

A Beginner's Guide to Simulation Studies

Barbara Foster, Ph.D.
Department of Academic Computing Services
UT Southwestern Medical Center
Dallas, TX

Why Do a Simulation Study?

To determine patterns or results from hypothetical situations which may be applied to real situation.

Overview

- Defining the question
- Design the study
- Program the design
- Run the program
 - to generate data
 - to obtain and accumulate statistic(s) of interest
 - to analyze accumulated information
- Draw Conclusions

Defining the Question

The Central Limit Theorem in the field of statistics states that regardless of the underlying pattern of the data, that

- if a number of samples of size n are randomly selected from a specified population, and
- if the average is calculated for each of the samples

then the sample averages will have a bell shaped distribution as the sample size gets larger.

Defining the Question

How large does the sample size have to be for the bell shaped curve to occur?

Designing the Study

- What is the statistic of interest?
the average (also known as the mean)

Example of data from a random sample
of size 5: 2, 4, 3, 5, 4

$$\bar{X} = \frac{\sum X}{n} = \frac{2+4+3+5+4}{5} = \frac{18}{5} = 3.6$$

Designing the Study

- What sample sizes are of interest?

5, 10, 25, 50, 100

Designing the Study

- What shapes (also known as distributions) will be sampled from?

Distributions

Normal (bell shaped)

Uniform

Exponential

Gamma

Reason for Selection

Well-studied/Ideal

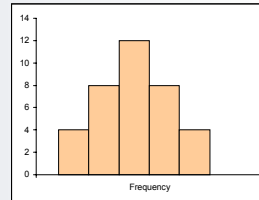
Constant frequency

Extreme values

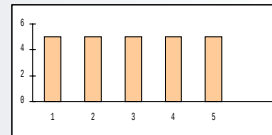
Non-symmetric

Designing the Study

- Normal Distribution

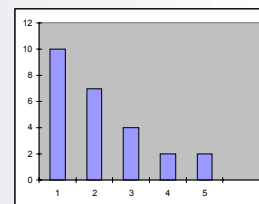


- Uniform Distribution

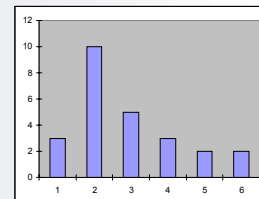


Designing the Study

- Exponential Distribution



- Gamma Distribution



Program the Design

- Start with the most basic unit
- When the basic unit is working correctly then add the repeats

Program the Design

- *Start with the simplest unit:*

```
DATA NUMBERS;  
  SXSZ=5;  
  DO A=1 TO SXSZ;  
    X=RANNOR(0);  OUTPUT;  
  END;  
  
PROC MEANS MEAN NOPRINT; VAR X; OUTPUT OUT=MEANS  
  MEAN=AVG /AUTOLABEL; RUN;  
  
DATA M;          SET MEANS;  
  SXSZ=20;  
  KEEP SXSZ MEAN;
```

Program the Design

■ Consolidate results of several runs:

```
LIBNAME CLT 'C:\PRESENTATIONS\SCSUG';                                RUN;
%macro CLT;
  %DO REPEATS=1 %TO 5;

  DATA NUMBERS;
    SXSZ=5; DO A=1 TO SXSZ;X=RANNOR(0); OUTPUT;      END;

  PROC MEANS MEAN NOPRINT; VAR X; OUTPUT OUT=MEANS MEAN=AVG /AUTOLABEL; RUN;

  DATA M;      SET MEANS;      SXSZ=20;      KEEP SXSZ AVG;

  PROC APPEND BASE=CLT.AVG_DATA;

  %END;
%MEND;

%CLT; RUN;

PROC PRINT DATA=CLT.AVG_DATA;                                RUN;
```

Program the Design

■ Add the graphing portion

```
LIBNAME CLT 'C:\PRESENTATIONS\SCSUG';                                RUN;
%macro CLT;
  %DO REPEATS=1 %TO 5;

  DATA NUMBERS;
    SXSZ=5; DO A=1 TO SXSZ;X=RANNOR(0);      OUTPUT;      END;

  PROC MEANS MEAN NOPRINT; VAR X; OUTPUT OUT=MEANS MEAN=AVG /AUTOLABEL; RUN;

  DATA M;      SET MEANS;      SXSZ=20;      KEEP SXSZ AVG;

  PROC APPEND BASE=CLT.AVG_DATA;
  %END;

  PROC chart DATA=CLT.AVG_DATA;
    VBAR AVG; RUN;

  %MEND;

%CLT; RUN;
```

Program the Design

- **The basic unit is completed.
Now add the multiple desired
sample sizes.**

Program the Design

- **Now add the desired
distributions.**

Run the Program

- **The time it takes to run the program is dependent upon**
 - The size of the samples
 - The number of repetitions (samples)
 - The number of distributions selected
 - The requested statistics and/or graphs

Draw Conclusions

- From which distributions do the sample means form a normal distribution with small sample sizes?
- From which distributions do the sample means require a large sample size to form a normal distribution?

Recommended Reading:

- SAS for Monte Carlo Studies: A Guide for Quantitative Researchers (2002)

Xitao Fan

Akas Felsovalyi

Stephen A. Sivo

Sean C. Keenan