Environmental Data Analysis
Part I: The Linear Model

Denis Allard[1]

Biostatistique et Processus Spatiaux (BioSP), INRA, Avignon
http://informatique-mia.inra.fr/biosp/content/homepage-denis-allard

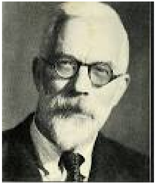
Doctoral program in Environmental Sciences
Università Ca' Foscari Venice
2016-2017

[1] With the help of Carlo Gaetan (Ca' Foscari) who allowed me to use some material from his lecture notes

Unit 0
Introduction

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

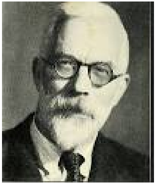
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

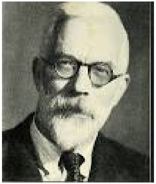
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

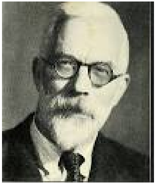
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

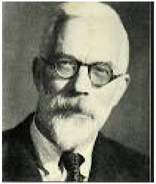
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:


R.A Fisher 1890–1962

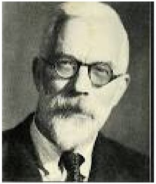

C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

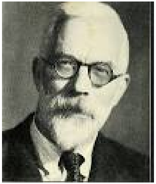
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

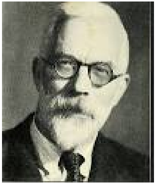
C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

History

Some scientific fields cannot go without statistics:



R.A Fisher 1890–1962

C.E. Spearman, 1863–1945

- Agronomy (field trials, genetics, seed selection, ...)
- Psychology (tests, ...)
- Medical trials
- Economics, political sciences (polls, surveys, ...)
- Environment and Geosciences

Historically, statistics was founded by non mathematicians

Statistical Triangle

What is statistics ?

- Statistics is about describing and analyzing data (samples)
- Using mathematic methods derived from probability theory
- In view of testing scientific hypothesis

Statistical Triangle

Data

Mathematics

Scientific hypothesis

Statistical Triangle

What is statistics ?

- Statistics is about describing and analyzing data (samples)
- Using mathematematic methods derived from probability theory
- In view of testing scientific hypothesis

Statistical Triangle

Data

Mathematics Scientific hypothesis

Statistical Triangle

What is statistics ?

- Statistics is about describing and analyzing data (samples)
- Using mathematic methods derived from probability theory
- In view of testing scientific hypothesis

Statistical Triangle

Data

Mathematics Scientific hypothesis

Statistical Triangle

What is statistics ?

- Statistics is about describing and analyzing data (samples)
- Using mathematic methods derived from probability theory
- In view of testing scientific hypothesis

Statistical Triangle

Data

Mathematics Scientific hypothesis

Objectives

- Introduction to statistical methods for dealing with data correlated in time and in space
- Focus on ideas and intuition
- Graphical inspection of data
- Estimating characteristics of a population, based on samples
- Quantifying causes of variations
- Testing scientific hypothesis

Objectives

- Four main chapters:
    1. Exploratory statistical analysis and inference
    2. Regression analysis
    3. Time series analysis
    4. Spatial analysis
- Not too formulas . . .
- . . . a full understanding of statistical methods requires technical details (formulas !)
- Practicals with R

Some definitions

Population

”In statistics, a population is a set of similar items or events which is of interest for some question or experiment.

A statistical population can be a group of actually existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker).

A common aim of statistical analysis is to produce information about some chosen population.”

Some definitions

Sample

A sample, $X_1, X_2, \ldots, X_n$ is a subset of a population

Random Sample

A sample is random if each individual in the sample is drawn randomly

- randomly
- independently to each other

Sampling bias

A random sample is biased when samples are collected in such a way that some members of the intended population are less likely to be included than others.

Examples:

- Internet surveys
- Survivorship bias
- Sampling in specific area or in ”interesting areas”

Some definitions

Data

Data: set of statistical variables measured on the statistical units of a population (or of a sample)

- numerical variables
  - discrete integer values
  - continuous real values
- categorical variables assume categories not numbers
  - nominal categories (or levels) without a natural order
  - ordinal categories (or levels) with a natural order

Some famous quotes and sayings

"The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."

(John Tukey)

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."

(John Tukey)

"All models are wrong, but some are useful."

"Statisticians, like artists, have the bad habit of falling in love with their models."

(Georges Box)

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

(Ronald Fisher)

Some famous quotes and sayings

And a last one, more than 100 years old...

”The great body of physical science, a great deal of the essential fact of financial science,and endless social and political problems are only accessible and only thinkable to those who have had a sound training in mathematical analysis, and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide States that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write.”

(H.G. Wells, 1911 — Mankind in the making.)

Unit 1
Exploratory analysis and R environment

Data frames

- Data are organized in tables (matrix-like format) where rows correspond to statistical units and columns to measured variables
- In `R` data matrices are called data frames
- Data frame `airquality` available in `R`: daily air quality measurements in NYC during period May-Sept 1973

```
> data(airquality)
> class(airquality)
[1] "data.frame"
> help(airquality)
```

Data frames

- `airquality` contains six variables (i.e. 6 columns) measured on 153 statistical units (i.e. 153 lines)

```
> dim(airquality)
[1] 153   6
```

- Measured variables are

```
> names(airquality)
[1] "Ozone"   "Solar.R" "Wind"    "Temp"
"Month"   "Day"
```

- First rows of `airquality`

```
> head(airquality)
Ozone Solar.R Wind Temp Month Day
1     41      190  7.4   67      5   1
2     36      118  8.0   72      5   2
3     12      149 12.6   74      5   3
4     18      313 11.5   62      5   4
5     NA       NA 14.3   56      5   5
6     28       NA 14.9   66      5   6
```

`NA` means not available and it is used to denote a missing observation

- Single variables can be accessed via operator $

```
> airquality$Ozone
[1]  41  36  12  18  NA  28  23  19 (...)
> airquality$Wind
[1]  7.4  8.0 12.6 11.5 14.3 14.9 (...)
```

- All the variables in this data frame are numeric

```
> is.numeric(airquality$Ozone)
[1] TRUE
> is.numeric(airquality$Wind)
[1] TRUE
```

- Another data frame: `CO2`

```
> data(CO2)
> help(CO2)
```

- `CO2` regards an experiment on the cold tolerance of the grass species *Echinochloa crusgalli*

```
> names(CO2)
[1] "Plant"     "Type"      "Treatment" "conc"
"uptake
> head(CO2)
Plant   Type  Treatment conc uptake
1   Qn1 Quebec nonchilled   95   16.0
2   Qn1 Quebec nonchilled  175   30.4
3   Qn1 Quebec nonchilled  250   34.8
4   Qn1 Quebec nonchilled  350   37.2
5   Qn1 Quebec nonchilled  500   35.3
6   Qn1 Quebec nonchilled  675   39.2
```

- `CO2` contains both numeric and categorical variables
- Categorical variables are also termed factors

- The first column of `CO2` is variable `Plant`

```
> CO2$Plant
[1] Qn1 Qn1 Qn1 Qn1 Qn1 Qn1 Qn1 Qn2 Qn2 ...
> is.numeric(CO2$Plant)
[1] FALSE
> is.factor(CO2$Plant)
[1] TRUE
> is.ordered(CO2$Plant)
[1] TRUE
> levels(CO2$Plant)
[1] "Qn1" "Qn2" "Qn3" "Qc1" "Qc3" "Qc2" "Mn3" "Mn2" "Mn1"
[10] "Mc2" "Mc3" "Mc1"
> nlevels(CO2$Plant)
[1] 12
```

- Variable `Plant` is an ordered factor with levels $\texttt{Qn1} < \texttt{Qn2} < \texttt{Qn3} < \ldots < \texttt{Mc1}$

- `Type` is an unordered factor (aka nominal factor)

```
> CO2$Type[c(1, 12, 45)]
[1] Quebec      Quebec      Mississippi
Levels: Quebec Mississippi
> is.numeric(CO2$Type)
[1] FALSE
> is.factor(CO2$Type)
[1] TRUE
> is.ordered(CO2$Type)
[1] FALSE
```

- `CO2` has two levels

```
> levels(CO2$Type)
[1] "Quebec"      "Mississippi"
```

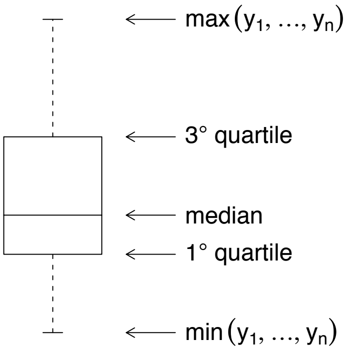
- Factors with two levels are also termed binary variables

- The data frame `CO2` includes another factor and also two numeric variables

```
> is.numeric(CO2$Treatment)
[1] FALSE
> is.factor(CO2$Treatment)
[1] TRUE
> is.ordered(CO2$Treatment)
[1] FALSE
> levels(CO2$Treatment)
[1] "nonchilled" "chilled"
> is.numeric(CO2$conc)
[1] TRUE
> is.factor(CO2$conc)
[1] FALSE
> is.numeric(CO2$uptake)
[1] TRUE
> is.factor(CO2$uptake)
[1] FALSE
```

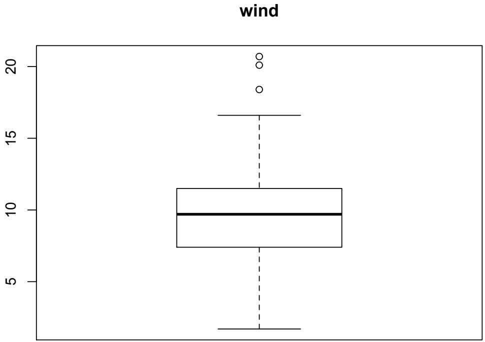
Boxplots

- First step of any statistical analysis is data visualization
- The box-and-whiskers plot is a graphical display of a numeric data vector



- In presence of outliers, the whiskers are shortened to a length of 1.5 times the box length
- Any point beyond the whiskers is a potential outlier

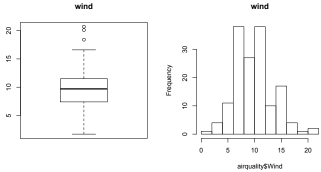
```
> boxplot(airquality$Wind, main="wind")
```



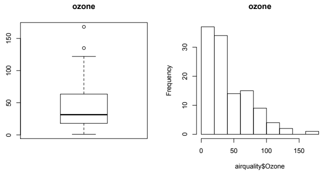
Plots

Pairing histograms and box plots can also be useful

```
> par(mfrow=c(1,2))
> boxplot(airquality$Wind, main="wind")
> hist(airquality$Wind, main="wind")
```



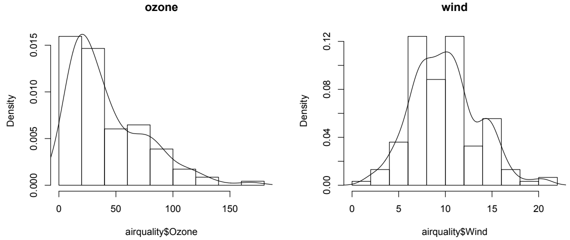

```
> boxplot(airquality$Ozone, main="ozone")
> hist(airquality$Ozone, main="ozone")
```



Smoothing the histogram

- With limited data, histograms may look rather irregular
- Irregularities may reflect:
  - sample uncertainty
  - measurement errors
- Smoothing the histogram is helpful to detect regularities obscured by sample uncertainty or measurement errors

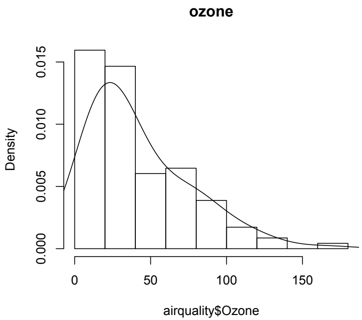
```
> par(mfrow=c(1,2))
> hist(airquality$Ozone, main="ozone", freq=FALSE)
> lines(density(airquality$Ozone, na.rm=TRUE))
> hist(airquality$Wind, main="wind", freq=FALSE)
> lines(density(airquality$Wind))
```

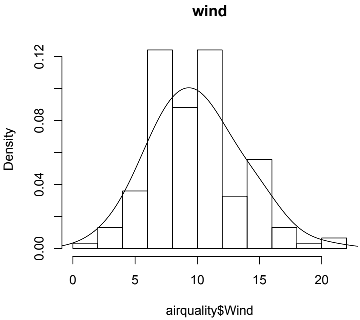


- The smoothing curve is called density

- The degree of smoothing is regulated by a parameter called bandwith
- The larger the bandwith, the smoother the density

```
> hist(airquality$Ozone, main="ozone", freq=FALSE)
> lines(density(airquality$Ozone, na.rm=TRUE, bw=13))
> hist(airquality$Wind, main="wind", freq=FALSE)
> lines(density(airquality$Wind, bw=2))
```



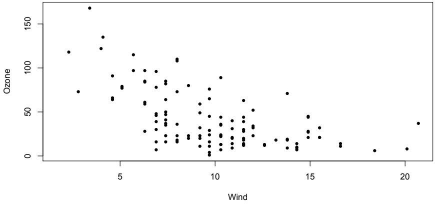


Scatterplots

- Scatterplots are used to display two numeric variables

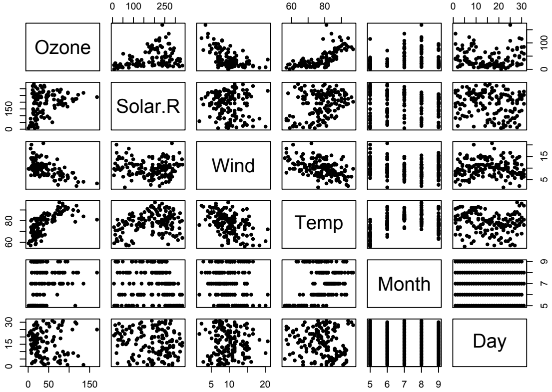
```
> plot(Ozone~Wind, data=airquality, pch=20)
```

- Notation A $\sim$ B is a formula. It means that A is explained as a function of B



Function `pairs(x)` can be used to draw the scatterplots between any pair of variables contained in the dataframe `x`

```
> pairs(airquality)
```



Correlation

- The correlation between two variables is a measure of their linear relationship
- Let $(x_i, y_i), i = 1, \ldots, n$, be n pairs of numeric variables jointly observed
- The correlation between x and y is defined as

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)}_{\text{standardized}} \underbrace{\left(\frac{y_i - \bar{y}}{s_y}\right)}_{\text{standardized}}$$

where:
- $\bar{x}$ and $\bar{y}$ are the means of x and y
- s_x and s_y are the standard deviations of x and y

- The correlation can be shown to be equal to

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} is the covariance between x and y

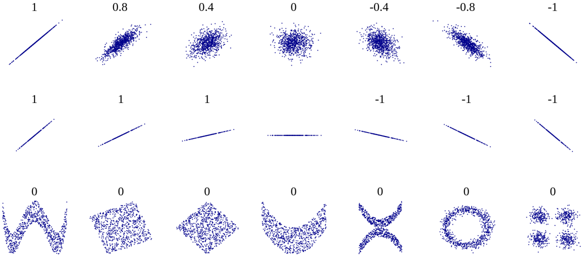
$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

- An important result known as the Cauchy-Swartz inequality states that

$$-1 \leq r_{xy} \leq 1$$

- Comments:
  - if $r_{xy} = 0$, then x and y are said to be uncorrelated which means that there is no linear relationship between them
  - $r_{xy} = 0$ does not mean there is no relationship but only no linear relationship
  - $r_{xy} = 1$ means perfect positive linear relationship between x and y
  - $r_{xy} = -1$ means perfect negative linear relationship between x and y

The various possibilities are illustrated below[2]



[2] This picture comes from the Wikipedia page
http://en.wikipedia.org/wiki/Correlation_and_dependence

- Functions `cor()` and `cov()` are used to compute r_{xy} and s_{xy}

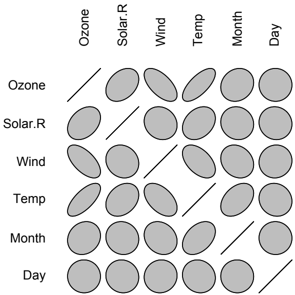
```
> cor(airquality$Ozone, airquality$Wind,
+ use="complete.obs")
[1] -0.6015465
```

  where option `use="complete.obs"` means that pairs with one or both missing observations are removed
- **R** computes correlations and covariances dividing by $n-1$ instead of n (in way to correct for the so-called "small-sample bias")
- If function `cor()` is applied to an entire data frame, then the correlation matrix containing all pairwise correlation is returned

```
> round( cor(airquality, use="complete.obs"), 2)
         Ozone Solar.R  Wind  Temp Month   Day
Ozone     1.00    0.35 -0.61  0.70  0.14 -0.01
Solar.R   0.35    1.00 -0.13  0.29 -0.07 -0.06
Wind     -0.61   -0.13  1.00 -0.50 -0.19  0.05
Temp      0.70    0.29 -0.50  1.00  0.40 -0.10
Month     0.14   -0.07 -0.19  0.40  1.00 -0.01
Day      -0.01   -0.06  0.05 -0.10 -0.01  1.00
```

- Function `plotcorr()` in the optional package **ellipse** (Murdoch and Chow, 2007) provides a nice representation of a correlation matrix

```
> install.packages("ellipse")
> library(ellipse)
> plotcorr( cor(airquality, use="complete.obs") )
```



- See `example(plotcorr)` for coloured examples

Random samples – uncertainty

- In many practical contexts it is not possible to observe an entire population
- A random sample is chosen from the population by a selection procedure with an unpredictable component
- If samples are random, then inferential statistical methods based on the theory of probability can be used to infer about the unobserved population the samples come from
- Conclusions based on non-random sampling are usually unreliable – or need more sophisticated methods that account for the specific selection scheme

Sampling in **R**

- Function `sample()` takes a random sample from a given population

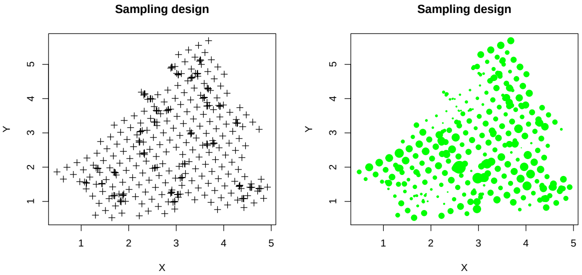
```
> data <- 1:100
> sample(data, size=5)
[1] 21 24 45 33 39
> sample(data, size=5)
[1] 44 95 22 36 85
> sample(data, size=5)
[1] 53 58 26 85 31
```

- Computer are deterministic machines and thus they cannot produce random samples
- In fact, `sample()` draws a pseudorandom sequence that looks like a random sequence, although it is not
- "Looks like a random sequence" means that the sequence is tested to check whether it is possible to predict its future values

- Pseudorandom sequences start from a certain seed
- If we choose the same seed, then we have the same sequence.
- In **R** the seed of the pseudo random sequence is set by command `set.seed(x)` where `x` is a number

```
> set.seed(543)
> sample(data, size=5)
[1] 92 81 57 11 65
> sample(data, size=7)
[1] 89 43 21 17 29 74 42
>
>
> set.seed(543)
> sample(data, size=7)
[1] 92 81 57 11 65 84 40
```

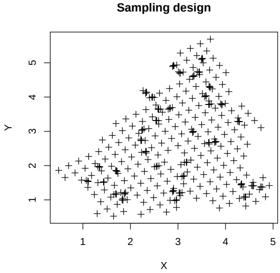
The Swiss Jura data set

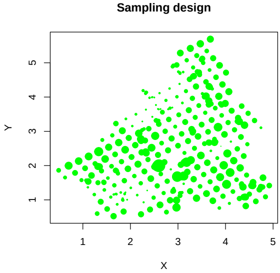


- 359 data, sampled in Swiss Jura, approx. 25 km^2 study area
- Sampling design: regular grid + random local densification
- Content of 7 heavy metals: Cd, Co, Cr, Cu, Ni, Pb, Zn
- 5 rock types: Argovian,Kimmeridgian, Sequenian, Portlandian, Quaternary
- 4 Land Use: forests, pastures, grasslands, tillage

The Swiss Jura data set

```
> jura = read.table("jura.txt",header=TRUE)
> jura[1,]
x       y lu rt    Cd    Co     Cr     Cu     Ni     Pb     Zn
1 2.386 3.077  3  3 1.74 9.32 38.32 25.72 21.32 77.36 92.56
> par(mfrow=c(1,2))
> plot(jura$x,jura$y,main="Sampling design",xlab="X",ylab="Y",pch=3)
> plot(jura$x,jura$y,main="Sampling design",xlab="X",ylab="Y",
> pch=19,cex=jura$Ni/20,col="green")
```





Sampling Variability of the mean

- Consider the average level of `Ni`

```
> true <- mean(jura$Ni, na.rm=TRUE)
> true
[1] 20.01822
```

- Suppose that, for some reason, we cannot observe all the data but only a random sample of size 30
- We want to use this sample to estimate the true average level

```
> mean(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 20.08267
> mean(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 18.648
> mean(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 19.72133
> mean(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 19.968
> mean(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 20.46
```

- Estimates based on random samples fluctuate around the true value

- To get further insight into random sampling, we can repeat the sampling process a large number of times, say 100

```
> all.samp <- replicate(100, sample(jura$Ni, size=30))
```

- Object `all.sim` is a matrix with 30 rows and 1,000 columns

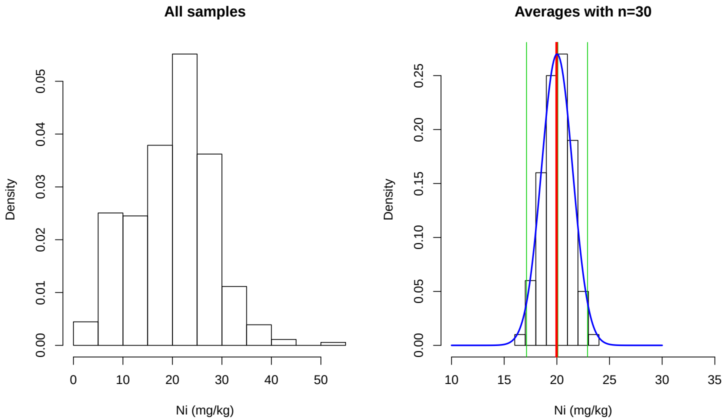
```
> dim(all.sim)
[1]  30 100
```

- Now, we compute the 100 estimates corresponding to the 100 random samples by function `apply()` which allows to apply a function to matrix rows or columns
- The syntax of `apply` is

```
apply(x, margin, function)
```

where `x` is a matrix, `margin` is 1 if `function` has to be applied to the rows and 2 if it has to be applied to the columns
- In our case, we have

```
> xbar <- apply(all.samp, 2, mean, na.rm=TRUE)
> xbar[1:20]
[1]  19.45333 19.47933 19.57600 19.80800 20.30267 21.02800 23.06133
[8]  20.72600 20.04800 20.35600 17.35067 18.05933 20.52933 21.20400
[16] 19.17733 18.92267 17.72533 21.28400 20.86800 19.69600
```



The distribution of the estimates is centered around the true value and looks symmetric

```
> summary(est)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
15.69   19.12   20.03   20.04   20.98   24.22
> true
[1] 20.01822
```

Histogram of the 100 estimates with a vertical red line corresponding to the true value of the Ni level

```
> par(mfrow=c(1,2))
> hist(jura$Ni,xlab="Ni (mg/kg)",main="All samples",
>      xlim=c(0,max(Ni)),probability=TRUE)
> hist(xbar,xlab="Ni (mg/kg)",main="Averages with n=30",
>      xlim=c(10,35),probability=TRUE)
> abline(v=mean(Ni),col=3,lwd=3)
> abline(v=mean(xbar),col=2,lwd=3)
> abline(v=mean(Ni)-1.96*sqrt(var(Ni)/ndata),col=3,lwd=1)
> abline(v=mean(Ni)+1.96*sqrt(var(Ni)/ndata),col=3,lwd=1)
> z = seq(10,30,by=0.1)
> f = dnorm(z,mean=mean(Ni),sd=sqrt(var(Ni)/ndata))
> lines(z,f,type="l",col="blue",lwd=2)
```

- If the sample size is larger, then the sample distribution of the estimates is more concentrated around the true value

```
> all.samp2 <- replicate(100, sample(jura$Ni, size=60))
> xbar2 <- apply(all.samp2, 2, mean, na.rm=TRUE)
```

- In fact, the 90% of the estimates with samples size 30 lies in

```
> quantile(xbar, probs=c(0.05, 0.50, 0.95))
      5%      50%      95%
17.69567 19.79067 22.10180
```

- The same interval with samples of size 60 is shorter

```
> quantile(xbar2, probs=c(0.05, 0.50, 0.95))
      5%      50%      95%
18.62182 20.06967 21.27593
```

Can we quantify this estimation uncertainty ?

We will need probability theory and Gaussian random variables

Unit 3
Estimation and Tests

Inference: definitions and basic principles

- A parameter is a quantity describing a theoretical probability distribution (for the population)
- A statistic is a quantity computed from the sample, in order to estimate the parameter
- An estimate of the parameter is derived from these statistics.

- The sampling error is the chance difference between an estimate and the population parameter being estimated
- The bias is a systematic discrepancy between estimates and the true population characteristic
- The standard error of an estimate is the standard deviation of the estimate' s sampling distribution
- The sampling distribution of a statistic is the probability distribution of values for an estimate that we might obtain when we sample a population.

Random Variables

- Before to proceed with the description of the normal distribution, we need to introduce random variables
- Informally, a random variable, usually denoted X, can be defined as *the future outcome of a measurement, before the measurement is taken*
- *A random variable does not have a specific value, but rather a collection of potential values with a distribution over these values* (Yakir, 2011)[3]
- Random variables can be either categorical or numerical, the latter further subdivided into discrete and continuous
- The normal variable (or Gaussian variable) is the most important example of continuous random variable

[3] Yakir, B. (2011). *Introduction to Statistical Thinking (With R, Without Calculus)*, available at url `http://pluto.huji.ac.il/~msby/StatThink/index.html`

Gaussian Random Variables

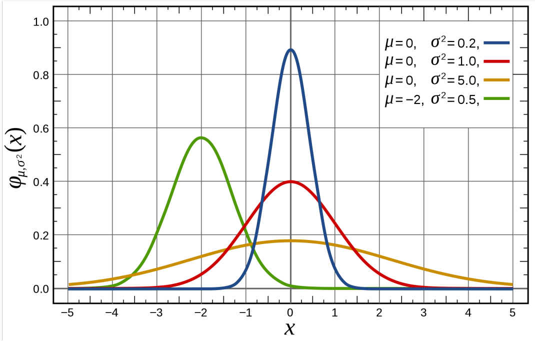
- We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to indicate that X is normal (or Gaussian) with mean μ and variance σ^2

- In order to specify a continuous random variable, we need to:
  - describe the range of possible outcomes (the support)
  - describe the probability of observing outcomes in a certain interval
- In the case of the normal variable we have:
  - the domain is the real line (from $-\infty$ to $+\infty$)
  - the probability of observing outcomes in a certain interval is described by the area under the normal density (Gaussian bell curve)

- Normal density function

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Characteristics:
  - symmetry, median and mode coincide with mean μ
  - variance σ^2 describes the spread around μ



Source: Wikipedia `http://en.wikipedia.org/wiki/Normal_distribution`

- Normal variables are identified by parameters μ and σ^2

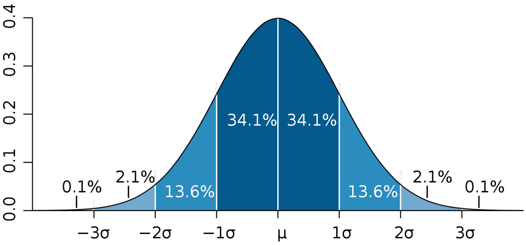
- Linear combinations of normal variables are still normal:

$$\text{if } X \sim \mathcal{N}(\mu, \sigma^2), \text{ then } Y = a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$$

- An important example of linear combination is standardization

$$Z = \frac{X - \mu}{\sigma} = \underbrace{-\frac{\mu}{\sigma} + \frac{1}{\sigma}X}_{a+bX} \sim \mathcal{N}(0, 1)$$

- Z is called the standard normal variable



Source: Wikipedia http://en.wikipedia.org/wiki/Normal_distribution

- If the “true” data generator mechanism follows a normal law, then about 95% of observations lies in $(\mu - 2\sigma, \mu + 2\sigma)$

- In R a normal random sample of size n is obtained by calling `rnorm(n, mean = 0, sd = 1)`
- Warning! **R** functions for the normal distribution use σ (`sd`) not σ^2!
- Next lines simulate an increasing number of observations from $X \sim \mathcal{N}(2, 1)$ and compute the frequency of observations smaller than 1

```
> set.seed(123)
> mean( rnorm(100, 2, 1) < 1 )
[1] 0.14
> mean( rnorm(1000, 2, 1) < 1 )
[1] 0.167
> mean( rnorm(10000, 2, 1) < 1)
[1] 0.159
```

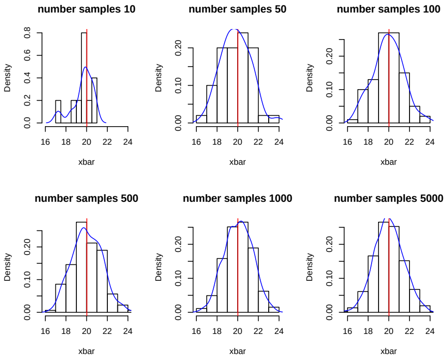
- Do these values make sense?
- Compare with $P(X \leq 1) = 0.1586553$

```
> pnorm(1,2,1)
[1] 0.1586553
```

The Normal Distribution

Increasing number of random samples of size 30

```
> for(i in c(10, 50, 100, 500, 1000, 5000)){
  all.sim <- replicate(i, sample(jura$Ni, size=30))
  xbar    <- apply(all.sim, 2, mean, na.rm=TRUE)
  hist(xbar, freq=FALSE, main=paste("number samples", i),xlim=c(16,24))
  abline(v=mean(jura$Ni), col="red")
  lines(density(xbar), col="blue")
  }
```



Estimating the mean

- Let $X_1, \dots, X_n$ be an i.i.d. n-sample, arising from a population with mean μ and variance σ^2.
- The arithmetic average

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is the natural estimator of μ.
- It is a random variable with

$$E[\bar{X}] = \mu; \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Central Limit Theorem

Provided $\sigma^2 < \infty$, as $n \to \infty$, we have

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1); \qquad \bar{X} \to \mu + \frac{\sigma}{\sqrt{n}}\mathcal{N}(0, 1)$$

Estimating the mean

Summary

- The estimate of mean, $\bar{X}$ is random, because the sampling is random
- It is unbiased: $E[\bar{X}] = \mu$
- $\mathrm{Var}[\bar{X}] = \sigma^2/n$
- The esimate of the mean is within

$$[\mu - 1.96\sigma/\sqrt{n}; \mu + 1.96\sigma/\sqrt{n}]$$

with probability 95%.

Confidence Interval for the mean

1. $X_1, \ldots, X_n$ i.i.d samples with $E[X] = \mu$
2. Let us find an interval $[\hat{\mu}_{inf}, \hat{\mu}_{sup}]$ containing the true value $\mu = E[X]$ with probability $1 - \alpha$: we call it the level.
3. We set the error on both sides

$$\mathbb{P}(\mu < \hat{\mu}_{inf}) = \mathbb{P}(\mu \geq \hat{\mu}_{sup}) = \alpha/2.$$

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then, as $n \to \infty$.

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}],$$

where u_p is the value such that

$$\mathbb{P}(\mathcal{N}(0,1) \leq p) = u_p$$

Confidence Interval for the mean

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then,

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Proof

Using CTL,

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(-u_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq u_{1-\alpha/2}\right) \\ &= \mathbb{P}\left(-u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} \leq -\mu \leq u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X}\right) \\ &= \mathbb{P}\left(\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Confidence Interval for the mean

Confidence Interval: σ^2 is known

Let us first assume that σ^2 is known. Then,

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

The width of the CI interval

- Increases with $1 - \alpha$:

$$\begin{aligned} \alpha = 10\% \quad &\Leftrightarrow \quad u_{0.950} = 1.64 \\ \alpha = 5\% \quad &\Leftrightarrow \quad u_{0.975} = 1.96 \\ \alpha = 1\% \quad &\Leftrightarrow \quad u_{0.995} = 2.58 \end{aligned}$$

- Increases with the variance
- Decreases as $1/\sqrt{n}$
- 1000 repetitions – samples of size 30. With $\alpha = 5\%$, one finds

$$\#\{\mu < \hat{\mu}_{inf}\} = 24; \quad \#\{\mu > \hat{\mu}_{sup}\} = 19,$$

where 25 expected.

Sampling Variability of the variance

- Consider the average level of `Ni`

```
> true.var<- var(jura$Ni, na.rm=TRUE)
> true.var
[1] 65.51511
```

- Suppose that, for some reason, we cannot observe all the data but only a random sample of size 30
- We want to use this sample to estimate the true average level

```
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 118.0867
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 49.55809
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 59.8795
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 50.19293
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 50.51859
> var(sample(jura$Ni, size=30), na.rm=TRUE)
[1] 131.4138
```

- Estimates based on random samples fluctuate (a lot) around the true value

- We compute the 100 estimates corresponding to the 100 random samples using `apply()`
- We have

```
> S2 <- apply(all.samp, 2, var, na.rm=TRUE)
> S2[1:20]
[1] 80.06701 89.42725 55.37910 67.03496 48.35586 56.89323 39.10971
[8] 87.23956 79.60954 37.37551 49.88000 70.77083 85.95101 61.96498
[15] 53.14278 43.22899 44.45698 68.54886 49.69491 58.54554
```

More probability facts

Estimation of the variance

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is unbiased, i.e.

$$E[\hat{\sigma}^2] = \sigma^2.$$

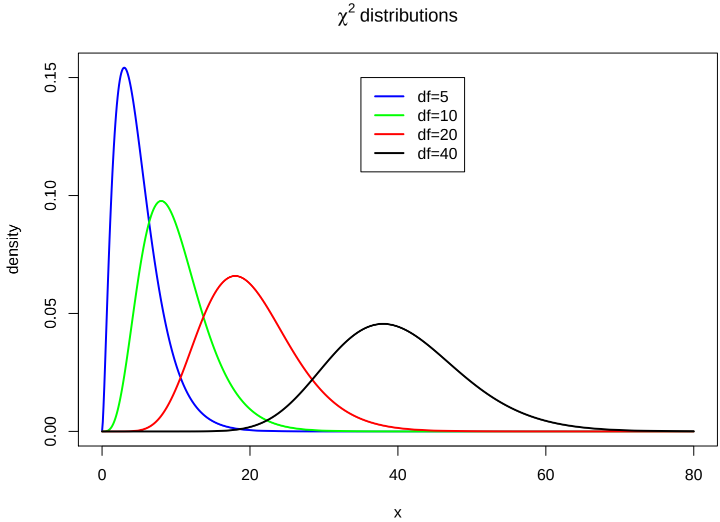
χ^2 distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Then,

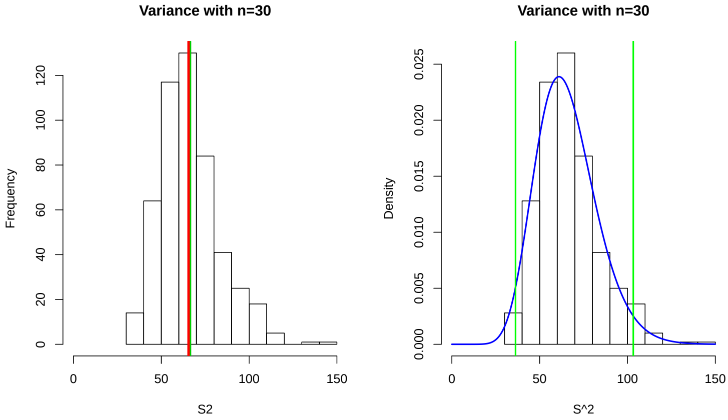
$$(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}.$$

There are $(n-1)$ independent RV (degrees of freedom) when computing S^2.

The χ^2 distributions



Illustration



Confidence Interval for the variance

Confidence Interval at level α

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Then,

$$[\hat{\sigma}^2_{inf}, \hat{\sigma}^2_{sup}] = [S^2(n-1)/x^{(n-1)}_{\alpha/2}, S^2(n-1)/x^{(n-1)}_{1-\alpha/2}],$$

where $x^{(n-1)}_p$ is such that $\mathbb{P}(\chi^2_{n-1} \leq p) = x^{(n-1)}_p$.

Proof

Using convergence towards χ^2,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(x^{(n-1)}_{\alpha/2} \leq (n-1)S^2/\sigma^2 \leq x^{(n-1)}_{1-\alpha/2}\right) \\
&= \mathbb{P}\left(1/x^{(n-1)}_{1-\alpha/2} \leq \sigma^2/(S^2(n-1)) \leq 1/x^{(n-1)}_{\alpha/2}\right) \\
&= \mathbb{P}\left(S^2(n-1)/x^{(n-1)}_{1-\alpha/2} \leq \sigma^2 \leq S^2(n-1)/x^{(n-1)}_{\alpha/2}\right)
\end{aligned}$$

Even more probability facts

Student t distribution

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_{(n)}$ be independent. Then,

$$\frac{X}{Y/\sqrt{n}} \sim t_n$$

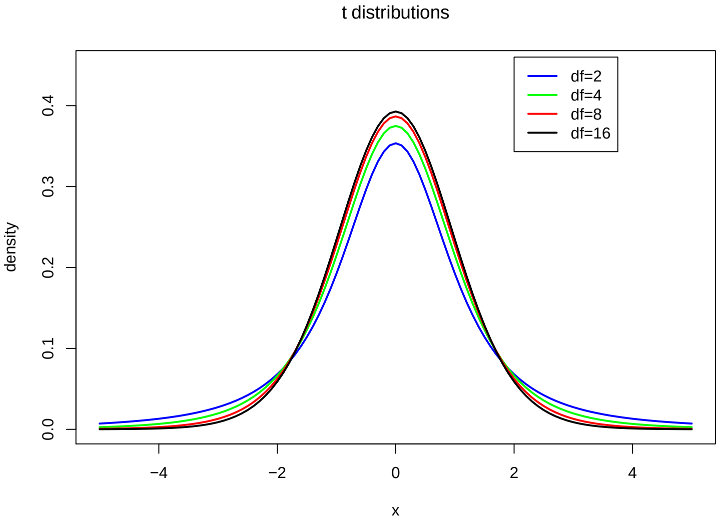
t distribution with n d.o.f.

Fisher F distribution

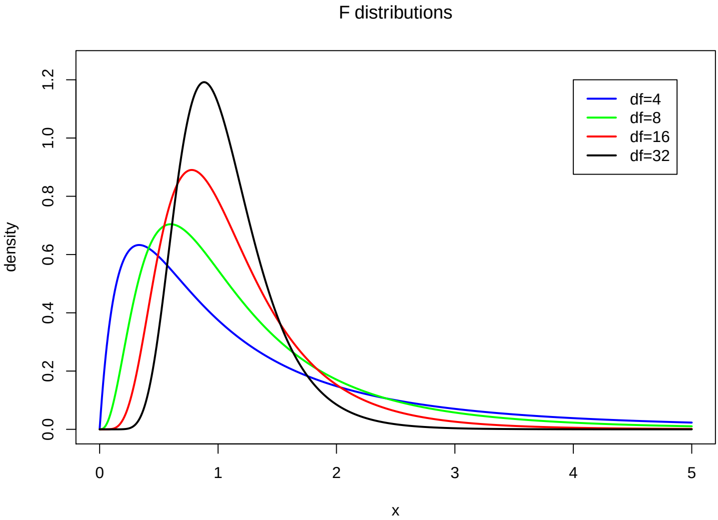
Let $X \sim \chi^2_{(n_X)}$ and $Y \sim \chi^2_{(n_Y)}$ be independent. Then,

$$\frac{X/\sqrt{n_X}}{Y/\sqrt{n_Y}} \sim F^{n_X}_{n_Y}.$$

The Student *t* distributions



The Fisher *F* distributions



Confidence interval Revisited

Confidence Interval: σ^2 is estimated

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a $\mathcal{N}(\mu, \sigma^2)$ RV. Using the definition of the t Student distribution:

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}]$$

where

$$\mathbb{P}(t_{(n-1)} \leq t_{1-\alpha/2}) = 1 - \alpha/2.$$

If $X_1, \ldots, X_n$ is not Gaussian, requires $n > 30$.

Confidence interval Revisited

Confidence Interval: σ^2 is estimated

Using the definition of the t Student distribution:

$$[\hat{\mu}_{inf}, \hat{\mu}_{sup}] = [\bar{X} - t_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}\frac{S}{\sqrt{n}}]$$

Since $t_{1-\alpha/2} \geq u_{1-\alpha/2}$, the interval is wider as compared to the case with known σ^2.
With $n = 30$:

$$\begin{aligned} \alpha = 10\% &\quad\Leftrightarrow\quad t_{0.950} = 1.70 \quad [u_{0.950} = 1.64] \\ \alpha = 5\% &\quad\Leftrightarrow\quad t_{0.975} = 2.04 \quad [u_{0.975} = 1.96] \\ \alpha = 1\% &\quad\Leftrightarrow\quad t_{0.995} = 2.76 \quad [u_{0.995} = 2.58] \end{aligned}$$

Same 1000 samples of size 30. One finds

$$\#(\mu < \hat{\mu}_{inf}) = 29 \quad [24]; \quad \#(\mu > \hat{\mu}_{sup}) = 24 \quad [19].$$

Expected value: 25.

Statistical tests

- According to former studies and/or expertise, one should have $\mu = 20$.
- A sample of size 30 provides $\bar{X} = 22.2$ and $S^2 = 52$.
- Is this a significant difference?

$\rightsquigarrow$ Need for formal statistical tests

Definition

Statistical test = Mathematical decision tool to check an hypothesis.

- Neutral, or ”null” hypothesis, H_0
- Alternative hypothesis, H_1

H_0 is not guilty unless proven otherwise.

Statistical tests

Test

$$H_0 \text{ vs. } H_1$$

We always test H_0 against an alternative. Both have to be clearly defined.

Two types of errors

	Decision	
	Do not reject H_0 Keep H_0	Reject H_0 Prefer H_1
H_0 true	Correct	Type I Error
H_1 true	Type II Error	Correct

- Level
$$\alpha = \mathbb{P}(\text{Type I Error})$$
(to be computed conditional on H_0 being true)
- Power
$$1 - \beta = \mathbb{P}(\text{No Type II Error})$$
(to be computed conditional on H_1 being true)

Statistical test: the very, very general procedure

H_0 is supposed to be true unless proven to be false.
$\Rightarrow$ computations are done conditional on H_0.

1. Define clearly the hypotheses H_0 and H_1
2. Set the level α
3. Use the relevant statistics (this is where the mathematical theory comes in), say T
4. Find the critical value of T, denoted t_c, as a function of n, α
5. Compute the value of T for the given sample, and compare to t_c
6. Conclude whether H_0 should be rejected or not

Power of a test

- The level α is set by the user.

$$1 - \alpha = \mathbb{P}(H_0 \mid H_0) = \mathbb{P}(\text{Not rejecting } H_0 | H_0 \text{ is true})$$

- Power

$$1 - \beta = \mathbb{P}(H_1 \mid H_1) = \mathbb{P}(\text{Rejecting } H_0 | H_1 \text{ is true})$$

Necessitates a complete specification H_1.

Example: testing the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

The power $1 - \beta$ is a function of μ_1.

Testing the mean

The average of Ni should be $\mu = 20$. In a sample of size 30, it is found that $\bar{X} = 22.2$ and $S^2 = 52$.

Should H_0 be rejected ?

1. Define the hypotheses $H_0 : \mu = 20$; $H_1 : \mu > 20$
2. Set a level: $1 - \alpha = 0.05$
3. Use the relevant distribution: $(\bar{X} - \mu)/(S/\sqrt{n-1}) \sim t_{(n-1)}$ with $n = 30$
4. If $(\bar{X} - \mu)/(S/\sqrt{n-1}) \sim t_{n-1}$ is “too large” I should reject H_0
5. One reads $\mathbb{P}(t_{(29)} \leq t_c) = 0.95$.

   t_c is the critical value. Here, $t_c = 1.70$.
6. $(\bar{X} - \mu)/(S/\sqrt{n-1}) = (22.2 - 20)/\sqrt{52/29} = 1.64 < 1.7$
7. The null hypothesis H_0 is not rejected.

   ”The sample was not able to prove H_0 was guilty”

Testing the mean: assessing the power

Example: testing the mean

$$H_0 : \ \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$

Some mathematics

$$\begin{aligned}
\mathbb{P}(H_1 \mid H_1) &= \mathbb{P}\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n-1}} \geq t_c\right) \\
&= \mathbb{P}\left(\frac{\bar{X}_0}{S/\sqrt{n-1}} \geq t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n-1}}\right) \\
&= 1 - F_{t_{n-1}}\left(t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n-1}}\right)
\end{aligned}$$

Testing the mean: assessing the power

Example: unilateral tests for the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$$

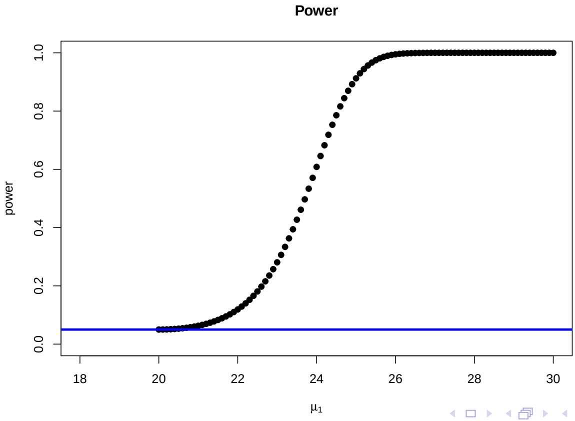
```
> delta = seq(0,10,by=0.1)
> tc    = qt(0.95,df=29)
> pow   = 1 - pt(tc - delta*delta/sqrt(var(Ni)),df=29)
> plot(20+delta,pow,main="Power",xlim=c(18,30),ylim=c(0,1),
>      xlab=expression(mu[1]),ylab="power",pch=19)
> abline(h=0.05,lwd=3,col="blue")
```

Testing the mean: assessing the power

Example: unilateral tests for the mean

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu = \mu_1 > \mu_0$$

$$X_1 = X_0 + (\mu_1 - \mu_0) \sim \mathcal{N}(\mu_0 + (\mu_1 - \mu_1), \sigma^2)$$



p-value

We do not set a level beforehand. Instead, one computes the probability of rejecting H_0, given the data.

Definition of the p-value

The probability of obtaining an ”equal or more extreme” test statistics than what was actually observed, assuming H_0 is true.

- A small p-value (≤ 0.05) indicates strong evidence against the null hypothesis, so it is rejected.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis (fail to reject).
- p-values very close to the cutoff (~ 0.05) are considered to be marginal (need attention).

Back to our first example

The average of Ni should be $\mu = 20$. In a sample of size 30, it is found that $\bar{X} = 22.2$ and $S^2 = 52$.

Should H_0 be rejected ?

$$
\begin{aligned}
p &= 1 - \mathbb{P}\left(t_{(n-1)} \leq (\bar{X} - \mu)/(S/\sqrt{n-1})\right) \\
&= 1 - \mathbb{P}\left(t_{29} \leq (22.2 - 20)/\sqrt{52/29}\right) \\
&= 1 - \mathbb{P}\left(t_{29} \leq 1.643\right) \\
&= 0.0556
\end{aligned}
$$

Fail to reject, but not by much. Requires attention.

```
> Ni.sample <- sample(jura$Ni,size=30)
> t.test(Ni.sample,alternative = "greater",mu=20)
```

Testing the variance

The variance of Ni should be $\sigma^2 = 65.5$ (H_0). In a sample of size 30, it is found that $S^2 = 90.2$.

Should H_0 be rejected ?

1. Define the hypotheses $H_0 : \sigma^2 = 65.5$; $H_1 : \sigma^2 > 65.5$
2. Use the relevant distribution:

   $$S^2/(\sigma^2/n) \sim \chi_{(n-1)},$$

   with $n = 30$
3. One reads $\mathbb{P}(\chi_{(29)} \leq 90.2 * 30/65.5) = 0.935$.

   The p-value is 0.065
4. The null hypothesis H_0 is not rejected.

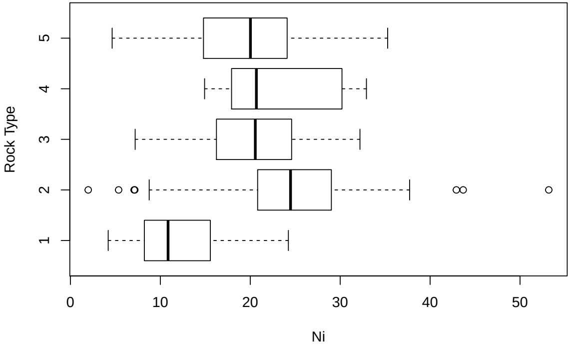
   But close

```
> install.packages("EnvStats")
> library(EnvStats)
> varTest(x = Ni.sample,alternative="greater",sigma.squared=65.5,conf.level = 0.
```

Back to data

Ni in different rock types

```
> boxplot(Ni ~ jura$rt, horizontal=T, xlab="Ni",ylab="Rock Type")
```



- Different means?
- Different variances?

Testing two means

Test

$$H_0 : \mu_1 = \mu_2; \quad H_1 : \mu_1 \neq \mu_2$$

i.e.

$$H_0 : \mu_1 - \mu_2 = 0; \quad H_1 : \mu_1 - \mu_2 \neq 0$$

with $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Under Gaussian hypothesis, we have

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}; \qquad \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

and

$$\bar{X}_1 \sim \mathcal{N}(\mu_1, \sigma^2/\sqrt{n_1}) \qquad \bar{X}_2 \sim \mathcal{N}(\mu_2, \sigma^2/\sqrt{n_2})$$

Testing two means

Hence,

$$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

and

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2 = 0, \sigma^2(1/n_1 + 1/n_2)\right).$$

Therefore, the test statistics is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)} \sqrt{n_1 + n_2 - 2} \sim t_{(n_1+n_2-2)}$$

Testing two means: example

Jura data:

rt	1	2	3	4	5
$\bar{X}$	12.3	25.0	20.4	22.9	18.8
S^2	31.0	54.6	32.0	50.5	57.2
n	76	124	89	6	64

Mean, variance and number of data, according to rock type

T-tests:

	2	3	4	5
1	0	0	$1.6\,10^{-5}$	$1.5\,10^{-5}$
2	–	$1.1\,10^{-5}$	0.25	$1.2\,10^{-7}$
3	–	–	0.16	0.07
4	–	–	–	0.10

p-value of T tests, assuming identical variance

```
> t.test(Ni[jura$Rock==1], Ni[jura$Rock==2], alternative = "two-sided")
```

Testing two variances

Test

$$H_0 : \sigma_1^2 = \sigma_2^2; \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

i.e.

$$H_0 : \sigma_1^2/\sigma_2^2 = 1; \qquad H_1 : \sigma_1^2/\sigma_2^2 \neq 1$$

Under Gaussian hypothesis, we have

$$\frac{n_1 S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}; \qquad \frac{n_2 S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

and

$$\frac{n_1 S_1^2}{n_1 - 1} \Bigg/ \frac{n_2 S_2^2}{n_2 - 1} \sim F_{n_1-1, n_2-1}.$$

Testing two variances: example

Jura data:

rt	1	2	3	4	5
S^2	31.0	54.6	32.0	50.5	57.2
n	76	124	89	6	64

Variance and number of data, according to rock type

F-tests:

	2	3	4	5
1	0.003	0.437	0.280	0.006
2	–	0.004	0.520	0.423
3	–	–	0.298	0.007
4	–	–	–	0.353

p-value of F tests

```
> var.test(x = jura.Ni[jura$Rock==1],y = jura.Ni[jura$Rock==1],
           alternative="two-sided" ,conf.level = 0.95)
```

Comparison test(s)

Gaussian hypothesis: not a problem for large n, thanks to CLT

- $\bar{X} \to \mathcal{N}$ as $n \to \infty$
- $nS^2/\sigma^2 \to \chi^2_{n-1} \to \mathcal{N}$ as $n \to \infty$
- $t_{n-1} \to \mathcal{N}$ as $n \to \infty$

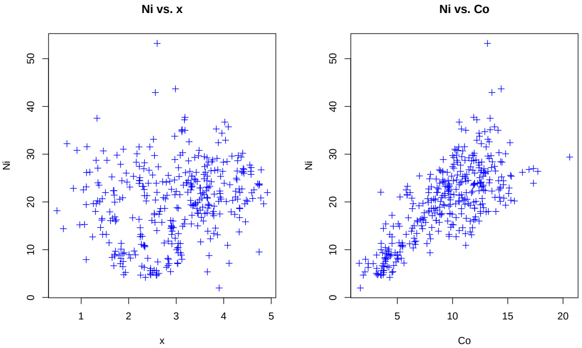
For moderate n, (say $n \leq 30$), the order is important:

1. Test for equal variance first;
2. If not rejected, test means

Unit 2
Regression models and ANOVA

Examples

Any relationship between Ni and x, or between Ni and Co ?



Objectives

We have two series of values, X and Y.

1. We wish to know whether there is some sort of relationship between X and Y
   Correlation coefficient, rank correlation, etc.
2. We wish to know whether Y can be predicted from X
   Linear regression, Generalized linear regression, etc.

Estimation, tests, predictions

Correlation coefficient

Definition from Probability

Let X and Y be two random variables. The linear correlation coefficient is

$$r = \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Property

$$\rho \in [-1, 1],$$

with

1. If $\rho = 1$, there is a linear relationship: $Y = a + bX$, with $b > 0$
2. If $\rho = 0$, there is no linear relationship at all
3. If $0 < \rho < 1$ there is *some amount* of linear relationship
4. If $\rho < 0$ the linear relationship is negative (Y decreases as X increases)

Estimation of the linear correlation coefficient

Estimator

Let (X_i, Y_i), with $i = 1, \ldots, n$, be a bivariate series of values.

$$\hat{\rho} = \frac{\hat{C}_{XY}}{S_X S_Y}$$

with

$$\hat{C}_{XY} = \frac{1}{n} \sum_{i=1^n} (X_i - \bar{X})(Y_i - \bar{Y})$$

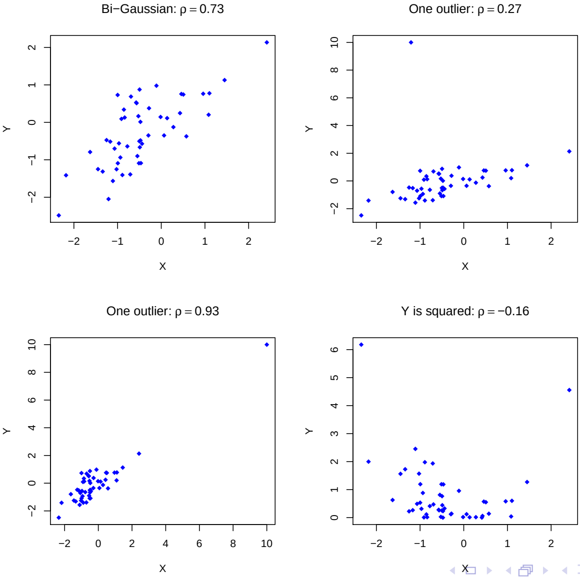
Sample of size 30, without repetitions. Correlation coefficient between Ni and Co

1. Cor #1: 0.73
2. Cor #2: 0.81
3. Cor #3: 0.79

   ⋮

4. Mean #100: 0.71

Correlation coefficient



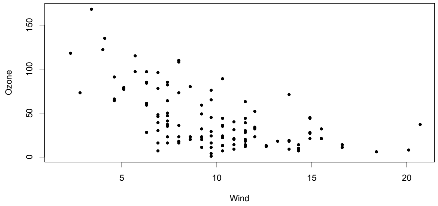
Correlation coefficient

- Correlation does not (always) mean causality. It could be
  - Spurious.
    e.g., presence of outliers, compositional data, ...
  - Due to a common cause
    e.g. life expectancy increases with the consumption of lobsters; Ni increases with Co
- Absence of Correlation does not necessarily mean absence of relationship. Only true for Gaussian vectors

Testing a correlation coefficient is difficult at this stage. Better within a regression context

Regression

- Target: identifying a model that relates variable `Ozone` to variable `Wind`
- Correlation between the two variables is (about) -0.6
- The scatterplot provides further insights about the negative relationship



- In other terms, we are interested in an asymmetric model where `Wind` is used to "predict" `Ozone`.

Remember

Statistics starts with a problem, proceeds with the collection of data, continues with the data analysis and finishes with conclusions.

It is a common mistake of inexperienced statisticians to plunge into a complex analysis without paying attention to the objectives or even whether the data are appropriate for the proposed analysis.

As Einstein said, the formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.

J.J. Faraway, (2015) Linear models with R

Linear regression

- Model

  $$\text{observed value} = \text{deterministic component} + \text{random component}$$

- Linear regression model: n observations $y_1, y_2, \ldots, y_n$:

  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

  where ϵ_i are error terms
- Variable y is termed response
- Variable x is termed covariate or explanatory variable or predictor
- β_0, β_1 are termed parameters or regression coefficients
- In a linear model the parameters enter linearly
  - $y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i$ (linear model)
  - $y_i = \beta_0 + \beta_1 x_i^{\beta_2} + \epsilon_i$ (nonlinear model)
- We suppose that $E(y_i) = \beta_0 + \beta_1 x_i$, $var(y_i)$ constant

- We want to predict the ozone concentration (the problem!)
- In our example, the response is `Ozone` and the predictor is `Wind`
- The linear regression model makes sense if
  - the errors are not systematic but they "fluctuate" around zero
  - the spread of the errors is more or less constant, that is the level of the fluctuations around zero does not depend on the observed values of the two variables
- In other terms, we ask that errors have zero mean and constant variance
- Quantities β_0 and β_1 are termed the intercept and the slope of the regression line, respectively
- The pair of parameters (β_0, β_1) are also termed regression coefficients

- Regression coefficients are parameters that need to be estimated from the observed data
- By varying the pair of regression coefficients, we obtain infinite possible regression lines
- The problem is how to select the line which better fits the data according to some criterion
- Many methods available to estimate β_0 and β_1, the most diffuse is ordinary least squares (OLS)
- Sum of squared residuals

$$\text{SSR}(\beta_0, \beta_1) = \sum_{i=1}^{n} \{\underbrace{y_i - (\beta_0 + \beta_1 x_i)}_{\text{raw residual}}\}^2$$

raw residuals $r_i^{\text{raw}} = y_i - (\beta_0 + \beta_1 x_i)$

- The pair $(\hat{\beta}_0, \hat{\beta}_1)$ that minimizes $\text{SSR}(\beta_0, \beta_1)$ identifies the best regression line in terms of the method of ordinary least squares

- Symbol hat ˆ denotes the data-based estimate of a parameter:
  - least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$
  - raw residuals computed at $(\hat{\beta}_0, \hat{\beta}_1)$ are errors estimates

$$\hat{\epsilon}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

  which are used for diagnostic (to answer questions like "does the chosen line really fit well the data?")

- In **R** linear regression computed with function `lm`

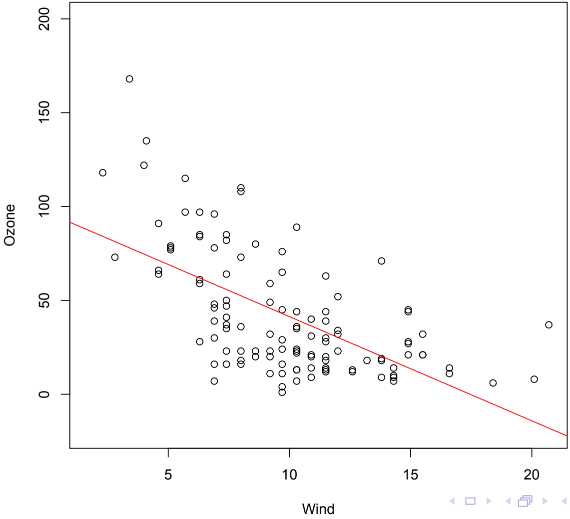
```
> fit <- lm(Ozone~Wind, data=airquality)
> fit
Call:
lm(formula = Ozone ~ Wind, data = airquality)

Coefficients:
(Intercept)          Wind
96.873        -5.551
```

- The (ordinary) least squares regression line is

$$\texttt{Ozone} = 96.87 - 5.55\,\texttt{Wind}$$

Useful to visualize observed points and the fitted model

```
> plot(Ozone~Wind, data=airquality, ylim=c(-20, 200))
> abline(fit, col="red")
```

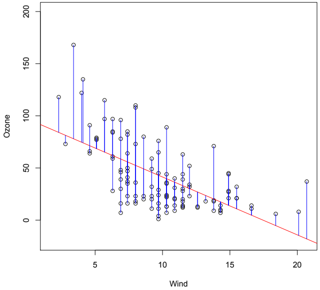


- Predicted values of `Ozone`

```
> ozone.hat <- predict(fit, newdat a= + data.frame(Wind=airquality$Wind))
```

- We can compare predictions versus observed values

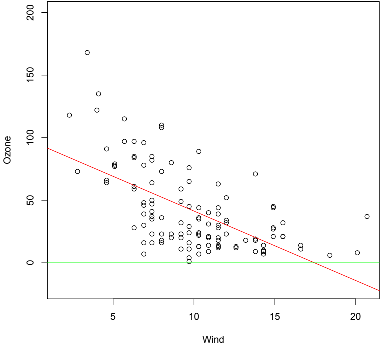
```
> plot(Ozone˜Wind, data=airquality, ylim=c(-20, 200))
> abline(fit, col="red")
> segments(x0=airquality$Wind, y0=airquality$Ozone, x1=airquality$Wind,
   y1=ozone.hat, col="blue")
```



- Blue segments are the raw residuals
- The regression line is chosen so to minimize the sum of the squared lengths of the blue segments

- Does the fitted model make sense? not really because it gives negative predictions for large values of `Wind`!

```
> plot(Ozone~Wind, data=airquality,
+ ylim=c(-20, 200))
> abline(fit, col="red")
> abline(h=0, col="green")
```

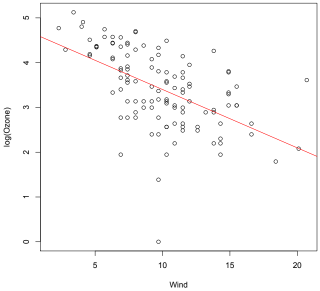


- The points pattern suggests that the relationship between `Ozone` and `Wind` is not well described by a regression line both at small and large values of `Wind`

Regression and Transformations

- Solution: transform the response so to
  - avoid non-sense negative predictions of `Ozone` levels
  - make the relationship between `Ozone` and `Wind` ”more” linear
- Try with a log-transformation of `Ozone`
- Logarithm maps positive numbers to unrestricted numbers, thus avoiding the risk of non-sense negative predictions

```
> fit2 <- lm(log(Ozone)~Wind, data=airquality)
> plot(log(Ozone)~Wind, data=airquality)
> abline(fit2, col="red")
```



Outliers

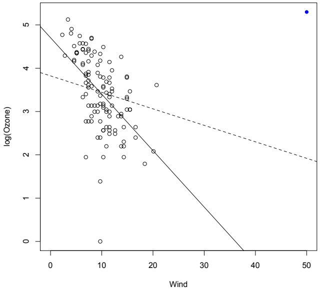
- Fit on log-scale not too bad, except for a few points
- The worse fitted point is the one with `Ozone` about 1 and `Wind` somehow smaller than 10

```
> id <- which.min(airquality$Ozone)
> id
[1] 21
> airquality[id,]
Ozone Solar.R Wind Temp Month Day
21      1       8  9.7   59     5  21
```

- Observation 21 of `Ozone` is an outlier on the log-scale
- The term outlier denotes an observation that is ”distant” from the rest of the data
- Is the presence of one or more outliers a problem? not much in this case, but sometimes outliers may have a strong impact
- An outlier which significantly affects the fitted regression line is called an influential point

As an example of influential point consider the hypothetical observation (`Ozone`=200, `Wind`=50)

```
> plot(log(Ozone)~Wind, data=airquality,
> ylim=c(0, log(200)), xlim=c(0, 50))
> points(50, log(200), col="blue", pch=16)
> abline(fit2, lty=1)
> fit3 <- lm( c(log(Ozone),log(200))~c(Wind, 50),
+ data=airquality)
> abline(fit3, lty=2)
```



- One single observation may have a substantial effect on the fitted regression line!
- Does this make sense? not really, as a good statistical model should fit well the great majority of the data and not be influenced too much from few isolated observations (which often have a ”special” meaning)
- Solution: use a fitting method that it is more resistant to outliers
- Robust Statistics. . .

Residuals

- Diagnostic is *very important* to validate the fitted model
- Helpful to visualize the residuals in way to check:
  - absence of systematic effects
  - stable variance
  - ...
- Residuals from an `lm`-fitted object are accessed by function `residuals()`

```
> res <- residuals(fit2)
> summary(res)
Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-3.44000 -0.49980  0.06051  0.00000  0.53750  1.60500
```

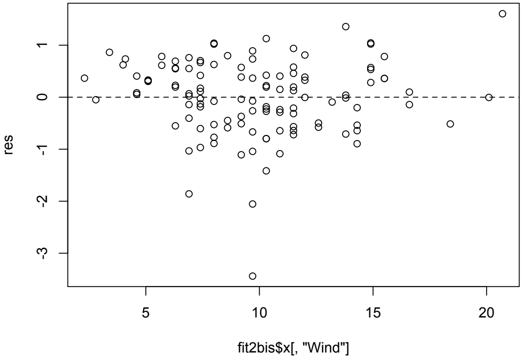
- A sensible diagnostic plot is the scatterplot of the residuals against the predictor
- Caution: there are missing values in `Ozone` which are discarded from the regression fitting. Hence, the number of residuals is not equal to the number of observed values of `Wind`

```
> length(res)
[1] 116
> length(airquality$Ozone)
[1] 153
> sum(is.na(airquality$Ozone)) #how many NAs?
[1] 37
> length(airquality$Wind)
[1] 153
```

- Hence, we need to compare the residuals with values of `Wind` corresponding to non-missing `Ozone` values

- We can refit the model with option `x=TRUE` to extract these values

```
> fit2bis <- lm(log(Ozone)~Wind, data=airquality, x=TRUE)
> head(fit2bis$x)
(Intercept) Wind
1            1  7.4
2            1  8.0
3            1 12.6
4            1 11.5
6            1 14.9
> airquality[1:8,]
Ozone Solar.R Wind Temp Month Day
1     41     190  7.4   67     5   1
2     36     118  8.0   72     5   2
3     12     149 12.6   74     5   3
4     18     313 11.5   62     5   4
5     NA      NA 14.3   56     5   5
6     28      NA 14.9   66     5   6
7     23     299  8.6   65     5   7
8     19      99 13.8   59     5   8
> plot(x=fit2bis$x[,"Wind"], y=res)
> abline(h=0, lty="dashed")
```



The scatterplot of the residuals versus `Wind` shows some problems for small values of `Wind` (in addition to the well-known outlier)

Multiple regression

- We may ask whether a more elaborated model can better fit the data
- For example the quadratic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, \ldots, n$$

where y_i is the log-transformed `Ozone`

```
> fit3 <- update(fit2bis, .~.+I(Wind^2) )
> fit3
Call:
lm(formula = log(Ozone) ~ Wind + I(Wind^2), data = airquality,
x = TRUE)

Coefficients:
(Intercept)          Wind     I(Wind^2)
5.83475     -0.36945       0.01116
```

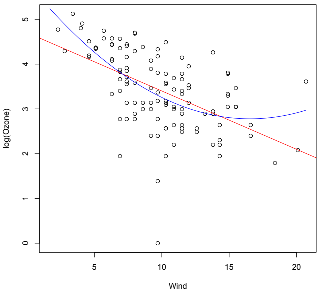
- Function `update()` is used to add the quadratic term which needs to be specified by function `I()`

- The coefficient of the quadratic term is very small: does this mean that it cannot be distinguished from zero (which means no quadratic effect)? or is the small value due to the scale of the squared `Wind`?

```
> range(airquality$Wind^2)
[1]   2.89 428.49
```

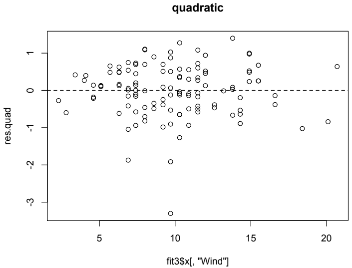
- Plot of the fitted quadratic model

```
> plot(log(Ozone) ~ Wind , data = airquality)
> abline(fit2bis, col="red")
> curve( coef(fit3)[1]+coef(fit3)[2]*x+coef(fit3)[3]*x^2,
+ col="blue", add=TRUE)
```



- Residuals of the the quadratic model

```
> res.quad <- residuals(fit3)
> summary(res.quad)
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.30200 -0.43220  0.05994  0.00000  0.49550  1.40000
> plot(x=fit3$x[,"Wind"], y=res.quad, main="quadratic")
> abline(h=0, lty="dashed")
```



- The residuals of the quadratic model improve on with respect to those of the linear model

Coefficient of determination

- The coefficient of determination R^2 is the proportion of variability in the response that is accounted for by the statistical model
- Ingredients:
  - the total sum of squares
  $$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
  this is n times the variance of y, $s_y^2 = \mathrm{SST}/n$
  - the residual sum of squares
  $$\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
  - the explained sum of squares
  $$\mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
- Decomposition of the sum of squares: $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$

- The R^2 index is defined as

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST}$$

- Properties of R^2 index
  - R^2 assumes values between 0 and 1
  - the better the model, the smaller the residual sum of squares (=R^2 closer to 1)
- The R^2 index can be obtained from the `summary` of an `lm` object

```
> summary(fit2)$r.squared
[1] 0.289894
```

- We can compare this value with the R^2 of the quadratic model

```
> summary(fit3)$r.squared
[1] 0.3431879
```

- The latter value is somehow larger, thus supporting the use of the quadratic model, but ...
- ...this conclusion requires care because it can be shown that inclusion of any further predictor yields to an improvement of R^2

- As an illustration, suppose we add a randomly generated predictor
- We can generate the random predictor with the following code

```
> n.obs <- nrow(airquality)
> set.seed(12345)
> simul <- rnorm(n.obs)
> simul[1:10]
[1]  0.5855288  0.7094660 -0.1093033 -0.4534972  0.6058875
[6] -1.8179560  0.6300986 -0.2761841 -0.2841597 -0.9193220
> cor(simul, airquality$Ozone, use="complete.obs")
[1] 0.03607404
```

- Although variable `simul` has been generated in way to be completely unrelated with `Ozone`, some (very) small degree of correlation is observed
- Now, consider the multiple regression model

$$\log(\text{Ozone}) = \beta_0 + \beta_1\,\text{Wind} + \beta_2\,\text{Wind}^2 + \beta_3\,\text{simul} + \epsilon,$$

where the term *multiple* indicates that more than one predictor is used

- Multiple regression

```
> fit4 <- update(fit3, .~.+simul)
> fit4

Call:
lm(formula = log(Ozone) ~ Wind + I(Wind^2) + simul, data = airquality,
x = TRUE)

Coefficients:
(Intercept)         Wind    I(Wind^2)        simul
5.831822    -0.369093     0.011154     0.006625
> summary(fit4)$r.squared
[1] 0.3432537
> summary(fit3)$r.squared
[1] 0.3431879
```

- The R^2 for the model with the random predictor is slightly larger than the one without the random predictor

- The adjusted R^2 index is constructed so that irrelevant increases of the R^2 are penalized

$$R^2_{\text{adj}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)},$$

where p is the number of regression coefficients
  - 2 in the linear model
  - 3 in the quadratic model
  - 4 in the quadratic plus random predictor model
- The adjusted R^2 indices for the three models are

```
> summary(fit2)$adj.r.squared
[1] 0.283665
> summary(fit3)$adj.r.squared
[1] 0.3315629
> summary(fit4)$adj.r.squared
[1] 0.3256622
```

- Correctly, the adjusted R^2 reveals that the improvement due to the random predictor is irrelevant and thus the quadratic model is preferred

Standardized Residuals

- Validation of linear regression models often based on standardized residuals (aka Pearson residuals)
- Standardization has two advantages:
  - removing scale effects
  - standardized residuals are realizations of a standard normal variable (if the model is correctly specified)
- Standardized residuals for the quadratic model

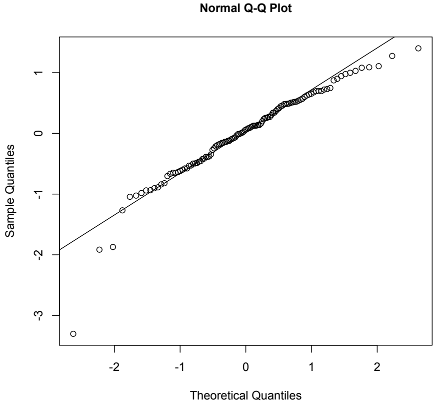
$$\log(\text{Ozone}) = \beta_0 + \beta_1 \,\text{Wind} + \beta_2 \,\text{Wind}^2 + \epsilon$$

```
> res.standard <- residuals(fit3)
> summary(res.standard)
Min.   1st Qu.   Median      Mean   3rd Qu.      Max.
-3.30200 -0.43220  0.05994  0.00000  0.49550  1.40000
```

- If residuals were realizations from $\mathcal{N}(0,1)$ then the probability of observing something smaller than -3 is 0.1%
- Thus, the residual equal to -3.3 is quite unusual
- Which one is the observation corresponding to this residual?

- Helpful checking the normality of the standardized residuals by normal probability plots

```
> qqnorm(res.standard)
> qqline(res.standard)
```



- Standardized residuals look rather OK (but not perfectly OK), except for the well-known outlier…

- Consider again the multivariate regression model

$$\log(\text{Ozone}) = \beta_0 + \beta_1\,\text{Wind} + \beta_2\,\text{Wind}^2 + \beta_3\,\text{simul} + \epsilon$$

where `simul` was a random normal sample completely unrelated to `Ozone` or `Wind`

```
> fit4
Call:
lm(formula = log(Ozone) ~ Wind + I(Wind^2) + simul,
data = airquality, x = TRUE)

Coefficients:
(Intercept)         Wind    I(Wind^2)        simul
5.831822     -0.369093     0.011154     0.006625
```

- We know that the true value of the coefficient for *simul* is zero. Its estimate is very small but not zero because of the sample uncertainty
- Also the estimated coefficient for `Wind`^{2} is quite small, but the R^2 index suggests that the squared term was useful
- In fact, we already said that the estimated coefficient for `Wind`^{2} is small as a consequence of the relative large values of `Wind`^{2}

Validation

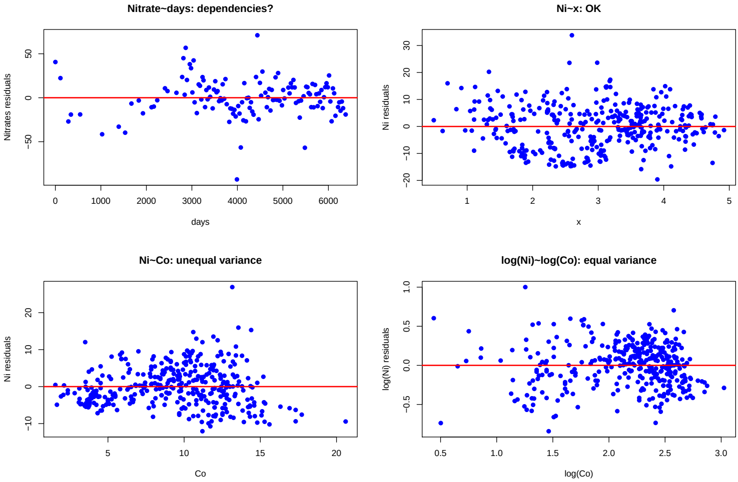
Checking the model:

- Checking the linearity
- Checking the assumptions on ϵ: equal variance, Gaussian, independent

If the model is not validated, more complex models should be found

- Transform the variables: squares, log, exp, cosine, etc...
- Add more covariates
- Introduce temporal or spatial dependencies [later !]

Validation


Nitrate~days: dependencies?
Nitrates residuals
days
Ni~x: OK
Ni residuals
x
Ni~Co: unequal variance
Ni residuals
Co
log(Ni)~log(Co): equal variance
log(Ni) residuals
log(Co)

Linear model

Some theory for the linear model

General notation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

- $\mathbf{Y}$ is a n vector of observed variables
- $\mathbf{X}$ is a $n \times p$ matrix of covariates (continuous or categorical). There are p covariates; one covariate is a column of one, accounting for the mean
- β is the $p \times 1$ vector of unknown parameters
- ϵ is a $n \times 1$ vector of i.i.d. random values, usually $\sim \mathcal{N}(0, \sigma^2)$

$\mathbf{X}$ is called the design matrix. We assume we can invert it.

Some theory for the linear model

Linear model

General notation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Attention ! Linear means linear combinations of covariates. Covariates could be t, $\cos t$, t^2, etc..

Remember log(Ni) vs. log(Co)

Some tasks in regression:

- Estimate β
- Test covariates
- Test models against each other
- Select the best model (if any)

Regression with two variables; ANOVA

Some theory for the linear model

Projection

Mathematically, a linear model is a projection onto the subspace spanned by the covariates, (where the constant function being one of them).

One seeks the vecteur $\hat{\beta}$ such that

$$||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2$$

is minimum.

Therefore, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ is a projection.

The relationship

$$SS_T = SS_E + SS_R$$

is nothing but the Pythagoras theorem in this abstract space.

Some theory for the linear model

Projection

In this framework one can show that

- The estimator $\hat{\beta}$ is unbiased, i.e.

$$E[\hat{\beta}] = \beta$$

- Under Gaussian hypothesis, it is also the ML estimator, with optimality properties
- The estimator of the variance is

$$\hat{\sigma}^2 = SS_R/(n-p)$$

- The coefficient of determination is

$$R^2 = SS_E/SS_T = 1 - SS_R/SS_T.$$

  It is the proportion of variance explained by the model.
- The adjusted coefficient of determination is

$$R^2_* = 1 - SS_R/(n-p-1)\Big/ SS_T/(n-1) < R^2$$

Coefficient of determination

The coefficient of determination, denoted R^2, is

$$R^2 = \frac{SS_E}{SS_T} = 1 - \frac{SS_R}{SS_T}$$

It measures how well the model fits the data. By definition, $O \leq R^2 \leq 1$.

Properties

- For a simple regression model, we have

$$R^2 = \rho^2$$

- Under Gaussian hypothesis for ϵ_i, we have for $\beta_1 = 0$,

$$\frac{SS_E}{SS_R} \sim \mathcal{F}(1, n-2)$$

Some theory for the linear model

Nested models

Model M_0 is nested in model M_1 if model M_0 can be obtained from M_1 by removing some covariates (i.e. some columns of $\mathbf{X}$).

e.g. Ni, as a function of Co only is a nested model of Ni as a function of Cd and Co.

- In regression, we would like to know whether a subset of variable is sufficient, or if additional variables are necessary
- In analysis of variance, we would like to know if one factor can be removed

Obviously M_1 has more parameters, it is likely to fit the data better (we add dimensions in the subspace in which we project):

$$SS_{M_1} > SS_{M_0}$$

Is this increase significant or is it due to chance? Do a statistical test!

Some theory for the linear model

Testing nested models

Let us test "H_0 : model M_0 is true" vs. "H_1 : model M_1 is true"

Theorem

Under Gaussian hypothesis for ϵ,

$$\frac{(SS_{M_1} - SS_{M_0})/(p_1 - p_0)}{SS_{R_1}/(n - p_1)} \sim \mathcal{F}(p_1 - p_0, n - p_1)$$

Note: this a generalization of the result seen for linear regression, with M_1 being for $a + bx$ and M_0 for intercept a only.

Regression with two variables

```
fit = lm(log(Ni) ~ long + log(Co), data = jura)
summary(fit)
Call:
lm(formula = log(Ni) ~ long + log(Co), data = jura)

Residuals:
Min        1Q    Median       3Q      Max
-0.83764 -0.19079  0.01704  0.19310  0.98855

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.44983    8.13293  -1.039    0.300
long         1.37677    1.18815   1.159    0.247
log(Co)      0.88247    0.03245  27.194   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2806 on 356 degrees of freedom
Multiple R-squared:  0.6909,Adjusted R-squared:  0.6891
F-statistic: 397.8 on 2 and 356 DF,  p-value: < 2.2e-16
```

Introduction to ANOVA: the analysis of variance

Sometimes, data are categorical, or ordinal with very few modalities: they are factors

- `Landuse` and `Rock` in the Swiss Jura data set
- Age, with few modalities: child, young, adult, senior [fish ??]
- Type of rocks, with few modalities according to e.g. porosity
- ...

A linear regression does not make much sense. We need to do something else.

Modalities are considered as levels of the factor. For example

- Factor = `Rock`: Levels = `Argovian`, `Kimmeridgian`, `Portlandian`, `Quaternary`, `Sequanian`
- Factor = `LandUse`: Levels = `Forest`, `Meadow`, `Pasture`, `Tillage`

Introduction to ANOVA: the analysis of variance

Several cases:

- One factor, several levels
- Two factors, several levels: balanced or unbalanced
- Many factors, two (or many) levels, generally unbalanced
- Optimal design

Notations:

- $i = 1, \dots, I$, are the levels of the factor
- $k = 1, \dots, n_i$, are the repetitions within level i of the factor
- There is a total of $n = \sum_{i=1}^{I} n_i$ data
- If there is a second factor, we use index $j = 1, \dots, J$,
- n_{ij} is the number of repetitions of data within level $(i, j) \in I \times J$.

ANOVA with one factor

The model

We write

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

with $i = 1, \ldots, I$, $k = 1, \ldots, n_i$ and

$$\epsilon_{ik} \sim \mathcal{N}(0, \sigma^2), \quad \text{i.i.d}$$

The values α_i are the effect of level i.

- There are $I + 1$ parameters for the mean $(\mu, \alpha_1, \ldots, \alpha_I)$. This is one too many. We will have to impose constraints, e.g; $\alpha_1 = 0$ (as in `lm()`)
- Equivalent to the model

$$Y_{ik} = \mu_i + \epsilon_{ik}$$

ANOVA with one factor

Marix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \epsilon,$$

with, for $I = 5$,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and $\boldsymbol{\theta} = (\mu, \alpha_1, \ldots, \alpha_5)^t$

The first column is the sum of all other columns. Matrix **X** is of rank I.
Need to impose constraints.

- A set of constraints reducing the number of parameters is called a contrast.
- The function `lm` uses $\alpha_1 = 0$: pretty simple
- An other natural possibility is to impose $\sum_{i=1}^{I} \alpha_i = 0$
- Estimates for each level will depend upon the contrast, when design is unbalanced

ANOVA with one factor

It is convenient to write

$$Y_{i.} = n_i^{-1} \sum_{k=1}^{n_i} Y_{ik}, \quad Y_{..} = n^{-1} \sum_{i=1}^{I} n_i Y_{i.}$$

Source	DF	Sum of squares	Mean Sum of Squares
Model	$I-1$	$SS_E = \sum_{i=1}^{I} n_i (Y_{i.} - Y_{..})^2$	$SS_M/(I-1)$
Residuals	$n-I$	$SS_R = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{ik} - Y_{i.})^2$	$SS_R/(n-I)$
Total	$n-1$	$SS_T = \sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{ik} - Y_{..})^2$	$SS_T/(n-1)$

Remember:

$$R^2 = SS_E/SS_T$$

ANOVA with one factor

Test

$$H_0 : \{\alpha_1 = \cdots = \alpha_I = 0\} = \{Y_{ik} = \mu + \epsilon_{ik},\ \forall i\}$$

vs.

$$H_1 : \{\exists i : \alpha_i \neq 0\} = \{Y_{ik} = \mu + \alpha_i + \epsilon_{ik}, \forall i\}$$

Test statistics

$$F = \frac{SS_M/(I-1)}{SS_R/(n-I)} = \frac{\text{Explained variance by Model}}{\text{Residual variance}}$$

Under H_0, and with a Gaussian hypothesis,

$$F \sim \mathcal{F}_{I-1,n-I} \quad \Rightarrow \quad \mathbb{P}(F > \mathcal{F}_{I-1,n-I})$$

Rock type in Swiss Jura

$$R^2 = 7738/23454 = 0.33; \quad F = \frac{7738/4}{15716/354} = 43.6$$

$$\mathbb{P}(\mathcal{F}_{4,355} > 43.6) = 0$$

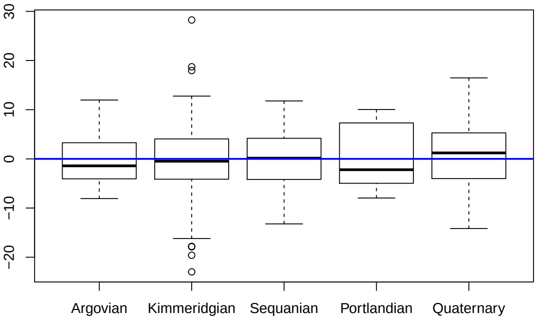
There is a highly significant effect of rock types

ANOVA: Ni ˜ Rock

```
> fit = lm(Ni ~Rock,data=jura)

> summary(fit)
Call:
lm(formula = Ni ~ Rock, data = jura)

Residuals:
Min        1Q    Median       3Q       Max
-22.9798  -4.1086  -0.3598   4.2306  28.2402

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        12.2784     0.7643  16.065  < 2e-16 ***
RockKimmeridgian   12.6814     0.9707  13.065  < 2e-16 ***
RockSequanian       8.1405     1.0407   7.822 6.03e-14 ***
RockPortlandian    10.6082     2.8255   3.754 0.000203 ***
RockQuaternary      6.5303     1.1304   5.777 1.67e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.663 on 354 degrees of freedom
Multiple R-squared:  0.3299,Adjusted R-squared:  0.3223
F-statistic: 43.57 on 4 and 354 DF,  p-value: < 2.2e-16
```

ANOVA: Ni ˜ Rock

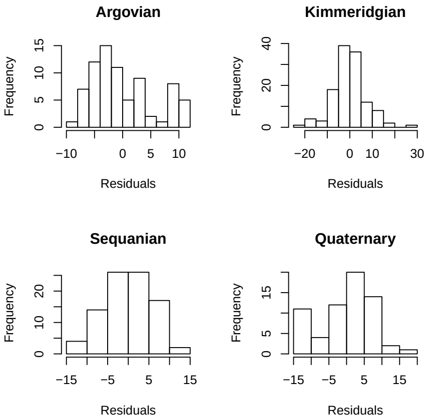
```
boxplot(lmNi$residuals ~ rt)
abline(h=0,col="blue",lwd=2)
```



Note: effect of unbalanced design.

ANOVA: Ni ~ rt

```
> par(mfrow=c(2,2))
> for (k in c(1,2,3,5)) hist(fit$residuals[jura$Rock==levels(jura$Rock)[k]],
                           xlab="Residuals",main=levels(jura$Rock)[k])
```



Note: Not quite Gaussian,

ANOVA with two factors

- $i = 1, \ldots, I$, are the levels of the first factor
- $j = 1, \ldots, J$, are the levels of the second factor
- $k = 1, \ldots, n_{ij}$, are the repetitions within levels $(i, j) \in I \times J$ of the factor
- There is a total of $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ data

The mathematics become cumbersome unless balanced design, i.e. $n_{ij} = K$.
Important mathematical properties follow.

This is not the case for Swiss Jura data set:

```
> freqJ
     [,1] [,2] [,3] [,4]
[1,]   11   10   53    2
[2,]   32   32   57    3
[3,]    5   31   51    2
[4,]    3    1    2    0
[5,]    0    8   55    1
> average
        [,1] [,2] [,3] [,4]
[1,] 0.03 0.03 0.15 0.01
[2,] 0.09 0.09 0.16 0.01
[3,] 0.01 0.09 0.14 0.01
[4,] 0.01 0.00 0.01 0.00
[5,] 0.00 0.02 0.15 0.00
```

ANOVA with two factors

The model

We write

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

with $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, n_{ij}$ and

$$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2), \quad \text{i.i.d}$$

The values α_i, β_j and γ_j are the effect respectively of level i, level j and interaction ij.

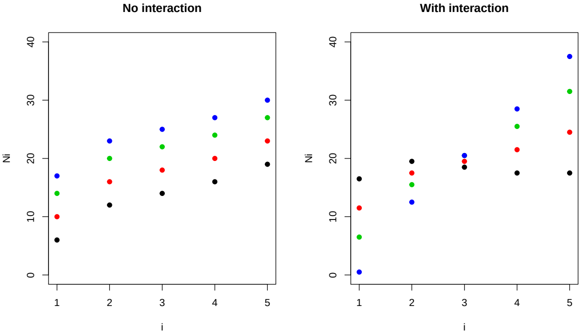
Interaction

When $\gamma_{ij} = 0$, we have for all $i = 1, \ldots, I$

$$\mu_{i1} - \mu_{i2}$$

ANOVA with two factors

Example inspired from Jura Swiss data set:



ANOVA with two factors

It is convenient to write

$$Y_{ij\cdot} = n_{ij}^{-1} \sum_{k=1}^{n_{ij}} Y_{ijk}, \quad Y_{i\cdot\cdot} = n_{i\cdot}^{-1} \sum_{j=1}^{J} n_{ij} Y_{ij\cdot}, \quad Y_{\cdots} = n^{-1} \sum_{i=1}^{I} n_{i\cdot} Y_{i\cdot\cdot},$$

with $n_{i+} = \sum_{j=1}^{n} n_{ij}$.

Source	DF	Sum of squares	Mean Sum of Squares
Model	$IJ - 1$	$SS_M = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(Y_{ij\cdot} - Y_{\cdots})^2$	$SS_M/(IJ-1)$
Residuals	$n - IJ$	$SS_R = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_i}(Y_{ijk} - Y_{ij\cdot})^2$	$SS_R/(n-IJ)$
Total	$n - 1$	$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_i}(Y_{ijk} - Y_{\cdots})^2$	$SS_T/(n-1)$

Remember: $n = IJK$

ANOVA with two factors

Test

$$H_0 : \{\alpha_1 = \cdots = \alpha_I = 0, \beta_1 = \cdots = \beta_J = 0, \gamma_{11} = \cdots = \gamma_{IJ} = 0, \}$$

vs.

$$H_1 : \{\exists(i,j) : \alpha_i \neq 0 \text{ or } \beta_j \neq 0 \text{ or } \gamma_{ij} \neq 0\}$$

- This is similar to a one factor analysis of variance with IJ levels.
- But there is more to it: can we decompose effect of A, B, and interaction?

Let us define decompose

$$S_E = SS_A + SS_B + SS_I$$

with

- $SS_E = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{...})^2$
- $SS_A = \sum_{i=1}^{I} n_{i+}(Y_{i..} - Y_{...})^2$
- $SS_B = \sum_{j=1}^{J} n_{+j}(Y_{.j.} - Y_{...})^2$
- $SS_I = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(Y_{ij.} - Y_{i..} - Y_{.j.} - Y_{...})^2$

ANOVA with two factors

```
> anova(lm(Ni ~ Rock + Landuse + Rock*Landuse,data=jura))
Analysis of Variance Table

Response: Ni
Df  Sum Sq Mean Sq F value    Pr(>F)
Rock            4  7738.0 1934.51 52.4628 < 2.2e-16 ***
Landuse         3  1116.3  372.12 10.0916 2.188e-06 ***
Rock:Landuse   10  2026.0  202.60  5.4944 1.394e-07 ***
Residuals     341 12574.0   36.87
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And we can test for each of this effect!

ANOVA with two factors: tests

Test for effect A:

$$H_0(A) = \{\alpha_1 = \cdots = \alpha_I = 0\}$$

The test statistics is

$$F_A = \frac{SS_A/(I-1)}{SS_R(n-IJ)} \sim \mathcal{F}_{(I-1),(n-IJ)}$$

and $n - IJ = IJ(K-1)$.

Test for interaction:

$$H_0(I) = \{\gamma_{ij} = 0,\ \forall (i,j) \in I \times J\}$$

The test statistics is

$$F_I = \frac{SS_I/(I-1)(J-1)}{SS_R(n-IJ)} \sim \mathcal{F}_{(I-1)(J-1),(n-IJ)}$$

ANOVA with two factors: unbalanced case

```
> jura.sel = jura[sample(1:359)[1:100],]
> anova(lm(Ni ~ Rock + Landuse + Rock*Landuse,data=jura.sel))
Analysis of Variance Table

Response: Ni
Df  Sum Sq Mean Sq F value    Pr(>F)
Rock          4 2187.05  546.76 14.8596 3.025e-09 ***
Landuse       3  230.38   76.79  2.0870   0.10795
Rock:Landuse  7  714.47  102.07  2.7739   0.01202 *
Residuals    85 3127.60   36.80
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> jura.sel = jura[sample(1:359)[1:100],]
> anova(lm(Ni ~ Rock + Landuse + Rock*Landuse,data=jura.sel))
Analysis of Variance Table

Response: Ni
Df Sum Sq Mean Sq F value    Pr(>F)
Rock          4 2046.9  511.72 12.8062 3.203e-08 ***
Landuse       3  234.7   78.24  1.9581  0.126327
Rock:Landuse  6 1148.2  191.37  4.7893  0.000294 ***
Residuals    86 3436.4   39.96
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Significances depend on the sample
- Conclusion might change if p-value close to threshold

ANOVA with two factors: unbalanced case

- When the design is unbalanced, the sum of squares do not sum up: we loose orthogonality and independence
- The test statistics F depend upon the order in which the factors are tested

Two methods are usually considered

- We fix an order, using priori information or expert knowledge. Results will depend on the order. Consider two very correlated factors, both significant. The second one will be considered as non significant.
- We can consider all orders, following the above set-up
- We consider each factor in turn, as the last factor. Here we lose the summation to SS_T. Attribution of fraction of variance is difficult. Also, two significant correlated factors will be considered as non-significant because always considered last.

Very carefull analysis when unbalanced design