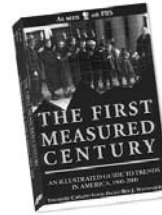


Introduction: Statistics, Data and Statistical Thinking

FREC 408
Dr. Tom Ilvento
213 Townsend Hall
ilvento@udel.edu
<http://www.udel.edu/FREC/ilvento>

The First Measured Century



<http://www.pbs.org/fmc/index.htm>

Statistics

- **Statistics** (Def 1.1 p24) is the science of data
- It refers to
 - Collecting data
 - Classifying, summarizing, and organizing data
 - Analysis of data
 - Interpretation of data

Statistics

- Statistics is both a field of study
- ...and a set of tools used by many disciplines
 - Social Sciences
 - Biological Sciences
 - Physical Sciences

We will focus on two types of statistical applications

- Descriptive
- Inferential

Descriptive Statistics

- Descriptive statistics uses summary measures, graphs, and measures of association to show relationships in data.
- The focus is on describing the data
- With an emphasis on **parsimony**

Descriptive Statistics

- Rather than looking at a set of numbers,
- 0, 0, 2, 2, 3, 3, 3, 4, 5, 2, 1, 3, 2, 2, 1, 1, 3, 1, 1, 2, 5, 7, 8, 10, 12

Descriptive Statistics

- we want to find summary measures which describe the data adequately and succinctly
- Be they a
 - Percentage
 - Average
 - Range from highest to lowest
 - mode

Descriptive Statistics

- Descriptive Statistics also involve relationships between variables or sets of variables
- And they can involve very sophisticated techniques – regression, principle components, factor analysis, Logistic Regression, Probit Analysis

Inferential Statistics

- Inferential statistics takes it a step further
- Now we use some of the same techniques to make estimates, decisions, predictions, or generalizations about a population from a smaller subset or sample

Inferential Statistics

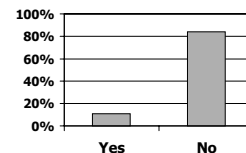
- Inferential statistics are a powerful tool for research
- It enables us to make statements about a large group from a much smaller sample.
 - We can survey 1,000 people and make statements about 280 million people

Did the public care if George W. Bush used cocaine in his 20s?

A Time/CNN Poll found:

If Bush did use cocaine in his 20s, should that disqualify him from being President?

Yes 11%
No 84%



Let's look closer at this survey example

- It was based on a telephone poll of 942 adult Americans taken for Time/CNN on August 19th by Yankelovich Partners, Inc.
- **The sampling error is $\pm 3.3\%$**
- **What does this mean?**

Here's my interpretation

- The survey is designed to represent adult Americans in August of 1999
- Because we are taking a sample, we have some error associated with our estimate.

Here's my interpretation

- Since the sample was taken randomly, we have a method to estimate the error of our estimate
- In this case, we are reasonably sure that the true percentage is within $\pm 3.3\%$ points of our estimate
- Which means our interval is 7.7% to 14.3%

We need some terms

- A **Population** (Def 1.6 P29) is the total number of units involved in the research question. The units are the members (or elements) of the population.
- Populations could be:
 - People
 - Animals
 - Plants
 - Courses
 - Objects

A POPULATION IS DEFINED BY

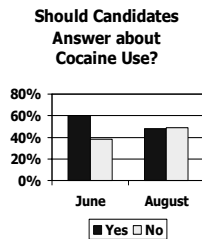
- Purpose of the study
- The units and elements involved
- Geographic coverage
- Time frame

Population Example

- If I was interested in understanding current household consumption of chicken in the Mid-Atlantic states, I might define the population as:
 - **All households in in the Mid-Atlantic states (DE, MD, PA, NJ, NY) in the Fall of 2002**

Does time matter for a population?

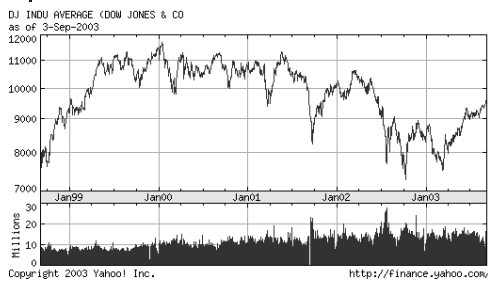
- The Time/CNN poll asked: *Should a candidate have to answer questions about whether he used cocaine in the past?*
- June 60% Yes
- August 48% Yes



The DOW over a One-Year Period (October 2002 to Sept. 2003)



The DOW over a Five-Year Period



Sampling

- When we collect data on all elements in a population, we take a census
- However, sometimes it is difficult to get information on the entire population
- So we take a sample of the population
- A **sample** (Def 1.8 P30) is a subset of the units or elements of a population

Why Sample?

- It saves time
- Money
- Other resources (computation time)
- It may actually be impossible to collect information on everyone
 - Every corn stalk in a field
 - Every dog who suffers from heart worm

Recent Census Debate

- Every 10 years we take a census
- It is mandated in the Constitution
- However, the Census Bureau knows that it doesn't get a complete count - some groups are difficult to contact
- So, the Census Bureau wants to take a really good sample to estimate the undercount, and then adjust the counts to reflect the missing people

More on Sample

- Samples are also defined in the terms we used for populations
 - purpose of the research,
 - the units and elements involved,
 - the geographic coverage, and
 - the time frame

More on Sampling

- A valuable property of a sample is that is representative of the population.
- The sample characteristics resemble those possessed by the population
- Inferential statistics require a sample to be representative of the population,
- And that can be done through a random process

More Terms

- A **random sample** (Def 1.11 p41) is when each element or unit has the same chance of being selected
 - If we select a random sample of 1,000 from a population of 100,000,
 - Each unit has a 1,000/100,000 or
 - 1/100th chance of being selected

More Terms

- A **variable** (Def 1.3 p24) is a characteristic of an individual unit of the population
- To be a variable the characteristics must **vary**
 - **It can't all be the same;**
 - **otherwise, it's a constant.**

More Terms

- Measurement is the process of assigning a number to variables of the individual units
 - Some measurement seems relatively straight-forward
 - years of age, dollars of income, cholesterol counts, parts per million of a chemical

Measurement

- Other concepts are more difficult to measure
 - Attitudes
 - Emotions
 - Intelligence
 - *LOVE*

Measurement

- The process of measurement is often complex – don't take it for granted
- It always comes with some error
- And perhaps Bias

Measurement

- With measurement we must also deal with
 - Validity – are we measuring what we think we are measuring
 - Reliability – is the measuring device consistent

Types of Data - the Book

- **Quantitative data** (Def 1.4 p24) are measures that are recorded on a naturally occurring scale
- **Qualitative data** (Def 1.5 p24) does not follow in natural numerical scale and thus are classified into categories

Types of Data

- I will use a more elaborate description of levels of measurement
 - Nominal
 - Ordinal
 - Continuous

Levels of Measurement

- **Nominal** (or categorical) – no implied order or superiority
 - Men and Women
 - Race
 - Species or genuses

Levels of Measurement

- **Ordinal** – an implied order or rank, but the distance between units is not well specified
 - Ranking
 - Strongly agree to Strongly disagree
 - On a scale from one to ten..

Levels of Measurement

- **Continuous** (combination of interval and ratio) – data that is measured on a scale where we can say something about the magnitude between numbers
 - Age
 - Income
 - Years of School

Why consider our level of measurement?

- Because our statistical techniques are predicated on certain levels of measurement.
- Each technique/formula assumes a certain level is used.
- Misusing a statistical technique on a variable can lead to results that are biased or misleading.

Sources of Data

- **Data from a published source** – also known as existing data. Someone else collected it and makes it available to you
 - Census of Population
 - Current Population Survey
 - Sports statistics
- **Caution** – data decisions are out of your control

Sources of Data

- **A designed Experiment** where the researcher has strict control over the units (people, objects and events).
 - Treatment and Control Groups
 - Randomized designs
- An experimental design allows you to control more factors and to extract more information from the data

Sources of Data

- **Surveys** are where a researcher samples a group of people, asks a set of questions, and records the answers
 - Face-to-Face
 - Telephone
 - Mail
 - Internet
- Social Surveys are **extremely popular** today

Sources of Data

- **Observational Studies** are when the researcher observes the units in their natural setting and records the variables of interest.
 - Animal studies in natural habitats
 - Studies of children's behaviors
- Observational Studies must deal with a number of methodological issues

Shere Hite Report Example

- Shere Hite began her work in 1968 on permissive sexual attitudes in the U.S.
- Her work tended to be controversial, not only for her topics, but because of her methods of collecting and analyzing data
- A second report was even more controversial in 1988, *Women and Love: A Cultural Revolution in Progress*

Key findings from Hite's 1988 book

- 84% of woman were not emotionally satisfied with their relationship
- 95% reported emotional and psychological harassment from their partners
- 70% of women married for 5 years or more were having extra-marital affairs
- Only 13% of women married for more than two years were in love.

Shere Hite Survey Methodology

- Her survey was a mail survey:
 - Mailed to 100,000 women in the U.S. over 7 years
 - The mailing list was a combination from a wide variety of organizations which were asked to circulate them to members. The groups tended to over-represent feminist groups and women in troubled circumstances
 - Approximately 4,500 people responded, a 4.5% response rate.

Shere Hite Survey

- Hite's survey used 127 open-ended questions
 - The instructions read: *It is not necessary to answer every question! Feel free to skip around and answer those questions that you choose.*
- The questions involved a complex set of issues with sub-questions and follow-ups

Statistical Critiques

- Sample was not random or representative of the population of all women
- Low response rate reflected a bias towards those most angry or eager to answer the survey
- Encouraging skipping questions would also lead to bias
- Open ended questions are often difficult to summarize

Critical Thinking and Statistics

- Statistics involves making critical decisions and rational thought to how a set of data are:
 - Sampled
 - Measured
 - Collected
 - Analyzed
 - Interpreted