STATISTICS APPLIED TO BUSINESS ADMINISTRATION (BAD) - Second Year   Academic Year 2018-19
STATISTICS APPLIED TO MARKETING (MD) - Second Year
STATISTICS APPLIED TO BUSINESS ADMINISTRATION (DD) - Third Year
Second Call. June 27, 2019

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, whereas each question incorrectly answered will not penalize your grade in any form. Therefore, you must answer all of the questions included in the exam**.
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 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 30 minutes)

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

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

Questions 2 to 3 refer to the following exercise:

The probability that a client that visits a specific travel agency finally purchases a given trip is 0.6. We assume independence between the different clients' purchases.

2. If twenty clients visit the travel agency, the probability that more than fifteen of them purchase a given trip is:

   (A) 0.0016 (B) 0.0510 (C) 0.8744 (D) 0.1256 (E) 0.9490

3. If one hundred clients visit the travel agency, what is the approximate probability that at most 65 of them purchase a given trip?

   (A) 0.8212 (B) 0.1314 (C) 0.8686 (D) 0.1788 (E) 0.5871

4. Let X be a r.v. having a Poisson distribution with parameter λ. It is known that $P(3) = 0.2138$ and that $P(4) = 0.1336$. The mean of this r.v. is:

   (A) 2 (B) 3 (C) 4 (D) 3.5 (E) 2.5

Questions 5 and 6 refer to the following exercise:

The number of labor absences registered in a specific day in a given firm follows a Poisson distribution, where we have that $P(3) = P(4)$. We assume independence between the labor absence distributions for different days.

5. The probability that, in a given day, there are more than eight absences is:

   (A) 0.0214 (B) 0.0511 (C) 0.9786 (D) 0.9489 (E) 0.7851

6. The approximate probability that in regular labor week of five working days the total number of absences is at most 28 is:

   (A) 0.9535 (B) 0.9713 (C) 0.8849 (D) 0.7823 (E) 0.0465

Questions 7 to 8 refer to the following exercise:

Let $X_1, \ldots, X_5$ be independent and identically distributed r.v. having a $\gamma(3,3)$ distribution.

7. The mean and the variance of the r.v. X_1 are, respectively:

   (A) $m = 3$, $\sigma^2 = 3$ (B) $m = 1$, $\sigma^2 = 1$ (C) $m = 1$, $\sigma^2 = \frac{1}{3}$
   (D) $m = 1$, $\sigma^2 = 9$ (E) $m = 3$, $\sigma^2 = \frac{1}{3}$

8. The distribution of the r.v. $Y = \frac{X_1+X_2+X_3+X_4+X_5}{5}$ is:

   (A) $\gamma(\frac{3}{5}, 15)$ (B) $\gamma(\frac{3}{5}, \frac{3}{5})$ (C) $\gamma(3, \frac{3}{5})$ (D) $\gamma(3, 15)$ (E) $\gamma(15, 15)$

Questions 9 and 11 refer to the following exercise:

The duration time for a specific type of engine component, in **thousands of hours**, follows an exponential distribution with mean 5. We assume independence between the different engine components.

9. The probability that a given component of this type lasts more than ten thousand hours is:

(A) 0.135 (B) 0.834 (C) 0.865 (D) 0.166 (E) 0.027

10. If we have a batch of ten components of this type, the expected number components of such pack that will have a duration larger than ten thousand hours is:

(A) 2 (B) 8.65 (C) 1.66 (D) 1.35 (E) 0.50

11. The approximate probability that in a batch of 200 components of this type there are less than 33 with duration larger than ten thousand hours is:

(A) 0.87 (B) 0.75 (C) 0.13 (D) 0.91 (E) 0.09

Questions 12 to 14 refer to the following exercise:

Let X, Y and Z be three independent r.v. such that their distributions are as follows: $X \in N(1, \sigma^2 = 4)$, $Y \in \chi^2_5$ and $Z \in \gamma(\frac{1}{2}, 4)$.

12. If we define the r.v. $V_1 = \left(\frac{X-1}{2}\right)^2 + Y$, then $P(V_1 > 2.20)$ is:

(A) 0.10 (B) 0.95 (C) 0.75 (D) 0.90 (E) 0.05

13. If we define the r.v. $V_2 = \dfrac{\sqrt{8}\,(X-1)}{2\sqrt{Z}}$, then $P\,(-1.40 < V_2 < 1.86)$ is:

(A) 0.70 (B) 0.05 (C) 0.85 (D) 0.15 (E) 0.95

14. If we define the r.v. $V_3 = \dfrac{8Y}{5Z}$, then the value of k such that $P\,(V_3 > k) = 0.95$ is:

(A) 0.207 (B) 3.69 (C) 0.366 (D) 2.73 (E) 4.82

Questions 15 and 16 refer to the following exercise:

Let $X_1, \ldots, X_n$ be a r.s. taken from a population with probability mass function given by:

$$P(X = 1) = 2\theta, \quad P(X = 2) = \theta, \quad P(X = 3) = 1 - 3\theta$$

In order to estimate the parameter θ, a r.s. of size $n = 10$ has been taken, providing the following sample values: 1, 1, 2, 2, 2, 2, 3, 3, 3, 3.

15. The method of moment estimate of θ is:

(A) 0.20 (B) 0.22 (C) 0.18 (D) 0.24 (E) 0.16

16. The maximum likelihood estimate of θ is:

(A) 0.18 (B) 0.22 (C) 0.16 (D) 0.24 (E) 0.20

Questions 17 and 18 refer to the following exercise:

Let X be a r.v. having a uniform distribution in the interval $[0, \theta + 1]$; that is, with probability density function:

$$f(x, \theta) = \begin{cases} \frac{1}{\theta+1} & \text{if } 0 \leq x \leq \theta + 1; \\ 0 & \text{otherwise} \end{cases}$$

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

17. The method of moments estimator of θ is:

    (A) $\min\{X_i\}$ (B) $2\overline{X}$ (C) $\max\{X_i\}$ (D) $2\overline{X} - 1$ (E) $\max\{X_i\} - 1$

18. The maximum likelihood estimator of θ is:

    (A) $\max\{X_i\}$ (B) $2\overline{X}$ (C) $\max\{X_i\} - 1$ (D) $2\overline{X} - 1$ (E) $\min\{X_i\}$

Questions 19 to 21 refer to the following exercise:

Let X be a r.v. having a normal distribution with known variance σ^2. In order to estimate the mean, m, a r.s. of size n is taken and the estimator $\hat{m} = \frac{X_1+X_2+\cdots+X_n}{n-1}$ is proposed.

19. The bias of the estimator $\hat{m}$ is:

    (A) $\frac{m}{n-1}$ (B) $\frac{nm}{n-1}$ (C) 0 (D) $\frac{(n-1)m}{n}$ (E) $\frac{m}{n}$

20. The variance of the estimator $\hat{m}$ is:

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

21. About the proposed estimator, we can state that:

    (A) It is unbiased and consistent (B) It is unbiased and not consistent (C) -
    (D) It is biased and not consistent (E) It is biased and consistent

Questions 22 to 24 refer to the following exercise:

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

$$f(x; \theta) = \begin{cases} (\theta + 1)x^\theta & \text{if } x \in (0, 1) \\ \\ 0 & \text{otherwise} \end{cases}$$

We wish to test the null hypothesis $\theta = 1$ against the alternative hypothesis $\theta = 2$. In order to do so, a r.s. of size $n = 1$ is taken.

22. Let us assume that the following decision rule is proposed: we reject the null hypothesis if $x \in (0.25, 0.5)$. The significance level for this test is:

    (A) 0.1875 (B) 0.8906 (C) 0.1094 (D) 0.8125 (E) 0.2500

23. For the above decision rule, the probability of type II error is:

    (A) 0.7500 (B) 0.1875 (C) 0.8125 (D) 0.1094 (E) 0.8906

24. The general form of the most powerful critical region for this test, for x, will be:

(A) $(0, k_2]$ (B) $[k_1, k_2]$ (C) All false (D) $(k_1, k_2)^C$ (E) $[k_1, 1)$

Questions 25 to 27 refer to the following exercise:

A firm in the bottling of soft drinks sector wishes to estimate the mean energy expense per working hour of the firm's bottling machine. In order to do so, a r.s. of 31 hours is taken, providing a sample average of 20 euros and a sample variance of 25 euros^2. We assume that the distribution of the energy expense per hour of the firm's bottling machine is normal.

25. A 95% confidence interval for the above mean energy expense is:

(A) $(18.21, 21.79)$ (B) $(18.14, 21.86)$ (C) $(18.82, 21.18)$ (D) $(16.42, 23.58)$ (E) $(18.56, 21.44)$

26. We wish to test the null hypothesis that mean energy expense per working hour of the form's bottling machine is at most 18. At the α significance level, the decision will be to reject the null hypothesis if:

(A) $\frac{\sqrt{30}(\overline{x}-18)}{5} \geq t_{\overline{30}|\alpha}$ (B) $\frac{\sqrt{30}(\overline{x}-18)}{5} \geq t_{\overline{30}|\frac{\alpha}{2}}$ (C) $\frac{\sqrt{31}(\overline{x}-18)}{5} \geq -t_{\alpha}$

(D) $\frac{\sqrt{31}(\overline{x}-18)}{5} \geq t_{\alpha}$ (E) $\frac{\sqrt{30}(\overline{x}-18)}{5} \leq -t_{\overline{30}|\alpha}$

27. A 95% confidence interval for the variance of the energy expense per working hour of the firm's bottling machine is:

(A) $(18.79, 39.84)$ (B) $(16.49, 46.13)$ (C) $(19.21, 37.82)$ (D) $(17.69, 41.89)$ (E) $(15.74, 48.32)$

Questions 28 to 30 refer to the following exercise:

A firm is evaluating the possibility of offering a new service to their clients and wishes to estimate the proportion that will eventually use this new service, with a 95% confidence and a maximum error of 2%. It is known that the total number of clients the firm has is 20000.

28. What is the minimum sample size of clients the firm will need to select if simple random sample with replacement is used?

(A) 1681 (B) 2144 (C) 1976 (D) 1551 (E) 2401

29. What if simple random sample without replacement is used instead?

(A) 2401 (B) 2144 (C) 1681 (D) 1551 (E) 1976

30. Finally, the firm decides to take a simple random sample without replacement of 200 clients, from which 30 indicate that they will use the new service. The 0.95 approximate confidence interval for the proportion of clients that will eventually use this service is:

(A) $(0.101, 0.199)$ (B) $(0.114, 0.186)$ (C) $(0.118, 0.182)$ (D) $(0.122, 0.178)$ (E) $(0.109, 0.191)$

EXERCISES (Time: 75 minutes)

A. (10 points, 25 minutes)

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

$$f(x,\theta) = \begin{cases} 2e^{2(\theta-x)} & \text{if } x \geq \theta,\ \theta > 0; \\ 0 & \text{otherwise} \end{cases}$$

It is known that the mean of this distribution is $m = \frac{1}{2} + \theta$, and its variance $\sigma^2 = \frac{1}{4}$. We wish to estimate the parameter θ and, in order to do so, a r.s. of size n, $X_1, \ldots, X_n$ is taken.

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) The estimator $\hat{\theta} = \overline{X}$ is proposed to estimate the parameter θ. Compute its bias, its variance and its mean square error.

B. (10 points, 25 minutes)

A TV channel is interested in knowing if the population preferences for the different types of programs it broadcasts vary depending on the age of its audience. In order to further investigate this issue the TV channel decides to take a r.s. of 500 individuals and ask them the type of program they prefer to watch, among sports programs, TV quiz programs and series. The results the TV channel obtained, classified by the age of its audience, are reported in the following table:

	Sports programs	TV quiz programs	Series	Total
18-30	130	40	30	200
31-50	85	40	25	150
> 50	75	35	40	150
Total	290	115	95	500

At the 5% significance level, test the null hypothesis that the audience preferences for the different types of programs and their age are independent variables.

C. (10 points, 25 minutes)

Let X be a r.v. having a Poisson distribution with unknown parameter λ. We wish to test the null hypothesis $\lambda = 2$ against the alternative hypothesis $\lambda = 3$. In order to do so, a r.s. of size n, $X_1, \cdots, X_n$ is taken.

i) Obtain, **providing all relevant details**, the form of the most powerful critical region for the above test and the test statistic $\sum_{i=1}^{n} X_i$.

ii) If we wish to carry out this test at the 5% significance level for a sample of only one single observation, what will be the corresponding decision rule for the test?

iii) Obtain the decision rule for the sane significance level if a r.s. of size 3 is taken instead. What will be the power of the test for the assumed significance level?

iv) If a sample of size 3 has provided the following results: 2, 2, 5, at the 5% significance level, what will be the decision of the test?

SOLUTIONS TO MULTIPLE CHOICE QUESTIONS (exam type 0)

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

SOLUTIONS TO EXERCISES

Exercise A

The probability density function for the r.v. X is:

$$f(x,\theta)=\begin{cases}2e^{2(\theta-x)} & \text{if } x\geq\theta,\ \theta>0;\\ 0 & \text{otherwise}\end{cases}$$

with mean and variance $m=\frac{1}{2}+\theta$ and $\sigma^2=\frac{1}{4}$, respectively.

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 to the first sample moment. That is,

$$\alpha_1=E(X)=a_1=\overline{X}\Rightarrow\frac{1}{2}+\theta=\overline{X}$$
$$\hat{\theta}_{\mathrm{MM}}=\overline{X}-\frac{1}{2}$$

ii) In order to be able to obtain the maximum likelihood method estimator for θ we must take into account that sample space for this probability density function depends upon the parameter θ. That is, the likelihood function for the sample is given by:

$$L(\vec{x};\theta)=2^n e^{2n\theta-2(x_1+\cdots+x_n)},\quad \text{si}\quad x_i\geq\theta,\quad \forall i=1,\ldots,n,$$

which is an increasing function of the parameter θ and, thus, it will reach its maximum when θ takes on its highest possible value.

Given that $\theta\leq X_i$ $\forall i=1,\ldots,n$, which is equivalent to $\theta\leq\min\{X_i\}$, which implies that:

$$\hat{\theta}_{\mathrm{ML}}=\min\{X_i\}$$

iii) The estimator $\hat{\theta}=\overline{X}$ is proposed.

Bias: The bias of an estimator of θ is defined as $\mathrm{b}(\hat{\theta})=\mathrm{E}(\hat{\theta})-\theta$.

To be able to obtain the bias for the proposed estimator, we compute its expected value.

$$\mathrm{E}(\hat{\theta})=\mathrm{E}\left(\overline{X}\right)=m=\frac{1}{2}+\theta$$

Therefore, its bias is:

$$\mathrm{b}(\hat{\theta})=\mathrm{E}(\hat{\theta})-\theta=\frac{1}{2}+\theta-\theta=\frac{1}{2}$$

Variance: We compute the variance of the proposed estimator.

$$\mathrm{Var}(\hat{\theta})=\mathrm{Var}\left(\overline{X}\right)=\frac{\mathrm{Var}(X)}{n}=\frac{\sigma^2}{n}=\frac{\frac{1}{4}}{n}=\frac{1}{4n}$$

Mean square error: The mean square error can be computed as:

$$\mathrm{MSE}\left(\hat{\theta}\right)=\mathrm{Var}\left(\hat{\theta}\right)+\left[\mathrm{b}(\hat{\theta})\right]^2=\frac{1}{4n}+\left(\frac{1}{2}\right)^2=\frac{1}{4n}+\frac{1}{4}=\frac{1+n}{4n}$$

Exercise B

It corresponds to a test of independence. The information provided by the TV channel is reported in the following table:

	Sports programs	TV quiz programs	Series	Total
18-30	130	40	30	200
31-50	85	40	25	150
> 50	75	35	40	150
Total	290	115	95	500

With this information, we build the following table:

	n_{ij}	$\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j}$	$n\hat{p}_{ij}$	$\frac{(n_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$
18-30, Sports	130	$0.4 \times 0.58 = 0.232$	116	1.6897
18-30, Quiz	40	$0.4 \times 0.23 = 0.092$	46	0.7826
18-30, Series	30	$0.4 \times 0.19 = 0.076$	38	1.6842
31-50, Sports	85	$0.3 \times 0.58 = 0.174$	87	0.0460
31-50, Quiz	40	$0.3 \times 0.23 = 0.069$	34.5	0.8768
31-50, Series	25	$0.3 \times 0.19 = 0.057$	28.5	0.4298
> 50, Sport	75	$0.3 \times 0.58 = 0.174$	87	1.6552
> 50, Quiz	35	$0.3 \times 0.23 = 0.069$	34.5	0.0072
> 50, Series	40	$0.3 \times 0.19 = 0.057$	28.5	4.6404
	500	1	500	$z = 11.8119$

The corresponding theoretical probabilities $\hat{p}_{i\bullet}$ and $\hat{p}_{\bullet j}$ are estimated from the information provided by the TV channel. That is,

$$\hat{p}_{\bullet\text{Sports}} = \frac{290}{500} = 0.58 \qquad \hat{p}_{\bullet\text{Quiz}} = \frac{115}{500} = 0.23 \qquad \hat{p}_{\bullet\text{Series}} = \frac{95}{500} = 0.19$$

$$\hat{p}_{18-30,\bullet} = \frac{200}{500} = 0.4 \qquad \hat{p}_{31-50,\bullet} = \frac{150}{500} = 0.3 \qquad \hat{p}_{>50,\bullet} = \frac{150}{500} = 0.3$$

Uner the null hypothesis of independence, the test statistic $\sum_{i=1}^{k'} \sum_{j=1}^{k''} \frac{(n_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}} \sim \chi^2_{(k'-1)(k''-1)}$, where k' and k'' are the number of classes in which each of the variables has been divided. In this case,

$$11.8119 > 9.49 = \chi^2_{(3-1)(3-1)\,,\,0.05},$$

so that, at the 5% significance level, we reject the null hypothesis of independence. That is, audience preferences for the different types of programs vary with their age.

Exercise C.

$$H_0 : \lambda = 2$$
$$H_1 : \lambda = 3$$

i) To be able to obtain the most powerful critical region for this test, we need to make use of the likelihood ratio test.

$$\frac{L(\vec{x}; \lambda = 2)}{L(\vec{x}; \lambda = 3)} \leq k \quad \text{for} \quad k > 0$$

In this case, the likelihood function for the sample will be:

$$L(\vec{x}; \lambda) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\Pi_{i=1}^{n} x_i!},$$

so that the critical region will be obtained from:

$$\frac{\left(\frac{e^{-2n}2^{\sum_{i=1}^{n} x_i}}{\Pi_{i=1}^{n} x_i!}\right)}{\left(\frac{e^{-3n}3^{\sum_{i=1}^{n} x_i}}{\Pi_{i=1}^{n} x_i!}\right)} \leq k \Longrightarrow e^n \left(\frac{2}{3}\right)^{\sum_{i=1}^{n} x_i} \leq k \Longrightarrow \left(\frac{2}{3}\right)^{\sum_{i=1}^{n} x_i} \leq k_1$$

$$S = \sum_{i=1}^{n} X_i \geq C$$

Therefore, the general form of the most powerful critical region for the test statistic $S = \sum_{i=1}^{n} X_i$ is:

$$\mathrm{CR} = [C, \infty)$$

ii) In the specific case of $n = 1$, the decision rule will be to reject the null hypothesis if $X \geq C$. In order to be able to compute the value of C, we use the fixed 5% significance level, so that, under H_0, $X \in \mathcal{P}(\lambda = 2)$, and, therefore, we have that

$$\begin{aligned}
\alpha = 0.05 \geq P(\text{reject } H_0 \mid \lambda = 2) &= \\
&= P(X \geq C \mid \lambda = 2) = \\
&= 1 - F_{\mathcal{P}(\lambda=2)}(C - 1) \\
F_{\mathcal{P}(\lambda=2)}(C - 1) &\geq 0.95 \\
C - 1 &= 5 \\
C &= 6
\end{aligned}$$

The decision rule will then be to reject H_0 if $X \geq 6$.

iii) In the specific case of $n = 3$, the decision rule will be to reject H_0 if $\sum_{i=1}^{3} X_i \geq C$. In order to be able to compute the value of C, we use the fixed 5% significance level, so that, under H_0, $\sum_{i=1}^{3} X_i \in \mathcal{P}(\lambda = 6)$, and, therefore, we have that

$$\begin{aligned}
\alpha = 0.05 \geq P(\text{reject } H_0 \mid \lambda = 2) &= \\
&= P\left(\sum_{i=1}^{3} X_i \geq C \mid \lambda = 2\right) = \\
&= 1 - F_{\mathcal{P}(\lambda=6)}(C - 1) \\
F_{\mathcal{P}(\lambda=6)}(C - 1) &\geq 0.95 \\
C - 1 &= 10 \\
C &= 11
\end{aligned}$$

The decision rule will the be to reject H_0 if $\sum_{i=1}^{3} X_i \geq 11$.

We now compute the power for this test. We know that, under H_1, $\sum_{i=1}^{3} X_i \in \mathcal{P}(\lambda = 9)$, so that,

$$\begin{aligned}
\text{Power}(\lambda = 3) &= P\left(\text{reject } H_0 \mid \lambda = 3\right) \\
&= P\left(\sum_{i=1}^{3} X_i \geq 11 \mid \lambda = 3\right) = \\
&= 1 - F_{\mathcal{P}(\lambda=9)}(10) = \\
&= 1 - 0.706 = \\
&= 0.294
\end{aligned}$$

iv) If we now use the values obtained from the sample, we have that

$$x_1 + x_2 + x_3 = 2 + 2 + 5 = 9$$

Given that $9 < 11$, at the 5% significance level, we do not reject the null hypothesis.