Introduction to Statistical Methods

Lecture 1: Basic Concepts & Descriptive Statistics

Theophanis Tsandilas theophanis.tsandilas@inria.fr

Course information

Web site: https://www.lri.fr/~fanis/courses/Stats2021

My email: theophanis.tsandilas@inria.fr

Slack workspace: stats-2021.slack.com

Course calendar

	Date (room)	Description
1	Nov 24 (0D19)	Basic concepts: data, populations, and samples. Why learning statistics? Types of data and descriptive statistics. Starting with R.
2	Dec 1 (4A312)	Discrete and continuous probability distributions: binomial, normal, and log-normal distributions. The sampling distribution of a statistic. The Central Limit Theorem.
3	Dec 8 (0D19)	Confidence intervals. Monte Carlo simulations. Experimental designs: independent groups, repeated measures.
4	Dec 15 (0D19)	Confidence intervals of non-normal distributions. Introduction to Null Hypothesis Significance Testing. Significance tests and p values. **Assignment handout.**
5	Jan 5 (0D19)	Significance tests: Type I and Type II errors. Statistical Power. The problem of multiple comparisons. Publication bias. p-hacking and criticisms of the NHST.
6	Jan 12 (0D19)	Covariance and correlation. Simple linear regression.
7	Jan 19 (0D19)	Final exam

Lecture overview

Part I

1. Basic concepts: data, populations & samples
2. Why learning statistics?
3. Course overview & teaching approach

Part II

4. Types of data
5. Basic descriptive statistics

Part III

6. Introduction to R

Part I. Basic concepts

Data vs. numbers

Numbers are abstract tokens or symbols used for counting or measuring[1]

Data can be numbers but represent « real world » entities

[1]Several definitions in these slides are borrowed from Baguley's book on *Serious Stats*

Example

Take a look at this set of numbers:

8 10 8 12 14 13 12 13

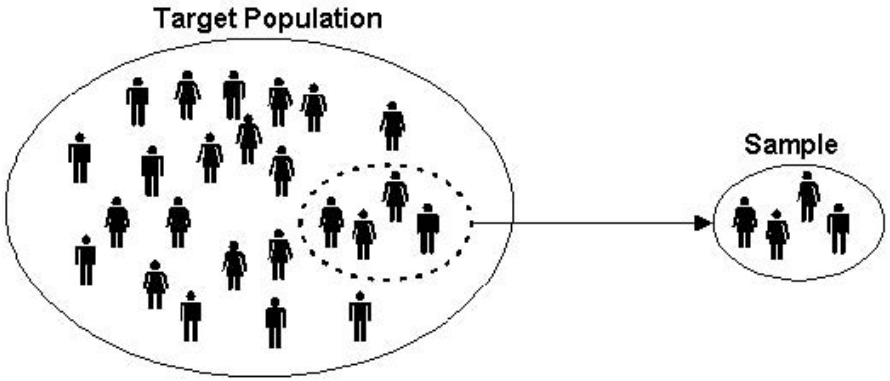
Context is important in understanding what they represent

- The age in years of 8 children
- The grades (out of 20) of 8 students
- The number of errors made by 8 participants in an experiment from a series of 100 experimental trials
- A memorization score of the same individual over repeated tests

Sampling

The context depends on the process that generated the data

- collecting a subset (**sample**) of observations from a larger set (**population**) of observations of interest



The idea of **taking a sample from a population** is central to understanding statistics and the heart of most statistical procedures.

Statistics

A sample, being a subset of the whole population, won't necessarily resemble it.

Thus, the information the sample provides about the population is **uncertain**.

The role of statistics is to find ways to **deal with this uncertainty**.

Probabilities

Uncertainty is commonly quantified and expressed in terms of **probabilities**

The probability of an event ***x*** can be written as ***P(x)*** or ***Pr(x)***

- It ranges from 0 (certain not to occur) to 1 (certain to occur)

A reasonable (but not the only) interpretation of a probability:

- as the relative frequency with which an event *x* occurs in the long run

Example: heads & tails

Consider a coin-tossing experiment with a fair coin.

- In the long run (e.g., after 1,000,000 trials), an (approximately) equal number of *Head* and *Tail* events are expected to be observed.
- Thus, *Pr(Head) = Pr(Tail) = 0.5*

Population & sampling procedure

The concept of a population is an abstraction

- rarely does it refer to a particular set of things, e.g., people
- a common assumption is that samples are drawn from an ***infinitely large, hypothetical*** population through a ***sampling procedure***

A well-designed study will use a sampling procedure that draws from a population **that is relevant to the aims of the research**

The sampling procedure is usually imperfect, e.g., due to bias into the sample

- a good study will minimize the impact of such problems

Example: heads & tails

A research team aims to estimate the probability of heads *Pr(Heads)* of real-world coins

- What is the population of interest?
- What are possible sources of bias in a sample?
- Describe a sampling procedure that minimizes the effect of such biases

Sample size

The **sample size *n*** is the number of observations (data points) in a sample

The larger the size *N* of a population, the less information (proportionately) a sample of size *n* provides

- Notice that if N is treated as infinite, then the size *n* of a sample is negligible.

How conclusions drawn from a sample generalize to the population of interest **depend on the adequacy of the sampling procedure in relation to the research goals**

Example

A researcher collects data from a **volunteer** or **opportunity sample** of 100 healthy people in Paris.

- Is this sample adequate?

Understanding the research domain is important!

- might be adequate if the goal is to assess the impact of caffeine on reaction time
- is inadequate if the goal is to assess the side-effects of a new substance
- or assess the average French family income

Independent vs. dependent observations

Observations in a sample can be assumed as **independent** if information about each observation provides no information about other observations.

Two observations are **dependent** (also **related**) if they are somehow connected

- nutrition habits of members of the same family
- repeated memory tests taken by the same individual

Example: heads & tails

Are observations on the occurrence of *Heads* or *Tails* independent and under which assumptions?

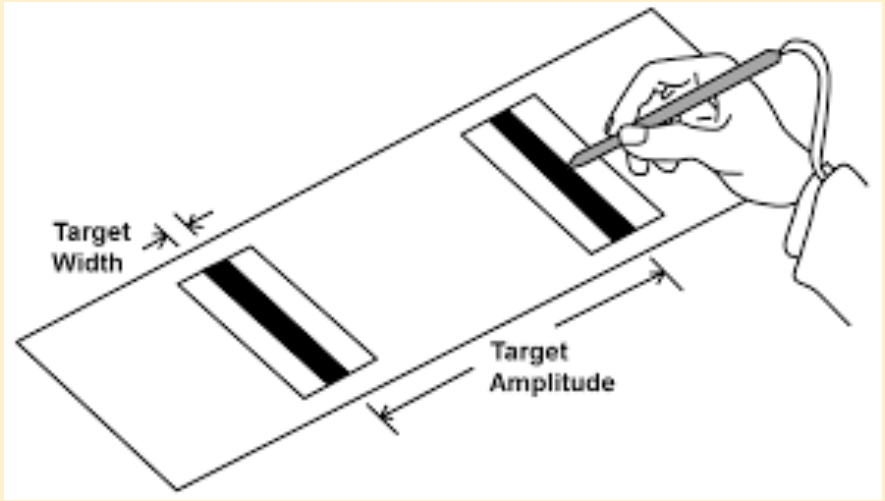
Statistical modeling

Samples are somehow related to the population from which their are drawn

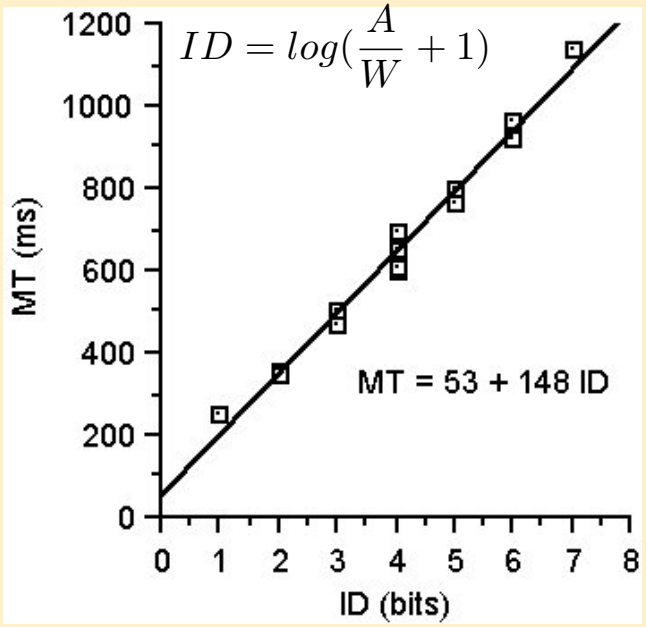
The goal of **statistical modeling** is to understand the process that generated the observed data and predict new observations

Example: Fitts' law experiments

A researcher is interested in creating a model that predicts how fast (*MT*) on average humans hit targets of varying widths *W* from varying distances (amplitudes) *A*



Experimental setup



Creating a statistical model

Statistical inference

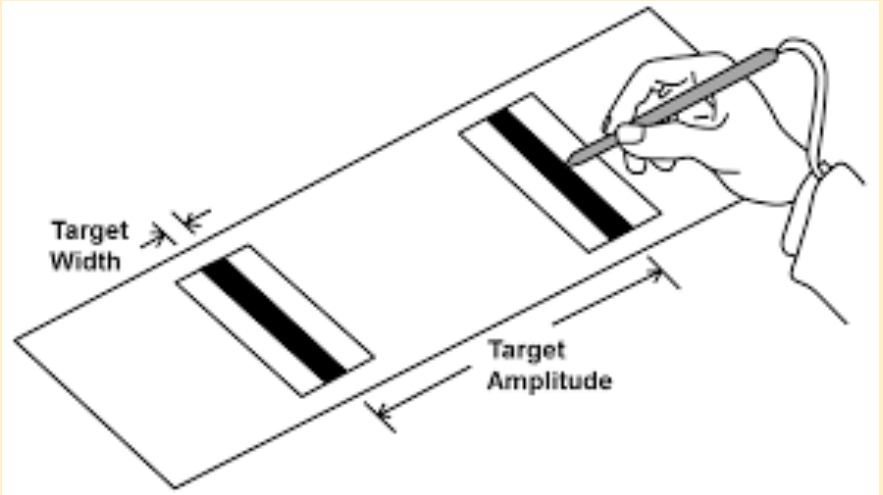
Statistical inference is a special case of statistical modeling, where the primary (or only) purpose of the model is to test a specific **hypothesis.**

Example hypotheses:

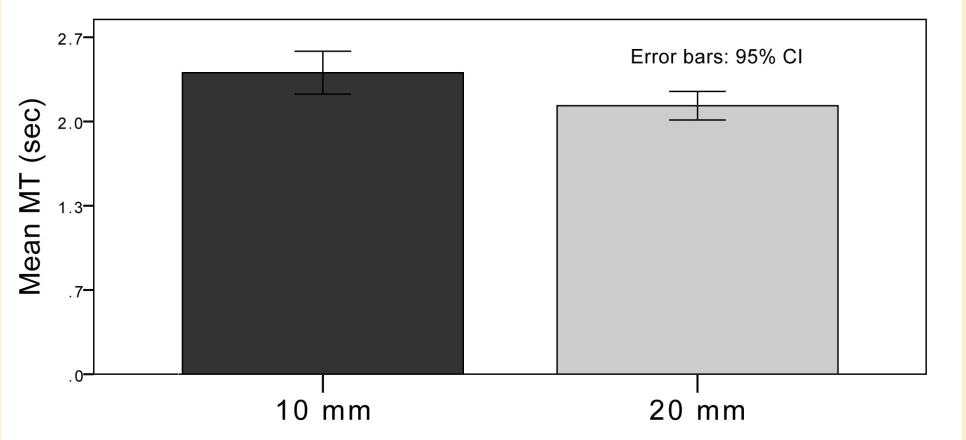
- Men are taller than women.
- Access to higher education positively affects income.
- Reading on paper leads to better memorization than reading on a tablet.

Example: Fitts' law experiments

The researcher is interested in testing whether wider targets are faster to hit.



Experimental setup



Observed data

Based on the statistical evidence, the researcher concludes that the hypothesis is true (with some level of quantified uncertainty): ***wider targets are faster to hit***.

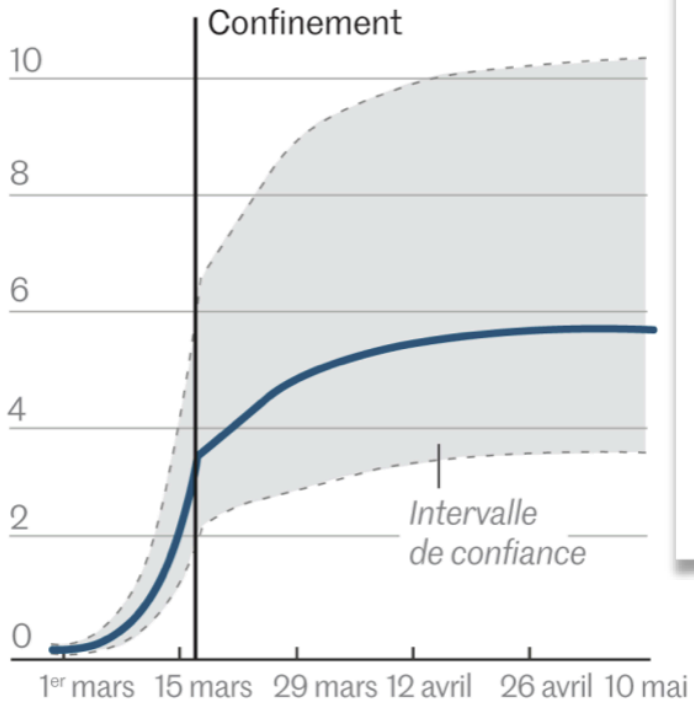
How informative is this result?

Why learning statistics?

5,7 % de la population seraient infectés par le Covid-19 au 11 mai

Part de la population qui **pourrait être infectée** par le Covid-19 à la date du 11 mai en France métropolitaine, en %

Dans la population totale



Comments by readers

Mickaël R. 14/05/2020 - 03H29

9,9 % (marge de 6,6 à 15,7 %) des habitants d'Ile-de-France auraient été contaminés au 11 mai et 9,1 % (marge 6,0 à 14,6 %) "
A ce niveau ce ne sont plus des marges, mais des abymes.
La marge comptent pour plus d'un tiers du pourcentage calculé.

Masquer la réponse Répondre Signaler un abus

Eric B 14/05/2020 - 11H40

Au moins ces incertitudes sont évaluées proprement. Avez-vous des données qui permettraient de réduire cette incertitude? J'imagine que non...

Le Monde, 21/4/2020

Why learning statistics?

Many claims and beliefs but also prejudices and stereotypes are founded on informal inferential statistics

…often based on inadequate or biased sampling

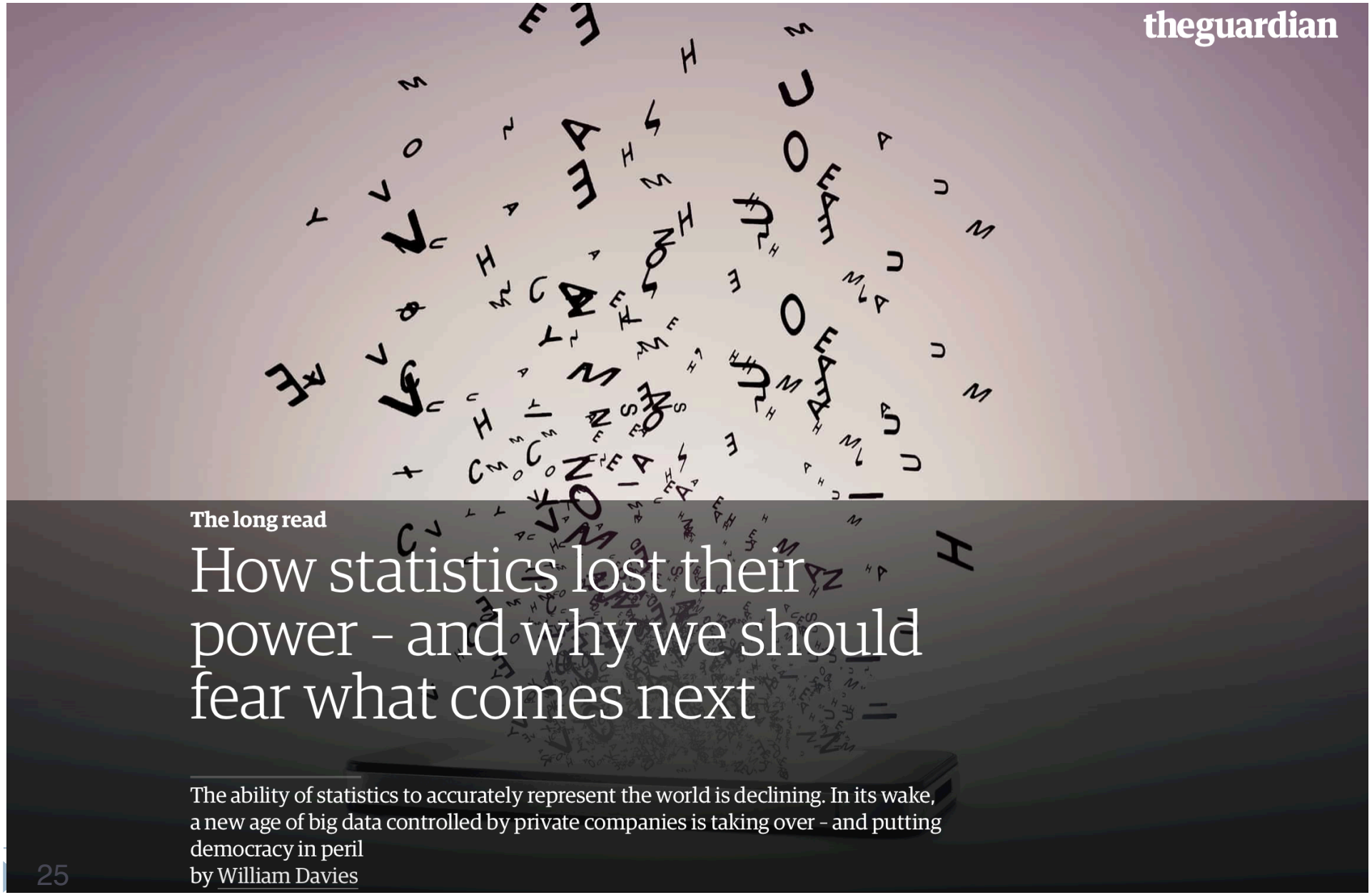
…based on incomplete or incorrect models

…that often fail to distinguish between correlation and causality

Why learning statistics?

Decision making and politics often rely or are justified based on true (or false) statistical evidence

Why learning statistics?



Why learning statistics?

> Yet in recent years, divergent levels of trust in statistics has become one of the key schisms that have opened up in western liberal democracies. Shortly before the November presidential election, a study in the US discovered that 68% of Trump supporters distrusted the economic data published by the federal government. In the UK, a research project by Cambridge University and YouGov looking at conspiracy theories discovered that 55% of the population believes that the government "is hiding the truth about the number of immigrants living here".

> Antipathy to statistics has become one of the hallmarks of the populist right, with statisticians and economists chief among the various "experts" that were ostensibly rejected by voters in 2016. Not only are statistics viewed by many as untrustworthy, there appears to be something almost insulting or arrogant about them.

https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

Why learning statistics?

Statistics is a fundamental research tool for many scientific disciplines

…but even in research, statistics are very often misused

Open access, freely available online

Essay https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.002012

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to
relationsh
field. In th
is less like
conducte
effect size
greater nu
of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of
is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity,
either there
(among
zed) or
any of the
ships. The
lationship
probability
tionship
reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV.

Ioannidis's 2005 paper has been the most downloaded paper in the PLoS Medicine journal.

nature International weekly journal of science

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

Brian Nosek's team set out to replicate scores of

https://doi.org/10.1136/bmj.323.7327.1450

Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial

BMJ 2001 ; 323 doi: https://doi.org/10.1136/bmj.323.7327.1450 (Published 22 December 2001)

Cite this as: *BMJ* 2001;323:1450

Leonard Leibovici (leibovic@post.tau.ac.il), *professor*

Author affiliations

Department of Medicine, Beilinson Campus, Rabin Medical Center, Petah-Tiqva 49100, Israel

https://doi.org/10.1136/bmj.323.7327.1450

Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: randomised controlled trial

BMJ 2001 ; 323 doi: https://doi.org/10.1136/bmj.323.7327.1450 (Published 22 December 2001)

Subjects: All 3393 adult patients whose bloodstream infection was detected at the hospital in 1990-6.

Intervention: In July 2000 patients were randomised to a control group and an intervention group. A remote, retroactive intercessory prayer was said for the well being and full recovery of the intervention group.

Main outcome measures: Mortality in hospital, length of stay in hospital, and duration of fever.

Results: Mortality was 28.1% (475/1691) in the intervention group and 30.2% (514/1702) in the control group (P for difference=0.4). Length of stay in hospital and duration of fever were significantly shorter in the intervention group than in the control group (P=0.01 and P=0.04, respectively).

Conclusions: Remote, retroactive intercessory prayer said for a group is associated with a shorter stay in hospital and shorter duration of fever in patients with a bloodstream infection and should be considered for use in clinical practice.

How could they arrive to such conclusions?

Statistics in Computer Science

Machine learning, data mining, computer networks, information visualization, human-computer interaction, etc.

The course focuses on experimental design and the analysis of experimental data

Course overview

Goals

Understand core concepts and methods of statistical reasoning

Understand a range of statistical methods of practical interest

Learn how to work with real-world (and « messy ») data

Familiarize with the R statistical software

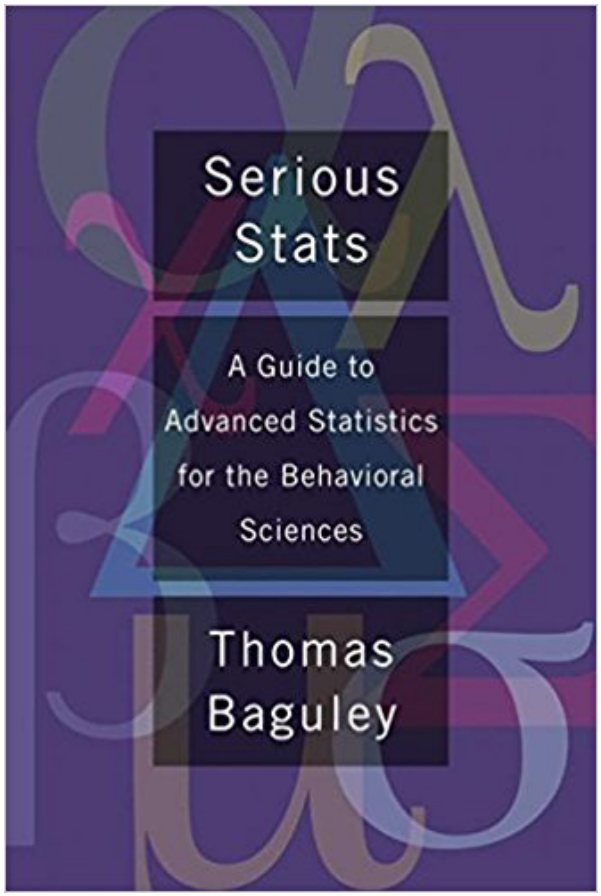
Approach

Focus on the scope, assumptions, uses, and limitations of each statistical method.

Avoid complex mathematical models. It is important to understand the intuition behind the mathematics.

Computers are very helpful! We will rely on computational methods when analytical methods cannot help.

Approach


Serious Stats
A Guide to Advanced Statistics for the Behavioral Sciences
Thomas Baguley

Approach

« *Statistical modeling* ***is not a set of recipes or instructions****. It is the search for a model or set of models that capture the regularities and uncertainties in data, and help us to understand what is going on.* »

[Baguley]

Types of data

Discrete vs. continuous data

Discrete data are restricted in the values that can legitimately occur. Examples:

- **Binary** data can take on only two possible values (e.g., 0 or 1)
- **Frequency** or **count** data are used to count things (e.g., 5 heads vs. 7 tails)

Continuous data can take on intermediate values within a range. Examples:

- Physical measures such as time and distance can take on any value from *0* to ∞
- The difference between two times or two distances can range between $-\infty$ to ∞

Scales of measurement (Stevens 1946;1951)

Nominal or **categorical**. They can be represented by numbers but their assignment is arbitrary

- Color, participant identifier, profession

Ordinal. Preserve information about the relative (not absolute) magnitude of what is measured

- Ranking of a collection of items, age group (child, teenager, adult)

Interval. Preserve continuous, linear relationships between what is measured. **Distances between subsequent points on the scale are equal.** But the presence of zero is arbitrary.

- Temperature in degrees Celsius

Ratio. Interval that have a 'true' zero

- Temperature in degrees Kelvin, weight in kilograms,
- Response time in seconds

Scales of measurement (Stevens 1946;1951)

Athens: 20 degrees Celsius

Paris: 10 degrees Celsius

Can we say that it is twice as hot in Athens than in Paris?

Notice:

Athens: 68 degrees Fahrenheit

Paris: 50 degrees Fahrenheit

Other **interval data** examples:

IQ scores, time in 12-hour format, dates (2019)

Scales of measurement (Stevens 1946;1951)

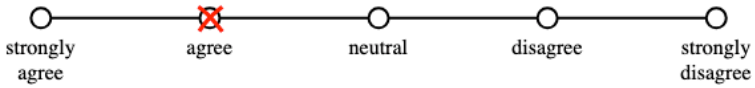
Examples of ordinal scales

B1. Which of the following video-communication tools do you currently use or have you used in the past? *

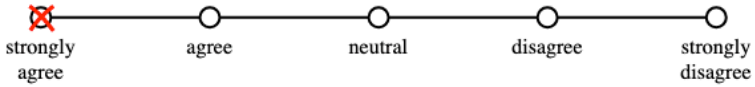
Une seule réponse possible par ligne.

	never	rarely	occasionally	frequently
Skype	◯	◯	◯	◯
Zoom	◯	◯	◯	◯
Jitsi	◯	◯	◯	◯
Google Meet	◯	◯	◯	◯
Blackboard	◯	◯	◯	◯
Discord	◯	◯	◯	◯
Teams	◯	◯	◯	◯

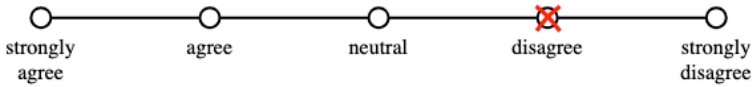
1. The website has a user friendly interface.



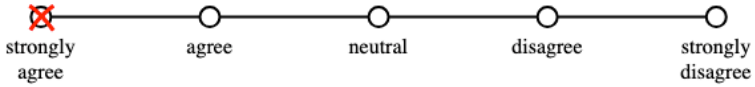
2. The website is easy to navigate.



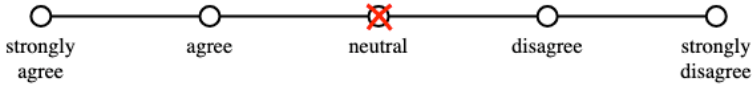
3. The website's pages generally have good images.



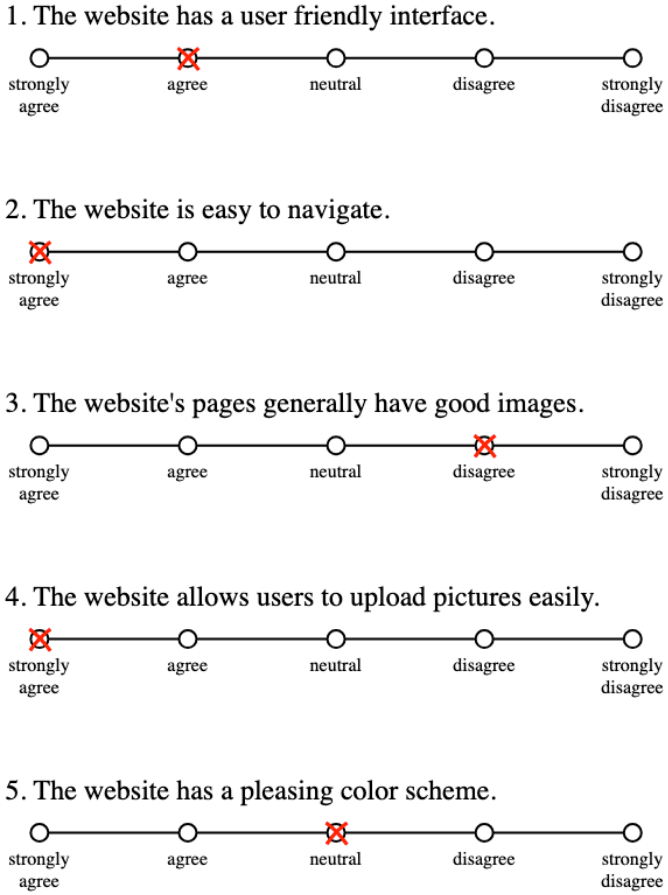
4. The website allows users to upload pictures easily.



5. The website has a pleasing color scheme.



Scales of measurement (Stevens 1946;1951)



One may decide to assign numerical values:

2 for *strongly agree*

1 for *agree*

0 for *neutral*

-1 for *disagree*

-2 for *strongly disagree*

…then treat the data as ratio or interval.

Problem:

Intervals between subsequent values may not be perceived as equal. For example, the perceived difference between *agree* and *strongly agree* may be much larger than the difference between *agree* and *neutral*.

Scales of measurement (Stevens 1946;1951)

Scales limit the mathematical operations that are permitted on data of a given type:

- Nominal data are limited to operations such as counting
- Ordinal data are limited to operations such as ranking
- Interval data also permit addition and subtraction (but not multiplication or division)
- Ratio scales permit the full range of arithmetic operation and allow for ratios between numbers (10/5 = 2 implies than 10 is twice as large than five)

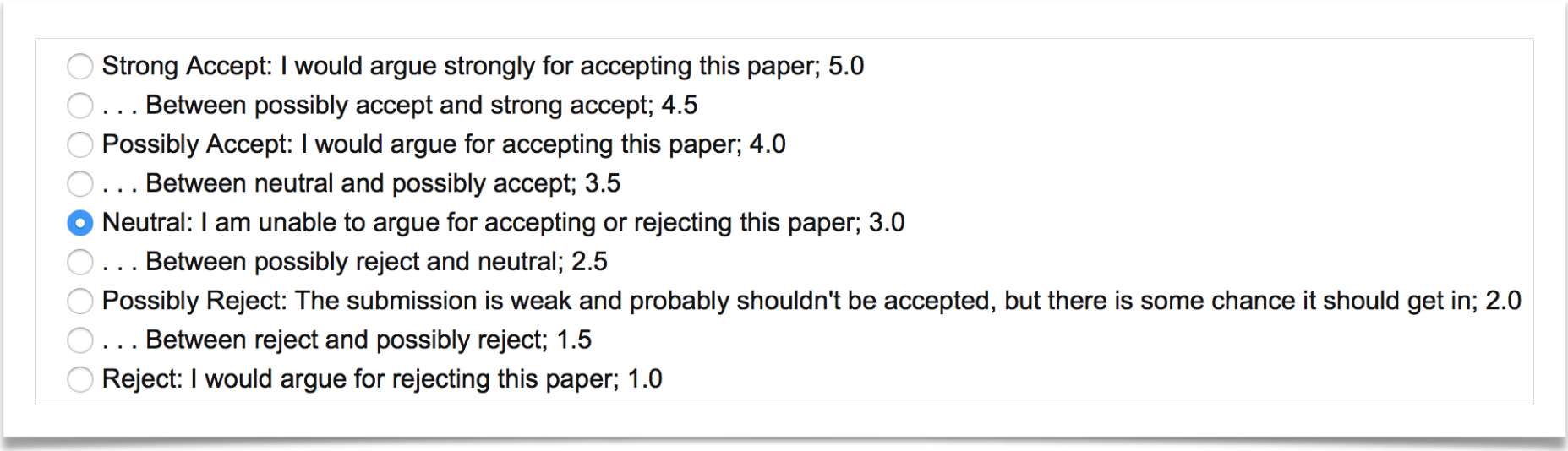
Scales of measurement are widely used.

However, they are controversial.

https://chi2018.acm.org/blog/

Review scores

The mean of the mean scores given to papers for the 2018 conference was 2.56 (SD=0.74). (Yeah, we know we took means of ordinal data here, but luckily there is no R2 of this blog post!). In total, 1356 papers (53%) received



Critiques of scales of measurements

Understanding the context of the data is important

- Classification schemes (such as scales of measurements) inevitably loose information about the context.

« the single unifying argument against proscribing statistics based on scale type is that it does not work » [Velleman and Wilkinson, 1993]

Alternative approach advocated by Baguley:

Consider a range of factors of data that impact on the statistical model. For example:

- Are data continuous or discrete?
- What is the probability distribution being assumed?
- What is the sample size?
- etc.

Descriptive statistics

Descriptive (or summary) statistics

Common descriptive statistics: *minimum (min)*, *maximum (max)*, *mean, median, standard deviation, etc.*

Excellent starting point of most statistical analyses

- A good way to summarize and communicate information about a dataset
- « Get a feel » for a data set
- Sometimes confirm some clear patterns in the data
- Identify irregularities and problems in the data collection process
- Guide the selection of an appropriate statistical model

Measures of central tendency

How to describe a data set with a single number?

- e.g., by means of a « typical », the « most common » or an « average »

Common measures of central tendency:

- mode (the most common value)
- median (the central or middle value)
- mean (or average)

The mode, median or mean of a sample will most often differ from those of the population being sampled

Parameters vs. statistics

A **parameter** is a property of the population

A **statistic** is a property of a sample. It provides a way to **estimate** a population parameter.

- As the size *n* of the sample approaches the size *N* of the population, sample statistics tend to resemble population parameters

One convention is to use a Greek letter for the population parameter and a Latin letter for the sample statistic

- e.g., μ designates the population mean and *M* designates the sample mean

Parameters vs. statistics

A **parameter** is a property of the population

A **statistic** is a property of a sample. It provides a way to **estimate** a population parameter.

- As the size *n* of the sample approaches the size *N* of the population, sample statistics tend to resemble population parameters

One convention is to use a Greek letter for the population parameter and a Latin letter for the sample statistic

- e.g., μ designates the population mean and *M* designates the sample mean

Another convention is to use the same symbol but differentiate a **sample estimate** by the « hat » symbol (ˆ), e.g., $\hat{\mu}$ designates the mean estimate given from the sample

Mode

The most common value

- The mode of the following data set is 10:

16 **10** 12 **10** 9 14 13

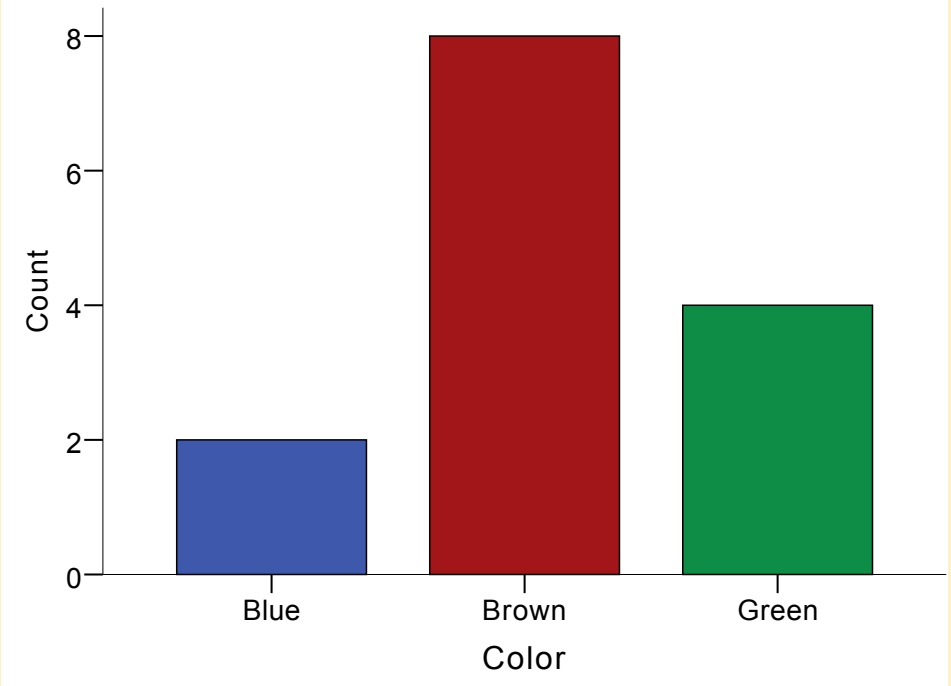
- A data set may have a single (unimodal) or multiple modes (multimodal)

The mode could be the best value to guess

- This works best for discrete rather than continuous data, where the mode may be far from typical
- Especially appropriate for categorical data, where there is no order relationship

Example

Consider the following plot that shows the responses of 14 students to the question: *What color eyes do you have?*



What's the mode of the sample?

What's the best color guess when picking a random student?

Median

The central value in a set of numbers

- If numbers are placed in order, the median is the middle value:

12 20 24 **34** 35 80 83

If the sample size is even, the convention is to take the mid-point between the two central values as the median

- Consider the following ordered set of values:

5 6 **8** **9** 12 15

Its median is (8+9)/2 = 8.5

The median is insensitive to extreme values

- Pros: Eliminates the effect of extremes
- Cons: Ignores vital information about non-central values

Arithmetic mean

It is the most widely used measure of central tendency. It is well-known as mean or average

- The mean of the following data set:

  16 10 12 10 9 14 13

  is *(16+10+12+10+9+14+13) / 7 = 12*

Generic formulation of the mean of a set of numbers x_i, where *i=1,2..n*:

$$M = \frac{\sum_{i=1}^{n} x_i}{n}$$

Unlike the mode or the median, the mean uses all the *n* numbers in its calculation

Example

What's the best measure of central tendency for each of the following?

1. Weight of 50 randomly sampled individuals
2. Income of French families
3. Housing expenses for 100 university students, where expenses are classified into three ranges:
   (a) lower than $300
   (b) between $300 and $600
   (c) higher than $600

Trimmed mean

A trimmed mean is a measure of central tendency designed to reduce the influence of extreme values

- This is achieved by **discarding the smallest and largest *k* numbers**:

$$M_{trim} = \frac{\sum_{i=k+1}^{n-k} x_{(i)}}{n - 2k}$$

where the notation $x_{(i)}$ indicates that values have been ordered from highest to lowest

A trimmed mean is a compromise between the mean and the median!

Example

Consider the following sample:

12 10 8 12 13 18 5 37 15

Let's reorder it:

5 8 10 12 12 13 15 18 37

Median = 12

Mean = 14.4

Trimmed Mean$_{(k = 2)}$ = 12.4

Measures of dispersion

Compare these two datasets:

D_1: 12 13 13 14 15 14

D_2: 5 9 12 15 20 20

They have identical means and medians but are very different

- The numbers in D_2 are more spread out. They have greater **dispersion.**

Measures of dispersion:

range, quartiles & quantiles, variance, standard deviation, etc.

Range

It is the difference between the *maximum* and the *minimum* values of the sample

D_1:	12	13	13	14	15	14
D_2:	5	9	12	15	20	20

The range of D_1 is: 15 - 12 = 3

The range of D_2 is: 20 - 5 = 15

The measure is very vulnerable to extreme values

Quartiles

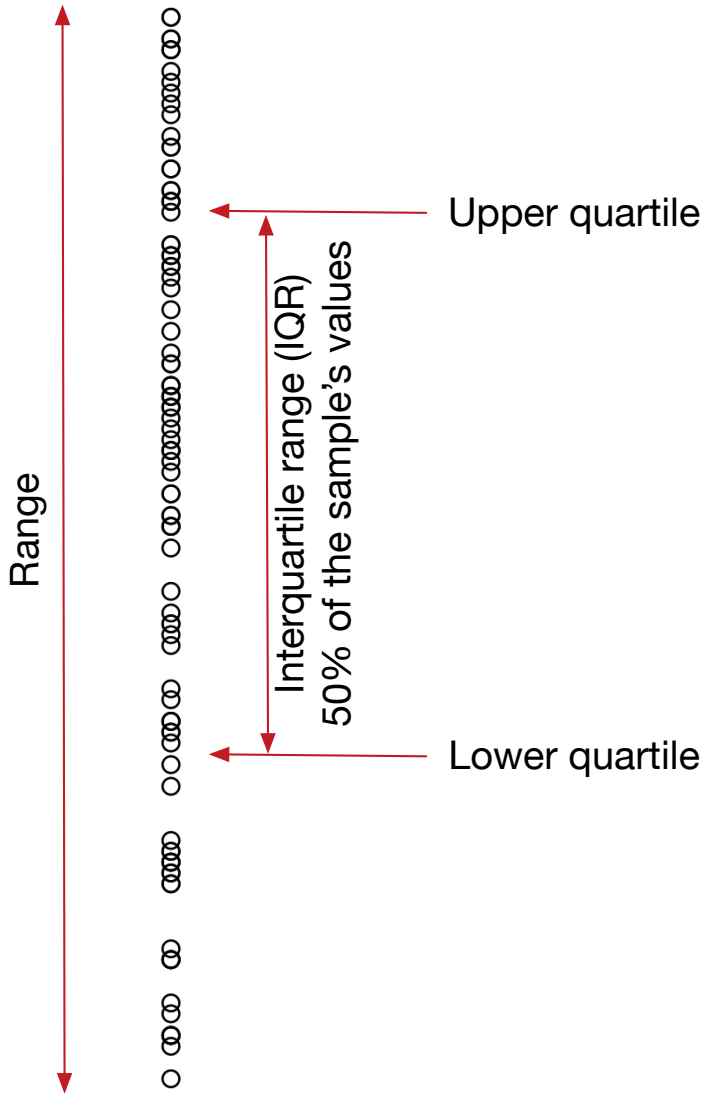
What about computing the range on a trimmed sample?

- This will better describe the spread of less extreme, more central values

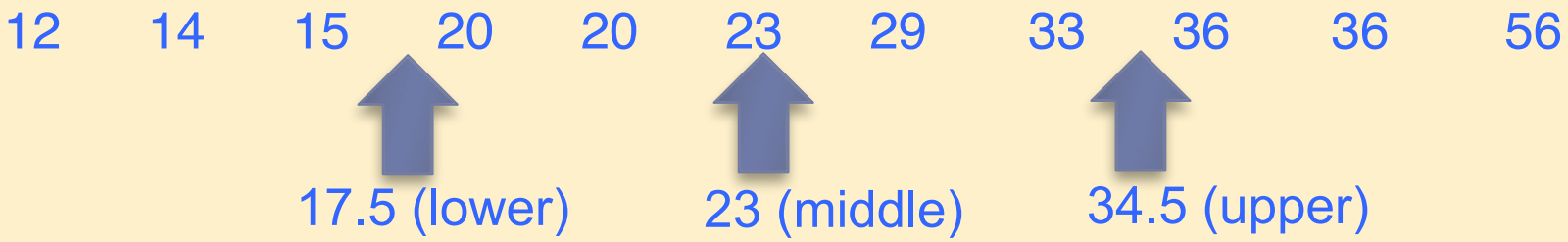
Quartiles are the **three points** that separate a set of *n* ordered numbers **into four equal subsets**

- The first (lower) quartile separates the lowest 25% numbers
- The second (middle) quartile is the median
- The third (upper) quartile separates the highest 25% numbers

Quartiles


Range
Interquartile range (IQR)
50% of the sample's values
Upper quartile
Lower quartile

Example: quartiles of small samples



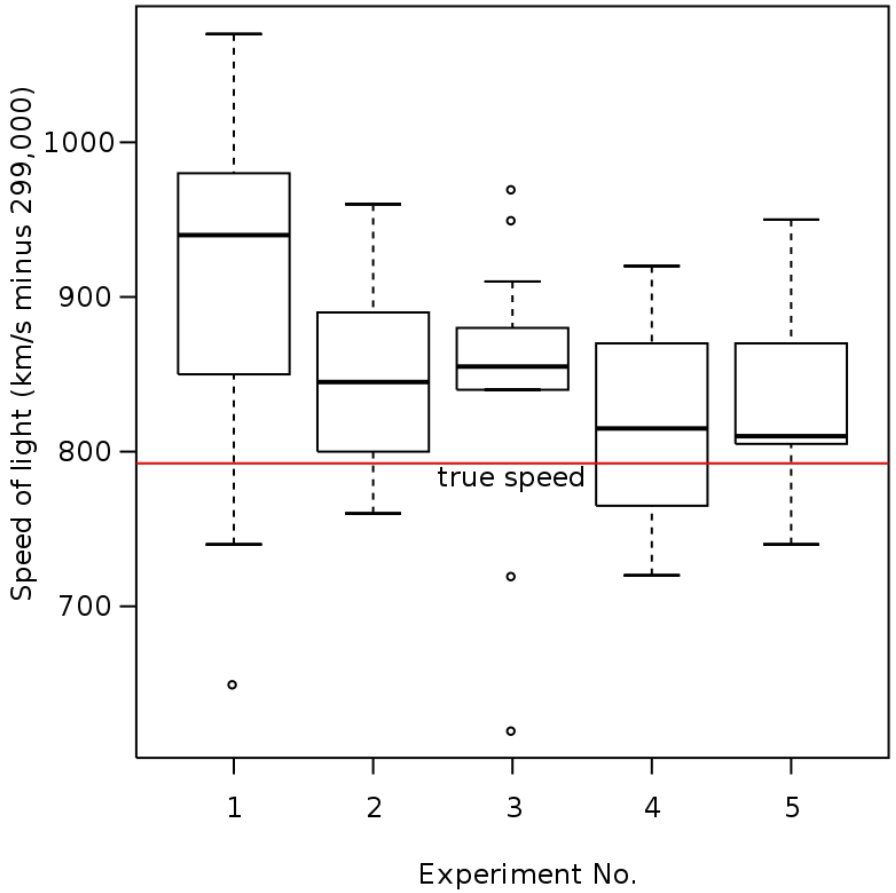
IQR (Interquartile range) = 34.5 – 17.5 = 17

The above calculations (from R) may seem awkward. Other statistical software (e.g., SPSS) may give a different result.

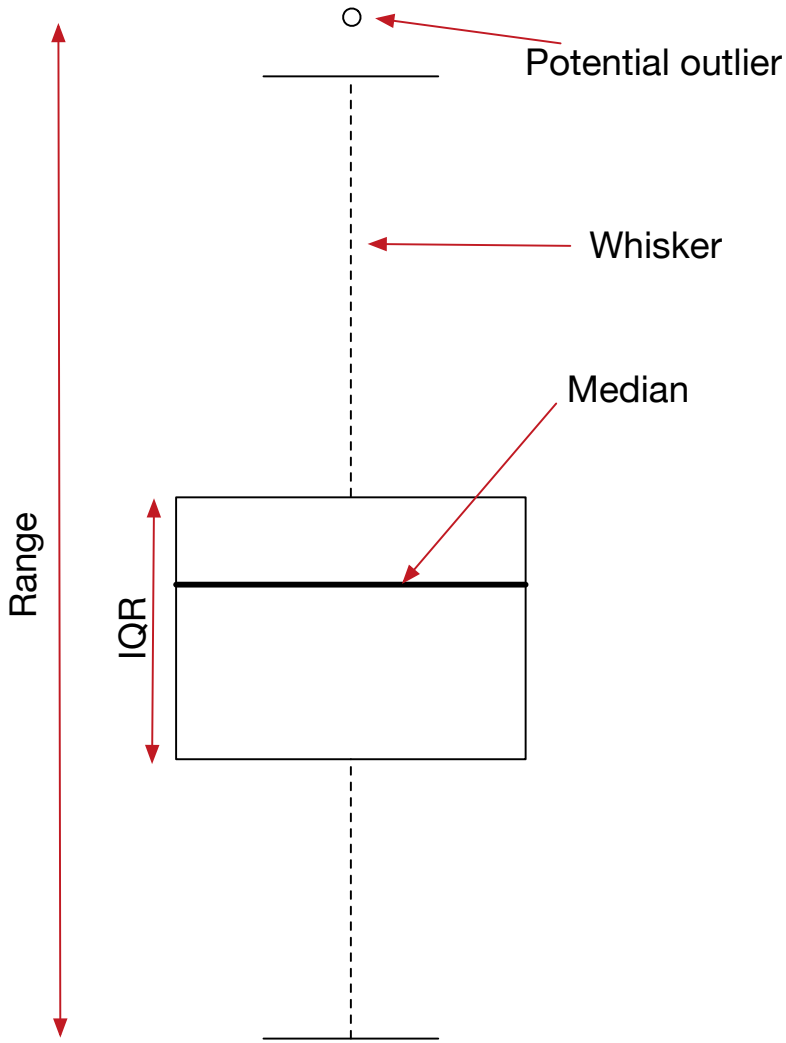
- There are different approaches for calculating quartiles of small samples

Boxplots

An example of multiple boxplots comparing measured speed of light for five different experiments (source: Wikipedia)



Boxplot



The maximum size of a whisker is usually 1.5 x IQR (it depends on the software)

Quantiles

The points that divide a set of numbers into *q* subsets of equal size

- Quartiles are a special case of quantiles, where q = 4

Another common choice is the **centile** (or **percentile**) where *q = 100*

- The 25th percentile is the first (lower) quartile
- The 50th percentile is the median
- The 75th percentile is the third (upper) quartile

Variance and standard deviation

Variance

sum of squares

$$Var = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n}$$

sample size

Standard deviation (scaled to use the same units as the original data)

$$SD = \sqrt{Var} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}}$$

SD is possibly the most common measure of dispersion

Example

Consider the following dataset that gives the weight of six 15-month old babies (in kilograms):

8 10 10 12 9 11

$$M = \frac{(8+10+10+12+9+11)}{6} = 10$$

$$Var = \frac{(8-10)^2+(10-10)^2+(10-10)^2+(12-10)^2+(9-10)^2+(11-10)^2}{6} = 1.667$$

$$SD = \sqrt{1.667} = 1.29$$

Example

Consider the following dataset that gives the weight of six 15-month old babies (in kilograms):

8 10 10 12 9 11

$$M = \frac{(8+10+10+12+9+11)}{6} = 10$$

$$Var = \frac{(8-10)^2+(10-10)^2+(10-10)^2+(12-10)^2+(9-10)^2+(11-10)^2}{6} = 1.667$$

$$SD = \sqrt{1.667} = 1.29$$

One can expect that the **typical** weight of a 15-month old baby is **roughly** 10 ± 1.29 kilograms

Residuals

The raw deviations from the mean $x_i - M$ are called **residuals**

For the following dataset (M = 10)

8 10 10 12 9 11

The residuals are as follows:

-2 0 0 2 -1 1

Exercise

Consider an experiment measuring the reaction time (in ms) of 6 participants over 5 repeated trials.

	T1	T2	T3	T4	T5
P1:	137	180	156	130	126
P2:	130	122	110	124	122
P3:	140	128	124	110	112
P4:	133	144	123	121	130
P5:	122	118	117	115	116
P6:	155	148	128	142	137

How would you proceed to calculate and report the mean, the median, and the standard deviation of this sample?

Estimators of population parameters

How do we estimate a population's mean, variance or standard deviation from a sample?

For example:

- Is a sample's mean a good estimator of the population's mean?
- Is a sample's standard deviation a good estimator of the population's standard deviation?

Efficient & unbiased estimators

A good statistic should be an **efficient** and **unbiased** estimator of the relevant population parameter

An efficient statistic has less error

- tends to be close to the population parameter
- fluctuates less from sample to sample

An unbiased statistic has no bias

- in the long run, it does not (consistently) overestimate neither underestimate the true population parameter

Unbiased vs. biased estimators

Sample statistics of central tendency such as means, medians, and trimmed means are unbiased estimators

Thus, we will often use $\hat{\mu}$ to designate the mean of the sample (*M*), as well as the sample's estimate of the population mean (μ)

...but **statistics of dispersion are biased**

- They tend to underestimate the true population parameter
- A small sample is unlikely to capture the extremes of a population

Estimation of population variance & SD

The population variance is usually represented as σ^2 and its unbiased estimator is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$$

$n-1$ degrees of freedom

The unbiased estimator of the population standard deviation σ is:

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}}$$

$n-1$ degrees of freedom

Degrees of freedom

degrees of freedom = the number of parameters in the calculation of a statistic that are free to vary

n-1 are the degrees of freedom of the residuals $x_i - \hat{\mu}$

(or the degrees of freedom of the parameter estimates)

Given $\hat{\mu}$ we only need n-1 independent observations to calculate the variance (or the standard deviation)

$$x_n = \hat{\mu} - (x_1 + x_2 + ... + x_{n-1})$$

Descriptive vs. inferential statistics

The unbiased estimates of a population variance (or standard deviation) are known as **inferential variance** or **inferential standard deviation**

Descriptive statistics simply describe the sample.

With **inferential statistics,** we try to infer the population parameters from a sample.

Comment on measures of dispersion

Why do common measures of dispersion (variance and standard deviation) use **sums of squares**

$$\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

instead of **sums of absolute residuals**?

$$\sum_{i=1}^{n}|x_i - \hat{\mu}|$$

Comment on measures of dispersion

Working with absolute values can be difficult, but this is not the main reason.

The measure of central tendency that minimizes the sums of absolute differences **is the median, not the mean.**

And since the mean is the prevalent measure of central tendency, we commonly use sums of squares.

However, for statistical methods that rely on medians, sums of absolute differences can be more appropriate.

Introduction to R

Why R?

R is an open source programming language and software environment for statistical computing and graphics

It is available for most computing platforms (Windows, Mac OS, Linux)

There is a wide range of statistical packages written in R. They cover nearly everything you may need for your statistical analyses.

R is widely used by the research community

Drawbacks

Challenging to learn for many, when compared to commercial statistical software with rich user interfaces, such as SPSS, Statistica, and JMP

R supports several related data types (lists, vectors, matrixes, frames). It is easy to forget how to pass from one data type to another.

Generating a good graph can be quite laborious, e.g., figuring out which parameters to specify

R vs. Python

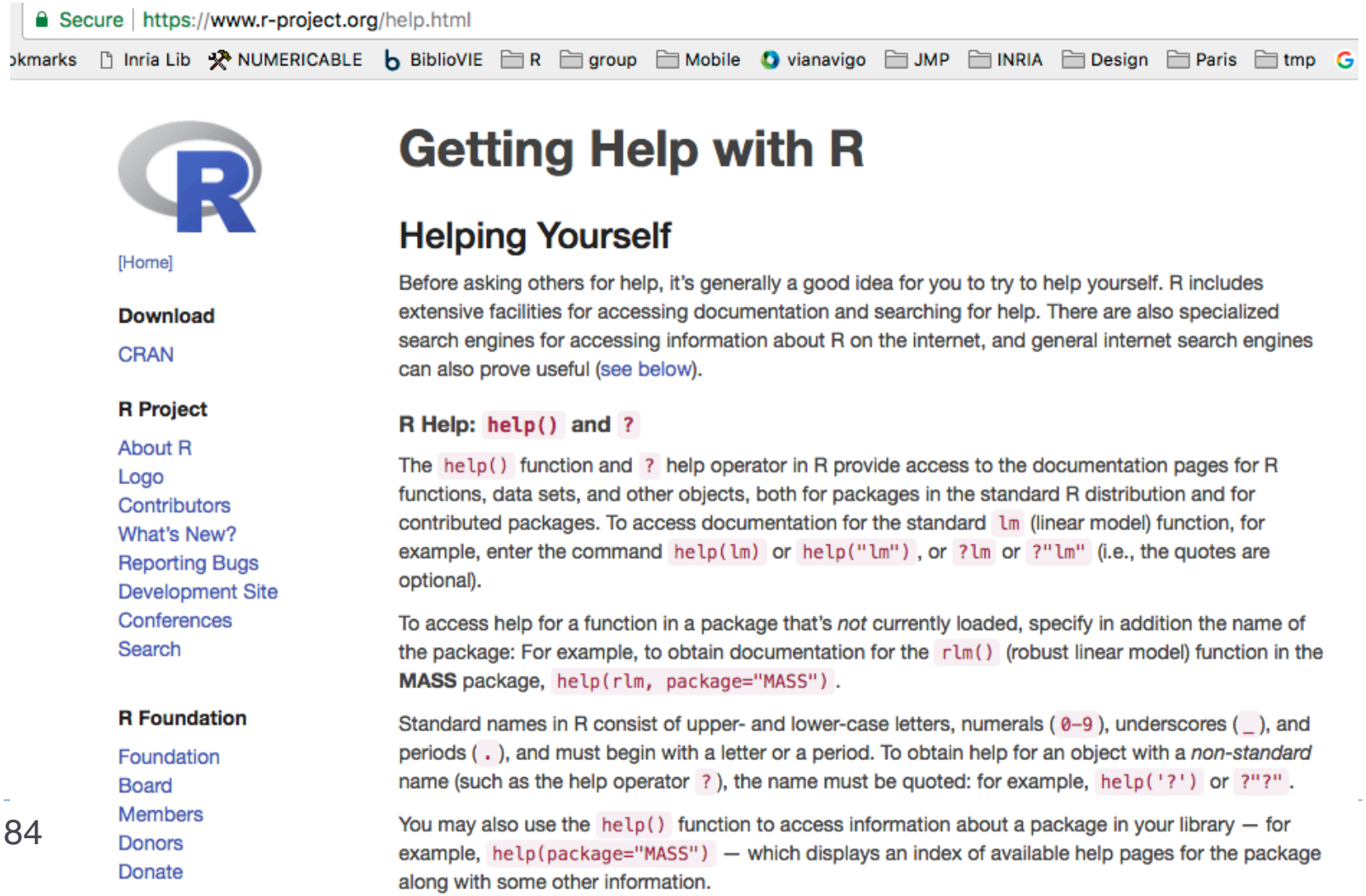
Python is a general-purpose language that is well-supported with libraries for statistical analysis.

For a detailed comparison:

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Getting help

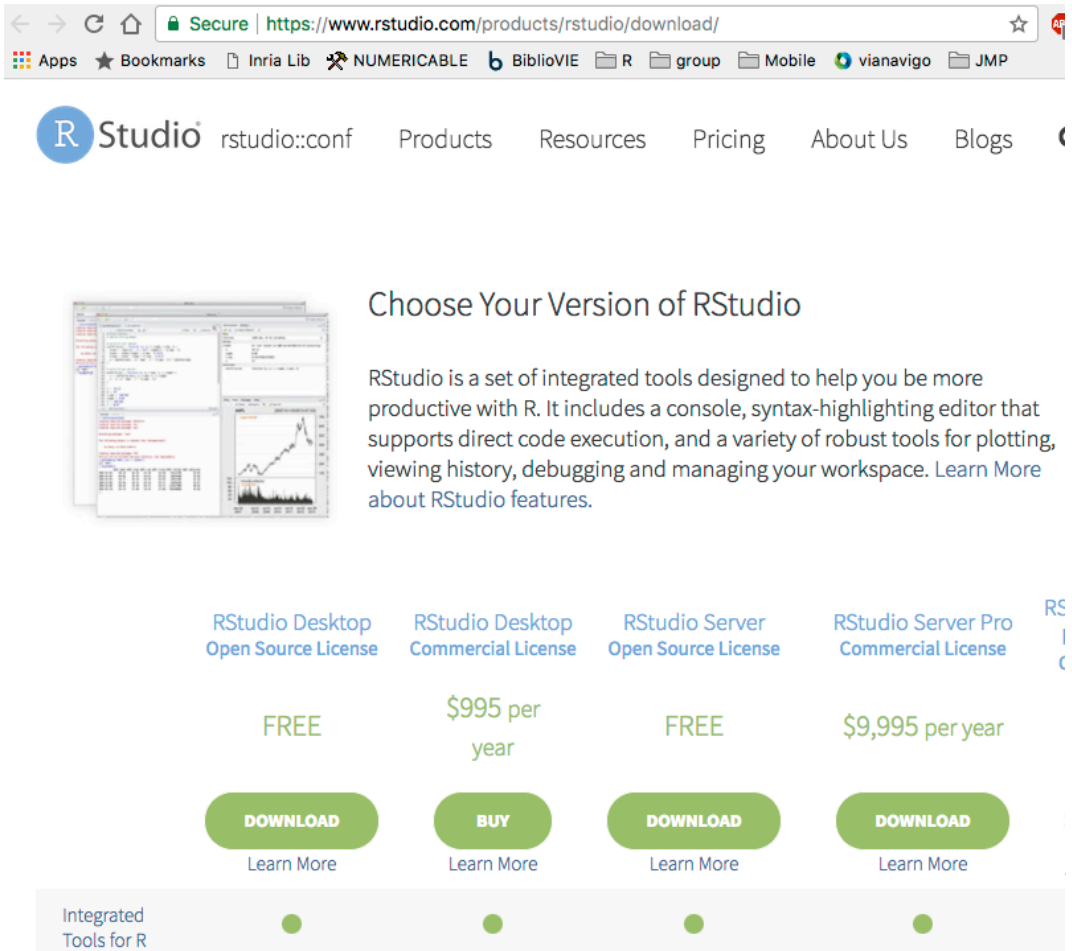
Fortunately, there are plenty of online resources that can help you find quick solutions



Getting started

Go to https://cran.r-project.org/ and download R for your platform

RStudio (not required)



Vectors

```
> val <- 2
> my.vector <- c(2,5,1)
> my.vector
[1] 2 5 1
> my.vector <- c(val, 3, c(8,8,8))
> my.vector
[1] 2 3 8 8 8
> rm(my.vector)
> my.vector
Error: object 'my.vector' not found
```

Arithmetic

```
> 3+4
[1] 7
> val <- 4 * 6^2
> val
[1] 144
> num.vect <- c(2,4,5)
> num.vect + 10 + num.vect*2
[1] 16 22 25
```

Measures of central tendency

```
> data <- c(10, 12, 8, 9, 11, 20)
> median(data)
[1] 10.5
> mean(data)
[1] 11.66667
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   8.00    9.25   10.50   11.67   11.75   20.00 
> 
```

Measures of dispersion

```
> population <- 1:1000
> sample <- sample(population, 15)
> sample
 [1] 380  32 144 705 971 326 974 632 396  22 203 295
[13] 114  44 733
> range(sample)
[1]  22 974
> IQR(sample)
[1] 539.5
> sd(sample)
[1] 329.6428
> sd(population)
[1] 288.8194
> mean(population)
[1] 500.5
> mean(sample)
[1] 398.0667
```


Unbiased estimator

Measures of dispersion

Biased estimator

```
> (sum((sample - mean(sample))^2)/length(sample))^.5
[1] 318.4652
```

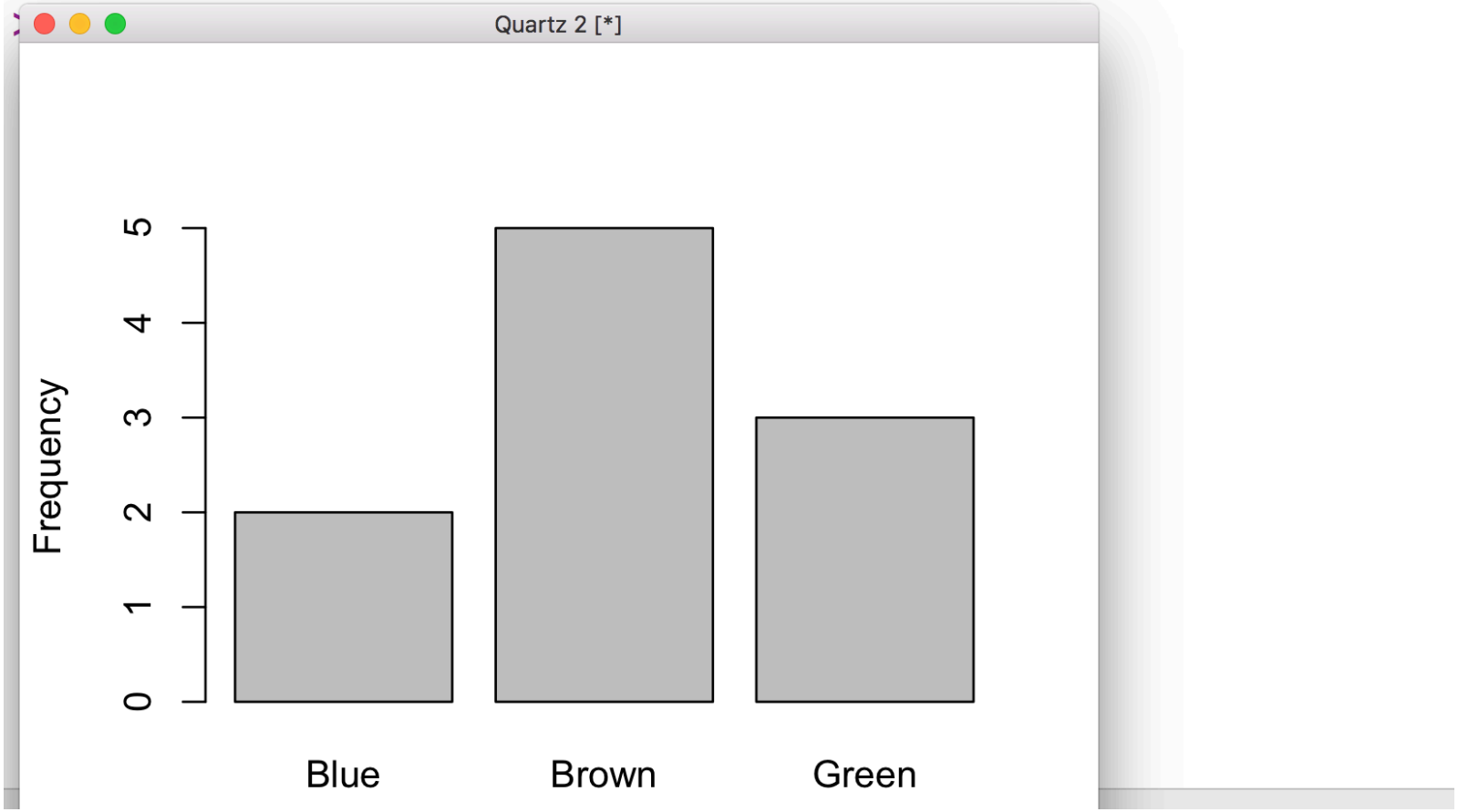
Minimizing the sum of absolute residuals

```
> data <- c(12, 13, 20, 16, 30, 7, 29)
> sum(abs(data - mean(data)))
[1] 49.14286
> sum(abs(data - median(data)))
[1] 47
```

```
> data <- c(12, 13, 20, 16, 30, 7, 29)
> sum((data - mean(data))^2)
[1] 454.8571
> sum((data - median(data))^2)
[1] 487
```

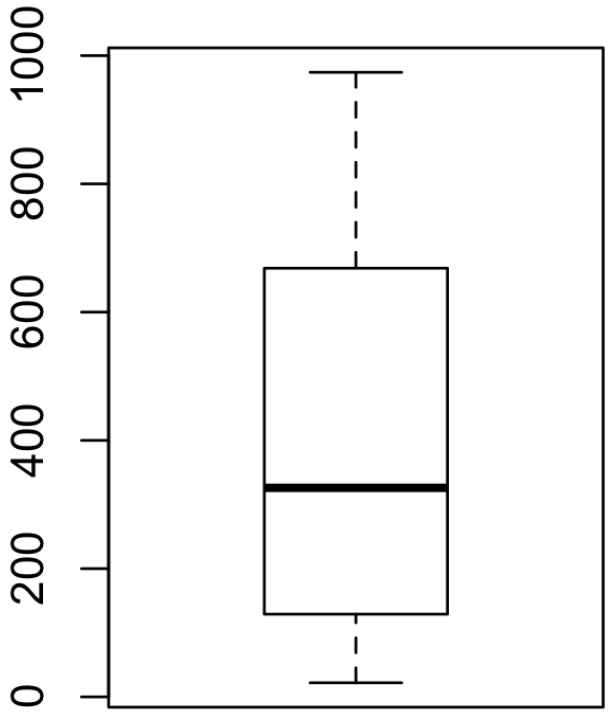
Plotting

```
> eyes <- c(2, 5, 3)
> labels <- c("Blue", "Brown", "Green")
> barplot(eyes, names.arg = labels, ylab = "Frequency")
```



Plotting

```
> boxplot(sample)
>
```



Creating larger programs - version 1

```
population <- 1:10000

samplesNum <- 10 # Number of samples
sampleSize <- 20 # Size of each sample

# Create a matrix of samplesNum x sampleSize
samples <- matrix(, nrow = samplesNum, ncol = sampleSize)

# Repetitively create samples
for(i in 1:samplesNum){
    samples[i,] <- sample(population, sampleSize)
}

# Transpose the samples matrix
samplesTrans <- t(samples)

# And plot it
boxplot(samplesTrans)
```

Creating larger programs - version 2

```
population <- 1:10000

samplesNum <- 10 # Number of samples
sampleSize <- 20 # Size of each sample

# Produce a matrix of samples
samples <- replicate(samplesNum, sample(population, sampleSize))

# And plot it
boxplot(samples)
```

Creating larger programs - version 3

```
population <- 1:10000

# Define a function that creates a matrix of random samples
createRandomSamples <- function(pop, num = 10, size = 10){
    replicate(num, sample(population, size))
}

boxplot(createRandomSamples(population, size = 20))
```