Endterm CSE1210

Friday June 24th, 2022, 9:00 — 11:00 Responsible lecturer: Christophe Smet

You can use a non-graphing, non-programmable calculator. You only have to hand in the answer sheet. Circle the correct option for the multiple choice questions. 27 marks can be earned, if you score M of them your grade will be $1 + \frac{M}{3}$.

1) 4p

Consider the random variable X with density function $f(x) = Ca^{1/3}e^{-ax^3}$ for $x > 0$. Here a is an unknown positive parameter and $C \approx 0.77$ is a constant that does not depend on a. For a random sample $\{X_i\}$ of size n from this distribution, consider the sum of cubes $S = \sum_{i=1}^{n} X_i^3$. You can use that $Ca^{1/3} \int_0^\infty x^3 e^{-ax^3} dx = \frac{1}{3a}$.

(a) Use S to construct an estimator T which is an unbiased estimator for $\frac{1}{a}$.

(b) Let us now construct an estimator for a. First, give the likelihood function $L(a)$ for a random sample x_1, x_2, x_3, x_4, x_5 of size 5.

(c) Find the maximum likelihood estimate for a if the dataset (for $n = 5$) consists of

$$x_1 = 0.2 \qquad x_2 = 0.3 \qquad x_3 = 0.4 \qquad x_4 = 0.5 \qquad x_5 = 0.8.$$

2) 2p

A standard deck of cards contains 52 cards, four of which are aces. I draw four cards and I get four aces!

(a) According to the maximum likelihood idea, which approach is most likely?

(b) What is the value of the likelihood function in the case with replacement?

3) 3p

Someone wants to construct a 95% confidence interval for the mass of an object, based on independent measurements of this mass. The measured mass is the real mass, plus a measurement error $Y_i \sim N(0, (0.5)^2)$.

(a) How large should the sample size be, in order to obtain an interval with width equal to 0.2? Round to the nearest integer. Note: the width of an interval $[a, b]$ is $b - a$.

(b) Which of the following would lead to a narrower (Dutch: smaller) interval? Indicate all correct answers. Note: $\alpha = 5\%$ in the starting situation.

4) 3p

A random sample of size 8 is being taken from a Pareto distribution with parameter α. You want to test the null hypothesis $H_0 : \alpha = 3$ against the alternative hypothesis $H_1 : \alpha < 3$. Perform this test on a 10% significance level, using as a test statistics the minimum of these 8 observations. The observed value for this minimum is 1.2. You can use that the minimum M_n of n independent random variables from a Pareto(α) distribution, has a Pareto($n\alpha$) distribution: $M_n \sim$ Pareto($n\alpha$).

(a) Compute the p-value for this hypothesis test.

(b) What is your conclusion?

5) 5p

We want to check whether the mean of a normally distributed random variable, is equal to 12. We use a two-sided t-test with significance level 5%. In a random sample of 14 observations, we find a sample mean of 12.6 with a sample standard deviation of 1.3.

(a) Give the test statistic T.

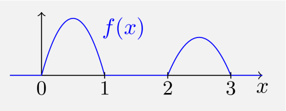
(b) Give its realization t.

(c) What is the distribution of T if H_0 is true?

(d) Make a sketch of the density function of the test statistic T and indicate the following: critical value(s), critical region(s), realization(s) of the test statistic.

6) 3p

Consider a random sample $\{X_i\}$ of size 2000 from a distribution with density function f. The graph of f is given: f is zero everywhere, except on the interval $[0, 1]$, where there is an area 0.6 below the graph, and on the interval $[2, 3]$, where there is an area 0.4 below the graph. Suppose we make a boxplot of the obtained dataset. The sample size is sufficiently high that you can assume that sample parameters are very close to the population parameters.



(a) Do you expect there to be any outliers?

(b) Where do you expect the median to be?

(c) Where do you expect the sample mean to be?

7) 2p

In order to predict a patient's blood pressure, a multiple linear regression analysis with constant term (i.e. with intercept) is being performed on data of 200 patients. The explanatory variables are the patient's age, whether they smoke (yes/no) and their blood type (four options: A, B, AB, O). Write down the model and explain the possible values of all variables x_i.

8) 2p

Consider a random sample of size n from a $U(-\theta, 2\theta)$ distribution, where $\theta > 0$ is an unknown parameter. We use $T = 2\bar{X}_n$ as an estimator for θ.

(a) Compute the bias of T.

(b) Compute the variance of T.

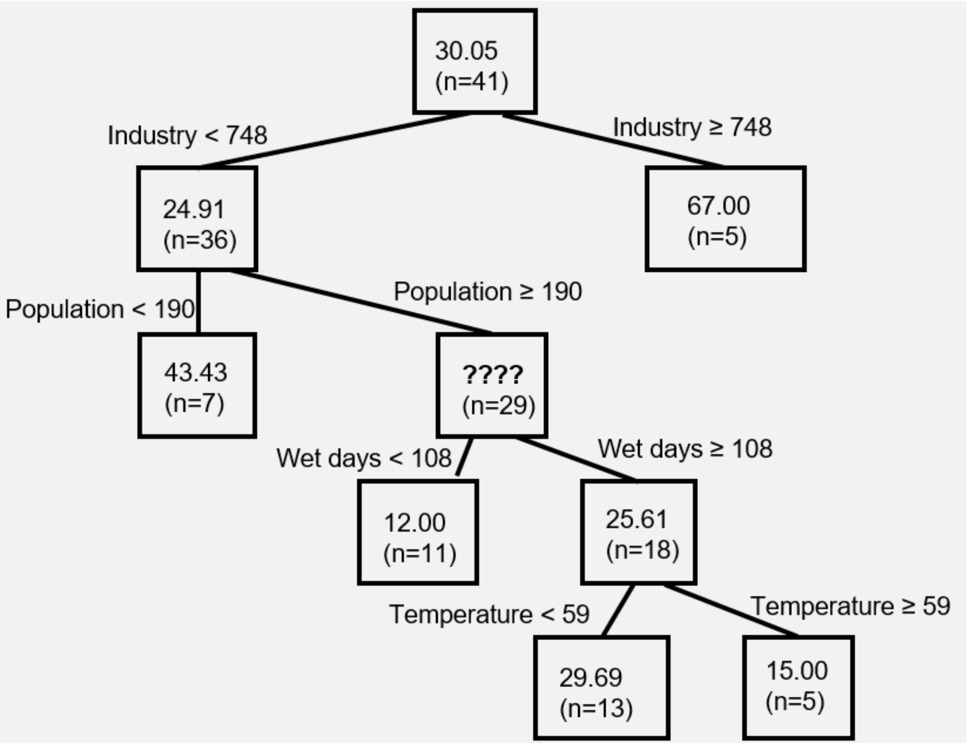
9) 3p

In order to find how pollution levels depend on certain explanatory variables, a regression tree has been made (given below).

(a) What is the predicted pollution level for a town with industry level equal to 800, population equal to 150, temperature level equal to 65, and 135 wet days?

(b) Compute the value that has been replaced by question marks.

(c) The nodes-sum-of-squares technique (covered in the lecture) was used to make this tree. Explain this technique in at most three lines (you can use the splitting at the bottom, based on temperature, as an example).



Answer sheet Endterm CSE1210 - June 24th, 2022

Name: **Student number:**

1. (a) $T =$ __________ (b) $L(a) =$ __________

   (c) $\hat{a} =$ ● 1.352 ● 1.509 ● 1.822 ● 2.264 ● 2.592 ● 3.030 ● 3.577

2. (a) ● With replacement (putting the card back in the deck after taking one out)

   ● Without replacement ● With and without replacement are equally likely.

   (b) $L =$ __________

3. (a) $n =$ ● 24 ● 45 ● 62 ● 96 ● 144 ● 225 ● 331

   (b) ● Increase α ● Decrease α

   ● Increase n ● Decrease n

4. (a) ● 0.0003 ● 0.0126 ● 0.0228 ● 0.0318 ● 0.0533 ● 0.0726 ● 0.0892

   (b) ● Reject H_0 in favour of H_1 ● Do not reject H_0 in favour of H_1

5. (a) $T =$ __________ (d) Sketch:

   (b) $t =$ __________

   (c) Under H_0, $T \sim$ __________

6. (a) ● No ● Yes, at the top of the boxplot

   ● Yes, at the bottom ● Yes, both at the top and at the bottom.

   (b) ● Closer to the lower quartile ● Closer to the upper quartile

   (c) ● In $[0, 1]$ ● In $[1, 2]$ ● In $[2, 3]$

7. $Y =$ __________

8. (a) $\text{Bias}(T) =$ __________

   (b) $\text{Var}(T) =$ __________

9. (a) Predicted pollution level: __________

   (b) Value for question marks: __________

   (c)

Solutions

1. (a) The given integral is the expression for $E[X^3]$. Hence $E[S] = \frac{n}{3a}$. So with $T = \frac{3S}{n}$ you get that $E[T] = \frac{1}{a}$, meaning that this T is an unbiased estimator for $\frac{1}{a}$.

   (b) The likelihood function based on those five observations, is the product of the pdf, evaluated at those five points. This means that

   $$L(a) = Ca^{1/3}e^{-ax_1^3}Ca^{1/3}e^{-ax_2^3}\ldots = C^5a^{5/3}e^{-a(x_1^3+x_2^3+x_3^3+x_4^3+x_5^3)}.$$

   (c) The log-likelihood is

   $$\ell(a) = 5\ln(C) + \frac{5}{3}\ln(a) - a\sum x_i^3,$$

   with derivative

   $$\ell'(a) = \frac{5}{3a} - \sum x_i^3.$$

   This is zero if

   $$\hat{a} = \frac{5}{3\sum x_i^3} \approx 2.264.$$

2. With replacement, the likelihood is

   $$L = \frac{4}{52}\frac{4}{52}\frac{4}{52}\frac{4}{52} \approx 0.000035.$$

   Without replacement, the likelihood is

   $$L = \frac{4}{52}\frac{3}{51}\frac{2}{50}\frac{1}{49} \approx 0.0000037.$$

   Hence the approach with replacement is more likely.

3. (a) The width of a confidence interval for the mean of a normal random sample with known variance, is given by $\frac{2z_{0.025}\sigma}{\sqrt{n}}$. Put this equal to 0.2 and solve for n to find (rounded to the nearest integer) $n = 96$.

   (b) Increasing α gives a smaller z-score and hence a narrower interval. Increasing n has this same effect.

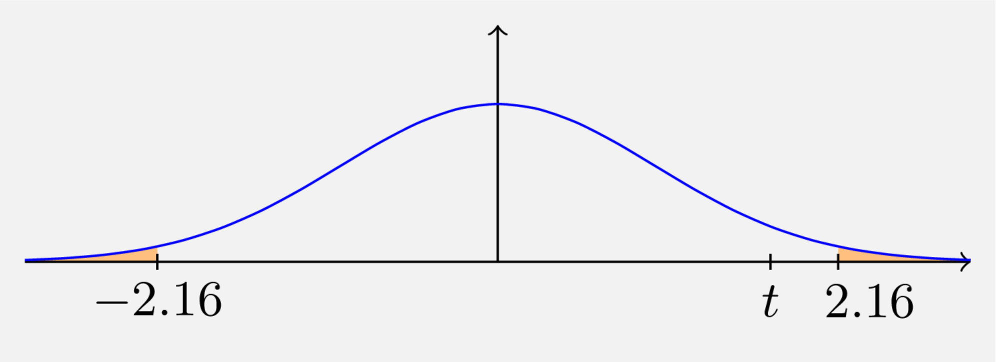
4. Under the assumption of H_0, M_8 has a Pareto distribution with parameter $8 \cdot 3 = 24$. The observed value for the minimum was 1.2. The higher the parameter, the closer to 1 this quantity will be. Hence greater values for the minimum will give evidence for a smaller value for α, which was stated by H_1. This means the the p-value is the probability that a Pareto(24) distribution takes the observed value 1.2, or even larger. This probability is given by $1 - F(1.2) = 1.2^{-24} \approx 0.0126$. Since this is less than the significance level 0.1, the null hypothesis gets rejected in favour of the alternative one (as would be the case for all options stated in the first subquestion).

5. (a) $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ (or with μ instead of μ_0, or with $n = 14$ in particular).

   (b) $t = \frac{0.6\sqrt{14}}{13} \approx 1.73$.

   (c) This T has a t_{13} distribution.

   (d) Sketch see below, the critical region indicated in orange.



6. (a) Since you expect $Q1$ to be between 0 and 1, and $Q3$ between 2 and 3, no outliers are to be expected.

   (b) The median will also be between 0 and 1, since 60% of the area is there as well. Hence it will be closer to the lower quartile.

   (c) The sample mean will be close to $0.6 \cdot 0.5 + 0.4 \cdot 2.5 = 1.3$ so certainly between 1 and 2.

7. The model will be something like $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + U$. Here x_1 is the age of the patient, $x_2 = 1$ if they smoke, else 0. $x_3 = 1$ if the blood type is A, else $x_3 = 0$. Similar for x_4 (type B) and x_5 (type AB). Different encodings can be chosen, but there should be one variable for age, one for smoking and three for the blood type.

8. (a) Since $E[X_i] = \frac{\theta}{2}$, we see that $E[T] = \theta$ so the bias is zero.

   (b) $\text{Var}(T) = \frac{4}{n}\text{Var}(X_i) = \frac{4}{n}\frac{(3\theta)^2}{12} = \frac{3\theta^2}{n}$ where the variance of a uniform random variable has been used.

9. (a) Follow the tree to spot a predicted pollution level of 67.

   (b) Each node contains the group average for the pollution level and the number of observations in that node. The deleted group average can then be found either by looking up or by looking down. By looking down you get that $\bar{x}_{29} = \frac{11 \cdot 12.00 + 18 \cdot 25.61}{29} \approx 20.45$.

   (c) Choose a temperature T and split the 18 observations into two groups, $P1$ and $P2$ with higher/lower temperature than T. Do this in such a way that the node-sum-of-squares, $\sum_{P_1}(y_i - \bar{y}_{P_1})^2 + \sum_{P_2}(y_i - \bar{y}_{P_2})^2$ is minimal. It is also fine if you wrote the rule of thumb that follows from this, which is that the node above gets split into two subnodes, in such a way that the sizes are more or less equal and the averages are as distinct as possible.