Hypothesis testing

Session 2

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

Outline

Variability

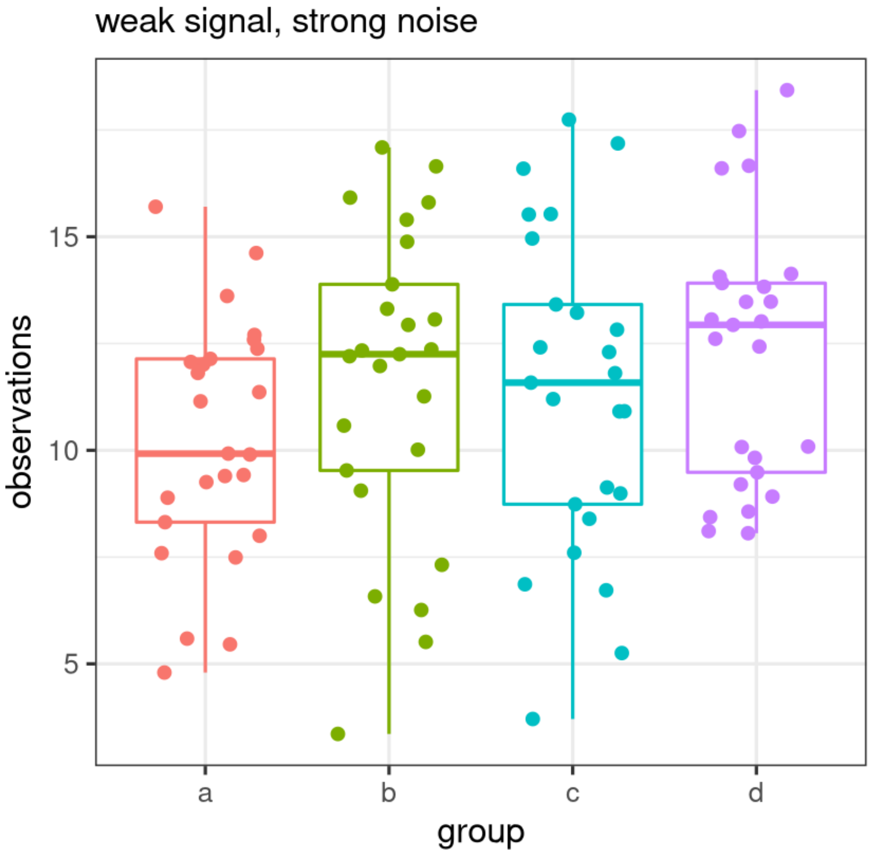
Hypothesis tests

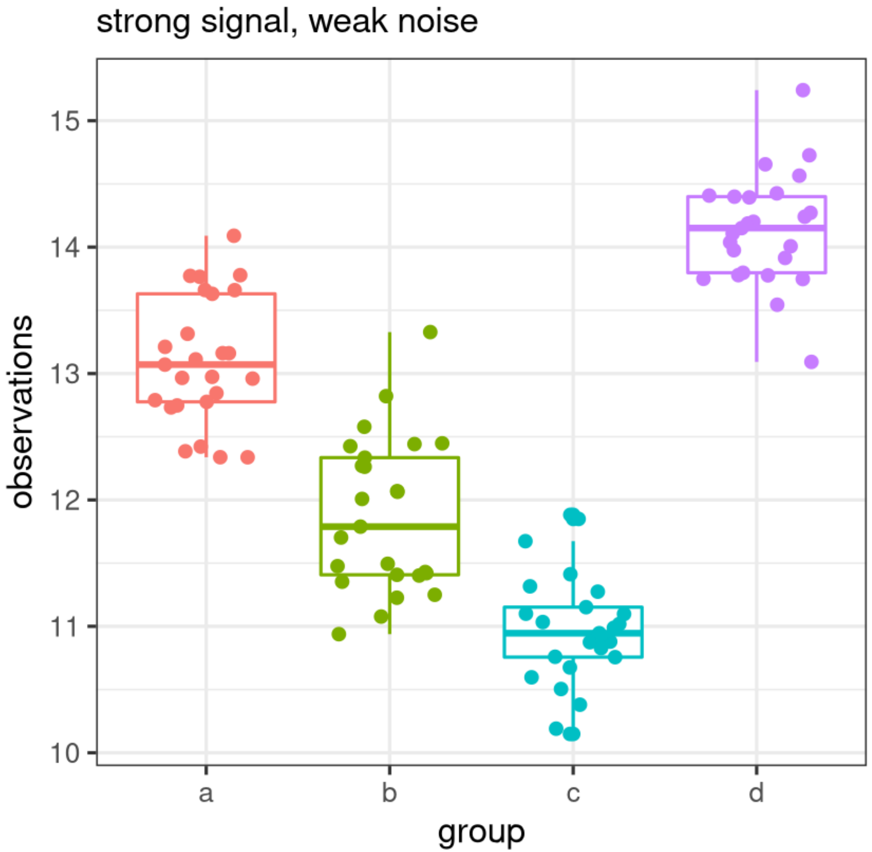
Power

R examples

Signal versus noise

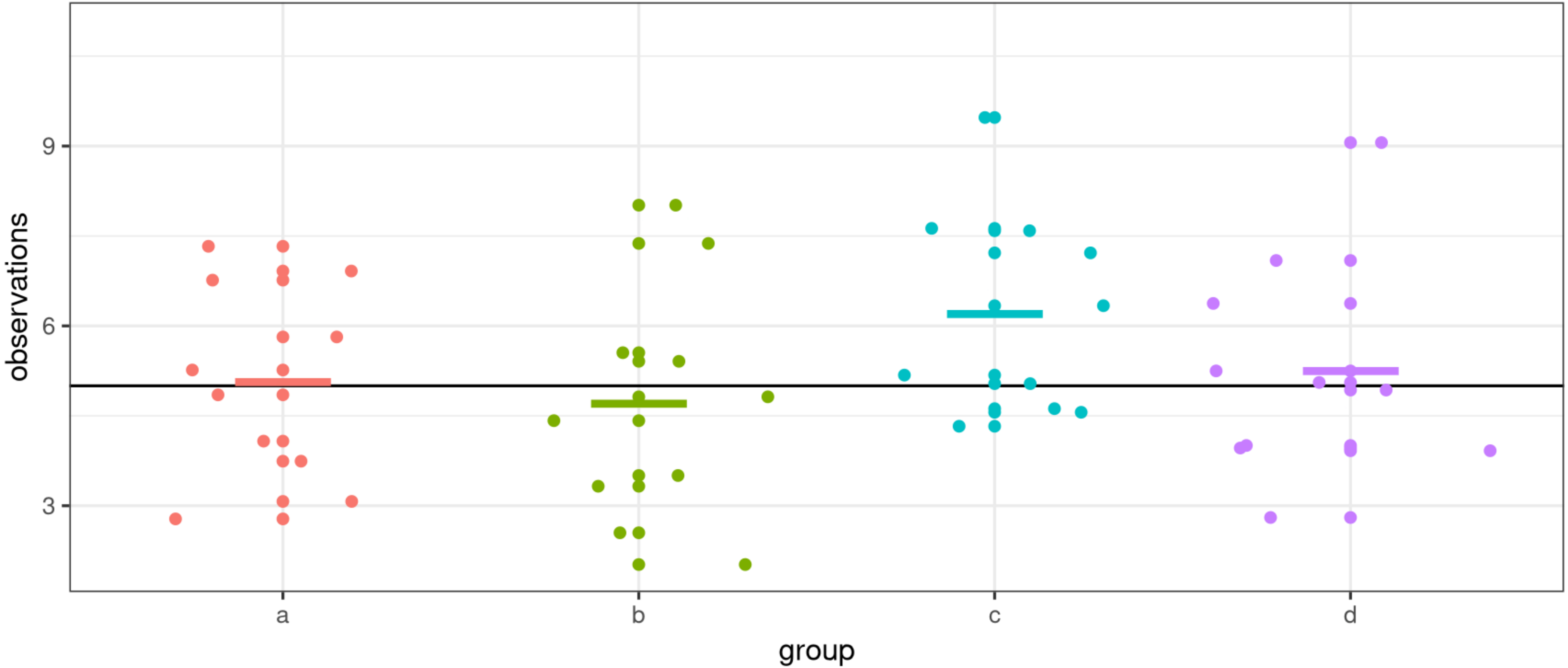
The signal and the noise





Can you spot the differences?

Sampling variability



Hypothesis tests

The general recipe of hypothesis testing

1. Define variables
2. Write down hypotheses ($\mathscr{H}_0$ / $\mathscr{H}_a$)
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.

Exercise

In pairs, identify

- the experimental and observational units
- the treatment levels
- the response variable
- the null and alternative hypothesis

03:00

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…
```

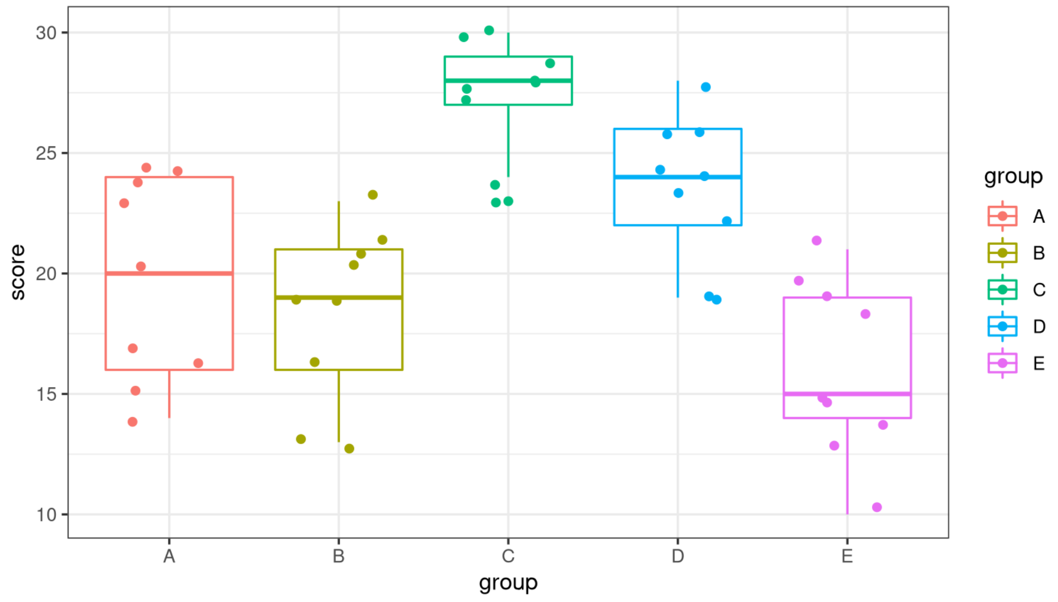
Load data | **Summary statistics** | Plot

```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

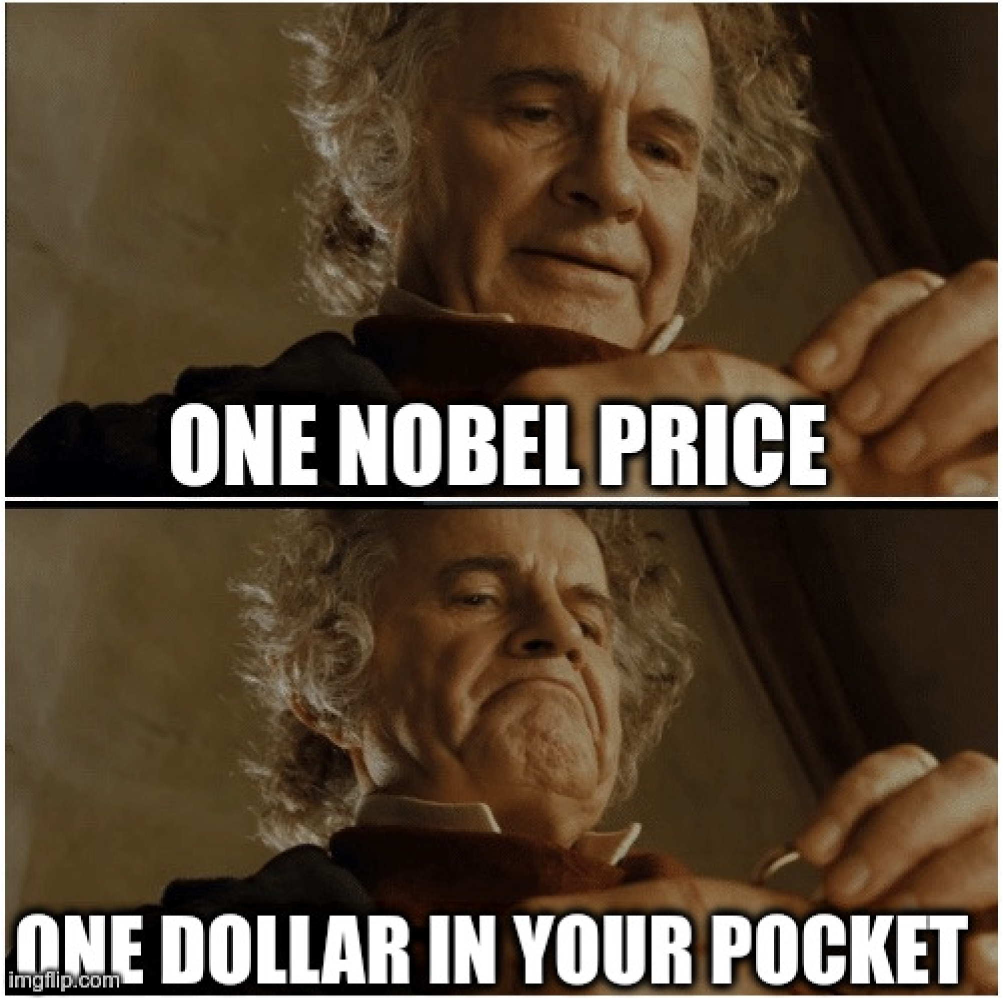
One-way analysis of variance uses an F statistic.

```
#one way analysis of variance
aov(data = arithmetic,
         formula = y ~ group)
# can replace 'aov' by 'lm'
```

- 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



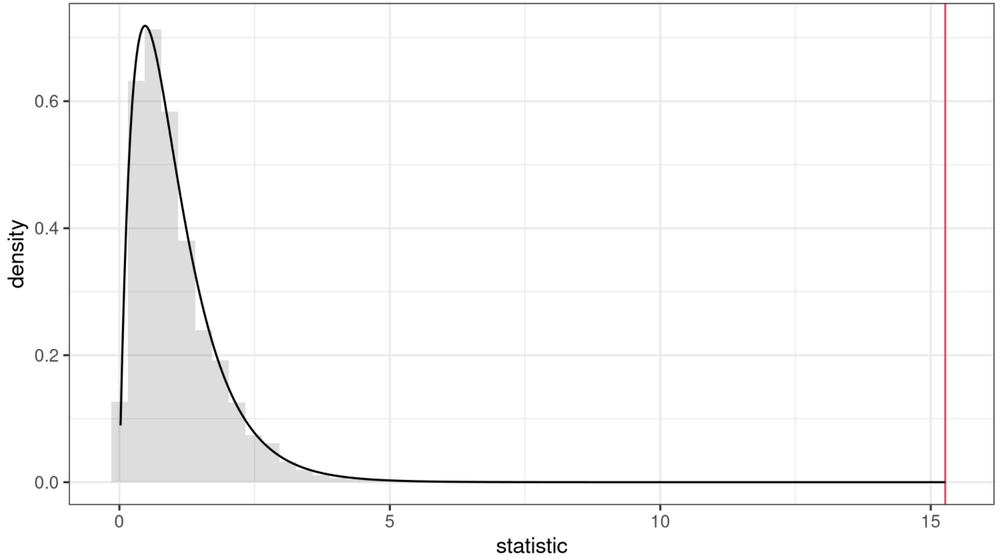
Benchmarking

- 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.

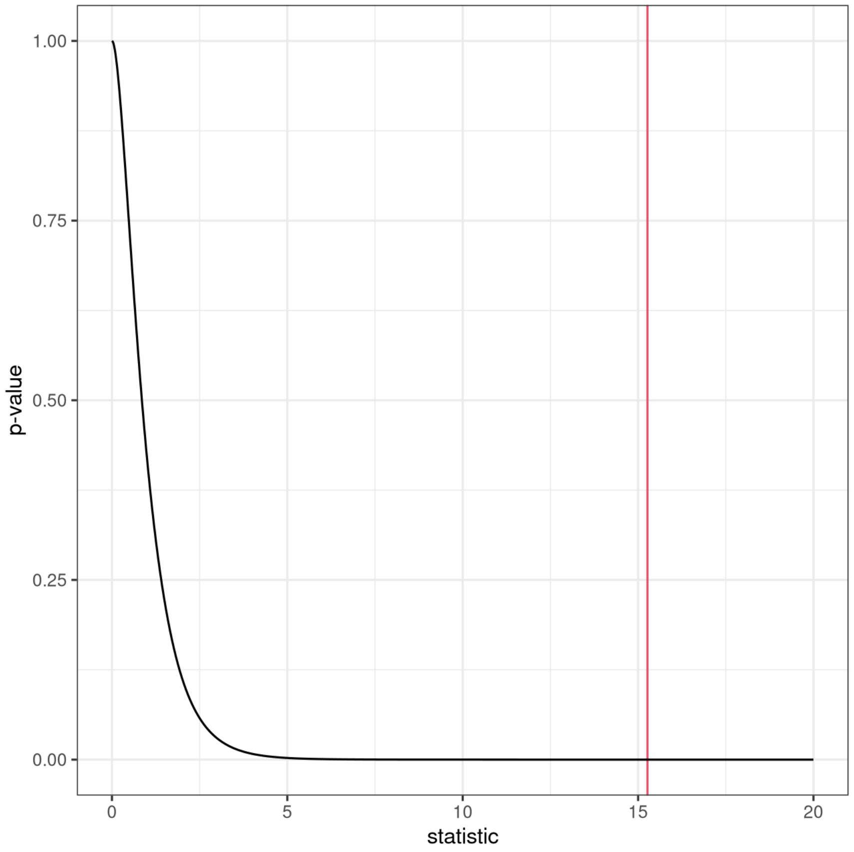


P-value

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)
```



Power

I cried power!

The null alternative corresponds to a single value: $\mu_a = \cdots = \mu_e$.

whereas there are infinitely many alternatives...

Power is the ability to detect when the null is false, for a given alternative.