## §1.1 Overview

When someone says statistics most people think of baseball type statistics – numbers that summarize. There is actually more to statistics than just summary numbers.

**Statistics** is a method of analysis involving:

- Designing an experiment
- Gathering data
- Organizing the data
- Summarizing the data
- Analyzing & interpreting the data
- Drawing conclusions based upon the analysis of the data
- Presenting the data & findings

**Data** is observations collected. These observations can be of 4 types, which will be discussed in §1.3. Some examples of data that we will later classify are:

- Gender
- Good/Better/Best Responses on a survey
- Temperatures
- Incomes of a specific group

How data is collected is an important part of the statistical process and we will talk about its importance in some detail. Suffice it to say that data must be collected by a process that statisticians call **random selection**. Data is collected from the **population** (all elements in a group with a particularly defined set of characteristics), but it is generally impossible to get a **census** (data from every member in the population), and therefore we talk about a **sample** (a portion of the population) and require this sample to be a **random sample** (a randomly chosen portion of the population) so that we may derive valid conclusions based upon the data.

## §1.2 Statistical Thinking

In this section Triola is trying to make you think like a good consumer. He wants you to think about the following:

- What is the context of the data?
- How was the data collected?
- Who collected the data?
- Are the conclusions drawn easily understood by everyone?
- What practical knowledge might be taken away from what was learned?
- Differentiate between practical and what is actually statistically significant.

Many of these concepts are elaborated upon in other sections, so I am not going to lecture on this section.

## §1.3 Types of Data

We just discussed a population and a sample and ∴ we know that a **population is every element in a set** and a **sample is a subset**. Now we must talk about summarizing a set and a subset. A **statistic** which we most often hear about, is the word used for *summarization of a sample* (the numerical measure describing the characteristics of a sample)! It is what we most often hear, because data most often refers to a sample. If and when we summarize information from a *population* then we would talk about the **parameters** (the numerical measure describing the characteristics of a population). It may be helpful to use alliteration (words that sound the same) to help remember this:

Sample → Statistics

Population → Parameters

Data can have 2 main types of characteristics – **qualitative** (Categorical/Attribute) **quantitative**. *Quantitative* data describes *quantities* where as *qualitative* describes *qualities*.

**Quantitative** Examples**:** Height, Weight, Other Measurements or Counts
Specific Ex: # of people in a room
Length of time between events

**Qualitative** Examples**:** Good, Bad, Excellent; Yes or No; Non-countable Qualities
Specific Ex: 0's & 1's denoting on/off in computers
Ratings on a scale of 0 to 5

For *quantitative* data there are 2 types – **discrete** and **continuous**. *Discrete* data has no in-between values, they are countable or finite and *continuous* data has infinitely many continuous values throughout a scale with no jumps or breaks.

**Discrete** Examples: Positive Integers – Can't be 2.982
# of People in a Room
**Continuous** Examples: Real Numbers (can it be viewed on a number line?)
Length of time between events

So far we have discussed 4 ways of classifying data. There are four more classifications of data. These are called the **Levels of Measure** by Triola. They are as follows:

2 Types that come under the Qualitative/Categorical/Attribute Heading

**Nominal** – Names that have no meaning with respect to order
Ex: Gender
Party Orientation
Nationality

**Ordinal** – Ordered Naming Scheme, but differences make no sense
Ex: Rating of item
Placing a spelling bee
Grades

2 Types that come under the Quantitative Heading

**Interval** – Numeric where differences make sense, but no **natural** zero
Ex: Temperatures
Years

**Ratio** – Numeric where differences and ratios make sense & there is a natural zero.
Ex: Incomes of politicians
Number of yes votes

These 4 levels of measure always give students a bit of trouble at first, especially the difference between interval and ratio. The **ratio test** may help. Ask yourself does a ratio make sense; does twice a value really mean twice?

Ex: Temperatures are not ratio data because 30° does not mean twice as hot as 15°.

For a nice summary of these classifications see the table on **page 15**.

## §1.4 Critical Thinking

Statistics and statisticians have a bad reputation because of the misuse of statistics by those who wish to deceive or by those who don't know any better. It is our hope that we will give you the skills to tell when statistics have been improperly used and to know how to correctly use and interpret them.

The most important thing in statistics, as previously mentioned, is **sampling**. Some common errors in sampling include:

**Small Samples**: Too small a subpopulation based upon the population.

Example: TV ads claiming 9 out of 10 mom's prefer a certain product

*Problem*: Were only 10 moms chosen or is the claim a ratio of 1000's of moms.

**Voluntary Response Sample:** People who respond are of a specific sub- population in the population desired.

Example: Poll taken by a radio station via call-ins from listeners

*Problem*: A certain type of person is likely to respond to such a poll and therefore the results can't be extrapolated back to the entire population of listeners, but rather are only valid to a subpopulation.

Many of the familiar visuals used by media and publications to **report statistics** can also be misleading – either intentionally or unintentionally.

**Graphs & Pictographs** can be used in misleading ways.

Graph Ex: Showing only a portion of a graph exaggerating the difference between the things being shown.

Pictograph Ex: Pictures that don't represent the intended difference. Such as using a cube with double side length to represent 2 times the amount.

**Percentages** can also be misleading and hard to interpret if the definition is not understood

- A percentage is a part of 100.
- 100% is ***all*** of the sample or population.

Working with percentages:

**% → #**

1. Sample Size Known
2. Convert % to Decimal or Fraction
   % → Decimal move the decimal 2 places to the left
   % → Fraction write # in % over 100. If decimal in #, then multiply by a factor of 100.
3. Multiply Decimal/Fraction by Sample Size

Ex: According to a survey, 82% of adults go to bed after 9 pm. So, in a group of 150 adults we should expect how many to go to bed after 9 pm.

The **Survey Tool** (the instrument) can also be a source for error.

**Person Giving** (Rosenthall Effect) – Elicit a certain response due to body language, verbal context, etc.

**Biased Questions** (Loaded Q's) – Questions asked to elicit the response desired by the developer.

**Ordering of Q's** – Order of questions can change people's responses – for the good or bad. For instance, surveys do not start with personal questions in general because people are not comfortable giving personal info until they know a person better or until they have begun to feel more comfortable with the survey in general.

**Self-Preservation** – Answering a question to make oneself look better

The **Interpretation** of data can also be flawed.

**Correlation & Causality** (Cause & Effect) – Because 2 things are proven to be related does not mean that they have a cause and effect relationship.Many times when data is interpreted or reported this type of cause and effect is implied to mislead. Sometimes it is interpreted by people incorrectly. (More discussion of this will be seen in Ch. 10: Correlation & Regression)

**Self-Interest Study** – A group should not conduct studies that could lead to monetary gain – we should be wary of such studies.

**Precise Numbers** – Although a precise number may sound very strong, it should have the opposite effect. We can never be precise when gathering info. Think of it like the $\Re$ -- when we think we've found the last number we can always find one more.

**Partial Pictures** – Sometimes a statistic is completely valid, but if you do not know the population for which this statistic is valid, then it is not a very good statistic.

**Deliberate Distortions –** Sometimes people deliberately rely on statistics to tell their story in the best light and they have used some of the described methods to do so. Sometimes they just use the power of a survey's findings to sway people and they don't even have a survey to back up their claims.

## §1.5 Collecting Sample Data

Good statistics rely heavily on 2 basic things: **Good Design & Correct Sampling** (random sampling!!)

**4 Elements of Good Design**

1) Define exactly the question at hand & the population it encompasses
2) Develop a plan to collect data
3) Collect data being aware of problems
4) Analyze and draw conclusions, pointing out sources of error

**There are 2 types of studies:**

**Observational**

Observe & measure specific characteristics with no attempt to modify subjects under study.

Ex. Survey of adult population about whether they suffer from migraines.

**Experimental**

A treatment (some times lack of treatment, placebo) is applied and then observations are made on the effects.

Ex. Migraine medicine given to sufferers and observed the results .

**Types of Observational Studies:**

**Cross-Sectional** (most common) – Done at one pointing time.

**Retrospective** (case-control) – Information is collected about the past via records, interviews, etc.

**Prospective** (longitudinal/cohort) – Data is collected from groups (cohorts) through a period of time (usually to show some type of cause and effect relationship).

**Designing a Good Experiment**

1. First an experiment must be **well run**.
2. It should **contain**:
   a. **Control Group** – To which no treatment is applied
   b. **Treatment Group** – To which treatment is given

**Problems**

1) **Placebo Effect**

This is change to the **experimental units** (person of think receiving treatment) receiving no treatment, be it real or imagined.

2) **Confounding**

This is the effect of different factors on the actual thing being studied.

**Control for Placebo Effect**

**Blind Experiment** – Groups don't know which group they are in

**Double-Blind Experiment** – Neither the groups nor the experimenter knows which group the experimental units are in

**Controls for Confounding**

**Block Designs:** A ***block*** is a group with similar characteristics.

**Completely Randomized** – In this type the blocks are chosen in a random way so that there is a mixture of all confounding variables within each block. (Preferred)

**Rigorously Controlled** – In this type the blocks are specifically chosen to contain experimental units with similar characteristics that may effect the experiment

**Randomness** – This means that the sample is chosen in a completely random manner.

**Replication** – This can refer to a sufficient number of subjects being used to truly recognize different treatment effects as well as being able to replicate the experiment and achieve the same results.

Methods of sampling that give us sufficient randomness are VERY important. A **random sample** is the most common.

**Random Sample** – A sample created in a random manner such that all subpopulations are expected to exist in proportion to their actual existence within the population. Use of random numbers is common to generate a sample .

**Simple Random Sample** – A sample created in a random manner such that a *sample of size n* has the *same chance* to be chosen as *any sample of size n.*

The book will refer to these two types of sampling throughout the remainder of the course.

## Types of Sampling

**Random** – See above

**Systematic** – A list is compiled and then an experimental unit is chosen at consistent intervals from the list.

**Stratified** – Different characteristics are noted and thus a proportional mount of each of the subpopulations is randomly sampled. Key is subdivided in 2 or more subpopulations. (Random sample from groups) Won't result in **simple r.s.**

**Cluster** – Groups are divided and then some groups are chosen at random (Random sample of groups) Won't result in **simple r.s.**

**Convenience** – Choosing people based upon ease of sampling (many surveys)

For examples and a nice pictorial break down please see p. 29 of Triola's Essentials Ed 4

Another type of sampling used by professionals and the government is a **multistage sampling** design. In this design a combination of the discussed types of sampling are used.

With all the above as clear as mud, it remains to be said that a sample needs to control error.

## 2 Types of Error

**Sampling Error** – Error that we ***can't control***. The difference that exists between a sample and the true population. (Due to the individuality of people)

**Non-Sampling Error** – Errors that we ***can control***. Errors from incorrect sampling procedure, biased instruments, Rosenthall Effect, poorly designed instruments, etc. In short the things that we have spent 2 sections discussing!

Much of our study will revolve around how to analyze the error that we can't control – Sampling Error!