Collaboration in Higher Education for Digital Transformation in European Business

Brief Guide for Data Analysis

(Working paper)

Karel Doubravský, doubravsky@fbm.vutbr.cz & **Jan Luhan**, luhan@fbm.vutbr.cz, Brno University of Technology/Czech Republic;
Luca Alfieri, luca.alfieri@ut.ee, Tartu University/Estonia;
Margareta Teodorescu, margareta.teodorescu@fh-bielefeld.de
University of Applied Sciences Bielefeld/Germany

This working paper is a result of the EU Erasmus Plus project "Collaboration in Higher Education for Digital Transformation of Corporate Businesses" (CHEDTEB) and essential ideas come from the intensive exchange with IT experts and company managers of the working groups "Big Data" and "Blockchain technology". Further information about Big Data, algorithms, AI and Blockchain technology can be found on our project webpage.

http://www.chedteb.eu/

The authors are grateful for the lively discussions within various Big Data workshops and would like to thank in particular following colleagues for their valuable ideas, and suggestions:
Jan Budik, Martin Fridrich, Lukáš Novák – all of Brno University of Technology/Czech Republic,
Viire Täks, Sherif Sakr both from Tartu University/Estonia,
Rainer Lenz, Bernd Kleinheyer, Tahir Lushi - all of University of Applied Sciences Bielefeld/Germany

Co-funded by the Erasmus+ Programme of the European Union

PREFACE

We live in a digital society where most sectors undergo the process of digital transformation. Almost every business faces numerous challenges as to how to be more effective, profitable, eco-friendly, secure and reliable. To be able to face these challenges and succeed in such a turbulent and changing world, an effective work with data seems to be a crucial precondition for surviving in upcoming digital era. The key task, which erases for managers in this context, is how to adapt and prepare businesses towards this situation.

Moreover, in this brief guide we try to enhance your readiness and understanding the issue of vast area of data analysis (including big data analysis) and link relevant topics with recommendations of appropriate didactical methods.

Our primary goal was to briefly introduce concepts of data science, statistics and machine learning as selected areas for building a better understanding of data analysis.

Authors' attention was paid to the introduction of handling cross-sectional and time series data as well as the basic concepts of data science, statistics and machine learning needed for proper examinations and estimations of data that could also be partly used in the area of Big Data.

One of our main intentions was to prepare a fully customized learning material for those who are at the beginning of this exciting path of gaining knowledge in this area and as well for those who are already on the way.

A brief presentation of the different problem areas is (where possible) complemented by practical cases and the use of R language to perform the building and testing of selected models. We have also added useful references for resources that could deepen the mentioned areas.

The authors believe that combination of diverse approaches and a solid base of references allow them to enlarge the range of potential readers from beginners to experienced users and all of them will benefit from the content and recommended sources selection that can be found by the end of each chapter.

Welcome and enjoy working with this guide.

Jan Luhan

(for the authors)

Content

Data Science

As the world entered the era of big data, the need for its storage also grew. The main focus was on building framework and solutions to store data. Now the focus has shifted to the processing of this data. *Data Science* is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems. Simply said, data science is a blend of skills from three major areas:

- Mathematics Expertise (Mathematical analysis, Linear algebra, Statistics, etc.),
- Technology and Hacking (SQL, Hadoop, Python, R and SAS),
- Business/Strategy acumen.

The main aspect of data science is all about uncovering findings from data. Data Scientists not only analyse past behaviour, but also uses various advanced machine learning algorithms to identify the occurrence of an event in the future. A Data Scientist will look at the data from many angles, sometimes angles not known earlier.

Significance of Data Science in today's digital world

Traditionally, the data that we had was mostly *structured* and small in size, which could be analysed by using the simple BI tools. Today the data is generated from different sources, e.g. text files, multimedia forms, sensors, etc. In this respect, today most of the data is *unstructured* and therefore simple BI tools are not capable of processing this huge volume and variety of data.

Applications of Data Science

As it is clear by now, Data Science is a broad term, and so are its applications. The earliest applications of data science were in Finance. Companies complained of bad debts and losses every year. Now Data Science has spread to other areas, such as

- Internet Search,
- Recommendation Systems,
- Image/Speech/Character Recognition,
- Gaming,
- Price Comparison Websites,
- Virtual assistance for patients and customer support,
- Augmented Reality,
- etc.

Data Science and R language

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R foundation.

What is R used for?

- Programming and statistical language,
- Data analysis and visualization.

The main benefits of R:

- Simple and easy to learn,
- Free and open source.

References

ALONSO-FERNÁNDEZ, Cristina, Antonio CALVO-MORATA, Manuel FREIRE, Iván MARTÍNEZ-ORTIZ and Baltasar FERNÁNDEZ-MANJÓN. Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*. 2019, 103612. ISSN 0360-1315. Available at doi:10.1016/j.compedu.2019.103612.

This paper presents a systematic literature review on how authors have applied data science techniques on game analytics data and learning analytics data from serious games to determine: (1) the purposes for which data science has been applied to game learning analytics data, (2) which algorithms or analysis techniques are commonly used, (3) which stakeholders have been chosen to benefit from this information and (4) which results and conclusions have been drawn from these applications.

BRUNNER, Robert J. and Edward J. KIM. Teaching Data Science. Procedia Computer Science. 2016, 80, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA, 1947–1956. ISSN 1877-0509. Available at doi:10.1016/j.procs.2016.05.513.

The course introduced general programming concepts by using the Python programming language with an emphasis on data preparation, processing, and presentation. The course had no prerequisites, and students were not expected to have any programming experience. This introductory course was designed to cover a wide range of topics, from the nature of data, to storage, to visualization, to probability and statistical analysis, to cloud and high performance computing, without becoming overly focused on any one subject.

CHAMIKARA, M. A. P., P. BERTOK, D. LIU, S. CAMTEPE and I. KHALIL. Efficient privacy preservation of big data for accurate data mining. *Information Sciences*. 2019. ISSN 0020-0255. Available at doi:10.1016/j.ins.2019.05.053.

Computing technologies pervade physical spaces and human lives, and produce a vast amount of data that is available for analysis. However, there is a growing concern that potentially sensitive data may become public if the collected data are not appropriately sanitized before being released for investigation. This paper addresses these issues by proposing an efficient and scalable nonreversible perturbation algorithm, PABIDOT, for privacy preservation of big data via optimal geometric transformations.

GIUDICI, Paolo. Financial data science. Statistics & Probability Letters. 2018, 136, The role of Statistics in the era of big data, 160–164. ISSN 0167-7152. Available at doi:10.1016/j.spl.2018.02.024.

This paper ~~we~~ provides a description of Financial data science, which involves the application of data science to technologically enabled financial innovations (FinTech), often driven by data science itself. We show that one of the most important data science models, correlation networks, can play a significant role in the advancements of Fintech developments.

SANCHEZ-PINTO, L. Nelson, Yuan LUO and Matthew M. CHURPEK. Big Data and Data Science in Critical Care. *Chest*. 2018, **154**(5), 1239–1248. ISSN 0012-3692. Available at doi:10.1016/j.chest.2018.04.037.

The present article reviews the definitions, types of algorithms, applications, challenges, and future of big data and data science in critical care.

Edureka.co
https://www.edureka.co/blog/data-science-tutorial/

Online Data Science Tutorial

Datajobs.com
https://datajobs.com/what-is-data-science

Online Data Science Tutorial

Introduction to Statistics

Statistics can be a powerful tool when performing Data Science. Using statistics, we can gain deeper and more precise view into how exactly our data is structured and based on that structure how we can optimally apply other data science techniques to get even more information.

Categories of Statistics for Data Science:

- Probability, Conditional probability,
- Random variable,
- Descriptive statistics,
- Mathematical statistics,
- Regression analysis.

Probability, Conditional probability

Probability is the measure of the likelihood that an event will occur. Simply said, probability is the chance that something will happen. Uncertainty and randomness occur in many aspects of our life and having a good knowledge of probability helps us. The knowledge of probability helps us make informed judgments on what is likely to happen, based on a set of data collected previously.

Probability is quantified as a number between 0 and 1. Probability of event A is

$$\mathrm{P(A)} = \frac{m(\mathrm{A})}{m(\Omega)} \tag{1}$$

where $m(\Omega)$ denotes number of all possible outcomes of an experiment (number of elements of a sample space Ω) and $m(\mathrm{A})$ denotes number of favourable outcomes of event A.

Conditional probability is a measure of the probability of an event given another event has occurred. Probability of event B given event A is

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(A)}}. \tag{2}$$

Many data science techniques rely on Bayes' theorem. Bayes' theorem is a formula describing how to update the probabilities of hypotheses when given evidence. For example, using conditional probability it is possible to build a learner that predicts the probability of the response variable belonging to some class, given a new set of attributes.

Random variable

A random variable is a function that assigns a numerical value to each outcome of an experiment. A random variable is called random because its possible values are outcomes of a random phenomenon.

Random variables can be:

- the *discrete* random variable which can only take certain (discrete, isolated) values,
- the *continuous* random variable which can take all values of an interval.

Behaviour of a random variable is described by distribution laws and empirical characteristics. *Distribution laws* are functions which determine the probability with which a random variable takes on a certain value.

Type of random variable	*Discrete*	*Continuous*
Distribution laws	Distribution function F(k)	Distribution function F(k)
	Probability mass function P(X=k)	Probability density function f(k)

For a discrete random variable, the probability distribution is defined by a *probability mass function*, which provides the probability for each value of the random variable X.

For a continuous random variable, the probability that a continuous random variable will lie within a given interval is considered. The probability distribution is defined by a *probability density function*.

Both probability functions must satisfy two requirements:

1) the probability function must be non-negative for each value of a random variable,
2) the sum of the probabilities for each value (or integral over all values) of a random variable always equals one.

For both type of random variables, a distribution function is a function, which is denoted by F(x) and defined for every x as follows

$$\mathrm{F}(k) = \mathrm{P}(X \leq k). \tag{3}$$

The distribution function F(x) expresses the probability that random variable X takes values from the interval $(-\infty, k\rangle$.

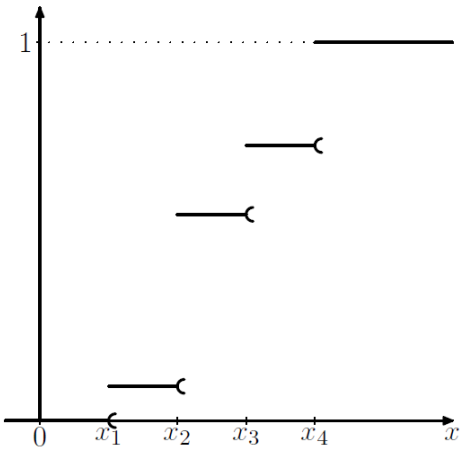

Figure 1 The distribution function of a discrete random variable

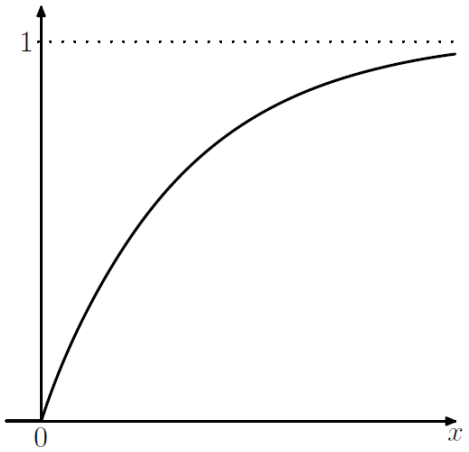

Figure 2 The distribution function of a continuous random variable

The distribution function and the probability function describe the probability distribution of values of a random variable. However, sometimes Data Scientists require a summary of the overall information of a random variable in a few numbers that characterize its other properties and allow us to compare it with the other random variable. These numbers are called empirical characteristics.

The most important characteristic of the random variable X is the *expected value* which we denote E(X). The expected value represents the number around which averages, calculated from a series of observed values, fluctuate.

The most important of the central moments is the second central moment, which we call the *variance* and denote D(X). Since the variance has a square dimension of the random variable X, for practical interpretation the *standard deviation* is the preferable characteristic. The standard deviation is denoted by $\sigma(X)$.

List of probability distributions

It has been found that some random variables behave according to the same rules. These rules have been described and probability distributions defined. Many probability distributions that are important in theory or applications have been given specific names.

Binomial distribution

Let's do an experiment, the event A may occur with the probability p. We repeat the experiment under the same conditions n-times, while the number p is the same in each experiment. This sequence of experiments is called the Bernoulli sequence of n independent experiments.

The discrete random variable X represents the number of occurrences of the event A in this sequence of experiments.

Hyper-geometric distribution

Let's have a file consisting of N elements which M elements have a certain property and $N - M$ elements have another property. From this set is randomly selected n elements, either simultaneously or consecutively without replacement.

The discrete random variable X represents the number of selected elements having a certain property.

Poisson distribution

Poisson distribution describes the probability laws for the "infrequently occurring events", which are the number of occurrences of events within a certain time interval, in certain parts of the area, etc.

Normal distribution

The *normal distribution*, also known as the *Gaussian distribution*, is a probability distribution that is symmetric about the expected value, showing that data near the expected value are more frequent in occurrence than data far from the expected value. It has following properties:

- the normal curve is symmetrical about the expected value μ,
- the expected value is at the middle and divides the area into halves,
- the total area under the curve is equal to 1,

- it is completely determined by its expected value and standard deviation σ (or variance σ^2).

The probability density function *f*(*x*) has a typical "bell curve" shape, see Figure 3.

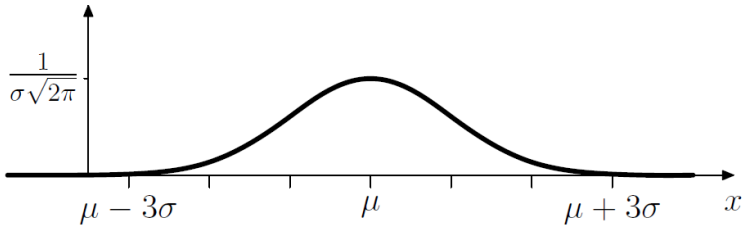

Figure 3 Graph of the probability density function

The distribution function F(*x*) has a typical "S curve" shape, see Figure 4.

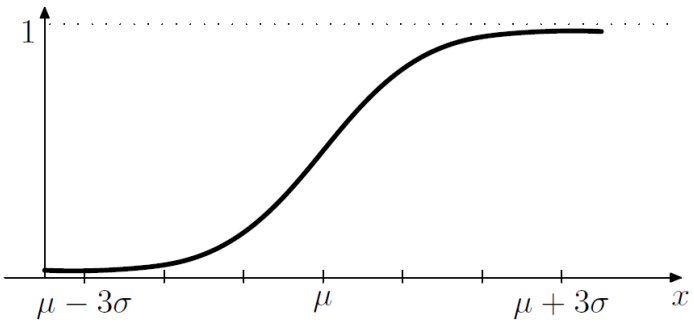

Figure 4 Graph of the distribution function

Exponential distribution

The exponential distribution has wide applicability in queuing theory, in reliability theory, and the theory of recovery. The exponential distribution is sometimes called the distribution without memory and is suitable to describe the distribution of the lifetime of a device, where a failure occurs completely randomly, i.e. external causes.

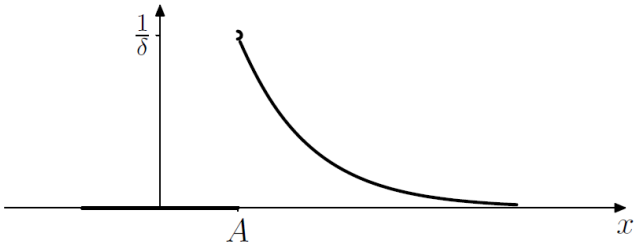

Figure 5 Graph of the probability density function

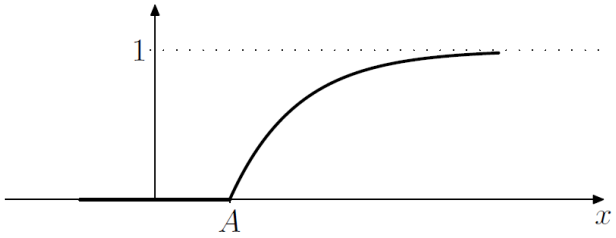

Figure 6 Graph of the distribution function

Probability distributions in R

Binomial distribution – **Bi**(*n*,*p*).

Mathematical description	Command in R
$P(X = k)$	dbinom (*k, n, p*)
$P(X \leq k)$	pbinom (*k, n, p*)
$P(X > k)$	pbinom (*k, n, p,* FALSE)

Hypergeometric distribution – **H**(*N, M, n*).

Mathematical description	Command in R
$P(X = k)$	dhyper (*k, M, N – M, n*)
$P(X \leq k)$	phyper (*k, M, N – M, n*)
$P(X > k)$	phyper (*k, M, N – M, n,* FALSE)

Poisson distribution – **Po**(λ).

Mathematical description	Command in R
$P(X = k)$	dpois (*k,* λ)
$P(X \leq k)$	ppois (*k,* λ)
$P(X > k)$	ppois (*k,* λ, FALSE)

Graphs of the probability function and the distribution function

Teh procedure how to draw charts of both functions will be demonstrated on the example of a random variable X which takes values 0, 1, 2, 3, 4, 5 and has the binomial distribution Bi(5, 0.95)

$X = 0 : 5$	# a random variable X takes values 0, 1, 2, 3, 4, 5
plot(*x*,dbinom(*x*, 5, 0.95), "h")	# the graph of the probability function
plot(*x*,pbinom(*x*, 5, 0.95), "s")	# the graph of the distribution function

Normal (Gauss) distribution – N(μ, σ)

Mathematical description	Command in R
$f(X = k)$	dnorm (k, μ, σ)
$P(X \leq k)$	pnorm (k, μ, σ)
$P(X > k)$	pnorm (k, μ, σ, FALSE)

Exponential distribution – E(δ)

Mathematical description	Command in R
$f(X = k)$	dexp (k, $1/\delta$)
$P(X \leq k)$	pexp (k, $1/\delta$)
$P(X > k)$	pexp (k, $1/\delta$, FALSE)

Graphs of the probability density function and the distribution function

The procedure how to draw charts of the probability density function and the distribution function will be demonstrated on the example of a random variable X which has the normal distribution N(35,102)

x = seq(5, 65, length = 100)	# a random variable which takes all values from the interval <5; 65> = $<\mu - 3\sigma; \mu + 3\sigma>$
plot(*x*,dnorm(*x*, 35, 10),"l")	# the graph of the probability density function
plot(*x*,pnorm(*x*, 35, 10), "l")	# the graph of the distribution function

Example of a discrete random variable

The test contains 10 questions. Four answers are offered to each question and only one is correct. At least eight correct answers are required to pass the test. What is the probability that a student, who randomly answers, will pass the test?

The experiment: answering a question.
The event A: an answer is correct.
The random variable X: the number of correct answers.

This random variable can be described by the binomial distribution because we know the number of repetition of the experiment (n = 10) and the probability of the event A (P(A) = ¼). A student will pass the test if he has at least eight correct answers. The probability of this event we can calculate as P($X \geq 8$).

$$\mathrm{P}(X \geq 8) = \mathrm{P}(X = 8) + \mathrm{P}(X = 9) + \mathrm{P}(X = 10). \quad (4)$$

Solution with R: dbinom(8, 10, 0.25) + dbinom(9, 10, 0.25) + dbinom(10, 10, 0.25)

$$\mathrm{P}(X \geq 8) = 0.00042. \quad (5)$$

Example of a continuous random variable

A content of automatically filled bottles has a normal distribution N(0.7; 0.02^2). How many bottles in the supply of 2,000 bottles have the contents less than 0.68 litre?

The random variable X: the contents of a bottle.

We want to calculate the probability of the contents is less than 0.68, P(X < 0.68) or P(0 < X < 0.68). This probability we can calculate by the distribution function F(x).

$$\mathrm{P}(0 < X < 0.68) = \mathrm{F}(0.68) – \mathrm{F}(0). \quad (6)$$

Solution with R: pnorm(0.68,0.7,0.02) – pnorm(0,0.7,0.02)

$$\mathrm{P}(0 < X < 0.68) = 0.159. \quad (7)$$

Now we have to determine how much is 15.9 % of 2,000 bottles.

$$0.159 \cdot 2{,}000 = 318. \quad (8)$$

318 bottles of 2,000 have the content less than 0.68 litre.

References

ALLEN, Arnold O. *Probability, Statistics, and Queueing Theory*. B.m.: Academic Press, 2014. ISBN 978-0-08-057105-8.

This is a textbook on applied probability and statistics with computer science applications for students at the upper undergraduate level. It may also be used as a self-study book for the practising computer science professional.

DataJobs.com
https://datajobs.com/what-is-data-science

Online Data Science Tutorial

HERBSTRITT, Michele and Michael FRANKE. Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*. 2019, 186, 50–71. ISSN 0010-0277. Available at doi:10.1016/j.cognition.2018.11.013.

Presents novel experimental data pertaining to the use and interpretation of simple probability expressions (such as possible or likely) and complex ones (such as possibly likely or certainly possible) in situations of higher-order uncertainty, i.e., where speakers may be uncertain about the probability of a chance event.

LINTON, Oliver. Probability, Statistics and Econometrics. Academic Press, 2017. ISBN 978-0-12-810496-5.

The book covers much of the groundwork for probability and inference before proceeding to core topics in econometrics.

R-project.org
https://www.r-project.org/.

The R Project for Statistical Computing

Descriptive statistics

A descriptive statistic is a summary statistic that summarizes features of a collection of information. Descriptive statistics are simply descriptive.

When performing statistical analysis we deal with events and processes, which occur on a mass scale and can be found in a large set of individual objects such as products or persons. We call this set the *population*. The objects under investigation are called *statistical items* and we observe them focusing on certain properties *statistical variables*.

By the type of outcomes variables are either *quantitative*, with numerical outcomes such as a length, strength, price, service life and the like, or *qualitative* which are not numeric and can only be expressed by words such as a colour, quality class, operation condition.

Quantitative variables are

- *discrete* if they only take on discrete (isolated) values (a number of defective products, a number of faults, a number of pieces),
- continuous if they take on all values of an interval (a size of a product, a time to failure).

Qualitative variables are

- *ordinal* if there is a point in ordering their outcomes expressed in words such as quality classes,
- *nominal* there is no point in ordering them such as colour, form, suppliers.

Statistics methods are based on fact that information on the population is not taken from all its elements but rather from a subset of the population. The subset is called the *sample*. The number of statistical items in the sample is called the size of the sample. If the size is less than or equal 50 we say that the sample is small, if the size is greater than 50 we say that the sample is large.

Values of a statistical variable depend on the random phenomena so a statistical variable is a random variable. Therefore, the necessary information is obtained from a sample data using

- empirical characteristics (the sample mean $\bar{x}$, the sample variance s^2, the sample standard deviation s, etc.),
- empirical distribution laws (the empirical distribution function $F_n(x)$).

Central tendency

Central tendency is a *central* or *typical* value for a distribution. It may also be called a *centre* of the distribution. The most common measures of central tendency are the *arithmetic mean*, the *median* and the *mode*.

The m*ean* is the numerical average of all values. Because the mean is sensitive to extreme values it is important to find out extreme values of the sample data. For this, we can use the box plot. The *median* is directly in the middle of the sample data and the *mode* is the most frequent value in the sample data.

Spread or variance

Spread (dispersion) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the *variance* and the *standard deviation*.

Empirical distribution function $F_n(x)$

The empirical distribution function represents the probability distribution which is obtained from the sample data. Knowledge of this distribution allows us to correctly interpret the empirical characteristics.

The empirical distribution function has two main meanings:

1) Values of the empirical distribution function $F_n(x)$ serve as an estimate of values of the distribution function $F(x)$, i.e. $F_n(x) \approx P(X \leq x)$.
2) According to a shape of a graph of the empirical distribution function we can assume the concrete type of the probability distribution of the statistical variable, such as the normal distribution, exponential distribution, Poisson distribution, etc.

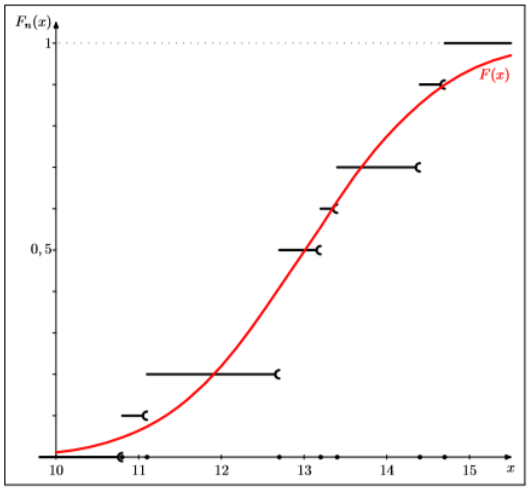

Figure 7 Graph of the empirical and theoretical distribution function

Figure 7 shows the empirical distribution function (the black curve) and the theoretical distribution function of the normal distribution (the red curve). Because the red curve intersects the black curve we can assume that the behaviour of a statistical variable can be described by the normal distribution.

A large data file can be processed by sorting. The sorting sample data provides information about

- the concrete type of the probability distribution of the statistical variable,
- interventions to the sample data, errors in measurement, etc.

The most commonly used graphical representation of a sorting sample data is the histogram, see Figure 8 and Figure 9.

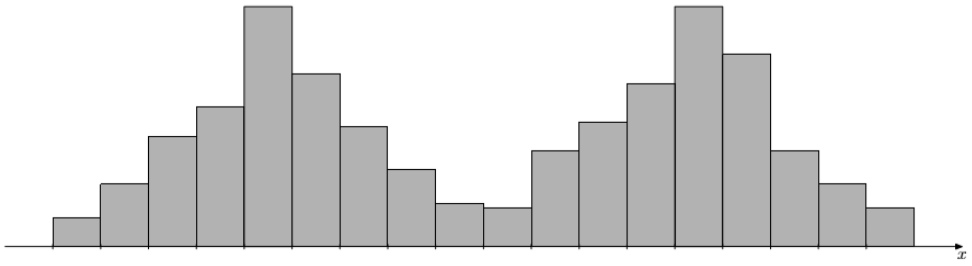

Figure 8 Two data files are mixed

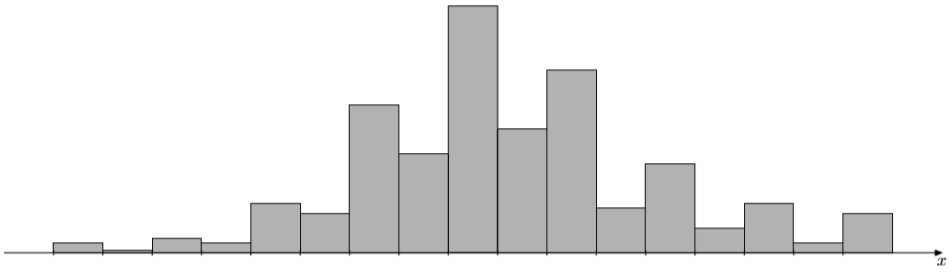

Figure 9 The rounding problem when measuring

Descriptive statistics in R

Entering values

In the following lists of the commands the symbol x represents a one-dimensional data (statistical) file. In the R environment measured values can be easily input using the following ways.

Way	Command in R
Direct input	*x* = c(4.1,4.0,3.8,3.9,3.8,3.8,3.5,3.7,4.0,4.0)
Loading form external file	*x* = read.table ("path/file_name.txt")

Empirical characteristics

Characteristic	Designation	Command in R
Size of a sample data	*n*	length(x)
Sample mean	$\overline{x}$	mean(x)
Sample variance	s^2	var(x)
Sample standard deviation	*s*	sd(x)
Minimal value	*min*	min(x)
Maximal value	*max*	max(x)
Median	$\tilde{x}_{0.5}$	median(x), quantile(x,0.5,type=6)
Quantile	$\tilde{x}_p$	quantile(x,p,type=6)

Empirical distribution function

The following procedure shows how to draw graph of the empirical distribution function.

f = ecdf(*x*)	# calculating of the empirical distribution function
f(*k*)	# calculating value of the empirical distribution function for a value k
knots(*f*)	# ascending order values from data sample
f(knots(*f*))	# values of the empirical distribution function
plot(*f*)	# the graph of the empirical distribution function

Sorting of a large data sample

The following part shows to sort a large one-dimensional data sample.

Class	Command in R
The number of classes is determined by the program	hist(*x*)
Entered the boundaries of individual classes	hist(*x*, br=seq(38,46,by=1)
Another way to specify the boundaries of individual classes	hist(*x*, br=c(38,39,40,41,42,43,45,46)

The standard output of the command *hist* is a histogram. To obtain more information you can use the following ways.

y = *hist* (*x*)	# the variable *y* contains results of the command *hist*
y$counts	# a frequency for each class
y$density	# a relative frequency for each class
y$breaks	# boundaries of individual classes
y$mids	# midpoints of individual

Class borders can be defined by a parameter *right* in the following way.

Intervals	Parameter *right*
$(c_i;c_{i+1}>$	right=TRUE (implicit value)
$<c_i;c_{i+1})$	right=FALSE

Example of a small data sample

It is required to normalize the time needed to a repair defective components. For this purpose we randomly selected 10 mechanics who carried out repairs. The times (in minutes) to form a data sample:

11.1, 12.7, 10.8, 13.2, 12.7, 13.4, 12.7, 14.4, 14.7, 14.4.

Determine basic characteristics $\bar{x}$, s^2, $\tilde{x}_{0.5}$ and a kind of distribution.

The population: all mechanics.

The statistical item: a mechanic.

The Statistical variable: the time needed to a repair.

The sample: 10 mechanics.

Calculation of basic empirical characteristics.

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10} \cdot 130.1 = 13.01 \text{ (}\textit{Solution in R:}\text{ mean(}x\text{))},$$
$$s^2 = \frac{1}{10-1}\left[\sum_{i=1}^{10} x_i^2 - 10 \cdot \bar{x}^2\right] = \frac{1}{9}[1708.33 - 10 \cdot 13.01^2] \doteq 1.75 \quad (9)$$
(*Solution in R:* var(*x*)),
$$s = \sqrt{s^2} = \sqrt{1.75} \doteq 1.32 \text{ (}\textit{Solution in R:}\text{ sd(}x\text{))}.$$

The ordered data file (*Solution in R:* sort(*x*)),

i	1	2	3	4	5	6	7	8	9	10
x(*i*)	10.8	11.1	12.7	12.7	12.7	13.2	13.4	14.4	14.4	14.7

Since the data set has an even number of elements, we calculate the median by the formula

$$\tilde{x}_{0.5} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{12.7+13.2}{2} = 12.95 \text{ (}\textit{Solution with R}\text{:} \quad (10)$$
median(*x*)).

The median says 50 % of repairs will end up within 12.95 minutes and the remaining 50 % of repairs will take longer.

Since the median value is close to the average value, we can assume that the distribution of the statistical variable is symmetrical, e.g. the normal distribution. The particular kind of the distribution we can find out by the empirical distribution function $F_n(x)$. The empirical distribution function values can be calculated by Table 1.

Table 1 The empirical distribution function

x	10.8	11.1	12.7	13.2	13.4	14.4	14.7
l	1	1	3	1	1	2	1
k	1	2	5	6	7	9	10
$F_n(x)$	0.1	0.2	0.5	0.6	0.7	0.9	1

The value of $F_n(13.4) = 0.7$ indicates that about 70% of repairs are completed within 13.4 minutes.

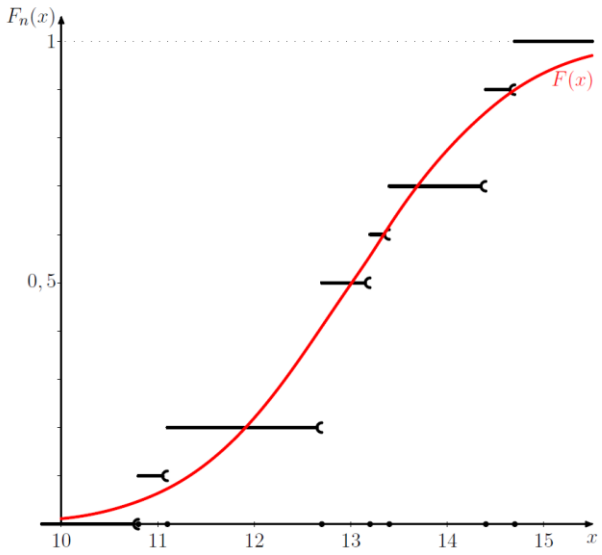

Figure 10 The graph of the empirical distribution function and the distribution function of the normal distribution

Solution with R: F = ecdf(*x*), plot(F)

Figure 10 shows that the shape of the empirical distribution function is close to the shape of the theoretical distribution function of the normal distribution. Therefore, we can assume that the behaviour of the observed variable can be described by the normal distribution.

Example of a large data sample

The time to repair certain defects of TV was measured. During 50 repairs were determined the following times (in minutes):

44.0, 40.2, 41.9, 43.4, 42.8, 42.3, 43.2, 45.0, 41.5, 42.7,
43.9, 40.1, 43.3, 41.1, 42.5, 42.4, 41.4, 40.8, 42.0, 44.7,
41.2, 41.9, 42.4, 44.4, 40.5, 39.7, 41.1, 41.0, 41.9, 40.8,
43.0, 42.8, 42.9, 42.7, 43.3, 42.2, 39.2, 41.5, 41.6, 42.7,
45.0, 42.3, 43.6, 43.2, 38.8, 43.0, 44.2, 43.0, 40.0, 44.4.

The size of the data sample is 50. We don't see any information from all 50 values. If we want to obtain some information, we have to class the data sample. Firstly, we have to choose the number of classes. Next, we have to determine how many values of the data sample lie in each class. The choice of the number of classes and finding out how many values from the data sample lies in each class are laborious. Therefore, some statistical software is recommended to use.

Solution with R: hist(*x*)

Table 2 Processed data sample

Interval/class	Representative	Frequency	Relative frequency
$\langle 38, 39)$	38.5	1	0.02
$\langle 39, 40)$	39.5	3	0.06
$\langle 40, 41)$	40.5	6	0.12
$\langle 41, 42)$	41.5	11	0.22
$\langle 42, 43)$	42.5	15	0.3
$\langle 43, 44)$	43.5	8	0.16
$\langle 44, 45)$	44.5	6	0.12
	∑	50	1

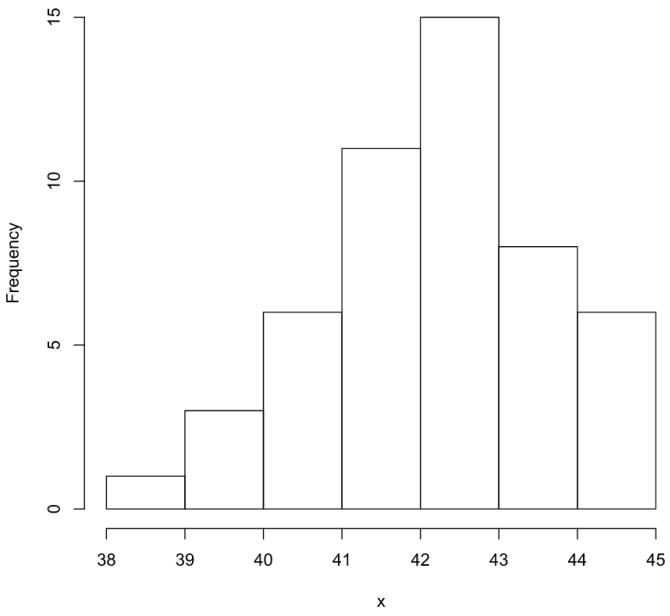

Figure 11 Histogram

Figure 11 shows there is only one peak and that the shape of the histogram is close to the shape of the Gauss curve. Therefore, we can assume that the behaviour of the observed variable can be described by the normal distribution.

References

BROWNSTEIN, Naomi C., Andreas ADOLFSSON and Margareta ACKERMAN. Descriptive statistics and visualization of data from the R datasets package with implications for clusterability. Data in Brief. 2019, 104004. ISSN 2352-3409. Available at doi:10.1016/j.dib.2019.104004.

The manuscript describes and visualizes datasets from the datasets package in the R statistical software, focusing on descriptive statistics and visualizations that provide insights into the clusterability of these datasets.

FIELD, Andy, Jeremy MILES and Zoë FIELD. Discovering Statistics Using R. SAGE, 2012. ISBN 978-1-4462-5846-0.

The journey begins by explaining basic statistical and research concepts before a guided tour of the R software environment. Next you discover the importance of exploring and graphing data, before moving onto statistical tests that are the foundations of the rest of the book (for example correlation and regression). You will then stride confidently into intermediate level analyses such as ANOVA, before ending your journey with advanced techniques such as MANOVA and multilevel models. Although there is enough theory to help you gain the necessary conceptual understanding of what you're doing, the emphasis is on applying what you learn to playful and real-world examples that should make the experience more fun than you might expect.

LINTON, Oliver. *Probability, Statistics and Econometrics.* Academic Press, 2017. ISBN 978-0-12-810496-5.

The book covers much of the groundwork for probability and inference before proceeding to core topics in econometrics.

MARSHALL, Gill and Leon JONKER. An introduction to descriptive statistics: A review and practical guide. Radiography. 2010, 16(4), e1–e7. ISSN 1078-8174. Available at doi:10.1016/j.radi.2010.01.001.

This paper, the first of two, demonstrates why it is necessary to understand basic statistical concepts both to assimilate the work of others and also in their own research work.

NYE, John V. C., Maksym BRYUKHANOV and Sergiy POLYACHENKO. Descriptive statistics and regressions of 2D:4D and educational attainment based on RLMS data. Data in Brief. 2017, 12, 552–583. ISSN 2352-3409. Available at doi:10.1016/j.dib.2017.04.009.

We document the descriptive statistics and detailed regression outputs for educational attainment and measured 2D:4D ratios

Inferential statistics

Inferential statistics is the second main branch of statistics which use a random sample of data taken from a population to describe and make inferences about the population. For example, to measure the lifetime of each battery that is manufactured is impractical. You can measure the lifetime of a representative random sample of batteries. You can use the information from the sample to make generalizations about the lifetime of all of the batteries.

Central Limit Theorem

The *Central Limit Theorem* is used to help us understand the following facts:

1) the mean of the sample is the same as the population mean,
2) the standard deviation of the sample means is always equal to the standard error,
3) the distribution *of* sample means will become increasingly more normal as the sample size increases.

Point and interval estimates

We usually do not know the real value of a parameter β of a probability distribution and we try to estimate it using a sample data. We can make the point estimate and the interval estimate.

The *point estimate* of parameter β is the value $t = \mathrm{T}(x_1,...,x_n)$ an estimator assumes for a sample data $(x_1,...,x_n)$. The point estimates of population characteristics are calculated

$$\mathrm{E}(X) = \bar{X}, \mathrm{D}(X) = S^2, \sigma(X) = S.$$

The interval estimate for parameter β with the confidence level $1 - \alpha$ is a pair of statistics $(\mathrm{T}_1,\mathrm{T}_2)$.

$$\mathrm{P}(\mathrm{T}_1 \leq \beta \leq \mathrm{T}_2) = 1 - \alpha. \qquad (11)$$

Hypothesis Testing

Hypothesis testing is a type of *statistical inference* that involves asking a question, collecting data, and then examining what the data tells us about how to proceed. The hypothesis to be tested is called the *null hypothesis* and given the symbol H. We test the null hypothesis against an *alternative hypothesis*, which is given the symbol $\bar{H}$.

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses using experimental data (a data sample).

When testing statistical hypotheses the following procedure is recommended:

1. We formulate the null hypothesis H and the alternative hypothesis $\overline{H}$.
2. We calculate the value g of a test statistics G.
3. We choose the number α, called level of significance and determine the so-called critical range W_α.
4. Depending on how the value of the test statistics g is realized in the critical range, we say one of following decision:
    - if $g \in W_\alpha$, then we reject the null,
    - if $g \notin W_\alpha$, then we accept the null hypothesis.

When testing hypotheses using a computer, statistical programs show the p-value instead of the critical range W_α. The *p-value* represents the size of the probability that a random variable T is in a certain relation to the calculated value of the test statistics t. The specific relationship depends on the selected variant of the null and alternative hypotheses.

Comparing the p-value and the level of significance, leads to one of the following decisions:

- If the p-value is less than the level of significance α, we reject the null hypothesis and accept the alternative hypothesis.
- If the p-value is greater or equal than the level of significance α, we accept (leave) the null hypothesis.

Testing can ended up with two errors, which are called error of the first type and error of the second type.

Table 3 Testing errors

H	*True*	*False*
Reject	the first type error	
Accept		the second type error

The level of significance α is a probability of the first type error.

Parametric tests

Parametric tests make an assumption about the shape of the population distribution, see page - 10 -.

The one sample t-test

The one sample t-test tells you how significant the differences between two groups are. In other words, it lets you know if those differences could have happened by chance or not. For example, does the change in production technology affect the lifetime of a product?

Requirements:

- A variable X is continuous with the Normal distribution.
- The parameters μ and σ^2 of this distribution are unknown.
- The observations are independent of one another.

The one sample t-test assesses the relationship between the selected number μ_0 and the unknown value of the parameter μ which is estimated by the mean $\bar{x}$. The one sample t-test can be written in the following table

Table 4 The one sample t-test

H	$\bar{H}$	Critical range W_α
$\mu \leq \mu_0$	$\mu > \mu_0$	$\left\{t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}; t \geq t_{1-\alpha}(n-1)\right\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\left\{t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}; \lvert t\rvert \geq t_{1-\frac{\alpha}{2}}(n-1)\right\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\left\{t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}; t \leq -t_{1-\alpha}(n-1)\right\}$

Where $t_{1-\alpha}(n-1)$ or $t_{1-\frac{\alpha}{2}}(n-1)$ is the quantile of the Student's t distribution.

There are also the t-test modification such as the paired t-test and the two sample t-test.

The one sample proportion test

The one sample proportion test is used to assess whether a population proportion P is significantly different from a hypothesized value P_0. This is called the hypothesis of inequality. For example, suppose that the current treatment for a disease cures 74 % of all cases. A new treatment method has been proposed and studied. In a sample of 80 subjects with the disease that were treated with the new method, 61 were cured. Do the results of this study support the claim that the new method has a higher success rate?

Requirements:

- A variable X is discrete with the Alternative distribution.
- The parameters p is unknown.
- The observations are independent of one another.

The one sample t-test assesses the relationship between the selected number p_0 and the unknown value of the parameter p which is estimated by the mean $\bar{x}$. The one sample proportion test can be written in the following table

Table 5 The one sample proportion test

H	$\bar{H}$	Critical range W_α
$p \leq p_0$	$p > p_0$	$\left\{u = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}; u \geq u_{1-\alpha}\right\}$
$p = p_0$	$p \neq p_0$	$\left\{u = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}; \lvert u\rvert \geq u_{1-\frac{\alpha}{2}}\right\}$
$p \geq p_0$	$p < p_0$	$\left\{u = \frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}; u \leq -u_{1-\alpha}\right\}$

Where $u_{1-\alpha}$ or $u_{1-\frac{\alpha}{2}}$ is the quantile of the Standard normal distribution.

Parametric tests in R

Entering values

In the following lists of the commands the symbol x represents a one-dimensional data (statistical) file. In the R environment measured values can be easily input using the following ways.

Way	Command in R
Direct input	*x* = c(4.1,4.0,3.8,3.9,3.8,3.8,3.5,3.7,4.0,4.0)
Loading form external file	*x* = read.table (”path/file_name.txt”)

The t-test

If we want to compare for the significance level α a parameter μ, estimated by a sample mean $\overline{x}$, with an entered value μ_0, then we can use different variants of this test the following commands.

H	$\overline{H}$	Command in R
$\mu \leq \mu_0$	$\mu > \mu_0$	t.test(x, alternative="greater", mu=μ_0, conf.level=1-α)
$\mu = \mu_0$	$\mu \neq \mu_0$	t.test(x, alternative="two.sided", mu=μ_0, conf.level=1-α)
$\mu \geq \mu_0$	$\mu < \mu_0$	t.test(x, alternative="less", mu=μ_0, conf.level=1-α)
		α - the level of significance

The two sample t-test

If we have two data sets then we can compare their sample means using the following commands.

H	$\overline{H}$	Command in R
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	t.test(x, y, alternative="greater", var.equal=T/F, conf.level=1-α)
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	t.test(x, y, alternative="two.sided", var.equal=T/F, conf.level=1-α)
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	t.test(x, y, alternative="less", var.equal=T/F, conf.level=1-α)
		α - the level of significance

The variant TRUE of the parameter *var.equal* denotes an assumption of equality of variances both data files, the variant FALSE indicates an assumption of diversity of these variances.

In the event that we have no knowledge of the equality or diversity of variances, we can use the following test.

H	$\overline{H}$	Command in R
$\sigma_1 \leq \sigma_2$	$\sigma_1 > \sigma_2$	var.test(x, y, alternative="greater", conf.level=1-α)
$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	var.test(x, y, alternative="two.sided", conf.level=1-α)
$\sigma_1 \geq \sigma_2$	$\sigma_1 < \sigma_2$	var.test(x, y, alternative="less", conf.level=1-α)
		α - the level of significance

The Proportional test

We can use for different variants of this test the following commands.

H	$\overline{H}$	Command in R
$p \leq p_0$	$p > p_0$	prop.test(k, n, p=p_0, alternative="greater", conf.level=1-α)
$p = p_0$	$p \neq p_0$	prop.test(k, n, p=p_0, alternative="two.sided", conf.level=1-α)
$p \geq p_0$	$p < p_0$	prop.test(k, n, p=p_0, alternative="less", conf.level=1-α)
		α - the level of significance; k number of positive answers; n number of all answers

Example of the one sample t-test

The bus line had an average driving time of 12 minutes on a route. Changes have been made to the route. Assess whether these changes have impacted on the driving time, if nine driving times were measured (in minutes)

12.5, 13.5, 11.9, 12.2, 13.0, 14.3, 12.2, 11.8, 14.0.

The population: all rides.

The statistical item: a ride.

The Statistical variable: the driving time (in minutes).

The sample: 9 rides.

Calculation of basic empirical characteristics.

$$\begin{aligned} \bar{x} &\doteq 12.82 \text{ (\textit{Solution in R:} mean}(x)), \\ s &\doteq 0.92 \text{ (\textit{Solution in R:} sd}(x)), \\ \tilde{x}_{0.5} &\doteq 12.5 \text{ (\textit{Solution in R:} median}(x)). \end{aligned} \tag{12}$$

Because the median is close to the mean, we can assume that the distribution of the driving time is symmetric, e.g. the normal distribution. The other way to find out the type of distribution is to draw the empirical distribution function, see page - 17 -, - 20 -.

The driving time is the continuous variable and if we assume the normal distribution of this variable then we can use it for assessment of driving time changes of a bus line the one sample t-test.

Firstly, we have to formulate hypotheses. If we want to prove that the changes have negative impact to the driving time (the driving time increases) so we choose the first line from Table 4. Then we can choose the level of significance, e.g. α = 0.05.Now we can use some statistical software.

Solution with R: t.test(x,alternative="greater",mu=12)

We obtain following result.

```
data:  x
t = 2.6685, df = 8, p-value = 0.01421
alternative hypothesis: true mean is greater than 12
95 percent confidence interval:
 12.24926      Inf
sample estimates:
mean of x
 12.82222
```

Because the p-value 0.01421 is less than the level of significance, e.g. α = 0.05, we reject the null hypothesis H and accept the alternative hypothesis $\bar{H}$. We can say the changes have negative impact to the driving time (the driving time increases).

Normality tests

Normality tests are used to determine if a data set is well-modelled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

Normality tests in R

The Kolmogorov-Smirnov test

The universal test which can be used for continuous distributions. Using this test is shown in an example of a normal (Gauss) distribution.

ks.test(x,"pnorm", μ, σ)	# μ, σ are parameters of the normal (Gauss) distribution

The Shapiro-Wilk test

This test is designed only for testing whether the data (statistical) file is selected from the population which has the *normal* (Gauss) distribution.

shapiro.test(*x*)

Analysis of variance

Analysis of variance (ANOVA) can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

$$H: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k, \tag{13}$$

where μ represents group mean, k represents number of groups.

Assumptions of ANOVA:

- all populations involved follow a normal distribution,
- all populations have the same variance or standard deviation (Bartlett's test),
- the samples are randomly selected and independent of each other.

To use the F-test to determine whether group means are equal, it's just a matter of including the correct variances in the ratio. In one-way ANOVA, the F-statistic is this ratio

$$F = \frac{\text{variation between sample means}}{\text{variation within the samples}}. \qquad (14)$$

.

Analysis of variance in R

Measured values

Compact cars	643	655	702
Mid-size cars	469	427	515
Full-size cars	484	456	402

Entering values

Way	Command in R
Direct input	*response* = c(643,655,702,469,427,525,484,456,402)
Loading form external file	*response* = read.table ("path/file_name.txt")

Data preparation

Way	Command in R
Identification of factors	*factor* = c(rep("Compact",3), rep("Mid",3), rep("Full",3))
Data	*data_name* = data.frame*(response.factor)*

One-Way ANOVA

Way	Command in R
One-Way ANOVA	*Results*=aov(*response ~ factor,* data=*data_name)*
Results of ANOVA	*summary* (*results*)

Bartlett's test

Bartlett's test is used to test if *k* samples are from populations with equal variances. Equal variances across populations is called *homoscedasticity* or *homogeneity* of variances. The analysis of variance assumes that variances are equal across groups or samples. The Bartlett's test can be used to verify that assumption.

Way	Command in R
Bartlett's test	bartlett.test(*response ~ factor, data=data_name)*

Turkey's test

Turkey's test is a single-step multiple comparison procedure and statistical test. It can be used in conjunction with an ANOVA (post-hoc analysis) to find means that are significantly different from each other.

Way	Command in R
Turkey's test	TurkeyHSD(*results,* conf.level=0.95*)*

Example of one-way ANOVA

A management of a company wants to compare four employee training programs. Twenty volunteers are randomly divided into four groups with the same number of participants. Each training group is trained according to one of the programs prepared. At the end of the training, each participant does the same task. The monitored variable is the time (in minutes) needed to complete the task.

Program	Time (in minutes)				
1	9	12	14	11	11
2	10	6	9	9	10
3	12	14	11	13	11
4	9	8	11	7	8

The statistical variable: the time needed to complete task,
the type of training program (factor).

We assume the training programs are independent and the time has the normal distribution (it can be verified with normality tests, see page - 31 -). Then we have to verify the homoscedasticity condition. We use the Bartlett's test where the null hypothesis says the variances are equal (there is homoscedasticity), the alternative hypothesis says the variances are different (there is heteroscedasticity). We choose the level of significance 0.05.

Solution with R: bartlett.test(time~factor,data)

We obtain following result.

```
data:  time by factor
Bartlett's K-squared = 0.41462, df = 3, p-value = 0.9372
```

Because the p-value 0.9372 is greater than the level of significance, e.g. α = 0.05, we accept the null hypothesis H. We can say there is homoscedasticity. So we can go to the one way ANOVA test.

Solution with R: summary(aov(time~factor,data))

We obtain following result.

```
            Df Sum Sq Mean Sq F value  Pr(>F)
factor       3  49.75   16.58   6.633 0.00404 **
Residuals   16  40.00    2.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value 0.00404 is less than the level of significance, e.g. α = 0.05, we accept the alternative hypothesis $\bar{H}$. It means there is a difference between the training programs.

Nonparametric tests

When using t-test or analysis of variance, the assumption of data normality should be met. To select larger ranges, a slight violation of normality has no significant impact on the outcome. However, in the case of small selections, a violation of the normality condition may lead to misleading results. So-called non-parametric tests were created for such cases, i.e. small selections with a markedly non-normal distribution.

Nonparametric tests do not require the assumption of a particular type of distribution of a variable, e.g. normal distribution. Mostly, it is sufficient to assume that the distribution function is continuous. These tests can also be used in situations where the data being investigated are not interval or proportional in character but merely ordinal in nature. In comparison with parametric tests, non-parametric tests are weaker, i.e. they reject the false hypothesis less likely than parametric tests.

These tests usually serve as a substitute for one sample t-test, pair t-test, two sample t-test and ANOVA, such as the sign test, the Wilcoxon test, the Wilcoxon two-sample test, the Kruskal-Wallis test, the median test and the Friedman's test.

The one sample Wilcoxon test as a nonparametric equivalent of the one sample t-test

The test assesses the deviation between the median and the real constant *c*.

H	$\overline{H}$	Command in R
$x_{0.5} \leq c$	$x_{0.5} > c$	wilcox.test(x, alternative="greater", mu=c, conf.level=1-α)
$x_{0.5} = c$	$x_{0.5} \neq c$	wilcox.test(x, alternative="two.sided", mu=c, conf.level=1-α)
$x_{0.5} \geq c$	$x_{0.5} < c$	wilcox.test(x, alternative="less", mu=c, conf.level=1-α)
		α - the level of significance (implicitly 0.05) x - the data sample

The one sample Wilcoxon test as a nonparametric equivalent of the paired t-test

The test assesses the deviation between the median difference and the real constant *c*.

H	$\overline{H}$	Command in R
$x_{0.5} - y_{0.5} \leq c$	$x_{0.5} - y_{0.5} > c$	wilcox.test(x,y, alternative="greater", paired=TRUE, mu=c, conf.level=1-α)
$x_{0.5} - y_{0.5} = c$	$x_{0.5} - y_{0.5} \neq c$	wilcox.test(x,y, alternative="two.sided", paired=TRUE, mu=c, conf.level=1-α)
$x_{0.5} - y_{0.5} \geq c$	$x_{0.5} - y_{0.5} < c$	wilcox.test(x,y, alternative="less", paired=TRUE, mu=c, conf.level=1-α)
		α - the level of significance (implicitly 0.05) x, y - the paired data sample

The two sample Wilcoxon test as a nonparametric equivalent of the two sample t-test

The test assesses the deviation between the median difference and the real constant *c*.

H	$\overline{H}$	Command in R
$x_{0.5} \leq y_{0.5}$	$x_{0.5} > y_{0.5}$	wilcox.test(x,y, alternative="greater", conf.level=1-α)
$x_{0.5} = y_{0.5}$	$x_{0.5} \neq y_{0.5}$	wilcox.test(x,y, alternative="two.sided", conf.level=1-α)
$x_{0.5} \geq y_{0.5}$	$x_{0.5} < y_{0.5}$	wilcox.test(x,y, alternative="less", conf.level=1-α)
		α - the level of significance (implicitly 0.05) x, y - the data samples

The Kruskal-Wallis test

It is a generalization of the Wilcoxon test. It is used as a nonparametric analogy to the one way ANOVA. We test the null hypothesis that all selections come from the same distribution.

Way	Command in R
The Kruskal-Wallis test	kruskal.test(quantitative variable ~ *factor, data)*

The median test

It is used as a nonparametric analogy to the one way ANOVA. We test the null hypothesis that all selections come from the same distribution.

Way	Command in R
Load the package The median test	Library(agricoale) Median.test(quantitative variable, *factor)*

The Friedman's test

A nonparametric analogy to the two way ANOVA. We test the null hypothesis that the distribution functions of variable X_{i1}…..X_{ik} are the same.

Way	Command in R
The Friedman's test or equivalent notation	Friedman.test(quantitative variable ~ factor A | factor B, data) Friedman.test(quantitative variable, factor A, factor B, data)

References

FIELD, Andy, Jeremy MILES and Zoë FIELD. Discovering Statistics Using R. SAGE, 2012. ISBN 978-1-4462-5846-0.

The journey begins by explaining basic statistical and research concepts before a guided tour of the R software environment. Next you discover the importance of exploring and graphing data, before moving onto statistical tests that are the foundations of the rest of the book (for example correlation and regression). You will then stride confidently into intermediate level analyses such as ANOVA, before ending your journey with advanced techniques such as MANOVA and multilevel models. Although there is enough theory to help you gain the necessary conceptual understanding of what you're doing, the emphasis is on applying what you learn to playful and real-world examples that should make the experience more fun than you might expect.

GLINER, JEFFREY A., GEORGE A. MORGAN, ROBERT J. HARMON and Robert J. HARMON. Introduction to Inferential Statistics and Hypothesis Testing. Journal of the American Academy of Child & Adolescent Psychiatry. 2000, 39(12), 1568–1570. ISSN 0890-8567. Available at doi:10.1097/00004583-200012000-00022.

Introduction to Inferential Statistics and Hypothesis Testing

LINTON, Oliver. Probability, Statistics and Econometrics. Academic Press, 2017. ISBN 978-0-12-810496-5.

The book covers much of the groundwork for probability and inference before proceeding to core topics in econometrics.

MATHEWS, Paul G. *Design of Experiments with MINITAB*. Milwaukee: ASQ Quality Press, 2005. ISBN 978-0-87389-637-5.

Paul Mathews presents the basic types and methods of designed experiments appropriate for engineers, scientists, quality engineers, and Six Sigma Black Belts and Master Black Belts.

OLHEDE, Sofia C. and Patrick J. WOLFE. The future of statistics and data science. *Statistics & Probability Letters* [online]. 2018, 136, The role of Statistics in the era of big data, 46–50. ISSN 0167-7152. Available at doi:10.1016/j.spl.2018.02.042.

The ubiquity of sensing devices, the low cost of data storage, and the commoditization of computing have together led to a big data revolution. We discuss the implication of this revolution for statistics, focusing on how our discipline can best contribute to the emerging field of data science.

REID, Nancy. Statistical science in the world of big data. *Statistics & Probability Letters*. 2018, 136, The role of Statistics in the era of big data, 42–45. ISSN 0167-7152. Available at doi:10.1016/j.spl.2018.02.049.

This essay considers the role of the statistical sciences in the world of big data, data science, machine learning, and artificial intelligence.

ITfeature.com
http://itfeature.com.
Basic Statistics and Data Analysis (Lecture notes, MCQS of Statistics).

Stack Exchange Network
https://stats.stackexchange.com.

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization.

Machine learning

Machine learning is a term closely associated with data science. It refers to a broad class of methods that revolve around data modelling to algorithmically make predictions ("supervised machine learning") or understand the underline structure of a dataset ("unsupervised machine learning"). Machine learning methods are considered very useful to handle Big Data and complex datasets.

Machine learning for making predictions

Core concept is to use tagged data to train predictive models. *Tagged data* means observations where ground truth is already known. *Training models* means automatically characterizing tagged data in ways to predict tags for unknown data points. Common methods for training models range is regression analysis.

Regression analysis

Very often we get into a situation where we need to determine whether the variables being monitored show a dependence. If the monitored variables are quantitative type, we can use:

- Correlation coefficient, which is
  - easy and fast calculation,
  - simple interpretability,
  - but it takes into account only one type of dependence (linear dependence),
- Regression analysis, which is
  - a universal tool that allows us to look at dependence using different types of functions (lines, parabolas, exponentials, etc.).

The relationship (dependence) between the observed variables can be expressed as a general function

$$y = f(x) + e, \tag{15}$$

where the function $f(x)$is unknown, x is an independent variable, y is a dependent variable and e represents random events (noise).

Examples of use of regression analysis are:

- How the size of household expenses for food depends on the number of households members,
- How the price of car depends on its age,
- Trends in time series.

To begin investigating whether or not there is a relationship between these two variables, we would begin by plotting these data points on a chart (a scatter plot), which would look like on Figure 12.

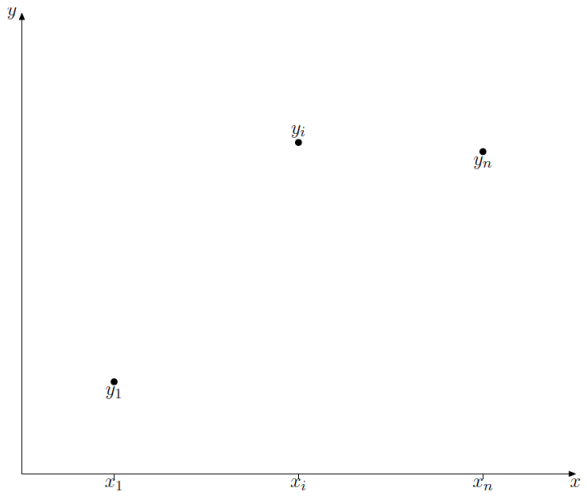

Figure 12 Scatter plot

The role of regression analysis is select for entered data (x_i, y_i), $i = 1, \ldots, n$, suitable function $\eta(x)$ so that this function reflects these entered data the best.

Regression functions (regression models) can be divided into two main types:

- Linear regression function (e.g. a regression line, etc.),
- Nonlinear regression function
    - Linearizable regression function which can be transform to a linear regression function. E.g. an exponential function, a power function, etc.
    - Non-linearizable regression function which cannot be transform to a linear regression function. There are a special approaches how to find out an unknown regression coefficients. E.g. S-curves.

We'll use a theoretical chart once more to depict what a regression line should look like.

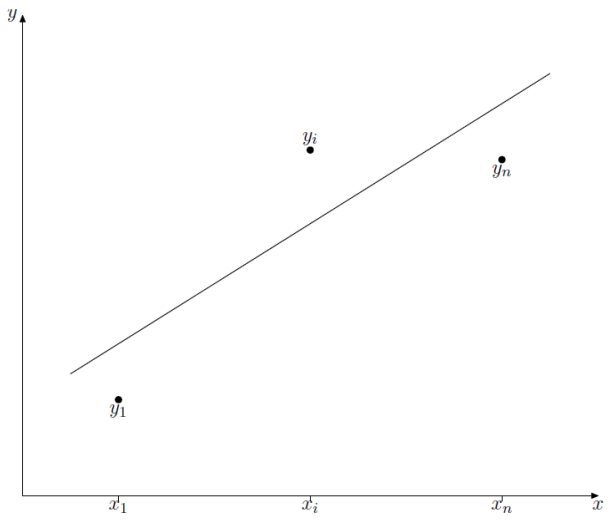

Figure 13 Scatter plot with a regression line

The formula of this regression line is

$$\eta(x) = \beta_1 + \beta_2 x \tag{16}$$

where β_1 and β_2 are regression coefficients.

These coefficients we can find out using the *least square method.* It works by making the total of the square of the errors e_i, see Figure 14, as small as possible (that is why it is called "least squares").

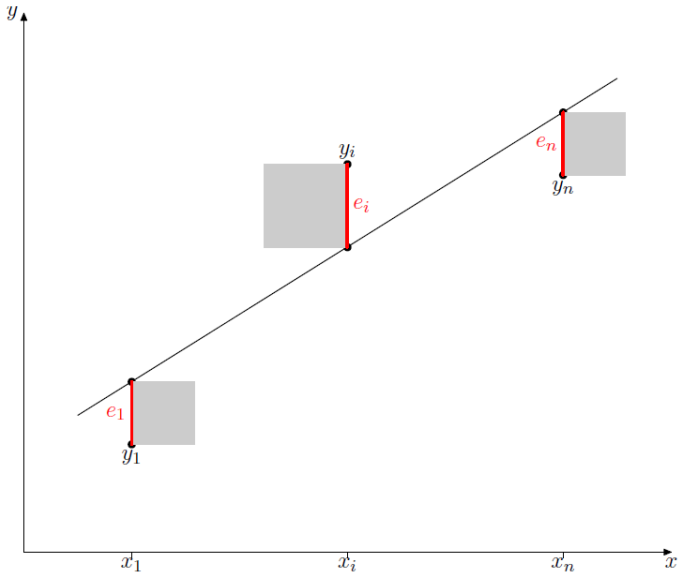

Figure 14 Scatter plot with squares

The least square method of a regression line leads to the system of normal equations.

$$\begin{aligned} n \cdot b_1 + \sum_{i=1}^{n} x_i \cdot b_2 &= \sum_{i=1}^{n} y_i, \\ \sum_{i=1}^{n} x_i \cdot b_1 + \sum_{i=1}^{n} {x_i}^2 \cdot b_2 &= \sum_{i=1}^{n} y_i x_i. \end{aligned} \tag{17}$$

By solving this system we get the estimates b_1 and b_2 of an unknown coefficients β_1 and β_2.

Some of nonlinear regression functions can be transformed into linear ones using a linearization.

Procedure to determine the coefficients of the linearizable function:

1. The transformation of a nonlinear regression function to a linear (e.g. regression line).
2. The determination of coefficients of linear regression function using the least square method.
3. Using back transformation of obtained coefficients we get coefficients of original nonlinear model.

Table 6 Linearization examples

Nonlinear function	y	x	a	b	Linear function
$u = c_1 e^{c_2 t}$	$\ln u$	t	$\ln c_1$	c_2	$y = a + bx$
$u = c_1 t^{c_2}$	$\ln u$	$\ln t$	$\ln c_1$	c_2	$y = a + bx$
$u = \frac{1}{c_1 + c_2 t}$	$\frac{1}{u}$	t	c_1	c_2	$y = a + bx$
$u = c_1 e^{\frac{c_2}{t}}$	$\ln u$	$\frac{1}{t}$	$\ln c_1$	c_2	$y = a + bx$

Linear regression function in R

Measured values

Number of members	Weekly cost [CZK]
1	490
2	820
3	1230
4	1570
5	1950
6	2320

Entering values

Way	Command in R
Loading form external file Transform table to matrix	*d* = read.table ("path/file_name.txt") *d* = as.matrix(*d*)

Data preparation

Way	Command in R
Identification of a independent variable Identification of a dependent variable	*x* = d[,1] *y* = d[,2]
Data	*data_name* = data.frame*(x,y)*

Graphical representation

Way	Command in R
Graphical representation of data	plot(*data_name*)

Linear model

Way	Command in R
Regression line $y = b_1 + b_2 . x$ Regression parabola $y = b_1 + b_2 . x + b_3 . x^2$	*model* = lm(*y~x*) *model* = lm(*y~x* + I(x^2))
Results	summary (*model*)

Nonlinear regression function in R

Measured values

Year	Profit [thous. CZK]
1	112
2	149
3	238
4	354
5	580
6	867

Entering values

Way	Command in R
Loading form external file	*d* = read.table ("path/file_name.txt")
Transform table to matrix	*d* = as.matrix(*d*)

Data preparation

Way	Command in R
Identification of a independent variable	*x* = d[,1]
Identification of a dependent variable	*y* = d[,2]
Data	*data_name* = data.frame*(x,y)*

Nonlinear model

Way	Command in R
Initializing nonlinear models	library(nls2)
Exponential function $y = b_1 e^{b_2 x}$	$model$=nls2$(y \sim b1 * \exp(b2 * x))$
Exponential function $y = b_1 b_2^x$	$model$=nls2$(y \sim b1 * (b2\hat{}x))$
Results	summary$(model)$

Examples of standard machine learning techniques

Regression trees and decision trees are becoming relevant instruments for predictions and classification aims. The concept of the trees is to partition the predictor space in distinct and non-overlapping regions. The method splits the most important variables in order of importance for predicting or classifying the target variable. The splitting creates new "leaves". The new "terminal nodes" are subsequently splitting too. The number of splitting can be fixed by the users. The advantages of using trees are principally its capacity to handle non-linear relationships, as the ones seen above, among the variables and the easy interpretability, see Figure 15. On the other side trees can overfit easily and have problems to handle a large amount of data.

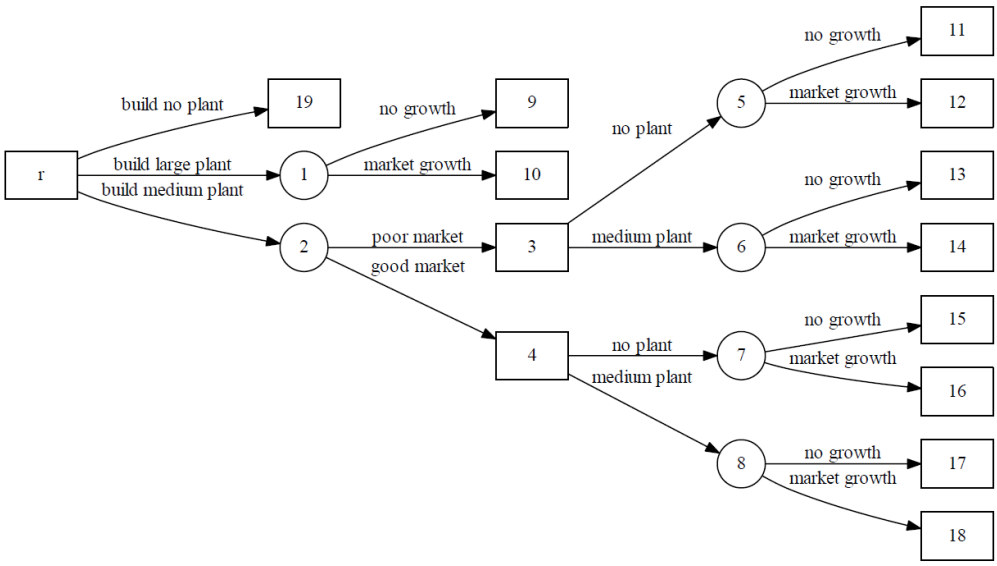

Figure 15 An example of Decision Tree (Source: Doubravsky and Dohnal, 2015)

Evaluations of realistic complex decision trees attract attention. Decision makers / field experts in their desperation to satisfy increasingly demanding laws and regulations (e.g. safety and environmental engineering) and/or pressure of competition (e.g. exchange rates hedging) are ready to believe that there is a theoretical answer to their needs. However, the only solution is to increase data/knowledge inputs into decision making processes. It means that no available information item may be ignored. Therefore known isolated fuzzy probabilities must be meaningfully incorporated into the decision making tasks.

The key reconciliation problem is the choice of the probabilities generation heuristic. If this heuristic is not accepted by a decision maker then some modifications of this heuristic are inevitable to cover specific requirements of the decision making problem under study. This is, however, an ad hoc procedure.

One of a new common sense heuristic is based on a strong analogy between a water flow through a one root tree system of pipes and the decision tree of the same topology. The heuristic solves decision problems under total ignorance, i.e. the decision tree topology is the only information available. However, isolated information items e.g. some vaguely known probabilities (e.g. fuzzy probabilities) are usually available. It means that a realistic problem is analysed under partial ignorance.

K-nn (K-Nearest Neighbors) is one of the most common machine learning methods used to separate in clusters the data. K-nn involves different

possibilities to measure the "distance" (the degree of differentiation or similarity) among the data points. K is the number of closest neighbours that the algorithm considers before passing to another cluster of similar points, see Figure 16.

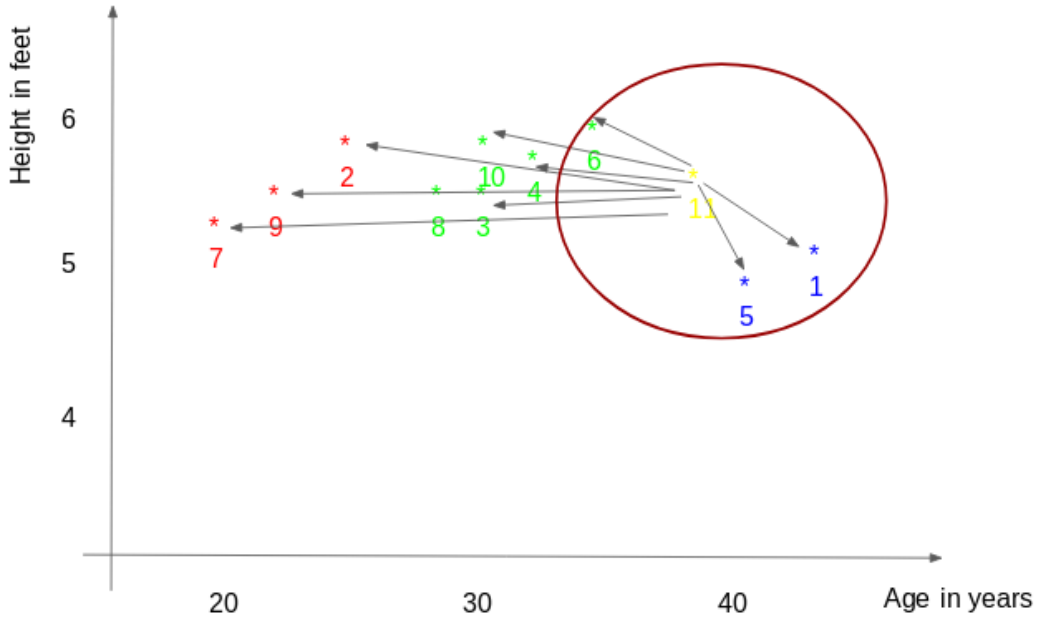

Figure 16 K-nn estimation (Source: https://www.analyticsvidhya.com)

Other very important methods are the "Shrinkage methods" like, Ridge and Lasso (Least Absolute Shrinkage and Selection Operator). These methods applied penalizations to the estimations in different ways. In particular Lasso is also a variable selection method to the estimations that can bring to solve multicollinearity problems.

A short introduction on ensemble methods

Ensemble methods are some machine learning methods applied especially to predictions that mostly, but not exclusively, involve regression or decision trees. These methods are used to overcome the risk of overfitting simple machine learning techniques like trees. The main idea is to bootstrap trees (or other simple techniques) and taking the average of the results obtained from the process. This allows researchers to handle a

higher dimensional space in terms of variables used. Some examples of ensemble methods are: Bagging, Random Forest, Boosting.

A good description of both simple and ensemble methods with applications with R can be found in *"An Introduction to Statistical Learning: with Applications in R"* by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) and the connected internet site of the book.

Time series analysis

A time series is a sequence of data points in a given interval of time. Normally, but not always, the intervals that separate the data points are equal. Most of the data available are yearly, quarterly and monthly data. However, certain financial data can be even weekly, daily or even intradaily data.

Components of the time series and Seasonality problem

The time series have four components (Figure 17). The secular trend is the long term tendency of the time series. The seasonal trend is a regular fluctuations observable in time series (i.e. the sales during Christmas). The cyclic movements are fluctuations that vary over time. This is normally the most studied component in economics. The last component is the residual term or irregular term that captures all the other events not observable in the other components.

In order to measure the time series without the seasonal trend, different techniques can be used to remove it what are called seasonal adjustment. An example is dividing the series by a proper seasonal index that takes into account the observations in normal times and applies a good differentiation process in case it is out of the average observations. This particular method is called multiplicative seasonal adjustment.

The time series can be divided into the short and long term time series. For the short-term time series, the trend can be determined using basic regression analysis methods. For the long-term time series, advanced methods are used, which are described in the text below.

Stationarity

Before deciding which kind of model could fit better, another important aspect to analyze is to determine if the model is stationary or non-stationary (see Figure 18) for an example of stationary time series and non-stationary time series). Indeed, most of the time series models work properly only in case of stationary time series and most of the real-life data in the world are not stationary.

The Augmented Dicky-Fuller (ADF) is a test used to measure the existence of non-stationary issues. If the test can reject the null hypothesis, the time series is stationary. If the test cannot reject the null, the test has a unit root and it is probably non-stationary. A unit root is an unpredictable stochastic trend and can change a shift in the distribution of the time series.

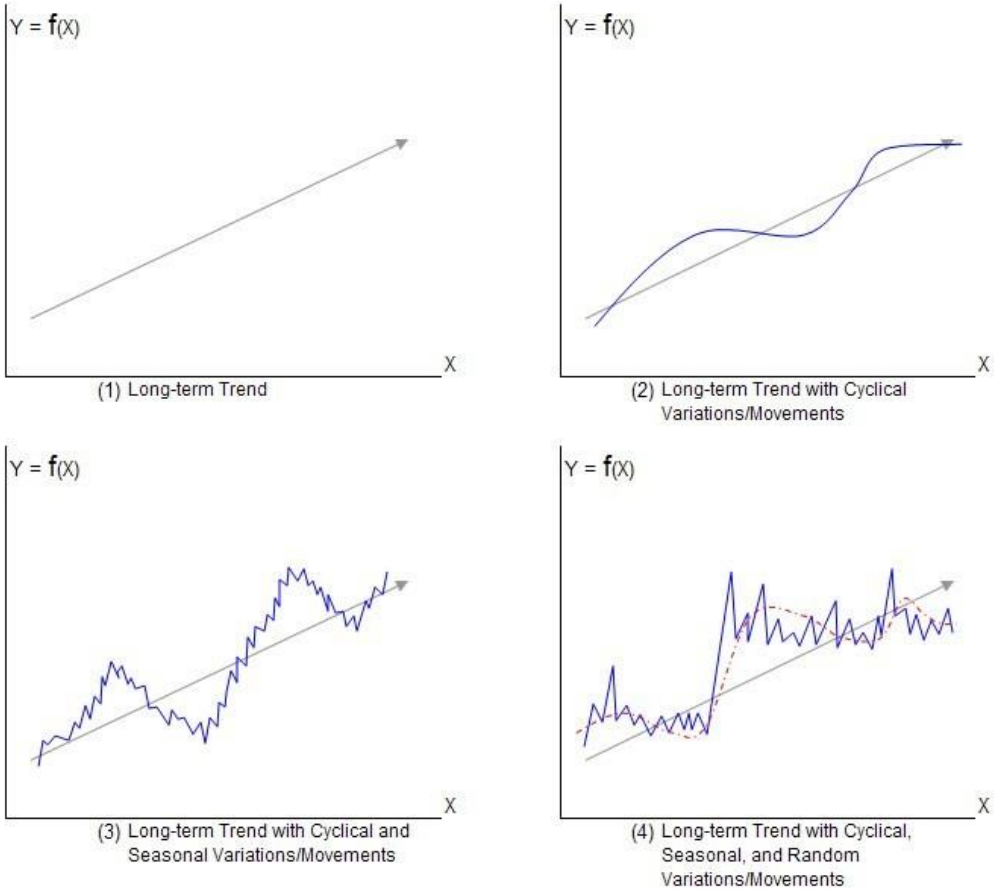

Figure 17 Components of time series (Source: http://itfeature.com)

Another possible test is the KPSS (Kwiatkowski–Phillips–Schmidt–Shin). In this case the interpretation of the result is the opposite of the ADF test. If I can reject the null, the time series is non-stationary. If I cannot reject, it is stationary. Using multiple tests for checking the stationarity of the time series is a good strategy for guarantee the robustness of the results.

Transformations

If time series are not stationary a proper transformation is needed. Some possible common transformations are taking the logarithm, the square root or differencing the data. The decision of what transformation is needed it is decided case by case after preliminary tests (like the ADF). Differencing the

data is a quite common tool and it can be used more than one time if the data required, for example a differentiation bigger than order one.

Example of differencing:

$$y_t' = y_t - y_{t-1}, \tag{18}$$

the series represents now the change between consecutive observations of our time series.

Most of the time the differencing method is used with the logarithm transformation in order to obtain more smooth results.

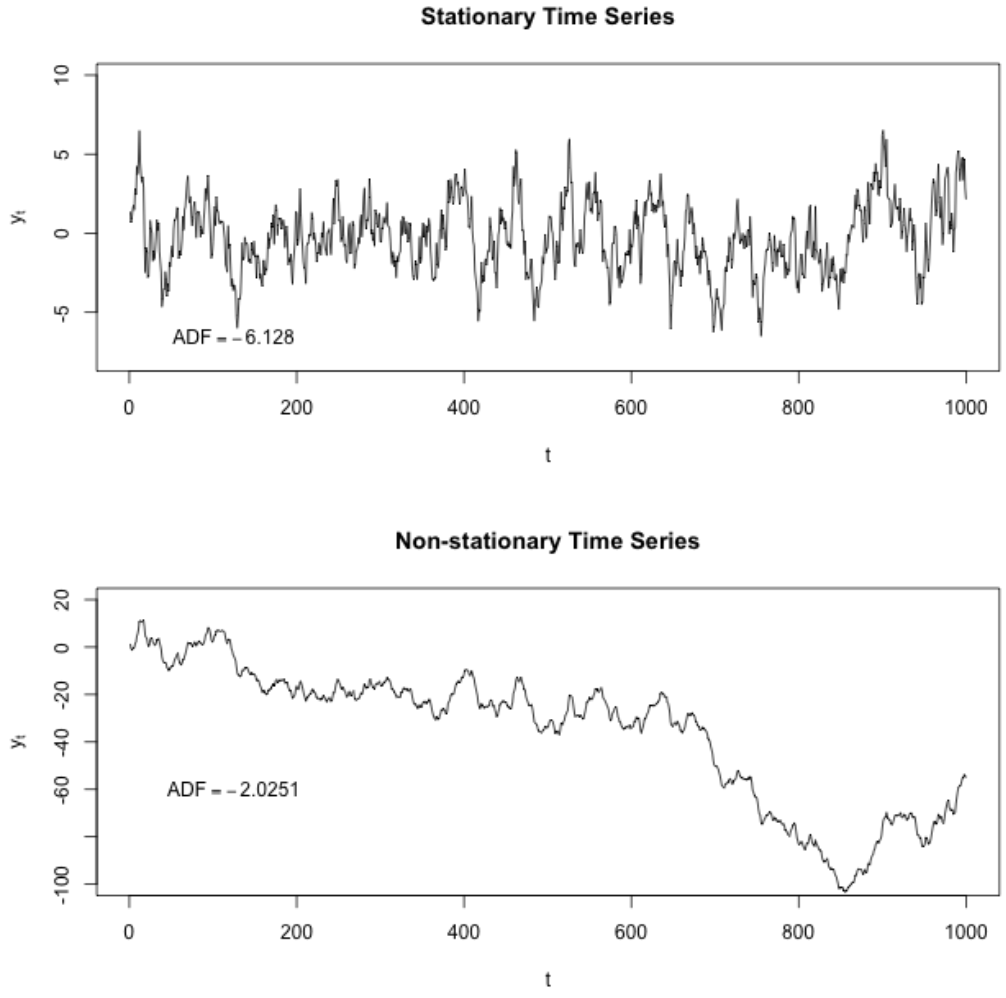

Figure 18 Stationary and non-stationary time series examples (Source: https://stats.stackexchange.com)

ARMA and ARIMA models

A univariate analysis of a time series involves only one variable and its lags. A multivariate analysis instead consists on the analysis of the effects of different variables on another variable(s) over time.

A standard model used for univariate time series analysis is the ARMA model (and is differentiated version, the ARIMA). The autoregressive model (AR) is simply the linear regression of one variable and one or more of its lags. The number of the lags takes into account is the order p of the AR model.

An AR(p) model can be defined in the following way:

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + e_t, \quad (19)$$

where y_t is the conditional mean of the variable of interest and $\phi_1 y_{t-1} + \cdots + \phi_p y_{t-p}$ are the past observations with their parameters and e_t is a white noise. A white noise is a random process of random numbers that are uncorrelated with mean zero and finite variance.

The moving average (MA) is the other part of the ARMA model. In this case the linear regression is on the white noise error term and its lags on the right side of the equation. The distribution is normally assumed normal distributed. The order of MA is defined by a q.

A typical MA(q) can be described as:

$$y_t = c + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}, \quad (20)$$

where $\theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$ are the past white noises and their parameters.

The autocorrelation and partial autocorrelation (ACF and PACF) can help to understand if the better model to use should be a AR(p), a MA(q) or their combination, the ARMA(p,q) model. Autocorrelation is also known as serial correlation and denotes the presence of a correlation between the variable and its lags. Partial autocorrelation is the autocorrelation after removing any linear dependence. The Figure 19 shows a possible example of correlograms, the typical tools to visualize ACF and PACF.

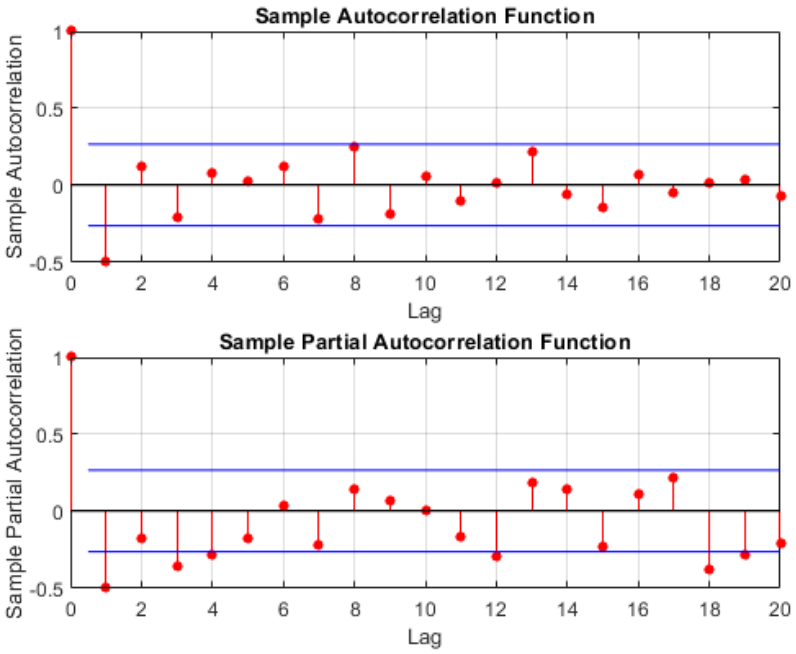

Figure 19 Correlograms of ACF and PACF (Source: mathworks.com)

The results of the correlograms, see Figure 19, illustrate significant autocorrelation. The ACF shows significant autocorrelation at lag 1. The PACF instead illustrates that lags 1, 3, and 4 have significant autocorrelation. The fast decrease of the ACF and the more gradual situation in the PACF combined could suggest to use a MA(1) model.

ARIMA is an ARMA model that is differentiated to make it stationary where the "I" stated for "integrating". The ARIMA model has not only the p and q orders of the ARMA model but also the value d that defined how many times the time series are differentiated.

Smoothing techniques

Smoothing techniques are useful to remove from the data random variations. This is useful to understand better the different components of the time series. A simple method is to take the average of the values of the different data points. This can eventually work if we assume that the data has no trend. Moreover, it can be useful in certain cases to give different weighs? in different time. Instead, the average gives the same weighs? all the times. Exponential smoothing (Figure 20) instead implies different weighs? in different time. More specifically recent data received higher weighs? than the oldest ones. The exponential smoothing method can be simple, double or triple. The double is used in case of presence of a trend in the time series. The triple if both trend and seasonality is detected.

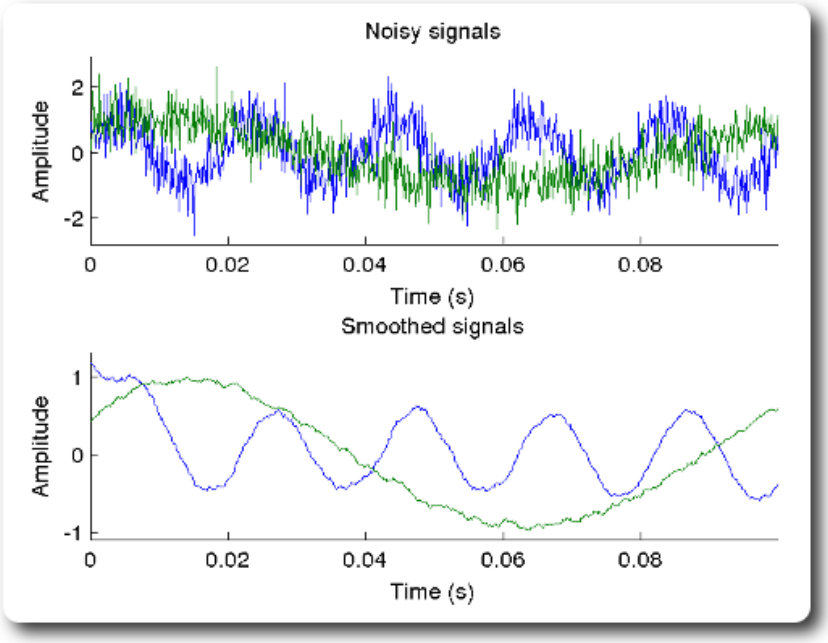

Figure 20 Exponential smoothing (Source: mathworks.com)

A short introduction on multivariate time series analysis

The multivariate time series is when we analyse more than one time series and its lags but many time series. For example, I want to forecast the unemployment rate of a country taking into account its inflation and industrial production. The most common models to make predictions are the vector autoregression (VAR), its Bayesian version (BVAR) and the factor models.

An example of a VAR(p) model can be the following:

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-1} + e_t, \tag{21}$$

where $y_t$is the vector of response (or target) time series variables with n elements at time t. *c* is a constant vector with *n* elements. φ_i are *n*-by-*n* autoregressive matrices for each *i*. The *p is the number of* autoregressive matrices. Some of those can be completely composed of zeros. ε_t is a vector of serially white noise with length *n*.

These models are more complex to treat given the intricacy of the process under observation.

R commands for time series analysis

Command in R	Test/model/operation
as.ts(x, …)	Coerce an object in a time series. Different option are possible, like frequency, start and end date of the series
log(x)	Logarithm transformation
diff(x, …)	Differencing. Possible insert as options which lag using of differencing and the order of the difference
adf.test(x,…)	Augmented Dicky-Fuller test with possible options *as the lag order with default to calculate the test statistic.*
kpss.test(x,…)	Kpss test with options as the possibility to indicate if the null hypothesis should be on level or trend
acf(x,..) and pacf(x,..)	Compute autocorrelation and partial autocorrelation. In the options possibility to plot the result. Other more complex plots can be found in the package corrgram
auto.arima(x,…)	This is the one of the most common command for fit the best arima model to univariate time series. Many options available
es(data,..)	Exponential smoothing command. Many options available.

References

BROCKWELL, Peter J. and Richard A. DAVIS. Introduction to Time Series and Forecasting. 2nd. ed. New York: Springer-Verlag, 2002. Springer Texts in Statistics. ISBN 978-1-4757-7750-5. Availabe at https://www.springer.com/gp/book/9781475777505.

The core of the book covers stationary processes, ARMA and ARIMA processes, multivariate time series and state-space models, with an optional chapter on spectral analysis. Additional topics include harmonic regression, the Burg and Hannan-Rissanen algorithms, unit roots, regression with ARMA errors, structural models, the EM algorithm, generalized state-space models with applications to time series of count data, exponential smoothing, the Holt-Winters and ARAR forecasting algorithms, transfer function models and intervention analysis.

DAS, Sourish, Sasanka ROY and Rajiv SAMBASIVAN. Fast Gaussian Process Regression for Big Data. Big Data Research. 2018, 14, 12–26. ISSN 2214-5796. Available at doi:10.1016/j.bdr.2018.06.002.

Gaussian Processes are widely used for regression tasks. A known limitation in the application of Gaussian Processes to regression tasks is that the computation of the solution requires performing a matrix inversion. The solution also requires the storage of a large matrix in memory. These factors restrict the application of Gaussian Process regression to small and moderate size datasets. We present an algorithm that combines estimates from models developed using subsets of the data obtained in a manner similar to the bootstrap.

DOUBRAVSKY, Karel and Mirko DOHNAL. Reconciliation of Decision-Making Heuristics Based on Decision Trees Topologies and Incomplete Fuzzy Probabilities Sets. *PLoS ONE*. 2015, 10(7), e0131590. Available at doi:10.1371/journal.pone.0131590.

This paper presents a relatively simple algorithm how some missing III (input information items) can be generated using mainly decision tree topologies and integrated into incomplete data sets. The algorithm is based on an easy to understand heuristics, e.g. a longer decision tree sub-path is less probable. This heuristic can solve decision problems under total ignorance, i.e. the decision tree topology is the only information available. But in a practice, isolated information items e.g. some vaguely known probabilities (e.g. fuzzy probabilities) are usually available. It means that a realistic problem is analysed under partial ignorance. The proposed algorithm reconciles topology related heuristics and additional fuzzy sets using fuzzy linear programming.

GUJARATI, Damodar N. *Basic econometrics*. McGraw Hill, 2003. ISBN 978-0-07-112342-6.

Basic Econometrics provides an elementary but comprehensive introduction to econometrics without resorting to matrix algebra, calculus, or statistics beyond the elementary level.

GUJARATI, Damodar N. and Dawn C. PORTER. Essentials of Econometrics. McGraw-Hill Education, 2009. ISBN 978-0-07-337584-7.

The primary objective of the fourth edition of Essentials of Econometrics is to provide a user-friendly introduction to econometric theory and techniques. This text provides a simple and straightforward introduction to econometrics for the beginner. The book is designed to help students understand econometric techniques through extensive examples, careful explanations, and a wide variety of problem material.

HYNDMAN, Rob J. and George ATHANASOPOULOS. Forecasting: principles and practice. OTexts, 2018. ISBN 978-0-9875071-1-2.

This textbook provides a comprehensive introduction to forecasting methods and presents enough information about each method for readers to use them sensibly.

COWPERTWAIT, Paul S. P. and Andrew V. METCALFE. *Introductory Time Series with R.* Springer Science Business Media, 2009. ISBN 978-0-387-88698-5.

This book gives you a step-by-step introduction to analysing time series using the open source software R. Each time series model is motivated with practical applications, and is defined in mathematical notation. Once the model has been introduced it is used to generate synthetic data, using R code, and these generated data are then used to estimate its parameters.

SHUMWAY, Robert H. and David S. STOFFER. Time Series Analysis and Its Applications: With R Examples. New York: Springer New York, 2010. ISBN 978-1-4419-7864-6.

This book presents a balanced and comprehensive treatment of both time and frequency domain methods with accompanying theory.

ITFeature.com
http://itfeature.com.

Basic Statistics and Data Analysis (Lecture notes, MCQS of Statistics).

Stack Exchange Network
https://stats.stackexchange.com.

Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization.

Machine learning and Big Data in economics and econometrics: the next frontier

The studies on the use of Machine learning (ML) methods and Big Data in economics and econometrics at the moment are interested to understand principally what are the limits and potentialities for economic predictions and analysis.

A major concern is that the methods to understand the significance of the results could be no more valid with Big Data. The main idea is that Big Data are pushing the economic studies to focus more on the "economic significance" and related less to classical methods to evaluate "statistical significance" of the results. Moreover, some recent and very advanced studies are trying to create new methods to validate robust results using different datasets as sources.

Another issue is the one related to the bias of ML methods in analysis of causal inference. Currently, most of the economic and econometric studies that exploit ML methods are either forecasting or descriptive studies. Some researchers are studying possible solutions to solve the issue of the standard bias of ML methods. In particular, Lasso and Post-Lasso methods seem the more relevant for this scope.

ML methods are attracting scholars not only for handling Big Data better than standard econometric methods but also for their ability to detect and analyse non-linear relationships among variables. In the past attempts to create non-linear problem failed because the results were not particularly different than standard linear models or in any case difficult to interpret. A part of the state-of-the-art literature is exploring how Big Data and ML methods could make non-linear relationships a relevant topic again.

Finally, text mining and the availability of the Unstructured Big Data (texts, video, audio, pictures) gave new opportunities to the scholars. The effects of the decisions and the communications from policy-makers are some of the topics related to these new fields of research.

References

ATHEY, Susan. The Impact of Machine Learning on Economics. *The Economics of Artificial Intelligence: An Agenda*. 2018, 507–547.

This paper provides an assessment of the early contributions of machine learning to economics, as well as predictions about its future contributions.

CHEN, Jeffrey C., Abe DUNN, Kyle K. HOOD, Alexander DRIESSEN and Andrea BATCH. Off to the Races: A Comparison of Machine Learning and Alternative Data for Predicting Economic Indicators. In: *NBER Chapters: National Bureau of Economic Research, Inc*, 2019. Available at https://ideas.repec.org/h/nbr/nberch/14268.html.

In this paper, we explore how ML and alternative data sources can play a role in producing official national statistics. We consider the case of reducing revisions to the services portion of Personal Consumption Expenditures (PCE Services) by way of predicting the U.S. Census Bureau's Quarterly Services Survey (QSS).

CHERNOZHUKOV, Victor, Denis CHETVERIKOV, Mert DEMIRER, Esther DUFLO, Christian HANSEN, Whitney NEWEY and James ROBINS. *Double/Debiased Machine Learning for Treatment and Causal Parameters*. arXiv:1608.00060. 2016. Available at http://arxiv.org/abs/1608.00060.

Most modern supervised statistical/machine learning (ML) methods are explicitly designed to solve prediction problems very well. Achieving this goal does not imply that these methods automatically deliver good estimators of causal parameters. Specifically, we can form an orthogonal score for the target low-dimensional parameter by combining auxiliary and main ML predictions. The score is then used to build a de-biased estimator of the target parameter which typically will converge at the fastest possible 1/root(n) rate and be approximately unbiased and normal, and from which valid confidence intervals for these parameters of interest may be constructed. The resulting method thus could be called a "double ML" method because it relies on estimating primary and auxiliary predictive models. In order to avoid overfitting, our construction also makes use of the K-fold sample splitting, which we call cross-fitting. This allows us to use a very broad set of ML predictive methods in solving the auxiliary and main prediction problems.

COULOMBE, Philippe Goulet, Maxime LEROUX, D. Miroslav STEVANOVIĆ and Stéphane SURPRENANT. *How is Machine Learning Useful for Macroeconomic Forecasting*? Working paper. 2019. Available at https://www.semanticscholar.org/paper/How-is-Machine-Learning-Useful-for-Macroeconomic-%E2%88%97-Coulombe-Leroux/4b7704f15f05195e791edc5e3486dc0751d9b551.

The current forecasting literature has focused on matching specific variables and horizons with a particularly successful algorithm. To the contrary, we study a wide range of horizons and variables and learn about the usefulness of the underlying features driving ML gains over standard macroeconometric methods.

FERRARA, Laurent, Massimiliano MARCELLINO and Matteo MOGLIANI. Macroeconomic forecasting during the Great Recession: The return of non-linearity? International Journal of Forecasting. 2015, 31(3), 664–679. ISSN 0169-2070. Available at doi:10.1016/j.ijforecast.2014.11.005.

The debate on the forecasting ability of non-linear models has a long history, and the Great Recession episode provides an interesting opportunity for a re-assessment of the forecasting performances of several classes of non-linear models.

HANSEN, Stephen, Michael MCMAHON and Andrea PRAT. Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. dp1276. The Quarterly Journal of Economics, 133(2), *801–870.* Available at doi: 10.1093/qje/qjx045.

How does transparency, a key feature of central bank design, affect the deliberation of monetary policymakers? We exploit a natural experiment in the Federal Open Market Committee in 1993 together with computational linguistic models (particularly Latent Dirichlet Allocation) to measure the effect of increased transparency on debate. Commentators have hypothesized both a beneficial discipline effect and a detrimental conformity effect. A difference-in-differences approach inspired by the career concerns literature uncovers evidence for both effects. However, the net effect of increased transparency appears to be a more informative deliberation process.

JOSEPH, Andreas. Shapley regressions: A framework for statistical inference on machine learning models. arXiv:1903.04209. 2019. Available at http://arxiv.org/abs/1903.04209.

This paper proposes the Shapley regression framework as an approach for statistical inference on non-linear or non-parametric models.

REICHLIN, L., C. DE MOL, E. GAUTIER, D. GIANNONE, S. MULLAINATHAN, H. VAN DIJK and J. WOOLDRIDGE. Big data in economics: evolution or revolution? In: L. MATYAS, ed. Economics without borders. Cambridge: Cambridge University Press, 2017, s. 612–632. ISBN 978-1-316-63640-4. Available at doi: 10.1017/9781316636404.

The Big Data Era creates a lot of exciting opportunities for new developments in economics and econometrics. At the same time, however, the analysis of large datasets poses difficult methodological problems that should be addressed appropriately.

VARIAN, Hal R. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives. 2014, 28(2), 3–28. Available at doi: 10.1257/jep.28.2.3.

Computers are now involved in many economic transactions and can capture data associated with these transactions, which can then be manipulated and analyzed. Conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships.
