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

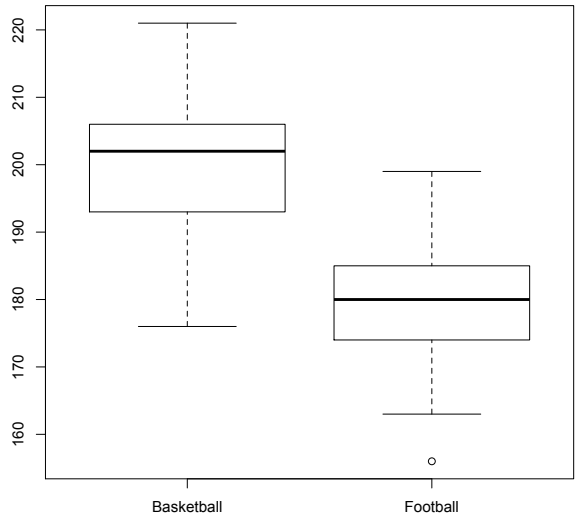
Question 1: Height of London Olympic games athletes

(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?

Model solution:

```
olympic <- read.table("olympic_height.txt")

summary(olympic) # allows us to see the variable names
# can alternatively be done by looking at the file

# as suggested in the hint, create an indicator vector
attach(olympic)
subset <- (sport=="Football"|sport=="Basketball")&sex == "M"

# as suggested in the hint, use the subset command
# (covered in Examples Class 3 - section on box plots)
olympicMFB <- subset(olympic, subset)

# to remove the other sports from the box plot, this command
# is needed. It was mentioned in Ex Class 3, section on box plots

olympicMFB <- droplevels(olympicMFB)

boxplot( height~ sport, data=olympicMFB)
```

This gives the following output:



Conclusion: on average, male basketball players are taller than male footballers. There is some possible indication of skewness in the heights of male basketball players. There is a single outlier among male footballers.

Marks (max 2):

- For this question be lenient as regards typos etc. if it is clear that it is just an error copying across from R (e.g. if the plots were correct).
- 1 mark for a completely correct plot

  0.5 marks for any of the following:
  - correctly reading in the data and creating a logical vector to extract the correct subset(s)
  - plotting a correct box plot without dropping unused levels
  - giving correct but separate plots for male Basketballers and Footballers (i.e. not together on same axes).
- If both males and females are included in their plot, the heights will be lower (check the axes): in this case 0 marks for the plot.
- 1 mark for at least two correct points from: basketballers are taller than footballers, commenting on symmetry/skewness, outliers. If only one point is made, only 0.5 marks.
- If no code given, do not award marks for the plot - only for interpretation.

ESSENTIALLY BOOKWORK. The only 'new' part is computing the logical vector, which is slightly more complicated than those seen previously.

Question 2: Confidence intervals for a binary proportion

(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 $n = 10$, $X = 5$, $\alpha = 0.05$.

Model solution:

```
# Wald interval
n <- 10
phat <- 5/10
alpha <- 0.05
z <- qnorm(1-alpha/2)
upper <- phat + z*sqrt(phat*(1-phat)/n)
lower <- phat - z*sqrt(phat*(1-phat)/n)
c(lower,upper)
 0.1901025 0.8098975
```

Hence $I(5) = [0.19, 0.81]$.

Marks (max 2):

- 2 marks for any correct solution featuring R code in terms of `phat`, `n` and `alpha`.
- 1.5 marks if $z = 1.96$ used instead of general `alpha`.
- Overall 1 mark for correct numerical answer only.
- If an incorrect z-value is used, award 0 overall.

BOOKWORK - I give R code to do essentially this in Ch 7 of the lecture slides. A very slight extension of the generality is required (to compute `z` for general `alpha`).

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

Solution:

```
> # adjusted Wald interval
>
> n <- 10 + 4
> phat <- (5+2)/(10+4)
```

```
> alpha <- 0.05
> z <- qnorm(1-alpha/2) # or use 1.96
> upper <- phat + z*sqrt(phat*(1-phat)/n)
> lower <- phat - z*sqrt(phat*(1-phat)/n)
> c(lower,upper)
[1] 0.2380888 0.7619112
```

The interval $[0.238, 0.762]$ is slightly narrower.

Marks (max 1):

- 0.5 marks for correctly adapting the R code they gave in the previous response. E.g. if the answer is only wrong because they used the wrong z-value in the previous question, then still award the mark.
- 0.5 marks for the correct comment, if it matches their interval (even if the interval itself is wrong).

UNSEEN - but only a minor extension of the previous part.

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

The completed lines are:

```
z <- qnorm(p=1-alpha/2) #these lines need to be filled in
upper  <- phat + z*sqrt(phat*(1-phat)/n)
lower  <- phat - z*sqrt(phat*(1-phat)/n)
k[i] <- (lower<=p)&(p<=upper)
bin.prob[i] <- choose(n,x) * (p^x) * (1-p)^(n-x)
```

or alternatively for the last line:

```
bin.prob[i] <- dbinom(x,size=n,prob=p)
```

Marks (max 2)

- For this part, do not be lenient with typos in the code that would prevent it working.
- 0.5 marks for correctly computing the lower and upper endpoints
- 0.5 marks for correctly computing k (allow <,> rather than <=, >=).
- 0.5 marks for correctly computing the binomial probability
- 0.5 for completely correct working function.
- If everything is correct except an incorrect z-value award 1 mark overall.
- If everything correct except $z = 1.96$ used, award 1.5 marks overall.

UNSEEN. Students have seen all of the requisite commands in computer classes in Weeks (logical vectors - Week 1 - and distribution functions - Week 5).

(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?

Solution:

```
> cover.prob.wald(n=10,p=0.5,alpha=0.05)
[1] 0.890625
```

This is substantially lower than the value 0.95 expected by theory. The reason is that the asymptotic approximation is not accurate (n is not large enough).

Marks (max 1):

- 0.5 marks for the plugging the correct inputs into the function from (c), even if the function is incorrect.
- 0.5 marks for correct interpretation (both points needed: anticipated value and reason).

UNSEEN - but students have seen how to use custom functions in the Week 7 Computer class, and have seen that the derivation of the CI depends on a large n approximation (in Chapter 7 of notes/slides).

(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)?

```
cover.prob.adj <- 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+2)/(n+4) # changed
 z <- qnorm(p=1-alpha/2)
 upper  <- phat + z*sqrt(phat*(1-phat)/(n+4)) # changed
 lower  <- phat - z*sqrt(phat*(1-phat)/(n+4)) # changed
 k[i] <- (lower<=p)&(p<=upper) # remains same
 bin.prob[i] <- choose(n,x) * (p^x) * (1-p)^(n-x)   # remains the same
  }
 C <-  sum(k*bin.prob)
 return(C)
}
```

```
> cover.prob.adj(n=10,p=0.5,alpha=0.05)
[1] 0.9785156
```

The coverage probability is closer to the expected 0.95.

Marks (max 1):

- 1 mark for completely correct solution.
- For this part, do not award carry through marks if e.g. z value is incorrect.

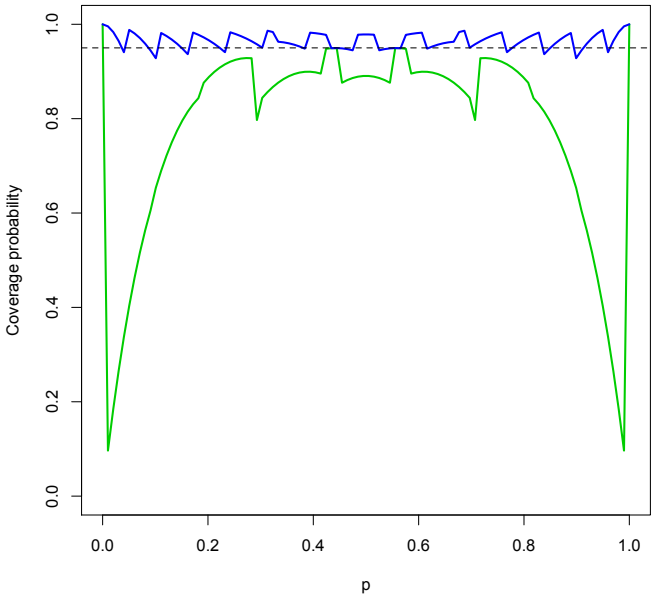
UNSEEN - minor extension of the earlier part.

(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.

Solution:

```
# final part: plots
cvg.wald <- NULL
cvg.adj <- NULL
ps <- seq(from=0,to=1,length=100)
for(i in 1:length(ps)) {
cvg.wald[i] <- cover.prob.wald(n=10,p=ps[i], alpha=0.05)
cvg.adj[i] <- cover.prob.adj(n=10,p=ps[i], alpha=0.05)
}

plot(ps, cvg.wald,type="l", xlab="p",
   ylab="Coverage probability",ylim=c(0,1),col=3,lwd=2)
lines(ps, cvg.adj,col=4,lwd=2)
abline(h=0.95,lty=2)
```



Interpretation: the adjusted interval is preferable because the coverage probability is closer to 0.95 across the range of values of p.

Marks (max 1):

- 0.5 marks for correctly plotting their functions, even if the functions are incorrect.
- 0.5 marks for the correct interpretation, which must also convincingly match their plot. It is hard to get this point if the function for the adjusted Wald coverage is incorrect.

UNSEEN. Students have seen how to do line plots of functions in the Week 3 and 5 computer classes.