Bus 701:
Advanced Statistics

Harald Schmidbauer

About These Slides

- The present slides are not self-contained; they need to be explained and discussed. This will be done in the lectures.

- Even though being a "work in progress" and subject to revision, the slides constitute copyrighted material.
  If you want to reproduce or copy anything from the slides, please ask:

Harald Schmidbauer	**harald** at **hs-stat** dot **com**
Angi Rösch	**angi** at **angi-stat** dot **com**

- The slides were produced using LaTeX and R (the R project; website: www.R-project.org) on a GNU/Linux system.

- R files used for this course are available upon request.

A Few Highlights

From the History of Statistics

(Fragments)

A Few Highlights From the History of Statistics

Outline in three parts. . .

- Part 1: The Origins
- Part 2: The Rise of Statistical Thinking, 1820–1900
- Part 3: Statistics in the 20th Century

Part 1: The Origins

Counting the Population

- 2600 BC: construction of pyramids in Egypt; censuses to determine the power of the state
- population censuses for recruitment and taxation
- USA: population census to represent the states according to population (1787 Constitution)

Part 1: The Origins

The study of social numbers

- 1660s: first systematic numerical studies of land and inhabitants
- main objective: provide information for state policy
- in line with the philosophy of Francis Bacon (1561–1626)
- focus on measurable information
- "Wealth of a state depends on the number of its inhabitants."
- members of society can be manipulated at will

Part 1: The Origins

Political arithmetic

- William Petty (1623–1687):
  "Essays on Mankind and Political Arithmetics" (1672)
  - on the growth of the city of London
  - observations upon the Dublin bills of mortality
  - a scheme to setting the poor to work

- John Graunt (1620–1674):
  "Bills of Mortality of the City of London" (1662)
  - risk of death by causes
  - regularity of sex ratio at birth

Part 1: The Origins

Political arithmetic

- Edmond Halley (1656–1742):
  life table for Breslau (1693)

- Johann Peter Süßmilch (1707–1767):
  "The divine order in the changes in the human sex from birth, death and reproduction of the same" (1762)
  - objective: maximize population
  - provide for medical care, discourage emigration
  - large-scale regularity emerging from local chaos
  - Immanuel Kant (1724–1804):
    ". . . regular movement in freedom of will in the large. . . like weather maintaining the growth of plants. . . natural goal. . . "

Part 1: The Origins

Political arithmetic

- Thomas Robert Malthus (1766–1834):
  "An Essay on the Principle of Population, as it Affects the Future Improvement of Society" (1798)
  - population grows geometrically, food supply arithmetically
  - constraint on governments and social institutions by laws of population
  - struggle, development of population
  - society as an unstable force

Part 1: The Origins

Meanwhile elsewhere: "University statistics"

- Hermann Conring (1606–1681):
  systematic description of state affairs

- Gottfried Achenwall (1719–1772):
  "Staatsmerkwürdigkeiten"
  ("phenomena of particular interest of a country or a people")
  $\longrightarrow$ "Statistik"

Part 1: The Origins

Games of Chance

- Gerolamo Cardano (1501–1576):
  "De ludo aleae" (1560s; published 1663)
- Antoine Chevalier de Méré (1607–1685):
  questions concerning gambling
- Blaise Pascal (1623–1662)
- Pierre de Fermat (1601–1665)
- Christiaan Huygens (1629–1695)

Part 1: The Origins

Probability

- Jacob Bernoulli (1655–1705):
  "Ars Conjectandi" (published 1713)
- Abraham de Moivre (1667-1754):
  "The Doctrine of Chances" (1718)
- Thomas Bayes (1702–1761): "Essay Towards Solving a Problem in the Doctrine of Chances" (1764)

Part 1: The Origins

Probability

- Pierre-Simon Laplace (1749–1827)
  - dispensing with hypothesis of divine intervention
  - strong belief in causal determinism ("Laplace's daemon")
  - error curve, method of least squares
  - probability: subjective; expressing human belief
  - "principle of indifference"
- Siméon Denis Poisson (1781–1840)
  - "law of large numbers"
  - application of probability to judicial decisions

Part 1: The Origins

Early inductive statistics

- constant cause for difference in morning and afternoon barometer readings
  (1820, by Pierre-Simon Laplace (1749–1827))

- standard error for the mean
  (1821, by Joseph Fourier (1768–1830))

- Variation had to be taken into account!

Part 1: The Origins

The situation ca. 1800–1830

- national statistical offices established in many states
- statistics often used to confirm preconceived ideas or political programs
- statistics: accumulation of facts; complete enumeration
- intermingling causation with statistics not desired
- statistics as a tool in search for reform
- numerical social studies under the name "statistics"
- opposition to statistics (e.g., Jean-Baptiste Say, 1767–1832)

Part 2: Statistical Thinking, 1820–1900

Quetelet: Background and scientific credo

- Adolphe Quetelet (1796–1874); originally astronomer
- numerical social science produces laws, not just facts
- a single method is enough for physical and social sciences
- investigation of seaonal phenomena (births, marriages)
- belief in social order; reformist attitude

Part 2: Statistical Thinking, 1820–1900

Quetelet: Regularity in social phenomena

- law-like certainty in social phenomena, when applied to masses
- no effect of chance or free will on collective events
- mass-regularity to be expected everywhere
  (physicalist/theological cosmology)
- society: an entity in its own right

Part 2: Statistical Thinking, 1820–1900

Quetelet: Physique sociale (1835)

- name pirated from August Comte (1798–1857)
- l'homme moyen physique / l'homme moyen moral
- penchant for crime — crime as social phenomenon
- adapt methods and notions from physics to social sciences (e.g., constant forces — nature; perturbational forces — man)

Part 2: Statistical Thinking, 1820–1900

Quetelet: L'homme moyen and the error law

- error law also applicable to distribution of human features (1844)
- human variation should be understood as "errors"
- human beings identified with copies of a statue
- average represents a moral idea; deviation is ugly
- progress of civilization: narrow limits / reduce variation

Part 2: Statistical Thinking, 1820–1900

Historical determinism: Henry Thomas Buckle (1821–1862)

- "History of Civilization in England" (1857–1861)
- laws of history rather than government meddling
- each individual act is a necessary consequence of social law: "Autonomy of free will is fiction."
- quota of crime in society
- propagated Quetelet's ideas

Part 2: Statistical Thinking, 1820–1900

A new interpretation of probability

- classical interpretation: probability as expression of imperfect knowledge
- 1840s:
  - use of priors is perceived with distrust
  - explain probability in terms of regularity of mass phenomena (invoking the law of large numbers)
- frequentist interpretation of probability

Part 2: Statistical Thinking, 1820–1900

Statistics and thermodynamics

- James Clerk Maxwell (1831–1879)
  - new kind of regularity for molecules, which are not individually sensible
  - error curve used in kinetic gas theory (1872) (equilibrium velocity distribution)
  - imperfection of human knowledge ("Maxwell's daemon")
- Ludwig Boltzmann (1844–1906)
  - reference to Buckle
  - second law of thermodynamics can be understood in terms of probability

Part 2: Statistical Thinking, 1820–1900

The theory of evolution

- Charles Darwin (1809–1882):
  "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life" (1859)

- summary:
  - a species can reproduce beyond replacement level
  - limited resources, struggle for life
  - variation (partly inherited), natural selection of most fitted
  - formation of new species

- chance variation substituted for teleology
  ("chance": undirected, but not uncaused)

Part 2: Statistical Thinking, 1820–1900

Galton — the background

- Francis Galton (1822–1911); originally meteorologist
- overwhelmed with Darwin's theory of evolution
- empirical evidence for the theory of evolution?
- founder of eugenics — "race improvement"
- science (in particular, eugenics) as replacement for religion

Part 2: Statistical Thinking, 1820–1900

Galton — investigation of heredity

- speculative theories on how genetic material is passed on
- opposed Quetelet's idea that "variation $\equiv$ error", methods of astronomy therefore unsuitable
- exceptional more interesting than average (evolution based on "sports" of nature)
- "Hereditary Genious" (1896)
- statistical theory of heredity (after 1875)

Part 2: Statistical Thinking, 1820–1900

Galton — statistical studies in heredity

- experiment on peas
  offspring of very large (small) peas not so large (small)
- reversion to mean while preserving variability
- plot of diameter of mother seed against diameter of offspring: first regression line
- measurement on humans (mid-parents against adult children)
- correlation between physical/psychological quantities

Part 2: Statistical Thinking, 1820–1900

Pearson — the background

- Karl Pearson (1857–1936),
  philosopher of science, mathematician, statistician, eugenicist
- vision: create statistical biology as the basis of effective eugenics
- made innumerable measurements (animals and humans)
- created neologisms ("normal distribution", "standard deviation")
- founder of *Biometrika*

Part 2: Statistical Thinking, 1820–1900

Pearson — the biometrical statistics

- skewed distributions (1895)
- regression and correlation (1896)
- χ^2 distribution
- contingency tables

Part 2: Statistical Thinking, 1820–1900

The situation around 1900

- Francis Galton and Karl Pearson have ushered in mathematical statistics
- "erosion of determinism" (Hacking, 1983)
- focus on mass phenomena
- "statistical revolution"? (Salsburg, 2002)

Part 3: Statistics in the 20th Century

- few observations ($\sim$1910)
- sampling theory ($\sim$1910, 1930–)
- maximum likelihood, theory of estimation (1912–)
- design of experiments ($\sim$1920)
- hypothesis testing ($\sim$1925)
- confidence intervals ($\sim$1935)
- probit, logit (1934, 1944)
- extreme value theory (1930s)
- non-parametric methods
- Bayesian statistics (1960s–)
- time series analysis
- quality control
- EDA
- data mining