EM2B: R solutions

Robert Lung

2018/19

This document contains solutions to the exercises from the `R` Workbook for EM2B in 2018/19. They are not written by the composers of the exercises and while I hope this is not the case on too many occasions it is perfectly possible that some parts of the exercises have been misunderstood by me and are meant to be solved in a different way. For many exercises there are alternative solutions that give slightly different answers that can be just as good or even better. That means, if you did it differently it might not be a mistake but rather an alternative solution. Any typos, serious errors or suggestions regarding this document can be directed to s1778022@sms.ed.ac.uk.

This note has been created using R-Markdown (and the source file should be available) and compiled to a PDF with LaTeX. R-studio does this automatically but in order to produce a PDF document from a .Rmd file you need to have LaTeXinstalled on your system. This can be a good way for you to write you homework report. If you decide to do that we can help you in the tutorials with it. If you don't want to install LaTeXon you machine to produce the PDF you can always compile the document to a HTML file and at the end do the final compilation to PDF on a University machine (e.g. in the computing labs) which have LaTeXinstalled on them. Of course, you can decide to write the report in another way.

Chapter 1

I have not included solutions to the exercises of this chapter in this note. There are many great **free** online learning ressources available today (even though the free MOOC trend is taking a slight turn). In particular, for learning the basics of a programming language this can be good way to go if you want to learn some more `R`. Of course, you can always ask your question regarding `R` in the tutorials (that's why we have them). These notes are mainly meant to guide you through the statistical part assisted by the `R` lenaguage.

Chapter 2

Discrete Distributions

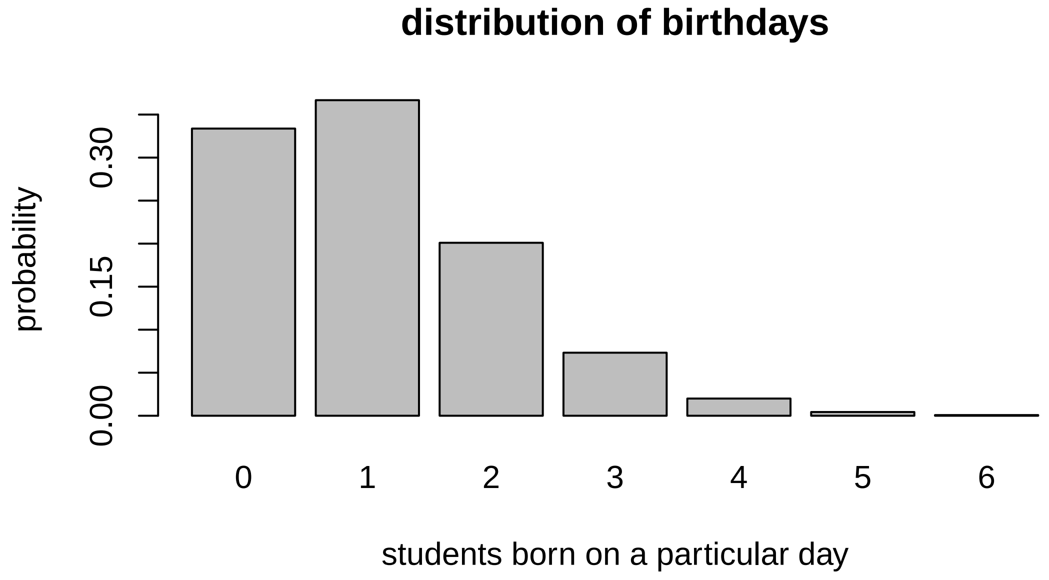
Exercise 1: Calculate the number of students in a class of 400 Engineers who share a birthday today.

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

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

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)

```r
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")
```



Note that this is not the probability that there are two people with the same birthday! For 400 individuals it is trivial that

$$\mathbb{P}(\text{two out of 400 individuals are born on the same day}) = 1$$

The above probability for less than 365 individuals is non-trivial and studied in what is called the Birthday Problem/Paradox (at a party with 30 people there will be two born on the same day about 70% of the time).

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

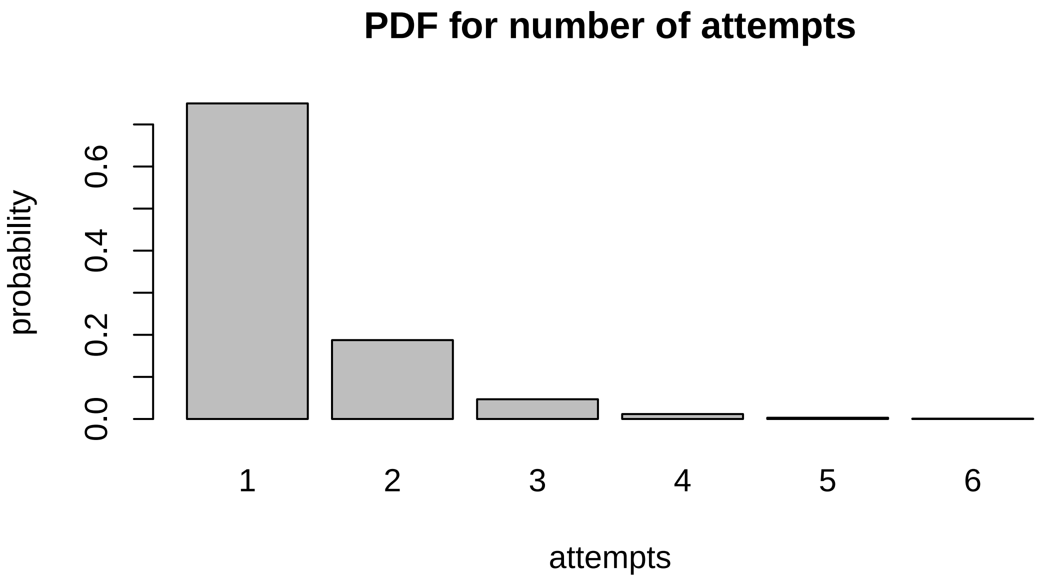
---

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)
```

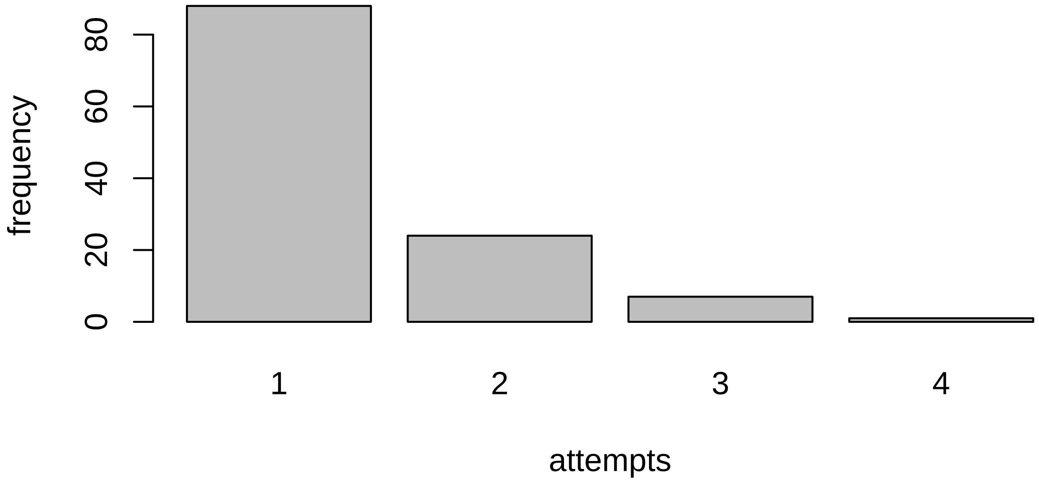
```
## [1] 1.341667
```

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)
```

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

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

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 $Y_i \overset{\text{iid}}{\sim} \mathsf{Bernoulli}(p_0)$. 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

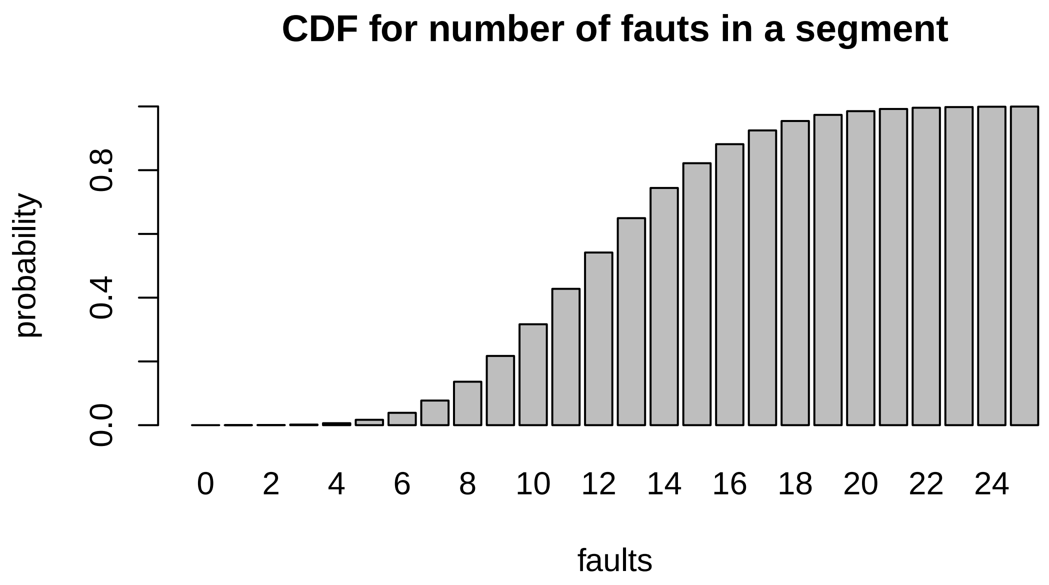
```r
p.0 <- dpois(0, lambda = 12.3)
expected.loss <- 1000*p.0
print(expected.loss)
```

```
## [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

```r
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

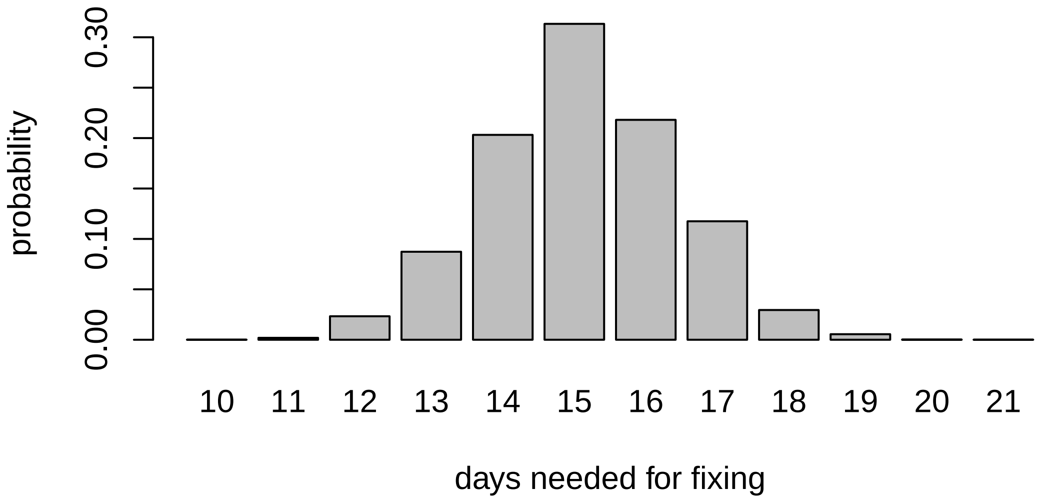
$$T := \sum_{i=1}^{10} X_i \sim \mathsf{Poisson}(123)$$

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)
```

10	11	12	13	14	15	16	17	18	19	20	21
1	19	233	873	2032	3133	2181	1175	295	55	2	1

```r
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.

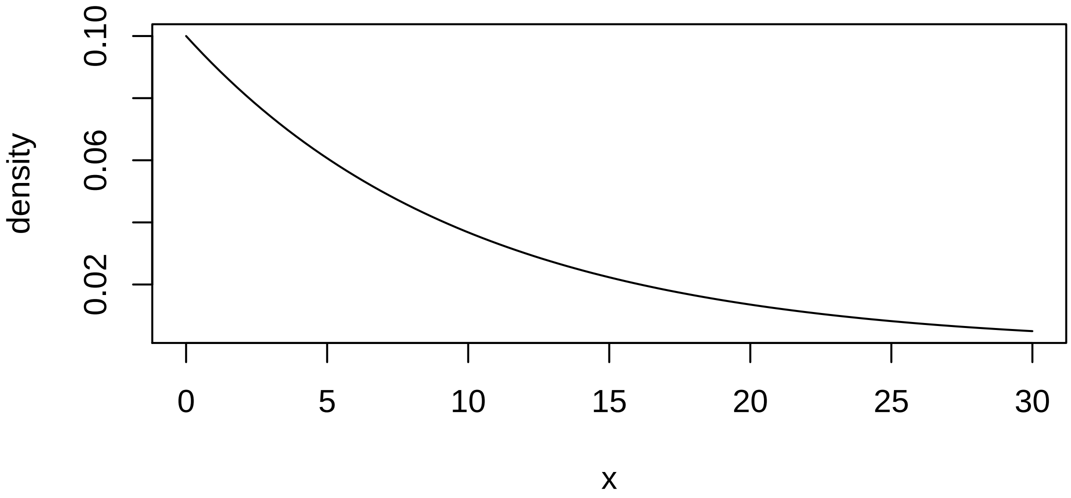
(a) Determine the failure rate and plot the PDF of reliability.
(b) Find the probability that the machine operates for at least 12 hours before breaking down.
(c) If the machine has already been operating 8 hours, what is the probability that it will last another 4 hours? [hint: remember that the exponential 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 guess what is has to be.

```r
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 isn't the best practice because 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

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

```r
# 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

```r
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?

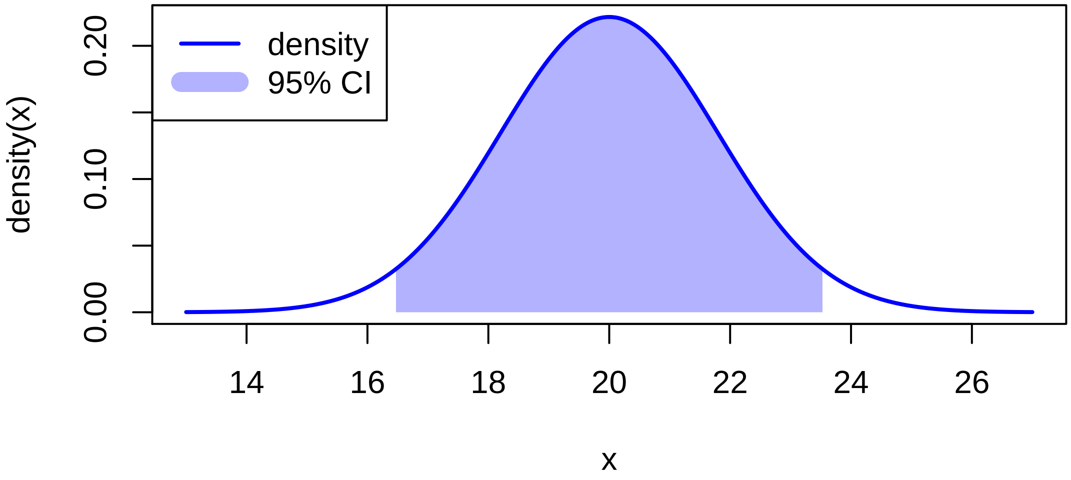
Solution: Since we know the distribution of the temperature we can obtain confidence interval right away in R

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

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

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

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

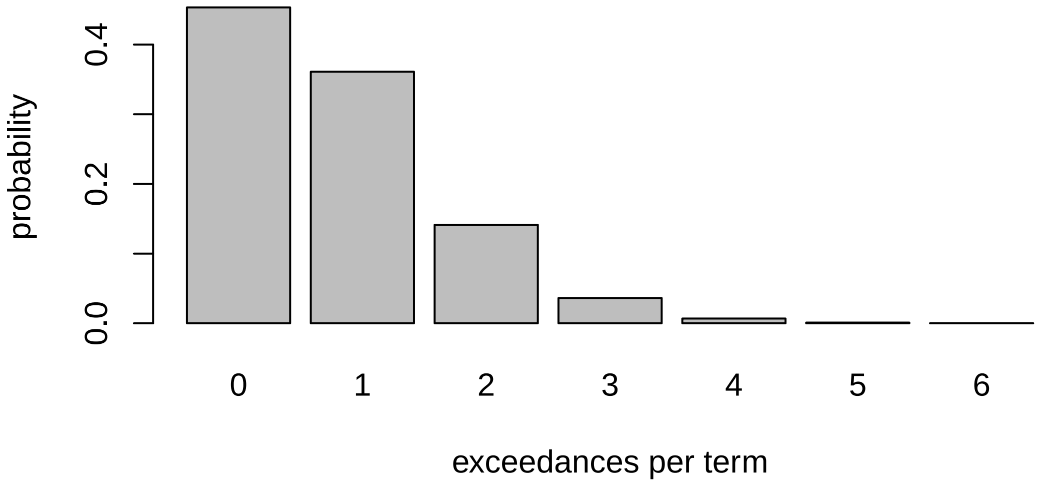
```
## [1] 0.7880487
```

Assuming that the term has 12 weeks and 5 working days each week (whatever numbers you take here is fine) 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

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

We can visualise this as before

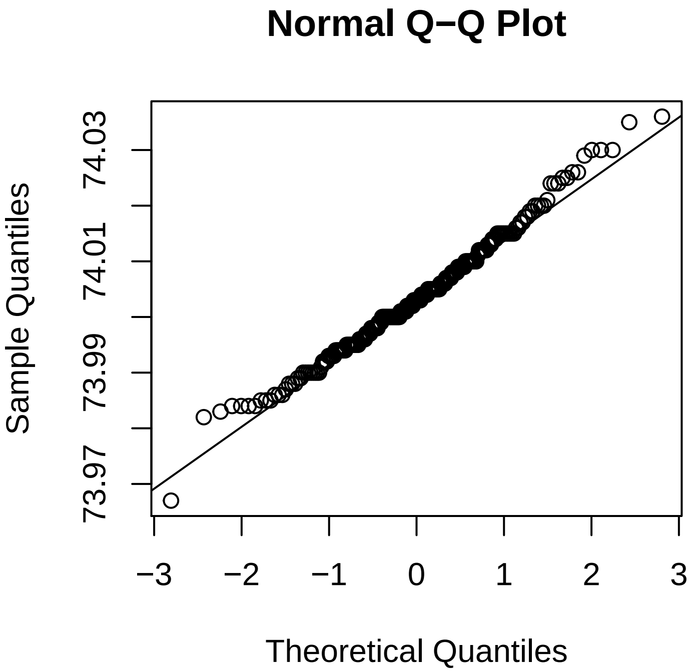
```
barplot(dbinom(0:6, size = 60, prob = 0.0131),
        names.arg = 0:6,
        ylab = "probability",
        xlab = "exceedances per term")
```



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

Solution: 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.

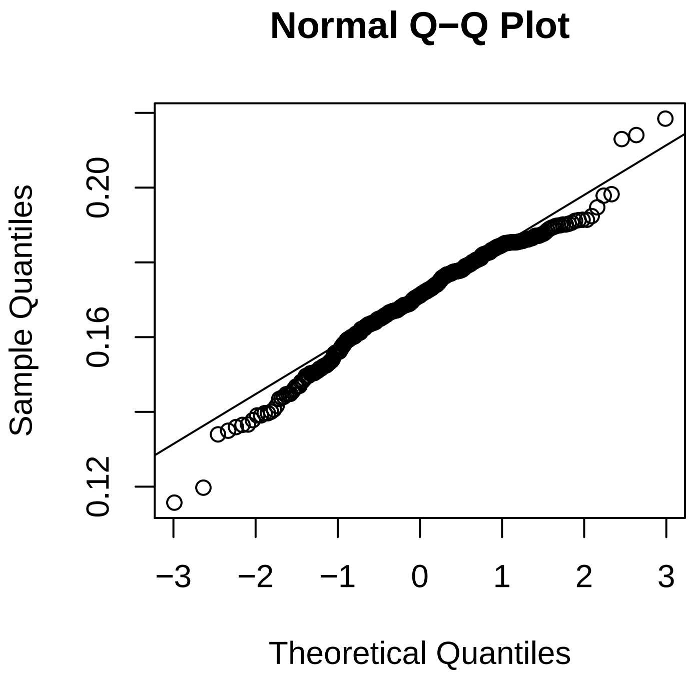
```
shapiro.test(pistonrings$diameter)
```

```
##
##  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

Exercise 1: The number of trucks arriving at a warehouse in over 500 one hour interval has been recorded. Counts of zero up to eight arrivals are recorded respectively on 52, 151, 130, 102, 45, 12, 5, 1 and 2 occasions. Test the hypothesis that these have a Poisson distribution and estimate how often there will be nine or more arrivals per hour.

Solution 1 (a): The exercise doesn't specify a parameter for the Poisson distribution so we need to guess one ourselves. This is a subtle issue. If we assume that an oracle told us some parameter that is believed to be the true one (i.e. what will form the null hypothesis) then we could use that. This isn't the case so in some way the best we can do is calculate the average from the data and take that. In R this can be done with the following command

```
(52*0 + 151*1 + 130*2 + 102*3 + 45*4 + 12*5 + 5*6 + 1*7 + 2*8) / 500
```

```
## [1] 2.02
```

and we could test whether the data fits a Poisson(2.02) distribution. **BUT** we have estimated a parameter from the data so we loose one degree of freedom (don't worry if that sounds new to you - it hasn't been treated in class as far as I know). This situation is briefly described in the book on page 947. Another issue is that the built-in R routine doesn't have a way to handle this (since it can't possibly know what we have estimated from the data there isn't really a way to implement this). Let's first consider the situation where an oracle told us that the parameter we should consider is $\lambda_0 = 2.5$. We can group the last three observations into a "> 5"- bin and calculate probabilities

```
observed.trucks <- c(52, 151, 130, 102, 45, 12, 5 + 1 + 2)
names(observed.trucks) <- c(as.character(0:5), ">5")
print(observed.trucks)
```

```
##   0   1   2   3   4   5  >5
##  52 151 130 102  45  12   8
```

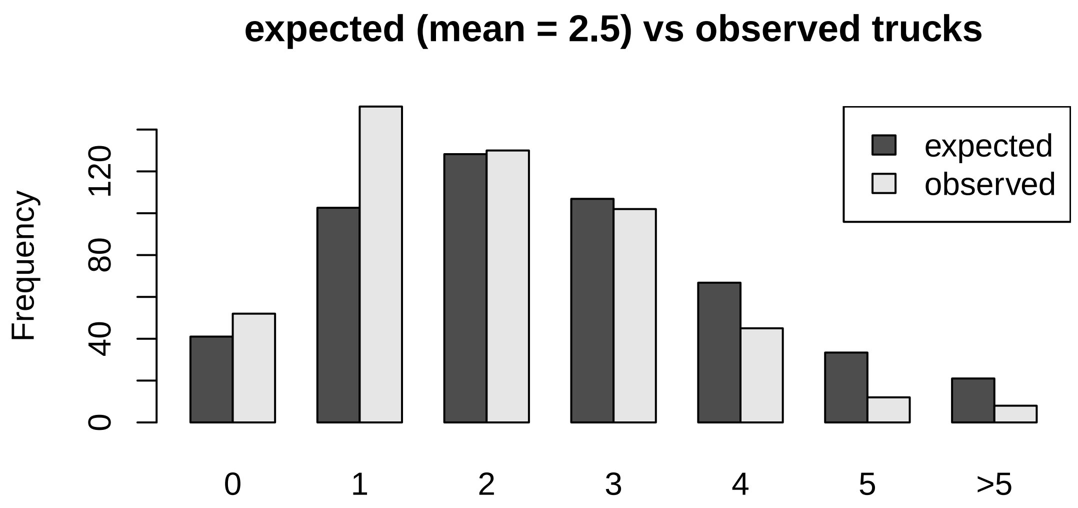
```
probabilities.H0 <- c(dpois(0:5, 2.5), 1-ppois(5,2.5))
names(probabilities.H0) <- c(as.character(0:5), ">5")
print(probabilities.H0)
```

```
##          0          1          2          3          4          5
## 0.08208500 0.20521250 0.25651562 0.21376302 0.13360189 0.06680094
##         >5
## 0.04202104
```

```
expected.trucks <- 500*probabilities.H0
print(expected.trucks)
```

```
##         0         1         2         3         4         5        >5
##  41.04250 102.60625 128.25781 106.88151  66.80094  33.40047  21.01052
```

```
barplot(t(cbind(expected.trucks,observed.trucks)), beside = TRUE,
        ylab="Frequency", main = "expected (mean = 2.5) vs observed trucks")
legend("topright", legend = c("expected", "observed"), fill=gray.colors(2))
```



```
# perform standard chisq-test in R (reject with very strong evidence)
chisq.test(x = observed.trucks, p = probabilities.H0)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed.trucks
## X-squared = 54.88, df = 6, p-value = 4.901e-10
```

Let's now do it as in the book and take away one degree of freedom as well as use the estimated parameter. A very similar procedure as before will give

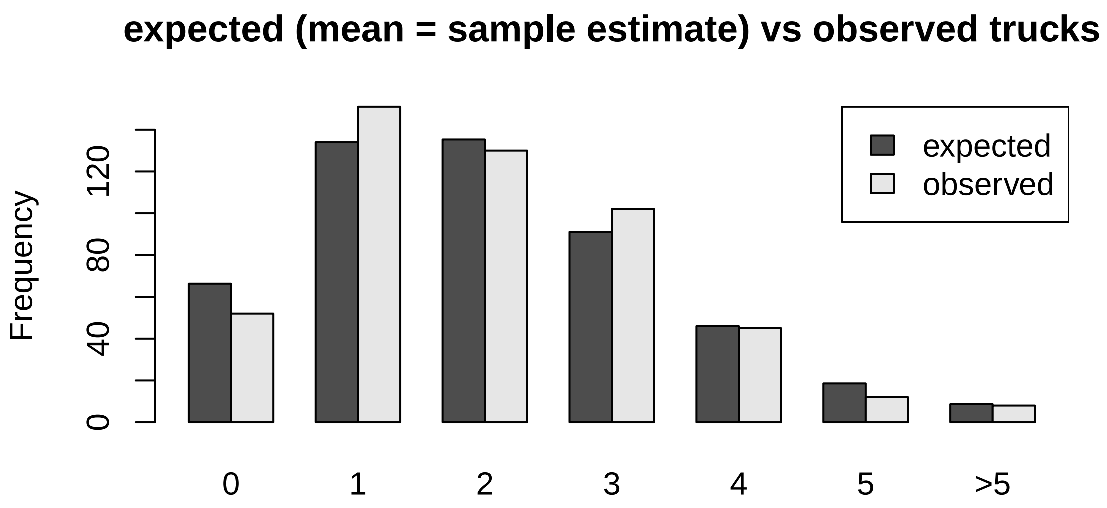
```
probabilities.H0 <- c(dpois(0:5, 2.02), 1-ppois(5,2.02))
names(probabilities.H0) <- c(as.character(0:5), ">5")
print(probabilities.H0)
```

```
##          0          1          2          3          4          5
## 0.13265547 0.26796404 0.27064368 0.18223341 0.09202787 0.03717926
##         >5
## 0.01729627
```

```
expected.trucks <- 500*probabilities.H0
print(expected.trucks)
```

```
##          0          1          2          3          4          5
##  66.327733 133.982020 135.321840  91.116706  46.013936  18.589630
##         >5
##   8.648136
```

```
barplot(t(cbind(expected.trucks,observed.trucks)), beside = TRUE,
        ylab="Frequency", main = "expected (mean = sample estimate) vs observed trucks")
legend("topright", legend = c("expected", "observed"), fill=gray.colors(2))
```



In this case we must manually compare the χ^2-statistic to the quantiles as R will by default have one degree of freedom too much. We can still use the built-in routine to calculate the statistic but we don't (ever) have to. We don't reject in either of the two cases (corrected and uncorrected degrees of freedom).

```
chisq.test(x = observed.trucks, p = probabilities.H0) # uncorrected
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed.trucks
## X-squared = 9.1726, df = 6, p-value = 0.1641
```

```
chisq.statistic <- sum((observed.trucks-expected.trucks)^2/expected.trucks)
df.uncorrected <- length(observed.trucks)-1
print(chisq.statistic) # compare this to X-squared in the above
```

```
## [1] 9.172598
```

```
p.value.chisq <- pchisq(chisq.statistic, df = df.uncorrected, lower.tail = FALSE)
print(p.value.chisq) # compare this to p-value
```

```
## [1] 0.1641014
```

```
# corrected test
chisq.statistic <- sum((observed.trucks-expected.trucks)^2/expected.trucks)
print(chisq.statistic) # statistic is unchanged
```

```
## [1] 9.172598
```

```
p.value.chisq <- pchisq(chisq.statistic, df = df.uncorrected-1, lower.tail = FALSE)
print(p.value.chisq) # compare this to above p-value (it is smaller)
```

```
## [1] 0.1023747
```

Solution 1 (b): Since we didn't reject the null hypothesis of having a $\mathsf{Poisson}(2.02)$ distribution we can assume this is the truth. Under that assumption we find for the probability of it being at least 9

```
paste("The probability of having 9 or more arrivals in one hour is",
      round(100*(1-ppois(8, lambda = 2.02)), digits = 4), "%", sep = " ")
```

```
## [1] "The probability of having 9 or more arrivals in one hour is 0.0255 %"
```

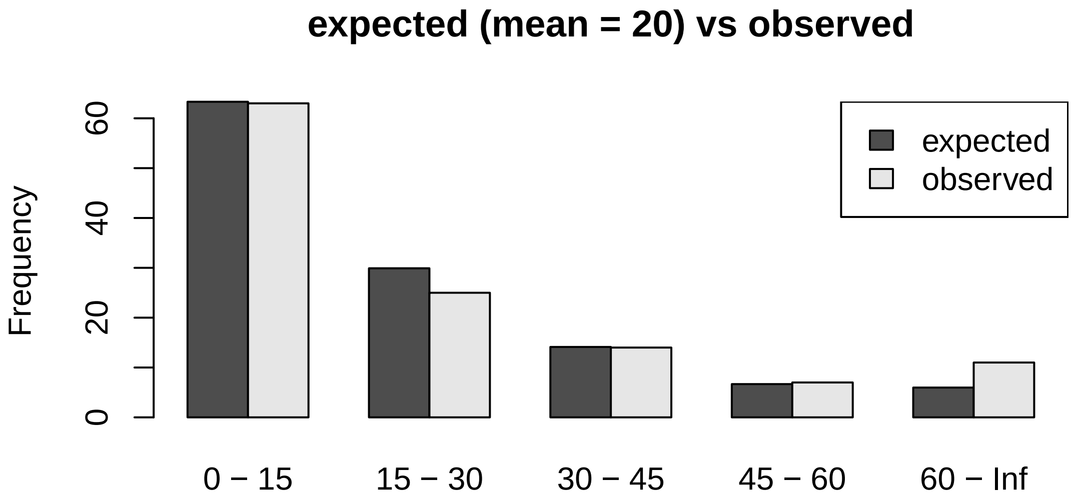
Exercise 2: Traffic is passing freely along a section of the A702, well away from traffic lights, and the interval (in seconds) between successive vehicles has been mea- sured by a roadside sensor as follows [see Workbook for table]: (a) Fit an exponential distribution to the data (b) Test to see if the exponential distribution explains the time interval be- tween cars.

Solution: As before the exercise doesn't specify a parameter for the Exponential distribution. Rather than that it even asks us to fit a parameter to the data. In principle we are free to do whatever we deem appropriate. Since we have binned data we cannot take simply the average in order to estimate the mean which means that you don't know how to properly fit a parameter to this data (there are ways to do this!) we will simply guess the value $\lambda = \frac{1}{20}$. After "fit-by-guess", which we assume doesn't come from the data but rather common sense (so we don't remove a degree of freedom), we perform the same steps as before

```r
# observed data
observations <- c(63,25,14,7,6+3+2)
# probabilities for the bins under H0
lower.bin <- seq(from = 0, by = 15, length.out = 5)
upper.bin <- c(seq(from = 15, by = 15, length.out = 4), Inf)
prob.exp <- pexp(upper.bin, rate = 1/20)-pexp(lower.bin, rate = 1/20)
names(prob.exp) <- paste(lower.bin, upper.bin, sep = " - ")
print(prob.exp)
```

```
##     0 - 15    15 - 30    30 - 45    45 - 60   60 - Inf
## 0.52763345 0.24923639 0.11773094 0.05561216 0.04978707
```

```r
barplot(t(cbind(prob.exp*sum(observations),observations)), beside = TRUE,
        ylab="Frequency", main = "expected (mean = 20) vs observed")
legend("topright", legend = c("expected", "observed"), fill=gray.colors(2))
```


expected (mean = 20) vs observed

We can now perform the standard test with the built-in routine. We do not reject the null hypothesis of having an exponential distribution with rate $\lambda = \frac{1}{20}$, i.e. mean 20.

```r
chisq.test(x = observations, p = prob.exp)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observations
## X-squared = 5.0516, df = 4, p-value = 0.282
```

Exercise 3: Cars produced at a factory are chosen at random for a thorough inspection. The number inspected and the number found to be unsuitable for shipment each month is as follows [see Workbook for table!]. Is there any significant variation in quality throughout the year?

Solution: We will do the test by hand and in R in order to see what is actually going on. Note that the inspected cars contain the sum of defective cars and non-defective ones. The contingency table that we need for the test will contain defective and non-defective ones so this needs to be adjusted (if you assumed that the tested mean tested and non-defective that's fine too).

```
inspected.cars <- c(450,550,550,400,600,450,450,200,450,600,600,550)
defective.cars <- c(8,14,6,3,7,8,16,5,12,6,15,9)
M <- as.table(cbind(defective.cars, inspected.cars-defective.cars))
dimnames(M) <- list(month = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                              "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"),
                    status = c("defective", "non-defective"))

chisq.test.cars <- chisq.test(M)
```

```
## Warning in chisq.test(M): Chi-squared approximation may be incorrect
```

```
print(chisq.test.cars)
```

```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 20.544, df = 11, p-value = 0.03842
```

```
# manually calculating it
Total.cars <- sum(inspected.cars)
r.count <- inspected.cars # monthly counts
c.count <- c(sum(defective.cars),
             sum(inspected.cars - defective.cars)) # total (non-)defective
count.product <- outer(r.count, c.count)
X.squards.statistic <- sum((M-count.product/Total.cars)^2/(count.product/Total.cars))
print(X.squards.statistic)
```

```
## [1] 20.54385
```

```
df.cars.contingency <- (length(r.count)-1) * (length(c.count)-1)
print(df.cars.contingency)
```

```
## [1] 11
```

```
P.cars.contingency <- 1-pchisq(X.squards.statistic, df = df.cars.contingency)
print(P.cars.contingency)
```

```
## [1] 0.03841678
```

```
# # obtain expected and residual matrices
# # expected
# chisq.test.cars$expected
# count.product/Total.cars
#
# # residuals
# chisq.test.cars$residuals
# (M-count.product/Total.cars)/sqrt(count.product/Total.cars)
```

at 5% we will reject the homogeneity hypothesis but at 1% we will not.

Exercise 4: The increased availability of light, high strength, materials has revolutionised the design and manufacture of drivers for playing golf. Clubs with hollow heads and thin faces result in much longer tee shots due to improved transfer of momentum between the club and the ball. The momentum transfer is measured using the coef- ficient of restitution (the ratio between the velocity of the club and the velocity of the ball). In an experiment 15 clubs (of the same type, from the same manufacturer) were selected at random and their coefficient of restitution measured. The measurements were taken by firing a standard golf ball from an air cannon (so the spin and incident velocity could be precisely controlled) at the face of the club (held firmly in a clamp) and measuring the exit velocity. Determine if there is statistical evidence to sup- port the manufactures claim that the coefficient of restitution is greater than 0.82. [Remember to check that the data is normally distributed.]

Solution: We use R in order to see whether the data is normally distributed first.

```r
data.set <- c(0.8411,0.8191,0.8182,0.8125,0.8750,
              0.8580,0.8532,0.8483,0.8276,0.7983,
              0.8042,0.8730,0.8282,0.8359,0.8660)
# test normality first
shapiro.test(data.set) # qqnorm(data.set) and qqline(data.set) might be useful
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data.set
## W = 0.96087, p-value = 0.7075
```

Since we didn't reject normality we can perform a t-test. Note that if we had rejected that the data has a Normal distribution this would be inconsistent (that means you shouldn't do that!). Since we didn't we can use R to perform the test.

In hypothesis testing, there is always the question what the null hypothesis should be. In some ases there can be a natural choice but in some cases there isn't. In principle, there is no single right answer to that. Many times the null hypothesis is chosen such that a test exists, i.e. we are happy that we can test anything at all. A general guidline is that one should try to choose the null hypothesis in such a way that the type I error (reject H_0 when true) is worse than the type II error (accept H_0 when false) because the former is controlled by the significance level α and the latter is often hard to determine (it can, and often is, be quite large!). In the present case we pick (and reject)

$$H_0 : \mu \leq 0.82 \qquad\qquad H_1 : \mu > 0.82$$

```r
t.test(data.set, mu = 0.82, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  data.set
## t = 2.719, df = 14, p-value = 0.008313
## alternative hypothesis: true mean is greater than 0.82
## 95 percent confidence interval:
##  0.8260722       Inf
## sample estimates:
## mean of x
##   0.83724
```

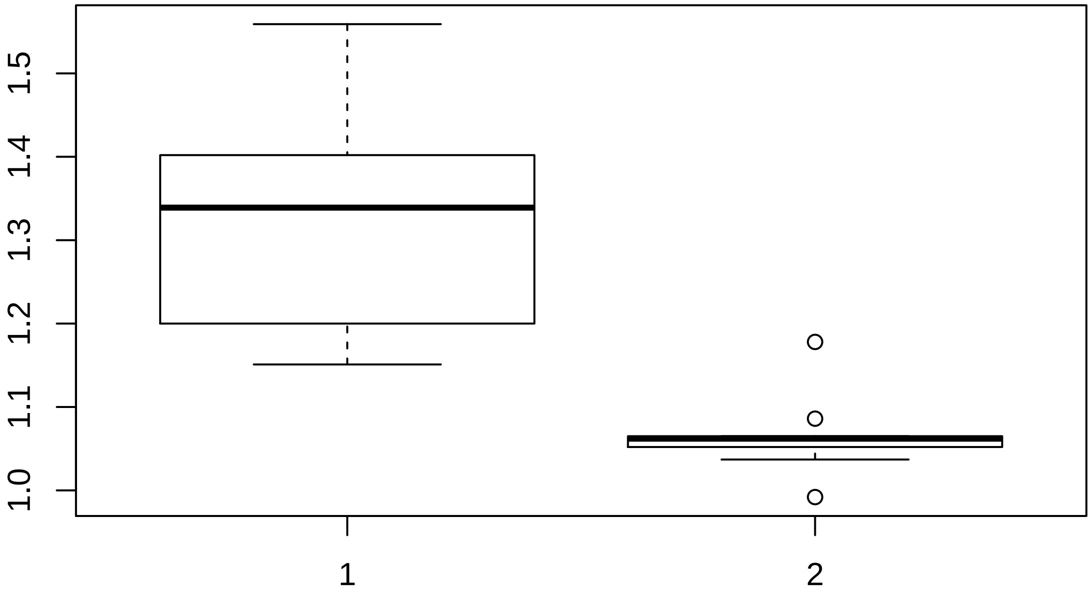
From the output we see that there is enough evidence to reject H_0 at 5% significance, i.e. to support the claim of the manufacturer.

Exercise 5: An article in the Journal of Strain Analysis (1983, vol 18, no 2) compares several meth- ods for predicting the shear strength of a steel plate girder. Data for the Karlsruhe and Lehigh methods are shown bellow applied to nine specific girders. Is there, on average, any statistically significant difference between the two methods? [See Workbook for table!].

Solution: We first get the data into R.

```
strength.data <- c(1.186, 1.061, 1.151, 0.992, 1.322, 1.063, 1.339, 1.062, 1.200,
                   1.065, 1.402, 1.178, 1.365, 1.037, 1.537, 1.086, 1.559, 1.052)

strength.data <- matrix(strength.data, byrow = TRUE, ncol = 2)
# print(strength.data)
boxplot(strength.data) # boxplot suggests different means !!
```



```
t.test(strength.data[,1], strength.data[,2], paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  strength.data[, 1] and strength.data[, 2]
## t = 6.0819, df = 8, p-value = 0.0002953
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1700423 0.3777355
## sample estimates:
## mean of the differences
##               0.2738889
```

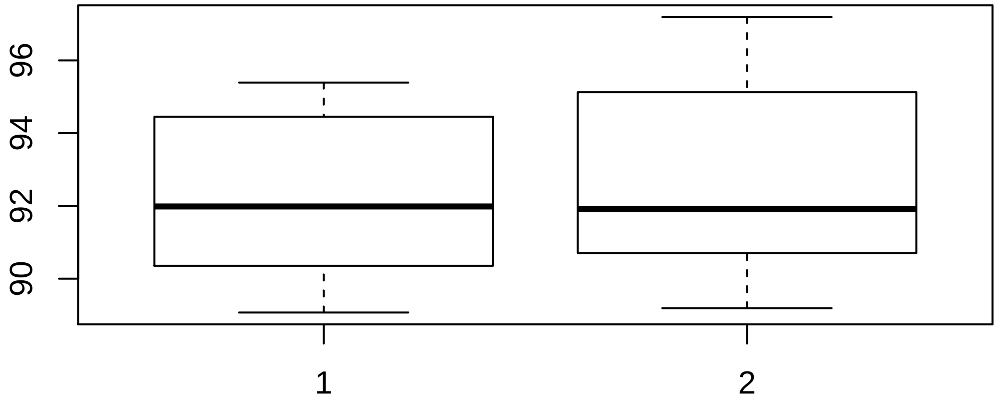
We conclude that there is sufficient statistical evidence to believe that there is a difference between the two methods in question.

Exercise 6: Two catalysts are being tested to determine how they affect the yield of a particular chemical process. Currently catalyst1 is being used but catalyst2 is cheaper and will save the company a lot of money if it does not affect the process yield. A series of test runs have been conducted using a pilot plant and the data is shown in Table 3.4, test the hypothesis that the choice of catalyst dose not affect the yield. Since all the test runs were conducted in the same plant under the same conditions (except for the choice of catalysts) you may assume equal variances. You should make the test using categorised data in a data-frame. [See Workbook for table!].

Solution: We first get the data into R. After that we could test the hypothesis whether

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$

```
Catalyst.data <- c(91.50, 89.19, 94.18, 90.95, 92.18,
                   90.46, 95.39, 93.21, 91.79, 97.19,
                   89.07, 97.04, 94.72, 91.07, 89.21, 92.75)
Catalyst.category <- rep(factor(c(1,2)), times = 8)
test.data <- data.frame(cbind(Catalyst.category, Catalyst.data))
boxplot(test.data$Catalyst.data ~ test.data$Catalyst.category)
```



```
# specify equal variances
t.test(test.data$Catalyst.data ~ test.data$Catalyst.category, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  test.data$Catalyst.data by test.data$Catalyst.category
## t = -0.35359, df = 14, p-value = 0.7289
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.373886  2.418886
## sample estimates:
## mean in group 1 mean in group 2
##         92.2550         92.7325
```

We conclude that there is no sufficient evidence to reject the null hypothesis that both catalysts perform equally. However, in the given situation it might be actually "safer" to test another hypothesis. If the company is thinking about replacing a working solution, then they want to be sure that the replacement won't cause problems, i.e. decrease the yield, as this might cost them a lot more money than they will save. Instead of testing the above hypothesis we could also test the hypothesis whether catalyst 2 affect the yield negatively. In order to be on the safe side we pick

$$H_0 : \mu_1 \geq \mu_2 \qquad\qquad H_1 : \mu_1 < \mu_2$$

```
t.test(test.data$Catalyst.data ~ test.data$Catalyst.category,
       var.equal = TRUE, alternative = "less")
```

```
##
##  Two Sample t-test
##
## data:  test.data$Catalyst.data by test.data$Catalyst.category
## t = -0.35359, df = 14, p-value = 0.3645
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.901027
## sample estimates:
## mean in group 1 mean in group 2
##         92.2550         92.7325
```

This hypothesis cannot be rejected either at 5% significance level. Note that in the first case we would have replaced the catalyst whereas in the second case we would assume that catalyst two is worse than the current solution and we wouldn't replace catalyst 1 by it. As we see in this example, the decisions we make based on hypothesis tests are highly dependent on the hypotheses. Those should be specified according to the (real world) risk associated with the type I and type II errors whenever this is feasible.

Example of a 2-way ANOVA (incl. formulas): This fully worked out example is maily there to provide you with the basic formulas and ideas behind ANOVA. Consider the example on page 59 of the Workbook. We can get the data in `R` by typing it in manually using the values provided in the table.

```r
FeedStock <- rep(c("I","II","III","IV","V"), each = 2, times = 3)
Process <- rep(c("A","B","C"), each = 10)
Process.FeedStock <- paste(Process, FeedStock, sep ="-")
Yield <- c(106,110,95,100,94,107,103,104,100,102,110,112,98,99,100,
           101,108,112,105,107,94,97,86,87,98,99,99,101,94,98)
full.chem.data <- data.frame(FeedStock = factor(FeedStock),
                             Process = factor(Process),
                             Yield = Yield)
full.chem.data # we can have a look at it
```

FeedStock	Process	Yield
I	A	106
I	A	110
II	A	95
II	A	100
III	A	94
III	A	107
IV	A	103
IV	A	104
V	A	100
V	A	102
I	B	110
I	B	112
II	B	98
II	B	99
III	B	100
III	B	101
IV	B	108
IV	B	112
V	B	105
V	B	107
I	C	94
I	C	97
II	C	86
II	C	87
III	C	98
III	C	99
IV	C	99
IV	C	101
V	C	94
V	C	98

We can produce an ANOVA table easily from that

```
anova(chemmod <- lm(Yield~FeedStock+Process+FeedStock:Process))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
FeedStock	4	449.4667	112.366667	12.393382	0.0001189
Process	2	512.8667	256.433333	28.283088	0.0000081
FeedStock:Process	8	143.1333	17.891667	1.973346	0.1220900
Residuals	15	136.0000	9.066667	NA	NA

But how did we arrive at these values? In order to understand what a 2-way ANOVA analysis is and how it works it might be useful to look at the formulas that `R` used to arrive at these values. Let $\mathcal{I} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ be the set of lables for the Process category and $\mathcal{J} = \{\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}, \mathrm{V}\}$ the ones for the Feed Stock category. For any pair of lables $i \in \mathcal{I}$ and $j \in \mathcal{J}$, i.e. a fixes Process and Feed Stock label, we denote by $\mathcal{K}_{ij}$ the set of observation, i.e. rows in the above table, with the right labels. In the given examples $\mathcal{K}_{ij}$ always contains two elements. For $i = \mathrm{A}, j = \mathrm{III}$ that would be the two rows with Yield 94 and 107. by $|\cdot|$ we denote the cardinality of a set, i.e. the number of elements in it. We will assume that all $\mathcal{K}_{ij}$ are of equal size.

Sum of squares partitioning: The key in ANOVA are different types of "sums of squares". Let's start by defining the following averages

$$\begin{aligned}
\bar{y}_{\bullet\bullet} &= \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{K}_{ij}|}\sum_{k\in\mathcal{K}_{ij}} y_{ijk} \\
\bar{y}_{i\bullet} &= \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\frac{1}{|\mathcal{K}_{ij}|}\sum_{k\in\mathcal{K}_{ij}} y_{ijk} \\
\bar{y}_{\bullet j} &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\frac{1}{|\mathcal{K}_{ij}|}\sum_{k\in\mathcal{K}_{ij}} y_{ijk} \\
\bar{y}_{ij} &= \frac{1}{|\mathcal{K}_{ij}|}\sum_{k\in\mathcal{K}_{ij}} y_{ijk}
\end{aligned}$$

Next we can define the actual sums of squares. Unsurprisingly, all of them can be written as an actual sum of squares. Some of these formulas may not be very intuitive and you don't have to memorise them.

$$\begin{aligned}
SS_{\text{tot}} &= \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_{ij}} (y_{ijk} - \bar{y}_{\bullet\bullet})^2 \\
SS_{\bullet\mathcal{J}} &= \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_{ij}} (y_{\bullet j} - \bar{y}_{\bullet\bullet})^2 \\
SS_{\mathcal{I}\bullet} &= \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_{ij}} (y_{i\bullet} - \bar{y}_{\bullet\bullet})^2 \\
SS_{\mathcal{I}\mathcal{J}} &= \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_{ij}} (\bar{y}_{ij} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{y}_{\bullet\bullet})^2 \\
SS_{\text{res}} &= \sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}_{ij}} (y_{ijk} - \bar{y}_{ij})^2
\end{aligned}$$

It is always true that the total sum of squares is the sum of the other, i.e.

$$SS_{\text{tot}} = SS_{\bullet\mathcal{J}} + SS_{\mathcal{I}\bullet} + SS_{\mathcal{I}\mathcal{J}} + SS_{\text{res}}$$

and that these sums of squares (which are of course realisations of random variables) satisfy certain independence relations under certain null hypotheses that allow us to make the test which we make. If you are interested (but you don't need to know this), the basis for all of this ANOVA theory is Cochran's Theorem.

Degrees of freedom: The next quantity that appears in the ANOVA table is the degrees of freedom. They are exactly the degrees of freedom associated to the χ^2 random variables that form the sums of squares. Let $M = |\mathcal{J}|, N = |\mathcal{I}|, K = |\mathcal{K}_{ij}|$. Recall that all $\mathcal{K}_{ij}|$ were assumed to have equal size. They can be calculated as follows

$$\begin{aligned} df_{\text{tot}} &= MNK - 1 \\ df_{\bullet\mathcal{J}} &= N - 1 \\ df_{\mathcal{I}\bullet} &= M - 1 \\ df_{\mathcal{I}\mathcal{J}} &= MN - (df_{\bullet\mathcal{J}} + df_{\mathcal{I}\bullet}) - 1 \\ df_{\text{res}} &= MNK - MN \end{aligned}$$

As was the case with the sums of squares, the degrees of freedom also sum to their total amount, i.e.

$$df_{\text{tot}} = df_{\bullet\mathcal{J}} + df_{\mathcal{I}\bullet} + df_{\mathcal{I}\mathcal{J}} + df_{\text{res}}$$

The Mean Square: The mean square is nothing other than a normalised sum of squares. The idea of ANOVA is to compora the magnitude of sums of squares. But the χ^2 distribution takes typically values around its degrees of freedom (which is nothing but their mean). In order to compare them objectively we need to normalise these quantities. This is what is formally done with the mean square. It is defined for each of the symbols above as

$$MS = \frac{SS}{df}$$

The total mean square can technically be defined but is usually left out because, unlike with the quantities before, it isn't the sum of all the others. **F Statistic and p-values:** As you can see in the table produced from `R` those are not defined for the residuals. The reason for that is evident from the formula. The F Statistic for each row is defined as

$$\begin{aligned} F_{\bullet\mathcal{J}} &= \frac{MS_{\bullet\mathcal{J}}}{MS_{\text{res}}} \\ F_{\mathcal{I}\bullet} &= \frac{MS_{\mathcal{I}\bullet}}{MS_{\text{res}}} \\ F_{\mathcal{I}\mathcal{J}} &= \frac{MS_{\mathcal{I}\mathcal{J}}}{MS_{\text{res}}} \end{aligned}$$

The F statstic for the residuals would always be equal to 1 (and that's silly) so one never writes it. The p-values are defined as always in hypothesis testing. The distribution of the F statistic is $\mathsf{F}(df_\circ, df_{\text{res}})$ distributed (yes, that's a distribution) where the symbol takes the three possible values $\circ \in \{\bullet\mathcal{J}, \mathcal{I}\bullet, \bullet\bullet\}$.

The below code gives you an example of how to produce the ANOVA table by hand with those formulas. As you can see the values match those produced by `R` precisely (except that we additionally have the total sum of squares in the last row).

```r
# single SumOfSquares
# decomposition
CrossMeans <- rep(tapply(Yield, Process.FeedStock, mean), each = 2)
ProcessMeans <- rep(tapply(Yield, Process, mean), each = 10)
FeedStockMeans <- rep(tapply(Yield, FeedStock, mean), each = 2, times = 3)
null.mean <- mean(Yield)

SS.FeedStock <- sum((FeedStockMeans - null.mean)^2)
df.FeedStock <- length(unique(FeedStock))-1
MS.FeedStock <- SS.FeedStock/df.FeedStock

SS.Process <- sum((ProcessMeans - null.mean)^2)
df.Process <- length(unique(Process))-1
MS.Process <- SS.Process/df.Process

SS.Cross <- sum((CrossMeans - ProcessMeans - FeedStockMeans + null.mean)^2)
df.Cross <- length(unique(Process.FeedStock)) - (df.FeedStock + df.Process) - 1
MS.Cross <- SS.Cross/df.Cross

SS.error <- sum((Yield - CrossMeans)^2)
df.error <- length(Yield) - length(unique(Process.FeedStock))
MS.error <- SS.error/df.error

SS.total <- sum((Yield - mean(Yield))^2)
df.total <- length(Yield) - 1

F.FeedStock <- MS.FeedStock/MS.error
P.FeedStock <- 1-pf(F.FeedStock, df1 = df.FeedStock, df2 = df.error)
F.Process <- MS.Process/MS.error
P.Process <- 1-pf(F.Process, df1 = df.Process, df2 = df.error)
F.Cross <- MS.Cross/MS.error
P.Cross <- 1-pf(F.Cross, df1 = df.Cross, df2 = df.error)


data.frame(SumOfSquares = c(SS.FeedStock, SS.Process, SS.Cross, SS.error, SS.total),
           DegreesOfFreedom = c(df.FeedStock, df.Process, df.Cross, df.error, df.total),
           MeanSquare = c(MS.FeedStock, MS.Process, MS.Cross, MS.error, NA),
           FStatistic = c(F.FeedStock, F.Process, F.Cross, NA, NA),
           PValue = c(P.FeedStock, P.Process, P.Cross, NA, NA))
```

SumOfSquares	DegreesOfFreedom	MeanSquare	FStatistic	PValue
449.4667	4	112.366667	12.393382	0.0001189
512.8667	2	256.433333	28.283088	0.0000081
143.1333	8	17.891667	1.973346	0.1220900
136.0000	15	9.066667	NA	NA
1241.4667	29	NA	NA	NA

Exercise 4: In Dr Tom Bruce's experiments on sea walls (see Section 1.12) the mean over-topping rate was measured on a model sea wall with 13 different types of armour unit.]

(a) Use 1-way ANOVA to answer the questions "Does the armour unit affect the overtopping volume?". You should use a linear model based on

$$\log(q_{ij}) = \mu + U_i + \varepsilon_{ij}$$

where U_i is the effect of armour unit type, i and ε_{ij} is the error.

(b) Repeat the analysis also including the relative crest height, R^*, as a predictor, using 2-way ANOVA and the linear model

$$\log(q_{ij}) = \mu + U_i + R^* + \varepsilon_{ij}$$

Solution 4 (a): We first get the data into R and remove the zeros (since we can't take a logarithm of these) and fit a regression with structure only.

```
clash.data <- clash.2d[clash.2d$q..cumecs. > 0,]
lm.clash <- lm(log(clash.data$q..cumecs.) ~ clash.data$Structure)
anova(lm.clash)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
clash.data$Structure	13	84.47659	6.498200	1.17723	0.2948504
Residuals	316	1744.29085	5.519908	NA	NA

In part (b), the exercise ask us to do the regression without interaction. This makes a difference. If we want to use height as a category then we need to convert it to a factor before giving it to the regression.

```
lm.clash <- lm(log(clash.data$q..cumecs.) ~ clash.data$Structure+as.factor(clash.data$Rc))
anova(lm.clash)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
clash.data$Structure	13	84.47659	6.498200	1.233082	0.2544624
as.factor(clash.data$Rc)	12	142.24582	11.853819	2.249351	0.0097923
Residuals	304	1602.04502	5.269885	NA	NA

Given that output it seems as if the crest height has some influence on the overtopping but the structure doesn't.