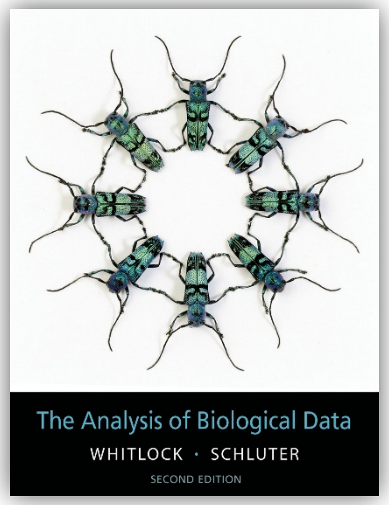
Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon



Week 08

Midterm assignment (by FRIDAY)
Palle says *** ZIP your html! ***

TODAY: ANOVA
THURSDAY: recaps + exp design & ANOVA again

Exam 14-15 January 2019

Requires your mid term and final projects are approved

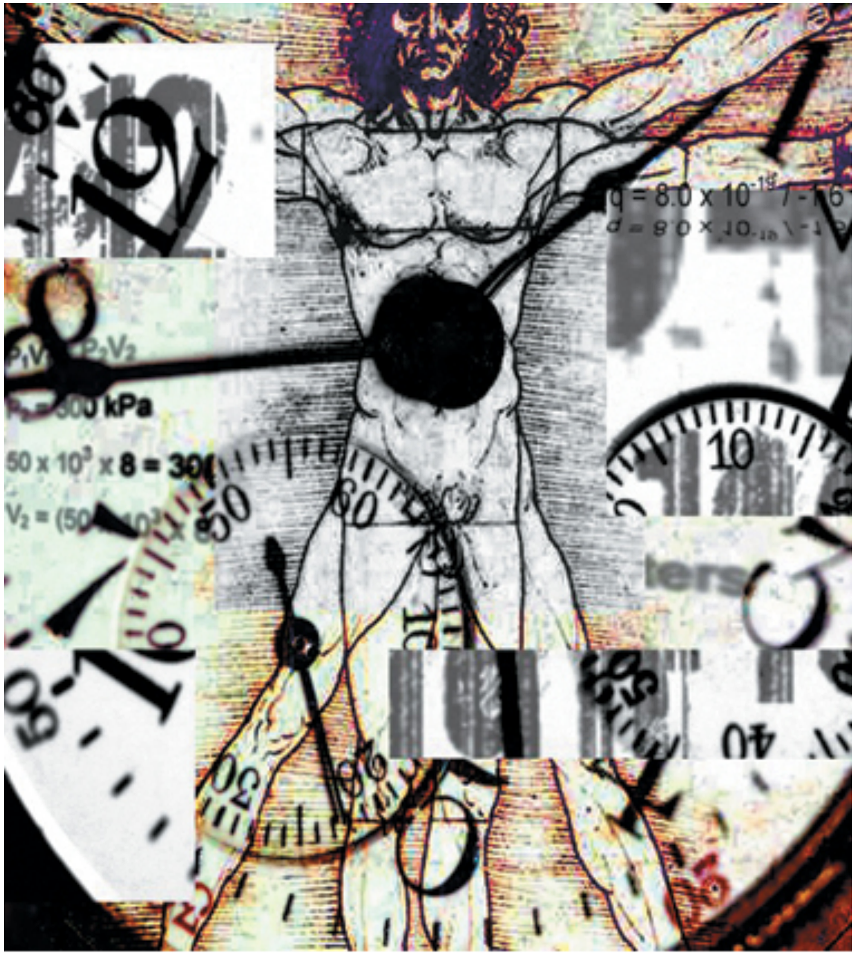
20 mins with no preparations

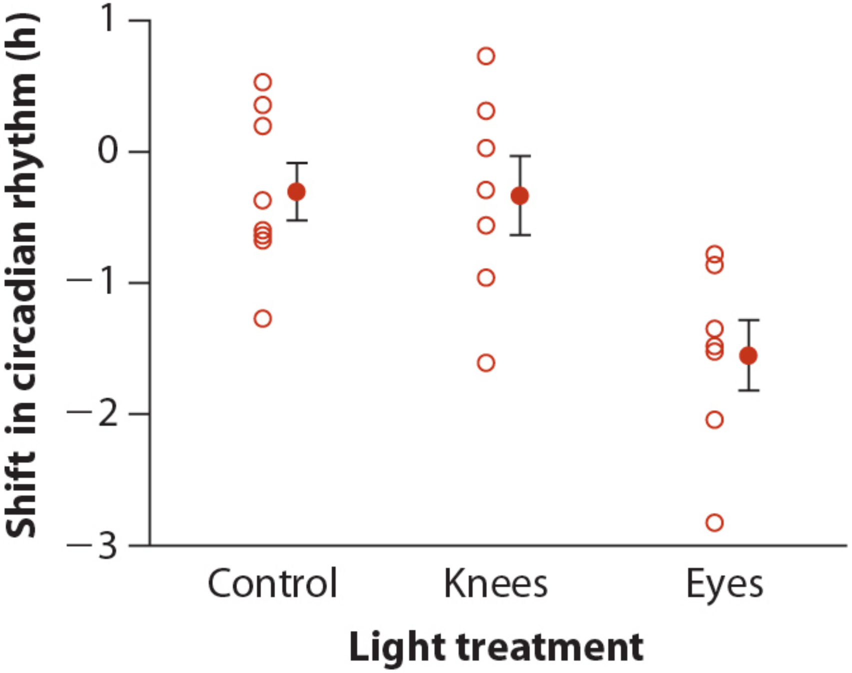
Pre set of questions given in advance + extra Qs

ANOVA fundamental Qs

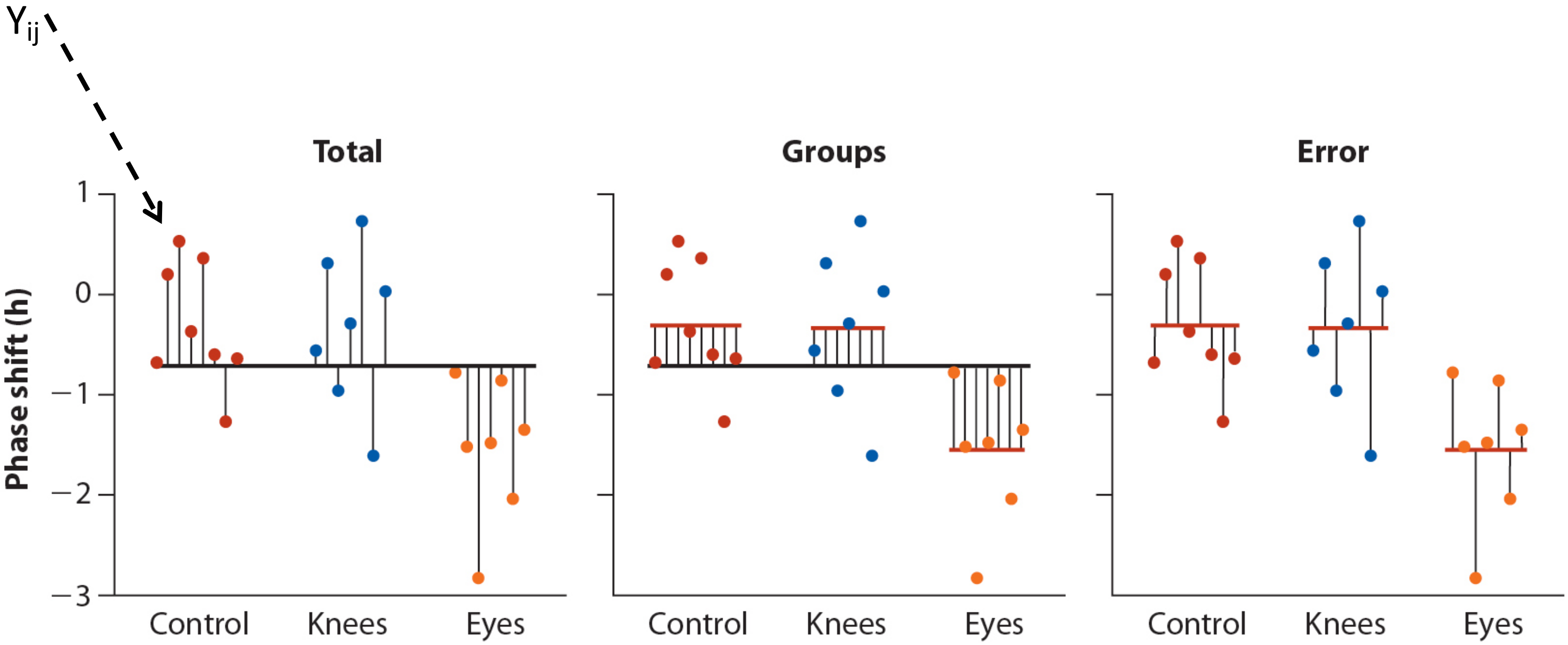
- Are differences among **a priori groups** real ?
- How much does one factor explain of the variation of one “response” variable
- Fixed or Random effects?
  - Are groups pre-determined, of direct interest ?
  - Are groups a random sample among many possible groups (for instance families in a larger population) ?

Shift in circadian rhythm (3 groups: 2 treatments + 1 control)





ANOVA fundamental intuition: partition the **variation**

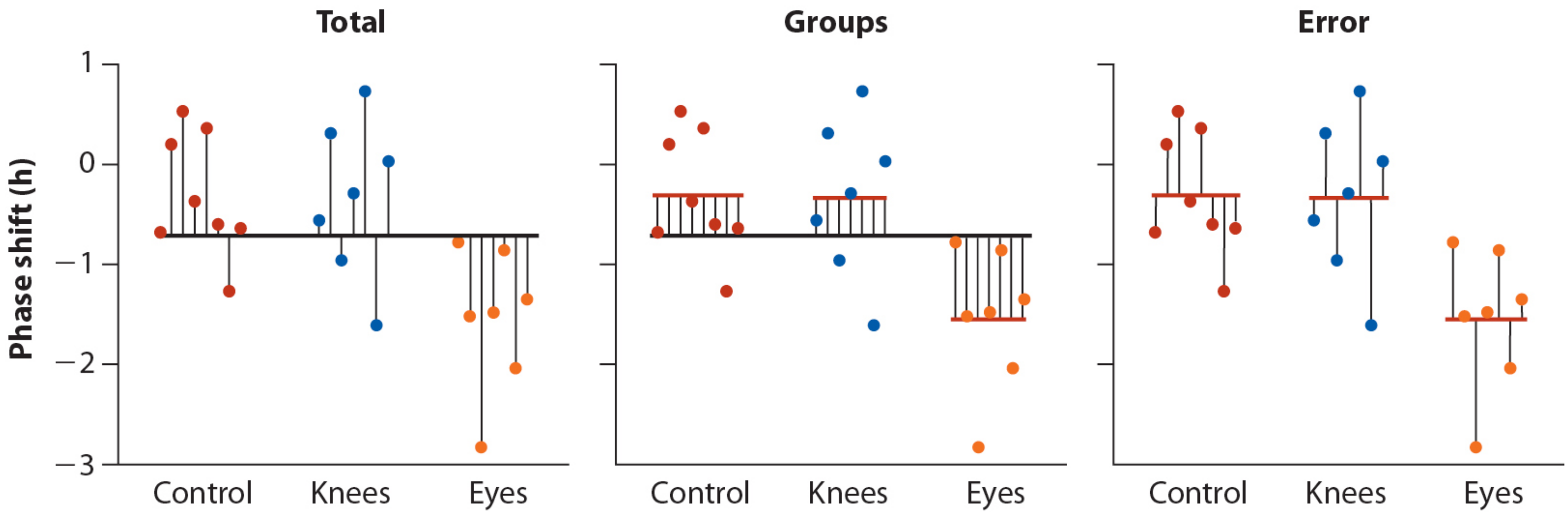


1 obs Y_{ij} – GrandMean = (Mean$_i$ – GrandMean) + (Y_{ij}- Mean$_i$)

all data SS$_{total}$ = SS$_{groups}$ + Ss$_{error}$

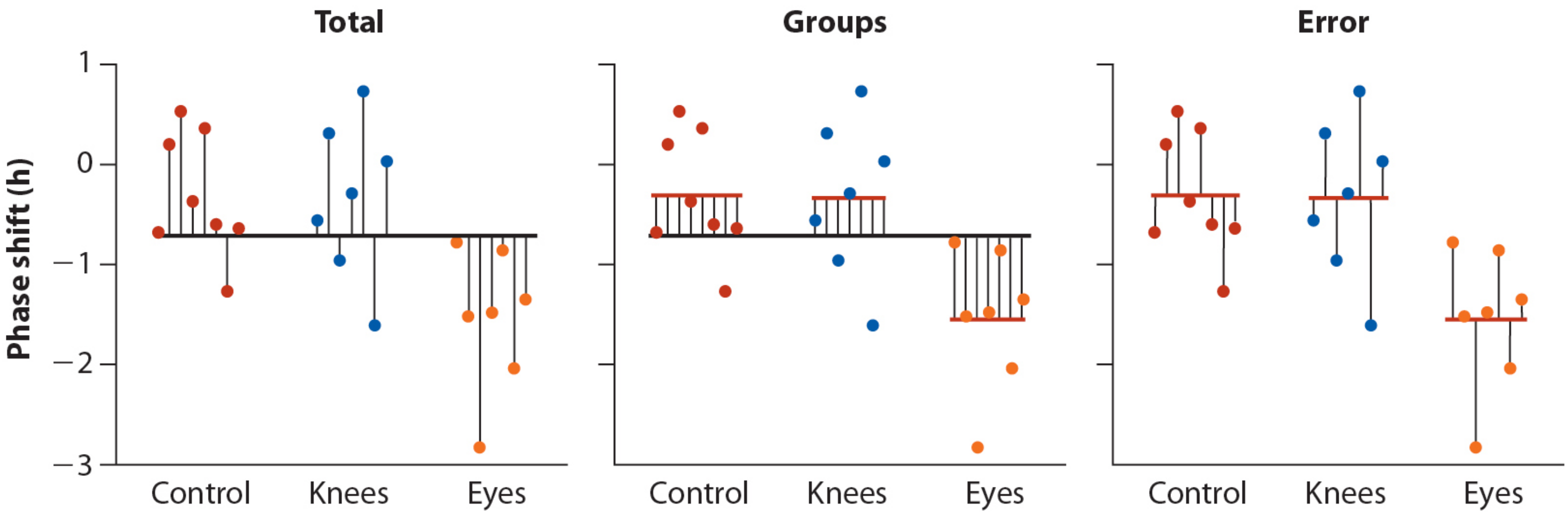
NB: Variation is measured in squares of deviation

ANOVA fundamental intuition: partition the **variation**



$$SS_{total} \quad = \quad SS_{groups} \qquad\qquad + \qquad\qquad SS_{error}$$

ANOVA fundamental intuition: R^2 is the portion of **variation explained**



$R^2 \quad = \; SS_{groups} \, / \, SS_{total}$ (here 0.43)

$SS_{total} \; = \; SS_{groups} \quad + \quad SS_{error}$

How do we do a statistical test on the data ?

Re-use the comparing means (planned comparisons)

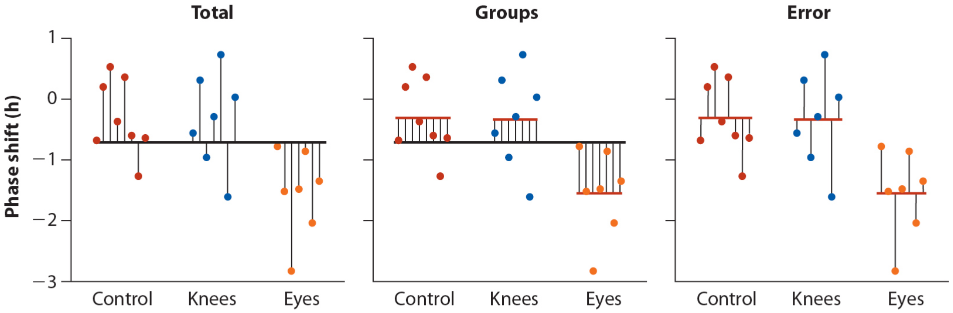
Do a ”global” test for an effect of group

- use R^2 ?
- use something else … (The F-ratio)

The group and error mean squares

MS_{group} represents the amount of variation explained by groups

MS_{error} represents the amount of variation within groups



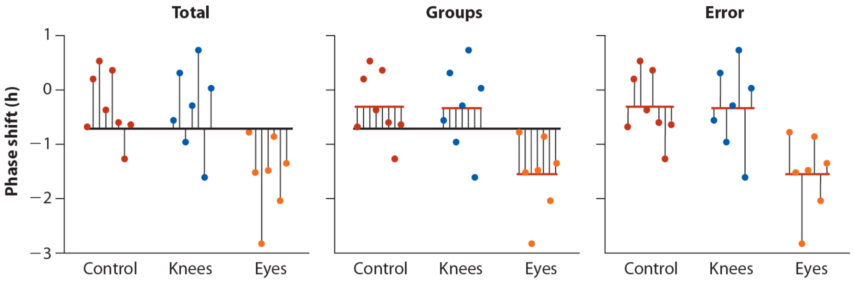
IF H0 is true:

observations in groups are actually samples from the same single population

--> MS_{error} ~ MS_{groups}

Partition of sum of squares and F test

$SS_{total} = SS_{groups} + SS_{error}$



$MS_{groups} = SS_{groups} / df_{groups}$

$MS_{error} = SS_{error} / df_{errors}$

$F_{obs} = MS_{groups} / MS_{error}$

H_o is true: we expect F ~1,

$F \sim F(df_{groups,}\ df_{error})$

H_o is false: we expect F >1

What is the distribution of F under H0?

Probability / MATH

Yij is a random variable

Yij ~ N(μ_i,σ_i)

SS_{group} is also a r.v.

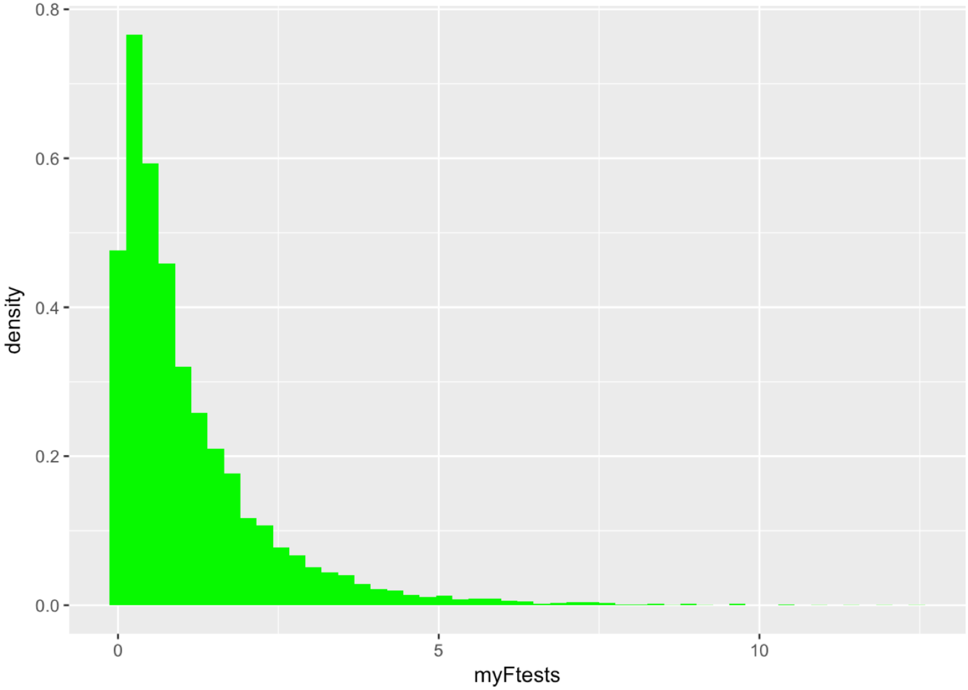
The F ratio is also a r.v. …

Simulation of H0 in R

Generate data from **one** population

Assign them randomly to **3 groups**

Calculate the F test



Null distribution for the F statistic

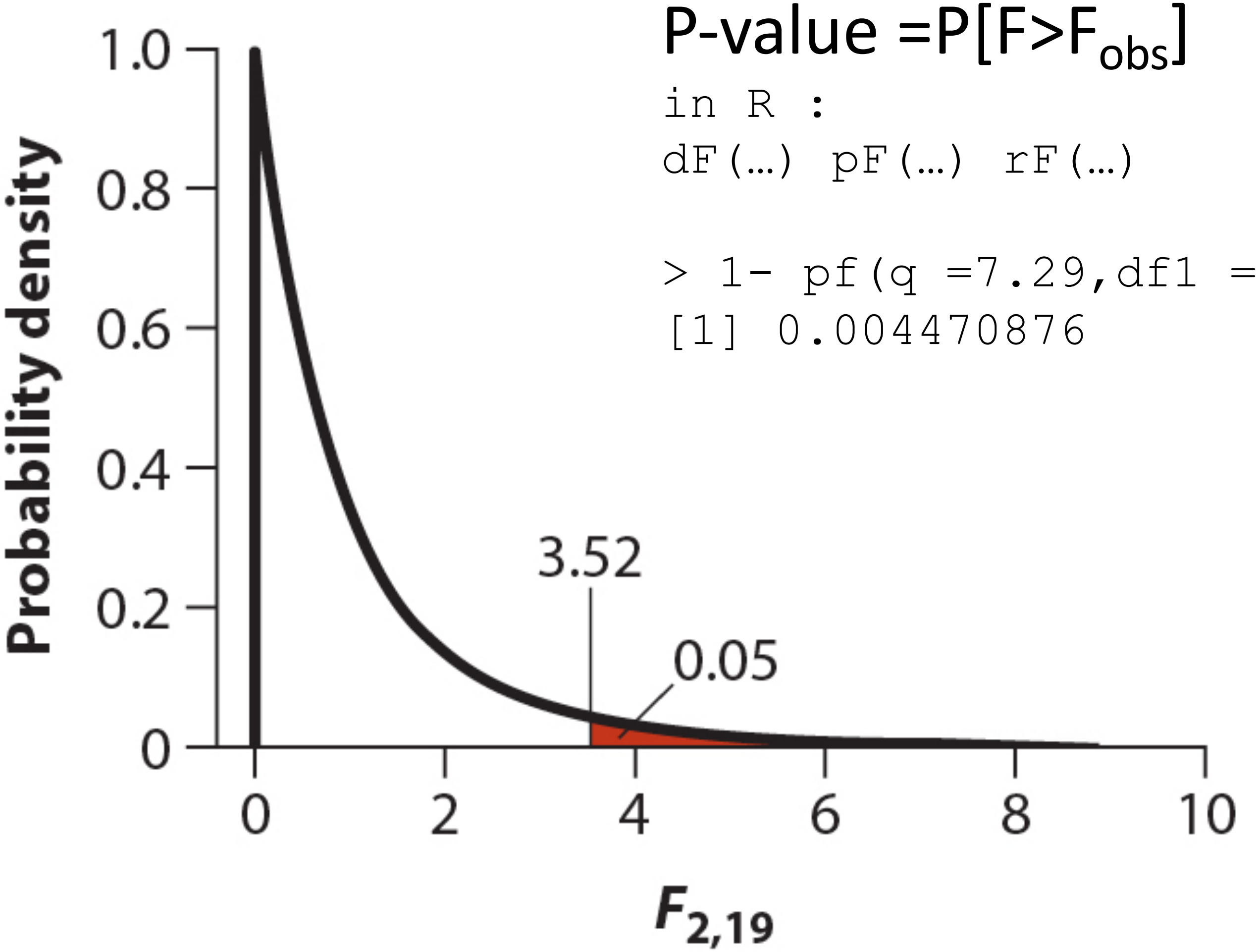
H_o is true: we expect F ~1,

More formally F~ F($df_{groups,}$ df_{error})

P-value =P[F>F_{obs}]

```
in R :
dF(…) pF(…) rF(…)

> 1- pf(q =7.29,df1 = 2, df2 = 19)
[1] 0.004470876
```



the F distribution in R

Random number generation for the F distribution

```
myDeviates <- rf(n=10^4,df1= 2, df2=19)
```

Getting the tail (aka p-value) for an F distribution

```
1- pf(q =7.29,df1 = 2, df2 = 19)
```

n   number of random observations to generate
q   the observed value
df1 number of degrees of freedom in numerator
df2 number of degrees of freedom in denominator

Internal "check”: what proportion of myDeviates exceed 7.29?

ANOVA Assumptions & model check

Observations are randomly drawn from several groups
Obs in each group are normally distributed
Each group has same variance

Model check (often visual)

- trend in residuals
- normality of residuals
- presence of point with “influence”

ALTERNATIVES to parametric ANOVA

- Transform (see week 07)
- Non parametric ANOVAs
- Build your own F-test by resampling (see next weeks).

The mammals dataset

We could also use an ANOVA setting to ask how much ”species” explains the variation in dn/ds or gene expression

Is ”species” a fixed or a random effect ?

How much variation is found between vs among species?

Basic to do list with ANOVA in R … (see also the R code)

Identify the design

- Fixed Factor
- Random Factor

Fit the model accordingly: y ~ x

- lm()
- lme

Test hypothesis

- F-tests and their dfs
- OR permutations (see in coming weeks)

Check the model

- trend in residuals
- normality of residuals