S520 Instructor's Solutions

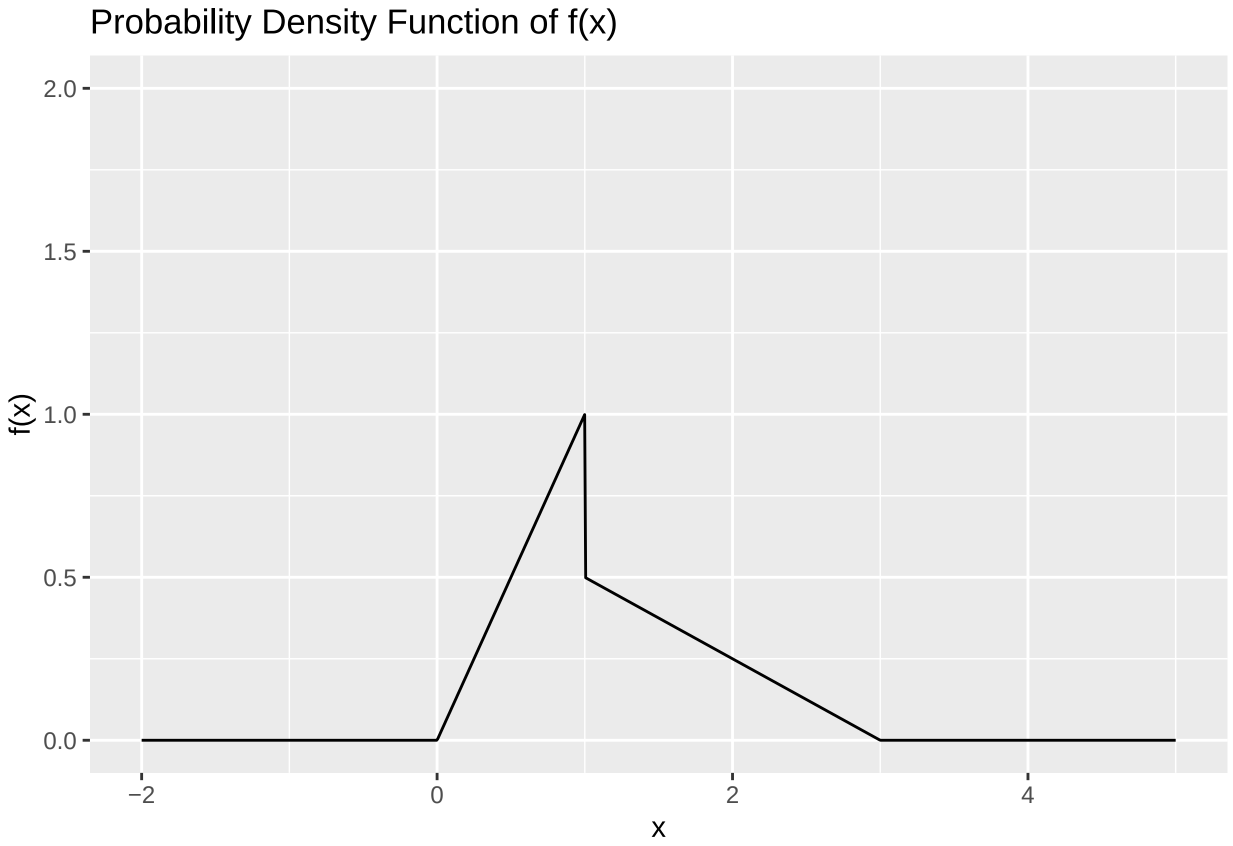
Spring 2023 STAT-S 520

2023-02-21

1

1a.

Here is a plot of the PDF of X



Using the above graph, the area on either side of the median should be equal to 0.5. Calculating the area under the graph between 0 and 1, we get $0.5 * 1 * 1 = 0.5$
Therefore, $q_2(X) = 1$.

Using integrals, for the region of the range $(0, 1)$

$$P(X \leq q) = \int_0^q x dx = \frac{x^2}{2}\bigg|_0^q = \frac{q^2}{2}$$

so for $q = 1$, $P(X \leq 1) = 1/2$, and $q_2(X) = 1$.

1b.

Method 1: Using geometry, we need the weighted average of the balance points for each triangle, but since the weight (area) of each triangle is the same, 0.5, the simple average provides the same result. The balance points for the left and right triangles are 2/3 and 5/3, respectively, so the average is $7/6 \approx 1.167$, so EX is greater than q_2.

Method 2: Using integrals:

$$EX = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x^2dx + \int_1^3 \frac{1}{4}[3x - x^2]dx$$

$$\begin{aligned} EX &= \frac{x^3}{3}\bigg|_0^1 + \frac{1}{4}\left(3\frac{x^2}{2} - \frac{x^3}{3}\right)\bigg|_1^3 \\ &= 1/3 + 1/4(27/2 - 27/3 - 3/2 + 1/3) \qquad = 1/3 + 1/4(12 - 9 + 1/3) = 14/12 = 7/6 \approx 1.167 \end{aligned}$$

1c.

Method 1: The area in the middle of the triangles, can be found as 1 - area of the triangles on the edges:

$$(0.5 - 0) * 0.5/2$$

```r
1 - (0.5 - 0)*0.5/2 - (3 - 1.5)*((3-1.5)/4)/2
```

```
## [1] 0.59375
```

Method 2: Using integrals:

$$\begin{aligned} P(0.5 < x < 1.5) &= \int_{0.5}^{1.5} xf(x)dx \\ &= \int_{0.5}^{1} xdx + \int_1^{1.5} \frac{3-x}{4}dx \\ &= \frac{x^2}{2}\bigg|_{0.5}^{1} + \frac{3}{4}x - \frac{x^2}{8}\bigg|_1^{1.5} \end{aligned}$$

Solving the above equation we get $P(0.5 < x < 1.5) = 0.59375$.

1d.

Method 1: The first quartile is located in the region below the left triangle, and the area should be 0.25, so $(q_1 - 0)q_1/2 = 0.25$ or $q_1 = \sqrt{0.5} \approx= 0.71$. Similarly, the third quartile is located in the region below the right triangle: $(3 - q_3)((3 - q_3)/4)/2 = 0.25$ so $(3 - q_3)^2 = 2$ and $q_3 = 3 - \sqrt{(}2)\sqrt{0.5} \approx= 1.59$ so $IQR = q_3 - q_1 \approx 1.59 - 0.71 = 0.88$.

Method 2: Using integrals, we can calculate q_1 as follows:

$$\int_{-\infty}^{q_1} xf(x)dx = \frac{1}{4}$$

However, we know that $q_1 < q_2$ and $q_2 = 1$ as calculated above. Therefore,

$$\int_0^{q_1} x dx = \frac{1}{4}$$

$$\frac{x^2}{2}\bigg|_0^{q_1} = \frac{1}{4}$$

$$\frac{{q_1}^2}{2} = \frac{1}{4}$$

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

Similarly, to find q_3

$$\int_1^{q_3} f(x)dx = \frac{1}{4}$$

$$\Rightarrow \int_1^{q_3} \frac{3-x}{4}dx = \frac{1}{4}$$

$$\Rightarrow [\frac{3}{4}x - \frac{x^2}{8}]_1^{q_3} = \frac{1}{4}$$

Solving the above equation we get,

$${q_3}^2 - 6q_3 + 7 = 0$$

Therefore, $q_3 = 3 \pm \sqrt{2}$.
However, $q_3 = 3 + \sqrt{2}$ is not possible.
Hence, $q_3 = 3 - \sqrt{2}$

$$\therefore IQR = q_3 - q_1$$

$$= 3 - \sqrt{2} - \frac{1}{\sqrt{2}}$$

$$= 3 - \frac{3}{\sqrt{2}} = 0.87868$$

2

a. **True**. This statement is true by definition of a symmetric random variable. The median is the value that separates the lower and upper halves of the distribution, and the first and third quartiles also divide the distribution into quarters. For a symmetric distribution, the median and the average of the first and third quartiles will be the same.

b. **False**. This statement is not necessarily true. The reason is that large spread or variation of extreme values (in the range of X) will affect the standard deviation directly but may not have any influence on the IQR, so the standard deviation could be as large as we would like it to be without changing the IQR. Here is an example where one single value makes all the difference (`x2` is the counterexample needed):

```
sample1 = rbinom(99, 50, 0.5)
x1 = c(sample1, 10)
x2 = c(sample1, 10^5)
IQR(x1)
```

```
## [1] 5
```

```
sqrt(mean(x1^2) - mean(x1)^2)
```

```
## [1] 3.492435
```

```
IQR(x2)
```

```
## [1] 4.25
```

```
sqrt(mean(x2^2) - mean(x2)^2)
```

```
## [1] 9947.345
```

c. **False**. The expected value is affected by all the observations. One extreme observation can make it as large (or small) as desired, while one extreme observation wouldn't have any influence on the location of the first and third quartiles. The example before also serves as illustration (`x2` is the counterexample needed):

```
summary(x1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.00   23.00   25.00   25.27   28.00   32.00 
```

```
summary(x2)
```

```
##      Min.    1st Qu.    Median      Mean    3rd Qu.      Max. 
##     19.00     23.75     25.00   1025.17     28.00 100000.00 
```

d. **True**. If the standard deviation of a random variable equals zero, then the variance of the random variable is also zero, which means that the random variable takes on only one value with probability 1. In this case, the IQR of the distribution is also zero, since the first and third quartiles are the same as the only value that the random variable can take.

e. **False**. Example. Let's say, for a discrete random variable that is not symmetric, where the mean is greater than the median, changing any one value in the range of X that is smaller than the median, and making the value even smaller than what it is, would not change the median at all, but would reduce the expected value to any value we want, in particular one equal to the median. As shown in the example below (moving from `x1` to `x2`):

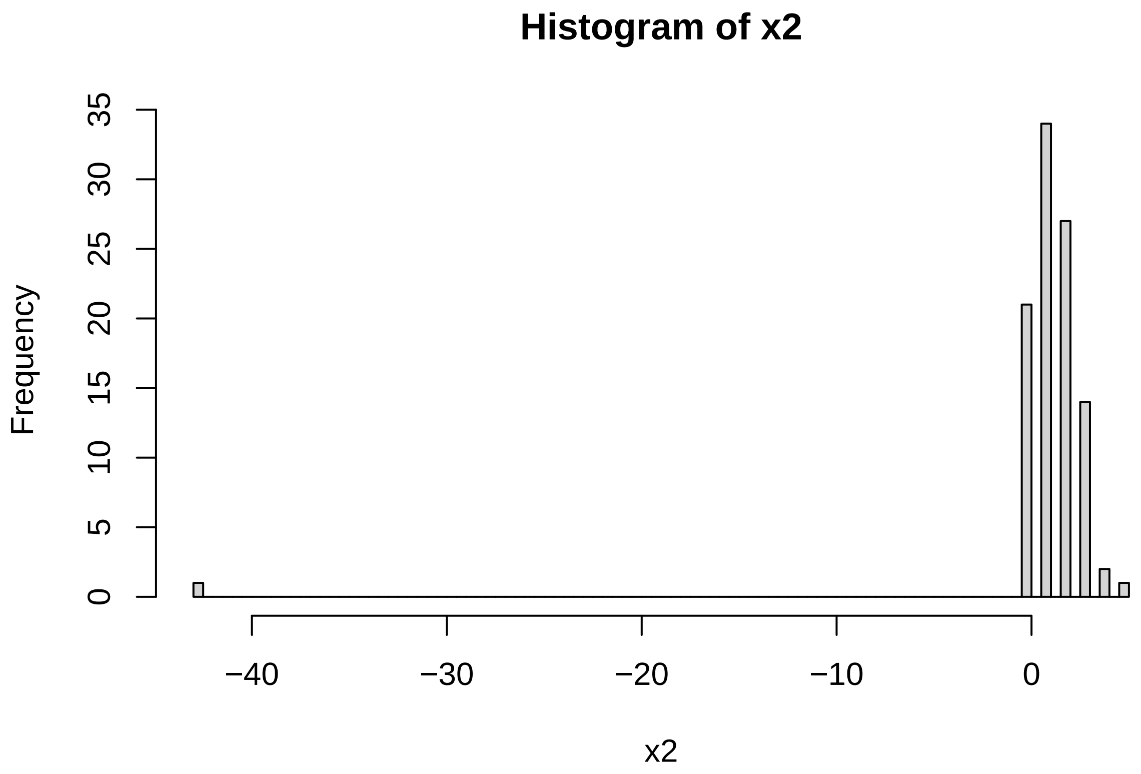
```
set.seed(126)
x = rbinom(99, 30, 0.05)
x1 = c(0,x)
summary(x1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    1.00    1.43    2.00    5.00
```

```
x2 = c(-43,x)
summary(x2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -43       1       1       1       2       5
```

```
hist(x2, breaks = 100)
```

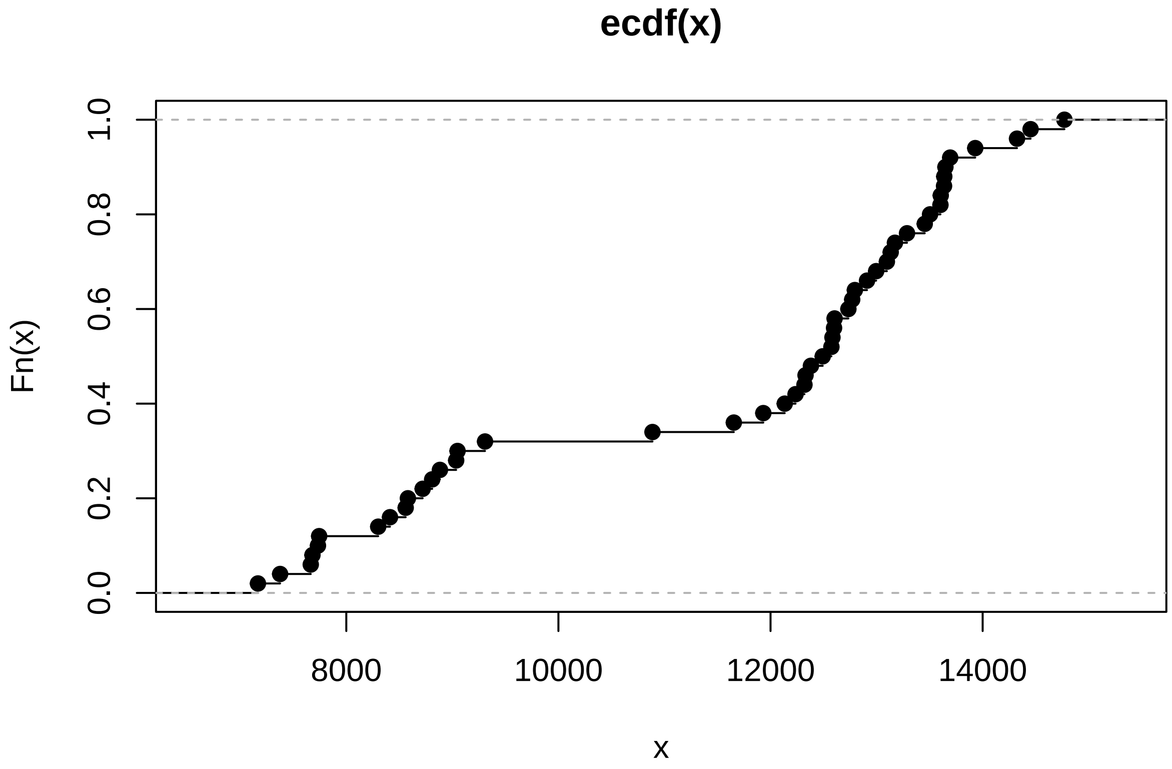


3

```
set.seed(520)
x <- sample(US_births_2000_2014$births,50)
```

3a.

```
plot(ecdf(x))
```



3b.

```
EX <- mean(x)
VX <- mean(x^2) - mean(x)^2
c(EX,VX)
```

```
## [1]   11498.14 5343986.92
```

3c.

```
m <- median(x)
iqr <- IQR(x)
c(m,iqr)
```

```
## [1] 12533.00  4337.25
```

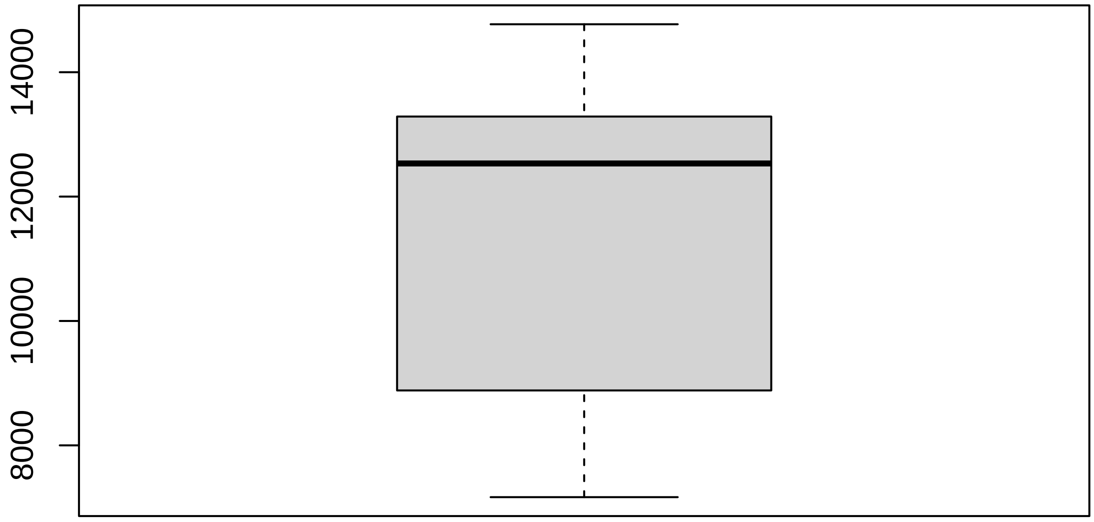
3d.

```
iqr / sqrt(VX)
```

```
## [1] 1.876211
```

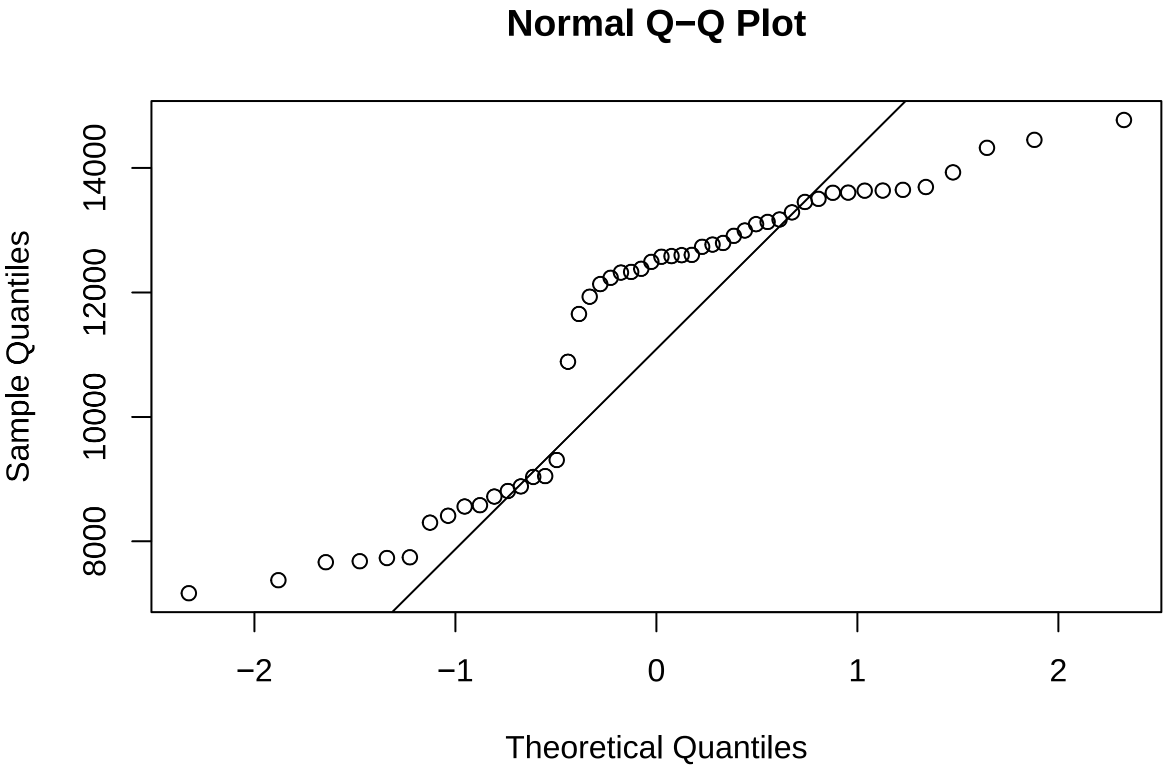
3e.

```
boxplot(x)
```



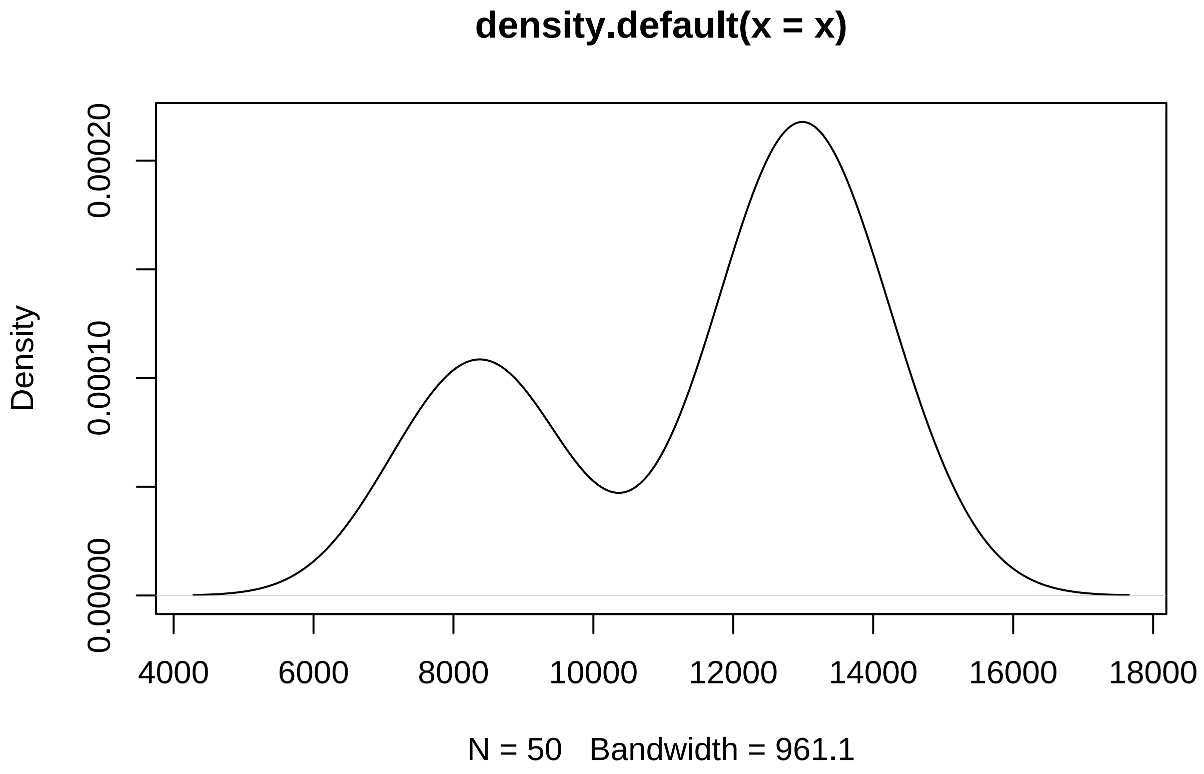
3f.

```
qqnorm(x)
qqline(x)
```



3g.

```
plot(density(x))
```



3h.

The sample doesn't seem to be drawn from a normal distribution because the QQ-plot has many deviation from the 45 degree line and the kernel density plot doesn't have the typical bell-shaped form observed on a normal curve (it actually has two modes or peaks).

4

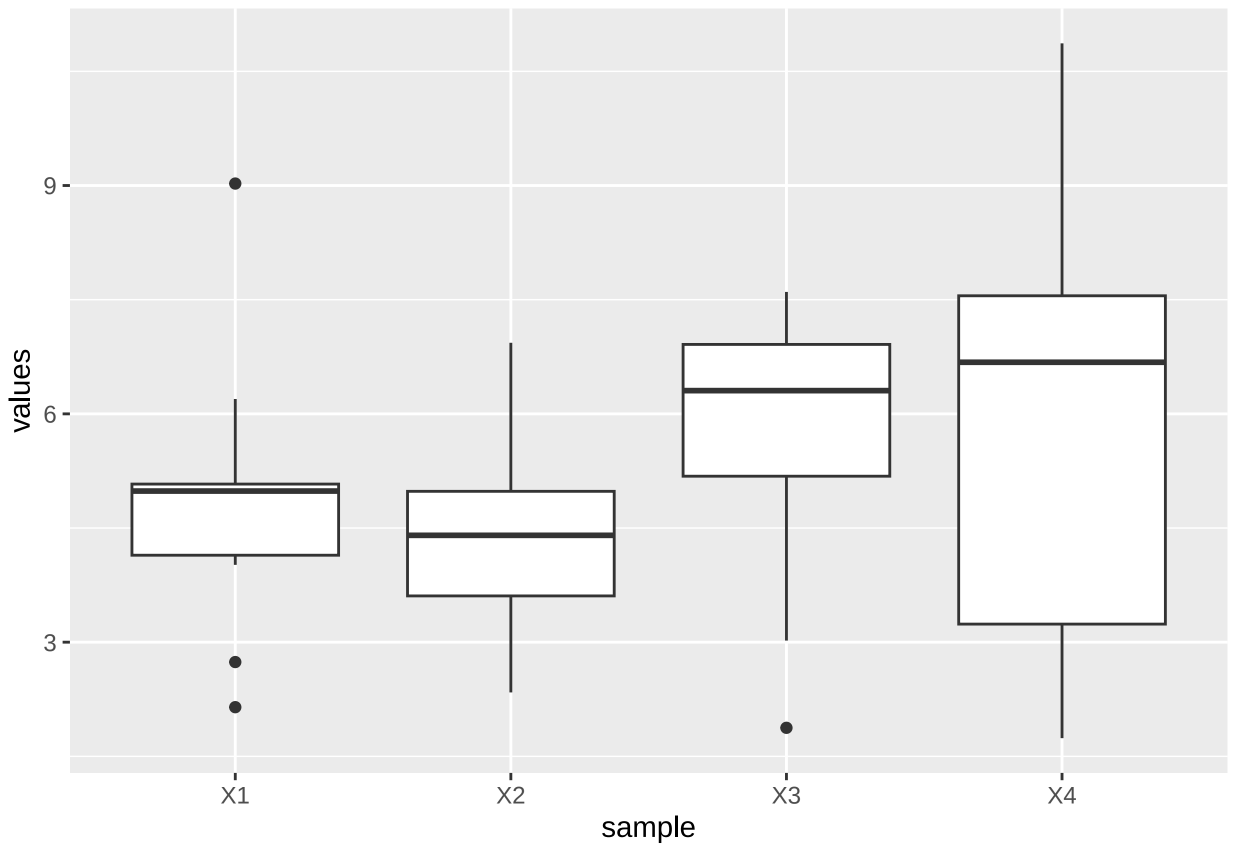
```
library(tidyverse)
x = scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/sample773.dat")
mat1 = matrix(x, nrow = 10, ncol = 4, byrow = T)
mat1
```

```
##        [,1]  [,2]  [,3]   [,4]
##  [1,] 5.098 4.627 3.021  7.390
##  [2,] 2.739 5.061 6.173  5.666
##  [3,] 2.146 2.787 7.602  6.616
##  [4,] 5.006 4.181 6.250  7.868
##  [5,] 4.016 3.617 1.875  2.428
##  [6,] 9.026 3.605 6.996  6.740
##  [7,] 4.965 6.036 4.850  7.605
##  [8,] 5.016 4.745 6.661 10.868
##  [9,] 6.195 2.340 6.360  1.739
## [10,] 4.523 6.934 7.052  1.996
```

```
df1 = data.frame(mat1)
df1
```

```
##       X1    X2    X3     X4
## 1  5.098 4.627 3.021  7.390
## 2  2.739 5.061 6.173  5.666
## 3  2.146 2.787 7.602  6.616
## 4  5.006 4.181 6.250  7.868
## 5  4.016 3.617 1.875  2.428
## 6  9.026 3.605 6.996  6.740
## 7  4.965 6.036 4.850  7.605
## 8  5.016 4.745 6.661 10.868
## 9  6.195 2.340 6.360  1.739
## 10 4.523 6.934 7.052  1.996
```

```
df.long = df1 |> pivot_longer(cols = X1:X4,names_to = "sample", values_to = "values")
ggplot(df.long, aes( x = sample,y = values)) +
  geom_boxplot()
```



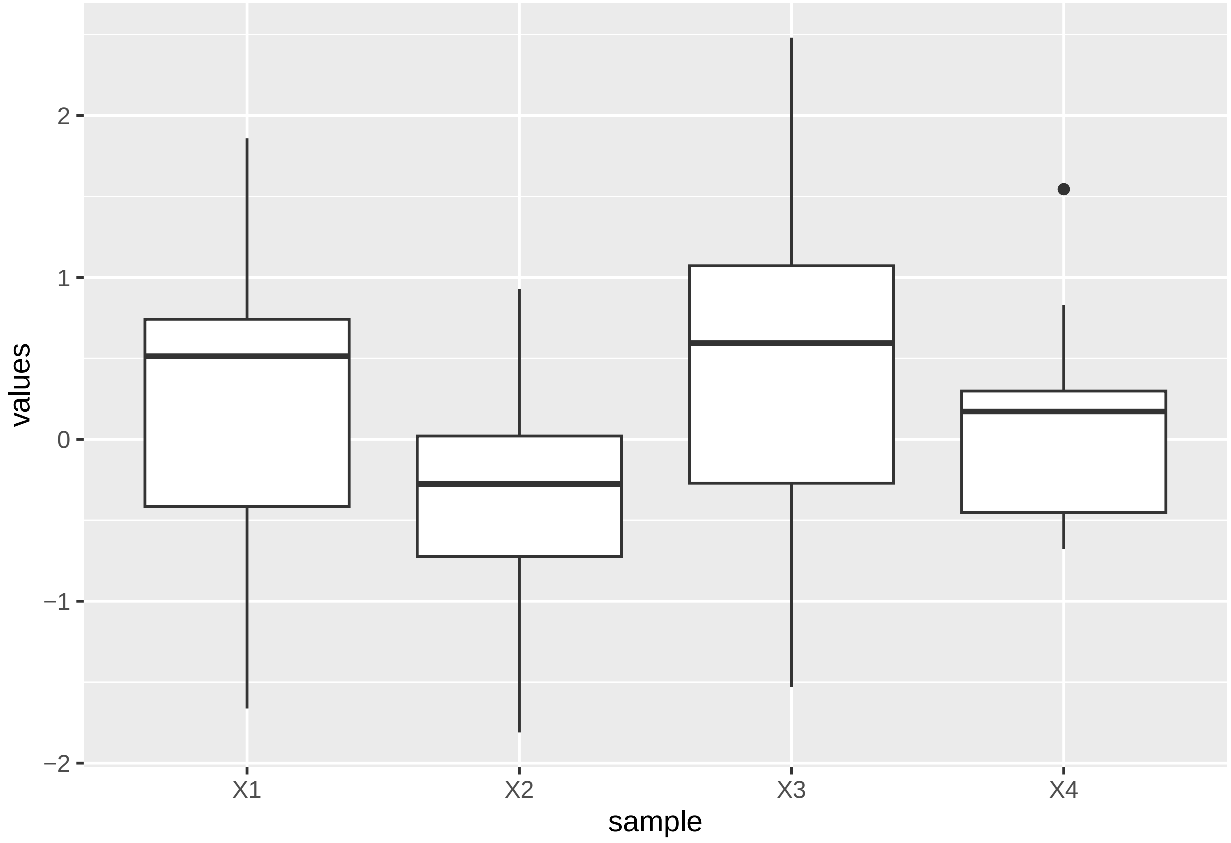
The samples clearly don't seem to come from the same distribution as the median is different and the dispersion is very different.

4b.

```
y = rnorm(40)
mat2 = matrix(y, nrow = 10, ncol = 4, byrow = T)
df2 = data.frame(mat2)
df2
```

```
##            X1          X2         X3         X4
## 1   0.5126612 -1.81063796 -0.4300173  0.8309330
## 2  -1.6626429 -0.68183108  1.1472526 -0.4894252
## 3   1.8588908 -0.04135104  0.8176767  0.2408158
## 4  -0.3036491  0.43834125 -1.5313678 -0.6787932
## 5   0.5134266  0.04074749  0.3701544  1.5445790
## 6   0.8040278 -0.73686673  0.2059935  0.1022925
## 7  -1.5601210  0.92926208  2.4808067  0.2901084
## 8   0.5545426 -1.24911395 -0.5398893  0.3007138
## 9  -0.4517852 -0.32886800  0.8437721 -0.4982010
## 10  0.8612707 -0.22283843  1.3370894 -0.3387699
```

```
df2.long = df2 |> pivot_longer(cols = X1:X4,names_to = "sample", values_to = "values")
ggplot(df2.long, aes( x = sample,y = values)) +
  geom_boxplot()
```

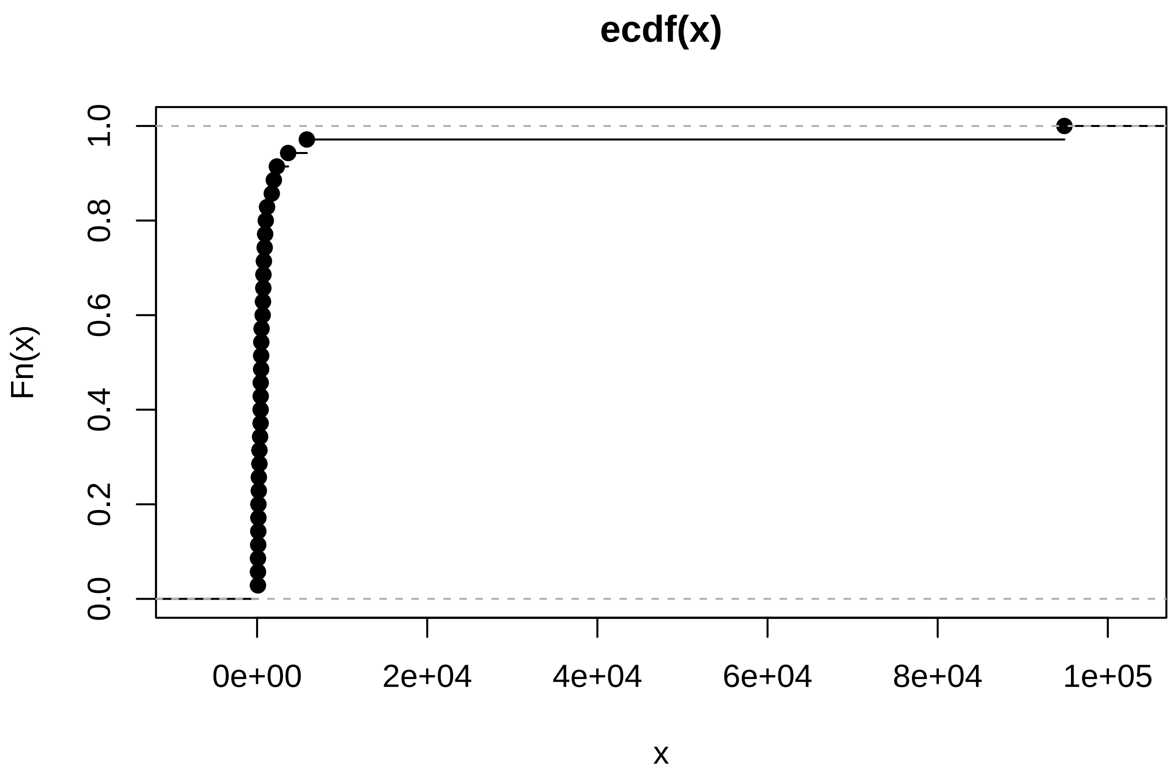


There is quite a bit of variation for these samples as well, although perhaps not as much as in the previous case.

5

5a

```
pop1 = unisex_names$total
set.seed(100)
x = sample(pop1, 35, T)
plot(ecdf(x))
```



5b.

The plug-in estimates are:

```
mean(x) #mean
```

```
## [1] 3541.43
```

```
mean(x^2)- mean(x)^2  #variance
```

```
## [1] 246716139
```

```
median(x)  #median
```

```
## [1] 480.3995
```

```
iqr <- unname(quantile(x,0.75)-quantile(x,0.25))
iqr
```

```
## [1] 679.4816
```

5c

```
#Ratio of the data
iqr/sqrt(mean(x^2)-mean(x)^2)
```

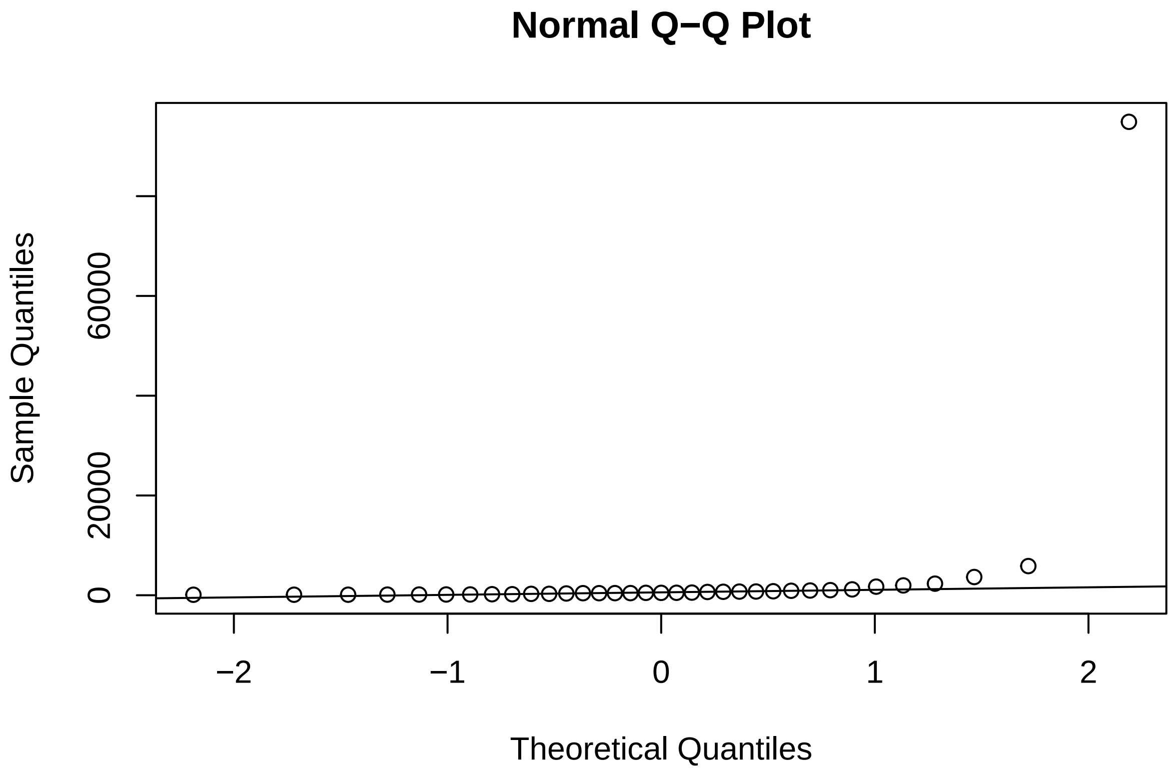
```
## [1] 0.04325924
```

```
ratio.normal = (qnorm(0.75)-qnorm(0.25))/1 #the ratio iqr/sigma for the standard normal
```

We can conclude that the ratio for normal distribution is 1.35 and the ratio for the data is 0.043. Since these both ratios do not match the data was likely not drawn from a normal distribution.

5d

```
qqnorm(x); qqline(x)
```



The QQ plot doesn't follow the straight line and clearly deviates for some observations. These data doesn't seem to be drawn from a normal distribution.

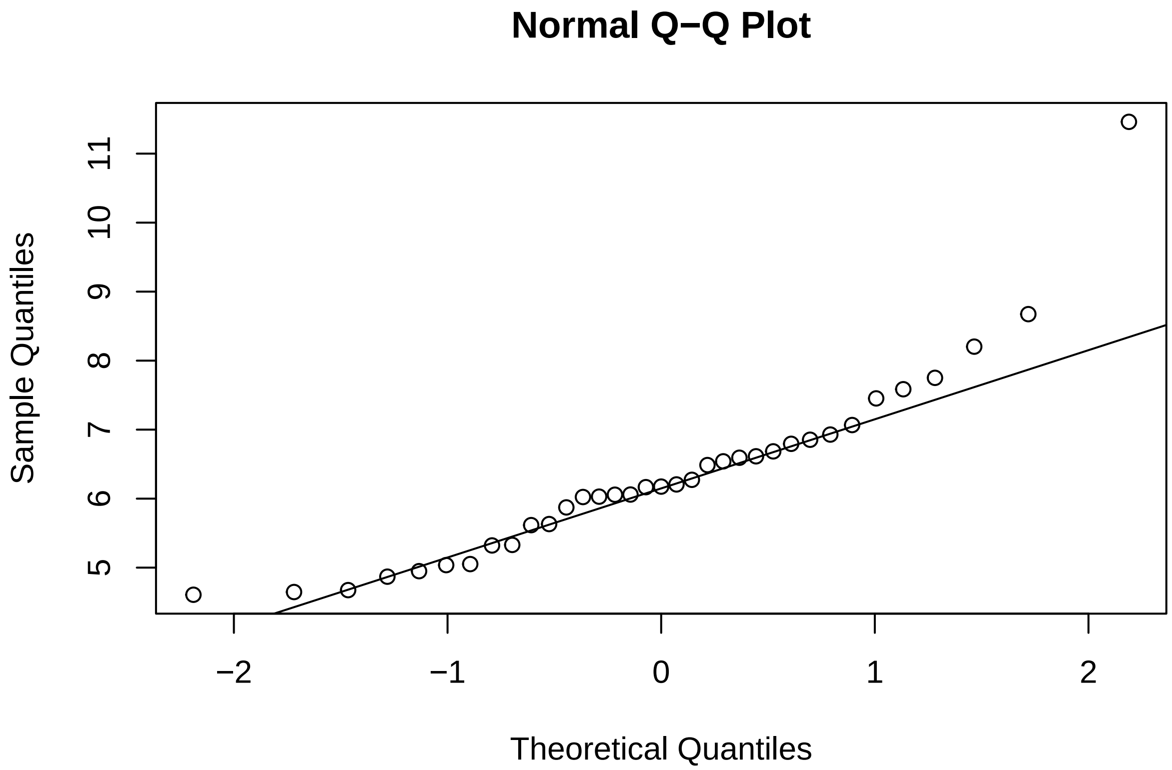
5e

```
y <- log(x)
iqr <- unname(quantile(y,0.75)-quantile(y,0.25))
iqr/sqrt(mean(y^2)-mean(y)^2)
```

```
## [1] 1.020015
```

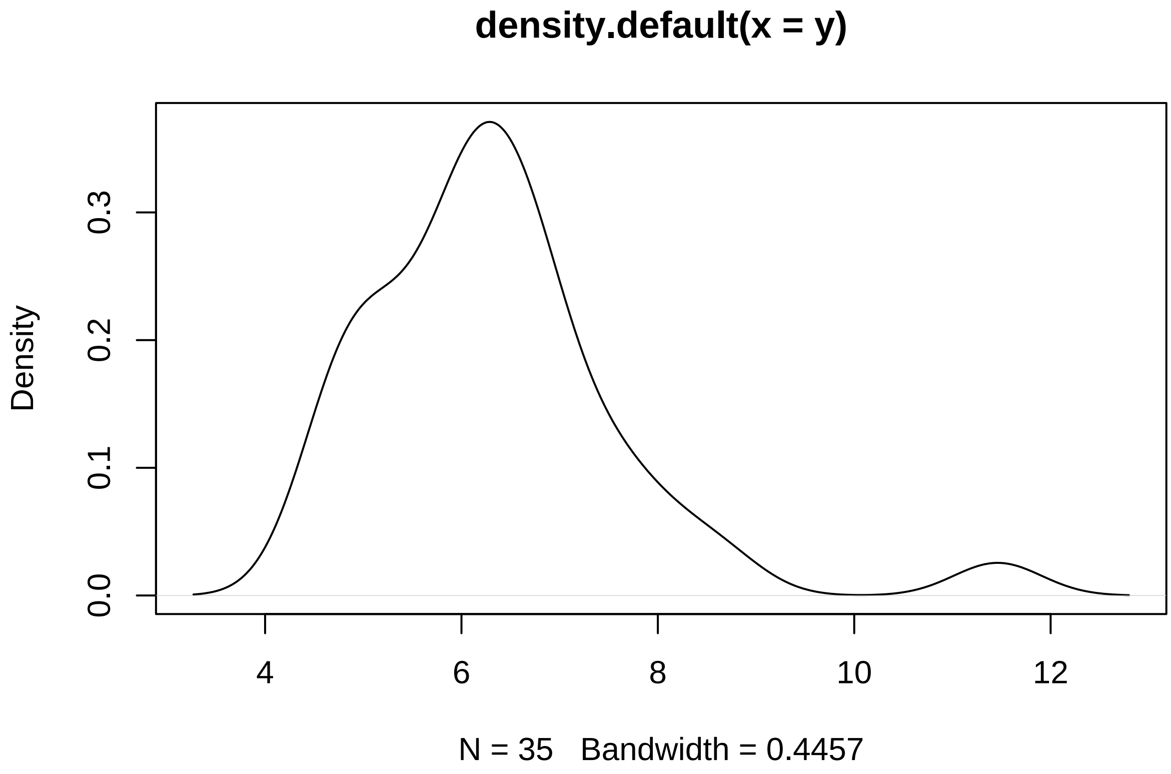
The ratio is closer to the normal now, but not close enough to thing the transformed data was drawn from a normal distribution.

```
#qqplot
qqnorm(y); qqline(y)
```



Similarly, the QQ plot still doesn't show the data were drawn from a normal distribution, although the values are not as extreme as before.

```
#plotting density of y
plot(density(y))
```


density.default(x = y)
Density
0.0
0.1
0.2
0.3
4
6
8
10
12
N = 35 Bandwidth = 0.4457