REVSTAT – Statistical Journal
Volume 3, Number 1, June 2005, 77–98

A LIKELIHOOD APROACH TO DIAGNOSTIC TESTS IN CLINICAL MEDICINE

Authors: Basilio de Bragança Pereira
– Faculdade de Medicina, Universidade Federal do Rio de Janeiro – UFRJ, Brazil (basilio@nesc.ufrj.br)

Carlos Alberto de Bragança Pereira
– Instituto de Matemática e Estatística, Universidade de São Paulo – USP, Brazil (cpereira@ime.usp.br)

Received: November 2004 Revised: February 2005 Accepted: March 2005

Abstract:

- This paper presents a pure likelihood approach to statistical inference and its relation to diagnostic tests in clinical medicine. The standard antagonistic frequentist approaches of Fisher and Neyman–Pearson–Wald, and the Bayesian perspective are discussed. We advocate that in medicine, usually, the likelihood is the only source of information available. Also, it is shown that there is no difference of interpretation of the concept of likelihood in statistics and in clinical diagnostic tests. This contradicts what is usually stated.

Key-Words:

- *credibility; diagnosability; likelihood; plausibility; sensitivity; significance; specificity.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

The likelihood function plays a central role in parametric statistical inference since it contains all the information in the observed data. It is used in both frequentist antagonistic approaches, Fisherian and Neyman–Pearson–Wald (NPW), but in neither methodology it is the main tool. On the other hand, the only experimental source of information to the ones fond of Bayesian methodology is exactly the likelihood function. Hence, for Bayesians using uniform (proper or improper) densities, the only relevant tool for analysis is the likelihood function.

Most Bayesians and frequentists may disagree with the views presented here (see [2] and [21]) since they are close to the ideas described by Ronald Fisher in his last and controversial book, [12]. We believe that it is closer to the Bayesian perspective than to the standard frequentist approaches. A recent revival of interest in the likelihood approach is in action; see [23], [25], [30], [39], [40], and [41] for instance. The site http://www.stat.unipd.it/LIKASY/biblio.html presents a comprehensive list of references.

A brief history is presented in Section 2. The likelihood perspective is presented and discussed in Sections 3 and 4. In Section 5, diagnostic tests results are compared to the contingencies of statistical results of the different views. In Section 6 we present an index for the diagnostic ability of a clinical test. Section 7 contains the likelihood view of a diagnostic test with a graphical illustration. Finally, in Section 8 we present a real example to illustrate the ideas discussed in previous sections.

2. STATISTICAL TESTS — A BRIEF HISTORY

Some of the material of this section can be found in [42]. The idea of significance tests was proposed by Fisher, who introduced the *p-value* as an index of agreement between the data and the null hypothesis: the greater the *p-value*, the greater the evidence in favor of the null hypothesis. A *p-value* of 5% is commonly used as a standard threshold for deciding against $\mathbf{H}$ ($p < 0.05$) or in favor of $\mathbf{H}$ ($p > 0.05$). However, we strongly support the idea that the choice of the threshold should depend on the problem currently faced by the scientist, the sample size, and the amount and type of information being collected. This is in fact the idea of significance tests as prescribed by [7] and [22].

The subjective judgment of an observed *p-value* to decide against or in favor of $\mathbf{H}$ led Neyman and Pearson ([29]) and Wald ([43] and [44]) to proposing the theory of Test of Hypotheses. This theory, contrarily to Fisher's significance

tests, was designed to replace the subjective judgment of the strength of evidence in favor of the null hypothesis, provided by a *p-value* judgment, with an objective decision-theoretical approach. By fixing, in advance, the Type I error rate, α, and minimizing the Type II error rate, β, the number of wrong decisions, made over many different repetitions of the experiment, would be limited. This may generate some controversy since only in very few medical applications repetitions are possible.

Originally, the NPW theory required the specification of single point null, **H**, and alternative, **A**, hypotheses. By fixing Type I and Type II error rates, the sample size could be determined. Sample size determination is an area in which NPW theory has been appropriately used in medicine (and also in industrial quality control), although a confuse mixture of the Fisher and NPW approaches to hypothesis testing may be found in the medical literature. Statements such as "*p-values* smaller than 5% were considered statistically significant", without specifying the alternative hypothesis and the Type II error rate, are common. It is usual to have a table with *p-values* and intervals obtained by summing and subtracting twice the sample standard error from the sample mean.

Jeffreys [20] attacked the problem under a Bayesian perspective. Let x denote the observations, π and $\pi(x)$ the prior and posterior probabilities for **H**. Alternatively the corresponding probabilities for **A** are $(1-\pi)$ and $[1-\pi(x)]$. Defining the prior and posterior odds by

$$\rho = \pi(1-\pi)^{-1} \qquad \text{and} \qquad \rho(x) = \pi(x)\big[1-\pi(x)\big]^{-1} \,,$$

Jeffreys proposed to look at the posterior odds, also called Bayes Factor, as the index of evidence in favor of **H**.

In the case of single point hypotheses, let $f_H(x)$ and $f_A(x)$ be the two alternative densities being compared. The likelihood ratio is $R(x) = f_H(x)/f_A(x)$. Hence, one can easily prove that $\rho(x) = \rho R(x)$. Also, for $\pi = 1/2$ we would have $\rho(x) = R(x)$. Hence, for the case of single point hypotheses, judging **H** based on the likelihood ratio corresponds to a Bayesian judgment with very particular prior choices. On the other hand, recall that the likelihood ratio is the function used by the Neyman–Pearson theorem of optimal decision. Also, note that one can use $R(x)$ to order the sample space, [8], [28] and [38]. If the computation of the *p-value* were performed under this ordering, the alternative hypothesis would be taking into consideration. As one may see, the three methods have their conclusions based on the likelihood ratio, $R(x)$.

Real controversial problems emerge with the consideration of composite hypotheses. Many of the practical problems in medicine involve sharp null hypotheses. That is, the dimension of the subspace where **H** is defined is smaller than the dimension of the subspace where **A** is defined. Let us consider the well-known standard problem of the test for independence in a 2×2 contingency table.

Let C_1 and C_2 be two populational characteristics and $x = (x_{11}, x_{12}, x_{21}, x_{22})$ be the vector of the sample frequencies for the respective combination of the levels of categories C_1 and C_2. The parameter space associate with this experiment is the simplex

$$\Theta \;=\; \left\{ \left(\theta_{11}, \theta_{11}, \theta_{11}, \theta_{11}\right) \;\middle|\; \theta_{ij} > 0,\; \sum_{i,j=1}^{2} \theta_{ij} = 1 \right\}$$

and the null hypothesis is defined by the subset

$$\Theta_H \;=\; \left\{ \Big[pq,\, p(1-q),\, (1-p)q,\, (1-p)\,(1-q)\Big] \;\middle|\; 0<p,q<1 \right\}.$$

Note that the two hypotheses are composite and that $p = \theta_{11} + \theta_{12}$ and $q = \theta_{11} + \theta_{21}$. The sets that define the null and the alternative hypotheses, Θ_H and $\Theta_A = \Theta - \Theta_H$, have different dimensions, i.e., $\dim(\Theta) = 3 > \dim(\Theta_H) = 2$.

Letting $f(x|\theta)$ denote the likelihood function, frequentists will define $S_H(x)$ and $S_A(x)$ as the suprema of $f(x|\theta)$ under **H** and **A**, respectively. The profile likelihood ratio is defined as $PR(x) = S_H(x)/S_A(x)$. Bayesians, on the other hand, in addition to the prior probabilities for **H** and **A**, namely $\pi(\mathbf{H})$ and $[1-\pi(\mathbf{H})] = \pi(\mathbf{A})$, define densities over Θ_H and Θ_A. Considering these densities as weighing systems — systems indexes that defines a preference order on the points of the space — and taking the weighted likelihood averages, $M_H(x)$ and $M_A(x)$, under Θ_H and Θ_A respectively, they define the Bayes Factor $BF(x) = \rho MR(x)$ where $\rho = \pi(1-\pi)^{-1}$ is the prior odds and $MR(x) = M_H(x)/M_A(x)$ is the weighted likelihood ratio. To compute the weighted averages one must uses the weighing systems considered for Θ_H and Θ_A. [18] uses this approach for a Bayesian version of the McNemar test for also comparing two composite hypotheses of different dimensions in a 2×2 contingency table. NPW (Jeffreys's Bayesian) approach for hypothesis testing consists of the evaluation of $PR(x)$ $[BF(x)]$. The Fisher approach for testing independence is a modification based on a conditional distribution of the data in the basic cells of the table given the marginal cells. It does not seem appropriate to consider that the marginal cells are known before the data were observed. For example, consider an overall frequency of 20 for the contingency table. The number of possible tables (the sample space size) in this case is 1771. If a marginal total is 5, for instance, the number of possible tables with this marginal is 6. That is, for considering a given marginal we reduce our sample space from 1771 possibilities to only 6 possibilities and the *p-value* could be much greater than it should be. For a detailed discussion on this matter see [17] and [35].

The fourth approach to hypothesis tests is that of (pure) likelihood, which is described in the next section.

3. LIKELIHOOD APPROACH

The deductive nature of probability versus the inductive nature of statistical inference is clearly reflected in the dual concepts of probability distributions and likelihood ([24] and [11]). Given a probability model and the corresponding

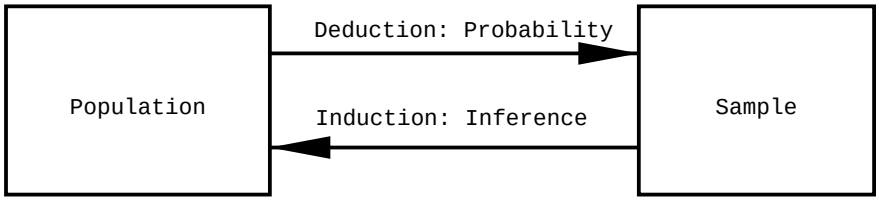

Figure 1: Probability and Statistics Harmonization.

parameter values, we may calculate the probabilities associated to all possible observations, x. Suppose that before observing the performance of the next 10 patients that will be submitted to a drug known to have efficacy of 60%, a doctor writes his probability model function for "the number of recovered patients, X" as:

$$f(x|.6) \;=\; \Pr\{X = x|\theta = 0.6\} \;=\; \binom{10}{x} (.6)^x \, (.4)^{10-x} \; .$$

The probability of having 7 recovered patients is $f(7|.6) = .215$. Note that $f(x|\theta)$ is a function of two variables: x, the observation, and θ, the parameter. For fixed θ, f is a probability function of x and for fixed x, f is a function of θ called likelihood function associated to the observed value, x. Suppose that we observe 7 success and 3 failures for this sample of 10 patients. The likelihood function is

$$L\big(\theta|X = 7\big) \;=\; \Pr\{X = 7|\theta\} \;=\; \binom{10}{7} \theta^7 (1-\theta)^3 \;=\; (120)\, \theta^7 (1-\theta)^3 \; .$$

In order to illustrate the differences between probability and likelihood functions, in Figure 2 we present the corresponding probability functions for $\theta = .6$ and for $\theta = .3$, while in Figure 3 we present the likelihood functions for $x = 7$ and for $x = 2$.

Note that the two probability functions in Figure 2 are discrete. Since the parameter space Θ is the interval $[0; 1]$, the likelihood functions depicted in Figure 3 are continuous. A statistical model has two arguments, the possible observations and the possible values of the parameter. The likelihood function is not a probability density function. However, dividing it by its integral over the

parameter space (whenever this integral exists), the resulting normalized likelihood is a probability density over Θ, and corresponds to the Bayesian posterior density under a uniform prior. Areas under this curve define probabilities of subsets of the parameter space.

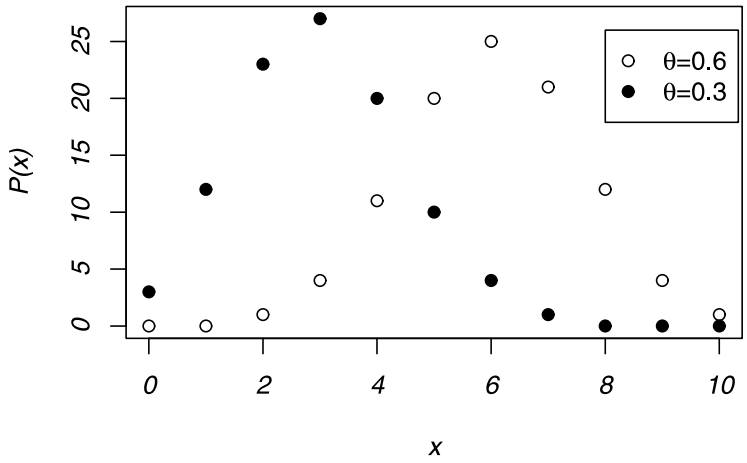

Figure 2: Binomial probability functions for $n = 10$.

The likelihood function, L, induces an ordering of preferences about the possible parameter points. Note that this order is not changed if a proportional function is defined. This means that we can divide L by any constant without modifying the conclusions about parameter point preferences. We can divide L by its integral obtaining the normalized likelihood, the Bayesian way, or divide it by the maximum value of L whenever it exists, obtaining what we call relative likelihood. Comparing two parameter values, we would say that the one with higher (normalized or relative) likelihood is more plausible than the other.

An important feature of the Likelihood approach is that it is independent of stopping rules. That is, it does not violate the likelihood principle, [1], [3] and [5]. For instance, suppose that another doctor in another clinic decided to start his analysis only when he obtain 3 failures, i.e., 3 patients that do not recover. As soon he obtained his 3rd failure, corresponding to the 10^{th} patient, he realizes that he had 10 patients with 7 successes and 3 failures. Although he has the same results as his colleague, the underlying statistical model is completely different but his (normalized) relative likelihood is equal to the one obtained from the previous models. Here the probability model is a negative binomial distribution. That is, the random variable is the number Y of failures to be observed since the number of failures k was fixed in advance. The model here is given by

$$P\{Y = y|\theta\} = \binom{y+k-1}{y} \theta^y (1-\theta)^k \ .$$

For the sample with $k = 3$ and $y = 7$, the likelihood is proportional to the one illustrated in Figure 3. Figure 4 shows the negative binomial probability distributions for $k = 3$, $\theta = .6$ and $\theta = .3$.

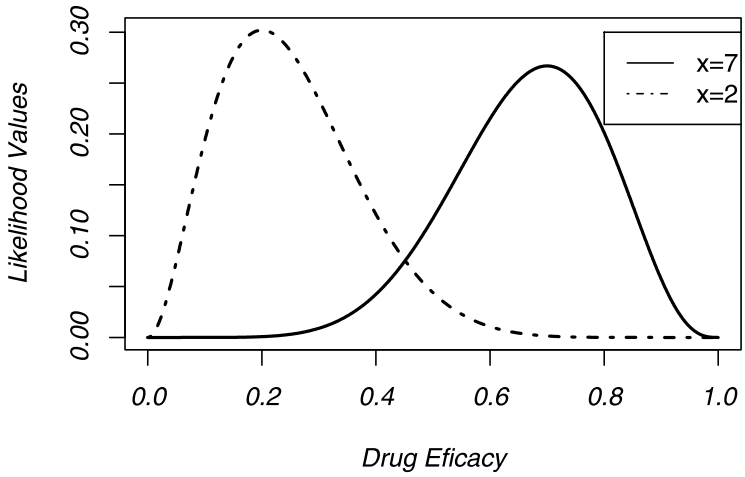

Figure 3: Binomial likelihood functions for $n = 10$.

Note that for both Figures 2 and 4, the probabilistic models, Binomial and Negative Binomial, have their sample space well defined since the stopping rules were defined previously. However there are many cases in medical statistics

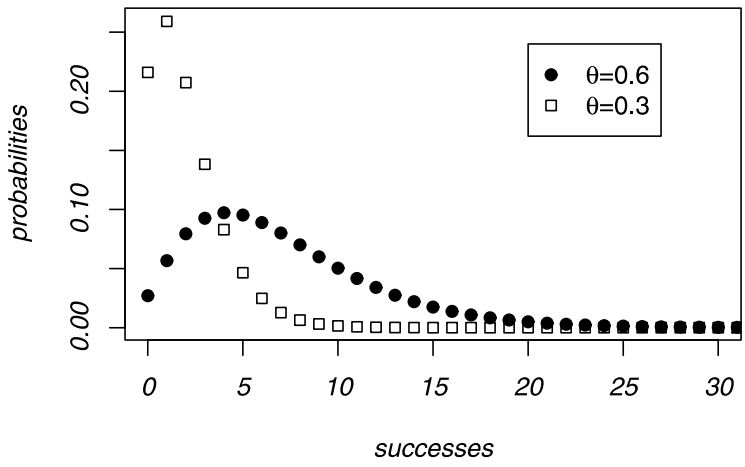

Figure 4: Negative Binomial probability function for $k = 3$.

where the sample space is not well defined. For instance, suppose that a doctor wants to write a paper and decides to look at the data he has collected up to

that moment. In this case, neither the sample size nor the number of success (or failures) was fixed a priori. However, if he had observed 7 recoveries in 10 patients, his likelihood would be proportional to $\theta^3(1-\theta)^7$, which is proportional to both observed Binomial and Negative Binomial likelihoods. Hence, in all 3 cases, the relative (normalized) likelihoods are exactly the same and then the inference would be the same as prescribed by the likelihood principle. We emphasize that the normalized likelihood for the example of 3 failures and 7 successes is a beta density with parameters $a = 4$ and $b = 8$. The relative likelihood is the beta density divided by the density evaluated at its mode, which is the maximum likelihood estimate, $3/10 = .3$.

In Figure 5 we illustrate the relative likelihood for 3 failures and 7 successes, with a solid line intercepting it at points with plausibility equal to 1/3 (relative to the maximum) and a dotted line at points with plausibility equal to .8057.

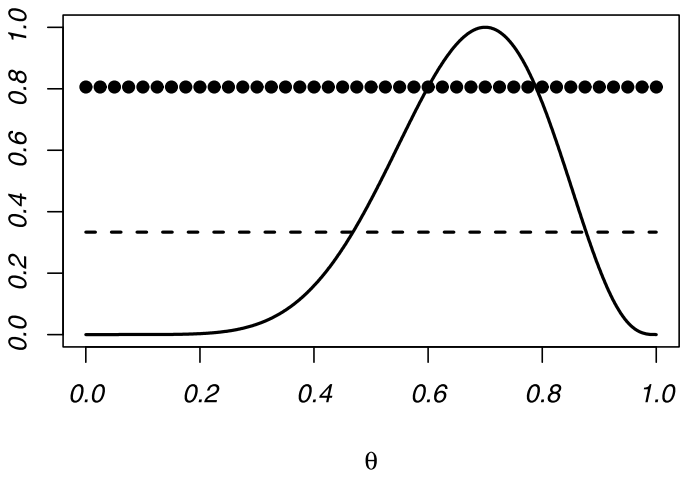

Figure 5: Relative Likelihood, and 1/3 and .8057 Plausible Levels.

Recall that the maximum of the likelihood function is attained at $\theta = .7$. Also, at $\theta = .6$, the suggested drug efficacy, the plausibility is .8057. Note that both $\theta_1 = .4681$ and $\theta_2 = .8770$ have plausibility equal to 1/3. Any parameter point inside (outside) the interval $I(1/3) = [.4681; .8771]$ has palusibility larger (smaller) than 1/3. If one uses the normalized likelihood as the posterior density, the (posterior) probability that the unknown parameter θ lies in $I(1/3)$ is equal to .8859. That is, $I(1/3)$ is a credible interval for θ with credibility 88.59%. This probability (or credibility) is calculated by computing the area under the curve limited by the vertical segments at .4681 and .8771 divided by the total area under the curve.

Consider the other point, $\theta_{00} = .7886$, with the same plausibility as the suggested drug efficacy, $\theta_0 = .6$. These two points have plausibility equal to .8057 and the interval $I(.8057) = [.6000; .7886]$ has credibility 51.65%. Considering now $\theta_0 = .4$, the corresponding parameter point with the same plausibility is $\theta_{00} = .9124$. These points have plausibility equal to .1592 and the interval $I(.1592) = [.4000; .9124]$ has credibility 95.90%.

Observing the low (high) probability of having a parameter value with more plausibility than .6 (.4), we would say that the hypothesis $\mathbf{H}\colon\ \theta = .6$ ($\mathbf{H}\colon\ \theta = .4$) should be not rejected (accepted). We suggest that the credibility of the interval $[\theta_0; \theta_{00}]$ may be interpreted as an index of evidence against the null sharp hypotheses $\mathbf{H}\colon \theta = \theta_0$ or $\mathbf{H}\colon \theta = \theta_{00}$. The probability of the complement of this credibility interval is an index (like a *p-value*) of evidence in favor of $\mathbf{H}$; see [37] and [27] for more on this measure of evidence. For the two cases presented here, the evidence in favor of $\mathbf{H}$ is 48.35% for $\mathbf{H}\colon \theta = .6$ and 4.10% for $\mathbf{H}\colon \theta = .4$.

We end this section by stating a rule to be used by Pure Likelihood followers.

Pure Likelihood Law: *If the relative likelihood function of two points, θ_0 and θ_1, satisfy $RL(\theta_0) >(<) RL(\theta_1)$, we say that θ_0 is more (less) plausible than θ_1. We say they have the same plausibility if equality of the likelihood functions holds. For single point hypotheses $\mathbf{H}\colon\ \theta = \theta_0$ versus $\mathbf{A}\colon\ \theta = \theta_1$ if $RL(\theta_0) <(>) RL(\theta_1)$, we reject (accept) $\mathbf{H}$. The strength of evidence of the data x in favor of $\mathbf{H}$ against $\mathbf{A}$ is measured by the likelihood ratio, $LR(\theta_0; \theta_1) = RL(\theta_0)/RL(\theta_1)$.*

For the example above, we have $LR(.6; .7) = .8057$ and $LR(.6; .4) = 5.0625$.

4. LADDER OF UNCERTAINTY AND CONTROVERSIES

Tests of hypotheses are decision procedures based on judgments and one can only judge something in relation to the alternatives. The concept of statistical evidence of some data, x, in favor or against some hypothesis must be relative in nature. We should not talk about evidence for or against $\mathbf{H}$ without mentioning the alternative $\mathbf{A}$. Pereira & Wechsler ([38]) show how to build a *p-value* that takes the two antagonistic hypotheses into consideration.

An implication of the pure law of likelihood is that: "uncertainty about x given θ" and "statistical evidence in x about θ" have different mathematical forms. The statistical model is based on a trinity of mathematical elements: the sample space $\mathbf{X}$, the parameter space Θ and a function $f(\cdot|\cdot)$ of two arguments

$(x, \theta) \in \mathbf{X} \times \Theta$. For every fixed $\theta \in \Theta$, $f(\cdot|\theta)$ is a probability (density) function on $\mathbf{X}$ and for every fixed $x \in \mathbf{X}$, $f(x|\cdot) = L(\cdot|x)$ is the likelihood function. The following sets characterize the statistical model:

i) $\Im = \{f(x|\theta) \mid x \in \mathbf{X},\ \theta \in \Theta\}$ is the overall statistical model,
ii) $\forall\, \theta \in \Theta,\ \Im_\theta = \{f(x|\theta) \mid x \in \mathbf{X}\}$ are the probability models, and
iii) $\forall\, x \in \mathbf{X},\ \Im_x = \{f(x|\theta) = L(\theta|x) \mid \theta \in \Theta\}$ are the likelihood functions.

Uncertainty is measured by probabilities, $\Im_\theta$, and evidence is measured by the likelihood, $\Im_x$. This is a critical insight: the measure of the strength of evidence and the frequency with which such evidence occurs are distinct mathematical quantities, [6]. [39] clearly explains alternative areas of Statistics where these concepts appear. Suppose a patient has a positive result in a diagnostic test, the physician might draw one of the following conclusions:

1. The person probably has the disease,
2. The person should be treated for the disease,
3. The test results are evidence that the person has the disease.

These possible attitudes front the tests results may represent, respectively, answers to different questions:

1′. What should I believe?
2′. What should I do?
3′. How should I interpret this body of observation as evidence about having the disease against not having the disease?

These questions involve distinct aspects of statistical methods, namely: frequentist or Bayesian inference, decision theory and, lastly, interpretation of statistical data as containing evidence, the significance test of hypothesis.

The correctness of the answer for the first question requires, the additional information of the behavior of the test in other (exchangeable) patients or the personal opinion about the probability of the disease before the test (prior probability). For the second question, in addition to the requirements of the first, one also needs knowledge about the costs or utilities of the decisions to be made. Only the third one does not require additional information other than data. [4] considers these arguments to suggest that the role of the likelihood in Statistics is equivalent to the role of diagnostic tests used in Medicine.

Royall ([39]) also discusses a possible paradox in the use of the pure likelihood approach through the following example:

> *"We pick a card at random out of a deck of 52 cards and observe an ace of clubs. Then consider two alternative hypotheses* **H***: it is a deck with 52 aces of clubs or* **A***: it is a standard deck of cards. The likelihood ratio of* **H** *against* **A** *is 52. Some find this disturbing.*

What this result shows is that this strong evidence is not strong enough to overcome the prior improbability of **H**. *A Martian faced with this problem would find* **H** *most appealing."*

Clearly, the Martian's ignorance about card decks does not permit him to use the tools used by both Bayesian and frequentist statisticians. These people may achieve stronger results than pure likelihood statisticians do, but at the price of more assumptions in their applications. [30] tentatively tries some reconciliation among the different approaches using the Fisherian idea of ladder of uncertainty. It remains to be proved that his ideas will succeed in Statistics by means of practical applications.

5. DIAGNOSTIC TESTS AND STATISTICAL VEREDICTS

The inadequacy in relying only and strongly on *p-values* in medicine has been widely emphasized in recent years. Worst yet, is the lack of understanding of what *p-values* are. In this section we present quantities that may be of more interest to medicine than the *p-values* are. For more discussion on the subject we refer to [9] and [32]. We use the following notation: $D^+ = Disease$, $D^- = No\ Disease$, $T^+ = Positive\ test\ result$ and $T^- = Negative\ test\ result$. For the populational parameters let $N(++)$ be the frequency of units in category (D^+T^+), $N(+-)$ the units in category (D^+T^-), $N(-+)$ the units in category (D^-T^+), and $N(--)$ the units in category (D^-T^-). $N(+\bullet)$ denote the number of units with the disease, $N(-\bullet)$ the number of units without the disease, $N(\bullet+)$ the number of units with positive test result, and $N(\bullet-)$ the number of units with negative test result.

The following quantities are of great interest for physicians evaluating patients. For a randomly selected unit from the population we define the following quantities:

a. ***Sensitivity*** is the conditional probability of responding positively to the test given that the patient has the disease, i.e., $S = \Pr\{T^+|D^+\} = N(++)/N(+\bullet)$.

b. ***Specificity*** is the conditional probability of responding negatively to the test given the absence of the disease, i.e., $E = \Pr\{T^-|D^-\} = N(--)/N(-\bullet)$.

c. ***Prevalence*** is the probability that the patient has the disease, i.e., $\pi = \Pr\{D^+\} = N(+\bullet)/N$. Alternatively, $(1-\pi) = \Pr\{D^-\}$ is the probability that the patient does not have the disease.

d. ***Test Positivity*** and ***Test Negativity*** are the probabilities of positive and negative test results, i.e., $\tau = \Pr\{T^+\} = N(\bullet+)/N$ and $(1-\tau) = N(\bullet-)/N$.

e. ***Diagnostic Parameters*** are the posterior probabilities of the states of a patient given the response to the clinical test:

PPV: Positive Predictive Value is the conditional probability of presence of disease given positive test response: $\pi(T^+) = \Pr\{D^+|T^+\} = N(++)/N(\bullet+)$ and

NPV: Negative Predictive Value is the conditional probability of absence of disease given negative test response: $[1-\pi(T^-)] = \Pr\{D^-|T^-\} = N(--)/N(\bullet-)$.

The quantities of higher interest in clinical practice are the predictive values, **PPV** and **NPV**. Using Bayes formula, we obtain important relations between the predictive values and the other terms of the model, namely

$$\mathbf{PPV} = \pi(T^+) = \frac{\pi S}{\pi S + (1-\pi)(1-E)} = \left\{1 + \left[\left(\frac{\pi}{1-\pi}\right)\left(\frac{S}{1-E}\right)\right]^{-1}\right\}^{-1}$$

and

$$\mathbf{NPV} = \big[1-\pi(T^+)\big] = \frac{(1-\pi)E}{(1-\pi)E + \pi(1-S)} = \left\{1 + \left[\left(\frac{1-\pi}{\pi}\right)\left(\frac{E}{1-S}\right)\right]^{-1}\right\}^{-1}.$$

Denoting the likelihood ratio for positive results by $LR(+) = S/(1-E)$, the likelihood ratio for negative results by $LR(-) = (1-S)/E$ and the prevalence odds by $\rho = \pi/(1-\pi)$ we have:

$$\mathbf{PPV} = \Big\{1 + \big[\rho\, LR(+)\big]^{-1}\Big\}^{-1} \qquad \text{and} \qquad \mathbf{PPV} = \Big\{1 + \rho\, LR(-)\Big\}^{-1}.$$

Considering ρ as the prior odds in favor of the disease and $1/\rho$ as the prior odds against it, the posterior odds in favor and against the disease become $\rho(+) = \mathbf{PPV} \div (1 - \mathbf{PPV})$ and $\rho(-) = \mathbf{NPV} \div (1 - \mathbf{NPV})$. Relating all these quantities we obtain the following interesting formulas:

$$\rho(+) = \rho\, LR(+) = \Big[(\textit{prior odds}) \times (\textit{likelihood ratio for} +)\Big],$$

$$\rho(-) = \big[\rho\, LR(-)\big]^{-1} = \Big[(\textit{prior odds}) \times (\textit{likelihood ratio for} -)\Big]^{-1},$$

$$\rho(+) = \frac{\textit{prevalence}}{1 - \textit{prevalence}} \times \frac{\textit{sensitivity}}{1 - \textit{specificity}},$$

$$\rho(-) = \frac{1 - \textit{prevalence}}{\textit{prevalence}} \times \frac{\textit{specificity}}{1 - \textit{sensitivity}},$$

$$\mathbf{PPV} = \rho(+)\big[1 + \rho(+)\big]^{-1} \qquad \text{and} \qquad \mathbf{NPV} = \rho(-)\big[1 + \rho(-)\big]^{-1}.$$

The important question for a physician working with diagnostic tests is to decide what to do when the result is positive (or negative). In fact, measures of sensitivity and specificity, when available, would be of great help to him since they may yield other valuable quantities, see [9] and [32]. Note that if there is a big change from prior to posterior odds the test will be considered of great value. In the next section we discuss a way of defining diagnostic power of clinical evaluations. This index is of great value to state an order of preference in a set of clinical procedures

6. DIAGNOSABILITY

In this section we discuss the diagnostic power of a medical test. To evaluate the diagnostic ability of a test T, we should focus on the change from ρ to $\rho(+)$ and from $(1-\rho)$ to $[1-\rho(-)]$. This is related with the weight of evidence provided by T^+ (T^-) in favor of D^+ (D^-) and denoted by $\omega^+ = \omega(D^+;T^+)$ $[\omega^- = \omega(D^-;T^-)]$. Good ([14]) showed that the function ω, to follow reasonable requirements, ought to be an increasing function of the odds ratio — the ratio of posterior to prior odds — or, equivalently, an increasing function of the likelihood ratio. That is, ω^+ and ω^- must be increasing functions of $\rho(x)\,\rho^{-1} = LR(+) = S(1-E)^{-1}$ and $\rho\rho(-) = [LR(-)]^{-1} = E(1-S)^{-1}$, respectively.

The usual cross-product ratio (in the context of contingency tables), useful in measuring association, is simply

$$R = \frac{LR(+)}{LR(-)} = \frac{SE}{(1-S)\,(1-E)} = \frac{(\mathbf{PPV})\,(\mathbf{NPV})}{(1-\mathbf{PPV})\,(1-\mathbf{NPV})}\,.$$

As we will see in the sequel, the larger R is, the better the test for detecting disease D, i.e., the better its **diagnosability**.

As a consequence of the requirement of additivity of information, [13] proves that the weights of evidence, ω^+ and ω^-, are the natural logarithms of $LR(+)$ and $LR(-)$. [13] also points out that the expected value of the weight of evidence is more meaningful than the likelihood ratio. Hence, the measure of the ability of a medical test, T, to discriminate in favor of D^+ (D^-), given that the true state of nature is D^+ (D^-) is the conditional expectation of ω^+ (ω^-) given S, E and the state of the patient, D^+ or D^-. We denote these conditional expectations by ϵ^+ and ϵ^- Finally, the diagnosability of T is by definition $\Delta = \epsilon^+ + \epsilon^-$. Let us explicitely introduce these formulas:

Weight of Evidence

a) In favor of D^+,
$\omega(D^+;T^+) = \omega^+ = \ln[LR(+)]$ and $\omega(D^+;T^-) = -\omega^- = \ln[LR(-)]$.

b) In favor of D^-,
$\omega(D^-;T^+) = -\omega^+ = -\ln[LR(+)]$ and $\omega(D^-;T^-) = \omega^- = -\ln[LR(-)]$.

Average Weight of Evidence

c) In favor of $D^+ = \epsilon^+ = S\omega^+ - (1-S)\,\omega^-$.
d) In favor of $D^- = \epsilon^- = E\omega^- - (1-E)\,\omega^+$.

Diagnosability Index

e) $\Delta = (S+E-1)\ln R$.

We would like to call the attention to the fact that all these indices depend strongly on the values of many parameters that are in fact not completely known. Usually the prevalence, the sensitivity and the specificity have to be estimated with sample data. [36] introduced Bayesian techniques for such purposes. They also consider the case where a set of clinical tests are observed in the same subject and show how a combination of them improves the diagnosability of the medical procedure. In a predictivist context, [33] and [34] show that if we look at a particular patient, the computation of her/his posterior probability of having the disease simplifies significantly the diagnostic calculus.

In order to decide if a new (possibly expensive) test must be considered in lieu of some other test, one must collect, observe, and analyze a new sample. Usually the size of a sample of patients, known to have the disease, is the number of patients under treatment at the clinic and the test is applied to all possible patients. A control group of units without the disease is also selected and tested after all ethical procedures have been fulfilled. Based on the two samples, S and E are estimated. Estimates of $LR(+)$, $LR(-)$, and R are then obtained.

The association measure R plays the most important role in the determination of the diagnostic power of a test T. In the next section, we present plots that will help to use only the likelihood ratios to define situations where a test is of interest for the clinician. We end this section with an analogy linking different schools of statistics and the clinician's interest in the properties of diagnostic tests:

- A Fisherian clinician would be mainly concerned with the false positive rate, cases where the treatment is harmful for the patients (e.g. prescribing a surgery when it is not necessary).
- A Neyman–Pearson–Wald clinician would be concerned with the false positive and false negative rates.
- A Bayesian clinician would be concerned with the positive and negative predictive values.
- A likelihood clinician view would be concerned with positive and negative likelihood ratios, which will be discussed further in the next section.

7. LIKELIHOOD ANALYSIS OF A DIAGNOSTIC TEST AND LIKELIHOOD RATIO PLOTS

For a given diagnostic test we have defined, respectively, the likelihood ratios of positive and negative test results as $LR(+)$ and $LR(-)$. We also show how to measure the diagnosability of a test, which is based on the change of the pre-test to the post-test odds ratios. According to [19], the directions and magnitudes of the pre to post changes using likelihood ratio values as a rough guide are as follows:

1. LR's larger than 10 or smaller than 0.1 generate conclusive changes.
2. LR's in the interval $(5; 10]$ or $[0.1; 0.2)$ generate moderate shifts.
3. LR's in the interval $(2; 5]$ or $[0.2; 0.5)$ generate small (important sometimes) shifts.
4. LR's in the interval $(1; 2]$ or $[0.5; 1)$ generate small (rarely important) shifts.

Jaeschke et al. ([19]) also presented a modification of a monogram suggested by [10]. The monogram is as an old calculus rule where in the left side we have values for the prevalence, in the middle the likelihood ratio and in the right side the **PPV** values. By drawing a straight line from the prevalence value throughout the likelihood ratio value and ending the line at the right side, the value obtained at this end is just the **PPV** observed.

Biggerstaff ([4]) presented another interesting graphical method for comparing diagnostic tests. A large value of $LR(+)$ indicates that the test has good sensitivity and a small value of $LR(-)$ means that the test has good specificity. If both situations hold we have that R is large and the test has a high diagnostic ability or equivalently high diagnosability. In many situations, due to costs or the health conditions of a patient, one must choose among a set of diagnostic tests a subset that will be performed. In this way ordering the tests by their diagnosability becomes important. To order a set of diagnostic tests according to their diagnostic ability one should have in mind the risks, the costs and the likelihood ratio values. Note that ordering the tests according to $LR(+)$, high to low values, is equivalent to ordering them based on the values of their **PPV**'s. On the other hand, ordering the tests according to $LR(-)$, low to high values, is equivalent to ordering them based on the values of their **NPV**'s.

Similarly to the ROC (Receiver Operator Characteristic Curve), in Figure 6 we plot, for a diagnostic test T_1, the point $A = (1-E_1; S_1)$. That is, the false positive rate, $X = (1-E_1)$, against the true-positive rate, $Y = S_1$. Additionally we draw two lines through this point; (i) a solid line-segment through $(0; 0)$ and A, ending in the horizontal line $(X; 1)$ and (ii) a dotted line-segment through $(1; 1)$

and A, ending in the vertical line $(0; Y)$. It is not difficult to prove that the slopes of the solid and the dotted lines are, respectively, $LR_1(+)$ and $LR_1(-)$, the likelihood ratios for the test T_1. The diagonal line delimitates the area where

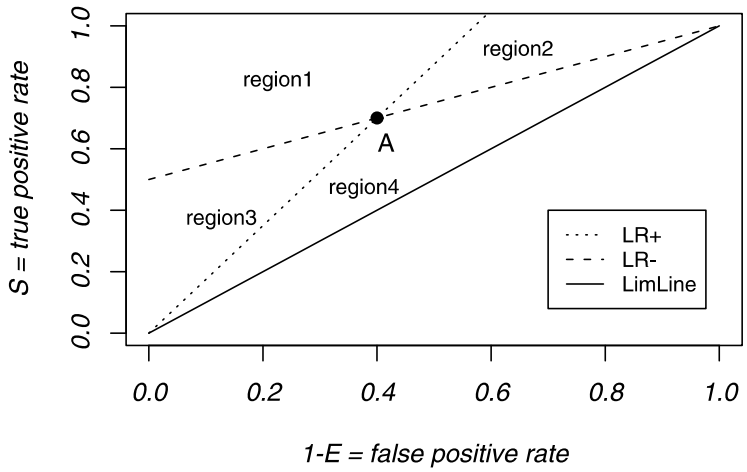

Figure 6: Regions of Preference: $A = (1-E; S) = (.4, .7)$.

a test is useful. Also, it is easy to show that, for a test, if the point A is below the diagonal line the test is useless. We end this section with the following example:

Example: Consider a diagnostic test T_1 where $S_1 = .7$ and $E_1 = .6$. For this case we have $A = (.4; .7)$, the solid line is $Y = 1.75\,X$ and the dashed line is $Y = (1+X)/2$. We have then $LR_1(+) = 1.75$ and $LR_1(-) = .5$. If a new test T_2 is considered we have four possible locations for the point $A_2 = (1-E_2; S_2)$:

i. $A_2 \in$ *Region 1*, which implies that T_2 is better than T_1 overall, since

$$LR_2(+) > LR_1(+) \qquad \text{and} \qquad LR_2(-) < LR_1(-)\ ;$$

ii. $A_2 \in$ *Region 2*, which implies that T_2 is better (worse) than T_1 for confirming absence (presence) of the disease, since

$$LR_2(-) < LR_1(-) \quad \big[\text{and} \quad LR_2(+) < LR_1(+)\big]\ ;$$

iii. $A_2 \in$ *Region 3*, which implies that T_2 is better (worse) than T_1 for confirming presence (absence) of the disease, since

$$LR_2(+) > LR_1(+) \quad \big[\text{and} \quad LR_2(-) > LR_1(-)\big]\ ;$$

iv. $A_2 \in$ *Region 4*, which implies that T_2 is worse than T_1 overall, since

$$LR_2(+) < LR_1(+) \qquad \text{and} \qquad LR_2(-) > LR_1(-)\ .$$

8. FINAL REMARKS

We would like to end this report with an optimistic view for the future of pure likelihood approach of Statistics. Let us recall that the work of a statistician lies in a trinity of problems; design of experiments, estimation, and hypotheses testing. We want to show how the likelihood approach works well for the three problems.

In the domain of design of experiments, consider the problem of determination of number of patients to be tested in order to estimate S, the sensitivity of a clinical test. The maximum of the likelihood is the prescribed estimate. However, we would also need to fix an interval around this estimate in order to guarantee the control of our sampling error. For this purpose we use the normalized likelihood and would like to have the smallest interval with relative plausibility (or credibility) around 95%. Since the binomial distribution is an adequate model, the normalized likelihood follows a beta distribution with parameter $(X+1; Y+1)$ where X (Y) is the number of true positive (false negative) results in a sample of size n, to be determined. Recall that the mean and the variance of this beta distribution are, respectively, $m = (X+1)/(n+2)$ and $v = m(1-m)/(n+3)$. Note that $v \leq [4(n+3)]^{-1}$ since $0 \leq m \leq 1$. Hence, the worst case ($m = 1 - m = .5$) is a symmetric beta distribution; i.e. $X = Y$. In this case the mean and the mode (the maximum likelihood estimate) are equal to .5. Adding and subtracting twice the standard deviation to m, we obtain a fair plausible interval (as usually we do when considering normal distributions). Let us represent this interval by $[I_1; I_2]$, where

$$I_1 = .5 - (n+3)^{.5} \qquad \text{and} \qquad I_2 = .5 + (n+3)^{.5} \ .$$

Let us now fix the length of the interval of highest plausibility as $I_2 - I_1 = 2(n+3)^{.5} = .1$. For this value we obtain $n = 397$. In order to satisfy the restriction $X = Y$, we would take $n = 298$ as the sample size. Note that, for $n = 398$ the normalized likelihood would be a beta density with parameter $(200; 200)$; that is, $X = Y = 199$. Considering this case, the interval $[.45; .55]$ would have credibility 95.49% and length .1. Now suppose that we perform the experiment and observe that $X = 53 = 398 - Y$. The parameter of the corresponding beta density is $(54; 346)$. This is not a symmetric density around its maximum, $53/398$, and the smallest interval with a fixed credibility has equal plausibility in its limits, I_1 and I_2. For this non-symmetric case we would have the interval $[.1033; .1703]$ with credibility 95.01% and length .067. To obtain this interval we recall that a beta distribution with parameters larger than 1 is uni-modal. Hence, to every parameter point there is a corresponding one with the same plausibility. Considering a pair, say I_1 and I_2, with the same plausibility in such a way that the interval $[I_1; I_2]$ has posterior probability equal to the fixed credibility, say 95%, we ob-

tain our interval. For bi-dimensional parameter spaces, obtaining a set of 95% of credibility, corresponds to obtaining a level curve where its interior has posterior probability of 95%.

In the above discussion we have shown how a likelihood approach will solve the sample size determination and both point and interval estimation problems. We now discuss the testing problem. We use here real data presented in [36]. Two samples of size 150 were taken respectively from a subpopulation of patients having a disease D and from a healthy control group. A new clinical test was applied to these samples. For the patients, we observed $x = 20 = 150 - y$ true positive cases and for the control sample we obtained $x' = 3 = 150 - y'$ false positive cases. We have here two likelihood functions, one for the sample of patients and another for the control sample. We want to compare this new test, T_1, with a standard one, T_0, know to have sensitivity $S_0 = .15$ and specificity $E_0 = .91$. To replace T_1 for T_0, we would like to have $S_1 > S_0$ and $E_1 > E_0$. To make a decision about the use of the new test we first identify the set of parameter points with plausibility higher than $S_1 = .15$ in the sample of patients and then compute its credibility. For the control sample we identify the set of parameter points with plausibility higher than $E_1 = .91$ and then compute its credibility. Note that the normalized likelihood for S_1 (E_1) obtained in the patient (control) sample is a beta density with parameters 21 and 131 (148 and 4). Before we describe the computations let us recall that $LR_0(+) = 5/3 = 1.67$ and $LR_0(-) = 85/91 = .93$. On the other hand, the maximum likelihood estimates for the likelihood ratios of the new test are $LR_1(+) = 20/3 = 6.67$ and $LR_1(-) = 130/147 = .88$. The odds ratio for the standard test is $R_0 = 1.78$ and the maximum likelihood estimate for the odds ratio of the new test is $R_1 = 7.54$. The Good's weights of evidence are $\Delta_0 = .0347$ and $\Delta_1 = .2289$. These values already provide evidence that the new test is superior. However, to quantify this superiority we proceed as follows:

1. For the sample of patients, the set of possible values of S_1 with plausibility higher than $S_0 = .15$ is the open interval $(.1178; .1500)$; this set has credibility 43.92%. Hence, the evidence in favor of **H**: $S_1 = .15$ is 56.09%. With these figures we cannot reject the hypothesis that the two tests have equivalent sensitivities;

2. For the control sample, the set of possible values of E_1 with plausibility higher than $E_0 = .91$ is the interval $(.910; .999)$; this interval has credibility 99.95%. Hence, the evidence in favor of **H**: $E_1 = .91$ is .05%. The conclusion here is that the new test is far more specific than the old one; and

3. Finally, constructing a plot like in Figure 2 with $A = (1 - E_0; S_0) = (.09; .15)$, one would show that the estimated value of $A_1 = (1 - E_1; S_1)$, which is $\left(\frac{1}{50}; \frac{2}{15}\right)$, belongs to *Region 1*, supporting the superiority of the new test, T_1.

We believe to have covered the three problems without using other elements than the likelihood function. We did not have to bring into consideration sample points that could be observed but were not, as in the usual frequentist techniques of unbiased estimation, confidence interval construction or standard significance and hypothesis testing. The most important feature of the methods described in this paper is that the likelihood principle is never violated.

We finalize the paper by presenting *p-values* for the hypothesis **H**: $S_1 = .15$ and **H**: $E_1 = .9$. In the first case we have 64.78% and in the second case .02% as exact *p-values*. Had we used the chi-square test, we would have 56.76% and .42%. Recall that our evidence values, based only on the likelihood function (defined on the parameter space, not on possible sample points), for these two hypotheses are 56.09% and .05%.

ACKNOWLEDGMENTS

This paper was written while the first author was visiting Prof. C.R. Rao at the Center of Multivariate Analysis, Department of Statistics, Pennsylvania State University in 2003. He was on leave from Federal University of Rio de Janeiro (UFRJ) under the financial support of a grant of CAPES, a Brazilian agency for research and graduate studies. Prof. J.M. Singer kindly read and discussed the controversial aspects of the paper. We thank him for his patience and interest.

REFERENCES

[1] Barnard, G.A. (1949). Statistical Inference (with discussion), *J. Royal Statistical Society*, **11**(2), 115–149.

[2] Basu, D. (1988). *Statistical Information and Likelihood*, in: "A Collection of Critical Essays by Dr. D. Basu" (J.K. Ghosh, Ed.), *Lecture Notes in Statistics* Vol. 45, Berlin, Springer.

[3] Berger, J.O. and Wolpert, R. (1988). *The likelihood Principle: A Review and Generalizations*, "IMS Monograph Series" (2nd edition), Hayward, California.

[4] Biggerrstaff, B.J. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios, *Statistics in Medicine*, **19**, 649–663.

[5] Birnbaum, A. (1962). On the foundations of statistical inference (with discussion), *J. Amer. Statist. Assoc.*, **32**, 414–435.

[6] Blume, J.D. (2002). Likelihood methods for measuring statistical evidence, *Statistics in Medicine*, **21**, 2563–2599.

[7] Cox, D.R. (1977). The role of significance tests, *Scand. J. Statist.*, **4**, 49–70.

[8] Dempster, A.P. (1997). The direct use of likelihood for significance testing, *Statistics and Computing*, **7**, 247–252.

[9] Diamond, G.A. and Forrester, J.S. (1983). Clinical trials and statistical verdicts: probable grounds for appeal, *Annals of Internal Medicine*, **98**, 385–394.

[10] Fagan, T.J. (1975). Nomograms for Bayes theorem, *New England J. of Medicine*, **293**, 257.

[11] Fenders, A.J. (1999). *Statistical Concepts*, in: "Intelingent Data Analysis" (M. Berthold and D. Hand, Eds.), Chapter 2, N. York, Springer.

[12] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, London, Oliver & Boyd.

[13] Good, I.J. (1950). *Probability and The Weighing of Evidence*, London, Griffin.

[14] Good, I.J. (1968). Corroboration, explanation, involving probability, simplicity and a sharpened razor, *Br. J. Phil. Sci.*, **19**, 123–143.

[15] Good, I.J. (1983). *Good Thinking: The Foundations of Probability and its Applications*, Minneapolis: University of Minnesota Press.

[16] Hill, G.; Forbes W.; Kozak J. and MacNeill, I. (2000). Likelihood and clinical trials, *J. of Clinical Epidemiology*, **53**, 223–227.

[17] Irony, T.Z. and Pereira, C.A. de B. (1986). Exact tests for equality of two proportions: Fisher vs. Bayes, *J. of Statistical Computation and Simulation*, **25**, 93–114.

[18] Irony, T.Z.; Pereira, C.A. de B. and Tiwari, R.C. (2000). Analysis of opinion swing: comparison of two correlated proportions, *The American Statistician*, **54**(1), 57–62.

[19] Jaeschke, R.; Guyatt, G. and Sackett, D. (1994). User's guide to the medical literature: III. How to use an article about diagnostic test: B. What are the results and will they help me in caring for my patients?, *J. of The American Medical Association*, **271**(9), 703–707.

[20] Jeffreys, H. (1939). *Theory of Probability*, Oxford, Claredon Press.

[21] Kempthorne, O. and Folks, L. (1971). *Probability, Statistics and Data Analysis*, Ames, The Iowa University Press.

[22] Kempthorne, O. (1976). Of what use are tests of significance and tests of hypothesis, *Commu. Statist. Theory Methods*, **8**(A5), 763–777.

[23] King, G. (1998). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*, Mineapolis, The University of Michigan Press.

[24] Lindsey, J.K. (1995). *Introductory Statistics: A Modeling Approach*, N. York, Claredon.

[25] Lindsey, J.K. (1996). *Parametric Statistical Inference*, Oxford, Oxford University Press.

[26] Lindsey, J.K. (1999). Relationship among sample size, model selection and likelihood regions and scientifically important differences, *The Statistician*, **48**, 4001–4011.

[27] Madruga, M.R.; Pereira C.A. de B. and Stern, J.M. (2003). Bayesian Evidence Test for Precise Hypotheses, *J. of Statistical Planning and Inference*, in press.

[28] Montoya-Delgado, L.E.; Irony, T.Z.; Pereira, C.A. de B. and Whittle, M.R. (2001). An unconditional exact test for the Hardy–Weimberg equilibrium law: sample space ordering using the Bayes factor, *Genetics*, **158**, 875–883.

[29] Neyman, J. and Pearson, E.S. (1936). Sufficient statistics and uniformly most powerful tests of statistical hypotheses, *Stat. Res. Memoirs*, **1**, 133–137.

[30] Pawitan, Y. (2000). *Likelihood: consensus and controversies*, in: "Conference in Applied Statistics in Ireland" (Pawitan's HP).

[31] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford, Oxford University Press.

[32] Pereira, B. de B. and Louzada-Neto, F. (2002). *Statistical inference* (in Portuguese), in: "Epidemiologia" (R.A. Medronho, Ed.), Chapter 19, Rio de Janeiro, Atheneu.

[33] Pereira, C.A. de B. (1990). *Influence diagrams and medical diagnosis*, in: "Influence Diagrams, Belief Networks and Decision Analysis" (R.M. Oliver & J.Q. Smith, Eds.), N. York, Wiley.

[34] Pereira, C.A. de B. and Barlow, R.E. (1990). Medical diagnosis using influence diagrams, *Networks*, **20**, 565–577.

[35] Pereira, C.A. de B. and Lindley, D.V. (1987). Examples questioning the use of partial likelihood, *The Statistician (J.R.S.S. D)*, **36**, 15–20.

[36] Pereira, C.A. de B. and Pericchi, L.R. (1990). Analysis of diagnosability, *J. Royal Statist. Soc. C (Applied Statistics)*, **39**, 189–204.

[37] Pereira, C.A. de B. and Stern, J.M. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses, *Entropy*, **1**, 69–80.

[38] Pereira, C.A. de B. and Wechsler, S. (1993). On the concept of P-value, *Brazilian J. of Probability and Statistics*, **7**, 159–177.

[39] Royall, R.M. (1997). *Statistical Evidence: A Likelihood Paradigm*, N. York, Chapman Hall.

[40] Severini, T.A. (2000). *Likelihood Methods in Statistics*, Oxford, Oxford University Press.

[41] Sprott, D.A. (2001). *Statistical Inference in Science*, N. York, Springer.

[42] Sterne, J.A.C. (2002). Teaching hypothesis test – time for significance change? (With discussion), *Statistics in Medicine*, **21**, 985–994.

[43] Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, *Annals of Probability and Statistics*, **10**, 299–326.

[44] Wald, A. (1950). *Statistical Decision Functions*, N. York, Wiley.