Chapter 1

The Where, Why, and How of Data Collection

Chapter Goals

After completing this chapter, you should be able to:

- Describe key data collection methods
- Learn to think critically about information
- Learn to examine assumptions
- Know key definitions

What is Statistics

Statistics is the *science* of data

The Scientific Method

1. Formulate a theory

2. Collect data to test the theory

3. Analyze the results

4. Interpret the results, and make decisions

Example

Exercise: Does the data always conclusively prove or disprove the theory?

The Scientific Method

The scientific method is an iterative process. In general, we **reject a theory** if the data were *unlikely* to occur if the theory were in fact *true*.

Tools of Business Statistics

- **Descriptive statistics**
- **Inferential statistics**

Statistical Inference

Statistical Inference

To use sample data to make generalizations about a larger data set (population)

Populations and Samples

- A **Population** is the set of all items or individuals of interest
- A **Sample** is a subset of the population under study so that inferences can be drawn from it
- *Statistical inference* is the process of drawing conclusions about the population based on information from a sample

Testing Theories

Hypotheses Competing theories that we want to test about a population are called *Hypotheses* in statistics. Specifically, we label these competing theories as *Null Hypothesis* (H_0) and *Alternative Hypothesis* (H_1 or H_A).

H_0 **:** The null hypothesis is the status quo or the prevailing viewpoint.

H_A **:** The alternative hypothesis is the competing belief. It is the statement that the researcher is hoping to prove.

Example

Taking an aspirin every other day for 20 years can cut your risk of colon cancer nearly in half, a study suggests. According to the American Cancer Society, the lifetime risk of developing colon cancer is 1 in 16.

- H_0:
- H_A:

You Do It 1.2

(*New York Times*, 1/21/1997) Winter can give you a cold because it forces you indoors with coughers, sneezers, and wheezers. Toddlers can give you a cold because they are the original Germs "R" Us. But, can going postal with the boss or fretting about marriage give a person a post-nasal drip?

Yes, say a growing number of researchers. A psychology professor at Carnegie Mellon University, Dr. Sheldon Cohen, said his most recent studies suggest that stress doubles a person's risk of getting a cold.

The percentage of people exposed to a cold virus who actually get a cold is 40%. The researcher would like to assess if stress increases this percentage. So, the population of interest is people who are under stress. State the appropriate hypothesis for assessing the researcher's theory regarding the population.

H_0:
H_A:

Deciding Which Theory to Support

Decision making is based on the "rare event" concept. Since the null hypothesis is the status quo, we assume that it is true unless the observed result is extremely unlikely (rare) under the null hypothesis.

- **Definition**: *If the data were indeed unlikely to be observed under the assumption that H_0 is true, and therefore we reject H_0 in favor of H_A, then we say that the data are **statistically significant**.*

YDI 1.3

Last month a large supermarket chain received many customer complaints about the quantity of chips in a 16-ounce bag of a particular brand of potato chips. Wanting to assure its customers that they were getting their money's worth, the chain decided to test the following hypothesis concerning the true average weight (in ounces) of a bag of such potato chips in the next shipment received from the supplier:

H_0:

H_A

Question

- Suppose you concluded H_A. Could you be wrong in your decision? What if you did not reject H_0? Could you be wrong in your decision?

Errors in Decision Making

In our current justice system, the defendant is presumed innocent until proven guilty. The null and alternative hypothesis that represents this is:

H_0:

H_A:

		Truth	
		H_0	H_A
Your decision based on data	H_0		
	H_A		

Definition

Rejecting the null hypothesis *H*0 when in fact it *is true* is called a **Type I** error. *Accepting* the null hypothesis *H*0 when in fact it *is not true* is called a **Type II** error.

Note: Rejecting the null hypothesis is usually considered the more serious error than accepting it.

Type I and II Errors

α = Type I error

= The chance of rejecting H_0 when in fact H_0 is true

= $P(H_A|H_0)$

β = Type II error

= The chance of accepting H_0 when in fact H_A is true

= $P(H_0|H_A)$

Fall 2006 – Fundamentals of Business Statistics 17

What's in the Bag?

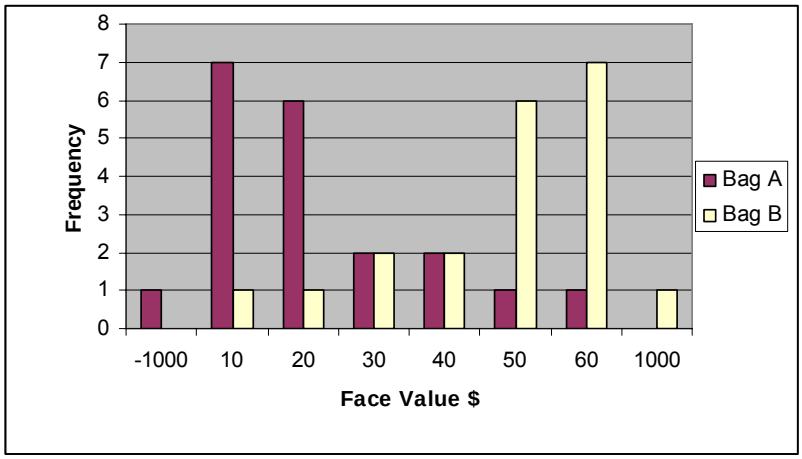
Objective To explore the various aspects of decision making

Problem statement There are two identical looking bags, Bag A and Bag B. Each bag contains 20 vouchers. The contents of the bag, i.e., the face value and the frequency of voucher values, are as follows:

Face Value ($)	Bag A	Bag B
-1000	1	0
10	7	1
20	6	1
30	2	2
40	2	2
50	1	6
60	1	7
1000	0	1
Total	20	20

Fall 2006 – Fundamentals of Business Statistics 18

Frequency Plot



Which bag would you choose?

Game Rules

- The objective is to pick Bag B.
- You will be shown only one of the bags.
- You will be allowed to gather some data from the bag, and based on that information, you must decide whether to take the shown bag (because you think that it is Bag B), or the other bag (because you think that the shown bag is Bag A).
- Initially, the data will consist of selecting just one voucher from the shown bag (without looking into it). In this case, we say that we are taking a sample of size n = 1.

Example (cont.)

H_0 : The shown bag is Bag A

H_A : The shown bag is Bag B

Type I error α =

Type II error β =

Exercise: If the voucher you selected was $60, what would you decide? What if the voucher was $10 instead

Forming a Decision Rule

- What values of the voucher (or in what direction of voucher values) support the alternative hypothesis H_A? That is, what is the direction of extreme?

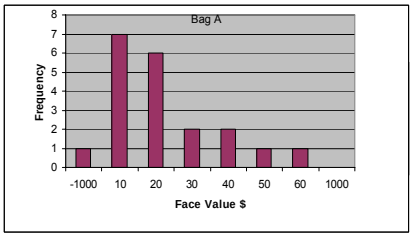
Face Value ($)	Chance if Bag A	Chance if Bag B
-1000	1/20	0
10	7/20	1/20
20	6/20	1/20
30	2/20	2/20
40	2/20	2/20
50	1/20	6/20
60	1/20	7/20
1000	0	1/20

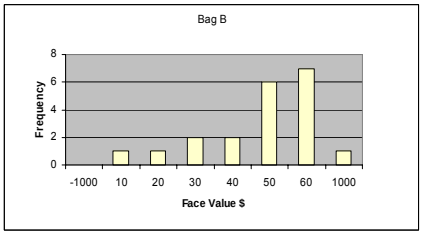
Decision Rule 1

Reject the null hypothesis in favor of the alternative hypothesis if the voucher value is ≥ $50.

Type I error α =

Type II error β =





Summary

Decision Rule Reject H_0 if voucher ≥ $50

Rejection Region $50 or more

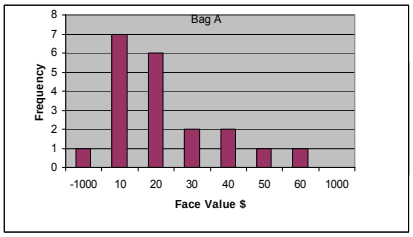
We say ... the cutoff is $50, and larger values are more extreme

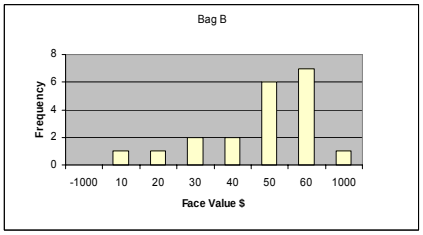
YDI: Decision Rule 2

Reject the null hypothesis in favor of the alternative hypothesis if the voucher value is ≥ $?

Type I error α =

Type II error β =



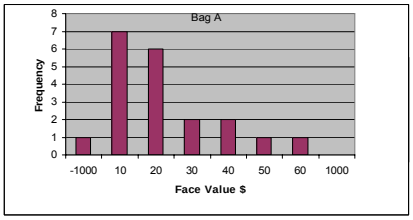


Fall 2006 – Fundamentals of Business Statistics 25

P-Values

Suppose we select a voucher. Assuming that H_0 is true, how likely is it that we would get the observed voucher value, or something more extreme?

Question: What kind of *p-values* support H_A?



Fall 2006 – Fundamentals of Business Statistics 26

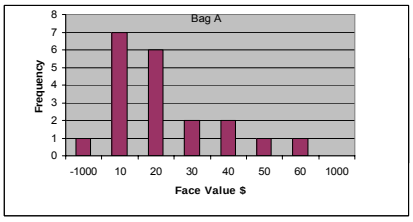
Decision Making and P-Values

Consider our earlier hypothesis:

H_0 : The shown bag is Bag A

H_A : The shown bag is Bag B

Using α=0.10, what is the decision rule?



If we draw a $30 voucher, which hypothesis would you conclude? For this voucher value, can you calculate the *p-value*?

Relationships between α and P-Values

If *p-values* ≤ *α,* Reject the null hypothesis H_0 in favor of the alternative hypothesis H_A

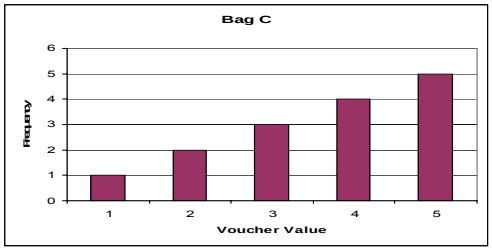
If *p-values* > *α,* Do Not Reject null hypothesis H_0.

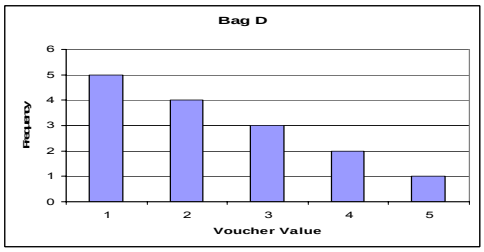
P-Values (continued)

Consider two identical bags C and D with the following distribution of voucher values:

	Bag C		Bag D	
Face Value	Frequency	Chance	Frequency	Chance
1	1	1/15	5	1/3
2	2	2/15	4	4/15
3	3	1/5	3	1/5
4	4	4/15	2	2/15
5	5	1/3	1	1/15

Bag C and D





YDI 1.6

H_0: The shown bag is Bag C

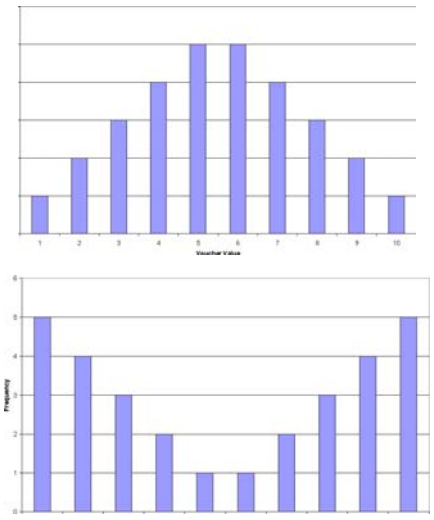
H_A: The shown bag is Bag D

- Suppose the observed voucher (*n=1*) is \$2. What is the *p-value*?

- Would you accept or reject the null hypothesis for the following levels of α = 0.10, 0.05, 0.01

P-Values (cont.)

Consider two identical bags E and F with the following distribution of voucher values:



YDI 1.7

H_0: The shown bag is Bag E
H_A: The shown bag is Bag F

- The decision rule is Reject H_0 if the selected voucher value is ≤ 1 or ≥ 10, then what are α and β?
- Suppose the observed voucher value is $2. What is the *p-value*?
- Would you accept or reject the null hypothesis for the following levels of α = 0. 10, 0. 05, 0. 01.

YDI 1.8

The following table summarizes the results of three studies:

Study A

H_0:The true average lifetime ≥ 54
H_A:The true average lifetime < 54
P-value = 0. 0251

Study B

H_0: The average time to relief for Treatment I is equal to the average time to relief for Treatment II
H_A: The average time to relief for Treatment I is not equal to the average time to relief for Treatment II
P-value = 0. 0018

Study C

H_0:The true proportion of adults who work 2 jobs is ≤ 0. 33
H_A:The true proportion of adults who work 2 jobs is > 0. 33
P-value = 0. 3590

YDI 1.8 (cont.)

1. For which study do the results show the most support for the null hypothesis?
2. Suppose Study A concluded that the data supported the alternative hypothesis that the true average lifetime is less than 54 months, but in fact the true average lifetime is greater than or equal to 54 months. Is this a Type I (α) or Type II (β) error?
3. For each of the three above studies, determine if the rejection region would be on the one-sided left tailed, one-sided right tailed, or two-sided.
   - Study A
   - Study B
   - Study C

Significant versus Important

- With a large enough sample size, even a small difference can be found statistically significant – that is, the difference is hard to explain by chance alone. This does not necessarily make the difference important.
- On the other hand, an important difference may not be statistically significant if the sample size is too small.

Why Sample?

A *Census* is a sample of the entire population

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF MANY YEARS

The Language of Sampling

- A *population* or universe is the total elements of interest for a given problem.
  - Finite population
  - Infinite population
- A *sample* is a part of the population under study selected so that inferences can be drawn from it about the population. Sample sizes are usually represented by *n*.
- *Sampling error (variation)* is the difference between the result obtained from a sample and the result that would be obtained from a census.
- *Parameters* are numerical descriptive measures of populations / processes.
- *Statistics* are numerical descriptive measures computed from the observations in a sample.

YDI 2.1

Exercise *Nine percent of the US population has Type B blood. In a sample of 400 individuals from the US population, 12.5% were found to have Type B blood. Circle your answer:*

- In this particular situation, the value 9% is a (parameter, statistic)
- In this particular situation, the value 12.5% is a (parameter, statistic)

Good Data?

A sampling method is *biased* if it produces results that systematically differ from the truth about the population.

Example Convenience samples and volunteer samples generally lead to biased samples.

Selection bias is the systematic tendency on the part of the sampling procedure to exclude or include a certain part of the population

Nonresponse bias is the distortion that can arise because a large number of units selected for the sample do not respond.

Response bias is the distortion that arises because of the wording of a question or the behavior of the interviewer.

Example

In the election of 1936 the Literary Digest magazine predicted that challenger Alf Landon would beat the incumbent, Franklin Roosevelt. They based their prediction on a survey of ten million citizens taken from lists of car and telephone owners, of whom over 2.3 million responded. This was the largest response to any poll in history, and based on this, the Literary Digest predicted that Landon would win 57% to 43%. In reality, Roosevelt won 62% to 38%. What went wrong? At the same time, a young man known as George Gallup surveyed 50,000 people and correctly predicted that Roosevelt would win the election.

YDI 2.3

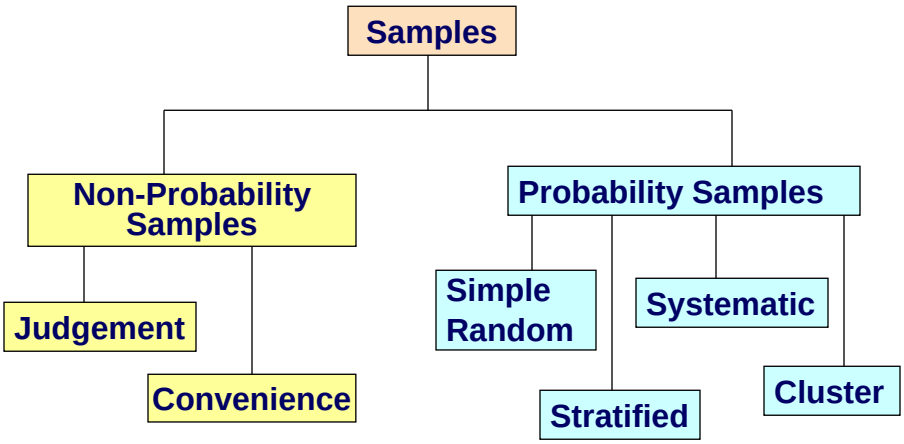
A study was conducted to estimate the average size of households in the US. A total of 1000 people were randomly selected from the population and they were asked to report the number of people in their household. The average of these 1000 responses was found to be 4.6.

1. What is the population of interest?

2. What is the parameter of interest?

3. An average computed in this manner tends to be larger than the true average size of households in the US. True or false? Explain.

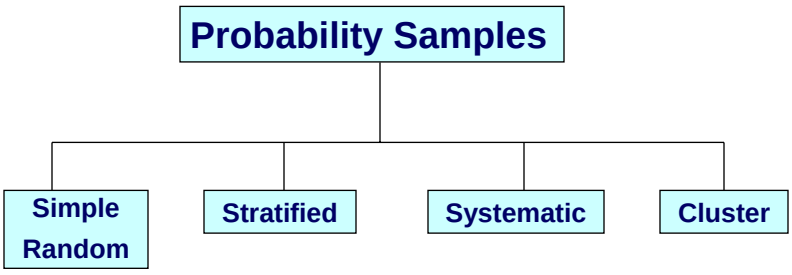
Sampling Techniques



Fall 2006 – Fundamentals of Business Statistics 43

Statistical Sampling

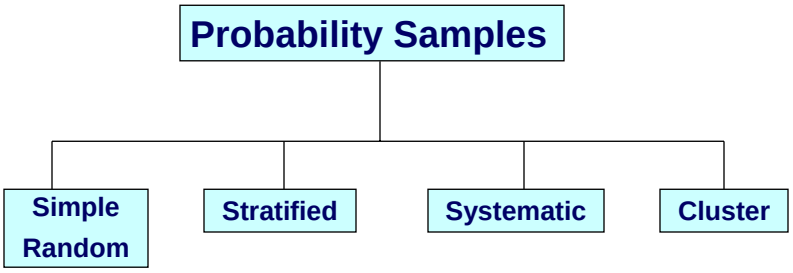
- Items of the sample are chosen based on known or calculable probabilities



Fall 2006 – Fundamentals of Business Statistics 44

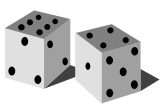
Statistical Sampling

*A sampling method that gives each unit in the population a known, non-zero chance of being selected is called a **probability sampling method** (statistical sampling).*



Simple Random Samples

- Every individual or item from the population has an equal chance of being selected



Stratified Samples

*A **stratified random sample** is selected by dividing the population into **mutually exclusive** subgroups, and then taking a simple random sample from each subgroup. The simple random samples are then combined to give the full sample.*

- allows us to obtain information about each Subgroup
- can be more efficient than simple random sampling

Example

Systematic Samples

For a **1-in-k systematic sample**, you order the units of the population in some way and randomly select one of the first k units in the ordered list. This selected unit is the first unit to be included in the sample. You continue through the list selecting every kth unit from then on.

- Convenient
- Fast
- Could be biased

Cluster Samples

In **cluster sampling**, the units of the population are grouped into clusters. One or more clusters are then selected at random. If a cluster is selected, that all units of that cluster are part of the sample.

Think about it

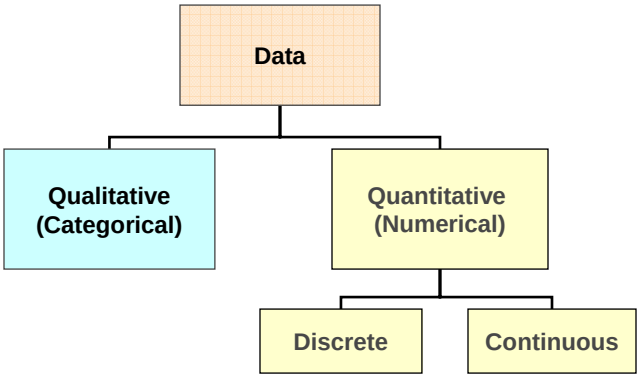
- Is a cluster sample a simple random sample?
- Is a cluster sample a stratified random sample?
- Were you to form clusters, how should the variability of the units within each cluster compare to the variability between the clusters?
- Is this criterion the same as in stratified random sampling?

YDI 2.13

Identify the sampling method for each of the following scenarios:

1. A shipment of 1000 3 oz. bottles of cologne has arrived to a merchant. These bottles were shipped together in 50 boxes with 20 bottles in each box. Of the 50 boxes, 5 boxes were randomly selected. The average content for these 100 bottles was obtained.
2. A faculty member wishes to take a sample from the 1600 students in the school. Each student has an ID number. A list of ID numbers is available. The faculty member selects an ID number at random from the first 16 ID numbers in the list, and then every sixteenth number on the list from then on.
3. A faculty member wishes to take a sample from the 1600 students in the school. The faculty member decides to interview the first 100 students entering her class next Monday morning.

Data Types



Data Types

- **Time Series Data**
  - Ordered data values observed over time
- **Cross Section Data**
  - Data values observed at a fixed point in time

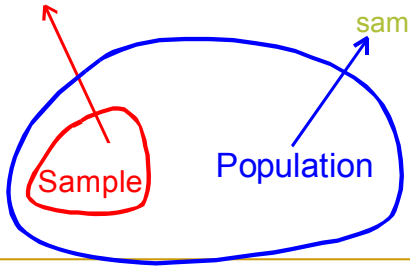
Key Definitions

- A **population** is the entire collection of things under consideration
  - A **parameter** is a summary measure computed to describe a characteristic of the population
- A **sample** is a portion of the population selected for analysis
  - A **statistic** is a summary measure computed to describe a characteristic of the sample

Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics (known) **Inference** Population parameters (unknown, but can be estimated from sample evidence)



Fall 2006 – Fundamentals of Business Statistics 55