# Statistical modelling

## 01. Statistical Inference
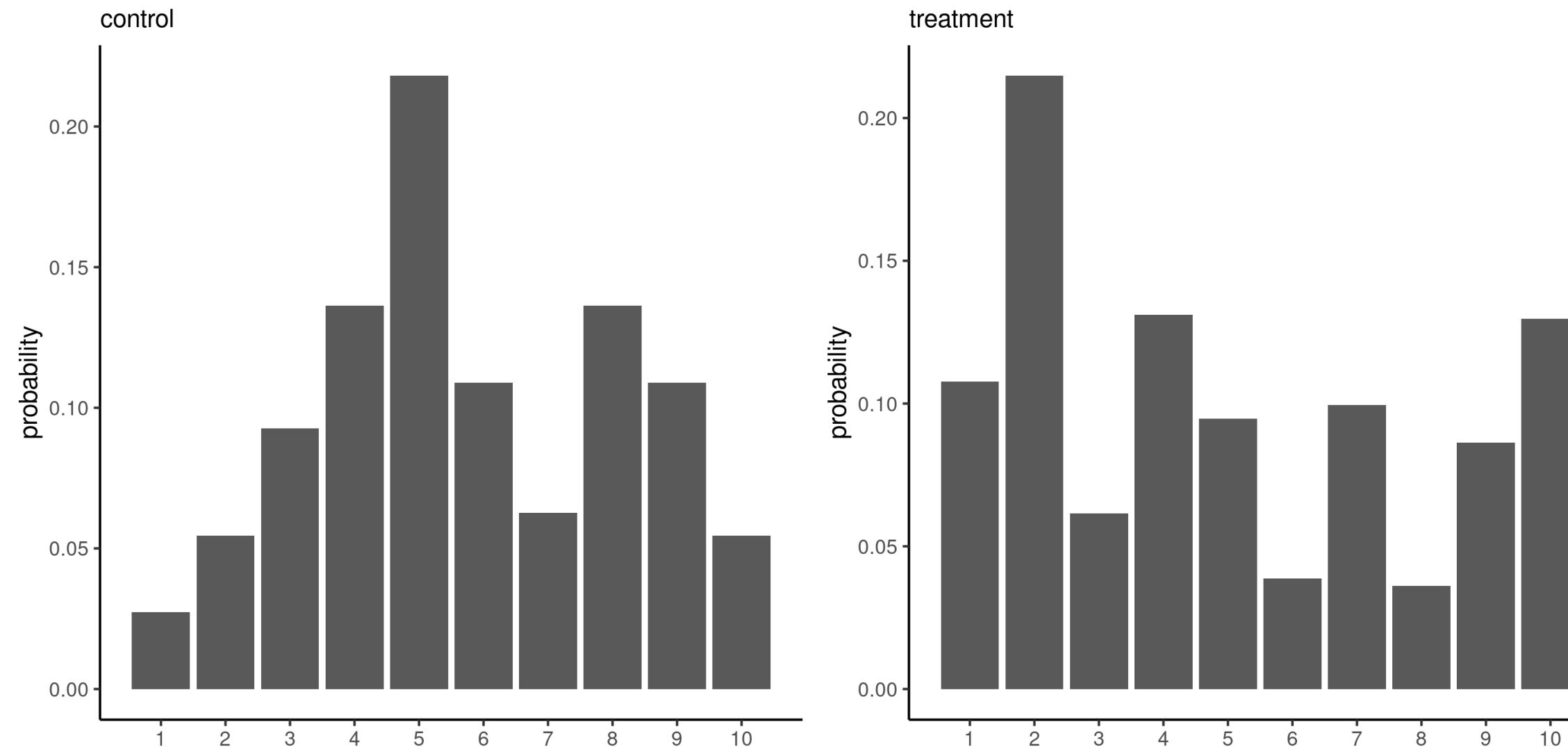
Léo Belzile, HEC Montréal

2024
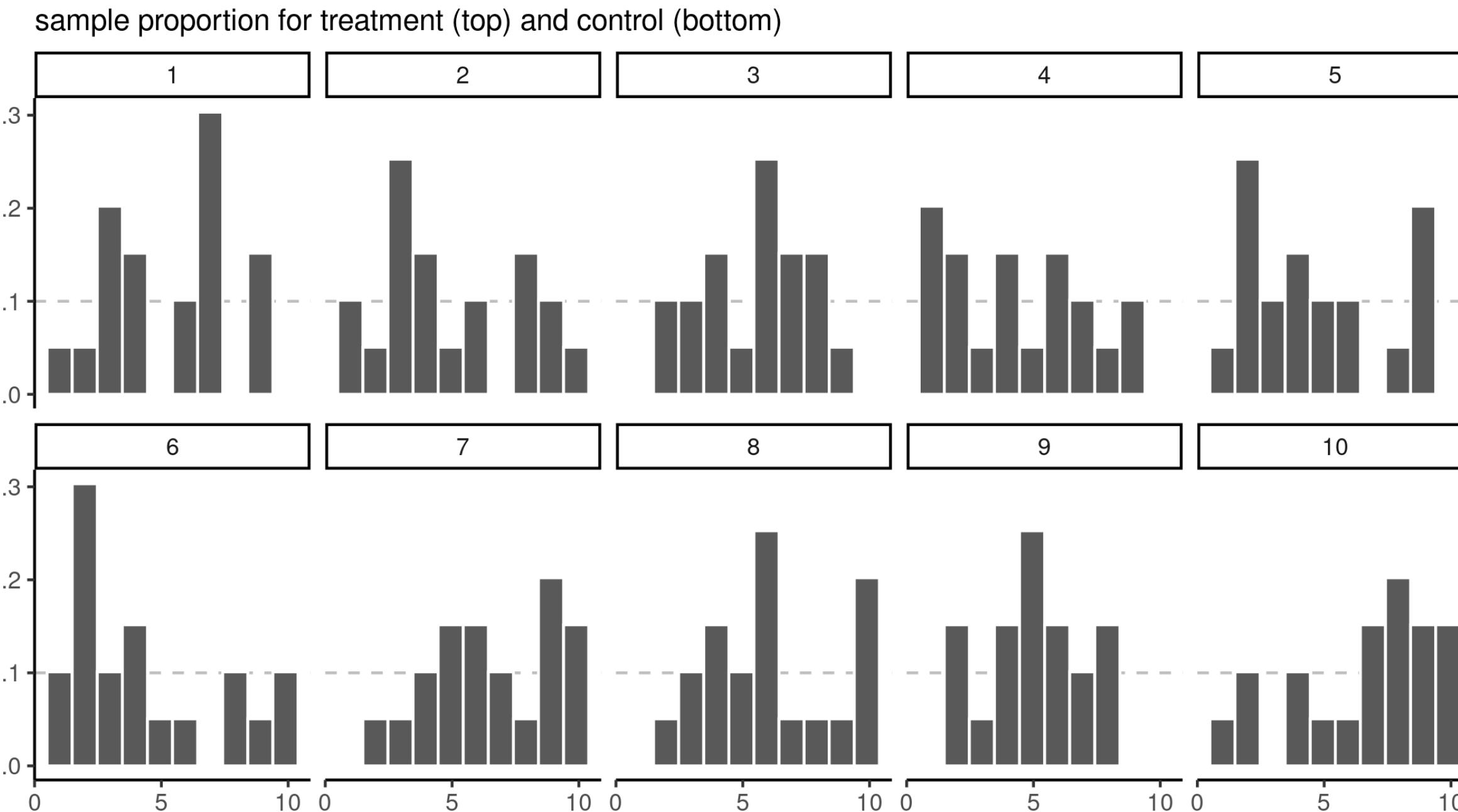
Statistical inference

Studying a population

The population distribution (describing possible outcomes and their frequencies) encodes everything we could be interested in.

Sampling variability

sample proportion for treatment (top) and control (bottom)

Histograms for 10 random samples of size 20 from a discrete uniform distribution.

Decision making under uncertainty

Statistics is concerned with decision making under uncertainty.

- Data collection is costly.
    - we only have limited information about the population.
    - focus on moments (mean μ, standard deviation σ, etc.)
- Samples are typically too small to reliably estimate the whole distribution.
    - we rely instead on a statistical model to gain insights about the data.

HEC MONTRÉAL

Why use models?

A famous quote attributed to George Box claims that

> All models are wrong, but some are useful.

Peter McCullagh and John Nelder wrote in the preamble of their book (emphasis mine)

> Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. The first is that all models are wrong; **some, though, are better** than others and we can **search for the better ones**. At the same time we must recognize that eternal truth is not within our grasp.

More insights on modelling

This quote by David R. Cox adds to the point:

> …it does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological or sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that **capture important stable aspects of such systems** is, however, a vital part of general scientific analysis and statistical models, especially substantive ones, do not seem essentially different from other kinds of model.

HEC MONTRÉAL

What is in a model?

A stochastic model typically combines

- a distribution for the data

- a formula linking the parameters or the mean of a response variable Y conditional on explanatory variables $\mathbf{X}$

Models are "golem" for obtaining answers to our questions.

How to model

0. Determine a research question and collect some data/harvest data.

1. Conduct an **exploratory data analysis**.

- what are the salient features of the data?

2. Build a preliminary model.

3. Check the goodness-of-fit or adjustment.

4. Rinse and repeat if necessary.

How do we determine the model?

We need to know

1. how were the data collected

- what is the sampling mechanism?

- are variables observed or experimentally manipulated?

2. what is the nature of the response?

- count data, proportions, etc.

Nature of data
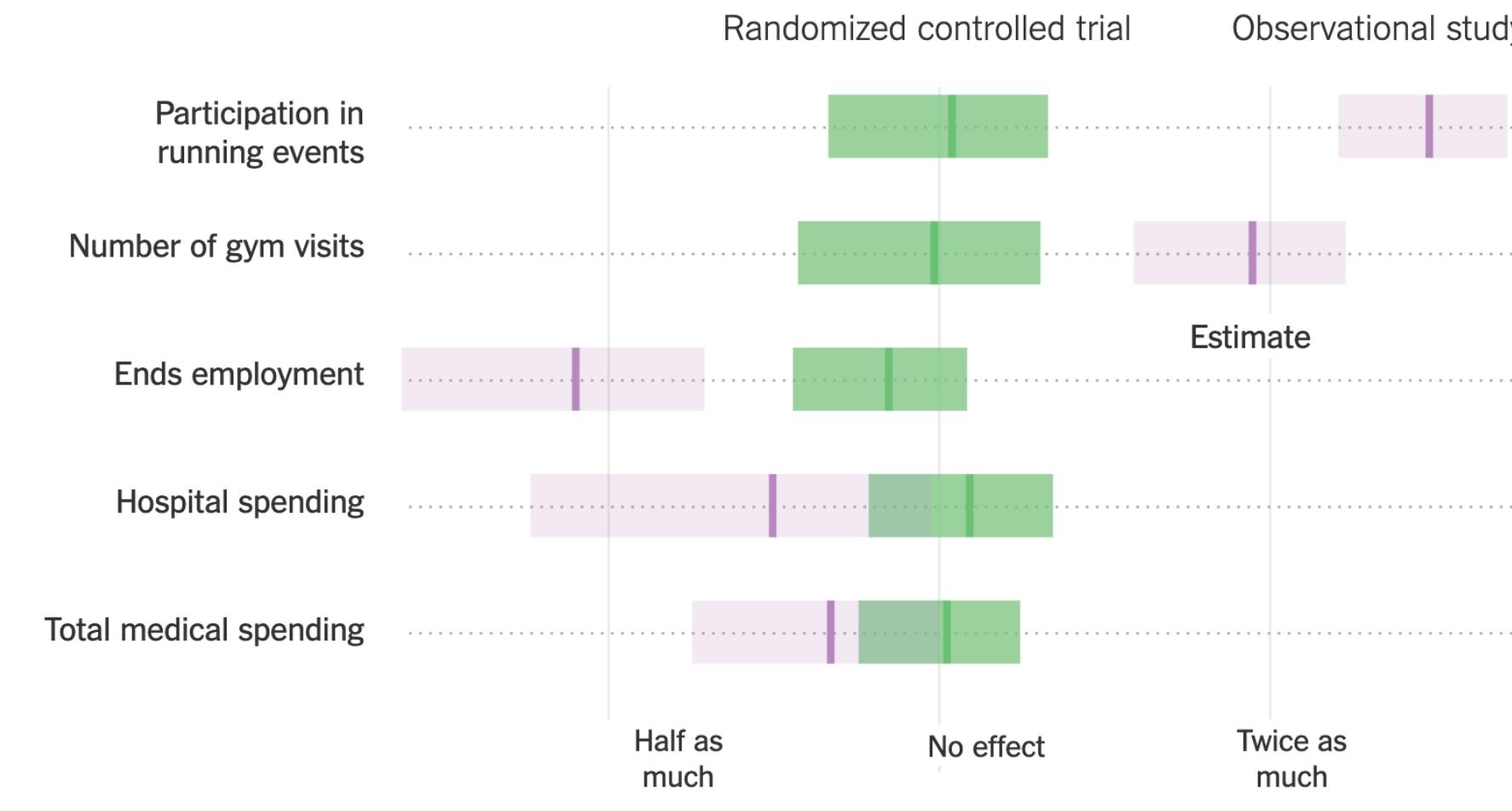
- Are data from a **random sample of the population** or not?
  - if so, can generalize conclusions.
- Is there a "treatment" randomly allocated?
  - if so, data are experimental (as opposed to observational).

Observational vs experimental

Without further adjustment, we cannot draw **causal** statements from observational data.

How the Illinois Wellness Program Affected . . .

Source: What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study

Motivating examples

1. Driving exams in Great Britain

Are driving tests easier if you live in a rural area? Source: The Guardian, August 23rd, 2019

An easy ride? Scottish village fuels debate on driving test pass rates
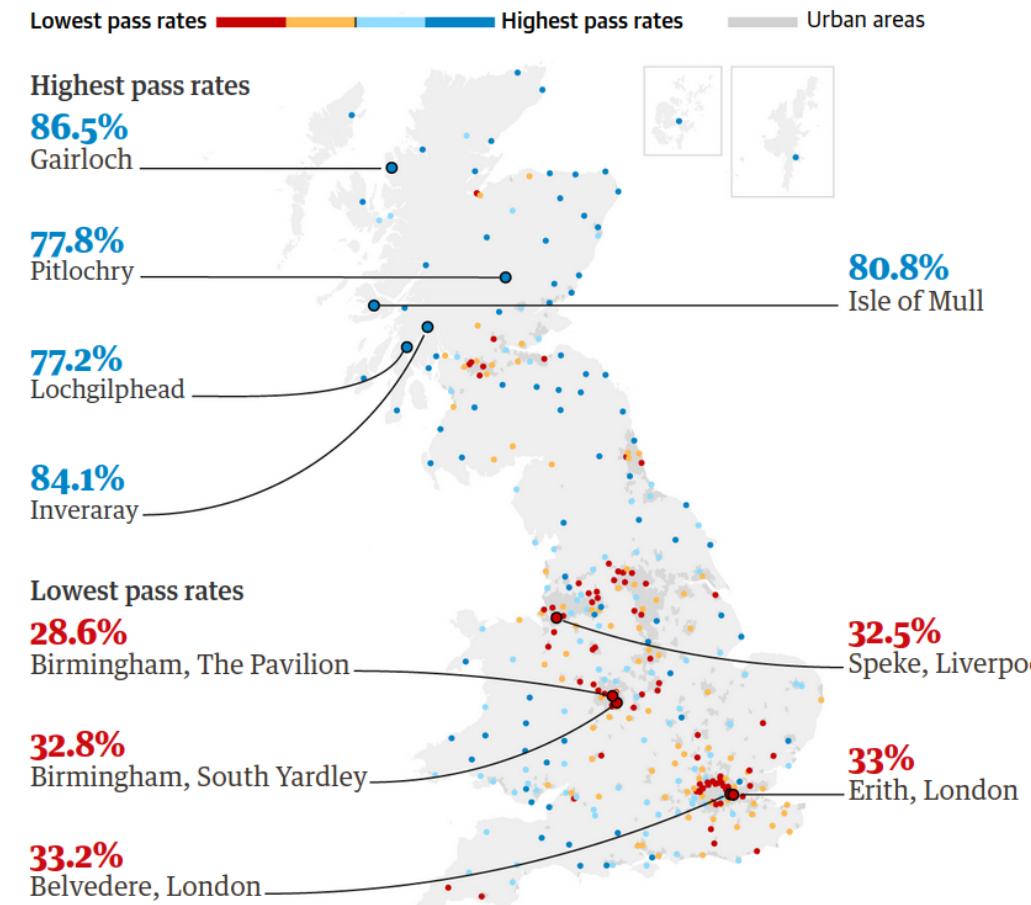
New figures reveal that it is far easier to pass driving test in Highlands than in inner city Birmingham

Simon Murphy

Fri 23 Aug 2019 13.00 BST

Share

GAIRLOCH
Please drive carefully

The 86.5% pass rate in Gairloch is stratospheric compared with urban areas, while the national average is 45.8%. Photograph: Peter Jolly/for the Guardian

Rural test centres tend to have higher pass rates than ones in cities

Lowest pass rates — Highest pass rates — Urban areas

Highest pass rates
86.5%
Gairloch

77.8%
Pitlochry

80.8%
Isle of Mull

77.2%
Lochgilphead

84.1%
Inveraray

Lowest pass rates
28.6%
Birmingham, The Pavilion

32.5%
Speke, Liverpool

32.8%
Birmingham, South Yardley

33%
Erith, London

33.2%
Belvedere, London

Guardian graphic. Source: gov.uk

Model: binomial logistic model. Data `gbdriving`, **R** package `hecstatmod`.

HEC MONTRÉAL

2. Road safety and distraction due to smartwatches

Brodeur et al. (2021)

> A within-subject experiment was conducted in a driving simulator where 31 participants received and answered text messages under four conditions: they received notifications (1) on a mobile phone, (2) on a smartwatch, and (3) on a speaker, and then responded orally to these messages. They also (4) received messages in a "texting" condition where they had to reply through text to the notifications.

Models: Within-subject ANOVA (repeated measures) with pairwise paired t-tests or nonparametric tests (Friedman + Wilcoxon signed-rank test). Dataset `BRLS21_T3`, package `hecedsm`.

HEC MONTRÉAL

3. Perceived environmental friendliness of packages

TABLE 1

EXAMPLES OF PLASTIC-PLUS-PAPER- AND PLASTIC-PACKAGED PRODUCTS

Sokolova, Krishna, and Döring (2023)

Eight studies (N = 4103) document the perceived environmental friendliness (PEF) bias whereby consumers judge plastic packaging with additional paper to be more environmentally friendly than identical plastic packaging without the paper.

Model: linear regression/ANOVA with custom contrasts. Dataset SKD23_S2A, package hecedsm

HEC MONTRÉAL

4. A/B tests and news headlines
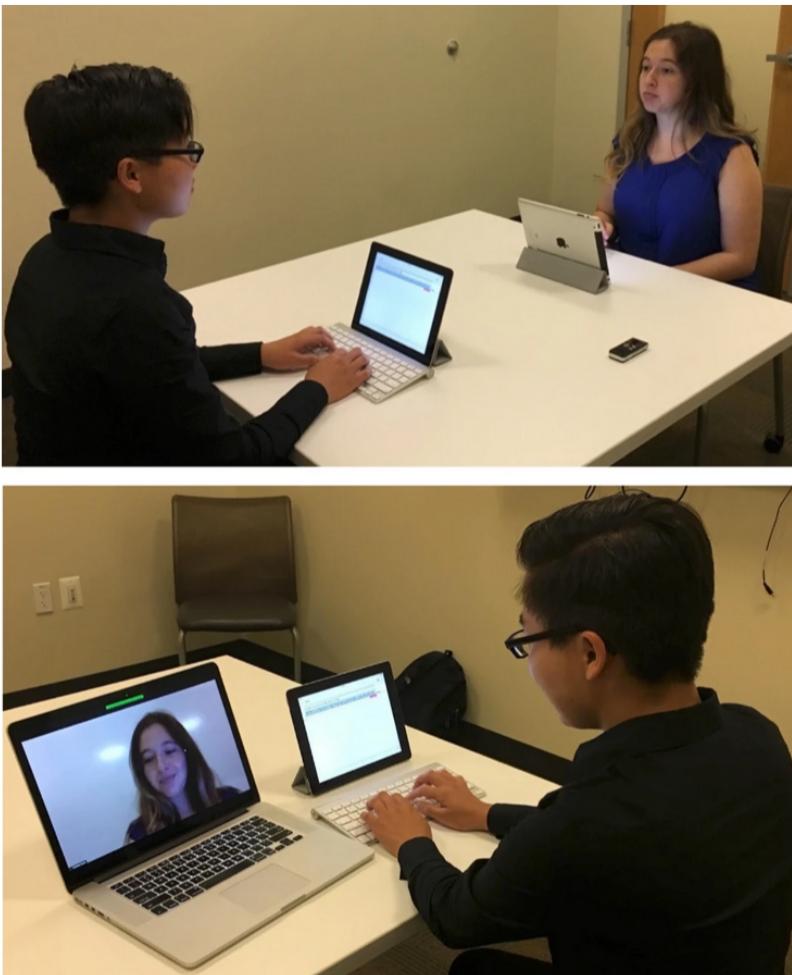
Upworthy.com, a US media publisher, revolutionized headlines online advertisement by running systematic A/B tests to compare the different wording of headlines, placement of text and image to figure out what catches attention the most.

The Upworthy Research Archive (Matias et al. 2021) contains results for 22743 experiments, with a click through rate of 1.58% on average and a standard deviation of 1.23%.

Model: Poisson regression with offset. Data `upworthy_sesame`, package `hecbayes`.

5. Impact of videoconferencing on idea generation

In the laboratory experiment, we randomly assigned half of the pairs to work together in person and the other half to work together in separate, identical rooms using videoconferencing. The pairs in the virtual condition interacted with a real-time video of their partner's face displayed on a 15-inch retina-display screen with no self-view. The image was taken during the first batch of data collection in the laboratory. Consent was obtained to use these images for publication.

Brucks and Levav (2022)

> In a laboratory study and a field experiment across five countries (in Europe, the Middle East and South Asia), we show that videoconferencing inhibits the production of creative ideas […]

> we demonstrate that videoconferencing hampers idea generation because it focuses communicators on a screen, which prompts a narrower cognitive focus. Our results suggest that virtual interaction comes with a cognitive cost for creative idea generation.

- Model: Linear regression with compound symmetry covariance/MANOVA. Dataset `BL22_E`, package `hecedsm`
- Model: Binomial regression, dataset `BL22_L` from package `hecedsm`.

HEC MONTRĒAL
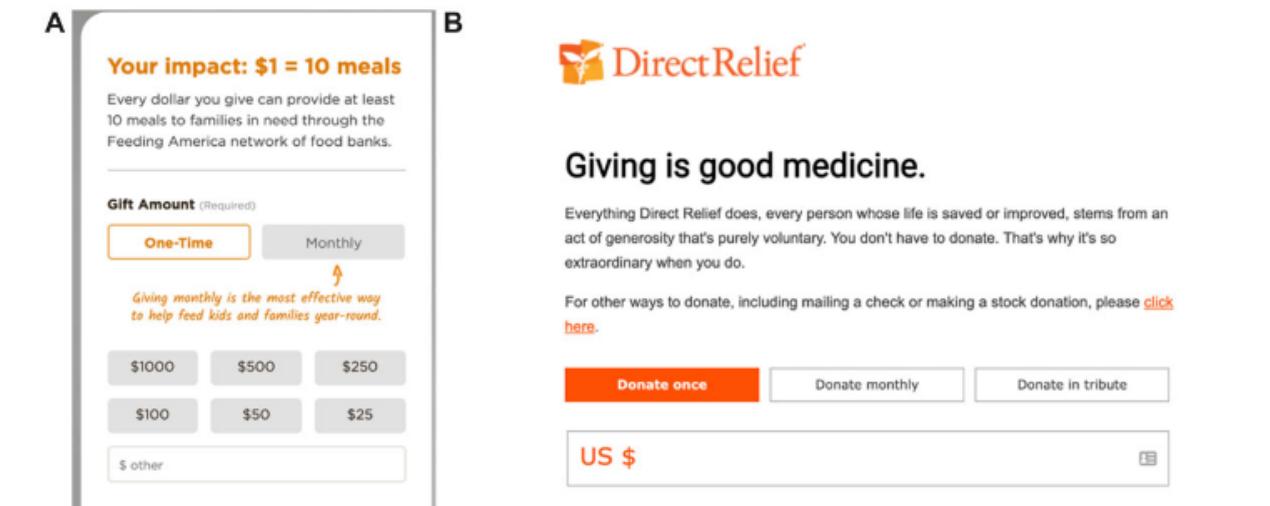
6. Suggesting amounts for donations

Moon and VanEpps (2023)

> Across seven studies, we provide evidence that quantity requests, wherein people consider multiple choice options of how much to donate (e.g., $5, $10, or $15), increase contributions compared to open-ended requests.

> Our findings offer new conceptual insights into how quantity requests increase contributions as well as practical implications for charitable organizations to optimize contributions by leveraging the use of quantity requests.

FIGURE 1

DONATION REQUESTS IN THE REAL WORLD

NOTE.—A quantity request used by Feeding America with donation choice options ranging from $25 to $1,000 (A) and an open-ended request used by Direct Relief (B).

Model: Tobit type **II** regression and Poisson regression (independence test), data `MV23_S1` from package `hecedsm`.

HEC MONTRÉAL

7. Integrated decisions in online shopping

Duke and Amir (2023)

Customers must often decide on the quantity to purchase in addition to whether to purchase. The current research introduces and compares the quantity-sequential selling format, in which shoppers resolve the purchase and quantity decisions separately, with the quantity-integrated selling format, where shoppers simultaneously consider whether and how many to buy. Although retailers often use the sequential format, we demonstrate that the integrated format can increase purchase rates.

A field experiment conducted with a large technology firm found that quantity integration yielded considerably higher sales, amounting to an increase of more than $1 million in annual revenue.
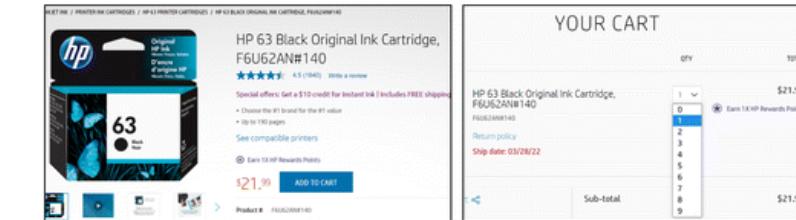

Figure 1. (Color online) Example Quantity-Sequential Selling Format, HP Printer Ink
Notes. (Left) Screenshot of a product information page shown to customers on HP's website. (Right) Screenshot of the shopping cart page after adding this item to the cart. Web locations: https://store.hp.com/us/en/pdp/hp-63-black-original-ink-cartridge and https://store.hp.com/us/en/AjaxOrderItemDisplayView. Both screenshots were taken on March 25, 2022.

Figure 2. (Color online) Hypothetical Quantity-Integrated Selling Format, HP Printer Ink
Note. To accommodate additional quantity options, the website could also consider a fourth button such as "Add 4 or more to cart"; alternatively, customers could simply click one of the presented "add" buttons multiple times.

Model: logistic regression, dataset DA23_E1.

HEC MONTRÉAL

8. Price of oil in the Gaspe Peninsula

Mayors requested an inquiry by the **Régie de l'énergie**, a regulating agency in charge of energy prices. The report found that prices were indeed more expensive, but pointed out that there were more retailers per capita, and lower volume so their margins were higher.

Source: Radio-Canada 1, 2

Model: Linear regression with autoregressive errors, pairwise comparisons. Dataset `renergy`, package `hecstatmod`.

Les Gaspésiens paient l'essence plus cher que les automobilistes des autres régions

C'est en Gaspésie que l'essence coûte le plus cher au Québec.
PHOTO : RADIO-CANADA / JEAN-FRANÇOIS DESCHÊNES

Joane Bérubé

Publié le 8 janvier 2020 à 15 h 45 HAE
Mis à jour le 8 janvier 2020 à 17 h 25 HAE

HEC MONTRÉAL

Hypothesis testing

Sampling variability

We cannot compare summaries without accounting for their uncertainty inherent to our estimation which is due to random sample.
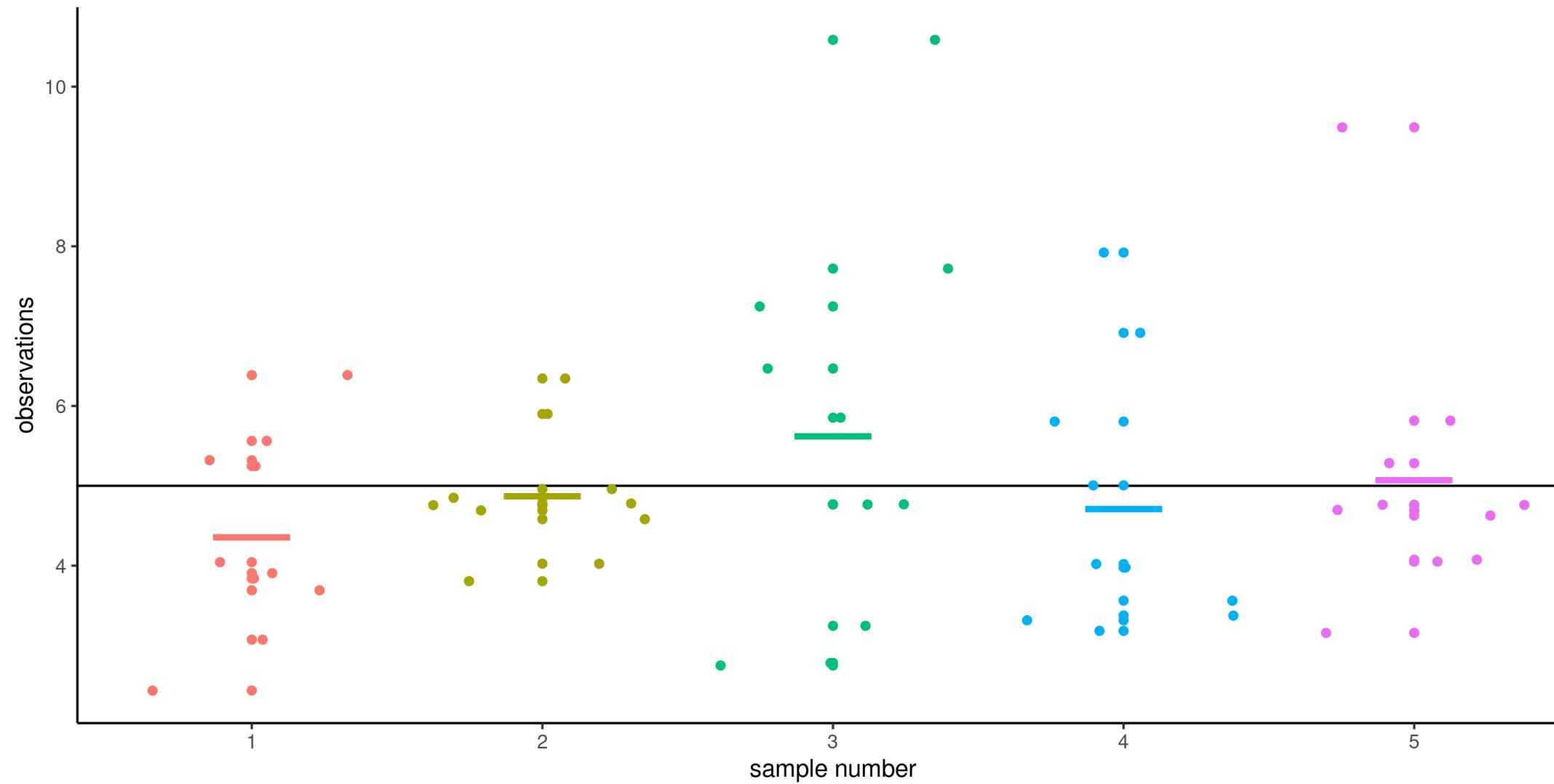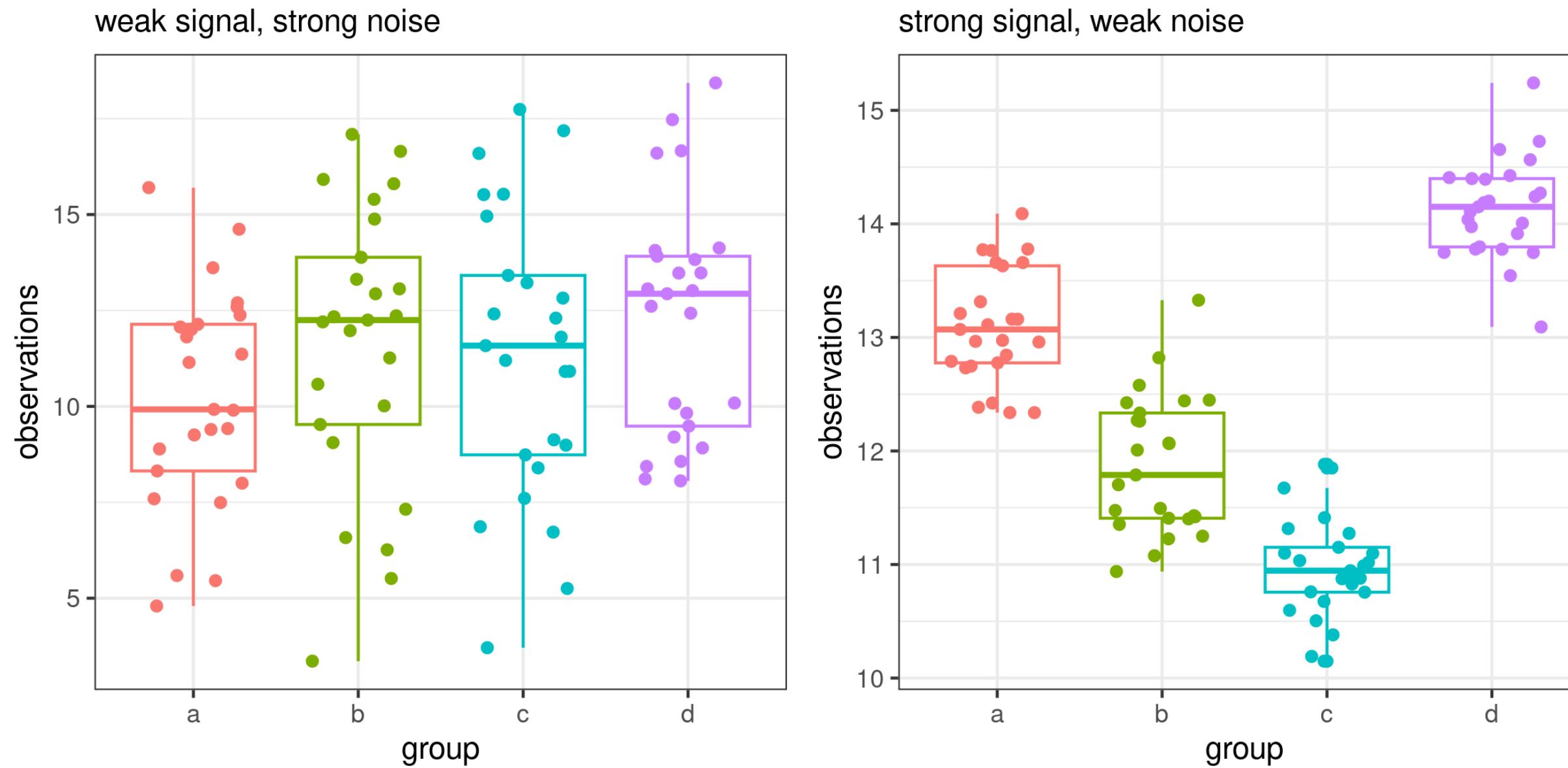
Figure 1: Five samples of size $n = 10$ drawn from a common population with mean μ (horizontal line). The colored segments show the sample means of each sample.
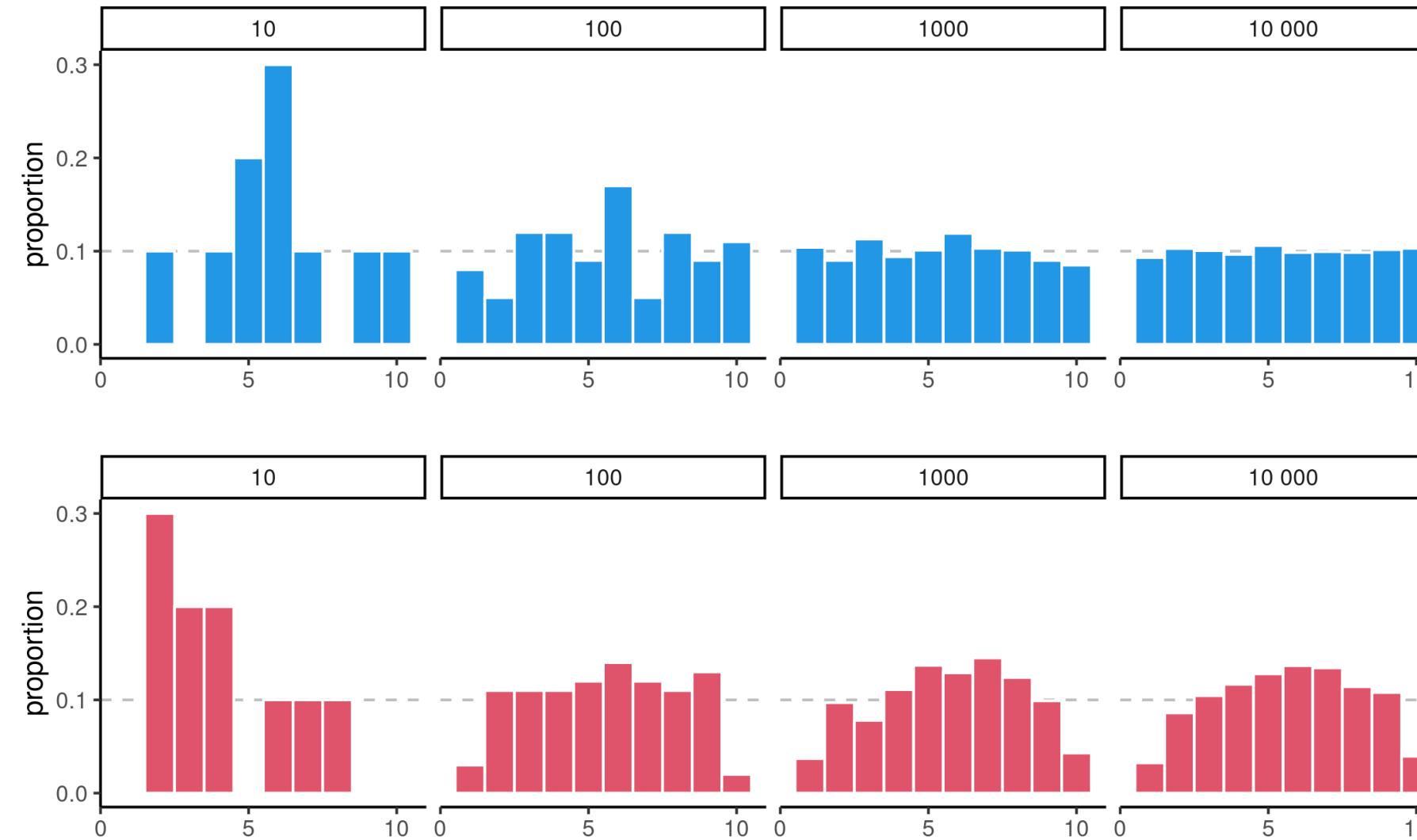
The signal and the noise

The stronger the signal-to-noise ratio, the larger our ability to detect differences when they truly exist.

Accumulation of information

As we gather more observations (sample size increases), we can better discriminate between scenarios.

Histograms of data from uniform (top) and non-uniform (bottom) distributions with increasing sample sizes.

Recipe for hypothesis test

An hypothesis test is a binary decision rule (reject/fail to reject)

Below are the different steps to undertake:

1. Define the variables of interest.

2. Formulate the alternative and the null hypotheses, $\mathscr{H}_a$ and $\mathscr{H}_0$.

3. Choose the test statistic and compute the latter on the sample.

4. Compare the numerical value with the null distribution.

5. Obtain the p-value or a confidence interval.

6. Conclude in the setting of the problem.

Running example: texting while walking

Tech3Lab, HEC Montreal's User Experience (UX) lab, studied the impact of texting on distraction.

Study details

- 35 participants took part in the study.

- Each person had to walk on a treadmill in front of a screen where obstacles were projected.

- In one of the sessions, the subjects walked while talking on a cell phone, whereas in another session, they walked while texting.

- The order of these sessions was determined *at random*.

- Different obstacles were randomly projected during the session.

- We are only interested in one kind of scenario: a cyclist riding towards the participant.

Characteristics

- Population: adults

- Sample: 35 individuals (convenience sample, recruited)

- Variables:

    - Time to perceive an obstacle: quantitative

    - Distraction type (cellphone call or texting): nominal variable

- Response variable: time (in seconds) that it takes for a person to notice the obstacle when walking while texting or talking on a cell phone (measured through an encephalogram)

Hypothesis testing as a trial analogy

Figure 2: Screenshot of the courtoom drama *Twelve Angry Men* (1957)

Trial analogy

The presumption of innocence applies (look at everything as if the null hypothesis is true)

- Evidences are judged under this optic: are they remotely plausible if the person was innocent?

- The burden of the proof lies with the prosecution to avoid as much as possible judicial errors.

- The null hypothesis $\mathscr{H}_0$ is *not guilty*, whereas the alternative $\mathscr{H}_a$ is *guilty*.

- If there is a reasonable doubt, the verdict of the trial will be not guilty.

1. Define the variables of interest

- Let μ_c be the average reaction time (in seconds) during a call (c)

- and let μ_t be the average reaction time (in seconds) while texting (t)

2. Formulate the null and alternative hypothesis

- Hypothesis of interest (alternative): does texting increases distraction?

  - $\mathscr{H}_a : \mu_{\mathsf{t}} > \mu_{\mathsf{c}}$ (one-sided)

- Null hypothesis (Devil's advocate)

  - $\mathscr{H}_0 : \mu_{\mathsf{t}} \leq \mu_{\mathsf{c}}$

Express the hypothesis in terms of the difference of means

$$\mathscr{H}_a : \mu_{\mathsf{t}} - \mu_{\mathsf{c}} > 0.$$

We only ever assess the null hypothesis at a single value.

How to assess evidence?

- A **statistic** is a function of the data that returns a numerical summary.

- There are general principles for constructing statistics: for example, **Wald statistics** are of the form

$$W = \frac{\text{estimated qty} - \text{postulated qty}}{\text{std. error (estimated qty)}} = \frac{T - T_0}{\text{se}(T)}$$

- Statistic are typically unitless quantities.

- This standardization is useful for benchmarking purposes.

- The **standard error** measures the uncertainty of the statistic.

Do not confuse standard error (variability of statistic) and standard deviation (variability of observation from the population)

HEC MONTRĒAL

3.Choose the test statistic

We compare the difference of the mean reaction time.

- one-sample t-test for $\mathbf{t} - \mathbf{c}$ (paired t-test)

$$T_D = \frac{\overline{D} - \mu_0}{\mathsf{se}(\overline{D})}$$

- $\overline{D}$ is the mean difference in the sample.

- Under $\mathscr{H}_0 : \mu_0 = \mu_{\mathbf{t}} - \mu_{\mathbf{c}} = 0$.

- The standard error of $\overline{D}$ is $\mathsf{se}(\overline{D}) = S_D/\sqrt{n}$, where S_D is the standard deviation of the variables D_i and n the sample size.

R code

```r
1  data(distraction, package = "hecstatmod")
2  ttest <- with(distraction,
3      t.test(x = t,
4             y = c,
5             paired = TRUE,
6             alternative = "greater"))
```

```
##
##  Paired t-test
##
## data:  t and c
## t = 3, df = 34, p-value = 0.003
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.131    Inf
## sample estimates:
## mean difference
##           0.313
```

Null distribution

The null distribution tells us what values we would obtain under a null, and their relative frequency.

- We typically rely on **large sample approximations**, that are justified by the central limit theorem.

- Alternative, use simulation-based inference (check permutation tests).

- The null distribution is a benchmark to establish whether our outcome is extreme.

Null distributions
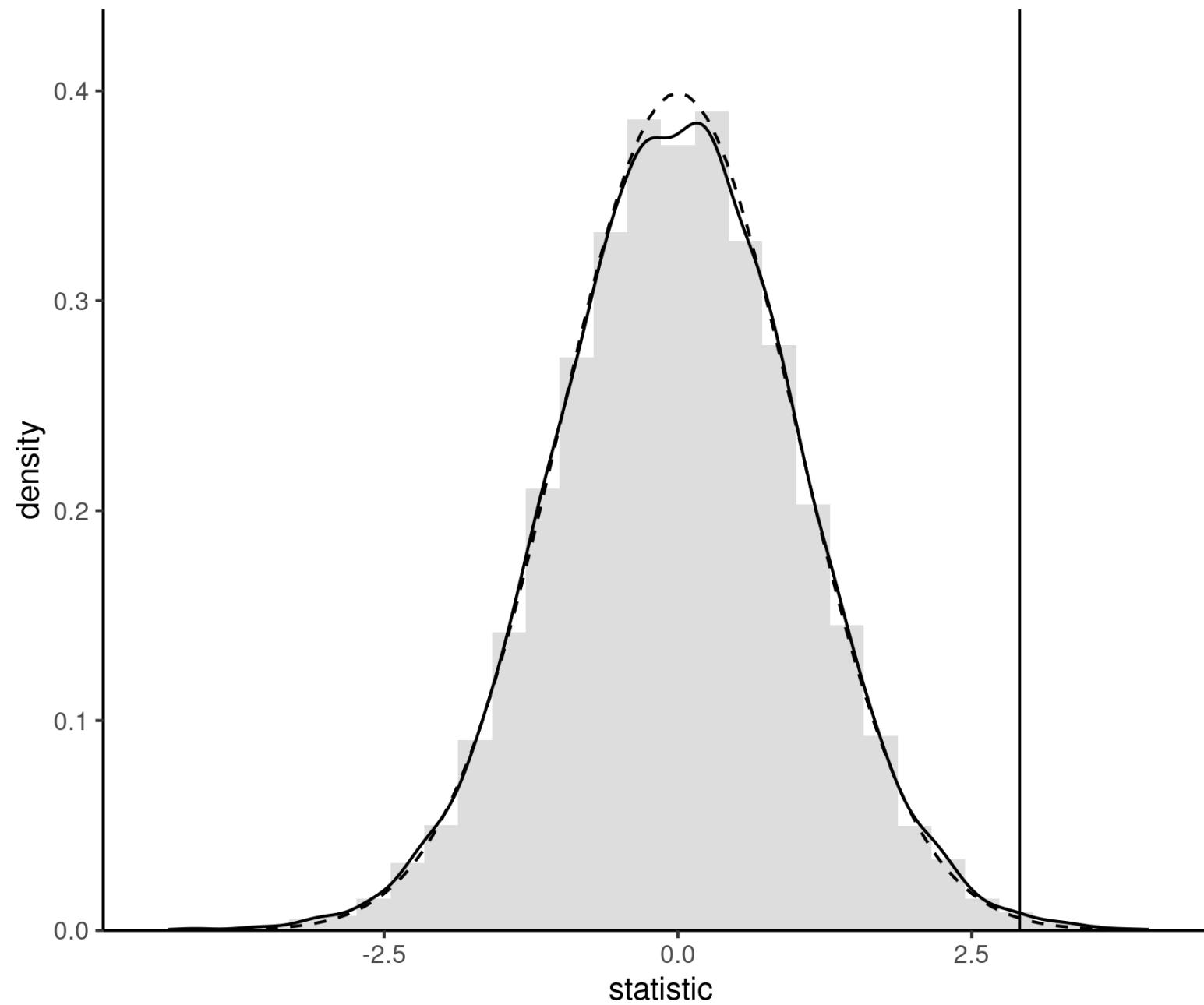
Figure 3: Sign permutation-based approximation to the null distribution (full) and standard normal approximation (dashed curve). The value of the test statistic calculated using the original sample is represented by a vertical line.

4. Compare the statistic with the null distribution

- The test statistic value is $t_D = 2.91$.

- We are only interested in the probability that $T > 2.91$ under $\mathscr{H}_0$ (one-sided test).

- Use as null distribution a standard Student-t distribution with 34 degrees of freedom, $T \sim \text{Student}(34)$.

- The 0.05 quantile of T is $t_{0.05} = -1.691$.

HEC MONTRĒAL

p-values

Apply distribution function of the null distribution to map the test statistic to the $[0, 1]$ interval.

The p-value is the probability that the test statistic is equal or more extreme to the estimate computed from the data, assuming $\mathscr{H}_0$ is true.

> ⚠️ **Caution**
>
> The American Statistical Association (ASA) published a list of principles guiding (mis)interpretation of p-values, including:
>
> - 2. P-values do not measure the probability that the studied hypothesis is true.
> - 5. p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

What can we expect of p-values?

If we repeat the experiment with random samples, we expect p-values to be uniform if $\mathscr{H}_0$ is true **and** the null hypothesis benchmark is properly calibrated.

Under the alternative, smaller p-values occur more often than α.
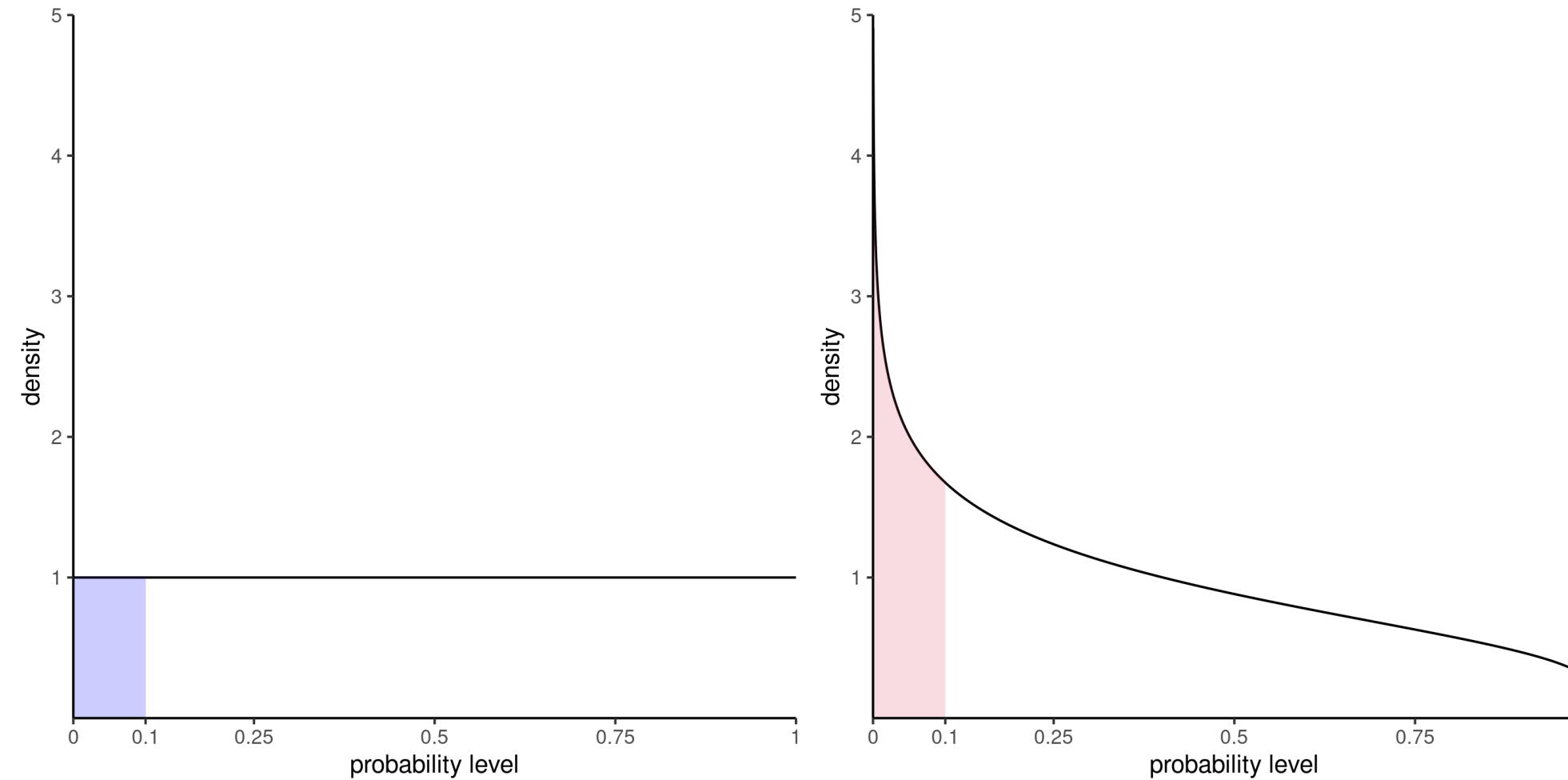
Figure 4: Density of p-values under the null hypothesis (left) and under an alternative with a signal-to-noise ratio of 0.5 (right).

HEC MONTRÉAL

Level of a test

If the null hypothesis $\mathscr{H}_0$ is true, the p-value follows a uniform distribution if our benchmark is properly calibrated.

- There are always small p-values, even if they are less likely.

- Since we want to rule out **against** the null hypothesis, we need to pre-determine a **level** α to decide when to reject $\mathscr{H}_0$.

To make a decision, we compare our p-value P with the level of the test α:

- if $P < \alpha$, we reject $\mathscr{H}_0$;

- if $P \geq \alpha$, we fail to reject $\mathscr{H}_0$.

The value of $\alpha \in (0, 1)$ is the probability of rejecting $\mathscr{H}_0$ when $\mathscr{H}_0$ is in fact true.

Statistical errors

We seek to avoid error of type I: we reject $\mathscr{H}_0$ when $\mathscr{H}_0$ is true.

- trial analogy: avoid condemning an innocent.

Decision \ true model	$\mathscr{H}_0$	$\mathscr{H}_a$
fail to reject $\mathscr{H}_0$	✓	type II error
reject $\mathscr{H}_0$	type I error	✓

Since we fix the level α, we have no control over the type II error.

Power

We want to be able to detect and reject $\mathcal{H}_0$ when it is false.

The **power of a test** is the probability of rejecting $\mathcal{H}_0$ when it is false, i.e.,

$$\mathrm{Pr}_a(\text{reject } \mathcal{H}_0),$$

where Pr_a is the probability under a **given** alternative of falling in the rejection region.

Illustration of power
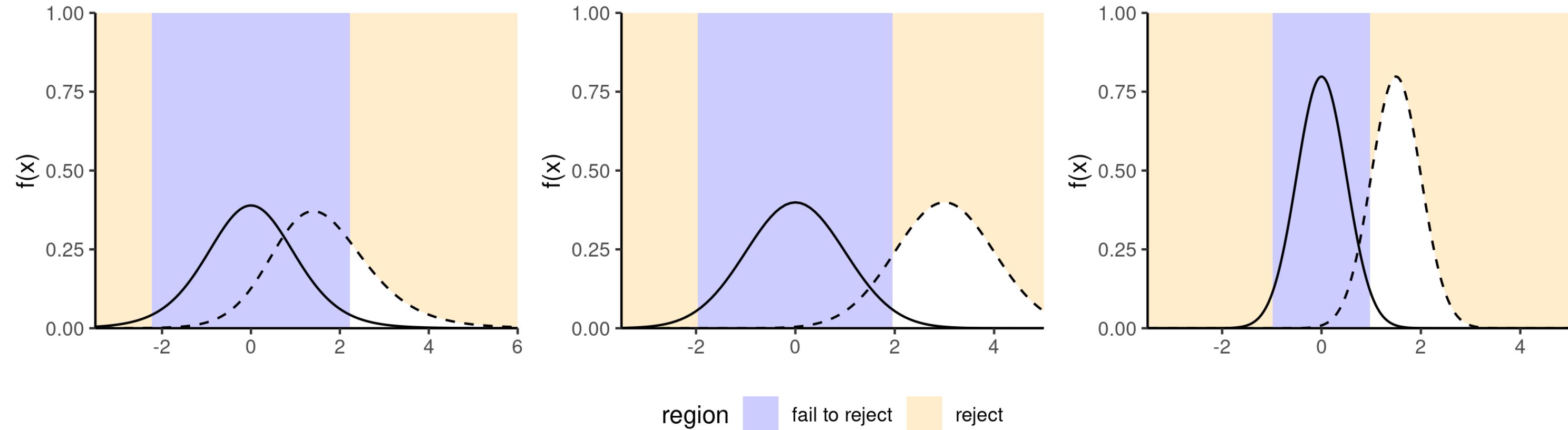
Figure 5: Comparison between null distribution (full curve) and a specific alternative for a *t*-test (dashed line). The power corresponds to the area under the curve of the density of the alternative distribution which is in the rejection area (in white). The middle panel shows an increase in power due to an increase in the mean difference, whereas the right panel shows the change due to a decrease in variability of increase in the sample size.

HEC MONTRÉAL

Criteria determining power

- the effect size: the bigger the difference between the postulated value for θ_0 under $\mathscr{H}_0$ and the observed behavior, the easier it is to detect it, as in the middle panel of Figure 5;

- variability: the less noisy your data, the easier it is to detect differences between the curves

- the sample size: the more observation, the smaller the standard error

- the choice of test statistic

Minimally, the power of the test should be α because we reject the null hypothesis α fraction of the time even when $\mathscr{H}_0$ is true.

HEC MONTRĒAL

Confidence interval

A **confidence interval** is an alternative way to present the conclusions of an hypothesis test performed at significance level α with the **same data units**.

Wald-based $(1 - \alpha)$ confidence intervals for a scalar parameter θ are of the form

$$[\hat{\theta} + \mathfrak{q}_{\alpha/2}\mathrm{se}(\hat{\theta}), \hat{\theta} + \mathfrak{q}_{1-\alpha/2} \times \mathrm{se}(\hat{\theta})]$$

corresponding to a point estimate plus or minus the margin of error.

Estimation: a piece of cake

We distinguish between our target (estimand, e.g., population mean), the recipe or formula (estimator) and the output (estimate).

(a) Estimand

(b) Estimator
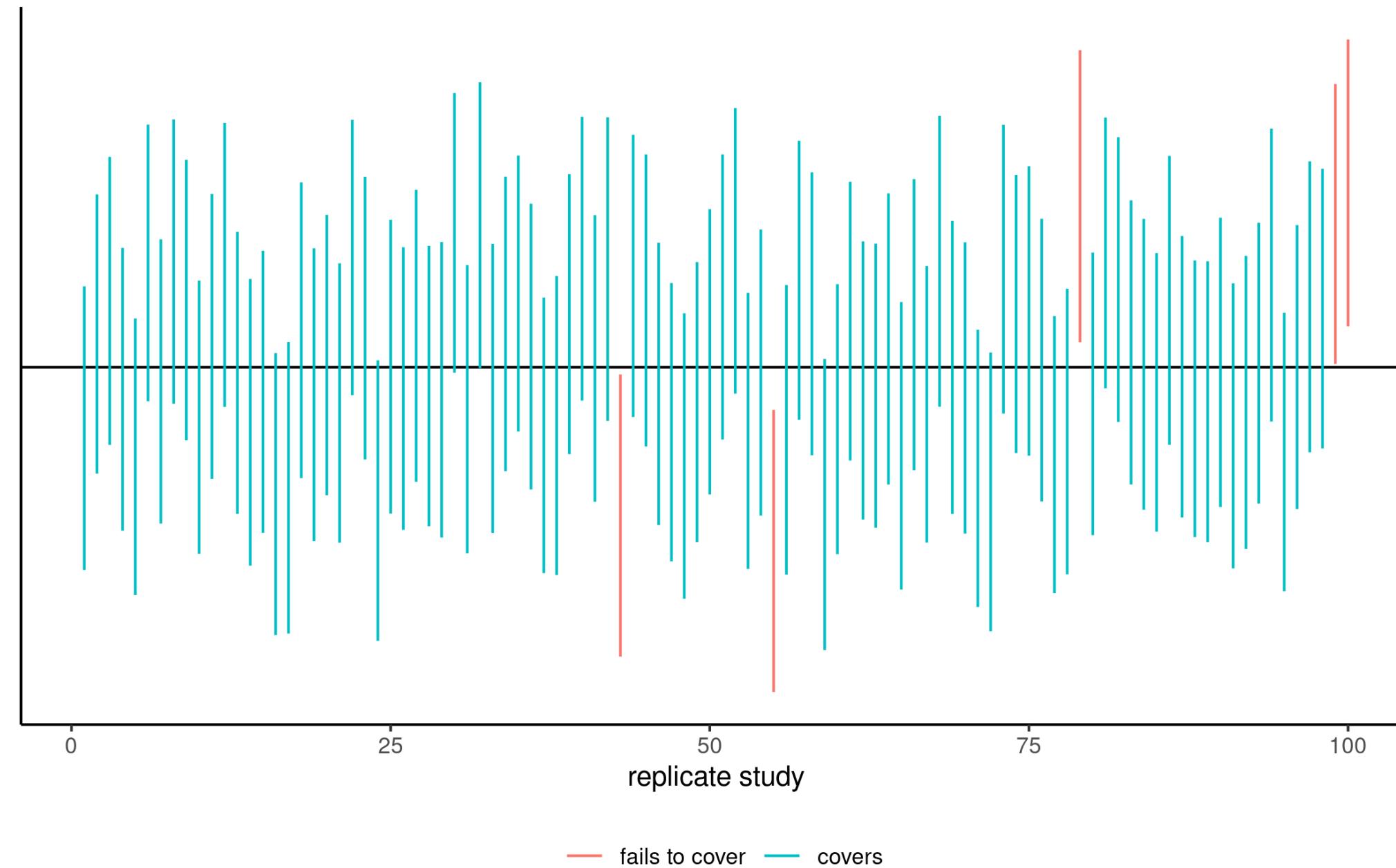
(c) Estimate

Figure 6: Estimand (left), estimator (middle) and estimate (right) illustrated with cakes and based on an original idea of Simon Grund. Cake photos shared under CC BY-NC 2.0 license.

Interpretation of confidence interval

Since the inputs of the confidence interval (estimator) are random, the output is also random and change from one sample to the next: even if you repeat a recipe, you won't always get the exact same result.

- If we calculate the confidence interval for a sample, the true unknown value of the parameter θ is either in the confidence interval or not: there is no notion of probability!

- Our confidence is in the *procedure* we use to calculate confidence intervals and not in the actual values we obtain from a sample.

- If we were to repeat the experiment multiple times, and calculate a $1 - \alpha$ confidence interval each time, then roughly $1 - \alpha$ of the calculated confidence intervals would contain the true value of θ in repeated samples.

Frequentist properties of confidence intervals

95% confidence intervals for the mean of a standard normal population for 100 random samples. On average, 5% of these intervals fail to include the true mean value of zero (in red).

Equivalence of confidence intervals and p-values

The $1 - \alpha$ confidence interval gives us all values for which we fail to reject $\mathcal{H}_0$ at level α

- Binary decision rule with confidence interval: if the postulated value lies inside the interval, we fail to reject the null hypothesis.

- The p-value based on the same procedure would be $p \geq \alpha$ if you fail to reject (and vice-versa $p < \alpha$ if the value is outside the interval).

But confidence intervals are in terms of the data units, so easier to interpret.

5. Obtain a *p*-value or a confidence interval

The *p*-value is $p = \mathrm{Pr}_0(T > t_D)$, where $T \sim \mathbf{Student}(34)$. Using **R**, we find $p = 0.0032$, which is smaller than $\alpha = 5\%$.

```
1  ttest$p.value
2  ## [1] 0.00319
```

The lower bound of the confidence interval is $\overline{D} + \mathbf{se}(\overline{D}) \times t_{0.05}$.

```
1  ttest$conf.int
2  ## [1] 0.131    Inf
3  ## attr(,"conf.level")
4  ## [1] 0.95
```

The one-sided confidence interval is $[0.131, \infty]$. The postulated null value, 0, is outside the interval.

R — calculations by hand

```
1  d <- with(distraction, t - c) # time difference text vs conversation
2  n <- length(d) # sample size
3  (mean_d <- mean(d)) # mean difference
4  ## [1] 0.313
5  (se_d <- sd(d)/sqrt(n)) # standard error of sample mean
6  ## [1] 0.108
7  (stat <- mean_d/se_d) # t-test statistic
8  ## [1] 2.91
9  dof <- n - 1L # degrees of freedom
10 crit <- qt(p = 0.05, df = dof) # critical value, "q" for quantile
11 (pval <- pt(q = stat, df = dof, lower.tail = FALSE)) # Pr(T > stat)
12 ## [1] 0.00319
13 (conf_low <- mean_d + se_d*crit) # lower bound of Wald confidence interval
14 ## [1] 0.131
```

R uses the convention `d/p/q/r` for density/distribution function/quantile function/random number generation. The name of the distribution (e.g., `norm`, `t`, `chisq`, `f`, `unif`) is appended to that letter, e.g.,

- `pnorm` is the distribution function of a normal variable,

- `qt` is the quantile function of a Student-t variable,

- `runif` generates uniform variables.

HEC MONTRĒAL

6. Conclude in the setting of the problem

The estimated mean difference is 0.313 seconds (std. error of 0.108 seconds).

We reject $\mathscr{H}_0$, meaning that the reaction time is significantly higher (at level 5%) when texting than talking on the cellphone while walking (p-value of 0.003).

Learning objectives

ⓘ Learning objectives

- Understanding the role of uncertainty in decision making.

- Understanding the importance of signal-to-noise ratio as a measure of evidence.

- Knowing the basic ingredients of hypothesis testing and being capable of correctly formulating and identifying these components in a paper.

- Correctly interpreting p-values and confidence intervals for a parameter.

References

Brodeur, Mathieu, Perrine Ruer, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. "Smartwatches Are More Distracting Than Mobile Phones While Driving: Results from an Experimental Study." *Accident Analysis & Prevention* 149: 105846. https://doi.org/10.1016/j.aap.2020.105846.

Brucks, Melanie S., and Jonathan Levav. 2022. "Virtual Communication Curbs Creative Idea Generation." *Nature* 605 (7908): 108–12. https://doi.org/10.1038/s41586-022-04643-y.

Duke, Kristen E., and On Amir. 2023. "The Importance of Selling Formats: When Integrating Purchase and Quantity Decisions Increases Sales." *Marketing Science* 42 (1): 87–109. https://doi.org/10.1287/mksc.2022.1364.

Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. "The Upworthy Research Archive, a Time Series of 32,487 Experiments in U.S. Media." *Scientific Data* 8 (195). https://doi.org/10.1038/s41597-021-00934-7.

Moon, Alice, and Eric M VanEpps. 2023. "Giving Suggestions: Using Quantity Requests to Increase Donations." *Journal of Consumer Research* 50 (1): 190–210. https://doi.org/10.1093/jcr/ucac047.

Sokolova, Tatiana, Aradhna Krishna, and Tim Döring. 2023. "Paper Meets Plastic: The Perceived Environmental Friendliness of Product Packaging." *Journal of Consumer Research* 50 (3): 468–91. https://doi.org/10.1093/jcr/ucad008.

HEC MONTRÉAL