DATA SAMPLING

MOTIVATING EXAMPLE

- You work at a manufacturing company producing climbing ropes.
- All the ropes produced must satisfy certain strength conditions to be certified.
- The only way to test the strength is by actually forcing the rope until it breaks and then get a measurement.

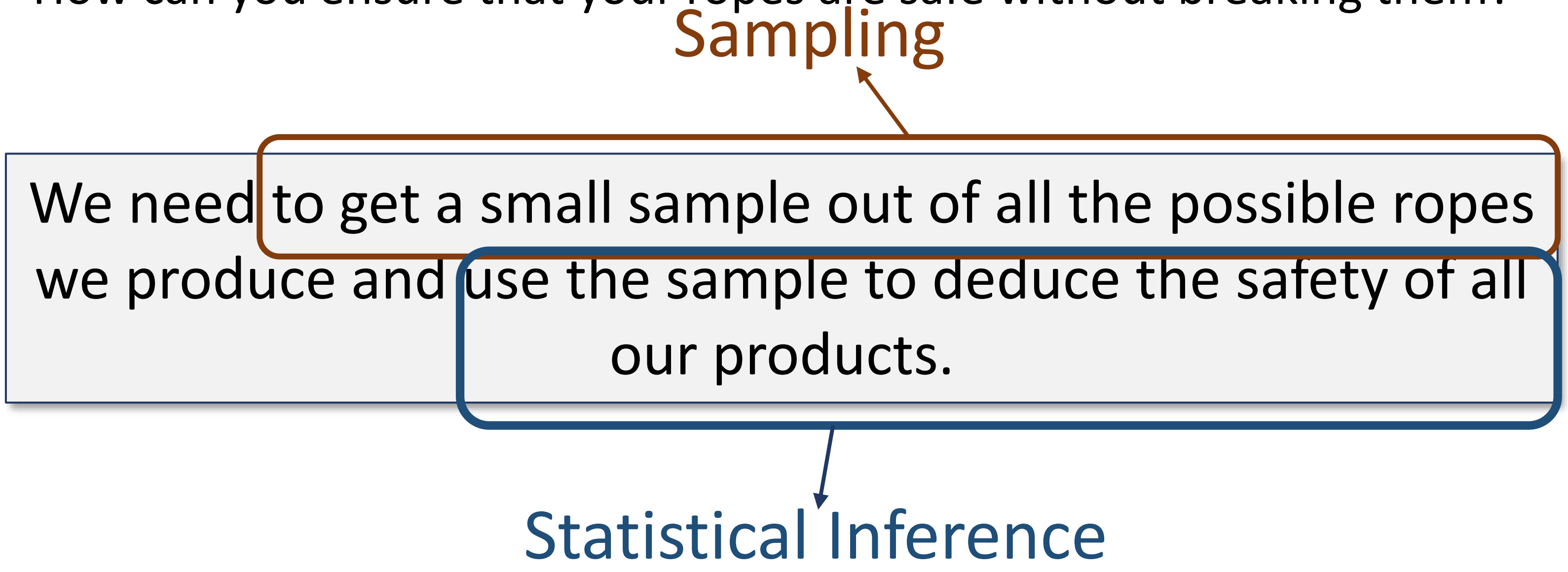
- How can you ensure that your ropes are safe without breaking them?

We need to get a small sample out of all the possible ropes we produce and use the sample to deduce the safety of all our products.

MOTIVATING EXAMPLE

- You work at a manufacturing company producing climbing ropes.
- All the ropes produces must satisfy certain strength conditions to be certified.
- The only way to test the strength is by actually forcing the rope until it breaks and then get a measurement.

- How can you ensure that your ropes are safe without breaking them?



Sampling

We need to get a small sample out of all the possible ropes we produce and use the sample to deduce the safety of all our products.

Statistical Inference

SAMPLING IS THE BASIS OF STATISTICAL INFERENCE

- Sampling consists in extracting or measuring a random subset of data coming from a larger population.

- Using this subsample of data, we would like to estimates features or statistics about the whole data.

- It is a form of prediction, (but instead of predicting the relationship between two variables, we are predicting some intrinsic properties of one single variable, for instance its average value).

- All our conclusions need to be given with a certain level of confidence. The larger the sample, the higher our confidence in the inference.

DATABASES VS PHYSICAL DATA COLLECTION

If we already have a lot of data, the process of sampling may consist simply in extracting a subset of our database (example: recommendation system in Amazon).

However, for some studies we need to physically collect the data. For instance:

- Asking people on the street to test a new coffee brand;
- Physically testing a manufactured piece.

The way we collect this data is going to affect the quality and generality of my model.

DIFFERENT TYPES OF SAMPLING

Make a subsample suitable for inference is much harder than it looks. Here are some sampling methodologies, and their benefits and potential biases:

Method	Definiton	Benefit	Potential bias	Example
Random Sampling	Random selection of sample	Easy to do. Cost efficient. Objective	Can ignore cases with low representation	Make database query based on a random selection of IDs.
Systematic sample	Data is selected at regular intervals	Simple. Representative	If there are cycles in the data, the sample can't be representative.	Select sales data for every Monday for the last 4 years. (maybe Monday is not a typical day?)
Quota sampling	Data deterministic, but make sure the subsample has same quota for a number of characteristics	Make sure all subgroups are represented	Subjectivity when identifying what groups should be considered.	When studying consumer behaviour, identify profession of buyer, and make sure that the sample has the same proportions as the population.