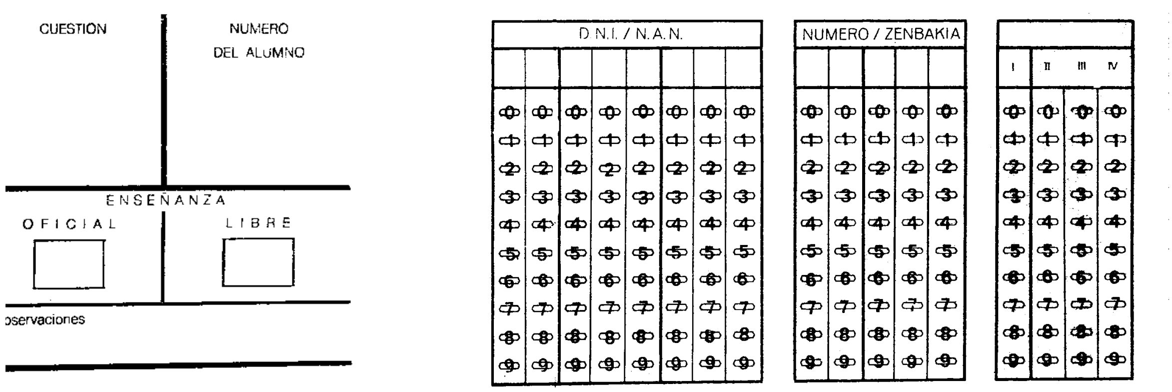
STATISTICS APPLIED TO BUSINESS ADMINISTRATION (BAD) - Second Year Academic Year 2016-17
STATISTICS APPLIED TO MARKETING (MD) - Second Year
STATISTICS APPLIED TO BUSINESS ADMINISTRATION (DD) - Third Year
Second Call. July 5, 2017

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. You will need to obtain 15 points in each part of the exam 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.

Example:

ID: $\underbrace{12345678}$ Exam type $\underbrace{0}$



MULTIPLE CHOICE QUESTIONS (Time: 1 hour and 30 minutes)

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

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

Questions 2 to 4 refer to the following exercise:

The probability that, in a given airport, a traveler entering the Duty Free store buys something is equal to 30%. We assume independence between the different travelers' purchases.

2. The probability that, out of 8 randomly selected travelers, 5 of them buy something at the Duty Free store is:

   (A) 0.0113 (B) 0 (C) 1 (D) 0.2541 (E) 0.0467

3. If 15 travelers are randomly selected, the probability that at least 10 of them buy something at the Duty Free store is:

   (A) 0.0300 (B) 0.0037 (C) 0.0007 (D) 0.9963 (E) 0.9993

4. There is a flight delay and its 180 passengers enter the Duty Free store. The mean number of buyers this store expects to have due to this flight delay is:

   (A) 10 (B) 126 (C) 37.8 (D) 54 (E) 180

5. The probability of having a defective component is equal to 5%. The approximate probability that in a pack of 180 randomly selected components less than 10 of them are defective is:

   (A) 0.4126 (B) 0.2940 (C) 0.5874 (D) 0.7060 (E) 0.8030

6. The number of days per academic year that a given high school closes due to bad weather follows a Poisson distribution with mean equal to 2.5. The probability that, during the upcoming academic year, this high school closes at least one day due to bad weather is:

   (A) 0.9179 (B) 0.3297 (C) 0.2052 (D) 0.6703 (E) 0.0821

7. Let X be a random variable having a Poisson distribution with variance equal to 25. The approximate probability that X takes on values larger than 25 is:

   (A) 0.492 (B) 0.608 (C) 0.250 (D) 0.540 (E) 0.460

8. The period of time between two consecutive phone calls to an emergency service is a random variable following an exponential distribution with mean equal to 2 minutes. The probability that the time between two consecutive phone calls is less than 1 minute is:

   (A) 0.6065 (B) 0.8647 (C) 0.1353 (D) 0.3935 (E) 0.6988

9. The service lifetime for a plasma TV **in years** is a random variable following an exponential distribution with mean equal to 10. The service lifetime for this plasma TV **in months** is a r.v. with distribution:

   (A) χ^2_{24} (B) $\gamma(12, 0.1)$ (C) $\exp(1.2)$ (D) $\exp(0.0083)$ (E) $\gamma(0.1, 12)$

10. Let $X \in \mathcal{F}_{5,10}$. The probability $P(X \leq 3.33)$ is equal to:

(A) 0.10 (B) 0.95 (C) 0.80 (D) 0.90 (E) 0.05

11. Let X_1 and X_2 be two independent r.v. with distributions $X_1 \in \gamma\left(\frac{1}{2}, 3\right)$ and $X_2 \in \chi^2_3$. If we define the r.v. $Y = X_1 + X_2$, Y is a r.v. with distribution:

(A) $\gamma(3,3)$ (B) χ^2_9 (C) χ^2_6 (D) $\gamma(\frac{3}{2},3)$ (E) $\gamma(\frac{1}{2},6)$

12. Let $X \in t_5$. The value of k such that $P(X \leq k) = 0.9$ is:

(A) -1.48 (B) -2.02 (C) 1.48 (D) 0.868 (E) 2.02

Questions 13 to 15 refer to the following exercise:

Let X be a discrete r.v. with probability mass function given by: $P(0) = \theta$, $P(1) = \frac{\theta}{2}$, $P(2) = 1 - \frac{3\theta}{2}$. In order to estimate the parameter θ, a r.s. of size n, $X_1, \ldots, X_n$, has been taken.

13. The method of moments estimator of θ is:

(A) $20 - 10\bar{X}$ (B) $\frac{4-2\bar{X}}{5}$ (C) $\frac{4-\bar{X}}{5}$ (D) $\bar{X}$ (E) $\frac{5}{4-2\bar{X}}$

14. The bias for the method of moments estimator of θ is:

(A) Positive (B) 0 (C) Negative (D) - (E) -

15. A r.s. of size $n = 8$ has been taken, rendering the following results: 0, 0, 0, 1, 1, 2, 2, 2. The maximum likelihood estimate of θ is:

(A) $\frac{12}{5}$ (B) $\frac{5}{12}$ (C) $\frac{4}{12}$ (D) $\frac{2}{5}$ (E) $\frac{3}{5}$

16. A r.s. is taken from a population having a uniform distribution on the interval $[\theta, 10]$, rendering the following results: 3, 4, 7.5 and 10. Under the maximum likelihood criterion, which one of the following distributions would we select as the one generating such a sample?

(A) $U[2,10]$ (B) $N(3,10)$ (C) $U[6,10]$ (D) $U[3,10]$ (E) $N(6,3)$

17. Let X be a r.v. having probability density function given by $f(x) = \theta^2\, x\, e^{-\theta\, x}$, $x > 0$. The maximum likelihood estimator of θ obtained from a r.s. of size n is:

(A) $\dfrac{2}{\overline{X}}$ (B) $\dfrac{2}{\sum_{i=1}^{n} X_i}$ (C) $\overline{X}$ (D) $\overline{X} - 2$ (E) $2\overline{X}$

Questions 18 and 19 refer to the following exercise:

Let θ be the unknown parameter of a given distribution. In order to estimate θ, a r.s. of size n is taken. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ, so that:

$$\mathrm{E}(\hat{\theta}_1) = \theta \qquad \mathrm{var}(\hat{\theta}_1) = \tfrac{2}{n} \qquad\qquad \mathrm{E}(\hat{\theta}_2) = \theta + \tfrac{2}{n} \qquad \mathrm{var}(\hat{\theta}_2) = \tfrac{1.5}{n}$$

18. The proposed estimator $\hat{\theta}_2$ is:

    (A) Biased and not consistent

    (B) Asymptotically biased and consistent

    (C) Unbiased and not consistent

    (D) Unbiased and consistent

    (E) Biased and consistent

19. In order to estimate the parameter θ from a r.s. of size $n = 6$, the estimator $\hat{\theta}_1$ is selected because:

    (A) - (B) $\mathrm{MSE}(\hat{\theta}_1) > \mathrm{MSE}(\hat{\theta}_2)$ (C) $\mathrm{MSE}(\hat{\theta}_1) = \mathrm{MSE}(\hat{\theta}_2)$
    (D) $\mathrm{MSE}(\hat{\theta}_1) < \mathrm{MSE}(\hat{\theta}_2)$ (E) -

Questions 20 and 21 refer to the following exercise:

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

$$f(x, \theta) = \theta x^{\theta - 1}, \qquad 0 < x < 1, \qquad \theta > 0$$

In order to test the null hypothesis $H_0 : \theta = 1$ against the alternative hypothesis $H_1 : \theta = 2$, a random sample of size $n = 1$ (i.e., X), is taken.

20. The most powerful critical region for this test is of the form:

    (A) $X \geq K$ (B) $X \leq K$ (C) $X \in (0, 1)$ (D) $X \in (K_1, K_2)$ (E) $X \in (K_1, K_2)^{\mathrm{c}}$

21. At the 10% significance level, the most powerful critical region for this test is:

    (A) $X \in (0.25, 0.75)^{\mathrm{c}}$ (B) $X \leq 0.10$ (C) $X \geq 0.9$ (D) $X \geq 0.10$ (E) –

Questions 22 to 24 refer to the following exercise:

An individual is interested in buying a telephone. Before doing so, s/he asks for its price in 31 different stores, obtaining a sample mean price of 750 euros with a sample standard deviation of 60 euros. We assume normality.

22. With a 98% confidence level, we can state that the mean telephone price is in the interval:

    (A) (723.05, 776.95) (B) (745.08, 754.92) (C) (719.88, 780.12)
    (D) (731.38, 768.62) (E) (727.65, 772.35)

23. If, in the previous confidence interval, we decide to decrease its confidence level, the width of the resulting interval will be:

    (A) The same (B) Smaller (C) Larger (D) It cannot be determined (E) -

24. With a 95% confidence level, we can state that the **standard deviation** of the telephone price is in the interval:

(A) (41.74, 63.86) (B) (48.73, 81.50) (C) (54.21, 79.87) (D) (46.82, 73.60) (E) (39.57, 110.71)

25. The number of times that an office printer fails during the day follows a Poisson distribution. In order to test the null hypothesis that the mean number of failures per day is at least 3, a random sample of 3 days is taken. At the 5% significance level, the null hypothesis is rejected if the total number of failures in the three-day period, z, is such that:

(A) $z \leq 4$ (B) $z \geq 12$ (C) $z \leq 3$ (D) $z \geq 14$ (E) $z \geq 15$

Questions 26 and 27 refer to the following exercise:

We wish to verify if the weekly sales in the four branches a given firm has follow the same distribution. Weekly sales are classified in five different categories: very bad, bad, normal, good, and very good. In each of these branches, a r.s. of 100 weeks is taken.

26. The most appropriate test to verify this hypothesis is:

(A) Tets of homogeneity

(B) Test of independence

(C) A χ^2 goodness of fit test to a normal distribution

(D) A χ^2 goodness of fit test to a completely specified distribution

(E) All false

27. Under the null hypothesis H_0, the distribution of the test statistic is:

(A) χ^2_{12} (B) $F_{5,4}$ (C) χ^2_{20} (D) t_{20} (E) $N(0,1)$

Questions 28 to 30 refer to the following exercise:

We wish to estimate the annual mean expense (in thousands of euros) for buildings maintenance in a given city. In order to do so, the population has been divided into two groups. The first group includes 20000 buildings with less that 50 years with $\sigma_1^{*2} = 1$, and the second group includes 5000 buildings with 50 of more years with $\sigma_2^{*2} = 4$. It has been decided that a stratified random sampling will be taken, with a total sample size of 300 buildings.

28. If proportional allocation is used, the corresponding sample sizes, n_1 and n_2, will be:

(A) 150 and 150 (B) 200 and 100 (C) 240 and 60 (D) 250 and 50 (E) 50 and 250

29. If optimal allocation is used, the corresponding sample sizes, n_1 and n_2, will be:

(A) 50 and 250 (B) 200 and 100 (C) 150 and 150 (D) 250 and 50 (E) 240 and 60

30. If we compare the mean estimators variances under proportional and optimal allocation, the largest one corresponds to:

(A) - (B) proportional allocation (C) they are equal (D) optimal allocation (E) -

EXERCISES (Time: 75 minutes)

A. (10 points, 25 minutes)

In the magazine "XL Semanal" published on May 7, 2017, we find the following news related to airport delays: "45% of the delays are due to the fact that the incoming airplane is late; 30% are due to unexpected luggage-related issues; 7% are due to airport-related issues; and 18% are due to weather, security or traffic restriction-related issues."

A r.s. of 900 delayed flights has been taken at a given airport. For this specific sample, causes were classified as: 346 flights were delayed because the incoming airplane was late; 230 were because of unexpected luggage-related issues; 54 were delayed because of airport-related issues; and 270 were late because of weather, security or traffic restriction-related issues. At the 5% significance level, test the hypothesis that the airport delays for this specific airport follow the general statement for airport delays included in the magazine mentioned above.

B. (10 points, 25 minutes)

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

$$P(x) = e^{(-5\lambda)} (5\lambda)^x \left(\frac{1}{x!}\right), \qquad x = 0, 1, \ldots, \qquad \lambda > 0.$$

We wish to estimate the parameter λ and, in order to do so, a r.s. of size n, $X_1, \ldots, X_n$ is taken.

It is known that the mean and variance for this distribution are both equal to 5λ.

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

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

iii) Is the method of moments estimator of λ unbiased? Is it consistent? Is it efficient?

Remark: The Cramer-Rao lower bound for a regular and unbiased estimator of λ obtained from a r.s. from the population under study is:

$$L_c = \frac{1}{nE\left[\frac{\partial \ln P(X,\lambda)}{\partial \lambda}\right]^2}$$

C. (10 points, 25 minutes)

From the 2401 individuals polled for the year 2014 in a given region, 415 of them boycotted or stopped buying some specific product.

i) If the sample was selected by using simple random sampling with replacement, obtain the 95% confidence interval for the proportion of individuals from that region who boycotted or stopped buying some specific product.

ii) At the 5% significance level, test the null hypothesis that the proportion of individuals who boycotted or stopped buying some specific product in that region was smaller than or equal to 10%.

iii) In the year 2017, a new poll is going to be conducted in the same region, and it has been decided to use simple random sampling without replacement instead. What should be the minimum sample size if the population size (15 years or older residents in that region) is 395000 people, the maximum error allowed for to estimate this proportion is ±0.02 and the confidence level is 95%?

SOLUTIONS TO MULTIPLE CHOICE QUESTIONS (exam type 0)

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

SOLUTIONS TO EXERCISES

Exercise A

It corresponds to a goodness of fit test to a completely specified distribution.

H_0 : Airport delays follow the general statement included in the magazine

H_1 : Airport delays do not follow the general statement included in the magazine

With the information provided in the exercise, we build the table below:

	n_i	p_i	np_i	$\frac{(n_i-np_i)^2}{np_i}$
Case 1	346	0.45	405	8.595
Case 2	230	0.30	270	5.926
Case 3	54	0.07	63	1.286
Case 4	270	0.18	162	72
Total	$n = 900$	1	$n = 900$	$z = 87.807$

Under the null hypothesis, the test statistic $Z = \sum_{i=1}^{k} \frac{(n_i-np_i)^2}{np_i}$ converges to a $\chi^2_{(k-1)}$ distribution, where k represents the number of classes in which the variable under study is divided.

In this case:

$$z = 87.807 > 7.81 = \chi^2_{(4-1)\,,\,0.05} = \chi^2_{3,0.05},$$

so that, at the approximate 5% significance level, we reject the null hypothesis that airport delays follow the general statement included in the magazine.

Exercise B

$$P(x,\lambda) = e^{-5\lambda}\,(5\lambda)^x \left(\frac{1}{x!}\right) \qquad x = 0, 1, 2, \ldots, \qquad \lambda > 0$$

$$\mathrm{E(X)} = \mathrm{Var(X)} = 5\lambda$$

i) **Method of moments estimator**

To obtain the method of moments estimator of λ, we make the first sample and population moments equal. That is,

$$a_1 = \alpha_1$$

Given that, in this case, $\alpha_1 = 5\lambda$ and $a_1 = \overline{X}$, we have that: $5\lambda = \overline{X}$, so that,

$$\hat{\lambda}_{\mathrm{MM}} = \frac{\overline{X}}{5}$$

ii) **Maximum likelihood estimator**

$$\begin{aligned}
L(\vec{x};\lambda) &= P(x_1;\lambda)\ldots P(x_n;\lambda) = \\
&= \left(e^{-5\lambda}\,(5\lambda)^{x_1}\,\frac{1}{x_1!}\right)\cdots\,\left(e^{-5\lambda}\,(5\lambda)^{x_n}\,\frac{1}{x_n!}\right) = \\
&= e^{-5n\lambda}\,(5\lambda)^{x_1+\cdots+x_n}\,\frac{1}{x_1!\,x_2!\cdots x_n!} = e^{-5n\lambda}\,(5\lambda)^{\sum_{i=i}^n x_i}\,\frac{1}{\prod_{i=1}^n x_i!} \\
\ln L(\vec{x};\lambda) &= -5n\lambda + \left(\sum_{i=1}^n x_i\right)\ln(5\lambda) - \ln\left(\prod_{i=1}^n x_i!\right) \\
\frac{\partial \ln L(\vec{x},\lambda)}{\partial\lambda} &= -5n + \left(\frac{\sum_{i=1}^n x_i}{\lambda}\right) = 0 \\
\hat{\lambda}_{ML} &= \frac{\sum_{i=1}^n X_i}{5n} = \frac{\overline{X}}{5}
\end{aligned}$$

iii) **Unbiasedness.** The estimator is unbiases because, in this case, we have that

$$\mathrm{E}\left(\hat{\lambda}_{\mathrm{MM}}\right) = \mathrm{E}\left(\frac{\overline{\mathrm{X}}}{5}\right) = \frac{1}{5}\mathrm{E}(\overline{\mathrm{X}}) = \frac{1}{5}\mathrm{E}(\mathrm{X}) = \frac{1}{5}\,(5\lambda) = \lambda$$

Consistency. It is a consistent estimator because the required sufficient conditions for it are satisfied:

1) $\hat{\lambda}$ is an unbiased estimator, and

2) $\lim_{n\to\infty}\left[\mathrm{Var}\left(\hat{\lambda}\right)\right] = \lim_{n\to\infty}\mathrm{Var}\left(\frac{\overline{X}}{5}\right) = \lim_{n\to\infty}\left(\frac{1}{25}\right)\left(\frac{\mathrm{Var}(X)}{n}\right) = \lim_{n\to\infty}\left(\frac{1}{25}\right)\left(\frac{5\lambda}{n}\right) = \lim_{n\to\infty}\left(\frac{\lambda}{5n}\right) = 0$

Efficiency. In order to verify if the estimator is efficient, we compute the Cramer-Rao lower bound.

$$Lc = \frac{1}{n\mathrm{E}\left[\frac{\partial \ln \mathrm{P}(\mathrm{X},\lambda)}{\partial\lambda}\right]^2}$$

$$\ln P(x,\lambda) = -5\lambda + x\ln(5\lambda) - \ln x!$$

$$\frac{\partial \ln P(x,\lambda)}{\partial\lambda} = -5 + \frac{x}{\lambda} = \frac{1}{\lambda}(x-5\lambda)$$

$$\mathrm{E}\left[\frac{\partial \ln \mathrm{P}(\mathrm{X},\lambda)}{\partial\lambda}\right]^2 = \mathrm{E}\left[\frac{1}{\lambda}(\mathrm{X}-5\lambda)\right]^2 = \frac{1}{\lambda^2}\,\mathrm{E}(\mathrm{X}-5\lambda)^2 = \frac{1}{\lambda^2}\,\mathrm{Var}(X) = \frac{1}{\lambda^2}\,(5\lambda) = \frac{5}{\lambda}$$

Therefore, we have that:

$$Lc = \frac{1}{n\left(\frac{5}{\lambda}\right)} = \frac{\lambda}{5n}$$

The variance of the estimator coincides with the Cramer-Rao lower bound, and thus, the estimator is efficient.

Exercise C

From the 2401 individuals that have been polled for the year 2014 in a given region, 415 boycotted or stopped buying some specific product.

i) For the 95% confidence interval for the proportion under simple random sampling with replacement, we have that

$$\mathrm{CI}_{1-\alpha}(\mathrm{p}) = \left(\frac{\mathrm{z}}{\mathrm{n}} \pm \mathrm{t}_{\frac{\alpha}{2}} \sqrt{\frac{\mathrm{z}(\mathrm{n}-\mathrm{z})}{\mathrm{n}^3}} \right)$$

In this case, $z = 415$, $n = 2401$ and $t_{0.025} = 1.96$:

$$\begin{aligned} \mathrm{CI}_{0.95}(\mathrm{p}) = & \left(\frac{415}{2401} \pm 1.96 \sqrt{\frac{415(2401-415)}{(2401)^3}} \right) \\ & = (0.1728 \pm 1.96 \cdot 0.007717) \\ & = (0.1577,\ 0.1879) \end{aligned}$$

ii) $H_0 : p \leq 0.1$

$H_1 : p > 0.1$

Under H_0 :

$$\frac{\frac{Z}{n} - p_0}{\sqrt{\frac{p_0\, q_0}{n}}} \longrightarrow N(0,1)$$

At the α significance level, the decision rule would then be to reject H_0 if:

$$\frac{\frac{z}{n} - p_0}{\sqrt{\frac{p_0\, q_0}{n}}} > t_\alpha$$

In this case:

$$\frac{\frac{415}{2401} - 0.1}{\sqrt{\frac{(0.1)\cdot(0.9)}{2401}}} \;=\; 11.898 > \;1.64 \;=\; t_{0.05}$$

Therefore, at the approximate 5% significance level, we reject the null hypothesis that the proportion of individuals who boycotted or stopped buying some specific product was smaller than or equal to 10%.

iii) The general formula to compute the sample size under simple random sampling without replacement is:

$$n = \frac{N\, t^2_{\alpha/2}}{4\,(N-1)\,\delta^2 \,+\, t^2_{\alpha/2}}$$

If the population under study includes 395000 individuals and we wish to estimate the proportion of individuals with a maximum error of ± 0.02 and a 95% confidence level, $t_{\frac{0.05}{2}} = 1.96$ and $\delta = 0.02$, so that:

$$n = \frac{395000\,(1.96)^2}{4\,(394999)\,(0.02)^2 \,+\, (1.96)^2} = 2386.4997$$

Therefore, it would be necessary to take a sample of 2387 individuals.