**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 3 - Handout 3 R and Handout 3 SAS**

**Part 3 – R Handout**

Exercise 1

1. Make a histogram of men’s heights and a histogram of women’s heights

> fhgtmean <- mean(fdims$hgt)

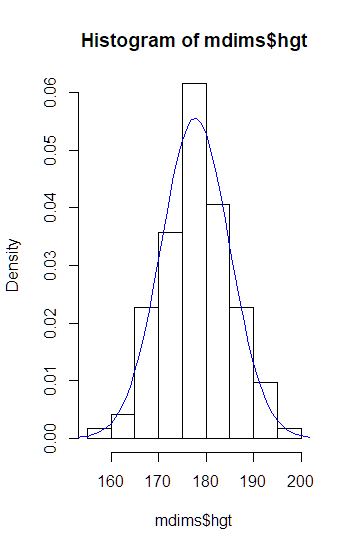
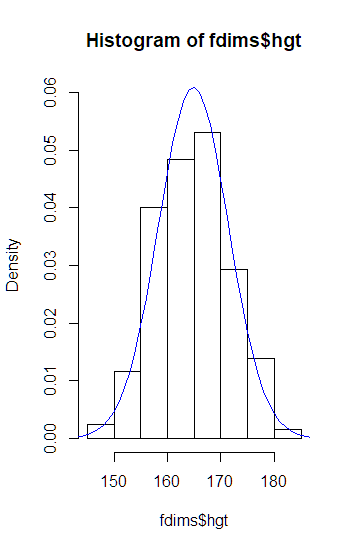
> fhgtsd <- sd(fdims$hgt)

> x <- 140:190

> y <- dnorm(x=x, mean=fhgtmean, sd=fhgtsd)

> hist(fdims$hgt, probability = TRUE, ylim = c(0,0.06))

> lines(x=x, y=y, col= "blue")



> mhgtmean <-mean(mdims$hgt)

> mhgtsd <- sd(mdims$hgt)

> x <- 150:210

> y <- dnorm(x=x, mean=mhgtmean, sd=mhgtsd)

> hist(mdims$hgt, probability = TRUE, ylim = c(0,0.06))

> lines(x=x, y=y, col= "blue")

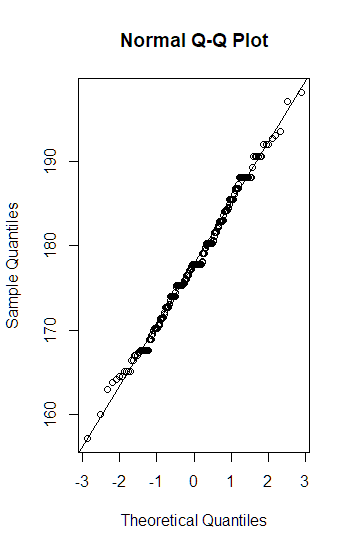
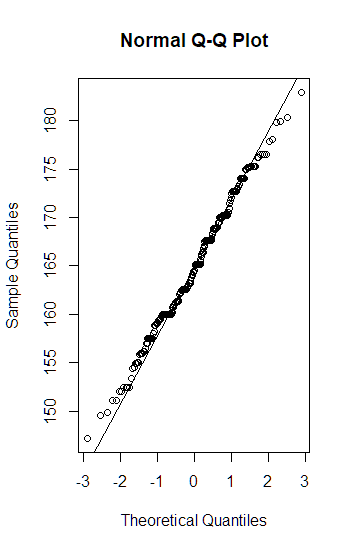
1. How would you compare the various aspects of the two distribution.
   1. I would review each of the histograms and view the curve and compare the shape and distribution to compare them or I could run qqnorm / qqline functions to generate plots and line and verify the plots are linear and follow the line, which would show the normality.

Exercise 2

1. Based on this plot, does it appear that the data follows a nearly normal distribution.
   1. I would say both follow the linear line and are normalized.

> qqnorm(fdims$hgt)

> qqline(fdims$hgt)



> qqnorm((mdims$hgt))

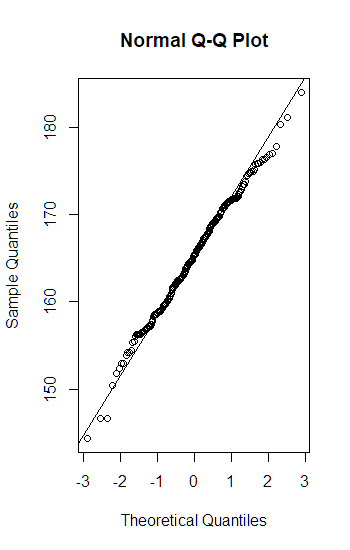
> qqline(mdims$hgt)

Exercise 3

1. Make a normal probability plot of sim\_norm.

> qqnorm(sim\_norm)

> qqline(sim\_norm)

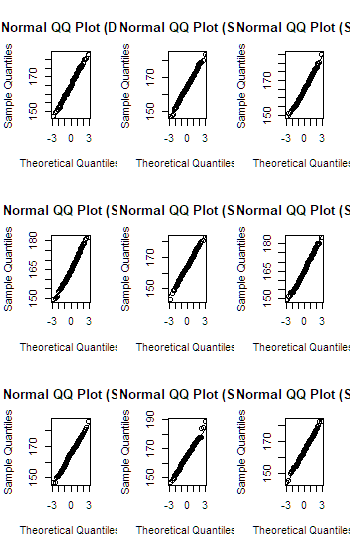


1. Do all of the points fall on the line.
   1. no
2. How does this plot compare to the probability plot for the real data.
   1. Similar, with majority of the plots on or around the linear line.

Exercise 4

1. Does the normal probability plot for fdims$hgt look similar to the plots created for the simulated data.

qqnormsim(fdims$hgt)



1. That is do the plots provide evidence that the female heights re nearly normal.
   1. Yes, each plot has slightly different results but all of them look to follow the linear line.

Exercise 5

1. Using the same technique determine whether or not female weights appear to come from a normal distribution

> fwgtmean <-mean(fdims$wgt)

> fwgtmean <-mean(fdims$wgt)

> x <- 30:150

> y <- dnorm(x=x, mean=fwgtmean, sd=fwgtsd)

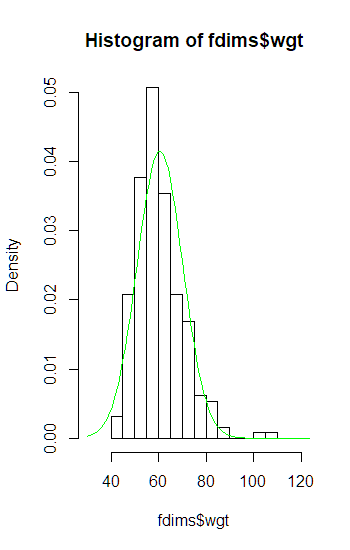
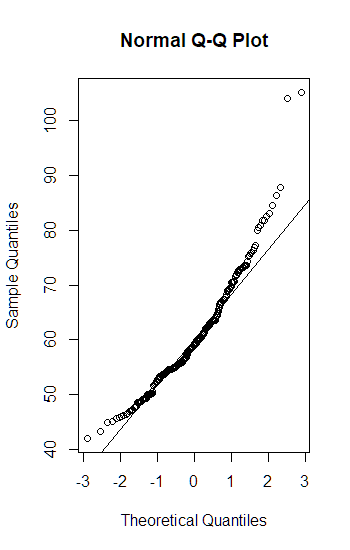
> hist(fdims$wgt, probability = TRUE, xlim = c(30,120), ylim = c(0,0.05))

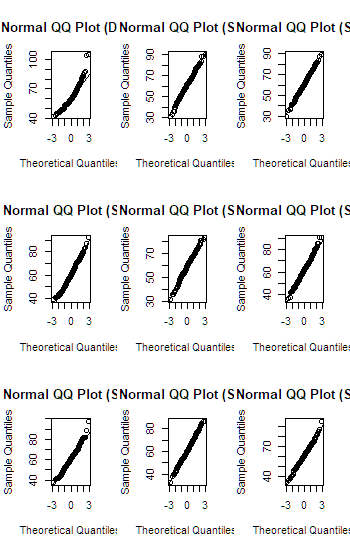
> lines(x=x, y=y, col="green")

> qqnorm((fdims$wgt))

> qqline(fdims$wgt)

> qqnormsim(fdims$wgt)



Yes, female’s weights also appear to come form a normalized distribution

Exercise 6

1. Write out two probability questions that you would like to answer; one regarding female heights and one regarding female weights.
   1. What is prob a random female is shorter than 5.2 ft (62.4 in or 158.496 cm)
   2. What is the prob a random female weight’s more than 126 lbs.(57.153)
2. Calculate the probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all).
   1. What is prob a random female is shorter than 5.2 ft (62.4 in or 158.496 cm)

> pnorm(q=158.496, mean=fhgtmean, sd=fhgtsd)

[1] 0.1649575

> sum(fdims$hgt > 158.496)/length(fdims$hgt)

[1] 0.8576923

* 1. What is the prob a random female weight’s more than 126 lbs.(57.153)

> 1-pnorm(q=57.153, mean=fwgtmean, sd=fwgtsd)

[1] 0.6400215

> sum(fdims$wgt > 57.153)/length(fdims$wgt)

[1] 0.5846154

1. Which variable, height or weight, had a closer agreement between the two methods.
   1. Females weight was < 126 lbs.

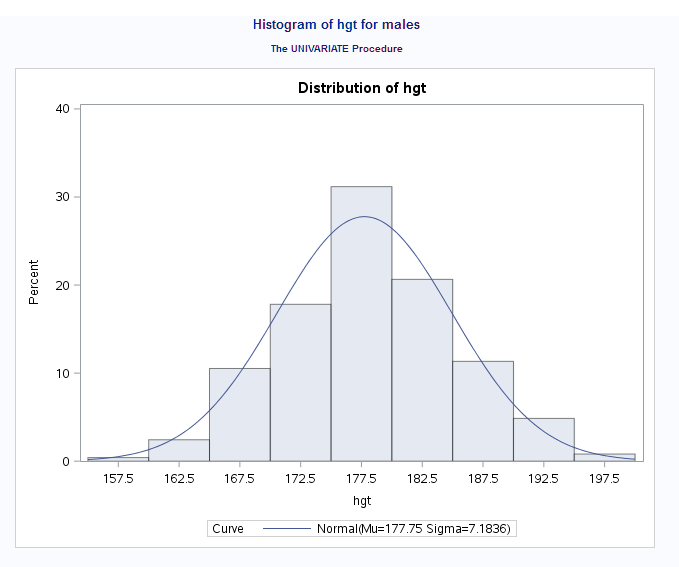
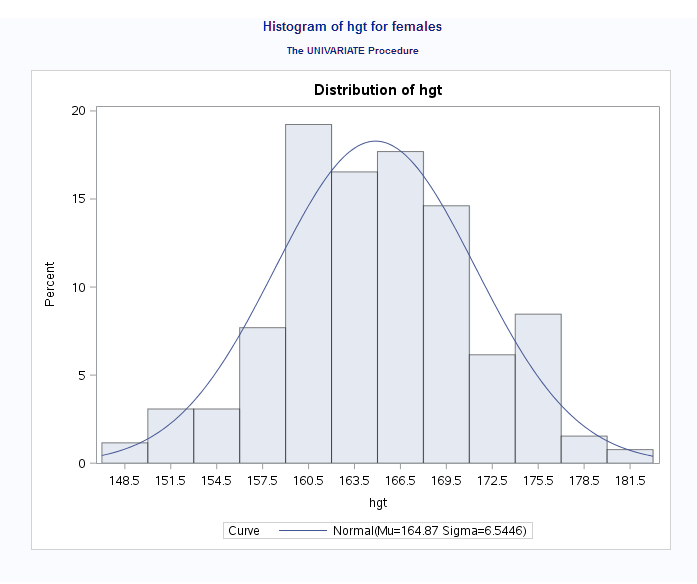
**Part 3 - SAS Handout**

Exercise 1

1. Make a histogram of men’s heights and a histogram of women’s heights

title 'Histogram of hgt for males';  
ods select histogram;  
proc univariate data=mdims;  
 var hgt;  
 histogram/ normal;  
 output out=estimates n=n mean=mean std=std;  
run;

title 'Histogram of hgt for females';  
ods select histogram;  
proc univariate data=fdims;  
 var hgt;  
 histogram/ normal;  
 output out=estimates n=n mean=mean std=std;  
run;

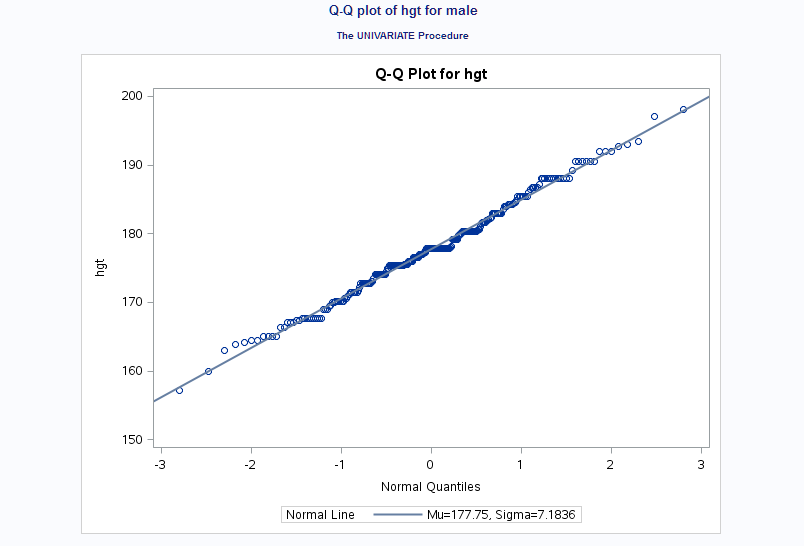
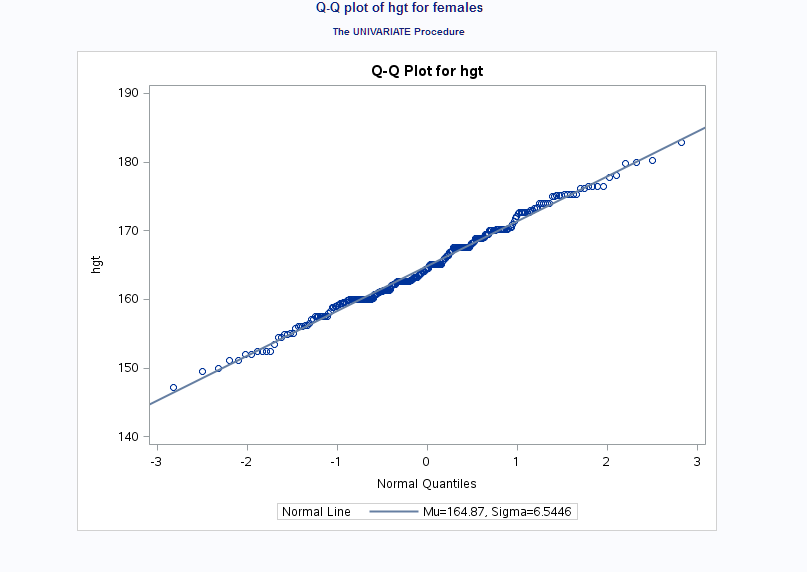
 

1. How would you compare the various aspects of the two distribution.
   1. I would perform a visual comparison first, then do linear model making sure plots follow the linear line and compare spread, shape and separation to confirm if they model follow a normal distribution.

Exercise 2

1. Based on this plot, does it appear that the data follows a nearly normal distribution.

Title 'Q-Q plot of hgt for females';  
ODS select qqplot;  
proc univariate data=fdims;  
 var hgt;  
 qqplot / normal (mu=est sigma=est);  
run;  
  
Title 'Q-Q plot of hgt for male';  
ODS select qqplot;  
proc univariate data=mdims;  
 var hgt;  
 qqplot / normal (mu=est sigma=est);  
run;

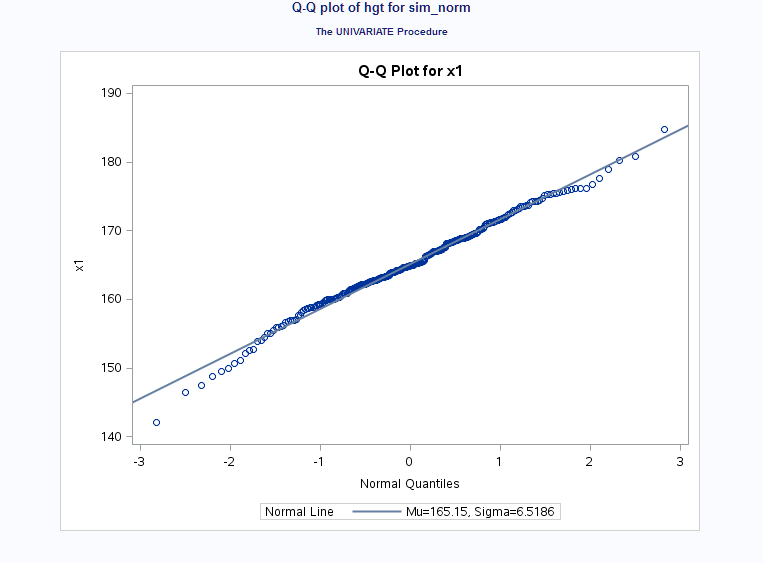


* 1. Yes, based on the Q-Q plot both male and female data have an approximate normal distibrution.

Exercise 3

1. Make a normal probability plot for variable **x1** in the data set **sim\_norm**.

data sim\_norm;  
 do i=1 to &n;  
 x1= rand('NORMAL', &mean, &std);  
 output;   
 end;  
run;  
  
Title 'Q-Q plot of hgt for sim\_norm';  
ODS select qqplot;  
proc univariate data=sim\_norm;  
 var x1;  
 qqplot x1/ normal (mu=est sigma=est);  
run;



1. Do all of the points fall on the line?
   1. Most of the points fall approximately on the line
2. How does this plot compare to the probability plot for the real data?
   1. Similar approximation when it comes to following the linear line.

Exercise 4

1. Does the normal probability plot for hgt for females look similar to the plots created for the simulated data?

data simulated;  
 array x {9} x1-x9;  
 do i=1 to &n;  
 do j=1 to 9;  
 x[j]=rand('NORMAL', &mean, &std);  
 end;  
 output;  
 end;  
run;  
  
Title 'Q-Q plot of hgt for simulated x1 - x9';  
ODS select qqplot;  
proc univariate data=simulated;  
 var x1-x9;  
 qqplot x1-x9/ normal (mu=est sigma=est);  
run; 







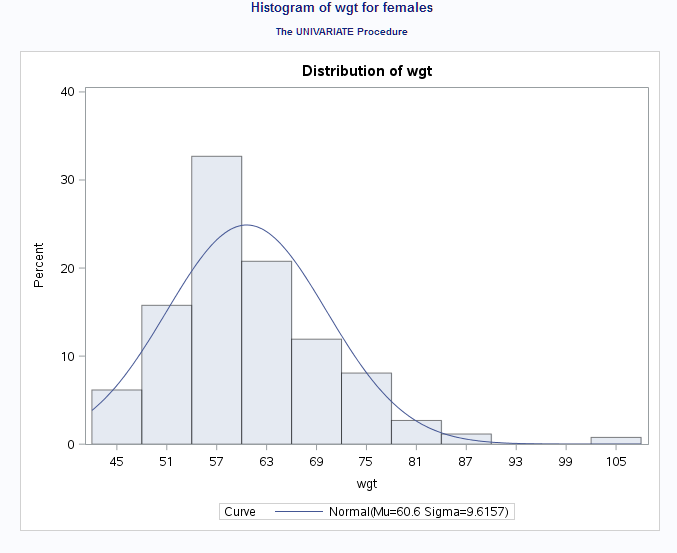


1. That is, do plots provide evidence that the female heights are nearly normal?
   1. Yes, this simulation shows the distribution for female heights are approximately distributed along the linear line.

Exercise 5

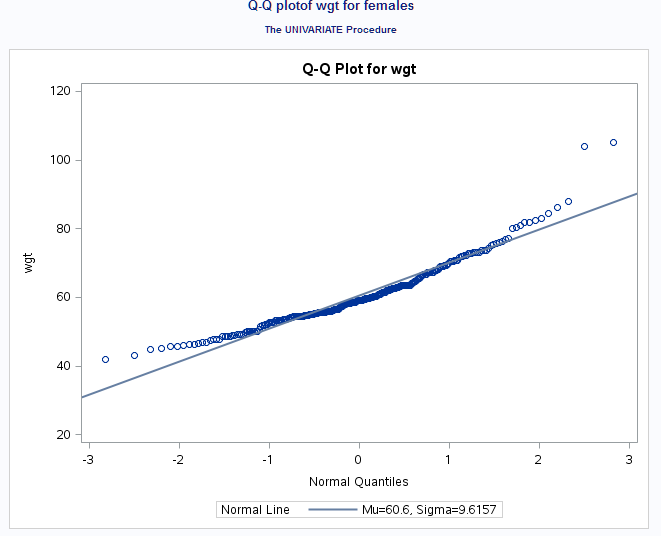
1. Using the same technique, determine whether female weights appear to come from a normal distribution.
   1. Yes, female wgt looks to come from an approximately normal distribution

title 'Histogram of wgt for females';  
ods select histogram;  
proc univariate data=fdims;  
 var wgt;  
 histogram/ normal;  
 output out=estimates n=n mean=mean std=std;  
run;



data \_NULL\_;  
 set estimates;  
 call symputx('n',n);  
 call symputx('mean', mean);  
 call symputx('std', std);  
run;

Title 'Q-Q plotof wgt for females ';  
ODS select qqplot;  
proc univariate data=fdims;  
 var wgt;  
 qqplot wgt/ normal (mu=est sigma=est);  
run;



data sim\_norm\_wgt;  
 do i=1 to &n;  
 x1 = rand('NORMAL', &mean, &std);  
 output;  
 end;  
run;

data fdims\_simulated\_wgt;  
 array x {9} x1-x9;  
 do i=1 to &n;  
 do j=1 to 9;  
 x[j]=rand('NORMAL', &mean, &std);  
 end;  
 output;  
 end;  
run;

Title 'Q-Q plot of Simulated for female wgt';

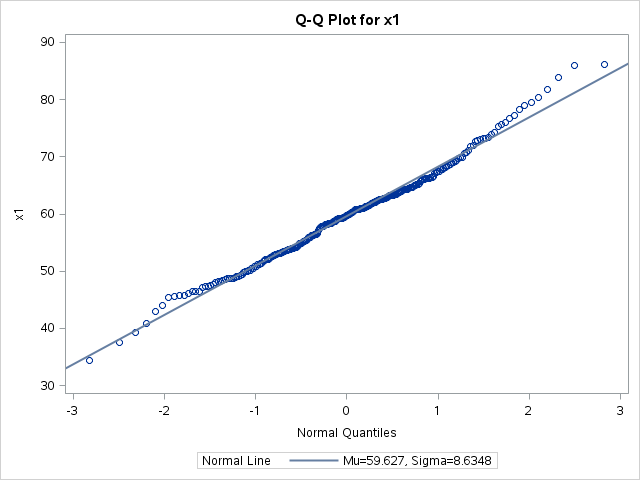
ODS select qqplot;

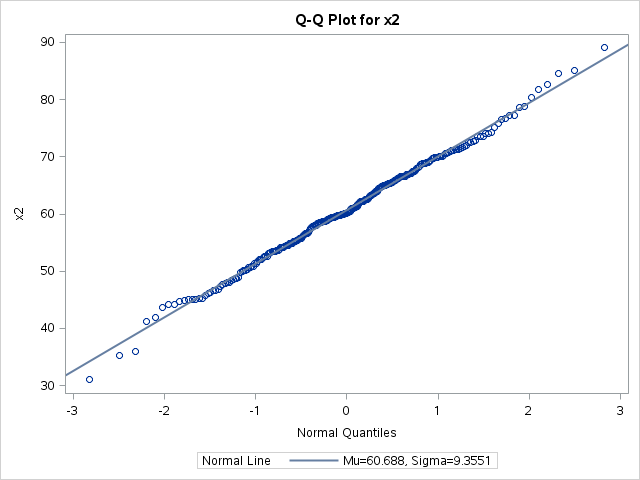
proc univariate data=fdims\_simulated\_wgt;

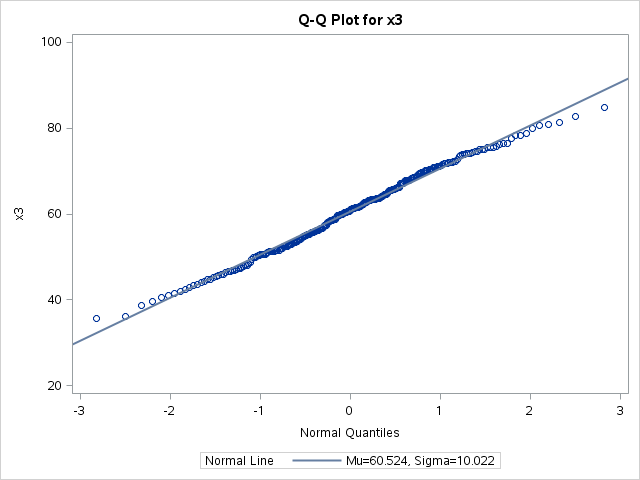
var x1-x9;

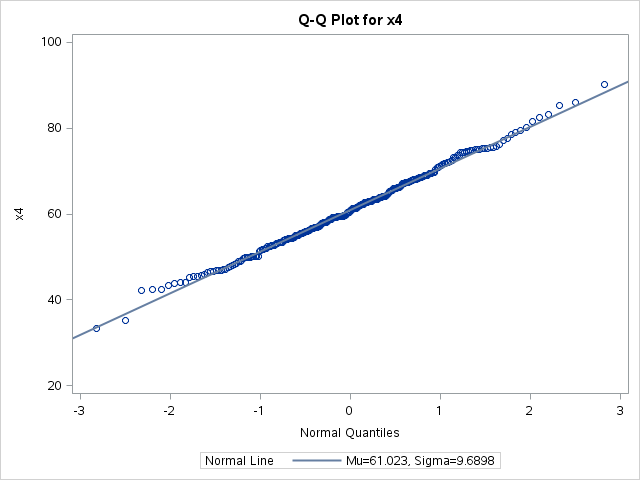
qqplot x1-x9/ normal (mu=est sigma=est);

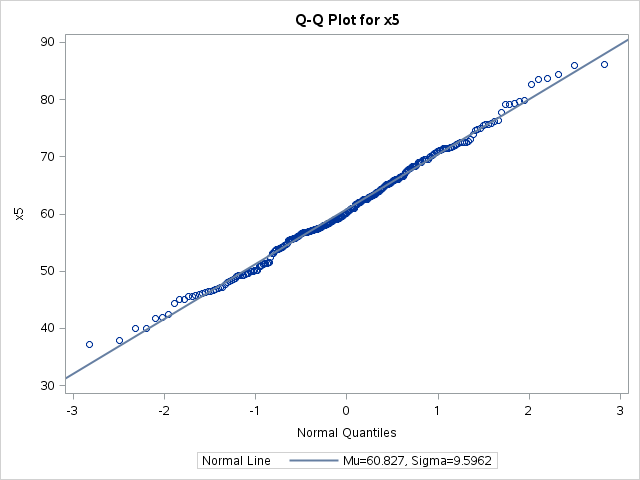
run;

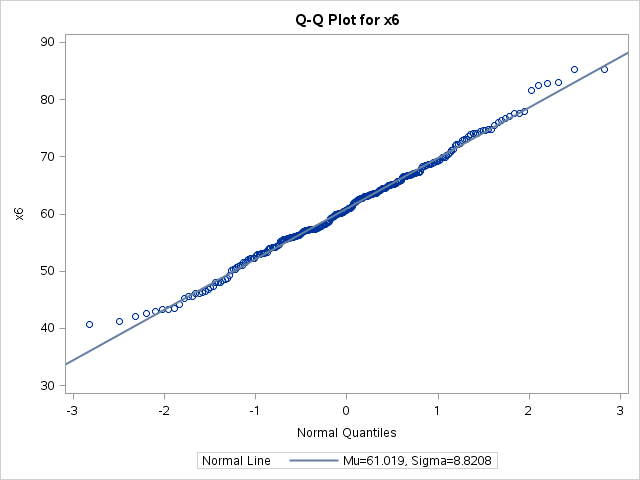


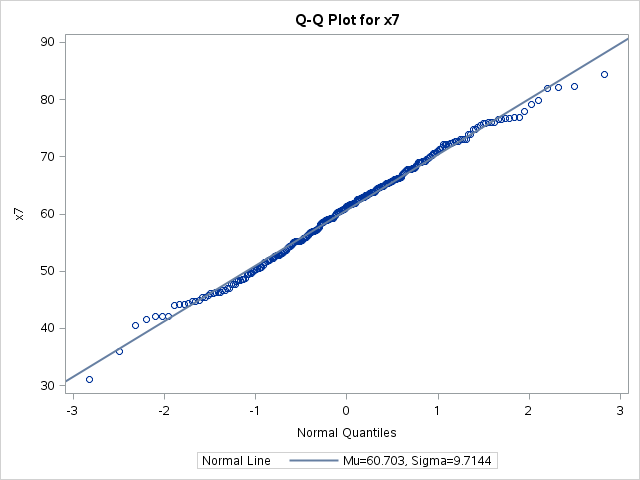


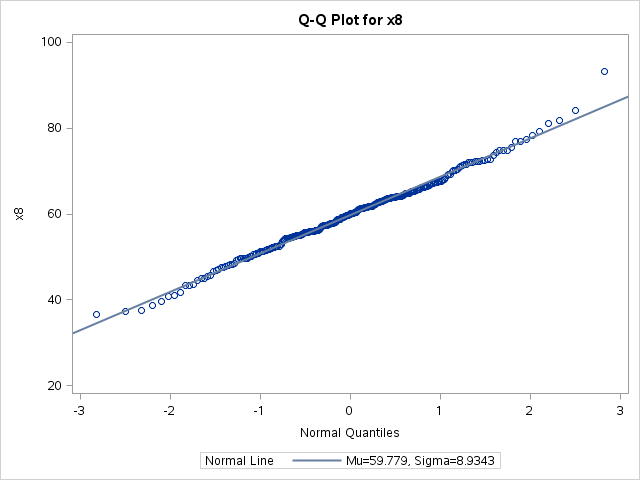


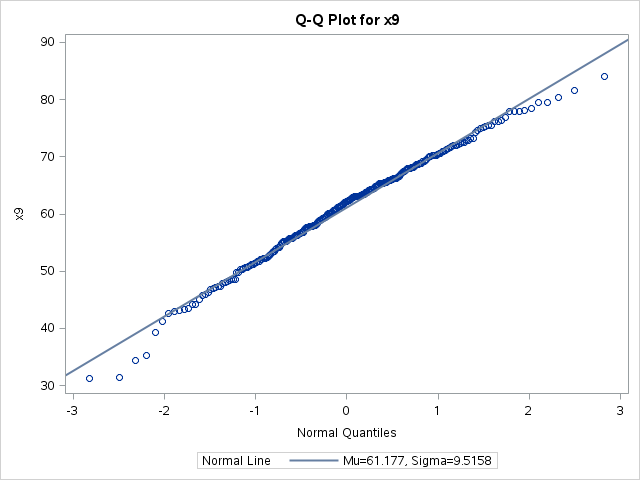








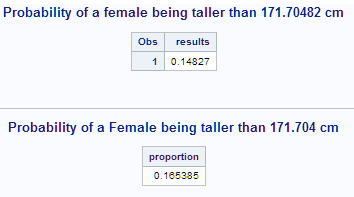




Exercise 6

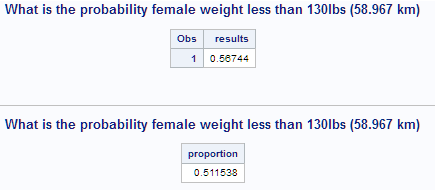
1. Write out two probability questions that you would like to answer; one regarding female heights and one regarding female weights.
   1. What is the probability females are taller than 5.2 ft. (171.704 cm)
   2. What is the probability female weight less than 130lbs (58.967 km)
2. Calculate the probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all).
   1. What is the probability females are taller than 5.2 ft. (171.704 cm)

data temp;  
 results =1 - cdf('NORMAL', 171.704, &mean, &std);  
run;  
  
Title 'What is the probability female weight less than 130lbs (58.967 km’;  
proc print data=temp;  
 var results;  
run;  
  
  
title 'What is the probability female weight less than 130lbs (58.967 km’;  
proc sql;  
 select b.n\_tall /a.n as proportion  
 FROM (select count(hgt)as n from fdims) as a,   
 (select count(hgt)as n\_tall from fdims  
 where hgt > 171.704) as b;  
 quit;  
  
title;



* 1. What is the probability female weight less than 130lbs (58.967 km)

data temp;  
 results =1 cdf('NORMAL', 58.967, &mean, &std);  
run;  
  
Title 'What is the probability female weight less than 130lbs (58.967 km)';  
proc print data=temp;  
 var results;  
run;  
  
  
title 'What is the probability female weight less than 130lbs (58.967 km)';  
proc sql;  
 select b.n\_tall /a.n as proportion  
 FROM (select count(hgt)as n from fdims) as a,   
 (select count(hgt)as n\_tall from fdims  
 where wgt > 58.967) as b;  
 quit;  
  
title;



1. Which variable, height or weight, had a closer agreement between the two methods.
   1. What is the probability female weight less than 130lbs (58.967 km)