Stockholm University

BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

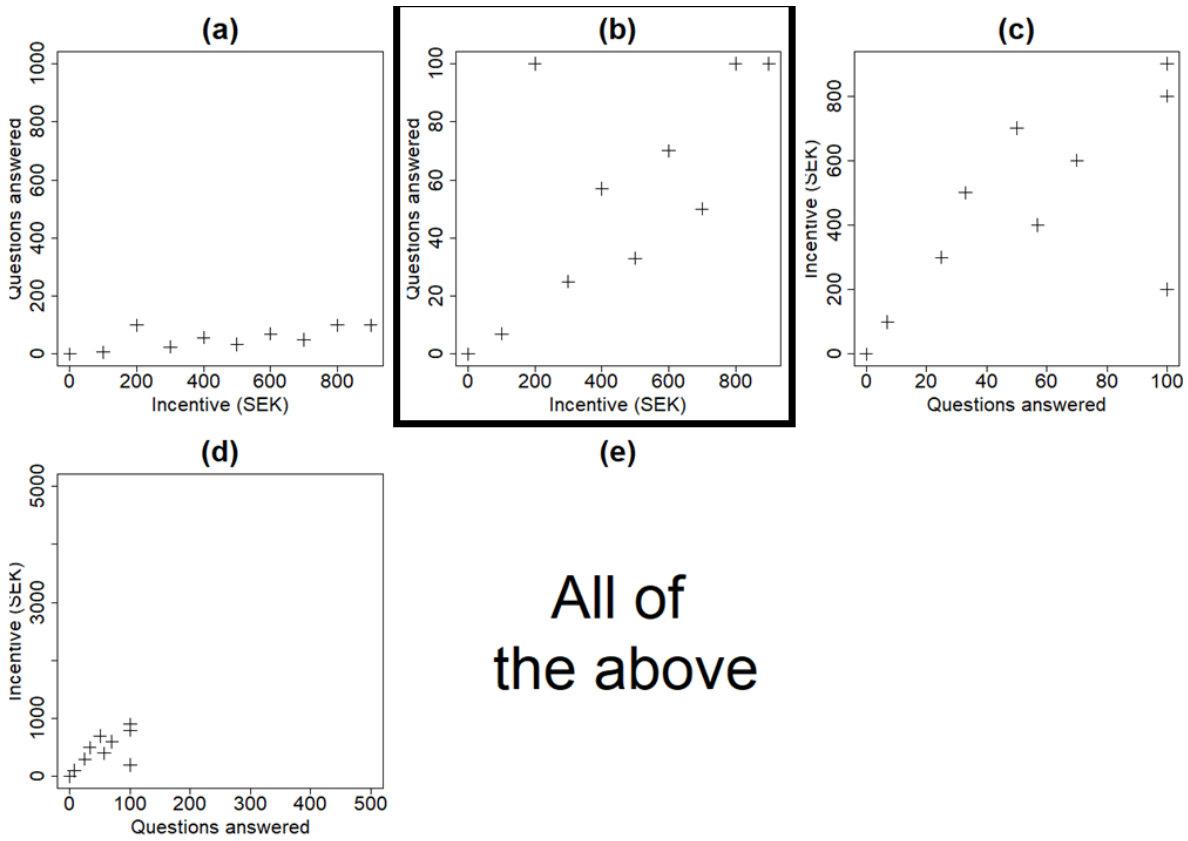
Department of statistics
Edgar Bueno and Jonas Bjermo
2023–02–16

Part one. Multiple choice

1. 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. The results were as follows:

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

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



2. The set of basic outcomes of a random experiment is called:

(a) sample mean;
(b) sample space;
(c) sample frame;
(d) sample size;
(e) none of the above.

3. In hypothesis testing, in which of the following situations is the null hypothesis **rejected**:

(a) If the p-value is smaller than the critical value;
(b) If the p-value is smaller than the significance level α;
(c) If the p-value is smaller than the test statistic;
(d) If the p-value is larger than the significance level α;
(e) If the p-value is larger than the test statistic.

4. 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, $R^2 = r_{xy,s}$;
(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, $R^2 = r^2_{\hat{y}y,s}$.

5. The owner of an electronic store wants to find out if there is any association between the brand of cell phone sold and the day of the week in which the sale is made. Which of the following is an appropriate method to this end:

(a) goodness-of-fit test;
(b) test of independence;
(c) simple linear regression;
(d) multiple linear regression;
(e) time-series analysis.

6. The probability that four students A, B, C and D get a passing grade in an exam of statistics is, respectively, 0.8, 0.7, 0.5 and 0.4. However, A and B have been studying together, and the probability that both of them pass the exam is 0.6. C and D have also been studying together, and the probability that both of them pass the exam is 0.3. Finally, A and B do not know C and D, so their grades can be considered to be independent. What is the probability that all four students pass the exam?

(a) 0.112;
(b) 0.18;
(c) 0.9;
(d) 1;
(e) 1.5.

7. In a card game, the player has three possible outcomes: win, tie or lose. If the player wins (which happens with probability 0.19), he gets two dollars; if the player loses (which happens with probability 0.47), he loses one dollar; in the case of a tie, the player neither wins nor loses any money. What is the expected amount of money of the player at the end of one game?

(a) -0.39;
(b) -0.09;
(c) 0.00;
(d) 0.33;
(e) 1.00.

8. Let X be a continuous random variable with *cumulative distribution function* —cdf— given by

$$F_X(x) = \begin{cases} \frac{x^2}{1000} & \text{if } 0 \leq x \leq 10 \\ \frac{200x - x^2 - 1000}{9000} & \text{if } 10 < x \leq 100 \end{cases}$$

What is the probability that X is larger than 71.54, $P(X > 71.54)$?

(a) 0.0900;
(b) 0.2846;
(c) 0.4882;
(d) 0.5118;
(e) 0.9100;

9. It is known that the lifetime of the light bulbs produced by a company has an expected value of 40 000 hours and a variance of 25 000 000. A random sample of 250 bulbs has been selected. What is the (approximated) probability that the average lifetime of the bulbs in the sample is less than 39 500 hours?

(a) 0.0569;
(b) 0.1056;
(c) 0.4372;
(d) 0.4980;
(e) 0.4999;

10. It is known that the weight in grams of the boxes produced in a packaging line follows a normal distribution with variance equal to 25. A sample of nine boxes has been selected and its weight has been measured:

497.9	493.8	483.8	500.9	506.1	498.5	495.9	487.8	509.2

A 95% confidence interval for the expected weight of the boxes in this packaging line is:

(a) (487.3 , 506.9);
(b) (488.9 , 505.3);
(c) (491.7 , 502.5);
(d) (493.8 , 500.4);
(e) (494.4 , 499.8).

Table 1, which will be used in Exercises 11 and 12, 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

11. What is the approximated mean of the 170 scores given in Table 1?

    (a) 24.3;

    (b) 51.6;

    (c) 52.9;

    (d) 57.8;

    (e) 62.9.

12. The teacher of the course wants to test whether the scores in Table 1 can be considered as a random sample from a (truncated) normal distribution. If it was, the probability in each class would be as given in the following table. (**Hint:** This is a goodness of fit test.)

Points	[0 , 40)	[40 , 50)	[50 , 60)	[60 , 70)	[70 , 80)	[80 , 90)	[90 , 100)
Probability	0.33	0.17	0.17	0.14	0.10	0.06	0.03

    What is the value of the test statistic:

    (a) 1.97;

    (b) 12.59;

    (c) 14.07;

    (d) 18.51;

    (e) 389.92.

Part one. Multiple choice

1. In (b) the points occupy the whole plot region. Also, as the number of questions answered are being considered as dependent on the incentive, the former is taken to be y whereas the latter is taken to be x.

2. See Section 3.1 in Newbold et. al or in the lecture notes.

3. See Section 9.2 in Newbold et. al or Section 10 in the lecture notes.

4. In (a), R^2 would have been equal to 1. In (b), $R^2 = r^2_{xy,s}$. In (c), as R^2 tends to increase with every variable that is added to the model, $\bar{R}^2$ is often preferred. In (d), R^2 cannot decrease when adding more variables to the model.

5. The aim is to determine if two categorical variables are associated or not.

6. Let A, B, C and D be, respectively, the probabilities that students A, B, C and D pass the exam. We have

$$P(A \cap B \cap C \cap D) = P((A \cap B) \cap (C \cap D)) \underset{\text{by ind.}}{=} P(A \cap B) \cdot P(C \cap D) = 0.6 \cdot 0.3 = 0.18.$$

7. Let $X =$ "Amount of money of the player at the end of one game". We have $P_X(2) = 0.19$, $P_X(-1) = 0.47$ and $P_X(0) = 0.34$. Then

$$\mu_X = \sum_x x P_X(x) = 2 \cdot 0.19 + -1 \cdot 0.47 + 0 \cdot 0.34 = -0.09$$

8. We have

$$P(X > 71.54) = 1 - F_X(71.54) = 1 - \frac{200 \cdot 71.54 - 71.54^2 - 1000}{9000} = 0.09.$$

9. Let $X_i =$ "lifetime of the ith randomly chosen light bulb". We know that $\mu_X = 40\,000$ and $\sigma_X^2 = 25\,000\,000$. Therefore, by the Central Limit Theorem, we have $\bar{X} \underset{\text{approx}}{\sim} N(40\,000\,, 100\,000)$, which yields

$$P(\bar{X} < 39\,500) = P(Z < -1.5811) = 0.0569.$$

10. We have

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} = \frac{25}{9}.$$

The confidence interval is then

$$\bar{x}_s \pm z_{\alpha/2}\sigma_{\bar{X}} = 497.1 \pm 1.96 \cdot \frac{5}{3} = (493.8\,, 500.4).$$

11. Taking into account that data has been grouped into seven groups, we can approximate the mean as

$$\bar{x} \approx \frac{1}{N}\sum_{k=1}^{7} f_k m_k = \frac{1}{170}\left(51 \cdot 20 + 17 \cdot 45 + \cdots 8 \cdot 95\right) = \frac{8985}{170} = 52.9.$$

12. Taking into account that the expected number of observations in each category is $E_k = nP_k^0$, the test statistic is

$$\chi_{obs}^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \frac{(51 - 56.1)^2}{56.1} + \frac{(17 - 28.9)^2}{28.9} + \cdots + \frac{(8 - 5.1)^2}{5.1} = 18.51.$$

Part two. Complete solution

13. The teacher of a course in statistics wants to fit the regression that explains the scores obtained in the exam (variable *Exam*) in terms of the number of exercises submitted throughout the course (variable *Exercises*). The following table shows the results for the ten students in the course:

Exercises	0	0	6	12	24	37	43	53	63	79
Exam	64	28	26	83	78	35	77	55	80	62

(a) Calculate the intercept and the slope of the regression of interest. (5p.)

(b) Calculate the ten fitted values and the ten residuals. (5p.)

(c) Calculate the *sum of squares total* —SST— and the *sum of squares error* —SSE—. (5p.)

(d) Calculate and interpret the coefficient of determination. (5p.)

Solution

(a). Taking into account that the explanatory variable is $x =$ "Exercises" and the dependent variable is $y =$ "Exam", we obtain that the slope is

$$b_1 = \frac{\sum_s x_i y_i - n\bar{x}_s\bar{y}_s}{\sum_s x_i^2 - n\bar{x}_s^2} = \frac{20483 - 10 \cdot 31.7 \cdot 58.8}{16993 - 10 \cdot 31.7^2} = \frac{1843.4}{6944.1} = 0.2655.$$

The intercept is

$$b_0 = \bar{y}_s - b_1\bar{x}_s = 58.8 - 0.2655 \cdot 31.7 = 50.38.$$

(b). The fitted values are obtained as $\hat{y}_i = b_0 + b_1 x_i$ and the residuals are $e_i = y_i - \hat{y}_i$ (for all $i = 1, 2, \cdots, 10$), this yields:

x	0	0	6	12	24	37	43	53	63	79
y	64	28	26	83	78	35	77	55	80	62
$\hat{y}$	50.38	50.38	51.98	53.57	56.76	60.21	61.80	64.45	67.11	71.36
e_i	13.62	-22.38	-25.98	29.43	21.24	-25.21	15.20	-9.45	12.89	-9.36

(c). We get

$$SST = \sum_s (y_i - \bar{y}_s)^2 = (64 - 58.8)^2 + (28 - 58.8)^2 + \cdots + (62 - 58.8)^2 = 4377.6$$

and

$$SSE = \sum_s (y_i - \bar{y}_s)^2 = \sum_s e_i^2 = 13.62^2 + (-22.38)^2 + \cdots + (-9.36)^2 = 3888.2.$$

(d). The coefficient of determination is $R^2 = 1 - SSE/SST = 1 - 3888.2/4377.6 = 11.18\%$, which means that around 11% of the variation in the exam' scores can be explained by the number of exercises submitted.

14. A box contains a large number of balls of four different colors: black, white, blue and red. We believe that one quarter of the balls in the box are of each color, i.e. we believe that 25% of the balls are black, 25% are white, 25% are blue and 25% are red. A random sample of 100 balls was drawn. We observe that 19 of them are black, 31 are white, 26 are blue and 24 are red.

    (a) State the hypothesis of interest. (5p.)

    (b) Compute the test statistic and the critical value (using a significance level of 5%). (5p.)

    (c) What is the conclusion regarding the hypothesis? (5p.)

    (d) Would the decision change if the significance level was 1%? Why? (5p.)

Solution

(a). Let P_b, P_w, P_u and P_r be the proportion of black, white, blue and red balls in the box, respectively. The hypothesis is:

$$H_0 : P_b = P_w = P_u = P_r = 0.25 \qquad \text{vs.} \qquad H_1 : \text{At least one } P_k \neq 0.25.$$

(b). We have $O_b = 19$, $O_w = 31$, $O_u = 26$ and $O_r = 24$ and $E_b = E_w = E_u = E_r = 100 \cdot 0.25 = 25$. Then the test statistic is

$$\chi^2_{obs} = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} = \frac{(19-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(24-25)^2}{25} = 2.96.$$

The critical value is

$$\chi^2_{K-1,\alpha} = \chi^2_{4-1,0.05} = 7.815.$$

(c). As $\chi^2_{obs} < \chi^2_{K-1,\alpha}$, the null hypothesis is not rejected.

(d). No. With $\alpha = 0.01$ we get a critical value equal to $\chi^2_{4-1,0.01} = 11.345$, so still the test statistic is smaller than the critical value.