**Tor Thogersen**

**DA 460 – Fall 2017**

**Lab 4 - Handout 4 R and Handout 4 SAS**

**Part 4 – R Handout**

**Exercise 1**

1. Describe this population distribution
   1. The population distribution looks to be approximately normally distributed with skew to the right due to 1.272805 and the upward “U” shape of the distribution, it looks like there is a correlation between living area and sale price, as one goes up or down the other variable responds accordingly.

> summary(area)

Min. 1st Qu. Median Mean 3rd Qu. Max.

334 1126 1442 1500 1743 5642

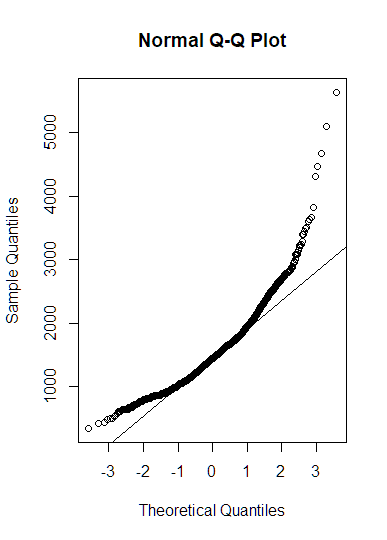
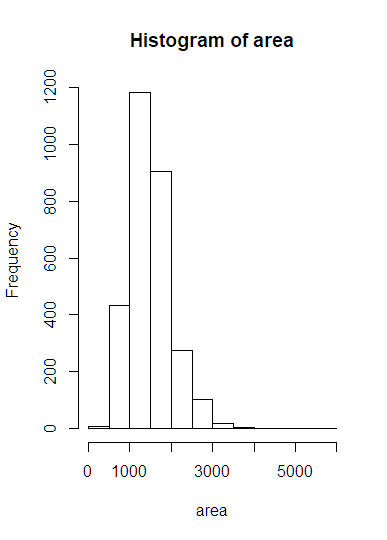
> hist(area)

> skewness(area)

[1] 1.272805

> qqnorm(area)

> qqline(area)



**Exercise 2**

1. Describe the distribution of this sample.
   1. The sample dist is still approximately normalized with a few outliers that may need to be researched and modified for fit, I see a slight skewness to the right.
2. How does it compare to the distribution of the population.
   1. The bulk of the sample data is within the range of 1000 to 1700, the bell shape spread is wider versus the population shape with a slightly smaller skewness.

> samp1 <-sample(area, 50)

> mean(samp1)

[1] 1514.7

> summary(samp1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

848 1176 1456 1515 1788 2698

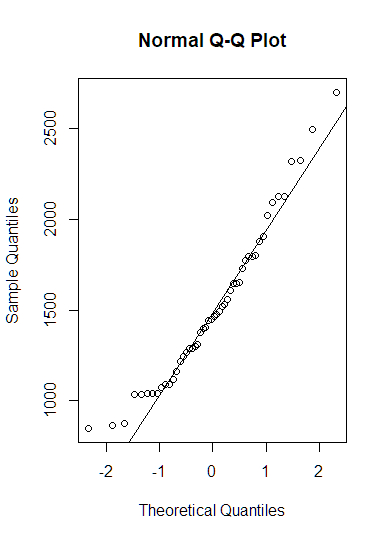
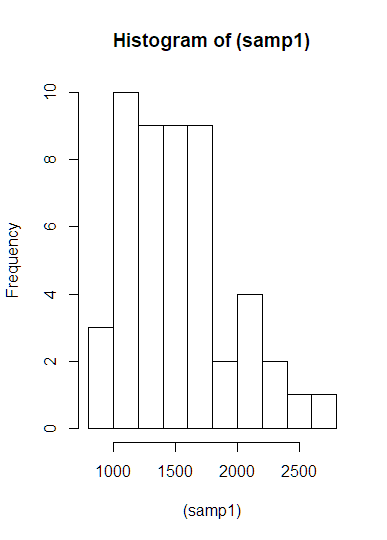
> skewness((samp1))

[1] 0.6636443

> hist((samp1))

> qqnorm(samp1)

> qqline(samp1)



**Exercise 3**

1. Take a second Sample (size 50) and call it samp2.

> samp2 <- sample(area, 50)

> mean(samp1)

[1] 1514.7

> mean(samp2)

[1] 1506.06

1. How does the mean of samp2 compare with mean of samp1.
   1. Smaller
2. Suppose we took two more samples, one of size 100 and one of size 1000. Which would you think would provide a more accurate estimate of the population mean.
   1. I would normally say the larger sample of 1000 would give you closer approximation of the population distribution because the sample size would be closer to the actual population data sample.

> samp100 <-sample(area, 100)

> samp1000 <-sample(area,1000)

> summary(samp100)

Min. 1st Qu. Median Mean 3rd Qu. Max.

630 1139 1436 1496 1762 2855

> summary(samp1000)

Min. 1st Qu. Median Mean 3rd Qu. Max.

407 1114 1430 1496 1754 5642

> mean(samp100)

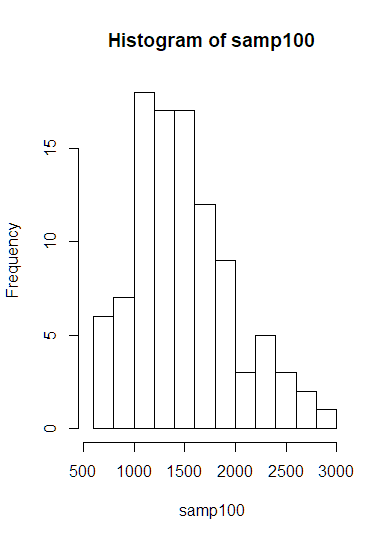
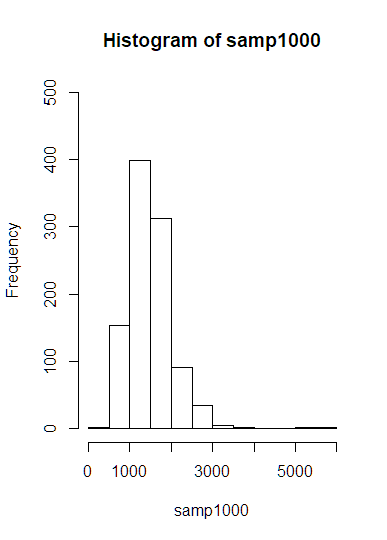
[1] 1495.71

> mean(samp1000)

[1] 1495.523

> hist(samp100, xlim = c(550,3000))

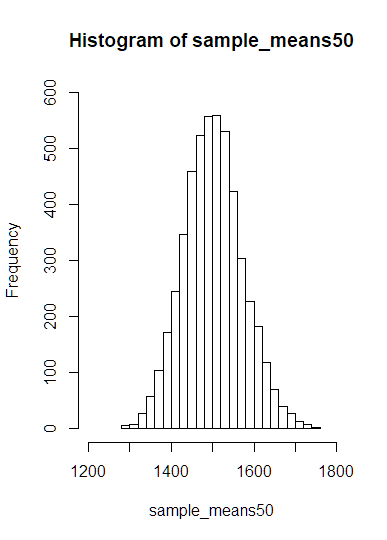
> hist(samp1000, xlim = c(0,6000))

**Exercise 4**

1. How many elements are there in sample\_means50.
   1. 5000 elements
2. Describe the sampling distribution and be sure specifically note its center.
   1. The Sample dist of the sample means has very good Bell Curve shape, with a spread that is balanced on from center, with center being at 1500.
3. Would you expect the distribution to change if we instead collected 50,000 sample means
   1. I would think the larger sample distribution would look very similar to this current distribution with exception of having a larger dataset.

> hist(sample\_means50, breaks = 25, xlim = c(1200, 1800), ylim = c(0,600))



**Exercise 5**

1. To make sure you understand what you’ve done in this loop, try running a smaller version. Initialize a vector of 100 zeros called sample means small. Run a loop that takes a sample of size 50 from area and stores the sample mean in sample means small, but only iterate from 1 to 100. Print the output to your screen (type sample means small into the console and press enter).

> sample.means.small <-rep(0, 100)

> for(i in 1:100) {

+ samp <-sample(area, 50)

+ sample.means.small[i] <-mean(samp)

+ }

> sample.means.small

[1] 1464.28 1373.56 1538.62 1557.84 1535.88 1605.90 1635.78 1381.36

[9] 1488.28 1475.24 1511.96 1515.76 1569.06 1454.10 1494.36 1509.62

[17] 1439.48 1557.24 1541.32 1401.36 1435.34 1675.64 1526.24 1520.66

[25] 1579.82 1541.10 1470.08 1562.80 1541.30 1575.68 1428.62 1502.58

[33] 1363.60 1387.54 1433.22 1435.80 1416.96 1492.02 1409.48 1478.88

[41] 1452.82 1516.14 1534.60 1448.42 1387.88 1475.30 1559.60 1497.30

[49] 1459.42 1494.50 1365.64 1540.30 1429.90 1495.80 1482.34 1446.68

[57] 1599.72 1392.14 1469.94 1523.22 1481.96 1600.64 1463.58 1618.64

[65] 1462.88 1433.88 1368.10 1410.50 1505.14 1488.54 1432.80 1578.54

[73] 1523.82 1560.14 1563.74 1493.34 1545.84 1479.98 1429.80 1505.54

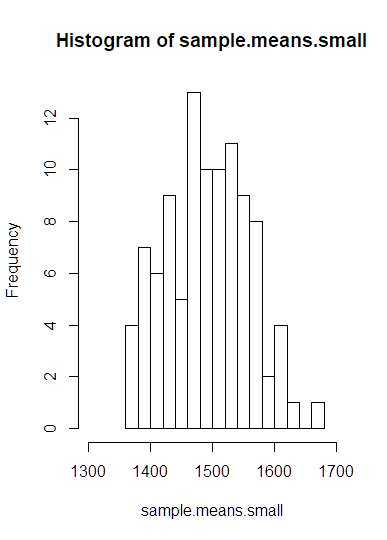
[81] 1548.34 1507.22 1460.42 1478.92 1401.56 1382.58 1508.54 1466.14

[89] 1389.50 1411.48 1522.42 1399.10 1539.40 1533.36 1594.32 1502.36

[97] 1607.70 1476.72 1577.72 1537.70

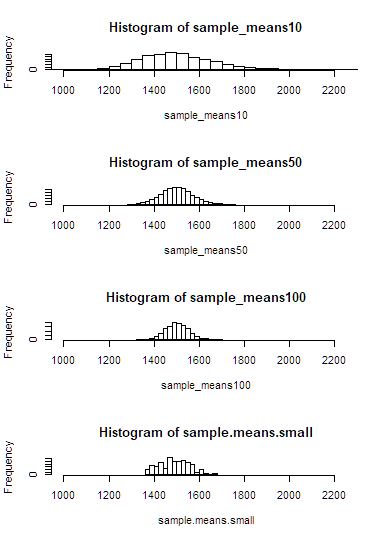
> par(mfrow = c(1,1))

> hist(sample.means.small, breaks = 20, xlim = c(1300, 1700))



1. How many elements are there in this object called sample means mall?
   1. 100 samples of 50 homes in the area
2. What does each element represent?
   1. One sample of the sample mean of 50 homes in the area

**Exercise 6**

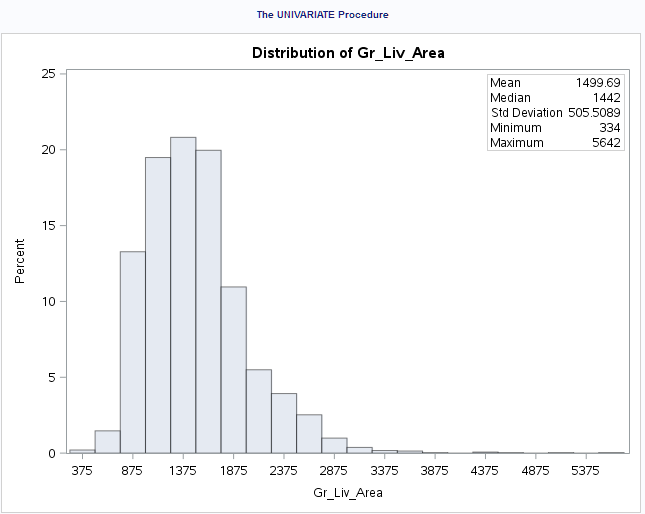
1. When the sample size is larger, what happens to the center?
   1. When the sample size is larger the center will be closer to the true population mean and the overall distribution will look evenly distributed with the center being the middle point of the bell curve.
2. What about the spread?
   1. Spread will disperse evenly on both sides of the mean and spread out wider

**Part 4 - SAS Handout**

**Exercise 1:**

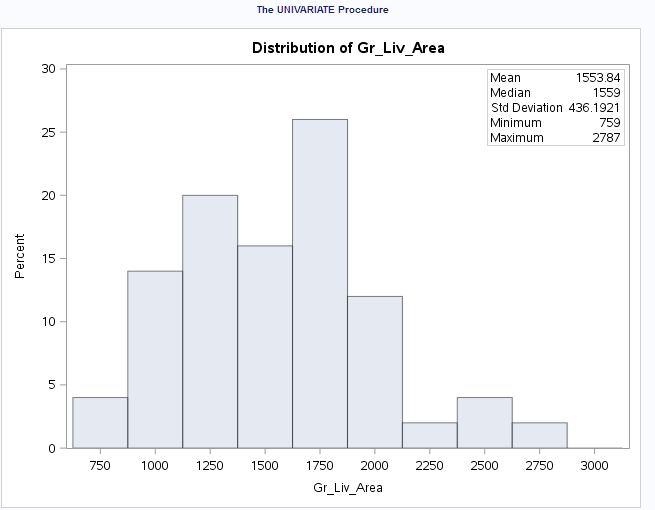
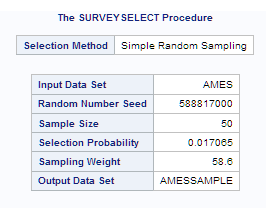
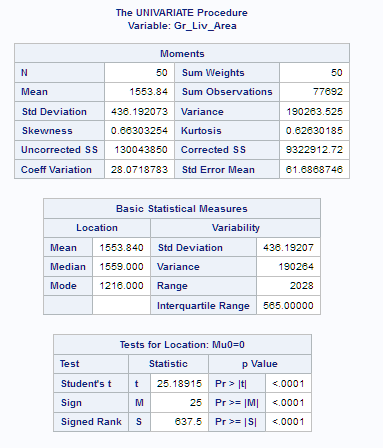
1. Describe this population distribution.
   1. Population dist for Gr\_Liv\_Area has an approximant normal dist that is right skewed with center mass in the 1375 range.

proc univariate data=work.ames;  
 var Gr\_Liv\_Area;  
 histogram Gr\_Liv\_Area;  
 inset mean median std min max / pos=ne;  
  
run;



**Exercise 2:**

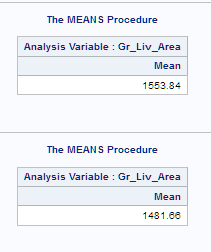
1. Describe the distribution of this sample.
   1. Sample dist histogram is approximently distributed with evenly spread from center, the center mass is located at 1500.
2. How does it compare to the distribution of the population?
   1. Sample Dist mirrors the population dist but with more evenly spread from center.

proc surveyselect   
 data=work.ames  
 out=work.amessample  
 sampsize=50  
 method=srs ranuni;  
RUN;  
  
Proc print data=work.amessample;  
run;  
  
proc means data=work.amessample mean;  
 var Gr\_Liv\_Area;  
RUN;  
  
proc univariate data=amessample;  
 var Gr\_Liv\_Area;  
 histogram Gr\_Liv\_Area / nmidpoints=10;  
 inset mean median std min max / pos=ne;  
run;  

**Exercise 3:**

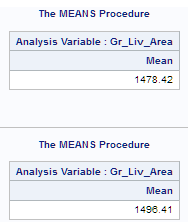
1. Take a second sample, also of size 50, and call it work.amessample2

proc surveyselect   
 data=work.ames  
 out=work.amessample2  
 sampsize=50  
 method=srs ranuni;  
RUN;  
  
proc means data=amessample mean;  
 var Gr\_Liv\_Area;  
run;  
  
proc means data=amessample2 mean;  
 var Gr\_Liv\_Area;  
run;



1. How does the mean of work.amessample2 compare with the mean of work.amessample?
   1. Smaller, due to different random data being used in the new sample
2. Suppose we took two more samples, one of size 100 and one of size 1000,
   1. The data from the larger sample will be closure to the population mean
3. which would you think would provide a more accurate estimate of the population mean?
   1. The larger the sample the closer to the population sample thus it would provide a more complete estimate.

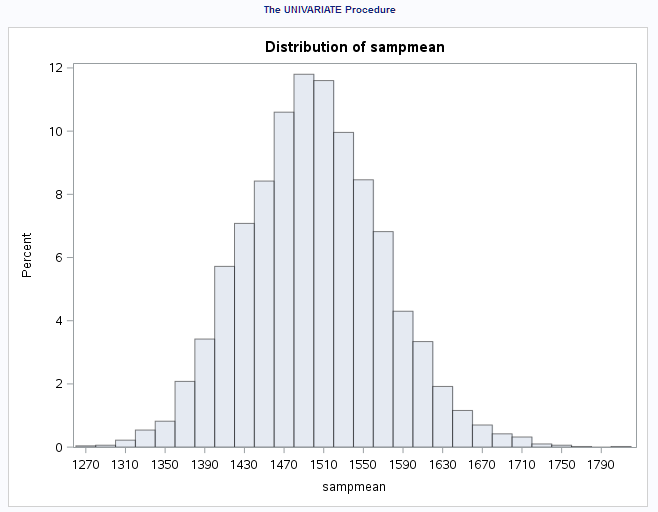
proc surveyselect   
 data=ames  
 out=amessample100  
 sampsize=100  
 method=srs ranuni;  
RUN;  
  
proc surveyselect   
 data=ames  
 out=amessample1000  
 sampsize=1000  
 method=srs ranuni;  
RUN;  
  
proc means data=amessample100 mean;  
 var Gr\_Liv\_Area;  
run;  
  
proc means data=amessample1000 mean;  
 var Gr\_Liv\_Area;  
run;



**Exercise 4:**

1. How many observations are there in work.reprun?
   1. 5000
2. Describe the sampling distribution, and be sure to specifically note its center.
   1. The Sample dist of the sample means has very good Bell Curve shape, with a spread that is balanced from center, with center being between 1470 to 1510.
3. Would you expect the distribution to change if we instead collected 50,000 sample means?
   1. Yes, it would again assume approximation closure to Population data and population mean.

proc surveyselect   
 data=ames  
 out=work.amessampler  
 sampsize=50  
 method=srs   
 reps=5000  
 ranuni;  
   
RUN;  
  
Proc means data=amessampler mean noprint;  
 by replicate;  
 var Gr\_Liv\_Area;  
 output out=reprun  
 mean = sampmean;  
run;  
  
Proc univariate data=reprun;  
 var sampmean;  
 histogram sampmean;  
run;

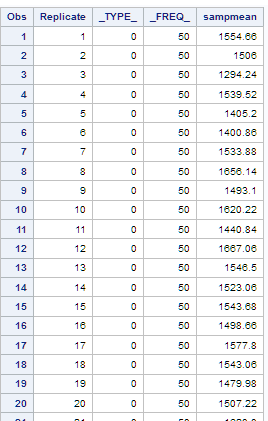
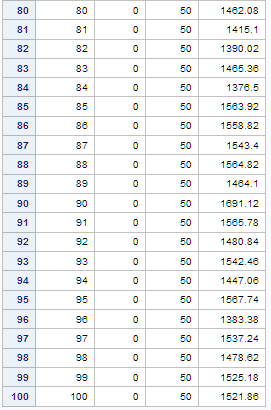


**Exercise 5:**

1. To make sure that you understand what you’ve done, try running a smaller version. This time, take only 100 samples of size 50 from the population. Change the data set by saving the sample means to work.reprunsmall. Print the storage data set to your screen.

proc surveyselect   
 data=ames  
 out=work.reprunsmall  
 sampsize=50  
 method=srs   
 reps=100  
 ranuni;  
   
RUN;  
  
Proc means data=reprunsmall mean noprint;  
 by replicate;  
 var Gr\_Liv\_Area;  
 output out=reprunsmall  
 mean = sampmean;  
run;  
  
Proc univariate data=reprunsmall;  
 var sampmean;  
 histogram sampmean;  
run;

proc print data=reprunsmall;  
run;

**Exercise 6:**

1. When the sample size is larger, what happens to the center?
   1. When the sample size is larger the center will be closer to the true population mean and the overall distribution will look evenly distributed with the center being the middle point of the bell curve.
2. What about the spread?
   1. When the sample size is larger, the spread becomes smaller

proc univariate data=work.reprun;   
 title 'Sample Size = 50';   
 var sampmean;   
 histogram sampmean;   
run;   
proc univariate data=work.reprun10;   
 title 'Sample Size = 10';   
 var sampmean;   
 histogram sampmean;   
run;   
proc univariate data=work.reprun100;   
 title 'Sample Size = 100';   
 var sampmean;   
 histogram sampmean;   
run;   
title;   
  
  
data work.allsamples;   
 set work.reprun (IN=Sample1)   
 work.reprun10 (IN=Sample2)  
 work.reprun100 (IN=Sample3);   
 group = 1\*(Sample1) + 2\*(Sample2) + 3\*(Sample3);   
run;   
proc sgpanel data=work.allsamples;   
 panelby group;   
 histogram sampmean;   
run;

