EPFL

Fundamentals in Ecology Practicals

Week 8

Statistical analyses in R

Christoph Bachofen

Aim of the course

At the end of this lecture you should know:

1. Why we should do statistical analyses
2. What different kind of analyses there are
3. How to chose which analysis to do
4. How to do the analyses well

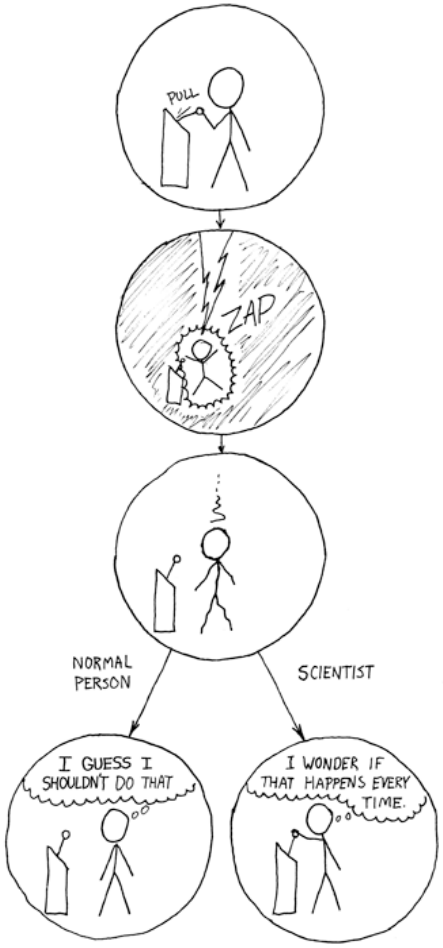
Why should we do statistical analyses?

Goal: understand an underlying **general pattern or relationship** we observed in nature

For instance:

"Plant leaves acclimated to shade have a higher leaf area and lower leaf thickness"

→ Is this generally true (across all leaves on the planet)?



Why should we do statistical analyses?

Goal: understand an underlying **general pattern or relationship** we observed in nature

For instance:

"Plant leaves acclimated to shade have a higher leaf area and lower leaf thickness"

→ Sample the environment

shade leaf

vs.

sun leaf

Why should we do statistical analyses?

Goal: understand an underlying **general pattern or relationship** we observed in nature

For instance:

"Plant leaves a
area and lowe

→ Sample the environment

Problems:

1. We have variability in nature
2. We cannot measure all plants (too many!)

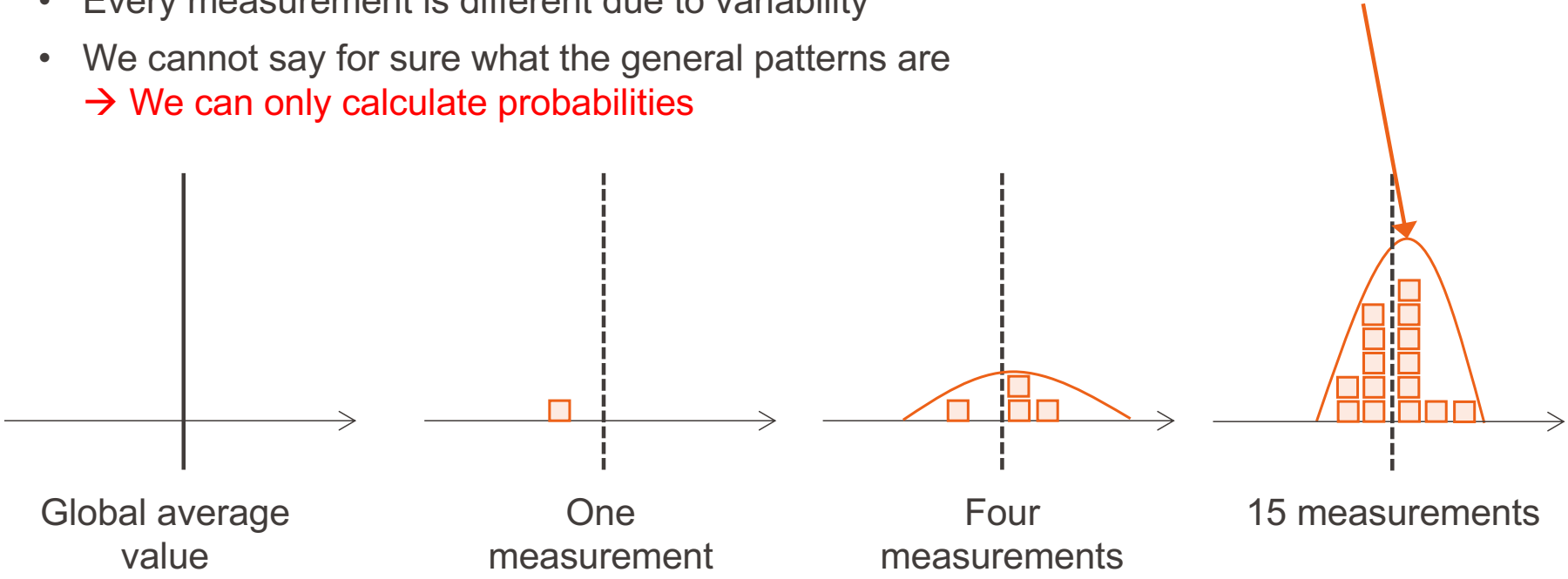
shade leaf

vs.

sun leaf

Why should we do statistical analyses?

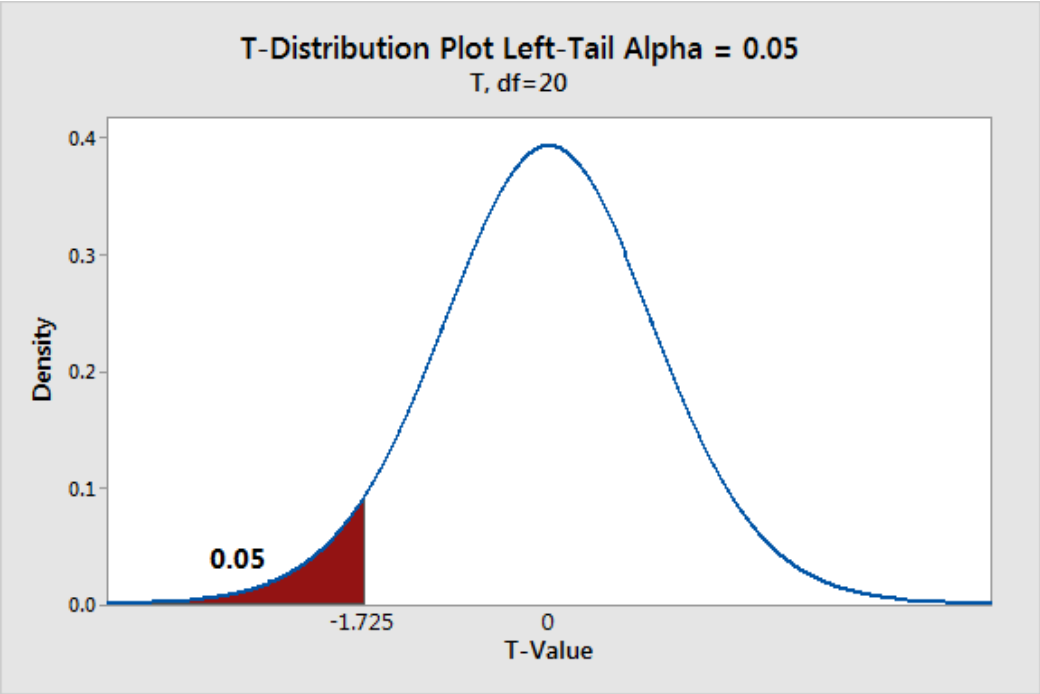
- **Is our sample representative of the general pattern?**
- Every measurement is different due to variability
- We cannot say for sure what the general patterns are
  → We can only calculate probabilities



The more measurements we do, the higher the probability that we get the global pattern

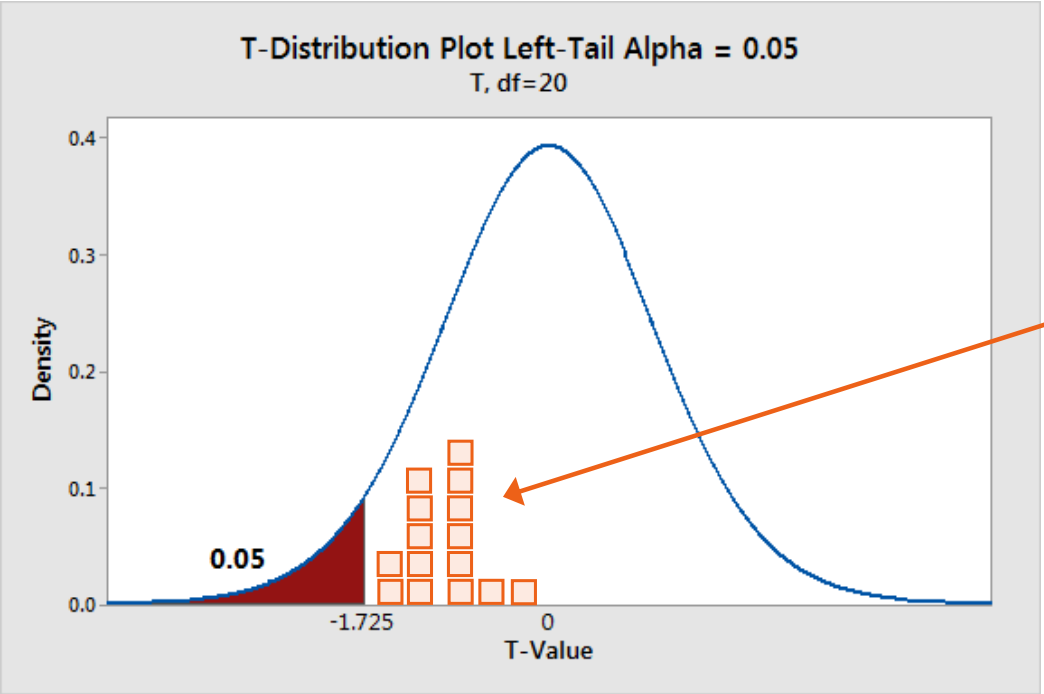
Why should we do statistical analyses?

Frequentist stats: assume an underlying distribution of your data (e.g. t-distribution) to calculate where your measurements are within the distribution area.



Why should we do statistical analyses?

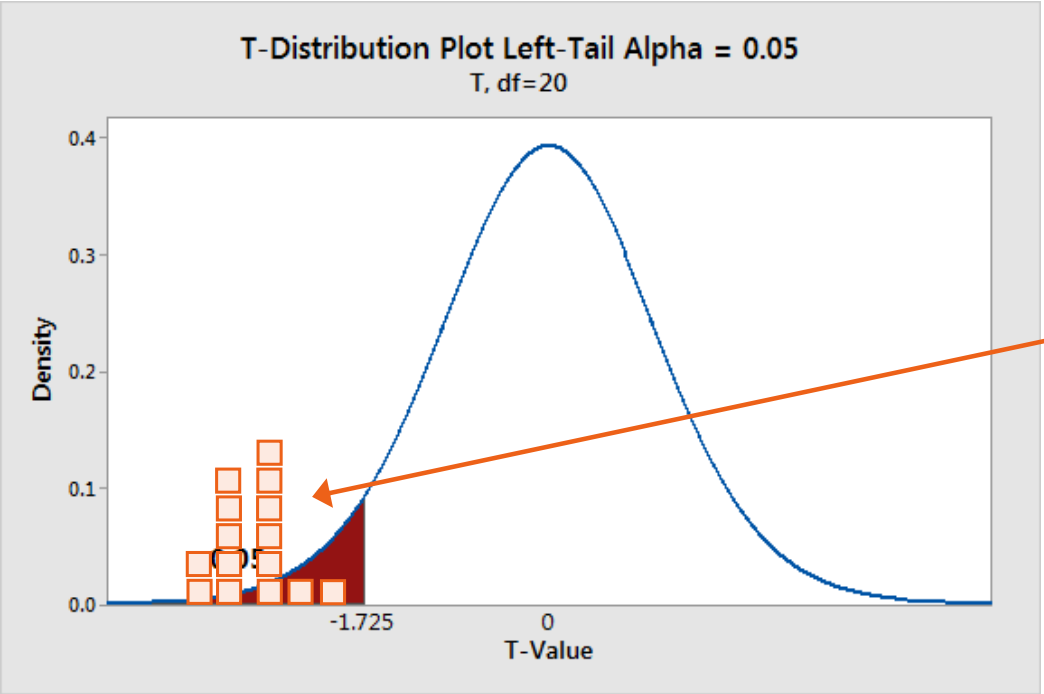
Frequentist stats: assume an underlying distribution of your data (e.g. t-distribution) to calculate where your measurements are within the distribution area.



Your data falls **within** 95% of the area of the t-distribution. The hypothesis that your measurements are different from 0 cannot be rejected

Why should we do statistical analyses?

Frequentist stats: assume an underlying distribution of your data (e.g. t-distribution) to calculate where your measurements are within the distribution area.



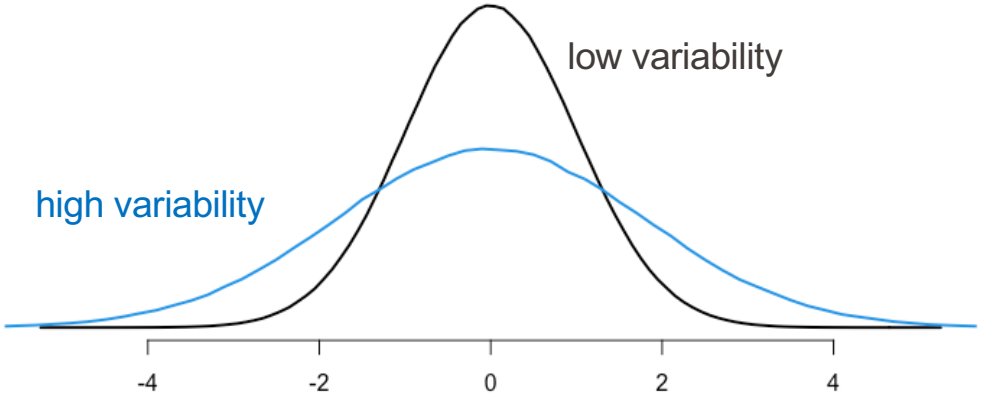
Your data falls **outside** 95% of the area of the t-distribution. The hypothesis that your measurements are different from 0 can be rejected (if you consider 5% sufficient)

Why should we do statistical analyses?

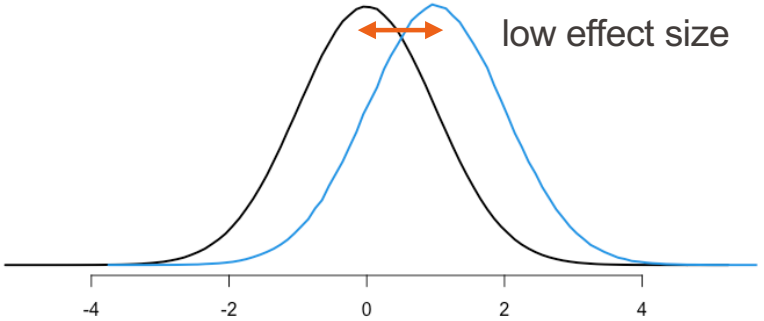
We need to measure repeatedly to have a good outcome

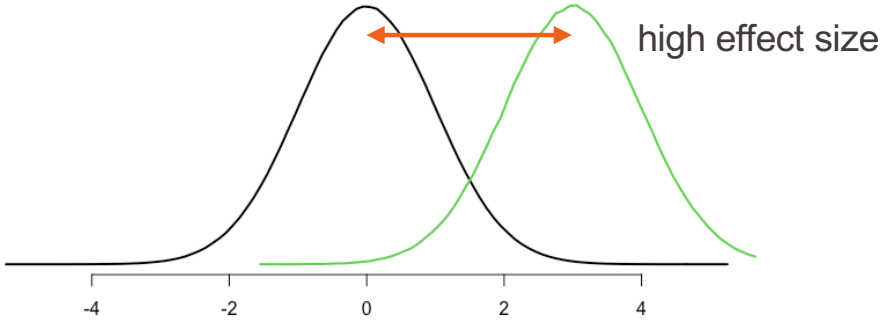
→ How often do we need to measure to have a high certainty?

→ Depends on the **variability** of the data and the **effect size** of the pattern or relationship




```
> power.t.test(...)
```





What kind of statistical analyses are there?

- Descriptive statistics
- Correlation of measurements
- Comparison of measurements between groups
- Regression models

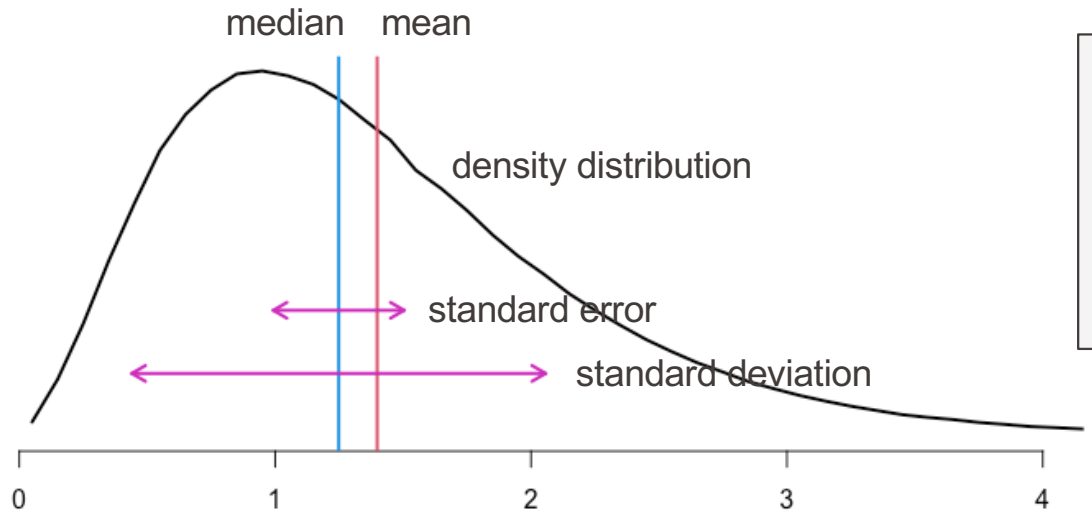
Descriptive statistics

- Mean, median, quantiles
- Standard deviation, standard error
- Distribution

→ **Standard deviation (SD): "**how far does each value within your dataset lie from the mean"

→ **Standard error (SE):** "how accurately does your sample data represent the whole population"
SE = SD / sqrt(n)

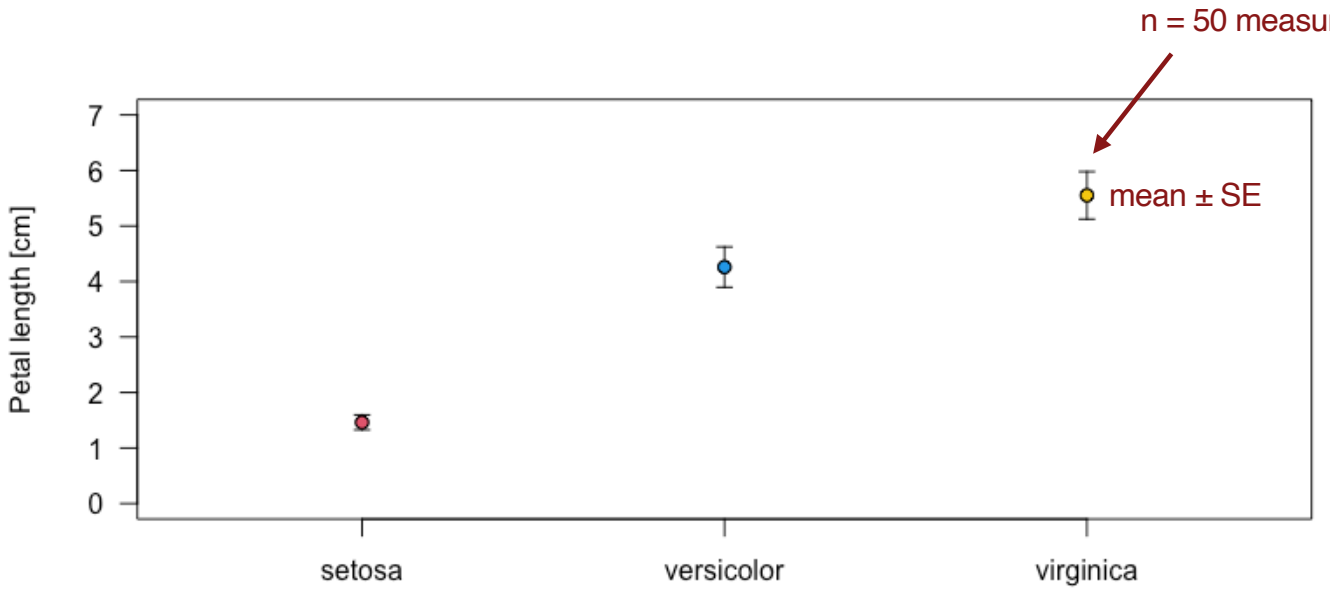
→ The standard error can be used to compare two different samples




```
> mean(my_data$my_measurements)
> median(my_data$my_measurements)
> sd(my_data$my_measurements)
> sd(my_data$my_measurements)/
  sqrt(length(my_data$my_measurements))
> hist(my_data$my_measurements)
```

Descriptive statistics

Show us your analyses in summarised form:
means and standard errors (SE)



→ SE provides a "gut feeling" of differences, but is not a statistical test

Iris versicolor

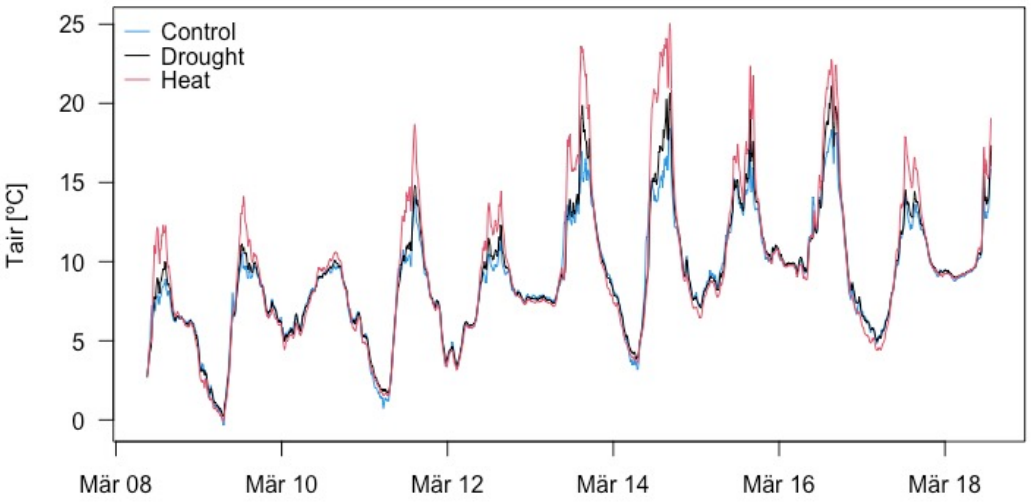
Iris setosa

Iris virginica

Descriptive statistics

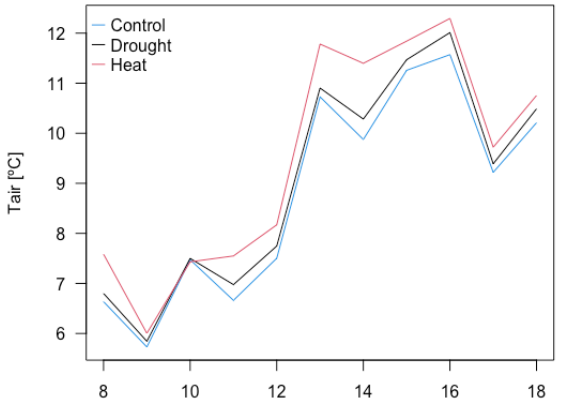
How to deal with meteodata?

→ 10 minute interval measurement too detailed for the report!

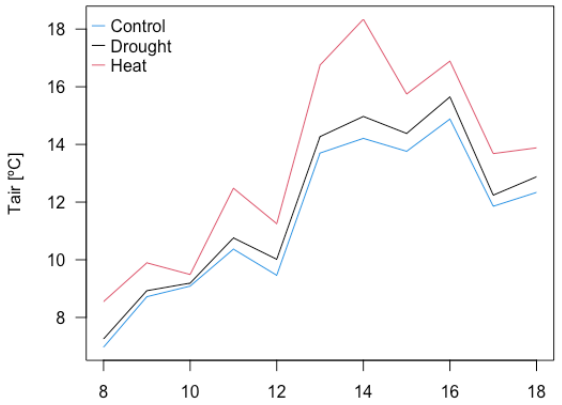


→ Don't do more than descriptive stats on meteodata!

Daily means



Daytime means (09:00–18:00)



Descriptive statistics

Basic code

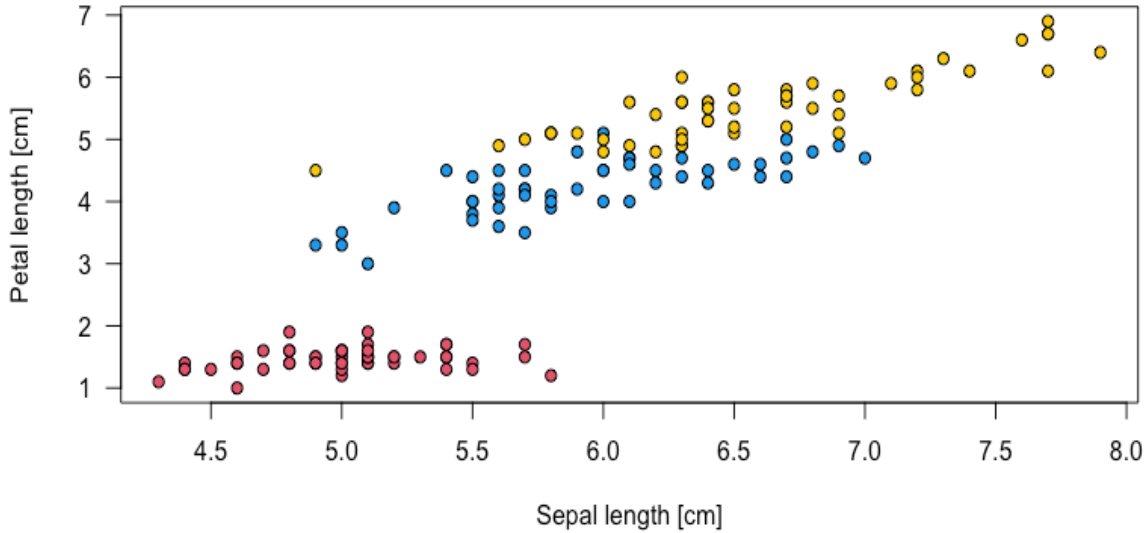
```
> data_agg_day <- aggregate(Tair ~ day + treatment,
         FUN=mean, data=my_data)

> data_agg_total <- aggregate(Tair ~ treatment,
         FUN=mean, data=my_data)
```

Advanced way

```
> install.package("dplyr")

> library(dplyr)

> data_agg_day <- my_data %>%
         group_by(treatment, day) %>%
         summarise(n=n(),
            Tair_m = mean(Tair, na.rm=T))


> data_agg_total <- my_data %>%
         group_by(treatment) %>%
         summarise(n=n(),
            Tair_m = mean(Tair, na.rm=T))
```

Correlation of measurements

Show relationships between measurements



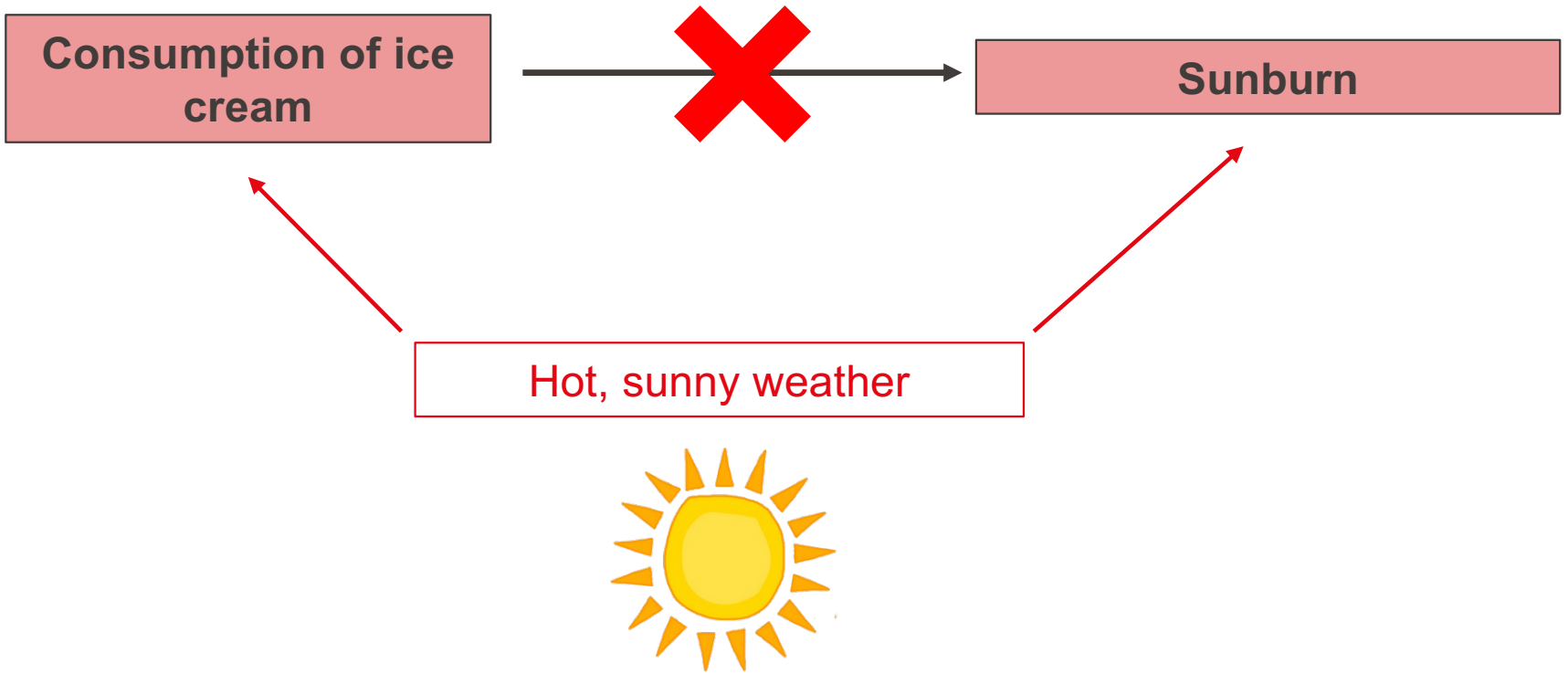

```
> cor(my_data$my_measurements1,
  my_data$my_measurements2)

[1] 0.8717538

> cor.test(my_data$my_measurements1,
  my_data$my_measurements2,
  method = "pearson")

t = 21.646, df = 148, p-value < 2.2e-16
```

→ no causality inferred! Both measurements might be related to something else (fertilisation, species, etc.)

Correlation of measurements

Confounding effect

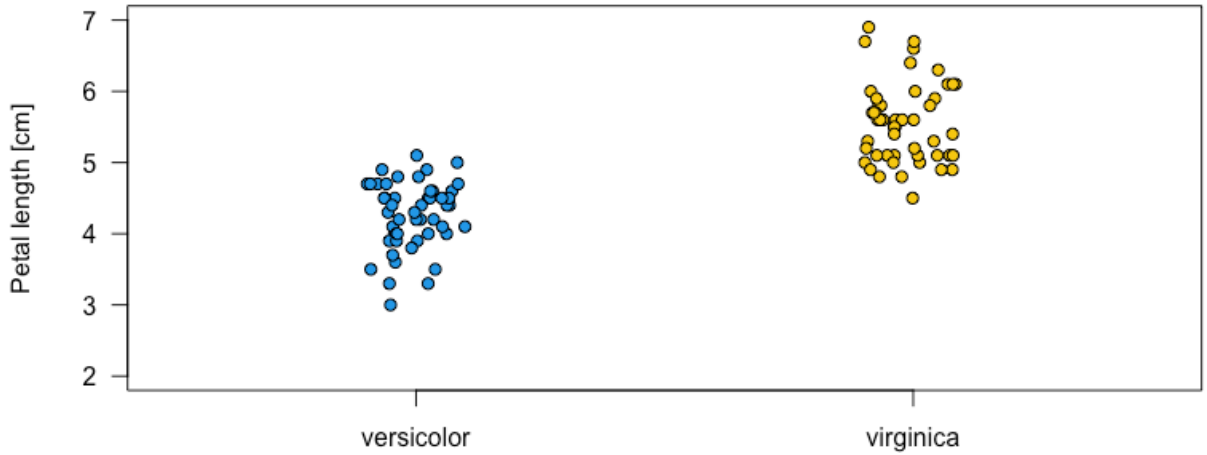
No direct causal link between predictor and response!



Group comparisons

Is the difference in petal length between the two species a true difference, or did we just measure it by chance?

→ **what is the probability that we measured the difference by chance?**



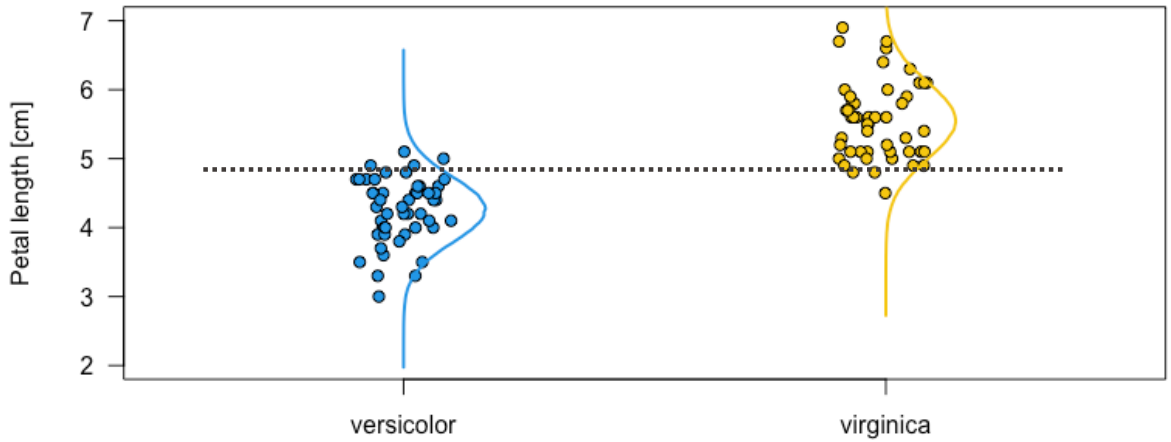
Who thinks the two species differ significantly?

Group comparisons

Assumption: the measurements are distributed with a t-distribution (similar to a normal distribution)

t-test: the probability that the difference is measured by chance is less than 0.001

→ the species differ significantly

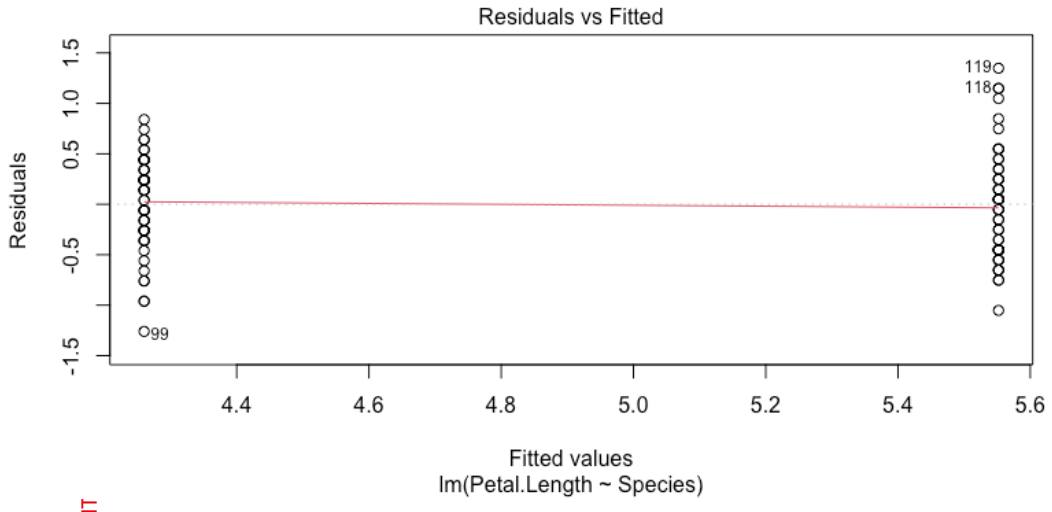


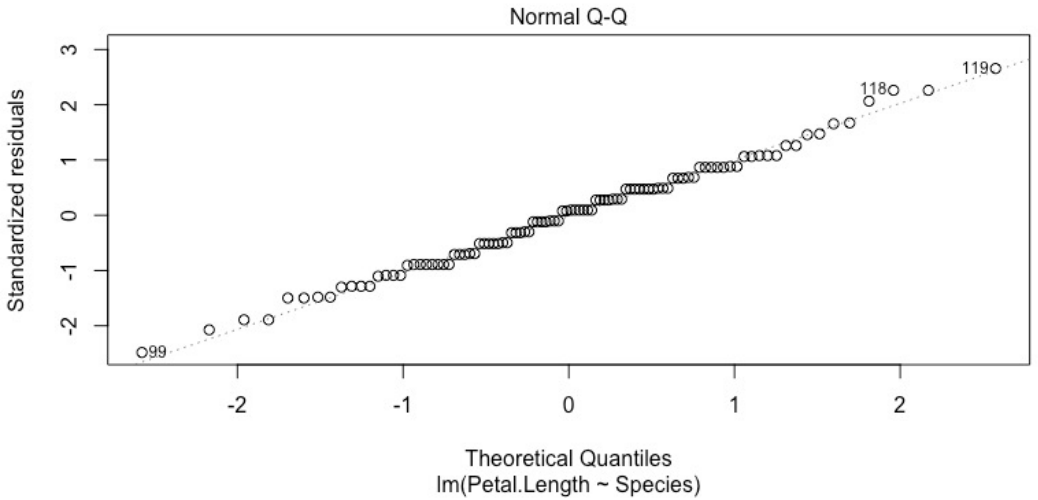

```
> t.test(iris$Petal.Length[iris$Species%in%"versicolor"], iris$Petal.Length[iris$Species%in%"virginica"])
t = -12.604, df = 95.57, p-value < 2.2e-16


> my_model <- lm(Petal.Length ~ Species, data=iris[iris$Species%in%c("versicolor", "virginica"), ])
> summary (my_model)
Speciesvirginica  1.29200    0.10251   12.60   <2e-16 ***
```

Group comparisons

Test the assumption that the data is more or less normally distributed (Homoscedasticity)

```
> plot(my_model)
> shapiro.test(iris$Petal.Length[iris$Species%in%"versicolor"])
W = 0.966, p-value = 0.1585
```





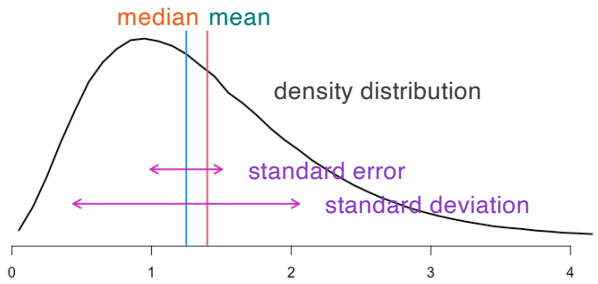
→ Always test these assumptions, but don't show us in the report

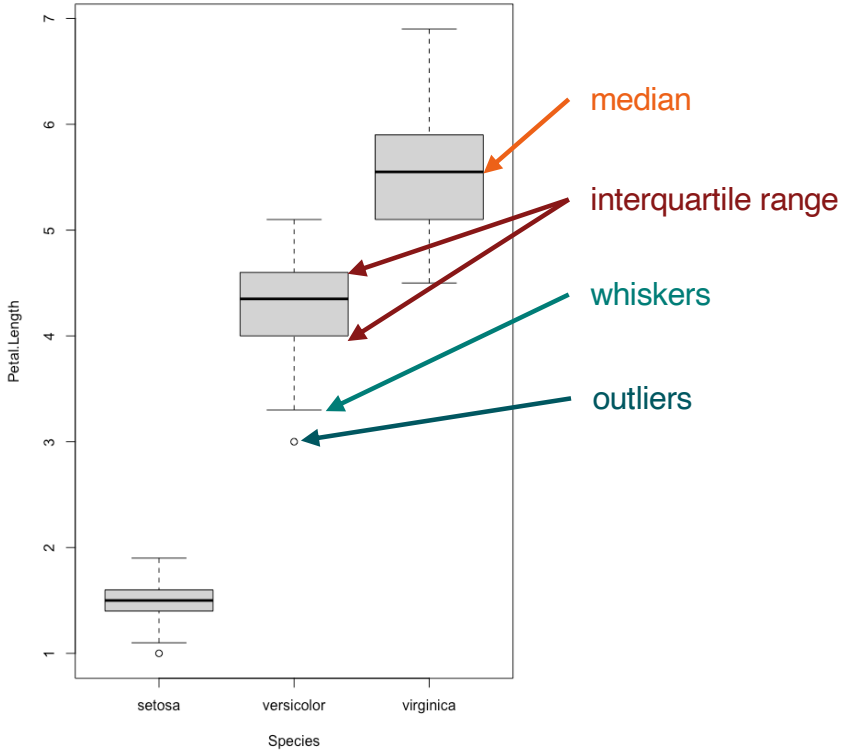
Group comparisons

Test the assumption that the data is more or less normally distributed (Homoscedasticity)

```
> boxplot(Petal.Length ~ Species, data=iris2)
```

When the median is in the middle of the interquartile range, the whiskers more or less symmetric and not much outliers, your data is pretty well normally distributed

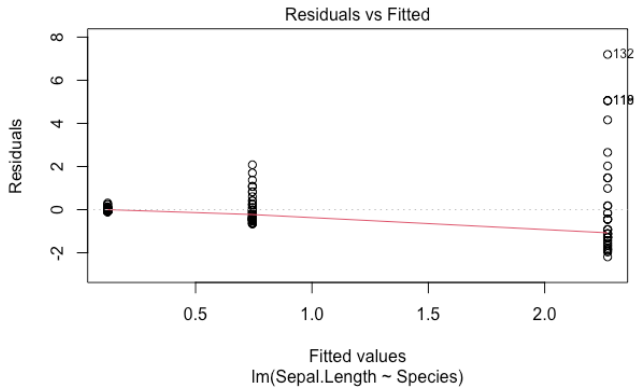


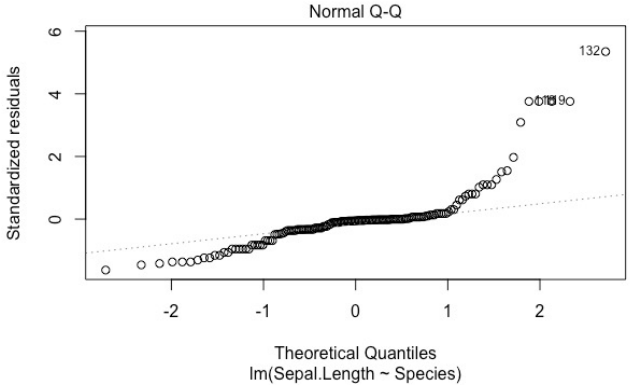


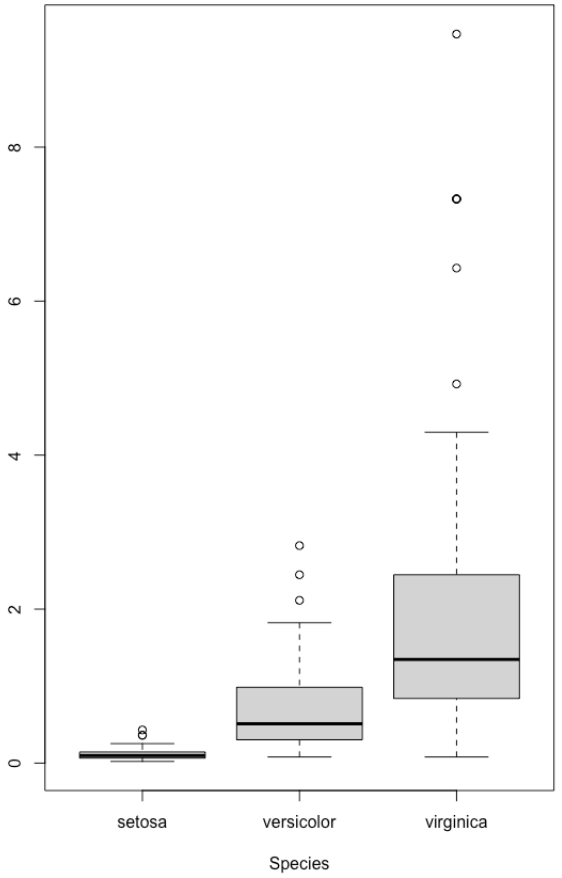
Group comparisons

What to do if the data is not normally distributed?

→ transform the data: log-transformation or sqrt-transformation








```
> my_data$log_measurement <- log(my_data$measurement)

> my_data$sqrt_measurement <- sqrt(my_data$measurement)
```

→ Don't hesitate to transform if you think it improves the analyses

Group comparisons

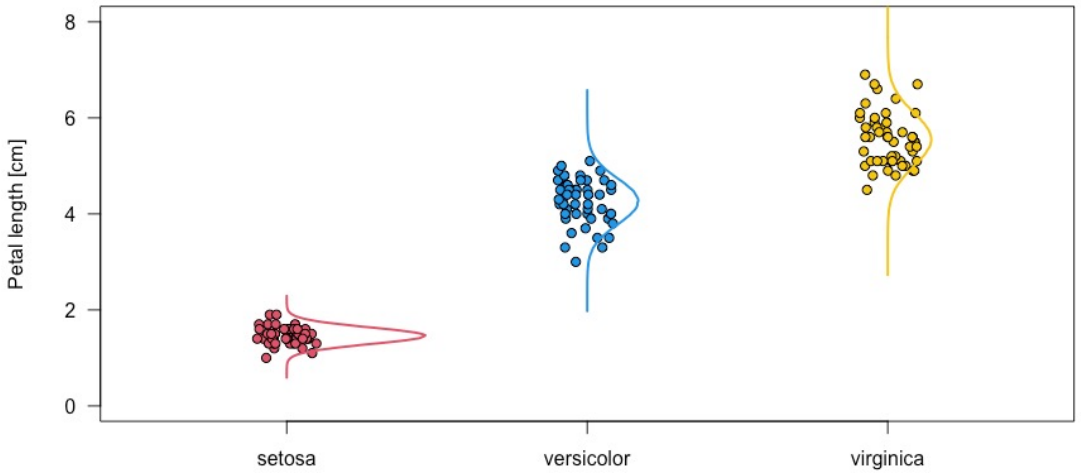
What do we do if we compare multiple groups (more than two)?

→ **Analysis of variance (ANOVA)**

```
> mymodel <- aov(Petal.Length ~ Species, data=iris)
> summary(model)

             Df   Sum Sq    Mean Sq   F value    Pr(>F)
Species       2    437.1     218.55      1180   <2e-16 ***
Residuals   147     27.2       0.19
```

The probability that the measured difference between the species is observed by chance is very small.

So, the species probably differ. But all of them? We don't know with this anova.

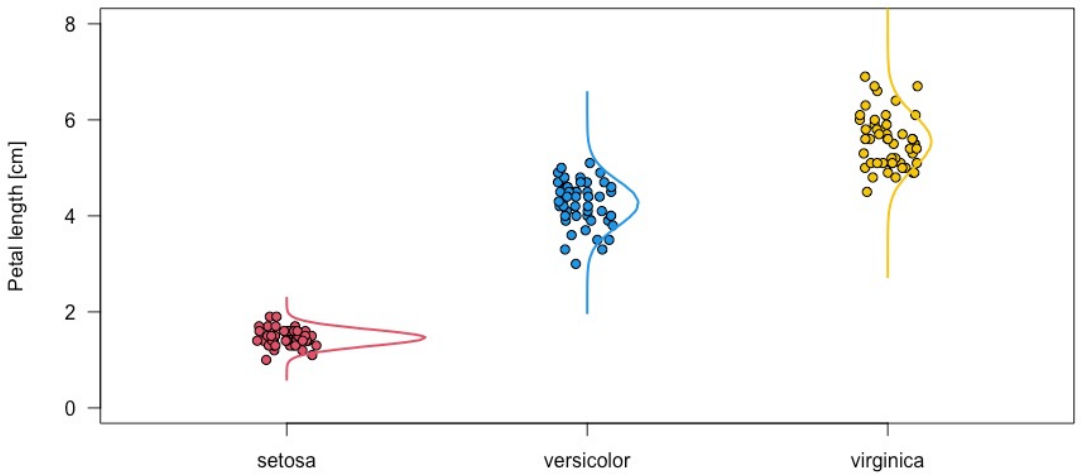


Group comparisons

Pairwise comparison: Tukey test

```
> my_model <- aov(Petal.Length ~ Species, data=iris)
> TukeyHSD(my_model)

                       diff     lwr       upr     p adj
versicolor-setosa    2.798     2.59422   3.00178     0
virginica-setosa     4.090     3.88622   4.29378     0
virginica-versicolor 1.292     1.08822   1.49578     0
```

All species probably differ from each other.

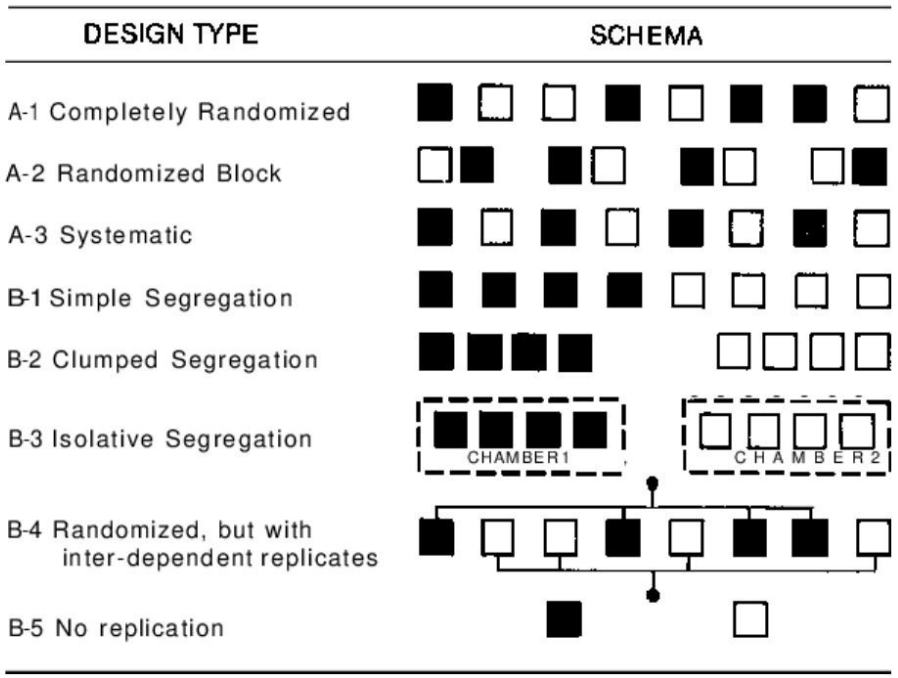


Group comparisons

Big question: which groups do I compare?

→ The analysis always has to reflect your experimental design!

Date	Treatment	Pot	Plant	Shoot length	Root length
28.3.2024	Control	1	Y	12.5	7.2
28.3.2024	Control	1	B	13.1	5.7
28.3.2024	Control	1	R	8.5	3.5
28.3.2024	Control	2	Y	7.9	4.7
28.3.2024	Control	2	B	12.5	7.2
28.3.2024	Drought	4	Y	8.5	3.5
28.3.2024	Drought	4	B	7.9	4.7
28.3.2024	Drought	4	R	12.5	7.2
28.3.2024	Drought	5	Y	13.1	5.7
28.3.2024	Drought	5	B	8.5	3.5
11.4.2024	Control	1	Y	12.5	7.2
11.4.2024	Control	1	B	13.1	5.7
11.4.2024	Control	1	R	8.5	3.5
11.4.2024	Control	2	Y	7.9	4.7



Group comparisons

Big question: which groups do I compare?

→ The analysis always has to reflect your experimental design!

Date	Treatment	Pot	Plant	Shoot length	Root length
28.3.2024	Control	1	Y	12.5	7.2
28.3.2024	Control	1	B	13.1	5.7
28.3.2024	Control	1	R	8.5	3.5
28.3.2024	Control	2	Y	7.9	4.7
28.3.2024	Control	2	B	12.5	7.2
28.3.2024	Drought	4	Y	8.5	3.5
28.3.2024	Drought	4	B	7.9	4.7
28.3.2024	Drought	4	R	12.5	7.2
28.3.2024	Drought	5	Y	13.1	5.7
28.3.2024	Drought	5	B	8.5	3.5
11.4.2024	Control	1	Y	12.5	7.2
11.4.2024	Control	1	B	13.1	5.7
11.4.2024	Control	1	R	8.5	3.5
11.4.2024	Control	2	Y	7.9	4.7

Pseudoreplication. Just to have more statistical power. Keep them as they are (don't average)

Replication. Also just keep as is.

Treatment. Test this for each date separately. (ANOVA)

Repeated measurements. Test them for each treatment separately to see the development over time.

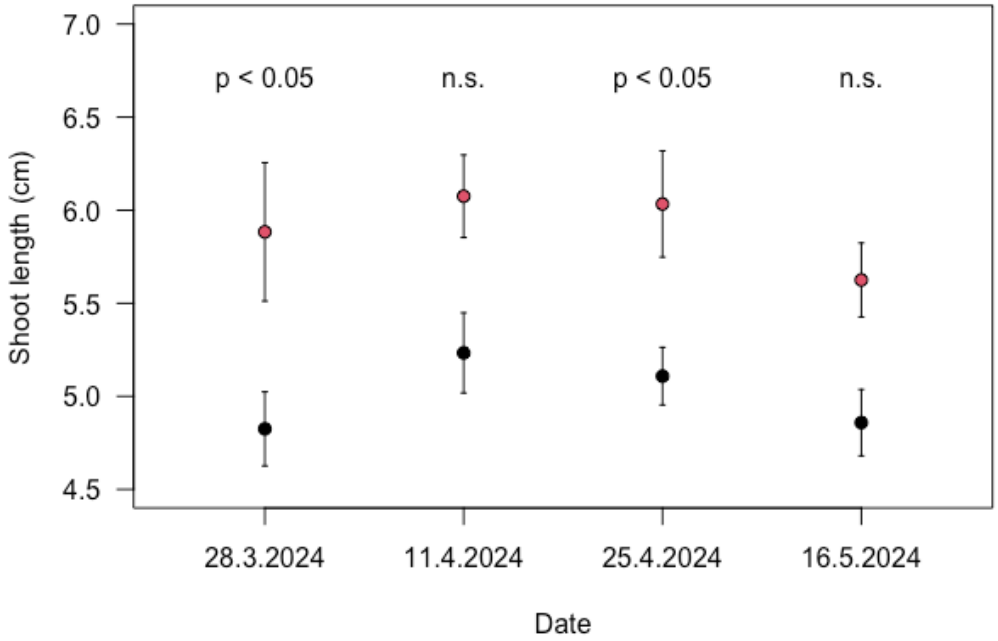
(ANOVA with Tukey)

4 x 2 x 3 x 3 = 72 measurements

Group comparisons

Big question: which groups do I compare?

→ The analysis always has to reflect your experimental design!

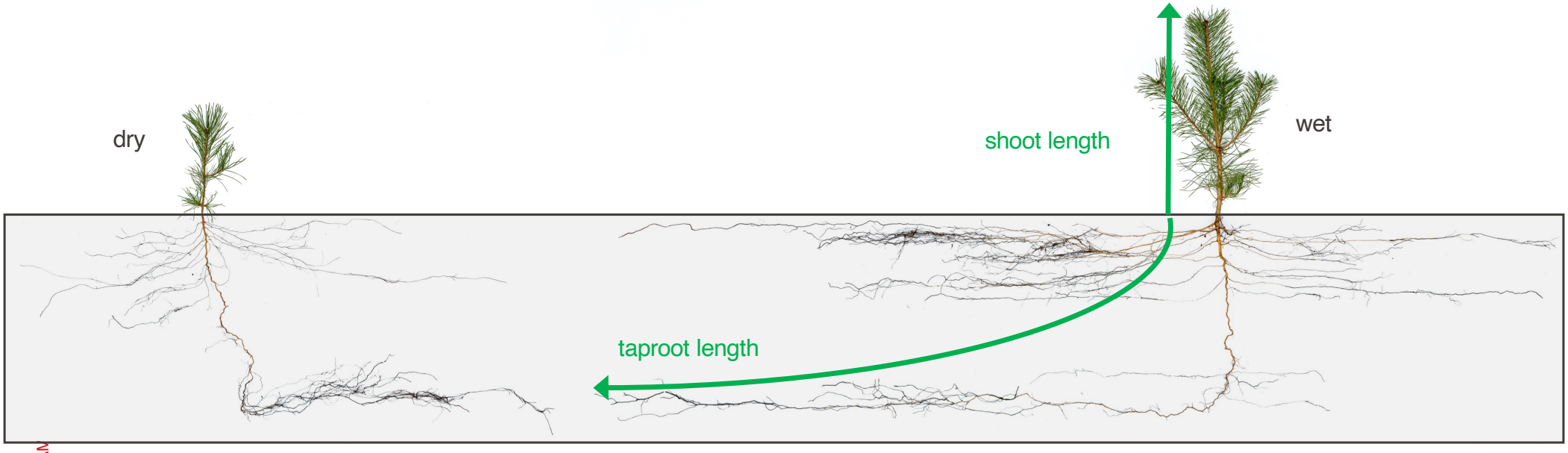



```
> my_model_treatment <- aov(Shoot_length ~
Treatment, data=my_ivy[my_ivy$date=="28.3.2024", ])


> my_model_date <- aov(Shoot_length ~ Date,
data=my_ivy[my_ivy$Treatment=="Control", ])
```

Group comparisons

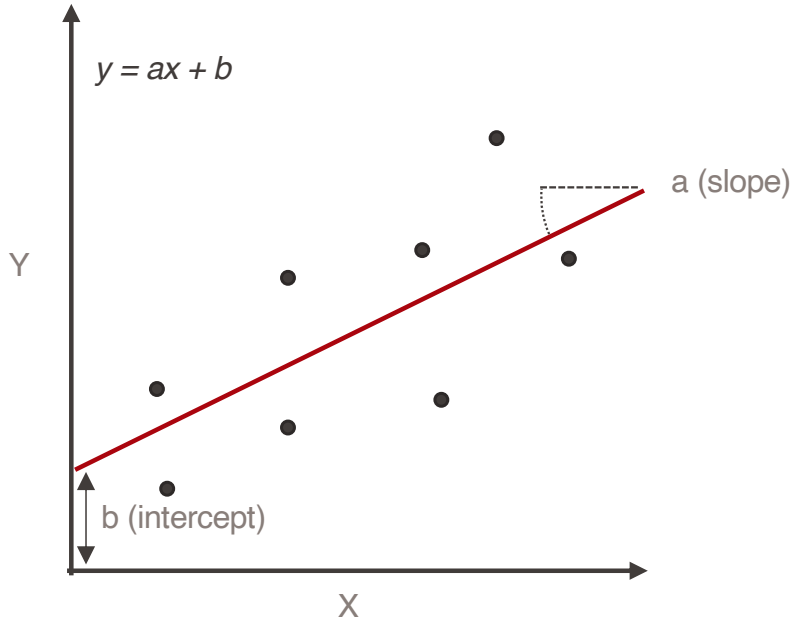
Calculate ratios

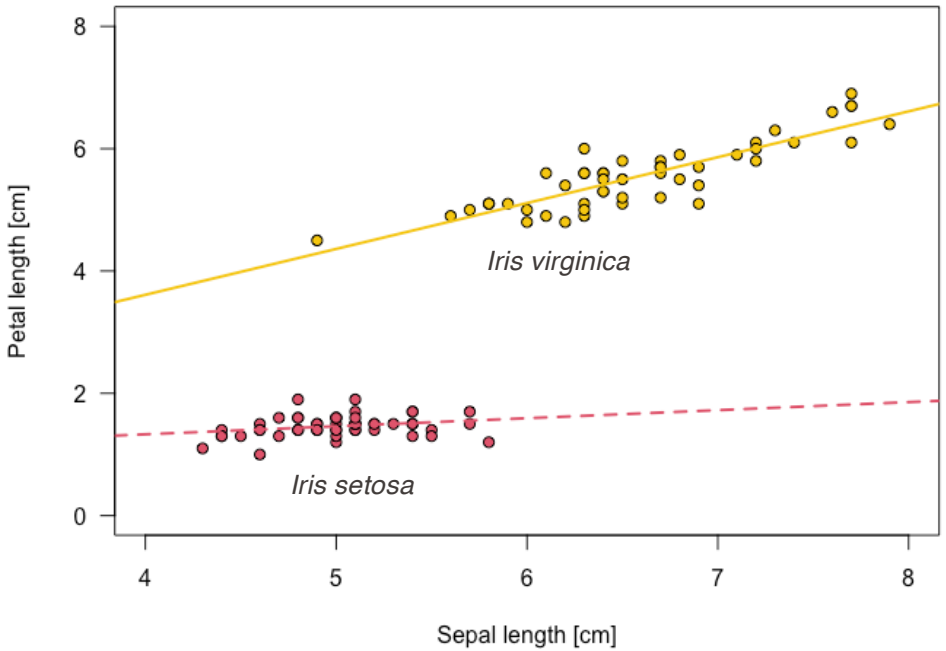
- Root : shoot biomass ratio
- Root : shoot length ratio
- Leaf area : leaf mass (specific leaf area, SLA)



Linear regression

Test the linear effect of X on Y

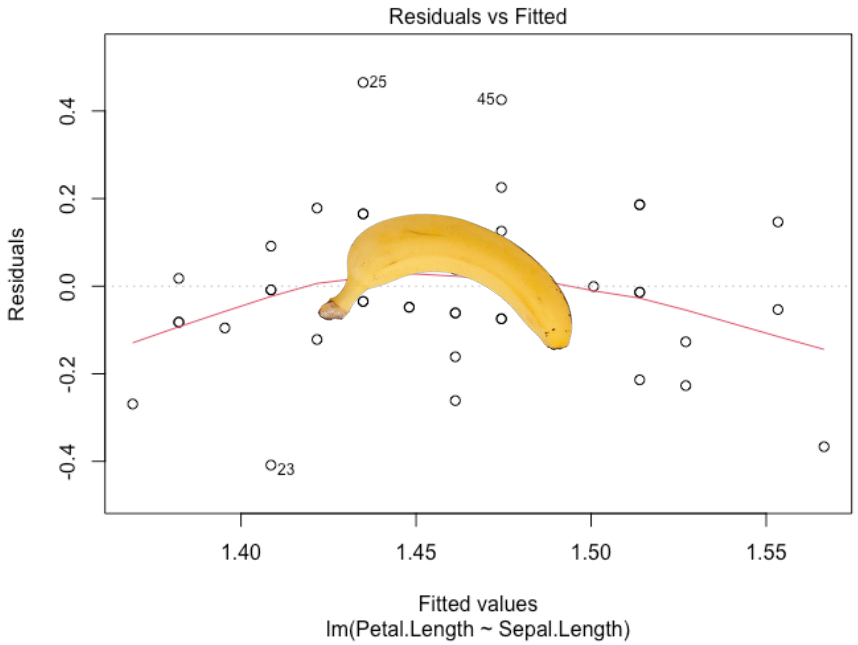




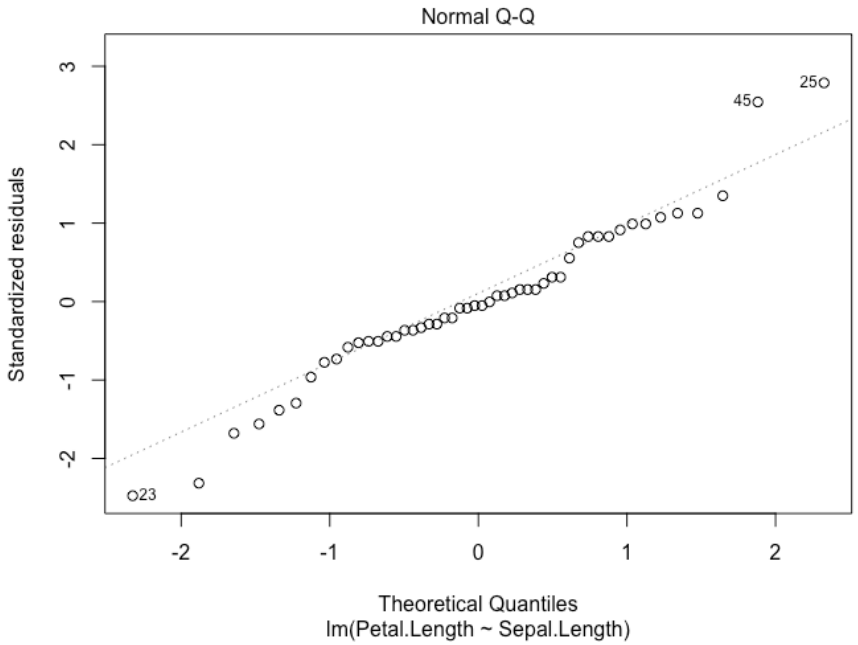

```
> model1 <- lm(Petal.Length ~ Sepal.Length, data=iris[iris$Species=="setosa", ])
> summary(model1)
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   0.80305      0.34388   2.335     0.0238 *
Sepal.Length  0.13163      0.06853   1.921     0.0607 .
```

Linear regression

Again, test the assumptions of normality



→ Not amazing, you want to avoid bananas!



→ Not too bad! Maybe no need for transformation. Up to you to decide...

Non-parametric tests

What should I do if my data is not normally distributed, and transformation does not help?

→ Use a Wilcoxon rank sum test!

→ Do this for counted data (e.g. leaf number)

→ Or whenever you feel like ☺

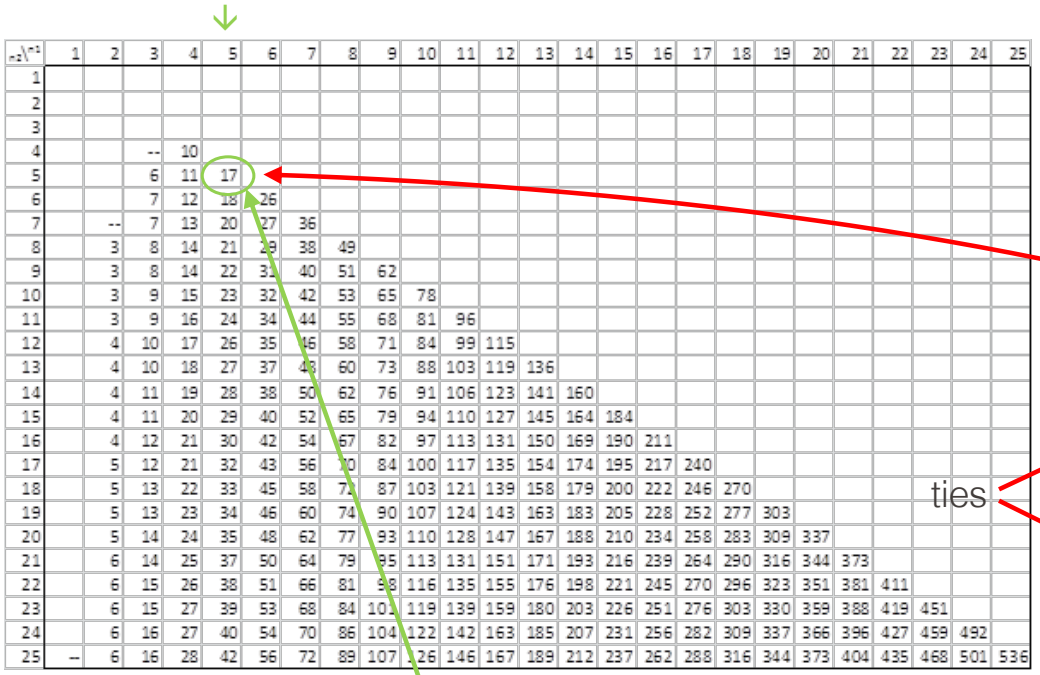
The Wilcoxon rank-sum test is a nonparametric alternative to the two- sample t-test which is based solely on the order in which the observations from the two samples fall.

It is very robust for comparing two independent samples and quite powerful.

plant number	shoot length (cm)	treatment	rank
1	20.5	drought	5
2	15	drought	3
3	12.5	drought	2
4	18.5	drought	4
5	11.5	drought	1
6	37.5	control	9
7	30	control	7.5
8	43.5	control	10
9	30	control	7.5
10	24	control	6

Non-parametric tests

test for alpha = 0.05, two-tailed
("are the treatments different?")



plant number	shoot length (cm)	treatment	rank
1	20.5	drought	5
2	15	drought	3
3	12.5	drought	2
4	18.5	drought	4
5	11.5	drought	1
6	37.5	control	9
7	30	control	7.5
8	43.5	control	10
9	30	control	7.5
10	24	control	6

ties

rank sum of drought: 5 + 3 + 2 + 4 + 1 = 15
rank sum of control: 9 + 7.5 + 10 + 7.5 + 6 = 40

significant difference ($p < 0.05$): plants have different lengths in control and drought

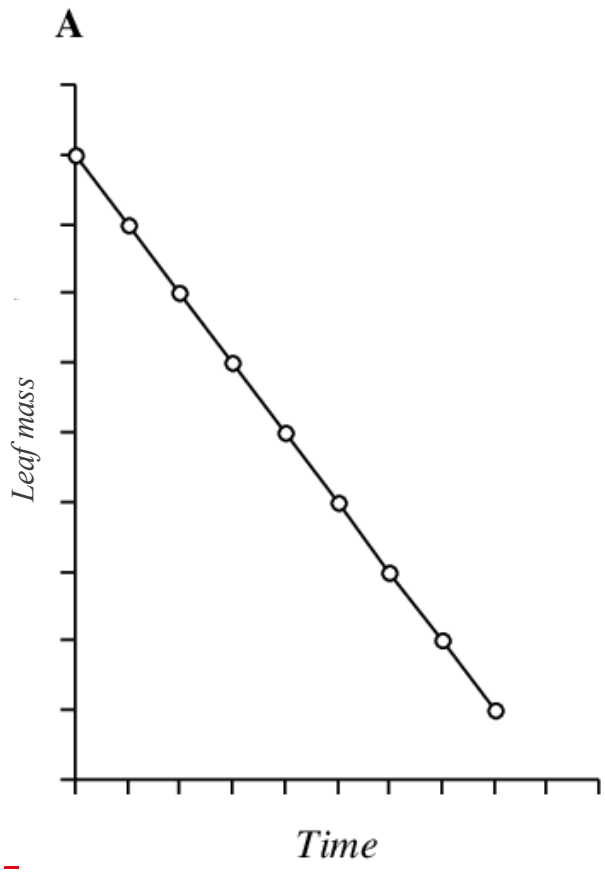
Non-parametric tests

test for alpha = 0.05, two-tailed
("are the treatments different?")

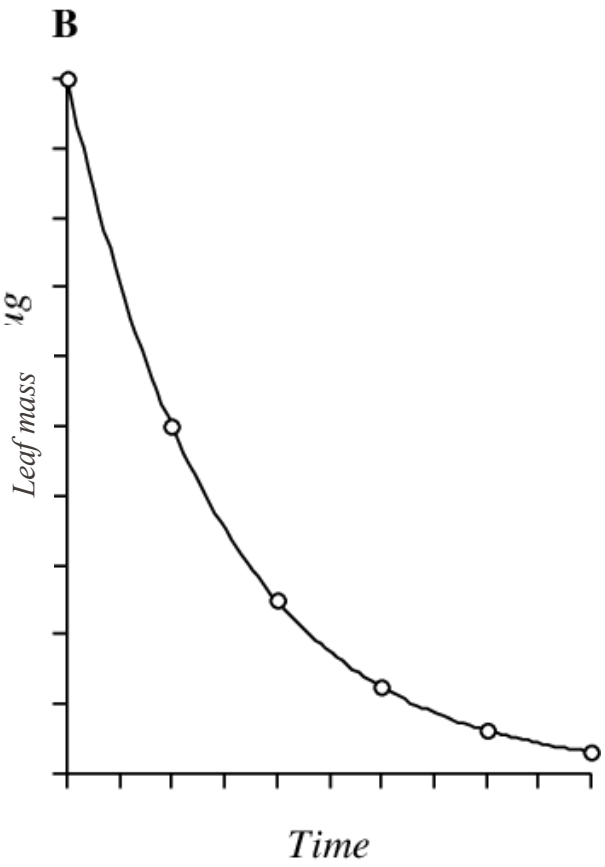
```
> wilcox.test(iris$Petal.Length[iris$Species=="setosa"],
iris$Petal.Length[iris$Species=="virginica"], alternative="two.sided")


          Wilcoxon rank sum test with continuity correction


data:  iris$Petal.Length[iris$Species == "setosa"] and
iris$Petal.Length[iris$Species == "virginica"]
W = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

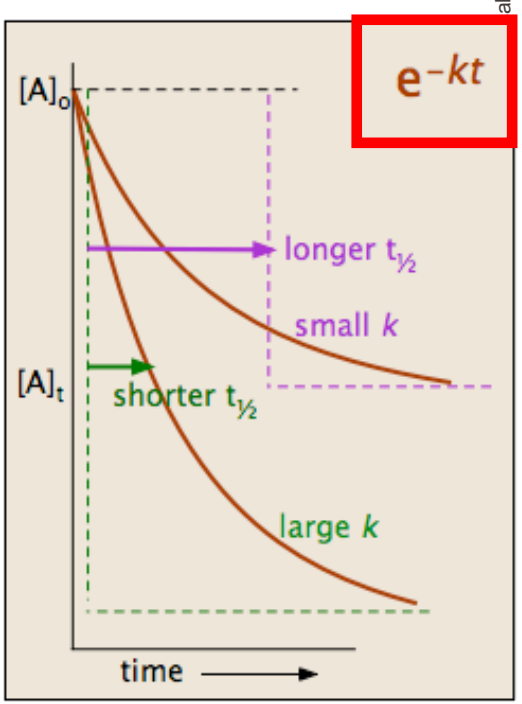
Kinetics

Zero-order kinetics
A constant <u>amount</u> of mass is removed per unit time



First-order kinetics
A constant <u>proportion</u> of mass is removed per unit time





Half-life time $t_{1/2}$
(time it takes to reach
50% of initial DM)

How to estimate breakdown rate (k)

1) **Zero-order kinetics**: regress DM against days of exposure. The slope of the regression line (negative) equals the decomposition rate k.
2) **First-order kinetics**: Regress the natural log (ln) of DM (y-axis) against days of exposure (x-axis). The slope of the regression line (negative) equals the breakdown rate constant k.

$\Rightarrow$ Compare fits of zero- and first-order kinetics: R^2

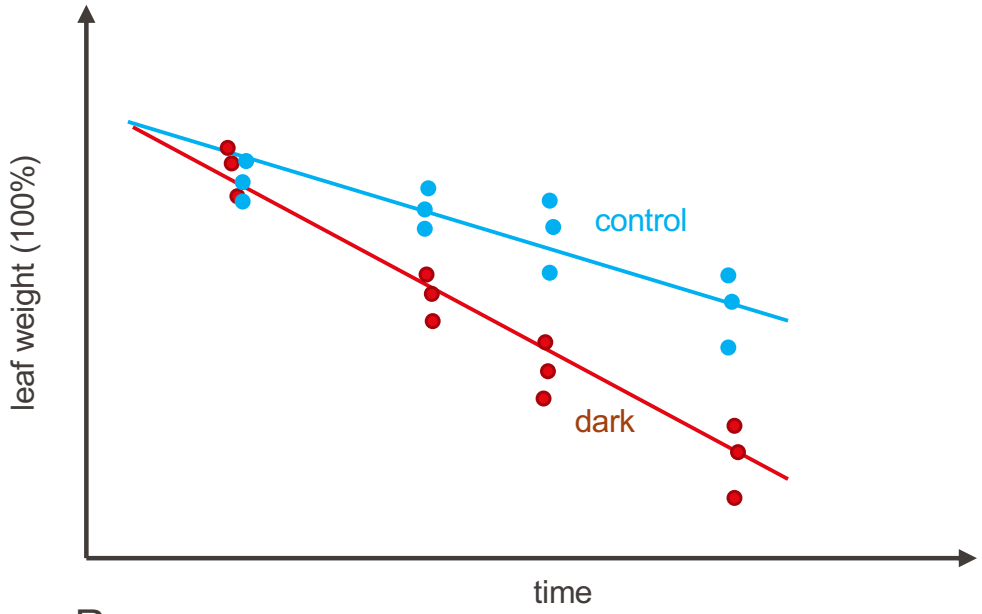
R: mod1<-lm(weigth~time)

summary(mod1) -> R^2

mod2<-lm(log(weigth)~time)

summary(mod2) –>R^2

Is there a significant effect of treatment?



R:
library(rstatix)
anova_test(weight~time*treatment)

```
## ANOVA Table (type II tests) ## ##
Effect      DFn    DFd      F                p                  p<.05 ges ##
1 time      2      39       209.314  1.40e-21 *        0.915 ##
2 treatment         1       39       572.828  6.36e-25 *        0.936 ##
3 time:treatment 2          39       0.27                8.81e-03          0.006
```