University of California, Los Angeles
Department of Statistics

Statistics M11 **Instructor: Nicolas Christou**
Introduction to Statistical Methods for Business and Economics **Winter 2001**

The distribution of the sample mean and the Central Limit Theorem

In this lab you will understand (hopefully) through a simulation how the sample mean $\bar{x}$ is distributed. We discussed in class earlier that if a sample of size n is taken from a population that follows the normal distribution with mean μ and standard deviation σ then regardless of how large the sample is, the sample mean $\bar{x}$ also follows the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

If a large sample (usually n>=30) is taken from a nonnormal distribution with mean μ and standard deviation σ then according to the central limit theorem the sample mean $\bar{x}$ follows the normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$. This is a very important theorem because knowing the distribution of $\bar{x}$ we can make inferences about the population's mean, even if this population does not follow the normal distribution.

We are going to see through a simulation study how the above statement works…

You must remember from earlier discussion that when a die is rolled and X represents the number that occurs then the probability distribution of X is given as follows:

X	*P(X)*
1	*1/6*
2	*1/6*
3	*1/6*
4	*1/6*
5	*1/6*
6	*1/6*

And you must know by now that the mean of X is μ=3.5, and the standard deviation of X is σ=1.71. Make sure you know how this two numbers are computed.

You are going to select many samples (obviously with replacement) of size n=80 from this distribution. In other words, each time you are going to roll the die 80 times (of course the computer will do this) and you will compute the sample mean of these 80 numbers. At the end you will have a list of many sample means.

You have already used the program `dice2`. ***This program will roll the die for you. As a reminder if you type*** `. dice2` ***you will receive the following information:***

```
. dice2
Here is how to use dice2
dice2 rolls [numdice numside, save]
  rolls  = number of rolls of the dice
  numdice= the number of dice rolled, default=2
  numside= the number of sides on the dice, default=6
  The save option saves the resulting data, and
    clears out the data currently in memory.
```

We want to roll one die that has 6 sides (1-6) 80 times. This is what you type:

`. dice2 80 1 6, save` ***80 is the number of rolls, 1 is the number of dice, 6 is the number of sides.***

The program will save the 80 numbers (the result of these 80 rolls) and also will construct a histogram of these 80 values (ignore the histogram). You can also see the 80 numbers by typing `. list` ***or by typing*** `. edit` **.**

As it was mentioned earlier we are going to generate many samples of size n=80. Let's say that we want to generate 500 samples. Each sample has 80 observations. Therefore we must roll the die total of 80x500=40000 times. Let's do this…

```
. dice2 40000 1 6, save
```

If you type `. list` ***or*** `. edit` ***you can see these 40000 numbers. Now we must group these 40000 numbers into groups of 80 each (this is the sample size of each sample) in order to generate 500 samples. This is how we do this…***

```
. generate dummy=group(500)
```

This command will generate a variable called "dummy" next to the variable "sumdice" that has the following values:
1's from 1-80
2's from 81-160
……
……
500's from 39921-40000.

Type `. list` ***or*** `. edit` ***to see the original outcomes of the 40000 rolls and the values of the "dummy" variable. Therefore the values of "sumdice" that correspond to the values of "dummy" from 1-80 will be your first sample, the values of "sumdice" that correspond to the values of "dummy" from 81-160 will be your second sample etc.***

The next step is to compute the mean of each sample. First save your data as it is!

Use the following command to compute the sample mean of each sample:

```
. collapse sumdice, by(dummy)
```

This command will "destroy" your dataset (but it is ok because you have saved it!). It will also compute the sample mean of each sample. Type `. list` ***or*** `. edit` ***to see the 500 sample means. Since we are dealing with sample means let's rename "sumdice" to "xbar" using the following command.***

`. rename sumdice xbar` ***The variable that contains the 500 sample means is now called "xbar."***

You have finished with the sample generation. You have created 500 sample means. Congratulations! Save them as a Stata file (give a name xbar.dta). Before you answer the following questions make sure you understand what we did so far. Go back and read each step carefully…

Questions:

- ***The population from where the samples were taken has μ=3.5 and σ=1.71. According to the central limit theorem what is the distribution of the sample mean (n=80).***
- ***Construct the histogram of these 500 sample means and ask Stata to draw a normal curve on top of that. Here is the command:*** `. graph xbar, normal` ***. You can also change the number of bins. What do you observe?***
- ***Compute the mean of these 500 means. What do you find?***
- ***Compute the standard deviation of these 500 sample means. What do you find?***
- ***Find the probability that the sample mean $\bar{x}$ is larger than 3.6. Remember μ=3.5, σ=1.71, n=80.***
- ***What proportion of these 500 sample means is larger than 3.6 (count how many sample means are larger than 3.6 out of the 500)? It will be helpful if you sort the sample means by using*** `. sort xbar` ***. Compare this empirical probability with the probability that you found in the previous question.***

Save these values (the 500 sample means) because you will need them again later.

Greek alphabet:

Lower case:
α β γ δ ε ζ η θ ι κ λ μ ν ξ ο π ρ σ τ υ φ χ ψ ω

Upper case:
Α Β Γ Δ Ε Ζ Η Θ Ι Κ Λ Μ Ν Ξ Ο Π Ρ Σ Τ Υ Φ Χ Ψ Ω