<u>WISB263 Mathematical Statistics</u>
Exam V2
25th June 2020
Lecturer: Dr. Wioletta Ruszel
Teaching assistants: Leandro Chiarini, Ben Balkenende, Esther Steenkamer

Exercise 1 (25 p.) The average energy expenditures for eight elderly subjects were estimated on the basis of information received from a battery-powered heart rate monitor that each subject wore. Two overall averages were calculated for each subject, one for the summer months and one for the winter months, as shown in the following table. Let μ_D denote the difference between the summer and winter energy expenditure populations.

V2		
Average daily energy expediture		
Subject	Summer x_	Winter y_i
1	1424	1458
2	1501	1353
3	1495	2209
4	1739	1804
5	2031	1912
6	934	1366
7	1401	1598
8	1339	1406

1. (10 p.) Perform a hypothesis test "TEST 1" for $(H_0) : \mu_D = 0$ against $(H_1) : \mu_D \neq 0$ at $\alpha = 0.05$ and determine the exact rejection region. Is there a smallest level of significance to reject H_0?
2. (10 p.) Perform a hypothesis test "TEST 2" for $(H_0) : \mu_D = 0$ against $(H_1) : \mu_D \neq 0$ now based on the assumption that the samples are normally distributed. The sample of summer months expeditures $X_1, ..., X_8$ has distribution $N(\mu_X, \sigma_X^2)$ and the sample of winter months expeditures $Y_1, ..., Y_8$ has distribution $N(\mu_Y, \sigma_Y^2)$ with σ_X, σ_Y unknown and $\mu_D = \mu_X - \mu_Y$. For which α do you reject/accept H_0? Compute the p-value. Determine the rejection region for $\alpha = 0.05$.
3. (5 p.) Compare the tests "TEST 1" and "TEST 2". Which one is more appropriate?

Solution:

1. We perform a Wilcoxon signed rank test (2 points) because we do not assume anything about the distribution and the observations x_i, y_i are not independent (1 point). The ranks of the differences are:

V2						
Average daily energy expediture						
Subject	Summer x_	Winter y_i		D	Rank	Sign
1	1424	1458		-34	1	-
2	1501	1353		148	5	+
3	1495	2209		-714	8	-
4	1739	1804		-65	2	-
5	2031	1912		119	4	+
6	934	1366		-432	7	-
7	1401	1598		-197	6	-
8	1339	1406		-67	3	-

(2 points)

We get that $w^+ = 9$ (1 point). For $\alpha = 0.05$ and $n = 8$, $\mathbb{E}(W^+) = 18$ (1 point) we obtain a rejection region of $\{W^+ < 5\} \cup \{W^+ > 31\}$. (1 points) For our case we do not have enough evidence to reject H_0. We reject if w^+ is in the rejection region, which will not happen at any positive probability smaller than 0.5 (1 point). For a two-sided test we find $p > 0.05$. (1 point)

2. We perform a pair test as in Theorem 12.4 (1 point) The test statistic is equal to

$$t(\mathbf{x}, \mathbf{y}) = \frac{-155.25}{\sqrt{10506.13}} \approx -1.51. \text{ (3 points)}$$

We have to compare it with a t-quantile with 7 degrees of freedom. We see that $t_7(0.9) = 1.415$ (1 point) and $t_7(0.95) = 1.895$. (1 point) For $\alpha \leq 0.1$ we accept H_0 and for $\alpha \geq 0.2$ we reject H_0 (2 points). The p-value is equal to 0.18. For $\alpha = 0.05$ we have $t_7(0.975) = 2.36$ so the rejection region is equal to $(-\infty, -2.36) \cup (2.36, \infty)$ (2 points).

3. The first test is more appropriate because we do not assume anything about the distribution (2 points). Moreover the p-value for the first test is much smaller than the one for the second test. (3 points)

Exercise 2 (25 p.) Consider an i.i.d. sample $X_1, ..., X_n$ on some probability space with common density given by

$$h_\theta(x_1) = \frac{g(\theta)}{x_1^4}, \ x_1 \geq \theta,$$

where $\theta > 0$.

1. (2 p.) Determine $g(\theta)$ as a function of θ.
2. (5 p.) Determine the MLE $\hat{\theta}_n$ of θ and $\hat{\theta}_{n,1}$ for θ^3.
3. (5 p.) Calculate the cumulative distribution function of $\hat{\theta}_n$ for all $x \in \mathbb{R}$ and determine its density.
4. (4 p.) Is $\hat{\theta}_n$ biased? Show that $\hat{\theta}_n$ is asymptotically unbiased.
5. (4 p.) Find the MoM estimator θ_n^* of θ and compute the variance.
6. (5 p.) Which estimator is more efficient for $n \geq 2$? Can we apply Cramer-Rao Theorem in this case?

Solution:

1. Since $\int_\theta^\infty h_\theta(x_1)dx_1 = 1$ we have that $g(\theta) = 3\theta^3$. (2 points)
2. We cannot take derivatives here. The likelihood function is equal to
$$L(\theta;\mathbf{x}) = \frac{3^n\theta^{3n}}{\prod_{i=1}^n x_i^4}\prod_{i=1}^n \mathbb{1}_{x_i\geq\theta} = \frac{3^n\theta^{3n}}{\prod_{i=1}^n x_i^4}\mathbb{1}_{x_{(1)}\geq\theta}.\text{(1 point)}$$
so the MLE of θ is $\hat{\theta}_n = X_{(1)}$ (2 points) and by invariance principle the MLE of θ^3 is $X^3_{(1)}$ (2 points).
3. We have for $x > \theta$
$$F_{X_{(1)}}(x) = 1 - \mathbb{P}(X_{(1)} > x) = 1 - (1 - F_{X_1}(x))^n \text{ (2 points)}$$
and
$$F_{X_1}(x) = \int_\theta^x \frac{3\theta^3}{y^4}dy = 1 - \frac{\theta^3}{x^3} \text{ (1 point)}$$
so we have for general $x \in \mathbb{R}$
$$F_{X_{(1)}}(x) = \begin{cases} 0, & \text{if } x < \theta \\ 1 - \frac{\theta^{3n}}{x^{3n}}, & \text{if } \theta \leq x. \end{cases} \text{ (1 point)}$$
The density is $f_{X_{(1)}}(x) = 3n\frac{\theta^{3n}}{x^{3n+1}}$ for $x > \theta$ note that we take the derivative so f is defined for $x > \theta$ and not for $x \geq \theta$. (1 point)
4. We have that
$$\mathbb{E}(X_{(1)}) = \int_\theta^\infty x f_{X_{(1)}}(x)dx = \frac{3n\theta}{3n-1}\text{(2 points)}$$
so the estimator is biased (1 point). $\lim_{n\to\infty}\mathbb{E}(X_{(1)}) = \theta$, (1 point).
5. We have that $\mathbb{E}(X_1) = \frac{3}{2}\theta$ (1 point) so the estimator is equal to $\theta_n^* = \frac{2}{3}\overline{X}_n$ (1 point). For the variance note that
$$\mathbb{E}(X_1^2) = \int_\theta^\infty x^2\frac{3\theta^3}{x^4}dx = 3\theta^2$$
so $Var(X_1) = \frac{3}{4}\theta^2$ (1 point) and $Var(\theta_n^*) = \frac{\theta^2}{3n}$ (1 point).
6. We cannot use the Cramer-Rao Theorem because the support of f depends on θ (1 point). The MSE for $\hat{\theta}_n$ is equal to
$$MSE(\hat{\theta}_n) = Var(\hat{\theta}_n) + b^2(\theta) = \frac{3n\theta^2}{(3n-2)(3n-1)^2} + \frac{\theta^2}{(3n-1)^2} = \frac{2\theta^2}{2-9n+9n^2} \text{ (2 points)}$$
and of the MoM is equal to $MSE(\theta_n^*) = \frac{\theta^2}{3n}$(1 point) hence the MLE is more efficient (1 point).

Exercise 3 (25 p.) Assume we have an estimator $\hat{X}_n$ for a parameter $\theta > 0$ of the density of an i.i.d. sample $X_1, ..., X_n$. The density of the estimator is
$$f_\theta(x) = 3n\frac{\theta^{3n}}{x^{3n+1}},\ x > \theta,$$
which means that $\mathbb{P}(\hat{X}_n \in (a,b)) = \int_a^b f_\theta(x)dx$ for $\theta < a < b$. The expected value is equal to $\mathbb{E}(\hat{X}_n) = \frac{3n\theta}{3n-1}$ and variance $Var(\hat{X}_n) = \frac{3n\theta^2}{(3n-2)(3n-1)^2}$.

1. (4 p.) Show that $\hat{X}_n$ is consistent.

2. (4 p.) Prove that $\frac{\hat{X}_n - \mathbb{E}(\hat{X}_n)}{\sqrt{Var(\hat{X}_n)}} \xrightarrow{d} Y$ as $n \to \infty$, where Y has density $f_Y(y) = e^{-(1+y)}$. (Hint: You can use that $\lim_{n\to\infty} \left(1 + \frac{1+a}{n}\right)^{-n} = \lim_{n'\to\infty} \left(1 + \frac{1+a\sqrt{\frac{n'}{n'-2}}}{n'-1}\right)^{-n'}$, where $n' = c \cdot n$ for some $c \geq 1$ and $a \in \mathbb{R}$.)

3. (2 p.) Show that the support of $f_Y(\cdot)$ is $(-1, \infty)$.

4. (3 p.) Let $\alpha \in (0,1)$. Find constants c_1, c_2 such that $\mathbb{P}(Y < c_1) = \alpha$ and $\mathbb{P}(Y > c_2) = \alpha$. What is the probability $\mathbb{P}(Y \in [c_1, c_2])$?

5. (4 p.) Use the asymptotic distribution of $\frac{\hat{X}_n - \mathbb{E}(\hat{X}_n)}{\sqrt{Var(\hat{X}_n)}}$ to construct an asymptotic random confidence interval for θ with confidence $1-\alpha$ of the form (C, ∞).

6. (4 p.) Define a formal hypothesis test with significance α for testing $(H_0) : \theta \geq \theta_0$ against $(H_1) : \theta < \theta_0$. Describe the rejection region implicitely.

7. (1 p.) Determine now the asymptotic rejection region using knowledge from (2) and (5).

8. (3 p.) Assume we get from our data $t(\mathbf{x}) = \ln(10) - 1$. What is the smallest level of significance to reject H_0 in this case? How do we call this value? (Hint: You can use the asymptotic distribution of the test statistic).

Solution:

1. Note that the estimator is the MLE from the first exercise. For showing consistency we compute

$$\mathbb{P}_\theta(|\hat{X}_n - \theta| > \epsilon) \leq \frac{\mathbb{E}_\theta((\hat{X}_n - \theta)^2)}{\epsilon^2} = \frac{Var_\theta(\hat{X}_n) + b^2(\theta)}{\epsilon^2} = \frac{2\theta^2}{\epsilon^2(3n-1)(3n-2)} \to 0$$

as $n \to \infty$. (4 points)

2. To determine the limit compute for $y > \theta$

$$\mathbb{P}_\theta\left(\frac{\hat{X}_n - \mathbb{E}_\theta(\hat{X}_n)}{\sqrt{Var_\theta(\hat{X}_n)}} \leq y\right) = \mathbb{P}_\theta\left(\hat{X}_n \leq \mathbb{E}_\theta(\hat{X}_n) + y\sqrt{Var_\theta(\hat{X}_n)}\right) \text{ (2 points)}$$

$$= 1 - \left(1 + \frac{1 + y\sqrt{\frac{3n}{3n-2}}}{3n-1}\right)^{-3n} \longrightarrow 1 - e^{-(1+y)} \text{ (1 point)}$$

provided $\mathbb{E}_\theta(\hat{X}_n) + y\sqrt{Var_\theta(\hat{X}_n)} > \theta$ (1 point).

3. The support of Y is determined by the previous equation: $\mathbb{E}_\theta(\hat{X}_n) + y\sqrt{Var_\theta(\hat{X}_n)} > \theta$ (1 point) is equivalent to

$$y > \frac{-\theta}{\theta\sqrt{\frac{3n}{3n-2}}} = -\sqrt{\frac{3n-2}{3n}}$$

for $n \to \infty$ we have that $y > -1$ (1 point).

4. First of all note that $c_1, c_2 > -1$ otherwise the probabilities are 0. Compute

$$\alpha = \int_{-1}^{c_1} e^{-(1+y)} dy \;\Leftrightarrow c_1 = \ln\left(\frac{1}{e(1-\alpha)}\right) \text{ (1 point)}$$

and

$$\alpha = \int_{c_2}^{\infty} e^{-(1+y)} dy \;\Leftrightarrow c_2 = \ln\left(\frac{1}{e\alpha}\right). \text{ (1 point)}$$

The probability is equal to $\mathbb{P}(Y \in [c_1, c_2]) = 1 - 2\alpha$. (1 point)

5. Since we have asymptotically

$$\mathbb{P}_\theta\left(\frac{\hat{X}_n - \mathbb{E}_\theta(\hat{X}_n)}{\sqrt{Var_\theta(\hat{X}_n)}} \geq c_2\right) \approx \mathbb{P}(Y \geq c_2) = 1 - \alpha, \text{ (1 point)}$$

the random confidence interval is equal to $\mathscr{G}(\mathbf{X}) = \left[\frac{(3n-1)\hat{X}_n}{3n + c_2\sqrt{\frac{3n}{3n-2}}}, \infty\right)$. (3 points)

6. Define the test statistic

$$T(\mathbf{X}) := \frac{\hat{X}_n - \mathbb{E}_\theta(\hat{X}_n)}{\sqrt{Var_\theta(\hat{X}_n)}}. \text{ (1 point)}$$

We can define a hypothesis test ϕ at significance α by

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{if } t(\mathbf{x}) < c' \\ 0 & \text{otherwise (1 point)} \end{cases}$$

where c' are such that $\alpha = \mathbb{P}_{\theta_0}(T(\mathbf{X}) < c'_1)$ (1 point). The rejection region is equal to

$$\{T(\mathbf{X}) < c'_1\}. \text{ (1 point)}$$

7. The asymptotic rejection region is equal to $\{T(\mathbf{X}) < c_2\}$. (1 point)

8. We are looking for the p-value.(1 point) From Definition 9.3 we have that

$$p(\mathbf{x}) = F(t(\mathbf{x})) = \frac{9}{10}. \text{ (2 points)}$$

Exercise 4 (25 p.) Let $X_1, ..., X_n$ be an i.i.d. random sample on some probability space with common density

$$f_\theta(x_1) = \frac{x_1}{\theta^2} e^{-\frac{x_1^2}{2\theta^2}}, \; x_1 \geq 0,$$

and $\theta > 0$. We want to compare two hypothesis tests for $(H_0): \theta = 1$ against $(H_1): \theta \neq 1$ using two different test statistics, namely

$$T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i^2, \text{ and } \; T_2(\mathbf{X}) = 2\ln(R(\mathbf{X})),$$

where $R(\mathbf{X})$ is the generalized likelihood ratio. You can use that the MLE is equal to $\hat{\theta}_n = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2}$ and is asymptotically normal.

1. (3 p.) Define a formal hypothesis test based on statistic $T_1(\mathbf{X})$ at significance α.
2. (2 p.) Show that the MGF of X_1^2 is equal to $\frac{1}{1-2t\theta^2}$ for $|t| < \frac{1}{2}$.
3. (3 p.) Determine the distribution of $T_1(\mathbf{X})$ under H_0 and show that it equal to $\frac{1}{\theta^2}T_1(\mathbf{X})$ under H_1.
4. (4 p.) Let $\alpha = 0.05$ and $n = 60$ determine the critical region for this test. Compute the power when $\theta = \sqrt{2}$. If you did not find the distribution before use the normal distribution $N(\theta, 1)$.
5. (4 p.) Consider now the second test statistic $T_2(\mathbf{X})$. Show that $T_2(\mathbf{X})$ is of the form $c_1 \ln(c_2 \sum_{i=1}^n X_i^2) + c_3 \sum_{i=1}^n (X_i^2 - c_4)$ where c_1, c_2, c_3, c_4 are some constants.
6. (4 p.) What is the asymptotic distribution of $T_2(\mathbf{X})$? Determine the asymptotic rejection region for $\alpha = 0.05$.
7. (5 p.) Assume we observe $\sum_{i=1}^{60} x_i^2 = 85$. Perform both tests at $\alpha = 0.05$ and decide whether to reject H_0 or not (you can use the second asymptotic test). If the null hypothesis would be 1-sided, which test is more powerfull? Argue without computing the power of the test based on $T_2(\mathbf{X})$.

Solution:

1. The test ϕ is defined by

$$\phi(\mathbf{x}) = \begin{cases} 1, & \text{if } t_1(\mathbf{x}) < c_1 \text{ or } t_1(\mathbf{x}) > c_2 \\ 0, & \text{otherwise} \end{cases} \quad \text{(1 point)}$$

where $\frac{\alpha}{2} = \mathbb{P}_{\theta=1}(T_1(\mathbf{X}) < c_1)$ and $\mathbb{P}_{\theta=1}(T_1(\mathbf{X}) > c_2) = \frac{\alpha}{2}$. (2 points)

2. We compute for $|t| < \frac{1}{2}$

$$\mathbb{E}(e^{tX_1^2}) = \int_0^\infty \frac{x_1}{\theta^2} e^{tx_1^2} e^{-\frac{x_1^2}{2\theta^2}} dx_1 = \frac{1}{\theta^2} \int_0^\infty x_1 e^{-x_1^2\left(\frac{1-2t\theta^2}{2\theta^2}\right)} dx_1 = \frac{1}{1-2t\theta^2}. \quad \text{(2 points)}$$

3. From computing the MGF we see that $T_1(\mathbf{X}) \sim \chi^2_{2n}$ under H_0 (2 points) which is the same using MGF for $\frac{1}{\theta^2}T_1(\mathbf{X})$ (1 point).

4. The critical region is equal to $[0, 91.57) \cup (152.21, \infty)$ (1 point). The power is equal to, using $\frac{1}{2}T_1(\mathbf{X}) \sim \chi^2_{2n}$ (1 point),

$$\begin{aligned}\beta_\phi(\sqrt{2}) &= \mathbb{P}_{\theta=\sqrt{2}}(T_1(\mathbf{X}) < 91.57) + \mathbb{P}_{\theta=\sqrt{2}}(T_1(\mathbf{X}) > 152.21) \\ &= \mathbb{P}_{\theta=\sqrt{2}}(0.5T_1(\mathbf{X}) < 45.785) + \mathbb{P}_{\theta=\sqrt{2}}(0.5T_1(\mathbf{X}) > 76.105) = 0 + 0.99 = 0.99. \quad \text{(2 points)}\end{aligned}$$

5. The MLE is equal to $\hat{\theta}_n = \sqrt{\frac{1}{2n}\sum_{i=1}^n X_i^2}$. The likelihood ratio is equal to

$$\begin{aligned}T_2(\mathbf{X}) &= 2\ln\left(\frac{\sup_{\theta\in\Theta} f_\theta(\mathbf{X})}{f_1(\mathbf{X})}\right) = 2\ln\left(\frac{2n}{\sum_{i=1}^n X_i^2} e^{-n+\frac{1}{2}\sum_{i=1}^n X_i^2}\right) \\ &= -2\ln\left(\frac{1}{2n}\sum_{i=1}^n X_i^2\right) + \sum_{i=1}^n (X_i^2 - 2). \quad \text{(4 points)}\end{aligned}$$

6. We want to use Theorem 9.1 to show that $T_2(\mathbf{X})$ is asymptotically χ_1^2-distributed (1 point). The MLE is asymptotically normal. Compute

$$l(\theta; x_1) = -2\ln(\theta) - \frac{x_1^2}{2\theta^2} \text{ (1 point) } .$$

It remains to show that

$$\left|\frac{\partial^2}{\partial\theta^2} l(\theta; x_1)\right| \leq \frac{2}{\theta^2} + \frac{3}{\theta^4}x_1^2,$$

is square-integrable (1 point). Since the density consists of Gaussian integrals and they converge for all powers of x_1 we have that $\mathbb{E}(X_1^4) < \infty$. The rejection region is equal to $(3.84, \infty)$ (1 point).

7. For this value we would reject using the first test (1 point) and not reject using the second test (1 point). The first test is more powerful since the family of densities is an exponential family the test based on T_1 is a u.m.p. test. (3 points)