Stockholm University

BASIC STATISTICS FOR ECONOMISTS, STE101. EXAM SOLUTIONS

Department of statistics
Edgar Bueno
2024–02–23

Part one. Multiple choice

1. Which of the following is **not** a characteristic of an estimator:

   (a) variance,

   (b) expectation,

   (c) mean square error,

   (d) bias,

   (e) reproducibility.

2. A real estate agent has estimated the regression that explains the closing price (variable *price*, in SEK) of the housing units in a region of interest with respect to their size (variable *size*, in m^2). The fitted regression line is

$$\widehat{value} = 2\,500\,000 + 5000\, size.$$

   Which of the following is **not correct**:

   (a) the slope of the fitted regression is 5000;

   (b) the intercept of the fitted regression is 2 500 000;

   (c) the closing price of a housing unit of size $100m^2$ is expected to be around 3 000 000;

   (d) if housing unit A has one more square meter than housing unit B, we expect the closing price of A to be around 5000 SEK higher than the closing price of B;

   (e) the mean closing price of the housing units in the region of interest is 2 500 000.

3. 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 a partition of S;

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

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

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

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

4. 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 spread of the outcomes of the random variable around its expectation;

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

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

   (e) it is only defined for continuous random variables, not for discrete random variables.

5. Let $X_1, \cdots, X_n$ be a random sample from a chi-square distribution with expectation μ_X and variance σ_X^2. Let also $\bar{X} = \sum_{i=1}^{n} X_i/n$ be the sample mean. Which of the following sentences is **correct**:

   (a) the expectation of $\bar{X}$ coincides with the expectation of X_1, i.e. $\mu_{\bar{X}} = \mu_X$;

   (b) the variance of $\bar{X}$ coincides with the variance of X_1, i.e. $\sigma_{\bar{X}}^2 = \sigma_X^2$;

   (c) the distribution of $\bar{X}$ coincides with the distribution of X_1, i.e. $\bar{X}$ follows a chi-square distribution with expectation μ_X and variance σ_X^2;

   (d) the probability that $\bar{X}$ is smaller than a given value x coincides with the probability that X_1 is smaller than x, i.e. $P\left(\bar{X} < x\right) = P\left(X_1 < x\right)$;

   (e) none of the above.

6. A researcher has asked the thirteen married women in a small community about the dowry they had to pay to the groom's family when they got married. The dowry values (in USD) are

   20000 3000 10000 20000 13000 0 31000

   20000 63000 8000 3000 12000 4000

   What is the **coefficient of variation**, CV_x, of the dowry?

   (a) 1.01;

   (b) 16041;

   (c) 16159;

   (d) 32082;

   (e) 257 300 000.

7. 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. As they do not know each other, their grades can be considered to be independent. What is the probability that all four students pass the exam?

   (a) 0;

   (b) 0.112;

   (c) 0.6;

   (d) 1;

   (e) 2.4.

8. 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;

9. The amount of money spent on clothing by students at 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 between 2200 and 2300 on clothing plus course literature?

   (a) 0;

   (b) 0.0040;

   (c) 0.0214;

   (d) 0.0872;

   (e) 0.0967.

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. 37 out of the 100 voters in the sample claim that they will vote for Mrs. A. The value of the 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;

11. Fitting a regression that explains the score of students in the final exam of a course in statistics (variable *exam*) in terms of the score in a previous home assignment (variable *assignment*), yields an intercept $b_0 = -25.2$ and a slope $b_1 = 1.5$. The following table shows the scores of the eight students in the course:

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

    The *sum of squares error* —SSE— is:

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

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

    Having a significance level $\alpha = 0.05$, what is the critical value:

    (a) 1.960;
    (b) 5.321;
    (c) 7.815;
    (d) 10.648;
    (e) 145.461;

Part one. Multiple choice

1. See Section 7.1 in Newbold et al. or 6.2 in the lecture notes.
2. See Chapter 11 in Newbold et al. or Section 10.1 in the lecture notes.
3. See Chapter 3 in Newbold et. al or in the lecture notes.
4. See Section 4.3 in Newbold et al. or 4.2 in the lecture notes.
5. See Section 6.2 in Newbold et al. or Chapter 7 in the lecture notes.
6. We have $\sum_U x_i = 20000+3000+\cdots+4000 = 207\,000$ and $\sum_U x_i^2 = 20000^2+3000^2+\cdots+4000^2 =$

$6.642 \cdot 10^9$, then $\bar{x} = \sum_U x_i/N = 207\,000/13 = 15923.08$ and

$$S_x^2 = \frac{1}{N-1}\left(\sum_U x_i^2 - \frac{\left(\sum_U x_i\right)^2}{N}\right) = \frac{1}{13-1}\left(6.642 \cdot 10^9 - \frac{(207\,000)^2}{13}\right) = 278\,743\,590.$$

Therefore

$$CV_x = \frac{S_x}{\bar{x}} = \frac{278\,743\,590^{0.5}}{15923.08} = 1.05.$$

7. Let A, B, C and D be, respectively, the probabilities that students A, B, C and D pass the exam. By independence, we have

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B) \cdot P(C) \cdot P(D) = 0.8 \cdot 0.7 \cdot 0.5 \cdot 0.4 = 0.112.$$

8. 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

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

9. 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

$$P\left(2200 < W < 2300\right) = P\left(2 < Z < 3\right) = 0.9987 - 0.9772 = 0.0214.$$

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. 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 sum of squares error is then

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

12. $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{3,0.05} = 7.815$.

Part two. Complete solution

13. (a) According to the manufacturer, when fully charged, the battery of the IRL15 calculator has a life that can be adequately modeled by a normal distribution with expectation of 25 hours and variance of 100 hours. **What is the probability that one battery lasts less than five hours?** (5p.)

Let $X =$ "life of the battery". We have that $X \sim N(\mu_X\,, \sigma^2_X)$ with $\mu_X = 25$ and $\sigma^2_X = 100$. Then

$$P(X < 5) = P(Z < -2) = 0.02275,$$

where Z denotes the standard normal distribution.

(b) All 200 students taking a five-hour exam in statistics are going to use the IRL15 calculator with batteries fully charged. Assuming that batteries are independent of each other, **what is the probability that more than nine calculators run out of battery?** (**Note:** If you did not solve exercise 13a you can use a probability of 0.03.) (6p.)

Let $Y =$ "number of batteries that run out of battery". We have that $Y \sim Bin(n\,, P)$ with $n = 200$ and $P = 0.02275$. As n is large, the approximation $Y \overset{\text{approx.}}{\sim} = N(nP\,, nP(1 - P)) = N(4.55\,, 4.4465)$ holds. Then

$$P(Y > 9) = 1 - P(Y \leq 9) \approx 1 - P(Z \leq 2.11) = 0.01741,$$

where, again, Z denotes the standard normal distribution.

(c) It turns out that 3 out of the 200 calculators run out of battery during the exam. Assuming that the sample is indeed a random sample, **construct a 95% confidence interval for the proportion of batteries that last less than 5 hours.** (6p.)

We have $\hat{p} = 3/200 = 0.015$, $n = 200$ and $t_{n-1,\alpha/2} = t_{199,0.025} = 1.972$, then

$$\hat{p} \pm t_{n-1,\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.015 \pm 1.972\sqrt{\frac{0.015(1-0.015)}{200}} = 0.015 \pm 0.01695 = (-0.2\%\,, 3.2\%).$$

(d) **Is the answer in (c) consistent with the manufacturer's claim in (a)? Justify.** (3p.)

Yes. If the manufacturer's claim is true, we expect around 2.3% of the batteries to last less than five hours. This value belongs to the interval $(-0.2\%\,, 3.2\%)$. Therefore the evidence provided by the data in (c) is consistent with the "hypothesis" in (a).

14. The Air Quality Index (AQI) is a measure of the quality of the air. It is measured on a scale from 0 to 500 in which lower values indicate a higher air quality. For instance, values above 100 are considered unhealthy whereas values below 50 are considered safe.

    In order to improve the air quality of a city, the Mayor is considering to implement free public transport. Before making a decision, the Mayor wants to evaluate if such a policy will indeed yield a significant improvement in air quality. To this end, free public transport is offered during one particular day. The AQI is measured at 16 different locations of the city on two occassions: on a regular day and during the free public transportation day. The measurements are shown in Table 1.

Location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Regular day	59	46	87	78	80	21	61	50	53	46	53	53	25	57	68	53
Free transport day	51	35	74	64	72	6	52	38	45	34	48	46	17	41	59	41

Table 1: AQI measurements in 16 locations during a regular day and during the free transport day

Test the alternative that free transport improves air quality, in other words, test the alternative that AQI is reduced during the free transport day.

(a) **State the hypothesis of interest.** (5p.)

Let X_i be the AQI on the ith location on the regular day and Y_i be the AQI on the ith location on the free transport day ($i = 1, \cdots, 16$). The hypothesis of interest is

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad \mu_X > \mu_Y$$

Or, equivalently, letting $\mu_D = \mu_X - \mu_Y$,

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

(b) **Compute the test statistic and the critical value (using a significance level of 1% and assuming that AQI measurements are adequately described by a normal distribution).** (10p.)

First we need to calculate the differences $d_i = x_i - y_i$. They are shown in the Table 2.

Location	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Regular day, x_i	59	46	87	78	80	21	61	50	53	46	53	53	25	57	68	53
Free transport day, y_i	51	35	74	64	72	6	52	38	45	34	48	46	17	41	59	41
Difference, d_i	8	11	13	14	8	15	9	12	8	12	5	7	8	16	9	12

Table 2: Differences of AQI in 16 locations during a regular day and during the free transport day

We have $\bar{d} = 10.44$, $S_d^2 = 9.863$, $\hat{\sigma}^2_{\overline{D}} = S_d^2/n = 9.863/16 = 0.6164$. Therefore the test statistic is

$$t_{obs} = \frac{\bar{d} - \mu_0}{\hat{\sigma}_{\overline{D}}} = \frac{10.44 - 0}{0.7851} = 13.29.$$

The critical value is $t_{n-1,\alpha} = t_{16-1,0.01} = 2.602$.

(c) **What is the conclusion regarding the hypothesis?** (5p.)

As $13.29 = t_{obs} > t_{n-1,\alpha} = 2.602$ the null hypothesis is rejected. In other words, the evidence supports the hypothesis that AQI significantly decreased during the free transport day compared to a regular day.