Introduction to Statistics
Assignment 5
Anusha Ramamurthy

1. Suppose that X is a continuous random variable with probability density function (pdf) f defined as follows:

$$f(x) = \begin{cases} 0, & x < 1 \\ 2(x-1), & 1 \le x \le 2 \\ 0, & x > 2 \end{cases}$$

Compute the following two quantities: q2 (X), the population median; and iqr (X), the population interquartile range.

Answer:

Since X is a continuous random variable with probability density function, in order to find the median, we have to find the cumulative density function for some m,

$$F(m) = F(q_2) = ½$$

q_2 is the second quartile

$$\int_1^m 2(x-1)\,.\,dx = 2\left[\frac{x^2}{2} - x\right]_1^m$$

$$2\left[\frac{x^2}{2} - x\right]_1^m = [x^2 - 2x]_1^m$$

$$[x^2 - 2x]_1^m = [(m^2 - 2m) - (1^2 - 2)]$$

$$[(m^2 - 2m) - (1^2 - 2)] = [m^2 - 2m + 1]$$

$$CDF = F(m) = [m^2 - 2m + 1]$$

$$Solving\ for\ m\ in\ [m^2 - 2m + 1] = 1/2$$

$$m^2 - 2m + 1/2 = 0$$

$$m = \frac{-(-2) \pm \sqrt{\left[(-2^2) - 4(1)\left(\frac{1}{2}\right)\right]}}{2} = \frac{2 \pm \sqrt{2}}{2} = 1 \pm \frac{1}{\sqrt{2}}$$

$$m = 1 \pm \frac{1}{\sqrt{2}}$$

Since $1 - \frac{1}{\sqrt{2}} = 0.707$ doesn't lie in the range $1 \le x \le 2$, we can conclude that $1 + \frac{1}{\sqrt{2}} = 1.707$ is the value of q_2 or median

$$q_2 = 1 + \frac{1}{\sqrt{2}}$$

b.

To find q_1 the first quartile, $F(q_1) = ¼$

$$\int_1^m 2(x-1)\,.\,dx = 2\left[\frac{x^2}{2} - x\right]_1^m$$

$$2\left[\frac{x^2}{2} - x\right]_1^m = [x^2 - 2x]_1^m$$

$$[x^2 - 2x]_1^m = [(m^2 - 2m) - (1^2 - 2)]$$

$$[(m^2 - 2m) - (1^2 - 2)] = [m^2 - 2m + 1]$$

$$CDF = F(m) = [m^2 - 2m + 1]$$

$$Solving\ for\ m\ in\ [m^2 - 2m + 1] = 1/4$$

$$m^2 - 2m + 3/4 = 0$$

$$m = \frac{-(-2) \pm \sqrt{\left[(-2^2) - 4(1)\left(\frac{3}{4}\right)\right]}}{2} = \frac{2 \pm 1}{2} = 1 \pm \frac{1}{2}$$

$$m = 1 \pm \frac{1}{2}$$

Since $1 + \frac{1}{2} = 1.5$ is very close to the median 1.7, we conclude that $1 - \frac{1}{2} = 0.5$ cannot be the value of q_1, as the area of the curve is present in the range $1 \leq x \leq 2$. Hence

$$q_1 = 1 + \frac{1}{2}$$

To find q_3, such that $F(q_3) = ¾$

$$\int_1^m 2(x-1)\,.\,dx = 2\left[\frac{x^2}{2} - x\right]_1^m$$

$$2\left[\frac{x^2}{2} - x\right]_1^m = [x^2 - 2x]_1^m$$

$$[x^2 - 2x]_1^m = [(m^2 - 2m) - (1^2 - 2)]$$

$$[(m^2 - 2m) - (1^2 - 2)] = [m^2 - 2m + 1]$$

$$CDF = F(m) = [m^2 - 2m + 1]$$

$$Solving\ for\ m\ in\ [m^2 - 2m + 1] = 3/4$$

$$m^2 - 2m + 1/4 = 0$$

$$m = \frac{-(-2) \pm \sqrt{\left[(-2^2) - 4(1)\left(\frac{1}{4}\right)\right]}}{2} = \frac{2 \pm \sqrt{3}}{2}$$

$$m = 1 \pm \frac{\sqrt{3}}{2}$$

Since $1 - \frac{\sqrt{3}}{2} = 0.1339$, which doesn't not lie in the range $1 \leq x \leq 2$, we can conclude that $1 + \frac{\sqrt{3}}{2} = 1.866$

$$q_3 = 1 + \frac{\sqrt{3}}{2}$$

Inter quartile range =

$$q_3 - q_1 = \left(1 + \frac{\sqrt{3}}{2}\right) - \left(1 + \frac{1}{2}\right) = 0.366$$

2. Consider the function g: $\mathfrak{R} \to \mathfrak{R}$ defined by let f(x) = cg(x), where c is an undetermined constant.

$$g(x) = \begin{cases} 0, & x < 0 \\ x, & x \in [0,1] \\ 1, & x \in [1,2] \\ 3 - x, & x \in [2,3] \\ 0, & x > 3 \end{cases}$$

a. For what value of c is f a probability density function?
answer:

$$f(x) = \int cg(x)\,dx = c\left[\int_0^1 x\,.dx + \int_1^2 1\,.dx + \int_2^3 3 - x\,.dx\right]$$

$$c\left[\int_0^1 x\,.dx + \int_1^2 1\,.dx + \int_2^3 3 - x\,.dx\right] = c\left[\left[\frac{x^2}{2}\right]_0^1 + [x]_1^2 + \left[3x - \frac{x^2}{2}\right]_2^3\right]$$

$$c\left[\left[\frac{x^2}{2}\right]_0^1 + [x]_1^2 + \left[3x - \frac{x^2}{2}\right]_2^3\right] = c\left(\frac{1}{2} + (2-1) + \left[9 - \frac{9}{2} - 6 + \frac{4}{2}\right]\right)$$

$$c[-2 + 4] = 1$$

$$c = \frac{1}{2}$$

b. Suppose that a continuous random variable X has probability density function f. Compute P(1.5 < X < 2.5).
Answer:

$$\mathrm{P}(1.5 < \mathrm{X} < 2.5) = c\left[\int_{1.5}^2 1\,.dx + \int_2^{2.5} 3 - x\,.dx\right] = \frac{1}{2}\left[[x]_{1.5}^2 + \left[3x - \frac{x^2}{2}\right]_2^{2.5}\right]$$

$$\frac{1}{2}\left[0.5 + 7.5 - \frac{6.25}{2} - 6 + 2\right] = 0.4375$$

c. Compute EX.

Answer:

$$EX = c\int x.g(x)\,.dx = c\left[\int_0^1 x^2\,.dx + \int_1^2 x\,.dx + \int_2^3 (3x - x^2)\,.dx\right]$$

$$c\left[\int_0^1 x^2\,.dx + \int_1^2 x\,.dx + \int_2^3 (3x - x^2)\,.dx\right] = c\left(\left[\frac{x^3}{3}\right]_0^1 + \left[\frac{x^2}{2}\right]_1^2 + \left[\frac{3x^2}{2} - \frac{x^3}{3}\right]_2^3\right)$$

$$\frac{1}{2}\left(\left[\frac{x^3}{3}\right]_0^1 + \left[\frac{x^2}{2}\right]_1^2 + \left[\frac{3x^2}{2} - \frac{x^3}{3}\right]_2^3\right) = \frac{1}{2}\left(\frac{1}{3} + \frac{3}{2} + \frac{15}{2} + \frac{8}{3} - 9\right) = \frac{3}{2}$$

$$EX = 1.5$$

d. Let F denote the cumulative distribution function of X. Compute F(1).

Answer:

$$F(1) = P(X \le 1) = c\left[\int_0^1 x\,.dx\right]$$

$$c\left[\int_0^1 x\,.dx\right] = \frac{1}{2}\left[\frac{x^2}{2}\right]_0^1 = \frac{1}{4}$$

$$F(1) = \frac{1}{4}$$

e. Determine the 0.90 quartile of f.

Answer:

$$\text{F(q)} = 0.9$$

$$\int_2^q 3 - x\,.dx = 0.3$$

$$\left[3x - \frac{x^2}{2}\right]_2^q = 0.3$$

$$3 * q - \frac{q^2}{2} - 6 + 2 = 0.3$$

$$3q - \frac{q^2}{2} - 4 = 0.3$$

$$3q - \frac{q^2}{2} - 4 - 0.3 = 0$$

$$3q - \frac{q^2}{2} - 4.3 = 0$$

$$-3q + \frac{q^2}{2} + 4.3 = 0$$

$$\frac{q^2}{2} - 3q + 4.3 = 0$$

$$m = \frac{-(-3) \pm \sqrt{[(-3^2) - 4(1/2)(4.3)]}}{1} = 2.36$$

Therefore 0.90 quartile of f is 2.36

3. A random variable X is Uniform(5, 15) has population mean EX = 10 and population variance VarX = 100/12. Let Y denote a normal random variable with the same mean and variance.

a. Consider X. What is the ratio of its interquartile range to its standard deviation?

```
set.seed(3000)
uniform_x <- sort(runif(120,min=5,max=15))
q <- as.vector(quantile(uniform_x,probs=c(.25,.75)))
iqr_uniform = q[2]-q[1]
ratio_uniform = iqr_uniform /sqrt(100/12)
ratio_uniform

## [1] 1.641318
```

b. Consider Y . What is the ratio of its interquartile range to its standard deviation?

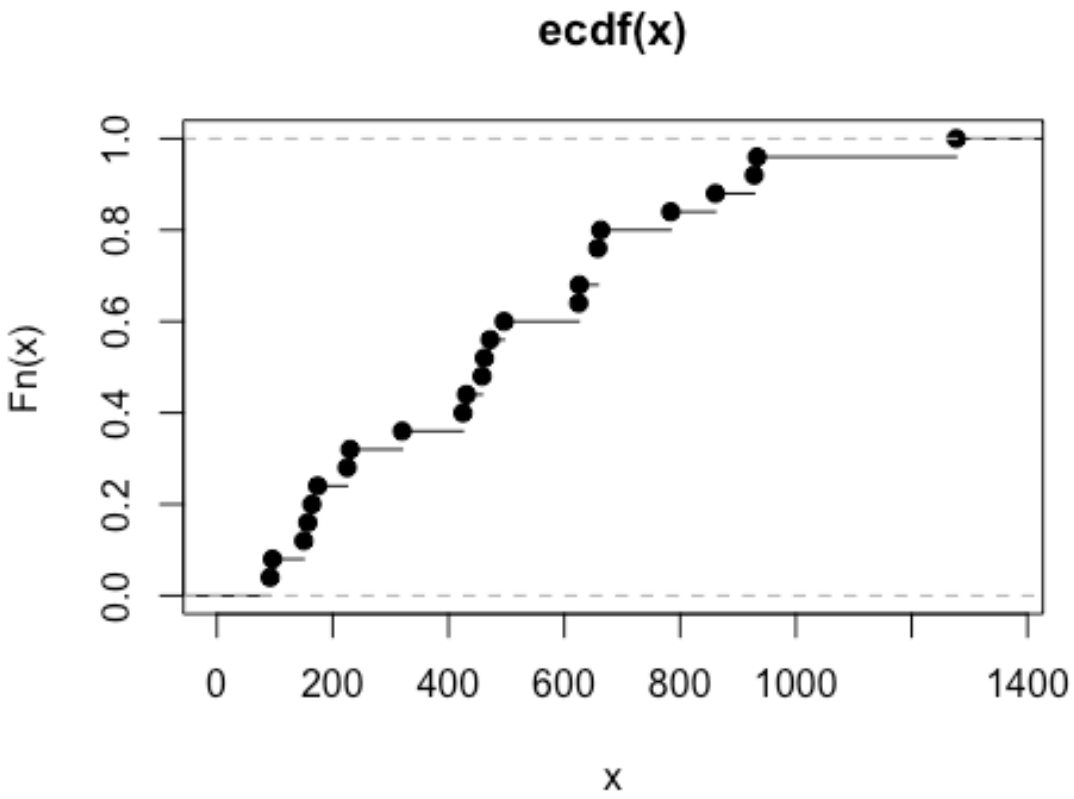
```
set.seed(3000)
norm_x <- sort(rnorm(120,mean=10,sd=sqrt(100/12)))
q <- as.vector(quantile(norm_x,probs=c(.25,.75)))
iqr_norm = q[2]-q[1]
ratio_norm = iqr_norm /sqrt(100/12)
ratio_norm

## [1] 1.29859
```

4. Let x denote the sample for an unknown population

a. Graph the ecdf of x

```
x = c (462,425,164,784,625,472,658,658,663,928,92,230,96,626,1277,225,150,
    320,496,157,458,933,861,174,431)
plot.ecdf(x)
```



b. Compute the plug-in estimates of population mean and variance

```
plug.mean <- mean(x)
plug.mean

## [1] 494.6

plug.variance <- mean(x^2)- plug.mean^2
plug.variance

## [1] 91078.72
```

c. Compute the plug-in estimates of population median and interquartile range

```
median(x)

## [1] 462

q <- as.vector(quantile(x,probs=c(.25,.75)))
iqr <- q[2]-q[1]
iqr

## [1] 433
```

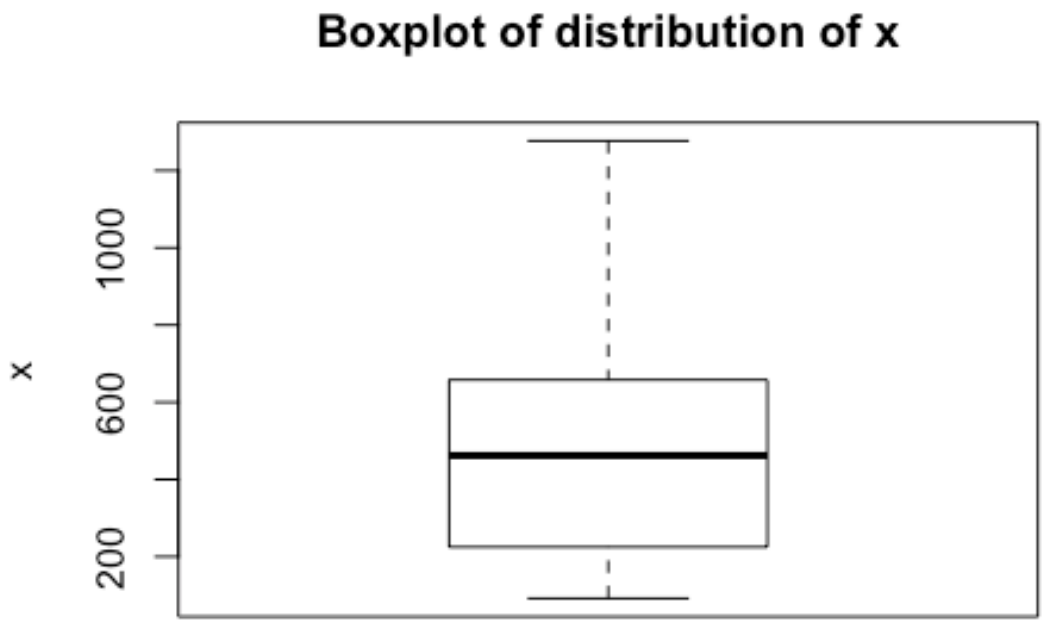
d. Compute the ratio of plug-in estimate of interquartile range to the square root of plug in estimate of variance

```
ratio_iqr_to_sd <- iqr / sqrt(plug.variance)
ratio_iqr_to_sd
```

```
## [1] 1.434761
```

e. Construct the boxplot

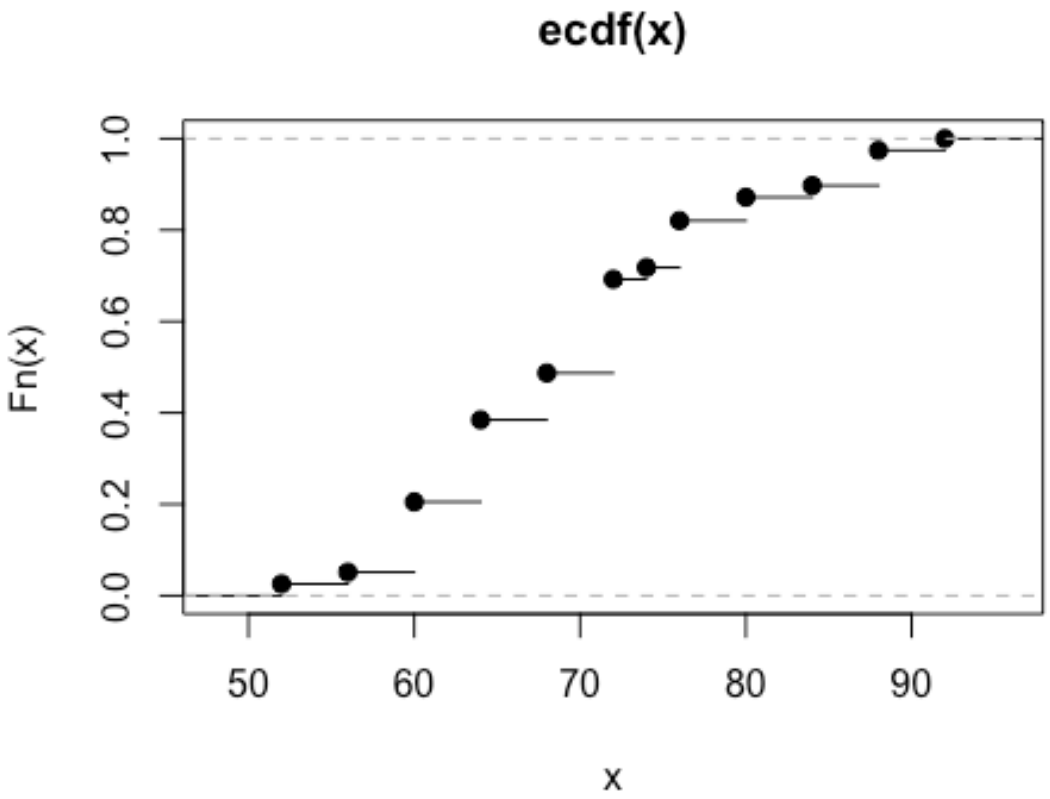
```
boxplot(x,range=0,
main="Boxplot of distribution of x",
ylab="x")
```



5. Let x denote the sample for pulse rates of the Peruvian Indians

a. Graph the ecdf of x

```
x = c (88,76,84,64,60,64,60,64,68,74,68,68,72,76,72,52,72,64,60,56,72,88,80,76,64,72,60,76,88,72,64,60,60
,72,92,80,72,64,68)
plot.ecdf(x)
```



b. Compute the plug-in estimates of population mean and variance

```
plug.mean <- mean(x)
plug.mean

## [1] 70.30769

plug.variance <- mean(x^2)- plug.mean^2
plug.variance

## [1] 87.90533
```

c. Compute the plug-in estimates of population median and interquartile range

```
median(x)

## [1] 72

q <- as.vector(quantile(x,probs=c(.25,.75)))
iqr <- q[2]-q[1]
iqr

## [1] 12
```

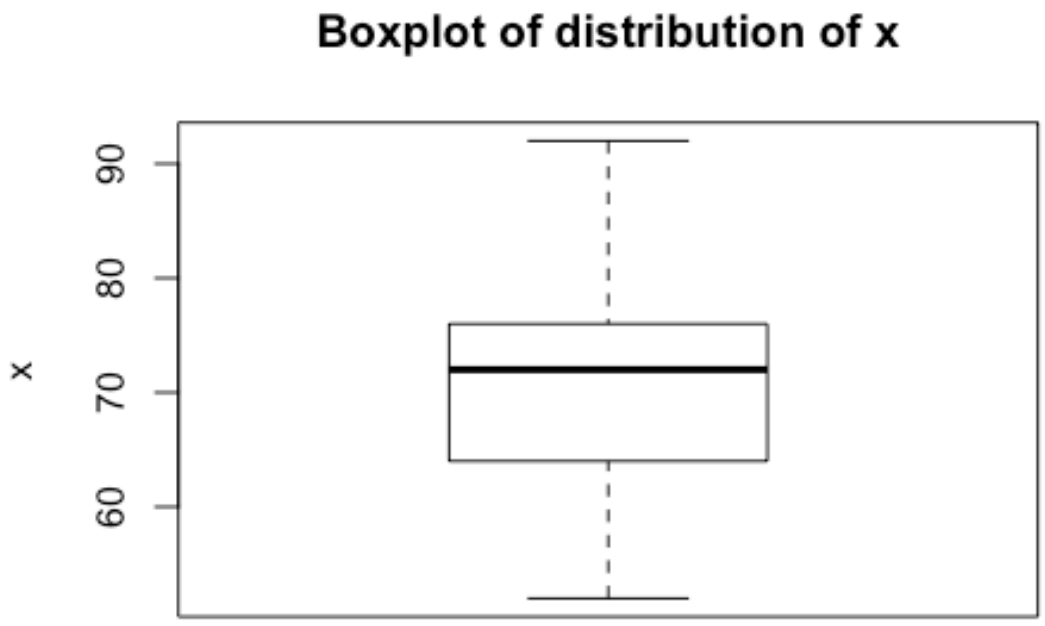
d. Compute the ratio of plug-in estimate of interquartile range to the square root of plug in estimate of variance

```
ratio_iqr_to_sd <- iqr / sqrt(plug.variance)
ratio_iqr_to_sd
```

```
## [1] 1.279893
```

e. Construct the boxplot

```
boxplot(x,range=0,
main="Boxplot of distribution of x",
ylab="x")
```



6. In Major League Baseball, there are 15 teams in the American League, and 15 teams in the National League. The standings for the 2015 season can be found at http://mlb.mlb.com/mlb/standings/#20151004 The W column gives the number of wins for each team.

a. Using R, draw two boxplots side-by-side on the same graph: one for the wins of American League teams in 2015, and one for the wins of National League teams in 2015. Your graph should be labeled clearly, so that a statistician should be able to interpret the graph without any further explanation.

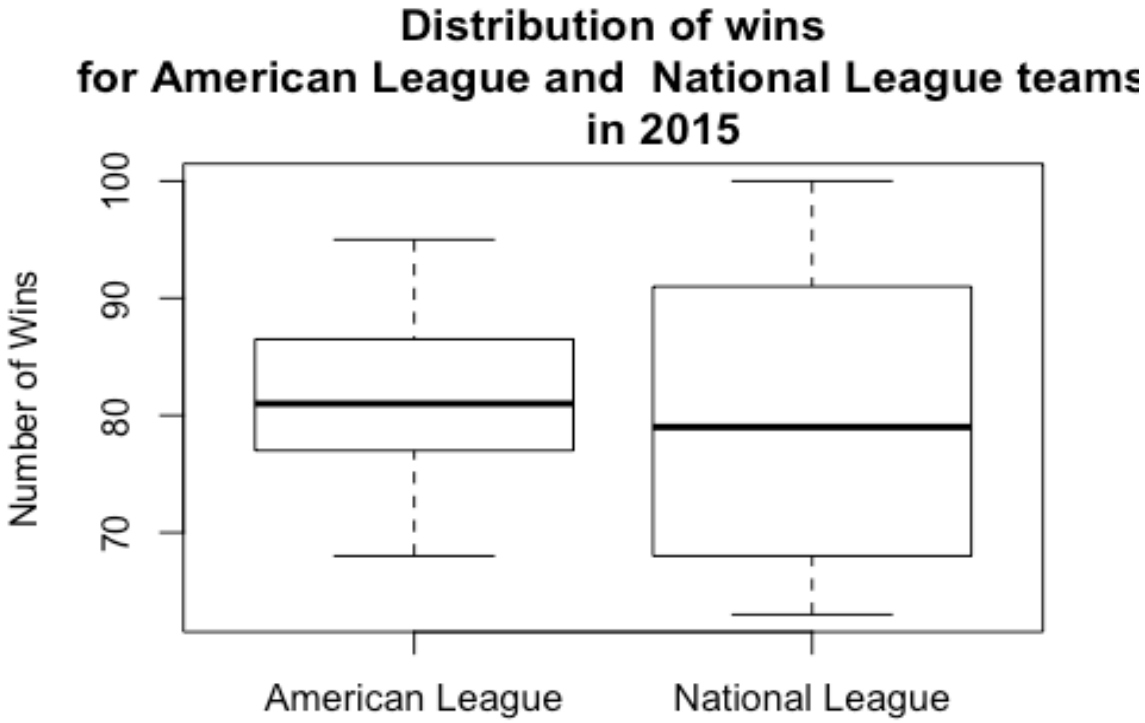
```
american_league = c(93,87,81,80,78,95,83,81,76,74,88,86,85,76,68)
national_league = c(90,83,71,67,63,100,98,97,68,64,92,84,79,74,68)
summary(american_league)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.00   77.00   81.00   82.07   86.50   95.00
```

```
summary(national_league)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   63.00   68.00   79.00   79.87   91.00  100.00

boxplot(american_league, national_league, range=0,
      main="Distribution of wins
 for American League and  National League teams
      in 2015",
      ylab="Number of Wins",
      names=c('American League','National League'))
```



b. Describe (in sentences) the differences between the distributions of wins for American League and National League teams in 2015.

We can observe that each of the teams have played 162 matches

1. The National League has the team(s) that have had the highest as well as the lowest number of wins in the season or the year 2015
2. The interquartile range is larger in the case of National League Teams when compared to the interquartile range of American League teams
3. The average number of wins of the American League teams is higher than the National League Teams
4. The spread of values is more in the case of National League Teams
5. While the American League teams have number of wins very close to each other