The University of Sydney
School of Mathematics and Statistics

Computer Exercise Week 11

STAT3023: Statistical Inference — Semester 2, 2021

Lecturers: Neville Weber and Michael Stewart

> Prepare your computer report and submit to the appropriate Canvas portal by **11:59pm Sunday 31st October**. Please include in your report all the code, plots and any comments required by the questions. The permitted format of the file that you upload will be restricted to PDF, HTML or Word, and should be created in a reproducible fashion (e.g. compiled from an Rmarkdown file). **Please do not have your name visible in your report.**

Interval estimation of an exponential rate parameter

Suppose we have a small sample of iid observations from an exponential distribution with unknown rate $\theta \in \Theta = (0, \infty)$. Consider the formal decision problem where the decision space is $\mathcal{D} = \Theta = (0, \infty)$ and the loss function is $L(d|\theta) = 1\{|d - \theta| > C\}$ for some fixed, known $C > 0$. In other words, we are interested in forming a fixed-width (of $2C$) interval estimate of θ (rather than a "confidence interval"; the main difference is that we are fixing the width, not the coverage probability of the procedure). The risk is the non-coverage probability of this interval; the estimator is the midpoint of the interval.

We are interested in the risk functions of two "decisions".

1. An interval based on the maximum likelihood estimator $\hat{\theta}_{\text{ML}} = 1/\bar{X}$: a "naïve" interval would simply be $\hat{\theta}_{\text{ML}} \pm C$, although this would not make complete sense if $\hat{\theta}_{\text{ML}} < C$ since then part of the interval would be below zero; we thus modify the interval by defining the centre as

$$d_1(\mathbf{X}) = \max\left(\hat{\theta}_{\text{ML}}, C\right) = \begin{cases} \hat{\theta}_{\text{ML}} & \text{if } \hat{\theta}_{\text{ML}} \geq C \\ C & \text{otherwise.} \end{cases}$$

   Then the interval estimate is

$$d_1(\mathbf{X}) \pm C = \begin{cases} \left(\hat{\theta}_{\text{ML}} - C, \hat{\theta}_{\text{ML}} + C\right) & \text{if } \hat{\theta}_{\text{ML}} \geq C \\ (0, 2C) & \text{otherwise.} \end{cases}$$

2. A Bayes procedure for this decision problem using the weight function $w(\theta) \equiv 1$ (the "flat prior"). Since the likelihood is

$$f_\theta(\mathbf{X}) = \prod_{i=1}^n \left[\theta e^{-\theta X_i}\right] = \theta^n e^{-T\theta}, \quad \text{where } T = \sum_{i=1}^n X_i,$$

   the product of the likelihood and prior/weight function is a multiple of a gamma density with shape $n+1$ and rate T; that is, the posterior density is

$$p(\theta|\mathbf{X}) = \frac{\theta^{(n+1)-1} e^{-T\theta} T^{n+1}}{\Gamma(n+1)}.$$

   The Bayes estimator $d_2(\mathbf{X})$ is the midpoint of the level set of this density of width $2C$; that is, $d_2 = d_2(\mathbf{X})$ satisfies

$$p(d_2 - C|\mathbf{X}) = p(d_2 + C|\mathbf{X}). \tag{1}$$

   Equivalently,

$$\begin{aligned} (d_2 - C)^n e^{-T(d_2 - C)} &= (d_2 + C)^n e^{-T(d_2 + C)} \\ e^{2TC} &= \left(\frac{d_2 + C}{d_2 - C}\right)^n \\ e^{\frac{2TC}{n}} &= \frac{d_2 + C}{d_2 - C} \\ (d_2 - C)e^{2\bar{X}C} &= d_2 + C \\ d_2 &= C\left(\frac{e^{2\bar{X}C} + 1}{e^{2\bar{X}C} - 1}\right). \end{aligned}$$

The endpoints of the interval are given by

$$
\begin{aligned}
d_2 - C &= \frac{2C}{e^{2\bar{X}C} - 1}\,, \quad \text{and} \\
d_2 + C &= \frac{2Ce^{2\bar{X}C}}{e^{2\bar{X}C} - 1}\,.
\end{aligned}
$$

For simplicity we shall take $n = 4$ and $C = 1.5$. We shall make use of the fact that $\bar{X}$ has a gamma distribution with shape n and rate $n\theta$. We shall firstly approximate the risk functions via simulation, then compute the exact risk functions.

1. (a) Write a function `mle.int()` that computes the mle-based interval. It should look like

```
mle.int=function(x,C) {
    ...
    d1=...
    c(d1-C,d1+C)
}
```

where, of course, the `...` parts are replaced with proper R code!

(b) Test it out by generating a test sample:

```
n=4
C=1.5
th0=2
x=rexp(n,rate=th0)
mle.int(x,C)
```

2. (a) Now write a function `bayes.int()` which computes the Bayes interval based on the flat prior. It should look like

```
bayes.int=function(x,C) {
    ...
    d2=...
    c(d2-C,d2+C)
}
```

(b) Try it out on your test sample to compare the two intervals.

(c) Visualise the interval by

- plotting the posterior density (using `curve()`);
- adding vertical lines (using `lty=2`) indicating the interval endpoints;
- adding a horizontal line (using `lty=3`) at the (common) height of the posterior density at the two interval endpoints;

(see week 10 2(c) for a similar plot). **Hint:** define `m=mean(x)`, use `n` and `m` to define the `shape` and `rate` of the posterior (gamma) density, and use `curve(...,from=0,to=3/m)`.

3. Define `th=(1:500)/50`, `L=length(th)` and `B=1000`. Also define

```
noncoverage.mle=noncoverage.bayes=0
```

Perform a double loop: at the `i`-th iteration of the outer loop

- define matrices `mle.mat=matrix(0,B,2)` and `bayes.mat=matrix(0,B,2)`;
- perform the inner loop: at the `j`-th iteration of the inner loop
  - generate a pseudo-random sample `x` of size `n=4` from an exponential distribution with rate `th[i]`;

  - obtain mle-based and Bayes intervals (with `C=1.5`), saving them in the `j`-th row of `mle.mat` and `bayes.mat` respectively;
- save in the `i`-th element of `noncoverage.mle` the number of times the mle-based interval did **not** cover `th[i]`; similarly for the Bayes interval in `bayes.mat`;

Convert the counts in the noncoverage vectors to proportions. Plot (as lines) these proportions against `th` (red for the mle-based, blue for Bayes). Add an informative heading and legend.

4. The precise form of the risk function for the mle-based interval is different for $0 < \theta < 2C$ and $\theta \geq 2C$, because for $0 < \theta < 2C$, the interval can only miss if the mle $\hat{\theta}_{\text{ML}}$ is too large, whereas for $\theta \geq 2C$, the interval can miss on both the high side and the low side.

Therefore, for $0 < \theta < 2C$,

$$R(\theta|d_1) = P_\theta\left\{\theta < \hat{\theta}_{\text{ML}} - C\right\} = P_\theta\left\{\bar{X} < \frac{1}{\theta + C}\right\}$$

and for $\theta \geq 2C$,

$$R(\theta|d_1) = P_\theta\left\{\theta < \hat{\theta}_{\text{ML}} - C\right\} + P_\theta\left\{\theta > \hat{\theta}_{\text{ML}} + C\right\} = P_\theta\left\{\bar{X} < \frac{1}{\theta + C}\right\} + P_\theta\left\{\bar{X} > \frac{1}{\theta - C}\right\}$$

The first probability here is of *overestimating*, and applies to all values of θ. The second is the probability of *underestimating* and only applies to $\theta \geq 2C$.

Since $\bar{X}$ has a gamma distribution with shape n and rate $n\theta$, we can compute this risk exactly. Compute this risk function for each value in `th`, save it in a vector called `m.risk` and plot it against `th`. Your code should look something like this:

```
risk.overest = ...                # this applies to all values in th
big = (th>=(2*C))                 # logical indicator of th values >= 2*C
risk.underest = 0*risk.overest    # start with a vector of zeroes
risk.underest[big] =  ...         # something involving th[big]
m.risk = risk.overest + risk.underest
```

5. The risk of the Bayes interval is given by

$$R(\theta|d_2) = P_\theta(\theta < d_2 - C) + P_\theta(\theta > d_2 + C)\,.$$

The first probability is

$$P_\theta\left(\theta < \frac{2C}{e^{2\bar{X}C} - 1}\right) = P_\theta\left(\bar{X} < \frac{1}{2C}\log\left(1 + \frac{2C}{\theta}\right)\right)\,.$$

Again, because $d_2 \geq C$ (the Bayes interval's lower endpoint never drops below zero!), the event in the second probability can only happen for $\theta \geq 2C$. For such θ it is

$$P_\theta\left(\theta > \frac{2Ce^{2\bar{X}C}}{e^{2\bar{X}C} - 1}\right) = P_\theta\left(\bar{X} > \frac{1}{2C}\log\left(\frac{\theta}{\theta - 2C}\right)\right)\,.$$

Compute this risk function for each value in `th` (save it in a vector `b.risk`) and plot it (in blue) against `th` (you will need to make use of the `big` indicator and compute `risk.overest` and `risk.underest` separately again as you did in the previous question). Add a line plot of the risk of the mle-based interval (in red). Add an informative heading and legend.

6. For which values of `th`

- does the mle-based interval do better;
- does the Bayes interval do better;
- do the two intervals have similar performance?

Can you describe and explain the behaviour of both curves near $\theta = 3$?