Boxplots

15 marks

Boxplots. The values used to create a boxplot are based on an underlying Gaussian (or Normal) distribution. In this question, you will explore the choices of these values.

In `R` the function `qnorm(p)` returns the quantile (i.e. $z = Q(p)$) of a standard normal distribution that corresponds to the cumulative probability `p`.
Similarly, `pnorm(z)` returns the value of the cumulative distribution (i.e. $p = F(z)$) for a standard normal distribution at z.

a. **(1 mark)** Using these functions as appropriate, what is the interquartile range for standard normal?

```
qnorm(0.75)-qnorm(0.25)
```

```
## [1] 1.34898
```

The interquartile range for standard normal is 1.34898.

b. **(2 marks)** Recall the definition of the upper and lower fences for a box plot,

$$\text{upper fence} = Q3 + c \times IQR$$

$$\text{lower fence} = Q1 - c \times IQR$$

where $c = 1.5$. Applying these to the $N(0, 1)$ distribution, what would be the theoretical values of the lower and upper fences?

```
IQR <- qnorm(0.75) - qnorm(0.25)
upper <- qnorm(0.75) + 1.5*IQR
lower <- qnorm(0.25) - 1.5*IQR
c(upper , lower)
```

```
## [1]  2.697959 -2.697959
```

The theoretical values of the upper and lower fences are (2.697959, -2.697959).

c. **(2 marks)** Having just determined the numerical values of the theoretical upper and lower fences, determine the probability that a $N(0, 1)$ random variate, say Z, lies outside of one of these fences (i.e. **either larger** than the upper fence **or lower** than the lower fence)? That is, determine the numerical value of

$$p = Pr((Z < \text{lower fence}) \textbf{ or } (Z > \text{upper fence}))$$

```r
2*pnorm(lower)
```

```
## [1] 0.006976603
```

Therefore, $p = 0.006976603$.

d. **(3 marks)** Suppose that in the previous part of this question, you found the numerical value of p. In a sample of size n from $N(0,1)$, what is the expected number, m say, of values to lie outside the theoretical fences? What is the value of m when $n = 50$?

The expected value is m and the sample size is n. We have the equation $m = n * p$. Since p is from part c and $n = 50$, we have

```r
50*(2*pnorm(lower))
```

```
## [1] 0.3488302
```

Therefore, the expected value when $n = 50$ is 0.3488302.

e. For the standard boxplot c (the constant multiplier of the IQR) is taken to be $c = 1.5$. Suppose we wish to have c change with the size n of the sample.

   Recall from above that m is the expected number of values in a sample of size n which will lie outside the theoretical fences.

   i. **(2 marks)** Write down an expression for the number m as a function of c and n.

$$m = n * 2Pr(Z < Q1 - c \times IQR)$$

   ii. **(2 marks)** Using this expression, show how c can be written as a function of m and n.

$$c = \frac{Q1 - Q_Z(\frac{m}{2n})}{IQR}$$

   iii. **(3 marks)** Write a function `getc <- function(m, n) { ... }`, hand it in. Use your function to determine c when $m = 0.35$ for $n = 50, 100, 1000, 10000$.

```r
getc <- function(m, n){
  IQR <- qnorm(0.75)-qnorm(0.25)
  c <- (qnorm(0.25)-qnorm(m/(2*n)))/IQR
  return(c)
}
```

```r
getc(0.35, 50)
```

```
## [1] 1.499174
```

```r
getc(0.35, 100)
```

```
## [1] 1.66462
```

```r
getc(0.35, 1000)
```

```
## [1] 2.150278
```

```r
getc(0.35, 10000)
```

```
## [1] 2.567672
```