Ben Juarez
UID: 2134561

Problem Set 1

1 Data Summaries and Linear Transformations

We want to express $\bar{y}, \tilde{y}, s_y, IQR_y$ in terms of $\bar{x}, \tilde{x}, s_x, IQR_x$ (given $y_i = \alpha + \beta x_i$): Let us first deal with $\bar{y}$.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \alpha + \beta x_i = \frac{1}{n}(n\alpha + \beta\sum_{i=1}^{n} x_i) = \frac{1}{n}(n\alpha + \beta n\bar{x}) = \alpha + \beta\bar{x}$$

For $\tilde{y}$ let us handle first handle when n is odd:

$$\tilde{x} = x_{\frac{n+1}{2}} \implies \tilde{y} = y_{\frac{n+1}{2}} = \alpha + \beta \cdot x_{\frac{n+1}{2}} = \alpha + \beta \cdot \tilde{y}$$

When n is even:

$$\tilde{x} = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \implies \tilde{y} = \frac{1}{2}(y_{\frac{n}{2}} + y_{\frac{n}{2}+1}) = \frac{1}{2}((\alpha + \beta \cdot x_{\frac{n}{2}}) + (\alpha + \beta \cdot x_{\frac{n}{2}+1}))$$

$$\implies \tilde{y} = \frac{1}{2}(2\alpha + \beta \cdot x_{\frac{n}{2}} + \beta \cdot x_{\frac{n}{2}+1}) = \alpha + \beta \cdot (\frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})) = \alpha + \beta\tilde{x}$$

Thus, we have shown

$$\implies \bar{y} = \alpha + \beta\bar{x} \quad \tilde{y} = \alpha + \beta\tilde{x}$$

Furthermore, we know $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$:

$$\implies s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \implies s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i - (\alpha + \beta\bar{x}))^2}$$

$$\implies s_y = \sqrt{\beta \cdot \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \implies s_y = \sqrt{\beta} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\implies s_y = \sqrt{\beta} \cdot s_x$$

Finally, we know $IQR_x = Q_{3x} - Q_{1x}$. We can see that if $\beta > 0$, then $Q_{1y} = \alpha + \beta Q_{1x}$ and $Q_{3y} = \alpha + \beta Q_{3x}$. If $\beta < 0$, then $Q_{1y} = \alpha + \beta Q_{3x}$ and $Q_{3y} = \alpha + \beta Q_{1x}$.

$$(\beta > 0) \implies IQR_y = Q_{3y} - Q_{1y} = \alpha + \beta Q_{3x} - (\alpha + \beta Q_{1x}) = \beta \cdot (Q_{3x} - Q_{1x}) = \beta \cdot IQR_x$$

$$(\beta < 0) \implies IQR_y = Q_{1y} - Q_{3y} = \alpha + \beta Q_{1x} - (\alpha + \beta Q_{3x}) = -\beta \cdot (Q_{3x} - Q_{1x}) = -\beta \cdot IQR_x$$

Therefore, we have $IQR_y = |\beta| \cdot IQR_x$.

2 Optimization Interpretation of $\bar{x}$ and $\tilde{x}$

Let us show that $\bar{x} = \arg\min_\alpha \sum_{i=1}^n (x_i - \alpha)^2$ and $\tilde{x} = \arg\min_\alpha \sum_{i=1}^n |x_i - \alpha|$. Let us first differentiate the first expression (that is being minimized) by x and set it to 0:

$$\implies 0 = \frac{d}{dx}(\sum_{i=1}^n (x_i - \alpha)^2) = \sum_{i=1}^n 2 \cdot (x - \alpha) = \sum_{i=1}^n 2x - \sum_{i=1}^n 2\alpha$$

$$\implies 2 \cdot \sum_{i=1}^n \alpha = 2 \cdot \sum_{i=1}^n x_i \implies n\alpha = \sum_{i=1}^n x_i$$

$$\implies \alpha = \frac{1}{n} \sum_{i=1}^n x_i$$

Clearly, this is the equation for the sample mean. Thus, as desired, we have shown that $\bar{x}$ is the value of α that minimizes the objective function $\sum_{i=1}^n (x_i - \alpha)^2$.

Let us perform this same process for the median (where $\frac{d}{dx}|x| = \frac{x}{|x|}$):

$$\implies 0 = \frac{d}{dx}(\sum_{i=1}^n |x_i - \alpha|) = \sum_{i=1}^n \frac{x_i - \alpha}{|x_i - \alpha|} = \sum_{i=1}^n \text{sign}(x_i - \alpha)$$

With this, we have two cases depending on when n is even or odd. If n is odd (and each x_i is distinct), then the sum expression is equal to 0 when $\frac{n}{2}$ of the x_i values are less than α and the other $\frac{n}{2}$ of the x_i values are greater than α. If each x_i is not distinct, then it is possible that the sum might not be equal to zero, but the function will still be minimized by taking the middle element as α (if we move α higher or lower, the sum $\sum_{i=1}^n |x_i - \alpha|$ can never be lower than it would be if $\alpha = x_{(\frac{n}{2}+1)}$, but it may be equivalent at the lowest). Clearly, this is the definition of the median $\tilde{x}$. Thus, when n is odd, we have shown that $\tilde{x}$ is the value of α that minimizes the objective function $\sum_{i=1}^n |x_i - \alpha|$, as desired.

Let us show that this still holds when n is even such that $\alpha \in [x_{(\frac{n}{2})-1}, x_{\frac{n}{2}}]$. We can imagine splitting the n values in half such that we can pick out the two "inside" values $x_{(\frac{n}{2})-1}, x_{\frac{n}{2}}$. Both of these values are valid medians $\tilde{x}$, and we can see that setting either one of these values as α minimizes the objective function $\sum_{i=1}^n |x_i - \alpha|$ equivalently. Clearly, any value for α that is not one of the middle values will have a greater (or equal) total sum from $\sum_{i=1}^n |x_i - \alpha|$. Thus, when n is even, the median $\tilde{x}$ is in the set of global minimizers for the objective function such that $\tilde{x} \in [x_{(\frac{n}{2})-1}, x_{\frac{n}{2}}]$. Let us note that this clearly also holds for $x_{(\frac{n}{2})-1} \leq \alpha \leq x_{\frac{n}{2}}$.

3 Interpretation of QQ Plots

Let us consider the distribution of the given sample if points $\{(z_{\frac{k}{n+1}}, x_{(k)})\}$ fall on the line $y = ax + b$ (instead of $y = x$).

As given, we know that if the points $\{(z_{\frac{k}{n+1}}, x_{(k)})\}$ fall on $y = x$, then the sample has approx. the standard normal distribution ($\mu = 0, \sigma = 1$). If these points now fall on $y = ax + b$, then we have $x_{(i)} \sim az_{\frac{i}{n+1}} + b$ instead of $x_{(i)} \sim z_{\frac{i}{n+1}}$. So, we are still dealing with a normal distribution, but the parameters of mean μ and standard deviation σ are altered such that μ is shifted by b and σ is scaled by a. Thus, we can say that if the points fall on $y = ax + b$, then the sample is distributed by

$$\mathcal{N}(b, a^2)$$

4 Readability of QQ Plots

See MATLAB section at end of file.

5 Playing with Data Summaries

See MATLAB section at end of file.

6 Simple Random Sampling

6.1 Part a

Let us compute $\mathbb{P}(s_1 = N), ..., \mathbb{P}(s_n = N)$. Each case has an equal probability such that with a target population of N, the probability that any one of the random samples is N (or any single population unit) is simply $\frac{1}{\text{population size}} = \frac{1}{N}$:

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

6.2 Part b

Let us compute $\mathbb{P}(\text{the } N\text{th population unit is in the sample})$. Following the logic of part a, for n random samples, the probability of a random sample being N is $\frac{1}{N}$. So, we must add up this probability for all of the n random samples:

$$\implies \mathbb{P}(\text{the } N\text{th population unit is in the sample}) = \mathbb{P}(s_1 = N) = \frac{1}{N} + ... + \mathbb{P}(s_n = N)$$

$$\implies \mathbb{P} = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N}$$

6.3 Part c

Let us compute $\mathbb{E}[s_1]$. Following previous logic and knowing the formula for expected value as $\mathbb{E}[X] = \sum x_i \cdot \mathbb{P}(x_i)$:

$$\implies \mathbb{E}[s_1] = \sum_{i=1}^{N} i \cdot \mathbb{P}(s_1 = i) = \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{1}{N} \sum_{i=1}^{N} i = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}$$

6.4 Part d

Let us compute $\mathbb{P}(s_1 = N, s_2 = 1)$. We can see that these two events are not independent (occurs without replacement), so we compute the probability as follows such that the probability of s_2 being any population unit (that is not equal to s_1) is simply $\frac{1}{N-1}$:

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

6.5 Part e

Let us compute $\mathbb{P}(s_i = i, \text{for all } i = 1, ..., n)$. Following the logic of the previous problem, we repeat the process of multiplying the probabilities for each sample (n times):

$$\implies \mathbb{P}(s_i = i, \text{for all } i = 1, ..., n) = \frac{1}{N} \cdot \frac{1}{N-1} \cdot \frac{1}{N-2} \cdot ... \cdot \frac{1}{N-(n-1)} = \prod_{i=1}^{n} \frac{1}{N-(i-1)}$$

7 Weighted Sample Mean

Let us consider a more general class of estimators:

$$\bar{X}_n^w = \sum_{i=1}^{n} w_i X_i \quad (4)$$

where w_i are some weights and $\bar{X}_n$ is a special case with $w_i = \frac{1}{n}$

7.1 Part a

We want to find what condition on the weights makes (4) an unbiased estimate of μ. We know that (4) is an unbiased estimate of μ if $\mathbb{E}[\bar{X}_n^w] = \mu$:

$$\implies \mathbb{E}[\bar{X}_n^w] = \mathbb{E}[\sum_{i=1}^{n} w_i X_i] = \sum_{i=1}^{n} w_i \mathbb{E}[X_i]$$

Clearly, the condition such that $\sum_{i=1}^{n} w_i = 1$ must hold in order for (4) to be an unbiased estimate of μ.

7.2 Part b

Let us find the estimate with the smallest standard error. So, we will need to minimize

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

We can express $\mathbb{V}[\bar{X}_n^w]$ as

$$\mathbb{V}[\bar{X}_n^w] = \mathbb{V}[\sum_{i=1}^{n} w_i X_i] = \sum_{i=1}^{n} \mathbb{V}[w_i X_i] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov(w_i X_i, w_j X_j)$$

$$\implies \mathbb{V}[\bar{X}_n^w] = \sum_{i=1}^{n} w_i^2 \cdot \mathbb{V}[X_i] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j \cdot Cov(X_i, X_j)$$

Since we can assume that we have SRS (confirmed from Piazza question 18), we can use the following following formula for covariance (Lemma 2) since $i \neq j$ (without replacement):

$$Cov(X_i, X_j) = -\frac{\sigma^2}{N-1}$$

This give us the following (also using $\mathbb{V}[X_i] = \sigma^2$):

$$\implies \mathbb{V}[\bar{X}_n^w] = \sum_{i=1}^{n} w_i^2 \cdot \sigma^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j \cdot (-\frac{\sigma^2}{N-1}) \quad (*)$$

Let us find a way to isolate the term w_i in order to proceed further. Since $(\sum_{i=1}^{n} w_i)^2 = \sum_{i=1}^{n} w_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j$ and $\sum_{i=1}^{n} w_i = 1$ (from condition in part a) we can express $2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j$ as

$$2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} w_i w_j = (\sum_{i=1}^{n} w_i)^2 - \sum_{i=1}^{n} w_i^2 = 1 - \sum_{i=1}^{n} w_i^2$$

Let us now plug back into $(*)$

$$\implies \mathbb{V}[\bar{X}_n^w] = \sum_{i=1}^{n} w_i^2 \cdot \sigma^2 + (1 - \sum_{i=1}^{n} w_i^2) \cdot (-\frac{\sigma^2}{N-1})$$

$$\implies \mathbb{V}[\bar{X}_n^w] = \sum_{i=1}^{n} w_i^2 \cdot \sigma^2 - \frac{\sigma^2}{N-1} + (\sum_{i=1}^{n} w_i^2 \cdot \frac{\sigma^2}{N-1})$$

$$\implies \mathbb{V}[\bar{X}_n^w] = \sum_{i=1}^{n} w_i^2 \cdot \sigma^2 + (\sum_{i=1}^{n} w_i^2 \cdot \frac{\sigma^2}{N-1}) - \frac{\sigma^2}{N-1}$$

$$\implies \mathbb{V}[\bar{X}_n^w] = (\sigma^2 + \frac{\sigma^2}{N-1})\sum_{i=1}^{n} w_i^2 - \frac{\sigma^2}{N-1}$$

At this point, we can focus on minimizing $\sum_{i=1}^{n} w_i^2$ on the condition $\sum_{i=1}^{n} w_i = 1 \implies \sum_{i=1}^{n} w_i - 1 = 0$. We can use Lagrange multipliers to find a critical value for w and then we can use a bordered Hessian matrix in order to prove that it minimizes $\sum_{i=1}^{n} w_i^2$:

$$L(w_1, w_2, ..., w_n, \lambda) = \sum_{i=1}^{n} w_i^2 - \lambda(\sum_{i=1}^{n} w_i - 1)$$

$$\implies \frac{\partial}{\partial w_k}(L) = 0$$

$$\implies \frac{\partial}{\partial w_k}(\sum_{i=1}^{n} w_i^2 - \lambda(\sum_{i=1}^{n} w_i - 1)) = 0$$

$$\implies 2 \cdot w_k - \lambda = 0 \implies w_k = \frac{\lambda}{2}$$

Therefore, we know that all w_k (for $k \in \{1, ..., n\}$) are equal to each other since we arrive at a constant. With this, plus the condition $\sum_{i=1}^{n} w_i = 1$, we can confirm that a critical point is when $w = \frac{1}{n}$. By utilizing bordered Hessian matrices, we can show that this value of w minimizes the variance, thus minimizing the standard error (this process could also be shown using the Cauchy-Schwarz inequality). With $g(w_1, ..., w_n) = \sum_{i=1}^{n} w_i$ and $f(w_1, ..., w_n) = \sum_{i=1}^{n} w_i^2$, let us show that $\frac{1}{n}$ minimizes $\sum_{i=1}^{n} w_i^2$.

For $n = 1$:

$$\implies \begin{vmatrix} 0 & g_{w_1} \\ g_{w_1} & L_{w_1w_1} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & 2 \end{vmatrix} = 0 - 1 = -1 < 0 \implies \text{min}$$

For $n = 2$:

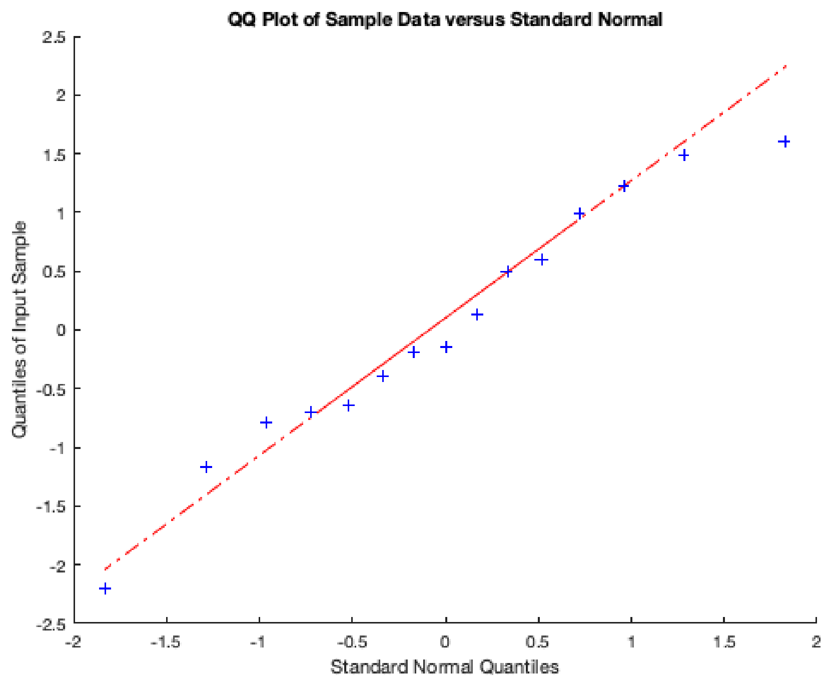
$$\implies \begin{vmatrix} 0 & g_{w_1} & g_{w_2} \\ g_{w_1} & L_{w_1w_1} & L_{w_1w_2} \\ g_{w_2} & L_{w_2w_1} & L_{w_2w_2} \end{vmatrix} = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix}$$

$$\implies \begin{vmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix} = 0 \begin{vmatrix} 2 & 0 \\ 0 & 2 \end{vmatrix} - 1 \begin{vmatrix} 1 & 0 \\ 1 & 2 \end{vmatrix} + 1 \begin{vmatrix} 1 & 2 \\ 1 & 0 \end{vmatrix} = 0 - 2 - 2 = -4 < 0 \implies \text{min}$$

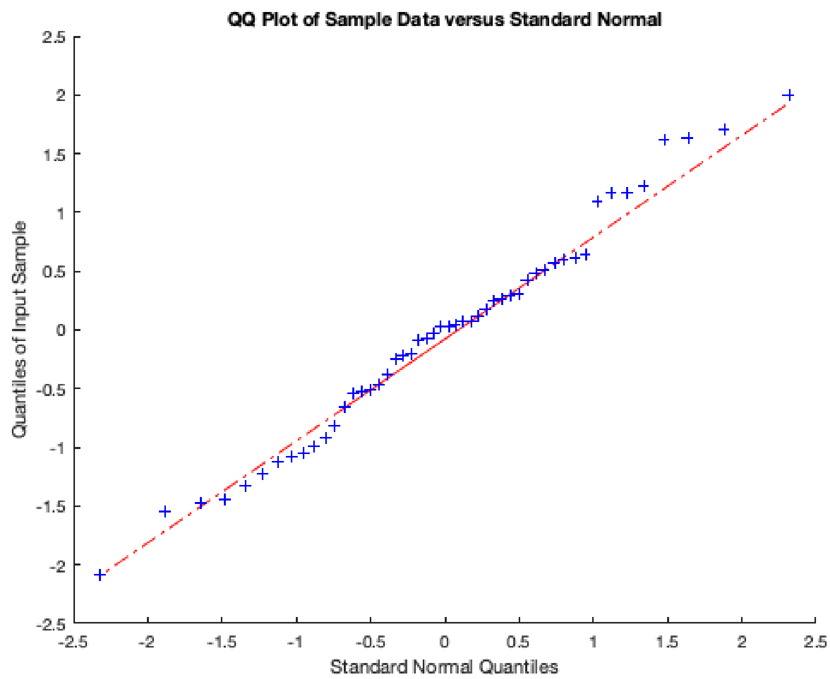
Extending this argument for all possible values of n (since the same structure is maintained), with the determinant of the bordered Hessian matrix being less than 0, we know that $\frac{1}{n}$ minimizes $\sum_{i=1}^{n} w_i^2$, thus minimizing the standard error overall. Let us also note that the weight condition $\sum_{i=1}^{n} w_i = 1$ clearly still holds. Therefore, we can say that the most efficient estimate among all unbiased estimates of $\bar{X}_n^w$ is when each $w_i = \frac{1}{n}$. In other words, the most efficient estimate is the sample mean $\bar{X}_n$.

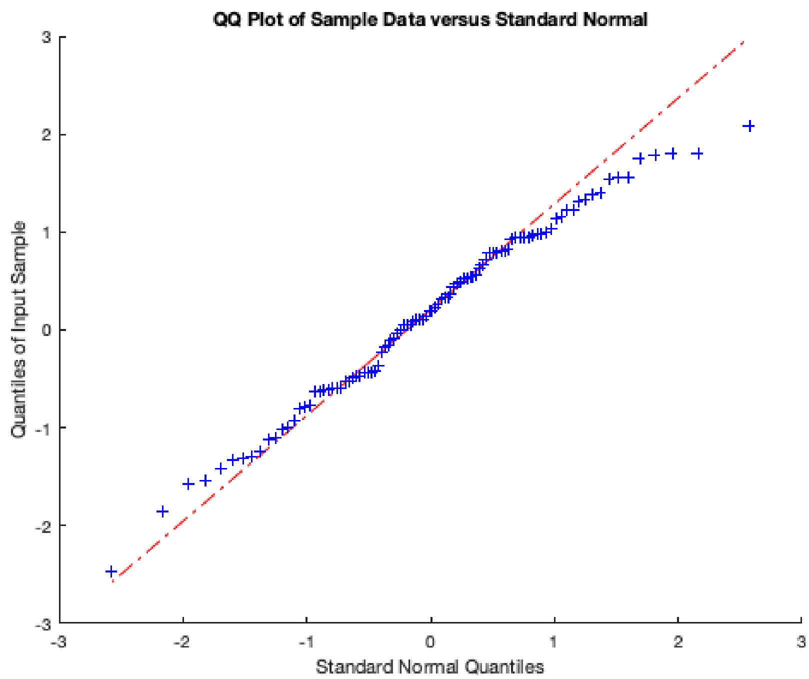
8 MATLAB

```
% Ben Juarez
% PS1Q4 - Readability of QQ Plots

% Part a
n = 15;
sample_a = normrnd(0,1,[1,n]);
histogram(sample_a);
qqplot(sample_a);
snapnow
% The points on the QQ plot do not appear to fall on a straight line.
% The histogram is neither symmetric, unimnodal, or bell-shaped (perhaps it
% could be interpretted as unimodal in some cases, but there does not seem
% to be enough data points).

% Part b
n = 50;
sample_bi = normrnd(0,1,[1,n]);
histogram(sample_bi);
qqplot(sample_bi);
snapnow
% The points do not appear to fall on a straight line (although, closer
% than part a).
% The histogram is unimodal, but not quite symmetric or bell-shaped.

n = 100;
sample_bii = normrnd(0,1,[1,n]);
histogram(sample_bii);
qqplot(sample_bii);
snapnow
% The points are closer to falling on a straight line, but this condition
% is still not quite met.
% The histogram is unimodal and essentially symmetric as well as generally
% bell-shaped.

n = 1000;
sample_biii = normrnd(0,1,[1,n]);
histogram(sample_biii);
qqplot(sample_biii);
snapnow
% The points appear to fall on a straight line for the most part.
% The histogram is unimodal, symmetric, and bell-shaped.

% Part c
n = 500;
sample_c = normrnd(0,1,[1,n]);
histogram(sample_c);
qqplot(sample_c);
snapnow
% 500 seems to be an appropriate estimate for n* since for n > n* = 500,
% the normal-quantile plots do not deviate substantially from linearity.
```

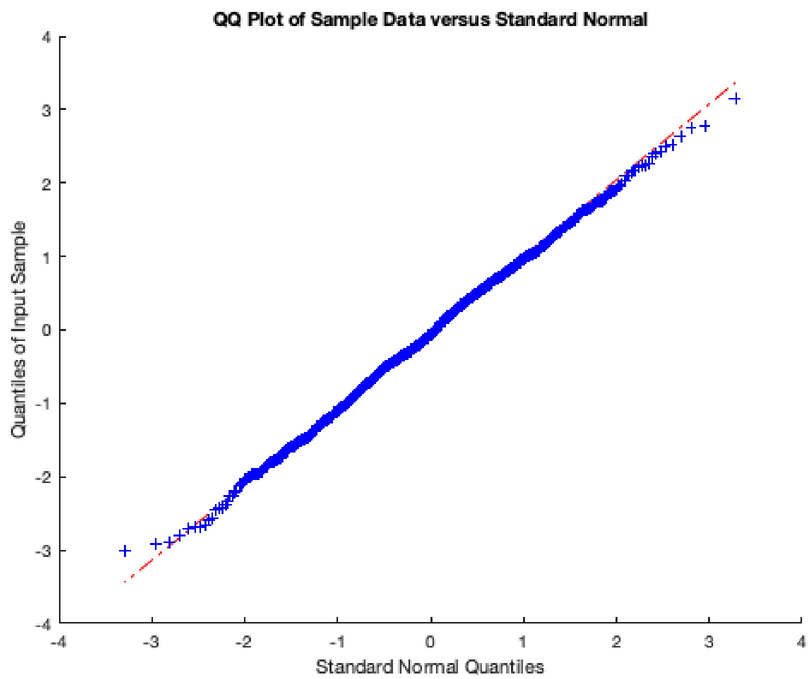

QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

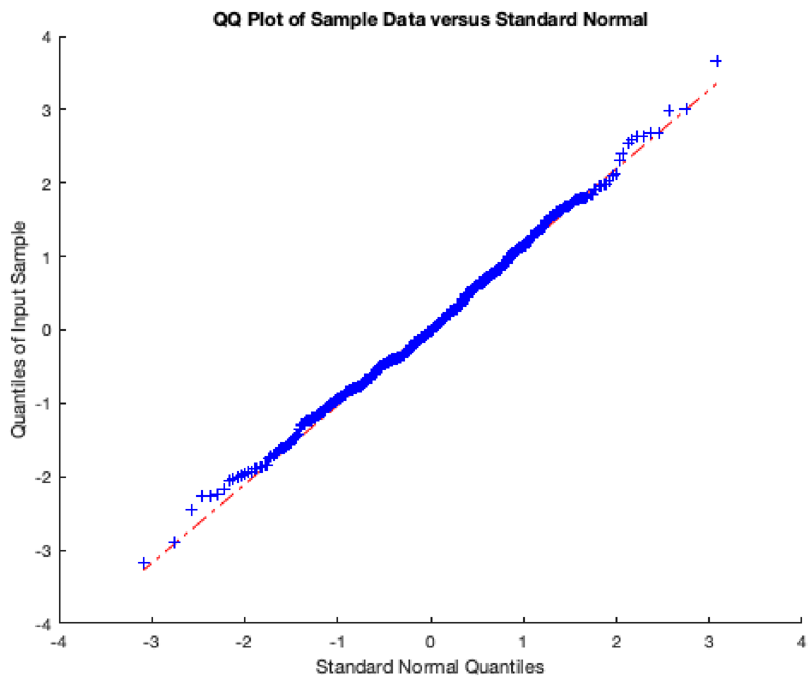
Problem Set 1




QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

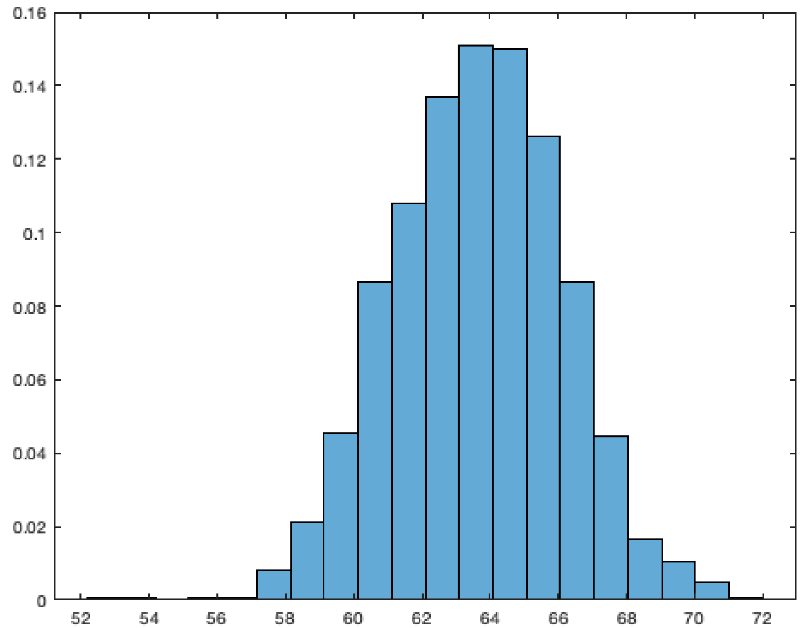
4


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

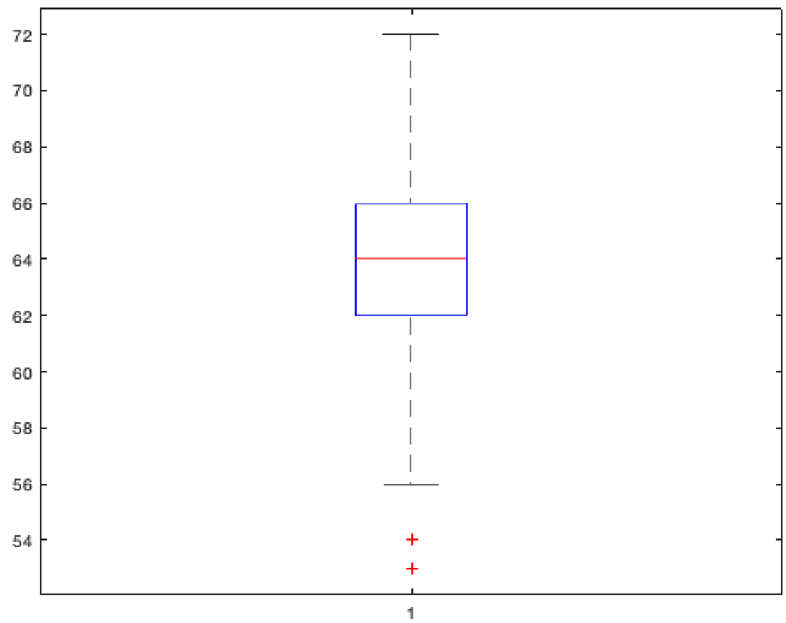


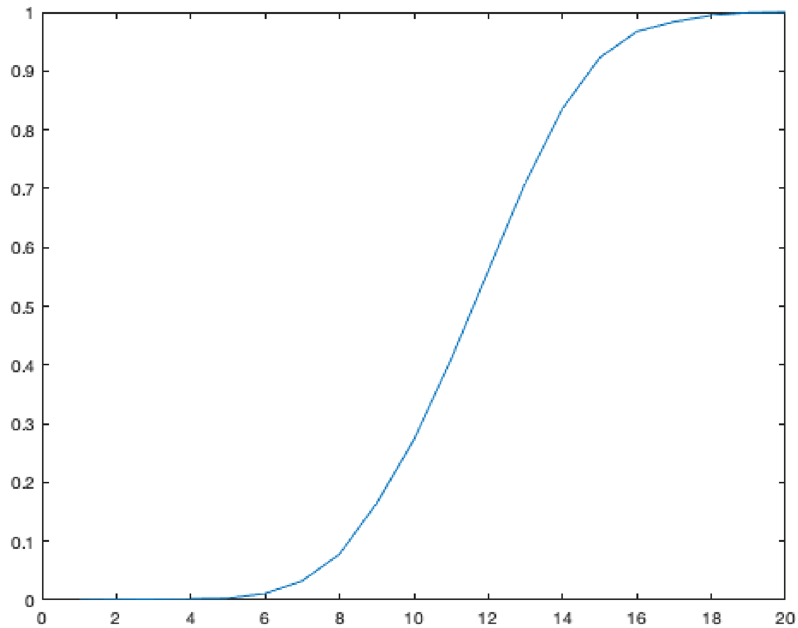
Published with MATLAB® R2021b

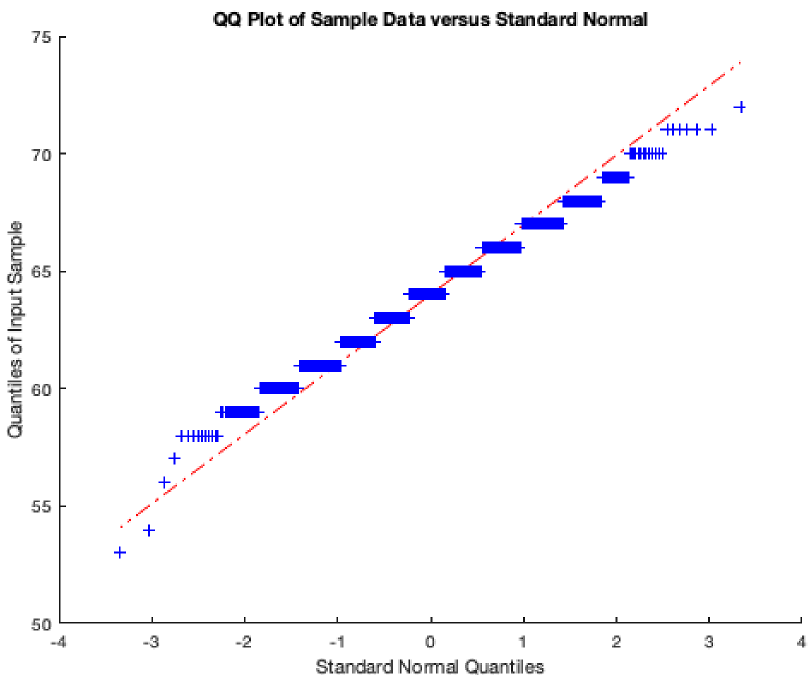
```
% Ben Juarez
% PS1Q5 - Playing With Data Summaries

% Part a
birth = importdata("birth.txt");
mother_heights = birth(:,5);
mother_heights = mother_heights(mother_heights<99); % removing 99 unknown
histogram(mother_heights, 20, "Normalization", "probability")
snapnow
% Around 20 bins appears to be optimal for representing the shape of the
% distribution.  If the number of bins was greater than 20, the bars do not
% always touch.

% Part b
mean(mother_heights)
median(mother_heights)
std(mother_heights)
iqr(mother_heights)
% Yes, the center of the sample seems well-defined because the mean and
% median are both approximately 64.

% Part c
boxplot(mother_heights)
snapnow
plot(ecdf(mother_heights))
snapnow
qqplot(mother_heights)
snapnow
% Yes, it appears that this sample is approximately normal considering the
% boxplot (placement of mean relative to Q1/Q3, proper spacing), eCDF plot
% (matches well with normal eCDF), QQ plot (linearity mostly preserved),
% and histogram (decently strong bell-shape).  The parameter for mean would
% simply be 64.04 and the parameter for st. dev. would be 2.53.

% Part d
smokers = birth(birth(:,7)==1,:);
smokers_heights = smokers(:,5);
smokers_heights = smokers_heights(smokers_heights<99);

nonsmokers = birth(birth(:,7)==0,:);
nonsmokers_heights = nonsmokers(:,5);
nonsmokers_heights = nonsmokers_heights(nonsmokers_heights<99);

boxplot(smokers_heights)
snapnow
boxplot(nonsmokers_heights)
snapnow
% Comparing the boxplots, the heights of smokers vs. nonsmokers in this
% case are very similar.  The means and medians appear to be the same.
% Thus, we cannot say that one of the two groups has definitively higher
% average heights.
```

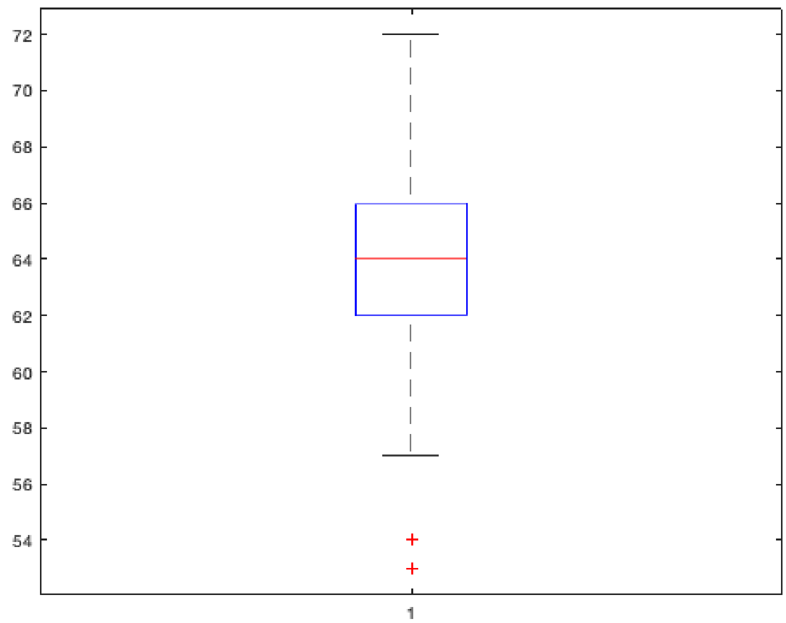


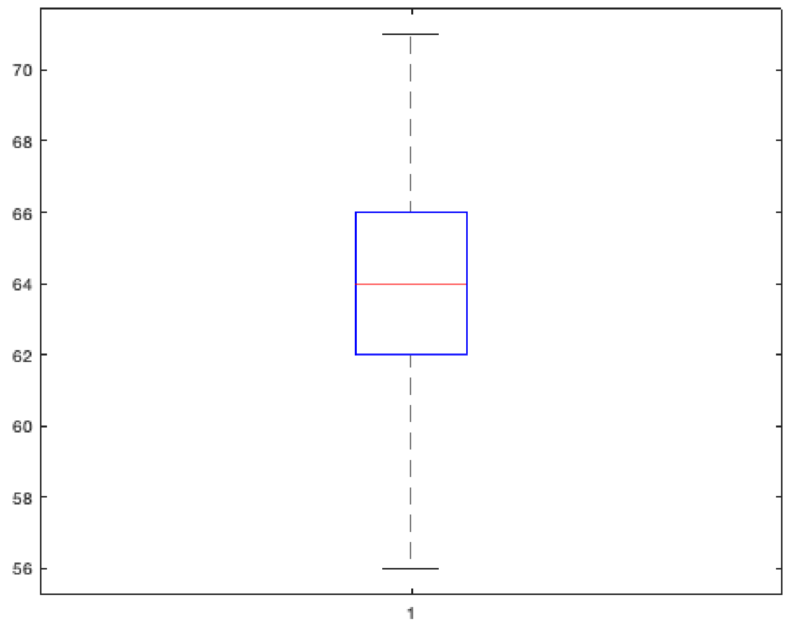

```
ans =

   64.0478

ans =

    64

ans =

    2.5334

ans =

     4
```


72
70
68
66
64
62
60
58
56
54
1


0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14
16
18
20


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
75
70
65
60
55
50
-4
-3
-2
-1
0
1
2
3
4


72
70
68
66
64
62
60
58
56
54
1



Published with MATLAB® R2021b