STATISTICS APPLIED TO BUSINESS ADMINISTRATION (BAD) - Second Year   Academic Year 2016-17
STATISTICS APPLIED TO MARKETING (MD) - Second Year
STATISTICS APPLIED TO BUSINESS ADMINISTRATION (DD) - Third Year
First Call. June 9, 2017

INSTRUCTIONS

1. The exam contains multiple choice questions that must be answered in the orange code sheet we have provided you with, together with three exercises that must be answered in detail in separate sheets of paper.
2. To select an answer, all you need to do is mark in the orange code sheet, **filling the rectangle over which the selected answer is located appropriately**. Please make sure you know the answer you wish to mark before doing it. Even though you can always erase your mark if you have used a pencil (number 2 or similar), any mark that has not been completely erased could be read by the machine. Therefore, we advice you to first mark your selected answers in the exam and to use only the last ten minutes or so from the time assigned to the multiple choice questions-part of the exam to copy them into the code sheet.
3. In the multiple choice questions-part of the exam there is always **only one correct answer** for every question. Every question correctly answered is worth 1 point, while each question incorrectly answered will penalize your grade by 0.2 points. Questions that have not been answered do not penalize your grade in any form.
4. Each one of the exercises, A, B and C, must be answered in a separate sheet of paper. We will collect the different parts of the exam at the indicated times and in this order: first, we will collect the code sheet for the multiple choice questions part of the exam and, then, and in this order, exercises A, B and C will be collected.
5. The exam has six numbered sheets, going from 0.1 to 0.6. Please make sure that you have all sheets and contact your professor if this is not the case. There are different exam types. This exam is of type 0. Mark a 0 in the column labelled with I in your code sheet.
6. The maximum final grade for each of the parts of the exam (i.e., the multiple choice questions part and the exercises part) is 30 points. If you have participated in the on-going evaluation process, you will need to obtain 12 points in each part of the exam to pass it. Otherwise, 15 points points in each part of the exam are required to pass it.
7. Please fill in your personal information in the appropriate places both in the code sheet and in the sheets provided for the exercises.

MULTIPLE CHOICE QUESTIONS (Time: 1 hour and 45 minutes)

1. FREE-QUESTION. The capital of Spain is:

   (A) Paris (B) Sebastopol (C) Madrid (D) London (E) Pekin

Questions 2 to 5 refer to the following exercise:

In a given high school, 15 students have been randomly selected. It is known that the number of students that will pass all of the courses they have registered for follows a binomial distribution, $Z \in b(p, n)$, with characteristic function $\psi_Z(u) = (0.90 + 0.10e^{iu})^{15}$. We assume independence between the different students in that high school.

2. The probability that exactly 2 students from that high school will pass all of the courses they have registered for is:

   (A) 0 (B) 0.2669 (C) 0.5490 (D) 0.8159 (E) 1

3. The probability that at least 7 students from that high school will pass all of the courses they have registered for is:

   (A) 1 (B) 0.0003 (C) 0 (D) 0.0022 (E) 0.9997

4. If we now have that 60 students have been randomly selected, the approximate probability that at most 6 of them will pass all of the courses they have registered for is:

   (A) 0.6063 (B) 0.7440 (C) 0.5543 (D) 0.3937 (E) 0.4457

5. If we finally have that 200 students have been randomly selected, the approximate probability that less than 25 of them will pass all of the courses they have registered for is :

   (A) 0.0968 (B) 0.3445 (C) 0.9032 (D) 0.1446 (E) 0.8554

Questions 6 to 8 refer to the following exercise:

The number of vehicles that arrives at a toll paying site in a given highway follows a Poisson distribution with mean equal to 2 vehicles per minute. We assume independence between the vehicles arriving at different minutes.

6. The probability that, in a given minute, no vehicle arrives at the highway toll paying site is:

   (A) 0.1353 (B) 0.8647 (C) 0.3025 (D) 0.5940 (E) 0.4060

7. The probability that, in a given minute, at least one vehicle arrives at the toll paying site is:

   (A) 0.1353 (B) 0.3025 (C) 0.5940 (D) 0.8647 (E) 0.4060

8. The approximate probability that, in a given hour, more than 130 vehicles arrive at the toll paying site is:

   (A) 0.8315 (B) 0.9128 (C) 0.1365 (D) 0.0872 (E) 0.1685

Questions 9 and 10 refer to the following exercise:

Let X be a r.v. with a Poisson distribution, and such that $P(X = 1) = P(X = 2)$.

9. The probability that the r.v. X takes on values smaller than 2 or larger than 3 is:

(A) 0.73 (B) 0.82 (C) 0.55 (D) 0.00 (E) 1.00

10. $P(2 < X < 6)$ is:

(A) 0.3068 (B) 0.1383 (C) 0.1263 (D) 0.3188 (E) 0.0583

Questions 11 and 12 refer to the following exercise:

Let X be a r.v. such that $X \in \gamma(a, r)$, and that its mean and variance are equal to 2 and 4, respectively.

11. The probability that the r.v. X takes on values smaller than 4.61 is:

(A) 0.95 (B) 0.05 (C) 0.90 (D) 0.10 (E) 0.975

12. If we define the r.v. $Y = \frac{X}{4}$, the distribution of the r.v. Y is:

(A) χ^2_2 (B) $\exp(0.5)$ (C) χ^2_4 (D) $\gamma(4, 1)$ (E) $\exp(2)$

Questions 13 to 15 refer to the following exercise:

Let X, Y, Z and V be independent r.v. so that their probability density functions are as follows: $X \in N(0, 1)$, $Y \in N(-1, 4)$, $Z \in \gamma(\frac{1}{2}, 4)$ and $V \in \gamma(\frac{1}{2}, 1)$.

13. The probability that the r.v. $W_1 = Z + V$ takes on values between 3.94 and 4.87 is:

(A) 0.95 (B) 0.05 (C) 0.90 (D) 0.10 (E) 0.025

14. The probability that the r.v. $W_2 = \frac{10\left[4X^2+(Y+1)^2\right]}{8W_1}$ takes on values smaller than or equal to 4.10 is:

(A) 0.90 (B) 0.10 (C) 0.01 (D) 0.05 (E) 0.95

15. If we define the r.v. $W_3 = \frac{2X}{\sqrt{(Y+1)^2}}$, then the value of k such that $P(W_3 \geq k) = 0.90$ is:

(A) 3.08 (B) -3.08 (C) 6.31 (D) -6.31 (E) 2.92

Questions 16 and 17 refer to the following exercise:

Let X be a discrete r.v. with probability mass function given by:

$$P(X=0) = \frac{\theta}{8} \qquad P(X=1) = \frac{\theta}{2} \qquad P(X=2) = 1 - \frac{3\theta}{4} \qquad P(X=3) = \frac{\theta}{8}$$

In order to estimate the parameter θ, a r.s. of size n, $X_1, X_2, \ldots, X_n$, has been taken.

16. The method of moments estimator of θ is:

(A) $\frac{8(2-\overline{X})}{5}$ (B) $\frac{2-\overline{X}}{5}$ (C) $\frac{2-5\overline{X}}{5}$ (D) $\frac{5}{8(2-\overline{X})}$ (E) $\overline{X}$

17. In order to be able to obtain an estimate of the parameter θ, a random sample of size $n = 20$ has been taken providing the following results: 3, 0, 3, 0, 3, 0, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2, 2, 2, 2. The maximum likelihood estimate of θ is equal to:

(A) 0.80 (B) 0.28 (C) 0.35 (D) 0.60 (E) 0.72

Questions 18 to 20 refer to the following exercise:

Let X be a r.v. having an exponential distribution with parameter $\frac{1}{\theta}$. We wish to estimate the parameter θ and, in order to do so, a r.s. of size n, $X_1, \ldots, X_n$, has been taken.

18. The maximum likelihood estimator of θ, $\hat{\theta}_{\text{ML}}$, will be:

(A) $\overline{X}$ (B) $n\overline{X}$ (C) $\frac{\overline{X}}{(1-\overline{X})}$ (D) $\frac{\overline{X}}{n}$ (E) $\frac{1}{\overline{X}}$

19. The bias for the maximum likelihood estimator of θ is:

(A) $\frac{1}{n}$ (B) $-\frac{\theta}{n}$ (C) 0 (D) $\frac{\theta}{n}$ (E) $\frac{1}{n^2}$

20. The mean square error for the maximum likelihood estimator of θ is:

(A) $\frac{\theta^2}{2n}$ (B) $\frac{\theta^2}{n}$ (C) $\frac{1}{2}$ (D) $\frac{1}{n}$ (E) θ^2

21. Let X be a r.v. with probability density function $f(x;\theta) = 4\ e^{-4(x-\theta)},\ x > \theta$. We wish to estimate the parameter θ and, in order to do so, a r.s. of size n, $X_1, \ldots, X_n$, has been taken. The maximum likelihood estimator of θ, $\hat{\theta}_{\text{ML}}$, will be:

(A) $4\overline{X}$ (B) $\overline{X}$ (C) $\max(X_1, \ldots, X_n)$ (D) $\frac{\overline{X}}{4}$ (E) $\min(X_1, \ldots, X_n)$

Questions 22 and 23 refer to the following exercise:

Let X be a r.v. with probability density function given by:

$$f(x;\theta) = (\theta + 1)x^{\theta}, \quad 0 < x < 1, \quad \theta > 0$$

Based on a r.s. of size $n = 1$, X_1, we wish to test the null hypothesis $H_0 : \theta = 2$ against the alternative hypothesis $H_1 : \theta = 1$, and reject the null hypothesis H_0 if $X_1 \leq 0.40$.

22. The probability of type I error for this test is:

(A) 0.16 (B) 0.042 (C) 0.936 (D) 0.064 (E) 0.021

23. The probability of type II error for this test is:

(A) 0.84 (B) 0.40 (C) 0.92 (D) 0.60 (E) 0.16

Questions 24 and 25 refer to the following exercise:

Before jumping, a parachutist has to decide if his/her parachute is or is not in good condition. If s/he believes it is in good condition, s/he will jump and, s/he will not do it otherwise. The following hypotheses are stated: H_0: the parachute is defective (i.e., it is not in good condition); H_1: The parachute is in good condition.

24. Which one of the following situations represents a type I error?

(A) The parachutist does not jump because s/he suspects that the parachute is defective. Another parachutist jumps and the parachutes opens up

(B) The parachutist does not jump because s/he believes that the parachute is defective. A careful inspection confirms that the parachute is not in good condition and has a problem

(C) The parachutist jumps and the parachute opens up

(D) The parachutist jumps and the parachute does not open up

(E) All false

25. Which one of the following situations represents the power?
    (A) The parachutist does not jump because s/he believes that the parachute is defective. A careful inspection confirms that the parachute is not in good condition and has a problem
    (B) The parachutist jumps and the parachute opens up
    (C) The parachutist does not jump because s/he suspects that the parachute is defective. Another parachutist jumps and the parachutes opens up
    (D) The parachutist jumps and the parachute does not open up
    (E) All false

Questions 26 to 28 refer to the following exercise:

An individual is interested in buying an iPhone7. Before doing so, s/he decides to ask for its price at 16 different stores, obtaining a mean sample price of 950 euros with a sample standard deviation of 80 euros. We assume normality.

26. At the 95% confidence level, we can state that the iPhone7's mean price is included in the interval:

    (A) (906.0, 994.0) (B) (917.2, 982.8) (C) (870.0, 1030.0)
    (D) (916.12, 983.88) (E) (913.85, 986.15)

27. At the 95% confidence level, we can state that the **standard deviation** of the iPhone7's price is included in the interval:

    (A) (46.55, 204.47) (B) (61.02, 127.90) (C) (64.0, 118.76)
    (D) (65.4, 94.6) (E) (56.8, 108.4)

28. The individual, independently, decides to ask for the iPhone's price in 14 stores in another city, obtaining a mean sample price of 850 euros with a sample standard deviation of 65 euros. The individual wishes to estimate the price difference between the two cities. We assume normality and that the price population variances for the two cities are unknown but equal. At the 95% confidence level, we can state that the iPhone7's mean price difference between the first city and the city in this item is included in the interval:

    (A) (48.08, 151.92) (B) (73.05, 126.95) (C) (62.45, 137.55)
    (D) (43.01, 156.99) (E) (52.74, 147.26)

Questions 29 and 30 refer to the following exercise:

We wish to estimate the mean family concert attendance expenses in a given city. It has been observed that, with regard to concert attendance expenses, there are two different types of families: families having only adult members and those having at least one child. There are 8000 families in the city, 6000 of them belonging to the first type of family, and 2000 belonging to the other type. In addition, it is also known that the population quasivariances for concerts attendance expenses for each of this type of families are 2000 and 1200, respectively. It has been decided to use a stratified random sampling of 800 families.

29. If we decide to use optimal allocation, the corresponding sample sizes for each one of the two strata will be:

    (A) 600 and 200 (B) 636 and 164 (C) 400 and 400 (D) 667 and 133 (E) 680 and 120

30. For these sample sizes, the estimated variance for the mean family concert attendance expenses will be:

    (A) 2.78 (B) 2.30 (C) 1.94 (D) 2.62 (E) 2.00

EXERCISES (Time: 60 minutes)

A. (10 points, 20 minutes)

We wish to know if the level of foreign language knowledge provided at a given high school is the one that would be considered desirable or acceptable. If this is the case, at most 10% of the students who have finished their studies would fail the exam provided for that foreign language. Given that it could be quite expensive to give this exam to all of the potential students, the following test procedure is proposed: 10 students are randomly selected and if two or more fail the exam, we conclude that the level of foreign language knowledge is not the one that would be considered desirable or acceptable.

i) What is the significance level for this test?

ii) If the proportion of students failing the exam in that high school were equal to 30%, what would the power for this test be?

iii) An alternative test procedure is proposed, in which 20 students are randomly selected and, if four or more fail the exam, we conclude that the level of foreign language knowledge is not the one that would be considered desirable or acceptable. Compute the significance level for this test. Compute the power for this test for a students' failing proportion of 30%. Is this alternative test better that the one mentioned above? Provide all relevant details to justify your response.

B. (10 points, 20 minutes)

It is believed that the weekly sales volume a given firm has in metric Tons (mT) follows a uniform distribution in the interval $(8, 12)$. In order to verify this claim, a r.s. of 100 weeks has been taken, providing the following results: in 20 of those weeks, sales were between 8 and 9 mT; in 31 of them, sales were between 9 and 10 mT; in 19 of them, sales were between 10 and 11 Tm, and, finally, in 30 of those weeks, sales were between 11 and 12 mT. At the 5% significance level, test the hypothesis that the weekly sales volume for this firm is the one claimed by the firm.

C. (10 points, 20 minutes)

Let X be a r.v. with probability mass function given by

$$P_X(x;\theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots$$

It is known that the mean and variance for the r.v. X are $\mathrm{E}(X) = \frac{1}{\theta}$ y $\mathrm{Var}(X) = \frac{(1-\theta)}{\theta^2}$. Let $X_1, \ldots, X_n$ be a r.s. from this r.v.

i) Obtain, **providing all relevant details**, the method of moments estimator of θ.

ii) Obtain, **providing all relevant details**, the maximum likelihood estimator of θ.

iii) If a r.s. of size $n = 10$ has been taken, providing the values $6, 4, 3, 7, 2, 1, 5, 4, 5, 9$, obtain a maximum likelihood estimate of θ.

SOLUTIONS TO MULTIPLE CHOICE QUESTIONS (exam type 0)

1: C	11: C	21: E
2: B	12: E	22: D
3: B	13: B	23: A
4: A	14: E	24: D
5: E	15: B	25: B
6: A	16: A	26: A
7: D	17: A	27: B
8: E	18: A	28: D
9: C	19: C	29: B
10: A	20: B	30: E

SOLUTIONS TO EXERCISES

Exercise A

We have a test for a binomial proportion, where, if p denotes the proportion of students failing the exam, we have to test the null hypothesis $\mathrm{H}_0 : p \leq 0.10$ against the alternative hypothesis $\mathrm{H}_1 : p > 0.10$.

i) In this case, we reject the null hypothesis if $Z = \sum_{i=1}^{10} X_i \geq 2$, where $Z \in b(p, 10)$. In this way, we will have that the significance level for this test is computed as:

$$\alpha = P[Z \geq 2 | Z \in b(0.10, 10)] = 1 - F_Z(1) = 1 - 0.7361 = 0.2639$$

ii) To compute the power for this test, we will have that:

$$\text{Power}(p = 0.30) = P[Z \geq 2 | Z \in b(0.30, 10)] = 1 - F_Z(1) = 1 - 0.1493 = 0.8507$$

iii) In this case, for this alternative test, we reject the null hypothesis if $Z = \sum_{i=1}^{20} X_i \geq 4$, where $Z \in b(p, 20)$. In this way, we will have that the significance level for this test is computed as:

$$\alpha = P[Z \geq 4 | Z \in b(0.10, 20)] = 1 - F_Z(3) = 1 - 0.8670 = 0.1330,$$

and the power for this alternative test will be:

$$\text{Power}(p = 0.30) = P[Z \geq 4 | Z \in b(0.30, 20)] = 1 - F_Z(3) = 1 - 0.1071 = 0.8929$$

Given that this alternative test improves both the significance level and the power when compared to the first test proposed above, we conclude that the alternaive test is a better test.

Exercise B

We have that the r.v. represents the weekly sales volume (in mT) for a given firm. Therefore, it is a **goodness of fit test to a completely specified distribution**, where we are testing the null hypothesis $\mathrm{H}_0 : X \in U(8,12)$ against the alternative hypothesis $\mathrm{H}_1 : X \notin U(8,12)$. The information we have to perform this test is given below:

Sales volume interval (mT)	(8,9)	(9,10)	(10,11)	(11,12)
Number of weeks	20	31	19	30

First of all, we have that the theoretical probabilities p_i for each one of the intervals in the table above will be:

$$p_1 = (9-8)/(12-8) = 0.25; \qquad p_2 = (10-9)/(12-8) = 0.25$$
$$p_3 = (11-10)(/12-8) = 0.25; \qquad p_4 = (12-11)/(12-8) = 0.25$$

Given that we have not estimated any parameter, we have that $h = 0$. In addition, as we have four possible sales volume intervals, $k = 4$. With this information and in order to perform the test, we build the following table:

Sales interval (mT)	n_i	p_i	np_i	$\frac{(n_i-np_i)^2}{np_i}$
$(8,9)$	20	0.25	25	1.0
$(9,10)$	31	0.25	25	1.44
$(10,11)$	19	0.25	25	1.44
$(11,12)$	30	0.25	25	1.0
	$n = 100$	1	$n = 100$	$z = 4.88$

Under the null hypothesis that the probability distribution for the firm's weekly sales volume is the correct one, the test statistic $\sum_i \frac{(n_i-np_i)^2}{np_i} \sim \chi^2_{k-h-1}$, where k is the number of possible intervals for the sales volume ($k = 4$) and h is the number of estimated parameters ($h = 0$).

The decision rule, at a 5% approximate significance level, will be to reject the null hypothesis if:

$$z > \chi^2_{k-h-1\,,\,0.05} = \chi^2_{3,0.05}$$

In this case:

$$z = 4.88 < 7.81 = \chi^2_{3,0.05}$$

so that, at a 5% approximate significance level, we do not reject the null hypothesis that the probability distribution the firm has for its weekly sales volume follows a uniform distribution in the interval $(8,12)$.

Exercise C

The probability mass function for the r.v. X is

$$P(x,\theta) = \theta(1-\theta)^{x-1}, \quad x = 1, 2, \ldots$$

In order to estimate the parameter θ, we have that the r.v. $X_1, \ldots, X_n$ are independent and identically distributed (I.I.D.), having the probability mass function provided above.

i) In order to be able to obtain the method of moments estimator of the parameter θ, we need to equate the first population moment $\alpha_1 = \mathrm{E}(X) = m$ to the first sample moment $a_1 = \overline{X}$. That is,

$$\alpha_1 = \mathrm{E}(X) = a_1 = \frac{\sum_{i=1}^n X_i}{n} = \overline{X}$$

Therefore, we will have that:

$$\alpha_1 = \mathrm{E}(X) = \frac{1}{\theta} = a_1 = \frac{\sum_{i=1}^n X_i}{n} = \overline{X} \Longrightarrow \hat{\theta}_{\text{MM}} = \frac{1}{\overline{X}}$$

ii) To obtain the likelihood function, we have that:

$$L(\vec{x},\theta) = P(x_1,\theta) \ \ldots \ P(x_n,\theta) = \left[\theta(1-\theta)^{x_1-1}\right] \cdots \left[\theta(1-\theta)^{x_n-1}\right] = \theta^n \left[1-\theta\right]^{\sum_{i=1}^n x_i - n}$$

In order to be able to obtain the maximum likelihood estimator of θ, we have to maximize the logarithm of the likelihood function. We now compute its natural logarithm to obtain:

$$\ln L(\vec{x},\theta) = n\ln(\theta) + \left[\sum_{i=1}^n x_i - n\right]\ln(1-\theta)$$

If we take derivatives with respect to θ and equate this to zero, we will have that:

$$\frac{\partial \ln L(\vec{x},\theta)}{\partial\theta} = \frac{n}{\theta} - \frac{\left[\sum_{i=1}^n x_i - n\right]}{(1-\theta)} = 0$$

so that,

$$\frac{n}{\theta} = \frac{\left[\sum_{i=1}^n x_i - n\right]}{(1-\theta)} \Longrightarrow n - n\theta = \theta\sum_{i=1}^n x_i - n\theta \Longrightarrow n = \theta\sum_{i=1}^n x_i \Longrightarrow \hat{\theta}_{\text{ML}} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X}}$$

iii) If a r.s. of size $n = 10$ has provided the values $6, 4, 3, 7, 2, 1, 5, 4, 5, 9$, a maximum likelihood estimate of θ will be:

$$\hat{\theta}_{\text{ML}} = \frac{n}{\sum_{i=1}^n x_i} = \frac{10}{(6+4+3+7+2+1+5+4+5+9)} = \frac{10}{46} = 0.2174$$