Problem 1. *Let $X_1, \dots, X_n$ be independent random variables from the uniform distribution on the interval $[0, \theta]$, where $\theta > 0$ is an unknown parameter.*

1. *Find the expected value and variance of the estimator $\hat{\theta} = 2\bar{X}$.*
2. *Find the expected value of the estimator $\tilde{\theta} = X_{(n)}$, ($X_{(n)}$ is the largest observation).*
3. *Find an unbiased estimator of the form $\check{\theta} = cX_{(n)}$ and calculate its variance.*
4. *Compare the mean square error of $\hat{\theta}$ and $\check{\theta}$.*
5. *Which of these estimators would you recommend and why?*
6. *Is one of the discussed estimators the maximum likelihood estimator? Justify your answer mathematically.*

Solution. We know that the expected value of a uniform random variable on $[0, \theta]$ is $\theta/2$ (Why?). Thus $\mathbb{E}\,\hat{\theta} = \theta$, i.e. $\hat{\theta}$ is an unbiased estimator of θ.

In order to find the expected value of the maximum of uniform random variables, let us first compute its density. To this end, note that for $u \in [0, \theta]$ the cdf is given by

$$F_{\tilde{\theta}}(u) = \mathbb{P}(\max_{1 \le i \le n} U_i \le u) = \mathbb{P}(U_1 \le u, ..., U_n \le u) = \mathbb{P}(Y_1 \le u)^n = (u/\theta)^n.$$

Taking the derivative with respect to u we obtain the density

$$f(u) = nu^{n-1}/\theta^n, \; u \in [0, \theta].$$

The expected value is thus

$$\mathbb{E}\,\tilde{\theta} = \int_0^\theta nu^n/\theta^n \; du = \frac{n}{n+1}\frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta.$$

Consequently, by taking $\check{\theta} = \frac{n+1}{n}\theta$ we obtain an unbiased estimator of θ.

To get the variance of this estimator, let us first compute the variance of the maximum of all observations

$$\begin{aligned}
\mathbb{V}ar\,\tilde{\theta} &= \int_0^\theta nu^{n+1}/\theta^n \; du - \frac{n^2}{(n+1)^2}\theta^2 \\
&= \theta^2 \left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2} \right) \\
&= \theta^2 \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} \\
&= \theta^2 \frac{2n^2 + n - 2n^2}{(n+2)(n+1)^2} \\
&= \theta^2 \frac{n}{(n+2)(n+1)^2}.
\end{aligned}$$

Thus the variance of $\check{\theta}$ is $\theta^2/(n(n+2))$.

The variance of $\hat{\theta}$ is $4\mathbb{V}ar(U_1)/n = 4\theta^2/(12n) = \theta^2/(3n)$.

We arrive to the conclusion that $\check{\theta}$ is an essentially better estimator than $\bar{\theta}$. It can be even argued that $\tilde{\theta}$ is a better (although biased) estimator than $\bar{\theta}$ (How?).

Let $\mathbf{1}_{[0,\theta]}(x)$ denote an indicator function of the interval $[0,\theta]$, i.e. it is one if x is in the interval and zero otherwise. The likelihood function for $\theta > 0$ and $x_i > 0$ is given by

$$\begin{aligned} l(\theta) &= \mathbf{1}_{[0,\theta]}(x_1)\cdots\mathbf{1}_{[0,\theta]}(x_n) \\ &= \mathbf{1}_{[0,\theta]}(\max x_i) \\ &= \mathbf{1}_{[\max x_i,\infty)}(\theta). \end{aligned}$$

It is clear that this function although discontinuous attains maximum at $\tilde{\theta} = \max x_i$, thus $\tilde{\theta}$ is the MLE.

Problem 2. *Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the shifted exponential distribution with probability density function*

$$f(x|\theta,\mu) = \frac{1}{\theta} e^{-(x-\mu)/\theta}, \qquad \mu < x < \infty,$$

where $\theta > 0$ and $-\infty < \mu < \infty$. Both θ and μ are unknown, and $n > 1$. The sample range W is defined as $W = X_{(n)} - X_{(1)}$, where $X_{(n)} = \max_i X_i$ and $X_{(1)} = \min_i X_i$.

1. *Show that the joint probability density function of $X_{(1)}$ and W is given by*

$$f_{X_{(1)},W}(x,w) = n(n-1)\theta^{-2} e^{-n(x-\mu)/\theta} e^{-w/\theta} (1 - e^{-w/\theta})^{n-2},$$

for $x > \mu$ and $w > 0$.

2. *Obtain the marginal density function of W and compute the cumulative distribution function of W.*
3. *Show that W/θ is a pivotal quantity for θ. Explain how this result may be used to construct a confidence interval for θ at the confidence level $100(1-\alpha)\%$, $\alpha \in (0,1)$.*
4. *Consider a sample of 10 values $5.9, 7.5, 12.7, 6.3, 5.7, 18.5, 6.0, 27.3, 6.8, 12.4$ and evaluate the 95% confidence interval for θ as discussed above.*

Solution. Note that for arbitrary jointly continuous r.v.'s X and Y, if

$$G(x,y) = P(X > x, Y < y),$$

then their joint density is given by

$$g(x,y) = -\frac{\partial^2 G}{\partial x \partial w}(x,y).$$

One can notice that for $X = X_{(1)}$, $Y = X_{(n)}$, and for $x < y$ we have

$$G(x,w) = P(x < X_1 < w)^n = \left(\int_x^w f(u)\, du \right)^n,$$

where f is the density of X_i's. Thus

$$g(x,w) = n(n-1) \left(\int_x^w f(u) du \right)^{n-2} f(x) f(w).$$

This gives joint density of $X_{(1)}$ and $X_{(n)}$ in the shifted exponential case

$$g(x,y) = n(n-1) e^{n\mu/\theta} e^{-(x+y)/\theta} \left(e^{-x/\theta} - e^{-y/\theta} \right)^{n-2}.$$

The transformation theorem for $h(x,y) = (y-x, x)$ leads straightforward to

$$f(x,w) = n(n-1)\theta^{-2}e^{-n(x-\mu)/\theta}e^{-w/\theta}(1-e^{-w/\theta})^{n-2}.$$

We have

$$f_W(w) = n(n-1)\theta^{-2}e^{-w/\theta}(1-e^{-w/\theta})^{n-2}\int_\mu^\infty e^{-n(x-\mu)/\theta}\mathrm{d}x.$$

By putting $v = (x-\mu)$ so that $\mathrm{d}v = \mathrm{d}x$, we have

$$\begin{aligned}
f_W(w) &= n(n-1)\theta^{-2}e^{-w/\theta}(1-e^{-w/\theta})^{n-2}\int_\mu^\infty e^{-nv/\theta}\mathrm{d}x \\
&= n(n-1)\theta^{-2}e^{-w/\theta}(1-e^{-w/\theta})^{n-2}\left[-\frac{\theta}{n}e^{-nv/\theta}\right]_{v=0}^\infty \\
&= \frac{(n-1)}{\theta}e^{-w/\theta}(1-e^{-w/\theta})^{n-2}.
\end{aligned}$$

Next, $P(W \le w) = \int_0^w \frac{(n-1)}{\theta}e^{-y/\theta}(1-e^{-y/\theta})^{n-2}\mathrm{d}y = \left[(1-e^{-y/\theta})^{n-1}\right]_0^w = (1-e^{w/\theta})^{n-1}$, for $0 < w < \infty$. Let $Z = \frac{W}{\theta}$. Then $F_Z(z) = P(Z \le z) = P(W \le z\theta) = (1-e^{-z})^{n-1},\ 0 < z < \infty$. Since Z is a random variable depending on the sample and θ whose distribution does not depend on θ. Hence Z is a pivotal quantity.

Let us fix $\alpha \in (0,1)$. For $p \in [0,1]$ let z_1 be such that $P(Z \le z_1) = p\alpha$ and z_2 be such that $P(Z \ge z_2) = (1-p)\alpha$, i.e. z_1 and z_2 are given by

$$\begin{aligned}
\left(1-e^{-z_1}\right)^{n-1} &= p\alpha, \\
\left(1-e^{-z_2}\right)^{n-1} &= 1-(1-p)\alpha.
\end{aligned}$$

From these

$$\begin{aligned}
z_1 &= -\ln\left(1-(p\alpha)^{1/(n-1)}\right), \\
z_2 &= -\ln\left(1-(1-(1-p)\,\alpha)^{1/(n-1)}\right).
\end{aligned}$$

Then the interval $[z_1, z_2]$, is such that $\int_{z_1}^{z_2} f_Z(z)\mathrm{d}z = 1-\alpha$ for $0 < \alpha < 1$. Then, given the range $W = w$, we have $z_1 \le \frac{w}{\theta} \le z_2$, and a $100(1-\alpha)\%$ CI for θ is $[w/z_2,\ w/z_1]$. One natural choice of z_1 and z_2 is to take $p = 1/2$ so that

$$\begin{aligned}
z_1 &= -\ln\left(1-(\alpha/2)^{1/(n-1)}\right), \\
z_2 &= -\ln\left(1-(1-\alpha/2)^{1/(n-1)}\right).
\end{aligned}$$

For the particular data set and the confidence level 95% we have $W = 27.3 - 5.7 = 21.6$ and

$$\begin{aligned}
z_1 &= -\ln\left(1-(0.025)^{1/9}\right) \approx 1.09, \\
z_2 &= -\ln\left(1-(0.975)^{1/9}\right) \approx 5.88,
\end{aligned}$$

resulting in the confidence interval $[21.6/5.88, 21.6/1.09] \approx [3.67, 19.82]$.

Problem 3. *Suppose that we have data $x_1, x_2, \ldots, x_n$ which are iid observations from a $N(\mu, 1)$ density where μ is unknown. Consider testing $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ using the test statistic $T = |\bar{x}|$.*

1. *Describe a testing procedure using a rejection region for the test statistic T.*
2. *Define the p-value for the proposed test.*
3. *Express the significance test procedure at level $\alpha = 0.05$ using the obtained p-value.*
4. *Find the power of this test as a function of μ.*
5. *Calculate the power of the test for $n = 25$ and $\mu = -1, -0.5, -0.1, +0.1, +0.5, +1$ and sketch the graph of the power function. In your calculations, you may find useful that Φ (the cdf of the standard normal distribution) at points 3.04, 0.54, -1.46, -1.96, -2.46, -4.46, -6.96 takes approximately the values 1.00 0.71 0.07 0.02 0.01 0.00 0.00, respectively.*
6. *Define the notion of a uniformly most powerful test and argue that the test discussed above is not uniformly most powerful among all possible tests in the considered problem.*
7. *Write a relation from which by using tables or statistical software one could determine how large n would have to be in order for the power of the test to be equal to 0.95 for $\mu = +1$?*

Solution. It is natural to reject H_0 if T is too big so we consider the rejection region $R_\alpha = [a, \infty)$ for some $a > 0$ to be determined from type 1 error given by

$$\begin{aligned}\mathbb{P}(T > a|\mu = 0) &= \mathbb{P}(\bar{X} \leq -a|\mu = 0) + \mathbb{P}(\bar{X} \geq a|\mu = 0)\\ &= \Phi(-a\sqrt{n}) + 1 - \Phi(a\sqrt{n}) = 2\Phi(-a\sqrt{n})\end{aligned}$$

and which is supposed to be at most α. Equalling it to α gives $a = -z_{\alpha/2}/\sqrt{n}$, where z_p stand for p-quantile of the standard normal distribution.

For the observed $\bar{x}$, the p-value is determined as $\hat{\alpha}$ such that $a = -z_{\hat{\alpha}/2}/\sqrt{n}$ being equal to $|\bar{x}|$. This is equivalent to $\hat{\alpha} = 2\Phi(-|\bar{x}|\sqrt{n})$.

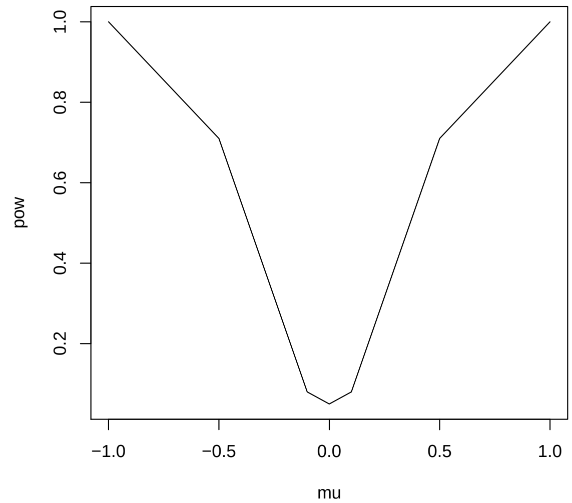
The testing procedure can be equivalently defined using the p-value to reject H_0 if the p-value is smaller than the significance level α.

The power function of the test is given by the probability of rejecting H_0, when H_1 is true, i.e. for $\mu \neq 0$ we have

$$\begin{aligned}P(\mu) &= \mathbb{P}(T > a|\mu)\\ &= \mathbb{P}(\bar{X} \leq -a|\mu) + \mathbb{P}(\bar{X} \geq a|\mu)\\ &= \Phi(z_{\alpha/2} - \mu\sqrt{n}) + 1 - \Phi(z_{1-\alpha/2} - \mu\sqrt{n}).\end{aligned}$$

For $\alpha = 0.5$, $z_{\alpha/2} = -z_{1-\alpha/2} \approx -1.96$.

For $n = 25$, we have $P(\mu) = \Phi(5(-1.96 - \mu)) + 1 - \Phi(5(1.96 - \mu))$, which for the specified values of μ takes values 1.00, 0.71, 0.08, 0.05, 0.08, 0.71, 1.00. The graph of the power function is given below.



A test for the problem of testing $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ at significance level α is called most powerful if its power is larger at each $\theta \in \Theta_1$ from the power of any other test for the same problem and at the same significance.

The Neyman-Pearson lemma guarantees that the test which is based on the likelihood ratio is most powerful for the testing problem $H_0 : \mu = 0$ vs. $H_1 : \mu = \mu_1$, where μ_1 is a fixed non-zero real value. Let us consider $\mu_1 > 0$, then the likelihood ratio that is the basis for the Neyman-Pearson test for the problem is given by the rejection region

$$\begin{aligned} R_\alpha &= \left\{ \mathbf{x}; \frac{l(\mu = 0)}{l(\mu_1)} \right\} \\ &= \left\{ \mathbf{x}; e^{-\mu_1 \sum x_i + \mu_1^2/2} < k \right\} \\ &= \left\{ \mathbf{x}; \bar{x} > \tilde{k}(\mu_1) \right\}, \end{aligned}$$

where $k(\mu_1)$ is determined from the equality

$$\mathbb{P}(\bar{X} \geq k(\mu_1) | \mu = 0) = \alpha,$$

from which it is clear that $k(\mu_1) = z_{1-\alpha}/\sqrt{n}$ which in fact does not depend on μ_1 given that the latter is positive. The power of this test for $\mu_1 > 0$ is given by

$$\begin{aligned} \tilde{P}(\mu_1) &= \mathbb{P}(\bar{X} \geq z_{1-\alpha}/\sqrt{n}) \\ &= 1 - \Phi(z_{1-\alpha} - \mu_1\sqrt{n}). \end{aligned}$$

For $\mu_1 > 0$ we have clearly (compare respective areas under the normal density curve)

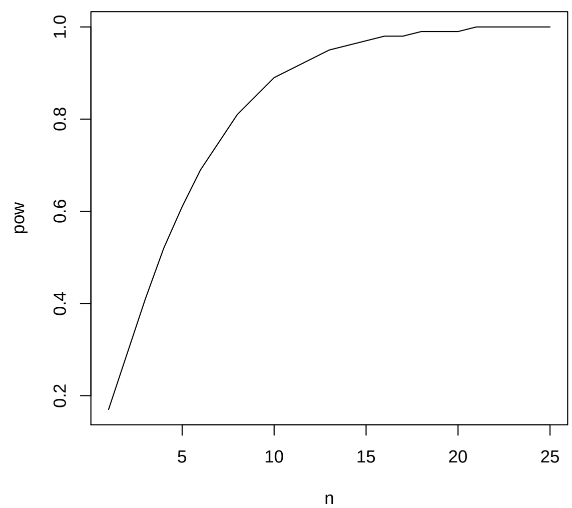
$$\tilde{P}(\mu_1) \geq P(\mu_1),$$

which means that the test is not uniformly most powerful.

In order to determine a sample size for which the power reaches the level 95% the equality

$$\Phi(-1.96 - \sqrt{n}) + 1 - \Phi(1.96 - \sqrt{n}) = 0.95$$

would have to be solved for n. The graph of the power as a function of n is presented below



Problem 4. *Suppose that household incomes in a certain country have a Pareto distribution with probability density function*

$$f(x) = \frac{\theta v^{\theta}}{x^{\theta+1}}, \qquad v \leq x < \infty \ ,$$

where $\theta > 0$ is unknown and $v > 0$ is known. Let $x_1, x_2, \ldots, x_n$ denote the incomes for a random sample of n such households. We wish to test the null hypothesis $\theta = 1$ against the alternative that $\theta \neq 1$.

1. *Derive an expression for $\hat{\theta}$, the MLE of θ.*
2. *Show that the generalised likelihood ratio test statistic, $\lambda(\mathbf{x})$, satisfies*

$$\ln\{\lambda(\mathbf{x})\} = n - n\ln(\hat{\theta}) - \frac{n}{\hat{\theta}}.$$

3. *Show that the test accepts the null hypothesis if*

$$k_1 < \sum_{i=1}^{n} \ln(x_i) < k_2 \ ,$$

and state how the values of k_1 and k_2 may be determined. Hint: Find the distribution of $\ln(X)$, where X has a Pareto distribution.

Solution. The likelihood function is

$$l(\theta) = \prod_{i=1}^{n} \left(\frac{\theta v^{\theta}}{x_i^{\theta+1}} \right) = \frac{\theta^n v^{n\theta}}{\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}}$$

for $v \leq x < \infty$ and $\theta > 0$. Therefore $\ln l(\theta) = n\ln\theta + n\theta\ln v - (\theta+1)\sum\ln(x_i)$. Differentiating we get the score function $S(\theta) = (n/\theta) + n\ln v - \sum\ln x_i$ and $I(\theta) = n/\theta^2 > 0$. The MLE $\hat{\theta}$ is found by $S(\hat{\theta}) = 0$, implying $(n/\hat{\theta}) = \sum\ln x_i - n\ln v = \sum\ln\left(\frac{x_i}{v}\right)$ so that $\hat{\theta} = n/\left[\sum_{i=1}^{n}(x_i/v)\right]$.

For the null hypothesis $\theta = 1$, the generalised likelihood ratio is $\lambda = L(1)/L(\hat{\theta})$ so that $\ln(\lambda(\mathbf{x})) = \ln l(1) - \ln l(\hat{\theta})$. Thus by direct algebra

$$\begin{aligned}
\ln(\lambda(\mathbf{x})) &= n\ln v - 2\sum_{i=1}^{n}\ln(x_i) - n\ln(\hat{\theta}) - n\hat{\theta}\ln v + (\hat{\theta}+1)\sum_{i=1}^{n}\ln(x_i) \\
&= n\ln v + (\hat{\theta}-1)\sum_{i=1}^{n}\ln(x_i) - n\ln(\hat{\theta}) - n\hat{\theta}\ln v \\
&= -\frac{n}{\hat{\theta}} + n - n\ln(\hat{\theta}) \\
&= n\left(1 - \ln\hat{\theta} - \frac{1}{\hat{\theta}}\right).
\end{aligned}$$

Let $u = 1/\hat{\theta}$. Then $\ln(\lambda(\mathbf{x})) = -n(u - 1 - \ln u)$ and $\frac{\mathrm{d}}{\mathrm{d}u}(\ln \lambda) = -n(1 - \frac{1}{u})$. Clearly $\ln \lambda$ has a maximum at $u = 1$. The null hypothesis $H_0: \ \theta = 1$ will be rejected if $\lambda(\mathbf{x}) \leq c$, for some c; i.e. if $u \leq k_1'$ or $u \geq k_2'$.

Reject H_0 if $\sum\limits_{i=1}^{n} \ln(x_i) \leq k_1$ or $\sum\limits_{i=1}^{n} \ln(x_i) \geq k_2$, where $k_1 = n\{k_1' = \ln v\}$ and $k_2 = n\{k_2' = \ln v\}$. For a significance α, choose k_1, k_2 to satisfy

$$\mathbb{P}\left\{\sum_{i=1}^{n} \ln(x_i) \leq k_1 \text{ or } \sum_{i=1}^{n} \ln(x_i) \geq k_2 | \theta = 1\right\} = \alpha.$$

The distribution of $\ln X$, where X has Pareto distribution is given by the density

$$h(y|\theta) = \theta v^{\theta} e^{-y\theta},\ y > \ln v,$$

in which we recognize the exponential distribution with parameter θ shifted by $\ln v$. Consequently, $\sum\limits_{i=1}^{n} \ln(X_i)$ has the gamma distribution with parameters n and θ shifted by $n \ln v$. Thus the null hypothesis is not rejected if

$$\gamma_{\alpha/2}(n, 1) < \sum_{i=1}^{n} \ln(x_i) - n \ln v < \gamma_{1-\alpha/2}(n, 1),$$

where $\gamma_p(n, 1)$ is the p-quantile of gamma distribution with parameters n and 1.