Scand. J. Statist. 2: 161–165, 1975

Reply to Sverdrup's Polemical Article Tests without Power

PER MARTIN-LÖF

University of Stockholm

Received June 1974

The title of Sverdrup's article is Tests without power, by which he apparently means Tests constructed without power considerations. For, although my exact tests, against which it is directed, are constructed without power considerations, they certainly do not lack power in the Neyman-Pearson sense.

A principle of statistical inference, like unbiasedness, invariance, the likelihood principle, or the principle that I have proposed, is a codification of statistical practice. Thus, what makes such a principle correct, is that it adequately accounts for well-established parts of statistical practice. There is no question of staring at the principle as it stands and trying to feel whether or not it is convincing. Conversely, to refute such a principle, we must show that it conflicts with the way statisticians actually behave or would behave in a particular case. This is what Sverdrup tries to do with the principle for testing reductive statistical hypotheses that I have called exact. To counter his criticism, I shall have to go through all his examples, showing that, in some of them, the hypothesis that he considers, although a hypothesis in the Neyman-Pearson sense, is not reductive, and that, in the remaining ones, despite of what Sverdrup says, it is the exact test, and neither the test favoured by him, nor the uniformly most powerful unbiased one, which is in agreement with common sense. (Thus these latter examples, rather than refuting the exact test, refute the principle of unbiasedness.) Whether this sense is in fact common, can of course only be decided by public discussion. If the publication of this polemic can contribute to such a discussion, it will have served its purpose.

Before turning to the examples, however, I want to make two comments on the introduction. First, Sverdrup says that my principle "disregards the alternatives except for the purpose of constructing the minimal sufficient statistic X under the a priori assumptions." (Here X is his notation for my $t(x)$.) This is misleading. In the Neyman-Pearson theory, a statistical model is a family H of probability distributions on a sample space X, and a hypothesis H_0 is a subset of H. In my theory, the sample space X is retained, but H is replaced by a statistic $t(x)$ and $H_0 \subset H$ by a statistic $u(x) = u(t(x))$ which factors through $t(x)$. Hence the alternatives $H_1 = H - H_0$ are in no way disregarded in my theory: they are just differently specified. Instead of specifying H_1 or $H = H_0 \cup H_1$, I specify a statistic $t(x)$ which induces a finer partitioning of the sample space than $u(x) = u(t(x))$. Cf. also Cox's question on p. 15 and my answer to it on p. 17 of Martin-Löf (1974*a*). Second, Sverdrup asks if it is meant as an assumption that the conditional (in my terminology, microcanonical) distribution of x given that $t(x) = t$ is uniform, and that, under the hypothesis, the distribution of x given that $u(x) = u(t(x)) = u$ is also uniform. Yes, this is an assumption. And it is this assumption which allows me to specify a statistical model by simply giving the sample space X (with its Riemannian metric in the continuous case) and the statistic $t(x)$. Otherwise, I would also have to specify a measure on X, and it would be mysterious where that measure should come from if it were not simply the uniform measure, that is, in the continuous case, the Riemannian measure determined by the metric. That an observation x can be described by the model whose sample space is X and whose statistic is $t(x)$, means precisely that x can be considered as drawn at random from (that is, according to the uniform distribution on) the set X_t of all outcomes y for which $t(y) = t =$ the observed value of $t(x)$. And the hypothesis is that x can be considered as drawn at random from the larger set X_u of all outcomes y for which $u(y) = u(t(y)) = u =$ the observed value of $u(x) = u(t(x))$.

Example 1. Card dealing in bridge. The sample space consists of all possible deals, and Sverdrup wants to test the hypothesis that

the mixing is perfect (no cheating)

against the alternative that

the deal is favourable to the dealer.

He also considers the reductive hypothesis that the statistic

$$t(x) = t \quad \text{if} \quad x \text{ belongs to } G_t$$

can be reduced to the trivial statistic

$$u(x) = \text{constant}.$$

This latter hypothesis is rejected by the exact test provided the ratio $N_{t(x)}/N$ is sufficiently small.

No doubt, choosing $u(x) = \text{constant}$ is the correct way of expressing the hypothesis of perfect mixing as a reductive hypothesis. But the exact test considered by Sverdrup tests this hypothesis against the alternative that the deal x can be considered as drawn at random from the set $G_{t(x)}$ of all deals which are as favourable to the dealer as the particular deal x, quite independently of the numerical value of $t(x)$. This is not the alternative in which Sverdrup is interested, namely, that the deal is favourable to the dealer.

If we index the sets G_t in such a way that $t(x)$ is the gain of the dealer measured in a suitable monetary unit, it makes sense to introduce the parametric model

$$p_a(x) = \frac{e^{at(x)}}{\sum_{t=1}^{r} N_t e^{at}}, \quad -\infty \leqslant a \leqslant +\infty.$$

In terms of this model, the exact test considered by Sverdrup tests the hypothesis $a = 0$ against the alternative $a \neq 0$, whereas the alternative which he actually has in mind is $a > 0$. Such one-sided hypotheses are not reductive and hence fall outside of my framework. And I have never suggested that the exact test should be applied to them.

Example 2. Non-paying passengers on tramcars. In this example, Sverdrup wants to test the hypothesis $p = 0.001$ or $p \leqslant 0.001$ (it is not quite clear which, but it does not matter) where p is the parameter of the geometric distibution $p(1-p)^{x-1}$, $x = 1, 2, \ldots, 0 < p \leqslant 1$, against the alternative $p > 0.001$. Again, this is a one-sided and not a reductive hypothesis, and hence it falls outside the scope of my theory. Changing the geometric distribution into

$$\frac{\binom{N-x}{a-1}}{\binom{N}{a}} = a\frac{(N-x)(N-x-1)\ldots(N-x-a+1)}{N(N-1)\ldots(N-a+1)}$$

does not remedy the situation as long as the alternative remains one-sided, that is, as long as we are testing $p = a/N = p_0$ (or $\leqslant p_0$) against $p > p_0$.

Example 3. This is an elaboration of the previous example. We have one sample $x_1, \ldots, x_m$ from the geometric distribution $p(1-p)^{x-1}$ and another sample $y_1, \ldots, y_n$ from the geometric distribution $q(1-q)^{y-1}$ and want to test the hypothesis $p = q$. If we consider the one-sided alternative $p > q$, the hypothesis is not reductive and hence falls outside of my framework. On the other hand, when considered against the two-sided alternative $p \neq q$, $p = q$ is nothing but the parametric specification of the reductive hypothesis

$$t = \left(\sum_1^m x_i, \sum_1^n y_j\right) \to \sum_1^m x_i + \sum_1^n y_j = u.$$

Putting

$$x = \sum_1^m x_i, \quad y = \sum_1^n y_j = u - x,$$

the exact test rejects this hypothesis in the tail of the distribution

$$p_u(x) = \frac{\binom{x-1}{m-1}\binom{u-x-1}{n-1}}{\binom{u-1}{m+n-1}}$$

which is related to the negative binomial distribution in the same way as the hypergeometric distribution is related to the ordinary binomial distribution.

Now, specialize, as Sverdrup does, to the case $m = 1$. If, in addition, $n = 1$, then

$$p_u(x) = \frac{1}{u-1}, \quad x = 1, 2, \ldots, u-1.$$

Hence, in this case, the exact test never rejects, that is, the critical level is constantly equal to one. This is entirely as it should be. When $m = n = 1$, whatever be the values of x and y, there is simply no information in the observations as to whether $p = q$ or $p \neq q$. Any values of x and y are compatible with the hypothesis $p = q$ provided p and q are sufficiently small. Sverdrup, on the other hand, seems to consider it as a defect that the exact test gives no guidance when $m = n = 1$. I cannot understand why.

When $m = 1$ and $n > 1$,

$$p_u(x) = \frac{\binom{u-x-1}{n-1}}{\binom{u-1}{n}}$$

decreases monotonically as x increases from 1 to $u-n$, and hence the exact test rejects the hypothesis

when x is sufficiently large, that is, close to $u-n$. Sverdrup finds it paradoxical that, although the hypothesis is two-sided, the critical region becomes one-sided, and thinks that we ought of course to reject the hypothesis when x is close to either 1 or $u-n$. It is also easy to see that a uniformly most powerful unbiased test exists and has a (randomized, in general) criterial region of the form favoured by Sverdrup. Now, since $p_u(x)$ is monotonically decreasing, rejecting when x is close to 1 amounts to *rejecting for the values which are most probable under the hypothesis*. Against this, I have two objections. First, I know of no case when one has rejected a statistical hypothesis after having observed the most probable value under the hypothesis, and I would not do so myself. Second, in order that it should be at all possible to reject the hypothesis for the most probable value of the distribution without randomization, the probability of this value must not exceed the level of significance ε ($=0.01$, say). This means that the distribution must contain at least $1/\varepsilon$ ($=100$ if $\varepsilon=0.01$) possible values, that is, be very much spread out and hence, on an appropriate scale, practically continuous. Considerations about the measuring accuracy of the kind that I shall make in connection with Example 4 then become crucial. Before turning to it, however, I must show how the notions of statistical model, reductive hypothesis and exact test are defined in the finite-dimensional continuous case.

The sample space X is then assumed to be an n-dimensional Riemannian manifold, that is, a manifold endowed with a Riemannian metric $G=(g_{ij})$ which determines the distance

$$ds=\sqrt{dx'Gdx}=\sqrt{\sum_{i,j} g_{ij}\,dx_i\,dx_j}$$

between two infinitesimally close points (whose local coordinate vectors are) x and $x+dx$. Here and in the following the sign $'$ denotes transposition. The metric, in turn, determines the Riemannian measure on X

$$d\lambda=\sqrt{\det G}\,dx_1\ldots dx_n$$

which is invariant under smooth coordinate changes. A statistic $t(x)$ is a smooth mapping from the sample space X onto a manifold T of dimension $p\leqslant n$ which is subjected to the following two conditions. First, letting

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

be the tangent map, it is required that $JG^{-1}J'$ should be invertible, that is,

$$\det JG^{-1}J'>0,$$

and constant on the surface X_t where $t(x)=t$, so that we can write

$$(JG^{-1}J')(x)=(JG^{-1}J')(t(x)).$$

Geometrically, this means that the infinitesimally close surfaces X_t and X_{t+dt} are parallel. Second,

$$f(t)=\int_{X_t} d\lambda_t=\lambda_t(X_t)<+\infty$$

for all t in T. Here λ_t is the surface measure on X_t, that is, the measure which is determined by the metric which X_t inherits from the metric on X.

Because of the first condition, the given metric G on X induces in a natural way a metric on T, namely, the metric

$$(JG^{-1}J')^{-1}.$$

Geometrically, the distance

$$\sqrt{dt'(JG^{-1}J')^{-1}dt}$$

between two infinitesimally close points t and $t+dt$ in this metric, is the perpendicular distance between the infinitesimally close surfaces X_t and X_{t+dt} measured in the metric G. This distance is welldefined since, by assumption, X_t and X_{t+dt} are parallel. The measure on T determined by the metric $(JG^{-1}J')^{-1}$ is

$$d\mu=\sqrt{\det (JG^{-1}J')^{-1}}\,dt_1\ldots dt_p=\frac{dt_1\ldots dt_p}{\sqrt{\det JG^{-1}J'}}.$$

This measure is invariantly defined (under smooth coordinate changes) since it is the Riemannian measure derived from the metric $(JG^{-1}J')^{-1}$ which, in turn, is determined by the given metric G on X and the statistic $t(x)$. The measure $d\lambda$ on X can be decomposed according to the (geometrically obvious) formula

$$d\lambda=d\lambda_t d\mu$$

Integrating out with respect to $d\lambda_t$, we see that the function $f(t)$ on T is the density with respect to $d\mu$ of the measure induced by the measure $d\lambda$ on X under the statistic $t(x)$. It is the function $f(t)$ which, in the discrete case, is simply the number of outcomes x such that $t(x)=t$. The above definition of $f(t)$ in the continuous case should replace the defective definition which I have given earlier in Martin-Löf (1970) and (1974*b*).

A reductive hypothesis states that

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

where $u(t)$ is a smooth mapping from T onto a manifold U of dimension $q \leqslant p$ which is such that the composite function $u(x) = u(t(x))$ is a statistic. (The difference $p - q$ is the number of degrees of freedom of the hypothesis.) With

$$K = \frac{du}{dt} = \left(\frac{\partial u_i}{\partial t_j}\right)$$

so that

$$\frac{du}{dx} = \frac{du}{dt}\frac{dt}{dx} = KJ,$$

this means, first, that $(KJ)G^{-1}(KJ)'$ should be invertible and constant on the surface X_u where $u(x) = u$, or, equivalently, that $K(JG^{-1}J')K'$ should be so on the surface T_u where $u(t) = u$, and, second, that

$$g(u) = \int_{X_u} d\lambda_u = \int_{T_u} f(t)\, d\mu_u < +\infty$$

for all u in U. Here $d\mu_u$ is the surface measure on T_u, that is, the measure which is determined by the metric which T_u inherits from the metric $(JG^{-1}J')^{-1}$ on T. Just as in the discrete case, the exact test rejects the hypothesis that $t(x)$ can be reduced to $u(x) = u(t(x))$ if the ratio

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

which satisfies

$$\int_{T_u} \frac{f(t)}{g(u)}\, d\mu_u = 1,$$

is sufficiently small. This ratio is the conditional probability density of $t(x)$ given that $u(x) = u$ with respect to the surface measure $d\mu_u$.

Example 4. We have samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ from normal distributions with means ξ and η and standard deviations σ_1 and σ_2, respectively, and want to test the hypothesis $\sigma_1 = \sigma_2$. The condition $\sigma_1 = \sigma_2$ is the parametric specification of the reductive hypothesis

$$t = (\bar{x}, \bar{y}, z_1, z_2) \to (\bar{x}, \bar{y}, z_1 + z_2) = u,$$

where

$$\bar{x} = \frac{1}{m}\sum_1^m x_i, \quad \bar{y} = \frac{1}{n}\sum_1^n y_j, \quad z_1 = \sum_1^m (x_i - \bar{x})^2,$$

$$z_2 = \sum_1^n (y_j - \bar{y})^2,$$

and hence the exact test can be applied. According to it, we shall reject the hypothesis when the ratio $f(t)/g(u)$ is sufficiently small, the functions $f(t)$ and $g(u)$ being defined as above. Computation yields

$$\frac{f(t)}{g(u)} = \frac{2}{B((m-1)/2, (n-1)/2)\, z^{(m+n-3)/2}} \times z_1^{(m-2)/2}(z - z_1)^{(n-2)/2}, \quad z = z_1 + z_2,$$

which is the conditional probability density of t for fixed u with respect to the measure

$$d\mu_u = \frac{1}{2}\sqrt{\frac{z}{z_1(z - z_1)}}\, dz_1.$$

That the exponents of z_1 and $z - z_1$ in the expression for $f(t)/g(u)$ are $(m-2)/2$ and $(n-2)/2$ and not $(m-3)/2$ and $(n-3)/2$ as in Sverdrups expression, is a consequence of my correction of the definition of $f(t)$ in the continuous case. Fortunately, the correction does not affect Sverdrup's argument in an essential way: we merely have to put $m = 2$ instead of $m = 3$. Indeed, when $m = 2$,

$$\frac{f(t)}{g(u)} = \frac{2}{B(1/2, (n-1)/2)\, z^{(n-1)/2}}(z - z_1)^{(n-2)/2}$$

which is a decreasing function of z_1 provided $n > 2$. Thus the exact test rejects the hypothesis $\sigma_1 = \sigma_2$ when z_1 is sufficiently large. Just as in Example 3, Sverdrup thinks that, since the hypothesis is two-sided, the critical region ought to be two-sided as well, that is, that we ought to reject if z_1 is either large or small. Now, there are at least three different ways of distributing the total level of significance $\varepsilon = \varepsilon_0 + \varepsilon_1$ between the two tails. Here ε_0 and ε_1 are the probabilities of rejecting the hypothesis when z_1/z is close to 0 and 1, respectively.

(1) Sverdrup proposes the rule of thumb $\varepsilon_0 = \varepsilon_1 = \varepsilon/2$. To me, this seems wholly arbitrary unless the distribution is symmetric, that is, $m = n$, in which case the principle of uniformly most powerful unbiasedness leads to precisely this result. If $m = n > 2$, the distribution is both symmetric and unimodal, and hence the exact test also yields $\varepsilon_0 = \varepsilon_1 = \varepsilon/2$.

(2) The uniformly most powerful unbiased test is always two-sided (even when m or $n = 2$) and

$$\varepsilon_0 \sim \frac{n-1}{m+n-2}\varepsilon, \quad \varepsilon_1 \sim \frac{m-1}{m+n-2}\varepsilon,$$

when $\varepsilon \to 0$. In the maximally asymmetric case $m = 2$ that Sverdrup considers,

$$\varepsilon_0 \sim \frac{n-1}{n}\varepsilon, \quad \varepsilon_1 \sim \frac{1}{n}\varepsilon.$$

Hence, if $m=2$ and n is not too small, the uniformly most powerful unbiased test places almost all of the significance level in the left tail, that is, the tail where the probability density is high. This must, I think, lead us to reject the principle of unbiasedness.

(3) The critical region of the exact test is two-sided when m and $n>2$ and one-sided when $m=2$ and $n>2$. Hence, in the latter case, $\varepsilon_0=0$ and $\varepsilon_1=\varepsilon$. This is what Sverdrup finds paradoxical and leads him to reject the principle on which the exact test is based.

My arguments for the exact test and against unbiasedness as well as the rule of thumb favoured by Sverdrup, are essentially the same as in Example 3. First, I know of no case where there has been agreement among statisticians that a hypothesis should be rejected after having observed the most probable outcome under the hypothesis. Second, one may wonder if this is at all possible in the continuous case when the inevitable limitations in the measuring accuracy are taken into account. A part of the critical region which is located where the probability density is close to its maximal value M must have length $\leqslant \varepsilon/M$. So, if it is at all to be possible to reject the hypothesis where the probability density is high, the class width h must satisfy $h \leqslant \varepsilon/M$. Choosing $\varepsilon=0.01$, we get, for the normal distribution with standard deviation σ, $h \leqslant 0.025\,\sigma$ and, for the exponential distribution with mean μ, $h \leqslant 0.01\,\mu$. These measuring accuracies are completely irreal. For example, it does not make sense to measure statures, which are normally distributed with $\sigma=6$ cm, with an accuracy $h \leqslant 1.5$ mm. So, when $m=2$, Sverdrup's test as well as the uniformly most powerful unbiased one give practically, even if not mathematically, the same result as the exact test carried out on the smaller level of significance ε_1. For Sverdrup, $\varepsilon_1=\varepsilon/2$ and, for the uniformly most powerful unbiased test, ε_1 is even smaller.

References

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

(1974*a*). 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.

(1974*b*). Exact tests, confidence regions and estimates. Pp. 121–138 in *Proceedings of the Conference on Foundational Questions in Statistical Inference*, Aarhus, May 7–12, 1973, Edited by Ole Barndorff-Nielsen, Preben Blaesild and Geert Schou, Memoirs No. **1**, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Sverdrup, E. (1974). Tests without power. *Scand. J. Statist.* **2**, 158–160.

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