Homework assignment 4 – Mathematical Statistics

Because there is some R involved, students should hand in the answers via an assignment on canvas. Deadline is October 5 at 11 am.

Let $X_1, \ldots, X_n$ be i.i.d. with probability density function given by

$$f(x) = \begin{cases} \frac{4\theta^4}{x^5}, & \text{if } x \geq \theta \\ 0, & \text{if } x < \theta \end{cases} \qquad \text{where } \theta > 0 \text{ is an unknown parameter}$$

a) Determine the **moment estimator of θ**, based on the dataset $X_1, \ldots, X_n$, and determine its expectation (is it unbiased?) and its variance (expressed in θ and n).

b) Show that $\widehat{\theta} = \min(X_1, \ldots, X_n)$ is the **maximum likelihood estimator (MLE) of θ**. (Start by deriving the likelihood function and explicitly mention its domain).

c) Show that the MLE $\widehat{\theta}$ in part b. is an asymptotically unbiased estimator and a consistent estimator of θ.

d) Which approximate distribution does the moment estimator (in part a.) have for large n? Use this approximate distribution to construct an approximate confidence interval for θ, at level 1-α

To detect doping in professional sports, urine samples with ratio testosterone/epitestosterone > 4 are suspicious and are subject to additional testing. Suppose that this ratio for a "clean" professional athlete follows a normal distribution with mean μ and standard deviation σ.

e) Show that the probability that a clean athlete will have a suspicious ratio is

$$1 - \Phi\left(\tfrac{4-\mu}{\sigma}\right)$$

with Φ the cumulative distribution function (c.d.f.) of the standard normal distribution.

f) Given i.i.d. data $X_1, \ldots, X_n$ from clean professional athletes, provide the formula for the 95%-confidence interval of the mean ratio and the formula for the 90%-confidence interval of the standard deviation. For these formulas you do not need to insert the explicit expressions of the quantiles.

g) Read the dataset in the supplementary file athletes.csv into R. Estimate the probability in part (e) by replacing μ and σ by the corresponding MLE and compute the confidence intervals in part f. (The best would be if you print the R script and attach it to your homework.)

Grading:	a	b	c	d	e	f	g	Total
	1	2	½ + 1½	½ +1	1	1	1+1/2	**10**

Solutions:

a) HWA 1: $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{\theta}^{\infty} x \cdot \frac{4\theta^4}{x^5} dx = \left[4\theta^4 \cdot -\frac{1}{3}x^{-3}\right]_{x=\theta}^{\infty} = 0 + \frac{4}{3}\theta = \frac{4}{3}\theta$
Or: $\theta = \frac{3}{4}E(X)$. Hence the moments estimator of θ is $\frac{3}{4}\overline{X}$

b) Consider the observations $x_1, \ldots, x_n$of the random sample, then the maximum likelihood function is
$L(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \prod_{i=1}^{n} \frac{4\theta^4}{x_i^5} = \frac{4^n\theta^{4n}}{\prod x_i^5}$, for $x_1 \geq \theta \geq 0$ and…and $x_n \geq \theta$
$L(\theta) = \begin{cases} \frac{4^n\theta^{4n}}{\prod x_i^5}, & \text{if } \theta \leq \min(x_1, .., x_n) \\ 0\ , & \text{elsewhere} \end{cases}$
$\Rightarrow\ L'(\theta) = \frac{4^n \cdot 4n\theta^{4n-1}}{\prod x_i^5} > 0$, for all $\theta \in (0, \min(x_1, .., x_n)\ ]$,
implying that L attains its maximum at the largest possible value of $\theta = \min(x_1, .., x_n)$. Hence the maximum likelihood estimator of θ is $\widehat{\theta} = \min(X_1, .., X_n)$.

c) We will use: $P(X_1 \geq m) = \int_{\theta}^{\infty} \frac{4\theta^4}{x^5} dx = \left[4\theta^4 \cdot -\frac{1}{4}x^{-4}\right]_{x=m}^{\infty} = \frac{\theta^4}{m^4}$, for all $m \geq \theta$.
$M = \min(X_1, \ldots, X_n) \Rightarrow F_M(m) = P(\min(X_1, \ldots, X_n) \leq m) = 1 - P(\min(X_1, \ldots, X_n) \geq m)$
$= 1 - P(X_1 \geq m)^n = 1 - \frac{\theta^{4n}}{m^{4n}}$, for all $m \geq \theta$ $(F_M(m) = 0,\ m < \theta)$
$\Rightarrow\ f_M(m) = \frac{4n\theta^{4n}}{m^{4n+1}}\quad (m \geq \theta)$,
$E(M) = \int_{-\infty}^{\infty} m f_M(m)\,dm = \int_{\theta}^{\infty} m \cdot \frac{4n\theta^{4n}}{m^{4n+1}} dm = \left[4n\theta^{4n} \cdot -\frac{1}{4n-1}m^{-(4n-1)}\right]_{x=\theta}^{\infty} = \frac{4n}{4n-1}\theta$
$E\left(M^2\right) = \int_{-\infty}^{\infty} m^2 f_M(m)\,dm = \int_{\theta}^{\infty} m^2 \cdot \frac{4n\theta^{4n}}{m^{4n+1}} dm = \left[4n\theta^{4n} \cdot -\frac{1}{4n-2}m^{-(4n-2)}\right]_{x=\theta}^{\infty} = \frac{4n}{4n-2}\theta^2$
and $var(M) = E\left(M^2\right) - (EM)^2 = \frac{4n}{4n-2}\theta^2 - \left(\frac{4n}{4n-1}\theta\right)^2 = \theta^2\left(\frac{4n}{4n-2} - \frac{16n^2}{(4n-1)^2}\right) = \frac{4n}{(4n-2)(4n-1)^2}\theta^2$.
Since $\lim_{n\to\infty} E(M) = \lim_{n\to\infty} \frac{4n}{4n-1}\theta = \theta$, M is asymptotically unbiased.
In addition $\lim_{n\to\infty} var(M) = \lim_{n\to\infty} \frac{4n}{(4n-2)(4n-1)^2}\theta^2 = 0$, so that $M = \min(X_1, \ldots, X_n)$ is a consistent estimator of θ.

d) HWA 1: $E(X) = \frac{4}{3}\theta$ and $var(X) = E\left(X^2\right) - (EX)^2 = 2\theta^2 - \frac{16}{9}\theta^2 = \frac{2}{9}\theta^2$.
According to the CLT $\overline{X}$ is approximately normally distributed, and so

is $\frac{3}{4}\overline{X}$, with parameters $\mu = E\left(\frac{3}{4}\overline{X}\right) = \frac{3}{4}E(X) = \theta$ and $\sigma^2 = var\left(\frac{3}{4}\overline{X}\right) = \frac{9}{16}\cdot\frac{var(X)}{n} = \frac{9}{16}\cdot\frac{2}{9}\theta^2/n = \frac{\theta^2}{8n}$

$\frac{3}{4}\overline{X} \sim N\left(\theta, \frac{\theta^2}{8n}\right) \Rightarrow Z = \frac{\frac{3}{4}\overline{X}-\theta}{\theta/\sqrt{8n}} = \frac{3\overline{X}\sqrt{8n}}{4\theta} - \sqrt{8n} \sim N(0,1)$ (approximately, for large n)

For a 1-α confidence level we can choose c, such that $P(-c < Z < c) \approx 1-\alpha$

or $c = \Phi^{-1}\left(1-\frac{1}{2}\alpha\right)$. $P\left(-c < \frac{3\overline{X}\sqrt{8n}}{4\theta} - \sqrt{8n} < c\right) \approx 1-\alpha \Leftrightarrow P\left(-c+\sqrt{8n} < \frac{3\overline{X}\sqrt{8n}}{4\theta} < c+\sqrt{8n}\right) \approx 1-\alpha$

$\Leftrightarrow P\left(\frac{1}{c+\sqrt{8n}} < \frac{4\theta}{3\overline{X}\sqrt{8n}} < \frac{1}{-c+\sqrt{8n}}\right) \approx 1-\alpha$ *(assuming that* $\sqrt{8n} > c$*)*

$\Leftrightarrow P\left(\frac{\frac{3}{2}\overline{X}\sqrt{2n}}{c+\sqrt{8n}} < \theta < \frac{\frac{3}{2}\overline{X}\sqrt{2n}}{-c+\sqrt{8n}}\right) \approx 1-\alpha$

$\left(\frac{\frac{3}{2}\overline{X}\sqrt{2n}}{c+\sqrt{8n}}, \frac{\frac{3}{2}\overline{X}\sqrt{2n}}{-c+\sqrt{8n}}\right)$ is an approximate (1-α)100% confidence interval for θ (if n is large enough).

<u>*Alternative approach:*</u> *analogously to e.g. the construction of a confidence interval for the proportion p one might estimate the standard deviation* $\frac{\theta}{\sqrt{8n}}$ *of* $\frac{3}{4}\overline{X}$*, by replacing* θ *by its unbiased estimator* $\frac{3}{4}\overline{X}$*, finding the standard error* $SE\left(\frac{3}{4}\overline{X}\right) = \frac{\frac{3}{4}\overline{X}}{\sqrt{8n}}$ *. Then:*

$P\left(-c < \frac{\frac{3}{4}\overline{X}-\theta}{SE\left(\frac{3}{4}\overline{X}\right)} < c\right) \approx 1-\alpha \iff P\left(\frac{3}{4}\overline{X} - c\cdot SE\left(\frac{3}{4}\overline{X}\right) < \theta < \frac{3}{4}\overline{X} + c\cdot SE\left(\frac{3}{4}\overline{X}\right)\right) \approx 1-\alpha$

<u>*Note:*</u> comparison of the numerical intervals if $n = 32$, $\overline{x} = 8$ and $c = 1.96$ $(1-\alpha = 0.95)$: $\left(\frac{\frac{3}{2}\overline{X}\sqrt{2n}}{c+\sqrt{8n}}, \frac{\frac{3}{2}\overline{X}\sqrt{2n}}{-c+\sqrt{8n}}\right) \approx (5.34,\ 6.84)$ and $\left(\frac{3}{4}\overline{X} - c\cdot\frac{\frac{3}{4}\overline{X}}{\sqrt{8n}} < \theta < \frac{3}{4}\overline{X} + c\cdot\frac{\frac{3}{4}\overline{X}}{\sqrt{8n}}\right) = (5.265,\ 6.735)$

e) If X is testosterone-epitestosterone ratio and Z a standard normal distribution, then $P(X > 4) = P\left(Z > \frac{4-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{4-\mu}{\sigma}\right)$

f) Formula for the confidence interval for μ (for unknown σ^2),

95%-CI(μ) = $\left(\overline{x} - c_{0.975}\frac{s}{\sqrt{n}}, \overline{x} + c_{0.975}\frac{s}{\sqrt{n}}\right)$,

where $c_{0.975}$ denotes the 0.975-quantile of the t_{99}-distribution. For the confidence interval of the standard deviation, we use

90%-CI(σ) = $\left(\sqrt{\frac{(n-1)s^2}{c_2}}, \sqrt{\frac{(n-1)s^2}{c_1}}\right)$

With c_1 and c_2 the 0.05 and 0.95 quantiles of the χ^2_{100-1} distribution.

g) See attached R script TE_ratios.R .