MATH10282 Introduction to Statistics
Semester 2, 2015-16
Coursework using R

The deadline for submitting this coursework is 4.00pm on Friday 15 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 2016

Question 1: Height of London Olympic games athletes

The data comprise the heights (cm) of all athletes who participated in the 2012 Olympic games in London. The data are in the file `'olympic_height.txt'` which contains four columns:

- Column 1: case number
- Column 2: height (cm)
- Column 3: sport
- Column 4: sex (female=F, male=M)

Each row corresponds to a different athlete.

(a) Read the data into R and write code to produce a box plot comparing the height distributions among (i) male basketball players and (ii) male football players. To facilitate comparison, include both sports on a single plot but no others. What do you conclude?

[*Hint: create a logical vector identifying whether a given athlete is of interest for the box plot, and use the* `subset` *command.*]

[2 marks]

Question 2: Confidence intervals for a binary proportion

Let $X \sim \text{Binomial}(n, p)$ record the number of successes in n binary trials with equal probability p of success, where n is known. In this question, we use R to conduct a numerical comparison of the performance of two different interval estimators for p.

Recall from lectures that for large n the 'Wald interval', which is defined by

$$I(X) = \left[\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\,\right],$$

is an approximate $100(1-\alpha)\%$ confidence interval for p, where $\hat{p} = X/n$ and z_α is defined by $\Phi(z_\alpha) = 1 - \alpha$.

(a) Write a sequence of R commands to calculate the upper and lower end-points of $I(X)$. The commands should be written in terms of variables `phat`, `n` and `alpha` that are assumed to store the values of $\hat{p}$, n and α respectively. [*You are not required to write a function here.*]

By assigning appropriate values to `phat`, `n`, and `alpha`, use your commands to calculate the confidence interval when $n = 10$, $X = 5$, $\alpha = 0.05$.

[2 marks]

An alternative interval estimator for p is given by the 'adjusted Wald interval', which is obtained by the following two steps:

Step 1: create a new data set with two additional successes and two additional failures, i.e. $\tilde{X} = X + 2$, $\tilde{n} = n + 4$

Step 2: compute the original Wald interval for the new dataset $(\tilde{X}, \tilde{n})$.

You are not required to justify this definition.

(b) Calculate the adjusted Wald interval when $n = 10$, $X = 5$, $\alpha = 0.05$. How does it compare with your answer to (a)?

[1 mark]

Let $J(X) = [l(X), u(X)]$ be a general interval estimator for p. The **coverage probability**, $C_J(p)$, of J is the probability that $J(X)$ contains the true value of p.

For binomial data X, the coverage probability can be calculated via

$$C_J(p) = \sum_{x=0}^{n} k(x,p)\ \mathrm{P}(X = x)\,, \tag{1}$$

where

$$k(x,p) = \begin{cases} 1 & \text{if } l(x) \le p \le u(x) \\ 0 & \text{otherwise}\,. \end{cases}$$

Note that the value of $k(x,p)$ indicates whether or not the data-dependent endpoints $l(x)$ and $u(x)$ contain p. Thus, (1) above is the sum of $\mathrm{P}(X = x)$ over all values x such that the resulting confidence interval $J(x)$ contains p.

(c) Using the formula (1) above, complete the R code below to create a function that computes the coverage probability of the Wald interval from part (a).

[*Hint: for each possible value of X the function calculates the interval and assesses whether it contains the true value of p.*]

```
cover.prob.wald <- function(n, p, alpha) {
   k <- rep(0,n+1)
   bin.prob <- rep(0,n+1)
   for (i in 1:(n+1)) {
      x      <-  i-1
      phat   <-  x/n
      z      <-              #  these lines need to be completed
      upper  <-              #  .
      lower  <-              #  .
      k[i]   <-              #  .
      bin.prob[i] <-         #  .
   }
   C <- sum(k*bin.prob)
   return(C)
}
```

[2 marks]

(d) Calculate the coverage probability of the Wald interval when $n = 10$, $p = 0.5$ and $\alpha = 0.05$. From theoretical results, what would you have anticipated as an approximate value for the coverage probability? Why is there a difference?

[1 mark]

(e) Write an R function to calculate the coverage probability of the adjusted Wald interval, and calculate its coverage probability for $n = 10$, $p = 0.5$, $\alpha = 0.05$. How does this compare to your answer in part (d)?

[1 mark]

(f) Plot the coverage probability of both the Wald and adjusted Wald intervals as a function of p, when $n = 10$, $\alpha = 0.05$. [*Hint: compute $C_J(p)$ for $p = 0, 0.01, 0.02, \ldots, 1.00$.*]

Which interval is preferable? Explain your reasoning.

[1 mark]

[END OF COURSEWORK QUESTIONS]