ISSN 1645-6726

INSTITUTO NACIONAL DE ESTATÍSTICA
PORTUGAL

REVSTAT
Statistical Journal

Volume 3, No.1
June 2005

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E., 2003- . - 30 cm
Semestral. - Continuação de : Revista de Estatística = ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726

- **PUBLISHER**
  - *Instituto Nacional de Estatística (INE)*
    Av. António José de Almeida, 2
    1000-043 LISBOA
    PORTUGAL
    Tel.: (0351) 21 842 61 00
    Fax: (0351) 21 842 63 64
    Web site: http://www.ine.pt
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística*
- **EDITION**
  - *400 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*

PRICE

[VAT 5% included]

- Single issue .... **€ 12**
- Anual subscription ... **€ 19,20**

INDEX

REVSTAT – Statistical Journal
Volume 3, Number 1, June 2005, 1–18

ROBUSTNESS OF TWO-PHASE REGRESSION TESTS

Authors: Carlos A.R. Diniz
– Departamento de Estatística, Universidade Federal de São Carlos, São Paulo, Brazil (dcad@power.ufscar.br)

Luis Corte Brochi
– Departamento de Estatística, Universidade Federal de São Carlos, São Paulo Brazil (dcad@power.ufscar.br)

Received: February 2003 Revised: July 2004 Accepted: December 2004

Abstract:

- This article studies the robustness of different likelihood ratio tests proposed by Quandt ([1]) and ([2]), (Q-Test), Kim and Siegmund ([3]), (KS-Test), and Kim ([4]), (K-Test), to detect a change in simple linear regression models. These tests are evaluated and compared with respect to their performance taking into account different scenarios, such as, different error distributions, different sample sizes, different locations of the change point and departure from the homoscedasticity. Two different alternatives are considered: i) with a change in the intercept from one model to the other with the same slope and ii) with a change in both the intercept and slope.

  The simulation results reveal that the KS-Test is superior to the Q-Test for both models considered while the K-Test is more powerful than the other two tests for nonhomogeneous models with a known variance.

Key-Words:

- *segmented regression models; likelihood ratio tests; robustness.*

AMS Subject Classification:

- 62J02, 62F03.

1. INTRODUCTION

The use of models involving a sequence of submodels has been widely applied in areas such as economics, medicine and biology, among others. These types of models, denoted by segmented (or switching or multi-phase) regression models, are useful when it is believed that the model parameters change after an unknown time or in some region of the domain of the predictor variables.

A simple segmented regression model, in the case that a sequence of observations (x_i, y_i), $i=1,2,...,n$, is considered, can be written in the following way

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & \text{if } x_i \leqslant r \\ \alpha_2 + \beta_2\, x_i + \varepsilon_i, & \text{if } x_i > r \ , \end{cases} \tag{1.1}$$

where α_1, α_2, β_1, β_2 and r are unknown parameters and the errors ε_i have distribution $N(0,\sigma^2)$. The submodels in (1.1) are referred to as regimes or segments and the point r as the change point.

Segmented regression models are divided into two types. One where the model is assumed to be continuous at the change point, and the other where it is not. The inferential theory is completely different for each type of model (Hawkins ([5])). The emphasis in this article is on the discontinuous model.

The linear-linear segmented regression model proposed by Quandt ([1]) is similar to model (1.1), except that the change point is identified by the observation order instead of the observation value as above. Moreover, the model (1.1) assumes homoscedasticity while the Quandt model assumes heteroscedasticity.

Considering a sequence of observations (x_i, y_i), $i = 1,2,...,n$, the Quandt two-phase regression model is given by

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i, & \text{if } i = 1,...,k \\ \alpha_2 + \beta_2\, x_i + \varepsilon_i, & \text{if } i = k+1,...,n \ , \end{cases} \tag{1.2}$$

where the ε_i have independent normal distributions with mean zero and variance σ_1^2 if $i \leqslant k$ and variance σ_2^2 if $i > k$. The parameters α_1, α_2, β_1, β_2 and k are all unknown.

Various tests for the presence of a change point based on the likelihood ratio are discussed in the statistics literature. Quandt ([1], [2]) was the first to propose a likelihood ratio test to detect the presence of a change point in a simple linear regression model. Hinkley ([6]) derived the asymptotic distribution of the maximum likelihood estimate of a change point and the asymptotic distribution of the likelihood ratio statistics for testing hypotheses of no change in (1.2), where the independent observations $x_1, ..., x_n$, are ordered, the change point k is unknown and the errors ε_i are considered uncorrelated $N(0,\sigma^2)$. Furthermore, it is assumed that $x_k \leqslant \gamma \leqslant x_{k+1}$, where $\gamma = (\alpha_1 - \alpha_2)/(\beta_2 - \beta_1)$ is the intersection of the two regression lines. Hinkley ([7]) discussed inference concerning a change point considering the hypothesis $H_0\colon\ k = k_0$ versus the one-sided alternative $H_1\colon\ k > k_0$ or versus the two-sided alternative $H_2\colon\ k \neq k_0$. Brown *et al.* ([8]) described tests for detecting departures from constancy of regression relationships over time and illustrated the application of these tests with three sets of real data. Maronna and Yohai ([9]) derived the likelihood ratio test for the hypothesis that a systematic change has occurred after some point in the intercept alone. Worsley ([10]) found exact and approximate bounds for the null distributions of likelihood ratio statistics for testing hypotheses of no change in the two-phase multiple regression model

$$y_i = \begin{cases} \mathbf{x}_i' \beta + \varepsilon_i, & \text{if } \ i = 1, ..., k \\ \mathbf{x}_i' \beta^* + \varepsilon_i, & \text{if } \ i = k+1, ..., n \ , \end{cases} \tag{1.3}$$

where $p \leqslant k \leqslant n-p$, $\mathbf{x}_i$ is a p-component vector of independent variables and β and β^* are p-component vectors of unknown parameters. His numerical results indicated that the accuracy of the approximation of the upper bound is very good for small samples. Kim and Siegmund ([3]) consider likelihood ratio tests for change point problems in simple linear regression models. They also present a review on segmented regression models and some real problems that motivated research in this area. Some of these problems are examined using change point methods by Worsley ([11]) and Raferty and Akman ([12]). Kim ([4]) derived the likelihood ratio tests for a change in simple linear regression with unequal variances. Kim and Cai ([13]) examined the distributional robustness of the likelihood ratio tests discussed by Kim and Sigmund ([3]) in a simple linear regression. They showed that these statistics converge to the same limiting distributions regardless of the underlying distribution. Through simulation the observed distributional insensitivity of the test statistics is observed when the errors follow a lognormal, a Weibull, or a contaminated normal distribution. Kim ([4]), using some numerical examples, examined the robustness to heteroscedasticity of these tests.

In this paper different likelihood ratio tests (Quandt ([1]) and ([2]), Kim and Siegmund ([3]) and Kim ([4])) to detect a change on a simple linear regression, are presented. The tests are evaluated and compared regarding their performance in different scenarios. Our main concern is to assist the user of such tests to decide which test is preferable to use and under which circumstances. The article is organized as follows. In Section 2, the likelihood ratio tests proposed by Quandt ([1]) and ([2]), Kim and Siegmund ([3]) and Kim ([4]) will be described. In Section 3, via Monte Carlo simulations, the performance of the tests discussed in Section 2 will be assessed and compared. Final comments on the results, presented in Section 4, will conclude the paper.

2. TEST STATISTICS

In this Section the likelihood ratio tests proposed by Quandt ([1]) and ([2]), Kim and Siegmund ([3]) and Kim ([4]) are described in more detail. In all the cases the model (1.2) is considered.

2.1. Likelihood Ratio Test by Quandt (Q-Test)

The test described by Quandt ([1]) and ([2]) is used for testing the hypothesis that no change has occurred against the alternative that a change took place. That is, $H_0\colon \alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\sigma_1 = \sigma_2$ against $H_1\colon \alpha_1 \neq \alpha_2$ or $\beta_1 \neq \beta_2$ or $\sigma_1 \neq \sigma_2$. The error terms ε_i are independently and normally distributed $N(0, \sigma_1^2)$ for $i = 1, ..., k$ and $N(0, \sigma_2^2)$ for $i = k+1, ..., n$.

The likelihood ratio λ is defined as

$$\lambda = \frac{l(k)}{l(n)} \ ,$$

where $l(n)$ is the maximum of the likelihood function over only a single phase and $l(k)$ is the maximum of the likelihood function over the presence of a change point. That is,

$$\begin{aligned}
\lambda &= \frac{\exp\left[-\log(2\pi)^{\frac{n}{2}} - \log\hat{\sigma}^n - \frac{n}{2}\right]}{\exp\left[-\log(2\pi)^{\frac{n}{2}} - \log\hat{\sigma}_1^k - \log\hat{\sigma}_2^{n-k} - \frac{n}{2}\right]} \\
&= \frac{\hat{\sigma}_1^k\,\hat{\sigma}_2^{n-k}}{\hat{\sigma}^n} \ ,
\end{aligned}$$

where $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the estimates of the standard errors of the two regression lines, $\hat{\sigma}$ is the estimate of the standard error of the overall regression based on all observations and the constant k is chosen in order to minimize λ. On the basis of empirical distributions resulting from sampling experiments, Quandt ([1]) concluded that the distribution of $-2\log\lambda$ can not be assumed to be χ^2 distribution with 4 degrees of freedom.

2.2. Likelihood Ratio Tests by Kim and Siegmund (KS-Test)

Kim and Siegmund ([3]), assuming the model (1.2) with homoscedasticity, consider tests of the following hypotheses:

H_0: $\beta_1 = \beta_2$ and $\alpha_1 = \alpha_2$ against the alternatives

(i) H_1: $\beta_1 = \beta_2$ and there exists a k $(1 \leqslant k < n)$ such that $\alpha_1 \neq \alpha_2$ or

(ii) H_2: there exists a k $(1 \leqslant k < n)$ such that $\beta_1 \neq \beta_2$ and $\alpha_1 \neq \alpha_2$.

The alternative (i) specifies that a change has occurred after some point in the intercept alone and alternative (ii) specifies that a change has occurred after some point in both intercept and slope.

The likelihood ratio test of H_0 against H_1 rejects H_0 for large values of

$$\max_{n_0 \leqslant i \leqslant n_1} |U_n(i)| / \hat{\sigma} ,$$

where $n_j = n t_j$, $j = 0, 1$, for $0 < t_0 < t_1 < 1$, and

$$U_n(i) = (\widehat{\alpha}_i - \widehat{\alpha}_i^*) \left(\frac{i\,(1 - i/n)}{1 - i(\bar{x}_i - \bar{x}_n)^2 / \{Q_{xxn}(1 - i/n)\}} \right)^{1/2} .$$

The likelihood ratio test of H_0 against H_2 rejects H_0 for large values of

$$\sigma^{-2} \max_{n_0 \leqslant i \leqslant n_1} \left\{ \frac{n\, i\, (\bar{y}_i - \bar{y}_n)^2}{n - i} + \frac{Q_{xyi}^2}{Q_{xxi}} + \frac{Q_{xyi}^{*2}}{Q_{xxi}^*} - \frac{Q_{xyn}^2}{Q_{xxn}} \right\} ,$$

where, following Kim and Siegmund ([3]) notation,

$$\bar{x}_i = i^{-1} \sum_{k=1}^{i} x_k , \qquad \bar{y}_i = i^{-1} \sum_{k=1}^{i} y_k , \qquad \widehat{\alpha}_i = \bar{y}_i - \widehat{\beta}\, \bar{x}_i , \qquad \widehat{\alpha}_i^* = \bar{y}_i^* - \widehat{\beta}\, \bar{x}_i^* ,$$

$$\bar{x}_i^* = (n-i)^{-1} \sum_{k=i+1}^{n} x_k , \quad \bar{y}_i^* = (n-i)^{-1} \sum_{k=i+1}^{n} y_k , \quad Q_{xyi} = \sum_{k=1}^{i} (x_k - \bar{x}_i)\,(y_k - \bar{y}_i) ,$$

$$Q_{xxi} = \sum_{k=1}^{i} (x_k - \bar{x}_i)^2 , \qquad Q_{xxi}^* = \sum_{k=i+1}^{n} (x_k - \bar{x}_i^*)^2 , \qquad \ldots ,$$

$$Q_{xxn} = \sum_{k=1}^{n} (x_k - \bar{x}_n)^2 , \qquad Q_{xyn} = \sum_{k=1}^{n} (x_k - \bar{x}_n)\,(y_k - \bar{y}_n) ,$$

$$\widehat{\beta} = Q_{xyn} / Q_{xxn} \qquad \text{and} \qquad \hat{\sigma}^2 = n^{-1} \left(Q_{yyn} - Q_{xyn}^2 / Q_{xxn} \right) .$$

In these tests and in the tests by Kim, the values for t_0 and t_1 depend on the feeling we have concerning the location of the change point. This impression comes from a scatterplot of y and x. In this study we will use $t_0 = 0.1$ and $t_1 = 0.9$.

2.3. Likelihood Ratio Tests by Kim (K-Test)

Kim ([4]) studied likelihood ratio tests for a change in a simple linear regression model considering the two types of alternatives presented in the previous subsection. It is assumed that the error variance is non-homogeneous, that is, the error terms follow $N(0,\sigma_i^2)$, where $\sigma_i^2=\sigma^2/w_i$ and the w_i's are positive constants. The likelihood ratio statistics is denoted by *weighted likelihood ratio statistics*.

The weighted likelihood ratio statistics to test H_0 against H_1, ($WLRS_1$), is given by

$$\hat{\sigma}^{-1} \max_{n_0\leqslant i\leqslant n_1} |U_{w,n}(i)| \tag{2.1}$$

where

$$U_{w,n}(i) = \left(\frac{\sum_{k=1}^{i} w_k \sum_{k=1}^{n} w_k}{\sum_{k=i+1}^{n} w_k}\right)^{\frac{1}{2}} \left(\frac{\bar{y}_{w,i} - \bar{y}_{w,n} - \widehat{\beta}_w \cdot (\bar{x}_{w,i} - \bar{x}_{w,n})}{\left[1 - \left\{\sum_{k=1}^{i} w_k \sum_{k=1}^{n} w_k \Big/ \sum_{k=i+1}^{n} w_k\right\} (\bar{x}_{w,i} - \bar{x}_{w,n})^2 \Big/ Q_{xxn}\right]^{\frac{1}{2}}}\right) .$$

The weighted likelihood ratio statistics to test H_0 against H_2, ($WLRS_2$), is given by

$$\hat{\sigma}^{-2} \max_{n_0\leqslant i\leqslant n_1} |V_{w,n}(i)| \ , \tag{2.2}$$

where

$$V_{w,n}(i) \;=\; \frac{\sum_{k=1}^{i} w_k \sum_{k=1}^{n} w_k}{\sum_{k=i+1}^{n} w_k} \, (\bar{y}_{w,i} - \bar{y}_{w,n})^2 + \frac{Q_{xyi}^2}{Q_{xxi}} + \frac{Q_{xyi}^{*2}}{Q_{xxi}^*} - \frac{Q_{xyn}^2}{Q_{xxn}} \ .$$

In both tests $n_j=nt_j$, $j=0,1$, for $0<t_0<t_1<1$, and following Kim ([4]) notation,

$$\bar{x}_{w,i} \;=\; \left(\sum_{k=1}^{i} w_k\right)^{-1} \sum_{k=1}^{i} w_k\, x_k \ , \qquad \bar{x}_{w,i}^* \;=\; \left(\sum_{k=i+1}^{n} w_k\right)^{-1} \sum_{k=i+1}^{n} w_k\, x_k \ ,$$

$$\bar{y}_{w,i} \;=\; \left(\sum_{k=1}^{i} w_k\right)^{-1} \sum_{k=1}^{i} w_k\, y_k \ , \qquad \bar{y}_{w,i}^* \;=\; \left(\sum_{k=i+1}^{n} w_k\right)^{-1} \sum_{k=i+1}^{n} w_k\, y_k \ ,$$

$$Q_{xxi} = \sum_{k=1}^{i} w_k \left(x_k - \bar{x}_{w,i}\right)^2 , \qquad Q^*_{xxi} = \sum_{k=i+1}^{n} w_k \left(x_k - \bar{x}^*_{w,i}\right)^2 ,$$

$$Q_{yyi} = \sum_{k=1}^{i} w_k \left(y_k - \bar{y}_{w,i}\right)^2 , \qquad Q^*_{yyi} = \sum_{k=i+1}^{n} w_k \left(y_k - \bar{y}^*_{w,i}\right)^2 ,$$

$$Q_{xyi} = \sum_{k=1}^{i} w_k \left(x_k - \bar{x}_{w,i}\right)\left(y_k - \bar{y}_{w,i}\right) ,$$

$$Q_{xxn} = \sum_{k=1}^{n} w_k \left(x_k - \bar{x}_{w,n}\right)^2 , \qquad Q_{xyn} = \sum_{k=1}^{n} w_k \left(x_k - \bar{x}_{w,n}\right)\left(y_k - \bar{y}_{w,n}\right) ,$$

$$\widehat{\beta}_w = Q_{xyn}/Q_{xxn} \qquad \text{and} \qquad \hat{\sigma}^2 = n^{-1}\left(Q_{yyn} - Q^2_{xyn}/Q_{xxn}\right) .$$

Kim ([4]) also presents approximations of the p-values of the *WLRS* (2.1) and (2.2).

3. PERFORMANCE OF THE TESTS

To assess and compare the performance of the tests described in the previous section we carried out a Monte Carlo simulation taking into account different scenarios, such as, different error distributions, different sample sizes, different locations of the change point and departure from the homoscedasticity assumption. In all the cases two different alternative hypotheses were considered: i) with a change in the intercept from one model to the other with the same slope, and ii) with a change in both intercept and slope.

In the simulation process, for each model, the sequence of observations $(x_1, y_1), ..., (x_n, y_n)$, where $x_i = i/n,\ i = 1, ..., n$ and y_i satisfying the model (1.2), were generated 5,000 times. To calculate the distributional insensitivity of the tests the following distributions for the errors were considered: $N(0,1)$, $N(0,1)$ for one regime and $N(0,4)$ for the other, normal with variance given by $1/w_i$, where $w_i = (1 + i/n)^2$, contaminated normal using the mixture distribution $0.95\,N(0,1) + 0.5\,N(0,9)$, exponential(1), Weibull(α,γ), with $\alpha = 1.5$ and $\gamma = 1/(2^{1/2\alpha}\{\Gamma(2/\alpha + 1) - \Gamma^2(1/\alpha + 1)\}^{0.5})$, and lognormal($\alpha, \gamma$), with $\alpha = 0.1$ and $\gamma = \{\exp(2\alpha) - \exp(\alpha)\}^{-0.5}$. For more details for the choice of these parameters interested readers can see Kim and Cai (1993).

To generate the Weibull and lognormal random errors we use the methods presented in Kim and Cai (1993). The Weibull random errors are generated by the transformation $\varepsilon = \gamma(Z_1^2 + Z_2^2)^{1/\alpha}$, where Z_1 and Z_2 are independent and follow standard normal distributions. The lognormal random errors are generated considering $\varepsilon = \gamma \exp(\sqrt{\alpha}\, Z_1)$. The exponential and the normal random variables are generated using SAS/IML functions. The $N(0, 1/w_i)$ distribution for the errors implies that each observation has a different variance determined by the value w_i; it is used to compare the K-Test with the other tests. The K-Test is applied only in models in which a change can occur in the intercept alone.

Following Zhang and Boss ([14]), Monte Carlo estimates of critical values are used to create adjusted power estimates. It allows for comparisons among the competing tests under the same scenarios.

3.1. A Change in the Intercept

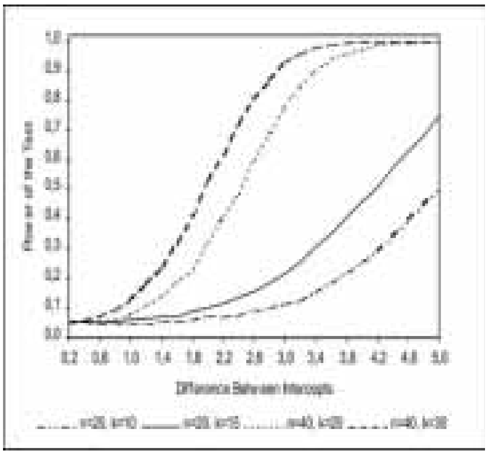
We start with the analysis of the performance of the Q-Test for models with a change in the intercept alone. Figure 1 – panel (a) shows the power of the Quandt test for different sample sizes and different locations of the change point considering models with errors following normal distributions. In all the cases the critical values are the 95$^{\text{th}}$ percentiles under H_0, estimated by Monte Carlo simulations. The best performance of the test occurs when the change point is not at the central position, that is, when the number of observations in a regime is much smaller than the number of observations in the other, and also when the sample size increases.

Figure 1 – panel (b) shows the performance of the Q-Test for homogeneous error variance models with different distributions for the errors. In these cases the sample size is 20 and the change point is located at $k = 15$. It can be concluded that the likelihood ratio test of Quandt achieves almost the same power for the four different distributions for the errors. Similar results are reached for other different sample sizes and different locations of the change point. The performance of the Q-Test for non-homogeneous error variance models is showen in Figure 1 – panel (c), for $n = 20$ and $k = 15$, when the errors have normal distributions with different variances from one regime to the other and when the errors have contaminated normal distribution with each observation having a different variance. The standard normal distribution model is also presented. The power of the likelihood ratio test of Quandt achieves almost the same power for the three distributions when the difference between the intercepts of the two regimes is less than 2, after that the behavior of the power functions are different. Other sample sizes and locations of the change point were explored and the results were similar.

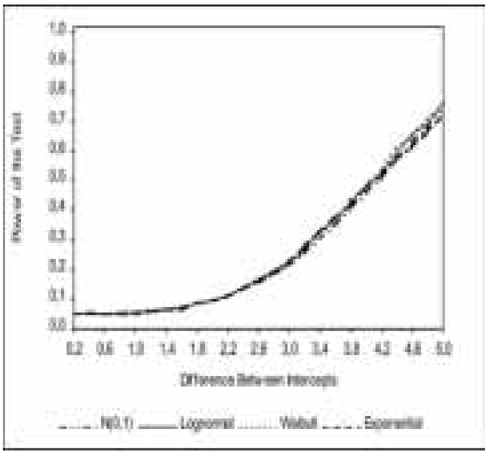
The performance of the KS-Test for models with a change in the intercept alone is now analyzed. Figure 2 − panel (a) shows the power of the KS-Test for models with errors following $N(0,1)$ considering different sample sizes and different locations of the change point in both regimes. This test performs well in those cases where the changes occurred far from the center. This is in broad agreement with the results of Kim and Cai ([13]). The Figure 2 − panel (b) shows the robustness of the KS-Test concerning different distributions of the errors but with homogeneous variances. However, for non-homogeneous models, as shown in Figure 2 − panel (c), the performance of the KS-Test depends on the distribution of the errors. For the cases where the errors follow non-homogeneous normal (different variances for each regime) and contaminated normal distribution the performance of the test is clearly inferior to the performance when the errors follow homoscedastic distributions. When the considered distribution is the heteroscedastic normal the performance of the test is worse than in the contaminated normal distribution case.

The non-homogeneous models considered in this study were the models in which the variances of the error terms are proportional to the square of a linear combination of the regressor variables. That is, the models have errors following a $N(0,1/w_i)$, with $w_i = (1+i/n)^2$. Models with this type of heteroscedasticity, known as additive heteroscedastic error models, have been discussed by Rutemiller and Bowers ([15]) and Harvey ([16]). These models were submitted to the Q, KS and K-Tests. The test powers are shown in Figure 3. Note that the Q-Test does not present a good performance in these cases. The KS-Test is better than the Q-Test but sensibly worse than the K-Test. Comparing such results with the results presented previously, the non robustness of the Q and KS-Tests is evident when applied to non-homogeneous models, mainly when the variances of the errors are different from an observation to another and not only from a regime to another. The K-Test is more powerful in the case where the variances of the error terms are different but known. In the cases of contaminated normal and heteroscedastic normal distributions, where the w_i is unknown, the application of the K-Test can be accomplished by taking $w_i = 1$, for $i = 1, 2, ..., n$, which would make such a test equivalent to the KS-Test, or by estimating these weights from data.

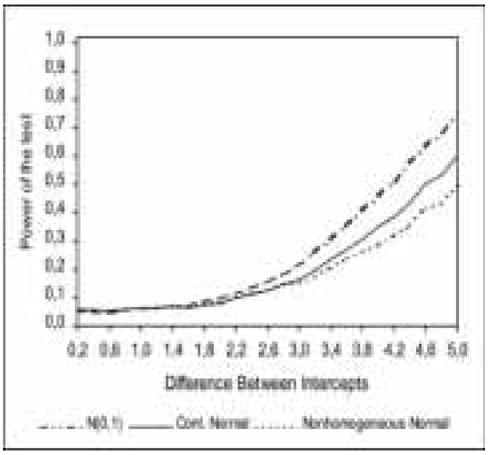
Comparatively, the Q-Test and KS-Test presented very different results for samples of size 20. The performance of the KS-Test is superior to the performance of the Q-Test. That superiority is noticed practically in all the considered differences between intercepts. It is important to point out that even when we considered the non-homogeneous error variance models submitted to the KS-Test, the results are better than the one of the Q-Test in homogeneous error variance models. Analyzing the Figures 1 and 2, a similar performance of the Q-Test and KS-Test when the samples are of size $n = 40$ is noticed. The K-Test is more efficient than the others in non-homogeneous error variance models with known variance.



(*a*)



(*b*)



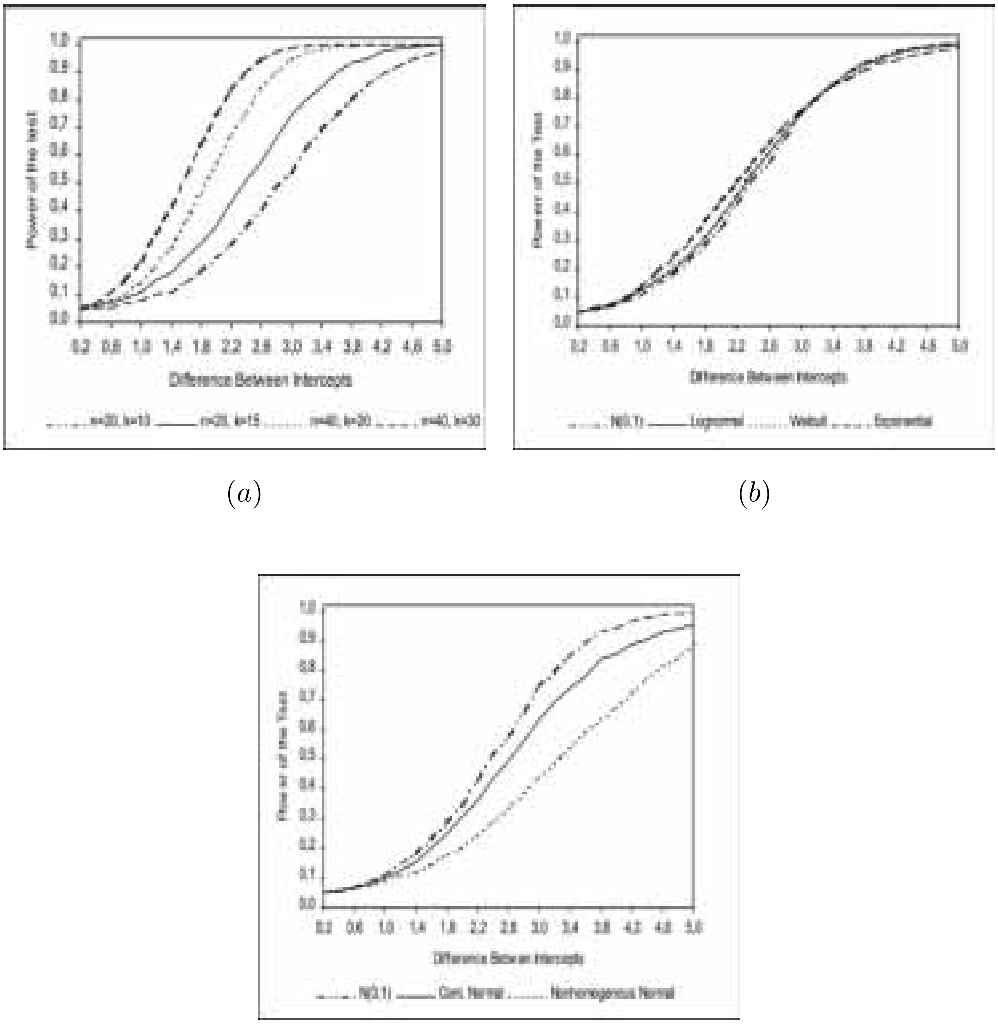
(*c*)

Figure 1: Power of the Q-Test in three scenarios:

Panel (*a*) for models with error terms following $N(0,1)$ considering different sample sizes and different locations of the change point. In this case the best performance of the test occurs when the change point is not at the central position.

Panel (*b*) for homogeneous error variance models supposing the error terms have normal, lognormal, Weibull and exponential distribution, with $n = 20$ and $k = 15$. Note that the likelihood ratio test of Quandt achieves almost the same power for the four different distributions.

Panel (*c*) for models with error terms following N(0,1), nonhomogenous normal and contaminated normal distribution, with $n = 20$ and $k = 15$. The power of the Q-Test achieves almost the same power for the three distributions when the difference between the intercepts of the two regimes is less than 2. Afterwards that, the behavior of the power functions are different.



(a) (b) (c)

Figure 2: Power of the KS-Test in three scenarios:

Panel (a) for models with error terms following N(0,1) considering different sample sizes and different locations of the change point. This test performs well in those cases where the changes occurred far from the center.

Panel (b) for homogeneous error variance models supposing the error terms follow N(0,1), lognormal, Weibull or exponential distribution, with $n = 20$ and $k = 15$. It shows the robustness of the KS-Test regarding the four different distributions.

Panel (c) for models with error terms following N(0,1), non-homogenous normal and contaminated normal distributions, with $n = 20$ and $k = 15$. In this case the performance of the KS-Test depends on the distribution of the errors.

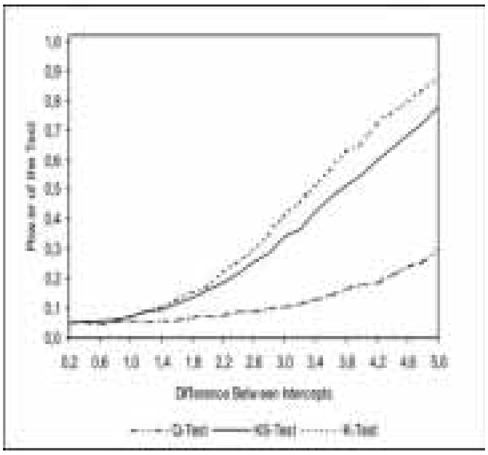


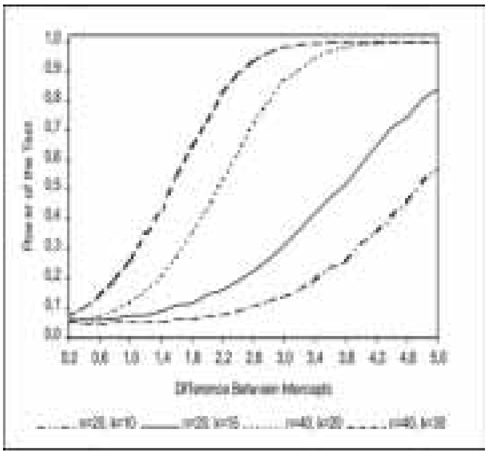
Figure 3: Power of the Q, KS and K-Tests for models with errors following $N(0, 1/w_i)$, where $w_i = (1 + i/n)^2$, for $i = 1, 2, ..., n$. The Q-Test does not present a good performance in these cases. The KS-Test is better than the Q-Test but sensibly worse than the K-Test.

3.2. Change in both Intercepts and Slope

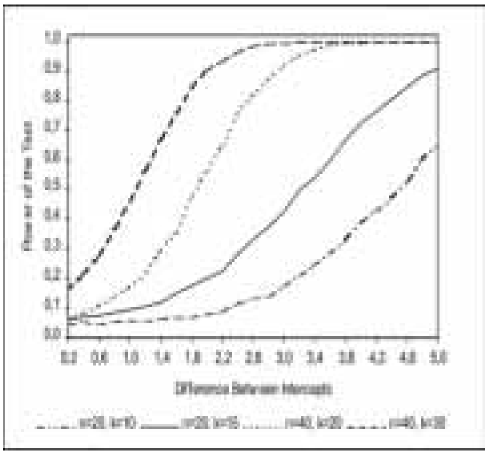
The cases considered in the previous section are analyzed here but with different slopes on the first and second regimes of each model. The difference between these slopes considered in the simulations are 0.5 and 1.0. The tests will continue to be denominated Q-Test and KS-Test but taking into account the versions that consider models with different intercepts and slopes between the regimes. Recall that the K-Test is applied only to models in which a change can occur in the intercept alone.

The Figures 4 − panel (a) and 4 − panel (b) present the results obtained with the Q-Test in models with distribution $N(0, 1)$, considering the differences between the slopes 0.5 and 1.0, respectively. The test is sensitive to the change in the difference between such slopes. Besides, as it happened with models where only the intercepts change, the results are clearly better when the change point is on a non central region. Moreover in these cases, the performance of the Q-Test improves as the sample size increases from 20 for 40.

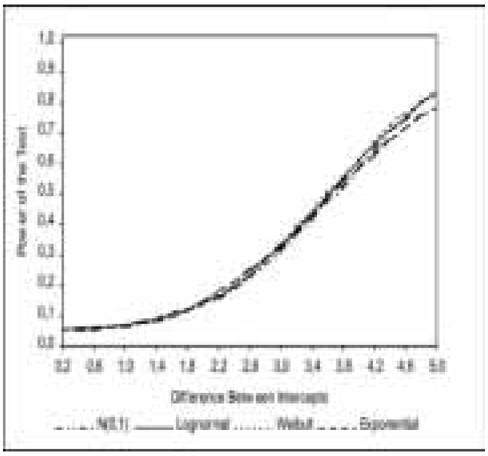
The Figures 4 − panel (c) and 4 − panel (d) present the results obtained with the Q-Test in homogeneous error variance models considering several distributions for the error terms with $n = 20$ and $k = 15$. Here it is also clear that the test is sensitive to the increase of the difference between the slopes. Both Figures evidence the robustness of the test in relation to the considered distributions of the errors.



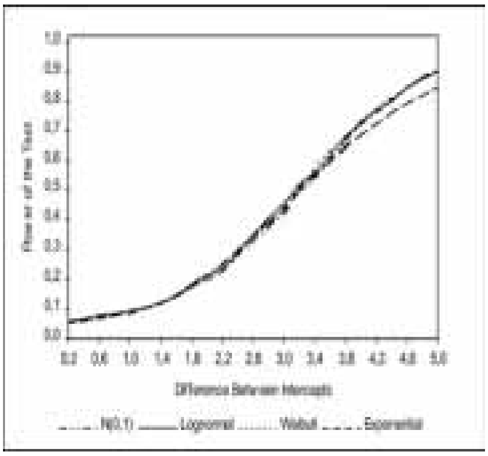
(a)



(b)



(c)



(d)

Figure 4: Power of the Q-Test in four scenarios:

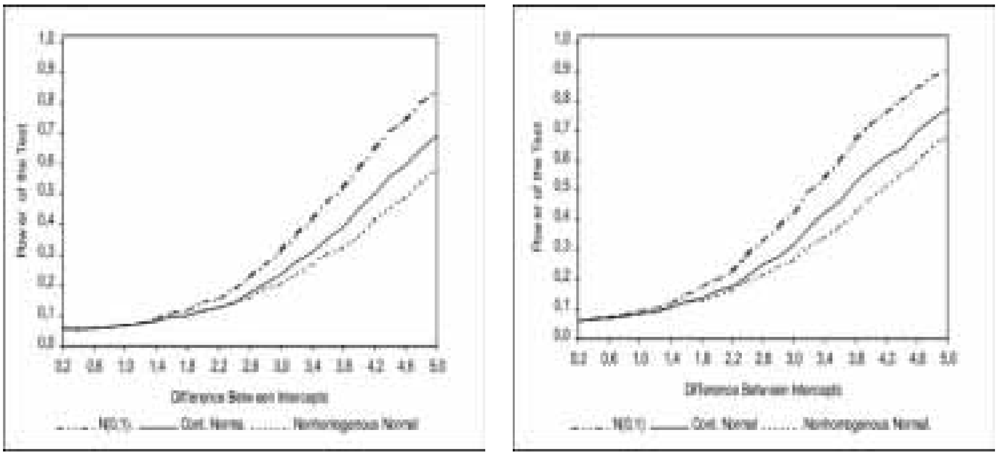
Panel (a) for models with error terms following $N(0,1)$ considering different sample sizes, different locations of the change point and $\beta_2 - \beta_1 = 0.5$.

Panel (b) for models with error terms following $N(0,1)$ considering different sample sizes, different locations of the change point and $\beta_2 - \beta_1 = 1.0$. In both cases the test is sensitive to the change in the difference between such slopes and improves as the sample size increases from 20 for 40.

Panel (c) for homogeneous error variance models supposing the error terms follow distribution $N(0,1)$, lognormal, Weibull and exponential, $n = 20$, $k = 15$ and $\beta_2 - \beta_1 = 0.5$.

Panel (d) for homogeneous error variance models supposing the error terms follow distributions $N(0,1)$, lognormal, Weibull and exponential, $n = 20$, $k = 15$ and $\beta_2 - \beta_1 = 1.0$. Both Figures evidence the robustness of the test in relation to the considered distributions of the errors.

The performance of the Q-Test in non-homogeneous error variance models in comparison to the case $N(0,1)$ can be seen in Figures $5-$panel$\,(a)$ and $5-$panel$\,(b)$, these also evidence that the test is sensitive to the increase of the difference between the slopes. Once again, the non robustness of the test studied in relation to the presence of heteroscedasticity can be clearly seen.



(a) (b)

Figure 5: Power of the Q-Test in two scenarios:
Panel (a) for models with error terms following $N(0,1)$, non-homogenous normal and contaminated normal distributions, $n=20$, $k=15$ and $\beta_2-\beta_1=0.5$.
Panel (b) for models with error terms following $N(0,1)$, non-homogenous normal and contaminated normal distributions, $n=20$, $k=15$ and $\beta_2-\beta_1=1.0$. In both cases there is evidence that the test is sensitive to the increase of the difference between the slopes.

The KS-Test when applied to models in which changes occurs in both intercept and slope is sensitive to the increase in the difference between the slopes and its best performance happens when the structural change occurs in a non central region. Another property of the KS-test is the non robustness concerning the heteroscedasticity of the model. Such properties were verified by the analysis of the plots (not shown here) which present the power of the KS-Test applied to the same models submitted to the Q-Test.

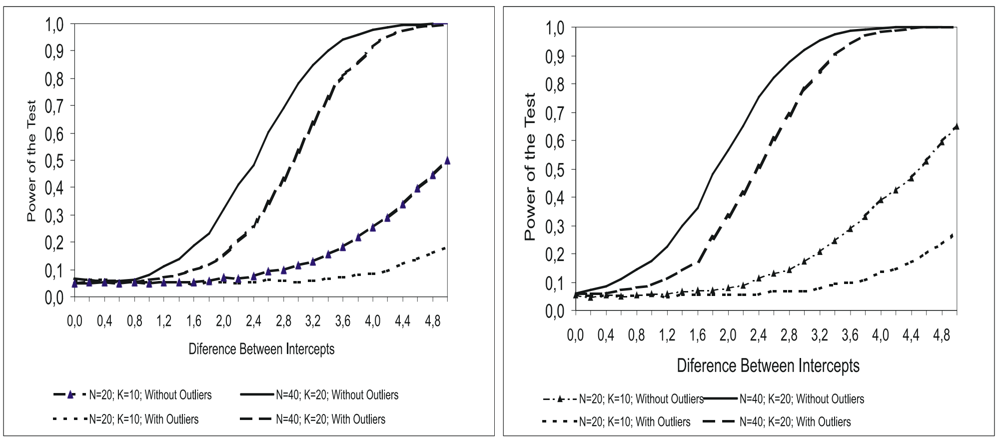
When the difference between intercepts is less than 1.0, the KS-Test presents superior performance, in both cases $\beta_2-\beta_1=0.5$ and $\beta_2-\beta_1=1.0$, in non-homogeneous normal models in relation to the $N(0,1)$ models. However, when the difference between intercepts is larger than 1.0, such superiority is total and completely reverted in favor of the cases $N(0,1)$, also in both cases $\beta_2-\beta_1=0.5$ and $\beta_2-\beta_1=1.0$.

Q-Test and KS-Test are sensitive to the alterations of differences between intercepts and differences between slopes, besides both tests present better performances when the structural change happens in a non central region. Another characteristic of the Q-Test and KS-Test is the non robustness of their performance in non-homogeneous error variance models in relation to the homogeneous models.

In all the presented cases the KS-Test performs better than the Q-Test. Both are shown to be robust regarding the different distributions considered in homogeneous models, except the KS-Test when applied in models with error following exponential distribution. However, even presenting inferior power in the exponential case, the KS-Test has more power than the Q-Test in all the homogeneous cases.

4. WORKING WITH OUTLIERS

In this section a small investigation of the sensitivity of the tests referring to outlying observations is presented. New data sets, contaminated with outliers, are simulated from the model (1.2) and the power results for the three tests, for some scenarios, are explored. This investigation involves the power of the Q-test, KS-Test and K-test considering data sets with and without outliers, different sample sizes and different locations of the change point, for errors following normal distributions. In all the cases the critical values are the 95$^{\text{th}}$ percentiles under H_0, estimated by Monte Carlo simulations.



(*a*) (*b*)

Figure 6: Power of the Q-Test for models with error terms following $N(0,1)$ considering data sets with and without outliers, different sample sizes and different locations of the change point. Panel (a) $\beta_2-\beta_1=0$. Panel (b) $\beta_2-\beta_1=1.0$.

For the Q-Test and KS-Test, if the outliers are clearly present in one of the regimes, inspection of the results reveals that the cases without outliers have slightly more power when compared to the cases with outliers, but for the K-test the power of cases with and without outliers are comparable.

For the three studied tests, the cases where the outliers are clearly in the change point region are slightly more powerful than those cases without outliers. The reason for that can reside in the fact that the simulated outliers reinforced the presence of change points.

5. CONCLUDING REMARKS

The robustness of different likelihood ratio tests was investigated under different scenarios via a simulation study as presented in Section 3. With the exception of the study by Kim and Cai ([13]) there has been little work done for a comprehensive discussion of the performance of these tests including the important question of deciding which test should be considered and under which circumstances. The simulation results suggested that the KS-Test is superior to the Q-Test when small to moderate sample sizes are considered for both homogeneous and non-homogeneous models with a change in the intercept alone. However, the K-Test is more powerful than the other two tests for non-homogeneous models with a known variance. The Q-Test and KS-Test are both robust regarding different distributions of the errors for homogeneous models. When there is a change point in both intercept and slope the KS-Test is superior to the Q-Test in all investigated scenarios.

ACKNOWLEDGMENTS

The authors would like to thank the editorial board and the referees for their helpful comments and suggestions in the early version of this paper.

REFERENCES

[1] Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes, *J. Amer. Statist. Assoc.*, **53**, 873–880.

[2] Quandt, R.E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes, *J. Amer. Statist. Assoc.*, **55**, 324–330.

[3] Kim, H.J. and Siegmund, D. (1989). The likelihood ratio test for a change point in simple linear regression, *Biometrika*, **76**, 409–423.

[4] Kim, H.J. (1993). Two-phase regression with non-homogeneous errors, *Commun. Statist.*, **22**, 647–657.

[5] Hawkins, D.M. (1980). A note on continuous and discontinuous segmented regressions, *Technometrics*, **22**, 443–444.

[6] Hinkley, D.V. (1969). Inference about the intersection in two-phase regression, *Biometrika*, **56**, 495–504.

[7] Hinkley, D.V. (1971). Inference in two-phase regression, *Am. Statist. Assoc.*, **66**, 736–743.

[8] Brown, R.L.; Durbin, J. and Evans, J.M. (1975). Techniques for testing the constancy of regression relationships over time, *Journal of Royal Statistical Society*, **B**, 149–192.

[9] Maronna, R. and Yohai, V.J. (1978). A bivariate test for the detection of a systematic change in mean, *J. Amer. Statist. Assoc.*, **73**, 640–645.

[10] Worsley, K.J. (1983). Test for a two-phase multiple regression, *Technometrics*, **25**, 35–42.

[11] Worsley, K.J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables, *Biometrika*, **73**, 91–104.

[12] Raferty, A.E. and Akman, V.E. (1986). Bayesian analysis of a Poisson process with a change-point, *Biometrika*, **73**, 85–89.

[13] Kim, H.J. and Cai, L. (1993). Robustness of the likelihood ratio for a change in simple linear regression, *J. Amer. Statist. Assoc.*, **88**, 864–871.

[14] Zhang, J. and Boss, D.D. (1994). Adjusted power estimates in Monte Carlo experiments, *Commun. Statist.*, **23**, 165–173.

[15] Rutemiller, H.C. and Bowers, D.A. (1968). Estimation in a heteroscedastic regression models, *J. Amer. Statist. Assoc.*, **63**, 552–557.

[16] Harvey, A.C. (1974). *Estimation of parameters in a heteroscedastic regression models*, Paper presented at the European Meeting of the Econometrics Society, Grenoble.

REVSTAT – Statistical Journal
Volume 3, Number 1, June 2005, 19–44

ASYMPTOTIC BEHAVIOUR OF REGULAR ESTIMATORS

Authors: Jean Diebolt
– CNRS, Université de Marne-la-Vallée,
Équipe d'Analyse et de Mathématiques Appliquées,
5, boulevard Descartes, Batiment Copernic, Champs-sur-Marne,
77454 Marne-la-Vallée Cedex 2 (jean.dieb@wanadoo.fr)

Armelle Guillou
– Université Paris VI, Laboratoire de Statistique Théorique et Appliquée,
Boîte 158, 175 rue du Chevaleret
75013 Paris (guillou@ccr.jussieu.fr)

Received: November 2003 Revised: October 2004 Accepted: December 2004

Abstract:

- The P.O.T. (Peaks-Over-Threshold) approach consists of using the generalized Pareto distribution (GPD) to approximate the distribution of excesses over thresholds. We use the maximum likelihood estimators, or some other ones satisfying regularity conditions, to estimate the parameters of the approximating distribution. We study the asymptotic bias of these estimators and also the functional bias of the estimated GPD.

Key-Words:

- *Extreme values; domain of attraction; excesses; generalized Pareto distribution; maximum likelihood estimators.*

AMS Subject Classification:

- 60G70, 62G20.

1. INTRODUCTION

In many statistical applications it is necessary to make inferences about the tail of a distribution, where little data is available. For instance, one is often interested in the probability that the maximum of n random variables exceeds a given threshold or, *vice versa*, one wants to determine a level such that the exceedance probability is below a given small value. As an example, an hydraulics engineer has to estimate the necessary height of a dike such that the probability of a flooding in a given year is less than 10^{-4} (cf. Dekkers and de Haan, 1989). This interest has given rise to a rapid development of extreme value theory in the last thirty years (see e.g. Galambos, 1978, Leadbetter *et al.*, 1983). The traditional approach to the analysis of extreme values in a given population is based on the family of generalized extreme value (GEV) distributions. More precisely, Gnedenko (1943) showed that the limit distribution of the maximum $X_{n,n}$ of a sample of independent and identically distributed variables $X, X_1, ..., X_n$ from a distribution F, properly centred and normalized, is necessarily of extreme value type, i.e. for some $\gamma \in \mathbb{R}$, there exists sequences of constants $\sigma_n > 0$ and $\alpha_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - \alpha_n}{\sigma_n} \le x\right) \longrightarrow H_\gamma(x) \,, \tag{1.1}$$

for all continuity points of the extreme value distribution function H_γ, defined as

$$H_\gamma(x) = \begin{cases} \exp\bigl(-(1+\gamma x)^{-\frac{1}{\gamma}}\bigr) & \text{for} \quad \gamma \neq 0 \quad \text{and} \quad 1+\gamma x > 0\,, \\ \exp\bigl(-\exp(-x)\bigr) & \text{for} \quad \gamma = 0 \quad \text{and} \quad x \in \mathbb{R}\,. \end{cases}$$

The distribution function F is said to belong to the maximum domain of attraction of H_γ. The real-valued parameter γ is referred to as the extreme value index (EVI) of F. Most common continuous distribution functions satisfy this weak condition. Distributions with $\gamma > 0$ are called heavy-tailed, as their tail $\bar{F}$ typically decays slowly as a power function. Examples in this Fréchet class are the Pareto, Burr, Student's t, α-stable ($\alpha < 2$) and loggamma distributions. The Gumbel class of distributions with $\gamma = 0$ encompasses the exponential, normal, lognormal, gamma and classical Weibull distributions, the tail of which diminishes exponentially fast. Finally, the Weibull class consists of distributions with $\gamma < 0$, which all have a finite right endpoint $s_+(F) := \sup\{x : F(x) < 1\}$. Examples in this class are the uniform and reverse Burr distributions.

The problem of estimating the so-called extreme value index γ, which determines the behaviour of the underlying distribution function F in its upper tail, has received much attention in the literature. An extensive motivation of this estimation problem can be found in Galambos (1978). The GEV distribution is appropriate when the data consist of a set of maxima. However, there has been some criticism of this approach, because using only maxima leads to the loss of information contained in other large sample values in a given period. This problem is remedied by considering several of the largest order statistics instead of just the largest one: that is, considering all values larger than a given threshold. The differences between these values and a given threshold are called excesses over the threshold. Denote by $F_u(x) := \mathbb{P}(X - u \le x \mid X > u)$ the distribution of the excesses of X over u, given that u is exceeded, and by $G_{\gamma,\sigma}$ the generalized Pareto distribution (GPD) defined, for all $x \ge 0$, as

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \dfrac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{for} \quad \gamma \neq 0 \;\text{ and }\; 1 + \gamma x/\sigma > 0\,, \\ 1 - \exp\left(-\dfrac{x}{\sigma}\right) & \text{for} \quad \gamma = 0\;, \end{cases}$$

where σ and γ are the scale and shape parameters.

Pickands' and Balkema and de Haan's result (see Pickands (1975) and Balkema and de Haan (1974)) on the limiting distribution of excesses over a high threshold states that condition (1.1) holds if and only if

$$\lim_{u \to s_+(F)} \sup_{0 < x < s_+(F) - u} \left| F_u(x) - G_{\gamma,\sigma(u)}(x) \right| = 0$$

for some positive scaling function $\sigma(u)$ depending on u.

Thus, if, for some n, one fixes a high threshold u_n and selects from a sample $X_1, ..., X_n$ only those observations $X_{i_1}, ..., X_{i_{N_n}}$ that exceed u_n, a GPD with parameters γ and $\sigma_n = \sigma(u_n)$ is likely to be a good approximation for the distribution F_{u_n} of the N_n excesses $Y_j := X_{i_j} - u_n$, $j = 1, ..., N_n$. This is called the Peaks-Over-Threshold (P.O.T.) method.

Several methods have been proposed for estimating the parameters of the GPD, for example the method of moments, of the probability-weighted moments introduced by Hosking and Wallis (1987) or the maximum likelihood method (Smith, 1987). In this paper, we look in more details at the maximum likelihood estimators, but also we derive the more general conditions required on the estimators $(\hat{\gamma}_n, \hat{\sigma}_n)$ in order to obtain our results.

In all the sequel, we denote by

$$\mathbb{A}_{n,u_n}(x) = \frac{1}{N_n} \sum_{j=1}^{N_n} \mathbb{1}_{\{Y_j \le x\}}$$

the empirical distribution function of the excesses.

It is of course very important to measure the error between $\bar{F}_{u_n} := 1 - F_{u_n}$ (unknown) and its estimator $\bar{G}_{\hat{\gamma}_n,\hat{\sigma}_n} := 1 - G_{\hat{\gamma}_n,\hat{\sigma}_n}$. This error can be splitted into two parts: an approximation error and an estimation error. The first one, also called *bias of approximation*, is justified by the fact that the distribution of the excesses over u_n is only approximated by a GPD, which implies a systematic error studied in Worms (2000). Since the distribution of the excesses over u_n is F_{u_n} and not G_{γ,σ_n}, the error due to the estimation of (γ, σ_n) is also divided into an approximation error due to the bias of approximation and a random term due to fluctuations.

Note that N_n follows a binomial distribution $\mathcal{B}(n, 1-F(u_n))$. We suppose, in all the sequel, that $n(1-F(u_n)) \to \infty$ as $n \to \infty$, that is $N_n \to \infty$ in probability. In such a case, $\frac{N_n}{n(1-F(u_n))} \to 1$ in probability, as $n \to \infty$.

Let

$$F^*_{u_n}(y) = F_{u_n}(\sigma_n y) \qquad \text{and} \qquad \mathbb{A}^*_{n,u_n}(y) = \mathbb{A}_{n,u_n}(\sigma_n y)$$

for all $y \in \mathbb{R}_+$. We will study the asymptotic behaviour of

$$\bar{F}_{u_n}(x) - \bar{G}_{\hat{\gamma}_n,\hat{\sigma}_n}(x)$$

where $\hat{\gamma}_n$ and $\hat{\sigma}_n$ are the maximum likelihood estimators, or other regular estimators with properties specified later on.

In what follows, we suppose that F is twice differentiable and that its inverse F^{-1} exists. Let V and A be two functions defined as

$$V(t) = \bar{F}^{-1}(e^{-t}) \qquad \text{and} \qquad A(t) = \frac{V''(\ln t)}{V'(\ln t)} - \gamma\ .$$

We suppose the following first and second order conditions:

$$\lim_{t\to+\infty} A(t) = 0\ , \tag{1.2}$$

and

$$A \text{ is of constant sign at } \infty \text{ and there exists } \rho \leq 0 \text{ such that } |A| \in RV_\rho\ , \tag{1.3}$$

(see Bingham *et al.*, 1987).

Under these assumptions, it is proved in Worms (2000) (Theorem 1.4, p. 43) that as $u_n \to s_+(F)$

$$\bar{F}_{u_n}(\sigma_n y) - \bar{G}_\gamma(y) \;=\; a_n\, D_{\gamma,\rho}(y) + o(a_n), \qquad \text{as } n \to +\infty\ , \tag{1.4}$$

for all y, when

$$\sigma_n := \sigma(u_n) = V'\big(V^{-1}(u_n)\big)\ , \qquad a_n := A\big(e^{V^{-1}(u_n)}\big)$$

$$\bar{G}_\gamma(y) := 1 - G_{\gamma,1}(y)\ ,$$

and

$$D_{\gamma,\rho}(y) \;=\; \begin{cases} C_{0,\rho}(y), & \text{if} \quad \gamma = 0\,, \\ C_{\gamma,\rho}\Big(\dfrac{1}{\gamma}\,\ln(1+\gamma y)\Big) & \text{if} \quad \gamma \neq 0\ , \end{cases}$$

where

$$C_{\gamma,\rho}(y) := \mathrm{e}^{-(1+\gamma)y}\, I_{\gamma,\rho}(y) \qquad \text{and} \qquad I_{\gamma,\rho}(y) := \int_0^y \mathrm{e}^{\gamma u} \int_0^u \mathrm{e}^{\rho s}\, ds\, du \;.$$

We also assume that

$$\lim_{n\to\infty} \sqrt{n\,(1-F(u_n))}\; a_n \;=\; \lambda \;\in\; \mathbb{R} \;. \tag{1.5}$$

This is equivalent to suppose that $\sqrt{N_n}\, a_n$ tends to λ in probability, as $n \to \infty$.

The main result of this paper is the following. For a regular class of estimators, when ρ is equal to 0, the error due to the fact that $\bar{F}_{u_n}$ is replaced by $\bar{G}_{\hat{\gamma}_n,\hat{\sigma}_n}$ is of smaller order than the same error in the case $\rho \neq 0$. This result is closely linked to the penultimate approximation for the distribution of the excesses established in Worms (2002) (Gomes and de Haan (2000), generalizing Cohen (1982), Gomes (1984) and Gomes and Pestana (1987), studied penultimate approximation for the distribution of the maximum). At first sight, it can appear a bit strange since it is well known that, if we consider only the problem of the estimation of the index γ, the smaller $|\rho|$, the more difficult it is to estimate the index. This problem of bias in the estimation of the index has been widely studied recently in the literature and justified in particular the work on regression model by Beirlant *et al.* (1999). This paper proves that, on the contrary, if we consider the problem of the estimation of the tail distribution, we do not need to construct asymptotically unbiased estimators, which is essential in the other estimation problem.

The remainder of our paper is organized as follows. In Section 2, we give our main results and the general conditions on the estimators $(\hat{\gamma}_n, \hat{\sigma}_n)$ that we need to obtain our results. Then, in Section 3, we study the asymptotic bias and also the functional bias of the estimated GPD in the case of maximum likelihood estimation with $\gamma > 0$. The details of the proofs are postponed to the appendix.

2. MAIN RESULT

We restrict our attention to a class of estimators that we call "regular estimators" of the couple (γ, σ). We say that an estimator is regular if it has the form $T(\bar{\mathbb{A}}_{n,u_n}) =: (T_1(\bar{\mathbb{A}}_{n,u_n}), T_2(\bar{\mathbb{A}}_{n,u_n}))$, where T satisfies:

($\mathbf{A}_1$) $T(\bar{G}_{\gamma,\sigma}) = \big(T_1(\bar{G}_{\gamma,\sigma}), T_2(\bar{G}_{\gamma,\sigma})\big) = (\gamma, \sigma)$.

($\mathbf{A}_2$) A form of Hadamard differentiability, namely the existence of linear forms $DT(\bar{G}_\gamma) =: (DT_1(\bar{G}_\gamma), DT_2(\bar{G}_\gamma))$, where

$$DT_k(\bar{G}_\gamma)[H] \;=\; \int_0^\infty H\, d\mu_{k,\gamma} \;, \qquad k = 1, 2 \;, \tag{2.1}$$

for some Borelian measures $\mu_{k,\gamma}$ and all $H \in L^1(\mathbb{R}^+, \mu_{1,\gamma}) \cap L^1(\mathbb{R}^+, \mu_{2,\gamma})$ such that under assumption (1.5) we have

$$\lim_{n\to+\infty} \frac{T(\bar{F}^*_{u_n}) - T(\bar{G}_\gamma)}{a_n} = DT(\bar{G}_\gamma)[D_{\gamma,\rho}] \ , \tag{2.2}$$

$$\lim_{n\to+\infty} \sqrt{N_n}\left(T\left(\bar{F}^*_{u_n} + \frac{1}{\sqrt{N_n}}\,\alpha_{N_n}\circ \bar{F}^*_{u_n}\right) - T(\bar{F}^*_{u_n})\right) = DT(\bar{G}_\gamma)\,[\mathbb{B}\circ\bar{G}_\gamma] \ , \tag{2.3}$$

in distribution, where α_k denotes the uniform empirical process and $\mathbb{B}$ a Brownian bridge, and

$$\lim_{n\to+\infty} \frac{T(\bar{G}_{\gamma+a_n}) - T(\bar{G}_\gamma)}{a_n} = DT(\bar{G}_\gamma)\left[\frac{\partial\bar{G}_\gamma}{\partial\gamma}\right] . \tag{2.4}$$

Clearly, condition (A_2) requires that $D_{\gamma,\rho}$, $\mathbb{B}\circ\bar{G}_\gamma$ and $\frac{\partial\bar{G}_\gamma}{\partial\gamma}$ are in $L^1(\mathbb{R}^+, \mu_{1,\gamma}) \cap L^1(\mathbb{R}^+, \mu_{2,\gamma})$.

(**A**$_3$) A scale invariance property, namely for all $\bar{F}$ such that $T(\bar{F})$ is defined and all $\sigma > 0$,

$$T_1\left(\bar{F}\left(\frac{\bullet}{\sigma}\right)\right) = T_1(\bar{F}) \qquad \text{and} \qquad T_2\left(\bar{F}\left(\frac{\bullet}{\sigma}\right)\right) = \sigma\, T_2(\bar{F}) \ . \tag{2.5}$$

As in the Introduction, we use the notation $\sigma_n = \sigma(u_n)$. Then, if we denote by $(\gamma_n^b,\, \sigma_n^b) = T(\bar{F}_{u_n})$ the values of γ and σ obtained when $\bar{G}_{\gamma,\sigma_n}$ has been substituted by the true distribution function $\bar{F}_{u_n}$ of the excesses, we deduce from (2.2)–(2.5) that

$$\frac{1}{a_n}\left(\gamma_n^b - \gamma,\, \frac{\sigma_n^b}{\sigma_n} - 1\right) \longrightarrow DT(\bar{G}_\gamma)[D_{\gamma,\rho}] =: \big(\mathcal{L}_1(\gamma,\,\rho),\, \mathcal{L}_2(\gamma,\,\rho)\big) \ , \tag{2.6}$$

and

$$\sqrt{N_n}\left(\hat{\gamma}_n - \gamma_n^b,\, \frac{\hat{\sigma}_n}{\sigma_n} - \frac{\sigma_n^b}{\sigma_n}\right) \xrightarrow{d} \tag{2.7}$$

$$\xrightarrow{d} \left(\int_0^\infty \mathbb{B}\circ\bar{G}_\gamma\; d\mu_{1,\gamma} \ ,\ \int_0^\infty \mathbb{B}\circ\bar{G}_\gamma\; d\mu_{2,\gamma}\right) =: (Z_1, Z_2) \ .$$

Consequently,

$$\sqrt{N_n}\left(\hat{\gamma}_n - \gamma,\, \frac{\hat{\sigma}_n}{\sigma_n} - 1\right) \xrightarrow{d} (Z_1, Z_2) + \lambda\big(\mathcal{L}_1(\gamma,\,\rho),\, \mathcal{L}_2(\gamma,\,\rho)\big) \ . \tag{2.8}$$

Note that (2.6) contains the bias of approximation of γ and σ, whereas (2.7) involves the limiting distribution of $\hat{\gamma}_n$ and $\hat{\sigma}_n$. This shows that the bias of approximation on the parameters is of order a_n and under (1.5), $\hat{\gamma}_n$ and $\hat{\sigma}_n/\sigma_n$ are asymptotically biased when $\lambda \neq 0$. In this paper, we will focus on the *asymptotic functional bias of approximation*, defined as:

$$A_E(x) := \lim_{n\to\infty} \frac{\bar{F}^*_{u_n}(x) - \bar{G}_{\gamma_n^b,\, \sigma_n^b/\sigma_n}(x)}{a_n} .$$

This quantity is important since it measures the first order non stochastic discrepancy between the unknown target tail function, $\bar{F}^*_{u_n}$, and its observable counterpart, $\bar{G}_{\hat{\gamma}_n,\hat{\sigma}_n/\sigma_n}$. This bias is important to statisticians who are more interested in estimating small tail probabilities than in estimating γ (as Drees (1998) and Drees *et al.* (2004) who have studied the asymptotic behaviour of the maximum likelihood estimators $(\hat{\gamma}_n, \hat{\sigma}_n)$).

Using (1.4) and a Taylor expansion, it can easily be proved that

$$A_E(x) \;=\; D_{\gamma,\rho}(x) \;-\; \mathcal{L}_1(\gamma,\rho)\,\frac{\partial\bar{G}_\gamma}{\partial\gamma}(x) \;+\; \mathcal{L}_2(\gamma,\rho)\,x\,\frac{\partial\bar{G}_\gamma}{\partial x}(x) \;. \tag{2.9}$$

This expression contains both the bias of approximation (1.4) and the error of approximation on the parameters (2.6).

This result has been first established in Diebolt *et al.* (2003) in the special case of the probability-weighted moments estimators of Hosking and Wallis (1987).

Our main result is summarized in the following theorem.

Theorem 1. *Under assumptions* (A_1)–(A_3) *and (1.5), we have*

$$A_E \equiv 0 \qquad \text{when} \;\; \rho = 0 \;.$$

Proof: From (A_1), we clearly have that $\frac{\partial T(\bar{G}_{\gamma,\sigma})}{\partial\gamma} = (1,0)$ and from (2.4) we deduce that $\frac{\partial T(\bar{G}_{\gamma,1})}{\partial\gamma} = DT(\bar{G}_\gamma)\Big[\frac{\partial\bar{G}_\gamma}{\partial\gamma}\Big]$. Therefore, in the case $\rho = 0$, since $D_{\gamma,0} = \frac{\partial\bar{G}_\gamma}{\partial\gamma}$, we derive that $\mathcal{L}_1(\gamma,0) = 1$ and $\mathcal{L}_2(\gamma,0) = 0$. Consequently, $A_E(x) = D_{\gamma,0}(x) - \frac{\partial\bar{G}_\gamma}{\partial\gamma}(x) = 0$ in the case $\rho = 0$. This explains why as soon as we use estimators which satisfy (A_1)–(A_3), the function A_E becomes the null function for $\rho = 0$. □

This result is particularly remarkable. Indeed:

- This result means that in the case $\rho = 0$, although the bias of $\widehat{\gamma}_n$ is of order a_n, its contribution is cancelled by compensation with $D_{\gamma,0}$ in the expression of the function A_E. It can be seen that A_E remains small whenever $|\rho|$ is small.

- The second order parameter ρ is zero for many usual distributions in the Gumbel domain of attraction ($\gamma = 0$): e.g., the normal, lognormal, gamma and classical Weibull distributions. Hence, our result applies to all of these distributions. In the Frechet domain of attraction, we also have distributions with $\rho = 0$, such as the loggamma distribution.
- This result is closely linked to penultimate approximation in Worms (2002).
- This remarkable behaviour of the error of functional approximation A_E when $\rho = 0$ contrasts strongly with theoretical and experimental results concerning the semiparametric estimators of γ, for example the Hill estimator (Hill, 1975). Indeed, the bias of these estimators explodes when ρ tends to 0. This motivates many recent works on exponential regression model, where estimators with reduced bias are proposed (see, for example, Beirlant *et al.*, 1999).

3. MAXIMUM LIKELIHOOD ESTIMATION FOR $\gamma > 0$

In this section, we will prove the following theorem.

Theorem 2. *For $\gamma > 0$, the maximum likelihood estimators are regular in the sense that they satisfy conditions (A_1)–(A_3).*

With this aim, we first establish the local existence and unicity of these estimators. Then, we prove the regularity of T in the maximum likelihood setting with $\gamma > 0$.

We will need some notation. First, we denote by $g_\theta = g_{\gamma,\sigma}$ the density of the GPD distribution with parameters θ. Then the score function $s_\theta(x)$ is the gradient of $\ln g_\theta(x)$ with respect to θ. It is a function of x taking its values in $\mathbb{R}^2$. The score function $s_\theta(x)$ is of the form

$$s_\theta(x) \;=\; \big(s_1(x;\theta),\, s_2(x;\theta)\big)^T$$

where, denoting $y = x/\sigma$,

$$s_1(x;\gamma,\sigma) \;=\; \begin{cases} \dfrac{(1+\gamma y)\,\ln(1+\gamma y) - (\gamma+1)\gamma y}{\gamma^2(1+\gamma y)} & \text{if} \;\; \gamma > 0\,, \\[2ex] \dfrac{y(y-2)}{2} & \text{if} \;\; \gamma = 0\,, \end{cases} \tag{3.1}$$

and

$$s_2(x;\gamma,\sigma) \;=\; \frac{y-1}{\sigma(1+\gamma y)} \qquad \text{if} \;\; \gamma \geq 0\,. \tag{3.2}$$

Let also $\psi_\theta(x)$ denote the derivative of $s_\theta(x)$ with respect to x:

$$\psi_\theta(x) \;:=\; (s_\theta)'(x) \;:=\; \big(\psi_1(x;\theta),\, \psi_2(x;\theta)\big)^T = \left(\frac{x-\sigma}{(\sigma+\gamma x)^2},\, \frac{1+\gamma}{(\sigma+\gamma x)^2}\right) .$$

3.1. Existence and local unicity

We consider the Küllback–Leibler divergence between two densities of probabilities h and g related to a measure of reference ν:

$$\mathrm{Ent}_g\left(\frac{h}{g}\right) \;:=\; \int \ln\left(\frac{h(x)}{g(x)}\right) h(x)\; \nu(dx) \;.$$

It takes, by convention, the value ∞ when the integral is not finite. Let $\theta \in \Theta = \{(\gamma,\sigma)\colon \gamma>0,\ \sigma>0\}$. Under our assumptions, F admits a density f and the two Küllback–Leibler divergences between g_θ and f are defined as (note that they can take the value ∞)

$$d_{KL}(g_\theta \,|\, f) \;:=\; \mathrm{Ent}_f\left(\frac{g_\theta}{f}\right) \;=\; \int_0^\infty \ln\left(\frac{g_\theta(x)}{f(x)}\right) g_\theta(x)\; dx$$

and

$$d_{KL}(f \,|\, g_\theta) \;:=\; \int_0^\infty \ln\left(\frac{f(x)}{g_\theta(x)}\right) f(x)\; dx \;.$$

These quantities are ≥ 0 and $d_{KL}(f|\, g) = 0$ if and only if $f = g$ a.e. Similar properties exist for $d_{KL}(g|\, f)$. Splitting the logarithm into two parts, we obtain

$$d_{KL}(f \,|\, g_\theta) \;=\; \int_0^\infty \ln f(x)\, f(x)\; dx \;+\; \Delta(\theta,\bar F) \;,$$

where

$$\Delta(\theta,\, \bar F) \;:=\; \int_0^\infty \ln g_\theta(x)\; d\bar F(x) \;=\; -\int_0^\infty \ln g_\theta(x)\, f(x)\; dx \;. \tag{3.3}$$

Thus, $d_{KL}(f|\, g_\theta)$ is minimal if and only if $\Delta(\theta,\bar F)$ is minimal. Here, the family $\{g_\theta\colon\ \theta\in\Theta\}$ is identifiable: the distance d_{KL} between g_θ and $g_{\theta'}$ is equal to zero if and only if $\theta = \theta'$.

For each $\bar F$, the function $\theta \mapsto \Delta(\theta,\bar F)$ is continuous on Θ as soon as

$$\int_1^\infty \ln x\;\, f(x)\; dx \;<\; \infty \;. \tag{C$_1$}$$

Lemma 1 below guarantees the local existence of the maximum likelihood estimators.

Lemma 1. *Under (C_1), the restriction of $\theta \mapsto \Delta(\theta, \bar{F})$ to each closed ball $\mathcal{K}$ contained in Θ reaches its minimum value in $\mathcal{K}$ and at each point where this minimum is reached, we have*

$$\int_0^\infty s_\theta(x)\, d\bar{F}(x) \;=\; \underline{0}\ . \tag{3.4}$$

The proof of this lemma is straightforward. Remark that (3.4) constitutes the likelihood equations.

We now consider the local unicity. First define

$$W(\theta, \bar{F}) \;:=\; \int_0^\infty \bar{F}(x)\, \psi_\theta(x)\, dx\ .$$

Integrating by parts yields the following result:

Lemma 2. *Under (C_1),*

$$\bar{F}(x)\, \ln x \underset{x\to\infty}{\longrightarrow} 0\ , \tag{C$_2$}$$

and if $\bar{F}(0) = 1$, we have

$$W(\theta, \bar{F}) \;=\; \left(0, \frac{1}{\sigma}\right)^T - \int_0^\infty s_\theta(x)\, d\bar{F}(x)\ . \tag{3.5}$$

Remark that if (C_2) is satisfied, then (C_1) can be rewritten as $\int_1^\infty (\bar{F}(x)/x)\, dx < \infty$. Moreover, if we assume the mild condition that there exists an $\varepsilon > 0$ such that

$$\bar{F}(x)\, (\ln x)^{1+\varepsilon} \underset{x\to\infty}{\longrightarrow} 0\ , \tag{C$_3$}$$

then (C_1) and (C_2) are satisfied. Note that in Appendix (4.3), we will prove that (C_3) is satisfied by $\bar{F}^*_{u_n}$.

In all the sequel, we use the notation $\theta^\star = (\gamma^\star, 1)$, $\gamma^\star > 0$ denoting the true value of γ and $\bar{G} := \bar{G}_{\theta^\star}$. The local unicity is established in the following lemma.

Lemma 3. *There exists a closed ball $\mathcal{V}^\star$ centered at $\theta^\star$ such that the restriction of the application $\theta \mapsto \Delta(\theta, \bar{F}^*_{u_n})$ to $\mathcal{V}^\star$ is strictly convex for all n sufficiently large, as $\bar{F}^*_{u_n}$ converges to $\bar{G}$ in the sense described in Worms (2000, 2002).*

Proof of Lemma 3: We consider the second order differential $D^2 \ln g_\theta(x)$ with respect to θ, represented by a 2×2 matrix-valued function of x. For each suitable $\bar{F}$ and each $\theta \in \Theta$, the matrix

$$I(\theta, \bar{F}) := \int_0^\infty D^2 \ln g_\theta(x)\, d\bar{F}(x)$$

is a Fisher-type information matrix. Recall that the symmetric matrix $I(\theta, \bar{G}_\theta)$ is definite positive for each $\gamma > -1/2$.

We show in Appendix 4.1 via an integration by parts that the matrix $I(\theta, \bar{F}^*_{u_n})$ converges, uniformly in θ in some compact neighbourhood $\mathcal{V}^\star$ of $\theta^\star$, to a matrix $I(\theta, \bar{G})$, as $n \to \infty$. Consequently, $I(\theta, \bar{F}^*_{u_n})$ is definite positive for all $\theta \in \mathcal{V}^\star$ and all n sufficiently large. In this case, we have for all n sufficiently large, unicity of the minimum of $\Delta(\theta, \bar{F}^*_{u_n})$ for $\theta \in \mathcal{V}^\star$, i.e. local unicity. We denote by $\theta_n = T(\bar{F}^*_{u_n})$ the point in $\mathcal{V}^\star$ minimizing $\Delta(\theta, \bar{F}^*_{u_n})$. $\square$

Remark that the functional T is sequentially continuous in the following sense. Since the $T(\bar{F}^*_{u_n})$'s are in the compact $\mathcal{V}^\star$, the sequence that they form admits at least an adherence value, which belongs to $\mathcal{V}^\star$. We deduce from compactness and identifiability that any adherence value of this sequence is necessarily $\theta^\star$. It therefore follows that $T(\bar{F}^*_{u_n})$ converges to $T(\bar{G}) = \theta^\star$.

3.2. Regularity of T in the maximum likelihood case

In this section, we will essentially prove the Hadamard differentiability of T, with a differential given by

$$H \longmapsto \Big[I\big(T(\bar{G}), \bar{G}\big)\Big]^{-1} \int_0^\infty H(x)\, \psi_{T(\bar{G})}(x)\, dx \,. \tag{3.6}$$

Starting from (3.5), we obtain

$$\begin{aligned}\int_0^\infty \bar{F}^*_{u_n}(x)\, D\psi_\theta(x)\, dx &= \begin{bmatrix} 0 & 0 \\ 0 & -\frac{1}{\sigma^2} \end{bmatrix} - \int_0^\infty Ds_\theta(x)\, d\bar{F}^*_{u_n}(x) \\ &=: M - \int_0^\infty Ds_\theta(x)\, d\bar{F}^*_{u_n}(x) \,,\end{aligned}$$

that is

$$I(\theta, \bar{F}^*_{u_n}) = -\int_0^\infty \bar{F}^*_{u_n}(x)\, D\psi_\theta(x)\, dx + M \,.$$

By local existence and unicity, for all n sufficiently large

$$-\left(0, \frac{1}{\sigma}\right)^T + W(\theta, \bar{F}^*_{u_n}) = \underline{0} \qquad \text{if and only if} \quad \theta = T(\bar{F}^*_{u_n}) \,.$$

Therefore

$$-\Big(0,\ \frac{1}{T_2(\bar F^*_{u_n})}\Big)^T + W\big(T(\bar F^*_{u_n}), \bar F^*_{u_n}\big) \;=\; \underline{0}$$

and

$$-\Big(0,1\Big)^T + W\big(T(\bar G), \bar G\big) \;=\; \underline{0}\ .$$

Thus,

$$\underline{0} \;=\; \frac{1}{a_n}\Bigg[W\big(T(\bar F^*_{u_n}), \bar F^*_{u_n}\big) \;-\; W\big(T(\bar G), \bar G\big) \;-\; \Big(0,\ \frac{1}{T_2(\bar F^*_{u_n})}\Big)^T + \Big(0,1\Big)^T\Bigg]$$

$$\text{(3.7)} \qquad = \frac{1}{a_n}\Bigg[\int_0^\infty \bar F^*_{u_n}(y)\,\psi_{T(\bar F^*_{u_n})}(y)\;dy \;-\; \int_0^\infty \bar G(y)\,\psi_{T(\bar G)}(y)\;dy$$
$$-\;\Big(0,\ \frac{1}{T_2(\bar F^*_{u_n})}\Big)^T + \Big(0,1\Big)^T\Bigg]\ .$$

We know that $\lim_{n\to\infty} T(\bar F^*_{u_n}) = T(\bar G) = (\gamma^\star, 1)$. For each y, we use a Taylor expansion of order 2 with integral remainder of $\psi_\theta(y)$ around $\psi_{T(\bar G)}(y)$:

$$\text{(3.8)} \qquad \psi_{T(\bar F^*_{u_n})}(y) \;=\; \psi_{T(\bar G)}(y) + D\psi_\theta(y)\big|_{\theta=(\gamma^\star,1)}\big[T(\bar F^*_{u_n}) - T(\bar G)\big] \;+\; \text{remainder}\,.$$

Recall the principle of Taylor expansions of order 2 with integral remainder. Let f be a function of two variables. Denoting $g(t) = f(a_1 + th_1,\ a_2 + th_2)$ and using

$$g(1) \;=\; g(0) \;+\; g'(0) \;+ \int_0^1 (1-t)\,g''(t)\;dt\ ,$$

it follows that

$$f\big(a_1 + h_1,\ a_2 + h_2\big) \;=$$
$$= f(a_1,a_2) \;+\; h_1\,\partial_1 f(a_1,a_2) \;+\; h_2\,\partial_2 f(a_1,a_2)$$
$$+ \int_0^1 (1-t)\,\big(h_1^2\,\partial_{11} f + 2\,h_1\,h_2\,\partial_{12} f + h_2^2\,\partial_{22} f\big)\,(a_1 + th_1,\ a_2 + th_2)\;dt\ .$$

We will apply this formula to the two functions $\psi_j(x;\gamma,\sigma)$, $j=1,2$, at a fixed x, as functions of (γ,σ). The point (a_1,a_2) will be $(\gamma^\star,1)$ and the point (a_1+h_1, a_2+h_2) will be (γ,σ) close to $(\gamma^\star,1)$. Denote by $\Delta\gamma := \gamma - \gamma^\star$ and $\Delta\sigma := \sigma - 1$. We have ($\partial_1$ denoting the derivative in γ, and ∂_2 the derivative in σ)

$$\partial_1\psi_1(x;\,\gamma,\sigma) \;=\; -\,\frac{2x\,(x-\sigma)}{(\sigma+\gamma x)^3} \qquad \text{and} \qquad \partial_2\psi_1(x;\,\gamma,\sigma) \;=\; -\,\frac{(\gamma+2)x-\sigma}{(\sigma+\gamma x)^3}\ ,$$

$$\partial_1\psi_2(x;\,\gamma,\sigma) \;=\; -\,\frac{(\gamma+2)x-\sigma}{(\sigma+\gamma x)^3} \qquad \text{and} \qquad \partial_2\psi_2(x;\,\gamma,\sigma) \;=\; -\,\frac{2(1+\gamma)}{(\sigma+\gamma x)^3}\ .$$

We have similar expressions for the second order partial derivatives. As x tends to infinity, there are all of order $\mathcal{O}(1)$. For example,

$$\partial_{11}\psi_1(x;\,\gamma,\sigma) \;=\; \frac{6x^2(x-\sigma)}{(\sigma+\gamma x)^4}\ .$$

Thus

$$(\Delta\gamma)^2\int_0^1 (1-t)\,\partial_{11}\psi_1\big(x;\,\gamma^\star+t\,\Delta\gamma,\,1+t\,\Delta\sigma\big)\ dt \;=$$

$$=\;(\Delta\gamma)^2\int_0^1 (1-t)\,\frac{6x^2(x-1-t\,\Delta\sigma)}{\big[\,1+\gamma^\star x+(\Delta\sigma+x\,\Delta\gamma)\,t\,\big]^4}\ dt\ .$$

We are interested in

$$\int_0^\infty \bar F^*_{u_n}(y)\,\psi_{T(\bar F^*_{u_n})}(y)\ dy\ -\int_0^\infty \bar G(y)\,\psi_{T(\bar G)}(y)\ dy\ .$$

In the development $\psi_{T(\bar F^*_{u_n})}(y)-\psi_{T(\bar G)}(y)$, the contribution to the integral remainder due to $\partial_{11}\psi_1$ is of the form

$$(\Delta\gamma)^2\int_0^\infty \bar F^*_{u_n}(y)\,\int_0^1 (1-t)\,\frac{6y^2(y-1-t\,\Delta\sigma)}{\big[\,1+\gamma^\star y+(\Delta\sigma+y\,\Delta\gamma)\,t\,\big]^4}\ dt\,dy\ ,$$

which is

$$(\Delta\gamma)^2\int_0^1 (1-t)\,\int_0^\infty \bar F^*_{u_n}(y)\,\frac{6y^2(y-1-t\,\Delta\sigma)}{\big[\,1+\gamma^\star y+(\Delta\sigma+y\,\Delta\gamma)\,t\,\big]^4}\ dy\,dt\ .$$

We now show that this term is $O((\Delta\gamma)^2)$. The range y small is trivial, since the dominating term in the ratio is $-6y^2$. Therefore, we will look only at the range $y\ge y_0>0$. We have therefore to study

$$\int_0^1 (1-t)\,\int_{y_0}^\infty \bar F^*_{u_n}(y)\,\frac{6\big(1-(1+t\,\Delta\sigma)/y\big)}{y\big[\,1/y+\gamma^\star+(\Delta\sigma/y+\Delta\gamma)\,t\,\big]^4}\ dy\,dt\ ,$$

which can be reduced to study the quantity

$$\int_0^1 (1-t)\,\int_{y_0}^\infty \frac{\bar F^*_{u_n}(y)}{y}\ dy\,dt\ ,$$

which is integrable under our conditions and converges to $\int_0^1(1-t)\int_{y_0}^\infty(\bar G(y)/y)\,dy\,dt$. It is thus a bounded sequence.

From (3.7) and (3.8), we obtain that

$$\left(I(\theta^\star, \bar F^*_{u_n}) - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{T_2(\bar F^*_{u_n})} - 1 \end{bmatrix}\right) \left[\frac{T(\bar F^*_{u_n}) - T(\bar G)}{a_n}\right] =$$

$$= \int_0^\infty \frac{\bar F^*_{u_n}(y) - \bar G(y)}{a_n}\, \psi_{\theta^\star}(y)\, dy \;+\; \text{remainder} .$$

Using the Appendix 4.3, we have

$$\int_0^\infty \frac{\bar F^*_{u_n}(y) - \bar G(y)}{a_n}\, \psi_{\theta^\star}(y)\, dy \underset{n\to\infty}{\longrightarrow} \int_0^\infty D_{\gamma^\star,\rho}(y)\, \psi_{\theta^\star}(y)\, dy .$$

Moreover, $T_2(\bar F^*_{u_n}) \to 1$ and we have established that $I(\theta^\star, \bar F^*_{u_n}) \to I(\theta^\star, \bar G)$ which is definite positive, thus

$$I(\theta^\star, \bar F^*_{u_n}) - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{T_2(\bar F^*_{u_n})} - 1 \end{bmatrix}$$

is inversible for all n sufficiently large. We conclude therefore that

$$\begin{aligned} \frac{T(\bar F^*_{u_n}) - T(\bar G)}{a_n} &= \left[I(\theta^\star, \bar F^*_{u_n}) - \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{T_2(\bar F^*_{u_n})} - 1 \end{bmatrix}\right]^{-1} \\ &\qquad \cdot \int_0^\infty \frac{\bar F^*_{u_n}(y) - \bar G(y)}{a_n}\, \psi_{\theta^\star}(y)\, dy \;+\; \text{remainder} \quad\longrightarrow \\ &\longrightarrow\quad \Big[I\big(T(\bar G), \bar G\big)\Big]^{-1} \int_0^\infty D_{\gamma^\star,\rho}(y)\, \psi_{\theta^\star}(y)\, dy . \end{aligned} \tag{3.9}$$

Therefore (3.6) is now established for $\bar F^*_{u_n}$ and a similar study is carried out for sequences $\big(\bar F^*_{u_n} + k_n^{-1/2} \alpha_{k_n}(\bar F^*_{u_n})\big)_{n\ge1}$ in Appendices 4.4 and 4.5. □

We have now to compute the couple of biases, to check that the principal term of the functional bias of approximation cancels for $\rho = 0$, and to compute it for $\rho \neq 0$. We know that

$$I^{-1} := \Big[I\big(T(\bar G), \bar G\big)\Big]^{-1} = (\gamma+1) \begin{bmatrix} \gamma+1 & -1 \\ -1 & 2 \end{bmatrix} .$$

The column 2×1 of biases (first the bias concerning γ, then the bias concerning σ close to 1, i.e. for σ/σ_n), is

$$\text{(3.10)} \qquad \frac{a_n}{\bigl(1+|\rho|\bigr)\,\bigl(1+\gamma+|\rho|\bigr)} \begin{bmatrix} \gamma+1 \\ |\rho| \end{bmatrix} .$$

This formula is also applicable, at least formally, in the case $\gamma = 0$. When we choose $\rho = 0$, we find $a_n\,(1,0)^T$, as expected.

Remark 1. If we consider the Hall model (1982) defined by

$$1 - F(t) \;=\; C\,t^{-1/\gamma}\Bigl(1 + D\,t^{\rho/\gamma}\bigl(1+o(1)\bigr)\Bigr) \;, \qquad t \to \infty \;,$$

where $\rho < 0$, $C > 0$ and $D \in \mathbb{R}$, direct computations lead to $A(t) = \rho\,D\,C^\rho(\gamma+\rho)\cdot t^\rho(1+o(1))$ and so $a_n = \rho\,D\,(\gamma+\rho)\,u_n^{\rho/\gamma}(1+o(1))$. The vector of biases (3.10) is the one given in Smith (1987). Remark that this verification is direct for the bias of γ. On the other hand, it is not the case for the bias of σ/σ_n, taking into account the fact that Smith (1987) took a σ_n different from ours. It is thus necessary to take this difference into account. In the same way, Drees *et al.* (2004) have also, but in a different way, obtained the vector of biases for a standardization different from ours and from that of Smith (1987), but as previously mentioned a repercussion of this difference gives again (3.10).

Remark 2. We have just shown that the maximum likelihood estimators satisfy conditions (A_1)–(A_3) with $\mu_{k,\gamma}(dx) = \psi_k(x;\gamma,1)\,dx$, $k = 1,2$.

Since $(\gamma_n^b, \sigma_n^b) = T(\bar F_{u_n})$, we have $(\gamma_n^b, \frac{\sigma_n^b}{\sigma_n}) = T(\bar F^*_{u_n})$. Then, by (3.6) and (3.7),

$$\begin{aligned} T(\bar F^*_{u_n}) - T(\bar G_{\gamma,1}) \;&=\; \left(\gamma_n^b - \gamma,\, \frac{\sigma_n^b}{\sigma_n} - 1\right) \\ &=\; a_n\,I^{-1}\int_0^\infty D_{\gamma,\rho}(x)\,\psi_{\gamma,1}(x)\;dx \;+\; o(a_n) \\ &=\; \frac{a_n}{\bigl(1+|\rho|\bigr)\,\bigl(1+\gamma+|\rho|\bigr)} \begin{bmatrix} \gamma+1 \\ |\rho| \end{bmatrix} + \; o(a_n) \;. \end{aligned}$$

It follows that

$$\begin{aligned} &\bar F^*_{u_n}(x) - \bar G_{\gamma_n^b}\!\left(\frac{\sigma_n^b\,x}{\sigma_n}\right) \;= \\ &\quad= \; a_n\left(D_{\gamma,\rho}(x) + \frac{\gamma+1}{(1{+}|\rho|)\,(1{+}\gamma{+}|\rho|)}\,\frac{\partial G_\gamma}{\partial\gamma}(x) - \frac{|\rho|\,x}{(1{+}|\rho|)\,(1{+}\gamma{+}|\rho|)}\,\frac{\partial G_\gamma}{\partial x}(x)\right) \\ &\qquad + \; o(a_n) \;. \end{aligned}$$

Therefore, it can be directly checked that, when $\rho{=}0$, $A_E(x){=}0$ for all x.

4. APPENDIX

4.1. Convergence of the information matrices for $\bar{F}^*_{u_n}$

We use the following notations in all the sequel: $\bar{G} := \bar{G}_{\theta^\star}$ and $\Psi(\cdot\,|\theta) := D\psi_\theta(x)$. We would like to prove that, $\forall\,\varepsilon > 0$,

$$\left| \int_0^\infty \bar{G}(y)\, \Psi(y|\theta)\ dy - \int_0^\infty \bar{F}^*_{u_n}(y)\, \Psi(y|\theta)\ dy \right| \ \le\ \varepsilon\ ,$$

for n sufficiently large, $\forall\,\theta \in \mathcal{V}^\star = \mathcal{V}(\theta^\star)$.

This proof can be divided into two parts. First, we consider an $\varepsilon > 0$ and we will prove that it is possible to choose a number A such that each quantity

$$\left| \int_A^\infty \bar{G}(y)\, \Psi(y|\theta)\ dy \right| \qquad \text{and} \qquad \left| \int_A^\infty \bar{F}^*_{u_n}(y)\, \Psi(y|\theta)\ dy \right|$$

is smaller than $\varepsilon/2$, $\forall\,\theta \in \mathcal{V}^\star$ and $\forall\, n \ge n_1(A)$.

Then, using this number A, we will establish our result by reasoning on

$$\int_0^A \bar{F}^*_{u_n}(y)\, \Psi(y|\theta)\ dy$$

and using the mean value theorem.

In order to establish the first part of this result, we use the classical change of variables

$$\left| \int_A^\infty \bar{G}(y)\, \Psi(y|\theta)\ dy \right| \ =\ \left| \int_B^\infty e^{-s}\, \Psi\big(g(s)\big|\theta\big)\, g'(s)\ ds \right|$$

and

$$\left| \int_A^\infty \bar{F}^*_{u_n}(y)\, \Psi(y|\theta)\ dy \right| \ =\ \left| \int_{B_n}^\infty e^{-s}\, \Psi\big(g_n(s)\big|\theta\big)\, g_n'(s)\ ds \right|\ , \tag{4.1}$$

where $g_n(s) = (e^{\gamma^\star s} - 1)/\gamma^\star + a_n R_n(s)$ and $g(s) = (e^{\gamma^\star s} - 1)/\gamma^\star$, and we prove that the latter quantity is smaller than $\varepsilon/2$ for an $\varepsilon = \varepsilon(B_n)$, uniformly in $\theta = (\gamma, \sigma)$ close to $\theta^\star = (\gamma^\star, 1)$, and for n sufficiently large. Remark that the first quantity can be treated similarly and recall that $g_n'(s)$ is of the form $e^{\gamma^\star s} + a_n r_n(s)$, an expression that we will use later on. All the proof will be done in the case $\gamma^\star > 0$ and we will use the first component of $\Psi(\cdot|\theta)$, the other ones can be treated similarly (and in fact more easily). Thus, we have to bound the quantity

$$\frac{g_n(s)\, |g_n(s) - \sigma|}{\big(\sigma + \gamma g_n(s)\big)^3}\ |g_n'(s)| \qquad \text{for} \quad s \ge B_n\ .$$

Recall that $y = \bar F^{*\,-1}_{u_n}(\mathrm{e}^{-s})$. Since $\bar F^{*\,-1}_{u_n}$ and $s \mapsto \mathrm{e}^{-s}$ are decreasing, y is increasing. Consequently, since $g_n(s) \nearrow \infty$ for $s \nearrow \infty$ $\forall\, n$, we can use the bound

$$\mathrm{cst}\, \frac{g_n^2(s)}{g_n^3(s)}\; |g_n'(s)|$$

uniformly for $\theta \in \mathcal{V}^\star$. We have therefore to study

$$\int_{B_n}^{\infty} e^{-s}\, \frac{|g_n'(s)|}{g_n(s)}\; ds \;=\; \int_{B_n}^{\infty} e^{-s}\, \frac{g_n'(s)}{g_n(s)}\; ds \tag{4.2}$$

since $g_n' \geq 0$. By integrating by parts, this integral is equal to

$$-\, e^{-B_n} \ln\bigl(g_n(B_n)\bigr) \;+\; \int_{B_n}^{\infty} e^{-s} \ln\bigl(g_n(s)\bigr)\; ds\;. \tag{4.3}$$

Using the fact that the first term is negative and that if $s \geq B_n$ then $g_n(s) \geq g_n(B_n)$, we obtain the first part of the proof.

Concerning the second part of the proof, we first prove that

$$\int_0^A \bar F_n^*(y)\, \Psi(y|\theta)\; dy \;\longrightarrow\; \int_0^A \bar G(y)\, \Psi(y|\theta)\; dy$$

as $n \to \infty$, uniformly in θ close to $\theta^\star = (\gamma^\star, 1)$. We have

$$\int_0^A \bar F_n^*(y)\, \Psi(y|\theta)\; dy \;=\; \int_0^{B_n} e^{-s}\, \Psi\bigl(g_n(s)\big|\theta\bigr)\, g_n'(s)\; ds$$

and

$$\int_0^A \bar G(y)\, \Psi(y|\theta)\; dy \;=\; \int_0^{B} e^{-(1-\gamma^\star)s}\, \Psi\bigl(g(s)\big|\theta\bigr)\; ds\;.$$

Thus we look at

$$\begin{aligned}
\sup_{\theta\in\mathcal{V}^\star} &\left| \int_0^A \bar F_n^*(y)\, \Psi(y|\theta)\; dy - \int_0^A \bar G(y)\, \Psi(y|\theta)\; dy \right| \;= \\
&= \sup_{\theta\in\mathcal{V}^\star} \left| \int_0^B e^{-(1-\gamma^\star)s} \Bigl[\Psi\bigl(g_n(s)\big|\theta\bigr) - \Psi\bigl(g(s)\big|\theta\bigr)\Bigr]\, ds \right.\\
&\qquad\qquad \left. + \, a_n \int_0^B e^{-s}\, r_n(s)\, \Psi\bigl(g_n(s)\big|\theta\bigr)\; ds \right| \\
&\quad + \sup_{\theta\in\mathcal{V}^\star} \left| \int_B^{B_n} e^{-s}\, \Psi\bigl(g_n(s)\big|\theta\bigr)\, g_n'(s)\; ds \right| \\
&= \sup_{\theta\in\mathcal{V}^\star} \left| \int_0^B e^{-(1-\gamma^\star)s}\, a_n\, R_n(s)\, \frac{\partial \Psi}{\partial x} \left(\frac{e^{\gamma^\star s} - 1}{\gamma^\star} + \xi_n(s) \,\middle|\, \theta \right) ds \right.\\
&\qquad\qquad \left. + \, a_n \int_0^B e^{-s}\, r_n(s)\, \Psi\bigl(g_n(s)\big|\theta\bigr)\; ds \right| \\
&\quad + \sup_{\theta\in\mathcal{V}^\star} \left| \int_B^{B_n} e^{-s}\, \Psi\bigl(g_n(s)\big|\theta\bigr)\, g_n'(s) \right| ,
\end{aligned}$$

where $\xi_n(s) \in (0, a_n R_n(s))$ and where $\mathcal{V}^\star$ is a compact neighbourhood of $\theta^\star$. Therefore, we have to study two supremum separately. We have some very useful properties (Potter bounds) that we recall below (see Worms, 2000):

- $$\frac{V\big(s + V^{-1}(u_n)\big) - u_n}{\sigma_n} \;=\; \frac{e^{\gamma^\star s} - 1}{\gamma^\star} + a_n R_n(s) \;, \qquad a_n \in \mathbb{R} \;,$$
  $$|R_n(s)| \;\le\; \text{cst}\; e^{(\gamma^\star + \eta)s} \int_0^s e^{\rho t}\, dt \;,$$
  $$R_n(s) \underset{n\to\infty}{\longrightarrow} \int_0^s e^{\gamma^\star z} \int_0^z e^{\rho t}\, dt\, dz \;.$$

- $$\frac{V'\big(s + V^{-1}(u_n)\big)}{\sigma_n} \;=\; e^{\gamma^\star s} + a_n r_n(s) \;,$$
  $$|r_n(s)| \;\le\; \text{cst}\; e^{(\gamma^\star + \eta)s} \int_0^s e^{\rho t}\, dt \;,$$
  $$r_n(s) \underset{n\to\infty}{\longrightarrow} e^{\gamma^\star s} \int_0^s e^{\rho t}\, dt \;.$$

For the first supremum, we study the two quantities of the sum separately. Concerning the first one and taking into account the fact that

$$\sup_{\theta \in \mathcal{V}^\star,\; x \in [0,A]} \big|\Psi(x|\theta)\big| \;<\; \infty \qquad \text{and} \qquad \sup_{\theta \in \mathcal{V}^\star,\; x \in [0,A]} \left|\frac{\partial \Psi}{\partial x}(x|\theta)\right| \;<\; \infty \;,$$

we can restrict ourself to a compact in x, i.e. work with the product of the two compacts. This is what we do below, where we have to study

$$\left| e^{-(1-\gamma^\star)s}\, a_n\, R_n(s)\, \frac{\partial \Psi}{\partial x}\left(\frac{e^{\gamma^\star s} - 1}{\gamma^\star} + \xi_n(s) \,\middle|\, \theta \right) \right| \;\le$$
$$\le\; \text{cst}\; |a_n|\, e^{-(1 - 2\gamma^\star - \eta)s} \sup_{\theta \in \mathcal{V}^\star,\; x \in [0,A]} \left| \frac{\partial \Psi}{\partial x}(x|\theta) \right| \;,$$

taking the fact that $|R_n(s)| \le \mathrm{K}\, e^{(\gamma^\star + \eta)s}$ into account. Consequently, the first quantity in the first supremum tends to 0 uniformly.

Concerning now the second quantity in the first supremum,

$$|a_n| \left| \int_0^B e^{-s}\, r_n(s)\, \Psi\big(g_n(s) \big| \theta\big)\; ds \right| \;,$$

we use again the fact that $|r_n(s)| \le \mathrm{K}\, e^{(\gamma^\star + \eta)s}$ and $\sup_{\theta \in \mathcal{V}^\star,\; x \in [0,A]} |\Psi(x|\theta)| < \infty$ in order to conclude.

The second supremum, related to

$$\left| \int_B^{B_n} e^{-s} \Psi\big(g_n(s) \big| \theta\big)\, g_n'(s)\; ds \right| ,$$

can be bounded by

$$\int_B^{B_n} e^{-s} \Big| \Psi\big(C_1\, e^{(\gamma^\star - \eta)s} \,\big|\, \theta\big) \Big|\, C_2\, e^{(\gamma^\star + \eta)s}\; ds \;=\; \mathcal{O}\bigg(\int_B^{B_n} e^{-(1-2\eta)s}\, ds \bigg) ,$$

where C_1 and C_2 are two constants. This last equality comes from the fact that B and B_n are large, then $\Psi(\cdot|\theta)$ is decreasing and $\Psi\big(C_1\, e^{(\gamma^\star-\eta)s} \,\big|\, \theta\big)$ is of order $\mathcal{O}\big(e^{-(\gamma^\star-\eta)s}\big)$ uniformly, which achieves the proof.

4.2. Conditions of integrability for $\bar{F}^*_{u_n}$

We would like to show that (C_3) is satisfied by $\bar{F}^*_{u_n}$. With this aim, let $e^{-s} = \bar{F}^*_{u_n}(y)$ which implies that $y = \bar{F}^{*\,-1}_{u_n}(e^{-s}) = g_n(s)$. Using the Potter bounds, we obtain that

$$\mathrm{cst}\, e^{(\gamma^\star - \eta)s} \;\le\; \bar{F}^{*\,-1}_{u_n}(e^{-s}) \;\le\; \mathrm{cst}\, e^{(\gamma^\star + \eta)s} \;.$$

Since $\bar{F}^*_{u_n}$ is decreasing, we deduce that

$$\bar{F}^*_{u_n}\big(\mathrm{cst}\, e^{(\gamma^\star + \eta)s}\big) \;\le\; e^{-s} \;\le\; \bar{F}^*_{u_n}\big(\mathrm{cst}\, e^{(\gamma^\star - \eta)s}\big) \;.$$

Let $y = \mathrm{cst}\, e^{(\gamma^\star \pm \eta)s}$ according to what we want to obtain. We have then $e^{-s} = (y/\mathrm{cst})^{-1/(\gamma^\star \pm \eta)}$, which implies that $\bar{F}^*_{u_n}(y) = \mathcal{O}(y^{-\beta})$ for some $\beta > 0$. The integrability condition (C_3) is then clearly satisfied by $\bar{F}^*_{u_n}$.

4.3. Hadamard differentiability of $\bar{F}^*_{u_n}$

Our aim in this appendix is to deal with

$$\int_0^\infty \frac{\bar{F}^*_{u_n}(y) - \bar{G}(y)}{a_n}\, \psi_{\theta^\star}(y)\; dy \;=$$

$$= \int_0^\infty e^{-s}\, \psi_{\theta^\star}(g_n(s))\, r_n(s)\; ds \;+ \int_0^\infty e^{-(1-\gamma^\star)s}\, \frac{\psi_{\theta^\star}(g_n(s)) - \psi_{\theta^\star}(g(s))}{a_n}\; ds$$

with $|r_n(s)| \le \mathrm{cst}\, e^{(\gamma^\star + \eta)s}$.

We will use the fact that all the functions of the matrix $\psi_{\theta^\star}(x)$ are continuous and of order $\mathcal{O}\big(\frac{1}{x}\big)$ as $x \to \infty$. Denoting by ϕ such a function, we will first prove that

$$\int_0^\infty e^{-s}\,\phi(g_n(s))\,r_n(s)\;ds \underset{n\to\infty}{\longrightarrow} \int_0^\infty e^{-(1-\gamma^\star)s}\int_0^s e^{\rho t}\,dt\;\;\phi(g(s))\;ds\;.$$

With this aim, we will split the integral into two parts: the first one from 0 to A and the second one from A to infinity. The first part does not pose any problem. Therefore, we will look at the second one.

Using the properties of ϕ, we have

$$\big|\phi(g_n(s))\big| \le \frac{\mathrm{cst}}{g_n(s)} \quad \text{for } s \ge A \quad \text{and for all } n \text{ sufficiently large}\,.$$

Using the lower Potter bound for $g_n(s)$, we have:

$$g_n(s) \ge \mathrm{cst}\,e^{(\gamma^\star-\eta)s}\;, \qquad \mathrm{cst} > 0, \quad \text{for } s \ge A \;\; \text{and for all } n \text{ sufficiently large}\,.$$

Therefore, for the same s and n, we have

$$\big|\phi(g_n(s))\big| \le \mathrm{cst}\,e^{(-\gamma^\star+\eta)s}\;,$$

which implies the domination of the function in the integral by $\mathrm{cst}\,e^{-s(1-2\eta)}$.

We thus obtain the desired convergence using the fact that ϕ is continuous and that $r_n(s) \to e^{\gamma^\star s}\big(\int_0^s e^{\rho t}\,dt\big)$ as $n \to \infty$.

Now we have to study

$$\int_0^\infty e^{-(1-\gamma^\star)s}\;\frac{\phi(g_n(s)) - \phi(g(s))}{a_n}\;ds\;.$$

We have

$$\begin{aligned}\phi(g_n(s)) - \phi(g(s)) \;&=\; \phi\Big(g(s) + a_n R_n(s)\Big) - \phi(g(s))\\ &=\; a_n\,\phi'\Big(g(s) + \xi_n(s)\,a_n\,R_n(s)\Big)R_n(s)\;, \qquad 0 \le \xi_n(s) \le 1\;.\end{aligned}$$

We can use the Potter bound for $g(s) + \xi_n(s)\,a_n R_n(s)$. Therefore

$$\Big|\phi'\Big(g(s) + \xi_n(s)\,a_n\,R_n(s)\Big)\Big| \le\; \mathrm{cst}\,e^{(-2\gamma^\star+2\eta)s}\;.$$

Recall that for R_n, we have the following bound

$$|R_n(s)| \le\; \mathrm{cst}\,e^{(\gamma^\star+\eta)s}\;.$$

By gathering these various results, we obtain that the function in the integral is bounded by $\operatorname{cst} e^{-s(1-3\eta)}$.

Moreover, ϕ' is continuous. Therefore, since $\forall\, s,\ g_n(s) \to g(s)$ as $n \to \infty$ and $g(s) + \xi_n(s)\, a_n R_n(s)$ is located between $g(s)$ and $g_n(s)$, we have

$$\phi'\Big(g(s) + \xi_n(s)\, a_n\, R_n(s)\Big) \ \longrightarrow\ \phi'(g(s))\ .$$

Finally, $R_n(s) \to \int_0^s e^{\gamma^\star u} \int_0^u e^{\rho t}\, dt\, du$.

By gathering the two limiting integrals, we obtain

$$\int_0^\infty e^{-(1-\gamma^\star)s}\Big(\int_0^s e^{\rho t}\, dt\Big)\, \phi(g(s))\ ds\ + \\ +\ \int_0^\infty e^{-(1-\gamma^\star)s}\Big(\int_0^s e^{\gamma^\star z} \int_0^z e^{\rho t}\ dt\, dz\Big)\, \phi'(g(s))\ ds\ \ =:\ \ A+B\ .$$

By integrating by parts, we derive the following limit

$$B\ =\ -\int_0^\infty \phi(g(s))\, e^{-s}\Big(-\int_0^s e^{\gamma^\star z} \int_0^z e^{\rho t}\ dt\, dz\ +\ e^{\gamma^\star s}\int_0^s e^{\rho t}\, dt\Big)\, ds\ ,$$

which implies that

$$A+B\ =\ \int_0^\infty \phi(g(s))\, e^{-s} \int_0^s e^{\gamma^\star z} \int_0^z e^{\rho t}\ dt\, dz\, ds\ .$$

Now, using the notations of the Introduction, it follows that

$$\begin{aligned} A+B\ &=\ \int_0^\infty \phi(g(s))\, e^{-s} I_{\gamma^\star,\rho}(s)\ ds \\ &=\ \int_0^\infty \phi(g(s))\, e^{\gamma^\star s}\, C_{\gamma^\star,\rho}(s)\ ds \\ &=\ \int_0^\infty \phi(y)\, D_{\gamma^\star,\rho}(y)\ dy\ , \end{aligned}$$

by the change of variable $y = g(s)$.

4.4. Existence and almost surely unicity of $\hat\theta_n$ for $\bar F^*_{u_n} + \frac{1}{\sqrt{k_n}}\,\alpha_{k_n}\big(\bar F^*_{u_n}\big)$

According to the proof for $\bar F^*_{u_n}$, it is sufficient to establish that

$$\frac{1}{\sqrt{k_n}} \int_0^\infty \alpha_{k_n}\big(\bar F^*_{u_n}(x)\big)\, D\psi_\theta(x)\ dx\ \ \underset{n\to\infty}{\longrightarrow}\ \ 0 \qquad \text{a.s.}\ ,$$

uniformly in $\theta \in \mathcal{V}^\star$. We use the fact that $D\psi_\theta(x)$ is bounded uniformly in $\theta \in \mathcal{V}^\star$. We split the integral, as in the proof for $\bar{F}^*_{u_n}$, into two integrals, one from 0 to A and the other one from A to infinity.

Concerning the integral from A to infinity, we use the classical change of variable, leading to (see (3.9))

$$\frac{1}{\sqrt{k_n}} \int_{B_n}^{\infty} \big|\alpha_{k_n}(\mathrm{e}^{-s})\big| \, \frac{g_n'(s)}{g_n(s)} \; ds \; .$$

We use Mason's theorem (1981) (see Shorack and Wellner, 1986, p. 425) which implies that for a small $\varepsilon > 0$, we have

$$\frac{1}{\sqrt{k_n}} \, |\alpha_{k_n}(\mathrm{e}^{-s})| \;=\; \frac{(\ln k_n)^{\frac{1}{2}+\varepsilon}}{\sqrt{k_n}} \; \mathcal{O}\big(\mathrm{e}^{-\frac{s}{2}}\big) \qquad \text{a.s.}$$

when $s \to \infty$, as $k_n \to \infty$. To conclude, we apply to the term $\mathcal{O}\big(\mathrm{e}^{-\frac{s}{2}}\big)$ an integration by parts technique similar to the one used in (3.10).

Concerning the integral from 0 to A, Mason's theorem (1981) implies that

$$\frac{1}{\sqrt{k_n}} \, \big|\alpha_{k_n}\big(\bar{F}^*_{u_n}(x)\big)\big| \;=\; \frac{(\ln k_n)^{\frac{1}{2}+\varepsilon}}{\sqrt{k_n}} \; \mathcal{O}(1) \qquad \text{a.s.}$$

for $0 \le x \le A$, as $n \to \infty$.

4.5. Hadamard differentiability of $\bar{F}^*_{u_n} + \frac{1}{\sqrt{k_n}}\,\alpha_{k_n}\big(\bar{F}^*_{u_n}\big)$

We must check that the condition of integrability (C_3) is satisfied. We showed that $\bar{F}^*_{u_n}(y) = \mathcal{O}(y^{-\beta})$ when $y \to \infty$, uniformly for n sufficiently large, and for a constant $\beta > 0$. We thus deduce, via Mason (1981), that

$$\frac{1}{\sqrt{k_n}} \, \big|\alpha_{k_n}\big(\bar{F}^*_{u_n}(y)\big)\big| \;=\; \frac{(\ln k_n)^{\frac{1}{2}+\varepsilon}}{\sqrt{k_n}} \; O\big(y^{-\frac{\beta}{2}}\big) \qquad \text{a.s.}$$

as $y \to \infty$, uniformly for n sufficiently large.

Now, we will show that there exists some versions $\widetilde{\alpha}_{k_n}$ of α_{k_n} and a brownian bridge $\mathbb{B}$ on $[0,1]$ such that

$$\int_0^\infty \widetilde{\alpha}_{k_n}\big(\bar{F}^*_{u_n}(x)\big)\, \psi_{\theta^\star}(x) \; dx \;=\; \int_0^\infty \mathbb{B}\big(\bar{G}(x)\big)\, \psi_{\theta^\star}(x) \; dx \;+\; o_P(1)$$

as $n \to \infty$. With this aim, we will use the fact that $|\psi_{\theta^\star}(x)| = \mathcal{O}(1/x)$ as $x \to \infty$, and the invariance principle for the weighted empirical process given in Einmahl and Mason (1992). We again split the initial integral into an integral from 0 to A and an integral from A to infinity, and we carry out the usual change of variable.

We start with $\int_{B_n}^{\infty} \alpha_{k_n}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\, ds$. Since $|\psi_{\theta^\star}(x)| = \mathcal{O}(1/x)$ as $x \to \infty$, this integral is of order $\mathcal{O}\Big(\int_{B_n}^{\infty} |\alpha_{k_n}(\mathrm{e}^{-s})|\, \frac{g_n'(s)}{g_n(s)}\, ds\Big)$. Here, we will change α_{k_n} into $\widetilde{\alpha}_{k_n}$ (Einmahl and Mason, 1992) for $B_n \le s \le \ln k_n$, with an error term of order $\mathcal{O}_P\big(k_n^{-\nu}\, \mathrm{e}^{-(\frac{1}{2}-\nu)s}\big)$ for $0 \le \nu < 1/4$. We obtain therefore

$$\begin{aligned}
&\int_{B_n}^{\ln k_n} \widetilde{\alpha}_{k_n}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\; ds \;= \\
&\qquad = \int_{B_n}^{\ln k_n} \mathbb{B}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\; ds \\
&\qquad\quad + \mathcal{O}_P\left(k_n^{-\nu} \int_{B_n}^{\ln k_n} \mathrm{e}^{-(\frac{1}{2}-\nu)s}\, \big|\psi_{\theta^\star}(g_n(s))\big|\, g_n'(s)\; ds\right)
\end{aligned} \tag{4.4}$$

as $n \to \infty$. The error term is

$$\mathcal{O}_P\left(k_n^{-\nu} \int_{B_n}^{\ln k_n} \mathrm{e}^{-(\frac{1}{2}-\nu)s}\, \frac{g_n'(s)}{g_n(s)}\; ds\right),$$

and we conclude that it tends to 0 by integrating by parts.

Now, we will study the integral from $\ln k_n$ to infinity by using again the fact that $|\psi_{\theta^\star}(x)| = \mathcal{O}(1/x)$ as $x \to \infty$. According to Jaeschke's theorem (see Shorack and Wellner, 1986, p. 600), this integral is of order

$$\sqrt{\ln\ln k_n}\; \mathcal{O}_P\left(\int_{B_n}^{\infty} \mathrm{e}^{-s/2}\, \frac{g_n'(s)}{g_n(s)}\; ds\right), \tag{4.5}$$

and again by integrating by parts, the result follows.

Combining (4.1) with (4.2), we obtain that

$$\int_{B_n}^{\infty} \widetilde{\alpha}_{k_n}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\; ds \;=\; \int_{B_n}^{\ln k_n} \mathbb{B}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\; ds \;+\; o_P(1)$$

as $n \to \infty$.

Therefore, we only have to study $\int_0^{B_n} \alpha_{k_n}(\mathrm{e}^{-s})\, \psi_{\theta^\star}(g_n(s))\, g_n'(s)\, ds$. Again, we split into two integrals, one from 0 to $-\ln(1-\frac{1}{k_n})$ and the other from $-\ln(1-\frac{1}{k_n})$ to B_n. The first one is clearly $o_P(1)$. For the second one, we use the fact that $\psi_{\theta^\star}(x)$ is bounded for $0 \le x \le A$, and we replace $\widetilde{\alpha}_{k_n}(\mathrm{e}^{-s})$ by $\mathbb{B}(\mathrm{e}^{-s})$, which leads to an error term of order

$$\mathcal{O}_P\left(k_n^{-\nu} \int_{-\ln\left(1-\frac{1}{k_n}\right)}^{B_n} \mathrm{e}^{-(\frac{1}{2}-\nu)s}\, g_n'(s)\; ds\right)$$

as $n \to \infty$. Remark now that $(B_n)_{n\ge 1}$ is bounded. Indeed, the Potter bounds imply that $0 \le g_n(s) \le \mathrm{cst}\, e^{(\gamma^\star+\eta)s}$. If we note $s = g_n^{-1}(y)$, we have $0 \le y \le \mathrm{cst}\, e^{(\gamma^\star+\eta)s}$, so $s \le \frac{\ln(y/\mathrm{cst})}{\gamma^\star+\eta}$. This leads to $B_n \ge \frac{\ln(A/\mathrm{cst})}{\gamma^\star+\eta}$, with a similar result in the other side.

Consequently, using again the Potter bounds, but this time for $g'_n(s)$, and the fact that $(B_n)_{n\geq 1}$ is bounded, the preceding error term is $o_P(1)$.

Finally, we obtain the following result

$$\int_0^\infty \widetilde{\alpha}_{k_n}(\mathrm{e}^{-s})\,\psi_{\theta^\star}(g_n(s))\,g'_n(s)\;ds \;=\; = \int_{-\ln\left(1-\frac{1}{k_n}\right)}^{\ln k_n} \mathbb{B}(\mathrm{e}^{-s})\,\psi_{\theta^\star}(g_n(s))\,g'_n(s)\;ds \;+\; o_P(1)\;.$$

Now, we only have to let n tending to infinity in the integral of the second member. Since $\mathbb{B}$ is a.s. continuous, $\ln k_n \to \infty$, $-\ln(1-\frac{1}{k_n}) \to 0$, $g_n(s) \to g(s)$ and $g'_n(s) \to g'(s)$ for all s as $n \to \infty$, we only have to establish an a.s. domination for $\mathbb{B}(\mathrm{e}^{-s})\,|\psi_{\theta^\star}(g_n(s))|\,g'_n(s)$ in order to apply the Lebesgue dominated convergence theorem. We use again the fact that $|\psi_{\theta^\star}(g_n(s))| = \mathcal{O}(1/g_n(s))$ as $s \to \infty$ for n sufficiently large, and we conclude using the Potter bounds on $g_n(s)$ and $g'_n(s)$, and the law of iterated logarithm for $\mathbb{B}(t)$, $t \to 0$.

ACKNOWLEDGMENTS

We thank the referees for valuable comments.

REFERENCES

[1] Balkema, A. and de Haan, L. (1974). Residual life time at a great age, *Annals of Probability*, **2**, 792–801.

[2] Beirlant, J.; Dierckx, G.; Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes*, **2**(2), 177–200.

[3] Bingham, N.H.; Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*, Cambridge, U.K.: Cambridge University Press.

[4] Cohen, J.P. (1982). Convergence rates for the ultimate and penultimate approximations in extreme-value theory, *Advances in Applied Probability*, **14**, 833–854.

[5] Dekkers, A.L.M. and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation, *Annals of Statistics*, **17**, 1795–1832.

[6] Diebolt, J.; Guillou, A. and Worms, R. (2003). Asymptotic behaviour of the probability-weighted moments and penultimate approximation, *ESAIM, Probability and Statistics*, **7**, 219–238.

[7] Drees, H. (1998). On smooth statistical tail functionals, *Scandinavian Journal of Statistics*, **25**(1), 187–210.

[8] Drees, H.; Ferreira, A. and de Haan, L. (2004). On maximum likelihood estimation of the extreme value index, *Annals of Applied Probability*, **14**(3), 1179–1201.

[9] Einmahl, U. and Mason, D.M. (1992). Approximation to permutation and exchangeable processes, *Journal of Theoretical Probability*, **5**, 101–126.

[10] Galambos, J. (1978). *Asymptotic Theory of Extreme Order Statistics*, Krieger, Malabar, Florida.

[11] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.

[12] Gomes, M.I. (1984). Penultimate limiting forms in extreme value theory, *Annals of the Institute of Statistical Mathematics*, **36**, 71–85.

[13] Gomes, I. and de Haan, L. (2000). Approximation by penultimate extreme value distributions, *Extremes*, **2**, 71–85.

[14] Gomes, M.I. and Pestana, D.D. (1987). *Non standard domains of attraction and rates of convergence*, In "New perspectives in theoretical and applied statistics", (Wiley Ser. Probab. Math. Statist. Probab. Math. Statist.), Wiley, New York, 467–477.

[15] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *Journal of the Royal Statistical Society Serie B*, **44**, 37–42.

[16] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

[17] Hosking, J. and Wallis, J. (1987). Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, **29**, 339–349.

[18] Leadbetter, M.R.; Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, Berlin.

[19] Mason, D.M. (1981). Bounds for weighted empirical distribution functions, *Annals of Probability*, **9**, 881–884.

[20] Pickands III, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics*, **3**, 119–131.

[21] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Processes with Applications to Statistics*, Wiley New York.

[22] Smith, R.L. (1987). Estimating tails of probability distributions, *Annals of Statistics*, **15**, 1174–1207.

[23] Worms, R. (2000). *Vitesses de convergence pour l'approximation des queues de distributions*, Thèse de doctorat de l'Université de Marne-la-Vallée.

[24] Worms, R. (2002). Penultimate approximation for the distribution of the excesses, *ESAIM, Probability and Statistics*, **6**, 21–31.

REVSTAT – Statistical Journal
Volume 3, Number 1, June 2005, 45–60

EXACT FORMULAS FOR THE MOMENTS OF THE FIRST PASSAGE TIME OF REWARD PROCESSES

Authors: G.A. Parham
– Department of Statistics, Shahid Chamran University, Ahwaz, Iran
(Parham_g@cua.ac.ir)

A.R. Soltani
– Department of Statistics, Shiraz University, Shiraz, Iran

Received: November 2004 Revised: January 2005 Accepted: January 2005

Abstract:

- Let $\{\mathcal{Z}_\rho(t),\ t \geq 0\}$ be a reward process based on a semi-Markov process $\{\mathcal{J}(t),\ t \geq 0\}$ and a reward function ρ. Let T_z be the first passage time of $\{\mathcal{Z}_\rho(t),\ t \geq 0\}$ from $\mathcal{Z}_\rho(0) = 0$ to a prespecified level z. In this article we provide the Laplace transform of the $E[T_z^k]$ and obtain the exact formulas for ET_z, ET_z^2 and $\operatorname{var}(T_z)$. Formulas for certain type I counter models are given.

Key-Words:

- *Semi-Markov process; reward process; Laplace transform; first passage time.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

Let $\{\mathcal{J}(t),\ t \geq 0\}$ be a semi-Markov process with a Markov renewal process $\{(\mathcal{J}_n, \mathcal{T}_n),\ n = 0, 1, 2, ...\}$. The state space of $\{\mathcal{J}_n\}$ is assumed to be $\mathcal{N} = \{0, 1, 2, ..., N\}$. A reward process is a certain functional that is defined on a semi-Markov process (Markov renewal process) by

$$\mathcal{Z}_\rho(t) \;=\; \sum_{n:\ \mathcal{T}_{n+1}<t} \rho(\mathcal{J}_n,\, \mathcal{T}_{n+1} - \mathcal{T}_n) \;+\; \rho(\mathcal{J}(t), X(t)) \;, \tag{1}$$

where $X(t)$ is the age process. The function ρ in (1) is a real function of two variables; $\rho : \mathcal{N}{\times}R \to R$, and $\rho(i,\tau)$ measures the excess reward when time τ is spent in the state i. The process $\mathcal{Z}_\rho(t)$ given by (1) provides the cumulative reward at time t, under the given reward function ρ. This process was introduced and studied in [4], for general ρ. For $\rho(i,\tau) = i\tau$, the reward process $\mathcal{Z}_\rho(t)$ has been treated by different authors, see [1] [2] [5]. Let T_z be the first passage time of $\mathcal{Z}_\rho(t)$ from $\mathcal{Z}_\rho(0) = 0$ to a prespecified level z. Asymptotic behaviors of ET_z, ET_z^2 as $z \to \infty$, were obtained in [5] for $\rho(i,x) = ix$, and in [3] for general ρ. In this article we provide exact formulas for ET_z, ET_z^2 and $\mathrm{var}(T_z)$, under general ρ. We apply our formulas to certain type I counter models and provide precise results. The main results are Theorems 2.1, 3.1, Corollary 3.1, Remark 3.1, and formulas (23), (24).

2. NOTATION AND PRELIMINARIES

Let $\{\mathcal{J}(t),\, t \geq 0\}$ be a semi-Markov process and $\{(\mathcal{J}_n, \mathcal{T}_n),\, n = 0, 1, 2, ...\}$ be a Markov renewal process, where $\mathcal{J}_n$ is a Markov chain in discrete time on state space $\mathcal{N} = \{0, 1, 2, ..., N\}$, and $\mathcal{T}_n$ is the n-th transition epoch with $\mathcal{T}_0 = 0$. The behavior of the Markov renewal process is governed by a semi-Markov matrix $A(x) = [A_{ij}(x)]$, where

$$A_{ij}(x) \;=\; P\Big\{\, \mathcal{J}_{n+1} = j,\ \mathcal{T}_{n+1} - \mathcal{T}_n \;\leq\; x \;\mid\; \mathcal{J}_n = i \Big\} \;. \tag{2}$$

We assume that the stochastic matrix $P = [P_{ij}] = A(\infty)$ governing the embedded Markov chain $\{\mathcal{J}_n : n = 0, 1, 2, ...\}$ is aperiodic and irreducible. For convenience let,

$$\begin{aligned} A_{k:ij} \;&=\; \int_0^\infty x^k A_{ij}(dx) \;, \\ A_{k:i} \;&=\; \int_0^\infty x^k A_i(dx) \;, \qquad k = 0, 1, 2, \ldots \;, \end{aligned} \tag{3}$$

if they exist, where

$$A_i(x) = \sum_{j\in\mathcal{N}} A_{ij} \;, \qquad \overline{A}_i(x) = 1 - A_i(x) \;.$$

We note that $A_i(x) = P\{\mathcal{T}_{n+1} - \mathcal{T}_n \leq x \mid \mathcal{J}_n = i\}$ is the cumulative distribution function of the dwell time of the semi-Markov process at state i, and $\overline{A}_i(x)$ is the corresponding survival function. Let $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. We define,

$$A_D(x) = [\delta_{ij} A_j(x)] \ , \qquad \overline{A}_D(x) = [\delta_{ij} \overline{A}_j(x)] \ , \tag{4}$$
$$A_k = [A_{k:ij}], \quad A_{D:k} = [\delta_{ij} A_{k:i}] \ , \qquad k = 0, 1, 2, \ldots \ .$$

Note that $A_{D:0} = I$. The Laplace–Stieltjes transform of $A(x)$ is denoted by

$$\alpha(s) = [\alpha_{ij}(s)] \ , \qquad \alpha_{ij}(s) = \int_0^\infty e^{-sx} A_{ij}(dx) \ , \tag{5}$$

Laplace–Stieltjes transforms $\alpha_i(s)$, $\alpha_D(s)$, etc. are defined similarly. We define n-fold convolution $A(x)$ by

$$A^{(n)}(x) \ = \ \int_0^x A(dx')\, A^{(n-1)}(x - x') \ ,$$

$$A_{jk}^{(0)}(t) \ = \ \begin{cases} 0 & \text{if } t < 0 \\ \delta_{jk} & \text{if } t \geq 0 \end{cases}$$

and

$$A_{jk}^{(n)}(t) \ = \ \begin{cases} 0 & \text{if } t < 0 \\ \displaystyle\sum_\nu \int_0^t A_{j\nu}(dy)\, A_{\nu k}^{(n-1)}(t - y) & \text{if } t \geq 0 \end{cases}$$

if M is a matrix of measures and N is a matrix of measurable functions, the convolution of M and N (written $M * N$) is defined by $M * N(t) = [(M * N)_{ik}(t)]$, where

$$M * N_{jk}(t) \ = \ \sum_\nu \int_0^t M_{j\nu}(dy)\, N_{\nu k}(t - y) \ .$$

Let $A(x)$ be a semi-Markov matrix. Then

$$\mathcal{A}(x) \ = \ \sum_{n=0}^{\infty} A^{(n)}(x)$$

is called the Markov renewal matrix corresponding to $A(x)$. Also denote the Laplace transform of the Markov renewal matrix by

$$\mathcal{L}_s[\mathcal{A}] = \frac{1}{s} \, [I - \alpha(s)]^{-1} \ .$$

The transition probability matrix of $J(t)$ is denoted by $P(t)$, i.e.,

$$P(t) = [P_{ij}(t)] \ , \qquad P_{ij}(t) = P\Big\{ J(t) = j \ \mid \ J(0) = i \Big\} \ . \tag{6}$$

The state probability vector at time t, $p'(t) = (\ p_0(t),\ p_1(t), ..., p_N(t)\ )$, is given by $p'(t) = p'(0)P(t)$, where $p'(0)$ is the initial probability vector. In this article $\underline{e}$ is the unit vector, i.e., $\underline{e} = (1, ..., 1)'$.

Let $X(t)$ be the age process, i.e., the time elapsed at time t since the last transition of $J(t)$, $X(t) = t - \mathcal{T}_n$, where $n = \sup\{m : \mathcal{T}_m \le t\}$. The joint distributions corresponding to the bivariate process $\{(\mathcal{J}(t), X(t)),\ t \ge 0\}$ and the trivariate process $\{(\mathcal{J}(t), X(t), \mathcal{Z}_\rho(t)),\ t \ge 0\}$, respectively, are given by

$$\tag{7} \begin{aligned} G_{ij}(x,t) &= P\Big\{\mathcal{J}(t) = j,\ X(t) \le x \mid \mathcal{J}(0) = i\Big\} ,\\ F_{ij}(x,z,t) &= P\Big\{\mathcal{J}(t) = j,\ X(t) \le x,\ \mathcal{Z}_\rho(t) \le z \mid \mathcal{J}(0) = i\Big\} . \end{aligned}$$

The Laplace transform of $F_{ij}(x,z,t)$ is denoted by

$$\tag{8} \phi_{ij}(v,\omega,s) = \int_0^\infty \int_{-\infty}^\infty \int_0^\infty e^{-vx-\omega z-st}\, F_{ij}(dx,dz,t)\, dt ,$$

in the matrix form $\phi(v,\omega,s) = [\phi_{ij}(v,\omega,s)]$. It is demonstrated in [4] that the following informative transform formula plays a crucial role in studying the statistical properties of the reward process (1), see also [5],

$$\tag{9} \phi(v,\omega,s) = [I - C(\omega,s)]^{-1}\, E_D(\omega, v+s) ,$$

where

$$\tag{10} \begin{aligned} C(w,s) &= [C_{kj}(\omega,s)] , \qquad C_{kj}(w,s) = \int_0^\infty e^{-\omega\rho(k,x)-sx} A_{kj}(dx) ,\\ E_D(\omega,s) &= [\delta_{kj}E_j(\omega,s)] , \qquad E_j(\omega,s) = \int_0^\infty e^{-\omega\rho(j,x)-sx}\, \overline{A}_j(x)\, dx . \end{aligned}$$

Let z be a given level, then the first passage time of the level z for $\mathcal{Z}_\rho(t)$, given $\mathcal{Z}_\rho(0) = 0$, is defined by

$$T_z = \inf\Big\{t > 0 : \mathcal{Z}_\rho(t) = z \mid \mathcal{Z}_\rho(0) = 0\Big\} .$$

Clearly

$$\tag{11} P\{T_z > t\} = P\{\mathcal{Z}_\rho(t) < z\} .$$

Let $H(z,t)$ be the distribution of T_z, and denote the Laplace transform of $E[e^{-sT_z}]$ by

$$\tag{12} \psi(\omega,s) = \int_0^\infty e^{-\omega z}\, E[e^{-sT_z}]\, dz .$$

Similarly we denote the Laplace transform of the survival function $\overline{H}(z,t) = P\{T_z > t\}$ by $\overline{\psi}(\omega,s)$. We recall from [5] that,

$$\tag{13} \psi(\omega,s) = 1 - s\,\overline{\psi}(\omega,s) ,$$

where $s \in D_0 = \{u\colon \mathrm{Re}(u) > 0\}$ and $\omega \in \mathrm{Im} = \{u\colon u = it,\ t \in R\}$. The following theorem was provided in [2] for $\rho(i,x) = ix$, and in [3] for general ρ.

Theorem 2.1.

$$\overline{\psi}(\omega,s) \;=\; \frac{1}{\omega}\, p'(0)\,[I - C(\omega,s)]^{-1}\, E_D(\omega,s)\,\underline{e}\;, \qquad \omega,s \in D_0\;. \tag{14}$$

For deriving the moments of T_z, we first note that

$$\left(\frac{\partial}{\partial s}\right)^k \overline{\psi}(\omega,s) \;=\; (-1)^k \int_0^\infty e^{-\omega z}\left(\int_0^\infty t^k\, e^{-st} P\{T_z > t\}\, dt\right) dz\;,$$

$$(-1)^k\,(k+1)\left(\frac{\partial}{\partial s}\right)^k \overline{\psi}(\omega,s)\big|_{s=0} \;=\; \int_0^\infty e^{-\omega z} E[T_z^{k+1}]\, dz\;,$$

$0 \le k \le K$, where $E[T_z^{K+1}] < \infty$ is assumed. Hence from the formula given above and Theorem 2.1,

$$\int_0^\infty e^{-\omega z}\, E[T_z^{k+1}]\, dz \;=\; (k+1)\,(-1)^k\, \frac{1}{\omega}\, p'(0)\left\{\left(\frac{\partial}{\partial s}\right)^k \phi(0,\omega,s)\big|_{s=0}\right\} \underline{e}\;. \tag{15}$$

In next section we use (15) to derive exact formulas for $E[T_z^k]$.

3. EXACT FORMULAS

In this section we apply (15) in order to derive formulas for ET_z, ET_z^2, and $\mathrm{var}(T_z)$. Throughout this section we assume that ρ satisfies the following condition.

(**A**) For each k, $\rho(k,x)\colon [0,\infty) \to [0,\infty)$ is one to one, admits a continuously differential inverse, and $\rho(k,0) = 0$.

We also introduce the following matrices:

$$F(t) \;=\; \Big[\delta_{kj}(\rho^{-1}(k,t))'\, \overline{A}_k(\rho^{-1}(k,t))\Big]\;,$$

$$\mathcal{B}(t) \;=\; \sum_{n=0}^{\infty} B^{(n)}(t)\;,$$

$$K(t) \;=\; \Big[\delta_{kj}\, \rho^{-1}(k,t)\, (\rho^{-1}(k,t))'\, \overline{A}_k(\rho^{-1}(k,t))\Big]\;,$$

$$D(t) \;=\; \left[\int_0^t \rho^{-1}(k,x)\, dB_{kj}(x)\right],$$

where B is the matrix with entries $B_{kj}(z) = A_{kj}(\rho^{-1}(k,z))$ and $B^{(n)}$ is n-fold convolution of B.

Theorem 3.1. *Let T_z be the first passage time of the reward process $\mathcal{Z}_\rho(t)$, $t \ge 0$, given by (1) with a reward function $\rho(k,x)$, $k \in \mathcal{N}$, $x \ge 0$, that satisfying condition (A). If $\mathcal{B}(t)$ exist then*

$$\textbf{(a)} \qquad ET_z \;=\; p'(0) \left\{ \int_0^z \mathcal{B} * F(x)\, dx \right\} \underline{e} \;,$$

$$\textbf{(b)} \qquad ET_z^2 \;=\; 2\, p'(0) \left\{ \int_0^z \mathcal{B} * K(x)\, dx \right\} \underline{e} \,+\, 2\, p'(0) \left\{ \int_0^z \mathcal{B} * D * \mathcal{B} * F(x)\, dx \right\} \underline{e} \;,$$

$$\textbf{(c)} \quad \mathrm{var}(T_z) \;=\; 2\, p'(0) \left\{ \int_0^z \mathcal{B} * K(x)\, dx \right\} \underline{e} \,+\, 2\, p'(0) \left\{ \int_0^z \mathcal{B} * D * \mathcal{B} * F(x)\, dx \right\} \underline{e} - \left\{ p'(0) \left\{ \int_0^z \mathcal{B} * F(x)\, dx \right\} \underline{e} \right\}^2 .$$

Proof: (**a**): By using (15) and (9) we obtain that

$$\mathcal{L}_\omega(ET_z) \;=\; \frac{1}{\omega}\, p'(0)\, [I - C(\omega,0)]^{-1}\, E_D(\omega,0)\, \underline{e} \;, \tag{16}$$

where

$$C(\omega,0) = [C_{kj}(\omega,0)] \;,$$

with

$$C_{kj}(\omega,0) \;=\; \int_0^\infty e^{-\omega\rho(k,x)}\, dA_{kj}(x) \;.$$

Now for each k, j, let $B_{kj}(\Delta) = A_{kj}\{x \in [0,\infty) \colon \rho(k,x) \in \Delta\}$, $\Delta \subset [0,\infty)$, then $B_{kj}(.)$ is a probability distribution on $[0,\infty)$ and it follows by change of variable that,

$$\begin{aligned} C_{kj}(\omega,0) \;&=\; \int_0^\infty e^{-\omega t}\, dB_{kj}(t) \\ &=\; \beta_{kj}(\omega) \;. \end{aligned}$$

Therefore $C_{kj}(\omega,0)$ is the Laplace transform of the distribution B_{kj}, and in matrix form

$$[I - C(\omega,0)]^{-1} \;=\; [I - \beta(\omega)]^{-1} \;. \tag{17}$$

Also note that

$$E_D(\omega,0) = [\delta_{ij} E_j(\omega,0)] \;,$$

where

$$E_j(\omega,0) \;=\; \int_0^\infty e^{-\omega\rho(j,x)}\, \overline{A}_j(x)\, dx \;,$$

and it follows by change of variable that

$$\begin{aligned} E_j(\omega,0) \;&=\; \int_0^\infty e^{-\omega t} (\rho^{-1}(j,t))'\, \overline{A}_j(\rho^{-1}(j,t))\, dt \\ &=\; \int_0^\infty e^{-\omega t} F(j,t)\, dt \;. \end{aligned}$$

Therefore in matrix form we have

$$E_D(\omega,0) \;=\; \int_0^\infty e^{-\omega t} F(t)\,dt \;. \tag{18}$$

If we replace (17) and (18) in (16) we obtain

$$\begin{aligned}\mathcal{L}_\omega(ET_z) \;&=\; p'(0)\,\frac{1}{\omega}\,[I-\beta(\omega)]^{-1}\,\mathcal{L}_\omega(F(t))\,\underline{e}\\ &=\; p'(0)\,\frac{1}{\omega}\,\mathcal{L}_\omega(\mathcal{B}(t))\,\mathcal{L}_\omega(F(t))\,\underline{e}\;,\end{aligned}$$

or equivalently

$$ET_z \;=\; p'(0)\left\{\int_0^z \mathcal{B} * F(t)\,dt\right\}\underline{e}\;,$$

giving (a).

(**b**): It follows from (15) that

$$\mathcal{L}_\omega E[T_z^2] \;=\; -2\,\frac{1}{\omega}\;p'(0)\left\{\frac{\partial}{\partial s}\phi(0,\omega,s)\big|_{s=0}\right\}\underline{e}\;. \tag{19}$$

But from (9),

$$\begin{aligned}\frac{\partial\phi(0,\omega,s)}{\partial s} \;&=\; [I-C(\omega,s)]^{-1}\,\frac{\partial C(\omega,s)}{\partial s}\;[I-C(\omega,s)]^{-1}\,E_D(\omega,s)\\ &\quad+\;[I-C(\omega,s)]^{-1}\,\frac{\partial E_D(\omega,s)}{\partial s}\;,\end{aligned} \tag{20}$$

where

$$\begin{aligned}C_{kj}(\omega,s) \;&=\; \int_0^\infty e^{-\omega\rho(k,x)-sx}\,dA_{kj}(x)\;,\\ \frac{\partial C_{kj}(\omega,s)}{\partial s}\Big|_{s=0} \;&=\; -\int_0^\infty x\,e^{-\omega\rho(k,x)}\,dA_{kj}(x)\;.\end{aligned}$$

Again it follows by change of variable that

$$\frac{\partial C_{kj}(\omega,s)}{\partial s}\Big|_{s=0} \;=\; -\int_0^\infty e^{-\omega t}\rho^{-1}(k,t)\,dB_{kj}(t)\;.$$

Therefore in matrix form

$$\begin{aligned}\frac{\partial C(\omega,s)}{\partial s}\Big|_{s=0} \;&=\; -\int_0^\infty e^{-\omega t}\rho_D^{-1}(t)\,dB(t)\\ &=\; -\mathcal{L}_\omega(D)\;,\end{aligned} \tag{21}$$

where

$$\begin{aligned}D(\Delta) \;&=\; \int_\Delta \rho_D^{-1}(t)\,dB(t)\;,\\ \rho_D^{-1}(t) \;&=\; \left[\delta_{kj}\,\rho^{-1}(k,t)\right]\;.\end{aligned}$$

On the other hand

$$\frac{\partial E_k(\omega,s)}{\partial s}\Big|_{s=0} = -\int_0^\infty x\, e^{-\omega\rho(k,x)}\, \overline{A}_k(x)\, dx \ ,$$

and using change of variable

$$\begin{aligned}\frac{\partial E_k(\omega,s)}{\partial s}\Big|_{s=0} &= -\int_0^\infty e^{-\omega t}\rho^{-1}(k,t)\,(\rho^{-1}(k,t))'\ \overline{A}_k(\rho^{-1}(k,t))\ dt \\ &= -\int_0^\infty e^{-\omega t} K(k,t)\, dt \ .\end{aligned}$$

Therefore in matrix form

$$\begin{aligned}\frac{\partial E_D(\omega,s)}{\partial s}\Big|_{s=0} &= -\int_0^\infty e^{-\omega t} K(t)\, dt \\ &= -\mathcal{L}_\omega(K) \ .\end{aligned} \tag{22}$$

By replacing (17), (18), (21) and (22) in (20), we obtain from (19) that

$$\begin{aligned}\mathcal{L}_\omega(ET_z^2) &= 2\,p'(0)\,\frac{1}{\omega}\,[I-\beta(\omega)]^{-1}\,\mathcal{L}_\omega(D(t))\,[I-\beta(\omega)]^{-1}\mathcal{L}_\omega(F(t))\,\underline{e} \\ &\quad + 2\,p'(0)\,\frac{1}{\omega}\,[I-\beta(\omega)]^{-1}\mathcal{L}_\omega(K(t))\,\underline{e} \ ,\end{aligned}$$

or

$$ET_z^2 = 2\,p'(0)\left\{\int_0^z \mathcal{B} * K(x)\, dx\right\}\underline{e} + 2\,p'(0)\left\{\int_0^z \mathcal{B} * D * \mathcal{B} * F(x)\, dx\right\}\underline{e} \ .$$

Part (**c**) Follows from (a) and (b). □

Corollary 3.1. *Let $\rho(k,x) = g_n(k)x^n,\ k \in \mathcal{N},\ x \in [0,\infty)$ and $g_n(k) > 0$. If $\mathcal{B}(t)$ exists, then the formulas (a), (b) and (c) of Theorem 3.1 are satisfied. Moreover*

$$\begin{aligned}F(t) &= \left[\delta_{ij}\,\frac{1}{n\,\sqrt[n]{\rho_j\, t^{n-1}}}\ \overline{A}_j\!\left(\sqrt[n]{\frac{t}{\rho_j}}\right)\right] , \\ B(t) &= [B_{ij}] \ , \qquad B_{ij}(t) = A_{ij}\!\left(\sqrt[n]{\frac{t}{\rho_j}}\right) , \\ K(t) &= \left[\delta_{ij}\,\frac{1}{n\,\sqrt[n]{\rho_j^2\, t^{n-2}}}\ \left(1 - A_j\!\left(\sqrt[n]{\frac{t}{\rho_j}}\right)\right)\right] , \\ D(t) &= \left[\int_0^t \frac{1}{n\,\sqrt[n]{\rho_j^2\, x^{n-2}}}\, dA_{ij}\!\left(\sqrt[n]{\frac{x}{\rho_j}}\right)\right] .\end{aligned}$$

Proof: The reward function satisfies condition (A), therefore Theorem 3.1 can be applied. □

Remark 3.1. Let $n = 1$ in Corollary 3.1, i.e., the reward function is linear. Then Corollary 3.1 holds with $n = 1$.

4. APPLICATIONS TO CERTAIN TYPE I COUNTERS MODELS

Arrivals at a counter form a Poisson process with rate q. An arriving particle that finds the counter free gets registered and locks it for a random duration with distribution function $F(t)$. Arrivals during a locked periods have no effect whatsover. Suppose a registration occurs at $T_0 = 0$, and write $T_0, T_1, T_2, ...$ for the successive epochs of changes in the state of the counter. Write $X_n = 1$ or 0 according as the n-th change locks or frees the counter. Clearly $X_0 = 1$, $X_1 = 0$, $X_2 = 1$, $X_3 = 0$, ... and (X_n, T_n) is a Markov renewal process. Its semi-Markov matrix is

$$A(x) = \begin{bmatrix} 0 & 1-e^{-qx} \\ F(x) & 0 \end{bmatrix}.$$

Let $F(x) = 1 - e^{-2qx}$ and $\mathcal{Z}_\rho(t)$ be the reward process that is defined by (1) with reward function $\rho(k, x) = \rho_k\, x$, $\rho_0 = 1$, $\rho_1 = 2$. Let T_z be the first passage time reward process $\mathcal{Z}_\rho(t)$ from $\mathcal{Z}_\rho(0) = 0$ to a prespecified level z. We apply the formulas of the previous section to give explicit expressions for ET_z and ET_z^2. Note that for each k, j

$$\begin{aligned} B_{kj}(t) &= A_{kj}\left(\frac{t}{\rho_k}\right), \\ B(t) &= \begin{bmatrix} 0 & 1-e^{-qt} \\ 1-e^{-qt} & 0 \end{bmatrix}, \\ B^{(0)}(t) &= I\ . \end{aligned}$$

By induction it follows that

$$B^{(2n+1)}(t) = \begin{bmatrix} 0 & B_{01}^{(2n+1)} \\ B_{10}^{(2n+1)} & 0 \end{bmatrix},$$

where

$$B_{01}^{(2n+1)} = B_{10}^{(2n+1)} = 1 - e^{-qt} - qte^{-qt} - \frac{q^2\, t^2}{2!}e^{-qt} - \cdots - \frac{q^{2n}\, t^{2n}}{2\, n!}e^{-qt}\ ,$$

and

$$B^{(2n)}(t) \;=\; \begin{bmatrix} B_{00}^{(2n)} & 0 \\ 0 & B_{11}^{(2n)} \end{bmatrix},$$

$$B_{00}^{(2n)} \;=\; B_{11}^{(2n)} \;=\; 1-e^{-qt}-qte^{-qt}-\frac{q^2\,t^2}{2}e^{-qt}-\frac{q^3\,t^3}{3!}e^{-qt}-\cdots-\frac{q^{2n-1}\,t^{2n-1}}{(2n-1)!}e^{-qt}\;.$$

Therefore

$$\begin{aligned}
\mathcal{B}_{00}(t) &= \sum_{n=0}^{\infty} B_{00}^{(n)}(t) \;=\; 1+\sum_{n=1}^{\infty}\left[1-e^{-qt}\sum_{k=0}^{2n-1}\frac{(qt)^k}{k!}\right],\\
\mathcal{B}_{00}(t) &= \mathcal{B}_{11}(t)\;,\\
\mathcal{B}_{01}(t) &= \sum_{n=0}^{\infty} B_{01}^{(n)}(t) \;=\; \sum_{n=0}^{\infty}\left[1-e^{-qt}\sum_{k=0}^{2n}\frac{(qt)^k}{k!}\right],\\
\mathcal{B}_{01}(t) &= \mathcal{B}_{10}(t)\;,\\
\mathcal{B}_{00}(t) &= 1+\sum_{n=0}^{\infty}\Big[1-P(Y\le 2n+1)\Big]\\
&= 1+\sum_{n=0}^{\infty}P(Y>2n+1)\;,\\
\mathcal{B}_{01}(t) &= \sum_{n=0}^{\infty}\Big[1-P(Y\le 2n)\Big]\\
&= \sum_{n=0}^{\infty}P(Y>2n)\;,
\end{aligned}$$

where Y is a Poisson random variable with $\lambda = qt$. Therefore

$$\mathcal{B}(t) \;=\; \begin{bmatrix} 1+\sum_{n=0}^{\infty}P(Y>2n+1) & \sum_{n=0}^{\infty}P(Y>2n) \\ \sum_{n=0}^{\infty}P(Y>2n) & 1+\sum_{n=0}^{\infty}P(Y>2n+1) \end{bmatrix}.$$

The derivation of $\mathcal{B}(t)$ can be simplified by noting that if

$$p_k \;=\; \frac{\lambda^k\,e^{-\lambda}}{k!}$$

where $\lambda = qt$, then

$$P_E \;\equiv\; P\{Y\text{even}\} \;=\; \sum_{k\in\{0,2,4,\ldots\}} p_k$$

and

$$P_O \;\equiv\; P\{Y\text{odd}\} \;=\; \sum_{k\in\{1,3,5,\ldots\}} p_k$$

implying that (after simplication)

$$\begin{aligned}\sum_{n=0}^{\infty} P(Y > 2n+1) &= (p_2 + 2p_4 + 3p_6 + ...) + (p_3 + 2p_5 + 3p_7 + ...) \\ &= \frac{\lambda}{2} P_O + \left\{ \frac{\lambda}{2}(P_E - e^{-\lambda}) - \frac{1}{2}(P_O - \lambda e^{-\lambda}) \right\} \\ &= \frac{\lambda}{2} - \frac{P_O}{2} ,\end{aligned}$$

similarly

$$\begin{aligned}\sum_{n=0}^{\infty} P(Y > 2n) &= (p_1 + 2p_3 + 3p_5 + ...) + (p_2 + 2p_4 + 3p_6 + ...) \\ &= \frac{1}{2} \left\{ \lambda P_E + P_O \right\} + \frac{\lambda}{2} P_O \\ &= \frac{\lambda}{2} + \frac{P_O}{2} .\end{aligned}$$

Now if $P(s) = \sum_{k=0}^{\infty} p_k \, s^k = e^{-\lambda + \lambda s}$, then

$$\begin{aligned}P(1) &= p_0 + p_1 + p_2 + p_3 + ... = 1 = P_O + P_E , \\ P(-1) &= p_0 - p_1 + p_2 - p_3 + ... = e^{-2\lambda} = P_E - P_O ,\end{aligned}$$

implying $P_E = \frac{1}{2}(1 + e^{-2\lambda})$ and $P_O = \frac{1}{2}(1 - e^{-2\lambda})$. Hence

$$\begin{aligned}\sum_{n=0}^{\infty} P(Y > 2n+1) &= \frac{\lambda}{2} - \frac{1}{4} + \frac{e^{-2\lambda}}{4} = \frac{qt}{2} - \frac{1}{4} + \frac{e^{-2qt}}{4} , \\ \sum_{n=0}^{\infty} P(Y > 2n) &= \frac{\lambda}{2} + \frac{1}{4} - \frac{e^{-2\lambda}}{4} = \frac{qt}{2} + \frac{1}{4} - \frac{e^{-2qt}}{4} ,\end{aligned}$$

and

$$\mathcal{B}(t) = \begin{bmatrix} \frac{qt}{2} + \frac{3}{4} + \frac{e^{-2qt}}{4} & \frac{qt}{2} + \frac{1}{4} - \frac{e^{-2qt}}{4} \\ \frac{qt}{2} + \frac{1}{4} - \frac{e^{-2qt}}{4} & \frac{qt}{2} + \frac{3}{4} + \frac{e^{-2qt}}{4} \end{bmatrix} ,$$

$$F(t) = \begin{bmatrix} e^{-qt} & 0 \\ 0 & \frac{1}{2} e^{-qt} \end{bmatrix} , \qquad K(t) = \begin{bmatrix} te^{-qt} & 0 \\ 0 & \frac{t}{4} e^{-qt} \end{bmatrix} ,$$

$$dD(t) = \begin{bmatrix} 0 & qte^{-qt} \\ \frac{qt}{2} e^{-qt} & 0 \end{bmatrix} ,$$

$$\mathcal{B} * F(t) = \int_0^t d\mathcal{B}(x) \, F(t-x) ,$$

hence

$$\mathcal{B} * F(t) \;=\; \begin{bmatrix} \frac{1}{2}\{1 - 2e^{-qt} + e^{-2qt}\} & \frac{1}{4}\{1 - e^{-2qt}\} \\ \frac{1}{2}\{1 - e^{-2qt}\} & \frac{1}{4}\{1 - 2e^{-qt} + e^{-2qt}\} \end{bmatrix},$$

and

$$\int_0^z \mathcal{B} * F(x)\,dx \;=\; \begin{bmatrix} \frac{z}{2} - \frac{3}{4q} + \frac{1}{q}e^{-qz} - \frac{1}{4q}e^{-2qz} & \frac{z}{4} - \frac{1}{8q} + \frac{1}{8q}e^{-2qz} \\ \frac{z}{2} - \frac{1}{4q} + \frac{1}{4q}e^{-2qz} & \frac{z}{4} - \frac{3}{8q} + \frac{1}{2q}e^{-qz} - \frac{1}{8q}e^{-2qz} \end{bmatrix}.$$

In the example $X_0 = 1$, the initial probability vector is clearly $p'(0) = (1, 0)$, then

$$ET_z \;=\; \frac{3}{4}z - \frac{7}{8q} + \frac{1}{q}\,e^{-qz} - \frac{1}{8q}\,e^{-2qz}\,. \tag{23}$$

$$\mathcal{B} * D * \mathcal{B} * F(x) \;=\; \begin{bmatrix} \mathcal{B} * D * \mathcal{B} * F_{00}(x) & \mathcal{B} * D * \mathcal{B} * F_{01}(x) \\ \mathcal{B} * D * \mathcal{B} * F_{10}(x) & \mathcal{B} * D * \mathcal{B} * F_{11}(x) \end{bmatrix},$$

where

$$\mathcal{B} * D * \mathcal{B} * F_{00}(x) \;=\; \frac{1}{8}\left\{3x - \frac{9}{q} + \frac{9}{q}e^{-2qx} + 12xe^{-qx} + 3xe^{-2qx}\right\},$$

$$\mathcal{B} * D * \mathcal{B} * F_{01}(x) \;=\; \frac{1}{16}\left\{3x - \frac{10}{q} + \frac{12}{q}e^{-qx} - \frac{10}{q}e^{-2qx} - 3xe^{-2qx} + 4qx^2e^{-qx}\right\},$$

$$\mathcal{B} * D * \mathcal{B} * F_{10}(x) \;=\; \frac{1}{8}\left\{3x - \frac{8}{q} + \frac{16}{q}e^{-qx} - \frac{8}{q}e^{-2qx} - 3xe^{-2qx} + 2qx^2e^{-qx}\right\},$$

$$\mathcal{B} * D * \mathcal{B} * F_{11}(x) \;=\; \frac{1}{16}\left\{3x - \frac{9}{q} + \frac{9}{q}e^{-2qx} + 12xe^{-qx} + 3xe^{-2qx}\right\}.$$

Also

$$\mathcal{B} * K(x) \;=\; \begin{bmatrix} \frac{1}{2}\left\{\frac{1}{q} - \frac{1}{q}e^{-2qx} - 2xe^{-qx}\right\} & \frac{1}{8}\left\{\frac{1}{q} - \frac{2}{q}e^{-qx} + \frac{1}{q}e^{-2qx}\right\} \\ \frac{1}{2}\left\{\frac{1}{q} + \frac{1}{q}e^{-2qx} - \frac{2}{q}e^{-qx}\right\} & \frac{1}{8}\left\{\frac{1}{q} - 2xe^{-qx} - \frac{1}{q}e^{-2qx}\right\} \end{bmatrix}.$$

If we replace $\mathcal{B} * D * \mathcal{B} * F(x)$ and $\mathcal{B} * K(x)$ in formula (b) of Corollary 3.1, we get

(24)

$$ET_z^2 \;=\; \frac{1}{16}\left\{9z^2 - \frac{36}{q}z + \frac{103}{2q^2} - \frac{48}{q^2}e^{-qz} - \frac{7}{2q^2}e^{-2qz} - \frac{32}{q}ze^{-qz} + \frac{3}{q}ze^{-2qz} + 8z^2e^{-qz}\right\}.$$

Remark 4.1. The asymptotic behaviors of ET_z, ET_z^2 were derived in [5] for $\rho(k,x) = \rho_k x$, and in [3] for general ρ. For the case considered in the Example given above,

$$ET_z \;=\; \frac{m_1}{m_1^{**}}z \,+\, p'(0)\left\{H_0^{**}A_{D:1} - \frac{1}{2}H_1^{**}\rho_{D:1}A_{D:2}\right\}\underline{e} \,+\, o(1)\,,$$

$$ET_z^2 \;=\; \left\{\frac{m_1}{m_1^{**}}\right\}^2 z^2 \,-\, p'(0)\left\{2\,V_1^{**}A_{D:1} - [V_2^{**} + H_1^{**}A_{D:2}]\right\}\underline{e}\,z \,+\, o(z)\,,$$

as $z \to \infty$, where $m_1 = \pi' A_1 \underline{e}$, $\rho_{D:1}$ = diagonal matrix of ρ_i,

$$B_k = \rho_{D:k} A_k \ , \qquad m_1^{**} = \pi' B_1 \underline{e} \ , \qquad H_1^{**} = \frac{1}{m_1^{**}} \underline{e} \pi' \ , \qquad Z_0 = [I - P + e\pi']^{-1} \ ,$$

$$H_0^{**} = \frac{1}{m_1^{**}} \underline{e}\,\pi' \left\{ -B_1 + \frac{1}{2m_1^{**}} B_2 \,\underline{e}\,\pi' \right\} + \left\{ Z_0 - \frac{1}{m_1^{**}} \underline{e}\,\pi' B_1 Z_0 \right\} \left\{ P - \frac{1}{m_1^{**}} B_1 \,\underline{e}\,\pi' \right\},$$

$$V_1^{**} = (H_1^{**} \rho_{D:1} A_2 - H_0^{**} A_1) H_1^{**} - H_1^{**} A_1 H_0^{**} \ ,$$

$$V_2^{**} = -H_1^{**} A_1 H_1^{**} \rho_{D:1} A_{D:2} \ .$$

For the semi-Markov $A(x)$ defined above

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} , \qquad A_1 = \begin{bmatrix} 0 & \frac{1}{q} \\ \frac{1}{2q} & 0 \end{bmatrix} , \qquad A_{D:1} = \begin{bmatrix} \frac{1}{q} & 0 \\ 0 & \frac{1}{2q} \end{bmatrix} .$$

$$A_2 = \begin{bmatrix} 0 & \frac{2}{q^2} \\ \frac{1}{2q^2} & 0 \end{bmatrix} , \qquad A_{D:2} = \begin{bmatrix} \frac{2}{q^2} & 0 \\ 0 & \frac{1}{2q^2} \end{bmatrix} ,$$

$$\pi' P = \pi' \implies \pi' = (0.5,\, 0.5) \ ,$$

$$m_1 = \pi' A_1 \,\underline{e} = \frac{3}{4q} \ ,$$

$$\rho_{D:1} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} ,$$

$$B_1 = \rho_{D:1}\, A_1 = \begin{bmatrix} 0 & \frac{1}{q} \\ \frac{1}{q} & 0 \end{bmatrix} ,$$

$$m_1^{**} = \pi' B_1 \,\underline{e} = \frac{1}{q} \ ,$$

$$B_2 = \rho_{D:2}\, A_2 = \begin{bmatrix} 0 & \frac{2}{q^2} \\ \frac{2}{q^2} & 0 \end{bmatrix} ,$$

$$Z_0 = [I - P + e\,\pi']^{-1} \ ,$$

therefore

$$Z_0 = \frac{1}{2} \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{bmatrix} ,$$

$$H_1^{**} = \frac{1}{m_1^{**}} \underline{e}\,\pi' \ ,$$

therefore

$$H_1^{**} = q \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} ,$$

$$H_0^{**} \;=\; \frac{1}{m_1^{**}}\,\underline{e}\,\pi'\left\{-B_1+\frac{1}{2m_1^{**}}\,B_2\,\underline{e}\,\pi'\right\}+\left\{Z_0-\frac{1}{m_1^{**}}\,\underline{e}\,\pi'\,B_1 Z_0\right\}\left\{P-\frac{1}{m_1^{**}}\,B_1\,\underline{e}\,\pi'\right\},$$

hence

$$H_0^{**} = \begin{bmatrix} -\frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{4} \end{bmatrix},$$

$$V_1^{**} \;=\; (H_1^{**}\rho_{D:1}A_2 - H_0^{**}A_1)\,H_1^{**} - H_1^{**}A_1 H_0^{**}\;,$$

therefore

$$V_1^{**} = \begin{bmatrix} \frac{12}{16} & \frac{14}{16} \\ \frac{10}{16} & \frac{12}{16} \end{bmatrix},$$

$$V_2^{**} = -H_1^{**}A_1 H_1^{**}\rho_{D:1}A_{D:2}\;,$$

$$V_2^{**} = -\begin{bmatrix} \frac{3}{4q} & \frac{3}{8q} \\ \frac{3}{4q} & \frac{3}{8q} \end{bmatrix}.$$

In the example, $X_0=1$, so that the initial probability vector is clearly $p'(0)=(1,0)$. Then by replacing values in ET_z, ET_z^2, we have

$$\begin{aligned} ET_z \;&=\; \frac{3}{4}z - \frac{7}{8q} + o(1)\;, \\ ET_z^2 \;&=\; \frac{9}{16}z^2 - \frac{18}{8q}z + o(z)\;, \end{aligned}$$

as $z\to\infty$, which also can be observed from the formulas (23), (24), as $z\to\infty$.

Remark 4.2. If one wishes to compare ET_z with the asymptotic behaviour it is sensible to allow for a general initial probability vector say $p'(0) = (p_0(0),\ p_1(0))$. In this case

$$\begin{aligned} ET_z \;&=\; \frac{3}{4}z - \frac{7p_0(0)+5p_1(0)}{8q} + \frac{2p_0(0)+p_1(0)}{2q}\,e^{-qz} - \frac{p_1(0)+p_0(0)}{8q}\,e^{-2qz} \\ &=\; \frac{3}{4}z - \frac{7p_0(0)+5p_1(0)}{8q} + o(1)\;. \end{aligned}$$

This last result is also obtained for the asymptotic expression for ET_z with a general initial probability vector.

ACKNOWLEDGMENTS

We acknowledge the valuable suggestions of the referees.

REFERENCES

[1] Mclean, R.A. and Neuts, M.F. (1967). The integral of a step function defined on a semi Markov process, *SIAM J. Appl. Math.*, **15**, 726–737.

[2] Masuda, Y. and Sumita, U. (1991). A multivariate reward processes defined on a semi-Markov process and its first passage time distributions, *J. Appl. Prob.*, **28**, 360–373.

[3] Parham, G.A. and Soltani, A.R. (1998). First passage time for reward processes with nonlinear reward function: asymptotic behavior, *Pak. J. Statist.*, **14**, 65–80.

[4] Soltani, A.R. (1996). Reward Processes with nonlinear Reward Functions, *J. Appl. Prob.*, **33**, 1011–1017.

[5] Sumita, U. and Masuda, Y. (1987). An Alternative Approach to the Analysis of Finite Semi-Markov and Related Processes, *Commun. Statist.-Stochastic Models*, **3**, 67–87.

REVSTAT – Statistical Journal
Volume 3, Number 1, June 2005, 61–75

PREDICTION OF TREATMENTS EFFECTS IN A BIASED ALLOCATION MODEL

Authors: Fernando J.M. Magalhães
– ISCAP / Inst. Politécnico do Porto, Univ. Portucalense and CEAUL, R. Prof. Jaime Lopes Amorim, 4465-111 S. Mamede de Infesta, Portugal

Ian R. Dunsmore
– Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, U.K.

David J. Robson
– Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, U.K.

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

Abstract:

- Robbins and Zhang [15] provide consistent estimators of multiplicative treatment effects under a biased treatment allocation scheme, and illustrate their methodology within Poisson and binomial models. Here we use predictive criteria to assess the differential treatment effects, and develop predictive distributions for the Poisson errors in variables models. With a hierarchical prior structure, various approximations are investigated, and an illustrative example is included.

Key-Words:

- *biased allocation; errors in variables; Gibbs sampling; Laplace approximation; Poisson model; predictive distributions; treatment effect.*

AMS Subject Classification:

- 62F15, 62F10, 62P10, 62E15, 62E17, 62E20.

1. INTRODUCTION

Robbins and Zhang ([13], [14], [15]) consider the estimation of a multiplicative treatment effect under biased allocation. For example, with a slight change from their notation to allow for generalization, suppose that within a Poisson errors in variables model (θ_i, X_i, Y_i), $i=1,2,...,n$, are independent random vectors such that

$$\text{(1.1)}\qquad \begin{aligned} &\textbf{(i)}\ \ \text{given } \theta_i, X_i \text{ is } Po(\theta_i)\,;\\ &\textbf{(ii)}\ \ \text{given } \beta_1,\ \beta_2,\ \theta_i \text{ and } x_i,\\ &\qquad\qquad Y_i \text{ is } Po(\beta_1\theta_i) \ \text{ if treatment } T_1 \text{ is used}\,;\\ &\qquad\qquad Y_i \text{ is } Po(\beta_2\theta_i) \ \text{ if treatment } T_2 \text{ is used}\,;\\ &\textbf{(iii)}\ \ \text{given } a,\ T_1 \text{ is used if } x_i < a \ \text{ and } \ T_2 \text{ is used if } x_i \geq a\ ; \end{aligned}$$

where $Po(\mu)$ represents a Poisson distribution with mean μ. No distributional assumptions about the θ_i's are made, and their values are not observed. The unknown parameters β_1 and β_2 could be thought of as multiplicative treatment effects. An alternative parameterization would be through logarithmic link functions with additive treatment effects.

Robbins and Zhang [15] discuss two scenarios for this model. The first concerns the number of accidents at road junctions. Suppose that X_i counts the number of night accidents during year 1 at junction i, $i=1,2,...,n$. Extra lights are installed at the beginning of year 2 at those junctions for which $x_i \geq a$, with no change being made to the light system at other junctions. Then, Y_i records the number of night accidents at junction i during year 2. Of particular interest is whether or not the extra lights reduce the frequency of night accidents.

The second, more controversial, application is in the context of clinical trials in which the allocation of treatments is based on the screening variable X. Robbins and Zhang [15] then seek to estimate the differential treatment effects based on this biased allocation of treatments to patients.

In both situations Robbins and Zhang [15] consider the problem as one of estimation, and take the difference $\beta_2-\beta_1$ or the ratio β_2/β_1 as a measure of the differential treatment effect. Based on data $(\boldsymbol{x}^n, \boldsymbol{y}^n) = \{(x_i, y_i)\colon\ i=1,2,...,n\}$, they derive the following consistent estimates for β_1 and β_2:

$$\text{(1.2)}\qquad \beta_{1,n} = \frac{\sum_{i=1}^{n} y_i\, I(x_i<a)}{\sum_{i=1}^{n} x_i\, I(x_i<a+1)}\ , \qquad \beta_{2,n} = \frac{\sum_{i=1}^{n} y_i\, I(x_i\geq a)}{\sum_{i=1}^{n} x_i\, I(x_i\geq a+1)}\ ,$$

where I represents the indicator function. They suggest the use of the central limit theorem to obtain confidence intervals for β_1, β_2, $\beta_2-\beta_1$ or β_2/β_1

with coverage probabilities that tend to 0.95, say, as $n \to \infty$, but omit details. They also note that the Poisson assumption for the conditional distribution of Y_i in (ii) of (1.1) is not required for the consistency of the estimates in (1.2).

Godambe and Kunte [6] provide an alternative semi-parametric solution for the estimation of β_1 and β_2 through the use of optimum estimating functions (Godambe and Thompson, [7]). Their model does not require the Poisson assumptions in (1.1) for X_i or Y_i, but only the mean value specifications. However, they do require an additional assumption, namely that

(iv) given x_i the mean value of θ_i is $f(x_i)$,
f being a specified function of x_i .

The assumption that the θ_i are (unobservable) random variables distinguishes the model from one in which they are unknown parameters. In this latter case, one might then consider the θ_i as "incidental" parameters, following the terminology of Neyman and Scott [10], as opposed to the "structural" parameters β_1 and β_2. Kiefer and Wolfowitz [8] discuss problems of consistency with maximum likelihood estimation in such cases, and illustrate how these may be overcome if the θ_i are independent chance variables with a common distribution, as in here.

We develop here, in Section 2, an approach to treatment comparisons based on predictive criteria, which perhaps seem more relevant for answering, in the medical context for example, the question "Which of the two treatments do I give to the next patient?". The approach extends the models used in Dunsmore and Robson [2] for other Poisson errors in variables models. We concentrate attention on the outcomes $Y_{n+1,1}$ and $Y_{n+1,2}$ from separate applications of the two treatments, T_1 and T_2, applied to patient $n+1$, and seek to make predictions for future values $y_{n+1,1}$ and $y_{n+1,2}$ based on $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ and x_{n+1}.

An illustrative example is provided in Section 3, and extensions to several treatments and other distributional models are discussed briefly in sections 4 and 5.

2. POISSON PREDICTIVE MODELS

2.1. Predictive distribution

Consider the Poisson errors in variables model specified in (1.1). Suppose further that for a future individual, labelled by $n+1$, we observe x_{n+1} from a $Po(\theta_{n+1})$ distribution and model potential outcomes from the two treatments through

$$Y_{n+1,1} \text{ is } Po(\beta_1\theta_{n+1}) \; ; \qquad Y_{n+1,2} \text{ is } Po(\beta_2\theta_{n+1}) \; . \tag{2.1}$$

The dependence between the outcomes for this individual from the two treatments is modelled through the common (unobserved) θ_{n+1}.

Such an individual will only be given one of the two treatments, and the predictive paradigm suggests that the choice centres around properties of the joint predictive function $p\big(y_{n+1,1}, y_{n+1,2} \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$, or perhaps considerations of $Y_{n+1,2} - Y_{n+1,1}$.

We denote the treatment given to an individual, for $i = 1, 2, ..., n$, by

$$\delta_{ij} \;=\; \begin{cases} 1, & \text{if individual } i \text{ gets treatment } T_j\,, \\ 0, & \text{otherwise}\,, \end{cases}$$

so that $\delta_{i1} + \delta_{i2} = 1$ for each i; and let, for $j = 1, 2$,

$$\begin{aligned} n_j &= \sum_{i=1}^{n} \delta_{ij}\;; & n &= n_1 + n_2\;; \\ S_{xj} &= \sum_{i=1}^{n} \delta_{ij}\, x_i\;; & T_x &= \sum_{j=1}^{2} S_{xj} \;= \sum_{i=1}^{n} x_i\;; \\ S_{yj} &= \sum_{i=1}^{n} \delta_{ij}\, y_i\;; & T_y &= \sum_{j=1}^{2} S_{yj} \;= \sum_{i=1}^{n} y_i\,. \end{aligned}$$

The maximum likelihood estimates of the parameters are given by

$$\hat{\theta}_i = \frac{x_i + y_i}{1 + \sum\limits_{j=1}^{2} \delta_{ij}\, \hat{\beta}_j}, \quad i = 1, 2, ..., n; \quad \hat{\theta}_{n+1} = x_{n+1}; \quad \hat{\beta}_j = \frac{S_{yj}}{S_{xj}}, \; j = 1, 2\,. \tag{2.2}$$

In this notation the Robbins and Zhang [15] estimates (1.2) based on $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ are given by

$$\beta_{1,n} = \frac{S_{y1}}{S_{x1} + aN_a}\,, \qquad \beta_{2,n} = \frac{S_{y2}}{S_{x2} - aN_a}\,, \tag{2.3}$$

where N_a are the number of x_i's equal to a. The additional information provided by x_{n+1} could be used to amend these estimates to

$$\begin{aligned} \beta^*_{1,n} &= \left(1 + \frac{1}{n}\right) \frac{S_{y1}}{S_{x1} + aN_a + x_{n+1}\, I\,(x_{n+1} < a+1)}\,, \\ \beta^*_{2,n} &= \left(1 + \frac{1}{n}\right) \frac{S_{y2}}{S_{x2} - aN_a + x_{n+1}\, I\,(x_{n+1} \geq a+1)}\,. \end{aligned} \tag{2.4}$$

2.2. Plug-in estimates

Predictive approaches within the classical framework typically involve plug-in estimates, pivotal statistics or tolerance regions. A simple plug-in estimate for the probability function of $Z = Y_{n+1,2} - Y_{n+1,1}$, for example, would be given by

$$P(Z\!=\!z) \;=\; \sum_{i=\max(0,-z)}^{\infty} \prod_{j=1}^{2} \frac{(\hat\beta_j\,\hat\theta_{n+1})^{i+(j-1)z}\,\exp(-\hat\beta_j\,\hat\theta_{n+1})}{\big(i+(j\!-\!1)z\big)!}\;. \tag{2.5}$$

Other such estimates are available if specification of the underlying distribution for the θ_i's is provided. As an illustration, we take $p(\theta_i)$ to be $\gamma\exp(-\gamma\theta_i)$ with unknown parameter $\gamma > 0$. The model specification in (1.1) then reduces to

$$p\big(x_i, y_i \mid \beta_1, \beta_2, \gamma\big) \;=\; \frac{(x_i+y_i)!}{x_i!\; y_i!}\;\frac{\gamma\,\beta_1^{\delta_{i1}y_i}\,\beta_2^{\delta_{i2}y_i}}{\big(1+\gamma+\delta_{i1}\beta_1+\delta_{i2}\beta_2\big)^{x_i+y_i+1}}\;, \tag{2.6}$$

whilst Godambe and Kunte's [6] condition (iv) above is satisfied by $f(x) = (x+1)/(\gamma+1)$. Both Robbins and Zhang [15] and Godambe and Kunte [6] consider this case. The former demonstrate that their estimates $\beta_{1,n}$ and $\beta_{2,n}$ in (1.2) compete well with the maximum likelihood estimates based on $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ from this fully parametric model; whilst the latter's solution coincides with them.

With the additional information from x_{n+1}, the maximum likelihood estimates are now given by

$$\breve\beta_1 = (1+\breve\gamma)\,\frac{S_{y1}}{S_{x1}+n_1}\;;\quad \breve\beta_2 = (1+\breve\gamma)\,\frac{S_{y2}}{S_{x2}+n_2}\;;\qquad \breve\gamma = \frac{n+1}{T_x+x_{n+1}}\;. \tag{2.7}$$

A simple plug-in estimate for the probability function of $Z = Y_{n+1,2} - Y_{n+1,1}$ would then be given by

$$P(Z\!=\!z) \;=\; \sum_{i=\max(0,-z)}^{\infty} \frac{(z+2i)!}{i!\,(z+i)!}\;\frac{\breve\gamma\,\breve\beta_1^{i}\,\breve\beta_2^{i}}{\big(\breve\gamma+\breve\beta_1+\breve\beta_2\big)^{z+2i+1}}\;. \tag{2.8}$$

2.3. Hierarchical prior structure

Within a Bayesian framework for the model specified by (1.1) and (2.1), the central feature is the predictive function $p\big(y_{n+1,1}, y_{n+1,2} \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ given by

$$\int \prod_{j=1}^{2} \big\{p(y_{n+1,j} \,|\, \beta_j, \theta_{n+1})\big\}\; p\big(\theta_{n+1}, \beta_1, \beta_2 \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)\; d\theta_{n+1}\, d\beta_1\, d\beta_2\;, \tag{2.9}$$

where $p\big(\theta_{n+1}, \beta_1, \beta_2 \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ is the posterior density function. Notice here that $\theta^n = (\theta_1, \theta_2, ..., \theta_n)$ behaves in the same way as a nuisance parameter, and we only require the posterior distribution of $(\theta_{n+1}, \beta_1, \beta_2)$ — or indeed only of $(\beta_1\theta_{n+1}, \beta_2\theta_{n+1})$.

Following the ideas in Gelfand and Smith [5], we adopt a Bayesian hierarchical prior structure. At the first stage we take

$$p\big(\theta^n, \theta_{n+1}, \beta_1, \beta_2 \mid \gamma, \eta_1, \eta_2\big) \;=\; \prod_{i=1}^{n+1} p(\theta_i \,|\, \gamma) \;\prod_{j=1}^{2} p(\beta_j \,|\, \eta_j) \;,$$

whilst at the second stage we assume

$$p(\gamma, \eta_1, \eta_2) \;=\; p(\gamma) \prod_{j=1}^{2} p(\eta_j) \;.$$

An appropriate structure here would be of the form

$$\begin{aligned} &\theta_i \sim Ga(k, \gamma) \;, &&\beta_j \sim Ga(g_j, \eta_j) \;, \\ &\gamma \sim Ga(\ell, m) \;, &&\eta_j \sim Ga(u_j, v_j) \;, \end{aligned}$$

for $i=1, 2, ..., n, n{+}1$ and $j=1, 2$, where $Ga(a, b)$ represents a gamma distribution with density proportional to $\theta^{a-1} \exp(-b\theta)$, $\theta > 0$, and where k, g_1, g_2, ℓ, m, u_1, v_1, u_2 and v_2 are known constants. Gaver and O'Muircheartaigh [3] and Gelfand and Smith [5] suggest, in a similar framework, that k, g_1 and g_2 might be treated as tuning parameters or estimated in an empirical Bayes spirit. Notice that the distributional assumptions about θ^n and θ_{n+1} in Section 2.2 are a special case of the above.

The posterior density function $p\big(\theta^n, \theta_{n+1}, \beta_1, \beta_2, \gamma, \eta_1, \eta_2 \mid x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ is proportional to

$$\begin{aligned} &\prod_{i=1}^{n} \left[\exp\left\{ -\theta_i \left(1 + \sum_{j=1}^{2} \delta_{ij}\, \beta_j + \gamma \right) \right\} \theta_i^{x_i+y_i+k} \right] \;\times \\ &\qquad \times\; \exp\big\{ -\theta_{n+1}(1{+}\gamma) \big\}\, \theta_{n+1}^{x_{n+1}+k}\, \gamma^{(n+1)k+\ell} \exp\{-m\gamma\} \;\times \\ &\qquad \times\; \prod_{j=1}^{2} \Big\{ \beta_j^{S_{yj}+g_j} \exp(-\eta_j\beta_j)\, \eta_j^{g_j+u_j} \exp(-v_j\eta_j) \Big\} \;. \end{aligned} \tag{2.10}$$

Eliminating θ^n, η_1 and η_2 we have that $p\big(\theta_{n+1}, \beta_1, \beta_2, \gamma \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ is proportional to

$$\frac{\exp\big\{ -\theta_{n+1}(1{+}\gamma) \big\}\, \theta_{n+1}^{x_{n+1}+k} \displaystyle\prod_{j=1}^{2} \big\{ \beta_j^{S_{yj}+g_j} \big\}\, \gamma^{(n+1)k+\ell} \exp\{-m\gamma\}}{\displaystyle\prod_{j=1}^{2} \big\{ (1+\beta_j+\gamma)^{W_j}\, (v_j+\beta_j)^{g_j+u_j} \big\}} \;, \tag{2.11}$$

where $W_j = S_{xj} + S_{yj} + k\,n_j$, $j=1,2$. Elimination of γ cannot be undertaken explicitly, but we find that $p\big(y_{n+1,1}, y_{n+1,2} \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ is proportional to

$$\int \frac{\Gamma(x_{n+1} + y_{n+1,1} + y_{n+1,2} + k)}{y_{n+1,1}!\; y_{n+1,2}!} \times \\ \times \frac{\prod_{j=1}^{2} \{\beta_j^{S_{yj}+y_{n+1,j}+g_j}\}}{(1+\beta_1+\beta_2+\gamma)^{x_{n+1}+y_{n+1,1}+y_{n+1,2}+k}} \times \\ \times \frac{\gamma^{(n+1)k+\ell}\exp\{-m\gamma\}}{\prod_{j=1}^{2}\big\{(1+\beta_j+\gamma)^{W_j}\,(v_j+\beta_j)^{g_j+u_j}\big\}}\; d\beta_1\, d\beta_2\, d\gamma \tag{2.12}$$

for $y_{n+1,1} = 0, 1, \ldots$ and $y_{n+1,2} = 0, 1, \ldots$. The joint predictive probability function may then be found numerically through three dimensional integration techniques.

Although no simple analytical form is available for (2.12) here, it is possible to obtain the marginal (but dependent) predictive probabilities in the case of vague second stage priors $(\ell, m, u_1, v_1, u_2, v_2 \to 0)$ explicitly, namely, for $j=1,2$,

$$p\big(y_{n+1,j} \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big) \;\propto\; \frac{B(x_{n+1}+ y_{n+1,j}+k,\, W_j)}{B(y_{n+1,j}+1,\, S_{ij}-1)}\;, \quad y_{n+1,j} = 0, 1, \ldots, \tag{2.13}$$

and these can easily be compared graphically.

2.4. Approximations

Alternatively, we might consider approximations through the use of, for example, posterior normality assumptions, Gibbs sampling or Laplace approximations to evaluate the predictive probabilities.

Noting that, as the sample size increases, the number of parameters in our model also increases, we surmise that problems may arise over assumptions of asymptotic normality of the overall posterior distribution, especially for the usual asymptotic normal approximation for the full posterior distribution in (2.10) — see, for example, Bernardo and Smith ([1, pp. 285–97]). A better result is likely from following O'Hagan's ([11, pp. 208]) suggestion of using a normal approximation for the reduced posterior $p\big(\theta_{n+1}, \beta_1, \beta_2, \gamma \,|\, x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big)$ alone, based on the posterior mode and modal dispersion matrix. We do not pursue this approach here, but further details can be found in Magalhães [9].

2.4.1. Gibbs sampling

The conditional distributions of θ_{n+1}, β_1, β_2 and γ follow from (2.11) in a straightforward manner, and, using rejection sampling with t iterations in each cycle, we obtain M random samples

$$\left(\theta^{(t)}_{n+1(\ell)}, \beta^{(t)}_{1(\ell)}, \beta^{(t)}_{2(\ell)}, \gamma^{(t)}_{(\ell)}\right) , \qquad \ell = 1, 2, ..., M \; .$$

The prediction function (2.9) can then be estimated using

$$\hat{p}\big(y_{n+1,1}, y_{n+1,2} \mid x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big) \;=\; \frac{1}{M} \sum_{\ell=1}^{M} \prod_{j=1}^{2} \frac{\mu_{j\ell}^{y_{n+1,j}} \, e^{-\mu_{j\ell}}}{y_{n+1,j}!} \;, \tag{2.14}$$

where $\mu_{j\ell} = \beta^{(t)}_{j\ell} \, \theta^{(t)}_{n+1(\ell)}$, $j=1,2$. If interest lies, say, in $Z = Y_{n+1,2} - Y_{n+1,1}$, we then need to derive the predictive distribution of Z.

Notice that, although it is necessary to generate values of $\gamma^{(t)}_{(\ell)}$ in this Gibbs routine, the values of this hyperparameter are not required further for our prediction problem.

2.4.2. Laplace approximation

Since the joint predictive probability function in (2.9) is a posterior expectation, which may be written, in generic form, as

$$E\big\{g(\psi) \,|\, data\big\} \;=\; \frac{\displaystyle\int g(\psi)\, L(data)\, p(\psi)\; d\psi}{\displaystyle\int L(data)\, p(\psi)\; d\psi} \;=\; \frac{\displaystyle\int \exp\{-nh^*(\psi)\}\; d\psi}{\displaystyle\int \exp\{-nh(\psi)\}\; d\psi} \;,$$

we may also use the Laplace approximation method; see, for example, Bernardo and Smith ([1, pp. 340–5]). In the posterior expectation above, ψ represents an unknown parameter and $L(data)$ is the likelihood function. Also, functions $h(\psi)$ and $h^*(\psi)$ are defined such that

$$-nh(\psi) = \ln p(\psi) + \ln L(data) \quad \text{and} \quad -nh^*(\psi) = \ln g(\psi) + \ln p(\psi) + \ln L(data) \,.$$

Again, we present the results for the special case of vague second stage priors. A good approximation for (2.9) is given by

$$\frac{1}{y_{n+1,1}!} \, \frac{1}{y_{n+1,2}!} \left(\frac{\sigma^*}{\tilde{\sigma}}\right) \exp\!\left[-n\Big\{h^*(\theta^*_{n+1}, \beta^*_1, \beta^*_2, \gamma^*) - h(\tilde{\theta}_{n+1}, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma})\Big\}\right] , \tag{2.15}$$

where

$$
\begin{aligned}
-nh(\theta_{n+1},\beta_1,\beta_2,\gamma) \;=\; & -\theta_{n+1}(1+\gamma) + (x_{n+1}+k)\,\ln\theta_{n+1} + (n+1)\,k\,\ln\gamma \\
& + \sum_{j=1}^{2}\Big\{S_{yj}\ln\beta_j - W_j\ln(1+\beta_j+\gamma)\Big\}\;,
\end{aligned}
$$

$$
\begin{aligned}
-nh^*(\theta_{n+1},\beta_1,\beta_2,\gamma) \;=\; & \\
=\; & -nh(\theta_{n+1},\beta_1,\beta_2,\gamma) + \sum_{j=1}^{2}\Big\{y_{n+1,j}(\ln\theta_{n+1}+\ln\beta_j) - \theta_{n+1}\,\beta_j\Big\}
\end{aligned}
$$

and where $\tilde\theta_{n+1}$, $\tilde\beta_1$, $\tilde\beta_2$, $\tilde\gamma$ and θ^*_{n+1}, β^*_1, β^*_2, γ^* are the modes of $-h$ and $-h^*$, respectively. The former are given by

$$
\text{(2.16)}\qquad
\begin{aligned}
\tilde\theta_{n+1} \;&=\; \frac{(x_{n+1}+T_x)\,(x_{n+1}+k)}{x_{n+1}+T_x+(n+1)k}\;, \qquad \tilde\gamma = \frac{(n+1)k}{x_{n+1}+T_x}\;, \\
\tilde\beta_j \;&=\; \frac{\big\{x_{n+1}+T_x+(n+1)k\big\}S_{yj}}{(x_{n+1}+T_x)\,(W_j-S_{yj})}\;, \quad j=1,2\;;
\end{aligned}
$$

whilst the latter are found iteratively from

$$
\begin{aligned}
\theta_{n+1}\left(1+\sum_{j=1}^{2}\beta_j+\gamma\right) \;&=\; x_{n+1} + \sum_{j=1}^{2} y_{n+1,j} + k\;, \\
\beta_j\left(\theta_{n+1}+\frac{W_j}{1+\beta_j+\gamma}\right) \;&=\; y_{n+1,j} + S_{yj}\;, \qquad j=1,2\;, \\
\gamma\left(\theta_{n+1}+\sum_{j=1}^{2}\frac{W_j}{1+\beta_j+\gamma}\right) \;&=\; (n+1)k\;.
\end{aligned}
$$

Finally, $\tilde\sigma$ and σ^* are the square roots of the inverse of the determinants of the appropriate hessian matrices of second order derivatives, namely

$$
\begin{aligned}
\tilde\sigma \;&=\; \big|n\,\nabla^2 h(\tilde\theta_{n+1},\tilde\beta_1,\tilde\beta_2,\tilde\gamma)\big|^{-\frac12}\;, \\
\sigma^* \;&=\; \big|n\,\nabla^2 h^*(\theta^*_{n+1},\beta^*_1,\beta^*_2,\gamma^*)\big|^{-\frac12}\;.
\end{aligned}
$$

Full details can be found in Magalhães [9].

3. ILLUSTRATION

In order to illustrate the different approximations, we consider the data shown in Table 1.

Table 1: Simulated data set of size $n = 20$.

x_i	δ_{i1}	y_i	x_i	δ_{i1}	y_i
7	0	3	11	0	22
8	0	8	6	0	13
9	0	1	9	0	10
13	0	16	6	0	10
5	1	1	10	0	16
2	1	2	17	0	16
13	0	12	3	1	0
4	1	2	2	1	2
6	0	12	2	1	1
7	0	4	8	0	11

These $n = 20$ data values were simulated from models with $\beta_1 = 0.3$ and $\beta_2 = 1.4$, with $a = 6$, and for a random selection of θ_i values. Note that $\hat{\beta}_1 = 0.44$ and $\hat{\beta}_2 = 1.25$ from (2.2), whilst the equivalent Robbins and Zhang [15] are $\beta_{1,n} = 0.22$ and $\beta_{2,n} = 1.46$ from (2.3).

Predictions are given for $y_{21,1}$ and $y_{21,2}$ corresponding to $x_{21} = 4$. The amended maximum likelihood estimates are now $\breve{\beta}_1 = 0.38$ and $\breve{\beta}_2 = 1.29$ from (2.7), whilst the Robbins and Zhang values are updated to $\beta^*_{1,n+1} = 0.21$ and $\beta^*_{2,n+1} = 1.53$ from (2.4).

In the analyses we assume a vague second stage prior $(\ell, m, u_1, v_1, u_2, v_2 \to 0)$. For such a case, specification of g_1 and g_2 is not necessary. We take $k = 6$ based on matching the first two marginal moments of the X_i's; see Dunsmore and Robson [2].

A clear picture emerges if we consider the marginal predictive functions for $y_{21,1}$ and $y_{21,2}$ separately. Figures 1 and 2 show the approximations from the two methods together with the exact forms from (2.13). Clearly, the Gibbs and Laplace methods provide excellent approximations to the exact distribution. Also shown in Figures 1 and 2 are the marginal predictive functions with the posterior normal approximations mentioned in Section 2.4. Normal approximation 1 refers to the Bernardo & Smith [1] approach and normal approximation 2 refers to O'Hagan's [11] approach. The anticipated problems with a sample size of only 20 manifest themselves, although perhaps not surprisingly O'Hagan's [11] suggestion, based on only four parameters, seems superior to the more usual posterior normal approximation, based on 26 parameters. Figures 1 and 2 also show that the predictive approach leads to more disperse distributions than the ones obtained through the plug-in method. This fact is not surprising because the predictive approach incorporates uncertainty about the parameters.

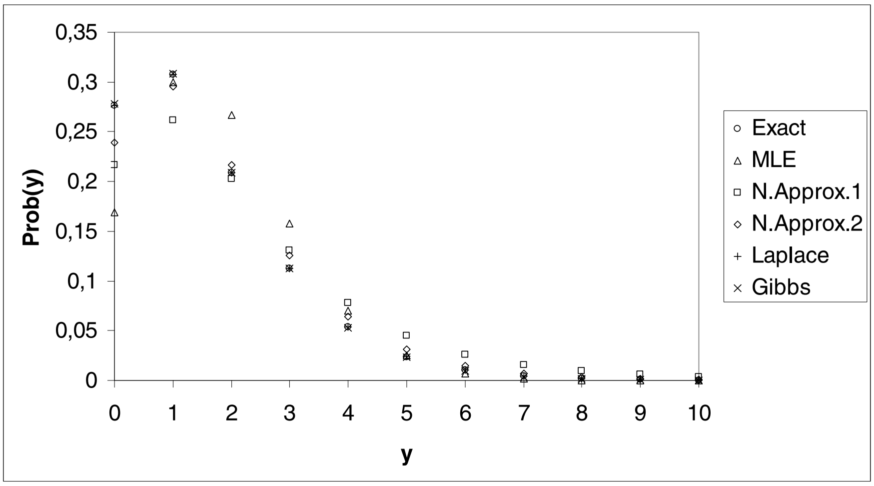


Figure 1: Comparison of the predictive functions $p\big(y_{21,1} \,|\, x_{21}, \boldsymbol{x}^{20}, \boldsymbol{y}^{20}\big)$ from the four approximations with the exact form in (2.13).

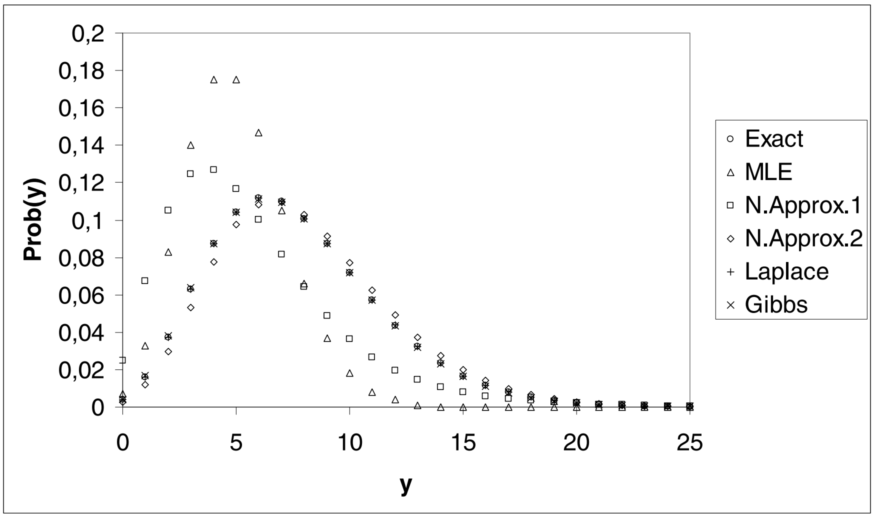


Figure 2: Comparison of the predictive functions $p(y_{21,2} \mid x_{21}, \boldsymbol{x}^{20}, \boldsymbol{y}^{20})$ from the four approximations with the exact form in (2.13).

We may conclude that the Gibbs and Laplace methods lead to excellent results when compared to the exact predictive distribution. The speed of the Laplace method, in comparison to Gibbs sampling, is a strong point in its favour.

Figure 3 compares the predictive functions for $Z = Y_{21,2} - Y_{21,1}$ for the Laplace method and the plug-in method of (2.5), and illustrates the unsatisfactory nature of the latter.

Similar conclusions were drawn in several other simulations.

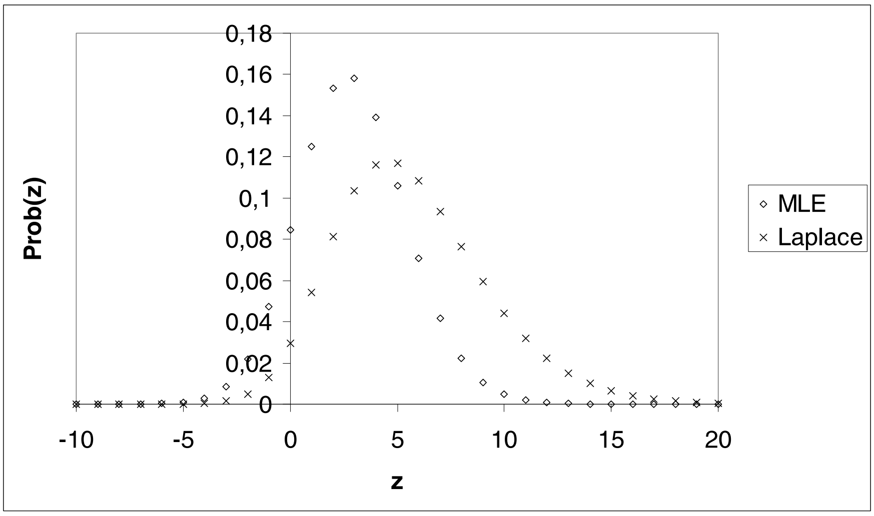


Figure 3: Comparison of the predictive functions for $Z = Y_{21,2} - Y_{21,1}$ from the Laplace method and the plug-in form in (2.8).

4. GENERALISATION TO $J \geq 2$ TREATMENTS

The models can be extended to the case of J treatments in a straightforward manner. Suppose that we can define mutually disjoint and exhaustive subsets C_1, C_2, ..., C_J of the non negative integers, such that treatment T_j is used for individual I if $x_i \in C_j$. We assume that Y_i is $Po(\beta_j \theta_i)$ if treatment T_j is used for individual i, $(i = 1, 2, ..., n{+}1;\ j = 1, 2, ..., J)$. Notice that the identification of subsets through cut-off points $a_1 < a_2 < ... < a_{J-1}$ is only one possible partition.

Robbins and Zhangs [15] method generalises (1.2) above to give consistent estimates

$$\beta_{jn} = \frac{\sum_{i=1}^{n} y_i\, I(x_i \in C_j)}{\sum_{i=1}^{n} x_i\, I(x_i{-}1 \in C_j)}$$

for $j = 1, 2, ..., J$. Similarly, with an obvious extension of the notation for δ_{ij}, n_j, S_{xj}, S_{yj}, T_x, T_y, the maximum likelihood estimates corresponding to (2.2) generalise in a simple way.

Numerical integration for the joint predictive probability

$$p\big(y_{n+1,1}, y_{n+1,2}, ..., y_{n+1,J} \mid x_{n+1}, \boldsymbol{x}^n, \boldsymbol{y}^n\big) ,$$

corresponding to (2.9) becomes impractical, but the Gibbs and Laplace methods provide approximations. Full details are again given in Magalhães [9].

Of interest now might be predictive probabilities associated with $\max(y_{n+1,1},\, y_{n+1,2},\, ...,\, y_{n+1,J})$. Within the Gibbs framework, one way of deriving these would be to consider the $y_{n+1,1},\, y_{n+1,2},\, ...,\, y_{n+1,J}$ as missing data and within each cycle to generate values of $y_{n+1,j}$ from a $Po(\beta_j\theta_{n+1})$ distribution, $j = 1, 2, ..., J$.

From the resulting samples $\big(y^{(t)}_{n+1,1(\ell)},\, y^{(t)}_{n+1,2(\ell)},\, ...,\, y^{(t)}_{n+1,J(\ell)}\big)$ it is then straightforward to approximate the probability that treatment T_j, say, provides the maximum response.

5. CONCLUSIONS

We have developed Bayesian predictive models for a Poisson errors in variables situation in which there are simple, multiplicative effects. Whilst standard numerical integration techniques, here in three dimensions, might be suitable for the determination of the appropriate predictive distributions, we have found that Laplace approximation and Gibbs sampling can provide alternative and reliable approaches. The use of the posterior normal approximations can be suspect because of the high dimensionality of the parameters, although O'Hagan's [11] approach improves matters somewhat.

Robbins and Zhang [15] also consider estimation in a binomial model, whilst Robbins [12] discusses the exponential case. Similar predictive frameworks can be developed for these situations. For example, in the model specification in Section 2 we might replace the Poisson assumptions (i) and (ii) by

(i) given r and θ_i, X_i is $Bi\left(r, \dfrac{\theta_i}{1+\theta_i}\right)$;

(ii) given s $\beta_1, \beta_2, \theta_i$ and x_i,

$$Y_i \text{ is } Bi\left(s, \frac{\beta_1\theta_i}{1+\beta_1\theta_i}\right) \text{ if treatment } T_1 \text{ is used ;}$$

$$Y_i \text{ is } Bi\left(s, \frac{\beta_2\theta_i}{1+\beta_2\theta_i}\right) \text{ if treatment } T_2 \text{ is used .}$$

Here, the odds ratios are θ_i for X_i and $\beta_1\theta_i$ or $\beta_2\theta_i$ for Y_i. Details of the predictive distributions for the binomial and exponential cases can be found in Magalhães [9].

ACKNOWLEDGMENTS

The authors acknowledge the support for F.J.M. Magalhães from Instituto Politécnico do Porto and Universidade Portucalense, Portugal, during his period of postgraduate studies at the University of Sheffield.

REFERENCES

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

[2] Dunsmore, I.R. and Robson, D.J. (1997). Prediction in some Poisson errors in variables models, *Scand. J. Statist.*, **24**, 543–554.

[3] Gaver, D.P. and O'Muircheartaigh, I.G. (1987). Robust empirical Bayes analyses of event rates, *Technometrics*, **29**, 1–15.

[4] Gelfand, A.E.; Hills, S.E.; Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, *J. Amer. Statist. Ass.*, **85**, 972–985.

[5] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Ass.*, **85**, 398–409.

[6] Godambe, V.P. and Kunte, S. (1993). Optimal estimation under biased allocation of treatments, *Biometrika*, **80**, 797–806.

[7] Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion), *J. Statist. Plan. Inf.*, **22**, 137–172.

[8] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *Ann. Math. Statist.*, **27**, 887–906.

[9] Magalhães, F.J.M. (1997). *Prediction in some errors in variables in variables models*, Unpublished Ph.D. Thesis, University of Sheffield.

[10] Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations, *Econometrica*, **16**, 1–32.

[11] O'Hagan, A. (1994). *Bayesian Inference: Kendall's Advanced Theory of Statistics, Volume 2B.*, Edward Arnold, London.

[12] Robbins, H. (1988). *The u, v method of estimation.* In "Statistical Decision Theory and Related Topics IV, **1**" (S.S. Gupta and J.O. Berger, Eds.), Springer-Verlag, New York, 265–270.

[13] Robbins, H. and Zhang, Cun-Hui (1988). Estimating a treatment effect under biased sampling, *Proc. Natl. Acad. Sci. USA*, **85**, 3670–3672.

[14] Robbins, H. and Zhang, Cun-Hui (1989). Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug, *Proc. Natl. Acad. Sci. USA*, **86**, 3003–3005.

[15] Robbins, H. and Zhang, Cun-Hui (1991). Estimating a multiplicative treatment effect under biased allocation, *Biometrika*, **78**, 349–354.

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 **H** ($p < 0.05$) or in favor of **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 **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

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

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 $\mathbf{H}$ and $\mathbf{A}$, 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 $\mathbf{H}$ and $\mathbf{A}$, 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

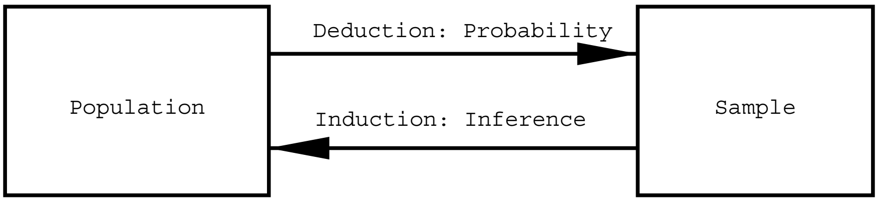


Figure 1: Probability and Statistics Harmonization.

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:

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

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

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

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.

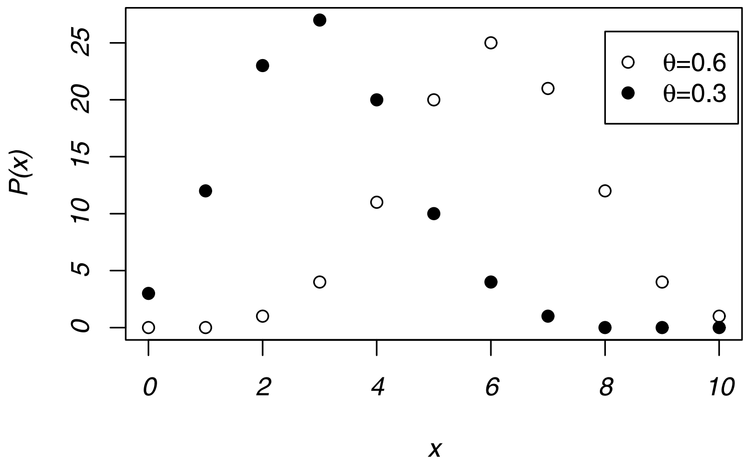


Figure 2: Binomial probability functions for $n = 10$.

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 $\theta = .3$.

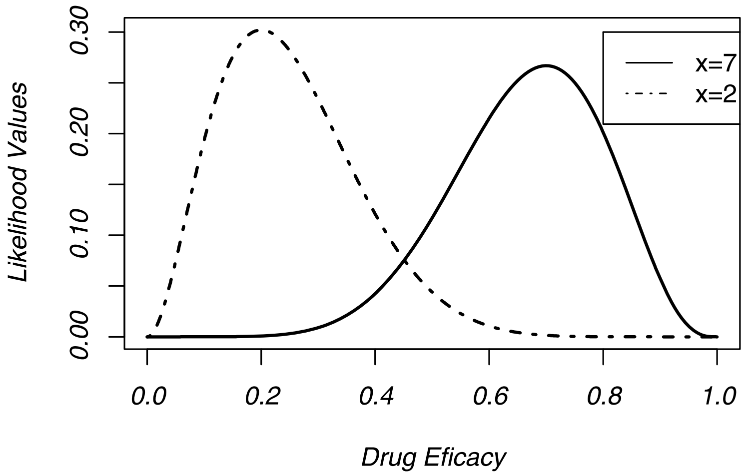


Figure 3: Binomial likelihood functions for $n = 10$.

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

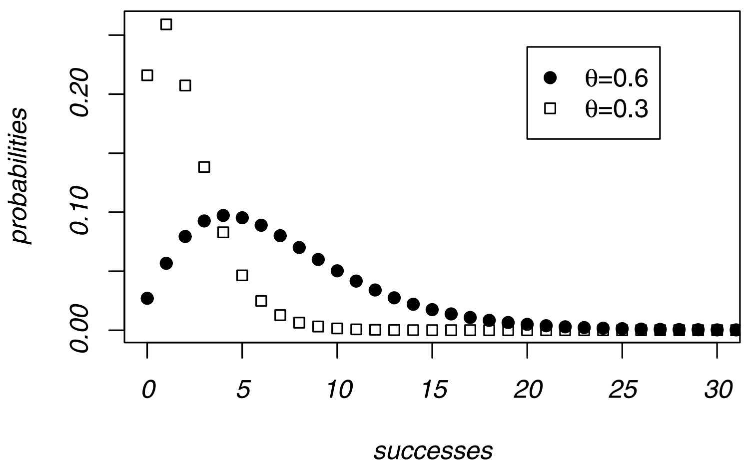


Figure 4: Negative Binomial probability function for $k = 3$.

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.

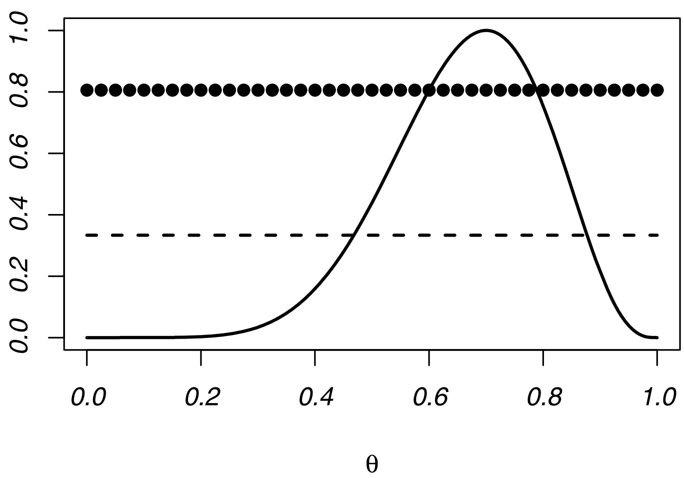


Figure 5: Relative Likelihood, and 1/3 and .8057 Plausible Levels.

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 $\mathbf{H}: \theta = \theta_0$ or $\mathbf{H}: \theta = \theta_{00}$. The probability of the complement of this credibility interval is an index (like a *p-value*) of evidence in favor of **H**; see [37] and [27] for more on this measure of evidence. For the two cases presented here, the evidence in favor of **H** is 48.35% for $\mathbf{H}: \theta = .6$ and 4.10% for $\mathbf{H}: \theta = .4$.

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) **H**. The strength of evidence of the data x in favor of **H** against **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 **H** without mentioning the alternative **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

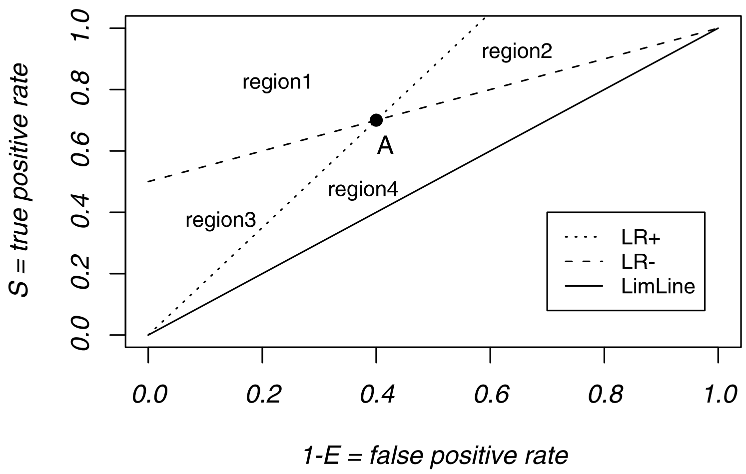


Figure 6: Regions of Preference: $A = (1{-}E; S) = (.4, .7)$.

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 obtain 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" (2^{nd} 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.

REVSTAT – STATISTICAL JOURNAL

Background

Statistical Institute of Portugal (INE), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, publishing three times a year papers containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided to publish papers also in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work.

At the time, the Editorial Board was mainly composed by Portuguese university professors, being now composed by national and international university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal.

In 2001, the *Revista de Estatística* published three volumes special issue containing extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

The name of the Journal has been changed to REVSTAT – STATISTICAL JOURNAL, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

— The editorial policy will focus on publishing research articles at the highest level in the domains of Probability and Statistics with emphasis on the originality and importance of the research.

— All research articles will be refereed by at least two persons, one from the Editorial Board and another, external.

— The only working language allowed will be English.

— Two volumes are scheduled for publication, one in June and the other in November.

— On average, four articles will be published per issue.

Aims and Scope

The aim of REVSTAT is to publish articles of high scientific content, in English, developing innovative statistical scientific methods and introducing original research, grounded in substantive problems.

REVSTAT covers all branches of Probability and Statistics. Surveys of important areas of research in the field are also welcome.

Abstract/indexed in

REVSTAT is expected to be abstracted/indexed at least in *Current Index to Statistics*, *Mathematical Reviews*, *Statistical Theory and Method Abstracts*, and *Zentralblatt für Mathematic*.

Instructions to Authors, special-issue editors and publishers

Papers may be submitted in two different ways:

— By sending a paper copy to the Executive Editor and one copy to one of the two Editors or Associate Editors whose opinion the author(s) would like to be taken into account, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

— By sending a paper copy to the Executive Editor, together with a postscript or a PDF file of the paper to the e-mail: revstat@fc.ul.pt.

Submission of a paper means that it contains original work that has not been nor is about to be published elsewhere in any form.

Submitted manuscripts (text, tables and figures) should be typed only in black, on one side, in double spacing, with a left margin of at least 3 cm and not have more than 30 pages.

The first page should include the name, affiliation and address of the author(s) and a short abstract with the maximum of 100 words, followed by the key words up to the limit of 6, and the AMS 2000 subject classification.

Authors are obliged to write the final version of accepted papers using LaTeX, in the REVSTAT style.

This style (REVSTAT.sty), and examples file (REVSTAT.tex), which may be download to *PC Windows System* (Zip format), *Mackintosh, Linux* and *Solaris Systems* (StuffIt format), and *Mackintosh System* (BinHex Format), are available in the REVSTAT link of the National Statistical Institute's Website: http://www.ine.pt/revstat.html

Additional information for the authors may be obtained in the above link.

Accepted papers

Authors of accepted papers are requested to provide the LaTeX files and also a postscript (PS) or an acrobat (PDF) file of the paper to the Secretary of REVSTAT: liliana.martins@ine.pt.

Such e-mail message should include the author(s)'s name, mentioning that it has been accepted by REVSTAT.

The authors should also mention if encapsulated postscript figure files were included, and submit electronics figures separately in .tiff, .gif, .eps or .ps format. Figures must be a minimum of 300 dpi.

Also send always the final paper version to:

Maria José Carrilho
Executive Editor, REVSTAT – STATISTICAL JOURNAL
Instituto Nacional de Estatística
Av. António José de Almeida
1000-043 LISBOA
PORTUGAL

Copyright and Reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, the INE, in order to ensure the widest possible dissemination of information, namely through the National Statistical Institute's Website (http://www.ine.pt).

After assigning the transfer copyright form, authors may use their own material in other publications provided that the REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Authors of articles published in the REVSTAT will be entitled to one free copy of the respective issue of the Journal and twenty-five reprints of the paper are provided free. Additional reprints may be ordered at expenses of the author(s), and prior to publication.