Chapter 1

1. Which of the below measures can be calculated from the five-number summary?

A. The mean
B. The interquartile range
C. The standard deviation
D. The variance

2. Person *X* made a lot of practice exams for statistics. Because of this, *X* understands the material well and passes the exam. The variable ‘hour spent studying’ is an example of

A. A dependent variable
B. A normally distributed variable
C. An independent variable
D. Qualitative variable

3. A teacher made a stemplot from the scores of 23 students on their statistics exam (range 0-100). In this stemplot it can be seen that de mode equals 61. Which of the below stemplots may be applicable?

A.

3	8				
4	2	8			
5	4	5	6	7	
6	1	1	1	6	
7	3	3	8	8	
8	0	2	2	5	9
9	3	5	9		

B.

3	8				
4	2	3	8		
5	4	5	5	5	
6	0	0	1	6	
7	3	3	8	8	9
8	0	2	5		
9	3	5	9		

C. Non of the above.
D. Both.

4. Which figure can best be used to check if a variable is normally distributed?

A. Q-Q plot
B. Barplot
C. Timeplot
D. Histogram

5. Given are the scores on Statistics 1 for Psychology students. The five-number summary of these scores is given below.

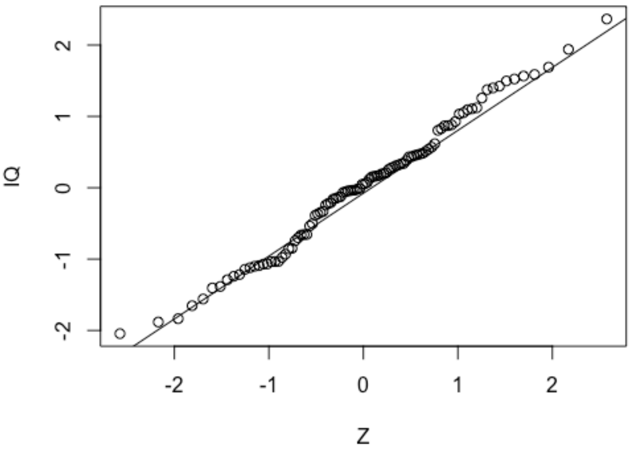
4 5 6 7 9

Which statement is true?

A. The scores above the mode are less spread than the scores below the mode.
B. The scores above the mode are more spread than the scores below the mode.
C. The scores above the median are less spread than the scores below the median.
D. The scores above the median are more spread than the scores below the median.

6. What can not be deduced from a boxplot, when the distribution of a variable is skewed?
A. The mean
B. The median
C. The interquartile range
D. The minimum

7. What kind of plot is depicted below?



A. Density plot
B. Normal Quantile plot
C. Line plot
D. Residual plot

8. The scores on 400 participants on an IQ-test provide a mean of 300 and a standard deviation of 30. The researcher wants to linearly transform the scores, so that the mean is 100 and the standard deviation 15. What should the researcher do to obtain this?
A. Divide all scores by 2
B. Divide all scores by 3
C. Divide all scores by 2 and subtract 50 from each value
D. Divide all scores by 2 and subtract 100 from each value

9. Which of the statements below is or are true?

I. The standard deviation is resistant.
II. De standard deviation is zero when no outliers are present.

A. Only statement I is true
B. Only statement II is true
C. Both statements are true
D. Both statements are false

10. The distribution of house selling prices appears to be right skewed. The mean house price is 223500 euros. Hence, the median is

A. Lower than 223500
B. Equal to 223500
C. Higher than 223500
D. The median can not be determined based on this information alone

11. Given are the test scores with mean equal to 100 and standard deviation equal to 30. A researcher wishes to transform the data in such a way that the standard deviation becomes 15, but that the mean remains equal to 100. Which transformation should the researcher use to achieve this?

A. Y = 0.50X
B. Y = 0.50X + 50
C. Y = 2X
D. That is not possible

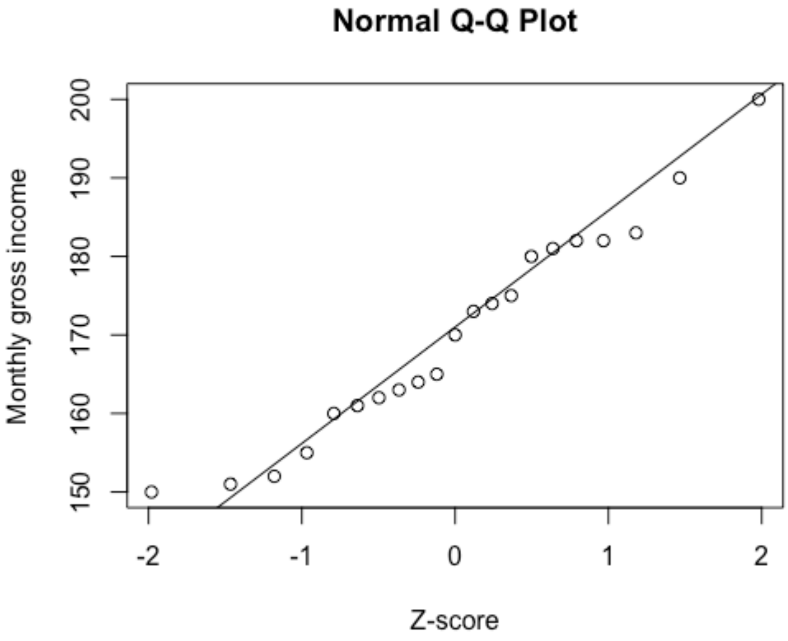
12. Given is the following five-number summary: 20, 25, 28, 35 en 55. Which of the following scores can be regarded as outlier according to the 1.5-IQR criterium?

A. 15
B. 55
C. Both 15 and 55
D. None of the above

13. A researcher wishes to describe his data with two summary measures: one center measure and one measure for spread. Which measures could he use best, if he strives to use robust measures?

A. Mean and standard deviation
B. Mean and IQR
C. Median and standard deviation
D. Median and IQR

14. A researcher collected data of 500 participants about their monthly gross income and gasoline costs per month. A Q-Q plot has been made for these data. Which of the following conclusions is true?



A. The monthly gross income correlates strongly with the monthly gasoline costs
B. The monthly gross income appears to be normally distributed
C. The monthly gross income does not correlate strongly with the monthly gasoline costs
D. The monthly gross income appears not to be perfectly normally distributed

15. A researcher collected data about the living situation of students and assigned them to four categories: independent (studio), living together with partner, living together with other students (student home), with parents. The researcher wants to display the data graphically. What figure can best be used to display the data?

A. Boxplot
B. Stemplot
C. Bar chart
D. Scatterplot

16. In an international research containing men and women from several countries, it is examined to what extent gross income can be predicted from education. What is the independent variable here?

A. Nationality
B. Sex
C. Gross income
D. Education

17. What does an IQR of 16 imply?

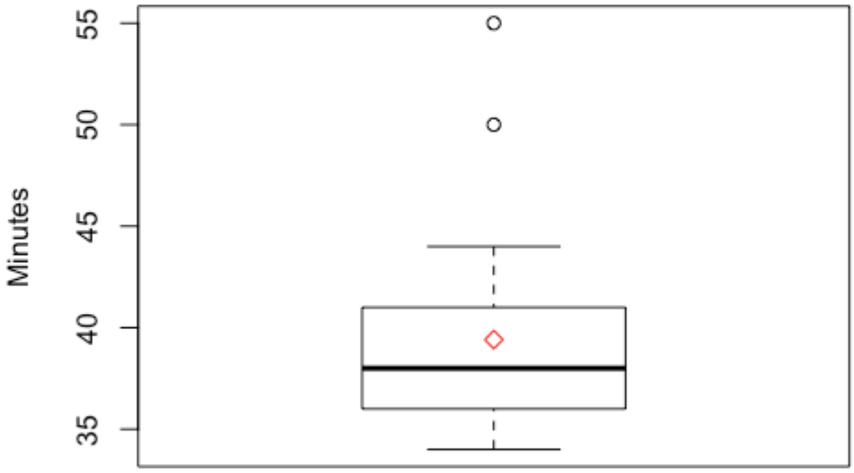
A. That the mean 50% of the scores are spread over a scale of 4 points.
B. That the mean 50% of the scores are spread over a scale of 8 points.
C. That the mean 50% of the scores are spread over a scale of 16 points.
D. That the mean 50% of the scores are spread over a scale of 32 points.

18. Data are collected for 1500 children about a writing test. It is measured how long each child needs to write a certain text. It is assumed that the variable 'time' is normally distributed in the population. From a random sample of 2500 children, 95% children score between 5 and 9 minutes. Which of the below statements is true?

I. The standard deviation in the sample is likely to be 1.
II. The mean in the sample is likely to be 7.

A. Only statement 1 is true
B. Only statement 2 is true
C. Both statements are true
D. Both statements are false

19. Due to falling leaves, travelers from NS (*Nederlandse Spoorwegen*) had to deal with much delay last weekend. Given is the delay in minutes in the past weekend for a random sample of 100 travelers. The data are displayed in a boxplot below. What does the red square display?



A. The median
B. The position of the median after removing outliers
C. The IQR
D. The mean

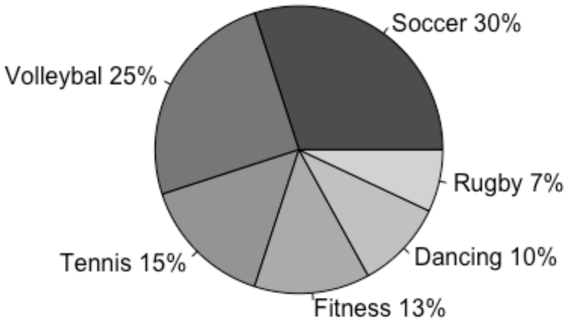
20. Given are the scores on variable *X*. A researcher wants to linearly transform the raw data by multiplying each score by 1 and then adding 20. What does change because of this transformation, and what does not change?

A. The shape of the distribution and the mean do not change, but the standard deviation becomes 20 points higher.
B. The shape of the distribution and the standard deviation do not change, but the mean becomes 20 points higher.
C. The shape of the distribution does not change, but the mean and standard deviation become 20 points higher.
D. The distribution will be more normally distributed, the mean and standard eviation become 20 points higher.

21. In a questionnaire the following item is present: 'How often did you wash your hair in the past week?'. This MC-question consists of the following response categories: 1 = not, 2 = once, 3 = twice, 4 = three times, 5 = four times or more. What is the highest meaningful measurement level of this variable?

A. Nominal
B. Ordinal
C. Interval
D. Ratio

22. For 800 students, data are collected about which sports they primarily play. Results are presented in the pie chart below. Based on this information, how many students play rugby?



A. 7
B. 56
C. 80
D. 560

23. The age of 500 participants of a Justin Bieber concert are displayed in the table below. What is the median of age?

Age	9	10	11	12	13	14	15	16	22
No. of participants	32	83	90	100	87	32	16	56	4

A. 11
B. 11,5
C. 12
D. 12,2

24. Three children of age 1, 3 and 5 are present in a room. If a 3-year old enters the room, how does this influence the mean and variance?

A. The mean remains equal, the standard deviation increases
B. The mean remains equal, the standard deviation decreases
C. The mean and standard deviation remain equal
D. The mean and standard deviation decrease

25. A teacher receives the following grades from 5 students: 4, 6, 7, 7, 8. What is the variance for these scores?

A. 0
B. 0.76
C. 1.40
D. 2.30

26. When is it better to use the five-number summary instead of the mean and standard deviation to describe the distribution of a variable?

A. Never, the mean and standard distribution are always better
B. When the distribution of the variable is fairly symmetric
C. When the distribution of the variable is strongly skewed with strong outliers
D. When the distribution of the variable is slightly skewed without outliers

27. A random variable X has a mean of 10 and a standard deviation of 2. The variable X is multiplied by 2 to create Y: Y = 2X. What is the variance of the new variable Y?

A. 2
B. 4
C. 16
D. 32

28. In a study it appears that people who drink more beer, are less often sick. In addition, it appears that people that drink more beer, also drink more orange juice. The variables “drinking beer” and “drinking orange juice” are ………… variables as explanation for being less often sick.

A. Skewed
B. Normally distributed
C. Explanatory
D. Confounding

29. A group of students thinks that drinking orange juice is good for physical recovery. To test this hypothesis, the students visit a retirement home weekly and talk with the elderly while drinking some orange juice. After a couple of weeks, the elderly are happy and healthy. What is the explanatory variable in this study?

A. Orange juice
B. The living situation (retirement home)
C. The emotional well-being of the elderly
D. All of the above answers

30. In a large-scale study in The United States, various variables have been measured. Which of the following variables is a nominal variable?

A. The state in which one lives

B. The age of the respondent
C. The number of people within a household
D. The annual income of a household per year

31. What can one use best to examine to what extent the scores on two variables are equal?

A. The correlation
B. Kendall’s tau
C. The IQR
D. The mean absolute difference

32. Kees has put the scores of 10 participants on a certain test in a stemplot. He now wants to expand the figure by adding the distinction between men and women. Which figure can Kees use best?

A. A scatterplot
B. A histogram
C. A time plot
D. A back-to-back stemplot

Answers Chapter 1

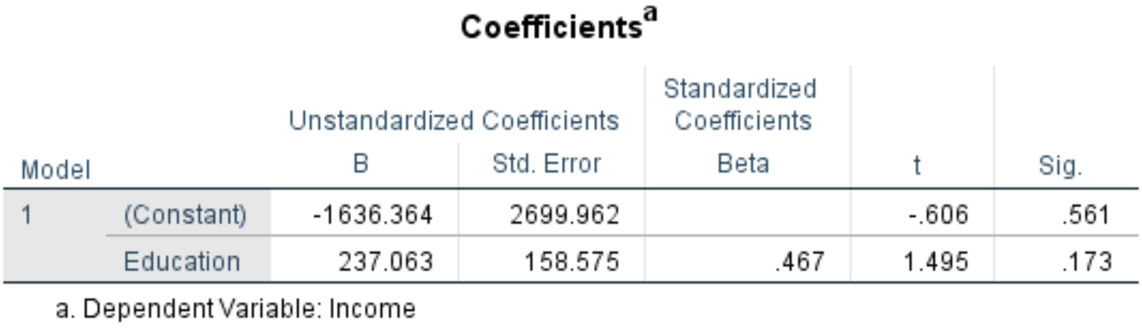
1	B	The interquartile range is the third quartile minus the first quartile, i.e.: IQR = $Q_3 – Q_1$
2	C	The variable 'hours spent studying' explains (partly) whether or not someone passes the exam and is therefore an independent variable (also called: explanatory variable). However, this does not reveal anything about the distribution of the variable. Hence, no claims can be made regarding the distribution of the variable.
3	B	For the first stemplot, the median (middle number) is 73 and de mode 61 (most frequent number). For the second stemplot, the median is 66 and the mode 55.
4	A	
5	D	The median is 6. The minimum score is 4 and the maximum score is 9. This implies that all possible values below the median vary from 4 to 6. All values above the median vary from 6-9. Hence, the spread is larger above the median. The five-number summary does not provide direct information about the mode.
6	A	A boxplot shows the median, Q1 and Q3, and outliers if present. If a variable is not (perfectly) normally distributed, the mean does not equal the median and hence the mean is not directly deducible from the boxplot.
7	B	
8	C	$x_{new} = a + bx$ Multiplying each observation with *b* (here: b = 0.5) results in a multiplication of both the center measures (e.g. mean) and spread measures (e.g. variance) with *b*. Adding the same number *a* to all observations adds *a* to the center measures, but does not change the measures of spread.
9	D	The standard deviation is influenced by outliers and hence not resistant; a few outliers can make the standard deviation very large. The standard deviation is zero, when there is no spread. That does not imply that all observations have the same value.
10	A	The mean is 'pushed' towards the side of the tail, because the mean is influenced more by extreme scores. The median is influences less by extreme scores and hence is lower than the mean.
11	B	Start with adapting the standard deviation: S_{new}= SD * |b| gives b = 0.5 Next, only adapt the mean: 100 = 0.5*100 + a gives a = 50
12	B	IQR = 35 – 25 = 10 points 1.5*IQR = 15, so outliers are below 25-15 = 10 and above 35+15 = 50.
13	D	Median and IQR are relative robust measures.
14	D	A Q-Q plot is used to display the (normal) distribution. The line is not perfectly diagonal, so only D is true.
15	C	This is a qualitative (categorical) variable. Only bar charts can be used to display categorical variables; all other figures are used for quantitative measures.

16	D	The independent variable is the variable that one uses to try to explain the dependent variable.
17	C	
18	C	When the population is normally distributed, the sample is likely to be normally distributed as well. According to the 65-95-99.7 rule, 2 standard deviations left and right from the main comprise 95% of the scores. Thus, 1 standard deviation equals approximately 1. The mean in the sample lies around 7.
19	D	A is false, because the median is the middle dash. B is wrong, because the median would be lower when removing the outliers. C is wrong, because the IQR is the middle box. D is right. We are facing a right skewed distribution, which implies that the mean is right from (i.e. higher than) the median.
20	B	
21	B	We are facing a categorical variable, so C and D are false. Because there is a rank order in the categories, ordinal is the highest measurement level.
22	B	7%, so 0.07 * 800 = 56
23	C	The median is on number (250+1)/2 = 250,5 so between 250 and 251. This is in accordance with age 12.
24	B	The extra child has exactly the mean age, so the mean does not change. Although the sum of squared deviations remains equal, dividing it by a larger number, results in a lower variance.
25	B	First, calculate the mean: $\bar{x} = \frac{4+6+7+7+8}{5} = 6.4$ Next, calculate the squared sum of deviation of each score from the mean: $(x-\bar{x})^2 = (4-6.4)^2 + (6-6.4)^2 + (7-6.4)^2 + (7-6.4)^2 + (8-6.4)^2 = (-2.4)^2 + (-0.4)^2 + (0.6)^2 + (0.6)^2 + (1.6)^2 =$ 5.76 + 0.16 + 0.36 + 0.36 + 2.56 = 9.2 Take the square root and divide by *n-1.* Thus: $var = \frac{1}{4}\sqrt{9.2} \approx 0.76$
26	C	
27	C	$\sigma_{a+bX} = b\,\sigma_x$, so: $\sigma^2_{a+bX} = b^2\sigma^2_x$ Hence: $\sigma^2 = 2^2 * 2^2 = 4 * 4 = 16$
28	D	
29	A	
30	A	
31	D	The correlation and Kendall’s tau provide information about the association between variables, this does not per se imply equal scores. The IQR provides information about the spread of scores. Again, this does not imply equal scores.
32	D	

Chapter 2

1. A regression analysis is performed in SPSS with the variables ‘education’ (in years) and income. The output is presented in the table below. What are the *a* and *b* here in the regression formula $\hat{y} = a + bx$?

A. *a* = -1636.364 and *b* = 237.063
B. *a* = 237.063 and *b* = -1636.364
C. *a* = -0.606 and *b* = 1.495
D. *a* = -1636.364 and *b* = -0.606



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1636.364	2699.962		-.606	.561
	Education	237.063	158.575	.467	1.495	.173

a. Dependent Variable: Income

2. What tries one to minimize in a scatterplot of the regression of Y on X?

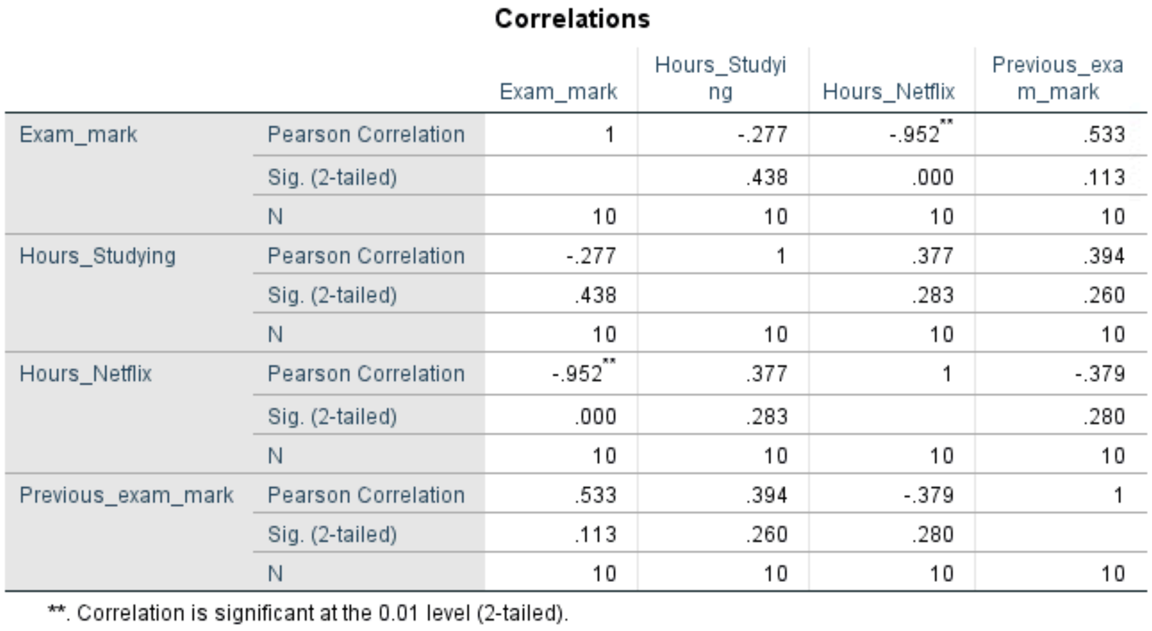
A. The sum of squares of the horizontal distances of the points till the regression line.
B. The sum of squares of the vertical distances of the points till the regression line.
C. The sum of squares of the shortest distances of the points till the regression line.
D. The sum of squares of the horizontal and vertical distances of the points till the regression line.

3. Given is that the correlation between *X* and *Y* equals 0.6. Furthermore, the mean of X equals 3, and the mean of *Y* equals 5. The standard deviation of both *X* and *Y* equals 1. What are *a* and *b* in the regression equation $\hat{y} = a + bx$?

A. *a* = 0 and *b* = 0.6
B. *a* = 0.6 and *b* = 0
C. *a* = 0.6 and *b* = 3.2
D. *a* = 3.2 and *b* = 0.6

4. The correlations between four variables are calculated and displayed in the table below. A researcher wants to make a linear regression equation to predict the exam mark on the basis of one other variable. Considering the output below, which variable is the best predictor of the exam mark?

A. Hours_studied
B. Hours_Netflix
C. Previous_exam_mark
D. That can not be determined based on correlational values only



Correlations

		Exam_mark	Hours_Studying	Hours_Netflix	Previous_exam_mark
Exam_mark	Pearson Correlation	1	-.277	-.952**	.533
	Sig. (2-tailed)		.438	.000	.113
	N	10	10	10	10
Hours_Studying	Pearson Correlation	-.277	1	.377	.394
	Sig. (2-tailed)	.438		.283	.260
	N	10	10	10	10
Hours_Netflix	Pearson Correlation	-.952**	.377	1	-.379
	Sig. (2-tailed)	.000	.283		.280
	N	10	10	10	10
Previous_exam_mark	Pearson Correlation	.533	.394	-.379	1
	Sig. (2-tailed)	.113	.260	.280	
	N	10	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

5. In a study regarding the relation between being overweight and visiting the G.P. (General Practitioner) it is shown that people who are overweight visit the G.P. more often than people with a healthy weight. This finding indicates that

A. Being overweight causes visiting the G.P.
B. People who are overweight will visit the G.P. less frequently when they lose weight.
C. There is a connection between being overweight and visiting the G.P.
D. Among people who are overweight, many people visit the G.P.

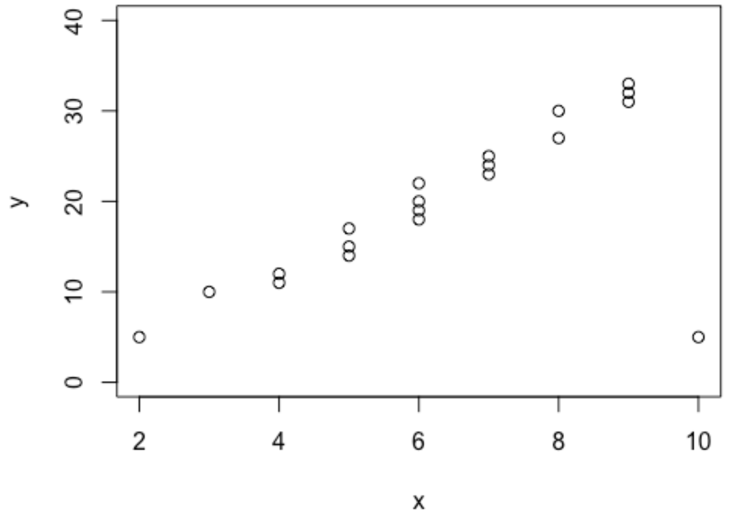
6. Given are the scores of 100 participants on variables *X* and *Y.* It is known that the variance of *X* equals 4, and that the variance of *Y* equals 9. The covariance of *X* and *Y* equals 3. What is the correlation between *X* and *Y*?

A. 0.08
B. 0.25
C. 0.50
D. 0.75

7. In a study regarding the relation between teeth and memory (Algemeen Dagblad, 2004) it is found that people who still have their own teeth, have a better memory than people with a denture. Based on this finding, the researchers conclude that 'teeth are of utmost important for our memory'. However, a critic argues that the connection that is found can be explained easily by *lurking variables* (third variables). Which of the variable(s) can play the role of third variable in this case?

A. Having a denture (fake teeth)
B. Age
C. Memory
D. All three of the above variables

8. The scores of 20 persons on variables *X* and *Y* are plotted in the figure below. Of these 20 persons, one person is quite striking. Are the scores of this person an influential point?



A. Yes, because removing this person results in a considerable change for the correlation between *X* and *Y.*
B. Yes, because the score of this person on variable *Y* is clearly an outlier.
C. No, because removing this person does not result in a change for the correlation between *X* and *Y.*
D. No, because the scores of this person on *X* and *Y* are clearly no outliers.

9. The correlation between variables *X* and *Y* appears to be exactly 1.0. What can you conclude, based on this information?

A. The mean absolute difference equals 0
B. The slope of the regression equation equals 0
C. The scores on *X* equal the scores on *Y*
D. The scores on *Y* are a linear transformation of the scores on *X*

10. Given are two variables *X* and *Y*. To predict *Y* from *X*, the following regression equation is made: $\hat{y} = -9 + 3.2X$. The correlation between *X* and *Y* is 1.0. Consider that someone scores -9 on Y. What can be said about the residual $y - \hat{y}$?

A. The residual is positive
B. The residual is negative
C. The residual will be zero
D. No statement can be made about the residual based on this information

11. The correlation between variables *X* and *Y* equals -0.40. Both *X* and *Y* have a mean of 30. The standard deviation of *X* equals 6. The standard deviation of *Y* equals 3. What is the intercept in the regression equation of *Y* on *X*?

A. 6
B. 24
C. 36
D. 54

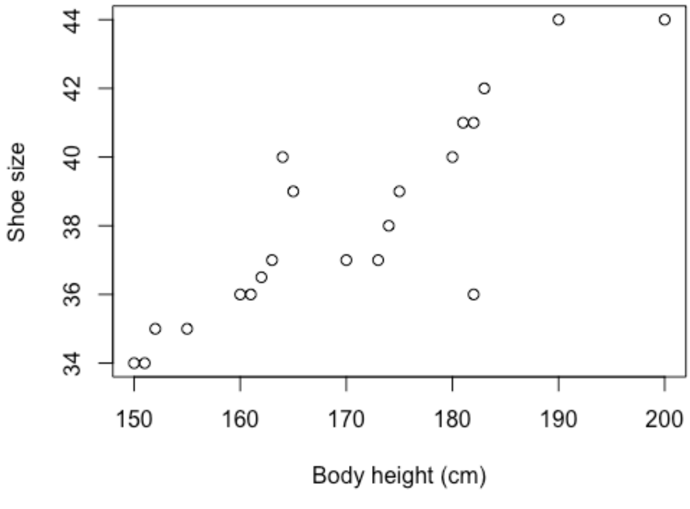
12. The correlation between variables *X* and *Y* equals 0. Below are four conclusions that are drawn based on this information. Which conclusion is false?

A. There is no linear relation between *X* and *Y*
B. The scores on *X* and *Y* are identical
C. The regression equation provides a horizontal line (slope equals zero)
D. There is 0% explained variance for a linear regression of *Y* on *X*

13. In which situation is there a Simpson's paradox present?

A. Hospital X has a lower death rate for terminal patients, whereas hospital Y has a lower death rate for non-terminal patients. When we do not consider whether the patient is terminal or not, hospital X has a lower death rate.
B. Hospital X has a lower death rate for terminal patients, whereas hospital Y has a lower death rate for non-terminal patients. When we do not consider whether the patient is terminal or not, hospital Y has a lower death rate.
C. Hospital X has a lower death rate for terminal patients, and hospital X has a lower death rate for non-terminal patients. When we do not consider whether the patient is terminal or not, hospital X has a lower death rate.
D. Hospital X has a lower death rate for terminal patients, and hospital X has a lower death rate for non-terminal patients. When we do not consider whether the patient is terminal or not, hospital Y has a lower death rate.

14. What is a reasonable estimate of the correlation between body height (in centimeters) and shoe size, according to the scatterplot that is displayed below?



A. -0.70
B. -0.10
C. 0.10
D. 0.70

15. The following linear regression equation is set up: $y = 10 + 0.8x$ in which *y* is the end score on a test, and *x* the partial score. Marleen scored 80 on her partial test. What is her predicted end score?

A. 64
B. 72
C. 74
D. 80

16. Someone examines the association between body height of women and their date partner. The table below displays the body height of six women and their data in inches (1 inch ≈ 2.5 cm).

Lengte vrouw	64	65	65	66	66	68
Lengte date	68	69	69	70	72	73

Which of the following statements is true?

A. Each body height above 66 inches should be considered an outlier
B. There is a strong positive association between the body height of the women and the body height of their date
C. There is a strong negative association between the body height of the women and the body height of their date
D. If the body height of the women and their data would have been expressed in centimetres, the correlation would be 2.5 times larger

17. In a study about the association between gender and income, the correlation between these two variables appears to be $r = -0.61$. Which statement is true?

A. Women earn more than men
B. Men earn more than women
C. A mistake has been made; the correlation should be positive
D. The measurement is pointless; *r* can only be determined for two quantitative variables

18. Many high-school students in The United States make the SAT-test and/or the ACT-test as admission for further education. Data are collected for 60 students who made both tests.

- The SAT had an average of 888 with a standard deviation of 180
- The CAT had an average of 25 with a standard deviation of 5
- The correlation between the SAT and CAT is 0.851

A researcher wants to predict the ACT from the SAT-test results by using a linear regression equation. What is the least sum of squares regression line $y = a + bx$ for these data?

A. $y = 122.10 + 30.636x$
B. $y = 30.636 + 122.10x$
C. $y = 0.024 + 3.725x$
D. $y = 3.725 + 0.024x$

19. A least squares regression line is estimated for a variable. One of the data-points has a positive residual. Which statement is true?

A. The correlation between all predicted and observed data points is positive
B. This data-point lies above the regression line
C. This data-point has to be an influential point
D. This data-point lies at the right side of the scatterplot

Answers Chapter 2

1	A	*a* is the intercept, *b* is the slope.
2	B	
3	D	$b = r_{xy}\frac{s_y}{s_x} = 0.6\ \frac{1}{1}$, thus *b* = 0.6 $a = \bar{y} - b * \bar{x} = 5 - 0.6 * 3 = 3.2$, thus *a* = 3.2
4	B	$r^2 = (-0.952)^2$ = 0.906. Thus, the hours spent watching Netflix explain about 90% of the variance of the exam mark.
5	C	A en B implicate a causal relation. D is wrong, because it may be that among people who are overweight, only a small proportion visits the G.P. but that this is more than among people with a healthy weight. Thus, it tells you something about the relative number of G.P. visits, not about the absolute number.
6	C	$r_{xy} = \frac{cov(x,y)}{s_x s_y} = \frac{3}{\sqrt{4}*\sqrt{9}} = \frac{3}{2*3} = \frac{3}{6} = 0.5$
7	B	A lurking variable is a variable –other than an exploratory or response variable- that influences the relation between the studied variables.
8	A	
9	D	Correlation tells you to what extent all points lie on one line: a correlation of 1 means that all points lie perfectly on one line. This, however, does not per se imply that all scores are equal, or that the slope is 1. When the scores are not equal, the mean difference does not have to be zero.
10	C	A correlation of 1 implies that all points lie perfectly on one line (see also question 9). This implies that all residuals are zero.
11	C	$b = \frac{S_y}{S_x} * r_{xy} = \frac{3}{6} * -.40 = -0.20$ $a = \bar{y} - b * \bar{x} = 30 - 0.20 * 30 = 30 - -6 = 30 + 6 = 36$
12	B	A correlation of zero implies that there is no linear association between the variables, so A and C are true. The proportion explained variance is r^2 and hence is also zero. B is false.
13	D	Discussed in class. See also page 143-145 of the book for a detailed explanation and different example. Moral: a causal relationship that seems to be present, switched when you add a third (lurking) variable.
14	D	The regression line is positive; so there is a positive correlation. Moreover, there is a reasonable association between body height and shoe size. Answer D is the best approximation.
15	C	y = 10 + 0.8*80 = 74
16	B	
17	D	
18	A	$b = r * \frac{s_{SAT}}{s_{CAT}} = 0.851 * \frac{180}{5} = 30.636$ $a = SAT - b * ACT = 888 - 30.636 * 25 = 122.1$
19	B	

Chapter 3

1. What is an example of a *matched-pairs design* with two conditions?

A. Each participant is matched to a similar participant. These two participants are allocated randomly to a condition and compared.
B. Each participant is allocated to both conditions. The order of the allocation is randomly selected per participant.
C. None of the above
D. Both

2. A random sample is a sample in which

A. The participants are drawn randomly from the population
B. The conditions are allocated randomly to participants
C. The conditions are selected randomly
D. The conditions are allocated in a random order to participants

3. Which of the following statements about experimental research is true?

I. The independent variable is manipulated by the researcher.
II. It is possible to examine a causal relationship with an experimental design

A. Only statement I is true
B. Only statement II is true
C. Both statements are true
D. Both statements are false

4. A research examines the association between income and education. When collecting the data, the researcher wishes to take into account the 50-50 distribution for male/female that is present in the population as well as the 30-60-10 distribution for social economical status (SES). Therefore, the researcher divides the population according to sex and SES, and draws a random sample with the numbers of each group equal to the proportions that are present in the population. What kind of sample does the researcher use?

A. Convenient sample
B. Stratified sample
C. Multistage sample
D. Paired sample

5. Anneloes has a cold. Her room mate uses a garlic tablet every day and has not had a cold for over a year now. The aunt of Anneloes knows someone who also uses garlic tablets daily and has not had a cold for a year. Based on this, Anneloes decides to use garlic tablets as soon as she is recovered from her cold. On which kind of study is her decision based?

A. Anecdotic evidence
B. An observational study based on available evidence
C. An observational study based on a sample
D. An experiment

6. The association between drinking Pepsi and weight gain is examined. The study divided 25 participants into two groups: one group followed a Pepsi-free diet and one group followed a Pepsi-rich diet. After 8 weeks the weight gain of each participant is determined. This study is an example of a(n)

A. Observational study
B. Survey
C. Matched-pairs experiment
D. Experiment, which is not double-blind

7. Geertje wants to examine price differences of coffee milk between Albert Heijn, Jumbo and De Spar. How can Geertje best select the products to prevent bias as good as possible?

A. Buy the most bought coffee milk
B. Buy coffee milk of the famous brands
C. Buy both the most bought and famous brands
D. Randomly select a number of available products

8. In a study regarding Ritalin, 100 participants are first divided according to gender. Next, half of the male participants (randomly selected) is assigned the Ritalin, and the other half is assigned a placebo. Equally, half of the female participants (randomly selected) is assigned the Ritalin, and the other half is assigned a placebo. This is an example of

A. Replication
B. Matched-pairs design
C. Entanglement, because the effect of gender is entangled with the effect of Ritalin
D. Block-design

Answers Chapter 3

1	D	A matched pairs design can be related to both the allocation (order) of participants to both conditions, and the allocation of *matched* participants to different conditions.
2	A	
3	C	
4	B	The population is subdivided in 'strata'. Next, a sample is drawn from each stratum. Due to this, the population proportions remain.
5	A	
6	D	
7	D	
8	D	

Chapter 4

1. Given are the scores on variable *X* with a mean of 10 and a standard deviation of 2. Based on this information, we can calculate *Y* as *Y* = 10 – 2*X*. The standard deviation of Y equals

A. 2
B. 4
C. 16
D. 32

2. Given are two events *A* and *B*. It is known that *P(B)* = 0.6, *P(A* and *B)* = 0.3 and *P(A* or *B)* = 1.0. What is the chance that A occurs, i.e. *P(A)*?

A. 0.1
B. 0.3
C. 0.6
D. 0.7

3. Given are two events *A* and *B*. It is known that *P*(A) = 0.3 and *P(B)* = 0.5 and *P(B|A)* = 0.8. What is the chance of *P(A* and *B*)?

A. 0.15
B. 0.24
C. 0.40
D. 0.48

4. A fair dice is thrown twice. What is the chance that the sum of these two throws equals 12?

A. 1/36
B. 2/36
C. 4/36
D. 1/12

5. Data are collected for a group of elderly about their living situation and loneliness, as can be seen in the table below.

	Being lonely		Total
	Yes	No	
Living in a retirement home	40	30	70
Living independently	10	20	30
Total	50	50	100

What is the chance that an elderly person, of which it is known that he or she lives in a retirement home, is lonely?

A. 40/70
B. 40/100
C. 50/100
D. 70/100

6. People who are psychotic, are often depressed too. To examine this relationship, we collected information of 100 patients. In this sample 30% of the patients are psychotic. Of the psychotic patient, 80% is depressed. Of the patients that are not psychotic, only 20% is depressed. How many patients from this sample are psychotic and depressed?

A. 20
B. 24
C. 30
D. 80

7. When event A and B are independent, then:

A. P(A|B) = 0
B. P(A and B) = 0
C. Both A and B
D. None of the above

8. It is given that 25% of the people has a vitamin deficiency. Moreover, of the people with a vitamin deficiency, 80% is truly tested positively on a certain test. Of the people with no vitamin deficiency, 10% somehow still has a positive test result. What is the chance that someone with a positive test result, actually has a vitamin deficiency?

A. 20%
B. 73%
C. 80%
D. 90%

9. Given is the below chance distribution of variable *X*. The mean of *X* equals 2.5. What is the expected standard deviation of *X*?

X	1	2	3	4
P	.30	.20	.20	.30

A. 1.20
B. 1.45
C. 1.80
D. 2.00

10. The next information is provided: P(A) = 0.40 and P(B) = 0.30. Moreover, is it given that A and B are independent events. What is the chance on A, given B?

A. 0.12
B. 0.30
C. 0.40
D. More information is needed to determine this

11. With a certain drinking game, you have to drink you throw a 1. Someone joins this game for three rounds, and throws with the same, equal dice. What is the chance that a person has to drink exactly one?

A. $1 * (\frac{1}{6}^{1} * \frac{5}{6}^{2})$

B. $3 * (\frac{1^1}{6} * \frac{5^2}{6})$
C. $\binom{3}{2}$
D. More information is needed to determine this

12. Imagine two independent events A and B with P(A) = 0.5 and P(B) = 0.2. What is the chance that both A and B do not happen?
A. 0.1
B. 0.3
C. 0.4
D. 0.7

13. Consider that you throw a fair dice twice. What is the chance that you throw the same number both times?
A. 1/6
B. 1/12
C. 1/18
D. 1/36

14. Below you find the chance distribution of X. X is the number of courses attended by full-time students in the last period.

X	1	2	3	4
P	.20	.30	.20	.30

What is the average number of attended courses by full-time students in the last period?
A. 0.65
B. 2
C. 2.6
D. 3

15. And what is the standard deviation of X, as presented in question 14?
A. 0.32
B. 0.64
C. 1.04
D. 1.10

16. Hans is often hired to fix computer problems, such as debugging viruses. Recently, two viruses are present: Dummy and Smarty. The following information is provided:

- 65% of the customers has problems with virus Dummy, and 35% of the customers has problems with virus Smarty
- If the computer is infected with Smarty, there is an 80% chance that Hans can fix the problems
- If the computer is infected with Dummy, there is an 30% chance that Hans can fix the problems

If we randomly select a computer, of which we know that Hans fixed the problems, what is then the chance that this computer was infected with Dummy?

A. 0.52
B. 0.53
C. 0.63
D. 0.83

17. Given are two disjoint events A and B. The chance on A is 0.2 The chance on B is 0.8. What is P(A or B)?

A. 0.6
B. 0.8
C. 1.0
D. More information is needed

Answers Chapter 4

1	C	$var(Y) = (-2)^2 * var(X) = (-2)^2 * (2)^2 = 4 * 4 = 16$ $sd(Y) = \sqrt{var(Y)} = \sqrt{16} = 4$
2	D	$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$. Thus: $1.0 = x + 0.6 - 0.3$. $x = 1.0 - 0.6 + 0.3 = 0.7$
3	B	$P(A\ and\ B) = P(B\|A) * P(A) = 0.24$
4	A	The sum of two times equals 12 only when both times 6 is thrown. Hence, $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$
5	A	The question implies the conditional chance (= given living situation in retirement home).
6	B	30% is psychotic, that is 30/100 * 100 = 30 patients. Of these 30 patients, 80% is depressed. That is: 80/100 * 30 = 24 patients. Tip: draw a tree diagram.
7	D	Independent means that event A does not influence the chance on event B, and vice versa. So P(A|B) = P(A) and P(B|A) = P(B).
8	B	Make a tree diagram. Consider *n* = 1000 participants. Then, approximately *n* = 275 test positively and *n* = 200 actually have a vitamin deficiency (250*0.8 = 200). Thus, 200/275 * 100 = 73%
9	A	$Var = 0.30 * (1 - 2.5)^2 + 0.20 * (2 - 2.5)^2 + 0.20 * (3 - 2.5)^2 + 0.30 * (4 - 2.5)^2$ = 1.45 $SD = \sqrt{1.45} \approx 1.20$
10	C	A and B are independent, so B does not predict anything about A.
11	B	The chance to drink exactly one means that you have to drink the first, second or third round. This implies three above 1 (=3) times the chance on each of the drinking possibilities (i.e. 1/6 * 5/6)
12	C	P(A not and B not) = P(A not)* P(B not) = (1 – 0.5) * (1 – 0.2*) = 0.5 * 0.8 = 0.4
13	A	The chance to throw any number is 1/6. The chance to throw that number twice is 1/6 * 1/6 = 1/36 This can happen for all 6 numbers, so multiply by 6 (= 6/36 thus 1/6)
14	C	$\mu = 1 * 0.2 + 2 * 0.3 + 3 * 0.2 + 4 * 0.3 = 2.65$

15	A	Mean = 2.6 (see question 14).
		Variance = (0.20 * 1-2.6)^2 + (0.30 * 2-2.6)^2 + (0.20 * 3-2.6)^2 + (0.30 * 4-2.6)^2 = (-0.32)^2 + (-0.18)^2 + (0.08)^2 + (0.42)^2 = 0.3176 ≈ 0.32

16	D	Make a tree diagram 100*0.65 = 65 > 65*0.8 = 52 (computers with Dummy, fixed by Hans) 100*0.35 = 35 > 35*0.3 = 10.5 (computers with Smarty, fixed by Hans) So, 52 / (52+10.5) = 0.8333 ≈ 0.84
17	C	Disjoint implies P(A or B) = 1.0

Chapter 5

1. The scores on the Cito-test are approximately normally distributed with a mean of 535 and a standard deviation of 5. What percentage of students scored higher than 545?

A. 1%
B. 2.5%
C. 5%
D. 10%

2. Given are the the scores on the normally distributed variable 'Time needed to fall asleep' for 100 children with a mean of 1500 seconds and a standard deviation of 300 seconds. What is the proportion of children that needs more than 1000 seconds to fall asleep?

A. 0.0475
B. 0.1423
C. 0.8577
D. 0.9525

3. Which of the below statements about sampling variability is/are true?

I. The sampling variability can be lowered by increasing the sample size.
II. The sampling variability is the degree of distribution of a statistic when the statistic is calculated for many randomly drawn samples from the same population.

A. Only statement I is true
B. Only statement II is true
C. Both statements are true
D. Both statements are false

4. The scores on a developmental test for toddlers are normally distributed with mean 100 and standard deviation 10. What is the chance that a random toddler scores 115 or higher?

A. .0068
B. .4404
C. .5596
D. .9332

Use the following information for questions 5 and 6. The population Dutch Psychology students has a skewed distribution for sex: only 20% is male, and 80% is female. We are interested in the population of male Dutch Psychology students (so $p = 0.20$).

5. What is the chance on less than 2 male students in a random sample of 8?

A. .1678 + .3355
B. .1678 + .3355 + .2936
C. 1 – (.1678 + .3355)
D. More information is needed

6. What is the chance on at least 30 male students in a random sample of 120 students? Use a normal approximation of the binomial distribution.

A. $P(Z > 1.15)$

B. $P(Z > 1.26)$
C. $P(Z > 1.37)$
D. $P(Z > 1.48)$

7. Given are the scores on the Cito-test. The scores are normally distributed in the population with mean 100. In a random sample of n = 25, the mean equals 25. The standard deviation in the sample is 3. Which of the statements below is true?
A. 100 is a parameter, 25 is a statistic
B. 100 is a parameter, 105 is a statistic
C. 25 is a parameter, 3 is a statistic
D. 25 is a parameter, 105 is a statistic

8. An unbiased statistic implies that for a large number of similar, representative samples from the same population and with the same sample size *n* …
A. All statistics are closely together
B. The mean of the statistics equals the mean of the parameter
C. The variance of the statistics is zero
D. The mean of the statistics is zero

9. What is P(-0.55 < Z < 1.21) if we use Table A for standard normal distributions?
A. 0.2912
B. 0.5957
C. 0.7088
D. 0.8869

10. The scores of students on the American College Test (ACT) are normally distributed in the population with mean 18 and standard deviation 6. 50 students from a certain school make the ACT. Assume that these 50 scores follow the same distribution as in the population. What is the sampling distribution of the mean on the ACT for samples of n = 50?
A. About normal, but the approximation is bad
B. Exactly normal
C. Skewed to the right
D. Skewed to the left

11. Birth weight of babies is normally distributed with a mean of 7 pound and a standard deviation of 0.8 pound. What is the chance that a randomly selected baby weights more than 7.6 pound?
A. 0.23
B. 0.75
C. 0.77
D. More information is needed

12. X has a binomial distribution with parameters n = 10 and p = 0.7. What is the average number of successes, and what is the standard deviation?
A. μ = 1.45, σ = 7
B. μ = 1.45, σ = 2.1
C. μ = 7, σ = 2.1

D. μ = 7, σ = 1.45

13. Given is that 30% of the marriages in The Netherlands results in a divorce within 15 years. A large study examined hundreds of marriages for the past 15 years. Imagine that 100 of these marriages are selected at random, what is then the chance that less than 20 of these marriages result in a divorce?

A. .011
B. .110
C. .890
D. .989

14. Given is that variable X is heavily skewed to the left in the population. What does the sampling distribution of X look like for samples of *n* = 100 from this population?

A. Heavily skewed to the left, in accordance with the population
B. More normally distributed than in the population
C. Exactly normally distributed
D. More information is needed

15. An assumption of the binomial distribution is that all observations are

A. Independent
B. Random
C. Dependent
D. Positive

16. A singular random sample is drawn from a large population. The percentage of respondents in the sample with a certain characteristic is determined. What is the best description of this percentage?

A. It is a parameter
B. It is a statistic
C. It is a lurking variable
D. None of the above answers is correct

Answers Chapter 5

1	B	545 – 535 = 10. That means the score 545 is 2 standard deviations higher than the mean. Two standard deviations left and right of the mean summarizes 95% of all observations. Of the remaining 5%, 2.5% is left (< 525) and 2.5% is right (> 545). Draw a normal distribution with vertical lines for the mean and critical values to provide more insight into the question.
2	D	$Z > \frac{x-\mu}{\sigma} = \frac{1000-1500}{300} = \frac{-500}{300} = -1.67$ Z = -1.67. Looking at Table A, we find for this z-value *p* = .0475. This is the left exceedance probability. Beause the question is how many children need *more* than 1000 seconds, we need 1 - .0475 = .9525
3	C	
4	A	$Z > \frac{x-\mu}{\sigma} = \frac{115-100}{10} = \frac{15}{10} = 1.5$ Look up Z = 1.5 in Table A. This provides you with a left exceedance probability of *p* = .9932. We want to know the right exceedance probability, i.e. 1 – 0.9932 = 0.0068
5	A	Table C: P(*X* < 2 | p = 0.20, *n* = 8) = P(*X* = 0 | p = 0.20, *n* = 8) + P(*X* = 1 | p = 0.20, *n* = 8) .
6	B	First, calculate the mean and standard deviation $\bar{x} = 120 * 0.20 = 24$ $SD = \sqrt{n * p * (1-p)} = \sqrt{120 * 0.20 * 0.80} \approx 4.38$ Then, use the continuity correction for a normal approximation of the binomial distribution. That means here that you have to use 29.5 instead of 30. $P(X \geq 30 \mid p = 0.20, n = 120) = P\left(Z > \frac{29.5-24}{4.38}\right) = P(Z > 1.26)$
7	B	Population > parameter and Sample > statistic PP - SS
8	B	Unbiased means that there is no structural distortion. While a single sample may deviate from the population (parameter), the statistic is on average equal to the parameter.
9	B	P(-0.55 < Z < 1.21) = P(Z < 1.21) – P(Z < -0.55) = 0.8869 – 0.2912 = 0.5957
10	B	
11	B	$Z > \frac{x-\mu}{\sigma} = \frac{7.6-7}{0.8} = \frac{0.6}{0.8} = 0.75$ gives .7734 We want to know the right exceedance probability, so P = 1 - .7734 = 0.2266
12	D	$\mu = np = 10 * 0.7 = 7$ $\sigma = \sqrt{(np(1-p)} = \sqrt{2.1} \approx 1.45$
13	A	Use the normal approximation of the binomial distribution $\mu = np = 0.30 * 100 = 30$ $\sigma = \sqrt{(np(1-p)} = \sqrt{30(0.70)} = \sqrt{21} \approx 4{,}58$ $P\ Z < \frac{19.5-30}{4.58} \approx -2.29$, looking up in Table A provides P < .0110
14	D	Central limit theorem (Chapter 5, page 300 in Moore, McCabe and Craig).
15	A	
16	B	

Chapter 6

1. Given are the years of education for a random sample of 100 participants from the population of Dutch man. Next, a 95% confidence interval is made for the first quartile. This 95% confidence interval consists of

A. The lowest 25% of the scores on 'years of education' in the sample
B. The lowest 25% of the scores on 'years of education' in the population
C. With 95% confidence the value of the first quartile in the sample
D. With 95% confidence the value of the first quartile in the population

2. The mean on a variable *X* has been calculated for 100 students from the population of students in Groningen. A 95% confidence interval is made for the mean. In this case, the 95% confidence interval is the interval is which we find

A. 95% of the means from the sample
B. 95% of the means from the population
C. With 95% certainty the sample mean of *X*
D. With 95% certainty the population mean of *X*

3. Dutch employees work on average 30 hours a week. Assume a normal distribution and a standard deviation of 3 in the population. What percentage of Dutch employees works between 24 and 36 hours a week?

A. 5%
B. 32%
C. 68%
D. 95%

4. Rimmer examines the satisfaction of Psychology students with their exam grade for statistics. He uses a 0-100 range and assumes that the scores are normally distributed. Rimmer makes a 95% confidence interval for the mean from a random sample. The confidence interval is [60-75]. What does this imply?

A. That 95% of the scores in the sample lie between 57 and 63
B. That 95% of the scores in the population lie between 57 and 63
C. That there is a 95% chance that this interval contains the parameter
D. That there is a 95% chance that this interval contains the statistic

5. One hundred students are asked how much beers they have drunken in the past week. The scores are skewed to the left with mean 5 and standard deviation 3. How many beers does a student have to drink to be in the top 2.5%?

A. At least 8
B. At least 11
C. At least 14
D. More information is needed

6. The scores on an exam are normally distributed with mean 60 and standard deviation 8. What is the score that one has to get in order to be in the lowest 5% of the scores?

A. Approximately 44 or lower
B. Approximately 44 or higher

C. Approximately 47 or lower
D. Approximately 47 or higher

7. The time to finish an exam is normally distributed with mean 50 and standard deviation 10. What is approximately the percentage of students that finishes the exam within an hour?

A. 68%
B. 84%
C. 95%
D. 99.7%

8. In a study it is found that Dutch citizens spend on average 1200 euros per year on clothing, with a standard deviation of 14.83. Given is that the margin of error equals 30. What is the minimum required sample size to obtain a 95% confidence interval?

A. 5
B. 6
C. 33
D. 34

Answers Chapter 6

1	D	A confidence interval is used to say something about the population with a certain degree of (un)certainty. The sample is only a means to an end.
2	D	See question/answer 1.
3	D	Take 2 SD's left and right from the mean (according to 68-95-99.7 rule-of-thumb).
4	C	
5	D	It is a *left skewed* distribution. Therefore, one cannot make statements according to the 68-95-99.7 rule-of-thumb for normal distributions.
6	C	Be aware that we can not use the rule-of-thumb as we we used in the previous question. We now need to look up the z-score in table A. The chance of .05 lies between z = -1.64 and z = -1.65, so we use z = -1.645. This results in: $60 - 1.645 * 8 = 46.84$ which is 47 after rounding.
7	B	Within an hour implies +1 SD to the right. +/- 1 SD comprises 68% Add half of the remaining 32%, so 68 + 16 = 84%
8	D	$n = (\frac{z*\sigma}{m})^2 = (\frac{1.96*14.83}{5})^2 = (\frac{29.07}{5})^2 = 5.813^2 = 33.80$ so at least 34

Chapter 7

1. Given are two independent variables *X* and *Y.* It is known that the mean of *X* equals 20 and the standard deviation equals 10. Variable *Y* has a mean of 10 and a standard deviation of 5. What is the standard deviation of the variable (*X* – *Y*)?

A. 5
B. 15
C. 75
D. 125

2. Given are two independent random variables X and Y. Which of the following statements is not true?

A. The variance of the difference X – Y equals the difference between the variances
B. The variance of the sum X + Y equals the sum of the variances
C. The mean of the sum X + Y equals the sum of the means
D. The mean of the difference X – Y equals the difference of the means

Answers Chapter 7

1	D	The variance of the difference variable equals the sum of the variances of both variables: $var(X - Y) = var(X) + var(Y) = 10^2 + 5^2 = 125.$
2	A	

Additional material (connected to Chapter 2)

1. Two assessors estimated a large number of stimuli on a certain characteristic.
Their judgements are given as scores on X and Y. To calculate Kendall's tau between *X* and *Y*, we need concordant and discordant pairs. What is a concordant pair?

A. A pair of stimuli that is mutually equal.
B. A pair of assessors that judges the stimuli equally on a certain characteristic.
C. A pair of variables for which the correlation between the scores is positive.
D. A pair of stimuli for which the rank order of the scores on X equals the rank order of the scores on Y.

2. Spearman's rho indicates to what extent

A. The scores on two variables are equal
B. There is a linear relation between the scores on two variables
C. The absolute values of the scores on two variables are equal
D. There is agreement in rank order between scores on two variables

3. What do you know when Kendall's tau between X and Y is exactly 1.0?

A. That each participant scored exactly the same on X as on Y
B. That each participant scored exactly 1 point higher on X than on Y
C. That the Spearman's correlation between X and Y equals 1.0
D. That the order of participants on X is exactly the same as the order of the participants on Y

4. Can Cohen's kappa be a negative number?

A. No
B. Yes, but only when the observed proportion of agreement is negative
C. Yes, but only when the observed proportion of agreement is lower than 0.50
D. Yes, but only when the observed proportion of agreement is lower than the expected proportion of agreement

5. The height (*H*) and length (*L*) of six boxes are determined. It appears that of the 15 mutual comparisons, the longest box is the highest as well in 12 cases. In 3 cases, the longest box is the lowest. What is Kendall's tau between *H* and *L*?

A. 0.20
B. 0.25
C. 0.60
D. 0.80

Answers additional material (connected to Chapter 2)

1	D	
2	D	
3	D	Kendall's tau only provides information about the rank order. This does not imply that the scores on different variables are equal, or that they lie on the same line. It also does not imply anything about a difference between the scores.
4	D	The kappa has a range of [-1; 1] and is negative when when the observed proportion of agreement is lower than the expected proportion of agreement.
5	C	Kendall's tau = $\frac{\#C-\#D}{n\ pairs}=\frac{12-3}{15}=0.60$