Stockholm University

BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2023–06–01

Part one. Multiple choice

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

   (a) Cluster bar chart;

   (b) Histogram;

   (c) Component bar chart;

   (d) Scatter plot;

   (e) All of the above.

2. Which of the following is **correct** regarding the variance of a random variable:

   (a) if the expectation of the random variable is negative, the variance will be negative;

   (b) it indicates the difference between the largest and the smallest outcome of the random variable;

   (c) it indicates how spread are the outcomes of the random variable around its expectation;

   (d) it is only defined for continuous random variables, not for discrete random variables;

   (e) it is measured in the same units as the random variable itself.

3. Which of the following sentences is **not correct**:

   (a) an estimate is the specific value taken by the estimator under the observed sample;

   (b) among two unbiased estimators, the one with largest variance shall be preferred;

   (c) an estimator is a statistic that approximates the parameter of interest;

   (d) an estimator is said to be unbiased if its expected value equals the parameter of interest;

   (e) an estimator is a random variable.

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

   (a) the coefficient of determination R^2 indicates the proportion of variability of the dependent variable y that is explained by the independent variable x;

   (b) the coefficient of determination R^2 is equal to the coefficient of correlation between the independent variable x and the dependent variable y;

   (c) the intercept b_0 indicates the expected value of the dependent variable y when the independent variable x equals zero;

(d) the slope b_1 indicates the expected increment in the dependent variable y associated to a one unit increment in the independent variable x;

(e) the least squares regression is the one that minimizes the *sum of squares error*.

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 **mode** of the brideprice?

(a) 12000;

(b) 15500;

(c) 15923;

(d) 20000;

(e) 31000.

6. An ice-cream shop offers 10 different flavors. How many combinations of 2 scoops can be made if the order is not important and the flavors 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 variance of the amount of money of the player at the end of one game?

(a) -0.09;

(b) 0;

(c) 1.22;

(d) 5;

(e) 5.2;

8. Coffee Inc. is a company that imports two types of coffee to Sweden. The number of sacks of the type *Arabica* imported every month can be described by a normally distributed random variable with expectation 50 and variance 4. The number of sacks of the type *Robusta* imported every month can be described by a normally distributed random variable with expectation 45 and variance 9. The covariance between the number of sacks of both types is 5. The price of one sack of the type arabica is 400 SEK and the price of one sack of the type robusta is 200 SEK. What is the probability that the total sales during one month are less than 30000 SEK?

(a) 0.77;

(b) 0.80;

(c) 0.84;

(d) 0.87;

(e) 1.00.

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 (approximated) probability that, out of the next 100 customers, the number of customers renting a suv is larger than eight but at most eleven?

(a) 0.03;

(b) 0.13;

(c) 0.30;

(d) 0.38;

(e) 0.62;

10. 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) $(39.5\%\,, 40.5\%)$;

11. 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. 45 out of the 100 voters in the sample claim that they will vote for Mrs. B. With a significance level of 1%, which of the following is **correct**. (**Hint:** Use the alternative $P > 0.3$):

(a) the critical value is 2.36 and the test statistic is 3.02, therefore the null hypothesis is rejected;

(b) the critical value is 2.36 and the test statistic is 3.02, therefore the null hypothesis is not rejected;

(c) the critical value is 2.33 and the test statistic is 60.61, therefore the null hypothesis is rejected.

(d) the critical value is 3.02 and the test statistic is 2.36, therefore the null hypothesis is rejected;

(e) the critical value is 3.02 and the test statistic is 2.36, therefore the null hypothesis is not rejected;

12. The teacher of a course in statistics wants to test whether $X =$ "grade in the first assignment" (Pass or Fail) and $Y =$ "grade in the exam" (A, C, E or F) are independent. The following table summarizes the results of the 120 students in the course:

		Y			
		A	C	E	F
X	Pass	14	17	34	42
	Fail	2	1	1	9

What is the value of the test statistic:

(a) 1.960;

(b) 5.321;

(c) 7.815;

(d) 10.648;

(e) 145.461;

Part one. Multiple choice

1. See Chapter 1 in Newbold et al. or Chapter 2 in the lecture notes.

2. See Section 4.3 in Newbold et al. or 4.2 in the lecture notes.

3. See Section 7.1 in Newbold et al. or 6.2 in the lecture notes.

4. See Chapter 11 in Newbold et al. or Section 10.1 in the lecture notes.

5. The mode is 2000, as this is the most frequently occurring value.

6. We are selecting $x = 2$ flavors out of $n = 10$, which gives $\frac{(n+x-1)!}{x!(n-1)!} = 55$.

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

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

and

$$\sigma_X^2 = \sum_x x^2 P_X(x) - \mu_X^2 = \left(2^2 \cdot 0.19 + (-1)^2 \cdot 0.47 + 0^2 \cdot 0.34\right) - (-0.09^2) = 1.22.$$

8. Let $X =$ "Number of sacks of the type *Arabica* imported during one month" and $Y =$ "Number of sacks of the type *Robusta* imported during one month". We have $X \sim N(\mu_X\,, \sigma_X^2)$ with $\mu_X = 50$ and $\sigma_X^2 = 4$ and $Y \sim N(\mu_Y\,, \sigma_Y^2)$ with $\mu_Y = 45$ and $\sigma_Y^2 = 9$. We have also that $\sigma_{XY} = 5$.

   Let $W = 400X + 200Y =$ "Total sales during one month". We have $W \sim N(\mu_W\,, \sigma_W^2)$ with $\mu_W = 400\mu_X + 200\mu_Y = 29000$ and $\sigma_W^2 = 400^2\sigma_X^2 + 200^2\sigma_Y^2 + 2 \cdot 400 \cdot 200\sigma_{XY} = 1\,800\,000$. Then $\sigma_W = 1342$. This yields

$$P(W < 30000) = P(Z < 0.75) = 0.77.$$

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. The confidence interval is

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

11. We have

$$\hat{\sigma}^2_{\bar{X}} = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.45(1-0.45)}{100} = 0.002475.$$

Therefore the test statistic is

$$t_{obs} = \frac{\bar{x}_s - \mu_0}{\hat{\sigma}_{\bar{X}}} = \frac{0.45 - 0.3}{0.002475^{0.5}} = 3.02.$$

Regarding the critical value, we have $t_{n-1,\alpha} = t_{99,0.01} = 2.36$.

As $t_{obs} > t_{n-1,\alpha}$ the null hypothesis is rejected.

12. We want to test if these two categorical variables are independent or not. First, let us add the marginal totals to the table of observations O_{ij} given in the exercise:

		Y				
		A	C	E	F	R_i
X	Pass	14	17	34	42	107
	Fail	2	1	1	9	13
	C_j	16	18	35	51	120

Now, let us find the expectations $E_{ij} = R_i C_j / n$. This yields

		Y			
		A	C	E	F
X	Pass	14.27	16.05	31.21	45.48
	Fail	1.73	1.95	3.79	5.53

Therefore, we obtain

$$\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 5.321.$$

Part two. Complete solution

13. The owner of a record store has advertised his business in a popular search engine. Before advertising, he collected a sample of the sales during eight randomly selected days. The observed sales (in SEK) were:

8100 6000 10600 10000 5800 8900 7000 8700

After advertising, he collected a sample of the sales during twelve randomly selected days. The observed sales (in SEK) were:

13500 11500 12800 12400 7800 11300 10000 10000 3500 11000 9400 8900

The owner knows, by experience, that the sales can be adequately described by a normal distribution. In addition, he considers it safe to assume that the variance before and after advertising is the same.

(a) Find a 95% confidence interval for the expected change in sales (after minus before). (5p.)

(b) Using a significance level of 5%, test the null hypothesis that there was no change in sales against the two-sided alternative. **i.** State the hypothesis of interest; **ii.** Compute the test statistic and the critical value; **iii.** What is the conclusion regarding the hypothesis? (5p.)

(c) Consider the null hypothesis that, after advertising, the sales increased by μ_0 SEK against the two-sided alternative. Using a significance level of 5%, for what values of μ_0 would the null hypothesis not been rejected. (5p.)

(d) Compare your answers in (a) and (d). What do you conclude? (5p.)

Solution

(a). Let y_i be the sales during the ith random day before advertising and x_i be the sales during the ith random day after advertising. We are interested in a confidence interval for $\mu_D = \mu_X - \mu_Y$. The confidence interval is of the form $\bar{d}_s \pm t_{n_x+n_y-2,\alpha/2}\hat{\sigma}_{\bar{D}}$ with $\hat{\sigma}^2_{\bar{D}} = S_p^2/n_x + S_p^2/n_y$ and $S_p^2 = \left((n_x-1)S^2_{x,s} + (n_y-1)S^2_{y,s}\right)/(n_x+n_y-2)$. We obtain

$$S_p^2 = \frac{(n_x-1)S^2_{x,s} + (n_y-1)S^2_{y,s}}{n_x+n_y-2} = 5\,602\,292$$

and

$$\hat{\sigma}^2_{\bar{D}} = S_p^2/n_x + S_p^2/n_y = 1\,167\,144.$$

We also have $\bar{d}_s = \bar{x}_s - \bar{y}_s = 2038$ and $t_{n_x+n_y-2,\alpha/2} = t_{18,0.025} = 2.101$. This yields

$$\bar{d}_s \pm t_{n_x+n_y-2,\alpha/2}\hat{\sigma}_{\bar{D}} = 2038 \pm 2.101 \cdot 1080 = (-232\,,4307).$$

(b). **i.** The hypothesis of interest is

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X \neq \mu_Y$$

or equivalently

$$H_0 : \mu_D = \mu_X - \mu_Y = 0 \quad \text{vs.} \quad H_1 : \mu_D = \mu_X - \mu_Y \neq 0.$$

ii. The test statistic is

$$t_{obs} = \frac{\bar{d}_s - 0}{\hat{\sigma}_{\bar{D}}} = \frac{2038-0}{1080} = 1.886.$$

And the critical value is $t_{n_x+n_y-2,\alpha/2} = 2.101$.

iii. As $1.886 = |t_{obs}| < t_{n_x+n_y-2,\alpha/2} = 2.101$, the null hypothesis is not reject and we conclude that the data provides no evidence for saying that advertising has change the expected sales.

(c). The hypothesis of interest is

$$H_0 : \mu_D = \mu_X - \mu_Y = \mu_0 \quad \text{vs.} \quad H_1 : \mu_D = \mu_X - \mu_Y \neq \mu_0.$$

The null hypothesis is not rejected if $|t_{obs}| < t_{n_x+n_y-2,\alpha/2}$, this means that the null hypothesis is not rejected if

$$\left|\frac{\bar{d}_s - \mu_0}{\hat{\sigma}_{\bar{D}}}\right| < t_{n_x+n_y-2,\alpha/2} \longrightarrow -t_{n_x+n_y-2,\alpha/2} < \frac{\bar{d}_s - \mu_0}{\hat{\sigma}_{\bar{D}}} < t_{n_x+n_y-2,\alpha/2} \longrightarrow$$

$$\bar{d}_s - t_{n_x+n_y-2,\alpha/2}\hat{\sigma}_{\bar{D}} < \mu_0 < \bar{d}_s + t_{n_x+n_y-2,\alpha/2}\hat{\sigma}_{\bar{D}} = 2038 - 2.101\cdot 1080 < \mu_0 < 2038 + 2.101\cdot 1080 = (-232\,,4307).$$

In other words, the null hypothesis is not rejected for values of μ_0 in the interval $(-232\,,4307)$.

(d). We obtain the same results in (a) and (c). This means that there is a relation between confidence interval estimation and hypothesis testing: in a two-sided test, the null hypothesis is not rejected if and only if the hypothesized value μ_0 is included in the confidence interval.

14. In order to fit the linear regression that explains $y =$ "sleeping time" (variable *sleep*, in hours per week) in terms of $x_1 =$ "working time" (variable *work*, in hours per week), data on $n = 700$ individuals was collected. Some summary statistics of the collected data are shown below:

$$\sum_s x_{1i} = 24980 \qquad \sum_s y_i = 38430 \qquad \sum_s x_{1i}^2 = 1\,060\,000 \qquad \sum_s x_{1i}y_i = 1\,333\,000$$

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

(b) Predict the number of hours sleeping for a person working forty hours per week. (5p.)

A second explanatory variable

$$x_{2i} = \begin{cases} 1 & \text{if the } i\text{th individual has kids younger than 3} \\ 0 & \text{otherwise} \end{cases}$$

(variable *young_kid*) was added to the model. The estimated coefficients and their estimated standard errors are shown in the following table.

	Coefficients	Standard Error
Intercept	59.80	0.6561
work	-0.1507	0.0168
young_kid	-13.82	47.33

(c) Using a significance level of 5%, test the hypothesis that *young_kid* is significant in the model. (5p.)

(d) Interpret the three estimated coefficients. (5p.)

Solution

(a). We have $\bar{x}_1 = \sum_s x_{1i}/n = 24980/700 = 35.69$ and $\bar{y} = \sum_s y_i/n = 38430/700 = 54.90$, which yields

$$b_1 = \frac{\sum_s x_{1i}y_i - n\bar{x}_1\bar{y}}{\sum_s x_{1i}^2 - n\bar{x}_1^2} = -0.2278 \qquad \text{and} \qquad b_0 = \bar{y} - b_1\bar{x}_1 = 63.03.$$

(b). We have $\hat{y} = b_0 + b_1 x_0 = 53.92$.

(c). The model is of the type $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. The hypothesis of interest is

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_2 \neq 0.$$

We have $t_{obs} = (b_2 - 0)/\hat{\sigma}_{b_2} = (-13.82 - 0)/47.33 = -0.292$ and $t_{n-K-1,\alpha/2} = t_{697,0.025} = 1.963$.

As $0.292 = |t_{obs}| < t_{n-K-1,\alpha/2} = 1.963$ the null hypothesis is not rejected and we conclude that *young_kid* is not significant in the model.

(d). We expect a person who does not work ($x_1 = 0$) and does not have young kids ($x_2 = 0$) to sleep around 59.80 hours per week.

Having the same status regarding having young kids, we expect a person who works one hour more than another one, to sleep 0.1507 hours (nine minutes) less.

Among two persons who work the same amount of time, we expect one person having young kids to sleep 13.82 hours less than one person who does not have young kids. Note, however, that this effect is not significantly different than zero.