MS -Business Intelligence & Analytics

Fall 2015

BIA – 652 C

September 30, 2015

Mohit Ravi Ghatikar

CWID - 10405877

Multivariate Data Analytics – Homework 1

Ethics Statement

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature _Mohit Ravi Ghatikar_____ Date: 9/30/2015___

1) A) Prove E(a+bx) = a + bE(x)

We can write E(a+bx) = E(a) + E(bx)

We know that E(a) = a (i.e The mean of a constant is the constant itself) and ,

E(bx) = $\sum$ bx * p(bx) (From the definition of expectation of a variable)

= b $\sum$ x * p(x)

= b * E(x)

Therefore, E(a+bx) = a + b*E(x)

B) Prove Var(a+bx) = b^2 Var(x)

We know that Var(x) = $E[x^2] - E[x]^2$

Var(a+bx) = $E[(a+bx) - E(a+bx)]^2$ …… (Var(x) = $\sum(x-u)^2$)

= $E [a - E(a) + bx - E(bx)]^2$

= $E [a - a + bx - E(bx)]^2$

= $E [b (x - E(x)]^2$

= $b^2 E(x - E(x)]^2$

= b^2 Var(x)

2) A) If X and Y are independent, Prove that Covariance(x,y) = 0

Cov(x,y) = $\sum (x - u_x) (y - u_y) p(x,y)$

= $\sum (x - u_x) (y - u_y) p(x) p(y)$ ……(P(x and Y) = P(X) * P(Y) since both X and Y are independent)

= $\sum (x - u_x) p(x) \sum (y - u_y) p(y)$

= $\sum (xp(x) - u_x) \sum yp(y) - u_y$

= $\sum (u_x - u_x) \sum (u_y - u_y)$ …… (Since E(x) = u_x = xp(x))

= $\sum (0 -0) \sum (0 -0)$

= 0

B) If X and Y are independent, Prove that Var(x+y) = Var(x) + Var(y)

We know that $Var(x) = E[x^2] – E[x]^2$

$Var(x+y) = E[(x+y)^2] – E[(x+y)]^2$

$= E[(x^2+2xy+y^2) – (u_x + u_y)^2$

$= E[(x^2+2xy+y^2) - u_x^2 – u_y^2 - 2 u_x u_y$

$= E(x^2) - u_x^2 + E(y^2) – u_y^2 + 2 E(xy) - 2 u_x u_y$

$= E(x^2) - u_x^2 + E(y^2) – u_y^2 + 2 u_x u_y - 2 u_x u_y$

$= Var(x) + Var(y)$

3) Calculate the mean and standard deviation of a continuous uniform distribution of x between a and b.

PDF of Uniform distribution = 1 / (b-a) for X belonging to (a,b)

0 if otherwise

The mean of continuous distribution = $\int_a^b x * p(x)\, dx$

$= \int_a^b x * (1/b-a)\, dx$

$= 1 / (b-a) * (x^2 / 2\, dx)$ from a to b

$= 1 / (b-a) * (b^2/2 – a^2/2)$

$= 1 / (b-a) * ((b+a) (b-a)) / 2$

$= (b+a) / 2$……. ((b-a) gets cancelled)

The variance of continuous distribution = $E[x^2] – E[x]^2$

$E[x^2] = \int_a^b x^2 * p(x)\, dx$

$= \int_a^b x^2 * (1/b-a)\, dx$

$= (1/b-a) * (x^3 / 3\, dx)$ from a to b

$= (b^3 - a^3) / 3 (b - a)$

$= (b - a)(a^2 + ab + b^2) / 3 (b-a)$

$= (a^2 + ab + b^2) / 3$

$Var(x) = (a^2 + ab + b^2) / 3 - (a+b)^2 / 4$ (Since $E[x] = (a+b) / 2$)

$= (a^2 + ab + b^2) / 3 - (a^2 + 2ab + b^2) / 4$

$= (4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2) / 12$

$= (b^2 - 2ab + a^2) / 12$

$= (b - a)^2 / 12$

4) The code and results for the SAS dataset is explained below

a) The mean of the first sample

```
* We define the library name where the permanent SAS datasets
are stored;
Libname SASHW1 "C:\Users\mohit\Documents\MySAS
Files\9.4\Solution";

* We copy the S_twentyfive__1 dataset into a temporary dataset
called First_dataset which is stored in the work folder;

data First_dataset;
set SASHW1.S_twentyfive__1;
run;

*/ We find the mean, standard deviation and variance of the
first_datatset and store the output of the mean into a separate
dataset called means_of_first_dataset /*;

title "Mean of First sample";
proc means data=First_dataset n mean std var vardef=n;
* We set the divisor of standard deviation to be the number of
observations since the default value is the degrees of freedom;
var income;
output out=means_of_first_dataset mean=income_mean;
run;
```

Result window : The mean is **47792.64**

Mean of First sample

The MEANS Procedure

Analysis Variable : income			
N	Mean	Std Dev	Variance
25	47792.64	9700.83	94106139.67

b) The variance is **94106129.67** and Standard deviation **is 9700.83**

c) To estimate the mean, Variance and Standard deviation of the population:

The sample mean will be approximately equal to the population mean. The population Variance can be estimated by taking the divisor of sample variance to be N-1. We can calculate the population Standard deviation by taking the square root of the population variance.

```
title " Estimating the variance of the population";
proc means data=First_dataset n mean std var;
* The default value of divisor for standard deviation is the
degrees of freedom (n-1);
var income;
output out=means_of_first_dataset mean=income_mean;
run;
```

Result window:

Estimating the variance of population

The MEANS Procedure

Analysis Variable : income			
N	Mean	Std Dev	Variance
25	47792.64	9900.87	98027228.82

We can estimate the population Variance = **98027228.82** and population Standard deviation = **9900.87.** The population mean is estimated to be = **47792.64.**

d) Estimating the mean, variance and standard deviation of the sample means.

The mean of the sample means will be approximately equal to population mean. We can estimate the variance of the sample means from the central limit theorem.

Variance of sample means = variance of the population / Sample size.

Standard deviation of sample means = Standard deviation of the population / (sample size) ^ 0.5

Variance of sample means = **98027228.82 / 25**

= **3921089.153**

Standard deviation of sample means = **9900.87 / 25^0.5**

= **1980.174**

Mean of sample means = **47792.64**

e) Calculate the means of the remaining samples: (From 2nd dataset till 10th dataset)

```
*/ To calculate the remaining means of the datasets, we store the
results of the means in temporary datasets. This step is repeated
for the 2nd dataset till the 10th dataset;

data Second_dataset;
set SASHW1.S_twentyfive__2;
run;

proc means data=Second_dataset noprint;
var income;
output out=means_of_second_dataset mean=income_mean ;
run;
```

```
data Third_dataset;
set SASHW1.S_twentyfive__3;
run;

proc means data=Third_dataset noprint;
var income;
output out=means_of_third_dataset mean=income_mean ;
run;

data Fourth_dataset;
set SASHW1.S_twentyfive__4;
run;

proc means data=Fourth_dataset noprint;
var income;
output out=means_of_Fourth_dataset mean=income_mean ;
run;

data Fifth_dataset;
set SASHW1.S_twentyfive__5;
run;

proc means data=Fifth_dataset noprint;
var income;
output out=means_of_fifth_dataset mean=income_mean ;
run;

data Sixth_dataset;
set SASHW1.S_twentyfive__6;
run;


proc means data=Sixth_dataset noprint;
var income;
output out=means_of_sixth_dataset mean=income_mean ;
run;

data Seventh_dataset;
set SASHW1.S_twentyfive__7;
run;

proc means data=Seventh_dataset noprint;
var income;
output out=means_of_Seventh_dataset mean=income_mean ;
run;
```

```
data Eighth_dataset;
set SASHW1.S_twentyfive__8;
run;

proc means data=Eighth_dataset noprint;
var income;
output out=means_of_eighth_dataset mean=income_mean ;
run;

data Ninth_dataset;
set SASHW1.S_twentyfive__9;
run;

proc means data=Ninth_dataset noprint;
var income;
output out=means_of_ninth_dataset mean=income_mean ;
run;

data tenth_dataset;
set SASHW1.S_twentyfive__10;
run;

proc means data=Tenth_dataset noprint;
var income;
output out=means_of_tenth_dataset mean=income_mean ;
run;
```

```
*/ we append the means stored in individual datasets into a single
dataset called total_means using the set function;

data total_means;

set
means_of_second_dataset means_of_third_dataset
means_of_Fourth_dataset means_of_fifth_dataset
means_of_sixth_dataset means_of_Seventh_dataset
means_of_eighth_dataset means_of_ninth_dataset
means_of_tenth_dataset
;
        run;
```

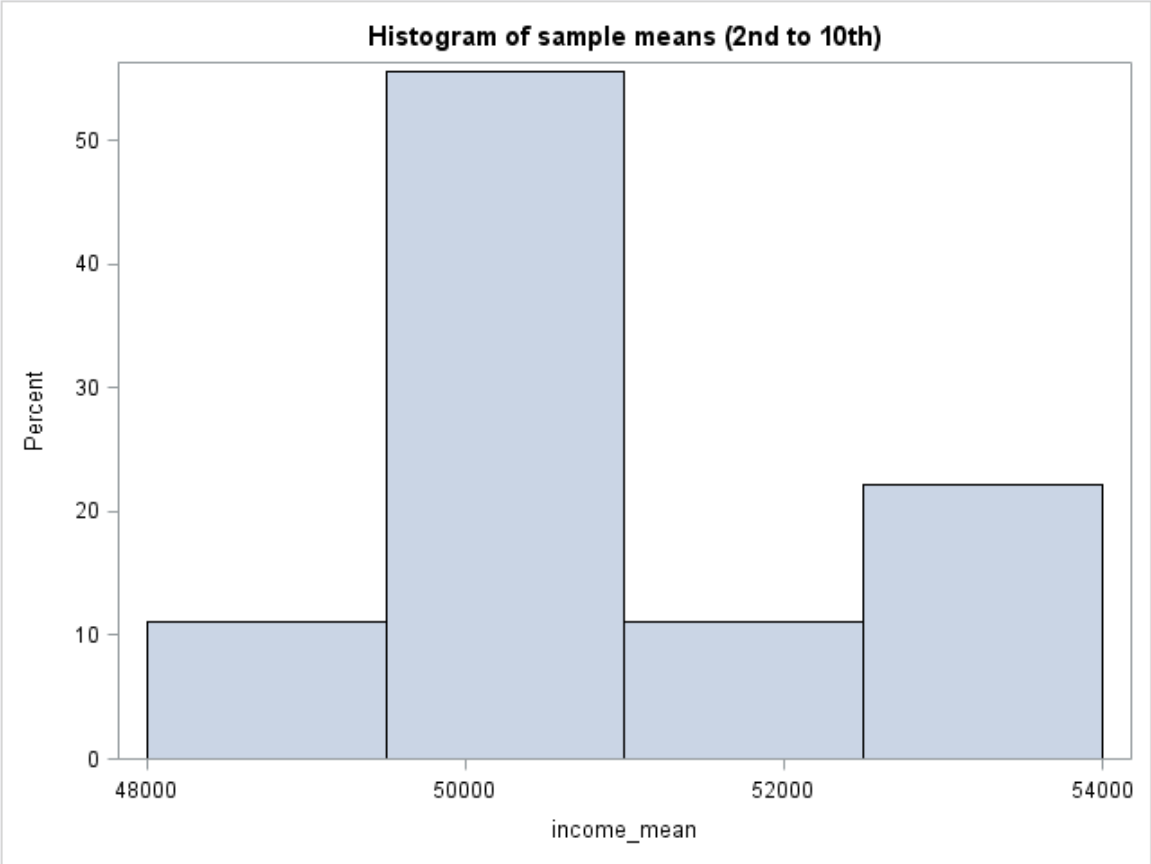
```
*/ Plot the histogram of the sample means from 2nd dataset to 10th
dataset;

title "Histogram of sample means (2nd to 10th)";
proc sgplot data=total_means;
histogram income_mean ;
        run;
```

Results window:

Total_means dataset containing the income_means from 2nd till 10th datasets.

	FREQ	income_mean
1	25	50398.72
2	25	50938.76
3	25	51760.04
4	25	50820.28
5	25	48174.04
6	25	50232.96
7	25	53240.32
8	25	50234.72
9	25	52762.28

The histogram plot of sample means from 2nd till 10th datasets.



f) To plot the means of the samples from 1st dataset to 10th dataset.

```
*/ we append the means stored in individual datasets (1st to
10th) into a single dataset called total_means1
using the set function ;
data total_means1 ; drop _type_;
set means_of_first_dataset  means_of_second_dataset
means_of_third_dataset means_of_Fourth_dataset
means_of_fifth_dataset means_of_sixth_dataset
means_of_Seventh_dataset means_of_eighth_dataset
means_of_ninth_dataset means_of_tenth_dataset
;
run;

*/ Plot the histogram of the sample means from 1st dataset to
10th dataset;
title "Histogram of sample means (1st to 10th)";
proc sgplot data=total_means1;
histogram income_mean ;
```

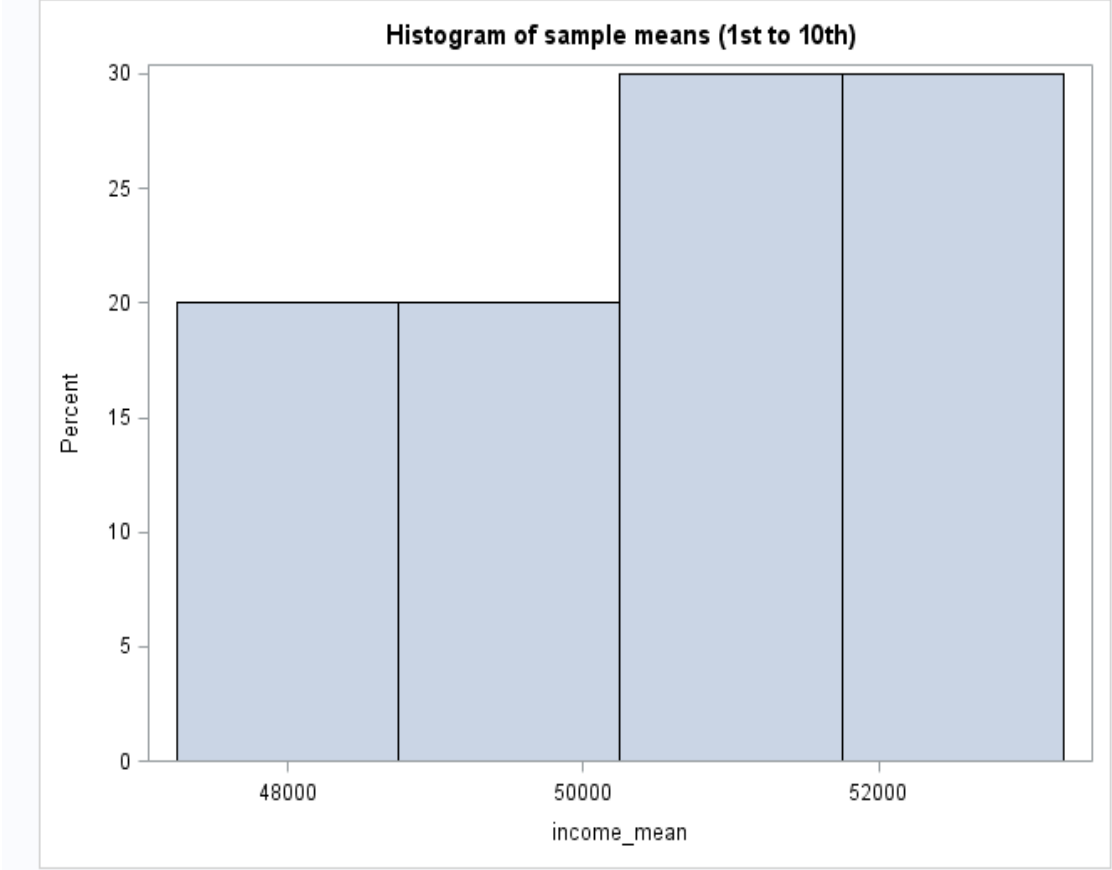
```
run;
```

Results Window:

Total_means1 dataset containing the income_means from 1st till 10th datasets.

	FREQ	income_mean
1	25	47792.64
2	25	50398.72
3	25	50938.76
4	25	51760.04
5	25	50820.28
6	25	48174.04
7	25	50232.96
8	25	53240.32
9	25	50234.72
10	25	52762.28

The histogram plot of sample means from 1st till 10th datasets.



g) To calculate the variance and standard deviation of the sample means from 2nd dataset till 10th dataset.

```
* To calculate the Variance and standard deviation of
sample means from 1st till 10th datasets;
title " Variance and standard deviation of sample means ";
proc means data=total_means n mean std var vardef=n;
* We set the divisor of standard deviation to be the number of
observations since the default value is the degrees of freedom;
var income_mean ;
        run;
```

Results window:

Variance and standard deviation of sample means

The MEANS Procedure

Analysis Variable : income_mean			
N	Mean	Std Dev	Variance
9	50951.35	1423.56	2026518.00

We calculate the Standard deviation = **1423.56** and Variance = **2026518.00**

h) Comparing the results of f) and g) to d)

We compare the variance and standard deviation of the sample means obtained from 2nd dataset till 10th dataset in **g)** to the estimated variance and standard deviation in **d).**

The mean of the sampling distribution is **505951.35,** whereas the estimated mean was **47792.64.** Therefore there is a sampling error in this case.

Similarly, the variance and standard deviation of the sampling distribution is **2026518.00** and **1423.56.** We estimated the variance and standard deviation to be equal to **3921089.153** and

1980.174. There is a variation in the values of variance and standard deviation of the sampling distribution when compared to the estimated parameters. **This could be possible because we haven't included the first sample in the sampling distribution**.

Again, we compare the compare the variance and standard deviation of the sample means obtained from 1st dataset till 10th dataset in **f)** to the estimated variance and standard deviation in **d).**

We can find out the various parameters of dataset containing all the sample means.

```
*/ We find the mean, standard deviation and variance of all the
samples stored in total_means1 dataset /*;
title "Means,Variance and Standard deviation of all the
samples";
proc means data=total_means1 n mean std var vardef=n;
* We set the divisor of standard deviation to be the number of
observations since the
default value is the degrees of freedom;
var income_mean;

run;
```

Result window:

Means,Variance and Standard deviation of all the samples

The MEANS Procedure

Analysis Variable : income_mean			
N	Mean	Std Dev	Variance
10	50635.48	1649.80	2721834.70

The mean of this sampling distribution which is equal to **50635.48** , whereas the estimated mean is **47792.4.**

Similarly, the Variance and standard deviation of this sampling distribution is equal to **2721834.70** and **1649.80**. We estimated the variance and standard deviation to be equal to **3921089.153** and **1980.174.** There is a variation in the parameters estimated with that of the sampling distribution parameters.

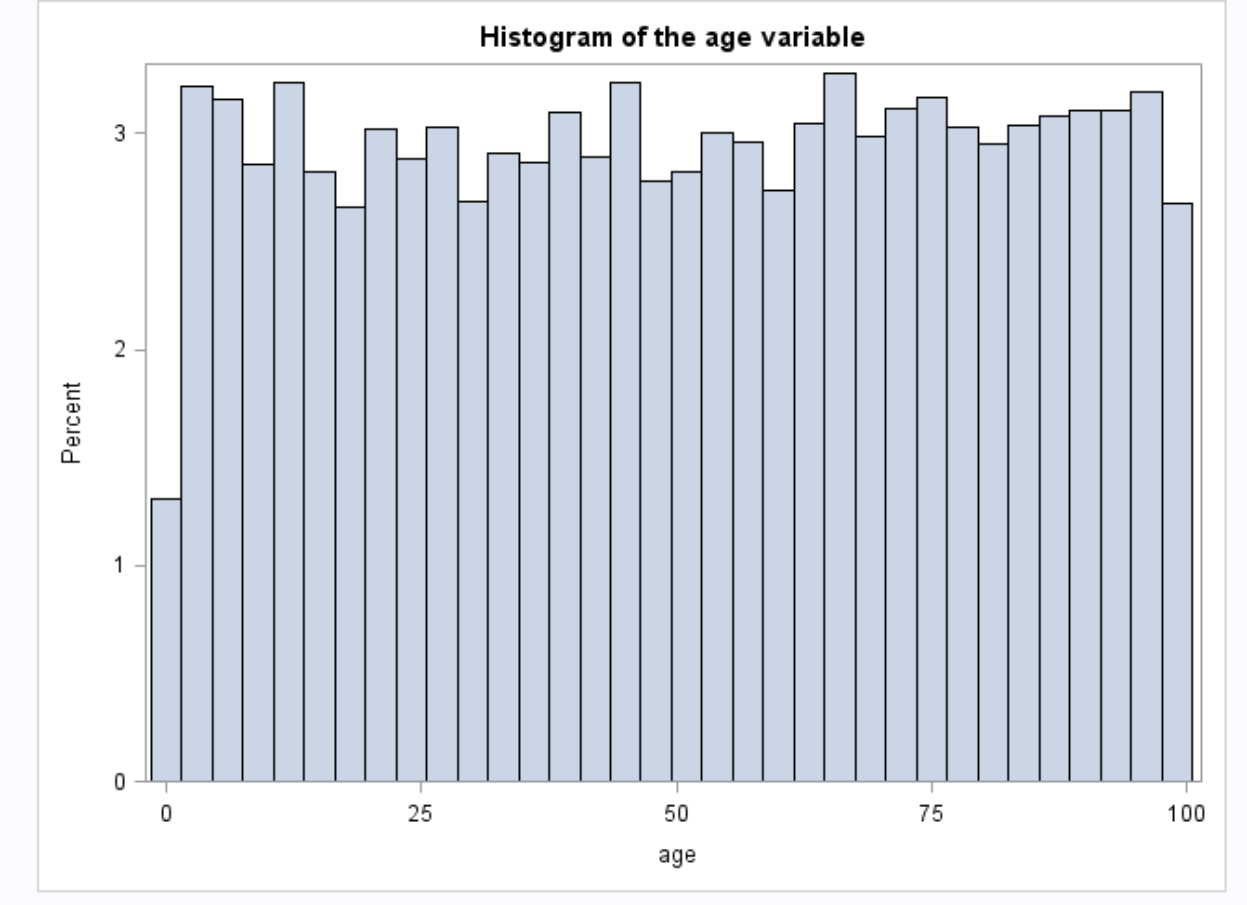
5) Performing operations on age_population dataset.

a) To plot the histogram of the age of the population.

```
* We define the library name where the age_popultion dataset is
stored;
libname SASHW1 "C:\Users\mohit\Documents\My SAS
Files\9.4\Solution";

* We copy the age_population dataset into a temporary dataset
called age_dataset which is stored in the work folder;
data age_dataset;
set SASHW1.age_population;
run;

*/ Plot the histogram of the age of the population;
title "Histogram of the age variable";
proc sgplot data=age_dataset;
histogram age ;
run;
```

Result Window:

Histogram of the Age variable:



b) To Take 20 samples of size 10 and calculate the mean and the variance of each sample

```
*/ We use proc surveyselect option to generate random samples.
Here we select the method of sampling to be simple random
sampling. It generates random samples without replacement. The
sample size is chosen to be 10 and this
process is replicated 20 times ;

proc surveyselect data = age_dataset
method = SRS rep = 20
sampsize = 10 out = Random_samples_of_age;
id age;
run;


* The random samples of age are stored in the
Random_samples_of_age dataset. The replicate coloumn has a range
from 1 to 20 since we are taking 20 samples. The size of each
```

```
replicate ID is 10. Therefore we calculate the mean, standard
deviation and variance of the samples with the same replicate
number. This can be done using the by statment.
We store these values in a new dataset called
Parameters_of_age.;
proc means data = Random_samples_of_age n mean std var;
var age;
output out = Parameters_of_age mean=average_of_age std=
standard_deviation_of_age var=variance_of_age;
by replicate;
run;
```

Result Window:

Random_samples_of_age dataset containing the random samples of age.

	Sample Replicate Number	age
1	1	29
2	1	1
3	1	12
4	1	45
5	1	17
6	1	27
7	1	98
8	1	88
9	1	29
10	1	99
11	2	81
12	2	88
13	2	3
14	2	87
15	2	94
16	2	15
17	2	41
18	2	78
19	2	52
20	2	73

Parameters_of_age dataset containing the mean and variance of each sample.

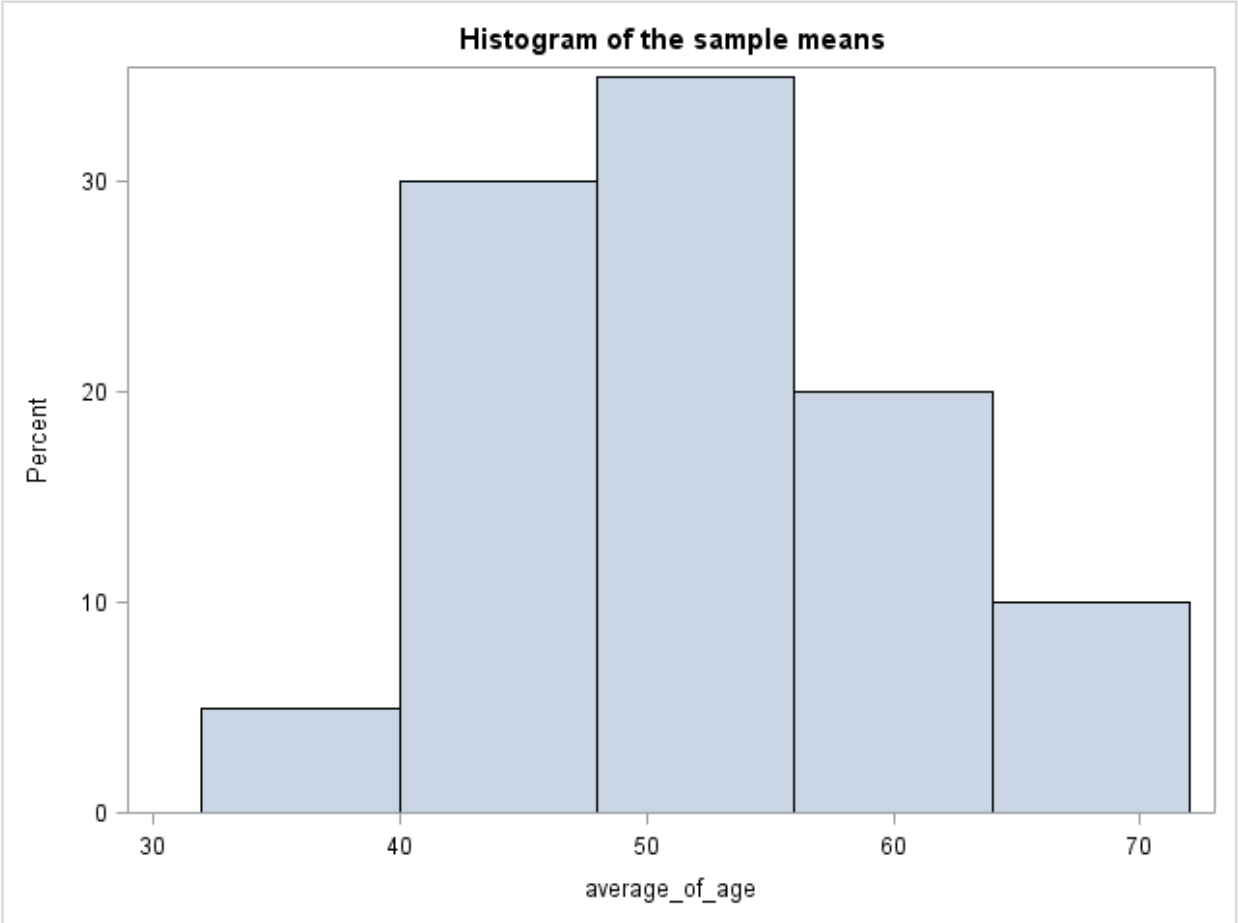
	Sample Replicate Number	_TYPE_	_FREQ_	average_of_age	standard_deviation_of_age	variance_of_age
1	1	0	10	44.5	36.842758975	1357.3888889
2	2	0	10	61.2	32.124065053	1031.9555556
3	3	0	10	57.5	39.691168903	1575.3888889
4	4	0	10	59.7	35.618503306	1268.6777778
5	5	0	10	57.5	26.395496249	696.72222222
6	6	0	10	45.1	35.381884882	1251.8777778
7	7	0	10	33.4	19.51181978	380.71111111
8	8	0	10	55.7	25.219260717	636.01111111
9	9	0	10	52.1	22.213359344	493.43333333
10	10	0	10	53.1	15.234828519	232.1
11	11	0	10	44.6	32.479737273	1054.9333333
12	12	0	10	50.5	29.740544716	884.5
13	13	0	10	47.5	37.865991778	1433.8333333
14	14	0	10	47.8	26.355686715	694.62222222
15	15	0	10	69.5	22.272304675	496.05555556
16	16	0	10	51.8	34.476400946	1188.6222222
17	17	0	10	47.5	25.321269057	641.16666667
18	18	0	10	52.7	28.952259555	838.23333333
19	19	0	10	64.4	23.852323437	568.93333333
20	20	0	10	50.7	36.478456351	1330.6777778

c) To plot the sample means and calculate the mean and standard deviation of the means

```
* To plot the histogram of sample means;
title"Histogram of the sample means";
proc sgplot data=Parameters_of_age;
histogram average_of_age;
run;

* To calculate the mean and standard deviation of the sample
means;
title" Mean and Standard deviation of the sample means";
proc means data = Parameters_of_age mean std var;
var average_of_age;
run;
```

Result Window:

Histogram of the sample means



The mean and Standard deviation of the sample means

Mean and Standard deviation of the sample means

The MEANS Procedure

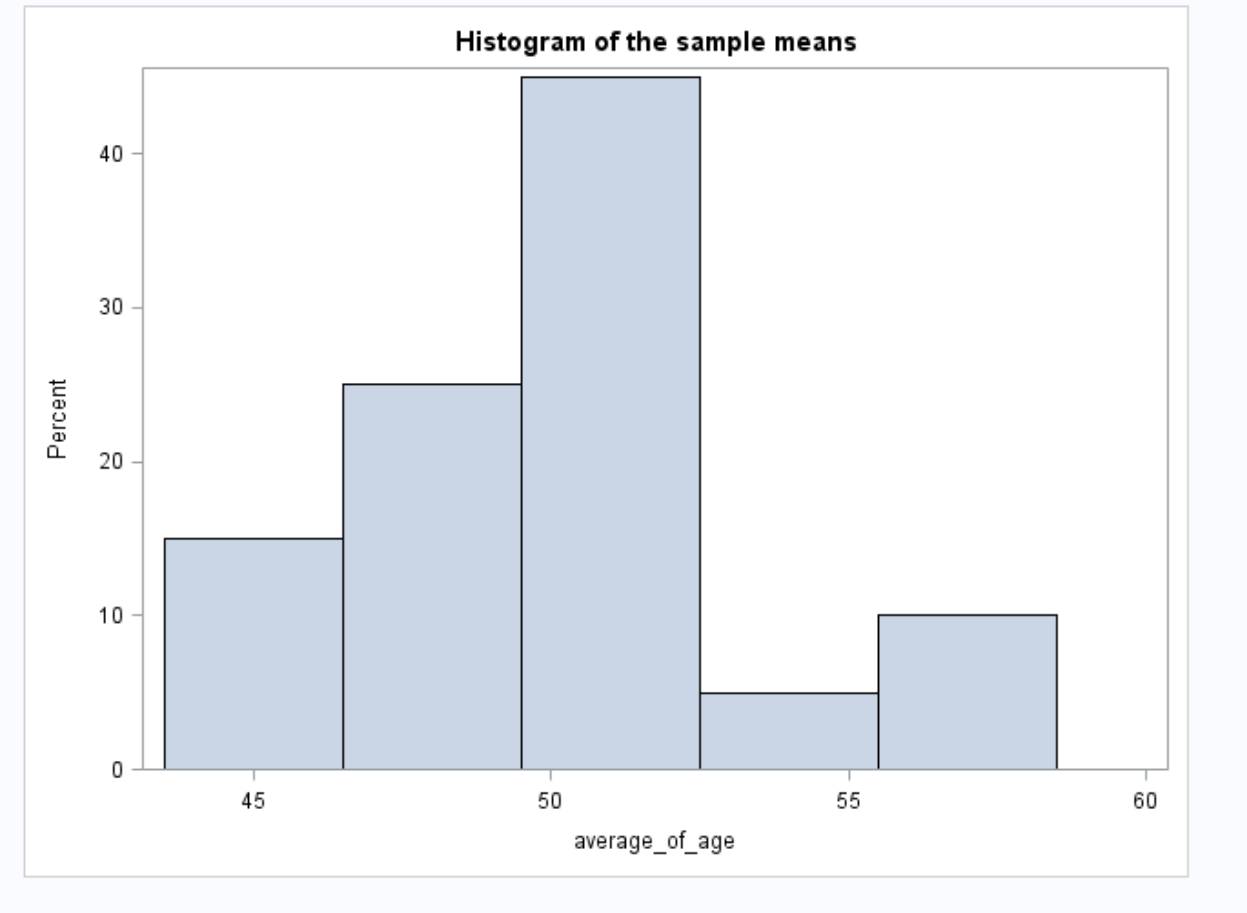
Analysis Variable : average_of_age		
Mean	Std Dev	Variance
52.3400000	8.0923225	65.4856842

d) The above steps are repeated for a sample size of 45.

```
*/ We use proc surveyselect option to generate random samples.
Here we select the method of sampling to be simple random
sampling. It generates random samples without replacement. The
sample size is chosen to be 45 and this
process is repicated 20 times ;

proc surveyselect data = age_dataset
method = SRS rep = 20
sampsize = 45 out = Random_samples_of_age_45;
id age;
run;

* The random samples of age are stored in the
Random_samples_of_age dataset. The replicate coloumn has a range
from
1 to 20 since we are taking 20 samples. The size of each
replicate ID is 45. Therefore we calculate the mean, standard
deviation and variance of the samples with the same replicate
number. This can be done using the by statment.
We store these values in a new dataset called
Parameters_of_age.;
proc means data = Random_samples_of_age_45 n mean std var;
var age;
output out = Parameters_of_age_45 mean=average_of_age std=
standard_deviation_of_age var=variance_of_age;
by replicate;
run;

* To plot the histogram of sample means;
title"Histogram of the sample means";
proc sgplot data=Parameters_of_age_45;
histogram average_of_age;
run;

* To calculate the mean and standard deviation of the sample
means;
title" Mean and Standard deviation of the sample means";
proc means data = Parameters_of_age_45 mean std var;
var average_of_age;
run;
```

Result window:

Parameters_of_age_45 dataset containing the Histogram of the sample means:



The mean and Standard deviation of the sample means :

Mean and Standard deviation of the sample means

The MEANS Procedure

Analysis Variable : average_of_age		
Mean	Std Dev	Variance
50.0244444	3.3663233	11.3321326