Experimental design and hypothesis testing: From association to causality

Methods 1, E2021 - Lecture 6
Tuesday 11/10/2021
Fabio Trecca


QUIZ
TIME

Quiz time (1)

1. $cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$

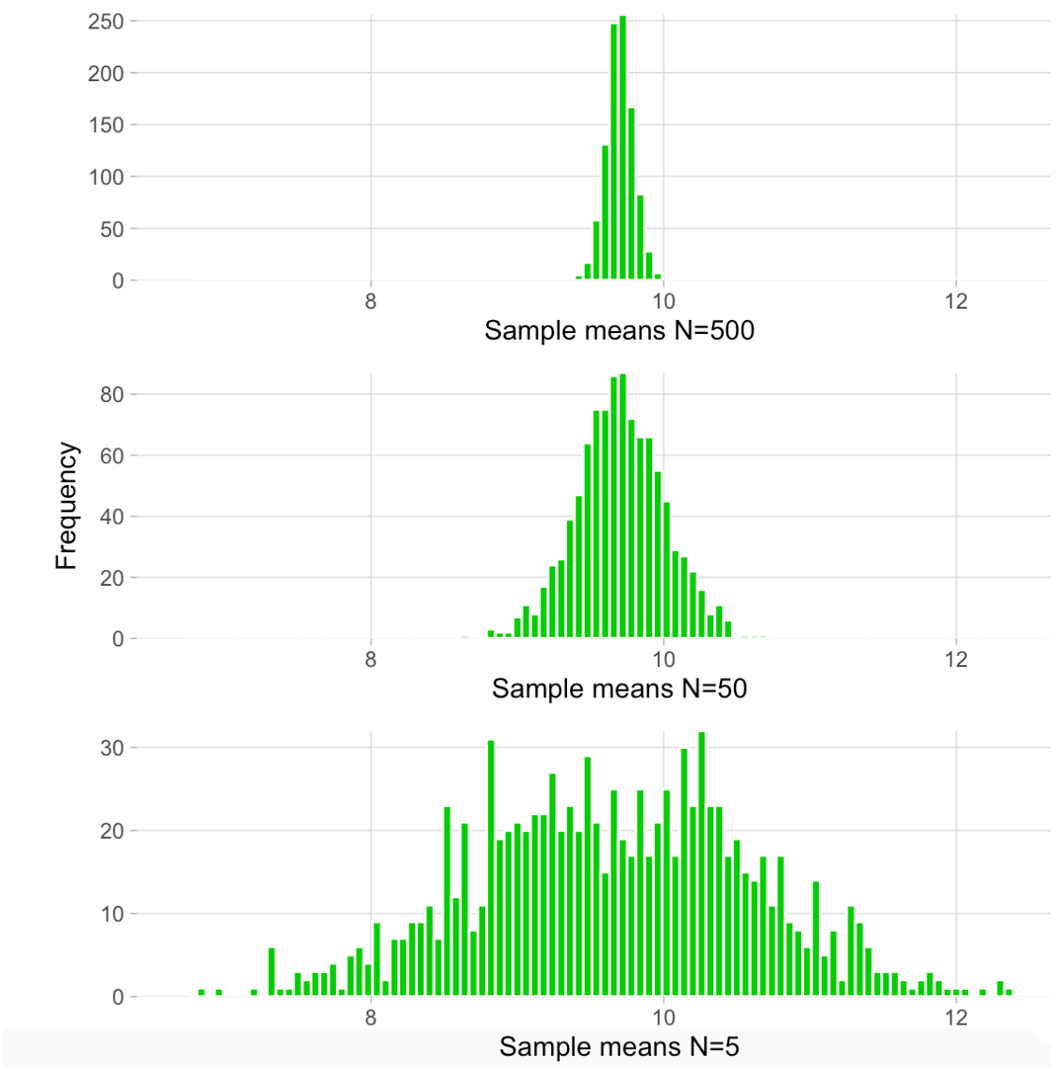
2. $r = \frac{cov(x, y)}{s_x s_y}$

- r, ρ
- R^2

Quiz time (2)

- $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N} \rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

- $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$



Two main types of empirical studies in CogSci

- Quasi-experimental studies
- Full experimental studies

True experiments:	Quasi-experiments:
Emphasize internal validity ▪ Assess cause & effect (in relatively artificial environment) ▪ Test clear, a priori hypotheses	**Emphasize external validity** ▪ Describe "real" / naturally occurring events ▪ Clear *or* exploratory hypotheses
Participants randomly assigned to exp. or control groups ▪ Participants & experimenter Blind to assignment	**Non-equivalent groups** ▪ Existing groups ▪ Non-random assignment ▪ Participants not blind ▪ Self-selection
Control study procedures ▪ Manipulate independent variable ▪ Control procedures & measures	**Full control may not be possible** ▪ May not be able to manipulate the independent variable ▪ Partial control of procedures & measures

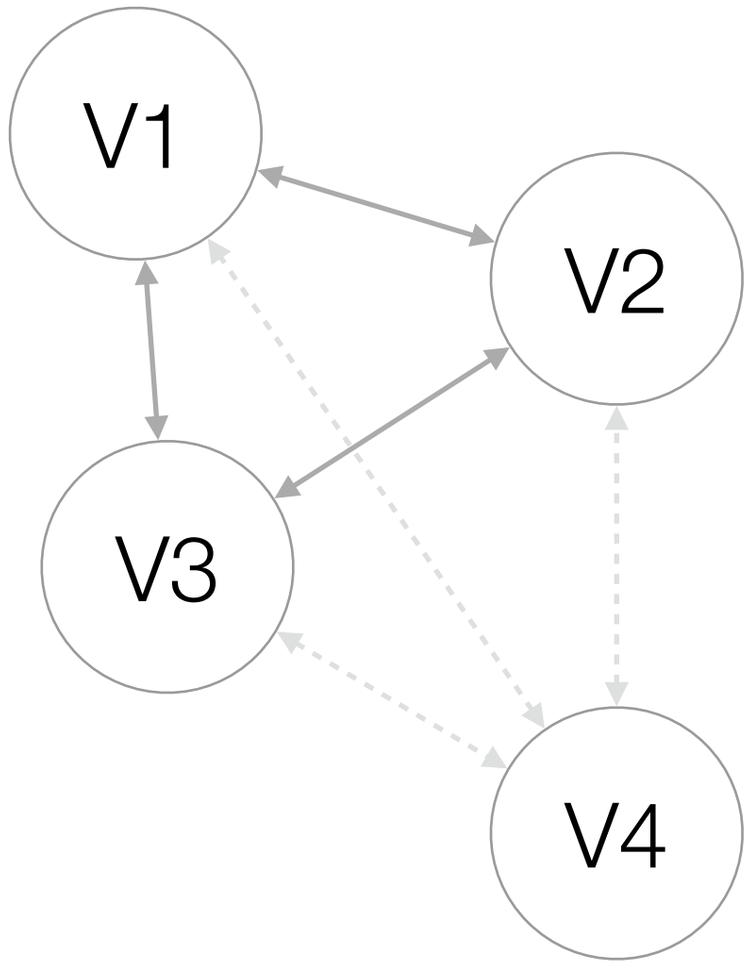
Quasi-experiments (1)

- Collection of data about two or more naturally occurring variables in the world/ lab (= correlation)

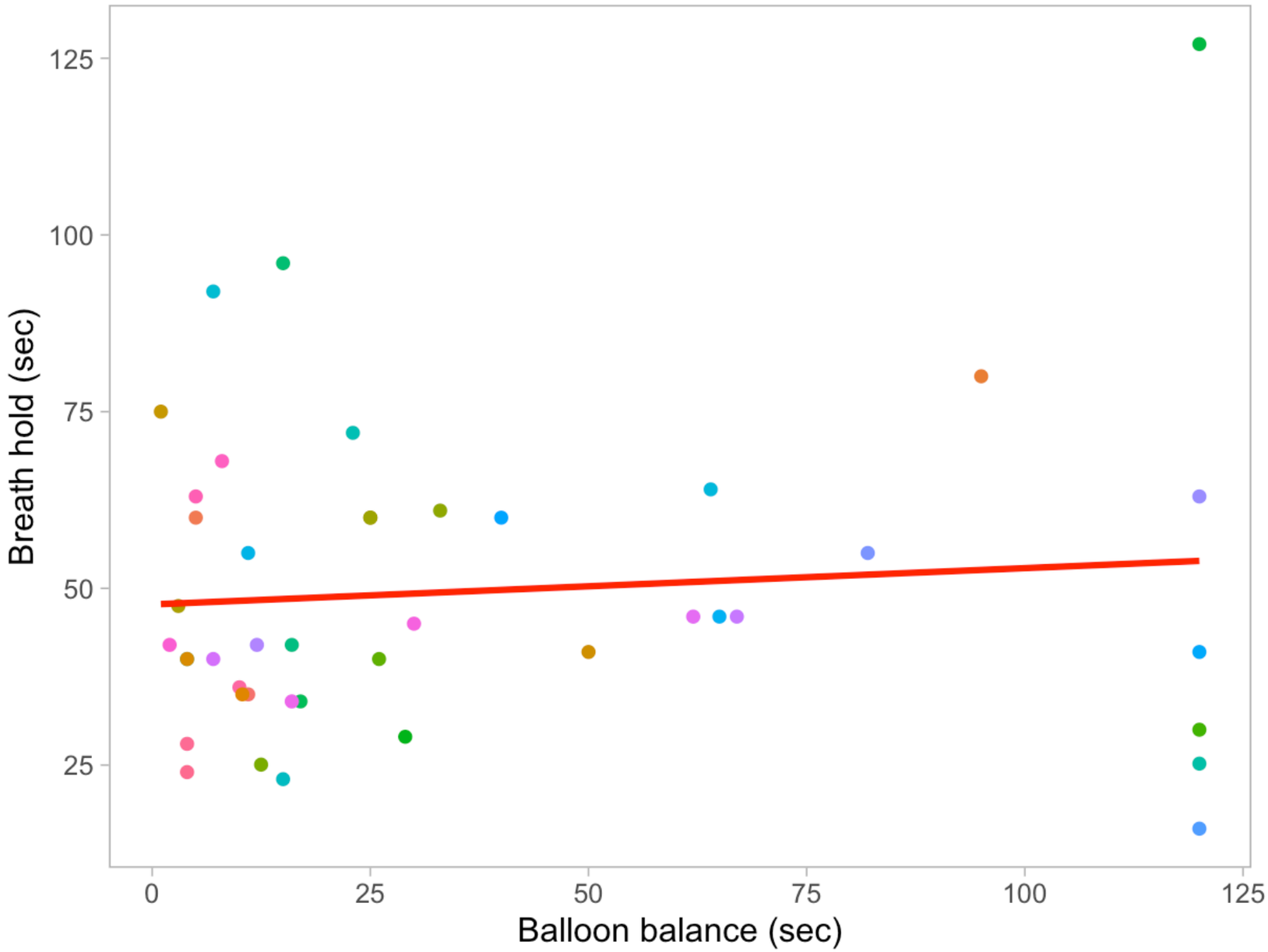
- *e.g., mother-child MLU, shoe size/breath hold, word length/reading times*

- Advantages:
  ecological validity

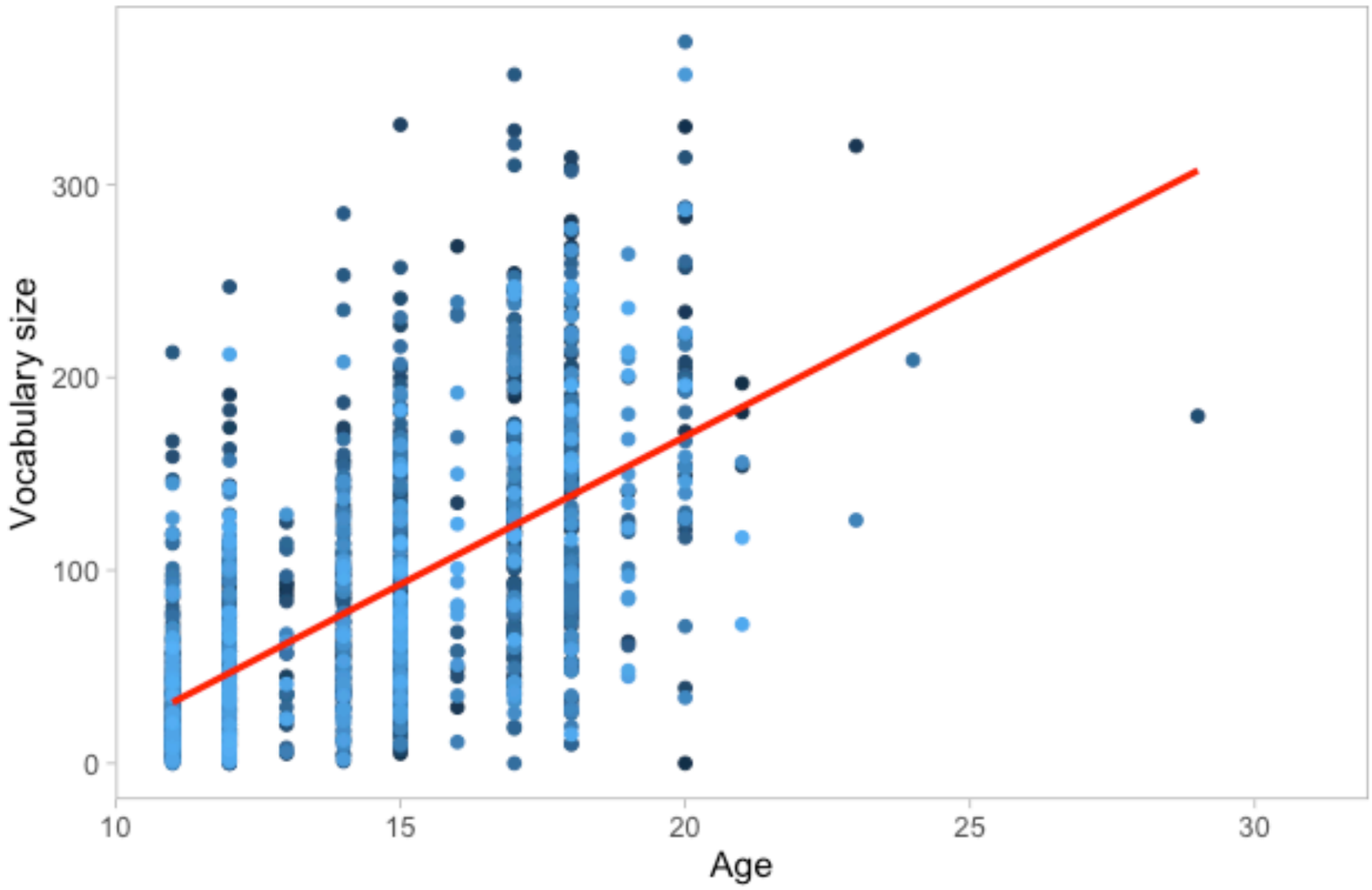
- Disadvantages:
  correlation ≠ causation, hidden variables?



Quasi-experiments (2)

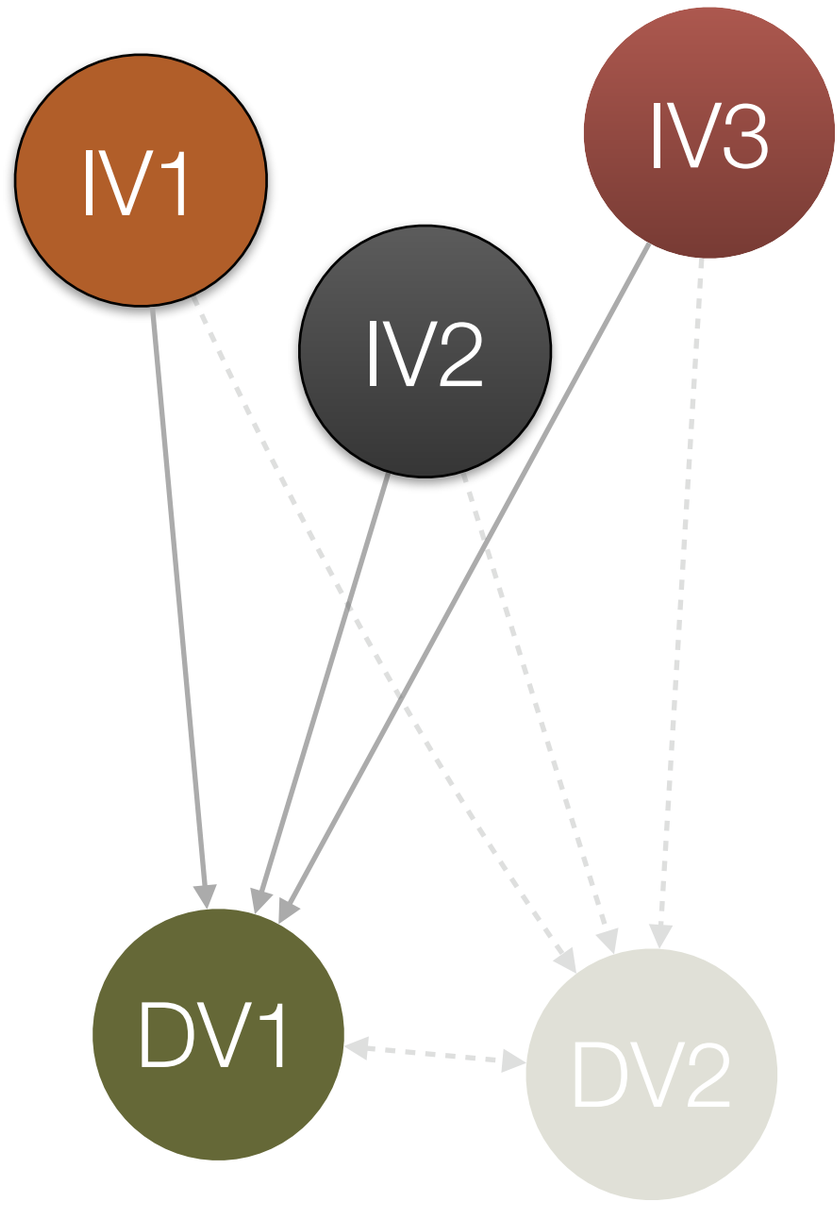


Quasi-experiments (3)



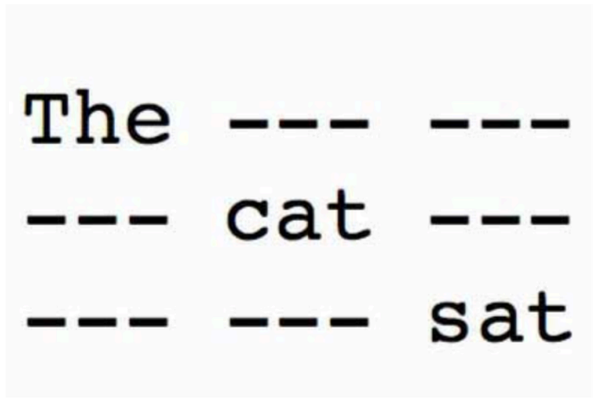
Full-fledged experiments (1)

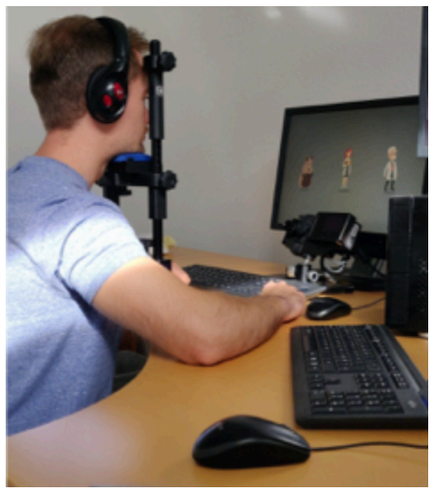
- Variables are manipulated systematically to observe changes in their relation
  - causality and directionality
  - dependent variable(s) and independent variable(s)
  - randomized control
- Advantages:
  easier to isolate causal relation
- Disadvantages:
  lower ecological validity

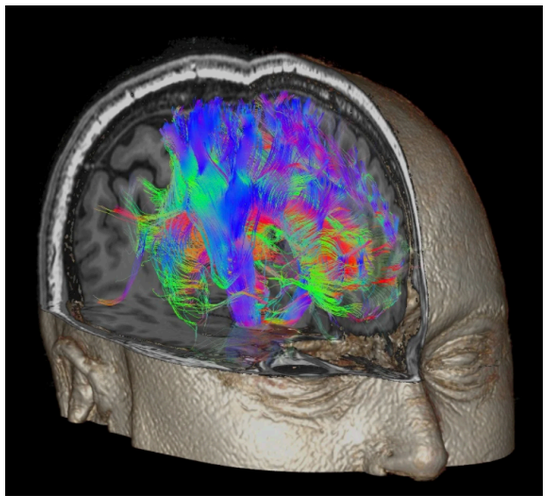


Full-fledged experiments (2)

- Different methods







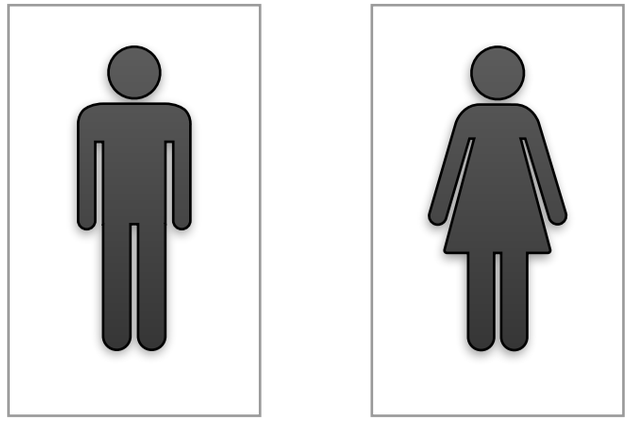
… and many more!

Full-fledged experiments (3)

- Brainstorming:
  - does breath hold capability affect balloon balance skills?
  - testing the effect of insomnia on cognitive performance?
  - investigating whether larger pupil sizes make people more attractive?
  - are apes less responsive than humans to collaborative signals?
  - does it take longer time to read an unexpected word in a text?

Experimental manipulation (1)

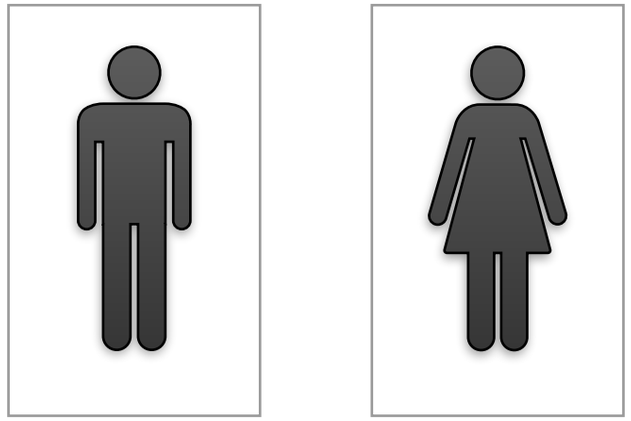
- **Independent measures design/ Between-participant design:**
  - Groups of participants
  - One group, one condition
  - How do the different groups fare?
  - Advantages: manipulation is not explicit
  - Disadvantages: hard to balance groups perfectly



Condition 1 **Condition 2**

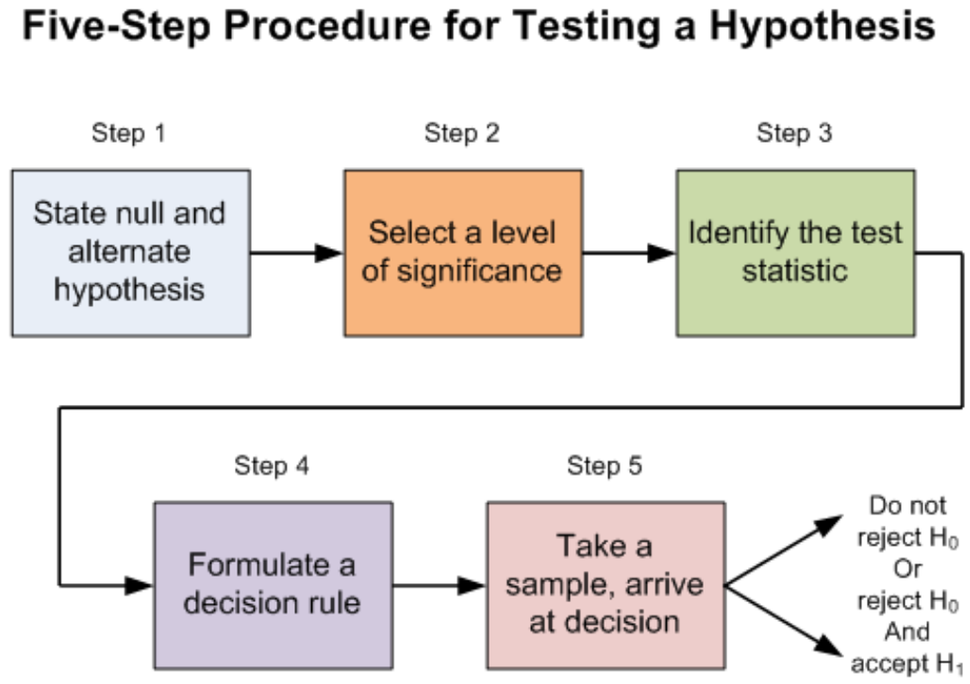
Experimental manipulation (2)

- **Repeated measures design/ Within-participant design:**
  - All participants go through all conditions
  - How do people fare on the different conditions?
  - Advantages: good for controlling individual differences
  - Disadvantages: order effects, explicitness of manipulation



Hypothesis testing (1)

- When we run studies, we typically want to test hypotheses
- Null hypothesis significance testing (NHST)
  - **H_0** (null hypothesis) = No difference between the means
  - **H_1** (alternative hypothesis) = Difference between the means
- We can't prove the H1, but we can reject the H0



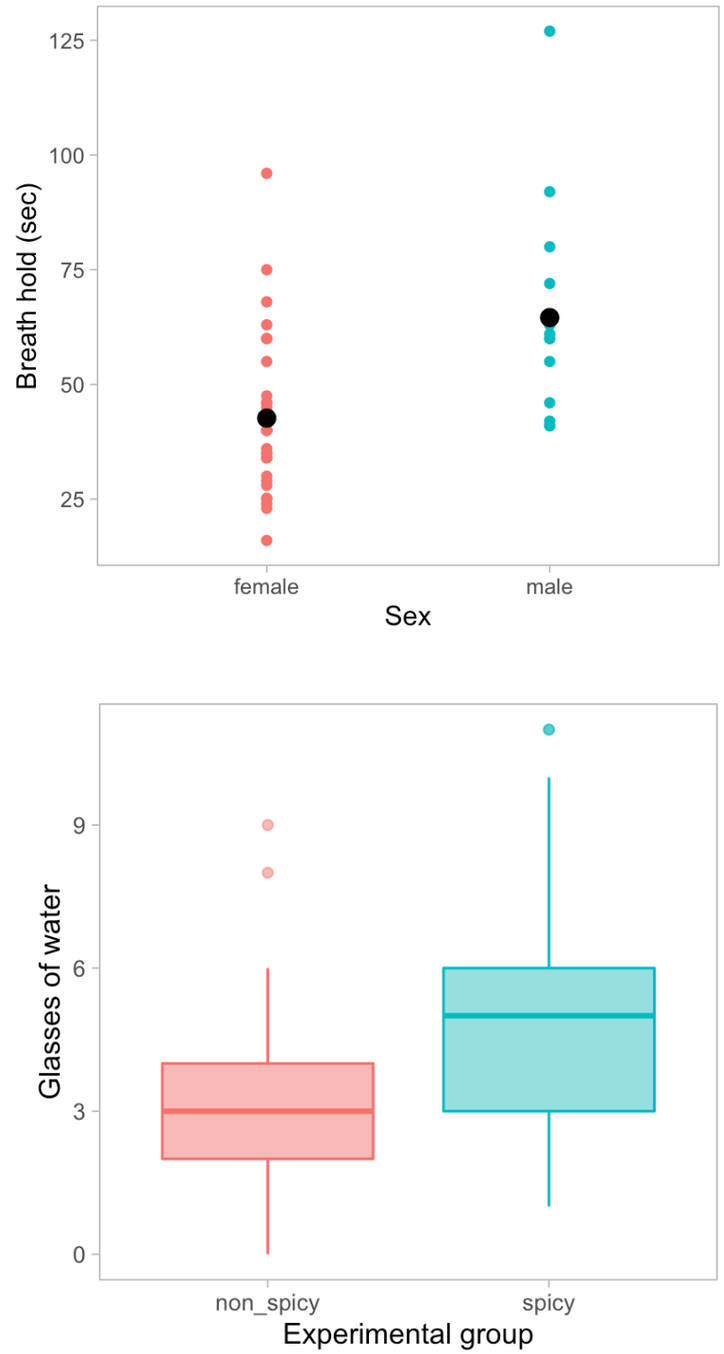
Hypothesis testing (2)

- Often we want to compare two (or more) groups

- eg., as a consequence of our experimental design

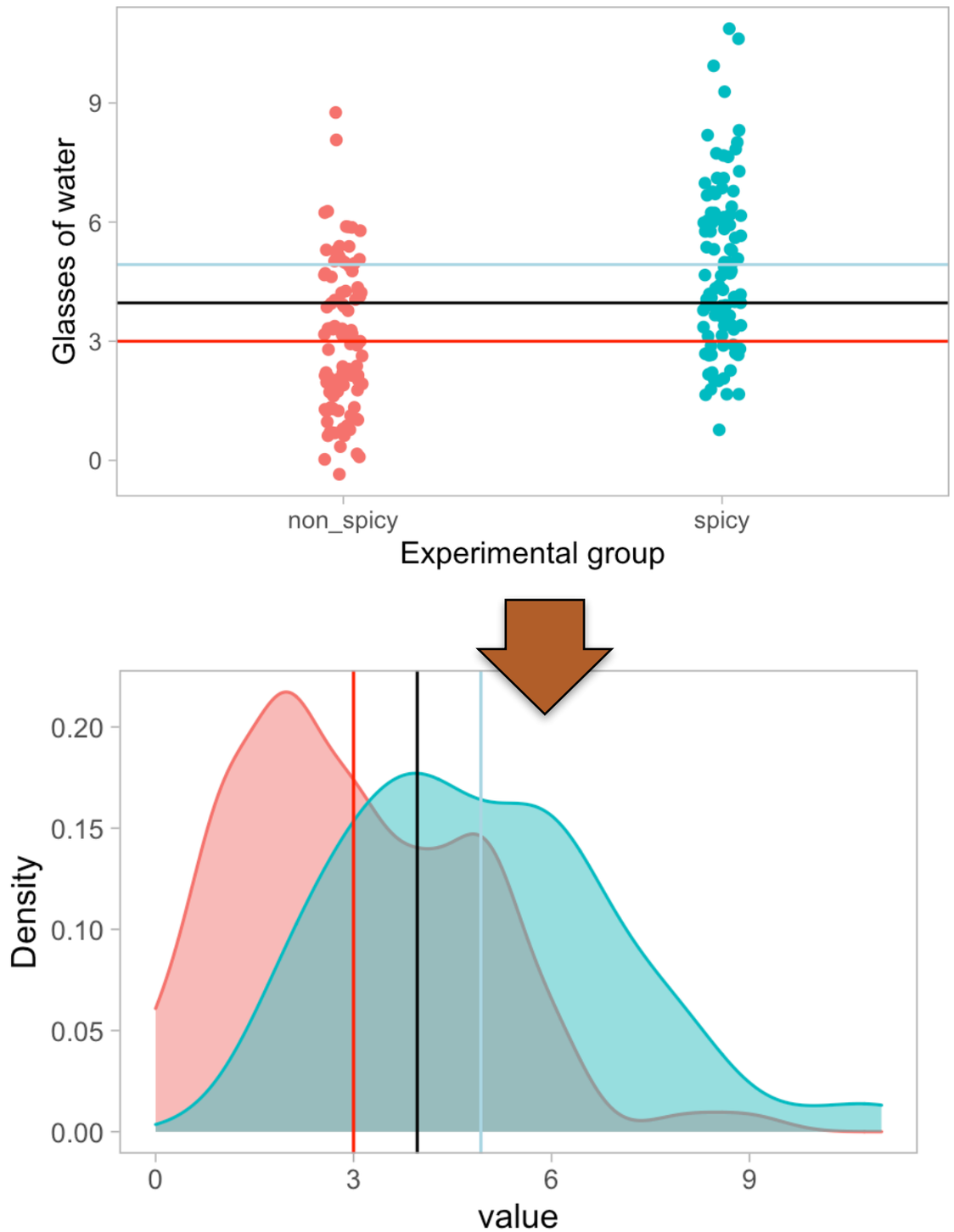
- How do we know whether the difference is real/reliable or just due to noise?

- Is this result likely to happen again if we rerun the experiment?



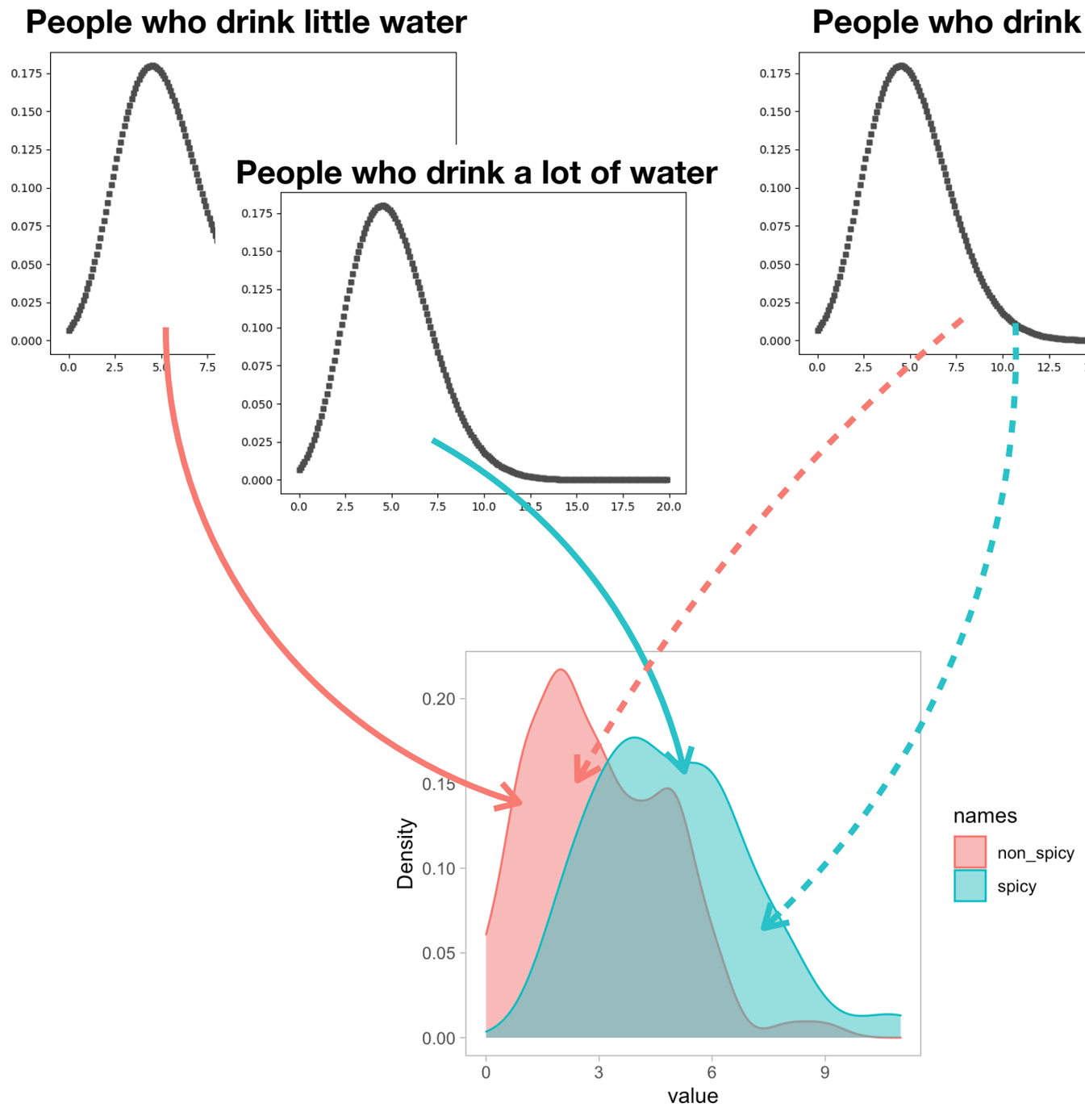
Example: Does spicy food make you thirsty?

- We formulate our hypotheses:
  - **H_0** = Spicy food makes you drink just as much as non-spicy food
  - **H_1** = Spicy food makes you drink more than non-spicy food
- We collect data and plot them



Concept: Same vs different populations

- Do the differences in our samples reflect differences that also exist in the population **(→ true effect)**
  - *"Do our groups come from different populations"?*
- Or from the same underlying population? **(→ sampling noise?)**
  - *"Do our groups come from the same population?"*



The t-test (*Student's test of statistical significance*)

- Inferential statistic used to determine if there is a **statistically significant difference** between the means of two groups

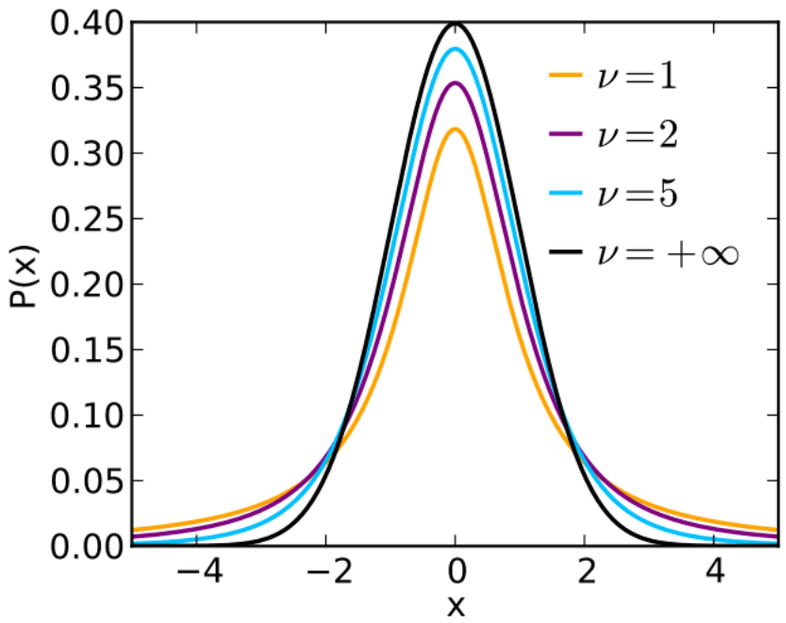
- Compares means/variances of two data sets and determine if they came from the same population

- t-test → t-statistic

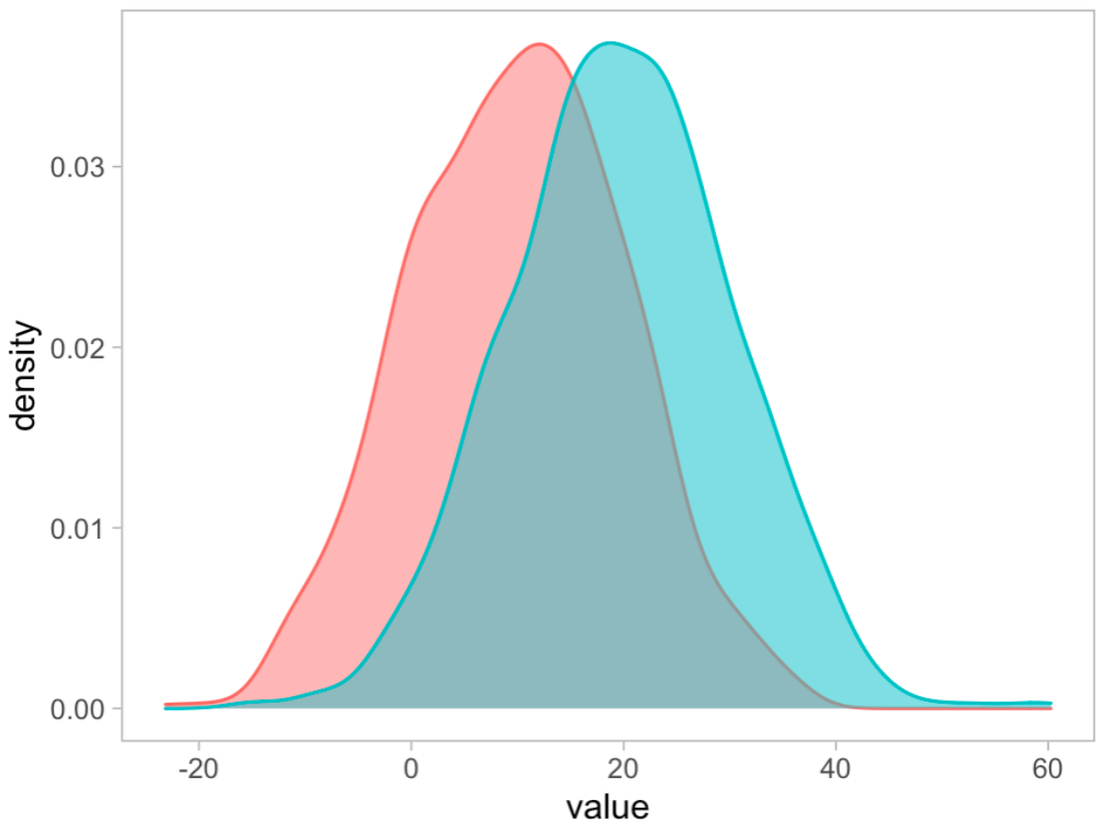
  - Under the null hypothesis, the t-statistic follows the t-distribution

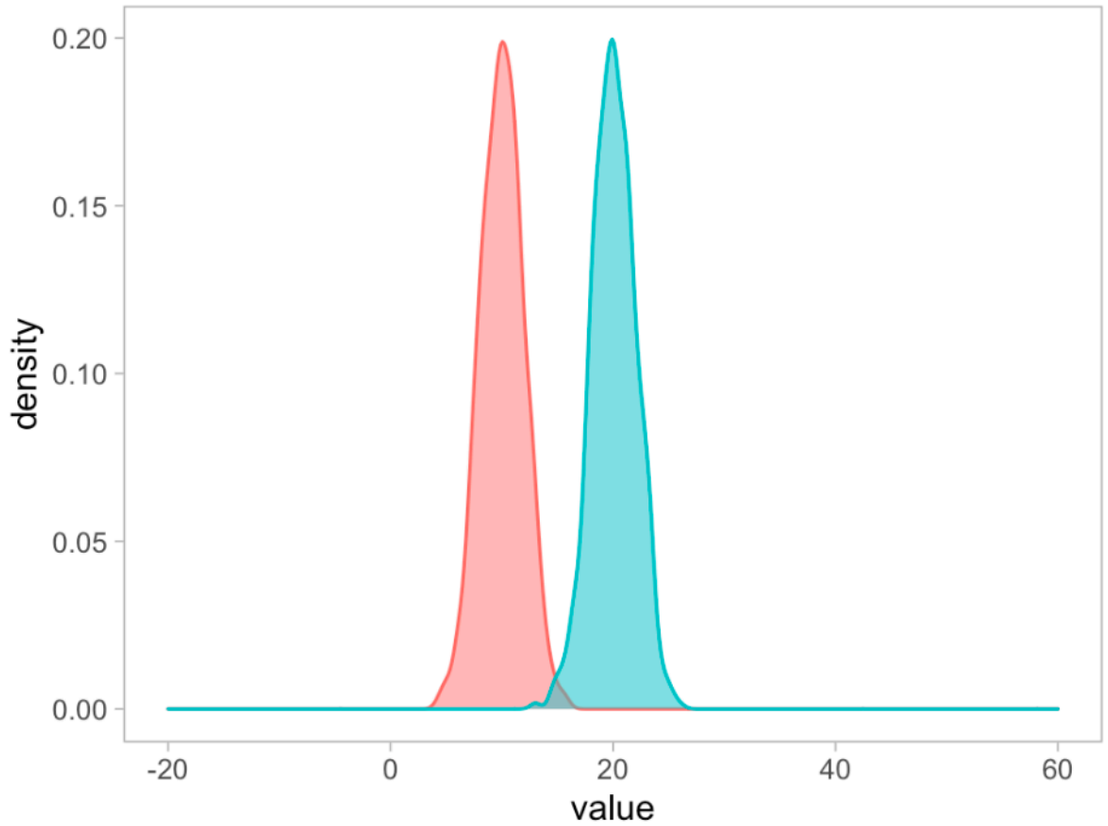
  - This allows us to estimate whether differences in means are statistically significant by looking at how "common" vs "rare" the t-statistic is

William S. Gosset a.k.a. "Student"
Head Experimental Brewer at Guinness
1876-1937



Which difference is more likely to be “true”?





The t-value (1)

- Conceptually, the t-value is a ratio of variance explained to variance unexplained

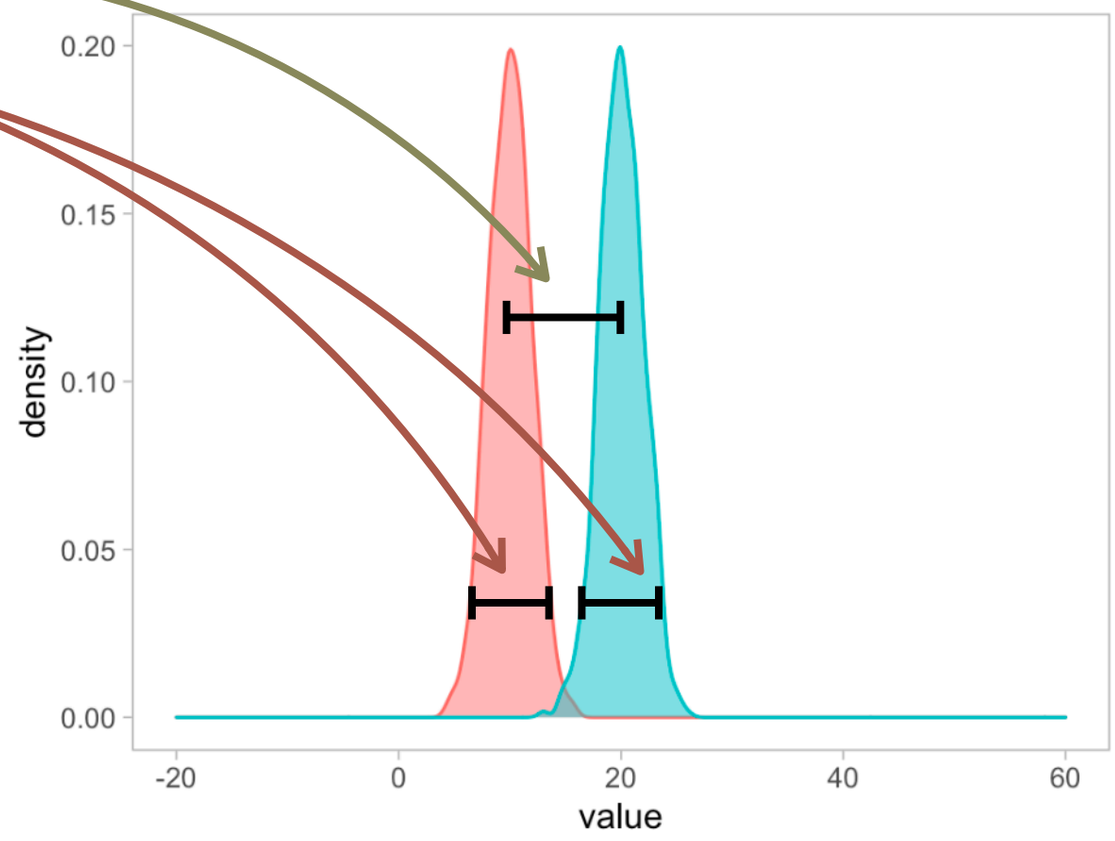
$$t = \frac{variance\ between\ groups}{variance\ within\ groups}$$

$$\frac{effect}{error}$$

$$\frac{mean}{standard\ deviation}$$

$$\frac{variance\ explained}{variance\ unexplained}$$

$$\frac{systematic\ variance}{unsystematic\ variance}$$



The t-value (2)

- More formally:

$$t = \frac{\textit{observed difference between sample means} - \textit{expected difference between population means (when } H_0 \textit{ is true)}}{\textit{SE of the difference between the two sample means}}$$

The t-value (2)

- More formally:

= 0 in H_0

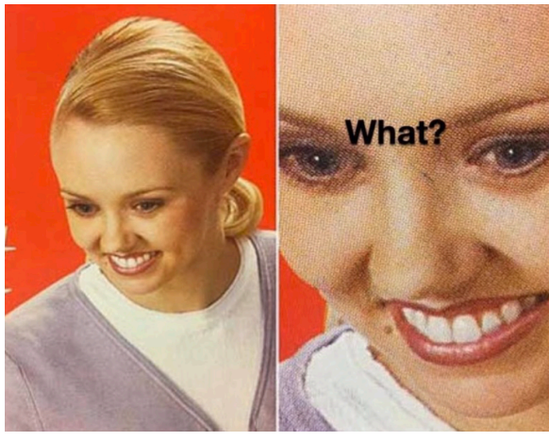
$$t = \frac{\textit{observed difference between sample means} - \textit{expected difference between population means (when } H_0 \textit{ is true)}}{\textit{SE of the difference between the two sample means}}$$

The t-value (2)

- More formally:

= 0 in H_0

$$t = \frac{\textit{observed difference between sample means} - \textit{expected difference between population means (when } H_0 \textit{ is true)}}{\textit{SE of the difference between the two sample means}}$$

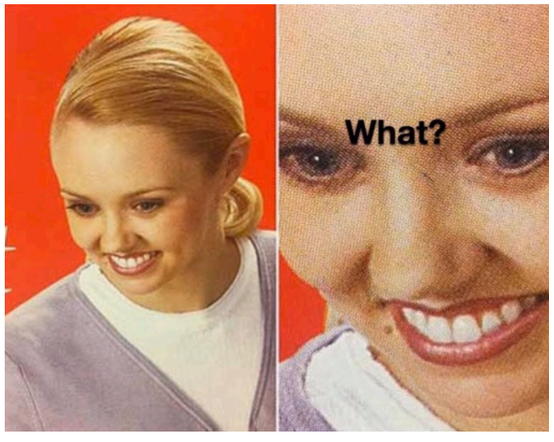


The t-value (2)

- More formally:

$$t = \frac{\textit{observed difference between sample means} - \textit{expected difference between population means (when } H_0 \textit{ is true)}}{\textit{SE of the difference between the two sample means}}$$

= 0 in H_0



estimate of the amount of variance/error in the population distribution

The t-value (3)

= 0 if H_0 is TRUE

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\textit{SE of the difference between the two sample means}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\textit{SE of the difference between the two sample means}}$$

SE of the difference between the two means?

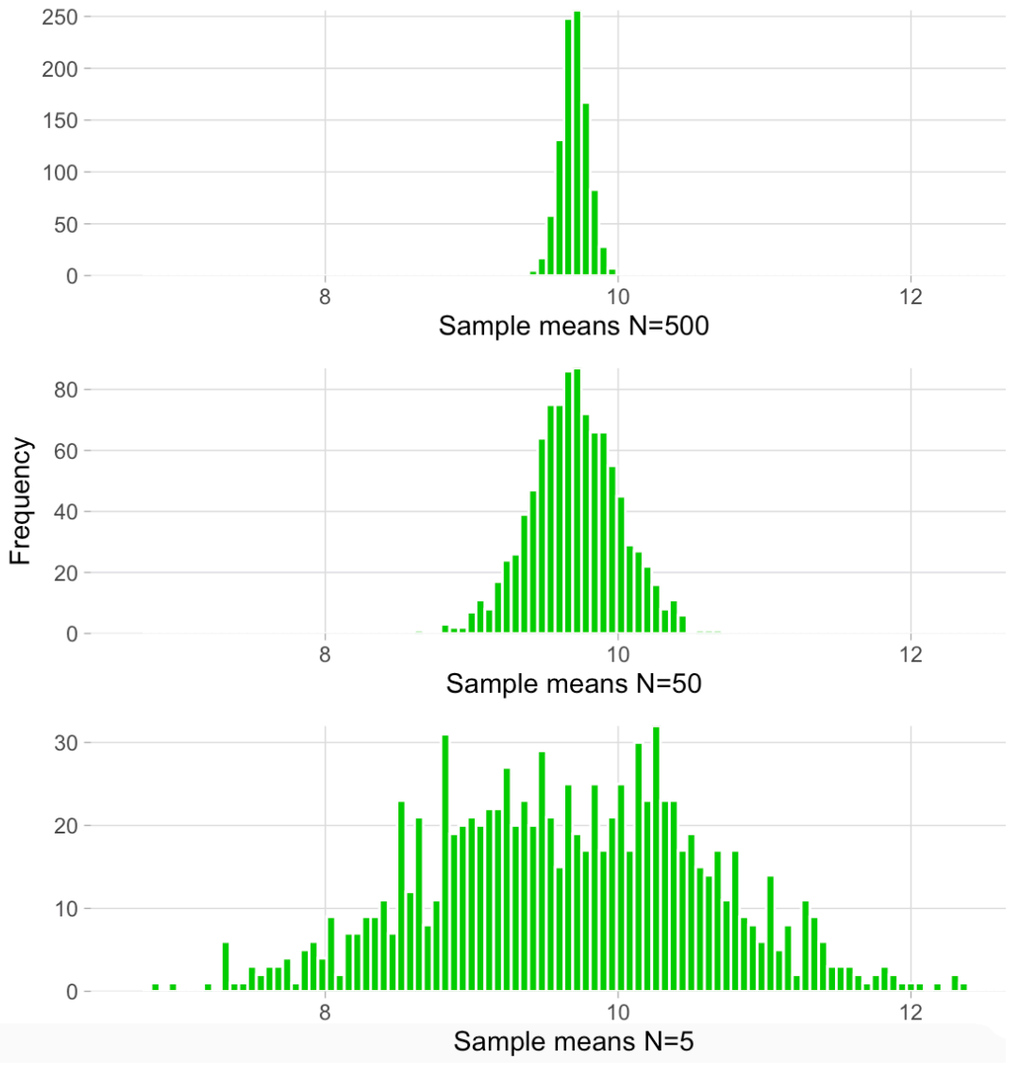
- $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \rightarrow \sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$ Standard Error of the Mean

- **Variance sum law:**

- $(\frac{s}{\sqrt{N}})^2 = \frac{s^2}{N}$ Can easily be transformed to the variance of the sampling distribution of the mean

- $\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$ To find the variance of the sampling distribution of the differences of the means, we sum the means

- $\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$ Its square root is the SE of the sampling distribution of the differences of the means



The t-value (4)

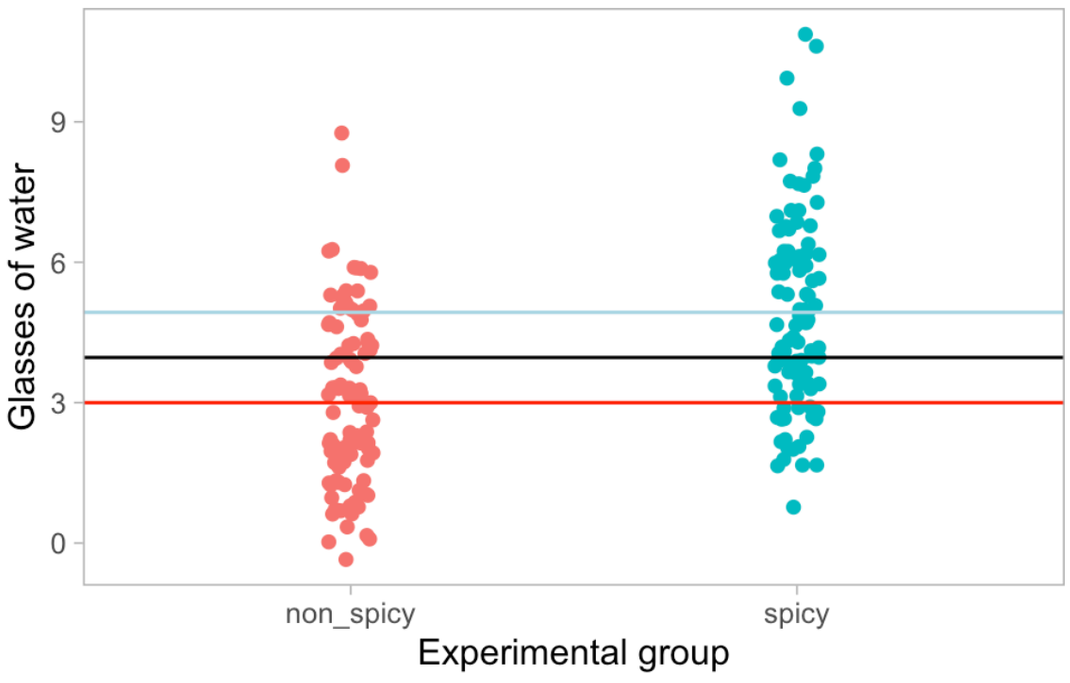
- $$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

The t-value (4)

- $$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$t = \frac{(3 - 4.93)}{\sqrt{\frac{1.83}{100} + \frac{2.06}{100}}} =$$

$$= -9.78$$



What does the t-value tell us?

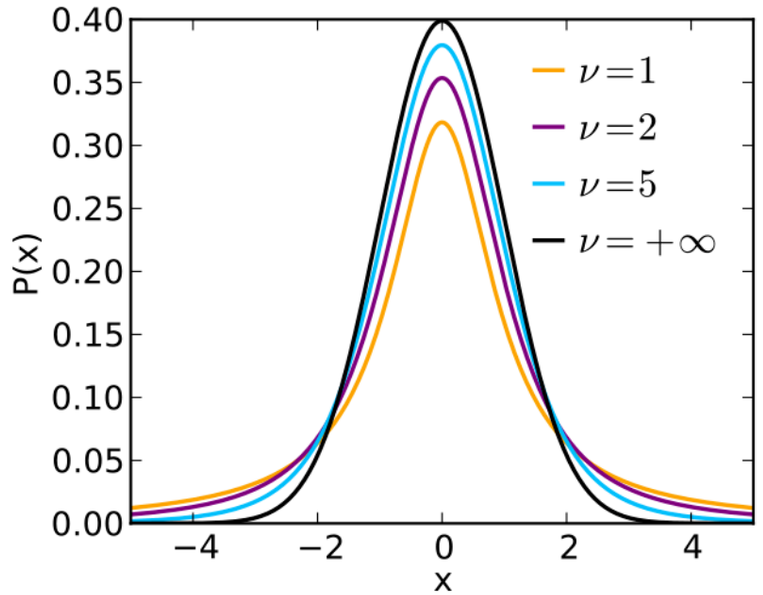
- The *t*-value informs me about how different the two populations/conditions/groups are

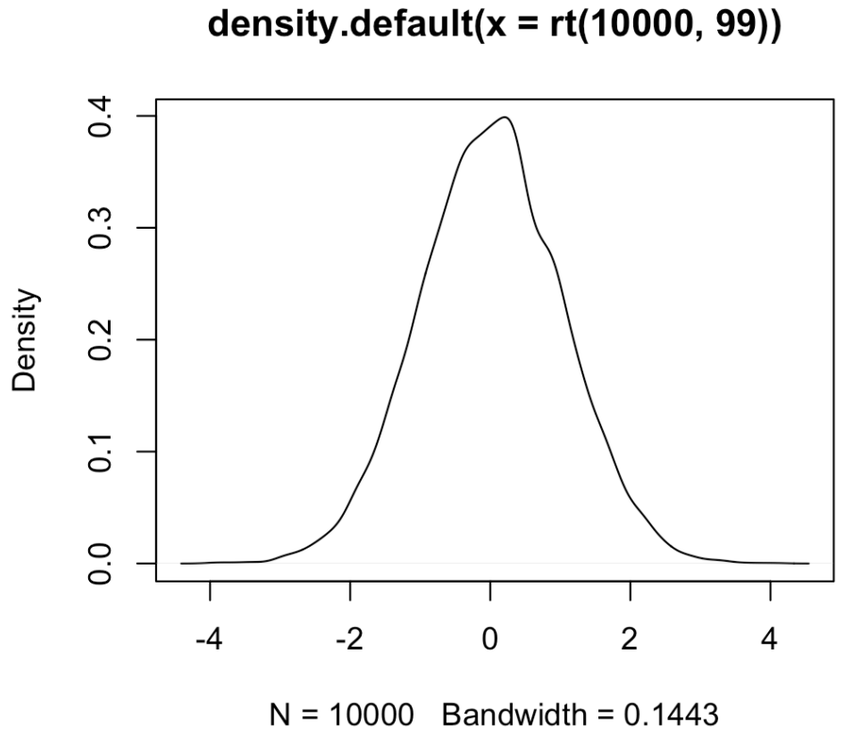
- Each sample size (*df*) has a different *t* distribution

- *t* = 0 = no difference

- A value of t = -9.78 is highly unlikely…

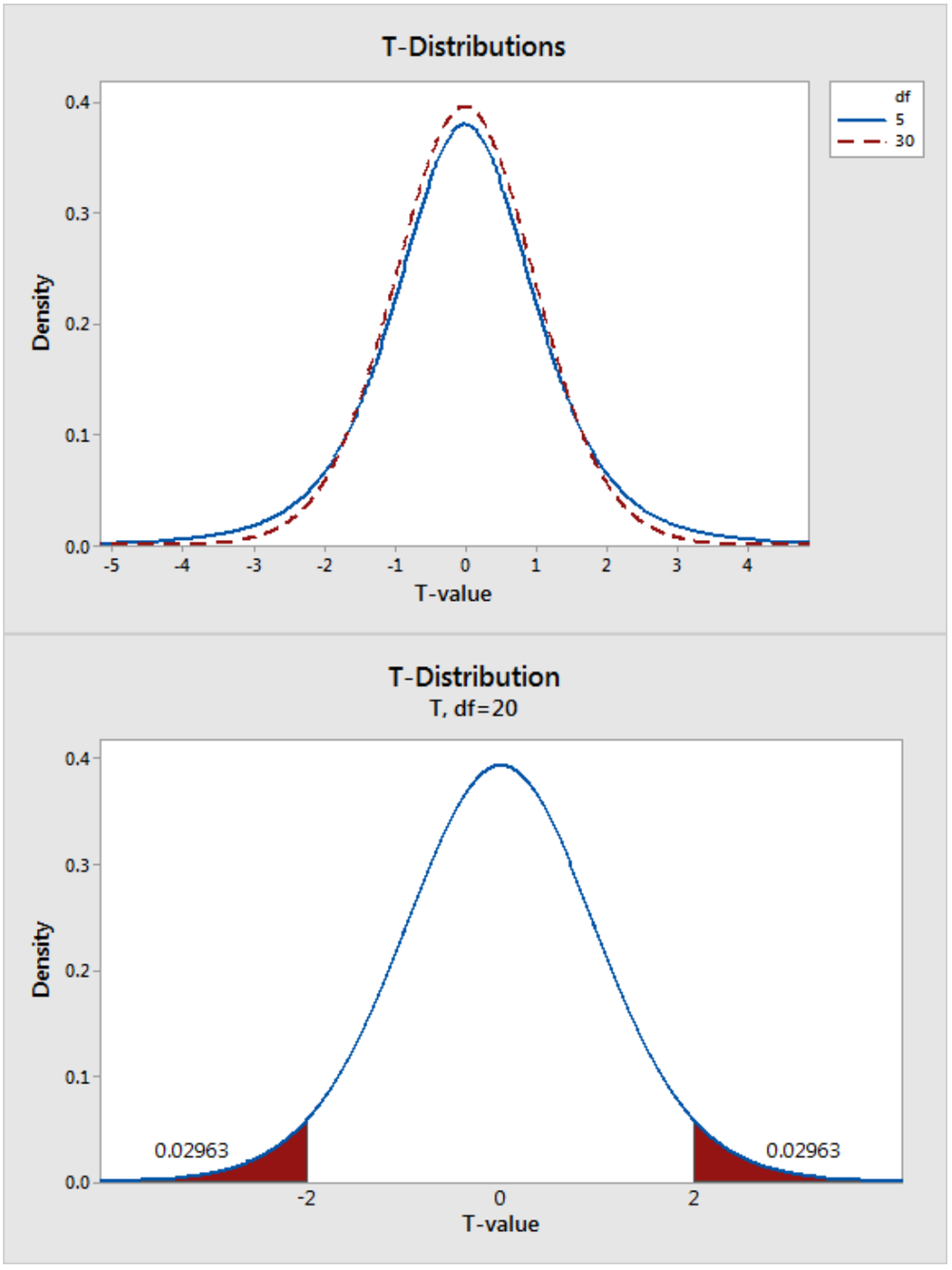
- … but how unlikely?





t-values → p-values

- We can use the *t*-distribution to derive probabilities
- What is the P of getting a t-value *at least as extreme* as -9.78?
- P-values for the t-distribution have been calculated before just as for the z-distribution
- t = -9.78 → p = 4e-11
- Do we accept or reject H_0?



Three types of t-test

- **Independent samples *t*-test**
  - for between-participant designs
  - assumes independent data points and equal variances between groups
- **Paired samples *t*-test**
  - for within-participant designs
  - assumes dependent data points and unequal variances between groups
- **One-sample *t*-test**
  - for comparing the mean of one sample to a specific value (μ)

Independent samples t-test (1)

- H_0: The difference between the two means is 0
- H_1: The difference between the two means is ≠ 0
- $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Independent samples t-test (2)

- In R:

```
> t.test(spicy, non_spicy, var.equal = TRUE)

    Two Sample t-test

data:  spicy and non_spicy
t = 7.0003, df = 198, p-value = 3.872e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.386313 2.473687
sample estimates:
mean of x mean of y
     4.93      3.00
```

Paired samples t-test (1)

- H_0: The difference between the two means is 0
- H_1: The difference between the two means is ≠ 0

$$t = \frac{\overline{D} - \mu_D}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Paired samples t-test (1)

- H_0: The difference between the two means is 0
- H_1: The difference between the two means is ≠ 0

$$t = \frac{\overline{D} - \mu_D}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

the mean of the difference of the responses between the two conditions for each participant

the difference in population means we'd expect if H0 were true

Paired samples t-test (2)

- In R:

```
> t.test(spicy, non_spicy, paired = TRUE)

    Paired t-test

data:  spicy and non_spicy
t = 7.164, df = 99, p-value = 1.411e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.39545 2.46455
sample estimates:
mean of the differences
                   1.93
```

One-sample t-test (1)

- H_0: The sample mean is equal to the population mean
- H_1: The sample mean is different from the population mean

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

One-sample t-test (1)

- H_0: The sample mean is equal to the population mean
- H_1: The sample mean is different from the population mean
- $$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}}$$

the sample mean

the population mean

the SE of the mean

One-sample t-test (2)

- In R:

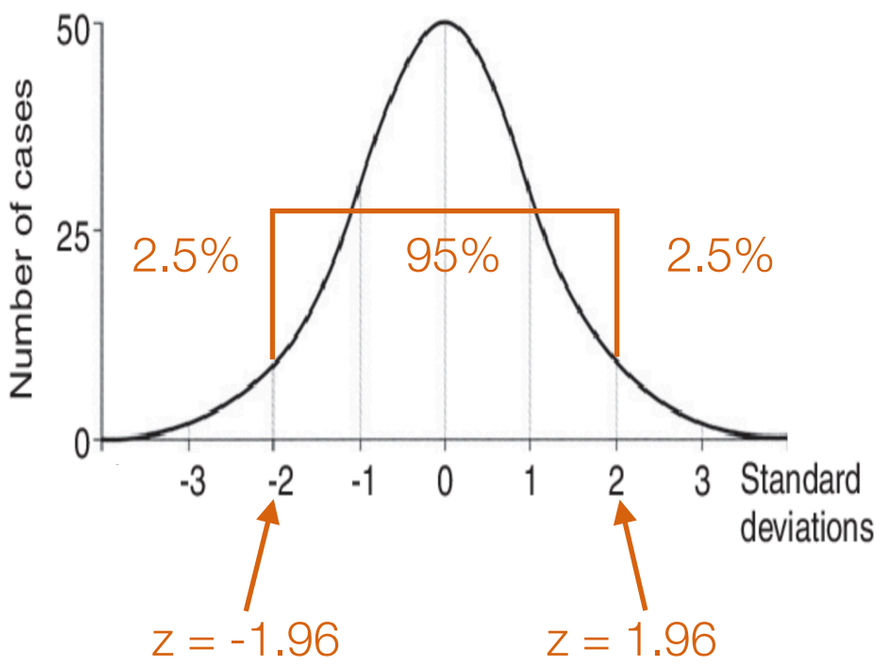
```
> t.test(spicy, mu = 3)

    One Sample t-test

data:  spicy
t = 9.3645, df = 99, p-value = 2.662e-15
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 4.52106 5.33894
sample estimates:
mean of x
     4.93
```

Recap: How to interpret the p-value (1)

- The *p*-value is the probability of obtaining test results at least as extreme as the observed ones when the null hypothesis of no difference between the means is true

- For instance:
  - $p = 0.32$ If the H0 is correct, we expect to see this specific test statistic value 32% of the time
  - $p = 0.02$ If the H0 is correct, we expect to see this specific test statistic value only 2% of the time

- Traditionally we use thresholds of either $p = 0.05$ (5%) or $p = 0.001$ (0.1%)



z	Large	Small	y
.12	.54776	.45224	.3961
.13	.55172	.44828	.3956
.14	.55567	.44433	.3951
.15	.55962	.44038	.3945
.16	.56356	.43644	.3939
.17	.56749	.43251	.3932
.18	.57142	.42858	.3925

Recap: How to interpret the p-value (2)

- The p-value is **not**:
  - the probability that the null hypothesis is true
  - the probability that the alternative hypothesis is false
  - the probability that the observed effects were produced by random chance alone
  - a measure of the size or importance of the observed effect

Recap: How to interpret the p-value (3)

Getting to a Post "p<0.05" Era

What Have We (Not) Learnt from Millions of Scientific Papers with *P* Values?

John P. A. Ioannidis

Pages 20-25 | Received 25 Nov 2017, Published online: 20 Mar 2019

Download citation https://doi.org/10.1080/00031305.2018.1447512 Check for updates

"*P* values linked to null hypothesis significance testing (NHST) **is the most widely (mis)used method of statistical inference**"

Reporting statistical significance on a t-test

- "On average, participants that were fed spicy food drank a significantly larger number of glasses of water (M = 4.93, SD = 2.06), compared to participants that ate non-spicy food (M = 3, SD = 1.83), $t(198) = -9.78$, $p < .001$."

Assumptions of the t-test

- Parametric test:
  - Assumes the dependent variable are normally distributed and that the variances are equal
    - Equality of variance is often violated in between-participant designs (different N)
      - → Welch's t-test: `> t.test(spicy, non_spicy, var.equal = TRUE)`
- If assumptions of normality is violated:
  - → Dependent test: WRS2::yuend(x, y, tr = 0.2)
  - → Independent test: WRS2::yuen(x ~ y, data = data)

Take-home message

- In cognitive science, we often rely on full-fledged experiments
- These imply experimental manipulations
- Often we want to compare results across conditions
- If we have two conditions/groups, the *t*-test is what we want to use
- The *t*-test (and its associated *p*-value) can tell us how much the difference between two groups/manipulations is statistically significant (ie. observable by chance less than X% of the time if H0 = TRUE) or not