One way ANOVA

Session 3

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

Outline

Hypothesis tests for ANOVA

Power

Model assumptions

F-test for one way ANOVA

Global null hypothesis

No difference between treatments

- $\mathscr{H}_0$ (null): all of the K treatment groups have the same average μ
- $\mathscr{H}_a$ (alternative): at least two treatments have different averages

Tacitly assume that all observations have the same standard deviation σ.

Building a statistic

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

Decomposing variability into bits

$$\underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu})^2}_{\text{total sum of squares}} = \underbrace{\sum_i \sum_k (y_{ik} - \widehat{\mu}_k)^2}_{\text{within sum of squares}} + \underbrace{\sum_k n_i (\widehat{\mu}_k - \widehat{\mu})^2}_{\text{between sum of squares}}.$$

null model **alternative model** **added variability**

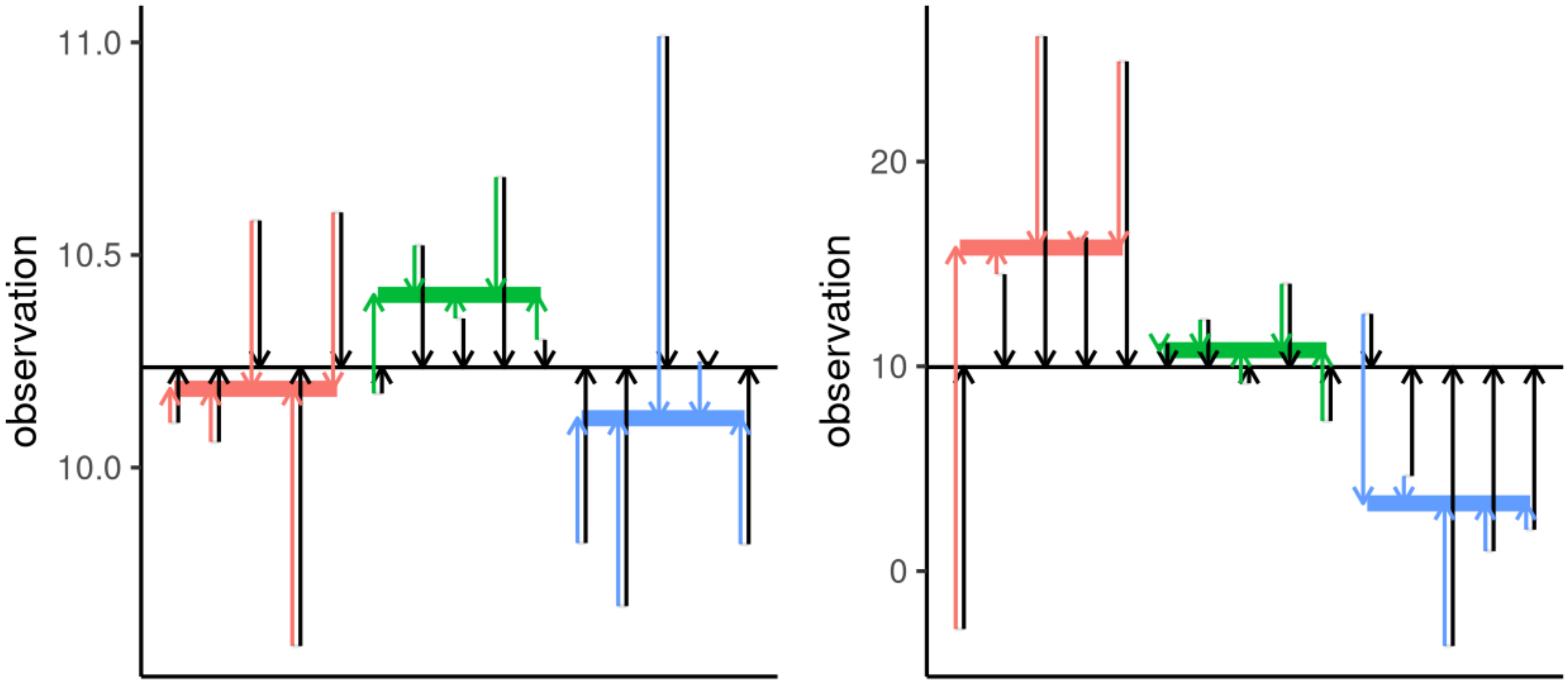
F-test statistic

Omnibus test

With K groups and n observations, the statistic is

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

Ratio of variance



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

Null distribution and degrees of freedom

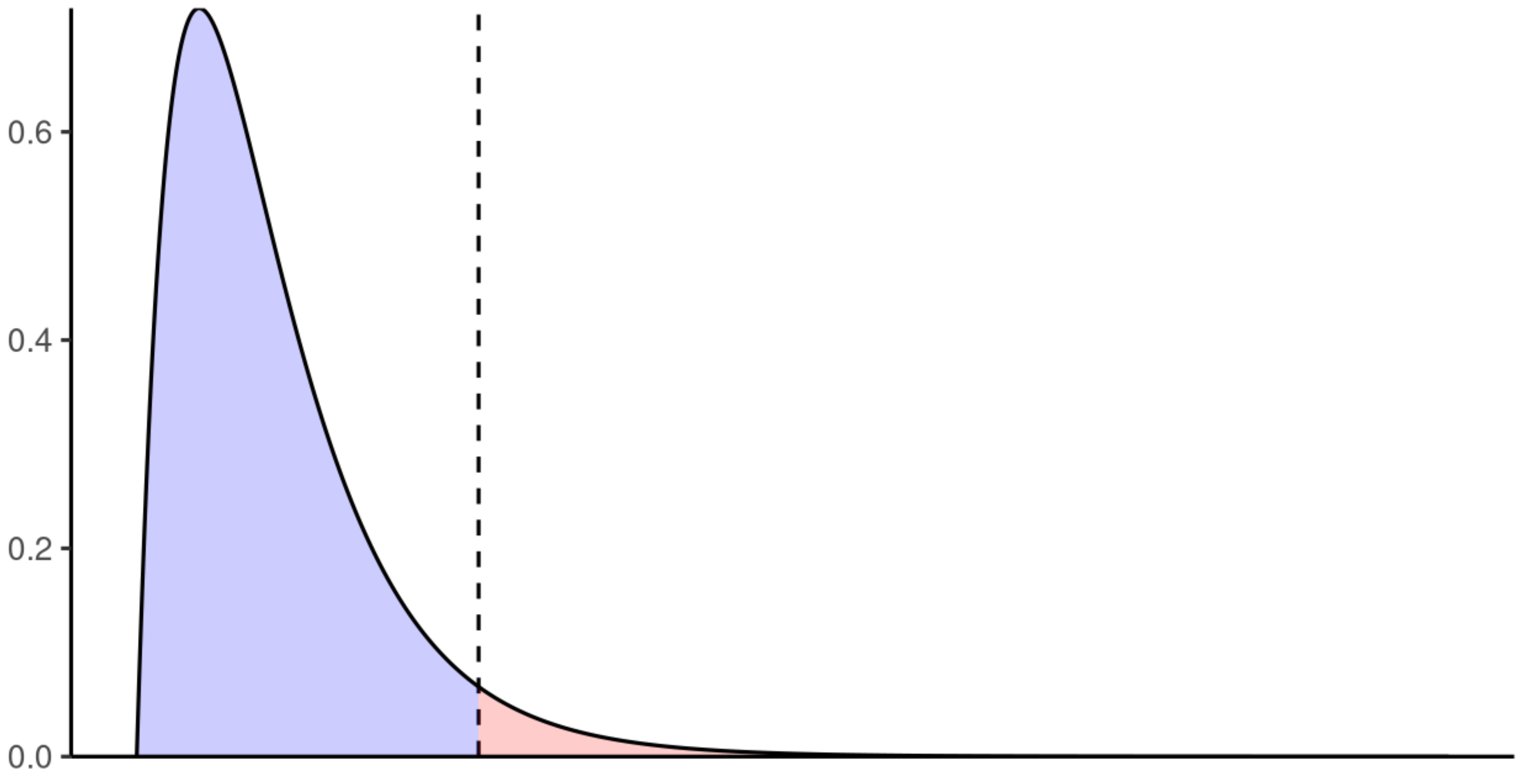
The null distribution (benchmark) is a Fisher distribution $\mathsf{F}(\nu_1, \nu_2)$.

The parameters ν_1, ν_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 $\mathsf{F}(\nu_1, \nu_2)$ distribution is indistinguishable from $\chi^2(\nu_1)$ for ν_2 large.

Intuition behind F-distribution

If all groups have the same mean, both numerator and denominator are estimators of σ^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.

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

The null hypothesis is

$$\mathscr{H}_0 : \mu_A = \mu_B = \cdots = \mu_E$$

against the alternative $\mathscr{H}_a$ that at least one of the population average is different.

F statistic

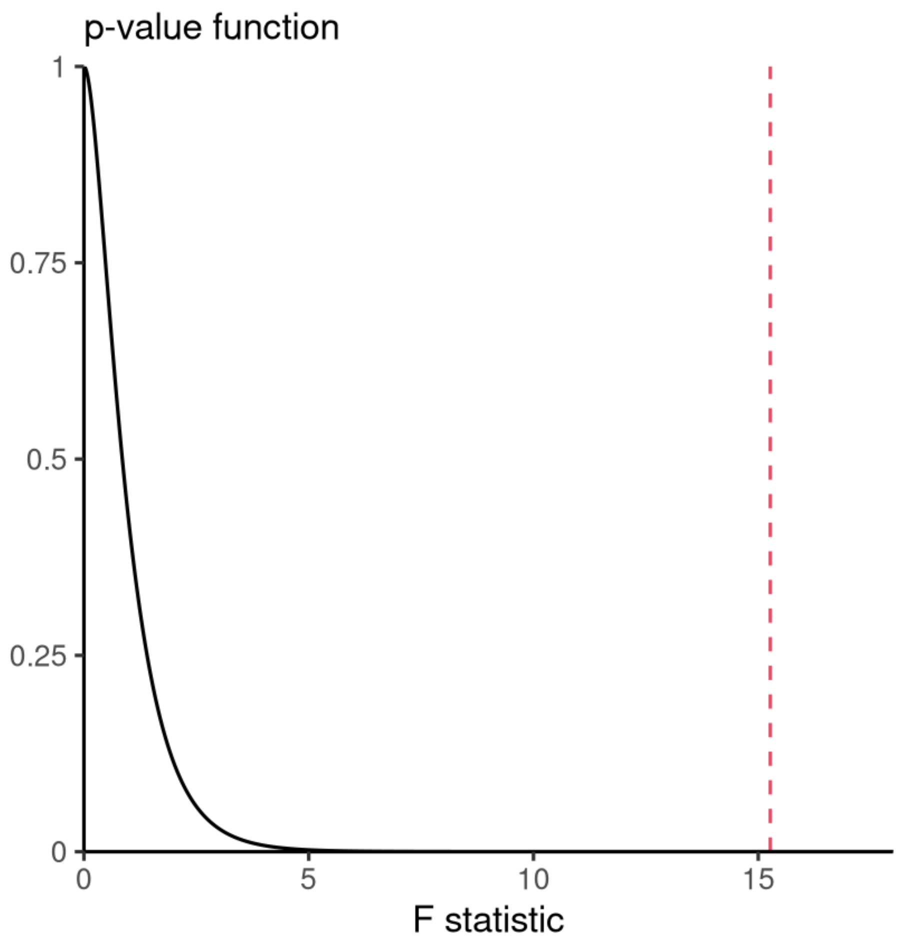
```r
#Fit one way analysis of variance
test <- aov(data = arithmetic,
            formula = score ~ group)
anova(test) #print anova table
```

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**.

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



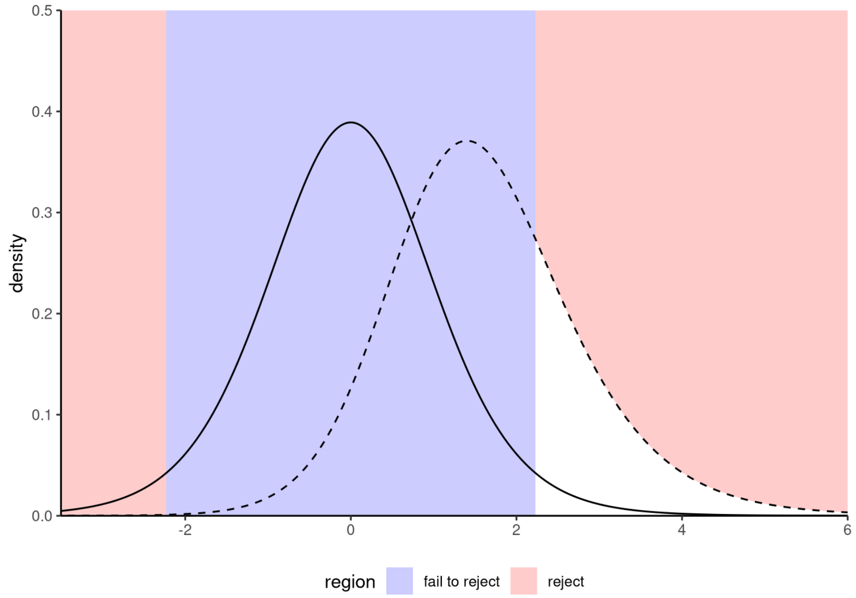
Power

I cried power!

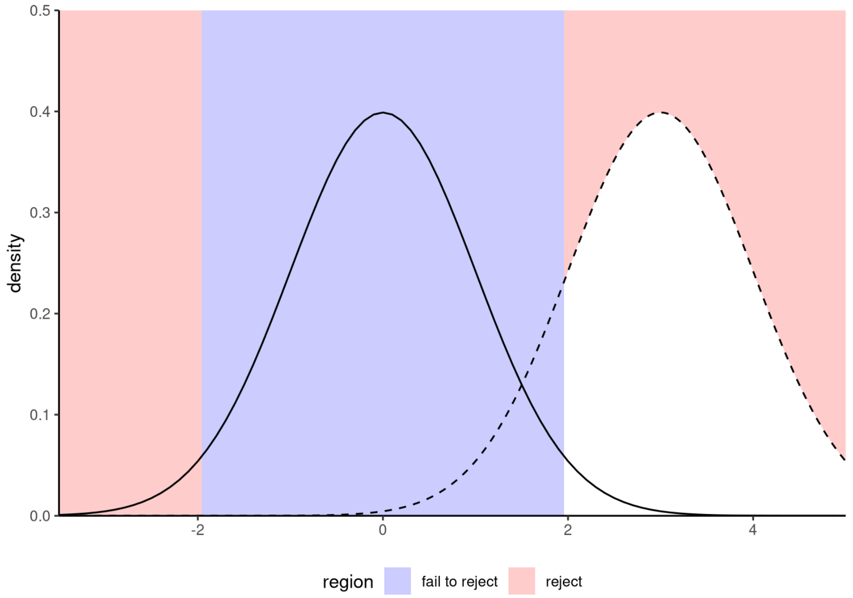
- **Power** is the ability to detect when the null is false, for a given alternative
- It is the *probability* of correctly rejecting the null hypothesis under an alternative.
- The larger the power, the better.

Power of an alternative

There are infinitely many alternatives...



Null distribution (full) and given alternative distribution (dashed).



Power is the area in white under the dashed curved, beyond the cutoff.

Living in an alternative world

How does the F-test behaves under an alternative?

Thinking about power

What do you think is the effect on **power** of an increase of the

- group sample size $n_1, \ldots, n_K$.
- variability σ^2.
- true mean difference $\mu_j - \mu$.

What happens under the alternative?

The peak of the distribution shifts to the right.

Why? on average, the numerator of the F-statistic is

$$\mathsf{E}(\text{between-group variability}) = \sigma^2 + \frac{\sum_{j=1}^{K} n_j(\mu_j - \mu)^2}{K-1}.$$

Under the null hypothesis, $\mu_j = \mu$ for $j = 1, \ldots, K$

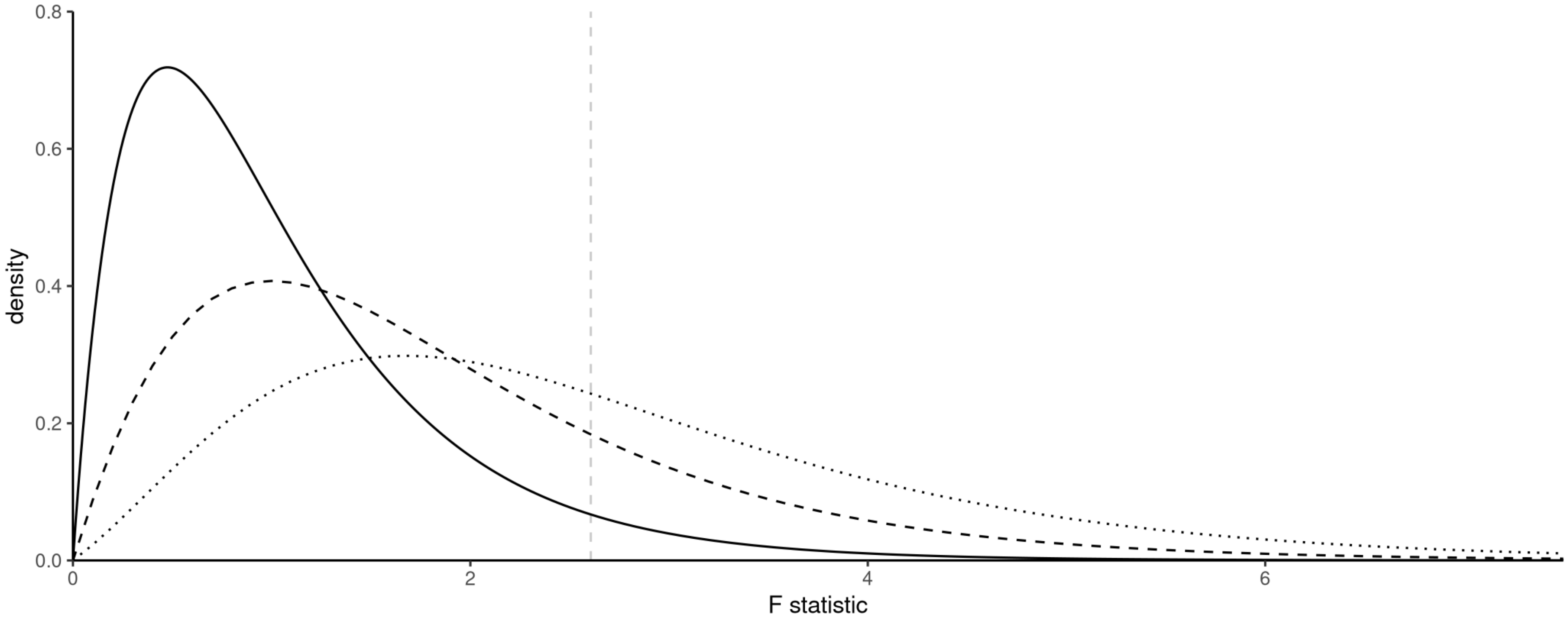
- the rightmost term is 0.

Noncentrality parameter and power

The alternative distribution is $F(\nu_1, \nu_2, \Delta)$ distribution with degrees of freedom ν_1 and ν_2 and noncentrality parameter

$$\Delta = \frac{\sum_{j=1}^{K} n_j(\mu_j - \mu)^2}{\sigma^2}.$$

Impact of noncentrality parameter



F distribution with $\Delta = 0$ (solid line), $\Delta = 3$ (dashed) and $\Delta = 6$ (dotted).

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 ith observation of kth experimental group as

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

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

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

1: Additivity

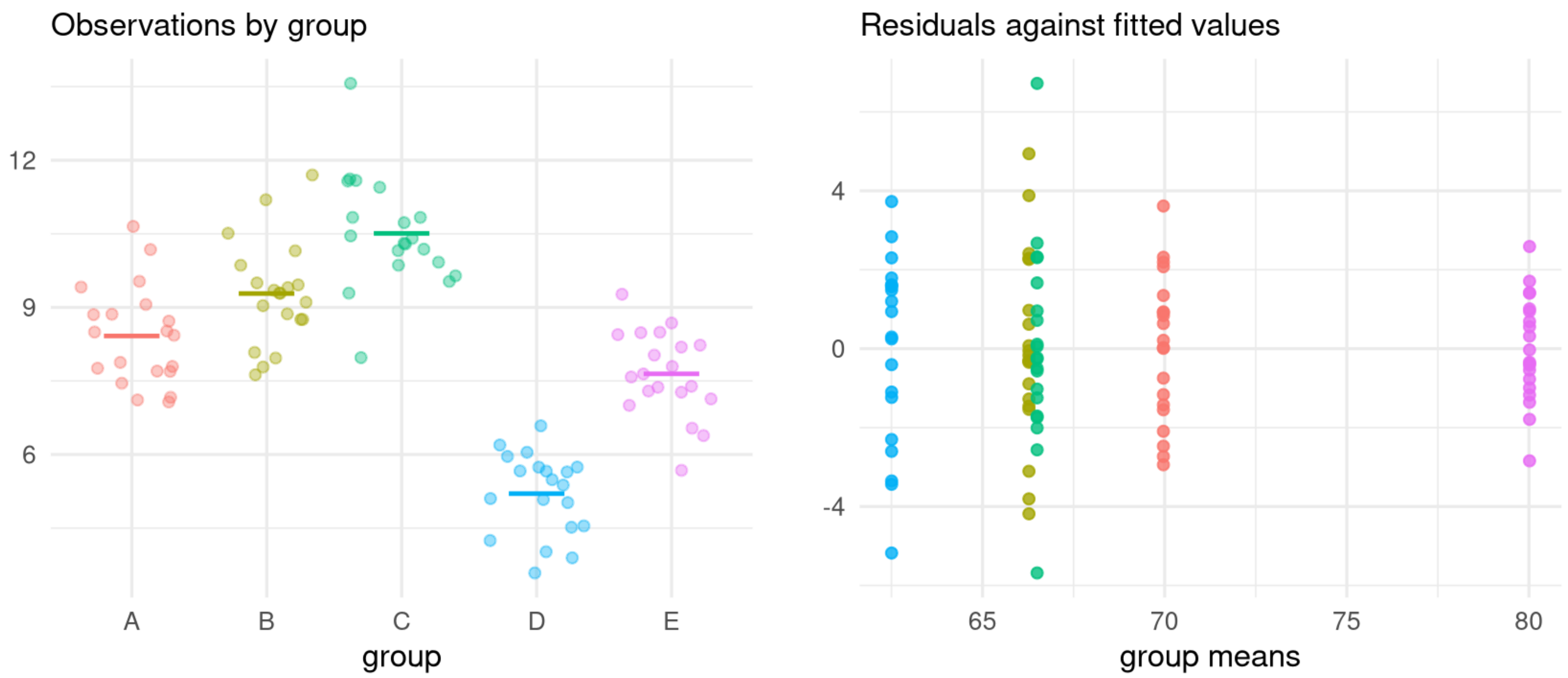
Additive decomposition reads:

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

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

Diagnostic plots for additivity

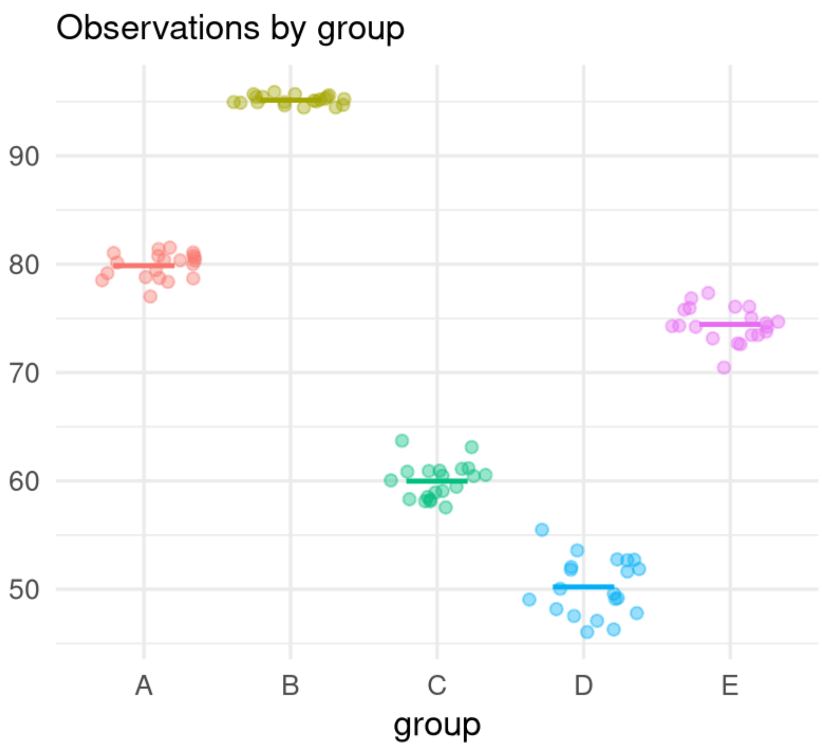
Plot group averages $\{\widehat{\mu}_k\}$ against residuals $\{e_{ik}\}$, where $e_{ik} = y_{ik} - \widehat{\mu}_k$.

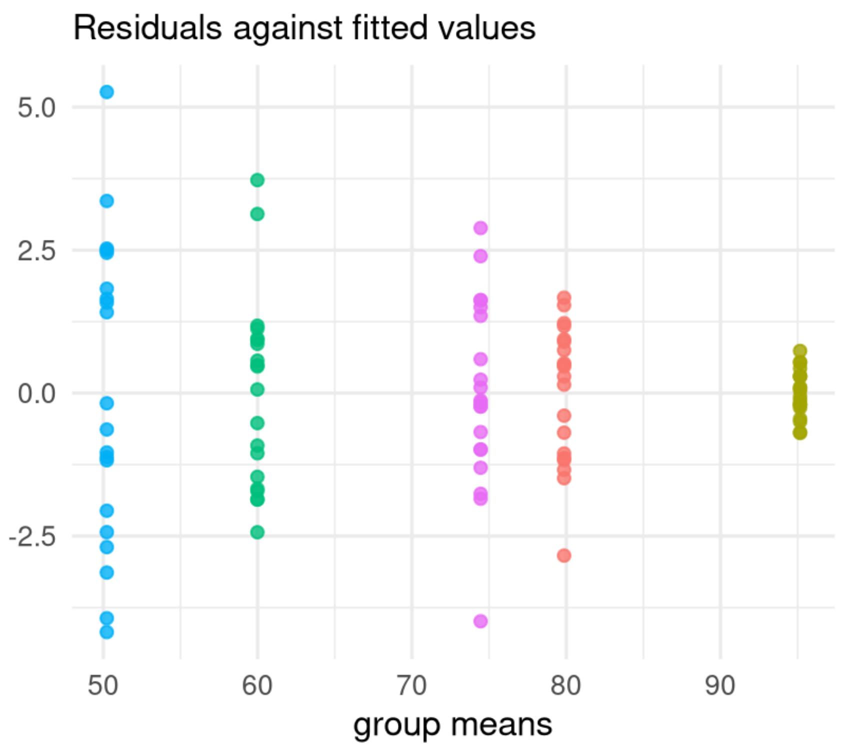


By construction, sample mean of e_{ik} is **always** zero.

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

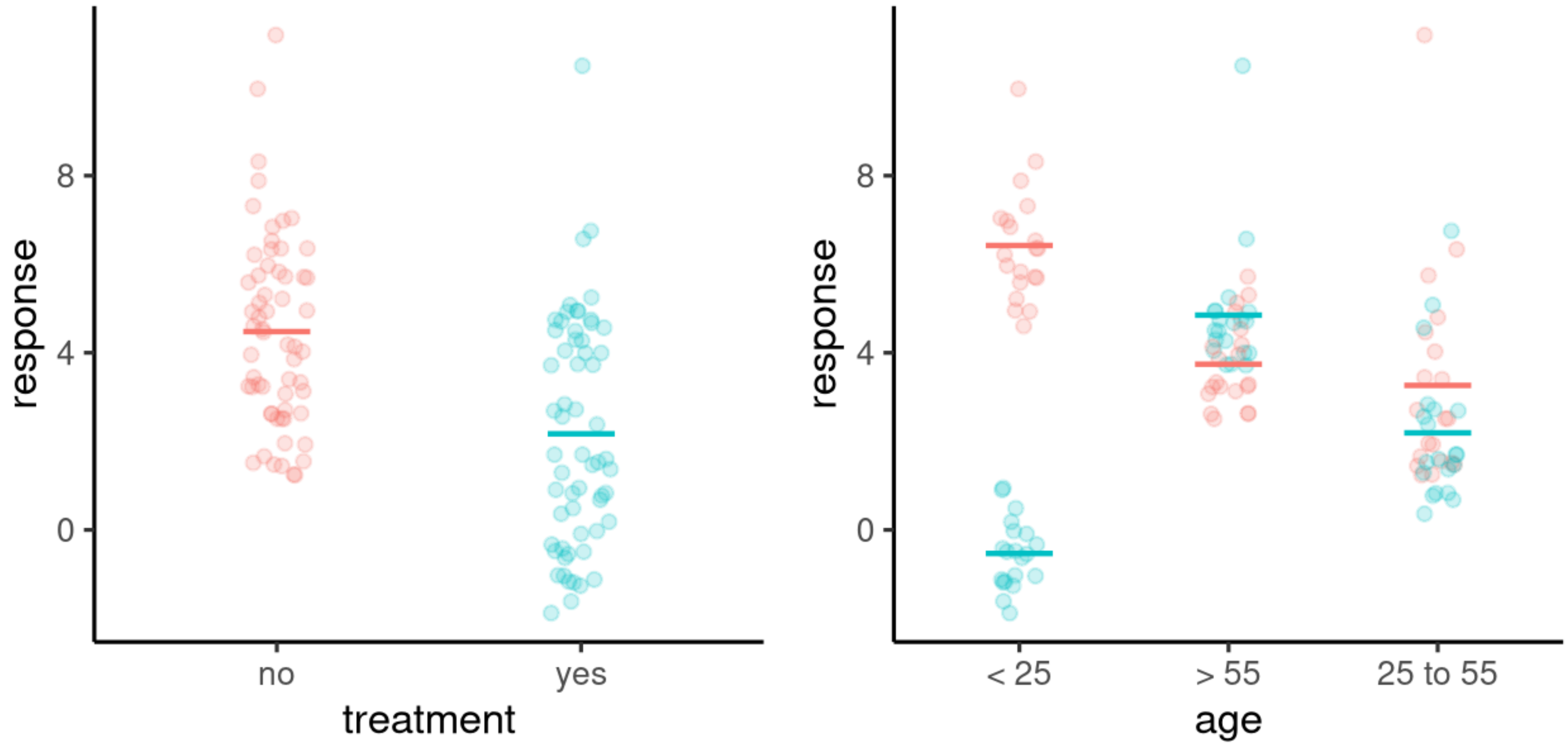
$$\ln\left\{\frac{x-a+\delta}{b+\delta-x}\right\}$$

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 σ.

ANOVA is quite sensitive to this assumption!

Graphical diagnostics

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

```
data(arithmetic, package = "hecedsm")
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 $\mathscr{H}_0 : \sigma_1 = \cdots = \sigma_K$.

- Bartlett's test (assumes normal data)
- Levene's test: a one-way ANOVA for $|y_{ik} - \widehat{\mu}_k|$
- 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

```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

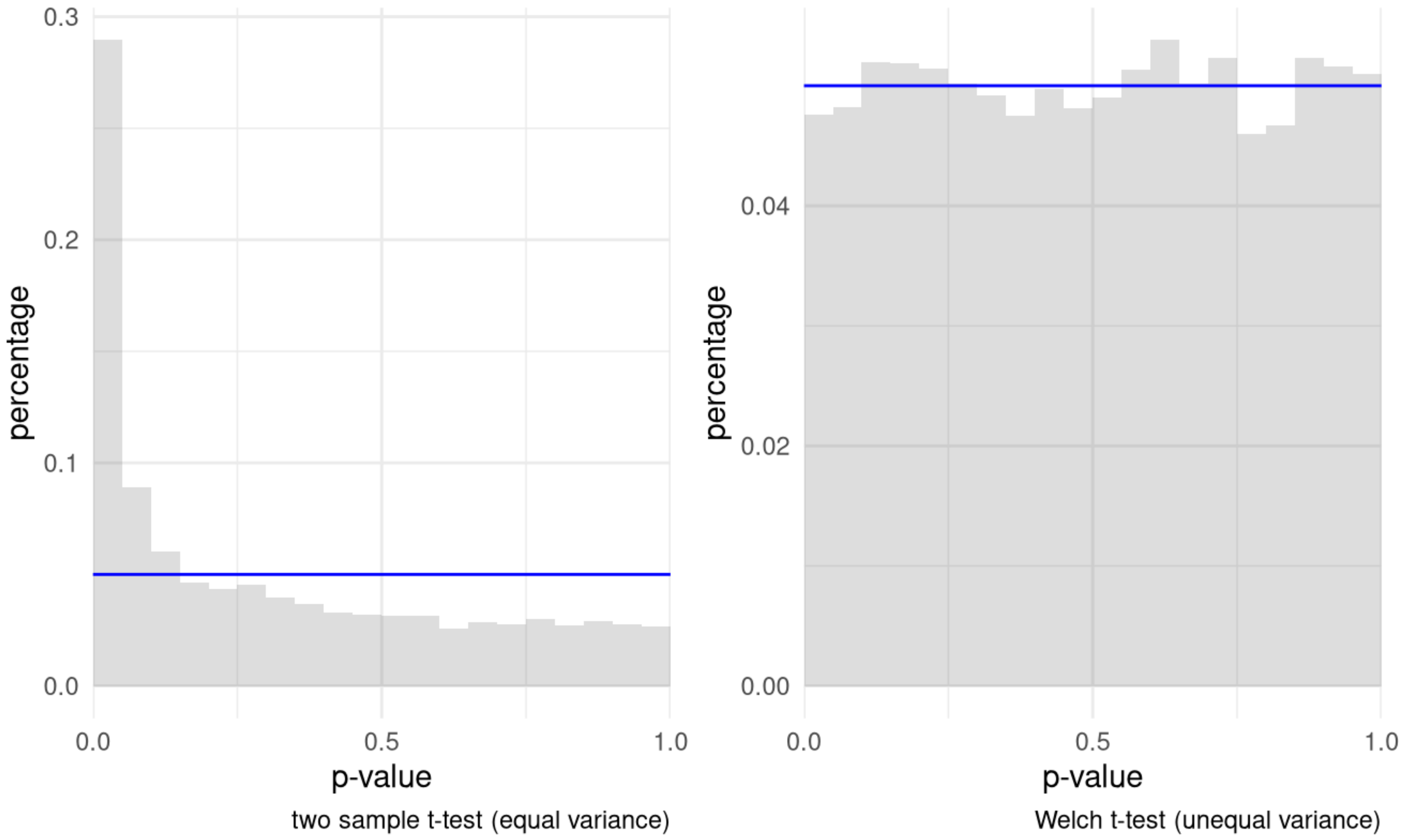
> To make the preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!

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 $\mathscr{H}_0 : \sigma_1 = \cdots = \sigma_K$.
- 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!



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.

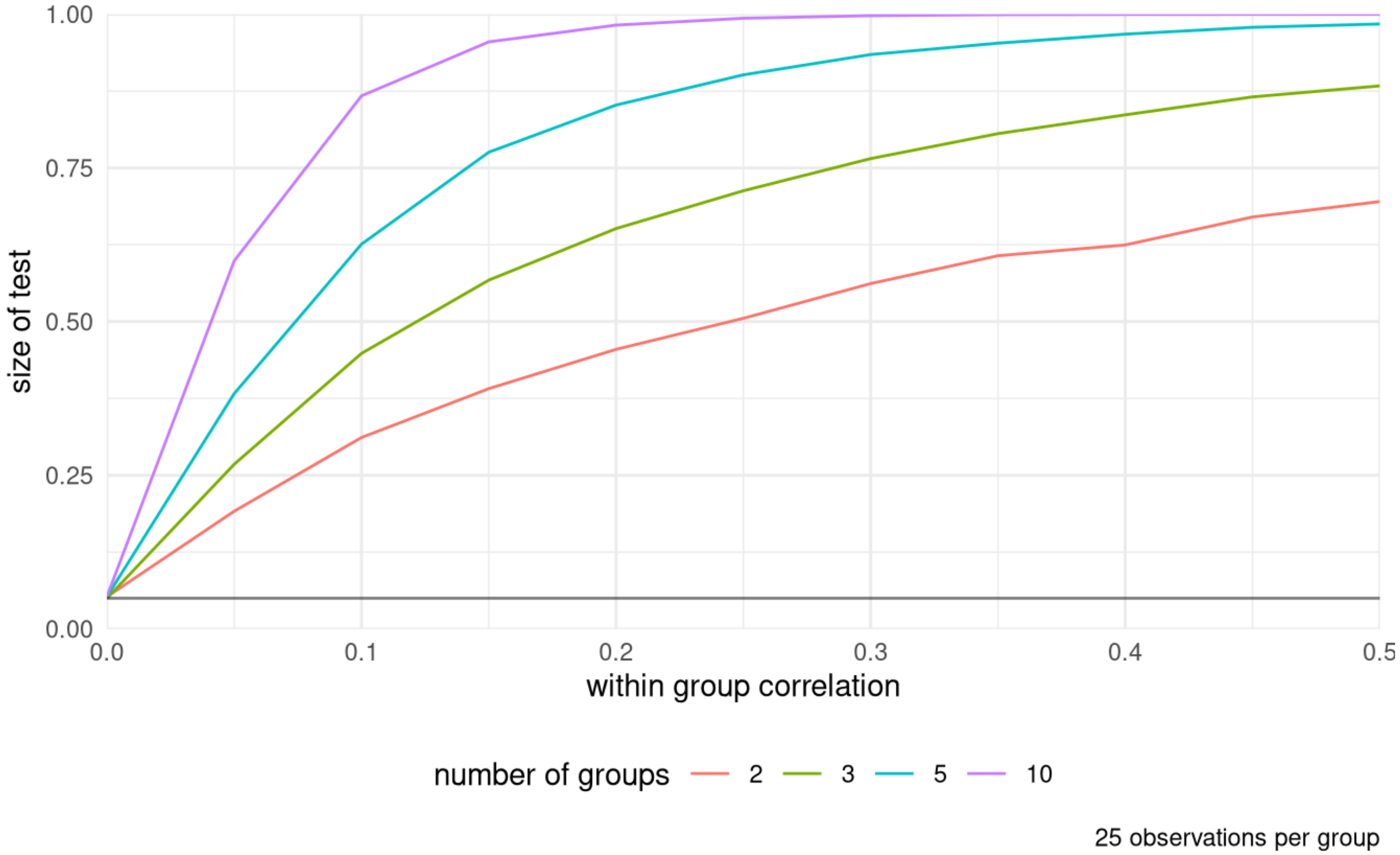
Checking independence

- 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).

Impact of (lack of) independence

Probability of rejecting null hypothesis when there is no difference in mean as a function of correlation between observations in a group.



4: Sample size (normality?)

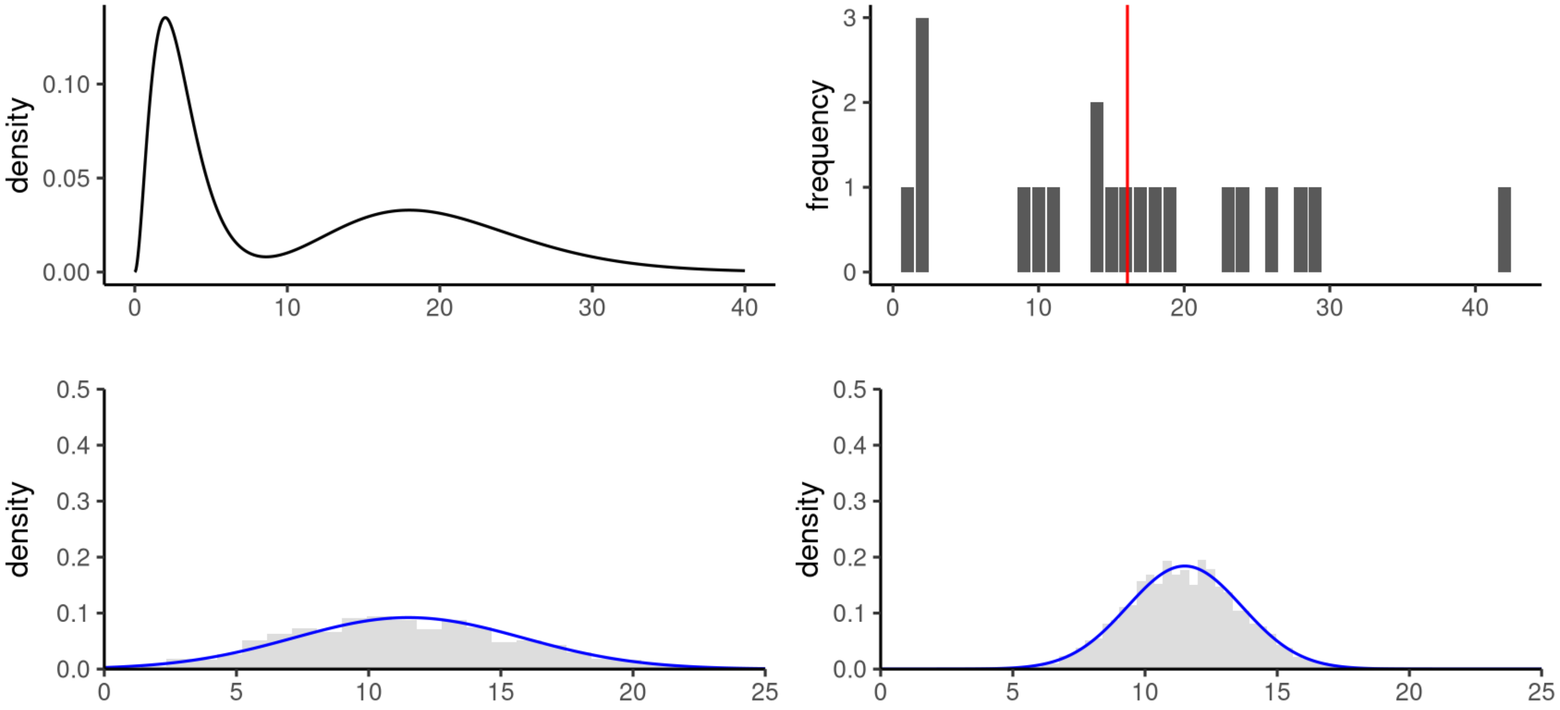
Where does the F-distribution come from?

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.

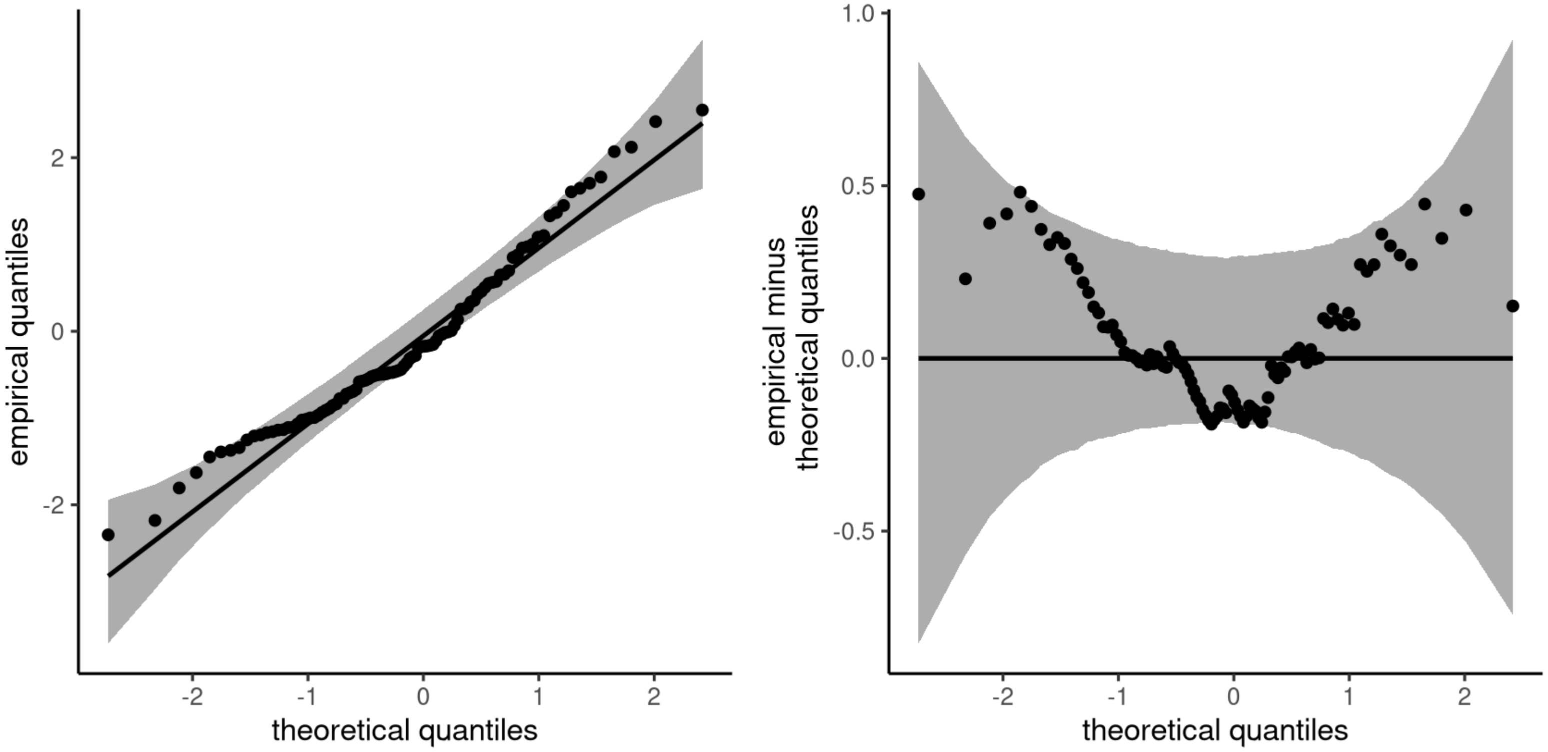
Can check normality via quantile-quantile plot with standardized residuals r_i:

- on the x-axis, the theoretical quantiles $\widehat{F}^{-1}\{\mathrm{rank}(r_i)/(n+1)\}$ 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).

Loss to follow-up

- In longitudinal studies, may lose subjects to attrition
- Creates missing values
- Matters less if we It may be necessary to exclude subjects

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.