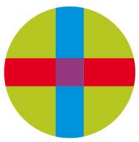
STATISTICS PROBLEMS

Subject: Statistics Applied to Health Sciences
Course: 2nd
Degree: Physiotherapy
Year: 2021-2022
Authors: 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) Continuous.
(b) Nominal.
(c) Continuous.
(d) Ordinal.
(e) Discrete.
(f) Continuous.
(g) Ordinal.
(h) Discrete.

★ 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

Compute:

(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 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 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 a 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.

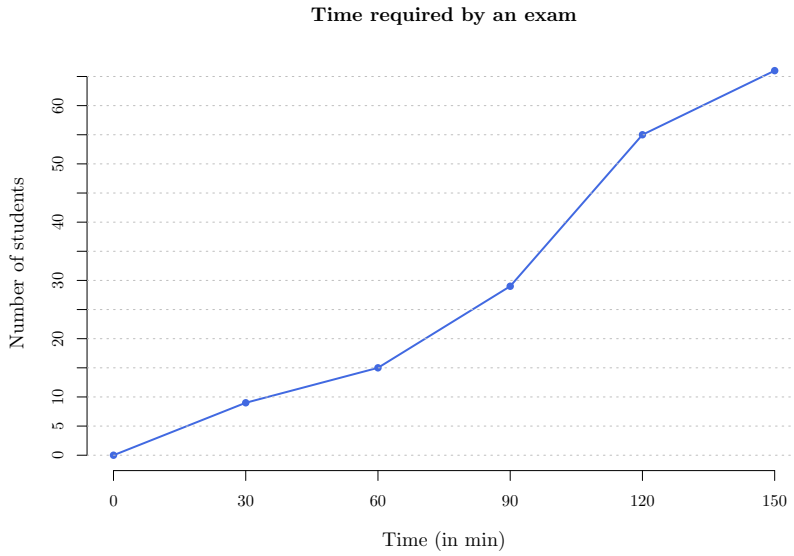
(b) $Me = 1$ injuries.

(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) At what time have half of the students finished? And 90% of students?

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

(c) What 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 children's growth, two samples were drawn, one for newborn babies and the other for one year old infants. The heights in cm of children in each of the 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 the mean more representative? Justify your answer.

Solution

Let X be the height of a newborn child and Y the height of a one year old child $\bar{x} = 50.4$ 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 of hematocrit in plasma, in percentage, were

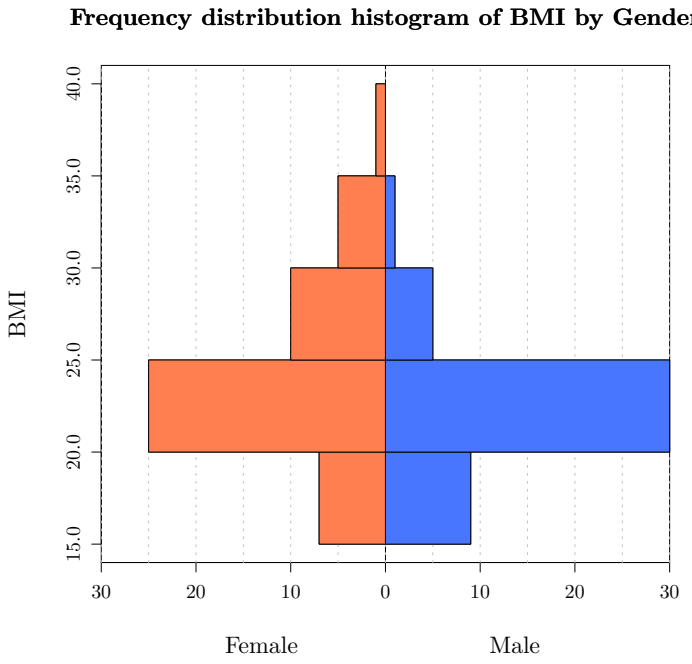
42.2 42.1 41.9 41.8 42 42.1 41.9 42

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
Females: $\sum x_i n_i = 1160$ kg/m^2 $\sum x_i^2 n_i = 29050$ kg^2/m^4
Males: $\sum x_i n_i = 1002$ kg/m^2 $\sum x_i^2 n_i = 22781$ kg^2/m^4

Solution

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

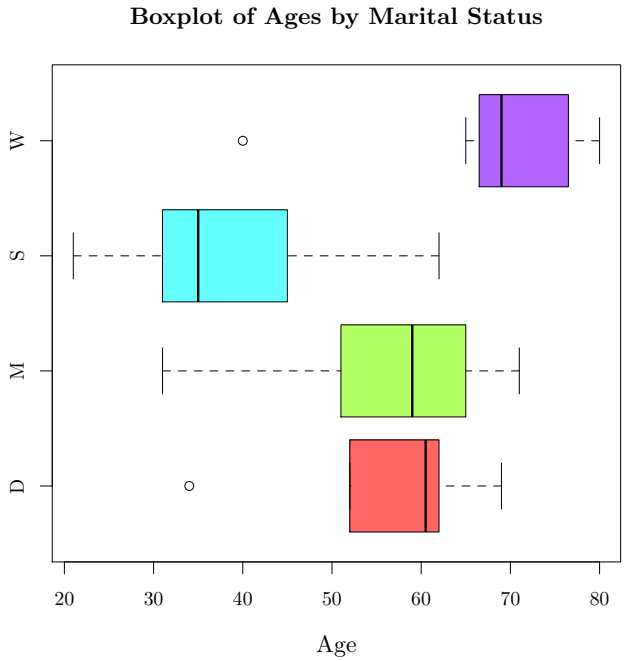
(c) $\bar{x} = 23.25$ kg/m^2.

★ 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. Study 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) At which group is the age distribution 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: $\sum x_i n_i = 4275$ years, $\sum (x_i - \bar{x})^2 n_i = 7462$ years^2, $\sum (x_i - \bar{x})^3 n_i = -18249$ years^3, $\sum (x_i - \bar{x})^4 n_i = 2099636$ years^4.

Solution

$\bar{x} = 65.769$ years, $s^2 = 114.823$ years^2, $s = 10.716$ 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 each 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: $\sum x_i n_i = 3040$ days, $\sum (x_i - \bar{x})^2 n_i = 14568$ days^2, $\sum (x_i - \bar{x})^3 n_i = 17011.2$ days^3, $\sum (x_i - \bar{x})^4 n_i = 9989603$ days^4
B: $\sum x_i n_i = 3020$ days, $\sum (x_i - \bar{x})^2 n_i = 16992$ days^2, $\sum (x_i - \bar{x})^3 n_i = -42393.6$ days^3, $\sum (x_i - \bar{x})^4 n_i = 12551516$ 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 $y = 1.2x - 5$, what are the statistics values of parts (a) to (d) for the new, corrected distribution?

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

Let X be the systolic blood pressure.

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

(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 your answer.

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

Solution

Let X be 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.

★ 18. The gene of a rat species has been modified to help the metabolization of cholesterol in blood. To check the effectiveness of this genetic modification two samples of 20 rats were drawn, ones with the gene modified and the others not, and they were fed with the same diet with different concentrations of palm oil during one month. The following sums summarize the results:

Palm oil quantity in gr (the same in both samples)
$\sum x_i = 640.6467$, $\sum x_i^2 = 23508.6387$, $\sum (x_i - \bar{x})^3 = -5527.08$, $\sum (x_i - \bar{x})^4 = 792910$
Cholesterol level in blood in mg/dl of non genetically modified rats
$\sum y_j = 2945.8545$, $\sum y_j^2 = 439517.5975$, $\sum (y_j - \bar{y})^3 = 604.08$, $\sum (y_j - \bar{y})^4 = 3717331.07$
Cholesterol level in blood in mg/dl of genetically modified rats
$\sum y_j = 2126.5899$, $\sum y_j^2 = 226824.5373$, $\sum (y_j - \bar{y})^3 = -629.4$, $\sum (y_j - \bar{y})^4 = 48248.29$

(a) In which sample the cholesterol has a more representative mean, genetically modified or non modified rats?

(b) In which sample the distribution of cholesterol is more skew?

(c) In which sample the kurtosis of the distribution of cholesterol is less normal?

(d) Which rat has a cholesterol level relatively bigger, a genetically modified rat with a cholesterol level of 130 mg/dl, or a non genetically modified rat with a cholesterol level of 145 mg/dl?

Solution

(a) Non genetically modified rats: $\bar{y} = 147.2927$ mg/dl, $s_y^2 = 280.7332$ (mg/dl)^2, $s = 16.7551$ mg/dl and $cv_y = 0.1138$.
Genetically modified rats: $\bar{y} = 106.3295$ mg/dl, $s_y^2 = 35.265$ (mg/dl)^2, $s = 5.9384$ mg/dl and $cv_y = 0.0558$.
Thus, the mean of genetically modified rats is more representative since the coef. of variation is smaller.

(b) Non genetically modified rats: $g_1 = 0.0064$.
Genetically modified rats: $g_1 - 0.1503$.
Thus, the distribution of genetically modified rats is more skew since the coef. of skewness is further from 0.

(c) Non genetically modified rats: $g_2 = -0.6416$.
Genetically modified rats: $g_2 - 1.0602$.
Thus, the kurtosis of the distribution of genetically modified rats is less normal since the coef. of kurtosis is further from 0.

(d) Non genetically modified rats: $z(145) = -0.1368$.
Genetically modified rats: $z(130) = 3.986$.
Thus, a cholesterol level of 130 mg/dl in genetically modified rats is relatively greater than 145 mg/dl in non genetically modified rats.

2 Regression and Correlation

19. Give some examples of:

(a) Non related variables.

(b) Variables that are increasingly related.

(c) Variables that are decreasingly related.

★ 20. In a 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 each variable.

(c) Compute the covariance and interpret it.

Solution

Let X be 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.

★ 21. 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: $\sum x_i n_i = 16960$ mg/dl, $\sum y_j n_j = 11160$ mmHg, $\sum x_i^2 n_i = 3627200$ (mg/dl)^2, $\sum y_j^2 n_j = 1576800$ mmHg^2 y $\sum x_i y_j n_{ij} = 2378800$ 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.

22. 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? Which prediction is more reliable? Justify the answer.

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·%.
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.

★ 23. 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) How much 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.

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.

---

24. 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) $r = 0.5$.

---

25. 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 $y = 5$.

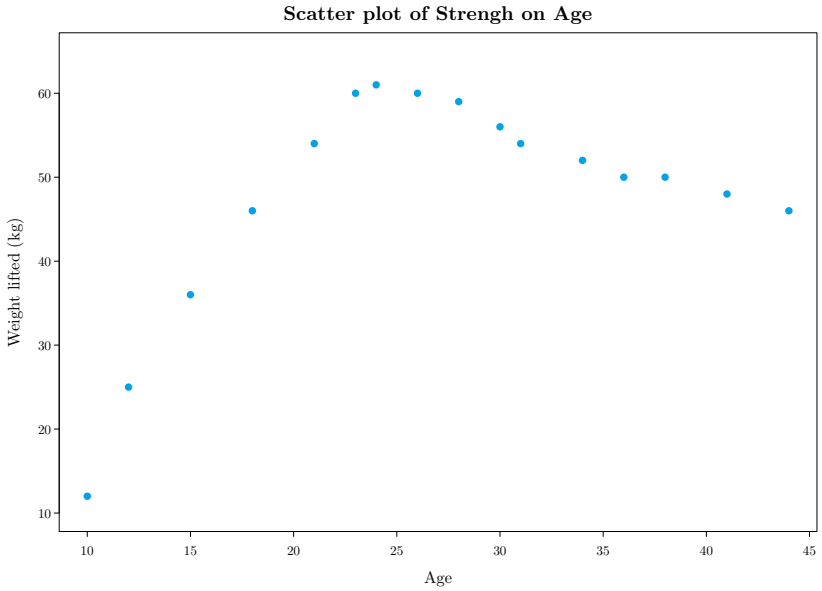
(c) Plot both regression lines.

Solution

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

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

26. 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) For which age group the relation between age and 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) $\bar{x} = 17.5714$ 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) $\bar{x} = 34.2222$ 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.

★ 27. 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.

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

Time	2	3	4	5	6	7	8
Drug concentration	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): $\sum t_i = 35$ h, $\sum \log(t_i) = 10.6046$ log(h), $\sum c_j = 541$ mg/dl, $\sum \log(c_j) = 29.147$ log(mg/dl), $\sum t_i^2 = 203$ h^2, $\sum \log(t_i)^2 = 17.5206$ log(h)^2, $\sum c_j^2 = 56937$ (mg/dl)^2, $\sum \log(c_j)^2 = 124.0131$ log(mg/dl)^2, $\sum t_i c_j = 3328$ h·mg/dl, $\sum t_i \log(c_j) = 154.3387$ h· log(mg/dl), $\sum \log(t_i) c_j = 951.6961$ log(h)·mg/dl, $\sum \log(t_i) \log(c_j) = 46.08046$ log(h) · log(mg/dl).

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.

★ 29. 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	77	30	51	75	62	45	90	20
Pain threshold	10	12	11.5	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, the linear or the logarithmic model?

(b) According to the best regression model, what is the response to pain expected for a person with an obesity of 50%? Is this prediction 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): $\sum x_i = 629$, $\sum \log(x_i) = 40.4121$, $\sum y_j = 85.5$, $\sum \log(y_j) = 20.4679$, $\sum x_i^2 = 45445$, $\sum \log(x_i)^2 = 165.6795$, $\sum y_j^2 = 854.75$, $\sum \log(y_j)^2 = 44.08906$, $\sum x_i y_j = 6124$, $\sum x_i \log(y_j) = 1390.14$, $\sum \log(x_i) y_j = 360.0725$, $\sum \log(x_i)\log(y_j) = 84.80687$.

Solution

(b) $\bar{x} = 62.9$, $s_x^2 = 588.09$, $\bar{y} = 8.55$, $s_y^2 = 12.3725$ and $s_{xy} = 74.605$.
Regression line of response to pain on obesity: $y = 0.5705 + 0.1269x$. $r^2 = 0.765$.
$\overline{\log(x)} = 4.0412$, $s^2_{\log(x)} = 0.2366$ and $s_{\log(x)y} = 1.45491$. Logarithmic model of response to pain on obesity: $y = -16.303 + 6.150\log(x)$. $r^2 = 0.7232$.
Using the linear model: $y(50) = 6.9155$.

(c) Regression line of obesity on response to pain: $x = 11.344 + 6.030y$.
$x(10) = 71.644$.

★ 30. A blood bank keeps plasma at a temperature of 0ºF. When it is required for a blood transfusion, it is heated in an oven at a constant temperature of 120ºF. 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 (ºF)	25	50	86	102	110	114	118	120

(a) Plot the scatter plot. Which type of regression model do you think 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 will the temperature of the plasma be after 15 minutes of heating? Is this prediction reliable? Justify your answer.

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

Solution

(a) A logarithmic model.

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

(c) $\bar{z} = 3.0661$ log(min), $s_z^2 = 0.6577$ log^2(min), $\bar{y} = 90.625$ ºF, $s_y^2 = 1080.2344$ ºF^2 and $s_{zy} = 26.0969$ log(min)ºF.
Logarithmic model of temperature on time: $y = -31.0325 + 39.6781 \log(x)$.

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

★ 31. A study tries to determine the effect of smoking during the pregnancy in the weight of newborns. The table below shows the daily number of cigarretes smoked by mothers (X) and the weight of the newborn (all of them are males) (Y).

Daily num cigarettes	10.00	14.00	8.00	11.00	7.00	6.00	2.00	5.00	9.00	9.00	4.00	6.00
Weight (kg)	2.55	2.44	2.68	2.65	2.71	2.85	3.45	2.93	2.67	2.59	3.02	2.72

(a) Give the equation of the regression line of the weight of newborns on the daily number of cigarettes and interpret the slope.

(b) Which regression model is better to predict the weight of newborns, the logarithmic or the exponential?

(c) Use the best of the two previous regression models to predict the weight of a newborn whose mother smokes 12 cigarettes a day. Is this prediction reliable?

Use the following sums for the computations:
$\sum x_i = 91$ cigarettes, $\sum \log(x_i) = 23.0317$ log(cigarettes), $\sum y_j = 33.26$ kg, $\sum \log(y_j) = 12.1857$ log(kg),
$\sum x_i^2 = 809$ cigarettes^2, $\sum \log(x_i)^2 = 47.196$ log(cigarettes)^2, $\sum y_j^2 = 92.9708$ kg^2, $\sum \log(y_j)^2 = 12.4665$ log(kg)^2,
$\sum x_i y_j = 243.61$ cigarettes·kg, $\sum x_i \log(y_j) = 89.3984$ cigarettes· log(kg), $\sum \log(x_i) y_j = 62.3428$ log(cigarettes)kg, $\sum \log(x_i) \log(y_j) = 22.8753$ log(cigarettes) log(kg).

Solution

(a) $\bar{x} = 7.5833$ cigarettes, $s_x^2 = 9.9097$ cigarettes^2.
$\bar{y} = 2.7717$ kg, $s_y^2 = 0.0654$ kg^2.
$s_{xy} = -0.7176$. cigarettes·kg
Regression line: $y = -0.0724x + 3.3208$.

(b) $\overline{\log(x)} = 1.9193$ log(cigarettes), $s^2_{\log(x)} = 0.2492$ log(cigarettes)^2.
$\overline{\log(y)} = 1.0155$ log(kg), $s^2_{\log(y)} = 0.0077$ log(kg)^2.
$s_{x\log(y)} = -0.2508$ cigarettes·log(kg), $s_{\log(x)y} = -0.1245$ log(cigarettes)·kg
Logarithmic coef. determination: $r^2 = 0.9499$
Exponential coef. determination: $r^2 = 0.8268$
Therefore, the logarithmic models fits better the data and is better to predict the weight.

(c) Logarithmic regression model: $y = 3.7301 + -0.4994 \log(x)$.
Prediction: $y(12) = 2.4892$ kg. The coefficient of determination is high but the sample size small, so the prediction is not enterely reliable.

★ 32. A study tries to determine the relationship between two substances X and Y in blood. The concentrations of these substances have 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 your answer.

Use the following sums: $\sum x_i = 52$ μg/dl, $\sum \log(x_i) = 13.1955$ $\log(\mu$g/dl), $\sum y_j = 11.2$ μg/dl, $\sum \log(y_j) = 3.253$ $\log(\mu$g/dl), $\sum x_i^2 = 451.36$ $(\mu$g/dl$)^2$, $\sum \log(x_i)^2 = 26.9397$ $\log(\mu$g/dl$)^2$, $\sum y_j^2 = 18.1$ $(\mu$g/dl$)^2$, $\sum \log(y_j)^2 = 1.5878$ $\log(\mu$g/dl$)^2$, $\sum x_i y_j = 86.57$ $(\mu$g/dl$)^2$, $\sum x_i \log(y_j) = 26.3463$ μg/dl $\cdot \log(\mu$g/dl), $\sum \log(x_i) y_j = 21.7087$ $\log(\mu$g/dl$) \cdot \mu$g/dl, $\sum \log(x_i) \log(y_j) = 6.5224$ $\log(\mu$g/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 record the gender and whether she or he is smoker or not.

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

(c) Pick a random person and record 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 head.

(b) Get exactly 2 tails.

(c) Get two or more heads.

(d) Get some tails.

Solution

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

(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.

    Compute again the above 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 a 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 a same sample space, such that $P(A) = 0.6$ and $P(A \cup B) = 0.9$. Compute $P(B)$ under the following assumptions:

    (a) A and B are incompatible.

    (b) A and B are independent.

    Solution

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

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

★ 41. A study about smoking has found 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) $P(SM|SF) = 0.325$.

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

    (c) They are not independent.

★ 42. 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,

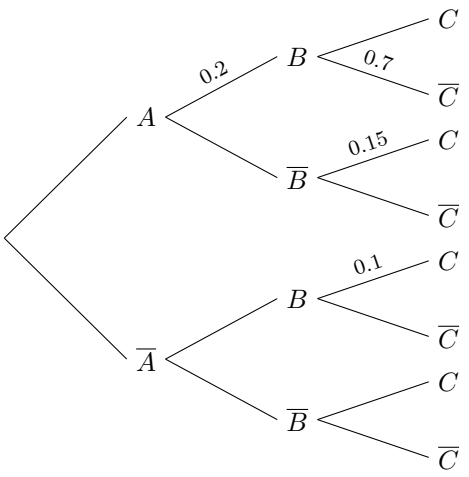
    (a) $P(B \cap \overline{A}) = 1/6$.

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

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

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

43. The tree below represents the probability space of a random experiment that consists in drawing a random person from a population and cheching if he or she suffered or not three diseases A, B and C.



Knowing that 1.5% of the population suffered the three diseases, that 54% suffered none of them and that diseases A and B are independent:

(a) Complete the tree labelling the branches with their probabilities.

(b) Compute the probability of suffering disease C.

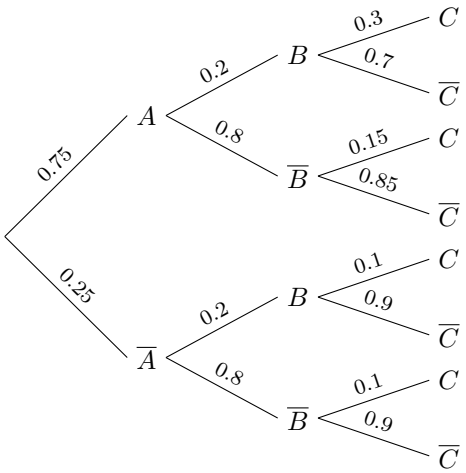
(c) Compute the probability of suffering disease B if it has been suffered disease C.

(d) Compute the probability of not suffering disease A it the other two have not been suffered.

(e) Are the diseases B and C independent?

Solution

(a)



(b) $P(C) = 0.12$.

(c) $P(B|C) = 0.25$.

(d) $P(\overline{A}|\overline{B} \cap \overline{C}) = 0.7606$.

(e) No, because $P(B|C) \neq P(B)$.

★ 44. A student has to take a test where each questions has 3 possible answers. The student knows 40% of the questions and he answers the other ones randomly. If we take a random question, what is the probability that the student does not know the answer of the question assuming 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 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 suffers from vomiting?

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

Solution

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

46. In a population with 10000 persons there was 15 persons with tuberculosis at a partircular moment. After 5 years, 5 of them died and 3 new persons got the tuberculosis. Assuming that the population size does not change in this 5 years, compute

(a) The prevalence of tuberculosis at the begining.

(b) The incidence proportion of tuberculosis in this 5 years period.

(c) The indicente rate (absolute riks) of tuberculosis in this period.

(d) The prevalence of tuberculosis at the end of this period.

Solution

(a) $P(T) = 0.0015$.

(b) $R(T) = 0.0003$, that is, 3 persons per 10000 persons in 5 years.

(c) $R(T) = 0.00006$, that is, 0.6 persons per 10000 persons-year.

(d) $P(T) = 0.0013$.

47. In an outbreak of varicella (chickenpox), varicella was diagnosed in 18 of 160 vaccinated children compared with 30 of 70 unvaccinated children.

(a) Compute and interpret the relative risk of varicella of vaccinated people.

(b) Compute and interpret the odds ratio of varicella of vaccinated people.

(c) Which association measure is more suitable for this type of study, relative risk or odds ratio?

Solution

(a) $RR(V) = 0.2625$.

(b) $OR(V) = 0.169$.

(c) The Odds ratio, because it is a restrospective study.

48. A ramdomized clinical trial tries to determine the effect of a program of preparation exercises for the birth on perineal traumas. The table below summarizes the results.

	Episotomy	No episotomy
Mothers following the program of exercises	62	358
Mothers not following the program of exercies	328	512

(a) Compute the absolute and the relative risk of episotomy following the program of exercises. What can you say about the effectiveness of the program of preparation exercises for the birth?

(b) Compute an interpret the odds ratio of episotomy following the program of exercises.

(c) Which association measure is more suitable for this type of study, relative risk or odds ratio?

Solution

(a) $R(E) = 0.1476$ and $RR(E) = 0.378$.

(b) $OR(E) = 0.2703$.

(c) Both are suitable.

★ 49. A study tries to determine the effectiveness of an occupational risk prevention program in jobs that require to be sit a lot of hours. A sample of 500 individuals between 40 and 50 years that spent more than 5 hours sitting was drawn. Half of the individuals followed the prevention program (treatment group) and the other half not (control group). After 5 years it was observed that 12 individuals suffered spinal injuries in the group following the prevention program while 32 individuals suffered spinal injuries in the other group. In the following 5 years it was observed that 21 individuals suffered spinal injuries in the group following the prevention program while 48 individuals suffered spinal injuries in the other group.

(a) Compute the cumulative incidence of spinal injuries in the total sample after 5 years and after 10 years.

(b) Compute the absolute risk of suffering spinal injuries in 10 years in the treatment and control groups.

(c) Compute the relative risk of suffering spinal injuries in 10 years in the treatment group compared to the control group. Interpret it.

(d) Compute the odds ratio of suffering spinal injuries in 10 years in the treatment group compared to the control group. Interpret it.

(e) Which statistics, the relative risk or the odds ratio, is more suitable in this study? Justify the answer.

Solution

(a) Cumulative incidence after 5 years: $R(D) = 0.088$. Cumulative incidence after 10 years: $R(D) = 0.226$.

(b) Risk in the treatment group: $R_T(D) = 0.132$. Risk in the control group: $R_C(D) = 0.32$.

(c) $RR(D) = 0.4125$. Thus, the risk of suffering spinal injuries is less than half following the prevention program.

(d) $OR(D) = 0.3232$. Thus, the odd of suffering spinal injuries is less than one third following the prevention program.

(e) Since the study is prospective and we can estimate the prevalence of D, both statistics are suitable, but relative risk is easier to interpret.

★ 50. A test was applied to a sample of people in order to evaluate its effectiveness; the results are as follows:

	Test +	Test −
Sick	2020	140
Healthy	80	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) $P((D \cap +) \cup (\overline{D} \cap -)) = 0.978$.

---

51. We know, from a research study, that 10% of people over 50 suffer a particular type or arthritis. We have developed a new method to detect the disease and after clinical trials we observe that if we apply the method to people with arthritis we get a positive result in 85% of cases, while if we apply the method to people without arthritis, we get a positive result in 4% of cases. Answer the following questions:

(a) What is the probability of getting a positive result after applying the method to a random person?

(b) If the result of applying the method to one person has been positive, what is the probability of having arthritis?

Solution

Naming A to the event of having arthritis,

(a) $P(+) = 0.121$.

(b) $P(A|+) = 0.7025$.

★ 52. A new test for detecting the Down syndrome in newborn babies has a sensitivity of 80% and a specificity of 90%. If 1% of newborn babies have Down syndrome, and after applying the test to one newborn baby the outcome of the test is positive, what is the proability that the baby has the Down syndrome? Should 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 $P(-|\overline{S}) = 0.9919$.

★ 53. 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) $P(-|\overline{D}) = 0.9783$.

★ 54. 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) Which test is better to confirm the disease?

(b) Which 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*. Which 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 we combine the 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 thest, 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.
$P(D| + A \cap +B) = 0.7878$.

★ 55. A severe pain without effusion in a particular zone of the knee joint is a symptom of sprained lateral collateral ligament (SLCL). If the sprains in that ligament are classified into grade 1, when there is only distension (60% of cases), grade 2 when there is a partial tearing (30% of cases) and grade 3 when there is a complete tearing (10% of cases). Taking into account that the symptom appears in 80% of cases of grade 1 sprains, 90% of cases of grade 2 and 100

(a) If a person has SLCL what is the probability that he or she present severe pain without effusion?

(b) What is the diagnosis for a person with severe pain without effusion?

(c) From a total of 10000 people with severe pain without effusion, how many are expected to have a grade 1 sprain? How many are expected to have a grade 2 sprain? And a grade 3 sprain?

Solution

Naming S to the event of presenting severe pain without effusion, and $G1$, $G2$ and $G3$ to the events of having a SLCL of grade 1, 2 and 3 respectively,

(a) $P(S) = 0.85$.

(b) $P(G1|S) = 0.5647$, $P(G2|S) = 0.3176$ and $P(G3|S) = 0.1176$, so the diagnosis is a SLCL of grade 1.

(c) 5647.0588 will have a grade 1 sprain, 3176.4706 will have a grade 2 sprain and 1176.4706 will have a grade 3 sprain.

★ 56. A physiotherapist uses two techniques A and B to cure an injury. It is known that the injury is 3 times more frequent in people over 30 than in people under 30. It is also known that in people over 30 technique A works in 30% of cases and technique B in 60%, while in people under 30 technique A works in 50% of cases and technique B in 70%. If both techniques are applied with the same probability, no matter the age,

(a) What is the probability that a random person under 30 is cured? And for a people over 30?

(b) What is the probability that a random person is cured?

(c) If after applying a technique to a person over 30, the person does not cure, what is the probability that the technique applied was A?

Solution

Let J be the event of being under 30, let C be the event of being cured, and let A and B be the events of applying techniques A and B respectively.

(a) $P(C|J) = 0.6$. and $P(C|\overline{J}) = 0.45$.

(b) $P(C) = 0.4875$.

(c) $P(A|\overline{J} \cap \overline{C}) = 0.636$.

4 Discrete Random Variables

57. Let X be a discrete random variable with the following probability distribution

X	4	5	6	7	8
$f(x)$	0.15	0.35	0.10	0.25	0.15

(a) Compute and represent graphically the distribution function.

(b) Compute the following probabilities

i. $P(X < 7.5)$.
ii. $P(X > 8)$.
iii. $P(4 \leq X \leq 6.5)$.
iv. $P(5 < X < 6)$.

Solution

(a)

$$F(x) = \begin{cases} 0 & \text{if } x < 4, \\ 0.15 & \text{if } 4 \leq x < 5, \\ 0.5 & \text{if } 5 \leq x < 6, \\ 0.6 & \text{if } 6 \leq x < 7, \\ 0.85 & \text{if } 7 \leq x < 8, \\ 1 & \text{if } 8 \leq x. \end{cases}$$

(b) $P(X < 7.5) = 0.85$, $P(X > 8) = 0$, $P(4 \leq x \leq 6.5) = 0.6$ and $P(5 < X < 6) = 0$.

58. Let X be a discrete random variable with the following probability distribution

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1/5 & \text{if } 1 \leq x < 4, \\ 3/4 & \text{if } 4 \leq x < 6, \\ 1 & \text{if } 6 \leq x. \end{cases}$$

(a) Compute the probability function.
(b) Compute the following probabilities
i. $P(X = 6)$.
ii. $P(X = 5)$.
iii. $P(2 < X < 5.5)$.
iv. $P(0 \leq X < 4)$.
(c) Compute the mean.
(d) Compute the standard deviation.

Solution

(a)

X	1	4	6
$f(x)$	1/5	11/20	1/4

(b) $P(X = 6) = 1/4$, $P(X = 5) = 0$, $P(2 < X < 5.5) = 11/20$ and $P(0 \leq X < 4) = 1/5$.
(c) $\mu = 3.9$.
(d) $\sigma = 1.6703$.

★ 59. An experiment consist in injecting a virus to three rats and checking if they survive or not. It is known that the probability of surviving is 0.5 for the first rat, 0.4 for the second and 0.3 for the third.

(a) Compute the probability function of the variable X that measures the number of surviving rats.
(b) Compute the distribution function.

(c) Compute $P(X \leq 1)$, $P(X \geq 2)$ and $P(X = 1.5)$.

(d) Compute the mean and the standard deviation. Is representative the mean?

Solution

(a)

X	0	1	2	3
$f(x)$	0.21	0.44	0.29	0.06

(b)

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.21 & \text{if } 0 \leq x < 1, \\ 0.65 & \text{if } 1 \leq x < 2, \\ 0.94 & \text{if } 2 \leq x < 3, \\ 1 & \text{if } 3 \leq x. \end{cases}$$

(c) $P(X \leq 1) = 0.65$, $P(X \geq 2) = 0.35$ y $P(X = 1.5) = 0$.

(d) $\mu = 1.2$ ratas, $\sigma^2 = 0.7$ ratas^2 y $\sigma = 0.84$ ratas.

---

60. The chance of being cured with certain treatment is 0.85. If we apply the treatment to 6 patients,

(a) What is the probability that half of them get cured?

(b) What is the probability that a least 4 of them get cured?

Solution

Let X be the number of cured patients in the sample of 6 treated patients, we have $X \sim B(6, 0.85)$.

(a) $P(X = 3) = 0.041$.

(b) $P(X \geq 4) = 0.9526$.

---

61. Ten persons came into contact with a person infected with tuberculosis. The probability of being infected after contacting a person with tuberculosis is 0.10.

(a) What is the probability that nobody is infected?

(b) What is the probability that at least 2 persons are infected?

(c) What is the expected number of infected persons?

Solution

Let X be the number of persons infected with tuberculosis, we have $X \sim B(10, 0.1)$.

(a) $P(X = 0) = 0.3487$.

(b) $P(X \geq 2) = 0.2639$.

(c) $\mu = 1$.

---

62. It is known that the probability having a bacteria in one mm^3 of a dissolution is 0.002. Assuming that in one mm^3 can not be more than one bacteria, compute the probability of having 5 bacteria at most in one cm^3 of the dissolution.

Solution

Let X be the number of bacteria in one cm^3 of dissolution, we have that $X \sim B(1000, 0.002) \approx P(2)$. $P(X \leq 5) = 0.9834$.

63. The probability of suffering an adverse reaction to a vaccine is 0.001. If 2000 persons are vaccinated, what is the probability of suffering some adverse reaction?

Solution

Let X be the number of adverse reactions, we have that $X \sim B(2000, 0.001) \approx P(2)$, and $P(X \geq 1) = 0.8648$.

★ 64. The average number of calls per minute received by a telephone switchboard is 120.

(a) What is the probability of receiving less than 4 calls in 2 seconds?

(b) What is the probability of receiving at least 3 calls in 3 seconds?

Solution

(a) If X is the number of calls in 2 seconds, then $X \sim P(4)$ and $P(X < 4) = 0.4335$.

(b) If Y is the number of calls in 3 seconds, then $Y \sim P(6)$ and $P(Y \geq 3) = 0.938$.

65. A test contains 10 questions with 3 possible options each. For every question you get a point if you give the right answer and lose half a point if the answer is wrong. A student knows the right answer for 3 of the 10 questions and answers the rest randomly. What is the probability of passing the exam?

Solution

Let X be the number of right questions in the 7 questions randomly answered, we have that $X \sim B(7, 1/3)$ and $P(X \geq 4) = 0.1733$.

★ 66. It has been observed experimentally that 1 of every 20 trillions of cells exposed to radiation mutates becoming carcinogenic. We know that the human body has approximately 1 trillion of cells by kilogram ot tissue. Compute the probability that a 60 kg person exposed to radiation develops cancer. If the radiation affects 3 persons weighing 60 kg, what is the probability that a least one of them develops cancer?

Solution

Let X be the number of mutations, we have that $X \sim B(60 \cdot 10^{12}, 1/20 \cdot 10^{-12}) \approx P(3)$ and $P(X > 0) = 0.9502$.
Let Y be the number of persons that develops cancer in the group of 3 persons, we have that $Y \sim B(3, 0.9502)$ and $P(Y \geq 1) = 0.9999$.

★ 67. A diagnostic test for a disease returns 1% of positive outcomes, and the positivie and negative predictive values are 0.95 and 0.98 respectively.

(a) Compute the prevalence of the disease.

(b) Compute the sensitivity and the specificity of the test.

(c) If the test is applied to 12 sick persons, what is the probability of getting at least a wrong diagnostic?

(d) If the test is applied to 12 persons, what is the probability of getting a right diagnostic for all of them?

Solution

(a) $P(D) = 0.0293$.

(b) Sensitivity $P(+|D) = 0.3242$ and specificity $P(-|\overline{D}) = 0.9995$.

(c) Let X be the number of wrong diagnostics in 12 sick individuals, we have that $x \sim B(12, 0.6758)$ and $P(X \geq 1) = 1$.

(d) Let Y be the number of right diagnostics in 12 individuals, we we have that $Y \sim B(12, 0.9797)$ and $P(Y = 12) = 0.7818$.

★ 68. In a study about a parasite that attacks the kidney of rats it is known that the average number of parasites per kidney is 3.

(a) Compute the probability that a rat has more than 3 parasites.

(b) Compute the probability of, in a sample of 10 rats, at least 9 are infected.

Solution

(a) If X is the number of parasites in a rat, then $X \sim P(6)$ and $P(X > 8) = 0.1528$.

(b) If Y is the number of infected rats in a sample of 10 rats, then $Y \sim B(10, 0.9975)$ and $P(Y \geq 9) = 0.9997$.

---

69. In a classroom there are 40 students of which 35% smoke. If we draw a random sample with replacement of 4 students, what is the probability of having at least a smoker student? Compute the same probability if the random sampling is without repalacement.

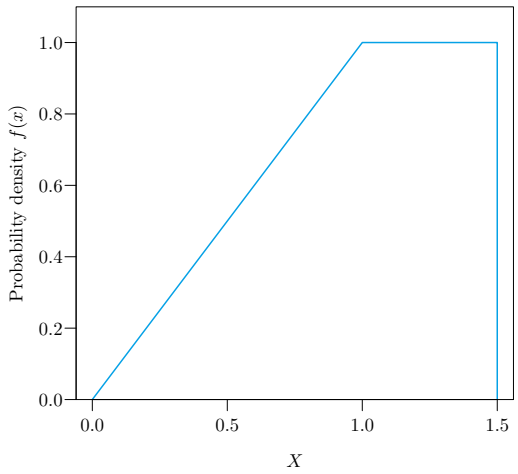
Solution

If X is the number of smoker students in a random sample with replacement of size 4, then $X \sim B(4, 0.35)$ and $P(X \geq 1) = 0.8215$.
If the random sampling is without replacement then $P(X \geq 1) = 0.8364$.

---

5 Continuous Random Variables

★ 70. Given the continuous random variable X with the probability density function chart below,



(a) Check that $f(x)$ is a probability density function.

(b) Compute the following probabilities

i. $P(X < 1)$
ii. $P(X > 0)$
iii. $P(X = 1/4)$
iv. $P(1/2 \leq X \leq 3/2)$

(c) Compute the distribution function.

Solution

(b) $P(X < 1) = 0.5$, $P(X > 0) = 1$, $P(X = 1/4) = 0$ and $P(1/2 \leq X \leq 3/2) = 0.875$.

(c)

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x^2}{2} & \text{if } 0 \leq x \leq 1, \\ x - 0.5 & \text{if } 1 < x \leq 1.5, \\ 1 & \text{if } 1.5 < x. \end{cases}$$

71. A worker can arrive to the workplace at any moment between 6 and 7 in the morning with the same likelihood.

(a) Compute and plot the probability density function of the variable that measures the arrival time.

(b) compute and plot the distribution function.

(c) Compute the probability of arriving quarter past six and half past six.

(d) What is the expected arrival time?

Solution

(a)

$$f(x) = \begin{cases} 0 & \text{if } x < 6, \\ 1 & \text{if } 6 \leq x \leq 7, \\ 0 & \text{if } 7 < x. \end{cases}$$

(b)
$$F(x) = \begin{cases} 0 & \text{if } x < 6, \\ x - 6 & \text{if } 6 \leq x \leq 7, \\ 1 & \text{if } 7 < x. \end{cases}$$

(c) $P(6.25 < X < 6.5) = 0.25$.

(d) $\mu = 6.5$.

★ 72. Let Z be a random variable following a standard normal distribution model. Calculate the following probabilities using the table of the distribution function:

(a) $P(Z < 1.24)$

(b) $P(Z > -0.68)$

(c) $P(-1.35 \leq Z \leq 0.44)$

Solution

(a) $P(Z < 1.24) = 0.8925$.

(b) $P(Z > -0.68) = 0.7517$.

(c) $P(-1.35 \leq Z \leq 0.44) = 0.5815$.

★ 73. Let Z be a random variable following a standard normal distribution model. Determine the value of x in the following cases using the table of the distribution function:

(a) $P(Z < x) = 0.6406$.

(b) $P(Z > x) = 0.0606$.

(c) $P(0 \leq Z \leq x) = 0.4783$.

(d) $P(-1.5 \leq Z \leq x) = 0.2313$.

(e) $P(-x \leq Z \leq x) = 0.5467$.

Solution

(a) $x = 0.3601$.

(b) $x = 1.5498$.

(c) $x = 2.0198$.

(d) $x = -0.5299$.

(e) $x = 0.7499$.

★ 74. Let X be a random variable following a normal distribution model $N(10, 2)$.

(a) Compute $P(X \leq 10)$.

(b) Compute $P(8 \leq X \leq 14)$.

(c) Compute the interquartile range.

(d) Compute the third decile.

Solution

(a) $P(X \leq 10) = 0.5$.

(b) $P(8 \leq X \leq 14) = 0.8186$.

(c) $RI = 2.698$.

(d) $D_3 = 8.9512$.

---

75. It is known that the glucose level in blood of diabetic persons is normally distributed with mean 106 mg/100 ml and standard deviation 8 mg/100 ml.

(a) Compute the probability that a random diabetic person has a glucose level less than 120 mg/100 ml.

(b) What percentage of persons have a glucose level between 90 and 120 mg/100 ml?

(c) Compute and interpret the first quartile of the glucose level.

Solution

(a) $P(X \leq 120) = 0.9599$.

(b) $P(90 < X < 120) = 0.9371$, that is, 93.71%.

(c) 100.64 mg/100 ml.

---

76. It is known that the cholesterol level in males 30 years old is normally distributed with mean 220 mg/dl and standard deviation 30 mg/dl. If there are 20000 males 30 years old in the population,

(a) how many of them have a cholesterol level between 210 and 240 mg/dl?

(b) If a cholesterol level greater than 250 mg/dl can provoke a thrombosis, how many of them are at risk of thrombosis?

(c) Compute the cholesterol level above which 20% of the males are?

Solution

(a) $P(210 \leq X \leq 240) = 0.3781 \rightarrow 7561.3$ males.

(b) $P(X > 250) = 0.1587 \rightarrow 3173.1$ males.

(c) $P_{80} = 245.2486$ mg/dl.

★ 77. In an exam done by 100 students, the average grade was 4.2 and only 32 students passed. Assuming that the grade is normally distributed, how many students got a grade greater than 7?

Solution

$P(X > 7) = 0.0508 \rightarrow 5.1$ students.

---

78. In a population with 40000 persons, 2276 have between 0.80 and 0.84 milligrams of bilirubin per deciliter of blood, and 11508 have more than 0.84. Assuming that the level of bilirubin in blood is normally distributed,

(a) Compute the mean and the standard deviation.

(b) How many persons have more than 1 mg of bilirubin per dl of blood?

Solution

(a) $\mu = 0.7$ mg and $\sigma = 0.25$ mg.

(b) $P(X > 1) = 0.1151 \rightarrow 4604$ persons.

★ 79. It is known that the blood pressure of people in a population with 20000 persons follows a normal distribution model with mean 13 mm Hg and interquartile range 4 mm Hg.

(a) How many persons have a blood pressure above 16 mm Hg?

(b) How much has to decrease the blood pressure of a person with 16 mm Hg in order to be below the 40% of people with lowest blood pressure?

Solution

(a) $P(X > 16) = 0.1587 \rightarrow 3174$ persons.

(b) It must decrease at least 3.75 mmHg.

★ 80. A study tries to determine the effect of a low fat diet in the lifetime of rats. The rats where divided into two groups, one with a normal diet and another with a low fat diet. It is assumed that the lifetimes of both groups are normally distributed with the same variance but different mean. If 20% of rats with normal diet lived more than 12 months, 5% less than 8 moth, and 85% of rats with low fat diet lived more than 11 moths,

(a) what is the mean and the standard deviation of the lifetime of rats following a low fat diet?

(b) If 40% of the rats were under a normal diet, and 60% of rats under a low fat diet, what is the probability that a random rat die before 9 months?

Solution

Let X be the lifetime of rats.

(a) $\mu = 12.6673$ months and $s = 1.6087$ months.

(b) $P(X < 9) = 0.068$.

★ 81. The time required to cure a basketball injury with a rehabilitation technique follows a normal distribution with quartiles $Q_1 = 22$ days and $Q_2 = 25$ days.

(a) Calculate the mean and standard deviation of the curation time.

(b) If a player has just been injured and has to play a match in 30 days, what is the probability that he will miss it?

(c) Calculate the interquartile range of the curation time distribution.

Solution

(a) Let X be the time required to cure the injury, then $X \sim N(25, 4.4478)$.

(b) $P(X > 30) = 0.1305$.

(c) $IQR = 6$ days.

★ 82. A diagnostic test to determine doping of athletes returns a positive outcome when the concentration of a substance in blood is greater than 4 μg/ml. If the distribution of the substance concentration in doped athletes follows a normal distribution model with mean 4.5 μg/ml and standard deviation 0.2 μg/ml, and in non-doped athletes is normally distributed with mean 3 μg/ml and standard deviation 0.3 μg/ml,

(a) what is the sensitivity and specificity of the test?

(b) If there is a 10% of doped athletes in a competition, what is the positive predicted value?

Solution

Naming D to the event of being doped, X to the concentration of the substance in doped athletes and Y to the concentration of the substance in non-doped athletes.

(a) Sensitivity $P(+|D) = P(X > 4) = 0.9938$ and specificity $P(-|\overline{D}) = P(Y < 4) = 0.9996$.

(b) PPV $P(D|+) = 0.9961$.

★ 83. According to the central limit theorem, for big samples ($n \geq 30$) the sample mean $\bar{x}$ follows a normal distribution model $N(\mu, \sigma/\sqrt{n})$, where μ is the population mean and σ the population standard deviation.

It is known that in a population the sural triceps elongation has a mean 60 cm and a standard deviation 15 cm. If you draw a sample of 30 individuals from this population, what is the probability of having a sample mean greater than 62 cm? If a sample is atypical if its mean is below the 5th percentile, is atypical a sample of 60 individuals with $\bar{x} = 57$?

Solution

(a) Naming $\bar{X}$ to the variable that measures the sural triceps elongation in samples of size 30, $P(\bar{X} > 62) = 0.2327$.

(b) Naming $\bar{Y}$ to the variable that measures the sural triceps elongation in samples of size 60, $P_5 = 56.8$ cm, so the sample is not atypical.

★ 84. The curing time of a knee injury in soccer players follows a normal distribution model with mean 50 days and standard deviation 10 days. If there is a final match in 65 days, what is the probability that a player that has just injured his knee will miss the final? If the semifinal match is in 40 days, and 4 players has just injured the knee, what is the probability that some of them can play the semifinal?

Solution

Let X be the curing time, $P(X > 65) = 0.0668$.
Let Y be the number of injured players that could play the semifinal, $P(Y \geq 1) = 0.4989$.

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