INTRODUCTION TO STATISTICS
MID-TERM EXAM

December 9, 2020, 09:00 – 12:00CET

Solutions

Question 1 *(2 points)* **– LECTRUE 1, PROGRESS TEST 1**

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?

Inferential

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

Descriptive

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

Inferential

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

Descriptive

Question 2 *(2 points)* **– LECTRUE 1, PROGRESS TEST 1, GRADED ASSIGNMENT 1**

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

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

Question 3 *(2 points)* **– LECTRUE 1, PROGRESS TEST 1, GRADED ASSIGNMENT 1**

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?

Mean:
(600+50+700+5000+400+350+400+300+550+550)/10 = 890

Median:
50 300 350 400 **400 550** 550 600 700 5000
(400 + 550)/2 = 475

The difference is due the presence of an outlier: Harry Potter has much more copies sold than any other book in question

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

Best Books & Co	600 + 400 + 300 + 550 + 550 = 2400
Good Books	50 + 700 + 5000 + 350 + 400 = 6500

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)* **– LECTRUE 3, PROGRESS TEST 2**

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 $E(Z)$?

$$E(Z) = E(2X + 3Y) = 2E(X) + 3E(Y) = 2 * 1 + 3 * 2 = 8$$

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

$$Var(X) = 3^2 = 9, \quad Var(Y) = 4^2 = 16$$

$$Var(Z) = Var(2X + 3Y) = 2^2 Var(X) + 3^2 Var(Y) = 4 * 9 + 9 * 16 = 180$$

$$std(Z) = \sqrt{Var(Z)} = \sqrt{180} \sim 13.4$$

Question 5 *(2 points)* Suppose that random variables $X_1, X_2, \dots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2),\ i = 1 \dots n.$ **– LECTRUE 6, PROGRESS TEST 2**

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?

$Y \sim N(\mu_Y, \sigma_Y^2)$
(sum of independent normally distributed random variables is normally distributed)

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

$$\mu_Y = E(Y) = E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n)$$
$$= \sum_{i=1}^{n} a_i\mu_i$$

$$\sigma_Y^2 = Var(Y) = Var(a_1X_1 + \cdots + a_nX_n) + a_1^2Var(X_1) + \cdots + a_n^2Var(X_n)$$
$$= \sum_{i=1}^{n} a_i^2\sigma_i^2$$

$$\sigma_Y = \sqrt{\sum_{i=1}^{n} a_i^2\sigma_i^2}$$

Question 6 *(3 points)* – **LECTRUE 4**

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

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

$$E(X) = \frac{(0+1)}{2} = 0.5, \quad Var(X) = \frac{(1-0)^2}{12} = \frac{1}{12}$$

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

$$E(e^x) = \int_{-\infty}^{+\infty} e^t \cdot p(t)dt$$

$p(x)$ - probability density function of $X \sim U(0,1)$:

$$p(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

Therefore,

$$E(e^x) = \int_{-\infty}^{+\infty} e^t \cdot p(t)dt = \int_0^1 e^t \cdot 1dt = e^1 - e^0 = e - 1 \approx 1.72$$

Question 7 *(3 points)* – **LECTRUE 6, PROGRESS TEST 2**

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) $P(Z \leq 1.53)$

$$P(Z \leq 1.53) = P(Z < 0) + P(0 \leq Z \leq 1.53) = \mathbf{0.5} + \boldsymbol{G}(\mathbf{1.53})$$

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

$$P(Z \leq -0.5) = P(Z \geq 0.5) = P(Z \geq 0) - P(0 < Z \leq 0.5)$$
$$= \mathbf{0.5} - \boldsymbol{G}(\mathbf{0.5})$$

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

$$P(-1 < Z \leq 0.5) = P(Z \leq 0.5) - P(Z \leq -1)$$

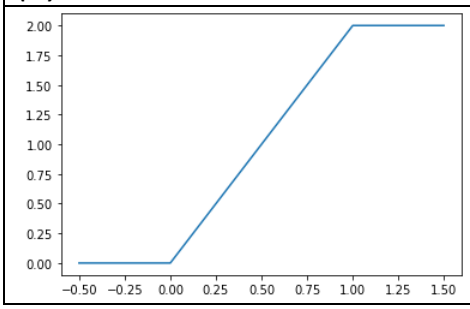
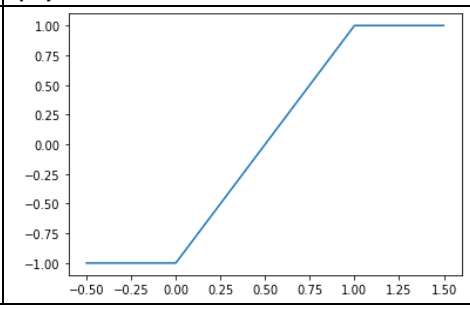
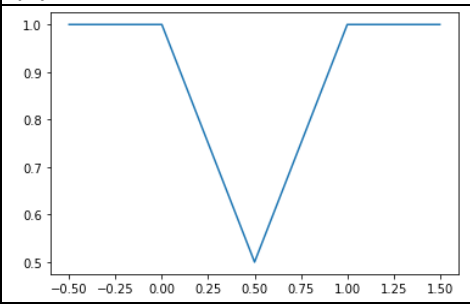
$$P(Z \leq 0.5) = P(Z < 0) + P(0 \leq Z \leq 0.5) = 0.5 + G(0.5)$$

$$P(Z \leq -1) = P(Z > 1) = P(Z \geq 0) - P(0 < Z \leq 1) = 0.5 - G(1)$$

$$P(-1 < Z \leq 0.5) = 0.5 + G(0.5) - 0.5 + G(1) = \boldsymbol{G}(\mathbf{1}) + \boldsymbol{G}(\mathbf{0.5})$$

Question 8 *(3 points)* **– LECTRUE 3, PROGRESS TEST 2**

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)*
		

(a) **Isn't a valid CDF**: CDF shows probability, and therefore should take values between 0 and 1, while the unction on the plot goes up to 2.

(b) **Isn't a valid CDF**: CDF shows probability, and therefore should take values between 0 and 1, while the unction on the plot goes down to -1.

(c) **Isn't a valid CDF**: CDF must be a non-decreasing function, while the function on the plot is non-increasing between -0.5 and 0.5 and non-decreasing between 0.5 and 1.5

Question 9 *(3 points)* **– LECTRUE 4**

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.

There're two basic properties of a PDF f(x):

1. PDF is a non-negative function: $f(x) \geq 0$
This property is satisfied

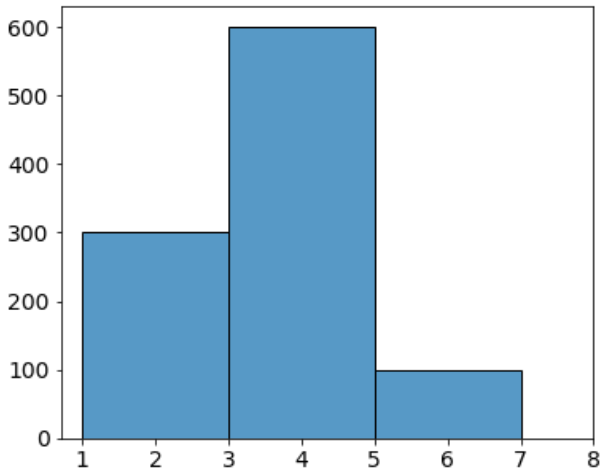
2. Area under the PDF curve should be 1: $\int_{-\infty}^{+\infty} f(x)dx = 1$
Let's check this:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{0}^{10} 0.2dx = 0.2x|_0^{10} = 0.2 * 10 - 0.2 * 0 = 2 \neq 1$$

Thus, such a function isn't a valid PDF.

Question 10 *(3 points)* – **LECTRUE 4**

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.

Currently, the bin heights equal raw frequency: 300, 600 and 100, and the width of each bin is 3.

To get a histogram that approximates the PDF of the underlying distribution, we need to transform them into relative frequencies normalized by the bin width:

$$Height_{bin1} = \frac{300}{1000} \cdot \frac{1}{2} = 0.15$$

$$Height_{bin1} = \frac{600}{1000} \cdot \frac{1}{2} = 0.3$$

$$Height_{bin1} = \frac{100}{1000} \cdot \frac{1}{2} = 0.05$$

When bin heights are normalized like that, the total area sums up to 1, as it should for the PDF:

$$0.15 * 2 + 0.3 * 2 + 0.0 * 2 = 0.3 + 0.6 + 0.1 = 1$$

Note that this is exactly the histogram you'd get if you plot it in Python with the appropriate normalization (*stats='density'*).

Question 11 *(5 points)* – **LECTRUE 2, GRADED ASSIGNMENT 2**

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.

Suppose that we have samples $X_1, X_2, \dots, X_n$.

A likelihood function $L(\theta)$ is a joint probability of the data given the model:

$$L(\theta) = \prod_{i=1}^{n} P(X = X_i|\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{X_i} = \theta^n \prod_{i=1}^{n} (1-\theta)^{X_i}$$

Log-likelihood:

$$\log L(\theta) = \log\left[\theta^n \prod_{i=1}^{n} (1-\theta)^{X_i}\right] = n\log\theta + \log(1-\theta)\sum_{i=1}^{n} X_i$$

Maximizing log-likelihood:

$$\frac{d}{d\theta}\log L(\theta) = \frac{n}{\theta} - \frac{\sum_{i=1}^{n} X_i}{1-\theta} = 0$$

$$(1-\theta)n - \theta\sum_{i=1}^{n} X_i = 0$$

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

(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?

$$\hat{\theta} = \frac{n}{n + \sum_{i=1}^{n} X_i} = \frac{1000}{1000 + 980} = \frac{1000}{1980} \sim 0.505$$

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

$$E(X) = \frac{1-\theta}{\theta} = \mu$$

$$\hat{\theta} = \frac{n}{n + \sum_{i=1}^{n} X_i} \Longrightarrow$$

$$\hat{\mu} = \frac{1-\hat{\theta}}{\hat{\theta}} = \frac{(\sum_{i=1}^{n} X_i)(n + \sum_{i=1}^{n} X_i)}{(n + \sum_{i=1}^{n} X_i)n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{980}{1000} = 0.98$$

Question 12 *(5 points)* – **LECTRUE 6, GRADED ASSIGNMENT 3**

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.

As has been derived in class, an MLE of the μ parameter is the sample mean: $\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and

And the MLE of the variance σ^2 is the sample variance:
$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$

Therefore, in this example

$$\hat{\mu}_{MLE} = \frac{1}{10}\sum_{i=1}^{10} X_i = \frac{1}{10} * 846 = 84.6$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{10}\sum_{i=1}^{10} X_i^2 - \left(\frac{1}{10}\sum_{i=1}^{10} X_i\right)^2 = \frac{1}{10} * 71607 - 84.6^2 = 3.54$$

(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.

$$P(X \le 83) = P(X - \mu \le 83 - \mu) = P\left(\frac{X-\mu}{\sigma} \le \frac{83-\mu}{\sigma}\right) = \Phi\left(\frac{83-\mu}{\sigma}\right)$$

If we use the MLE of the parameters μ and σ obtained at the previous step, we'll get:

$$P(X \le 83) \approx \Phi\left(\frac{83 - 84.6}{\sqrt{3.54}}\right) \approx \Phi(0.85)$$

Question 13 *(5 points)* – **LECTRUE 7, GRADED ASSIGNMENT 3**

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.

An estimator $T(X)$ is unbiased if it's expected value equals the parameter it's estimating. Let's check this property:

$$E(T_1) = E\left(\frac{X_1 + X_3}{2}\right) = \frac{1}{2}\big(E(X_1) + E(X_3)\big) = \frac{1}{2}(\mu + \mu) = \mu \Longrightarrow$$

T_1 is an unbiased estimator of the μ parameter

$$E(T_2) = E\left(\frac{X_1 + 2_{X_2} + 3X_3}{6}\right) = \frac{1}{6}\big(E(X_1) + 2E(X_2) + 3E(X_3)\big)$$
$$= \frac{1}{6}(\mu + 2\mu + 3\mu) = \mu \Longrightarrow$$

T_2 is an unbiased estimator of the μ parameter

(b) Compare the variances of T_1 and T_2.

$$Var(T_1) = Var\left(\frac{X_1 + X_3}{2}\right) = \frac{1}{4}\big(Var(X_1) + Var(X_3)\big) = \frac{1}{4}(\sigma^2 + \sigma^2)$$
$$= \frac{1}{2}\sigma^2$$

$$Var(T_2) = Var\left(\frac{X_1 + 2X_2 + 3X_3}{6}\right)$$
$$= \frac{1}{36}\big(Var(X_1) + 4Var(X_2) + 9Var(X_3)\big) = \frac{1}{36}14\sigma^2 = \frac{7}{18}\sigma^2$$

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

Both estimators are unbiased, but T_2 has a smaller variance than T_1:

$$Var(T_2) = \frac{7}{18}\sigma^2 < \frac{1}{2}\sigma^2 = Var(T_1)$$

Therefore, T_2 should be preferred over T_1, as it's a more 'stable' estimator of the unknown parameter μ.