Stat 202-2015 W1-Tues

Statistics is the sciene of learning from data.

Let's get some data: survey
what is you favorite color

~~primary colors~~ red, orange, yellow, green, blue, purple
black, brown, grey, white, other (pink?) (maroon?)

Cases are the objects described by a set of data

What are the cases here
students

This was confusing because we collected the data as a summary instead of

red	3
blue	1

bill	red
sally	blue
joe	red
pat	red

underlying data

~~StatCrunch~~

A variable is a chacacteristic of a case

What are variables here
name
favorite color.

Stat 202-2015S-W1-Tues

A label is a special variable used in some data sets to distinguish cases

Each case must have a unique label (otherwise it is not called a label)

What are labels in favorite color data set

name

Could also be (if we collected this data)

possible labels because unique

- Student number
- SSN
- phone number (unless two students share the same phone #)
- email address
- made up number (1...30)

Stat 202-2015S-W1-Tues (pg 3)

Types of variables

categorical variable
a variable that places a case into one of several groups or categories

quantitative variable
a variable that takes numerical values for which arithmetic operations such as adding and averaging make sense

Favorite color? Which is it? categorical
Weight in pounds of students in class? quantitative
Age? quantitative
Height quantitative
Result of following survey
I like to wear pink. categorical
- strongly agree
- agree
- neutral
- disagree
- strongly disagree.

Party Affiliation
Green, Democrat, Republican, Libertarian, Unaffiliated

Gender Identity: Male, Female, Neither, Both.

Stat 202 - 2015S - W1 - Tues

There is also another distinction between categorical variables

ordinal-categorical

have a natural order

eg. Strongly agree
agree
Neutral
disagree
Strongly disagree

nominal-categorical

in name only red, green, blue, black, grey, white

Non natural order.

? Grades A B C D F

nominal-categorical
ordinal-categorical
quantitative

note can be converted to categorical
A=4 B=3 C=2 D=1 F=0

conversion imposes equal distance

But this operation isn't natural. IS the distance Between A and B the same as between D and F=?

Stat 202 20155 – W1-Tues (pg 5)

The distribution of a variable tells us what values it takes and how often it takes these values.

Exploratory data analysis

(1) When we study a data set we study each variable by itself

Then

(2) move on to study the relationships among the variables

For each of these steps we

(a) Start with a graph or graphs

(b) Then add numerical summaries

Graphs For single variables (categorical)
* bar graph } shows
* pie chart } distribution

I am going to show you how with StatCrunch.

Stat 202-2015S-W1-Tues Pg 6

StatCrunch

* pie chart with summary with data

it is important that the pie add up to whole if deleting rows add other row

* bar chart with summary with data

not important with bar chart

* Loading files
* Value assending etc

Homework

1.14
1.16 < 1.22
1.23
1.27