Scand J Statist 1: 3–18, 1974

The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data

PER MARTIN-LÖF

University of Stockholm

Received March 1973

ABSTRACT. It is proposed to supplement the critical level, as used in ordinary significance testing, by a measure of the magnitude of the departure of the observed set of data from the hypothesis to be tested. This measure, which is called the redundancy, appears in two versions, one microcanonical (or combinatorial) and the other canonical (or parametrical). The microcanonical redundancy is obtained by dividing minus the logarithm of the critical level by the Boltzmann entropy of the experiment and the canonical redundancy by dividing minus the logarithm of the likelihood ratio by the Gibbsian entropy. An approximation theorem shows that the former may be approximated asymptotically by the latter. The problem of calibrating the redundancy scale is discussed in connection with a series of examples, and, finally, certain considerations concerning the size of a statistical experiment are given which are based on the redundancy rather than the power function.

Key word: redundancy

Introduction

In statistical practice, we are faced with the following dilemma. When the number of observations is small, that is, when we have little information about the random phenomenon that we are studying, we easily get a positive result: this or that model fits the data satisfactorily, whereas with large sets of data our results are purely negative: no matter what model we try, we are sure to find significant deviations which force us to reject it. Exceptions are perhaps afforded by randomizing machines specially devised for the purpose of producing random sampling numbers. Even die casting leads to significant deviations from the hypothesis of equal probabilities for the six faces if one is sufficiently persistent, like Weldon with his 26 306 throws of 12 dice (see Fisher, 1925).

This indicates that for large sets of data it is too destructive to let an ordinary significance test decide whether or not to accept a proposed statistical model, because, with few exceptions, we know that we shall have to reject it even without looking at the data simply because the number of observations is so large. In such cases, we need instead a quantitative measure of the size of the discrepancy between the statistical model and the observed set of data which will allow us to decide whether this discrepancy, although highly significant, that is, not attributable to chance, is nevertheless so small that the model must be considered as providing a satisfactory approximate description of the data.

For certain special models or classes of models various such measures have indeed been introduced. For instance, in an $r \times s$ contingency table with a total of n observations, the quantity

$$\frac{\chi^2}{n\,(\min(r,s)-1)},$$

which has been normalized so as to take its values in the closed unit interval (see Cramér, 1945), measures the deviation from the hypothesis of independence. But it is not at all clear that it is meaningful to compare the values of this quantity for different contingency tables on a common scale. In fact, from the point of view adopted in the present paper, this will turn out not to be the case. To make different values of the mean square contingency χ^2/n comparable, it should instead be normalized by dividing by twice the sum of the entropies of the marginal distributions of the contingency table for which it has been calculated.

Microcanonical redundancy

Let X be a discrete sample space. By a *statistic*, I shall understand a function $t(x)$ which is defined on X and takes its values in some discrete set T and which is such that the set X_t, which consists of all outcomes x such that $t(x) = t$, is finite for all choices of t in T. The sets X_t were called *isostatistical regions* by Fisher (1921). Let $f(t)$ denote the

number of elements in X_t, that is, the number of outcomes x such that $t(x) = t$. In statistical mechanics, the function $f(t)$ is called the *microcanonical partition function* or, as in Khinchin (1949), the *structure function*. Also, the uniform distribution on X_t,

$$p_t(x) = \begin{cases} \dfrac{1}{f(t)} & \text{if } t(x) = t, \\ 0 & \text{otherwise,} \end{cases}$$

is called the *microcanonical distribution* after Gibbs. It is defined, of course, only if X_t is nonempty, that is, if $f(t) \neq 0$.

A *statistical description* of the outcome (or data) x consists of the observed value of $t(x)$ together with the information that x can be considered as drawn at random from the set X_t of all those outcomes x' for which $t(x') = t =$ the observed value of $t(x)$.

By a *(reductive) hypothesis*, I mean a hypothesis of the form

the data x can be described not only by the statistic $t(x)$ but already by the simpler statistic $u(x) = u(t(x))$.

Here $u(t)$ is a function defined on T with values in some discrete set U. Now, suppose that the hypothesis is true, that is, that x can be considered as drawn at random from the set X_u of all outcomes x' for which $u(x') = u =$ the observed value of $u(x)$. The corresponding microcanonical distribution is

$$p_u(x) = \begin{cases} \dfrac{1}{g(u)} & \text{if } u(x) = u, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, under the hypothesis, the distribution of $t(x)$ becomes

$$p_u(t) = \begin{cases} \dfrac{f(t)}{g(u)} & \text{if } u(t) = u, \\ 0 & \text{otherwise,} \end{cases}$$

where, of course,

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

is the structure function determined by the statistic $u(x) = u(t(x))$.

I regard it as a *fundamental principle* that the smaller the number $f(t(x))$ of outcomes is that realize the observed value of $t(x)$, the more does our observation x contradict the hypothesis that the statistic $t(x)$ can be reduced to $u(x) = u(t(x))$. By fundamental principle, I mean that it does not seem possible to reduce it to any other more basic or convincing principles.

According to the fundamental principle, we should define the *critical level* by

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

and reject the hypothesis for the outcome x on the level ε provided $\varepsilon(t(x)) \leqslant \varepsilon$. I have proposed to call this test, which was introduced in Martin-Löf (1970), the *exact test* of a reductive hypothesis since it is a general formulation of the procedure used by Fisher (1934) in his so-called exact treatment of a 2×2 contingency table.

The *statistical interpretation* of the critical level is as usual that

$\varepsilon(t(x))$ is the probability of getting an outcome which deviates at least as much from the hypothesis as the observed outcome x.

Here the probability is with respect to the microcanonical distribution determined by $u =$ the observed value of $u(x)$. However, in addition to the statistical interpretation, the critical level allows an *information theoretic interpretation* which says that

$-\log_2 \varepsilon(t(x))$ is the absolute decrease in the number of binary units needed to specify the outcome x when we take into account the regularities that we detect by means of the exact test.

Let us namely order the $g(u)$ outcomes x for which $u(x) = u$ according to their associated values of $f(t(x))$

$$\begin{array}{ccccc} x_1 & & x_2 & & \ldots \\ f(t(x_1)) & \leqslant & f(t(x_2)) & \leqslant & \ldots \end{array}$$

and give them binary codes as follows

x_1	x_2	x_3	x_4	...
1	10	11	100	...

Then the length of the binary code of an outcome x for which $u(x) = u$ is at most roughly

$$\log_2 \sum_{\substack{x' \\ f(t(x')) \leqslant f(t(x))}} 1 = \log_2 \sum_{\substack{t' \\ f(t') \leqslant f(t(x))}} f(t').$$

This should be compared with

$$\log_2 g(u)$$

which is roughly the number of binary units that we need in general to code an outcome x for which

we only know that $u(x) = u$. Hence the (absolute) decrease is

$$\log_2 g(u) - \log_2 \sum_{\substack{t' \\ f(t') \leqslant f(t(x))}} f(t') = -\log_2 \varepsilon(t(x)).$$

It is hardly astonishing that for large sets of data, that is, for large values of $\log_2 g(u)$, say of the order of magnitude 10^6, it will only very exceptionally be the case that $-\log_2 \varepsilon(t(x)) < 10$ which is required for acceptance of the hypothesis on the level of significance 0.001. This is remedied by considering, instead of the critical level, the *microcanonical redundancy*

$$R(t) = -\frac{\log \varepsilon(t)}{\log g(u)}, \quad u(t) = u,$$

which is, of course, independent of the base of the logarithms. Since

$$\frac{1}{g(u)} \leqslant \varepsilon(t) \leqslant 1$$

we have

$$0 \leqslant R(t) \leqslant 1$$

with $R(t) = 0$ and 1 corresponding to $\varepsilon(t) = 1$ and $1/g(u)$, that is, perfect and worst possible fit, respectively. The *interpretation* of the microcanonical redundancy is obtained directly from the information theoretic interpretation of the critical level. Thus

$R(t(x))$ is the *relative* decrease in the number of binary units needed to specify the outcome x when we take into account the regularities that we detect by means of the exact test.

Example 1. 2×2 contingency table. Each of n items is classified according to two dichotomous properties so that the outcome of the whole experiment may be represented in the form

$$x = ((2, 1), (1, 2), \ldots, (1, 1)).$$

The data are summarized in the usual four fold table shown in Table 1.

Put

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

and

$$u(x) = u(t(x)) = (n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2})$$
$$= (n_{11} + n_{12}, n_{21} + n_{22}, n_{11} + n_{21}, n_{12} + n_{22}).$$

Then

Table 1

	1	2	Total
1	n_{11}	n_{12}	$n_{1\cdot}$
2	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot} = n$

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

and

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

The hypothesis that $t(x)$ can be reduced to $u(x) = u(t(x))$ is the usual hypothesis of independence. It is rejected by the exact test if the hypergeometric probability

$$\frac{f(t)}{g(u)} = \frac{n_{1\cdot}!\, n_{2\cdot}!\, n_{\cdot 1}!\, n_{\cdot 2}!}{n!} \cdot \frac{1}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}$$

is too small.

For Lange's data concerning the criminality among the twin brothers or sisters of criminals reproduced in Table 2 taken from Fisher (1934), we get the values of $f(t)/g(u)$ shown in Table 3 so that the critical level becomes

$$\varepsilon(t) = \frac{13!\, 18!}{30!} \cdot (1 + 102 + 476 + 2992) = 0.00054$$

which is slightly larger than Fisher's value since he neglected the term corresponding to $n_{11} = 0$. The corresponding microcanonical redundancy is

$$R(t) = -\frac{\log \varepsilon(t)}{\log g(u)} = 0.20.$$

Table 2

	Convicted	Not convicted	Total
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

Table 3

n_{11}	0	1	...	10	11	12
$\frac{f(t)}{g(u)} \cdot \frac{30!}{13!\, 18!}$	476	12376	...	2992	102	1

Canonical redundancy

We shall now assume that the statistic $t(x)$ takes its values in Z^r and that the sum

$$\varphi(a) = \sum_x e^{a \cdot t(x)} = \sum_t e^{a \cdot t} f(t)$$

converges in the neighbourhood of at least some point a in R^r. The *parameter space* $A \subseteq R^r$ is defined to be the largest open set on which the sum converges. The function $\varphi(a)$, which is the Laplace transform of $f(t)$, is called the *canonical partition function.* It is positive and analytic on A. For a in A, the *canonical distribution* of x is defined by

$$p_a(x) = \frac{1}{\varphi(a)} e^{a \cdot t(x)}.$$

The induced canonical distribution of $t(x)$ is then clearly

$$p_a(t) = \frac{1}{\varphi(a)} e^{a \cdot t} f(t)$$

and the first two moments of $t(x)$ are given by

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

and

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

Like any variance matrix, $V(a)$ is positive definite, and it is *strictly* positive definite if and only if the range of $t(x)$ or, what amounts to the same, the support of $f(t)$ is not contained in a coset of a subgroup of Z^r of lower dimension. Note that this is a condition which does not depend on a. By replacing $t(x)$ by $t(x) - t_0$ and diminishing r, if necessary, we can and shall in the following assume that this condition is fullfilled. It implies no restriction of generality, because the passage from $t(x)$ to $t(x) - t_0$ and the decrease of r does not alter the induced partitioning of the sample space.

The function $\log \varphi(a)$ is analytic and, under the assumption just made, strictly convex. Therefore grad $\log \varphi(a)$ is one-to-one and analytic and has an inverse $\hat{a}(t)$ which is defined and analytic on the image of A under grad $\log \varphi(a)$. $\hat{a}(x) = \hat{a}(t(x))$ is the *maximum likelihood estimate* of the parameter a, because

$$\log p_a(x) = a \cdot t(x) - \log \varphi(a)$$

is a strictly concave function of a which assumes its unique maximum when

$$t(x) = \text{grad} \log \varphi(a)$$

provided $t(x)$ belongs to the range of grad $\log \varphi(a)$.

The *canonical* or *Gibbs entropy* is the quantity

$$H(a) = E_a(-\log p_a(x)) = \log \varphi(a) - a \cdot m(a).$$

Using it, we obtain the following simple expression for the attained maximum value of the likelihood,

$$\max_a p_a(x) = \frac{1}{\varphi(\hat{a}(t(x)))} e^{\hat{a}(t(x)) \cdot t(x)} = e^{-H(\hat{a}(t(x)))},$$

assuming, of course, that x is such that $t(x)$ belongs to the domain of $\hat{a}(t)$.

We shall now turn to the problem of testing a reductive hypothesis of the form

$t(x)$ can be reduced to $u(x) = u(t(x))$

where $u(t)$ is a *homomorphism* from Z^r to Z^p with $p < r$. Since a subgroup of a finitely generated free abelian group is again free and has at most as many generators (see Lang, 1965, p. 45), we can assume that the homomorphism $u(t)$ is actually *onto* Z^p. But then, after a change of basis in Z^r, if necessary, we can write

$$Z^r = Z^p \times Z^q,$$

where $q = r - p$ is the number of *degrees of freedom* of the reduction,

$$t = (u, v)$$

and assume that the homomorphism $u(t)$ is simply the associated left projection (again, see Lang, 1965, p. 44).

Partition the parameter vector $a = (b, c)$ in the same way as $t = (u, v)$ so that $a \cdot t = b \cdot u + c \cdot v$, and assume that the parameter space A contains at least one point of the form $(b, 0)$. Then the canonical distribution associated with the statistic $u(x)$ exists,

$$p_b(x) = \frac{1}{\psi(b)} e^{b \cdot u(x)}$$

where

$$\psi(b) = \sum_x e^{b \cdot u(x)} = \sum_x e^{b \cdot u(x) + 0 \cdot v(x)} = \varphi(b, 0),$$

and is obtained from the canonical distribution associated with $t(x)$ by putting $c = 0$. Hence the associated parameter space B consists of all those values of b for which $(b, 0)$ belongs to A. The condition

$$c = 0$$

will be referred to as the *parametric specification* of the hypothesis that $t(x)$ can be reduced to $u(x) = u(t(x))$.

The *canonical redundancy* is defined by

$$R(a) = 1 - \frac{H(a)}{H(b(a), 0)}$$

where $b(a)$ is the solution of the equation

$$P(m(b(a), 0)) = P(m(a)).$$

Here P denotes the left projection from $R^p \times R^q$ to R^p. If we compare this equation with the maximum likelihood equation for b under the hypothesis $c = 0$,

$$u = \text{grad log } \psi(b) = \text{grad log } \varphi(b, 0) = P(m(b, 0)),$$

we see that

$$b(a) = \hat{b}(P(m(a))).$$

The domain of $R(a)$ equals the domain of $b(a)$ and consists of all values of a in A for which $P(m(a))$ belongs to the domain of $\hat{b}(u)$. It is an open subset of A which contains all points in A of the form $(b, 0)$, because, if $a = (b, 0)$, then $b(a)$ is clearly defined and equal to b.

Whenever defined, $R(a)$ satisfies the inequality

$$0 \leqslant R(a) < 1,$$

and, furthermore,

$$R(b, c) = 0 \text{ if and only if } c = 0.$$

To see this, suppose that t and u belong to the domains of $\hat{a}(t)$ and $\hat{b}(u)$, respectively, where $t = (u, v)$. Then

$$H(\hat{a}(t)) = \min_a (\log \varphi(a) - a \cdot t)$$
$$\leqslant \min_b (\log \varphi(b, 0) - b \cdot u) = H(\hat{b}(u), 0).$$

Now, a belongs to the domain of $R(a)$ if and only if both $m(a)$ and $P(m(a))$ belong to the domains of $\hat{a}(t)$ and $\hat{b}(u)$, respectively. Hence we can put $t = m(a)$ in the above inequality, use the fact that $\hat{a}(m(a)) = a$, and conclude

$$H(a) \leqslant H(b(a), 0),$$

that is,

$$R(a) \geqslant 0.$$

Since $\log \varphi(a) - a \cdot t$ is a strictly convex function of a under the assumption that the structure function $f(t)$ is not concentrated on a coset of a subgroup of lower dimension, equality holds if and only if a is of the form $(b, 0)$. Finally, the inequality

$$R(a) < 1$$

follows immediately from the fact that, under the assumption about the support of the structure function, the distribution $p_a(x)$ is non degenerate so that its entropy satisfies

$$H(a) > 0.$$

The canonical redundancy $R(a)$ is a measure of the deviation of the parameter vector $a = (b, c)$ from the hypothesis $c = 0$. Its relation to the microcanonical redundancy will be established in the next section.

Example 1 (continued). For a 2×2 contingency table with probabilities as indicated in Table 4, the canonical redundancy with respect to the hypothesis of independence $p_{ij} = p_{i\cdot} p_{\cdot j}$ becomes

Table 4

	1	2	Total
1	p_{11}	p_{12}	$p_{1\cdot}$
2	p_{21}	p_{22}	$p_{2\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot} = 1$

$$1 - \frac{H(p_{11}, p_{12}, p_{21}, p_{22})}{H(p_{1\cdot}, p_{2\cdot}) + H(p_{\cdot 1}, p_{\cdot 2})}$$

where

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i.$$

Example 2. Multinomial distribution. The canonical redundancy with respect to the hypothesis that all the multinomial probabilities are equal,

$$p_1 = p_2 = \ldots = p_n = \frac{1}{n},$$

becomes

$$1 - \frac{H(p_1, \ldots, p_n)}{\log n},$$

which is the redundancy as defined by Shannon (1948).

The likelihood ratio with respect to the hypothesis $c = 0$ is by definition

$$\lambda(x) = \lambda(t(x)) = \frac{\max_b p_{b,0}(x)}{\max_a p_a(x)} = e^{H(\hat{a}(t(x))) - H(\hat{b}(u(x)), 0)},$$

assuming of course that $t(x)$ and $u(x)$ belong to the domains of $\hat{a}(t)$ and $\hat{b}(u)$, respectively. It allows us to give the following simple expression for the value of the canonical redundancy $R(a)$ for the argument $a = \hat{a}(t)$,

$$R(\hat{a}(t)) = 1 - \frac{H(\hat{a}(t))}{H(\hat{b}(u), 0)} = -\frac{\log \lambda(t)}{H(\hat{b}(u), 0)}.$$

Thus $R(\hat{a}(t))$ is $-\log \lambda(t)$ normalized by dividing by the entropy $H(\hat{b}(u), 0)$. Taylor expansion of $\log \lambda(t) = \log \lambda(u, v)$ in v around the point $v = Q(m(\hat{b}(u), 0))$, where Q denotes the right projection from R^{p+q} to R^q, yields

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

where, using matrix notation and putting for brevity $\hat{b} = \hat{b}(u)$,

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

Hence

$$R(\hat{a}(t)) = \frac{\chi^2}{2H(\hat{b}(u), 0)} + \text{terms of third and higher order}$$

which is a convenient formula to use for approximate computation of the redundancy when the value of χ^2 is either known or easier to compute than $H(\hat{a}(t))$.

If the structure function $f(t)$ is chosen in a pathological way, it can actually happen that the canonical redundancy $R(a)$ is defined only on a proper part of the parameter space A. The following example is due to Thomas Höglund. Take $p = q = 1$ and put

$$f(t) = f(u, v) = \begin{cases} \left[\frac{e^{u-v}}{(u-v)^3}\right] & \text{if } v < u \text{ and } v = 0 \text{ or } 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\varphi(a) = \varphi(b, c) = \sum_{u, v} e^{bu+cv} f(u, v)$$
$$= \left(\sum_{u=1}^{\infty} \left[\frac{e^u}{u^3}\right] e^{bu} \right) (1 + e^{b+c})$$

and the parameter space A is the half plane determined by the inequalities $b < -1$ and $-\infty < c < +\infty$. The range of

$$P(m(b, 0)) = \frac{\partial \log \varphi(b, 0)}{\partial b}$$

is the open interval from $-\infty$ to

$$\frac{\sum_{u=1}^{\infty} \left[\frac{e^u}{u^3}\right] u e^{-u}}{\sum_{u=1}^{\infty} \left[\frac{e^u}{u^3}\right] e^{-u}} + \frac{1}{e+1},$$

whereas the range of

$$P(m(a)) = \frac{\partial \log \varphi(b, c)}{\partial b}$$

is the open interval from $-\infty$ to

$$\frac{\sum_{u=1}^{\infty} \left[\frac{e^u}{u^3}\right] u e^{-u}}{\sum_{u=1}^{\infty} \left[\frac{e^u}{u^3}\right] e^{-u}} + 1,$$

the right end point being approached when b and c tend to -1 and $+\infty$, respectively. Hence there is no solution $b(a)$ to the equation

$$P(m(b(a), 0)) = P(m(a))$$

when the components b and c of the parameter vector $a = (b, c)$ are sufficiently close to -1 and $+\infty$.

All exponential families that occur in practice turn out to be such that the range of $m(a) = \text{grad} \log \varphi(a)$, which is always contained in the interior of the convex support of $f(t)$, actually equals it. As shown by Barndorff-Nielsen (1970), this is equivalent to $\log \varphi(a)$ being *steep* in his terminology. Now, suppose that the family of canonical distributions determined by $u(x)$ satisfies this regularity condition, that is, that $\hat{b}(u)$ is defined on the whole of the interior of the convex support of $g(u)$ or, what amounts to the same, that $\log \psi(b) = \log \varphi(b, 0)$ is steep (which, in turn, is guaranteed by $\log \varphi(a)$ being steep). Then $R(a)$ is defined on the whole of A, because for an arbitrary choice of a in A,

$m(a)$ belongs to the interior of the convex support of $f(t)$

which implies that

$P(m(a))$ belongs to the interior of the convex support of $g(u)$

which, in turn, implies that

$b(a) = \hat{b}(P(m(a)))$ is defined.

This argument, which is due to Ole Barndorff-Nielsen, shows that a counterexample such as that of Höglund has to be pathological.

Approximation theorem

In this section, we shall see that, in the case of a large number of independent repetitions of one and the same experiment, the microcanonical redundancy may be approximated by the canonical redundancy evaluated for the maximum likelihood estimate of the parameter.

Consider a sequence of sample spaces

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

and statistics

$$t_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} t(x_i)$$

where, as before, $t(x)$ takes its values in Z^r. Then, with obvious notation,

$$f_n(t) = f^{n*}(t), \qquad \varphi_n(a) = \varphi(a)^n,$$

$$m_n(a) = n \cdot m(a), \quad V_n(a) = n \cdot V(a),$$

$$H_n(a) = n \cdot H(a), \quad \hat{a}_n(t) = \hat{a}(t/n).$$

On the other hand, the parameter space A is the same regardless of the value of n.

Assume that the support of the structure function $f(t)$ is not contained in a coset of a proper subgroup of Z^r. Note that this condition is stronger than the previous condition that the support of $f(t)$ not be contained in a coset of a subgroup of lower dimension, which is equivalent to $V(a)$ being strictly positive definite. However, it clearly implies as little restriction of generality. Under the stronger assumption, we have the following *saddle point approximation* of the structure function,

$$f_n(t) = \frac{e^{n H(\hat{a}(t/n))}}{(2\pi n)^{r/2} \sqrt{\det V(\hat{a}(t/n))}} \left(1 + O\left(\frac{1}{n}\right)\right) \text{ as } n \to \infty$$

uniformly as long as $\hat{a}(t/n)$ stays within a fixed compact subset of the parameter space A. For a proof, see Martin-Löf (1970).

We shall be concerned with a reductive hypothesis of the form

$t_n(x_1, \ldots, x_n)$ can be reduced to $u(t_n(x_1, \ldots, x_n))$

$$= \sum_{i=1}^{n} u(t(x_i))$$

where $u(t)$ is a homomorphism from $Z^r = Z^{p+q}$ onto Z^p which we may assume to be simply the left projection. The parametric specification of the hypothesis is then $c = 0$, assuming of course that the parameter space A contains at least one point of the form $(b, 0)$. Let $R_n(t)$ denote the microcanonical redundancy with respect to this hypothesis and $R_n(a) = R(a)$ the corresponding canonical redundancy.

Theorem. *As* $n \to \infty$,

$$R_n(t) = R(\hat{a}(t/n)) + O\left(\frac{\log n}{n}\right)$$

uniformly when $\hat{a}(t/n)$ *stays within a fixed compact subset of the domain of* $R(a)$.

Proof. By definition,

$$R_n(t) = R_n(u, v) = 1 - \frac{\log \sum_{v'} f_n(u, v')}{\log g_n(u)}.$$

The inequality

$$f_n(u, v) \leqslant \sum_{f_n(u, v') \leqslant f_n(u, v)} f_n(u, v')$$

$$\leqslant f_n(u, v) \text{ (no. of } v\text{'s such that } f_n(u, v) \neq 0)$$

is trivial. The saddle point approximation gives

$$\log f_n(t) = n H(\hat{a}(t/n)) + O(\log n)$$

and

$$\log g_n(u) = n H(\hat{b}(u/n), 0) + O(\log n)$$

as $n \to \infty$ uniformly when $\hat{a}(t/n)$ and $\hat{b}(u/n)$ belong to compact subsets of A and B, respectively. Hence both of these asymptotic relations hold when $\hat{a}(t/n)$ stays within a compact subset of the domain of $R(a)$. It remains to estimate the size of the support of $f_n(u, v)$ regarded as a function of v. By assumption, the parameter space A contains at least one point of the form $(b, 0)$. But then, being open, it contains all of the $2q$ points (b, e_j) and $(b, -e_j)$ for $j = 1, \ldots, q$ where

$$e_j = (0, \ldots, \underset{j\text{th place}}{\varepsilon}, \ldots, 0)$$

provided ε is a sufficiently small positive number. The trivial inequality

$$f_n(u, v) \cdot e^{b \cdot u + c \cdot v} \leqslant \varphi(b, c)^n$$

implies that, if $f_n(u, v) \neq 0$, then

$$e^{b \cdot u + c \cdot v} \leqslant \varphi(b, c)^n.$$

Applying this to $c = e_j$ and $-e_j$, we can conclude that, if $f_n(u, v) \neq 0$, then

Table 5

Redundancy	Fit	p	
1	Worst possible	0.000	1.000
0.1	Very bad	0.316	0.684
0.01	Bad	0.441	0.559
0.001	Good	0.482	0.518
0.0001	Very good	0.494	0.506

$$-\frac{n}{\varepsilon}\left(\log \varphi(b, -e_j) - b \cdot \frac{u}{n}\right) \leqslant v_j \leqslant \frac{n}{\varepsilon}\left(\log \varphi(b, e_j) - b \cdot \frac{u}{n}\right)$$

for $j = 1, ..., q$ where $v = (v_1, ..., v_q)$. Hence the number of points in the support of $f_n(u, v)$ regarded as a function of v is $O(n^q)$ provided u/n is bounded as it is if $\hat{b}(u/n)$ belongs to a compact subset of B which, in turn, is guaranteed by the assumption that $\hat{a}(t/n)$ belongs to a compact subset of the domain of $R(a)$. Summing up,

$$R_n(t) = 1 - \frac{H(\hat{a}(t/n))}{H(\hat{b}(u/n), 0)} + O\left(\frac{\log n}{n}\right)$$

as was to be proved.

That the error term is best possible can be seen by considering the hypothesis that a binomial probability $p = 1/2$, because then

$$R_n(t) = 1 - \frac{1}{n} \log_2 \sum_{\binom{n}{t'} \leqslant \binom{n}{t}} \binom{n}{t'},$$

$$R(p) = 1 - (-p \log_2 p - (1-p) \log_2 (1-p)),$$

and a simple calculation shows that $R_n(t) - R(t/n) = \log_2 n/2n + O(1/n)$ when t/n is bounded away from 0, 1/2 and 1.

Example 1 (continued). For Lange's data, the canonical redundancy computed for the maximum likelihood estimates of the parameters equals 0.17 which should be compared with the value of the microcanonical redundancy which was found earlier to be 0.20. Thus the agreement is good even in this rather unfavourable case, especially as we shall only be interested in the order of magnitude of the redundancy.

Calibration of the redundancy scale

The redundancy enables us to measure quantitatively the discrepancy between a statistical hypothesis and a given set of data on an absolute scale. That is, whatever model and reductive hypothesis we consider, the interpretation of the redundancy is the same: it is the relative decrease in the number of binary units needed to specify the given set of data when we take into account the regularities that we detect by means of the exact test. Being the relative decrease of something, the redundancy takes its values in the closed unit interval.

We shall now turn to the problem of giving a qualitative interpretation of the various quantitative values of the redundancy. This is a problem which is similar in nature to the problem of where to write very cold, cold, cool, mild, warm, hot, etc. along an ordinary thermometer scale. In both cases, the solution has to be found through case studies. Table 5 contains for certain values of the redundancy, which are taken to be negative powers of ten, my proposed qualitative interpretation and also, in the last column, the values of a binomial probability p that produce the redundancy in question with respect to the hypothesis $p = 1/2$. Thus the last column is a table of the inverse of the function

$$R(p) = 1 - (-p \log_2 p - (1-p) \log_2 (1-p)).$$

The qualitative scale is admittedly tentative and needs to be corroborated by further case studies, but it is not as arbitrary as it may seem. So much is clear already from the few examples considered below, that it would be too liberal to admit a redundancy of 0.01 as good and that, in the other direction, if only redundancies of at most 0.0001 were accepted as good, then statistical inference for large data sets would become almost wholly impossible. That is, we would be able to fit statistical models only to very exceptional kinds of data, obtained, say, by coin tossing, die casting or observing a randomizing machine.

Example 2 (continued). Consider an English text without spaces and punctuation marks. If the text is long, consisting of 10 000 letters, say, we obtain for the redundancy with respect to the hypothesis of complete randomness

$$1 - \frac{-\hat{p}_a \log \hat{p}_a - ... - \hat{p}_z \log \hat{p}_z}{\log 26} = 0.12$$

where $\hat{p}_a, ..., \hat{p}_z$ are the relative frequencies of the letters $a, ..., z$. This corresponds to a very bad fit on the proposed qualitative scale.

Example 3. Out of 88 273 children born in Sweden in 1935, 45 682 were boys (see Cramér, 1945). The relative frequency of boys equals 0.5175 and differs of course highly significantly from 0.5. However, the redundancy with respect to the hypothesis of equal probabilities for boys and girls is only 0.0009 which corresponds to the value good on the qualitative scale.

Example 4. Test of independence in a 4×5 contingency table showing the distribution of 25 263

married couples according to annual income and number of children (see Cramér, 1945) gives $\chi^2 = 568.5$ for 12 degrees of freedom, indicating a highly significant deviation. The corresponding redundancy, obtained by dividing the mean square contingency by twice the sum of the entropies of the marginal distributions, equals 0.005 which is a bit higher than one would be willing to accept.

Example 5. The distribution of the head hair and eyebrow colours (light or red versus dark or medium) of 46 542 Swedish conscripts is shown in a 2×2 contingency table in Cramér (1945). $\chi^2 = 19\,288$ for 1 degree of freedom so that the deviation is highly significant. The corresponding redundancy is 0.17 indicating a very high degree pf association.

Example 6. Weldon's dice data (see Fisher, 1925). 12 dice were thrown 26 306 times and, in each throw, the number of dice scoring 5 or 6 was recorded. Let p_i denote the probability of exactly i dice scoring 5 or 6. The first hypothesis is that

$$p_i = \binom{12}{i} p^i (1-p)^{12-i}, \quad i = 0, 1, \ldots, 12,$$

for some p, giving $\chi^2 = 13.2$ for 11 degrees of freedom. Thus there is no significant deviation and no need to compute the redundancy. With respect to the second hypothesis, namely, that the dice are true,

$$p = 1/3$$

we get $\chi^2 = 27.1$ for 1 degree of freedom which is highly significant. The corresponding redundancy is nevertheless as small as

$$\frac{27.1}{2 \cdot 26306 \cdot 12 \cdot \log_e 6} = 0.000024$$

which falls well below the value corresponding to a very good fit on the proposed redundancy scale. The relative frequency of dice scoring 5 or 6 equals 0.3377 and is hence very close to 0.3333 ...

Example 7. Testing independence of sex and hair colour in the 2×5 contingency table reproduced by Fisher (1925) (Tocher's data) gives $\chi^2 = 10.48$ for 4 degrees of freedom which corresponds to a critical level between 0.02 and 0.05 (almost significance in Cramér's terminology). The number of observations is 3 883 and this makes the redundancy as low as 0.0007 which corresponds to a good fit on the qualitative scale.

Example 8. Traffic accidents. Let the index i range over the years 1961, ..., 1966, the index j over 92 consecutive days from the end of May till the end of August and the index k over the speed limits 90 km/hr, 100 km/hr and free speed that were tried in Sweden during the period in question. Assume that the number of accidents involving injured people and reported by the police year i and day j follows a Poisson distribution with mean value λ_{ij} and that the different accident numbers are independent. The test of the hypothesis

$$\lambda_{ij} = \alpha_i \beta_{jk_{ij}},$$

where k_{ij} is the speed limit year i and day j, gives for the Swedish accident data (see Jönrup & Svensson, 1971) $\chi^2 = 565$ with 446 degrees of freedom, indicating a highly significant deviation from the hypothesis. The corresponding redundancy equals 0.0038 and falls between good and bad on the proposed qualitative scale. Also, the test of the hypothesis that there are no effects of the speed limits,

$$\beta_{jk} = \beta_j,$$

yields $\chi^2 = 85$ for 9 degrees of freedom which is again highly significant, but the redundancy is now only 0.0006. Thus the effects of the speed limits are almost drowned by the bad fit of the model.

Example 9. Wilson's model. The Stockholm region has been divided into 41 districts and for each of 407 063 persons living and working in the region has it been recorded in which district the person lives and in which district he works. Thus the data (from Marksjö, 1970) appear in the form of a quadratic contingency table. Let p_{ij} be the probability that a person lives in district i and works in district j, and assume the different persons to be independent. The hypothesis to be tested is that

$$p_{ij} = \alpha_i \beta_j \gamma^{c_{ij}}$$

where c_{ij} is the cost of transportation from district i to district j. This model has been proposed by Wilson (1967). For his data, Marksjö obtained χ^2/degrees of freedom = 16.4 which is of course highly significant. However, because of the very large number of observations, the redundancy is still as low as 0.0041, a value which falls between good and bad on the proposed qualitative scale and corresponds to a deviation of a binomial probability from 0.5 by the amount 0.04. On the other hand, the hypothesis

$$\gamma = 1$$

of no sensitivity to the cost of transportation leads to the redundancy 0.024 which is six times as high and worse than bad on the qualitative scale.

Critical size of an experiment

The following procedure for testing a (reductive) statistical hypothesis is suggested. To begin with,

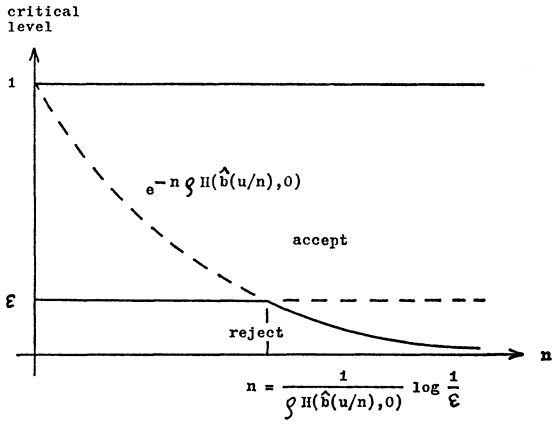

Fig. 1

compute the critical level $\varepsilon(t)$. If $\varepsilon(t) > \varepsilon$, where ε is the level of significance, we accept the hypothesis. If $\varepsilon(t) \leqslant \varepsilon$, we compute in addition the (microcanonical) redundancy

$$R(t) = -\frac{\log \varepsilon(t)}{\log g(u)}$$

and check whether $R(t) < \varrho$ or $R(t) \geqslant \varrho$ where ϱ is the limit of the redundancies that we are willing to tolerate. If $R(t) < \varrho$, we accept the model although the observed deviation is significant because we think that it nevertheless describes the data with sufficient accuracy. Finally, if $\varepsilon(t) \leqslant \varepsilon$ and $R(t) \geqslant \varrho$, we reject the model because the observed deviation is both significant and unacceptably large.

To be sure that an unacceptably large value of the redundancy is significant, that is, has probability $\leqslant \varepsilon$, the experiment has to be so big that

$$g(u) \geqslant \left(\frac{1}{\varepsilon}\right)^{1/\varrho}$$

or, equivalently,

$$\log g(u) \geqslant \frac{1}{\varrho} \log \frac{1}{\varepsilon}.$$

Indeed, the inequality $R(t) \geqslant \varrho$ is equivalent to $\varepsilon(t) \leqslant g(u)^{-\varrho}$ and, to be sure that the probability of this event is $\leqslant \varepsilon$, we have to have $g(u)^{-\varrho} \leqslant \varepsilon$ which is equivalent to the inequality above. Note that the expression $(1/\varrho) \log (1/\varepsilon)$ is much more sensitive to changes in ϱ than changes in ε.

In the special case of n independent repetitions of one and the same experiment, the saddle point approximation gives

$$\log g_n(u) = nH(\hat{b}(u/n), 0) + O(\log n)$$

uniformly when $\hat{b}(u/n)$ stays within a compact subset of B. Neglecting the error term, the above inequality is then transformed into the inequality

$$n \geqslant \frac{1}{\varrho H(\hat{b}(u/n), 0)} \log \frac{1}{\varepsilon}$$

which allows us to determine the number of observations that we have to make in order to be sure that an observed redundancy $\geqslant \varrho$ is significant. See Fig. 1.

Example 10. Suppose that we want to test whether a coin can be regarded as ideal by tossing it n times. Then $g_n(u) = 2^n$ so that the above inequality specializes to

$$n \geqslant \frac{1}{\varrho \log 2} \log \frac{1}{\varepsilon}.$$

In particular, for $\varepsilon = 0.01$ and $\varrho = 0.001$ we get $n \geqslant 6\,644$ which is roughly the number of times that we have to toss the coin in order to be able to detect substantial deviations from the hypothesis of equal probabilities for head and tail.

References

Barndorff-Nielsen, O. (1970). *Exponential families. Exact theory*. Various Publication Series No. 19. Matematisk Institut, Aarhus Universitet.

Cramér, H. (1945). *Mathematical methods of statistics*. Almqvist & Wiksell, Stockholm.

Fisher, R. A. (1921). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222**, 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers*. First edition. Oliver and Boyd, Edinburgh.

Fisher, R. A. (1934). Ibid. Fifth edition.

Jönrup, H. & Svensson, Å. (1971). *Effekten av hastighetsbegränsningar utanför tätbebyggelse*. Statens Trafiksäkerhetsråd. Meddelande 10.

Khinchin, A. I. (1949). *Mathematical foundations of statistical mechanics*. Dover, New York.

Lang, S. (1965). *Algebra*. Addison-Wesley, Reading, Massachusetts.

Marksjö, B. (1970). *Gravitationsmodellen i trafikplanering*. Skrifter i planeringsteori och tillämpad matematik, PT 1970: 2. Forskningsgruppen för planeringsteori, Matematiska institutionen, Tekniska högskolan, Stockholm.

Martin-Löf, P. (1970). *Statistiska modeller*. Anteckningar från seminarier läsåret 1969–70 utarbetade av R. Sundberg. Institutionen för matematisk statistik, Stockholms universitet.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423 and 623–656.

Wilson, A. G. (1967). A statistical theory of spatial distribution models. *Transportation Research* **1**, 253–269.

Per Martin-Löf
Barnhusgatan 4
S-111 23 Stockholm
Sweden

DISCUSSION

The previous paper was presented at the Conference on Foundational Questions in Statistical Inference, held at the Department of Theoretical Statistics, Aarhus University, May 7–12, 1973. After the presentation of the paper, the following discussion took place.

F. Abildgård (Copenhagen; specially invited contribution): I think that is is wellknown to everybody that testing in large sets of data almost always provides significant deviations from the hypothesis, whether this is a model or e.g. some sort of homogeneity. In any case this is wellknown from the literature and to those doing practice it is also wellknown from their daily life. The question to be discussed here is which conclusions should be drawn from this experience.

It is clear from the paper and the lecture we have just heard what is Martin-Löf's conclusion. This is already stated on page 3 in the paper: "In such cases we need instead (of the classical tests) a quantitative measure of the size of the discrepancy between the statistical model and the observed set of data which will allow us to decide whether this discrepancy, although highly significant, that is, not attributable to chance, is nevertheless so small that the model must be considered as providing a satisfactory approximate description of the data."

Martin-Löf's answer to this need is the measure of redundancy we have now heard about. This proposal may be evaluated from different points of view. As far as I am able to follow the mathematics I enjoy it very much as a stimulating contribution to the theory of statistics. At the same time the exposition is at such a level of clarity that I find it difficult to point out problems for discussion on the purely technical or mathematical aspect of the paper.

But there is another aspect of the idea of testing on the basis of the measure of redundancy which should be scrutinized, namely the results which come out. Let us to this end regard e.g. Example 4 in the paper: The data concern the distribution of married couples according to number of children in different income groups, and Karl Pearson's classical χ^2-test yields a highly significant deviation between these distributions. On the other hand the redundancy "equals 0.005 which is a bit higher than one would be willing to accept". I looked up the reference in Cramér's book, and found Table 1.

Table 1

No. of children	Income, Sw. kr. $\times 10^{-3}$			
	0–1	1–2	2–3	3–
0	35	33	42	54
1	45	47	43	35
2	15	16	12	10
3	4	4	2	1
$\geqslant 4$	1	1	1	1
Σ	100	101	100	101
n	6 116	10 928	5 173	3 046

Inspection of the relative distributions in Table 1 reveals an obvious, systematic trend towards fewer children in families with high income. I had the impression that this example was meant as an illustration of an unjustified significance caused by a large n. I think the example is better suited to illustrate the claim that one should study the structure of the data before putting model reducing hypotheses forward for testing.

Of course there are many situations where one feels that the deviation between the model and the data even if it is highly significant is statistically irrelevant and only causes trouble. Let us consider another example. It is wellknown from statistical textbooks how one can test the linearity of the regression of X on z by means of the statistic

$$F = \frac{s_2^2}{s_1^2},$$

$$s_1^2 = \frac{1}{n_\cdot - k} \sum_{ij} (X_{ij} - X_{i\cdot})^2,$$

$$s_2^2 = \frac{1}{k-2} \sum_i n_i (X_{i\cdot} - a - bz_i)^2.$$

Under H_0: $\xi_i = \alpha + \beta z_i$ F will follow the F-distribution. It is also wellknown to everybody in this audience that this test has nothing at all to do with linearity. So it is easy to construct examples where a perfect linearity leads to highly significant F-statistics, namely by decreasing s_1^2. It is also easy to construct examples to the opposite effect e.g. with clearly curved mean structure but with s_1^2 sufficiently large.

I have the impression that the first one of these two possibilities in principle is of the same nature as, and even may be considered as a generalization of the problem with large sets of data dealt with by Martin-Löf. The essential problem is that the "precision" of the data may be too high compared to the precision with which the data fit the model. Sometimes this damagingly high precision is provided by means of a large set of observations. What do we do

with this problem in the regression case. The usual way to remedy the bad fit is by introducing a model of iterated sampling. In this house this is called the JOK-principle. The idea is that you have a model until you choose a new one. Here you choose a new model in which the ξ's are considered stochastic with their own variance. This is a generally accepted approach and it may well be worth-while to consider its use also in the cases discussed by Martin-Löf.

One may ask why one shall choose a new model e.g. in the regression example above. Of course not for the purpose of describing the data. They are already satisfactorily described by the graphical display of the observations. This is a perfect reduction of the data. One may, however, be interested in testing a certain position of the regression line or in comparing two regression lines. In these cases one needs a model that fits and may use a model with iterated sampling. What is said by this is merely that the methods of theoretical statistics do not provide tools that are suitable for discovering the structure and essential features of the data. Their application lies in two other functions.

The first is that statistical models sometimes may be used for obtaining a more comprehensive picture of the structure of the data than can be obtained by other means. The second function is that the appropriate application of statistical methods may safeguard against overinterpreting the tendencies of the data. In vulgar language one may say that the statistical methods tell whether or not this or that feature of the data is likely to reproduce if the observation is repeated.

The next question is why I seem to prefer the application of e.g. models with iterated sampling to inference of Martin-Löf's type when describing data and testing. First of all I must admit, referring to a remark on page 5 in the paper, that it actually is astonishing to me "that for large sets of data ... it will only very exceptionally be the case that $-\log_2 \varepsilon(t(x)) < 10$" if you have a serious belief in the model on trial.

But there is also another thing, namely that I still seem to retain some sentimental feelings for classical significance testing. I think these are closely related to the second of the two above-mentioned functions of statistics: the distinction between those features of the data which should be considered incidental and those which one would bet will occur again if the experiment is repeated, even if factors not entering into the structure are changed. My feeling is that this purpose is in some way or another much better achieved by means of classical testing procedures than e.g. by means of Martin-Löf's method based on the redundancy measure.

I happen to have some experience from work in two of the fields mentioned in Martin-Löf's examples: traffic accidents under speed limits (Example 8) and traffic modelling (Example 9). In both cases I got the same result as Martin-Löf: an extremely bad fit with the first model I tried. In both cases I drew the conclusion that the data contained something not contained in the model, and that this something had something to do with inhomogeneities, and that the apparent structure might be influenced by factors not controlled by the data, so that I really would hesitate to bet even one penny on the reliability or reproductiveness of this structure. Consequently I started to search for factors generating such inhomogeneities. This work is not yet finished but it has up till now in both cases resulted in the discovery of a factor which seems to be of importance for the description of the data. In the traffic accident case one obtains a highly improved fit by means of splitting the accidents according to part-combinations (e.g. single driver's accidents, collissions between two drivers in crossroads). In the traffic model example the trick seems to be to split the total frequency matrix according to different categories of travellers, e.g. according to trade, appointment, and income.

The main trait in the argument here, however, is not how to split or according to what, but the idea that if you, by means of splitting the data and thereby in some sense making the model more detailed and complex, have obtained a satisfactory fit in classical terms, you feel that you have finished the description and got to the bottom of the data.

In cases where one is prevented from tracing the fundamental structure in this way e.g. due to certain limitations in the data the question arises whether to use Martin-Löf's redundancy measure, models with iterated sampling, or simply to abstain from reducing the data and drawing further conclusions. One reason for preferring the last possibility is that the uncontrolled factors producing inhomogeneities may act in the data in such a way that they guide your conclusions in a way you are unable to observe.

Even if I in principle agree with the idea that the proposed qualitative scale for the evaluation of the redundancy measure (cf. p. 10) as well as the entire idea of using such a measure in statistical inference need to be corroborated by further case studies, it is not at all clear to me what this means in practice and I am afraid that I see problems in this, possibly even more than the author does. For the moment I shall consider a bad fit as indicating that one has not finished the investigation and the research work must proceed. And that you must be most hesitating and make all reservations when drawing conclusions.

(These considerations also seem to apply to the social sciences.)

A. P. Dempster (Harvard University): Martin-Löf has shown in detail how the results of traditional tail area testing at a fixed size ε diverge from the results of redundancy testing at a fixed level ϱ as sample size increases, suggesting that the latter may sometimes be more reasonable with large data sets when tail area tests are almost certain to reject. Another kind of test which is known to accept much more often in large samples than does the traditional test is the straightforward Bayesian procedure which rejects where the posterior probability of the null hypothesis is less than some small value δ. This procedure requires specification of prior probabilities of null and alternative hypotheses, and genuine prior densities over the parameters within the null and alternative hypotheses. I wish to ask Martin-Löf: what is the relative behaviour as sample size increases of redundancy testing and Bayesian testing with fixed prior distribution? Or, put differently, how must the Bayesian let his prior distribution change as sample size increases in order to reproduce the results of redundancy testing?

D. Basu (Indian Statistical Institute): That classical null-hypothesis testings and Bayesianism do not go well together is seen from the following simple example. For further comments on this see section 11 of my essay on likelihood.

Suppose n independent observations on a variable $X \sim N(\theta, 1)$ with the null-hypothesis $H_0 = \text{Hyp}\ (\theta = 0)$ yield the 'highly significant' mean observation $\bar{x}_{(n)} = 3/\sqrt{n}$. With a uniform prior for θ over a reasonable interval around the origin, a Bayesian will then work out the posterior distribution for θ as $N(3/\sqrt{n}, 1/\sqrt{n})$. (A fiducialist or structural probabilist will arrive at the same distribution for θ.) If we denote by H_0^* the composite hypothesis $-1/10 < \theta < 1/10$, then a Bayesian will work out the posterior probability of H_0^* as

$$\Pr\left(-\frac{\sqrt{n}}{10} - 3 < N(0, 1) < \frac{\sqrt{n}}{10} - 3\right).$$

Note that the above is less than 0.025 if $n = 100$ but is greater than 0.999 if $n = 10\,000$.

D. R. Cox (Imperial College): (i) Significant 'rejection' of a null hypothesis means that the data are inconsistent with that hypothesis and provide evidence of the direction of departure. This is quite a different issue from whether to procede with the hypothesis, for which consideration of the practical importance of the magnitude of departures is normally required; Newton's law of gravitation is an often-quoted example. The present very interesting paper provides a measure of importance of departures that does not involve considerations outside the null hypothesis itself. A crucial question is whether such external considerations can really be avoided.

(ii) Following Professor Barnard, significance tests can be classified according as they

(*a*) involve explicit probabilistically formulated alternatives, as in Neyman-Pearson theory;

(*b*) are "simple" tests, in which a test statistic is defined measuring departures from the null hypothesis, and the tail area of its null hypothesis distribution calculated;

(*c*) are "absolute" tests, in which, as in the present paper the ordinates of the distribution under the null hypothesis define the test statistic.

Is (*c*) really a viable idea, independently of some considerations of type (*b*)? If, as in some permutation tests, the ordinate varies irregularly with a "natural" test statistic, would one sum over all points with small ordinates?

(iii) The denominator of Dr Martin-Löf's ratio seems especially sensitive to the precise definition of the data. For example, if x were supplemented by binary noise, ε would be unaffected, but not R.

A. W. F. Edwards (Cambridge University): I do not myself accept the notion of an absolute measure of goodness-of-fit, and therefore I am not worried by precisely this problem. However, if one uses a relative measure, such as the support or log-likelihood, one has a comparable situation in which two hypotheses, hardly distinguishable from one another, differ enormously in the support they attract, owing to the great size of the data.

A parallel argument to yours would then say that such a small difference between hypotheses should not 'matter' because taking the better hypothesis makes a relatively negligible contribution to 'explaining' the data (in the information-theory sense). But surely in science we are frequently concerned with such hypotheses, that require extensive experiments or series of observations to put them to the test. For example, atmospheric tides have been demonstrated from very long series of barometric observations. Your test would dismiss the tides as making a negligible contribution to explaining the daily variation in atmospheric pressure. But surely that was not the point.

D. A. Sprott (University of Waterloo): Other discussants have pointed out the necessity of examining the sources and possible reasons for departures

of the data, however numerous, from the model, particularly in sciences in which reproducible measurements can be taken under controlled experimental conditions. It is interesting in this regard to draw attention to a brief, somewhat critical, paper by R. A. Fisher (1943) entitled "Note on Dr Berkson's criticism of tests of significance". In it he quotes Dr Berkson as saying of the results of a genetics experiment that the line is as straight as any in biology. Fisher states that this attitude, in respect of the particular genetics example cited, would have precluded finding an important fact. Or, that if the deviation could be due to an error in experimental technique, this error would never be uncovered, as the ignoring of the results of the significance test essentially denies the existence of such an error.

G. A. Barnard (University of Essex): Since everyone so far has been critical of Martin-Löf's proposal, I would like to say something in support. The fact is, that when we use tests like χ^2 we are using "blunderbuss" procedures which are almost always capable of refinement if we take more thought. Nonetheless the χ^2 test remains a very useful tool for statisticians who may not be able to give all sets of data the individual treatment they ought to receive. Similarly the redundancy may be regarded as a "blunderbuss" procedure to check against rejection when the model gives fair approximate fit.

For these reasons I see this procedure as potentially very useful for social science application, though less useful for physicists.

I wonder whether there could be any "partitioning" of R, to correspond with the partitioning of χ^2?

O. Barndorff-Nielsen (Aarhus University): Tying up with some of the previous remarks, and with regard to the question of what discrepancies to expect between model and data for large data sets, I wish to draw attention to a paper by Berkson (1966) in which he examines the fit to the Poisson hypothesis for a series of 10 220 observations of waiting times for α-particle emissions. Berkson found excellent agreement between the data and the Poisson model as judged by usual χ^2 and dispersion index tests.

J. D. Kalbfleisch (The State University of New York at Buffalo): My remarks are closely related to the comments of some of the previous discussants. A significance test answers the question "Are the observed observations significantly different from those that are expected under the hypothesis?" where the word significantly has a well defined technical meaning. As such it is reasonable to form an absolute calibration of the significance scale. It is certainly true, however, that "significant" in its technical sense does not necessarily mean "important". This paper, it seems to me, is concerned with the more difficult problem of assessing when the data indicate that the departures from the specified model are important. But, what is an important departure depends critically on the type of model being considered and more specifically on the use to be made of the model. It would seem, therefore, that any calibration of the redundancy scale should depend on these factors and specifically on the size of the deviations from the proposed model which are deemed to be important.

G. Rasch (University of Copenhagen): Let me first put a technical question to the speaker: From the approximation on only the exponential family of distributions was considered. It certainly offers facilities that follow from the additivity of the relevant statistics, but is that quite decisive? Are similar results available for other types of distributions?

Next I wish to make it quite explicit, that the reason for using both significance and redundancy lies in the contention that *every model is basically wrong*, i.e. it is bound to fail, given enough data.

When you are in the possession of a set of data you may then either be in the position that your significance test tells you that the model fails, or you may not have got enough observations for that purpose. In the latter case you *cannot yet reject* the model on statistical grounds—which of course should not be construed as meaning that you really *accept* it. In the former case you have to realize that the model fails—and I have no sympathy for relaxing the significance requirement for the reason that the data are substantial enough to show it—but that does not mean that the model is too bad to be applied in the actual case.

To take a parallel from elementary physics: A "mathematical pendulum" is defined as "a heavy point, swinging frictionless in a weightless string in vacuum". A contraption like that was never seen; thus as a model for the motion of a real pendulum it is "unrealistic". Notwithstanding, it works quite well for a short time interval, but it begins soon to show a systematic decrease of the oscillation angle. To the model—a second order differential equation —thus requiring an amendment, a friction term is added, and now it works perfectly well for a long time, even during a few days, until another systematic deviation shows. If needed, a further correction, for air resistance, say, should be attempted—but as a matter of fact, *this is not needed*, because it has worked well enough for the purpose of the geophysicist, which was to measure the gravity constant ("g") with 7 decimal places!

It is exactly at this point Martin-Löf's redundancy sets in: the model fails—that being demonstrated by some significance test—but does it matter for its purposes?

Taking his cue from Information Theory, Martin-Löf uses the redundancy, as there defined, for measuring the deviation of the model from the data, in the sense of determining the relative decrease of the amount of information in the data which is caused by the departure from the null hypothesis.

Taken literally, the redundancy as a tool may be a rather gross evaluation of the loss suffered by replacing the data by the model. Even if it seems small *the parts lost* may effect some of the use of the model quite appreciably. Therefore it may be necessary to undertake a careful analysis in order to localise the losses and consider what to do about them.

In this connection I may touch upon Weldons dice throwing experiment with a redundancy of 0.000024. But what if we on several repetitions found the same result and it turned out, that the deviations of the observed distributions from the model distributions persisted in the same parts of them?

I do not know of any repetition of the experiment, neither of any detailed report on fractions of it as they were produced during some years, but I do happen to know (see Steffensen, 1923) that in a similar case the deviations were taken sufficiently seriously by statisticians to attempt fitting them with a number of alternative distributions, any particular justification of which I do not recall having seen.

Let me end up with the scale of redundancies presented by the speaker. It did leave me with the notion of new horrors of conventional limits! In *this* connection we may, however, have a chance of doing it more rationally by analyzing just which sort of damage and how much of it is invoked by using the model for specified purposes.

I do look forward to the contribution of the redundancy concept to articulating my vague thesis, that we should never succumb to the illusion that any of our models are correct, but we should certainly aim at *making them adequate for our purposes* —the redundancy possibly being a useful measuring instrument in that connection.

Author's reply: Dempster asks how the Bayesian would have to let his prior distribution change as sample size increases in order to reproduce the results of redundancy testing. I do not know exactly how, but it is clear that the change would have to be quite drastic. It would probably be more reasonable to ask how he would have to change his δ in order to reproduce the results in question.

It is true that the definition of the redundancy involves no power function considerations, but I cannot agree with Cox that it involves no considerations at all outside the null hypothesis itself. The *null hypothesis* asserts that the data x can be described by the statistic $u(x)$, and we are testing this hypothesis against the *alternative* determined by a statistic $t(x)$ through which $u(x)$ factors, so that we may write $u(x) = u(t(x))$.

Cox also asks if it is a viable idea to let the ordinate of the distribution of t under the null hypothesis define the test statistic even if this distribution varies irregularly. As a typical example, we may consider the hypothesis of absence of a trend in a permutation $x_1, ..., x_n$ of the integers $1, ..., n$. The statistic

$$t = \sum_1^n i x_i$$

has a distribution which, although asymptotically normal, varies irregularly for moderate values of n. See Fig. 16.1 on p. 398 of Kendall (1943), where it is plotted for $n = 8$. The exact (or, in Cox's terminology, absolute) test rejects if

$$\frac{f(t)}{n!}$$

is too small where $f(t)$ is the number of permutations with $\Sigma_1^n i x_i = t$. In this particular example, Cox's question is whether the test statistic $f(t)/n!$ might not be unnatural compared with one like

$$\left| t - \frac{n(n+1)^2}{4} \right|$$

which measures the deviation of t from the mean value of its (symmetrical) distribution under the null hypothesis. Now, in the extreme case when $f(t) = 0$, the hypothesis should no doubt be rejected even if the value of t falls close to the center of its distribution under the null hypothesis. And there seems to me to be a difference not of substance but merely of degree between $f(t)$ actually vanishing and $f(t)$ being very small. Thus I think that it is in agreement with intuition that the exact test rejects the hypothesis if $f(t)$ is sufficiently small irrespective of the numerical value of t.

Finally, Cox points out that, if we supplement our outcome x by n digits of binary noise, then the microcanonical redundancy changes from

$$-\frac{\log \varepsilon(t)}{\log g(u)} \text{ to } -\frac{\log \varepsilon(t)}{\log g(u) + n \log 2}$$

and the canonical redundancy from

$$\frac{H(b(a),0)-H(a)}{H(b(a),0)} \text{ to } \frac{H(b(a),0)-H(a)}{H(b(a),0)+n\log 2}.$$

Thus the denominator is very sensitive to the precise definition of the data. However, this seems to be unavoidable for a quantity which, like the redundancy, attempts to measure the overall discrepancy between the hypothesis and the observed set of data. Indeed, if we supplement the data by perfect binary noise, then the overall fit is improved and becomes perfect in the limit when the amount of noise increases indefinitely.

The difficulty that Cox points out in his last remark is more acute in the case of continuous distributions, for which the choice of the class width is to some extent arbitrary. Suppose, for example, that $x_1, ..., x_n$ is a sample from a normal distribution with mean value μ and unknown variance σ^2 and that we want to test the hypothesis $\mu=0$. The entropy of a discretized normal distribution with respect to the natural logarithm base equals

$$\tfrac{1}{2}(1+\log 2\pi)+\log\frac{\sigma}{h}$$

where h is the class width. Hence the canonical redundancy, evaluated for the maximum likelihood estimates of the parameters, equals

$$\frac{\log\frac{\hat\sigma_0}{\hat\sigma_1}}{\tfrac{1}{2}(1+\log 2\pi)+\log\frac{\hat\sigma_0}{h}}$$

where

$$\hat\sigma_0^2=\frac{1}{n}\sum_1^n x_i^2 \quad \text{and} \quad \hat\sigma_1^2=\frac{1}{n}\sum_1^n (x_i-\bar x)^2.$$

Again it is the denominator that causes trouble, in this case by depending on the class width h. However, when h changes from the very large value $\hat\sigma_0$ to the rather small value $\hat\sigma_0/10$, the denominator changes from 1.4 to 3.7 which means that, for h in the specified range, the redundancy becomes determined up to a factor three approximately. Since one step on my proposed qualitative scale corresponds to a factor ten, this makes the redundancy quite well-determined for practical purposes despite its dependence on the class width, the choice of which will probably always remain somewhat arbitrary.

Barnard asks whether there might be a partitioning of the redundancy analogous to the wellknown partitioning of χ^2. Indeed, there is. Consider first a reduction from a statistic t_1 to t_2 and then a subsequent reduction from t_2 to t_3. Let the corresponding canonical redundancies be denoted R_{12} and R_{23}, respectively. Also, let R_{13} be the canonical redundancy with respect to the composite reduction from t_1 to t_3. Then it follows immediately from the definition of the canonical redundancy that

$$(1-R_{13})=(1-R_{12})(1-R_{23})$$

or

$$R_{13}=R_{12}+R_{23}-R_{12}R_{23},$$

the last term being negligible if R_{12} and R_{23} are both small. Thus, in a sequence of reductions, the values of $1-R$ multiply. By the approximation theorem, the same relation holds approximately (though not exactly) for the microcanonical redundancies. Note that the above relation between the redundancies in a chain of reductions implies, in particular, that

$$R_{13}\geqslant \max(R_{12}, R_{23}).$$

Hence, if the redundancy in any one of the links of a chain of reductions exceeds a critical value ϱ, so does the redundancy of the composite reduction. Compare this with ordinary significance testing on a fixed level ε which may very well lead us to reject the reduction from t_1 to t_2 or from r_2 to t_3 but accept the composite reduction from t_1 to t_3.

Rasch points out that, while the microcanonical redundancy is defined quite generally for reductive hypotheses, the canonical redundancy is defined and the approximation theorem proved for exponential families only, and he asks if similar results are available for other types of distributions. Not so far, but on the other hand the limitation to the exponential family (or, equivalently, to additive statistics) is only dictated by the fact that it covers most applications and is the only class of distributions for which the necessary analytical machinery has been developed.

References to the discussion

Berkson, J. (1966). Examination of randomness of α-particle emissions. *Research papers in statistics.* Festschrift for J. Neyman (ed. F. N. David), pp. 37–54. Wiley, London.

Fisher, R. A. (1943). Note on Dr. Berkson's criticism of tests of significance. *J. Amer. Statist. Assoc.* **38**, 103–104.

Kendall, M. G. (1943). *The advanced theory of statistics*, vol. 1, 1st ed. Griffin, London.

Steffensen, J. F. (1923). *Matematisk iagttagelseslære.* G. E. C. Gad, Copenhagen.