

BI Norwegian Business School

FINAL EXAM: **EBA 2904 – Statistics with programming**

WITH: **Emil Aas Stoltenberg**

DAY OF EXAMINATION: **June 6, 2023**

EXAMINATION HOURS: **Five hours**

PERMITTED AIDS: **A bilingual dictionary**

INSTRUCTIONS: **Write with a pen, NOT a pencil. Be concise.**

This exam set contains three exercises and comprises five pages. An appendix with results that can be pointed to in solving the exercises is included on pages three and four.

Exercise 1. Consider the sample space associated with flipping a coin and rolling a die,

$$\Omega = \{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\},$$

where H and T stand for heads and tails, respectively, while the numbers indicate the number of eyes shown by the die. For any event A , let $\#(A)$ be the number of elements in A . Consider the function

$$\Pr(A) = \frac{\#(A)}{12},$$

defined for all events A , that is, for all subsets A of the sample space Ω . Note that the elements ω of Ω are of the form

$$\omega = (\omega_1, \omega_2) = (\text{what the coin shows}, \text{what the die shows}).$$

- (a). Show that \Pr is a probability function.
- (b). Consider the event $A = \{\text{the coin shows heads}\}$. Find the probability of the event A .
- (c). Consider the event $B = \{\text{die shows an even number of eyes}\}$. What are the elements of the event B ? Show that A and B are independent events.
- (d). Introduce the random variable $X: \Omega \rightarrow \{0, 2, 4, 6\}$, giving the outcome of the die if the outcome is even, and zero otherwise. More formally,

$$X(\omega) = I_B(\omega)\omega_2,$$

where B is as defined in (c). Find the probability mass function of X . In other words, compute the probabilities $\Pr(X = x)$ for all $x \in \mathbb{R}$.

(e). Let X be as in (d). Make a sketch of the cumulative distribution function $F(x) = \Pr(X \leq x)$. Clearly indicate the relevant numbers on the x -axis and on the y -axis.

(f). Let the event A be as defined in Ex. (b), and consider the random variable $Y: \Omega \rightarrow \{-1, 1\}$ defined by

$$Y(\omega) = 2I_A(\omega) - 1,$$

where I_A is the indicator function of A . Find the expectation and the variance of Y .

(g). Let X and Y be as defined in (d) and (f), respectively. Show that

$$\Pr(X = x, Y = 1) = \Pr(X = x) \Pr(Y = 1),$$

for all $x \in \mathbb{R}$, and explain why this entails that X and Y are independent.

(h). What is the covariance of X and Y ?

Exercise 2. Consider the function

$$f(x) = \frac{3}{\theta} \left(\frac{x}{\theta} \right)^2, \quad \text{for } x \in [0, \theta],$$

and $f(x) = 0$ for x outside of $[0, \theta]$, where $\theta > 0$ is an unknown parameter that we want to estimate.

(a). Verify that $f(x)$ is a probability density function (pdf).

(b). Let X be a random variable with distribution determined by the pdf $f(x)$. Find the cumulative distribution function (cdf) $F(x)$ of X .

(c). Show that the expectation of X is

$$\mathbb{E}(X) = \frac{3}{4}\theta.$$

The variance of X is $\text{Var}(X) = \frac{3}{80}\theta^2$, but this you are not asked to show.

(d). Let X_1, \dots, X_n be $n \geq 2$ independent random variables with the same distribution as X . Let $\bar{X}_n = (X_1 + \dots + X_n)/n$ be the empirical mean of these, and consider the estimator of θ given by

$$\hat{\theta}_n = \frac{4}{3}\bar{X}_n.$$

Find the expectation and the variance of this estimator.

(e). Let Y_n be the maximum of the n random variables X_1, \dots, X_n , that is

$$Y_n = \max(X_1, \dots, X_n).$$

Since X_1, \dots, X_n are independent random variables, the events $\{X_1 \leq y\}, \dots, \{X_n \leq y\}$ are also independent, for any y . Show that the cdf of Y is given by

$$G(y) = \begin{cases} 0, & \text{for } y < 0 \\ \left(\frac{y}{\theta}\right)^{3n}, & \text{for } 0 \leq y < \theta \\ 1, & \text{for } y \geq \theta. \end{cases}$$

(f). Find an expression for the probability density function, say $g(y)$, of Y .

(g). Let $k \geq 1$ be some number. Show that the expectation $\mathbb{E}(Y_n^k)$ is

$$\mathbb{E}(Y_n^k) = \frac{3n}{3n+k}\theta^k.$$

(h). Find expressions for the expectation and the variance of Y_n .

(i). Another estimator for the unknown parameter θ is

$$\tilde{\theta}_n = \frac{3n+1}{3n} Y_n.$$

In terms of the mean squared error, which of the two estimators $\hat{\theta}$ and $\tilde{\theta}$ is the best one for estimating θ , and why?

(j). The Python code below generates a random sample of size n , contained in `xx`, from the distribution with pdf $f(x)$ and parameter θ . Explain why these simulations can set us on track of an answer to Ex. 2(i). Comment also on the limitations of this approach. Points (A.11) and (A.13) are of relevance for these questions.

```
import numpy as np
n = 23
theta = 2.34
sims = 10**3
thetahats = np.zeros(sims)
thetatildes = np.zeros(sims)

for jj in range(0,sims):
    uu = np.random.uniform(0,1,n)
    xx = theta*uu**(1/3.) # the data
    thetahats[jj] = (4./3.)*np.mean(xx)
    thetatildes[jj] = (3*n + 1)/(3.*float(n))*np.max(xx)

mse_hat = np.mean((thetahats - theta)**2)
mse_tilde = np.mean((thetatildes - theta)**2)
```

(k). Let $\varepsilon > 0$ be some number, and Y_n be as defined in Ex. (e). Show that

$$\Pr(|Y_n - \theta| \geq \varepsilon) \rightarrow 0,$$

as n tends to infinity. In doing so, (A.11) and (A.12) in the appendix is of help.

Exercise 3. The municipality of Oslo has two instruments at its disposal for testing whether the water at Sørenga sjøbad – a popular place to take a bath, close to the Opera and Munch – is clean or dirty. Instrument *A* has a ninety percent accuracy in detecting dirty water, but, unfortunately for the sunbathers of Oslo, its probability of falsely claiming that the water is dirty is twenty percent. The corresponding numbers for Instrument *B* are eighty and ten percent, respectively. Recent records for the month of July, due to more intensive lab investigations, show that the water at Sørenga is dirty in one out of one hundred July days.

(a). On the morning of July 1, the water is dirty according to the test performed with Instrument *A*. What is the probability that the water is dirty? Given reports of a positive test (i.e., a test saying dirty water) on the radio, would you go swim?

(b). Which of the two instruments minimises the probability of a wrong test result?

(c). The summer intern testing the water for the municipality of Oslo is sloppy, and forgets which instrument she brought with her on the morning of July 14. Since both instruments are stored the same place, she assess that there is a fifty-fifty chance that what she brought with her is Instrument *A*. According to the test performed by what could be Instrument *A* or *B*, the water is dirty. What is the probability that the water is dirty?

APPENDIX

(A.0) The function $\Pr(\cdot)$ taking events as its arguments, is a probability function if (i) $\Pr(A) \geq 0$ for all events A ; (ii) $\Pr(\Omega) = 1$ for the sample space Ω ; and (iii) if A and B are disjoint events (i.e. $A \cap B = \emptyset$), then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

(A.1) For any probability function and events A and B , we have

- (a) $\Pr(\emptyset) = 0$.
- (b) $\Pr(A) \leq 1$.
- (c) $\Pr(A) = 1 - \Pr(A^c)$.
- (d) $\Pr(B \setminus A) = \Pr(B) - \Pr(A \cap B)$.
- (e) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.
- (f) If $A \subset B$ then $\Pr(A) \leq \Pr(B)$.

(A.2) The conditional probability of A given B is

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \quad \text{provided } \Pr(B) > 0.$$

This definition gives the multiplication rule $\Pr(A \cap B) = \Pr(A | B)\Pr(B)$. A conditional probability function $B \mapsto \Pr(B | A)$ is a probability function.

(A.3) Let A and B be two events, such that $0 < \Pr(B) < 1$. The law of total probability says that

$$\Pr(A) = \Pr(A | B)\Pr(B) + \Pr(A | B^c)\Pr(B^c),$$

and if B_1, \dots, B_k is a partition of the sample space Ω , i.e., $B_i \cap B_j = \emptyset$ whenever $i \neq j$ and $B_1 \cup \dots \cup B_k = \Omega$, then $\Pr(A) = \sum_{j=1}^k \Pr(A | B_j)\Pr(B_j)$.

(A.4) Let A and B be two events such that $\Pr(A) > 0$ and $\Pr(B) > 0$. Bayes' formula reads

$$\Pr(B | A) = \frac{\Pr(A | B)\Pr(B)}{\Pr(A)}.$$

(A.5) Two events A and B are independent if $\Pr(A \cap B) = \Pr(A)\Pr(B)$. The events A_1, \dots, A_n are mutually independent (or just independent) if

$$\Pr(A_{j_1} \cap \dots \cap A_{j_k}) = \Pr(A_{j_1}) \cdots \Pr(A_{j_k}),$$

for any subset $\{j_1, \dots, j_k\}$ of $\{1, \dots, n\}$.

(A.6) The expectation of a discrete random variable W taking the values $\{w_1, w_2, \dots\}$ is

$$\mathbb{E}(W) = \sum_{\omega \in \Omega} W(\omega)\Pr(\omega) = \sum_{j=1}^{\infty} w_j \Pr(W = w_j).$$

The expectation of a continuous random variable Z with probability density function $f(z)$ is

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} z f(z) dz.$$

The variance of any random variable Z is

$$\text{Var}(Z) = \mathbb{E}\{(Z - \mathbb{E}[Z])^2\} = \mathbb{E}\{Z^2\} - (\mathbb{E}\{Z\})^2.$$

(A.7) The indicator function of an event A , denoted I_A , is

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation of an indicator function is $E(I_A) = \Pr(A)$. For any two events A and B , the product of their indicator functions is $I_A I_B = I_{A \cap B}$.

(A.8) Let X_1, \dots, X_k be random variables. They are independent if

$$\Pr(X_1 \leq x_1, \dots, X_n \leq x_n) = \Pr(X_1 \leq x_1) \cdots \Pr(X_n \leq x_n),$$

for all x_1, \dots, x_n in \mathbb{R} .

(A.9) If X and Y are two independent random variables, then $E(XY) = E(X)E(Y)$.

(A.10) The cumulative distribution function of a random variable X is the function $F(x) = \Pr(X \leq x)$. If X is a discrete random variable taking the values $\{x_1, x_2, \dots\}$, then

$$F(x) = \sum_{j: x_j \leq x} \Pr(X = x_j).$$

If X is a continuous random variable with pdf $f(x)$, then

$$F(x) = \int_{-\infty}^x f(y) \, dy.$$

In particular, the pdf of a continuous random variable can be found by differentiating its cdf, i.e.,

$$f(x) = F'(x).$$

(A.11) If $\hat{\theta}_n$ is an estimator of θ , the mean squared error of $\hat{\theta}_n$ is

$$\text{mse} = E\{(\hat{\theta}_n - \theta)^2\}.$$

It is always true that $\text{mse} = \text{Var}(\hat{\theta}_n) + \{E(\hat{\theta}_n) - \theta\}^2$. Notice that the mse is a function of θ . Small mse is good.

(A.12) Markov's inequality: Let X be a nonnegative random variable. For any $\varepsilon > 0$, we have that $\Pr(X \geq \varepsilon) \leq E(X)/\varepsilon$. We deduce that for any random variable Y , we have that for any $\varepsilon > 0$

$$\Pr(|Y| \geq \varepsilon) \leq E(Y^2)/\varepsilon^2.$$

(A.13) The law of large numbers: If X_1, \dots, X_n are independent and identically distributed random variables with expectation θ , then $(1/n) \sum_{i=1}^n X_i$ converges in probability to θ .