Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 12

11-1-2004

On Comparison of Hypothesis Tests in the Bayesian Framework without Loss Function

Vladimir Gercsik
Russian Academy of Science

Mark Kelbert
University of Wales-Swansea, United Kingdom, M.Kelbert@swansea.ac.uk

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 399-405

1538 – 9472/04/$95.00

On Comparison of Hypothesis Tests in the Bayesian Framework without Loss Function

Vladimir Gercsik
International Institute of Earthquake
Prediction Theory & Mathematical Geophysics
Russian Academy of Science

Mark Kelbert
Department of Mathematics
University of Wales-Swansea
United Kingdom

The problem is how to compare the quality of different hypothesis tests in a Bayesian framework without introducing a loss function. Three different linear orders on the set of all possible hypothesis tests are studied. The most natural order estimates the Fisher information between indicators of event and decision.

Key words: Bayesian risk, Neyman-Pearson test, Fisher information, tests of independency

Introduction

It is well-known that no universal measure of a hypothesis test quality exists in statistics. In the Bayesian framework a linear ordering of tests is possible for a given loss function. It is said that a test is optimal if it minimizes the Bayesian risk.

However, the selection of a loss function is often somewhat arbitrary, and it is not always natural to measure the losses under different types of errors in the same units. For example, the cost of prevention measures in earthquake prediction is naturally expressed in money units. However, the losses from an earthquake including the psychological traumas, maiming and even the loss of human lives could be hardly expressed in money terms. Even if this expression is imposed, any estimation of these losses in money terms would depend a great deal on the variable economical and political situation. This loss function hardly looks as neutral and scientifically unbiased.

The subject of interest is in a situation when the loss function is unknown but the quality of any two hypothesis tests should be quantitatively compared. It turned out that all hypothesis tests could be linearly ordered at least in three different ways:

Let p and $1-p$ be the Bayesian probabilities of random experiments $\omega\in\Omega$ with cumulative distribution functions F_1 and F_2, respectively. A test is defined by a function $\mathbf{\Phi}(\omega) = 1$ on the critical set B and 0 on B^c. If $\mathbf{\Phi}(\omega) = 1$ then the alternative F_2 is accepted, and the hypothesis F_1 is accepted in the case $\mathbf{\Phi}(\omega) = 0$. Clearly, the problem is symmetric with respect to interchange of hypothesis F_1 and F_2 and a simultaneous interchange of $\mathbf{\Phi}(\omega)$ and $1 - \mathbf{\Phi}(\omega)$.

The type I error is denoted (i.e., the probability to accept F_2 when F_1 is true) by α_1, and the type II error (i.e., the probability to accept F_1 when F_2 is true) by α_2. Considered are only *unbiased* tests (Barra, 1981), i.e., assume that $\alpha_1 + \alpha_2 \leq 1$. (If this condition is violated one could get an unbiased test by selecting B^c instead of B as a critical set.) Consider a random variable $X_1 = 0$ if F_1 is true, $X_1 = 1$ if F_2 is true, and call it *an indicator of events*. In a similar manner we define a random variable $X_2 = 0$ if a test accepts F_1, and $X_2 = 1$ if the test accepts F_2, called *an indicator of decision*. The joint distribution $\mathbf{P}(x_1, x_2)$ is defined by the relations $\alpha_1 = \mathbf{P}(X_2 = 1 \mid X_1 = 0)$, $\alpha_2 = \mathbf{P}(X_2 = 0 \mid X_1 = 1)$, in particular,

Vladimir Gercsik is a Senior Researcher in geophysics and earthquake prediction. Mark Kelbert is a Reader in statistics. Email him at M.Kelbert@swansea.ac.uk.

$$\mathbf{P}(X_2 = 0, X_1 = 0) = p(1 - \alpha_1),$$
$$\mathbf{P}(X_2 = 1, X_1 = 0) = p\alpha_1, \quad (1)$$

$$\mathbf{P}(X_2 = 0, X_1 = 1) = (1 - p)\,\alpha_2,$$
$$\mathbf{P}(X_2 = 1, X_1 = 1) = (1 - p)(1 - \alpha_2). \quad (2)$$

The marginal one-dimensional probabilities take the form:

$$\mathbf{P}(X_1 = 0) = p,\ \mathbf{P}(X_1 = 1) = (1 - p),$$
$$\mathbf{P}(X_2 = 0) = p(1 - \alpha_1) + \alpha_2 (1 - p), \quad (3)$$

$$\mathbf{P}(X_2 = 1) = p\alpha_1 + (1 - p)(1 - \alpha_2). \quad (4)$$

Clearly, the worst possible unbiased test is determined by the condition that indicator of event X_1 and indicator of decision X_2 are independent. On the other extreme, an ideal test (normally, it does not exist) is one that provides the correct solution without any errors. Generically, the quality of a test is measured by some non-negative function of X_1 and $X_2 = \mathbf{\Phi}(\omega)$ which takes the value 0 iff X_1, X_2 are independent, and the value 1 iff $X_1 = X_2$.

Measuring the quality of a test

Any of the following well-known Rachev (1991) functions used in tests of independency is acceptable as a measure of the quality of a test.

$$\beta_1(\mathbf{\Phi}) = \sup_{x_1, x_2} \left|\mathbf{P}(x_1, x_2) - \mathbf{P}(x_1)\mathbf{P}(x_2)\right|, \quad (5)$$

$$\beta_2(\Phi) = \mathbf{E}_{x_1}\left[\sup_{x_2}\left|\mathbf{P}(x_2 \mid x_1) - \mathbf{P}(x_2)\right|\right] \quad (6)$$

($\mathbf{E}_{x_1}$ stands for the expectation with respect of distribution of X_1). It is easy to check that

$$\beta_1(\mathbf{\Phi}) = \beta_2(\mathbf{\Phi}) = p(1 - p)(1 - \alpha_1 - \alpha_2). \quad (7)$$

The quality of a test is measured by $\beta = (1 - \alpha_1 - \alpha_2)$. This is quite popular in practice as the Bayesian risk $\mathbf{R} = \mathbf{E}[w] = \alpha_1 + \alpha_2$ appears for the simplest loss function $w(x_1, x_2)$: $w(0,0) = w(1,1) = 0$, $w(0,1) = w(1,0) = 1$.

Another possibility to test the independence is to consider the maximal correlation coefficient

$$\rho(\mathbf{\Phi}) = \sup \mathbf{E}\left[\phi_1(X_1)\phi_2(X_2)\right] \quad (8)$$

where sup is taken over the set of functions ϕ_1, ϕ_2 such that $\mathbf{E}[\phi_i(X_i)] = 0$, $\sigma^2[\phi_i(X_i)] = \mathbf{E}[\phi_i(X_i)^2] = 1$, $i = 1,2$. Clearly, $\rho(\mathbf{\Phi}) = 0$ iff X_1 and X_2 are independent. As function $\phi_i(x)$, i =1,2 could take only two values $\phi_i(0)$ and $\phi_i(1)$, and these values are defined in a unique way by the conditions imposed, the following relation holds

$$\rho(\mathbf{\Phi}) = \mathbf{E}\left[\frac{(X_1 - \mathbf{E}(X_1))}{\sigma(X_1)}\frac{(X_2 - \mathbf{E}(X_2))}{\sigma(X_2)}\right] \quad (9)$$

A straightforward computation yields:

$$\rho(\mathbf{\Phi}) = \beta\sqrt{\frac{p(1-p)}{P(1-P)}} \quad (10)$$

where

$$\mathbf{P} = \mathbf{P}(X_2 = 0) = \alpha_2(1 - p) + (1 - \alpha_1)p = \alpha_2 + \beta p.$$

The correlation coefficient $\rho(\mathbf{\Phi})$ is non-negative for any unbiased test; it equals to 1 when $X_1 = X_2$.

Perhaps, the most interesting way to measure the quality of a test is to consider an information $I(\mathbf{\Phi})$ in the indicator of solution $X_2 = \mathbf{\Phi}$ about the indicator of event X_1. To formalize this idea consider

$$I(\mathbf{\Phi}) = \sum_{x_1=0,1}\sum_{x_2=0,1}\mathbf{P}(x_1, x_2)\log_2\frac{\mathbf{P}(x_1, x_2)}{\mathbf{P}(x_1)\mathbf{P}(x_2)} \quad (11)$$

which equals to $I(\mathbf{\Phi}) = S(X_1) - \mathbf{E}[S(X_1 \mid X_2)]$.

$S(X_1)$ stands for the Fisher information of the prior distribution $\mathbf{P}(x_1)$, $S(X_1) = p\log_2 p + (1 - p)\log_2(1 - p)$. Intuitively, it means that in a random trial with n outcomes where hypothesis F_1 and F_2 appear with probabilities p and $1 - p$, respectively, there are $\approx 2^{nS}$ quite

probable outcomes, and the rest could be neglected as $n \to \infty$. Next, $S(X_1 \mid X_2)$ is the *conditional entropy* of the conditional distribution $\mathbf{P}(x_1 \mid x_2)$ under condition that the decision x_2 is taken.

Therefore, the information $I(\mathbf{\Phi})$ equals to the mean reduction of uncertainty obtained by the using of the test $\mathbf{\Phi}$. Clearly, for an ideal test without any errors $I(\Phi)$ takes its maximal value $S(X_1)$, and $I(\mathbf{\Phi}) = 0$ iff X_1 and X_2 are independent.

It is interesting to note that a sequence of random events $X_1^{(n)}$ and decisions $X_2^{(n)}$ =1,2,… can be treated as a message transmission over a channel without noise. In this interpretation the observation $x_1^{(n)}(\omega)$ could be treated as the coding of the random outcome F_i, $i = 1,2$, and the decision $x_2^{(n)}(\omega)$ as its decoding. The maximization of $I(\mathbf{\Phi})$ means that the optimal decoding is applied.

Now following is proved:

Lemma 1. Fix the Type I error probability α_1. Then, the Neyman-Pearson test $\mathbf{\Phi}^*$ minimizing the Type II error probability α_2, maximizes also the information $I(\mathbf{\Phi}^*)$among all unbiased tests.

Proof. A straightforward computation yields

$$\frac{\partial I}{\partial \alpha_2} = (1-p)\left[\log_2\left(1-p+\frac{\alpha_1 p}{1-\alpha_2}\right)-\log_2\left(1-p+\frac{(1-\alpha_1)p}{\alpha_2}\right)\right] \tag{13}$$

This derivative is non-positive for an unbiased test with $\alpha_1 + \alpha_2 \le 1$. Hence, the information $I(\mathbf{\Phi})$ is maximal for a minimal possible value of α_2, i.e. for the test $\mathbf{\Phi}^*$. A symmetric statement with interchanging of α_1 and α_2 is also true.•

The same property holds also for $\beta(\mathbf{\Phi})$ and $\rho(\mathbf{\Phi})$; this is immediate for $\beta(\mathbf{\Phi})$, and follows from the equality

$$\frac{\partial \rho}{\partial \alpha_2} = -\frac{\alpha_1 P + (1-\alpha_1)(1-P)}{P(1-P)}\sqrt{\frac{p(1-p)}{P(1-P)}} \le 0 \tag{14}$$

in the case of $\rho(\mathbf{\Phi})$.

Next, observe that the Fisher information $I(\mathbf{\Phi})$ is a convex function of α_1 and α_2 for all $0 < \alpha_1 < 1$, $0 < \alpha_2 < 1$, $0 < p < 1$. The prove this, it suffices to compute

$$\frac{\partial^2 I}{\partial \alpha_2^2} = p\left[\frac{P(1-P) - p\alpha_1(1-\alpha_1)}{\alpha_1(1-\alpha_1)P(1-P)}\right], \tag{15a}$$

$$\frac{\partial^2 I}{\partial \alpha_1^2} = (1-p)\left[\frac{P(1-P) - p\alpha_2(1-\alpha_2)}{\alpha_2(1-\alpha_2)P(1-P)}\right], \tag{15b}$$

$$\frac{\partial^2 I}{\partial \alpha_1 \partial \alpha_2} = \frac{p(1-p)}{P(1-P)}, \tag{15c}$$

$$\det\left\|\frac{\partial^2 I}{\partial \alpha_i \partial \alpha_j}\right\| = \frac{p(1-p)(1-\alpha_1-\alpha_2)^2}{P(1-P)\alpha_1(1-\alpha_1)\alpha_2(1-\alpha_2)} > 0. \tag{15d}$$

This convexity property implies that a randomized test $\overline{\mathbf{\Phi}} = \int_{\mathbf{\Phi}\in A} \mathbf{\Phi}\mathbf{P}(d\mathbf{\Phi})$ where $\mathbf{P}(d\Phi)$ is a probability measure on a suitable set A, $\mathbf{P}(A)=1$, can not be optimal in the sense of Fisher information. Indeed, Jensen's inequality yields

$$I(\overline{\mathbf{\Phi}}) = I\left(\int_{\mathbf{\Phi}\in A} \mathbf{\Phi}\mathbf{P}(d\mathbf{\Phi})\right) \le \int_{\mathbf{\Phi}\in A} I(\mathbf{\Phi})\mathbf{P}(d\mathbf{\Phi}) \le \sup_{\Phi\in A} I(\mathbf{\Phi}) \tag{16}$$

Hence, there always exists a non-randomized test Φ' such that $I(\overline{\mathbf{\Phi}}) \le I(\mathbf{\Phi}')$. Clearly, a similar statement holds for $\beta(\Phi)$ as well. However, it is not true generically for $\rho(\Phi)$.

Fig.1 demonstrates this presenting the surface $\rho(\alpha_1,\alpha_2)$ for $p = 1/3$. Fig.2 presents the surface $I(\alpha_1, \alpha_2)$ for $p = 1/3$. Each of the characteristics $\beta(\Phi)$,$\rho(\Phi)$ and $I(\Phi)$ divides the set of all tests into equivalence classes with the same value of this characteristic inside the class, and defines a linear order between different equivalence classes.

Numerical Examples

Figure 3 presents the results of computation of optimal tests with respect of three criteria as above. As hypothesis F_1 and F_2, select the normal distributions with pdf's

$$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x+0.5)^2}{2}\right), \qquad f_2(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-0.5)^2}{2}\right), \tag{17}$$

respectively. The prior probability of appearing F_1 is $p = 1/3$. Fig.3 presents the type II error $\alpha_2 = g(\alpha_1)$ for Neyman-Pearson's problem in the cases $\sigma^2 = 0.2$, $\sigma^2 = 0.6$ and $\sigma^2 = 1$. Each of these curves represents all the tests of the form

$$\mathbf{\Phi}(x) = \begin{cases} 1, & \text{if } f_2(x) \ge \lambda f_1(x) \\ 0, & \text{if } f_2(x) < \lambda f_1(x) \end{cases} \tag{18}$$

with different λ which are optimal for Neyman-Pearson problem, i.e., minimizes α_2 for a given value of α_1. These curves of errors serve as a boundary of a convex domain of errors for all possible tests. The points are indicated (α_1, α_2) on the boundary where each of three characteristics of quality as above achieves its maximum. Clearly, the position of maximum is distinct in all three cases.

Checking Hypothesis of Uselessness of a Test

It is desirable to use empirical data for checking the uselessness of a test $\mathbf{\Phi}$. In the case of independence of the indicator of event and the indicator of decision any of three characteristics of quality as above equals 0. To reject the hypothesis of useless of test $\mathbf{\Phi}$ with some level of confidence it is sufficient to demonstrate that $\alpha_1 + \alpha_2 < 1$ with this level of confidence.

Consider a series of n independent trials where the number M of appearance of distribution F_2 equals m, the number L of selection of distribution F_2 by the test Φ equals l, and the number of correct choices K of F_2 by the test Φ equals k. The hypothesis of uselessness is formalized in the following form of H_0: $\alpha_1+\alpha_2=1$. If H_0 is true, then (with unknown probability p)

$$\mathbf{P}(X_1 = 0) = p, \qquad \mathbf{P}(X_2 = 0) = \alpha_2(1-p) + (1-\alpha_1)p \equiv \alpha_2 \tag{19}$$

and X_1 and X_2 are independent.

Probabilities $\mathbf{P}(X_2=0, X_1=0)$, $\mathbf{P}(X_2=1, X_1=0)$, $\mathbf{P}(X_2=0, X_1=1)$ and $\mathbf{P}(X_2=1, X_1=1)$ are estimated by the empirical frequencies $(n - m - 1 + k) / n$, $(1 - k) / n$, $(m - k) / n$ and k / n, respectively. Hence, the estimates of conditional probabilities $\alpha_1 = \mathbf{P}(X_2 = 1 \mid X_1 = 0)$ and $\alpha_2 = \mathbf{P}(X_2 = 0 \mid X_1 = 1)$ looks like $(1 - k) / (n - m)$ and $(m - k) / m$, respectively.

Clearly, the inequality $k > ml / n$ corresponds to the alternative H_1: $\alpha_1 + \alpha_2 < 1$. It means that the critical sets have the form $\{K > K^*\}$ for given values $M = m$ and $L = 1$.

If hypothesis H_0 is true then the independence of X_1 and X_2 and independence of different trials imply an explicit expression for test size

$$\mathbf{P}(K > K^*, M = m, L = l) = \binom{n}{m} p^m (1-p)^{n-m} \times \sum_{k=K^*}^{\min\{m,l\}} \binom{m}{k} q^k (1-q)^{m-k} \binom{n-m}{l-k} q^{l-k} (1-q)^{n-m-l+k} = \binom{n}{m} p^m (1-p)^{n-m} q^l (1-q)^{n-l} \sum_{k=K^*}^{\min\{m,l\}} \binom{n-m}{l-k}\binom{m}{k}. \tag{20}$$

Figure 1. Dependence of correlation coefficient on errors.

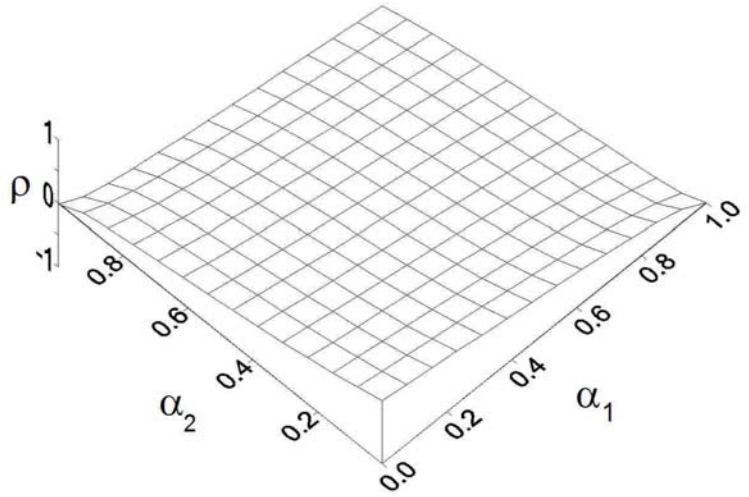

Figure 2. Dependence of Fisher information coefficient on errors.

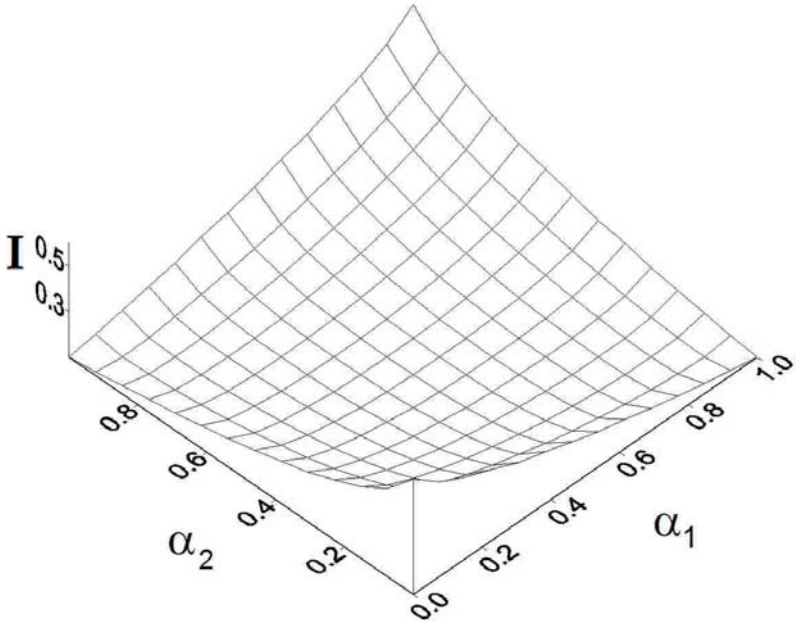

Figure 3. Location of maximum points for different characteristics of test efficiency.

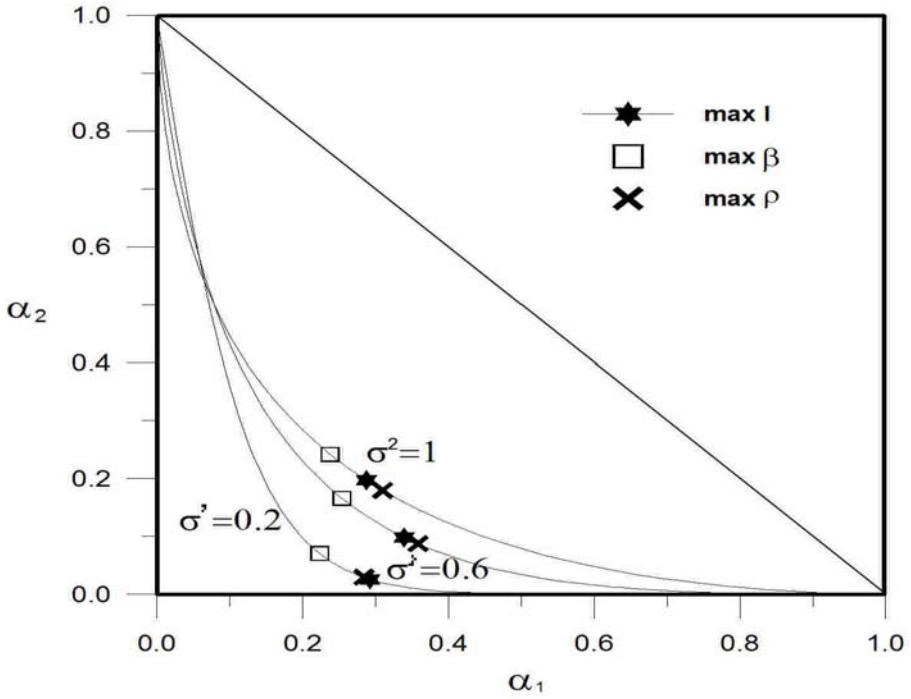

Figure 4. Dependence of maxima locations for different characteristics of test efficiency on variance.

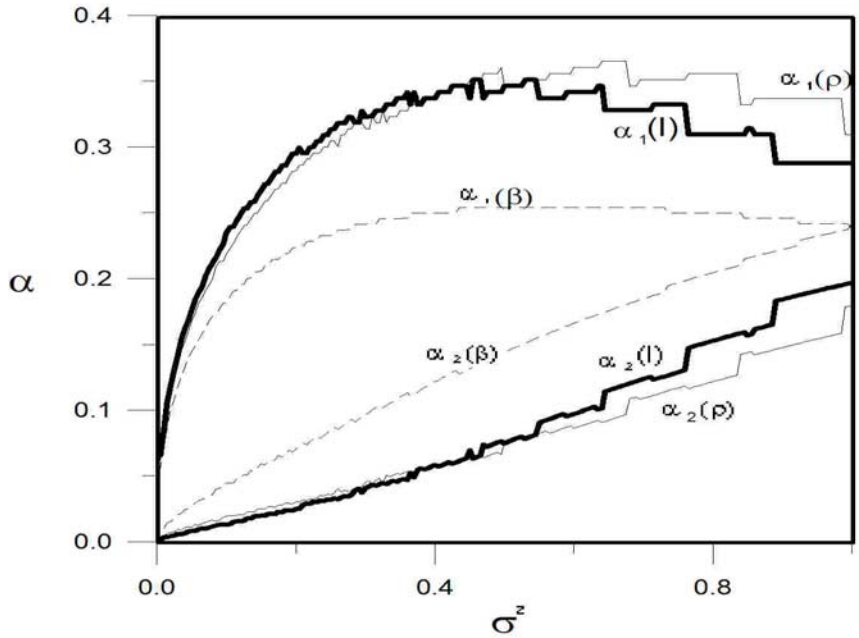

Fig. 4 presents the dependence of locations of maximum points $\alpha_i(\beta)$, $\alpha_i(\rho)$ and $\alpha_i(I)$, $i = 1,2$, for three characteristics of quality as above as functions of σ^2. Observe that the Type II error tends to 0 as $\sigma^2 \to 0$, as the PDF f_2 tends to a delta-function located at the point $x = 0.5$.

Here $q = \alpha_2$. Using the equalities

$$\mathbf{P}(M = m) = \binom{n}{m} p^m (1-p)^{n-m}, \qquad \mathbf{P}(L = l) = \binom{n}{l} q^l (1-q)^{n-l}, \tag{21}$$

obtain

$$\mathbf{P}(K \geq K^* \mid M = m, L = l) = \sum_{k=K^*}^{\min\{m,l\}} \frac{\binom{m}{k}\binom{n-m}{l-k}}{\binom{n}{l}} \tag{22}$$

This conditional probability represents the level of confidence for the critical region $\{K > K^*\}$ for given values $M = m$ and $L = l$.

References

Barra J. R. (1981). *Mathematical basis of statistics*. New York: Academic Press.

Rachev S. (1991). *Probability metrics and the stability of stochastic models*. Chichester: John Wiley.