MATH10282 Introduction to Statistics
Semester 2, 2016-17
Coursework using R

The deadline for submitting this coursework is 4.00pm on Friday 28 April 2016. You should upload your report to the assessment entitled 'R Coursework Submission' on the MATH10282 Blackboard site by this time and date. Please note that a similarity report on your work will be generated by Turnitin to detect plagiarism.

This coursework comprises 10% of the overall marks for the module.

Instructions

(a) You should prepare your coursework report using Microsoft Word. A PDF report from another program is also acceptable. You can include code and numerical results from R by copying and pasting into your Word document. Comments and discussion of the results should be added as required. You can save any plots created in R, for example as a PDF, and import these into your final report.

Include in your report all R commands used to generate results.

(b) To facilitate anonymous marking, please do not include your name in your report. You should include your Student ID in the title of your submission on Turnitin, and on the first page of your report.

(c) If you have any queries or problems, please contact me as soon as possible.

Tim Waite, March 2017

Question 1: Weather station data

The data[1] are contained in the files `durham.txt` and `eastbourne.txt` available on Blackboard. They comprise various measurements collected at two weather stations, Durham and Eastbourne. Each row corresponds to a different month.

- Column 1: year (labelled `yyyy`)
- Column 2: month (labelled `mm`), coded 1 (January), . . . , 12 (December).
- Column 3: average daily maximum temperature (for that month) (labelled `tmax`)
- Column 4: average daily minimum temperature (for that month) (labelled `tmin`)

(a) Read the data into R.

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

(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.

[1] adapted from public data released under an Open Government Licence at `https://data.gov.uk/dataset/historic-monthly-meteorological-station-data`

(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)?

(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: try passing two data vectors as arguments to an appropriate function]

[12 marks]

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 $I(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$ is defined as $\mathrm{P}[a(\mathbf{X}) \le \mu \le b(\mathbf{X})]$, i.e. the probability that the random end-points 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?

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 $n = 9$.

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

(i) simulate a number, `nsims`, of different datasets each of size `n` from a $N(\mu, \sigma^2)$ distribution

(ii) for each simulated dataset $\mathbf{X}$ compute the lower and upper endpoints of $I(\mathbf{X})$

(iii) create a logical vector called `covered` whose ith element records whether, for the ith simulated dataset, the interval $I(\mathbf{X})$ contains μ

(iv) return the vector `covered` as output

(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].

(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).

[8 marks]

```
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 <-         # these lines need to be completed
      s    <-         # .
      zalpha <-       # .
      upper <-        # .
      lower <-        # .
      covered[i] <-   # .
      }
      return(covered)
}
```

[END OF COURSEWORK QUESTIONS]