Introduction to Statistics

CSU, Chico Math 314

2018-08-27

outline

outline

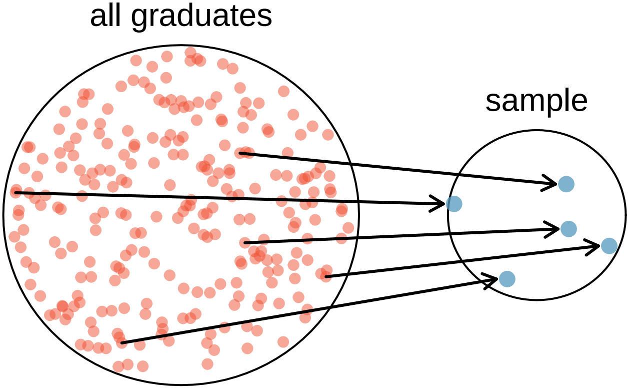
The Goal of Statistics

Statistics seeks to describe characteristics of a broad group using only a subset of information. To do so we will study how to analyze and interpret data.

The Goal of Statistics

Example

Making statements about all of Chico's graduates would be difficult; we'd first have to find them all and then extract data from each person. Instead, statistics uses a sample of all graduates to infer characterstics about them.



Population

Statistics generalizes this idea of the broader group with the word **population**.

population

The set of all relevant objects of interest.

Population

Example

- all of Chico's graduates
- all U.S. adults
- all Gmail users
- ...

Sample

Statistics generalizes the idea of the subset of objects of interest, with the word **sample**.

sample

Any subset of the relevant objects of interest, preferably chosen randomly.

Sample

Example

- alumni from the class of 2008
- California citizens
- Gmail users before 2007
- ...

The Goal of Statistics, redux

Statistics seeks to describe characteristics of a **population** using a **random sample** from that population. To do so we will study how to analyze and interpret data.

Nota Bene (N.B.)

Random samples naturally lead to **sampling variability** – the fact that characteristics of interest will vary from sample to sample.

outline

Data Types

We'll focus on two broad types of data.

categorical

Categorical data can be categorized or placed into non-overlapping groups. Possible values of a categorical variable are called levels.

numerical

Numerical data is quantitative; e.g. it takes on numerical values and all mathematical operations $(+, -, *, /, <, >, =, \ldots)$ make sense with these values.

Data Types in R

Consider a dataset about and named `email`. Here is one of many ways to read data into R

```
url <- "https://roualdes.us/data/email.csv"
suppressMessages(email <- read.csv(url))
```

Data Types in R

Here are some columns from the dataset `email`

	spam	num_char	format	number
1	0	11.37	1	big
2	0	10.50	1	small
3	0	7.77	1	small
50	0	14.43	1	small

Table 1: Four rows of data from the email data set

Data Types in R

categorical

R calls a categorical variable a factor, and maintains the word **levels** to mean the values the factor can take on.

```
is.factor(email$number)

## [1] TRUE
```

Data Types in R

The variable spam should reasonably be considered a factor.

```
# R doesn't agree and that's probably good
is.factor(email$spam)

## [1] FALSE

is.numeric(email$spam)

## [1] TRUE

head(email$spam) # 0 = not spam, 1 = spam

## [1] 0 0 0 0 0 0

is.factor(factor(email$spam)) # coerce to factor

## [1] TRUE
```

Numerical Data Types in R

There are two sub-types of numerical data: discrete and continuous. R calls a numerical variable of either sub-type numeric.

```
is.numeric(email$num_char)

## [1] TRUE
```

Data Storage in R

The data in Table 1 represent a **data frame**, which is the way to organize data for statistical analysis[1]. Each row represents a new **observation** (or **case**) and each column represents a new **variable**. More observations are added to the data set by appending rows, and more variables are added by appending columns.

[1] The article *Tidy Data* by Wickham [2014] provides an excellent, and thorough, discussion of proper data organization.

outline

The Goal of Statistics, redux

Statistics (the discipline) seeks to describe characteristics of a **population** using a random **sample** from that population.

parameter

Characteristics that describe a population are called population parameters.

statistic

Characteristics that describe, and are calculated from, a sample are called statistics.

Summarizing Data with (Summary) Statistics

Different data types require different statistics.

- categorical data
  - proportions
- numerical data
  - mean
  - median
  - percentiles
  - variance / standard deviation
  - IQR

Proportion

The **sample proportion** summarizes multiple observations of a categorical variable; divide the number of "successes" by the number of observations:

$$\hat{p} = \frac{\#successes}{\#observations}.$$

Proportion, example

Remember that `R` didn't think the variable `spam` was a factor. Here's why that helps.

```
phat <- sum(email$spam)/length(email$spam)
phat == mean(email$spam)

## [1] TRUE
```

Mean

The **sample mean** gives a summary of the middle of multiple observations of a numeric variable; add up all the numbers and divide by however many there are:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Mean, example

From the dataset `email`, we can estimate the (population) mean number of characters in an email using the sample mean

```
mean(email$num_char) # sample mean, xbar

## [1] 10.70659
```

Median

The **sample median** gives a (different) summary of the middle of multiple observations of a numeric variable. There is no simple mathematical expression for it. Hence, we think of the median as the number in the middle of the ordered observations, dependent on the sample size n:

- odd – median is $(n+1)/2$th observation
- even – median is the mean of the $n/2$th and $(n/2)+1$th terms.

Median, example

From the dataset `email`, we can estimate the (population) median number of characters in an email using the sample median

```
median(email$num_char) # sample median

## [1] 5.856
```

Motivating The Percentile

The median puts half, 50% of the data below it. Let's generalize this idea. One could just as easily put 25%, 75%, 33%, or any other percentage of (ordered) data below the number of interest.

Percentile

The **sample percentile** is the number that puts $p\%$ of the (ordered) observations below it.

- The 10% percentile puts 10% of the observations below it.
- The 33% percentile puts 33% of the observations below it.
- The 98% percentile puts 98% of the observations below it.

Quartiles

We reserve special names for the 25th, 50th, and 75th percentiles. We call these **quartiles**. Further, the quartiles are often denoted Q_1 for the 25th percentile, Q_2 for the median, and Q_3 for the 75th percentile.

Percentiles, example

R has the function quantile, which accepts the probabilities associated with the percentiles of interest.

```
quantile(email$num_char, probs=c(0.25, 0.5, 0.75))

##    25%    50%    75%
##  1.459  5.856 14.084
```

Quickly Summarize Data

The R function summary will produce a 6 number summary of a variable for you:

```
summary(email$num_char)

##    Min. 1st Qu.  Median    Mean 3rd Qu. 
##   0.001   1.459   5.856  10.707  14.084 
##    Max. 
## 190.087
```

Different Statistics Tell Us Different Things

Up until now, all the statistics we've considered measured the center of the data, i.e. where the data is located.

- re `num_char`: something near 5 or 10
- U.S. adult heights
- weight of dogs

Different Statistics Tell Us Different Things

Next, we'll look at how wide the data are. We generally refer to these statistics as measures of spread. These statistics no longer tell us about where, instead about how variable.

Variance / Standard Deviation

The most common measures of spread are the **variance** and the **standard deviation**. Think of the variance as the average squared distance away from the mean. The standard deviation is the square root of the variance, which essentially gives us interpretable units (no longer squared).

Sample Variance / Standard Deviation

Assume we have a sample of n data points $x_1, x_2, \ldots, x_n$.

sample variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

sample standard deviation

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Sample Variance / Standard Deviation, take 2

You might see elsewhere, e.g. in `R`, the following definition for the sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

This is technically called the unbiased sample variance – we'll mention bias a bit later.

Sample Standard Deviation, in R

R will happily do calculate the s, the sample standard deviation, for you.

```
sd(email$num_char) # sample Standard Deviation

## [1] 14.64579
```

Interquartile Range

The `interquartile range` (IQR) is defined as the difference between the third and the first quartile:

$$IQR = Q_3 - Q_1.$$

Interquartile Range

In R, we can calculate the IQR with the function IQR.

```
IQR(email$num_char)

## [1] 12.625
```

This *one* number tells us how far away Q_3 is from Q_1, hence it measures spead (width).

outline

Summarizing Data with Plots

Different data types require different plot types

- categorical data
  - table
  - bar chart
- numerical data
  - histogram
  - box plot
  - scatter plot
  - density plot (later)

Tables

Categorical data doesn't support mathematical operations; e.g. orange + apple = ?. The best we can do is count observations

```
table(email$spam)

##
##    0    1
## 3554  367

  prop.table(table(email$spam))

##
##          0          1
## 0.90640143 0.09359857
```

Two-way Tables

Two-way tables are similarly simple to make.

```r
round(prop.table(table(email$spam, email$number)), 2)

##
##      big none small
##   0 0.13 0.10  0.68
##   1 0.01 0.04  0.04
```

Towards Plots

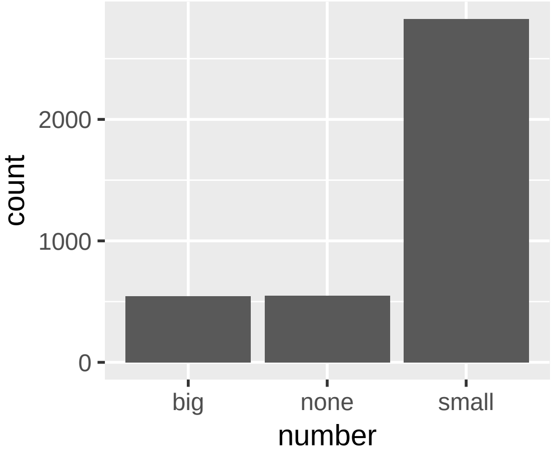
Throughout this course, we'll use the `R` library `ggplot2`. Specifically, we'll call on most the function `ggplot()`, contained within `ggplot2`, which the `R` community writes idiomatically as `ggplot2::ggplot`.

```
library(ggplot2) # first load the library
# ?ggplot        # help files
```

Bar Charts

A **bar chart** plots the number of observations of each level of a given categorical variable.

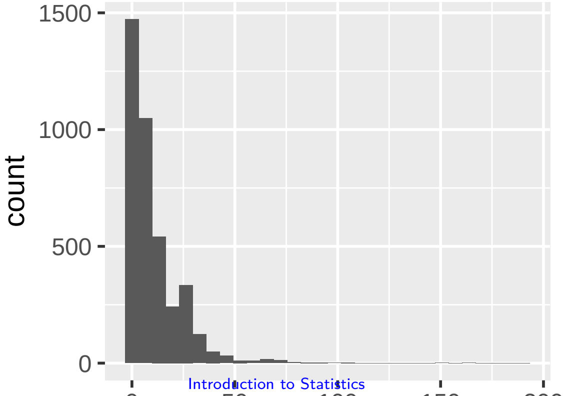
```
ggplot(data=email, aes(number)) + geom_bar()
```



Histogram

A **histogram** plots the numbers of observations that fall into numerical bins, the width of which is user determined.

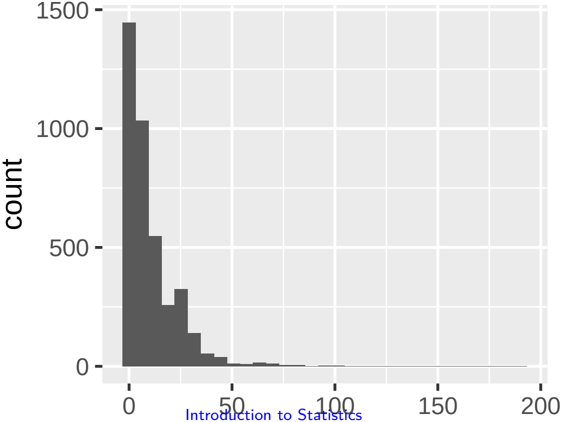
```
ggplot(data=email, aes(num_char)) + geom_histogram()

## `stat_bin()` using `bins = 30`.  Pick
## better value with `binwidth`.
```



Histogram

As you can see, ggplot reminds us that we need to choose a binwidth.

```r
# guess where 31 comes from
ggplot(data=email, aes(num_char)) +
    geom_histogram(bins=31)
```



Histogram, notes

Some key facts about histograms:

- bin width choice is inherently subjective, choose well
- the x-axis is numeric, not categorical as for bar charts
- great way to measure center (location), spread (width), and `skew`

Skew

The **skew** of a dataset measures asymmetry. Data with long right tails are said to be **right skewed**, while data with long left tails are said to be **left skewed**. What is the variable `num_char` above?

Mean v. Median, skew

Skew tells us things about the relationship between the mean and the median. The variable `num_char` is right skewed, and notice

```
mean(email$num_char)

## [1] 10.70659

median(email$num_char)

## [1] 5.856
```

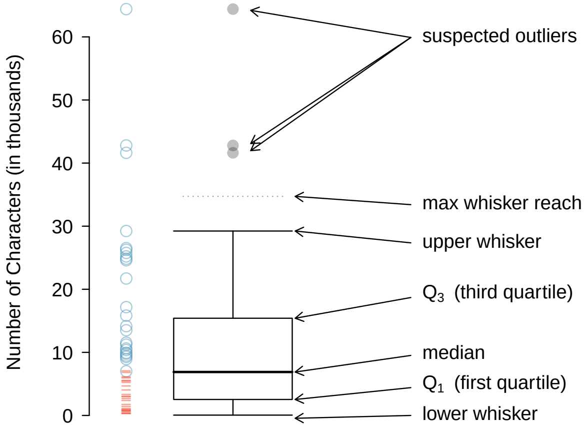
Mean v. Median, skew

More on skew:

- What is a variable called where there is no skew?
- If the data were left skewed, which is bigger the mean or median?
- Notice that the mean is pulled toward the extreme data. Why?

Box Plot

A **box plot**[2] is a great way to visualize something close to the 6 number summary produced by `R`'s function `summary`.

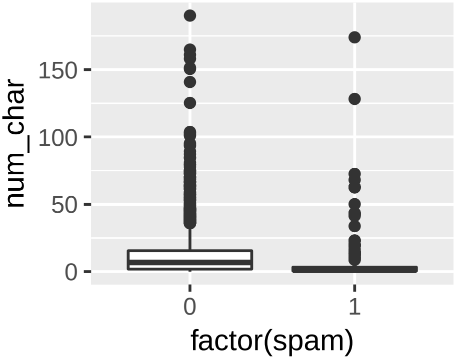


[2] Wikipedia provides some of the more common variations on the whiskers of a box plots.

Box Plot

It is more common for box plots to have a numerical variable on the y-axis and a categorical (recognized as a factor) variable on the x-axis.

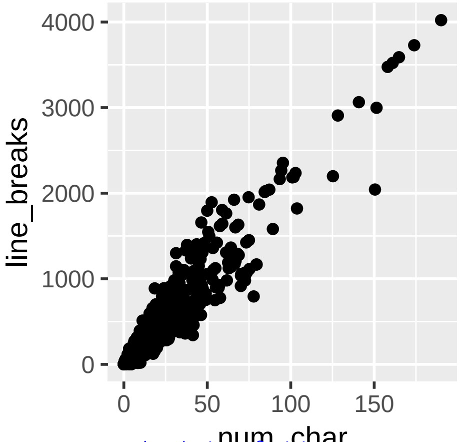
```
ggplot(data=email, aes(factor(spam), num_char)) +
    geom_boxplot()
```



Scatter Plot

A **scatter plot** plots two numerical variables, observation by observation, against each other.

```
ggplot(data=email, aes(num_char, line_breaks)) +
    geom_point()
```



Relations Between 2 Numeric Variables

We often describe the relationship between two numeric variables. The previous plot is described as positive and linear. What does a negative nonlinear relationship look like?

Plots, advanced

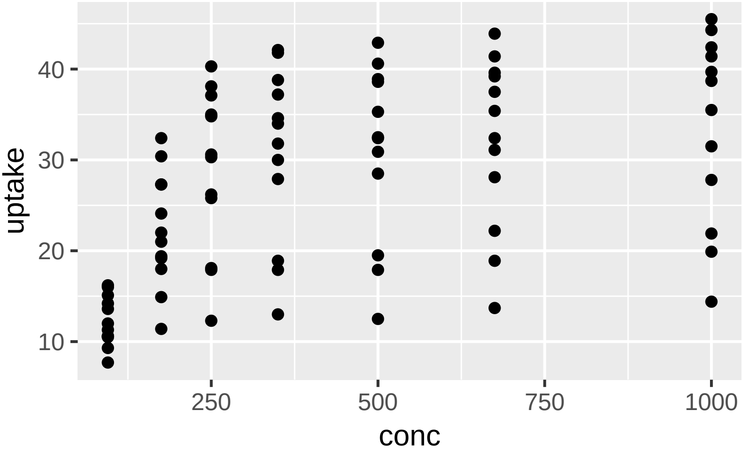
Meshing the ideas of two different plots into one plot can help elucidate the story behind the data. For instance, consider the dataset `datasets::CO2` where some ecologists measured the carbon dioxide `uptake` ($\mu mol/m^2$) of different grass species as related to ambient CO2 concentrations (`conc` mL/L).

```
head(CO2[,c("uptake", "conc")]) # only first 6 obs

##   uptake conc
## 1   16.0   95
## 2   30.4  175
## 3   34.8  250
## 4   37.2  350
## 5   35.3  500
## 6   39.2  675
```

Plots, advanced

Because we have two numeric variables, `uptake` and `conc`, we should think scatter plot.

```
ggplot(data=CO2, aes(conc, uptake)) + geom_point()
```



Plots, advanced

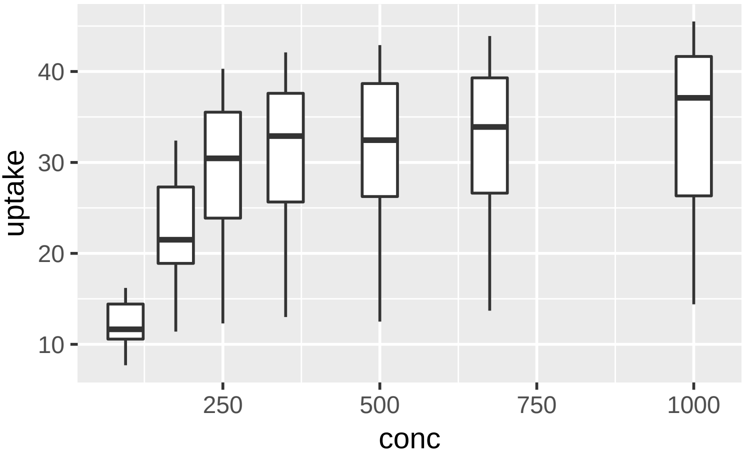
But, `conc` only has 7 values so we can think of it as a categorical variable.

```
length(unique(CO2$conc))

## [1] 7
```

Plots, advanced

So, how about box plots by “group”?

```
ggplot(data=CO2, aes(conc, uptake, group=conc)) +
    geom_boxplot()
```



outline

Comparative Studies

So far, we've described the words / tools researchers use relative to general data sets. More often researchers want to navigate the world of causality and they have a set of special words to help them compare differences between groups.

Explanatory and Response Variables

In studying the relationship between two variables, the variables are often viewed as either a **response variable** or an **explanatory variable**. To identify the explanatory variable in a pair of variables, ask yourself which of the two explains the other. Often, many variables will explain/predict the response variable.

Explanatory and Response Variables

response variable

The response variable is the variable or characteristic of the data that we are wanting to learn about (to explain, to predict, or to estimate).

explanatory variable

The explanatory variable is the variable that does the explaining, or whose effect on the response variables is of interest.

Explanatory and Response, examples

Identify the explanatory and response variables from the following.

1. fertilizer and growth
2. college grade point average and high school grade point average
3. average federal spending and counties with high rates of poverty
4. police department budget and crime rate
5. ...

Explanatory and Response, caution

Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables. We use these labels only to keep track of which variable we suspect affects the other.

Explanatory Variable $\xrightarrow{\text{might affect}}$ Response Variable

Observational and Experimental Studies

There are two primary types of data collection: observational studies and experiments.

observational study

The researcher simply monitors and collects data on things as they are, by observing. There is no manipulation of the study by the researcher.

experiment

The researcher assigns the value of the explanatory variable for each unit. In other words, the researcher controls which subjects go into which treatment groups.

Observational Studies, examples

In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

Experimental Studies, examples

Experimental studies, through the direct manipulation from the researcher, can provide cause-and-effect relationships between the response and explanatory variable.

- (insert context) ... researchers collect a sample of individuals and split them into groups. The individuals in each group *assign* a treatment, one group per level of the explanatory variable.
- To study the effect of tar contained in cigarettes researchers painted tobacco tar on the back of some mice but not others, and recorded if the painted mice had cancer at a higher rate than those not exposed to the tar [Wynder et al., 1953].

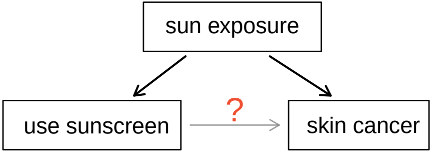
Observational and Experimental Studies, identified

For each of the following situations, identify if it is an observational study or an experiment.

- Review medical or company records to attempt to identify fraud.
- Follow a group of many similar individuals to study why certain diseases might develop.
- Plant a specific type of native grass in select areas to see if the native species will out-compete an invasive species.

Confounding Variables

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer? Or is there another variable we aren't accounting for?



Confounding Variables

confounding variable

Confounding variables, or confounders, are variables that are correlated with both the explanatory variable and the response variable. These variables are also known as lurking variables, because they are not always easily seen.

outline

Take Away

- **random samples** help us make inferences about the **population** of interest
  - we'll later use the inherent variation across random samples
- Data has types analogous to CS types – be able to reconcile them
- **statistics** are functions calculated from random samples
  - know at least all the statistics mentioned in this presentation
- many plot types
  - know at least all the plots mentioned in this presentation
- The language of statistical studies is crucial for communicating across disciplines

outline

references I

Hadley Wickham. Tidy data. *Journal of Statistical Software*, 2014.

Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. " O'Reilly Media, Inc.", 2016.

Ernest L Wynder, Evarts A Graham, and Adele B Croninger. Experimental production of carcinoma with cigarette tar. *Cancer Research*, 13(12):855–864, 1953.