Welcome to Stat*2040

Statistics I

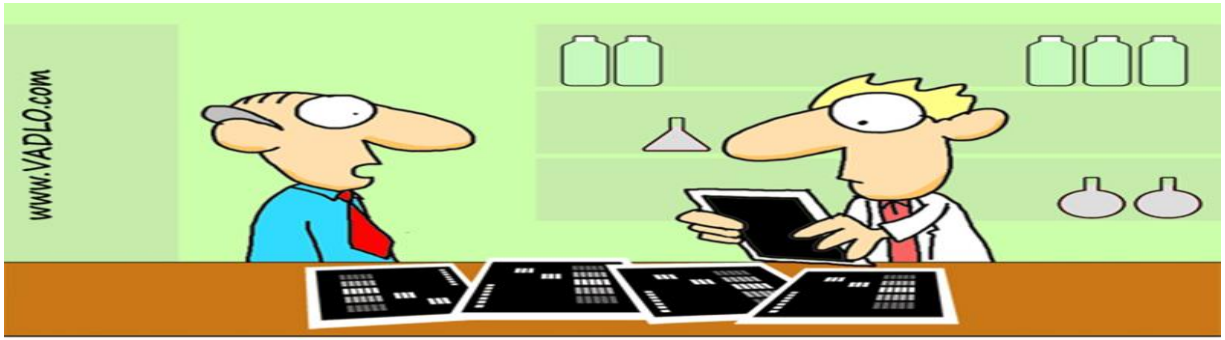
Instructor: Dr. Faisal Khamis

Introductory Statistics: A Problem-Solving Approach

by Stephen Kokoska

Chapter 1

An Introduction to Statistics and Statistical Inference



"Data don't make any sense,
we will have to resort to statistics."

Statistics Today

- Statistics are everywhere: in newspapers, magazines, the Internet, the evening weather forecast, medical studies, and even sports reports.
- They are used by professionals in many different disciplines in a variety of settings to make decisions that directly affect our lives.
- The goal in this course is how to use statistics to understand and make decisions about data.

No matter how you are employed or where you live, you will have to make decisions based on available data. some questions you may have to consider:

1. Do you have enough data to make a confident decision? how will these data be gathered?
2. How are the data summarized? Are the graphical and/or numerical techniques used appropriate?
3. What is the appropriate statistical technique for analyzing the data?

What is Statistics?

Statistics is the science of collecting and interpreting data, as well as drawing logical conclusions from available information to solve real-world problems.

- The science of learning from (or making sense out of) data.
- The theory and methods of extracting information from observational data for solving real-world problems.
- The science of uncertainty.
- The art of telling a story with [numerical] data.

Applications of Statistics

Definitions

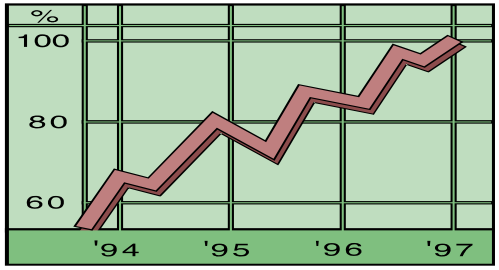
Descriptive statistics: Graphical and numerical methods used to describe, organize, and summarize data.

Inferential statistics: Techniques and methods used to analyze a small, specific set of data (sample) to draw a conclusion about a large, more general collection of data (population).

Applications of Statistics

Descriptive Procedures:

- **Collect data**
  - e.g., Survey, Observation, Experiments
- **Present data**
  - e.g., Charts and graphs



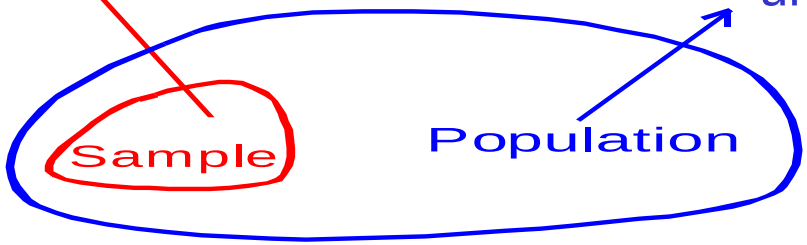
- **Characterize data**
  - e.g., Sample mean = $\frac{\sum x_i}{n}$

Inferential Procedures:

- Making statements about a population by examining sample results

Sample statistics (known) → **Inference** → Population parameters

unknown, but can be estimated from sample evidence



Population Versus Sample

Definitions:

A **population** is the entire collection of individuals or objects to be considered or studied.

A **sample** is a subset of the entire population, a small selection of individuals or objects taken from the entire collection.

A **variable** is a characteristic of an individual or object in a population of interest.

Example: A major pharmaceutical company is conducting clinical trials on a new drug it wants to bring to market. The company surveys 2000 people who have the particular condition that the drug is designed to treat. A portion of the people surveyed receive a high dose of the drug, a portion of the people surveyed receive a medium dose of the drug, and the remaining portion of the people surveyed receive a placebo (sugar pill). What are the population, sample, and variable in this experiment

The population is all people with the condition that the drug is intended to treat.

The sample is the 2000 people surveyed.

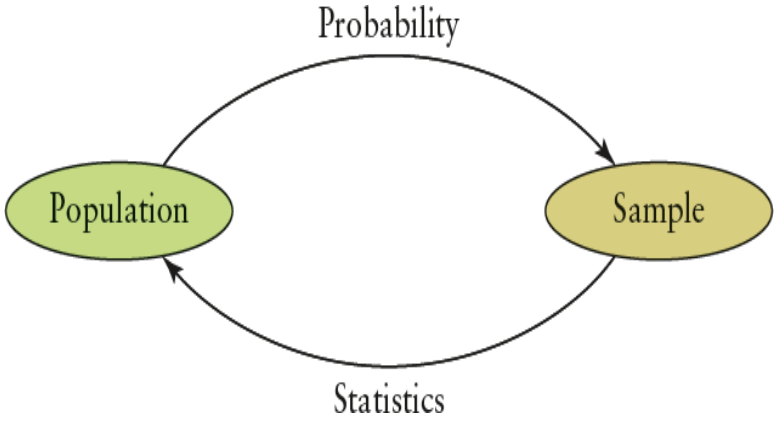
The variable is the dosage of drug administered to each person.

Probability Versus Statistics

Definition

To solve a **probability** problem, certain characteristics of a population are assumed to be known. We then answer questions concerning a sample from that population.

In a **statistics** problem, we assume very little about a population. We use the information about a sample to answer questions concerning the population.



Relationships among probability, statistics, population, and sample

EXAMPLE 1.8: Apple Pay is a method for making secure purchases and for sending and receiving money in stores, in apps, and on the web. According to a recent article, Apple Pay is now used by 35% of U.S. retail stores. Consider the population consisting of all U.S. retail stores and a random sample of 100 from this population.

A probability question: Suppose 35% of all U.S. retail stores use Apple Pay. What is the probability that at most 30 (of the 100) retail stores in the sample use Apple Pay? We know something about the population and try to answer a question about the sample.

A statistics question: Of the 100 U.S. retail stores selected, 45 use Apple Pay. What does this suggest about the proportion of **all** U.S. retail stores that use Apple Pay? We know something about the sample and try to answer a question about the population.

Observational Vs. Experimental Studies

Definition

In an **observational study**, we observe the response for a specific variable for each individual or object.

In an **experimental study**, we investigate the effects of certain conditions on individuals or objects in the sample.

What is the difference between observational and experimental studies?

An experiment that simply records the direction a rat turns at a particular junction in a maze is an observational study.

An experiment that forces rats to take a particular direction at a junction in a maze and then observes the differences in times for the rats that went left and rats that went right is an experimental study.

Observational Vs. Experimental Studies

EXAMPLE 1: Time for Breakfast?

A guidance counselor at Kerr Elementary School in Allen, Texas, is interested in the amount of time each student spends in the morning eating breakfast. The guidance counselor decides to measure the amount of time from wake-up to school bus arrival. A random sample of students is selected, and each is asked for the amount of school-day preparation time.

EXAMPLE 2: Gardening Advice

The manager of Gardener's Supply Company claims that a new organic fertilizer, in comparison with the leading brand, increases the yield and size of tomatoes. To test this claim, tomato plants are randomly assigned to one of two groups. One group is grown using the leading fertilizer; the other is cultivated using the new product. At harvest time, the size and weight of each tomato are recorded, along with the total yield per plant. The data collected during this experiment are used to compare the two fertilizers.

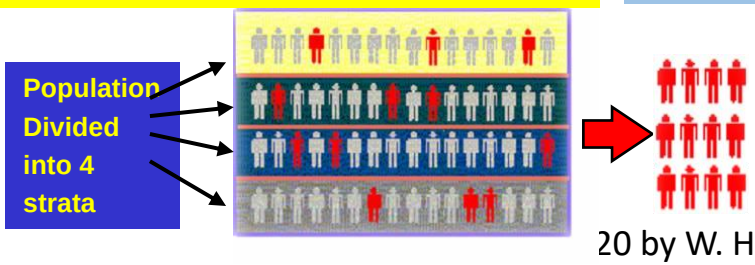
Goal: Convert data into meaningful information!

Sampling Techniques

Statistical Sampling (Probability Sampling)

Simple Random

- ✓ Divide population into subgroups (called strata) according to some common **characteristic**. e.g., gender, income level
- ✓ Select a simple random sample from **each** subgroup
- ✓ Combine samples from subgroups into one



Stratified

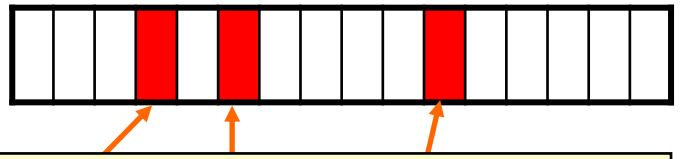
- ✓ Decide on sample size: n
- ✓ Divide ordered (e.g., alphabetical) frame of N individuals into groups of k individuals: k=N/n
- ✓ Randomly select one individual from the 1st group
- ✓ Select every kth individual thereafter

e.g.
k=100/10

Systematic

- ✓ Divide population into several "clusters," each representative of the population (e.g., county)
- ✓ Select a simple random sample of clusters

Population divided into 16 clusters.



Randomly selected clusters for sample

Simple Random Sample

Definition

A (**simple**) **random sample** (SRS) of size *n* is a sample selected in such a way that every possible sample of size *n* individuals has the same chance of being selected and the members of sample are chosen independently of each other (the chance of a given member of the population being chosen does not depend on which other members are chosen).

- In practice, a random sample may be very difficult to achieve. Statisticians employ various techniques, including random number tables and random number generators(such as www.random.org), to select a random sample.
- If a sample is not random, then it is biased. Many kinds of bias are possible, and many different factors may contribute to a biased sample.
- Nonresponse bias is very common when data are collected using surveys. Most people who receive a survey in the mail simply discard it. The original collection of people receiving the survey may be random, but the final sample of completed surveys is not. Because the sample is biased, it is impossible to draw a valid conclusion.

Simple Random Sample

- Self-selection bias occurs when the individuals (or objects) choose to be included in the sample, as opposed to being randomly selected. For example, a television news program may ask viewers to respond to a yes/no question by dialing one of two phone numbers to cast their vote. Viewers choose to participate, and usually those with strong opinions (either way) vote. Many more did not have the opportunity to respond, so every single sample is not equally likely. Certainly, this sample is biased, and hence no valid conclusion is possible.

- Before doing any analysis, you should always ask how the data were obtained. If any evidence of a pattern in selection is found, or if the observations are associated or linked in some way, or if the observations share some connection, then the sample is not random. There is simply no way to transform bad data into good statistics.

In summary the sample should be representative to our population

Simple Random Sample

How do we gather data?

Sampling Example:

How could the university's administration use simple random sampling to estimate how many UG undergraduates plan to apply to Statistics school?

One possible sampling design:

- Get a list of all currently enrolled undergraduates from the registrar
- Assign each undergraduate a number
- Use a random number generator to select a sample of undergraduates
- Email a survey to the sampled undergraduates

Statistical Inference Procedure

The process of checking a claim can be divided into four parts:

- **Claim**
- **Experiment**
- **Likelihood**
- **Conclusion**

- **Claim**
  This is a statement of what we assume to be true.
- **Experiment**
  To check the claim, we conduct a relevant experiment.
- **Likelihood**
  We consider the likelihood of occurrence of the observed experimental outcome, assuming the claim is true. We will use many techniques to determine whether the experimental outcome is a reasonable observation (subject to reasonable variability) or a rare occurrence.
- **Conclusion**: There are only two possible conclusions.
  1. If the outcome is reasonable, then we cannot doubt the claim. We have no evidence that suggests the claim is false.
  2. If the outcome is rare, we discount the lucky alternative and question the claim. A rare outcome is a contradiction: It shouldn't happen frequently if the claim is true. There is evidence to suggest that the claim is false.

Statistical Inference Procedure Example

A professor claims that 199 students are passed the course and only 1 is not (due to probably health conditions). A researcher reaches into the students and selects one at random, asked him/her, and finds that he/she is the one who didn't pass!

Claim: There were 199 passed students and 1 didn't in the class.
Experiment: The researcher selected one student from the class, asked him/her, and finds that he/she is the one who didn't pass!
Likelihood: One of two things has happened.

1. The researcher could be incredibly lucky. Intuitively, the chance of selecting the student who didn't pass from among the 200 total students is very small. It is possible to select the one who didn't pass, but it is very unlikely. We have found evidence that the claim is false by showing that the observed experimental outcome is unreasonable, an outcome so rare that it should almost never happen if the claim is really true.
2. The claim (199 passed students, 1 didn't) is false. Because the chance of selecting the student who didn't pass is so small, it is more likely the professor was mistaken about the number of students who didn't pass in the class (Perhaps there are really 199 didn't pass and only one good student in the class.)

Conclusion: Selecting the student who didn't pass is an extremely rare occurrence. Therefore, there is evidence to suggest the professor's claim is false