The Annals of Mathematical Statistics
1971, Vol. 42, No. 3, 1003-1009

A SEQUENTIAL STUDENT TEST

BY GIDEON SCHWARZ

The Hebrew University

1. Introduction. Wald has suggested a sequential test [6] for testing the mean of a normal variable with unknown variance. Like his other sequential tests of composite hypotheses, this test has optimality properties only when the indifference region separating the hypotheses is removed altogether from the sample space. In reality this is rarely feasible, and those tests lead to considerable oversampling whenever the true parameter value lies in the indifference region (see Bechofer [1]). Here the problem is treated by the Asymptotic Shapes Method, which still requires the hypotheses to be separated, but admits the points between them as possible parameter points.

The Asymptotic Shapes Method for large-sample sequential testing of composite hypotheses was first developed in our previous paper ([3], 1962). In a later paper ([4], 1969) we extended the result to higher dimensional exponential families and to truncation parameters. The Student Problem had actually been treated in the unpublished version of our dissertation (Columbia University, 1960), but since convergence to the asymptotic shape was very slow (Fushimi [2]), the practical applicability of special cases was rather limited, and therefore only the general theory was included in [4]. A second-order correction, found by Fushimi (*loc. cit.*) for special distributions and loss functions and extended by us to one-dimensional exponential families ([5], 1969), made the method applicable to "real" problems. In this paper we extend the correction term to higher dimensions, show how it applies to the general problem of large-sample sequential testing, and finally implement the general result to testing the mean of a normal variable with unknown variance.

2. The correction term in s dimensions. The evaluation of the correction term involves essentially the behavior of L_p-norms for large p. We begin by studying a special case:

LEMMA 1. *Let $w_1, \cdots, w_k$ be independent linear functionals on E^s, $0 \leqq k \leqq s$, and denote by V the set where all the w_i are nonnegative. Let g be a linear functional which is nonnegative on V, and q a positive quadratic form. By w, denote one of the w_i, and by ρ a real number > -1, or, if $k = 0$, put $w \equiv 1$ and $\rho = 0$.*

Then, as $n \to \infty$

$$\log \int_V w^\rho \exp(-(q+g)n)\, d\tau = (\rho/(i+1)+s-j/2)\log n + O(1),$$

where $d\tau$ is the volume element in E^s, $j = \dim V \cap \{g = 0\}$, and $i = 0$ if $w \equiv 0$ on $V \cap \{g = 0\}$ and $i = 1$ otherwise.

Received November 18, 1969.

PROOF. (1) Since $g \geqq 0$ on V, $\{g = 0\}$ is a supporting hyperplane of V at the origin, and g is a nonnegative linear combination of the w_i, the dimension j of the set of contact is simply the number of w_i missing in the combination.

(2) For q, it suffices to consider the spherical case $q = q_0 \sum_{i=1}^{s} w_i^2$, where $w_1, \cdots, w_s$ is some completion of $w_1, \cdots, w_k$ to a basis. Any other positive quadratic form is bounded above and below by such spherical forms.

(3) The volume element $d\tau$ may be replaced by $dw_1 \times \cdots \times dw_s$. This can only change the integral by a constant factor.

(4) According to whether $i = 0$ or 1, w is one of the w_i occurring or missing in $g = \Sigma g_i w_i$.

(5) The integral now factors into s one-dimensional integrals. Of these $j-i$ are of the form

$$\int \exp(-q_0 w^2 n)\, dw$$

over the whole line or the positive half line, and in either case the logarithm of the integral is $-\frac{1}{2}\log n + O(1)$. There are also $s-j+i-1$ factors of the form

$$\int_0^\infty \exp[-(g_i w + q_0 w^2)n]\, dw,$$

and one more factor, which is of the form

$$\int_0^\infty w^\rho \exp[-(g_i w + q_0 w^2)n]\, dw$$

when $i = 0$, and of the form

$$\int_0^\infty w^\rho \exp(-q_0 w^2 n)\, dw,$$

when $i = 1$. The logarithms of the last three integrals are $-\log n + O(1)$, $-(\rho+1)\log n + O(1)$ and $-\frac{1}{2}(\rho+1)\log n + O(1)$, which can be seen by applying Lemmas 2 and 3 of [5], or by direct evaluation.

(6) Collecting the terms, the statement of the lemma follows.

The application of this lemma to the evaluation of the correction term depends on the fact, rigorously stated and proved in Lemma 1 of [5], that the asymptotic behavior of the L_p-norm of a function depends only on the behavior of the function in a neighborhood of its maximum. An accordingly strengthened version of our Lemma 1 is the following:

LEMMA 2. *Let u be a sphere around the origin of E^s. A set W is given such that $W \cap u = V \cap u$ with V defined as in Lemma 1. Also, let f be a strictly convex and twice-differentiable function whose maximum in W is attained at the origin. Define i, j, w and ρ as in Lemma 1, with the differential of $-f$ at the origin playing the role of g. Then as $n \to \infty$*

$$\log \int_W w^\rho e^{fn} d\tau = f(0)n - (\rho/(i+1) + s - j/2)\log n + O(1).$$

PROOF. The Taylor expansion of f begins with $f(0) - g - q$, where q is a positive quadratic form. In $W \cap u$ we must have $g \geqq 0$, or f would not attain the maximum in W at the origin. So the conditions of Lemma 1 are fulfilled by V, g, q, w and ρ, and its conclusion holds. By Lemma 1 of [5], we may replace the domain of integration V by W. Since $f - f(0)$ is bounded in a sufficiently small sphere around 0 between two positive multiples of $-g-q$, the conclusion of the lemma follows.

We are now given a sequence of independent observations ω_i whose common distribution forms an s-dimensional exponential family. With respect to some fixed measure, this distribution has a density of the form $\exp(\boldsymbol{\theta}\cdot\mathbf{x}(\omega)-b(\boldsymbol{\theta}))$. Here $\boldsymbol{\theta}=(\theta_1,\cdots,\theta_s)$ is a vector parameter, and $\mathbf{x}=(x_1,\cdots,x_s)$ a vector statistic. The dot denotes scalar product, and $b(\boldsymbol{\theta})$ a scalar function. The parameter $\boldsymbol{\theta}$ ranges over the parameter space $\Omega\subset E^s$. Two nonnegative loss functions $l_0(\boldsymbol{\theta})$ and $l_1(\boldsymbol{\theta})$ are also given on Ω. They describe the "loss incurred when rejecting H_i", $i=0, 1$. The closure in Ω of the set where l_i is positive is denoted by H_i, and the H_i are assumed to be d-testable (see [4]). We restrict our attention for the time being to H_0 and l_0.

To facilitate the treatment we now put some restrictions on Ω, H_0 and l_0.

The "natural parameter space" is the set of all $\boldsymbol{\theta}$ in E^s where $\exp(\boldsymbol{\theta}\cdot\mathbf{x})$ has a finite integral. We assume that Ω is the intersection of the natural parameter space with a polyhedron that has only s faces intersecting at each vertex. The polyhedron may be unbounded, and may even be all of E^s. The loss function l_0 is assumed to be of the form $l_0(\boldsymbol{\theta})=d(\boldsymbol{\theta})((w(\boldsymbol{\theta})-w_0)^+)^\rho$, with $d(\boldsymbol{\theta})$ locally bounded away from 0 and ∞, $w(\theta)$ a linear functional, and $\rho>-1$. Thus $H_0=\{\boldsymbol{\theta}\in\Omega \mid w(\boldsymbol{\theta})\geqq w_0\}$. Finally, assume that no vertex of Ω lies on the *loss boundary* $\{\boldsymbol{\theta} \mid w=w_0\}$.

These assumptions are fulfilled in many practical cases, and they could be somewhat relaxed, at the expense of brevity, to include other cases. We also assume an *a priori* distribution dF on Ω, and restrict ourselves to the case where it has, with respect to Lebesgue measure, a density locally bounded away from 0 and ∞. Many other prior distributions could be handled after the model of [5], and we shall not carry out this generalization.

The *a posteriori* risk of rejecting H_0 after having observed $\omega_1,\cdots,\omega_n$ such that $(1/n)\Sigma\mathbf{x}(\omega_i)=\mathbf{k}$ is

$$R_0(n,\mathbf{k})=\int_{H_0} l_0(\boldsymbol{\theta})\exp(\boldsymbol{\theta}\cdot\mathbf{k}-b(\boldsymbol{\theta}))dF/\int_\Omega \exp(\boldsymbol{\theta}\cdot k-b(\boldsymbol{\theta}))dF.$$

Denote by $\boldsymbol{\theta}(\mathbf{k})$ the unique $\boldsymbol{\theta}$ in E^s where grad $b(\boldsymbol{\theta})=E_\theta(\mathbf{x})=\mathbf{k}$, and by $\boldsymbol{\theta}^*(\mathbf{k})$ and $\boldsymbol{\theta}^0(\mathbf{k})$ the maximum likelihood estimates of $\boldsymbol{\theta}$ in Ω and H_0 respectively. Clearly if $\boldsymbol{\theta}(\mathbf{k})\in\Omega$, $\boldsymbol{\theta}^*(\mathbf{k})=\boldsymbol{\theta}(\mathbf{k})$.

THEOREM. *Under the above conditions, when* $n\to\infty$ *for a fixed* $\mathbf{k}$ *such that* $\boldsymbol{\theta}(\mathbf{k})\notin H_0$ *we have* $\log R_0=-n\log\lambda_0(\mathbf{k})-(\rho/(i+1)+(j-j_0)/2)\log\log n+O(1)$, *where* λ_0 *is the maximum likelihood statistic* $(\max_\Omega/\max_{H_0})\exp(\boldsymbol{\theta}\cdot\mathbf{k}-b(\boldsymbol{\theta}))$, $j=j(\mathbf{k})$ *is the dimension of contact between* Ω *and the hyperplane orthogonal to* $E_{\theta^*}(\mathbf{x})$ *that supports it at* $\boldsymbol{\theta}^*(\mathbf{k})$, *and* $j_0=j_0(\mathbf{k})$ *is similarly defined with* 0^* *and* Ω *replaced by* 0^0 *and* H_0. *If* $E_{\theta^*}(\mathbf{x})$ *is orthogonal to the loss boundary*, $i=1$; *otherwise* $i=0$.

PROOF. The function $\boldsymbol{\theta}\cdot\mathbf{k}-b(\boldsymbol{\theta})$ is strictly convex, since its second derivatives are given by the covariance matrix of the vector $\mathbf{x}$. Both integrals occurring in R_0 have the form appearing in Lemma 2, provided we shift the origin, separately in the numerator and the denominator, to the point where $\boldsymbol{\theta}\cdot\mathbf{k}-b(\boldsymbol{\theta})$ attains its maximum in H_0 and in Ω respectively, and ignore $d(\boldsymbol{\theta})$ and the a priori density

(the latter are bounded, and so do not affect the result). The domains of integration also have the required shape; hence, the conclusion of Lemma 2 applies to each of the integrals, and the statement of the theorem follows.

3. The test procedure. An analogous expansion holds for R_1, the risk of stopping and rejecting H_1. The stopping risk $R = \min(R_0, R_1)$ is the risk of stopping and deciding optimally, having observed $\omega_1, \cdots, \omega_n$ with given $\mathbf{k}$. The sample size that yields, for given $\mathbf{k}$, stopping risk r, is found by solving $\log R(n, \mathbf{k}) = \log r$ for n. Using the theorem, and defining $\Lambda = \max(\lambda_0, \lambda_1)$, the two-sided maximum likelihood statistic, we obtain after some manipulation

$$n = (\log \Lambda(k))^{-1} (\log r^{-1} - (\rho/(i+1) + (j-j')/2) \log \log r^{-1}) + O(1).$$

Here j' is the one among j_0 and j_1 whose index agrees with that of the larger among λ_0 and λ_1. According to Lemma 2 of [4] sampling should continue if the stopping risk is greater than $C\, c \log c^{-1}$, where c is the cost of an observation, and C a constant depending on the problem, but not on c.

On the other hand, if the stopping risk is less than c, even one more observation would be a waste. Therefore optimal stopping occurs for given $\mathbf{k}$, somewhere between the two values obtained if the above formula for n is applied once to $r = c$ and once to $r = C\, c \log c^{-1}$. In the first case, $\log r^{-1} = \log c^{-1}$ and $\log \log r^{-1} = \log \log c^{-1}$; in the second case $\log r^{-1} = \log c^{-1} - \log \log c^{-1} + O(1)$ and $\log \log r^{-1} = \log \log c^{-1} + O(1)$. The two expressions for n will therefore differ by $(\log \Lambda)^{-1} \log \log c^{-1} + O(1)$. A bounded difference would have been preferable, but this cannot be achieved unless Lemma 2 of [4] is improved, to push up the lower n, or stopping is shown to be called for at stopping risks larger than c, which would push down the upper n. We have not succeeded in doing either. However, some computations performed by Fushimi [2], as well as some heuristic considerations seem to indicate that the lower bound yields better approximations. We therefore substitute $r = c \log c^{-1}$ in the formula for n, and obtain the procedure "continue sampling as long as

$$n \leqq (\log \Lambda(\mathbf{k}))^{-1} (\log c^{-1} - (1 + \rho/(i+1) + (j-j')/2) \log \log c^{-1}) + O(1)."$$

In accordance with the term "asymptotic shape", we call $(\log \Lambda(k))^{-1}$ the *shape factor* and the expression in the following parentheses the *size factor*. The first depends only on $\mathbf{k}$. The second depends on c, and for fixed c its $\log c^{-1}$ part is fixed, and its $\log \log c^{-1}$ part can change with $\mathbf{k}$ only when the point where $\exp(\boldsymbol{\theta} \cdot \mathbf{k} - b(\boldsymbol{\theta}))$ attains its maximum on Ω, H_0 or H_1 passes to a different face.

4. The Student case. Now consider the problem of testing the hypothesis H_0: $\mu \leqq \mu_0$ sequentially against H_1: $\mu \geqq \mu_1 > \mu_0$ on the basis of observations normally distributed with expectation μ and unknown variance σ^2. To ensure d-testability of the hypotheses, the parameter space is restricted by putting an upper bound, say K, on the variance. The family of distributions takes on the required exponential form if we define $\mathbf{X}(\omega) = (\omega, \omega^2)$, $\boldsymbol{\theta} = (\mu/\sigma^2, -1/2\sigma^2)$ and $b(\boldsymbol{\theta}) =$

$-\theta_1^2/4\theta_2 - \frac{1}{2}\log(-\theta_2)$. We then have $\Omega = \{\boldsymbol{\theta} \mid -\infty < \theta_1 < \infty, -\infty < \theta_2 < -1/2K\}$ and $H_0 = \{\boldsymbol{\theta} \mid -\infty < \theta_1 \leqq -2\theta_2\mu_0, -\infty < \theta_2 < -1/2K\}$.

So Ω is a half-plane, and H_0 is the intersection of Ω with another half-plane. The loss function l_0 is assumed to have the form $(\mu_0 - \mu)^\rho(\sigma^2)^\tau d$ where d is bounded away from 0 and ∞, τ is arbitrary, and $\rho > -1$. In terms of $\boldsymbol{\theta}$, this becomes $(\theta_1 + 2\theta_2\mu_0)^\rho[(-2\theta_2)^{-\rho-\tau}d]$, with the factor in brackets locally bounded away from 0 and infinity throughout Ω. If we assume a locally bounded a priori joint density for μ and σ^2, this ensures such a density also for θ_1 and θ_2, since the Jacobian of the transformation is $2\sigma^6$. Thus all the conditions for the validity of the theorem are fulfilled, and we may return to regard the problem in terms of the more conventional parameters μ and σ^2.

Similarly, we pass from the sufficient statistics $\Sigma\omega_i$ and $\Sigma\omega_i^2$ to an equivalent pair, the sample mean $\bar{\omega}$ and the sample variance S^2. The logarithm of the joint density of the observations now takes on the form

$$-\frac{n}{2}\left(\frac{(\bar{\omega}-\mu)^2+S^2}{\sigma^2}+\log\sigma^2\right).$$

Maximizing this expression over all (μ, σ^2) with $\sigma^2 \leqq K$, we find:

(1) If $S^2 \leqq K$, the maximum is attained at $\mu = \bar{\omega}$, $\sigma^2 = S^2$, and equals $-\frac{1}{2}n(1+\log S^2)$;

(2) If $S^2 \geqq K$, the maximum is attained at $\mu = \bar{\omega}$, $\sigma^2 = K$, and equals $-\frac{1}{2}n(S^2/K+\log K)$.

Maximizing the expression with the additional constraint that $\mu \leqq \mu_0$, we find for $\bar{\omega} > \mu_0$

(3) If $(\bar{\omega}-\mu_0)^2+S^2 \leqq K$, the maximum is attained at $\mu = \mu_0$, $\sigma^2 = (\bar{\omega}-\mu_0)^2 + S^2$, and equals $-\frac{1}{2}n(1+\log((\bar{\omega}-\mu_0)^2+S^2))$;

(4) If $(\bar{\omega}-\mu_0)^2+S^2 \geqq K$, the maximum is attained at

$$\mu = \mu_0, \sigma^2 = K, \quad \text{and equals} \quad -\frac{n}{2}\left(\frac{(\bar{\omega}-\mu_0)^2+S^2}{K}+\log K\right).$$

Combining conditions (1) through (4) yields the following expressions for $\log\lambda_0(\omega_1, \cdots, \omega_n)$, valid for $\bar{\omega} > \mu_0$:

(1) If $(\bar{\omega}-\mu_0)^2+S^2 \leqq K$, $\log\lambda_0 = \frac{1}{2}\log(1+((\bar{\omega}-\mu_0)/S)^2)$;

(2) If $S^2 \leqq K \leqq (\bar{\omega}-\mu_0)^2+S^2$, $\log\lambda_0 = \dfrac{1}{2}\left(\dfrac{(\bar{\omega}-\mu_0)^2+S^2}{K}-1-\log\dfrac{S^2}{K}\right)$;

(3) If $S^2 \geqq K$, $\log\lambda_0 = (\bar{\omega}-\mu_0)^2/2K$ for $\bar{\omega} \leqq \mu_0$, we have $\boldsymbol{\theta}(\mathbf{k}) \in H_0$ and $\lambda_0 = 0$.

For λ_1, three similar formulae, with μ_0 replaced by μ_1 are valid when $\bar{\omega} < \mu_1$; when $\bar{\omega} \geqq \mu_1$, we have $\boldsymbol{\theta}(\mathbf{k}) \in H_1$ and $\log\lambda_1 = 0$.

The formulae for the two-sided maximum likelihood statistic $\Lambda = \max(\lambda_0, \lambda_1)$ are best expressed in terms of S^2 and the statistic $z = S^2 + \max((\bar{\omega}-\mu_0)^2,$

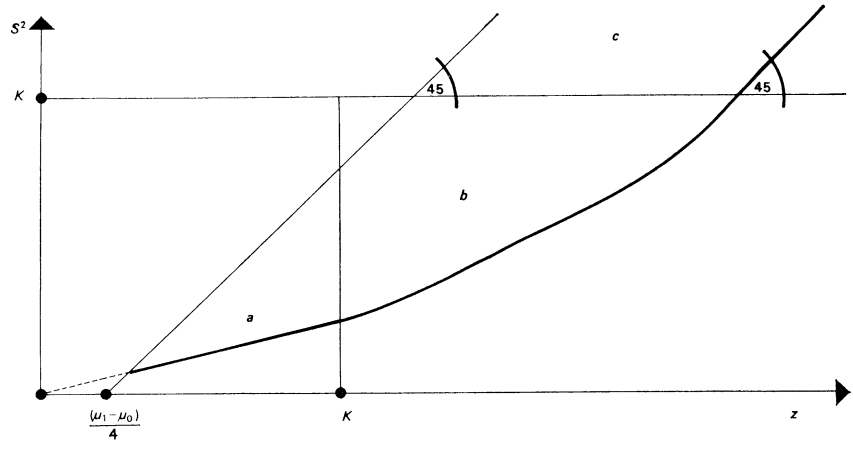
$(\bar{\omega}-\mu_1)^2$). This is simply the larger among the second sample moments around μ_0 and μ_1, and has the advantage that the level lines of Λ in the (z, S^2)-plane are explicitly describable.

(a) If $z \leqq K$ $\quad \log\Lambda = \frac{1}{2}\log\dfrac{z}{S^2}$.

(b) If $S^2 \leqq K \leqq z$ $\quad \log\Lambda = \dfrac{1}{2}\left(\dfrac{z}{K}-1-\log\dfrac{S^2}{K}\right)$.

(c) If $S^2 \geqq K$ $\quad \log\Lambda = \dfrac{z-S^2}{K}$.

The following diagram shows the respective domains of (*a*), (*b*) and (*c*), and a typical level-line of Λ.



In domain *a* the level-lines are straight lines emanating from the origin, given by $S^2 = \Lambda^{-2}z$.

In domain *c* they are parallel lines given by $S^2 = K\log\Lambda^{-1}+z$.

In the middle domain they are the exponential curves $S^2 = K\Lambda^{-1}\exp(z-K)/K)$, and are horizontal translates of each other.

As is easily seen, where the domains meet, the tangents match.

Above the line $S^2 = z-\frac{1}{4}(\mu_1-\mu_0)^2$ no point corresponds to a possible sample. Samples with $\bar{\omega} = \frac{1}{2}(\mu_0+\mu_1)$ yield $(z-S^2)$-points on this line, and all other samples yield points below the line.

The diagram can be used to find $\log\Lambda$, and hence the shape factor $(\log\Lambda)^{-1}$, for a given sample. We plot the point (z, S^2), and follow the level line until it hits the line $S^2 = K$. At that point, $z = K(1+\log\Lambda)$.

For the evaluation of the size factor, we return to the (θ_1, θ_2)-plane, and find

for $z \leqq K$, $\quad j = 2, j' = 1$ and $i = 0$;

for $z > K \geqq S^2$, $\quad j = 2, j' = 0$ and $i = 0$;

for $S^2 > K$, $\quad j = 1, j' = 0$ and $i = 0$.

(This time the three regions appear with their boundaries definitely assigned to one or the other region, since, unlike the shape factor, the size factor is not continuous at the boundaries.)

Thus the size factor is equal to $\log c^{-1} - (\rho + 2) \log \log c^{-1}$ in the "middle zone" $z > K \geqq S^2$, and $\log c^{-1} - (\rho + 3/2) \log \log c^{-1}$ in the rest of the sample space.

The test is now performed as follows: at each stage of sampling, the shape factor is multiplied by the size factor (where the former is determined by the diagram, and the latter is one of two possible values, that are computed in advance), and if their products exceed the present sample size, another observation is taken.

REFERENCES

[1] BECHOFER, R. (1960). A note on the limiting relative efficiency of the Wald sequential probability test. *J. Amer. Statist. Assoc.* **55** 660–663.

[2] FUSHIMI, M. (1967). On the rate of convergence of asymptotically optimal Bayes tests. *Rep. Statist. Appl. Res. Un. Japan. Sci. Engrs.* **14** 1–7.

[3] SCHWARZ, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist.* **33** 224–236.

[4] SCHWARZ, G. (1968). Sequential testing of truncation parameters. *Ann. Math. Statist.* **39** 2038–2043.

[5] SCHWARZ, G. (1969). A second-order approximation to optimal sampling regions. *Ann. Math. Statist.* **40** 313–315.

[6] WALD, A. (1947). *Sequential Analysis.* Wiley, New York.