TOPIC: Introduction to Statistics

WELCOME TO MY CLASS!

Two statisticians were traveling in an airplane from Los Angeles to New York City. About an hour into the flight, the pilot announced that although they had lost an engine, there was no need for worry as the plane had three engines left. However, instead of 5 hours travel time it would now take them 7 hours to get to New York. A short while later, the pilot announced that a second engine failed. They still had two left, but it would take 10 hours to get to New York. Somewhat later, the pilot announced that a third engine had died. Never fear, he announced, because the plane could fly on a single engine. However, it would now take 18 hours to get to New York. At this point, one statistician turned to the other and said, "Gee, I hope we don't lose that last engine, or we'll be up here forever!"

Key terms used in the field of statistics

Population: Universe. The entire category under consideration. It is the data which we have not completely examined but to which our conclusions refer. If your company manufactures one million laptops, they might take a sample of say, 500, of them to test quality. The population N = 1,000,000 and the sample n= 500.

Examples:

- all pediatricians in the US;
- incomes of all adult NYC residents.

Sample: That portion of the population that is available, or to be made available, for analysis. A good sample is representative of the population. We will learn about probability samples and how they provide assurance that a sample is indeed representative. The sample size is shown as lower case n.

Parameter: A characteristic of a population. The population mean, μ and the population standard deviation, σ, are two examples of population parameters. If you want to determine the population parameters, you have to take a census of the entire population. Taking a census is very costly. The population size is usually indicated by a capital N.

Statistic: A statistic is a measure that is derived from the sample data. For example, the sample mean, $\bar{X}$, and the sample standard deviation, s, are statistics. They are used to estimate the population parameters.

Statistical Inference: The process of using sample statistics to draw conclusions about population parameters is known as statistical inference. For instance, using $\bar{X}$ (based on a sample of, say, n=1000) to draw conclusions about μ (population of, say, 300 million). This is a measure of performance in which the sample measurement is used to estimate the population parameter.

Note that pollsters do not call every adult who can vote for president. This would be very expensive. What pollsters do is call a representative sample of about 2,000 people and use the sample statistics to estimate who is going to win the election.

EXAMPLE: Nielsen television ratings

The Nielsen ratings are based on a sample, not the population.
The sample consists of about 5,000 TV households
Population of more than 115,000,000 TV households
For example, if a show has a 10.0 rating, this means that 10% of the entire sample were watching that show. [Note: "Share of audience" is the percentage of those who have the TV on , i.e. of those actually watching TV.]

EXAMPLE: market share of a product

Sample of supermarkets throughout the US to determine what percentage of people who buy a type of product (e.g., detergent) buy a specific brand (e.g., Tide).

Both of these examples are of statistics that are used to make inferences about the population.

Descriptive Statistics: Those statistics that summarize a sample of numerical data in terms of averages and other measures for the purpose of description, such as the mean and standard deviation.

Descriptive statistics, as opposed to inferential statistics, are not concerned with the theory and methodology for drawing inferences that extend beyond the particular set of data examined, in other words from the sample to the entire population. All that we care about are the summary measurements such as the average (mean). Thus, a teacher who gives a class, of say, 35 students, an exam is interested in the descriptive statistics. What was the class average, the median grade, the standard deviation, etc.? The teacher is not interested in making any inferences.

[For example, after grading an exam, a teacher may calculate the average grade to summarize the overall performance of the class. No inferences being made here.]

This includes the presentation of data in the form of graphs, charts, and tables.

Outline of this lecture

A. Sources of data
    1. Primary
        a. Surveys
            i. mail / email / web
            ii. telephone
            iii. personal interview
    2. Secondary

B. Survey errors
    1. Response errors
    2. Nonresponse error

C. Types of samples
    1. Nonprobability samples
        a. Convenience (chunk) sample
        b. Judgment sample
        c. Quota sample
    2. Probability samples
        a. Simple random sample
        b. Other types of probability samples
            i. systematic sample
            ii. stratified sample
            iii. cluster sample

D. Data
    1. Types of Data
        a. Qualitative Data
        b. Quantitative Data
            i. Discrete vs. Continuous
    2. Levels of Data
        a. Nominal, Ordinal, Interval, Ratio

A. Sources of Data

1. Primary data: This is data that has been compiled by the researcher.

Surveys, experiments, depth interviews, observation, focus groups.

Much data is obtained via surveys (uses a questionnaire).

Types of surveys:

Mail: lowest rate of response; usually the lowest cost
Personally administered: can "probe"; most costly; interviewer effects (the interviewer might influence the response)
Telephone: fastest
Web: fast and inexpensive

2. Secondary data: This is data that has been compiled or published elsewhere, e.g., census data. The trick is to find it. Also, the data was probably collected for some purpose other than helping to solve the researcher's problem at hand.

Advantages: It can be gathered quickly and inexpensively. It enables researchers to build on past research.

Problems: Data may be outdated. Variation in definition of terms. Different units of measurement. May not be accurate (e.g., census undercount).

Typical Objectives for secondary data research designs:

(1) Fact Finding, eg- amount spend by industry and competition on advertising; market share; # of computers with modems in U.S., Japan, …

(2) Model Building - specify relationships between two or more variables. Often using descriptive or predictive equations. Used, eg, to measure market potential, as per capita income + # cars bought in various countries.

Longitudinal vs. static studies.

B. Survey Errors

I. Response Errors

a) subject lies – question may be too personal or subject tries to give the socially acceptable response (example: "Have you ever used an illegal drug? "Have you even driven a car while intoxicated?")

b) subject makes a mistake – subject may not remember the answer (e.g., "How much money do you have invested in the stock market?"

c) interviewer makes a mistake – in recording or understanding subject's response

d) interviewer cheating – interviewer wants to speed things up so s/he makes up some answers and pretends the respondent said them.

e) interviewer effects – vocal intonation, age, sex, race, clothing, mannerisms of interviewer may influence response. An elderly woman dressed very conservatively asking young people about usage of illegal drugs may get different responses than young interviewer wearing jeans with tattoos on her body and a nose ring.

II Nonresponse error

If the rate of response is low, the sample may not be representative. The people who respond may be different from the rest of the population. Usually, respondents are more educated and more interested in the topic of the survey. Thus, it is important to achieve a reasonably high rate of response. Use follow-ups.

Which is better?

Sample 1	Sample 2
n = 2,000	n = 1,000,000
rate of response = 90%	rate of response = 20%

A small but representative sample can be useful in making inferences. But, a large and probably unrepresentative sample is useless. No way to correct for it. Thus, sample 1 is better than sample 2.

Example:

The Literary Digest, based on n= 2,000,000 – more than 2 million returned postcards – predicted a landslide win for Republican Alfred Landon over FDR in 1936.

The study had 2 biases:
1- large proportion of nonrespondents (at least 10 million "ballots" had been mailed out).
2) The questionnaires were mailed to people listed in telephone directories and on automobile registration lists (higher income people in 1936). But in that Depression year, millions of voters had no phones or cars.

C. Types of Samples

I Nonprobability Samples – based on convenience or judgment

1. Convenience (or chunk) sample
students in a class, mall intercept

2. judgment sample
based on the researcher's judgment as to what constitutes "representativeness"
e.g., he/she might say these 20 stores are representative of the whole chain.

3. quota sample
interviewers are given quotas based on demographics for instance, they are each told to interview 100 subjects – 50 males and 50 females. Of the 50, say, 10 nonwhite and 40 white.

The problem with a nonprobability sample is that we do not know how representative our sample is of the population.

II Probability Samples

Probability Sample: A sample collected in such a way that every element in the population has a known chance of being selected.

1. Simple Random Sample: A sample collected in such a way that every element in the population has an equal chance of being selected.

The following are other kinds of probability samples but beyond the scope of this course.

2. systematic random sample
3. stratified sample
4. cluster sample

Optional Topic: How to draw a simple random sample:

N= population size n= sample size

Suppose N=800 and n=80.

First, number all the elements in the population from 001 to 800. Then go to the random number table – or, more likely, use a random number generator – and select 80 3-digit random numbers. Discard numbers greater than 800.

Select a row randomly. For example, going across row 1:
661
942 (discard)
892 (discard)
699
547
166
254
551
567
953 (discard)
…

Note that every element in the population has an equal chance of being selected for the sample: n/N or 80/800 = 10%.

TABLE OF RANDOM NUMBERS

	Column							
	00000	00001	11111	11112	22222	22223	33333	33334
Row	12345	67890	12345	67890	12345	67890	12345	67890
01	66194	28926	99547	16625	45515	67953	12108	57846
02	78240	43195	24837	32511	70880	22070	52622	61881
03	00833	88000	67299	68215	11274	55624	32991	17436
04	12111	86683	61270	58036	64192	90611	15145	01748
05	47189	99951	05755	03834	43782	90599	40282	51417
06	76396	72486	62423	27618	84184	78922	73561	52818
07	46409	17469	32483	09083	76175	19985	26309	91536
08	74626	22111	87286	46772	42243	68046	44250	42439
09	34450	81974	93723	49023	58432	67083	36876	93391
10	36327	72135	33005	28701	34710	49359	50693	89311
11	74185	77536	84825	09934	99103	09325	67389	45869
12	12296	41623	62873	37943	25584	09609	63360	47270
13	90822	60280	88925	99610	42772	60561	76873	04117
14	72121	79152	96591	90305	10189	79778	68016	13747
15	95268	41377	25684	08151	61816	58555	54305	86189
16	92603	09091	75884	93424	72586	88903	30061	14457
17	18813	90291	05275	01223	79607	95426	34900	09778
18	38840	26903	28624	67157	51986	42865	14508	49315
19	05959	33836	53758	16562	41081	38012	41230	20528
20	85141	21155	99212	32685	51403	31926	69813	58781
21	75047	59643	31074	38172	03718	32119	69506	67143
22	30752	95260	68032	62871	58781	34143	68790	69766
23	22986	82575	42187	62295	84295	30634	66562	31442
24	99439	86692	90348	66036	48399	73451	26698	39437
25	20389	93029	11881	71685	65452	89047	63669	02656
26	39249	05173	68256	36359	20250	68686	05947	09335
27	96777	33605	29481	20063	09398	01843	35139	61344
28	04860	32918	10798	50492	52655	33359	94713	28393
29	41613	42375	00403	03656	77580	87772	86877	57085
30	17930	00794	53836	53692	67135	98102	61912	11246
31	24649	31845	25736	75231	83808	98917	93829	99430
32	79899	34061	54308	59358	56462	58166	97302	86828
33	76801	49594	81002	30397	52728	15101	72070	33706
34	36239	63636	38140	65731	39788	06872	38971	53363
35	07392	64449	17886	63632	53995	17574	22247	62607
36	67133	04181	33874	98835	67453	59734	76381	63455
37	77759	31504	32832	70861	15152	29733	75371	39174
38	85992	72268	42920	20810	29361	51423	90306	73574
39	79553	75952	54116	65553	47139	60579	09165	85490
40	41101	17336	48951	53674	17880	45260	08575	49321
41	36191	17095	32123	91576	84221	78902	82010	30847
42	62329	63898	23268	74283	26091	68409	69704	82267
43	14751	13151	93115	01437	56945	89661	67680	79790
44	48462	59278	44185	29616	76537	19589	83139	28454
45	29435	88105	59651	44391	74588	55114	80834	85686
46	28340	29285	12965	14821	80425	16602	44653	70467
47	02167	58940	27149	80242	10587	79786	34959	7…..

II. more probability samples

2. systematic random sample. Choose the first element randomly, then every kth observation, where k = N/n

3. stratified random sample. The population is sub-divided based on a characteristic and a simple random sample is conducted within each stratum

4. cluster random sample. First take a random sample of clusters from the population of cluster. Then, a SRS within each cluster. Example, election district, orchard.

D. Data

1. Types of Data

Qualitative data result in categorical responses.
Called *Nominal*, or *categorical* data

Example:

Sex ☐MALE ☐FEMALE

Quantitative data result in numerical responses, and may be discrete or continuous.

Discrete data arise from a counting process.

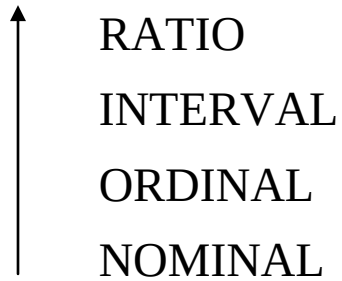
Example:
How many courses have you taken at this College? _____

Continuous data arise from a measuring process.

Example:
How much do you weigh? _____

One way to determine whether data is continuous, is to ask yourself whether you can add several decimal places to the answer. You may weigh 150 pounds but in actuality may weigh 150.23568924567 pounds. On the other hand, if you have 2 children, you do not have 2.3217638 children.

2. Levels of Data



NOMINAL
same as Qualitative

Classification – categories
When objects are measured on a nominal scale, all we can say is that one is different from the other

Examples: sex, occupation, ethnicity, marital status, etc.

> [Question: What is the average SEX in this room? What is the average RELIGION?]

Appropriate statistics: mode, frequency
We cannot get an average. The "average sex" in this class makes no sense!

Example:
Say we have 20 males and 30 females. The mode – the data value that occurs most frequently - is 'female'.
Frequencies: 60% are female.
Say we code the data – 1 for male and 2 for female:
(20 x 1 + 30 x 2) / 50 = 1.6
Is the average sex = 1.6? What are the units? 1.6 what? What does 1.6 mean?

Note: Roger Ebert's 2 thumbs up / down is nominal type of data. Nominal data is weak. Better if we use all five fingers. ☺

ORDINAL

Ranking, but the intervals between the points are not equal

We can say that one object has more or less of the characteristic than another object when we rate them on an ordinal scale. Thus, a category 5 hurricane is worse than a category 4 hurricane which is worse than a category 3 hurricane, etc.

Examples: social class, hardness of minerals scale, income as categories, class standing, rankings of football teams, military rank (general, colonel, major, lieutenant, sergeant, etc.), hurricane rankings (category 1, 2, …, category 5)

Example:
Income
Under $20,000 – checked by, say, John Smith
$20,000 – $49,999 – checked by, say, Jane Doe
$50,000 and over – checked by, say, Bill Gates

Bill Gates checks box 3 even though he earns several billion dollars. Distance between Gates and Doe is not the same as the distance between Doe and Smith.

Appropriate statistics
– same as those for nominal data, plus
– median, but not mean

ranking scales are obviously ordinal. There is nothing absolute here. Just because someone chooses a "top" choice does not mean it is really a top choice.

Example:

Please rank from 1 to 4 each of the following:
__ being hit in the face with a dead rat
__ being buried up to your neck in cow manure
__ failing this course
__ having nothing to eat except for chopped liver for a month

INTERVAL

Equal intervals, but no "true" zero.
Examples: IQ, temperature, GPA.

Since there is no true zero – the complete absence of the characteristic you are measuring – you cannot speak about ratios.

Example:
Suppose
New York temperature = 40 degrees
Buffalo temperature = 20 degrees
Does that mean it is twice as cold in Buffalo as in NY? No.

Appropriate statistics
- same as for nominal
- same as for ordinal
  plus,
- mean

RATIO

Equal intervals and a "true" zero.

Examples: height, weight, length, units sold

All scales, whether they measure weight in kilograms or pounds, start at 0.
The 0 means something and is not arbitrary.

100 lbs is double 50 lbs (same for kilograms)
$100 is half as much as $200

How to Choose what type of data to collect?

The goal of the researcher is to use the highest level of measurement possible.

Example:
(A) Do you smoke? __ yes __ no
vs.
(B) How many cigarettes did you smoke in the last 3 days (72 hours)?
_____

(A) is nominal, we can get frequencies
(B) is ratio, we can get mean, median, mode, frequencies

Example:
(A) Please rank the taste of the following soft drinks (1=best, 2= next best, etc.)
__ Coke
__ Pepsi
__ 7Up
__ Sprite
__ Dr. Pepper

vs.

(B) Please rate each of the following brands of soft drink:

Coke:	(1) excellent	(2) very good	(3) good	(4) fair	(5) poor	(6) very poor	(7) awful
Pepsi:	(1) excellent	(2) very good	(3) good	(4) fair	(5) poor	(6) very poor	(7) awful
7Up:	(1) excellent	(2) very good	(3) good	(4) fair	(5) poor	(6) very poor	(7) awful
Sprite:	(1) excellent	(2) very good	(3) good	(4) fair	(5) poor	(6) very poor	(7) awful
Dr Pepper:	(1) excellent	(2) very good	(3) good	(4) fair	(5) poor	(6) very poor	(7) awful

This scale (B) is almost interval and is usually treated so – means are computed. We call this a rating scale. By the way, if you hate all five soft drinks, we can determine this by your responses. With scale (A), we have no way of knowing whether you hate all five soft drinks.

Rating Scales – what level of measurement? Probably better than ordinal, but not necessarily exactly interval. Certainly not ratio.

Are the intervals between, say, "excellent" and "good" equal to the interval between "poor" and "very poor"? Probably not. Researchers will assume that a rating scale is interval.

This course will help you learn to think for yourself. Knowledge of statistics will allow you to see the difference between junk science and real science.

EXAMPLE: An interesting article you may want to read is by John Tierney (*NY Times*, Science, October 9, 2007, pp. F1-F2 "Diet and Fat: A Severe Case of Mistaken Consensus"). Apparently, the "research" indicating that a diet rich in fatty foods leads to a shortened life because it causes heart disease and other ailments is wrong. Doctors have been recommending low-fat diets for many years and fell into the trap of what is known as "informational cascade." If one person states something with a great deal of confidence, and then another agrees with this information, then the third person will probably go along with this since s/he will assume the first two must know what they are talking about. By the time the cascade has run its course, everyone will assume that the information is correct. The doctor who started the informational cascade with fatty foods was Dr. Ancel Keys in the 1950s. He made several blunders in his research:
(1) It is not clear that traditional diets were very lean. Ancient man did a great deal of hunting and had considerably more fat in his diet than we do today.
(2) There are more cases of heart disease being reported today probably because people live longer and are more likely to see a doctor; it is not due to worse health.
(3) Keys correlated diet and heart disease in 6 countries (one was the US) and found a relationship between fat in the diet and heart disease. Had Dr. Keys studied 22 countries for which data was available, he would have found no correlation.

The moral of the above is that you have to learn to think for yourself and not believe information that may just be the view of one person and a case of mistaken cascade.