Introduction to Statistics

Introduction

Let’s say you want to determine Nashvillian’s favorite hot chicken restaurant.

You stop 30 people on 2nd Avenue and ask whether they prefer Prince’s or Hattie B’s and get the following response:

Restaurant	Number of Responses
Prince’s	18
Hattie B’s	12

Can we conclude that Nashville’s preferred hot chicken restaurant is Prince’s?

Statistics

- The study of how best to collect, analyze, and draw reliable inferences from incomplete, noisy, or otherwise imperfect data
- The science of learning from data in the presence of **uncertainty** (meaning we cannot know the answer for sure)
- A set of procedures to make inferences that extend beyond the data to the broader population.

Statistics

Descriptive Statistics: describing/summarizing a data set using just a single number or a few numbers

Inferential Statistics: saying something about a population based on examining a subset (sample) from it (and expressing how confident we are in our inferences)

Sample vs Population

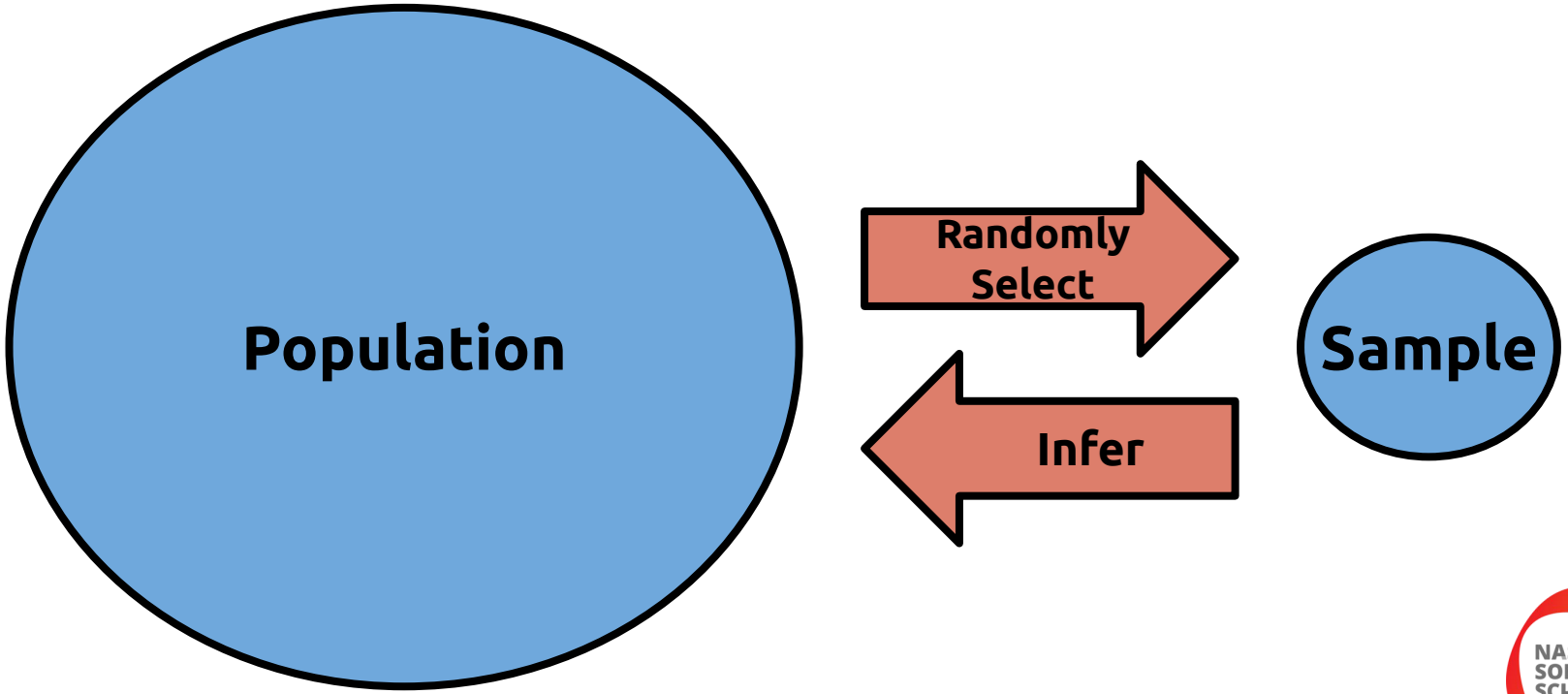
Population:

- The entire collection of items **of interest**
- What we wish we knew (How effective will this drug be?)

Sample:

- A subset of the population selected for study (or available for study)
- What we know (How effective was the drug on the participants in our study?)

Sample vs Population



NASHVILLE SOFTWARE SCHOOL

Sample vs Population

Example: We want to ensure that we have enough emergency responders scheduled for this upcoming Friday. How many accidents can we expect there to be reported on Friday?

Population of interest: Crashes that take place on (all) Fridays

We select 12 Fridays at random (our *sample*) and observe the following number of crashes:

98	101	105	107
108	113	115	116
117	128	153	154

Sample vs Population

Parameter:

- The exact thing we wish to know about the population
- Eg. The average number of crashes on Friday in Davidson County

Statistic:

- An estimate of the parameter based on the sample
- Here, we could look at a number of *descriptive statistics:*
  - Mean: 117.9
  - Median: 114
  - Standard Deviation: 18.4
  - 80th percentile: 126

Sample vs Population

Big idea: We are trying to glimpse the population through the “keyhole” of our sample.

We want to say something that extends beyond the data that we have, but do so in a *rigorous* way.

It is easy to find patterns in sample data. The real question is whether those patterns exist beyond the sample data and in the general population.

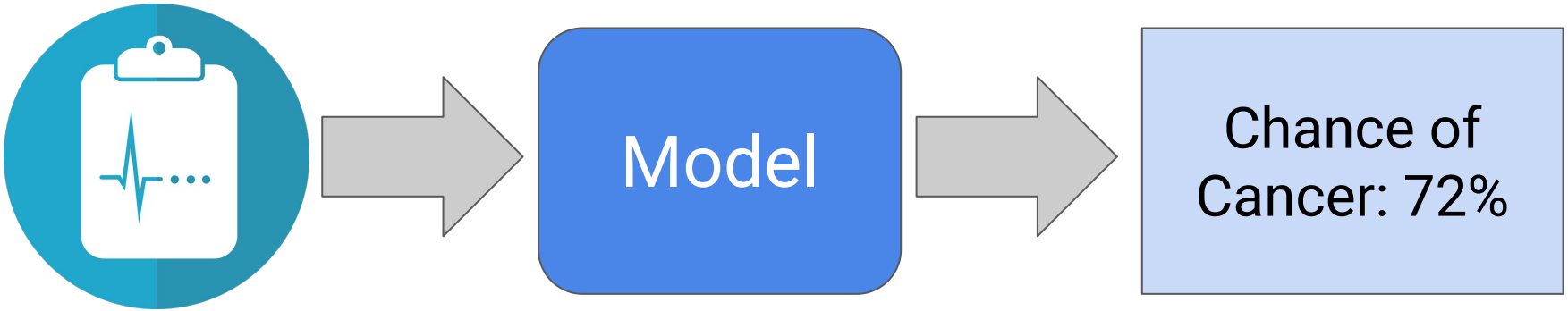
Data Analysis vs Statistics

Data Analysis	Statistics
Looks at what has happened historically	Infers/generalizes/extrapolates
Deals with facts	Deals with uncertainty
Uncovers patterns in the data	Determines if patterns carry over to the broader population of interest
Forms hypotheses about data	Tests hypotheses about data
Encounters and surfaces potential insights in data	*Safely* comes to conclusions beyond the data at hand

Statistics vs. Data Science

Both data scientists and statisticians want to extract knowledge from data.

Both data scientists work with models.



Input: patient's medical information

Output: Predicted probability of cancer

Statistics vs. Data Science

The key distinction comes from how the model is treated.

Statisticians care about *inference*:

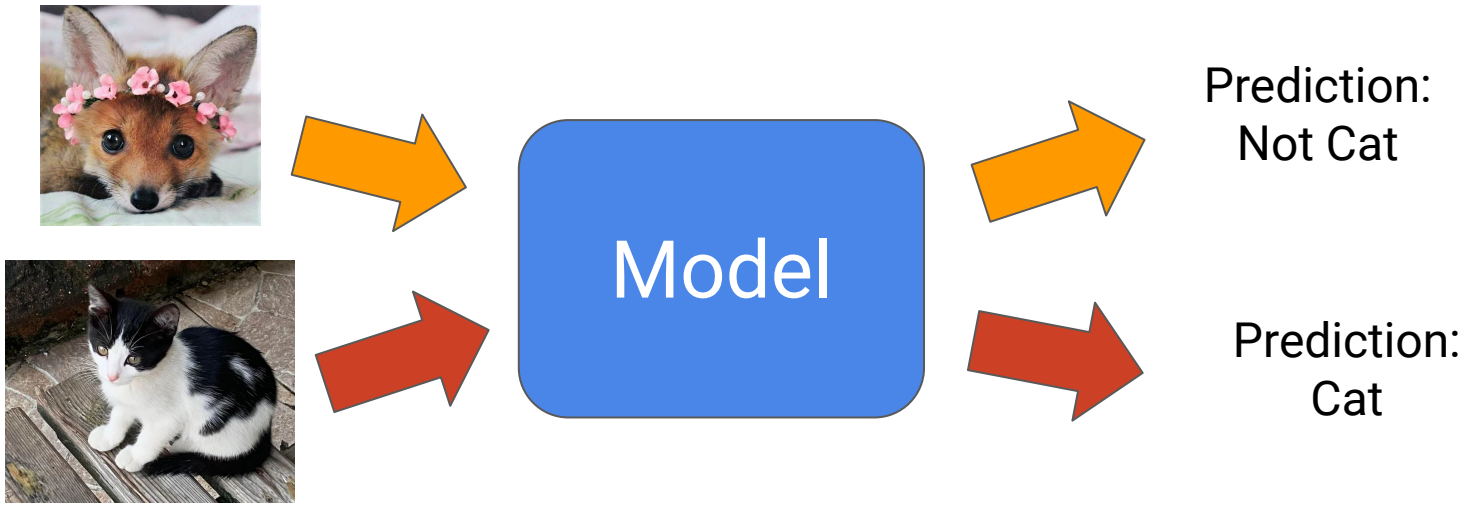
- How does “nature” associate the inputs to the output?
- What is the *mechanism* that generates the outputs?

Data Scientists care about making accurate *predictions*:

- What will the output be for future input values?

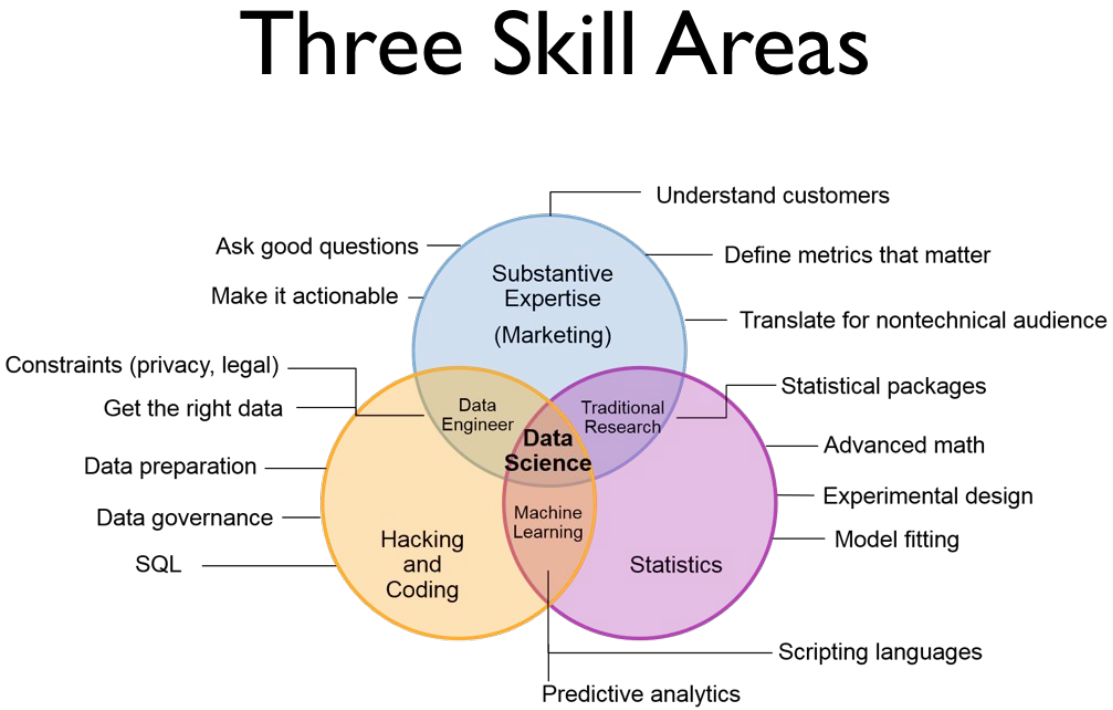
Statistics vs. Data Science

The focus on prediction over inference allows data scientists to create models for outputs for which a statistical model would be very difficult.



Statistics vs. Data Science

A data scientist must also be knowledgeable about coding and must be proficient in dealing with data: gathering, storing, cleaning, etc.



Designed Experiment vs. Observational Study

Designed/Randomized Experiment:

- Involve a treatment (eg. a new drug) and a response to that treatment (Did the patient recover?)
- The researcher *imposes* the treatment on the subjects
- Eg. The researcher randomly assigns one group to receive a new drug and another to receive a placebo. The research will compare the outcomes for each group.

Designed Experiment vs. Observational Study

Observational Study:

- The treatment is not assigned by the researcher, but is simply observed by the researcher
- Data is collected in a way that does not interfere with the data generation process
- Eg. Having participants fill out a survey and self report health outcomes

Designed Experiment vs. Observational Study

When doing data analysis/data science, you are often working with data where you have no control over the generation process.

That is, you will almost always be working with observational data.

Hence, a data analyst/data scientist must think carefully about the source of their data to avoid reaching incorrect conclusions.

Causation

Usually we want to be able to make conclusions about causation.

- Does higher salt consumption *cause* elevated blood pressure?
- Does our new drug *cause* patients to recover from a particular disease?

Randomized Experiment vs. Observational Study

Observational studies can provide evidence of an *association* between variables, but cannot be used to prove causality.

Why? In observational studies, there may be some systematic difference between the group of subjects that received the treatment and those that did not.

Randomization “spreads out” the differences between groups.

Determining Causation

Does smoking *cause* lung cancer?

We look at the data and observe that people who smoke have higher rates of lung cancer, so there is definitely an association.

But, maybe there's a gene that predisposes people to both smoking and getting lung cancer.

Or maybe people who smoke tend to follow unhealthy lifestyles, and the other lifestyle choices are the real cause of lung cancer.

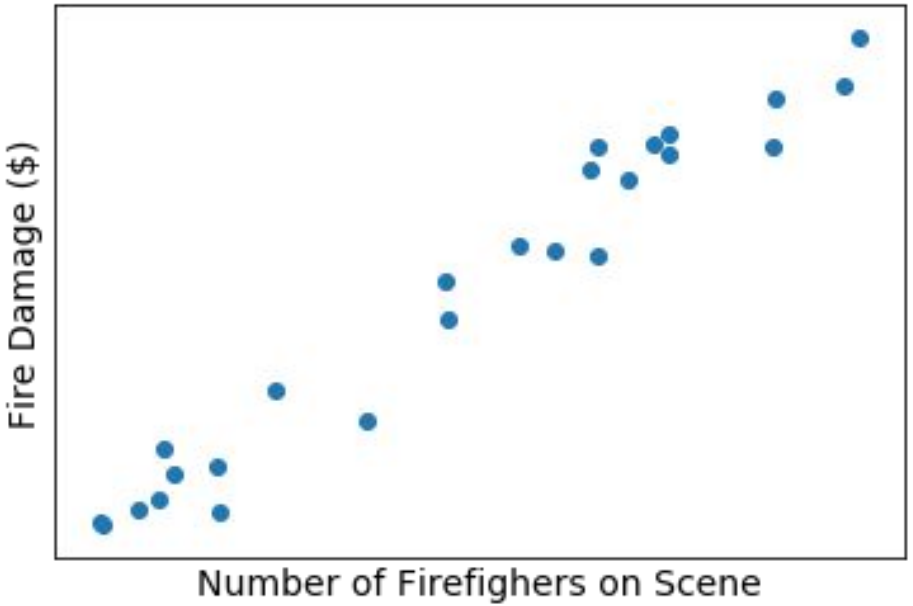
Determining Causation

Does smoking *cause* lung cancer?

Unfortunately, for this type of question, we are pretty much limited to doing an observational study.

After all, we can't randomly assign people to smoke or not smoke!

Correlation and Causation



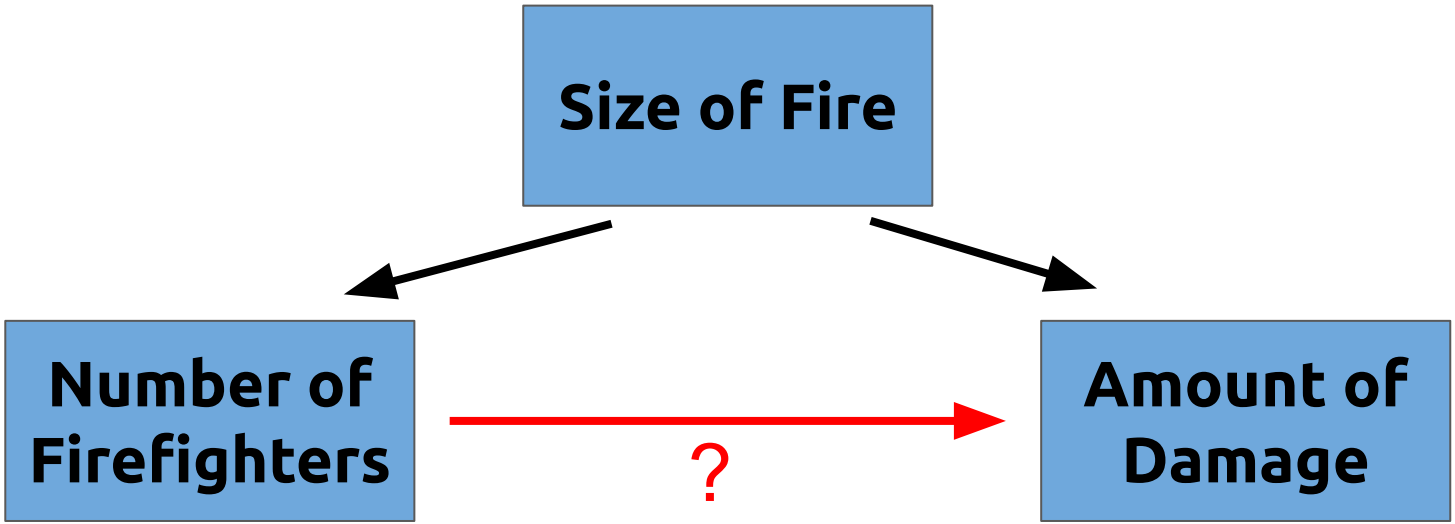
There is a strong correlation between number of firefighters and fire damage. That is, the fires with more firefighters tend to have more fire damage.

Can we conclude that having more firefighters on scene *causes* more fire damage?

If so, we should start dispatching fewer firefighters!

Correlation and Causation

Of course, this is ridiculous.
For larger fires, there will be more firefighters on scene, and more fire damage.



Confounding Factors

These hidden third variables (genes, lifestyles, or fire size) are called **confounding factors** or **lurking variables**.

Confounding factors are correlated with both the treatment and the effect, and make it difficult to infer any relationship between the two.

We can mitigate the effect of confounding variables through careful randomized experiment designs.

Bias

Bias (in the context of statistics) is the degree to which a procedure *systematically* overestimates or underestimates a population value.

A lot of bias can be eliminated through careful research design and proper sampling.

A random sample where all members of the population are equally likely to be part of the sample is likely (but not guaranteed) to produce a mix of observations which looks like a miniature version of the population.

Bias

Bias can come from not properly sampling or by choosing a sample that does not represent your *entire* population of interest (**selection bias**)

Experiments in the cognitive sciences often have samples which overrepresent W.E.I.R.D. individuals (White, Educated, Industrial, Rich, Developed)


Science News

from research organizations

Brain imaging results skewed by biased study samples

Neuroimaging researchers must consider who their samples represent, study says

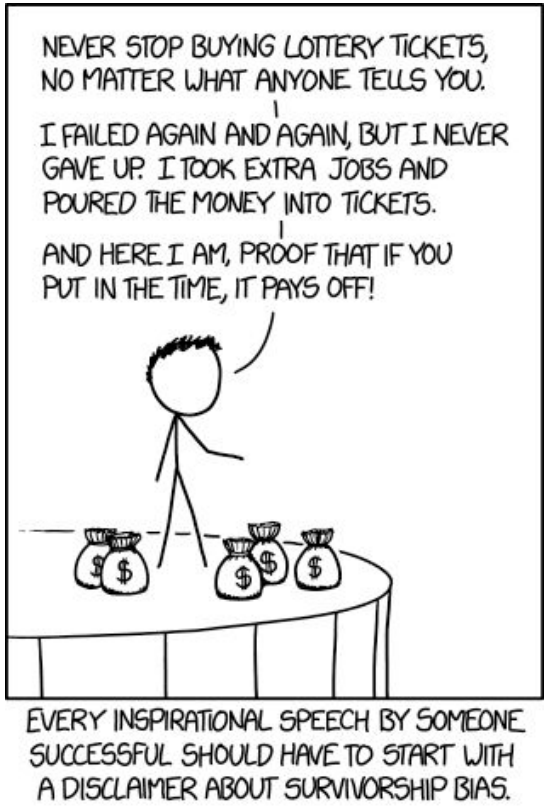
https://www.sciencedaily.com/releases/2017/10/171012090954.htm

NASHVILLE SOFTWARE SCHOOL

Bias

Survivorship Bias is a form of selection bias, where a study concentrates on the people or things that made it past some selection process and overlooks those that did not.

It can also lead to the false belief that the successes in a group have some special property, rather than just coincidence.



NASHVILLE SOFTWARE SCHOOL

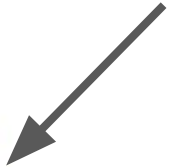
Bias

Highrise Syndrome in Cats (Cats that Fall from Heights)

To a cat, a window may look like a path to freedom, but to many it leads to injury and death from upper story apartments. Each year, many cats fall from windows and balconies. The trauma sustained from a fall of over two stories (24 to 30 feet) is known as “high-rise syndrome.”

As you would guess, high-rise syndrome is more common in urban settings. Studies done on cats that have fallen from 2 to 32 stories show that the overall survival rate is a surprising 90 percent. Strangely, cats that fall from a height under 6 stories have more severe trauma than those that fall from over 6 stories. One theory is that cats reach terminal velocity at about 5 stories, and at this point they relax, allowing a more distributed force of impact and less severe injuries. When cats land before reaching top speed, they are rigid and flexed and prepared for the landing. This results in most of the force impacting the parts of the body that hit initially.

Falling from a greater height leads to less severe injuries???



NASHVILLE SOFTWARE SCHOOL

Bias

Self-selection bias occurs whenever participants can choose whether or not to participate in a study.

Rotten Tomatoes Score (Based on Critics)

NASHVILLE SOFTWARE SCHOOL

Bias

Self-selection bias occurs whenever participants can choose whether or not to participate in a study.

Amazon Review Scores

NASHVILLE SOFTWARE SCHOOL