Statistical Methods and Computing, STAT:2010:4200
Chapter 11

Part I: Online Demo of the sampling distribution

1 Demonstration of the sampling distribution of $\overline{x}$

Go to

`http://onlinestatbook.com/stat_sim/sampling_dist/`

Then click "Begin" on the upper left corner.

1.1 What are the plots on the screen

- The distribution portrayed at the top of the screen is the population from which samples are taken. The mean of the distribution is indicated by a small blue line.
- The second histogram displays the sample data. This histogram is initially blank.
- The third plot is used to display statistics calculated from samples from the population. (Ignore the fourth plot for now.)

1.2 Draw a random sample of size 5 each time. Repeat this 10 times.

Note the buttons to the right of the third plot. "**N=5**" means that we want to draw a sample of size 5; "**Mean**" means that we are going to calculate the sample mean, that is, the average of the N=5 sample points.

Now click on the button "**Animated**" once. You will see that a random sample of size 5 is drawn from the population. The blue bar indicates the sample mean. Let us denote the sample mean by $\overline{x}^{(1)}$, where the superscript (1) indicates that this statistic is calculated from the first sample that we draw.

Now click on the button "**Animated**" a second time. Another random sample of size 5 is drawn from the population. A new blue bar indicates the sample mean of this new sample. We denote this sample mean by $\overline{x}^{(2)}$. Its value is added to the histogram displayed in the third plot.

Click on the button "**Animated**" a third, a fourth, ... and a tenth time. Make sure you wait long enough between clicks in order for the applet to complete the simulation each time.

QUESTION: How does the center and the spread of the histogram of $\overline{x}^{(1)}, \overline{x}^{(2)}, ..., \overline{x}^{(10)}$ compare to that of the population? That is, compare plot 3 to plot 1.

1.3 Draw a lot more random samples, each of size 5.

Suppose we want to draw 50 random samples, each of size 5. Instead of clicking on the "**Animated**" button 40 more times, we click on the "**5**" button 8 more times.

Suppose we want a lot more random samples, each of size 5. Click on the "**1000**" button once. (How many samples have we drawn so far?)

QUESTION: How does the center and the spread of the histogram of the sample means compare to that of the population? That is, compare plot 3 to plot 1.

QUESTION: What is the shape of the histogram of the sample means?

Select the box "**Fit to normal**" and see for yourself.

1.4 Changing the population distribution

Note the tab to the right of plot 1. Change "**Normal**" to "**Skewed**".
Repeat the steps in 1.2-1.3.

QUESTION: Is the histogram of the sample means (plot 3) look like a symmetric bell?

Repeat the above for $N = 2$ and then for $N = 25$. Pay attention to the plot 3.

1.5 Summary and implication of the demonstration

To study the a certain population, suppose you are going to draw a random sample of size N from it and calculate its sample mean $\overline{x}$. Note that $\overline{x}$ is a random thing, that is, if you draw another random sample of size N, you'll probably get a different $\overline{x}$. Hence we can talk about the distribution of $\overline{x}$, which we call the **sampling distribution** of $\overline{x}$.

The histogram in plot 3 shows approximately what the **sampling distribution** of $\overline{x}$ looks like. We observed that:

- If the samples (of size N) are drawn from a population that has normal distribution. Then for any N, the sampling distribution has a normal distribution as well.
- If the samples (of size N) are drawn from a population that has a skewed distribution. Then the sampling distribution is skewed when N is small (N=2, 5). The sampling distribution looks much more symmetric and has an approximately normal distribution when N is large (for eg, when N=25).

Part II: SAS

2 Using `proc tabulate` to summarize the distributions of quantitative variables in different groups

2.1 Gulanick data

Gulanick (*Heart and Lung*, 1991) studied patients who were recovering from heart surgery. She was interested in whether different combinations of supervised exercise or teaching would affect patients' self-efficacy (or confidence) to perform physical activity.

Patients were randomly assigned to one of three groups. Group 1 received teaching, treadmill exercise testing, and exercise training three times per week. Group 2 received only teaching and exercise testing. Group 3 received only routine care without supervised exercise or teaching. After 4 weeks, each patient was scored on self-efficacy.

Self-efficacy was measured on a continuous scale and scores were assumed to be distributed normally in each of the populations of interest. Her results are in the dataset "gulanick.dat." We wish to produce a table that shows the number of observations and the mean and standard deviation of scores within each of the three groups.

```
proc format ;
value grpfmt 1 = 'Teaching and Training' 2 = 'Teaching' 3 = 'Neither' ;
run ;

data gulan;
* here we assume that gulanick.dat is downloaded on the desktop on the local disk;
infile '\\Client\C$\Users\ShengWang\Desktop\gulanick.dat';
```

```
input score group;
format group grpfmt.;
run;

proc tabulate data = gulan ;
class group ;  * class statement identifies qualitative variables ;
var score ;    * var statement identifies quantitative variables ;
tables group , score * (n mean std) ;
run ;
```

	score		
	N	Mean	Std
group			
Teaching and Training	11.00	126.82	24.25
Teaching	12.00	128.42	25.04
Neither	13.00	103.92	17.71

We can use *proc means* to get various summary statistics in a more compact format than *proc univariate* provides. The default statistics provided are

- n = number of observations
- mean
- std dev = standard deviation
- minimum
- maximum

```
proc means data = gulan ;
var score ;
where group eq 1 ;   * restricts to those records with group = 1 ;
run ;

The MEANS Procedure

Analysis Variable : score

N            Mean         Std Dev         Minimum         Maximum
-----------------------------------------------------------------
11     126.8181818      24.2520852     100.0000000     170.0000000
```

2.2 billion data

Read in the billion data set in the desired format.

```
proc format ;
value $regfmt 'A' = 'Asia' 'E' = 'Europe' 'M' = 'Middle East'
'O' = 'Other' 'U' = 'US' ;
value amtfmt low-<5 = '<5' 5-<10 = '5-<10' 10-<20 = '10-<20' 20-high = '20+' ;
run ;

filename url "http://homepage.divms.uiowa.edu/~kcowles/Datasets/billion.dat";
data billion ;
infile bill;
input wealth age region $ ;
format region $regfmt. wealth amtfmt. ;
label wealth = 'Wealth in Billion $'
        age = 'Age in Years' ;
run ;
```

Proc tabulate for the billionaire data.

```
proc tabulate data = billion ;
class region ;
var age ;
tables region, age * (mean stddev) ;
run ;
```

	Age in Years	
	Mean	StdDev
region		
Asia	63.65	9.96
Europe	64.03	14.70
Middle East	64.23	19.53
Other	64.14	13.08
US	64.15	11.88