The Annals of Mathematical Statistics
1971, Vol. 42, No. 5, 1742–1747

THE ASYMPTOTIC BEHAVIOR OF THE SMIRNOV TEST COMPARED TO STANDARD "OPTIMAL PROCEDURES"

By George Kalish and Piotr W. Mikulski
Morgan State College and University of Maryland

1. Summary. Let $X_1, X_2, \cdots, X_m, Y_1, Y_2, \cdots, Y_n$ be independent random samples from absolutely continuous distributions F and G respectively. Several standard tests of the hypothesis $H:F = G$ against the one-sided shift alternative $A:G(v) = F(v-\theta);(\theta > 0)$, are defined in terms of F. If, however, the true distributions of X's and Y's are $\Psi(v)$ and $\Psi(v-\theta)$ respectively, with Ψ not necessarily equal to F, these tests are no longer optimal. It will be shown that there exist continuous distributions Ψ (with density ψ), which are quite similar to F but for which the Smirnov test—in terms of generalized Pitman efficiency (defined below) is considerably superior.

2. Assumptions, definitions, notation. Let $N = m+n$ and $\tau = m/n$. Assume that $f = F'$ is unimodal (with mode assumed without loss of generality, to be at the origin) with finite variance. Suppose that assumptions 1, 2, 3, 5 of [6] are satisfied. Furthermore, assume that $g = -f'/f$ (as defined in [6]) is twice continuously differentiable, $\int_{-\infty}^{+\infty} g'(x)f(x)\,dx < \infty$ and g'' is F–integrable and uniformly continuous.

The standard tests which will be considered are the locally most powerful rank tests, the "Neyman" tests [7] and the likelihood ratio tests with test statistics $T_N^* = \sum_{j=1}^{N} E[g(V^{(j)})]Z_j$ and $T_N = [\tau\sum_{j=1}^{n} g(Y_j) - \sum_{j=1}^{m} g(X_i)]/(1+\tau)$ ([6], page 24). Note that the "Neyman" tests are locally equivalent to large sample likelihood ratio tests, hence the same test statistic can be used for both. Let

$$S_N = [mn/(m+n)]^{\frac{1}{2}} \sup_z (F_m(z) - G_n(z))$$

be the two-sample one-sided Smirnov statistic.

Let $e_{TT^*}(F; \Psi)$ denote the Pitman efficiency computed under Ψ of the "Neyman" test for F, to the LMP rank test for F. Generalizing the Pitman efficiency we shall define $\underline{e}_{ST}(F; \Psi) = \liminf_{i\to\infty} N_i(T)/N_i(S)$, where $N_i(T)$ and $N_i(S)$ are sample sizes of corresponding tests T_N and S_N, needed to achieve the same power β for the alternative $A:\theta = \theta_i$ with the same significance level $\alpha < \beta$, where $\theta_i \to 0$. The distribution F is used to define the test statistic T_N; then the calculations are carried out assuming that the true distribution is Ψ. Similarly define $\underline{e}_{ST^*}(F; \Psi)$.

3. Main results and proofs. Under present assumption we have

I. $$\sup_\Psi \underline{e}_{ST^*}(F;\Psi) = +\infty$$

II. $$\sup_\Psi \underline{e}_{ST}(F;\Psi) = +\infty.$$

Received November 13, 1969; revised July 29, 1970.

To show I, we first consider the following

LEMMA 1. *If Ψ is such that $\sigma^2(\theta)/\sigma^2(0) \to 1$ as $\theta \to 0$, where $\sigma^2(\theta_i) = \mathrm{Var}_{\Psi;\theta_i} T_{N_i}$, g' and g'' are Ψ integrable, then*

$$\lim_{i\to\infty} N_i(T)\theta_i^2 = \frac{(K_\beta - K_\alpha)^2(1+\tau)^2}{\tau[E_\Psi g'(X)]^2}\mathrm{Var}_\Psi\, g.$$

K_α denotes the root of the equation $(1/(2\pi)^{\frac{1}{2}})\int_{K_\alpha}^{\infty} e^{-t^2/2}\,dt = \alpha$.

PROOF. It can be shown that T_{N_i} appropriately standardized is asymptotically normal. For each sample size $N_i(T)$ there is a critical point C_i determined by $\Pr[T(N_i(T), 0) \geqq C_i] = \alpha$.

Under the hypothesis, we have $\lim_{i\to\infty}(1+\tau)C_i/(\tau N_i(T)\,\mathrm{Var}_\Psi\, g)^{\frac{1}{2}} = K_\alpha$ while under the alternatives

$$\lim_{i\to\infty}(1+\tau)\{C_i + [mn/(m+n)]\,.\,E_\Psi[g(X)-g(Y)]\}/(N_i(T)\tau\,\mathrm{Var}_\Psi\, g)^{\frac{1}{2}} = K_\beta$$

and hence

$$\lim_{i\to\infty}(N_i(T))^{\frac{1}{2}}E_\Psi[g(X)-g(Y)] = \frac{(K_\beta-K_\alpha)(1+\tau)}{\tau^{\frac{1}{2}}}(\mathrm{Var}_\Psi\, g)^{\frac{1}{2}}. \tag{3.1}$$

It should be pointed out that throughout this paper the sample sizes $N_i(T)$, $N_i(S)$, etc., are determined by the power β, which depends on the true distribution Ψ. Hence, in fact, we have $N_i(T;\Psi)$, which for short is denoted by $N_i(T)$. More explicitly, to justify the first limit we can observe that

$$C_i \simeq (\tau N_i(T;F)\,\mathrm{Var}_F\, g)^{\frac{1}{2}}K_\alpha/(1+\tau)$$

and hence

$$\frac{(1+\tau)C_i}{[\tau N_i(T;\Psi)\,\mathrm{Var}_\Psi\, g]^{\frac{1}{2}}} \simeq \left[\frac{N_i(T;F)\,\mathrm{Var}_F\, g}{N_i(T;\Psi)\,\mathrm{Var}_\Psi\, g}\right]^{\frac{1}{2}} K_\alpha \to K_\alpha,$$

since (as is well-known)

$$\frac{N_i(T;F)}{N_i(T;\Psi)} \to \frac{\mathrm{Var}_\Psi\, g}{\mathrm{Var}_F\, g}.$$

Considering the expansion $g(x+\theta) = g(x) + \theta g'(x) + g''(\bar\theta)\theta^2/2$, and applying Fatou's lemma, we find

$$\begin{aligned}
\lim_{i\to\infty}(N_i(T))^{\frac{1}{2}}E_\Psi[g(X)-g(Y)] &= \lim_{i\to\infty}(N_i(T))^{\frac{1}{2}}E_\Psi[g(X)-g(X+\theta_i)]\\
&= \lim_{i\to\infty}(N_i(T))^{\frac{1}{2}}E_\Psi\{-\theta[g'(X)+\theta_i/2g''(\bar\theta)]\}\\
&= -\lim_{i\to\infty}(N_i(T))^{\frac{1}{2}}\cdot\theta_i\cdot\lim_{i\to\infty}E_\Psi[g'(X)+\theta_i/2g''(\bar\theta)]\\
&= -\lim_{i\to\infty}(N_i(T))\cdot\theta_i\cdot E_\Psi[g'(X)].
\end{aligned}$$

Using (3.1) we find

$$\lim_{i\to\infty} N_i(T)\theta_i^2 = \frac{(K_\beta-K_\alpha)^2(1+\tau)^2\,\mathrm{Var}_\Psi\, g}{\tau[E_\Psi g'(X)]^2},$$

which proves the lemma.

LEMMA 2. $\limsup_{i\to\infty} N_i(S)\theta_i^2 \leqq K/\psi^2(0)$.

PROOF. Observe that $S_N \geqq S_N' = [mn/(m+n)]^{\frac{1}{2}}[F_m(0)-G_n(0)]$. S_N' is asymptotically normal with mean $ES_N' = n^{\frac{1}{2}}[\tau/(1+\tau)][\Psi(0)-\Psi(-\theta_i)]$ if $\theta = \theta_i$. Therefore we have

$$\beta = \beta(\theta_i) = \Pr(S_N > s_i \mid \theta_i) \geqq \Pr(S_N' > s_i \mid \theta_i) = \Pr\left(\frac{S_N'-ES_N'}{\sigma_N} > \frac{s_i - ES_N'}{\sigma_N} \middle| \theta_i\right),$$

where $s_i \to s$ are defined by $\Pr(S_N > s_i \mid H) = \alpha$. We observe that

$$\sigma_N = (\operatorname{Var}_{\theta_i} S_N')^{\frac{1}{2}}_{i\to\infty} \to \sigma = \{\psi(0)[1-\psi(0)]\}^{\frac{1}{2}}.$$

Hence we have $\liminf_{i\to\infty}(s_i - ES_N')/\sigma_N \geqq K_\beta$. Upon expanding $\Psi(-\theta_i) = \Psi(0)-\theta_i\psi(0)+\psi'(\bar{\theta})\theta^2/2$ and replacing $n = N_i(S)/(1+\tau)$ in ES_N' we have $\limsup_{i\to\infty} N_i(S)\theta_i^2 \leqq (1+\tau)^3(s-K_\beta\sigma)^2/\tau^2\psi^2(0)$, which proves the lemma.

THEOREM 1. *Let F and Ψ be cumulative distribution functions. Then*

$$e_{ST}(F;\Psi) \geqq \frac{K_1\psi^2(0)}{[E_\Psi g'(X)]^2}\operatorname{Var}_\Psi g.$$

PROOF. Observe that $e_{ST}(F;\Psi) \geqq \liminf_{i\to\infty} N_i(T)\theta_i^2/\limsup_{i\to\infty} N_i(S)\theta_i^2$ and apply the previous two lemmas.

THEOREM 2. *For any distribution F*, $\sup_\Psi e_{ST}(F;\Psi) = +\infty$.

PROOF. Let $\psi(x) = \gamma f(x)+(1-\gamma)\sigma f(x\sigma)$. We observe that $\psi(x)$ is a density function of a random variable $W = [U+(1-U)/\sigma]X$ where X has distribution F and U is a Bernoulli random variable independent of X and such that $\Pr(U = 1) = \gamma$.

We can easily see that

$$\operatorname{Var}_\Psi g(X) = \operatorname{Var}_F g(W) \geqq E_U\{\operatorname{Var}[g(W) \mid U]\} \geqq \gamma \operatorname{Var}_F g(X).^1$$

Furthermore we observe that $\psi^2(0) \geqq (1-\gamma)^2\sigma^2 f^2(0)$.

Now we consider

LEMMA 3. *There is a K and σ^* such that if $\sigma \geqq \sigma^*$ then $E_\Psi g'(X) \leqq K$.*

PROOF. Under present assumptions it can be shown that

$$\lim_{\sigma\to\infty}\sigma\int_{-\infty}^{\infty} g'(x)f(\sigma x)\,dx = g'(0).$$

Hence $\lim_{\sigma\to\infty} E_\Psi g'(X) = \gamma E_F g'(X)+(1-\gamma)g'(0)$ and the lemma follows.

Substituting these results into Theorem 1, we find

$$e_{ST}(F;\Psi) \geqq [K_1(1-\gamma)^2\sigma^2 f^2(0)\gamma \operatorname{Var}_F g]/K^2 = K^*(1-\gamma)^2\sigma^2$$

which completes the proof of Theorem 2, since σ can be arbitrarily large.

---

[1] The authors wish to thank the referee for his suggestion of a very short and elegant proof of this portion of Theorem 2, which was much longer in the original paper.

THEOREM 3. *For any distribution F there is a distribution Ψ such that the lower bound of the relative asymptotic efficiency of the Smirnov test to the likelihood ratio test derived for F exceeds C (where C is an arbitrary constant).*

PROOF. Follows from Theorem 2 and the asymptotic equivalence of "Neyman" tests and likelihood ratio tests (see [2], page 1137).

THEOREM 4. *For any distributions F and Ψ there exists a constant K_2 such that*

$$\underline{e}_{ST^*}(F;\Psi) \geqq \frac{K_2\psi^2(0)}{\{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx\}^2}.$$

PROOF: It is known [6] that

$$e_{T^*T}(F;\Psi) = \left[\frac{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx}{\int_{-\infty}^{\infty} g'(x)\psi(x)\,dx}\right]^2 \frac{\mathrm{Var}_\Psi\, g}{\mathrm{Var}_F\, g}$$

where $J(z) = g(F^{-1}(z))$.

Observing that $\underline{e}_{ST^*}(F;\Psi) \geqq \underline{e}_{ST}(F;\Psi)\,.\,e_{TT^*}(F;\Psi)$, the result follows by applying Theorem 1, where $K_2 = K_1\,\mathrm{Var}_F\, g$.

THEOREM 5. *For any distribution F $\sup_\Psi \underline{e}_{ST^*}(F;\Psi) = +\infty$.*

PROOF. In order to prove the theorem it is sufficient to show that for any C there is a Ψ such that

$$\frac{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx}{\psi(0)} \leqq C.$$

This can be done by making $\psi(x) = f(x)$ outside a fixed interval, while replacing $f(x)$ inside an interval by a density with a sharp spike. One such construction is the following:

Arbitrarily select points u_1 and u_2 satisfying $f(u_1) = f(u_2)$. Since f is assumed to be unimodal with mode at the origin it may be assumed $u_1 < 0$ and $u_2 > 0$.

Let $A = \int_{u_1}^{u_2} f(x)\,dx - (u_2-u_1)f(u_1)$. Define D by $\frac{1}{2}(u_2-u_1)D = A$.

Let ε be a real number satisfying $\varepsilon < \min\{u_2; -u_1; D^2\}$ and define P by $D-P = \varepsilon^{\frac{1}{2}}$. Let $\varepsilon K = A+\varepsilon P-\frac{1}{2}(u_2-u_1)P$; $V_1 = P+\varepsilon P/u_1$; $V_2 = P-\varepsilon P/u_2$;

$$\begin{aligned}
\psi(x) &= f(x), && x < u_1,\\
&= \frac{-P}{u_1}x+P+f(u_1), && u_1 \leqq x \leqq -\varepsilon,\\
&= \frac{K-V_1}{\varepsilon}x+K+f(u_1), && -\varepsilon < x < 0,\\
&= -\frac{K-V_2}{\varepsilon}x+K+f(u_1), && 0 \leqq x \leqq \varepsilon,\\
&= -\frac{P}{u_2}x+P+f(u_1), && \varepsilon < x < u_2,\\
&= f(x), && u_2 \leqq x.
\end{aligned}$$

Since $F(x) = \Psi(x)$ if $x \in (-\infty, u_1]$ or $x \in [u_2, \infty)$,

$$\frac{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx}{\psi(0)} \leqq \frac{\int_{-\infty}^{\infty} J'[F(x)]f^2(x)\,dx}{\psi(0)} + \frac{\int_{u_1}^{u_2} J'[\Psi(x)]\psi^2(x)\,dx}{\psi(0)}$$

$$\leqq \frac{\int_{-\infty}^{\infty} g'(x)f(x)\,dx}{\psi(0)} + \sup_{x \in [u_1,u_2]} \frac{g'}{f}[F^{-1}(\Psi(x))] \cdot \frac{\int_{u_1}^{u_2} \psi^2(x)\,dx}{\psi(0)}.$$

Consider $\int_{u_2}^{u_1} \psi^2(x)\,dx$. Upon omitting negative terms and using the fact $P = D - \varepsilon^{\frac{1}{2}} < D$, simple integration yields

$$\int_{u_1}^{u_2} \psi^2(x)\,dx \leqq \frac{(D+f(u_1))^3}{3(D-\varepsilon^{\frac{1}{2}})} \cdot (u_2 - u_1) + \frac{2K}{3} \frac{\varepsilon K}{1-(D/K)} \left(\frac{f(u_1)}{K} + 1 \right)^3.$$

Also, since $\Psi(x)$ and $F(x)$ are both continuous increasing functions on $[u_1, u_2]$ and $\Psi(u_1) = F(u_1)$, $\Psi(u_2) = F(u_2)$, it follows that

$$\sup_{x \in [u_1,u_2]} \frac{g'}{f}[F^{-1}(\Psi(x))] = \sup_{x \in [u_1,u_2]} \frac{g'}{f}(x).$$

Hence

$$\frac{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx}{\psi(0)} < \frac{\int_{-\infty}^{\infty} g'(x)f(x)\,dx}{K} + \sup_{x \in [u_1,u_2]} \frac{g'}{f}(x) \cdot \left[\frac{(D+f(u_1))^3}{3K(D-\varepsilon^{\frac{1}{2}})}(u_2 - u_1) + \frac{2}{3} \frac{\varepsilon K}{1-(D/K)} \left(\frac{f(u_1)}{K} + 1 \right)^3 \right].$$

Since $K > (u_2 - u_1)/2\varepsilon$ and $\varepsilon K < (\frac{1}{2})(u_2 - u_1)\varepsilon^2 + \varepsilon D$,

$$\frac{\int_{-\infty}^{\infty} J'[\Psi(x)]\psi^2(x)\,dx}{\psi(0)} \leqq \frac{2\varepsilon^{\frac{1}{2}} \int_{-\infty}^{\infty} g'(x)f(x)\,dx}{u_2 - u_1} + B(\varepsilon) \sup_{x \in [u_1,u_2]} \frac{g'}{f}(x)$$

where

$$B(\varepsilon) = \tfrac{2}{3}\varepsilon^{\frac{1}{2}} \frac{(D+f(u_1))^3}{D-\varepsilon^{\frac{1}{2}}} + \frac{\varepsilon^{\frac{1}{2}}}{3} \frac{[u_2 - u_1 + 2D\varepsilon^{\frac{1}{2}}]}{[1 - 2D\varepsilon^{\frac{1}{2}}/(u_2 - u_1)]} \left[\frac{2f(u_1)\varepsilon^{\frac{1}{2}}}{u_2 - u_1} + 1 \right]^3.$$

Clearly $\lim_{\varepsilon \to 0} B(\varepsilon) = 0$. Hence the theorem follows.

COROLLARY. *For any constant C there is a distribution Ψ such that the relative asymptotic efficiency of the Smirnov test to*: (a) *Student's* (b) *Fisher–Yates* (c) *Wilcoxon tests exceeds C.*

It is interesting to observe that the graphs of F and Ψ can be very similar; nevertheless, the Smirnov test can be much more efficient than any of the above mentioned tests optimal for corresponding F.

REFERENCES

[1] Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain non-parametric test statistics. *Ann. Math. Statist.* **29** 972–994.

[2] Hájek, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Statist.* **33** 1124–1147.

[3] Hodges, J. L., Jr. and Lehmann, E. L. (1956). The efficiency of some non-parametric competitors of the *t*-test. *Ann. Math. Statist.* **27** 329–335.

[4] Kalish, G. (1964). The relative asymptotic efficiency of optimal non-parametric tests. M.A. Thesis, Univ. of Maryland.

[5] Lehmann, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[6] Mikulski, P. W. (1963). On the efficiency of optimal nonparametric procedures in the two sample case. *Ann. Math. Statist.* **34** 22–32.

[7] Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics. The Harald Cramér Volume.* Wiley, New York, 213–234.

[8] Ramachandramurty, P. V. (1966). On the Pitman efficiency of one-sided Kolmogorov and Smirnov tests for normal alternatives. *Ann. Math. Statist.* **37** 940–944.