# Section 5A

Fundamentals of Statistics

**Statistics** (singular) – science of collecting, organizing, and interpreting data

**Statistics** (plural) – the data that describe or summarize something

The **population** is a statistical study is the *complete* set of people or things being studied.

Most populations are too large or too spread out to actually obtain data from all of them, so instead we use a sample.

The **sample** is the subset of the population from which the raw data are actually obtained.

**Population parameters** are specific characteristics of the population that a statistical study is designed to *estimate*.

Since we usually can't use the entire population, we are required to estimate the population parameter(s).

**Sample statistics** are numbers or observations that summarize the raw data.

Astronomers typically determine the distance to a galaxy by measuring the distances to just a few stars within it and taking the mean (average) of these distance measurements. Describe the population, sample, population parameters, and sample statistics.

**Population** – all stars in the galaxy

**Sample** – the few stars selected for measurement

**Population parameter** – mean (average) distances between *all* stars and Earth

**Sample statistic** – mean (average) distances between the stars in the sample and Earth

# Basic Steps in a Statistical Study

1. State the goal of your study precisely. Determine the population you want to study and exactly what you'd like to learn about it.
2. Choose a representative sample from the population.
3. Collect raw data from the sample and summarize by finding sample statistics.
4. Use the sample statistics to infer the population parameters.

   - sample must be representative of the population for these to be reasonable
5. Draw conclusions. Determine what you've learned and whether you achieved your goal.

# Choosing a Sample

Choosing a sample can be the most important step in a statistical study.

If the sample is not representative of the population, the inferences made about the population will not be valid.

Example: If we wanted to find out about the average height of all AC students, would it be reasonable to go to the gym and choose the basketball players to be our sample?

A **representative sample** is a sample in which the relevant characteristics of the sample members match those of the population.

# Common Sampling Methods

**Simple random sampling** – choose a sample of items in such a way that every sample of a given size has an equal chance of being selected

- many calculators & computers have random number generators

- can also find random number tables or can be as simple as drawing names (or numbers) out of a hat

**Systematic sampling** – use a simple system to choose the sample

- choose every $5^{th}$ person, etc.

# Common Sampling Methods

**Convenience sampling** – use a sample that is convenient to select

- could be people in the same room

- not necessarily the best method, but sometimes the only way to be able to choose a sample

**Stratified sampling** – identify the subgroups and then draw a simple random sample within each subgroup. The total sample consists of all the samples from the individual subgroups.

- this method is used when we are concerned about differences among subgroups (strata) within a population

There are different situations where some of the common sampling methods cannot be used.

Sometimes, we have to use methods that wouldn't normally be the most preferable method.

It is important to remember that no matter which method is used:

- the study can only be successful if the sample is representative of the population

- it is still just a sample

A statistical study suffers from **bias** if its design or conduct tends to favor certain results.

Bias can be introduced in a study intentionally or non-intentionally.

When looking at results from a study conducted by others, you should always consider any possible bias.

# Types of Statistical Studies

**Experiment** – a study where researchers apply a treatment to some or all of the sample members and then look to see whether the treatment has any effects

**treatment group** – the group of sample members who receive the treatment being tested

**control group** – the group of sample members who do **not** receive the treatment being tested

It is important for the treatment and control groups to be selected randomly and to be alike in all respects except for the treatment.

A **placebo** lacks the active ingredients of a treatment being tested in a study, but is identical in appearance to the treatment.

The participants should not be able to distinguish the placebo from the real treatment.

The **placebo effect** refers to the situation in which patients improve simply because they believe they are receiving a useful treatment.

# Blinding in Experiments

An experiment is **single-blind** if the participants do not know whether they are members of the treatment group or the control group, but the *experimenters do know*.

An experiment is **double-blind** if *neither* the participants nor the experimenters know who belongs to the treatment group or the control group.

# Types of Statistical Studies

**Observational study** – a study where researchers observe or measure characteristics of the sample members but do **not** attempt to influence or modify these characteristics

A **case-control study** is an observational study that resembles an experiment because the sample naturally divides into two (or more) groups.

- The participants who engage in the behavior under study are called the **cases**, which makes them like a treatment group in an experiment.

- The participants who do not engage in the behavior are called the **controls**, making them like a control group in an experiment.

# Surveys and Opinion Polls

The **margin of error** in a statistical study is used to describe a **confidence interval** that is likely to contain the true population parameter.

We find this interval by adding and subtracting the margin of error from the sample statistic obtained in the study.

**From** (sample statistic – margin of error)

**to** (sample statistic + margin of error)

A poll is conducted the day before an election. The poll shows that 53% of the voters surveyed favor candidate A, with a margin of error of ±2.5 percentage points. Find the confidence interval.

The sample statistic from the survey is 53%.

The margin of error is ±2.5%.

(53% - 2.5%) = 50.5%

(53% + 2.5%) = 55.5%

The confidence interval is **from 50.5% to 55.5%**