Introduction to probability

Intro to Data Science

Shawn Santo

Today's agenda

- Introduce probability vocabulary and concepts
- Estimate probabilities through simulation

What's the use?

What we've done so far...

- Use visualization techniques to *visualize* data
- Use descriptive statistics to *describe* and *summarize* data
- Use data wrangling tools to *manipulate* data
- ...all using the reproducible, shareable tools of R and git

That's all great, but what we eventually want to do is to *quantify uncertainty* in order to make **principled conclusions** about the data.

The statistical process

Statistics is a process that converts data into useful information, whereby practitioners

1. form a question of interest,
2. collect and summarize data,
3. and interpret the results.

The population of interest

The **population** is the group we'd like to learn something about. For example:

- What is the prevalence of diabetes among **U.S. adults**, and has it changed over time?
- Does the average amount of caffeine vary by vendor in **12 oz. cups of coffee at Duke coffee shops**?
- Is there a relationship between tumor type and five-year mortality among **breast cancer patients**?

The **research question of interest** is what we want to answer - often relating one or more numerical quantities or summary statistics.

If we had data from every unit in the population, we could just calculate what we wanted and be done!

Sampling from the population

Unfortunately, we (usually) have to settle with a **sample** from the population.

Ideally, the sample is **representative**, allowing us to make conclusions that are **generalizable** to the broader population of interest.

In order to make a formal statistical statement about the broader population of interest when all we have is a sample, we need to use the tools of probability and statistical inference.

Interpreting probabilities

Interpretations of probability

"There is a 1 in 3 chance of selecting a white ball"

Interpretations of probability

"There is a 75% chance of rain tomorrow"

Interpretations of probability

"The surgery has a 50% probability of success"

Interpretations of probability

Long-run frequencies vs. **degree of belief**

Formalizing probabilities

What do we need?

To talk about probabilities, we need three components. These components, when taken together, allow us to think of probabilities as objects that model random experiments:

1. The **sample space** - the set of all possible **outcomes**
2. Subsets of the sample space, called **events**, which comprise any number of possible outcomes (including none of them!)
3. Some way to assign **probabilities** to events

An event is said to **occur** if the outcome of the random experiment is contained in that event.

Sample spaces

Sample spaces depend on the random experiment in question

- Tossing a single fair coin
- Sum of rolling two fair six-sided dice
- The proportion of successful surgeries performed in a given week

What are the sample spaces for the random experiments above?

Events

Events are subsets of the sample space that comprise all possible outcomes from that event.

- Tossing a single fair coin
- Sum of rolling two fair six-sided dice
- The proportion of successful surgeries performed in a given week

What are some examples of events for the random experiments above?

Probabilities

Consider the following possible events and their corresponding probabilities:

- Getting a head from a single fair coin toss: **0.5**
- Getting a prime number sum from rolling two fair six-sided dice: **5/12**
- Having more than 80% of surgeries performed in a given week be successful: **...way more difficult to quantify**

Don't worry about how we calculated these probabilities for now. Just know that probabilities are numbers describing the likelihood of each event's occurrence, which map events to a number between 0 and 1, inclusive.

Working with probabilities

Set operations

Remember that events are (sub)sets of the outcome space. For two sets (in this case events) A and B, the most common relationships are:

- **Intersection** $(A \cap B)$: A **and** B both occur
- **Union** $(A \cup B)$: A **or** B occurs (including when both occur)
- **Complement** (A^c): A does **not** occur

In probability, the union is satisfied when at least one of the events in the union is satisfied.

Two sets A and B are said to be **disjoint** or **mutually exclusive** when $A \cap B = \emptyset$. This means they have no outcomes in common.

Can you think of an experiment with two well-defined events that are disjoint?

How do probabilities work?

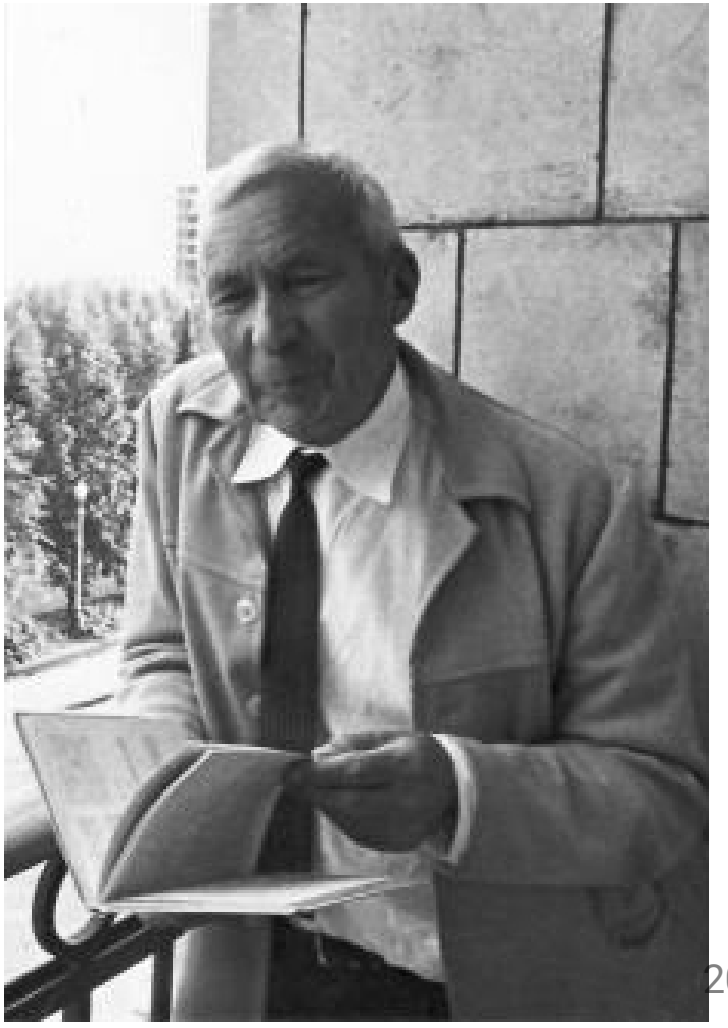
Kolmogorov axioms:

1. The probability of any event in the sample space is a non-negative real number.

2. The probability of the entire sample space is 1.

3. If A and B are disjoint events, then the probability of $A \cup B$ occurring is the sum of the individual probabilities that they occur.

The Kolmogorov axioms lead to probabilities being between 0 and 1 inclusive, and also give rise to some important rules.



Two important rules

Suppose we have events A and B, with probabilities $P(A)$ and $P(B)$ of occurring. The Kolmogorov axioms give us two important rules:

- **Complement Rule**: $P(A^c) = 1 - P(A)$
- **Inclusion-Exclusion**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Practicing with probabilities

ORIGINAL RESEARCH

Annals of Internal Medicine

Coffee Drinking and Mortality in 10 European Countries
A Multinational Cohort Study

	Did not die	Died	Sum
Does not drink coffee	5438	1039	6477
Drinks coffee occasionally	29712	4440	34152
Drinks coffee regularly	24934	3601	28535
Sum	60084	9080	69164

Define the events A = died and B = non-coffee drinker. Calculate the following probabilities for a randomly selected person in the cohort. What are these events in plain English?

- $P(A)$
- $P(A \cap B)$
- $P(A \cup B)$
- $P(A \cup B^c)$

Computing probabilities

Intuitively, we can think of the probability of an outcome (set of outcomes) as the proportion of times the outcome (set of outcomes) would occur if we observed the random process infinitely many times.

If all the outcomes in our random process (sample space - $\mathcal{S}$) are equally likely, then for some event E

$$P(E) = \frac{\#\text{ of outcomes in } E}{\#\text{ of outcomes in } \mathcal{S}}$$

This intuition and the above expression will serve as motivation for how we simulate probabilities with a computer algorithm.

Getting some more practice

- Create your personal private repository by clicking https://classroom.github.com/a/Shv2Eok5
- Follow the steps as we have done previously to clone this and create a new RStudio project in the RStudio Docker containers.