Hypothesis testing

Session 2

MATH 80667A: Experimental Design and Statistical Methods
for Quantitative Research in Management
HEC Montréal

Outline

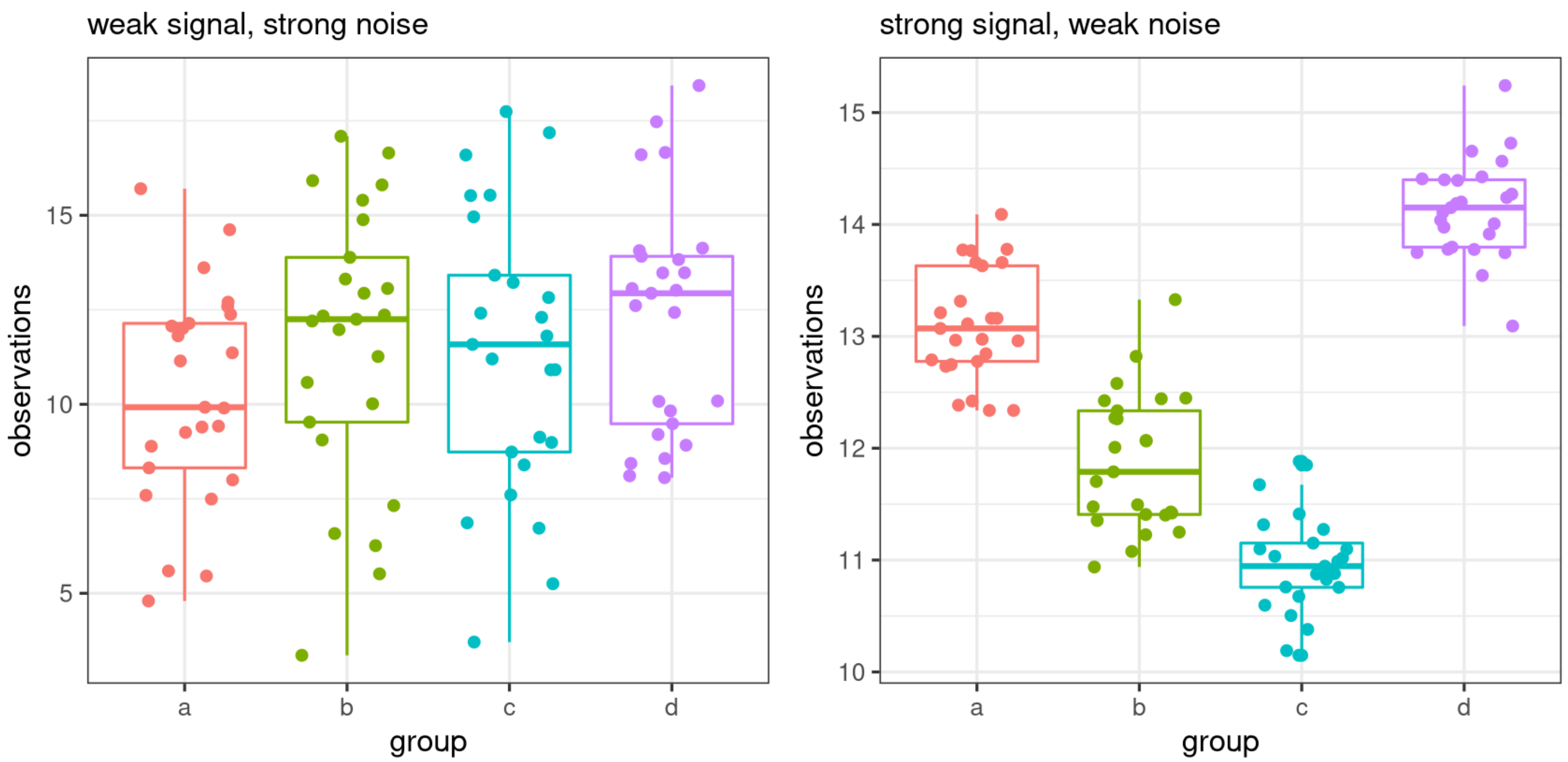
Variability

Hypothesis tests

R examples

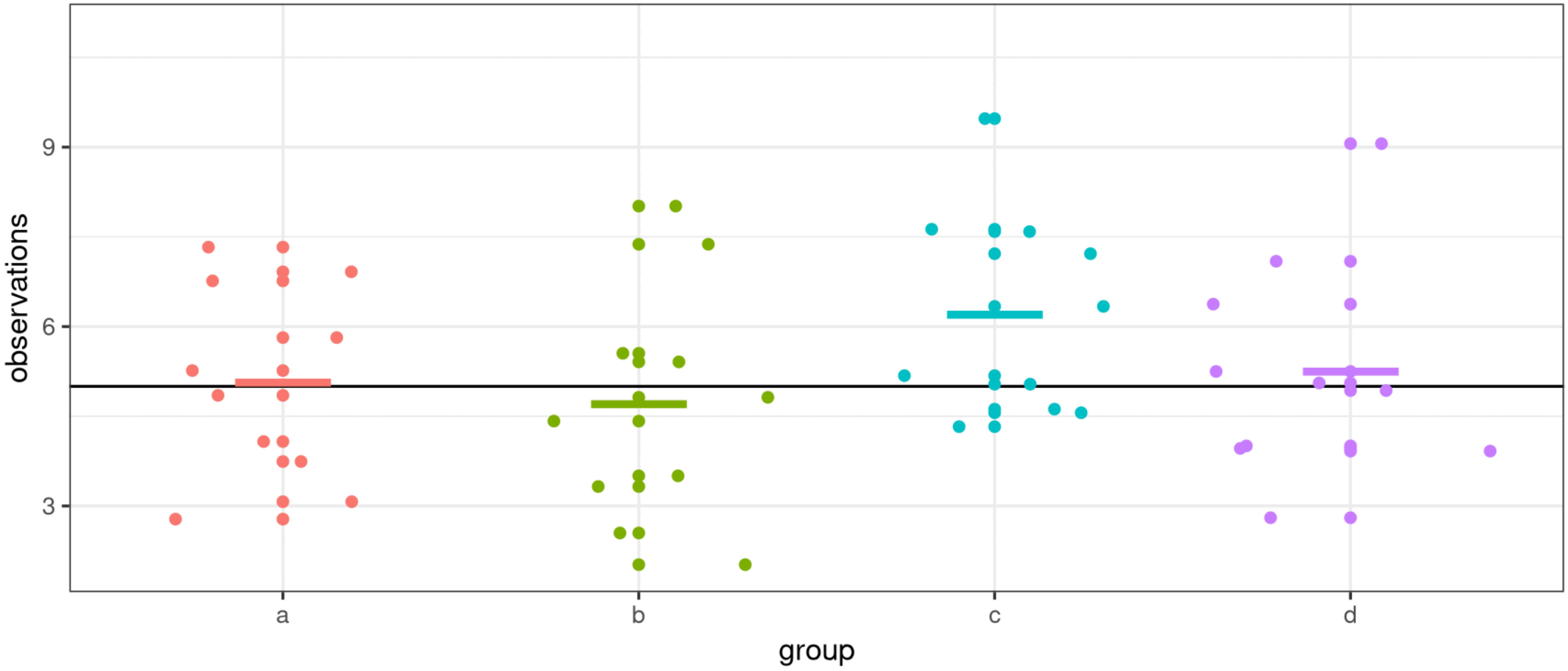
Signal versus noise

The signal and the noise



Can you spot the differences?

Sampling variability



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

Scene from "12 Angry Men" by Sidney Lumet

Trial

- 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

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.

Load data | Summary statistics | Plot

```r
# Load libraries
library(tidyverse)
# Load and reformat data
url <- "https://edsm.rbind.io/dat
arithmetic <-
 read_csv(url) %>%
  mutate(group = factor(group))
# categorical variable == factor
glimpse(arithmetic)
```

```
## Rows: 45
## Columns: 2
## $ group <fct> A, A, A, A, A,…
## $ score <dbl> 17, 14, 24, 20…
```

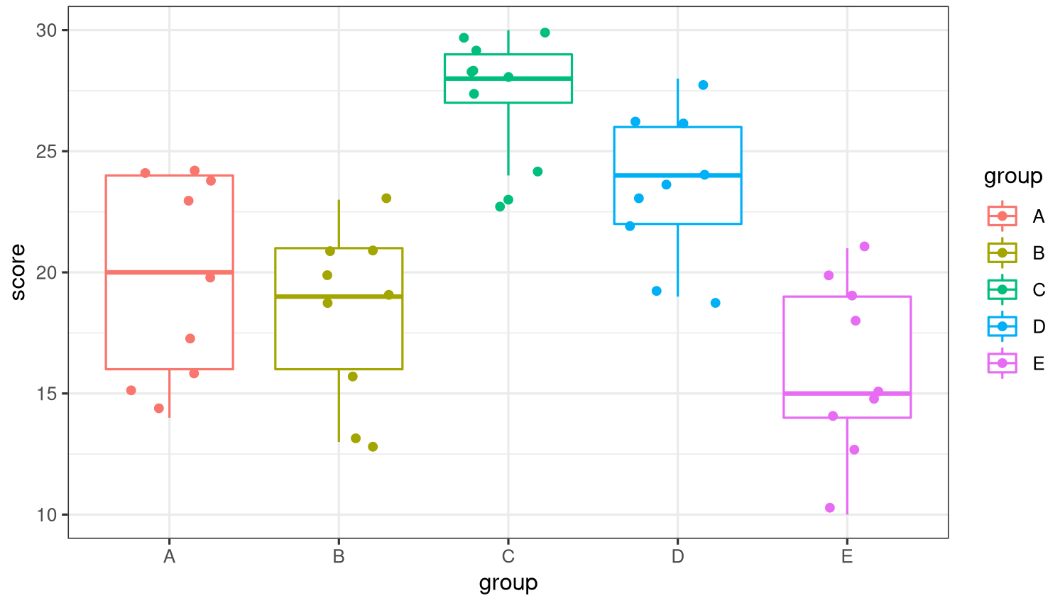
Load data | **Summary statistics** | Plot

```r
# compute summary statistics
arithmetic %>%
  group_by(group) %>%
  summarize(mean = mean(score),
            sd = sd(score))
```

```
## # A tibble: 5 × 3
##   group  mean    sd
##   <fct> <dbl> <dbl>
## 1 A      19.7  4.21
## 2 B      18.3  3.57
## 3 C      27.4  2.46
## 4 D      23.4  3.09
## 5 E      16.1  3.62
```

Load data   Summary statistics   **Plot**

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



Pick a test, compute its value

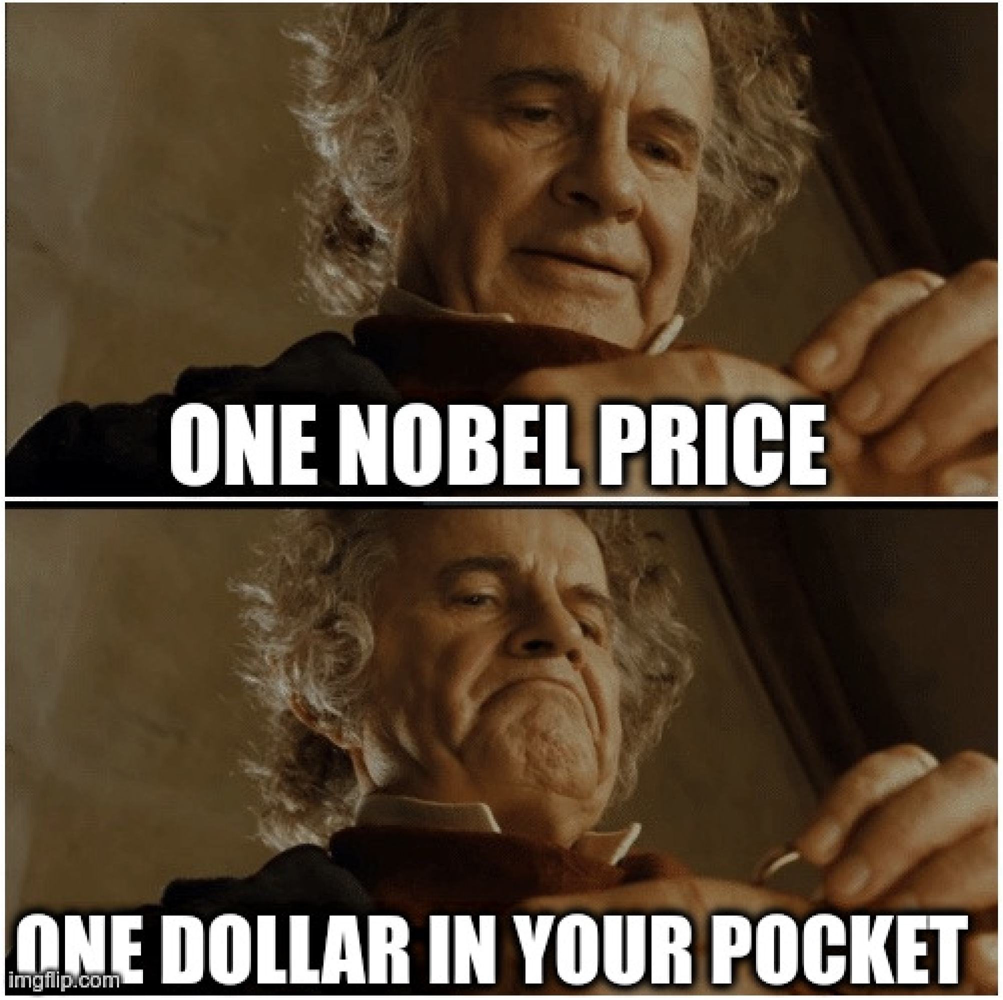
One-way analysis of variance uses an F statistic.

```r
#one way analysis of variance
aov(data = arithmetic,
    formula = score ~ group)
```

- In **R**, the function `anova` prints the analysis of variance table.
- The value of the statistic is 15.268.

How 'extreme' is this number?

Assessing evidence



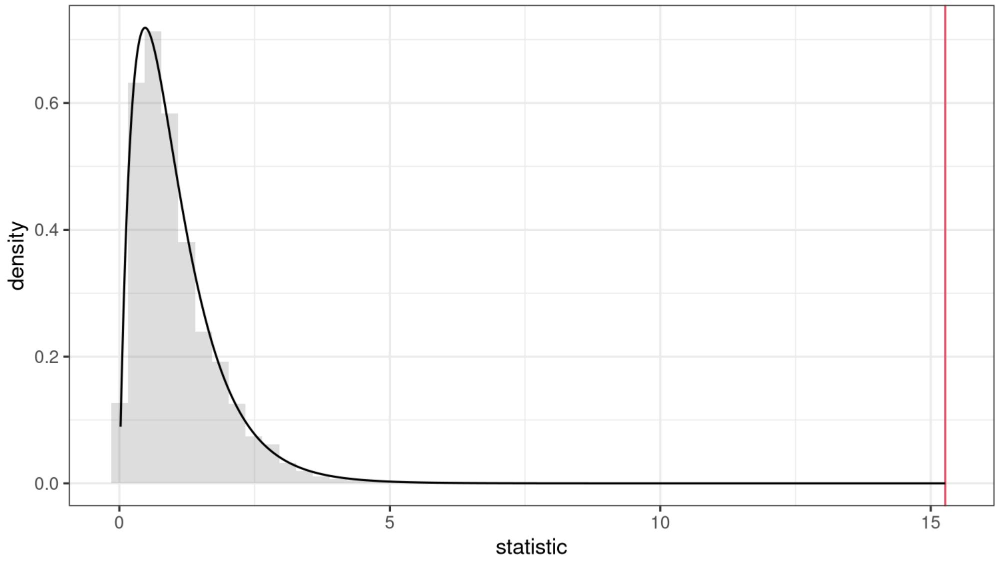
Benchmarking

- 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 there relative frequency

- what can we expect to see **by chance** if there is **no difference** between groups.

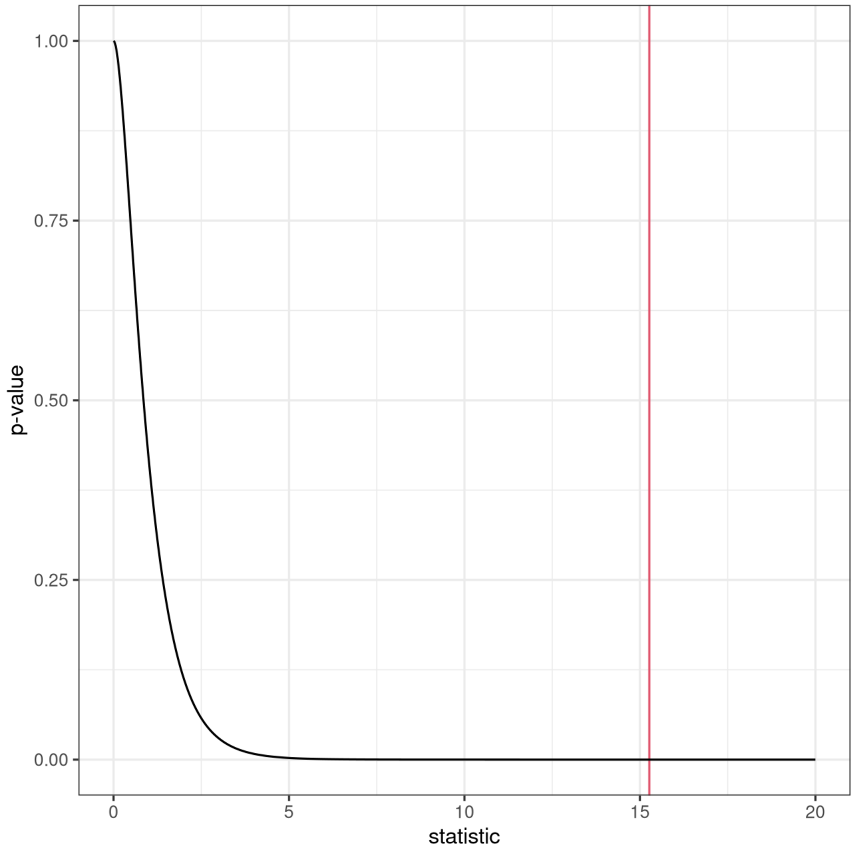


P-value

Null distributions are different, which makes comparisons uneasy.

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

```
pf(stat,
   df1 = 4,
   df2 = 40,
   lower.tail = FALSE)
```



t-tests

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

The larger the values of t (positive or negative), the more evidence against the null hypothesis.

Example

Consider the pairwise average difference in scores between the praised (group C) and the reproved (group D) of the `arithmetic` study.

- Sample averages are $\widehat{\mu}_C = 27.4$ and $\widehat{\mu}_D = 23.4$
- The estimated pooled standard deviation for the five groups is 1.15
- The estimated average difference between groups C and D is $\hat{\delta}_{CD} = 4$.
- The standard error for the difference is $\mathsf{se}(\hat{\delta}_{CD}) = 1.6216$

Example

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

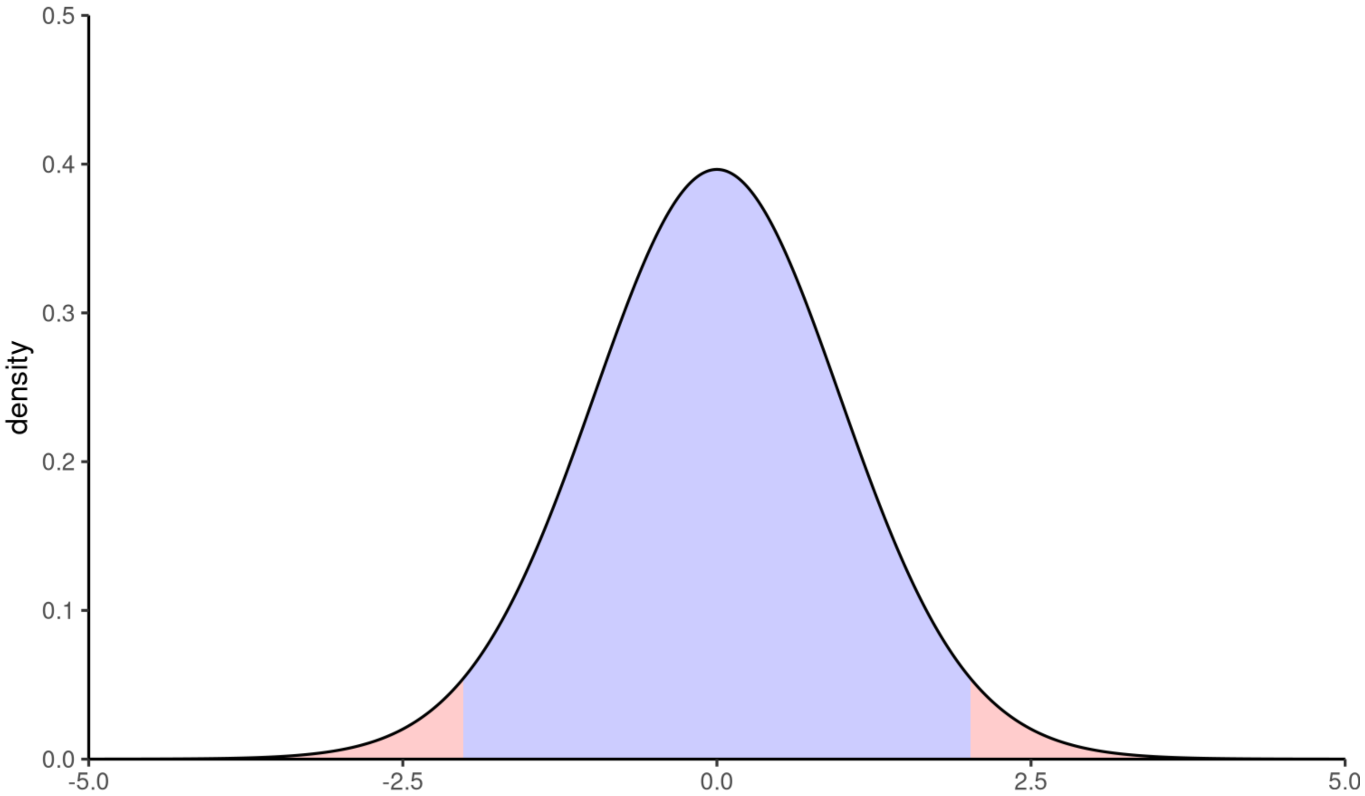
$$t = \frac{\hat{\delta}_{CD} - 0}{\mathsf{se}(\hat{\delta}_{CD})} = \frac{4}{1.6216} = 2.467$$

- 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.

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%.



Critical values

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

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

Because of 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**, `qt(1-alpha/2, df = n - k)` where `n` is the number of observations and `k` the number of groups

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

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

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

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

Interpretation of confidence intervals

The reported confidence interval is

$$[\hat{\delta}_{jk} + \mathsf{se}(\hat{\delta}_{jk})\mathrm{t}_{n-k}(\alpha/2), \hat{\delta}_{jk} + \mathsf{se}(\hat{\delta}_{jk})\mathrm{t}_{n-k}(1-\alpha/2)].$$

Each bound is of the form

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

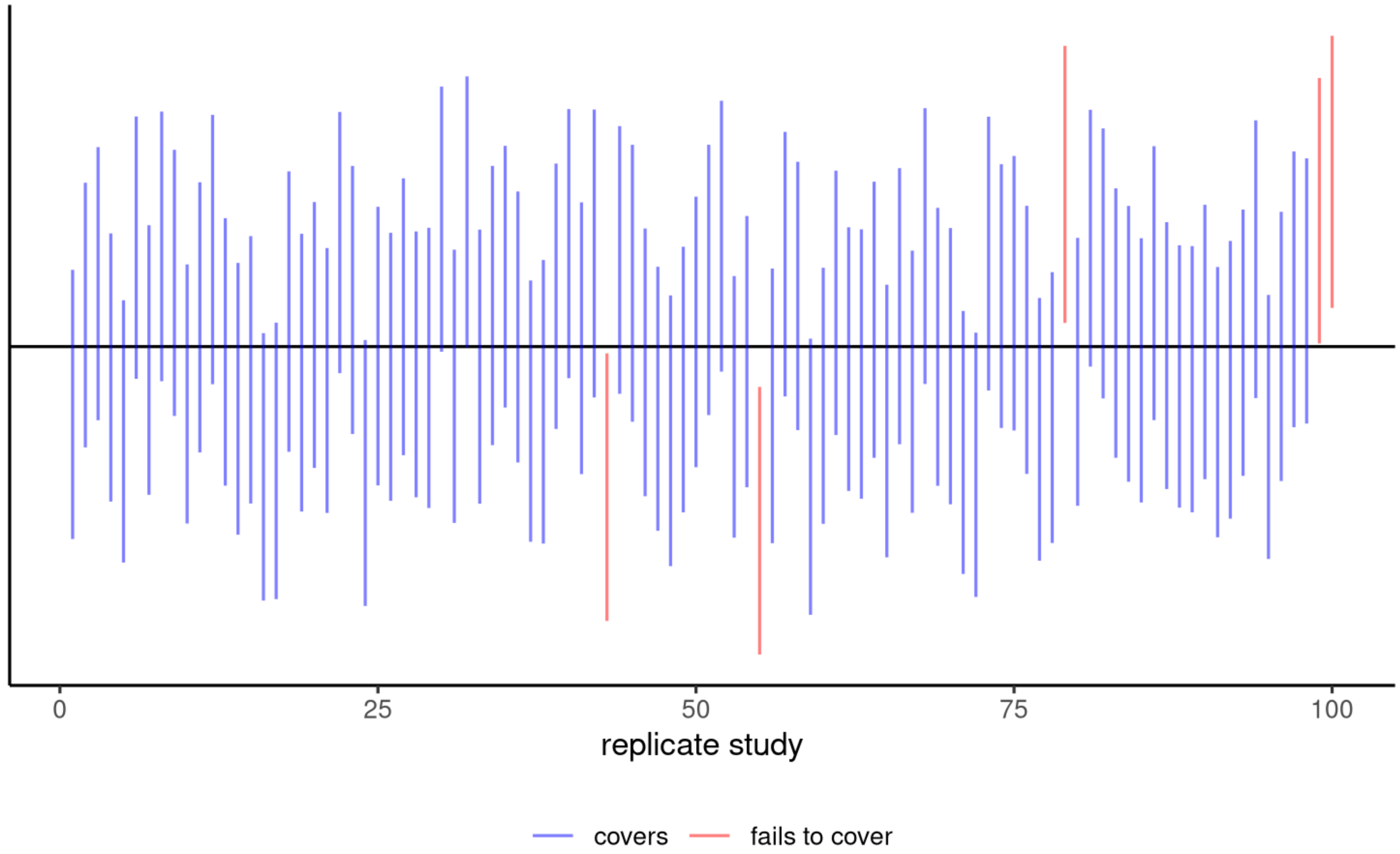
confidence interval = [lower, upper] units

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 $\mathsf{se}(\hat{\delta}_{CD}) = 1.6216$.
- 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

$$[4 - 1.6216 \times 2.021, 4 + 1.6216 \times 2.021] = [0.723, 7.28]$$

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

Pairwise differences in R

```r
library(tidyverse) # data manipulation
library(emmeans) # marginal means and contrasts
url <- "https://edsm.rbind.io/data/arithmetic.csv"
# load data, define column type (factor and integer)
arithmetic <- read_csv(url, col_types = "fi")
# fit one-way ANOVA model
model <- lm(score ~ group, data = arithmetic)
# Compute average of groups with model specification
margmeans <- emmeans::emmeans(model, specs = "group")
# Contrasts (default to pairwise comparisons) - no adjustment
contrast(margmeans, adjust = 'none', infer = TRUE)
#infer = TRUE for confidence intervals
```