BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2024–10–11

Part one. Multiple choice

1. The households' size in Sweden can be adequately modeled by a distribution with expectation 2.15 individuals. A consulting firm is going to draw a random sample of n households. Which of the following sentences is **correct**?

(a) the expectation of the sample mean is 2.15 individuals **only** if the distribution is symmetric;

(b) the expectation of the sample mean is 2.15 individuals **only** if the sample size n is large enough;

(c) the expectation of the sample mean is 2.15 individuals;

(d) the expectation of the sample mean is 2.15 individuals **only** if the distribution is the normal distribution;

(e) the expectation of the sample mean is **not** 2.15 individuals.

2. A real estate agent has estimated the regression that explains the closing price (in SEK) of the housing units in a region of interest with respect to their size (variable *size*, in m^2), the number of rooms (variable *room*) and a dummy variable indicating whether the housing unit has a balcony (variable *balcony*). The output is shown below.

Regression Statistics						
Multiple R	0.2363					
R Square	0.0558					
Adjusted R Square	0.0495					
Standard Error	720474					
Observations	450					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	1.369E+13	4.563E+12	8.791	1.129E-05	
Residual	446	2.315E+14	5.191E+11			
Total	449	2.452E+14				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2000000	200000	10.00	0.0000	1606943	2393057
SIZE	-5000	3000	-1.67	0.0963	-10896	896
ROOM	400000	80000	5.00	0.0000	242777	557223
BALCONY	30000	70000	0.43	0.6684	-107570	167570

Considering a significance level of 1%, which of the following is **not correct**:

(a) around 5% of the variability of the closing price is being explained by the three explanatory variables;

(b) the combined effect of the three explanatory variables is significant for explaining the closing price;

(c) for two housing units with the same size and the same number of rooms, it is expected that one with a balcony will be sold at around 30000 SEK more than one without a balcony;

(d) the number of rooms has a significant effect on the closing price;

(e) having a balcony has a significant effect on the closing price.

3. 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) time-series analysis;

(b) multiple linear regression;

(c) test of independence;

(d) simple linear regression;

(e) goodness-of-fit test.

4. Which of the following is **not** one of the components of a time series:

   (a) Smooth component;

   (b) Cyclical component;

   (c) Trend component;

   (d) Seasonal component;

   (e) Irregular component.

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 **median** of the brideprice?

   (a) 12000;

   (b) 15500;

   (c) 15923;

   (d) 20000;

   (e) 31000.

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. The amount of money spent on clothing by students on Stockholm University during 2024 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. A car rental company knows by experience that 10% of the customers rent a *sport utility vehicle* —suv— and that the customers' choice is independent of each other. What is the probability that, out of the next 100 customers, the number of customers renting a suv is larger than eight but at most eleven?

   (a) 0.0300;
   (b) 0.1322;
   (c) 0.3000;
   (d) 0.3822;
   (e) 0.6178;

10. A teacher knows by experience that the number of points students get in the final exam can be adequately modeled by a normal distribution with variance equal to 441. At the end of one semester, a sample of nine students has been selected and their score in the exam has been measured. The sample mean is $\bar{x}_s = 53.6$. At the end of the next semester, a sample of nine students has been selected independently and their score in the exam has been measured. The sample mean is $\bar{y}_s = 54.6$. A 99% confidence interval for the difference of the expected number of points between the first and the second semester is:

    (a) $(-253.4\,, 251.4)$;
    (b) $(-229.0\,, 227.0)$;
    (c) $(-50.5\,, 48.5)$;
    (d) $(-26.5\,, 24.5)$;
    (e) $(-24.0\,, 22.0)$.

11. One week before the local elections of a city, a candidate, Mrs. A, 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. A. The value of the test statistic for testing the alternative that the proportion of voters for Mrs. A is larger than 30% is:

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

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) 67.505;
   (d) 17.000;
   (e) none of the above.

Part one. Multiple choice

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

2. The p-value for the test that the coefficient associated to balcony is zero, $\beta_3 = 0$, is 0.6684 therefore (e) is not correct.

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

4. See Section 16.1.

5. $c = (N+1)p/100 = (13+1)50/100 = 7$ then $a = 7$ and $b = 0$. Therefore $\breve{x}_{50} = (1-b)x_{(a)} + bx_{(a+1)} = x_{(7)} = 12000$.

6. Let A, B, C and D be, respectively, the events 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. 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. Let Y = "Number of customers renting a suv". We have $Y \sim Bin(n, P)$ with $n = 100$ and $P = 0.1$. As n is large, we have that, approximately, $Y \sim N(nP, nP(1-P)) = N(10, 9)$, thus

$$P(8 < Y \leq 11) = P(Y \leq 11) - P(Y \leq 8) \approx P(Z \leq 0.33) - P(Z \leq -0.67) = 0.62930 - 0.25143 = 0.38.$$

10. Let x_i be the exam's score of the ith student in the first sample and y_i be the exam's score of the ith student in the second sample. We have

$$\bar{d}_s = \bar{x} - \bar{y} = -1 \qquad \text{and} \qquad \sigma_{\bar{D}}^2 = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y} = 98.$$

The confidence interval is

$$\bar{d}_s \pm z_{\alpha/2}\sigma_{\bar{D}} = -1 \pm 2.57 \cdot 9.899 = (-26.5, 24.5).$$

11. 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 $t_{obs} = (\bar{x} - \mu_0)/\hat{\sigma}_{\bar{X}} = (0.37 - 0.3)/0.04828 = 1.45$.

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

Part two. Complete solution

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" $(i = 1, \cdots, 6)$. It is known that the times using the service are adequately described by a normal distribution.

Some summary statistics of the collected data are shown below:

$$\sum_s x_i = 803, \qquad \sum_s y_i = 753, \qquad \sum_s x_i^2 = 139\,933, \qquad \sum_s y_i^2 = 110\,221, \qquad \sum_s x_i y_i = 121\,769$$

13. **Calculate the following statistics:**

i. the mean of x, $\bar{x}_s = \sum_s x_i/n = 133.8$;

ii. the mean of y, $\bar{y}_s = \sum_s y_i/n = 125.5$;

iii. the variance of x, $S_{x,s}^2 = \left(\sum_s x_i^2 - \left(\sum_s x_i\right)^2/n\right)/(n-1) = 6493$;

iv. the variance of y, $S_{y,s}^2 = \left(\sum_s y_i^2 - \left(\sum_s y_i\right)^2/n\right)/(n-1) = 3144$;

v. the covariance between x and y, $S_{xy,s} = \left(\sum_s x_i y_i - \left(\sum_s x_i\right)\left(\sum_s y_i\right)/n\right)/(n-1) = 4199$.

14. Confidence interval estimation:

   (a) **Find a 95% confidence interval for the expected change (*before* minus *after*) in time using the service.**

   We have $\bar{d}_s = \bar{x}_s - \bar{y}_s = 8.3$, $S^2_{d,s} = S^2_{x,s} - 2S_{xy,s} + S^2_{y,s} = 1240$, $t_{n-1,\alpha/2} = 2.571$ and $\hat{\sigma}^2_{\bar{D}} = S^2_{d,s}/n = 206.7$. Thus, an (approximately) 95% confidence interval for the expected change is

   $$\bar{d}_s \pm t_{n-1,\alpha/2}\hat{\sigma}_{\bar{D}} = 8.3 \pm 2.571 \cdot 14.38 = (-28.6\,, 45.3).$$

   (b) **Based on your results on 14a, has there been a significant change in the time using the service? Justify your answer.**

   No. The confidence interval contains the value zero, which means that it is likely that the actual change was equal to zero. In other words, there is no evidence for concluding that there was any significant change in the time using the service after the new logo was released.

15. Simple linear regression:

   (a) **Estimate 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.**

   $$b_1 = \frac{S_{xy,s}}{S^2_{x,s}} = 0.6466 \qquad \text{and} \qquad b_0 = \bar{y}_s - b_1\bar{x}_s = 38.96.$$

   (b) **Taking into account that $\sum_s e_i^2 = 2145$, estimate the variance of the intercept and the slope coefficients, $\hat{\sigma}^2_{b_0}$ and $\hat{\sigma}^2_{b_1}$.**

   We have $\hat{\sigma}^2_\epsilon = \sum_s e_i^2/(n-2) = 536.3$, thus

   $$\hat{\sigma}^2_{b_0} = \frac{\hat{\sigma}^2_\epsilon \sum_s x_i^2}{n(n-1)S^2_{x,s}} = 385.2 \qquad \text{and} \qquad \hat{\sigma}^2_{b_1} = \frac{\hat{\sigma}^2_\epsilon}{(n-1)S^2_{x,s}} = 0.0165.$$

   (c) **Find a 95% confidence interval for the intercept, β_0, and the slope, β_1.**

   An (approximately) 95% confidence intervals for β_0 is

   $$b_0 \pm t_{n-2,\alpha/2}\hat{\sigma}_{b_0} = 38.96 \pm 2.776 \cdot 19.63 = (-15.53\,, 93.45).$$

   An (approximately) 95% confidence intervals for β_1 is

   $$b_1 \pm t_{n-2,\alpha/2}\hat{\sigma}_{b_1} = 0.6466 \pm 2.776 \cdot 0.1285 = (0.2898\,, 1.0035).$$

   (d) **i. Does the confidence interval for the intercept contain the value zero?** Yes.

   ii. Does the confidence interval for the slope contain the value one? Yes.

   iii. In terms of the linear model, what does it mean for the intercept to be equal to zero and, at the same time, the slope to be equal to one in this particular problem?

   The linear model is of the form $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. If $\beta_0 = 0$ and $\beta_1 = 1$ the model reduces to $Y_i = x_i + \epsilon_i$, which means that, for any user, the time using the service after the new logo was released (y) is the same as the time using the service before (x) plus some random noise. In other words, there has been no change in the using times due to the release of the new logo. This is consistent with what we found in Exercise 14a.