old.

    (d) Which model explains better the relation between the age and the strength?

    Use the following sums (X=Age and Y =Weight lifted).

    - Whole sample: $\sum x_i = 431$ years, $\sum y_j = 769$ Kg, $\sum x_i^2 = 13173$ years^2, $\sum y_j^2 = 39675$ Kg^2 and $\sum x_i y_j = 21792$ years·Kg.
    - Young people: $\sum x_i = 123$ years, $\sum y_j = 294$ Kg, $\sum x_i^2 = 2339$ years^2, $\sum y_j^2 = 14418$ Kg^2 and $\sum x_i y_j = 5766$ years·Kg.
    - Old people: $\sum x_i = 308$ years, $\sum y_j = 475$ Kg, $\sum x_i^2 = 10834$ years^2, $\sum y_j^2 = 25257$ Kg^2 and $\sum x_i y_j = 16026$ years·Kg.

26. A dietary center is testing a new diet in sample of 12 persons. The data below are the number of days of diet and the weight loss (in Kg) until them for every person.

    (33 , 3.9), (51 , 5.9), (30 , 3.2), (55 , 6.0), (38 , 4.9), (62 , 6.2),
    (35 , 4.5), (60 , 6.1), (44 , 5.6), (69 , 6.2), (47 , 5.8), (40 , 5.3)

    (a) Draw the scatter plot. According to the point cloud, what type of regression model explains better the relation between the weight loss and the days of diet?

    (b) Construct the linear regression model and the logarithmic regression model of the weight loss on the number of days of diet.

(c) Use the best model to predict the weight that will lose a person after 100 days of diet. Is this prediction reliable?

Use the following sums (X=days of diet and Y=weight loss): $\sum x_i = 564$ days, $\sum \log(x_i) = 45.8086$ log(days), $\sum y_j = 63.6$ Kg, $\sum x_i^2 = 28234$ days^2, $\sum \log(x_i)^2 = 175.6603$ log(days)^2, $\sum y_j^2 = 347.7$ Kg^2, $\sum x_i y_j = 3108.5$ days·Kg, $\sum \log(x_i) y_j = 245.4738$ log(days)·Kg.

27. The concentration of a drug in blood, in mg/dl, depends on time, in hours, according to the data below.

Drug concentration	2	3	4	5	6	7	8
Hours	25	36	48	64	86	114	168

(a) Construct the linear regression model of drug concentration on time.

(b) Construct the exponential regression model of drug concentration on time.

(c) Use the best regression model to predict the drug concentration after 4.8 hours? Is this prediction reliable? Justify the answer.

Use the following sums (C=Drug concentration and T=time): $\sum c_i = 35$ mg/dl, $\sum \log(c_i) = 10.6046$ log(mg/dl), $\sum t_j = 541$ hours, $\sum \log(t_j) = 29.147$ log(hours), $\sum c_i^2 = 203$ (mg/dl)^2, $\sum \log(c_i)^2 = 17.5206$ log(mg/dl)^2, $\sum t_j^2 = 56937$ hours^2, $\sum \log(t_j)^2 = 124.0131$ log(hours)^2, $\sum c_i t_j = 3328$ mg/dl·hours, $\sum c_i \log(t_j) = 154.3387$ mg/dl· log(hours), $\sum \log(c_i) t_j = 951.6961$ log(mg/dl)·hours, $\sum \log(c_i) \log(t_j) = 46.08046$ log(mg/dl) · log(hours).

28. A researcher is studying the relation between the obesity and the response to pain. Te obesity is measured as the percentage over the ideal weight, and the response to pain as the nociceptive flexion pain threshold. The results of the study appears in the table below.

Obesity	89	90	75	30	51	75	62	45	90	20
Pain threshold	10	12	4	4.5	5.5	7	9	8	15	3

(a) According to the scatter plot, what model explains better the relation of the response to pain on the obesity?

(b) According to the best regression model, what is the response to pain expected for a person with an obesity of 50%? Is this prection reliable?

(c) According to the best regression model, what is the expected obesity for a person with a pain threshold of 10? Is this prediction reliable?

Use the following sums (X=Obesity and Y=Pain threshold): $\sum x_i = 627$, $\sum \log(x_i) = 40.3858$, $\sum y_j = 78$, $\sum \log(y_j) = 19.4119$, $\sum x_i^2 = 45141$, $\sum \log(x_i)^2 = 165.4516$, $\sum y_j^2 = 738.5$, $\sum \log(y_j)^2 = 40.0458$, $\sum x_i y_j = 5538.5$, $\sum x_i \log(y_j) = 1306.051$, $\sum \log(x_i) y_j = 327.3887$, $\sum \log(x_i) \log(y_j) = 80.1831$.