STATISTICS AND DATA ANALYSIS, EXAM–PLE

Department of statistics
Edgar Bueno
2024–11–15

Approved aid: Hand-held calculator with no stored text, data or formulas
Provided aid: Formula Sheet and Probability Distribution Tables, returned after the exam.

Problems 1 — 12: Multiple choice questions (max 60 points):

- A total of 12 multiple choice questions with five alternative answers per question one of which is the correct answer. Mark your answers on the attached **answer form**.
- Marking more than one alternative will result in zero points for that question.
- Each correct answer is worth 5 points.
- Written solutions should not be submitted; only your answers on the answer form will be considered in the assessment and final grading.

Problems 13 — 14: Complete written solutions (max 40 points):

- Use only the provided answer sheets when submitting your solutions and answers.
- For full marks, clear, comprehensive and well-motivated solutions are required. Unclear and unexplained solutions will result in point deductions even if the final answer is correct.
- Check your calculations and solutions before submitting. Careless mistakes will result in unnecessary point deductions.

The maximum total number of points is $60 + 40 = 100$. At least 50 points are required to pass (grades A-E). The grading scale is as follows:

Points	0—39	40—49	50—59	60—69	70—79	80—89	90—100
Grade	F	Fx	E	D	C	B	A

NOTE: Fx and F are failing grades that require re-examination. Students who receive the grade Fx or F cannot supplement extra assignments for a higher grade.

Part one. Multiple choice

1. A salesperson has classified each of her potential customers regarding how likely they are to buy her product. The categories are: high, medium and low. She wants to summarize the data in an adequate chart. Which of the following is a type of chart that is adequate for this situation?

   (a) Bar chart;

   (b) Dotplot;

   (c) Histogram;

   (d) Scatter plot;

   (e) Box-and-whisker plot.

2. Which of the following charts describes the information of **only one** variable?

   (a) Histogram;

   (b) Stacked bar chart;

   (c) Mosaic plot;

   (d) Scatter plot;

   (e) All of the above.

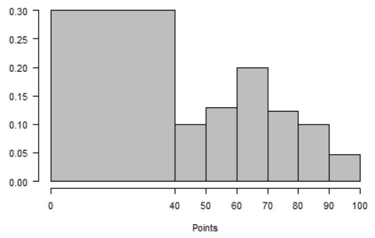
3. Table 1 summarizes the scores of 170 students in an exam of statistics:

Points	$[0\,,40)$	$[40\,,50)$	$[50\,,60)$	$[60\,,70)$	$[70\,,80)$	$[80\,,90)$	$[90\,,100)$
Frequency	51	17	22	34	21	17	8

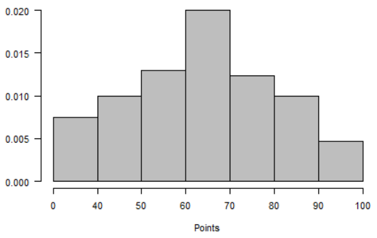
Table 1: Scores of 170 students in an exam of statistics

Which of the following is a histogram that correctly represents the scores in Table 1?

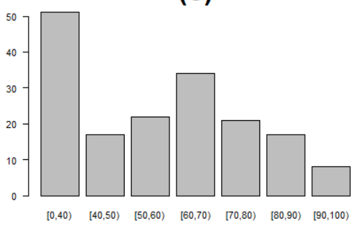
(a)



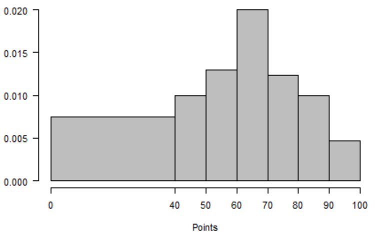
(b)



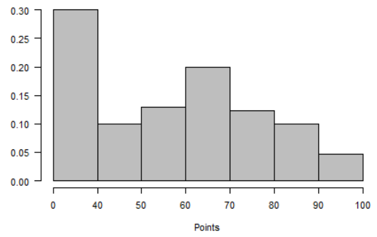
(c)



(d)



(e)



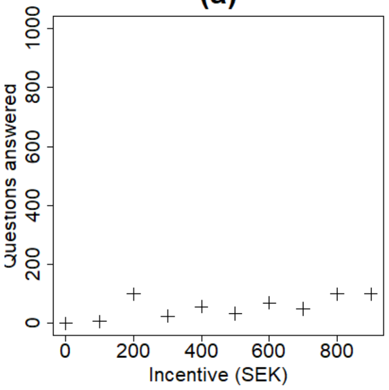
4. A researcher in survey methodology is studying the effect of incentives on item nonresponse. To this end she has selected a sample of ten individuals, offered them different amounts of money and submitted them to a long questionnaire. Then she has measured how many questions they answer before they get tired and decide to stop. Table 2 shows the results.

Incentive (in SEK)	0	100	200	300	400	500	600	700	800	900
Questions answered	0	7	100	25	57	33	70	50	100	100

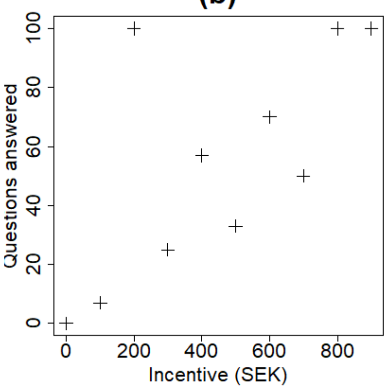
Table 2: Incentives offered to and number of questions answered by a sample of ten respondents

Which of the following is a scatter plot that adequately represents the measurements in Table 2?

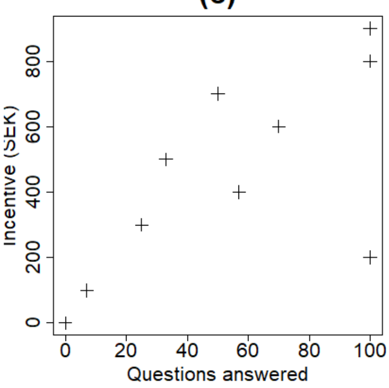
(a)



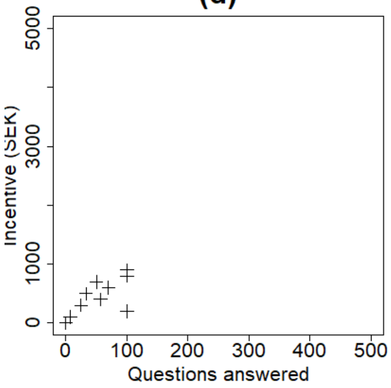
(b)



(c)



(d)



(e)

All of the above

5. Which of the following is **correct** as an interpretation of the mean, $\bar{x}$:

(a) It is the most frequently occurring value;

(b) It is the middle observation;

(c) It is the most likely value;

(d) It is the center of gravity of the observations;

(e) None of the above.

6. A researcher has asked the thirteen married men in a small community about the brideprice they had to pay to the bride's family when they got married. The brideprice values (in USD) are

   20000 3000 10000 20000 13000 0 31000 20000 63000 8000 3000 12000 4000

   What is the **interquartile range** of the brideprice?

   (a) -12000;
   (b) 11000;
   (c) 12000;
   (d) 16000;
   (e) 63000.

7. Which of the following sentences is **correct** regarding the coefficient of determination R^2:

   (a) In *simple* linear regression, if there is a perfect negative linear association between the independent variable x and the dependent variable y, R^2 is equal to -1 (minus one);
   (b) In *simple* linear regression, R^2 is equal to the coefficient of correlation between the independent variable x and the dependent variable y, that is, $R^2 = r_{xy}$;
   (c) In *multiple* linear regression, R^2 should always be preferred over the adjusted coefficient of determination $\bar{R}^2$;
   (d) R^2 may decrease when more variables are added to a model;
   (e) R^2 is equal to the square coefficient of correlation between the predictions $\hat{y}$ and the dependent variable y, that is, $R^2 = r^2_{\hat{y}y}$.

8. In the context of simple linear regression, which of the following is **not** correct?

   (a) the least squares regression is the one that minimizes the *sum of squares error*;
   (b) the intercept b_0 indicates the expected value of the dependent variable y when the independent variable x equals zero;
   (c) the slope b_1 indicates the expected increment in the dependent variable y associated to a one unit increment in the independent variable x;
   (d) the coefficient of determination R^2 indicates the proportion of variability of the dependent variable y that is explained by the independent variable x;
   (e) the coefficient of determination R^2 is equal to the coefficient of correlation between the independent variable x and the dependent variable y.

9. A real estate agent has estimated the regression that explains the closing price (variable *price*, in SEK) of the housing units in a region of interest with respect to their size (variable *size*, in m^2). The fitted regression line is

$$\widehat{price} = 2\,500\,000 + 5000\,size.$$

Which of the following is **not correct**:

(a) if housing unit A has one more square meter than housing unit B, we expect the closing price of A to be around 5000 SEK higher than the closing price of B;

(b) the closing price of a housing unit of size $100m^2$ is expected to be around 3 000 000;

(c) the mean closing price of the housing units in the region of interest is 2 500 000;

(d) the intercept of the fitted regression is 2 500 000;

(e) the slope of the fitted regression is 5000.

10. The teacher of a course in statistics wants to explain the score of students in the final exam (variable *exam*) in terms of the score in a previous home assignment (variable *assignment*) through a linear regression of the form:

$$exam = \beta_0 + \beta_1\,assignment + \epsilon$$

The following table shows the scores of the eight students in the course:

Assignment	42	48	50	50	51	55	59	67
Exam	38	43	57	33	81	50	48	84

The estimated intercept of the regression line of interest is:

(a) -548.7;

(b) - 25.2;

(c) 0.6;

(d) 1.5;

(e) 11.4.

11. Fitting a regression that explains the score of students in the final exam of a course in statistics (variable *exam*) in terms of the score in a previous home assignment (variable *assignment*), yields an intercept $b_0 = -25.2$ and a slope $b_1 = 1.5$. The following table shows the scores of the eight students in the course:

Assignment	42	48	50	50	51	55	59	67
Exam	38	43	57	33	81	50	48	84

The *sum of squares error* —SSE— is:

(a) 0;

(b) 656;

(c) 1594;

(d) 1881;

(e) 2508;

12. Fitting a regression that explains the score of students in the final exam of a course in statistics in terms of the score in a previous home assignment, yields an intercept $b_0 = -25.2$ and a slope $b_1 = 1.5$. The predicted score in the final exam for a student with 50 points in the home assignment is:

(a) -1258.5;

(b) -1185.0;

(c) 12.2;

(d) 49.8;

(e) 75.0.

Part two. Complete solution

13. On September 2024 a company that offers audio streaming services released a new logo. A random sample of six users was drawn and their time using the service before and after the release was measured. Let x_i = "time (in minutes) spent by the ith user using the service one week *before* the new logo was released" and y_i = "time (in minutes) spent by the ith user using the service one week *after* the new logo was released" $(i = 1, \cdots, 6)$. The measurements are shown in the table below:

User	1	2	3	4	5	6
x	40	53	123	139	205	243
y	64	92	124	90	171	212

(a) Calculate the following parameters: **i.** the mean of x, $\bar{x}$; **ii.** the median of x, $\breve{x}$; **iii.** the standard deviation of x, S_x; and **iv.** the interquartile range of x, IQR_x. (10p.)

(b) It is known that the mean and variance of y are, respectively, $\bar{y} = 125.5$ and $S_y^2 = 3144$. It is also known that the correlation between x and y is $r_{xy} = 0.93$. Using this and your results from (a) find the intercept and the slope of a regression that explains the time using the service *after* the release of the new logo in terms of the time using the service *before* the release of the new logo. (**Note:** If you did not solve part (a) use $\bar{x} = 125.5$ and $S_x^2 = 3144$) (5p.)

(c) Using the fitted regression in (b), predict the time (in minutes) that a person using the service 100 minutes *before* the release of the new logo will spend using the service *after* the release of the new logo. (5p.)

(d) Using the fitted regression in (b), find the six residuals e_i. (10p.)

(e) Find the coefficient of determination of the fitted regression. (10p.)