Stat 202-2015XD-W1-Mon

What is Statistics?

Statistics is the science of learning from data.

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

red, orange, yellow, green, blue, purple, black, brown, grey, white other, e.g. pink, maroon, magenta,

Cases are the objects described by a set of data

What are the cases here?
Students present in stat 202 today,

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

summary

red 3
blue 4
white 2

instead of

full data set

bill → red
sally → blue
pat → red
erin → white

A variable is a characteristic of a case.

What are the variables here (in the full data set)

name and favorite color

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 labels are in full data set

name.

But two students can have the same name, especially first name) If we are collecting data we might want to use a different label

Student number, SSN, phone number email address, made up number (e.g. 1 to 30)

Types of variables

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

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

Favorite colors? which is it? Categorical
Weight in pounds of students in class
quantitative
Age; quantitative
Height: quantitative

Result of following survey:

I like pink: Strongly agree
agree
disagree
Strongly disagree
Party Affiliation: Green, Democrat, Republican, Libertarian
Unaffiliated
Gender Identy: Male, Female, Both, Neither

Another distinction is made between two different types of categorical variables

Ordinal - categorical variables have a natural order

Nominal - categorical variable are categories in name only, Lacks a natural order

Strongly agree
agree
disagree
strongly disagree
} have a natural order

colors: pink, grey, black, red are nominal categorical, because they lack a natural order

You can assign numbers to categories But if the number are arbitrary adding and averaging won't make sense

Consider grades A, B, C, D, F, A+ A⁻, B⁺, B⁻ Ordinal categorical

Or should we consider grades as quantitative, after all we compute GPAs.

To the extent that we compute GPAs grades are quantitative But many people argue that averaging for GPAs doesn't make sense Is a F and a D or a D and a C the same distance from each other as a A and a A?

They argue that grades should be considered as ordinal categorical and care should be taken interpreting GPAs.

Active Learning 1

Problems 1.14 and 1.16 on Homework #1

key concept

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

Exploratory data analysis – when we study a data set, we

(1) Study each variable by itself

then move on to

(2) study the relationships among variables

For both of these steps we

(a) start with a graph or graphs
(b) Then add numerical summaries.

The graphs chosen depend on whether the variable(s) are categorical or quantitative.

For single variables that are categorical there are two options

* bar plot } both gives
* pie chart } distributions (see defn)

StatCrunch

* pie chart with summary with data
  - Favorite Colors

It is important that the pie add up to a whole.

In other words if there are too many color lump the smaller ones into other

* bar ~~graph~~ Plot with summary with data

No longer need to add to whole

* Loading files
* Value ascending/etc

Homework: 1.22 – 1.27

Exploratory data analysis

Graphs for single quantitative variables

Stemplot, Histogram } again shows distribution

Stemplots

Homework #2

Explain and do.