Interactions of Statistics and Probability with Algebra and Analysis

Odile Pons

Abstract. This note presents some historical examples of the link between the main areas of mathematics and the statistical theory. The research in statistics has an impact on algebra and analysis as much as the innovations due to the probability theory, while algebra and analysis improve the statistical methods.

1. Introduction

The interactions between all domains of the mathematics with the theories of probability and statistics can be found at the origin of new results in probability and statistics as well as in algebra or analysis. Tables and charts for the quantiles of free tests were established at the beginning of the 20th century using approximations of integrals by sums and other classical computational methods. This has been a very active domain for several decades in all countries where the statistics was developed, sometimes errors due to the round-off have been corrected by applications of the probability theory. The second step after building new statistical estimators and tests is their comparison [2].

The asymptotic behaviour of the maximum likelihood tests provides here an example where optimisation methods, probability theory and algebra are applied to the theory of the statistical tests.

2. Variance of Random Variables

In a probability space $(\Omega, \mathcal{F}, P)$, let X be a real variable with a density f_θ indexed by a real parameter set Θ and such that for every x, $f_\theta(x)$ belongs to $C_2(\Theta)$. The mean $\mu_\theta = \int_{\mathbb{R}} x f_\theta(x)\,dx$ of X has the derivative

$$\mu'_\theta = \int x\dot{f}_\theta(x)\,dx = \int_{\mathbb{R}} (x-\mu_\theta)\dot{f}_\theta(x)\,dx\,,$$

where $\dot{f}_\theta$ denotes the first derivative of f_θ, and its variance is $\sigma^2_\theta = \int_{\mathbb{R}} (x-\mu_\theta)^2\,dF_\theta(x)$. By the Cauchy–Schwarz inequality, we get the Cramer–Rao bound

$$\{\mu'_\theta\}^2 \le \sigma^2_\theta \int_{\mathbb{R}} f_\theta^{-2}\dot{f}_\theta^2\,dF_\theta\,.$$

With a random vector X, the inequality applies to $\mu_\theta^T a$ for every real a. With a vector parameter the integral in the bound is a matrix and the inequality applies with the trace of the matrix.

A random vector X has a symmetric variance matrix and it is singular if X has linearly dependent components. Let $I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$ be a block decomposition of a non-singular symmetric matrix, with non-singular square sub-matrices I_{11} and I_{22}. Using the notations

$$\begin{aligned} A &= I_{11} - I_{12}I_{22}^{-1}I_{21}\,, \\ B &= I_{21}I_{11}^{-1}\,, \\ C &= I_{22} - I_{21}I_{11}^{-1}I_{12} \end{aligned}$$

and the relationship $AI_{11}^{-1}I_{12} = I_{12}I_{22}^{-1}C$, one can write a block decomposition of the inverse of I in the two following forms

$$I^{-1} = \begin{pmatrix} I_{11}^{-1} + B^T C^{-1} B & -I_{11}^{-1}I_{12}C^{-1} \\ -C^{-1}I_{21}I_{11}^{-1} & C^{-1} \end{pmatrix}$$

and

$$I^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}I_{12}I_{22}^{-1} \\ -C^{-1}I_{21}I_{11}^{-1} & C^{-1} \end{pmatrix}.$$

By the uniqueness of the inverse, it follows that $A^{-1} = I_{11}^{-1} + B^T C^{-1} B$ which is equivalent to

$$\begin{aligned} &I_{11}(I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}I_{11} \\ &\quad = I_{11} + I_{12}(I_{22} - I_{21}I_{11}^{-1}I_{12})^{-1}I_{21}. \end{aligned}$$

3. Log-likelihood Ratio Test

Let $X_1, \ldots, X_n$ be a sample with distribution function F belonging to a class $\mathcal{G}$ of distribution functions. Let $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$ be a parametric subset of $\mathcal{G}$ indexed by a bounded open Θ set of $\mathbb{R}^{d_1}$, $d_1 \ge 1$. Goodness of fit test statistics for the hypothesis $H_0 : F$ belongs to $\mathcal{F}$ include the likelihood ratio test for the parametric class of functions $\mathcal{F}$ and nonparametric tests. The general alternative of the test is $K : F$ belongs to $\mathcal{G} \setminus \mathcal{F}$ where $\mathcal{G}$ can be parametric or not.

Assuming that F has a uniformly continuous density f, the likelihood ratio test for the density of H_0 against an alternative K is defined by maximising the density ratio of the sample under K and H_0

$$T_n = 2\log\frac{\sup_{f:F\in\mathcal{G}}\prod_{i=1,\ldots,n}f(X_i)}{\sup_{f:F\in\mathcal{F}}\prod_{i=1,\ldots,n}f(X_i)}.$$

Let $\mathcal{H} = \{F_\gamma, \gamma\in\Gamma\}$ be a parametric class of distribution functions of $\mathcal{G}$ including $\mathcal{F}$ and indexed by a d_2-dimensional parameter γ belonging to a set Γ including the parameters of Θ. The alternative is then $K : F$ belongs to $\mathcal{H}\setminus\mathcal{F}$ and the statistic is written

$$T_n = 2\sum_{i=1}^{n}\{\log f_{\widehat{\gamma}_n}(X_i) - \log f_{\widehat{\theta}_n}(X_i)\},$$

where $\widehat{\gamma}_n = \arg\max_{\gamma\in\Gamma}\sum_{i=1}^{n}\log f_\gamma(X_i)$ is the estimator of the parameter in Γ and $\widehat{\theta}_n$ is the estimator of the parameter in Θ. Under H_0, there exists a parameter value θ_0 belonging to the interior of Θ such that X has the density $f_0 = f_{\theta_0}$. Under the assumptions that every $F = F_\theta$ of $\mathcal{F}$ has a twice continuously differentiable density f_θ with respect to θ and that the Fisher information matrix $I_\theta = -E_\theta\{f_\theta^{-1}(X)\dot{f}_\theta(X)\}^2$ is finite and non-singular, $E_0\log f_\theta(X)$ is locally concave in a neighbourhood of θ_0 and $\widehat{\theta}_n$ is a consistent estimator of θ_0 under H_0. At θ_0, let

$$U_n = n^{-\frac{1}{2}}\sum_{i=1,\ldots,n}\frac{\dot{f}_{\theta_0}}{f_0}(X_i),$$

$$-I_n = n^{-1}\sum_{i=1,\ldots,n}\left\{\frac{\ddot{f}_{\theta_0}}{f_0}(X_i) - \left(\frac{\dot{f}_{\theta_0}}{f_0}\right)^2(X_i)\right\},$$

be the first two derivatives of the log-likelihood under H_0, $n^{\frac{1}{2}}(\widehat{\theta}_n - \theta_0) = I_n^{-1}U_n + o_p(1)$. Since $\mathcal{H}$ includes $\mathcal{F}$, the true parameter value γ_0 in Γ under H_0 can be written as a vector with components including θ_0 and other parameters with value zero, $\gamma_0 = (\theta_0^T, 0^T)^T$. Let $\widetilde{U}_n$ and $\widetilde{I}_n$ be the first two derivatives of the log-likelihood under distributions of $\mathcal{H}$, their d_1 first components are respectively U_n and I_n.

Assuming that, under the alternative, the Fisher information matrices $\widetilde{I}_\gamma = -E_\gamma\{f_0^{-1}(X)\dot{f}_\gamma(X)\}^2$ are finite and non-singular, $\widetilde{I}_n$ converges in probability to I, uniformly in Θ. The estimator of γ is consistent for every parameter of Γ and $n^{\frac{1}{2}}(\widehat{\gamma}_n - \gamma_0) = \widetilde{I}_n^{-1}\widetilde{U}_n + o_p(1)$, when X has the distribution function F_γ.

Under the condition that $I_n(\theta)$ and $\widetilde{I}_n(\gamma)$ have an inverse for every θ in Θ and γ in Γ, the expansion of the log-likelihood ratio statistic relies on the inversion by blocks of the matrix $\widetilde{I}_n$. Let $d = \dim\Gamma - \dim\Theta$ and let π_Θ be the projection from a set Γ into Θ, such that $\dim\Gamma > \dim\Theta$.

Proposition 1. *Under H_0, $T_n = Y_n^TY_n + o_p(1)$ where Y_n is a d-dimensional vector of independent and centred variables with variance 1 and it converges weakly to a χ_d^2 variable. Let $(K_n)_{n\geq1}$ be a sequence of local alternatives indexed by a sequence of parameters $(\gamma_n)_{n\geq1}$ in sets $(\Gamma_n)_{n\geq1}$ for which there exists θ in Θ and γ_a in K such that $\gamma_a = \lim_{n\to\infty}n^{\frac{1}{2}}(\gamma_n - \gamma)$. Under K_n, the limiting distribution of the statistic T_n is $\chi_d^2 - \gamma_a^T\widetilde{I}(\gamma)\gamma_a$.*

Proof. From the consistency of the estimator under H_0, expanding $\log f_{\widehat{\theta}_n} - \log f_{\theta_0} = \log\{1 + f_{\theta_0}^{-1}(f_{\widehat{\theta}_n} - f_{\theta_0})\}$ as n tends to infinity, we obtain

$$\begin{aligned}\sum_{i=1}^{n}&\{\log f_{\widehat{\theta}_n} - \log f_{\theta_0}\}(X_i)\\ &= n^{\frac{1}{2}}(\widehat{\theta}_n - \theta_0)^TU_n\\ &\quad - \frac{n}{2}(\widehat{\theta}_n - \theta_0)^TI_n(\widehat{\theta}_n - \theta_0) + o_p(1)\\ &= \frac{1}{2}U_n^TI_n^{-1}U_n + o_p(1)\end{aligned}$$

and a similar expansion is written for $\log f_{\widehat{\gamma}_n} - \log f_{\gamma_0} = \log\{1 + f_{\gamma_0}^{-1}(f_{\widehat{\gamma}_n} - f_{\gamma_0})\}$. The test statistic is then written

$$T_n = \widetilde{U}_n^T\widetilde{I}_n^{-1}\widetilde{U}_n - U_n^TI_n^{-1}U_n + o_p(1).$$

Let $\widetilde{U}_n = (U_n^T, \widetilde{U}_{n2}^T)^T$, its variance matrix $\widetilde{I}_n$ is split into blocks according to the components of the parameter inside or outside Θ. It follows that

$$T_n = (I_{n21}I_n^{-1}U_n - \widetilde{U}_{n2})^TA_n^{-1}(I_n^{-1}U_n - \widetilde{U}_{n2}) + o_p(1)$$

where the variance of $I_{n21}I_n^{-1}U_n - \widetilde{U}_{n2}$ is A_n hence T has an asymptotically free distribution and T_n converges weakly to a χ_d^2 variable.

Let $\gamma = (\theta^T, 0^T)^T$, under the sequence of local alternatives $(K_n)_{n\geq1}$ the sequence of parameters of $(K_n)_{n\geq1}$ is such that there exists a limit γ_a for the sequence $(n^{\frac{1}{2}}(\gamma_n - \gamma))_{n\geq1}$. Since $f_\gamma = f_\theta$, an expansion of T_n under K_n is obtained from the expansions of $\log\{f_{\widehat{\gamma}_n}f_\theta^{-1}\} = \log\{f_{\gamma_n}f_\gamma^{-1}\} + \log\{1 + f_{\gamma_n}^{-1}(f_{\widehat{\gamma}_n} - f_{\gamma_n})\}$ and $\log\{f_{\widehat{\theta}_n}f_\theta^{-1}\} = \log\{1 + f_\theta^{-1}(f_{\widehat{\theta}_n} - f_\theta)\}$ as n tends to infinity.

The statistic is now written

$$\begin{aligned} T_n &= n\{(\widehat{\gamma}_n - \gamma_n)^T \widetilde{I}_n(\gamma_n)(\widehat{\gamma}_n - \gamma_n) - \gamma_a^T \widetilde{I}_n(\gamma_n)\gamma_a \\ &\quad - (\widehat{\theta}_n - \theta)^T I_n(\theta)(\widehat{\theta}_n - \theta)\} + o_p(1) \\ &= \{I_{n21}(\gamma_n) I_n^{-1} U_n - \widetilde{U}_{n2}(\gamma_n)\}^T \\ &\quad A_n^{-1}(\gamma_n)\{I_{n21}(\gamma_n) I_n^{-1} U_n - \widetilde{U}_{n2}(\gamma_n)\} \\ &\quad - \gamma_a^T \widetilde{I}_n(\gamma_n)\gamma_a + o_p(1)\,. \end{aligned}$$

The variance of $I_{n21}(\gamma_n)I_n^{-1}U_n - \widetilde{U}_{n2}(\gamma_n)$ under the sequence of local alternatives is $A_n(\gamma_n)$ and γ_n tends to γ, therefore the asymptotic distribution of T_n under the sequence of local alternatives is $\chi_d^2 - \gamma_a^T \widetilde{I}(\gamma)\gamma_a$. □

The critical value of the test of level α is c_α such that $\alpha = P(\chi_d^2 > c_\alpha)$ and its asymptotic power under the sequence of local alternatives $(K_n)_{n\geq 1}$ for which there exist θ in Θ and γ_a in K such that $\gamma_a = \lim_{n\to\infty} n^{\frac{1}{2}}(\gamma_n - \gamma)$ is

$$\beta_{\gamma,\gamma_a}(\alpha) = P\{\chi_d^2 > c_\alpha + \gamma_a^T \widetilde{I}(\gamma)\gamma_a\}\,.$$

Therefore, $\inf_{\gamma,\gamma_A\in\Gamma} \beta_{\gamma,\gamma_a} > \alpha$, so the test is unbiased.

Let us consider a parameter space defined by a constraint $\theta^T x \leq 1$, x in $\mathbb{R}^{d_1}$, which split Θ into disjoint sub-spaces, $\Theta_1 = \{\theta \in \Theta : \theta^T x \leq 1\}$ and $\Theta \setminus \Theta_1$. The estimators of the parameters in Θ_1 and the tests for the null hypothesis $H_0 : \theta \in \Theta_1$ against the alternative $K_1 : \theta \in \Theta \setminus \Theta_1$ are built according to the same arguments with indicators of the parameter sets. In the estimation procedure, the log-likelihood is replaced by the Lagrangian

$$l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i) - \lambda(\theta^T x - 1)\,,$$

where $\lambda = 0$ if θ belongs to Θ_1. Under H_0, the constraint implies $0 \geq (\widehat{\theta}_n - \theta)^T x + (\theta^T x - 1)$ hence $U_n^T I_n^{-1} x \leq 0$ and the estimator of θ has the expansion $n^{\frac{1}{2}}(\widehat{\theta}_n - \theta)^T = U_n^T I_n^{-1} 1_{\{U_n^T I_n^{-1} x \leq 0\}} + o_p(1)$. Under the alternative, the constraint is $U_n^T I_n^{-1} x > 0$ and $n^{\frac{1}{2}}(\widehat{\theta}_n - \theta)^T = U_n^T I_n^{-1} 1_{\{U_n^T I_n^{-1} x < 0\}} + o_p(1)$. Let $Z_n^T = U_n^T I_n^{-\frac{1}{2}}$, the test statistic for H_0 against K_1 is

$$T_n = Z_n^T Z_n 1_{\{Z_n^T I_n^{-\frac{1}{2}} x > 1\}} - Z_n^T Z_n 1_{\{Z_n^T I_n^{-\frac{1}{2}} x \leq 1\}} + o_p(1)$$

and Z_n converges to a normal variable Z in $\mathbb{R}^d$ as n tends to infinity. The limiting distribution under H_0 of the log-likelihood ratio tests with the constraint of Θ_1 is then a difference of truncated χ_d^2 variables

$$T = Z^T Z 1_{\{Z^T I^{-\frac{1}{2}} x > 1\}} - Z^T Z 1_{\{Z^T I^{-\frac{1}{2}} x \leq 1\}}.$$

Other constraints modify the limiting distribution of the LR test statistic. Let $(p_1, \ldots, p_K)$ belong to $[0,1]$ and such that $\sum_{k=1}^K p_k = 1$, and let f_k be densities of the variable X conditionally on K classes, belonging to the same set of densities. Tests about the number of components in a mixture model with a density $g = \sum_{k=0}^K p_k f_k$ is non standard since the information matrix is singular. The expansions of the estimators and of the LR statistic do not apply without more conditions. A condition of separation of the parameters of the densities of the mixture was used by Gosh and Sen [1], Chernoff (1995) proposed a reparametrisation of the likelihood ratio, Pons [4] adopted the same approach.

4. Tests and Large Deviations

In a probability space $(\Omega, \mathcal{F}, P)$, let X be a real variable with distribution function F. The Laplace transform of a variable X is

$$L_X(t) = \int_{\mathbb{R}} e^{tx}\, dF(x)\,.$$

On $(\Omega, \mathcal{F}, P)$, let $(X_i)_{i=1,\ldots,n}$ be a sequence of independent and identically distributed real random variables with mean zero, having a finite Laplace transform L_X, and let $S_n = \sum_{i=1}^n X_i$. By Chernov's large deviations theorem, for every $a > 0$ and for every $n > 0$, $\log P(S_n > a) = \inf_{t>0}\{n \log L_X(t) - at\}$ and it is strictly negative. A d-dimensional variable X has the same property, for every a in $\mathbb{R}^d$.

The score variable $U_n(\theta_0)$ of the LR test for $H_0 : F = F_0$ against the alternative $K : F \neq F_0$ converges weakly to a centred Gaussian variable U_0 with variance $I_0 = I(\theta_0)$, under the hypothesis. For distributions with a scalar parameter, the level of the test with rejection domain $D_n(\alpha) = \{|I_n^{-\frac{1}{2}} U_n(\theta_0)| > c_\alpha\}$ is $\alpha = P_0(|N(0,1)| > c_\alpha) = 2e^{-\frac{1}{2}c_\alpha^2}$, by Chernov's theorem, and the critical value is deduced. If $d > 1$, let $\|\cdot\|_{2,d}$ be the norm of $l_2(\mathbb{R}^d)$, a score test has a rejection domain $D_n(\alpha) = \{\|I_n^{-\frac{1}{2}} U_n(\theta_0)\|_{2,d} > c_\alpha\}$ and it is asymptotically equivalent to the LR test. The χ^2 variable $\|I_0^{-\frac{1}{2}} U_0\|_{2,d}^2$ has the Laplace transform $L(t) = (1-2t)^{-\frac{d}{2}}$, t in $]0, \frac{1}{2}[$, and $\inf_{0<t<.5}\{\log L(t) - c_\alpha^2 t\} = \inf_{0<t<.5}\{-\frac{d}{2}\log(1-2t) - c_\alpha^2 t\} = d \log c + \frac{1}{2}(d - d\log d - c^2) := k_\alpha < 0$. Applying Chernov's theorem to the asymptotic level of the test, it follows that $\alpha = e^{k_\alpha}$. Its asymptotic power against the sequence of alternatives of Proposition 1 is

$$\beta_{\gamma,\gamma_a}(\alpha) = P\{\|I_0^{-\frac{1}{2}} U_0 - \widetilde{I}^{-\frac{1}{2}}(\gamma)\gamma_a\|_{2,d} > c_\alpha\}\,.$$

Consider the score test of $H_0 : F \in \mathcal{F}$ against the alternative $K : F \notin \mathcal{F}$, with a parametric class of distributions $\mathcal{F} = \{F_\theta, \theta \in \Theta\}$. The score process $(U_n(\theta))_{\theta\in\Theta}$ of the LR test converges weakly to a centred Gaussian process with variance $I(\theta)$ and covariance function $E[\{f_{\theta_1}^{-1}\dot{f}_{\theta_1} f_{\theta_2}^{-1}\dot{f}_{\theta_2}\}(X)]$. A score test has a rejection domain

$$D_n(\alpha) = \{\sup_{\theta\in\Theta} \|I_n^{-\frac{1}{2}}(\theta)U_n(\theta)\|_{2,d} > c_\alpha\}$$

with c_α such that $\sup_{\theta\in\Theta} P_\theta\{D_n(\alpha)\}$ converges to α, as n tends to infinity, applying Chernov's theorem. Its asymptotic power is the limit of the sequence $\beta_n(\alpha) = \inf_{F\notin\mathcal{F}} P_F\{D_n(\alpha)\}$ and it is larger than α.

Boundaries of the tail probabilities of empirical processes obtained by large deviations also apply to the asymptotic behaviour of estimators [5].

References

[1] J. Ghosh and P. Sen, On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results, *Proc. Berkeley Conference in Honor of J. Neyman and J. Kiefer, II*, eds. L. M. Lecam and R. A. Olshen, (1985) pp. 789–806.

[2] J. L. Hodges and E. L. Lehman, Comparison of the normal scores and Wilcoxon tests, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics* **t1** (1961) pp. 307–318.

[3] E. L. Lehman, *Elements of Large-Sample Theory* (Springer, New York, 1999).

[4] O. Pons, *Estimation et tests dans les modèles de mélanges de lois et de ruptures* (Hermès-Science Lavoiser, Paris, 2009).

[5] O. Pons, *Inequalities in Analysis and Probability* (World Scientific, Singapore, 2012).

[6] C. Rao, *Linear Statistical Inference and Its Applications* (Wiley, New York, 1967).

Odile Pons

INRA Mathematics, France
odile.pons78@orange.fr

Odile Pons is the director of research at the French National Institute for Agronomical Research. She graduated from the University Paris 11, Orsay and obtained a PhD in 1979 in mathematical statistics and a habilitation to supervise theses in 1987. She published more than 30 papers about new nonparametric estimators and tests for censored variables and processes and about 10 articles on statistics in mixture models, as well as applications to agronomy, toxicity and genetics with biologists. Odile Pons' main books are *Nonparametric Estimation for Renewal and Markov Processes* (2008), *Estimation and tests in Distribution Mixtures and Change-Points Models* (2009), *Functional Estimation for Density, Regression Models and Processes* (2011), *Inequalities in Analysis and Probability* (2012).