EM2B: R solutions

Robert Lung

2018/19

This document contains solutions to the exercises from the `R` Workbook for EM2B in 2018/19. For many exercises there are alternative solutions that give slightly different answers that can be just as good or even better. Any typos or suggestions regarding this document should be directed to s1778022@sms.ed.ac.uk.

Chapter 1

tbd

Chapter 2

Discrete Distributions

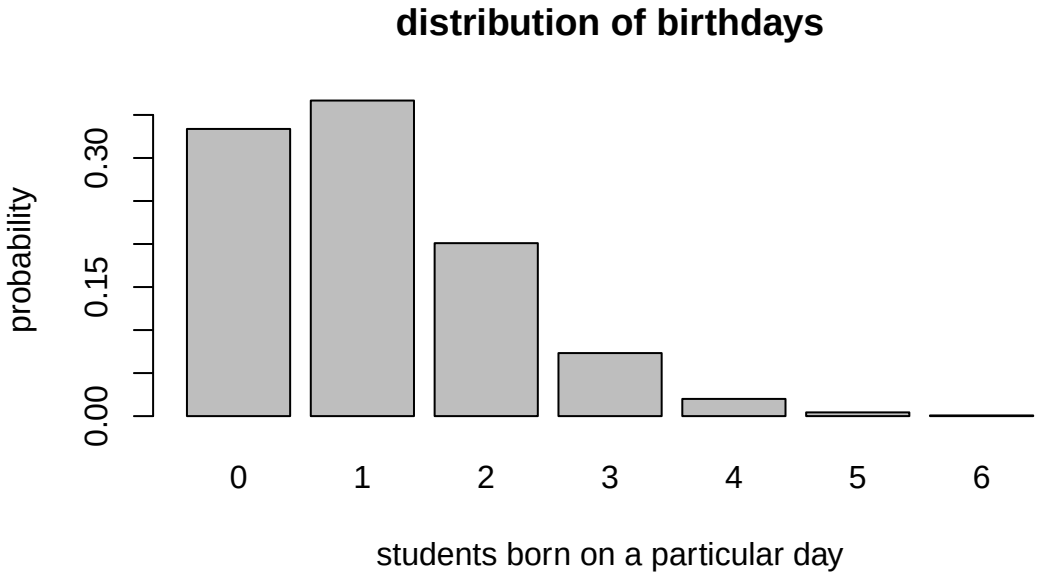
Exercise 1: Calculate the number of students in a class of 400 Engineers who share a birthday today.

Solution: Assuming that birthdays are independent and that being born on each day is equally likely, disregarding the 29th of February, we find that the sought number has a Binomial distribution with $n = 400$ and $p = \frac{1}{365}$, i.e. if X is the number of students sharing a birthday today then

$$X \sim \text{Binomial}\left(400, \frac{1}{365}\right)$$

Consequently $\mathbb{E}[X] = \frac{400}{365} \approx 1.1$ which is the number of people sharing a birthday at a particular day on average. If we are interested in the distribution of X we can look for example at a plot of its probability mass function (see below)

```
barplot(dbinom(0:6, size = 400, prob = 1/365),
        names.arg = 0:6,
        ylab = "probability",
        xlab = "students born on a particular day",
        main = "distribution of birthdays")
```



Exercise 2: In a workshop exercise a class of mechanical engineering students are asked to assemble a two stroke engine and test it. The lecturer has decided that since this is so important the students must perform the task successfully. Given that the students have, on average, a 75% chance of assembling the engine so it works:

(a) Plot the PDF for the number of attempts made by the students.
(b) Calculate the percentage of students who take more than 2 attempts.
(c) By simulating a class of 120 students calculate the mean number of attempts and produce a suitable graph to illustrate the number of attempts

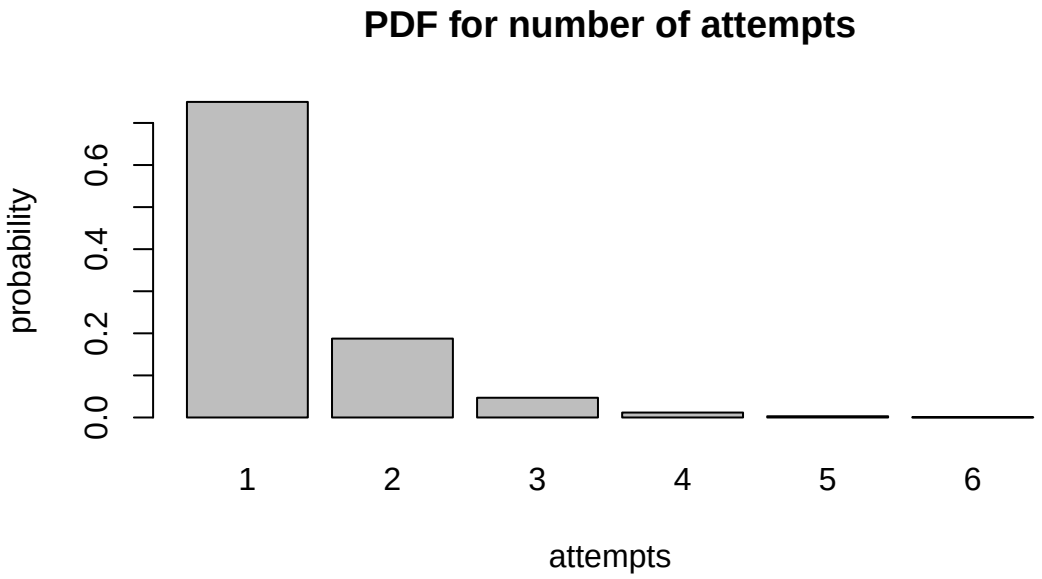
---

Solution (a): The number of students is irrelevant here since each student does the task by himself and whatever the number of students is we just need to multply the attemps of a single student by the total amount of students once we are obtain information about that quantity. If X is the number of attempts a student makes then $X-1$ is the number of times a student failed and we have

$$X - 1 \sim \mathsf{Geometric}(0.75)$$

This can be plotted in R with the commands

```
barplot(dgeom(0:5, prob = .75),
        names.arg = 1:6,
        ylab = "probability",
        xlab = "attempts",
        main = "PDF for number of attempts")
```



Solution (b): The percentage of students who take more than 2 attempts is equal to the probability that a student needs more than 2 attempts. In ordert to find that probability we need to calculate

$$\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - \mathbb{P}(X - 1 \leq 1)$$

where, as before, X is the number of attempts a student needs so that $X - 1 \sim \mathsf{Geometric}(0.75)$ and the last quantity can be easily found. We can calculate it by hand or use R and find

```
1-pgeom(1, prob = .75)
```

```
## [1] 0.0625
```

which means that 6.25% will need more than two attempts, i.e. fail more than once.

Solution (c): In order to simulate a class of 120 students we simulate 120 geometric random variables (the number of failed attempts) and add one (the successful attempt). This can be done by

```
set.seed(333) # for reproducibility
class.of.120 <- 1+rgeom(120, prob = .75)
# mean of the class
mean(class.of.120)
```

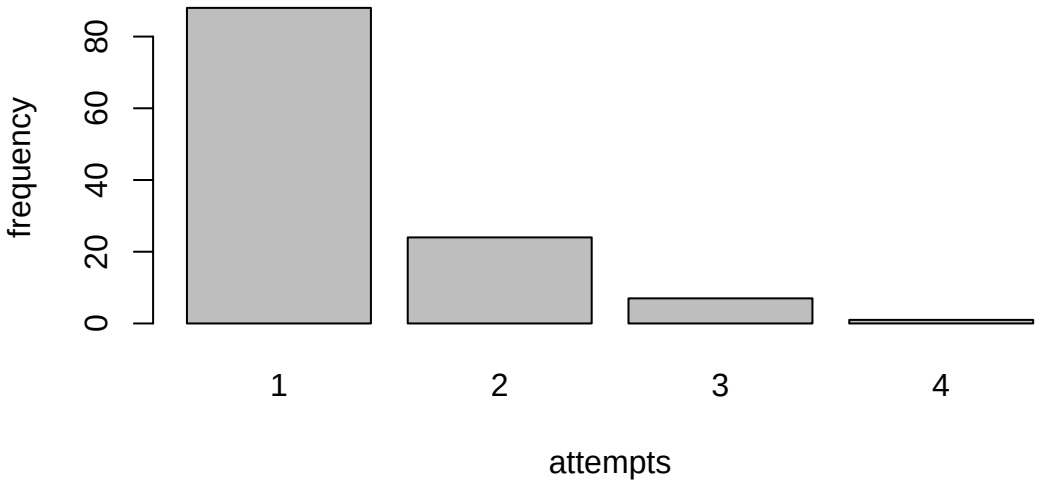
```
## [1] 1.341667
```

For visualising the attempts of the class a suitable graph is for example a histogram or barplot showing the amount of students that needed a certain number of of attempts.

```
table(class.of.120)
```

```
## class.of.120
##  1  2  3  4
## 88 24  7  1
```

```
barplot(table(class.of.120), ylab = "frequency", xlab = "attempts")
```



Exercise 3: A preservation railway use volunteer plate layers to install track. The engineering manager (Mrs C Miles) for the railway estimates that an initial inspection reveals on average 12.3 faults per mile of track.

(a) Mr JB Portly, the railway controller has said he will give free train tickets worth GBP 100 to any volunteer teams who lay a fault free mile. Given that this winter the railway will lay 10 miles of new track along the Sodar branch line, How much money should Mr Portly expect to loose?
(b) Plot the CDF of the number of faults per mile.
(c) The railway inspector form the Department of Transport will not require a full inspection if there are 15 or less faults on any given mile of track. What is the probability that ministry will not require a full inspection of the 10 miles of the new branch line?
(d) Mrs Miles has two engineers who inspect the track for her and work with a skilled team to repair any faults. Each team can fix 4 faults per day. How many days will be required to declare the 10 miles of track 95% fault free?

Solution 3 (a): We make the simplifying assumption that Mr Portly is referring to the 10 segments $0 \to 1, 1 \to 2, \ldots, 9 \to 10$. Denote by X_i the number of faults on segment $i-1 \to i$. In that case we can use a Poisson distribution with the correct mean to model the number of faults

$$X_i \overset{\text{iid}}{\sim} \text{Poisson}(12.3)$$

and if we write

$$Y_i = \begin{cases} 1 & \text{if } X_i = 0 \\ 0 & \text{if } X_i \neq 0 \end{cases}$$

If we define $p_0 := \mathbb{P}(X_i = 0)$then $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_0)$. The loss of Mr Portly (in GBP) can then be written as

$$100 \times \sum_{i=1}^{10} Y_i \quad \Longrightarrow \quad \text{Mr Portly's expected loss} = \mathbb{E}\left[100 \times \sum_{i=1}^{10} Y_i\right] = 1000 p_0$$

In order to compute this quantity we only have to find p_0 and calculated the expected loss. In R this can be done with the following commands

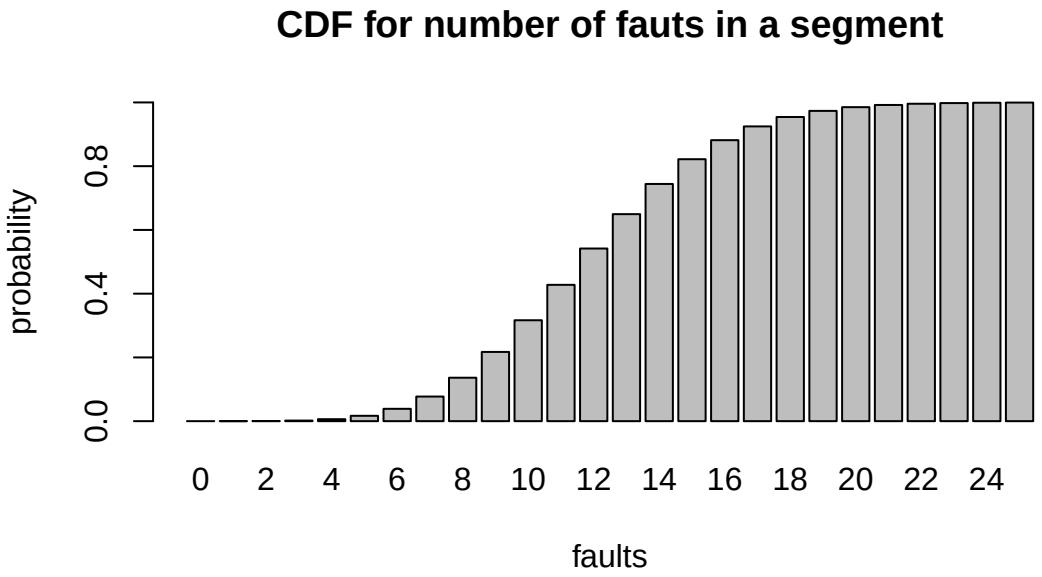
```
p.0 <- dpois(0, lambda = 12.3)
expected.loss <- 1000*p.0
print(expected.loss)
```

```
## [1] 0.004551744
```

which is essentially nothing or slightly less than half a Penny on average to be precise. We conclude that Mr Portly won't loose any money in most cases. This comes from the fact that under the Poisson assumption it is very unlikely that the volunteers will have a fault free segment.

SOlution 3 (b): Since we identified the distribution we just need to plot the cdf of a Poisson(12.3) random variable. Similarly as before we can do this with

```
barplot(ppois(0:25, lambda = 12.3),
        names.arg = 0:25,
        ylim = c(0,1),
        ylab = "probability",
        xlab = "faults",
        main = "CDF for number of fauts in a segment")
```



Solution 3 (c): In order to find this probability we should recall first that the number of faults in each segment is independent. Using the notation from before we have

$$\mathbb{P}\left(X_i \leq 15 : i = 1, \ldots, 10\right) = \prod_{i=1}^{10} \mathbb{P}\left(X_i \leq 15\right)$$

The last identity even simplifies further because all X_i have the same distribution, i.e. $X_i \sim \mathsf{Poisson}(12.3)$. We can use `R` to find

```
# probability of segment having no more than 15 faults
p.15 <- ppois(15, lambda = 12.3)
# since all segments have equal probability of having no more than 15 faults
# we are looking for the 10th power of this (each factor in the above product is equal)
print(p.15^10)
```

```
## [1] 0.1406638
```

So the probability that the ministry won't require a full inspection is roughly 14%.

Solution 3 (d): Since analytical calculation of the distribution of this quantity is difficult, we will use a simulation based approach to this question but this isn't the only way of answering this question. Using that the sum of independent Poisson random variables is again Poisson distributed we have for the total number of faults T on the whole track

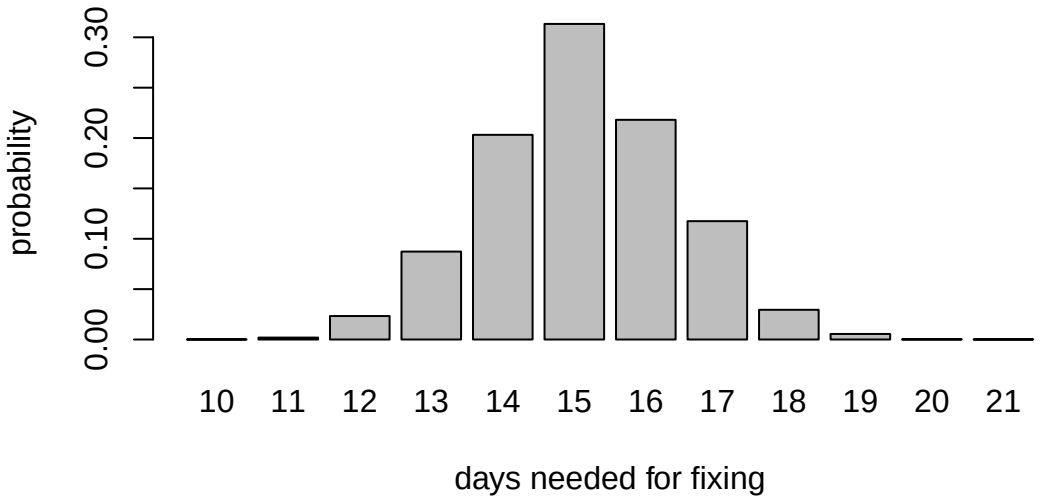
$$T := \sum_{i=1}^{10} X_i \sim \mathsf{Poisson}(123)$$

We can simulate easily obtain 10000 samples from T in `R`

```
set.seed(333)
# get samples
N <- 10000
T.samples <- rpois(N, lambda = 123)
# 95% fault free means T.95 faults are fixed where T.95 is given by
T.95 <- ceiling(0.95*T.samples)
# two engineers, each with a team, can fix 8 faults per day
days.to.fix <- ceiling(T.95/8)
table(days.to.fix)
```

```
## days.to.fix
##   10   11   12   13   14   15   16   17   18   19   20   21
##    1   19  233  873 2032 3133 2181 1175  295   55    2    1
```

```
barplot(table(days.to.fix)/N, xlab = "days needed for fixing", ylab = "probability")
```



Continuous Distributions

Exercise 1: Studies of a single-machine-tool system showed that the time the machine operates before breaking down is exponentially distributed with a mean time before failure of 10 hours.

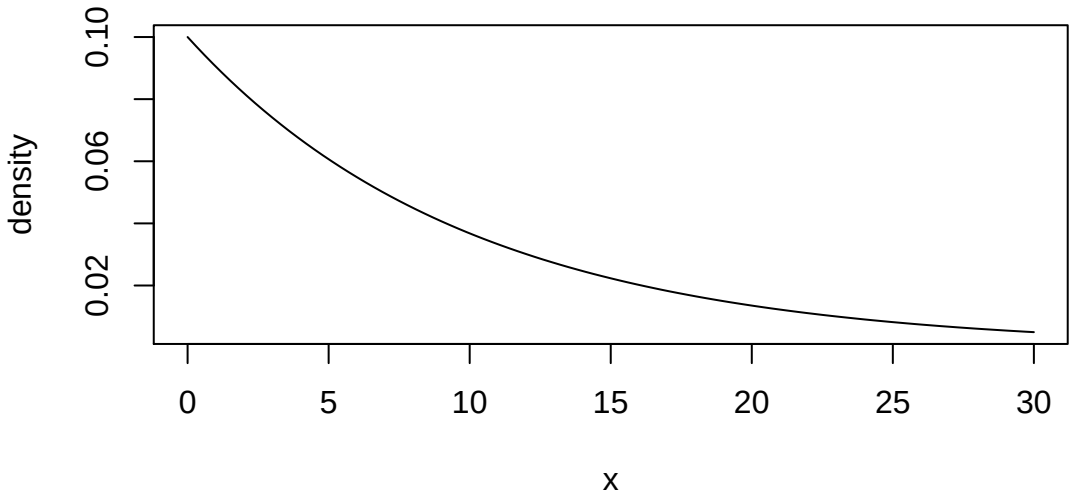
(a) Determine the failure rate and plot the PDF of reliability.
(b) Find the probability that the machine operates for at least 12 hours before breaking down.
(c) If the machine has already been operating 8 hours, what is the probability that it will last another 4 hours? [hint: remember that the exponential distribution has no memory!]

Solution (a): The failure rate is the reciprocal mean so we get

$$\text{failure rate} = \frac{1}{10h}$$

where h denotes hours. Admittedly, it is not absolutely clear what is meant by PDF of reliability but since we only have one distribution to work with we can pretty much narrow it down to what is has to be.

```
curve(dexp(x, rate = 1/10), 0, 30, ylab = "density", xlab = "x")
```



In the `R` workbook the y-axis is labelled with "probability" (see e.g. Figure 2.5, which displays rate 1 and not 1/8, and Figure 2.6 therein). This shouldn't be done. Unlike in the discrete case, the value of the density cannot be interpreted as a probability since probability densities are not bounded by 1 in general.

Solution (b): In the setting from before let τ be the time until breakdown (in hours). Then $\tau \sim \mathsf{Exponential}(\frac{1}{10})$ and the sought quantity is

$$\mathbb{P}(\tau > 12) = 1 - \mathbb{P}(\tau \leq 12)$$

```
# this can be found using the upper tail cdf
pexp(12, rate = 1/10, lower.tail = FALSE)
```

```
## [1] 0.3011942
```

Solution (b): We are given as a hint that τ has no memory which means $\mathbb{P}(\tau > 12 | \tau > 8) = \mathbb{P}(\tau > 4)$ and we get

```
pexp(4, rate = 1/10, lower.tail = FALSE)
```

```
## [1] 0.67032
```

Exercise 2: A group of students have found that over the winter, Temperature readings taken in the EngInn have a mean of 20 degrees Celcius and a standard deviation of 1.8 degrees Celsius. Their understanding of instrumentation leads them to believe these values will be normally distributed. What range of temperatures should the expect 95% of the time?

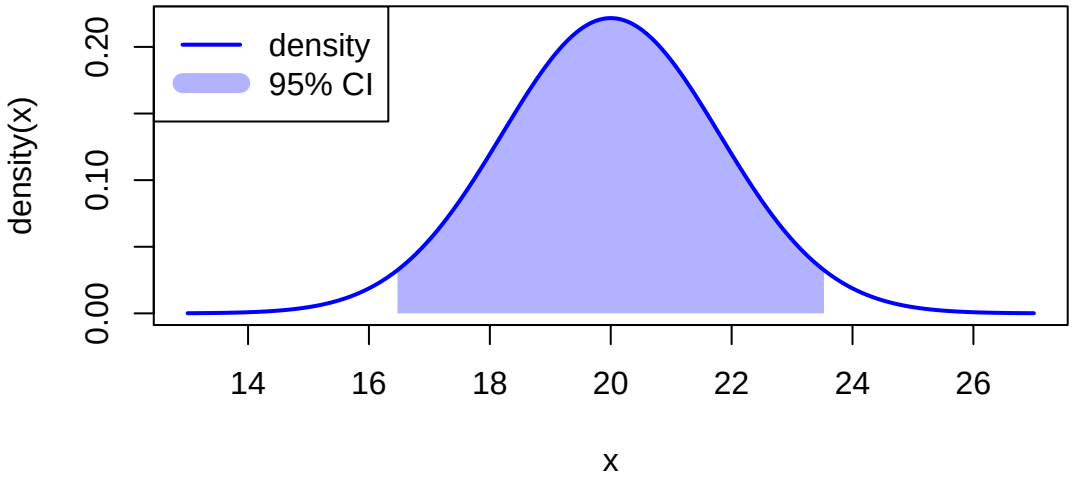
Solution: The failure rate is the reciprocal mean so we get

```
CI.95 <- qnorm(c(0.025, 0.975), mean = 20, sd = 1.8)
print(CI.95)
```

```
## [1] 16.47206 23.52794
```

Thus expect the temperature to be between 16.5 and 23.5 degrees Celsius 95% of the time. In addition to these numbers we can create a plot that visualisies this a little nicer and in a more informative way.

```
CIblue <- adjustcolor("blue", alpha.f = 0.3)
polygon.range <- seq(from = CI.95[1], to = CI.95[2], by = 0.001)
y <- seq(from = 13, to = 27, by = 0.01)
corners.polygon.X <- c(CI.95[1], polygon.range, CI.95[2])
corners.polygon.Y <- c(0, dnorm(polygon.range, mean = 20, sd = 1.8), 0)
plot(y, dnorm(y, mean = 20, sd = 1.8),
     type = "l", lwd = 2, col = "blue", xlab = "x", ylab = "density(x)")
polygon(corners.polygon.X,
        corners.polygon.Y,
        col = CIblue,
        lty = 1, lwd = 1, border = NA)
legend("topleft", legend = c("density", "95% CI"),
       col = c("blue", CIblue),
       lty = c(1, 1), lwd = c(2, 10))
```



Exercise 3: The same students have been told that the tostie machine must be shut off if the temperature exceeds 24 degrees. How many, term-time, days per year is this expected to occur?

Solution 3: We first want to find the probability

$$\mathbb{P}(X > 24) = 1 - \mathbb{P}(X \leq 24)$$

for $X \sim \mathsf{Normal}(20, 1.8^2)$.

```r
pnorm(24, mean = 20, sd = 1.8, lower.tail = FALSE)
```

```
## [1] 0.01313415
```

```r
60*pnorm(24, mean = 20, sd = 1.8, lower.tail = FALSE)
```

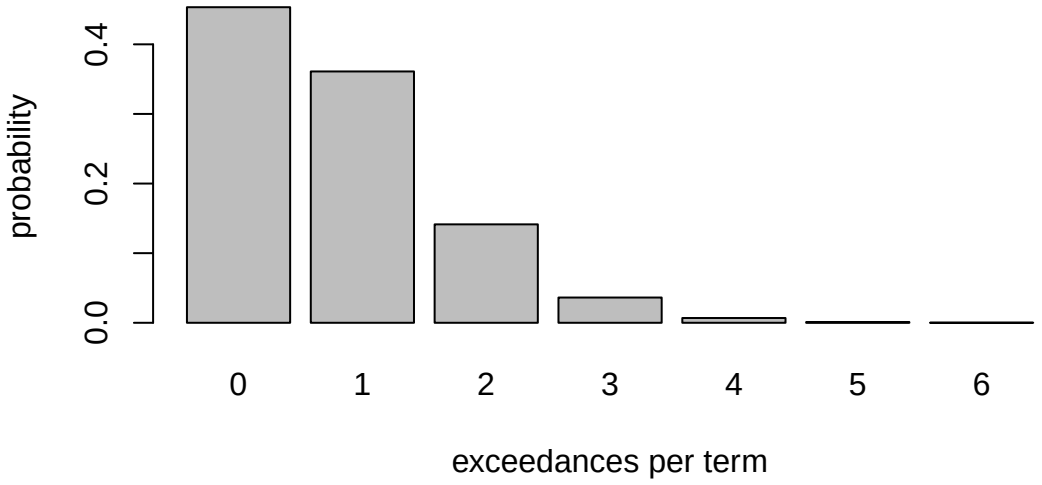
```
## [1] 0.7880487
```

Assuming that the term has 12 weeks and 5 working days each week we have that the expected number of days is $60 \times \mathbb{P}(X > 24) \approx 0.79$. More generally, assuming that the exceedances are independent on each day (which isn't necessarily realistic), the number of days D that this happens satisfies

$$D \sim \text{Binomial}(60, 0.0131)$$

We can visualise this as before

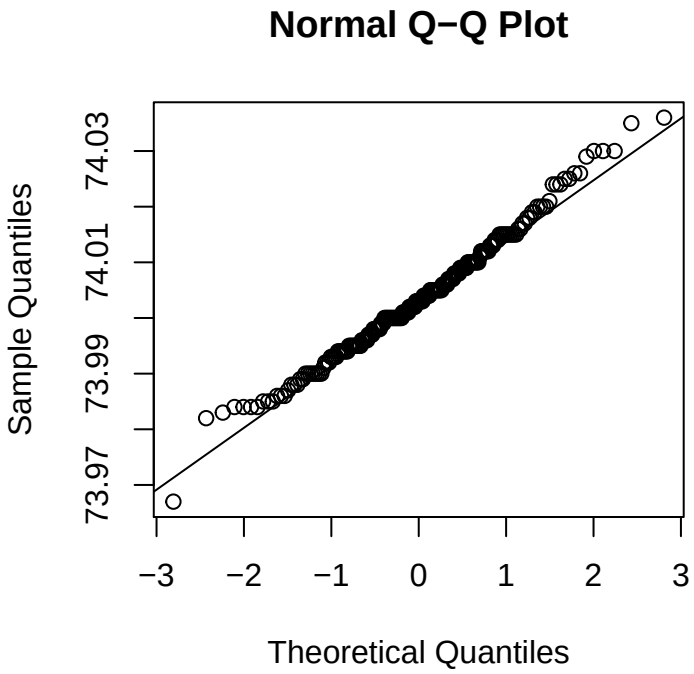
```r
barplot(dbinom(0:6, size = 60, prob = 0.0131),
        names.arg = 0:6,
        ylab = "probability",
        xlab = "exceedances per term")
```



Exercise 4: Is the piston ring diameter data normally distributed?

Solution 4: Let's examince the normal Q-Q plot for that

```
qqnorm(pistonrings$diameter)
qqline(pistonrings$diameter)
```



It is always a subjective matter to decide whether a normal Q-Q plot looks good, i.e. if the data is normally distributed. In this case there doesn't seem to be strong tendency towards something that isn't normal. In the next section you will be learning about hypothesis testing. A test for normality, i.e. a quantitative criterion that can be used for assessing normality is the Shapiro-Wilk test. In this case the we cannot reject normality based on this test.

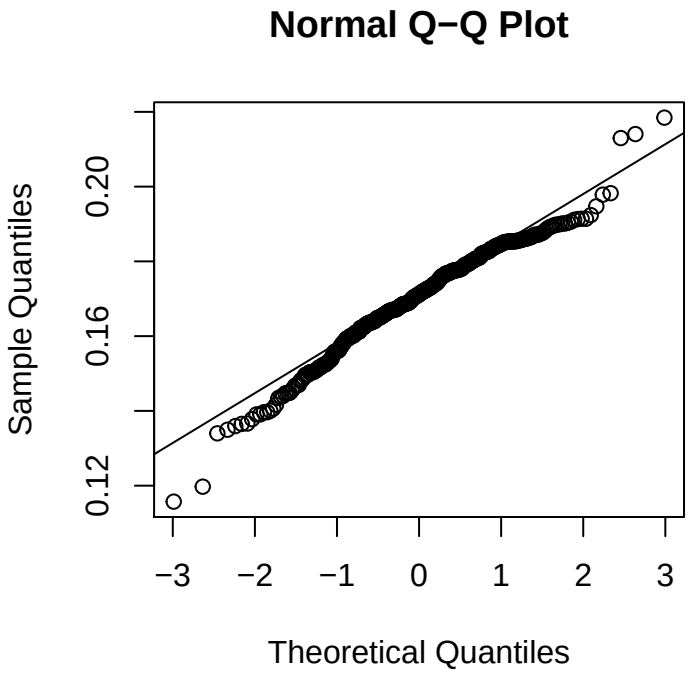
```
shapiro.test(pistonrings$diameter)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pistonrings$diameter
## W = 0.98968, p-value = 0.1607
```

Exercise 4: Are Prof Ingram's wave heights normally distributed?

Solution 4: Let's examince the normal Q-Q plot for that

```
qqnorm(waves$Height)
qqline(waves$Height)
```



Unlike in the previous excercise here the samples deviate from the line not only at the tails (where they will almost never fit. This provides some evidence that the waves height isn't normally distributed. If we consulted the Shapiro-Wilk statistic, as in the previous exercise, instead of a Q-Q plot we would reject normality.

```
shapiro.test(waves$Height)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  waves$Height
## W = 0.97609, p-value = 1.225e-05
```

Chapter 3