Exercise 1

Use R to do the following:

(a) Create a vector named `x` with elements -1, 2, 3, 0, 2, -1, 1, 4.

(b) Extract the 4th element of the vector `x`.

(c) Extract as a single vector the 2nd, 3rd, 4th and 5th elements of `x`.

(d) Try the effect of `x[1 <= x & x <= 3]` and `x[x < 1 | x > 3]`.

The first command in (d) extracts those elements of `x` that are at least 1 and at most 3, and the second one those that are less than 1 or greater than 3.

(e) Create a vector named `y` with elements $\frac{1}{101}, \frac{2}{101}, \ldots, \frac{100}{101}$.

(f) Create a vector called `z` that has 100 elements (you can make these anything you like — try to find a simple command to do this).

(g) The command `sum(v)` returns the sum of the elements of the vector `v`. Find the sum of squares of the elements of `y` above. (Answer: 33.16832.) Type `sum(c(1,12)^3)-sum(c(9,10)^3)` — a simple answer, but why?

(h) Create vector `a` with elements (1,2,3,4,5) and vector `b` with elements (0,6,0,4,6). Try the commands `a>3`, `a >= 3`, `a>b`, `max(a,b)` and `pmax(a,b)`. What would be a neat way to sum those elements of `a` that are strictly greater than the corresponding elements of `b`?

(i) The natural exponential and logarithm functions are called `exp` and `log`. Use these to evaluate $e^{(\ln 2)^2}$. (Answer: 1.616807.)

(j) **Recycling.** When `x` is a vector and `h` is a number, `x+h` will add `h` to each component of `x`. If `h` is a vector but `x` is longer than `h`, R *recycles* the value of `h` as many times as necessary to make a vector of the same length as `x`. Try the results of the following: `c(1,1,1,1,1,1)+c(2,3)` and `c(1,1,1,1,1)+c(2,3)`.

Exercise 2

The R command `dbinom(x,n,prob)` computes $\mathrm{P}[X = x]$, where X is a Binomial RV with number of trials `n` and probability of success `prob`. The R

command `pbinom(q,n,prob)` computes $\mathrm{P}[X \leq q]$ for the same X. Thus, if X is Binom(n, p), with $n = 5$ and $p = 0.3$, then $\mathrm{P}[X = 3]$ can be computed using `dbinom(3,5,0.3)`, and $\mathrm{P}[X \leq 3]$ is computed using `pbinom(3,5,0.3)`.

A balanced die is rolled 100 times. Let X denote the number of times it shows a 6. Find two different ways of computing the probability that X is at least 25, one based on `pbinom`, and one based on `dbinom`. (Answer: 0.02170338.)

Information for Exercise 3

The R command `rbinom(1,1,prob)` returns the value `1` with probability `prob`, and `0` with probability `1-prob`. Use the command `rbinom(1,1,0.6)` to simulate the toss of a coin that lands H with probability 0.6.

The command `rbinom(nreps,1,prob)` simulates `nreps` *independent* trials, each with probability of success `prob`. So, tossing the same coin as above 10 times can be simulated with the command `rbinom(10,1,0.6)`. Try it.

Recall that the total number of successes in n independent trials, each with probability of success p is a Binom(n, p) RV. This total can be simulated directly, without generating results of the individual trials. The R command `rbinom(1,n,prob)` simulates a Binomial RV with `n` trials and probability of success `prob`. For example, the total number of Hs in 10 tosses of the coin above can be simulated with `rbinom(1,10,0.6)`. The result of this is an integer in the set $\{0, 1, 2, \ldots, 10\}$. Try it. Note that above we had `rbinom(1,1,prob)` as the special case of a single trial.

Finally, we may sometimes be interested in simulating `nreps` independent Binomial RVs, each with the same number of trials `n` and probability of success `prob`. This can be done using `rbinom(nreps,n,prob)`: the result is a vector with `nreps` integer elements, each between `0` and `n`.

Exercise 3

One hundred children rolled a die 10 times and counted the number of sixes.

(a) Simulate the outcome of this experiment.

(b) If `x` and `y` are vectors of length n, the R command `plot(x,y)` plots n points with the specified x and y coordinates. The command `plot(x,y,pch=20)` uses a prettier plotting character. Plot your results for the 100 children, with the x coordinates being $1, \ldots, 100$, and the y coordinates the number of sixes obtained by each child.

(c) Plot the probability mass function of a Binom$(10, 1/6)$ RV.

(d) Does the plot in (c) match with the result of the experiment?

Exercise 4

(a) Let X be the number of Heads in 100 tosses of a fair coin. Plot the probability mass function of X over the range $20 - 70$.

(b) Find the probabilities of the events $E_1 = \{X \leq 40\}$, $E_2 = \{X \geq 70\}$ and $E_3 = \{X \text{ is even}\}$ when X is the number of Heads in 100 tosses of a fair coin.

(c) Find the probabilities of the same events, $E_1 = \{X \leq 40\}$, $E_2 = \{X \geq 70\}$ and $E_3 = \{X \text{ is even}\}$, when X is the number of Heads in 100 tosses of a coin which lands Heads with probability $2/3$.

(d) In 100 tosses, Coin 1 lands Heads 40 times, and Coin 2 lands Heads 70 times. If it is known that each coin is either fair or has $P(\text{Heads}) = 2/3$, draw your conclusions about the nature of each coin.

Exercise 5

(a) In R, define a vector with elements $1, \ldots, 6$, use this to create a vector with elements 10, 10^2, 10^3, 10^4, 10^5 and 10^6, and hence compute

$$\left(1 - \frac{1}{n}\right)^n \tag{1}$$

for $n = 10, 10^2, \ldots, 10^6$.

(b) It is possible to represent the limit of (1) as $n \to \infty$ by a simple mathematical expression. Do this, guessing the answer if necessary. (Hint: It may help to examine the reciprocals of your answers in (a).)

(c) Each student attending a lecture course is asked to roll a die and choose a card at random from a standard pack of playing cards. If the student rolls a 6 and chooses the Ace of Spades, they win a chocolate egg. Suppose there are 312 students in the class, what is the probability that nobody wins a chocolate egg? Explain how your answer relates to part (b) of this question.

(d) Over a period of five years, the above experiment is conducted with a new cohort of 312 students each year.

(i) What is the probability that not a single chocolate egg is won over the five year period?

(ii) State the distribution of the total number of chocolate eggs awarded as prizes over the five years. Draw a sample of 20 realisations from this distribution and display this sample.

Exercise 6

(a) Use R to create a vector `x` containing 1000 independent realisations from the $N(3, 4^2)$ distribution.

(b) Plot a histogram of `x`. Use the option `prob=TRUE` in `hist` to show **probability density** on the y-axis: now, the scale implies that the area of each block in the histogram represents a probability. Plot the histogram again but this time give a title by including `main="Observed X values"` in the `hist` command, and use the `xlab` and `ylab` options to label the axes.

(c) Add the PDF of the $N(3, 4^2)$ distribution to the plot in red. The PDF should match the histogram. Give commands to save this plot as a .png file.

Exercise 7

A fair die is rolled 1000 times. Find the probability that 4 appears between 150 and 180 times inclusively.

(1) Use `pbinom` to obtain the binomial probability exactly. (The command `pbinom(x,n,p)` returns $\mathrm{P}(X \leq x)$ for $X \sim \mathrm{Binom}(n, p)$.)

(2) Approximate the distribution of a $\mathrm{Binom}(n, p)$ RV by that of a $N(np, np(1-p))$ RV and use the function `pnorm` to evaluate this approximation. Do this with and without the continuity correction.

Exercise 8

Each of 100 squirrels buries an acorn in a $10m$ by $10m$ square of ground. For each squirrel, the co-ordinates X and Y of the location of its acorn (measured in metres) are independent $\mathrm{Uniform}(0, 10)$ RVs.

(a) Simulate this process by generating vectors `Xcoord` and `Ycoord`, each of length 100, containing the x-co-ordinates and y-co-ordinates, respectively.

(b) Plot `Ycoord` against `Xcoord` to produce a map of the 100 acorn locations.

(c) Examine this plot for any patterns of clustering (a group of acorns which are close to each other) or linearity (a number of acorns that lie close to a straight line). If you find some such patterns, what do they indicate?

(d) Put the set of commands used in (c) in a `for` loop and run this to produce 4 independent plots of 100 acorns each. If you give the command `par(mfrow=c(2,2))` first, you can create a graphic containing all four plots (you can type `par(mfrow=c(1,1))` later to revert to having just one plot in the graphics window). Remember you can use the command `windows()` to start a new graphics window while keeping the old one.

Exercise 9

(a) Generate a vector `x` containing 500 independent replicates of a $N(25, 3^2)$ RV. Draw a histogram of `x` using the `prob=TRUE` option. I suggest you also set `breaks=15` in the `hist` command — see `help(hist)` for an explanation of what this does.

Superimpose the probability density function of the $N(25, 3^2)$ distribution on your histogram.

(b) Generate a vector `y` containing 500 independent realisations of a $N(5, 3^2)$ RV. Set `w=x-y` and draw a histogram of `w` using the `prob=TRUE` option. Use the `breaks` option again to get a suitable number of bars in the histogram.

(c) By inspecting the histogram of `w`, guess the distribution of the RV $W = X - Y$ when $X \sim N(25, 3^2)$ and $Y \sim N(5, 3^2)$ are independent.

Re-draw the histogram of `w` and, this time, superimpose the probability density function for your choice of distribution to check that your guess is reasonable. Remember, in the `curve` command you have to give a function of x — not w — to be plotted (see "Plotting commands" on page 1).

(d) Carry out further simulations to give a more definitive check that your chosen PDF really is correct.

Exercise 10

An experiment is defined as follows

Step 1. Generate independent random variables $X_1, \ldots, X_5$, each from an Exponential distribution with rate parameter 1. Set $M = X_1 + \ldots + X_5$.

Step 2. Generate Y from a Poisson distribution with mean M.

Step 3. Declare the outcome of the experiment to be the value of Y.

(a) Write R commands to run the experiment once.

Create a loop to run the experiment 500 times. Store the 500 values of M generated in Step 1 in a vector *Msample* and store the final outcomes, Y, in a vector *Ysample*. Draw histograms of the data in *Msample* and *Ysample*.

(b) We denote by NB$(5, 0.5)$ the Negative Binomial distribution with parameters $n = 5$ and $p = 0.5$. Typing `rnbinom(1,5,0.5)` produces one NB$(5, 0.5)$ random variable taking values in the set $\{0, 1, \ldots\}$. Use `rnbinom` to generate a vector *Nsample* containing 500 samples from the distribution NB$(5, 0.5)$. Draw a histogram of the data in *Nsample*.

(c) Professor Plum claims that the data generated in *Ysample* in part (a) should follow a NB$(5, 0.5)$ distribution. Do your histograms of *Ysample* and *Nsample* agree with this theory? You may find the option `xlim=(a,b)` in the `hist` command useful to set limits for the x-axis.

(d) Find the proportions of values of *Ysample* less than or equal to each of the integers $0, 1, \ldots, 20$. Do the same for *Nsample*. Plot one set of proportions against the other. What does this plot show?

(e) Repeat the comparison of *Ysample* and *Nsample* for larger data sets and state whether you believe that Professor Plum's claim is correct.

Exercise 11

As an angler sits by a river, the time (in minutes) between catches follows an Exp(0.1) distribution (i.e., with "rate" $\lambda = 0.1$).

Type `help(rexp)` and use the information given to write code that simulates

(a) The waiting time, $W1$, until the first fish is caught,

(b) The time, $W2$, between the first and second fish,

(c) The time, $W3$, between the second and third fish.

The angler starts at 10am. Let A be the event that she catches exactly 2 fish in the period 10am to 10.40am. Using the times $W1$, $W2$ and $W3$ you have generated, calculate the value of X, the indicator variable for the event A.

Make a loop to repeat these commands 1000 times and use the results to estimate $p = \mathrm{P}(A)$. We refer to your estimate as $\widehat{p}$. Use the formula

$$\sqrt{\mathrm{V}ar(\widehat{p})} \;=\; \sqrt{p(1-p)/n}$$

to find the standard deviation of your estimate. Since you do not know p precisely, substitute your estimate $\widehat{p}$ in the right hand side of the formula.

Is your estimate of $p = \mathrm{P}$(Catch 2 fish in 40 minutes) consistent with the value that would follow from the assumption that the number of fish caught in 40 minutes follows a Poisson distribution with mean 4?

Run a higher number of replications to check agreement with this Poisson model more thoroughly.

Exercise 12

A factory produces electronic components whose lifetimes (in hours) follow a Weibull distribution. Suppose the lifetime distribution is Weib(λ, β), as defined in lectures, with $\lambda = 0.01$ and $\beta = 2$.

Important: Type `help(rweibull)` for information on the R command that simulates Weibull data. The PDF and CDF stated there have a different parameterisation from that used in lectures. To get our Weib(λ, β) the R parameters a (shape) and b (scale) must satisfy $a = \beta$ and $b^a = 1/\lambda$, which means taking

$$a = \beta \quad \text{and} \quad b = \frac{1}{\lambda^{1/\beta}}.$$

Write R code to simulate the lifetimes of 15 components and compute the average observed lifetime (the sample mean of these 15 observations).

Write a loop to repeat these commands 1000 times. Draw a histogram of the set of 1000 sample means. If the mean of a sample of 15 components is used to estimate the expected component lifetime, how variable is this estimate?

Repeat your loop to generate another 1000 data sets of 15 components each and calculate the sample mean for each set. You should see similar results.

Exercise 13

Dr Barchan, the geography teacher has a globe of radius $1m$ in her office. Eric the fly is located at a random point on the globe's surface. Here is a neat method to generate such a random point, expressed in terms of coordinates (x, y, z), that represent distances, in m, from the centre of the globe:

> Step 1. Generate independent, Uniform$(-1, 1)$ random variables X, Y and Z. If the vector (X, Y, Z) lies inside the sphere of radius 1, accept it, otherwise, start again. Continue until you have accepted a vector (X, Y, Z).
>
> Step 2. Divide (X, Y, Z) by $\sqrt{\{X^2 + Y^2 + Z^2\}}$ to get a point $1m$ from the centre of the globe.

A 3D version of the argument in Question 14 of the theoretical exercises implies that Step 1 produces a random point in the sphere of radius 1. Then, by spherical symmetry, Step 2 yields a uniformly distributed point on the sphere's surface.

(a) Write a function `sample3d` that produces a sample of random vectors (X, Y, Z), each of which is a point from a uniform distribution on the globe's surface. Calling this function by the command `sample3d(n)` should produce an $n \times 3$ array, where each row is a vector (X, Y, Z).

(b) Generate coordinates (X, Y, Z) of 400 random points on the globe's surface and store these in a 400×3 array.

Plot the y-coordinates of these points against their x-coordinates.

Typing `curve(sqrt(a^2-x^2),from=-a,to=a,col="red",add=TRUE)` and `curve(-sqrt(a^2-x^2),from=-a,to=a,col="red",add=TRUE)` should add a circle of radius a to your plot.

Compare the density of points (X, Y) between circles of radius $0.9m$ and $1m$ with the density of points (X, Y) between circles of radius $0.4m$ and $0.5m$. Explain the differences that you see in these densities.

(c) Draw histograms of the values of X, Y and Z for the points found in (b).

Generating more data as appropriate, make a conjecture as to the form of these marginal distributions.

(d) Dr Barchan makes 600 independent recordings of Eric's coordinates (X, Y, Z), selects the cases where $X \in (0.45, 0.55)$, and draws a histogram

of the Y values for these cases. By construction, these values of Y follow the conditional distribution of Y given $X \in (0.45, 0.55)$. Use your function `sample3d` to mimic this process and draw the resulting histogram. How many samples of Y are displayed in this histogram?

We can argue that the conditional distribution of Y given $X \in (0.45, 0.55)$ approximates the conditional distribution of Y given $X = 0.5$ — and this approximation is improved if we make the interval of X values smaller.

Repeat the above simulations selecting cases where $X \in (0.5 - \delta, 0.5 + \delta)$, using a suitably chosen δ and a large enough sample size to give a reliable picture of the conditional distribution of Y given $X = 0.5$.

Dr Arbuthnot, the maths master, believes the conditional distribution of Y given $X = 0.5$ has probability density

$$f_{Y|X}(y|0.5) \;=\; \begin{cases} \frac{1}{\pi\sqrt{0.75-y^2}} & -\sqrt{0.75} < y < \sqrt{0.75} \\ 0 & \text{otherwise.} \end{cases}$$

Do your results confirm that this is the case?

(e) Eric the fly has a friend, Ernie. Assume that the two flies sit at independent locations, uniformly distributed on the globe's surface. Let D denote the Euclidean distance between Eric and Ernie (i.e., on a straight line through the interior of the globe).

Use your function `sample3d` to investigate the distribution of D.

Make a conjecture about the probability density function of D and give an estimate of its expected value, $E(D)$.

Exercise 14: The Inverse CDF method

Follow the steps (a) to (c) below to write an R function called `ex1` that simulates a random sample of size n from the distribution with PDF

$$f_X(x) \;=\; \begin{cases} \frac{x}{8} & 0 < x < 4 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Note that the CDF of X is

$$F_X(x) \;=\; \begin{cases} 0 & \text{if } x \le 0 \\ \frac{x^2}{16} & \text{if } 0 < x < 4 \\ 1 & \text{if } x \ge 4. \end{cases}$$

(b) Invert the relationship $u = F_X(x)$, that is, solve for x in the equation

$$u \;=\; x^2/16$$

for values $u \in (0, 1)$.

(c) Write a function `ex1` that applies the transformation found in (b) to a sample of n Uniform$(0, 1)$ RVs. Calling this function by the command `ex1(n)` should return a vector containing n independent realisations of X.

(d) Write a function `ex1_check` with argument n, to be called by the command `ex1_check(n)`, that

(i) calls `ex1` to obtain a random sample of size n,

(ii) plots a histogram of the sample,

(iii) and uses the `curve` command to overlay the PDF.

Include commands of the form `main="Main title"`, `xlab="X label"` and `ylab="Y label"` as arguments to `hist()` to give your histogram a suitable title and to label the axes as "X values" and "Probability density".

Run the function `ex1_check` with $n = 5,000$.

(e) Let X_1, X_2 and X_3 be independent RVs, each with the PDF f_X given above. Write a function `ex1_prob` to estimate $p = \mathrm{P}(X_1 + X_2 + X_3 < 8)$. Run this program once using 1,000 simulations, then repeat it using enough simulations to estimate the probability p to 2 decimal places.

Exercise 15: Rejection sampling

Suppose we wish to sample a RV from the distribution with PDF

$$f_X(x) \;=\; \begin{cases} \frac{2}{(\pi-2)} \left[\{2 - (x-1)^2\}^{1/2} - 1\right] & 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Draw a plot of $y = f_X(x)$ for $0 < x < 2$. In `plot()` set `xlim=c(0,2)` and `ylim=c(0,1)`. Denote the area under the curve $f_X(x)$ by A.

(A geometric argument involving the circle with centre $(1, -1)$ and radius $\sqrt{2}$ can be used to show that A has area 1, so $f_X(x)$ integrates to 1.)

(b) Write a function `ex2_gen_xy` that generates points (X, Y) uniformly distributed in $(0, 2) \times (0, 1)$ until one lies in the region A. *We have seen that this produces a point distributed uniformly in A.*

(c) Use your function `ex2_gen_xy` to generate a sample of 1,000 points (X, Y) and add these to your plot of the function $y = f_X(x)$.

(d) Let X be the x co-ordinate of a random point in A. Then, $\mathrm{P}(a < X < b)$ is the area under the curve $y = f_X(x)$ from $x = a$ to $x = b$. Thus,

$$\mathrm{P}(a < X < b) \;=\; \int_a^b f_X(x)\,dx$$

and we see that X is a random sample from the distribution with PDF $f_X(x)$.

Draw a histogram (with title and axis labels) of your sample of 1,000 X values. Superimpose the PDF $f_X(x)$. Does the sample agree with the PDF?

Exercise 16

(a) Follow steps (a), (b) and (c) of Exercise 14 to write an R function called `pr1` that simulates a random sample of size n from the distribution with PDF

$$f_X(x) \;=\; \begin{cases} 0 & x \le 10 \\ 3(x-10)^2/1000 & 10 < x < 20 \\ 0 & x \ge 20. \end{cases}$$

(b) Follow the approach taken in part (d) of Exercise 14 to write a function `pr1_check` and use this to check that a sample of 10,000 RVs produced by `pr1` matches the PDF $f_X(x)$.

(c) Follow the approach of part (e) of Exercise 14 to estimate $\theta = \mathrm{P}(X_2 > X_1 + 2)$ where X_1 and X_2 are independent RVs with PDF $f_X(x)$.

Choose the number of simulations so that the standard deviation of your estimate of θ is approximately 0.0005.

Exercise 17

(a) It is time for George and Fred's cars to be serviced. George arrives at the garage at noon. The time taken to service his car is S_1 minutes, where S_1 is distributed as 30 plus an exponential RV with mean 15.

Fred arrives at F_1 minutes after noon where $F_1 \sim \text{Uniform}(0, 60)$. If the mechanic has already finished servicing George's car, she starts servicing Fred's car as soon as he arrives. Otherwise, she starts when she has finished work on George's car. The time taken to service Fred's car is S_2 minutes, and S_2 follows the same distribution as S_1.

Let X and Y denote the lengths of time (in minutes) that George and Fred, respectively, have to wait at the garage. Thus, $X = S_1$ while Y is equal to S_2 plus any time Fred has to wait before the start of his car's service.

Write a function `gentimes()` that generates `nreps` realisations of the pair of times (X, Y) and produces an `nreps` $\times 2$ array containing these times.

(b) Use your function `gentimes()` to generate 250 pairs (X, Y).

Plot Y against X. What features of this plot suggest that X and Y are not independent?

(c) Denote the values seen in a sample of n randomly generated pairs of times by (x_i, y_i), $i = 1, \ldots, n$. If n is large, you may expect that

$$\frac{1}{n}\sum_{i=1}^{n} x_i \approx \text{E}(X), \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 \approx \text{E}(X^2),$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i \approx \text{E}(Y), \quad \frac{1}{n}\sum_{i=1}^{n} y_i^2 \approx \text{E}(Y^2)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i \approx \text{E}(XY).$$

Create a sample of 10,000 pairs (X, Y). Using the above facts, estimate $\text{E}(X)$, $\text{E}(X^2)$, $\text{E}(Y)$, $\text{E}(Y^2)$ and $\text{E}(XY)$ from your sample of data.

Hence, compute an estimate of the correlation between the times X and Y in a pair (X, Y). (You should write your own code for this calculation, without using any special R functions for finding a correlation.)

Exercise 18

In R, type

```
library(boot)
data(coal)
coaldates=coal[1:106,1]
```

The first command gives you access to data sets associated with a package called "boot". The second and third commands load a data item called `coal` and create a vector `coaldates`.

The vector `coaldates` contains the dates between 1851 and 1882 of the 106 explosions in coal mines in the UK which resulted in 10 or more fatalities. The integer part of each date gives the year and the day is represented as the fraction of the year that had elapsed on that date.

Compute the intervals, in days, between successive explosions by typing

```
x=365*(coaldates[2:106]-coaldates[1:105])
```

(a) Plot a histogram of the data in `x` using options `prob=TRUE, breaks=20`.

(b) Assuming these data follow an exponential distribution, find the method of moments estimate of the rate parameter.

(c) Overlay the PDF of this exponential distribution on the histogram.

(d) **Quantile-quantile plots** Suppose $X_1, \ldots, X_n$ are believed to follow the distribution with CDF F and we wish to check this is the case. Define the vector q by $q(i) = F^{-1}(i/(n+1))$, for $i = 1, \ldots, n$. Let y be the vector containing the observed values $x_1, \ldots, x_n$ arranged in increasing order — you can create this with the R command `y=sort(x)`. We expect

$$F(y_i) \approx \frac{i}{n+1} \quad \text{and, thus,} \quad y_i \approx F^{-1}\left(\frac{i}{n+1}\right).$$

A quantile-quantile (Q-Q) plot is a plot of the ordered data in y against the elements of q. If the distribution F is the correct model for the data, points in the Q-Q plot should lie close to a line of slope 1 through the origin.

Draw a Q-Q plot to check whether the exponential distribution is a good model for the data in `x`.

(e) Is the largest observation reasonable for the fitted distribution?

Exercise 19

Four statisticians are discussing how to estimate the mean of an exponential distribution. They have an odd number, n, of independent observations from an Exp(λ) distribution where λ is the rate parameter, so the expected value of each observation is $\theta = 1/\lambda$.

Denote the observations by $X_1, \ldots, X_n$ and write their ordered values as $X_{[i]}$ so, for example, $X_{[1]}$ is the smallest of the n observations and, in general, $X_{[1]} < X_{[2]} < \ldots < X_{[n]}$.

Alice suggests estimating θ by: $\widehat{\theta}_A = nX_{[1]}$,

Bob notes that the median of the Exp(λ) distribution is $\log(2)/\lambda$ and, hence, proposes estimating $\theta = 1/\lambda$ by the sample median divided by log(2), i.,e. $\widehat{\theta}_B = X_{[(n+1)/2]}/\log(2)$ (remember that n is odd),

Carol proposes the sample mean: $\widehat{\theta}_C = \overline{X} = (X_1 + \ldots + X_n)/n$,

Ted suggests a modification of Carol's estimator: $\widehat{\theta}_D = (n\,\overline{X})/(n+1)$.

The R code in the file *lab5-exercise2.R* provides a function `dgen` with arguments `theta`, `n` and `nreps`. This function generates *nreps* data sets, each containing n independent Exp(λ) random variables, where $\lambda = 1/\theta$. It then computes, for each data set, the four estimates defined above. The output from `dgen` is a $4 \times nreps$ array in which row 1 contains the values of estimator $\widehat{\theta}_A$ from the *nreps* data sets and, similarly, rows 2 to 4 contain the values of the estimators $\widehat{\theta}_B$, $\widehat{\theta}_C$ and $\widehat{\theta}_D$.

Additional commands in *lab5-exercise2.R* run `dgen` with one particular set of arguments and display histograms of the sets of estimates produced.

Use the output from `dgen` to estimate the bias, $\mathrm{E}(\widehat{\theta}) - \theta$, the variance, $\mathrm{V}ar(\widehat{\theta})$, and the mean square error, $\mathrm{E}\{(\widehat{\theta} - \theta)^2\} = \{\mathrm{E}(\widehat{\theta}) - \theta\}^2 + \mathrm{V}ar(\widehat{\theta})$, of each of the estimators $\widehat{\theta}_A$, $\widehat{\theta}_B$, $\widehat{\theta}_C$ and $\widehat{\theta}_D$ when $\theta = 10$ and $n = 21$.

Obtain further results for other values of θ and n and see whether your conclusions about the qualities of the four estimators remain the same.

If you were given the task of estimating an exponential mean from a sample of 21 observations, which estimator would you use?

Exercise 20

Charles has taken over a market stall where he sells fresh fish. Each morning he buys fish from the wholesaler at a cost of £4 per kg up to the first 100 kg and £2.5 per kg for anything above 100 kg. He sells the fish at £8 per kg. If any fish remains unsold at the end of the day, he sells it to the cat food factory for £1 per kg. Charles would like to know how much fish to buy each day in order to maximise his average daily profit.

Charles' predecessor, Elizabeth, bought 250 kg of fish per day. This was always enough to satisfy demand and no sales were lost. Elizabeth's sales figures for her last 150 days are in the R script *lab5-problem.R*. Open this file and run the command in it to create a vector `nfsold` of daily sales (in kg).

(a) Charles asks you to model the data in `nfsold` as a set of 150 independent observations from a Gamma(λ, k) distribution. Use the Method of Moments, as described in lectures, to obtain estimates of k and λ.

Draw a histogram of the data and superimpose the PDF of your fitted gamma distribution as a preliminary check that this distribution matches the observed data. Note that you can type `dgamma(x,k,rate=lambda)` to obtain the density of a Gamma(λ, k) distribution at x.

(b) Draw a quantile-quantile (Q-Q) plot to check whether the gamma distribution is a good model for these data. See part (d) of Exercise 18 for information about Q-Q plots and create your Q-Q plots in a similar manner: do ***not*** use R commands such as `qqplot` to do the work for you. You may use the command `qgamma(p,k,rate=lambda)` to find the value y such that $\mathrm{P}(Y \leq y) = p$ when $Y \sim \text{Gamma}(\lambda, k)$.

To judge the goodness of fit in this Q-Q plot, draw Q-Q plots for three sets of 150 observations generated from your fitted Gamma distribution. What do you conclude about the suitability of your fitted model for Elizabeth's data?

(c) Suppose Charles buys n kg of fish each day. Write functions

(i) `cost(n)` to give the cost of buying n kg of fish,

(ii) `income(n,d)` to give the income from sales (including selling surplus fish to the cat food factory) when there are potential sales of d kg and Charles has bought n kg,

(iii) `avg.profit(n,k,lambda,nreps)` to estimate the average daily profit when buying n kg of fish per day if potential sales follow a

Gamma(λ, k) distribution.

Apply these functions with selected parameter values to check they give the correct answers in particular cases.

(d) Use the functions written in (c) to find the optimal quantity of fish (an integer number of kg) that Charles should buy each day in order to maximise his average daily profit. *Hint:* You might find it helpful to use a relatively small number of simulations in an initial search, then run more simulations to get really accurate results as you close in on the optimal value,

(e) The analysis you have carried out in part (d) and your advice to Charles is based on certain assumptions. State one key assumption that you believe should be checked. How might you modify your analysis or fit a better model, in order to increase Charles' daily profit?