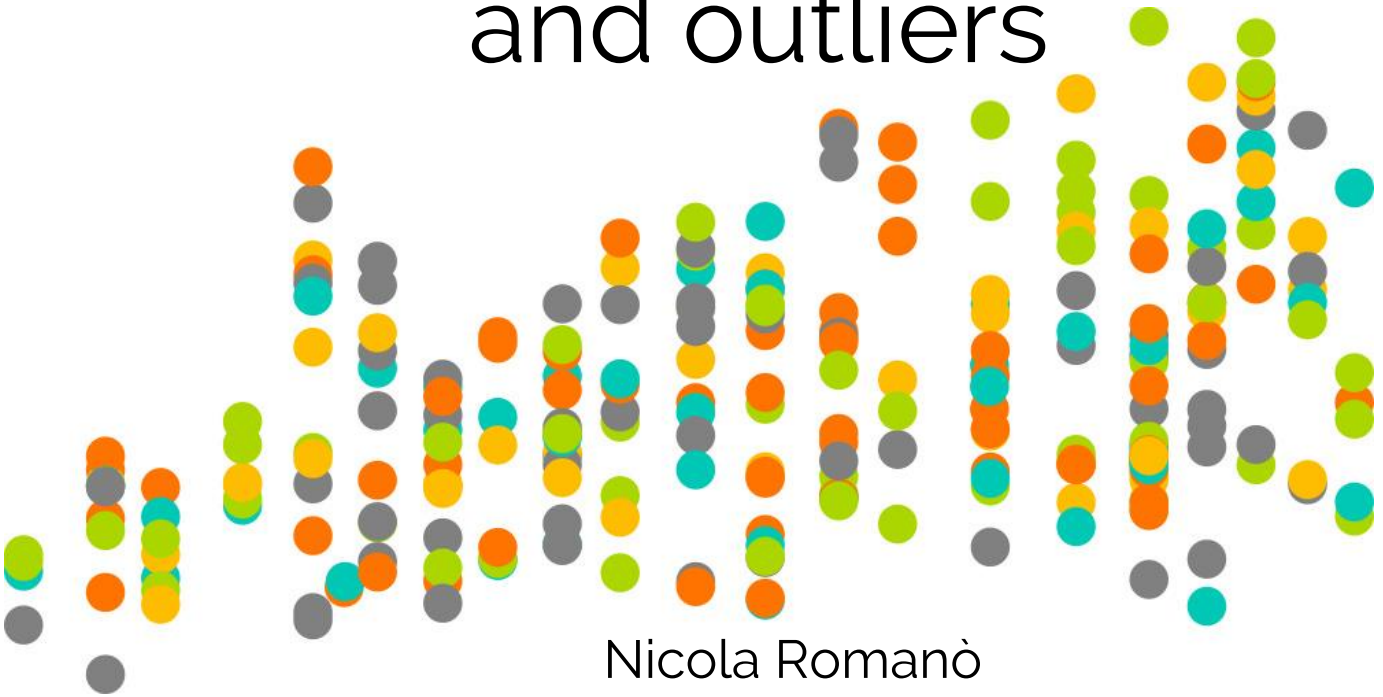
Experimental design #1
Data formats, missing data, and outliers



Nicola Romanò
nicola.romano@ed.ac.uk
11-09-2019

浙江大学爱丁堡大学联合学院
ZJU-UoE Institute

Learning objectives

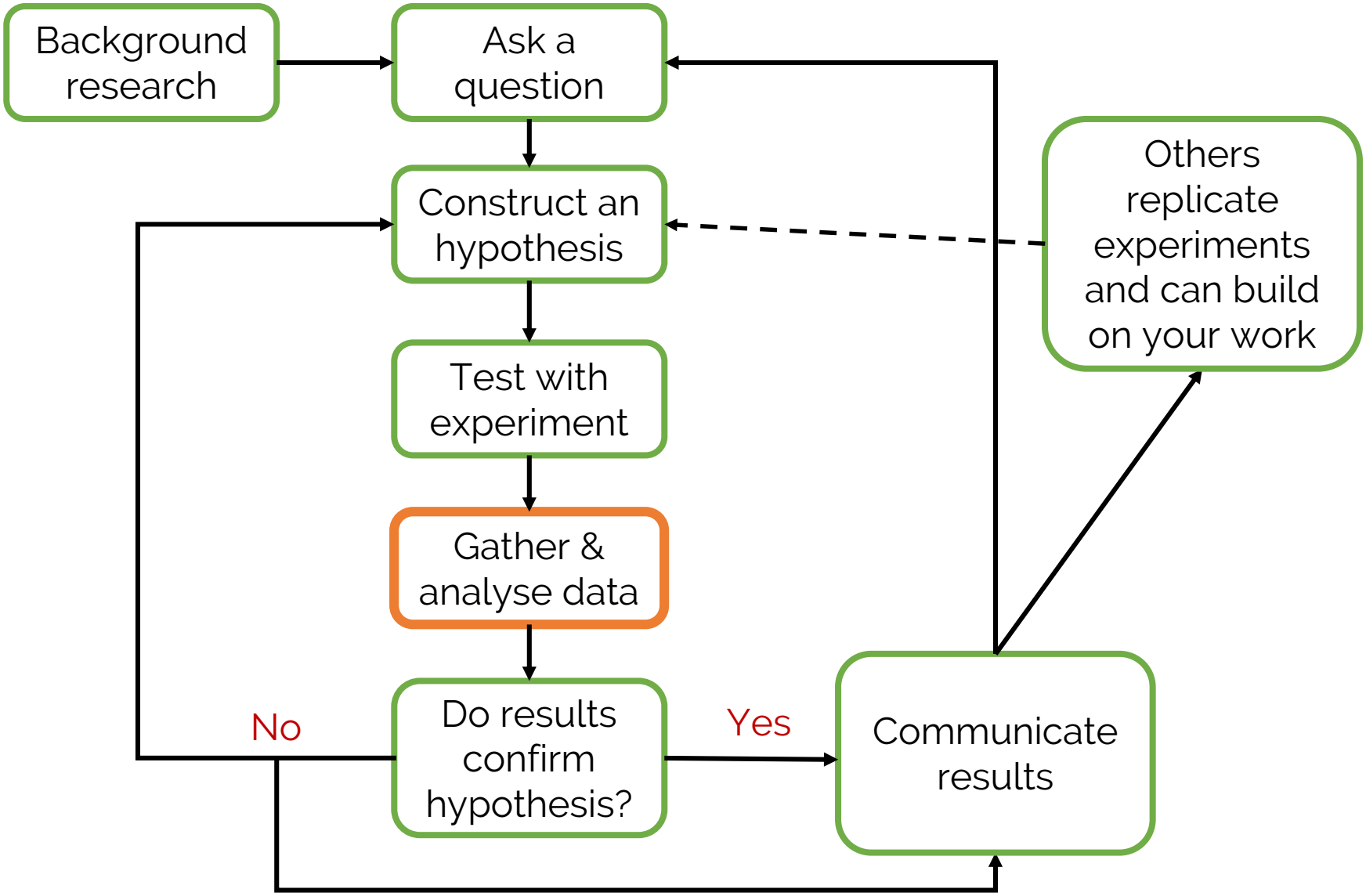
At the end of this lecture you will be able to:

- Describe the process of scientific method and explain the various steps involved

- Define the types of missing data and how to handle it

- Define what an outlier is and decide how to deal with it

The scientific method



Asking a question…

The very basis of experimental work.

- Background reading is important.
  - What do we already know about the subject?
  - What is missing?
  - What is unclear and why?

- Often different studies come to different conclusions.

- Things to be careful of:
  - “Generic press” articles are often imprecise and “amplify” the results of studies.
  - Reviews are good to get an idea of the field, but they may miss details. Always look at primary articles.
  - Beware of commercial interests… (although that does not make those studies false!)

Formulating an hypothesis

A good hypotheses needs to be **clearly defined, logic and testable** (or falsifiable).

Does not mean false!!!

Are these good hypotheses?

1. Ultraviolet light could cause cancer.
2. Activation of the cAMP pathway increases epithelial cell proliferation.
3. Red grapes taste better than white grapes.
4. When given the choice, mice prefer drinking water with added sucrose to plain water.
5. Rats can survive in the absence of oxygen.

Hypothesis-driven vs data-driven research

Recent advances in technology allow us to get a lot of data quickly (so called “big-data”).

This has created much debate on whether we still need hypotheses driven research

Hypothesis: mutation in gene X cause condition Y

↓

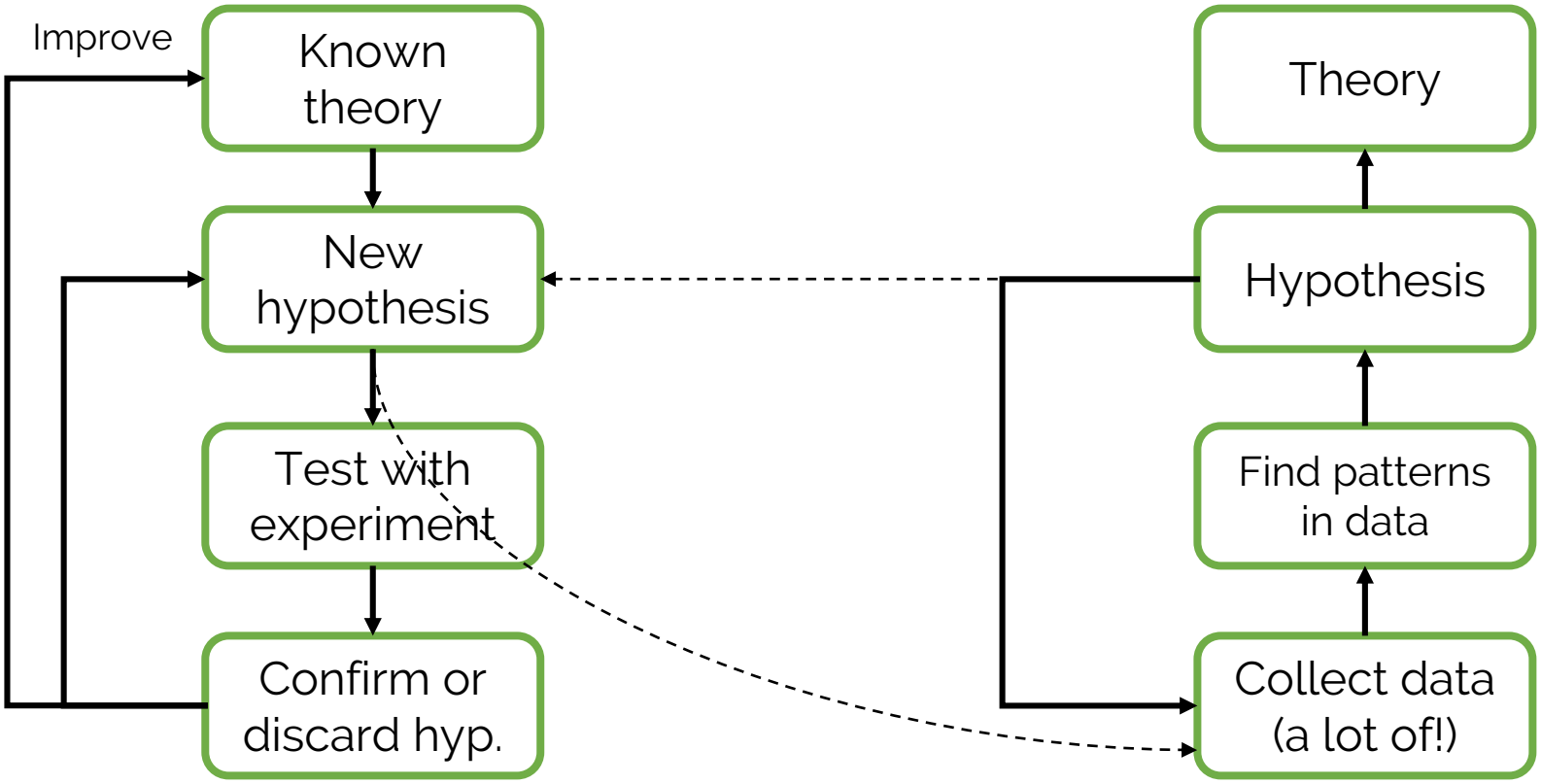
Experiments: - take people with Y and check their gene X
- make mice with mutant X and see if they develop Y
- …

But I could also:

- take cells from patients with Y or from a control group
- sequence their whole genome or transcriptome
- see what changes between the two
- formulate an hypothesis

Unbiased but more complex!

Deduction vs induction



Experimental design

Not always straightforward.

Many things to consider:

- What are the possible outcomes?
- What do they mean?
- Is the experiment reproducible?

Recording data...

11/7/2018

First trial on 6 rats.
Analysed data, M vs F statistically different

	Sex	Age	P1	P2 (adjusted)
145	M	38	28	2.39
200	F	42	33	10.91
133		41	11	2.40
115	M	50	34	6.27
122	M	28	44	1.03
138	F	55	22	1.41

Can you comment this record of data from a PhD student lab book?

Data format – wide format

Wide data

Subject	Monday	Tuesday	Wednesday	Thurdsday	Friday
1	36.6	36.6	37.3	36.9	36.8
2	36.1	36.2	36.1	36.3	36.4
3	37.0	37.1	36.9	36.8	36.9

One row per subject

One column for each level of the factor (e.g. day of the week)

The measured variable (body temperature) is in each cell

Data format – long format

Long (narrow) data

Subject	Temp	Day
1	36.6	Monday
1	36.6	Tuesday
1	37.3	Wednesday
1	36.9	Thursday
1	36.8	Friday
2	36.1	Monday
2	36.2	Tuesday
2	36.1	Wednesday
…	…	…

One row per measurement

One column for each factor or measured variable

Multiple rows per subject

Wide vs long format

Wide data is often easier to read for humans

Long data makes it easier to process data with software such as R

For example with the long data you can write:

```
boxplot(temp ~ day, tempdata)
```

Although it is possible to deal with wide data in R often it becomes very complex

Workshop #1 will show you how to convert wide data into long data
(and other things as well!)

Missing data

Ideal situation → all the data is collected and recorded

In real life some data points may be missing e.g.:

- Some data was not recorded by error (avoidable)
- Data (or part of) was lost (should be avoidable)

- Unavoidable:
    - A participant withdraws from the study during the course of an experiment
    - One subject dies before the end of the experiment
    - A participant doesn't answer an item in a questionnaire
    - Machine breaks during the course of the experiment

Missing data – what is the problem?

Reduce the statistical power of the study

Can produce biased results if data is not missing at random.

Makes it more difficult to analyse the data

Data missing completely at random (MCAR): the probability of missing the measurement of a variable is independent from the value of the variable itself or any other variable in the study

Data missing at random (MAR): the probability of missing is independent from the value of the variable itself but is related to some other variable.

Data missing not at random (MNAR): the probability of missing data depends on the value of the variable itself.

MCAR

Data missing completely at random is the less problematic case.

Example:

- Blood sample left at room temperature and damaged
- Questionnaire lost in the post

Reduces power of experiment
Does not introduce a particular bias

MAR

Data missing at random: unrelated to the value of what is measured but can depend on other variables

Example:

- Men may be less likely to complete a questionnaire about depression compared to women.

  In this case the missing data (about depression) depends on the variable "Sex" which is fully observed (i.e.: we know for each participant whether they are male or female).

- Decreases power of the experiment
- Can lead to misleading interpretation of results

MNAR

Data missing not at random: this is the most problematic case, and may indicate problems with the experimental design

Example:

- An athlete does not attend a drug test because he/she took drugs the night before

- Patients with severe depression are less likely to complete the questionnaire about depression.

  In this case the missing data will not be in a random subset of the sample.

- Results in that particular subset may be difficult to interpret, since they will probably not be representative of reality.

What to do? Omission

Remove all observations containing missing data points. This is referred to as "list omission". **It's probably the best option in many cases.**
Can be problematic if many missing points and/or few observations.

Handy function in R – *complete.cases*

Example

```
> mydata
   Sex Age       Value
1    F  51  0.54546745
2    F  37 -0.09543286
3    F  56 -1.79838831
4    M  36  1.59942971
5    M  50          NA
6    M  60  0.65504155
7    M  35 -0.40093197
8    F  48 -1.13259371
9    M  47  0.26713380
10   M  45  0.46326495
```

```
> mydata[complete.cases(mydata), ]
   Sex Age       Value
1    F  51  0.54546745
2    F  37 -0.09543286
3    F  56 -1.79838831
4    M  36  1.59942971
6    M  60  0.65504155
7    M  35 -0.40093197
8    F  48 -1.13259371
9    M  47  0.26713380
10   M  45  0.46326495
```

NA indicates a missing data point in R

Omission

Some R functions have specific parameters e.g. *na.rm*

```
> mydata

   Sex Age       Value
1    F  51  0.54546745
2    F  37 -0.09543286
3    F  56 -1.79838831
4    M  36  1.59942971
5    M  50          NA
6    M  60  0.65504155
7    M  35 -0.40093197
8    F  48 -1.13259371
9    M  47  0.26713380
10   M  45  0.46326495
```

```
> mean(mydata$Value)

[1] NA

> mean(mydata$Value, na.rm = TRUE)

[1] 0.0114434
```

Other functions omit missing values by default (e.g. *lm*)

```
> lm(Value ~ Age, mydata)

Call:
lm(formula = Value ~ Age, data = df, na.action = na.omit)

Coefficients:
(Intercept)          Age
    1.42537     -0.03066
```

The na.action parameter defines what to do

Observation #5 has not been used for this calculation

Pairwise omission

Some functions allow for “pairwise omission”. This means that the if the missing variable is not used in a certain calculation the observation is retained, otherwise discarded.

```
> mydata
       Var.1      Var.2    Var.3
1 -1.7033758 -0.1824386       NA
2  0.2632812  2.2787125 2.848954
3 -0.7896782  2.8033309 6.047446
4  1.1121877  3.1650596 7.388105
5 -0.7489375  0.8057609 2.971650
```

Note: specific syntax depends on the command. Read the help!

NA not omitted

```
> cor(mydata)
          Var.1     Var.2 Var.3
Var.1 1.0000000 0.8028439    NA
Var.2 0.8028439 1.0000000    NA
Var.3        NA        NA     1
```

Observation 1 is removed

```
> cor(mydata, use = "complete.obs")
          Var.1     Var.2     Var.3
Var.1 1.0000000 0.5903603 0.4447850
Var.2 0.5903603 1.0000000 0.8000003
Var.3 0.4447850 0.8000003 1.0000000
```

Observation 1 is removed only when calculating correlations involving Var.3

```
> cor(mydata, use = "pairwise.complete.obs")
          Var.1     Var.2     Var.3
Var.1 1.0000000 0.8028439 0.4447850
Var.2 0.8028439 1.0000000 0.8000003
Var.3 0.4447850 0.8000003 1.0000000
```

What to do? Imputation

Imputation allows you to “fill” the missing value with a likely value.

This sounds great in theory, but in practice can be very tricky

Various methods exist. Some examples:

- **Mean substitution** – substitute the missing value with the mean for that value in the sample. Implies variable is normally distributed in the population. Generally a bad idea, especially for small samples…

- **“Hot deck” substitution** – substitute with a value from a similar observation. Problem: how to choose the record? Can bias results

- **Regression imputation** – predict the likely value for the variable using some regression method from other variables measured in the study. Better, but can overestimate correlations.

- **Other more advanced statistical methods** – e.g. multiple imputations, beyond the scope of this course…

Problems with imputation

Simple imputation methods likely bias the analysis

Even with “fancy” methods, imputation only works well with large sample

Probably best to design the experiment so to maximise data collection, then resort to imputation

As with any statistical technique, it’s important to be aware of the pitfalls of the technique.

Outliers

What is an outlier?

An observation lying at an abnormal distance from other values in a random sample from a population.

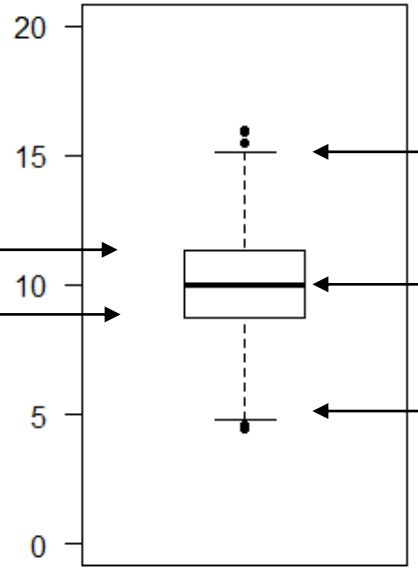
What is “abnormal distance” ?

Outliers – visual inspection

Many software packages mark as outliers those points that are outside of certain limits in the distribution.

Example R *boxplot*

```
boxplot(x, pch = 20, ylim = c(0, 20))
```



3^{rd} quartile (11.32)

1^{st} quartile (8.7)

IQR = 11.32-8.7 = 2.62

3^{rd} quart + 1.5 * IQR
(closest point)

Median (10.03)

1^{st} quart - 1.5 * IQR
(closest point)

Outliers – formal tests

There are formal tests to define outliers

Most commonly used:

Grubb's test

H_0: There are no outliers in the data set
H_A: There is exactly one outlier in the data set

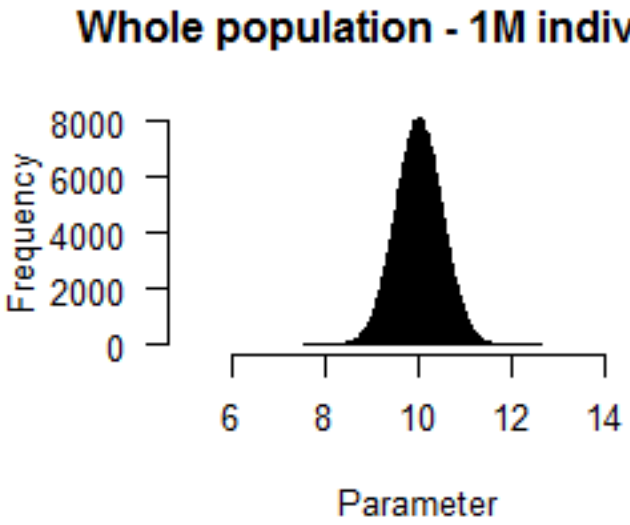
Grubb's test is based on the assumption of normality

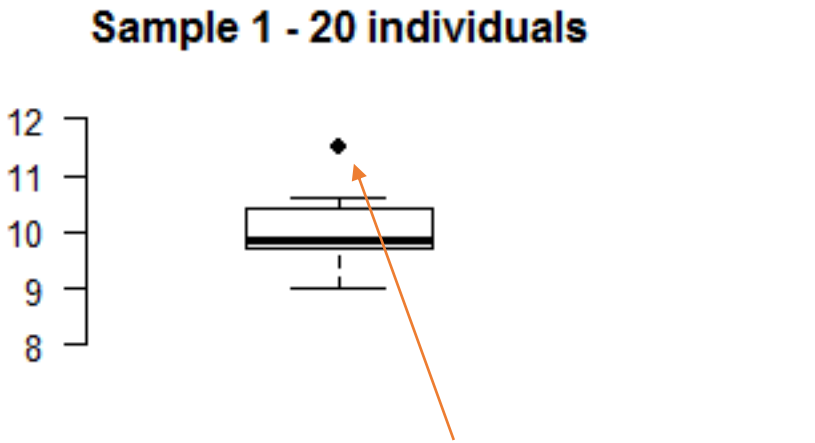
R function *grubbs.test* in package *outliers*.

```
> install.packages(outliers) # Need to do this only once!
> library(outliers) # The package needs to be loaded every time
> grubbs.test(x)

    Grubbs test for one outlier

data:  x
G = 3.43240, U = 0.87979, p-value = 0.02048
alternative hypothesis: highest value 3.65 is an outlier
```

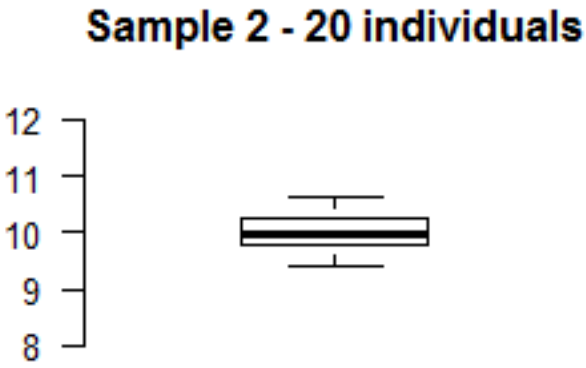
Outliers – what to do?

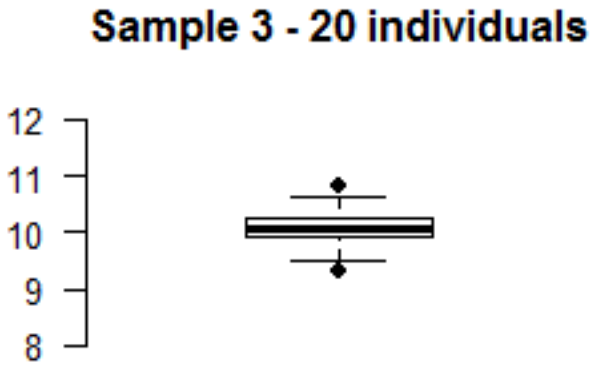
Should we remove outliers? **Generally speaking – not a good idea**





Grubb's test marks this as outlier





Outliers – what to do?

In some cases it is clear that a value has been mis-recorded.
In those cases you can safely remove it or in some instances correct it

Remove:

- Impossible values, such as a negative age

- A value which is 1000 times bigger than what expected

Correct (with caution):

- A value which has clearly been swapped with another one in a spreadsheet

THIS DOES NOT MEAN YOU CAN CORRECT A VALUE BECAUSE IT IS NOT THE RESULT THAT YOU WOULD LIKE TO HAVE!

The best solution is to leave the outlier be… and repeat the experiment more times!

Summary

Producing good results depends on good experimental design

This involves many steps: background research, formulating an hypothesis, collecting data and analyse/interpret them!

Each of these steps is critical

Issues in data collection:

- How to collect data?
- What format to save it?
- What analysis to run?
- Dealing with missing data
- Dealing with outliers