Hypothesis testing

Session 2

MATH 80667A: Experimental Design and Statistical Methods
HEC Montréal

Outline

Variability

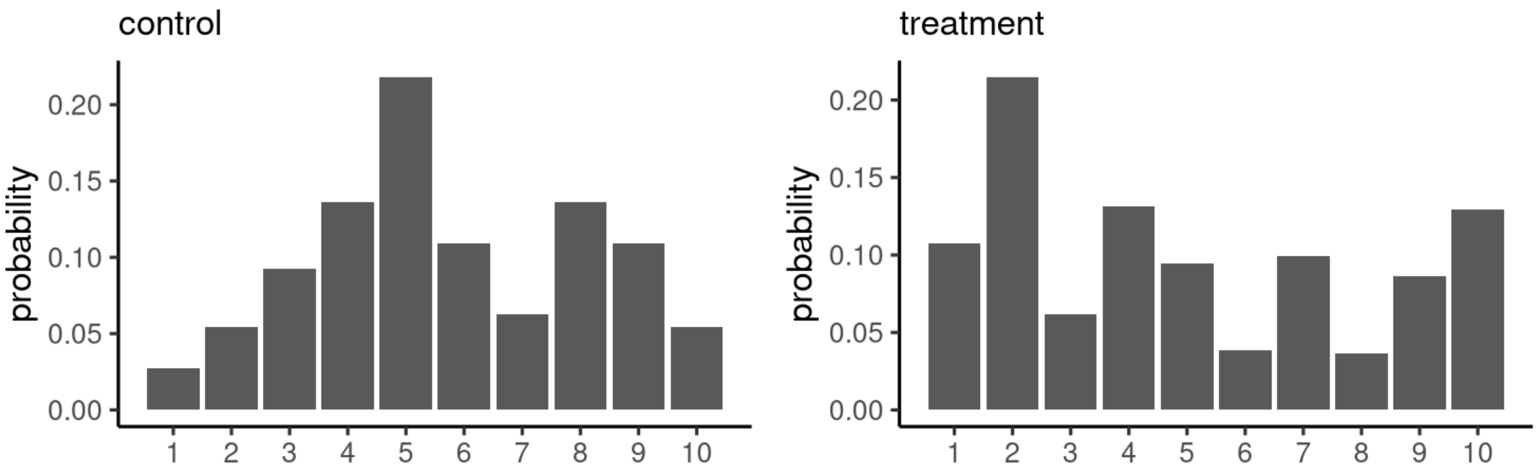
Hypothesis tests

Pairwise comparisons

Sampling variability

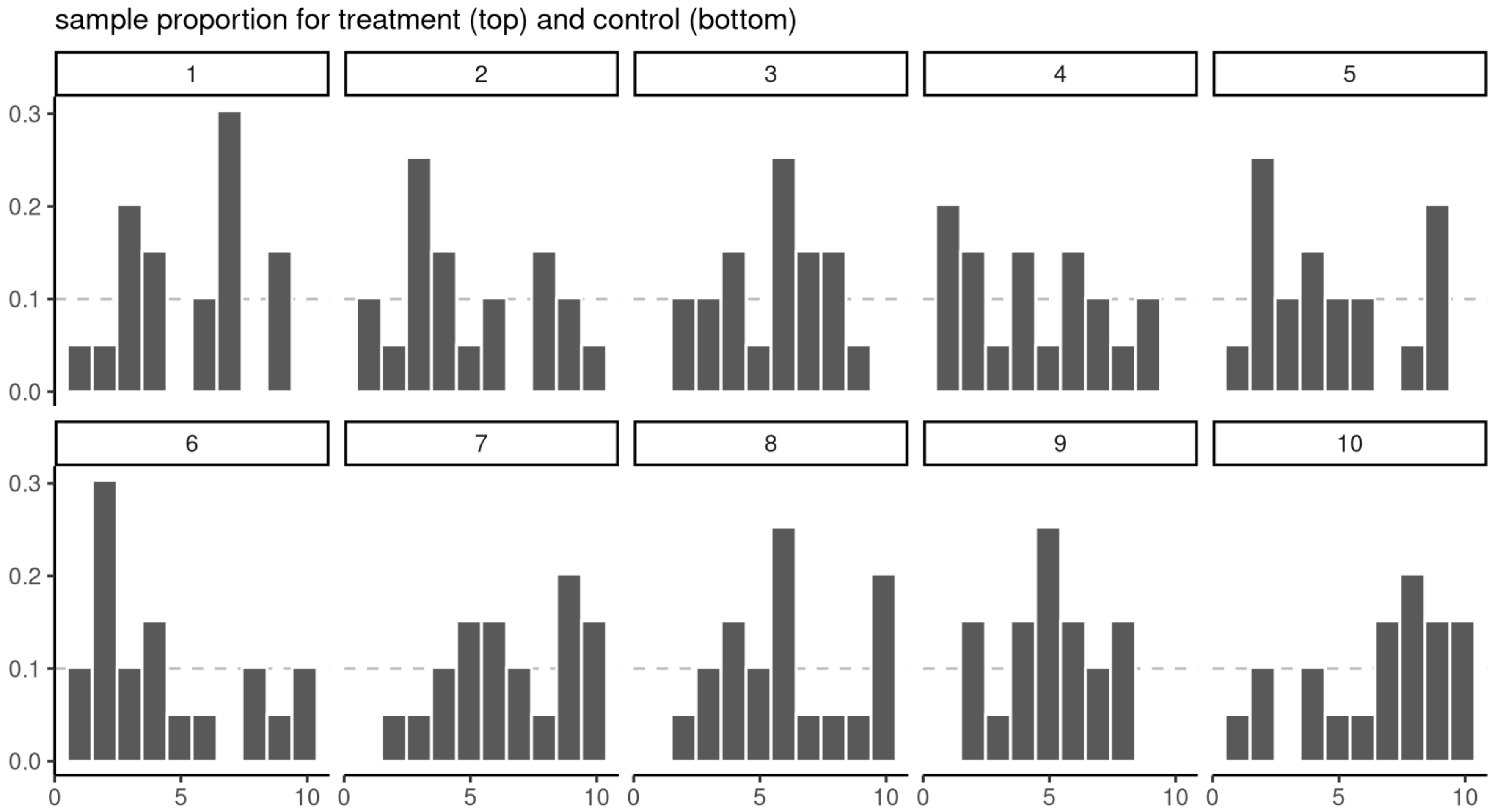
Studying a population

Interest in impacts of intervention or policy



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

Sampling variability



Decision making under uncertainty

- Data collection costly

  $\rightarrow$ limited information available about population.
- Sample too small to reliably estimate distribution
- Focus instead on particular summaries

  $\rightarrow$ mean, variance, odds, etc.

Population characteristics

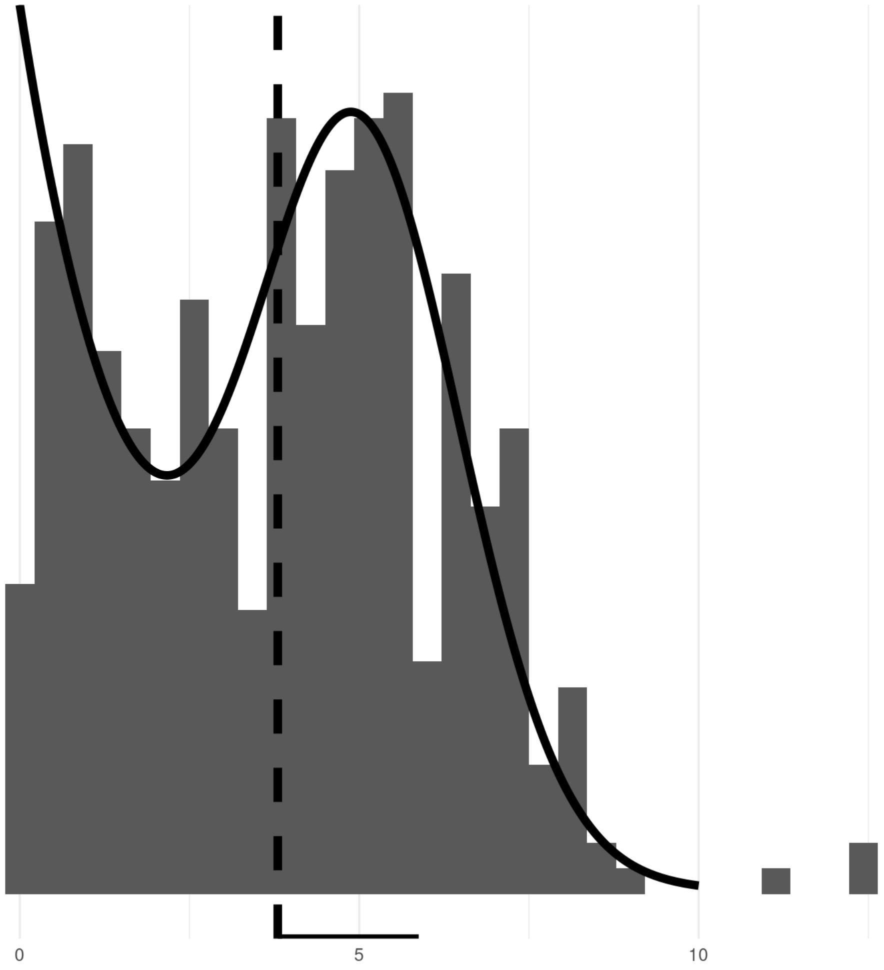
mean / expectation

$$\mu$$

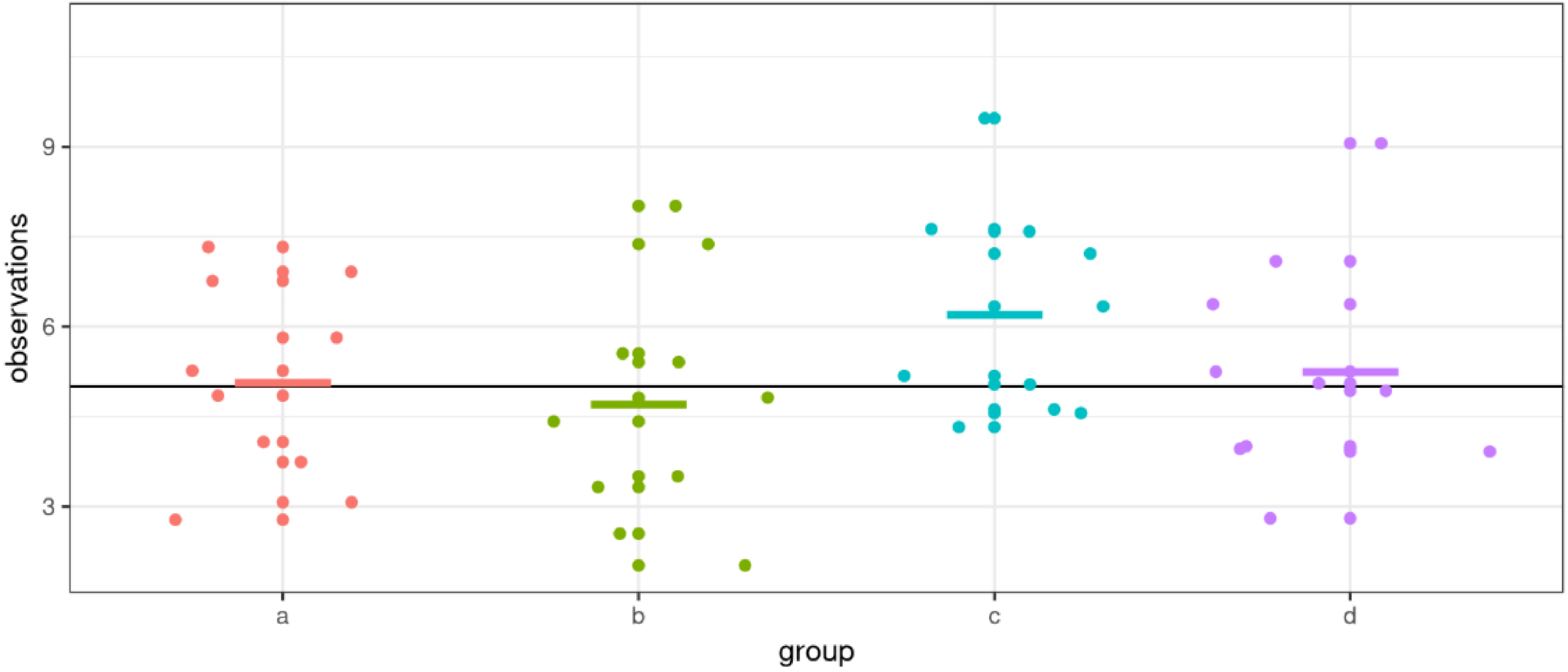
standard deviation

$$\sigma = \sqrt{\text{variance}}$$

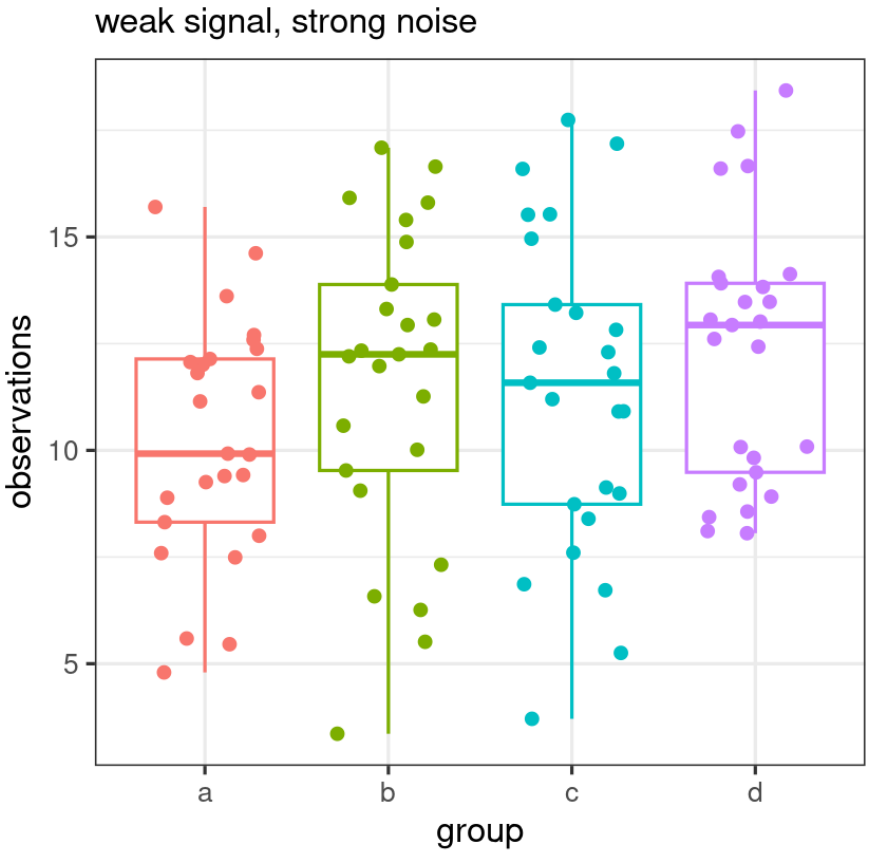
same scale as observations

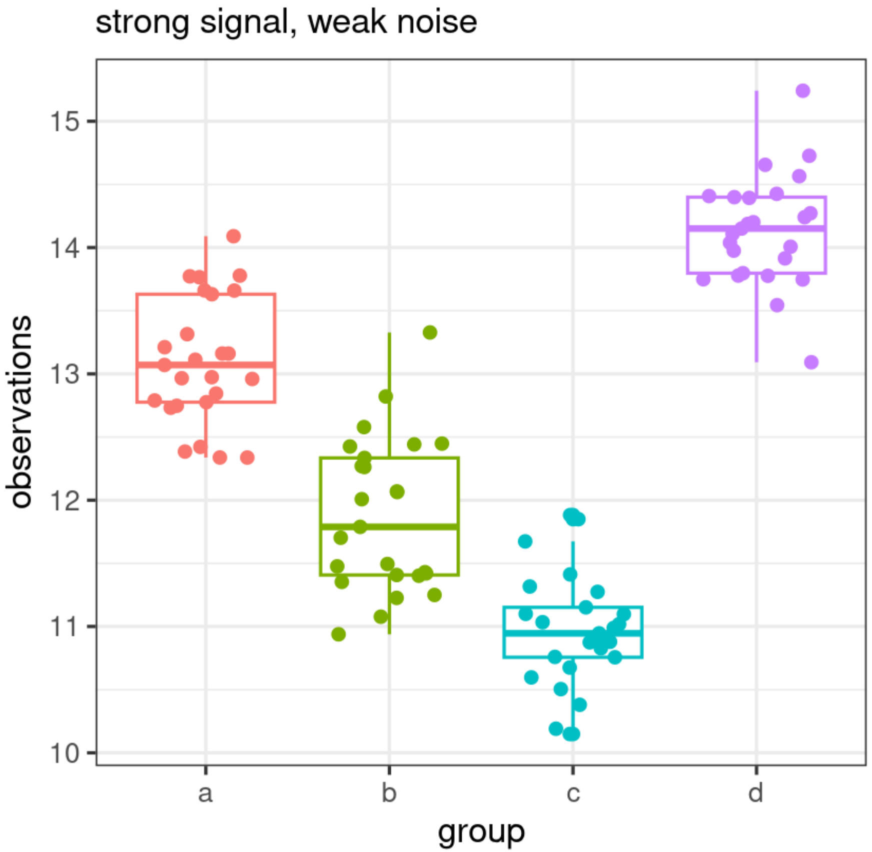


Sampling variability



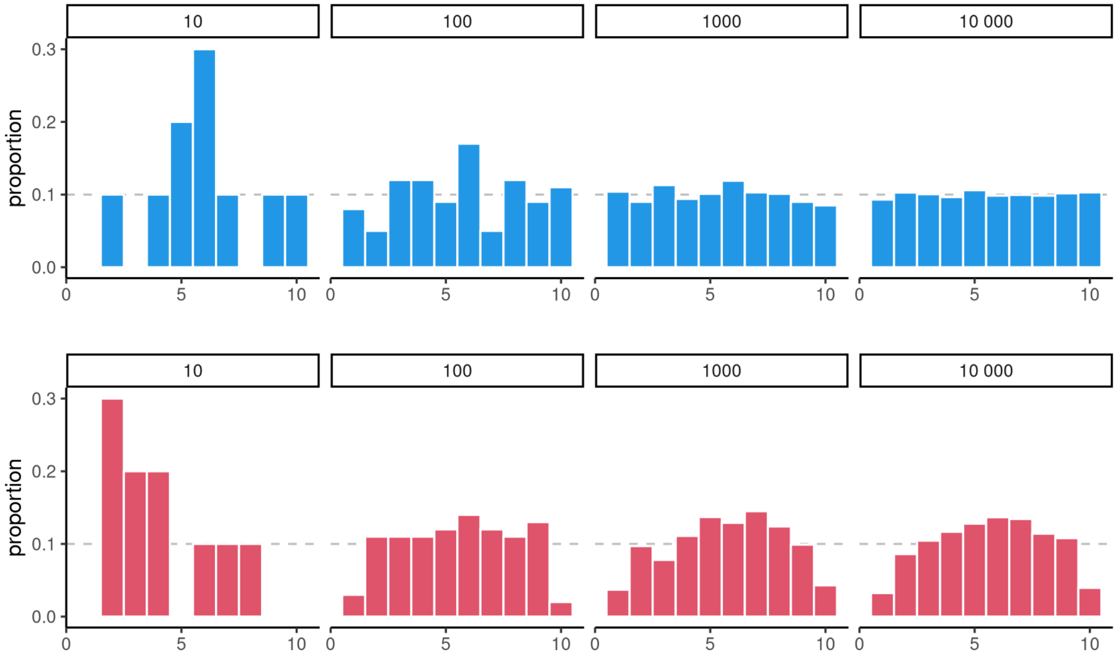
The signal and the noise





Can you spot the differences?

Information accumulates



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

Hypothesis tests

The general recipe of hypothesis testing

1. Define variables
2. Write down hypotheses (null/alternative)
3. Choose and compute a test statistic
4. Compare the value to the null distribution (benchmark)
5. Compute the p-value
6. Conclude (reject/fail to reject)
7. Report findings

Hypothesis tests versus trials

- Binary decision: guilty/not guilty
- Summarize evidences (proof)
- Assess evidence in light of **presumption of innocence**
- Verdict: either guilty or not guilty
- Potential for judicial mistakes

How to assess evidence?

statistic = numerical summary of the data.

requires benchmark / standardization

typically a unitless quantity

need measure of uncertainty of statistic

General construction principles

Wald statistic

$$W = \frac{\text{estimated qty} - \text{postulated qty}}{\text{std. error}\left(\text{estimated qty}\right)}$$

standard error = measure of variability (same units as obs.)

resulting ratio is unitless!

Impact of encouragement on teaching

From Davison (2008), Example 9.2

> In an investigation on the teaching of arithmetic, 45 pupils were divided at random into five groups of nine. Groups A and B were taught in separate classes by the usual method. Groups C, D, and E were taught together for a number of days. On each day C were praised publicly for their work, D were publicly reproved and E were ignored. At the end of the period all pupils took a standard test.

Basic manipulations in **R**: load data

```r
data(arithmetic,
     package = "hecedsm")
# categorical variable = factor

# Look up data
str(arithmetic)
```

```
## 'data.frame':	45 obs. of  2 variables:
##  $ group: Factor w/ 5 levels "control 1","control 2",..: 1 1 1 1 1 1 1 1
##  $ score: num  17 14 24 20 24 23 16 15 24 21 ...
```

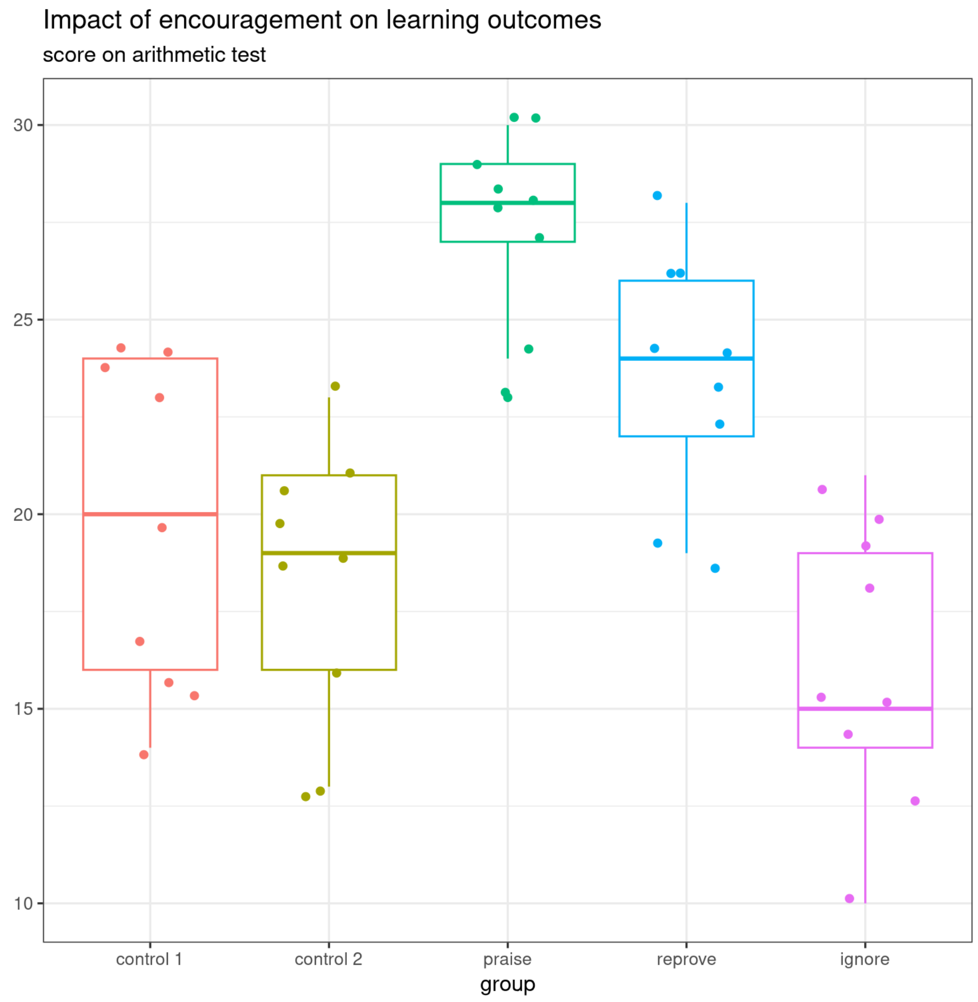
Basic manipulations in **R**: summary statistics

```r
# compute summary statistics
summary_stat <-
  arithmetic |>
  group_by(group) |>
  summarize(mean = mean(score),
            sd = sd(score))
knitr::kable(summary_stat,
             digits = 2)
```

group	mean	sd
control 1	19.67	4.21
control 2	18.33	3.57
praise	27.44	2.46
reprove	23.44	3.09
ignore	16.11	3.62

Basic manipulations in **R**: plot

```
# Boxplot with jittered data
ggplot(data = arithmetic,
       aes(x = group,
           y = score,
           color = group)) +
  geom_boxplot() +
  geom_jitter(width = 0.3) +
  theme_bw()
```



Formulating an hypothesis

Let μ_C and μ_D denote the population average (expectation) score for praise and reprove, respectively.

Our null hypothesis is

$$\mathscr{H}_0 : \mu_C = \mu_D$$

against the alternative $\mathscr{H}_a$ that they are different (two-sided test).

Equivalent to $\delta_{CD} = \mu_C - \mu_D = 0$.

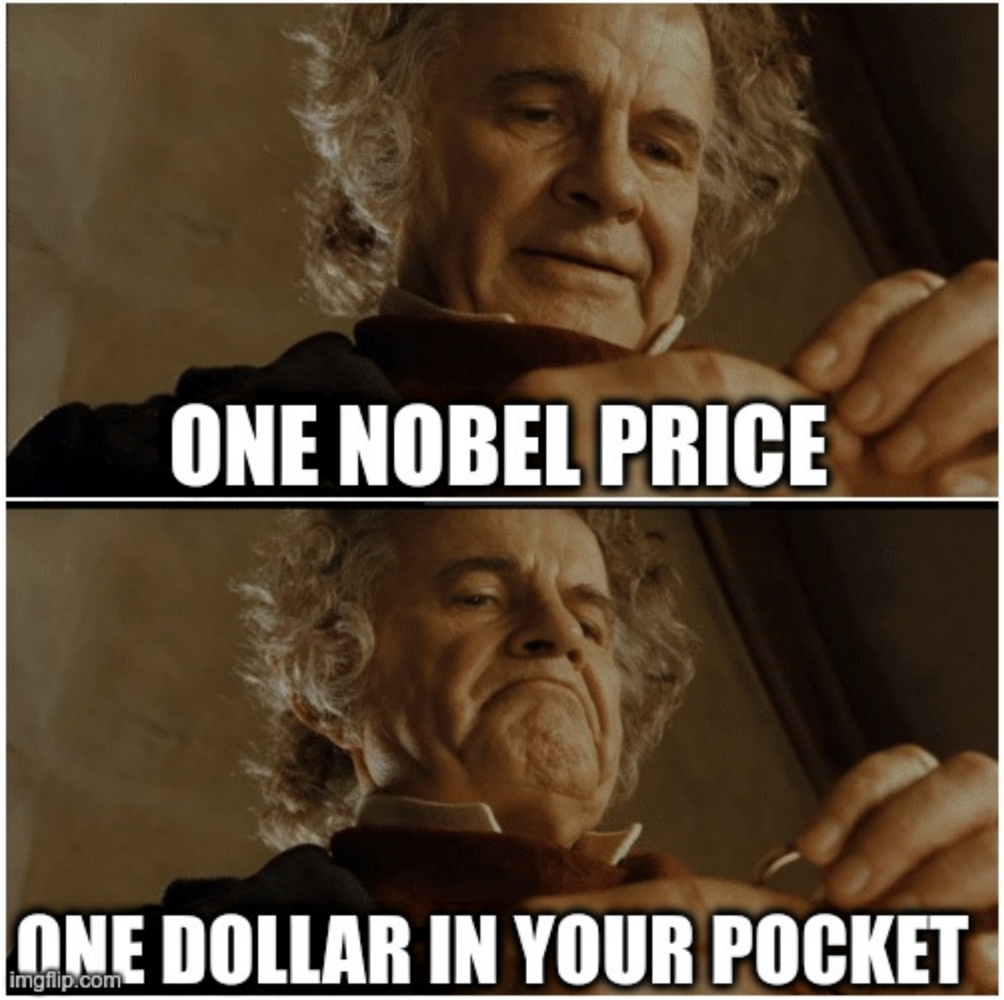
Test statistic

The value of the Wald statistic is

$$t = \frac{\hat{\delta}_{CD} - 0}{\mathsf{se}(\hat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

How 'extreme' is this number?

Assessing evidence



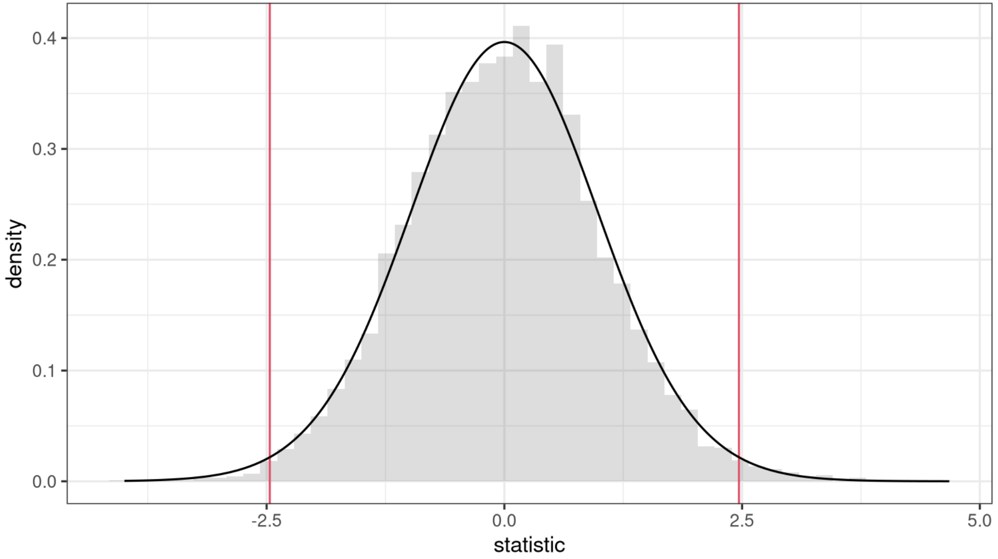
Benchmarking

- The same number can have different meanings
    - units matter!
- Meaningful comparisons require some reference

Possible, but not plausible

The null distribution tells us what are the *plausible* values for the statistic and their relative frequency if the null hypothesis holds.

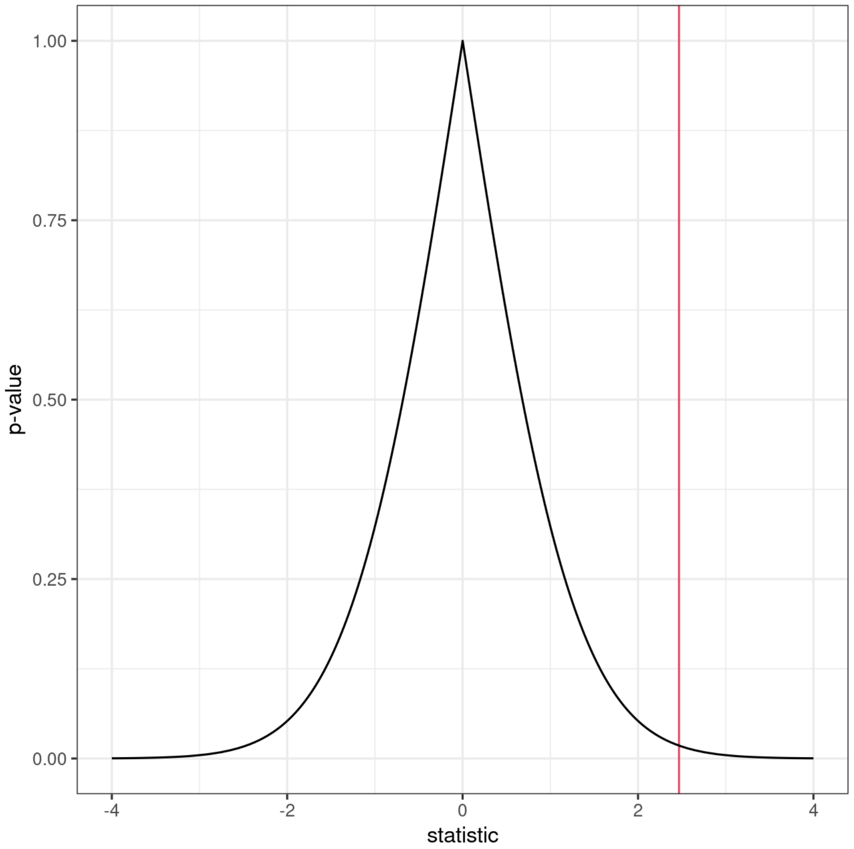
What can we expect to see **by chance** if there is **no difference** between groups?



P-value

Null distributions are different, which makes comparisons uneasy.

- The *p*-value gives the probability of observing an outcome as extreme **if the null hypothesis was true**.



Level = probability of condemning an innocent

Fix level α before the experiment.

Choose small α (typical value is 5%)

Reject $\mathscr{H}_0$ if p-value less than α

What is really a *p*-value?

The American Statistical Association (ASA) published a statement on (mis)interpretation of p-values.

> (2) P-values do not measure the probability that the studied hypothesis is true
>
> (3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
>
> (4) P-values and related analyses should not be reported selectively
>
> (5) P-value, or statistical significance, does not measure the size of an effect or the importance of a result

Reporting results of a statistical procedure

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☐	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☐	A description of all covariates tested
☐	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☐	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☐	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Nature's checklist

Pairwise comparisons

Pairwise differences and t-tests

The pairwise differences (p-values) and confidence intervals for groups j and k are based on the t-statistic:

$$t = \frac{\text{estimated} - \text{postulated difference}}{\text{uncertainty}} = \frac{(\widehat{\mu}_j - \widehat{\mu}_k) - (\mu_j - \mu_k)}{\mathsf{se}(\widehat{\mu}_j - \widehat{\mu}_k)}.$$

In large sample, this statistic behaves like a Student-t variable with $n-K$ degrees of freedom, denoted $\mathsf{St}(n-K)$ hereafter.

Note: in an analysis of variance model, the standard error $\mathsf{se}(\widehat{\mu}_j - \widehat{\mu}_k)$ is based the pooled variance estimate (estimated using all observations).

Pairwise comparison

Consider the pairwise average difference in scores between the praise (group C) and the reprove (group D) of the `arithmetic` data.

- Group sample averages are $\widehat{\mu}_C = 27.4$ and $\widehat{\mu}_D = 23.4$
- The estimated average difference between groups C and D is $\hat{\delta}_{CD} = 4$
- The estimated pooled *standard deviation* for the five groups is 1.15
- The *standard error* for the pairwise difference is **se**$(\hat{\delta}_{CD}) = 1.6216$
- There are $n = 45$ observations and $K = 5$ groups

t-tests: null distribution is Student-t

If we postulate $\delta_{jk} = \mu_j - \mu_k = 0$, the test statistic becomes

$$t = \frac{\hat{\delta}_{jk} - 0}{\mathsf{se}(\hat{\delta}_{jk})}$$

The p-value is $p = 1 - \Pr(-|t| \leq T \leq |t|)$ for $T \sim \mathsf{St}_{n-K}$.

- probability of statistic being more extreme than t

Recall: the larger the values of the statistic t (either positive or negative), the more evidence against the null hypothesis.

Critical values

For a test at level α (two-sided), we fail to reject null hypothesis for all values of the test statistic t that are in the interval

$$\mathfrak{t}_{n-K}(\alpha/2) \leq t \leq \mathfrak{t}_{n-K}(1-\alpha/2)$$

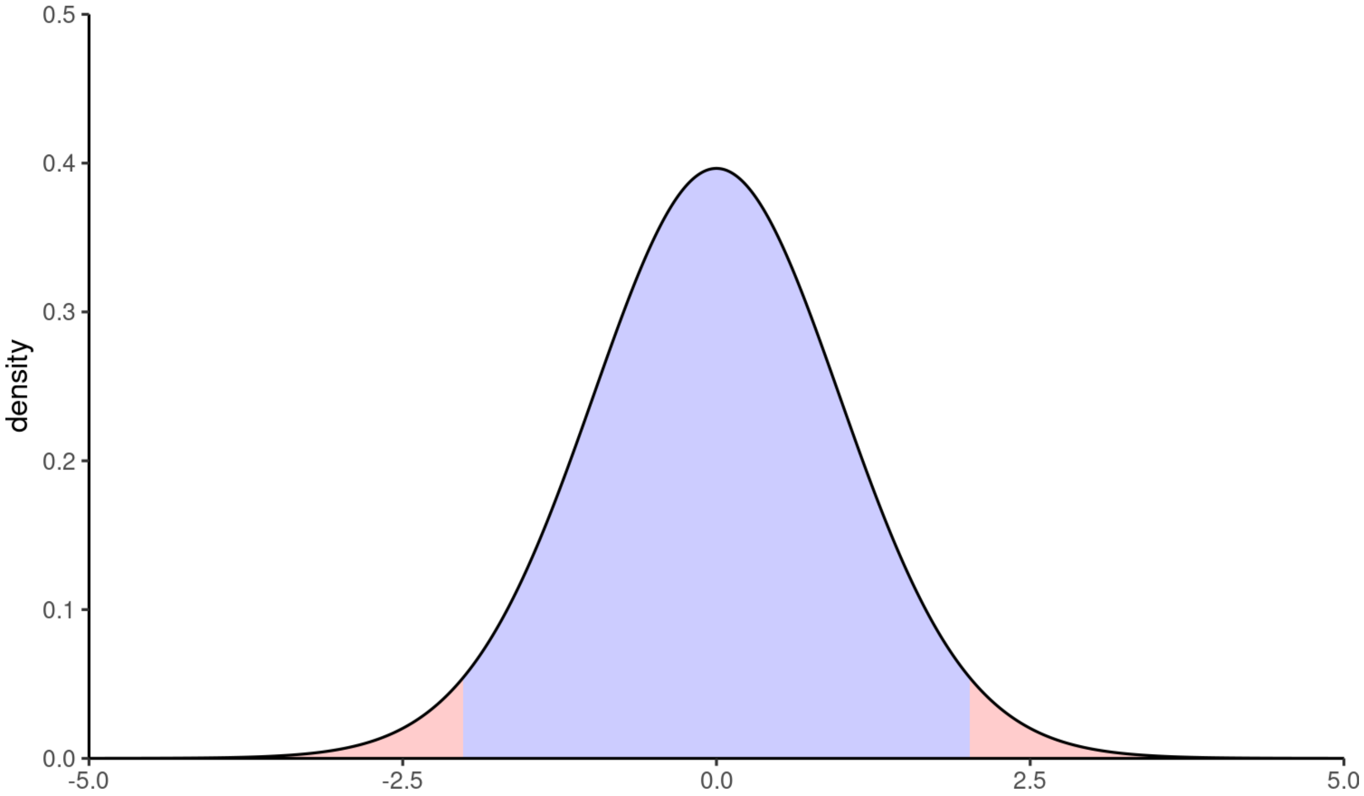
Because of the symmetry around zero, $\mathfrak{t}_{n-K}(1-\alpha/2) = -\mathfrak{t}_{n-K}(\alpha/2)$.

- We call $\mathfrak{t}_{n-K}(1-\alpha/2)$ a **critical value**.
- in **R**, the quantiles of the Student-t distribution are obtained from `qt(1-alpha/2, df = n - K)` where `n` is the number of observations and `K` the number of groups.

Null distribution

The blue area defines the set of values for which we fail to reject null $\mathscr{H}_0$.

All values of t falling in the red area lead to rejection at level 5%.



Example

- If $\mathscr{H}_0 : \delta_{CD} = 0$, the t statistic is

$$t = \frac{\hat{\delta}_{CD} - 0}{\text{se}(\hat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

- The p-value is $p = 0.018$.
- We reject the null at level $\alpha = 5\%$ since $0.018 < 0.05$.
- Conclude that there is a significant difference at level $\alpha = 0.05$ between the average scores of subpopulations C and D.

Confidence interval

Let $\delta_{jk} = \mu_j - \mu_k$ denote the population difference, $\hat{\delta}_{jk}$ the estimated difference (difference in sample averages) and $\mathsf{se}(\hat{\delta}_{jk})$ the estimated standard error.

The region for which we fail to reject the null is

$$-\mathfrak{t}_{n-K}(1-\alpha/2) \leq \frac{\hat{\delta}_{jk} - \delta_{jk}}{\mathsf{se}(\hat{\delta}_{jk})} \leq \mathfrak{t}_{n-K}(1-\alpha/2)$$

which rearranged gives the $(1-\alpha)$ confidence interval for the (unknown) difference δ_{jk}.

$$\hat{\delta}_{jk} - \mathsf{se}(\hat{\delta}_{jk})\mathfrak{t}_{n-K}(1-\alpha/2) \leq \delta_{jk} \leq \hat{\delta}_{jk} + \mathsf{se}(\hat{\delta}_{jk})\mathfrak{t}_{n-K}(1-\alpha/2)$$

Interpretation of confidence intervals

The reported confidence interval is of the form

$$\text{estimate} \pm \text{critical value} \times \text{standard error}$$

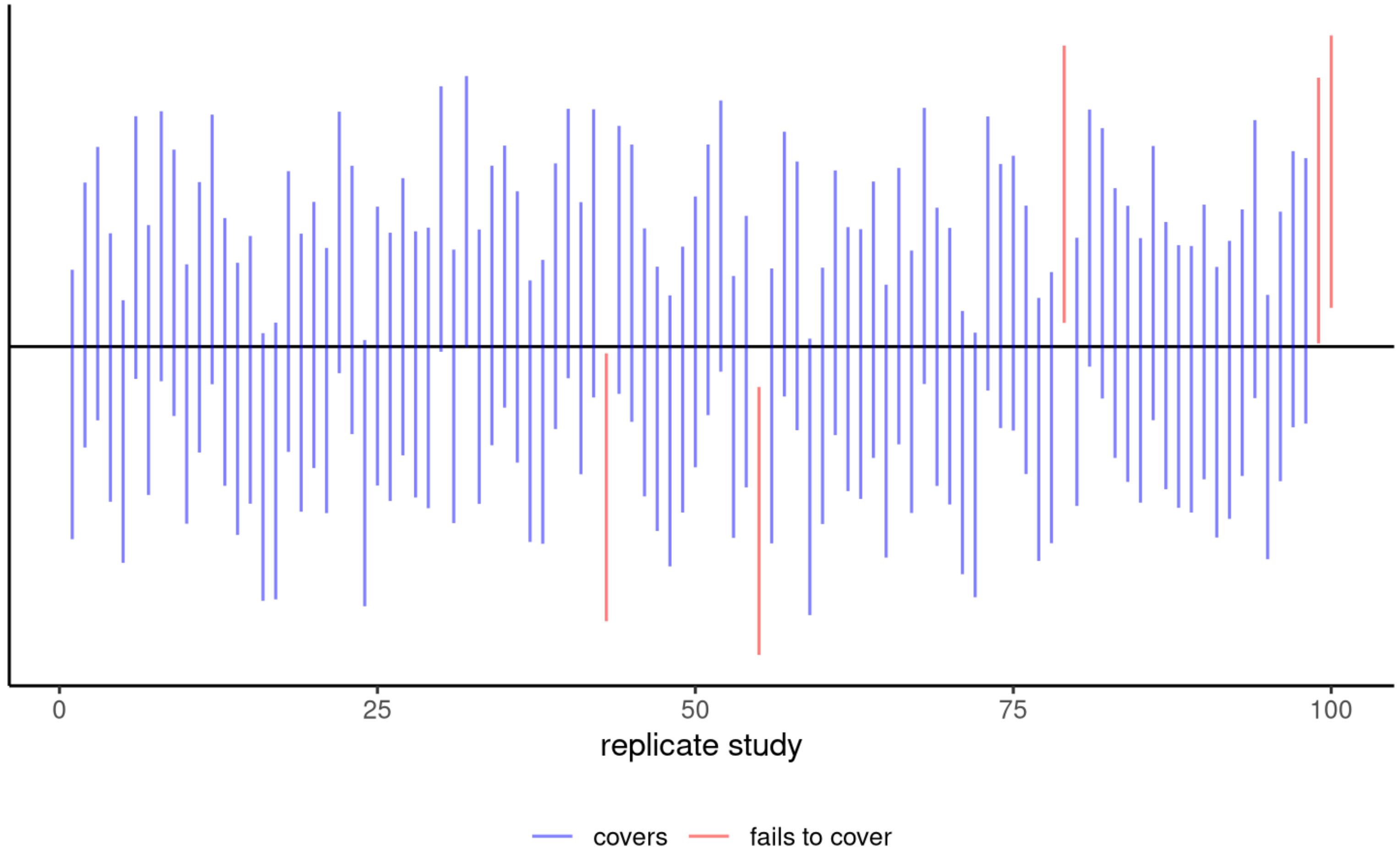
confidence interval = [lower, upper] units

If we replicate the experiment and compute confidence intervals each time

- on average, 95% of those intervals will contain the true value if the assumptions underlying the model are met.

Interpretation in a picture: coin toss analogy

Each interval either contains the true value (black horizontal line) or doesn't.



Why confidence intervals?

Test statistics are standardized,

- Good for comparisons with benchmark
- typically meaningless (standardized = unitless quantities)

Two options for reporting:

- p-value: probability of more extreme outcome if no mean difference
- confidence intervals: set of all values for which we fail to reject the null hypothesis at level α for the given sample

Example

- Mean difference of $\hat{\delta}_{CD} = 4$, with **se**$(\hat{\delta}_{CD}) = 1.6216$.
- The critical values for a test at level $\alpha = 5\%$ are -2.021 and 2.021
  - `qt(0.975, df = 45 - 5)`
- Since $|t| > 2.021$, reject $\mathscr{H}_0$: the two population are statistically significant at level $\alpha = 5\%$.
- The confidence interval is

$$[4 - 1.6216 \times 2.021, 4 + 1.6216 \times 2.021] = [0.723, 7.277]$$

The postulated value $\delta_{CD} = 0$ is not in the interval: reject $\mathscr{H}_0$.

Pairwise differences in **R**

```r
library(emmeans) # marginal means and contrasts
model <- aov(score ~ group, data = arithmetic)
margmeans <- emmeans(model, specs = "group")
contrast(margmeans,
         method = "pairwise",
         adjust = 'none',
         infer = TRUE) |>
  as_tibble() |>
  filter(contrast == "praise - reprove") |>
  knitr::kable(digits = 3)
```

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
praise - reprove	4	1.622	40	0.723	7.277	2.467	0.018

Recap 1

- Due to sampling variability, looking at differences between empirical measures (sample mean, etc.) is not enough.
- Testing procedures factor in the uncertainty inherent to sampling.
- Adopt particular viewpoint: null hypothesis (simpler model, e.g., no difference between group) is true. We consider the evidence under that optic.

Recap 2

- p-values measures compatibility with the null model (relative to an alternative)
- Tests are standardized values,

The output is either a p-value or a confidence interval

- confidence interval: on scale of data (meaningful interpretation)
- p-values: uniform on [0,1] if the null hypothesis is true

Recap 3

- All hypothesis tests share common ingredients
- Many ways, models and test can lead to the same conclusion.
- Transparent reporting is important!