An Introduction to Statistical Data Analysis

presented by
Christoph Scherber

Georg-August-University of Goettingen
Department of Crop Science (DNPW)
Agroecology

This manual is based on a lecture series that was given by C.S.
at the Max-Planck-Institute for Chemical Ecology in Jena (Germany).

The content closely follows Mick Crawley´s famous "The GLIM Course"
that is held every year at Silwood Park, London, UK.

I thank Prof. Mick Crawley (Imperial College, London, UK) for
introducing me to statistical data analysis.

Statistics - Lecture 1

Statistics = Techniques of

- Collecting
- Analysing data
- Drawing conclusions from

"A mode of thought"

⇨ will change the way you do science

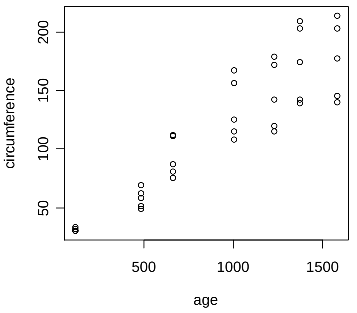
Getting started

Vocabulary: Response variable, Explanatory Variable

- Which are the response and explanatory variables?
- Type of explanatory variable
- Type of response variable

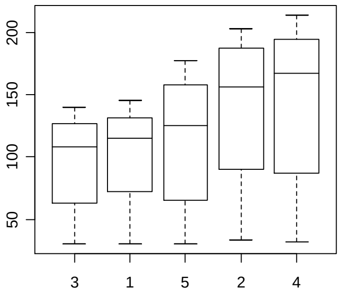
Response Variable		
Continuous	Weight, height, length, temperature, concentration	
Count (Whole numbers, integers)	Number of individuals, days, cells; zero is a common value	
Proportion	Percentage mortality, infection rate, proportion responding to a treatment; percent leaf area eaten.	
Explanatory Variable	Continuous	Weight, height, length, temperature, concentration

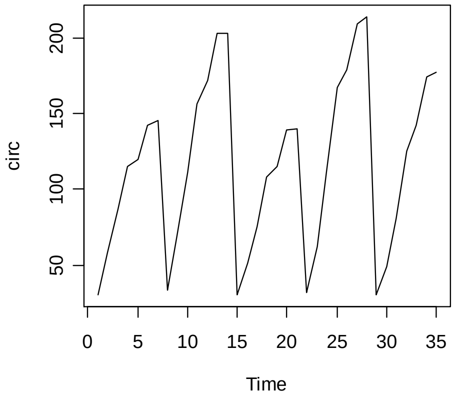
Scatterplot



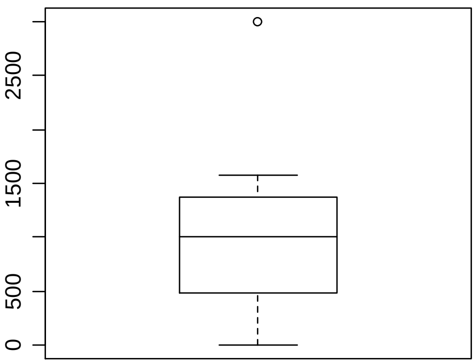
Response Variable		
Continuous	Weight, height, length, temperature, concentration	
Count (Whole numbers, integers)	Number of individuals, days, cells; zero is a common value	
Proportion	Percentage mortality, infection rate, proportion responding to a treatment; percent leaf area eaten.	
Explanatory Variable	Categorical (Factor with levels)	Species, Clone, Genotype,Treatment, Diet, Growth Chamber, Habitat

Box-and-whisker-plot





Explanatory variable = time => Time series



outliers

1.5 inter-quartile range

75%

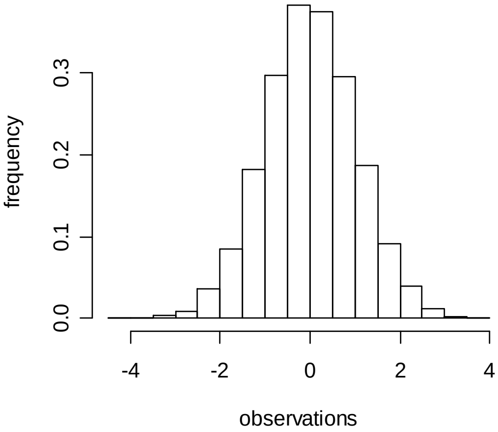
50% (Median)

25%

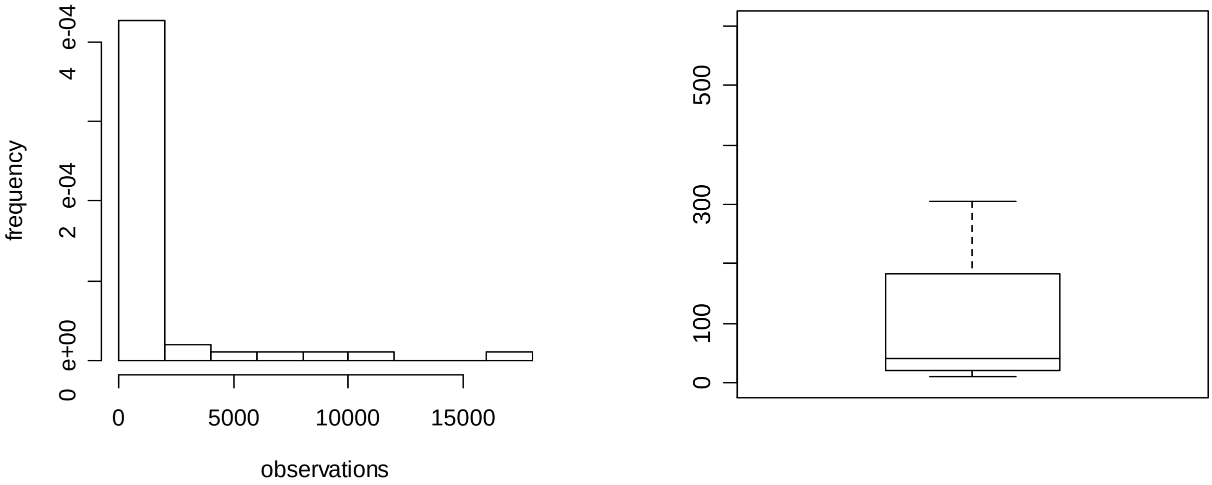
1.5 inter-quartile range

outliers

A typical box-and-whisker-plot
=> detect skewness; here: skewed to the left!



Normally distributed
mean=median=0



positive skew = skew to the right (long right tails)

Parameters + variables in models

Model = a way of describing the behaviour of a process in order to predict its future or understand its past

- Words
- Clay or Wood
- a mathematical relationship (mostly equations), e.g. $y = a + b\ x$
- **Statistical models** are fitted to data; they are used to **describe** a given set of data

- response variable ~ explanatory variable(s)

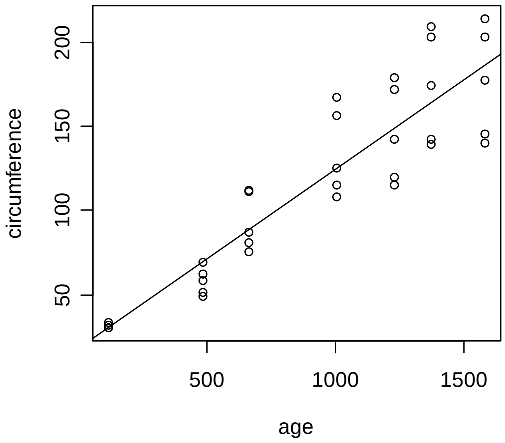
A model should be as simple as possible (but no simpler).

⇨ Minimal adequate model, model simplification
⇨ Prefer linear to non-linear models
⇨ as few parameters as possible
⇨ prefer simple explanations to complex ones
⇨ factor level reduction (in ANOVA and ANCOVA)

Variables = those elements of the model that are changing and whose behaviour is to be predicted by the model; e.g. x, y

Parameters = usually constants in the model, e.g. a, b

y=a + bx => Type of model: two-parameter linear model



y = a + bx
a...intercept
b...slope (steep, shallow...)
b = $\frac{\Delta y}{\Delta x}$

What test to do?

Explanatory Variable	**Response Variable**		
	Continuous	*Count*	*Proportion*
Continuous	Regression	• square-root transformation • GLM (Log-linear Regression)	• arcsine transformation • GLM (Logistic regression)
Categorical	Student´s t; ANOVA	Contingency Table	• arcsine transformation • GLM (Logistic analysis of deviance)
Continuous and Categorical	ANCOVA	• square-root transformation • GLM (Poisson Errors)	• arcsine transformation • GLM (Binomial Errors)
Time	Time Series Analysis		

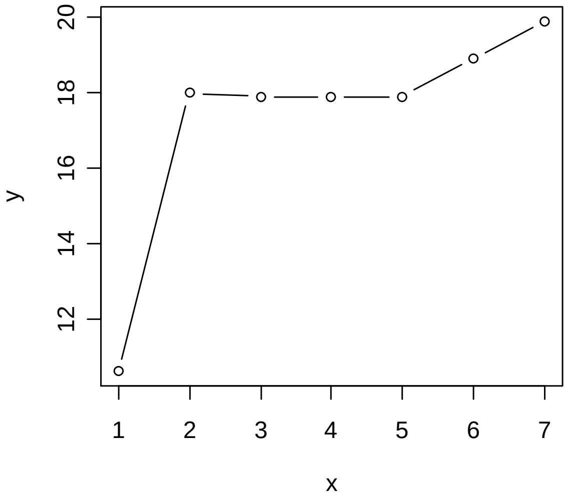
Experimental Design

Everything varies!
So finding differences is simply uninteresting

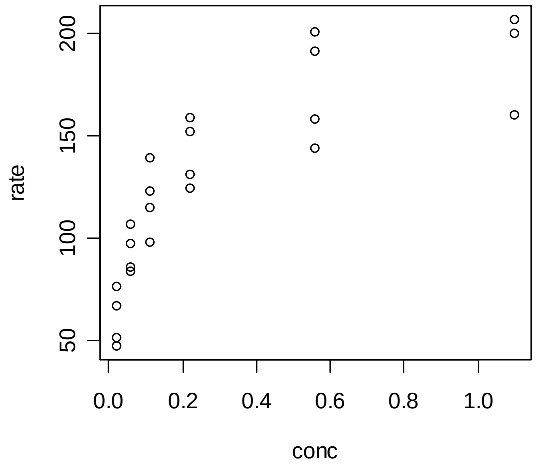
We need techniques to distinguish between interesting & uninteresting variation

e.g. “Dolly gets arthritis aged 5”

- ⇨ n=1
- ⇨ no hypothesis
- ⇨ no control (similar but not cloned)
- ⇨ no randomization



e.g. ion uptake in plants
=> we see some pattern, but there´s no replication!



In order to be certain about an observed pattern,
we need information about variation in the data

<u>The 2 R´s:</u>
Replication and Randomization!

<u>Replication</u>

⇨ to improve reliability

Replicates are <u>independent</u> repeats

How many? Do a pilot study to find out about

⇨ variance

⇨ most suitable response variable

⇨ magnitude of responses to treatment(s)

Two rules of thumb:

⇨ let n be >5

⇨ if possible, let n be at least 30

Power Analysis

For a **power of 0.8** to detect a significant difference (with Type I Error rate of 0.05), we get:

$n = 8\frac{\text{variance}}{(\text{difference})^2}$ i.e. we need to know the **variability**, and the **size of the difference** we want to detect.

so, e.g.,
difference = 2.0;
variance 10.0
=> n=8 (10/4) = 20

Blocking to reduce unexplained variation

SST

- ⇨ treatment
- ⇨ Block effects
- ⇨ unexplained (Error)

Randomization

- ⇨ to reduce bias

In practice:

- ⇨ Tables
- ⇨ Flip a coin, throw a dice
- ⇨ Random number generator
- ⇨ number each item !
- ⇨ alternative: stratified sampling; matched pairs

e.g. grow plants in the greenhouse
group them in

small – medium - large

draw at random from each group and transplant together

Aim: The unexplained variation should be as small as possible!

Important issues:

- ✓ replication
- ✓ randomization
- ✓ controls: no control, no conclusions
- ✓ avoid pseudoreplication

  temporal: repeated measures from the same individual
  spatial: several measurements from the same vicinity
  Look at error degrees of freedom!

  e.g. feeding trial with 10 petri dishes
  5 controls, 5 JA-treated
  each 4 leaf pieces (exchanged once)

  4 x 2 x 10 = 80 replicates??? No! its´s only 10!

  solutions: average it away; separate analyses; time-series analysis; mixed-effects models

- ✓ measure initial conditions:

demonstrate that experimental units really were alike at the beginning of the experiment!
use this also to check efficiency of randomization

Types of Experimental Design

Aim: Interspersion of replicates and treatments!

„Bad Designs“

□■ **no replication**

□□□□■■■■ **clumped segregation**
(totally uninformative)

□□□□ ■■■■ **isolative segregation**
(growth chambers etc)

□■□■ □■□■ **systematic**
(problem: periodic variations)

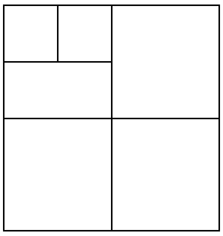
“Good” Designs

□■ □■ ■□ □■ **randomized block**
(matched pairs)

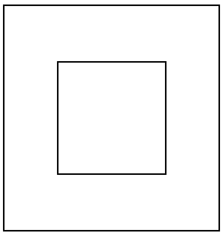
□■□■□□■■ **completely randomized**
(if enough time, space, money)

A	C	D	B
D	A	B	C
B	D	C	A
C	B	A	D

Latin Square
(in case of 2 gradients)



Split-Plot
very common!
Lab benches, Greenhouses,
Petri Dishes, Microarrays



Nested Design
especially in medicine: liver
samples from individual mice

Problem: small experiments => symmetrical by chance

Analysing Experimental Data

⇨ Estimate parameters of models
⇨ Hypothesis testing: Are estimated parameters significantly different from one another (from theory) ?

Estimates of Central Tendency

(1) **Mode**: most common value
(2) **Median**: splits data (y) into two equal halves

e.g. {3,7,9,11,15} => Median = 9
e.g. {3,7,9,11,15,18} => Median = 10

(3) **Mean**: $\bar{y} = \frac{\sum y}{n}$

e.g. aphids on plants

1	2	3	4	5
1	10	10	1	1000 => mean=204.4

poor at measuring central tendency!

the arithmetic mean is highly sensitive to outliers!

Calculator:
SD mode; DATA entry with “M+”
“shift 1” gives mean

(4) The geometric mean:
Log-transform the data:

y	1	10	10	1	1000
log10 y	0	1	1	0	3

$$\sum \log_{10} y = 5; n = 5 \rightarrow \overline{\log_{10}} y = 1 \rightarrow anti\log = 10^1 = 10$$

$$\overline{y_{GM}} = \sqrt[n]{\prod y} = anti\log\left(\frac{\sum \log_{10} y}{n}\right)$$

(5) The Harmonic Mean

e.g. body size in competition experiments
Michaelis-Menten and Lineweaver-Burk!
take the reciprocal of the data and average it

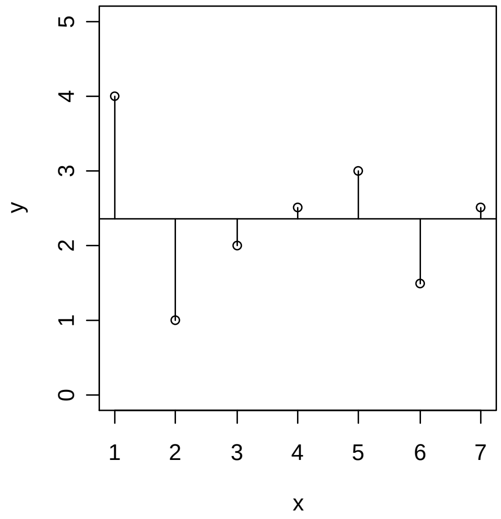
$$\bar{y} = \frac{n}{\sum \frac{1}{y}}$$

y	1	2	4	1
1/y	1	0.5	0.25	1

$$\overline{\frac{1}{y}} = 0.687 \Rightarrow \left(\overline{\frac{1}{y}}\right)^{-1} = 1.45$$

Estimation of variability

arithmetic mean, plus
differences from it



order of measurement

(1) The Range of the data: minimum, maximum

⇨ increases monotonically with sample size

⇨ not a good to measure variability

(2) Differences between y and $\bar{y}$

$d = y - \bar{y}$ individual differences

$$\sum \mathrm{d} = \sum (\mathrm{y} - \bar{\mathrm{y}}) = \sum (y - n\bar{y}) = \sum (y - \frac{n\sum y}{n}) = 0$$

$\sum d = 0$ Always! → <u>no</u> measure of variability!

$\sum |d|$ Absolute value of d → a good measure!

$\sum d^2 \succ 0$ The sum of squares

$\sum d^2$ The sum of squares
One of the most important things to remember!

Problem: SS increases ever and ever

Possible solution: Divide it by n
But: We need to know $\bar{y}$ first (i.e. estimate it from the data)

A new concept: **Degrees of freedom**

Suppose we have n=5 numbers with mean=4
They´d have to sum up to 20

So let´s fill in some possible numbers:

2	7	4	0	7

Free choice (“freedom”) until we arrive at the last number!

We therefore have n-1 degrees of freedom if we estimated the mean from a sample size of n.

D.f. = the sample size, n, minus
the number of parameters, p, estimated from the data

Our formula for the variance is therefore:

$$\boxed{\text{Variance} = \frac{\text{Sum of Squares}}{\text{Degrees of freedom}}}$$

Or, put more mathematically:

$$\boxed{s^2 = \frac{\sum(y-\bar{y})^2}{n-1}}$$

Statistics – Lecture 2

Assumptions of parametric statistics

...most important

(1) Random samples
(2) Constant variance
(3) Independent errors
(4) Normal errors
(5) Additivity of Treatment effects

...least important

(1) Random samples: if they´re not: nothing can be done to cure this!

(2) Constant variance:

> Never compare two means, when the variances are significantly different!

Use Fisher´s F test to decide (F≈smaller than 4)

Non-constant variance = Heteroscedasticity
What to do against it?

- ⇨ transform the response variable (log, square root...)
- ⇨ pick an appropriate error distribution: Generalized Linear Models

(3) Independent Errors: Errors must not be correlated (pseudoreplication); cure:

- ⇨ average it away
- ⇨ use a better model: Time series analysis, mixed effects models

⇨ Rule of thumb to spot pseudoreplication: Look at the Error degrees of freedom; these mustn´t be too large.

(4) Normal Errors:

It´s the errors that need to be normally distributed (not the data!).

The errors are often called **residuals** – after a model has been fit to the data

(5) Additive Treatment Effects: We usually assume there are no interactions between factors; this must be shown explicitly. Cure: Transform the response.

e.g. $y = a \times x \times z$	non-linear, non-additive
$\ln y = \ln a + \ln x + \ln z$	linear, additive

Never ever test for main effects before you test for interaction effects!

Describing Data: The Histogram

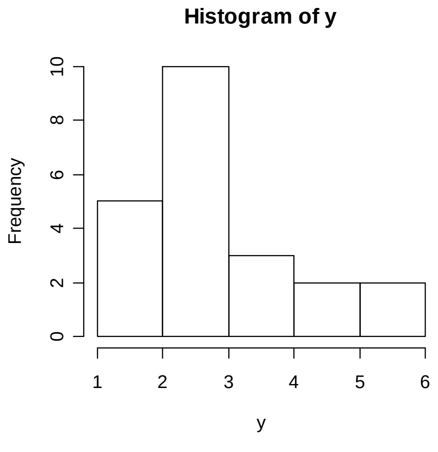
Suppose we have counted the number of feeding holes in 22 leaves; we see:

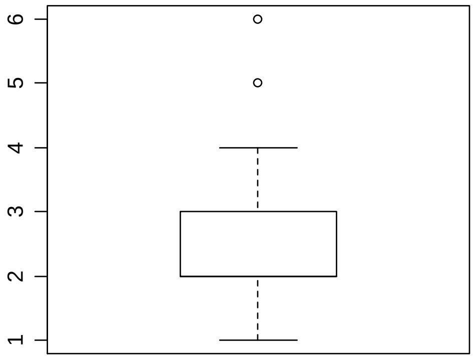
- ⇨ the sample size is 22
- ⇨ every leaf has at least one hole
- ⇨ the maximum number of holes was 6

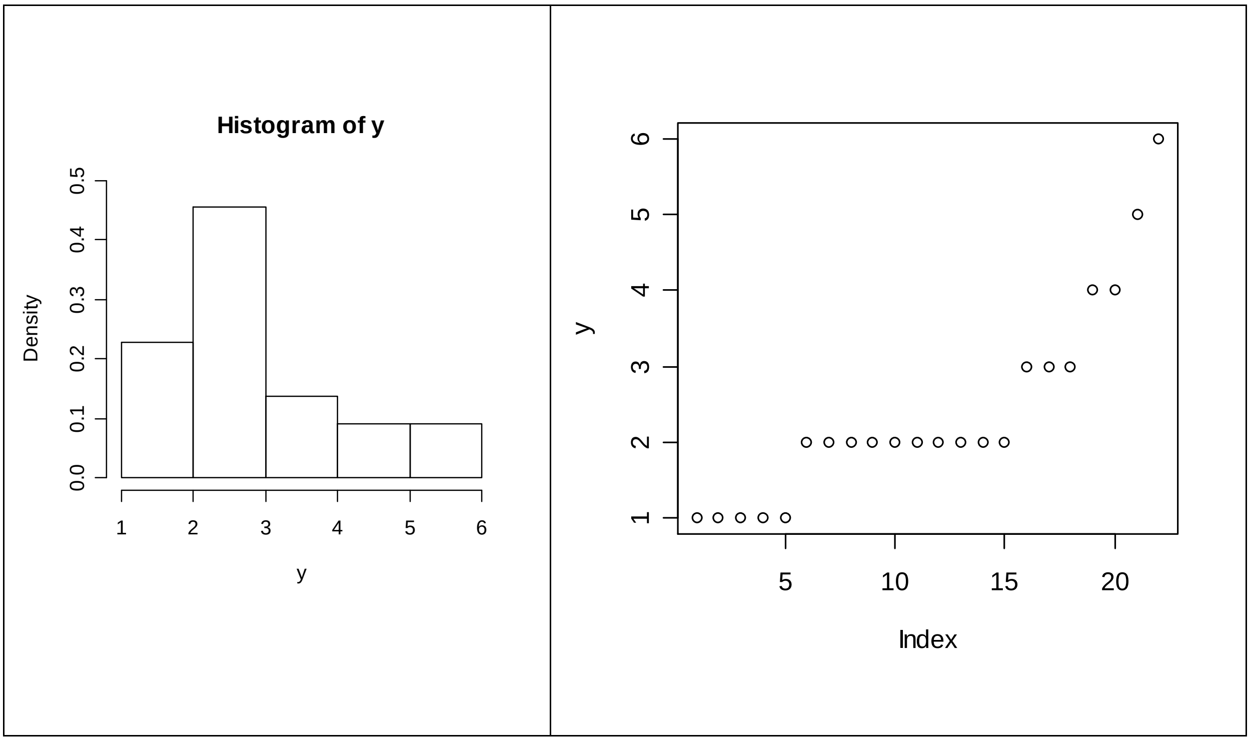
y	f(y)	y×f(y)	cum f(y)	cum p(y)	percent
1	5	5	5	5/22=0,23	0-23
2	10	20	15	15/22=0,68	23-68
3	3	9	18	18/22=0,82	68-82
4	2	8	20	20/22=0,91	82-91
5	1	5	21	21/22=0,95	91-95
6	1	6	22	22/22=1,00	95-100
		$\sum y \times f(y) = 53$			

Calculating the mean from a frequency distribution:

$$\bar{y} = \frac{\sum f_i y_i}{n} = \frac{53}{22} = 2.41$$






Histogram of y
Density
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
y
y
1
2
3
4
5
6
5
10
15
20
Index

A second example (not presented in the lecture) :

y	f(y)	cum f(y)	cum p(y)	percent
2	21	21	0,32	0-32
5	3	24	0,36	32-36
6	22	46	0,70	36-70
13	15	61	0,92	70-92
18	5	66	1,00	92-100

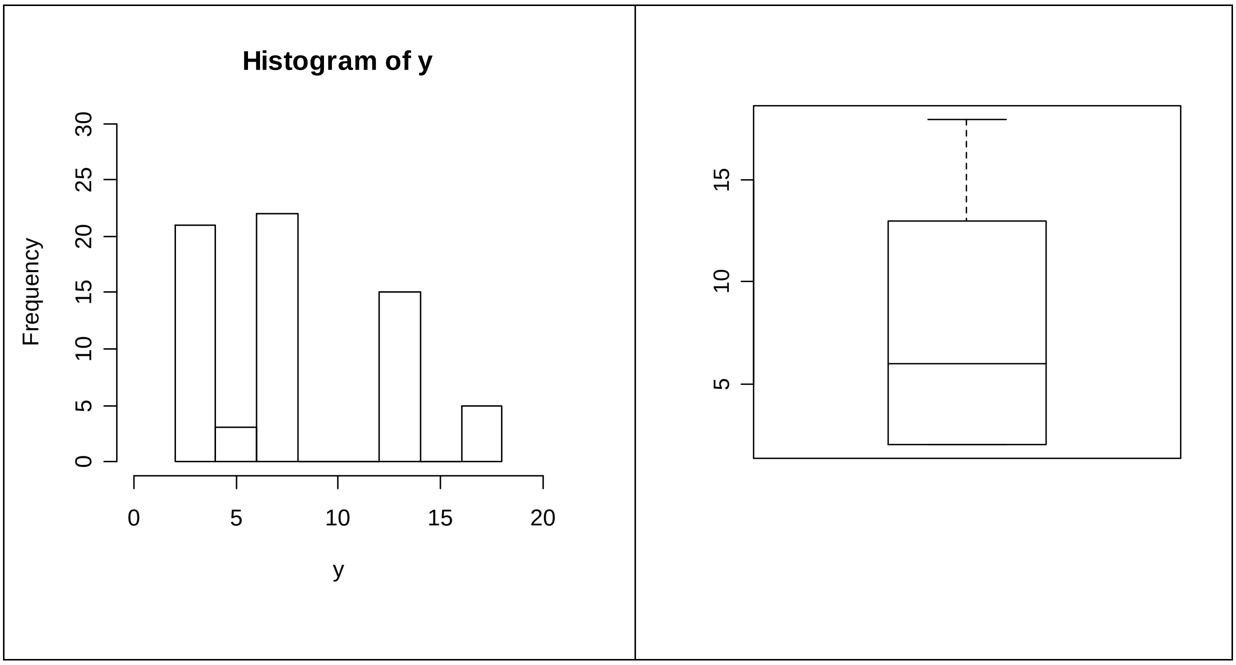
⇨ n=66

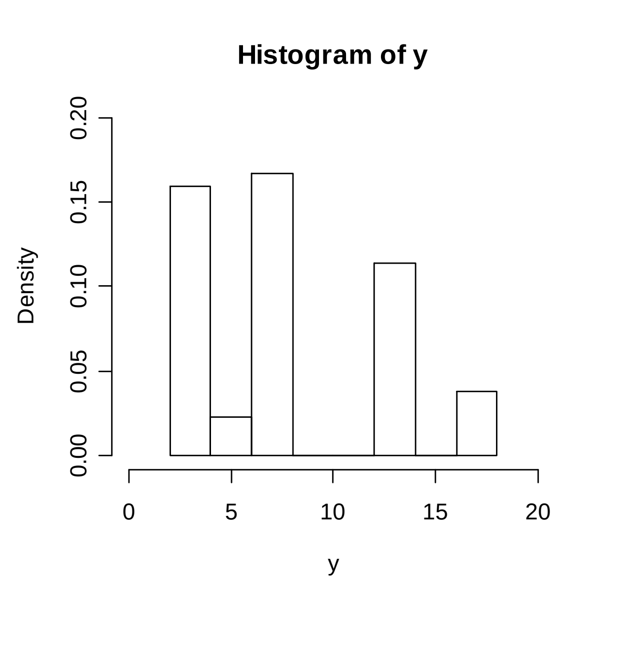
⇨ min(y)=2

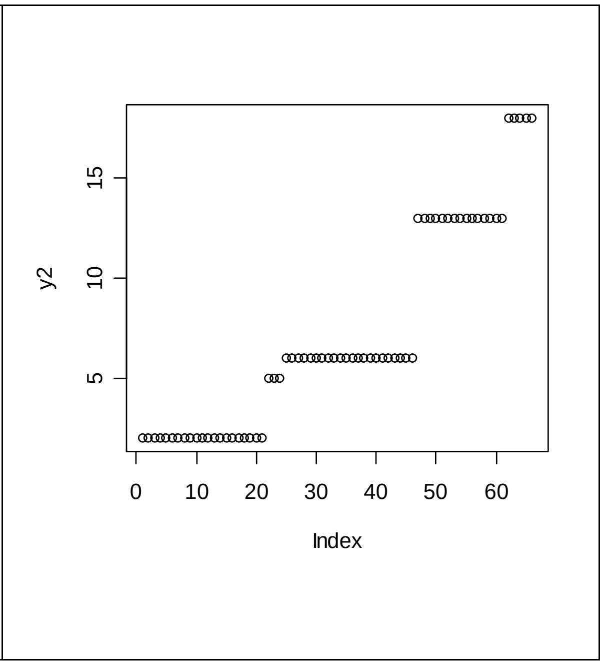
⇨ median(y)=6

⇨ mean(y)=7.18

⇨ max(y)=18







The variance:

$$\boxed{s^2 = \frac{\sum (y-\bar{y})^2}{n-1}}$$

“Sum of Squares” / “Degrees of freedom”

Let´s try an example:

- We have measured the height of five plants
- We want to know the variability in plant height

Plant	Height	Deviations, $y_i - \bar{y}$	Deviations² $(y_i - \bar{y})^2$	$\bar{y} = 230/5 = 46$
1	50	50-46= 4	16	$\sum (y-\bar{y})^2 =$ 250 $s^2 = \frac{SS}{df} = \frac{250}{4} = 62.5$
2	49	49-46= 3	9	
3	48	48-46= 2	4	
4	51	51-46= 5	25	
5	32	32-46= -14	196	
$\sum y$	230			

How can we make this calculation quicker (avoid all these subtractions)?

⇨ Short-cut formula for the sums of squares:

$$\boxed{\text{Sum of Squares} = \sum y^2 - \frac{(\sum y)^2}{n}}$$

Plant	Height	Height²	$SS=\sum y^2 - \frac{(\sum y)^2}{n}$ $SS=10830-\frac{(230)^2}{5}$ $SS=10830-\frac{52900}{5}$ SS=10830-10580=250 $s^2 = \frac{SS}{df} = \frac{250}{4} = 62.5$
1	50	2500	
2	49	2401	
3	48	2304	
4	51	2601	
5	32	1024	
	$\sum y$=230	$\sum y^2 = 10830$	

- We now know that our sample variance was 62.5
- How do we use this information?

Variance is used for
⇨ measuring unreliability
⇨ testing hypotheses

The Standard Deviation

is the square root of the variance: $s = \sqrt{s^2}$

The Standard Error

- ⇨ shall grow when variance grows (i.e. proportional to s²)
- ⇨ shall grow when sample size <u>de</u>creases (divide by n)
- ⇨ should have the same units as y (take square root)

$$SE_{\bar{y}} = \sqrt{\frac{s^2}{n}}$$

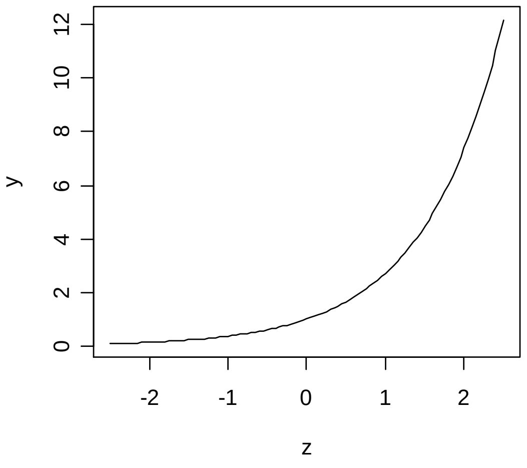
So, with our variance of 62.5, and knowing that n=5, we get;

$$SE_{\bar{y}} = \sqrt{\frac{62.5}{5}} = \sqrt{12.5} = 3.53$$

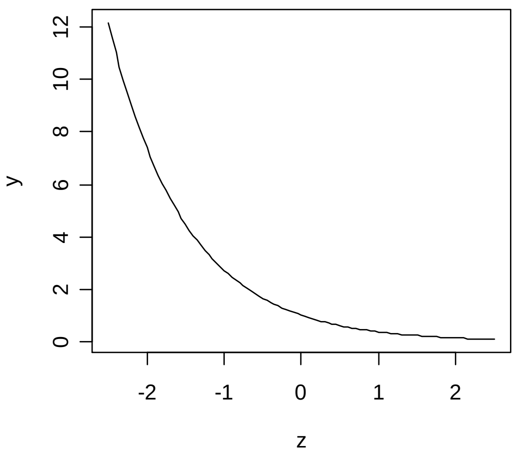
We write:
“The mean plant height was 46±3.5 (1 s.e., n=5)”

Probability Calculations: **The Normal Distribution**

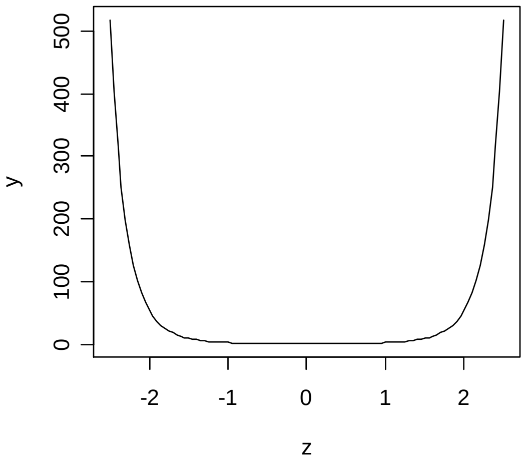
Consider a simple exponential function



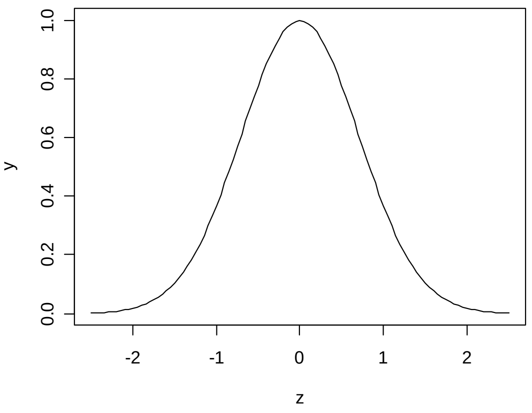
y=exp(z)



y=exp(-z)



$y=exp((-z)^2)$



$y=exp(-z^2)$

This leads us to a very important function, the Normal Distribution

- needs to be scaled, so that the area under the curve from minus to plus infinity becomes 1
- the scaling constants are the mean and the standard deviation
- The Normal Probability Density Function:

$$\boxed{p(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(z-\mu)^2}{2\sigma^2}}}$$

- ⇨ p(z) is the height on the y axis
- ⇨ z is the value on the x axis; large (small) z always means infinitely small p(z) (because $e^{-\text{large value}}$ approaches 0)
- ⇨ μ is the mean; for z=μ: e^0=1, p(z)=0.4
- ⇨ σ is the standard deviation
- ⇨ for μ=0 and σ=1, we get:

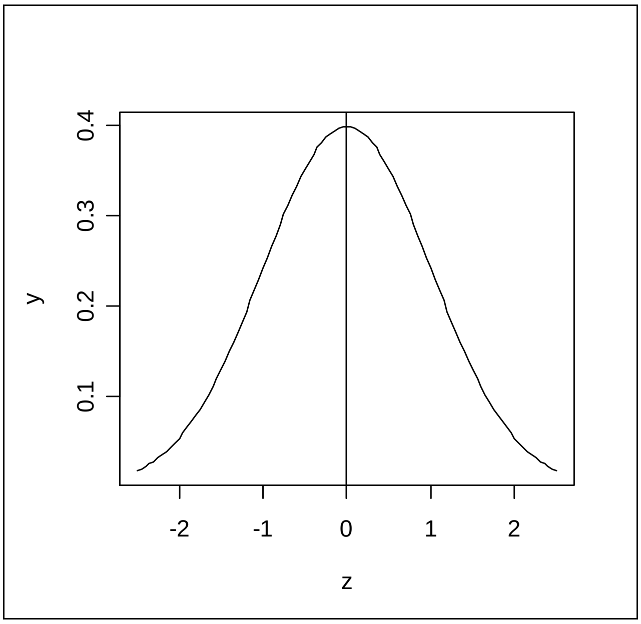
$$p(z) = \frac{1}{1\sqrt{2\pi}} e^{\frac{-(z-0)^2}{2\times1^2}}$$

Which makes our equation much simpler!

This is the **Standard Normal Distribution**

$$\boxed{p(z) \approx 0.4 \times e^{\frac{-z^2}{2}}}$$

Now let´s draw it; we know, for z=μ=0, p(z) is 0.4:



z now becomes –3; -2; -1; 0; 1; 2; 3 standard deviations from the mean

Statistics – Lecture 3

Converting a distribution to standard normal form:

Calculate z (a number of standard deviations)

$$\boxed{z = \frac{y - \overline{y}}{s}}$$

Let´s use our example from Lecture 2:

- ⇨ Mean plant height was $\overline{y} = 230/5 = 46$
- ⇨ The variance was $s^2 = \frac{SS}{df} = \frac{250}{4} = 62.5$
- ⇨ So the standard deviation = $\sqrt{s^2}$ =7.9

What is the probability for a plant being smaller than 60 cm?

$$\boxed{z = \frac{60 - 46}{7.9} = \frac{14}{7.9} = 1.77}$$

- ⇨ Look up the probability (i.e. the area under the standard normal distribution) for z=1.77. It is 0.96.
- ⇨ I.e., about 96% of our plants will be smaller than 60 cm
- ⇨ If we want to know how much will be taller, it will be 1-0.96=0.04 = 4%

The Confidence Interval

Up to now, we´ve done “less than” or “more than” tests, which means: one-sided tests.

Now, we want to do our first **two-tailed test**. We want to estimate **how certain** we can be about a mean value we have estimated from data.

Confidence Interval

Shows the **likely range** into which the mean would fall if the sampling exercise were to be repeated

- **more** confidence means the interval becomes **wider**
- e.g. we would like to be 99% confident (not 50%)
- the conficence interval is two-tailed
- to establish a 95% CI, we need to work out a special value for (100%-95%)/2=2.5%
- for n<30, we use **Student´s t** from tables to calculate the CI

CI =Student´s t from tables ×standard error

$$\boxed{CI_{95\%} = t_{(\alpha=0.025, d.f.=\gamma)}\sqrt{\frac{s^2}{n}}}$$

For large samples (n>30), we get

$$\boxed{CI_{95\%} = \pm 1.96 \times \sqrt{\frac{s^2}{n}}}$$

So, let´s take our plants example (n=5) again:

- The standard error of the mean was

$$SE_{\bar{y}} = \sqrt{\frac{62.5}{5}} = \sqrt{12.5} = 3.53$$

⇨ we have 5 plants, so there are 4 d.f.
⇨ t(0.025,4)=±2.77, thus CI=±2.77×3.53=±9.78
⇨ We could therefore write:

“The mean plant height was 46±3.5 (1 s.e., n=5)”
“The mean plant height was 46±9.78 (95% C.I.,n=5)”

Hypothesis testing

Karl Popper: “A good hypothesis is a falsifiable hypothesis.”

e.g.

“There is a rat in my kitchen”

= a bad hypothesis; not falsifiable (it could be that I always overlook the rat)

Alternative Hypothesis (H_1):
“Something is happening”

“There is **no** rat in my kitchen”

= a **good** hypothesis: as soon as we do see a rat, it will be falsified.

Null hypothesis (H_0):
“Nothing is happening”

We keep H_0 until there´s significant proof against it.

Mistakes to be made:

- ⇨ **Type I Error(α)**: Reject H_0 when there´s nothing going on; i.e. we conclude that something is different but in fact it isn´t; in biology: α should be low (two-tailed ~5%)
- ⇨ **Type II Error** (ß): Accept H_0 when in fact there´s something going on, i.e. we still believe in “nothing is happening” when in fact there is something happening. ß is 0.2 (for practical reasons)

The Power of a test:
The ability to find a significant difference when there really is one.

Power = 1-ß = 80% (→Power Analysis, see Lecture 1)

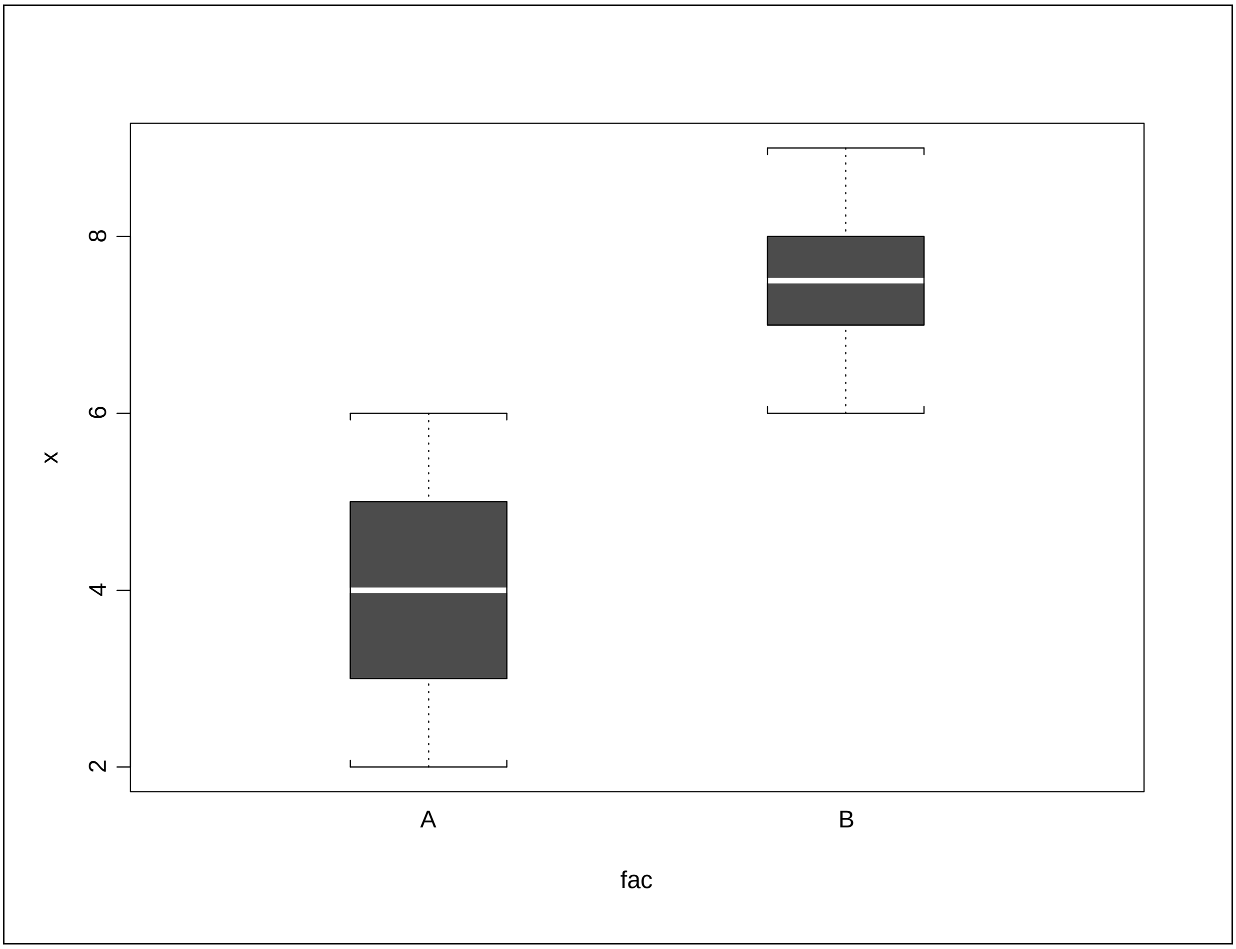
Hypothesis testing in practice

- ⇨ There are loads of tests to do (F test, t test etc.)
- ⇨ No matter what test we do, the principal is always the same:
- ⇨ Calculate a test statistic (e.g. z, t, F)
- ⇨ Look up the critical value in a table or in the computer
- ⇨ compare the calculated value with it

- ⇨ Only if test statistic $\geq$critical value: reject H_0

Student´s t test

- ⇨ When the explanatory variable is categorical (a factor with two levels)
- ⇨ and when the response variable is continuous



We use t test to find out if the two means for “A” and “B” are significantly different at α=0.05

t-Test:

“How many standard errors is our difference?”

$$t = \frac{\text{The difference between two means}}{\text{The standard error of that difference}}$$

$$\boxed{t = \frac{|\bar{y}_A - \bar{y}_B|}{\sqrt{\frac{s_A{}^2}{n_A} + \frac{s_B{}^2}{n_B}}}}$$

But this only holds as long as A and B are *not correlated*.
(If they´re correlated: **Paired** t-test)

An example:

We have measured the length of rabbit ears (cm) in male and female rabbits.

Males (A)	A^2	**Females (B)**	B^2
20	400	12	144
18	324	15	225
19	361	14	196
20	400	13	169
17	289	14	196
$\sum y$=94	$\sum y^2$=1774	$\sum y$=68	$\sum y^2$=930

A short-cut formula for the sum of squares:

$$\boxed{\text{SS}= \sum y^2 - \frac{(\sum y)^2}{n}}$$

i.e. we only need to estimate two quantities, $\sum y^2$ and $(\sum y)^2$, from the data.

$\bar{y}_A$ =94/5=18.8	$\bar{y}_B$ =68/5=13.6
$s^2_A = \dfrac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1} =$	$s^2_B = \dfrac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1} =$

$=\frac{1774-\frac{(94)^2}{5}}{4}$ $=\frac{1774-1767.2}{4}$ $=\frac{6.8}{4}=1.7$	$=\frac{930-\frac{(68)^2}{5}}{4}$ $=\frac{930-924.8}{4}$ $=\frac{5.2}{4}=1.3$
$n_A=5$	$n_B=5$
$\mathrm{SE}\overline{y_A}=\sqrt{\frac{s^2}{n}}=\sqrt{\frac{1.7}{5}}=0.58$	$\mathrm{SE}\overline{y_B}=\sqrt{\frac{s^2}{n}}=\sqrt{\frac{1.3}{5}}=0.51$

Before we start comparing our two samples, we need to find out if we are allowed to compare them:

Never compare two means, when their variances are (significantly) different!

How to compare two variances?

F-test:

$$\boxed{F=\frac{\text{larger variance}}{\text{smaller variance}}}$$

⇨ In our case: F=0.58/0.51= 1.14

⇨ Look up the critical value for F at $\gamma_{numerator}=\gamma_{deno\min ator}=4$ (6.38).

⇨ Our calculated value is smaller, so we conclude that the variances are <u>not</u> significantly different

Rule of thumb: If F>4, then the variances are significantly different.

In our case, we can go on with our analysis and calculate t:

$$t = \frac{|\overline{y}_A - \overline{y}_B|}{\sqrt{\frac{s_A{}^2}{n_A} + \frac{s_B{}^2}{n_B}}} = \frac{|18.8 - 13.6|}{\sqrt{\frac{1.7}{5} + \frac{1.3}{5}}} = \frac{5.2}{\sqrt{0.34 + 0.26}} = \frac{5.2}{\sqrt{0.6}} \approx \underline{\underline{6.75}}$$

Note that

- ⇨ our total sample size was $n_A+n_B=10$
- ⇨ we have estimated two parameters from the data
- ⇨ our total d.f.=10-2=8

Now compare this value with value from t tables at α=0.05 and γ=8: $t_{0.0.25,4}$=2.30

The value we have just calculated (6.75) is much bigger than the one from the t tables;

We conclude:

"Male and female rabbits differ significantly in their ear length ($t_{0.025,8}=2.30, n_A=n_B=5$, $p<0.05$)"

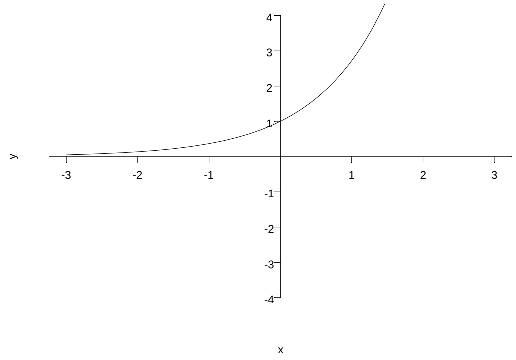
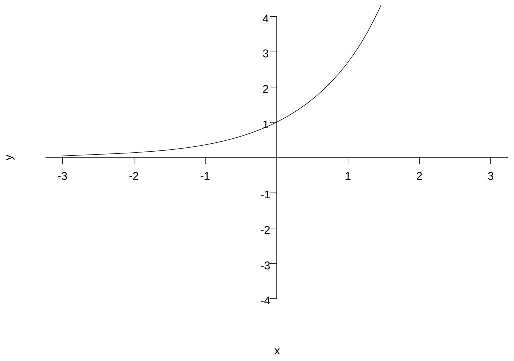
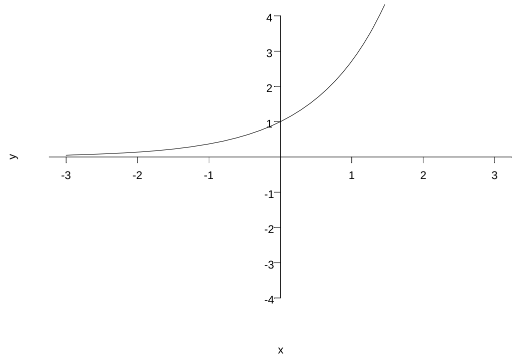
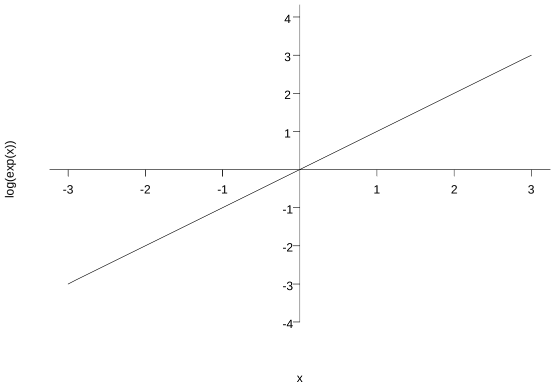
Transformation of the Response variable

Why?

- to deal with non-constant variance (heteroscedasticity)
- to deal with non-normal error distributions

How?

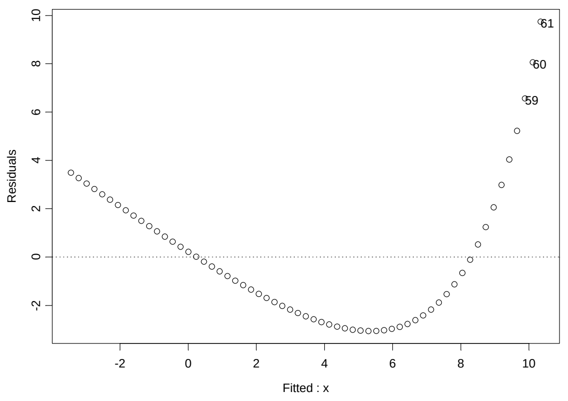
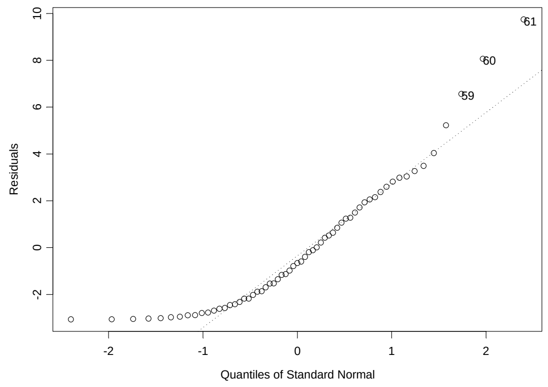
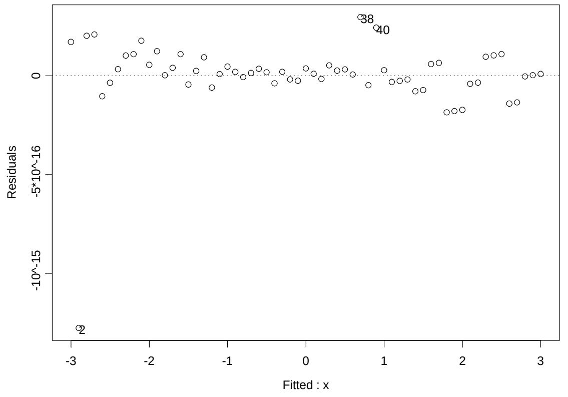
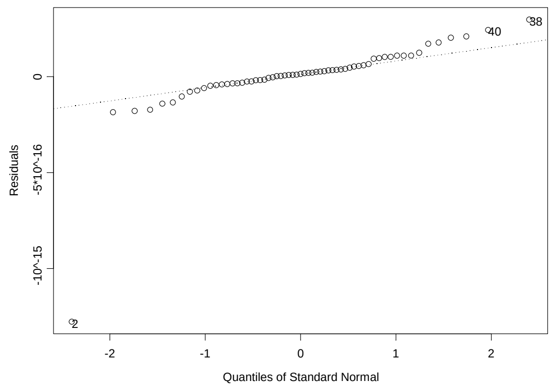
(1) Log Transformation

	
$y=e^{x}$	$y=(e^{x})'$
general model: $y=a\times e^{bx}$	general model: $y'=a\times e^{bx}$
	
$y=\int e^{x}$	y = ln a + b x here (special case): y=x

Transformation affects the error structure:

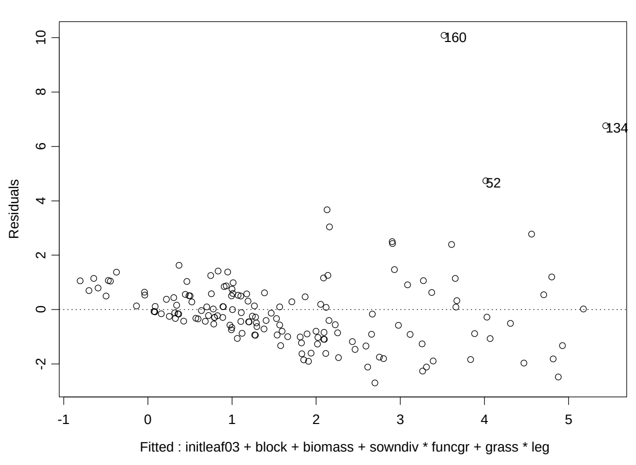
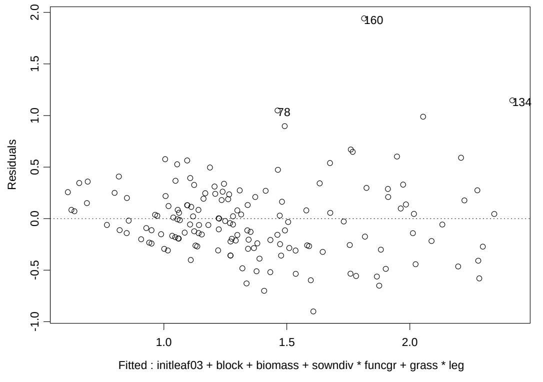
before transformation: Errors log-normally distributed

after transformation: Errors normally distributed

	
Heteroscedasticity before data have been transformed	Error distribution before data transformation
	
...and after log-transformation	...and after log-transformation

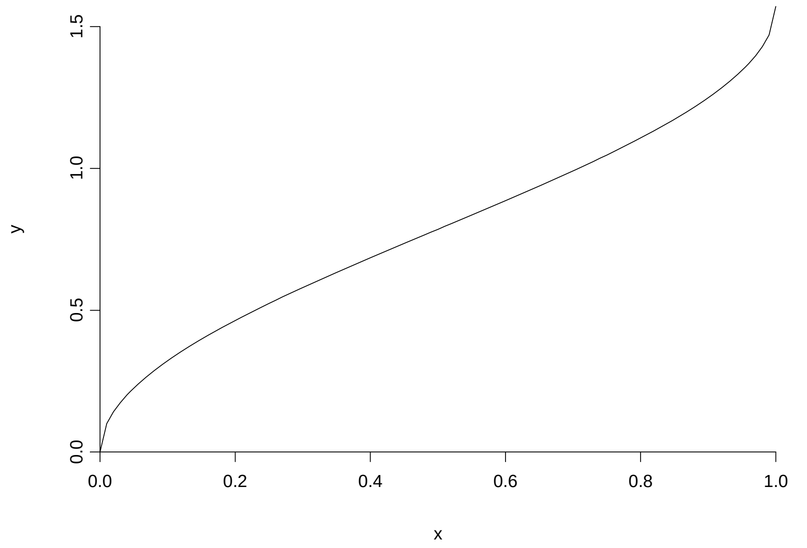
(2) The Square-root transformation

- Used with count data, where the errors follow a Poisson distribution
- We use $y' = \sqrt{y+0.5}$ or any other arbitrary constant added to y

	
Heteroscedasticity before transformation	...and after transformation

(3) The Arcsine-Square Root Transformation

- used with proportion data
- the formula is $y' = \sin^{-1}(\sqrt{proportion})$

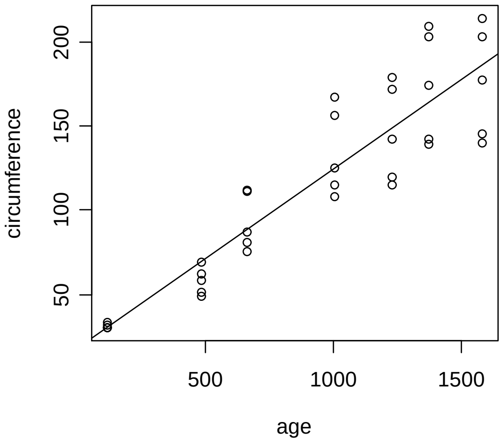


The arcsine transformation accounts for proportions that are bounded between 0 and 1.

Regression

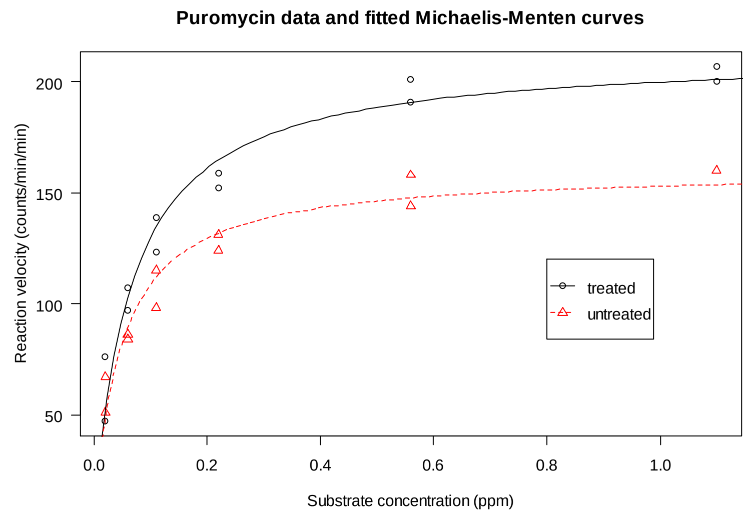
- when both response & explanatory variable are **continuous**
- graphic: **Scatterplot**

Different possible models:
(1) linear



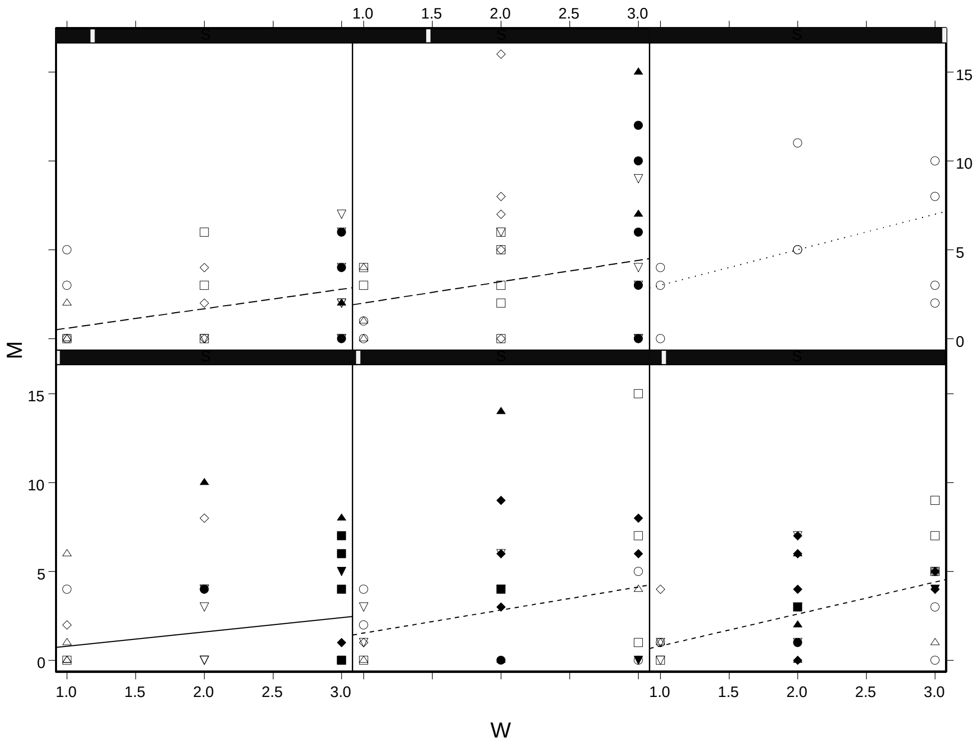
Model: y=a+b x

(2) non-linear (curved): polynomials



Model: e.g. $y = a + b\,x + c\,x^2$

(3) multiple explanatory variables



multiple linear regressions with several explanatory variables (x,z,...)

Model: $y = a + b\,x + c\,z$

Linear Regression

used for

- Describing data, e.g. $y = 2.2+0.4\ x$
- Hypothesis testing: is y really a function of x?
  H_0: There is no relationship between x and y
- Estimation, i.e. slopes and intercepts (a, b)
  unreliability of slopes and intercepts (SE_b, SE_a)
- Prediction, e.g. linear **interpolation** or **extrapolation**

Statistics – Lecture 4

Linear Regression (continued)

We are studying a relationship of the form

$$\boxed{\text{response variable = intercept + slope * explanatory variable}}$$

or, put in mathematical notaion:

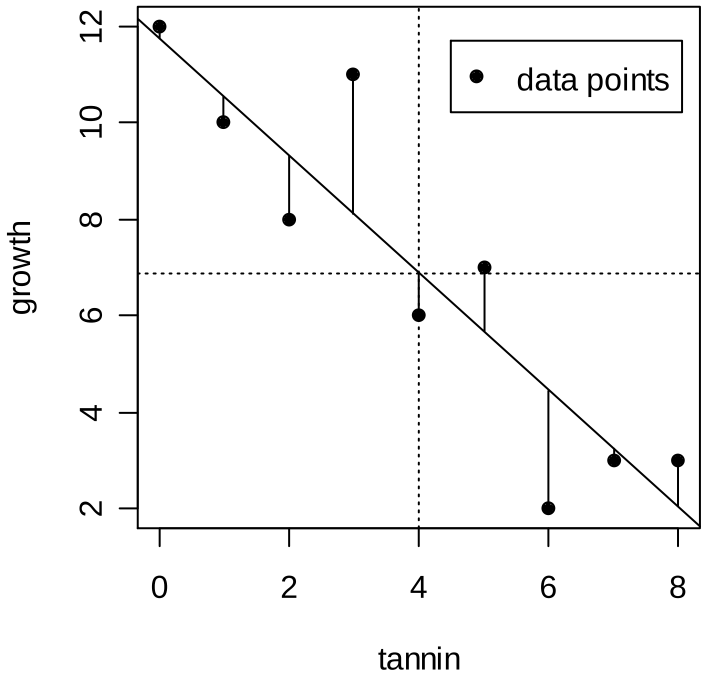
$$\boxed{y = a + b\,x}$$

A regression analysis consists of two parts:

a) estimate the parameters a, b and their standard errors
b) find out what fraction of the variation in y is explained by the model

Assumptions of Linear regression:

a) Normal errors
b) Constant variance
c) The explanatory variable is fixed and measured without error
d) All unexplained variation is confined to the response variable



This is how regression works:

- We define the best fit line as passing through $(\overline{x}, \overline{y})$

- we then rotate the line, until we find the sum of the individual departures:

$\sum d^2 = \sum (y - \hat{y})^2$ to reach a **minimum**

The minimum is found by the so-called **maximum likelihood** technique:

- given the data, and
- having selected a particular model:
- What values of the parameters
- make the data most likely?

This means: The data are fixed, and the values for each of the parameters are changed until the data are most likely.

Likelihood is the product of the probability densities for each of the values of the response variable, y.
Many likelihood functions involve the product sign:

L(a,b)=$\prod_{i=1}^{n} f(y_i|a,b)$

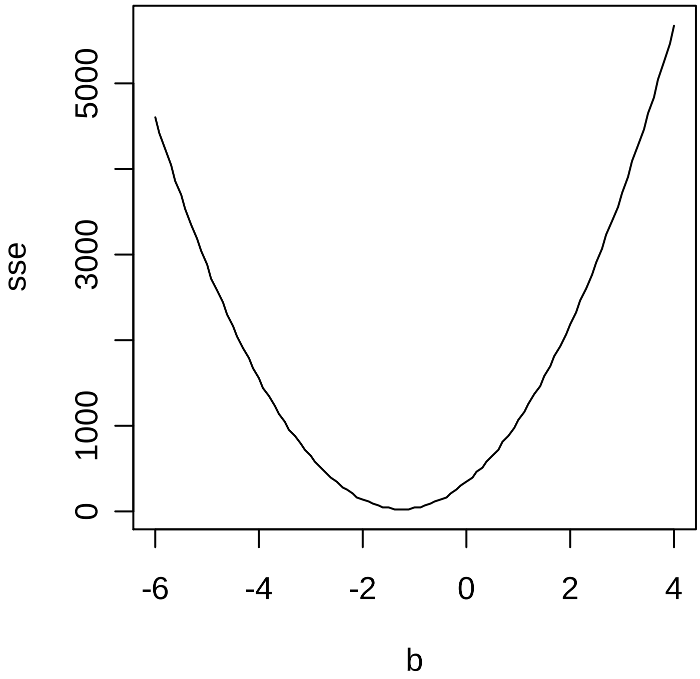
Estimation of the slope

We now want to find the **maximum likelihood** estimate of the slope b.

The maximum likelihood estimate of the slope is the **value of b** for which $\sum d^2 = \sum(y - \hat{y})^2$ reaches a **minimum.**

- ⇨ rotate the line around $(\overline{x}, \overline{y})$, until the error sums of squares is minimised.
- ⇨ The **error sums of squares** is the sum of squares of the individual departures between the data and the predicted values:

  SSE= $\sum(y - \hat{y})^2$



This means, we want to find the value of b for which SSE is minimal. We write:

SSE= $\sum(y-\hat{y})^2$

$\hat{y}=a+bx$

SSE = $\sum(y-a-bx)^2$ We want to find the minimum of SSE:

SSE = minimum $\sum(y-a-bx)^2$

So we form the first derivative:

$$\frac{dSSE}{db}=-2\sum x(y-a-bx)=-2\sum xy-ax-bx^2$$

$$\frac{dSSE}{db} = \sum xy - \sum ax - \sum bx^2 \overset{!}{=} 0$$

we know that

(1) $\sum ax = \mathrm{a}\sum x$

(2) $a = \bar{y} - b\bar{x} \rightarrow a = \frac{\sum y}{n} - b\frac{\sum x}{n}$

So we can conclude

$$\sum xy - \sum ax - \sum bx^2 = 0$$

$$\sum xy - \left(\frac{\sum y}{n} - b\frac{\sum x}{n}\right)\sum x - b\sum x^2 = 0$$

$$\sum xy - \left(\frac{\sum x \sum y}{n} - b\frac{(\sum x)^2}{n}\right) - b\sum x^2 = 0$$

now take all terms involving b on one side:

$$\sum xy - \frac{\sum x \sum y}{n} = b\frac{(\sum x)^2}{n} + b\sum x^2 \quad \text{divide by } \sum x^2 - \frac{(\sum x)^2}{n}$$

$$\frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}} = b \quad \text{which is exactly} \qquad \boxed{b = \frac{SSXY}{SSX}}$$

All statistics in regression are done withe the so-called **"famous five":**

$\sum x$

$\sum x^2$

$\sum y$

$\sum y^2$
$\sum xy$

Estimation of the intercept

Knowing that the slope is $b = \frac{SSXY}{SSX}$, we can calculate the intercept using

$$\bar{y} = a + b\bar{x} \rightarrow a = \bar{y} - b\bar{x} = \bar{y} - \frac{SSXY}{SSX}\bar{x}$$

$$\boxed{a = \bar{y} - \frac{SSXY}{SSX}\bar{x}}$$

How good is the fit of our regression line to the data?

We need to calculate

1. (1) The total sum of squares in the usual way (SST)
2. (2) The corrected sum of squares for x (SSX)
3. (3) A measure of covariation in x and y (SSXY)
4. (4) The regression sums of squares (SSR)

These are calculated as

$$SSY = \sum y^2 - \frac{(\sum y)^2}{n}$$

$$SSX = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SSXY = \sum xy - \frac{(\sum x \sum y)}{n}$$

The **regression sum of squares** is just

SSR=$\sum(\hat{y}-\bar{y})^2$; it can also be computed as $b \times SSXY$:

$$\frac{\sum xy-\frac{\sum x\sum y}{n}}{\sum x^2-\frac{(\sum x)^2}{n}}\sum xy-\frac{(\sum x\sum y)}{n}$$

To sum up, we now know:

$$\boxed{b=\frac{SSXY}{SSX}}$$

$$\boxed{a=\bar{y}-\frac{SSXY}{SSX}\bar{x}}$$

With an easy dataset, this gives:

x	y	x²	y²	xy
0 4 8 12	6 9 10 11	0 16 64 144	36 81 100 121	0 36 80 132
24	**36**	**224**	**338**	**248**

$\sum x$ $\sum y$ $\sum x^2$ $\sum y^2$ $\sum xy$; $\bar{y}=9; \bar{x}=6$

$SSY=\sum y^2-\frac{(\sum y)^2}{n}$=338-$\frac{36^2}{4}=14$

$SSX=\sum x^2-\frac{(\sum x)^2}{n}$=224-$\frac{24^2}{4}=80$

$SSXY=\sum xy-\frac{(\sum x\sum y)}{n}$=248-$\frac{24\times 36}{4}=32$

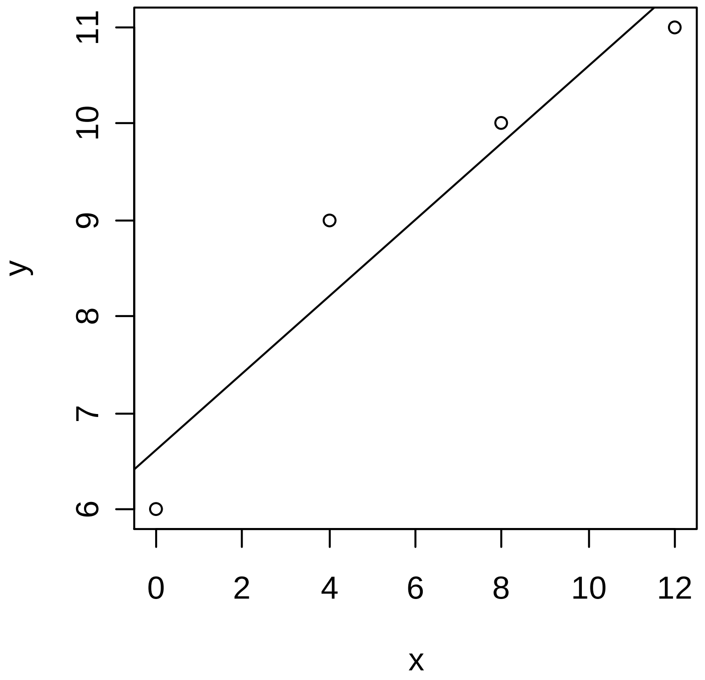
b=SSXY/SSX=
32/80=0.4

$a=\bar{y}-b\bar{x}$

a=9-0.4×6=6.6

So our regression equation would be

$$\boxed{y=6.6+0.4x}$$



It looks like all the data lie close to our regression line, i.e. there is good fit between observed and predicted values.

How can we proof this mathematically?

(1) SST= SSY = $\sum(y-\overline{y})^2$

(2) SSE = $\sum(y-\hat{y})^2=\sum(y-a-bx)^2$

(3) SSR = SST-SSE

SSR = b SSXY = 0.4×32 = 12.8

SST = SSY = 14

SSE= SST-SSR=14-12.8= 1.2

The r^2 value is SSR/SST; in our case, it is 12.8/14=0.91. 91% of our data are explained by the regression line.

ANOVA table for regression

Take SST and partition it into SSR (explained) and SSE (unexplained) variation. Compare the resulting variances using F tests:

Source	SS	df	MS (variances)	F
Regression	SSR	1	MSR= SSR/1=SSR	MSR/MSE
Error	SSE	n-2	MSE=s^2= SSE/n-2	
Total	SST	n-1		

Degree of freedom calculations:

SST=$\sum(y-\bar{y})^2$	1 parameter
SSE=$\sum(y-\hat{y})^2=\sum(y-a-bx)^2$	2 parameters (a and b)
SSR= SST-SSE	number of *extra* parameters from the null model (y=$\bar{y}$, 1 parameter) to the full model (y=a+bx; 2 parameters)

Source	SS	df	MS (variances)	F(df=1,2)
Regression	12.8	1	MSR=12.8	12.8/0.6=21.3
Error	1.2	4-2=2	s^2=0.6	

Total	14	4-1=3		

The so-called residual standard error is $\sqrt{0.6} = 0.775$

All we need now is the standard errors for slope and intercept:

$\boxed{y=6.6+0.4x}$

$$SE_b=\sqrt{\frac{s^2}{SSX}}=\sqrt{\frac{0.6}{80}}=0.086$$

$$SE_a=\sqrt{\frac{s^2\sum x^2}{n\times SSX}}=\sqrt{\frac{0.6\times 224}{4\times 80}}=0.648$$

"The slope was 0.4±0.08 (1 s.e., n=4)"
"The intercept was 6.6±0.65 (1 s.e., n=4)"

So we´re now ready to draw the observed and the predicted values:

x	y	$\hat{y}$	Residuals, y-$\hat{y}$
0	6	6.6+0.4×0=6.6	-0.6
4	9	6.6+0.4×4=8.2	0.8
8	10	6.6+0.4×8=9.8	0.2
12	11	6.6+0.4×12=11.4	-0.4

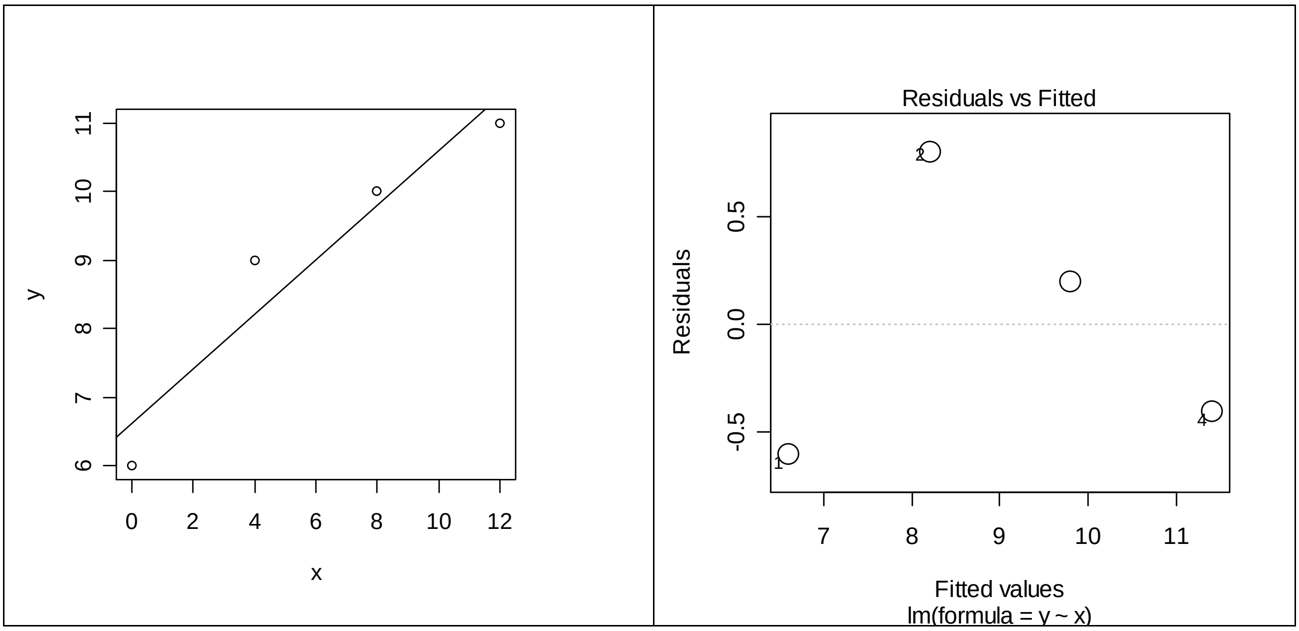
A typical call to a software program would yield the following output:

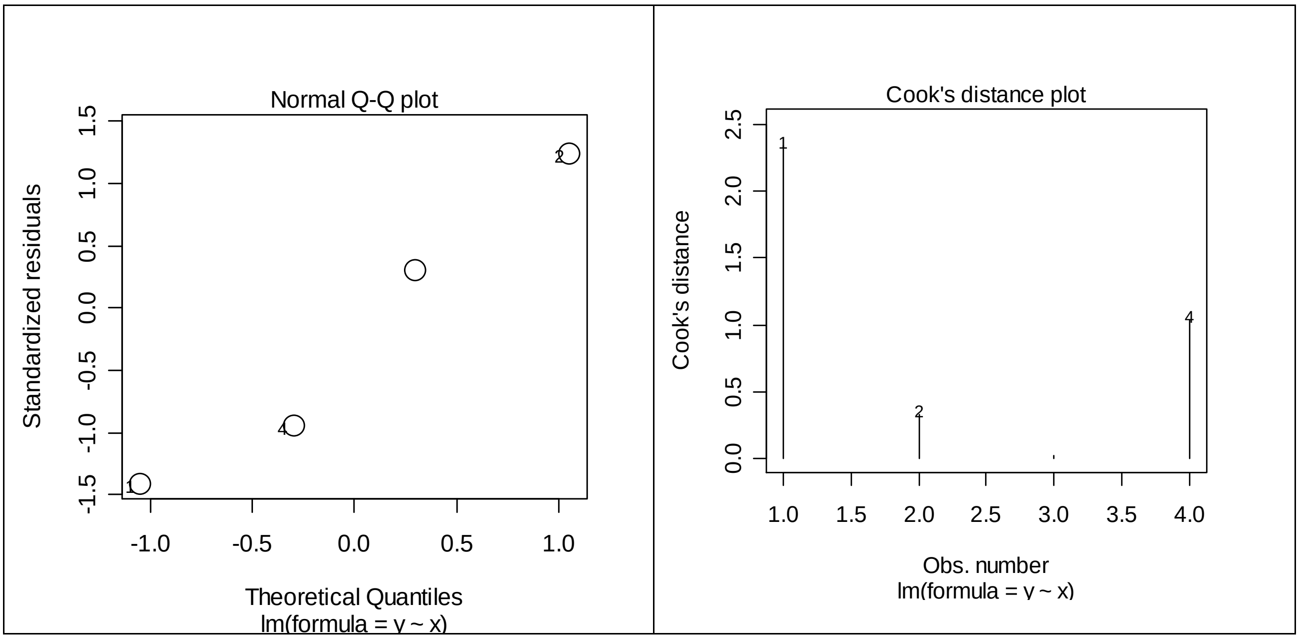
```
Call:
lm(formula = (y ~ x))

Residuals:
   1    2    3    4 
-0.6  0.8  0.2 -0.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   6.6000     0.6481  10.184   0.0095 **
x             0.4000     0.0866   4.619   0.0438 * 
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1 

Residual standard error: 0.7746 on 2 degrees of freedom
Multiple R-Squared: 0.9143,     Adjusted R-squared: 0.8714
```

Analysis of Variance Table

```
Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)  
x          1   12.8    12.8  21.333 0.04382 *
Residuals  2    1.2     0.6                  
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

The model diagnostic plots would then look like:





Now:

Analysis of Variance

- response variable: continuous
- explanatory variable: categorical!
- Null Hypothesis: All group means are the same.

Regression and ANOVA are identical approaches except for the nature of the explanatory variables.

E.g. Light intensity "low"-"medium"-"high" could also be expressed as 500 lx; 1,000 lx; 10,000 lx

Given the choice between ANOVA and regression: Always do regression. Regression and ANOVA can be combined to give analysis of covariance (ANCOVA).

Assumptions:

- random sampling
- equal variances

- independence of errors
- normal distribution of errors
- additivity of treatment effects

Model:
$y= a + bx_1 + cx_2 + ...$

Let´s stick to the simplest case: one factor with two levels
$y= a + b\ x_1 + c\ x_2$

! The factor levels enter the equation **as if they were separate explanatory variables**, x_1 and x_2

code the explanatory variables:
x_1:=1 for A and 0 for B
x_2:=0 for A and 1 for B

$\overline{y_A} = a + b \times 1 + c \times 0 = a + b$ for the first level of the factor
$\overline{y_B} = a + b \times 0 + c \times 1 = a + c$ for the second level.

a is the overall mean
b is a difference between the overall mean and $\overline{y_A}$
c is a difference between the overall mean and $\overline{y_B}$

-In Regression, "a" is the intercept and the other parameter is a slope.
-In ANOVA, "a" is the overall mean and the other parameters are differences between means

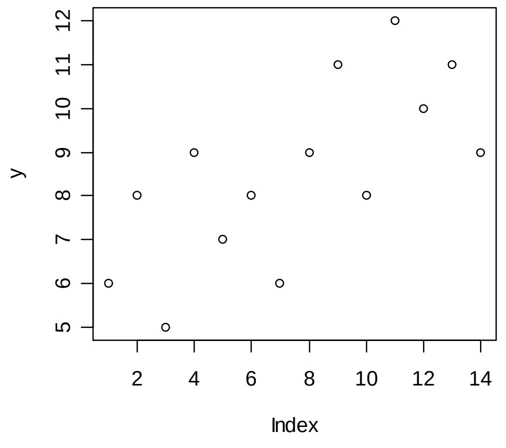
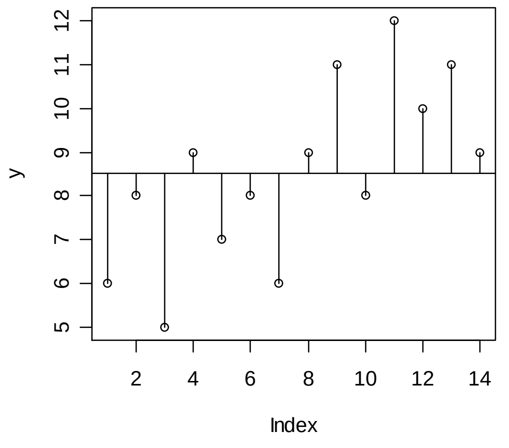
The individual observations xij can be written as

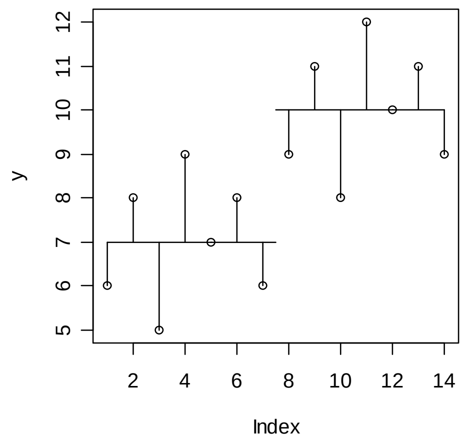
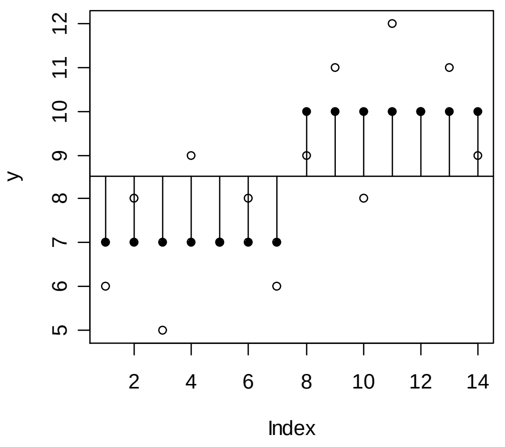
$x_{ij}=\mu + \alpha_i + \varepsilon_{ij}$ [i=1,...,a; j=1,...,n; $\varepsilon_{ij} = N(0,\sigma^2)$]

Every observed value is the sum of
(1) an overall mean μ
(2) a treatment or class deviation and
(3) a random element from a normally distributed population with mean 0 and standard deviation σ

Example:
Suppose we had n=14 observations in 2 groups (i.e. 7 observations per group).

ANOVA works like this:

 	
The individual datapoints	the overall mean, $\bar{\bar{y}}$ SST=$\sum(y-\bar{\bar{y}})^2$

	
the group means, $\overline{y_A}, \overline{y_B}$ SSE=$\sum(y_A - \overline{y_A})^2 + \sum(y_B - \overline{y_B})^2$	Difference between overall mean and group means: SSA=SST-SSE=$\sum(\hat{y} - \overline{\overline{y}})^2$

Now, a typical ANOVA table looks like this:

k is the number of factor levels
n is the number of replicates

Source	SS	df	MS	F	Critical F
Treatment	SSA	k-1	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{s^2}$	from tables (α=0.05; df= {k-1;k(n-1)})
Error	SSE	k(n-1)	$s^2 = \frac{SSE}{k(n-1)}$		
Total	SST	kn-1			

$$\text{F test} = \frac{\text{Treatments mean square}}{\text{Error mean square}} = \frac{\text{MS between classes}}{\text{MS within classes}}$$

A complex Three-way factorial ANOVA:

-There are three factors, A, B and C, each with a,b and c levels
-We are interested in interactions and main effects.

Never test for main effects before you test for interaction effects!

This is how it goes:
- calculate SSA,SSB,SSC
- calculate a so-called correction factor, CF
- Calculate the two-way interactions SSAB,SSAC,SSBC
- Calculate the three-way interaction SSABC
- Calculate SSE and SST

$\mathrm{CF}=\frac{(\sum y)^2}{abcn}$	$\mathrm{SSAB}=\frac{\sum Q^2}{n}-SSA-SSB-CF$
$\mathrm{SSA}=\frac{\sum A^2}{bcn}-CF$	$\mathrm{SSAC}=\frac{\sum Q^2}{n}-SSA-SSC-CF$
$\mathrm{SSB}=\frac{\sum B^2}{acn}-CF$	$\mathrm{SSBC}=\frac{\sum Q^2}{n}-SSB-SSC-CF$
$\mathrm{SSC}=\frac{\sum C^2}{abn}-CF$	

$$\mathrm{SSAB}=\frac{\sum T^2}{n}-SSA-SSB-SSC-SSAB-SSAC-SSBC-CF$$

SSE=SST-SSA-SSB-SSC-SSAB-SSAC-SSBC-SSABC

And the ANOVA table, then, looks like this:

Source	**SS**	**df**	**MS**	**F**
Factor A	SSA	a-1	SSA/(a-1)	MSA/s^2
Factor B	SSB	b-1	(...)	(...)
Factor C	SSC	c-1		
Interaction A:B	SSAB	(a-1)(b-1)		
Interaction A:C	SSAC	(a-1)(c-1)		
Interaction B:C	SSBC	(b-1)(c-1)		
Interaction A:B:C	SSABC	(a-1)(b-1)(c-1)		
Error	SSE	abc(n-1)	$s^2 = \frac{SSE}{abc(n-1)}$	
Total	SST	abcn-1		

Statistics, Lecture 5

Analysis of Covariance

Why do we do it?
(1) To increase precision in randomized experiments. Knowledge on initial conditions can be (and has to be!) included.

(2) To adjust for sources of bias in observational studies. E.g. include tree age as a covariate in a study on tree growth

Generally: ANCOVA is one of the most widely applicable techniques and can be extended using other modeling approaches (mixed effects models, generalized linear models etc).

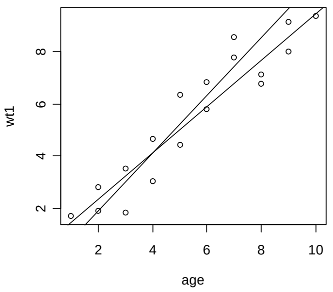
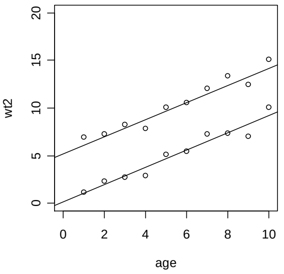
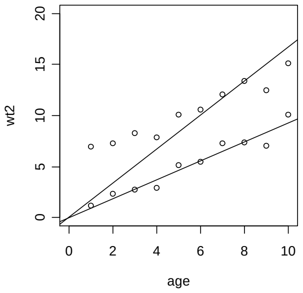
Response Variable: Continuous
Explanatory Variables: both categorical and continuous

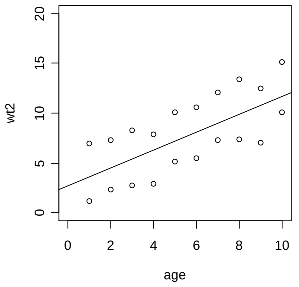
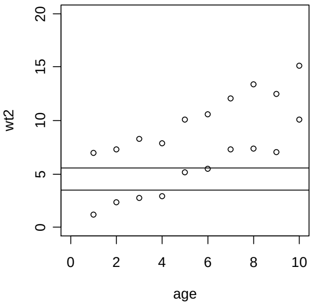
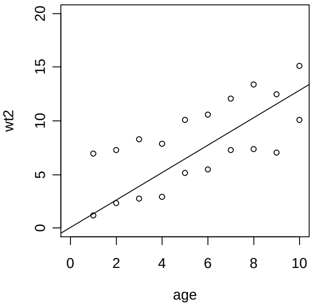
Model:
e.g. 1 categorical variable, sex
1 continuous variable: age
response variable: weight

Start with the most complex model

a) y=a+bx (males)
b) y=c+dx (females)

so in sum we have a 4-parameter model

		
Maximal model 4 parameters	3 parameters	3 parameters

		
2 parameters	2 parameters	1 parameter

Model formulae:

Regression: y=a+bx
ANOVA: y=a+bx+cz

Analysis of Covariance:

weight = a+b sex + c age + d sex:age

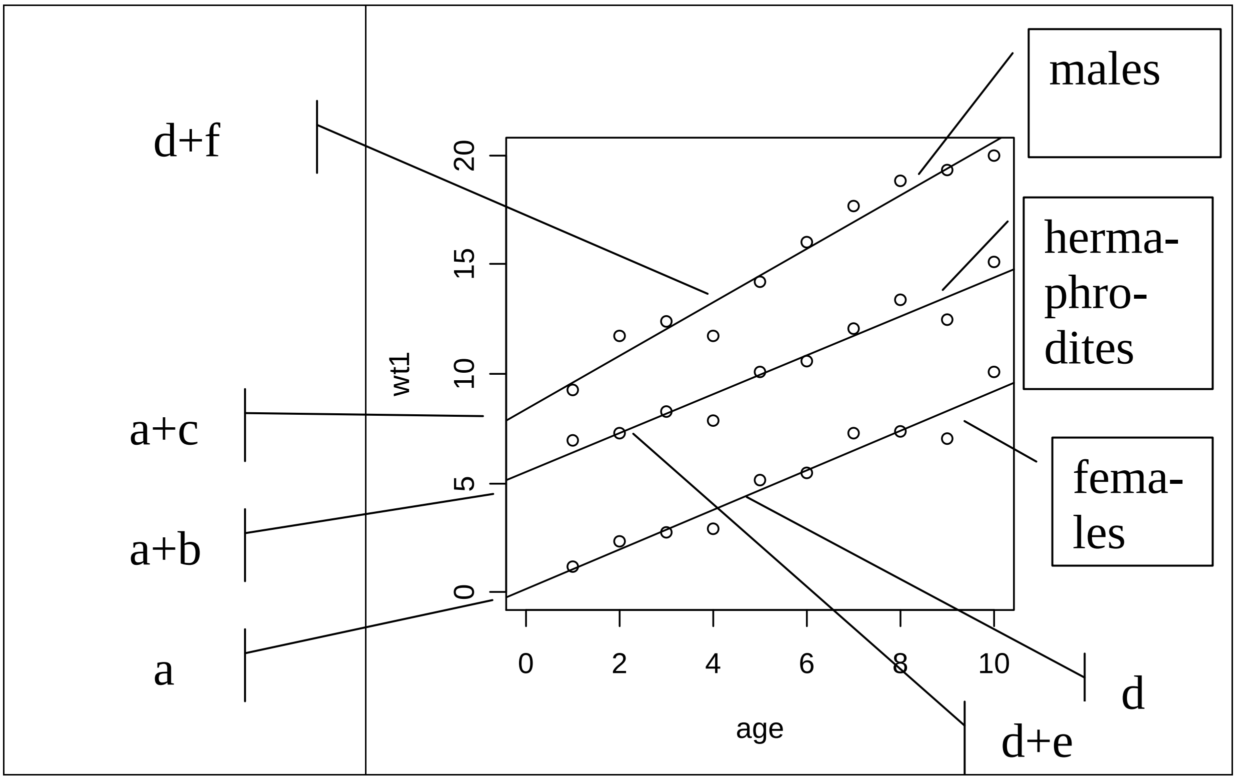
a is an intercept
b is a difference between two intercepts
c is a slope
d is a difference between slopes

sex:age is an interaction between a continuous and a categorical variable

What, if sex had three levels:
male, female, hermaphrodite?

y = a + b hermaphrodite + c male + d age +
e age:hermaphrodite + f age:male

a is an intercept
b is a difference between intercepts
c is a difference between intercepts
d is a slope
e is a difference between slopes
f is a difference between slopes



How does ANCOVA work?

split the total sums of squares, SST into

a) SSR (explained by continuous variable = a slope)
b) SSA (explained by categorical variable)
c) SSAR (Differences in slopes = an interaction term)
d) SSE the unexplained variation

A typical ANOVA table for ANCOVA:

Source

Intercept for Factor A
Differences between intercepts
overall slope for continuous variable (the regression part)
Differences between slopes

I.e. we fit a regression line for every level of the factor(s)

Contrasts

planned comparisons (a priori): part of the design
unplanned comparisons (a posteriori): after you´ve seen the analysis

Contrast coefficients:
To test hypotheses related to **levels of factors** in experiments e.g.
the ANOVA table says "Genotypes are significantly different"
we now want to know **which** Genotypes differ from which

How to do it?
(1) contrasted groups get opposite signs
(2) grouped means get the same sign

(3) the sum of the contrast coefficients, $\gamma\delta$ =0

First step: All treatments, compared with the control

Factor level	steps 1+2	step3	resulting coefficient
A	-	1	-1
B	-	1	-1
C	-	1	-1
D	+	4	4
E	-	1	-1

Further steps: Compare other subgroups

Note:
There is a large number of possible contrasts, but there are only k-1 orthogonal contrasts (where k is the number of treatments)

e.g. a factor with 5 levels a,b,c,d,e

possible contrasts could be

ab	a(b+c)	a(b+c+d)
ac	a(b+d)	a/c+d+e)
ad	a(b+e)	a(b+c+de)
ae	(...)	(...)

but there are only 4 orthogonal contrasts! I.e., those which have not been done (implicitly) already.

e.g. ab, ac: "bc" is already done implicitly.

So here comes an orthogonal contrast matrix:

	a	b	c	d	e
α	-1	-1	-1	4	-1
β	1	1	-1	0	-1
γ	0	0	1	0	-1
δ	-1	1	0	0	0

α , β and so on are the **comparisons** we want to make.

in ANOVA: How to use contrasts?

SSA is the sum of the sums of squares of the k-1 orthogonal contrasts

up to now, we´ve split up SST in SSA and SSE

now, we further split up SSA into k-1 orthogonal contrasts

Unplanned comparisons

Since there are loads of contrasts, if you do enough of them, you will find some false positives (by chance alone)

With multiple comparisons, we need to use a lower α value than usual (Bonferroni correction)

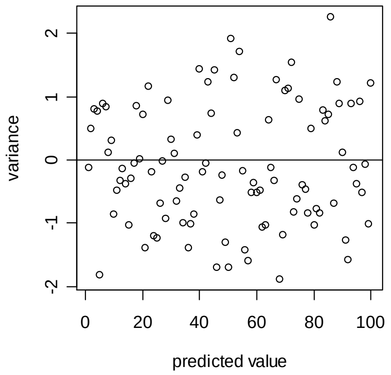
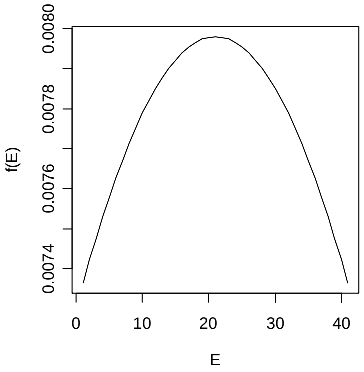
new α=old α/m (where m is the number of comparisons made)

e.g. 10 comparisons, new α is then 0.005

A new class of models:

Generalized linear models

Up to now, we had continuous response variables with

- constant variance	
- normal errors	
-additive effects	y=a+bx+cz
-independent errors	no spatial / temporal autocorrelation

New kinds of response variables:

(1) Count data.

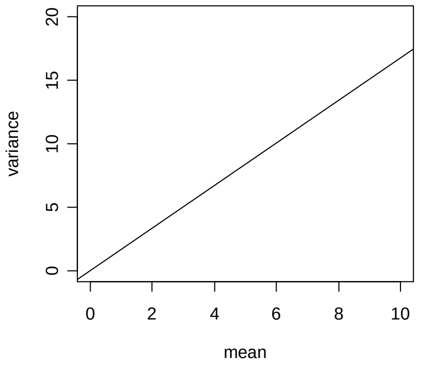
e.g. Lesions on a leaf; parasites in a host

- lots of zeros
- all values are integer

The variance will not be constant; the variability of a count increases with the mean!

A random count process is always a Poisson process.

We are studying a Poisson process, where the variance equals the mean.



If the variance: mean ratio is 1, the count data are randomly distributed.

Poisson processes show:
non-constant variance
non-normal errors
non additivity of treatment effects

(2) **Proportion Data**

a) count proportions

The number of individuals that did one thing and the number of individuals that did not do this thing

b) non-count proportions (e.g. percentage cover estimates)

bounded above and below (0-1)

E.g. "Percentage growth" or "Percentage increase"
23%, 41%, 12%

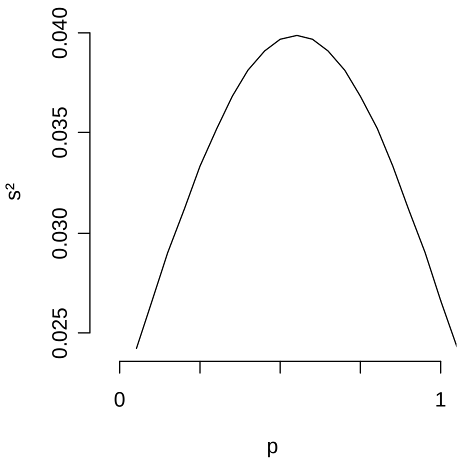
That´s not a good idea! Loss of information (2 numbers turned into one number)

Better: response variable=final mass,
initial mass = a covariate!

For real count proportions, we have:

e.g. deaths (dead or alive)
parasitology (infected or uninfected)
sex ration studies (number of males / females)

= Number of successes and failures



p=1: you always succeed, the variance is zero

p=0: you never succeed, the variance is zero

Non-constant variance

I.e. we have
- non-constant variance
- binomial errors

Generalized Linear Models

3 new things:
(1) The linear predictor
(2) The link function
(3) The Error structure

(1) Linear Predictor

so far, in

regression, we had	y=a+bx
multiple regression:	y=a+bx+cz

in general, one can say y=$\eta = \log(\frac{\mu}{1-\mu})$

$\sum x\beta$ is the **linear predictor**

> The number of rows in a linear predictor equals the number of parameters.

Generalized linear models are linear in the parameters; this means, also non-linear relationships can be modelled.

e.g. y=a+bx+cx^2

x^2:=z and hence y=a+bx-xz

(2) The Link function

η is the linear predictor: $\eta = \sum x\beta$

up to now: y = a+bx = η

new: y=f(η); f(η) is the reciprocal of the link function.

Canonical Link functions

The Link function for a Normal process is identity, y=η
The Link function for a Poisson process is log; y=exp(η)
The Link function for a Gamma process is the reciprocal
The Link function for a Binomial process is logit

logit: $\eta = \log(\frac{p}{1-p})$

(3) The Error structure

The components of the response variable have distributions that belong to the exponential distribution family

How to specify GLM´s?

- for normal data: use the Normal family with the identity link
- for count data: use the Poisson family with the log link
- for proportion data: use the Binomial family with the logit link
-Link functions and variance functions can also be used independently

e.g. we see that taking the square root makes our regression a straight line, and constant variance is achieved by taking logs

data transformation wouldn´t work

so we use a generalized linear model with
a) a square root link and
b) the variance proportional to the square of the mean

Outlook 1: Mixed effects models

random effects: things we can´t control (random variation, e.g. blocks in an experiment)

fixed effects: things we have applied (i.e. we know what we´ve done), e.g. treatments

Mixed effects models account for
- temporal autocorrelation
- spatial autocorrelation
- variance patterns (variance functions)
- correlated errors (correlation structure)

i.e. a good thing to try out!

Outlook 2: Generalized linear mixed effects models

- Still a controversial thing
- only offered in a few software programs such as SAS or R
- fit by so-called penalized quasi-likelihood or marginal quasi-likelihood
- use with caution.