Probability vs. statistics

- **Probability** is a set of tools to **describe** the uncertain world around us
- We model those uncertain things by RVs
- Assume a distribution for the RV. Then we can do some **computations**
  1. conditional mean
  2. covariance
  3. conditional probabilities
  4. ...
- These are meaningful IF our model is true. Theoretical.

Probability vs. statistics

- **Statistics**: use data to learn about the world
- Type of questions:
  1. Is a die fair or unfair? **hypothesis testing**
  2. If unfair, what is the probability of rolling a 3? **estimation**

Population / sample

The **two** main things I want you to take away from this course are:

1. Understanding the difference between a **population and a sample**
2. Theoretical and practical understanding of linear regression

Population / sample

- the **population** is the entire group of units of interest
- the **sample** is the part of the population that we actually have measurements for

Sampling: urn

Percentage of purple balls?

Example: light bulbs

- On a given day, we are doing quality control at a factory that produces light bulbs
- We are interested in how long these bulbs last on average
- The population consists of **all the bulbs** that the factory produces **that day**
- To get an idea about the duration of the bulbs, we randomly take a few and see how long it takes before they run out
- Obviously, we are not going to try all the bulbs!
- We will try to say something about the population of **all** bulbs produced that day, using the observations in our **sample**
- That is called **statistical inference**

Bulbs: population versus sample

- Let X be the duration for a light bulb produced on the day we are testing
- If we would test **all** the bulbs, we would **know** the mean $E[X] = \mu_X$ for our **population**
- We only have a **sample** of all light bulbs.
- The number μ_X exists, but we do not know it!
- Instead, we look at the **sample average**, $\bar{X}$
- **The sample average is a random variable!**

The sample average is a random variable!

Sampling: non-random

Average number of calories in a candy?

Expectation, sample average

- The mean is a **population** quantity. It is a **fixed** number.
- The average is a **sample** quantity. It is a **random variable**.

- In statistics, we **estimate**: use the sample average to make a guess (**estimate**) about the mean

Example: voting and polls

- Prior to **election** day, we are interested in predicting what percentage of votes each candidate gets
- The **population** is: all voters
- **54%** of voters is going to choose candidate A
- But we **do not know** this number: it is the **population** parameter
- It is not cost-effective (nor feasible) to ask **all** the people
- We can obtain a **sample** of 100 individuals, and ask them
- In that sample, we find that **59%** of individuals would vote for candidate A

Example: voting (2)

- Note 1: $59\% \neq 54\%$
- Note 2: Had I asked a different group of people, I would have had a **different** number
- Note 3: 59% is our **best guess.** It may be better to report a **confidence interval** rather than our best guess. "With 95% confidence, the percentage of people that would vote for candidate A is in **between** 53% and 65%."
- Note 4: The reason that this is important, is that the uncertainty could have been much **higher.** The research might have only been able to conclude that the 95%-CI is **(10%,90%)**

Random sampling

- Assumption: data is gathered using **simple random sampling.**
- i.i.d.: **i**ndependently and **i**dentically **d**istributed
- For more, read p. 43+44.