STAT 6600: Assignment 1

Due: September 10, 2019

Akeem Ajede

Question 1(a)

Find the sample mean $\bar{x}$ and standard deviation s. Then compute the intervals $[\bar{x}-s, \bar{x}+s]$, $[\bar{x}-2s, \bar{x}+2s]$ and $[\bar{x}-3s, \bar{x}+3s]$ and compute the percentage of sample values that fall within each interval.

Solution

$\mathbf{Approach}$: Stem and Leaf

The sample mean is equivalent to the 50th percentile P_{50}, which was computed as delineated below:

$(n+1)p = (97)\frac{1}{2} = 48.5 = 48 + \frac{1}{2}$,

The sample mean is located between the 48th and 49th integer.

$\hat{\pi}_{\frac{1}{2}} = (1-\frac{1}{2})x_{48} + (\frac{1}{2})x_{49} = (\frac{1}{2})16.6 + (\frac{1}{2})16.7 = 16.65$

From sigma rule, 95% of the data lies within $2s$ from μ. Hence, 5% of the data are beyond $2s$, which is approximately $96 * 5\% = 4.8$, (i.e., about 4 or 5 observations). From the 16th stem, 4 and 3 stems from the right and left tail, respectively will keep 4 observations beyond 2 standard deviations from the mean. By computing the average, $\frac{3+4}{2} = 3.5 = 2\sigma$, Thus, $\sigma = 1.75$

$\mathbf{Intervals:}$

$[\bar{x}-s, \bar{x}+s] = [16.7-1.75, 16.7+1.75] = [15.0, 18.5]$ which contains 65 values $\approx 67.7\%$ of the data.

$[\bar{x}-2s, \bar{x}+2s] = [16.7-2(1.75), 16.7+2(1.75)] = [13.2, 20.2]$ which contains 91 values $\approx 94.7\%$ of the data.

$[\bar{x}-3s, \bar{x}+3s] = [16.7-3(1.75), 16.7+3(1.75)] = [11.45, 21.95]$ which contains 95 values $\approx 98.9\%$ of the data.

$\mathbf{Note:}$ The sample percentages computed approximately corresponds to the established proportions of a standard normal distribution.

Question 1(b)

Find the minimum $x_{(1)}$ and the maximum $x_{(96)}$ order statistics, the 25th, 50th and 75th percentiles (also known as the 1st quartile $Q1$, the median $\tilde{x}$ and the third quartile $Q3$, respectively. Use the stem-and-leaf plot to do so. Also,find the interquartile range $IQR = Q3 - Q1$.

Solution

$x_{(1)} = 11.9$ and $x_{(96)} = 22.1$,

$P_{25} = Q_1$

$p = \frac{25}{100} = \frac{1}{4}, (n+1)p = (97)\frac{1}{4} = 22.25 = 22 + \frac{1}{4},$

$\hat{\pi}_{\frac{1}{4}} = (1 - \frac{1}{4})x_{22} + (\frac{1}{4})x_{23} = (\frac{3}{4})15.3 + (\frac{1}{4})15.4 = 15.3$

P_{50} $P_{50} = \bar{x} = 16.65$ (Previously computed)

$P_{75} = Q_3$

$p = \frac{75}{100} = \frac{3}{4}, (n+1)p = (97)\frac{3}{4} = 72.75 = 72 + \frac{3}{4},$

$\hat{\pi}_{\frac{3}{4}} = (1 - \frac{3}{4})x_{72} + (\frac{3}{4})x_{73} = (\frac{3}{4})17.8 + (\frac{1}{4})17.8 = 17.8,$

$IQR = Q_3 - Q_1 = 17.8 - 15.3 = 2.5$

Question 1(c)

Find the 5th and 95th percentiles p_5 and p_{95}

Solution

$\mathbf{P_5}$

$p = \frac{5}{100} = \frac{1}{20}, (n+1)p = (97)\frac{1}{20} = 22.25 = 22 + \frac{1}{4}$

$\hat{\pi}_{\frac{1}{20}} = (1 - \frac{1}{20})x_4 + (\frac{1}{20})x_5 = (\frac{19}{20})13.7 + (\frac{1}{20})14.1 = \mathbf{13.7}$

$\mathbf{P_{95}}$

$p = \frac{95}{100} = \frac{19}{20}, (n+1)p = (97)\frac{19}{20} = 92.15 = 92 + \frac{3}{20},$

$\hat{\pi}_{\frac{19}{20}} = (1 - \frac{19}{20})x_{92} + (\frac{19}{20})x_{93} = (\frac{1}{20})19.8 + (\frac{19}{20})20.2 = \mathbf{20.18}$

Question 1(d)

Assuming that the data are from a normal population, and an expression for the maximum likelihood estimator of the CDF,

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} exp\Big\{ - \frac{(t-\mu)^2}{2\sigma^2}\Big\}dt.$$

Solution

Using the invariant property of MLEs, an estimate of the CDF is given as

$$F_X(x|\hat{\theta}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}}\int_{-\infty}^{x} exp\Big\{ - \frac{(t-\hat{\mu})^2}{2\hat{\sigma}^2}\Big\}dt,$$

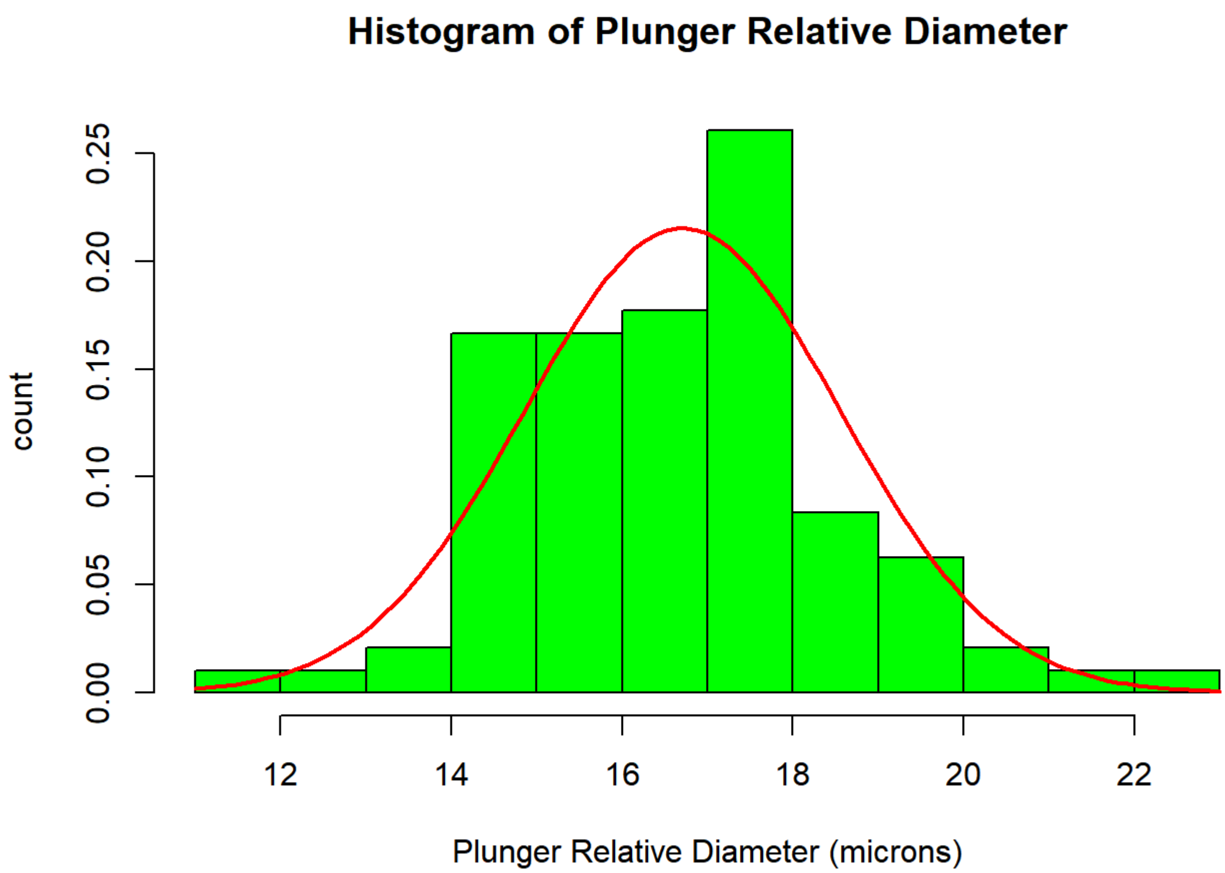
where, $\hat{\mu} = \bar{X} = 16.7$, and $\hat{\sigma}^2 = \frac{\sum_{i=1}^{96}(x_i - \bar{x})^2}{n} = \mathbf{3.396}$

Question 1(e)

Use R to generate a histogram of the sample distribution with the “theoretical” pdf overlaid.

Solution

```
Raw = read.table("Prob1InjectorPumps.txt", header = F)
y = Raw[order(Raw$V1),]
hist(y, prob = T, xlab = "Plunger Relative Diameter (microns)", ylab = "count", col = "green", m
ain = "Histogram of Plunger Relative Diameter")
m <-mean(y);std <-sqrt(var(y))
curve(dnorm(x, mean=m, sd=std), col= 2, lwd=2, add=TRUE)
```

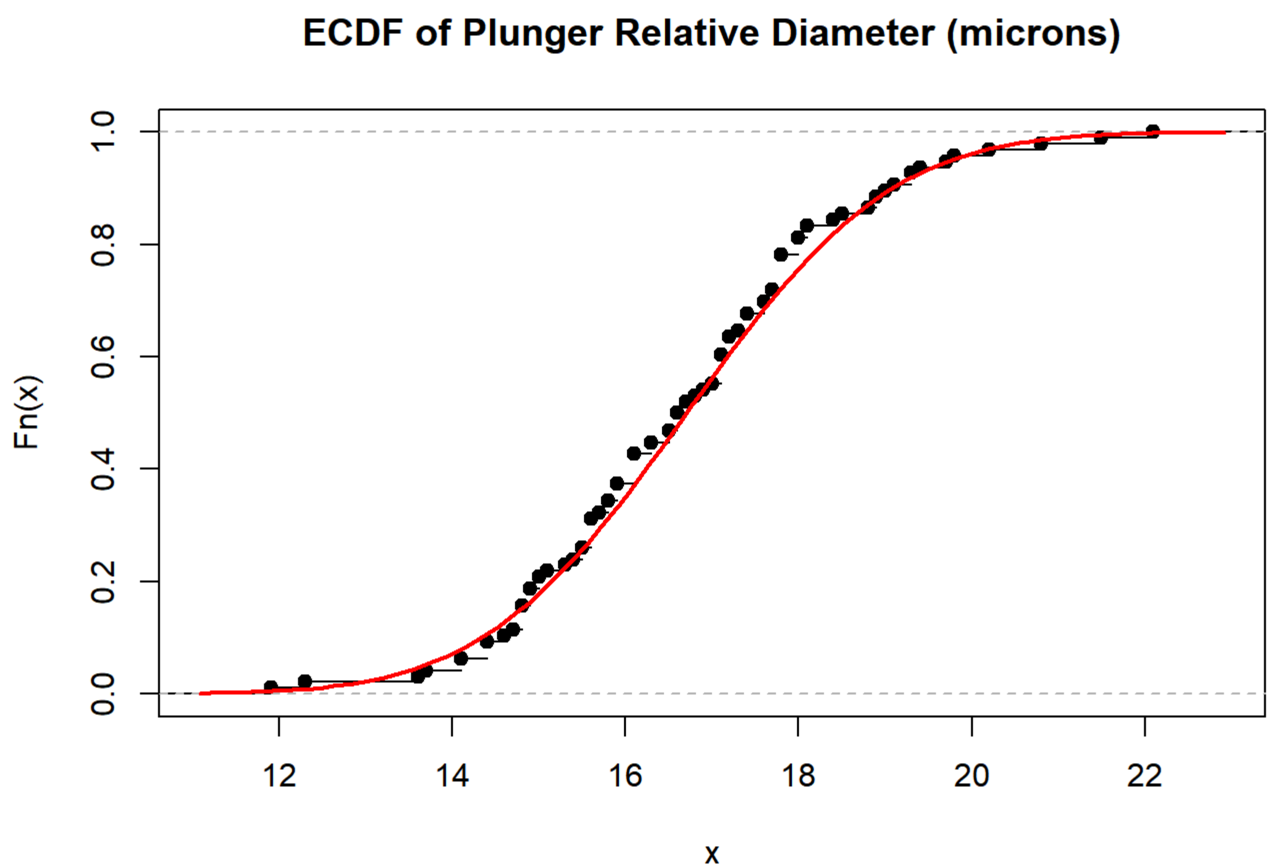


Question 1(f)

Use R to generate the empirical CDF with the “theoretical” CDF overlaid.

Solution

```
plot(ecdf(y), main = "ECDF of Plunger Relative Diameter (microns)")
curve(pnorm(x, m, std), col= 2, lwd=2, add=TRUE)
```

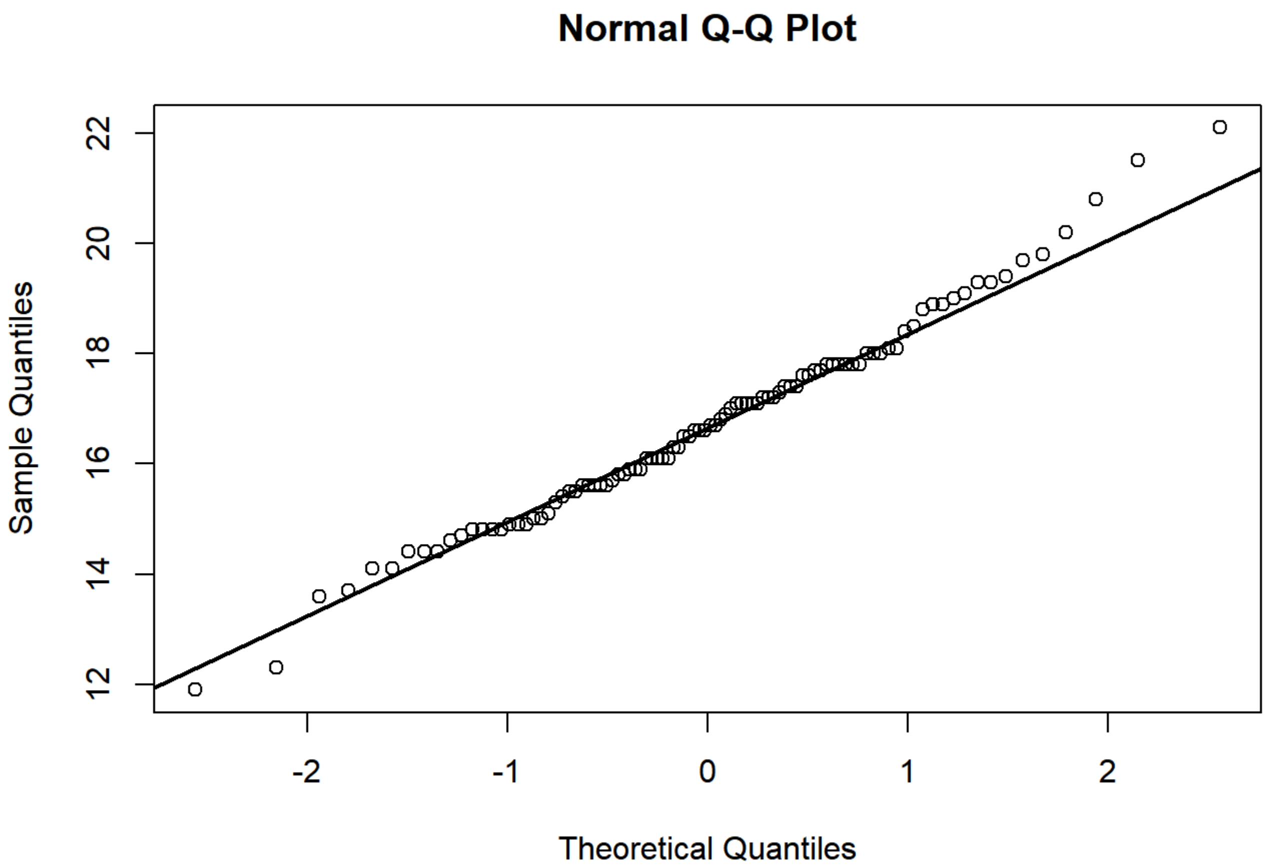


Question 1(g)

Use R to produce a normal QQ plot.

Solution

```
qqnorm(y)
qqline(y, col = "black", lwd = 2)
```



Question 1(h)

Does the data look like it came from a normal population?l

Solution

From the normality plot (i.e., qq - plot), the data form an approximately straight line along the normality line. Hence, it is safe to conclude that the data came from a normal population.

Question 1(i)

Recall that since $\bar{X} \approx N(\mu_{\bar{X}}, \sigma^2_{\bar{X}}) = N(\mu, \frac{\sigma^2}{n})$, the $SE(\hat{\mu}) = \frac{s}{\sqrt{n}}$. Compute a 90% confidence interval on the true mean plunger force.

Solution

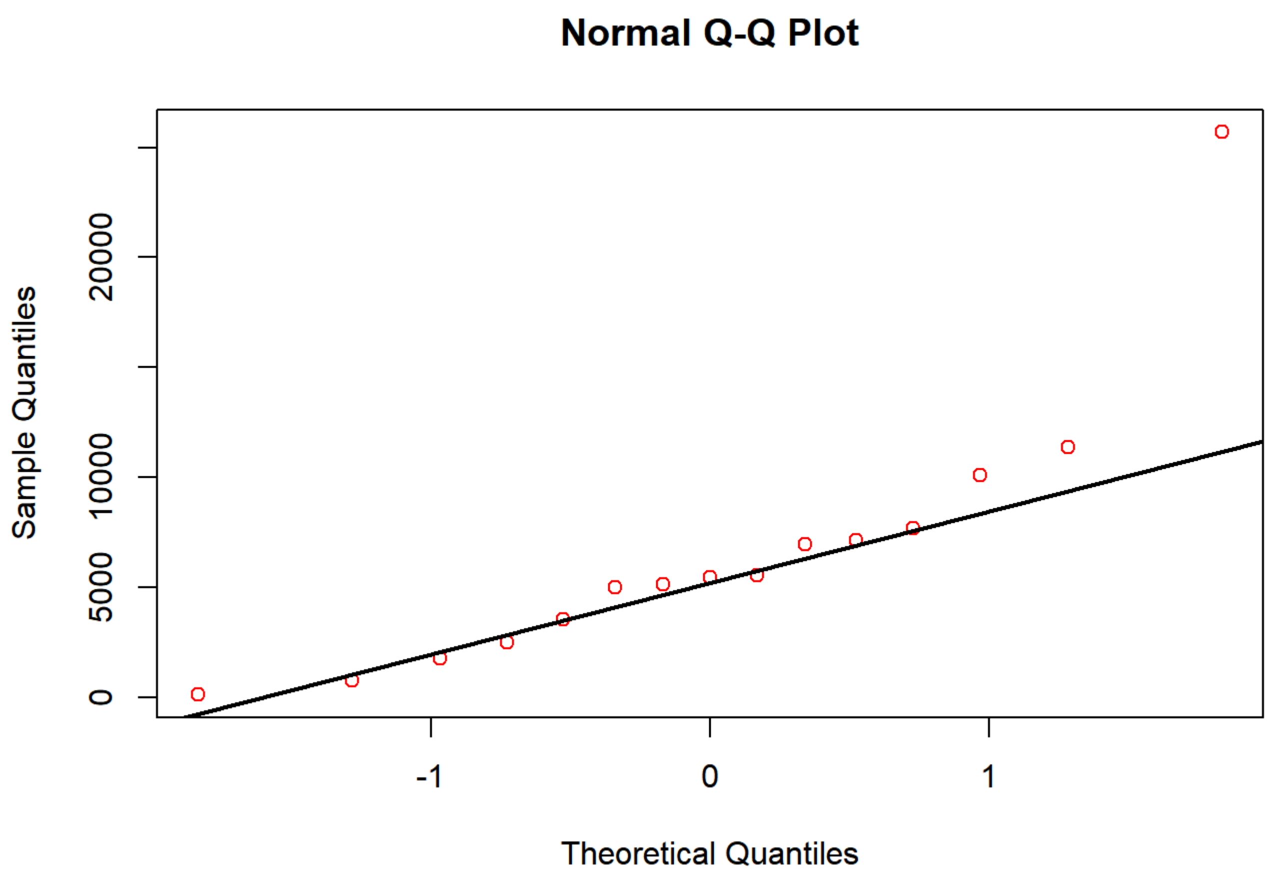
$\sum_{i=1}^{n} x_i = 1603.8$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{x} = \frac{1}{96} * 1603.8 = 16.706$, $\sum_{i=1}^{n} x_i^2 = 27119.48$,
$S = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{n}}$, $S = 1.843$, $t_{\frac{\alpha}{2}, n-2} = t_{0.05,94} = 0.397$, $S_n = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{n}}$,
$\sum_{i=1}^{n} x_i^2 = 27119.48$, $S_n = 1.843$

Interval: $\bar{x} \pm t_{\frac{\alpha}{2}, n-2} * \frac{S_n}{\sqrt{n}}$ = $16.706 \pm (-0.312)$,

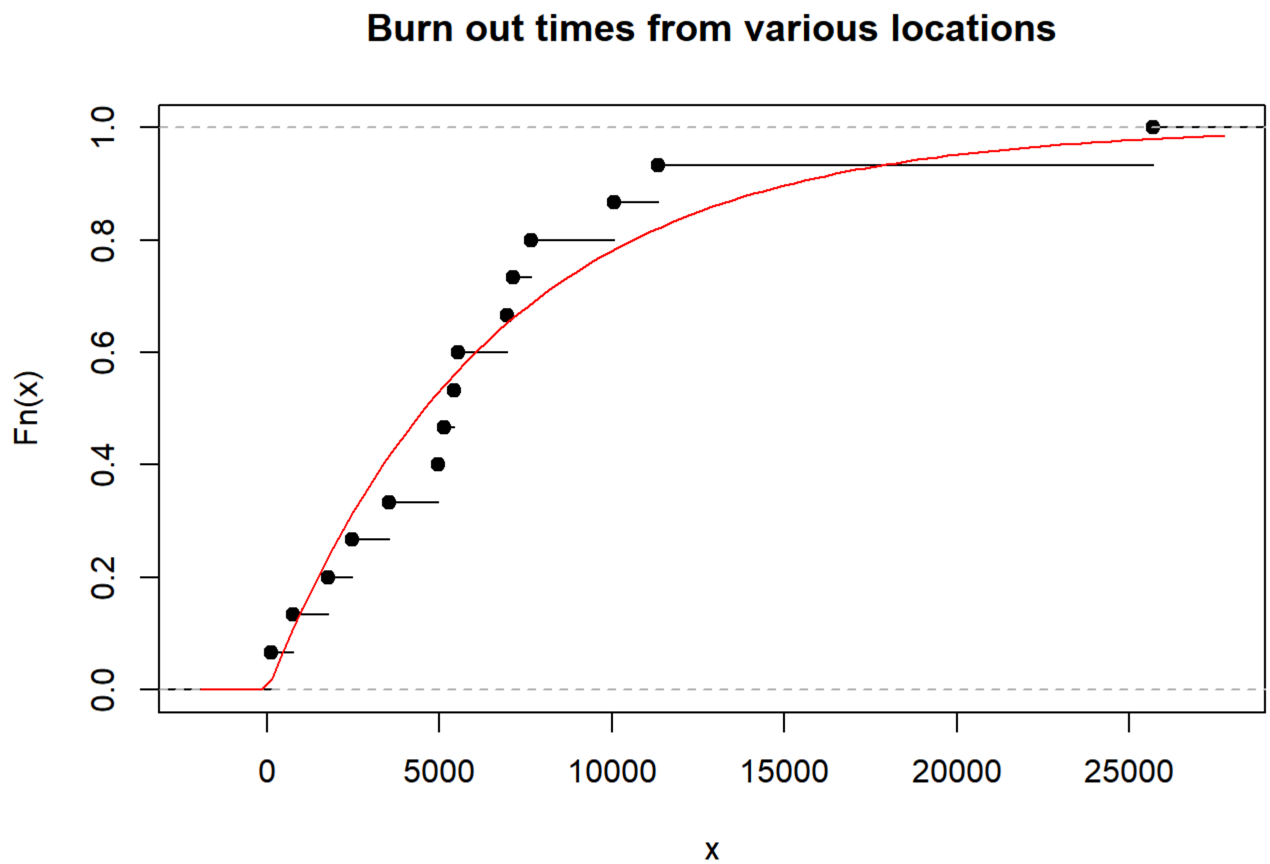
$(16.394 \leqslant \bar{x} \leqslant 17.018$

Question 2(e)

```
q2 = c(106, 4972, 7140, 7661, 1776, 2471, 5550, 6959, 3541, 747, 5142, 25691, 11345, 10067, 5434
)
qqnorm(q2, col = 2)
qqline(q2, col = "black", lwd = 2)
```




```
plot(ecdf(q2), main = "Burn out times from various locations")
curve(pexp(x,0.000152), add=TRUE, col= 2)
```



From the plot of the empiral CDF, it can be deduced that the sample data is from an exponential population data.