Homework 5 solutions

S520

Due at the beginning of class, Thursday 18th February

Trosset question numbers refer to the hardcover textbook. Show all working and give R code where appropriate.

For this homework, you may either upload your answers to Canvas or submit on paper in class. If you upload to Canvas, your answers must be typed and in PDF format.

1. *Trosset exercise 6.4.1*

   For $1 < y < 2$, the cdf $F(y)$ is the area of a triangle with base $y-1$ and height $f(y) = 2(y-1)$:

$$F(y) = \begin{cases} 0 & y < 1 \\ (y-1)^2 & 1 \le y < 2 \\ 1 & y \ge 2. \end{cases}$$

   To find the α-quantile, set $F(q_\alpha) = \alpha$:

$$\begin{aligned} (q_\alpha - 1)^2 &= \alpha \\ q_\alpha - 1 &= \sqrt{\alpha} \\ q_\alpha &= 1 + \sqrt{\alpha}. \end{aligned}$$

   To find the median, substitute $\alpha = 0.5$, giving $q_2 = 1 + \sqrt{0.5} \approx 1.71$.

   To find the IQR, substitute $\alpha = 0.25$ and 0.75 to find the quartiles, then take the difference:

$$IQR = (1 + \sqrt{0.75}) - (1 + \sqrt{0.25}) = \sqrt{0.75} - 0.5 \approx 0.37.$$

2. *Trosset exercise 6.4.2*

   (a) By geometry, the area under $g(x) = 2$. We require the area under $f(x)$ to be 1, so $c = 1/2$.

   (b) The probability is the area between $x = 1.5$ and $x = 2.5$. By symmetry, the area between $x = 1.5$ and $x = 3$ is $1/2$.

   The area between $x = 2.5$ and $x = 3$ is a triangle with base $1/2$ and height $f(2.5) = 1/4$, giving area $1/16$.

   The area between $x = 1.5$ and $x = 2.5$ is thus $1/2 - 1/6 = 7/16$. This is the probability we want.

   (c) The pdf is symmetric about $x = 1.5$, so $EX = 1.5$.

(d) This is the area between $x = 0$ and $x = 1$. This is a triangle with base 1 and height $f(1) = 1/2$, giving area 1.4.

(e) We need to find q such that the area to the left of the vertical line at q is 0.9, and the area to the right is 0.1. By inspection of the pdf, this will happen for some x-value between 2 and 3. The area to the right is a triangle with base $3 - q$ and height $f(q) = (3 - q)/2$. Setting this area equal to 0.1 gives:

$$\begin{aligned} 0.1 &= \frac{1}{4}(3-q)^2 \\ 0.4 &= (3-q)^2 \\ \sqrt{0.4} &= 3-q \\ q &= 3 - \sqrt{0.4} \approx 2.37. \end{aligned}$$

3. Trosset exercise 6.4.6

(a) X is uniformly distributed in the range $(5, 15)$, so its quartiles are 7.5, 10, and 12.5. The ratio of interquartile range to standard deviation is

$$\frac{12.5 - 7.5}{\sqrt{25/3}} = \frac{5}{5/\sqrt{3}} = \sqrt{3} \approx 1.73.$$

(b) Using R,

```
> sd = 5/sqrt(3)
> iqr = qnorm(.75, 10, sd) - qnorm(.25, 10, sd)
> iqr/sd
[1] 1.34898
```

(See also section 6.3 of Trosset.)

4. Trosset exercise 7.7.1 parts (a)–(e)

Here's some R code for the question.

```
x = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/sample771.dat")
plot.ecdf(x)
mean(x)
# Plug-in variance
mean(x^2) - mean(x)^2
median(x)
IQR = quantile(x, 0.75) - quantile(x, 0.25)
IQR / sqrt(mean(x^2) - mean(x)^2)
boxplot(x, main="Boxplot for Trosset exercise 7.7.1", ylab="x")
```

(a) See Figure 1.

(b) Mean is 494.6, plug-in variance is 91079, or 94974 if you use `var(x)` (which technically is not a plug-in estimate.)

(c) Median is 462, IQR is $658 - 225 = 433$

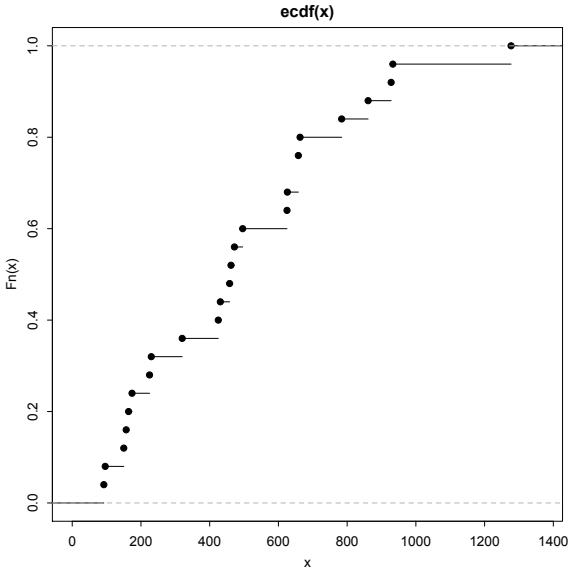

Figure 1: ECDF plot for Trosset exercise 7.7.1.

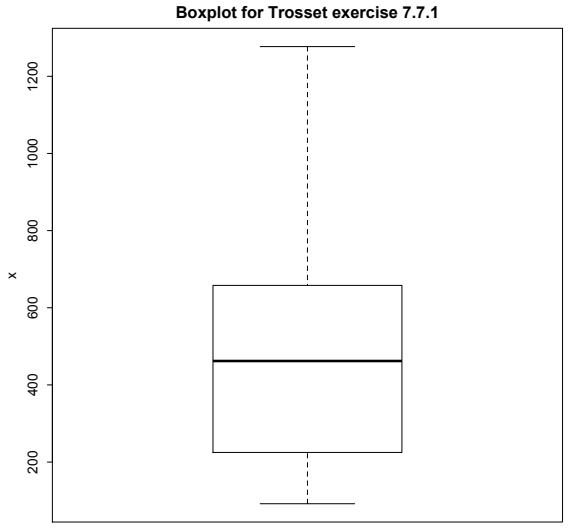

Figure 2: Boxplot for Trosset exercise 7.7.1.

(d) The ratio is 1.43 (or 1.41 if you use the `sd` function.)

(e) See Figure 2.

5. Trosset exercise 7.7.2 parts (a)–(e) Here's R code to produce the required graphs and and estimates.

```
pulses  = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/pulses.dat")
plot(ecdf(pulses), main="ECDF plot")
summary(pulses)
mean(pulses^2) - mean(pulses)^2 # Plug-in variance
IQR = quantile(pulses, 0.75) - quantile(pulses, 0.25)
IQR / sqrt(mean(pulses^2) - mean(pulses)^2) # IQR/SD
boxplot(pulses, main="Peruvian pulse rate boxplot")
# Not required:
qqnorm(pulses)
plot(density(pulses), main="Density plot")
```

The plug-in mean is 70.3 and the plug-in variance 87.9 (the sample variance, which is 90.2, is also acceptable). The median is 72 and the IQR. The ratio of the IQR to the plug-in SD is 1.28 (1.26 if you use the sample SD), about the same as the normal.

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.*

Since there are only two sets of 15 numbers, we can just enter the data manually.

```
AmericanLeague = c(93, 87, 81, 80, 78,
  95, 83, 81, 76, 74,
  88, 86, 85, 76, 68)
NationalLeague = c(90, 83, 71, 67, 63,
  100, 98, 97, 68, 64,
  92, 84, 79, 74, 68)
boxplot(AmericanLeague, NationalLeague,
  main="Major League Baseball wins, 2015",
  ylab="Wins",
  names=c("American League", "National League"))
dev.print(pdf, "MLBwins2015.pdf")
```

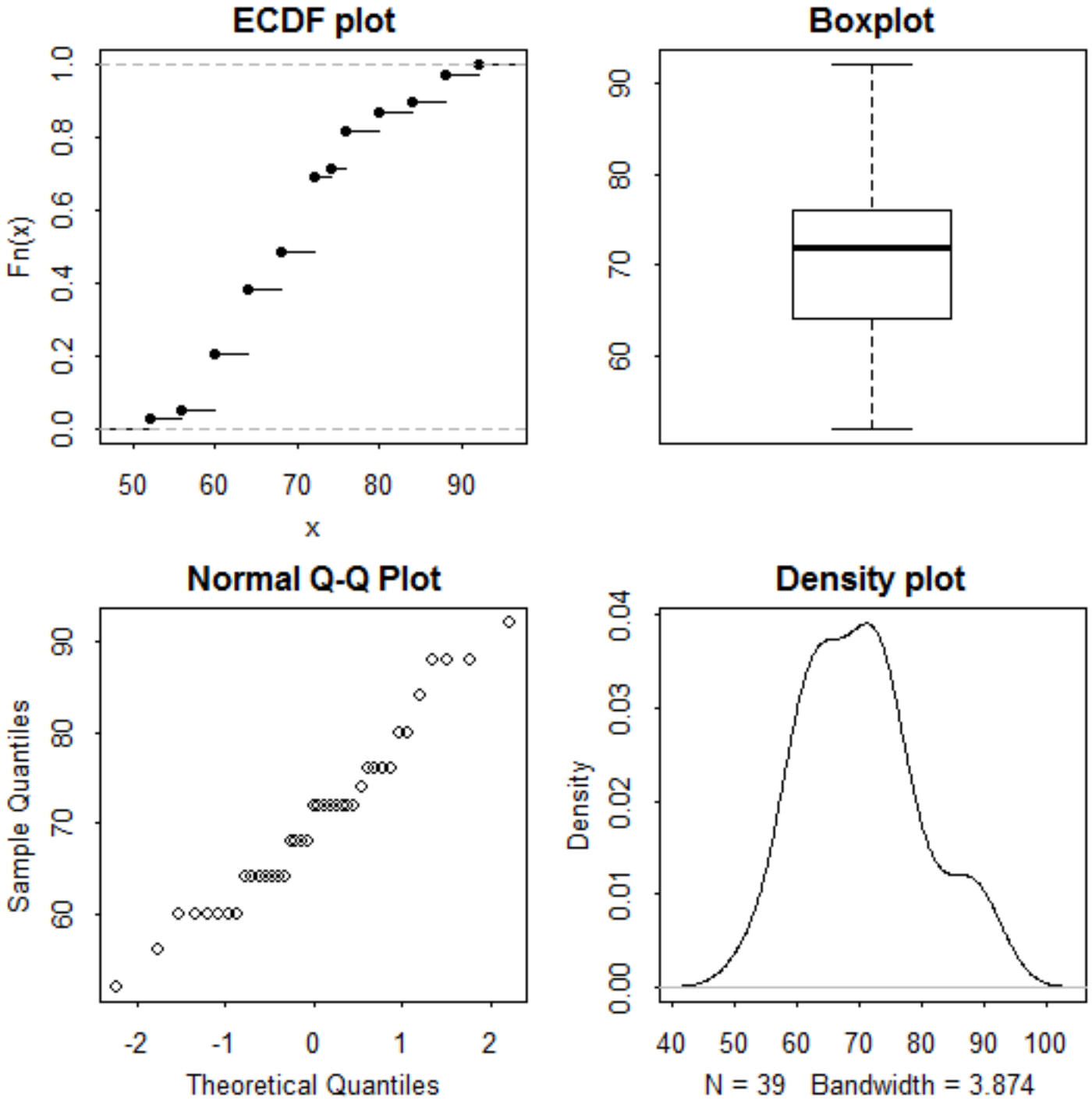

Figure 3: Plots of the distribution of Peruvian Indian pulse rates. (The QQ plot and density plot are not required but are shown for completeness.)

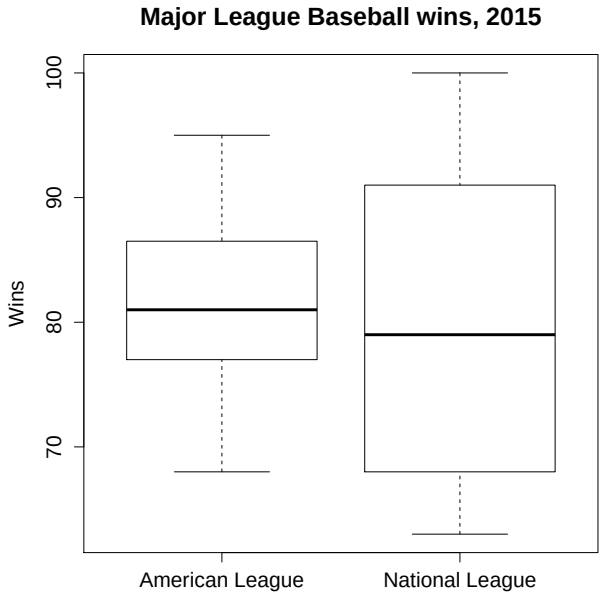

Figure 4: Boxplots of wins by American League and National League baseball teams in 2015.

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

The American League did a little better on average, with its mean and median a couple of wins higher than the National League. Perhaps more notably, the spread of the National League team wins is much bigger: their IQR is 23 and SD is 13, compared to the American League's 9.5 and 7 respectively. Further inspection (e.g. using QQ plots) shows that while the American League distribution is close to normal, the National League distribution is much less so (it's actually fairly uniform between the lowest value of 63 and the highest value of 100.)