STATISTICS PROBLEMS

Subject: Mathematics
Course: 1st
Degree: Pharmacy
Year: 2016-2017

Authors: Pablo Ares Gastesi (pablo.aresgastesi@ceu.es)
Eduardo López Ramírez (elopez@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.

Solution

(a) Variable cuantitativa continua.
(b) Atributo nominal.
(c) Variable cuantitativa continua.
(d) Atributo ordinal.
(e) Variable cuantitativa discreta.
(f) Variable cuantitativa continua.
(g) Atributo ordinal.
(h) Variable cuantitativa discreta.

★ 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 a 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. 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) Mean.

(b) Median.

(c) Mode.

(d) Quartiles.

(e) Percentile 32.

Solution

(a) $\bar{x} = 1.125$ injuries.

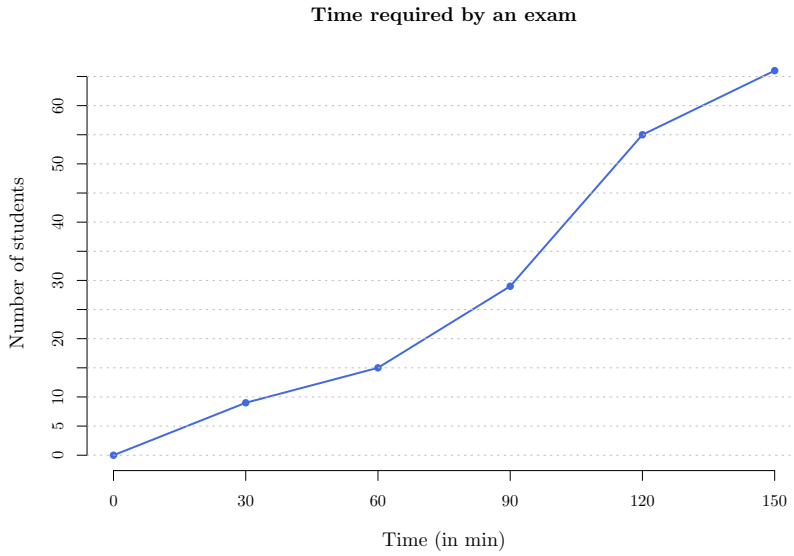
(b) $Me = 1$ injuries.

(c) $Mo = 1$ injuries.

(d) $C_1 = 0.5$ injuries, $C_2 = 1$ injuries y $C_3 = 2$ injuries.

(e) $P_{32} = 1$ injury.

★ 8. The chart below shows the cumulative distribution of the time (in min) required by 66 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?

Solution

(a) $Me = 94.6154$ min y $P_{90} = 132$ min.

(b) 57.08% of the students.

(c) $\bar{x} = 85.9091$ min, $s^2 = 1408.2645$ min^2, $s = 37.5268$ min y $cv = 0.4368$, so the dispersion is moderate.

★ 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.

Solution

Naming X to the height of a newborn children and Y to the height of a one year old children: $\bar{x} = 50.4$ cm, $s_x = 1.685$ cm, $cv_x = 0.034$, $\bar{y} = 66.875$ cm, $s_y = 2.713$ cm, $cv_y = 0.041$. Thus, both means are representative but the mean of newborn children is a little more representative.

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

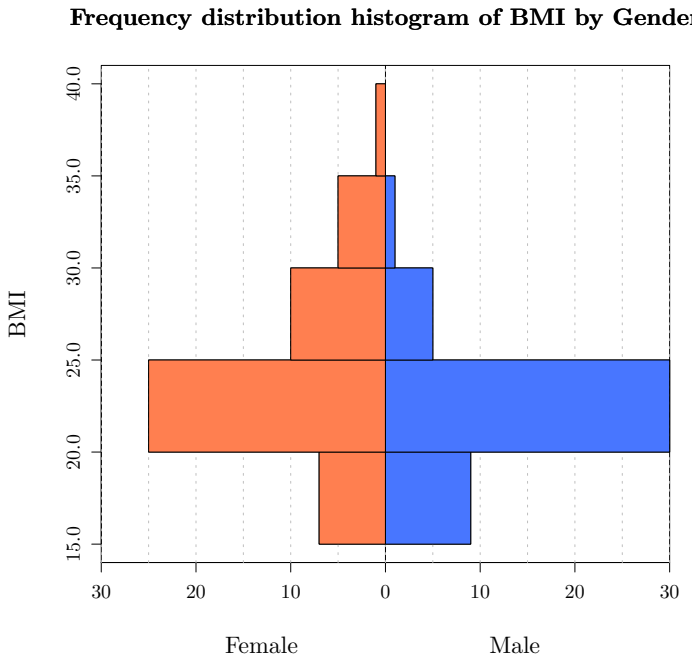
42.2 42.1 41.9 41.8 42 42.1 41.9 42

What do you think about the accuracy of the method?

Solution

$\bar{x} = 42\%$, $s^2 = 0.015\%^2$, $s = 0.1225\%$ and $cv = 0.003$, so the variability in the measurements is very low and the method has a high accuracy.

★ 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

Solution

(b) $\bar{m} = 24.17$ kg/m^2, $s_m^2 = 21.1806$ (kg/m^2)^2, $s_m = 4.6$ kg/m^2 and $cv_m = 0.19$.
$\bar{h} = 22.28$ kg/m^2, $s_m^2 = 9.9506$ (kg/m^2)^2, $s_m = 3.15$ kg/m^2 and $cv_h = 0.14$. Thus, the mean of males is more representative.

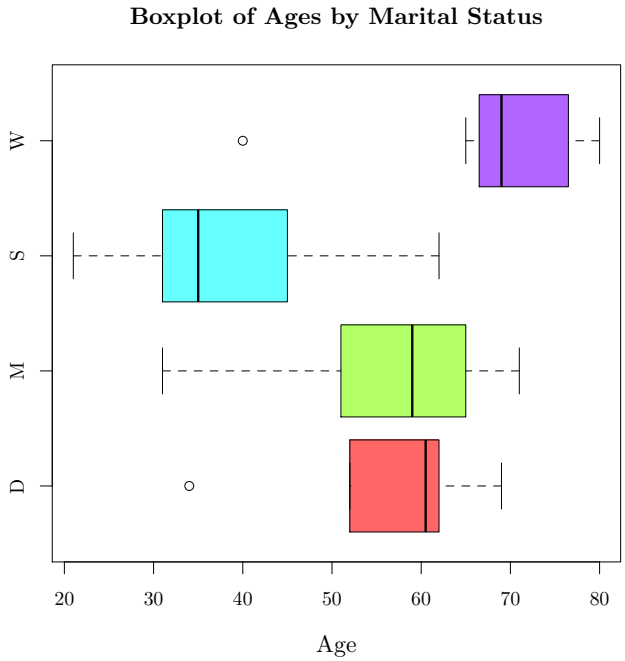
(c) $\bar{x} = 23.25$ kg/m^2.

★ 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?

Solution

(a) Widowers
(b) Divorced.
(c) Widowers and divorced.
(d) Divorced.

★ 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.

Solution

$\bar{x} = 65.769$ years, $s^2 = 114.823$ years, $s = 10.716$ years. $g_1 = -0.228$ and $g_2 = -0.55$. As the coefficient of skewness and kurtosis are both between -2 and 2, we can assume that both samples becomes from a normal population.

★ 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

Solution

(a) $\bar{x}_A = 60.8$ days, $s_A^2 = 291.36$ days^2, $s_A = 17.0693$ days and $cv_A = 0.28$.
$\bar{x}_B = 60.4$ days, $s_B^2 = 339.84$ days^2, $s_B = 18.4348$ days and $cv_B = 0.31$.
Thus, as $cv_A < cv_B$ the mean of treatment A is a little more representative.

(b) $g_{1_A} = 0.068$ and $g_{1_B} = -0.14$, so the distribution of treatment A is a little bit right skewed and the distribution of treatment B is a little bit left skewed, but both distributions are almost symmetric.

(c) $g_{2_A} = -0.65$ and $g_{2_B} = -0.83$, so the distribution of treatment A is flatter than a normal distribution (platykurtic) and the distribution of treatment B is more peaked than a normal distribution (leptokurtic).

★ 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.

Solution

Naming x to the systolic blood pressure.

(a) $\bar{x} = 127.1$ mmHg, $Me = 127.5$ mmHg, $Mo = 135$ mmHg.

(b) $s = 14.7949$ mmHg and $cv = 0.1164$.

(c) $g_1 = 0.0854$, so the distribution is almost symmetric.

(d) $g_2 = -0.8292$, so the distribution is flatter than a normal distribution (platykurtic).

(e) $\bar{y} = 147.52$ mmHg, $Me = 148$ mmHg, $Mo = 157$ mmHg, $s = 17.7539$ mmHg, $cv = 0.1203$, $g_1 = 0.0854$ and $g_2 = -0.8292$.

★ 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: $\sum x_i = 2450$, $\sum x_i^2 = 7370$.

Solution

Naming x to the number of pregnancies, y to the number of abortions and z to the number of births.

(a) 129 outliers.

(b) Pregnancies: $\bar{x} = 2.7858$, $s_x = 1.422$ and $cv_x = 0.5105$.
Abortions: $\bar{y} = 0.3333$, $s_y = 0.6697$ and $cv_y = 2.009$. Births: $\bar{z} = 2.4525$, $s_z = 1.1674$ and $cv_z = 0.476$.

(c) The standard score of 7 pregnancies is 2.9635, y de standard score of 4 abortions is 5.4754, so 4 abortions is relatively greater than 7 pregnancies.

---

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.

Solution

Naming X to the dose and Y to the curation time:

(c) $\bar{x} = 3.05$ mg, $\bar{y} = 4.55$ days, $s_x^2 = 0.648$ mg^2, $s_y^2 = 1.448$ days^2, $s_x = 0.805$ mg, $s_y = 1.203$ days and $s_{xy} = -0.678$ mg·days, so there is a decreasing relation.

★ 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) in mmHg.

$X \setminus Y$	$[110, 130)$	$[130, 150)$	$[150, 170)$	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.

Solution

(a) Frequency table

$X \setminus Y$	$[110, 130)$	$[130, 150)$	$[150, 170)$	n_x
$[170, 190)$	8	4	0	12
$[190, 210)$	10	12	4	26
$[210, 230)$	7	9	8	24
$[230, 250)$	1	5	12	18
n_y	26	30	24	80

(b) $\bar{x} = 212$ mg/dl, $\bar{y} = 139.5$ mmHg, $s_x^2 = 396$ (mg/dl)^2, $s_y^2 = 249.75$ mmHg^2 y $s_{xy} = 161$ mg/dl·mmHg. Regression line of cholesterol on pressure: $x = 122.0721 + 0.6446y$.

(c) $x(160) = 225.2152$ mg/dl.

(d) Regression line of pressure on cholesterol: $y = 0.4066x + 53.3081$.
$y(270) = 163.0808$ 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?

Solution

Naming T the time and A the drug activity:

(a) $\bar{t} = 3$ years, $\bar{a} = 72\%$, $s_t^2 = 2$ years^2, $s_a^2 = 264\%^2$, $s_{ta} = -22.8$ years$\cdot\%$.
Regression line of activity on time: $a = -11.4t + 106.2$.

(b) Regression line of time on activity: $t = -0.086a + 9.2182$.
$t(80) = 2.3091$ years and $t(0) = 9.2182$ years.

★ 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$\cdot$injuries.

Solution

(a) Regression line of Y on X: $y - 0.08x + 3.35$.

(b) For each minute more of stretching there will be 0.08 injuries less.

(c) To having no injuries at least 38.26 minutes of stretching are required. $r = -0.91$, and the prediction is quite reliable.

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.

Solution

(a) $\bar{x} = 8$ and $\bar{y} = 10$.

(b) $r = 0.5$.

---

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.

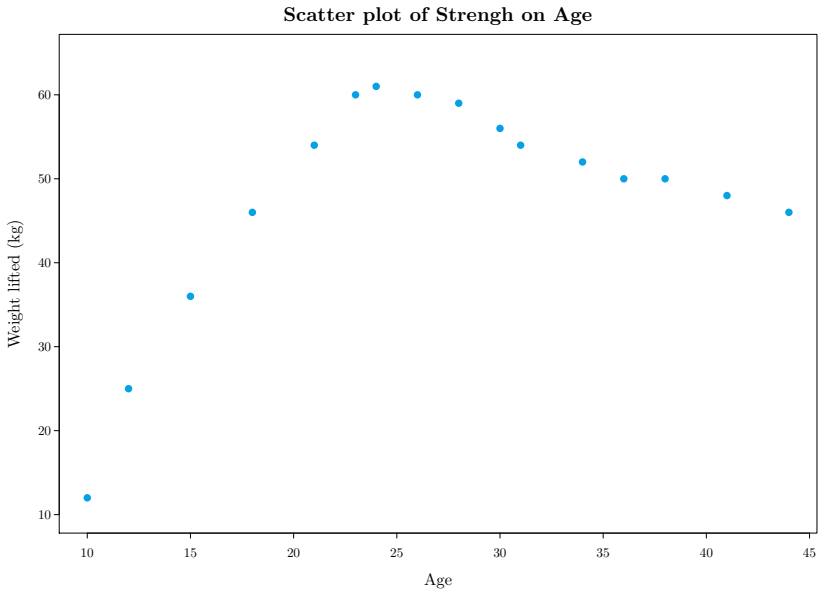
Solution

(a) $y(10) = 1$.

(b) $x(5) = 2$.

---

25. SA 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) For which group of ages the relation between the age and the strength is stronger?

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.

Solution

(a) $\bar{x} = 26.9375$ years, $s_x^2 = 97.6836$ years^2, $\bar{y} = 48.0625$ kg, $s_y^2 = 169.6836$ kg^2, $s_{xy} = 67.3164$ years·kg and $r^2 = 0.2734$.

(b) $\bar{x} = 15.5714$ years, $\bar{y} = 42$ kg, $s_x^2 = 25.3878$ years^2, $s_y^2 = 295.7143$ kg^2, $s_{xy} = 85.7143$ years·kg and $r^2 = 0.9786$.

(c) $\bar{x} = 35.2222$ years, $\bar{y} = 52.7778$ kg, $s_x^2 = 32.6173$ years^2, $s_y^2 = 20.8395$ kg^2, $s_{xy} = -25.5062$ years·kg and $r^2 = 0.9571$.

(d) The linear relation between age and physical strength is stronger in young people.

★ 26. A dietary center is testing a new diet in a 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 40 and 100 days of diet. Are these predictions 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.

Solution

Naming $Z = \log X$.

(b) $\bar{x} = 47$ days, $\bar{y} = 5.3$ kg, $s_x^2 = 143.833$ days^2, $s_y^2 = 0.885$ kg^2, $s_{xy} = 9.942$ days·kg.
Linear model: $y = 0.069x + 2.051$.
$\bar{z} = 3.82$ log(days), $s_z^2 = 0.07$ log^2(days), $s_{yz} = 0.22$ log(days) · kg.
Logarithmic model: $y = 3.4 \log y - 7.67$.

(c) Linear model: $r^2 = 0.78$, logarithmic model: $r^2 = 0.86$.
Predictions with the logarithmic model: $y(40) = 4.86$ kg and $y(100) = 7.98$ kg. The predictions are reliable since the coefficient of determination is high, although the prediction for 100 days is less reliable for being out of the range of observed values in the sample.

★ 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?

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).

Solution

Naming $Z = \log(C)$.

(a) $\bar{t} = 5$ hours, $\bar{c} = 77.2857$ mg/dl, $s_t^2 = 4$ hours^2, $s_c^2 = 2160.7755$ (mg/dl)^2, $s_{tc} = 89$ hours(mg/dl).
Linear model of C on T: $c = -33.9643 + 22.25t$.
$r^2 = 0.9165$.

(b) $\bar{z} = 4.1639$ log(mg/dl), $s_z^2 = 0.3785$ log^2(mg/dl), $s_{tz} = 1.2291$ hours· log(mg/dl).
Exponential model of C on T: $c = e^{0.3073t+2.6275}$.
$r^2 = 0.9979$.

(c) $c(4.8) = 60.498$ mg/dl and is quite reliable since the coefficient of determination is close to 1.

★ 28. A researcher is studying the relation between the obesity and the response to pain. The 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): $\sum x_i = 627$, $\sum \log(x_i) = 40.3858$, $\sum y_j = 78$, $\sum \log(y_j) = 19.4119$, $\sum x_i^2 = 45141$, $\sum \log(x_i)^2 = 165.4516$, $\sum y_j^2 = 738.5$, $\sum \log(y_j)^2 = 40.0458$, $\sum x_i y_j = 5538.5$, $\sum x_i \log(y_j) = 1306.051$, $\sum \log(x_i) y_j = 327.3887$, $\sum \log(x_i)\log(y_j) = 80.1831$.

Solution

(b) $\bar{x} = 62.9$, $s_x^2 = 588.09$, $\bar{y} = 9.22$, $s_y^2 = 11.0056$ and $s_{xy} = 82.0356$.
Regression line of response to pain on obesity: $y = 1.3232 + 0.1255x$. $r^2 = 0.8422$.
$\overline{\log(x)} = 4.0412$, $s_{\log(x)}^2 = 0.2366$ and $s_{\log(x)y} = 1.4973$. Logarithmic model of response to pain on obesity: $y = -16.3578 + 6.3293\log(x)$. $r^2 = 0.8611$.
$y(50) = 8.4023$.

(c) Exponential model of obesity on response to pain: $x = e^{2.7868+0.1361y}$.
$x(10) = 63.2648$.

★ 29. A blood bank keeps the plasma to a temperature of 0°F. When it is required for a blood transfusion, it is heated in an oven at a constant temperature of 120°F. In an experiment it has been measured the temperature of plasma at different times during the heating. The results are in the table below.

Time (min)	5	8	15	25	30	37	45	60
Temperature (°F)	25	50	86	102	110	114	118	120

(a) Plot the scatter plot. Which type of regression model do you think that explains better relationship between temperature and time?

(b) Which transformation should we apply to the variables to have a linear relationship?

(c) Compute the logarithmic regression of the temperature on time.

(d) According to the logarithmic model, what temperature there will be after 15 minutes of heating? Is this prediction reliable? Justify the answer.

Use the following sums (X=Time and Y=Temperature): $\sum x_i = 225$ min, $\sum \log(x_i) = 24.5289$ log(min), $\sum y_j = 725$ °F, $\sum \log(y_j) = 35.2051$ log(°F), $\sum x_i^2 = 8833$ min^2, $\sum \log(x_i)^2 = 80.4703$ log(min)^2, $\sum y_j^2 = 74345$ °F^2, $\sum \log(y_j)^2 = 157.1023$ log(°F)^2, $\sum x_i y_j = 24393$ min·°F, $\sum x_i \log(y_j) = 1048.0142$ min· log(°F), $\sum \log(x_i) y_j = 2431.7096$ log(min)°F, $\sum \log(x_i) \log(y_j) = 111.1165$ log(min) log(°F).

Solution

(a) A logarithmic model.

(b) Apply a logarithmic transformation to time, $z = \log(x)$.

(c) $\bar{z} = 28.125$ log(min), $s_z^2 = 0.6577$ log^2(min), $\bar{y} = 90.625$ °F, $s_y^2 = 1080.2344$ °F^2 and $s_{zy} = 26.0969$ log(min)°F.
Logarithmic model of temperature on time: $y = -31.0325 + 39.6781 \log(x)$.

(d) $y(15) = 76.4176$ °F. $r^2 = 0.9586$, that is close to 1, so the prediction is reliable.

30. The activity of a radioactive substance depends on time according to the data in the table below.

t (hours)	0	10	20	30	40	50	60	70
A (10^7 disintegrations/s)	25.9	8.16	2.57	0.81	0.25	0.08	0.03	0.01

(a) Represent graphically the data of radioactivity as a function of time. Which type of regression model explains better the relationship between radioactivity and time?

(b) Represent graphically the data of radioactivity as a function of time in a semi-logarithmic paper.

(c) Compute the regression line of the logarithm of radioactivity on time.

(d) Taking into account that radioactivity decay follows the formula

$$A(t) = A_0 e^{-\lambda t}$$

where A_0 is the number of disintegrations at the beginning and λ is a disintegration constant, different for each radioactive substance, use the slope of the previous regression line to compute the disintegration constant for the substance.

Use the following sums (X=Time and Y=Radioactivity): $\sum x_i = 280$ hours, $\sum y_j = 37.81$ 10^7 disintegrations/s, $\sum \log(y_j) = -5.9371$ log(10^7 disintegrations/s), $\sum x_i^2 = 14000$ hours^2, $\sum y_j^2 = 744.7265$ (10^7 disintegrations/s)^2, $\sum \log(y_j)^2 = 57.7369$ log(10^7 disintegrations/s)^2, $\sum x_i y_j = 173.8$ hours·10^7 disintegrations/s, $\sum x_i \log(y_j) = -680.9447$ hours· log(10^7disintegrations/s).

Solution

Naming $Z = \log(Y)$.

(c) $\bar{x} = 35$ hours, $\bar{z} = -0.7421$ 10^7 disintegrations/s, $s_x^2 = 525$ hours^2, $s_z^2 = 6.6664$ (10^7 disintegrations/s)^2 and $s_{xz} = -59.1434$ hours·10^7 disintegrations/s.
Regression line of the logarithm of radioactivity on time: $z = -0.1127x + 3.2008$.

(d) $\lambda = 0.1127$.

---

31. For oscillations of small amplitude, the oscillation period T of a pendulum is given by the formula

$$T = 2\pi\sqrt{\frac{L}{g}}$$

where L is the length of the pendulum and g is the gravitational constant. In order to check if the previous formula is satisfied, an experiment has been conducted where it has been measured the oscillation period for different lengths of the pendulum. The measurements are shown in the table below.

L (cm)	52.5	68.0	99.0	116.0	146.0
P (seg)	1.449	1.639	1.999	2.153	2.408

(a) Represent graphically the data of the period versus the length of the pendulum. Does a linear model fit well to the points cloud?

(b) Represent graphically the data of the period versus the length in a logarithmic paper. Which type of model fits better to the points cloud?

(c) Compute the regression line of the logarithm of period on the logarithm of length.

(d) Taking in to account the independent term of the previous regression line, compute the value of g.

Solution

Naming X to the logarithm of length and Y to the logarithm of period.

(c) $\bar{x} = 4.5025$ log(cm), $\bar{y} = 0.6407$ log(s), $s_x^2 = 0.1353$ log^2(cm), $s_y^2 = 0.0339$ log^2(s), $s_{xy} = 0.0677$ log(cm) log(s).
Regression line of Y on X: $y = 0.5006x - 1.6132$.

(d) $g = 994.4579$ cm/s^2.

★ 32. A study tries to determine the relationship between two substances X and Y in blood. The concentration of these substances has been measured in seven individuals (in μg/dl) and the results are shown in the table below.

X	2.1	4.9	9.8	11.7	5.9	8.4	9.2
Y	1.3	1.5	1.7	1.8	1.5	1.7	1.7

(a) Are Y and X linearly related?

(b) Are Y and X potentially related?

(c) Use the best of the previous regression models to predict the concentration in blood of Y for $x = 8$ μgr/dl. Is this prediction reliable. Justify the answer.

Use the following sums: $\sum x_i = 52$ μg/dl, $\sum \log(x_i) = 13.1955$ log(μg/dl), $\sum y_j = 11.2$ μg/dl, $\sum \log(y_j) = 3.253$ log(μg/dl), $\sum x_i^2 = 451.36$ (μg/dl)^2, $\sum \log(x_i)^2 = 26.9397$ log(μg/dl)^2, $\sum y_j^2 = 18.1$ (μg/dl)^2, $\sum \log(y_j)^2 = 1.5878$ log(μg/dl)^2, $\sum x_i y_j = 86.57$ (μg/dl)^2, $\sum x_i \log(y_j) = 26.3463$ μg/dl $\cdot$ log(μg/dl), $\sum \log(x_i) y_j = 21.7087$ log(μg/dl) $\cdot$ μg/dl, $\sum \log(x_i)\log(y_j) = 6.5224$ log(μg/dl)^2.

Solution

(a) Linear relation: $r^2 = 0.9696$, so there is a strong linear relation.

(b) Potential relation: $r^2 = 0.9688$, so there is a strong potential relation, however the linear relation is a little bit stronger.

(c) $y(8) = 1.6296$ μgr/dl.

---

3 Probability

33. Construct the sample space of the following random experiments:

    (a) Pick a random person and measure the gender and whether she or he is smoker or not.

    (b) Pick a random person and measure the blood type and whether she or he is smoker or not.

    (c) Pick a random person and measure the gender, the blood type and whether she or he is smoker or not.

★ 34. There are two boxes with balls of different colors. The first box contains 3 white balls and 2 black balls, and the second one contains 2 blue balls, 1 red ball and 1 green ball. Construct the sample space of the following random experiments:

    (a) Pick a random ball from every box.

    (b) Pick two random balls from every box.

★ 35. The Morgan's laws state that given two events A and B from the same sample space, $\overline{A \cup B} = \overline{A} \cap \overline{B}$ and $\overline{A \cap B} = \overline{A} \cup \overline{B}$. Proof both assertions graphically using Venn diagrams.

★ 36. Compute the probability of the following events of the random experiment consisting in tossing 3 coins:

    (a) Get exactly 1 heads.

    (b) Get exactly 2 tails.

    (c) Get two or more heads.

    (d) Get some tails.

Solution

(a) $P(1 \text{ heads}) = 0.375$.

(b) $P(2 \text{ tails}) = 0.375$.

(c) $P(2 \text{ or more heads}) = 0.5$.

(d) $P(\text{some tails}) = 0.875$.

---

37. In a laboratory there are 4 flasks with sulfuric acid and 2 with nitric acid, and in another laboratory there are 1 flask with sulfuric acid and 3 with nitric acid. A random experiment consist in picking two flask, one from every laboratory. Compute the probability of the following events:

    (a) The two drawn flasks are of sulfuric acid.

    (b) The two drawn flasks are of nitric acid.

    (c) The two drawn flasks contains different acids.

Calculate the same probabilities if the flask drawn in the first laboratory is put in the second laboratory before drawing the flask from it.

Solution

(a) 4/24.
(b) 6/24.
(c) 14/24.
(d) 8/30, 8/30 and 14/30 respectively.

38. Let A and B be two events of the same sample space, such that $P(A) = 3/8$, $P(B) = 1/2$ and $P(A \cap B) = 1/4$. Compute the following probabilities:

(a) $P(A \cup B)$.
(b) $P(\overline{A})$ and $P(\overline{B})$.
(c) $P(\overline{A} \cap \overline{B})$.
(d) $P(A \cap \overline{B})$.
(e) $P(A|B)$.
(f) $P(A|\overline{B})$.

Solution

(a) $P(A \cup B) = 5/8$.
(b) $P(\overline{A}) = 5/8$ and $P(\overline{B}) = 1/2$.
(c) $P(\overline{A} \cap \overline{B}) = 3/8$.
(d) $P(A \cap \overline{B}) = 1/8$.
(e) $P(A|B) = 1/2$.
(f) $P(A|\overline{B}) = 1/4$.

39. In a hospital the probability of getting hepatitis in a blood transfusion from a unit of blood is 0.01. A patient gets two units of blood while staying at the hospital. What is the probability of getting hepatitis?

Solution

0.0199.

40. Let A and B be two events of the same sample space, such that $P(A) = 0.6$ and $P(A \cup B) = 0.9$. Compute $P(B)$ with the following assumptions:

(a) A and B are incompatible.
(b) A and B are independent.

Solution

(a) $P(B) = 0.75$.
(b) $P(B) = 0.3$.

★ 41. A study about smoking has published that 40% of smokers have a smoker father, 25% have a smoker mother and 52% have al least one of the parents smoker. We pick a random person from this population. Answer the following questions:

(a) What is the probability of having a smoker mother if the father smokes?

(b) What is the probability of having a smoker mother if the father does not smoke?

(c) Are independent the events having a smoker father and having a smoker mother?

Solution

Naming SF to the event of having a smoker father, and SM tho the event of having a smoker mother,

(a) $P(SM|SF) = 0.33$.

(b) $P(SM|\overline{SF}) = 0.2$.

(c) They are not independent.

★ 42. In a study to determine the relation between hypercholesterolemia and hypertension a random sample of 1000 persons has been drawn. In the sample there was 180 persons with hypertension, 140 with hypercholesterolemia and 800 with none of them. Compute the probability that a random person,

(a) Have both diseases.

(b) Have hypertension if he or she does not have hypercholesterolemia.

Solution

Naming HT to the event of having hypertension and HC to the event of having hypercholesterolemia,

(a) $P(HT \cap HC) = 0.12$.

(b) $P(HT|\overline{HC}) = 0.0698$.

★ 43. The probability that an injury A is repeated is $4/5$, the probability that another injury B is repeated is $1/2$, and the probability that both injuries are repeated is $1/3$. Compute the probability of the following events:

(a) Only injury B is repeated.

(b) At least one injury is repeated.

(c) Injury B is repeated if injury A has been repeated.

(d) Injury B is repeated if injury A hasn't been repeated.

Solution

Naming A and B to the events of repetition of injuries A and B respectively,

(a) $P(B \cap \overline{A}) = 1/6$.

(b) $P(A \cup B) = 29/30$.

(c) $P(B|A) = 5/12$.

(d) $P(B|\overline{A}) = 5/6$.

★ 44. A student have to take a test where each questions has 3 possible answers. The student knows 40% of the questions an he answer the other ones randomly. I we take a random question, what is the probability that the student does not know the answer of the question if his answer was correct?

Solution

1/3.

★ 45. In a digestive clinic from every 1000 patients that arrive with stomach pain, 700 have gastritis, 200 have an ulcer and 100 have cancer. After analyzing the gastric symptoms, it is known that the probability of having vomiting is 0.3 in case of gastritis, 0.6 in case of ulcer and 0.9 in case of cancer. What is the diagnostic for a new patient with stomach pain that has vomiting?

Note: Assume that the only diseases are gastritis, ulcer and cancer and that are incompatible among them.

Solution

Naming G to having gastritis, U to having ulcer, C to having cancer and V to having vomiting, $P(G|V) = 0.5$, $P(U|V) = 0.286$ and $P(C|V) = 0.214$. Therefore, the diagnostic is gastritis.

★ 46. To evaluate the effectiveness of a diagnostic test, the test was applied to a sample of people with the following results:

	Test +	Test −
Sick	2020	80
Healthy	140	7760

Compute for this test:

(a) The sensitivity and the specificity.

(b) The positive and negative predictive values.

(c) The probability of a correct diagnostic.

Solution

Naming D and $\overline{D}$ to the events of being sick and healthy respectively,

(a) Sensitivity $P(+|D) = 0.9352$ and specificity $P(-|\overline{D}) = 0.9898$.

(b) PPV $P(D|+) = 0.9619$ and NPV $P(\overline{D}|-) = 0.9823$.

(c) $P((D \cap +) \cup (\overline{D} \cap -)) = 0.978$.

★ 47. A new test for detecting the Down syndrome in newborns has a sensitivity of 80% and a specificity of 90%. If there are 1% of newborns with the Down syndrome in the popultion, and after applying the test to one newborn the outcome of the test is positive, what is the proability that the newborn have the Down syndrome? Will we diagnose the syndrome? What must the minimum specificity of the test be to diagnose the syndrome after a positive outcome?

Remark: The *sensitivity* of a test is the proportion of people with the disease that have a positive outcome in the test, while the *specificity* of the test is the proportion of people without the disease that have a negative outcome in the test.

Solution

Naming S to the event of having the Down syndrome and $+$ to the event of having a positive outcome of the test, $P(S|+) = 0.0748$ and $P(\overline{S}|+) = 0.9252$, so we wont diagnose the syndrome.
The minimum specificity to diagnose the syndrome after a positive outcome is $P(-|\overline{S}) = 0.9919$.

★ 48. After applying a diagnostic test to a population we got 1% of sick persons with a negative outcome of the test, 2% of healthy persons with a positive outcome of the test and 90% of healthy persons with a negative outcome of the test.

(a) Compute the prevalence of the disease.

(b) Compute the sensitivity of the test.

(c) Compute the specificity of the test.

Solution

Naming D to having the disease and, $+$ to having a positive outcome of the test and $-$ to having a negative outcome of the test,

(a) $P(D) = 0.08$.

(b) $P(+|D) = 0.875$.

(c) $P(-|\overline{D}) = 0.9783$.

★ 49. The third part of a population has been vaccinated against the flu. After the winter, it is found that the probability of having been vaccinated if a person suffered the flu is 0.2, and that 10% of vaccinated people suffered the flu.

(a) Compute the incidence of the flu.
Remark: The incidence of a epidemic is the percentage of infected people.

(b) What is the probability that a non vaccinated person suffers the flu?

(c) Can we say that the vaccine is effective?

Solution

Naming F to having suffered the flu and V to having been vaccinated,

(a) $P(F) = 1/6$.

(b) $P(F|\overline{V}) = 0.2$.

(c) Yes, it is effective, but not to much.

★ 50. We have two different test A and B to diagnose a disease. Test A have a sensitivity of 98% and a specificity of 80%, while test B have a sensitivity of 75% and a specificity of 99%.

(a) What test is better to confirm the disease?

(b) What test is better to rule out the disease?

(c) Often a test is used to discard the presence of the disease in a large amount of people apparently healthy. This type of test is known as *screening test*. What test will work better as a screening test? What is the positive predictive value (PPV) of this test if the prevalence of the disease is 0.01? And if the prevalence of de disease is 0.2?

(d) The positive predictive value of a screening test used to be not too high. How can combine tests A and B to have a higher confidence in the diagnosis of the disease? Calculate the post-test probability of having the disease with the combination of both tests if the outcome of both test is positive for a prevalence of 0.01.

Solution

Naming D to the event of having the disease,

(a) The test B, since it has a greater specificity.

(b) The test A, since it has a greater sensitivity.

(c) The test A would work better as a screening test.
For a prevalence of 0.01 the PPV is $P(D|+) = 0.0472$ and the NPV is $P(\overline{D}|-) = 0.9997$.
For a prevalence of 0.2 the PPV is $P(D|+) = 0.5506$ and the NPV is $P(\overline{D}|-) = 0.9938$.

(d) Applying first the test A to everybody and then the test B to people with a positive outcome of A.
$P(D| + A \cap +B) = 0.7878$.

★ 51. A disease is treated with 3 different medicines: A in 50% of the cases, B in 30% of the cases and C in 20% of the cases, independently of the gender. If we know that medicine A produces side effects in 5% of males and 10% of females, medicine B in 15% of males and 5% of females, and medicine C in 8% of males and 13% of females,

(a) Which gender is more probable to have side effects? Justify the answer.

(b) Compute the probability that a male with side effects had been treated with medicine C, and that a female with no side effects had been treated with medicine A.

(c) If there are 65% of males and 35% of females in the population, what is the probability of being female if a person has no side efects?

Solution

Naming E to the event of having side effects, M to the event of being male and F to the event of being female,

(a) $P(E \cap M) = 0.086$ and $P(E \cap F) = 0.091$.

(b) $P(C|E \cap M) = 0.186$, and $P(A|\overline{E \cap F}) = 0.495$.

(c) $P(F|\overline{E}) = 0.349$.

Remark: The problems with a (★) are exam problems of previous years.