BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2024–08–20

Part one. Multiple choice

1. Let us consider a random experiment with sample space given by the months of the year, i.e. $S = \{jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec\}$. Let $E_1 = \{jan, feb, mar\}$, $E_2 = \{apr, may, jun, jul\}$, $E_3 = \{jun, jul, aug, sep\}$ and $E_4 = \{oct, nov, dec\}$. Which of the following is **correct**:

(a) E_1, E_2, E_3 and E_4 are disjoint;

(b) E_1, E_2, E_3 and E_4 are a partition of S;

(c) **E_1, E_2, E_3 and E_4 are collectively exhaustive;**

(d) $E_1 \cup E_2$ is the complement of $E_3 \cup E_4$;

(e) $E_1 \cap E_2$ is the complement of $E_3 \cap E_4$.

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

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

(b) it is measured in the same units as the random variable itself;

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

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

(e) **it indicates how spread are the outcomes of the random variable around its expectation.**

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

(a) the significance level is the probability of making a type I error;

(b) the type II error is the probability of not rejecting the null hypothesis when it is false;

(c) **the test is constructed by assuming that the alternative hypothesis is true;**

(d) if the p-value is smaller than the significance level, the null hypothesis is rejected;

(e) the type I error is the probability of rejecting the null hypothesis when it is true.

4. According to the latest census which took place in 2018, 41% of a country's households are of size one, 28% are of size two, 10% are of size three, 16% are of size four and the remaining 5% have a larger size. In order to test if the distribution has changed, a researcher draws a random sample of 100 households and measures their size. Using a statistical software, the researcher carries out a test of goodness–of–fit and obtains a p-value of 0.8185. Considering a significance level of 5%, which of the following is **correct** regarding the test that has been implemented:

   (a) the average size of the households has increased;
   (b) the average size of the households has decreased;
   (c) there is strong evidence for concluding that the distribution has changed;
   (d) there is no evidence for concluding that the distribution has changed;
   (e) the information provided is not enough for making a conclusion.

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. 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 ten customers, exactly one will rent a suv?

   (a) 0.00;
   (b) 0.10;
   (c) 0.40;
   (d) 0.75;
   (e) 1.00;

7. 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 spends more than 2000 on clothing plus course literature?

   (a) 0;
   (b) 0.0967;
   (c) 0.5;
   (d) 0.25;
   (e) 1.

8. 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;

9. 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) $(494.4\,,499.8)$.

10. 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 2.36 and the test statistic is 3.02, therefore the null hypothesis is rejected;
   (b) the critical value is 3.02 and the test statistic is 2.36, 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 not rejected;
   (e) the critical value is 2.33 and the test statistic is 60.61, therefore the null hypothesis is rejected.

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

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

The teacher of the course wants to explain the score in the final exam in terms of the score in the home assignment through a linear regression of the form:

$$exam_i = \beta_0 + \beta_1\, assignment_i + \epsilon_i \qquad \text{for } i = 1, \cdots, 8$$

where $\epsilon_1, \cdots, \epsilon_8$ are a random sample from a normal distribution with expectation $\mu_\epsilon = 0$ and variance σ^2_ϵ. Fitting the desired regression yields an intercept $b_0 = -25.2$ and a slope $b_1 = 1.5$. The model variance σ^2_ϵ is estimated as:

(a) 0;

(b) 109.3;

(c) 265.7;

(d) 313.5;

(e) 418.

12. Table 1 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

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. See Chapter 3 in Newbold et. al or in the lecture notes.
2. See Section 4.3 in Newbold et al. or 4.2 in the lecture notes.
3. See Section 9.1 in Newbold et al. or Section 9 in the lecture notes.

4. Let P_1 be the current proportion of households of size one, P_2 the proportion of size two, P_3 the proportion of size three, P_4 the proportion of size four and P_+ the proportion of households of a larger size. The hypothesis in this goodness–of–fit test is:

$$H_0 : P_1 = 0.41, \quad P_2 = 0.28, \quad P_3 = 0.10, \quad P_4 = 0.16, \quad P_+ = 0.05$$

vs.

$$H_1 : \text{At least one of the probabilities } P_i \text{ is not as stated}$$

As the p-value is larger than the significance level, the null hypothesis is not rejected.

5. The mode is 20000, as this is the most frequently occurring value.

6. Let X = "number of customers, out of the next ten, that rent a suv. We have $X \sim Bin(10\,,0.1)$, therefore

$$P(X=1) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{10}{1} 0.1^1 (1-0.1)^{10-1} = 0.39.$$

7. 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

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

8. 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

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

9. We have

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

The confidence interval is then

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

10. We have

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

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.

11. 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

$$\hat{y}_1 = -25.2 + 1.5 \cdot 42 = 37.8 \qquad \text{and} \qquad e_1 = y_1 - \hat{y}_1 = 38 - 37.8 = 0.02.$$

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	0.2	-3.8	7.2	-16.8	29.7	-7.3	-15.3	8.7

The estimated model variance is then

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n-2}\sum_s e_i^2 = \frac{1}{8-2}\left(0.2^2 + (-3.8)^2 + \cdots + 8.7^2\right) = 265.7.$$

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

13. Table 2 shows the number of passengers by class and survival status (Alive or Dead) after the sinking of the Titanic on April 14, 1912. We want to investigate whether class and survival status are independent or not.

		Survival status	
		Alive	Dead
	First class	201	123
Class	Second class	119	166
	Third	180	530
	Crew	212	677

Table 2: Passengers on board the Titanic by class and survival status

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

$$H_0 : \text{Class and survival status are independent}$$
$$\text{vs.}$$
$$H_1 : \text{Class and survival status are dependent}$$

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

Let us add the marginals to the table:

		Survival status		Total
		Alive	Dead	Total
	First class	201	123	324
Class	Second class	119	166	285
	Third	180	530	710
	Crew	212	677	889
	Total	712	1496	2208

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

		Survival status	
		Alive	Dead
	First class	104.5	219.5
Class	Second class	91.9	193.1
	Third	228.9	481.1
	Crew	286.7	602.3

The test statistic is

$$\chi^2_{obs} = \sum_{i=1}^{4} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 187.6.$$

The critical value is $\chi^2_{(r-1)(c-1),\alpha} = 7.815$.

(c) **What is the conclusion regarding the hypothesis?**

As $187.6 = \chi^2_{obs} > \chi^2_{(r-1)(c-1),\alpha} = 7.815$ the null hypothesis is rejected, which means that the chances of surviving the wreck were associated to the class of the ticket.

14. A multiple linear regression that explains "sleeping time" (variable *sleep*, in hours per week) in terms of "working time" (variable *work*, in hours per week) and a dummy variable indicating whether the individual has kids younger than three

$$young\_kid = \begin{cases} 1 & \text{if the individual has kids younger than 3} \\ 0 & \text{otherwise} \end{cases}$$

has been fitted and the Excel output is shown in Figure 1.

Regression Statistics	
Multiple R	0.2258
R Square	0.0510
Adjusted R Square	-0.2202
Standard Error	9.082
Observations	10

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	31.0	15.51	0.1881	0.8326
Residual	7	577.4	82.48		
Total	9	608.4			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57.28	7.40	7.74	0.0001	39.78	74.78
work	-0.1226	0.2028	-0.60	0.5647	-0.6022	0.3570
young_kid	1.01	7.19	0.14	0.8922	-16.00	18.02

Figure 1: Excel output for the regression in Exercise 14

(a) **i. How many individuals were included in the analysis?** $n = 10$;
ii. What is the value of the coefficient of determination? $R^2 = 0.0510$;
iii. What is the value of the sum of squares error? $SSE = 577.4$;
iv. What is the 95% confidence interval for the coefficient associated to *work*? $(-0.6022\,, 0.3570)$;
v. What is the p-value for the test that all coefficients in the model are equal to zero? 0.8326.

(b) **Predict the number of hours sleeping for a person working twenty hours per week who has two kids younger than three.**

$$\widehat{sleep} = 57.28 - 0.1226 \times 20 + 1.01 \times 1 = 55.8.$$

(c) **Using a significance level of 5%, is the hypothesis that all coefficients are equal to zero rejected or not? What does that mean in this particular problem?**
As $0.8326 = p$–value $> \alpha = 0.05$, the null hypothesis is not rejected. This means that the combined effect of having young kids and the number of hours working does not affect significantly the number of hours sleeping for the ten individuals being analyzed.

(d) **Is the following sentence correct or incorrect?: "a person who works one more hour per week than another one, sleeps 0.12 hours less". (Justify your answer.)**

It is incorrect for at least three reasons. First, it does not take into account the conditional interpretation of the (estimated) coefficient. Second, it makes a deterministic interpretation of the coefficient, not taking into account the random nature of the problem. A third reason is that as the coefficient is not significant it does not even makes sense to interpret its value.