Last updated October 25, 2016

Applied Statistics Hypothesis Testing and Simple Tests

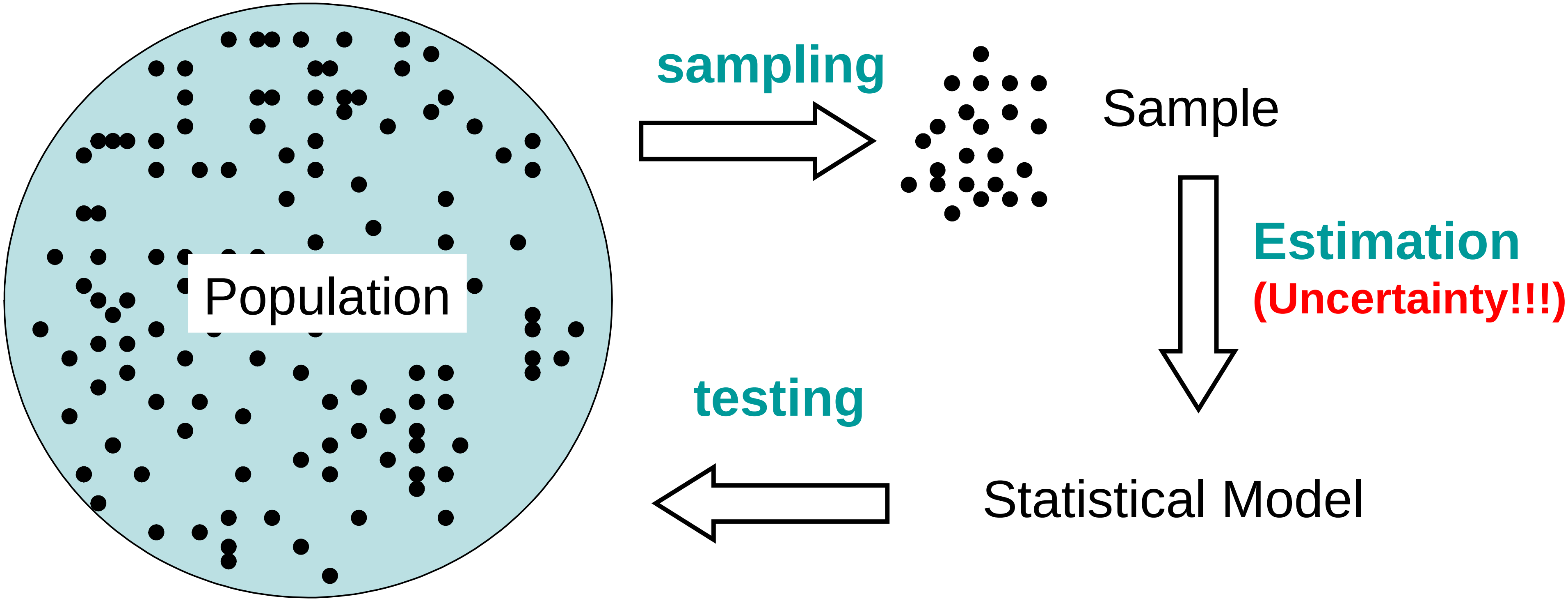
G. Bacaro

Design and Analysis of Environmental Monitoring and Experiments
Master Degree in Global Change Ecology
I Year, I term

Inference

Hypothesis testing

A **statistical hypothesis test** is a method of making statistical decisions from and about experimental data. Null-hypothesis testing just answers the question of "how well do the findings fit the possibility that chance factors alone might be responsible?".



Inferential statistics: logics

Hypothesis testing

Statistical testing in five steps:

1. Construct a null **hypothesis** (H0) (RESEARCH QUESTION)

E.g. H0: Does spruce productivity depend on soil fertility?

2. Choose a **statistical analysis**

E.g. Regression between N and P and biomass

3. Collect the data (**sampling**)

E.g. Sampling of 145 sites with different level of fertility

4. Calculate **P-value** and test statistic

Test of our regression model (F-test)

5. **Reject/accept (H0)** if p is small/large

(ANSWER THE QUESTION)

Common error

Sampling before (1) constructing the hypothesis and (2) choosing the statistical analysis

Key concepts

Hypothesis testing

Statistical testing in five steps:

1. Construct a null **hypothesis** (H0)
2. Choose a **statistical analysis (assumptions!!!)**
3. Collect the data (**sampling**)
4. Calculate **P-value** and test statistic
5. **Reject/accept (H0)** if P is small/large

Remember the order!!!

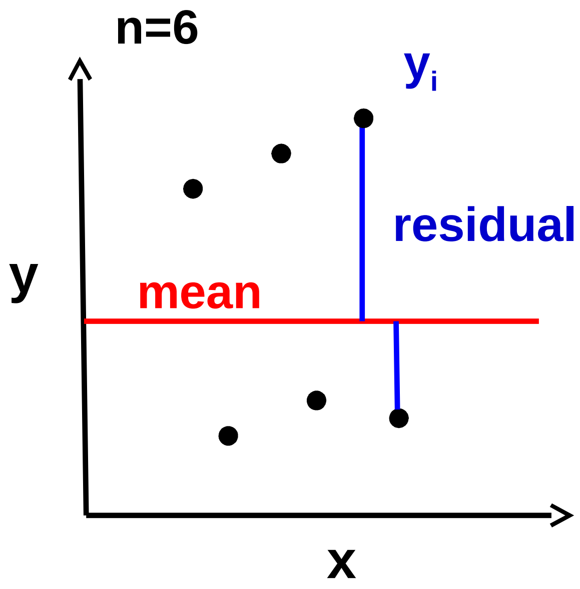
Concept of replication vs. pseudoreplication

1. Spatial dependence (e.g. spatial autocorrelation)
2. Temporal dependence (e.g. repeated measures)
3. Biological dependence (e.g. siblings)

Key quantities

$$mean = \frac{\sum y_i}{n} \qquad deviance = SS = \sum (y_i - mean)^2$$

$$\mathrm{var} = \frac{\sum (y_i - mean)^2}{(n-1)}$$



Hypothesis testing

Hypothesis testing

- 1 – Hypothesis formulation (Null hypothesis H0 vs. alternative hypothesis H1)
- 2 – Compute the probability P that H0 is false;
- 3 – If this probability is lower than a defined threshold we can reject the null hypothesis

Statistical Analyses

Hypothesis testing

Mean comparisons for 2 populations

Test the difference between the means drawn by two samples

Correlation

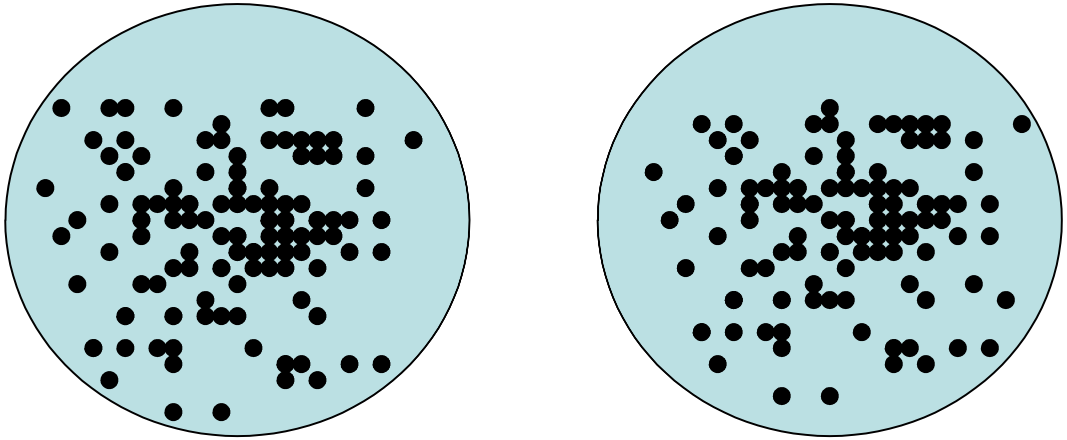
In probability theory and statistics, correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation refers to the departure of two variables from independence.

Introduction to Statistical Modelling

Basic concept

Mean Comparison for 2 Samples

Hypothesis testing



H0: means do not differ

H1: means differ

Assumptions

- Independence of cases - this is a requirement of the design.
- Normality - the distributions in each of the groups are normal
- Homogeneity of variances - the variance of data in groups should be the same (use Fisher test or Fligner's test for homogeneity of variances).
- These together form the common assumption that the errors are independently, identically, and normally distributed

Normality

Hypothesis testing

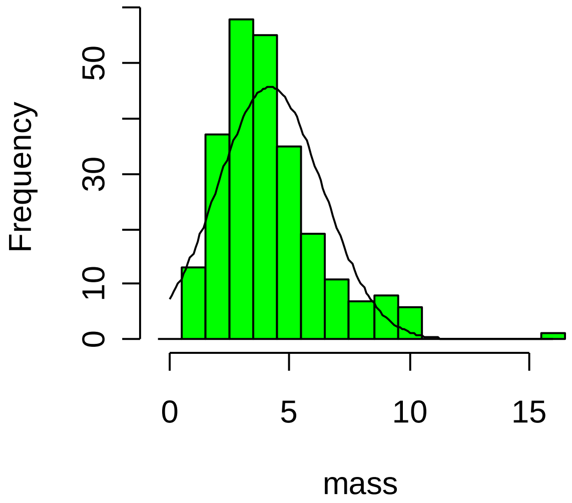
Before we can carry out a test assuming normality of the data we need to test our distribution (not always before!!!)

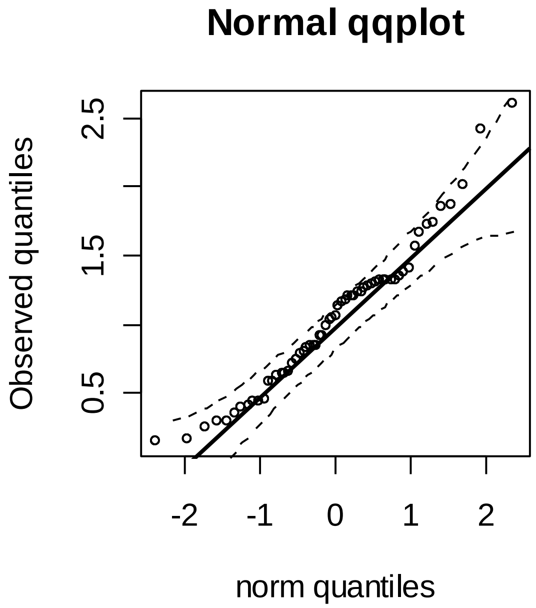
In many cases we must check this assumption after having fitted the model

(e.g. regression or multifactorial ANOVA)

RESIDUALS MUST BE NORMAL

Graphics analysis





```
hist(y)
lines(density(y))
```

```
library(car)
qq.plot(y) or qqnorm(y)
```

Test for normality

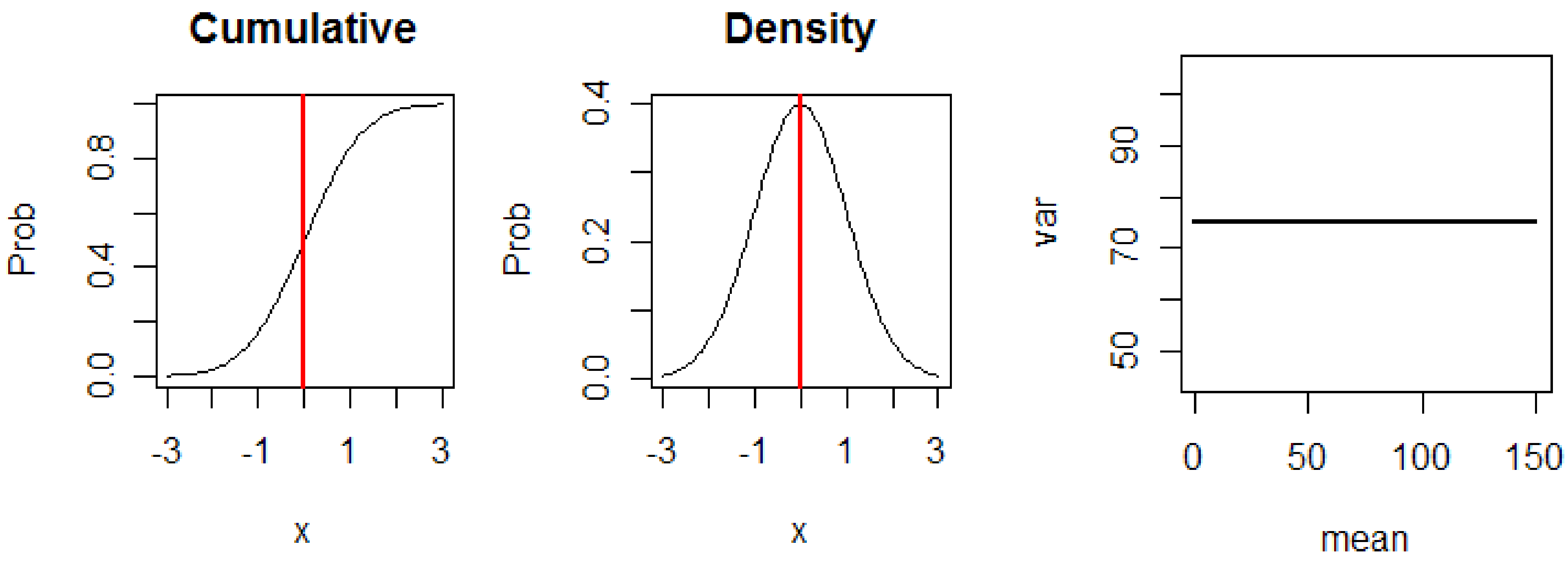
Shapiro-Wilk Normality Test `shapiro.test()`

UNIVERSITÀ
DEGLI STUDI DI TRIESTE

DIPARTIMENTO DI
SCIENZE DELLA VITA

Basic concepts: Normal distribution

The normal distribution is ubiquitous in nature and statistics due to the central limit theorem: every variable that can be modelled as a sum of many small independent variables is approximately normal.



Basic concepts: Poisson distribution

The Poisson distribution, which describes a very large number of individually **unlikely** events that happen (count data)

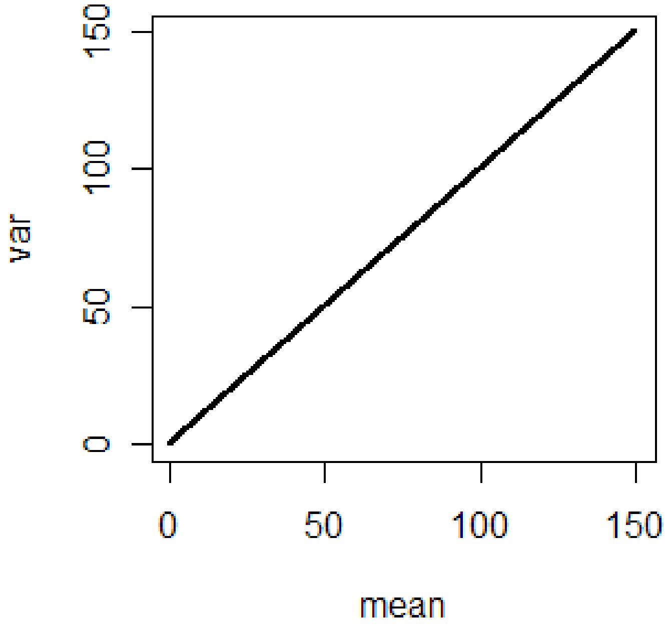
Non-negative values

Variance=mean

Right skewed

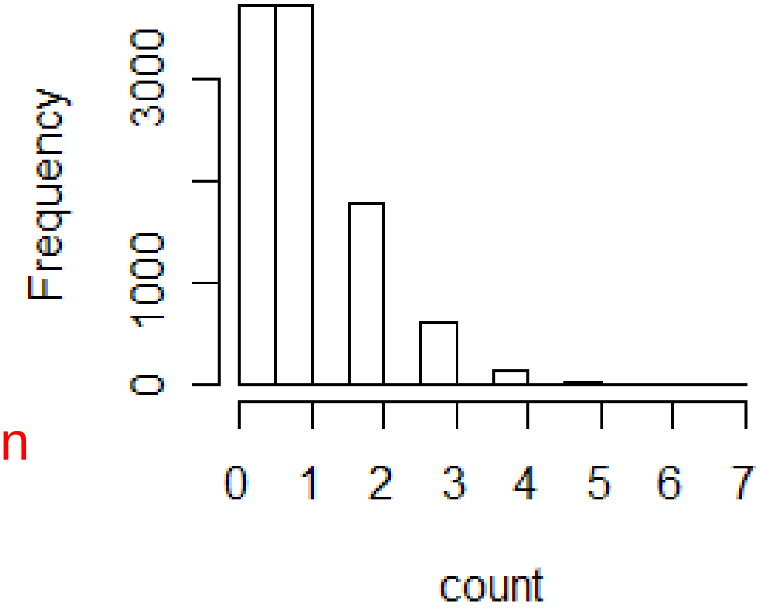
1 Parameter: λ (mean=variance)

Use: count data



var = mean

Sample from a Poisson distribution (n=1000, mean=variance=0.2)

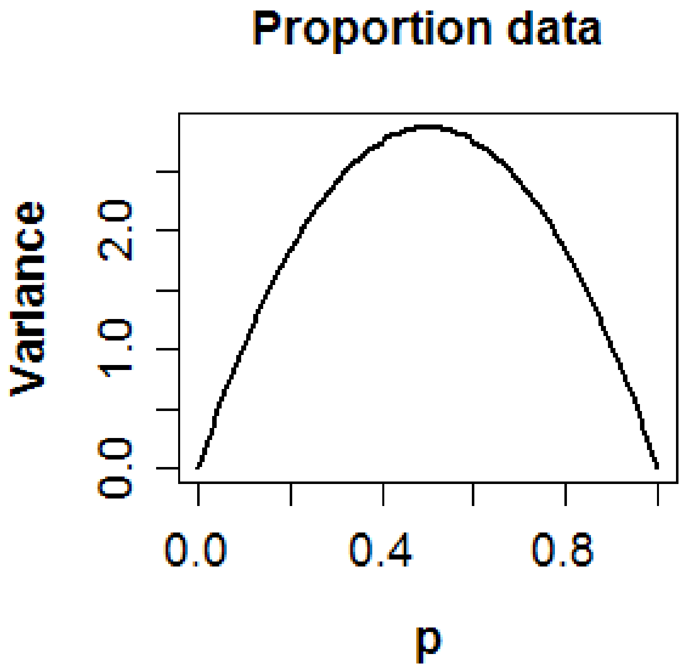


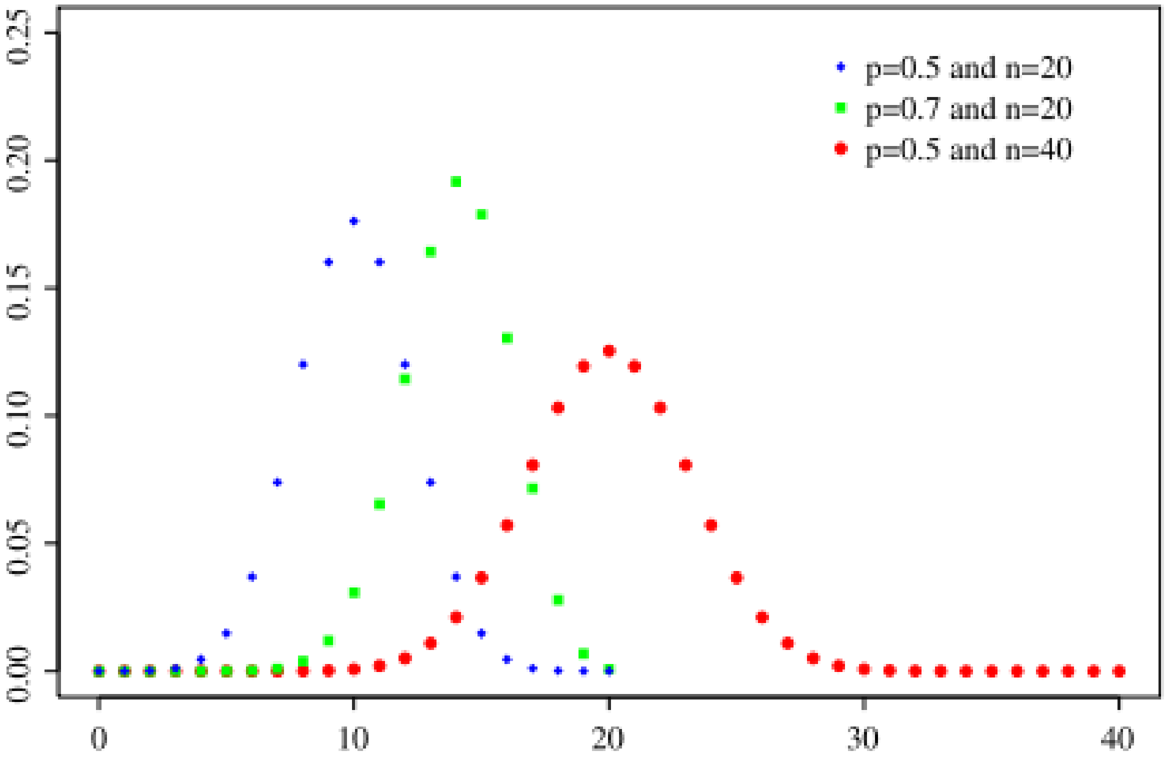
Basic concepts: Binomial distribution

The binomial distribution describes the **number of successes in a finite series of independent Yes/No experiments.**

2 Parameters: sample size, probability

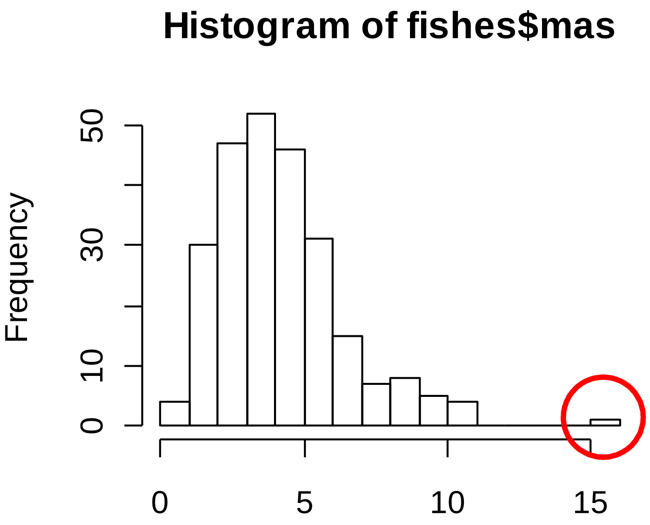
Use: proportion data and power analysis

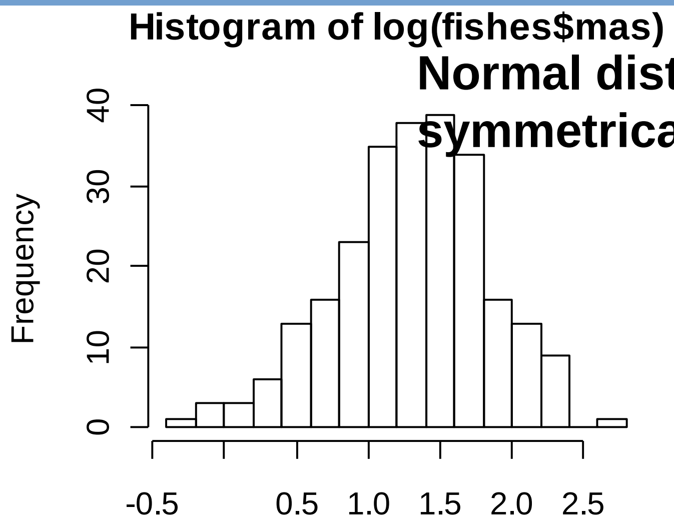




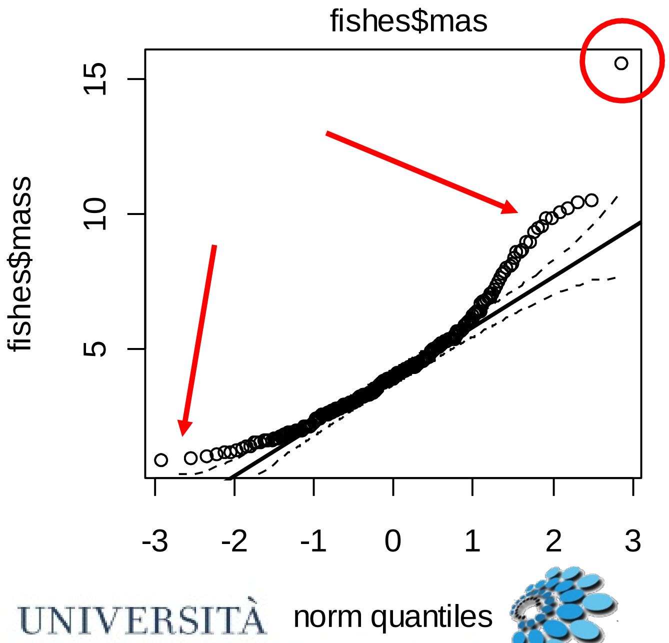
Normality: Histogram and QQ plot

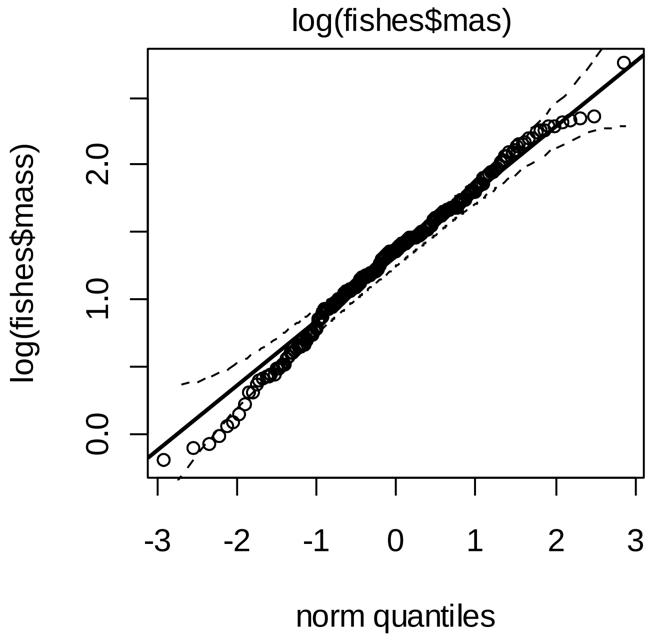
Hypothesis testing





Normal distribution must be symmetrical around the mean

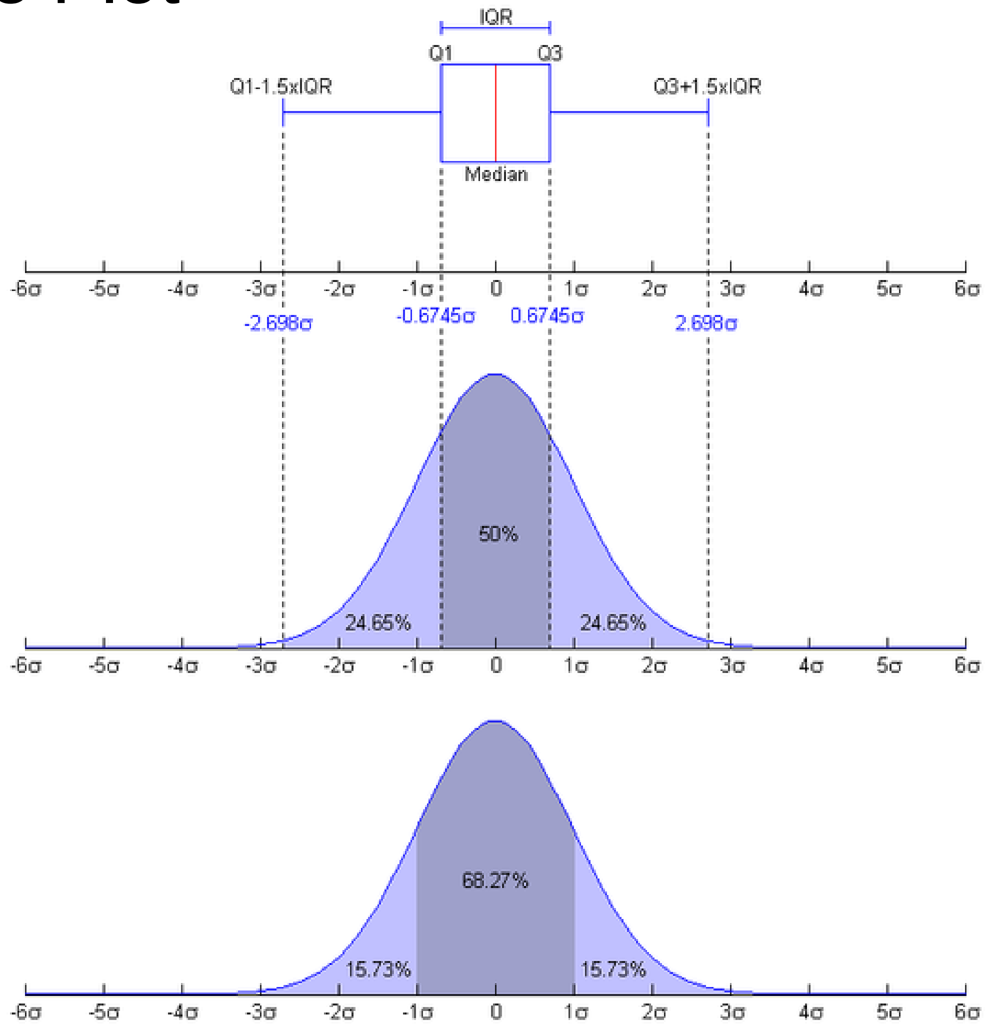




Normaliy: Quantile-Quantile Plot

Normality: Quantile-Quantile Plot

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. The quantiles are the data values marking the boundaries between consecutive subsets



Not Normal…what to do?

In case of non-normality: 2 possible approaches

1. Change the distribution (use GLMs)

Advanced statistics

E.g. Poisson (count data)

E.g. Binomial (proportion)

2. Data transformation

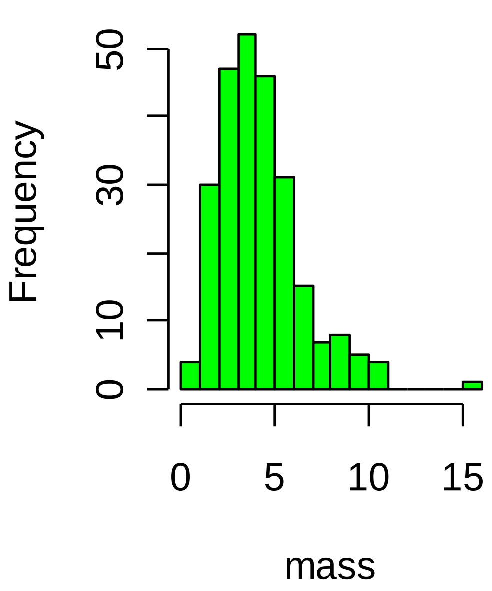
Log

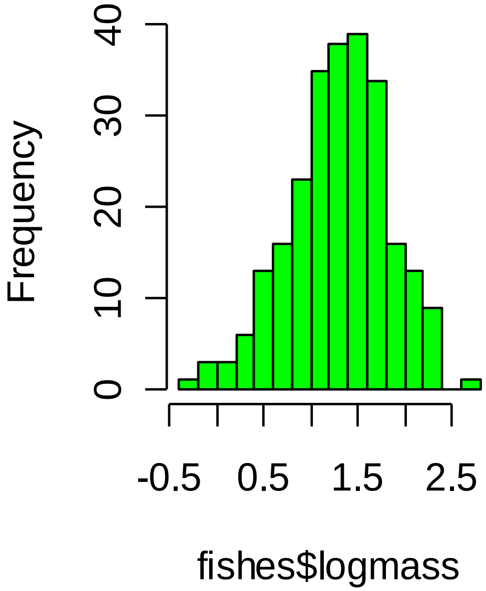
Square-root

Arcsin (percentage)

Probit (proportion)

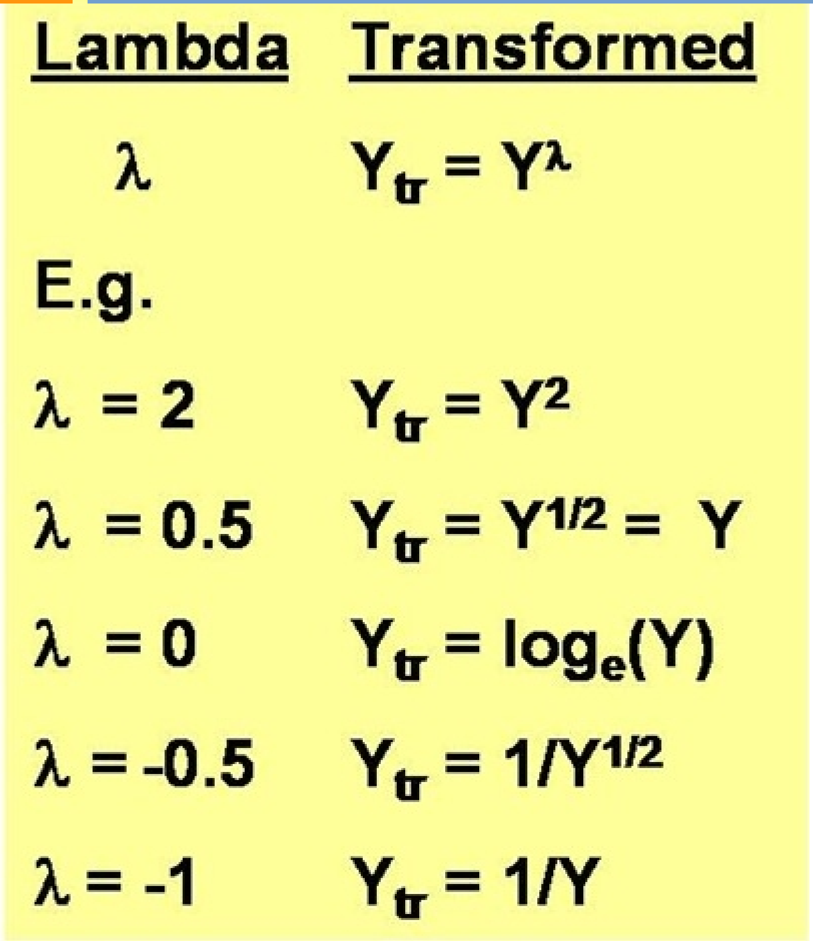
Box-Cox transformation



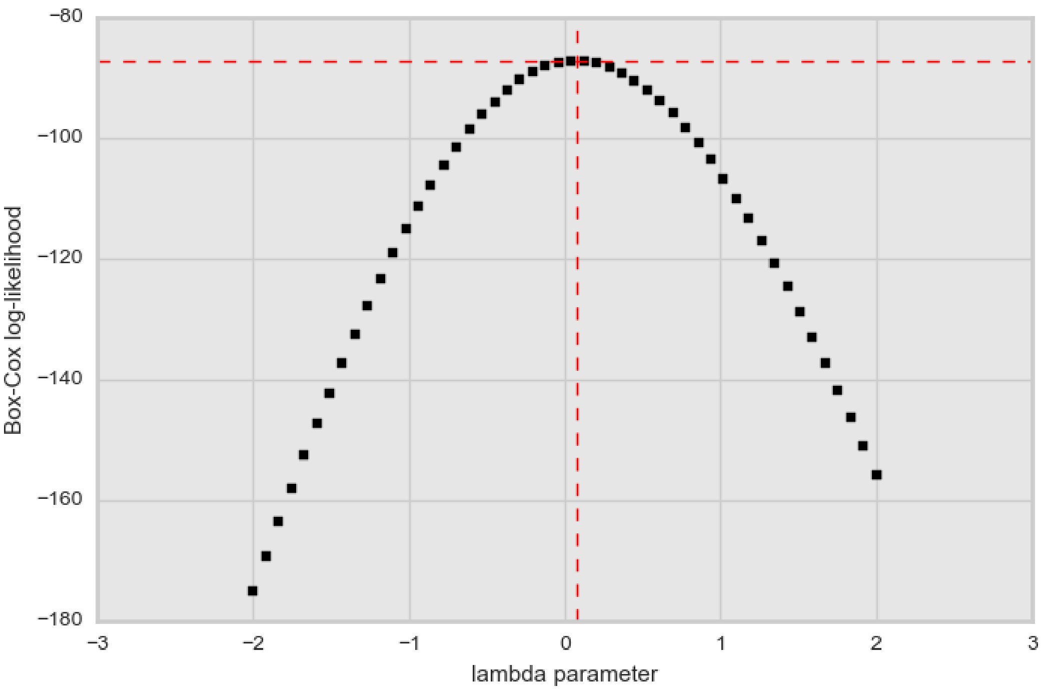


Box – Cox Normalization

Hypothesis testing



Lambda	Transformed
λ	$Y_{tr} = Y^{\lambda}$
E.g.	
$\lambda = 2$	$Y_{tr} = Y^2$
$\lambda = 0.5$	$Y_{tr} = Y^{1/2} = Y$
$\lambda = 0$	$Y_{tr} = \log_e(Y)$
$\lambda = -0.5$	$Y_{tr} = 1/Y^{1/2}$
$\lambda = -1$	$Y_{tr} = 1/Y$



$$L = -\frac{\nu}{2}\ln s^2_{TRAS} + (\lambda - 1)\frac{\nu}{n}\sum \ln X$$

$$X_{TRAS} = \frac{X^{\lambda} - 1}{\lambda}$$

$\lambda \neq 0$

$$X_{TRAS} = \log(X)$$

$\lambda = 0$

Homogeneity of Variance for two samples

Hypothesis testing

Before we can carry out a test to compare two sample means, we need to test whether the sample variances are significantly different. The test could not be simpler. It is called Fisher's *F*

To compare two variances, all you do is
divide the larger variance by the smaller variance.

E.g. Students from A group vs. Students from B group

```
F<-var(A)/var(B)
```

F calculated

```
qf(0.975,n_A-1,n_B-1)
```

F critical

if the calculated F is larger than
the critical value, we reject the null hypothesis

Test can be carried out with the **`var.test()`**

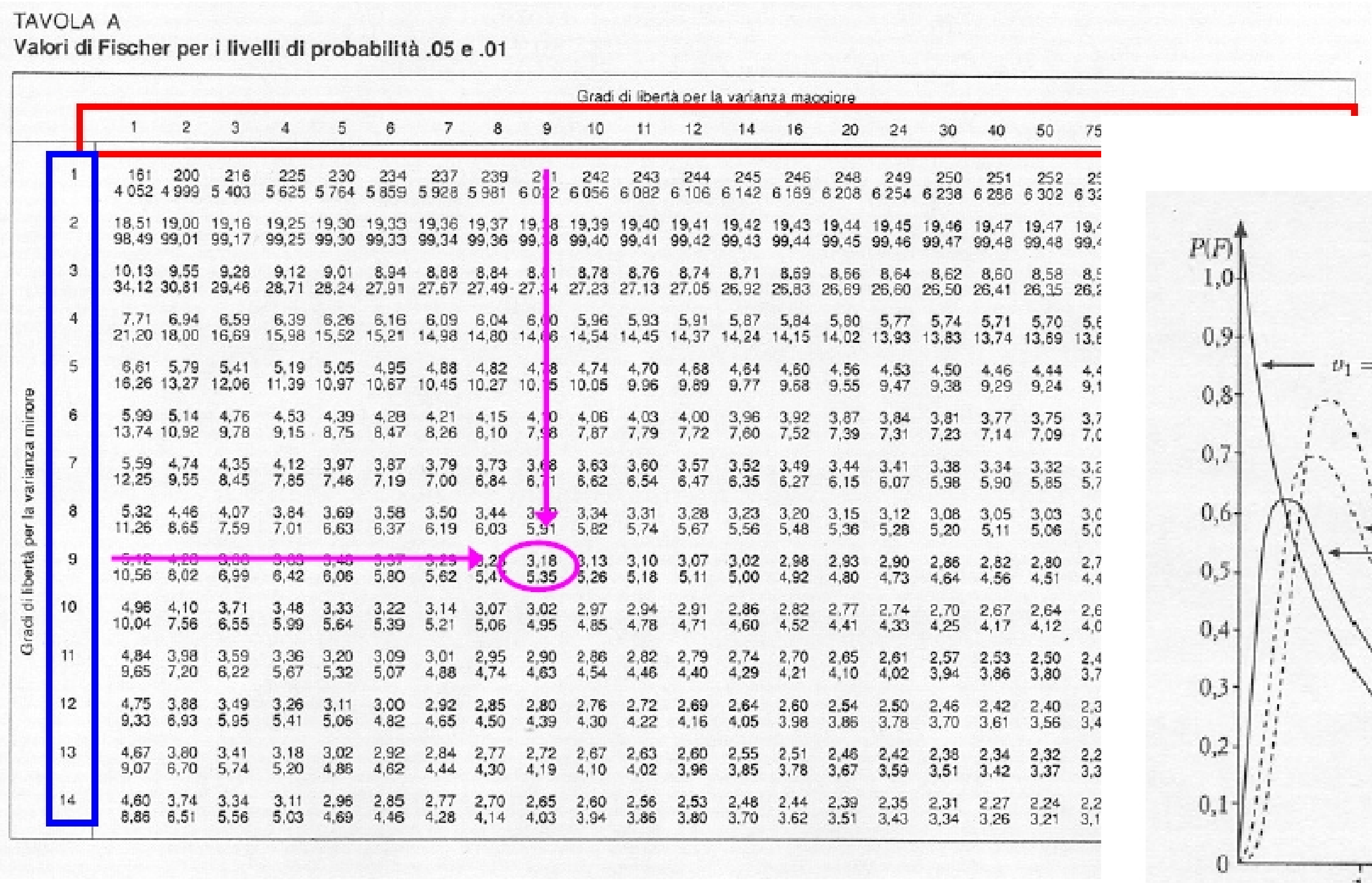
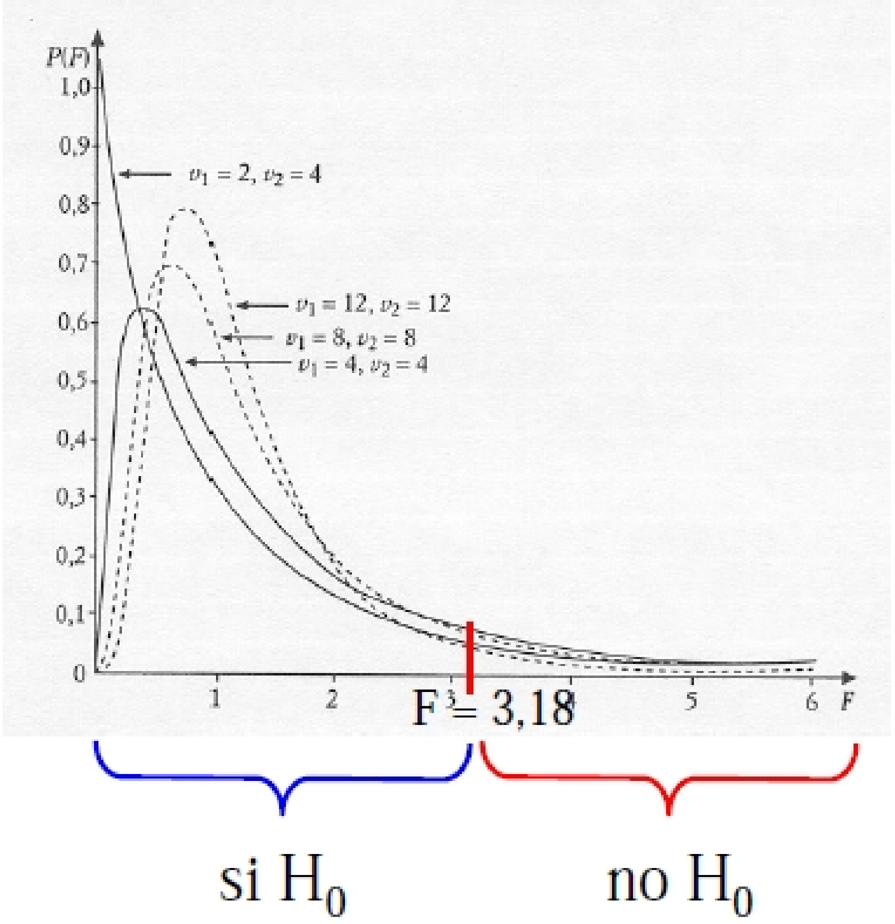
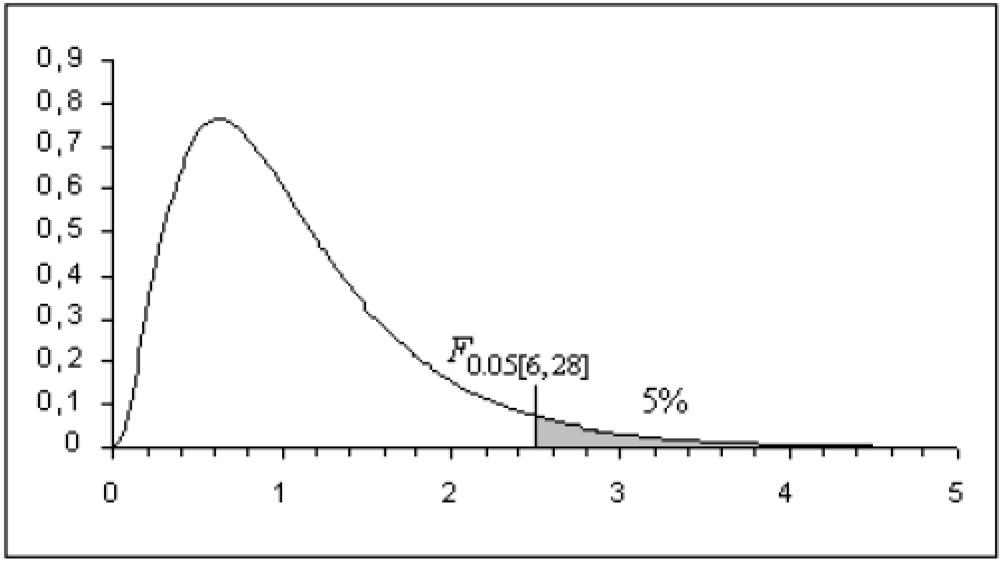


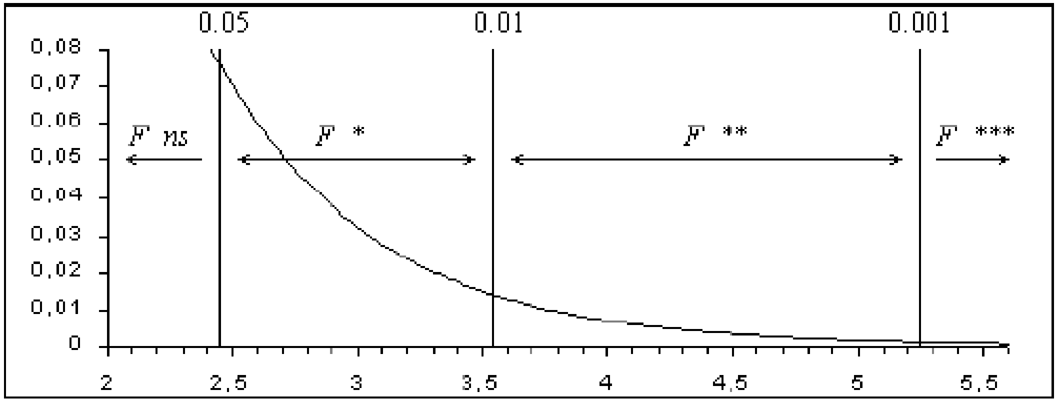
TAVOLA A

Valori di Fischer per i livelli di probabilità .05 e .01

Gradi di libertà per la varianza minore	Gradi di libertà per la varianza maggiore																			
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75
1	161 4 052	200 4 999	216 5 403	225 5 625	230 5 764	234 5 859	237 5 928	239 5 981	241 6 022	242 6 056	243 6 082	244 6 106	245 6 142	246 6 169	248 6 208	249 6 254	250 6 238	251 6 286	252 6 302	25 6 32
2	18,51 98,49	19,00 99,01	19,16 99,17	19,25 99,25	19,30 99,30	19,33 99,33	19,36 99,34	19,37 99,36	19,38 99,38	19,39 99,40	19,40 99,41	19,41 99,42	19,42 99,43	19,43 99,44	19,44 99,45	19,45 99,46	19,46 99,47	19,47 99,48	19,47 99,48	19,4 99,4
3	10,13 34,12	9,55 30,81	9,28 29,46	9,12 28,71	9,01 28,24	8,94 27,91	8,88 27,67	8,84 27,49	8,81 27,34	8,78 27,23	8,76 27,13	8,74 27,05	8,71 26,92	8,69 26,83	8,66 26,69	8,64 26,60	8,62 26,50	8,60 26,41	8,58 26,35	8,5 26,2
4	7,71 21,20	6,94 18,00	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,04 14,80	6,00 14,66	5,96 14,54	5,93 14,45	5,91 14,37	5,87 14,24	5,84 14,15	5,80 14,02	5,77 13,93	5,74 13,83	5,71 13,74	5,70 13,69	5,6 13,6
5	6,61 16,26	5,79 13,27	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,45	4,82 10,27	4,78 10,15	4,74 10.05	4,70 9,96	4,68 9,89	4,64 9,77	4,60 9,68	4,56 9,55	4,53 9,47	4,50 9,38	4,46 9,29	4,44 9,24	4,4 9,1
6	5,99 13,74	5,14 10,92	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,15 8,10	4,10 7,98	4,06 7,87	4,03 7,79	4,00 7,72	3,96 7,60	3,92 7,52	3,87 7,39	3,84 7,31	3,81 7,23	3,77 7,14	3,75 7,09	3,7 7,0
7	5,59 12,25	4,74 9,55	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 7,00	3,73 6,84	3,68 6,71	3,63 6,62	3,60 6,54	3,57 6,47	3,52 6,35	3,49 6,27	3,44 6,15	3,41 6,07	3,38 5,98	3,34 5,90	3,32 5,85	3,2 5,7
8	5,32 11,26	4,46 8,65	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,19	3,44 6,03	3,39 5,91	3,34 5,82	3,31 5,74	3,28 5,67	3,23 5,56	3,20 5,48	3,15 5,36	3,12 5,28	3,08 5,20	3,05 5,11	3,03 5,06	3,0 5,0
9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,62	3,23 5,47	3,18 5,35	3,13 5,26	3,10 5,18	3,07 5,11	3,02 5,00	2,98 4,92	2,93 4,80	2,90 4,73	2,86 4,64	2,82 4,56	2,80 4,51	2,7 4,4
10	4,96 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,21	3,07 5,06	3,02 4,95	2,97 4,85	2,94 4,78	2,91 4,71	2,86 4,60	2,82 4,52	2,77 4,41	2,74 4,33	2,70 4,25	2,67 4,17	2,64 4,12	2,6 4,0
11	4,84 9,65	3,98 7,20	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,88	2,95 4,74	2,90 4,63	2,86 4,54	2,82 4,46	2,79 4,40	2,74 4,29	2,70 4,21	2,65 4,10	2,61 4,02	2,57 3,94	2,53 3,86	2,50 3,80	2,4 3,7
12	4,75 9,33	3,88 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,92 4,65	2,85 4,50	2,80 4,39	2,76 4,30	2,72 4,22	2,69 4,16	2,64 4,05	2,60 3,98	2,54 3,86	2,50 3,78	2,46 3,70	2,42 3,61	2,40 3,56	2,3 3,4
13	4,67 9,07	3,80 6,70	3,41 5,74	3,18 5,20	3,02 4,86	2,92 4,62	2,84 4,44	2,77 4,30	2,72 4,19	2,67 4,10	2,63 4,02	2,60 3,96	2,55 3,85	2,51 3,78	2,46 3,67	2,42 3,59	2,38 3,51	2,34 3,42	2,32 3,37	2,2 3,3
14	4,60 8,86	3,74 6,51	3,34 5,56	3,11 5,03	2,96 4,69	2,85 4,46	2,77 4,28	2,70 4,14	2,65 4,03	2,60 3,94	2,56 3,86	2,53 3,80	2,48 3,70	2,44 3,62	2,39 3,51	2,35 3,43	2,31 3,34	2,27 3,26	2,24 3,21	2,2 3,1







Homogeneity of Variance > two samples

Hypothesis testing

It is important to know whether variance differs significantly from sample to sample. Constancy of variance (**homoscedasticity**) is the most important assumption underlying regression and analysis of variance. For multiple samples you can choose between the

Bartlett test and the **Fligner–Killeen test**.

```
Bartlett.test(response,factor)
```

```
Fligner.test(response,factor)
```

There are differences between the tests: Fisher and Bartlett are very sensitive to outliers, whereas Fligner–Killeen is not

Mean Comparisons

Hypothesis testing

In many cases, a researcher is interesting in gathering information about two populations in order to compare them. As in statistical inference for one population parameter, confidence intervals and tests of significance are useful statistical tools for the difference between two population parameters.

Ho: the two means are the same

H1: the two means differ

- All Assumptions met? Parametric `t.test()`

- t test with independent or paired sample

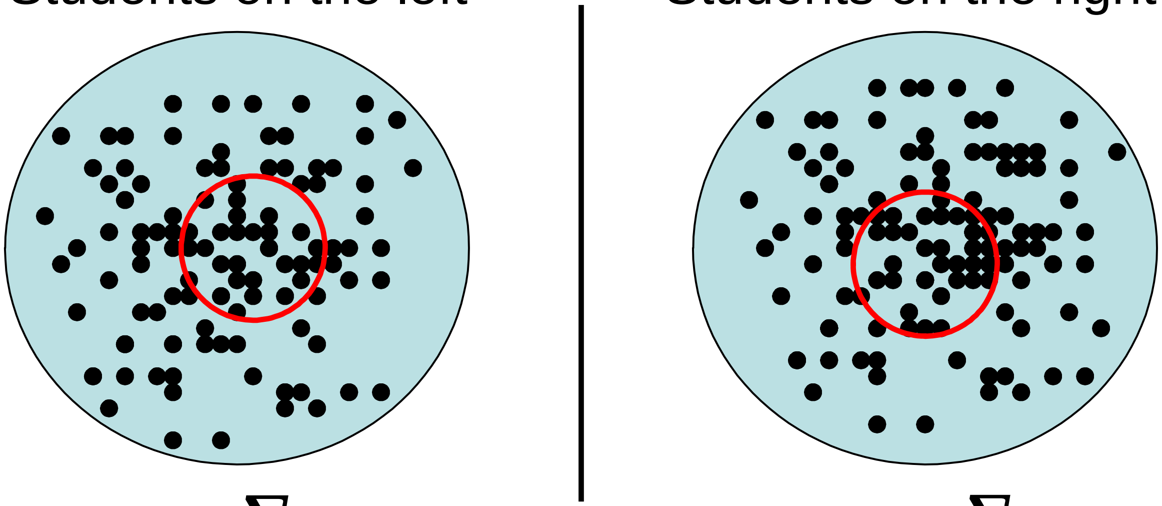
-Some assumptions not met? Non-parametric `Wilcox.test()`

- The **Wilcoxon signed-rank test** is a non-parametric alternative to the Student's t-test for the case of two samples.

Mean Comparison: Two independent sample

Students on the left

Students on the right



The two samples are statistically independent

$$mean_a = \frac{\sum y_i}{n}$$

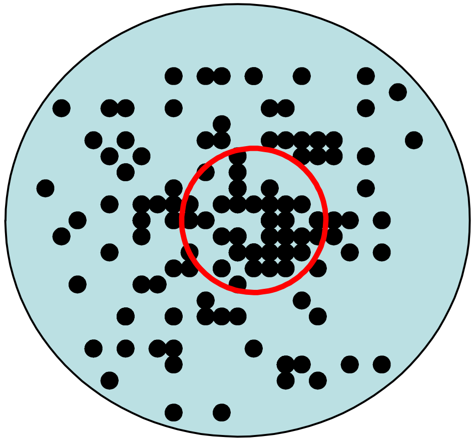
$$mean_b = \frac{\sum y_i}{n}$$

$$SE_{diff} = \sqrt{\frac{\text{var}_a}{n_a} + \frac{\text{var}_b}{n_b}}$$

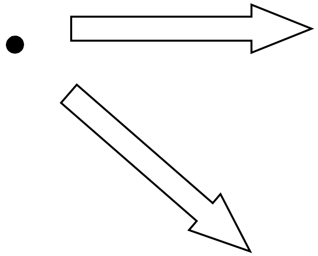
$$t = \frac{mean_a - mean_b}{SE_{diff}}$$

Test can be carried out with the `t.test()` function

T-Test for paired samples



E.g. Test your performance before or after the course. I measure twice on the same student



Time 1 a: 1, 2, 3, 2, 3, 2 ,2

Time 2 b: 1, 2, 1, 1, 5, 1, 2

$$t = \frac{(\sum a_i - b_i)/n}{SD_{diff} / \sqrt{n}}$$

If we have information about dependence, we have to use this!!!

We can deal with dependence

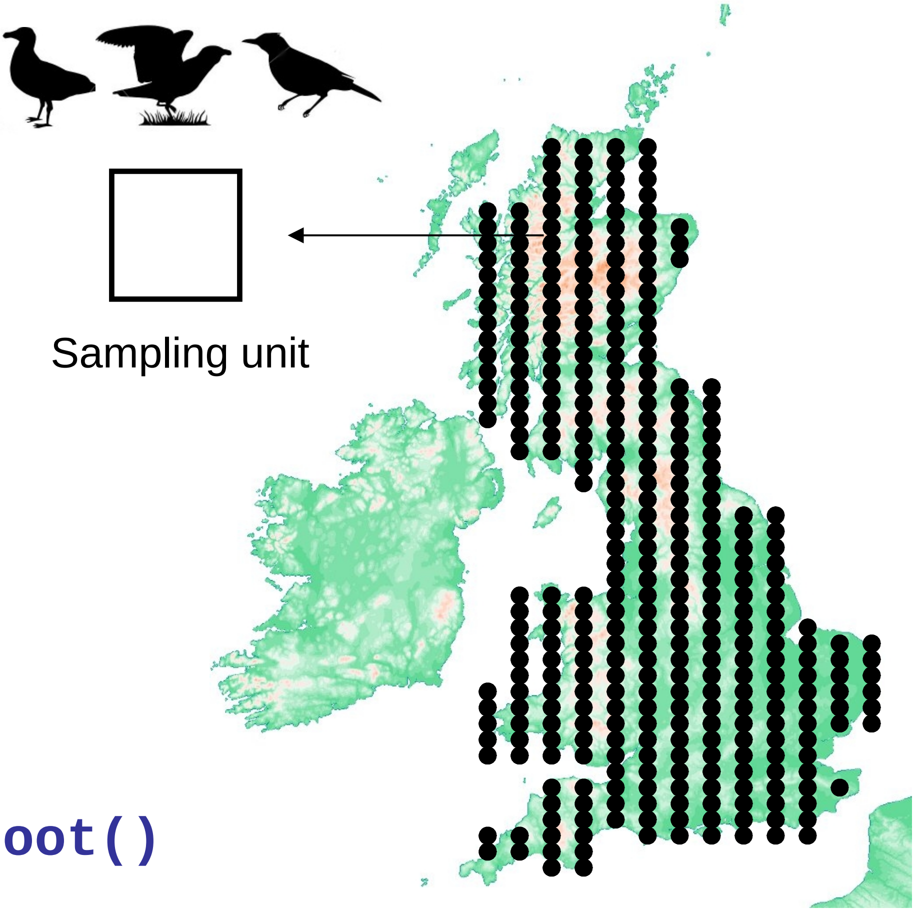
Test can be carried out with the `t.test()` function

Correlation

Hypothesis testing

Correlation, (often measured as a correlation coefficient), indicates the **strength and direction of a linear** relationship between two random variables

Bird species richness	Plant species richness
x_1	l_1
x_2	l_2
x_3	l_3
x_4	l_4
…	…
x_{458}	l_{458}



Three alternative approaches

1. Parametric - `cor()`
2. Nonparametric - `cor()`
3. Bootstrapping - `replicate(), boot()`

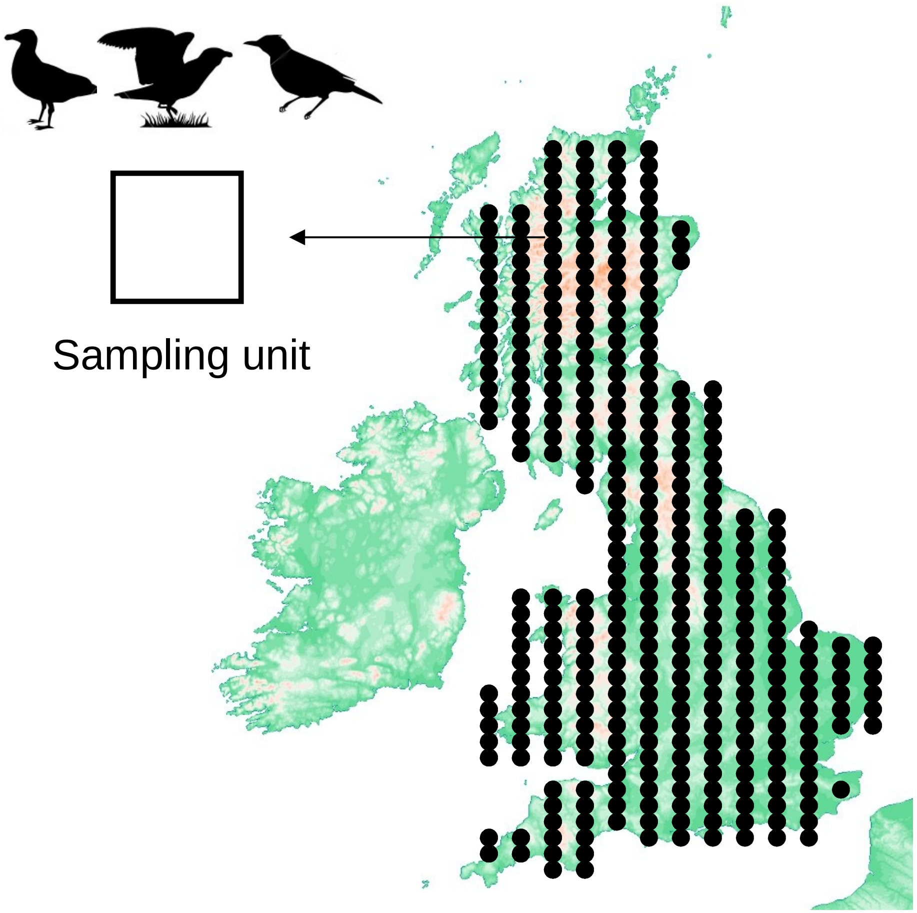
Correlation: causal relationship?

Hypothesis testing

Which is the response variable in a correlation analysis?

NONE

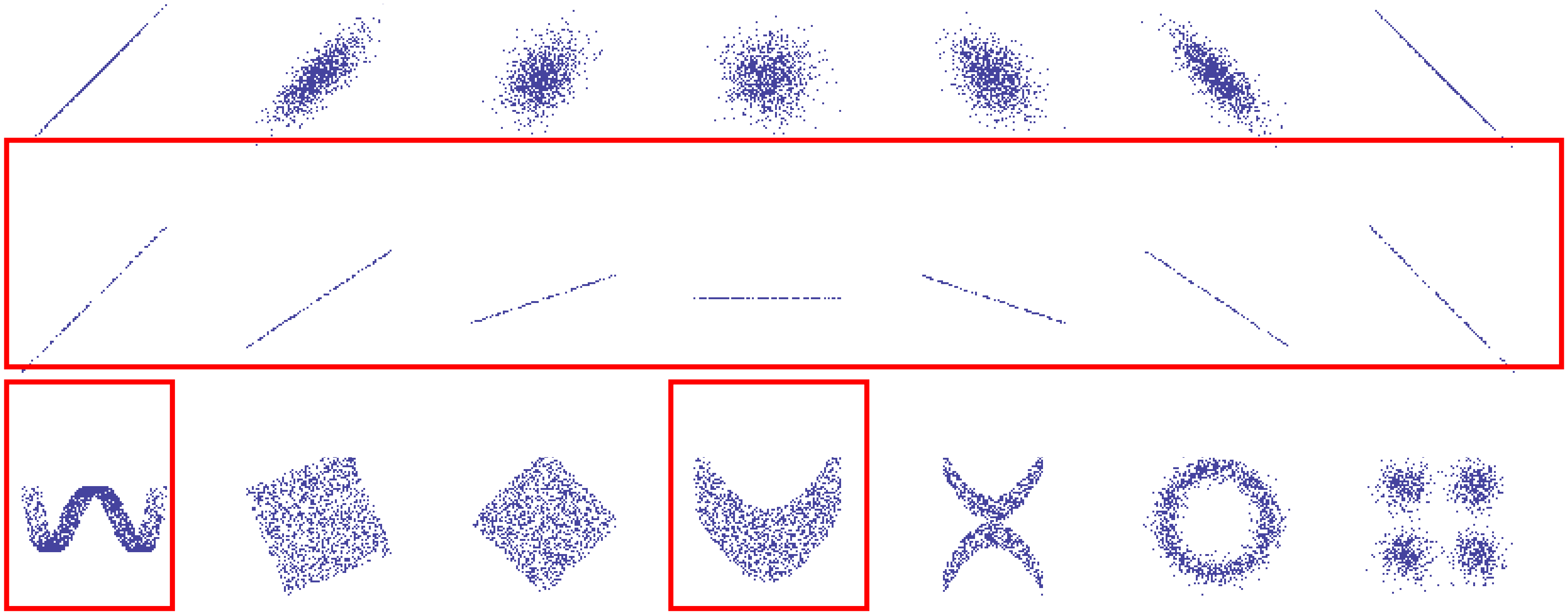
$$
\begin{array}{c c c}
 & \text{Bird species richness} & \text{Plant species richness} \\
1 & x_1 & l_1 \\
2 & x_2 & l_2 \\
3 & x_3 & l_3 \\
4 & x_4 & l_4 \\
\ldots & \ldots & \ldots \\
458 & x_{458} & l_{458}
\end{array}
$$



Correlation

Hypothesis testing

Plot the two variables in a Cartesian space

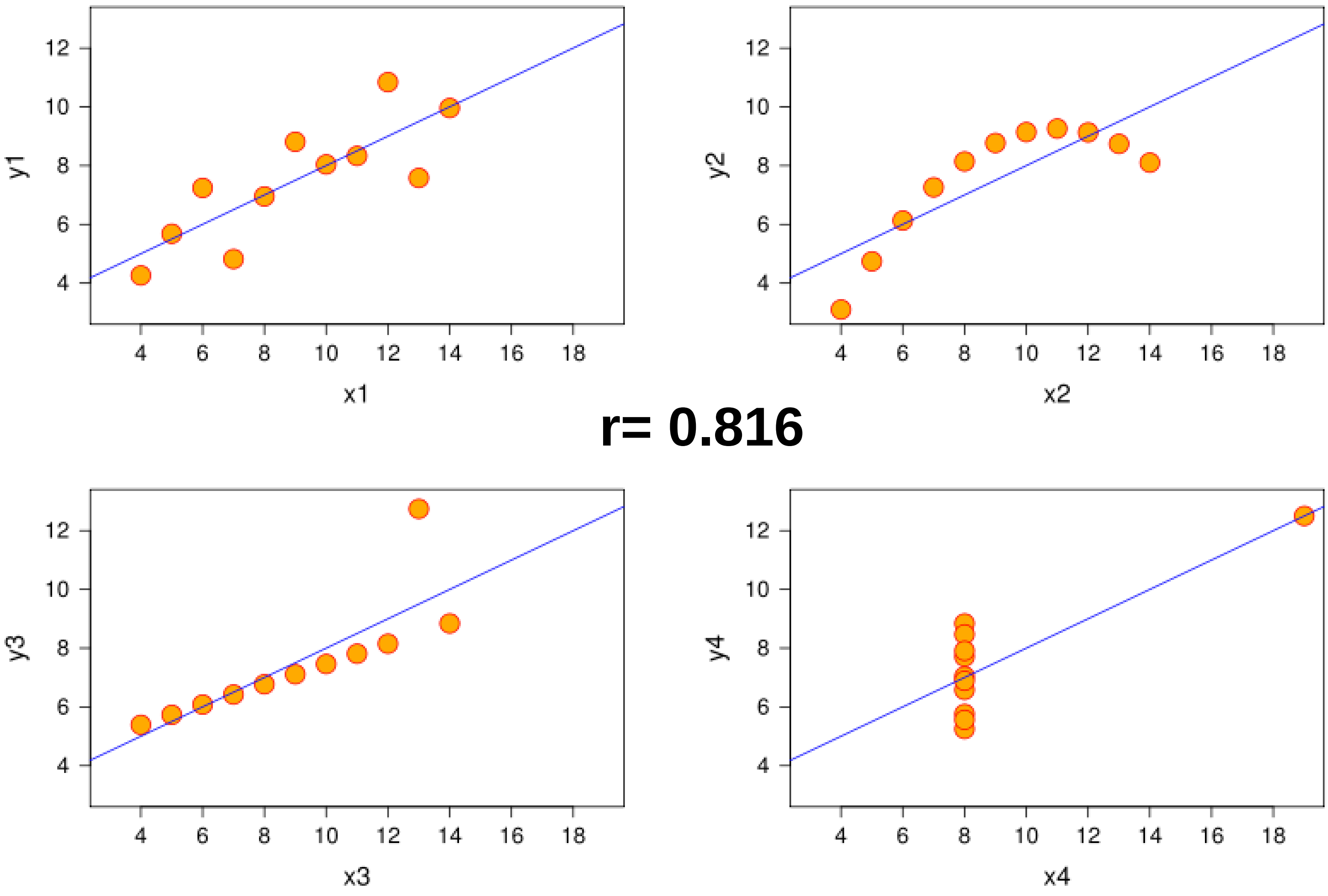


A correlation of +1 means that there is a **perfect positive LINEAR relationship** between variables.
A correlation of -1 means that there is a **perfect negative LINEAR relationship** between variables.
A correlation of 0 means there is **no LINEAR relationship** between the two variables.

Correlation

Hypothesis testing

Same correlation coefficient!



r= 0.816

Parametric correlation: when is significant?

Hypothesis testing

Pearson product-moment correlation coefficient

Correlation coefficient:

$$cor = \frac{\sum (xy)}{\sqrt{\sum x^2 \sum y^2}} \qquad SE_{cor} = \sqrt{\frac{(1 - cor^2)}{n - 2}}$$

Hypothesis testing using the t distribution:

Ho: Is cor = 0

H1: Is cor ≠ 0

$$t = \frac{cor}{SE_{cor}} \quad \Longrightarrow$$ **t critic value for d.f. = n-2**

Assumptions

- Two random variables from a random populations

- `cor()` detects ONLY linear relationships

Nonparametric correlation

Rank procedures

Distribution-free but less power

Spearman correlation index

$$cor.spearman = \frac{\sum (rank_x rank_y)}{\sqrt{\sum rank_x{}^2 \sum rank_y{}^2}}$$

The **Kendall tau rank correlation coefficient**

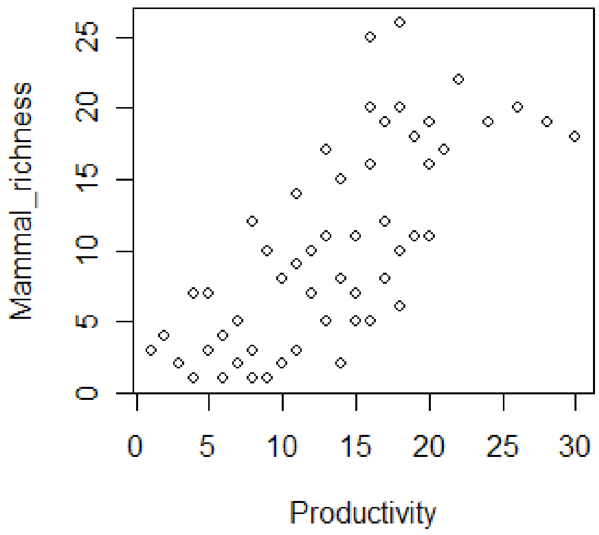
$$cor.kendall = \frac{4P}{n(n-1)} - 1$$

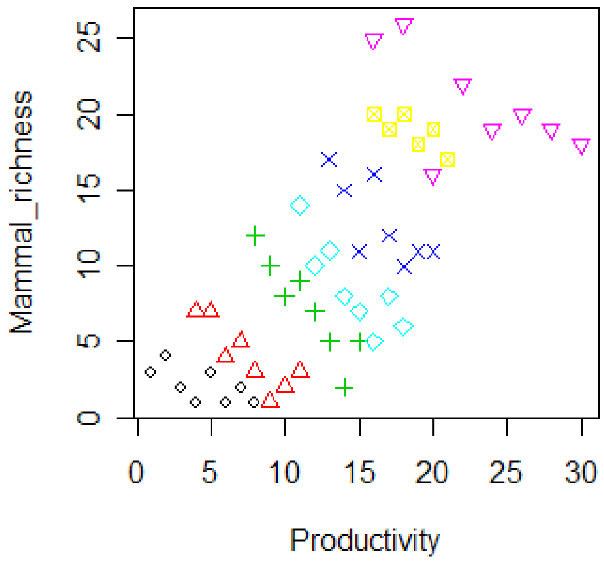
P is the number of concordant pairs
n is the total number of pairs

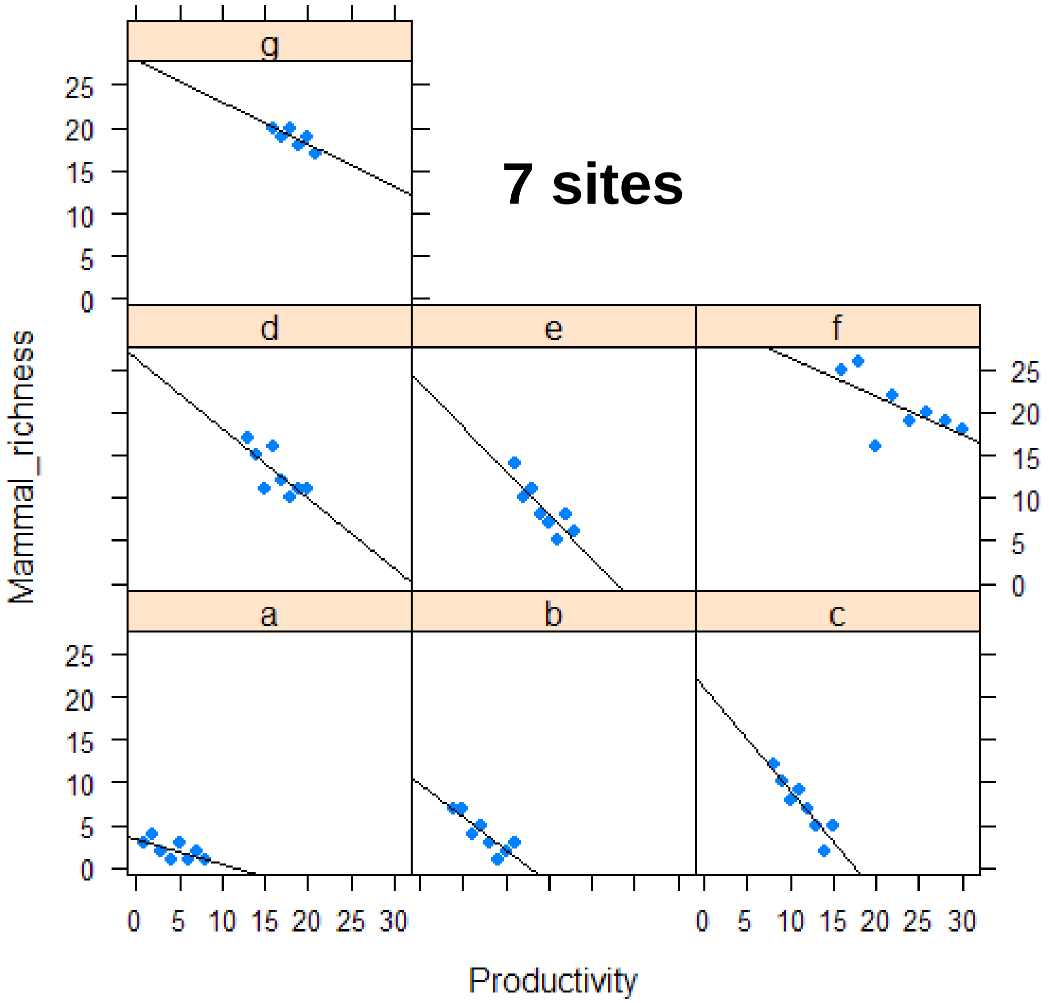
Scale-dependent correlation

Hypothesis testing

NB Don’t use grouped data to compute overall correlation!!!







7 sites

Statistical modelling

Hypothesis testing

MODEL

Generally speaking, a statistical model is a function of your explanatory variables to explain the variation in your response variable (y)

E.g. $Y=a+bx_1+cx_2+ dx_3$

Y= response variable (performance of the students)

x_i= explanatory variables (ability of the teacher, background, age)

The object is to determine the values of the parameters (a, b, c and d) in a specific model that lead to *the best fit of the model to the data*

The best model is the model that produces the least unexplained variation (the ***minimal residual deviance***), subject to the constraint that **all the parameters in the model should be statistically significant** (many ways to reach this!)

$$deviance = SS = \sum (y_i - mean)^2$$

Statistical modelling

Hypothesis testing

Getting started with complex statistical modeling

It is essential, that you can answer the following questions:

- Which of your variables is the response variable?

- Which are the explanatory variables?

- Are the explanatory variables continuous or categorical, or a mixture of both?

- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, or a category?

Statistical modelling: multicollinearity

Hypothesis testing

1. Multicollinearity

Correlation between predictors in a non-orthogonal multiple linear models

Confounding effects difficult to separate

Variables are not independent

This makes an important difference to our statistical modelling because, in orthogonal designs, the variation that is attributed to a given factor is constant, and does not depend upon the order in which factors are removed from the model.

In contrast, with non-orthogonal data, we find that the variation attributable to a given factor *does* depend upon the order in which factors are removed from the model

The order of variable selection makes a huge difference

Statistical modelling

Hypothesis testing

Getting started with complex statistical modeling

It is essential, that you can answer the following questions:

- Which of your variables is the response variable?

- Which are the explanatory variables?

- Are the explanatory variables continuous or categorical, or a mixture of both?

- What kind of response variable do you have: is it a continuous measurement, a count, a proportion, a time at death, or a category?

Statistical modelling

Each analysis estimate a MODEL

You want the model to be ***minimal*** (parsimony), and ***adequate*** (must describe a significant fraction of the variation in the data)

It is very important to understand that ***there is not just one model***.

- given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

Model building: estimate of parameters

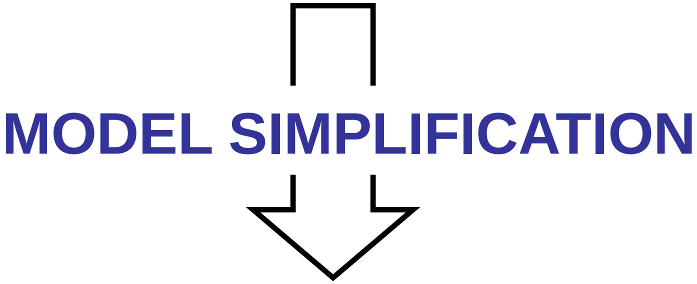
(slopes and level of factors)

Occam's Razor

Statistical modelling

Occam’s Razor

- Models should have as few parameters as possible;
- linear models should be preferred to non-linear models;
- experiments relying on few assumptions should be preferred to those relying on many;
- models should be pared down until they are *minimal adequate*;
- simple explanations should be preferred to complex explanations.



MODEL SIMPLIFICATION

The process of model simplification is an integral part of hypothesis testing in R. In general, **a variable is retained in the model only *if it causes a significant increase in deviance when it is removed from the current model***.

Statistical modelling: model simplification

Parsimony requires that the model should be as simple as possible. This means that the model should not contain any redundant parameters or factor levels.

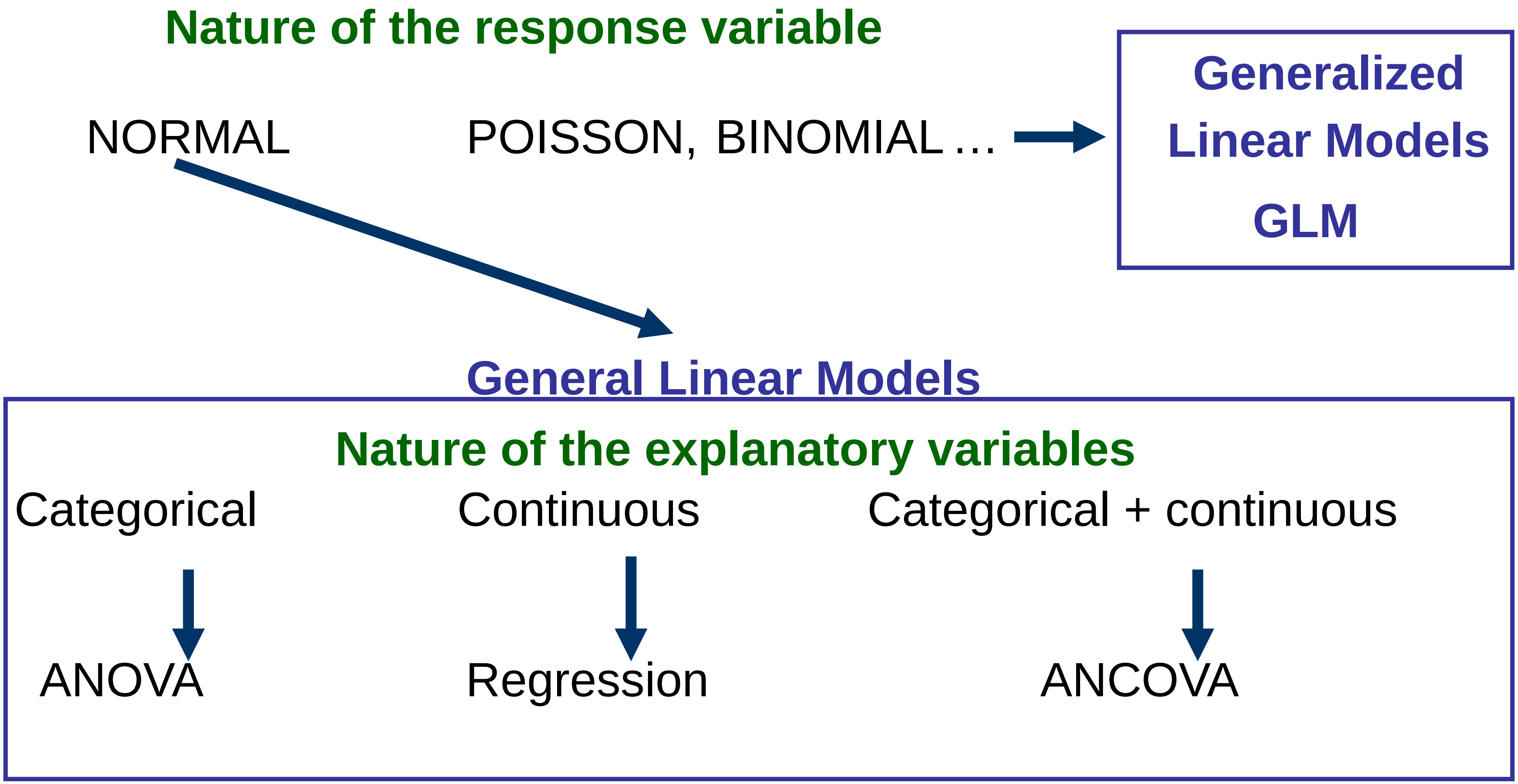
Model simplification

- remove non-significant interaction terms;
- remove non-significant quadratic or other non-linear terms;
- remove non-significant explanatory variables;
- group together factor levels that do not differ from one another;
- in ANCOVA, set non-significant slopes of continuous explanatory variables to zero.

Statistical modelling: model simplification

Step	Procedure	Interpretation
1	Fit the maximal model	Fit all the factors, interactions and covariates of interest. Note the residual deviance. If you are using Poisson or binomial errors, check for overdispersion and rescale if necessary.
2	Begin model simplification	Inspect the parameter estimates (e.g. using the R function `summary()`. Remove the least significant terms first (using `update -,)` **starting with the highest-order interactions.**
3	If the deletion causes an insignificant increase in deviance	Leave that term out of the model. Inspect the parameter values again. **Remove the least significant term remaining.**
4	If the deletion causes a significant increase in deviance	Put the term back in the model (using **`update +).`** These are the statistically significant terms as assessed by deletion from the maximal model.
5	Keep removing terms from the model	Repeat steps 3 or 4 until the model contains nothing but significant terms. **This is the minimal adequate model (MAM).** If none of the parameters is significant, then the minimal adequate model is the null model.

Statistical modelling: more than one parameter



Nature of the response variable

NORMAL

POISSON, BINOMIAL …

Generalized Linear Models

GLM

General Linear Models

Nature of the explanatory variables

Categorical

Continuous

Categorical + continuous

ANOVA

Regression

ANCOVA

LINEAR REGRESSION `lm()`

Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a **dependent variable** (response variable) and of one or more **independent continuous variables** (explanatory variables)

Assumptions

Independence: The Y-values and the error terms must be independent of each other.

Linearity between Y and X.

Normality: The populations of Y-values and the error terms are normally distributed for each level of the predictor variable x

Homogeneity of variance: The populations of Y-values and the error terms have the same variance at each level of the predictor variable x.

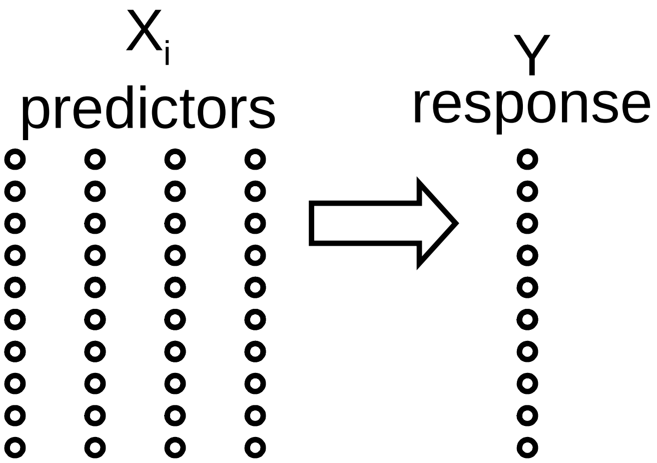
(**don't test for normality or heteroscedasticity**, **check the residuals instead!)**

LINEAR REGRESSION `lm()`

AIMS

1.To describe the linear relationships between Y and X_i (**EXPLANATORY APPROACH**) and to quantify how much of the total variation in Y can be explained by the linear relationship with X_i.

2. To predict new values of Y from new values of X_i (**PREDICTIVE APPROACH**)



$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

⇩

We estimate one INTERCEPT and one or more SLOPES

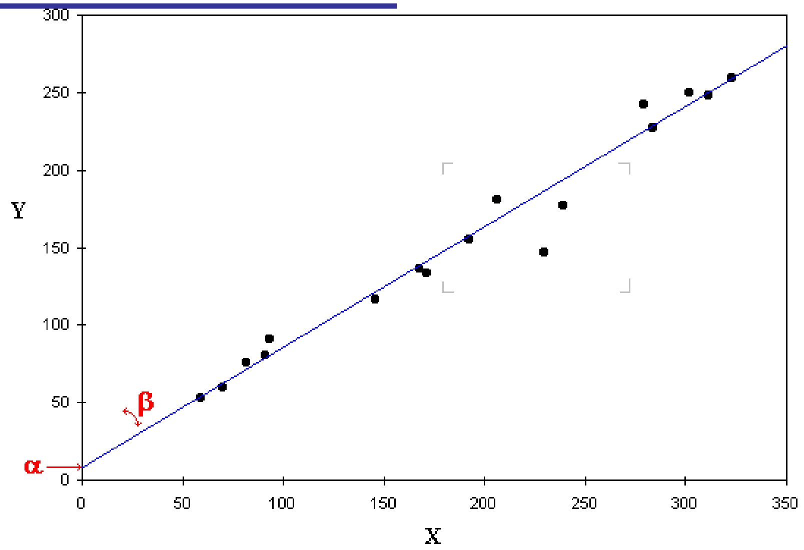
SIMPLE LINEAR REGRESSION

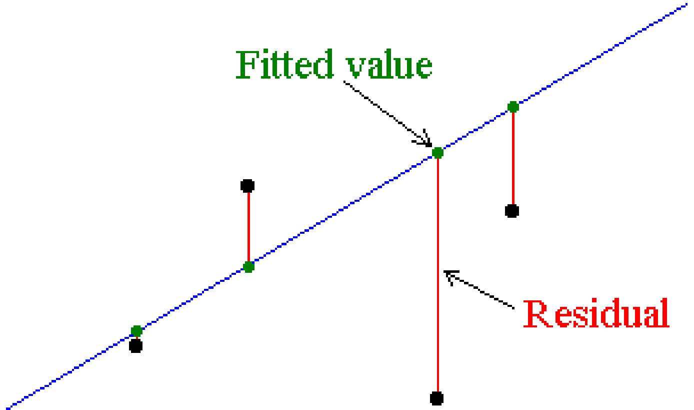
Estimating parameters

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\beta = \frac{\Sigma\,[(x_i - x_{mean})(y_i - y_{mean})]}{\Sigma\,(x_i - x_{mean})^2}$$

$$\alpha = y_{mean} - \beta * x_{mean}$$





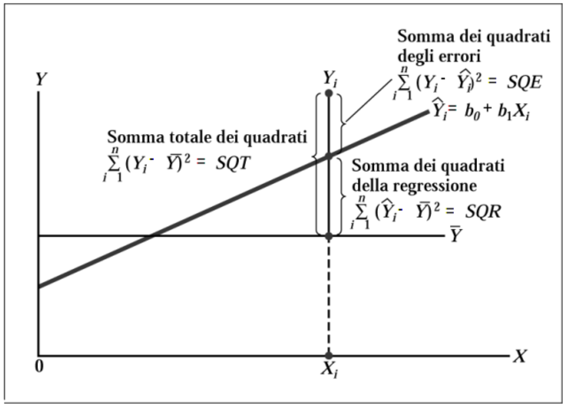
Measure of goodness-of-fit

Total SS = $\Sigma(y_{\text{fitted } i} - y_{mean})^2$

Model SS = $\Sigma(y_{\text{predicted } i} - y_{mean})^2$

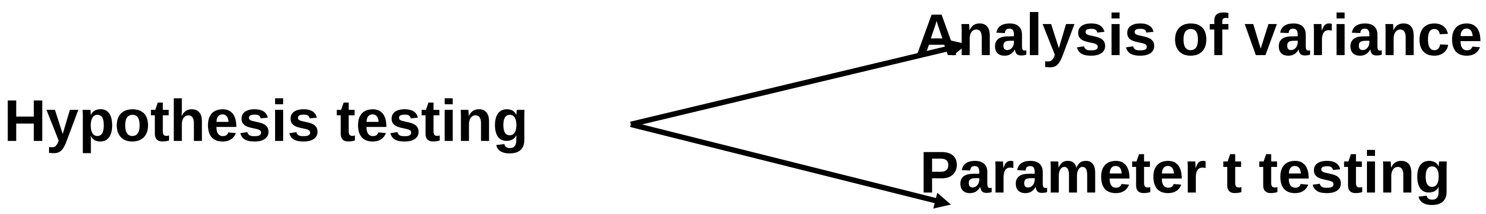
Residual SS = Total SS - Model SS

R² = Model SS /Total SS → **Explained variation**


Y
Somma dei quadrati degli errori
$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = SQE$
Y_i
$\hat{Y}_i = b_0 + b_1 X_i$
Somma totale dei quadrati
$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SQT$
Somma dei quadrati della regressione
$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = SQR$
$\bar{Y}$
0
X_i
X

Le misure di variabilità nel modello di regressione

SIMPLE LINEAR REGRESSION



Hypothesis testing

Analysis of variance

Parameter t testing

Ho: $\beta = 0$ (There is no relation between X and Y)

H1: $\beta \neq 0$

Analysis of variance (test the model!)

$F_{1,\, n-2}$ = (Model SS/1) / (Residual SS/n-2)

Parameter t testing (test the single parameter!)

We must measure the unreliability associated with each of the estimated parameters (i.e. we need the standard errors)

$SE(\beta) = [(\text{residual SS}/(n-2))/\Sigma(x_i - x_{mean})]^2$

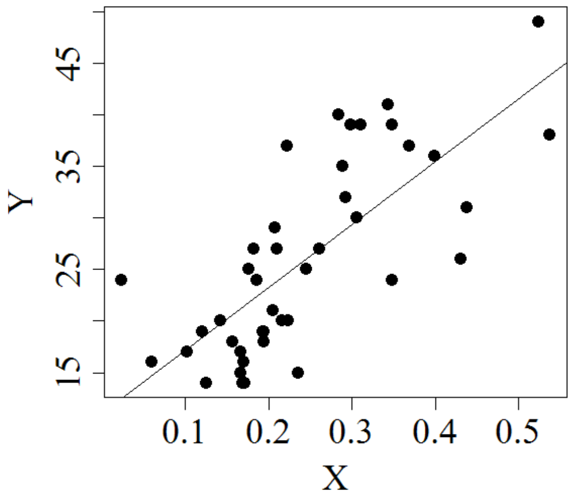
$t = (\beta - 0) / SE(\beta)$

SIMPLE LINEAR REGRESSION: example 1

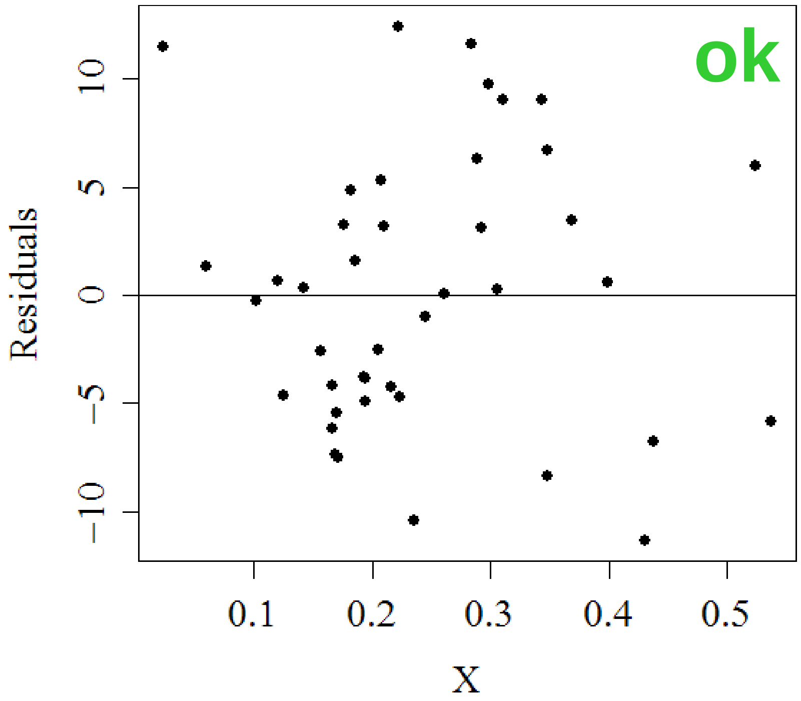
If the model is significant, then **model checking**

1. Linearity between X and Y?

Regression model



Residuals vs predictor



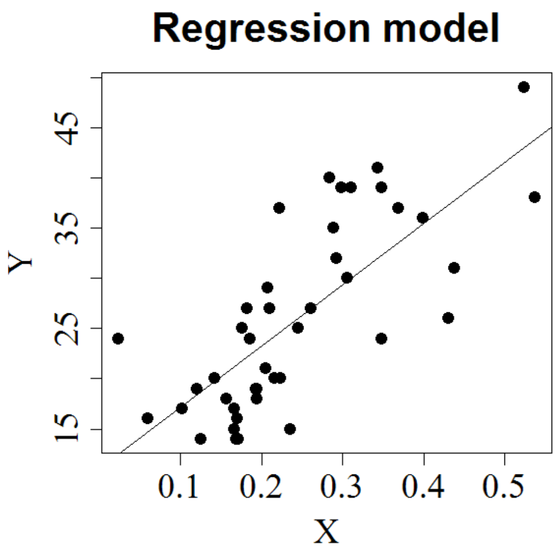
ok

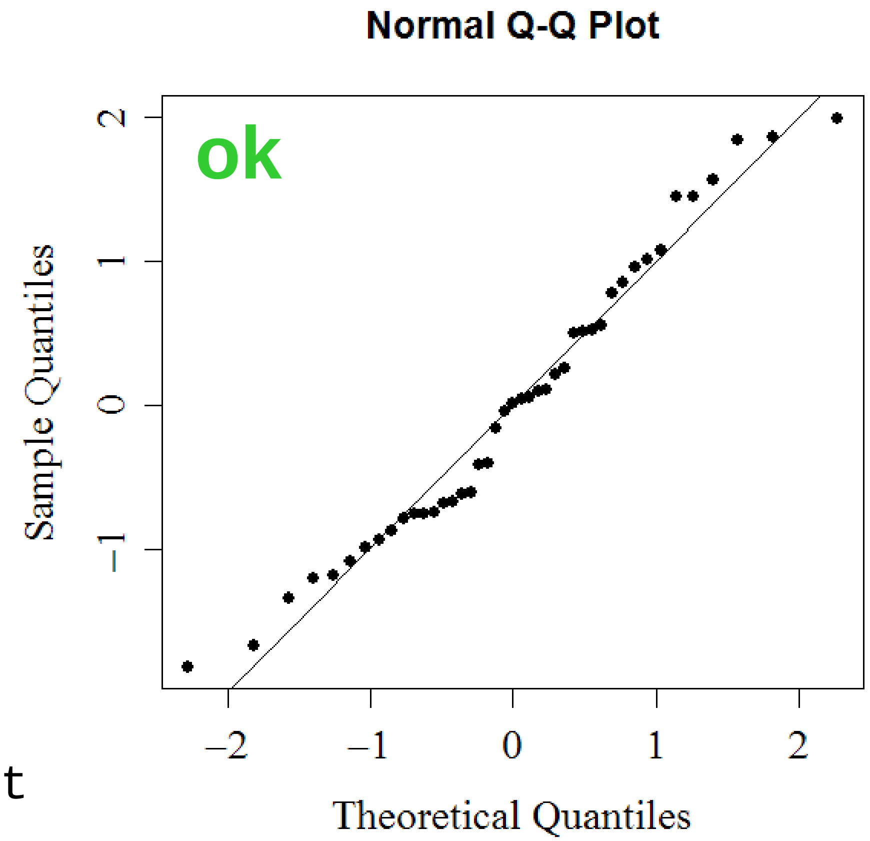
No patterns in the residuals vs. predictor plot

SIMPLE LINEAR REGRESSION: example 1

2. Normality of the residuals

Q-Q plot + Shapiro-Wilk test on the residuals





ok

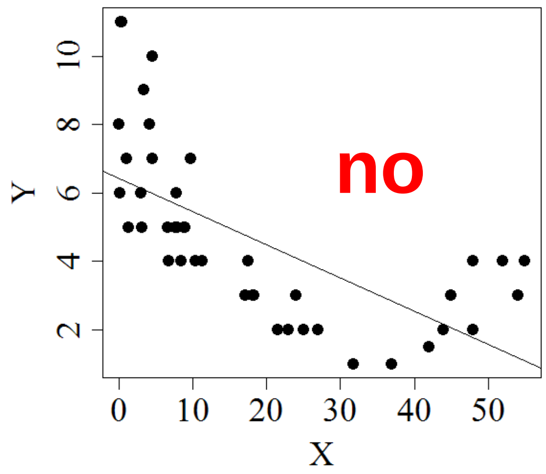
```
> shapiro.test(residuals)
Shapiro-Wilk normality test
data:  residuals
W = 0.9669, p-value = 0.2461
```

ok

SIMPLE LINEAR REGRESSION: example 2

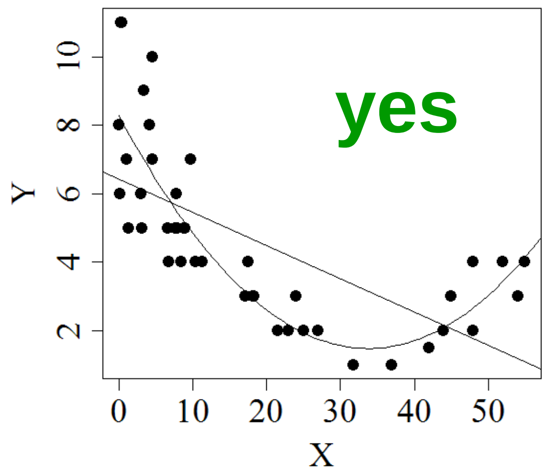
1. Linearity between X and Y?

Regression model



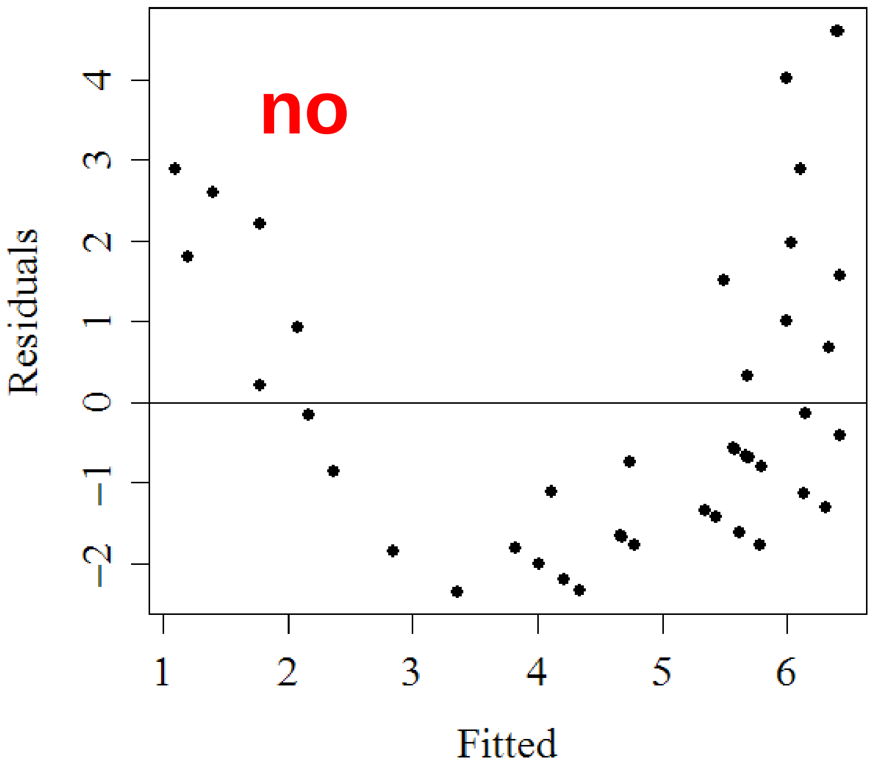
no

Quadratic regression



yes

Residuals vs fitted



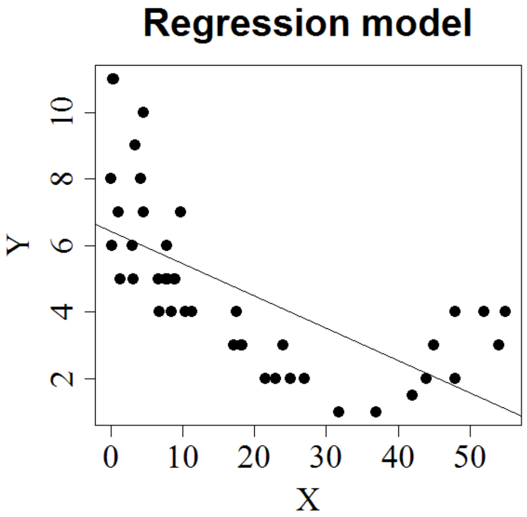
no

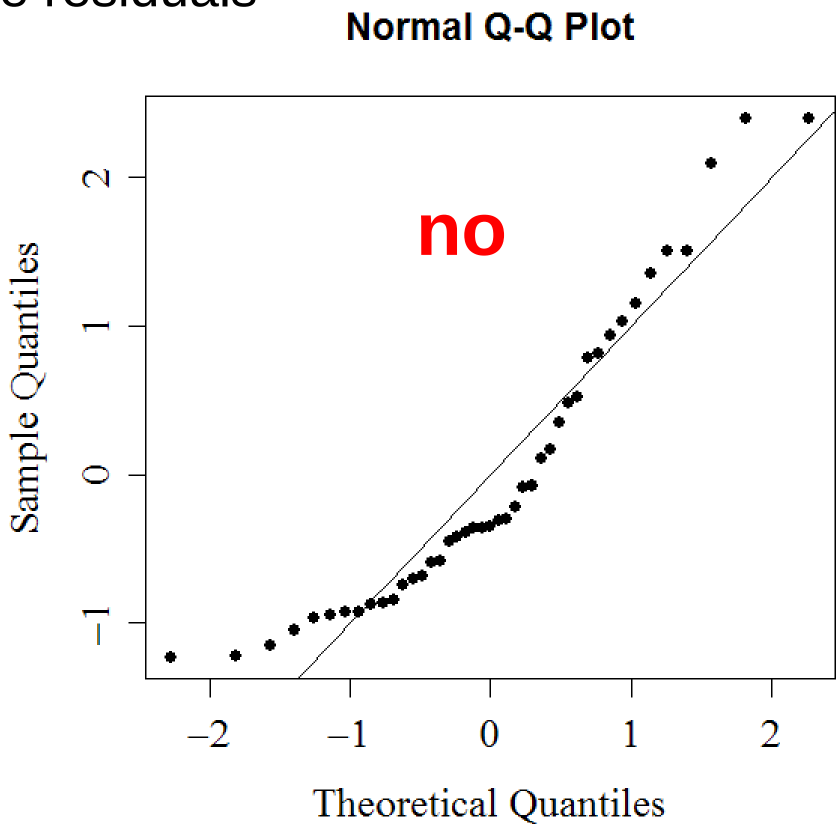
NO LINEARITY between X and Y

SIMPLE LINEAR REGRESSION: example 2

2. Normality of the residuals

Q-Q plot + Shapiro-Wilk test on the residuals





```
> shapiro.test(residuals)
Shapiro-Wilk normality test
data:  residuals
W = 0.8994, p-value = 0.001199
```

no

SIMPLE LINEAR REGRESSION: example 2

How to deal with non-linearity and non-normality situations?

Transformation of the data

-Box-cox transformation (power transformation of the response)

-Square-root transformation

-Log transformation

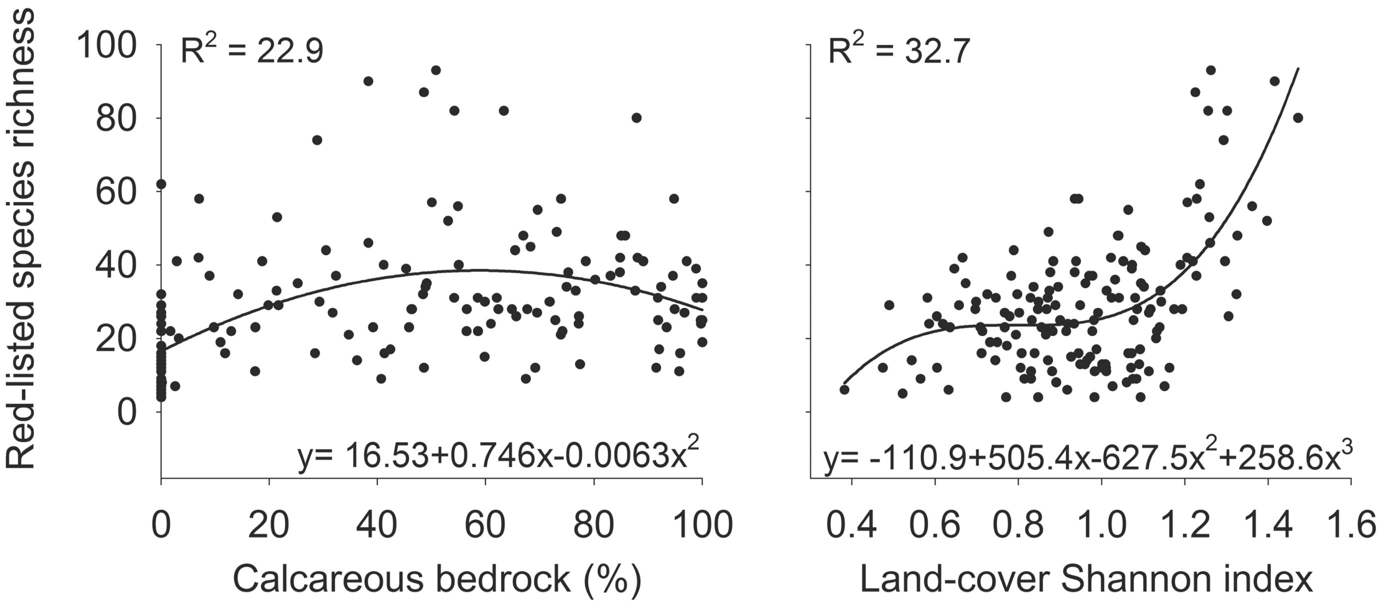
-Arcsin transformation

Polynomial regression

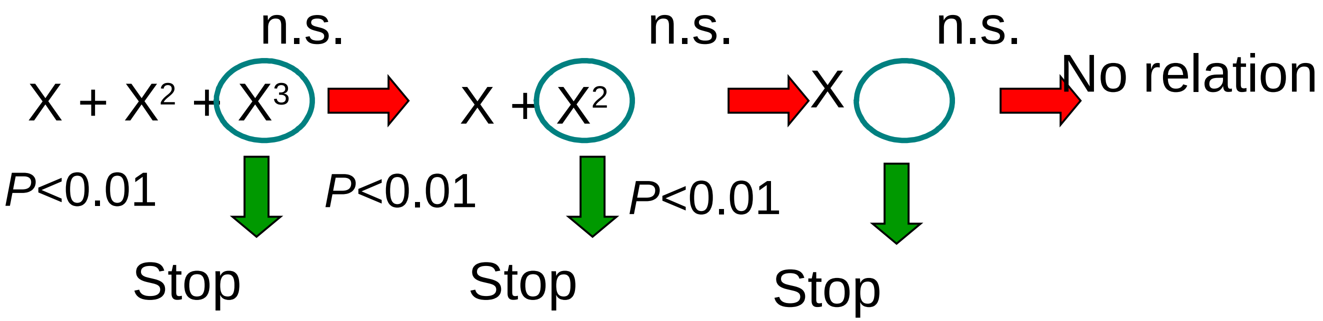
Regression with multiple terms (linear, quadratic, and cubic)

$Y = a + b_1X + b_2X^2 + b_3X^3 + \text{error}$ **X is one variable!!!**

POLYNOMIAL REGRESSION: one x, n parameters



Hierarchy in the testing (always test the highest)!!!!



n.s. n.s. n.s.

$X + X^2 + X^3$ → $X + X^2$ → X → No relation

$P<0.01$ $P<0.01$ $P<0.01$

Stop Stop Stop

NB Do not delete lower terms even if non-significant

MULTIPLE LINEAR REGRESSION: more than one x

Multiple regression

Regression with two or more variables

$Y = a + b_1X_1 + b_2X_2 + \ldots + b_iX_i$ + quadratic and cubic terms + interactions+ error

Assumptions

Same assumptions as in the simple linear regression!!!

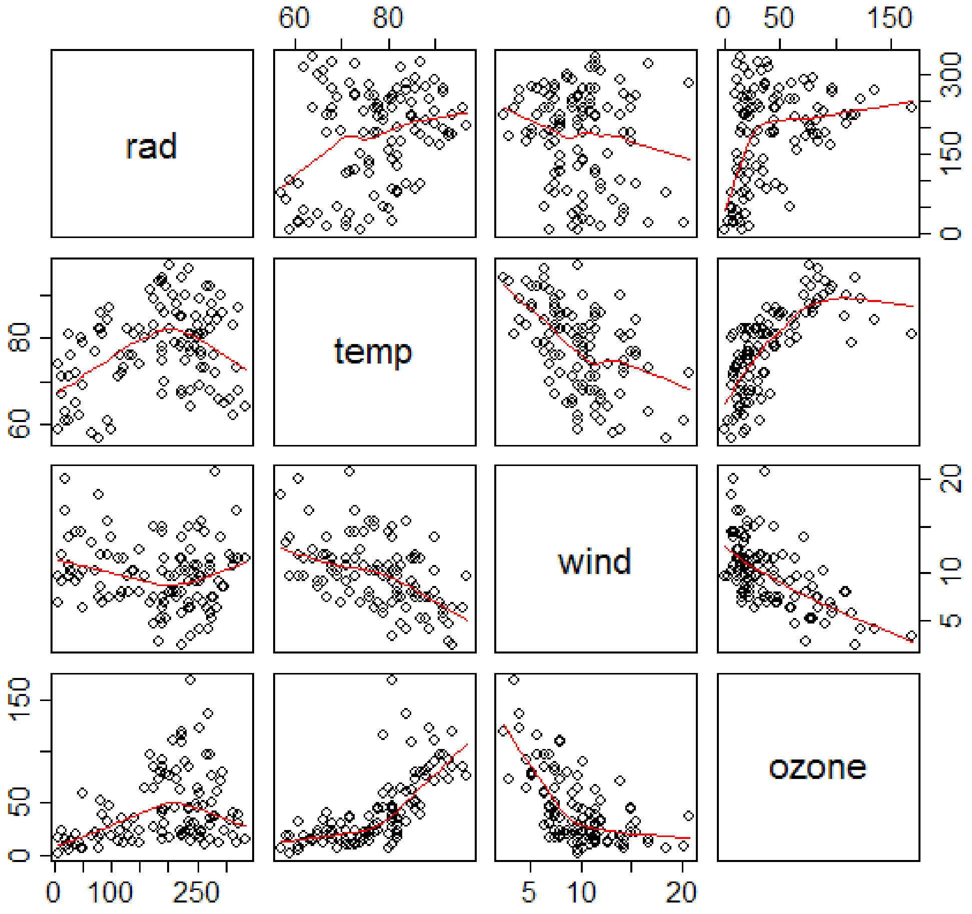
The Multiple Regression Model

There are important issues involved in carrying out a multiple regression:

- which explanatory variables to include (**VARIABLE SELECTION**);
- **NON-LINEARITY** in the response to the explanatory variables;
- **INTERACTIONS** between explanatory variables;
- correlation between explanatory variables (**COLLINEARITY**);
- **RELATIVE IMPORTANCE** of variables

MULTIPLE LINEAR REGRESSION: more than one x

Let’s begin with an example from air pollution studies. How is ozone concentration related to wind speed, air temperature and the intensity of solar radiation?



Presence of non-linearity and possible interactions

MULTIPLE LINEAR REGRESSION: more than one x

Start with a complex model with interactions and quadratic and cubic terms

Model simplification (Occam's razor)

Minimum Adequate Model

How to carry out a model simplification in multiple regression

1. Remove non-significant interaction terms.
2. Remove non-significant quadratic or other non-linear terms.
3. Remove non-significant explanatory variables.
4. Amalgamate explanatory variables that have similar parameter values.

MULTIPLE LINEAR REGRESSION: more than one x

Start with the most complicate model (it is one approach)

model1<lm(ozone ~ temp*wind*rad+I(rad^2)+I(temp^2+I(wind^2))

	Estimate	Std.Error	t	Pr(>t)	
(Intercept)	5.7E+02	2.1E+02	2.74	0.01	**
temp	-1.1E+01	4.3E+00	-2.50	0.01	*
wind	-3.2E+01	1.2E+01	-2.76	0.01	**
rad	-3.1E-01	5.6E-01	-0.56	0.58	
I(rad^2)	-3.6E-04	2.6E-04	-1.41	0.16	
I(temp^2)	5.8E-02	2.4E-02	2.44	0.02	*
I(wind^2)	6.1E-01	1.5E-01	4.16	0.00	***
temp:wind	2.4E-01	1.4E-01	1.74	0.09	
temp:rad	8.4E-03	7.5E-03	1.12	0.27	
wind:rad	2.1E-02	4.9E-02	0.42	0.68	
temp:wind:rad	-4.3E-04	6.6E-04	-0.66	0.51	

Delete only the highest interaction temp:wind:rad

**!!!!!!
We cannot delete these terms
!!!!!!!**

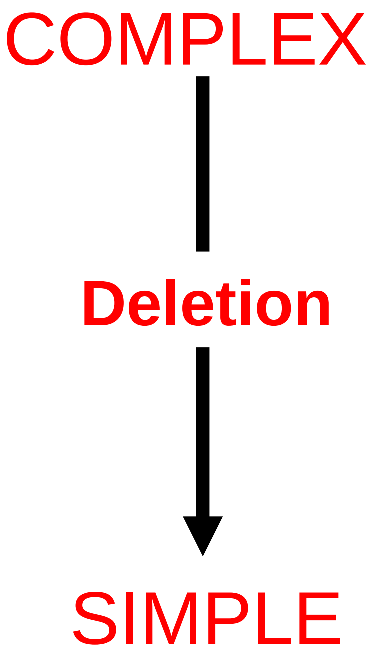
MULTIPLE LINEAR REGRESSION: more than one x

Manual model simplification

(It is one of the many philosophies)
Deletion the non-significant terms one by one:

Hierarchy in the deletion:
1. Highest interactions
2. Cubic terms
3. Quadratic terms
4. Linear terms

At each deletion test:
Is the fit of a
simpler model worse?


COMPLEX
Deletion
SIMPLE

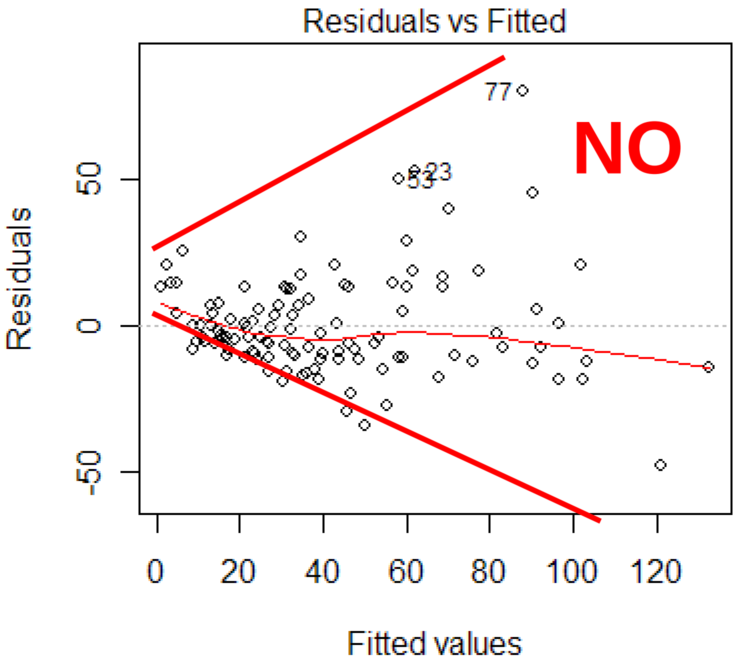
IMPORTANT!!!

If you have quadratic and cubic terms significant you cannot delete the linear or the quadratic term even if they are not significant

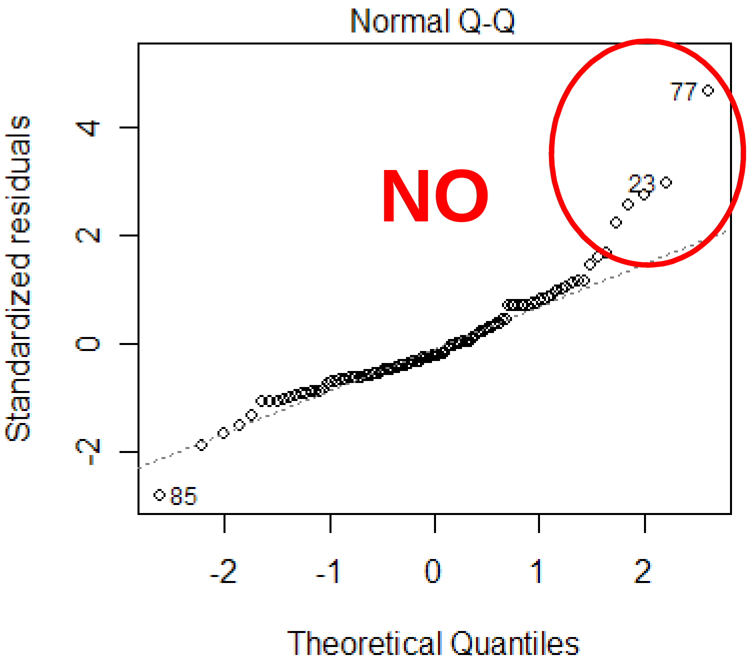
If you have an interaction significant you cannot delete the main terms even if they are not significant

MULTIPLE LINEAR REGRESSION: more than one x

Once we reached the MAM we must check the assumptions



Variance tends to increase with y

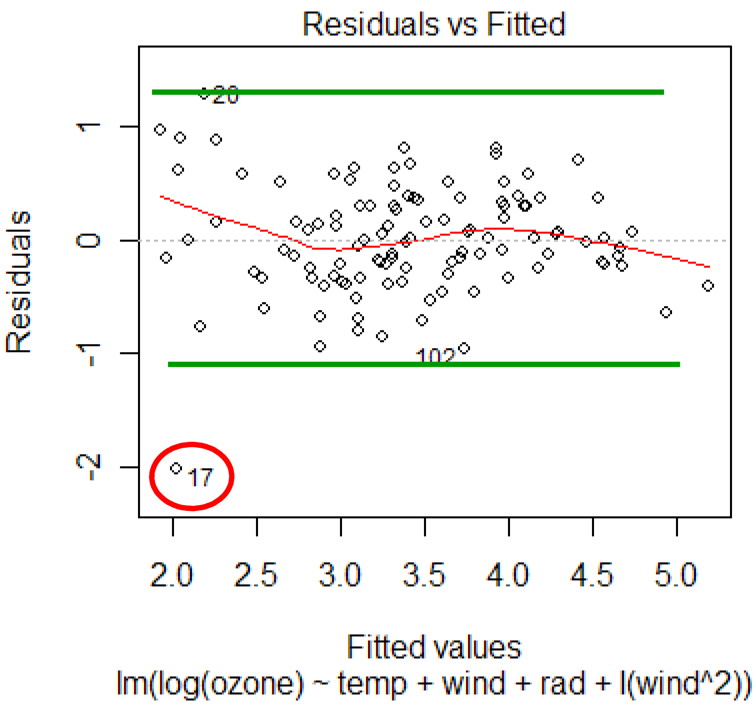


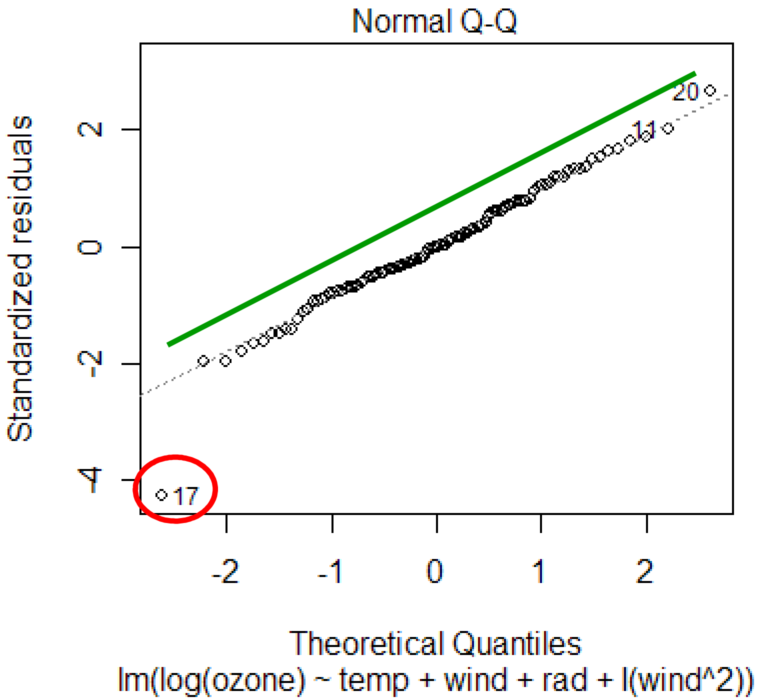
Non-normal errors

We can transform the data (e.g. Log-transformation of y)

```
model<lm( log(ozone) ~ temp + wind + rad + I(wind²))
```

MULTIPLE LINEAR REGRESSION: more than one x

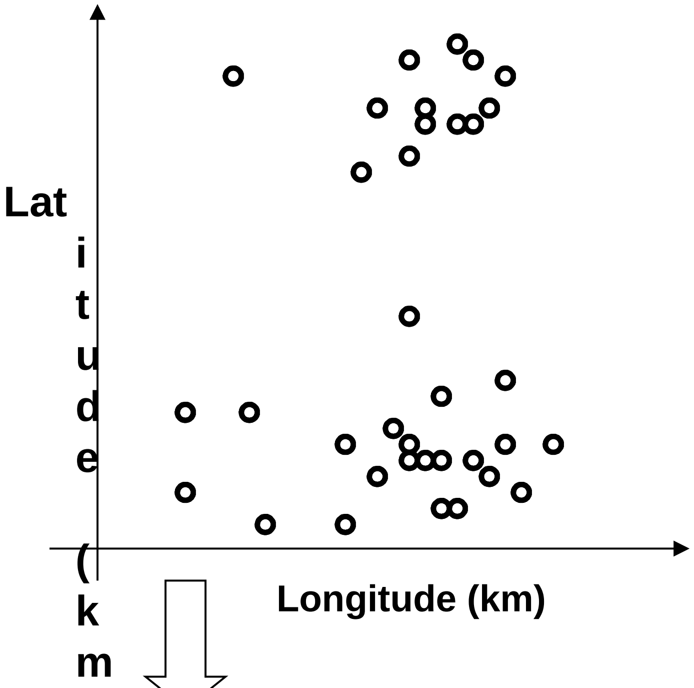




The log-transformation has improved our model but maybe there is an outlier

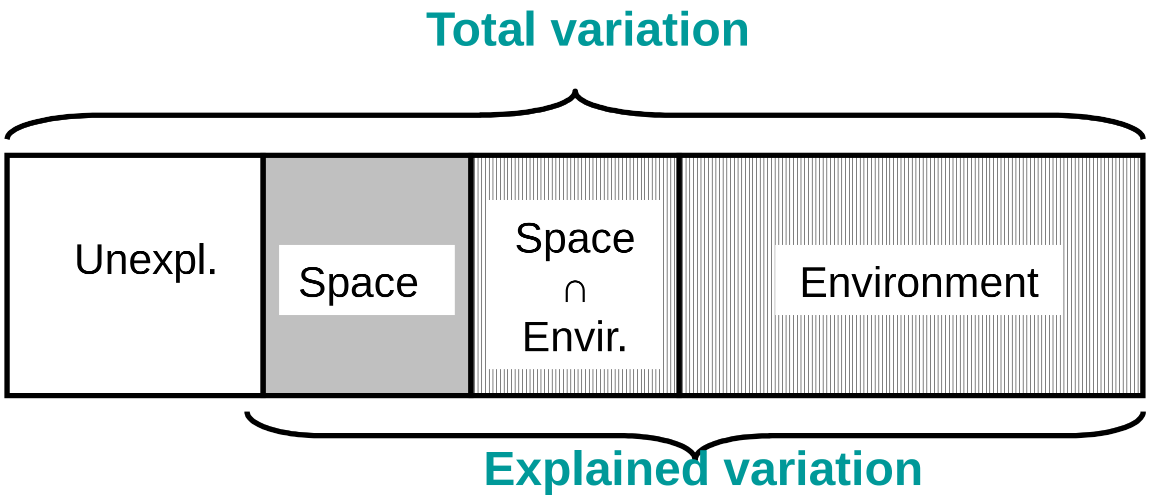
VARIATION PARTITIONING

Relative importance of groups of explanatory variables



R^2= 76% (**TOTAL EXPLAINED VARIATION**)

What is space and what is environment?



Total variation

Unexpl. | Space | Space ∩ Envir. | Environment

Explained variation

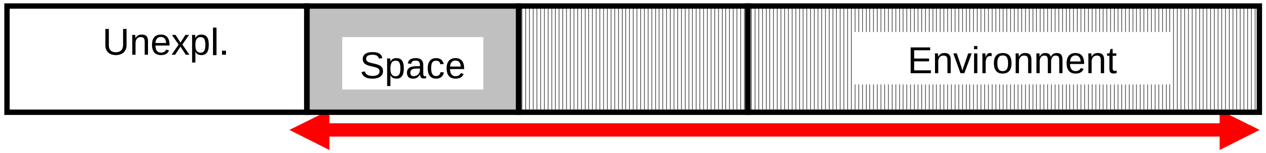
```
Full.model<lm(species ~ environment i + space i)
```

Response variable: orthopteran species richness

Explanatory variable: **SPACE (latitude + longitude)** +
ENVIRONMENT (temperature + land-cover heterogeneity)

VARIATION PARTITIONING: **`varpart(vegan)`**

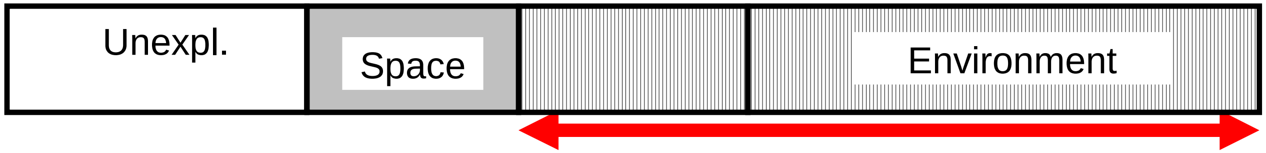
```
Full.model<lm(SPECIES ~ temp + het + lat + long)
```



TVE=76%

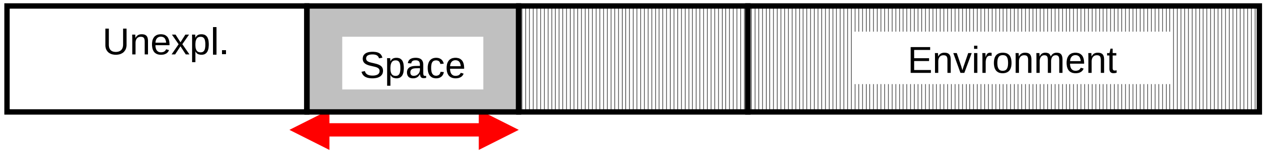
```
Env.model<lm(SPECIES ~ temp + het)
```

⇒ env.residuals



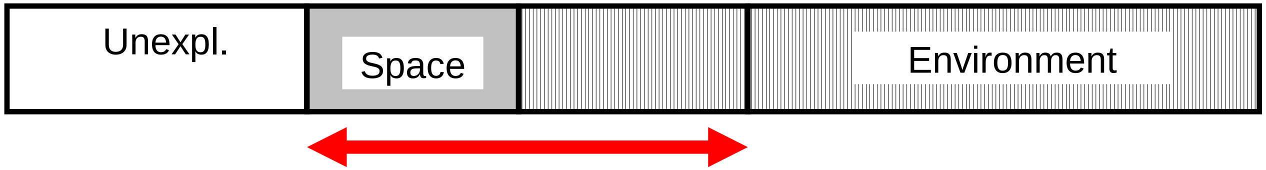
```
Pure.Space.model<lm(ENV.RESIDUALS ~ lat + long)
```

⇒ **VE=15%**



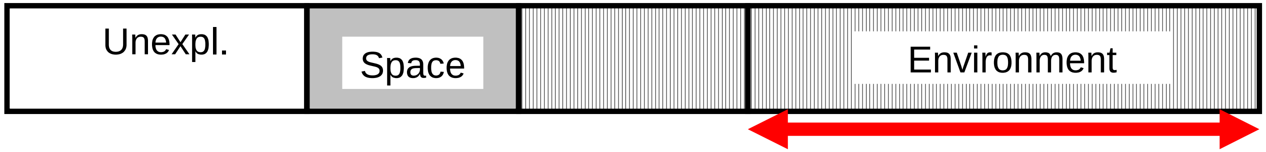
```
Space.model<lm(SPECIES ~ lat + long)
```

⇒ space.residuals



```
Pure.env.model<lm(SPACE.RESIDUALS ~ tem + het)
```

⇒ **VE=40%**


Unexpl.
Space
Environment