Introduction to Statistical Inference

Lecture 1: Course introduction

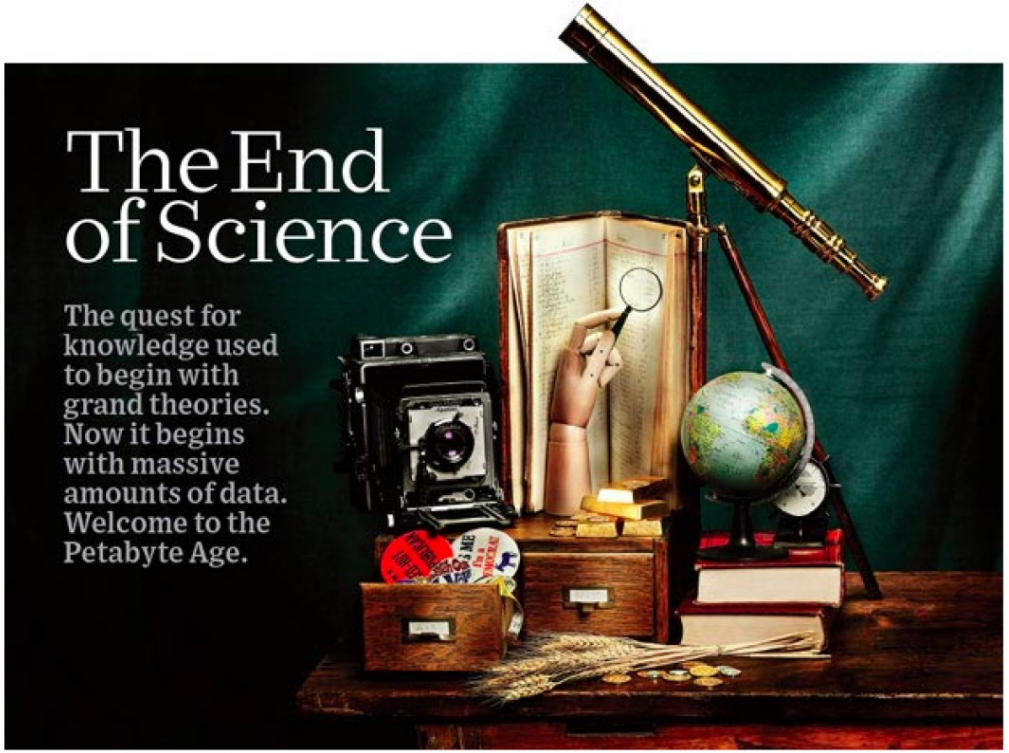
Mohammad-Reza A. Dehaqani

dehaqani@ut.a

We live in the era of data

New role of data in business

New role for data in science



Wired Magazine, issue 16.07

This is the religion of big data. No need to ask questions, just collect lots of data and let it speak.

Gil Press,
Forbes, September 2014

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

There are also signs of trouble...

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

(Submitted on 21 Jul 2016)

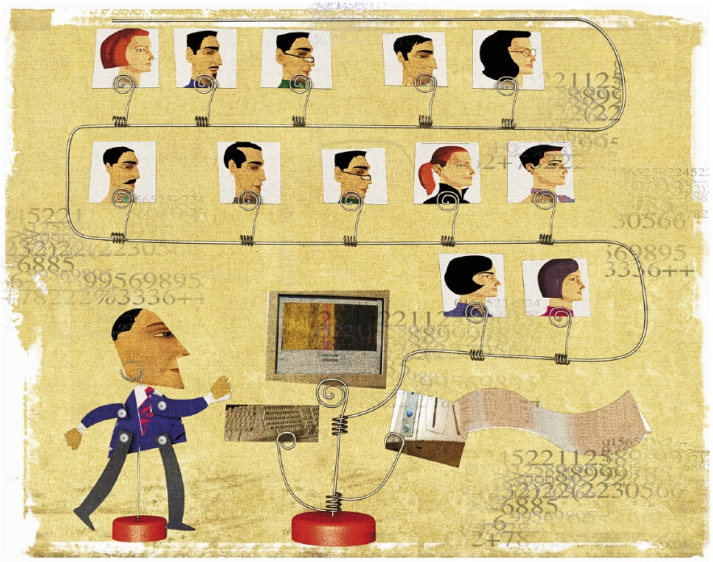
Technology

Liking curly fries on Facebook reveals your high IQ

By PHILIPPA WARR

Tuesday 12 March 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".



POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2]* Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

Crisis of science?

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

DECEMBER 13, 2010

Scientific knowledge based only upon inductive reasoning and careful observation of events in nature

Portrait by Pourbus the Younger, 1617

Lord High Chancellor of England

In office

7 March 1617 – 3 May 1621

The problem of induction

David Hume (1711 – 1776)

- The problem of induction: no matter how many instances of white swans we might have observed, this does not justify the conclusion that all swans are white
- natural instinct, rather than reason, explains the human practice of making inductive inferences.

Learning from data is not easy matter

One example; Educational level in US workforce

Suppose we are interested in learning about the level of education of individuals employed in the US in January 2018.

- We decide to describe level of education in terms of number of completed years in any educational institution
- The **population** of people employed in the US in January 2018 includes $N = 251{,}000{,}000$ individuals.
- Each individual $j \in \{1, \ldots, N\}$ has a certain level of education x_j

$$x_1 \;\; x_2 \;\; \cdots \;\; x_N$$

- The N values $x_1, \ldots, x_N$ will not be all different. Let's say that the number of years of education ranges from 0 to 30, for a total of $K = 31$ distinct values. We can describe the **distribution** of years of education using the distinct values $\xi_1, \ldots, \xi_K$ and their frequencies $f_1, \ldots, f_K$, with $\sum_{k=1}^{K} f_k = N$.

ξ_1	ξ_2	$\cdots$	ξ_K
f_1	f_2	$\cdots$	f_K

Educational level in US workforce

If we were to do a **census** and query every worker in the US, we would know the distribution

ξ_1	ξ_2	$\cdots$	ξ_K
f_1	f_2	$\cdots$	f_K

These are a lot of numbers and we might really want to summarize them in meaningful ways

- The **population mean** μ tells us about the "typical" level of education:

$$\mu = \frac{1}{N}\sum_{j=1} x_j = \frac{1}{N}\sum_{k=1}^{K} \xi_k f_k$$

- The **population variance** σ^2 tells us about how much variability there is around this typical level

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} (x_j - \mu)^2 = \frac{1}{N}\sum_{k=1}^{K} (\xi_k - \mu)^2 f_k$$

Variance: a refresher

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2 = \frac{1}{N}(\sum_{j=1}^{N} x_j^2 + \sum_{j=1}^{N}\mu^2 - 2\sum_{j=1}^{N} x_j\mu)$$

$$= \frac{1}{N}\sum_{j=1}^{N} x_j^2 + \frac{N}{N}\mu^2 - 2\mu\frac{1}{N}\sum_{j=1}^{N} x_j =$$

$$= \frac{1}{N}\sum_{j=1}^{N} x_j^2 - \mu^2$$

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{K}(\xi_k - \mu)^2 f_k = \frac{1}{N}\sum_{k=1}^{K}\xi_k^2 f_k - \mu^2$$

$$\mathsf{Var}(X) = E(X^2) - (E(X))^2$$

A random sample from the US workforce

Now suppose that we cannot do a census and nobody told us the values of $f_1, \ldots, f_K$. We could get some information from a **sample**, as those routinely used in polls. Let's say that we randomly choose $\boldsymbol{n = 1000}$ workers and get information on their educational level.

- X_i is the random variable (*note the use of upper case*) representing the number of education years of the i worker sampled

$$X_1 \quad X_2 \cdots X_n$$

- X_i is random because i is random: we do not know a priori which worker we will sample

- Since each worker in the population is equally likely to be the ith member of the sample

$$P(X_i = \xi_k) = \frac{f_k}{N}$$

The distribution of X_i is

$$X_i \sim \begin{cases} \xi_1 & f_1/N \\ \xi_2 & f_2/N \\ & \\ \xi_K & f_K/N \end{cases}$$

- The expected value $E(X_i)$ is the population mean

$$E(X_i) = \sum_{k=1}^{K} \xi_k \frac{f_k}{N} = \mu$$

- The variance $\mathrm{Var}(X_i)$ is the population variance

$$\mathrm{Var}(X_i) = E(X_i^2) - (E(X_i))^2 = \sum_{k=1}^{K} \xi_k^2 \frac{f_k}{N} - \mu^2 = \sigma^2$$

Sample mean $\bar{X}$

The sample mean is also a random quantity

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Expectation of the sample mean**

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

- **Variance of the sample mean**

$$\text{Var}(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = E(\bar{X}^2) - \mu^2$$

We need to work a bit on the first term

Variance of $\bar{X}$, continued

$$
\begin{aligned}
E(\bar{X}^2) &= E\left(\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right) = \frac{1}{n^2} E(\sum_{i=1}^{n} X_i^2 + 2\sum_{i \neq j} X_i X_j) = \\
&= \frac{1}{n^2} n E(X_1^2) + \frac{1}{n^2} 2 \binom{n}{2} E(X_1 X_2) = \\
&= \frac{E(X_1^2)}{n} + \frac{2n(n-1)(n-2)!}{2(n-2)! n^2} E(X_1 X_2) = \\
&= \frac{1}{n}(\sigma^2 + \mu^2) + \frac{n-1}{n} E(X_1 X_2)
\end{aligned}
$$

Now, X_1 and X_2 are not independent if our sample is done *without replacement* (if the first individual is of type ξ_k, the number of available subjects of this type for X_2 decreases by one). However, let's imagine we sample *with replacement* (which is approximately true when $N >> n$), so that we have independence and $E(X_1 X_2) = E(X_1)E(X_2) = \mu^2$.

$$
\text{Var}(\bar{X}) = E(\bar{X}^2) - \mu^2 = \frac{1}{n}(\sigma^2 + \mu^2) + \frac{n-1}{n}\mu^2 - \mu^2 = \frac{\sigma^2}{n}
$$

What do we learn from a sample?

- The sample mean is in expectation equal to the population mean. Every sample would potentially lead to a different sample mean, but "on average across samples", $\bar{X}$ gives us an *unbiased* guess for μ
- The size of the distance of $\bar{X}$ from μ decreases with the sample size n: the variance of $\bar{X}$ goes to 0 as $n \longrightarrow \infty$
- We can choose n such that the expected value of the "error" is within bounds that we choose

General framework: use samples to learn about populations

Population

- Ex. all the US employees
- Described by a **distribution** F; Ex.

ξ_1	ξ_2	$\cdots$	ξ_K
f_1	f_2	$\cdots$	f_K

- F is either totally **unknown**, or we might know its form up to some parameter θ: F_θ, with θ unknown
- We are interested in learning F or some characteristics of it, as the population mean.

Sample

We **observe** the realization of a random sample $X_1, \ldots, X_n$ from the population.

- $X_i \sim F$
- X_i and X_j are independent (sample with replacement)
- We want to study which functions of $X_1, \ldots, X_n$ are useful to make inference relative to the entire population, and with which error.

More about populations

- **Finite populations**: US workforce, etc. They are concrete populations, of which we can enumerate members. In this course we are going to consider them so much larger than the samples we observe that it is meaningful to think about sampling with replacement.

- **Infinite/conceptual populations**: we might be interested in the performance of a drug on patients of a disease. We do not want to restrict to patients at a certain time or in a certain location, but we are interested in the abstract notion of "people affected by a disease".

More about random sample

We are always going to assume that the data we observe is a realization of a random process. Where does the randomness come from?

- We introduce it by **design**: sampling subjects, assigning random treatments in a clinical trial...
- The observations are subject to random **measurement error**. For example, there might be a true value of the weight of a molecule, but every time we measure it we get a different reading due to measurement error.
- We are studying a phenomenon that is **inherently random** (in many cases "random" is just an approximation for "too complex for us to describe deterministically"). For example, we might want to study the average time between two earthquakes in the Bay Area $\longrightarrow$ the occurrences of earthquakes are random.

Probability and statistics

- In a typical "probability" problem you are given information on a process and you calculate the probability of an outcome

- In a typical "statistical" problem, you are given an outcome and you try to reconstruct something about the process that generated it.

Statistical inference

Statistical inference = Probability^{-1}

Probability: For a specified probability distribution, what are the properties of data from this distribution?

Example: $X_1, \ldots, X_{10} \overset{iid}{\sim} \mathcal{N}(2.3, 1)$. What is $\mathbb{P}[X_1 > 5]$? What is the distribution of $\frac{1}{10}(X_1 + \ldots + X_{10})$?

Statistical inference: For a specified set of data, what are properties of the distribution(s)?

Example: $X_1, \ldots, X_{10} \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ for some θ. We observe $X_1 = 3.67$, $X_2 = 2.24$, etc. What is θ?

Going back to the variance of $\bar{X}$

Let's now be more careful in evaluating $E(X_1X_2)$ for a random sample, considering the case of *sampling without replacement*.

$$\begin{aligned} E(X_1X_2) &= \sum_{k=1}^{K}\sum_{l=1}^{K}\xi_k\xi_l P(X_1=\xi_k, X_2=\xi_l) = \\ &= \sum_{k=1}^{K}\xi_k P(X_1=\xi_k)\sum_{l=1}^{K}\xi_l P(X_2=\xi_l|X_1=\xi_k) \end{aligned}$$

The distribution of $X_2|X_1=\xi_k$ is

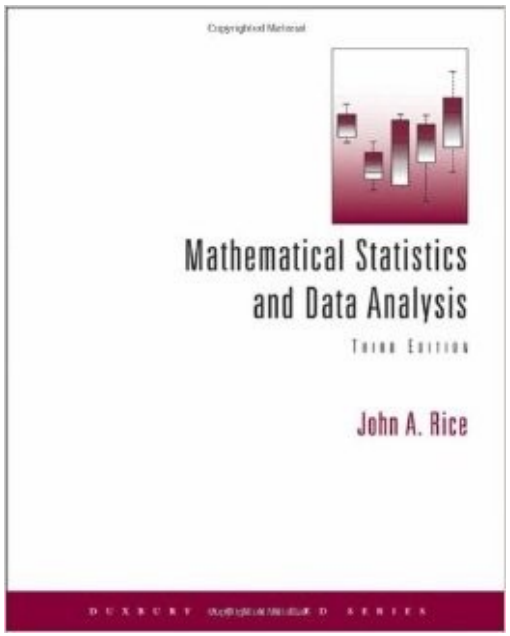
$$P(X_2=\xi_l|X_1=\xi_k) = \begin{cases} \dfrac{f_l}{N-1} & \text{for } l\neq k \\ \dfrac{f_k-1}{N-1} & \text{for } l=k \end{cases}$$

$$
\begin{aligned}
E(X_2|X_1=\xi_k) &= \sum_{l=1}^{K}\xi_l P(X_2=\xi_l|X_1=\xi_k) = \\
&= \sum_{l\neq k}\xi_l\frac{f_l}{N-1}+\xi_k\frac{f_k-1}{N-1} = \\
&= \sum_{l=1}^{K}\xi_l\frac{f_l}{N-1}-\xi_k\frac{1}{N-1}
\end{aligned}
$$

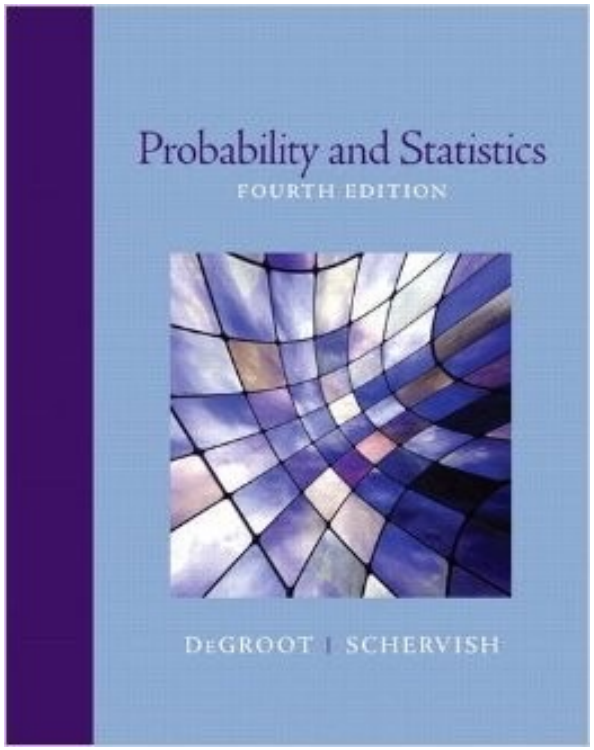
Putting this back in $E(X_1X_2)$,

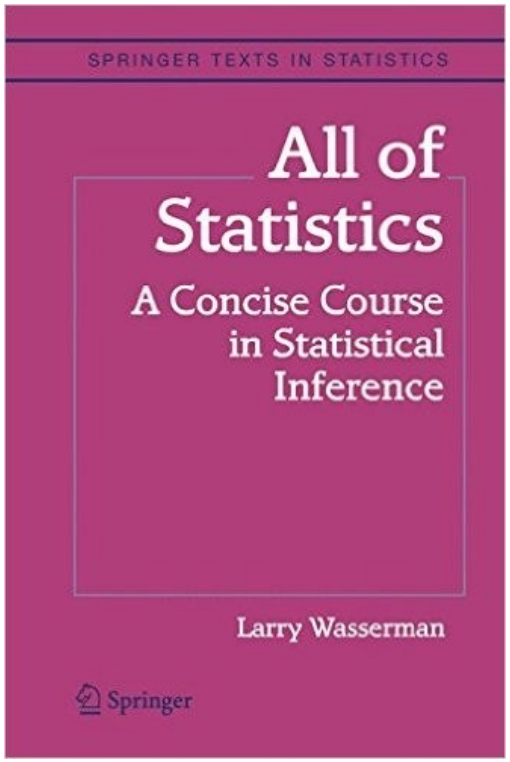
$$
\begin{aligned}
E(X_1X_2) &= \sum_{k=1}^{K}\xi_k\frac{f_k}{N}\left(\sum_{l=1}^{K}\xi_l\frac{f_l}{N-1}-\xi_k\frac{1}{N-1}\right) = \\
&= \frac{N}{N-1}\sum_{k=1}^{K}\xi_k\frac{f_k}{N}\sum_{l=1}^{K}\xi_l\frac{f_l}{N}-\frac{1}{N-1}\sum_{k=1}^{K}\xi_k^2\frac{f_k}{N} = \\
&= \frac{N}{N-1}\mu^2-\frac{1}{N-1}(\mu^2+\sigma^2) = \mu^2-\frac{\sigma^2}{N-1}
\end{aligned}
$$

Notes and textbook



For reference:





Morris H. DeGroot and Mark J. Schervish, *Probability and Statistics*
Larry Wasserman, *All of Statistics: A concise course in statistical inference*

"Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid."—Wasserman

Reference

Slide contents come from Stanford inference course