P. MARTIN-LÖF

EXACT TESTS, CONFIDENCE REGIONS AND ESTIMATES

ABSTRACT. This paper proposes a uniform method for constructing tests, confidence regions and point estimates which is called exact since it reduces to Fisher's so-called exact test in the case of the hypothesis of independence in a 2 x 2 contingency table. All the wellknown standard tests based on exact sampling distributions are instances of the exact test in its general form. The likelihood ratio and χ^2 tests as well as the maximum likelihood estimate appears as asymptotic approximations to the corresponding exact procedures.

A *statistic* $t(x)$ on a discrete sample space X is simply a function which is such that the set

$$X_t = \{x \mid t(x) = t\}$$

is finite for all choices of t in the range of $t(x)$. Let $f(t)$ denote the number of elements in X_t. To say that an outcome x is described by a statistic $t(x)$ is tantamount to saying that x can be considered as drawn at random from the set X_t where t = observed value of $t(x)$.

A *reductive hypothesis* asserts that for the purpose of describing an outcome x the statistic $t(x)$ may be reduced to a statistic $u(x) = u(t(x))$ which induces a coarser partitioning of the sample space. Let

$$g(u) = \sum_{u(t(x))=u} 1 = \sum_{u(t)=u} f(t)$$

denote the number of outcomes x such that $u(x) = u(t(x)) = u$.

According to the *exact test* criterion, the smaller the number $f(t(x))$ of outcomes is that realize the observed value of $t(x)$, the greater is the discrepancy between the outcome x and the hypothesis that $t(x)$ can be reduced to $u(x) = u(t(x))$. Thus the critical level with respect to the exact test equals $\epsilon(t(x))$ where

$$\epsilon(t) = \sum_{f(t(x'))\leqslant f(t)} \frac{1}{g(u)} = \sum_{f(t')\leqslant f(t)} \frac{f(t')}{g(u)}.$$

In the continuous case, X will typically be (an open subset of) a Euclidean space and $t(x)$ a continuously differentiable mapping into a Euclidean space

Synthese **36** (1977) 195–206.

of lower dimension such that the determinant $\det(JJ')$ where

$$J = \left(\frac{\partial t_i}{\partial x_j}\right)$$

vanishes at most on a set of Lebesgue measure zero which we assume to be deleted from X. The function $f(t)$ may then be defined by the formula

$$f(t) = \int_{X_t} \frac{1}{\sqrt{\det(JJ')}}\, d\sigma_t$$

where σ_t is the surface measure on $X_t = \{x \mid t(x) = t\}$. With this change in the definition of $f(t)$, what has been said above carries over immediately to the continuous case. In particular, the exact test orders the sample points x on the surface $u(x) = u(t(x)) = u$ according to their associated values of $f(t(x))$.

Now, it is a *fact* that all the wellknown standard tests based on exact sampling distributions can be derived from the exact test criterion formulated above. In particular, this is so for

the test of independence in a contingency table,

tests of homogeneity for the binomial, multinomial, Poisson, geometric, normal and exponential distributions as well as for Rasch's item analysis,

Student's t-test,

various tests in the analysis of variance,

Hotelling's multivariate T-test and

the Rayleigh test.

For a detailed verification of this fact, see Martin-Löf (1970). Here I shall only consider two very simple examples.

Example 1. Test of independence in a 2 x 2 contingency table. Let

	1	2	total
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	$n_{..} = n$

be an ordinary four fold table and put

$$t = (n_{11}, n_{12}, n_{21}, n_{22})$$

and

$$u = u(t) = (n_{1.}, n_{2.}, n_{.1}, n_{.2}).$$

Then

$$f(t) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}$$

and

$$g(u) = \frac{n!}{n_{1.}!n_{2.}!} \frac{n!}{n_{.1}!n_{.2}!}$$

so that the exact test criterion tells us to reject the hypothesis that t can be reduced to u, which is the usual hypothesis of independence, provided the hypergeometric probability

$$\frac{f(t)}{g(u)}$$

is too small. This is Fisher's so-called exact test of independence in a 2 x 2 table.

Example 2. Testing $\mu = 0$ for a normal sample. Let $x_1, \ldots, x_n$ be a sample from a normal distribution and put

$$t = (t_1, t_2) = \left(\sum_1^n x_i, \sum_1^n x_i^2\right)$$

and

$$u = \sum_1^n x_i^2.$$

Then

$$f(t) = \frac{\pi^{(n-1)/2}}{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)} \left(t_2 - \frac{t_1^2}{n}\right)^{(n-3)/2}$$

and

$$g(u) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}\right)} u^{(n/2)-1}.$$

The exact test criterion tells us to reject the hypothesis that t can be reduced to u if the ratio

$$\frac{f(t)}{g(u)}$$

is too small, or, equivalently, if

$$\frac{|\bar{x}|}{s/\sqrt{n}}$$

is too big. So we have recovered Student's t-test for the hypothesis that $\mu = 0$ for a normal sample.

I want to supplement the abovementioned fact by the following *proposal.* Simply accept as a fundamental principle that the smaller the value of $f(t(x))$, the more does the observation x contradict the hypothesis that $t(x)$ can be reduced to $u(x) = u(t(x))$. By fundamental principle, I mean a principle that it does not seem possible to reduce to any other more basic or convincing principles.

From the present point of view, the likelihood ratio and χ^2 criteria are important because they provide manageable asymptotic approximations to the exact test and not because they are immediately convincing in themselves. To derive these asymptotic approximations in the simplest case, suppose X to be a finite discrete sample space and put

$$X_n = X^n = \underbrace{X \times \ldots \times X}_{n}$$

and

$$t_n(x_1, \ldots, x_n) = \sum_1^n t(x_i)$$

where $t(x)$ is a fixed function from X into Z^r. Then

$$f_n(t) = f^{n*}(t).$$

Suppose we want to test the reductive hypothesis that $t_n(x_1, \ldots, x_n)$ can be reduced to

$$u(t_n(x_1, \ldots, x_n)) = \sum_1^n u(t(x_i))$$

where $u(t)$ is a homomorphism from Z^r to Z^p with $p < r$. By change of

coordinates, we may assume $u(t)$ to be simply the left projection which takes $t = (u, v)$ into u. Then

$$g_n(u) = g^n{*}(u)$$

where

$$g(u) = \sum_v f(u, v).$$

The exact test orders the sample points according to the value of the ratio

$$\frac{f_n(t)}{g_n(u)}.$$

Introduce the exponential families

$$\frac{1}{\varphi(a)^n} e^{a \cdot t_n(x_1, \ldots, x_n)}$$

and

$$\frac{1}{\psi(b)^n} e^{b \cdot u_n(x_1, \ldots, x_n)}$$

corresponding to the full and reduced models, respectively, and decompose the parameter vector $a = (b, c)$ in the same way as $t = (u, v)$. Then we see that the parametric distribution of the reduced model is obtained from the parametric distribution of the full model by putting $c = 0$ so that, in particular,

$$\psi(b) = \varphi(b, 0).$$

The likelihood ratio with respect to the hypothesis $c = 0$ equals

$$\lambda_n(t) = \left(\lambda\left(\frac{t}{n}\right)\right)^n$$

$$= e^{n(H(\hat{a}(t/n)) - H(\hat{b}(u/n), 0))}$$

where $\hat{a}_n(t) = \hat{a}(t/n)$ and $\hat{b}_n(u) = \hat{b}(u/n)$ are the maximum likelihood estimates in the respective models and

$$H(a) = \log \varphi(a) - a \cdot m(a)$$

with

$$m(a) = E_a(t) = \operatorname{grad} \log \varphi(a)$$

is the Gibbsian entropy.

That the exact test reduces asymptotically to the likelihood ratio test is apparent from the following approximation theorem which is proved in Martin-Löf (1970). As $n \to +\infty$

$$\frac{f_n(t)}{g_n(u)} = \lambda_n(t) \frac{1}{(2\pi n)^{q/2}} \cdot \sqrt{\frac{\det PV(\hat{b}_n(u), 0)P'}{\det V(\hat{a}_n(t))}} \left(1 + O\left(\frac{1}{n}\right)\right)$$

uniformly when $\hat{a}_n(t) = \hat{a}(t/n)$ is bounded. Here $q = r - p$ is the number of degrees of freedom of the hypothesis,

$$V(a) = \mathrm{Var}_a(t) = \left(\frac{\partial^2 \log \varphi}{\partial a_i \partial a_j}\right)$$

and P is the left projection from $R^r = R^p \times R^q$ to R^p. Furthermore, Taylor expansion of $\log \lambda(t) = \log \lambda(u, v)$ in v around the point $v = Qm(\hat{b}(u), 0)$, where Q is the right projection from $R^r = R^p \times R^q$ to R^q, shows that

$$\log \lambda_n(t) = n \log \lambda\left(\frac{t}{n}\right) = -\frac{n}{2}\chi^2\left(\frac{t}{n}\right)$$
$$+ \text{terms of third and higher order}$$

where

$$\chi^2(t) = (t - m(\hat{b}(u), 0))' V(\hat{b}(u), 0)^{-1} (t - m(\hat{b}(u), 0))$$
$$= (v - Qm(\hat{b}(u), 0))' Q V(\hat{b}(u), 0)^{-1} Q' (v - Qm(\hat{b}(u), 0)).$$

These results justify the asymptotic use of the likelihood ratio and χ^2 tests on the basis of the fundamental principle formulated above.

Suppose now that the range of $t(x)$ is of the form $U \times V$ and consider the hypothesis that $t(x) = (u(x), v(x))$ can be reduced to $u(x)$. As above, the critical level with respect to the exact test of this hypothesis equals (in the discrete case)

$$\epsilon(t) = \epsilon(u, v) = \sum_{\substack{v' \\ f(u,v') \leqslant f(u,v)}} \frac{f(u, v')}{g(u)}.$$

If the reduced model fits, we are $(1 - \epsilon)$–certain that v is such that $\epsilon(u, v) > \epsilon$, that is, that

$$v \in V_\epsilon(u) = \{ v \mid \epsilon(u, v) > \epsilon \}.$$

Conversely, suppose that we can observe v but not u. By the Neyman

principle, we are then $(1-\epsilon)$ –certain that

$$u \in U_\epsilon(v) = \{u \mid v \in V_\epsilon(u)\}$$
$$= \{u \mid \epsilon(u, v) > \epsilon\}$$

provided the reduced model fits. $U_\epsilon(v)$ for $0 < \epsilon < 1$ is the family of *exact confidence regions* for u.

We define the (set of) *exact estimate*(s) of u to be the intersection of the exact confidence regions,

$$\check{U}(v) = \bigcap_{\epsilon<1} U_\epsilon(v) = \{u \mid \epsilon(u, v) = 1\}$$
$$= \left\{ u \mid \text{the distribution } \frac{f(u, v)}{g(u)} \text{ has mode at } v \right\}.$$

This is what the fundamental principle (of ordering the sample points according to the value of the ratio $f(u, v)/g(u)$) leads to when applied to the problem of estimation (or prediction). In the discrete case, $\check{U}(v)$ normally consists of several points whereas, in the continuous case, it reduces to a single point $\check{u}(v)$.

The considerations leading to the exact estimate are similar to those motivating the notion of *universality* introduced by Barndorff-Nielsen (1973). The exact mathematical relationship is this. The family of distributions $f(u, v)/g(u)$ indexed by u is universal if and only if $\check{U}(v) \neq \phi$ for all values of v.

The exact estimate should be compared with the (set of) *maximum likelihood estimate*(s)

$$\hat{U}(v) = \left\{ u \,\middle|\, u \text{ maximizes } \frac{f(u, v)}{g(u)} \right\}.$$

Observe, however, that, when we apply the Neyman procedure to the exact test, it is not the maximum likelihood estimate but the exact estimate that we are led to.

Example 3. Suppose that we have made a sequence of Bernoulli trials $x_1, \ldots, x_n$ with a total of $\Sigma_1^n x_i = t$ successes and that we want to predict $\Sigma_1^N x_i = T$ for $N > n$. The exact estimate $\check{T}(t)$ consists of those values of T for which the hypergeometric distribution

$$\frac{\binom{n}{t}\binom{N-n}{T-t}}{\binom{N}{T}}$$

has its mode at the observed value of t. For $N = n + 1$ we get in particular

$$\check{T}(t) = \begin{cases} \{t\} & \text{if } t < \frac{n}{2}, \\ \{t, t+1\} & \text{if } t = \frac{n}{2}, \\ \{t+1\} & \text{if } t > \frac{n}{2}, \end{cases}$$

provided n is even, and

$$\check{T}(t) = \begin{cases} \{t\} & \text{if } t < \frac{n-1}{2} \\ \{t, t+1\} & \text{if } t = \frac{n-1}{2} \text{ or } \frac{n+1}{2}, \\ \{t+1\} & \text{if } t > \frac{n+1}{2}, \end{cases}$$

provided n is odd. The maximum likelihood predictor, on the other hand, is given by

$$\hat{T}(t) = \begin{cases} \{t\} & \text{if } t < \frac{n}{2}, \\ \{t, t+1\} & \text{if } t = \frac{n}{2}, \\ \{t+1\} & \text{if } t > \frac{n}{2}, \end{cases}$$

irrespective of whether n is even or odd.

Let us now turn to the more usual form of the problem of estimation, namely that of *parameter estimation*. Again, I shall consider the simple case of a sequence of outcomes

$$x_1, \ldots, x_n, x_{n+1}, \ldots, x_N$$

from a finite discrete sample space X and statistics of the form

$$t = \sum_1^n t(x_i), \quad T = \sum_1^N t(x_i),$$

where $t(x)$ is a fixed function from X to Z^r. Let as before $f(t)$ denote the number of outcomes x such that $t(x) = t$, and consider the hypothesis (of homogeneity between the two samples $x_1, \ldots, x_n$ and $x_{n+1}, \ldots, x_N$) that the pair of statistics (t, T) can be reduced to T alone. According to the exact test criterion, we shall consider the ratio

$$\frac{f_n(t) f_{N-n}(T-t)}{f_N(T)}$$

where

$$f_n(t) = f^{n*}(t),$$

and reject the hypothesis if this ratio is too small. Also, from the exact test, the Neyman procedure described above allows us to derive the family of exact confidence regions $T_\varepsilon(t)$ and the exact predictor $\check{T}(t)$.

Now, it is a theorem (proved in Martin-Löf (1970)) that, if $N \to \infty$ and T varies with N in such a way that

$$\frac{T}{N} \to m(a) = \text{grad log } \varphi(a),$$

then

$$\frac{f_n(t) f_{N-n}(T-t)}{f_N(T)} \to \frac{1}{\varphi(a)^n} e^{a \cdot t} f_n(t).$$

Hence the problem of predicting T is turned into the problem of estimating the parameter a in the exponential family that appears in the limit. In particular, the exact predictor $\check{T}(t)$ is turned into

$$\check{A}(t) = \check{A}_n(t) = \left\{ a \;\middle|\; \text{the distribution } \frac{1}{\varphi(a)^n} e^{a \cdot t} f_n(t) \text{ has mode at } t \right\}$$

which is the exact estimate as originally introduced by Höglund (1971).

Example 4. Let $x_1, \ldots, x_n$ be the outcomes of a sequence of Bernoulli trials with success probability θ, and let $t = \Sigma_1^n x_i$ be the total number of successes. The exact estimate $\check{\theta}(t)$ as defined above consists of those values of θ for which the binomial distribution

$$\binom{n}{t} \theta^t (1-\theta)^{n-t}$$

has its mode at t. Thus

$$\check{\theta}(t) = \left[\frac{t}{n+1}, \frac{t+1}{n+1}\right].$$

In the continuous case, $\check{A}(t)$ normally consists of the single point

$$\check{a}(t) = -\text{grad} \log f(t).$$

Example 5. Let $x_1, \ldots, x_n$ be a sample from a normal distribution and put $t = (t_1, t_2)$ where

$$t_1 = \sum_1^n x_i \quad \text{and} \quad t_2 = \sum_1^n x_i^2.$$

Then, as in Example 2 above

$$f(t) = \frac{\pi^{(n-1)/2}}{\sqrt{n}\,\Gamma\left(\frac{n-1}{2}\right)} \left(t_2 - \frac{t_1^2}{n}\right)^{(n-3)/2}$$

so that

$$\check{a}_1 = \frac{n-3}{n} \frac{t_1}{t_2 - \frac{t_1^2}{n}},$$

$$\check{a}_2 = -\frac{n-3}{2} \frac{1}{t_2 - \frac{t_1^2}{n}},$$

which, via the relations

$$a_1 = \frac{\mu}{\sigma^2} \quad \text{and} \quad a_2 = -\frac{1}{2\sigma^2}$$

gives

$$\check{\mu} = \bar{x} \quad \text{and} \quad \check{\sigma}^2 = \frac{1}{n-3} \sum_1^n (x_i - \bar{x})^2.$$

The asymptotic behaviour of the exact estimate $\check{A}_n(t)$ as compared with the maximum likelihood estimate $\hat{a}_n(t) = \hat{a}(t/n)$ has been determined by Höglund (1971). Restrict $\hat{a}_n(t) = \hat{a}(t/n)$ to an arbitrary compact subset of the interior of the natural parameter space of the exponential family. Then, in

the discrete case, $\check{A}_n(t)$ is nonempty and compact for n sufficiently large and

$$\max_{a \in \check{A}_n(t)} |a - \hat{a}_n(t)| = O\left(\frac{1}{n}\right)$$

uniformly as $n \to \infty$. And, in the continuous case, under a mild regularity condition, $\check{A}_n(t)$ consists of the single point $\check{a}_n(t) = -\text{grad} \log f_n(t)$ if n is sufficiently large and

$$\check{a}_n(t) = \hat{a}_n(t) + O\left(\frac{1}{n}\right)$$

uniformly as $n \to \infty$. This, together with the fact that the random error of $\hat{a}_n(t)$ is of the order of magnitude $1/\sqrt{n}$, shows that the exact estimate may be approximated asymptotically by the maximum likelihood estimate. From the point of view adopted in the present paper, Höglund's theorem should be considered as a justification of the asymptotic use of the maximum likelihood estimate on the basis of the fundamental principle formulated above conjoined with the Neyman principle for constructing confidence regions and point estimates. This justification of the maximum likelihood estimate is, of course, quite different from the usual justification in terms of asymptotic minimum variance properties.

Summing up, it has been my purpose to draw attention to certain (in the discrete case) combinatorially defined quantities as being conceptually more fundamental then their more wellknown parametrical counterparts.

	combinatorial	parametrical
hypothesis testing	exact test	likelihood ratio and χ^2 tests
estimation	exact estimate	maximum likelihood estimate
entropy	$H(t) = \log f(t)$	$H(a) = E_a(-\log p_a)$ $= \log \varphi(a) - a \cdot m(a)$
redundancy	$R(t) = \dfrac{-\log \epsilon(t)}{\log g(u)}$	$R(a) = 1 - \dfrac{H(a)}{H(b(a), 0)}$

In this table, the last row refers to the notion of redundancy as discussed in detail in Martin-Löf (1973). Corresponding to each row there is an approximation theorem which shows that the combinatorially defined quantity under suitable asymptotic conditions may be replaced by its parametric counterpart. As mentioned above, these approximation theorems

can be found, in the case of the first, third and fourth rows, in Martin-Löf (1970) and (1973),and, in the case of the second row, in Höglund (1971).

University of Stockholm

REFERENCES

Barndorff-Nielsen, O.: 1973 *Exponential Families and Conditioning*, Sc.D. thesis, University of Copenhagen.

Höglund, T.: 1974 'The Exact Estimate – A Method of Statistical Estimation', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **29**, pp. 257–271.

Martin-Löf, P.: 1970 'Statistiska modeller', anteckningar från seminarier läsåret 1969–70 utarbetade av R. Sundberg, Institutionen för Matematisk Statistik, Stockholms Universitet.

Martin-Löf, P.: 1973 'The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data', *Scand. J. Statist.* 1, 3–18.