Mandatory Exercise 3

Børge Olav Haug

November 15, 2018

Problem 1

a) The pdf of the triangle distribution is given by:

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)}, & c \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

or

$$f(x) = \begin{cases} 0, & x < a \\ \frac{2(x-a)}{(b-a)(c-a)}, & a \leq x \leq c \\ \frac{2(b-x)}{(b-a)(b-c)}, & c \leq x \leq b \\ 0, & x > b \end{cases}$$

The cdf of a continuous random variable is given by:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

When we apply this to the pdf of the triangle distribution, we have to deal with the 4 different cases:

(1) $x \leq a$:

(i) $x < a$:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} 0dt = 0$$

(ii) $x = a$:

$$F(a) = \int_{-\infty}^{a} f(t)dt = \int_{-\infty}^{a} 0dt + \int_{a}^{a} \frac{2(t-a)}{(b-a)(c-a)}dt = 0$$

(2) $a < x \leq c$:

$$\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{a} 0dt + \int_{a}^{x} \frac{2(t-a)}{(b-a)(c-a)}dt \\
&= \frac{1}{(b-a)(c-a)} \int_{a}^{x} (2t-2a)dt = \frac{\left[t^2 - 2at\right]_a^x}{(b-a)(c-a)} \\
&= \frac{x^2 - 2ax - a^2 + 2a^2}{(b-a)(c-a)} = \frac{x^2 - 2ax + a^2}{(b-a)(c-a)} = \frac{(x-a)^2}{(b-a)(c-a)}
\end{aligned}$$

(3) $c < x < b$:

$$\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(t)dt = \\
&= \int_{-\infty}^{a} 0dt + \int_{a}^{c} \frac{2(t-a)}{(b-a)(c-a)}dt + \int_{c}^{x} \frac{2(b-t)}{(b-a)(b-c)}dt \\
&= \frac{(c-a)^2}{(b-a)(c-a)} + \frac{1}{(b-a)(b-c)} \int_{c}^{x} (2b-2t)dt \\
&= \frac{c-a}{b-a} + \frac{\left[2bt - t^2\right]_c^x}{(b-a)(b-c)} = \frac{c-a}{b-a} + \frac{2bx - x^2 - 2bc + c^2}{(b-a)(b-c)} \\
&= \frac{c-a}{b-a} + \frac{b^2 - b^2 + 2bx - x^2 - 2bc + c^2}{(b-a)(b-c)} \\
&= \frac{c-a}{b-a} + \frac{(b^2 - 2bc + c^2) - (b^2 - 2bx + x^2)}{(b-a)(b-c)} \\
&= \frac{c-a}{b-a} + \frac{(b-c)^2 - (b-x)^2}{(b-a)(b-c)} = \frac{c-a}{b-a} + \frac{b-c}{b-a} - \frac{(b-x)^2}{(b-a)(b-c)} \\
&= \frac{c-a+b-c}{b-a} - \frac{(b-x)^2}{(b-a)(b-c)} = \frac{b-a}{b-a} - \frac{(b-x)^2}{(b-a)(b-c)} \\
&= 1 - \frac{(b-x)^2}{(b-a)(b-c)}
\end{aligned}$$

(4) $x \geq b$:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} f(t)dt = 1$$

$$\Big(F(x) = \int_{-\infty}^{a} 0dt + \int_{a}^{c} \frac{2(t-a)}{(b-a)(c-a)}dt + \int_{c}^{b} \frac{2(b-t)}{(b-a)(b-c)}dt + \int_{b}^{x} 0dt$$

$$= \frac{c-a}{b-a} + \frac{\left[2bt - t^2\right]_c^b}{(b-a)(b-c)} = \frac{c-a}{b-a} + \frac{2b^2 - b^2 - 2bc + c^2}{(b-a)(b-c)}$$

$$= \frac{c-a}{b-a} + \frac{b^2 - 2bc + c^2}{(b-a)(b-c)} = \frac{c-a}{b-a} + \frac{(b-c)^2}{(b-a)(b-c)}$$

$$= \frac{c-a}{b-a} + \frac{b-c}{b-a} = \frac{c-a+b-c}{b-a} = \frac{b-a}{b-a} = 1\Big)$$

To summarize:

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{(x-a)^2}{(b-a)(c-a)}, & a < x \leq c \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)}, & c < x < b \\ 1, & x \geq b \end{cases}$$

To be able to use the inverse transform method to simulate data from the triangle distribution, we first have to derive an expression for $F^{-1}(u)$. To find $F^{-1}(u)$, let $u = F(x)$, and solve for x. Again we have to deal with 4 different cases. The quadratic formula will be useful when deriving $F^{-1}(u)$. The formula for solving $ax^2 + bx + c = 0$ is given by:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

(1) $u \leq F(a) = 0 \Rightarrow u = 0$:

$$F^{-1}(0) = a$$

(2) $F(a) < u \leq F(c) \Rightarrow 0 < u \leq \frac{c-a}{b-a}$:

$$
\begin{aligned}
& u = \frac{(x-a)^2}{(b-a)(c-a)} \\
\Rightarrow\ & u = \frac{x^2 - 2ax + a^2}{(b-a)(c-a)} \\
\Rightarrow\ & \frac{1}{(b-a)(c-a)} x^2 - \frac{2a}{(b-a)(c-a)} x + \left(\frac{a^2}{(b-a)(c-a)} - u \right) = 0 \\
& \underset{\text{quadratic formula}}{\Rightarrow} \\
\Rightarrow\ & x = \frac{\frac{2a}{(b-a)(c-a)} \pm \sqrt{\left(-\frac{2a}{(b-a)(c-a)}\right)^2 - 4\frac{1}{(b-a)(c-a)}\left(\frac{a^2}{(b-a)(c-a)} - u\right)}}{\frac{2}{(b-a)(c-a)}} \\
\Rightarrow\ & x = \frac{\frac{2a}{(b-a)(c-a)}}{\frac{2}{(b-a)(c-a)}} \pm \frac{\sqrt{\frac{4a^2}{(b-a)^2(c-a)^2} - \frac{4}{(b-a)(c-a)}\left(\frac{a^2}{(b-a)(c-a)} - u\right)}}{\frac{2}{(b-a)(c-a)}} \\
\Rightarrow\ & x = a \pm \frac{\sqrt{\frac{4}{(b-a)(c-a)}\left(\frac{a^2}{(b-a)(c-a)} - \frac{a^2}{(b-a)(c-a)} + u\right)}}{\frac{2}{(b-a)(c-a)}} \\
\Rightarrow\ & x = a \pm \frac{\sqrt{\frac{4u}{(b-a)(c-a)}}}{\frac{2}{(b-a)(c-a)}} = a \pm 2\sqrt{\frac{u}{(b-a)(c-a)}}\ \frac{(b-a)(c-a)}{2} \\
\Rightarrow\ & x = a \pm \sqrt{\frac{u}{(b-a)(c-a)}}\ (b-a)(c-a) \\
\Rightarrow\ & x = a \pm \sqrt{\frac{u}{(b-a)(c-a)}}\ \sqrt{(b-a)^2(c-a)^2} \\
\Rightarrow\ & x = a \pm \sqrt{\frac{u(b-a)^2(c-a)^2}{(b-a)(c-a)}} \\
\Rightarrow\ & x = a \pm \sqrt{u(b-a)(c-a)}
\end{aligned}
$$

The number produced, x, must satisfy $a < x \leq c$.
$x = a - \sqrt{u(b-a)(c-a)}$ is smaller than a, so the inverse we are after is:

$$F^{-1}(u) = a + \sqrt{u(b-a)(c-a)}$$

(3) $F(c) < u < F(b) \Rightarrow \frac{c-a}{b-a} < u < 1$:

$$
\begin{aligned}
& u = 1 - \frac{(b-x)^2}{(b-a)(b-c)} \\
\Rightarrow\ & u = 1 - \frac{b^2 - 2bx + x^2}{(b-a)(b-c)} \\
\Rightarrow\ & -\frac{1}{(b-a)(b-c)}x^2 + \frac{2b}{(b-a)(b-c)}x + \Big(1 - \frac{b^2}{(b-a)(b-c)} - u\Big) = 0 \\
& \underset{\text{quadratic formula}}{\Rightarrow} \\
\Rightarrow\ & x = \frac{-\frac{2b}{(b-a)(b-c)} \pm \sqrt{\big(\frac{2b}{(b-a)(b-c)}\big)^2 - 4\big(-\frac{1}{(b-a)(b-c)}\big)\big(1 - \frac{b^2}{(b-a)(b-c)} - u\big)}}{-\frac{2}{(b-a)(b-c)}} \\
\Rightarrow\ & x = \frac{-\frac{2b}{(b-a)(b-c)}}{-\frac{2}{(b-a)(b-c)}} \pm \frac{\sqrt{\frac{4b^2}{(b-a)^2(b-c)^2} + \frac{4}{(b-a)(b-c)}\big(1 - \frac{b^2}{(b-a)(b-c)} - u\big)}}{-\frac{2}{(b-a)(b-c)}} \\
\Rightarrow\ & x = b \pm \frac{\sqrt{\frac{4}{(b-a)(b-c)}\big(\frac{b^2}{(b-a)(b-c)} + 1 - \frac{b^2}{(b-a)(b-c)} - u\big)}}{-\frac{2}{(b-a)(b-c)}} \\
\Rightarrow\ & x = b \pm \frac{\sqrt{\frac{4}{(b-a)(b-c)}(1-u)}}{-\frac{2}{(b-a)(b-c)}} = b \mp \frac{\sqrt{\frac{4}{(b-a)(b-c)}(1-u)}}{\frac{2}{(b-a)(b-c)}} \\
\Rightarrow\ & x = b \mp \frac{\sqrt{\frac{4}{(b-a)(b-c)}(1-u)}}{\frac{2}{(b-a)(b-c)}} = b \mp 2\sqrt{\frac{(1-u)}{(b-a)(b-c)}}\ \frac{(b-a)(b-c)}{2} \\
\Rightarrow\ & x = b \mp \sqrt{\frac{(1-u)}{(b-a)(b-c)}}\ (b-a)(b-c) \\
\Rightarrow\ & x = b \mp \sqrt{\frac{(1-u)}{(b-a)(b-c)}}\ \sqrt{(b-a)^2(b-c)^2} \\
\Rightarrow\ & x = b \mp \sqrt{\frac{(1-u)(b-a)^2(b-c)^2}{(b-a)(b-c)}} \\
\Rightarrow\ & x = b \mp \sqrt{(1-u)(b-a)(b-c)}
\end{aligned}
$$

The number produced, x, must satisfy $c < x < b$.
$x = b + \sqrt{(1-u)(b-a)(b-c)}$ is larger than b, so the inverse we are after is:

$$F^{-1}(u) = b - \sqrt{(1-u)(b-a)(b-c)}$$

(4) $u \geq F(b) = 1 \Rightarrow u = 1$:

$$F^{-1}(1) = b$$

To summarize:

$$F^{-1}(u) = \begin{cases} a, & u = 0 \\ a + \sqrt{u(b-a)(c-a)}, & 0 < u \leq \frac{c-a}{b-a} \\ b - \sqrt{(1-u)(b-a)(b-c)}, & \frac{c-a}{b-a} < u < 1 \\ b, & u = 1 \end{cases}$$

Assume that a number generator for generating numbers uniformly distributed on 0 to 1 is available. Since we have derived $F(x)$ and $F^{-1}(u)$, we can simulate data from the triangle distribution by applying the following algorithm:

1. Generate $u_1, u_2, \ldots, u_n$ from the Uniform$[0, 1]$ distribution.
2. Deliver $x_1 = F^{-1}(u_1), x_2 = F^{-1}(u_2), \ldots, x_n = F^{-1}(u_n)$.

b) See R-code.

Problem 2

$$X_1 \sim \mathrm{N}(\mu_1 = 75, \sigma_1^2 = 625)$$
$$X_2 \sim \mathrm{N}(\mu_2 = 46, \sigma_2^2 = 100)$$
$$X_3 \sim \mathrm{N}(\mu_3 = 18, \sigma_3^2 = 25)$$

$$\rho_{12} = \mathrm{Corr}(X_1, X_2),\ \rho_{13} = \mathrm{Corr}(X_1, X_3),\ \rho_{23} = \mathrm{Corr}(X_2, X_3)$$

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

a)

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 75 \\ 46 \\ 18 \end{bmatrix}$$

$$\sigma_1 = \sqrt{\sigma_1^2} = \sqrt{625} = 25,\ \sigma_2 = \sqrt{\sigma_2^2} = \sqrt{100} = 10,\ \sigma_3 = \sqrt{\sigma_3^2} = \sqrt{25} = 5.$$

The general covariance matrix for this problem is:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$$

ρ_{ij} is defined as $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$, so $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$.
We will also make use of the fact that $\sigma_{ji} = \sigma_{ij}$, and that $\sigma_{ii} = \sigma_i^2$.

i) $\rho_{12} = -0.75$, $p_{13} = 0$ and $p_{23} = 0$.

$$\sigma_{11} = \sigma_1^2 = 625$$
$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2 = -0.75 \cdot 25 \cdot 10 = -187.5$$
$$\sigma_{13} = \rho_{13}\sigma_1\sigma_3 = 0 \cdot 25 \cdot 5 = 0$$

$$\sigma_{21} = \sigma_{12} = -187.5$$
$$\sigma_{22} = \sigma_2^2 = 100$$
$$\sigma_{23} = \rho_{23}\sigma_2\sigma_3 = 0 \cdot 10 \cdot 5 = 0$$

$$\sigma_{31} = \sigma_{13} = 0$$
$$\sigma_{32} = \sigma_{23} = 0$$
$$\sigma_{33} = \sigma_3^2 = 25$$

$$\Rightarrow \ \mathbf{\Sigma} = \begin{bmatrix} 625 & -187.5 & 0 \\ -187.5 & 100 & 0 \\ 0 & 0 & 25 \end{bmatrix}$$

ii) $\rho_{12} = 0.75$, $p_{13} = 0$ and $p_{23} = 0$.

$$\sigma_{11} = \sigma_1^2 = 625$$
$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2 = 0.75 \cdot 25 \cdot 10 = 187.5$$
$$\sigma_{13} = \rho_{13}\sigma_1\sigma_3 = 0 \cdot 25 \cdot 5 = 0$$

$$\sigma_{21} = \sigma_{12} = 187.5$$
$$\sigma_{22} = \sigma_2^2 = 100$$
$$\sigma_{23} = \rho_{23}\sigma_2\sigma_3 = 0 \cdot 10 \cdot 5 = 0$$

$$\sigma_{31} = \sigma_{13} = 0$$
$$\sigma_{32} = \sigma_{23} = 0$$
$$\sigma_{33} = \sigma_3^2 = 25$$

$$\Rightarrow \ \mathbf{\Sigma} = \begin{bmatrix} 625 & 187.5 & 0 \\ 187.5 & 100 & 0 \\ 0 & 0 & 25 \end{bmatrix}$$

iii) $\rho_{12} = -0.75$, $p_{13} = 0.4$ and $p_{23} = -0.5$.

$$\sigma_{11} = \sigma_1^2 = 625$$
$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2 = -0.75 \cdot 25 \cdot 10 = -187.5$$
$$\sigma_{13} = \rho_{13}\sigma_1\sigma_3 = 0.4 \cdot 25 \cdot 5 = 50$$

$$\sigma_{21} = \sigma_{12} = -187.5$$
$$\sigma_{22} = \sigma_2^2 = 100$$
$$\sigma_{23} = \rho_{23}\sigma_2\sigma_3 = -0.5 \cdot 10 \cdot 5 = -25$$

$$\sigma_{31} = \sigma_{13} = 50$$
$$\sigma_{32} = \sigma_{23} = -25$$
$$\sigma_{33} = \sigma_3^2 = 25$$

$$\Rightarrow \ \mathbf{\Sigma} = \begin{bmatrix} 625 & -187.5 & 50 \\ -187.5 & 100 & -25 \\ 50 & -25 & 25 \end{bmatrix}$$

A positive correlation between X_i and X_j, (i.e. $0 < \rho_{ij} \leq 1$), means that X_j increases when X_i increases. The closer to 1, the stronger this relationship is. A negative correlation between X_i and X_j, (i.e. $-1 \leq \rho_{ij} < 0$), means that X_j decreases when X_i increases. The closer to -1, the stronger this relationship is.

In *i)* there is a strong negative correlation between X_1 and X_2, which means it's now more likely that $X_2 < 50$ when $X_1 > 80$ than if there was no correlation between the two. Similarly, the strong positive correlation between X_1 and X_2 in *ii)* increases the likelihood of $X_2 > 50$ when $X_1 > 80$. In *iii)* there is a negative correlation between X_1 and X_2, and between X_2 and X_3, and a (somewhat weak) positive correlation between X_1 and X_3. The two negative correlations will "dominate" the single positive correlation, which will result in an even smaller likelihood of generating a top tier character. So, to conclude, scenario *ii)* is most likely to generate a top tier character, scenario *iii)* the least likely, and scenario *i)* lies in the middle.

The main difference between *iii)* and the two first scenarios is that in *iii)* there is a dependency between every variable pair, while in the two first there is a dependency only between X_1 and X_2.

b) See R-code.

Problem 3

$$\int_0^{24} \lambda(t)dt = \int_0^{24} \big(5 + 50sin(\pi \cdot t/24)^2 + 190e^{-(t-20)^2/3}\big)dt$$

a) In the general case, we look at the integral $\int_a^b g(x)dx$. In crude Monte Carlo integration we treat X as a random variable uniformly distributed on $[a, b]$, and $g(X)$ as a funtion of this random variable. We then "force" the uniform density function, i.e. $\frac{1}{b-a}$, into the integral we started with by multiplying by 1, in this case in the form of $\frac{b-a}{b-a}$, and end up with $(b-a)\int_a^b g(x)\frac{1}{b-a}dx$. This is then the constant $(b-a)$ multiplied by the expectation/mean of the function $g(X)$, i.e. $(b-a)E(g(X))$. Instead of dealing with this integral, we make a crude estimate by generating numbers from the U$[a, b]$ distribution, apply the function $g(x)$ to these numbers, approximate the mean by calculating the average, and multiply by $(b-a)$, giving

$$\hat{\theta}_{CMC} = \frac{b-a}{n}\sum_{i=1}^{n} g(X_i)$$

where n is the amount of data we want to generate.

For this particular problem, we plug in $a = 0$, $b = 24$ and $g(\cdot) = \lambda(\cdot)$.

To calculate the required number of simulations needed to be at least 95% certain that the esimate strays no more than 10 from the true answer, we interpret $\frac{b-a}{n}\sum_{i=1}^{n} g(X_i) = \frac{1}{n}\sum_{i=1}^{n}(b-a)g(X_i)$ as the average of n random variables $Y_i = (b-a)g(X_i)$, with the same expecation and variance. As n grows large the central limit theorem applies, and

$$\begin{aligned} Z =& \frac{\frac{b-a}{n}\sum_{i=1}^{n} g(X_i) - \mathrm{E}\big(\frac{b-a}{n}\sum_{i=1}^{n} g(X_i)\big)}{\sqrt{\mathrm{Var}\big(\frac{b-a}{n}\sum_{i=1}^{n} g(X_i)\big)}} \\ =& \frac{\hat{\theta}_{CMC} - \mathrm{E}(\hat{\theta}_{CMC})}{\sqrt{\mathrm{Var}(\hat{\theta}_{CMC})}} \approx \mathrm{N}(0,1) \end{aligned}$$

We can use this result to create an approximate $(1-\alpha)100\%$ confidence interval. There is approximately a $(1-\alpha)100\%$ chance that $\hat{\theta}_{CMC}$ falls in

$$\Big[\mathrm{E}(\hat{\theta}_{CMC}) \pm z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_{CMC})}\Big]$$

For a $(1-\alpha)100\%$ probability that $\hat{\theta}_{CMC}$ strays no more than e from the true answer $E(\hat{\theta}_{CMC})$ we need:

$$z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_{CMC})} < e$$

We have that:

$$\sqrt{\mathrm{Var}(\hat{\theta}_{CMC})} = (b-a)\cdot\frac{\sqrt{\mathrm{Var}\big(g(X)\big)}}{\sqrt{n}} \approx (b-a)\cdot\frac{\sqrt{\widehat{\mathrm{Var}}\big(g(X)\big)}}{\sqrt{n}}$$

where

$$\widehat{\mathrm{Var}}\big(g(X)\big) = \frac{1}{n-1}\sum_{i=1}^{n}\big(g(X_i) - \overline{g(X)}\big)^2$$

So we need:

$$\begin{aligned}
&z_{\alpha/2}\frac{b-a}{\sqrt{n}}\sqrt{\widehat{\mathrm{Var}}\big(g(X)\big)} < e\\
\Rightarrow&\sqrt{n} > z_{\alpha/2}\frac{b-a}{e}\sqrt{\widehat{\mathrm{Var}}\big(g(X)\big)}\\
\Rightarrow&n > z_{\alpha/2}^2\frac{(b-a)^2}{e^2}\widehat{\mathrm{Var}}\big(g(X)\big)
\end{aligned}$$

For this problem we plug in $\alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$, $a = 0$, $b = 24$, $e = 10$ and $g(\cdot) = \lambda(\cdot)$, giving

$$n > 1.96^2\cdot\frac{24^2}{10^2}\cdot\widehat{\mathrm{Var}}\big(\lambda(T)\big) \approx 22.13\cdot\widehat{\mathrm{Var}}\big(\lambda(T)\big)$$

$\widehat{\mathrm{Var}}\big(\lambda(T)\big)$ is approximated by generating data of a moderate size from the U$[0, 24]$ distribution, applying $\lambda(\cdot)$ to these generated numbers, and calculating the sample variance. Multiplying this sample variance by 22.13 gives us n, the number of simulations needed to be at least 95% certain that the esimate strays no more than 10 from the true answer.

b) See R-code.

c) The approach where antithetic variables are used to improve precision requires that the function of interest, $g(x)$, is monotonic on the interval $[a, b]$ that we are integrating over, i.e. for all $x_1, x_2 \in [a, b]$ where $x_2 > x_1$, $g(x_2) \geq g(x_1)$ if $g(x)$ is monotonically increasing, or $g(x_2) \leq g(x_1)$ if $g(x)$ is monotonically decreasing.

If U and V are random variables and $g(x)$ is monotonic on $[a, b]$, then $\mathrm{Cov}\big(g(U), g(V)\big) < 0$, when $\mathrm{Cov}(U, V) < 0$. [1, chap5_part2, p. 4]

In the figure below, $\lambda(t)$ is plotted on the interval $[0, 24]$. We can see that $\lambda(t)$ starts increasing twice, and also starts decreasing twice. $\lambda(t)$ is therefore non-monotonic on $[a, b]$, and antithetic variables will not necessarily improve the precision of the estimated integral.

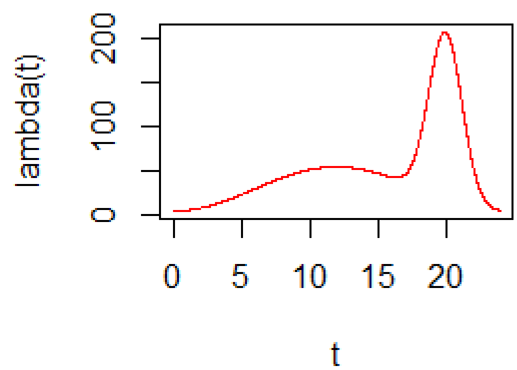

Figure 1: Plot of $\lambda(t)$ on the interval $[0, 24]$

We will use the triangle density as the importance function. It is closer to $\lambda(t)$ in shape than the uniform density, and should therefore lead to improved precision. Do as follows:

1. Generate $x_1, \ldots, x_n \sim \mathrm{Triangle}(a, b, c)$.
2. Calculate

$$\hat{\theta}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \lambda(x_i) / f(x_i; a, b, c)$$

   where $f(x_i; a, b, c)$ is the triangle density with parameters a, b and c.

$a = 0$, $b = 24$ and c can be calculated by generating a seqence of numbers between 0 and 24, $t_1, \ldots, t_m$, and selecting the t_i that maximizes $\lambda(t)$.

d) See R-code

Problem 4

a) The pdf of the exponential distribution is given by:

$$f(x;\lambda) = \lambda e^{-\lambda x}$$

The maximum-likelihood estimator for λ is the one that maximizes the probability of observing the data, i.e. the one that maximizes
$f(x_1,\ldots,x_n;\lambda) = \prod_{i=1}^n f(x_i;\lambda)$.

$$L(\lambda) = \prod_{i=1}^n f(x_i;\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

$$\begin{aligned} l(\lambda) &= \ln\big(L(\lambda)\big) = \ln(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}) = \ln(\lambda^n) + \ln(e^{-\lambda \sum_{i=1}^n x_i}) \\ &= n\ln(\lambda) - \lambda \sum_{i=1}^n x_i \end{aligned}$$

We want the maximum of $f(x_1,\ldots,x_n;\lambda)$, so we have to differentiate, set to zero and solve for λ.

$$\frac{\partial}{\partial\lambda} l(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

$$\begin{aligned} &\frac{\delta}{\delta\lambda} l(\lambda) = 0 \\ \Rightarrow\ &\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \\ \Rightarrow\ &\frac{n}{\lambda} = \sum_{i=1}^n x_i \\ \Rightarrow\ &\frac{\lambda}{n} = \frac{1}{\sum_{i=1}^n x_i} \\ \Rightarrow\ &\lambda = \frac{n}{\sum_{i=1}^n x_i} \end{aligned}$$

We should check that we have found a maximum by differentiating the log-likelihood function once more and checking if it is negative, at least at $\lambda = \frac{n}{\sum_{i=1}^{n} x_i}$:

$$\frac{\partial^2}{\partial\lambda^2} l(\lambda) = -\frac{n}{\lambda^2}$$

With $n > 0$, and $\lambda > 0$, $\frac{\delta^2}{\delta\lambda^2} l(\lambda) = -\frac{n}{\lambda^2}$ is always negative, and we have therefore found a maximum. We can then conclude that the maximum likelihood estimator for λ is:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Note that this can be rewritten as:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = 1/\overline{\mathrm{X}}$$

To use the bootstrap method to estimate the bias and standard deviation of $\hat{\lambda}$ and to calculate confidence intervals for λ, do as follows[1, chap7, pp. 9-15]:

1. Repeat steps 2-3 for $b = 1, 2, \ldots, B$.
2. Simulate data $x_1^{*(b)}, \ldots, x_n^{*(b)}$ by drawing n observations with replacement from the original data $x_1, \ldots, x_n$.
3. Calculate $\hat{\lambda}^{*(b)} = 1/\overline{X^{*(b)}}$ from $x_1^{*(b)}, \ldots, x_n^{*(b)}$, where $\overline{X^{*(b)}} = \frac{1}{n}\sum_{i=1}^{n} x_i^{*(b)}$.
4. Estimate the standard deviation of $\hat{\lambda}$ by the sample standard deviation of $\hat{\lambda}^{*(1)}, \ldots, \hat{\lambda}^{*(B)}$, i.e. calculate:

   $$\mathrm{SD}(\hat{\lambda}) \approx \widehat{\mathrm{SD}}(\hat{\lambda}^*) = \sqrt{\frac{1}{n-1}\sum_{b=1}^{B}(\hat{\lambda}^{*(b)} - \overline{\hat{\lambda}^*})^2}$$

   where $\overline{\hat{\lambda}^*} = \frac{1}{B}\sum_{b=1}^{B} \hat{\lambda}^{*(b)}$
5. Estimate the bias of λ as:

   $$\mathrm{bias}(\hat{\lambda}) \approx \widehat{\mathrm{bias}}(\hat{\lambda}) = \overline{\hat{\lambda}^*} - \hat{\lambda}$$

6. Calculate approximate confidence intervals for λ:

   (i) Standard normal bootstrap interval: $\left[\hat{\lambda} \pm z_{\alpha/2} \cdot \widehat{\mathrm{SD}}(\hat{\lambda}^*)\right]$

(ii) Basic bootstrap interval: $\left[2\hat{\lambda} - \hat{\lambda}^*_{1-\alpha/2}\ ,\ 2\hat{\lambda} + \hat{\lambda}^*_{\alpha/2}\right]$

(iii) Percentile bootstrap interval: $\left[\hat{\lambda}^*_{\alpha/2}\ ,\ \hat{\lambda}^*_{1-\alpha/2}\right]$

(iv) BCa interval: Use the boot library in R.

b) See R-code.

c) We want to test

$$H_0 : \lambda_X = \lambda_Y \text{ vs } H_1 : \lambda_X \neq \lambda_Y$$

based on time data $\mathbf{X}$ and $\mathbf{Y}$ obtained by running the experiment with two different temperatures, using

$$\hat{\lambda}_{\text{diff}}(\mathbf{X}, \mathbf{Y}) = \hat{\lambda}_X - \hat{\lambda}_Y$$

as our test statistic, where $\hat{\lambda}_X$ and $\hat{\lambda}_Y$ are the MLEs for λ, for the $\mathbf{X}$ and $\mathbf{Y}$ data, respectively.

Carry out the hypothesis test by doing as follows[1, chap8, p. 7]:

1. Calculate $\hat{\lambda}_{\text{diff}} = \hat{\lambda}_{\text{diff}}(\mathbf{X}, \mathbf{Y}) = \hat{\lambda}_X - \hat{\lambda}_Y$ for the original data.
2. Repeat steps 3-4 for $p = 1, \ldots, P$.
3. Simulate data $\mathbf{X}^{*(p)} = (x_1^{*(p)}, \ldots, x_{nX}^{*(p)})$, $\mathbf{Y}^{*(p)} = (y_1^{*(p)}, \ldots, y_{nY}^{*(p)})$, by drawing $n_X + n_Y$ observations without replacement from the original data $\mathbf{X}, \mathbf{Y}$.
4. Calculate $\hat{\lambda}^{*(p)}_{\text{diff}} = \hat{\lambda}^{*(p)}_{\text{diff}}(\mathbf{X}^{*(p)}, \mathbf{Y}^{*(p)}) = \hat{\lambda}_{X^{*(p)}} - \hat{\lambda}_{Y^{*(p)}}$ for the permuted data.
5. Calculate

$$p\text{-value} = \frac{1}{P} \sum_{p=1}^{P} I(|\hat{\lambda}^{*(p)}_{\text{diff}}| > |\hat{\lambda}_{\text{diff}}|)$$

where

$$I(|\hat{\lambda}^{*(p)}_{\text{diff}}| > |\hat{\lambda}_{\text{diff}}|) = \begin{cases} 1, & \text{if } |\hat{\lambda}^{*(p)}_{\text{diff}}| > |\hat{\lambda}_{\text{diff}}| \\ 0, & \text{if } |\hat{\lambda}^{*(p)}_{\text{diff}}| \leq |\hat{\lambda}_{\text{diff}}| \end{cases}$$

6. Reject the null hypothesis H_0 if p-value $< \alpha$.

d) See R-code.

e) We want to test

$$H_0 : \lambda \leq 0.0003 \text{ vs } H_1 : \lambda > 0.0003$$

based on time data $\mathbf{X}$.

We will use bootstrapping to generate many $\hat{\lambda}^{*(b)}$ values, and calculate the p-value as:

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} I(\hat{\lambda}^{*(b)} \leq 3)$$

Reject the null hypothesis H_0 if p-value $< \alpha$.

f) See R-code.

Problem 5

$$g(s) = \begin{cases} 0, & 0 \le s \le t_0 \\ \frac{\beta}{t_s}(s - t_0), & t_0 < s \le t_0 + t_s \\ \beta, & t_0 + t_s < s \le t_0 + t_s + t_p \\ \beta e^{-\gamma(s-(t_0+t_s+t_p))}, & s > t_0 + t_s + t_p \end{cases}$$

a) (i) $0 \le s \le t_0$:

$$G(s) = \int_0^s g(u)du = \int_0^s 0du = 0$$

(ii) $t_0 < s \le t_0 + t_s$:

$$\begin{aligned} G(s) &= \int_0^s g(u)du = \int_0^{t_0} 0du + \int_{t_0}^s \frac{\beta}{t_s}(u - t_0)du \\ &= \frac{\beta}{t_s}\big[\frac{1}{2}u^2 - t_0 u\big]_{t_0}^s = \frac{\beta}{t_s}(\frac{1}{2}s^2 - t_0 s - \frac{1}{2}t_0^2 + t_0^2) \\ &= \frac{\beta}{t_s}(\frac{1}{2}s^2 - t_0 s + \frac{1}{2}t_0^2) = \frac{\beta}{2t_s}(s^2 - 2t_0 s + t_0^2) \\ &= \frac{\beta}{2t_s}(s - t_0)^2 \end{aligned}$$

(iii) $t_0 + t_s < s \le t_0 + t_s + t_p$:

$$\begin{aligned} G(s) &= \int_0^s g(u)du = \int_0^{t_0} 0du + \int_{t_0}^{t_0+t_s} \frac{\beta}{t_s}(u - t_0)du + \int_{t_0+t_s}^s \beta du \\ &= \frac{\beta}{2t_s}(t_0 + t_s - t_0)^2 + \beta\big[u\big]_{t_0+t_s}^s = \frac{\beta}{2t_s}t_s^2 + \beta(s - (t_s + t_0)) \\ &= \frac{\beta}{2}t_s + \beta(s - (t_s + t_0)) \end{aligned}$$

(iv) $s > t_0 + t_s + t_p$:

$$\begin{aligned}
G(s) &= \int_0^s g(u)du = \\
&= \int_0^{t_0} 0du + \int_{t_0}^{t_0+t_s} \frac{\beta}{t_s}(u-t_0)du + \int_{t_0+t_s}^{t_0+t_s+t_p} \beta du \\
&+ \int_{t_0+t_s+t_p}^{s} \beta e^{-\gamma(u-(t_0+t_s+t_p))}du \\
&= \frac{\beta}{2}t_s + \beta(t_0+t_s+t_p-(t_s+t_0)) - \frac{\beta}{\gamma}\left[e^{-\gamma(u-(t_0+t_s+t_p))}\right]_{t_0+t_s+t_p}^{s} \\
&= \frac{\beta}{2}t_s + \beta t_p - \frac{\beta}{\gamma}(e^{-\gamma(s-(t_0+t_s+t_p))} - e^{-\gamma(t_0+t_s+t_p-(t_0+t_s+t_p))}) \\
&= \frac{\beta}{2}t_s + \beta t_p - \frac{\beta}{\gamma}(e^{-\gamma(s-(t_0+t_s+t_p))} - e^{0}) \\
&= \frac{\beta}{2}t_s + \beta t_p - \frac{\beta}{\gamma}(e^{-\gamma(s-(t_0+t_s+t_p))} - 1) \\
&= \frac{\beta}{2}t_s + \beta t_p + \frac{\beta}{\gamma}(1 - e^{-\gamma(s-(t_0+t_s+t_p))})
\end{aligned}$$

To summarize:

$$G(s) = \begin{cases} 0, & 0 \le s \le t_0 \\ \frac{\beta}{2t_s}(s-t_0)^2, & t_0 < s \le t_0 + t_s \\ \frac{\beta}{2}t_s + \beta(s-(t_s+t_0)), & t_0 + t_s < s \le t_0 + t_s + t_p \\ \frac{\beta}{2}t_s + \beta t_p + \frac{\beta}{\gamma}(1-e^{-\gamma(s-(t_0+t_s+t_p))}), & s > t_0 + t_s + t_p \end{cases}$$

b) See R-code.

c) To simulate the uncertainty in total production up to time s we will use the triangle distribution to express the uncertainty of the parameters. The algorithm described in Problem 1 a) and implemented in Problem 1 b) will be used to generate numbers from the triangle distribution.

1. Repeat steps 2-3 for $i = 1, \ldots, n$.

2. Generate

$$
\begin{aligned}
t_0^{(i)} &\sim \text{Triangle}(a = 0.85,\ b = 1.5,\ c = 1.1),\\
t_s^{(i)} &\sim \text{Triangle}(a = 0.7,\ b = 1.7,\ c = 1.0),\\
t_p^{(i)} &\sim \text{Triangle}(a = 4.0,\ b = 7.0,\ c = 5.0),\\
\beta^{(i)} &\sim \text{Triangle}(a = 7.5,\ b = 8.5,\ c = 8.0),\\
\gamma^{(i)} &\sim \text{Triangle}(a = 0.15,\ b = 0.3,\ c = 0.25)
\end{aligned}
$$

3. Calculate

$$G_s^{(i)} = G(s; t_0^{(i)}, t_s^{(i)}, t_p^{(i)}, \beta^{(i)}, \gamma^{(i)})$$

4. Deliver $G_s^{(1)}, \ldots, G_s^{(n)}$.

To calculate the probability $P\big(G(15) > 80\big)$, generate n values for $G(15)$, count how many times $G(s)$ exceeds 80 and divide by n, i.e. calculate

$$\frac{1}{n}\sum_{i=1}^{n} I\big(G_{15}^{(i)} > 80\big)$$

where

$$I\big(G_{15}^{(i)} > 80\big) = \begin{cases} 1, & G_{15}^{(i)} > 80 \\ 0, & G_{15}^{(i)} \leq 80 \end{cases}$$

$I\big(G_{15}^{(i)} > 80\big)$ is then a Bernoulli random variable that is equal to 1 with some probability p and 0 with probability $1 - p$. The expectation and variance of $I\big(G_{15}^{(i)} > 80\big)$ is $\mathrm{E}\big[I\big(G_{15}^{(i)} > 80\big)\big] = 1 \cdot p + 0 \cdot (1 - p) = p$ and $\mathrm{Var}\big[I\big(G_{15}^{(i)} > 80\big)\big] = \mathrm{E}\big[I\big(G_{15}^{(i)} > 80\big)^2\big] - \mathrm{E}\big[I\big(G_{15}^{(i)} > 80\big)\big]^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) - p^2 = p - p^2 = p(1 - p)$.

$\frac{1}{n}\sum_{i=1}^{n} I\big(G_{15}^{(i)} > 80\big)$ is then an average of n Bernoulli random variables. When

n grows large the central limit theorem applies, and

$$
\begin{aligned}
Z &= \frac{\frac{1}{n}\sum_{i=1}^{n} I\big(G_{15}^{(i)} > 80\big) - \mathrm{E}[\frac{1}{n}\sum_{i=1}^{n} I\big(G_{15}^{(i)} > 80\big)]}{\sqrt{\mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} I\big(G_{15}^{(i)} > 80\big)]}} \\
&= \frac{\overline{I\big(G_{15} > 80\big)} - \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[I\big(G_{15}^{(i)} > 80\big)]}{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[I\big(G_{15}^{(i)} > 80\big)]}} \\
&= \frac{\overline{I\big(G_{15} > 80\big)} - \frac{1}{n}np}{\sqrt{\frac{1}{n^2}np(1-p)}} = \frac{\overline{I\big(G_{15} > 80\big)} - p}{\sqrt{\frac{p(1-p)}{n}}} \\
&\approx \mathrm{N}(0,1)
\end{aligned}
$$

There is then a $(1-\alpha)100\%$ probability that $\overline{I\big(G_{15} > 80\big)}$ falls in

$$
\left[p \;\pm\; z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right]
$$

For there to be a $(1-\alpha)100\%$ probability that $\overline{I\big(G_{15} > 80\big)}$ strays no more than e from the true probability p, we need:

$$
\begin{aligned}
& z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < e \\
\Rightarrow\ & z_{\alpha/2}^2 \frac{p(1-p)}{n} < e^2 \\
\Rightarrow\ & n > z_{\alpha/2}^2 \frac{p(1-p)}{e^2}, \qquad \text{and with } p(1-p) \leq 0.25 \text{ for all } p \in [0,1] \\
\Rightarrow\ & n > z_{\alpha/2}^2 \frac{0.25}{e^2}
\end{aligned}
$$

Plugging in $e = 0.02, \alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$, gives

$$
n > 1.96^2 \frac{0.25}{0.02^2} = 2401
$$

To conclude, we can achieve the desired precision with 2401 simulations.

d) See R-code

e) To derive the time to the threshold production level, set $g(s) = 1$ and solve for s. We are interested in the solution s_c that satisfies $s_c > t_0 + t_p + t_s$, so we can skip the first three cases of $g(s)$, and focus on the last case $g(s) = \beta e^{-\gamma(s-(t_0+t_s+t_p))}$.

$$
\begin{aligned}
& g(s) = 1 \\
\Rightarrow\ & \beta e^{-\gamma(s-(t_0+t_s+t_p))} = 1 \\
\Rightarrow\ & \ln\left(\beta e^{-\gamma(s-(t_0+t_s+t_p))}\right) = \ln(1) \\
\Rightarrow\ & \ln(\beta) + \ln\left(e^{-\gamma(s-(t_0+t_s+t_p))}\right) = \ln(1) \\
\Rightarrow\ & \ln(\beta) - \gamma(s-(t_0+t_s+t_p)) = 0 \\
\Rightarrow\ & s - (t_0+t_s+t_p) = \frac{\ln(\beta)}{\gamma} \\
\Rightarrow\ & s = \frac{\ln(\beta)}{\gamma} + t_0 + t_s + t_p
\end{aligned}
$$

The uncertainty in the time to threshold production level and the uncertainty in the total volume produced until the threshold level is reached, is again expressed through using the triangle distribution to produce the parameters. The simulation algorithm is as follows.

1. Repeat steps 2-4 for $i = 1, \dots, n$.
2. Generate

$$
\begin{aligned}
t_0^{(i)} &\sim \text{Triangle}(a = 0.85,\ b = 1.5,\ c = 1.1), \\
t_s^{(i)} &\sim \text{Triangle}(a = 0.7,\ b = 1.7,\ c = 1.0), \\
t_p^{(i)} &\sim \text{Triangle}(a = 4.0,\ b = 7.0,\ c = 5.0), \\
\beta^{(i)} &\sim \text{Triangle}(a = 7.5,\ b = 8.5,\ c = 8.0), \\
\gamma^{(i)} &\sim \text{Triangle}(a = 0.15,\ b = 0.3,\ c = 0.25)
\end{aligned}
$$

3. Calculate

$$
s_c^{(i)} = \frac{\ln(\beta^{(i)})}{\gamma^{(i)}} + t_0^{(i)} + t_s^{(i)} + t_p^{(i)}
$$

4. Calculate

$$
G_{s_c}^{(i)} = G(s_c^{(i)}; t_0^{(i)}, t_s^{(i)}, t_p^{(i)}, \beta^{(i)}, \gamma^{(i)})
$$

5. Deliver $G_{s_c}^{(1)}, \dots, G_{s_c}^{(n)}$ and $s_c^{(1)}, \dots, s_c^{(n)}$.

To estimate the expected time to threshold production level, generate n values for s_c and approximate the expectation as the sample mean $\hat{\mu}_{s_c} = \frac{1}{n}\sum_{i=1}^{n} s_c^{(i)}$. This is the average of n random variables with the same expectation μ_{s_c} and variance σ_{s_c}. As n grows large the central limit theorem applies, and

$$Z = \frac{\hat{\mu} - \mathrm{E}(\hat{\mu}_{s_c})}{\sqrt{\mathrm{Var}(\hat{\mu}_{s_c})}} = \frac{\hat{\mu_{s_c}} - \mu_{s_c}}{\sigma_{s_c}/\sqrt{n}} \approx \mathrm{N}(0,1)$$

There is then approximately a $(1-\alpha)100\%$ probability that $\hat{\mu}_{s_c}$ falls in

$$\left[\mu_{s_c} \pm z_{\alpha/2}\frac{\sigma_{s_c}}{\sqrt{n}}\right]$$

. To have a $(1-\alpha)100\%$ probability that $\hat{\mu}$ strays no more than e (measured in years, 1 month equals $\frac{1}{12}$ years) from the true expected time to threshold production level, we need:

$$\begin{aligned} &z_{\alpha/2}\frac{\sigma_{s_c}}{\sqrt{n}} < e \\ \Rightarrow\ &\sqrt{n} > \frac{z_{\alpha/2}\sigma_{s_c}}{e} \\ \Rightarrow\ &n > \frac{z_{\alpha/2}^2\sigma_{s_c}^2}{e^2} \end{aligned}$$

Plugging in $e = \frac{1}{12}$, $\alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$, gives

$$n > \frac{1.96^2\sigma_{s_c}^2}{(\frac{1}{12})^2} \approx 553.19 \cdot \sigma_{s_c}^2$$

$\sigma_{s_c}^2$ is approximated by generating a moderate number of s_c values and calculating the sample variance. Multiplying this sample variance by 553.19 gives us n, the number of simulations needed to be at least 95% certain that the estimated expected time to threshold production level strays no more than 1 month from the true answer.

Estimating the expected total volume produced until the threshold level is done in a similar manner: generate n values for G_{s_c} and approximate the expectation as the sample mean $\hat{\mu}_{G_{s_c}} = \frac{1}{n}\sum_{i=1}^{n} G_{s_c}^{(i)}$.

This again gives the following formula for calculating n:

$$n > \frac{z_{\alpha/2}^2\sigma_{G_{s_c}}^2}{e^2}$$

And with $e = 0.25,\ \alpha = 0.05 \Rightarrow z_{\alpha/2} = z_{0.025} = 1.96$,

$$n > \frac{1.96^2 \sigma^2_{G_{s_c}}}{0.25^2} \approx 61.46 \cdot \sigma^2_{G_{s_c}}$$

$\sigma^2_{G_{s_c}}$ is approximated by generating a moderate number of G_{s_c} values and calculating the sample variance. Multiplying this sample variance by 61.46 gives us n, the number of simulations needed to be able to estimate the expected total volume produced until the threshold level with an error of at most 0.25 with 95% certainty.

f) See R-code.

List of Figures

1 Plot of $\lambda(t)$ on the interval $[0, 24]$ 13

References

[1] Jan Terje Kvaløy, *Notes, lecture slides, and code examples from the course STA510 Statistical modelling and simulation*, 2018

[2] Maria L. Rizzo, *Statistical Computing with R*, Chapman & Hall/CRC, London, 1st Edition, 2007