

Class Notes

- Class Notes are available at: <http://www.math.umt.edu/steele/Math444>

They are organized by Chapter. The best way to use them is to print them before class, read them, and bring them to class. Then, you can concentrate on the what is being said, rather than taking notes. **Do not** use them in place of going to class.

- MATH 117 is a prerequisite for the class. You need the material on probability.

Homework

- Read §1.1-1.3, 2.1-2.4
- Turn in on Friday, problems 1.2, 1.5 (p. 14), 2.2, 2.4, and 2.10 (p. 26)

Chapter 1: Introduction to Statistics

Statistics is the science of extrapolating from a sample, or an experiment, to a population, or process. Statistical methods are used extract information from a sample that applies to a larger population

Example 1 From Roberts, H.V. 1979. "Harris Trust and Savings Bank: An Analysis of Employee Compensation," Report 7946, Center for Mathematical Studies in Business and Economics, University of Chicago Graduate School of Business.

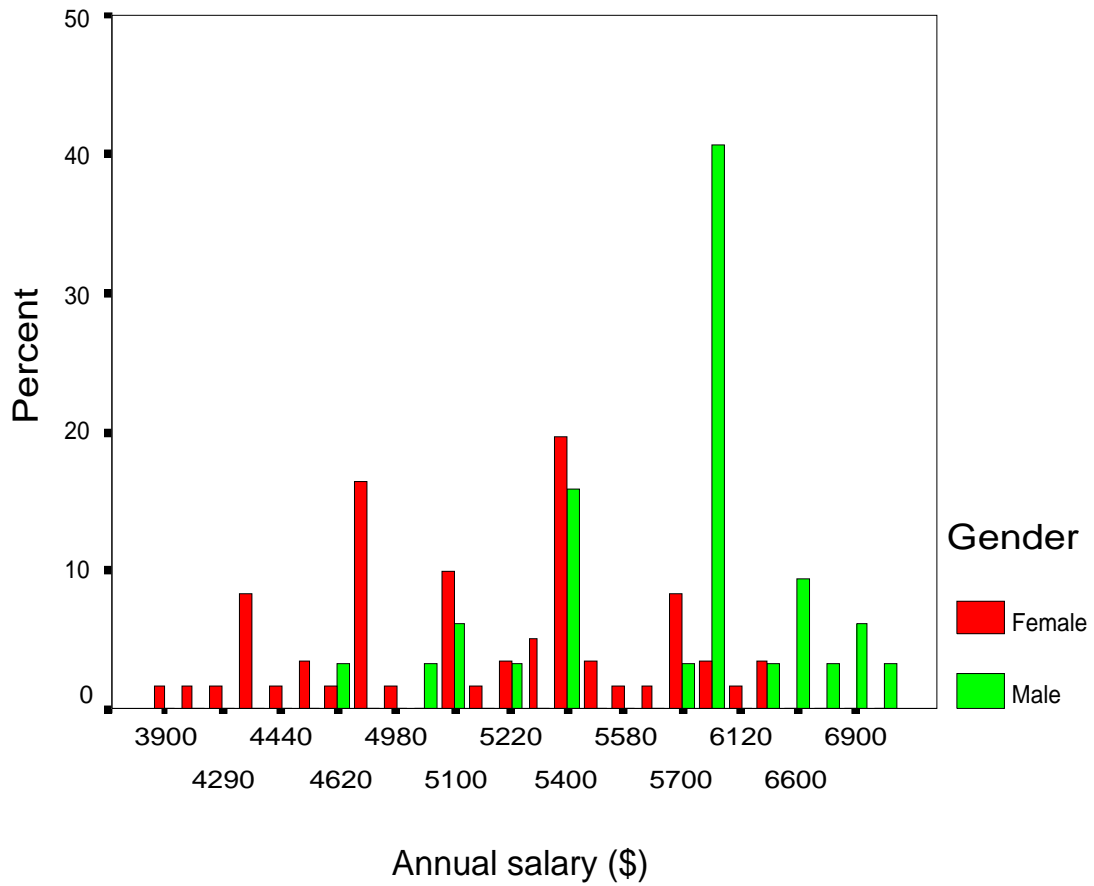
- A lawsuit was filed against the Harris Trust and Savings Bank claiming that as a group, new male employees received larger starting salaries than new women employees. Data are beginning salaries for 32 male and 61 female skilled, entry-level clerical workers hired by the bank from 1969 to 1977.

Statistical summaries:

- The average salary for males was \$5956 and the standard deviation was \$540
- The average salary for females was \$5139 and the standard deviation was \$691
- Are these differences indicative of differences among all employees? Salaries are highly variable; perhaps it is just by random chance that there is a \$817 difference in average salary in favor of the males

Graphical Statistics are used to give visual summaries of the data.

Figure 1. Bar chart showing male and female starting salaries of skilled, entry-level clerical workers hired by the bank from 1969 to 1977. ($n_m = 32$, $n_f = 61$)



Question: Do the data show evidence of differences?

- What do we mean? Only among the 93 salaries? Among all clerical workers at the bank, during that time span? What about other factors: experience, education, age?
- Going from the data, or sample, to a population is the process of drawing inferences
- An inference is a statement about a population which is drawn from a sample
- A statistical inference is an inference justified by a probability model linking the data to the broader context
- If the probability model is a good approximation of the true mechanism or process generating the data, then we can obtain a reliable inference, i.e., one that is *likely* to be correct.
- The scope of inference is the population which our conclusion are applicable. What is the scope of inference in the sex discrimination study?

Statistical Methods are used to

1. minimize error in extrapolating from the sample to the population
2. quantify the uncertainty in our extrapolations
 - A population is the set of all objects, units or individuals of interest. E.g., all clerical workers at the bank
 - A sample is *any* subset of a population. E.g., skilled, entry-level clerical workers hired by the bank from 1969 to 1977
 - A variable is an attribute of the population units (e.g., beginning salary)
 - Statistical inference is carried out by
 1. Gathering data: either by sampling or conducting an experiment
 2. Summarizing the data using graphs and summary statistics E.g., sample means and sample standard deviation
 3. Analyzing the data, usually using more sophisticated methods, often involving a probability model. E.g., two-sample *t*-test
 4. Reporting the results. Lack of effective communication of statistics is a major impediment to scientific success.

Chapter 2: Using Surveys, Scientific and Observational Studies to Collect Data

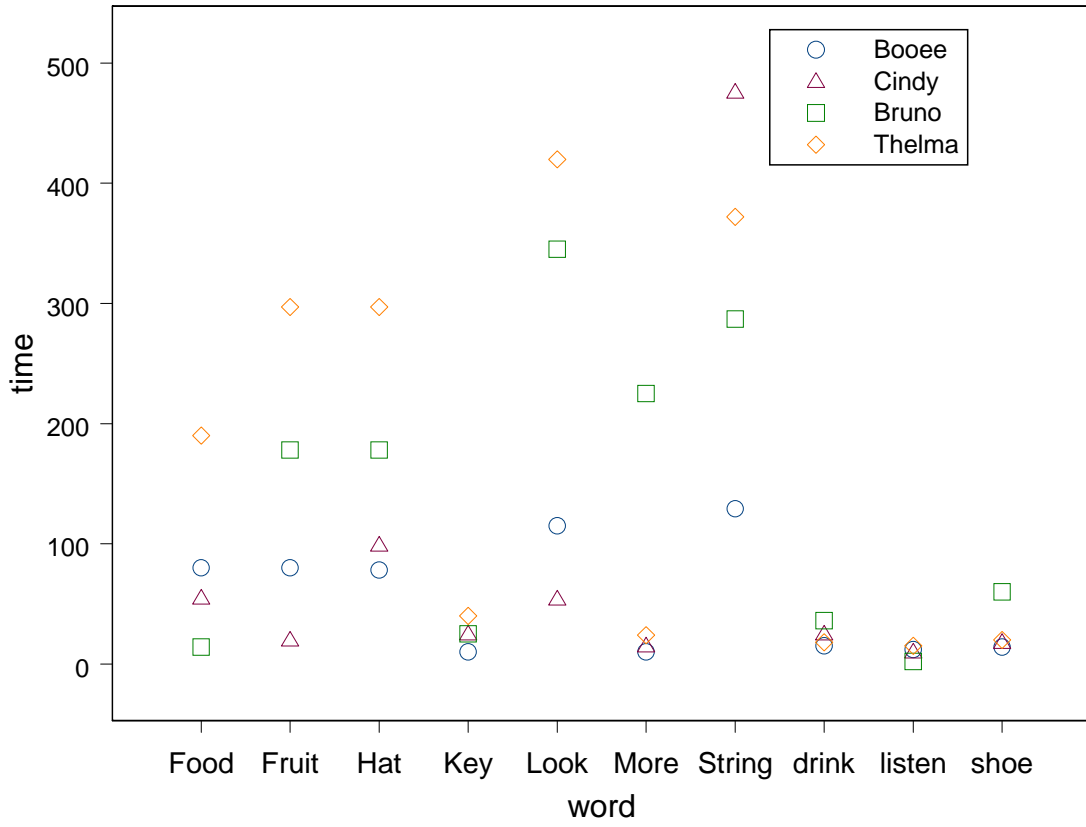
- This is a very brief introduction to the key ingredients of data collection. Recall that statistical inference is the process of extrapolating from a sample to a population using statistics.
- Statistical methods are used to minimize and quantify error
- The key ingredients for good data collection are
 1. clear objectives
 2. one or more variables (or attributes of interest)
 3. an appropriate design for collecting data
 4. successful data collection

- **Case Study:** Chimpanzee learning times.
 - **References:** Fouts, R.S. 1983. "Acquisition and testing of gestural signs in four young chimpanzees," *Science*, **180**, 978-980, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 398.
 - Fouts taught 4 young chimpanzees 10 signs of the American sign language with the intent of determining whether some signs are easier to learn, and whether some chimps tended to learn more quickly than others.
1. Objectives: determine whether some signs are easier to learn, and determine whether some chimps tended to learn more quickly than others.
 2. Variable of interest: time (minutes) to learn a sign
 3. Design: individually teach each of four (2 male, 2 female) chimps 10 words
 4. Data collection They used a system of rewards and threats to get the animals to make a sign. When a chimp produced the sign 5 consecutive times without prompting by the trainer, the word was declared to have been learned

- Possible problems: the four chimps are not necessarily representative of any well-defined population (e.g., adult chimps raised in captivity). Differences (among learning times) might be attributable to different teachers, rather than differences among chimps, or words. Teachers may have learnt how to teach better as more words were taught. Differences might be attributable to the order in which the words were taught.
- Data (time in minutes to learn a word):

Word	Chimpanzee			
	Booee	Cindy	Bruno	Thelma
Listen	12	10	2	15
Drink	15	25	36	18
Shoe	14	18	60	20
Key	10	25	25	40
More	10	15	225	24
Food	80	55	14	190
Fruit	80	20	177	297
Hat	78	99	178	297
Look	115	54	345	420
String	129	476	287	372

Figure. Learning times in minutes plotted against word.



Three Main Categories of Studies

1. Surveys
2. Scientific studies (experiments)
3. Observational Studies

Surveys

A survey is different from a scientific study and an observational study because

- the population of interest is passively sampled. The sample units are *not* manipulated in any way. The goal is to describe the population in its natural state
- For example, NBC conducted exit polls in Florida to estimate the proportion of votes for Bush, Gore, and all other candidates on election day

Scientific Studies (Experiments)

A scientific study is different from a survey and an observational study because

- treatments are *actively imposed* on a set of experimental units
- if successful, *causation* may be attributed to the treatments
- For example, (ignoring the obvious flaws in the design) Fouts can attribute differences in learning times to differences among words. It can be said that there is statistical evidence that string is more difficult to learn than listen

Observational Studies

An observational study is different from a scientific study and a survey because

- treatments are *passively imposed* on a set of experimental units. That is, each population unit can be classified as being in a treatment group, but it cannot be said that the scientist assigned units to treatment groups
- for example, the sex discrimination study is an observational study. The (treatment) groups are gender (male and female)
- the question of interest was: are there differences in starting salary between males and females? A researcher cannot assign gender to a human being

A Closer Look at Surveys

- The goal is to describe a population of interest. It is essential that the sample is representative of the population.

Some sampling techniques

1. Simple random sampling (SRS) - select n units in such a manner that any set of n units has the same probability of being sampled. This can be done by sequentially drawing units at random until n have been selected. All unsampled units have the same probability of being selected on each draw.

- To collect an SRS of registered voters, get a list of the names of registered voters (say N voters), label them $1, 2, 3, \dots, N$. Randomly permute these integers, and select the first n labels for sampling.

2. Stratified random sampling - stratify the population according to a variable (say, location of residence), and select separate random samples of size n_1, n_2, \dots, n_s from each of the s strata.

- To collect a stratified random sample of registered voters, get a list of the names of registered voters and their voting district, and select an SRS from each voting district.

3. Cluster sampling - divide the population into many small sets (clusters), choose a SRS of clusters, and sample all units in each selected cluster.

- To collect a cluster sample of annual incomes, get a list of the households, select a SRS of households, and sample all individuals in each sampled household. For what population is this a valid sample?

4. Systematic sampling - construct a list of population units (sometimes called a sampling frame), and select every k th unit in the list. E.g., randomly choose an integer between 1 and 100, say 56, sample unit 56, and every 100th after that (i.e., units 56, 156, 256, ...). Transect sampling is another example.

Sampling Difficulties

- every unit in the population must have a known, non-zero probability of being sampled. For example, if sampling is done by telephone, then some voters have a probability of 0 of being sampled.
- humans often give non-representative responses if questions are leading, e.g., "In light of the large number of reversals of murder convictions based DNA testing, are you in favor of the death penalty?"

A Closer Look at Scientific Studies

- Scientific studies are conducted by experimentation. Treatments are imposed on experimental units. Experimental units should be representative of the population of interest

Terminology:

- A *factor* is a variable with *levels* chosen by the experimenter and imposed on the experimental units. In the chimp experiment, there were two factors: word and chimp. There were 10 levels of word: listen, drink, . . . , string. Notice that the levels of word are in the control of the experimenter
- *Treatments* are all the combinations of factor levels used in the experiment. E.g., (Bruno,string), (Cindy,shoe)
- The *response* variable is the measured variable. In the chimp study, the response variable was learning time. The observations were the recorded learning times. There were $4 \times 10 = 40$ observations.

Another example: To determine the strongest alloy of iron (Fe), titanium (Ti), and aluminum (Al), 10 m rods were produced as mixtures of the metals shown in the following table. Each rod is destructively tested, and the force used to break the rod is the response variable.

Table. Number of rods tested, by treatment. Aluminum is the third metal.

	Fe (%)		
Ti (%)	25	50	75
25	20	10	20
50	10	20	0
75	20	0	0

- There are 3 factors: percent Ti, percent Fe, percent Al
- The levels are 25, 50, 75% for Ti, 25, 50, 75% for Fe, 0, 25, 50% for Al
- The 6 treatments (Ti, Fe, Al) are: (25, 25, 50), (25, 50, 25), (25, 75, 0), (50, 25, 25), (50, 50, 0), (75, 25, 0).
- The total number of observations is $20 + 10 + \dots + 20 = 100$

Random allocation of units to treatments is essential for a statistically valid experiment. The assignment of treatment to units must be random with respect to the treatment

- For example, suppose that the metals are melted in three large cauldrons, and the first 20 rods are assigned treatment (25, 25, 50), the next 10 are (25, 50, 25), and so on. As the metals cool, the mixing properties change, and the strength may change as well. The results will be statistically invalid without randomization
- Random allocation can be accomplished by labeling the first 20 rods to be in treatment (25, 25, 50), the next 10 in (25, 50, 25), and so on.

Label	Treatment
1	(25, 25, 50)
2	(25, 25, 50)
⋮	⋮
20	(25, 25, 50)
21	(25, 50, 25)
⋮	⋮
100	(75, 25, 0)

Then, randomly permute the integers 1, 2, 3, ..., 100 (e.g., 21, 53, 92, 1, ..., 65), and form the rods according to the permutation order. I.e., first produce rod 55, then 23, ..., lastly, 65.

A Closer Look at Observational Studies

- Data is collected by passively sampling a population, and the population can be classified according to a variable that is the same as a factor in a scientific study *except that* the scientist does not impose the treatments onto the population units.
- Formal, rigorous arguments for causation are not defensible because there may be other variables that cause the observed effect
- The classic example is smoking studies. For many years, it has been argued that smoking causes cancer based on studies that find higher rates of lung and oral cancers for smokers than nonsmokers. Yet, it cannot be said that cancer is caused by smoking. These individuals may be predisposed to cancer because they are innately weaker and more susceptible to habit-forming behaviors

Case Study Example - State Average SAT Scores

- Citations: B. Powell and L.C. Steelman, 1984. "Variations in state SAT performance: Meaningful or Misleading?" *Harvard Educational Review* **54**(4), 389-412, and Ramsey, F.L. and Schafer, D.W., 1997, *The Statistical Sleuth*, Duxbury Press, p. 327.
- In 1982, SAT scores were first published on a state-by-state basis. Huge variation between states was observed. This was a source of great pride for some states, and consternation for others. Out of 1600 points, average scores ranged from 790 (S. Carolina) to 1088 (Iowa)

These authors set out to "assess the extent to which the compositional/demographic and school-structural characteristics are implicated in SAT differences."

- Variables were opportunistically collected. They were:
 1. takers - (test-takers) percentage of total eligible students that took the exam
 2. income - median family income of the test-takers (hundreds of dollars)
 3. years - average number of years that the takers took formal courses in social sciences, humanities, and natural sciences
 4. public - percentage of takers that attended public secondary school
 5. expend - total state expenditure on secondary schools (hundreds of dollars per student)
 6. rank - median percentile ranking of the test takers among the students in their classes
- This is an observational study. It is probable that one or more variables (or factors) affect the average state scores. We will determine (later) the degree of association between these variables and average score. *But*, no matter how strong the association, we cannot say whether one variable or another causes high average scores