

# INTRODUCTION TO SAMPLING DISTRIBUTIONS

By Grace Thomson

## INTRODUCTION TO SAMPLING DISTRIBUTIONS

In this chapter we will learn about 3 important topics:

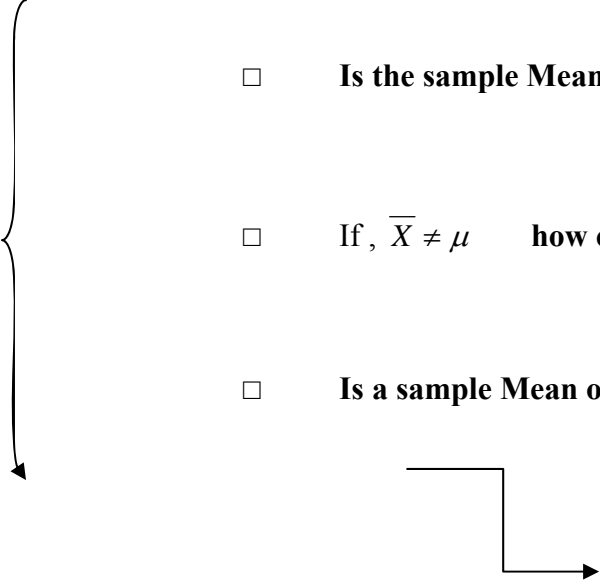
1. Sampling error
2. Sampling Distribution of the mean
3. Sampling Distribution of a proportion

This chapter introduces information about Sampling and its objectives. In Chapter 1 we had studied some techniques for sampling and data collection. Remember when we talked about the systematic random sampling, the stratified sampling, among others? Well, chapter 6 gets into the requirements to ensure that the sample that you have chosen meets quality and validity criteria.

### 1. Sampling error

We have discussed before how effective is to work with a sample instead of a large population, for economic and logistic reasons; but once you have your sample, **new questions arise**:

- Is the sample Mean equal to the population mean**       $\bar{X} = \mu$  ?
- If,  $\bar{X} \neq \mu$       **how close is sample mean** to the actual population?
- Is a sample Mean of size (n) a good estimate** of the population mean?

- 
- Samples are different
  - There are many combinations
  - Sample mean may be different
  - Sample % may be different

- **Do you need to increase n** to make sample mean closer to population mean?

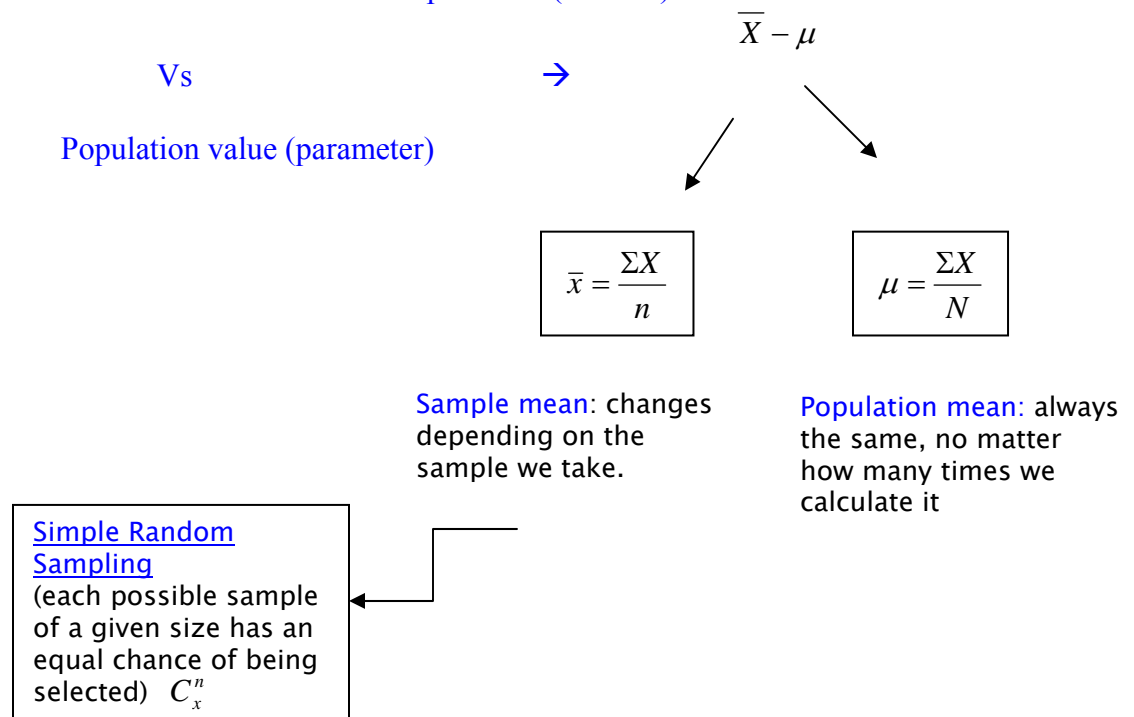
$$(\bar{X} \Rightarrow \mu)$$

**Objective of Sampling** → To gather data that mirrors a population

→ However, we would rarely know if objective data would be achieved!!! *We would need the population count information.*

→ Sampling needs to be chosen randomly to avoid bias: *to ensure that it reflects characteristics of the population*

**Sampling error** → Difference between Sample value (statistic)



**Potential for extreme sampling error is greater when smaller – sized sample are used**

However, there are cases when larger samples are no guarantee of smaller error

## 2. Sampling Distribution of the mean

Business applications use Simple random sampling  $C_x^n$

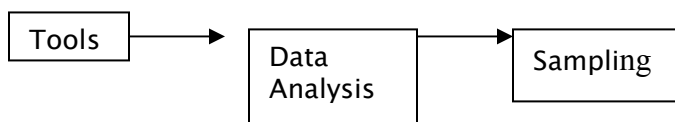
“True” Sampling Distribution

Distribution of the possible values of a statistic for a given-size random sample selected from a population

above  
population  
below

We can use Excel features for sampling; let’s remember the procedure. Let’s say that we want to pick random samples of 10 observations  $n=10$  out of a population of size 200. We know that the population mean is  $\mu=2.505$ , let’s proceed:

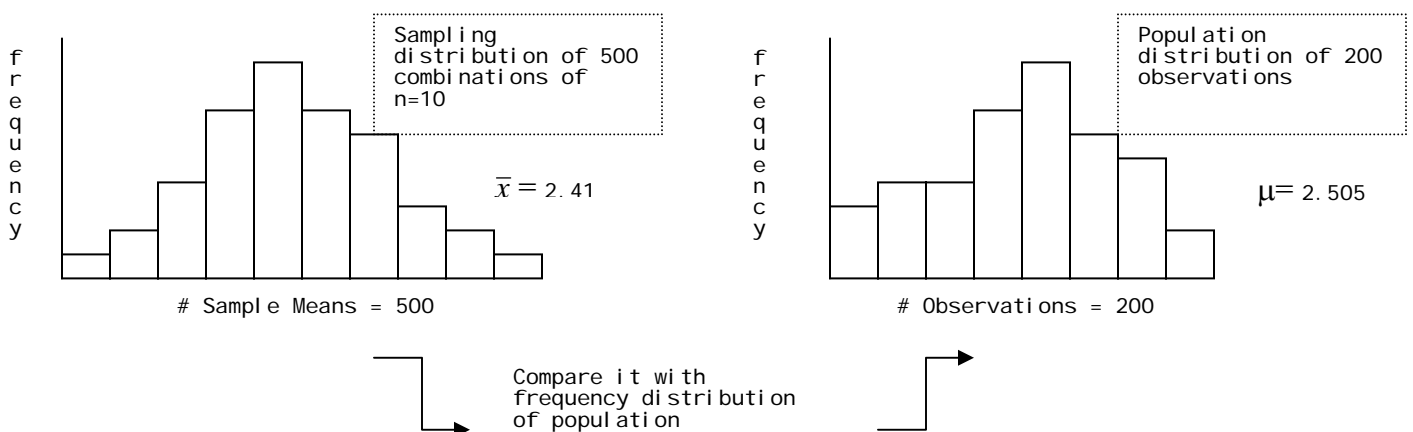
Excel → Select repeated samples



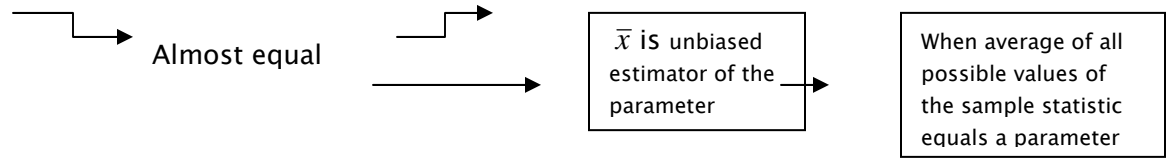
- Population → (X1...X200)
- “Random Sampling”
- $n = 10$
- Output option (in the same page)
- ok

You can calculate Sample mean, standard deviation and all the statistics that you have learned.

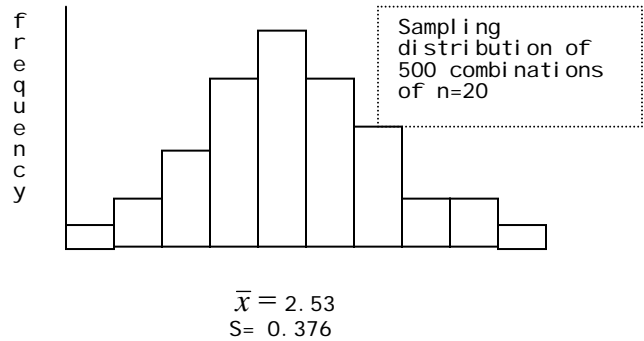
If you **repeat this same sampling operation 500 times**, you can build a histogram with the means of each sample, something like this:



1. Sampling Distribution takes the shape of a **bell curve**
2.  $\bar{x} = 2.41$  is the **Mean of sample means** vs.  $\mu_{\bar{x}} = 2.505$  **Mean of population**



3.  $\sigma_{\bar{x}} = 1.507 > S = 0.421$



If  $n \uparrow$  distribution of Sample mean will become shaped more like a normal

It's almost impossible to calculate a TRUE Sampling distribution, as there are so many ways to choose samples, and each one of them may have different means, standard deviations and statistics. We won't know which the right one is unless we compared it to the Population (if we get to have it available). Therefore, in order to make the process simpler we can use two theorems:

### Theorem 6-1

If population is normally distributed  
With mean  $\mu$  and  
standard deviation  $\sigma$

- Sample distribution of sample mean is also normally distributed with:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

*(used when population is*

*Normally distributed)*

We can use the Standard Normal Distribution, and easily make conclusions about the behavior of parameters, by looking at the Statistics. We use Z value to express the Sampling Distribution of  $\bar{x}$ .

Z tells how distant  $\bar{x}$  is from  $\mu$

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



However  
If  $n = 5\%N$  (large sample!)  
and sampling is w/o  
replacement, we use "Finite  
population correction factor"

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

$$\sqrt{\frac{N-n}{N-1}}$$

### Theorem 6-2 The Central Limit Theorem

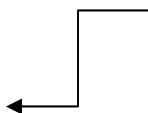


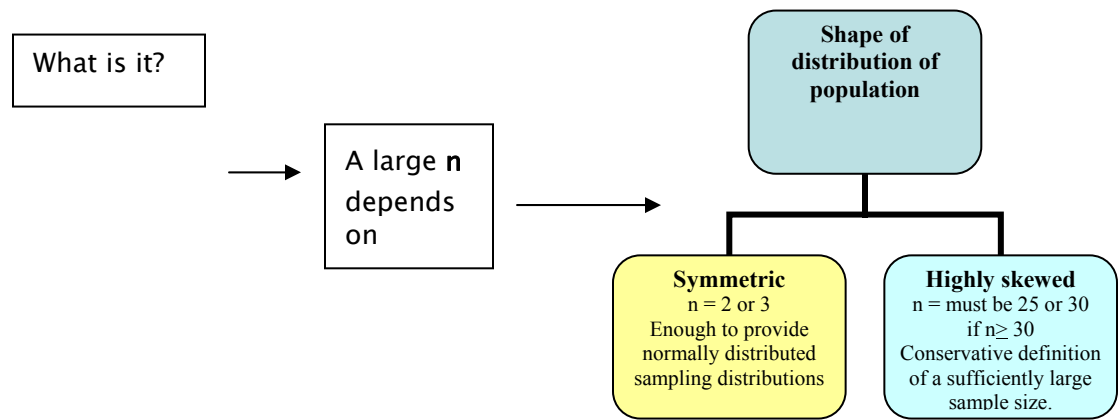
Any population with  $\mu, \sigma$ ; will  
result in a sample with mean  
 $\mu_{\bar{x}} = \mu$  and  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

*(used when population is not normally  
Distributed e.g. weight, income in a*

If n is sufficiently large.  
The larger the sample size, the better the  
approximation to Normal distribution

*region)*





### 3. Sampling Distribution for Proportions

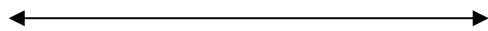
When information about population is given in proportions, the sampling procedure requires slight modifications to apply the Central Limit Theorem, let's explain it:

**Population proportion**

$$\rightarrow p = \frac{X}{N}$$

**Sample proportion**

$$\rightarrow \bar{p} = \frac{x}{n}$$



**Sampling error =**  
 $\bar{p} - p$

*Notice that population variables are capitalized.*

*Notice that Sample variables are lowercase.*

p has **BINOMIAL** as underlying distribution,  
but when  
**np** and **n(1-p)** are large  $\rightarrow$  **p** is treated as normal distribution

**Sampling Distribution of  $\bar{p}$**

$$\mu = p$$

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

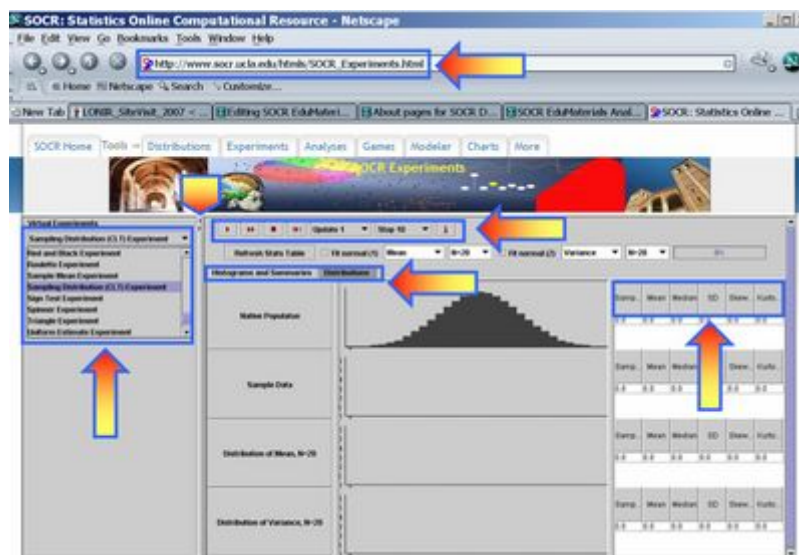
$$Z = \frac{\bar{p} - p}{\sigma_{\bar{p}}}$$

If  $np > 5\% N \rightarrow \sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$

## SOCR CLT Experiments

[http://wiki.stat.ucla.edu/socr/index.php/SOCR EduMaterials Activities GeneralCentralLimitTheorem](http://wiki.stat.ucla.edu/socr/index.php/SOCR_EduMaterials_Activities_GeneralCentralLimitTheorem)

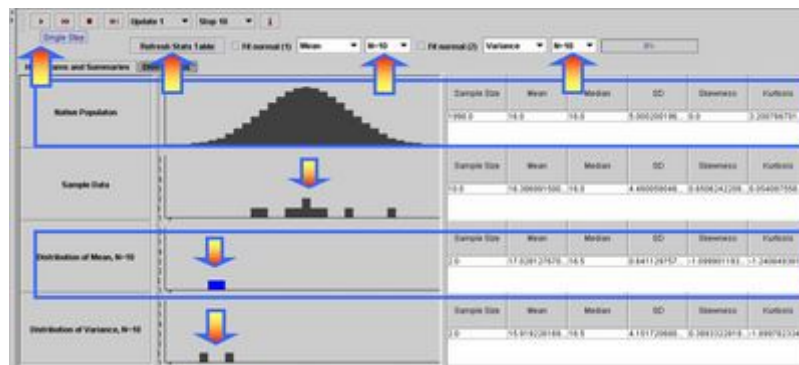
To start the this Experiment, go to [SOCR Experiments \(socr.ucla.edu/htmls/SOCR\\_Experiments.html\)](http://socr.ucla.edu/htmls/SOCR_Experiments.html) and select the SOCR Sampling Distribution CLT Experiment from the drop-down list of experiments in the left panel. The image below shows the interface to this experiment. Notice the main control widgets on this image (boxed in blue and pointed to by arrows). The generic control buttons on the top allow you to do one or multiple steps/runs, stop and reset this experiment. The two tabs in the main frame provide graphical access to the results of the experiment (Histograms and Summaries) or the Distribution selection panel (Distributions). Remember that choosing sample-sizes  $\leq 16$  will animate the samples (second graphing row), whereas larger sample-sizes ( $N > 20$ ) will only show the updates of the sampling distributions (bottom two graphing rows).





## Experiment 1

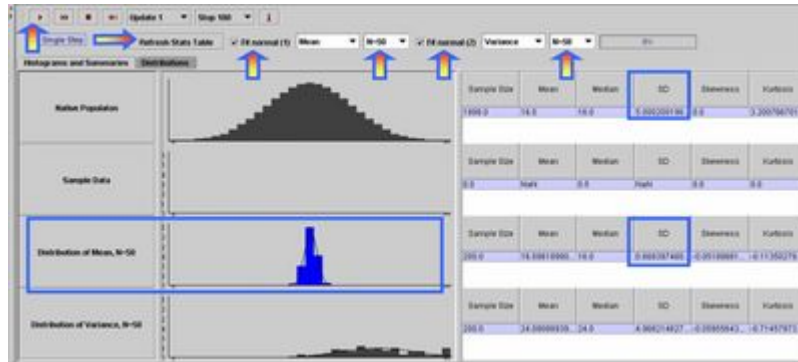
Expand your Experiment panel (right panel) by clicking/dragging the vertical split-pane bar. Choose the two sample sizes for the two statistics to be 10. Press the **step**-button a few of times (2-5) to see the experiment run several times. Notice how data is being sampled from the native population (the distribution of the process on the top). For each step, the process of sampling 2 samples of 10 observations will generate 2 sample statistics of the 2 parameters of interest (these are defaulted to *mean* and *variance*). At each step, you can see the plots of all sample values, as well as the computed sample statistics for each parameter. The sample values are shown on the second row graph, below the distribution of the process, and the two sample statistics are plotted on the bottom two rows. If we run this experiment many times, the bottom two graphs/histograms become good approximations to the corresponding sampling distributions. If we did this infinitely many times these two graphs become the sampling distributions of the chosen sample statistics (as the observations/measurements are independent within each sample and between samples). Finally, press the **Refresh Stats Table** button on the top to see the sample summary statistics for the native population distribution (row 1), last sample (row 2) and the two sampling distributions, in this case *mean* and *variance* (rows 3 and 4).



## Experiment 2

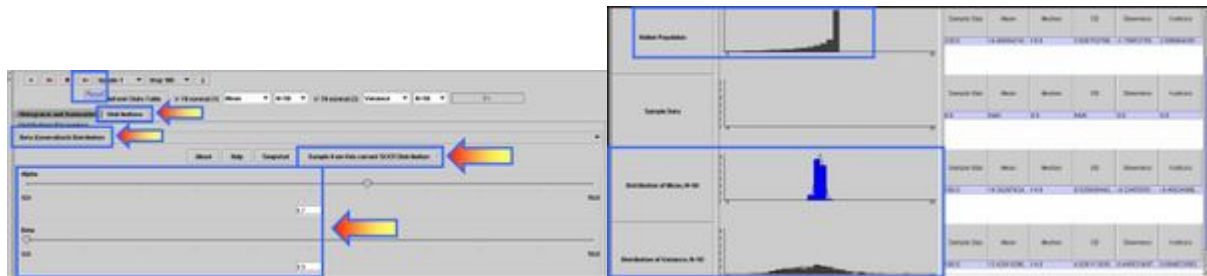
For this experiment we'll look at the mean, standard deviation, skewness and kurtosis of the sample-average and the sample-variance (these two sample data-driven statistical estimates). Choose sample-sizes of 50, for both estimates (mean and variance). Select the **Fit Normal Curve** check-boxes for both sample distributions. **Step** through the experiment a few times (by clicking the Run button) and then click **Refresh Stats Table** button on the top to see the sample summary statistics. Try to understand and relate these sample-distribution statistics to their analogues from the native population (on the top row). For example, the mean of the multiple sample-averages is about the same as the mean of the native population, but the standard deviation of the

sampling distribution of the average is about  $\frac{\sigma}{\sqrt{n}}$ , where  $\sigma$  is the standard deviation of the original native process/distribution.



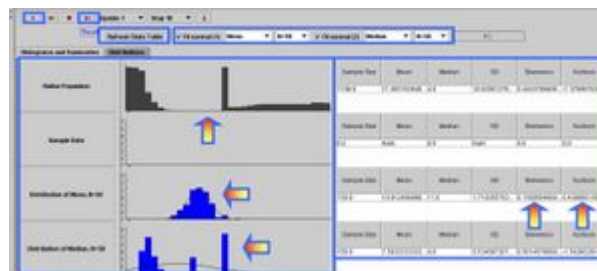
### Experiment 3

Now let's select any of the [SOCR Distributions](#), sample from it repeatedly and see if the central limit theorem is valid for the process we have selected. Try Normal, Poisson, Beta, Gamma, Cauchy and other continuous or discrete distributions. Are our empirical results in agreement with the CLT? Go to the **Distributions** tab on the top of the graphing panel. Reset the experiments panel (button on the top). Select a distribution from the drop-down list of distributions in this list. Choose appropriate parameters for your distribution, if any, and click the **Sample from this Current Distribution** button to send this distribution to the graphing panel in the **Histograms and Summaries** tab. Go to this panel and again run the experiment several times. Notice how we now sample from a Non-Normal Distribution for the first time. In this case we had chosen the Beta distribution ( $\alpha = 6.7, \beta = 0.5$ ).



### Experiment 4

Suppose the distribution we want to sample from is not included in the list of [SOCR Distributions](#), under the **Distributions** tab. We can then draw a shape for a hypothetical distribution by clicking and dragging the mouse in the top graphing canvas (Histograms and Summaries tab panel). This way you can construct contiguous and discontinuous, symmetric and asymmetric, unimodal and multi-modal, leptokurtic and mesokurtic and other [types of distributions](#). In the figure below, we had demonstrated this functionality to study differences between two data-driven estimates for the population center - sample [mean](#) and sample [median](#). Look how the sampling distribution of the sample-average is very close to Normal, where as the sampling distribution of the sample median is not.



## Questions

- What effects will asymmetry, gaps and continuity of the native distribution have on the applicability of the CLT, or on the asymptotic distribution of various sample statistics?
- When can we reasonably expect statistics, other than the sample mean, to have CLT properties?
- If a native process has  $\sigma_X = 10$  and we take a sample of size 10, what will be  $\sigma_{\bar{X}}$ ? Does it depend on the shape of the original process? How large should the sample-size be so that  $\sigma_{\bar{X}} = \frac{2}{3}\sigma_X$ ?