Mathematics for Big Data

Master in Modelling for Science and Engineering

UAB
Universitat Autònoma
de Barcelona

Statistics at the heart of Scientific Method

Statistics is the science of extracting information from data.

As such, statistical methods are intimately interwoven with the ideas of the methods used to formulate scientific explanations of natural phenomena.
So let's reformulate the previous definition:

Statistics is about answering questions based on

- data
- models
- a quantification of the link between them (*likelihood*).

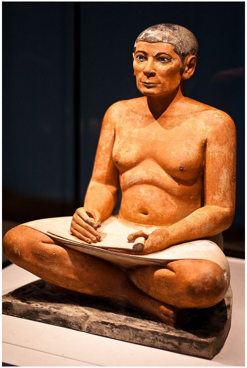
Data can have different nature and complexity, according to the way they are obtained and displayed, $e.g$, multidimensional data, functional data, dependent data, time-dependent data...

Data! Data! Data! he cried impatiently!
I can't make bricks without clay!

Sherlock Holmes [1]

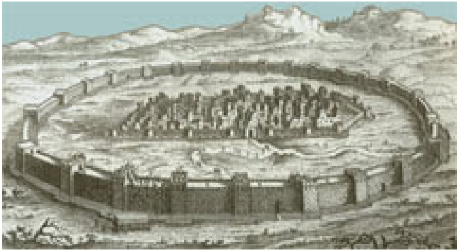
[1] The adventure of the copper beeches, Sir A.Connan Doyle

Statistics is a very old discipline.



The egypicians (2400-2600 AC) collected data on harvests, Nile river swells, population...

In the siege of Platea (429-427 AC) (Peloponnesian wars), the attackers calculated the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.



Flying bomb hits on London[2] The following data give the number of flying bomb hits recorded in each of 576 small areas of $\frac{1}{4}km^2$ in the south of London during World War II.

Example of some of the data used in the Bomb Sight project
1940s bomb census images reproduced by permission of The National Archives, London, England

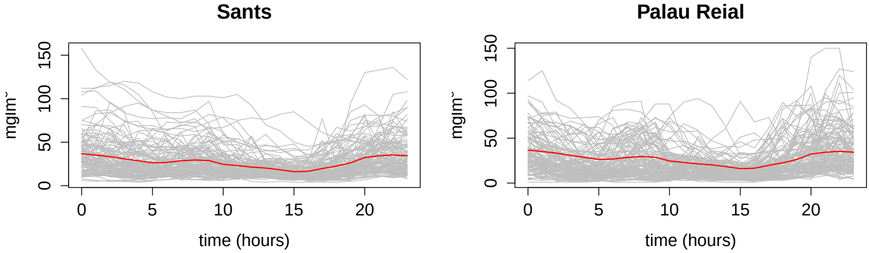
Flying bomb hits on London

Number of impacts	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be randomly distributed over the area and should therefore be fitted by a Poisson distribution. Is this the case?

[2] http://bombsight.org/data/

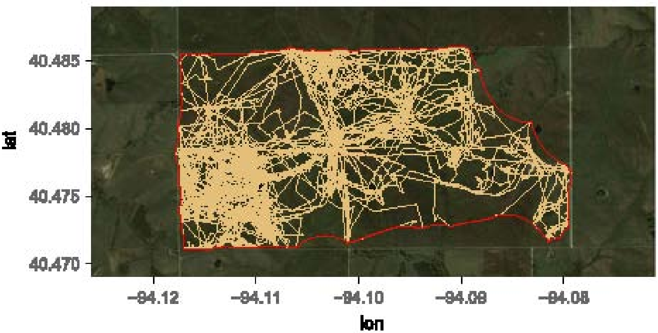
Levels of pollutants in non-working days in Sants and Palau Reial in 2014 in gray, and the respective pointwise mean functions in red[3].



Were these neighbourhoods equally polluted?

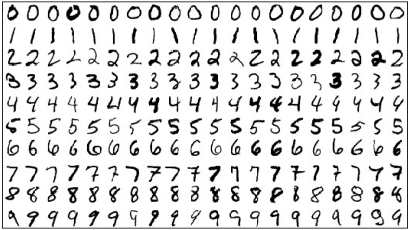
[3] A. Cabaña, A. Estrada, J. Peña and A. Quiroz (2016) *Permutation tests in the two-sample problem for functional data*

Imagine a bison wandering in Missouri meadows. Thanks to a GPS, we have a register of its whereabouts. Here is a picture of his trajectory.



Ecologists are interested in determining the *home range* of the species. The statistical problem consists in estimating a **set** from the data collected.

Handwritten digits



- You are provided a dataset containing images (16x16 grayscale image) of digits
- Each image contains a single digit
- Each image is labelled with the corresponding digit
- Think of each image as a vector $X \in R^{256}$ and the label as a scalar $Y \in \{0, \ldots, 9\}$
- With a large enough sample, we should be able to identify/predict digits

Gene expression data[4]: rows = genes, columns = sample

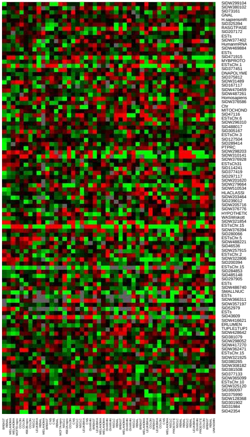


FIGURE 1.3. *DNA microarray data: expression matrix of* 6830 *genes (rows) and* 64 *samples (columns), for the human tumor data. Only a random sample of* 100 *rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*

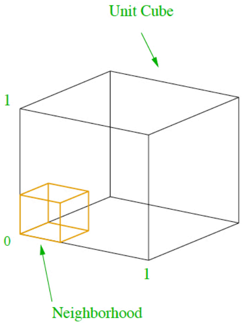
- DNA microarrays measure the expression of a gene in a cell.
- Nucleotide sequences for a few thousand genes are printed on a glass slide.
- Each spot contains millions of identical molecules which will bind a specific DNA sequence.
- A target sample and a reference sample are labeled with red and green dyes, and each are hybridized with the DNA on the slide.
- Through fluoroscopy, the log (red/green) intensities of RNA hybridizing at each site is measured.

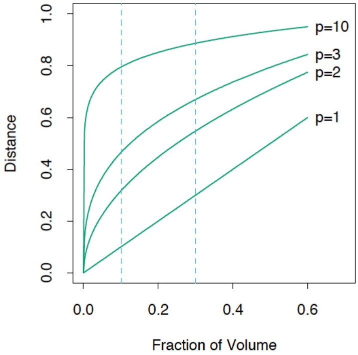
Question: do certain genes show very high (or low) expression for certain cancer samples?

[4] *The Elements of Statistical Learning*,T. Hastie, R. Tibshirani, J. Friedman, Springer(2001)

Classical statistical methods deal with *large* samples (n) with a *small* number of variables (p)

A few years ago, people began to work wit $p >> n$, and the *curse of dimensionality* entered the scene: Consider a hypercube with sides of length c along the axes in a unit hypercube. Its volume is c^p. To capture a fraction r of the unit hypercube, c must be $r^{1/p}$.





If $p = 10$, in order to capture (only) 10% of the volume, we need $c \approx 0.8$!

A linear regression problem: suppose we try to use linear regression to estimate Y (response) using X (predictors)

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- Classical statistical theory guarantees (under certain hypotheses) that we can recover the regression coefficients β if n is large enough (consistency).
- In modern problems n/p is often small.
- What if we assume only a small percentage of the "true" coefficients are nonzero?
- Obtain consistency results when $p, n \to \infty$ with $n/p = \mathrm{constant}$.
- How do we identify the "right" subset of predictors?
- We can't examine all the $\binom{p}{k}$ possibilities! For example, $\binom{1000}{25} \approx 2.7 \times 10^{49}$!

What's Big Data?

- It is a data set which is too large to fit into the computer's memory in one go[5]. See other possible definitions in the same paper.
- Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software[6].
- Gartner's " 3Vs": Big data is high volume (amount of data), high velocity (speed of data in and out), and high variety (range of data types and sources).
- The size in gigabytes, terabytes (10^3 gigabytes), petabytes (10^3 terabytes)... of a data set to be consider a *Big* Data set is changing constantly.

[5] Hand (2013), *Data, not Dogma: Big Data, Open Data and the Opportunities Ahead*, in LNCS, IDA XII, 1-12

[6] Snijders, et.al, (2012), *Big Data, Big gaps of knowledge in the field of internet science*, Intl. Jr. of Internet Science 7, 1-15

Where does Big Data come from?

- Automatic data capture as a side effect of other human activity, such as credit card transaction data, supermarket purchase pattern data, web search data, phone call data, and social network data, using tools such as Facebook and Twitter.
- Automatic monitoring of mechanical devices (such as aircrafts).
- Complex scientific experiments: Large Hadron Collider at CERN (Higgs' Bosson), astronomical data that are recorded automatically by powerful telescopes (Sloan Digital Sky Survey, Large Synoptic Survey Telescope), complex model simulations (Physics, Meteorology, Finance, ...).
- Satellite images registered every few seconds, Genomics, medical imaging (brain images), ...

In what sense is Big Data *big*?

- A classical data frame in Statistics is a matrix with dimensions

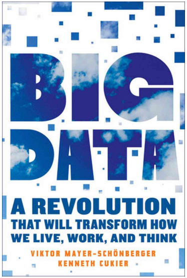
  n number of observations,
  p number of data (*variables*) recorded for each observation,

- A third dimension d can be added, indicating what is the dimension of each element in the data frame. For instance, if each observation is an image, p is the number of pixels in the image, which in turn are decribed by their RGB levels ($d = 3$).
- d is a measure of the data complexity.

A Big Data set should have at least one of these characteristics:

large n, large p, large d

Beware Big data Gurus!



- This is not an academic text.
- It contains a happy-ending business stories that stimulate the reader to imagining business oportunities related to big data.
- They advocate for black-box predictive models rather than substantive models.
- the important thing is to know *what* , not *why*.

Big Data big misconceptions!

- $n = all$ *In the era of Big Data we can analyse far more data, even all of it.* **Watch out!** - Possible bias selection
- *Sampling is an artifact of a period of information scarcity*
- *Messy (dirty) data are acceptable. Given that now we have no sampling error we can accept a larger measurement error.* **Watch out!** - A small carefully designed sample can be far better than a dirty Big Data set.
- *The end hypothesis testing. Before big data, the data analysis was limited to test a small number of hypothesis defined before the data collection. When we let speak the data, we discover unexpected knowledge.* **Watch out!** - The data do not speak! This is the end of the scientific method.

David Hand's "Data, not Dogma..."

In his 2013 paper, Hand (Department of Mathematics, Imperial College, London U.K.) gives a critical view of several aspects of Big Data.

- *No one wants data [...] what people want are answers. Data are of value only to the extent that they can lead to answers...*
- The extraction of information from data generally requires more than trivial exercises of data manipulation. It requires [analysis] statistical inference, machine learning, pattern recognition, data mining...
- Inference from the data we have to the data we might have had or might have in the future is a non-trivial exercise which requires very deep theory.

Beware Big data Gurus 2

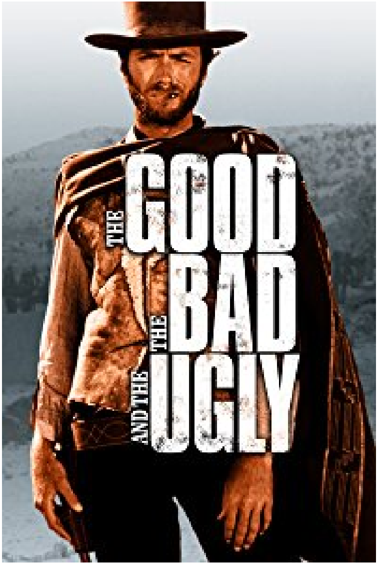
Some Big Data gurus state that when you have so huge amount of data you do not need any model. Hand does not agree:

The numbers don't speak for themselves. [A model is needed to] turn them into sounds.

He distinguishes two kinds of models:

- substantive models based on a theory about the mechanism or process underlying the phenomenon.
- Empirical models based solely on data, are very useful for prediction but essentially useless for enlightening us about underlying mechanisms.

Substantive models allow for generalisation to other contexts much more than empirical models. *[All] models must be wrong if they are to be in use. But they have to be wrong in the right way, and that's why data analysis is a skilled profession.*



https://www.youtube.com/watch?v=h_zeiKrRTuk&ytbChannel=null

The Good

- With a large enough data set, any slight [peculiarity] in the underlying generating mechanism [...] is likely to be statistically significant, for instance, Higgs' bosson has been found because the null hypothesis was rejected in a statistic test
- Telefonica's *Dynamic Insights* collects mobile data, anonymised and aggregated, to trace trends and the behaviour of crowds, not individuals.

  `https://www.youtube.com/watch?v=5VQVL9cctLE`

The Bad

- Algorithms will be able to predict the likelihood that one will get a heart attack (and pay more for health insurance), default on a mortgage (and be denied a loan), or commit a crime (and perhaps get arrested in advance)
- People might think twice before visiting websites of extreme sports or watching sitcoms glorifying couch potatoes if they felt this might result in higher insurance premiums.
- In the U.S.A and Britain, drivers can buy a car insurance priced according to where and when they actually drive.
- In the Big Data era, people and regions that are not well represented in the digital mainstream will be ignored not only by the market but also by the decision-makers.
- The observational nature of most Big Data implies that selection bias is likely to be present.

The Ugly

- Statistics will be forced to deal with problems of scale in order to remain relevant.
- The field of Computer Science is also currently poorly equipped to provide solutions to the inferential problems associated with Big Data[7].
- Database researchers rarely view the data in a database as noisy measurements on an underlying population about which inferential statements are desired.
- Instead of using a single algorithm for solving an inference problem, a hierarchy of algorithms is considered that are ordered by computational complexity. For small sample sizes complex algorithms (high quality results, low computational eficiency) are preferred, but when sample size increases the option are algorithms that run more quickly, even if they deliver poorer quality results. Can this trade-off be optimised?

[7] Jordan (2013), *On statistics, computation and scalability*, Bernoulli 19, 1378-1390