Statistical Interference
Set 1

Caltech IDS/ACM/CS 157
April 2022

Dallas Taylor

1 Problem 1

In this problem, you will explore how data summaries change under liner transformation of the data. Consider a sample $x_1, ..., x_n$. Let $\bar{x}$, $\tilde{x}$, s_x, and IQR_x denote the sample mean, median, standard deviation, and the interquartile range of the sample. Let

$$y_i = \alpha + \beta x_i$$

Express $\bar{y}$, $\tilde{y}$, s_y, and IQR_y in terms of $\bar{x}$, $\tilde{x}$, s_x, and IQR_x.

Solution : *Let us first examine $\bar{x}$ and $\bar{y}$:*

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} \alpha + \beta x_i \\
\therefore \bar{y} &= \frac{1}{n}\left(\sum_{i=1}^{n} \alpha + \sum_{i=1}^{n} \beta x_i\right) \\
\therefore \bar{y} &= \frac{1}{n}\left(\alpha n + \beta \sum_{i=1}^{n} x_i\right) \\
\therefore \bar{y} &= \frac{1}{n}\left(\alpha n + \beta n \bar{x}\right) \\
\therefore \bar{y} &= \alpha + \beta \bar{x}
\end{aligned}$$

Now, let us examine $\tilde{x}$ and $\tilde{y}$:

$$
\begin{aligned}
\tilde{x} &= \begin{cases} x_{(\frac{n+1}{2})} & n \textit{ is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \textit{ is even} \end{cases} \\
\tilde{y} &= \begin{cases} y_{(\frac{n+1}{2})} & n \textit{ is odd} \\ \frac{1}{2}(y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}) & n \textit{ is even} \end{cases} \\
\therefore \tilde{y} &= \begin{cases} \alpha + \beta(x_{(\frac{n+1}{2})}) & n \textit{ is odd} \\ \frac{1}{2}\left[(\alpha + \beta(x_{(\frac{n}{2})})) + (\alpha + \beta(x_{(\frac{n}{2}+1)}))\right] & n \textit{ is even} \end{cases} \\
\therefore \tilde{y} &= \begin{cases} \alpha + \beta(x_{(\frac{n+1}{2})}) & n \textit{ is odd} \\ \frac{1}{2}\left[2\alpha + \beta(x_{(\frac{n}{2})}) + \beta(x_{(\frac{n}{2}+1)})\right] & n \textit{ is even} \end{cases} \\
\therefore \tilde{y} &= \begin{cases} \alpha + \beta(x_{(\frac{n+1}{2})}) & n \textit{ is odd} \\ \alpha + \frac{\beta}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \textit{ is even} \end{cases} \\
\therefore \tilde{y} &= \alpha + \beta\left(\begin{cases} x_{(\frac{n+1}{2})} & n \textit{ is odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & n \textit{ is even} \end{cases}\right) \\
\therefore \tilde{y} &= \alpha + \beta\tilde{x}
\end{aligned}
$$

Now let us examine s_x and s_y:

$$\begin{aligned}
s_x &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2} \\
\therefore s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((\alpha+\beta x_i)-(\alpha+\beta\bar{x}))^2} \\
\therefore s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\beta x_i-\beta\bar{x})^2} \\
\therefore s_y &= \sqrt{\frac{(\beta)^2}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
\therefore s_y &= \beta\left(\sqrt{\frac{(1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \\
\therefore s_y &= \beta s_x
\end{aligned}$$

Now let us examine IQR$_x$ and IQR$_y$. Let us fi where we notice that the x_i that corresponds with with the 1st quartile is the mean of all the lowest 50% of data (and 3rd is 75%):

$$
\begin{aligned}
IQR_x &= \begin{cases} \frac{1}{2}\left[\left(x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}\right) - \left(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}\right)\right] & n, \frac{n}{2} \textit{ are even} \\ x_{(\frac{3n+2}{4})} - x_{(\frac{n+2}{4})} & n \textit{ is even, } \frac{n}{2} \textit{ is odd} \\ \frac{1}{2}\left[\left(x_{(\frac{3n+1}{4})} + x_{(\frac{3n+1}{4}+1)}\right) - \left(x_{(\frac{n-1}{4})} + x_{(\frac{n-1}{4}+1)}\right)\right] & n \textit{ is odd, } \frac{n+1}{2} \textit{ is even} \\ x_{(\frac{3n+3}{4})} - x_{(\frac{n+1}{4})} & n, \frac{n+1}{2} \textit{ are odd} \end{cases} \\
IQR_y &= \begin{cases} \frac{1}{2}\left[\left(y_{(\frac{3n}{4})} + y_{(\frac{3n}{4}+1)}\right) - \left(y_{(\frac{n}{4})} + y_{(\frac{n}{4}+1)}\right)\right] & n, \frac{n}{2} \textit{ are even} \\ y_{(\frac{3n+2}{4})} - y_{(\frac{n+2}{4})} & n \textit{ is even, } \frac{n}{2} \textit{ is odd} \\ \frac{1}{2}\left[\left(y_{(\frac{3n+1}{4})} + y_{(\frac{3n+1}{4}+1)}\right) - \left(y_{(\frac{n-1}{4})} + y_{(\frac{n-1}{4}+1)}\right)\right] & n \textit{ is odd, } \frac{n+1}{2} \textit{ is even} \\ y_{(\frac{3n+3}{4})} - y_{(\frac{n+1}{4})} & n, \frac{n+1}{2} \textit{ are odd} \end{cases} \\
\therefore IQR_y &= \begin{cases} \frac{1}{2}\left[2\alpha + \beta(x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}) - 2\alpha - \beta(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)})\right] & n, \frac{n}{2} \textit{ are even} \\ \alpha + \beta(x_{(\frac{3n+2}{4})}) - \alpha - \beta(x_{(\frac{n+2}{4})}) & n \textit{ is even, } \frac{n}{2} \textit{ is odd} \\ \frac{1}{2}\left[2\alpha + \beta(x_{(\frac{3n+1}{4})} + x_{(\frac{3n+1}{4}+1)}) - 2\alpha - \beta(x_{(\frac{n-1}{4})} + x_{(\frac{n-1}{4}+1)})\right] & n \textit{ is odd, } \frac{n+1}{2} \textit{ is even} \\ \alpha + \beta(x_{(\frac{3n+3}{4})}) - \alpha - \beta(x_{(\frac{n+1}{4})}) & n, \frac{n+1}{2} \textit{ are odd} \end{cases} \\
\therefore IQR_y &= \begin{cases} \frac{\beta}{2}\left[\left(x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}\right) - \left(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}\right)\right] & n, \frac{n}{2} \textit{ are even} \\ \beta(x_{(\frac{3n+2}{4})} - x_{(\frac{n+2}{4})}) & n \textit{ is even, } \frac{n}{2} \textit{ is odd} \\ \frac{\beta}{2}\left[\left(x_{(\frac{3n+1}{4})} + x_{(\frac{3n+1}{4}+1)}\right) - \left(x_{(\frac{n-1}{4})} + x_{(\frac{n-1}{4}+1)}\right)\right] & n \textit{ is odd, } \frac{n+1}{2} \textit{ is even} \\ \beta(x_{\frac{3n+3}{4}} - x_{\frac{n+1}{4}}) & n, \frac{n+1}{2} \textit{ are odd} \end{cases} \\
\therefore IQR_y &= \beta \begin{cases} \frac{1}{2}\left[\left(x_{(\frac{3n}{4})} + x_{(\frac{3n}{4}+1)}\right) - \left(x_{(\frac{n}{4})} + x_{(\frac{n}{4}+1)}\right)\right] & n, \frac{n}{2} \textit{ are even} \\ x_{(\frac{3n+2}{4})} - x_{(\frac{n+2}{4})} & n \textit{ is even, } \frac{n}{2} \textit{ is odd} \\ \frac{1}{2}\left[\left(x_{(\frac{3n+1}{4})} + x_{(\frac{3n+1}{4}+1)}\right) - \left(x_{(\frac{n-1}{4})} + x_{(\frac{n-1}{4}+1)}\right)\right] & n \textit{ is odd, } \frac{n+1}{2} \textit{ is even} \\ x_{\frac{3n+3}{4}} - x_{\frac{n+1}{4}} & n, \frac{n+1}{2} \textit{ are odd} \end{cases} \\
\therefore IQR_y &= \beta IQR_x
\end{aligned}
$$

Thus, as a reminder, we have determined that $\bar{y} = \alpha + \beta\bar{x}$, $\tilde{y} = \alpha + \beta\tilde{x}$, $s_y = \beta s_x$, and IQR$_y$ = βIQR$_x$.

2 Problem 2

The sample mean $\bar{x}$ and median $\tilde{x}$ have interesting optimization interpretations. Show that

$$\begin{aligned} \bar{x} &= \text{arg min}_\alpha \sum_{i=1}^{n} (x_i - \alpha)^2, && (1) \\ \tilde{x} &= \text{arg min}_\alpha \sum_{i=1}^{n} |x_i - \alpha|. && (2) \end{aligned}$$

Solution : *Let us first evaluate $\bar{x}$. Here, we can minimize the summation:*

$$\begin{aligned} 0 &= \sum_{i=1}^{n} -2(x_i - \alpha^*) \\ 0 &= \sum_{i=1}^{n} (x_i - \alpha^*) \\ 0 &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \alpha^* \\ \alpha^* n &= \sum_{i=1}^{n} x_i \\ \alpha^* &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \alpha^* &= \bar{x} \end{aligned}$$

The last line from above is trivial, as we find the definition of $\bar{x}$, as desired. Now, let us consider $\tilde{x}$:

$$0 = \sum_{i=1}^{n} \mathit{sign}(x_i - \alpha^*)$$

It is clear that the only instance in which the above can equal 0 is when there are an equal number of positive and negative values for $(x_i - \alpha^*)$. This can only occur when there are an equal number of $x_i < \alpha^*$ and $x_i > \alpha^*$, which is trivially the definition of $\tilde{x}$ and thus $\alpha^* = \tilde{x}$.

This can be shown to hold for two possible cases of our sample. Let us first consider the case where there are multiple values equal to the median (such as in the case $\{x_1, x_2, x_2, x_3, x_3\}$ where $x_1 < x_2 < x_3$). Thus, it is clear that the set of initial possible alpha$^* \in (x_1, x_3)$, however it is also clear that if $\alpha^* > 2$ that we produce $\sum_{i=1}^{n}$ sign$(x_i - \alpha*) = -1$ and if $\alpha^* < 2$ that we produce $\sum_{i=1}^{n}$ sign$(x_i - \alpha*) = +3$, and thus our function is *minimized at $\alpha* = 2$, as desired. It is clear that this process holds for any value of n, even or odd.*

When there are 0 or 1 values in our sample equal to the median, then it is trivial that our equality holds, as every x_i will cause sign$(x_i - \alpha*) = \pm 1$, with an equal amounts of x_i producing $+1$ as -1. This also holds for any value of n, even or odd (as only 1 sample equal to our median is indicative of n being trivially odd and 0 samples equal to our median is indicative of n being even and all values greater than 1 were dealt with above.

3 Problem 3

Let $x_1, ..., x_n$ be a sample. We know that if the normal-quantile plot, i.e. a collection of points $\{(z_{\frac{k}{n+1}}, x_{(k)})\}$, falls roughly on the line $y = x$, then the sample has approximately the standard normal distribution. What can you say about the distribution of the sample if points $\{(z_{\frac{k}{n+1}}, x_{(k)})\}$, instead if $y = x$, fall on the line $y = ax + b$?

Solution : *If the collection of points falls on the line $y = ax + b$, we can determine other information about the distribution. For instance, if we have that $a \neq 1$, then we can determine that the standard deviation of our sample varies from normal such that if $a > 1$, then our distribution is wider (larger range), and if $0 < a < 1$ then our distribution is thinner. If we have that $b \neq 0$, then we can determine that the mean of the sample distribution is higher (or lower if $b < 0$) than that of the normal distribution. We are thus able to say that our distribution follows $\mathcal{N}(b, a^2)$.*

4 Problem 4

Solution : *See attached scripts.*

5 Problem 5

Solution : *See attached scripts.*

6 Problem 6

This problem will test your basic understanding of simple random sampling. Let $\mathcal{P} = \{1, ..., N\}$ be the target population, and $\mathcal{S} = \{s_1, ..., s_n\}$ a simple random sample from $\mathcal{P}$. Compute the following and show your work.

Problem A: $\mathbb{P}(s_1 = N), ..., \mathbb{P}(s_n = N)$

Solution A: *We can clearly use Lemma 1 of Survey Sampling I to see the following:*

$$\mathbb{P}(s_1 = P_i) = \frac{n_i}{N} = \frac{1}{N},$$

since we clearly have that n_i (the number of population units valued P_i) is 1 for all i. Thus, we have the following:

$$\mathbb{P}(s_1 = N), ..., \mathbb{P}(s_n = N) = \frac{1}{N}, ..., \frac{1}{N}.$$

Problem B: $\mathbb{P}$(the N-th population unit is in the sample)

Solution B: *This is clearly equivalent to the summation of if any individual sample unit is the N-th population unit. In Part A, we solved for each of these values. Thus, we have:*

$$\begin{aligned} \mathbb{P}(\textit{the } N\textit{-th population unit is in the sample}) &= \sum_{i=1}^{n} \mathbb{P}(s_i = N) \\ &= \sum_{i=1}^{n} \frac{1}{N} \\ &= \frac{n}{N}. \end{aligned}$$

Problem C: $\mathbb{E}[s_1]$

Solution C: *We can also clearly use Lemma 1 here to say that:*

$$\mathbb{E}[s_1] = \mu = \sum_{i=1}^{N} \mathbb{P}(s_1 = i)(i) = \frac{1}{N} \sum_{i=1}^{N} i = \left(\frac{1}{N}\right) \left(\frac{N(N+1)}{2}\right) = \frac{N+1}{2}.$$

Problem D: $\mathbb{P}(s_1 = N, s_2 = 1)$

Solution D: *This statement is equivalent to finding the probability that $s_1 = N$ AND $s_2 = 1$. Thus, we produce the following:*

$$\mathbb{P}(s_1 = N, s_2 = 1) = \mathbb{P}(s_1 = N) \times \mathbb{P}(s_2 = 1) = \frac{1}{N} \times \frac{1}{N-1} = \frac{1}{N(N-1)}.$$

Problem E: $\mathbb{P}(s_i = i, \text{ for all } i = 1, ..., n)$

Solution E: *This statement is equivalent to finding the probability that $s_1 = 1$, ..., AND $s_n = n$. Thus, we produce the following:*

$$\mathbb{P}(s_i = i, \textit{ for all } i = 1, ..., n) = \mathbb{P}(s_1 = 1) \times ... \times \mathbb{P}(s_n = n) = \frac{1}{N} \times ... \times \frac{1}{N-n+1} = \frac{(N-n)!}{N!}.$$

Dallas Taylor

7 Problem 7

In lectures, we discussed the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimate of the population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$. Let us consider a more general class of estimators:

$$\bar{X}_n^w = \sum_{i=1}^{n} w_i X_i,$$

where w_i are some weights. Note that the sample mean $\bar{X}_n$ is a special case of the above with $w_i = \frac{1}{n}$.

Problem A: Under what condition on the weights, the estimate above is an unbiased estimate of μ?

Solution A: *We know that $\bar{X}_n^w$ is an unbiased estimate of μ when we have that $\mathbb{E}[\bar{X}_n^w] = \mu$. Thus, we can perform the following:*

$$\begin{aligned} \mathbb{E}[\bar{X}_n^w] &= \sum_{i=1}^{n} w_i X_i \\ \therefore \mu &= \sum_{i=1}^{n} w_i X_i \end{aligned}$$

From the above, it is clear that the condition that is required on the weights is the following:

$$\sum_{i=1}^{n} w_i = 1.$$

Problem B : Among all unbiased estimates of the form above, find the most efficient estimate (i.e. the estimate with the smallest standard error).

Solution B: *The most efficient estimate (the one with the smallest standard error) is when we have the following minimized:*

$$se[\bar{X}_n^w] = \sqrt{\mathbb{V}[\bar{X}_n^w]}.$$

Thus, we are minimizing $\bar{X}_n^w$:

$$\begin{aligned}
\mathbb{V}[\bar{X}_n^w] &= \sum_{i=1}^n \mathbb{V}[w_i X_i] + 2\sum_{i=1}^n \sum_{j=1}^n \mathit{Cov}(w_i x_i, w_j x_j) \\
&= \sum_{i=1}^n w_i^2 \mathbb{V}[X_i] + 2\sum_{i=1}^n \sum_{j=1}^n w_i w_j \mathit{Cov}(x_i, x_j) \\
&= \sum_{i=1}^n w_i^2 \sigma^2 + 2\sum_{i=1}^n \sum_{j=1}^n w_i w_j (\frac{-\sigma^2}{N-1}) \\
&= \sigma^2 \sum_{i=1}^n w_i^2 - (\frac{2\sigma^2}{N-1}) \sum_{i=1}^n \sum_{j=1}^n w_i w_j \\
&= \sigma^2 \sum_{i=1}^n w_i^2 - (\frac{\sigma^2}{N-1}) \left((\sum_{i=1}^n w_i)^2 - \sum_{i=1}^n w_i^2 \right) \\
&= \sigma^2 \sum_{i=1}^n w_i^2 - (\frac{\sigma^2}{N-1}) \left(1 - \sum_{i=1}^n w_i^2 \right) \\
&= \sigma^2 \sum_{i=1}^n w_i^2 - \frac{\sigma^2}{N-1} + \frac{\sigma^2 \sum_{i=1}^n w_i^2}{N-1} \\
&= \frac{\sigma^2 N \sum_{i=1}^n w_i^2}{N-1} - \frac{\sigma^2}{N-1}
\end{aligned}$$

It is clear that we are thus minimizing $\sum_{i=1}^n w_i^2$ as $\frac{\sigma^2 N}{N-1}$, $\frac{\sigma^2}{N-1}$ are positive constants. Thus, we can use Lagrange multipliers for minimization with conditional $\sum_{i=1}^n w_i = 1$:

$$\begin{aligned}
\mathcal{L}(w_1, w_2, ..., w_n, \lambda) &= \sum_{i=1}^{n} w_i^2 + \lambda\left(\sum_{i=1}^{n} w_i - 1\right) \\
\frac{\partial \mathcal{L}}{\partial w_k} &= 0 \\
\frac{\partial}{\partial w_k}\left[\sum_{i=1}^{n} w_i^2 + \lambda\left(\sum_{i=1}^{n} w_i - 1\right)\right] &= 0 \\
2w_k + \lambda &= 0 \\
w_k &= -\frac{\lambda}{2}
\end{aligned}$$

Therefore, we can see that our problem is minimized when each w_k is equal to some constant ($\frac{-\lambda}{2}$). The following is thus clear since each weight is the same:

$$\begin{aligned}
\sum_{i=1}^{n} w_i &= 1 \\
\sum_{i=1}^{n}(-\frac{\lambda}{2}) &= 1 \\
-\frac{n\lambda}{2} &= 1 \\
\lambda &= -\frac{2}{n} \\
\therefore w_1, ..., w_n &= -\frac{-\frac{2}{n}}{2} \\
\therefore w_1, ..., w_n &= \frac{1}{n}
\end{aligned}$$

This value of w_i can be verified to be the minimum through Cauchy-Schwarz:

$$\begin{aligned}
\left(\sum_{i=1}^{n} w_i(1)\right)^2 &\leq \sum_{i=1}^{n} w_i^2 \times \sum_{i=1}^{n}(1)^2 \\
\left(\sum_{i=1}^{n} w_i\right)^2 &\leq \sum_{i=1}^{n} w_i^2 \times n \\
(1)^2 &\leq \sum_{i=1}^{n} w_i^2 \times n \\
\frac{1}{n} &\leq \sum_{i=1}^{n} w_i^2 \\
\frac{1}{n} &\leq \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \\
\frac{1}{n} &\leq \frac{n}{n^2} = \frac{1}{n}
\end{aligned}$$

Thus, we clearly see that the minimized value of $w_1, ..., w_n$ is when all $w_i = 1/n \ \forall i$, and thus is our most efficient estimate.

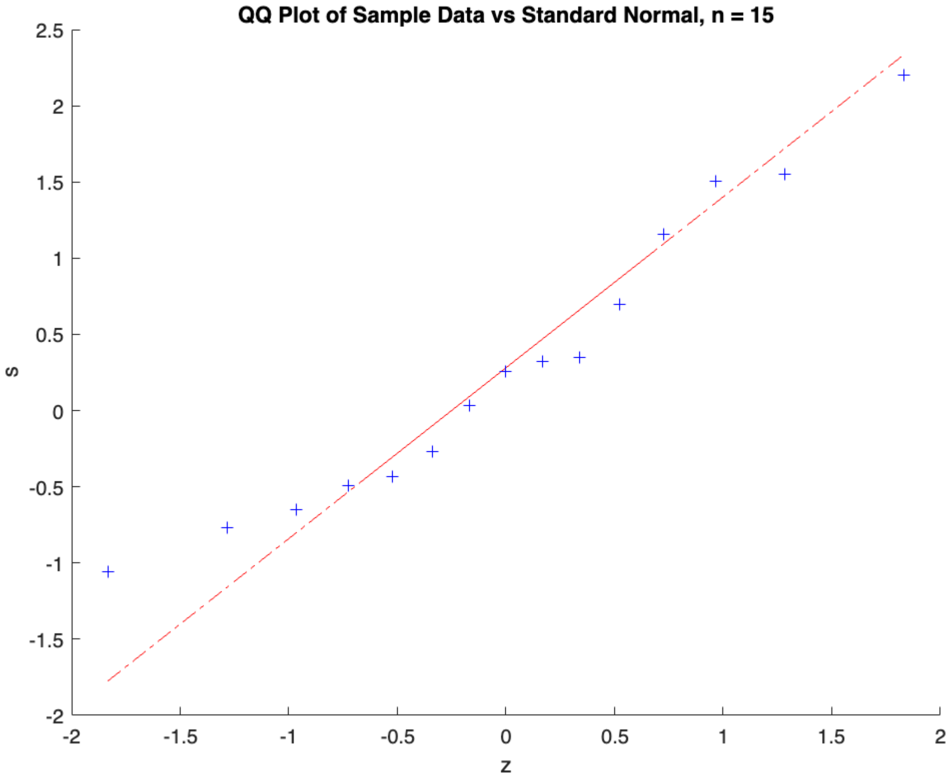
IDS 157: Problem 4

Part a

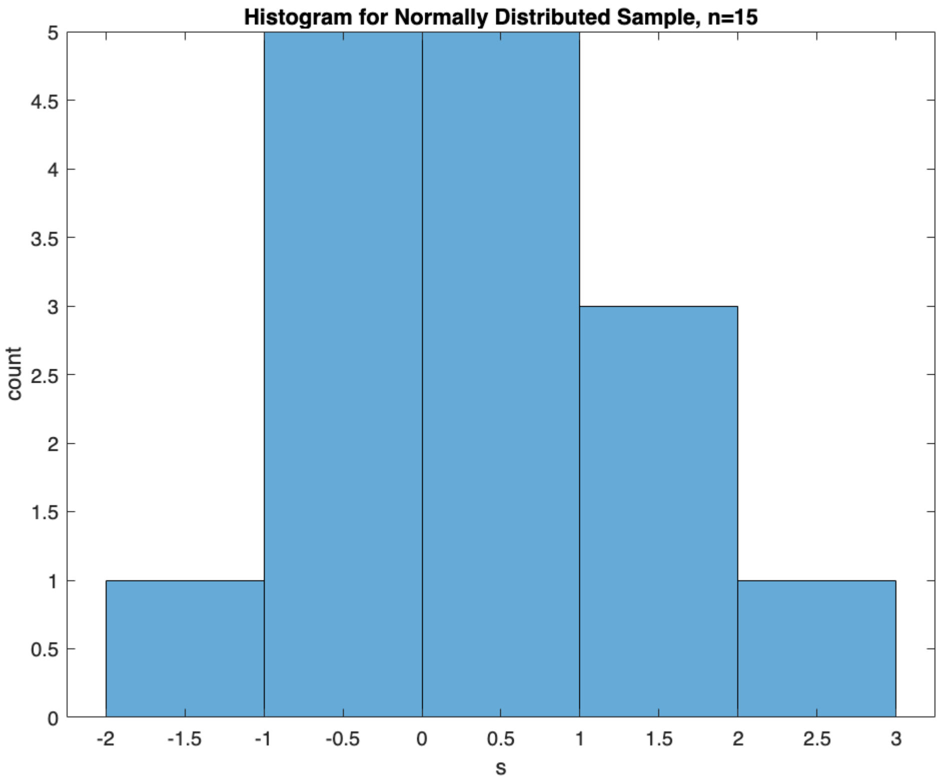
Draw a random sample of size $n = 15$ from $N(0,1)$ and plot both the normal-quantile plot and the histogram. Do the points on the QQ plot appear to fall on a straight line? Is the histogram symmetric, unimodal, and bell-shaped? Do this several times and summarize you observations.

```
% random sample
n = 15;
s = normrnd(0,1,[1,15]);

figure(1); clf
qqplot(s); xlabel('z'); ylabel('s'); title('QQ Plot of Sample Data vs Standard Normal,
```



```
figure(2); clf
histogram(s); xlabel('s'); ylabel('count'); title('Histogram for Normally Distributed S
```



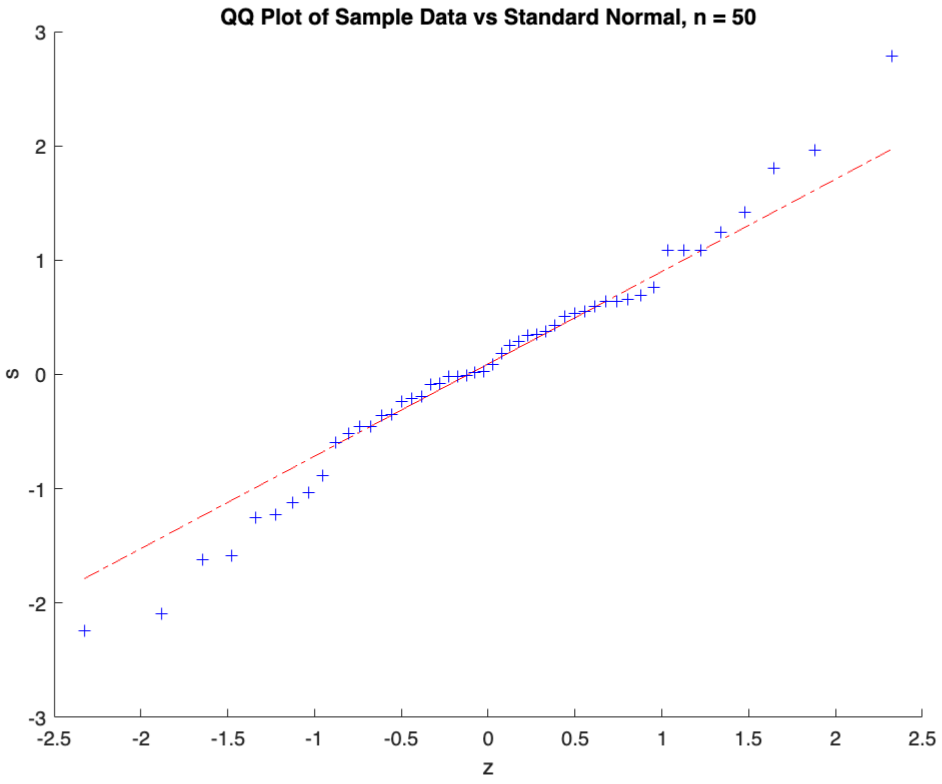
After several iterations of the above, I observed that most of the sample distributions closely matched the normal distribution benchmark, with most deviations that occurred occurring towards the tails, producing a distribution that more closely matched a different function y = ax + b. When considering the histogram, you see that for nearly normal distributions produce mostly symmetric, unimodal, and bell-shaped curves. When considering deviating distributions, you see skewing of the bell-shape (and thus loss of symmetry) and sometimes a loss of unimodality.

Part b

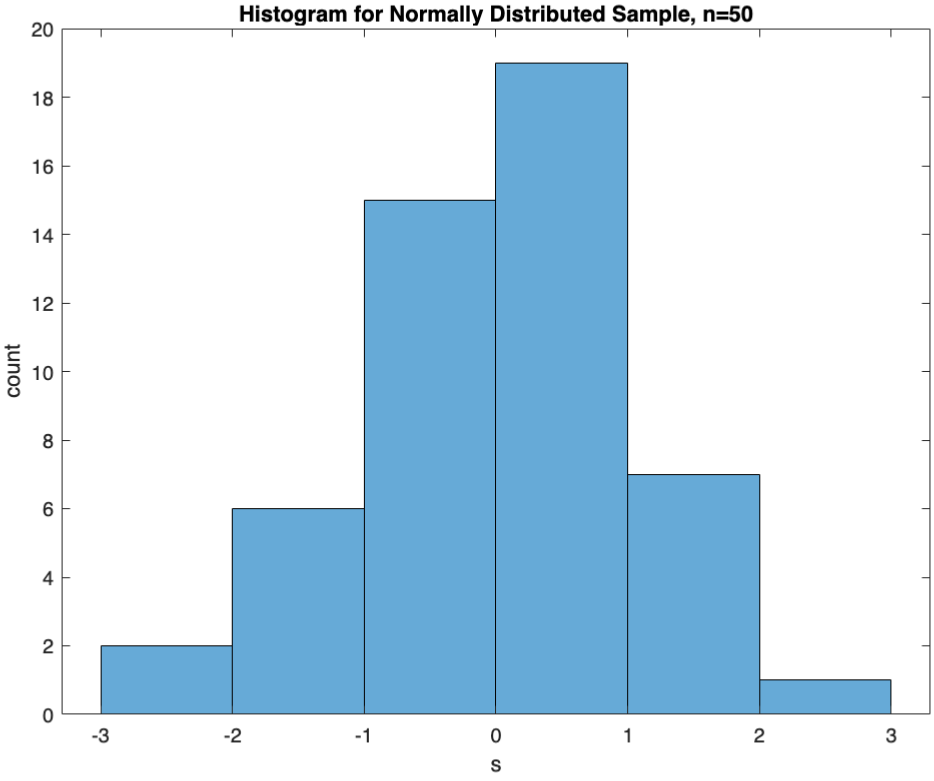
Repeat (a) for samples of sizes $n = 50$, $n = 100$, and $n = 1000$.

```
% n = 50
n = 50;
s50 = normrnd(0,1,[1,50]);

figure(3); clf
qqplot(s50); xlabel('z'); ylabel('s'); title('QQ Plot of Sample Data vs Standard Norma
```

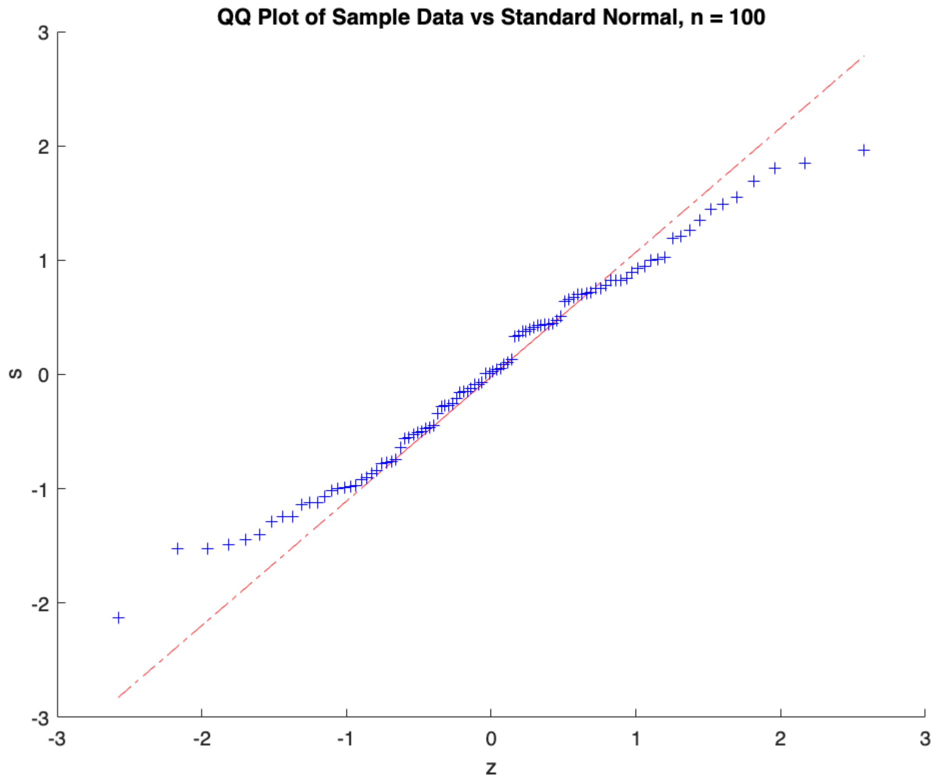


```
figure(4); clf
histogram(s50); xlabel('s'); ylabel('count')
title('Histogram for Normally Distributed Sample, n=50');
```

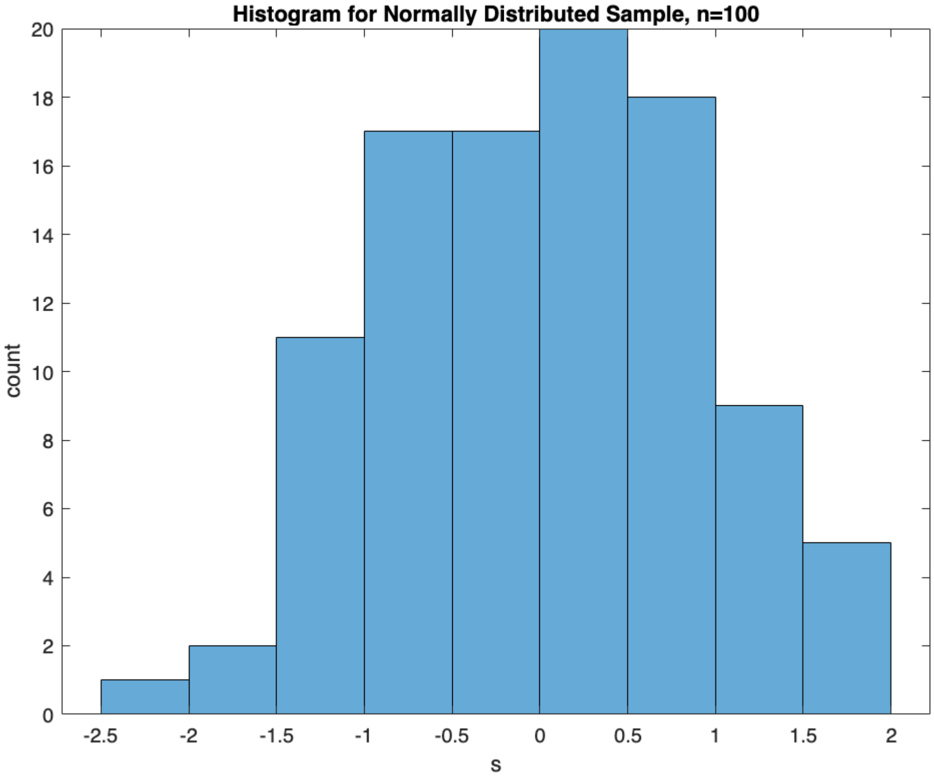


```
% n = 100
n = 100;
s100 = normrnd(0,1,[1,100]);

figure(5); clf
qqplot(s100); xlabel('z'); ylabel('s'); title('QQ Plot of Sample Data vs Standard Norma
```

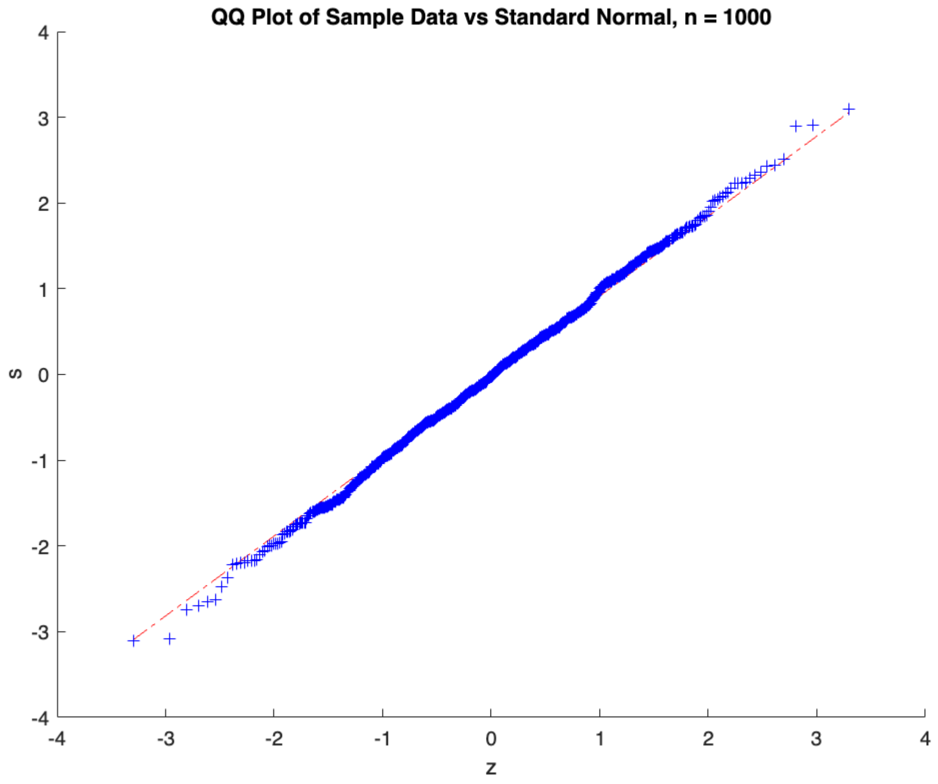


```
figure(6); clf
histogram(s100); xlabel('s'); ylabel('count')
title('Histogram for Normally Distributed Sample, n=100');
```

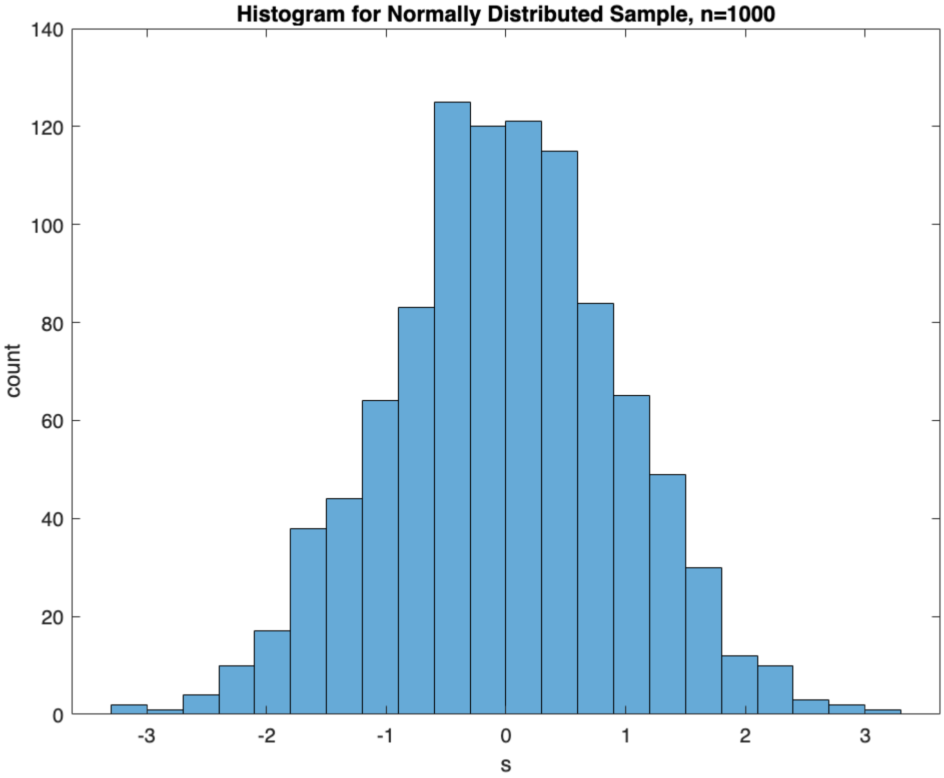


```
% n = 1000
n = 1000;
s1000 = normrnd(0,1,[1,1000]);

figure(7); clf
qqplot(s1000); xlabel('z'); ylabel('s'); title('QQ Plot of Sample Data vs Standard Nor
```



```
figure(8); clf
histogram(s1000); xlabel('s'); ylabel('count')
title('Histogram for Normally Distributed Sample, n=1000');
```



After several iterations of the above, I observed that as we increase n, we produce distributions that produce less variance from Standard normal, with more of the points in the QQ plot falling closer to $y = x$ and historgrams producing a more defined bell-shape, with reduced unimodality instances, and high symmetry. This held for multiple iterations of the code, with little variance from normal with the largest sample size $n = 1000$ and with the occurrence of unimodality, asymmetry, and skew still observed at $n = 50, 100$ (with the variance being worse for the former).

Part c

After experimenting with normal samples in (a) and (b), what would be your estimate for the "critical" sample size n^* , such that for samples of size larger than n^* , the normal-quantile plots are stable enough to be easily interpreted (i.e. do not deviate substantially from linearity)?

My estimate for the critical sample size n^* to be around $n^* = 500$. This is due to the fact that the behavior at $n = 1000$ is very desirable, and the behavior at $n = 100$ is very close to desirable.

IDS 157: Problem 5

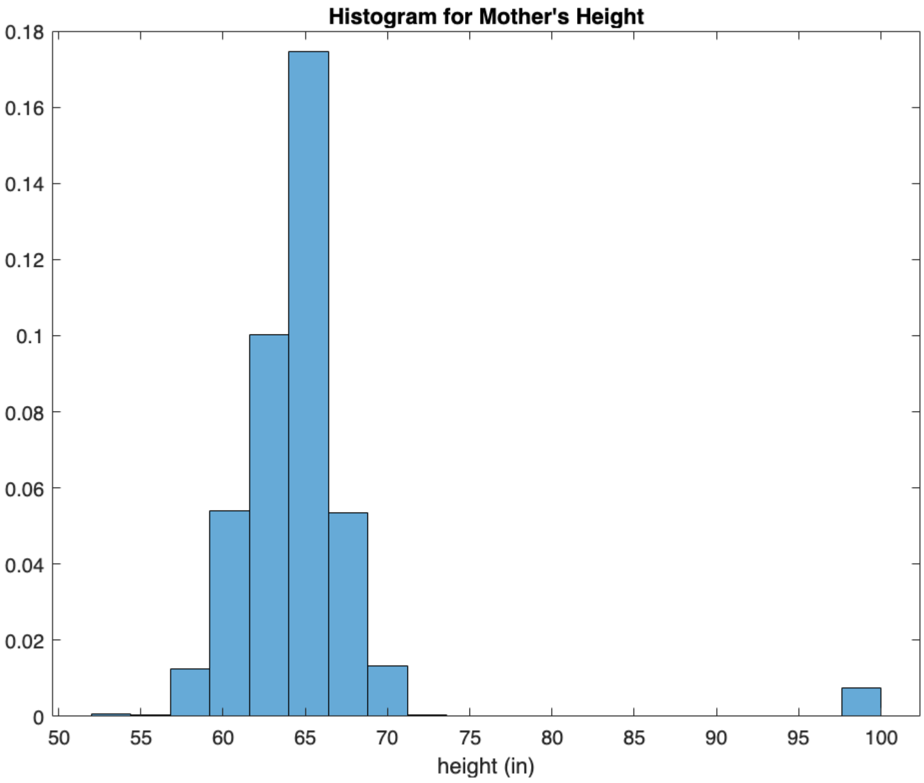
```
birth = readmatrix('./birth.txt');
```

Part a

Plot a normalized histogram for mothers' heights. What number of bins do you find optimal for representing the shape of the distribution?

```
motherHeight = birth(:,5);
nbins = 20;

figure(1); clf
histogram(motherHeight,nbins,'Normalization','pdf');
xlabel('height (in)');
title("Histogram for Mother's Height");
```



I find that the optimal number of bins for representing the distribution of the data is ~20 bins.

Part b

Compute the mean, median, standard deviation, and IQR of the heights. Is the center of the sample well-defined?

```
n = 1236; % n is even, n/2 is even

x_bar = 0; % calculating mean
```

```
for i = 1:n
    x_bar = x_bar + motherHeight(i);
end
x_bar = x_bar / n;
disp('Mean:'); disp(x_bar);
```

```
Mean:
   64.6699
```

```
% calculating median
x_tilde = 1/2 * (motherHeight(n/2) + motherHeight(n/2 + 1));
disp('Median:'); disp(x_tilde);
```

```
Median:
    62
```

```
% calculating standard deviation
s_x = 0;
for i = 1:n
    s_x = s_x + (motherHeight(i) - x_bar)^2;
end
s_x = sqrt(s_x / n);
disp('Standard Deviation:'); disp(s_x);
```

```
Standard Deviation:
    5.2589
```

```
% calculating IQR - n and n/2 are even
iq1 = n / 4;
iq3 = 3 * n / 4;
IQRx = 1/2 * ((motherHeight(iq3) + motherHeight(iq3+1)) - ...
    (motherHeight(iq1) - motherHeight(iq1+1)));
disp('IQR:');disp(IQRx);
```

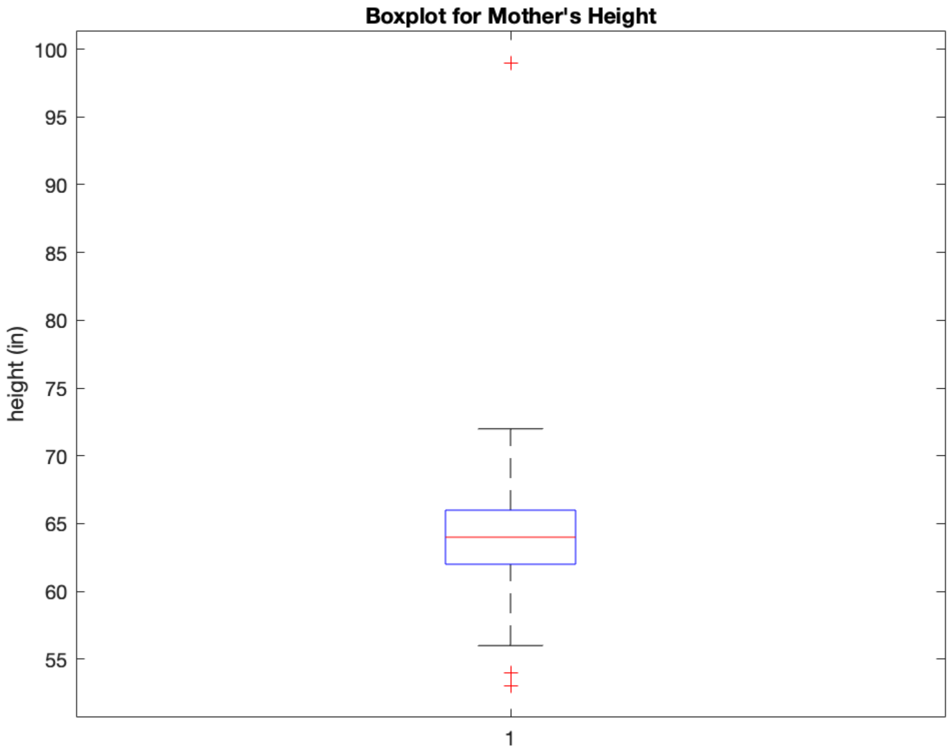
```
IQR:
   64.5000
```

We can see from the above that the center of the data is well defined, as it is clear that the trimmed mean is a slowly changing function of α.

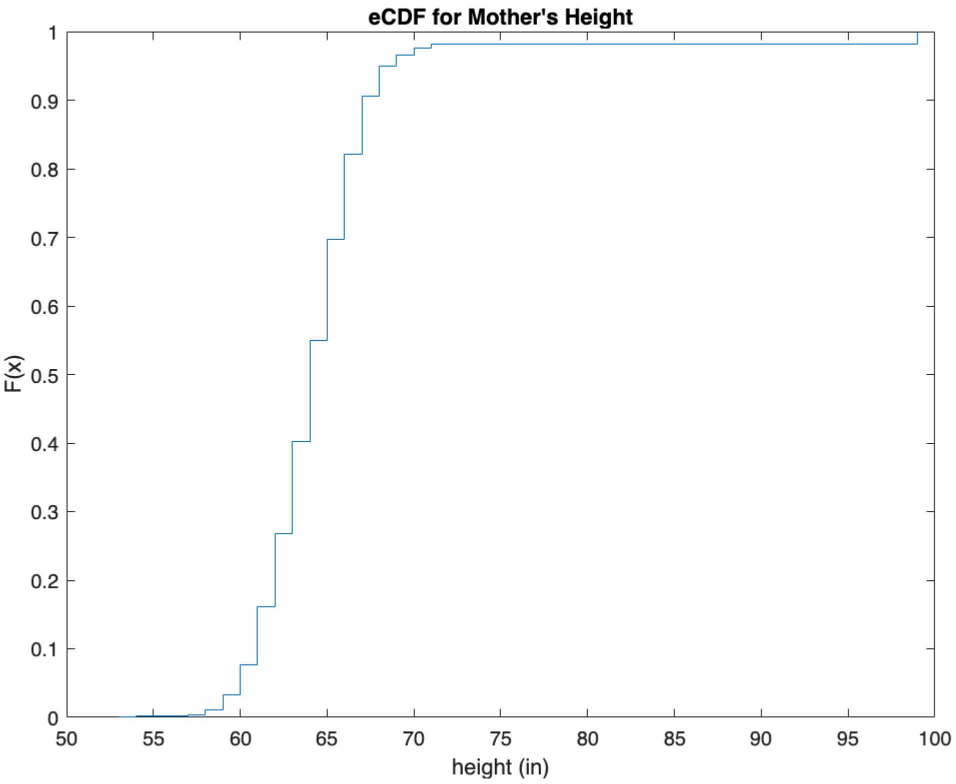
Part c

Plot the boxplot, eCDF, and QQ plot for the mothers’ heights. Based on these plots (and the histogram from (a)), would you consider the sample as being approximately normal? If yes, with what parameters (μ and σ^2)?

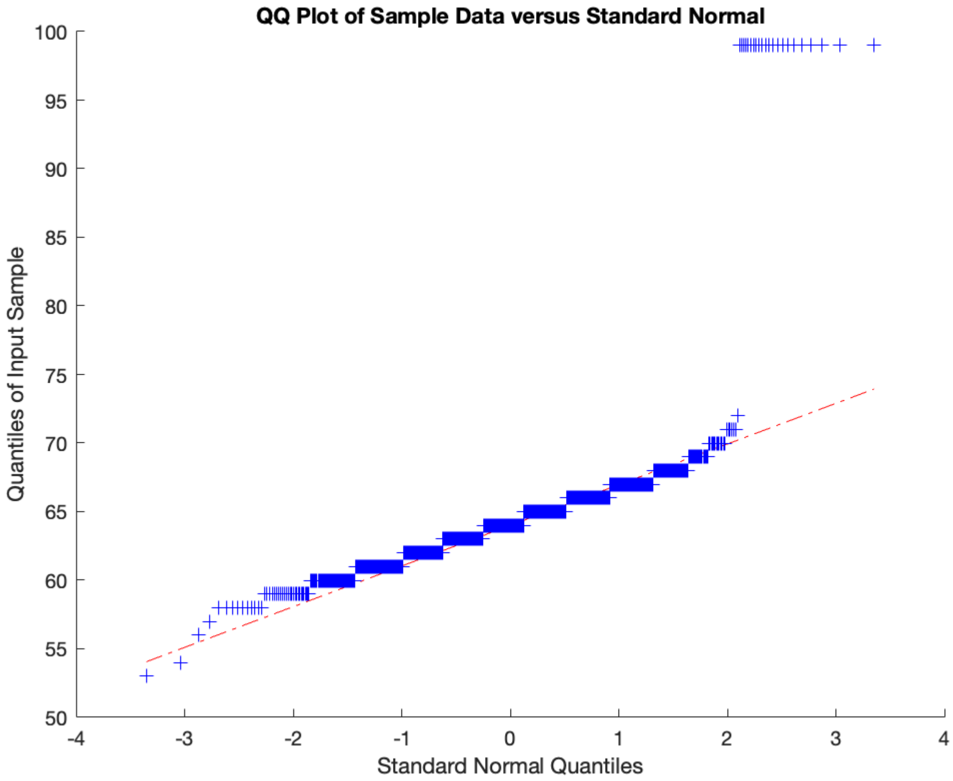
```
% plot boxplot
figure(2); clf
boxplot(motherHeight);
ylabel('height (in)');
title("Boxplot for Mother's Height");
```



```
% plot eCDF
figure(3); clf
ecdf(motherHeight);
xlabel('height (in)');
title("eCDF for Mother's Height");
```



```
% plot QQ
figure(4); clf
qqplot(motherHeight);
```



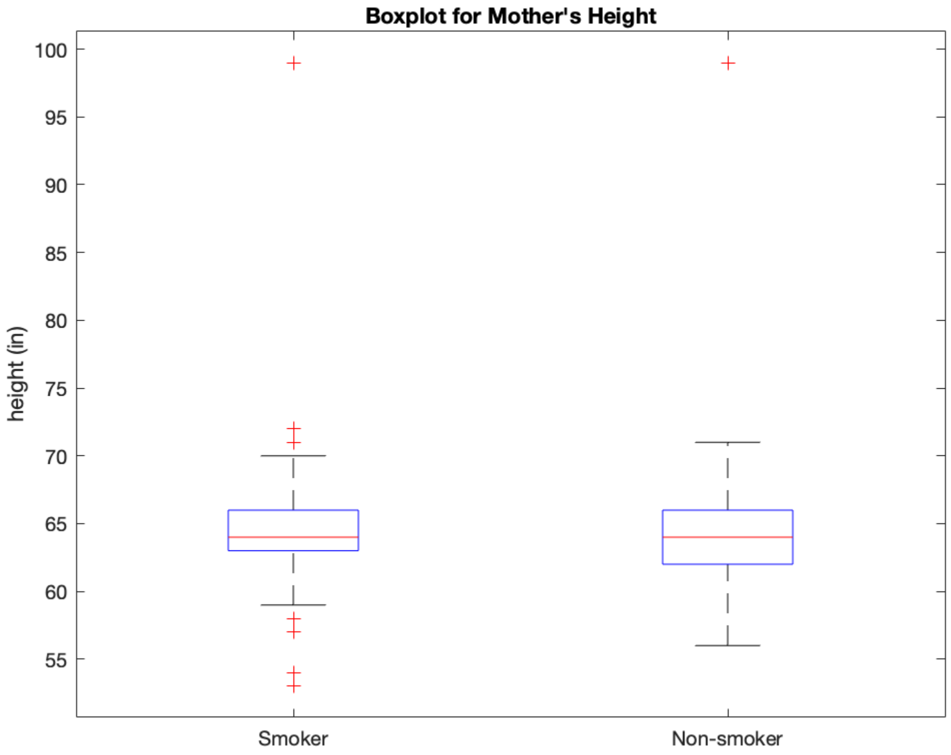
Based on the plots and histogram from (a), I would consider the sample as approximately normal. This is due to the fact that the only large deviations from the normal distribution are due to the large outliers, which do not have a large effect on the overall sample due to the large overall sample size. Thus, we have the parameters $\mu = 65, \sigma^2 = 5.3$

Part d

Compare the heights of mothers who smoked and who did not smoke using boxplots. Can you say that one of the two groups is higher (on average) than the other?

```
smokeHeight = birth(birth(:,7) == 1,5);
nosmokeHeight = birth(birth(:,7) == 0,5);

figure(5); clf
g = [ones(size(smokeHeight)); 2*ones(size(nosmokeHeight))];
boxplot([smokeHeight; nosmokeHeight],g,'Labels',{'Smoker', 'Non-smoker'});
ylabel('height (in)');
title("Boxplot for Mother's Height");
```



Based on the boxplots, I cannot determine whether mothers who smoke or mothers who do not smoke are taller. The distributions are extremely similar and are centered around the same area, with similar outliers.