

SPSS Tutorial 6: The Sampling Distribution of the Mean

Goals

1. To work through an example that demonstrates the following:
 - a. The mean of the sampling distribution of the mean = the mean of the population from which the samples are drawn ($\mu_{\bar{x}} = \mu_x$)
 - b. The SD of the sampling distribution of the mean, also called the standard error of the mean, is equal to the SD of the population from which the samples were drawn divided by the square root of the sample size $\left(\sigma_{\bar{x}} = \sigma_x / \sqrt{n} \right)$
2. To understand what the Central Limit Theorem tells us about the shape of the sampling distribution of the mean

The Central Limit Theorem

Imagine drawing (with replacement) all possible samples of size n from a population, and for each sample, calculating a statistic—e.g., the sample mean. The frequency distribution of those sample means would be the **sampling distribution of the mean** (for samples of size n drawn from that particular population).

The Central Limit Theorem (CLT) is an important theorem in statistics which tells us, among other things that:

1. The mean of the sampling distribution of the mean = the population mean
2. The SD of the sampling distribution of the mean = the standard error (SE) of the mean = the population standard deviation divided by the square root of the sample size

Putting these statements into symbols:

$$\mu_{\bar{x}} = \mu_x \quad \{ \text{mean of the sample means} = \text{the population mean} \}$$

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \quad \{ \text{SE of mean} = \text{population SD over square root of } n \}$$

The goal of this tutorial is to demonstrate, using all possible samples of $n=2$ from a small population of $N=5$, that these equations actually do work out as shown.

The Population

The example we will use in this tutorial is taken from Robert R. Pagano's book *Understanding Statistics in the Behavioral Sciences* (3rd Edition). In Pagano's example, the population consists

of only 5 scores: 2, 3, 4, 5, and 6. Let us begin by reading these scores into SPSS using the DATA LIST method.

```
DATA LIST FREE / X (f2.0) .
BEGIN DATA.
2 3 4 5 6
END DATA.
```

Now use the MEANS procedure to get the mean and SD for these 5 scores.¹

Report

X		
N	Mean	Std. Deviation
5	4.00	1.581

Use COMPUTE statements to create new variables N, MEAN.X, S.X, SIGMA.X, and SE.XBAR.

```
compute mean.x = 4.
compute n = 5.
compute s.x = 1.581138830084.          /* sample SD .
compute sigma.x = s.x * SQRT((n-1)/n). /* population SD .
compute se.xbar = sigma.x/sqrt(2).    /* SE of mean for samples of n=2 .
exe.

var lab
  n          'Population N'
  mean.x     'Population mean'
  s.x        'SD with n-1 (from SPSS) '
  sigma.x    'Population SD (n) '
  se.xbar    'SE of the mean'.
```

```
format n (f3.0) / mean.x s.x sigma.x se.xbar (f8.4).
list all.
```

Regarding the preceding COMPUTE statements:

- S.X is the sample SD (with division by $n-1$).² In order to take along all available decimal places, I copied the value of the SD (1.585538830084) from my Output window and pasted it into the COMPUTE statement for variable S.X.
- SIGMA.X is the population SD, with division by N.

¹ Remember what you learned in Tutorial 5 (bottom of first page) about using the MEANS command when you have only one variable. You cannot use the pull-down menus in this case. Rather, you must type the command into the syntax window directly (e.g., MEANS x.).

² Recall from earlier tutorials that SPSS always computes a sample SD, with division by $n-1$.

- SE.XBAR is the standard error of the mean, or the standard deviation of the distribution of all samples of $n=2$ drawn from this population of 5 scores

Your data file should now look like this:

X	MEAN.X	N	S.X	SIGMA.X	SE.XBAR
2	4.0000	5	1.5811	1.4142	1.0000
3	4.0000	5	1.5811	1.4142	1.0000
4	4.0000	5	1.5811	1.4142	1.0000
5	4.0000	5	1.5811	1.4142	1.0000
6	4.0000	5	1.5811	1.4142	1.0000

All Possible Samples of $n=2$

We now want to draw all possible samples of $n=2$ scores from this population of $N=5$ scores. Before proceeding, I need to make a distinction between two kinds of sampling. One can sample **with replacement**, or **without replacement**.

How to sample with replacement. Randomly select a score from the population of scores, and record it. Throw it back into the population (i.e., replace it). Repeat these two steps $n-1$ more times, where n is the desired sample size. Note that the same member of the population may be sampled more than once when you sample with replacement.

How to sample without replacement. The procedure is as described above, except that sampled members of the population are not returned to the population after being selected. That is, no member can be sampled more than once.

For this demonstration to work, we need to sample **with replacement**. With a sample size of $n=2$, there are 5^2 , or 25 possible samples. The following DATA LIST syntax can be used to read these possible samples into SPSS.

```
* -- Read in all possible samples of n=2 --- .
```

```
DATA LIST LIST / sample X1 X2 (3f2.0).
```

```
BEGIN DATA.
```

```
1      2      2
2      2      3
3      2      4
4      2      5
5      2      6
6      3      2
7      3      3
8      3      4
9      3      5
10     3      6
11     4      2
```

```

12      4      3
13      4      4
14      4      5
15      4      6
16      5      2
17      5      3
18      5      4
19      5      5
20      5      6
21      6      2
22      6      3
23      6      4
24      6      5
25      6      6
END DATA.

```

Now use a COMPUTE statement to create new variable XBAR, the mean of the samples. (See Tutorial 3 if you can't remember how to compute means.) Then use the MEANS procedure to get the mean and SD of the 25 sample means (i.e., the mean and SD of XBAR). The output from MEANS should look like this:

Report

XBAR		
N	Mean	Std. Deviation
25	4.0000	1.02062

As you can see, the mean of the 25 sample means is equal to the mean of the population from which we sampled. That is, $\mu_{\bar{X}} = \mu_X$, as it should be, according to the Central Limit Theorem (CLT).

Earlier, you found that the SD of the population (variable SIGMA.X) was equal to 1.4142, and the standard error of the mean was equal to 1 (which is the square root of 1.4142). But notice that the SD of the sample means shown above is equal to 1.02062, not 1. Why? SPSS has computed the SD of the sample means using division by $n-1$. We need to have the SD of the 25 sample means with division by N , the number of sample means. So, we need to use another COMPUTE statement to obtain this value, as shown below.

```

compute mu.xbar = 4.
compute sd.xbar = 1.02062072616 * sqrt(24/25).
exe.
list.

```

You can now see that the population SD of the 25 sample means is indeed equal to the SD of the population divided by the sample size. That is,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{1.4142}{\sqrt{2}} = 1$$

What the CLT tells us about the shape of the sampling distribution

We have already learned what the CLT tells us about the mean and SD of the sampling distribution of the mean. It also provides us with some very helpful information about the *shape of the sampling distribution of the mean*. Specifically, it tells us the conditions under which the sampling distribution of the mean is *normally distributed*, or at least *approximately normal*, where *approximately* means close enough to treat as normal for practical purposes.

The shape of the sampling distribution depends on two factors: the shape of the *population* from which you sampled, and sample size. The following statements illustrate how these factors interact in determining the shape of the sampling distribution:

1. If the population from which you sampled is itself normally distributed, then the sampling distribution of the mean will be normal, **regardless of sample size**. (Even for sample size = 1, the sampling distribution of the mean will be normal, because it will be an exact copy of the population distribution).
2. If the population distribution is reasonably symmetrical (i.e., not too skewed, reasonably normal looking), then the sampling distribution of the mean will be approximately normal for samples of **30 or greater**.
3. If the population shape is as far from normal as possible, the sampling distribution of the mean will still be approximately normal for sample sizes of **300 or greater**.

So, the general principle is that the more the population shape departs from normal, the greater the sample size must be to ensure that the sampling distribution of the mean is approximately normal.

For many of the distributions that we work with in areas like Psychology and biomedical research, sample sizes of 20 to 30 or greater are often adequate to ensure that the sampling distribution of the mean is approximately normal (but some more conservative textbook authors may recommend 50 or greater). Notice that even for the population of 5 scores we started with, the sampling distribution of the mean for samples of $n=2$ looks very symmetrical (Figure 1).

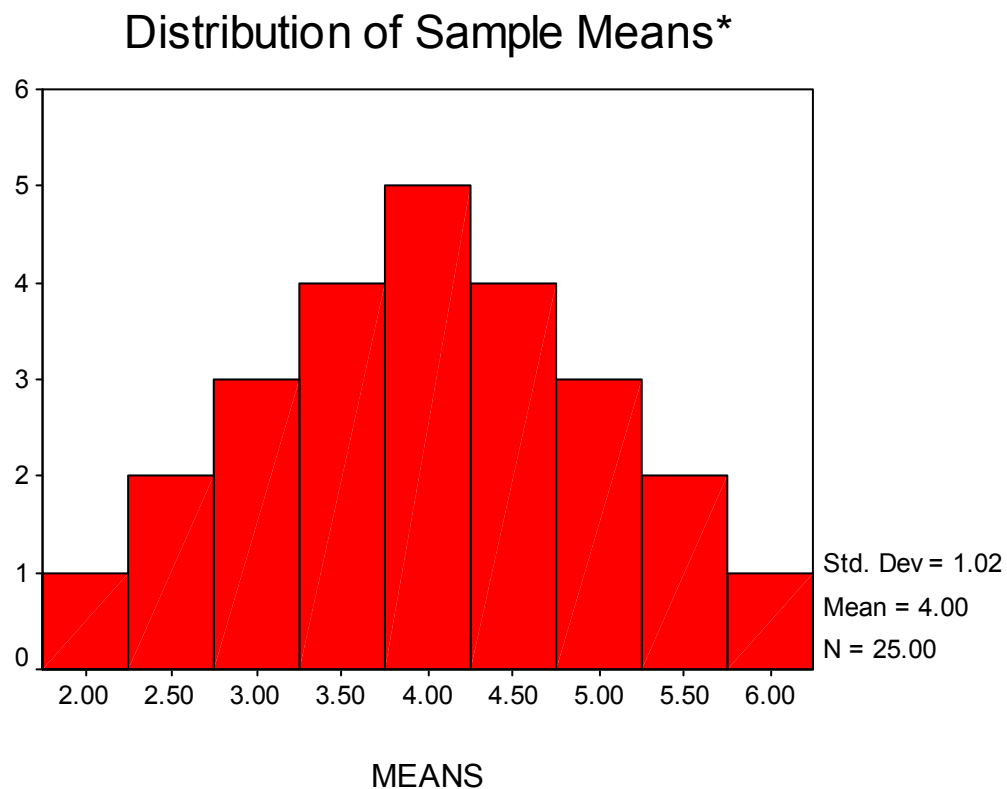
Plotting the distribution of sample means from our example

Use the FREQUENCIES command (Analyze→Descriptive Statistics→Frequencies) with the Histogram option to create a frequency distribution and histogram for variable XBAR. Your frequency table should look like this:

XBAR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2.00	1	4.0	4.0	4.0
	2.50	2	8.0	8.0	12.0
	3.00	3	12.0	12.0	24.0
	3.50	4	16.0	16.0	40.0
	4.00	5	20.0	20.0	60.0
	4.50	4	16.0	16.0	76.0
	5.00	3	12.0	12.0	88.0
	5.50	2	8.0	8.0	96.0
	6.00	1	4.0	4.0	100.0
	Total	25	100.0	100.0	

And your histogram should look something like this:



* or "Sampling Distribution of the Mean"

Notice that even though the population from which we sampled was very small ($N=5$), and not normally distributed (it was in fact a uniform distribution), the distribution of sample means for all samples of $n=2$ looks very nice and symmetrical.

Why is the CLT important?

When you perform a z-test, you use the table of the standard normal distribution (or better yet, a computer program) to find the probability of obtaining a z-value as large or larger than the observed value (i.e., a p-value). The p-value you calculate is only accurate if the sampling distribution of the mean is normal, or at least pretty close to it. The CLT tells you the conditions under which the sampling distribution of the mean is (approximately) normal.

That's all for this Tutorial. Save your syntax file for future reference.
