STAT 515

Statistical Methods I

Chapter 1

Data Types and Data Collection

Brian Habing
Department of Statistics
University of South Carolina

Redistribution of these slides without permission is a violation of copyright law.

Outline

- Introduction and Basic Terminology
- Collecting Data to Describe a Population
- Collecting Data to Compare Treatments

Chapter 1 - Statistics

Statistics Is the Science of Data…

- ***Gathering Data*: Design of Experiments and Surveys**
- ***Describing Data*: Descriptive Statistics**
- ***Using Data to Make Decisions*: Inference about Populations from Samples**
- ***The Theory Behind Analyzing Data*: Mathematical Statistics and Theoretical Probability**

Chapter 1 – Data in General

It is often easiest to picture data as being in the form of a spreadsheet

Student	College	Course Grade	Course %	Hmwk	Exam 1	Exam 2	Exam 3
1	440	B	83	85.60%	88	64	88
2	410	B+	87	94.40%	96	80	84
3	230	A	90	87.20%	84	76	100
4	112	B	84	69.60%	92	76	88
5	350	C	74	54.40%	72	56	88
6	390	D	65	40.80%	88	72	60

Chapter 1 – Cases and Variables

Each row is a different case, respondent, subject, participant, record, or experimental unit.

The variables are found in the columns and are the characteristics of the units.

Student	College	Course Grade	Course %	Hmwk	Exam 1	Exam 2	Exam 3
1	440	B	83	85.60%	88	64	88
2	410	B+	87	94.40%	96	80	84
3	230	A	90	87.20%	84	76	100
4	112	B	84	69.60%	92	76	88
5	350	C	74	54.40%	72	56	88
6	390	D	65	40.80%	88	72	60

Chapter 1 – Types of Data

Variables can be either quantitative (have a numerical value that works like a number) or qualitative (don't have a meaningful numerical value; aka categorical variables).

Student	College	Course Grade	Course %	Hmwk	Exam 1	Exam 2	Exam 3
1	440	B	83	85.60%	88	64	88
2	410	B+	87	94.40%	96	80	84
3	230	A	90	87.20%	84	76	100
4	112	B	84	69.60%	92	76	88
5	350	C	74	54.40%	72	56	88
6	390	D	65	40.80%	88	72	60

Chapter 1 – Types of Data

Qualitative variables can be divided into ordinal variables (ordered in a specific way, less than or greater than makes sense) or nominal variables (no intrinsic ordering).

Student	College	Course Grade	Course %	Hmwk	Exam 1	Exam 2	Exam 3
1	440	B	83	85.60%	88	64	88
2	410	B+	87	94.40%	96	80	84
3	230	A	90	87.20%	84	76	100
4	112	B	84	69.60%	92	76	88
5	350	C	74	54.40%	72	56	88
6	390	D	65	40.80%	88	72	60

Chapter 1 – Types of Data

Quantitative variables are often divided into discrete variables (there are jumps between the possible values) and continuous variables (there is another possible value between any two values).

Student	College	Course Grade	Course %	Hmwk	Exam 1	Exam 2	Exam 3
1	440	B	83	85.60%	88	64	88
2	410	B+	87	94.40%	96	80	84
3	230	A	90	87.20%	84	76	100
4	112	B	84	69.60%	92	76	88
5	350	C	74	54.40%	72	56	88
6	390	D	65	40.80%	88	72	60

Chapter 1 - Population and Sample

The population is the set of all units that we are interested in studying.

A numerical summary of a variable across the entire population is a parameter.

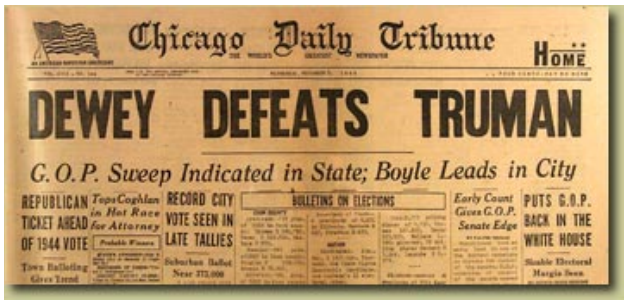
The sample is the subset of the population that we actually have observations from.

A numerical summary of a variable across the entire sample is a statistic.

Chapter 1 - Population and Sample

Statistical inference (that we're building up to) tells us how well the statistics that measure our sample capture the true parameter values that describe our population.

In order for the data from the sample to be useful, it must be representative of the population for the variable(s) we are interested in.


Chicago Daily Tribune

HOME

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Boyle Leads in City

REPUBLICAN TICKET AHEAD OF 1944 VOTE

Tops Coghlan in Hot Race for Attorney

RECORD CITY VOTE SEEN IN LATE TALLIES

BULLETINS ON ELECTIONS

Early Count Gives G.O.P. Senate Edge

PUTS G.O.P. BACK IN THE WHITE HOUSE

Chapter 1 – Collecting Data

Unfortunately, actually collecting data receives the most attention in STAT 110, and then is hardly mentioned in advanced courses (like this one).

This vastly understates how important collecting the right data in the right way is – if the data is bad, everything covered in any stats class is useless!

Chapter 1 – For Describing Populations

Random sampling (aka random selection) methods are designed to guarantee that our sample will be representative of the population on average for every variable.

Simple random samples (SRSs) are often used to describe a population. This is where every subgroup of individuals has the same chance of being chosen as any other subgroup.

Chapter 1 – For Describing Populations

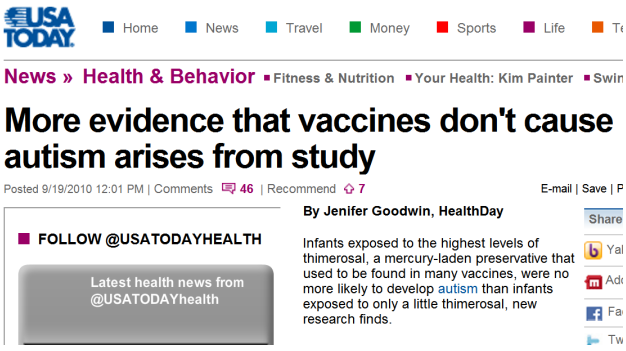
An SRS is easiest to work with mathematically, but other types of probabilistically based sampling methods can given even better results.

Actually getting an SRS and the data you really want can be very hard! (Do they have a phone? Is the question leading?)

Chapter 1 – For Describing Populations

A stratified sample is one where we have different sub-populations of known size. We take a simple random sample of each one and then combine the results in the right way.

The math is harder, but the results will have less variability (the statistic will usually be more close to the parameter) as long as the groups are related to the variable we're measuring.



USA TODAY Home News Travel Money Sports Life

News » Health & Behavior ▪ Fitness & Nutrition ▪ Your Health: Kim Painter ▪ Swin

More evidence that vaccines don't cause autism arises from study

Posted 9/19/2010 12:01 PM | Comments 46 | Recommend 7 E-mail | Save | P

FOLLOW @USATODAYHEALTH

Latest health news from @USATODAYhealth

By Jenifer Goodwin, HealthDay

Infants exposed to the highest levels of thimerosal, a mercury-laden preservative that used to be found in many vaccines, were no more likely to develop autism than infants exposed to only a little thimerosal, new research finds.

Share

Chapter 1 – For Determining Efficacy

The "gold standard" in statistics for determining the efficacy of something is a scientifically controlled, randomized, comparative experiment with a large sample size from a representative population in a realistic setting.

Chapter 1 – Experiment

An experiment is where the treatment is imposed upon the subjects by the experimenter. This allows for conclusions about whether the treatment causes a given effect or not.

Consider trying to judge the efficacy of charter schools – in many situations the parent has to be motivated enough to enroll their child in the charter school…

Chapter 1 – Comparative

Comparative means that several treatments are being compared.

Historically a large number of medical procedures were adopted because they seemed reasonable and some patients improved… thousands of people receive arthroscopic knee surgery each year, but a 2002 study showed it doesn't seem to give any better results than a "placebo surgery"!!

Knee Surgery Proves No Better Than Placebo

HOUSTON, Jul 10, 2002 (United Press International via COMTEX) -- For individuals suffering from osteoarthritis in their knees, a common type of knee surgery has been found to be no more beneficial than a placebo, a new study revealed Wednesday.

Researchers at the Houston VA Medical Center and at Baylor College of Medicine came to this surprising conclusion after comparing various knee treatments to placebo surgery on 180 patients with knee pain.

The patients were randomly divided into three groups. One group underwent debridement, in which the damaged or loose cartilage is the knee is surgically removed by an arthroscope, a pencil-thin tube that allows doctors to see inside the knee. The second group received arthoscopic lavage, which flushes out the bad cartilage from the healthier tissue. A third group underwent a placebo surgery. They were sedated by medication while surgeons simulated arthroscopic surgery on their knees by making small incisions on the leg, but not removing any tissue.

Chapter 1 – Randomization

All aspects that can be should be randomly assigned (aka randomly allocated).

Each subject in a drug trial is randomly assigned either the standard treatment or the proposed new treatment. But what if it is set up so that one doctor gives all the standard treatments, and a second doctor gives all the new treatments…

Chapter 1 – Scientifically Controlled

The effects of variables other than the one being investigated are minimized.

When examining the efficacy of a drug, what should it be compared to? What if the placebo has a better taste or is more easily digested or doesn't have some of the side effects?

Chapter 1 – Large Sample Size

The number of subjects needs to be as large as possible.

Consider looking for the occurrence of a rare side effect…

Chapter 1 – Representative Population

From a representative population.

Consider trying to judge fuel economy by using professional drivers… do the estimates have any relationship to the real world?

Chapter 1 – Realistic Setting

In a realistic setting.

Consider trying to judge fuel economy by using drivers on a closed track where they know fuel economy is the subject of the study… do the estimates have any relationship to the real world?

Chapter 1 – Reality Sets In

Of course, in many cases it's impossible to have a randomized comparative experiment with a large sample size from a representative population.

The effect of missing any of these conditions needs to be taken into account when determining what conclusions can be drawn from a study.