The Ways of Probable Truth

●

Philosophy of Science in the 21st Century — Challenges and Tasks

Dinis Pestana, and Fernando Sequeira

Abstract The famous aphorism of Lord Rutherford, "if your experiment needs statistics, make a better experiment", is undoubtly outdated, and nowadays anyone of his status would perhaps say something like "your research needs statistics, to start with planning an adequate experiment, choosing appropriate data gathering discipline and data analysis tools" followed by a caveat: Use statistics quantum satis, no more no less." Good statistics is a scientific crystal ball to peer into the future, but the eternal agon opposing accuracy and probability results in some kind of equilibrium, a probable truth, but neither the all truth nor certainty, something more fuzzy that we can conceptualize as probable truth. But the ways of establishing probable truth are frequently abused, and the statistics crystal ball is substituted by a void soap bubble of bad science. We present some examples of good science and of bad science achieved using statistics, and stress once again that the power of statistics stems out from its usefulness in rejecting false conjectures, with a caveat: there is no good statistics with insufficient data, but with too many data anything can be rejected, even truth.

1 Introduction

> *Linus Pauling was enormously fertile in the way he developed his ideas. He was asked a few years ago, "How do you get ideas?" And he gave I think what is the correct reply. He said, "If you want to have good ideas you must have many ideas. Most of them will be wrong, and what you have to learn is which ones to throw away."*
>
> Francis Crick, opening conference of *The Pauling Symposium*, Oregon State University.

The rejection of false hypotheses is a cornerstone of modern science methodology. The experimental sciences goal is to build knowledge from information, that is always partial. Information is often referred to as *data*, often a *sample* that we hope represents the *population*. But, as Abraham de Moivre cleverly remarked, in *The Doctrine of Chances*

> *Further, The same Arguments which explode the notion of Luck may, on the other side, be useful in some Cases to establish a due comparison between Chance and Design: We may imagine Chance and Design to be as if it were in Competition with each other, for the production of some sorts of Events, and may calculate what Probability there is, that those Events should be rather owing to one than to the other.*

that there always exists a share of chance on the evidence we gather, and this will indeed condition the induction from the sample to the population.

In fact, Probability is the instrument to tame chance, and now chance should be regarded as a friendly ally, not as malevolent enemy of truth. Probability must intervene in the sample selection, with a price: if the

Dinis Pestana
Universidade de Lisboa, DEIO–FCUL,
CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa,
Instituto de Investigação Científica Bento da Rocha Cabral, and
CFCUL — Centro de Filosofia das Ciências da Universidade de Lisboa, Portugal
e-mail: dinis.pestana@fc.ul.pt

Fernando Sequeira
Universidade de Lisboa, DEIO–FCUL,
CEAUL — Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
e-mail: fjsequeira@fc.ul.pt

sample strategy has been appropriate, and the sample size sufficient, the sample will be representative for our purposes, in the strict sense that conclusions will have a prescribed error bound *at a given probability level*. There is an unavoidable uncertainty that has to be clearly incorporated in scientific reasoning, even when the data collection has been firmly guided by good sampling methodology. cf. [16].

Rejecting false hypothesis is the central idea of the Neyman-Pearson [12] formalization of statistical hypothesis testing (as opposed to Fisher's significance tests, relying solely on the surprise carried by the statistic summarizing the sample, as measured by the p-value). Although Neyman-Pearson theory is less than a century old, the most outstanding ideas appeared for the first time three centuries ago, in an ingenious letter written by Arbuthnot [1] to the London Royal Society.

Arbuthnot reasoning is as follows; in the 72 years of birth register in his parish, year by year, more boys were born than girls. From this he concludes that the Divine Providence judiciously decided that more males than females should be born, since, as he explicitly states, only a very lunatic guy would claim that the probability of Head = probability of Tail when tossing a coin that in the 72 previous tosses always return Head!

The modern formalization would be the following; denote $\mathbb{P}(M)$ the probability of a male birth, $\mathbb{P}(F)$ the probability of a female birth; we suspect that $\mathbb{P}(M) > \mathbb{P}(F)$, something that we shall call alternative hypothesis H_A; on the other hand, we shall call null hypothesis H_0 the naïve idea that $\mathbb{P}(M) = \mathbb{P}(F) = \frac{1}{2}$.

The probability of observing something at least as extreme as that more boys have been born than girls in every one of $n = 72$ independent inspections (in this context non-overlapping inspection windows vouches for independence) is surely very low assuming that H_0 is true. In fact, denoting X the random variables "number of inspections returning more boys than girls in 72 random inspections", under H_0,

$$\mathbb{P}[X \geq 72] = \left(\frac{1}{2}\right)^7 2 = \frac{1}{4,722,366,482,869,650,000,000.00} \approx 0.00000000000000000000211758237$$

This *p-value* is so low that it seems safe to conclude that H_0 is false and should be rejected, and therefore our faith in H_a is reinforced. Observe however that there is no way of proving that H_A is truth, what we have just seen is that the probability of its negation[1] being true is so thin it doesn't deserve to be kept.

Often scientific matters are dealt along similar lines: some "degree zero" of knowledge is rejected in favor of some more interesting alternative, on the grounds that facts (observed data) would be "highly" improbable under the null hypothesis.

Due to computational limitations in the early years of statistics, it was usual to reject H_0 if the observed p-value $p < 0.05$ (significant). This makes some sense in science, since the criterion of repeatability implies that wrong decisions would be corrected at a later stage. Nowadays, scientific conclusions carry in general higher confidence, since the ambition is often to report p-vales less than 0.0001.

Observe however that apparently low values carry more or less conviction according to context: In any blood collection, health professionals must use as null hypothesis "this blood sample is contaminated", but none would feel safe knowing that rejection of this hypothesis could be wrong 1 in 10000 cases!. The test is performed so that the probability of wrongly rejecting the null hypothesis is less than 1 in 100 — but repeated at lest 4 times (so the probability of wrongly using contaminated blood is less than 1 in 100,000,000) or even 11 times in case the donor is considered a risk!

The danger, with the actual protocol are minimal. However, probability and certitude aren't the same thing, and we know that impossible events do happen, almost surely. In fact even when A is almost certain, in infinite independent replication of the random experiment

$$\mathbb{P}\left(\bigcap_{k=1}^{n} A_n\right) = \prod_{k=1}^{n} \mathbb{P}\left(\bigcap_{k=1}^{n} A_n\right) \quad \underset{n\to\infty}{\longrightarrow} \quad 0, \tag{1}$$

[1] Strictly speaking H_0 is not the negation of H_A. However, under H_0^*; $\mathbb{P}(M) < \frac{1}{2}$ the observed data would be even more improbable than under H_0, and so the test we performed carries the same conlusion that the composite *vs*. composite test H_0^* *vs*, H_A.

a most striking consequence of defining $\mathbb{P}$ as a measure taking values in $[0,1]$. This is sometimes referred to jokingly as the infinite number of monk(ey)s theorem (if an infinite number of monk(ey)s type randomly in the keyboards of an infinite number of typewriters, one will surely produce *Opus Majus* (in fact, monk Roger Bacon produced it, although not in the keyboard of a writing machine). The occurrence of almost impossible events seems at first sight farfetched, and a simple mathematical curiosity, but almost impossible events are easily exhibited in very trivial random experiments. For instance, when a point is chosen at random in the line segment $[0,10]$, the result of the experiment (some point $x_0 \in [0,10]$) is always a null probability event, while the almost certain result $[0,10]-\{x_0\}$ doesn't occur.

Statistical tests are surely the most evident feature os Statistics dealing with science methodology. On the other hand the most outstanding statistician, Sir Ronald Fisher, developed almost single handed, in the thirties, the theory of planning of experiments, cf. [4], that no doubt is the gold standard of modern research procedures.

In Section 2, we describe what we consider an remarkable example of research (Salk on the mother heartbeat) in contrast with bad science that perversely uses the fashions of good science.

In Section 3 we discuss Laplace's "justification" of induction, and briefly sketch some points about Probability and Statistics that have some immediate bearing on the methodology of experimental scientific research. This is followed by some trite remarks on the limits of "truth" in science.

2 Statistics, research methodology, and science

Statistics deals with an important share of research methodology. In fact, Statistics is more than data analysis and the ensuing interpretation of statistical findings. The most sophisticated data analysis will be a flop if the data are inappropriate, and this can happen for several reasons. To start with, few data can lead to a biased view (but on the other hand, too much data will exhibit irrelevant statistical significant results). But sample size isn't the only major issue in what concerns representativeness: the so-called "convenience sampling" can be very convenient in what regards easy and cheap data gathering, but is flawed and inconvenient for statistical purposes: randomness arises as one of the most faithful allies of researchers, since only random choice and random assignment of data to experimental and control groups can vouch for true representativeness and bias control. On the other hand, the reuse of data is inappropriate, since even soundly collected samples distort, to some extent, reality [2], and therefore reuse of samples will reinforce bias. In particular, when a sample is reused to perform a confirmatory data analysis of something suggested by the sample observation in exploratory data analysis stages, the results are abusive manipulation of statistics.

On the other hand, bad science abundantly abuses statistics to confer some degree of apparent plausibility to the most farfetched claims. Statistical experimental design is a standard in modern research, and unplanned experiments, or experiments without control, are prone to produce bad science. But on the other hand, experimental design can be used to investigate stupid questions, an abuse that often is at the root of bad science, many thanks to the improbable research team (`http://www.improbable.com`) and to the jury of the Ig Nobel prizes for the best use of bad science, i.e. a joyful guffaw.

We describe Salk's [17] research on heartbeat sound soothing effect on babies and its feedback on the way mothers hold them as an example of sound research carefully planned. On the other hand, we comment on several issues responsible for bad science.

[2] Only a population census, instead of sampling, would ideally produce a faithful representation, but census of infinite populations is impossible, and census of finite population is costly, and in general it is not a complete and faithful portrait of reality: some units cannot be found or contacted, and when a large number of census agents are hired, some of them fake results.

Fig. 1 Mother and Child.

Fig. 2 Pietà.

2.1 Heart Beating — a remarkable example of good research

Salk [17] is what we consider an excellent example of old planning and execution of an experiment to investigate a sound question. Salk was a gynaecologist who observed that mothers in general carry their infant with the head to the left of their breast. It is curious to observe that in *Piets* in general the dead body has his head to the right side of the *Mater Dolorosa*, cf. 2.

This he observed also to be a general pattern in *Mother and Child* representations, cf. Fig. 2, and his research in anthropology and in art books confirmed the "universality" of this preferred position, cf. Fig. 3

When asking why this choice, the majority would answer something like: in this position I have free use of my right arm, and this is useful to perform many tasks while I am holding my baby."

But a minority of left-handed women used the same position, justifying their choice saying something like "I am carrying my precious baby using my stronger arm, because it is much safer".

This conflicting evidence shows that the reason surely lies much deeper. Salk hypothesised that the true reason was that this way the baby would sense the mother's heartbeat, something that was familiar from his womb days, and this was embalming, a feedback reinforcing the choice of this position, that rapidly becomes prevalent.

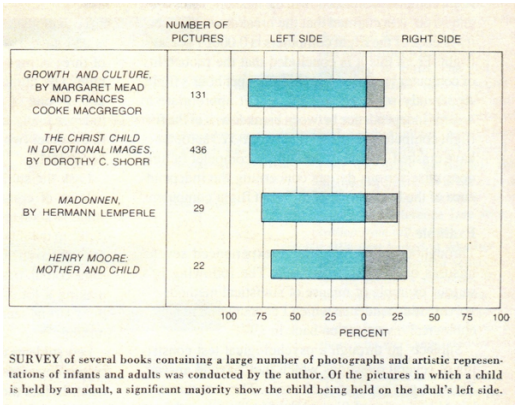

Fig. 3 Patterns of holding a baby.

Salk decided to measure a proxy of welfare of the baby, weight gain under controlled conditions (same feeding and room temperature), but in three rooms there was an artificial heartbeat, in three control rooms this artificial back sound didn't exist. As a whole, discounting statistical sampling fluctuation, babies subject to heartbeat developed better, cf. Fig. 4 although the quantity of food provided was the same: they were calmer, cried less, wasted less energy, and the increased weight gain was the consequence. (Salk further experienced with accelerated heart beat, normal heartbeat and slowed down heartbeat, and babies subject to normal heartbeat were calmer and developed better.)

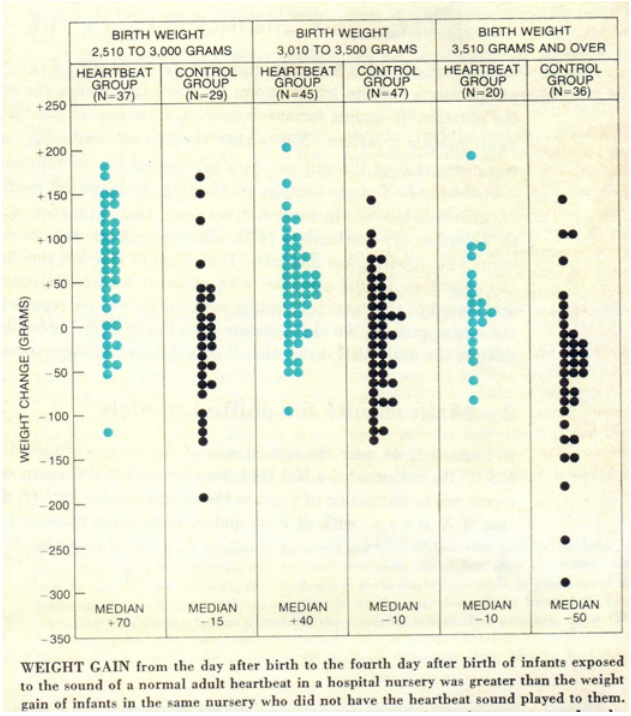

Fig. 4 Weight gain of babies under controlled feeding but in rooms with/without adult heartbeat sound.

Salk research also shows evidence that the pattern of helping babies is a feedback to the response of the newborns to the heartbeat hearing. In fact, mother and baby separation after birth distorts the general pattern, cf. Fig. 5 A clearly devised experimental strategy, excellent control of details, randomisation of the choice

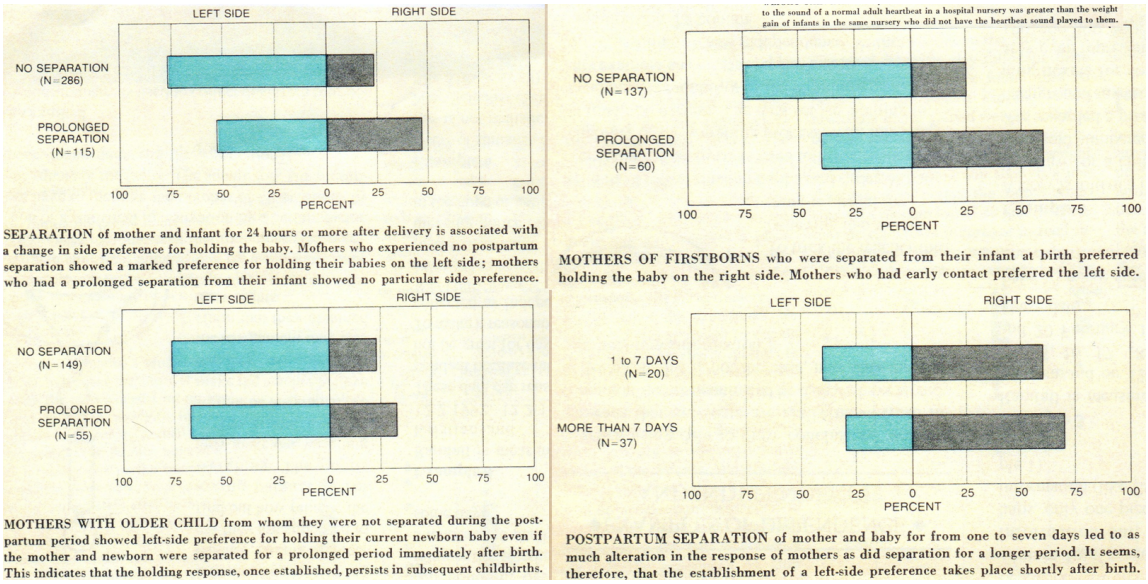

Fig. 5 Mother and child separation after birth precludes the feedback and pattern development

of babies having the experimental treatment and the control treatment, with balanced subsample sizes, have been the keys to this successful investigation, that produced clues favouring recovery of prematures.

2.2 Bad Science

It is very unfortunate, and a discredit to Science, that economic interests, bare fraud, foolishness, produce such quantity of bad science — a famous paper [8] plainly states in the title ”Why most of the published papers are wrong”. The pressure to publish is another source of bad science, and is more and more a cause for concern on the integrity of scientists, cf. [2]. Happily, the crew of *The Annals of Improbable Science* and the board awarding yearly IgNobel Prizes give us some hope that a good lough will contribute to clean so much dung.

The issue of bad science has produced remarkable books, for instance [5], and therefore we shall solely comment on three important questions that are frequently at the root of bad scientific output.

2.2.1 Inappropriate samples

The idea that chance should be considered an ally hasn’t yet penetrated the instruction of science professionals. Random sampling is the unique way of avoiding bias, and collecting representative data. And even

Chance is our best ally in gathering representative data. However, it seems that there is still some prevalence of prejudice against chance, and in many fields and institutions many reports detail that data have

been collected using volunteers. This is an important cause of bias, and very seldom conclusions from such type of research resist the test of time, since other researchers, using other volunteers, are bound to obtain very different results.

Self-selected samples should be avoided. But that doesn't mean that samples collected "at random" by a naïve scientist are truly random. If someones uses a sample of people he interviews in the street, questioning them about controversial issues, this is bound to produce bad data, bad statistics, and bad science — for instance, if someone conducts a simple survey on attitudes on sexual behaviour on a sunday morning near an important church, the results are bound to be different from the ones obtained by someone that asks exactly the same questions on a saturday night in Pigalle.

"At random" is a precise technical expression, that refers to very strict sampling strategies to gather good data, cost being naturally a concern [16]. It is interesting to observe that when someone is asked to extract cards at random from a shuffled deck, almost surely cards will be extracted from different positions in the deck, not realising that simply using the sequence of cards as it is in the shuffled deck is an extraction "at random" in the strict sense. Probability is in most aspects elusive to the layman, and the structures of chance yet mysterious for many people, even with advanced scientific training.

Even properly collected data can produce bad samples; this is in the nature of dealing with uncertainty, but this is the kind of bad samples we can deal with, in the sense that the repeatability criterion of science will surely limit the effects of bad samples. The important problem that we want to raise is the fact that an excellent strategy can be used to select a sample, but at the end an important share of selected sampling units is missing, doesn't respond, disappears from follow-up or from longitudinal studies —-and this is an important source of bias. However, many users of statistics report the prevalence of non-response, but seem unaware that this spoils the data they are using, and that the conclusions drawn are dubious.

2.2.2 Inadequate sample size

Even if the sampling strategy is fool-proof, a sample can be insufficient to produce the conclusions, namely because the sample size is smaller than it should be.

In fact, one of the main features of sampling is to determine the sample size needed to attain a prescribed accuracy, at a probability level. Small samples often produce reports that should, strictly, be inconclusive.

Unfortunately, the temptation to use "many" data — and nowadays automatic data collection, large data sets, resampling techniques, and the use of simulated data easily serve such a purpose — can as well produce bad statistics. In fact, with too many data the most irrelevant differences will be significant. Consider for instance some 2×2 contingency table appropriate to investigate independence of factors (i.e., table with free margins, resting from a dichotomous cross classification of a sample of size n), say

$$\begin{array}{cc|c} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & n(=a+b+c+d) \end{array}$$

The usual test statistic is Pearson's chi-square,

$$X^2_{2,2} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

For instance, if the table is

$$\begin{array}{cc|c} 33 & 54 & 87 \\ 40 & 42 & 82 \\ \hline 73 & 96 & 169 \end{array}$$

then the observed value of the test statistics is 2.025, corresponding to a p-value of 0.155, and at the usual level of significance independence is not rejected. However with the similar table

33×5 54×5	87×5
40×5 42×5	82×5
73×5 96×5	169×5

(strictly similar in the sense that the odds ratios are exactly the same) the observed value of the test statistic is 2.025×5, and the corresponding p-value is 0.0015, leading to straight rejection. It can be argued: OK, it's natural, with bigger size we have stronger evidence.

This is so — but the worrying question is: with a sample size big enough, at the end of the day ANY null hypothesis will be rejected!

This of course raises important philosophical questions that cannot be avoided, namely on meta-analysis and resampling. For a striking example of bad performance of computationally inflated samples, cf. [6].

Another fine point is that there is a time for exploratory data analysis, and there is another time for confirmatory analysis, and that no sample should be used to confirm hypotheses that it suggested. Thus, repeated use of the same sample should clearly be avoided, moreover because its reuse raises important independent questions. Take this also as a caveat against retrospective studies using *ad hoc* data, that in general are inadequate.

2.2.3 Stupid Questions, Stupid Answers

Fool proof methodologies can be perversely used to confer some apparent seriousness to utmost stupid research. Consider for instance the procedure that has been used to solve the ancestral question "what came first, the egg or the hen?"

Two members of a research team phoned one hundred hen farms in the UK, one of them ordering the delivery of a 3 pounds chicken, the other the delivery of one dozen eggs. According to the best practice in planning of experiments, there has been a random choice, with equal probabilities, whether the first call should order eggs or poultry.

In 97% of cases the eggs arrived first. The p-value 1.31543×10^{-25} clearly shows that the null hypothesis that eggs didn't arrive earlier than hens must be turned down!

Doesn't this striking facetious example devised by the *Annals of Improbable Research* team have an echo of some report you have seen? (If you never saw bad science, please read urgently some pearls in `http://www.improbable.com/ig/`.)

Fig. 6 What came first, the egg or the hen?

3 Probability, Statistics, the Experimental Method and Truth

Laplace on Induction

Consider Laplace's urn model: There are $N+1$ (unlabelled) urns, $U_0, U_1, ..., U_N$, the composition of the U_k-th urn being k white balls and $N-k$ black balls. One urn is randomly selected, and at each step a ball is drawn from the urn, and replaced after observation.

This sophisticated model is very special. On one hand, the probability of drawing a white ball in the first step is $\frac{1}{2}$, since in this universe there are, as a whole, as many white balls as black balls, and *conditionally* to the selected urn the extractions are independent, since there is always replacement of the drawn ball in the urn, after each draw. But as the urn is randomly chosen only before the first draw (i.e., the urn is not replaced in the set of unlabelled urns, those shuffled and in the next step random choice of the urn), globally there is dependence. In fact, although we do not know the "state of nature" (the composition of the selected urn), if in the past n draws in exactly w occasions white ball has been observed, then the probability of drawing a white ball in the next extraction is, in the long term, approximately[3]

$$\mathbb{P}(W_{n+1}) = \frac{w+1}{n+2}.$$

Thus, if n draws repeatedly result in white ball, the probability of observing white ball in the $n+1$-th draw is

$$\mathbb{P}(W_{n+1}) = \frac{n+1}{n+2} \quad \underset{n\to\infty}{\longrightarrow} \quad 1.$$

This ingenious model of states of nature is, in a very strict interpretation, an indication that induction based on accumulated evidence leads us to *probable truth*. Observe however that we are not dealing with strong convergence of random variables, we are not implying that $W_(n+1)$ converges almost surely to 1.

Observe also that as the experimenter ignores the state of nature he is working with (unlabelled urn), he is learning from experiment, and it is the information of the results of past draws that are used to compute probability; thus, Laplace's idea of probability is much more sophisticated than "number of favourable simple events over the number of simple events in the universe of reference": probability is *subjective*, always changing from step to step since the experimenter is accumulating information!.

Probability

Probability is a branch of Mathematics whose aim is to tame uncertainty, in the sense that the chances that some event E does happen are quantified by $\mathbb{P}(E) \in [0,1]$.

If $\mathbb{P}(E) = 0$ the event E is almost impossible, if $\mathbb{P}(E) = 1$ the event E is almost sure, and $\mathbb{P}(E) = \frac{1}{2}$ means that there is the same probability of E occurring or not occurring. In terms of taking decisions under uncertainty, we are prone to choose patterns whose probability is "high" (near 1), and to avoid patterns with "low" probability (near 0) , although high and low probabilities are vague concepts. Many times decisions are guided by "expected values", a weighted average of the possible values, using their probabilities as weights. Subjective considerations influence our attitudes towards risk — for instance the expected values when betting in games like euromilions are very low, and the probability of winning a valuable prize ex-

---

[3] This ingenious approximation results from $\frac{1}{N+1}\sum_{k=0}^{N}\binom{n}{w}\left(\frac{k}{N}\right)^{w}\left(\frac{N-k}{N}\right)^{n-w} \approx \binom{n}{w}B(w,n-w)$, where $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}\mathrm{d}x,\ p,q>0$ is the Euler integral of the first kind (beta function).

tremely low, but as each bet is cheap and the reward can be very high, many people are prone to bet regularly in such games.

Aside from the absolute probability of an event E, there is the very useful concept of conditional probability that E does occur given that another event A did happen, $\mathbb{P}(E|A)$; the definition is $\mathbb{P}(E|A) = \frac{\mathbb{P}(E \cap A)}{\mathbb{P}(A)}$. Conditional probability is in fact a probability measure itself, in the restricted sample space whose "universe" is A. Conditioning can either increase, decrease or leave unchanged the absolute probability; when $\mathbb{P}(E|A) = \mathbb{P}(E)$ — and hence $P(E \cap A) = \mathbb{P}(E) \times \mathbb{P}(A)$ — we say that A and E are independent events (the extension to three or more events is handsome, and in fact the very useful assumption of independence, which is a pervasive assumption in many probability theorems, is strictly that: one of the hypothesis assumed to develop some result).

The convention that the probability expresses a measure taking values in [0,1] is extremely useful, since the probability of an intersection of events is the product of conditional probabilities. Using the convention $A_1 \cap A_2 \cap \cdots A_n$ is written as $A_1A_2 \cdots A_n$, from the definition of conditional probability we get $\mathbb{P}(A_1A_2 \cdots A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1) \cdots \mathbb{P}(A_n|A_1A_2 \cdots A_{n-1})$. As the product os values less than 1 is always less than any of the factors, it follows that when we consider very complex occurrences, their probabilities tend to be low.

An interesting consequence is that the probability of an event A never occurring is in the limit 0. In other words, "miracles" do happen, true "predictions" do happen, even though the individual probability of a miracle or of a true prediction is very, very low.

The occurrence of impossible events is in fact something we must regard as quite common. For instance, anybody recognises easily that the probability that choosing a random point in the segment [0,10] and obtaining the exact value of $\pi \approx 3.1415....$ is 0. There is no need to stretch the imagination to realise that the probability of a random choice of a point in [0,10] having a special value x_0 is always zero — but every time we do it, we obtain a result, whose probability *a priori* is 0.

It is rather curious that dealing with uncertainty is a matter of Mathematics, a deductive science that is the paragon of rigour. Although evaluating risks is essential for survival, the mathematical theory began in the rather limited setup of finite sample spaces, assuming equiprobability of the elementary events — a severe limitation, since equiprobability is hardly possible, even in manmade artifacts such as dice or when design with well shuffled card decks —, Jacques Bernoulli (and after him Laplace) assumed a "principle of insufficient reason": if there is not a sufficient evidence that elementary events aren't equiprobable, assume that they are in fact equiprobable.

However, the same Jacques Bernoulli hugely increased the domain of applications of Probability by the end of the XVIIth century, when he discovered what he referred to as his "aureum theorema" (nowadays called the weak law of large numbers: under very broad assumptions, the averages of random samples of increasing size tend to stabilize, more precisely in probability they get closer and closer to the population mean (this result has been further improved by Abraham de Moivre, who discovered an elementary form of the central limit theorem, showing that the rate of convergence of the sample averages towards the population mean is of the order of $\frac{1}{\sqrt{n}}$.

An immediate consequence is the following: in a sequence of independent experiments, marking 1 the fact that the event A is observed, and marking 0 the fact that the event A is not observed, the average number of occurrences (i.e., the relative frequency of A occurrences) goes to the probability of A occurring. This "frequentist probability" can thus be applied to any kind of events, and in fact this discovery was crucial for the early developments of Probability applications (for instance, the construction of the first life table for insurance purposes, by Halley, or the discovery, by John Abuthnott, that the probability that a pregnant woman delivers a boy is higher than the probability that she delivers a girl (John Abuthnott paper is considered the early attempt to construct statistical tests of hypotheses).

There are of course other concepts of probability, for instance considering that probability is a "degree of belief". The idea that probability changes as a function of observed experiments comes from the early work of reverend Thomas Bayes on *a priory* and *a posteriori* probability, in his theorem on inverse probability (also caused probability of causes), and has been independently discussed by Laplace in his important

"Mémoire" on statistical inference presented to the Académie des Sciences de Paris. His very ingenious discussion of what is now known as "Laplace urns" is fascinating, and is a very important contribution towards justifying the use of inductive reasoning (although almost sure, i.e. probability 1, is not certainty).

After a golden period of intellectual interest in Probability, a collection of paradoxes built up by the French mathematician Bertrand, at the end of the XIXth century created the idea that Probability was essentially flawed. In his famous allocution at the 1900 World Congress of Mathematicians, Hilbert listed the need for an axiomatic rigorous construction of Probability among what he considered the great problems defying XXth century mathematicians. The problem has been solved by A. N. Kolmogoroff in his deep essay on the Foundation of Probability, in 1933. In fact, in t he also solves the Borel-Kolmogorov paradox, arising when conditioning to 0 probability events, pointed out by Borel when he started the development of probability in continuous spaces *circa* 1910.

After that the development of Probability Theory and its extension Theory of Stochastic Pocesses never ceased, and in particular the study of random elements in abstract structures (Banach algebras, Hilbert spaces) has clearly shown that many classical theorems rely on the geometry of metric spaces. Many results of the fertile activity of probabilists do not have so far any application in Statistics, they are the result of "pure" research, of the compulsion to create new consequences of this vast theory, namely by the creation of new definitions and the exploitation of alternative structures (for instance, while independence is a simple concept, dependence structures have infinite variations, and the study of robustness of results under mild forms of dependence will eventually be for ever a fertile area of research).

Statistics

Only very elementary "descriptive statistics" deals with statements about the observed sample; the same happens in some areas of what is called "official statistics", namely when there is a census of the whole population.

Modern "Mathematical Statistics", that started with the work of Galton on regression (1889), of his associate researcher Karl Pearson invention of correlation, and his presentation of the "chi-square test" (end of the XXth century), and of 'Student' (William Gosset) on "the Probable Error of the Mean" (1908), had crucial developments by Sir Ronald Fisher (likelihood, rediscovery of cumulants, ANalysis Of VAriance — ANOVA —, planing of experiments), and by the end of the thirties Mathematical Statistics had developed quite successfully the theory of confidence intervals, the theory of statistical test of hypothesis (Neyman and Egon Pearson), the elegant theory of exact result for small samples in 'normal" populations, and an alternative "nonparametric" inference when the normal assumption failed, or was dubious.

With the need to restrain costly (and often destructive) sampling, Wald developed sequential statistical analysis (considered a war secret during Worl War II); with the advent of routine inexpensive intensive calculation machines, more sophisticated models than the classical normal model entered the rich ground of statistical modeling, and stochastic simulation became also a tool for computationally increasing small samples with "fake" observations generated from the empirical distribution of the observed sample (something like the miracle of multiplication of bread and fish...). An ingenious consequence reassessment of the consequences of the law of large numbers led Ulam to create a numerical way of performing intricate computations, while he was working with the atomic bomb developers team, which he called Monte Carlo simulation, and that is in fact a clever use of averages of very large simulated samples. The recognition that many models can be adjusted to the data, with varying degrees of sophistication, clarified the idea that we should never think of "the model" for the data, bur rather of "one useful model" for the data. The possibility that the model we choose deviates from what is appropriate for the population led to a rich theory of "robustness" in statistics.

While in the early stage of the development of Statistics the main problem was the scarsseness of data, new devices for automatic gathering of data (for instance in meteorological plants, or stock exchange of

assets, options and futures) created the need for stochastic "data mining" of relevant data. Cooperation of diverse research groups led to the need of harmonizing conflicting evidence, or of investigation whether the combination of non-significant studies provided sound ground for some decision, the new flourishing field of Meta-Analyisis.

While the many branches of modern Mathematical Statistics have diverse aims and use eventually different mathematical tools in their development, there is an important common goal: to use limited information from a sample to inductively infer to the whole population, to forecast future results, to control the performance of complex devices. So, all Mathematical Statistics must take into account uncertainty inherent to any inductive construct. The goal is to tame, to limit, to quantify this inherent uncertainty unavoidable when dealing with the real world.

Hence, Probability Theory is the language of Statistics, since it is the tool tuned to understand and quantify uncertainty, namely in all processes of generalising for a whole population what has been observed in a limited sample, which of course must be considered as essentially 'random", in the non-technical sense that chance intervened in its collection.

So, Probability and Statistics are closely interwoven; inductive Statistics (the core of Statistics) cannot exist without Probability, and on the other way Statistics is always stimulating new probability research. But Probability and Statistics are clearly different subjects, with different aims; while probability is a pure Mathematics field, Statistics encompasses many real world issues and applications, and has to take into account many non-mathematical questions for cleverly dealing with many problems (for instance body language when collecting data in interviews, psychological issues in questionnaires). While Probability results are fool proof, the many abuses of Statistics clearly show that its use is ground in quicksand and prone to exhibition of human stupidity (the only thing that Einstein believed to be infinite).

However, good statistics must be considered an essential tool in knowledge building, in fact its interplay with the methodology of scientific research in experimental areas is crucial for the education of any researcher, and an important issue in The Philosophy of Science.

4 Conclusion

Richard von Mises [11] *Probability, Statistics and Truth* raises many questions on the role of Probability and Statistics in modern conceptions of Science; von Mises theory of collective is a sound basis of the frequentist paradigm of Statistics. On the other hand, de Finetti's [3] exchangeability has been a landmark on the bayesian construction of statistical theory. Jaynes [9] is a lively discussion of Probability and Statistics, and the fact that the author died before revising and completing the manuscript had perhaps some bearing on its challenging style. Perusing Pólga [14, 15] is also a pleasure.

The complexity of modern science had indeed a fertilising role on the development of Statistics, and *The Grammar of Science* [13] has been most influential, namely by pointing out that in complex matters causality should be abandoned in favour of statistical association, cf. also [10] for a modern discussion of causality.

Although Laplace's effort to justify induction are worth praise, his interpretation of urns composition as representing the "states of nature" is farfetched. Probability of obtaining the result invariably obtained in infinite previous repetitions of the experiment increases towards 1, but we are not dealing with almost sure convergence. And, as twice discussed in the previous sections, the impossible does happen. But the important point is that Laplace's construction is indeed artificial, since only in Mathematics can we envisage seriously endless repetitions of the experiment. In actual science research, there is a gross limitation on the sample size / number of experiments that can be carried out. Hence, there is always room for alterations, and the revissibility is a strong point upholding our trust in Science. Probable truth is perhaps what we can hope for in inductive reasoning, and any attempt to convert induction in deduction is pernicious, as Husserl [7] comments

"Descartes lui-même s'était donné d'avance un idéal scientifique, celui de la géométrie, ou, plus exactement, de la physique mathématique. Cet idéal a exercé pendant des siècles une influence néfaste."

(we do not read German, so we used the french translation we possess in the citation). The influential role of Probability and Statistics in scientific research, using high level confidence intervals, small p-values of the null hypothesis rejection region, maximum likelihood is deserved. But the value of those excellent tools should not blind us in what concerns the role of null-probability events, and truth is like the holly land (Pär Lagerkvist), something we search but never attain. Professor Rohl (*The Reader*, 2008) says a devastating truth when he states that the purpose of the courts of justice is to enforce the law, not to make justice. We think that we have also to recognise that science cannot establish truth, it only attains probable truth. Fortunately, the remissibility of scientific findings vouches for our confidence in the fact that scientific probable truth, like Zenon's arrow, will endlessly be closer and closer to the target,

Acknowledgements This research has been supported by National Funds through FCT — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011.

References

1. Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observed in the birth of both sexes. *Philosophical Transactions of the Royal Society of London* **27** 186–190.
2. Arnold, D. N. (2009). Integrity Under Attack: The State of Scholarly Publishing. *SIAM News* bf 42. `http://www.ima.umn.edu/~arnold//siam-columns/integrity-under-attack.pdf`
3. de Finetti, B. (1974). Theory of Probability. A Critical Introductory Treatment, Wiley, Chichester.
4. Fisher, R. A. (1995). *Statistical Methods, Experimental Design and Scientific Inference* (re-issue of *Statistical Methods for Research Workers*, *The Design of Experiments*, and *Statistical Methods and Scientific Inference*), Oxford Univ. Press, Oxford.
5. Goldacre, B. (2009). *Bad Science*, Harper Perennial.
6. Gomes, M.I., Pestana, D.D., Sequeira, F., Mendonca, S. and Velosa, S. (2009). Uniformity of Offsprings from Uniform and Non-Uniform Parents. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, p. 243-248.
7. Husserl, E (1953). *Mditations Cartsiennes. Introduction la Phnomnologie*, Vrin, Paris.
8. Ioannidis, J. P. A. (2005). Why most published research findings are false, *PLoS Med.* **2**(8):e124, 696–701.
9. Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge.
10. Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, Cambridge University Press, Cambridge.
11. von Mises, R. (1981). Probability, Statistics and Truth. Dover, New York.
12. Neyman, J., and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences***231** 289–337.
13. Pearson, K. (1896). *The Grammar of Science*, reprinted by Cosimo Classics (2007).
14. Pólya, G. (1990). *Mathematics and Plausible Reasoning, Volume 1: Induction and Analogy in Mathematics*, Princeton University Press, Princeton.
15. Pólya, G. (1990). *Mathematics and Plausible Reasoning, Volume II Patterns of Plausible Inference*, Princeton University Press, Princeton.
16. Pestana, D., Rocha, M. L. and Sequeira (2013). Sampling strategies and costs control. Notas e Comunicaes do CEAUL, 14/2013.
17. Salk L. (1973). The role of the heartbeat in the relations between mother and infant.. *Sci Am.* **228**, 24–29.