**Questions: Part I** (Simulation and Macros in SAS)

1. For each of the following five distributions, generate 1000 random samples:

(a) Normal distribution N(90,8),

(b) Exponential with parameter 8,

(c) Binomial with n=200 and p=0.2,

(d) Poisson distribution with parameter 8, and

(e) Gamma distribution (3,8).

Use PROC SGPLOT (pages 230-231 of textbook) to plot a histogram for each generated random sample. Print

the five histograms you got (clearly title them), but do NOT print the datasets. You are required to use

a Macro to simplify your coding | in other words, instead of writing five separate codes (one for each Distribution), write one Macro and call it five times with appropriate parameters.

**ANS: - Macro used to generate 1000 random samples using various distributions**

**%macro** Dist(type=,mean=,stddev=,lambda=,prob=,freq=,occur=,alpha=,beta=);

DATA HW2.sample;

\*//(a)-> for Normal distribution//;

%if &type=NORMAL %then %do;

Do i = **1** to **10000**;

x=rand("&type",&mean,&stddev);

output;

end;

%end;

\*//(b)-> for Exponential distribution//;

%if &type=EXPONENTIAL %then %do;

Do i = **1** to **1000**;

x=rand("&type")/&lambda;

output;

end;

%end;

\*//(c)-> for Binomial distribution//;

%if &type=BINOMIAL %then %do;

Do i = **1** to **1000**;

x=rand("&type",&prob,&freq);

output;

end;

%end;

\*//(d)-> for Poisson distribution//;

%if &type=POISSON %then %do;

Do i = **1** to **1000**;

x=rand("&type",&occur);

output;

end;

%end;

\*//(e)-> for Gamma distribution//;

%if &type=GAMMA %then %do;

Do i = **1** to **1000**;

x=rand("&type",&alpha,&beta);

output;

end;

%end;

PROC SGPLOT DATA=HW2.sample;

HISTOGRAM x;

TITLE "&type DISTRIBUTION";

RUN;

**%mend** Dist;

**run**;

%***Dist***(type=NORMAL,mean=**90**,stddev=**8**);

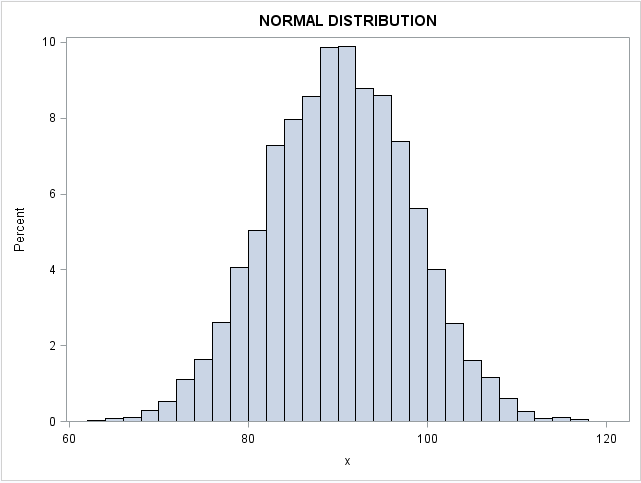
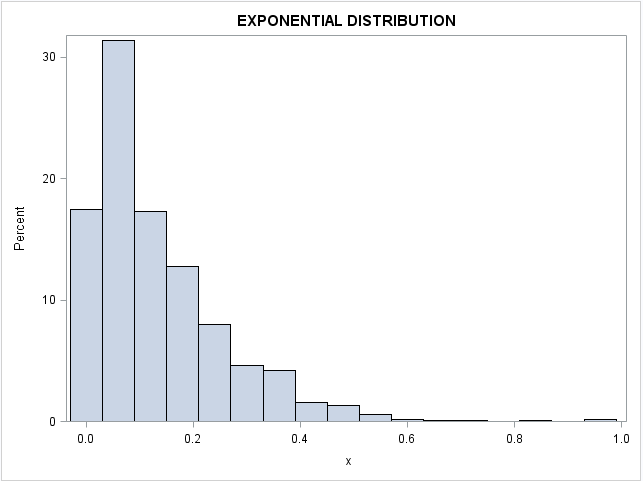
%***Dist***(type=EXPONENTIAL,lambda=**8**);

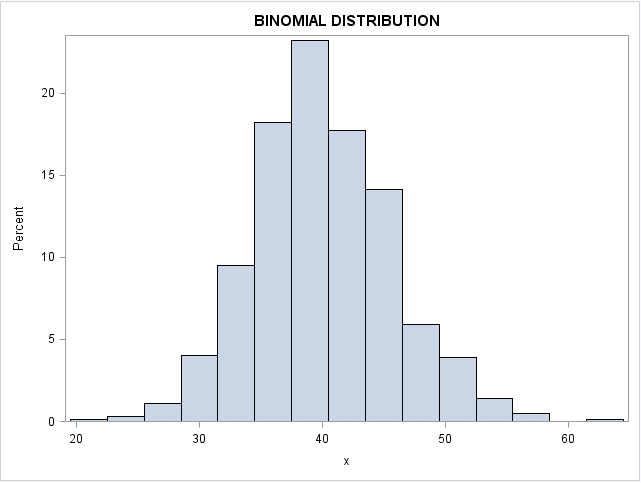
%***Dist***(type= BINOMIAL, prob=**0.2**,freq=**200**);

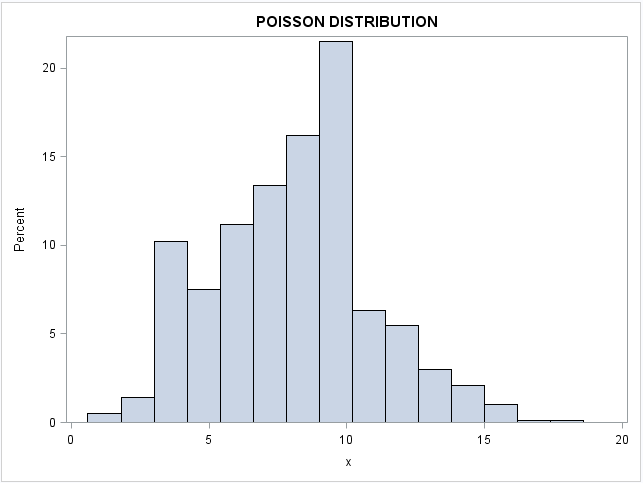
%***Dist***(type=POISSON,occur=**8**);

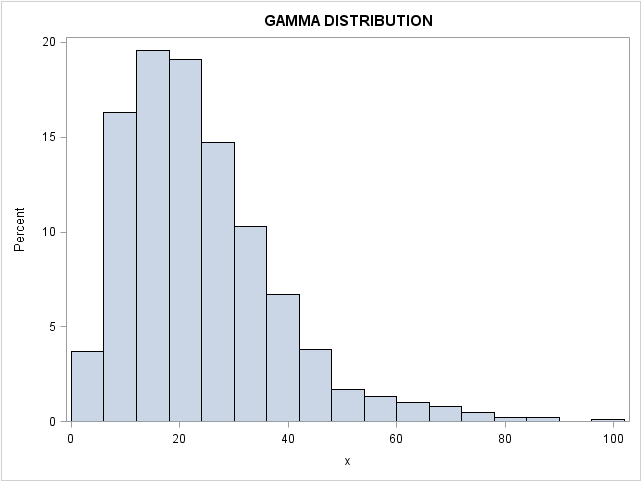
%***Dist***(type=GAMMA,alpha=**3**,beta=**8**);

**run**;









Alternate Solutions for Part I, Question 1 ***(Without using Macro)***

* For Normal Distribution with mean=90 and Standard deviation=8

**data** random1;

\*call streaminit(123);

do i=**1** to **1000**;

x=rand('normal',**90**,**8**);

output;

end;

**proc** **sgplot** data=random1;

histogram x;

**run**;

2. For exponential Distribution with lambda=8

**data** random1;

\*call streaminit(123);

do i=**1** to **1000**;

x=rand('exponential')/**8**;

output;

end;

**run**;

**proc** **sgplot** data=random1;

histogram x;

**run**;

3. For binomial distribution with success probability=0.2

**data** random1;

\*call streaminit(123);

do i=**1** to **1000**;

x=rand('binomial',**0.2**,**200**);

output;

end;

**run**;

**proc** **sgplot** data=random1;

histogram x;

**run**;

4. For Poisson’s distribution with occurance=8

**data** random1;

\*call streaminit(123);

do i=**1** to **1000**;

x=rand('poisson',**8**);

output;

end;

**run**;

**proc** **sgplot** data=random1;

histogram x;

**run**;

5. For Gamma distribution with alpha=3 and beta=8

**data** random1;

\*call streaminit(123);

do i=**1** to **1000**;

x=rand('gamma',**3**,**8**);

output;

end;

**run**;

**proc** **sgplot** data=random1;

histogram x;

**run**;

2. From mathematics, we know that a gamma distribution (1,5), where the first parameter is 1, is equivalent to

an exponential distribution with parameter 1/5. Can you use data to show that indeed they generate the same

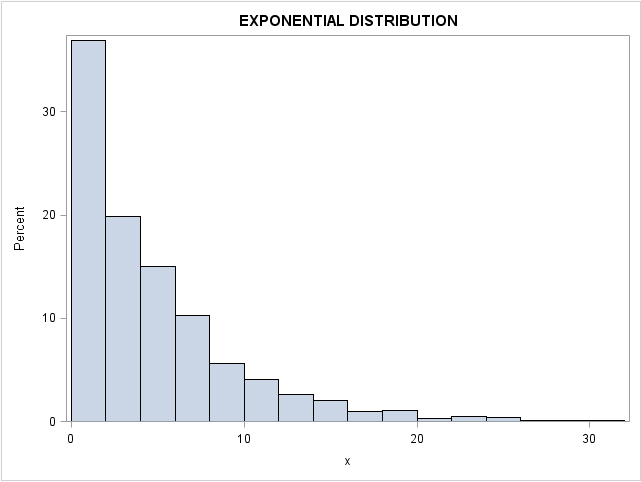
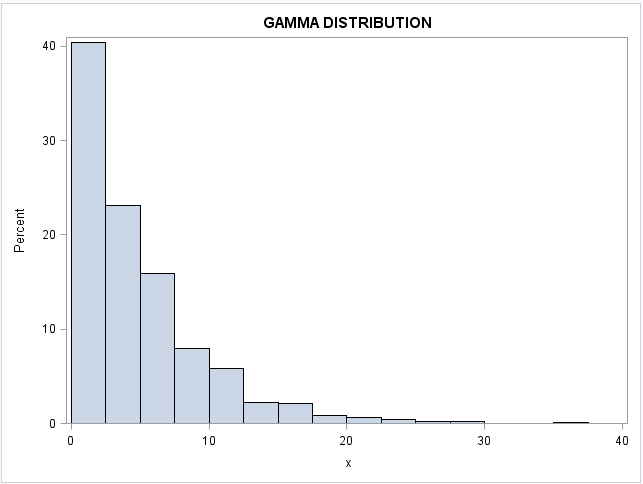
(or at least very similar) distribution of random numbers? Show your code and your finding. (Hint: re-use

your code from question 5, and see if you can show the histograms are similar.)

**ANS** ->

%***Dist***(type=EXPONENTIAL,lambda=**.2**);

%***Dist***(type=GAMMA,alpha=**1**,beta=**5**);



From the above graphs, both Exponential distributions (with lamba=8) and GAMMA distribution

(With alpha=1 & beta =5) generate similar distribution of random numbers and hence look similar.

Alternate code for Part I, Q2 (Without using macro):

Comparison of Exponential and Gamma function:

**data** random1;

call streaminit(**123**);

do i=**1** to **1000**;

x=rand('GAMMA', **1**,**5**);

output;

end;

**run**;

**proc** **sgplot** data=random1;

histogram x /binwidth=**1** showbins scale=percent;

title "gama";

**run**;

**data** random2;

call streaminit(**123**);

do i=**1** to **1000**;

x=rand('exponential')/**0.2**;

output;

end;

**run**;

**proc** **sgplot** data=random2;

histogram x /binwidth=**1** showbins scale=percent;

title "expo";

**run**;

**Questions: Part II** (Customer Retention)

* Using the survival dataset as the input, write a SAS program to create a new dataset called sALTERED (this is identical to the right one on slide 20 of the customer retention lecture).

**ANS** : -

/\*HW 2 PART II - After Corrections \*/

LIBNAME MIS6334 'F:\MIS6334’;

**DATA** MIS6334.sALTERED (drop = PrevCust);

SET MIS6334.Survival END = LastYear ;

RETAIN PrevCust;

Lost = PrevCust - Customers;

PrevCust = Customers ;

IF Year = **0** THEN DELETE ;

IF LastYear THEN

DO;

OUTPUT;

Year = Year+**1**;

Customers = **0**;

Lost = PrevCust - Customers;

END;

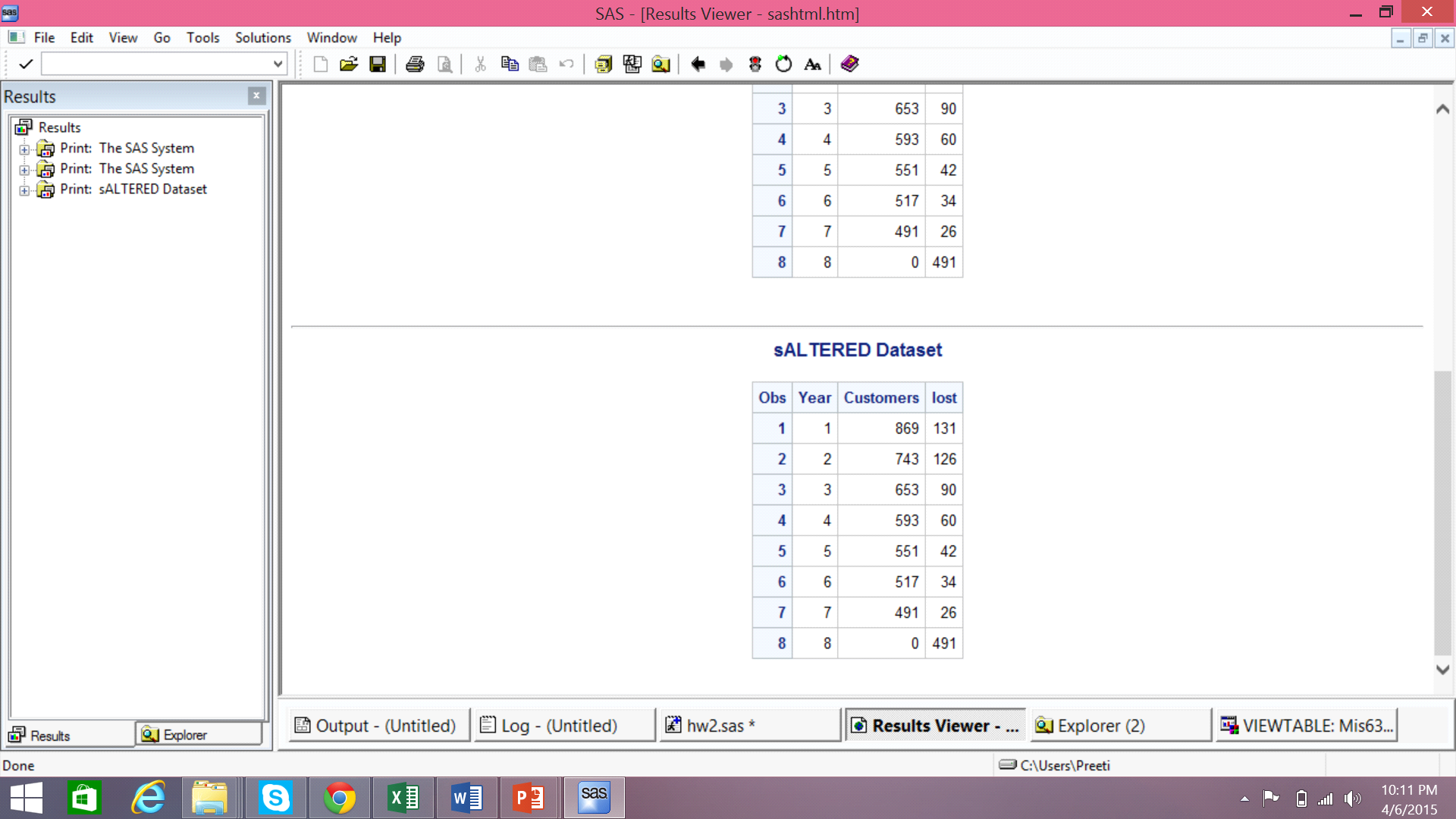
OUTPUT;

**RUN**;

**PROC** **PRINT** DATA = MIS6334.sALTERED;

**RUN**;

Output:- sALTERED Dataset Screenshot:-



* In class we analyzed the survival model using MLE by assuming that all customers have the same θ. However the fit turned out to be poor. In this question, use PROC NLMIXED to conduct maximum likelihood estimation (MLE) using the shifted Beta-Geometric (sBG) model, where each customer’s θ is a random draw from a Beta distribution. Print the “Fit Statistics” table and the “Parameter Estimates” table.

Hints:

• Start from the code on slide 20; and modify it to accomplish this.

• Use the formulas on slide 28, and the BETA function in SAS to simplify your job.

• Compare your results with the ones on slide 30. If they are different, something is wrong with your code.

ANS :-

/\* MLE analysis with theta produced by random Beta distribution \*/

**PROC** **NLMIXED** DATA=mis6334.sALTERED;

PARMS alpha =**1**, beta =**1**;

/\* initial values of Alpha and Beta Parameters are assumed 1 \*/

retain prevprob **0**;

retain sumprob **0**;

IF year = **1** THEN prob = alpha/(alpha+beta);

ELSE prob = prevprob\*(beta+year-**2**)/(alpha+beta+year-**1**);

prevprob = prob;

IF customers > **0** THEN

DO;

ll = lost\*log(prob);

sumprob = prevprob + sumprob;

END;

ELSE IF customers = **0** THEN ll = lost\*log((**1**-sumprob));

MODEL lost ~ general(ll);

**RUN**;

Proposed changes –

**PROC** **NLMIXED** DATA = MIS6334.sALTERED;

PARMS Alpha = **1**, Beta = **1**;

/\* initial values of Alpha and Beta Parameters are assumed 1 \*/

IF Customers > **0** THEN

DO;

Prob = Beta(Alpha+**1**,Beta+Year-**1**)/ Beta(aplha,Beta);

ll = Lost \* log(Prob);

END;

ELSE IF Customers = **0** THEN

DO;

Prob = Beta(Alpha,Beta+Year)/ Beta(Alpha,Beta);

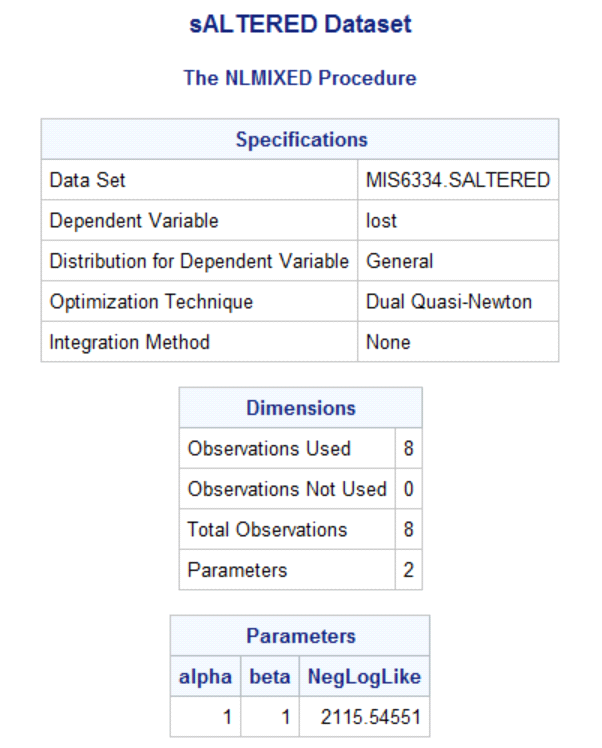
ll = Lost \* log(Prob);

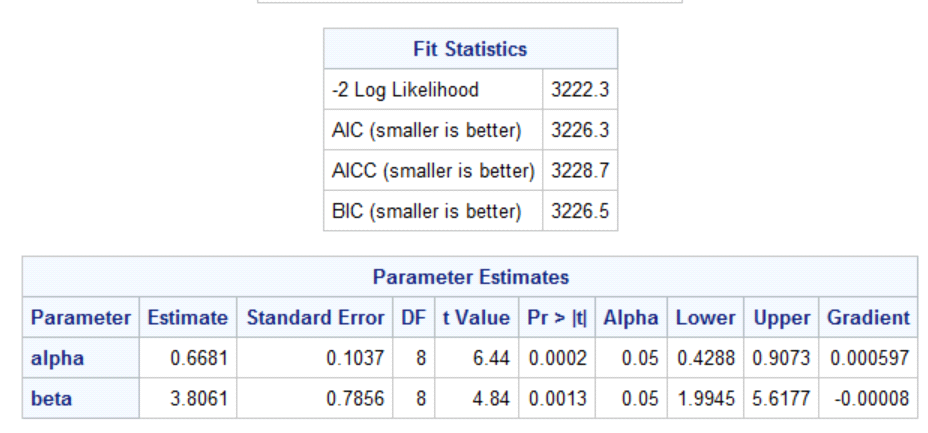
END;

MODEL Lost ~ GENERAL(ll);

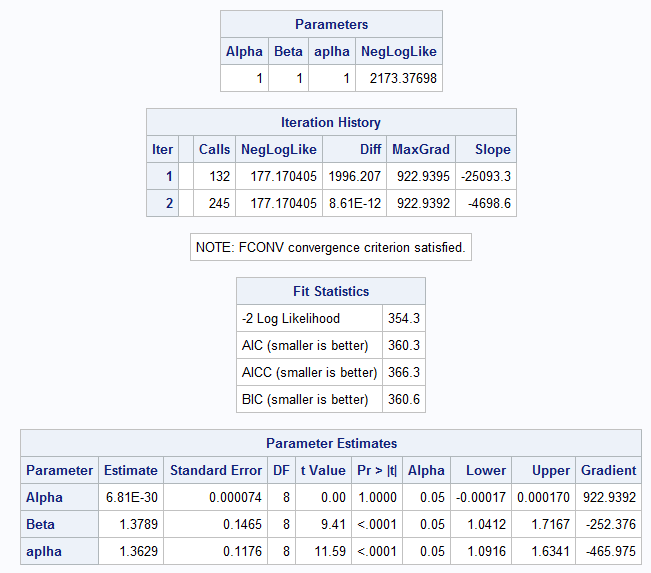
**RUN**;

SAS OUTPUT: -





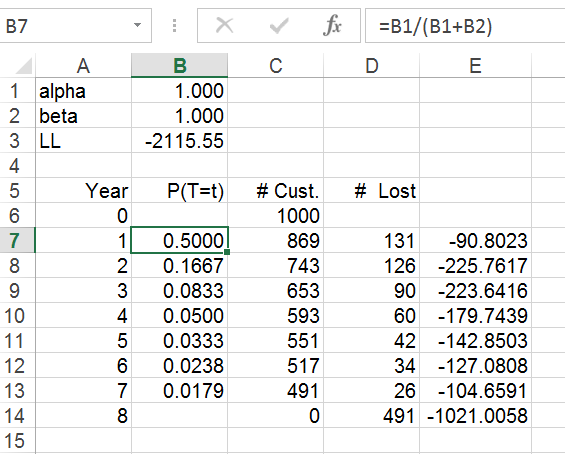
New Output



EXPLAINATION:-

Verified the output: Negative log likelihood calculated in SAS and Excel for this model are same i.e. 2115.54 when parameters of Beta distribution are assumed to be alpha = 1 and beta = 1.

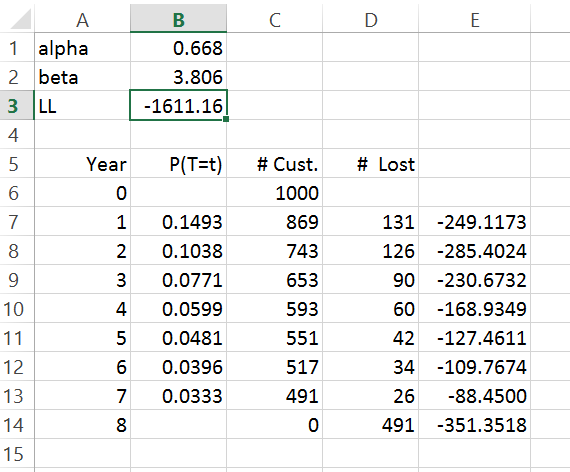
MLE Analysis in EXCEL:



Verified MLE output:

According to Fit Statistics in SAS, maximum log likelihood = 3222.3 / -2 = -1611.15 and estimated alpha = 0.6681 and estimated beta = 3.8061.

These values are verified using solver in Excel (as given below) and slide 34 has the same results for MLE.



3. Suppose you found out that the optimal parameters are α = 0.668 and β = 3.806. Using these parameters, write a SAS program for simulating the individual-level survival behavior of 1000 customers over a 12-year span. (Use the code on slide 22, and modify it to ﬁt the sBG model.)

**ANS**:- From the maximum likelihood estimation, we have obtained the values α = 0.668 and β = 3.806

This corresponds to a theta of 0.149 i.e. churn probability.

theta=alpha/(alpha+beta) = 0.668/(0.668+3.806)

Now, using this value of probability in the Bernoulli’s random distribution, we get the following result:

**DATA** simulated;

CALL streaminit(**123**);

DO i=**1** to **1000**; /\* simulate 1000 customers \*/

DO t=**0** to **11**; /\* t represent period \*/

x=rand('BERNOULLI',**0.149**);/\* use the theta obtained from MLE \*/

IF x=**1** THEN leave;

OUTPUT;

END;

OUTPUT;

END;

**RUN**;

Proposed changes –

**DATA** MIS6334.Simulated;

CALL streaminit(**123**);

DO i = **1** to **1000**; /\* Simulating 1000 Customers \*/

Theta = Beta(**1.3629**,**1.3789**); /\* Alpha and Beta values are from MLE\*/

DO t = **0** to **11**; /\* t represents Year. \*/

x = RAND('BERNOULLI',Theta);

IF x = **1** THEN LEAVE;

OUTPUT;

END;

OUTPUT;

END;

**RUN**;

The following code, results in a table that contains number of customers **retained** at each period (year)

**PROC** **SQL**; /\* count how many Customers LEAVE in each period \*/

CREATE TABLE MIS6334.Sumtable AS

SELECT t AS Year,

Count(i) AS CUSTOMER

FROM MIS6334.Simulated

GROUP BY t;

**QUIT**;

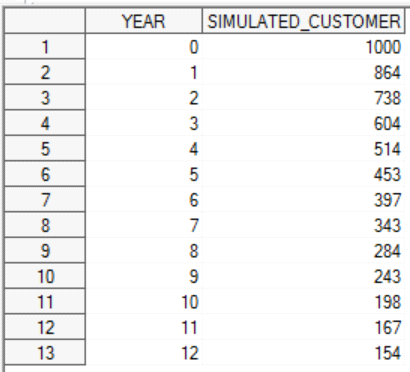


Fig. Model Simulated Dataset

The following code visualizes the behavior of randomly generated (1000) customers with the original data:

**DATA** MIS6334.MergedData;

MERGE MIS6334.Survival MIS6334.Sumtable;

BY Year;

**RUN**;

**PROC** **SGPLOT** DATA = MIS6334.MergedData;

SERIES x = Year y = Customer;

SERIES x = Year y = Customers;

**RUN**;

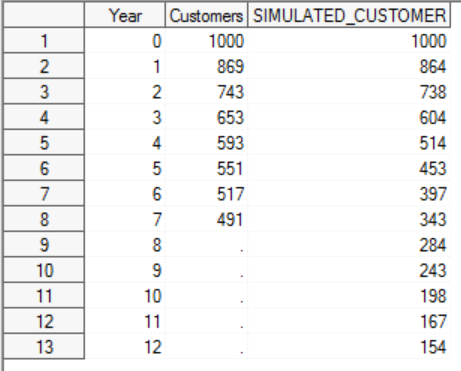
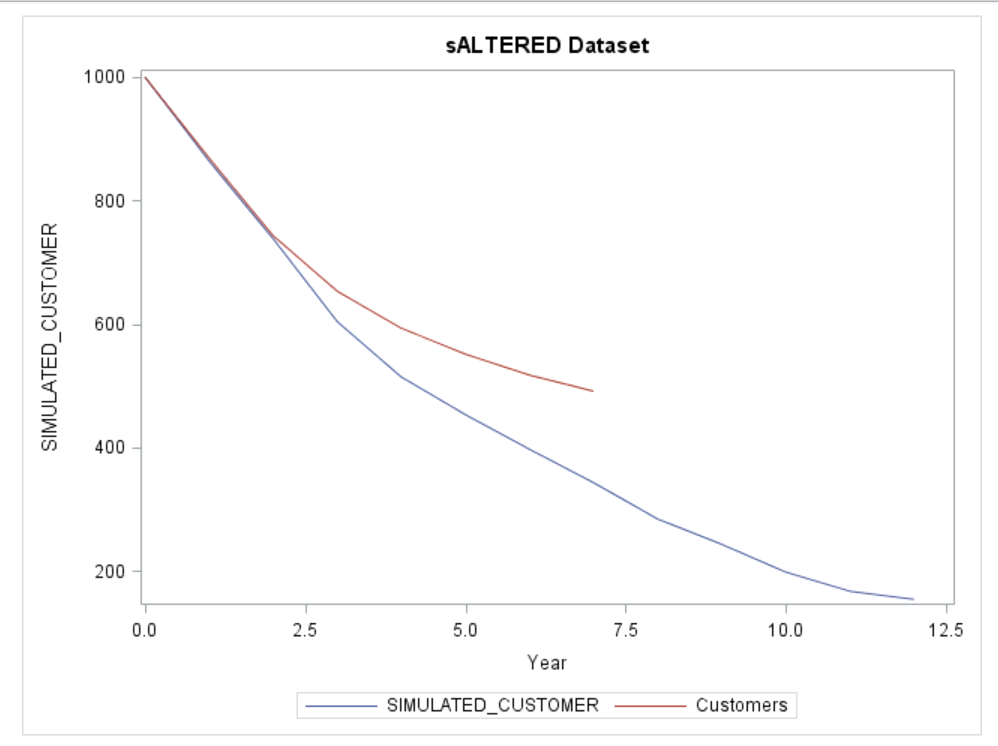
 

Fig : Merged Dataset – Original and Simulated Customers Fig : Plot of Original data and Modeled data

Note: We also tried using *rand(‘BERNOULI’, rand(‘BETA’,0.668,3.086))* function to simulate 1000 customers, but the fit was observed to be poor and hence we found the aforementioned solution to be the best.

New Outputs –

