Homework 6 solutions

S320/520

Due by 4pm, Thursday 15th October

Please write "S320" or "S520" at the top of your homework. Trosset question numbers refer to the hardcover textbook. Include R code as an appendix to your answers. Data for the questions from chapter 7 can be found at

`http://mypage.iu.edu/~mtrosset/StatInfeR.html`

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

1. (Trosset exercise 7.7.4.)

```
data = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/sample774.dat")
plot(ecdf(data), main="ECDF of data")
summary(data)
mean(data^2) - mean(data)^2 # Plug-in variance
sort(data)
(1.464+2.063)/2 - (.434+.530)/2 # Plug-in IQR
1.2815 / sqrt(mean(data^2) - mean(data)^2) # IQR/SD
qqnorm(data)
plot(density(log(data)), main="Density of log data")
qqnorm(log(data))
```

   Plug-in estimates are mean 1.49, variance 2.79 (2.93 for sample variance), median 1.076, IQR 1.28 (1.11 if you use R's default method for quartiles). The ratio of IQR to SD is somewhere from 0.65 to 0.77 depending on which methods you use; in any, it's far from the 1.35 you expect from a normal distribution. The QQ plot bends upwards, which confirms the distribution isn't normal. On the other hand, the QQ plot of the logged data is very close to a straight line, so the logged data is approximately normal.

2. (Trosset exercise 7.7.6.)

```
scores = scan("http://mypage.iu.edu/~mtrosset/StatInfeR/Data/test351.dat")
qqnorm(scores)
plot(density(scores), main="Density plot of Math 351 scores")
```

   (a) The QQ plot is curved, so the data is non-normal.

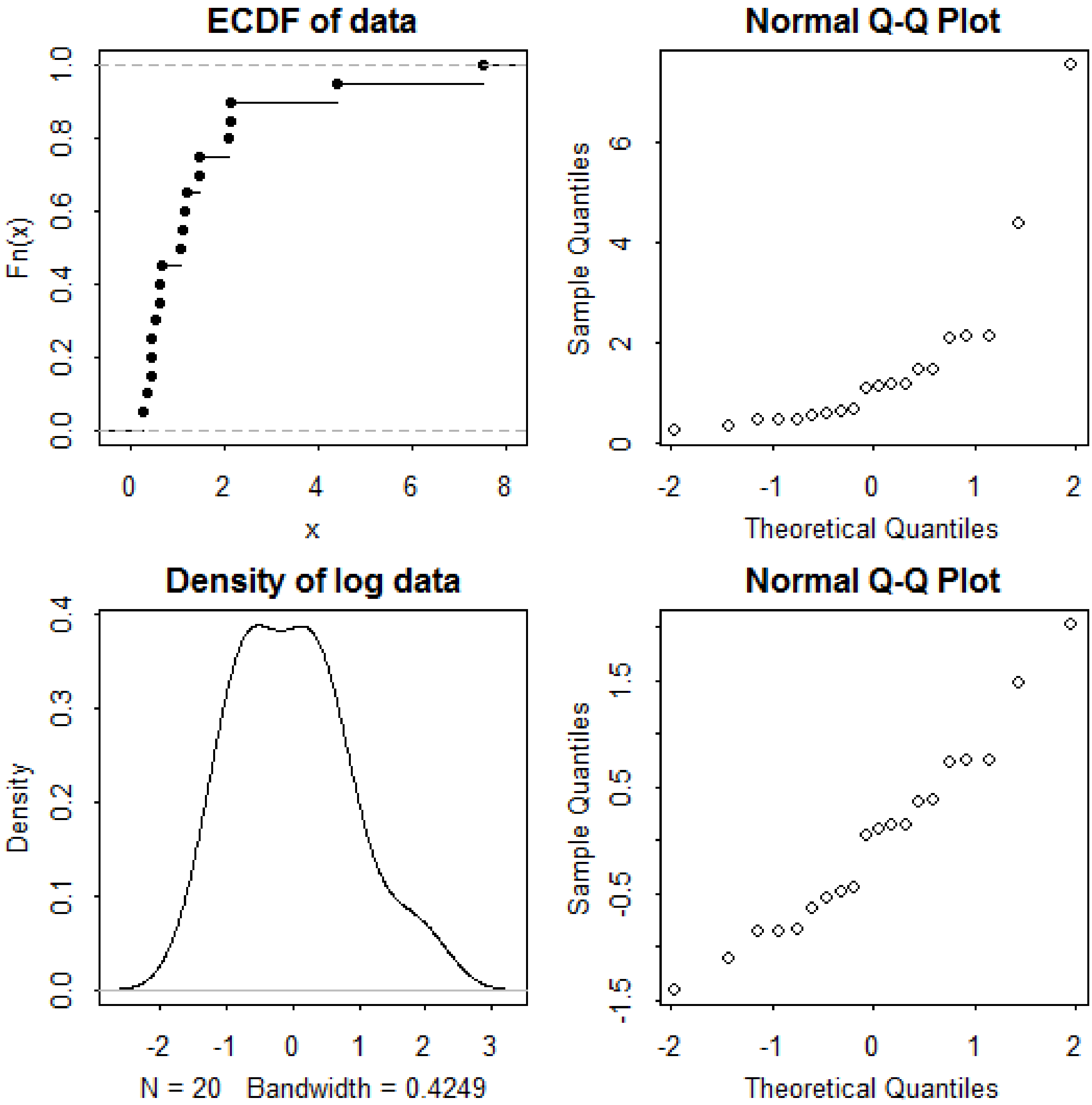

Figure 1: Plots of data for Trosset ex. 7.7.4. Top two are for raw data, bottom two are for logged data.

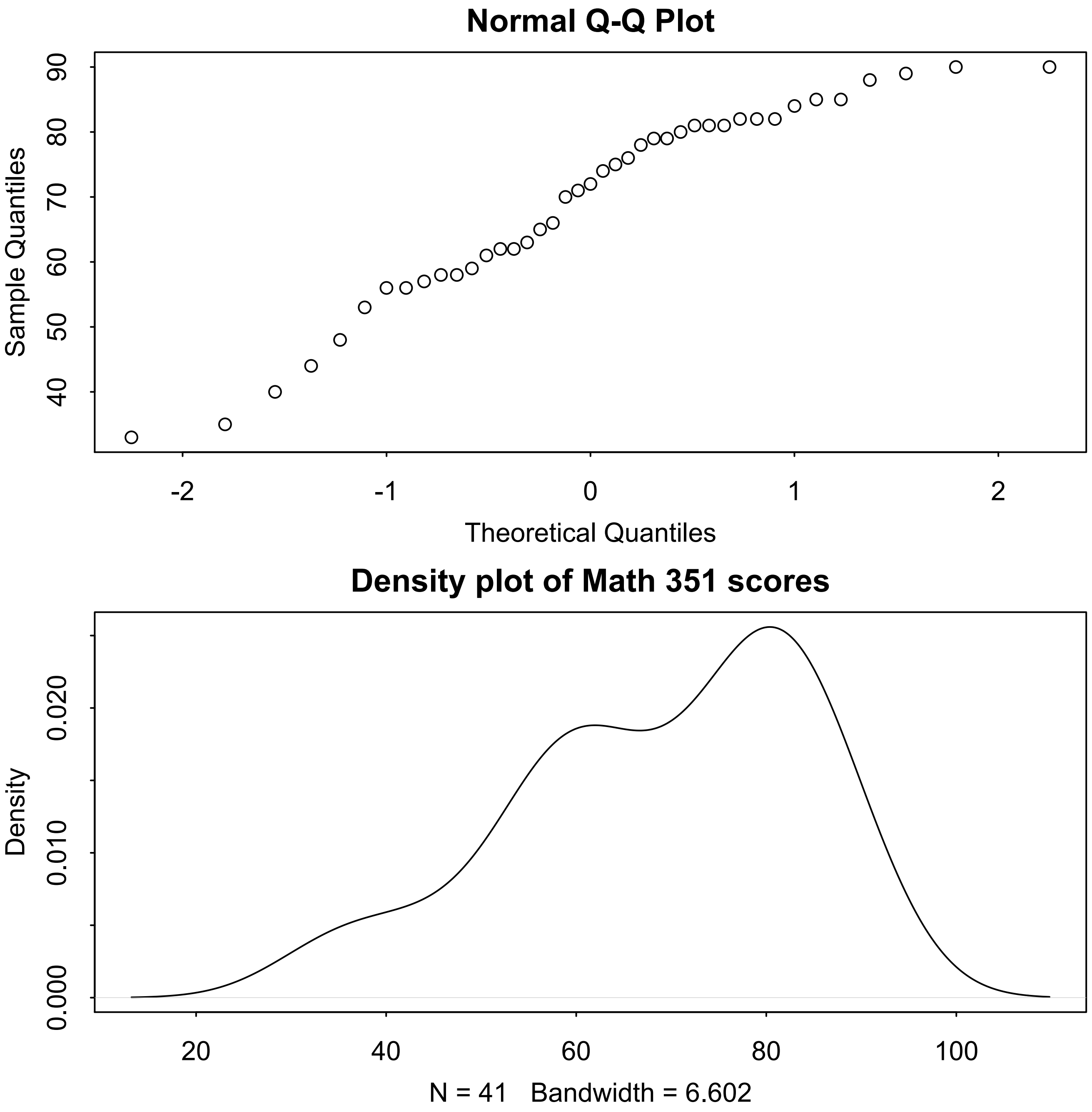

Figure 2: Math 351 test scores.

(b) A density plot shows the data is left-skewed, which is comparatively uncommon in real data.

3. (Trosset exercise 8.4.4.) The distribution of the life of one battery (in hours) has mean 5, standard deviation 0.5, and variance 0.25. So by the Central Limit Theorem, the distribution of the life of twenty batteries (in hours) is approximately normal with mean $20 \times 5 = 100$, variance $20 \times 0.25 = 5$, and standard deviation $\sqrt{5}$. The probability the batteries last at least 105 hours is thus

```
> 1 - pnorm(105, 100, sqrt(5))
[1] 0.01267366
```

It's a bit over 1% — not that likely.

4. *In the Powerball lottery, there are 59 white balls, numbered 1 to 59. Each week, five of the white balls are drawn, without replacement. In the past, the most frequently occurring white ball has been 23.*

(a) *In the next lottery, will the probability of drawing the number 23 be greater than* 5/59*, less than* 5/59*, or equal to* 5/59*?*

5/59. The Law of Large Numbers has no memory.

Before one season, the Oakland A's were considered to be an average major league baseball team, predicted to win half (81) of their 162 games. They win the first six games of the season.

(b) *True or false: After the first six games, the best prediction of the number of games the Oakland A's win that season is 162 out of 162.*

False. No Major League Baseball team has ever won more than 76% of their games over a season. Six games is too small a sample to make such a strong extrapolation.

(c) *True or false: After the first six games, the best prediction of the number of games the Oakland A's win that season is still 81 out of 162.*

False. They've already won six, so if they win half their remaining games, they would win 84 out of 162. Plus since they've won six out of six, we should entertain the possibility they're better than average.

(d) *I survey a simple random sample of 1000 U.S. households and find out their income. True or false: By the Central Limit Theorem, the incomes in the population will have an approximately normal distribution.*

False. The Central Limit Theorem only says the sample mean will be normal. It doesn't say anything about the population, which is likely to be right-skewed.

(e) *True or false: By the Central Limit Theorem, the incomes in the sample will have an approximately normal distribution.*

False. The sample will probably look like the population, i.e. right-skewed.

5. *Let X be a discrete random variable with probability mass function*

$$P(X = x) = \begin{cases} 0.3 & x = -2 \\ 0.6 & x = -1 \\ 0.1 & x = 12 \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_1, \ldots, X_n$ be an iid sequence of random variables with the same distribution as X. Let $\bar{X}$ be the sample mean (of $X_1, \ldots, X_n$.)

(a) *Find EX.*

$EX = (-2 \times 0.3) + (-1 \times 0.6) + (12 \times 0.1) = 0$.

(b) *Find Var(X).*

$$\begin{aligned} EX^2 &= (4 \times 0.3) + (1 \times 0.6) + (144 \times 0.1) = 16.2 \\ \operatorname{Var} X &= EX^2 - (EX)^2 = 16.2 \end{aligned}$$

(c) *What is the expected value of $\bar{X}$?*

This is the same as EX, i.e. zero.

(d) *What is the variance of $\bar{X}$? (Note: This will depend on n.)*

$\sigma^2/n = 16.2/n$.

(e) *Suppose $n = 100$. Use the R function `pnorm()` to find the approximate probability that $\bar{X}$ is greater than 0.5.*

`1 - pnorm(0.5, mean=0, sd=sqrt(16.2/100))` gives 0.107.

6. *(Extra credit for everyone.) I want to find out the average number of people per household in the U.S. I survey a simple random sample of U.S. households and obtain the results displayed in the following table.*

Household size	Number of households
1	27
2	34
3	16
4	13
5	6
6	3
7	1

(a) *Lacking any other information, our best estimate for the population mean household size is the sample mean. What is the sample mean of our data?*

```
households = c(rep(1,27), rep(2,34), rep(3,16), rep(4, 13), rep(5, 6), rep(6,3), 7)
mean(households)
```

This gives a sample mean of 2.5.

(b) *What is our estimate for the standard deviation of household sizes?*

`sd(households)` gives 1.41. (Alternatively, the square root of the plug-in variance is 1.40.)

(c) *What is the estimated standard error of the sample mean? (That is, based on our answer to (b), what is our estimate for the standard deviation of the distribution of the sample mean?)*

$1.41/\sqrt{100} = 0.141$.

(d) *Our error is the difference between the sample mean and the population mean. Using the normal distribution, find the approximate probability that the absolute value of the error in a survey of this form and size is less than 0.5.*

```
pnorm(0.5, 0, sd(households)/sqrt(100)) - pnorm(-0.5, 0, sd(households)/sqrt(100))
```

gives a probability of 99.96%.

(e) *Can we be reasonably sure that the average household size for all U.S. households is between 2 and 3?*

Yes. In well over 99% of similar surveys, the error will be less than 0.5, so we can be confident that the error in this survey is less than 0.5.