YData: Introduction to Data Science

Lecture 15: Sampling

Overview

Elementary probability

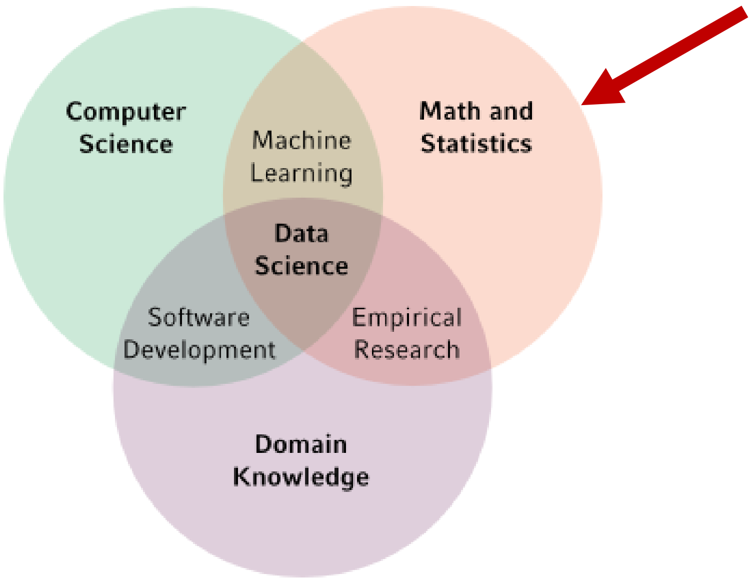
Sampling

Distributions

Large random samples

Parameters and statistics

Sampling distributions



Announcements

To give you more time to work on project 1, homework 5 is now a "practice homework"

- i.e., you will not turn it in, and it will not be graded

Keep working on project 1!

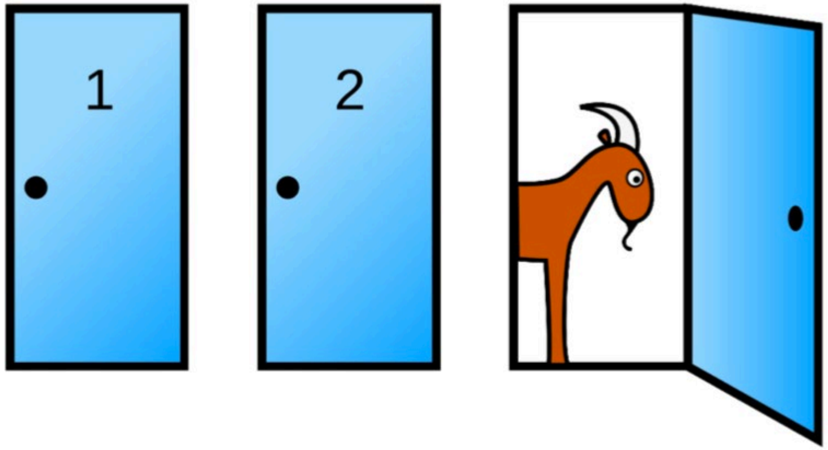
- Due at 11pm on Friday

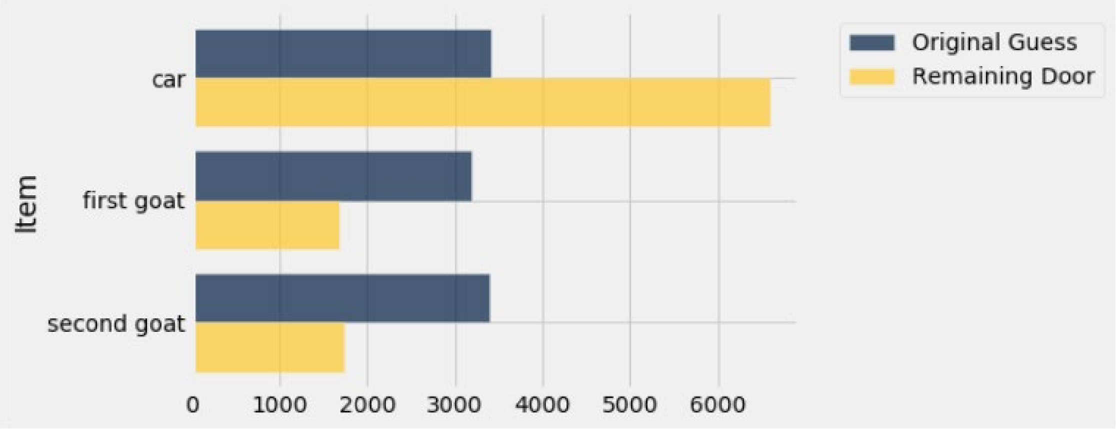
Probability

Probability

Last class we looked at the "Monty Hall Problem" which involved thinking through the probability of different events

Let's continue our exploration with a bit more formal definitions and explorations





The basics of probability

A probability model assigns values to random events

Lowest value: 0

- Chance of event that is impossible

Highest value: 1 (100% chance)

- Chance of event that is certain

The probability that an event **doesn't occur** is 1 minus the probability the event does occur:

- E.g., if there is a 0.7 change an event occurs, then the probability it doesn't occur is:
- 1 - 0.7 = 0.3

Equally likely outcomes

Assuming all outcomes are equally likely, the chance of an event A is:

$$P(A) = \frac{\text{number of outcomes that make A happen}}{\text{total number of outcomes}}$$

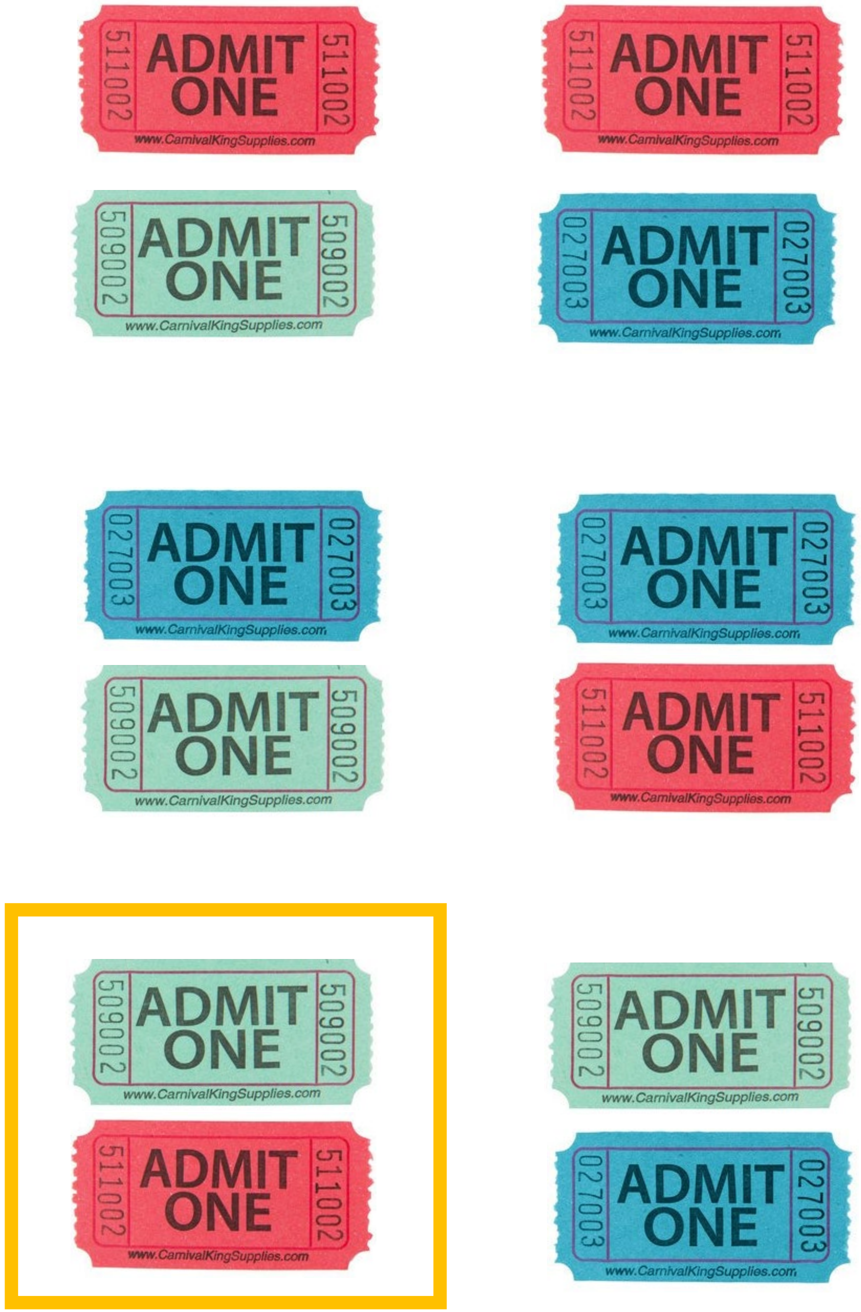
Example

Suppose there are three tickets: Red, Green, and Blue

When sampling **without replacement**, what's the chance of getting GR? i.e.,

- a green ticket on the **first** draw
- and then a red ticket on the **second** draw?

RG RB BG BR GR GB = P(GR) = 1/6



Multiplication rule

The chance that two events A and B both happen is:

= P(A happens) x P(B happens given that A has happened)

When sampling **without replacement**, what's the chance of getting GR?

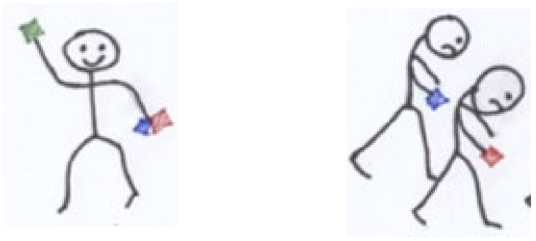
- RB RG BR BG GR GB = P(GR) = 1/6

P(G) = 1/3

P(R given G) = 1/2

P(GR) = 1/3 x 1/2 = 1/6

Stage 1: 1/3



Stage 2: 1/2

Addition rule

If event A can happen in exactly one of two (mutually exclusive) ways, then:

$$P(A) = P(\text{first way}) + P(\text{second way})$$

What is the chance of getting a red or a green on a single draw?

$$P(R \text{ or } G) = R \quad G \quad B = 2/3$$

$$P(R \text{ or } G) = P(R) + P(G) = 1/3 + 1/3 = 2/3$$

Example

To calculate the probability an event occurs, sometimes it is easier to calculate 1 minus the probability the event did not occur

Example: what is the probability of getting **at least** one head out of *k* coin flips?

In 3 tosses:

- Any outcome except TTT
- P(TTT) = (1/2) x (1/2) x (1/2) = 1/8
- P(at least one head) = 1 - P(TTT) = 7/8 = 87.5%

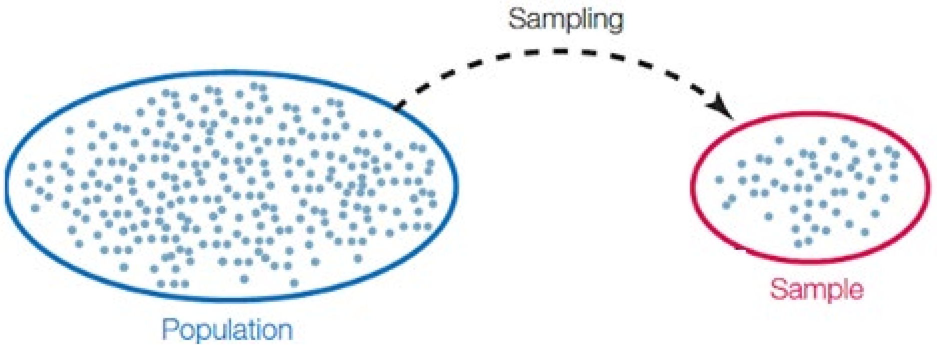
In 10 tosses:

- $1 - (1/2)^{10}$
- 99.9%

Let's explore this in Jupyter!

Sampling

Sampling



Sampling is the process of selecting a subset of items from a larger population

Deterministic sample:

- Specify which elements of a set you want to choose, without any chances involved

Probability sample:

- The probability of drawing each subset of the population can be calculated
- Not all members of the population have to have equal chance of being selected

Sample of Convenience

Convenience sampling is a type of non-probability sampling that involves the sample being drawn from that part of the population that is close to hand

Example: sample consists of whomever walks by

Just because you think you're "sampling at random", doesn't mean you are

If you can't figure out ahead of time

- what's the population
- what's the chance of selection, for each group in the population

then you don't have a random sample

Simple random sample

Simple random sample: each member in the population is equally likely to be in the sample

Allows for generalizations to the population!

Let's explore this in Jupyter!

Distributions

Probability Distributions

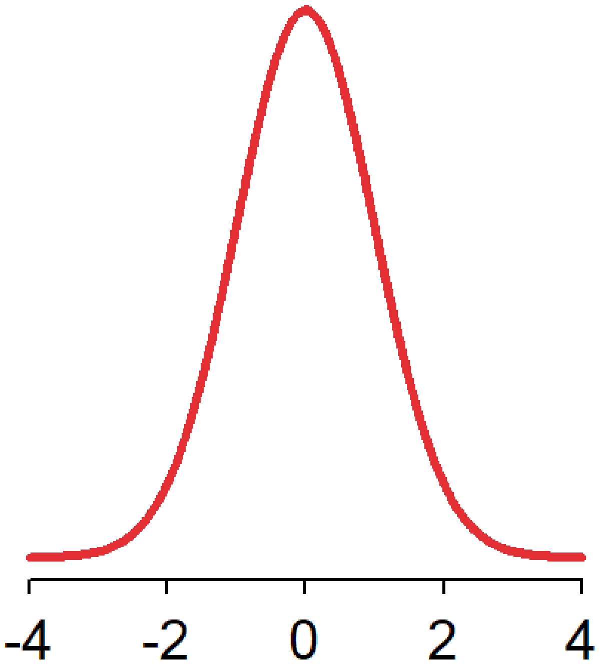
A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes

A probability distribution consist of:

- All possible values a random quantity can take
- The probability of each of those values

If you can do the math, you can work out the probability distribution without ever simulating the random quantity

- However, running simulations can often be easier



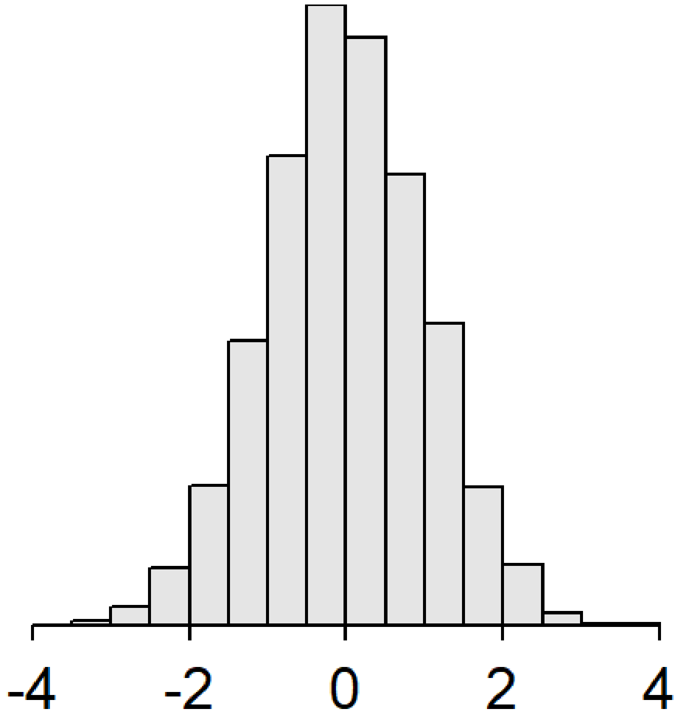
Empirical Distribution

In an **empirical distribution**, the probability of each outcome is based on observations

- Observations can be from repetitions of an experiment

An empirical distribution consists of

- All observed values
- The proportion of times each value appears

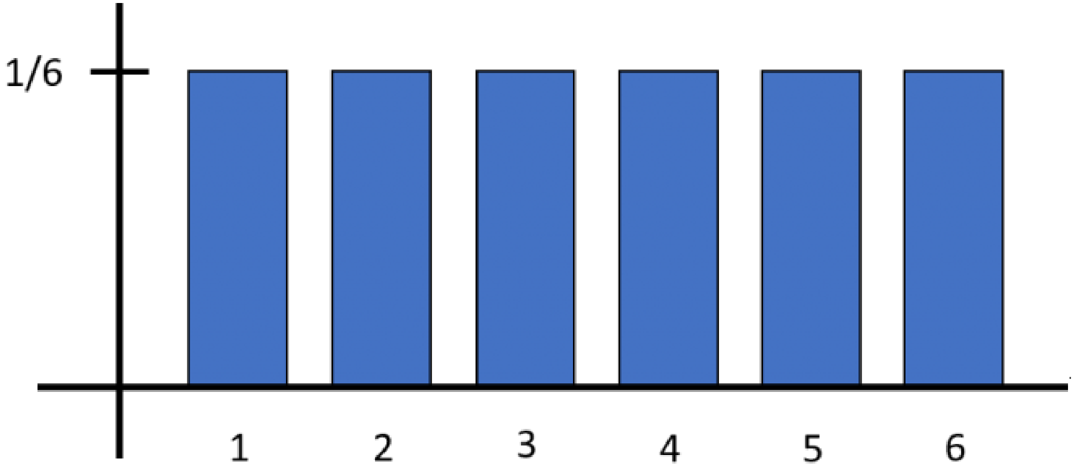


Let's explore this in Jupyter!

Large random samples

Law of large numbers

Law of large numbers: If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event



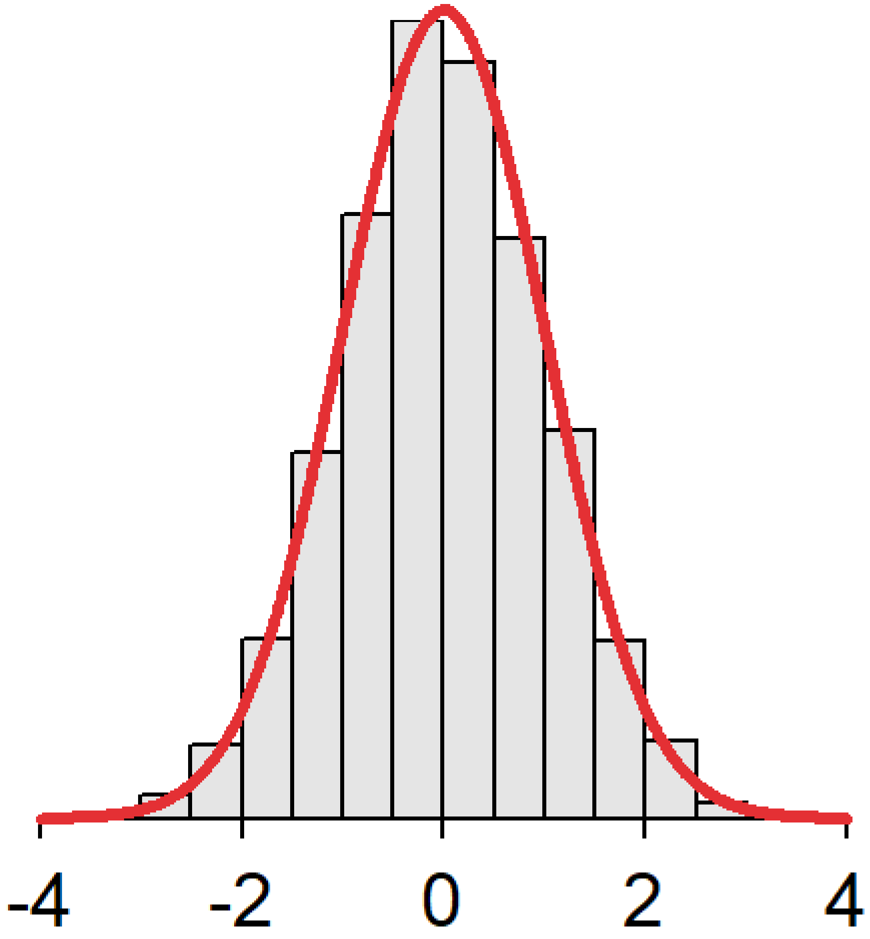
As you increase the number of rolls of a die, the proportion of times you see the face with five spots gets closer to 1/6

$\hat{p}_5 \rightarrow 1/6$

As the number of rolls get large

Empirical distribution of a sample

If the sample size is large, then the empirical distribution of a simple random sample resembles the distribution of the population, with high probability

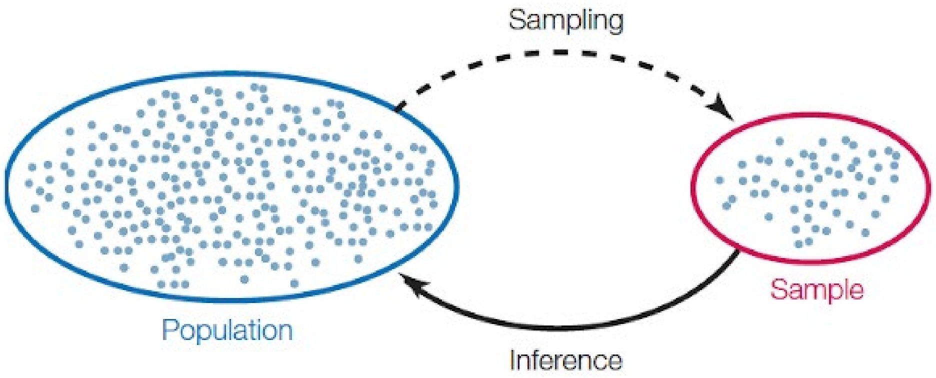


Let's explore this in Jupyter!

statistics

Inference

Statistical Inference: Making conclusions about a population based on data in a random sample



Frequently this involves using data in a sample to estimate the value of a fixed unknown number

Example:

- Estimating the average height of all humans on Earth from a random sample of 1,000 humans
  - Our estimate will vary from sample to sample

Terminology

A parameter is number associated with the population

- e.g., population mean μ
- e.g., average height of all humans

A **statistic** is number calculated from the sample

- e.g., sample mean $\bar{x}$
- e.g., average height of 1,000 people in our sample

A statistic can be used as an estimate of a parameter

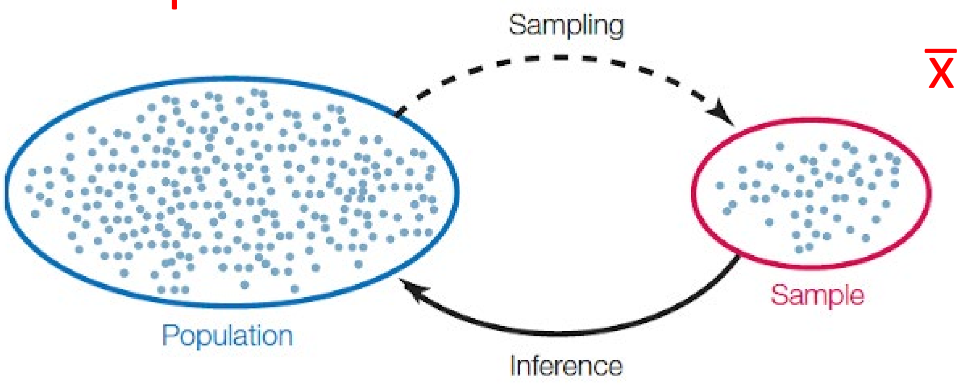
Let's explore this in Jupyter!

parameter

μ

statistic

$\bar{x}$



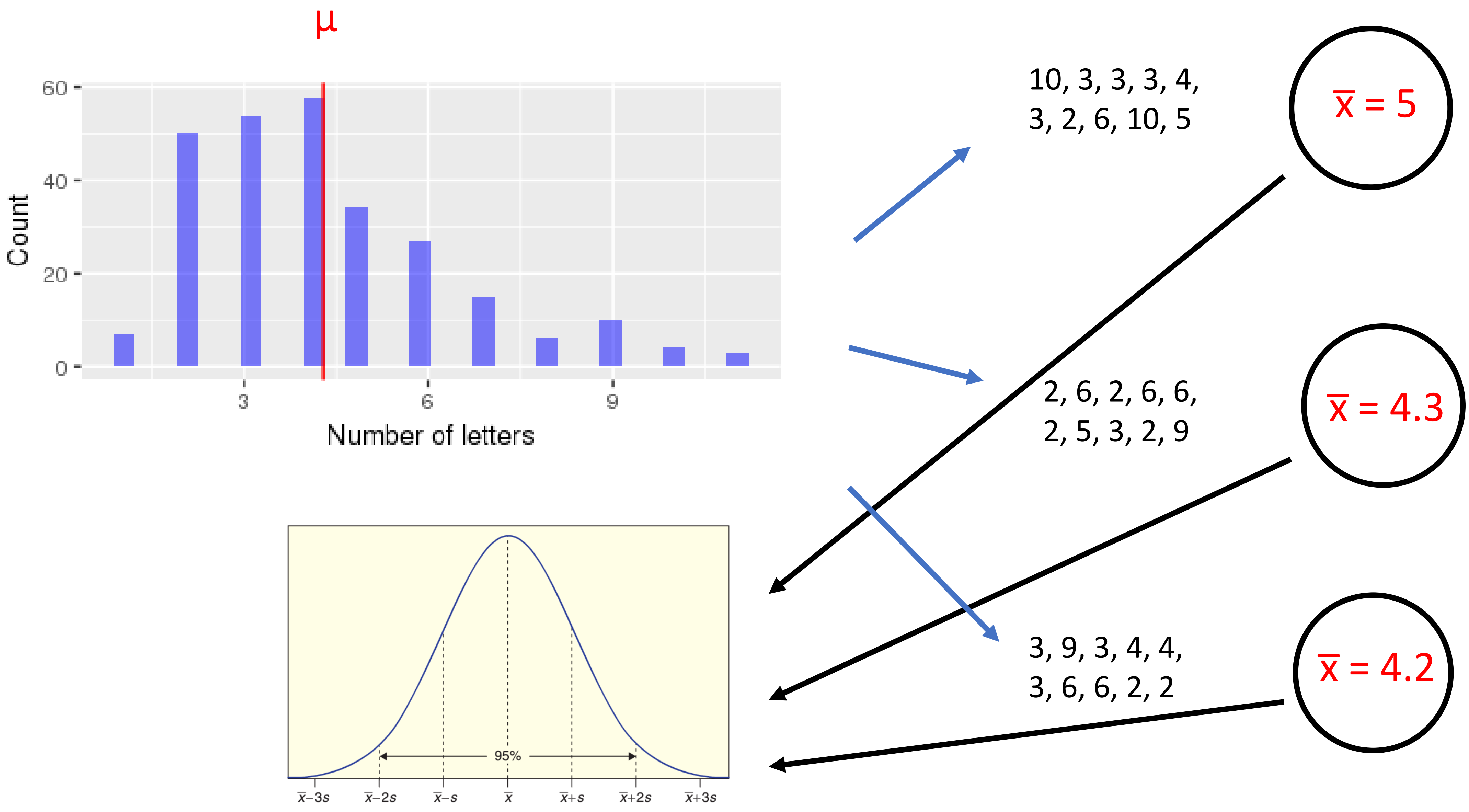
Probability distribution of a statistic

Values of a statistic vary because random samples vary

A **sampling distribution** is a probability distribution of *statistics*

- All possible values of the statistic and all the corresponding probabilities
- We can approximate a sampling distribution by a simulated statistics

Sampling distribution illustration



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

Sampling distribution!

Empirical distribution of a statistic

Empirical distribution of the statistic:

- Based on simulated values of the statistic
- Consists of all the observed values of the statistic, and the proportion of times each value appeared

Good approximation to the sampling distribution of the statistic if the number of repetitions in the simulation is large