ANOVA

One-way independent & F-distribution

Klinkenberg

13 okt 2017

Inhoud

- f distribution
  - ronald fisher
  - population distribution
  - more samples
- one way independent anova
  - total variance
  - model variance
  - error variance
  - variance components 1
  - f ratio

F-distribution

Ronald Fisher

The F-distribution, also known as Snedecor's F distribution or the Fisher–Snedecor distribution (after Ronald Fisher and George W. Snedecor) is, in probability theory and statistics, a continuous probability distribution. The F-distribution arises frequently as the null distribution of a test statistic, most notably in the analysis of variance; see F-test.

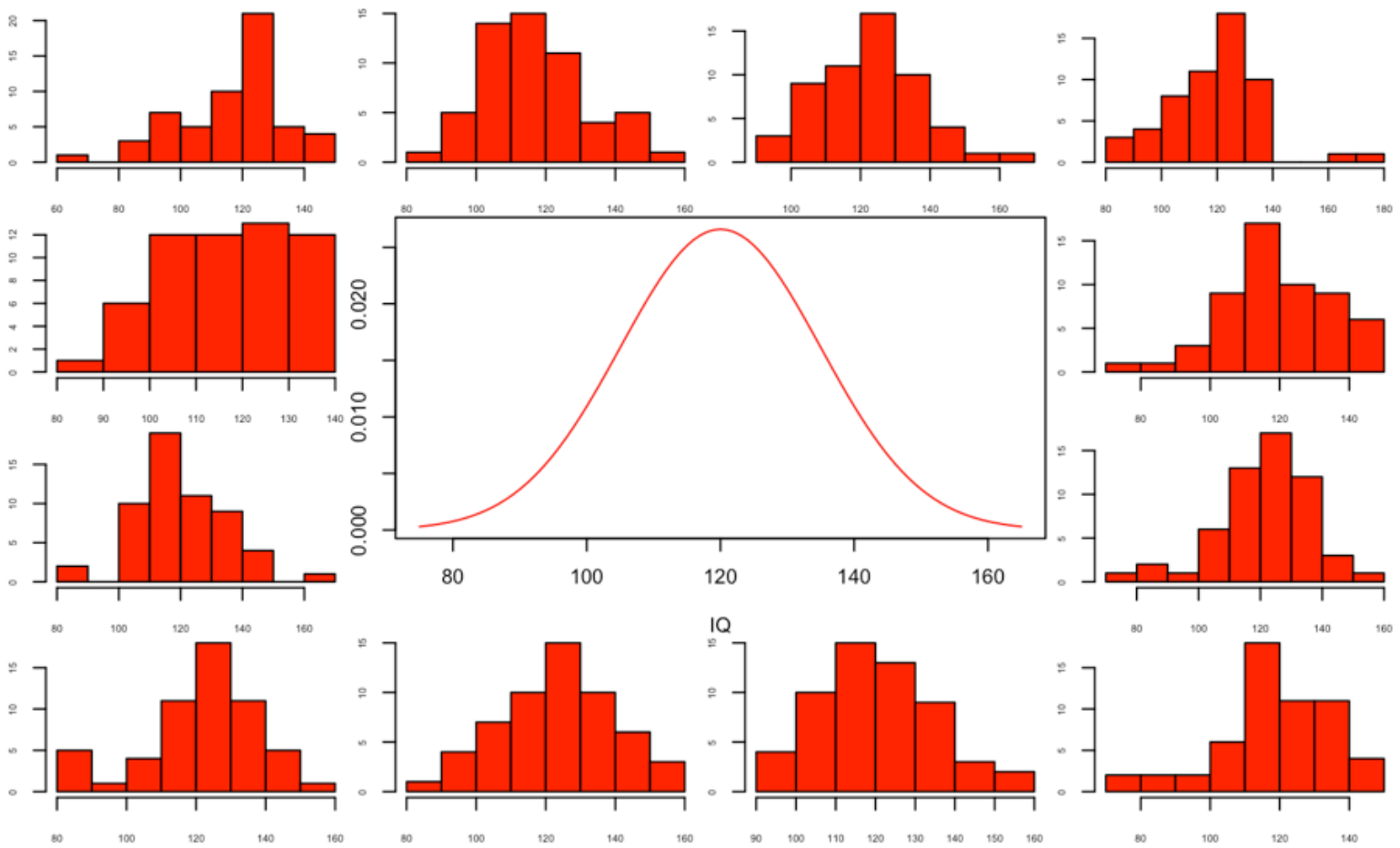
Wikipedia

Sir Ronald Aylmer Fisher FRS (17 February 1890 – 29 July 1962), known as R.A. Fisher, was an English statistician, evolutionary biologist, mathematician, geneticist, and eugenicist. Fisher is known as one of the three principal founders of population genetics, creating a mathematical and statistical basis for biology and uniting natural selection with Mendelian genetics.

Wikipedia

Population distribution

```
layout(matrix(c(2:6,1,1,7:8,1,1,9:13), 4, 4))

n  = 56    # Sample size
df = n - 1 # Degrees of freedom

mu    = 120
sigma = 15

IQ = seq(mu-45, mu+45, 1)

par(mar=c(4,2,0,0))
plot(IQ, dnorm(IQ, mean = mu, sd = sigma), type='l', col="red")

n.samples = 12

for(i in 1:n.samples) {

  par(mar=c(2,2,0,0))
  hist(rnorm(n, mu, sigma), main="", cex.axis=.5, col="red")

}
```


0.020
0.010
0.000
80
100
120
140
160
IQ

F-statistic

$$F = \frac{MS_{model}}{MS_{error}} = \frac{SIGNAL}{NOISE}$$

So the F-statistic represents a signal to noise ratio by deviding the model variance component by the error variance component.

A samples

Let's take two sample from our normal populatiion and calculate the F-value.

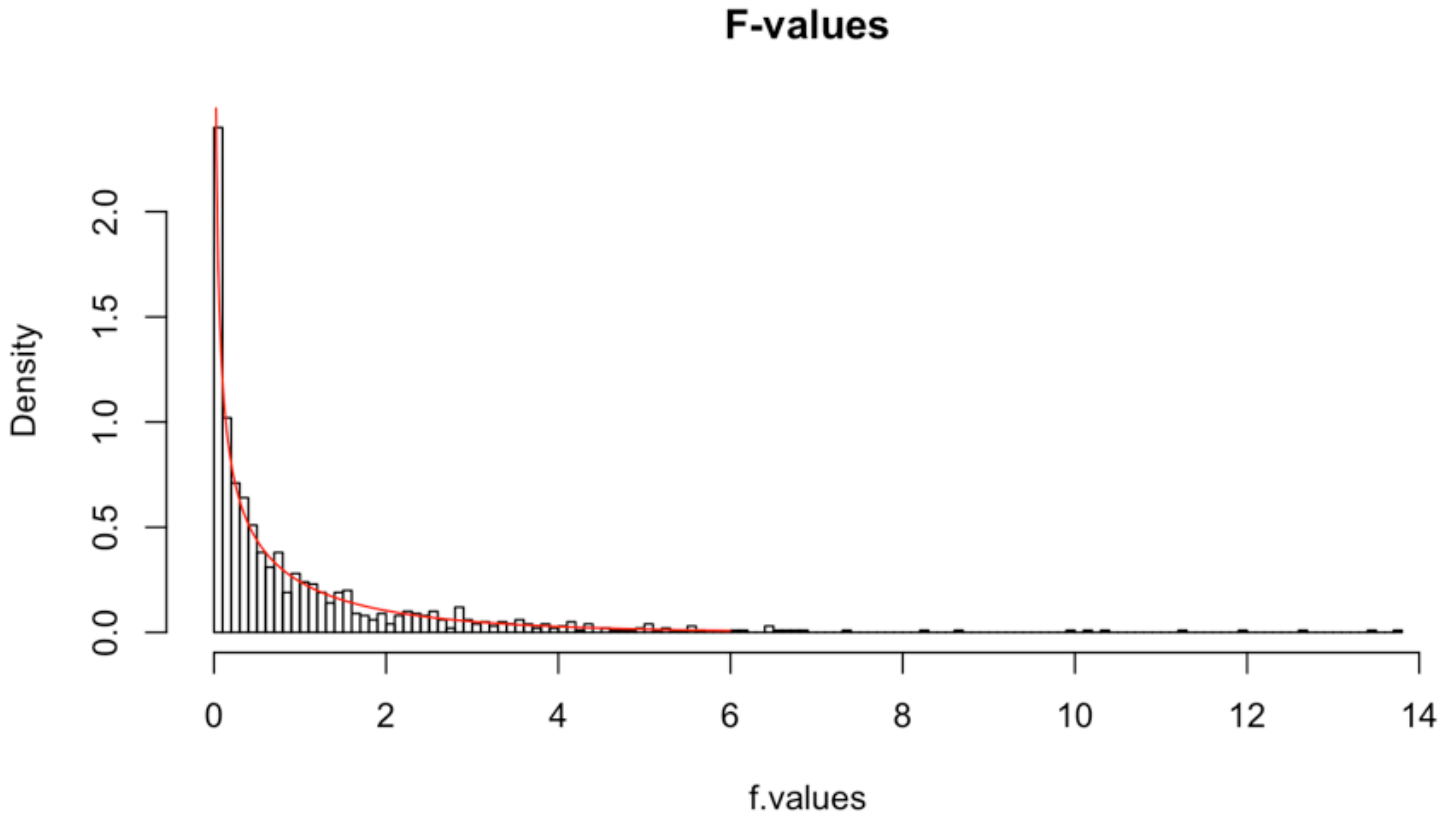
```
x.1 = rnorm(n, mu, sigma)
x.2 = rnorm(n, mu, sigma)

data <- data.frame(group = rep(c("s1", "s2"), each=n), score = c(x.1,x.2))

F = summary(aov(lm(score ~ group, data)))[[1]]$F[1]
F
```

```
## [1] 0.6114998
```

More samples

let's take more samples and calculate the F-value every time.

```
n.samples = 1000

f.values = vector()

for(i in 1:n.samples) {

  x.1 = rnorm(n, mu, sigma); x.1
  x.2 = rnorm(n, mu, sigma); x.2

  data <- data.frame(group = rep(c("s1", "s2"), each=n), score = c(x.1,x.2))

  f.values[i] = summary(aov(lm(score ~ group, data)))[[1]]$F[1]

}

k = 2
N = 2*n

df.model = k - 1
df.error = N - k
```


F-values
Density
0.0
0.5
1.0
1.5
2.0
0
2
4
6
8
10
12
14
f.values

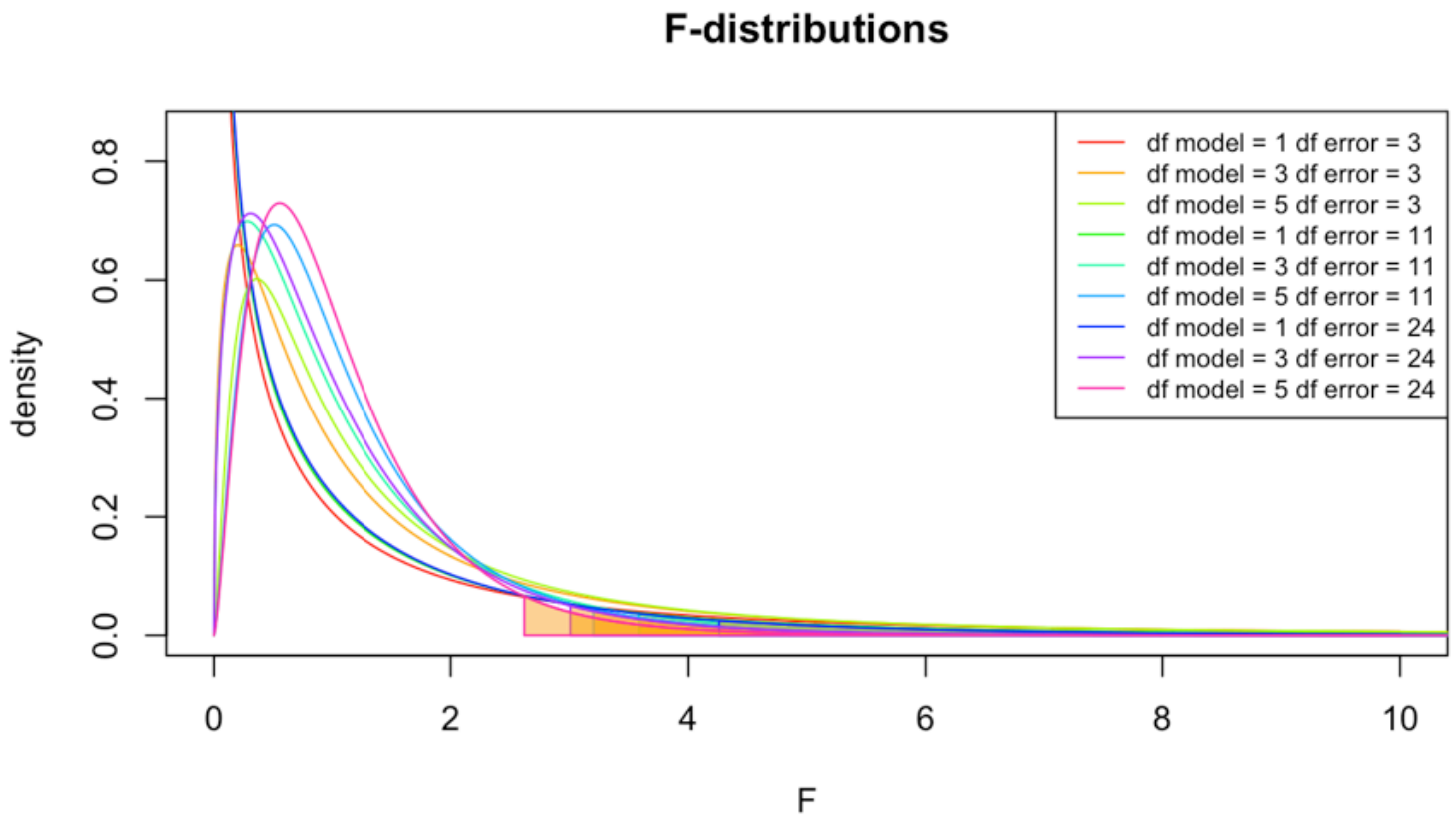
F-distribution

So if the population is normaly distributed (assumption of normality) the f-distribution represents the signal to noise ration given a certain number of samples ($df_{model} = k - 1$) and sample size ($df_{error} = N - k$).

The F-distibution therefore is different for different sample sizes and number of groups.

F-distribution

```
multiple.n  = c(5, 15, 30)
multiple.k  = c(2, 4, 6)
multiple.df.model = multiple.k - 1
multiple.df.error = multiple.n - multiple.k
col         = rainbow(length(multiple.df.model) * length(multiple.df.error))
F = seq(0, 10, .01)

plot(F,  df(F, multiple.df.model[1], multiple.df.error[1]), type = "l",
     xlim = c(0,10), ylim = c(0,.85),
     xlab = "F", ylab="density",
     col  = col[1], main="F-distributions" )

dfs = expand.grid(multiple.df.model, multiple.df.error)

for(i in 2:dim(dfs)[1]) {

  lines(F, df(F, dfs[i,1], dfs[i,2]), col=col[i])

  critical.f <- qf(.95, dfs[i,1], dfs[i,2])

  f.alpha <- seq(critical.f, 1000, .01)
```


F-distributions
density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
F
df model = 1 df error = 3
df model = 3 df error = 3
df model = 5 df error = 3
df model = 1 df error = 11
df model = 3 df error = 11
df model = 5 df error = 11
df model = 1 df error = 24
df model = 3 df error = 24
df model = 5 df error = 24

Compare 2 or more independent groups.

One-way independent ANOVA

Assumptions

Assuming th H_0 hypothesis to be true, the following should hold:

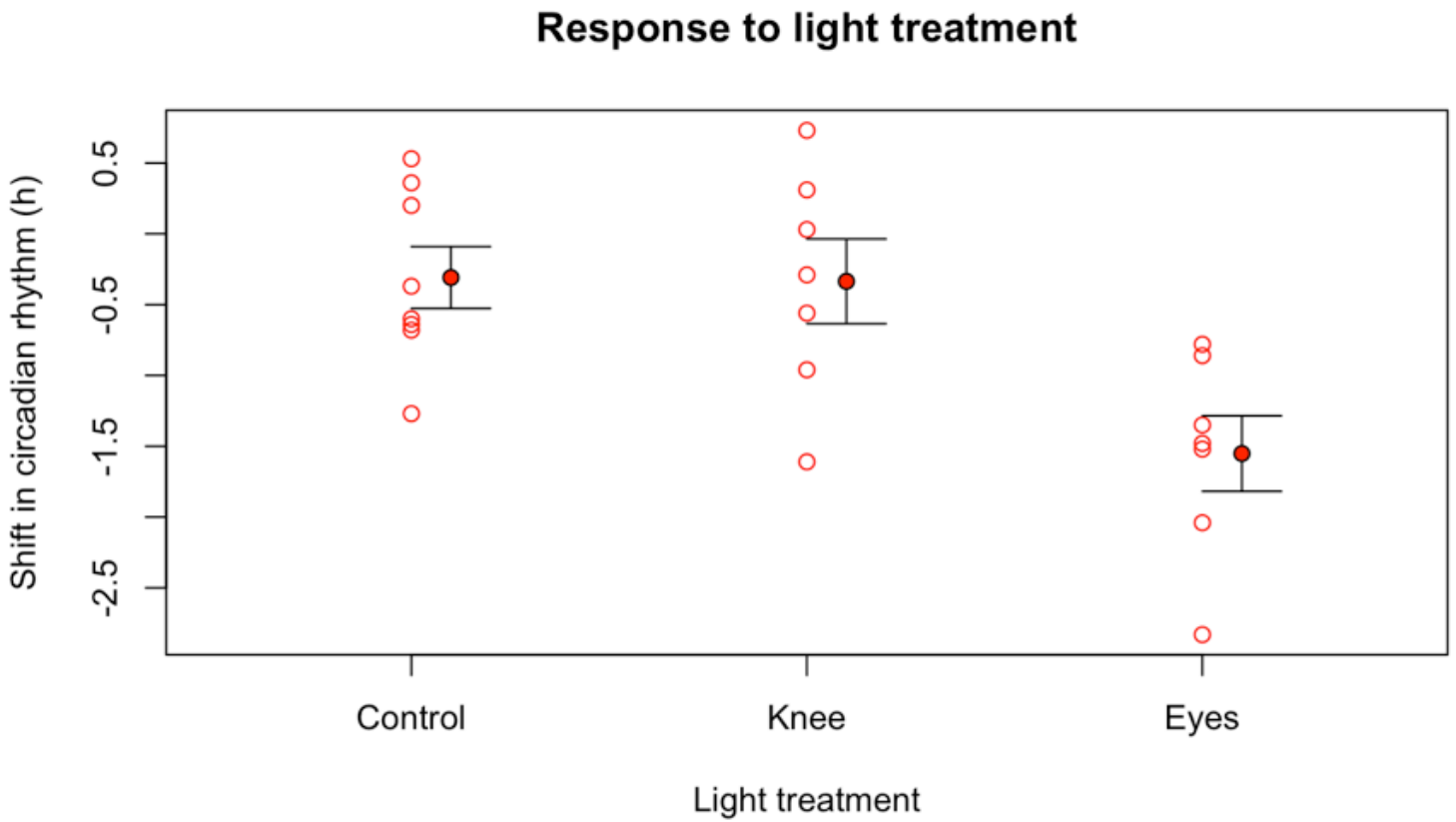
- Continuous variable
- Random sample
- Normaly distributed
    - Shapiro-Wilk test
- Equal variance within groups
    - Levene's test

Jet lag

Wright and Czeisler (2002) performed an experiment where they measured the circadian rhythm by the daily cycle of melatonin production in 22 subjects randomly assigned to one of three light treatments.

- Control condition (no light)
- Knees (3 hour light to back of knees)
- Eyes (3 hour light in eyes)

```
rm(list=ls())
x.c = c( .53, .36,  .2,  -.37, -.6,  -.64, -.68,-1.27) # Control
x.k = c( .73, .31,  .03, -.29, -.56, -.96,-1.61      ) # Knees
x.e = c(-.78,-.86,-1.35,-1.48,-1.52,-2.04,-2.83      ) # Eyes
x   = c( x.c, x.k, x.e )                               # Conditions combined
```


Response to light treatment
Shift in circadian rhythm (h)
0.5
-0.5
-1.5
-2.5
Control
Knee
Eyes
Light treatment

Variance components

Variance	Sum of Squares	DF	Mean Squares	F-ratio
Model	$SS_{model} = \sum n_k(\bar{X}_k - \bar{X})^2$	$k-1$	$\frac{SS_{model}}{df_{model}}$	$\frac{MS_{model}}{MS_{error}}$
Error	$SS_{error} = \sum s_k^2(n_k - 1)$	$N-k$	$\frac{SS_{error}}{df_{error}}$	
Total	$SS_{total} = SS_{model} + SS_{error}$	$N-1$	$\frac{SS_{total}}{df_{total}}$	

Where N is the total sample size, n_k is the sample size per category and k is the number of categories. Finally s_k^2 is the variance per category.

Total variance

$$MS_{total} = s_x^2$$

```
ms.t = var(x); ms.t
```

```
## [1] 0.7923732
```

```
sum( (x - mean(x))^2 ) / (length(x) - 1)
```

```
## [1] 0.7923732
```

$$SS_{total} = s_x^2(N-1)$$

```
N = length(x)

ss.t = var(x) * (N-1); ss.t
```

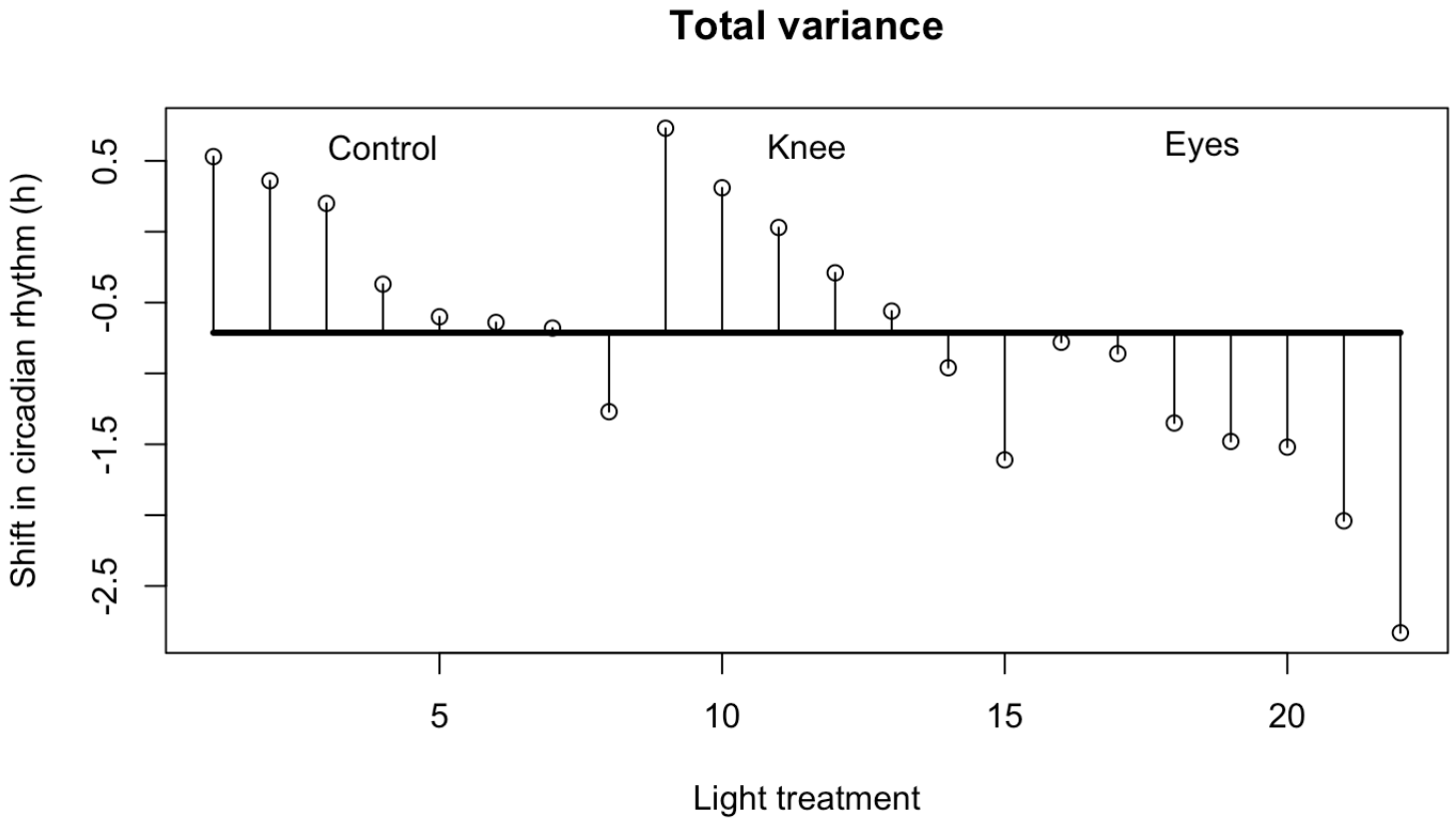
```
## [1] 16.63984
```

```
sum( (x - mean(x))^2 )
```

```
## [1] 16.63984
```

Visual SS_{total}

```
# Assign labels
lab = c("Control", "Knee", "Eyes")

# Plot all data points
plot(1:N,x,
     ylab="Shift in circadian rhythm (h)",
     xlab="Light treatment",
     main="Total variance")

# Add mean line
lines(c(1,22),rep(mean(x),2),lwd=3)

# Add delta lines / variance components
segments(1:N, mean(x), 1:N, x)

# Add labels
text(c(4,11.5,18.5),rep(.6,3),labels=lab)
```


Total variance
Control
Knee
Eyes
Shift in circadian rhythm (h)
0.5
-0.5
-1.5
-2.5
5
10
15
20
Light treatment

Model variance

$$MS_{model} = \frac{SS_{model}}{df_{model}}$$

$$df_{model} = k - 1$$

Where k is the number of independent groups and

$$SS_{model} = \sum_{k} n_k(\bar{X}_k - \bar{X})^2$$

```
k       = 3
n.c = length(x.c)
n.k = length(x.k)
n.e = length(x.e)
```

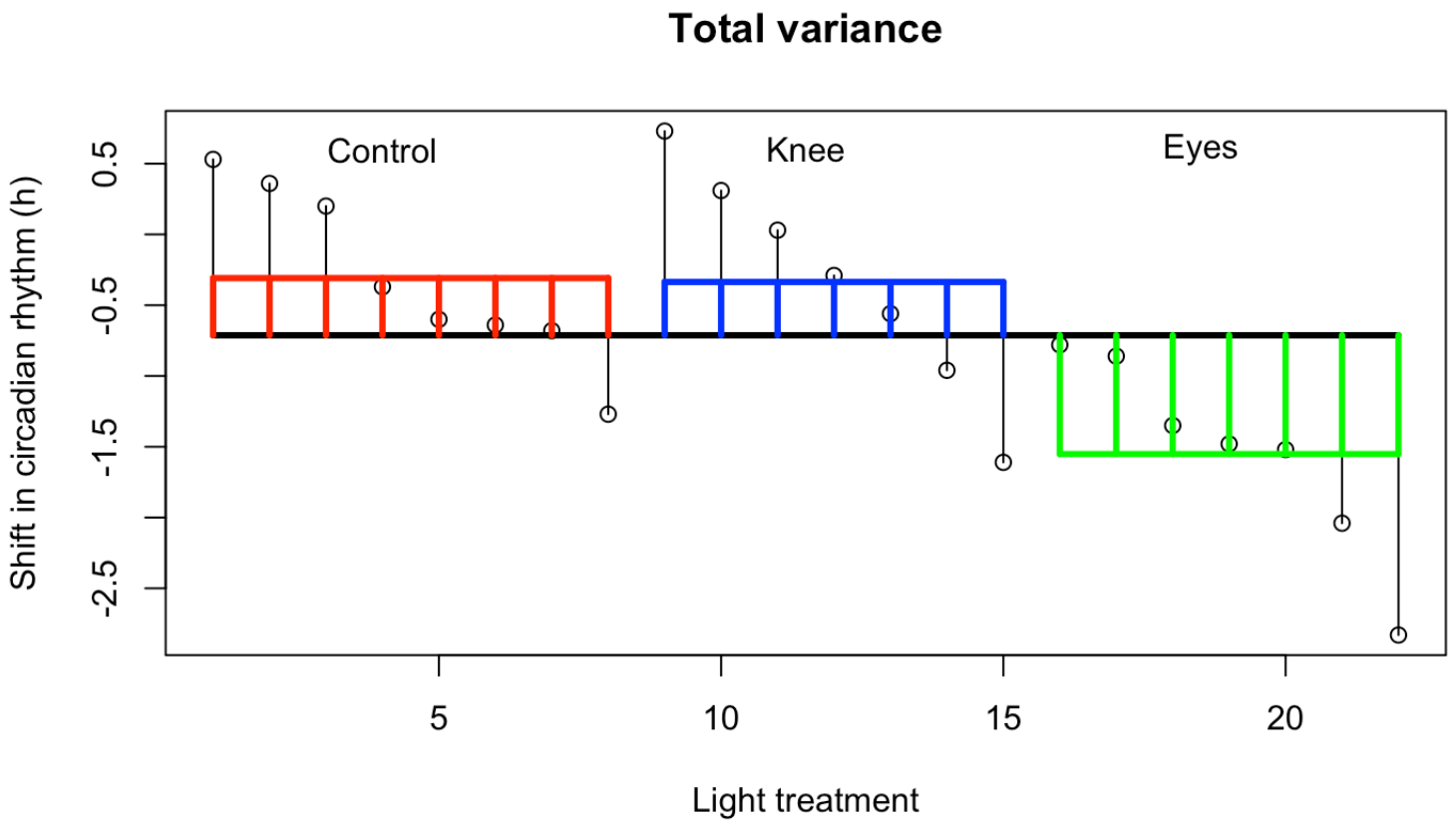
```
ss.m.c = n.c * (mean(x.c) - mean(x))^2
ss.m.k = n.k * (mean(x.k) - mean(x))^2
ss.m.e = n.e * (mean(x.e) - mean(x))^2

ss.m = sum(ss.m.c, ss.m.k, ss.m.e); ss.m
```

```
## [1] 7.224492
```

```
df.m = (k - 1)
ms.m = ss.m / df.m; ms.m
```

```
## [1] 3.612246
```

Visual SS_{model}


Total variance
Control
Knee
Eyes
Shift in circadian rhythm (h)
0.5
-0.5
-1.5
-2.5
5
10
15
20
Light treatment

Error variance

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$
$$df_{error} = N - k$$

where

$$SS_{error} = \sum_k s_k^2 (n_k - 1) = \sum_k \frac{\sum (x_{ik} - \bar{x}_k)^2}{(n_k - 1)} (n_k - 1)$$

```
ss.e.c = var(x.c) * (n.c - 1)
ss.e.k = var(x.k) * (n.k - 1)
ss.e.e = var(x.e) * (n.e - 1)

ss.e = sum(ss.e.c, ss.e.k, ss.e.e); ss.e
```

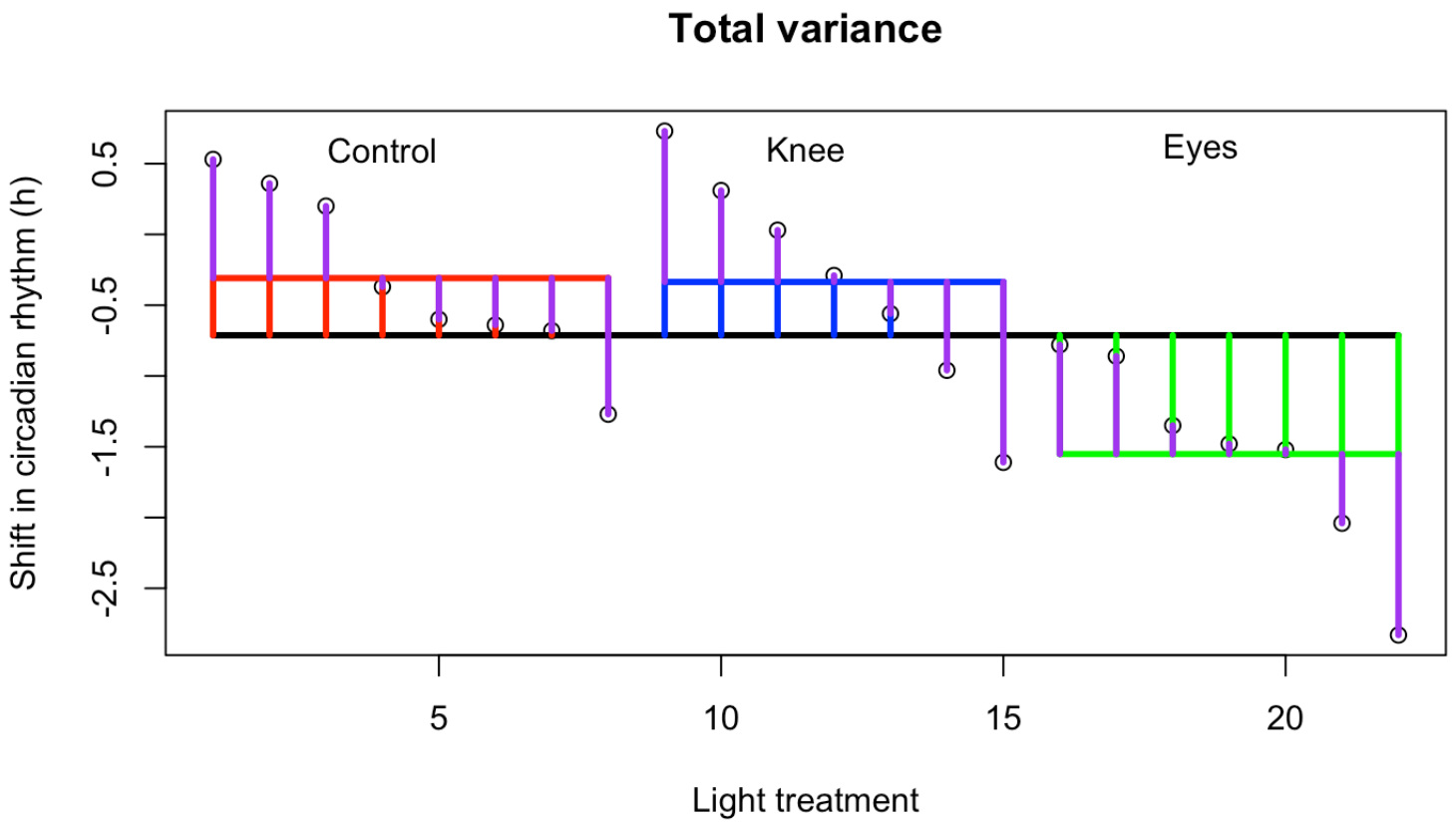
$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

$$df_{error} = N - k$$

```
df.e = (N - k)
ms.e = ss.e / df.e; ms.e
```

```
## [1] 0.4955445
```

Visual SS_{error}


Total variance
Control
Knee
Eyes
Shift in circadian rhythm (h)
0.5
-0.5
-1.5
-2.5
5
10
15
20
Light treatment

Variance components

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

$$16.6398364 = 7.2244917 + 9.4153446$$

$$MS_{total} = \frac{SS_{total}}{df_{total}} = 0.7923732$$

$$MS_{model} = \frac{SS_{model}}{df_{model}} = 3.6122459$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = 0.4955445$$

F-ratio

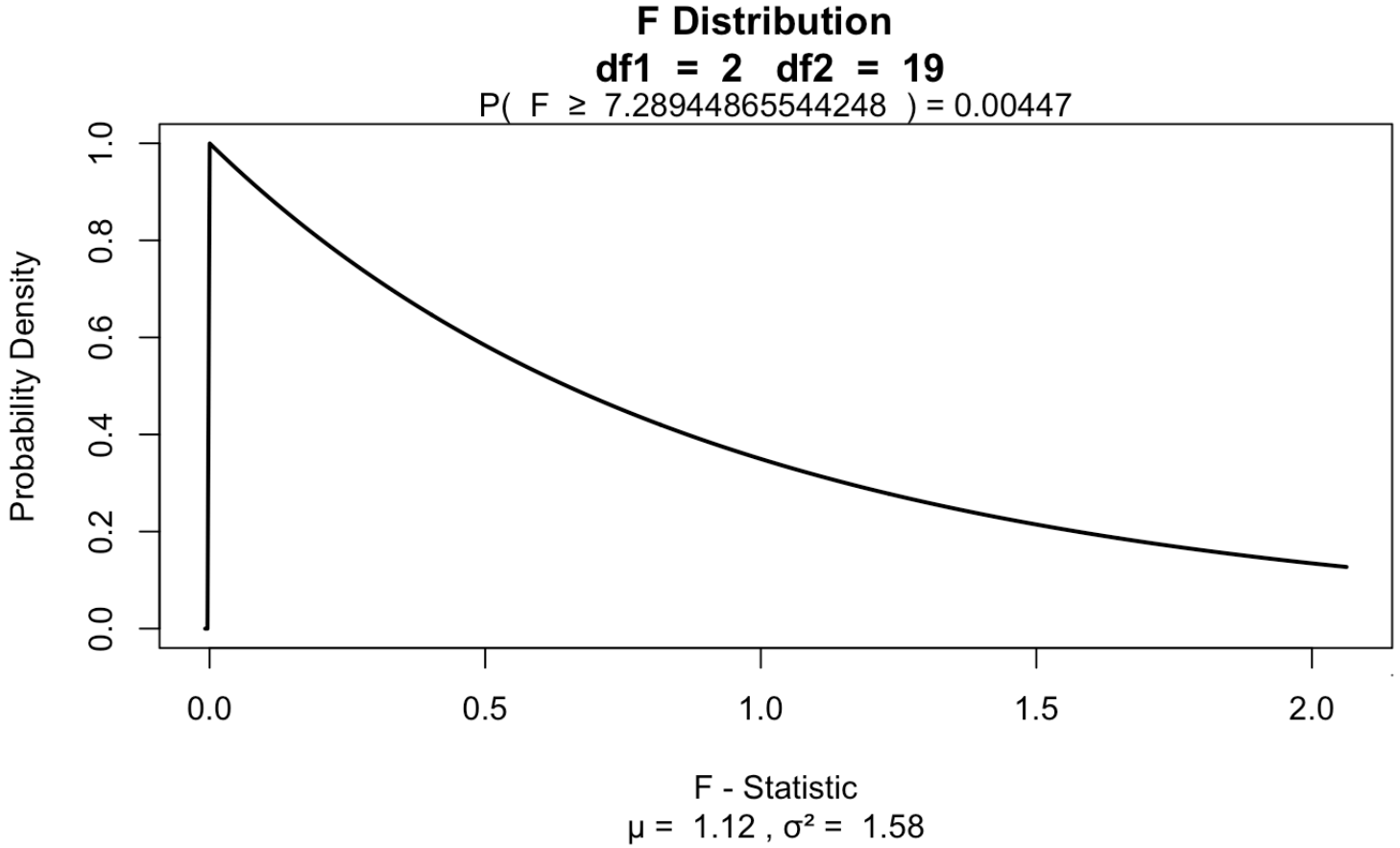
$$F = \frac{MS_{model}}{MS_{error}} = \frac{SIGNAL}{NOISE}$$

```
F = ms.m / ms.e; F
```

```
## [1] 7.289449
```

Reject H_0?

```
if(!"visualize" %in% installed.packages()) { install.packages("visualize") }
library("visualize")

visualize.f(F, df.m,df.e,section="upper")
```


F Distribution
df1 = 2 df2 = 19
P(F ≥ 7.28944865544248) = 0.00447
Probability Density
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
F - Statistic
μ = 1.12 , σ² = 1.58

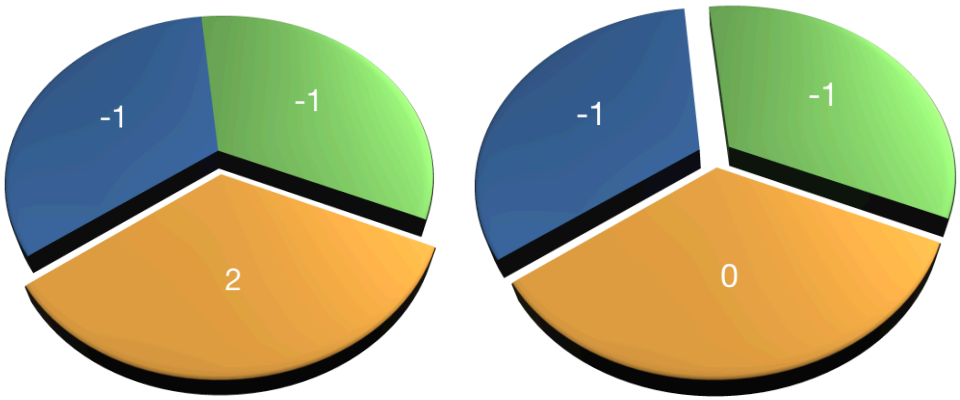
Contrasts

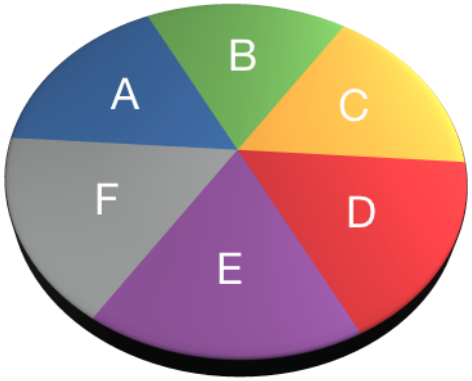
Planned comparisons

- Exploring differences of theoretical interest
- Higher precision
- Higher power

Contrasts

- Only use chunks once
- Values add up to 0



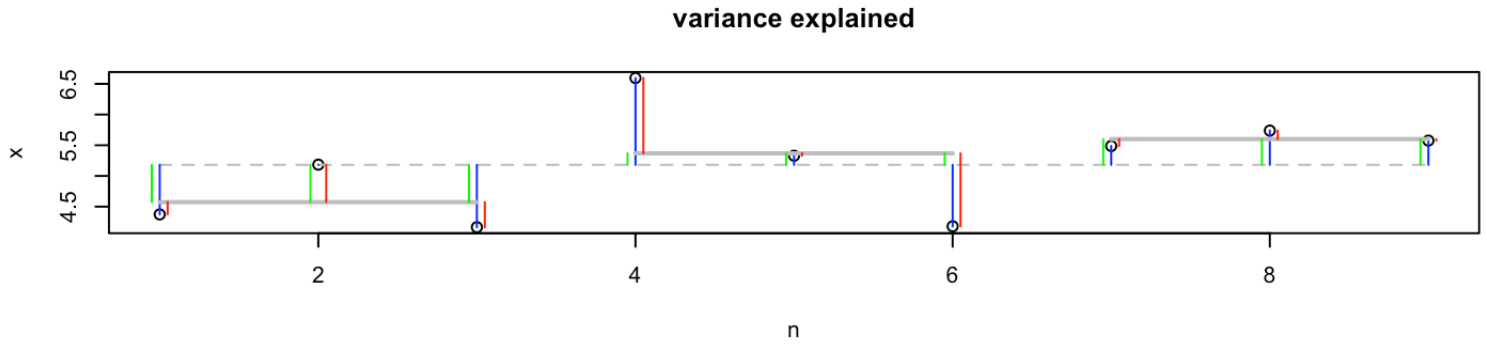


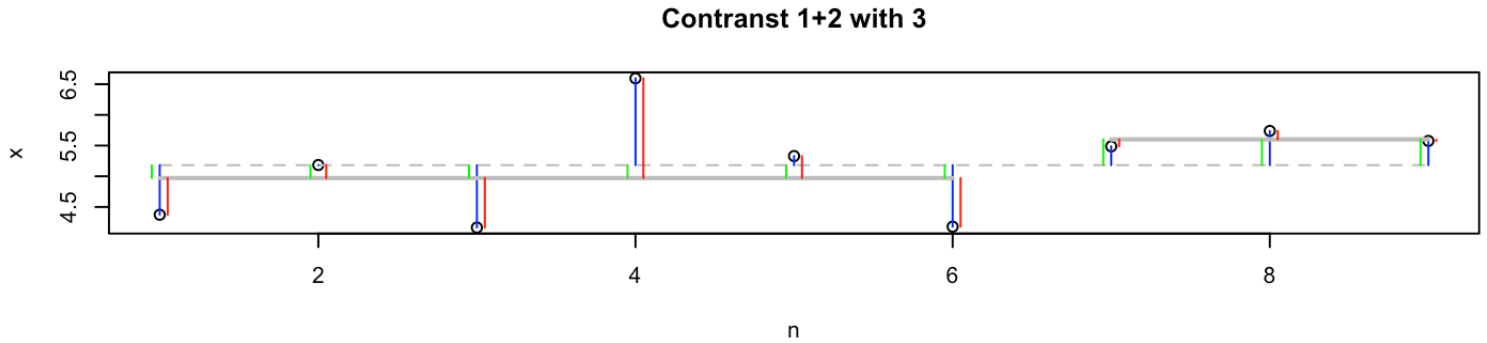
- AB-CDEF → A-B → CD-EF → C-D → E-F
- A-BCDEF → A-B → A-C
- A-BCDEG → BC-DEF → B-C → B-DEF
- ABC-DEF → BC-DEF → B-C

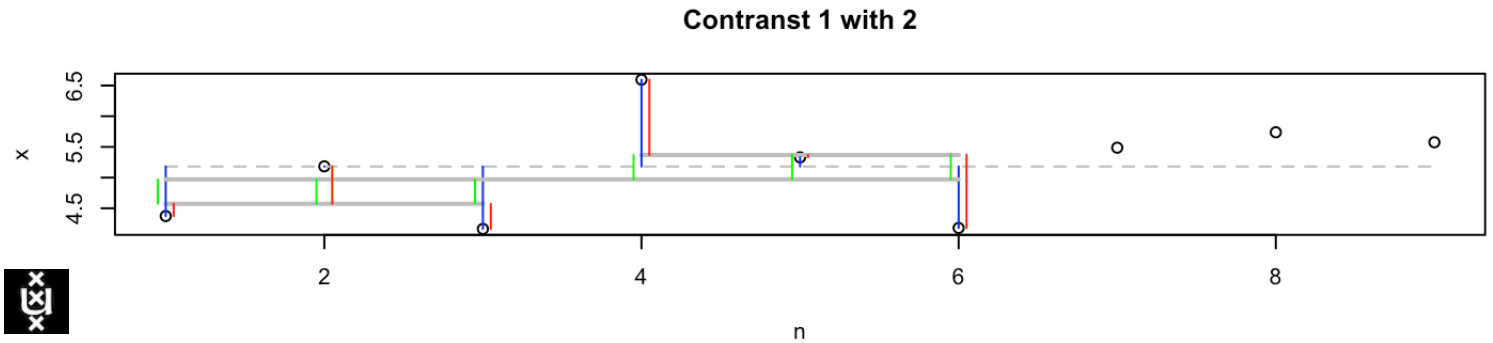
Assign values that combine to one. Same values define chunk.

- AB-CDEF → A-B → CD-EF → C-D → E-F

	A	B	C	D	E	F
Contrast 1						
Contrast 2						
Contrast 3						
Contrast 4						


variance explained
x
4.5
5.5
6.5
2
4
6
8
n


Contranst 1+2 with 3
x
4.5
5.5
6.5
2
4
6
8
n


Contranst 1 with 2
x
4.5
5.5
6.5
2
4
6
8
n

Post-hoc

Unplanned comparisons

- Exploring all possible differences
- Adjust T value for inflated type 1 error

	C	K	E
C		2	3
K	4		6
E	7	8	

Effect size η^2

The amount of explained variance R^2 as a general effect size measure.

$$R^2 = \frac{SS_{model}}{SS_{total}} = \eta^2$$

Taking the square root gives us Cohen's r.

Effect size ω^2

Less biased towards just the sample is omega squared ω^2.

$$\omega^2 = \frac{SS_{model} - (df_{model})MS_{error}}{SS_{total} + MS_{error}}$$

But what does it say?

Effect size r

A more interpretable effect size measure is $r_{Contrast}$. Which gives the effect size for a specific contrast.

$$r_{Contrast} = \sqrt{\frac{t^2}{t^2 + df}}$$

- @shklinkenberg
- Klinkenberg
- S.Klinkenberg@UvA.nl
- ShKlinkenberg

END