Formation R niveau 2

ANOVAs et ANCOVAs

B. Simon-Bouhet

La Rochelle Université

Avril 2023

What will we talk about?...

1. Comparing means of more than two groups
   One-factor ANOVA
   Assumptions and alternatives
   Post-hoc tests
   Fixed and random effects
   Two factors ANOVA
   Bloc design and nested factors

2. Introduction to linear models
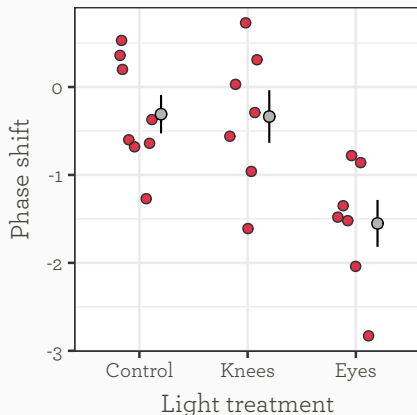   Linear regression
   ANCOVA: analysis of covariance

1. Comparing means of **more than two** groups

Outline

Some data to work with

Knees and jet lag

Does light treatment affect phase shift? Data from Wright & Czeisler (2002) in response to an earlier study by Campbell & Murphy (1998).

Treatment	$\overline{Y}$	s	n
Control	-0.309	0.618	8
Knees	-0.336	0.791	7
Eyes	-1.551	0.706	7

One-factor ANOVA
Knees and jet lag

> ### Definition
> Analysis of variance is the most powerful approach known for simultaneously testing whether the means of k groups are equal. It tests whether individuals from different groups are, on average, more different than individuals chosen from the same group.

1. State the hypotheses

- H_0: $\mu_1 = \mu_2 = \mu_3$.
- H_A: At least one mean is different from the others.

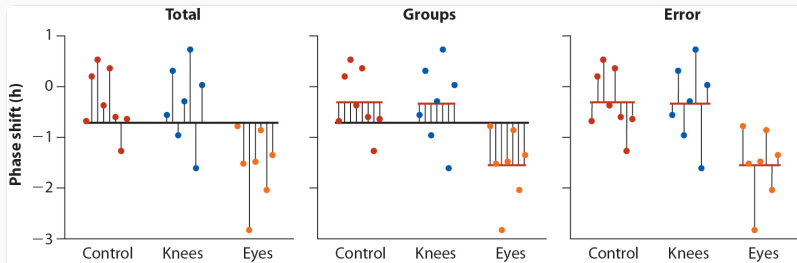> ### Important
> Rejecting H_0 does not mean that all means are different, but that at least one mean is different from the others.

One-factor ANOVA

Knees and jet lag

Source of variation	Sum of squares	df	Mean squares	F-ratio	P
Groups (treatment)	7.224	2	3.6122	7.29	0.004
Error	9.415	19	0.4955		
Total	16.639	21			

$$Y_{ij} - \overline{Y} = (Y_{ij} - \overline{Y}_i) + (\overline{Y}_i - \overline{Y})$$

One-factor ANOVA

Knees and jet lag

Here is what it looks like in R:

```
# Perform the ANOVA...
circadianAnova <- aov(shift ~ treatment, data = circadian)

# ... and print the ANOVA table
anova(circadianAnova)

Analysis of Variance Table

Response: shift
          Df Sum Sq Mean Sq F value   Pr(>F)
treatment  2 7.2245  3.6122  7.2894 0.004472 **
Residuals 19 9.4153  0.4955
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One-factor ANOVA

Knees and jet lag

Definition

R^2 measures the fraction of the variation in Y that is explained by group differences.

```
circadianAnovaSummary <- summary.lm(circadianAnova)
circadianAnovaSummary$r.squared

[1] 0.4341684
```

4. Draw the appropriate conclusions
Since $P < \alpha$, we reject H$_0$. At least one of the groups has a mean that is different from the others.

Outline

Assumptions of the ANOVA

The assumptions are the same as for the t-test, but they must hold for all k groups:

- ▶ The measurements in every group represent a random sample from the corresponding population.
- ▶ The variable is normally distributed in each of the k populations
- ▶ The variance is the same in all k populations

Very important!

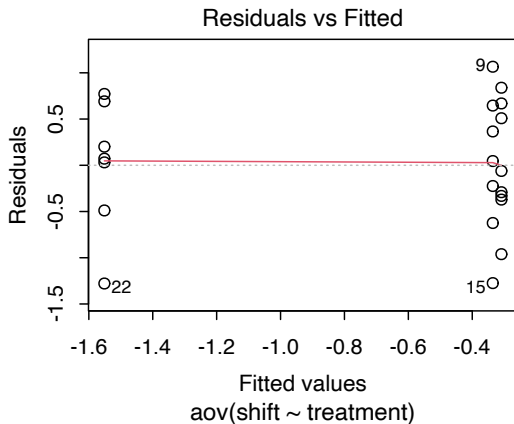In practice, contrary to all other tests described up to this point, the assumptions are verified :

- ▶ **after** the analysis is done.
- ▶ not on the data themselves, but on the **ANOVA residuals.**

This is true for **all linear models**.

Assumptions of the ANOVA

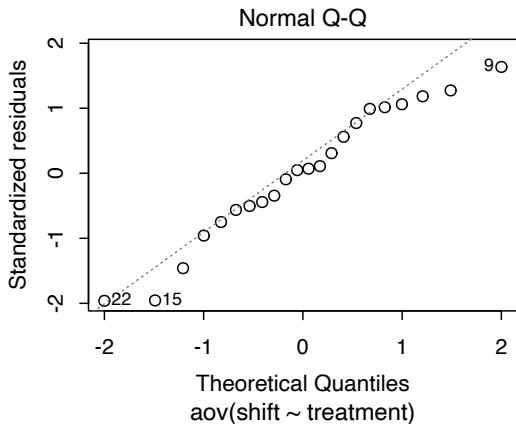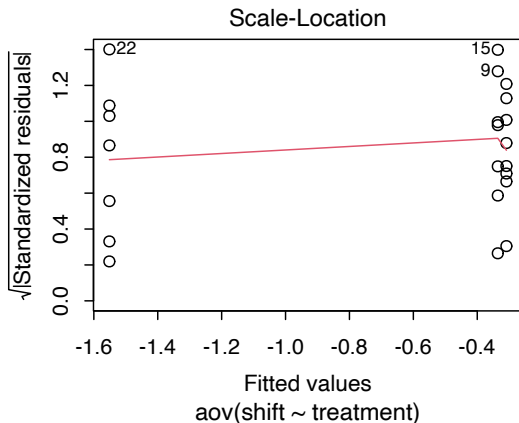## Checking the residuals

```
plot(circadianAnova)
```

Assumptions of the ANOVA

Checking the residuals

```
plot(circadianAnova)
```

Assumptions of the ANOVA

Checking the residuals

```
plot(circadianAnova)
```

Assumptions of the ANOVA

Checking the residuals

```
plot(circadianAnova)
```
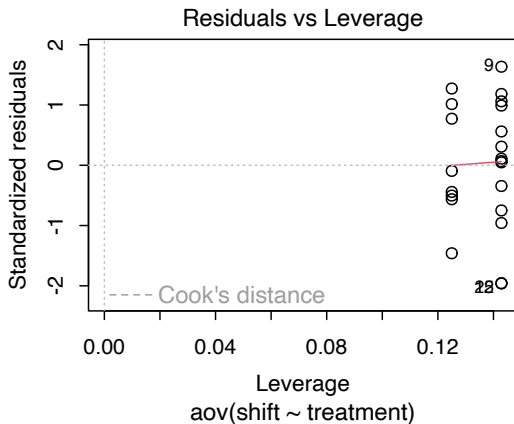
What if assumptions are not met?

> **Definition**
>
> The Kruskal-Wallis test is the nonparametric equivalent to the one-factor ANOVA. It should be used when data (*i.e.* residuals) are **not normally distributed**.

```
kruskal.test(shift ~ treatment, data = circadian)


^^IKruskal-Wallis rank sum test

data:  shift by treatment
Kruskal-Wallis chi-squared = 9.4231, df = 2, p-value
= 0.008991
```

Outline

We rejected H$_0$. Now what?

Differences between groups

> **Definition**
>
> Post-hoc tests perform all possible pairwise comparisons of means and apply corrections so that the type I error does not inflate.

We rejected H$_0$. Now what?

Differences between groups

Definition

Post-hoc tests perform all possible pairwise comparisons of means and apply corrections so that the type I error does not inflate.

```
library(broom)
tukey <- tidy(TukeyHSD(circadianAnova))
tukey

# A tibble: 3 x 7
  term      contrast      null.value estimate conf.low conf.high adj.p.value
  <chr>     <chr>              <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
1 treatment Knees-Control          0  -0.0270   -0.953     0.899       0.997
2 treatment Eyes-Control           0  -1.24     -2.17     -0.317       0.00787
3 treatment Eyes-Knees             0  -1.22     -2.17     -0.260       0.0117
```

We rejected H$_0$. Now what?

Differences between groups

> **Definition**
>
> Post-hoc tests perform all possible pairwise comparisons of means and apply corrections so that the type I error does not inflate.

```
tukey %>%
  ggplot(aes(x = estimate,
             y = contrast)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low,
                     xmax = conf.high),
                 height = 0.1) +
  geom_vline(xintercept = 0,
             linetype = 2) +
  labs(x = "Differences in means",
       y = "",
       title = "95% confidence level")
```
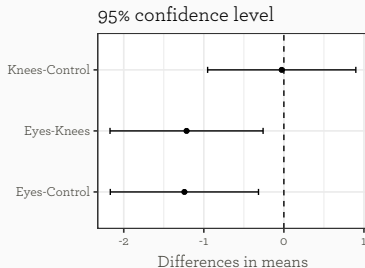
We rejected H$_0$. Now what?

Differences between groups

The pairwise t-test:

```
pairwise.t.test(circadian$shift, g=circadian$treatment)


^^IPairwise comparisons using t tests with pooled SD

data:  circadian$shift and circadian$treatment


      Control Knees
Knees 0.9418  -
Eyes  0.0088  0.0088


P value adjustment method: holm
```

We rejected H_0. Now what?

Differences between groups

The pairwise Wilcoxon test:

```
pairwise.wilcox.test(circadian$shift, g=circadian$treatment)


^^IPairwise comparisons using Wilcoxon rank sum exact test

data:  circadian$shift and circadian$treatment


      Control Knees
Knees 0.9551  -
Eyes  0.0037  0.0524


P value adjustment method: holm
```

Outline

ANOVA with random effects

> ## Definition
>
> An explanatory variable is called a **fixed effect** if the groups are predetermined and are of direct interest.
> An explanatory variable is called a **random effect** if the groups are randomly sampled from a population of possible groups.

femur length

Data from Nosil & Crespi (2006).

ANOVA with random effects
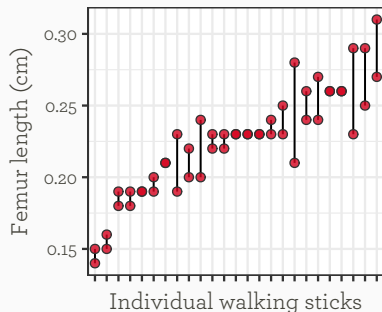
Walking-stick limb


```
walkingstick

# A tibble: 50 x 2
   specimen femurLength
   <fct>          <dbl>
 1 1               0.26
 2 1               0.26
 3 2               0.23
 4 2               0.19
 5 3               0.25
 6 3               0.23
 7 4               0.26
 8 4               0.26
 9 5               0.23
10 5               0.22
# i 40 more rows
```

```
dim(walkingstick)

[1] 50  2
```

ANOVA with random effects

Walking-stick limb

In R formulae, the `Error()` function is used to indicate random effects.

```
stickaov <- aov(femurLength ~ Error(specimen), data = walkingstick)
summary(stickaov)


Error: specimen
          Df  Sum Sq  Mean Sq F value Pr(>F)
Residuals 24 0.05913 0.002464

Error: Within
          Df Sum Sq  Mean Sq F value Pr(>F)
Residuals 25 0.0089 0.000356
```

Here, we don't have any P-value. Since we only have one variable of interest, our goal is to estimate the variance components: which part of the total variation in femur length is due to measurement errors and which part is due to real differences among individuals?

Random effects: variance components

> ### Definition
>
> In a random effect ANOVA, σ^2 and σ_A^2 are called variance components. They describe all the variance in the response variable Y:
>
> - σ^2 describe the variance within groups (i.e. here, the measurement error).
> - σ_A^2 describe the variance among groups (i.e. here, the differences between the true femur lengths of individual insects).

$$s_A^2 = \frac{\text{MS}_{\text{groups}} - \text{MS}_{\text{error}}}{n} = \frac{0.002464 - 0.000356}{2} = 0.00105\,\text{cm}^2$$

where n is the number of measurements taken within each group.

Random effects: variance components

Walking-stick limb

> ### Definition
>
> Repeatability measures the overall similarity of repeat measurements made on the same group.
>
> - When repeatability ≈ 0, all variance in the response variable comes from differences between separate measurements made on the same group.
> - When repeatability ≈ 1, repeated measurements on the same group give nearly the same answer every time.

Once we have s_A^2, we can calculate the repeatability:

$$\text{Repeatability} = \frac{s_A^2}{s_A^2 + \text{MS}_{\text{error}}} = \frac{0.00105}{0.00105 + 0.000356} = 0.75$$

Outline

An example with 2 fixed factors

Academic grades, levels and teaching method

Grades of undergrad and grad students depending of the teaching methods they have been exposed to.

	Teach.meth.1	Teach.meth.2
Undergrad	8.00	6.00
	11.00	7.00
	8.00	8.00
	9.00	7.00
Grad	16.00	12.00
	18.00	11.00
	17.00	12.00
	17.00	9.00

An example with 2 fixed factors

Here, we have 3 sets of hypotheses

- ▶ H_0: The level of students has no effect on the grades ($\mu_{\text{undergrad}} = \mu_{grad}$).
- ▶ H_A: The level of students has an effect on the grades ($\mu_{\text{undergrad}} \neq \mu_{grad}$).

- ▶ H_0: The teaching method has no effect on the grades ($\mu_{\text{method 1}} = \mu_{method2}$).
- ▶ H_A: The teaching method has an effect on the grades ($\mu_{\text{method 1}} \neq \mu_{method2}$)

- ▶ H_0: The teaching methods have the same influence on both groups of students.
- ▶ H_A: The teaching methods do not have the same influence on both groups of students.

An example with 2 fixed factors

Data format

```
   Grades Teaching      Level
1       8 Method.1 Undergrad
2      11 Method.1 Undergrad
3       8 Method.1 Undergrad
4       9 Method.1 Undergrad
5      16 Method.1      Grad
6      18 Method.1      Grad
7      17 Method.1      Grad
8      17 Method.1      Grad
9       6 Method.2 Undergrad
10      7 Method.2 Undergrad
11      8 Method.2 Undergrad
12      7 Method.2 Undergrad
13     12 Method.2      Grad
14     11 Method.2      Grad
15     12 Method.2      Grad
16      9 Method.2      Grad
```

An example with 2 fixed factors
The test

```
res <- aov(Grades ~ Teaching * Level, data = dat)
summary(res)

                Df Sum Sq Mean Sq F value   Pr(>F)
Teaching         1     64   64.00      48 1.59e-05 ***
Level            1    144  144.00     108 2.36e-07 ***
Teaching:Level   1     16   16.00      12  0.00468 **
Residuals       12     16    1.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
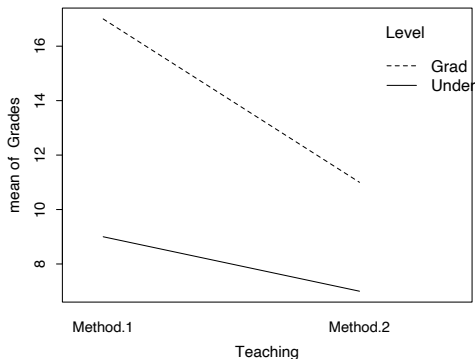
The 3 null hypotheses are rejected.

An example with 2 fixed factors

Interaction plot

```
attach(dat)
interaction.plot(Teaching, Level, Grades)
```

An example with 2 fixed factors
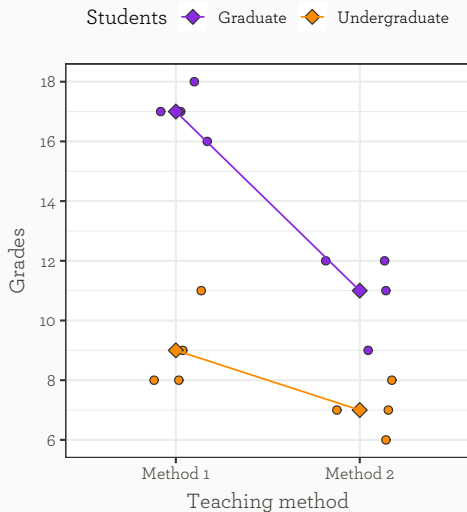
Interaction plot

An example with 2 fixed factors

Interaction plots
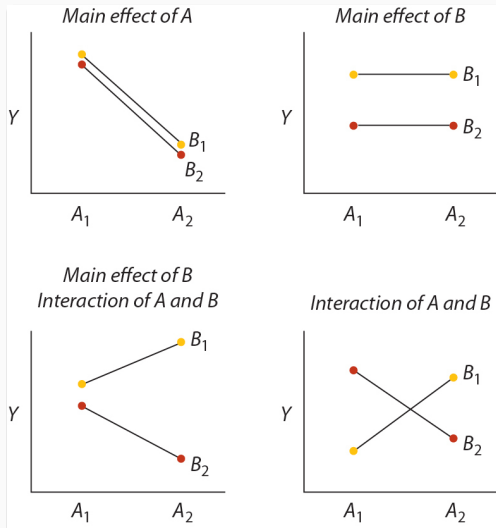
ANOVA and formulae

In R, the formulae syntax is not always very intuitive…

We saw that "`Error()`" is used to specify random effects.

The ":" operator is used to indicate an interaction term.

The "*" argument is used to specify a model with both the main effects and the interaction term.

The following commands specify the same model:

```
aov(Grades ~ Teaching * Level, data = dat)
aov(Grades ~ Teaching + Level + Teaching:Level, data = dat)
```

Two factors ANOVA: mixed effects
Zooplankton depredation

Data from Svanbäck & Bolnick (2007).

Treatment	1	2	3	4	5
control	4.1	3.2	3.0	2.3	2.5
low	2.2	2.4	1.5	1.3	2.6
high	1.3	2.0	1.0	1.0	1.6

Treatment	Diversity	Block
control	4.1	1
low	2.2	1
high	1.3	1
control	3.2	2
low	2.4	2
high	2.0	2
control	3.0	3
low	1.5	3
high	1.0	3
control	2.3	4
low	1.3	4
high	1.0	4
control	2.5	5
low	2.6	5
high	1.6	5

Two factors ANOVA: mixed effects

ANOVA the **wrong** way

```
model1 <- aov(diversity~treatment*block, data = zoopk)
summary(model1)

                Df Sum Sq Mean Sq
treatment        2  6.857   3.429
block            4  2.340   0.585
treatment:block  8  1.676   0.209
```

Two factors ANOVA: mixed effects
ANOVA the right way

```
model2 <- aov(diversity~treatment + Error(block), data = zoopk)
summary(model2)


Error: block
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4   2.34   0.585

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
treatment  2  6.857   3.429   16.37 0.00149 **
Residuals  8  1.676   0.209
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two factors ANOVA: mixed effects

Nonparametric alternative

For this specific design, i.e. two factors ANOVA with mixed effects, when:

- ▶ Normality is not met
- ▶ Data are not replicated

Two nonparametric tests can be used instead of the ANOVA.

- ▶ The Friedman rank sum test with unreplicated blocked data: `friedman.test()`.
- ▶ The Quade test with unreplicated blocked data: `quade.test()`.

- ▶ The Quade test is more powerful than the Friedman test when there are few treatments (3-4).
- ▶ Both tests are equally powerful when there are 5 or 6 treatments
- ▶ The Friedman test is more powerful than the Quade test when there are more than 6 treatments.

Two factors ANOVA: mixed effects

The assumption of sphericity

Treatment	Block				
	1	2	3	4	5
control	4.1	3.2	3.0	2.3	2.5
low	2.2	2.4	1.5	1.3	2.6
high	1.3	2.0	1.0	1.0	1.6

Two factors ANOVAs with mixed effects can be considered as one factor ANOVAs with repeated measurements.

Here, each `Block` provides one measurement for each level of the only variable of interest (`treatment`).

Two factors ANOVA: mixed effects

The assumption of **sphericity**

	Treatment			Differences		
Block	control	low	high	ctrl − low	ctrl − high	low − high
1	4.1	2.2	1.3	1.9	2.8	0.9
2	3.2	2.4	2.0	0.8	1.2	0.4
3	3.0	1.5	1.0	1.5	2.0	0.5
4	2.3	1.3	1.0	1.0	1.3	0.3
5	2.5	2.6	1.6	-0.1	0.9	1.0
			Variance	0.577	0.583	0.097

Definition

Sphericity is met when the differences between all possible pairs of within-subject conditions have equal variance. This is a **very important** condition of repeated-measures ANOVA.

Two factors ANOVA: mixed effects

The assumption of sphericity

The most frequently used test to verify the assumption of sphericity is Mauchly's test (`mauchlys.test`).

> **Important**
>
> Mauchly's test of sphericity should be avoided most of the times !

In practice, we multiply df by $\dfrac{1}{k-1}$, where k is the number of repeated measures (here, 3).

Hence:

```
1 - pf(16.37, df1 = 2 * 0.5, df2 = 8 * 0.5)
[1] 0.01552707
```

Two factors ANOVA: repeated measures

Actually, three factors with mixed effects

Here, each subject (random factor) provides a measurement for each combination of the two fixed factors of interest.

Subjects	Bg low		Bg high	
	Stim low	Stim high	Stim low	Stim high
S1	12	18	20	8
S2	9	20	24	10
S3	9	22	16	9
S4	10	22	18	11
S5	9	17	18	10
S6	11	23	22	12

Repeated measures ANOVA is way more powerfull than independent ANOVA because one source of variation (the subject error) can be isolated from the main and interaction effects.

Two factors ANOVA: repeated measures
Hypotheses

Here, we have 3 sets of hypotheses:

- H_0: background noise *has no effect* on the ability to detect the stimuli.
- H_A: background noise *has an effect* on the ability to detect the stimuli.

- H_0: the frequency of the stimulus *has no effect* on its detectability.
- H_A: the frequency of the stimulus *has an effect* on its detectability.

- H_0: the two types of background noises *have the same effect* on the detectability of the stimuli.
- H_A: the two types of background noises *do not have the same effect* on the detectability of the stimuli.

Two factors ANOVA: repeated measures
Data format

Data must be presented in the "long format".

	Score	Noise	Stimulus	Subject
1	12	Low	Low	Subject.1
2	9	Low	Low	Subject.2
3	9	Low	Low	Subject.3
4	10	Low	Low	Subject.4
5	9	Low	Low	Subject.5
6	11	Low	Low	Subject.6
7	18	Low	High	Subject.1
8	20	Low	High	Subject.2
9	22	Low	High	Subject.3
10	22	Low	High	Subject.4
11	17	Low	High	Subject.5
12	23	Low	High	Subject.6

	Score	Noise	Stimulus	Subject
13	20	High	Low	Subject.1
14	24	High	Low	Subject.2
15	16	High	Low	Subject.3
16	18	High	Low	Subject.4
17	18	High	Low	Subject.5
18	22	High	Low	Subject.6
19	8	High	High	Subject.1
20	10	High	High	Subject.2
21	9	High	High	Subject.3
22	11	High	High	Subject.4
23	10	High	High	Subject.5
24	12	High	High	Subject.6

Two factors ANOVA: repeated measures
The test

```
model <- aov(Score ~ Noise * Stimulus
    + Error(Subject / (Noise * Stimulus)), data = hearing)
summary(model)
```

Two factors ANOVA: repeated measures

```
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  5   32.5     6.5

Error: Subject:Noise
          Df Sum Sq Mean Sq F value Pr(>F)
Noise      1  0.667   0.667   0.177  0.691
Residuals  5 18.833   3.767

Error: Subject:Stimulus
          Df Sum Sq Mean Sq F value Pr(>F)
Stimulus   1  0.667   0.667   0.124  0.739
Residuals  5 26.833   5.367

Error: Subject:Noise:Stimulus
               Df Sum Sq Mean Sq F value   Pr(>F)
Noise:Stimulus  1  600.0   600.0     240 2.04e-05 ***
Residuals       5   12.5     2.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two factors ANOVA: repeated measures
Conclusions

▶ H_0: background noise *has no effect* on the ability to detect the stimuli.

▶ H_A: background noise *has an effect* on the ability to detect the stimuli.

▶ H_0: the frequency of the stimulus *has no effect* on its detectability.

▶ H_A: the frequency of the stimulus *has an effect* on its detectability.

▶ H_0: the two types of background noises *have the same effect* on the detectability of the stimuli.

▶ H_A: the two types of background noises *do not have the same effect* on the detectability of the stimuli.

Two factors ANOVA: repeated measures
Conclusions

▶ We have **no evidence** that background noise has an effect.
▶ We have **no evidence** that the frequency of the stimuli has an effect.
▶ We have **evidence** ($P < 0.01$) that the two types of background noises do not have the same effect on the detectability of the stimuli.

Outline

ANOVA 3 nested factors

Experimental design

ANOVA 3 nested factors

Experimental design

ANOVA 3 nested factors

Data format

```
yields <- read_delim("data/splityield.txt", "\t",
                     escape_double = FALSE, trim_ws = TRUE) %>%
         mutate_if(is.character, as_factor)
yields

# A tibble: 72 x 5
   yield block irrigation density fertilizer
   <dbl> <fct> <fct>      <fct>   <fct>
 1    90 A     control    low     N
 2    95 A     control    low     P
 3   107 A     control    low     NP
 4    92 A     control    medium  N
 5    89 A     control    medium  P
 6    92 A     control    medium  NP
 7    81 A     control    high    N
 8    92 A     control    high    P
 9    93 A     control    high    NP
10    80 A     irrigated  low     N
# i 62 more rows
```

ANOVA 3 nested factors

Data format

```
model <- aov(yield ~ irrigation * density * fertilizer +
             Error(block / irrigation / density / fertilizer),
             data = yields)
summary(model)
```

ANOVA 3 nested factors

Data format

```
Error: block
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  3  194.4   64.81

Error: block:irrigation
           Df Sum Sq Mean Sq F value Pr(>F)
irrigation  1   8278    8278   17.59 0.0247 *
Residuals   3   1412     471
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: block:irrigation:density
                   Df Sum Sq Mean Sq F value Pr(>F)
density             2   1758   879.2   3.784 0.0532 .
irrigation:density  2   2747  1373.5   5.912 0.0163 *
Residuals          12   2788   232.3
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA 3 nested factors

Data format
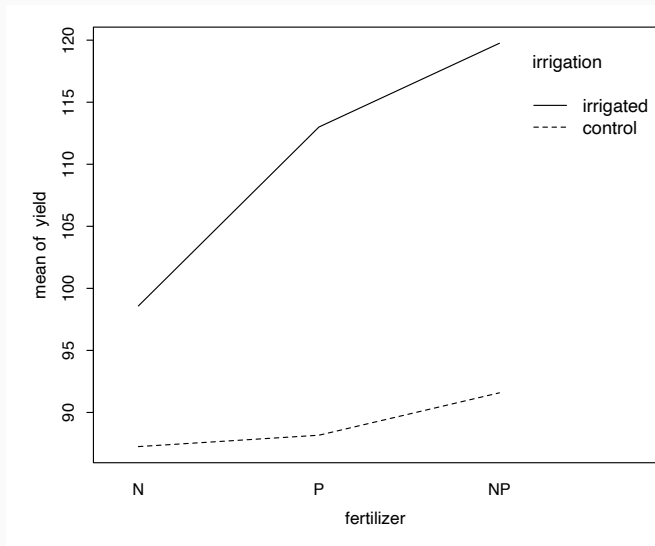
```
Error: Within
                             Df Sum Sq Mean Sq F value    Pr(>F)
fertilizer                    2 1977.4   988.7  11.449  0.000142 ***
irrigation:fertilizer         2  953.4   476.7   5.520  0.008108 **
density:fertilizer            4  304.9    76.2   0.883  0.484053
irrigation:density:fertilizer 4  234.7    58.7   0.680  0.610667
Residuals                    36 3108.8    86.4
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA 3 nested factors
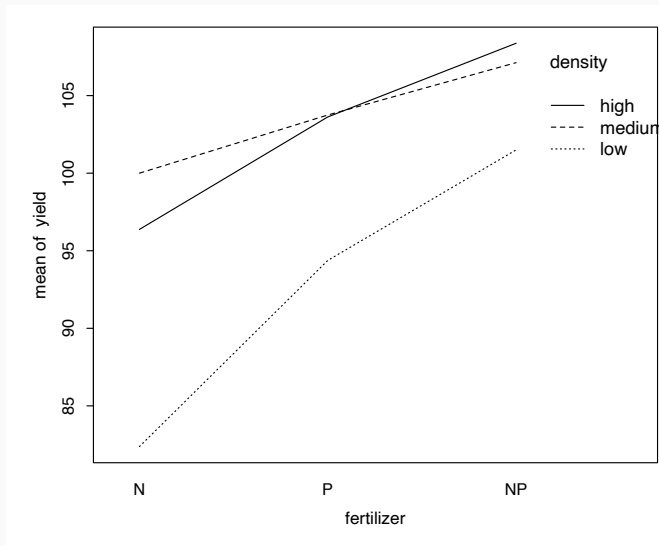
Understanding significant interactions

ANOVA 3 nested factors

Understanding significant interactions

ANOVA 3 nested factors

Understanding significant interactions

2. Introduction to linear models
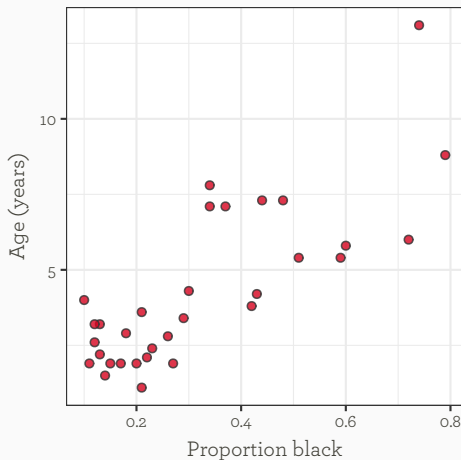
Outline

Simple linear regression
The lion's noses

Data from Whitman *et al.* (2004).

```
# A tibble: 32 x 2
   Prop.black    Age
        <dbl>  <dbl>
 1       0.21    1.1
 2       0.14    1.5
 3       0.11    1.9
 4       0.13    2.2
 5       0.12    2.6
 6       0.13    3.2
 7       0.12    3.2
 8       0.18    2.9
 9       0.23    2.4
10       0.22    2.1
# i 22 more rows
```

Simple linear regression

Specifying the model

```
lionReg <- lm(Age ~ Prop.black, data = lion)
lionReg


Call:
lm(formula = Age ~ Prop.black, data = lion)

Coefficients:
(Intercept)    Prop.black
      0.879        10.647

confint(lionReg)

                 2.5 %      97.5 %
(Intercept) -0.2826733    2.040686
Prop.black   7.5643082   13.729931
```

Simple linear regression
Testing the slope and intercept

Here, we have 2 sets of hypotheses:

- ▶ H_0: the intercept of the linear regression is equal to zero.
- ▶ H_A: the intercept of the linear regression is different from zero.

- ▶ H_0: the slope of the linear regression is equal to zero.
- ▶ H_A: the slope of the linear regression is different from zero.

Confidence intervals gave us a first clue but we can test these hypotheses with the `lm()` function.

Simple linear regression
Testing the slope and intercept

```
summary(lionReg)


Call:
lm(formula = Age ~ Prop.black, data = lion)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8790     0.5688   1.545    0.133
Prop.black   10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared:  0.6238,^^IAdjusted R-squared:  0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Simple linear regression

Confidence and prediction bands

```r
# Create a vector of regularly spaced values
x <- seq(from = min(lion$Prop.black),
         to = max(lion$Prop.black),
         length.out = 100)

# Calculate confidence interval
conf <- as_tibble(predict(object = lionReg,
                          newdata = data.frame(Prop.black = x),
                          interval = "confidence"))
names(conf) <- paste0("ci_", names(conf))

# Calculate prediction interval
pred <- as_tibble(predict(object = lionReg,
                          newdata = data.frame(Prop.black = x),
                          interval = "prediction"))
names(pred) <- paste0("pr_", names(pred))

# Merge in a unique tibble
bands <- bind_cols(tibble(x), conf, pred)
```

Simple linear regression
Confidence and prediction bands

```
bands

# A tibble: 100 x 7
        x ci_fit ci_lwr ci_upr pr_fit pr_lwr pr_upr
    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 0.1      1.94   1.03   2.86   1.94  -1.58   5.47
 2 0.107    2.02   1.12   2.91   2.02  -1.51   5.54
 3 0.114    2.09   1.21   2.97   2.09  -1.43   5.61
 4 0.121    2.17   1.30   3.03   2.17  -1.35   5.68
 5 0.128    2.24   1.39   3.09   2.24  -1.27   5.75
 6 0.135    2.31   1.48   3.15   2.31  -1.19   5.82
 7 0.142    2.39   1.57   3.21   2.39  -1.12   5.89
 8 0.149    2.46   1.66   3.27   2.46  -1.04   5.97
 9 0.156    2.54   1.75   3.33   2.54  -0.961  6.04
10 0.163    2.61   1.83   3.39   2.61  -0.884  6.11
# i 90 more rows
```

Simple linear regression

Plotting confidence and prediction bands

```r
# Create a scatter plot and add a regression line
pl <- lion %>%
  ggplot(aes(x = Prop.black, y = Age)) +
    geom_point(shape = 21,
               color = grey(0.2),
               fill = rgb(0.84, 0.01, 0.12, 0.8),
               alpha = 0.8) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Proportion black",
         y = "Age (years)")
```
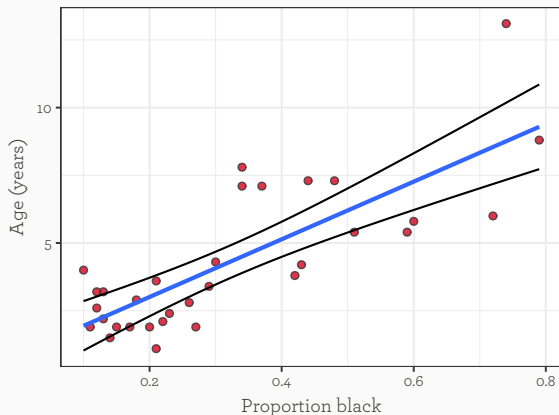
Simple linear regression

Plotting confidence and prediction bands

pl

Simple linear regression

Plotting confidence and prediction bands
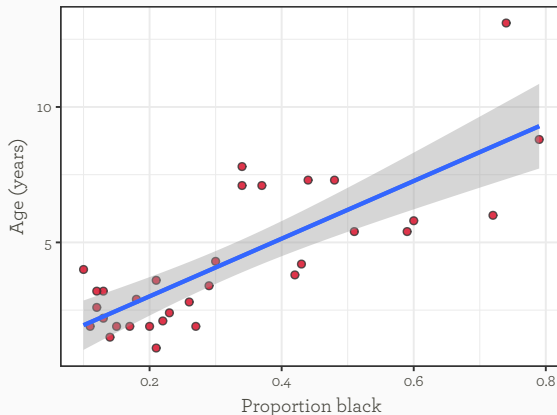
```
# Add confidence interval
pl + geom_line(data = bands, aes(x = x, y = ci_lwr)) +
    geom_line(data = bands, aes(x = x, y = ci_upr))
```

Simple linear regression

Plotting confidence and prediction bands

```
# Add confidence interval. Equivalent to :
pl + geom_smooth(method = "lm", se = TRUE)
```
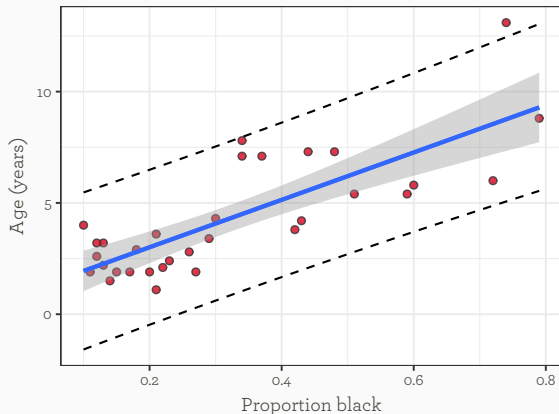
Simple linear regression

Plotting confidence and prediction bands

```
# Add prediction interval.
pl + geom_smooth(method = "lm", se = TRUE) +
    geom_line(data = bands, aes(x = x, y = pr_lwr), linetype = 2) +
    geom_line(data = bands, aes(x = x, y = pr_upr), linetype = 2)
```

Outline

Mixing regression and ANOVA

> **Definition**
>
> The Analysis of Covariance examines the influence of at least two explanatory variables on a single numerical response variable. The explanatory variables are a mix of continuous and categorical factors.

In 2011, Caurant *et al.* studied the relationship between the bioaccumulation of cadmium in the kidney of 244 common dolphins (*Delphinus delphis*) of various ages, coming from 5 distinct geographic regions.
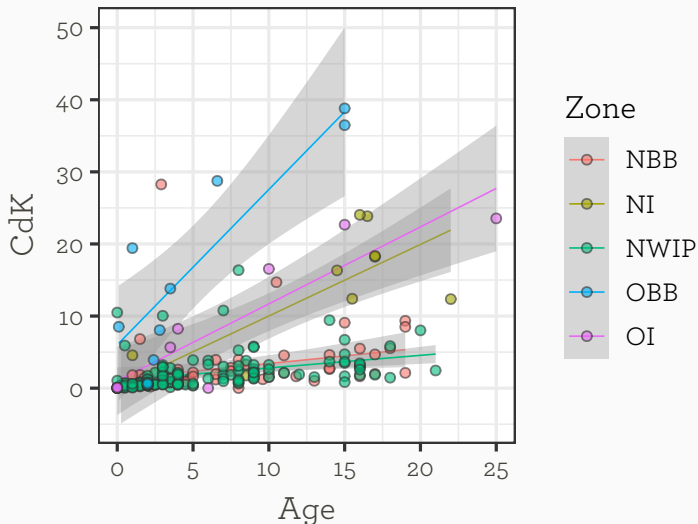
Mixing regression and ANOVA
Definition and example data

```
dauphin
# A tibble: 244 x 3
   Zone    Age    CdK
   <chr> <dbl>  <dbl>
 1 NBB    10.5  14.7
 2 NBB    14     2.64
 3 NBB     4.8   1.34
 4 NBB     0.2   0.31
 5 NBB     3     1.66
 6 NBB     1.5   1.83
 7 NBB     7.5   2.83
 8 NBB     1.8   1.43
 9 NBB    17     1.98
10 NBB     2.2   0.22
# i 234 more rows
```

Mixing regression and ANOVA

Definition and example data

Mixing regression and ANOVA

Understanding the results

```
# Performing the ANCOVA
ancov <- lm(CdK ~ Zone * Age, data = dauphin)

# Global results
anova(ancov)

Analysis of Variance Table

Response: CdK
           Df Sum Sq Mean Sq F value    Pr(>F)
Zone        4 2869.7  717.43  59.936 < 2.2e-16 ***
Age         1 1237.3 1237.29 103.367 < 2.2e-16 ***
Zone:Age    4 1569.7  392.43  32.785 < 2.2e-16 ***
Residuals 207 2477.8   11.97
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mixing regression and ANOVA

Understanding the results

```
summary(ancov)$coefficients

               Estimate Std. Error    t value      Pr(>|t|)
(Intercept)   1.28813839 0.68752959  1.8735752 6.239819e-02
ZoneNI       -1.20125790 1.75501252 -0.6844726 4.944424e-01
ZoneNWIP     -0.23286289 0.84014649 -0.2771694 7.819266e-01
ZoneOBB       4.71584521 1.76444627  2.6727055 8.123839e-03
ZoneOI       -0.28732736 1.68968591 -0.1700478 8.651386e-01
Age           0.21222491 0.07696870  2.7572886 6.349305e-03
ZoneNI:Age    0.78025773 0.14784391  5.2775777 3.293708e-07
ZoneNWIP:Age -0.03738055 0.09726805 -0.3843045 7.011473e-01
ZoneOBB:Age   1.94271615 0.22636524  8.5822193 2.215078e-15
ZoneOI:Age    0.85589636 0.16451143  5.2026558 4.713221e-07
```

Mixing regression and ANOVA

Understanding the results

```r
# Filter out NBB
dauphin2 <- dauphin %>%
  filter(Zone != "NBB")

# Perform new ANCOVA
ancov2 <- lm(CdK ~ Zone * Age, data = dauphin2)

# Print the results
summary(ancov2)$coefficients
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)   0.08688049  1.4826028  0.05859998 9.533509e-01
ZoneNWIP      0.96839501  1.5474703  0.62579229 5.324282e-01
ZoneOBB       5.91710311  2.1033812  2.81313874 5.582104e-03
ZoneOI        0.91393054  2.0509786  0.44560706 6.565412e-01
Age           0.99248264  0.1158993  8.56331803 1.408667e-14
ZoneNWIP:Age -0.81763828  0.1281186 -6.38188470 2.183300e-09
ZoneOBB:Age   1.16245842  0.2272369  5.11562423 9.681897e-07
ZoneOI:Age    0.07563863  0.1767889  0.42784719 6.693928e-01
```

References

Campbell SS, Murphy PJ (1998) Extraocular circadian phototransduction in humans. *Science*, **279**, 396–399.

Nosil P, Crespi BJ (2006) Experimental evidence that predation promotes divergence in adaptive radiation. *Proceedings of the National Academy of Sciences*, **103**, 9090–9095.

Svanbäck R, Bolnick D (2007) Intraspecific competition drives increased resource use diversity within a natural population. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **274**, 839–844.

Whitman K, Starfield AM, Quadling HS, Packer C (2004) Sustainable trophy hunting of african lions. *Nature*, **428**, 175–178.

Wright KPJ, Czeisler CA (2002) Absence of circadian phase resetting in response to bright light behind the knees. *Science*, **297**, 571.