INTRODUCTION TO STATISTICS
MID-TERM EXAM

December 9, 2020, 09:00 – 12:00CET

Guidelines

- This exam consists of 13 questions, and you can get a maximum of 40 points. The number of points per question is indicated at the beginning of each question.
- Complete the tasks listed in this document and fill in your answers in the answer sheet (Google doc *midterm-exam-answer-sheet* attached).
- If any of the questions is unclear, do not hesitate to contact me.
- Make sure that your answers always contain ***explanations*** of *how* you arrive to the solution. Do not just state the answer.
- If you are not sure how to compute a certain derivative, integral, etc., you can use WolframAlpha to do so.
- You have 3 hours to complete the exam. **Answer sheets should be submitted to Google Classroom by 12:00 Barcelona time.**
- This is an open-book exam. You can use any materials you wish.
- **This exam should be completed strictly individually**. It is not allowed to collaborate with your classmates or ask help from anyone.

Good luck!

Question 1 *(2 points)*

Identify which type of the Statistics – descriptive or inferential – is needed to answer each of the questions below. Explain your reasoning for every case.

(a) Given data on 20 fish caught in a lake, what is the average weight of all fish in that lake?

(b) Given data on every customer service request made, what is the average time it took to respond?

(c) After interviewing 100 customers, what percent of all the customers are satisfied with the product?

(d) Given data on all 100,000 people who viewed an ad, what percent of people clicked on it?

Question 2 *(2 points)*

Map variables listed below to one of the following data types: (1) continuous numeric, (2) discrete numeric or (3) categorical.

Air temperature, number of items in stock, zip code, kilowatts of electricity used, number of online courses taken, brand of a product, number of clicks on an ad, sex

Question 3 *(2 points)*

Consider a toy dataset below that contains information about last month's book sales in a local bookstore.

(a) Compute the mean and the median number of copies sold per book. Is there a big difference between the mean and the median that you obtained? If so, why?

(b) Construct a pivot table that highlights the *total* number of books sold *per publisher*.

Title	Publisher	Copies sold
Hamlet	Best Books & Co	600
Metamorphosis	Good Books	50
War and Peace	Good Books	700
Harry Potter	Good Books	5000
Don Quixote	Best Books & Co	400
Old Man and the Sea	Good Books	350
100 Years of Solitude	Good Books	400
Gone with the Wind	Best Books & Co	300
Great Gatsby	Best Books & Co	550
Pride and Prejudice	Best Books & Co	550

Question 4 *(2 points)*

Let X and Y be independent random variables and let $Z = 2X + 3Y$.

(a) Suppose that $E(X) = 1$ and $E(Y) = 2$. What is $E(Z)$?

(b) What is the standard deviation of Z if standard deviations of X and Y are 3 and 4 respectively?

Question 5 *(2 points)* Suppose that random variables $X_1, X_2, \dots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1 \dots n$.

Note: σ_i^2 refers to the variance of X_i, while σ_i is its standard deviation.

(a) Which distribution does a random variable $Y = \sum_{i=1}^{n} a_i X_i$ follow?

(b) Compute the values of the parameter(s) of the distribution of Y.

Question 6 *(3 points)*

Let random variable X follow a uniform distribution between 0 and 1.

(a) What is the expected value and variance of X?

(b) Compute the expected value of the random variable $Y = e^X$.

Question 7 *(3 points)*

Suppose that Z is a standard normal variable, i.e., $Z \sim N(0,1)$. Let us define a function $G(x) = P(0 < Z \leq x)$ for $x \geq 0$.

You can use $G(x)$ to compute probabilities from the standard normal distribution. For example, $P(Z > 1.53) = 0.5 - G(1.53)$.

Express the probabilities below in terms of $G(0.5)$, $G(1)$ and $G(1.53)$ in a similar way. Always explain how you obtain the result. Make drawings, if necessary.

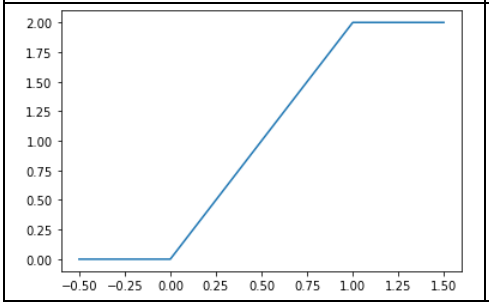
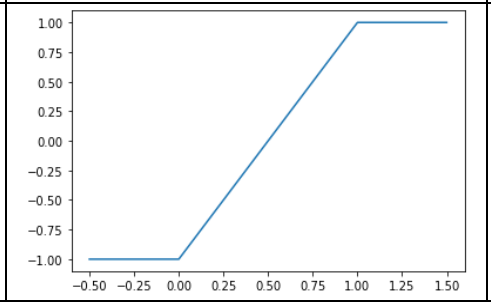
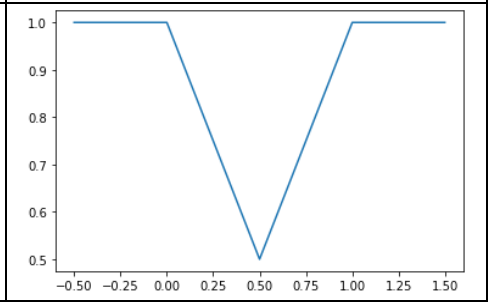
(a) $P(Z \leq 1.53)$

(b) $P(Z \leq -0.5)$

(c) $P(-1 < Z < 0.5)$

Question 8 *(3 points)*

Consider the functions below. Based on their plots, explain why each of them can or cannot be a valid cumulative distribution function (CDF).

(a)	*(b)*	*(c)*
		

Question 9 *(3 points)*

Consider the following function:

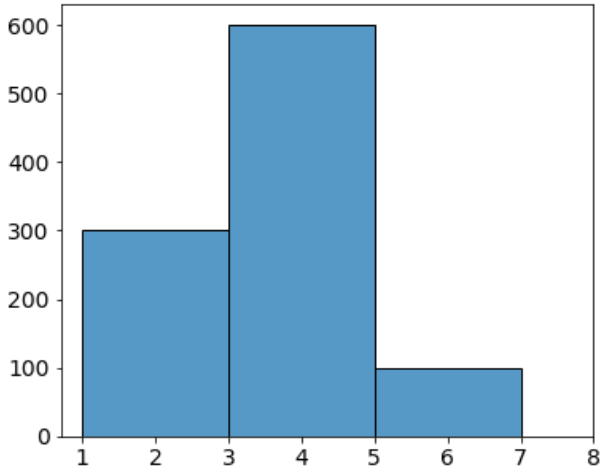
$$f(x) = \begin{cases} 0.2, & 0 < x < 10 \\ 0, & otherwise \end{cases}$$

Is it a valid probability density function (PDF) for some continuous distribution? Explain why or why not.

Hint: check the basic properties that a PDF should satisfy.

Question 10 *(3 points)*

Consider a histogram below that visualizes the distribution of 1000 random samples drawn from an unknown distribution. The heights of the bins correspond to the absolute frequency of the bin. For example, out of 1000 samples, 300 fall between 1 and 3, and so on.



Our goal is to study the underlying probability distribution. How should the heights of the bins be *normalized* if we want to approximate the true probability density function of the underlying distribution with such a histogram?

Your answer should be the normalized bin heights of the new histogram. Keep the number and size of the bins as illustrated on the plot.

Question 11 *(5 points)*

Consider a discrete random variable X that takes values 0, 1, 2, 3, … Its probability mass function is defined as follows:

$$P(X = k) = \theta(1 - \theta)^k, \qquad 0 < \theta < 1$$

(a) Derive the maximum likelihood estimator for the unknown parameter θ.

(b) A random sample of size 1000 gave

$$\sum_{i=1}^{1000} X_i = 980.$$

What is the estimate of θ that the estimator from the previous step would provide?

(c) It is known that $E(X) = \frac{1-\theta}{\theta}$. Based on the above, compute the maximum likelihood estimate of $E(X)$. Justify.

Question 12 *(5 points)*

A random variable X, denoting repair time (in minutes) of some machine, follows a Normal distribution with unknown parameters μ and σ^2. A random sample of size 10 was drawn from the distribution, and the following was observed:

$$\sum_{i=1}^{10} X_i = 846, \quad \sum_{i=1}^{10} X_i^2 = 71607$$

(a) Based on the observed data, compute the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the parameters μ and σ^2. You can use the formulas obtained in class.

(b) What is the probability that the repair time is not greater than 83 minutes? Express it using the CDF of the standard normal distribution.

Question 13 *(5 points)*

Consider a random sample of size 3 (X_1, X_2, X_3) from some distribution with mean μ and variance σ^2. In order to estimate the mean, the following estimators are proposed:

$$T_1 = \frac{X_1 + X_3}{2}, \qquad T_2 = \frac{X_1 + 2X_2 + 3X_3}{6}$$

(a) Are T_1 and T_2 unbiased estimators of the mean? Show.

(b) Compare the variances of T_1 and T_2.

(c) Based on the above, can you conclude that one of the proposed estimators is better than the other? Why? Explain.