Arvind Borde / MAT 19, Week 1: Basic Concepts of Statistics

(1) Why are you here? _____

(2) What's ? _____

a) _____

b) _____

c) _____ and

d) _____

_____

_____

1

The "information" that is collected, organized, etc., in statistics is called _____

can be _____

(3) Write down three types of data about yourselves that are numerical, and three that are not:

Num: 1. _____ Non: 1. _____

2. _____ 2. _____

3. _____ 3. _____

2

Misuses of Statistics: Examples

A] A talk show host uses the number of people who call in with an opinion to measure how widespread the opinion is.

(4) Why is this misleading?

_____

(a) _____

(b) _____

3

B] A study says that gray cars are involved in fewer accidents than cars of other colors. The media announces that gray cars are safer.

(5) Correct conclusion? _____

_____

_____

4

C] Studies show that breast-fed children have higher IQs than those who were not breast-fed. They say on TV that breast-feeding increases a child's IQ

(6) Correct conclusion? _____

_____

_____

5

D] An athlete with a very successful later career offers a defense against the charge that steroids were involved by providing examples of other athletes with similarly successful later careers.

(7) Is this statistically valid?

_____

_____

6

ADDITIONAL NOTES

Examples of Statistical Statements

New York Times, January 8, 2016

Seattle Seahawks' Defense Stands Out, Even Across Eras

"Over the last four years, Seattle has allowed 15.73 points a game, relative to the league average of 22.89. The standard deviation in points allowed by the league's 32 teams over this period has been 2.62 points per game. This means Seattle, which has been over 7 points a game better than average, has been 2.73 standard deviations better than average, a statistic known as a Z-score."

7

New York Times, January 15, 2016

Signs of a Retail Rebound as Consumers Take On More Debt

"Statistics from the Federal Reserve Bank of New York show that auto loans increased by more than 12 percent from the third quarter of 2014 to the third quarter of 2015. That left the total of auto credit at $1.05 trillion"

8

New York Times, September 5, 2015

The Collateral Victims of Criminal Justice

"Between 1991 and 2007, the percentage of children with mothers in prison more than doubled, according to federal data – and that does not count the many more mothers who spent time in jail. It doesn't take statistics to grasp how damaging separation can be, but even so, the data shows these children have more depression, aggression, delinquency, absenteeism, asthma and migraines."

9

New York Times, September 5, 2015

Awash in Data, Thirsting for Truth

". . . in a numbers-soaked era. . . questions arise: How important is data to reporting? And does it get readers closer to the truth or obscure it?

"I'm thinking about these questions because of two recent pieces in The New York Times,. . .

"The first was a major exposé of brutal working conditions at Amazon, based on six months of reporting . . . some, including Amazon brass, contested it . . .

10

"The second piece was a magazine cover story, built on data, arguing that a feared digital-age 'creative apocalypse' never happened. . . . Many readers disputed the article's conclusions and charged that the numbers it relied upon had been 'cherry-picked' to make a flawed case. . . ."

– Margaret Sullivan, The Public Editor

11

Understanding statistical tests in the medical literature

"When writing or reading articles, one should be aware whether the statistical tests performed were appropriate for the type of data collected and used, thereby avoiding misleading conclusions. The goal of all statistical tests is to determine whether two (or more) variables are associated with one another or independent from each other. . . "

"One of the first things to keep in mind is the type of data and outcomes the author wants to measure and correlate. . .

12

ADDITIONAL NOTES

"The second important thing to keep in mind is how the results are distributed. Do they follow a 'bell curve'. . . , similar to biological phenomena and exam grading techniques, or do the results tend to cluster resulting in a skewed distribution?"

– National Institute of Health, September 2008

13

Health & Health Profession Statistics

"Nursing is the nation's largest health care profession, with more than 3.1 million registered nurses nationwide. Of all licensed RNs, 2.6 million or 84.8% are employed in nursing.

"Registered Nurses comprise one of the largest segments of the U.S. workforce as a whole and are among the highest paying large occupations. Nearly 58% of RNs worked in general medical and surgical hospitals, where RN salaries averaged $66,700 per year."

– American Association of Colleges of Nursing
September 2015

14

New York Times, August 19, 2017

The Stock Market Has Been Magical. It Can't Last.

"The bull market that started in March 2009 seems to have gone on forever, but that's only a small part of the story. What is astonishing is the way that stocks have risen in 2017. . . .

"Several statistical measurements demonstrate how unusual this market environment has been."

– Jeff Sommer

15

Here are some of the concepts mentioned in these examples that we'll study in this course:

- _____
- _____
- _____
  _____
- _____

16

Some Statistical Questions

(8) Here are some real baseball batting averages:

	1995	1996
Derek Jeter	.250	.314
David Justice	.253	.321

Who, would you guess, had the higher average over the two years? _____

17

(9) How is this possible?

Here are the details of the years:

	1995		1996	
Jeter	12/48	.250	183/582	.314
Justice	104/411	.253	45/140	.321

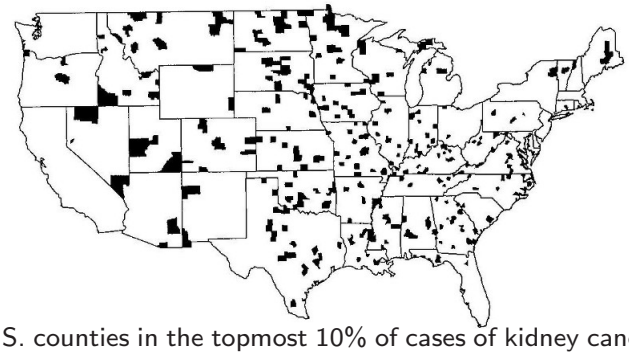
Now combine the two years:

Jeter: _____

18 Justice: _____

ADDITIONAL NOTES

(10) What pattern do you see here?



U.S. counties in the topmost 10% of cases of kidney cancer deaths among males, adjusted for age, 1980–89.

19

(11) _____

_____

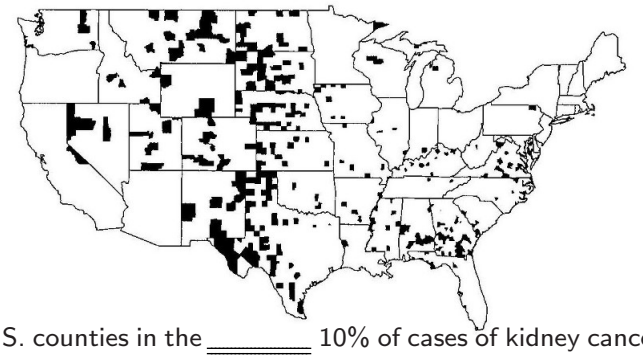
(12) _____

_____

_____

20

(13) What pattern do you see here?



U.S. counties in the _____ 10% of cases of kidney cancer deaths among males, adjusted for age, 1980–89.

21

_____

_____

(14) What gives?

_____

(15) _____

22

Some U.S. Counties, with population

1. LA, CA: _____
2. Cook IL: _____
5. SD, CA: _____
8. Kings, NY: _____
10. Queens, NY: _____

...

~3149. Loving, TX: _____

~3150. Kalawao, HI: _____

23

If, by pure chance, a county, with a population of 100 has one death in this category in the decade, as opposed to none, its rate will change from zero (low), to _____ (high).

A random fluctuation of a few deaths in LA, on the other hand, will not change its stats as much.

_____

_____

24

ADDITIONAL NOTES

(16) In WWII statistician Abraham Wald analyzed bullet damage on planes that returned from missions. He recommended that systematically damaged portions of aircraft be left unchanged, but that the *undamaged* parts be re-enforced. Why?

25

(17) I notice that when I get to a bus stop, I am more likely to see a bus going east than west. How might this be possible?

26

Some Statistical Terms

Definition

The entire group under study is the _____. An _____ is a person or object that is a member of the population. A _____ is a subset of the population.

27

Example A: I want to study the weekly spending habits of Post students. I survey this class.

(18) What is the population in this study?

(19) Who are the individuals?

(20) What is my sample?

28

Example B: I want to figure out who will win an election. I select 100 landline numbers at random and call them to ask who they plan to vote for.

(21) What is the population in this study?

(22) Who are the individuals?

(23) What is my sample? _____

29

Example C: I want to figure out who will win an election. I select 100 mobile numbers at random and call them to ask who they plan to vote for.

(24) What is the population in this study?

(25) Who are the individuals?

(26) What is my sample? _____

30

ADDITIONAL NOTES

Definition

_____ consist of organizing and summarizing data.

Typical ways to organize, summarize and display data are numerical summaries, tables, and graphs.

31

Definition

_____ uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

You want to describe and display information you've collected, and also to *infer* something from it. The process is never completely accurate, and you need a way to measure your confidence in the inference.

32

Definition

A _____ is a numerical summary of a sample.

If my survey of you (see slide 28) shows that you spend $102 per week on average, that number is a statistic.

33

Definition

A _____ is a numerical summary of a full population.

If *all* students at Post were asked, and the average weekly spending was found to be $93, that number is a parameter.

One goal of inferential statistics is to estimate parameters from statistics.

34

Variables in statistics are the characteristics of individuals within the population. They can be classified into two groups: _____ or _____

In my weekly spending example, students are the individuals. The amount of money they spend is the variable (varies from individual to individual).

35

Definition

_____ are descriptive characteristics.

These variables are often non-numerical, but don't have to be.

Example: _____

36

ADDITIONAL NOTES

Definition

_____ are numerical measures of individuals that can be manipulated mathematically (added, subtracted, etc.).

Example: _____

37

(27) If, instead of names, I identify you by your ID number, does that switch your identifier variable from qualitative to quantitative?

_____

_____

_____

38

(28) Identify these as quantitative or qualitative:

a) Your nationality. _____

b) Your temperature. _____

c) Your weight. _____

d) Your zip code. _____

39

Definition

A _____ is a quantitative variable that has either a finite number of possible values or a "countable" number of possible values.

The term countable means that the values result from counting, such as 0, 1, 2, 3, and so on. A discrete variable cannot take on every possible value between any two possible values.

40

Definition

A _____ is a quantitative variable that has an infinite number of possible values that are not countable.

A continuous variable may take on every possible value between any two values.

41

(29) One of the two quantitative variables we used as examples, height and weekly spending, is discrete and one is continuous. Which is which?

_____

_____

42

ADDITIONAL NOTES

These distinctions are not always clearcut.

(30) What type of variable is "handedness"?

Fill out the table on the next slide, as follows:

1 in R or L column if you *usually* use that hand.
2 if you *always* use that hand.
1 in *each* column if you often switch.

43

Handedness scoresheet

Task	Left	Right
Writing		
Drawing		
Throwing		
Scissors		
Toothbrush		
Cutting (knife)		
Spoon		
Hand on lid when opening jar		

44

R: Your total Right score:

L: Your total Left score:

Your handedness score: $\frac{\mathbf{R}-\mathbf{L}}{\mathbf{R}+\mathbf{L}} =$

45

Definition

An _____ measures the value of variables without attempting to influence them.

In an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome.

46

Definition

A _____ changes the value of a variable, and then records the value of the response.

In a designed experiment, the researcher changes the situation in order to see if it influences the outcome.

47

A study of the possible relationship between cell phone use and brain tumors that looks at people with and without tumors, then studies if their cell phone usage differs is observational.

A study that bombards rats with cell phone frequency radiation to see if they develop tumors is a designed experiment.

48

ADDITIONAL NOTES

It may seem that an observational study does less harm. But it may also cause confusion.

One observational study has shown that seniors who get 'flu shots are far less likely to get the 'flu, or associated diseases such as pneumonia.

(31) Does this prove that the 'flu shot makes people healthier, or are there variables that have not been taken into account? _____

49

Definition

A _____ is a variable that was not considered in a study, but that affects the value of the response in the study.

In our 'flu example, those who got the shots were often more mobile than ones who did not, so mobility was a lurking variable. (Must not confuse _____ and _____.)

50

Sampling

How do we pick samples?

Definition

_____ is the process of using chance to select individuals from a population to be included in the sample.

An opposing approach is _____

51

(32) In the weekly spending example on slide 28, was my sampling random or convenience?

_____

(33) How might I randomize the sample?

_____

52

Definition

A sample is a _____ if each
(i) _____
(ii) _____
_____ and
(iii) _____
_____

53

For example, I have 27 students in class. I decide to give out 10 A grades at random. I write on pieces of paper all possible lists of 10 students, then pick one list out of a hat.

In practice, one picks one student at random, then the next, etc., but this is not *exactly* the same.
(34) Why not? _____

54

ADDITIONAL NOTES

We'll concentrate on simple random sampling in this course, but (for the record), here's another kind of random sampling: _____

_____

It's used in situations where the population has _____, or _____, and it's important to ensure that each is proportionately represented in the sample.

55

(35) Can you give an example where such situations might occur? _____

_____

(36) What would you do in situations where strata are important?

_____

56 _____

(37) If 5% of the population has an income above $\$1{,}000{,}000$, 60% between that and $\$100{,}000$, and the rest under $\$100{,}000$, what would a good stratified sample of 200 contain?

_____

_____

_____

57

Even with "random" sampling biases can occur. In our polling examples (slides 29 and 30) picking samples based on landlines vs. mobile, each introduce biases.

(38) Who is more likely to still have a landline?

_____

(39) Who is more likely to have a mobile?

58 _____

Estimation

(40) How many school busses in the U.S.?

59

ADDITIONAL NOTES