STAT2203: Probability Models and Data Analysis for Engineering
Quiz 7

1. In the Jelinski and Moranda model of software reliability, a program contains N bugs. Each of these bugs cause the program to fail independently. Once a bug is found, it is immediately fixed and no longer causes any issues for the program. The time T_i for bug i to cause the program to fail is assumed to have an exponential distribution with rate parameter λ.

   (a) Suppose a program has $N = 10$ bugs and that $\lambda = 1$. Simulate this process 10 000 times to get an estimate for the mean and variance of the time for all bugs to be found. It may be useful to look at the help in for `exprnd`, `max`, `mean` and `var`.

   Solution:

```
m=10000;
BugTimes = exprnd(1,m,10);
mean(max(BugTimes,[],2))

ans =

    2.9309

var(max(BugTimes,[],2))

ans =

    1.5595
```

   (b) Let $T_{(i)}$ denote the time the program has failed for the i-th time and define $T_{(0)} = 0$. The memoryless property of the exponential distribution implies that $T_{(i+1)} - T_{(i)}$ has an exponential distribution with rate parameter $\lambda(N-i)$ and that $T_{(1)}, T_{(2)} - T_{(1)}, \ldots, T_{(N)} - T_{(N-1)}$ are independent. Use these properties to evaluate the mean and variance of $T_{(N)}$, that is the time for all bugs to be found.

   Solution:

$$\begin{aligned}
\mathbb{E}T_{(N)} &= \mathbb{E}\left[\sum_{i=0}^{N-1}\left(T_{(i+1)} - T - (i)\right)\right] \\
&= \sum_{i=0}^{N-1}\mathbb{E}\left(T_{(i+1)} - T - (i)\right) \\
&= \sum_{i=0}^{N-1}\lambda^{-1}(N-i)^{-1}.
\end{aligned}$$

$$\begin{aligned}\mathrm{Var}T_{(N)} &= \mathrm{Var}\left[\sum_{i=0}^{N-1}\left(T_{(i+1)} - T - (i)\right)\right] \\ &= \sum_{i=0}^{N-1}\mathrm{Var}\left(T_{(i+1)} - T - (i)\right) \\ &= \sum_{i=0}^{N-1}\lambda^{-2}(N-i)^{-2}.\end{aligned}$$

With $N = 10$ and $\lambda = 1$ the mean and variance of $T_{(N)}$ is 2.9290 and 1.5498, respectively.

2. Let $X_1, \ldots, X_5$ be a simple random sample from a $\mathsf{N}(\mu, \sigma^2)$ distribution. The probability that the interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{5}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{5}}\right)$$

contains μ is 0.95. If we do not know σ^2, then we might replace it with the estimator S^2. Simulate 10 000 simple random samples from a $\mathsf{N}(0, 1)$ distribution. For each sample construct the interval

$$\left(\bar{x} - 1.96\frac{s}{\sqrt{5}},\ \bar{x} + 1.96\frac{s}{\sqrt{5}}\right)$$

and evaluate how often this interval contains 0 . What effect does replacing σ^2 with S^2 have on the coverage probability. It may be useful to look at the help in for `normrnd`, `mean` and `std`.

Solution:

```
X = normrnd(0,1,10000,4);
Upper = mean(X,2) + 1.96*std(X,0,2)/sqrt(5);
Lower = mean(X,2) - 1.96*std(X,0,2)/sqrt(5);
(sum(Lower > 0) + sum(Upper<0))/10000

ans =

    0.1793
```

The coverage probability is approximately 0.82 which is much smaller than the desired level of 0.95.

3. Let $X_1, \ldots, X_n$ be a simple random sample from a distribution with pdf

$$f(x) = \begin{cases} 2x, & x \in [0, 1] \\ 0, & \text{else} \end{cases}$$

A sample can be simulated using function `betarnd`.

```
x = betarnd(2,1,1000,1);
```

A histogram is a simple estimate of the probability density based on the data.

```
histogram(x,'BinLimits',[0 1],'Normalization','pdf')
```

MATLAB has a number of ways of determining the number of bins to use in constructing the histogram. In the following we look at how many 'bins' should be used for the a histogram.

Suppose we construct a histogram with m bins. The histogram counts how many observations fall in intervals $[i/m, (i+1)/m]$. In the following we will just consider the interval $[1/2, 1/2 + 1/m]$.

(a) Let Y be the number of X_i that fall in the interval $[1/2, 1/2 + 1/m]$. What is the distribution of Y?
Solution: The cdf of X is $F(x) = x^2$ for $x \in [0, 1]$. So the probability that X lies in the interval $[1/2, 1/2 + 1/m]$ is

$$(1/2 + 1/m)^2 - (1/2)^2 = 1/m + 1/m^2.$$

Therefore, $Y \sim \mathsf{Binomial}(n, 1/m + 1/m^2)$.

(b) We normalise the histogram so that it forms a probability density function. This is achieved by dividing the counts by the sample size and bin width so we have $Y/(n/m)$. What is the expected value and variance of mY/n from part (a)?
Solution:

$$\mathbb{E}[mY/n] = \frac{m}{n}\mathbb{E}[Y] = 1 + 1/m$$

$$\mathrm{Var}(mY/n) = \frac{m^2}{n^2}\mathrm{Var}(Y) = \frac{m^2}{n^2}n(1/m+1/m^2)(1-1/m-1/m^2) = \frac{m}{n}(1+1/m)(1-1/m-1/m^2)$$

(c) The mean squared error is a measure of the quality of an estimator. We want to use mY/n as an estimator of $f(1/2) = 1$. Compute the mean squared error $\mathbb{E}[(mY/n - 1)^2]$.
Solution:

$$\begin{aligned}\mathbb{E}[(mY/n - 1)^2] &= \mathbb{E}[(mY/n - \mathbb{E}(mY/n) + \mathbb{E}(mY/n) - 1)^2] \\ &= \mathrm{Var}(mY/n) + (\mathbb{E}(mY/n) - 1)^2 \\ &= \frac{m}{n}(1 + 1/m)(1 - 1/m - 1/m^2) + 1/m^2.\end{aligned}$$

(d) For what value of m (approximately) is the mean squared error minimised?
Solution: The mean squared error is approximately $m/n + 1/m^2$ which is minimised with $m \approx n^{1/3}$.

(e) Look at the 'BinMethod' argument in the help for the `histogram` function. How does the choice above compare to the 'Scott' and 'fd' methods?
Solution: Both these methods take the bin width to be $cn^{-1/3}$. In part (d) we showed bin width for this density should be $n^{-1/3}$.