Stockholm University

BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2024–05–30

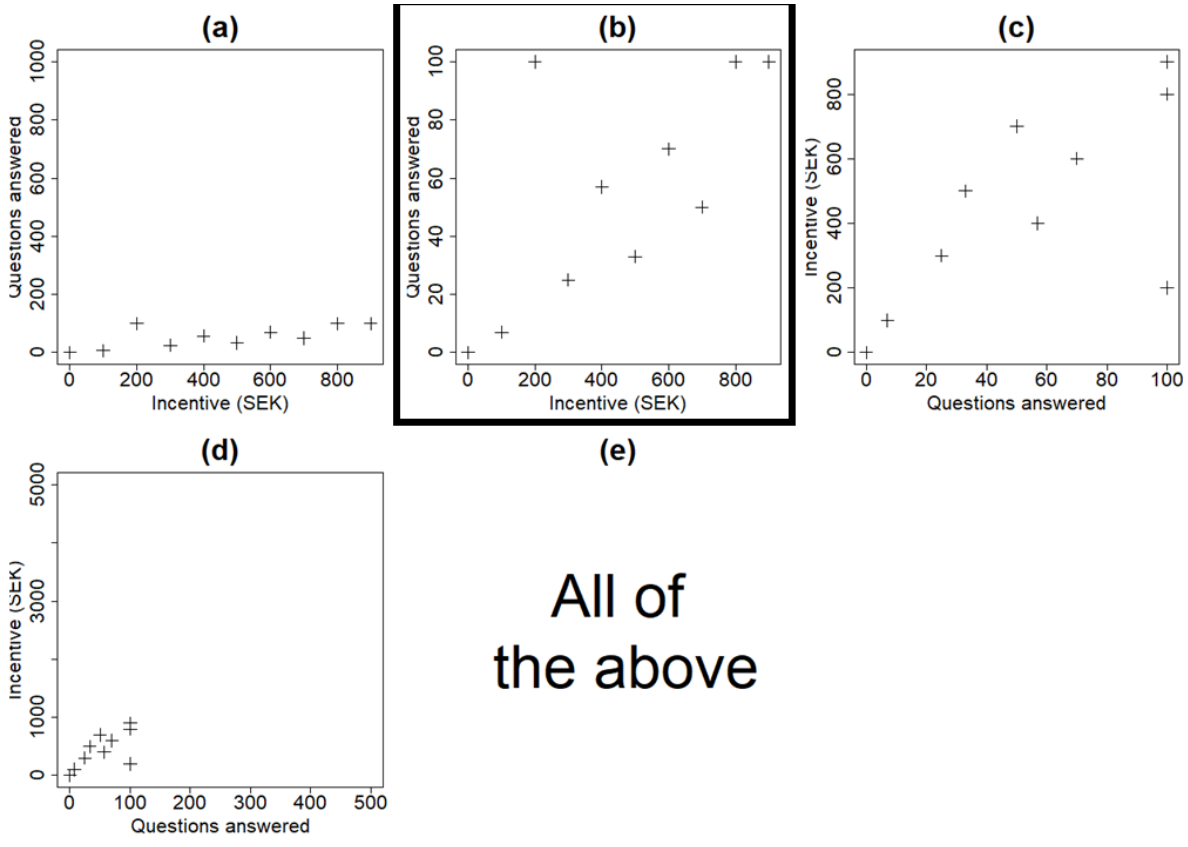
Part one. Multiple choice

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. Table 1 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 1: 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 1?



2. Which of the following sentences is **not** correct regarding the *cumulative probability distribution* of a discrete random variable X, $F_X(x)$:

(a) $\boxed{\lim_{x\to\infty} F_X(x) = 0;}$

(b) $\lim_{x\to-\infty} F_X(x) = 0;$

(c) it is a non-decreasing function;

(d) it takes values between 0 and 1, i.e. $0 \leq F_X(x) \leq 1$ for all x;

(e) it is a step function.

3. It is known that the weight of male african lions has an expectation of 204 kg and a standard deviation of 18 kg. A researcher will select a random sample of 10 male african lions. Which of the following is **correct**:

(a) the variance of the sample will be 324 kg^2;

(b) the mean of the sample will be 204 kg;

(c) the sample mean follows a normal distribution;

(d) $\boxed{\text{the sample mean has an expected value of 204 kg;}}$

(e) the sample mean has a variance of 324 kg^2.

4. The owner of an electronic store wants to verify the hypothesis that 60%, 30% and 10% of the customers buy, respectively, cell phones of the brands A, B and C. Which of the following is an appropriate method to this end:

(a) multiple linear regression;

(b) time-series analysis;

(c) $\boxed{\text{goodness-of-fit test;}}$

(d) simple linear regression;

(e) test of independence.

5. 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) -9500;

(b) 9500;

(c) 11000;

(d) $\boxed{16500;}$

(e) 63000.

6. An ice-cream shop offers 10 different flavors. How many combinations of 2 scoops can be made if the order is important and no flavor can be used more than once?

(a) 20;

(b) 45;

(c) 55;

(d) 90;

(e) 100.

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. The amount of money spent on clothing by students on Stockholm University during 2023 can be modeled by a normal distribution with expected value of 1200 and variance of 40 000. The amount of money spent on course literature can be modeled by a normal distribution with expected value of 800 and variance of 18 000. The covariance between money spent on clothing and course literature is $-24\,000$. What is the probability that one student chosen at random spent more than 2000 on clothing plus course literature?

(a) 0;

(b) 0.0967;

(c) 0.5;

(d) 0.25;

(e) 1.

9. One week before the local elections of a city, a poll is carried out by selecting a random sample of 100 voters. The proportion of individuals in the sample who will vote for the candidate of the party A is 0.4. A 99% confidence interval for the proportion of individuals who will vote for this candidate on the elections is:

(a) $(0\%\,, 99\%)$;

(b) $(27.4\%\,, 52.6\%)$;

(c) $(30.4\%\,, 49.6\%)$;

(d) $(39.4\%\,, 40.6\%)$;

(e) $(39.5\%\,, 40.5\%)$;

10. One week before the local elections of a city, a candidate, Mrs. B, believes that more than 30% of the voters support her. In order to verify her claim, the campaign has selected a sample of 100 voters. 37 out of the 100 voters in the sample claim that they will vote for Mrs. B. The value of the statistic for testing the alternative that the proportion of voters for Mrs. B is larger than 30% is:

(a) 1.45;
(b) 1.66;
(c) 1.98;
(d) 14.50;
(e) 30.03;

11. 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.

12. Two teachers, Teacher 1 and Teacher 2, are in charge of grading 100 exams of statistics. 50 exams are randomly assigned to teacher 1 and the remaining 50 are assigned to teacher 2. The following table summarizes the results:

	Grade						
Teacher	A	B	C	D	E	Fx	F
1	2	2	6	12	8	4	16
2	2	6	7	8	7	4	16

Having a significance level $\alpha = 0.05$, what is the critical value for the hypothesis that grade and teacher are independent:

(a) 2.944;
(b) 12.592;
(c) 17.000;
(d) 67.505;
(e) none of the above.

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. In fact, what is true is $\lim_{x\to\infty} F_X(x) = 1$.

3. See Section 6.2 in Newbold et al. or Chapter 7 in the lecture notes.

4. The aim is to determine if the categories of the variable have probabilities that are consistent with the hypothetical ones.

5. $c = (N+1)p/100 = (13+1)25/100 = 3.5$ then $a = 3$ and $b = 0.5$. Therefore $\breve{x}_{25} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.5)x_{(3)} + 0.5x_{(4)} = 3500$. Also, $c = (N+1)p/100 = (13+1)75/100 = 10.5$ then $a = 10$ and $b = 0.5$. Therefore $\breve{x}_{75} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.5)x_{(10)} + 0.5x_{(11)} = 2000$. Finally, $\text{IQR}_x = \breve{x}_{75} - \breve{x}_{25} = 16500$.

6. We are selecting $x = 2$ flavors out of $n = 10$, which gives $n!/(n-x)! = 90$.

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. Let X and Y be, respectively, the amount of money spent on clothing and on course literature by a randomly chosen student. We have $\mu_X = 1200$, $\sigma_X^2 = 40000$, $\mu_Y = 800$, $\sigma_Y^2 = 18000$ and $\sigma_{XY} = -24000$. We also have $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$. Let $W = X + Y =$ "amount of money spent on clothing plus course literature by a randomly chosen student". Then $W \sim N(\mu_W, \sigma_W^2)$ with $\mu_W = \mu_X + \mu_Y = 2000$ and $\sigma_W^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} = 10000$. Therefore

$$P(W > 2000) = P(Z > 0) = 1 - P(Z < 0) = 1 - 0.5 = 0.5.$$

9. The confidence interval is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.4 \pm 2.5758\sqrt{\frac{0.4(1-0.4)}{100}} = (27.4\%, 52.6\%).$$

10. We have $\bar{x} = \hat{p} = 37/100 = 0.37$ and $\hat{\sigma}_{\bar{X}}^2 = \hat{p}(1-\hat{p})/n = 0.37(1-0.37)/100 = 0.002331$. Therefore the test statistic is $t_{obs} = (\bar{x} - \mu_0)/\hat{\sigma}_{\bar{X}} = (0.37 - 0.3)/0.04828 = 1.45$.

11. $\widehat{exam} = -25.2 + 1.5 \cdot 50 = 49.8$.

12. $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{6,0.05} = 12.592$.

Part two. Complete solution

13. A Formula 1 team is considering to implement a new setup on its cars which will significantly improve the lap times. However, there is concern about the effect that this setup may have on the life of the tyres (i.e. the number of laps before the tyres need to be changed). Thus, the team will test the new setup and measure the life of the tyres.

    By experience, the team knows that the life of the tyres can be adequately described by a normal distribution with a standard deviation of 2.5 laps.

    (a) **What is the smallest sample size needed if they want to estimate the life of the tyres through a 95% confidence interval with a length no larger than one.** (**Note**: the length of a confidence interval is the difference between the upper and the lower limits) (5p.)

$$\left(\bar{x}_s + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) - \left(\bar{x}_s - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) \leq 1 \longrightarrow n \geq (2 \times z_{\alpha/2} \times \sigma_X)^2 = 96.04.$$

    Which is rounded up to 97.

(b) The team decided to test 100 sets of tyres and measure its life. The results are summarized in the following table.

Tyre's life (laps)	7	8	9	10	11	12	13	14	15	16
Frequency	7	6	11	6	10	11	22	11	9	7

Find the sample mean of the tyre's life, $\bar{x}_s$. (5p.)

$$\bar{x}_s = \frac{1}{n}\sum_{k=1}^{K} f_k x_k = \frac{1}{100}\left(7\times 7 + 6\times 8 + \cdots + 7\times 16\right) = 11.85.$$

(c) **Find a 95% confidence interval for the life of the tyres.** (**Note**: If you did not solve 13b use $\bar{x}_s = 11$.) (5p.)
We have $\sigma^2_{\bar{X}} = \sigma^2_X/n = 2.5^2/100 = 0.0625$, then $\sigma_{\bar{X}} = 0.25$ and

$$\bar{x}_s \pm z_{\alpha/2}\sigma_{\bar{X}} = 11.85 \pm 1.96 \cdot 0.25 = (11.36\,, 12.34).$$

(d) Based on historical records, it is known that the tyre's life expectation with the previous setup is equal to 12 laps. Using a significance level of 5%, **test the hypothesis that the new setup has had any effect on the tyre's life** (i.e. test the two-sided alternative). (**Note**: If you did not solve 13b use $\bar{x}_s = 11$.) (5p.)
The hypothesis is

$$H_0 : \mu_X = 12 \qquad \text{vs.} \qquad H_1 : \mu_X \neq 12.$$

The test statistic is

$$z_{obs} = \frac{\bar{x}_s - \mu_0}{\sigma_{\bar{X}}} = \frac{11.85 - 12}{0.25} = -0.6.$$

The critical value is $z_{\alpha/2} = 1.96$. As $0.6 = |z_{obs}| < z_{\alpha/2} = 1.96$ the null hypothesis is not rejected which means that there is no evidence that supports that there has been any change on the tyre's life with the new setup.

14. A machine produces, at random, pyramids and cubes of three different colors (black, white or red).

(a) The machine's operator sets the probability with which the machine will produce each combination of shape and color. During one particular day, the operator has set the following probabilities:

		Color		
		Black	White	Red
Shape	Pyramid	0.2	0.25	0.1
	Cube	0.0	0.05	0.4

Are the shape and the color independent of each other? (5p.)
Let us add the marginal distributions to the table:

		Color			
		Black	White	Red	Total
Shape	Pyramid	0.2	0.25	0.1	0.55
	Cube	0.0	0.05	0.4	0.45
	Total	0.2	0.3	0.5	1

Now let X = "shape" (p = "pyramid" and c = "cube") and Y = "color" (b = "black", w = "white" and r = "red"). As $0.2 = P(X = p, Y = b) \neq P(X = p)P(Y = b) = 0.55 \cdot 0.2 = 0.11$ we conclude that the shape and the color are not chosen independently by the machine.

(b) A customer (who does not know the probabilities set by the operator) wants to determine if the machine is producing objects whose shape and color are chosen independently. To this end, he asks the operator to activate the machine $n = 10$ times. The results are summarized below:

		Color		
		Black	White	Red
Shape	Pyramid	1	1	1
	Cube	0	1	6

Using a significance level of 5%, test if the machine is producing objects whose shape and color are chosen independently. (5p.)

Let us add the marginals to the table:

		Color			
		Black	White	Red	Total
Shape	Pyramid	1	1	1	3
	Cube	0	1	6	7
	Total	1	2	7	10

The hypothesis is

$$H_0 : X \text{ and } Y \text{ are independent} \qquad \text{vs.} \qquad H_1 : X \text{ and } Y \text{ are dependent.}$$

The expected frequencies $E_{ij} = R_i C_j / n$ are

		Color		
		Black	White	Red
Shape	Pyramid	0.3	0.6	2.1
	Cube	0.7	1.4	4.9

The test statistic is

$$\chi^2_{obs} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 3.537.$$

The critical value is $\chi^2_{(r-1)(c-1),\alpha} = 5.991$.

As $3.537 = \chi^2_{obs} < \chi^2_{(r-1)(c-1),\alpha} = 5.991$ we do not reject the null hypothesis, in other words, there is no evidence against the hypothesis that the shape and the color are independently chosen by the machine.

(c) A second customer (who does not know the probabilities set by the operator either) also wants to determine if the machine is producing objects whose shape and color are chosen independently. To this end, she asks the operator to activate the machine $n = 100$ times. The results are summarized below:

		Color		
		Black	White	Red
Shape	Pyramid	17	21	10
	Cube	0	4	48

Using a significance level of 5%, test if the machine is producing objects whose shape and color are chosen independently. (5p.)

Let us add the marginals to the table:

		Color			
		Black	White	Red	Total
Shape	Pyramid	17	21	10	48
	Cube	0	4	48	52
	Total	17	25	58	100

The hypothesis is the same as above.

The expected frequencies $E_{ij} = R_i C_j / n$ are

		Color		
		Black	White	Red
Shape	Pyramid	8.16	12	27.84
	Cube	8.84	13	30.16

The test statistic is

$$\chi^2_{obs} = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 53.38.$$

The critical value is the same as above, $\chi^2_{(r-1)(c-1),\alpha} = 5.991$.

As $53.38 = \chi^2_{obs} > \chi^2_{(r-1)(c-1),\alpha} = 5.991$ we reject the null hypothesis, in other words, the evidence indicates that the shape and the color are not chosen independently by the machine.

(d) **What are your conclusions on the results in 14a, 14b and 14c?** (**Hint:** Take into account that from 14a you know if the hypothesis holds or not.) (5p.)

From 14a we know that the shape and the color are not chosen independently by the machine, so the null hypothesis in both 14b and 14c is false. However, in 14b we fail to reject it whereas in 14c we correctly reject it. A sample of size $n = 10$ is not providing enough evidence for rejecting the hypothesis, whereas a sample of size $n = 100$ has given enough enough for rejecting it.

This is related to the power of a test, i.e. the probability of rejecting the null hypothesis when it should be rejected. It turns out that the power increases with the sample size.