Sampling and Sampling Distributions

- **Normal Distribution**
- **Aims of Sampling**
- **Basic Principles of Probability**
- **Types of Random Samples**
- **Sampling Distributions**
- **Sampling Distribution of the Mean**
- **Standard Error of the Mean**
- **The Central Limit Theorem**

Sampling

- **Population** – A group that includes all the cases (individuals, objects, or groups) in which the researcher is interested.
- **Sample** – A relatively small **subset** from a population (uses inferential statistics).

What is the purpose of sampling:

- The whole population is too large to question everyone (or all cities or whatever the unit of analysis is).
- It would cost too much money and take too long.
- Therefore, we choose to select a "sub-group" of the whole population to ask our questions. This will cost less and take less time.

So, what do we want our sample (i.e., sub-group of the whole population) to be able to do?

- Reflect the whole population
- We will be able to "infer" from our sample what is true for the whole population.

  Inferential statistics are used to generalize from our "sub-group" (or sample) to the whole population.

So, what is the aim of sampling?
to create a "sub-group" that will allow us to determine what is true of the population without having to question (or collect data on) the entire population.

Sampling Terms

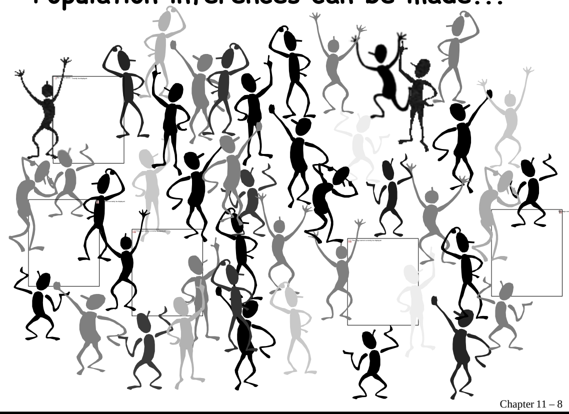
- **Parameter** – A measure used to describe a **population distribution** (for example, mean or standard deviation).
- **Statistic** – A measure used to describe a **sample** (for example, mean or standard error).

Notation

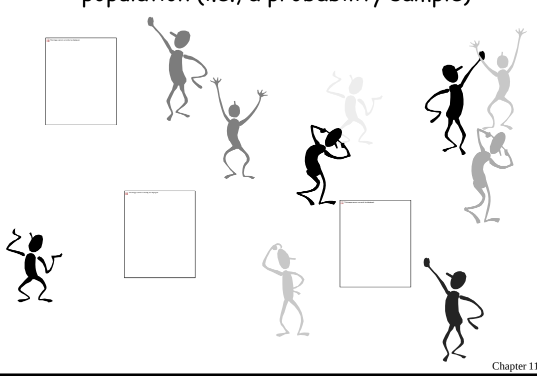
Table 11.1 **Sample and Population Notations**

Measure	Sample Notation	Population Notation
Mean	$\bar{Y}$	μ_Y
Proportion	p	π
Standard deviation	S_Y	σ_Y
Variance	S_Y^2	σ_Y^2

Population inferences can be made...



...by selecting a <u>representative</u> sample from the population (i.e., a probability sample)



Probability Sampling

Every case has an equal chance of being selected for the sample.

A probability sample is one where inferences can be made about the whole population with a "known amount" of confidence in our inference.
(That is, we may be highly confident in our inference about the population or not very confident.)

Probability Sampling

A probability sample is a representative sample of the whole population.

A probability sample is a random sample. There are various ways of obtaining a random sample.

Probability Sampling: Simple Random Sampling

A sample designed in such a way as to ensure that:

every member of the population has an **equal chance of being chosen**

(This can be done using a table of random numbers, computer, or other means; Appendix A in your book provides a Table of Random Numbers)

Probability Sampling: Systematic Random Sampling

A method of sampling in which **every *K*th member in the total population is chosen** for inclusion in the sample (for example every 10th member).

To determine the very first case selected use simple random sampling and include only the first k members of the population (e.g., if the skip interval is ten, use simple random sampling to choose the first case among the first 10 cases in the population).

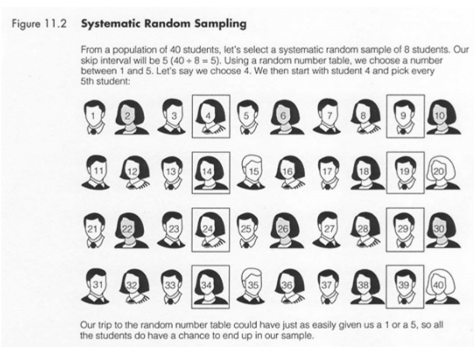
Probability Sampling: Systematic Random Sampling

The **specific skip interval used** (such as every 10th) **is typically determined based on the sample size desired**.

To determine the skip interval, divide the total number of cases in the population by the sample size desired.

For example, if there are 4000 cases in the population and you want a sample size of 400, you would divide 4000 by 400 for a skip interval of every 10th case.

Systematic Random Sampling



Probability Sampling: Stratified Random Sampling

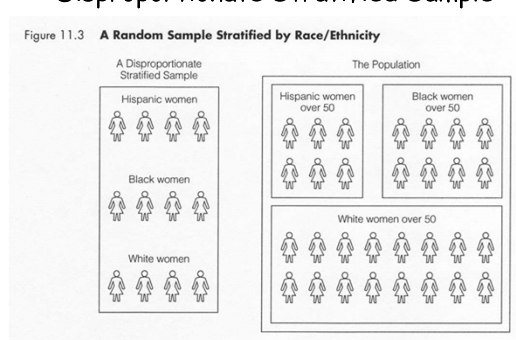
A method of sampling obtained by:

(1) **dividing the population into strata** (or sub-groups) based on one or more variables central to our analysis and

(2) then drawing a simple random **sample from each of the strata** (i.e., subgroups)

- **Proportionate stratified sample** – The size of the sample selected from each subgroup is proportional to the size of that subgroup in the entire population.
- **Disproportionate stratified sample** – The size of the sample selected from each subgroup is disproportional to the size of that subgroup in the population.

Disproportionate Stratified Sample



What is sampling error?

Or, we could ask:
when we select a sample, will the sample statistic (such as the sample mean) be identical to the population parameter (such as the population mean)? The answer is “probably not”. It will probably be similar but not “identical”.

The difference between the sample statistic (such as a sample mean) and the actual population parameter (such as the population mean) is called sampling error.

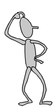
For example,

if we drew a sample of students and found the mean age for these students to be 26,

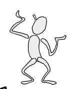
but we happen to know the mean age for the whole population of students (from which the sample was drawn) is 24,

the **sampling error** would be “?”

The Delimma



Each time we take a sample, our sample statistic (such as the sample mean) will vary somewhat from the actual population parameters (such as the population mean), resulting in some degree of sampling error.



So, when we draw a sample, how much confidence can we have that the sample’s statistics (such as the mean for a particular variable) are similar to that of the whole population’s parameters?

That is, what is the probability that the sample’s statistics are a good reflection of the total population.

The answer to this dilemma is to apply knowledge that we have about a device known as the ”sampling” distribution to our sample distribution.

The sampling distribution is a “theoretical” distribution involving multiple samples while a sample distribution involves only a single sample.

Since the distribution of statistics within a single sample have similar characteristics as the distribution of statistics within a sampling distribution, statisticians have found it useful to uncover the characteristics of the sampling distribution.

To create a sampling distribution, we would need to take as many different samples as is possible from the population and:

(1) determine the mean for each of these samples (or other statistic of interest), and

(2) plot all these sample means to create a distribution of the means known as the sampling distribution.

Why is the sampling distribution considered a “theoretical” distribution?

The Central Limit Theorem gives us information about the sampling distribution that we can then apply to a single sample.

The Central Limit Theorem tells us that:

When drawing samples to create the sampling distribution, the larger the size of the samples, the more likely that a perfect normal curve will be created and the resulting statistics will be identical to the population’s parameters (the samples drawn should be at least 50 cases and preferably 150).

This suggests that, as the size of our single sample is made larger, our single sample will also be more reflective of a normal curve (assuming the population distribution is at least somewhat normal).

Further, since we know that a sufficiently large single sample will typically resemble a normal curve,

we can apply characteristics of the normal curve to our single sample if it is large.

What do we mean by large?

Ideally, the sample would have at least 150 cases. However, some statisticians say 50 is plenty and, if the population is fairly normally distributed, 30 cases is enough.

So, when studying the distribution of a population we would calculate the standard deviation and then apply characteristics of the normal curve.

When studying the distribution of a sample, we would calculate the standard error and then apply characteristics of the normal.

The standard deviation for the population is referred to as the standard error when speaking of a sample.

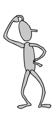
While the standard deviation and the standard error are conceptually the same thing, the former is used with populations and the latter with samples.

Further, while they are conceptually the same thing, the standard error (of a sample) is calculated slightly differently than the standard deviation (of a population).

Because we can apply characteristics of the normal curve to our single sample, we can obtain information about the total population.

For example, when considering our sample mean and our sample standard error, we can be 68% confident that the true population mean falls somewhere between +1 and -1 standard errors of our single sample mean.

Similarly, we can be 95% confident that the true population mean falls somewhere between +2 and -2 standard errors of our single sample mean.



<u>Lets demonstrate this with an example</u>: Lets suppose we drew a large single sample of prison inmates and gave them an aggression test.

We found our sample mean to be 74 and the standard error to be 2.

Applying characteristics of the normal curve to our sample, we could be 95% confident that the true population mean (i.e., the mean aggression score for all prison inmates) falls somewhere between 70 (minus two standard errors from the mean) and 78 (plus two standard errors from the mean).

凩訝

(zijian, see you later)