Problem Set . 6

Aditya Mhaske

1) X = continous random Variable with probability density Function.

$$F(x) = \begin{cases} 0 & x < 0 \\ x & x \in (0.1) \\ (3-x)/4 & x \in (1.3) \\ 0 & x > 3 \end{cases}$$

$$= \int_{-\infty}^{0} 0 \cdot dx + \int_{0}^{1} x \cdot dx + \int_{1}^{3} \left(\frac{3}{4} - \frac{x}{4}\right) \cdot dx + \int_{3}^{\infty} 0 \cdot dx$$

$$= \left[\frac{x^2}{2}\right]_0^1 + \left[\frac{3x}{4} - \frac{1}{8}x^2\right]_1^3 = 0.5$$

$$= \frac{1}{2} + \frac{3(c-1)}{4} - \frac{(c^2-1)}{8} = 0.5$$

$\Rightarrow c = 1$

Thus $q_2(X) = 1$

b) Mean of $f(x)$ =

$$\int_{-\infty}^{\infty} x f(x) \cdot dx = \int_{0}^{1} x^2 \cdot dx + \int_{1}^{3} (3-x)\frac{x}{4} \, dx$$

$$= \left[\frac{x^3}{3}\right]_0^1 + \left[\frac{3x^2}{8}\right]_1^3 - \left[\frac{x^3}{12}\right]_1^3$$

$$= \frac{1}{3} + \frac{24}{8} - \frac{26}{12} = \frac{7}{16}$$

$\therefore$ Mean > Median For $F(x)$

here graph is also right skewed

$\Rightarrow$ Suggests Mean (should be) > Median

c) $P(0.5 < x < 1.5) = \int_{0.5}^{1} x \cdot dx + \int_{1}^{1.5} \frac{(3-x)}{4} \cdot dx$

$= \left[\frac{x^2}{2}\right]_{0.5}^{1} + \left[\frac{3x}{4}\right]_{1}^{1.5} - \left[\frac{x^2}{8}\right]_{1}^{1.5}$

$= \frac{1}{2} - \frac{1}{8} + \frac{9}{8} - \frac{3}{4} - \frac{2.25}{8} + \frac{1}{8} = \boxed{\frac{4.75}{8}}$

d) $iqr = q_3(x) - q_1(x)$

For $q_1 \longrightarrow$ AOC $= 0.25$

$\frac{1}{2}x^2 = 0.25$

$\boxed{q_1 \Rightarrow x = 0.7}$

for q_3 area b/w $q_3(x)$ and $3 = 0.25$

$\Rightarrow \frac{1}{2}(3-x)\frac{(3-q_3)}{4} = 0.25$

$= (3-x)^2 = 2$

$\Rightarrow x = 3 - \sqrt{2} = 1.58$

$\therefore iqr = 1.58 - 0.7$

$= \boxed{0.88}$

Q. 2

7. Identify each of the following statements as explain each of your answers.True or False .

a) For every symmetric random variable X , the median of X equals the average of the first and third quartiles of X .

→ $\boxed{\text{True}}$: symmetric distribution $(x_1+x_2)/2$ mean lies at center i.e. $(q_1+q_3)/2 = (0.75+0.25)/2 = 0.5$

b) For every random variable X , the interquartile range of X is greater than the standard deviation of X .

→ $\boxed{\text{False}}$: iqr covers 50% distribution
SD covers 69%.

c) For every random variable X , the expected value of X lies between the first and third quartile of X .

→ $\boxed{\text{False}}$: When data is skewed , it is possible for mean to be outside q_1 & q_3

d) If the standard deviation of a random variable equals zero, then so does its interquartile range.

– $\boxed{\text{True}}$: OSD implies there is no spread but only single datapoint . ∴ All percentile overlap

e) If the median of a random variable equals its expected value, then the random variable is symmetric.

→ $\boxed{\text{False}}$ - For 1 or 1+ unique values for median this statement is False.

Problem Set: 06

Aditya Sanjay Mhaske

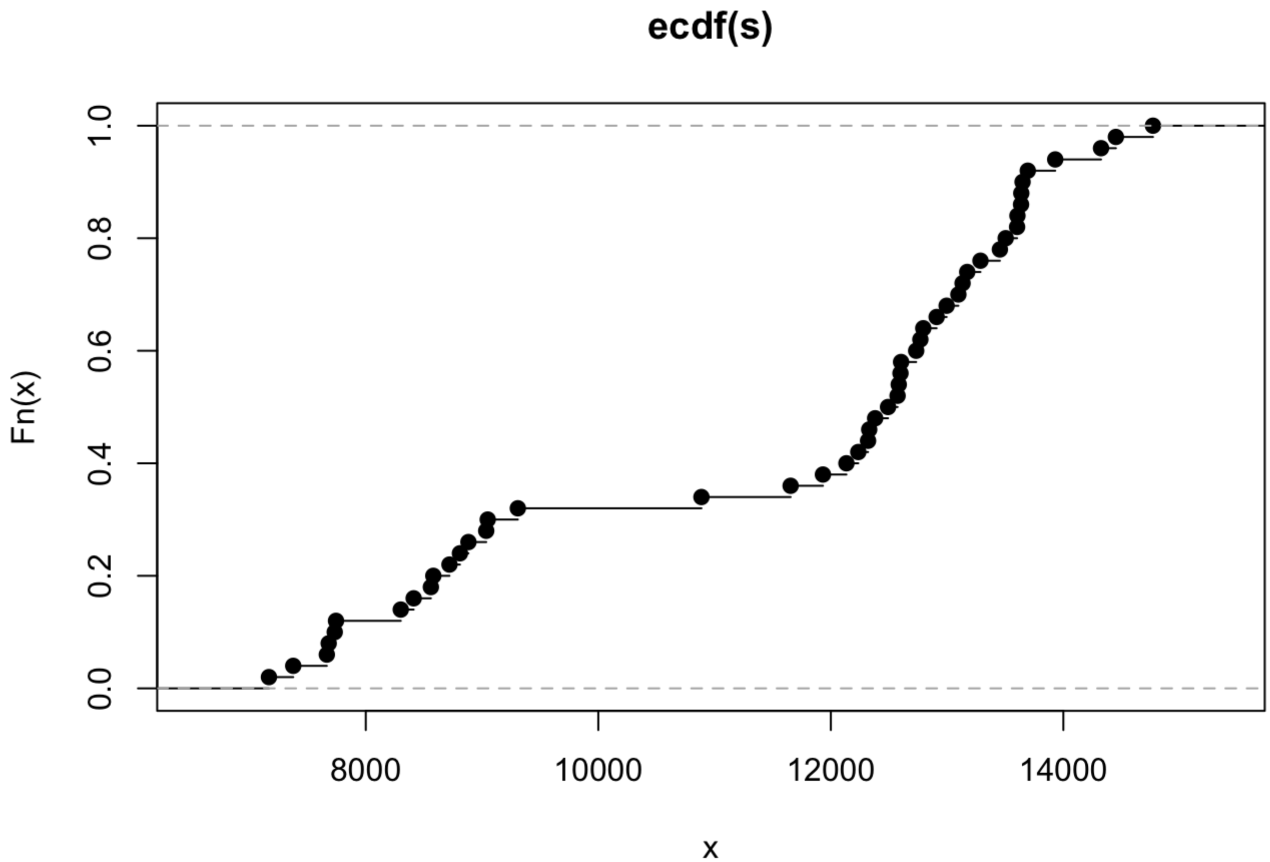
2023-02-20

Q3.

```
library(fivethirtyeight)

#getting the data
data("US_births_2000_2014")

set.seed(520)
#random samples
s <- sample(US_births_2000_2014$births, size = 50)
```

a. Empirical CDF:

```
ef = ecdf(s)

plot(ef)
```



b. Plug-in estimation of population mean and variance

i.e simple variance and simple mean:

```
n <- length(s)
n
```

```
## [1] 50
```

mean

```
p_mean <- mean(s)
p_mean
```

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

variance

```
p_var <- mean(s^2) - p_mean^2
p_var
```

```
## [1] 5343987
```

c. Plug-in estimates of population median and interquartile range

median:

```
p_median <- median(s)
p_median
```

```
## [1] 12533
```

IQR:

```
#0.25 and 0.75 quantiles
quantile(s, probs=c(.25, .75))
```

```
##      25%      75%
##  8921.25 13258.50
```

```
#summary [quantiles]
summary(s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7167    8921   12533   11498   13258   14771
```

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

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

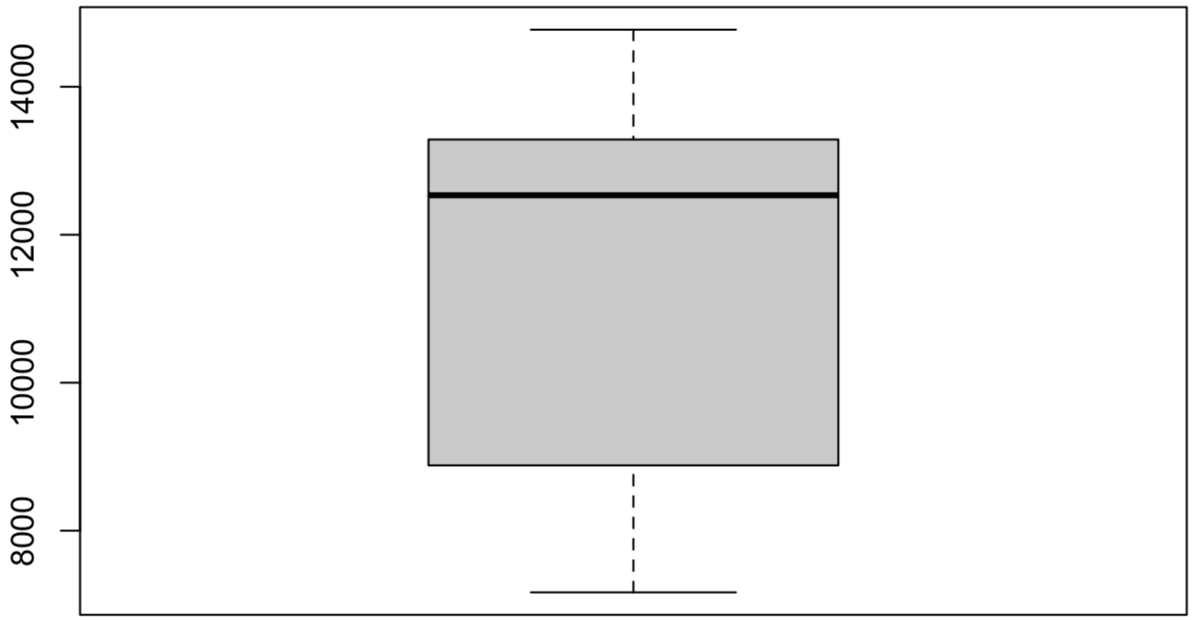
d. Compute the ratio of the plug-in estimate of the interquartile range to the square root of the plug-in estimate of the variance.

```
iqr / sqrt(p_var)
```

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

e. boxplot:

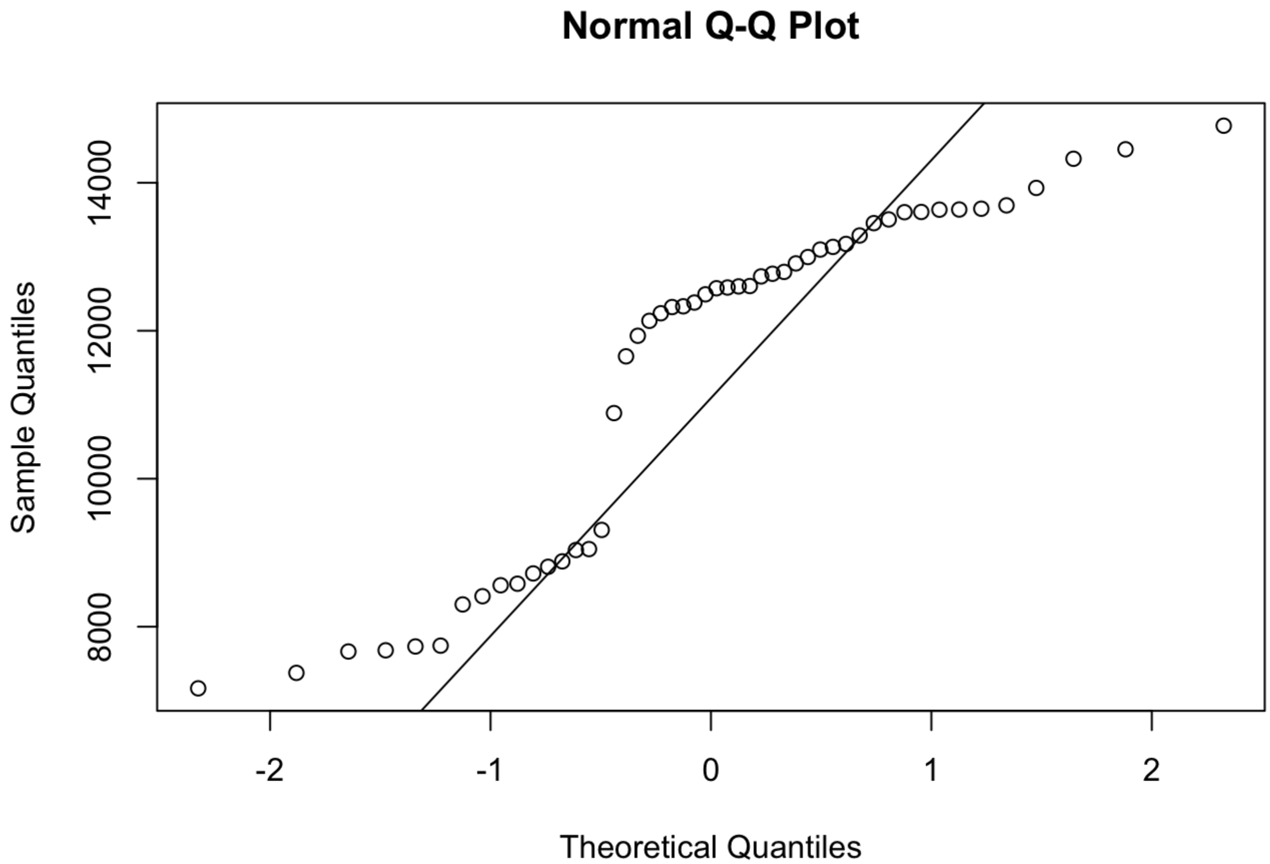
```
boxplot(s)
```



f. normal probability plot

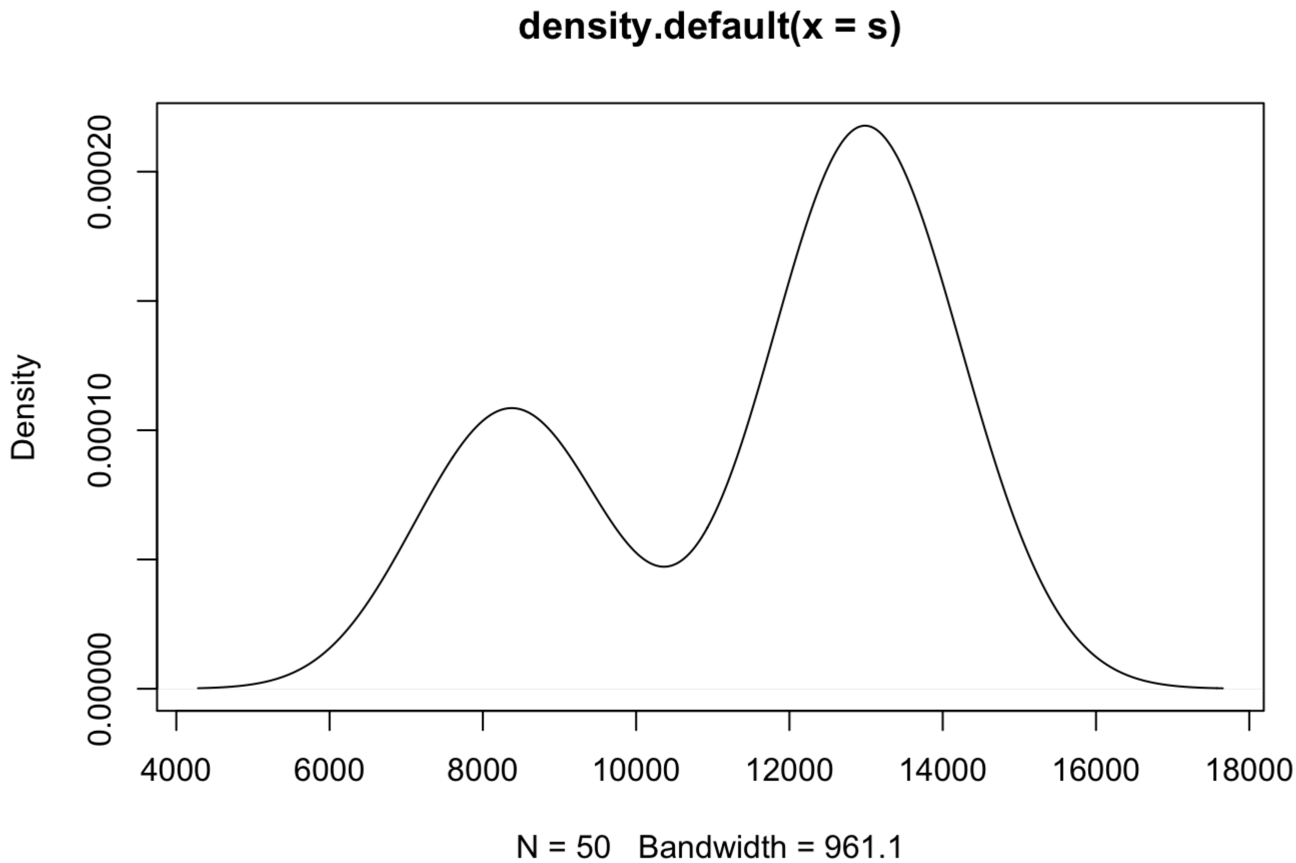
Q-Q plot

```
qqnorm(s)
qqline(s)
```



g. kernel density estimate

```
plot(density(s))
```



3. h.

The Q-Q plot demonstrates that the data points are not linear (i.e. not in straight line). As a result, this sample does not come from a normal distribution.

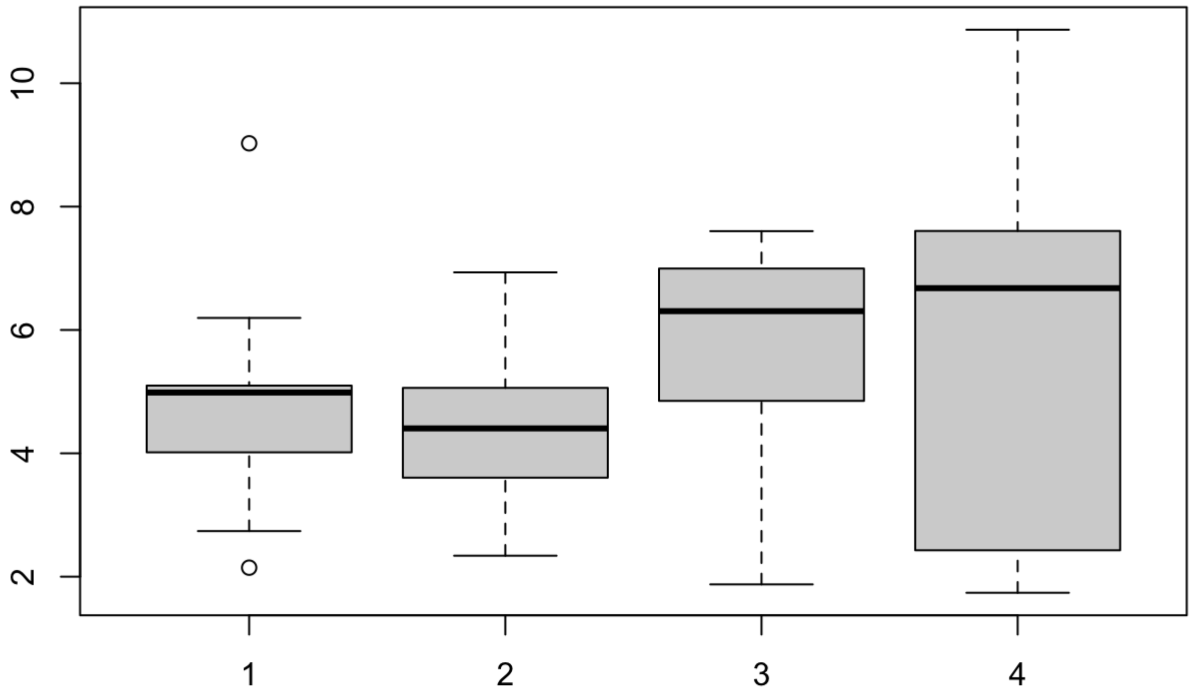
#-----------------------------

Q4. ISI 7.7: 3.

```
s1 <- c(5.098, 2.739, 2.146, 5.006, 4.016, 9.026, 4.965, 5.016, 6.195, 4.523)
s2 <- c(4.627, 5.061, 2.787, 4.181, 3.617, 3.605, 6.036 ,4.745, 2.340, 6.934)
s3 <- c(3.021, 6.173, 7.602, 6.250, 1.875, 6.996, 4.850, 6.661, 6.360, 7.052)
s4 <- c(7.390, 5.666, 6.616, 7.868, 2.428, 6.740, 7.605, 10.868, 1.739, 1.996)
```

a. boxplot

```
boxplot(s1, s2, s3, s4)
```

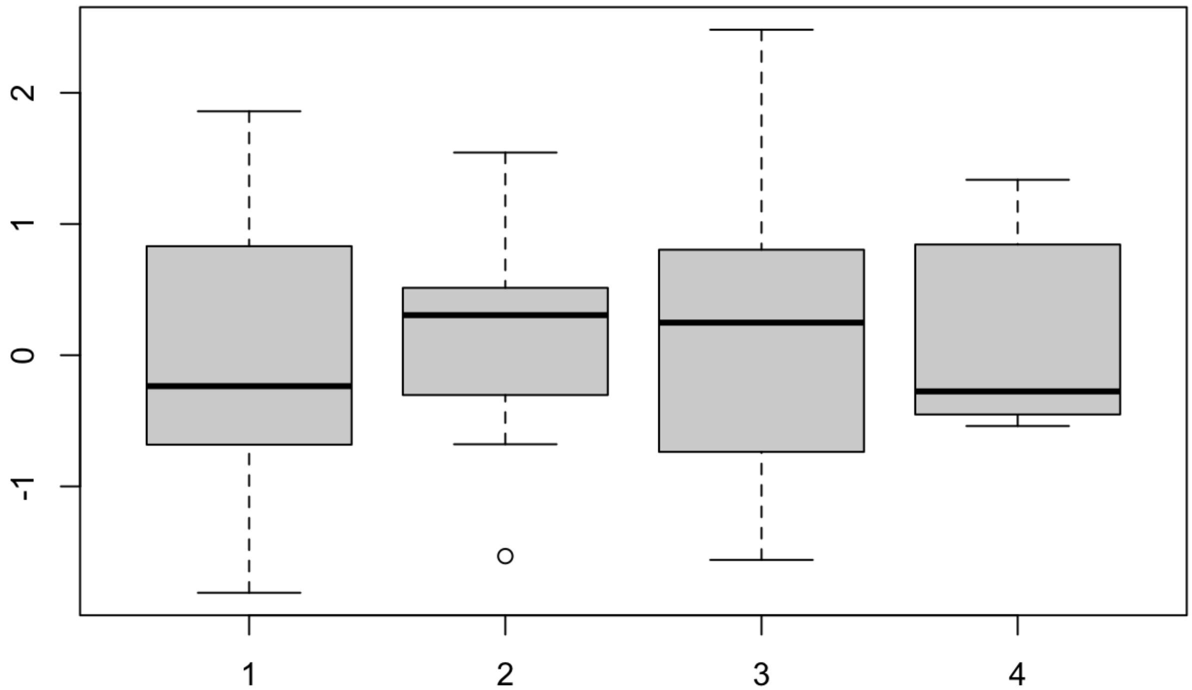


The boxplots for the four samples do not appear to be identical in the image above. Their skewness, range, and quartiles differ.

Thus, these samples are not drawn from same population.

b.

```
boxplot(rnorm(10), rnorm(10), rnorm(10), rnorm(10))
```



When the four samples obtained by the rnorm function are displayed, they are #' different from the other samples. These plots do not appear to be related. #' Their skewness, range, and quartiles are also different #' As a result, it is possible that the samples s1, s2, s3, and s4 were all taken from the same normal distribution. #'

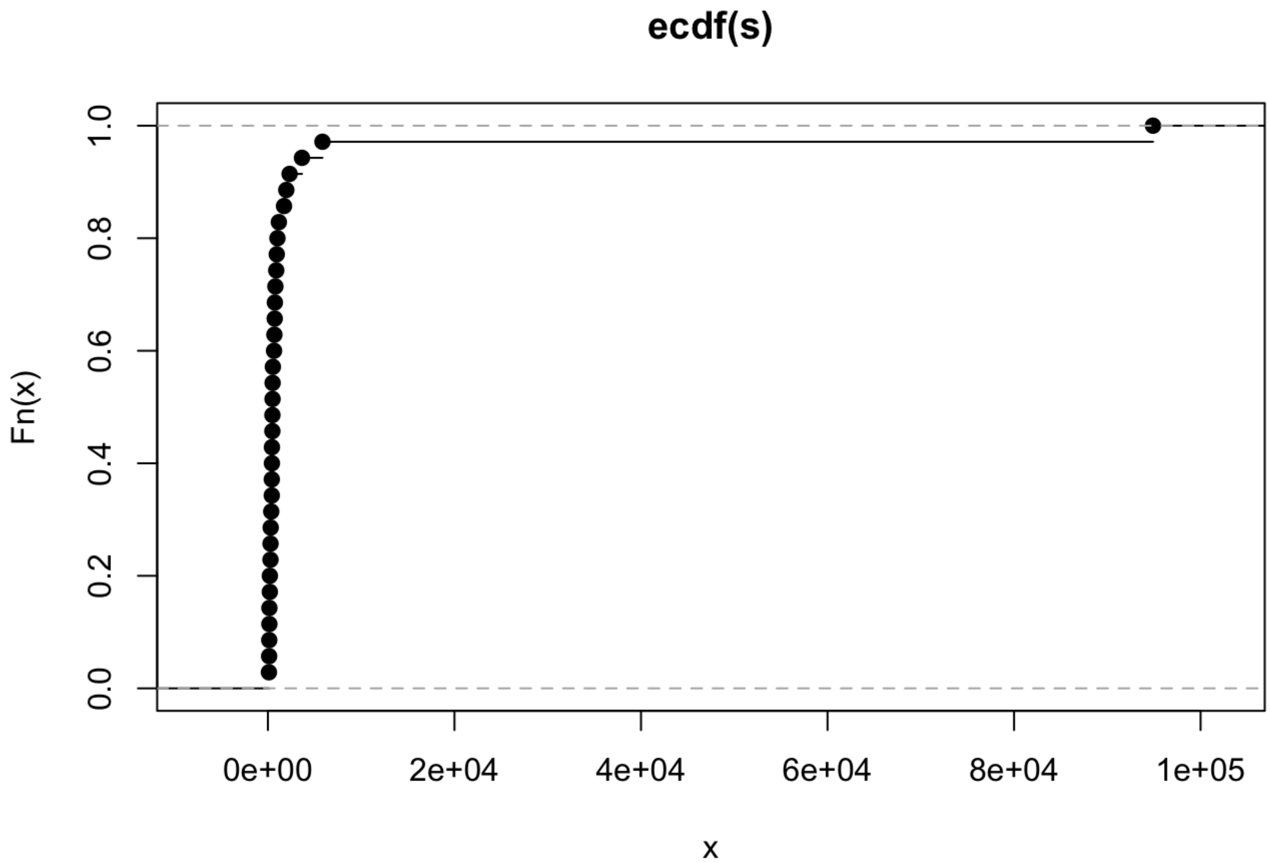
#——————————————————————————

Q5. ISI 7.7: 4.

```
#getting the data
data("unisex_names")

set.seed(100)
#random samples
s <- sample(unisex_names$total, size = 35)
```

a. Empirical CDF:

```
ef = ecdf(s)

plot(ef)
```



b. Plug-in estimation of mean and variance, median, IQR

mean

```
p_mean <- mean(s)
p_mean
```

```
## [1] 3558.419
```

variance

```
p_var <- mean(s^2) - p_mean^2
p_var
```

```
## [1] 246604651
```

median:

```
p_median <- median(s)
p_median
```

```
## [1] 495.7103
```

IQR:

```
#0.25 and 0.75 quantiles
quantile(s, probs=c(.25, .75))
```

```
##      25%      75%
## 287.7077 920.0639
```

```
#summary [quantiles]
summary(s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   107.2   287.7   495.7  3558.4   920.1 94896.4
```

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

```
## [1] 632.3562
```

5. c.

```
sqrt(p_var)
```

```
## [1] 15703.65
```

```
iqr
```

```
## [1] 632.3562
```

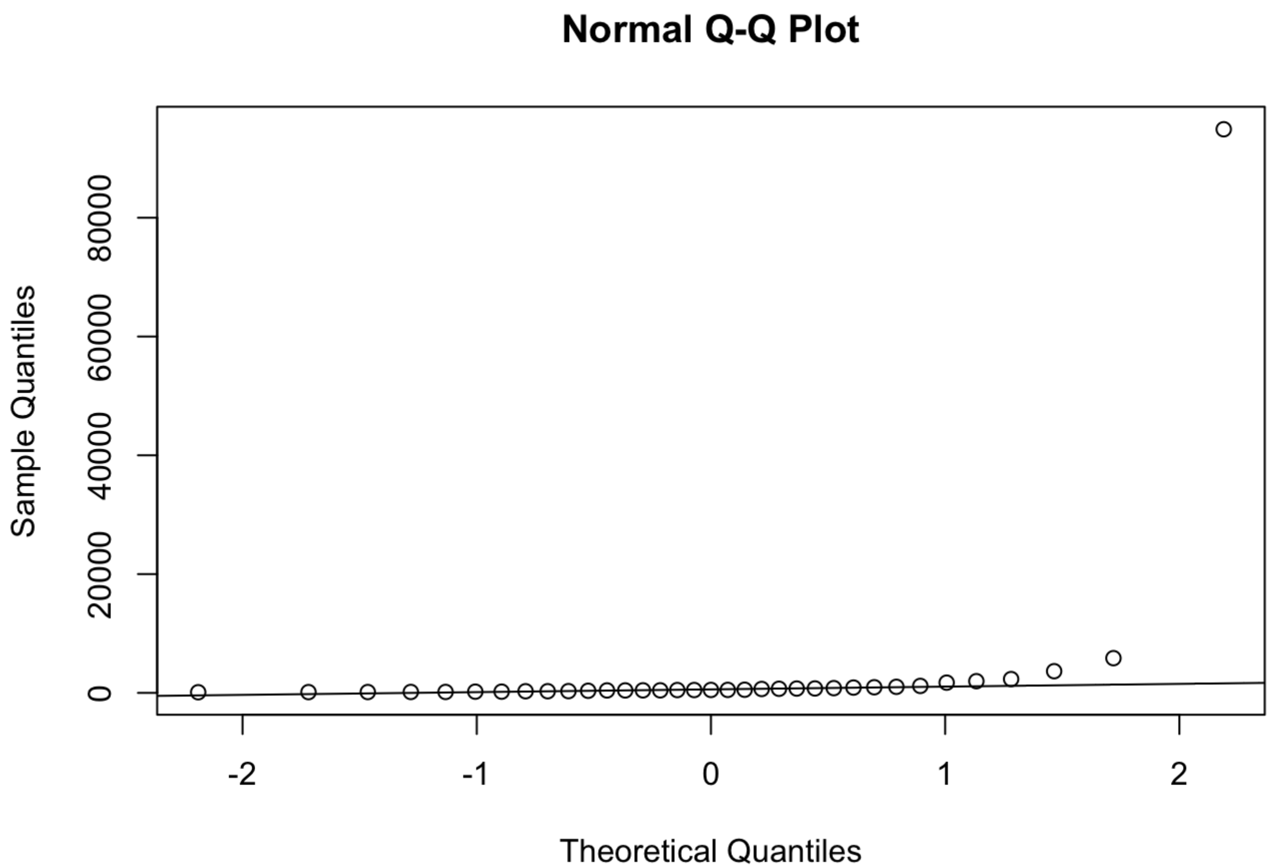
```
#ratio for comparing
iqr/sqrt(p_var)
```

```
## [1] 0.0402681
```

As compared to the square root of variance, IQR is quite modest. is an asymmetric distribution. The dispersion is also not consistent (not-even). As a result, the data does not come from a normal distribution. Moreover, the ratio implies that the samples were not taken from a normal distribution.

d. qqnorm: normal distribution?

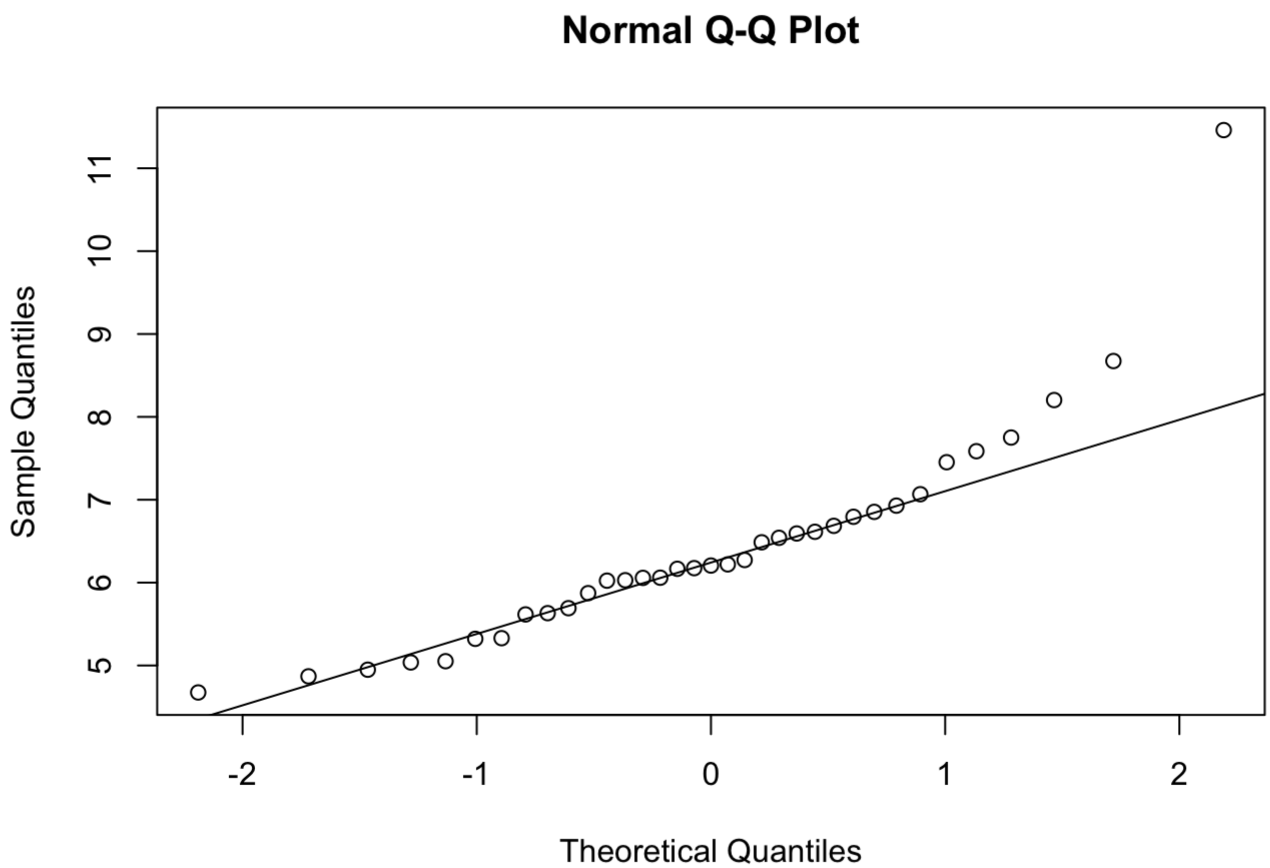
```
qqnorm(s)
qqline(s)
```



The Q-Q plot shows that the majority of the points fall on a straight line. This indicates that the data fits the normal distribution. Yet, because the final few points are away from the line, there is a slight chance that the data has a non-normal distribution. To understand the distribution, I believe that additional samples or other approaches must be employed.

e.

```
y <-log(s)

#q-q plot
qqnorm(y)
qqline(y)
```



The points are linear, according to the figure. As a result, the data may be pulled from the normal distribution. Yet, the last points are distant from the #' line, which may indicate non-normality. #' #' To calculate the distribution, more sample points are required.