# Statistical theory:
# Null hypothesis significance testing

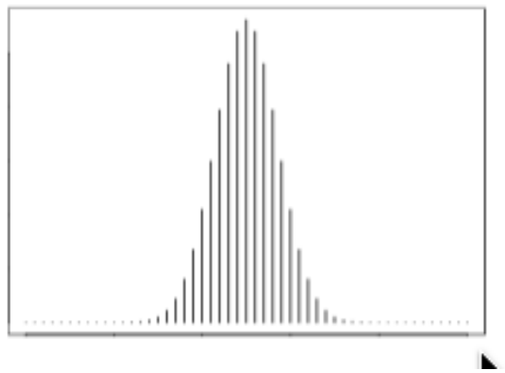
Research Methods for Human Inquiry
Andrew Perfors

# So far we have learned about:

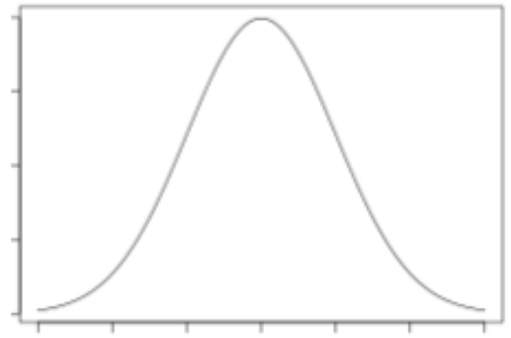
* Probability and different probability distributions

and how to calculate the probability of different data given those distributions

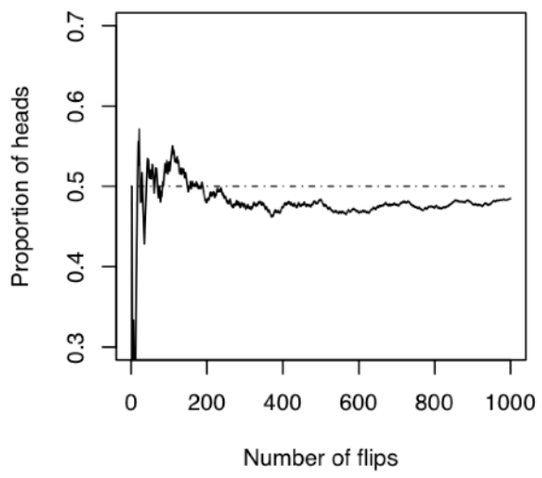
Binomial



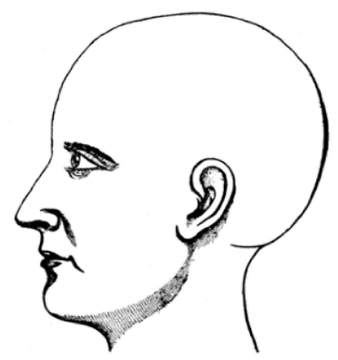
Normal



Frequentists




Bayesians



# So far we have learned about:

* Probability and different probability distributions

    and how to calculate the probability of different data given those distributions

* How experiments (ideally) involve randomly sampling from a population, and incorporate three kinds of statistics

| thing | "usual" symbol | thing | "usual" symbol | what is it? | do we know its value? |
|---|---|---|---|---|---|
| **true population mean** | $\mu$ | **true population sd** | $\sigma$ | the truth | no |
| **estimated population mean** | $\hat{\mu}$ | **estimated population sd** | $\hat{\sigma}$ | a statistical inference | yes |
| **sample mean** | $\bar{X}$ or $M$ | **sample sd** | $s$ | a description of our dataset | yes |

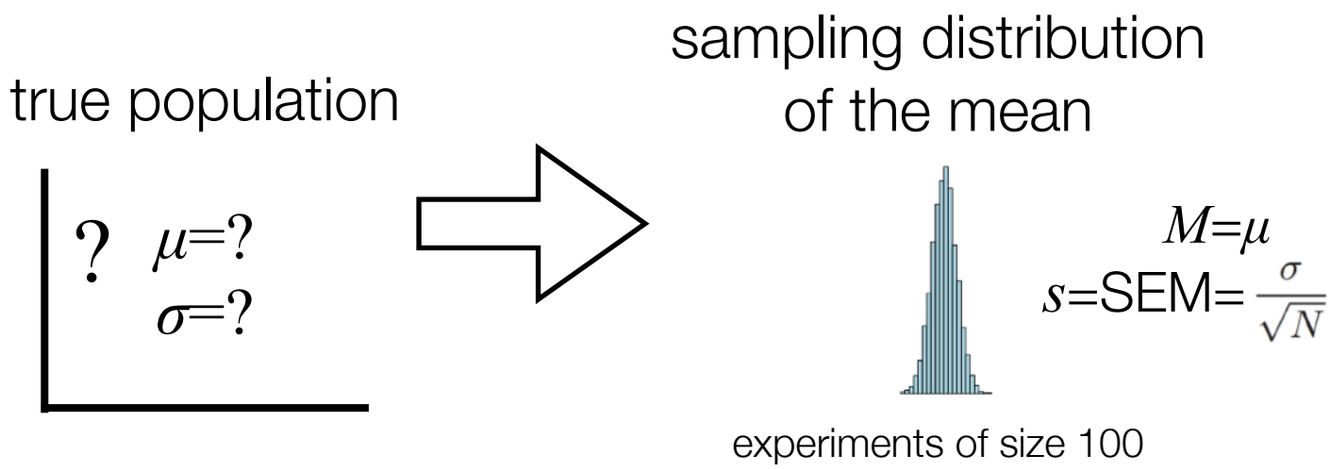
# So far we have learned about:

* Probability and different probability distributions

and how to calculate the probability of different data given those distributions

* How experiments (ideally) involve randomly sampling from a population, and incorporate three kinds of statistics

* The central limit theorem, which says the sampling distribution of the mean will always be normal (and its standard deviation given by the SEM) regardless of the shape of the true distribution



# So far we have learned about:

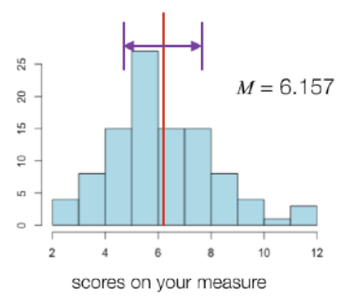
* Probability and different probability distributions

and how to calculate the probability of different data given those distributions

* How experiments (ideally) involve randomly sampling from a population, and incorporate three kinds of statistics

* The central limit theorem, which says the sampling distribution of the mean will always be normal (and its standard deviation given by the SEM) regardless of the shape of the true distribution

* How this means that we can calculate a confidence interval around our sample mean which covers the true mean 95% of the time




$$\mathrm{CI}_{95} = \bar{X} \pm 1.96 \frac{\hat{\sigma}}{\sqrt{N}}$$

```
ciMean(x)
```

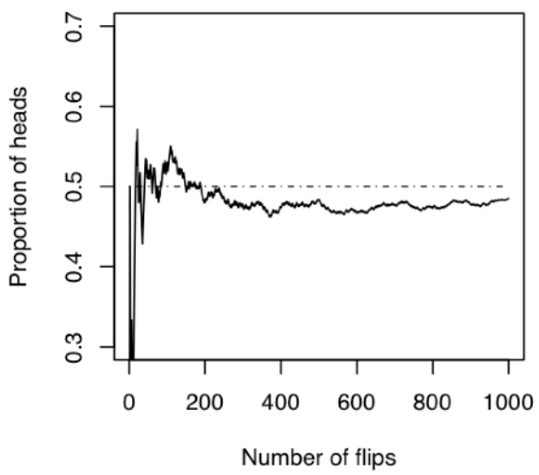
Now, let's bring it together: How do we use all these pieces of knowledge to form inferences (beyond the CI) about our data?

# Null hypothesis significance testing (NHST)

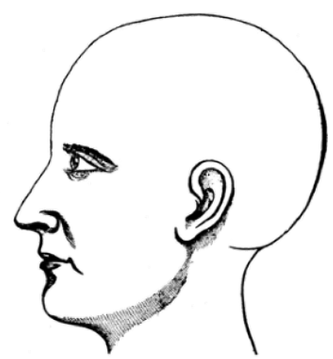
# Basic idea of hypothesis testing


# Two schools of thought…

## Frequentists



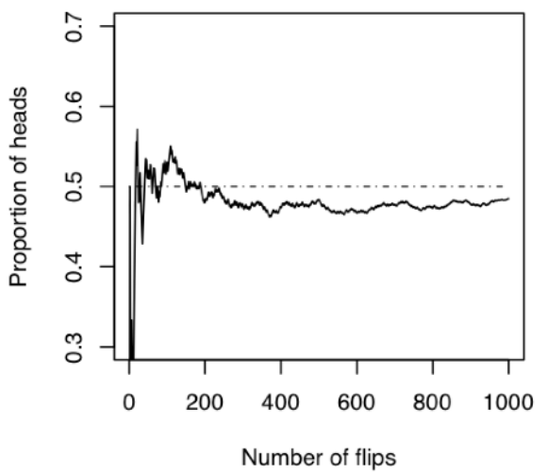

## Bayesians



# Basic idea of hypothesis testing


# Two schools of thought…

## Frequentists




We'll be focusing here, because this is mostly what people in psychology use now.

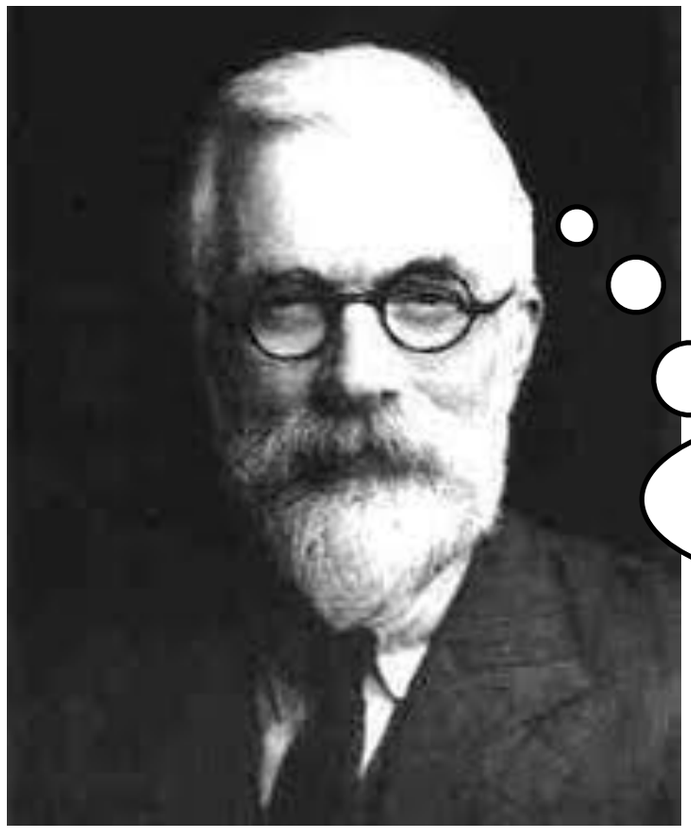
Called:
orthodox hypothesis testing
null hypothesis significance testing (NHST)

# The history is somewhat relevant

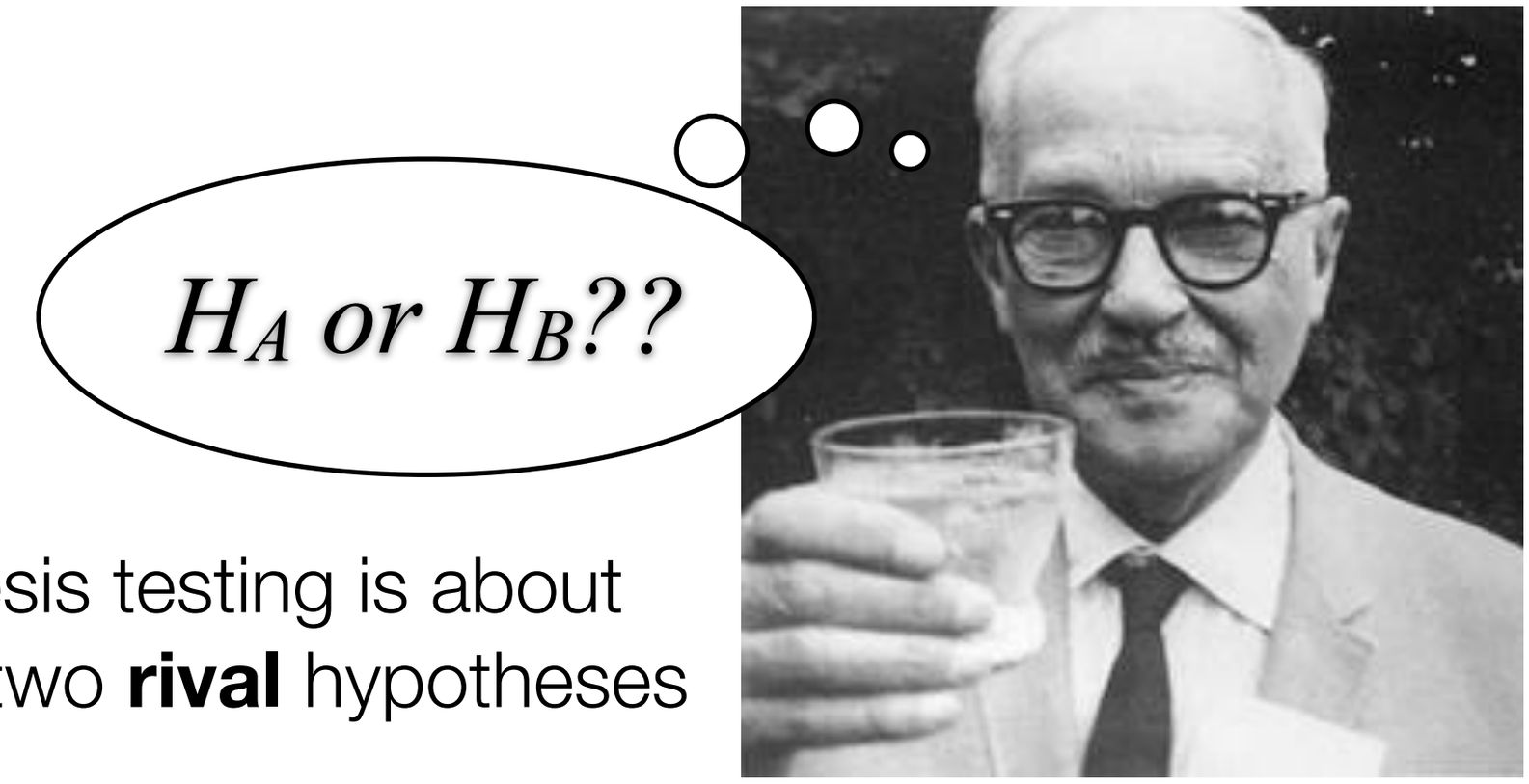
R. A. Fisher J. Neyman

- Big fight in the early 20th century
- Fisher vs Neyman
  - Both frequentists, but…
  - Different views on hypothesis testing
- “Orthodox NHST” is a weird hybrid
  - It borrows from both theories
  - It does so in a messy way
  - Neither author would agree with it!

# The fundamental difference




Fisher: hypothesis testing is about trying to falsify a **single** hypothesis




Neyman: hypothesis testing is about choosing between two **rival** hypotheses

# The fundamental difference

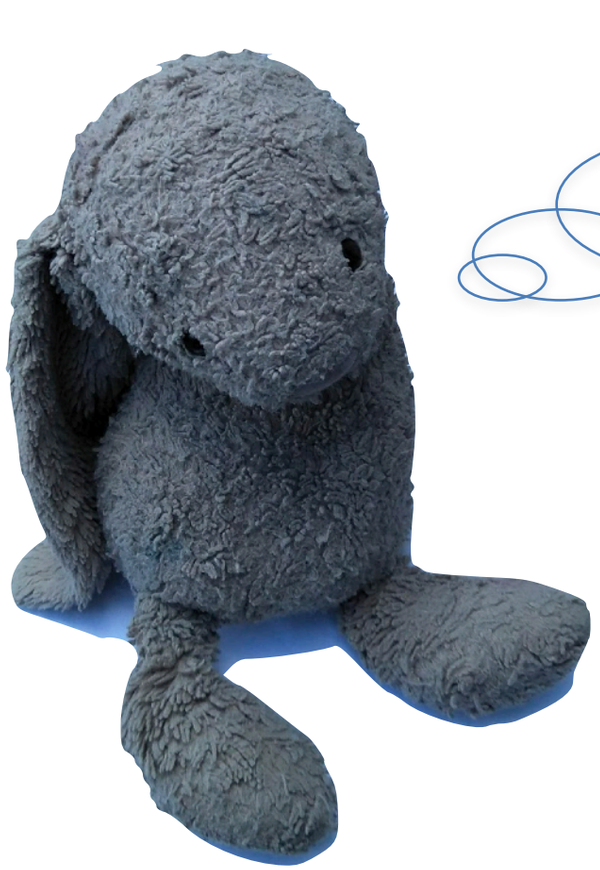
The details *really* don't matter for this class

I mention is because NHST is really weird and counterintuitive in some ways. This is because science is a human construction that (like everything) was shaped by historical choices, not simply what was objectively "best" in any real sense.

# Null hypothesis significance testing

(the way most people do it)

# Bunny has a research topic



I want to know if people agree with each other about whether we're running out of food

# Bunny and Gladly run a study to test it

- They ask N = 100 people this question: “Do you think we have less food now?”
- The task is to answer either YES or NO (no other options)

59


41

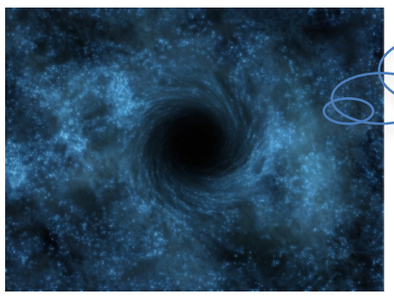
# Two different issues

- **Is the study valid?** (research methods)
  - Maybe the experimental design has holes in it?
  - Was the question badly phrased in any way?
  - How did they sample their participants?
- **Do the data suggest that there is a real effect?** (statistics)
  - i.e. How much evidence does “59 from 100” imply?
- To draw any conclusions we need both
  - Must have sound methodology (covered earlier)
  - Must have statistical evidence (let’s talk about this…)

# For any topic there are two kinds of hypotheses one could have…

The null hypothesis: there is no effect

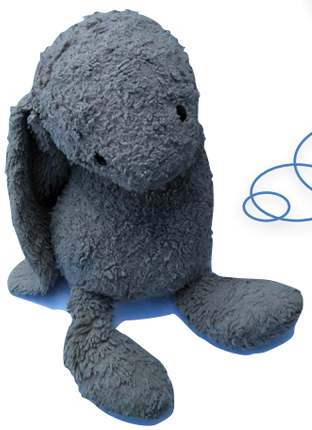
$H_0$



People are evenly split in their views

The alternative is the opposite of the null: there is some effect. Usually this is what I want to find (not always though!)

$H_1$



People agree with one another about whether we're running out of food

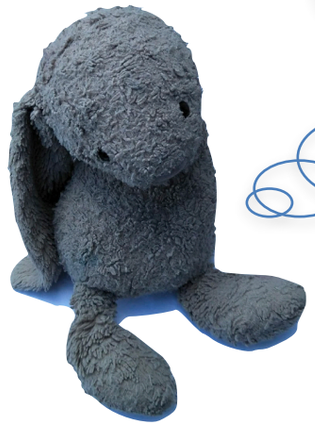
# But null hypothesis significance testing is really weird

You'd think we would just test the alternative hypothesis…

… but that is mathematically more difficult, and also there are often many alternatives (that I don't really care about the difference between, e.g., 80/20 vs 90/10, who cares)

The alternative is the opposite of the null: there is some effect. Usually this is what I want to find (not always though!)

$H_1$

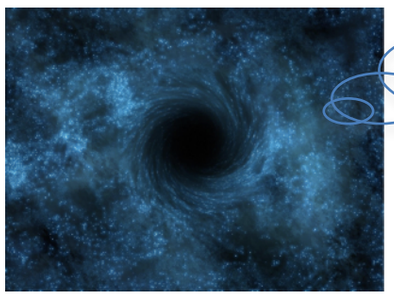


People agree with one another about whether we're running out of food

# But null hypothesis significance testing is really weird

The null hypothesis: there is no effect

$H_0$



People are evenly split in their views

**Instead we always test the null hypothesis.**

How likely were we to have seen the data we did if the null hypothesis was true?
If we weren't very likely, we reject the null hypothesis.

**Hugely important point:** *Any* statistical claim you make when doing NHST is a claim about the null hypothesis, *not* the alternative

# Research hypotheses and statistical hypotheses

# **Research hypotheses** are claims about psychological constructs

People agree with one another about whether we are running out of food

**Statistical hypotheses** are claims about population parameters

The probability $\theta$ that people answered YES is not 0.5

**Statistical hypotheses** are claims about population parameters

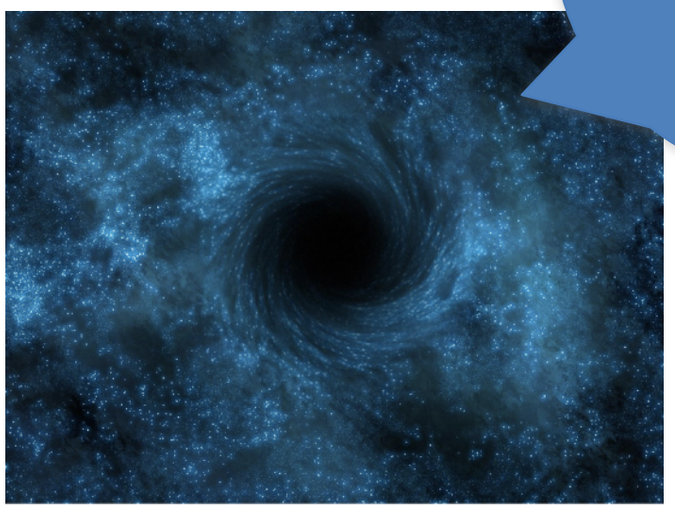
Alternative hypothesis…

$\theta \neq 0.5$

**Statistical hypotheses** are claims about population parameters

Null hypothesis…

$\theta = 0.5$



A critical part of designing a study is being able to specify how your *research hypothesis* (which is what you're interested in) maps onto your *statistical hypothesis* (which is what you can actually test)

A (real) scientist starts out with a research hypothesis

The statistical consequences of the research hypothesis

The exact opposite statistical hypothesis

| | | |
|---|---|---|
| Research hypothesis | | "people agree with each other about whether we're running out of food"   |
| Statistical hypothesis | $H_0 : \theta = 0.5$ (the null) | $H_1 : \theta \neq 0.5$ (the alternative) |

**This (in blue)** is the implicit competition set up by the hypothesis test

It maps onto an implicit competition between research hypotheses

| | | |
|---|---|---|
| Research hypothesis | "people disagree with each other about whether we're running out of food"   | "people agree with each other about whether we're running out of food"   |
| Statistical hypothesis | $H_0 : \theta = 0.5$ (the null) | $H_1 : \theta \neq 0.5$ (the alternative) |

We evaluate the competition by determining **how likely our data would be if the null is true**. If they are unlikely "enough", we reject the null.

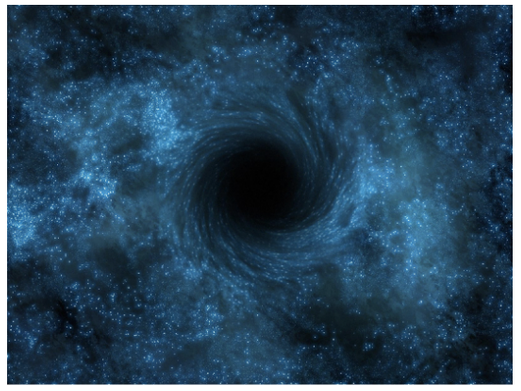
| | | |
|---|---|---|
| Research hypothesis | "people disagree with each other about whether we're running out of food"   | "people agree with each other about whether we're running out of food"   |
| Statistical hypothesis | $H_0 : \theta = 0.5$ (the null) | $H_1 : \theta \neq 0.5$ (the alternative) |

# Statistical decision making

$H_0$ is true

$H_0$ is false

Is the null hypothesis true? There's two possibilities about the actual state of the world

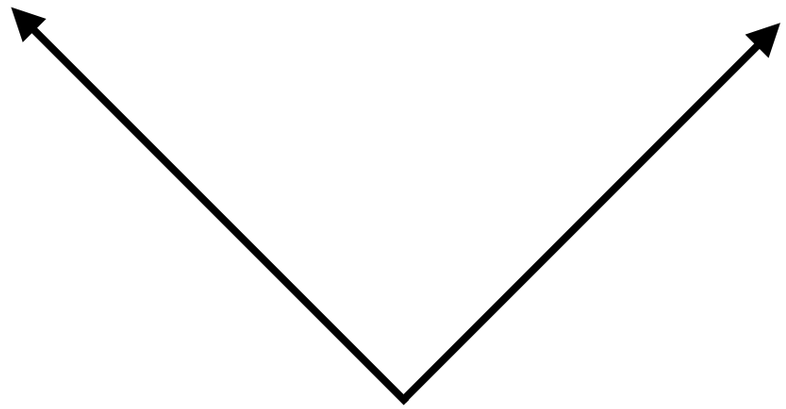


accept $H_0$

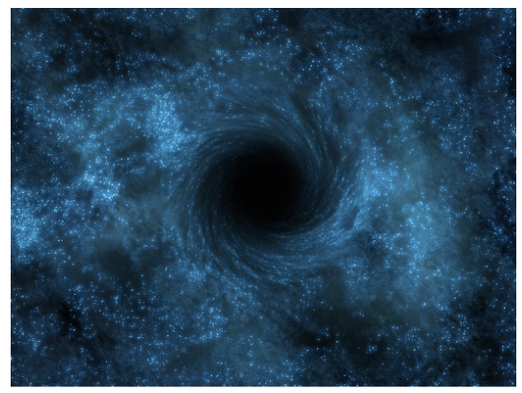
reject $H_0$

$H_0$ is true

$H_0$ is false



And do we decide to agree with it, or not? Again, two possibilities exist for what decision we make



| | accept $H_0$ | reject $H_0$ |
|---|---|---|
| $H_0$ is true | Correct! | Type I error: false positive |
| $H_0$ is false | Type II error: false negative | Correct! |

Central design principle of NHST:
**control** the Type I error rate

| | accept $H_0$ | reject $H_0$ |
|---|---|---|
| $H_0$ is true | | $\alpha$<br>Type I error rate |
| $H_0$ is false | $\beta$<br>Type II error rate | $1-\beta$<br>power |

There's a tradeoff here, because power depends on sample size, effect size, and $\alpha$: all else equal, lower $\alpha$ means higher $\beta$

Central design principle of NHST: **control** the Type I error rate

| | accept $H_0$ | reject $H_0$ |
|---|---|---|
| $H_0$ is true | | $\alpha$<br>significance level |
| $H_0$ is false | $\beta$<br>Type II error rate | $1-\beta$<br>power |

# Why is NHST designed this way?

- Essentially it's about controlling belief bias…
  - Type I error is when we incorrectly believe H1
  - i.e., falsely accepting a desirable conclusion (a false positive)
  - That's the "dangerous" form of belief bias!
- It forces a minimum evidentiary standard on you
  - Before you're "allowed" to endorse your preferred conclusion…
  - … you must demonstrate that you actually have sufficient evidence.
- (Aside: it's arguable whether it works, but that's for later)

# How are power, alpha, and sample size related again!?

- Let’s have a look at this tool that lets us play around with them! (also linked on the LMS)

https://rpsychologist.com/d3/nhst/

See the `w5day2exercises.Rmd` file for the exercises!