ECON 57 - Lecture 1

A Brief History of Probability and Statistics

Augusto Gonzalez-Bonorino

Pomona College

Fall 2023

Table of Contents

Introduction

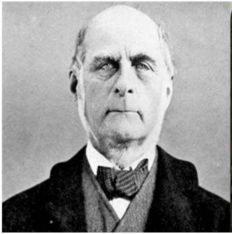
- Studying the history of any subject helps reveal details hidden by time, biases, assumptions, and set the context for the topics to be studied. It can help us avoid misinterpretations, identify long-term trends or interdisciplinary connections, and find inspiration to innovate.
- Logic and reasoning can then be employed to assess if:
  - The assumptions are still valid.
  - The biases are not strong enough to invalidate the methods.

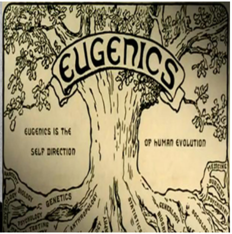
Collecting Experiences

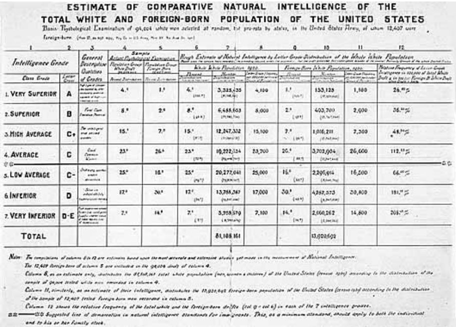
1. First evidence of data collection is from 19000 BC
2. Data analysis as a concept has existed since 1663
3. Think of data as a collection of experiences or observations
4. Statistics emerged from previous efforts to make sense of these experiences.

Eugenics and Early Statistics

- Francis Galton's pioneering role
- Birth of eugenics and its implications
- Statistical techniques developed by Galton:
  - Standard deviation
  - Correlation
  - Linear Regression
- Ethical considerations and lessons learned







ESTIMATE OF COMPARATIVE NATURAL INTELLIGENCE OF THE TOTAL WHITE AND FOREIGN-BORN POPULATION OF THE UNITED STATES

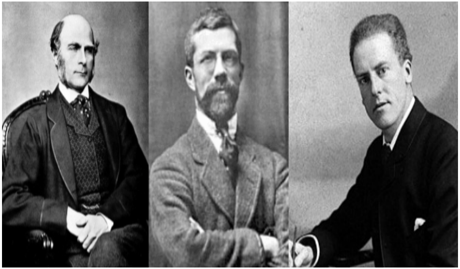
Basis: Psychological Examination of 94,004 white men selected at random, but pro-rata by states, in the United States Army, of whom 12,407 were foreign-born. [illegible]

1	2	3	4	5	6	7	8	9	10	11	12
Intelligence Grade		General Descriptive Qualities of Grades	Sample Actual Psychological Examination		Rough Estimate of Natural Intelligence by Letter Grade Distribution of the Whole White Population						
			Population Group White Draft 94,004	Population Group Foreign Born 12,407	White White Population 1920			Foreign Born White Population 1920			Relative Frequency of Letter Grade Intelligence in 100,000 of total White Draft & in [illegible] Foreign-B. White Draft
Class Grade	Letter Grade		Percent Distribution	Percent Distribution	Percent	Number	[illegible]	Percent	Number	[illegible]	
1. VERY SUPERIOR	A	[illegible]	4.1	1.3	4.1	3,325,435	4,100	1.3	153,123	1,100	26.[illegible]%
2. SUPERIOR	B	First Class [illegible]	8.0	2.0	8.0	6,655,653	8,000	2.9	403,700	2,900	36.[illegible]%
3. HIGH AVERAGE	C+	The intelligent [illegible]	15.1	7.3	15.1	12,247,332	15,100	7.3	1,016,211	7,300	48.[illegible]%
4. AVERAGE	C	Good Common [illegible]	23.7	26.6	23.7	10,222,324	23,700	26.6	3,702,904	26,600	112.[illegible]%
5. LOW AVERAGE	C-	Ordinary worker [illegible]	25.0	16.5	25.0	20,277,041	25,000	16.5	2,296,914	16,500	66.[illegible]%
6. INFERIOR	D	[illegible]	17.0	30.8	17.0	13,758,367	17,000	30.8	4,287,373	30,800	181.[illegible]%
7. VERY INFERIOR	D-E	[illegible]	7.1	14.8	7.1	5,758,579	7,100	14.8	2,060,262	14,800	208.[illegible]%
TOTAL						81,108,161			13,920,692		

Note: The computations of columns 6 to 12 are estimates based upon the most accurate and extensive studies yet made in the measurement of Natural Intelligence.
The 12,407 foreign born of column 5 are included in the 94,004 draft of column 4.
Column 6, as an estimate only, distributes the 81,108,161 total white population (men, women & children) of the United States (census 1920) according to the distribution of the sample of 94,004 tested white men recorded in column 4.
Column 11, similarly, as an estimate of their intelligence, distributes the 13,920,692 foreign-born population of the United States (census 1920) according to the distribution of the sample of 12,407 tested foreign-born men recorded in column 5.
Column 12 shows the relative frequency of the total white and the foreign-born drafts (col. 9 ÷ col. 6) in each of the 7 intelligence grades.
——— Suggested line of demarcation in natural intelligence standards for immigrants. This, as a minimum standard, should apply to both the individual and to his or her family stock.

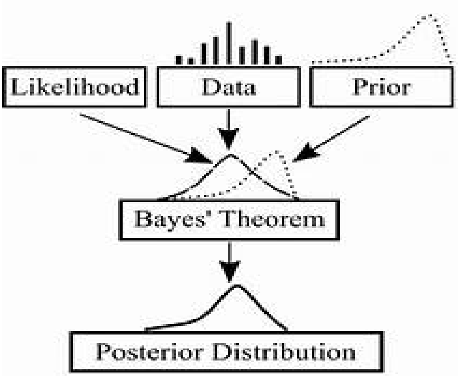
Ronald Fisher, Karl Pearson, and the Frequentist School

- Ronald Fisher and Karl Pearson's contributions
- Introduction of statistical significance and unbiased authority
- Correlation as a measure of causation
- Limitations of inferring causation from correlation



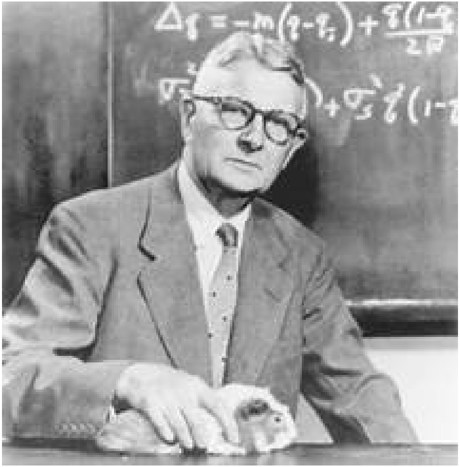
Emergence of Bayesian School of Thought

- Introduction to the Bayesian school of thought
- Thomas Bayes' theorem and incorporating prior beliefs
- Computational challenges and adoption of Bayesian methods
- Balancing uncertainty and complex analysis



Wright's Opposition

- American geneticist and statistician from the '20s.
- Proposed a new methodology for studying causal relationships
    - He accounted for the direction of the relationship, unlike correlation which as a general measure of association.
- His methodology culminated in the development of path diagrams



Path Diagrams

- Path diagrams are a reasoning tool. To help you break down relationships into relevant factors and map the direction of these relationshis.
- Moving beyond simplistic correlations.
- Help guide the formulation and testing of causal hypothesis.
- Fundamental tool in Structural Equation Modeling (SEM) and other causal inference techniques widely used by econometricians.

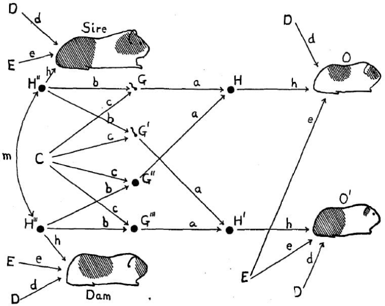


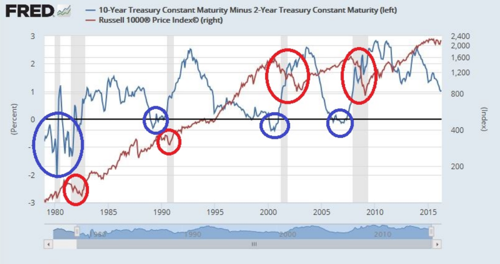
FIGURE 2.—A diagram illustrating the relations between two mated individuals and their progeny. H, H', H'' and H''' are the genetic constitutions of the four individuals. G, G', G'' and G''' are four germ-cells. E and D represent tangible external conditions and chance irregularities as factors in development. C represents chance at segregation as a factor in determining the composition of the germ-cells. Path coefficients are represented by small letters.

Correlation is not Causation

- Shifting from relying solely on correlation
- Highlighting confounding variables and chance
- Recognizing the complexity of cause-and-effect relationships
- Nevertheless, correlations have taken us a long way.

SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.

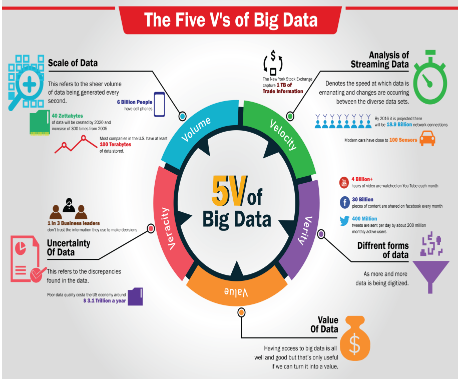
© marketoonist.com



research.stlouisfed.org

Big Data and the Birth of Data Science

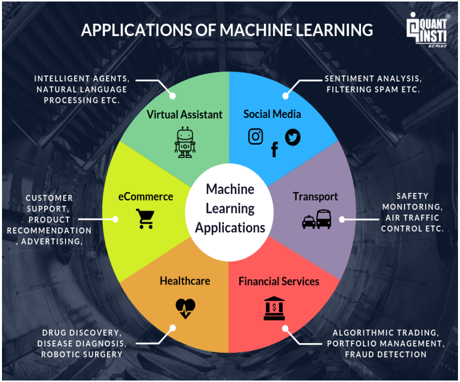
- Transformation with computing power and Big Data
- Rise of data science as an interdisciplinary field
- Integration of machine learning algorithms
- Extracting insights from large datasets



The Five V's of Big Data
5V of Big Data
Volume
Velocity
Verity
Value
Veracity
Scale of Data
This refers to the sheer volume of data being generated every second.
6 Billion People have cell phones
40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005
Most companies in the U.S. have at least 100 Terabytes of data stored.
The New York Stock Exchange capture 1 TB of Trade Information
Analysis of Streaming Data
Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.
By 2016 it is projected there will be 18.9 Billion network connections
Modern cars have close to 100 Sensors
4 Billion+ hours of video are watched on You Tube each month
30 Billion pieces of content are shared on facebook every month
400 Million tweets are sent per day by about 200 million monthly active users
Diffrent forms of data
As more and more data is being digitized.
1 in 3 Business leaders don't trust the information they use to make decisions
Uncertainty Of Data
This refers to the discrepancies found in the data.
Poor data quality costa the US economy around $ 3.1 Trillion a year
Value Of Data
Having access to big data is all well and good but that's only useful if we can turn it into a value.

Machine Learning and Neural Networks

- ML combines statistics with algorithms with the goal of creating reasoning engines.
- Machine learning's impact on diverse domains
- Machine learning vs Neural networks vs Deep learning
- Sparks of AGI: computer vision and NLP
- Python and R


APPLICATIONS OF MACHINE LEARNING
QUANT INSTI GO ALGO
Machine Learning Applications
Virtual Assistant
INTELLIGENT AGENTS, NATURAL LANGUAGE PROCESSING ETC.
Social Media
SENTIMENT ANALYSIS, FILTERING SPAM ETC.
Transport
SAFETY MONITORING, AIR TRAFFIC CONTROL ETC.
Financial Services
ALGORITHMIC TRADING, PORTFOLIO MANAGEMENT, FRAUD DETECTION
Healthcare
DRUG DISCOVERY, DISEASE DIAGNOSIS, ROBOTIC SURGERY
eCommerce
CUSTOMER SUPPORT, PRODUCT RECOMMENDATION , ADVERTISING,

Judea Pearl and Bayesian Networks

- Judea Pearl's contribution to causal inference
- Introducing Bayesian networks for complex systems
- Incorporating probability theory and causal reasoning
- Applications in artificial intelligence and epidemiology

Basic Concepts of Statistics

Populations and Samples

In statistics, we often deal with two main concepts: populations and samples.

Population: The entire group of individuals or items that we are interested in studying.
Sample: A subset of the population, carefully selected to represent the entire population.

For example, if we want to study the average income of all households in a country, the *population* would be all households in that country, and a *sample* would be a smaller group of households from different regions.

Parameter vs Statistic

In statistics, we use both parameters and statistics to describe data:

- **Parameter:** A numerical measure that describes a characteristic of a population. For instance, the average income of all households in a country is a population parameter.
- **Statistic:** A numerical measure that describes a characteristic of a sample. The average income of a sample of households from different regions is a sample statistic.

In practice, we often use statistics to estimate parameters. This involves collecting data from a sample and using sample statistics to make educated guesses about population parameters.

Random Sampling

Random sampling helps us representative samples:

- **Random Sample:** A sample selected from a population in such a way that each individual or item has an equal chance of being included in the sample. This helps to minimize bias and ensures that the sample is representative of the population.
- **Simple Random Sampling:** A type of random sampling where each possible sample of a given size has an equal chance of being selected. This is typically achieved using random number generators.
- **Stratified Sampling:** Dividing the population into subgroups (strata) based on some characteristic, and then selecting samples from each stratum. This ensures that each subgroup is represented in the sample.
- **Cluster Sampling:** Dividing the population into clusters, and then randomly selecting entire clusters to include in the sample. This can be more practical when the population is large.

Importance of Sampling in Economics

- **Resource Efficiency:** Sampling yields accurate insights using fewer resources than surveying the entire population.
- **Representativeness:** Well-designed samples mirror population characteristics, enhancing result reliability.
- **Inference:** Accurate inferences about the entire population are drawn from well-conducted samples, vital for policymaking.
- **Reducing Bias:** Sampling mitigates bias risks from over/underrepresentation in analysis.
- **Feasibility:** Sampling offers a practical approach when surveying the entire population is impossible.
- **Time Sensitivity:** Sampling enables frequent data collection for swift response to changing economic trends.
- **Statistical Analysis:** Sampling provides manageable datasets for complex economic modeling.

Statistical Data Types

Data can be classified into two main types:

1. **Categorical:** Data points that can be grouped into categories. For instance, food groups or gender groups or economic indicators (a special category of general statistical indicators that are related to economics). Remember the XOR problem from last class?
2. **Numerical:** Data points that are solely represented by numbers and do not belong to a group or category.
    1. **Discrete:** Data points that take distinct, separate values. For example, the number of students in a class, or the count of cars passing through a toll booth in a given time period.
    2. **Continuous:** Data points that can take any value within a range. These values are not limited to specific, separate points. Examples include height, weight, or temperature measurements.

R Data Types

1. **character/string:** The character data type in R is used to represent text, such as words, sentences, or any sequence of characters. In R, strings are enclosed within single or double quotes. For example, "Hello, World!" and 'R Programming' are both examples of character data.
2. **integer:** Integers are used to represent whole numbers in R without any decimal points. They can be either positive or negative.
3. **float:** Float numbers are used to represent decimals in R. They differ from integers in memory requirements, which must be left "floating" to account for the fractional parts.
4. **logical/boolean:** Logical data types in R represent binary values, either TRUE or FALSE, and are used for making decisions and logical operations. They are commonly used in conditional statements and filtering data.

R Data Structures

Unidimensional data structures

1. **vector:** A vector is a one-dimensional data structure in R that can hold elements of the same data type. It can be created using the 'c()' function, and elements can be accessed using numeric indices. Vectors are fundamental in R and are widely used for data manipulation and calculations.
2. **factor:** Factors are used to represent categorical data in R. They are particularly useful when dealing with data that have predefined categories or levels. Factors are created using the 'factor()' function, and each level represents a distinct category. Factors play a significant role in statistical modeling and are used for tasks such as creating bar charts and performing categorical data analysis.

More on numerical data

Qualitative vs Quantitative

We will be focusing primarily on numerical data in this course. In general, numerical data and variables can be of two broader categories: Qualitative and Quantitative

1. **Qualitative:** Qualitative data are non-numeric in nature and represent categories or attributes. They convey qualities or characteristics that cannot be measured using numbers alone, they simply represent different labels.
    - There is no inherent meaning in the difference of numbers. This means that comparing two qualitative numbers yields little to no actionable information. For instance, economic regions.
2. **Quantitative:** There is a difference, thus quantitative variables carry information. These values can be measured, compared, and subjected to mathematical operations. For instance, the height of the players in a basketball team (i.e., x is taller than y) or inflation and GDP growth rates.

More on numerical data

Scales of Measurement

- **Qualitative:**
  1. **Nominal Data:** Categories without any inherent order, such as colors or names.
  2. **Ordinal Data:** Categories with a natural order, but the differences between them are not well-defined, such as rankings.
- **Quantitative:**
  1. **Interval Data:** Numerical data with a consistent interval between values, but no true zero point, such as temperature in Celsius.
  2. **Ratio Data:** Numerical data with a consistent interval between values and a true zero point, such as height or weight.

Conceptual Exercises

1. You are analyzing the data related to countries' GDP (Gross Domestic Product) values. Discuss whether GDP values are nominal, ordinal, interval, or ratio data. Explain your reasoning and provide examples to support your answer.
2. Consider a dataset of student exam scores. The scores are recorded as "A," "B," "C," "D," and "F." Determine the level of measurement for these scores (nominal, ordinal, interval, or ratio). Also, discuss whether arithmetic operations like addition and subtraction are meaningful for this dataset.
3. A dataset contains the heights of students in centimeters. Classify the heights as interval or ratio data. Explain the difference between these two levels of measurement and provide an example for each.

R programming

Installing RStudio

Install R
R studio