TESTS OF FIT IN THE PRESENCE OF NUISANCE LOCATION AND SCALE PARAMETERS[1]

By Lionel Weiss

Cornell University

1. Summary. Certain functions of the sample spacings are shown to converge stochastically as the sample size increases. This leads to certain convenient tests of fit which are consistent against wide classes of alternatives.

2. The stochastic convergence of certain functions of sample spacings. Suppose $X_1, X_2, \cdots, X_n$ are n independent and identically distributed chance variables, each with density function $f(x)$. Let $Y_1 \leqq Y_2 \leqq \cdots \leqq Y_n$ denote the ordered values of $X_1, X_2, \cdots, X_n$, and let U_i denote $Y_{i+1} - Y_i$ $(i = 1, \cdots, n-1)$. Let $g(v)$ be a bounded nonnegative function of v defined for $0 \leqq v \leqq 1$, and r be a number greater than or equal to unity. Define the chance variable $U(r)$ as $n^{r-1}\sum_{i=1}^{n-1} g(i/n)U_i^r$. Then we have

THEOREM 2.1. *If* $f(x) = 1$ *for* $0 \leqq x \leqq 1$, *and* $f(x) = 0$ *elsewhere, then*

$$|U(r) - \Gamma(r+1)n^{r-1}[n!/\Gamma(n+r+1)]\sum_{i=1}^{n} g(i/n)|$$

converges stochastically to zero as n increases.

PROOF. It is shown in [1] that for any positive number s,

$$E\{U_i^s\} = [n!\Gamma(s+1)/\Gamma(n+s+1)]$$

for any i, and $E\{U_i^s U_j^s\} = [n!\Gamma^2(s+1)/\Gamma(n+2s+1)]$ for any $i \neq j$. From this, we find immediately that

$$E\{U(r)\} = \Gamma(r+1)n^{r-1}[n!/\Gamma(n+r+1)]\sum_{i=1}^{n-1} g(i/n),$$

and, remembering that $g(v)$ is bounded, we find that the variance of $U(r)$ approaches zero as n increases. This completes the proof.

COROLLARY. *If* $f(x) = 1$ *for* $0 \leqq x \leqq 1$, *and* $f(x) = 0$ *elsewhere, and* $\int_0^1 g(v)\,dv$ *exists (in the Riemann sense), then* $U(r)$ *converges stochastically to*

$$\Gamma(r+1)\int_0^1 g(v)\,dv$$

as n increases

PROOF. If $\int_0^1 g(v)\,dv$ exists, $n^{r-1}[n!/\Gamma(n+r+1)]\sum_{i=1}^{n-1} g(i/n)$ approaches it as n increases.

Let us denote $\int_{-\infty}^{x} f(y)\,dy$ by $F(x)$. Suppose that on the interval $[A, B]$, $f(x)$ is continuous and $f(x) \geqq D > 0$ for each x in $[A, B]$. Suppose $h(v)$ is a nonnegative bounded function of $v (0 \leqq v \leqq 1)$, and $\int_0^1 h(v)\,dv$ exists (in the Riemann sense). Define the chance variable R as $\sum_{nF(A)<j<nF(B)} h(j/n)(Y_{j+1} - Y_j)$, and S as $n\sum_{nF(A)<j<nF(B)} h(j/n)[Y_{j+1} - Y_j]^2$. Then we have

Received August 21, 1956.

[1] Research under contract with the Office of Naval Research.

THEOREM 2.2. *As n increases, R converges stochastically to $\int_{F(A)}^{F(B)} \{h(v) / f[F^{-1}(v)]\}\, dv$, and S converges stochastically to $2\int_{F(A)}^{F(B)}[h(v) / \{f[F^{-1}(v)]\}^2]\, dv$.*

PROOF. If Y_{j+1} and Y_j are both in the interval $[A, B]$, we may write

$$F(Y_{j+1}) - F(Y_j) = f(\theta_j)[Y_{j+1} - Y_j],$$

where $Y_j \leqq \theta_j \leqq Y_{j+1}$. Let $F_n^*(x)$ denote the empirical distribution function based on $X_1, X_2, \cdots, X_n$. That is, for each x, $F_n^*(x)$ equals the proportion of the values $(X_1, X_2, \cdots, X_n)$ which are no greater than x. Define λ_i as

$$|F(Y_i) - (i/n)|,$$

which is the same as $|F(Y_i) - F^*(Y_i)|$. It is well known that if δ is any positive number, then $\max_i n^{1/2-\delta}\lambda_i$ converges stochastically to zero as n increases. Define δ_i as $|Y_i - F^{-1}(i/n)|$. If Y_i and $F^{-1}(i/n)$ are both in the interval $[A, B]$, then, since $f(x) \geqq D$ on that interval, we have $\delta_i \leqq (\lambda_i/D)$. Then, if Y_j, Y_{j+1}, $F^{-1}(j/n)$, and $F^{-1}[(j+1)/n]$ are all in $[A, B]$, we have $|\theta_j - F^{-1}(j/n)| \leqq 1/nD + \delta_j + \delta_{j+1}$, and we can write

$$F(Y_{j+1}) - F(Y_j) = f\left[F^{-1}\left(\frac{j}{n}\right)\right](Y_{j+1} - Y_j) + \gamma_j(Y_{j+1} - Y_j), \tag{2.2.1}$$

where $\gamma_j = f(\theta_j) - f[F^{-1}(j/n)]$. But because of the uniform continuity of $f(x)$ in $[A, B]$, the inequality for $|\theta_j - F^{-1}(j/n)|$, and the Glivenko-Cantelli theorem, it is easily seen that $\max_{nF(A)<j<nF(B)}|\gamma_j|$ converges stochastically to zero as n increases. We denote $F(Y_{j+1}) - F(Y_j)$ by U_j, and note that $\{U_j\}$ has the same distribution as in Theorem 2.1. From (2.2.1) we have

$$\begin{aligned}\sum_{nF(A)<j<nF(B)} h\left(\frac{j}{n}\right)\frac{U_j}{f\left[F^{-1}\left(\frac{j}{n}\right)\right]} &= R \\ &+ \sum_{nF(A)<j<nF(B)} h\left(\frac{j}{n}\right)\gamma_j\left[\frac{Y_{j+1} - Y_j}{f\left[F^{-1}\left(\frac{j}{n}\right)\right]}\right].\end{aligned} \tag{2.2.2}$$

The expression on the left of (2.2.2) converges stochastically to

$$\int_{F(A)}^{F(B)} \{h(v) / f[F^{-1}(v)]\}\, dv,$$

by the corollary to Theorem 2.1. Let us denote the second term on the right of (2.2.2) by R'. As n increases, the probability that $|R'|$ will be no greater than $(\max_j|\gamma_j|/D)R$ approaches one. This means that $|R'|/R$ converges stochastically to zero as n increases. But $R + R'$ converges stochastically to

$$\int_{F(A)}^{F(B)} \{h(v) / f[F^{-1}(v)]\}\, dv$$

as n increases. This proves Theorem 2.2 as far as R is concerned. The proof for S is entirely similar.

3. Application to tests of fit. We need the following lemma.

LEMMA 3.1. *If $F(x)$ and $G(x)$ are two distribution functions, and u, $v(0 \leqq u <$*

$v \leqq 1$) are two given numbers, suppose $F^{-1}(u)$, $F^{-1}(v)$, $G^{-1}(u)$, $G^{-1}(v)$ are all uniquely determined. Also suppose that $F(x)$ has a derivative $f(x)$ between $F^{-1}(u)$ and $F^{-1}(v)$, and $G(x)$ has a derivative $g(x)$ between $G^{-1}(u)$ and $G^{-1}(v)$. Then a necessary and sufficient condition that $f[F^{-1}(r)] = kg[G^{-1}(r)]$ for almost all r in $[u, v]$ (where k is a positive constant) is that there are two constants C, $D(C > 0)$, such that $F(Cx + D) = G(x)$ for all x in the interval $[G^{-1}(u), G^{-1}(v)]$.

PROOF.

(a) *Sufficiency.* Suppose there are constants C, D such that $F(Cx + D) = G(x)$ for all x in $[G^{-1}(u), G^{-1}(v)]$. For any such x, $Cf(Cx + D) = g(x)$. There is an r in $[u, v]$ such that $x = G^{-1}(r)$. Then $Cx + D = F^{-1}(r)$, so that $Cf[F^{-1}(r)] = g[G^{-1}(r)]$. For any r in $[u, v]$ we can find a value x so $r = G(x)$, and this completes the proof of sufficiency.

(b) *Necessity.* Suppose $f[F^{-1}(r)] = kg[G^{-1}(r)]$ for almost all r in $[u, v]$. Since $F[F^{-1}(r)] = r$, we find by differentiation that $(d/dr)F^{-1}(r) = 1/\{f[F^{-1}(r)]\}$ wherever $f[F^{-1}(r)]$ is positive. Therefore, at each r in $[u, v]$ at which $f[F^{-1}(r)] > 0$, $(d/dr)G^{-1}(r) = k(d/dr)F^{-1}(r)$. This implies that for all r in $[u, v]$, $G^{-1}(r) = kF^{-1}(r) + b$, b a constant, or $F^{-1}(r) = KG^{-1}(r) + B$, B, K constants with K positive. There is a value x in $[G^{-1}(u), G^{-1}(v)]$ with $r = G(x)$. Then $F^{-1}[G(x)] = Kx + B$, or $G(x) = F(Kx + B)$ for all x in $[G^{-1}(u), G^{-1}(v)]$. This completes the proof of Lemma 3.1.

As an application, suppose we are to test the hypothesis to be described. $X_1, X_2, \cdots, X_n$ are known to be independent and identically distributed chance variables. Two known constants u, $v(0 \leqq u < v \leqq 1)$ are given. The hypothesis is that the common distribution function $F(x)$ of X_i is, for each x in the interval $[F^{-1}(u), F^{-1}(v)]$, equal to $G(Cx + D)$, where C, D are some unspecified constants $(C > 0)$, and the distribution function $G(x)$ is specified. We assume that for each x in the interval $[G^{-1}(u), G^{-1}(v)]$, $G(x)$ has a derivative $g(x)$, with $g(x) \geqq A > 0$, and $g(x)$ has at most a finite number of discontinuities in $[G^{-1}(u), G^{-1}(v)]$.

We propose to test the hypothesis just described by means of the following statistic:

$$Z_n = \frac{n\sum_{nu<j<nv} g^2\left[G^{-1}\left(\frac{j}{n}\right)\right](Y_{j+1} - Y_j)^2}{\left\{\sum_{nu<j<nv} g\left[G^{-1}\left(\frac{j}{n}\right)\right](Y_{j+1} - Y_j)\right\}^2}.$$

From Theorem 2.2, we know that if the true common distribution $F(x)$ has a derivative $f(x)$ on the interval $[F^{-1}(u), F^{-1}(v)]$ with at most a finite number of discontinuities in that interval, and if $f(x) \geqq A' > 0$ on the interval, then Z_n converges stochastically to

$$\frac{2\int_u^v \left\{\frac{g[G^{-1}(x)]}{f[F^{-1}(x)]}\right\}^2 dx}{\left[\int_u^v \left\{\frac{g[G^{-1}(x)]}{f[F^{-1}(x)]}\right\} dx\right]^2} \tag{3.1}$$

as n increases. From Lemma 3.1, we know that $f[F^{-1}(x)] = kg[G^{-1}(x)]$ almost everywhere on $[u, v]$ if and only if the hypothesis is true. Therefore if the hypothesis is true, Z_n converges stochastically to $2/(v - u)$. If the hypothesis is not true, but $F(x)$ satisfies the conditions that guarantee that Z_n converges to (3.1), we see that Z_n converges stochastically to

$$\frac{2\int_u^v h^2(x)\,dx}{\left[\int_u^v h(x)\,dx\right]^2} \tag{3.2}$$

as n increases, where $h(x)$ is a certain function not equal to a constant almost everywhere on $[u, v]$. But then (3.2) has a value greater than $2/(v - u)$. Therefore the test of the hypothesis which rejects when Z_n is "too large" is consistent against any alternative $F(x)$ with a density function bounded away from zero and with a finite number of discontinuities on the interval $[F^{-1}(u), F^{-1}(v)]$. If $F(x)$ assigns zero probability to a subinterval of $[F^{-1}(u), F^{-1}(v)]$ of positive length, while $F^{-1}(v) - F^{-1}(u)$ is finite, it is easily verified that Z_n approaches infinity with probability one as n increases. Thus the test based on Z_n is consistent against a very wide class of alternatives. Another advantage of the test is that the distribution of Z_n does not depend upon the unknown parameters when the hypothesis is true. Furthermore, the computation of Z_n is fairly easy if a table of values of $G(x)$ and $g(x)$ is available.

4. A conjecture about large-sample distributions. A reasonable conjecture seems to be that the numerator and the square root of the denominator of the chance variable Z_n have a limiting distribution which is bivariate normal. The remainder of this section will be a heuristic justification of this conjecture. We denote $n\sum_{nu<j<nv} g^2[G^{-1}(j/n)](Y_{j+1} - Y_j)^2$ by Q, and $\sum_{nu<j<nv} g[G^{-1}(j/n)](Y_{j+1} - Y_j)$ by W. Z_n was defined as Q/W^2. From the proof of Theorem 2.2, W and Q have about the same joint distribution as

$$\sum_{nu<j<nv} g\left[G^{-1}\left(\frac{j}{n}\right)\right]\left\{\frac{U_j}{f\left[F^{-1}\left(\frac{j}{n}\right)\right]}\right\}$$

and

$$n\sum_{nu<j<nv} g^2\left[G^{-1}\left(\frac{j}{n}\right)\right]\left\{\frac{U_j}{f\left[F^{-1}\left(\frac{j}{n}\right)\right]}\right\}^2,$$

where $\{U_i\}$ have the same joint distribution as in Theorem 2.1. Thus W is (approximately) a linear combination of $U_{nu}, \cdots, U_{nv}$, and Q is (approximately) a linear combination of $U_{nu}^2, \cdots, U_{nv}^2$. Next we show that the mixed moments of $\{nU_i\}$ approach the corresponding moments of $\{V_i\}$, where $V_1, V_2, \cdots$

are independent chance variables, each with density e^{-v} for $v > 0$. In fact,

$$E\{(nU_{i_1})^{a_1} \cdots (nU_{i_k})^{a_k}\} = \frac{n^{a_1+\cdots+a_k}\,\Gamma(a_1+1)\cdots\Gamma(a_k+1)\Gamma(n+1)}{\Gamma(n+a_1+\cdots+a_k+1)},$$

which approaches $\Gamma(a_1+1)\cdots\Gamma(a_k+1)$ as n increases, while

$$E\{V_{i_1}^{a_1}\cdots V_{i_k}^{a_k}\} = \Gamma(a_1+1)\cdots\Gamma(a_k+1).$$

Thus the chance variables W and Q are essentially linear combinations of chance variables which in important respects act like independent chance variables in the limit. The bivariate central limit theorem suggests the limiting normality of the joint distribution. If this conjecture is correct, then for large samples the approximate critical value for Z_n, as well as the power of the test against various alternatives, can be very easily computed.

REFERENCES

[1] B. F. KIMBALL, "Some basic theorems for developing tests of fit for the case of non-parametric probability distribution functions," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 540–548.