(1) $\bar{y} = \frac{\sum y_i}{n} = \frac{\sum \alpha + \beta x_i}{n} = \frac{\sum \alpha + \sum \beta x_i}{n} = \frac{\alpha n}{n} + \beta\left[\frac{\sum x_i}{n}\right] = \bar{x}$

$\Rightarrow \frac{\alpha n}{n} + \beta\bar{x} = \alpha + \beta\bar{x} \Rightarrow \boxed{\bar{y} = \alpha + \beta\bar{x}}$

Let $\hat{x}$ be median of sample $x_1 \dots x_n$, under linear transformation

the ~~~~ median value adjusts linearly as well (since ordering doesn't change)

thus $\boxed{\hat{y} = \alpha + \beta\hat{x}}$

$\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N}} = \sqrt{\frac{\sum(\alpha + \beta x_i - (\alpha + \beta\bar{x}))^2}{N}} = \sqrt{\frac{\sum(\beta x_i - \beta\bar{x})^2}{N}}$

$\Rightarrow = \sqrt{\frac{\sum \beta^2 (x_i - \bar{x})^2}{N}} = |\beta|\sqrt{\frac{\sum(x_i - \bar{x})^2}{N}} = |\beta\sigma_x| \Rightarrow \boxed{s_y = |\beta s_x|}$

$IQR_y \Rightarrow IQR_x = x_{75} - x_{25}$, under a linear transformation (even negative multiplication) x_{75} and x_{25} are still the edge values of the ~~50~~ center 50% of sample, since linear transformation can only reverse order not change,

thus $IQR_y = y_{75} - y_{25} = (\alpha + \beta x_{75}) - (\alpha + \beta x_{25}) = (\beta x_{75} - \beta x_{25}) = \beta(x_{75} - x_{25})$

$= \beta(IQR_x) \Rightarrow \boxed{y = |\beta| \cdot IQR_x}$

IQR

(6) $\rightarrow$ all sum

(2)

(a) $\bar{x} = \underset{\alpha}{\arg\min} \sum_{i=2}^{n} (x_i - \alpha)^2$, we will show this to be true by showing $\alpha = \bar{x}$ does indeed minimize the function. Notice the function is quadratic wrt α so a single minimum exists for $\sum_{i=2}^{n} (x_i - \alpha)^2$, differentiating we get $\frac{df}{d\alpha} = -\sum_{i=2}^{n} 2(x_i - \alpha) = -2\sum_{i=2}^{n}(x_i - \alpha)$

$\Rightarrow = -2\left[\sum_{i=2}^{n} x_i - \sum_{i=2}^{n} \alpha\right] = -2\left[n\bar{x} - n\alpha\right]$, to find the minimal value we just need to set $\frac{df}{d\alpha} = 0 \Rightarrow -2\left[n\bar{x} - n\alpha\right] = 0 \Rightarrow \underline{\alpha = \bar{x}}$ so we prove that $\underset{\alpha}{\arg\min} \sum_{i=2}^{n} (x_i - \alpha)^2 = \bar{x}$. 2^{nd} derivative $\Rightarrow \frac{d^2}{d\alpha^2} = 2 > 0$ thus MINIMUM

(b) $\tilde{x} = \underset{\alpha}{\arg\min} \sum_{i=2}^{n} |x_i - \alpha|$, we can also prove this by showing $\alpha = \tilde{x}$ does indeed minimize the function. Notice the absolute value function with only a linear term $|x_i - \alpha|$ only has one minimum, which we can find by differentiating $f = \sum |x_i - \alpha| \Rightarrow \frac{df}{d\alpha} = \sum \text{sign}(x_i - \alpha) = 0$, notice for $\frac{df}{d\alpha} = 0$ we need α s.t. $\sum_i \text{sign}(x_i - \alpha) = 0$, this is true for $\alpha = \tilde{x}$ since then half of the summative terms are -1 and the other half $+1$, resulting in their sum totalling to 0 (for odd cases the $x_i = \tilde{x}$ term $= 0$). Thus we see $\alpha = \tilde{x}$ sets the $\frac{df}{d\alpha} = 0$ meaning $f = \sum |x_i - \alpha|$ is minimized at $\alpha = \tilde{x}$ proving the equality.

meaning still equal terms of $-1, 1$

considering both odd/even cases

(3) If the QQ-Plot for a normal quantile on sample $\{X_1 \dots X_n\}$ falls on a line of the form $y=ax+b$ instead of $y=x$, it suggests that the distribution sampled from is still normally distributed but with a different spread and a shift of the mean or center from the standard normal distribution. This can be explained by how if the QQ-plot against the normal quantile falls on a line the quantiles are still linearly spread out and indicate the shape of the sample distribution to be the same (if the sample was divided by a and b subtracted) it would fall on $y=x$, ~~and~~ and since a and b are linear operators the distribution shape doesn't change so if via linear operations we can reach $y=x$ for a standard normal, the data on $y=ax+b$ is also normal just not standard normal. Specifically a lessens the spread if $|a|<1$ and widens the spread if $|a|>1$, but since the relative distance of points doesn't change under linear/scalar multiplication the distribution shape stays the same. Similarly adding b only shifts the distribution left or right, not its shape. In conclusion, if the QQ-Plot against normal quantile, has a sample falling on $y=ax+b$, it is still normally distributed but not the standard normal, as it is scaled by a and center/mean-shifted by b.

std scaled by a

(6)

(a) $P(s_1 = N) = \frac{1}{N}$, $P(s_i = N) = \frac{1}{N-(i-1)} \prod_{j=2}^{i-1} \frac{N-j}{N-(j-2)}$

$\hookrightarrow = \frac{1}{N}$

$\boxed{P(s_i = N) = 1/N}$

(b) $P(N \text{ in sample}) = 1 - P(N \text{ not in sample})$

$= 1 - \prod_{j=2}^{n} \frac{\cancel{N-j}\, N-j}{N-(j-2)} = 1 - \frac{N-n}{N} = \boxed{\frac{n}{N} = P(N \text{ in sample})}$

(c) $\mathbb{E}(s_2) = \sum P_i \cdot p(P_i) = \frac{1}{N} \sum P_i = \frac{1}{N}\left(\frac{N(N+1)}{2}\right)$

$\Rightarrow \boxed{\mathbb{E}(s_2) = (N+1)/2}$

(d) $P(s_1 = N, s_2 = 2) = \frac{1}{N} \cdot \frac{1}{N-1} = \boxed{\frac{1}{N^2 - N}}$

(e) $\boxed{\mathbb{P}(s_i = i, \forall i \in [1, n]) = \frac{1}{\left(\frac{N!}{(N-n)!}\right)}}$

$\hookrightarrow = \frac{(N-n)!}{N!}$

(7a) For an estimate to be unbiased:

$\mathbb{E}(\bar{\mu}) = \mu$, thus we need to find conditions on weights w_i such that the equality holds: for $\bar{X}_n^w = \bar{\mu} = \sum_{i=1}^{n} w_i x_i$:

$\mathbb{E}(\bar{\mu}) = \mathbb{E}\left[\sum_{i=1}^{n} w_i x_i\right]$, by linearity of expectation this rearranges to $\sum_{i=1}^{n} w_i \boxed{\mathbb{E}[x_i]}$

from which we know $\mathbb{E}[x_i] = \mu$, thus

$\mathbb{E}(\bar{\mu}) = \mu \sum_{i=1}^{n} w_i$, to set this as unbiased:

$\mathbb{E}(\bar{\mu}) = \mu \Rightarrow \mu \sum_{i=1}^{n} w_i = \mu \Rightarrow \sum_{i=1}^{n} w_i = 1$,

so we show for the weighted sample mean $\bar{X}_n^w$ to be unbiased, $\sum_{i=1}^{n} w_i = 1$ meaning the sum of all weights $= 1$, is the condition for the estimate to be unbiased.

(7b) Notice given $SE(\overline{x_n^w}) = SE(\sum w_i x_i) = \sqrt{\mathbb{V}[\sum w_i x_i]}$

thus to minimize $SE(\overline{x_n^w})$ we can minimize

$\mathbb{V}[\sum w_i x_i]$ since the terms are proportional.

$\mathbb{V}[\sum w_i x_i]$ is given by $= \sum_{i,j} w_i w_j \operatorname{cov}(x_i, x_j)$

$= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j \operatorname{cov}(x_i, x_j)$, for $j=i$ $\operatorname{cov}(x_i, x_j) = \mathbb{V}(x_i)$

$\Rightarrow = \sum_{i=1}^{n} \left[w_i^2 \mathbb{V}[x_i] + w_i \sum_{j \neq i} w_j \operatorname{cov}(x_i, x_j) \right]$

given for an SRS $\mathbb{V}[x_i] = \sigma^2$ and $\operatorname{cov}(x_i, x_j) = \frac{-\sigma^2}{N-1}$

we simplify to: $\sum_{i=1}^{n} \left[w_i^2 \sigma^2 + w_i \sum_{j \neq i} w_j \left(\frac{-\sigma^2}{N-1} \right) \right]$

$= \sum_{i=1}^{n} \left[w_i^2 \sigma^2 + w_i \left(\frac{-\sigma^2}{N-1} \right) \sum_{j \neq i} w_j \right]$ $\quad \sum_{j \neq 1}^{n} w_j = 1 - w_i$ under constraints of (a)

$= \sum_{i=1}^{n} f(\sigma, N) w_i^2 + g(\sigma, N) w_i$

$= f(\sigma, N) \sum_{i=1}^{n} w_i^2 + g(\sigma, N) \sum_{i=1}^{n} w_i = f(\sigma, N) \sum w_i^2 + g(\sigma, N)$

Thus $\mathbb{V}[\sum w_i x_i] \propto \sum_{i=1}^{n} w_i^2$, which under $\sum w_i = 1$

is minimized at $\boxed{w_i = 1/n}$

```
clc; clear; close all;
```

Problem 5

Importing data

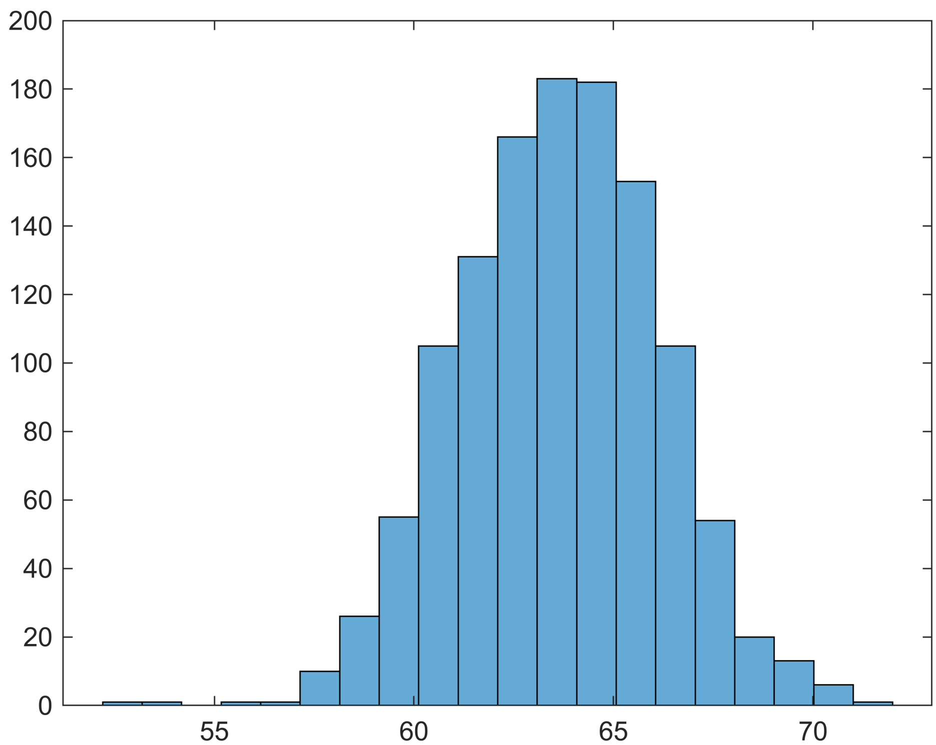
```
data = textread("C:\Users\sayuj\OneDrive - California Institute of
Technology\ACM_157\birth.txt");
```

Normalized histogram

Part (A)

```
heights = data(:,5);
heights = heights(heights ~= 99);

hist = histogram(heights, 20);
```



The histogram of the mothers' height is provided above, nbins = 20 appears to be the ideal parameter to show the symmetric bell curved shape of the distribution.

Part (B)

```
mu = mean(heights)
```

```
mu = 64.0478
```

```
med = median(heights)
```

```
med = 64
```

```
stdev = std(heights)
```

```
stdev = 2.5334
```

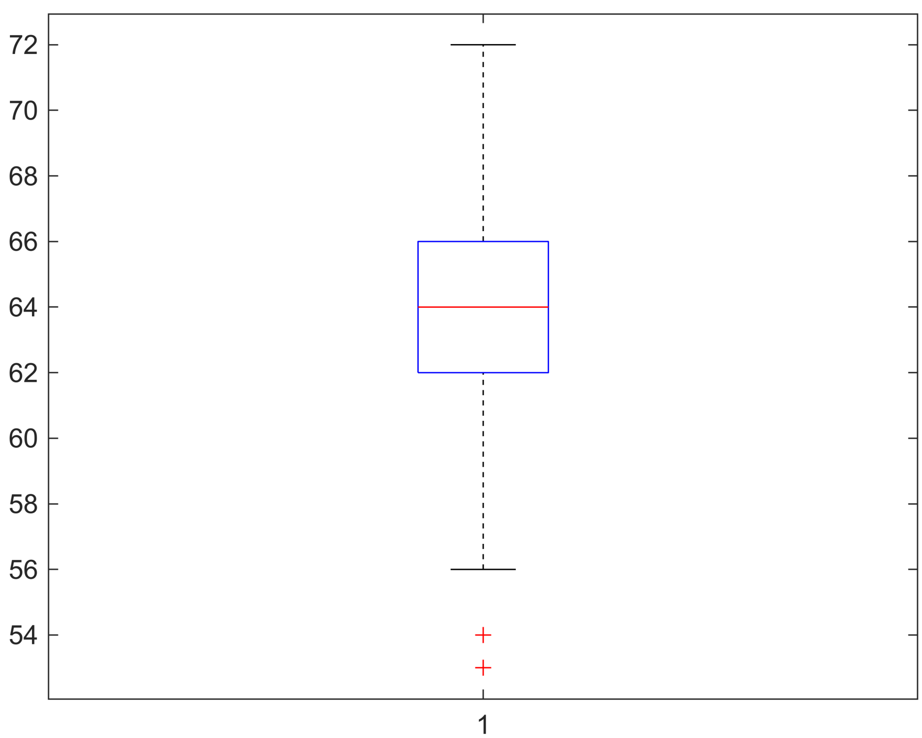
```
quart = iqr(heights)
```

```
quart = 4
```

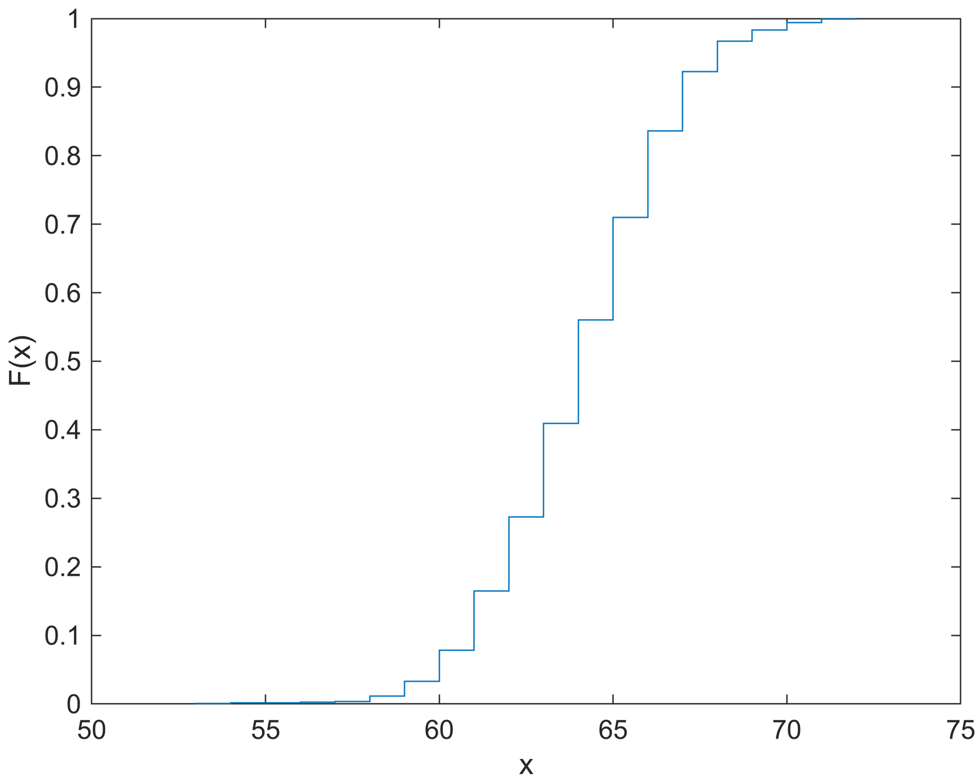
Above is the mean, median, standard deviationm and IQR of the mothers' heights, respectively. It appears that the center is well defined given that the mean and the median are nearly the same value and the standard deviation and interquartile range are relatively small in terms of the mean and median value and the context of heights (50% center of the distribution is in within 4 inches)

Part (C)

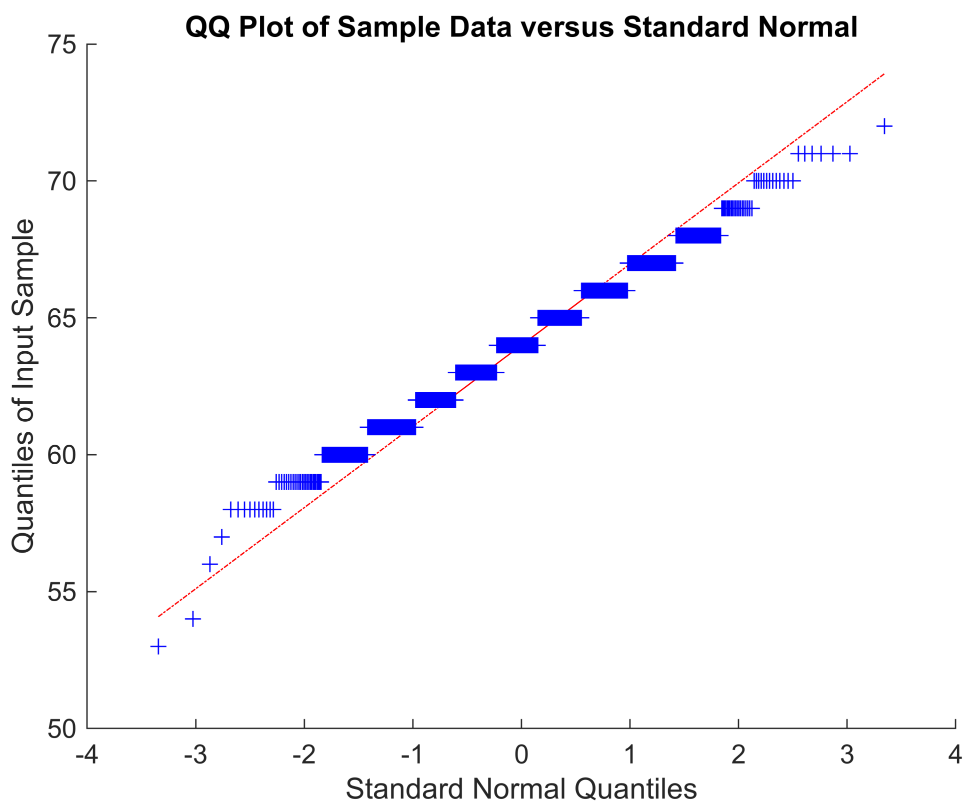
```
boxplot(heights)
```




```
ecdf(heights)
```




```
qqplot(heights)
```



The plots above provides the boxplot, the eCDF, and the normal QQ-plot of the mothers' height data. We can observe from the centered shape of the boxplot, the sigmoid shape of the eCDF, and the approximately linear (falls on the normal line) shape of the QQ-plot that the data appears to be normally distributed. The only concern is the slightly curved shape of the QQ-plot which may suggest is more concentrated around the mean (slimmer) than the normal bell curve (thin tailed). It still appears to have a normal distribution shape from all other plots so far, and appears to have a mean ~65 and variance of ~6.5 (stdev ^2).

Part (D)

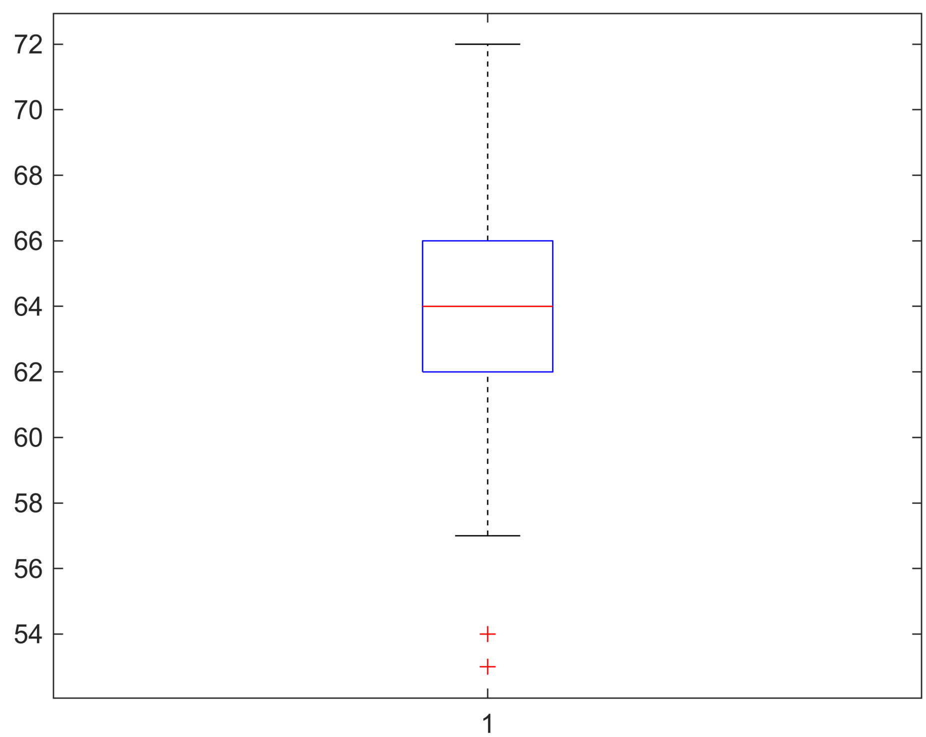
```
s_data = data(:,7);
s_data = s_data(data(:,5) ~= 99);

smoker_heights = heights(s_data == 1);

nonsmoker_heights = heights(s_data == 0);
```

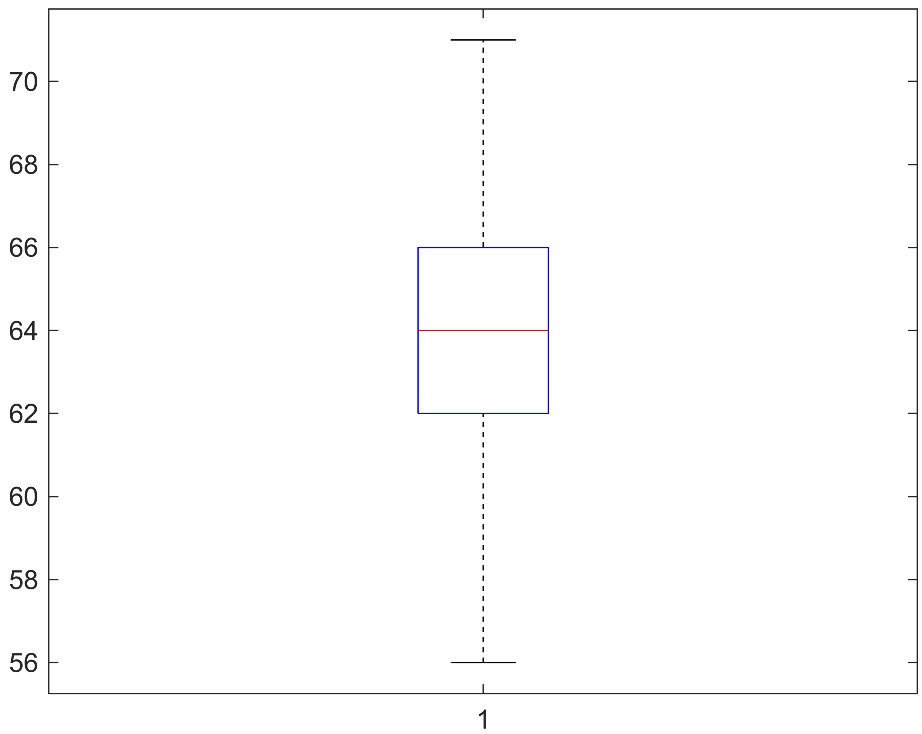
Smoker plots

```
boxplot(smoker_heights)
```



Non-smoker plots

```
boxplot(nonsmoker_heights)
```



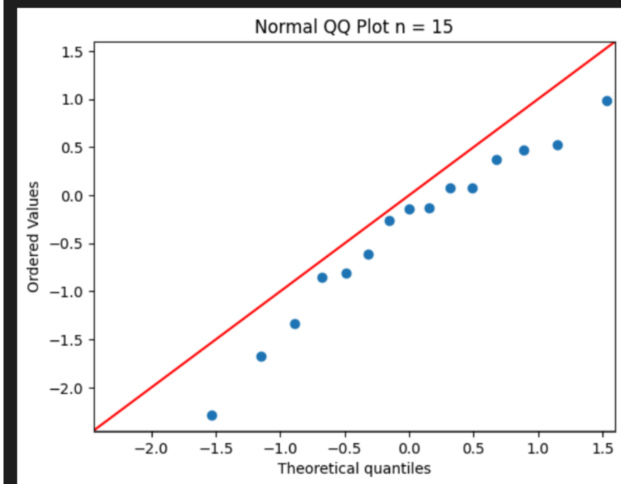
The boxplots of smoker heights vs non-smoker heights above do not appear to have a clear difference between the mean height or the range of heights. The smoker heights and non-smoker heights have very similar IQR's and quartile points, a small difference being the smoker heights having 2 outliers whereas the non smokers do not. Thus, it is difficult to conclude with convincing evidence that the heights between the tow groups vary.

Part(A)

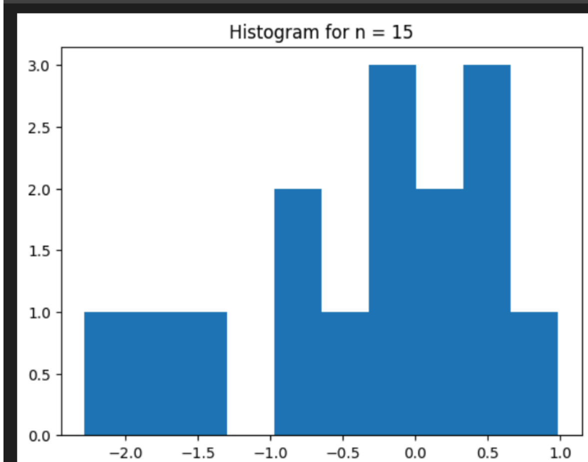
```
"""
It is diffucult to determine if the QQ plots fall on a straight line as the plots vary largely across iterations likely due to small sample size.
In general the points across many iterations average over the line but no singular plot directly falls on the line with n = 15. The histogram
is not unimodal mostly and has gaps across the range of samples, it also lacks symmetry/ bell shape given the small n.
"""
```

✓ 2.1s

```
<Figure size 640x480 with 0 Axes>
```




```
'\nIt is diffucult to determine if the QQ plots fall on a straight line as the plots vary largely across iterations likely due to small sample size.\nIn g
```

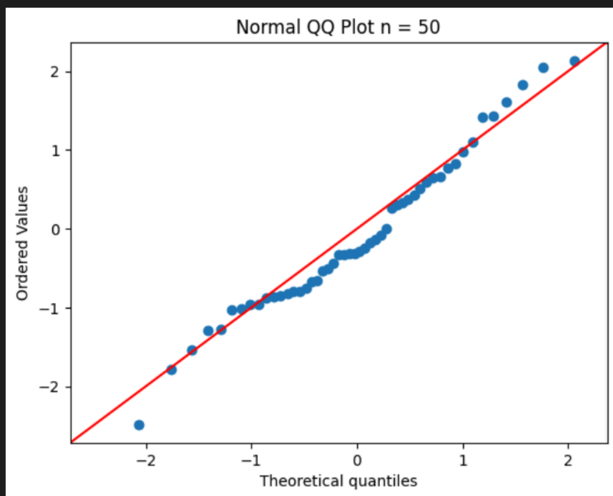


Part(B)

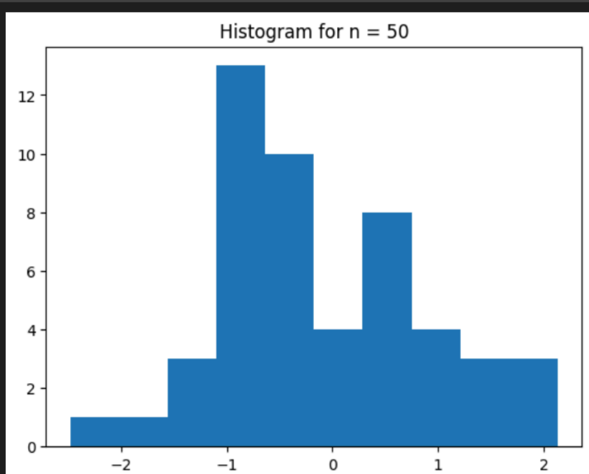
```
The QQ plot is more tightly close to the normal line but does not fit it to an extent that we can confidently conclude a normal distribution.
In addition, the histogram appears to be approaching a normal bell shape but still is not unimodal or symmetric.
'''
```

✓ 1.4s

```
<Figure size 640x480 with 0 Axes>
```




```
'\nThe QQ plot is more tightly close to the normal line but does not fit it to an extent that we can confidently conclude a normal distribution.\nIn
```

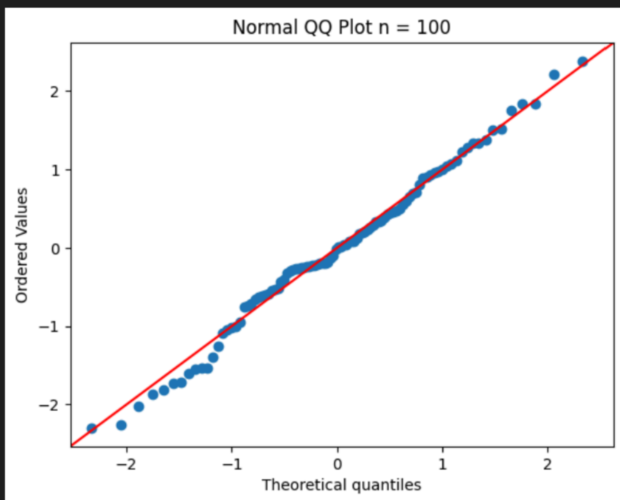


Part(B)

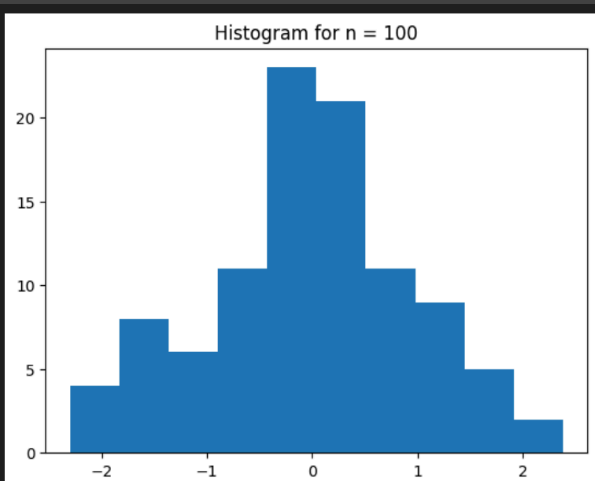
```
'''
The QQ plot now more tightly fits the normal line signalling a strong likelihood of the sample population being normal.
The histrogram is almost unimodal and closer to a bell curve although not fully symmetric. This seems to be approaching solid threshold of
n* to be confident in determining a normally distributed population from the sample.
'''
```

✓ 1.5s

```
<Figure size 640x480 with 0 Axes>
```




```
'\nThe QQ plot now more tightly fits the normal line signalling a strong likelihood of the sample population being normal.\nThe histrogram is unim
```

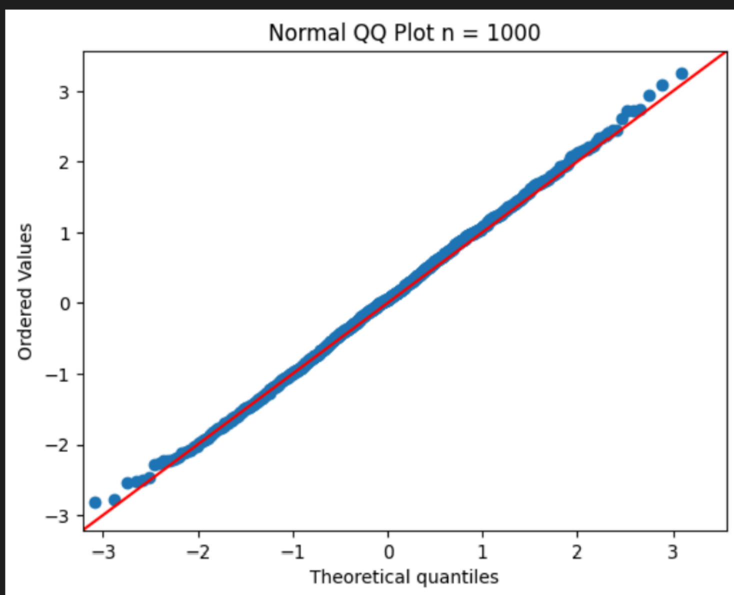


Part(B)

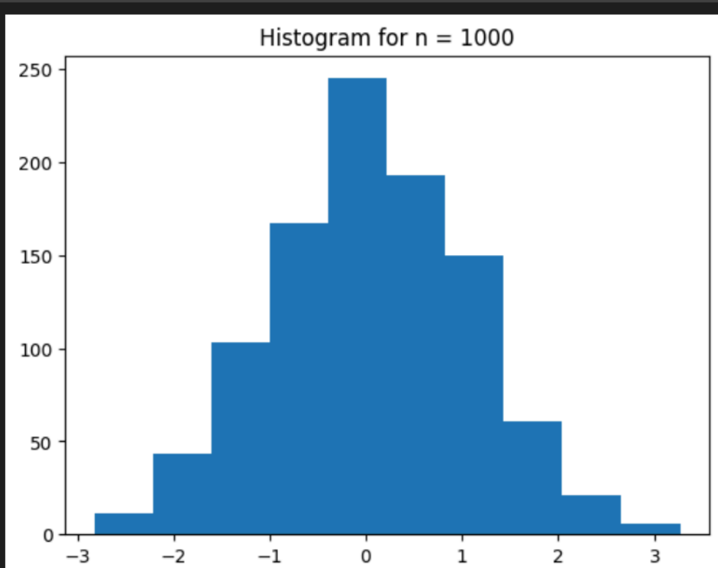
```
This QQ plot directly falls on the normal line except at end points providing strong support that the population
sampled form is normally distributed. The histogram is unimodal, symmetric and bell curved shaped further supporting
the evidence of a normal distribution.
'''
```

✓ 1.7s

```
<Figure size 640x480 with 0 Axes>
```




```
'\nThis QQ plot directly falls on the normal line except at end points providing strong support that the population\nsampled
```



```
'''
Part C

I would estimate  n* = 100 to be a solid threshold beyond which sample qq-plots and histograms are stable enough to make conclusions
on the distribution of the population sampled. Observed in the experiments on the data, n = 5 and  n = 50 produce qq plots that vary
in behavior along with the histograms where as for n = 100 and above the qq-plots fall consistently on the normal line and the histogram
consistently appears to follow the normal distribution curve, thus from that data it is reasonable to conclude n* = 100 is a threshold
for which the stability of the qq-plots and histograms to provide conclusive information on the population holds.
'''
```