STATISTICS LECTURE 1

⇒ This is a course in DATA ANALYSIS

⇒ When you work with data and calculate statistics you need to know HOW you calculated the statistics

⇒ You cannot obtain the correct result unless you follow the proper procedure

⇒ You will ALWAYS make mistakes, so you need a way to audit your work to catch the mistakes

⇒ Spreadsheets make it very difficult to audit your work because:

- formulas buried within a cell
- references are to cell numbers (e.g. "B2") instead of variable names
- difficult to place comments in the spreadsheet
- no history of commands

⇒ By contrast, programming languages make it EASY to audit your work

- formulas are explicitly written out
- references are to variable names
- easy to place comments in script
- the script is the history

⇒ Additionally, programming languages

- make it easy to set conditions
- create tables
- automate routine tasks
- work with large datasets

⇒ In my own work, I use Perl and R but those languages are not beginner friendly

⇒ For this course, we'll use Gretl

- point & click
- also has programming language
- written specifically for students

What is statistics?

- method of data analysis
- numbers computed from a data set

- methods include calculation of an average, plotting data, looking for relation between two or more variables

- Examples of "computed numbers"
  - Obama's approval rating
  - "Earned Run Average"
  - "chance of precipitation is 30%"

NOTE: Good data collection is critical

- result of Obama's approval rating very different depending on WHO you survey
- how ~~Who~~ do you select people for the survey?

→ It's also important to formulate good questions that you want to explore with the data

- Are people who took larger loans more/less likely to default on their home mortgage?
  - tricky because amount borrowed also depends on income
  - so perhaps we should look at loan-to-income ratio

→ Beware of the average

The distribution of the variable is very important

House #1	$80,000
House #2	35,000
House #3	70,000
House #4	15,000
House #5	2,300,000

average: $\frac{\$2,500,000}{5} = \$500,000$

median: $70,000

→ Stretching or shrinking horizontal/vertical axis can make trend look striking or insignificant

→ descriptive (explanatory) vs. inferential statistics

- descriptive – summarizes a particular set of data; does not try to draw conclusions beyond the observed sample
- inferential – assuming that you have a representative sample you can draw conclusions about the larger population

  → must be careful
  → how much confidence can we have in the estimate?

Types of Data

→ numerical vs. categorical

- numerical – a measurement
  - → income } continuous
  - → age }
  - → number of children ← discrete
- categorical
  - → male or female
  - → Republican or Democrat
  - → white, black, Asian, etc.

Note that you cannot take the average of categorical data

→ objectivity in measurement

- multiple choice exams – objective
- essay questions – subjective
- To obtain objective data, care must be taken so that the process of observation does not affect the measurement

→ level of measurement

- time + quantities are straightforward
- underlying variables are more difficult not directly measurable, so:

Likert Scale

1 – strongly agree
2 – agree
3 – indifferent
4 – disagree
5 – strongly disagree

} performance (behavioral) variable

but is the magnitude between "agree" and "indifferent" really the same as the magnitude between "disagree" and "strongly disagree"

- average Likert response may or may not have meaning

→ nominal-level measurement

- male / female
- Ford, Toyota, Hyundai
- counts of such variables are suitable for statistical analysis

→ ordinal-level measurement

- shortest, middle-height, tallest
- CEO, manager, worker
- once again, differences between adjacent classes in the rank order are not necessarily equal

→ interval-level measurement

- $10, $20, $30, $100
- difference between $20 & $30 is equal to difference between $60 & $70
- arithmetic average has meaning
- but ratio might not have meaning

degrees Fahrenheit	Celsius	Kelvin
-459,7	-273,15	0 ← absence of molecular activity
32	0	273,15
50	10	283,15
100	38	310,93
212	100	373,15

so 100°F is 1,10 times hotter than 50°F because 310,93°K is 1,10 times 283,15°K

⇒ ratio level measurement

- must have distinct classes, order among classes & equal intervals between classes AND
- must have absolute zero for the underlying variable

  - 200 pounds is twice as heavy as 100 pounds
  - 2 hours is twice as long as 1 hour

⇒ absolute scale measurement

- when we count the number of occurrences of a variable of any scale (nominal, ordinal, interval or absolute)

  - 7 men & 5 women
  - number of people who voted for Obama

  - 1 black pen
    3 red pens
    4 blue pens
    } 2,667 pens per color

→ continuity

10.893146 27 cm long ← continuous

$ 5.46 ← discrete

money has a minimal size unit whereas length does not

the average American family has 2.3 children Note: it is still possible to compute statistics for discrete valued variables

Empirical Frequency Distributions

⇒ Textbook has the following example

home state	number of students	
Indiana	250	48%
Illinois	110	22%
Ohio	55	11%
Kentucky	30	6%
Michigan	20	4%
Other US	37	7%
Foreign	12	2%
total	516	100%

note: nominal-level data

⇒ notice what's missing

- the total + percentages (added in red)
- which counties are the Indiana students from?
- which states are the "other US" students from?
- which countries (or at least regions of the world) are the "foreign" students from?

⇒ table also tells us only about geography

⇒ it may also be interesting to know how [income, race, religion, opinion, etc.] vary by geography

~~[illegible]~~

⇒ perhaps we have a second variable

response	frequency
strongly agree	10
agree	16
no opinion	20
disagree	36
strongly disagree	18
	100

ordinal level data

Textbook's Q:
"Too much money is being spent on national defense."

⇒ when we discuss cross tabulation, we'll see that the two tables can be combined to show how opinion may or may not depend on geography

⇒ EXAMPLE

	no LisP	LisP	total
no modification	29,318	5,644	34,962
HAMP mod	3,524	811	4,335
non-HAMP mod	4,023	697	4,720
total	36,865	7,152	44,017

But that's hard to interpret, so we may want to look at ~~[illegible]~~ percentages across rows or percentages across column

	no LisP	LisP	
no mod	84%	16%	100%
Hamp mod	81%	19%	100%
non-Hamp mod	85%	15%	100%
	84%	16%	

	no LisP	LisP	
no mod	80%	79%	79%
HAMP mod	10%	11%	10%
non-HAMP mod	11%	10%	11%
	100%	100%	

⇒ The previous example used categorical data (nominal-level measurement)

⇒ But what if we had dollar values? (interval-level measurement) Such variables are almost continuous (though not entirely)

⇒ We could create categories

monthly payment	no Lisp	Lisp	
under $2000	54%	42%	52%
$2000 and up	46%	58%	48%
total	36,865	7,152	44,017

⇒ In this particular case, the $2000 boundary is all that's relevant, but there could be cases when it's useful to look at the whole distribution

(e.g. under $1000,
$1000 to $1,499
$1500 to $1,999
$2000 to $2,500
⋮

9

loan amount	no PFF	PFF	
under 50	5%	3%	5%
50 to 99	17%	13%	16%
100 to 249	36%	28%	35%
250 to 399	26%	34%	26%
400 to 499	8%	13%	9%
500 + up	8%	10%	9%
total	1,544,118	130,722	1,674,840

↑ about 8% ↓

→ In this particular case, the ~~data~~ observations are "clustered in the middle" (think of a ~~bell~~ bell curve), but that won't always be the case

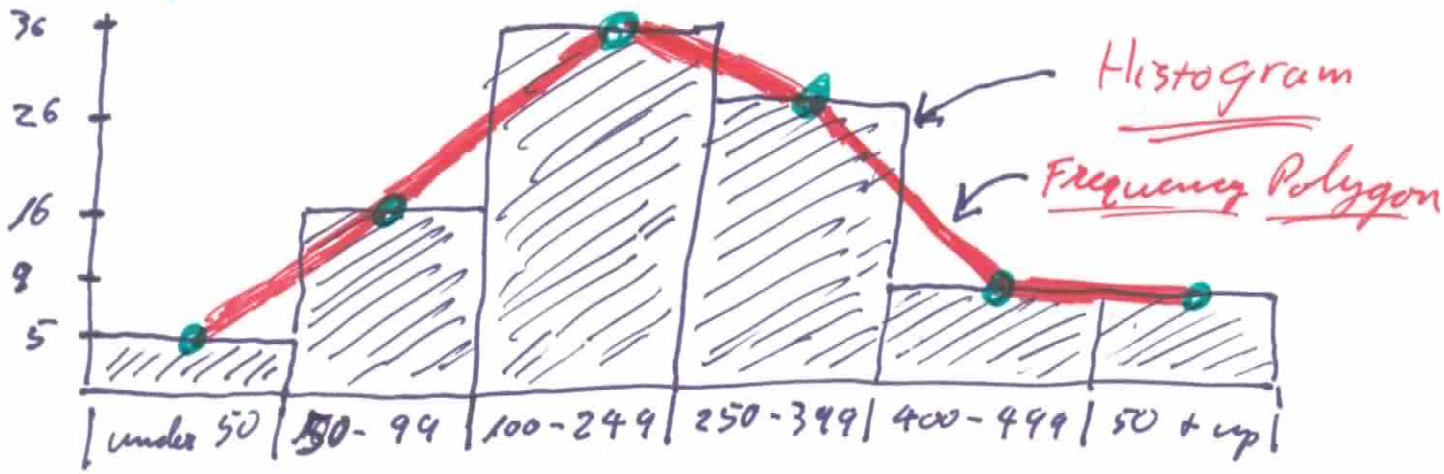
→ Here, few people borrow very small & very large amounts

→ Notice the ~~range~~ intervals of categories in table above — They're NOT even.

→ range of loan amounts is also quite large under 50 to 500 + up

→ "typical observation" — Is it 100 to 249? or is it 50 to 399? (78% in that range)

→ We can also graph the frequency distribution



loan amounts
(regardless of whether or not PFF)

→ Notice that the distribution is positively skewed (i.e. it's not symmetric)

→ when placing two frequency polygon on same graph, it's important to use the relative frequency (i.e. percentage)

otherwise one would be much higher than the other

percentiles

loan amount	relative frequency	cumulative relative frequency	loan amount
under 50	5%	5%	under 50
50 to 99	16%	21%	under 100
100 to 249	35%	56%	under 250
250 to 399	26%	82%	under 400
400 to 499	9%	91%	under 500
500 & up	9%	100%	all