BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2024–01–11

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) $\boxed{\text{Bar chart;}}$

   (b) Histogram;

   (c) Box-and-whisker plot;

   (d) Scatter plot;

   (e) Steam-and-leaf display.

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

   (a) it is a non-decreasing function;

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

   (c) $\boxed{\text{it is a step function;}}$

   (d) $\lim\limits_{x \to -\infty} F_X(x) = 0$;

   (e) $\lim\limits_{x \to \infty} F_X(x) = 1$.

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

   (a) $\boxed{\begin{array}{l}\text{the coefficient of determination } R^2 \text{ is equal to the coefficient of correlation} \\ \text{between the independent variable } x \text{ and the dependent variable } y\text{;}\end{array}}$

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

   (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*.

4. A researcher is studying the relation between the continent in which a country is located (Africa, Asia, Europe and Latin America) and its Human Development Index —HDI— (Very high, High, Medium, Low). Using a statistical software, the researcher carries out a test of independence between both variables and obtains a p-value smaller than 0.00001. Considering a significance level $\alpha = 0.01$, which of the following is **correct** regarding the test that has been implemented:

   (a) the continent and the HDI are positively correlated;

   (b) the continent and the HDI are negatively correlated;

   (c) the continent and the HDI are independent;

   (d) $\boxed{\text{the continent and the HDI are dependent;}}$

   (e) the information provided is not enough for making a conclusion.

5. Consider the experiment of randomly selecting one student from a course of statistics and measuring two variables: X = "number of hours solving exercises during the previous week" and Y = "number of hours solving home assignments during the previous week". The joint probability distribution of X and Y is given in the following table.

		Y		
		0	8	16
	5	0.2	0.1	0.1
X	8	0.1	0.1	0.1
	18	0.2	0.1	0

   Which of the following is **correct**?

   (a) X and Y are dependent because their expectations are not equal;

   (b) $\boxed{X \text{ and } Y \text{ are dependent because their covariance } \sigma_{XY} \text{ is not equal to } 0;}$

   (c) X and Y are independent because $P(X = 5, Y = 0) = P(X = 5)P(Y = 0)$;

   (d) X and Y are independent because one of the joint probabilities is equal to 0;

   (e) it is not possible to establish if X and Y are independent with the information provided.

6. 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. What is the probability that the total imports during one month exceed 100 sacks?

   (a) 0.00;

   (b) 0.08;

   (c) $\boxed{0.15;}$

   (d) 0.35;

   (e) 0.41.

7. 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.

8. 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. 45 out of the 100 voters in the sample claim that they will vote for Mrs. A. With a significance level of 1%, which of the following is **correct**. (**Hint:** Use the alternative $P > 0.3$):

   (a) the critical value is 2.33 and the test statistic is 60.61, 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 rejected;]
   (c) the critical value is 2.36 and the test statistic is 3.02, therefore the null hypothesis is not 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;

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 B 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. The following table shows the scores of the eight students in a course of statistics in the final exam (variable *exam*) and the score in a previous home assignment (variable *assignment*):

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

Fitting a regression that explains the score in the exam in terms of the score in the assignment yields an intercept $b_0 = -25.2$ and a slope $b_1 = 1.5$. The *sum of squares error* —SSE— is:

   (a) 0;
   (b) 656;
   (c) [✓ 1594;]
   (d) 1881;
   (e) 2508;

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) $\boxed{52.9;}$

    (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) $\boxed{18.51;}$

    (e) 389.92.

Part one. Multiple choice

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

2. See Section 5.1 in Newbold et. al or Section 5 in the lecture notes.

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

4. The null hypothesis in a test of independence is that both variables are independent. The alternative is that they are dependent. As p-value $< \alpha$, the null hypothesis is rejected.

5. It is true that $\mu_X \neq \mu_Y$, but that is neither a necessary nor a sufficient condition for X and Y being dependent, therefore (a) is incorrect. It is true that $P(X = 5, Y = 0) = P(X = 5)P(Y = 0)$, but that is not a sufficient condition for X and Y being independent, therefore (c) is incorrect. It is true that one of the joint probabilities is equal to 0, but that is neither a necessary nor a sufficient condition for X and Y being independent, therefore (d) is incorrect. It is true that σ_{XY} is not equal to 0, in addition, this is a necessary and sufficient condition for X and Y being dependent, therefore (b) is correct.

6. 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 $W = X + Y =$ "Total sacks imported during one month". We have $W \sim N(\mu_W\,,\sigma_W^2)$ with $\mu_W = \mu_X + \mu_Y = 95$ and $\sigma_W^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} = 23$. Then $\sigma_W = 4.8$. This yields

$$P(W > 100) = 1 - P(Z < 1.04) = 0.15.$$

7. 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

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

8. 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.

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. Let x_i and y_i be, respectively, the assignment's score and the exam's score associated to the ith student. The fitted value and the residual associated to the first student are

$$\hat{y}_1 = -25.2 + 1.5 \cdot 42 = 37.8 \qquad \text{and} \qquad e_i = y_1 - \hat{y}_1 = 38 - 37.8 = 0.02.$$

   The remaining fitted values and residuals are found in an analogous way. They are shown in the following table.

x	42	48	50	50	51	55	59	67
y	38	43	57	33	81	50	48	84
$\hat{y}$	37.8	46.8	49.8	49.8	51.3	57.3	63.3	75.3
e_i	0.2	-3.8	7.2	-16.8	29.7	-7.3	-15.3	8.7

   The sum of squares error is then

$$SSE = \sum_s (y_i - \hat{y}_i)^2 = \sum_s e_i^2 = 0.2^2 + (-3.8)^2 + \cdots + 8.7^2 = 1594.$$

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

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

Part two. Complete solution

A Swedish sociologist is conducting an experiment about discrimination. She wants to determine if the name of the applicant influences the likelihood of being called to an interview. To this end, she sends out job applications to 100 advertised entry–level jobs in Sweden using fictive Swedish-sounding names (like Karl Andersson). She also sends out job applications to 100 advertised entry–level jobs in Sweden using fictive foreign-sounding first names (like Shady Gamhour). It turns out that 40 out of the 100 applications with Swedish sounding names are called to an interview; whereas 21 out of the 100 applications with foreign sounding names are called to an interview.

13. Considering the samples of applications with Swedish–sounding names and foreign–sounding names as two independent samples from two populations, test the null hypothesis that the proportion of applicants called to an interview is the same in both populations.

(a) **State the hypothesis of interest. (4p.)**

$$\text{Let } X_i = \begin{cases} 1 & \text{if the } i\text{th Swedish–sounding application is called to an interview} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } Y_i = \begin{cases} 1 & \text{if the } i\text{th foreign–sounding application is called to an interview} \\ 0 & \text{otherwise} \end{cases}$$

Let also P_x and P_y be the expectation of X_i and Y_i, respectively ($i = 1, \cdots, 100$). The hypothesis of interest is

$$H_0 : P_x = P_y \qquad \text{vs.} \qquad H_1 : P_x \neq P_y.$$

Or, equivalently, letting $\mu_D = P_x - P_y$,

$$H_0 : \mu_D = 0 \qquad \text{vs.} \qquad H_1 : \mu_D \neq 0.$$

(b) **Compute the test statistic and the critical value (using a significance level of 1%). (8p.)**

We have $\bar{d}_s = \hat{p}_x - \hat{p}_y = 40/100 - 21/100 = 0.19$ and

$$\hat{\sigma}^2_{\bar{D}} = \frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y} = \frac{0.4(1-0.4)}{100} + \frac{0.21(1-0.21)}{100} = 0.004059.$$

Then the test statistic is

$$t_{obs} = \frac{\bar{d}_s - \mu_0}{\hat{\sigma}_{\bar{D}}} = \frac{0.19 - 0}{0.004059^{0.5}} = 2.982.$$

and the critical value is

$$t_{n_x+n_y-2,\alpha/2} = t_{198,0.005} = 2.601.$$

(c) **What is the conclusion regarding the hypothesis? (4p.)**

As $2.982 = |t_{obs}| > t_{n_x+n_y-2,\alpha/2} = 2.601$, the null hypothesis is rejected and we conclude that the proportion of applicants with Swedish–sounding names and foreign–sounding names that are called to an interview differ. In other words, the name matters.

14. Let X = "first name's origin" (Swedish or foreign) and Y = "the applicant is called to an interview" (Yes or No). Test the hypothesis that X and Y are independent.

   (a) **Summarize the provided information in a 2×2 contingency table. (4p.)**

		Y		
		Yes	No	Total
X	Swedish	40	60	100
	Foreign	21	79	100
Total		61	139	200

   (b) **State the hypothesis of interest. (4p.)**

   H_0 : X and Y are independent vs. H_1 : X and Y are dependent

   (c) **Compute the test statistic and the critical value (using a significance level of 1%). (8p.)**

   First, let us find the expectations $E_{ij} = R_i C_j / n$:

		Y	
		Yes	No
X	Swedish	30.5	69.5
	Foreign	30.5	69.5

   Then, the test statistic is

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 8.515.$$

   The critical value is $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{1,0.01} = 6.635$.

   (d) **What is the conclusion regarding the hypothesis? (4p.)**

   As $8.515 = \chi^2_{obs} > \chi^2_{(r-1)(c-1),\alpha} = 6.635$, the null hypothesis is rejected and we conclude that being called to an interview depends on the name of the applicant. In other words, the name matters.

15. **Are your conclusions in 13. and 14. consistent with each other? (Yes or No.) Explain why they should be or they should not be consistent. (4p.)**

   In Exercises 13. and 14. two different methods were used to analyze the same data with the same purpose: we want to determine if the name of the applicant influences the likelihood of being called to an interview. The conclusions are consistent: both methods indicate that the name matters.

   This is not a coincidence: it can be shown that the chi-square test on a 2×2 contingency table is (asymptotically) equivalent to the test of hypothesis for the difference between two proportions.