preliminaries: statistics and data analysis

Statistics. . .

started quite early. Or rather . . . rather with a considerable delay.
The first applications of statistics were the games of luck.
And the games – dice in the first place – were known already in ancient civilisations. Already Aristotle mentions dice games; Greek gods played dice in the Olympus (disputing partition of lands, seas etc.) Greek warriors played games to pass the long time of the Troy (Illium) siege.
No one tried any quantitative estimation(s) of chances to win; only Aristotle mentions dice games and writes about ,,more easy" combinations.
A nice story can be found in the Herodotus ,,Histories":
The Lybians had a 18-year long period of dry weather and famine so . . . they played dice to support more easily the hunger (every 2nd day).
Ancient Romans used four dice and some combinations had nice names: ,,Venus" (all four numbers different – the best result) and ,,dog" (4 identical numbers – the worst). Cesar Claudius wrote a witty booklet about the dice (he had a special table for playing dice in his portable chair).
In the middle ages knights played games again (like Greek heros) during long and hard sieges (crusades!); the king Louis IX had to issue a special law, defending knights to play too much.
In Renaissance cards entered. Printing workshops (very first of them) produced cards and . . . players used them. The well-known printer Gutenberg produced a magnificent Bible and . . . a deck of cards.

Probability pioneers: Antoine Gombaud, (Chevalier de Méré) (1607–1684) and Pierre Fermat

de Méré was quite a smart mathematician. He knew, that tossing a die 4 times in a row he could expect to have rather one ‘six’ than not. — $1 - (5/6)^4 > 0.5$.

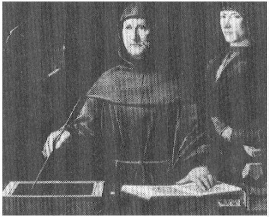
For 2 dice, he tried: $6 \times 6 = 36; \quad 36 \times 4/6 = 24$, so he was trying to bet to have two ‘sixes’ in 24 attempts. And …he was loosing money.

Furiuos, he wrote a letter to Blaise Pascal. Pascal combined his efforts with his friend Pierre Fermat and the two of them laid out mathematical foundations for the theory of probability. And so it started…

Fermat’s first entry into the subject of statistics was the long-awaited solution of the game *balla*.

We are in the Renaissance Europe

Three great mathematicians emerge: Fra Luca Paciolli (OFM), Tartaglia i Cardano.



the game *balla*

Paciolli wrote a book *Summa de Arithmetica et Geometrica* in which he defines the problem:
Two teams are competing and the game ends if one of the teams has won 5 times. Now, because of some reasons the game is interrupted with team A having 4 wins, and team B – 3. How should the prize (22 ducats) be divided? Cardano proposed (*De ludo alea*) an interesting (and almost correct) partition

$$\frac{1 + ... + (6-5)}{1 + 2 + ... + (6-3)} = 1/6.$$

Pascal: we have to examine *all* the possible scenarios

$$AAA \;\; AAB \;\; ABA \;\; ABB \;\; BAA \;\; BAB \;\; BBA \;\; BBB,$$

Conclusion: the prize should be split in the proportion 7(A) : 1(B).

Pierre Fermat

John Graunt and *Bills of mortality*

Royal Society of London, 1662
The first statistical sampling of the death registries
(London, 1604-1661) (also – estimates of London population)

Table: Percentage of persons versus the span of life

Age	Londyn	US
	17th c.	ca. 1990
0	100	100
6	64	99
16	40	99
26	25	98
36	16	97
46	10	95
56	6	92
66	3	84
76	1	70

John Graunt . . .

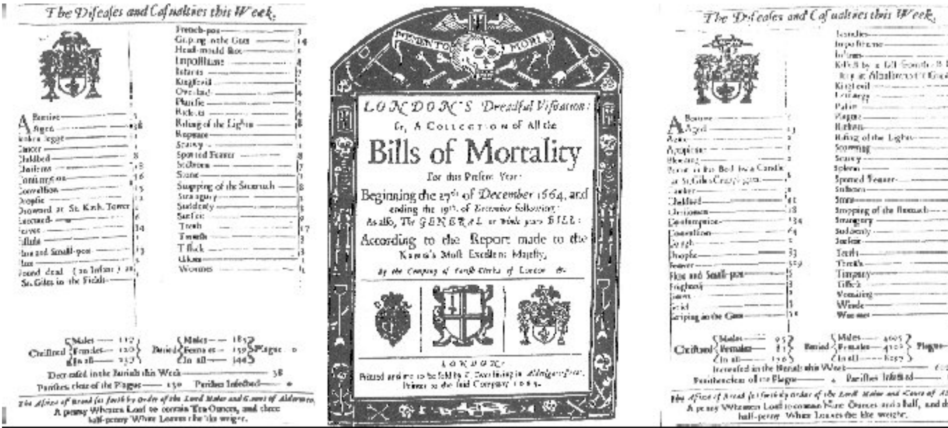
. . . had quite a hard task. The registers (started in 16th C.) were far from complete – they included the sex and the causes of death but not always . . . the age of deceased. However, John G. did quite a decent job.

Some 30 years later Edmund Halley repeated the scheme (in *Transactions of the Royal Society*) using the (more complete!) data sent by a Breslau minister Casper Neumann. On the basis of his study he tried to estimate the number of inhabitants (34 000).

About one hundred years later a similar study was executed by the 'French Newton' – Pierre Simon Laplace (the 1st textbook in statistics).

This is how the idea of 'statistical sample' was born.
The use of such reasoning was also obvious – already in the beginning of 16th C. the commerce of . . . insurances (rents) was already flowering.

John Graunt and *Bills of mortality*



The Diſeaſes and Caſualties this Week.

Abortive	5
Aged	38
Broken legge	1
Cancer	1
Childbed	8
Chriſomes	13
Conſumption	16
Convulſion	15
Dropſie	13
Drowned at St. Kath. Tower	1
[illegible]	6
Feaver	34
Fiſtula	1
Flox and Small-pox	13
Flux	1
Found dead (an Infant) at St. Giles in the Fields	1
French-pox	3
Griping in the Guts	14
Head-mould ſhot	1
Impoſthume	4
Infants	2
Kingſevil	2
Overlaid	4
Plurifie	2
Rickets	4
Riſing of the Lights	8
Rupture	1
Scurvy	1
Spotted Feaver	8
Stilborn	7
Stone	1
Stopping of the Stomach	8
Strangury	1
Suddenly	5
Surfeit	9
Teeth	17
Thruſh	3
Tiſick	3
Ulcer	1
Wormes	4

Chriſtned {Males — 112, Females — 120, In all — 232} Buried {Males — 185, Females — 159, In all — 344} Plague — 0

Decreaſed in the Burials this Week — 38

Pariſhes clear of the Plague — 130 Pariſhes Infected — 0

The Aſſize of Bread ſet forth by Order of the Lord Maior and Court of Aldermen,
A penny Wheaten Loaf to contain Ten Ounces, and three
half-penny White Loaves the like weight.

MEMENTO MORI

LONDONS Dreadful Viſitation:
Or, A COLLECTION of All the
Bills of Mortality
For this Preſent Year:
Beginning the 27th of *December* 1664. and
ending the 19th. of *December* following:
As alſo, The GENERAL or whole years BILL:
According to the Report made to the
King's Moſt Excellent Majeſty,
By the Company of Pariſh-Clerks of London, &c.

LONDON:
Printed and are to be ſold by E. Cotes living in *Aldersgate-ſtreet*,
Printer to the ſaid Company. 1665.

The Diſeaſes and Caſualties this Week.

Abortive	5
Aged	43
Ague	2
Apoplexie	1
Bleeding	2
Burnt in his Bed by a Candle at St. Giles Cripplegate	1
Canker	1
Childbed	42
Chriſomes	18
Conſumption	134
Convulſion	64
Cough	2
Dropſie	33
Feaver	309
Flox and Small-pox	5
Frighted	3
Gowt	1
Grief	3
Griping in the Guts	51
Jaundies	
Impoſthume	
Infants	
Killed by a fall down ſtairs [illegible] at Alhallows [illegible]	
Kingſevil	
Lethargy	
Palſie	
Plague	
Rickets	
Riſing of the Lights	
Scowring	
Scurvy	
Spleen	
Spotted Feaver	
Stilborn	
Stone	
Stopping of the Stomach	
Strangury	
Suddenly	
Surfeit	
Teeth	
Thruſh	
Timpany	
Tiſick	
Vomiting	
Winde	
Wormes	

Chriſtned {Males — 95, Females — 81, In all — 176} Buried {Males — 4095, Females — 4202, In all — 8297} Plague —

Increaſed in the Burials this Week —

Pariſhes clear of the Plague — 4 Pariſhes Infected —

The Aſſize of Bread ſet forth by order of the Lord Maior and Court of Al
A penny Wheaten Loaf to contain Nine Ounces and a half, and th
half-penny White Loaves the like weight.

So, one might say…

…that the major goal of statistics is to draw some conclusions (conjectures) and/or formulate some hypotheses on the basis of some statistical models (dice) and/or statistical sampling-data (bills of mortality).

But …

…there is another, quite important of aspect of statistics

this is estimation of uncertainty (-ies) of the results of our measurements

Science

in a way can be perceived as *collecting and analysing data*. These data can be of various nature:

- qualitative (or categorical) : this pencil is black (belongs to the category of **black** pencils)
- semi-quantitative: this pencil is long
- or quantitative: this pencil is 5 cm long.

We shall look upon our data as results of some *physical measurement* that can be, in turn, defined as a controlled physical experiment whose outcome can be more or less precisely quantified.

A **measurement** can be regarded as a *classification* of objects belonging to or forming a group with respect to an attribute or the degree of it that can be associated with a given member of the group. Again, such a classification can have different aspects:

- dichotomy: man/woman, head/tails, etc. . . .
- sorting by category: black/white/, . . . ; eagle/sparrow/tit/. . .
- sorting in order, in function of an attribute: a, a, a, a, . . .

continued

- **measuring** with the aid of a *relative scale* (relative standard), e.g. t_1, t_2 – centigrades (the ratio t_1/t_2 has no sense; the difference $t_1 - t_2$ has)
- **measuring** with the aid of an *absolute scale* (absolute standard), e.g. T_1, T_2 – kelvins (both the ratio t_1/t_2 and the difference $t_1 - t_2$ have sense)

The things we measure are labelled *physical quantities*, e.g. mass, length, current, voltage, etc. To carry out a measurement we use an apparatus (setup): a yardstick or an electronic microscope.

Every measurement is subject to an error – or *uncertainty*.
A measurement report listing only results (without their possible errors – uncertainties) is incomplete.
One of the major goals we want to achieve with the help of statistics is to learn how to estimate correctly the measurement errors.

continued

ERRORS (uncertainties): suppose we measure a quantity whose "true" value is $\tilde{x}$ and we obtain — as the result — x. We define the *absolute error* Δx as:

$$\Delta x = \tilde{x} - x.$$

(The true value $\tilde{x}$: a value known from some theoretical considerations or rather from a very, very accurate measurement.) We may define also the *relative error*:

$$\frac{\Delta x}{\tilde{x}} \approx \frac{\Delta x}{x}.$$

Statistically realistic goal: to find (in a measurement) x and a measure of the error, say σ_x, such as the interval $(x - \sigma_x, x + \sigma_x)$ for which we can state that the "true" value of $\tilde{x}$ can be found within it *with such–and–such degree of probability.*
(the concept of the probability will be discussed in the very next section).

continued

Errors can be divided into two major classes:

- **systematic errors** — they have usually their source in the finite possibilities of our tools (devices)
- **statistical errors** — they may have many sources, for instance:
    - variable conditions during the experiment (measurement) that could not be accounted for
    - difference(s) between the conceptual *model* of our measurement and the actual realisation

So, generally, we shall have:

$$\sigma = \sigma_{\text{systematic}} + \sigma_{\text{statistical}}.$$

or more reallistically

$$\sigma = \sqrt{\sigma^2_{\text{systematic}} + \sigma^2_{\text{statistical}}}.$$