**Chapter 7 Data for Decisions-part I *— Sample Survey and Experiment***

Statistics is the science and art dealing with techniques for

I organizing data (Chapters 5 and 6)

— Graphical presentation of data

— Numerical summary statistics

II collecting data through (Sections 7.1 - 7.6)

— sampling surveys

— experiments

*We need to get data first before we can work on it*.

III utilizing sample data to make (Chapter 6, Sections 7.7 - 7.8)

— estimations, predictions

— generalizations about a large population.

**1. Population and sample**

**• *The population in a statistical study is the entire group of individuals about which we want information*.**

**• *A sample is a part of the population from which we actually collect information used to draw conclusions about the whole population.***

***• The sample size is the number of individuals in a sample, denoted as n.***

Example 1. A sociologist hypothesizes that the average annual income of households in a large city is less than $40,000 per year. To test her hypothesis, she samples 500 households in the city and determines the income of each.

(a) Describe the population.

(b) Describe the sample.

(c) What is the sample size?

Example 2. In Oct 2009 the unemployment was 10.1%, the highest since January of 2000. How is unemployment rate obtained?

Answer: Bureau of Labor Statistics (http://www.bls.gov) estimates monthly unemployment rate by a sample of 50,000 households each month.

(a) Describe the population.

Answer: *Civilian labor force: (i) 16 or older; (ii) look for work; (iii) not in prison*

(b) Describe the sample.

(c) How is the unemployment rate calculated?

Answer:

**2. Biased Sampling Methods**

Since we will utilize sample data to make a decision, we need to minimize the bias in the sampling.

***The design of a statistical study is biased if it systematically favors certain outcomes****.*

Example 3. **Convenience Sampling**: I wish to know students’ opinion about cafeteria food service and I survey at an entrance of IUSB cafeteria. Is my sample possibly biased? Explain.

Example 4. Online and call-in polls.

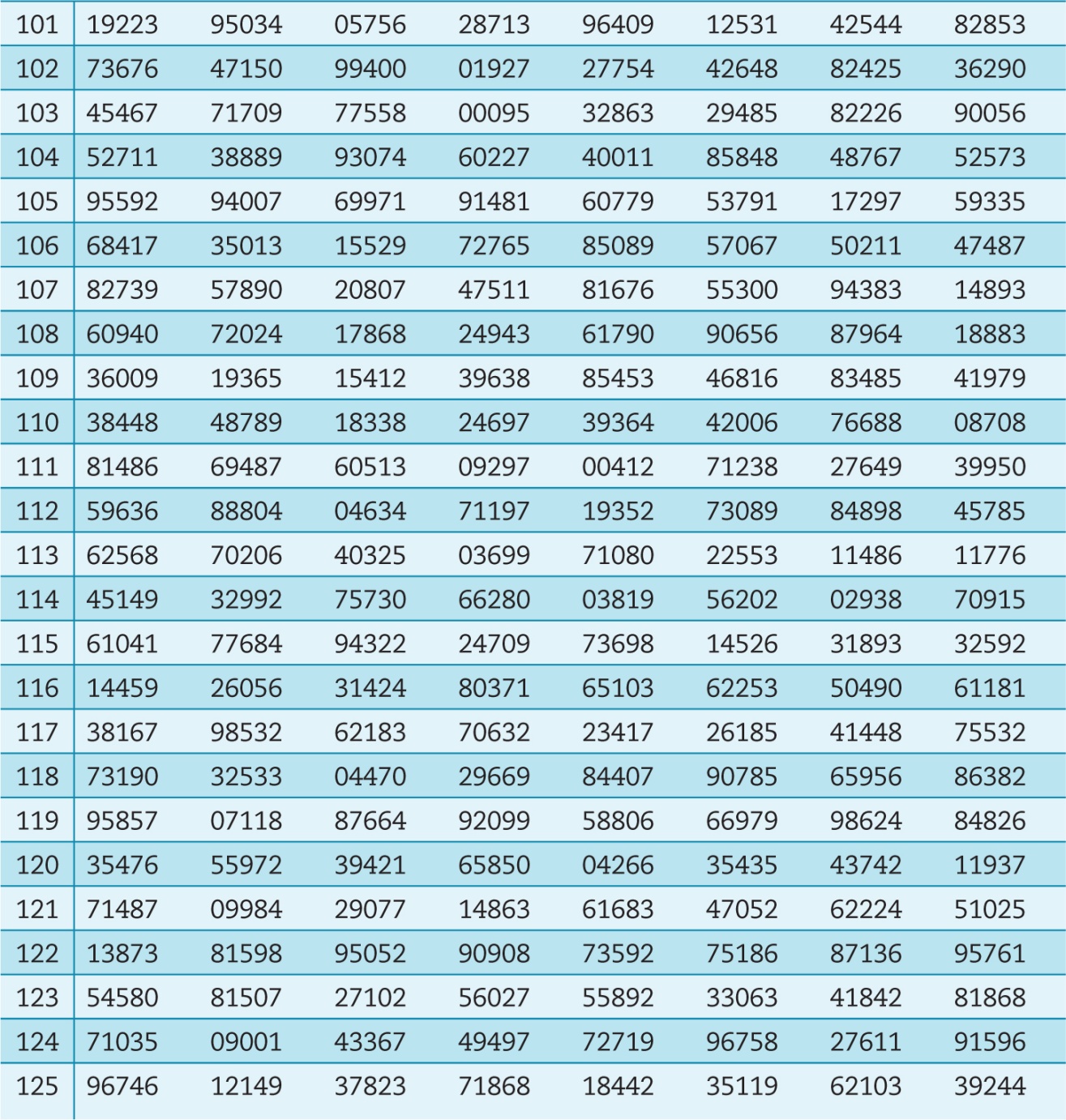
Online and call-in polls are particularly at risk of this error, because the respondents are self-selected. At best, this means the people who care most about an issue will answer; at worst, people listening to a particular radio host, or on a political mailing list, flood the poll.

***A voluntary response sample consists of people who choose themselves by responding to a general appeal. They are biased because people with strong opinions are most likely to respond***.

**3. Simple Random Samples (*SRS*)**

***A SRS of size n is to select n individuals from the population such that every set of n individuals has an equal chance to be selected.***

**Random Digits Table**: Table 7.1 on page 246:



(a) Each entry is equally likely to be one of: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

(b) The entries are independent of each other.

**Procedure of SRS using Table 7.1**:

**Step I**. Label all individuals in the population numerically.

**Step II**. Use Table 7.1 to select *n* individuals.

Example 5. Simple random sample five students from this class using

Table 7.1 and following the above procedure.

**4. Cautions about sample surveys**

• **Under coverage**: occurs when some groups in the population are left out of the process of choosing the sample.

• **Nonresponse**: occurs when an individual chosen for the sample cannot be contacted or refuses to participate.

Example 6 – Under coverage. We used a telephone book to randomly choose numbers to dial and ask “What brand of soap do you use most often?

— Population: All US adults

— Sampling Frame: All adults with listed phone numbers

— Under coverage: By using the telephone book, we have left out all those people who do not have land phones and all the people who have unlisted phone numbers.

What may cause sampling errors?

— Under coverage

— Nonresponse.

*Different groups have different rates of nonresponse.*

*Even if a sampling plan is unbiased, nonresponses could cause bias.*

— Wording

Example 7 - Wording of question.

• Question 1: “Should laws be passed to eliminate all possibilities of special interests giving huge sums of money to candidates?

• Question 2: “Should laws be passed to prohibit interest groups from contributing to campaigns, or do groups have a right to contribute to the candidate they support?

• The result: 80% Yes for question 1, while 40% Yes for question 2.

***The exact wording of the question influences the answers that are given. Small changes in how a question is worded can make a big difference in the response***.

**5. Experiments vs observational studies**

Example 8. A new drug is introduced. How should an experiment be designed to test its effectiveness?

• Treatment group: Subjects are treated with the new drug.

• Control group: Subjects are treated with the standard drug, or are not treated.

• Then the responses of two groups are compared.

***An experiment deliberately imposes some “treatment”***

***on individuals in order to observe their responses.***

***The purpose of an experiment is to study whether***

***the treatment causes a change, or has an effect,***

***in the response.***

**Attention**: An opinion survey is not an experiment. There is no “treatment” and the investigators do not attempt to influence the responses***.***

Example 9. A study of Americans aged 40 to 79 found that the lung cancer death rate was 11 times higher among smokers than among nonsmokers.

This is an observational study. The different groups are formed by the different lifestyles of subjects. The investigators just watch/observe what happens.

***An observational study, such as sample survey, observes individuals and measures variables of interest but does not attempt to influence the responses.The purpose of an observational study is to describe some group or situation.***

**Caution**: Observational studies can establish association; one thing is linked to another. But association is not necessarily causation.

The investigators should always think and judge:

How can one be sure that the comparison is fair?

Example 10. A university normally offers a test preparation course for those who plan to take Management Admission Test. Last semester, they offered the option of an on-line course as an experiment. The test result shows that students who took the on-line session scored 10% higher on average than that scored by the students from the in-class sessions. Can we conclude that the on-line teaching is better than the in-class teaching? This is an experiment with on-line teaching as a treatment. However, it is *an uncontrolled experiment,* because there is no control group to make fair comparison.

Further study found that

• on-line students are older on average,

• more on-line students are employed, and

• among those employed, more on-line students had management experience.

***Variables, whether part of a study or not, are said to be confounded when their effects on the outcome cannot be distinguished from one another.***

***The remedy for confounding is to do a randomized comparative experiment.***

**6. Randomized Comparative Experiments (Groups)**

• Subjects should be assigned to treatment or control groups at random.

– *Advantage*: The randomized comparative experimental study seeks to identify cause-and-effect relationships.

– *Disadvantage*: The randomized comparative experimental studies are not always feasible. (Can you assign people to, or not to, smoke?)

Example 11. Design a randomized comparative experiment to evaluate a new mathematics curriculum with 200 available students.

**7. Thinking about experiments**

• The experiment should be run double blind and placebo should be used when they can be applied.

***An experiment is called double-blind if neither the subjects nor the persons who interact with them know which treatment each subject received.***

***The effect of a dummy treatment on the response of subjects is called placebo effect.***

Example 12. In a study, 1000 patients who suffer chronicle pain participated in a clinical trial, which tests the effectiveness of new painkiller drug.

• 1000 patients are randomly divided into two groups, 500 in each.

• Each patient in the treatment group receives the new drug injection. Each patients in the control group receives placebo, which is saline solution (simple salt water) injections

• The trial was double-blind.

• Many patients in control group responded that their pain situation is improved.

• If the reports from two groups are similar regarding the pain relief, we cannot conclude that the new drug is effective.

• Only if the result from treatment group is significantly better that the result from control group, we can conclude that the new drug is effective.

Example 9. (Continue) A study of Americans aged 40 to 79 found that the lung cancer death rate was 11 times higher among smokers than among nonsmokers.

The connection between smoking and lung cancer is statistically significant.

Usually, an observational study cannot tell us what factors are at work.

But the statistical evidence that points to smoking as a cause of lung cancer is about as strong as nonexperimental evidence can be:

(a) Similar results are observed in many countries.

(b) Cigarette smoke contains tars that can be shown to cause tumors in animals.

(c) Studies in each race, gender, age group, working environment, etc, all show the significance.

(d) No plausible alternative explanation is available.

***Study Review Vocabulary on page 267-268***.

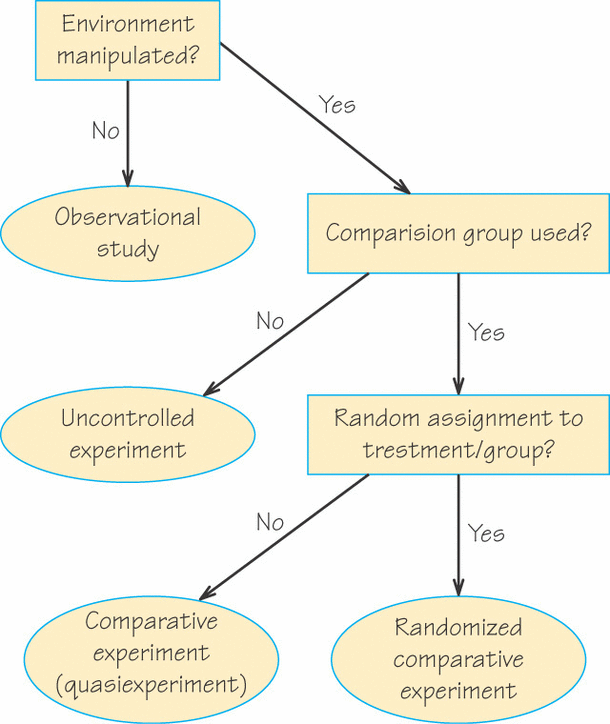
**Chapter 7 *Statistical Inference: From Sample to Population- part II***

Main topics: • Parameters and statistics

• Point estimate for a population proportion.

• Margin of error and confidence interval for a population proportion

The flowchart below gives a conceptual overview of the types of designs we have covered in these last two sections.



Example 1. (CNN) — 43% of 1,033 adults surveyed by OpinionResearch Corporation said they think the country is safer now than it was before the 9/11 attacks. 32% said they believe it is about as safe, while 25% said it is less safe.

The telephone poll was conducted August 18-20, 2006. The margin of error is plus or minus 3 percentage points.

Critical thinking: • Can we make an inference that the percent of people who think the country is safer now than it was before the 9/11 attack in our whole nation is also about 43%?

• What does 3 percentage points margin of error mean?

**1. Parameters and statistics**

• ***A number that describes a population is called a parameter (such as population proportion).*** A parameter is a fixed number, but in practice we usually do not know its value.

• ***A number that describes a sample is called a statistic (such as sample proportion).*** The value of a statistic is known, based on data from survey, or experiment, or observational study. But it changes from sample to sample.

Example 1. (Cont.) (CNN) — 43% of 1,033 adults surveyed by Opinion Research Corporation said they think the country is safer now than it was before the 9/11 attacks. The margin of error is 3 percentage points.

(a) Describe the population, sample, and sample size in the study.

The population = The sample = The sample size =

(b) Describe the parameter and statistic in the study.

The parameter = The statistic =

• The population proportion is denoted as *p*. The value of *p* is typically unknown.

• A sample mean is denoted as, read as “*p* hat”, which can be computed from survey results.

**2. Sampling distribution of a sample proportion**

Example 2.

• A sample of 100 adults was surveyed and the sample percentage of people who think the country is safer now than it was before the 9/11 attacks were computed.

• Another 100 adults were sampled and the sample percentage of people who think the country is safer now than it was before the 9/11 attacks were computed as well.

• Repeat this process for 36 times and therefore 36 sample percentages were computed and listed below: 35 48 48 47 36 36 42 46 38 44 42 42

38 42 47 43 43 43 47 50 42 41 50 49

36 42 50 44 44 45 39 41 39 47 44 38



• The median of the above 36 sample percentages with sample size 100 for each is 43%.

• Now, repeat the sampling process for 36 times but the sample size is 1033 for each sample this time. The following are the results of 36 sample percentages

41 43 43 43 42 42 41 43 43 40 43 45

42 44 45 44 44 41 44 44 44 41 41 43

44 43 43 42 44 43 43 46 40 41 42 44



• The median of the above 36 sample percentages with sample size 1033 for each is also 43%.

***The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population***

The above histogram shows a sampling distribution of a sample percentage with sample size

n = 1033.

• What can we observe based on these two histograms?

– Shape: Both histograms look normal.

– Center: Two histograms have about the same center.

– Spread: The histogram with larger sample size for each sample has less spread

***The sampling distribution of a sample percentage***

• Choosing a simple random sample of size *n* from a large population with population proportion of p of “success”. Let be the sample proportion,

(a) The sampling distribution of is approximately normal when the sample size *n* is large.

(n > 30)

(b) The mean/center of the sampling distribution is the population proportion *p*.

(c) The standard deviation of the sampling distribution is

Example 3. Assume that the population proportion is 0.43 (43%) and sample size *n* is 1033.

(a) Find the mean of the sampling distribution of sample proportion.

(b) Find the standard deviation of the sampling distribution of sample proportion. Round to 4 decimal places.

(c) In what range of values do the 95% of sample proportions fall? Draw the Normal curve. (Recall 68%-95%-99.7% rule). Round to 3 decimal places.

**3. Confidence intervals and margin of errors**

• The sample percentage gives a point estimate of population percentage p.

• A 95% confidence interval for *p* is

The formula is based on 68-95-99.7 rule.

• Furthermore, since sample proportion is different from sample to sample, we like to measure the variation.

is called **the margin of error**.

Workout 1. A recent FOX News poll finds that most people (756 out of 900) think a college education is more important to succeeding in life today than it was 25 years ago. This belief is consistent among demographic groups: men and women, young and old, high and low income, white and non-white all agree that having a college degree is vital these days.

Opinion Dynamics Corporation conducted a random sample of 900 registered voters for FOX News from June 13 to June 14, 2013.

Find a 95% confidence interval for the population proportion of people who think a college education is more important to succeeding in life today than it was 25 years ago.

Step 1. Compute the sample proportion.

Step 2. Compute the standard deviation.

Step 3. Compute the margin of error.

Step 4. Compute 95% confidence interval.

**Understand a confidence interval:**

A 95% confidence interval is an interval obtained from the sample data by a method that in 95% of all samples will produce an interval containing the true population parameter.

• If you draw one sample of size 900, you can compute one 95% confidence interval.

• If you draw another sample of size 900, you can compute another 95% confidence interval.

• If you draw 100 samples of size 900, for each sample you may compute a 95% confidence interval. Totally, you have 100 confidence intervals.

• Among 100 confidence intervals, about 95 intervals will contain the true population proportion *p* and about 5 intervals will miss it.

• Interpretation of a 95% confidence interval:

*We are 95% confident that the interval contains the true population proportion.*

• The margin of error measures the sampling variation of the simple random sample plan, because the sample percentage will vary from sample to sample.

• If a sample plan is biased, the error due to the bias is not included in the margin of error, and should be evaluated in a case by case way.

How does the sample size affect the margin of error and confidence interval?

Redo Workout 1 using a sample size of a) 400 b) 100

sample size std deviation margin of error 95% confidence interval

n = 900 .012 .012 0.816 to 0.864

n = 400 = .018 2(.018)= .036 .84+/-.036 =

n = 100

When sample size decreases, the margin of error \_\_\_\_\_\_\_\_\_\_\_\_\_.

When the margin of error increases, the confidence interval \_\_\_\_\_\_\_\_\_.

Workout 2. A recent Gallup poll shows that the plurality of Americans, 33%, think neither the Republicans nor the Democrats have a clear plan on Iraq. The result is based on interviews with a randomly selected national sample of 2,003 adults, aged 18 and older.

(a) Find the sample proportion of Americans who think neither the Republicans nor the Democrats have a clear plan on Iraq.

(b) Compute the standard deviation.

(c) Compute the margin of error.

(d) Compute a 95% confidence interval.

(e) Interpret the confidence interval: We are 95% confident that the percentage of American adults who think neither the Republicans nor the Democrats have a clear plan on Iraq is between \_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_. You can also use interval notation:

(f) (Critical thinking) if you increased sample size to 8000, and assume that the sample percentage is still 33%, how is the margin of error going to change? How is the width of the 95% confidence interval going to change?

Figure 7.8 A collection of 25 samples from the same population give these 95% confidence intervals. The red dot of each interval is its point estimate of the proportion. Note that 24 of these 25 intervals (i.e., 96%) cover or contain the true population proportion, marked by the vertical line. If this process were done many, many times, 95% of all 95% confidence intervals would cover or contain the true population proportion.

