Sankhyā : The Indian Journal of Statistics
1976, Volume 38, Series A, Pt. 2, pp. 124–130.

ON UNBIASEDNESS OF MANN-WALD-GUMBEL χ^2-TEST

By BIMAL KUMAR SINHA
Indian Statistical Institute

SUMMARY. In this note it is proved that the Mann-Wald Gumbel χ^2 test based on equal hypothetical probabilities (Kendall and Stuart, 1961) which was shown by Mann and Wald (*Ann. Math. Stat.*, 1942) to be locally unbiased is in fact uniformly so against all alternatives.

1. Introduction

Let $x_1, x_2, \ldots, x_n$ be independent observations on a random variable with distribution function (d.f.) $F(x)$ which is unknown and consider the goodness-of-fit problem of testing the hypothesis

$$Ho : F(x) = F_0(x) \quad \ldots \quad (1.1)$$

where $F_0(x)$ is a completely specified d.f. (continuous or discrete). One method of testing H_0 which depends on a very simple device consists in dividing the range of the variate into k ($\geqslant 2$) mutually exclusive classes and the test is based on the statistic

$$\chi^2 = \sum_{i=1}^{k} (n_i - np_{0i})^2/np_{0i} \quad \ldots \quad (1.2)$$

where p_{0i} is the probability (under H_0) of an observation falling in the i-th class and n_i in the actual number of observations falling in the i-th class, $i = 1, 2, \ldots, k$; $p_{0i} > 0, i = 1, 2, \ldots, k$; $\sum_1^k p_{0i} = 1$; $n_i \geqslant 0, i = 1, \ldots, k$; $\sum_1^k n_i = n$.

Regarding the method of how the classes for a fixed $k \geqslant 2$ would be constructed, Mann and Wald (1942) and Gumbel (1943) suggested the following rule : 'Given k, choose the classes so that the hypothetical probabilities p_{0i} are all equal to $\frac{1}{k}$'. Under this rule, the form of the χ^2-statistic becomes

$$\chi_0^2 = \frac{k}{n} \sum_1^k n_i^2 - n \quad \ldots \quad (1.3)$$

and the goodness-of-fit test of H_0 based on this statistic is given by

Reject H_0 if $\chi_0^2 > c_\alpha$ where c_α is such that

$$\Pr[\chi_0^2 > c_\alpha \mid H_0] = \Pr\left[\sum_1^k n_i^2 > \frac{n}{k}(n + c_\alpha) = c\,(\text{say})\right] = \alpha,\ 0 < \alpha < 1 \quad \ldots \quad (1.4)$$

We shall refer to the test outlined in (1.4) as Mann-Wald-Gumbel (M-W-G) χ^2-test. The test is consistent (Kendall and Stuart, 1961; Neyman, 1949) and has been proved to be locally unbiased by Mann and Wald (1942). The object of this note is to prove somewhat stronger result that the test is uniformly unbiased against all alternatives. Throughout this paper we assume that n and k ($\geqslant 2$) are arbitrary but fixed.

2. Explicit formulation of the problem

Let ξ_i be the random variable denoting the number of observations (out of n) falling in the i-th class ($i = 1, 2, \ldots, k$). It is then clear that $\xi_1, \ldots, \xi_k$ have a joint multinomial distribution and under H_0, the joint probability function is given by

$$\Pr[\xi_1 = n_1, \ldots, \xi_k = n_k \mid n, H_0] = \frac{n!}{n_1! \ldots n_k!} \left(\frac{1}{k}\right)^n \quad \text{for } 0 \leqslant n_i \leqslant n$$

$$1 \leqslant i \leqslant k;$$

$$\sum_1^k n_i = n$$

$$= 0, \quad \text{otherwise.} \qquad \ldots \quad (2.1)$$

Consider any arbitrary alternative $H_1 : F(x) = F_1(x)$ where $F_1(x)$ is an arbitrary c.d.f. and denote by p_i the probability (under H_1) of an observation falling in the i-th class, $i = 1, 2, \ldots, k$, $p_i \geqslant 0$, $i = 1, \ldots, k$ $\sum\limits_1^k p_i = 1$. It then follows that under this alternative H_1, the joint probability function of $\xi_1, \ldots, \xi_k$ is given by

$$\Pr[\xi_1 = n_1, \ldots, \xi_k = n_k \mid n, H_1]$$

$$= \frac{n!}{n_1! \ldots n_k!} p_1^{n_1}, \ldots, p_k^{n_k} \text{ for } \quad 0 \leqslant n_i \leqslant n, \ 1 \leqslant i \leqslant k; \sum_1^k n_i = n$$

$$= 0, \quad \text{otherwise.} \qquad \ldots \quad (2.2)$$

Let $n_{(k)}$, $p_{(k)}$ and $\delta_{(k)}$ stand for the k-vectors $(n_1, \ldots, n_k)$, $(p_1, \ldots, p_k)$ and $\left(\frac{1}{k}, \ldots, \frac{1}{k}\right)$ respectively. Here $(n_1, \ldots, n_k)$ is a k-vector of non-negative integral co-ordinates n_i's with $\sum\limits_1^k n_i = n$. Let us also denote the R.H.S. expression in (2.2) by $\pi_k(n_{(k)}/n, \ p_{(k)})$ so that $\pi_k \ (n_{(k)}/n, \ \delta_{(k)})$ stands for the R.H.S. expression in (2.1).

Our aim is to prove the following result directing to uniform unbiasedness of Mann-Wald-Gumbel χ^2-test :

Theorem : *Whatever* $c, p_1, \ldots, p_k \geqslant 0, \Sigma p_i = 1$

$$\sum_{n_{(k)} \in S_k} \pi_k(n_{(k)} | n; p_{(k)}) \geqslant \sum_{n_{(k)} \in S_k} \pi_k(n_{(k)} | n; \delta_{(k)}) \quad \ldots \quad (2.3)$$

where $S_k = \{n_{(k)} : n_i \geqslant 0,\ 1 \leqslant i \leqslant k;\ \Sigma n_i = n;\ \Sigma n_i^2 > c\}$.

Here strict inequality holds unless

$$\text{(i)} \quad p_{(k)} = \delta_{(k)} \quad \text{or} \quad \text{(ii)} \quad c < c_{n,k}, \geqslant n^2 \quad \ldots \quad (2.3.1)$$

with $c_{n,k} = min \left\{ \sum_1^k n_i^2 \right\}$ *for variations of integral* n_i*'s subject to* $\sum_1^k n_i = n$.

3. The case of $k = 2$

We first show that the theorem is true for $k = 2$ in which case it reduces to proving for arbitrary $p_1, p_2 \geqslant 0$, $p_1 + p_2 = 1$ and c

$$\sum_{n_{(2)} \in S_2} \pi_2(n_{(2)} | n; p_{(2)}) \geqslant \sum_{n_{(2)} \in S_2} \pi_2(n_{(2)} | n; \delta_{(2)}) \quad \ldots \quad (3.1)$$

where $S_2 = \{n_{(2)} : n_1, n_2 \geqslant 0,\ n_1 + n_2 = n,\ n_1^2 + n_2^2 > c\}$.
Here strict inequality will hold unless

$$\text{(i)} \quad p_{(2)} = \delta_{(2)} \quad \text{or} \quad \text{(ii)} \quad c < c_{n,2}, \geqslant n^2. \quad \ldots \quad (3.1.1)$$

We note that when (ii) of (3.1.1) holds, the L.H.S. of (3.1) is either zero or unity, independently of $p_{(2)}$ and hence (3.1) reduces to an equality. Assume, therefore, $c_{n,2} \leqslant c < n^2$. Observe also that if p_1 or p_2 is zero, the L.H.S. in (3.1) is unity but its R.H.S. is less than unity and hence (3.1) is trivially true. Assume, therefore, $p_1, p_2 > 0$. Also note that there is nothing to prove if (3.1.1) holds. So assume neither (i) nor (ii) of (3.1.1) holds i.e., $p_{(2)} \neq \delta_{(2)}$ and $c_{n,2} \leqslant c < n^2$. Writing $n_1 = x$ and $n_2 = n - x$, the L.H.S. of (3.1) comes out as

$$\sum_{x : x^2 + (n-x)^2 > c} \binom{n}{x} p_1^x (1-p_1)^{n-x}$$

$$= \sum_{x : \left(x - \frac{n}{2}\right)^2 > \frac{1}{2}\left(c - \frac{n^2}{2}\right)} \binom{n}{x} p_1^x (1-p_1)^{n-x}$$

$$= \sum_{\substack{x > \frac{n}{2} + \left\{\frac{1}{2}\left(c - \frac{n^2}{2}\right)\right\}^{1/2} \\ < \frac{n}{2} - \left\{\frac{1}{2}(c - n^2/2)\right\}^{1/2}}} \binom{n}{x} p_1^x (1-p_1)^{n-x} \quad \ldots \quad (3.2)$$

Since $c_{n,2} = \frac{n^2}{2}$ or $\frac{n^2+1}{2}$ according as n is even or odd and since we restrict to $c \geqslant c_{n,2}$, we may write (3.2) as

$$\sum_{x \geqslant n-r} \binom{n}{x} p_1^x (1-p_1)^{n-x} + \sum_{x \leqslant r} \binom{n}{x} p_1^x (1-p_1)^{n-x} \text{ where } r \leqslant \left[\frac{n}{2}\right]-1 \quad \ldots \ (3.2.1)$$

whatever n, odd or even. This shows that the expression in (3.2.1) is *not* equal to unity, independently of p_1. We now rewrite (3.2.1) in the form

$$\int_{1-p_1}^{1} \frac{z^r(1-z)^{n-r-1}dz}{B(r+1, n-r)} + \int_{p_1}^{1} \frac{z^r(1-z)^{n-r-1}dz}{B(r+1, n-r)} = g(p_1) \quad \text{(say)}. \quad \ldots \ (3.2.2)$$

It is now easy to verify that $g(p_1)$ has a unique minimum at $p_1 = \frac{1}{2}$ and hence (3.1) follows with strict inequality unless $p_{(2)} = \delta_{(2)}$.

The case of $k = 2$ is thus disposed of.

4. General case

Before taking up the general case, we record an algebraic result of subsequent interest. We begin with

Lemma 4.1 : (i) *$c_{n,k}$ is attained by $\sum_1^k n_i^2$ (subject to $\sum_1^k n_i = n$) at a unique point n^* where $n_i^* = \left[\frac{n}{k}\right]$ or $\left[\frac{n}{k}\right]+1$.*

(ii) *If $\sum_1^k n_i^2 > c_{n,k}$, then for some n_i and n_j, $|n_i - n_j| \geqslant 2$.*

The converse is also true. The proof is easy and hence omitted.

The setting for our algebraic result would be as follows :

We have $\quad S_k = \{n_{(k)} : n_i \geqslant 0,\ 1 \leqslant i \leqslant k; \sum n_i = n; \sum n_i^2 > c\}$.
For given c such that $c_{n,k} \leqslant c < n^2$, let c_0 be the largest value $\leqslant c$, actually attained by $\sum n_i^2$. Then the event $\sum n_i^2 > c$ is the same as the event $\sum n_i^2 > c_0$. Let c_0 be attained for $n_i = n_i^0$ $(1 \leqslant i \leqslant k)$. We write $n^0 = (n_1^0, n_2^0, \ldots, n_k^0)$.

We will make use of the above concept to prove the following.

Lemma 4.2 : *Given c such that $c_{n,k} \leqslant c < n^2$ $\exists$ x $(0 \leqslant x \leqslant n)$ such that*

$$c_{n-x,k-1} + x^2 \leqslant c < (n-x)^2 + x^2.$$

Proof: From the above consideration, for any given c, we determine the particular c_0. Then $c_{n,k} \leqslant c_0 < n^2$. Two cases are to be considered.

Case I : $n_i^0 \geqslant 1$ *for all* i, $1 \leqslant i \leqslant k$.

(a) When $c_0 = c_{n,k}$, $\exists\, x$ $\left(x = \left[\frac{n}{k}\right] \text{ or } \left[\frac{n}{k}\right]+1\right)$ such that $c_{n-x,k-1}+x^2 = c_{n,k}$. Since $n_i^* = n_i^0 \geqslant 1 \;\forall\; i$, we also have $(n-x)^2+x^2 > c_{n,k} \;\forall\; k \geqslant 3$ by the uniqueness part of Lemma 4.1. Hence we can find an x such that $(n-x)^2+x^2 > c \geqslant c_0 = c_{n,k} = x^2+c_{n-x,k-1}$ (the former inequality follows from the definition of c_0).

(b) When $c_0 > c_{n,k}$, by Lemma 4.1, we can find two integers n_i^0 and n_j^0 ($\geqslant 1$, $< n$) such that $|n_i^0-n_j^0| \geqslant 2$. Set $n_i^0 < n_j^0$. Next define $n_j^1 = n_j^0+1$, $n_i^1 = n_i^0-1$, $n_h^1 = n_h^0$, $1 \leqslant h \leqslant k$ ($h \neq i$, $h \neq j$) so that $c_1 = \sum_{l=1}^{k} (n_l^1)^2 = c_0+2(n_j^0-n_i^0+1)$ and also define $n_j^{-1} = n_j^0-1$, $n_i^{-1} = n_i^0+1$, $n_h^{-1} = n_h^0$, $1 \leqslant h \leqslant k$ ($h \neq i$, $h \neq j$) so that $c_{-1} = \sum_{l=1}^{k} (n_l^{-1})^2 = c_0+2(n_i^0-n_j^0+1)$.
It is easy to observe that $c_{-1} < c_0 < c_1$ and further that, for $k \geqslant 3$, n^0, n^1 and n^{-1} have a common coordinate, say x.

We write then $c_0 = x^2+\sum_l (n_l^0)^2$,

$$c_1 = x^2+\sum_l (n_l^1)^2 \text{ and } c_{-1} = x^2+\sum_l (n_l^{-1})^2.$$

From $c_{-1} < c_0 < c_1$ and definition of c_0, we get $c_{-1} < c < c_1$. Again, obviously, $c_{-1} \geqslant c_{n-x,k-1}+x^2$ and $c_1 \leqslant (n-x)^2+x^2$. Hence, $c_{n-x,k-1}+x^2 < c < (n-x)^2+x^2$ for a particular choice of x.

Case II : $n_i^0 = 0$ *for some* i, $1 \leqslant i \leqslant k$.

Certainly this time $c_{n,k} \leqslant c_0 \leqslant c < n^2$ where, again, $c_0 = \sum_{h(\neq i)} (n_h^0)^2 \geqslant c_{n,k-1}$. Hence $c_{n,k-1} \leqslant c_0 \leqslant c < n^2$ i.e., $c_{n,k-1} \leqslant c < n^2$. This is how we achieve the result with $x = 0$. The lemma is thus proved.

Now we attempt a proof for the general case based on induction.

Assume that the result is true for $k = m-1$. We then prove that the result is true for $k = m$. Note that there is nothing to prove if (2.3.1) holds. So assume (2.3.1) does not hold. We make use of the well-known fact that if $\xi_1, \ldots, \xi_m$ have a joint multinomial distribution with the parameters n and

$p_1, \ldots, p_m$ (< 1), then the conditional joint distribution of $\xi_1, \ldots, \xi_{m-1}$, given $\xi_m = x$, is again multinomial with the new prameters $n-x$ and $\frac{p_1}{1-p_m}, \ldots, \frac{p_{m-1}}{1-p_m}$. The L.H.S. expression in the theorem for $k = m$ can then be written as

$$\sum_{x=0}^{n} \left\{ \sum_{n_{(m-1)} \in S^*_{m-1}} \pi_{m-1} \left(n_{(m-1)} \,|\, n-x;\ \frac{p_{(m-1)}}{1-p_m} \right) \right\} \pi_1(x \,|\, n;\ p_m) \quad \ldots \quad (4.1)$$

[Here we have excluded the trivial case of any of the p_i's being equal to unity.]

Here $S^*_{m-1} = \Big\{ n_{(m-1)} : n_i \geqslant 0,\ 1 \leqslant i \leqslant m-1;\ \sum_1^{m-1} n_i = n-x;$

$$\sum_1^{m-1} n_i^2 > c - x^2 \Big\}.$$

By the induction hypothesis, the bracketted expression above is greater than or equal to

$$\left\{ \sum_{n_{(m-1)} \in S^*_{m-1}} \pi_{m-1}(n_{(m-1)} \,|\, n-x;\ \delta_{(m-1)}) \right\} \quad \ldots \quad (4.2)$$

with strict inequality unless

$$\frac{p_{(m-1)}}{1-p_m} = \delta_{(m-1)} \qquad \text{or} \quad c - x^2 < c_{n-x,k-1}, \geqslant (n-x)^2.$$

Suppose now that $c_{n,k} \leqslant c < n^2$. Then, by Lemma 4.2, we can find an integral $x (0 \leqslant x \leqslant n)$ such that $c_{n-x,k-1} \leqslant c - x^2 < (n-x)^2$. Therefore, for every $c \in [c_{n,k}, n^2)$, the L.H.S. expression in the theorem for $k = m$ is *strictly greater than*

$$\sum_{x=0}^{n} \left\{ \sum_{n_{(m-1)} \in S^*_{m-1}} \pi_{m-1}(n_{(m-1)} \,|\, n-x;\ \delta_{(m-1)}) \right\} \pi_1(x \,|\, n;\ p_m) \quad \ldots \quad (4.3)$$

unless $p_{(m-1)}/1 - p_m = \delta_{(m-1)}$ in which case (4.3) is attained.

Now suppose that the absolute minimum of the L.H.S. of (2.3) for $k = m$ is actually attained at a point $p^0_{(m)}$. Then (4.3) implies $p^0_1 = p^0_2 = \ldots = p^0_{m-1}$ (or else it would provide a value smaller than the absolute minimum). A similar argument shows $p^0_2 = \ldots p^0_m$. Hence, necessarily, $p^0_1 = \ldots = p^0_m = \frac{1}{m}$ i.e., $p^0_{(m)} = \delta_{(m)}$. Therefore, the inequality in (2.3) is strict unless $p_{(m)} = \delta_{(m)}$ whenever $c_{n,k} \leqslant c < n^2$.

This completes the proof of (2.3) for $k = m$ when it is true for $k = m-1$.

The proof of the Theorem is thus completed by the induction argument.

Remark. When the paper was being revised, the author came to know that recently Sethuraman *et al* (1974) have proved, among other results, a very strong result about the multinomial implying the result of this paper.

Acknowledgement

I wish to express my sincere thanks to one of the co-editors for making some helpful comments regarding the proof of the theorem. Most thankfully I acknowledge the generous help received from Dr. Bikas Kumar Sinha during the revision of the paper.

References

Gumbel, E. J. (1943) : On the reliability of the classical χ^2 test. *Ann. Math. Stat.* **14**, 253.

Kendall, M. G. and Stuart, A. (1961) : *The Advanced Theory of Statistics*, Volume 2.

Mann, H. B. and Wald, A. (1942) : On the choice of the number of intervals in the application of the chi-square test, *Ann. Math. Stat.* **13**, 306.

Nevius, E. S., Proschan, F. and Sethuraman, J. (1974): Schur functions in statistics (abstract), presented at the Mahalanobis Memorial Symposium held at Indian Statistical Institute in December, 1974.

Neyman, J. (1949) : Contribution to the theory of the χ^2 test, *Proc. (First) Berkeley Symposium on Math. Stat. and Prob.*, 239, Univ. California Press.

Paper received : June, 1974.
Revised : July, 1975.