FORENSIC STATISTICS: A GENERAL VIEW

Benito V. Frosini

Department of Statistical Sciences, Catholic University of Milan, Italy

Abstract*. This paper introduces many basic topics connected with applications of statistical methods to the domain of trials in a court of law. After resuming the different decision criteria applied to civil and to criminal trials, as well as the caution to be adopted in evaluating apparent correlations, and briefly commenting on the U.S. Federal rules of evidence, the paper exposes important topics such as: the application of Bayes formula and related inference approach, the fallacy of the transposed conditional, the role of so-called naked statistics, the meaning and domain of application of significance tests. The application of these concepts and approaches is illustrated throughout by many reported cases.*

Key Words*: Bayes' theorem, Fallacy of the transposed conditional, Naked statistics, Tests of significance.*

1. INTRODUCTION

By *forensic statistics* we mean the application of statistical methods, mainly in the inference domain, to the several problems encountered in civil and criminal trials. This kind of application is seldom commonplace or trivial; on the contrary, it has distinctive features, both for some special problems encountered in many law suits and criminal trials (of which some account will be given in the following sections), and for apposite ethical issues pertaining to the expert witnesses in trials. Writes Kadane (2008, p. 108): «I find it interesting that the legal context impinges on the data analysis in several places. While it is to be expected that the application would have a strong influence in every applied problem, it is somewhat surprising that an analysis done in a legal context might be substantially different from an analysis done with a solely scientific aim».

Actually, the distinction features hinted at above have risen – in the latest fifty years – quite an impressive amount of scientific literature, in books and papers; to this purpose it is worthwhile to mention at least some of the scientific journals which are devoted, or accept as a rule, papers concerned with applications of statistics in the forensic domain.

Forensic Science International, *Journal of Forensic Science, Science and Justice* play a relevant role in the forensic community; *Law, Probability and Risk* is the leading journal for jurists, and a variety of papers dealing with forensic statistics are regularly published in statistical journals such as *Journal of the Royal Statistical Society* (*A, B and C*).

In any case, we must point out that many law journals, as well as many statistical journals, usually accept quality papers which deal with typical problems of forensic statistics; some quotations of this kind will appear in the following sections.

Many problems presented in a court of law require to establish the personal responsibility for certain actions, thus they pertain to the old and hard domain of the search of the cause for a given effect (with an important overlapping with traditional philosophical research). However, although looking for the cause of a given event is common to many trials in court, there is a fundamental distinction between civil and criminal cases, as far as the conclusion or judgment is concerned. It is commonly accepted that – for ethical reasons – a guilt in a criminal case must be reached under quite a strong criterion, i.e. with a belief *beyond a reasonable doubt*; quite a weaker criterion is instead applied in civil suits, where (at least in principle) the judgment is decided on the basis of the *preponderance of probabilities* (which is equivalent to the criterion *more probable than not*); in other words, the reference probability in this case is 0.50.

It often happens, especially when quantitative variables are concerned, that an *apparent* link (although not deterministic) between two or more variables yields a large (absolute) *correlation*, which can be an indication of a possible cause-effect relation; however, statisticians are well aware that «correlation is not causation» (Barnard, 1982). In fact, many apparent correlations emerge from the associations of two variables in a multiple time series: for example, if we examine the per capita tobacco consumption and the life expectation in Italy in the latest eighty years, a very large positive correlation results, but nobody could derive that the increase in the length of human life is mainly ascribable to the corresponding increase in tobacco consumption (Frosini, 2009, pp. 342-343). As in this case, it often happens that the variables under consideration are subjected to a concomitant variation of a *common cause* (see e.g. Frosini, 2006, p. 310; Garbolino, 2014, pp. 80-81).

One of the most important concepts – often neglected – about making an inference about some population characteristic (or parameter) is the *reference set*, or *reference population*, in which the observed sample can be merged. This reference is essential in order to correctly computing event probabilities related to a possible perpetrator of a crime (Coleman and Walls, 1974; Garbolino, 2014, p. 71 and elsewhere), but it shares a much more wide domain of application (Fisher, 1937, 1959; Frosini, 1999; Meier, Sacks and Zabell, 1986; 7-8), pointing to useful subsets of the original sample space and to informative ancillary statistics.

Since 1973 the U.S. Supreme Court has introduced a new set of *Federal Rules of Evidence* (FRE), successively amended several times; in particular, by dealing in 1993 with the landmark case *Daubert v. Merrill Dow Pharmaceutical, Inc.* «the Court held that under Rule 702 of the Federal Rules of Evidence the trial judge was required to be a "gatekeeper" for scientific evidence. Before allowing such evidence to be heard by a jury, the judge has to determine that the proposed testimony was not only relevant but also "reliable". The personal opinion of a qualified expert was no longer enough; the expert's pronouncements had to be based on scientific knowledge" (Finkestein and Levin, 2004, p. 40). Concerning the effects of thousands of breast implant cases, and of millions of Bendectin users (some of which had children with defective limbs), an immediate outcome of the new FRE's was a much more rigorous request of acknowledgeable scientific research, and especially of serious epidemiological studies, approved by the scientific community. This new wave led to rejecting almost all claims of compensation in thousands of trial for lack in motivation; actually, practically all the available epidemiological studies were in favour of the defendant (Zeisel and Kaye, 1997; Stella, 2001, Chapters 3-5; Frosini, 2002, pp. 41-44; Finkelstein and Levin, 2004). Basically, the request of the U.S. Supreme Court, in order to sustain a given claim, was the availability of epidemiological studies approved by the scientific community.

Before passing to a more formal treatment of the application of probability calculus to some kinds of inference encountered in civil and criminal trials, it seems proper to make some hints to some classical works issued in the second half of the eighteenth century, which were destined to leave a permanent track in our history. The first work of this kind is the celebrated booklet *Dei delitti e delle pene* (*On crimes and punishments*) by C. Beccaria (1764), soon translated into many languages (1766 into French and German, 1767 into English, 1774 into Spanish). This work is especially cited for the many pages written by Beccaria against torture and death penalty, but it contains also several considerations on specific trial topics, such as *Witnesses* (Section XIII) and *Evidence and forms of judgments* (Section XIV). In this last section we find classifications of proofs into *dependent* and *independent*, and also into *perfect* and *imperfect* ones. About independence Beccaria writes (quoting from the translation by D. Young, p. 26): «When the proofs are independent of one another … then the likelihood of the fact increases as more proofs are adduced, because the flaws in one proof have no bearing on the others. I speak of probability in criminal cases, even though certainty ought to be required if punishment is to be inflicted… Moral certainty is only a sort of probability». And about the quality of proofs we find the following reflection: «I call perfect those proofs that exclude the possibility that a given person may be innocent; I call imperfect those that do not exclude it. A single proof of the first sort is sufficient for conviction; of the second sort, as many are required as are needed to form one perfect proof». This small work by Beccaria exercised an important impact on other juridical contributions of that time, to begin with the celebrated *Commentaries on the laws of England* by W. Blackstone (1769, in particular in Chapter 1 of Book IV, entitled *On the nature of crimes and punishments*).

The essay by Voltaire (1772) *Sur les probabilités en fait de justice* was mostly devoted to commenting on the so-called *Affaire Morangiés*, with exposition of a number of probabilities *in favour* and *against*; unfortunately Voltaire was not able to elaborate all these probabilities by means of a Bayesian network, such as the one worked out by Kadane and Schum (1996) for the Sacco and Vanzetti trials (to be resumed at Section 4.2 of this paper). Anyway, the Introduction of Voltaire to his essay contains a few wise words about trials in general, that we recognize of general validity even today. Voltaire makes a sharp distinction between *civil* and *criminal* trials: «Dans le *civil*, tous ce qui n'est pas soumis à une loi clairement énoncée est soumis au calcul des probabilités … alors la plus grande probabilité vous conduit [principle of *preponderance of probabilities*]. Il ne s'agit que d'argent. Mais il n'en est pas de même quand il s'agit d'ôter la vie et l'honneur à un citoyen. Alors la plus grande probabilité ne suffit pas … Il se peut que vingt apparences contre lui soient balancées par une seule en sa faveur [principle of *beyond a reasonable doubt*].»

2. PROBABILITIES OF CAUSES. BAYES' FORMULA

In most trials there is an interest in going back, from a known fact or event, to the *cause* which is deemed responsible for the occurrence of the same event. In the sequel we will generally assume that the given event can be the *effect* of one of k possible causes $A_1,\ldots,A_k$ (the general problem of the plurality of causes is dealt with by Stella, 2000, p. 297). It is generally granted that a given effect can be the outcome of different causes; however, the maintained plurality of causes is incompatible with deductive inference as usually shaped, where the cause itself is understood as a *necessary and sufficient condition* for a given effect (Copi, 1964, p. 407). If more than one possible cause is assumed, then no deductive inferences – from a given effect towards the generating cause – are admitted. It must be said, however, that the *apparent* plurality of causes usually disappears when the effect is specified with great precision. Moreover, if we possess sufficient information for excluding all possible causes but one, the cause surviving this scrutiny is just the cause we are looking for, beyond any doubt (Frosini, 2002, pp. 45-46).

In the sequel it will be assumed that the information at our disposal does not allow the identification of just one specific cause; therefore, for the effect *B* we generally admit the existence of *k* possible causes $A_1,\dots, A_k$. With the aim of performing an *inductive inference*, from the effect *B* towards every possible cause, it is nevertheless necessary to assume – and employ – another kind of information, namely a probabilistic information as to the efficiency of every cause in producing the given effect; for example, the cause A_1 is very efficient in producing the effect *B*, while the presence of A_2 is able to produce *B* only in a very limited fraction of the cases. With respect to the possible cause A_r such synthetic measure is the *probability of B given* A_r (for $r = 1,\dots, k$)

$$P(B \mid A_r) = L(A_r) \qquad r = 1, 2,\dots, k \tag{1}$$

also called the *likelihood* of A_r for *B*.

The above information concerning the likelihoods must be completed with another information, concerning the probabilities of the causes

$$P(A_r) \qquad r = 1, 2,\dots, k \tag{2}$$

also called prior, or *a priori* probabilities; the more probable is the cause A_r, the more must be high (other conditions held constant) the probability that the given effect comes out from the given cause. From the above, it is highly intuitive that the probability $P(A_r|B)$ (probability of cause A_r given the effect *B*) be defined as *proportional* to the product of the prior probability of A_r and the likelihood of A_r:

$$P(A_r \mid B) \propto P(A_r)\times P(B \mid A_r) \qquad r = 1, 2,\dots, k.$$

The same formula, comprising the proportionality coefficient, is the celebrated Bayes' formula

$$P(A_r \mid B) = \frac{P(A_r)\times P(B \mid A_r)}{P(B)} \qquad r = 1,\dots,k \tag{3}$$

where

$$P(B) = \sum_{r=1}^{k} P(A_r)\times P(B \mid A_r)\,.$$

This formula comes out very simply from the general formula for a conditional probability

$$P(A_r \mid B) = \frac{P(A_r \,\&\, B)}{P(B)}\,;$$

this formula allows to exchange the two events A_r and B with respect to the likelihood $P(B \mid A_r)$: from the probability of effect B given the cause A_r we pass to the probability of cause A_r given the effect B (for a general reference about the Bayesian approach to inference we may refer to the volume of Bernardo and Smith, 1994).

The comparisons between probabilities of kind $P(A_r \mid B)$ can be simplified when they are performed by means of ratios of kind (by making reference, to fix ideas, to causes A_1 and A_2):

$$\frac{P(A_1 \mid B)}{P(A_2 \mid B)} = \frac{P(A_1)}{P(A_2)} \times \frac{P(B \mid A_1)}{P(B \mid A_2)} \tag{4}$$

by calculating the above ratio the value $P(B)$ is eliminated, thus the same ratio results from the product of two ratios, i.e. between the prior probabilities and the likelihoods.

A particular application of the formula (4) concerns the two complementary events, of relevance in many trials, $A_1 = G$ (*guilty*, in the sense that a given individual is guilty), and $A_2 = \overline{G}$ (*not guilty*, or *innocent*[1]); in this kind of application the effect B is usually denoted by E, meaning that E is the factual *evidence* presented in the trial. Thus the above ratio is presented in the form of the following *odds ratio*:

$$\frac{P(G \mid E)}{P(\overline{G} \mid E)} = \frac{P(G)}{P(\overline{G})} \times \frac{P(E \mid G)}{P(E \mid \overline{G})} \tag{5}$$

which is equal to the product *odds ratio* between the prior probabilities and the likelihood ratio. Basically, the likelihood ratio transforms the original ratio between the prior probabilities (determined previously to the knowledge of the evidence E) into the ratio of final probabilities (after gaining the information provided by E). Thus the ratio between the prior probabilities, multiplied by the likelihood ratio (which can be >=< 1), can rise (in favor of guilt), or remain constant, or else diminish (in favor of innocence).

The above probabilities, namely the prior probabilities $P(A_r)$ and the likelihoods $P(E \mid A_r)$, may be – in particular applications – strictly objective or merely subjective. Of course, they can be graduated between the limits of purely objective

[1] Note that propositions generally put forward by parties at trial refer to source attribution, given activities, or crime commission. Here, for sake of illustration a *guilty/innocent* pair of propositions are used. Please refer to Cook et al., (1998) for a detailed discussion on the hierarchy of propositions.

and purely subjective probabilities. In many applications the likelihoods are eminently objective; on the contrary, the prior probabilities are mostly of a subjective character (although possibly based on reliable information about the trial evidence). This last observation relates to the fact that such probabilities refer to *unique events*; as already pointed out, this kind of events cannot properly belong to a sample space formally defined (Frosini, 2009, p. 190). In these cases, rather widespread, such probabilities essentially express the personalistic belief (graduated between 0 and 1) which a given individual attaches to the occurrence of an event (e.g. A_r, or $E \mid A_r$) (Bernardo and Smith, 1994, p. 4).

As a provisional conclusion about the application of the Bayesian approach to an inductive inference, we list in the sequel the assumptions and pieces of information deemed necessary for the formal enforcement of the Bayes's formula:

(1) we must be sure that the causes $A_1, \dots, A_k$ of the evidence E exhaust the whole set of possible causes (which is sometimes hard to assess); in the application to criminal trials this is formally ensured when the reference is to two complementary causes, i.e. G (*guilty*) or $\overline{G}$ (not guilty, or innocent);

(2) all prior probabilities $P(A_r)$, $r = 1, 2, \dots, k$, must be available from the outset, unless only a probability ratio like (5) is required;

(3) also the likelihoods $P(E \mid A_r)$, $r = 1, 2, \dots, k$, must be available (be they objective or subjective);

With all these pieces of information it is possible to calculate the *final probabilities*, or *posterior probabilities*, in application of the Bayes's formula. Such probabilities are usually of a subjective nature, but they can be evaluated as substantially objective when most prior probabilities and most likelihoods can be assessed as objective (i.e. practically accepted by everybody).

Before resuming this same topic at Section 6, when a radically different approach to statistical inference will be exposed, it will be expedient to mention an inherent weakness of the Bayesian approach (which is present also with other approaches), well summarized by Taruffo (1992, p. 181): «the application of the Bayesian methodology … is scarcely suited for the trial context, because it neglects the aspect which is most important in this contexts, namely the weight and meaning of the evidential proofs which are available in every case». This same judgment conforms to the evaluation of Keynes (1921, p. 313): «the degree of completeness of the information upon which a probability is based does seem to be relevant, as well as the actual magnitude of the probability, in making practical decisions… If, for one alternative, the available information is necessarily small, that does not seem to be a consideration which ought to be left out of account altogether».

3. THE FALLACY OF THE TRANSPOSED CONDITIONAL

In the statistical literature it is usual to call “fallacy of the transposed conditional” (or “prosecutor’s fallacy”) a true and genuine fallacy, whose knowledge is less widespread than the traditional fallacies that go back to the Aristotelian logic, mostly because its correct understanding implies the knowledge of the essential tools of the theory of probability (Aitken and Taroni, 2004, p. 112; Garbolino, 2014, p. 103). The practical effects ensuing from this fallacy, unfortunately recognized in some sentences of real criminal trials, can be ravaging, thus this fallacy must be acknowledged and possibly avoided.

With reference to the symbols introduced above, let E be the evidence (effect) presented at the trial, G the event that a given individual is guilty, and $\overline{G}$ the complementary event that such an individual is not guilty. Therefore the sum of the two probabilities $P(G \mid E)$ and $P(\overline{G} \mid E)$ is one. Let us fix our attention on the probability $P(\overline{G} \mid E)$, i.e. the probability of innocence given the evidence (from which the probability of guilt $P(G \mid E) = 1 - P(\overline{G} \mid E)$ can be immediately derived, if necesssary). It is obviously convenient that such probability $P(\overline{G} \mid E)$ must be distinguished, and not confused, with respect to the probability $P(E \mid \overline{G})$, i.e. the probability that the evidence E comes out from an innocent person. When the conditional role of the events E and $\overline{G}$ is reversed, the *fallacy of the transposed conditional* arises; for example, the fact that the evidence E is normally produced by a guilty person, is mistakenly interpreted as an indication of guilty for the given suspected individual.

As a more specific example, let us assume that we know the blood group E (rather uncommon) of an individual that was on the crime scene just before a crime was committed; this information *could* change the position of a given individual XY from vaguely suspected to probably guilty, if we make the following (pseudo) reasoning: it is rather uncommon that the evidence E be observed on an innocent individual, thus it is unlikely that XY is innocent. For example, from $P(E \mid \overline{G}) = 0{,}01$ we *derive* $P(\overline{G} \mid E) = 0{,}01$, and finally, for the complementary event G, $P(G \mid E) = 0{,}99$.

We may comment on this example by admitting that “it is really uncommon that the evidence E be observed on an innocent person”, however the above conclusion cannot ensue from the above pseudo-reasoning. Actually, from the *sole* knowledge of the probability $P(E \mid \overline{G})$ no deduction can be made about the other probability $P(\overline{G} \mid E)$; for example, one of the two probabilities can be near 1, and the other near 0. By using the formula of conditional probability, we can write

$$P(\overline{G} \,\&\, E) = P(\overline{G}) \times P(E \mid \overline{G})$$

hence

$$P(\overline{G}\,|\,E)=\frac{P(\overline{G}\,\&\,E)}{P(E)}=P(E\,|\,\overline{G})\times\frac{P(\overline{G})}{P(E)}.$$

The value of $P(\overline{G}\,|\,E)$ can thus be obtained from $P(E\,|\,\overline{G})$ after multiplication by the ratio $P(\overline{G})/P(E)$. Continuing this same example, we can observe that if the reference population (comprising all the people that could be possibly guilty of the given crime) is very large, the probabilities of observing the evidence E on an individual drawn from the whole population, or from this same population excluding the guilty person, are practically coincident: $P(E)\approx P(E\,|\,\overline{G})$; then with a very good approximation we can write $P(\overline{G}\,|\,E)\approx P(\overline{G})$, thus ensuing that the events $\overline{G}$ and E are practically independent, namely that the *sole* knowledge of E is practically useless for the identification of the guilty person.

4. SOME COMMENTS ABOUT TWO *CAUSES CÉLÈBRES:* DREYFUS AND SACCO-VANZETTI

The scientific literature about juridical and statistical aspects of two well known *causes célèbres* (Dreyfus in France and Sacco-Vanzetti in the United States) provides some useful elements for a short comment on relevant probabilistic aspects in the respective trials. The bibliographic references, sufficient to frame both trials, are Champod et al (1999), Frosini (2002), Aitken and Taroni (2004) and Garbolino (2014) for the Dreyfus case, and Kadane and Schum (1996) for the Sacco-Vanzetti case.

4.1 THE DREYFUS CASE

As already hinted, both cases are widely known, and a large literature (mostly of a narrative kind) has developed in the past decades. To begin with the first case, we may recall that Alfred Dreyfus, an officer in the General Staff of the French Army, was accused in 1894 of selling military secrets to the German embassy in Paris; the accusation relied on a document (called *borderau*), written by him, which – according to the accusation – was a forged document and contained a cipher message. In 1895 Dreyfus was convicted of high treason, and sentenced to life imprisonment on Devil's Island in French Guiana.

An authentic rebellion took place among the French intellectual people, to begin with the famous open letter *J'accuse!*, addressed by Émile Zola to the President of the French Republic. After another trial, Dreyfus was condemned in 1899 to *only* ten years of imprisonment; afterward he was pardoned by the President

E. Loubet. Full justice for Dreyfus only occurred when a new trial took place in 1904; as a conclusion of this last trial he was officially exonerated, and readmitted in the Army with a promotion to the rank of Major. Many years later the German Army acknowledged that the their informer was Walsin Esterhazy, Major in the French Army.

We pass now to comment on the two principal “inductive reasonings” that (incredibly) nailed Dreyfus in his first trial (Tribe, 1971; Frosini, 2002; Aitken and Taroni, 2004; Garbolino, 2014). A fact, then accepted as incontestable evidence against Dreyfus, was the number – not specially high but not really trifling – of coincidences in his letter (so called *borderau*) with respect to a *model* that Dreyfus *would* have followed in order to prepare a cipher message. An example, brought in the trial by the expert A. Bertillon, was the following: if we accept that the probability of a sole coincidence is 0,2, then the probability of four coincidences is $(0{,}2)^4 = 0{,}0016$, quite a small probability if we admit mere causality, and independence of the same coincidences. According to Bertillon we can be reasonably convinced that a like event, so less probable under mere causality, be instead intentionally forged by Dreyfus.

While provisionally accepting Bertillon’s viewpoint (which nevertheless was later checked as mistaken – see Frosini, 2002, Aitken and Taroni, 2004), we can immediately observe that the conclusion of his argument is not justified, as it is an evident case of the fallacy of the transposed conditional. In fact, the probability used by him is of kind $P(E\,|\,\overline{G}\,)$, while it is “interpreted” by Bertillon as $P(\overline{G}\,|\,E)$, with a logical bound wholly unjustified.

Another logical mistake – leading to the same conclusion, however allowing to comment on another kind of mistake – presented in the first trial against Dreyfus, was itself based on an undeniable fact: the letters of the French alphabet comprised in the *borderau* did not show the *normal* proportions observed in the French prose. In particular, it was pointed out that the observed proportions (in Dreyfus document) have a very small probability of occurring (and that was true); as a gratuitous conclusion, it was again assumed that the writing of the letter was prepared just to include a cipher message.

However, the authors of this “good idea” have omitted (!) to notify that *any* distribution of proportions (of the alphabet letters) has a very small probability, if the calculation is performed by assuming mere randomness in the choice of the letters (not a reasonable assumption, but this point is not under discussion here). (Aiken, 1995, p. 78; Frosini, 2002, pp. 81-82).

In any case, the above mistake is therefore another clear example of the *fallacy of the transposed conditional*: having observed that $P(E\,|\,\overline{G}\,)$ (probability of the evidence E under the hypothesis that the document has not been forged) is very small, the apparent *deduction* was that $P(\overline{G} \mid E)$, i.e. the probability that Dreyfus is innocent given the evidence, is equally small, hence that the probability of guilt $P(G \mid E)$ is very large.

4.2 THE SACCO AND VANZETTI CASE

Nicola Sacco and Bartolomeo Vanzetti were charged of the murder – happened during an armed robbery – of Alessandro Berardelli and Frederick A. Parmenter. The robbery and the shooting took place at South Braintree, Massachusetts, on 15th April 1920. Following police inquires, Sacco and Vanzetti were suspected, and then arrested. The formal act of indictment was signed by a judge on 11th September 1920. An immediate retaliation from the anarchist organization (to which Sacco and Vanzetti belonged) soon followed: on September 16 Mario Buda placed a bomb, set to go off at noon, at a corner of two streets in New York; thirty-three persons were killed and more than two hundred were injured. This bomb killed no members of the government so despised by the anarchists but secretaries, stenographers, and other innocents on their way to lunch. Passing through various nets designed to catch anarchists, Buda returned to Italy and was never apprehended (Kadane and Schumn, 1966, p. 9). In the first trial, lasted six weeks, the twelve jurors were unanimous in concluding that Sacco and Vanzetti were guilty as charged (Sacco because he actually shot at Berardelli and Parmenter, Vanzetti because he was an accomplice). Both Sacco and Vanzetti protested their innocence, as they had done while testifying at trial on their own behalf. The following appeals trial lasted for six years; it confirmed the first trial sentence; Sacco and Vanzetti were executed on 23rd August 1927.

Joseph B. Kadane and David A. Schum (1996) have performed a very deep and complete analysis of all the facts and all the testimonies resulting from both trials. Contrary to what happened in the Dreyfus case, where the “experts” employed – although in a distorted manner – objective probabilities, in the Sacco and Vanzetti case all probabilities assessed by Kadane and Schum were of a *subjective* kind (or *epistemic*, as they were usually called by the authors), as they were necessarily referred to *unique* events. Such probabilities have been associated with several elements of uncertainty, or doubt, found in the trials, with the aim to assess the probative or inferential *force*, *strength* and *weight* of the evidence. This has been made by following an atomistic approach, namely by spotting all the elementary components of complex facts, and then connecting such components by means of logical relations, or *chains of reasoning*. The authors clarify their procedure as follows (p. 26): «The degree of detail we employ in constructing our chains of reasoning regulates the “resolving power” of the conceptual microscope we have focused on the evidence in this case». Actually, an *atomistic* approach makes it easier to spot singular elements of doubt or uncertainty, with respect to a holistic approach.

All the *elements of evidence*, identified by the application of this atomistic procedure, have been linked by Kadane and Schum by means of suitable *inference networks,* having a DAG structure (DAG = *Directed Acyclic Graph*); the elements of a DAG are linked by means of directional segments – or arrows – however not allowing for loops (in fact they are *acyclic*) (Kadane and Schum, 1996; Frosini, 2006; Garbolino, 2014). To each arrow (linking two elements or events) is associated an epistemic probability; it is thus possible to run along linked events within a DAG by successive applications of the Bayes formula; in this manner, the probability of an event is determined according to the contributions of all the probabilities of its “parents” (Pearl, 2000; Spirtes, Glymour and Scheines, 2000; Frosini, 2006; Garbolino, 2014).

The proposal of inferential networks, with the aim of linking (by means of probabilistic relations) several events which are relevant in a trial, dates back to John H. Wigmore, for many years dean of the Law School at Northwestern University (Wigmore, 1913, 1937). In the past thirty years a number of commercial algorithms have been produced that allow to implement and control complex inferential networks, such as those actually employed by Kadane and Schum and applied for the Sacco and Vanzetti case analysis. An interesting aspect of these softwares concerns the possibility of performing some kind of *sensitivity analysis*, namely of checking the effects on the network (and particularly on the conclusions of the inferential process) when given changes are inserted in some epistemic probabilities; it is thus possible to *test the robustness* of the same networks.

Although the inferential procedure employed by Kadane and Schum is clearly aimed at evaluating final epistemic probabilities of guilt, these authors express a *qualitative* appreciation about the practical “interpretation” of the criterion BRD = *Beyond Reasonable Doubt*, quite in agreement with the viewpoint expressed by Tribe (1971), Cohen (1977, pp. 247-252) and Stella (2001). Actually, the two authors write: “On Cohen’s Baconian view, “beyond reasonable doubt” simply means that all relevant reasons for doubt have been eliminated, with no let-outs or qualifications. In our analysis of just the trial evidence, we believe there are significant doubts remaining about both Sacco and Vanzetti participation in the South Braintree crime. So we cannot say that the evidence at trial was complete in covering matters we judged relevant on the basis of specific arguments we constructed from the trial evidence itself” (Kadane and Schum, 1996, p. 282). The conclusion of Kadane and Schum is that Vanzetti was innocent; the judgment concerning Sacco is somewhat different, but in any case – according to the authors – a proof of guilt “beyond reasonable doubt” was not attained, thus also Sacco would have been acquitted.

As a final comment to the wide analysis carried out by Kadane and Schum on all the elements of evidence concerned with the Sacco and Vanzetti case, we call the attention to the correct reference used by the authors for the inferential procedure followed by them in many situations; in fact, they call such procedure *abduction*, as their aim was essentially *to generate hypotheses* out of empirical evidence. As the authors write (p. 39): «Deduction shows that something is *necessarily* true, induction shows that something is *probably* true, but abduction shows that something is *possibly* or *plausibly* true. Most human reasoning tasks, such as those encountered by the historian and criminal investigator, involve mixtures of these three forms of reasoning» (see also Peirce, 1901; Rizzi, 2004; Garbolino, 2014, pp. 47-49).

5. NAKED STATISTICAL EVIDENCE AND SMALL PROBABILITIES

The conceptual mistakes connected with the *naked statistics* (to be defined soon) are usually observed when such statistics are given very small probabilities. As pointed out in the comments about the Dreyfus case, many people are inclined – usually in an unconscious way – to assess as non random an event which happens very rarely under normal conditions of randomness. This topic is substantially resumed from the volume of Frosini (2002, pp. 65-79 and pp. 125-129), excepting the case of Sally Clark, commented at the end of this section.

A kind of apparent statistical terminology, which is foreign to statistical literature, but was introduced in the latest fifty years in the juridical literature of United Kingdom and United States, in nonetheless suitable in characterizing a certain kind of *statistical evidence* presented in a trial: the terms used are *naked statistical evidence*, and also *naked statistics*; in these cases the statistics presented in trials are mostly of the *base-rate* kind, i.e. their meaning does not ensue from the specific case of the trial, but from some related group or population. Most (but not all) comments about naked statistics are in the negative, meaning that the naked statistical evidence is devoid of any relevance in the assessment of the case at hand; however, there are exceptions. Actually, it will be seen that the widespread debate in the juridical literature on the theme of naked statistical evidence reveals the substantial difficulty to deal with naked statistics in trials; in any case – and this will be our viewpoint – it is not clear, outside fancy situations, if and when some base-rate statistics, presented in a trial, acquire meaning and relevance for the case at hand. The first two cases, commented on in the sequel, have been suggested by two celebrated scholars, and have aroused many comments in the juridical literature.

5.1 THE PARADOX OF THE GATECRASHER

The first case where a naked statistic does not appear decisive in resolving the case, although satisfying the criterion "more probable than not" usually required for the civil cases, was contrived by L.J. Cohen (1977, p. 75), as one of the many interesting examples contained in his famous volume "The probable and the provable". Let us consider a case in which 1,000 people were admitted to a rodeo, and that only 499 regularly paid for admission (perhaps because of a hole in the fence?). Moreover, let us assume that no tickets were issued, and there can be no testimony as to whether given individuals paid for the admission or climbed over the fence. Therefore we can assume a probability 0.501 that a given individual A did not pay; on this basis, the rodeo organizers would be entitled – by applying the criterion "more probable than not" – to ask A for the admission ticket. However, in the comments by Cohen, «if the organizers were really entitled to judgment against A, they would presumably be equally entitled to judgment against each person in the same situation as A», not knowing whether he has paid the admission ticket. Cohen's conclusion is as follows: «The absurd injustice of this suffices to show that there is something wrong somewhere. But where?».

Before making any comments on this example, we may say at once that all the scholars who have commented on this case – to begin with Cohen himself – agree that no ticket collection can be made on the people composing the rodeo audience. The general criterion, applied in this case, is that the probability of guilt (not having paid the ticket) does not relate to any particular individual; in the words of Lea Brilmayer (1986, p. 675) «the judge would probably not even allow the case to go to the jury. The explanation seems to lie in the law's unwillingness to base a verdict upon naked statistical evidence. The problem is that the evidence in question does not deal with each defendant's guilt individually». Nonetheless one must loyally acknowledge that the above example, contrived by Cohen, seems to satisfy the civil law criterion of the preponderance of evidence. For many other comments by juridical and statistical scholars on this case we refer to Frosini (2002, pp. 66-72) and Garbolino (2014, pp, 353-358).

5.2 THE BLU BUSES

The second example, devised by L. H. Tribe (1971, pp. 1340-1341), is certainly more interesting and realistic than the former (and actually its starting point was a real case). An individual XY was negligently run down by a blue bus; plaintiff was giving proof that defendant Z operates four-fifths of all the buses in the town (and specifically on the street where the accident occurred), and on such basis he asks for a compensation from the defendant. In this case the probability that the accident was caused by a Z bus – in the absence of other information – can be reasonably approximated by 4/5 = 0.80; such probability turns out to be decidedly greater than 0.50, namely the reference threshold in the civil trials (when the criterion “more probable than not” is employed).

The juridical literature is substantially in agreement on affirming that no compensation is due to the plaintiff from the defendant Z; the main reasons are strictly analogue to those already expounded for the case of the *gatecrasher*: the given information about the proportion of the Z blue buses is deemed as *naked*, or *base-rate*, or *background statistic*. On this point Tribe (1971, p. 1349) is even more extremist, as he writes: «the plaintiff does not discharge that burden [i.e. by showing a preponderance of the evidence in his case] by showing simply that four-fifths, or indeed ninety-nine percent, of all blue buses belong to the defendant. For, unless there is a satisfactory explanation for the plaintiff’s singular failure to do more than present this sort of general statistical evidence, we might well rationally arrive, once the trial is over, at *a subjective probability of less than .5*». Beyond that, Tribe shares the general juridical policy of rewarding «any incentive for plaintiffs to do more than establish the background statistics». Further considerations from juridical scholars are cited by Frosini (2002, pp. 73-79).

Some juridical scholars make some reservations with respect to the above viewpoint (although positively evaluated by most authors) (see e.g. Shaviro, 1989, p. 531; Allen, 1991, p. 1098). Actually, such authors evaluate as sufficient, or relevant anyhow, the naked statistics about the four-fifths of the buses; moreover, we could add – to the above naked statistic – other pieces of information (e.g. on the direction of the bus), which could restrict the *reference population* relevant for the case (Frosini, 2002, pp. 78-79). What does it mean? We could add other testimonies concerning other aspects of the accident, which allow to further circumscribe the relevant qualifications of the event. At what point are we able to acknowledge the bound, from the denial of using a background statistic as a genuine element of proof, to its acceptance as a valid proof in the inferential procedure? The answer provided by most scholars, both in the juridical and in the statistical domain, is well summarized by Fienberg and Schervish (1986, p. 783): «The decision to convict is not based solely on the probability that a reasonable person would adjudge guilt, but also upon the quality of the evidence on which that probability judgment is based». The *quality of the evidence*! It is only too obvious. But it is not of much assistance.

5.3 THE CHEDZEY CASE

A very interesting example is provided by the Australian case *R v. Chedzey* of 1987 (Robertson and Vignaux, 1995, pp. 82-85; Frosini, 2002, pp. 125-129). Chedzey was accused of having made a bomb-hoax call to Perth police station; the only evidence against Chedzey was that the call was traced to his home by means of the telephone company tracing equipment. Chedzey consistently denied having made the call (although he did change his account of his movements on the evening concerned). An unusual control was made about the correct running of the tracing equipment; evidence was given by an expert on 12,700 calls from known numbers: only five of the records were subject to error. The expert concluded that the tracing equipment was "99.96% accurate".

The jury of the first trial interpreted this percentage as a probability of guilt beyond reasonable doubt. Chedzey was convicted and appealed. The Western Australian Court of Criminal Appeal quashed the conviction, and the defendant was acquitted. Such decision by the Court of Appeal was attained *only* on the basis that Chedzey's telephone *could have* been spotted by a malfunction of the tracing equipment. Attaining a conclusion on this case is certainly a difficult and subtle problem; anyway Frosini (2002) has presented some probability calculations which confirm the above posterior probability of guilt. Therefore we could agree with Robertson and Vignaux that, if an objective probability of guilt – as high as 0.9996 – is *not* deemed sufficient for a sentence of conviction, then not even two or three independent and concordant testimonies could be judged sufficient, as they could generally be evaluated less than a hundred per cent trustworthy. Strangely enough, it is perhaps just the existence of an *objective* probability less than 1 that prevents a decision of conviction; it seems that a *subjective probability* very near to certainty would have led to a sentence of conviction.

5.4 THE CASE OF SALLY CLARK

The case of Sally Clark is among the most interesting ones, because it allows to compare conclusions induced from certain naked statistics, with opposite conclusions induced by a Bayesian reasoning (also itself based on naked statistics). Following the condemnation of Sally Clark in the first trial the English newspapers entitled «One is tragic, two is murder».

The first child of Sally Clark died unexpectedly at the age of about three months, when his mother was the only other person in the house. The death was registered as a case of SIDS (*Sudden Infant Death Syndrome*). The second child of Sally Clark died the following year in similar circumstances (Aitken and Taroni, 2004, p. 211; Garbolino, 2014, p. 388). She was arrested, and charged with murdering both her children; at trial in 1999 she was sentenced to 26 years of imprisonment.

The main argument of the accusation was the evaluation, by a professor of paediatrics, of the probability of natural sudden infant death in the first months of life; this probability was evaluated, on the basis of recent demographic statistics, as 1/8500. Then the probability of two such deaths in the same family was evaluated by squaring this value (adopting the assumption, actually inappropriate, of independence of the two events). The probability of both deaths was then estimated as $(1/8500)^2 = 0.000000013$, a very small probability which brought the judge in the first trial to evaluate the natural death of *both* children as a practically impossible event. However, as we know, from a very small probability we cannot infer that the associated event was not produced by the inherent random process.

The judge in the second appeal trail (in January 2003), which ended with Sally Clark' acquittal, admitted new medical evidence, and accepted an inferential reasoning of the Bayesian kind (Dawid, 2002, pp. 71-90; Aitken and Taroni, 2004, pp. 211-213; Garbolino, 2014, pp. 389-392). The probability of murdering a child in his first year of life was estimated (on data UK) as about 1/92000; again assuming the independence of the two events (although inappropriate, as underlined above), the prior probability of guilt is $(1/92000)^2$, while the probability of innocence is $P(\overline{G}) = (1/8500)^2$. Now we can resume Bayes's formula, in the *odds ratio* form:

$$\frac{P(G \mid E)}{P(\overline{G} \mid E)} = \frac{P(G)}{P(\overline{G})} \times \frac{P(E \mid G)}{P(E \mid \overline{G})} \cdot$$

Taking into account that $P(E \mid G) = P(E \mid \overline{G}) = 1$ (both hypotheses, of guilt and of innocence, wholly explain both deaths), the above ratio reduces to $P(G)/P(\overline{G})$ $\approx 0{,}0085$ implying that $P(\overline{G} \mid E) \approx 0{,}99$, namely *almost certainly* Sally Clark was innocent!

For a more correct approach, i.e. taking account of the dependence structure of both deaths, see Hill (2004) and Hand (2014, Chapter 7).

6. SAMPLING VARIABILITY AND HYPOTHESIS TESTING

We have already hinted, at the end of Section 2, at the strict conditions and at the difficulties often encountered for a correct application of the Bayesian paradigm; nonetheless it remains the *reference* rational paradigm, provided a whole justification be ensured. Another kind of statistical inference is however often encountered in trials and in the juridical and statistical literature, where the Bayesian approach is quite absent (although theoretically admissible). This other kind of inference, relevant in many trials, both civil and criminal, is mostly concerned with (a) the discrimination in employment, (b) the differential mortality of a group of workers – with respect to a suitable reference population – depending on special features of the production process. Such applications, and many others of similar kind, are characterized by (A) a statistical assessment summarized from a sample of individuals, (B) the existence of law rules to be respected, or of a reference population with which to compare the available sample.

The inferential approach hinted at in this section is the one most applied, in cases like the ones sketched above, in the latest fifty years; it derives from a *mix* of two inferential approaches, actually with distinctive features, proposed by Ronald A. Fisher from one side, and from Jerzy Neyman and Egon Pearson from the other side. Among the many textbooks which expound these approaches we may refer to the volumes of Freedman, Pisani and Purves, 2007, Chapters 26-29, and of Frosini, 2009, Chapters 11 and 12.

By a strict *test of significance* (or *significance test*) Fisher has suggested a general formalization of a rather common *inductive* reasoning, which was already used by many other researchers; even the first published statistical test, applied in 1710 by John Arbuthnot (Frosini, 1993), was substantially of this kind. This inductive argument tries *to mimic* a deductive argument, known from the Aristotelian logic as *modus tollens* (a special kind of syllogism); the structure of this argument contains two premises, of type “If *a* then *b*” and “Not *b*”, from which the conclusion “Not *a*” is thus derived (see e.g. Barker, 1965, p. 95).

The approximate translation of this deductive argument for the case of random phenomena, where $P(a \rightarrow b) = p < 1$ (i.e. *a* implies *b* only in a proportion *p* of cases), requires at the outset to spot a subset *A* of the sample space (which is the set of all possible events) which is deemed *to conform* to a hypothesis H_0 (usually called *null hypothesis)* which provides a complete configuration of the given random experiment. Such a subset *A*, called *acceptance region* of the hypothesis H_0, typically maintains a probability rather large under H_0; by calling P_0 the probability function under H_0, it is usually prescribed that $P_0(A) \geq 0.90$ (although this reference can vary according to the specific problem at hand). The complementary set of *A* with respect to the whole sample space, namely $\bar{A}$, is called the *rejection region* of *A*. Actually, when the observed sample $\boldsymbol{x}$ (of given size *n*) is judged to conform to H_0, which happens when $\boldsymbol{x} \in A$ (as formally established), the hypothesis H_0 is accepted; on the contrary, when the sample $\boldsymbol{x}$ is included in the critical region *C*, and could be preferably obtained under some hypothesis $H_1 \neq H_0$, the hypothesis H_0 is formally rejected. This does not mean that H_0 is certainly false, but only that the occurrence of the observed sample could more easily be explained with a different hypothesis about the probability distribution relating to the sample space (or to a statistic defined on the sample space, e.g. the sample average).

When applying a test of significance one must be well aware of the errors which may be done, namely (a) to reject H_0 when it is true (error of the first kind), and (b) to accept H_0 when it is false (error of the second kind). When H_0 (null hypothesis) and H_1 (alternative hypothesis) are *simple*, namely are able to exactly determine the probability of any subset of the sample space under each hypothesis, the above error probabilities are usually called $\alpha = P_0(C)$ (probability of rejecting H_0 under the validity of H_0) and $\beta = P_1(A)$ (probability of accepting H_0 under an alternative hypothesis H_1). The probability α is also called the *significance level* of the test. When H_0 and/or H_1 are *composite*, the probabilities of kind α and β can be determined as functions of the simple hypotheses contained in H_0 or H_1. One must be very careful about fixing such error probabilities, taking into account that α and β are inversely related, given the planned experiment and the sample size: if we want to lower α, we must be willing to bear an increase in β, and vice versa. In most cases a certain balance must be assured between the two error probabilities; e.g. if α is very small (near zero), and β is large (near one), we must be aware that that we could accept H_0 also in most cases when some alternative hypothesis holds.

For the above (a) case it must be reminded that in the USA legislation, Title VII of the Civil Rights Act of 1964, a number of precise rules are established against discrimination in employment based on race, sex, religion, national origin and age (40 and older). Since then many law-courts have coped (in thousands of trials) with problems which sometimes are difficult to solve, and generally require the consultation of experts in statistics. A similar protection for the equality treatment of individuals was ensured in Italy by the Law 10 Apr. 1991 n. 125.

6.1 THE CASE OF HAZELWOOD SCHOOL DISTRICT

To give an idea of such problems, let us briefly examine the suit discussed in *Hazelwood School District v. United States* (Meier, Sacks and Zabell, 1986, pp. 1-48). In this case the judge was called «to assess whether a difference between the proportion of black teachers employed by the Hazelwood School District and the proportion of black teachers in the relevant labour market was substantial enough to indicate discrimination». Recognizing such *substantial difference* is able to imply an effective *discrimination* towards the black teachers. While resuming the cited paper as concerns the appropriate relevant labor market, and also as concerns the *80% Rule* quoted in the sub-title of the same paper, let us dwell on the more methodological topics, concerned with the *tests of statistical significance*, whose application was discussed in the trial.

Just to discuss a numerical example, very near to the one employed in the trial at hand, let us admit that the proportion of black teachers in the reference population is 6%, and that the black teachers hired by the defendant in the two years 1972-73 and 1973-74 have been 400 on the whole, of which 16 black. It is quite evident that, if the black teachers hired in the District would have been 400×0.06 = 24, the proportion of black teachers in the sample and in the reference population would have been the same, hence no discrimination at all. It is also acceptable, however, that the case would be judged of *no discrimination* if *small differences* would be observed with respect to the value of 24 black teachers. The problem is: *how much small*? In other words, when can we say that an observed frequency smaller than 24 can be a clear indication of a discriminatory policy? Hence the problem is to single out a value $x_0 < 24$ such that, when the number of black teachers hired in the given period is ≤ 24, we have a clear indication of discriminatory policy.

The above threshold x_0 can be determined on examination of the *sample variability* (or *sample dispersion*) of the frequency of black teachers, in a random sample of n teachers (in this case $n = 400$) drawn from a large population in which the proportion of black teachers is p (in this case $p = 0.06$). As in this case the value of n is sufficiently large, we may employ the *normal approximation of the binomial random variable*; practically, the above value x_0 can be spotted as the greatest integer value which is smaller than two times the *standard deviation* σ of the corresponding binomial distribution; in this case $\sigma = \sqrt{0.06(1-0.06/400}$ = 0.01187. As 2×0.01187 = 0.02374, and (0.06 – 0.02374) = 0.03626, a sample proportion ≤ 0,03626 can be taken as an indication of a discriminatory policy against black teachers. We can express this same result in terms of the greatest integer value smaller than 0.03626×400 = 14.504, to be rounded down to 14. Under the hypothesis of random sampling from the reference population, a sampling observation of a frequency ≤ 14 has a probability of about 0.025. With this choice of the critical region C, the observed number 16 of black teachers is included in the acceptance interval, thus the null hypothesis of no discrimination is accepted.

A more precise determination can be attained by using exact probabilities, instead of approximate probabilities obtained by the normal approximation of the binomial distribution (by applying a statistical software, or simply by employing some statistical functions provided by a spreadsheet). Actually, as $p = 0.06$ is rather far from 0.5, and we need a probability computed on the left tail, the approximation is not so good as expected: calling by X the binomial variable with $n = 400$ and $p = 0.06$, the exact value of the probability $P(X \leq 14)$ turns out to be 0.01715 (instead of the approximation 0.025). Moreover, with a critical interval C of type $X \leq 14$, and consequently an acceptance interval A of type $X \geq 15$, we can compute values of the error probabilities $\beta(p)$ like the following: $\beta = 0.6363$ for $p = 0.04$; $\beta = 0.2252$ for $p = 0.03$; such error probabilities appear quite large (notice that the observed number 16 of black teachers hired corresponds to a relative frequency of 0.04).

Enlarging the critical interval C by including the values 15 and 16 (notice that in this case the observed frequency $x = 16$ is included in the critical interval, although just in its boundary), the significance level raises to $\alpha = 0.051$ – a standard value in this type of applications – while the error probability b(p) becomes: $\beta = 0.434$ for $p = 0.04$, $\beta = 0.098$ for $p = 0.03$, quite an acceptable balance between the probabilities of the two kinds of error.

6.2 A CASE ABOUT DIFFERENTIAL MORTALITY OF WORKERS IN A CHEMICAL PLANT

A rather rare kind of cancer was examined in a large cohort of 1652 workers in a chemical plant in Italy; such a cohort was followed for 35 years. According to the null hypothesis H_0 that the deaths inside the cohort, expected from this cancer, follow the same behavior as for the general population (of the region where the plant is located), such deaths have been modelled by a Poisson distribution Y with mean $\lambda = 7.5$ (cf Frosini, 2009, pp. 112-114). Taking into account that the cohort-sample could show either a decrease or an increase of mortality with respect to the reference population (both behaviors have been actually observed for some specific causes of death), a reasonable *acceptance region A* could include all the values (absolute frequencies) 3, 4, ... , 13; as $P_0(Y \leq 2) = 0.0203$, $P_0(Y \leq 13) = 0.9784$, hence $P_0(3 \leq Y \leq 13) = 0.9784 - 0.0203 = 0.9581$, with an error probability of the first kind $\alpha = 1 - 0.9851 = 0.0419$.

On the contrary, if there are defensible reasons which could exclude – in the given working conditions – a decrease in mortality with respect to the general population, the acceptance region could include all the values 0, 1, ... , 12, as $P_0(Y \leq 12) = 0.9573$, with an error probability $\alpha = 1 - 0.9573 = 0.0427$. In this second case it is interesting – and standard – to check for the error probability β when the mortality rate *doubles*, namely by using a Poisson distribution Y_1 with parameter l = 15; in this case the error probability of the second kind is $\beta = P_1(Y_1 \leq 12) = 0.2676$.

REFERENCES

Aitken, C.G.G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists* (First edition). Wiley, Chichester.

Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists* (Second edition). Wiley, Chichester.

Allen, R.J. (1991). On the significance of batting averages and strikeout totals: A clarification of the "naked statistical evidence". *Tulane Law Review* 65: 1093-1110.

Barker, S.F. (1965). *The Elements of Logic*. McGraw-Hill, New York,

Barnard, G.A. (1982). Causation. In *Wiley Encyclopedia of Statistical Sciences*, Vol. 1, Wiley, New York, 387-389.

Beccaria, C. (1764). *Dei delitti e delle pene*. Coltellini, Livorno. Translated by D. Young, *On Crimes and Punishments*, Hackett Publ. Co., Indianapolis, 1986.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.

Blackstone, W. (1769, 1899). *Commentaries on the Laws of England*, (four books). Banks and Co., Albany.

Brilmayer, L. (1986). Second-order evidence, and Bayesian logic. In *Boston University Law Review*. 66: 673-691.

Champod, C., Taroni, F. and Margot, P. (1999). The Dreyfus case – an early debate on experts' conclusions (an early and controversial case on questioned document examination). In *International Journal of Forensic Document Examiners.* 5: 446-459.

Cohen, L.J. (1977). *The Probable and the Provable*. Oxford University Press, Oxford.

Coleman, R. and Walls, H. (1974). The evaluation of scientific evidence. *Criminal Law Review*: 276-287.

Cook, R., Evett, I.W., Jackson, G., Jones, P.J. and Lambert, J.A. (1998). A hierarchy of propositions: deciding which level to address in casework. In *Science & Justice*. 38(4): 231-239.

Copi, I.M. (1964). *Introduzione alla logica*. Il Mulino, Bologna.

Dawid, A.P. (2002). Bayes's theorem and the weighing of evidence by juries. In R. Swinburne (ed.), *Proceedings of the British Academy*, Vol. 113, Oxford University Press, Oxford, 71-90.

Fienberg, S.E. and Schervish, M.J. (1986). The relevance of Bayesian inference for the presentation of statistical evidence and for legal decisionmaking. In *Boston University Law Review*. 66: 771-798.

Finkelstein, M.O. and Levin, B. (2004). Epidemiologic evidence in the silicone breast implant cases. In *Chance.* 17 (2): 39-43.

Fisher, R.A. (1937). On a point raised by M.S. Bartlett on fiducial probability. In *Annals of Eugenics.* 7: 370-375.

Fisher, R.A. (1959). *Statistical Methods and Scientific Inference* (Second edition). Oliver and Boyd, Edinburgh.

Freedman, D., Pisani, R. and Purves, R. (2007). *Statistics* (Fourth edition). Norton, New york.

Frosini, B.V. (1993). Caso e inferenza. In *Vita e Pensiero.* 76: 22-46.

Frosini, B.V. (1994). Regressione diretta e inversa: problemi di taratura e applicazioni legali. In *Statistica Applicata.* 6: 195-214.

Frosini, B.V. (1999). Conditioning, information and frequentist properties. In *Statistica Applicata.* 11: 165-184.

Frosini, B.V. (2002). *Le prove statistiche nel processo civile e nel processo penale*. Giuffré, Milano.

Frosini, B.V. (2006). Causality and causal models: A conceptual perspective. *International Statistical Review.* 74: 305-334.

Frosini, B.V. (2009). *Metodi Statistici*. Carocci, Roma.

Garbolino, P. (2014). *Probabilità e logica della prova*. Giuffré, Milano.

Hand, D.J. (2014). *The improbability principle*. Scientific American, Farrar Strauss and Giroux.

Hill, R. (2004). Coincidence or beyond coincidence? In *Paediatric and Perinatal Epidemiology*. 18: 320-326.

Kadane, J.B. (2008). *Statistics in the Law* (Online book). Oxford University Press, Cary, NC, USA.

Kadane, J.B. and Schum, D.A. (1996). *A Probabilistic Analysis of the Sacco and Vanzetti Evidence*. Wiley, New York.

Keynes, J.M. (1921). *A Treatise on Probability*. Macmillan, London.

Meier, P., Sacks, J. and Zabell, S.L. (1986) What happened in Hazelwood. Statistics, Employment, Discrimination, and the 80% Rule. In M.H. DeGroot, S.E. Fienberg and J.B. Kadane (eds.), *Statistics and the Law*, Wiley, New York, 1-48.

Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.

Peirce, C. (1901, 1955) Abduction and induction. In J. Buchler (ed.), *Philosophical Writings of Peirce*, Dover, New York, 150-156.

Rizzi, A. (2004). Abduzione e inferenza statistica induttiva. In *Statistica e Società*. 15-25.

Robertson, B. and Vignaux, G.A. (1995). *Interpreting Evidence. Evaluating Forensic Science in the Courtroom*. Wiley, Chichester.

Shaviro, D. (1989). Statistical-probability evidence and the appearance of justice. In *Harvard Law Review*. 103: 530-554.

Spirtes, P, Glaymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search* (Second edition). The MIT Press, Cambridge, MA.

Stella, F. (2000). *Leggi scientifiche e spiegazione causale nel diritto penale* (Seconda edizione), Giuffré, Milano.

Stella, F. (2001). *Giustizia e Modernità*. Giuffré, Milano.

Taruffo, M. (1992). *La prova dei fatti giuridici. Nozioni generali*. Giuffré, Milano.

Tribe, L.H. (1971). Trial by mathematics: Precision and ritual in the legal process. In *Harvard Law Review*. 84: 1329-1393.

Voltaire (1772). *Essai sur les probabilités en fait de justice*. Voltaire Foundation, Oxford.

Wigmore, J. (1913). The problem of proof. In *Illinois Law Review*. 8: 77-103.

Wigmore, J. (1937). *The Science of Proof: As given by Logic, Psychology, and General Experience, and Illustrated in Judicial Trials* (Third edition). Little and Brown, Boston.

Zeisel, H. and Kaye, D. (1997). *Prove It with Figures. Empirical Methods in Law and Litigation*. Springer, New York.