MATH10282 Introduction to Statistics
Semester 2, 2016-17
Coursework using R - Mark scheme (v3)

Question 1: Weather station data

(a) Read the data into R.

Download the data from Blackboard, save it to a folder, and change the working directory in R to that folder. Then do the following:

```
dur <- read.table("durham.txt",header=TRUE)
east <- read.table("eastbourne.txt",header=TRUE)
```

[2 marks]

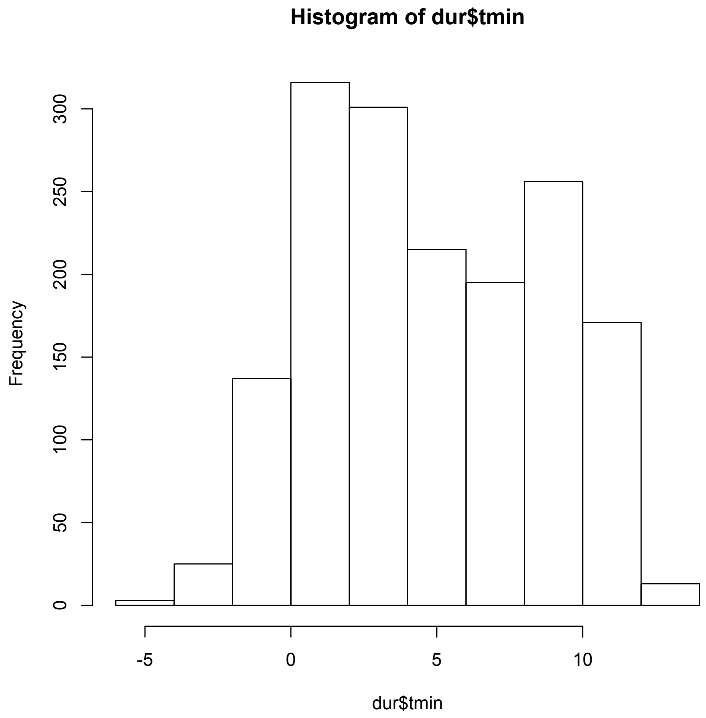
(b) When do records begin at the Eastbourne station? When do records begin at the Durham station?

From inspecting the text file, the records for Eastbourne begin in January 1959, whereas those for Durham begin in January 1880.

[2 marks; 1 for correct year, 1 for correct month]

(c) Plot a histogram of the distribution of the average daily minimum temperature at Durham. Comment on any special features of the shape of the distribution.

```
  hist(dur$tmin)
```



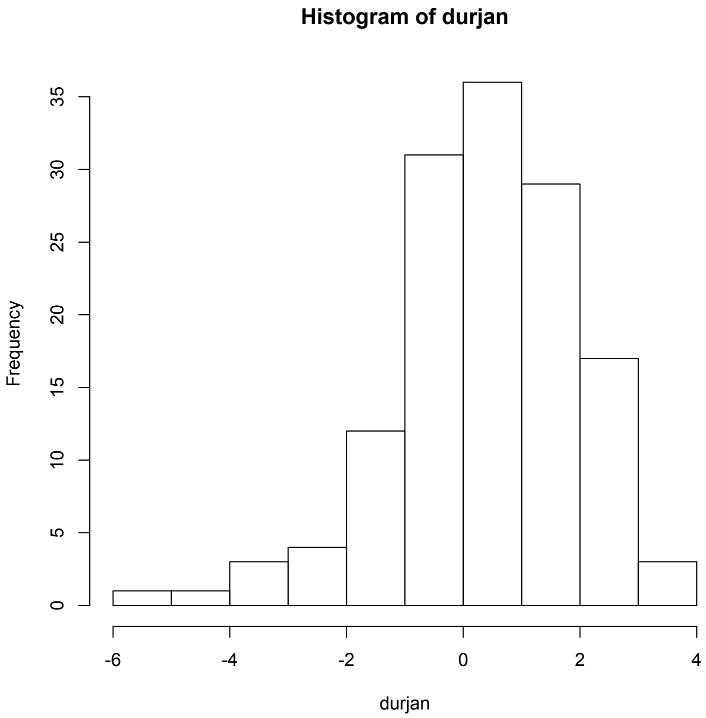
The distribution appears to be bimodal.

[4 marks; 2 for plot, 2 for comment.]

- Award 1 comment mark if they have the right idea but poorly expressed.

(d) Plot a histogram of the distribution of the average daily minimum temperatures at Durham in January. Do you notice anything interesting about the shape of the distribution compared to part (c)? [Hint: create a logical vector which takes the value `TRUE` when the month is January, and `FALSE` otherwise.]

```
dur$mm==1
durjan <- dur$tmin[dur$mm==1]
hist(durjan)
```

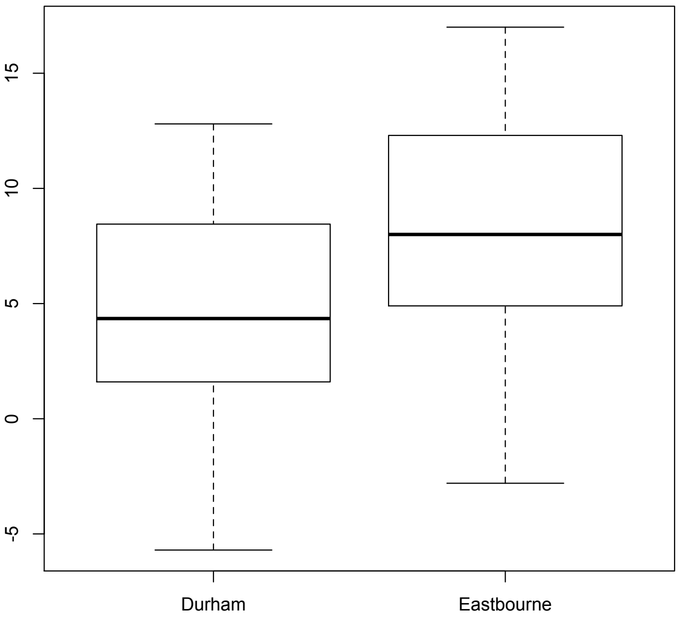


The distribution no longer appears bimodal.

[1 mark for plot, 1 mark for comment]

(e) Draw box plots comparing the average daily minimum temperatures at Durham with those at Eastbourne. Put both box plots on the same axes. Comment on what your box plots indicate. [Hint: an appropriate command will accept two data vector arguments]

```
boxplot(dur$tmin, east$tmin, names=c("Durham","Eastbourne"))
```



The main point is that the boxplots suggest it is generally colder in Durham than Eastbourne. The median temperature is around 3.65 degrees C lower in Durham. There is some indication of skewness in both distributions. However, these plots obscure the bimodal nature of the distributions.

[1 mark for a correct plot, 1 mark for sensible comments e.g. Durham colder, not necessarily all points]

[Total 12 marks for Question 1]

Question 2: Coverage of confidence intervals

In this question, you will investigate the coverage properties of confidence intervals using simulation. Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, and let $\mathbf{X} = (X_1, \ldots, X_n)$. Define an interval estimator for μ via

$$I(\mathbf{X}) = \left[\bar{X} - \frac{cS}{\sqrt{n}}, \bar{X} + \frac{cS}{\sqrt{n}} \right] ,$$

where c is a constant to be chosen appropriately. Recall from lectures that the *coverage probability* of $I(\mathbf{X})$ is defined as $\mathrm{P}[I(\mathbf{X}) \ni \mu]$, i.e. the probability that the random endpoints of the interval contain the true value of the parameter μ.

(a) What value of c should be chosen to ensure that the coverage probability of $I(\mathbf{X})$ is *exactly* $1-\alpha$, in other words to ensure that $I(\mathbf{X})$ is an exact $100(1-\alpha)\%$ confidence interval for μ?

A value $c = t_{\alpha/2}$ ensures this, where $t_{\alpha/2}$ is the upper $\alpha/2$ point of a t distribution with

$n-1$ degrees of freedom.

[1 mark. Must state number of degrees of freedom.]

For the remainder of this question, suppose that the value $c = z_{\alpha/2}$ is used instead, where $z_{\alpha/2}$ denotes the upper $\alpha/2$ point of a $N(0,1)$ distribution.

(b) What would you anticipate to be the approximate value of the coverage probability of $I(\mathbf{X})$ in this case? Explain why there might be a difference between your suggestion and the actual coverage probability if $n = 9$.

By the central limit theorem, if n is large, the coverage probability will be approximately $1-\alpha$. However, for small n the central limit theorem approximation may be poor and as a result coverage probability may be somewhat different.

[1 mark] **N.B. mark this jointly with 2(e)**. If between the two questions they get the correct approximate coverage, award 1 for 2(c). If between the two questions they get the reason (n is too small), then award the comment mark for 2(e).

(c) Write additional code to complete the R function given overleaf. Your completed function should do the following...

The solution is

```
coverage <- function(n=9, mu=5, sigma=2, alpha=0.05, nsims=10000) {

   xi <- numeric(n)
   xbar <- numeric(nsims)
   covered <- numeric(nsims)

   # repeat the following nsims times
   for (i in 1:nsims) {
      xi  <- rnorm(n, mean=mu, sd=sigma)  # simulate a new dataset of size n
      xbar <-  mean(xi)       # these lines need to be completed
      s    <-   sd(xi)      # .
      zalpha <- qnorm(1-alpha/2)      # .
      upper <-  xbar+zalpha*s/sqrt(n)      # .
      lower <-  xbar-zalpha*s/sqrt(n)   # .
      covered[i] <- (upper>=mu)&(lower<=mu)  # .
      }
      return(covered)
}
```

- correct end points (ok to compute $z_{0.025}$ a different way, e.g. `qnorm(0.975)` or 1.96) - 2 marks.

  Award 1 here if basically correct barring a minor slip.

- correct logical vector (ok to use `<` or `<` instead of `>=`). - 1 mark

  If they use `5` rather than `mu`, dock the mark here.

(d) Use your completed R function to simulate 10,000 datasets each of size $n = 9$, with $\mu = 5$, $\sigma^2 = 2^2$, and calculate for each dataset whether $I(\mathbf{X})$ contains μ, using $\alpha = 0.05$. [N.B. you do **not** need to record the individual data sets or confidence intervals, just whether or not $I(\mathbf{X})$ includes μ].

```
ans <- covered()   # or covered(9,5,2,10000)
```

[1 mark for correct R code, even if not able to complete part (c)]

(e) For what proportion of your simulated datasets does $I(\mathbf{X})$ contain the true value of μ? Comment on your result in relation to your answer to part (b).

```
mean(ans)
```

The answer is random but probably around 0.91 or 0.92.

[1 mark for correct R code, even if not able to complete part (c)]

As discussed in part (b), we would expect this proportion to be around 0.95, but it is somewhat lower because n is small and so the central limit theorem approximation is not particularly accurate.

[1 mark - **N.B. mark jointly with 2(b)**]

[Total for Question 2, 8 marks]

Total for coursework, 20 marks