

# STATISTICS 301

Text: Walpole, Myers, Myers, and Ye (9<sup>th</sup> Ed.)

## Chapter 1: Introduction

### 1.1 Basic Terms (Section 1.1, text)

Statistics - Science of collecting, summarizing, analyzing, and interpreting data (information, usually in the form of numbers).

Two major parts of Statistics:

Descriptive Statistics - Summarizing data.

Inferential Statistics - Analyzing and interpreting data.

The analysis depends on how the data was collected.

Inferential Statistics is built on top of some ideas in Probability.

Probability - Field of mathematics involved with determining the relative frequency of certain events.

Thought processes involved in Probability and Inferential Statistics:

Probability - Deductive thought process (known facts imply new facts.)

This is analogous to selecting some elements from a box with **known** contents and asking “what are the chances of getting certain results?”

Inferential Statistics - Inductive thought process (observations are used to infer properties).

This is analogous to selecting some elements from a box with **unknown** contents and making a guess about the totality of the contents of the box.

Basic (most important) terms in Statistics:

Population - Set of all elements that we want information about.

Sample - A subset of a population about which we actually collect information.

Parameter - A numerical characteristic of the population.

Statistic - A numerical characteristic of a sample.

Example: Coin Toss Experiment

To illustrate the above concepts we will conduct the following experiment:

Toss a coin eight times; Before each toss, the students will predict the outcome of the toss.

For this experiment,

population - All possible tosses of the coin. (An infinite number of tosses.)

In some cases, in order to define the population from which our data comes, we try to describe the set of observations that we would obtain if the process generating the data were repeated indefinitely.

sample - 8 tosses of the coin.

statistic - the number of heads tossed in 8 tosses.

parameter - probability (proportion) of heads in all possible tosses.

Let  $X = \#$  of heads so far. (This is a random variable in Probability.)

Let  $x$  = the actual number of heads observed so far.

Toss	Predict	$x$	Prob that $X = x$	Expected value of $X$
1				
2				
3				
4				
5				
6				
7				
8				

Probability - Deductive thought process - Assumes known facts.

What known fact did we assume? That the probability of heads was .5.

The answers to the Probability questions (the last two columns of the table) are true (exact - without any error).

Statistical Inference - Inductive thought process - Based on observation.

Is the probability of heads for this coin equal to one-half?

This answer is based on what we observed, what we expected, and whether the probability of what we observed is small. (The probabilities were based on the assumption that the probability of heads was .5.)

What is the probability of heads for this coin?

Do we know the true answers to the above inferential statistics questions?

No.

Another Example: See the **Fuel Pump Data** in exercise 1.19, p. 31.

For this experiment,

population - All fuel pumps produced over a certain time period by a certain company. (Let's assume that there were 500 such fuel pumps.)

Actually, the population can be thought of as the 500 lengths of life for the 500 fuel pumps.

sample - the 30 fuel pumps selected (or their lengths of life).

parameter - average length of life for the 500 fuel pumps.

statistic - average length of life for the 30 fuel pumps. This average is 2.8 years.

Two questions of interest:

1. What is your guess for the average length of life for the 500 fuel pumps?

Is this answer the true value of the parameter? No, but it should be close. How close will depend on the probabilities involved in this problem.

2. Does the data refute a claim that the average length of life for the 500 fuel pumps is less than or equal to 2 years?

This question becomes 'Is the **sample** average (2.8) large enough to conclude that the **population** average is greater than 2?'

Again, as in the coin example, we need to look at probabilities and expected values assuming that the person's claim (average  $\leq 2$ ) was correct.

Note: The questions in statistical inference can generally be classified as either estimation questions or hypothesis testing questions. The first question is an example of an estimation question and the second is an example of a hypothesis testing question.

## 1.2 Sampling Concepts (Section 1.2)

To determine probabilities and expected values, we must start with some known facts. One of the known facts is how the data was collected. A sampling scheme must be used which

1. gives reasonable assurance that the sample is representative of the population
2. allows for computing probabilities about the statistic.

Census - a 100% enumeration of a population

Why take a sample instead of a census?

Convenience - refers to time, money, and effort

Necessity - when the act of making an observation destroys the element

Accuracy - when data collection requires highly skilled workers

Some possible sampling schemes:

Random Sample - Each element in a population is equally likely to be selected on each draw from the population and the draws are independent of one another.

Practically speaking, the elements are taken "at random with replacement."

Simple Random Sample - A simple random sample of size  $n$  elements from a population of  $N$  elements is such that every subset of  $n$  elements are equally likely.

Practically speaking, the elements are taken "at random without replacement."

As the population size gets large, then simple random sampling and random sampling are essentially the same.

Collecting a random sample or simple random sample must be done using a randomization device (computer, calculator, or random number table) and requires a sampling frame.

Sampling Frame - List of the elements in the population from which the sample will be drawn. (Let  $N$  = number of elements in the population.)

Examples of using **Minitab** to select a sample (also refer to the **MINITAB – STA 301** handout):

1. Take a random sample (of tag "or ID" numbers) of size 4 from a population of size 5.
2. Take a simple random sample (of tag numbers) of size 4 from a population of size 5.

Minitab Session Window:

```

MTB > Set c1
DATA> 1( 1 : 5 / 1 )1
DATA> End.
MTB > Sample 4 C1 c2;
SUBC> Replace.
MTB > Sample 4 C1 c3.
MTB > Print C1 C2 C3.

```

Create a set of patterned data →  
Random Sample →  
Simple Random Sample →  
Print the data into the session window →

**Data Display**

Row	C1	C2	C3
1	1	5	1
2	2	1	2
3	3	2	5
4	4	1	4
5	5		

- Take a random sample (of tag numbers) of size 30 from a population of size 500 (as in the **Fuel Pump Data** example).

Minitab Session Window:

```

MTB > Set c4
DATA> 1( 1 : 500 / 1 )1
DATA> End.
MTB > Sample 30 C4 c5;
SUBC> Replace.
MTB > Print C5.

```

**Data Display**

C5

388	71	120	94	145	204	402	141	34	468	477	69	196
313	288	171	128	424	158	261	181	58	453	156	317	64
478	396	40	96									

- Take a random sample of size 7 from the population defined by the 30 lengths of life of fuel pumps in the **Fuel Pump Data** example.

Minitab Session Window:

```

MTB > Retrieve
"C:\DOCUME~1\DUNN\LOCALS~1\TEMP\TEMPOR~3.ZIP\013238~1\DATADI~1\CHAPTE~1\MINITA
B\EX01.19.MTP";
SUBC> Port.
Retrieving worksheet from file:
'C:\DOCUME~1\DUNN\LOCALS~1\TEMP\TEMPOR~3.ZIP\013238~1\DATADI~1\CHAPTE~1\MINITA
B\EX01.19.MTP'
Worksheet was saved on Mon Jun 05 2006

Results for: EX01.19.MTP

MTB > Print 'life'.

Data Display

life
  2.0   3.0   0.3   3.3   1.3   0.4   0.2   6.0   5.5   6.5   0.2   2.3   1.5
  4.0   5.9   1.8   4.7   0.7   4.5   0.3   1.5   0.5   2.5   5.0   1.0   6.0
  5.6   6.0   1.2   0.2

MTB > Sample 7 'life' c2;
SUBC> Replace.
MTB > Print C2.

Data Display

C2
  1.2   2.5   6.0   0.7   1.5   2.5   1.2

```

Retrieve the data from  
Prentice Hall Web Page

The statistical inference procedures we will study in this class require that the data be collected according to a random sample (or a simple random sample from a large population).

Additional points:

1. How does the random sampling model relate to reality?

In some cases, data cannot be collected by an actual random sample. Before applying a statistical analysis that requires the random sampling procedure, one must decide whether the collected data may be viewed **as if** it could have been obtained by random sampling.

Example:

A biologist plans to study the distribution of body length in a certain population of fish in Chesapeake Bay. The sample will be collected using a fishing net.

Is this a random sample?

Can the data be viewed as if it could have been obtained by random sampling?

This is an example of sampling bias: the systematic tendency for some elements to be more readily selected than others. One wouldn't want to apply a statistical analysis procedure that requires the random sampling to such data.

2. It often happens that the population actually studied is narrower than the population that is of real interest. In such cases, one must argue that the results from the narrower population can be extrapolated to the population of interest. This extrapolation is not based on statistical grounds but on the subject matter grounds.

Example:

In the **Fuel Pump Data** example, we assumed that the data came from the population of all fuel pumps produced in a certain time period by a certain company. Perhaps the population of real interest is all fuel pumps produced in the certain time period by all companies (or maybe all fuel pumps including those produced currently).

On statistical grounds, we might estimate that the average length of life produced in the certain time period by the certain company is 2.8 years.

Estimating that average length of life for all fuel pumps (produced by all companies up to today) is 2.8 years cannot be defended on statistical grounds.