Testing Simple Hypotheses

R.L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Box 90251
Durham, NC 27708, USA

Summary: Pre-experimental Frequentist error probabilities do not summarize adequately the strength of evidence from data. The *Conditional Frequentist* paradigm overcomes this problem by selecting a "neutral" statistic S to reflect the strength of the evidence and reporting a *conditional* error probability, given the observed value of S.

We introduce a neutral statistic S that makes the Conditional Frequentist error reports *identical* to Bayesian posterior probabilities of the hypotheses. In symmetrical cases we can show this strategy to be optimal *from the Frequentist perspective*. A Conditional Frequentist who uses such a strategy can exploit the consistency of the method with the Likelihood Principle—for example, the validity of sequential hypothesis tests even if the stopping rule is informative or is incompletely specified.

1. Introduction

1.1 Pre-experimental Frequentist Approach

The classical frequentist procedure for testing a simple hypothesis such as $H_0: \ X \sim f_0(\cdot)$ against a simple alternative $H_1: \ X \sim f_1(\cdot)$ upon observing some random variable X taking values in a space $\mathcal{X}$ is simple: select any measurable set $\mathcal{R} \subset \mathcal{X}$ (the *critical* or *rejection* region) and

Pre-exp Freq: If $X \in \mathcal{R}$, then **Reject** H_0 and report
error probability $\alpha \equiv \mathsf{P}_0[X \in \mathcal{R}]$
If $X \notin \mathcal{R}$, then **Fail to reject** H_0 and report
error probability $\beta \equiv \mathsf{P}_1[X \notin \mathcal{R}]$

The *Size* and *Power* of the test are, respectively, $\alpha = \mathsf{P}_0[X \in \mathcal{R}]$ and $1-\beta = \mathsf{P}_1[X \in \mathcal{R}]$, the probabilities of rejection under the null and alternative hypotheses. It is desirable to have both α and β small, since they represent the probabilities of making two possible kinds of errors: α is the probability of a *Type-I error*, erroneously rejecting the null hypothesis H_0 (by observing $x \in \mathcal{R}$), if in fact H_0 is true; while β is that of a *Type-II error*, erroneously failing to reject H_0 (by observing $x \notin \mathcal{R}$), if in fact H_1 is true.

The Neyman-Pearson lemma (see, e.g., Lehmann (1986)) shows that these error probabilities are minimized by rejection regions of the form

$$\mathcal{R} = \{x \in \mathcal{X} : B(x) \le r_c\}$$

Research supported by United States National Science Foundation grant DMS-9305699 and Environmental Protection Agency grant CR822047-01-0, and performed in collaboration with James O. Berger and Lawrence D. Brown.

for an arbitrary constant (the *critical value*) $r_c > 0$, where $B(x)$ denotes the *Likelihood Ratio statistic*

$$B(x) \equiv \frac{f_0(x)}{f_1(x)}.$$

The probability distribution for $B \equiv B(X)$ will depend on that of X, of course; if we denote by $F_\theta(b) \equiv \mathsf{P}_\theta[B(X) \le b]$ the distribution function for the likelihood ratio under the distribution $X \sim f_\theta(\cdot)$, for $\theta \in \{0, 1\}$, then the size and power of this test are $\alpha = F_0(r_c)$ and $1 - \beta = F_1(r_c)$, respectively. The critical value r_c is often chosen to attain a specified Type-I error probability α (by setting $r_c \equiv {F_0}^{-1}(\alpha)$) or to attain equal probabilities of the two kinds of errors (by setting $r_c \equiv (F_0 + F_1)^{-1}(1)$). The Neyman-Pearson test depends on the data only through $B(X)$, and its complete specification requires only knowledge of the distribution functions $F_0(b)$ and $F_1(b)$ for all $b > 0$.

1.2 Objections

The hallmark virtue of this method is its *Frequentist Guarantee*: the strong law of large numbers implies that in a long series of independent size-α tests of true hypotheses H_0, at most $100\,\alpha\%$ will be erroneously rejected.

Nevertheless many statisticians and investigators object to fixed-level hypothesis testing in this form, on the grounds that it does not distinguish between marginal evidence (for example, that on the boundary $\partial\mathcal{R}$ of the rejection region) and extreme evidence.

1.2.1 A Short "Aside" on p-Values

Some investigators address this problem by reporting not only the decision to reject the hypothesis but also the $p-$value, the "probability of observing evidence against H_0 as strong or stronger as that actually observed, if the hypothesis is true"; in this context, that is simply

$$p = \mathsf{P}_0[\text{ More extreme evidence} \mid X = x] = F_0\big(B(x)\big).$$

Although p-values do offer some indication of relative strengths of evidence for different outcomes, they suffer from two fatal flaws: they distort the strength of evidence (this is illustrated in Example 1 below, where data with $\bar{x}_n = -.18$ offer strong evidence in favor of H_0, yet $p = 0.05$ leading many investigators to reject H_0 in favor of H_1; or the case $\bar{x}_n = 0$, equally improbable under the two hypotheses, yet with $p = 0.0023$ suggesting strong evidence in favor of H_1). More alarming perhaps is that they are invalid from the Frequentist perspective, because

$$\mathsf{P}_0[p \equiv F_0\big(B(X)\big) \le \alpha' \;\mid\; X \in \mathcal{R}] > \alpha'$$

for $\alpha' < \alpha$ (i.e., the distribution of p for rejected but nevertheless true hypotheses is not subuniform), in violation (conditional on rejection) of the Frequentist Guarantee.

1.3 Example 1

Let X be a vector of n independent observations from a normal distribution with known variance (say, one) but uncertain mean, known only to be one of two possible values (say, $\mu_0 = -1$ or $\mu_1 = +1$). The Likelihood Ratio upon observing $X = x$ depends only on the average $\bar{x}_n \equiv \Sigma\, x_i/n$:

$$B(x) = \frac{f_0(x)}{f_1(x)} = \frac{(2\pi)^{-n/2} e^{-\Sigma(x_i+1)^2/2}}{(2\pi)^{-n/2} e^{-\Sigma(x_i-1)^2/2}} = e^{-2n\bar{x}_n},$$

so a Neyman-Pearson test would reject $H_0 : \mu = \mu_0$ for large values of $\bar{X}$. The size and power of a test that rejects H_0 when $\bar{X} \geq c$ would be $\alpha = \Phi\big(-\sqrt{n}(c+1)\big)$ and $1 - \beta = 1 - \Phi\big(\sqrt{n}(c-1)\big)$, respectively. With $n = 4$ and $c = 0$, for example, we have equal error probabilities $\alpha = \beta = \Phi(-\sqrt{4}) \approx 0.025$. But these *pre-experimental* reports of error probability do not distinguish between an observation with $\bar{X} = \bar{x}_n = 0.1$, offering at best marginal evidence against H_0, and and extreme observation like $\bar{x}_n = 1.0$, a compelling one; in each case the reported error probability is 0.025, while the p-value is $p = 1 - F_0\big(B(x)\big) = \Phi\big(-\sqrt{n}(1+\bar{x}_n)\big)$:

Data	**LHR**	**Pre-Exp Freq**		**p-Value**
$\bar{x}_n$	$B(x)$	α	β	p
-0.18	4.221	0.025	0.025	0.050
0.00	1.000	0.025	0.025	0.023
0.10	0.449	0.025	0.025	0.014
0.37	0.052	0.025	0.025	0.003
1.00	3.4×10^{-4}	0.025	0.025	3.2×10^{-5}

2. The Conditional Frequentist Paradigm

A broad effort to find valid frequentist tests that better distinguish extreme from marginal evidence was begun by Kiefer (1975, 1977) and advanced by Kiefer and Brownie (1977), Brown (1978), and others (see Berger and Wolpert (1988) for references). Briefly, the idea is to choose a "neutral" (possibly ancillary) statistic S intended to reflect how extreme the data are, without offering evidence either for or against H_0; in Example 1, for example, one might choose $S(X) = |\bar{X}|$ or $\bar{X}^2$. With the same rejection region as before, the *conditional* frequentist reports:

Cond'l Freq: If $X \in \mathcal{R}$, then **Reject** H_0 and report
error probability $\alpha(s) \equiv \mathsf{P}_0[X \in \mathcal{R} | S = s]$
If $X \notin \mathcal{R}$, then **Fail to reject** H_0 and report
error probability $\beta(s) \equiv \mathsf{P}_1[X \notin \mathcal{R} | S = s]$

Now the error-reports $\alpha(S)$, $\beta(S)$ are data-dependent random variables, and the frequentist guarantee takes the form $\mathsf{P}_0[\alpha(S) \leq p] \leq p$ for every $p \in (0,1)$. In Example 1, for $S(X) = |\bar{X}|$, we have

Data	**Pre-exp Freq**		**p-Value**	**Cond'l Freq**	
$\bar{x}_n$	α	β	p	$\alpha(s)$	$\beta(s)$
-0.18	0.025	0.025	0.050	0.81	0.19
0.00	0.025	0.025	0.023	0.50	0.50
0.10	0.025	0.025	0.014	0.31	0.69
0.37	0.025	0.025	0.003	0.05	0.95
1.00	0.025	0.025	3.2×10^{-5}	3.4×10^{-4}	1.00

2.1 Bayesian Tests

Bayesian decision-theoretic methods begin by quantifying the cost L_θ of making an error when $H_\theta: \; X \sim f_\theta$ is true, for $\theta \in \{0,1\}$. The "state of nature" θ is regarded as uncertain and therefore random, with some probability distribution $\pi_0 = \mathsf{P}[H_0] = \mathsf{P}[\theta = 0]$, $\pi_1 = \mathsf{P}[H_1] = 1 - \pi_0$. Upon observing the data $X = x$ the *posterior probability of* H_0 is computed,

$$\pi_0^\star = \pi[H_0 | X{=}x] = \frac{\pi_0 f_0(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} = \frac{\frac{\pi_0}{\pi_1} B(X)}{1 + \frac{\pi_0}{\pi_1} B(X)};$$

this would be the error probability if H_0 were rejected.

The optimal strategy for minimizing the expected loss depends only on the prior odds $\rho \equiv \pi_0/\pi_1$, the loss ratio $\ell \equiv L_0/L_1$, and the Likelihood Ratio (also called a *Bayes factor*) $B \equiv B(X)$:

Bayesian: If $\rho B \leq \ell$, then **Reject** H_0 and report
error probability $\pi_0^\star \equiv \mathsf{P}[H_0|X] = \rho B/(1+\rho B)$
If $\rho B > \ell$, then **Accept** H_0 and report
error probability $\pi_1^\star \equiv \mathsf{P}[H_1|X] = 1/(1+\rho B)$

Error reports for rejected hypotheses will never exceed $\ell/(1+\ell)$, so setting $\ell = \alpha/(1-\alpha)$ would lead to a test with a guaranteed level $\pi_0^\star \leq \alpha$; conversely, the choice $\ell = \rho$ will always prefer the more likely hypothesis. For example, with equal losses (so $\ell = 1$) and a priori equal probabilities for the two hypotheses (so $\rho = 1$) in Example 1, for $S(X) = |\bar{X}|$, we have:

Data	**LHR**	**Cond'l Freq**		**Bayesian**	
$\bar{x}_n$	$B(x)$	$\alpha(s)$	$\beta(s)$	$\pi_0^\star$	$\pi_1^\star$
-0.18	4.221	0.81	0.19	0.81	0.19
0.00	1.000	0.50	0.50	0.50	0.50
0.10	0.449	0.31	0.69	0.31	0.69
0.37	0.052	0.05	0.95	0.05	0.95
1.00	3.4×10^{-4}	3.4×10^{-4}	1.00	3.4×10^{-4}	1.00

In this example we always have $\alpha(S) = \pi_0^\star$ and $\beta(S) = \pi_1^\star$; indeed both $\alpha(S)$ and $\pi_0^\star$ are given by $B/(1+B) = 1/(1+e^{2n\bar{x}_n})$, so they must be equal for all n and x. Is it possible that in every example we can find some statistic S for which the Conditional Frequentist and Bayesian tests coincide? The answer is *almost* "yes."

3. The Proportional Tail Statistic S_ρ

Suppose that $B \equiv B(X)$ has an absolutely continuous distribution under both hypotheses, so both $F_0(\cdot)$ and $F_1(\cdot)$ are equal to the integrals of their derivatives $F_0'(\cdot)$ and $F_1'(\cdot)$.

Lemma. *For all $b > 0$, $F_0'(b) \equiv bF_1'(b)$.*

Proof.

$$\begin{aligned}
F_0(b) &= \int_0^b F_0'(y)dy = \mathsf{P}_0[B(X) \le b] \\
&= \int_{\{x:\ B(x)\le b\}} f_0(x)dx \\
&= \int_{\{x:\ B(x)\le b\}} B(x)f_1(x)dx \\
&= \int_0^b yF_1'(y)dy.
\end{aligned}$$

The last step follows from the change of variables $y = B(x) \equiv \frac{f_0(x)}{f_1(x)}$. Now differentiate both sides with respect to b (by our supposition above that F_0 and F_1 be absolutely continuous) to complete the proof. $\diamond$

For each $\rho > 0$ define the *proportional tail statistic*

$$S_\rho \equiv \min\big(F_1(B), \rho - \rho F_0(B)\big).$$

Notice that $F_1(\cdot) + \rho F_0(\cdot)$ increases continuously from from 0 to $(1+\rho)$ on $[0,\infty)$, so there is a smallest $b^* > 0$ at which $F_1(b^*) = \rho\big(1 - F_0(b^*)\big)$.

For any critical value $r_c \le b^*$ and likelihood ratio $b < r_c$, the conditional distribution of B given $S_\rho = F_1(b)$ is concentrated on exactly two points: $B = b$ and $B = b'$, where $b' > b^*$ satisfies

$$\frac{F_1(b)}{1 - F_0(b')} = \frac{\mathsf{P}_1[B \le b]}{\mathsf{P}_0[B > b']} = \rho$$

(this explains the statistic's name). Simple computation using the Lemma reveals that the conditional (on S_ρ) probabilities for B to take on the two values (b, b') are in the ratio $\rho b : 1$.

Let $a_c = {F_0}^{-1}\big(1 - F_1(r_c)/\rho\big)$ be the number $a > b^*$ satisfying $\frac{F_1(r_c)}{1-F_0(a)} = \rho$. Then similarly if $b > a_c$ then the conditional distribution of B, given $S_\rho = \rho\big(1 - F_0(b)\big)$, is concentrated on two points, with probabilities in the same ratio.

Thus the error-report for a Conditional Frequentist test using S_ρ as the neutral statistic, and $\mathcal{R} = \{x : B(x) \le r_c\}$ for a critical region, would be:

Prop Tail: If $B \le r_c$, then **Reject** H_0 and report
error probability $\alpha(s) = \rho B/(1 + \rho B)$
If $B > a_c$, then **Accept** H_0 and report
error probability $\beta(s) = 1/(1 + \rho B)$

If $r_c < B < a_c$ we cannot reject H_0 (since $X \notin \mathcal{R}$) but the conditional error report would be $\beta(s) \equiv \mathsf{P}_1[B > r_c | S_\rho] = 1$ (since $S_\rho^{-1}(s) = \{b, b'\} \subset (r_c, \infty)$ for $s = S_\rho(b)$), making "acceptance" of H_0 unappealing; we regard this evidence as insufficiently compelling to *either* reject or accept, and recommend in this case that judgment be deferred. Of course this situation does not arise if $r_c = b^* = a_c$.

This rejection region and both error reports are identical to those of the Bayesian method for loss-ratio $\ell = \rho r_c$; for that reason, call this test "$T_{\rho\ell}$". The Bayesian test, like $T_{\rho\ell}$, rejected H_0 for $B \le r_c = \ell/\rho$, but (perhaps) differed from $T_{\rho\ell}$ by accepting whenever $B > r_c$ while $T_{\rho\ell}$ can only accept for $B > a_c \ge r_c$; these can be made identical by setting $r_c = b^*$ and $\ell = b^* \rho$, whereupon r_c and a_c coincide with $b^* = \ell/\rho$. Example 1 is of this form, with $r_c = b^* = a_c = 1$.

4. Example 2: Sequential Tests

In a *sequential test* of a simple hypothesis on the basis of *i.i.d.* observations $X_i \sim f_\theta(\cdot)$ the number N of observations is itself random, under the control of the investigator, determined by a "stopping rule" of the form

$$\tau_n(x_1, \ldots, x_n) = \mathsf{P}[N = n | X_1 = x_1, \ldots, X_n = x_n],$$

whose distribution (conditional on the observations) does not depend on whether H_0 or H_1 is true. For each possible value n of N there is a critical or Rejection Region $\mathcal{R}_n$; the hypothesis H_0 is rejected if $(X_1, \ldots, X_N) \in \mathcal{R}_N$. Computing exactly the pre-experimental significance level

$$\begin{aligned} \alpha &= \mathsf{P}_0[(X_1, \ldots, X_N) \in \mathcal{R}_N] \\ &= \sum_{n=0}^{\infty} \mathsf{P}_0[\{(X_1, \ldots, X_n) \in \mathcal{R}_n\} \cap \{N = n\}] \end{aligned}$$

is prohibitively difficult, depending in the detail on the probability distribution for the stopping rule. The Likelihood Ratio and the Conditional Frequentist procedure for S_ρ remain simple: $B_n = \frac{f_0(X_1)\cdots f_0(X_n)}{f_1(X_1)\cdots f_1(X_n)}$, and $\alpha(s) = \rho B_N/(1 + \rho B_N)$; neither depends on the stopping rule at all.

4.1 The SPRT

In Abraham Wald's *Sequential Probability Ratio Test* (or SPRT), for example, one chooses numbers $R < 1 < A$ and continues taking samples until $B_n < R$, whereupon one stops and rejects H_0 in favor of H_1; or until $B_n > A$, whereupon one stops and accepts H_0. An elementary martingale argument shows that $N < \infty$ almost surely, and that *approximately*

$$\alpha = \mathsf{P}_0[B_N \le R] \approx \frac{R(A-1)}{A-R} \qquad \beta = \mathsf{P}_1[B_N \ge A] \approx \frac{1-R}{A-R}$$

Unfortunately the accuracy of these approximations depends critically on the probability distribution for the "overshoot," the amount $R - B_N$ or $B_N - A$ by which B_N jumps past Wald's boundary; see Siegmund (1985) for details. Our proposed test with $\rho = (A-1)/\big(A(1-R)\big)$ would give exactly the same error probabilities, in the absence of overshoot, and moreover corrects automatically for overshoot (by giving appropriately smaller error probabilities), without need for even knowing the stopping rule! In the symmetric case $R = 1/A$, for the SPRT, we have $\rho = 1$ and

$$\alpha(s) = \frac{B_N}{1 + B_N} \qquad \beta(s) = \frac{1}{1 + B_N},$$

while the pre-experimental error probabilities are $\alpha \approx \frac{R}{1+R}$ for $B_N \le R$ and $\beta \approx \frac{1}{1+A}$ upon accepting with $B_N \ge A$.

4.2 An Informative Stopping-Rule

Stopping rules are not always so tractable and well-motivated as Wald's. By the law of the iterated logarithm an investigator who continues sampling until reaching "significant" evidence against H_0 (say, at level $\alpha = .05$) will be able to do so, even if H_0 is true; for testing $H_0 : \mu = -1$ versus $H_1 : \mu = +1$ with $X_i \sim N(\mu, 1)$ (as in Example 1), for example, the random sequence $\alpha_n \equiv \Phi(-\sqrt{n}(1 + \bar{X}_n))$ is certain to fall below any preselected α, even for $\mu = -1$. While this will lead to fallacious error reports and inference if α_n or α are used for error reports, the report $\alpha(s) = \frac{\rho B_N}{1+\rho B_N} = (1 + e^{2N\bar{X}_N})^{-1}$ of test $T_{\rho\ell}$ will continue to be both valid and meaningful; the large value of n needed to reach $\alpha_n < \alpha$ will lead to an error report of $\alpha(s) \approx \frac{e^{2n}}{e^{2n}+e^{2Z_\alpha \sqrt{n}}}$, close to one if H_0 is rejected.

4.3 An Ambiguous Example

Suppose we are told, "Investigators observing *i.i.d.* $X_i \sim N(\mu, 1)$ to test $H_0 : \mu = 0$ against $H_1 : \mu = 1$ report stopping after $n = 20$ observations, with $\bar{x}_{20} = -0.7$." How are we to interpret this evidence? We are not even told whether this was conceived as a sequential or fixed-sample experiment; and, if sequential, what was the stopping rule. But for the Conditional Frequentist test $T_{\rho\ell}$, it doesn't matter; for example, we can select the symmetric $\rho = \ell = 1$ and report $B = f_0(x)/f_1(x) = e^{-20(\bar{x}_{20}-1/2)} \approx 0.018$, leading us to reject (since $B \le \ell/\rho = 1$) with $\alpha(s) = \frac{B}{1+B} = 0.018$. This will be Brown-optimal (see below) and hence better than *whatever* method the investigators used, no matter what their stopping rule.

5. Brown Optimality

Brown (1978) introduced an ordering and an optimality criterion for hypothesis tests: A test T_1 is to be preferred to T_2 (written $T_1 \succ T_2$) if for each increasing convex function $h(\cdot) : [0,1] \to R$, the error probabilities $\alpha_i \equiv \mathsf{P}_0[\text{ Rejection by } T_i]$ and $\beta_i \equiv \mathsf{P}_1[\text{ No rejection by } T_i]$ satisfy

$$
\begin{aligned}
U_0(T_1) \equiv \mathsf{E}_0\big[h\big(1 - \max(\alpha_1, \beta_1)\big)\big] \ &\ge U_0(T_2) \equiv \mathsf{E}_0\big[h\big(1 - \max(\alpha_2, \beta_2)\big)\big] \\
U_1(T_1) \equiv \mathsf{E}_1\big[h\big(1 - \max(\alpha_1, \beta_1)\big)\big] \ &\ge U_1(T_2) \equiv \mathsf{E}_1\big[h\big(1 - \max(\alpha_2, \beta_2)\big)\big].
\end{aligned}
$$

This criterion is designed to prefer a test with lower error probabilities (the monotonicity of h assures this); and, for tests with similar overall error probabilities, to prefer one that better distinguishes marginal from extreme evidence (the convexity of h assures this).

Under conditions of Likelihood Ratio Symmetry (where $B \equiv f_0(X)/f_1(X)$ has the same probability distribution under H_0 as does $1/B$ under H_1—

i.e., where $F_0(b) = 1 - F_1(1/b)$ for all $b > 0$), Brown (1978) proved that the symmetric ($\rho = \ell = 1$) version T_{11} of the Conditional Frequentist test based on S_ρ is optimal, i.e., preferred $T_{11} \succ T^*$ to every other test T^*. Even in the absence of Likelihood Ratio Symmetry, it is easy to show that the test $T_{\rho\ell}$ based on S_ρ is at least admissible, in the sense that $U_0(T_{\rho\ell}) \geq U_0(T^*)$ or $U_1(T_{\rho\ell}) \geq U_1(T^*)$ (or both) for every T^*.

6. Conclusions

The Conditional Frequentist test $T_{\rho\ell}$, the Neyman-Pearson test that rejects H_0 if $B \equiv \frac{f_0(x)}{f_1(x)} \leq \ell/\rho$ and reports conditional error probabilities (given the value of the Proportional Tail statistic $S_\rho \equiv \min\big(F_1(B), \rho(1 - F_0(B))\big)$) of $\alpha(s) = \frac{\rho B}{1+\rho B}$ (upon rejecting) or $\beta(s) = \frac{1}{1+\rho B}$ (upon accepting), is:

- A valid Bayesian test,
- A valid Likelihoodist test,
- A valid Conditional Frequentist test,
- Always admissible,
- Brown-Optimal, at least in symmetric cases,
- Unaffected by stopping rules,
- Flexible, and easy to compute and implement,
- Consistent with the Likelihood Principle.

Within all three paradigms it is superior to the commonly-used Neyman-Pearson test. I suggest that it should replace Neyman-Pearson tests for all tests of simple hypotheses against simple alternatives.

7. References

BERGER, J.O. (1985): *Statistical Decision Theory and Bayesian Analysis (2nd edn.)*. Springer-Verlag, New York.

BERGER, J.O., BROWN, L.D., and WOLPERT, R.L. (1994): A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann. Statist.* **22**, 1787–1807.

BERGER, J.O. and WOLPERT, R.L. (1988): *The Likelihood Principle: Review, Generalizations, and Statistical Implications (2nd edn.) (with Discussion)*. Institute of Mathematical Statistics Press, Hayward, CA.

BROWN, L.D. (1978): A contribution to Kiefer's theory of conditional confidence procedures. *Ann. Statist.* **6**, 59–71.

BROWNIE, C. and KIEFER, J. (1977): The ideas of conditional confidence in the simplest setting. *Commun. Statist. – Theory Meth.* **A6(8)**, 691–751.

KIEFER, J. (1975): Conditional confidence approach in multi-decision problems. In: P.R. Krishnaiah (ed.): *Multivariate Analysis IV*. Academic Press, New York.

KIEFER, J. (1977): Conditional confidence statements and confidence estimators (with discussion). *J. Amer. Statist. Assoc.* **72**, 789–827.

LEHMANN, E.L. (1986): *Testing statistical hypotheses (2nd edn.)*. John Wiley & Sons, New York; reprinted 1994 by Chapman & Hall, New York.

SIEGMUND, D. (1985): *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.