Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

12 September 2022

Name and Surname: _____ Matricola: _____

(TOT 6 pts)

1) Let us consider two types of electronic devices, A and B. The lifetime of type-A (type-B) devices is described by an exponentially distribute random variable with parameter λ_A(λ_B, respectively). The experiment has two steps:
    1. a number z is randomly extracted by a standard normal.
    2. if $z \leq 1.645$ then a type-A component is tested otherwise a type-B component is tested. The test provides the lifetime of the component as output.

One experiment is carried on. Given t_0, a positive real number

2 pts a) what is the probability that the lifetime of the selected component will be greater than t_0 ?

2 pts b) We observe that the lifetime of the selected component is greater than t_0. What is the probability that the component is of type A?

We test ten components of type A carrying on ten independent experiments.

2 pts c) What is the probability that exactly two of them have a lifetime greater than t_0?

(6pt)

(Tot 6 pts)

2) An urn contains n numbered balls (from 1 to n). The n balls are randomly extracted from the urn *without replacement*.

1 pt 1. Determine the probability that the ordered sequence of the extracted balls is $(1, \dots, n)$.

2. Let X = the number of balls that are extracted in the right position (a ball is extracted in the right position if the number on it is equal to its position in the extracted sequence. For example, for $n = 3$, if the extracted sequence is $(3,2,1)$ the only ball in the right position is the ball number 2 and then the observed value of X is $x = 1$).

    1 pt a. Determine the probability that X is $n - 1$, that is $P(X = n - 1)$.

    b. For $n = 4$ determine

    1.5 pts i. the probability mass function of the random variable X.

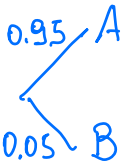
    1 pt ii. the expected value of X

1.5 pts 3. Now suppose that the n balls are randomly extracted from the urn with replacement. For $n = 4$ determine the probability mass function of the random variable X.

(6pt)

Exercise 1

X = lifetime of the component



a) $\mathbb{P}(X > t_0) = \mathbb{P}(A)\,\mathbb{P}(X > t_0 | A) + \mathbb{P}(B)\,\mathbb{P}(X > t_0 | B)$ (1)

If $X \sim \exp(\lambda)$, $t \geq 0$ then

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} \lambda e^{-\lambda x}\, dx = \lambda \left. \frac{e^{-\lambda x}}{-\lambda} \right|_0^t = 1 - e^{-\lambda t}$$

$$\Rightarrow \mathbb{P}(X > t) = e^{-\lambda t}$$

Then (1) becomes

$$\mathbb{P}(X > t_0) = 0.95\, e^{-\lambda_A t_0} + 0.05\, e^{-\lambda_B t_0}$$

b) $$\mathbb{P}(A | X > t_0) = \frac{\mathbb{P}(A)\,\mathbb{P}(X > t_0 | A)}{\mathbb{P}(X > t_0)} = \frac{0.95\, e^{-\lambda_A t_0}}{0.95 e^{-\lambda_A t_0} + 0.05 e^{-\lambda_B t_0}}$$

c) Y = number of components with lifetime greater than t_0 among the $n = 10$ selected

$Y \sim$ Binomial $(n = 10, p = e^{-\lambda_A t_0})$

$$\mathbb{P}(Y = 2) = \binom{10}{2} \left(e^{-\lambda_A t_0}\right)^2 \left(1 - e^{-\lambda_A t_0}\right)^8$$

EXERCISE 2

1) There are $n!$ possible outcomes

$$\mathbb{P}(\text{'the extracted sequence is } (1,2\dots n)\text{'}) = \frac{1}{n!}$$

2) a) $\mathbb{P}(X = n-1) = 0$. If $(n-1)$ balls are in the right positions it follow that also the remaing one is in the right position. Then "$n-1$" is not a possible value of $X \Leftrightarrow \mathbb{P}(X = n-1) = 0$.

b) $n = 4$

We can list the $n! = 24$ possible outcomes

	1st	2nd	3rd	4th	x
1	1	2	3	4	4
2	1	3	2	4	2
3	1	3	4	2	1
4	1	2	4	3	2
5	1	4	2	3	1
6	1	4	3	2	2
7	2	1	3	4	2
8	2	3	1	4	1
9	2	3	4	1	0
10	3	1	2	4	1
11	3	2	1	4	2
12	3	2	4	1	1
13	3	1	4	2	0
14	3	4	1	2	0
15	3	4	2	1	0
16	2	1	4	3	0
17	2	4	1	3	0
18	2	4	3	1	1
19	4	1	2	3	0
20	4	2	1	3	1
21	4	2	3	1	2
22	4	1	3	2	1
23	4	3	1	2	0
24	4	3	2	1	0

(i) $\Rightarrow$

x	0	1	2	3	4
$p(x)$	$\frac{9}{24}$	$\frac{8}{24}$	$\frac{6}{24}$	0	$\frac{1}{24}$

(ii)

x	$p(x)$
0	$\frac{8}{24}$
$\frac{12}{24}$	0
$\frac{4}{24}$	

$$\Rightarrow \mathbb{E}(X) = \sum_{x=0}^{4} x\,p(x) = 1$$

3) $X \sim \text{Bin}(n=4, p=1/4) \Rightarrow p(x) = \binom{4}{x}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{4-x} \quad x = 0,1,2,3,4$

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 1B (PART 2)

1pt **Quiz 1** Let X be an exponentially distributed population with parameter λ whose value is unknown. Given the sample $(X_1, X_2, X_3) = (5, 7, 8)$ then the corresponding log-likelihhod function for λ is:

[**a**] $3\log(\lambda) + \lambda^{-20}$

[**b**] $\log(3\lambda) - \lambda^{20}$

[**c**] $\log(3\lambda) - 20\lambda$

[**d**] $3\log(\lambda) - 20\lambda$

1pt **Quiz 2** Given the sample $(X_1, \ldots, X_n)$ with $n \geq 3$, consider the following estimators for the mean:

$$M_1 = \frac{X_1 + 2X_n}{2} \qquad M_2 = \frac{X_1 + 2X_2 + X_3}{3}.$$

$E[M_1] = \frac{3}{2}\mu$

$E[M_2] = \frac{4}{3}\mu$

Which of the following assertions is true?

[**a**] The estimator M_1 is consistent.

[**b**] The estimator M_2 is consistent.

[**c**] For $n = 4$ the bias of M_2 is smaller than the bias of M_1.

[**d**] For $n = 6$ the bias of M_1 is smaller than the bias of M_2.

$Bias[M_1] = E[M_1] - \mu = \frac{1}{2}\mu$

$Bias[M_2] = E[M_2] - \mu = \frac{1}{3}\mu$

1pt **Quiz 3** Let X_p be a Bernoulli distributed population with parameter p whose value is unknown and let Θ_n be an unbiased estimator of p based on a random sample extracted from X_p having a finite size n. Then by the Cramer-Rao's Theorem we can affirm that:

[**a**] the variance of Θ_n can not be greater than $(1 - p)$;

[**b**] the variance of Θ_n can not be smaller than $(1 - p)$;

[**c**] the variance of Θ_n can not be greater than $p(1 - p)/n$;

[**d**] the variance of Θ_n can not be smaller than $p(1 - p)/n$.

1pt **Quiz 4** Given a sample of size n from a population X, we want to perform a χ^2 test where the null hypothesis H_0 is that X has exponential distribution. For it, we use a statistical software, which return us a p-value equal 0.03. Which of the following assertions is true?

[**a**] We do not reject H_0 if the test is with confidence 95%.

[**b**] We reject H_0 if the test is with confidence 95%.

[**c**] The χ^2 goodness of fit test can not used in this case since the exponential distribution is continuous.

[**d**] To decide if reject or not H_0 we also need to know in how many classes the sample has been subdivided.

(Tot 6 pts) **Ex 1)** A number of 60 measurements concerning the monthly household income (in euros) are made for a given population of 100000 households. Assume that the distribution of these incomes is normally distributed, and that the sample returned a sample mean equal to 410 and a corrected sample variance equal to 2400.

1.5 pt (a) Provide an interval estimate of the mean of the monthly incomes with a confidence 90%.

1.5 pt (b) Assuming that the corrected sample variance remains the same, determine the minimum sample size to have a 90% confidence interval with a width less than 5 euros.

1.5 pt (c) Performing a test for the hypothesis $H_0 : \mu \leq 400$ against $H_1 : \mu > 400$ at a level $\alpha = 0.1$, is there statistical evidence to reject H_0?

1.5 pt (d) Recalling the the sample mean assumed the value 410, calculate the corresponding p-value in testing the hypothesis of the previous point.

(Tot 6 pts) **Ex 2)** You want to estimate the proportion of students at your university who answer *yes* to the question of whether governments should do more about global warming. In a random sample of 10 students, 8 answered *yes*, and 2 answered *no*. Provide a point estimate of the probability that the next student interviewed will answer *yes*

2 pts (a) if you use maximum likelihood estimation (showing the passages in the calculations);

2 pts (b) if you use a Bayesian estimate with a prior distribution of type Beta with $\alpha = 2$ and $\beta = 1$.

2 pts (c) Assume now that you have interviewed 100 students, and that 80 of them answered *yes*. Provide a one-sided confidence interval at level 90% (infinite to the left) of the probability that the next student interviewed will answer *yes*.

EX 1

(a) Large sample. $I = \left(\bar{X}_n - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}},\ \bar{X}_n + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_n^2}{n}}\right)$

$\bar{X}_n = 410$, $\hat{\sigma}_n^2 = 2400$, $n = 60$, $z_{1-\alpha/2} = z_{0.95} = 1.65$

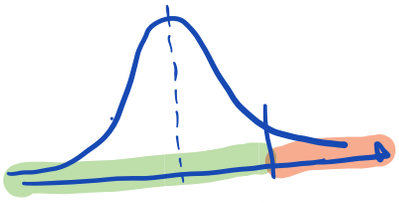
$\rightarrow I = (410 - 1.65 \cdot \sqrt{40},\ 410 + 1.65 \cdot \sqrt{40}) \underline{\sim} (400, 420)$

(b) $2 \cdot z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\sigma}_n^2}{n}} \leq 5$

$\rightarrow \sqrt{n} \geq \frac{2 \cdot z_{1-\alpha/2} \cdot \sqrt{\hat{\sigma}_n^2}}{5}$ $\rightarrow n \geq \left(\frac{2 \cdot 1.65 \cdot \sqrt{2400}}{5}\right)^2 = 1045.44$

$\rightarrow n \geq 1046$

(c)



$z_{0.9} = 1.29$

$Z_n = \frac{\bar{X}_n - 400}{\sqrt{40}} = \frac{410 - 400}{\sqrt{40}} = 1.58$

$Z_n \in$ rejection region

$\Rightarrow$ We reject H_0.

(d) The p-value is the prob in red:

$$P[Z > 1.58] = 1 - P[Z \leq 1.58]$$
$$= 1 - 0.94 = 0.06$$

EX 2

(a) $$L(p|\bar{x}) = p^{\sum x_i} \cdot (1-p)^{n-\sum x_i} = p^8 (1-p)^2$$

$$\log L(p|\bar{x}) = 8 \log p + 2 \log(1-p)$$

$$\frac{\partial \log L(p|\bar{x})}{\partial p} = \frac{8}{p} - \frac{2}{1-p} = \frac{8-8p-2p}{p(1-p)} = 0 \Rightarrow 8 - 10p = 0$$

$$\Rightarrow \hat{p} = \frac{8}{10} = 0.8$$

(b) $$\hat{p} = \frac{\sum x_i + \alpha}{n + \alpha + \beta} = \frac{8+2}{10+2+1} = \frac{10}{13} \approx 0,77$$

(c) $$I = \left(-\infty, \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \quad \text{with } \hat{p} = \bar{X}_{100} = \frac{80}{100}$$

(large sample)

$$= \left[0, 0.8 + z_{0.9}\sqrt{\frac{0.8 \cdot 0.2}{100}}\right) \quad \text{(can not be smaller than 0)}$$

$$= \left[0, 0.8 + 1.29 \cdot \sqrt{0.0016}\right)$$

$$\approx [0, 0.85)$$

1. Let us consider the following two-step experiment:

   Step 1. a fair coin is tossed;
   Step 2. the second step is conducted according to the results of the first step:
      i. if "head" has appeared in step 1, then n=5 extractions are made with replacement from an urn that contains 2 white and 8 black balls;
      ii. if "tail" has appeared in step 1, then n=5 extractions are made without replacement from an urn that contains 3 white and 7 black balls.

   a. What is the probability that, at the end of the experiment, exactly two white balls are observed?
   b. What is the probability that by repeating the two-step experiment 15 times, in exactly one case, two white balls are observed?
   c. Let us suppose that we replace the fair coin with a coin such that P[head]=α (P[tail]=1-α). Which is the value of α such that the probability that, at the end of the experiment, exactly two white balls are observed becomes 0.25?

(6pt)

i) W = number of white balls in $n=5$ trials

$$P(W=2) = P(H)\,P(W=2|H) + P(T)\,P(W=2|T)$$

$$P(H) = P(T) = 1/2$$

$$W|H \sim \text{Binomial}\left(n=5, p=\frac{2}{10}=\frac{1}{5}\right)$$

$$W|T \sim \text{Hypergeometric}(N=10, M=3, n=5)$$

$$P(W=2|H) = \binom{5}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^3 = 10\cdot\frac{4^3}{5^5} = 2\cdot\frac{4^3}{5^4} = \frac{128}{625} \simeq 0.2048$$

$$P(W=2|T) = \frac{\binom{3}{2}\binom{7}{3}}{\binom{10}{5}} = 3\cdot\frac{7!}{3!4!}\,\frac{5!5!}{10!} = 3\,\frac{7\cdot6\cdot5}{5\cdot2}\cdot\frac{5\cdot4\cdot3\cdot2}{10\cdot9\cdot8\cdot7\cdot6} =$$

$$= \frac{5}{12} \simeq 0.416667$$

$$P(W=2) = \frac{1}{2} \cdot 0.2048 + \frac{1}{2} 0.416667 \cong 0.310733$$

b) Y = number of times in which two balls are observed in 15 2-step experiments

$$Y \sim Bin(n=15, p=0.310733)$$

$$P(Y=1) = \binom{15}{1} 0.310733 (1-0.310733)^{14} =$$

$$= 15 \cdot 0.310733 \cdot 0.689267^{14} \cong 0.025462$$

c) $\alpha\, 0.2048 + (1-\alpha) 0.41667 = 0.25$

$$(0.2048 - 0.41667)\,\alpha = 0.25 - 0.41667$$

$$\alpha = \frac{0.25 - 0.41667}{0.2048 - 0.41667} = \frac{-0.16667}{-0.21187} = 0.786658$$

2. Suppose that for a certain individual, calorie intake at breakfast is a random variable X_B with expected value 500 and standard deviation 50, calorie intake at lunch is a random variable X_L with expected value 900 and standard deviation 100, and calorie intake at dinner is a random variable X_D with expected value 2000 and standard deviation 180.

a) Assuming that intakes at different meals are independent of one another, what is the approximate probability that average calorie intake per day over the next (365-day) year is at most 3500?
b) Assuming that intakes at different meals are correlated with correlation coefficients $\rho(X_B, X_L)=\rho(X_B, X_D)=0$ and $\rho(X_L, X_D)=0.9$, what is the approximate probability that average calorie intake per day over the next (365-day) year is at most 3500?
c) Assuming now that X_B, the intake at breakfast. is a uniform random variable $X_B \sim U(400,600)$ what is the probability that the total intake at breakfast of two randomly chosen days is less than 900 calories?

(6pt)

·	X_B	X_L	X_D
$E(\cdot)$	500	900	2000
$\sigma_\cdot$	50	100	180

a) $Y = X_B + X_L + X_D$

$E[Y] = 500 + 900 + 2000 = 3400$

$V(Y) = 50^2 + 100^2 + 180^2 = 44900$ $\quad (X_B, X_L, X_D)$ indep

$\bar{Y}_{365} \approx N\left(3400, \frac{44900}{365}\right) \equiv N(3400, 123.0137)$

Approx CLT

$P(\bar{Y}_{365} \leq 3500) = P\left(Z \leq \frac{3500-3400}{\sqrt{\frac{44900}{365}}}\right) =$

$Z \sim N(0,1)$

$= P\left(Z \leq \frac{100}{11.09115}\right) = P(Z \leq 9.016) \cong 1$

In general $V(X_1+X_2) = V(X_1)+V(X_2)+2\,cov(X_1,X_2)$ and $cov(X_1,X_2) = \rho_{1,2}\,\sigma_1\,\sigma_2$

b) $E[Y] = 3400$

$$V[Y] = \sigma_B^2 + \sigma_L^2 + \sigma_D^2 + 2\rho_{L,D}\sigma_L\sigma_D =$$

$$= 44900 + 2\cdot 0.9\cdot 100\cdot 180 =$$

$$= 44900 + 32400 = 77300$$

$$\bar{Y}_{365} \underset{\text{Approx CLT}}{\sim} N\left(3400, \frac{77300}{365}\right) \equiv N(3400, 211.7808)$$

$$P(\bar{Y}_{365} \le 3500) = P\left(Z \le \frac{100}{\sqrt{211.7808}}\right) = P\left(Z \le \frac{100}{14.55269}\right)$$

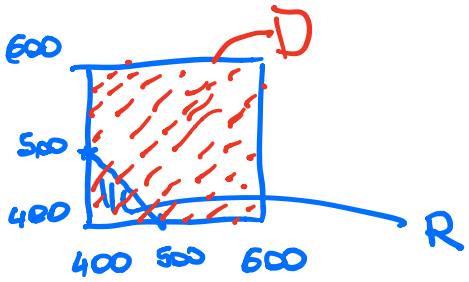
$$= P(Z \le 6.8716) \simeq 1$$

c) $X_B \sim U(400, 600)$

$$P(X_{B_1} + X_{B_2} < 900) =$$

$$= \frac{Area(R)}{Area(D)} = \frac{100\cdot 100}{2\cdot 200^2} =$$

$$= \frac{10^4}{2\cdot 2^2\cdot 10^4} = \frac{1}{8}$$



~~STOCHASTIC~~ STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 1 (PART 2)

Quiz 1 Which of the following statements is false?

[a] The sequence $\{X_n, n \in \mathbb{N}\}$ of random variables converges in almost sure sense to X iff the sequence $\{Y_n, n \in \mathbb{N}\}$, where $Y_n = X_n - X$, converges almost surely to the constant 0.

[b] The sequence $\{X_n, n \in \mathbb{N}\}$ of random variables converges in distribution to X iff the sequence $\{Y_n, n \in \mathbb{N}\}$, where $Y_n = X_n - X$, converges in distribution to the constant 0.

[c] The sequence $\{X_n, n \in \mathbb{N}\}$ of random variables converges in probability to X iff the sequence $\{Y_n, n \in \mathbb{N}\}$, where $Y_n = X_n - X$, converges in probability to the constant 0.

[d] All the other 3 statements are true.

Quiz 2 Given a distribution depending on a parameter θ, the estimator $\widehat{\theta}$ of θ is said to be unbiased if

[a] $\widehat{\theta}$ does not depend on the size n of the sample $(X_1, ..., X_n)$.

[b] $\widehat{\theta}$ is a function of the sample $(X_1, ..., X_n)$.

[c] $E[\widehat{\theta} - \theta]$ converges to 0 as the size n of the sample increases to ∞.

[d] $E[\widehat{\theta} - \theta] = 0$.

Quiz 3 Consider a bilateral interval estimate of the mean of normally distributed population, with confidence $1 - \alpha$, for which the variance σ^2 is known. Initially we take a sample of size n, then double the size.

[a] The second interval will be with confidence $1 - 2\alpha$.

[b] The second interval will be with confidence $1 - \alpha/2$.

[c] The second interval will be a subset of the first one.

[d] The first interval will be a subset of the second one.

Quiz 4 Let X be a population having a Bernoulli distribution with parameter p. We want to test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ (for a specified $p_0 \in (0, 1) \subseteq \mathbb{R}$) by using a sample with large size (say $n = 100$). Which of the following procedures CAN NOT be used to perform the test?

[a] The test for $H_0 : E[X] = p_0$ based on the normal approximation for the sample mean $\overline{X}_n$.

[b] The χ^2 (Chi-square) goodness of fit test for $H_0 : X \sim \text{Bernoulli}(p_0)$.

[c] The Kolmogorov-Smirnov goodness of fit test for $H_0 : X \sim \text{Bernoulli}(p_0)$.

[d] Actually, none of the other three statements is false, and all such test procedures can be used.

Ex 1) Let X be a population having density and defined as

$$f_{\alpha,\theta}(x) = \begin{cases} \left(\frac{1}{\theta}\right)^{\alpha} \frac{x^{\alpha-1}}{(\alpha-1)!}\, e^{-x/\theta} & \text{if } x \geq 0; \\ 0 & \text{if } x < 0; \end{cases}$$

where the parameters α and θ can assume values in $\mathbb{R}^+$. Note that with straightforward calculations one can verify that

$$E[X] = \alpha\theta \quad \text{and} \quad V(X) = \alpha\theta^2$$

(you can avoid these calculations). Assume you extract the sample $\overline{X} = (3, 5, 2, 6)$ from such a population.

(a) Find the estimators for θ and α using the method of moments, clearly writing all the passages you used to define them.

(b) Assuming $\alpha = 2$, find the MLE estimator for θ.

EX 2)

We have two populations: $X \sim N(\mu_x, \sigma^2)$ and $Y \sim N(\mu_x, \sigma^2)$, where σ^2 is know and equal 16. From X we extract a sample of size $n = 5$ that gives $\bar{X}_5 = 9.5$, while from Y we extract a sample of size $n = 10$ that gives $\bar{Y}_{10} = 10$.

(a) Test, with confidence 95%, the hypothesis $H_0 : \mu_X = \mu_Y$.

(b) Assume now that the difference $\mu_X - \mu_Y$ can assume only the values 0 or 1, and we want to test $H_0 : \mu_X - \mu_Y = 0$ (with the only alternative $H_1 : \mu_X - \mu_Y = 1$). To this aim, we decide to reject H_0 if $\bar{X}_5 \leq \bar{Y}_{10}$. Compute the probability of errors of both the I and the II type.

SOLUTIONS

QUIZZES: 1-B 2-D 3-C 4-C

Ex 1

(a) $E[X^2] = V[X] + E[X]^2 = \alpha\theta^2 + (\alpha\theta)^2 = \alpha(1+\alpha)\theta^2$

If $M_1 = \frac{\sum X_i}{n}$ and $M_2 = \frac{\sum X_i^2}{n}$, then

$$\begin{cases} \hat{\alpha}\hat{\theta} = M_1 \\ \hat{\alpha}(1+\hat{\alpha})\hat{\theta}^2 = M_2 \end{cases} \longrightarrow \hat{\theta} = \frac{M_1}{\hat{\alpha}}$$

$$\hat{\alpha}(1+\hat{\alpha})\frac{M_1^2}{\hat{\alpha}^2} = M_2 \longrightarrow \frac{1}{\hat{\alpha}} M_1^2 = M_2 - M_1^2$$

$$\longrightarrow \hat{\alpha} = \frac{M_1^2}{M_2 - M_1^2} \qquad \hat{\theta} = \frac{M_2 - M_1^2}{M_1}$$

Since here $M_1 = 4$, $M_2 = \frac{37}{2} \Rightarrow \hat{\alpha} = \frac{32}{5}$ $\hat{\theta} = \frac{5}{8}$

(b) Here $\alpha=2 \rightarrow f_\theta(x) = \left(\frac{1}{\theta}\right)^2 \cdot x \cdot e^{-x/\theta}$

$$L(\theta \mid \bar{X}) = \prod_{i=1}^{n} f_\theta(x_i) = \left(\frac{1}{\theta}\right)^{2n} \cdot \prod_{i=1}^{n} x_i \; e^{-\frac{1}{\theta}\cdot\Sigma x_i}$$

$$\log L(\theta,\bar{X}) = 2n \log\left(\frac{1}{\theta}\right) + \log\left(\prod_{i=1}^{n} x_i\right) - \frac{1}{\theta}\Sigma x_i$$

$$\frac{\partial \log L(\theta,\bar{X})}{\partial \theta} = 2n \cdot \frac{1}{1/\theta} \cdot \left(-\frac{1}{\theta^2}\right) + \frac{1}{\theta^2}\Sigma x_i$$

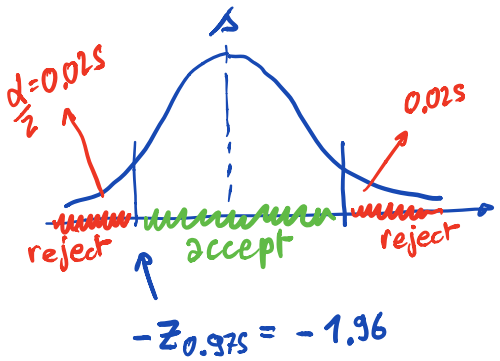
$$= -2n \cdot \frac{1}{\theta} + \frac{\Sigma x_i}{\theta^2} = \frac{-2n\theta + \Sigma x_i}{\theta^2} = 0$$

if $\hat{\theta} = \frac{\Sigma x_i}{2n}$

EX 2

$\bar{X}_5 = 9.5 \quad \bar{Y}_{10} = 10 \quad \sigma^2 = 16$ for both.

(a) $\bar{X}_{n_1} - \bar{Y}_{n_2} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right) = N(\mu_X - \mu_Y, 4.8)$



$$Z_{obs} = \frac{(\bar{X}_5 - \bar{Y}_{10}) - 0}{\sqrt{4.8}} = -0.23$$

$Z_{obs} \in$ Accept region $\Rightarrow$ Do not reject!

(b) $P[\text{err I}] = P[\text{reject } H_0 \mid H_0 \text{ true}] = P[\bar{X}_5 < \bar{Y}_{10} \mid \mu_X = \mu_Y]$

$$= P\left[\frac{(\bar{X}_5 - \bar{Y}_{10}) - 0}{\sqrt{4.8}} < 0\right] = P[Z < 0] = \frac{1}{2}$$

$P[\text{err II}] = P[\text{accept } H_0 \mid H_0 \text{ false}]$

$$= P[\bar{X}_5 \geq \bar{Y}_{10} \mid \mu_X - \mu_Y = 1]$$

$$= P\left[\frac{(\bar{X}_5 - \bar{Y}_{10}) - 1}{\sqrt{4.8}} \geq \frac{0-1}{\sqrt{4.8}}\right] = P[Z \geq -0.46] = 0.677$$

Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

30 June 2021

Name and Surname: _____ Matricola: _____

---

1) Consider randomly selecting a single individual and having that person test 3 different smartphones. Define events A_i="the selected person likes smartphone i", i=1,2,3.
Suppose that
$P(A_1) = 0.55, P(A_2) = 0.65, P(A_3) = 0.70,$
$P(A_1 \cup A_2) = 0.80, P(A_2 \cap A_3) = 0.40,$ and $P(A_1 \cup A_2 \cup A_3) = 0.88.$

a. What is the probability that the individual likes both smartphone 1 and smartphone 2?
b. Determine and interpret $P(A_2|A_3)$.
c. Are A_2 and A_3 independent events?
d. If you learn that the individual did not like smartphone 1, what now is the probability that he/she liked at least one of the other two smartphones?

(6pt)

2) After shuffling a deck of 12 cards (4 cards for each of 3 suits), a dealer deals out 3 cards. Let Y = the number of suits represented in the three-card hand.
   1. **(4pt)** Determine the probability mass function (pmf) of Y.
   2. **(2pt)** In the case of 52 cards (13 cards for each of 4 suits) let X = the number of suits represented in the five-card hand. The pmf of X is

x	1	2	3	4
$p(x)$	.002	.146	.588	.264

Compute the mean and the standard deviation of X, μ_X and σ_X respectively. The following computations are available.

x*p(x)	0,002	0,292	1,764	1,056
x^2*p(x)	0,002	0,584	5,292	4,224

(6pt)

a) $\mathbb{P}(A_1 \cap A_2)$

ES 7

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

$$\Rightarrow \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cup A_2)$$

$$= 0.55 + 0.65 - 0.80 = 0.40$$

b) $$\mathbb{P}(A_2 | A_3) = \frac{\mathbb{P}(A_2 \cap A_3)}{\mathbb{P}(A_3)} = \frac{0.40}{0.70} = \frac{4}{7} \simeq 0.571429$$

c) $\mathbb{P}(A_2) = 0.65$

$\mathbb{P}(A_2 | A_3) = 0.5714..$ $\Rightarrow$ non sono ind.p perchè valori diversi

d) $$\mathbb{P}(A_2 \cup A_3 | \bar{A}_1) = \frac{\mathbb{P}((A_2 \cup A_3) \cap \bar{A}_1)}{\mathbb{P}(\bar{A}_1)} = \frac{\mathbb{P}(A_1 \cup A_2 \cup A_3) - \mathbb{P}(A_1)}{1 - \mathbb{P}(A_1)} =$$

$$= \frac{0.88 - 0.55}{1 - 0.55} = \frac{0.33}{0.45} \simeq 0.733$$

$\Rightarrow A_1 \cup A_2 \cup A_3 \setminus A_1$

2.a

4	4	4
Λ_1	Λ_2	Λ_3

$\mathbb{P}(Y=1) = 3\,\mathbb{P}(\text{"all cards of suit 1"})$ (3 = number of suits)

$$= 3 \cdot \frac{\binom{4}{3}}{\binom{12}{3}} = 3 \cdot \left(4 \cdot \frac{3!\,9!}{12!}\right) = \cancel{3} \cdot \cancel{4} \cdot \frac{3 \cdot \cancel{2}}{\cancel{12} \cdot 11 \cdot \cancel{10}_5} = \frac{3}{55}$$

$\mathbb{P}(Y=2) = \binom{3}{2}\,\mathbb{P}(\text{"all cards of suit 1 or suit 2"})$ ($\binom{3}{2}$ = number of groups made by 2 suits)

$$= 3 \cdot \left(\frac{\binom{4}{1}\binom{4}{2} + \binom{4}{2}\binom{4}{1}}{\binom{12}{3}} \right) = 3 \cdot 2 \cdot 4 \cdot 6 \cdot \frac{3!\,9!}{12!} = \cancel{3} \cdot 2 \cdot \cancel{4} \cdot 6 \cdot \frac{3 \cdot \cancel{2}}{\cancel{12} \cdot 11 \cdot \cancel{10}_5} =$$

$$= \frac{36}{55}$$

$$\mathbb{P}(Y=3) = 1 - \mathbb{P}(Y=1) - \mathbb{P}(Y=2) = 1 - \frac{39}{55} = \frac{16}{55}$$

Verifice $\mathbb{P}(Y=3) = \dfrac{\binom{4}{1}\binom{4}{1}\binom{4}{1}}{\binom{12}{3}} = 4^3 \cdot \dfrac{3!\,9!}{12!} = 4^3 \cdot \dfrac{3 \cdot 2}{12 \cdot 11 \cdot 10} =$

$$= 4^2 \cdot \cancel{4} \cdot \frac{\cancel{3} \cdot \cancel{2}}{\cancel{12} \cdot 11 \cdot \cancel{10}_5} = \frac{16}{55}$$

2.b

x	1	2	3	4	Σ
$p(x)$	0.002	0.146	0.588	0.264	
$x\,p(x)$	0.002	0.292	1.764	1.056	3.114
$x^2 p(x)$	0.002	0.584	5.292	4.224	10.102

$\mu_X = 3.114$

$V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

$= 10.102 - (3.114)^2 = 0.405004$

$\Rightarrow \sigma_X = \sqrt{0.405004} = 0.636399$

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 3 (PART 2)

Quiz 1 Let X be a normally distributed random variable having mean $\mu = 1$ and variance $\sigma^2 = 4$. Then, by using the Chebishev's inequality, we can affirm that:

[a] $P[X^2 \leq 10] \geq 1/2$.

[b] $P[X^2 \leq 10] \geq 3/4$.

[c] $P[X^2 \geq 10] \leq 1/4$.

[d] $P[X^2 \geq 10] \geq 1/4$.

$P[|X| \leq t] \geq \frac{E[X^k]}{t^k}$

$P[X^2 \leq t] \geq \frac{E[X^{2k}]}{t^k} = \frac{E[X^2]}{t} = \frac{V[X] + E[X]^2}{t} = \frac{4+1}{t}$ (for $k=1$)

Quiz 2 Given a random variable X_θ whose distribution depends on a parameter θ, in the Bayesian setting we say that the family of distributions $\mathcal{B}$ is a coniugate for the family $\mathcal{A}$

[a] if X_θ has distribution in $\mathcal{A}$ and the posterior distribution for θ is in $\mathcal{B}$ whenever the prior is in $\mathcal{B}$.

[b] if X_θ has distribution in $\mathcal{B}$ and the posterior distribution for θ is in $\mathcal{A}$ whenever the prior is in $\mathcal{A}$.

[c] if the posterior distribution for θ is in $\mathcal{A}$ whenever the prior is in $\mathcal{B}$.

[d] if the posterior distribution for θ is in $\mathcal{B}$ whenever the prior is in $\mathcal{A}$.

Quiz 3 Let the sample $\{5, 5, 2\}$ be extracted from a population having normal distribution with mean μ and variance σ^2. Knowing that the quantiles of order 0.05 and 0.95 for a χ^2 distribution with 2 degrees of freedom are approximatively equal to 0.10 and 6.00, respectively, then the interval estimate for the variance σ^2 with confidence 90% is:

[a] $(1, 60)$.

[b] $(0.5, 30)$.

[c] $(2.90, 9)$.

[d] $(1.90, 8)$.

$\bar{X}_3 = 4$ $\quad \hat{S}^2 = \frac{1+1+4}{2} = 3$ $\quad I = \left(\frac{(n-1)\cdot \hat{S}^2}{\chi^2_{0.95}}, \frac{(n-1)\hat{S}^2}{\chi^2_{0.05}} \right)$

$\cong \left(\frac{6}{6}, \frac{6}{0.1} \right) = (1, 60)$

Quiz 4 Let X be normally distributed with unknown mean μ and known variance $\sigma^2 = 1$. Assume that μ can take only two possible values: 1 and 2. We perform a test for $H_0 : \mu = 1$ versus $H_1 : \mu = 2$ using a sample of size $n = 9$, the test statistic $\bar{X}_9$ and the acceptance region $C = (-\infty, 1.3)$. Then the probability of a type II error is approximately equal to

[a] 0.02

[b] 0.98

[c] 0.19

[d] 0.81

$P[\text{err I}] = P[\text{accept } H_0 \mid H_1 \text{ true}] = P[\bar{X}_n < 1.3 \mid \mu = 2]$

$= P\left[\frac{\bar{X}_n - 2}{\sqrt{1/9}} < \frac{1.3-2}{\sqrt{1/9}} \right] = P[Z < -2.1] = 1 - P[Z < 2.1]$

$= 1 - 0.98 = 0.02$

EX 1 - version A

Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random sample from a population X having Pareto distribution with parameter α, i.e., whose density is

$$f_\alpha(x) = \begin{cases} \frac{\alpha}{x^{\alpha+1}}, & \text{if } x \geq 1; \\ 0, & \text{if } x < 1. \end{cases}$$

Assume we know in advance that $\alpha > 2$.

(a) Knowing that for $\alpha > 1$ it holds $\mathbb{E}(X) = \frac{\alpha}{\alpha-1}$, find an estimator of α by using the method of moments (MME).

(b) Knowing that for $\alpha > 2$ it holds $\mathrm{Var}(X) = \frac{\alpha}{(\alpha-1)^2(\alpha-2)}$, find the limiting distribution of the MME estimator when $n \to \infty$ by using the Delta method.

(c) Find the maximum likelihood estimator (MLE) of α.

EX 2 - version A

Let X be a population having discrete distribution, assuming values in the set of non negative integers.

(a) For a sample having size $n = 100$, one gets the following observed frequencies:

observed value	0	1	2	3	4
frequency	20	40	20	10	10

Test the hypothesis that X has Poisson distribution, where the value of λ is the one estimated from the sample. For it, uses $\alpha = 0.05$.

(b) Assume you want to test the hypothesis $H_0 : \lambda \leq 1$ versus $H_1 : \lambda > 1$. To this aim, assume you take a sample of size $n = 10$ (different from the one above), and you decide to reject H_0 if the corresponding sample mean $\bar{X}_{10}$ assumes a value greater than 1.2. How can we find the probability of an error of type I? [Do not make all calculations, just explain how to find it]

Ex 1

(a) $\frac{\alpha}{\alpha-1} = E[X] \cong \bar{X}_n \longrightarrow \hat{\alpha} = \frac{\bar{X}_n}{\bar{X}_n - 1} = g(\bar{X}_n)$ where $g(x) = \frac{x}{x-1}$

(b) Since $\bar{X}_n \xrightarrow{n\to\infty} N\left(E[X], \frac{V[X]}{n}\right)$

$$g(\bar{X}_n) \xrightarrow{n\to\infty} N\left(g(E[X]), \frac{V[X]}{n} \cdot \left[g'(E[X])\right]^2\right)$$

$g'(x) = -\frac{1}{(x-1)^2}$ $\quad g\left(\frac{\alpha}{\alpha-1}\right) = \alpha$ $\quad g'\left(\frac{\alpha}{\alpha-1}\right) = -(\alpha-1)^2$

$$\frac{V[X]}{n} \cdot \left[g'\left(\frac{\alpha}{\alpha-1}\right)\right]^2 = \frac{1}{n} \frac{\alpha}{(\alpha-1)^2(\alpha-2)} \cdot (\alpha-1)^4 = \frac{1}{n} \frac{\alpha(\alpha-1)^2}{\alpha-2}$$

$$g(\bar{X}_n) \longrightarrow N\left(\alpha, \frac{\alpha(\alpha-1)^2}{n \cdot (\alpha-2)}\right)$$

(c) $L(\alpha|\bar{x}) = \alpha^n \cdot \prod_{i=1}^{n} (x_i)^{-(\alpha+1)} = \alpha^n \left(\prod_{i=1}^{n} x_i\right)^{-(\alpha+1)}$

$\log L(\alpha|\bar{x}) = n \cdot \log\alpha - (\alpha+1) \log\left(\prod_{i=1}^{n} x_i\right)$

$= n \log\alpha - (\alpha+1)\left(\sum_{i=1}^{n} \log x_i\right)$

$\frac{d \log L(\alpha|\bar{x})}{d\alpha} = \frac{n}{\alpha} - \sum_{i=1}^{n} \log x_i$

$= 0 \quad \text{if} \quad \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \log x_i}$

Ex 2

(a) $\hat{\lambda} = \bar{X}_n = \frac{0\cdot 20 + 1\cdot 40 + 2\cdot 20 + 3\cdot 10 + 4\cdot 10}{100} = 1.5$

observed value	0	1	2	3	≥ 4
frequency	20	40	20	10	10
p_i	0,225	0.335	0,25	0,125	0.065
$100 \cdot p_i$	22.5	33.5	25	12.5	6.5

$p_0 = \frac{1.5^0}{0!} e^{-1.5} \simeq 0.225$

$p_1 = \frac{1.5^1}{1!} e^{-1.5} \simeq 0.335$

$p_2 \simeq 0.25$

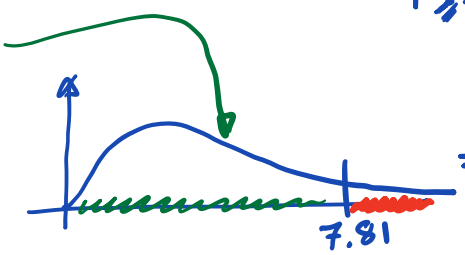
$p_3 \simeq 0.125$

$p_{\geq 4} = 1 - \sum_{i=0}^{3} p_i \simeq 0.065$

$W = \sum_{i=1}^{N} \frac{(f_i - np_i)^2}{np_i} \simeq 4.9$

$\chi^2_{5-1-1,\,0.95} = 7.81$

(1 parameter is estimated)



$\Rightarrow$ Do not reject!

(b) $\bar{X}_n > 1.2 \Rightarrow X_1 + \cdots X_{10} > 12$

$X_1 + \cdots + X_{10} \sim \text{Poiss}(10\lambda)$. If $\lambda = 1$, then $X_1 + \cdots + X_{10} \sim \text{Poiss}(10)$

$\alpha = P[\text{Poiss}(10) > 12] = 1 - F^{-1}(12)$ where F = cumulative of Poiss(10)

EX 1 - version B

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM JUNE 2021 (EX 1 - PART 2)

Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random sample from a population X having the following density (depending on a parameter $\beta \geq 0$):

$$f_\alpha(x) = \begin{cases} \frac{\beta+1}{x^{\beta+2}}, & \text{if } x \geq 1; \\ 0, & \text{if } x < 1. \end{cases}$$

Assume we know in advance that $\beta > 1$.

(a) Knowing that for $\beta > 0$ it holds $\mathbb{E}(X) = \frac{\beta+1}{\beta}$, find an estimator of β by using the method of moments (MME).

(b) Knowing that for $\beta > 1$ it holds $\mathrm{Var}(X) = \frac{\beta+1}{\beta^2(\beta-1)}$, find the limiting distribution of the MME estimator when $n \to \infty$ by using the Delta method.

(c) Find the maximum likelihood estimator (MLE) of β.

(a) $E[X] = \frac{\beta+1}{\beta} \longrightarrow \frac{\hat{\beta}+1}{\hat{\beta}} = \bar{X}_n \longrightarrow \hat{\beta} = \frac{1}{\bar{X}_n - 1}$

(b) $\hat{\beta} = g(\bar{X}_n)$ where $g(x) = \frac{1}{x-1}$ $\quad g'(x) = -\frac{1}{(x-1)^2}$

$g(E[X]) = \beta \qquad g'(E[X]) = -\beta^2$

$\frac{V[X]}{n}\left[g'(E[X])\right]^2 = \frac{1}{n}\,\frac{\beta^2(\beta+1)}{(\beta-1)}$

$\hat{\beta} = g(\bar{X}_n) \longrightarrow N\left(\beta, \frac{1}{n}\,\frac{\beta^2(\beta+1)}{(\beta-1)}\right)$

(c) $L(\beta|\bar{X}) = (\beta+1)^n \prod_{i=1}^n x_i^{-(\beta+2)}$

$\log L(\beta|\bar{X}) = n \log(\beta+1) - (\beta+2) \cdot \sum_{i=1}^n \log x_i$

$\frac{d \log L(\beta|\bar{X})}{d\beta} = \frac{n}{\beta+1} - \sum \log x_i \longrightarrow \hat{\beta} = \frac{n}{\sum \log x_i} - 1$

EX 2 - Version B

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM JUNE 2021 (EX 2 - PART 2)

Let X be a population having discrete distribution, assuming values in the set of non negative integers.

(a) For a sample having size $n = 100$, one gets the following observed frequencies

observed value	0	1	2	3	4
frequency	40	35	15	5	5

Test the hypothesis that X has Poisson distribution, where the value of λ is the one estimated from the sample. For it, uses $\alpha = 0.10$.

(b) Assume you want to test the hypothesis $H_0 : \lambda \leq 1.5$ versus $H_1 : \lambda > 1.5$. To this aim, assume you take a sample of size $n = 10$ (different from the one above), and you decide to reject H_0 if the corresponding sample mean $\bar{X}_{10}$ assumes a value greater than 2.0. How can we find the probability of an error of type I? [Do not make all calculations, just explain how to find it]

(a) $\hat{\lambda} = \bar{X}_n = \frac{0\cdot 40 + 1\cdot 35 + 2\cdot 15 + 3\cdot 5 + 4\cdot 5}{100} = 1$

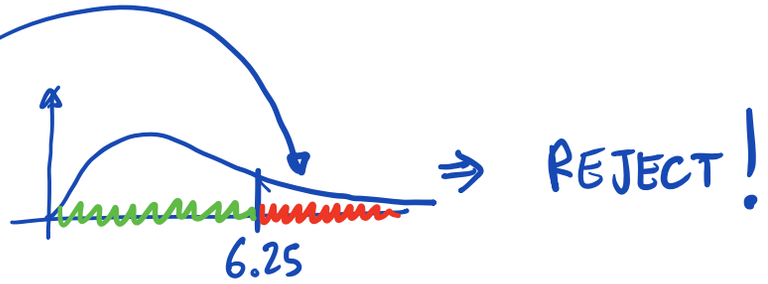
observed value	0	1	2	3	4
frequency	40	35	15	5	5
p_i	0.37	0.37	0.18	0.06	0.02
$100 \cdot p_i$	37	37	18	6	2

$p_0 = \frac{1^0}{0!} e^{-1} \simeq 0.37$

!
!
etc..

$W = \sum_{i=1}^{N} \frac{(f_i - np_i)^2}{np_i} \simeq 9.04$

$\chi^2_{5-1-1,\,0.90} = 6.25$



$\Rightarrow$ REJECT!

(b) $\bar{X}_n > 2.0 \Rightarrow X_1 + \cdots X_{10} > 20$

$X_1 + \cdots + X_{10} \sim \text{Poiss}(10\lambda)$. If $\lambda = 1.5$, then $X_1 + \cdots + X_{10} \sim \text{Poiss}(15)$

$\alpha = P[\text{Poiss}(15) > 20] = 1 - F^{-1}(20)$ where F = cumulative of Poiss(15)

1. Individual A has a circle of five close friends (B, C, D, E, and F). A has heard a certain rumor from outside the circle and has invited the five friends to a party to circulate the rumor. To begin, A selects one of the five at random and tells the rumor to the chosen individual. That individual then selects at random one of the four remaining individuals and repeats the rumor. Continuing, a new individual is selected from those not already having heard the rumor by the individual who has just heard it, until everyone has been told.
    a. What is the probability that the rumor is repeated in the order B, C, D, E, and F?
    b. What is the probability that F is the third person at the party to be told the rumor? What is the probability that F is the last person to hear the rumor?
    c. If at each stage the person who currently "has" the rumor does not know who has already heard it and selects the next recipient at random from all five possible individuals, what is the probability that F has still not heard the rumor after it has been told ten times at the party?

(6pt)

a) There are $5!$ possible ways of ordering B, C, D, E, F
Then the prob. of selecting (B, C, D, E, F) is $\frac{1}{5!} = \frac{1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{1}{120}$

b) There are $4!$ possible ways to obtain
$$(a_1, a_2, F, a_3, a_4)$$
where (a_1, a_2, a_3, a_4) is a permutation of B, C, D, E
Then the prob. $P(\text{"F is 3rd"})$ is $\frac{4!}{5!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{1}{5} = 20\%$

Similarly $P(\text{"F is last"}) = 20\%$

c) At each stage the prob that F is not selected is $\frac{4}{5}$
Then $P(\text{"F hasn't heard after 10 times"}) = \left(\frac{4}{5}\right)^{10} \cong 0.1074$

2. The accompanying table gives information on the type of coffee selected by someone purchasing a single cup at a particular airport kiosk.

	Small	Medium	Large
Regular	14%	20%	26%
Decaf	20%	10%	10%

Consider randomly selecting such a coffee purchaser.

a. What is the probability that the individual purchased a small cup? A cup of decaf coffee?
b. If we learn that the selected individual purchased a small cup, what now is the probability that he/she chose decaf coffee, and how would you interpret this probability?

Now consider randomly selecting ten coffee purchasers.

c. What is the probability that at least 2 of them purchase small cups?

a) $P(\text{"small cup"}) = 0.14 + 0.20 = 0.34$

$P(\text{"decaf"}) = 0.20 + 0.10 + 0.10 = 0.40$

b) $P(\text{decaf} \mid \text{small cup}) = \dfrac{P(\text{small cup} \cap \text{decaf})}{P(\text{small cup})} =$

$= \dfrac{0.20}{0.34} \simeq 0.588$

We have that 58.8% of people who purchase a small cup choose decaf coffee.

c) Let Y = number of purchasers of small cups among the 10 randomly selected.

$Y \sim \text{Binomial}(n = 10, p = 0.34)$

$P(Y \geq 2) = 1 - P(Y \leq 1) = 1 - P(Y=0) - P(Y=1) =$

$= 1 - 0.66^{10} - 10 \cdot 0.34 \cdot 0.66^{9} = 1 - 0.015683 - 0.080793$

$= 0.903524$

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 4 (PART 2)

Quiz 1 Let $\{X_n, n \in \mathbb{N}\}$ be a sequence of random variables that converges in distribution to a variable X having cumulative distribution F_X. Which of the following is surely true?

[a] $\lim_{n\to\infty} E[\arctan(X_n)] = E[\arctan(X)]$, whenever the expectations exist. because the function arctan is bounded and continuous

[b] $\lim_{n\to\infty} E[X_n] = E[X]$, whenever the expectations exist.

[c] $\lim_{n\to\infty} Var[X_n] = Var[X]$, whenever the variances exist.

[d] $\lim_{n\to\infty} E[\ |X_n - X|\ ] = 0$, whenever the expectations exist.

Quiz 2 Let a population X have an exponential distribution with unknown parameter λ. Given the sample $(X_1, X_2) = (3, 5)$ then the likelihood function for λ is:

[a] $L(\lambda|(3,5)) = \lambda^2 e^{-8\lambda}$. $L(\lambda|(3,5)) = f_\lambda(3) \cdot f_\lambda(5) = \lambda e^{-3\lambda} \cdot \lambda e^{-5\lambda} = \lambda^2 e^{-8\lambda}$

[b] $L(\lambda|(3,5)) = \lambda e^{-2\lambda}$.

[c] $L(\lambda|(3,5)) = 2\lambda e^{-8\lambda}$.

[d] $L(\lambda|(3,5)) = \lambda^8 e^{-\lambda}$.

Quiz 3 Let $\bar{x}_n$ and $\hat{s}_n^2$ be the unbiased estimated of the mean and the variance obtained from a sample $\{X_1, \ldots, X_n\}$ extracted from a normally distributed population. Then the interval estimate for the variance σ^2 with confidence $1 - \alpha$ is defined as:

[a] $\left(\frac{(n-1)\hat{s}_n^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)\hat{s}_n^2}{\chi^2_{n-1,\alpha/2}}\right)$. see slides lectures

[b] $\left(\frac{n\hat{s}_n^2}{\chi^2_{n,1-\alpha/2}}, \frac{n\hat{s}_n^2}{\chi^2_{n,\alpha/2}}\right)$.

[c] $\left(\hat{s}_n^2 - \frac{\chi^2_{n,\alpha/2}}{n}, \hat{s}_n^2 + \frac{\chi^2_{n,1-\alpha/2}}{n}\right)$.

[d] $\left(\hat{s}_n^2 - \frac{\chi^2_{n-1,\alpha/2}}{n-1}, \hat{s}_n^2 + \frac{\chi^2_{n-1,1-\alpha/2}}{n-1}\right)$.

Quiz 4 In a Kolmogorov-Smirnov goodness of fit test for the null hypothesis $H_0 : F = F_0$, with confidence $1 - \alpha$:

[a] the test statistics is defined as $D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$, where $\hat{F}_n$ is the empirical distribution, and one rejects H_0 whenever D_n is greater than the $(1 - \alpha)$-quantile available from specific tables for the K-S test.

[b] the test statistics is defined as $D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$, where $\hat{F}_n$ is the empirical distribution, and one rejects H_0 whenever D_n is smaller than the α-quantile available from specific tables for the K-S test.

[c] the test statistics is defined as $D_n = \inf_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$, where $\hat{F}_n$ is the empirical distribution, and one rejects H_0 whenever D_n is smaller than the α-quantile available from specific tables for the K-S test.

[d] the test statistics is defined as $D_n = \inf_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)|$, where $\hat{F}_n$ is the empirical distribution, and one rejects H_0 whenever D_n is greater than the $(1 - \alpha)$-quantile available from specific tables for the K-S test.

see slides lectures

Ex 1) Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random sample from a population X having Bernoulli distribution with parameter p. To estimate p, the following point estimator has been suggested by your best friend:

$$\hat{\Theta}_1 = \frac{1 + \sum_{i=1}^{n} X_i}{n+1},$$

(a) Is Θ_1 un unbiased estimators of p? Is it consistent?

(b) Find its MSE (mean square error).

(c) Compare it with the Cramer-Rao lower bound for the variance of an unbiased estimator of p, and comment what you find.

Ex 2) We have two populations: $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_Y, \sigma_y^2)$, where σ_x^2 and σ_y^2 are known and equal 16 and 9, respectively. From X we extract a sample of size $n = 10$ that gives $\bar{X}_{10} = 21.5$, while from Y we extract a sample of size $n = 20$ that gives $\bar{Y}_{20} = 20$.

(a) Provide a confidence interval with confidence 90% for the difference $\mu_x - \mu_y$.

(b) Test, with confidence 95%, the hypothesis $H_0 : \mu_x - \mu_y = 1$.

(c) What is the p-value for the test in point (b)?

Ex 1

a)

$$\Theta_1 = \frac{1}{n+1} + \frac{n}{n+1}\bar{X}_n$$

$$E[\Theta_1] = \frac{1}{n+1} + \frac{n}{n+1}p \neq p \rightarrow \text{not unbiased}$$

$$P[|\Theta_1 - p| < \varepsilon] = P[|1 + n\bar{X}_n - (n+1)p| < (n+1)\varepsilon]$$

$$\geq 1 - \frac{E[(1+n\bar{X}_n-(n+1)p)^2]}{(n+1)^2\varepsilon^2}$$

$$= 1 - \frac{V[(1+n\bar{X}_n-(n+1)p)] + E[(1+n\bar{X}_n-(n+1)p)]^2}{(n+1)^2\varepsilon^2}$$

$$V[(1+n\bar{X}_n-(n+1)p)] = V[X_1+\cdots X_n] = np(1-p)$$

$$E[(1+n\bar{X}_n-(n+1)p)] = 1 + nE[\underbrace{X_1+\cdots+X_n}_{n}] - (n+1)p = 1 + np - np - p = 1-p$$

$$= 1 - \frac{np(1-p) + (1-p)^2}{(n+1)^2\varepsilon^2} \xrightarrow{n\to\infty} 1$$

Thus $\Theta_1 \xrightarrow{p} p$, i.e., Θ_1 is consistent

b)

$$V[\Theta_1] = V\left[\frac{1}{n+1} + \frac{n}{n+1}\bar{X}_n\right] = \left(\frac{n}{n+1}\right)^2 \cdot V[\bar{X}_n] = \left(\frac{n}{n+1}\right)^2 \cdot \frac{p(1-p)}{n}$$

$$\text{Bias}(\Theta_1) = E[\Theta_1] - p = \frac{1}{n+1} + \frac{n}{n+1}p - \frac{n+1}{n+1}p = \frac{1+np-np-p}{n+1} = \frac{1-p}{n+1}$$

$$MSE(\theta_1) = V[\theta_1] + Bias(\theta_1)^2 = \left(\frac{1}{n+1}\right)^2 \left[np(1-p) - (1-p)^2 \right]$$

c) $$f_p(t) = \begin{cases} p & \text{if } t=1 \\ 1-p & \text{if } t=0 \end{cases} \qquad \ln f_p(t) = \begin{cases} \ln p & \text{if } t=1 \\ \ln(1-p) & \text{if } t=0 \end{cases}$$

$$\frac{\partial \ln f_p(t)}{\partial p} = \begin{cases} 1/p & \text{if } t=1 \\ -\frac{1}{1-p} & \text{if } t=0 \end{cases}$$

$$E\left[\left(\frac{\partial \ln f_p(X)}{\partial p} \right)^2 \right] = \left(\frac{1}{p}\right)^2 \cdot P[X=1] + \left(-\frac{1}{1-p}\right)^2 \cdot P[X=0]$$

$$= \frac{1}{p^2} \cdot p + \frac{1}{(1-p)^2} \cdot (1-p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

$$\text{C-R lower bound} = \left[n \cdot \frac{1}{p(1-p)} \right]^{-1} = \frac{p(1-p)}{n}$$

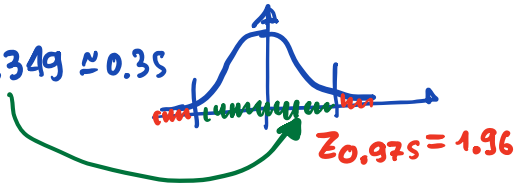
Thus, $V(\theta_1)$ is smaller than the C-R lower bound (since $\left(\frac{n}{n+1}\right)^2 < 1$). Thus, it seems that θ_1 is better than $\bar{X}_n$ to estimate p. But it is not, because θ_1 is not unbiased!

Ex 2

$$\bar{X}_n - \bar{Y}_m \sim N\left(\mu_X - \mu_Y, \frac{16}{10} + \frac{9}{20}\right) = N\left(\mu_X - \mu_Y, 2.05\right)$$

a) $$I = \left(\bar{X}_n - \bar{Y}_m - z_{0.95}\sqrt{2.05},\ \bar{X}_n - \bar{Y}_m + z_{0.95}\cdot\sqrt{2.05}\right) = (-0.86,\ 3.86)$$

b) $$z_{obs} = \frac{(\bar{X}_n - \bar{Y}_m) - 1}{\sqrt{2.05}} = 0.349 \simeq 0.35$$



$\Rightarrow$ Do not reject H_0

c) p-value (for two-sided test) $= P[|Z_{obs}| > 0.35]$

$$= 2 \cdot (1 - \Phi(0.35))$$

$$\simeq 0.727$$

Version A

Statistical Methods in Data Science

Master course in

DATA SCIENCE AND ENGINEERING

3/~~26~~ January 2022

Name and Surname: _____ Matricola: _____

1.1 One box contains two red pens and four black pens. Another box contains one red pen and three black pens. One pen is randomly extracted from the first box and another pen is independently extracted from the second box. What is the probability that both the extracted pens are red?

A	1/12	B	1/2	C	7/12	D	3/10

1.2 Let A and B be events. Suppose that $P(A) = 1/3$, $P(B)$=1/6, $P(B|A)$=1/2. Then $P(A \cap B)$ is

A	1/3	B	1/18	C	1/2	D	1/6

1.3 Let $X \sim Hypergeometric(N = 5, M = 3, n = 3)$ where N is the size of the population, M is the number of successes in the population, and n is the size of the sample. Which is the false statement between the following?

A	$P[X = 3] = 0$	B	$P[X = 2] = 3/5$	C	$P[X = 1] = 3/10$	D	$P[X = 0] = 0$

1.4 Let X be a Poisson random variable with mean E[X]=3. Then V[-2X+12], the variance of -2X+12 is

A	6	B	0	C	12	D	24

ANSWERS

The score for each correct answer is 1 point, 0 points for each wrong or not given answer.

Question	Your choice A	B	C	D	*Please briefly justify your answers.*
1.1	X				2R 4B | 1R 3B B_i = "the i-th pen is red" $i=1,2$ $P(B_1 \cap B_2) = P(B_1)P(B_2) = \frac{2}{6}\cdot\frac{1}{4} = \frac{1}{12}$ (↑ indep.)
1.2				X	$P(A \cap B) = P(A)P(B\|A) = \frac{1}{3}\cdot\frac{1}{2} = \frac{1}{6}$
1.3	X				$X \sim Hyp(N=5, M=3, n=3)$ succ. / failures $P[X=3] = \frac{\binom{3}{3}\binom{2}{0}}{\binom{5}{3}} = \frac{1}{10}$ $P(X=2) = \frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{3}} = \frac{6}{10} = \frac{3}{5}$ $P[X=1] = \frac{\binom{3}{1}\binom{2}{2}}{\binom{5}{3}} = \frac{3}{10}$ $P(X=0) = P(3-X=3) = 0$ (↑ $3-X \le 2$)
1.4			X		$V(-2X+12) = 4V(X)$ $V(X) = E(X) = 3$ $\Rightarrow V(-2X+12) = 12$

Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

31 January 2022

Name and Surname: _____ Matricola: _____

---

1) Suppose that the quantity of water that George consumes during
    1. the morning is a random variable (X_M) with an expected value of 700 and a standard deviation of 20 [ml];
    2. the afternoon is a random variable (X_A) with an expected value of 500 and a standard deviation of 20 [ml];
    3. the rest of the day is a random variable (X_R) with an expected value 300 and a standard deviation 10 [ml].

    Assume quantities of water at different times of the day and/or in different days are independent of one another.

a) What are the expected value and the standard deviation of the random variable that represents the total quantity of water that George consumes in one day?

b) What is the approximate probability that the average over the next 365 days of the total quantity of water consumed by George is greater than 1502.466 ml?

c) Suppose that the probability that water that is consumed in one morning is more than 695 ml is 0.6, $P[X_M>695] = 0.6$. What is the probability that, in the next 10 mornings, in exactly two mornings, the consumed water will be more than 695 ml?

(6pt)

2) A box contains four numbered balls:
    - ball number "k" means "win prize k", k=1,2,3;
    - ball number "4" means "win prizes 1,2, and 3".

    One ball will be randomly selected. Let A_i=" win prize i", i=1,2,3.

    a) Are A_1 and A_2 independent events? Are A_1 and A_2 mutually exclusive events?
    b) Are A_1, A_2 and, A_3 independent events?

    Let's consider another box which contains ten numbered balls:
    - ball number "k" means "win prize k", k=1,…9;
    - ball number "10" means "win prizes 1 and 10".

    c) A person selects three balls randomly without replacement. What is the probability that he/she will win prize 1?

    Justify your answers.

(6pt)

$\mathbb{E}(X_M) = 700 \quad \mathbb{E}(X_A) = 500 \quad \mathbb{E}(X_R) = 300$

$\sigma_{X_M} = 20 \quad \sigma_{X_A} = 20 \quad \sigma_{X_R} = 10$

a) $X_D = X_M + X_A + X_R \qquad \mathbb{E}(X_D) = \mathbb{E}(X_M) + \mathbb{E}(X_A) + \mathbb{E}(X_R) = 700 + 500 + 300 = \underline{1500}$

$V(X_D) = V(X_M) + V(X_A) + V(X_R) = 20^2 + 20^2 + 10^2 = 900 = 30^2$

$\underline{\sigma_{X_D} = 30}$

b) $\mathbb{P}(\overline{X}_D > 1502.466) \cong$

Central Limit Thm $\overline{X}_D \overset{CLT}{\approx} N\left(1500, \frac{900}{365}\right)$

$\cong \mathbb{P}\left(Z > \frac{1502.466 - 1500}{30/\sqrt{365}}\right) \cong \mathbb{P}(Z > 1.57) = 1 - \Phi(1.57) = 1 - 0.9418 = 0.0582$

$\uparrow$ $Z \sim N(0,1)$

c) $\mathbb{P}(X_M > 695) = 0.6$ $\qquad$ Y = "number of mornings in which the quality will be greater than 680 among the next 10"

$Y \sim Bin(n = 10, p = 0.6)$

$\mathbb{P}(Y = 2) = \binom{10}{2} 0.6^2 \, 0.4^8 = F_Y(2) - F_Y(1) = 0.012 - 0.002 = 0.010$

$\uparrow$ tables $\qquad$ (Excel 0.010617)

B_i = "the ball with number i is extracted" $i = 1, \dots 4$

$$P(A_1) = P(B_1 \cup B_4) = P(B_1) + P(B_4) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

($B_1 \cap B_4 = \emptyset$)

$$P(A_2) = P(B_2 \cup B_4) = \frac{1}{2}$$

$$P(A_3) = P(B_3 \cup B_4) = \frac{1}{2}$$

$$P(A_1 \cap A_2) = P(B_4) = \frac{1}{4}$$

$$P(A_1 \cap A_2 \cap A_3) = P(B_4) = \frac{1}{4}$$

2) Are A_1 and A_2 independent?

$$P(A_1 \cap A_2) \stackrel{?}{=} P(A_1) P(A_2)$$

$$\frac{1}{4} \stackrel{?}{=} \frac{1}{2} \cdot \frac{1}{2}$$ Yes, A_1 and A_2 are indep.

Are A_1 and A_2 mutually exclusive?

$P(A_1 \cap A_2) > 0$ then $A_1 \cap A_2 \neq \emptyset$, A_1 and A_2 are not mutually exclusive

b) Are A_1, A_2, A_3 indep.

A_1, A_2 are indep. (point a)

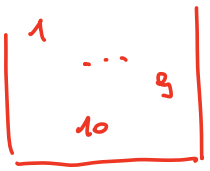
A_1, A_3 are indep } This can be proven as it has been done for A_1 and A_2

A_2, A_3 are indep }

$$P(A_1 \cap A_2 \cap A_3) \stackrel{?}{=} P(A_1) P(A_2) \cdot P(A_3)$$

$$\frac{1}{4} \stackrel{?}{=} \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}$$ No

$\Rightarrow A_1, A_2, A_3$ are not independent

c) 

ball 1 and ball 10 are successes

ball 2 ... ball 9 are failures

X = number of successes in 3 trials (extraction without replacement)

$X \sim Hyp(N=10, M=2, n=3)$

$P(\text{"person wins prize"}) = P(X=1) + P(X=2) =$

$$= \frac{\binom{2}{1}\binom{8}{2}}{\binom{10}{3}} + \frac{\binom{2}{2}\binom{8}{1}}{\binom{10}{3}} = \left(\frac{8 \cdot 7}{2} + 8\right) \cdot \frac{3!\,7!}{10!} =$$

$$= \frac{8}{\cancel{64}} \cdot \frac{\cancel{3} \cdot \cancel{2}}{\cancel{10} \cdot \cancel{9} \cdot \cancel{8}} = \frac{8}{15} \simeq 0.5\overline{3}$$

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM (PART 2 A)

Quiz 1. Given a sequence $\{X_n, n \in \mathbb{N}\}$ of random variables with cumulative distribution function F_{X_n}. We say that $\{X_n\}_n$ converges in distribution to the random variable X, having cumulative distribution function F_X if:

(a) $\forall \varepsilon > 0 \lim_{n\to\infty} P[\omega : |X_n(\omega) - X(\omega)| < \varepsilon] = 1$.
(b) $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ for all $t \in \mathbb{R}$.
(c) $\lim_{n\to\infty} P[X_n \leq t] = P[X \leq t]$ for all t continuity point for F_X.
(d) $\forall \varepsilon > 0 \lim_{n\to\infty} P[\omega : |X_n(\omega) - X(\omega)| < \varepsilon] = 0$.

Quiz 2. Given a sequence $\{X_n, n \in \mathbb{N}\}$ of independent and identically distributed random variables such that $X_i \sim Poisson(\lambda)$ (then $\mathbb{E}[X_i] = \lambda$ and $Var(X_i) = \lambda$). We define $S_n = X_1 + \cdots + X_n$. The Law of Large Numbers affirms that:

(a) $\frac{S_n - n\lambda}{\sqrt{n\lambda}}$ converges in distribution to a normal random variable $N(0, 1)$.
(b) $\frac{S_n - \lambda}{\sqrt{n\lambda}}$ converges in distribution to a normal random variable $N(0, 1)$.
(c) $\frac{S_n}{n} - \lambda$ converges in probability to λ.
(d) $\frac{S_n}{n} - \lambda$ converges in probability to 0.

Quiz 3. Given the sample data (5; 4.5; 5.2; 5.5; 6; 3.8) extracted from a normally distributed population. An interval estimate for the variance with a confidence level $\alpha = 0.1$ is

(a) $I = \left[\frac{5\times 0.6}{\chi^2_{5,0.95}}; \frac{5\times 0.6}{\chi^2_{5,0.05}}\right]$
(b) $I = \left[\frac{5\times 0.8}{\chi^2_{5,0.95}}; \frac{5\times 0.8}{\chi^2_{5,0.05}}\right]$
(c) $I = \left[5 - z_{0.95}\sqrt{\frac{0.6}{6}}, 5 + z_{0.95}\sqrt{\frac{0.6}{6}}\right]$
(d) $I = \left[\frac{6\times 0.8}{\chi^2_{6,0.95}}; \frac{6\times 0.8}{\chi^2_{6,0.05}}\right]$

Quiz 4. We consider an hypothesis test with rejection region $\bar{C}_1$. Then, we increase the rejection region to $\bar{C}_2$ which contains the set $\bar{C}_1$. We can say that:

(a) The probability of the second type error increases.
(b) The probability of the first type error decreases.
(c) The probability of the first type error increases.
(d) The probability of the second type error does not change.

Exercise 1). Let $(X_1, \ldots, X_n)$ a sample of size n of density function

$$f_X(x;\theta) = \begin{cases} (5\theta - 1)x^{5\theta-2} & \text{for } x \in (0,1) \\ 0 & \text{otherwise;} \end{cases} \tag{1}$$

with $\theta > 1$.

(a) Find an estimator for θ using the method of moments.

(b) Find the maximum likelihood estimator for θ.

Exercise 2). A company of electronic components decides to estimate the mean number of small imperfections on an assembled system. To this purpose it analyzes 16 systems and notes that the average number of imperfections is 32 and that the sample variance is 8.

(a) Assuming that the imperfections are normally distributed, find the confidence interval at 95 % for the average number of small imperfections on assembled systems.

(b) Assuming that the imperfections are normally distributed, test, with confidence 90% the hypothesis that the true mean is 31.

(c) Assuming the variance to be known, and equal to 9, find the p-value for the test in point (b).

Sol. GROUP A

1.c 2.d 3.a 4.c

Ex 1

$$f_X(x,\theta) = \begin{cases} (5\theta-1)x^{5\theta-2} & x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

a) Method of moments:

$$E[X] = \int_0^1 x(5\theta-1)x^{5\theta-2}dx = (5\theta-1)\int_0^1 x^{5\theta-1}dx$$

$$= (5\theta-1)\frac{x^{5\theta}}{5\theta}\Big|_0^1 = \frac{5\theta-1}{5\theta} = 1-\frac{1}{5\theta}$$

To find the estimator we set

$$1-\frac{1}{5\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\Rightarrow 1-\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{5\hat{\theta}}$$

$$\Rightarrow 5\hat{\theta} = \frac{1}{1-\bar{X}_n} \Rightarrow \hat{\theta} = \frac{1}{5}\frac{1}{1-\bar{X}_n}$$

b) Maximum likelihood estimator:

We look for the likelihood function

$$L(\theta | x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

$$= (5\theta - 1)^n \left(\prod_{i=1}^{n} x_i \right)^{5\theta - 2} \prod_{i=1}^{n} \mathbb{1}_{(0,1)}(x_i)$$

Now we take the logarithm

$$\ln L = n \ln(5\theta - 1) + (5\theta - 2) \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \ln(\mathbb{1}_{(0,1)}(x_i))$$

and now we look for the max.

$$\frac{\partial}{\partial \theta} \ln L = \frac{5n}{5\theta - 1} + 5 \sum_{i=1}^{n} \ln x_i = 0$$

$$\text{iff} \quad \frac{n}{5\theta - 1} = - \sum_{i=1}^{n} \ln x_i$$

$$\Leftrightarrow \quad 5\theta - 1 = \frac{-n}{\sum_{i=1}^{n} \ln x_i}$$

$$\Leftrightarrow \theta = \frac{1}{5} - \frac{n}{5 \sum_{i=1}^{n} \ln x_i} \quad \text{which is}$$

a point of maximum because the

second derivative is always negative:

$$\frac{\partial^2}{\partial\theta^2} \ln L = \frac{-25\,m}{(5\theta-1)^2} < 0$$

The maximum likelihood estimator for θ is

$$\hat{\theta}_{MLE} = \frac{1}{5} - \frac{m}{5\sum_{i=1}^{m} \ln X_i}$$

EX2 $m = 16$ $\bar{X}_m = 32$ $\hat{\sigma}_m^2 = 8$

a) $X \sim N(\mu, \sigma^2)$

$$I = \left[\bar{X}_m - t_{m-1,1-\alpha/2} \frac{\hat{\sigma}_m}{\sqrt{m}}, \bar{X}_m + t_{m-1,1-\alpha/2} \frac{\hat{\sigma}_m}{\sqrt{m}} \right]$$

$1-\alpha = 0.95$ $\alpha = 0.05$ $\alpha/2 = 0.025$

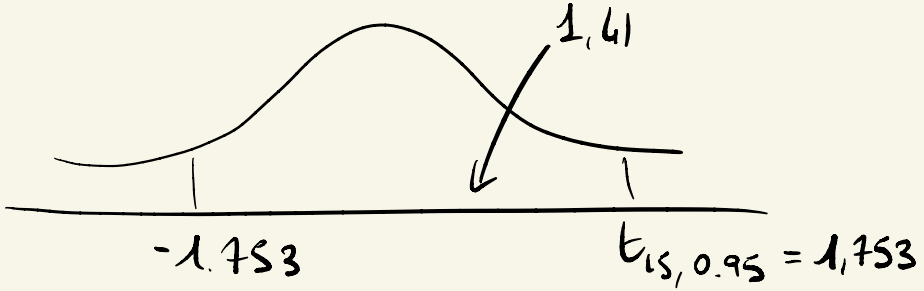
$t_{15,\,0.975} = 2.1315$

$$I = \left[32 - 2.1315\sqrt{\frac{8}{16}}, 32 + 2.1315\sqrt{\frac{8}{16}} \right] = [30.49; 33.51]$$

b) $H_0: \mu = 31$

$$T = \frac{\bar{X}_m - 31}{\sqrt{8/16}} = \frac{32-31}{\sqrt{1/2}} = \sqrt{2} = 1.41$$

$1-\alpha = 0.9 \quad \alpha = 0.1 \quad \alpha/2 = 0.05$



we do not reject H_0!

c) $\sigma^2 = 9$ $\quad T_{obs} = \dfrac{\bar{X}_n - 31}{\sqrt{9/16}} = \dfrac{4}{3} = 1,33$

$p\text{-value} = \mathbb{P}(T > 1.33) = 1 - \Phi(1.33) \simeq 0.09$

(for one-sided tests)

$p\text{-value} = 2 \min\{\mathbb{P}(t > 1.33), \mathbb{P}(t \leq 1.33)\}$

$= 2\left(1 - \Phi(1.33)\right) \simeq 0.18$

(for two-sided tests)

Version A

Statistical Methods in Data Science

Master course in

DATA SCIENCE AND ENGINEERING

21 February 2022

Name and Surname: _____ Matricola: _____

QUESTIONS

1.1 A die is rolled once. Let A="the number of dots in the upper face is 1", B="the number of dots in the upper face is 6", P(A)>0, and P(B)>0. Which is the true statement among the following?

A	A and B are independent	B	A and B are disjoint	C	$P(A\cap B)=1/36$	D	$P(A\cap B)=P(A)P(B)$

1.2 There are 7 people and 7 cats. In how many different ways can cats be assigned to people (exactly one cat to each person)?

A	$\binom{7}{7}$	*B*	$7\cdot 7$	*C*	7^7	*D*	7!

1.3 Let $X_1 \sim N(\mu, \sigma^2)$ and $X_2 \sim N(\mu, \sigma^2)$ two independent and identically distributed normal random variables. Let Φ denote the cdf of a standard normal. The probability p that the absolute value of the difference between X_1 and X_2 is less than 2σ is

A	B	C	D
$p = 2\Phi(\sqrt{2}) - 1$	p depends on μ, σ	$p = 2\Phi(\sqrt{2\sigma}) - 1$	$p = 0$

1.4 Let X and Y be two Poisson random variable with mean E[X]=E[Y]=3 and correlation ρ(X,Y)=0.1. Then Cov[X,Y] (the covariance between X and Y) is

A	0.3	B	0.1	C	0	D	0.9

ANSWERS

The score for each correct answer is 1 point, 0 points for each wrong or not given answer.

Question	Your choice A	B	C	D	*Please briefly justify your answers.*
1.1		X			$A\cap B=\phi \Leftrightarrow$ A and B are disjoint
1.2				X	There are 7 possible cats for person 1, 6 possible cats for person 2 .... $\Rightarrow 7\cdot 6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1 = 7!$
1.3	X				$X_1 - X_2 \sim N(0, 2\sigma^2)$ $P(\lvert X_1 - X_2\rvert < 2\sigma) = P(-2\sigma < X_1 - X_2 < 2\sigma) =$ $= P\left(-\frac{2\sigma}{\sqrt{2}\sigma} < Z < \frac{2\sigma}{\sqrt{2}\sigma}\right) = P(-\sqrt{2} < Z < \sqrt{2}) = 2\Phi(\sqrt{2}) - 1$ $\sim Z \sim N(0,1)$
1.4	X				$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \Rightarrow cov(X,Y) = \rho_{X,Y}\sigma_X\sigma_Y = 0.1\cdot\sqrt{3}\cdot\sqrt{3} = 0.3$ $V(X)=3$

Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

21 February 2022

Name and Surname: _____ Matricola: _____

---

1) A chemical test is available for determining whether a certain substance is present in a randomly chosen bottle of milk.
An experiment has a probability of 0.80 of detecting the substance if it is present. The probability of not detecting the substance if it is absent is 0.90. The prior probabilities of the substance being present and being absent are 0.40 and 0.60, respectively.

   a) What is the probability that one experiment results in a detection?
   b) Given that one experiment has resulted in a detection what is the probability that the substance is present?
   c) Three separate experiments which are conducted *on the same bottle of milk* result in two detections (and one non-detection). We assume that the results of the tests are independent. What is the posterior probability that the substance is present?

(6pt)

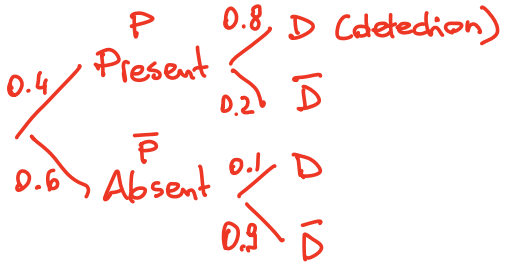
2) For a randomly selected product which is made by two parts (A and B), let X= "the weight of part A" and Y = "the weight of part B". Suppose that the joint pmf of X and Y is given in the table:

p(x,y)	y			
x	0	5	10	15
0	0,02	0,06	0,02	0,1
5	0,04	0,15	0,2	0,1
10	0,01	0,15	0,14	0,01

   a) What is the expected value of the total weight of the product?
   b) Let X_M=max(X,Y). What is the pmf of X_M?
   c) Let X_1 and X_2 denote the weight of part A of two independently selected products. Determine the exact sampling distribution of $X_1 + X_2$.

(6pt)

1)



2) $\mathbb{P}(D) = \mathbb{P}(P)\mathbb{P}(D|P) + \mathbb{P}(\bar{P})\mathbb{P}(D|\bar{P}) = 0.4 \cdot 0.8 + 0.6 \cdot 0.1 =$

$= 0.32 + 0.06 = 0.38$

b) $\mathbb{P}(P|D) = \dfrac{\mathbb{P}(P)\mathbb{P}(D|P)}{\mathbb{P}(D)} = \dfrac{0.4 \cdot 0.8}{0.38} = \dfrac{0.32}{0.38} = \dfrac{16}{19} \simeq 0.842105$

Bayes

c) Y_P number of detections in three tests when the substance is present

$Y_P \sim \text{Bin}(n=3, p=0.8)$

$Y_{\bar{P}}$ number of detections in three tests when the substance is absent

$Y_{\bar{P}} \sim \text{Bin}(n=3, p=0.1)$

$$\mathbb{P}(P \mid \text{"two detections in 3 tests"}) = \frac{\mathbb{P}(P \cap \text{"two det in 3 tests"})}{\mathbb{P}(\text{"two det in 3 tests"})} =$$

$$= \frac{\mathbb{P}(P)\mathbb{P}(Y_P = 2)}{\mathbb{P}(P)\mathbb{P}(Y_P = 2) + \mathbb{P}(\bar{P})\mathbb{P}(Y_{\bar{P}} = 2)} = \frac{0.4 \cdot \binom{3}{2} 0.8^2 \cdot 0.2}{0.4 \cdot \binom{3}{2} 0.8^2 \cdot 0.2 + 0.6 \binom{3}{2} 0.1^2 \, 0.9} =$$

$= 0.904594$

2) a) $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

x	0	5	10
$p_X(x)$	0.2	0.49	0.31
$x\,p_X(x)$	0	2.45	3.1

$\rightarrow \mathbb{E}(X) = 5.55$

y	0	5	10	15
$p_Y(y)$	0.07	0.36	0.36	0.21
$y\,p_Y(y)$	0	1.80	3.6	3.15

$\rightarrow \mathbb{E}(Y) = 8.55$

$\mathbb{E}(X+Y) = 5.55 + 8.55 = 14.10$

b) $X_m = \max(X, Y)$

x_m	0	5	10	15
$p_{X_m}(x_m)$	0.02	0.25	0.52	0.21

$\max(x,y)$ x \ y	0	5	10	15
0	0	5	10	15
5	5	5	10	15
10	10	10	10	15

c)

x_1	x_2	x_1+x_2	$p(x_1,x_2) = p_X(x_1)p_X(x_2)$
0	0	0	$0.2^2 = 0.04$
0	5	5	$0.2 \cdot 0.49 = 0.098$
0	10	10	$0.2 \cdot 0.31 = 0.062$
5	0	5	$0.49 \cdot 0.2 = 0.098$
5	5	10	$0.49^2 = 0.2401$
5	10	15	$0.49 \cdot 0.31 = 0.1519$
10	0	10	$0.2 \cdot 0.31 = 0.062$
10	5	15	$0.49 \cdot 0.31 = 0.1519$
10	10	20	$0.31^2 = 0.0961$

x_1+x_2	0	5	10	15	20
$p_X(x_1+x_2)$	0.04	0.196	0.3641	0.3038	0.0961

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 21/02/2022 (Quiz part 2, VERSION A)

Quiz 1. Given a sequence $\{X_n, n \in \mathbb{N}\}$ of independent and identically distributed random variables such that $X_i \sim Poisson(\lambda)$ (then $\mathbb{E}[X_i] = \lambda$ and $Var(X_i) = \lambda$). Which of the following statement concerning the sample mean $\bar{X}_n$ is false?

(a) $\bar{X}_n$ is a consistent estimator for λ.
(b) $\bar{X}_n$ converges in law to λ.
(c) $\forall \varepsilon > 0 \lim_{n\to\infty} P(\omega : |\bar{X}_n(\omega) - \lambda| < \varepsilon) = 1$
(d) $\bar{X}_n$ is biased for the variance.

Quiz 2. Let us consider a sample (X_1, X_2, X_3) extracted from a population X with expectation θ and variance σ^2. We consider the following two estimators of θ: $\hat{\theta}_1 = X_1 - X_2 + X_3$ and $\hat{\theta}_2 = \frac{X_1}{2} + \frac{X_2}{2}$. Which of the following is true?

(a) $MSE(\hat{\theta}_1) = MSE(\hat{\theta}_2)$
(b) $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$
(c) $MSE(\hat{\theta}_2) < MSE(\hat{\theta}_1)$
(d) $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$

$E[\hat{\theta}_1] = \theta \quad E[\hat{\theta}_2] = \theta$
$Var(\hat{\theta}_1) = 3\sigma^2 \quad Var(\hat{\theta}_2) = \frac{\sigma^2}{2}$
$Var(\hat{\theta}_2) < Var(\hat{\theta}_1)$
$MSE(\hat{\theta}_2) < MSE(\hat{\theta}_1)$

Quiz 3. In the Bayesian setting, which one among the following statement is false?

(a) the point estimate of the mean of a population X is defined as the mean of the random variable having as distribution the posterior distribution of random mean.
(b) the parameter to be estimated is considered a random variable.
(c) the posterior distribution is the maximum likelihood function.
(d) if the posterior distribution is in the same class of distributions of the prior we say that the two families of distributions are conjugate.

Quiz 4. Given the sample data $(X_1, \ldots, X_{100})$ where X_i are i.i.d. and $X_i \sim Bernoulli(p)$, let us suppose that $\hat{p} = \frac{\sum_{i=1}^{100} X_i}{100} = 0.8$. A confidence interval for $\hat{p}$ with confidence level 90% is given by

(a) $[0.8 - t_{0.95} \times 0.06, 0.8 + t_{0.95} \times 0.06]$
(b) $[0.8 - z_{0.95} \times 0.04, 0.8 + z_{0.95} \times 0.04]$
(c) $[0.8 - z_{0.95}\sqrt{\frac{0.8}{100}}, 0.8 + z_{0.95}\sqrt{\frac{0.8}{100}}]$
(d) $[0.8 - z_{0.95}0.09, 0.8 + z_{0.95}0.09]$

$I_{1-\alpha} = \left[\hat{p} - t_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + t_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 21/02/2022 (Exercises part 2, VERSION A)

Exercise 1). Let $X_1, \dots, X_n$ i.i.d. with density function

$$f(x;\theta) = 1 + \theta\left(x - \frac{1}{2}\right), 0 < x < 1,$$

and $-1 < \theta < 1$.

(a) Find an estimator for θ using the Method of the Moments. Is it unbiased?

(b) Is it consistent? What is its limiting distribution?
hint: use the Central Limit Theorem for the sample mean.

Exercise 2). Let us suppose we have a random sample of 36 drinks coming from a vending machine which cointains in mean 7.4 cl and the estimate of the standard deviation is 0.48 cl. Suppose that the sample is normally distributed.

(a) Provide a confidence interval for the mean with confidence 99%.

(b) Test the hypothesis $H_0 : \mu_0 = 7.5$ versus the alternative hypothesis $\mu_1 < 7.5$ at a significance level 95%. Do we reject H_0?

Sol. ex.1

$$E[X] = \int_0^1 x\left(1+\theta\left(x-\tfrac{1}{2}\right)\right)dx = \int_0^1 x + \theta x^2 - \frac{1}{2}\theta x \, dx$$

$$= \frac{x^2}{2} + \frac{\theta x^3}{3} - \frac{\theta}{2}\frac{x^2}{2}\Big|_0^1 = \frac{1}{2} + \frac{\theta}{3} - \frac{\theta}{4} = \frac{1}{2} + \frac{\theta}{12}$$

$$\Rightarrow \bar{X}_n = \frac{1}{2} + \frac{\hat{\theta}}{12} \Rightarrow \hat{\theta} = \left(\bar{X}_n - \frac{1}{2}\right)12$$

$$\hat{\theta} = 12\bar{X}_n - 6$$

Is it unbiased? $E[\hat{\theta}] = 12E[\bar{X}_n] - 6$

$$= 12\left(\frac{1}{2} + \frac{\theta}{12}\right) - 6$$

$$= 6 + \theta - 6 = \theta \quad \text{yes!}$$

Consistent?

For the Law of Large Numbers we have that

$$\bar{X}_n \xrightarrow{p} E[X] = \frac{1}{2} + \frac{\theta}{12}$$

$$\Rightarrow 12\bar{X}_n - 6 \xrightarrow{p} 12\left(\frac{1}{2} + \frac{\theta}{12}\right) - 6$$

$$= \theta$$

$$\Rightarrow \hat{\theta} \xrightarrow{p} \theta \quad \text{yes!}$$

Limiting distribution:

The Central Limit theorem implies that $\frac{S_n - nE[X]}{\sqrt{n\,\mathrm{Var}(X)}} \xrightarrow{d} Z$

where $Z \sim N(0,1)$ and

$S_n := X_1 + \dots + X_n$.

Then $\bar{X}_n \xrightarrow{d} N\left(E[\bar{X}], \frac{\mathrm{Var}(X)}{n}\right)$

We need to compute the variance.

$$E[X^2] = \int_0^1 x^2\left(1+\theta x - \frac{\theta}{2}\right)dx$$

$$= \frac{x^3}{3} + \frac{\theta x^4}{4} - \frac{\theta x^3}{6}\Big|_0^1$$

$$= \frac{1}{3} + \frac{\theta}{4} - \frac{\theta}{6} = \frac{1}{3} + \frac{\theta}{12}$$

$$\mathrm{Var}(X) = \frac{1}{3} + \frac{\theta}{12} - \left(\frac{1}{2} + \frac{\theta}{12}\right)^2$$

$$= \frac{1}{3} + \frac{\theta}{12} - \frac{1}{4} - \frac{\theta^2}{144} - \frac{\theta}{12} = \frac{1}{12} - \frac{\theta^2}{144}$$

- Note that the variance must be positive: $\frac{1}{12} - \frac{\theta^2}{144} \geq 0 \iff \theta^2 \leq 12$

$$\iff -2\sqrt{3} \leq \theta \leq 2\sqrt{3}$$

Since $\theta \in (-1, 1)$ this condition is satisfied.

- Note also that to apply the Law of large Numbers and the Central Limit Theorem we needed to check that $\mathrm{Var}(X_i) < \infty$.

Finally we get:

$$\bar{X}_n \xrightarrow{d} N\left(\frac{1}{2}, \frac{\theta}{12}, \frac{1}{12n}\left(1 - \frac{\theta^2}{12}\right)\right)$$

$$12\bar{X}_n - 6 \xrightarrow{d} N\left(\theta, \left(1 - \frac{\theta^2}{12}\right)\frac{12}{n}\right)$$

$$\hat{\theta} \xrightarrow{d} N\left(\theta, \frac{12 - \theta^2}{n}\right)$$

EX 2

a) $1-\alpha = 0.99 \quad \alpha = 0.01$

$\frac{\alpha}{2} = 0.005 \quad 1-\frac{\alpha}{2} = 0.995$

$t_{35,\,0.995} = 2.724$

$$CI = \left[\bar{X}_m - t_{35,\,0.995} \times \frac{0.48}{6} ,\ \bar{X}_m + t_{35,\,0.995} \times \frac{0.48}{6} \right]$$

$$= \left[7,182 ;\ 7,618 \right]$$

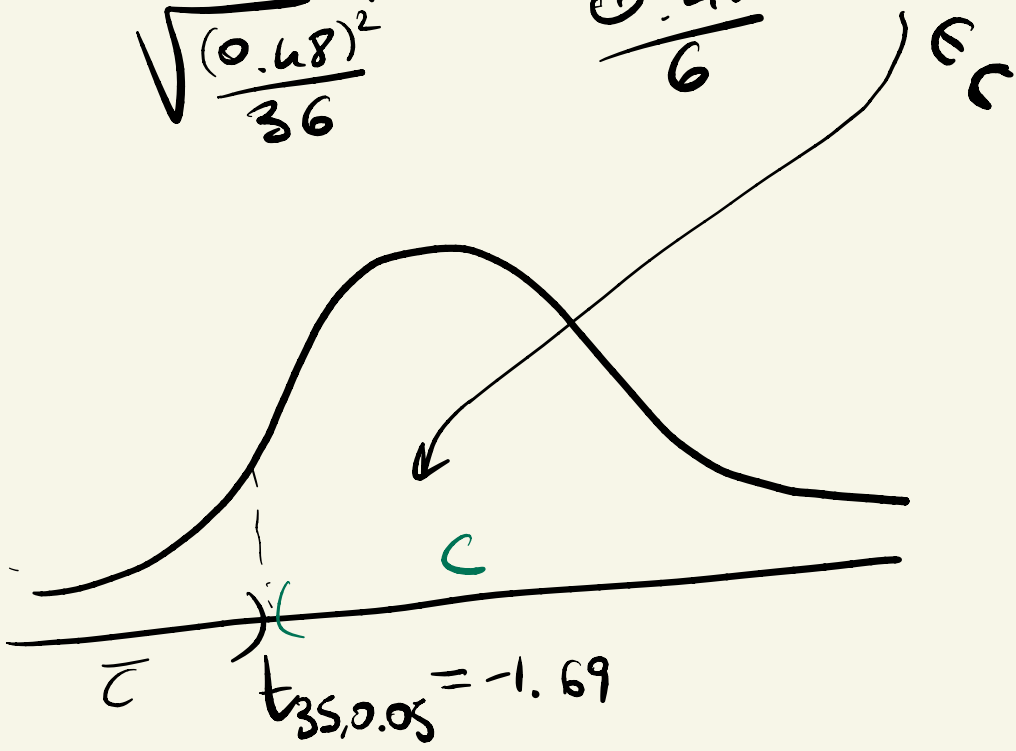
b) $H_0 : \mu_0 = 7.5 \quad H_1 : \mu < 7.5$

$1-\alpha = 0.95 \Rightarrow \alpha = 0.05$

this is a unilateral test

$t_{35,0.05} = -1.69$

$$T = \frac{\bar{x}_m - \mu_0}{\sqrt{\frac{(0.48)^2}{36}}} = \frac{7.4 - 7.5}{\frac{0.48}{6}} \simeq -1,25$$



we do not reject Ho!

Version A

Statistical Methods in Data Science

Master course in

DATA SCIENCE AND ENGINEERING

17 June 2022

Name and Surname: _____ Matricola: _____

QUESTIONS

1.1 A fair die is rolled till the event A="the number of dots in the upper face is 1" occurs. Let X be defined as the number of trials to be done till A occurs (and then the game is stopped). The probability that X=2 is

A	P(X = 2) = 1/36	B	P(X = 2) = 5/36	C	P(X=2)=1/216	D	P(X=2)=5/216

1.2 There are 7 cats. Three (different) cats must be randomly selected for testing a special diet: the first selected cat will be fed according to "DIET_A", the second one according to "DIET_B", the third one according to "DIET_C". How many different selections are there?

A	$7!$	B	$7 \cdot 6 \cdot 5$	C	7^3	D	$7 \cdot 5$

1.3 Let $X \sim N(\mu, 9\sigma^2)$, $Y \sim N(\mu, 16\sigma^2)$, X and Y independent, and Φ the cdf of a standard normal rv. The probability p that the absolute value of the difference between $X + Y$ and its mean μ_{X+Y} is less than 10 is

A	B	C	D
$p = 2\Phi(2) - 1$	p depends on μ, σ	$p = 2\Phi(\sqrt{2}) - 1$	$p = 2\Phi(2/\sigma) - 1$

1.4 Let X and Y be two Poisson random variables with mean E[X]=E[Y]=3 and correlation $\rho(X,Y)=0.1$. Then the variance of X+Y is

A	6	B	6.6	C	5.4	D	6.3

ANSWERS

The score for each correct answer is 1 point, 0 points for each wrong or not given answer.

Question	Your choice A	B	C	D	*Please briefly justify your answers.*
1.1		X			$P(X=2) = P(\bar{A}_1 \cap A_2) = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}$
1.2		X			There are 7 choices for the first cat, 6 choices for the second and 5 for the third $\Rightarrow 7 \cdot 6 \cdot 5$
1.3				X	$\mu_{X+Y} = E(X+Y) = E(X)+E(Y) = 2\mu$ $V(X+Y) = V(X)+V(Y) = 25\sigma^2$ $P(\lvert X+Y-2\mu \rvert < 10) = P(-10 < X+Y-2\mu < 10) = P(-\frac{10}{5\sigma} < Z < \frac{10}{5\sigma}) = 2\Phi(\frac{2}{\sigma})+1$ $Z \sim N(0,1)$
1.4		X			$V(X+Y) = V(X)+V(Y)+2\mathrm{cov}(X,Y) = 3+3+2\rho(X,Y)\sigma_X\sigma_Y = 6+2 \cdot 0.1 \cdot 3 = 6.6$

Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

17 June 2022

Name and Surname: _____ Matricola: _____

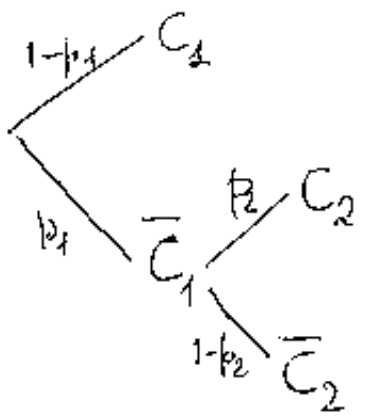
1) Let p_1 denote the probability that any particular code symbol is erroneously transmitted through a communication system. Assume that on different symbols, errors occur independently of one another. Suppose also that with probability p_2 an erroneous symbol is corrected upon receipt. Let X denote the number of correct symbols in a message block consisting of n symbols (after the correction process has ended).
   a) What is the probability that X=0?
   b) What is the probability distribution of X?
   c) What is the expected value of X?

(6pt)

2) Let X denote the vehicle speed on a road. Let us make the hypothesis that X is normally distributed. Assume that speeds are independent.
   a) If 5% of all vehicles travel less than 39.12 m/h and 10% travel more than 73.24 m/h, what are the mean and standard deviation of vehicle speed?
   b) What is the probability that the difference between the speed of two randomly selected vehicles is less than 16.5 m/h?
   c) What is the probability that among 10 randomly selected vehicle exactly eight have a vehicle speed between 39.12 m/h and 73.24 m/h?

(6pt)

A.1



C_1 = symbol is transmitted in a right way

$\overline{C_1}$ = symbol is erroneously transmitted

C_2 = ~~symbol~~ an erroneous symbol is corrected

$\overline{C_2}$ = an erroneous symbol is not corrected

$$\mathbb{P}(\text{symbol is correct}) = \mathbb{P}(C_1 \cup C_2) = \mathbb{P}(C_1) + \mathbb{P}(C_2) =$$

$$= 1 - p_1 + p_1 p_2$$

X = number of correct symbols in a message block of n symbols

b) $X \sim \text{Bin}(n, p = 1 - p_1 + p_1 p_2)$ ✓

a) $\mathbb{P}(X = 0) = (p_1 - p_1 p_2)^n$ ✓

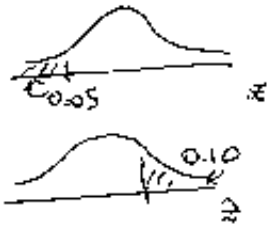
c) $\mathbb{E}(X) = n(1 - p_1 + p_1 p_2)$ ✓

A.2

a) $X \sim N(\mu, \sigma^2)$

$Z \sim N(0,1)$

$$\begin{cases} P(X < 39.12) = 0.05 \\ P(X > 73.24) = 0.10 \end{cases} \rightarrow \begin{cases} P\left(\frac{X-\mu}{\sigma} < \frac{39.12-\mu}{\sigma}\right) = 0.05 \\ P\left(\frac{X-\mu}{\sigma} > \frac{73.24-\mu}{\sigma}\right) = 0.10 \end{cases}$$



$$\begin{cases} \frac{39.12-\mu}{\sigma} \cong -1.645 \\ \frac{73.24-\mu}{\sigma} \cong 1.28 \end{cases} \rightarrow \begin{cases} \mu = 39.12 + 1.645\sigma \\ 73.24 - 39.12 - 1.645\sigma = 1.28\sigma = 0 \end{cases}$$

$$\begin{cases} \dots \\ 2.925\sigma = 39.12 \end{cases} \rightarrow \begin{cases} \mu \cong 58.309 \\ \sigma \cong 11.665 \end{cases}$$

b) $D = X_1 - X_2 \sim N(0, 2\sigma^2) \equiv N(0, 272.1445)$

$$P(|D| < 16.5) = P\left(-\frac{16.5}{\sqrt{272.1445}} < Z < \frac{16.5}{\sqrt{272.1445}}\right) \cong 2\Phi(1) - 1 =$$

$$= 2 \cdot 0.8413 - 1 = 0.6826$$

c) Y = number of vehicles with $39.12 < \text{speed} < 73.24$ among ten randomly selected

$Y \sim Bin(n = 10, p = 0.85)$

$$P(Y = 8) = \binom{10}{8} 0.85^8 \, 0.15^2 = 45 \cdot 0.85^8 \cdot 0.15^2 \cong 0.2759$$

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 17/06/2022 Version A (Quiz part 2)

Quiz 1. Let us consider a random sample $(X_1, \ldots, X_n)$ extracted from a population with density

$$f(x,p) = \begin{cases} \frac{1-p}{6} & x \in A = \{2,4,6\} \\ \frac{1+p}{6} & x \in B = \{1,3,5\} \end{cases}$$

with $p \in (0,1)$. The Likelihood function is

(a) $L(p; x_1, \ldots, x_n) = \left(\frac{1-p}{6}\right)^{\sum_{i=1}^n \mathbb{I}_A(x_i)} \left(\frac{1+p}{6}\right)^{\sum_{i=1}^n \mathbb{I}_B(x_i)}$

(b) $L(p; x_1, \ldots, x_n) = \left(\frac{1-p}{6}\right)^{2+4+6} \left(\frac{1+p}{6}\right)^{1+3+5}$

(c) $L(p; x_1, \ldots, x_n) = \left(\frac{1-p}{6}\right)^{3} \left(\frac{1+p}{6}\right)^{3}$

(d) $L(p; x_1, \ldots, x_n) = \left(\frac{1-p}{6}\right)^{\sum_{i=1}^n x_i} \left(\frac{1+p}{6}\right)^{\sum_{i=1}^n x_i}$

Quiz 2. Which of the following options is NOT correct for Cramer-Rao Theorem?

(a) Cramer-Rao inequality provides a lower bound to the variance of an unbiased estimator for a parameter

(b) The theorem requires the computation of the expected value of a specific function.

(c) The parameter space can have only positive values.

(d) The derivative of $\log L(x;\theta)$ exists for all θ

Quiz 3. If in a test the null hypothesis is rejected at the level of 2%, then

(a) it is not rejected at level 5%

(b) it is rejected also at level 5%

(c) the p-value=0.02

(d) none of the previous statements is correct.

Quiz 4. Let us consider a random sample of size $n < 30$, extracted from a normally distributed population with mean μ and unknown standard deviation σ. In the test $H_0 : \mu \geq 5$ versus $H_1 : \mu < 5$ at level α, we reject the null hypothesis if:

(a) $t < t_{n-1,1-\alpha}$

(b) $t < -t_{n-1,1-\alpha}$

(c) $|z| < z_{1-\alpha/2}$

(d) $z < -z_{\alpha}$

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 17/06/2022 (Exercises part 2)

Exercise 1). Let $(X_1, \dots, X_n)$ a random sample extracted from a population with density function

$$f_X(x;\theta) = \frac{2}{\theta}(1 - \frac{x}{\theta})I_{[0,\theta]}(x), \quad \theta > 0.$$

(a) Verify that $\mathbb{E}_\theta[X] = \frac{\theta}{3}$. Find an **unbiased** estimator of θ, which is a function of the sample mean $\bar{X}_n$.

(b) Compute the variance of the estimator found in the previous point and study its consistency.

Exercise 2). Let $(X_1, \dots, X_n)$ a random sample of dimension $n = 3$ extracted from a population which has a Poisson distribution of unknown parameter θ, namely with probability distribution:

$$f_X(x;\theta) = e^{-\theta}\frac{\theta^x}{x!} \quad x = 0, 1, 2, \dots \quad \theta > 0.$$

Let us consider the following hypothesis test:
$H_0 : \theta = \theta_0 = 2, \qquad H_1 : \theta = \theta_1 = 1.$
The rejection region of a likelihood ratio test is such that

$$\alpha = \mathbb{P}[\lambda(\bar{X}) < \lambda^* | H_0 \text{ is true}].$$

(a) Write explicitly the test statistic $\lambda(\bar{X})$.

(b) Now suppose you know that the rejection region can be written as

$$R = \{(x_1, x_2, x_3) : \sum_{i=1}^{3} x_i \leq k\}.$$

Set $k = 1$. Compute the probability of the first type error α and $1 - \beta$ (where β is the probability of the second type error).

hint: Note that in this case the distribution of $\sum_{i=1}^{n} x_i$ is known and then it is not necessary to use the asymptotic distribution for big n.

Quiz 1: (a) 2. (c) 3. (b) 4. (b)

EX 1 a) $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x \frac{2}{\theta}\left(1-\frac{x}{\theta}\right) I_{[0,\theta]}(x)\,dx = \int_0^\theta x \frac{2}{\theta}\left(1-\frac{x}{\theta}\right)dx = \frac{2}{\theta}\int_0^\theta x - \frac{x^2}{\theta}\,dx$

$= \frac{2}{\theta}\left(\frac{x^2}{2} - \frac{x^3}{3\theta}\right)_0^\theta = \frac{2}{\theta}\left(\frac{\theta^2}{2} - \frac{\theta^3}{3\theta}\right) = \frac{2}{\theta}\frac{\theta^2}{6} = \boxed{\frac{\theta}{3}}$

$\Rightarrow$ It follows that $\bar{X}_n$ is such that $\mathbb{E}[\bar{X}_n] = \frac{\theta}{3}$

Then an unbiased estimator which is function of the sample mean is $\boxed{\hat{\theta} = 3\bar{X}_n}$ (Indeed: $\mathbb{E}[\hat{\theta}] = 3\mathbb{E}[\bar{X}_n] = 3\frac{\theta}{3} = \theta$)

b) $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(3\bar{X}_n) = 9\,\mathrm{Var}(\bar{X}_n) = 9\frac{\sigma^2}{n}$ where $\sigma^2 = \mathrm{Var}(X)$

We look for σ^2:

$\sigma^2 = \mathrm{Var}(X) = E[X^2] - (E[X])^2$

$E[X^2] = \int_0^\theta x^2 \frac{2}{\theta}\left(1-\frac{x}{\theta}\right)dx = \frac{2}{\theta}\int_0^\theta x^2 - \frac{x^3}{\theta}\,dx = \frac{2}{\theta}\left(\frac{x^3}{3} - \frac{x^4}{4\theta}\right)\Big|_0^\theta = \frac{2}{\theta}\left(\frac{\theta^3}{3} - \frac{\theta^4}{4\theta}\right)$

$= \frac{2}{\theta}\frac{\theta^3}{12} = \frac{\theta^2}{6} \Rightarrow \mathrm{Var}(X) = \frac{\theta^2}{6} - \left(\frac{\theta}{3}\right)^2 = \frac{\theta^2}{18} \Rightarrow \sigma^2 = \frac{\theta^2}{18}$

$\Rightarrow \mathrm{Var}(\hat{\theta}) = \frac{9}{n}\frac{\theta^2}{18} = \frac{\theta^2}{2n}$ $\boxed{\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{2n}}$

Consistency: (1) way: $\bar{X}_n \xrightarrow{p} \frac{\theta}{3}$ for the Law of Large Numbers.

$\Rightarrow 3\bar{X}_n \xrightarrow{p} \theta \Rightarrow \hat{\theta} \xrightarrow{p} \theta$ it is consistent.

(2) way: $\begin{cases} \hat{\theta} \text{ unbiased} \\ \mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{2n} \xrightarrow[n\to\infty]{} 0 \end{cases}$ $\Rightarrow$ consistency because of Chebyshev inequality

Indeed: $\mathbb{P}(|\hat{\theta}-\theta|<\varepsilon) = \mathbb{P}(|\hat{\theta} - \mathbb{E}[\hat{\theta}]|<\varepsilon) \geq 1 - \frac{\mathbb{E}[(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2]}{\varepsilon^2}$

$= 1 - \frac{\mathrm{Var}(\hat{\theta})}{\varepsilon^2} = 1 - \frac{\theta^2}{2n\varepsilon^2} \xrightarrow[n\to\infty]{} 1$.

Then $1 \geq \mathbb{P}(|\hat{\theta}-\theta|<\varepsilon) \geq 1 \Rightarrow \mathbb{P}(|\hat{\theta}-\theta|<\varepsilon) \xrightarrow[n\to+\infty]{} 1$.

EX 2

a) $\Lambda(\bar{x}) = \dfrac{\sup_{\theta} L(\theta|\bar{x})}{\sup_{\theta} L(\theta|\bar{x})} = \dfrac{e^{-n\theta_0}\theta_0^{\sum_{i=1}^n x_i}}{e^{-n\theta_1}\theta_1^{\sum_{i=1}^n x_i}} = e^{-n(\theta_0-\theta_1)}\left(\dfrac{\theta_0}{\theta_1}\right)^{\sum_{i=1}^n x_i}$

$n=3 \quad \Lambda(\bar{x}) = e^{-3(\theta_0-\theta_1)}\left(\dfrac{\theta_0}{\theta_1}\right)^{\sum_{i=1}^3 x_i}$

b) $k=1$

$\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) = \mathbb{P}\left(\sum_{i=1}^3 X_i \le 1 \,\middle|\, \theta=2\right)$ (*)

Note that $\sum_{i=1}^3 X_i = X_1+X_2+X_3$ is a sum of [illegible] r.v. which are $\text{Pois}(\theta) \Rightarrow Y = X_1+X_2+X_3 \sim \text{Poisson}(3\theta)$

$\Rightarrow$ (*) $= \mathbb{P}(Y \le 1 | \theta=2) = \mathbb{P}(Y=0,\theta=2) + \mathbb{P}(Y=1,\theta=2)$

$= e^{-3\cdot 2}\dfrac{(3\cdot 2)^0}{0!} + e^{-2\cdot 3}\dfrac{(3\cdot 2)^1}{1!} = e^{-6} + 6e^{-6} = 0.017$

$\beta = \mathbb{P}(\text{accept } H_0 | H_0 \text{ false})$

$1-\beta = \mathbb{P}(\text{reject } H_0 | H_0 \text{ false}) = \mathbb{P}\left(\sum_{i=1}^3 X_i \le 1 \,\middle|\, \theta=1\right)$

$= \mathbb{P}(Y \le 1 | \theta=1) = \mathbb{P}(Y=1,\theta=1) + \mathbb{P}(Y=0,\theta=1)$

$= e^{-3} + 3e^{-3} = 0.2$

Version A

Statistical Methods in Data Science

Master course in

DATA SCIENCE AND ENGINEERING

12 September 2022

Name and Surname: _____ Matricola: _____

QUESTIONS

1.1 A fair die is rolled once. Let A="the number of dots in the upper face is 1" and B="the number of dots in the upper face is 2". Which is the false statement?

A	A and B are independent	B	$P(A \cap B) = 0$	C	$P(A\|B) = 0$	D	$P(A \cup B) = 1/3$

1.2 There are 4 cats (C_1,C_2,C_3, and C_4), 4 dogs (D_1,D_2,D_3, and D_4), 4 elephants (E_1,E_2,E_3, and E_4), and 4 giraffes (G_1,G_2,G_3, and G_4). How many different 4-element vectors (C_i,D_j,E_k,G_m) can be generated? Please note that only cats can be put in the first position, only dogs in the second position, only elephants in the third position, and only giraffes in the fourth position.

A	16!	*B*	$4 \cdot 3 \cdot 2 \cdot 1$	*C*	4^4	*D*	4! 4!

1.3 Let $X \sim N(\mu_1, \sigma_1^2 = 36)$, $Y \sim N(\mu_2, \sigma_2^2 = 64)$, X and Y independent, and Φ the cdf of a standard normal rv. The probability p that the absolute value of the difference between $X + Y$ and its mean μ_{X+Y} is less than 10 is

A	B	C	D
$p = 2\Phi(1) - 1$	p depends on μ_1, μ_2, σ	$p = 2\Phi(\sqrt{2}) - 1$	$p = 2\Phi(2) - 1$

1.4 Let X and Y be two exponential random variables with mean E[X]=E[Y]=3 and correlation ρ(X,Y)=0.1. Then the variance of X+Y-4 is

A	15.8	B	14	C	18	D	19.8

ANSWERS

The score for each correct answer is 1 point, 0 points for each wrong or not given answer.

	Your choice				*Please briefly justify your answers.*
Question	A	B	C	D	
1.1	X				The false statement is A and B independent. $P(A \cap B) = P(\emptyset) = 0$, $P(A\|B) = \frac{P(A \cap B)}{P(B)} = \frac{0}{1/6} = 0$, $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/6 + 1/6 = 2/6 = 1/3$
1.2			X		(C_i, D_j, E_k, G_m) There are $4 \times 4 \times 4 \times 4 = 4^4$ possible different vectors
1.3	X				$X+Y \sim N(\mu_1+\mu_2, 10^2)$ $P(\|X+Y-(\mu_1+\mu_2)\| < 10) =$ $= P(-10 < X+Y-(\mu_1+\mu_2) < 10) = P(-1 < Z < 1) = 2\Phi(1) - 1$
1.4				X	$V(X+Y-4) = V(X+Y) = V(X) + V(Y) + 2\,cov(X,Y)$, $Cov(X,Y) = \rho \sigma_X \sigma_Y$ $V(X) = 3^2 = V(Y)$ $Cov(X,Y) = 0.1 \cdot 3^2$ $\Rightarrow V(X+Y-4) = 2 \cdot 9 + 2 \cdot 0.1 \cdot 9 = 19.8$

Version A

Statistical Methods in Data Science

Master course in
DATA SCIENCE AND ENGINEERING

12 September 2022

Name and Surname: _____ Matricola: _____

1) The distribution of the diameter X of a certain type of objects is

$$f(x) = p\, f_1(x; \mu_1, \sigma) + (1 - p)\, f_2(x; \mu_2, \sigma)$$

where f_1 e f_2 are normal probability density functions and p=0.35, $\mu_1 = 4.4, \mu_2 = 5.0$, and $\sigma = 0.27$.

a) What is the expected value of X?
b) What is the probability that the diameter is between 4.4 and 5?
c) Assuming that the diameters of different items are independent of one another what is the probability that among n=10 randomly chosen items exactly two have the diameter between 4.4 and 5?

(6pt)

2) Consider a communication source that transmits packets containing digitized speech. After each transmission, the receiver sends a message indicating whether the transmission was successful or unsuccessful. If a transmission is unsuccessful, the packet is re-sent. Suppose a voice packet can be transmitted a *maximum* of 10 times. Assuming that the results of successive transmissions are independent of one another and that the probability of any particular transmission being successful is p,

1. determine the probability mass function of the random variable X = the number of times a packet is transmitted.
2. determine the probability $p(X_1 + X_2 = 4)$, i.e. the probability that the total number of transmissions required to send two packets is 4

(6pt)

1)

$X_1 \sim N(4.4, 0.27^2)$
$X_2 \sim N(5, 0.27^2)$
$p = 0.35$

a) $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x)\,dx =$

$= \int_{-\infty}^{\infty} x \left[p f_1(x) + (1-p) f_2(x) \right] dx =$

$= p \int_{-\infty}^{\infty} x f_1(x) dx + (1-p) \int_{-\infty}^{\infty} x f_2(x)\,dx =$

$= p\mu_1 + (1-p)\mu_2 = 0.35 \cdot 4.4 + 0.65 \cdot 5 = 1.54 + 3.25 = 4.79$

b) $\mathbb{P}(4.4 < X < 5) = \int_{4.4}^{5} f(x)\,dx =$

$= \int_{4.4}^{5} \left(p f_1(x) + (1-p) f_2(x) \right) dx =$

$= p \int_{4.4}^{5} f_1(x) dx + (1-p) \int_{4.4}^{5} f_2(x)\,dx =$

$= p\ \mathbb{P}(4.4 < X_1 < 5) + (1-p)\ \mathbb{P}(4.4 < X_2 < 5) =$ $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$

$= p\ \mathbb{P}\left(\frac{4.4-4.4}{0.27} < Z_1 < \frac{5-4.4}{0.27} \right) + (1-p)\ \mathbb{P}\left(\frac{4.4-5}{0.27} < Z_2 < \frac{5-5}{0.27} \right) =$

$\simeq p\ \mathbb{P}(0 < Z_1 < 2.22) + (1-p)\ \mathbb{P}(-2.22 < Z_2 < 0) =$ $\mathbb{P}(0<Z_1<2.22) = \mathbb{P}(-2.22<Z_2<0)$

$= \mathbb{P}(0 < Z_1 < 2.22) \left[p + (1-p) \right] = \mathbb{P}\left[0 < Z_1 < 2.22 \right] = \phi(2.22) - 0.5 =$

$= 0.9868 - 0.5 = 0.4868$

1/

9) Y = number of items with diameter between 4.4 and 5 among the ten randomly selected

$Y \sim \text{Binomial}(n=10, p=0.4868)$

$$P(Y=2) = \binom{10}{2}(0.4868)^2(1-0.4868)^8 = 45 \cdot 0.4868^2\, 0.5132^8 = \approx 0.051311$$

2) X = number of times a packet is transmitted

a)

sequences (s = success, f = failure)	x	$p(x)$
s	1	p
f, s	2	$(1-p)p$
f, f, s	3	$(1-p)^2 p$
f, f, f, s	4	$(1-p)^3 p$
...		
f, f, ... f, s (8 times)	9	$(1-p)^9 p$
f, ... f, s (9 times) or f, ... f, f (10 times)	10 ($x=10$ regardless of the 10th outcome)	$(1-p)^9$

Summarizing

$$p(x) = \begin{cases} (1-p)^{x-1} p & x = 1, \dots, 9 \\ (1-p)^9 & x = 10 \\ 0 & \text{otherwise} \end{cases}$$

b) $$\mathbb{P}(X_1 + X_2 = 4) = \sum_{x_1=1}^{3} \mathbb{P}(X_1 = x_1, X_2 = 4 - x_1) =$$

$$= \sum_{x_1=1}^{3} \mathbb{P}(X_1 = x_1)\,\mathbb{P}(X_2 = 4 - x_1) = p(1-p)^2 p + (1-p)^2 p^2 + (1-p)^2 p\, p =$$

($\uparrow$ x_1, x_2 indep)

$$= 3p^2(1-p)^2$$

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 12/09/2022 (Quiz part 2)

Quiz 1. Given $(X_n)_n$ a sequence of random variable which has distribution $Exp(\lambda)$. Then

(a) $\frac{S_n - \frac{n}{\lambda}}{\sqrt{\frac{n}{\lambda}}}$ converges in probability to λ.

(b) $\frac{S_n - n\lambda}{\sqrt{n\lambda}}$ converges in distribution to a normal random variable $N(0,1)$.

(c) $\frac{S_n - \frac{n}{\lambda}}{\sqrt{\frac{n}{\lambda^2}}}$ converges in distribution to a normal random variable $N(0,1)$.

(d) $\frac{S_n - \frac{n}{\lambda}}{\sqrt{\frac{n}{\lambda^2}}}$converges in probability to 0.

Quiz 2. Let X be a population having an Exponential distribution with unknown parameter λ. Given the sample $(X_1; X_2; X_3) = (2; 3; 5)$ then the Likelihood function for λ is:

(a) $\lambda^{10} e^{-3\lambda}$

(b) $\lambda^3 e^{-10\lambda}$

(c) $\dfrac{e^{-10\lambda}}{\lambda^3}$

(d) $\dfrac{e^{-3\lambda}}{\lambda^{10}}$

Quiz 3. In the confidence interval for the true value of the mean of one Gaussian population, the Student's t distribution is used

(a) if σ^2 is known

(b) if n is big

(c) if σ^2 is unknown

(d) always

Quiz 4. In a statistical test, the probability of rejecting the null hypothesis when it is true is the probability of

(a) not making an error of the first type

(b) making an error of the second type

(c) none of the other options is correct

(d) making an error of the first type

Name and Surname:

Matricola:

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 12/09/2022 (Exercises part 2)

Exercise 1). Let $\mathbf{X} = (X_1, \ldots, X_n)$, $n \geq 4$, be a random sample with *unknown* mean value $E[X_i] = \mu$ and variance $V[X_i] = \sigma^2$ *known* and finite.

Let us consider the class of estimators of μ obtained considering a generic linear combination of the random variables X_i :

$$T(\mathbf{X}) = \sum_{i=1}^{n} a_i X_i, \qquad a_i \in \mathbb{R}.$$

(a) Determine the expected value of $T(\mathbf{X})$ and establish the condition on the coefficients a_i so that the estimator is unbiased.

(b) Determine the expression of the $MSE[T(\mathbf{X})]$ for an unbiased estimator of the previous point.

(c) Write the expression of the two estimators $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$, obtained from $T(\mathbf{X})$ setting:

$$a_i = \frac{1}{n}, \qquad \text{for } i = 1, \ldots n \qquad \text{for } T_1(\mathbf{X})$$

$$a_1 = \frac{n-2}{n}, \quad a_2 = a_3 = \frac{1}{n} \text{ and } a_4 = \ldots a_n = 0 \qquad \text{for } T_2(\mathbf{X}).$$

(d) Prove that the two estimators $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are unbiased. Determine the expression of $MSE[T_1(\mathbf{X})]$ and $MSE[T_2(\mathbf{X})]$ and establish which one is more efficient.

Exercise 2). A sample survey of $n = 200$ customers of a restaurant revealed that 18 are dissatisfied with the service.

(a) Give a confidence interval of level $1 - \alpha = 0.90$ for the proportion of dissatisfied costumers.

Hint: the population of the sample $\sim$ Bernoulli(p), with p =proportion of dissatisfied costumers.

(b) To verify if the true proportion of dissatisfied customers can be considered equal to 0.1 or if it is less, test the hypothesis

$$H_0 : p = p_0 = 0.10 \quad \text{versus} \quad H_1 : p = p_1 < 0.10$$

at a level of significance of $\alpha = 0.02$. (Note that this is a unilateral test).

Hint: Use the Central Limit Theorem to approximate $\hat{p}$ under H_0

Solutions

1. (c) 2. (b) 3. (c) 4. (d)

EX. 1

a) $\mathbb{E}[T(\mathbf{X})] = \mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$

$= \mu \sum_{i=1}^{n} a_i$

T is unbiased if $\mathbb{E}[T(\mathbf{X})] = \mu$

then if $\mu \sum_{i=1}^{n} a_i = \mu$

then if $\boxed{\sum_{i=1}^{n} a_i = 1}$

b) We suppose T unbiased $\Rightarrow MSE(T) = Var(T)$

$MSE(T) = Var(T) = Var\left(\sum_{i=1}^{n} a_i X_i\right)$

$= \sum_{i=1}^{n} a_i^2 Var(X_i) = \sigma^2 \sum_{i=1}^{n} a_i^2$

c) $T_1(\mathbf{X}) = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X}_n$ the sample mean

$T_2(\mathbf{X}) = \frac{n-2}{n} X_1 + \frac{1}{n} X_2 + \frac{1}{n} X_3 = \frac{(n-2)X_1 + X_2 + X_3}{n}$

d) The coefficients $\{a_i\}_{i=1,\dots,n}$ of T_1 and T_2 satisfy condition in point a) and then T_1 and T_2 are unbiased.

Indeed for T_1: $\sum_{i=1}^{n} 1/n = 1$

for T_2: $\sum_{i=1}^{n} a_i = \frac{n-2}{n} + \frac{1}{n} + \frac{1}{n} + 0 + \dots + 0 = 1$

We can use the formula for the MSE found in point (b).

$$MSE(T_1) = \sigma^2 \sum_{i=1}^{n} \frac{1}{n^2} = \frac{\sigma^2}{n}$$

$$MSE(T_2) = \sigma^2 \left(\frac{(n-2)^2}{n^2} + \frac{1}{n^2} + \frac{1}{n^2} \right) = \sigma^2 \frac{n^2 - 4n + 6}{n^2}$$

$$MSE(T_1) < MSE(T_2)$$

T_1 is more efficient than T_2.

EX.2 $n = 200$ P = proportion of dissatisfied customers

(a) $1-\alpha = 0.90 \Rightarrow \alpha = 1-0.90 = 0.10$

$$\hat{p} = \frac{18}{200} = 0.09$$

$$I_{1-\alpha} = \left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

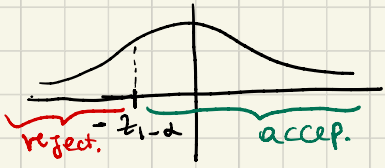
$$z_{1-\alpha/2} = z_{1-\frac{0.10}{2}} = z_{0.95} = 1.64$$

$$I_{(90\%)} = \left[0.09 - 1.64\sqrt{\frac{0.09(1-0.09)}{200}},\ 0.09 + 1.64\sqrt{\frac{0.09(1-0.09)}{200}}\right]$$

$$= (0.06, 0.12)$$

(b) $H_0: p = p_0 = 0.10$

$H_1: p < 0.10$



$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.09 - 0.10}{\sqrt{0.10(1-0.10)/200}} = -0.47$$

we reject H_0 if $z < -z_{1-\alpha}$

$z_{1-\alpha} = z_{1-0.02} = z_{0.98} = 2.05$

$-0.47 > -2.05 \Rightarrow$ we do not reject H_0

1) Ten Italian soccer teams playing in a world competition have been classified according to their performance in the last five seasons. The best team is assigned number 1, the second team number 2 and so on until the worst team is given number 10. With respect to the first match of the competition, the 10 teams are randomly divided into 5 *home* teams (i.e. they will play the first match at home) and 5 *away* teams (i.e. they will play the first match away).
   a) What is the probability that exactly two of the top 5 teams (i.e. teams with number 1, 2, 3, 4, 5) end up playing home?
   b) What is the probability that all of the top 3 teams (i.e. teams with number 1, 2, 3) end up playing the same way?
   c) If there are 2n teams, what is the probability mass function of T=the number among the top n teams who end up playing home? (If binomial coefficients appear in the formula, it is sufficient to leave them indicated without making any computation).

(6pt)

1) X number of top 5 teams that will play home among the five selected

2) $X \sim \text{Hypergeom}(N=10, M=5, n=5)$ (number of team selected for playing home)

(N: total number of teams; M: number of top teams)

$$P(X=2) = \frac{\binom{5}{2}\binom{5}{3}}{\binom{10}{5}} = 10 \cdot 10 \cdot \frac{5!5!}{10!} = 100 \cdot \frac{5 \cdot 4 \cdot 3 \cdot 2}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} = \frac{50}{14 \cdot 9} = \frac{25}{63} \simeq 0.396825$$

b) Y number of top 3 teams that will play home among the 5 selected

$Y \sim \text{Hypergeom}(N=10, M=3, n=5)$

$P(\text{"all three teams play in the same way"}) =$

$P(Y=3) + P(Y=0) =$

$$= \frac{\binom{3}{3}\binom{7}{2}}{\binom{10}{5}} + \frac{\binom{3}{0}\binom{7}{5}}{\binom{10}{5}} = \left(\frac{7 \cdot 6}{2} + \frac{7 \cdot 6}{2}\right) \frac{5 \cdot 4 \cdot 3 \cdot 2}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} =$$

$$\simeq \frac{7 \cdot 6^{2}}{9 \cdot 7 \cdot 2^{2}} = \frac{1}{6} \simeq 0.166667$$

c) $T \sim \text{Hyp}(N = 2n, M = n, n = n)$

$$\mathbb{P}(T = t) = \begin{cases} \dfrac{\binom{n}{t}\binom{n}{n-t}}{\binom{2n}{n}} & t = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

STATISTICAL METHODS IN DATA SCIENCE
FINAL EXAM 2 (PART 2)

Quiz 1 Let X_1 and X_2 be two independent and exponentially distributed random variables, both with rate $\lambda = 2$. Then, by using the Chebishev's inequality, we can affirm that:

[**a**] $P[(X_1 + X_2) \geq t] \leq 3/(2t^2)$ for all $t \in \mathbb{R}^+$.

[**b**] $P[(X_1 + X_2) \geq t] \leq 1/(2t^2)$ for all $t \in \mathbb{R}^+$.

[**c**] $P[(X_1 + X_2) \geq t] \leq 1/(2t)$ for all $t \in \mathbb{R}^+$.

[**d**] $P[(X_1 + X_2) \geq t] \leq 1/2$ for all $t \in \mathbb{R}^+$.

Quiz 1 Let X_1 and X_2 be two independent and Poisson distributed random variables, both with $\lambda = 1$. Then, by using the Chebishev's inequality, we can affirm that:

[**a**] $P[(X_1 + X_2) \geq n] \leq 6/n^2$ for all $n \in \mathbb{N}^+$.

[**b**] $P[(X_1 + X_2) \geq n] \leq 2/n^2$ for all $n \in \mathbb{N}^+$.

[**c**] $P[(X_1 + X_2) \geq n] \leq 1/n$ for all $n \in \mathbb{N}^+$.

[**d**] $P[(X_1 + X_2)^2 \geq n] \leq 1/n^2$ for all $n \in \mathbb{N}^+$.

Quiz 2 To estimate the mean of a population X we have a sample (X_1, X_2, X_3) of size $n = 3$. Using it, we consider two different estimators (for the mean): $\widehat{\mu}_1 = (2X_1 + X_2 + 2X_3)/5$ and $\widehat{\mu}_2 = (X_1 + 2X_2 + X_3)/4$. Which of the following is true?

[**a**] The bias of $\widehat{\mu}_1$ is strictly smaller than that of $\widehat{\mu}_2$.

[**b**] The bias of $\widehat{\mu}_2$ is strictly smaller than that of $\widehat{\mu}_1$.

[**c**] The MSE of $\widehat{\mu}_1$ is strictly smaller than that of $\widehat{\mu}_2$.

[**d**] The MSE of $\widehat{\mu}_2$ is strictly smaller than that of $\widehat{\mu}_1$.

Quiz 2 To estimate the mean of a population X we have a sample (X_1, X_2, X_3) of size $n = 3$. Using it, we consider two different estimators (for the mean): $\widehat{\mu}_1 = (X_1 + 3X_2 + X_3)/5$ and $\widehat{\mu}_2 = (3X_1 + X_2 + 3X_3)/7$. Which of the following is true?

[**a**] The bias of $\widehat{\mu}_1$ is strictly smaller than that of $\widehat{\mu}_2$.

[**b**] The bias of $\widehat{\mu}_2$ is strictly smaller than that of $\widehat{\mu}_1$.

[**c**] The MSE of $\widehat{\mu}_1$ is strictly smaller than that of $\widehat{\mu}_2$.

[**d**] The MSE of $\widehat{\mu}_2$ is strictly smaller than that of $\widehat{\mu}_1$.

Quiz 3 To provide an interval estimate of the mean of normally distributed population, with confidence 95%, we take a sample of size $n = 10$, that gives a sample mean $\widehat{\mu}$ equal 10.5 and an unbiased sample variance $\widehat{\sigma}_u^2$ equal 0.9. Then the interval is

[a] $(10.5 - 0.3 \cdot z_{0.95},\ 10.5 + 0.3 \cdot z_{0.95})$.

[b] $(10.5 - 0.3 \cdot t_{9,0.975},\ 10.5 + 0.3 \cdot t_{9,0.975})$.

[c] $(10.5 - \sqrt{0.3} \cdot t_{10,0.975},\ 10.5 + \sqrt{0.3} \cdot t_{10,0.975})$.

[d] $(10.5 - \sqrt{0.3} \cdot t_{10,0.95},\ 10.5 + \sqrt{0.3} \cdot t_{10,0.95})$.

Quiz 3 To provide an interval estimate of the mean of normally distributed population, with confidence 95%, we take a sample of size $n = 9$, that gives a sample mean $\widehat{\mu}$ equal 10.5 and an unbiased sample variance $\widehat{\sigma}_u^2$ equal 0.09. Then the interval is

[a] $(10.5 - 0.1 \cdot z_{0.95},\ 10.5 + 0.1 \cdot z_{0.95})$.

[b] $(10.5 - 0.1 \cdot t_{8,0.975},\ 10.5 + 0.1 \cdot t_{8,0.975})$.

[c] $(10.5 - \sqrt{0.1} \cdot t_{9,0.975},\ 10.5 + \sqrt{0.1} \cdot t_{9,0.975})$.

[d] $(10.5 - \sqrt{0.1} \cdot t_{9,0.95},\ 10.5 + \sqrt{0.1} \cdot t_{9,0.95})$.

Quiz 4 Let $X \sim N(\mu, 10)$, and assume that μ can assume only the values $\mu_1 = 1$ and $\mu_2 = 2$. We want to test $H_0 : \mu = \mu_1$ versus $H_1 : \mu = \mu_2$ by using a sample of size $n = 10$. We reject H_0 if $\overline{X}_{10} > 1.2$. Denoted with $\Phi(z)$ the cumulative distribution of a $Z \sim N(0, 1)$, then the probability of an error of the II type is:

[a] $\Phi(0.8)$

[b] $\Phi(0.2)$

[c] $1 - \Phi(0.8)$

[d] $1 - \Phi(0.2)$

Quiz 4 Let $X \sim N(\mu, 5)$, and assume that μ can assume only the values $\mu_1 = 1$ and $\mu_2 = 2$. We want to test $H_0 : \mu = \mu_1$ versus $H_1 : \mu = \mu_2$ by using a sample of size $n = 20$. We reject H_0 if $\overline{X}_{20} > 1.2$. Denoted with $\Phi(z)$ the cumulative distribution of a $Z \sim N(0, 1)$, then the probability of an error of the II type is:

[a] $1 - \Phi(1.6)$

[b] $\Phi(0.8)$

[c] $1 - \Phi(0.8)$

[d] $\Phi(1.6)$

(a) **Ex 1)** Consider a discrete random variable X that can assume values only in the set $\{1,2,3\} \subseteq \mathbb{N}$ and whose discrete density is

$$f_\theta(x) = \begin{cases} \frac{1}{3} - \theta, \text{ if } x = 1; \\ \frac{1}{3} + 2\theta, \text{ if } x = 2; \\ \frac{1}{3} - \theta, \text{ if } x = 3; \\ 0, \text{ otherwise,} \end{cases}$$

where the parameter θ has range in $[0, 1/3)$. A sample of size $n = 100$ gives 25 times the outcome 1, 42 times the outcome 2 and 33 times the outcome 3.

(a) Write the likelihood function, and use it to provide a point estimate of θ.

(b) Using the χ^2 test for goodness of fit, test with confidence 95% the hypothesis that the population's density is the one above with $\theta = 1/12$.

(b) **Ex 1)** Consider a discrete random variable X that can assume values only in the set $\{1,2,3\} \subseteq \mathbb{N}$ and whose discrete density is

$$f_\theta(x) = \begin{cases} \frac{1}{4} - \theta, \text{ if } x = 1; \\ \frac{1}{2} + 2\theta, \text{ if } x = 2; \\ \frac{1}{4} - \theta, \text{ if } x = 3; \\ 0, \text{ otherwise,} \end{cases}$$

where the parameter θ has range in $[0, 1/4)$. A sample of size $n = 100$ gives 12 times the outcome 1, 68 times the outcome 2 and 20 times the outcome 3.

(a) Write the likelihood function, and use it to provide a point estimate of θ.

(b) Using the χ^2 test for goodness of fit, test with confidence 95% the hypothesis that the population's density is the one above with $\theta = 0$.

EX 1 (a)

(a) $L(\theta|\bar{x}) = \prod f_\theta(x_i) = \left(\frac{1}{3}-\theta\right)^{58} \cdot \left(\frac{1}{3}+2\theta\right)^{42}$

$\log L(\theta|\bar{x}) = 58 \cdot \log\left(\frac{1}{3}-\theta\right) + 42 \log\left(\frac{1}{3}+2\theta\right)$

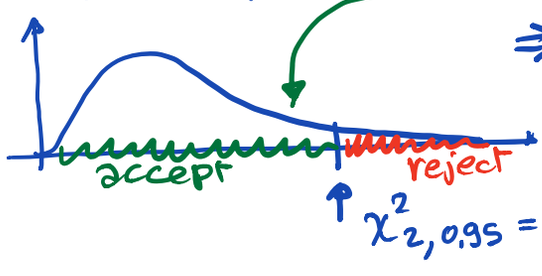
$\frac{\partial \log L(\theta|\bar{x})}{\partial \theta} = \frac{-58}{\frac{1}{3}-\theta} + \frac{84}{\frac{1}{3}+2\theta}$

$= 0 \Rightarrow -600\theta + 26 = 0 \Rightarrow \theta = \frac{26}{600} \simeq 0.043$

(b)

c	f_i	p_i	np_i
1	25	1/4	25
2	42	1/2	50
3	33	1/4	25

$W = \sum_{i=1}^{3} \frac{(f_i - np_i)^2}{np_i} = 3.84$



$\Rightarrow$ We do not reject !

EX 1 (b)

(a) $L(\theta|\bar{x}) = \prod f_\theta(x_i) = \left(\frac{1}{4}-\theta\right)^{32} \cdot \left(\frac{1}{2}+2\theta\right)^{68}$

$\log L(\theta|\bar{x}) = 32 \cdot \log\left(\frac{1}{4}-\theta\right) + 68 \log\left(\frac{1}{2}+2\theta\right)$

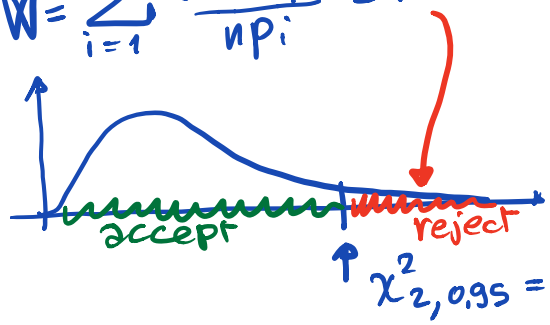
$\frac{\partial \log L(\theta|\bar{x})}{\partial \theta} = \frac{-32}{1/4-\theta} + \frac{132}{1/2+2\theta}$

$= 0 \Rightarrow -196\theta + 17 = 0 \Rightarrow \theta = \frac{17}{196} \approx 0{,}087$

(b)

C	f_i	p_i	np_i
1	12	1/4	25
2	68	1/2	50
3	20	1/4	25

$W = \sum_{i=1}^{3} \frac{(f_i - np_i)^2}{np_i} = 14.24$



$\Rightarrow$ reject H_0 !

(a) **Ex 2)** Let $X \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. For a sample having size $n = 10$, one gets $\overline{X}_{10} = 1.5$ and $\widehat{\sigma}_u^2 = 0.9$.

(a) Provide a confidence interval for the mean, with $1 - \alpha = 0.99$.

(b) Test, with confidence 90%, the hypothesis that the mean is equal 1.

(c) Assuming the variance to be know, and equal 1, find the p-value for the test in point (b).

(b) **Ex 2)** Let $X \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. For a sample having size $n = 9$, one gets $\overline{X}_9 = 0.52$ and $\widehat{\sigma}_u^2 = 0.09$.

(a) Provide a confidence interval for the mean, with $1 - \alpha = 0.99$.

(b) Test, with confidence 90%, the hypothesis that the mean is equal 0.5.

(c) Assuming the variance to be know, and equal 0.36, find the p-value for the test in point (b).

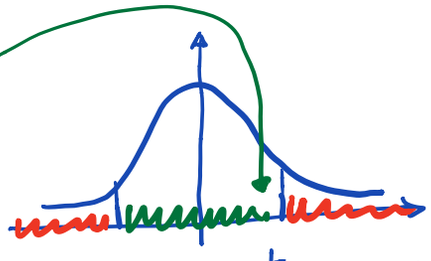
Ex 2 (a)

(a) $$I = \left(\bar{X}_n - t_{n-1,\,0.995} \cdot \sqrt{\frac{\hat{\sigma}^2}{n}},\ \bar{X}_n + t_{n-1,\,0.995} \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right)$$

$t_{9,\,0.995} = 3.25 \qquad \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{0.09} = 0.3$

$$I = (1.5 - 0.975,\ 1.5 + 0.975) = (0.525,\ 2.475)$$

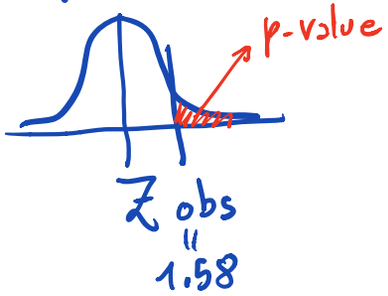
(b) $$T_{obs} = \frac{\bar{X}_n - 1}{\sqrt{\frac{\hat{\sigma}^2}{n}}} = \frac{0.5}{0.3} = 1.66$$



$T_{obs} \in$ accept. region

$\Rightarrow$ Do not reject H_0

(c) p-value



$$Z_{obs} = \frac{\bar{X}_n - 1}{\sqrt{1/10}} = 1.58$$

p-value $= P[Z > 1.58] = 1 - \Phi(1.58) \simeq 0.06$

(for one-sided tests)

p-value $= P[|Z| > 1.58] = 2(1 - \Phi(1.58)) \simeq 0.12$

(for two-sided tests)

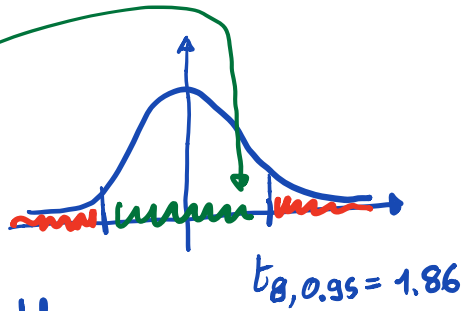
Ex 2

(a) $I = \left(\bar{X}_n - t_{n-1,\,0.995} \cdot \sqrt{\frac{\vartheta^2}{n}},\ \bar{X}_n + t_{n-1,\,0.995} \cdot \sqrt{\frac{\vartheta^2}{n}} \right)$

$t_{8,\,0.995} = 3.355 \qquad \sqrt{\frac{\vartheta^2}{n}} = \sqrt{\frac{0.09}{9}} = 0.1$

$I \approx (0.52 - 0.33,\ 0.52 + 0.33) \approx (0.19,\ 0.85)$

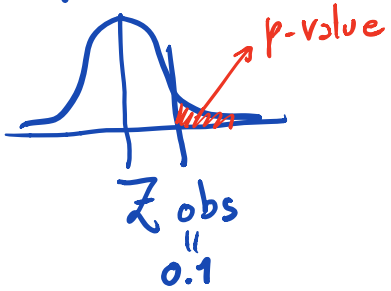
(b) $T_{obs} = \frac{\bar{X}_n - 0.5}{\sqrt{\frac{\vartheta^2}{n}}} = \frac{0.02}{0.1} = 0.2$



$T_{obs} \in$ accept. region

$\Rightarrow$ Do not reject H_0

(c) p-value



$Z_{obs} = \frac{\bar{X}_n - 0.5}{\sqrt{0.09}} = 0.1$

- p-value $= P[Z > 0.1] = 1 - \Phi(0.1) \approx 0.46$

(for one-sided tests)

p-value $= P[|Z| > \ ] \approx 2(1 - \Phi(0.1)) \approx 0.92$

(for two-sided tests)

2) Sugar of a certain brand is packaged in boxes whose net weight X is approximated with a normal random variable of mean 500 and standard deviation 2 (values in grams). It is assumed that the net weights of different boxes are independent.
   a) What is the probability that the net weight of a randomly chosen box is less than 498 grams?
   b) What is the probability that the total net weight of 100 boxes is less than 50020 grams?
   c) What is the probability that among five randomly selected boxes exactly two have net weight less than 498 grams?

(6pt)

2) $X \sim N(500, 2^2)$

$$P(X < 498) = P\left(\frac{X-500}{2} < \frac{498-500}{2}\right) =$$

$$= P(Z < -1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

$Z \sim N(0,1)$

b) $Y = X_1 + \dots + X_{100}$

Y is a linear combination of independent normal random variables. Then

$$Y \sim N(100 \cdot \mu_X, 100\,\sigma_X^2) = N(50000, 400)$$

$$\equiv N(50000, 20^2)$$

$$P(Y < 50020) = P\left(\frac{Y-50000}{20} < \frac{50020-50000}{20}\right)$$

$= P(Z < 1) = \Phi(1) = 0.8413$

$Z \sim N(0,1)$

c) T number of boxes with net weight less than 498 among five randomly selected

$T \sim \text{Binomial}(n = 5, p = 0.1587)$ ← from point e)

$$P(T = 2) = \binom{5}{2} 0.1587^2 \cdot 0.8413^3 =$$

$$= 10 \; 0.1587^2 \cdot 0.8413^3 \cong 0.149971$$