Chapter 1: Stats Starts Here

D. Raffle
5/18/2015

1.1 What is Statistics?

Wiktionary defines statistics as:

Noun

statistics (plural statistics)

- (*singular*) A mathematical science concerned with data collection, presentation, analysis, and interpretation.
- (*plural*) A systematic collection of data on measurements or observations, often related to demographic information such as population counts, incomes, population counts at different ages, etc.

… so why do you have to take this class?

Why Statistics?

Statistics is used in many fields, but generally speaking, we use statistics to:

- *Describe* and *investigate* the world around us using *data*.
- Give us *formal and reproducable methods* to deal *uncertainty* and *randomness*.
- Detect *patterns* or *features* in our data that we can generalize to larger populations.
- Test hypotheses.
- Make predictions.

An often used example is that we use statistics to separate the *signal* from the *noise* in our data.

Example: Gender and Height

- We might hypothesize that women are generally shorter than men.
- The only way to know this *for certain* would be to find the heights of all men and all women who have ever existed and compare the averages. Obviously, this isn't feasible.
- An alternative would be to get a *sample* of men and women and compare the average heights within the sample. To do this, we'd need:
- To come up with a strategy to get a representative sample
- To account for the fact that we could have randomly gotten a group of tall women and short men (or vice versa).
- This is something you will know how to do at the end of this course.

Example: Heights Across Generations

- In addition to gender, we might want to know if the heights of your parents have anything to do with your height. In addition to simply knowing *if* there is a relationship, we might want to use this information to predict the height of your children.
- A person's height is obviously determined almost completely by your genetics, and we might get a nearly perfect prediction by sequencing your genome. So what's the downside?
- Genetic analysis is expensive, and each person would need to be analyzed individually. A better option might be to take a sample of families and build a statistical model to analyze the relationship.
- We will cover this topic in chapters 6 and 7.

Example: Google

- Google makes money with AdSense, a technology that (is supposed to) only show you adds you might be interested in.
- To do this, Google keeps track of your search history, browsing history, keywords appearing in your emails (if you use Gmail), among other techniques.
- They use this information to build a model of your personality and demographics and choose the ads they think you will be interested in. For example, they might look at other users with similar profiles and see what they've clicked on, with the assumption that you are also likely to be interested in these ads.
- Facebook operates in a similar manner, looking at the pages you and your friends have Liked and using that information to show you sponsored pages.
- We will *not* be talking about these techniques in this course.

Example: Netflix

- While companies like Google and Facebook make money by selling your information, others use it to make recommendations and keep your business.
- Any time you rate a movie you've watched on Netflix, they use this information to predict what other movies you'll like so you keep watching (and paying for an account).
- Each movie has a list of attributes (Drama, Comedy, Suspenseful, etc.), and they will try to match other movies that share these characteristics.
- Pandora does the same thing, but using attributes of the songs (genre, key, mood, etc.).
- These are also models we will *not* get to cover in this course.

1.2 Data

- We've used the term data several times, and most people have some understanding of what data is.
- Since statistics is in many ways the science of data, we need a solid understanding of data to form a foundation for the course.
- The key characteristic of data is that it *answers questions*. A random collection of measurements is meaningless without *context*.
- In general, data is made up of *variables* that describe *individuals*, *observations*, or *cases*.

For example:

- The heights, weights, gender, and other medical stats of children born at a particular hospital
- The fuel efficiency, make, model, weight, class and horsepower of cars being sold in a given year

The Six W's

Anytime we look at a data set, we should be able to answer the "Six W's"

- **W**ho are the individuals (human or otherwise)?
- **W**hat is being measured (what are the variables)?
- **W**hen was the data collected?
- **W**here was the data collected?
- **W**hy was the data collected?
- Ho**w** was the data collected?

Without this information, it is impossible to know if the data is meaningful.

- Data collected by prohibitionists in the 1920s may not be informative about marijuana legalization today.

Example: Cars

Consider this table:

```
##      C_1 C_2 C_3 C_4   C_5   C_6 C_7 C_8
## R_1 21.0   6 160 110 2.620 16.46   1   4
## R_2 21.0   6 160 110 2.875 17.02   1   4
## R_3 22.8   4 108  93 2.320 18.61   1   4
## R_4 21.4   6 258 110 3.215 19.44   0   3
```

Is this data? How many Individuals are there? What about variables?

- Without context, this is just a collection of numbers.
- We don't know if there are 4 observations and 8 variables or if it's the opposite.

Example: Cars (cont.)

Now consider this table:

```
##                mpg cyl disp  hp    wt  qsec am gear
## Mazda RX4      21.0   6  160 110 2.620 16.46  1    4
## Mazda RX4 Wag  21.0   6  160 110 2.875 17.02  1    4
## Datsun 710     22.8   4  108  93 2.320 18.61  1    4
## Hornet 4 Drive 21.4   6  258 110 3.215 19.44  0    3
```

Is this data?

- There are four individuals (rows) and eight variables are recorded for each (columns). This is the typical convention in statistics.
- We can see that each individual is a specific car model, and each variable contains some performance measure.
- The measurements were collected by *Motor Trend* magazine in 1974.
- It's not perfect – we still don't know all the variable definitions.

1.3 Variables

We've seen that data is made up of observations (individuals), and characteristics of the observations are described by variables.

There are four major types of variables we will discuss:

- Categorical
- Numeric
- Identifiers (ID)
- Ordinal Variables

Why make the distinction?

- Each variable type needs to be treated differently when we describe and analyze our data.

Categorical Variables

Categorical variables divide our observations into groups or categories.
Categorical variables can also be called:

- Qualitative Variables
- Nominal Variables
- Factors

The values a categorical variable can take are often called *levels*, *classes*, or *labels*.

- Gender has the classes "Male" and "Female"
- Transmission has the levels "Manual" and "Automatic"

We typically summarize categorical variables by counting the number of individuals at each level or finding the proportions/percentages of individuals with that level.

Numeric Variables

Numeric Variables are measurements of our data with units. For example:

- The number of average customers Subway locations serve in an hour
- The weight of cars in 1000s of pounds
- The price of milk across different cities in US dollars
- Students' grades as percentage of total possible points

We typically summarize numeric variables by:

- Their center: What is a typical or average value?
- Their spread: How far apart can we expect individuals to be?

Identification (ID) Variables

ID variables are simply used to uniquely label each individual observation.

- Student ID numbers
- Social Security numbers
- First and last name
- Email address
- License plate numbers

We typically don't use ID variables in analysis, but to keep our data organized.

Ordinal Variables

Ordinal variables are similar to categorical variables, but they have some natural order to them.

- Shirt sizes S, M, L, XL
- Letter grades A, B, C, D, F
- Star ratings of restaurants
- Likert Scales (Strongly Agree, Mostly Agree, Agree, Neutral, Disagree, Mostly Disagree, Strongly Disagree)

Interpreting ordinal variables can be tricky.

- We can describe an average number of stars a movie receives on Netflix, but what exactly is a "star" as a unit?
- Is the difference between L and XL the same as the difference between S and M shirts?

Deriving Variables

We can often derive other variable types from numeric variables, or vice versa

- We can use individuals' ages to define them as "Child", "Teenager", "Young Adult", "Middle Aged", and "Senior"
- Grades as percentages can be used to define letter grades
- We can represent Automatic or Manual as 1s and 0s

Deciding how to treat variables depends on the questions we are asking and the types of analyses we are going to run.

Some variable types can be more difficult to identify:

- What type of variable is a phone number's area code?

Review

The car data set we saw earlier had eight variables, what are their variable types?

- Miles per Gallon
- Number of Cylinders
- Engine Displacement (Cubic Inches)
- Gross Horsepower
- Weight (lb/1000)
- 1/4 Mile Time (sec.)
- Transmission (0 = auto, 1 = manual)
- Number of Gears

What's Next?

In the next section, we will see how to summarize a data set by:

- Finding numerical summaries of numeric and categorical variables
- Visualizing categorical and numeric data