Statistics, models, & hypothesis testing

Introduction to data analysis: Lecture 6

Ori Plonsky

Spring 2023

Source: xkcd

Inference

- **Statistical Inference:** Making conclusions based on data in random samples
- **Example:**

fixed

Use the data to guess the value of an unknown number

depends on the random sample

Create an **estimate** of the unknown quantity

Terminology

Parameter

A number associated with the population

- Unknowable (unless we have data for the whole population)

Statistic

A number which is a function of (only) the sample

- Can be exactly calculated from the sample

A statistic can be used as an **estimate** of a parameter

(notebook)

Probability distribution of a Statistic

- Values of a statistic vary because random samples vary
  - In each possible sample, the statistic has a different value

Probability distribution of a Statistic

- Values of a statistic vary because random samples vary
  - In each possible sample, the statistic has a different value
- "Sampling distribution" or "probability distribution" of the statistic:
  - All possible values of the statistic,
  - and all the corresponding probabilities

Probability distribution of a Statistic

- Values of a statistic vary because random samples vary
  - In each possible sample, the statistic has a different value
- "Sampling distribution" or "probability distribution" of the statistic:
  - All possible values of the statistic,
  - and all the corresponding probabilities
- Can be hard to calculate
  - Either have to do the math
  - Or have to generate all possible samples and calculate the statistic based on each sample

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

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 probability distribution of the statistic
  - if the number of repetitions in the simulation is large
  - According to?

Simulating a Statistic

- Figure out the code to generate *one* value of the statistic
  - Based on a fixed sample size
- Create an empty array in which you will collect all the simulated values
- For each repetition of the process:
  - Simulate one value of the statistic
  - Append this value to the collection array
- At the end of all the repetitions, the collection array will contain all the simulated values

(notebook)

Models and Hypotheses

Biased coin?

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100
- Simulate the empirical distribution of the statistic

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100
- Simulate the empirical distribution of the statistic
- What do we need first?
  - Simulate one value of the statistic
    - How do we do this?

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100
- Simulate the empirical distribution of the statistic
- What do we need first?
  - Simulate one value of the statistic
    - How do we do this?
- We need a model of how we think the coin works!
  - To simulate the coin, we need to assume we know the (unknowable!) parameter

Biased coin?

- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100
- Simulate the empirical distribution of the statistic
- What do we need first?
  - Simulate one value of the statistic
    - How do we do this?
- We need a model of how we think the coin works!
  - To simulate the coin, we need to assume we know the (unknowable!) parameter

A model is a set of assumptions about the world

Biased coin?

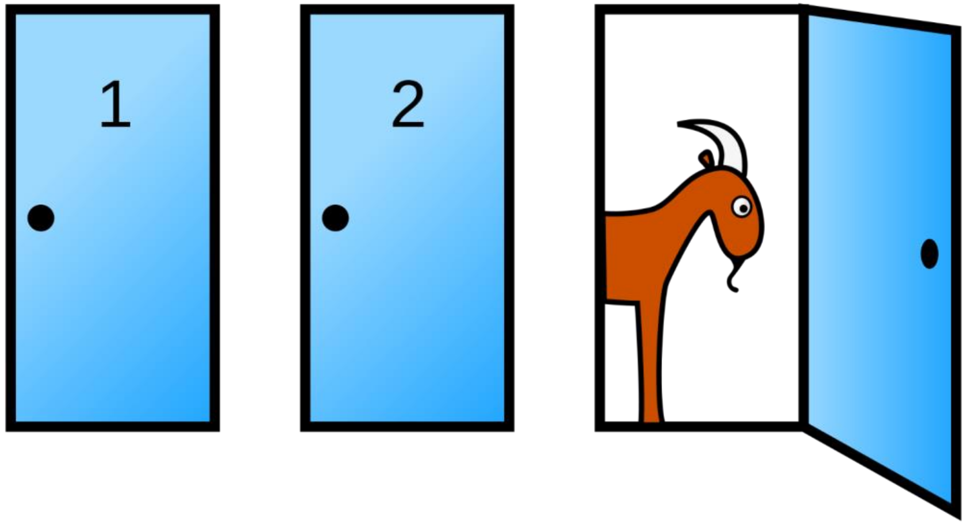
- Parameter: the probability that (a particular) coin yields Heads
- Statistic: (#Heads)/100
  - Sample size = 100
- Simulate the empirical distribution of the statistic
- What do we need first?
  - Simulate one value of the statistic
    - How do we do this?
- We need a model of how we think the coin works!
  - To simulate the coin, we need to assume we know the (unknowable!) parameter

A model is a set of assumptions about the world

- Often, models include assumptions about the chance processes used to **generate data**

Lying Monty?

- Parameter: chance for win given switch
- Statistic: #wins/#shows
  - (Given switch)
- To simulate the statistic, we need a model



Testing hypotheses

Testing hypotheses

- We have a question regarding the (unknowable) value of parameter(s)

Testing hypotheses

- We have a question regarding the (unknowable) value of parameter(s)
- We have access to data that gives us one estimate of this parameter

Testing hypotheses

- We have a question regarding the (unknowable) value of parameter(s)
- We have access to data that gives us one estimate of this parameter
- We know the parameter has some theoretical value
  - E.g., The coin is unbiased
  - E.g., Monty never lies
  - E.g., the two groups are equal

Testing hypotheses

- We have a question regarding the (unknowable) value of parameter(s)
- We have access to data that gives us **one** estimate of this parameter
- We know the parameter has some theoretical value
    - E.g., The coin is unbiased
    - E.g., Monty never lies
    - E.g., the two groups are equal
- We want to ask whether the data - an estimate from **one** random sample - is **consistent or inconsistent** with the theoretical value of the parameter

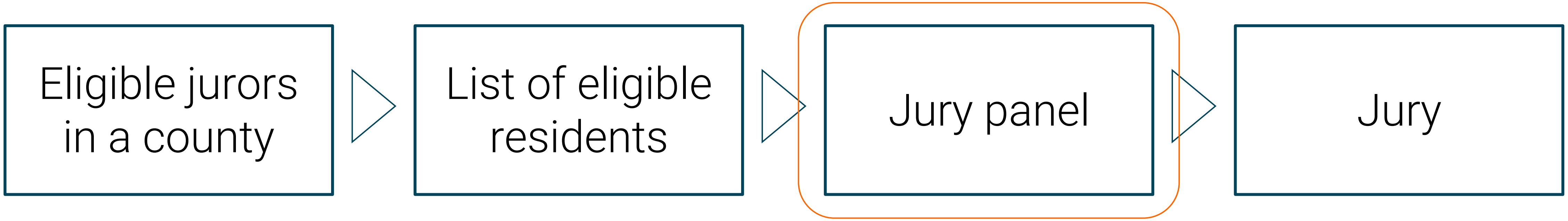
Approach to Assessment

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts as likely values of the statistic.
- Then, we compare the predictions of the model to the observed data
- If the observed data and the model's predictions are not consistent, it is evidence **against the model**.

Approach to Assessment

- If we can simulate data according to the assumptions of the model, we can learn what the model predicts as likely values of the statistic.
- Then, we compare the predictions of the model to the observed data
- If the observed data and the model's predictions are not consistent, it is evidence **against the model**.
  - We say we can **reject the** model (with some probability)
    - E.g., the coin is not unbiased
    - E.g., Monty may sometime lie
    - E.g., the two groups are not equal

Example inference problem: Jury Panels



Section 197 of California's Code of Civil Procedure:
All persons selected for jury service shall be selected **at random**, from a source or sources inclusive of a representative cross section of the population of the area served by the court.

Sixth Amendment to the US Constitution:
… the accused shall enjoy the right to a speedy and public trial, by an impartial jury of the state and district wherein the crime shall have been committed.

Example inference problem

- Talladega County, Alabama, 1965
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Population as required by law: all men 21 years or older
- 26% of this population were black
- Swain's jury panel consisted of 100 men
- 8 men on the panel were black

https://en.wikipedia.org/wiki/Swain_v._Alabama

Example inference problem

- Talladega County, Alabama, 1965
- Robert Swain, black man convicted of crime
- Appeal: one factor was all-white jury
- Population as required by law: all men 21 years or older
- 26% of this population were black
- Swain's jury panel consisted of 100 men
- 8 men on the panel were black
    - Is this likely when the jury panel is selected at random from the population?

https://en.wikipedia.org/wiki/Swain_v._Alabama

Example inference problem

- US Supreme Court:
  "... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes.“

https://en.wikipedia.org/wiki/Swain_v._Alabama

Example inference problem

- US Supreme Court:
  "... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes."
- That is, the court claims:
  - The panel was selected at random
  - Difference between 8% and 26% is due to chance

https://en.wikipedia.org/wiki/Swain_v._Alabama

Example inference problem

- US Supreme Court:
  "... the overall percentage disparity has been small and reflects no studied attempt to include or exclude a specified number of Negroes.“
- That is, the court claims:
  - The panel was selected at random
  - Difference between 8% and 26% is due to chance
- Model: “Selection is made at random”
  - We can simulate data according to the assumptions of the model
  - We’ll get an empirical distribution of how many black jurors we can expect
    - Under the assumptions of the model!
  - We’ll check if the observed data (8% black jurors) is **consistent** with random selection (from a population of which 26% is black)

(notebook)

https://en.wikipedia.org/wiki/Swain_v._Alabama

Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

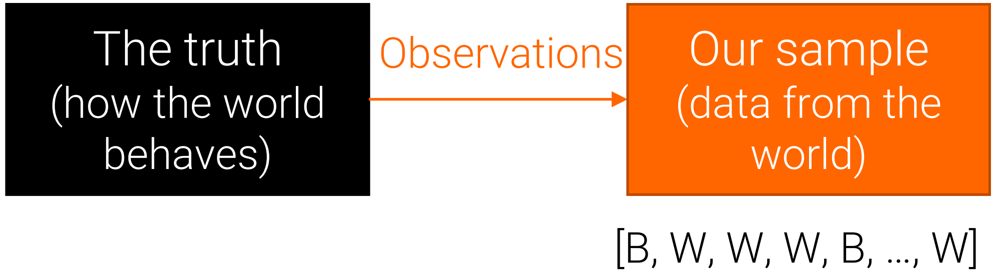
Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

The truth
(how the world
behaves)

Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

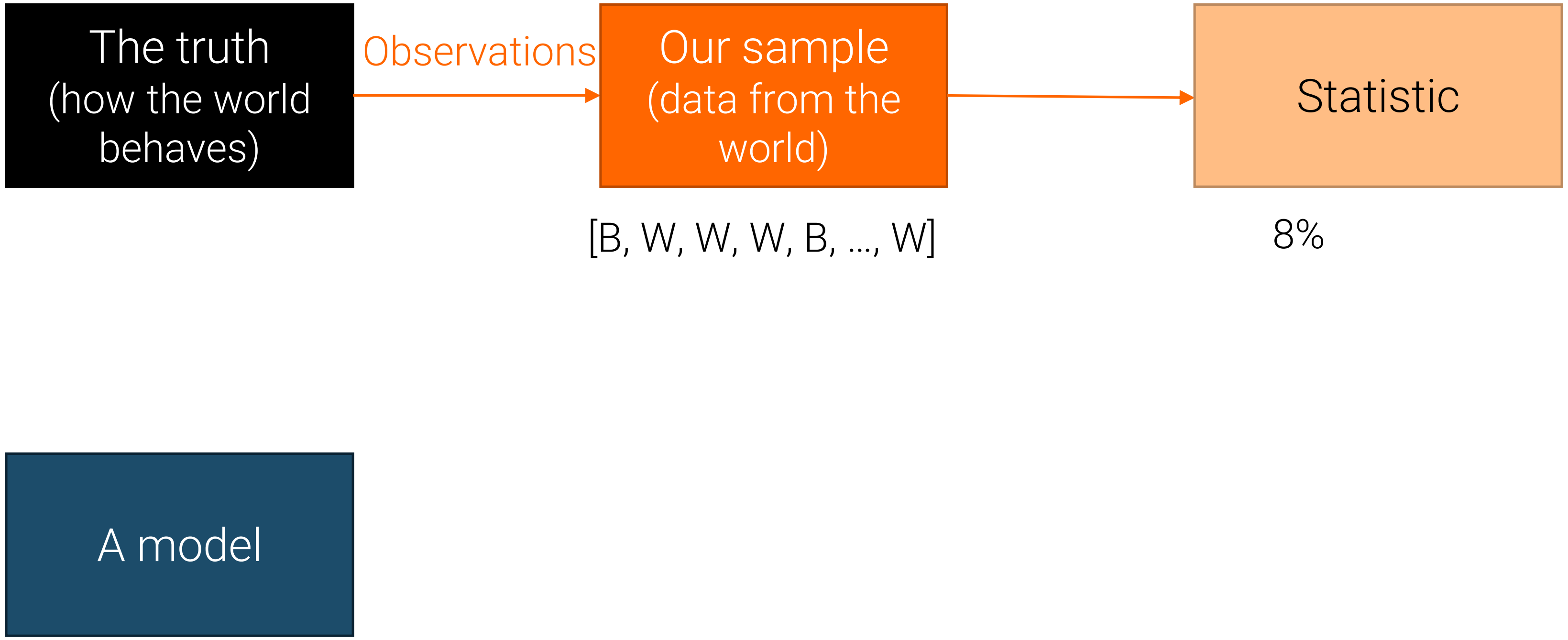


Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

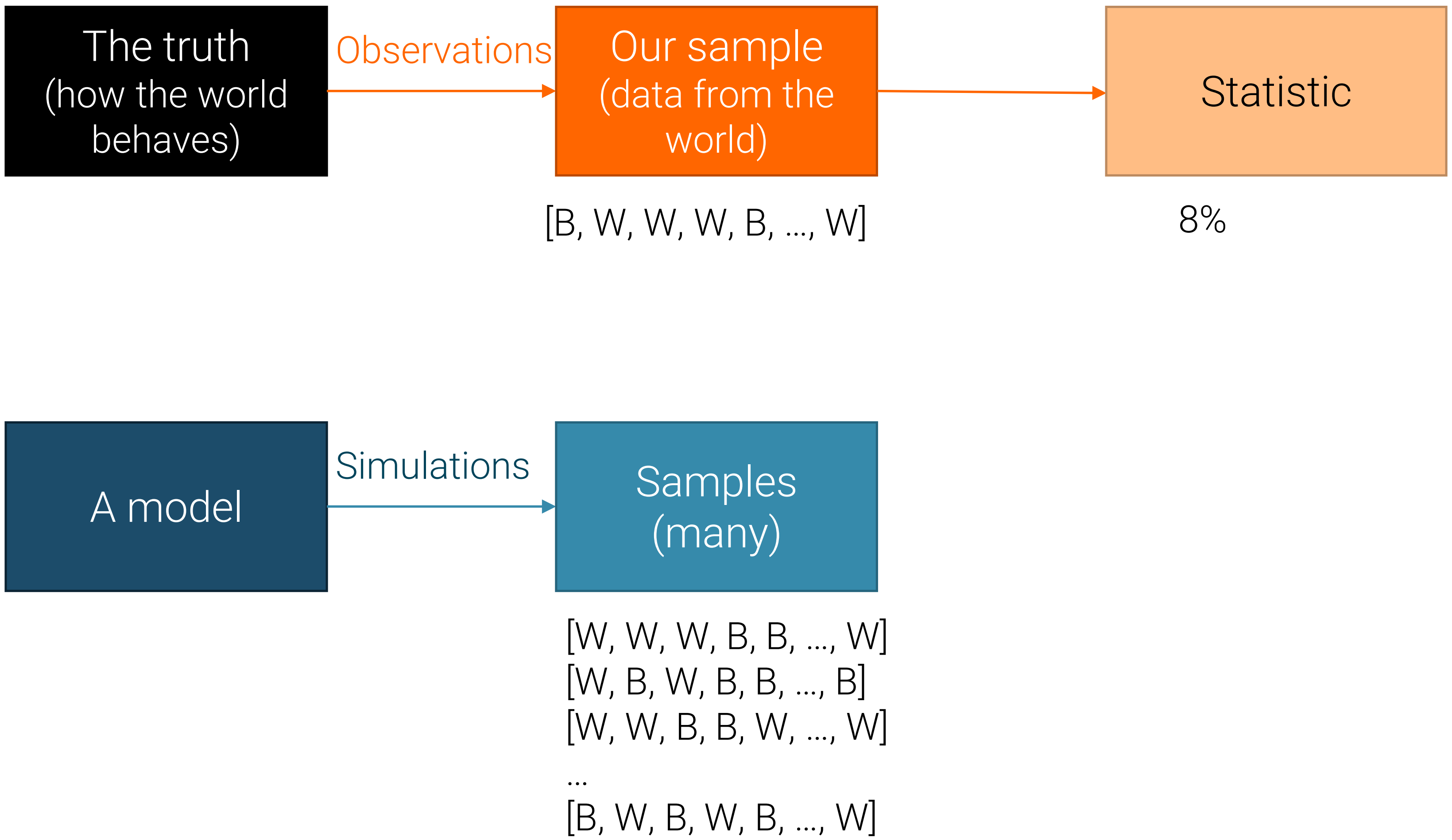
Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!



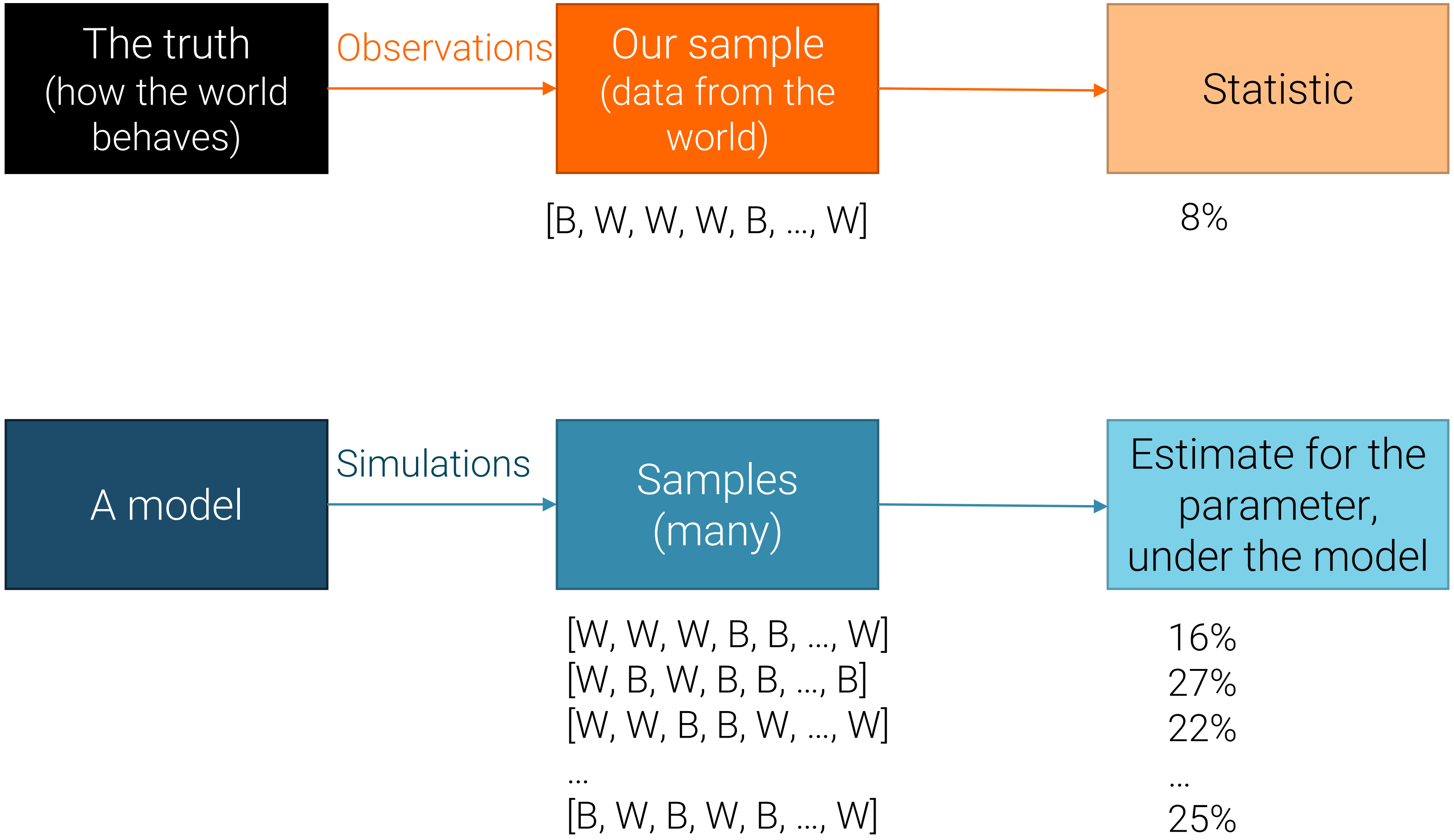
Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!



Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!



Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

The truth (how the world behaves) → Observations → Our sample (data from the world) → Statistic

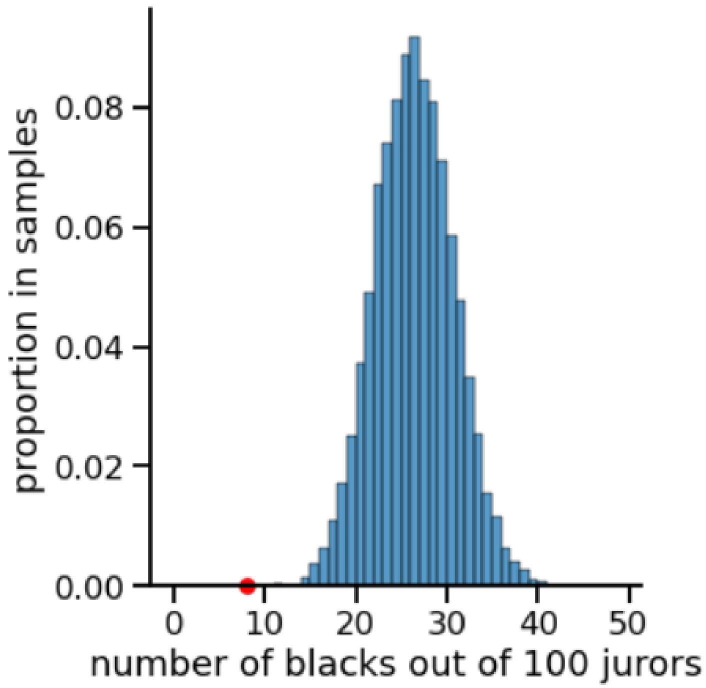
[B, W, W, W, B, ..., W]

8%

A model → Simulations → Samples (many) → Estimate for the parameter, under the model

[W, W, W, B, B, ..., W]
[W, B, W, B, B, ..., B]
[W, W, B, B, W, ..., W]
...
[B, W, B, W, B, ..., W]

16%
27%
22%
...
25%



Assessing models using simulations

We want to know whether our model is correct, but we only have a sample!

The truth (how the world behaves) → Observations → Our sample (data from the world) → Statistic

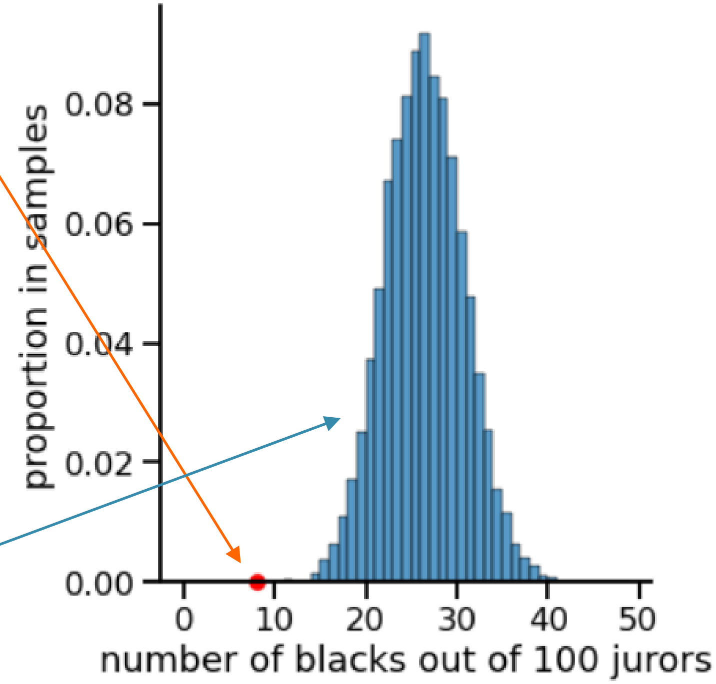
[B, W, W, W, B, …, W]

8%

A model → Simulations → Samples (many) → Estimate for the parameter, under the model

[W, W, W, B, B, …, W]
[W, B, W, B, B, …, B]
[W, W, B, B, W, …, W]
…
[B, W, B, W, B, …, W]

16%
27%
22%
…
25%



Incomplete information

- Two views of the world
  - “The model reflects the truth”
  - “The model does not reflect the truth”
- We want to choose between them based on data in a sample.
  - But random samples can turn out to be extreme (perhaps unlikely, but possible)
    - So both views *can* be consistent with the data
- We need a formal test to decide between the two views, using the incomplete information we have

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two hypotheses: views of how data were generated

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two **hypotheses**: views of how data were generated
- Null hypothesis – H_0 **(השערת האפס)**
  - Data were generated by a well-specified chance model
  - And any variation from the model which is observed in the data is due to chance

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two **hypotheses**: views of how data were generated
- Null hypothesis – H_0 **(השערת האפס)**
  - Data were generated by a well-specified chance model
    - The selection of jury was random
    - The coin is fair
    - And also, the coin is biased with P(H) = 0.75
  - And any variation from the model which is observed in the data is due to chance

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two **hypotheses**: views of how data were generated
- Null hypothesis – H_0 **(השערת האפס)**
  - Data were generated by a well-specified chance model
    - The selection of jury was random
    - The coin is fair
    - And also, the coin is biased with $P(H) = 0.75$
  - And any variation from the model which is observed in the data is due to chance
  - We need to be able to simulate data under the assumptions of the null hypothesis

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two **hypotheses**: views of how data were generated
- Null hypothesis – H_0 **(השערת האפס)**
  - Data were generated by a well-specified chance model
    - The selection of jury was random
    - The coin is fair
    - And also, the coin is biased with $P(H) = 0.75$
  - And any variation from the model which is observed in the data is due to chance
  - We need to be able to simulate data under the assumptions of the null hypothesis
- Alternative hypothesis – H_1
  - There is an effect other than chance: The model is incorrect
  - Variation from the model observed in the data is not only because of chance

Testing Hypotheses

Step 1: The Hypotheses

- A test chooses between two **hypotheses**: views of how data were generated
- Null hypothesis – H_0 (השערת האפס)
  - Data were generated by a well-specified chance model
    - The selection of jury was random
    - The coin is fair
    - And also, the coin is biased with P(H) = 0.75
  - And any variation from the model which is observed in the data is due to chance
  - We need to be able to simulate data under the assumptions of the null hypothesis
- Alternative hypothesis – H_1
  - There is an effect other than chance: The model is incorrect
  - Variation from the model observed in the data is not only because of chance
  - We may not have a clue on how the data *is* generated

Testing Hypotheses

Step 2: The Test Statistic

- A value that can be computed for the data and for samples
  - And that we can therefore simulate/compute analytically
- Questions to ask:
  - What values of the test statistic will make us favor the null hypothesis?
  - What values of the test statistic will make us favor the alternative?

Testing Hypotheses

Step 2: The Test Statistic

- A value that can be computed for the data and for samples
  - And that we can therefore simulate/compute analytically
- Questions to ask:
  - What values of the test statistic will make us favor the null hypothesis?
  - What values of the test statistic will make us favor the alternative?

Step 3: The Sampling distribution of the Test Statistic

- What the test statistic might be **if** the null hypothesis were true
- Approximate the sampling distribution by an empirical distribution

Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the observed value of the test statistic with its empirical distribution **under the null hypothesis**
- If the observed value is **not consistent** with the distribution, then the test favors the alternative – “rejects the null hypothesis”

Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the observed value of the test statistic with its empirical distribution **under the null hypothesis**
- If the observed value is **not consistent** with the distribution, then the test favors the alternative – "rejects the null hypothesis"

Consistency with the distribution:

- A visualization may be enough
- There are conventions about what is "consistent enough" (in a few slides)

Back to the jury example

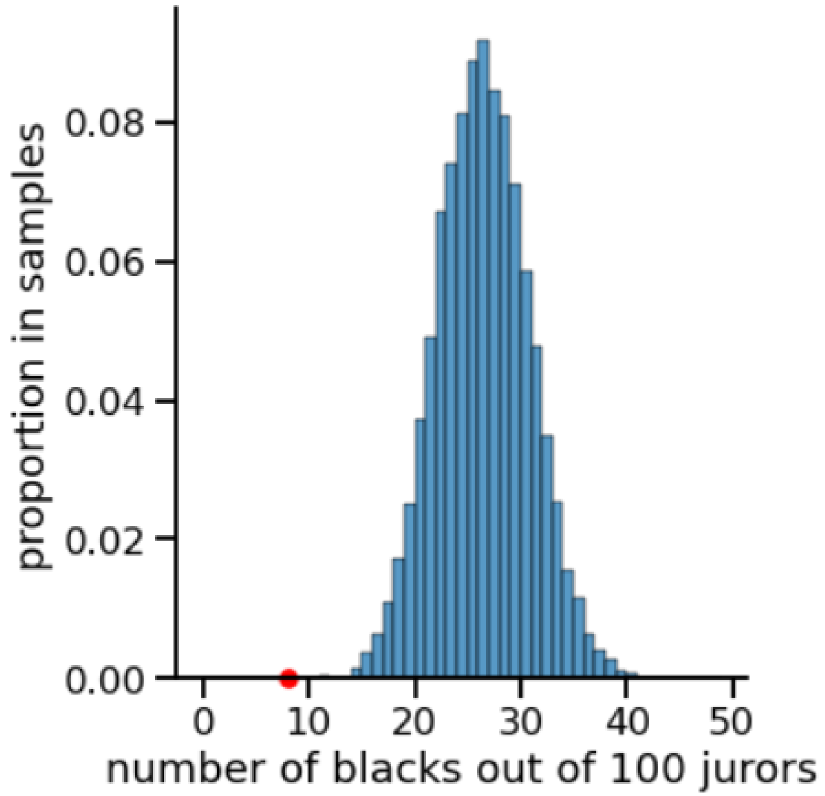
- What are the two hypotheses?
- What is the value of the test statistic?
- Was the statistic consistent with the distribution?
- What will be the conclusion from the test?

What is “consistent”?

- When the value of the test statistic is in the tails of the (empirical) distribution of the statistic simulated under the null hypothesis (the model), we say it is inconsistent with the null hypothesis

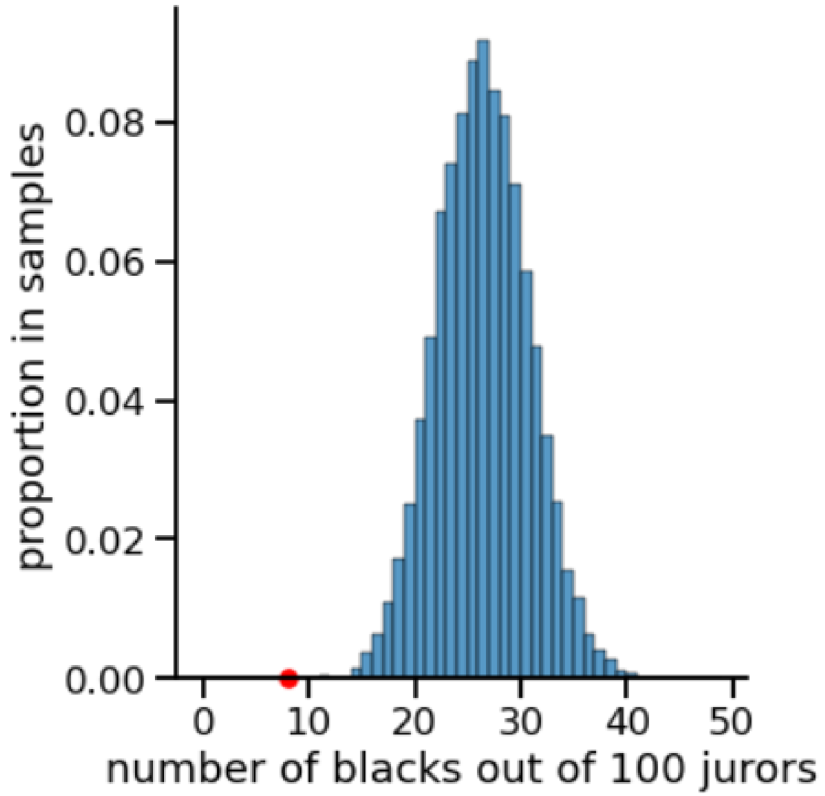
What is “consistent”?

- When the value of the test statistic is in the tails of the (empirical) distribution of the statistic simulated under the null hypothesis (the model), we say it is inconsistent with the null hypothesis



What is “consistent”?

- When the value of the test statistic is in the tails of the (empirical) distribution of the statistic simulated under the null hypothesis (the model), we say it is inconsistent with the null hypothesis



- But what if it is not this clear?

Conventions about inconsistency

- “Inconsistent”: The test statistic is in the tail of the empirical distribution under the null hypothesis
- “In the tail,” convention:
  - The area in the tail is less than 5%
  - The result is then “statistically significant”

Conventions about inconsistency

- “Inconsistent”: The test statistic is in the tail of the empirical distribution under the null hypothesis
- “In the tail,” convention:
    - The area in the tail is less than 5%
    - The result is then “statistically significant”
        - (Note this does not mean it is necessarily important)

Definition of the *P*-value

The *P*-value is the probability,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

Definition of the *P*-value

The *P*-value is the probability,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

ההסתברות לקבל במדגם (עבור סטטיסטי המבחן) ערך השווה לזה שקיבלנו בפועל, או קיצוני יותר בכיוון ההשערה האלטרנטיבית, **תחת ההנחה שהשערת האפס נכונה.**

(notebook)

Can the conclusion be wrong?

Can the conclusion be wrong?

- Yes.

	Null is true	Null is false
Test rejects the null	✗	✓
Test doesn't reject the null	✓	✗

Can the conclusion be wrong?

- Yes.

Different worlds, and we do not know in which of these worlds we live

	Null is true	Null is false
Test rejects the null	X	✓
Test doesn't reject the null	✓	X

Can the conclusion be wrong?

- Yes.

Different worlds, and we do not know in which of these worlds we live

Different samples may lead to different conclusions

	Null is true	Null is false
Test rejects the null	X	✓
Test doesn’t reject the null	✓	X

An error probability

- The cutoff we use for the *P*-value **is** an error probability.

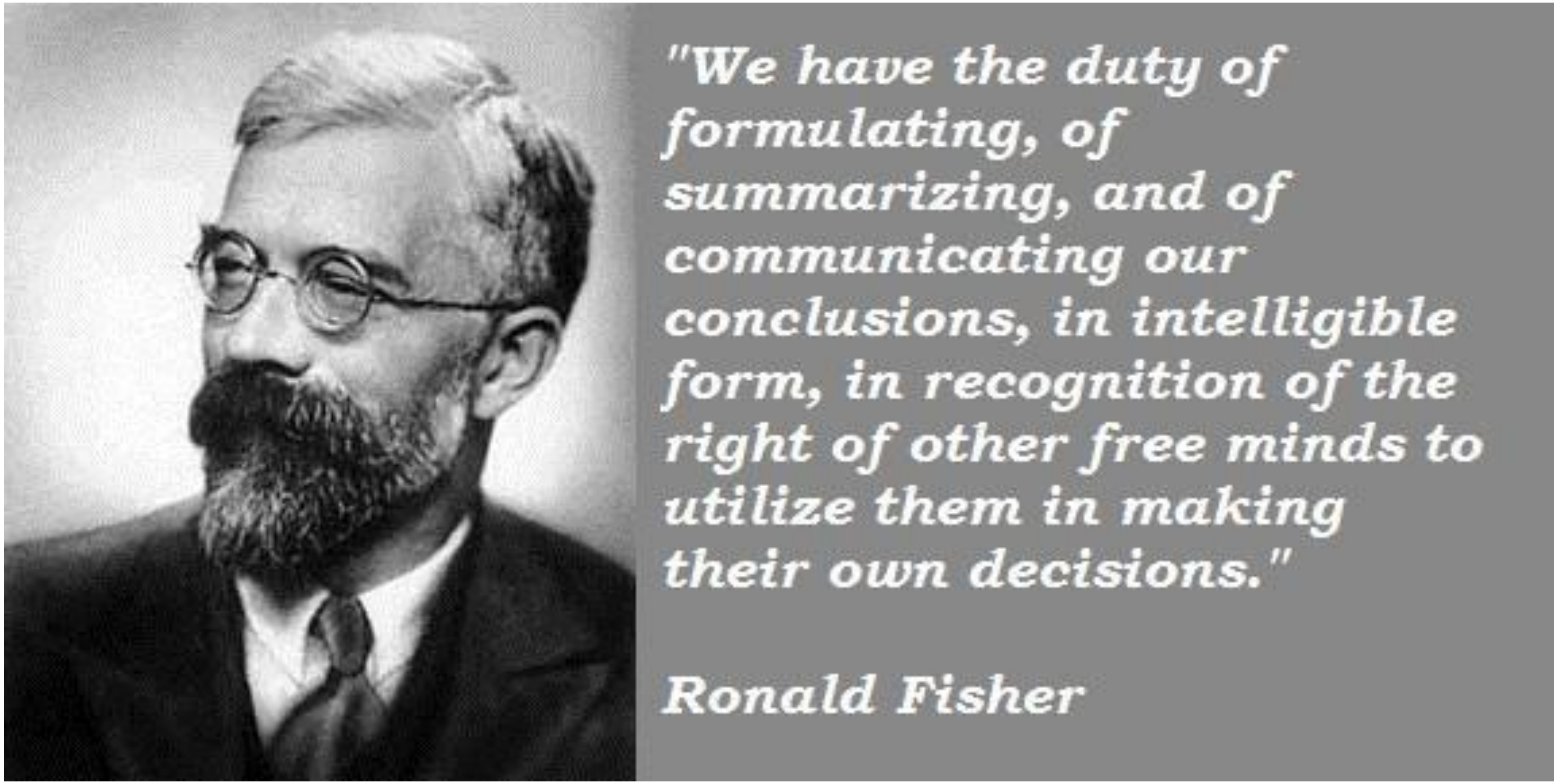
- If
  - your cutoff is 5%
  - and you live in a world in which the null hypothesis is true

- then there is about a 5% chance that your test will (wrongly) reject the null hypothesis.

An error probability

- The cutoff we use for the *P*-value is an error probability.
- If
  - your cutoff is 5%
  - and you live in a world in which the null hypothesis is true
- then there is about a 5% chance that your test will (wrongly) reject the null hypothesis.
- (there is also another type of possible error)

The origin of the 5% convention



Sir Ronald Fisher

“It is **convenient** to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not.”

–– *Statistical Methods for Research Workers*

“If one in twenty does not **seem** high enough odds, **we may, if we prefer it**, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). **Personally**, the author **prefers** to set a low standard of significance at the 5 percent point …”