YData: Introduction to Data Science

Lecture 16: Assessing models

Overview

Quick review of

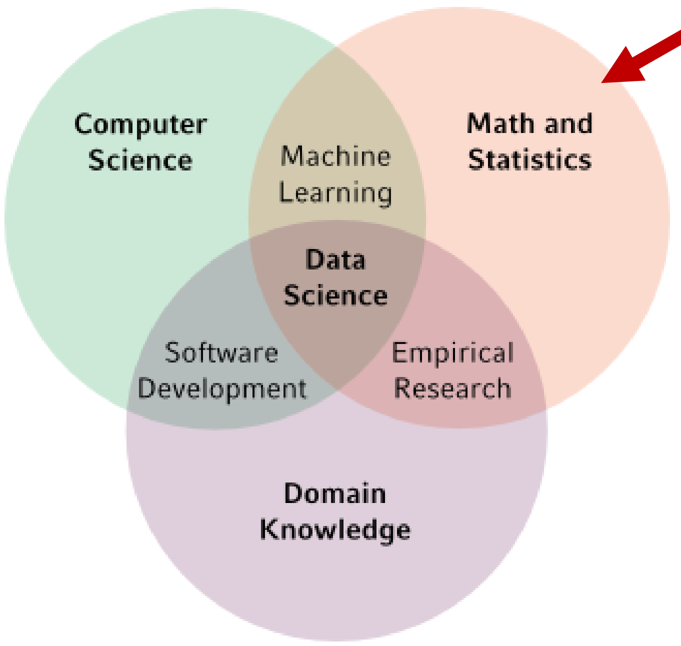
- Elementary probability
- Sampling
- Distributions
- Large random samples

Parameters and statistics

Sampling distributions

Testing hypotheses and assessing models

- Example: bias in jury selection
- Example: genetics of peas

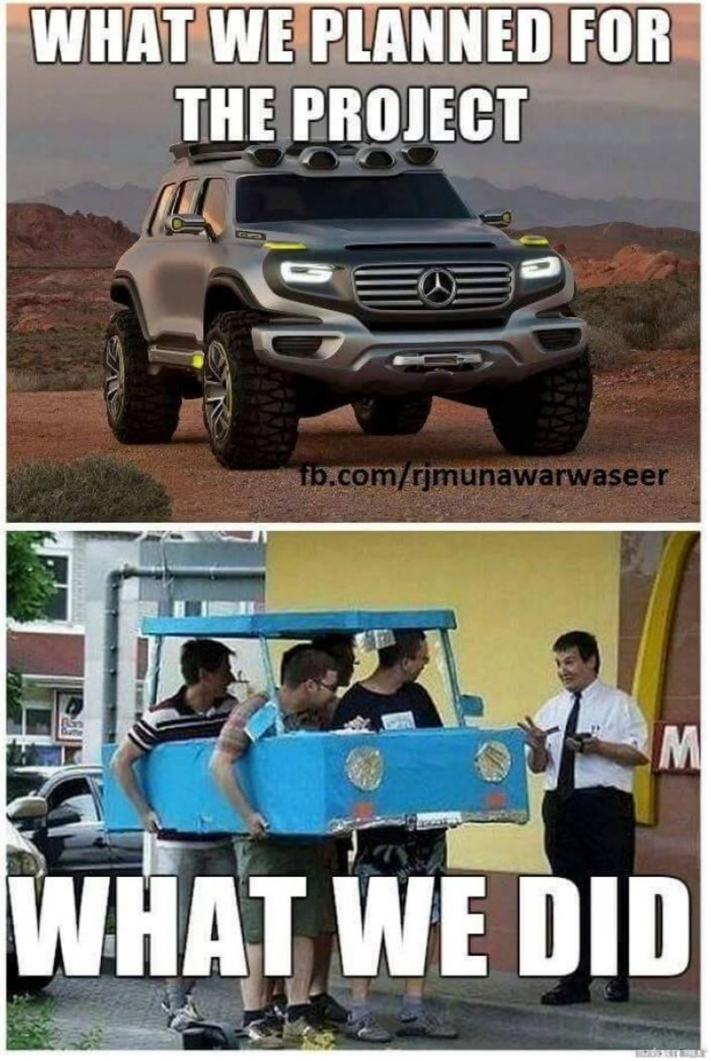


We are here

Announcements

Project 1 due at 11pm on Friday

- For people working together in pairs, only one person should submit the project.
- Be sure to mark both people's names on Gradescope.
    - Instructions on part of the project are incorrect. Only do one submission with both partners' names
- TAs are holding additional office hours this week
    - Check calendar on Canvas



Quick review of probability and sampling

Quick review: probability

Probability models assigns values between 0 and 1 to random events

For equally likely events, the probability of an event A, denoted P(A), is the number of ways A can happen divided by the total number of outcomes

- Example: Event A = rolling a 5 on a six-sided die
- 6 possible outcomes {1, 2, 3, 4, 5, 6} one of which is 5
- P(A = 5) = 1/6

Quick review: probability

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)

- Example: Probability of rolling two 5's in a row of a 6-sided die:

  $P(A = 5) \times P(B = 5) = 1/6 \times 1/6 = 1/36$

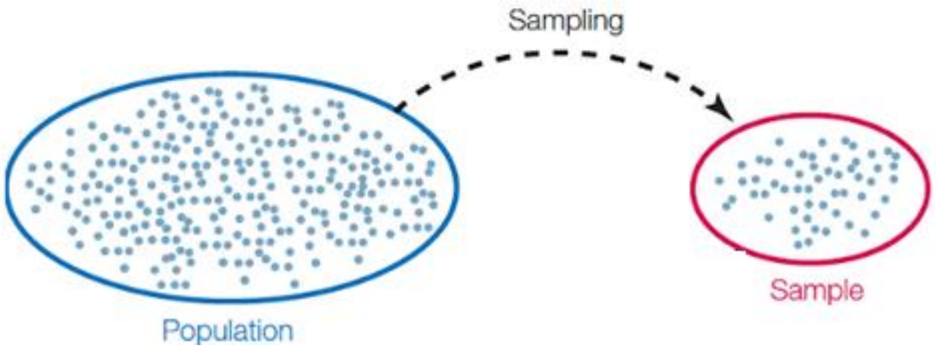
Addition rule: If event A can happen in exactly one of two ways, then:

P(A) = P(first way) + P(second way)

- Example: Probability of rolling a 5 **or** a 6 on a single roll:

  $P(A = 5) + P(A = 6) = 1/6 + 1/6 = 2/6$

Quick review: sampling



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

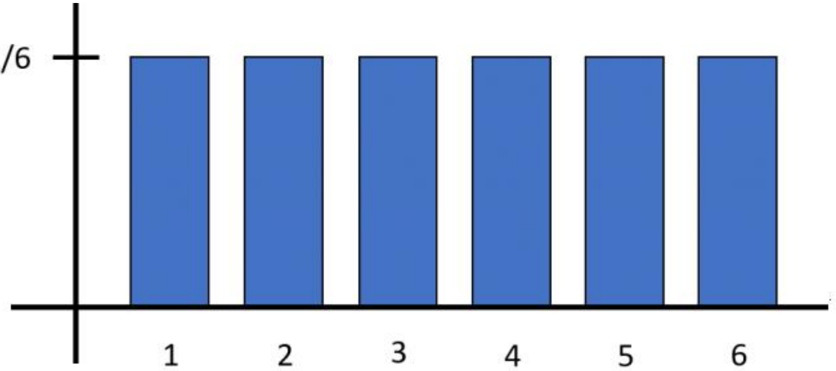
There are different ways to sample. Some are better than others:

- **Deterministic sample**: specify which elements of a set you want to choose
- **Convenience sample:** sample drawn from that part of the population that is close to hand
- **Simple random sample**: each member in the population is equally likely to be in the sample

Quick review: Probability Distributions

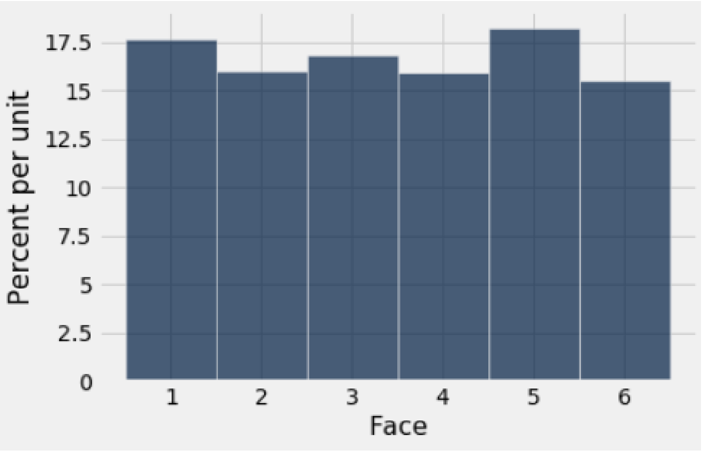
A **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible outcomes and consists of:

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



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

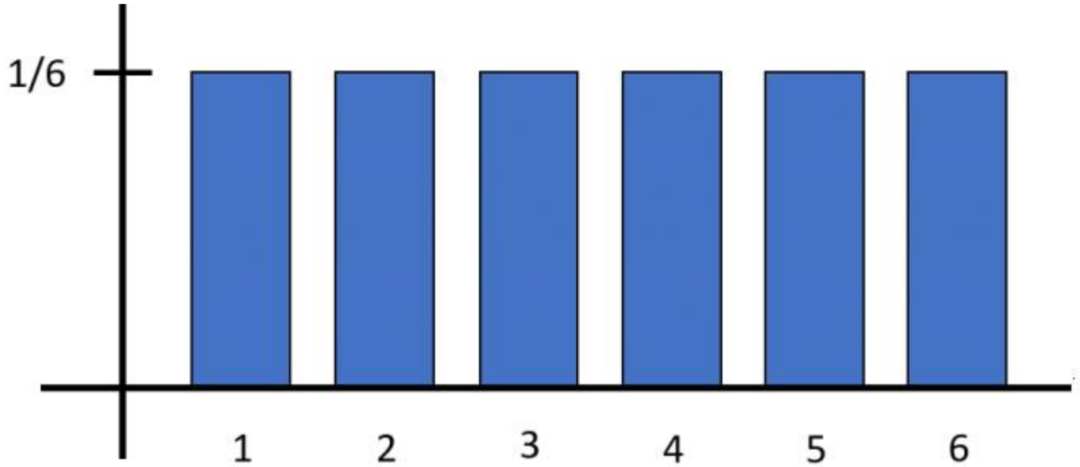
- All observed values
- The proportion of times each value appears



Law of large numbers

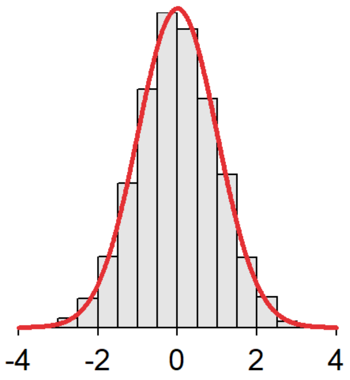
Law of large numbers: If a chance experiment is repeated many times, independently and under the same conditions, then the proportion of times that an event occurs gets closer to the theoretical probability of the event

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



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

As the number of rolls get large



Let's explore this in Jupyter!

Parameters, statistics and sampling distributions

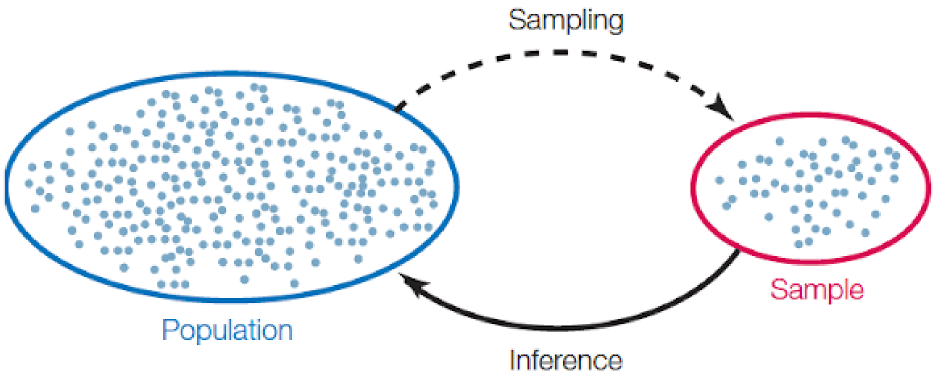
Inference

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

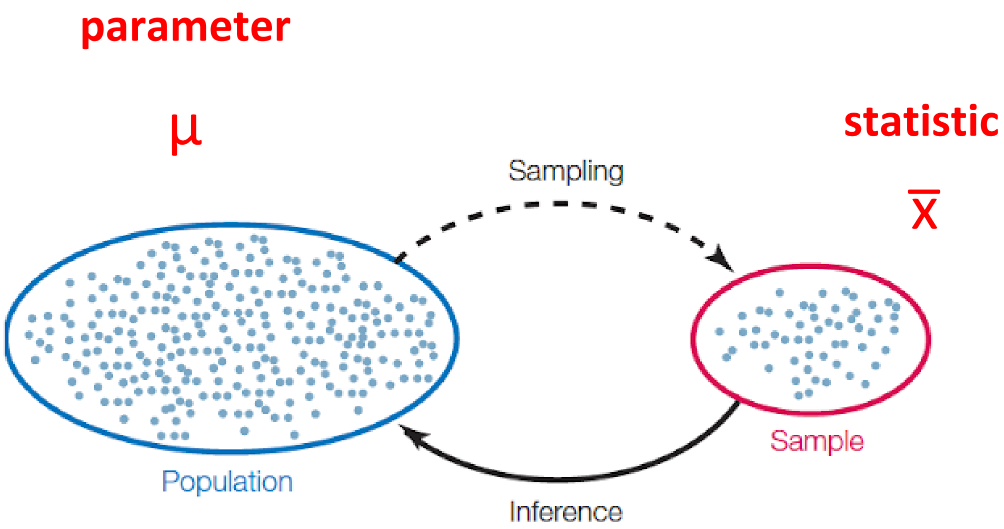
A parameter is number associated with the population

- 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



parameter

μ

statistic

$\bar{x}$

Sampling distributions

Values of a statistic vary because random samples vary

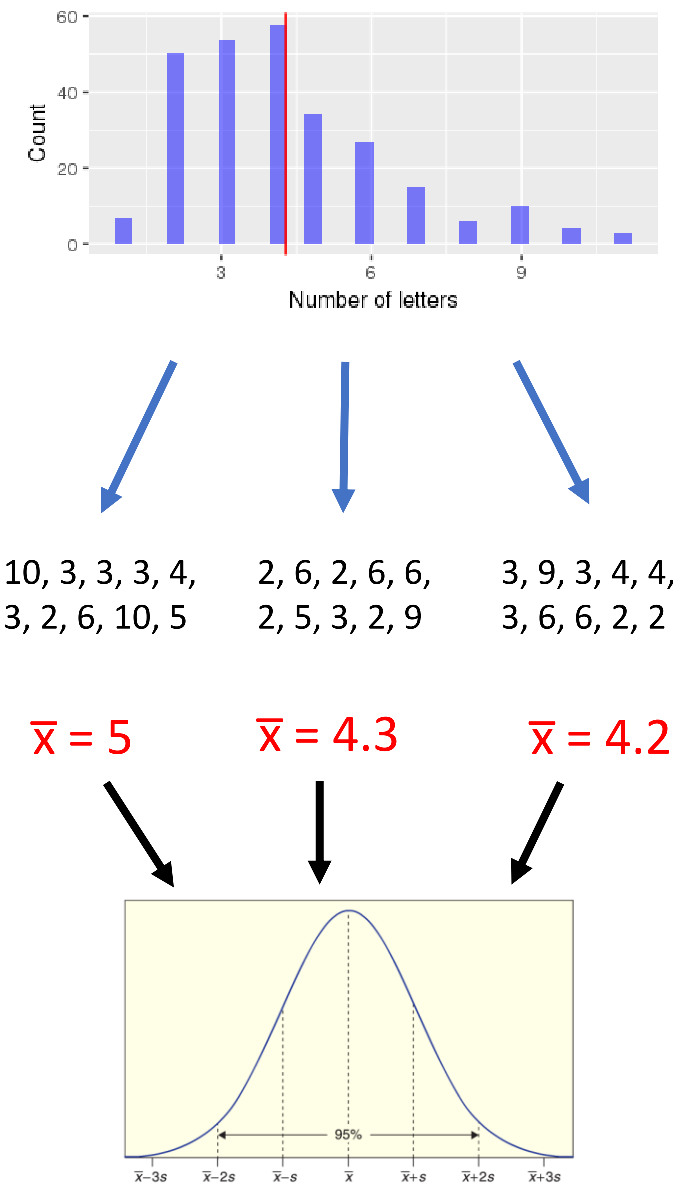
A **sampling distribution** is a probability distribution of *statistics*

- All possible values of the statistic and all the corresponding probabilities
  - Usually solved mathematically to get all possible values

We can approximate a sampling distribution by an **empirical distribution** of the statistic that is based on simulated values

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

Let's explore this in Jupyter!



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

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

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

$\bar{x} = 5$ $\bar{x} = 4.3$ $\bar{x} = 4.2$

Sampling distribution!

Testing hypotheses

Choosing one of two viewpoints

Example 1:

- "Chocolate has no effect on cardiac disease."
- "Yes, it does."

Example 2:

- "This jury panel was selected at random from eligible jurors."
- "No, it has too many people with college degrees."

How can we use data to decide which claim is correct?

Assessing Models

Models

A model is a description of some underlying phenomenon

- Based a set of assumptions about how a particular phenomenon works

In data science, many models describe processes that involve randomness

- Called "Chance" or "Stochastic" models

Approach to assessment

If we can simulate data according to the assumptions of the model, we can learn what the model predicts

We can then compare the predictions to the data that were observed

If the data and the model's predictions are not consistent, that is evidence against the model

Example: assessing bias in jury selection

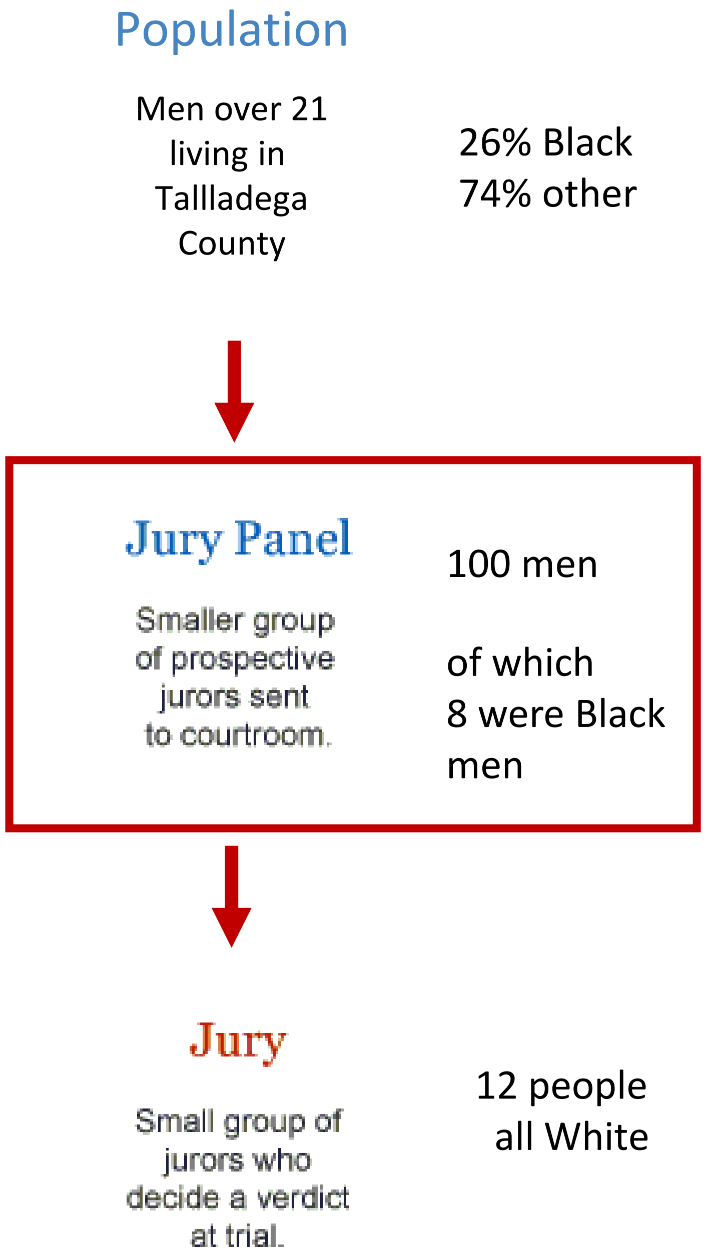
Swain vs. Alabama, 1965

Robert Swain was a Black man who was convicted of rape in the Circuit Court of Talladega County, Alabama, and sentenced to death by an all White jury

The case was appealed to the Supreme Court, in part, on the ground that there were no Black jurors on the jury that convicted him

Only men 21 years or older were allowed to serve

- 26% of this population were Black
- Swain's jury panel consisted of 100 men
    - These men were supposed to be randomly selected from the population
- 8 of the 100 men on the panel were Black



Population

Men over 21 living in Tallladega County

26% Black
74% other

Jury Panel

Smaller group of prospective jurors sent to courtroom.

100 men

of which
8 were Black men

Jury

Small group of jurors who decide a verdict at trial.

12 people
all White

Supreme court ruling

The Supreme Court denied Swain's appeal, claiming that the overall percentage disparity was small

- i.e., they claimed that while 8/100 is less than 26%, this difference is not large enough to show Black men were systematically excluded

Let's assess this claim…

Sampling from a distribution

We can randomly sample from a categorical distribution using:

sample_proportions(sample_size, pop_distribution)

- sample_size: the size of a sample
- pop_distribution: population proportions in different categories

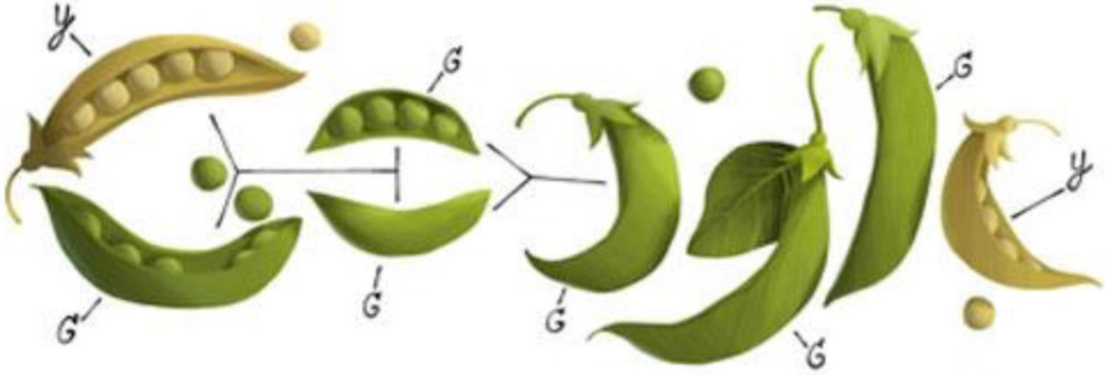
Samples *at random* from the population

- Returns an array containing the distribution of the categories in the sample

Let's explore this in Jupyter!

A Genetic Model

Gregor Mendel, 1822-1884



A model of pea flowers

Mendel examined the flowers of the common pea plant *Pisum sativum*

- Each plant has either purple flowers or white flowers

Mendel's model:

- Each plant is purple-flowering with chance 75%

Mendel grew 929 plants and 705 out of them had purple flowers

Choosing a statistic

Start with the percent of purple-flowering plants in a sample

If that percent is much larger or much smaller than 75, that is evidence against the model

statistic: | sample percent of purple-flowering plants - 75 |

If the statistic is large, that is evidence against the model

Let's explore this in Jupyter!