

# One way ANOVA

## Session 3

MATH 80667A: Experimental Design and Statistical Methods  
HEC Montréal

# Outline

**Hypothesis tests for ANOVA**

**Model assumptions**

# F-test for one way ANOVA

## Global null hypothesis

No difference between treatments

- $H_0$  (null): all of the  $K$  treatment groups have the same average  $\mu$
- $H_a$  (alternative): at least two treatments have different averages

Tacitly assume that all observations have the same standard deviation  $\sigma$ .

# Building a statistic

- $y_{ik}$  is observation  $i$  of group  $k$
- $\hat{\mu}_1, \dots, \hat{\mu}_K$  are sample averages of groups  $1, \dots, K$
- $\hat{\mu}$  is the overall sample mean

## Decomposing variability into bits

$$\sum_i \sum_k (y_{ik} - \hat{\mu})^2 \text{total sum of squares} = \sum_i \sum_k (y_{ik} - \hat{\mu}_k)^2 \text{within sum of squares} + \sum_k n_k (\hat{\mu}_k - \hat{\mu})^2$$

null model

alternative model

added variability

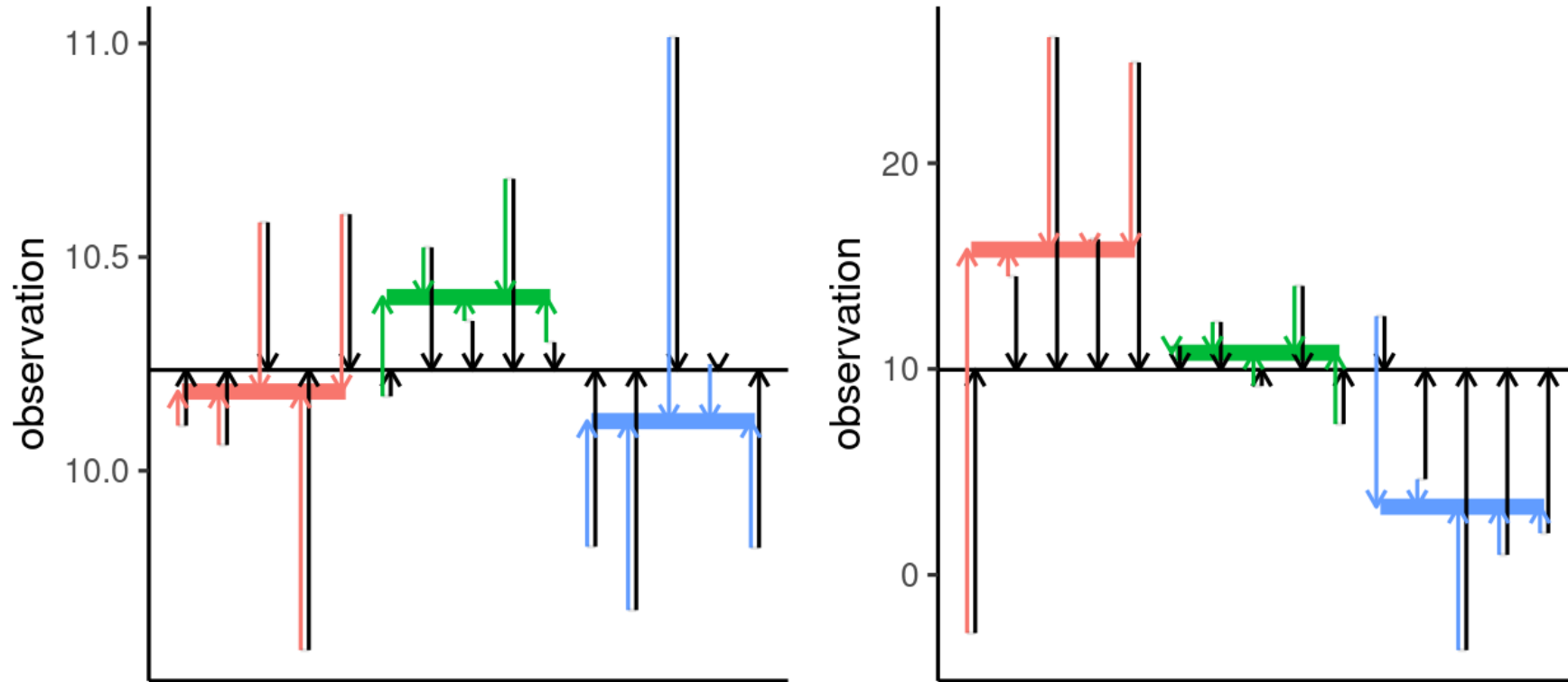
# F-test statistic

## Omnibus test

With  $K$  groups and  $n$  observations, the statistic is

$$\begin{aligned} F &= \frac{\text{between-group variability}}{\text{within-group variability}} \\ &= \frac{\text{between sum of squares} / (K - 1)}{\text{within sum of squares} / (n - K)} \end{aligned}$$

# Ratio of variance



Data with equal mean (left) and different mean per group (right).

# What happens under the null regime?

If all groups have the same mean, both numerator and denominator are estimators of  $\sigma^2$ , thus

- the  $F$  ratio should be 1 on average if there are no mean differences.
- but the numerator is more variable because it is based on  $K$  observations
  - benchmark is skewed to the right.

# Null distribution and degrees of freedom

The null distribution (benchmark) is a Fisher distribution  $F(\nu_1, \nu_2)$ .

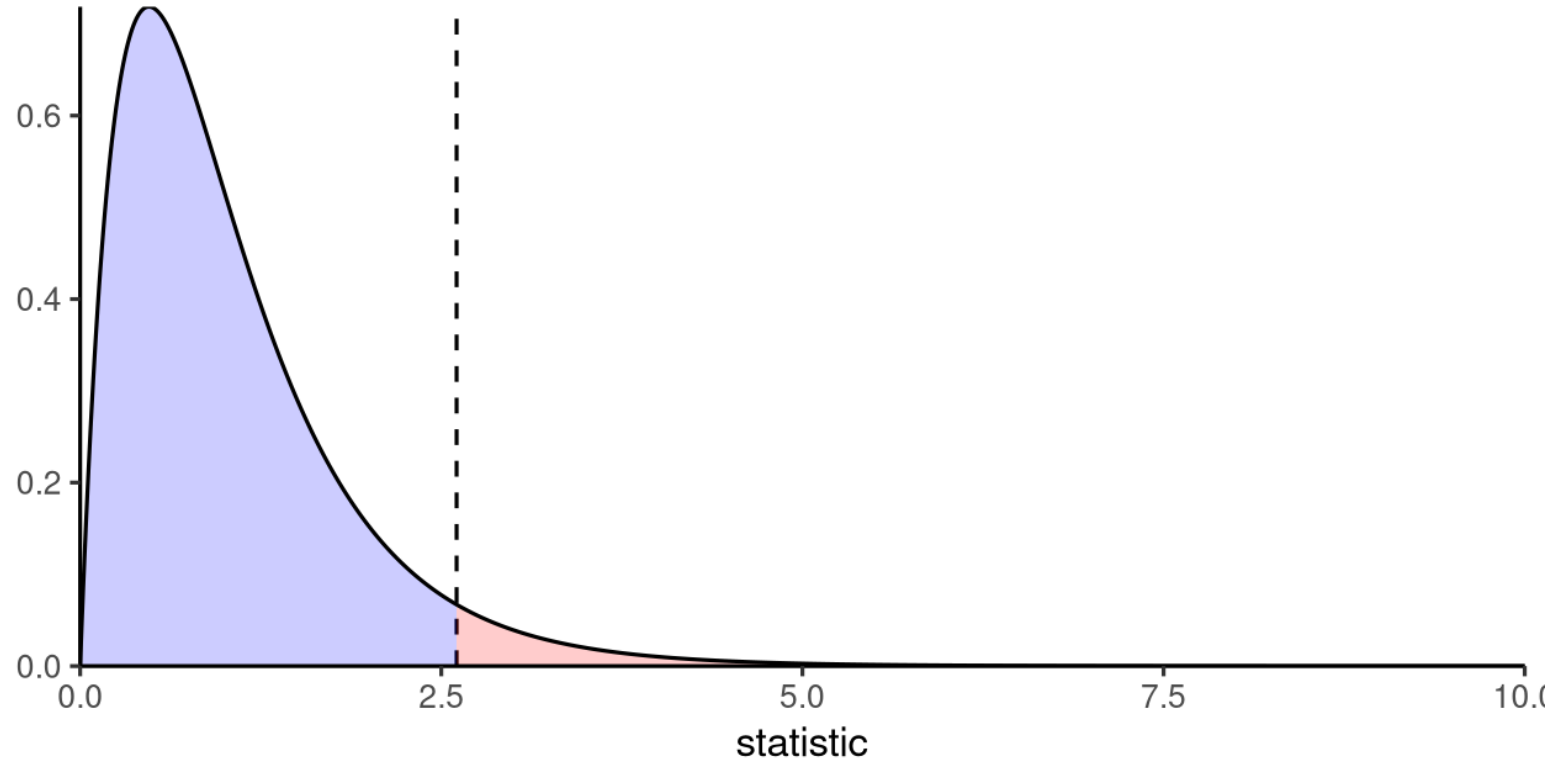
The parameters  $\nu_1, \nu_2$  are called **degrees of freedom**.

For the one-way ANOVA:

- $\nu_1 = K - 1$  is the number of constraints imposed by the null (number of groups minus one)
- $\nu_2 = n - K$  is the number of observations minus number of mean parameters estimated under alternative



# Fisher distribution



Note: the  $F(\nu_1, \nu_2)$  distribution is indistinguishable from  $\chi^2(\nu_1)$  for  $\nu_2$  large.

# 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.

# Formulating an hypothesis

Let  $\mu_A, \dots, \mu_E$  denote the population average (expectation) score for the test for each experimental condition.

The null hypothesis is

$$H_0: \mu_A = \mu_B = \dots = \mu_E$$

against the alternative  $H_a$  that at least one of the population average is different.

# F statistic

```
#Fit one way analysis of variance  
test <- aov(data = arithmetic,  
            formula = score ~ group)  
anova(test) #print anova table
```

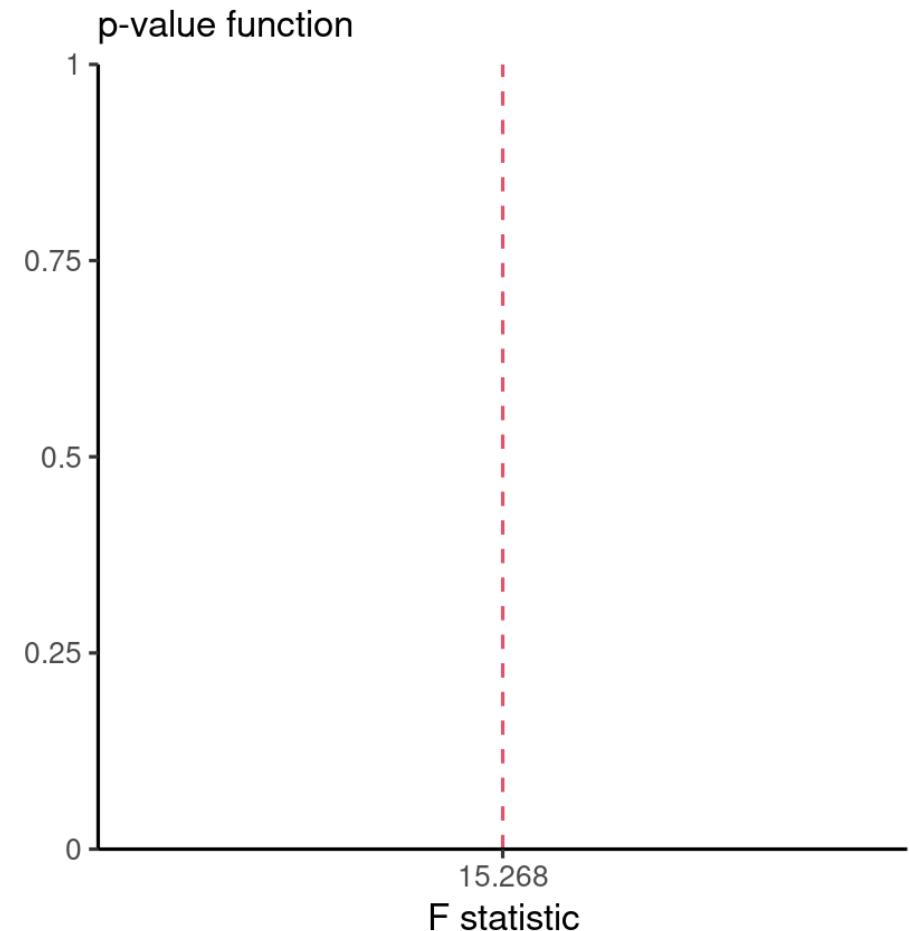
term	df	sum of square	mean square	statistic	p-value
group	4	722.67	180.67	15.27	< 1e-04
Residuals	40	473.33	11.83		

# P-value

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

```
# Compute p-value  
pf(15.27,  
   df1 = 4,  
   df2 = 40,  
   lower.tail = FALSE)
```

Probability that a  $F(4, 40)$  exceeds 15.27.





# Model assumptions

# Quality of approximations

- The null and alternative hypothesis of the analysis of variance only specify the mean of each group
- We need to assume more to derive the behaviour of the statistic

**All statements about  $p$ -values  
are approximate.**

Read the fine print conditions!



# Model assumptions

**Additivity and linearity**

**Equal variance**

**Independence**

**Large sample size**

# Alternative representation

Write  $i$ th observation of  $k$ th experimental group as

$$Y_{ik} \text{ observation} = \mu_k \text{ mean of group} + \varepsilon_{ik} \text{ error term,}$$

where, for  $i = 1, \dots, n_k$  and  $k = 1, \dots, K$ ,

- $E(\varepsilon_{ik}) = 0$  (mean zero) and
- $Va(\varepsilon_{ik}) = \sigma^2$  (equal variance)
- errors are independent from one another.

# # 1: Additivity

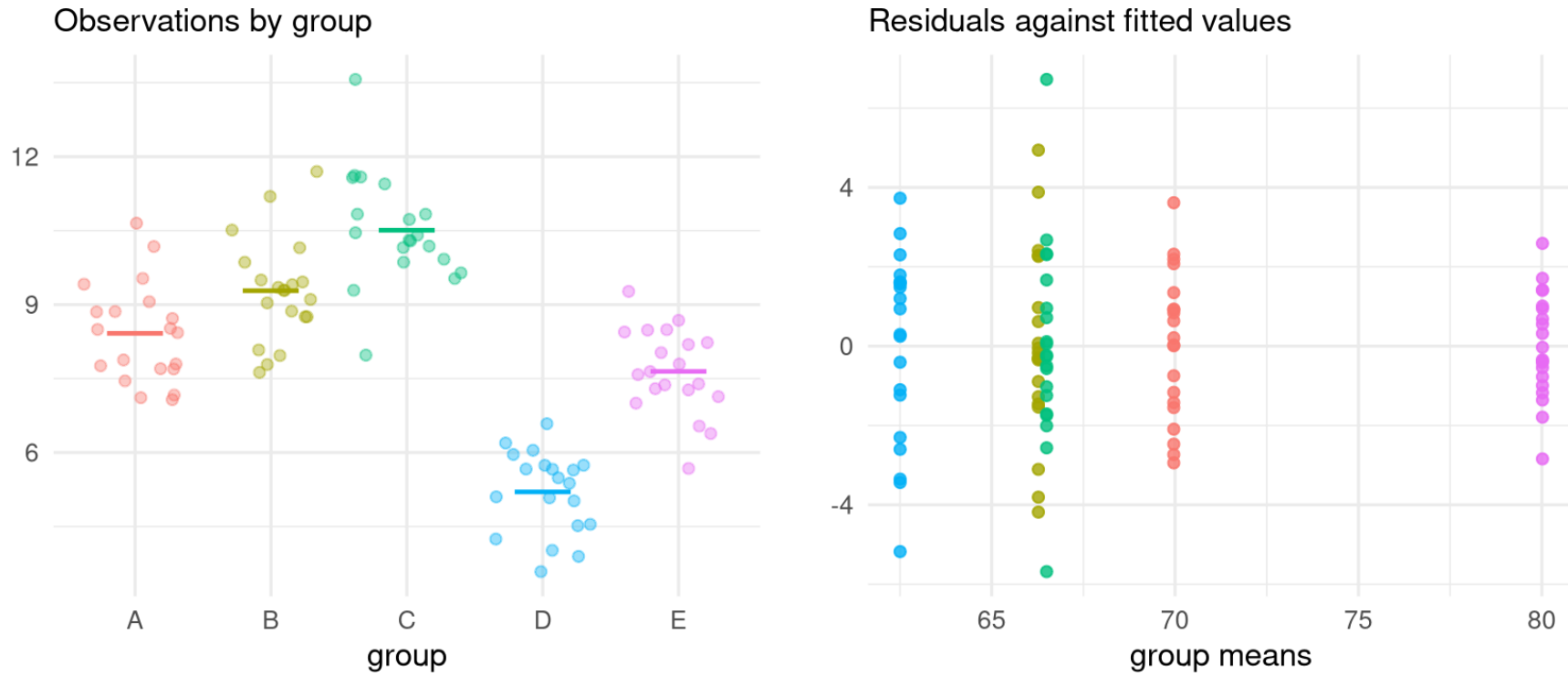
Additive decomposition reads:

$$\left( \begin{array}{c} \text{quantity depending} \\ \text{on the treatment used} \end{array} \right) + \left( \begin{array}{c} \text{quantity depending only} \\ \text{on the particular unit} \end{array} \right)$$

- each unit is unaffected by the treatment of the other units
- average effect of the treatment is constant

# Diagnostic plots for additivity

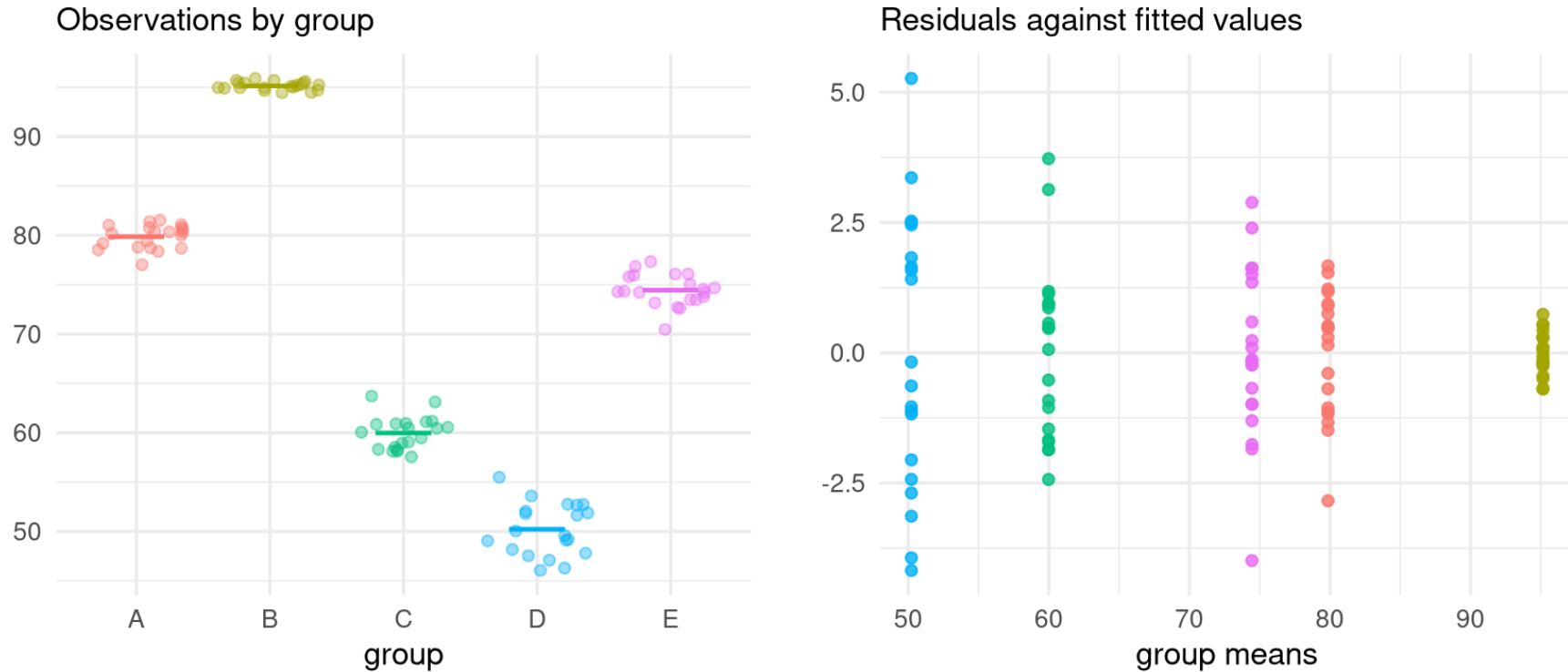
Plot group averages  $\{\hat{\mu}_k\}$  against residuals  $\{e_{ik}\}$ , where  $e_{ik} = y_{ik} - \hat{\mu}_k$ .



By construction, sample mean of  $e_{ik}$  is **always** zero.

# Lack of additivity

Less improvement for scores of stronger students.



Plot and context suggests multiplicative structure. Tempting to diagnose unequal variance.

# Interpretation of residual plots

**Look for potential patterns  
on the  $y$ -axis only!**

# Multiplicative structure

Multiplicative data of the form

$$\begin{pmatrix} \text{quantity depending} \\ \text{on the treatment used} \end{pmatrix} \times \begin{pmatrix} \text{quantity depending only} \\ \text{on the particular unit} \end{pmatrix}$$

tend to have higher variability when the response is larger.

# Fixes for multiplicative data

A log-transformation of response makes the model **additive**.

For responses bounded between  $a$  and  $b$ , reduce warping effects via

$$\ln \left\{ \frac{x - a + \delta}{b + \delta - x} \right\}$$

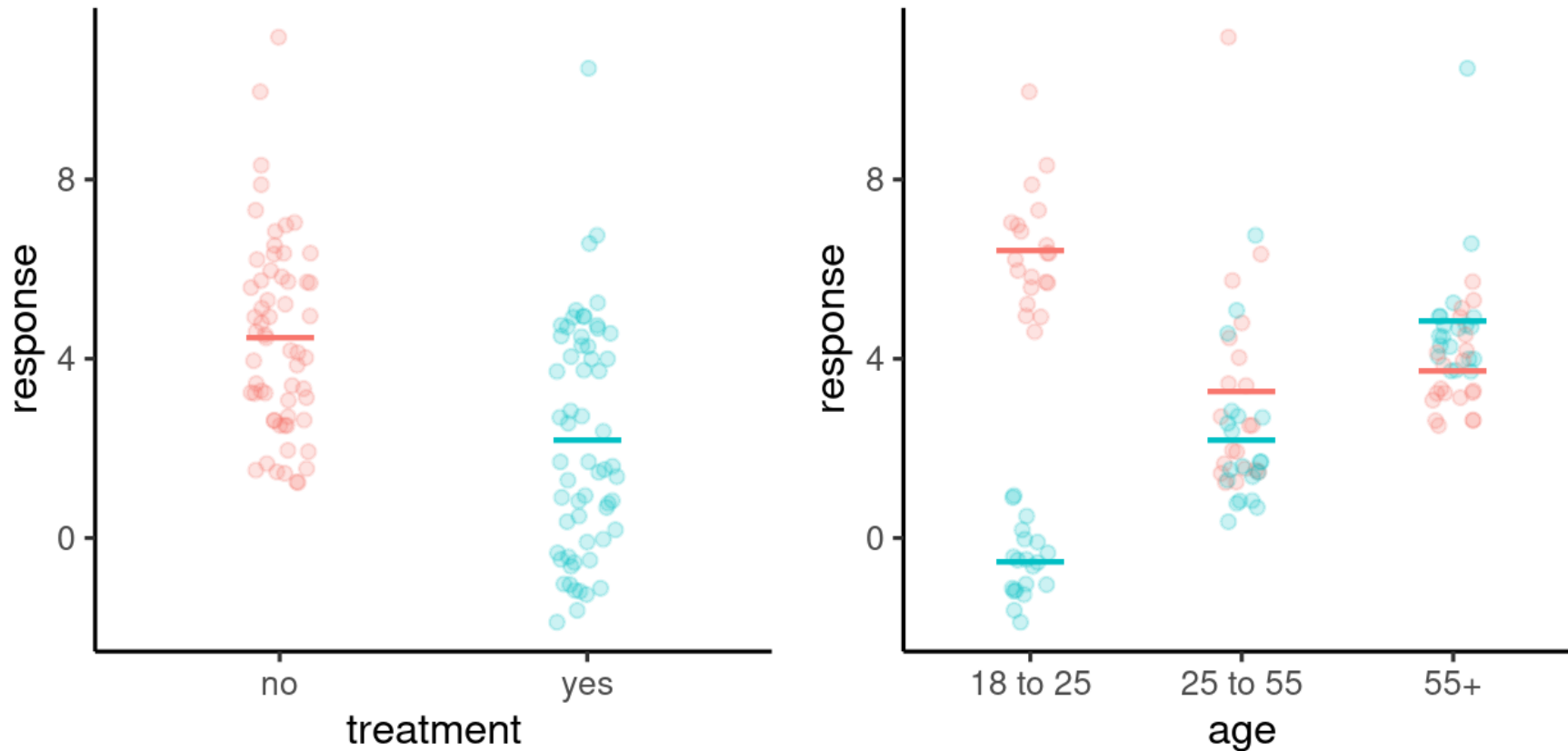
Careful with transformations:

- lose interpretability
- change of meaning (different scale/units).



# Interactions

Plot residuals against other explanatories.



# A note about interactions

An **interaction** occurs when the effect of experimental group depends on another variable.

In principle, randomization ensures we capture the average marginal effect (even if misleading/useless).

We could incorporate the interacting variable in the model capture it's effect (makes model more complex), e.g. using a two-way ANOVA.

## # 2: Equal variance

**Each observation  
has the *same*  
standard deviation  $\sigma$ .**

**ANOVA is quite sensitive to this assumption!**

# Graphical diagnostics

Plot *standardized* (`rstandard`) or *studentized residuals* (`rstudent`) against fitted values.

```
data(arithmetic, package = "hecmedsm")
model <- lm(score ~ group, data = arithmetic)
data <- data.frame(
  fitted = fitted(model),
  residuals = rstudent(model))
ggplot(data = data,
  mapping = aes(x = fitted,
                y = residuals)) +
  geom_point()
```

# Test diagnostics

Can use a statistical test for  $H_0: \sigma_1 = \dots = \sigma_K$ .

- Bartlett's test (assumes normal data)
- Levene's test: a one-way ANOVA for  $|y_{ik} - \hat{\mu}_k|$
- Brown–Forsythe test: a one-way ANOVA for  $|y_{ik} - \text{median}_k|$  (**more robust**)
- Fligner-Killeen test: based on ranks

**Different tests may yield different conclusions**

# Example in R

```
data(arithmetic, package = "hecedsm")
model <- aov(score ~ group, data = arithmetic)
car::leveneTest(model) #Brown-Forsythe by default
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value Pr(>F)
## group    4  1.2072 0.3228
##           40
```

Fail to reject the null: no evidence of unequal variance

# Box's take

To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!

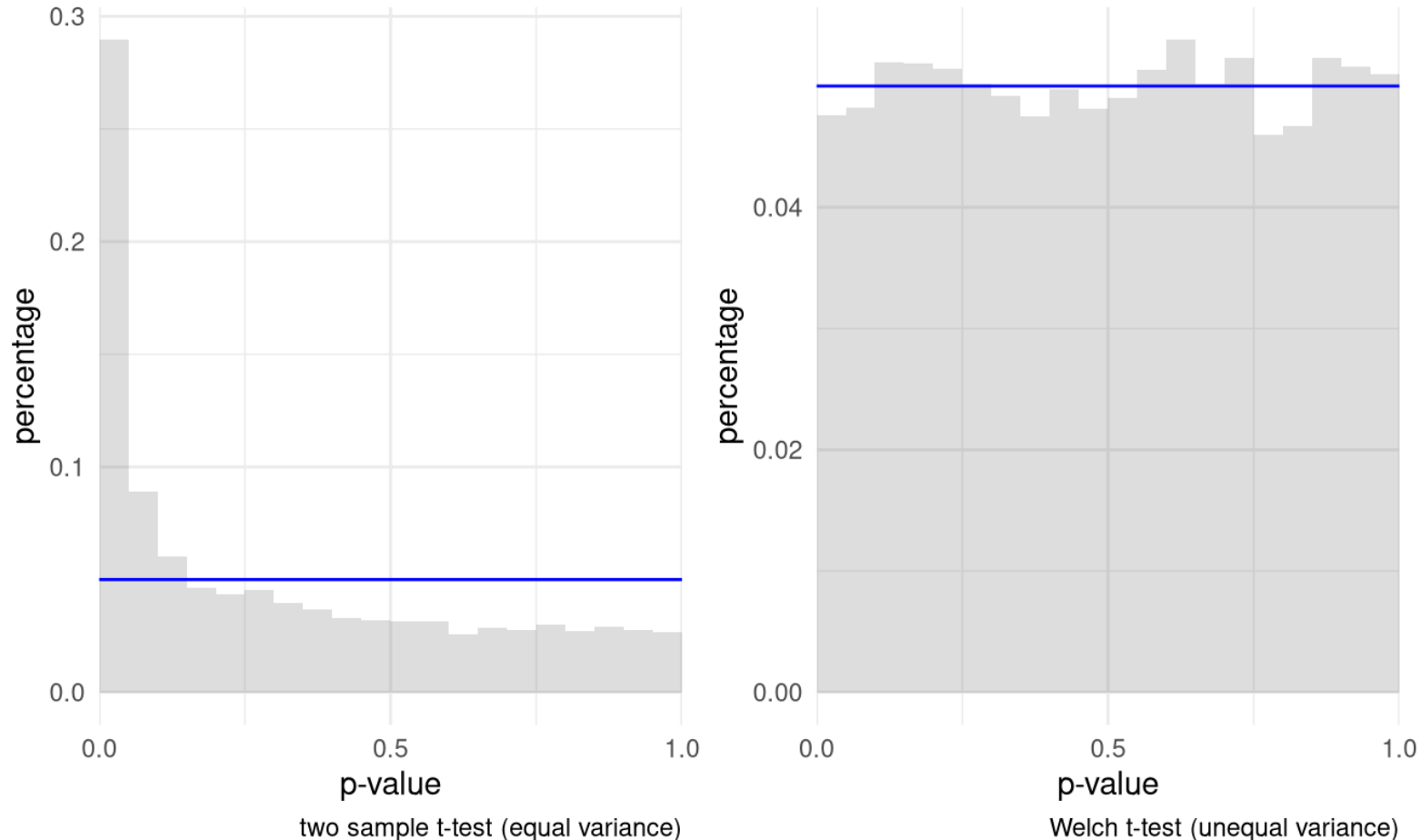
Box, G.E.P. (1953). *Non-Normality and Tests on Variances*. Biometrika 40 (**3**)-4: 318–335.

# Solutions

- In large sample, power is large so probably always reject  $H_0: \sigma_1 = \dots = \sigma_K$ .
- If heterogeneity only per experimental condition, use **Welch's ANOVA** (`oneway.test` in **R**).
- This statistic estimates the std. deviation of each group *separately*.
- Could (should?) be the default when you have large number of observations, or enough to reliably estimate mean and std. deviation.



# What can go wrong? Spurious findings!



Reject null hypothesis more often even if no difference in mean!

# More complex heterogeneity patterns

- Variance-stabilizing transformations (e.g., log for counts)
- Explicit model for trend over time, etc. may be necessary in more complex design (repeated measure) where there is a learning effect.

# # 3: Independence

**No visual diagnostic or test available.**

**Rather, infer from context.**

As a Quebecer, I am not allowed to talk about this topic.

# Checking independence

- Repeated measures are **not independent**
- Group structure (e.g., people performing experiment together and exchanging feedback)
- Time dependence (time series, longitudinal data).
- Dependence on instrumentation, experimenter, time of the day, etc.

Observations close by tend to be more alike (correlated).

# # 4: Sample size (normality?)

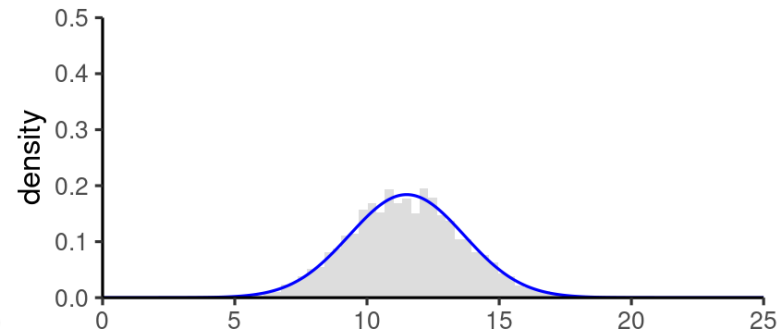
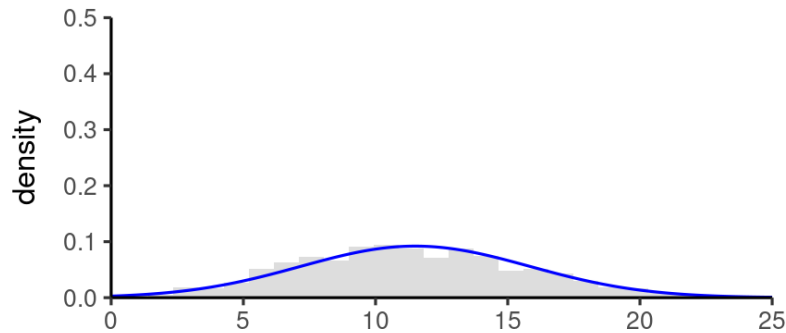
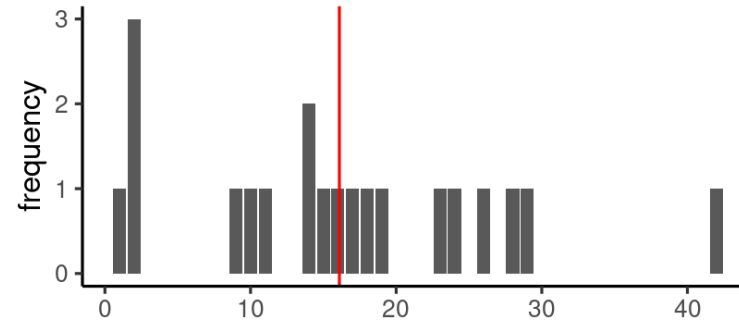
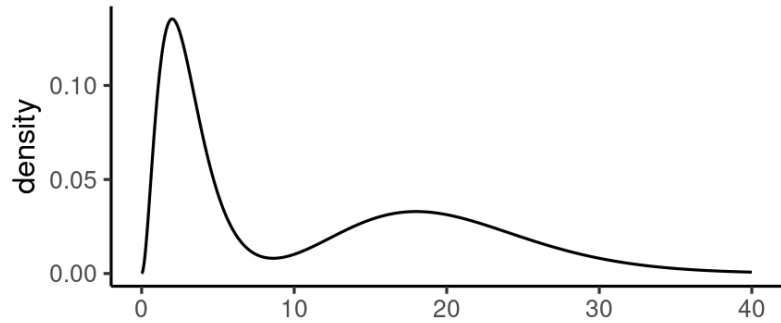
**Where does the F-distribution come from?**

**Normality of group average**

**This holds (in great generality)  
because of the  
central limit theorem**

# Central limit theorem

In large samples, the mean is approximately normally distributed.



# How large should my sample be?

**Rule of thumb: 20 or 30 per group**

**Gather sufficient number of observations.**

# Assessing approximate normality

The closer data are to being normal, the better the large-sample distribution approximation is.

Can check normality via quantile-quantile plot with standardized residuals  $r_i$ :

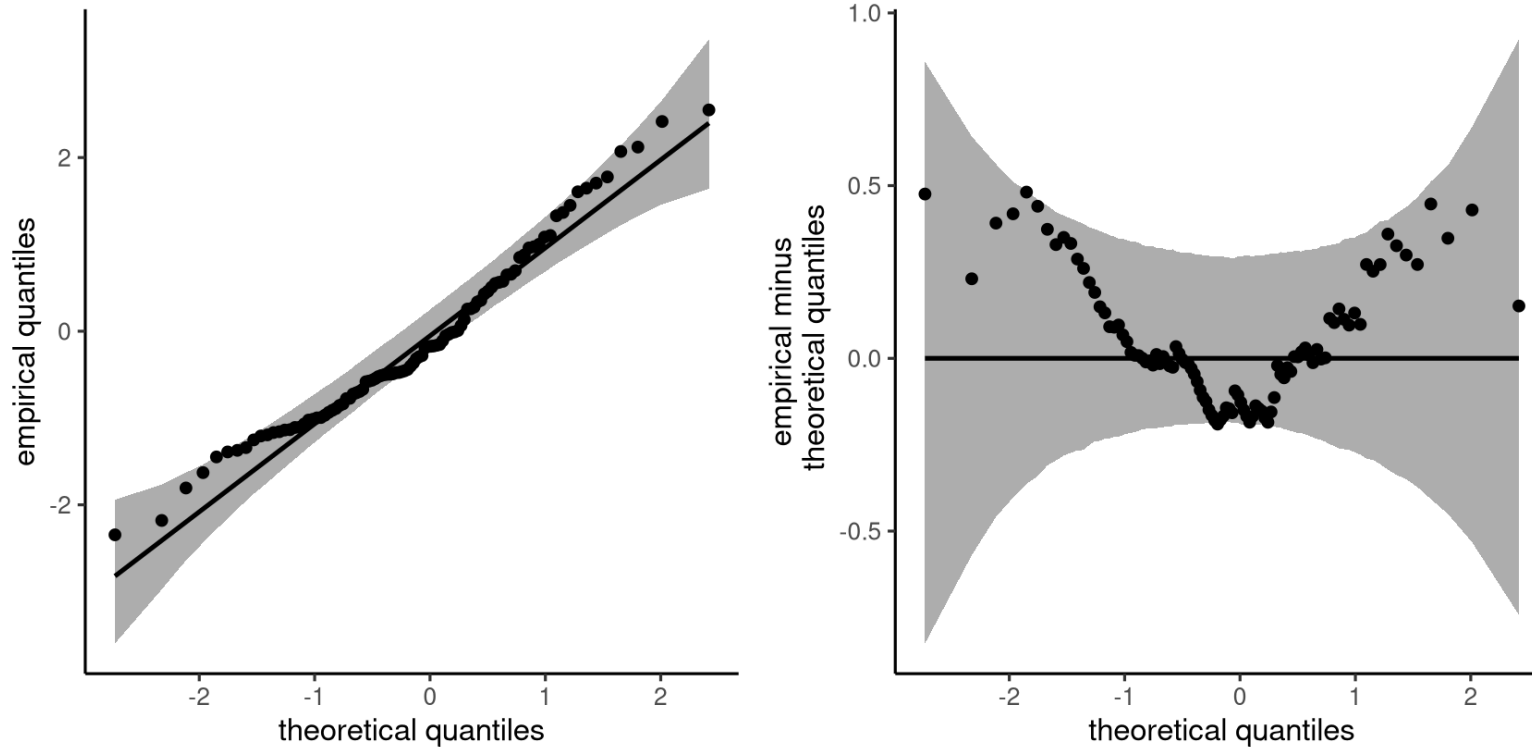
- on the  $x$ -axis, the theoretical quantiles  $\hat{F}^{-1}\{\text{rank}(r_i)/(n+1)\}$  of the residuals, where  $F^{-1}$  is the normal quantile function.
- on the  $y$ -axis, the empirical quantiles  $r_i$

In **R**, use functions `qqnorm` or `car::qqPlot` to produce the plots.



# More about quantile-quantile plots

The ordered residuals should align on a straight line.



Normal quantile-quantile plot (left) and Tukey's mean different QQ-plot (right).

# Recap 1

- One-way analysis of variance compares **average** of experimental groups
- Null hypothesis: all groups have the same average
- Easier to detect when the null hypothesis is false if:
  - large differences group average
  - small variability
  - large samples

# Recap 2

- Model assumes independent observations, additive structure and equal variability in each group.
- All statements are approximate, but if model assumptions are invalid, can lead to spurious results or lower power.