The Annals of Probability
1984, Vol. 12, No. 3, 794–804

NORMAL APPROXIMATIONS TO SUMS OF SCORES BASED ON OCCUPANCY NUMBERS

BY M. P. QUINE AND J. ROBINSON

University of Sydney

A central limit theorem and remainder term estimates are given for the distribution of the sum of scores based on the occupancy numbers resulting from the random allocation of N balls to n boxes. The proof involves bivariate characteristic functions, exploiting the equivalence of multinomial and conditioned Poisson variables. The results are shown to include the statistics for the empty cell test, the chi-squared test and the likelihood ratio test.

1. Introduction and results. Let $(U_1, \cdots, U_n)$ be a multinomial variable with parameters $(N; p_1, \cdots, p_n)$ such that $\sum_{k=1}^n p_k = 1$ and $\sum_{k=1}^n U_k = N$. Let $\{c(j, k): j \geq 0, k = 1, \cdots, n\}$ be sets of constants. The random variables, the parameters and the sets of constants all actually depend on N and n but for simplicity this dependence is suppressed in the notation.

Let $S = \sum_{k=1}^n c(U_k, k)$. We will obtain a central limit theorem and a bound on the rate of convergence for S as $n, N \to \infty$. It is well known that S has the conditional distribution of $T = \sum_{k=1}^n c(Y'_k, k)$ given $Y' = \sum_{k=1}^n Y'_k = N$, where $(Y'_1, \cdots, Y'_n)$ are independent Poisson variables with parameters $(\alpha_1, \cdots, \alpha_n)$, with $\alpha_k = Np_k$. We expect the joint limiting distribution of T and Y' to be bivariate normal, so instead of T we consider $Z' = T - \gamma Y'$, where γ is chosen so that Z' and Y' are asymptotically independent. To this end, let $Z'_k = c(Y'_k, k) - \gamma Y'_k$, where

$$\gamma = \operatorname{cov}(T, Y')/\operatorname{var}(Y') = \sum_{k=1}^n \operatorname{cov}(c(Y'_k, k), Y'_k)/N,$$

define $Z_k = (Z'_k - EZ'_k)/\sigma = d(Y'_k, k)$ and $Y_k = (Y'_k - \alpha_k)/\alpha^{1/2}$, where $\alpha = N/n$ and $\sigma^2 = n^{-1} \sum_{k=1}^n \operatorname{var}[c(Y'_k, k) - \gamma Y'_k]$, and let $Z = n^{-1/2} \sum_{k=1}^n Z_k$ and $Y = n^{-1/2} \sum_{k=1}^n Y_k$.

The conditional characteristic function of Z given $Y = 0$ can be obtained from the joint characteristic function of Y and Z. The central limit theorem will be proved by showing that this conditional characteristic function tends to the characteristic function of a standard normal variate. Rates of convergence are obtained from bounds on the difference of these characteristic functions. The result for the central limit theorem is now given; in the sequel C_1, C_2, c_1, c_2, C and c are positive constants which do not depend on n or N; C and c may change value at each appearance.

THEOREM 1. *If $\alpha_k \leq C_1\alpha$ for all k, if $n\alpha^2 \to \infty$, if there exists c_1 such that*

Received June 1983; revised November 1983.
AMS 1980 *subject classification.* Primary 60F05.
Key words and phrases. Central limit theorem, Berry-Esseen bound, occupancy schemes, multinomial sums.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Probability.* www.jstor.org

$\alpha < n^{c_1}$, if $n, N \to \infty$ and if for all $\varepsilon > 0$

$$n^{-1}\sum_{k=1}^{n} EZ_k^2 I(|Z_k| > n^{1/2}\varepsilon) \to 0, \tag{1.1}$$

then $\Delta \to 0$, where

$$\Delta = \sup_x |P(S - \sum_{k=1}^{n} \mu_k \leq n^{1/2}\sigma x) - \Phi(x)|$$

for Φ the standard normal distribution function and $\mu_k = Ec(Y'_k, k)$.

This result is similar to that of Theorem 4.2 and Corollary 4.1 of Morris (1975). However, our conditions neither imply nor are implied by his. The Lindeberg condition (1.1) is weaker than his condition (4.12) which implies a Liapounov condition on the fourth moments of Z_k. For example if $c(j, k) = j^{j/3}$ and $\alpha_k = 1$ for $1 \leq k \leq n$, then (1.1) holds but third moments do not exist. Also we use the condition $\alpha_k < C_1\alpha$ in place of the weaker conditions $\max_k p_k \to 0$ and either the very complicated (4.12) or the condition $\alpha_k > \varepsilon > 0$ for all k, which replaces (4.12) in his Corollary 4.1. It seems to us more natural to restrict the relative size of the α_k than to insist that none of them tend to zero. We also need the rather weak conditions that $n\alpha^2 \to \infty$ and $\alpha < n^{c_1}$; however these can be avoided at the cost of conditions on the scores given in Theorems 3 and 4 later. Theorem 1 implies the results of Holst (1972), those of Steck (1957) given in Morris (1975, page 182), for the particular case of a chi-squared statistic and those of Rényi (1962) for the empty cell statistic. Mann and Wald (1942) also state a limit result for the chi-squared statistic; other earlier partial results are cited in the above papers.

The major results of this paper are the bounds on the rates of convergence given in the following theorems. The only earlier results of this type of which we are aware are in Englund (1981) and Quine and Robinson (1982) (referred to as QR in the sequel) which concern the empty cell statistic, and Quine (1983) which considers the case of $\alpha_k = \alpha$, $k = 1, \cdots, n$, under restrictive conditions on the scores. The present results subsume all of these. The common method of proof used for the central limit theorem and for the rate results is closely related to that of Rényi (1962) and Holst (1972) but is very different to the methods of Morris (1975) which do not appear capable of giving results on rates.

The following results give rates in terms of

$$\ell_1 = n^{-3/2}\sum_{k=1}^{n} E|Y_k|^3 \quad \text{and} \quad \ell_2 = n^{-3/2}\sum_{k=1}^{n} E|Z_k|^3.$$

THEOREM 2. *If $\alpha_k \leq C_1\alpha$ and $c_2 < \alpha < n^{c_1}$, then*

$$\Delta < C\ell_2. \tag{1.2}$$

THEOREM 3. *If $\alpha_k \leq C_1\alpha$, $\alpha < C_2$ and*

$$|d(1, k) - d(0, k)| < Cn^{1/2}\ell_2, \quad \text{for all } k, \tag{1.3}$$

then

$$\Delta < C(\ell_1 + \ell_2 + \ell_2^{-2}e^{-cN}). \tag{1.4}$$

If in addition

$$\ell_2 \geq CN^{-1/2} \tag{1.5}$$

then (1.2) *holds.*

THEOREM 4. *If* $\alpha_k \leq C_1\alpha$, $\alpha > C$ *and*

$$E\,|\,d(Y_k' + 1, k) - d(Y_k', k)\,| < C\alpha_k^{-1/2}, \quad \textit{whenever } \alpha_k > \tfrac{1}{2}\,\alpha, \tag{1.6}$$

then (1.2) *holds.*

We will prove these results in Section 2 and discuss three important special cases in Section 3.

2. Proofs.

PROOF OF THEOREM 1. Let $Y = n^{-1/2}\sum_k Y_k$, $Z = n^{-1/2}\sum_k Z_k$, so that

$$Ee^{ivY+itZ} = \prod_{k=1}^{n} g_k(n^{-1/2}v, n^{-1/2}t),$$

where

$$g_k(n^{-1/2}v, n^{-1/2}t) = E\exp(in^{-1/2}vY_k + in^{-1/2}tZ_k).$$

According to a result of Bartlett (1938),

$$\psi_N(t) = E(e^{itZ}\,|\,\textstyle\sum_k Y_k' = N) = d\displaystyle\int_{-\pi N^{1/2}}^{\pi N^{1/2}} \prod_k g_k(n^{-1/2}v, n^{-1/2}t)\,dv,$$

where

$$d = [2\pi N^{1/2}P(\textstyle\sum_k Y_k' = N)]^{-1} = \dfrac{N!e^N}{2\pi N^{N+1/2}} = (2\pi)^{-1/2}(1 + O(N^{-1})).$$

Theorem 1 will be proved if we can show that $\psi_N(t) \to \exp(-\frac{1}{2}t^2)$ for all fixed t. To do this we consider the partition

$$\begin{aligned} &|\,\psi_N(t) - \exp(-\tfrac{1}{2}t^2)\,| \\ &\quad \leq d\int_{R_1} |\,\textstyle\prod_k g_k(n^{-1/2}v, n^{-1/2}t) - \exp(-\frac{1}{2}v^2 - \frac{1}{2}t^2)\,|\,dv \\ &\quad + d\int_{R_2} |\,\textstyle\prod_k g_k(n^{-1/2}v, n^{-1/2}t)\,|\,dv + CN^{-1} + C\exp(-c\ell_1^{-2}) \end{aligned} \tag{2.1}$$

where

$$R_1 = \{v\colon |\,v\,| < \lambda\ell_1^{-1}\} \text{ and } \quad R_2 = \{v\colon \lambda\ell_1^{-1} < |\,v\,| < \pi N^{1/2}\},$$

where $\lambda > 0$ is specified below. We will bound the first two integrals here using the following two lemmas.

LEMMA 1. *Under the conditions of Theorem* 1, *for fixed* t, ε *sufficiently small and* n *sufficiently large,*

$$|\prod_k g_k(n^{-1/2}v, n^{-1/2}t) - \exp(-\tfrac{1}{2}v^2 - \tfrac{1}{2}t^2)| < \varepsilon P(v,t)\exp(-(v^2+t^2)/12) \tag{2.2}$$

for $|v| < \lambda \ell_1^{-1}$ *with* $\lambda = \frac{1}{3} C_1^{-1/2}$, *where* $P(v, t)$ *is a polynomial in* $|v|$ *and* $|t|$ *of degree* 4.

We will defer the proof of this to the appendix, since it follows the same lines as that of Lemma 2 of QR.

LEMMA 2. *For fixed* t, $\alpha > c_2 > 0$ *and* $\lambda \ell_1^{-1} < |v| < \pi N^{1/2}$,

$$|\prod_k g_k(n^{-1/2}v, n^{-1/2}t)| < e^{-cn}. \tag{2.3}$$

PROOF. We have

$$\begin{aligned} &|g_k(n^{-1/2}v, n^{-1/2}t)| \\ &\quad \le |E \exp(in^{-1/2}vY_k)| + E|\exp(in^{-1/2}vY_k)(\exp(in^{-1/2}tZ_k) - 1)| \\ &\quad \le |\exp(\alpha_k(\exp(in^{-1/2}v/\alpha^{1/2}) - 1))| + n^{-1/2}|t|E|Z_k|. \end{aligned}$$

Since $\cos x - 1 \le -2x^2/\pi^2$ for $-\pi < x < \pi$,

$$\begin{aligned} &|g_k(n^{-1/2}v, n_{-1/2}t)| \\ &\qquad \le \exp[-2\alpha_k v^2/N\pi^2] + n^{-1/2}|t|E|Z_k| \\ &\qquad \le \exp[(\exp(-2\alpha_k v^2/N\pi^2) - 1) + n^{-1/2}|t|E|Z_k|] \\ &\qquad \le \exp[-c + n^{-1/2}|t|E|Z_k|] \end{aligned} \tag{2.4}$$

for $k \in B = \{k: \alpha_k \ge \frac{1}{2}\alpha\}$, since $|v| > \lambda \ell_1^{-1} > Cn^{1/2}$ for $\alpha > c_2$, from QR (equation (13)). Now from QR (page 669), B contains $m \ge n/(2C_1 - 1)$ elements, so

$$\begin{aligned} |\prod_k g_k(n^{-1/2}v, n^{-1/2}t)| &\le \exp[-cm + n^{-1/2}|t| \textstyle\sum_k E|Z_k|] \\ &\le \exp[-cn + |t|n^{1/2}] \end{aligned} \tag{2.5}$$

using the Hölder inequality. The lemma follows for any fixed t.

Now from (2.1), we get, by integrating (2.2) over R_1 and (2.3) over R_2,

$$|\psi_N(t) - \exp(-\tfrac{1}{2}t^2)| \le \varepsilon C + CN^{1/2}e^{-cn} + CN^{-1} + Ce^{-cn} \le \varepsilon C$$

if $c_2 < \alpha < n^{c_1}$, in which case the theorem follows since ε is arbitrarily small.

If we allow α to tend to zero, we can only expect the joint characteristic function of Y and Z to be $O(\exp(-cN))$ for $|v| > \lambda \ell_1^{-1} = O(N^{1/2})$, so Lemma 2 is replaced as follows.

LEMMA 3. *For fixed* t, $\alpha < (2C_1)^{-1}$ *and* $\lambda \ell_1^{-1} < |v| \le \pi N^{1/2}$,

$$|\prod_k g_k(n^{-1/2}v, n^{-1/2}t)| < \exp[-v^2/\pi^2 + n^{1/2}|t|]. \tag{2.6}$$

PROOF. If $\alpha < (2C_1)^{-1}$ then $\alpha_k < \frac{1}{2}$. In this case $2\alpha_k v^2/N\pi^2 \leq 1$, so from the second inequality of (2.4)

$$|g_k(n^{-1/2}v, n^{-1/2}t)| \leq \exp[-\alpha_k v^2/N\pi^2 + n^{-1/2}|t|E|Z_k|],$$

since $e^{-x} - 1 \leq \frac{1}{2}x$ for $0 < x < 1$. The lemma follows using the Hölder inequality again.

Integrating (2.6) over R_2 gives a value bounded above by

$$C \exp[-(\lambda \ell_1^{-1}/\pi)^2 + n^{1/2}|t|] \leq C \exp[-n^{1/2}(n^{1/2}C\alpha - |t|)]$$

which is arbitrarily small for fixed t if $n\alpha^2 \to \infty$. Using this result in (2.1) completes the proof of Theorem 1.

PROOF OF THEOREM 2. For $t > 0$,

$$\begin{aligned}
&\psi_N(t) - \exp(-\tfrac{1}{2}t^2) \\
&\quad = \exp(-\tfrac{1}{2}t^2) \int_0^t \frac{d}{ds} \exp(\tfrac{1}{2}s^2)\psi_N(s)\, ds \\
&\quad = \exp(-\tfrac{1}{2}t^2)\, d \int_0^t \left[\int_{-\pi N^{1/2}}^{\pi N^{1/2}} \frac{d}{ds} \exp(\tfrac{1}{2}s^2) \prod_k g_k(n^{-1/2}v, n^{-1/2}s)\, dv \right] ds.
\end{aligned}$$

So

$$\begin{aligned}
&t^{-1}|\psi_N(t) - \exp(-\tfrac{1}{2}t^2)| \\
&\quad \leq \exp(-\tfrac{1}{2}t^2) \int_{-\pi N^{1/2}}^{\pi N^{1/2}} \sup_{|s| \leq t} \left| \frac{d}{ds} \exp(\tfrac{1}{2}s^2) \prod_k g_k(n^{-1/2}v, n^{-1/2}s) \right| dv.
\end{aligned}$$

We will use the smoothing inequality as in Loève (1955, page 285), that for any $\delta > 0$,

$$\Delta \leq \frac{2}{\pi} \int_0^{\delta \ell_2^{-1}} t^{-1}|\psi_N(t) - \exp(-\tfrac{1}{2}t^2)|\, dt + C\ell_2/\delta. \tag{2.7}$$

From Lemma 2 of QR (which, as indicated in the proof, is true for an arbitrary score function $c(j, k)$) we have

$$\begin{aligned}
&|(d/ds)\exp(\tfrac{1}{2}s^2) \prod_k g_k(n^{-1/2}v, n^{-1/2}s)| \\
&\quad \leq C(|s| + |v| + 1)^3(\ell_1 + \ell_2)\exp\left(\frac{11t^2}{24} - \frac{s^2}{24}\right)
\end{aligned} \tag{2.8}$$

for $|v| < \frac{2}{9}\ell_1^{-1}$, $|s| < \delta\ell_2^{-1} \leq \frac{2}{9}\ell_2^{-1}$ so long as $\ell_1, \ell_2 < 12^{-3/2}$. This will enable us to bound the integral of the term on the left over the region R_1 with λ henceforth taken to be 2/9. To bound the integral over R_2, we note that

$$\begin{aligned}
|(d/ds)g_k(n^{-1/2}v, n^{-1/2}s)| &= |En^{-1/2}Z_k(\exp(in^{-1/2}vY_k + in^{-1/2}sZ_k) - 1)| \\
&\leq n^{-1}|v|E|Y_kZ_k| + n^{-1}|s|EZ_k^2
\end{aligned}$$

so that using the Hölder inequality

$$|(d/ds)\exp(\tfrac{1}{2}s^2)\prod_k g_k(n^{-1/2}v, n^{-1/2}s)| \le (|v| + 2|s|)\exp(\tfrac{1}{2}s^2)\prod{}^*, \tag{2.9}$$

where

$$\prod{}^* = \max_j |\prod_{k\in B, k\neq j} g_k(n^{-1/2}v, n^{-1/2}s)|.$$

The following lemma provides a bound on $\prod^*$ which we can use for $v \in R_2$ when $c_2 < \alpha < n^{c_1}$ and for $\frac{2}{9}\ell_1^{-1} < |v| < 4n^{1/2}$ when $\alpha > n^{c_1}$.

LEMMA 4. *For* $\alpha > c_2$,

$$\prod{}^* \le e^{-cn}. \tag{2.10}$$

PROOF. From (2.4)

$$\prod{}^* \le \exp(-c(m-1) + n^{-1/2}|s| \sum_k E|Z_k|) \le \exp(-cn + \delta n)$$

using the Hölder inequality and noting that $|s| \le \delta\ell_2^{-1} \le \delta n^{1/2}$. The lemma follows by choosing δ small enough.

To prove Theorem 2 we integrate the bound given by (2.8) over R_1 and the bound given by (2.9) over R_2, bearing in mind Lemma 4, to obtain for $t > 0$

$$t^{-1}|\psi_N(t) - \exp(-\tfrac{1}{2}t^2)| \le P(t)\exp(-t^2/24)(\ell_1 + \ell_2) + Nte^{-cn};$$

from this inequality and (2.7),

$$\Delta \le C(\ell_1 + \ell_2 + nN\,e^{-cn}) \le C\ell_2$$

since $c_2 < \alpha < n^{c_1}$ and $\ell_2 \ge n^{-1/2} \ge c\ell_1$.

PROOF OF THEOREM 3. For the same reasons given immediately prior to Lemma 3, (2.10) fails for $\alpha \to 0$. However in this case the same argument as above combined with the following lemma yields (1.4). If (1.5) holds then (1.2) follows easily from (1.4) using $\ell_1 \le cN^{-1/2}$ for $\alpha < 1$. Since the constant c_2 in Theorem 2 can be taken arbitrarily small, Theorem 3 follows.

LEMMA 5. *For sufficiently small* α, *if* (1.3) *holds*,

$$\prod{}^* < e^{-cN}.$$

PROOF. We have

$$|g_k(n^{-1/2}v, n^{-1/2}s)| \le e^{-\alpha_k}(M_{k1} + M_{k2})$$

where M_{k1} is derived from the first two terms of the series for the expectation and M_{k2} from the other terms. Then

$$\begin{aligned} M_{k1} &\le |1 + \alpha_k\exp(iN^{-1/2}v + in^{-1/2}s[d(1,k) - d(0,k)])| \\ &= (1 + 2\alpha_k\cos\tau + \alpha_k^2)^{1/2} \end{aligned}$$

where

$$\tau = N^{-1/2}v + n^{-1/2}s[d(1, k) - d(0, k)], \quad \text{and}$$

$$M_{k2} \leq e^{\alpha_k} - 1 - \alpha_k.$$

Using $\pi N^{1/2} > |v| > CN^{1/2}$, $|s| < \delta\ell_2^{-1}$ and (1.3), we can choose $\eta > 0$ such that

$$\eta < C - \delta C < |\tau| < \pi + \delta C < \pi + \tfrac{1}{5}C$$

if δ is chosen small enough. Since $\cos\tau - 1 \leq -\tau^2/5$ for $|\tau| < \pi + \frac{1}{50}$,

$$M_{k1} \leq (1 + 2(1 - \eta_1)\alpha_k + \alpha_k^2)^{1/2} \leq 1 + (1 - \eta_2)\alpha_k$$

with $\eta_1 = \eta^2/5$, if $\eta_2 < \eta_1$ and $\alpha_k < 2(\eta_1 - \eta_2)/[\eta_2(2 - \eta_2)] = \eta_3$. In this case, from the inequality $e^x < 1 + (1 + \theta)x$ for $0 < x < \theta < 1.5$,

$$M_{k2} < \eta_3\alpha_k;$$

so

$$|g_k(n^{-1/2}v, n^{-1/2}t)| \leq e^{-\alpha_k}(1 + (1 - \eta_2)\alpha_k + \eta_3\alpha_k)$$

$$< \exp(-(\eta_2 - \eta_3)\alpha_k) < 1$$

if $\eta_3 < \eta_2$, that is if $\eta_2^2 > 2(\eta_1 - \eta_2)/(2 - \eta_2)$, which can be achieved by choosing η_2 sufficiently close to η_1. Thus for small α,

$$\prod{}^* < \max_j \exp(-c \textstyle\sum_{k\in B, k\neq j} \alpha_k) < e^{-c(m-1)\alpha}$$

and the lemma follows.

PROOF OF THEOREM 4. Although (2.10) is true for all $\alpha > c_2$, a tighter bound is necessary when $\alpha > n^{c_1}$ for $Cn^{1/2} < |v| < \pi N^{1/2}$.

LEMMA 6. *Under the conditions of Theorem* 4,

$$\prod{}^* \leq (3n^{1/2}/|v|)^{cn}. \tag{2.11}$$

PROOF. Writing $p_j = e^{-\alpha_k}\alpha_k^j/j!$ and $d_j = d(j, k)$,

$$|g_k(\alpha^{1/2}a, b)|$$

$$= |\textstyle\sum_{j=0}^{\infty} \exp(iaj + ibd_j)p_j|$$

$$= |e^{ia} - 1|^{-1}|\textstyle\sum_{j=0}^{\infty} (e^{ia(j+1)} - e^{iaj})e^{ibd_j}p_j|$$

$$\leq (\pi/(2|a|))[p_0 + \textstyle\sum_{j=0}^{\infty} |\exp(ibd_{j+1}) - \exp(ibd_j)|\, p_j + \sum_{j=0}^{\infty} |p_{j+1} - p_j|],$$

using the summation formula

$$\textstyle\sum_{j=0}^{\infty}(f_{j+1} - f_j)g_j = -f_0g_0 - \sum_{j=0}^{\infty} (g_{j+1} - g_j)f_{j+1}$$

and the inequality

$$|e^{ia} - 1| = [2(1 - \cos a)]^{1/2} \geq 2|a|/\pi \quad \text{for} \quad |a| < \pi.$$

Now

$$\sum_{j=0}^{\infty} |\exp(ibd_{j+1}) - \exp(ibd_j)|\, p_j \le |b|\, E\,|d(Y_k' + 1, k) - d(Y_k', k)|$$

and

$$p_0 + \sum_{j=0}^{\infty} |p_{j+1} - p_j| = p_0 + \sum_{\ell=1}^{\infty} |p_\ell - p_{\ell-1}| = \sum_{\ell=0}^{\infty} p_\ell |1 - \ell/\alpha_k|$$
$$= \alpha_k^{-1} E\,|Y_k' - \alpha_k| \le \alpha_k^{-1/2}.$$

Thus for $k \in B$ and small δ, (1.6) and $|s| < \delta \ell_2^{-1}$ imply

$$|g_k(n^{-1/2}v, n^{-1/2}s)| \le \frac{\pi n^{1/2}}{2|v|}[C\delta + 1]\left(\frac{\alpha}{\alpha_k}\right)^{1/2} < 3n^{1/2}/|v|,$$

and the lemma follows.

To prove Theorem 4, we argue as before for $v \in R_1$, use the bound (2.10) for $\frac{2}{9}\ell_1^{-1} < |v| < 4n^{1/2}$ and the bound (2.11) for $4n^{1/2} < |v| < \pi N^{1/2}$ to get

$$t^{-1}|\psi_n(t) - e^{-(1/2)t^2}| \le P(t)\exp(-t^2/24)(\ell_1 + \ell_2) + (1 + t)(e^{-cn} + (\tfrac{3}{4})^{cn}).$$

Theorem 4 now follows from (2.7).

3. Some examples. We consider three examples from statistics, in which it is most natural to work with third moments. We therefore concentrate mainly on rate results.

The *empty cell test* is based on $c(j, k) = \delta_{0j}$. Rényi (1962) showed that when $\alpha_k = \alpha$, $k = 1, \cdots, n$, a necessary and sufficient condition for $\Delta \to 0$ is $n\sigma^2 \to \infty$. In QR it was shown that when $\alpha_k \le C_1\alpha$, $k = 1, \cdots, n$,

$$\Delta = O((n\sigma^2)^{-1/2}), \tag{3.1}$$

so that $n\sigma^2 \to \infty$ is sufficient for $\Delta \to 0$ in this case too. (3.1) also follows from our present theorems as we now show. Since (QR equation (14)) $\ell_2 = O((n\sigma^2)^{-1/2})$, Theorem 2 gives (3.1) immediately for $c_2 < \alpha < n^{c_1}$ and we need only check the conditions for $\alpha \to 0$ and $\alpha \to \infty$.

When $\alpha \to 0$ (so $\gamma \downarrow -1$) we have

$$n^{1/2}\ell_2 \ge \tfrac{1}{2}n^{-1}\sigma^{-3}\sum_k |e^{-\alpha_k} + \gamma(2 - \alpha_k)|^3\alpha_k^2 e^{-\alpha_k}$$
$$\ge C\alpha^2/\sigma^3 \ge C/\sigma \ge C|1 + \gamma|/\sigma = C|d(1, k) - d(0, k)|$$

using $\sigma^2 = O(\alpha^2)$. Similarly $n^{1/2}\ell_2 \ge C/\alpha$, so Theorem 3 applies. When $\alpha \to \infty$,

$$E\,|d(Y_k' + 1, k) - d(Y_k', k)| = (\gamma + e^{-\alpha_k})/\sigma = O(1/\alpha_k)$$

since $\sigma^2 \ge \frac{1}{2}\sum_k \alpha_k^2 e^{-2\alpha_k}$, so Theorem 4 applies.

The χ^2 *test* is based on $c(j, k) = (j - \alpha_k)^2/\alpha_k$. Morris (1975) showed that $\Delta \to 0$ if $\alpha_k \ge c > 0$, $k = 1, \cdots, n$, and $\max_k p_k \to 0$ (the latter being equivalent to $\ell_1 \to 0$). Our results are somewhat different as discussed in Section 1; apart from providing a bound on Δ, they also improve substantially the results of Steck (1957) as quoted in Morris (1975) and those of Holst (1972), as we now show.

It can be shown that $\gamma = 1/\alpha$, $\sigma^2 = 2 + n^{-1} \sum_k \alpha_k \beta_k^2$, where $\beta_k = 1/\alpha_k - 1/\alpha$, and that

$$\begin{aligned} & 8 + n^{-1} \textstyle\sum_k \alpha_k \beta_k^3 + 22 n^{-1} \sum_k \alpha_k \beta_k^2 + 4/\alpha \\ & \quad \le n^{-1} \textstyle\sum_k E(Z_k' - EZ_k')^3 \\ & \quad \le n^{-1} \textstyle\sum_k E \mid Z_k' - EZ_k' \mid^3 \\ & \quad \le n^{-1} \textstyle\sum_k E(Z_k' - EZ_k')^3 + 2 \sum_k E((Y_k' - \alpha_k)^2/\alpha + 1)^3 \\ & \quad \le 10 + n^{-1} \textstyle\sum_k \alpha_k \beta_k^3 + 22 n^{-1} \sum_k \alpha_k \beta_k^2 + 10/\alpha. \end{aligned} \tag{3.2}$$

In this and the next example, we make the simplifying assumption

$$c\alpha \le \alpha_k \le C_1 \alpha, \quad k = 1, \cdots, n, \tag{3.3}$$

which is equivalent to $\mid \beta_k \mid < C/\alpha$, $k = 1, \cdots, n$.

From (3.2),

$$\begin{aligned} n^{1/2} \ell_2 & \le C(1 + n^{-1} \textstyle\sum \alpha_k \mid \beta_k \mid^3 + n^{-1} \sum \alpha_k \beta_k^2 + 1/\alpha)/\sigma^3 \\ & \le C(1 + \max_k \mid \beta_k \mid + 1/\alpha) = O(1) + O(1/\alpha). \end{aligned} \tag{3.4}$$

Since (1.1) is implied by $\ell_2 \to 0$, it follows from Theorem 1 that $\Delta \to 0$ so long as $n\alpha^2 \to \infty$ and $\alpha \le n^{c_1}$. We remark that when $\alpha_k = \alpha$, $k = 1, \cdots, n$, and $\alpha \to 0$, $n\alpha^2 \to \infty$ is necessary and sufficient for $\Delta \to 0$ (Quine, 1980). Theorem 2 and (3.4) show that the rate of convergence is $O(n^{-1/2})$ for $c_2 < \alpha < n^{c_1}$, and for larger α the same result is implied by Theorem 4, since for $k \in B$

$$E \mid d(Y_k' + 1, k) - d(Y_k', k) \mid = E \mid 2(Y_k' - \alpha_k)/\alpha_k + \beta_k \mid /\sigma = O(\alpha_k^{-1/2}).$$

The *likelihood ratio test* is based on $c(j, k) = j \log(j/\alpha_k)$ with $0 \log 0 = 0$. Related results are the central limit theorems of Morris (1975) under the same conditions as his χ^2 result and Holst (1973) under $\alpha_k \le C_1 \alpha$ and α fixed. In this case the analysis is facilitated for $\alpha > c_2$ (which we now assume) by use of the function $I(j, \alpha) = j \log(j/\alpha) - j + \alpha$, some of whose properties are given in Lemma 5.1 of Morris (1975). In particular it gives

$$n\sigma^2 \ge \tfrac{1}{2} \textstyle\sum_k \alpha_k^2/(2 + \alpha_k)^2 \ge \tfrac{1}{8} \sum_{k \in B} \alpha^2/(2 + C_1\alpha)^2 \ge Cn, \tag{3.5}$$

and since $0 \le I(j, \alpha) \le (j - \alpha)^2/\alpha$,

$$\mid \gamma - 1 \mid = \mid N^{-1} \textstyle\sum_k \operatorname{cov}(I(Y_k', \alpha_k), Y_k') \mid \le 1/\alpha + C/\alpha^{1/2}. \tag{3.6}$$

Since $Z_k' = I(Y_k', \alpha_k) - (\gamma - 1)(Y_k' - \alpha_k)$, it follows that

$$E \mid Z_k' - EZ_k' \mid^3 = O(\alpha_k^{-2}) + O(1),$$

so that from Theorem 2, $\Delta \le Cn^{-1/2}$ if (3.3) holds and $c_2 < \alpha < n^{c_1}$. The same result is true for larger α as well from Theorem 4, as we now show. We have

$$\begin{aligned} & E \mid d(Y_k', k + 1, k) - d(Y_k', k) \mid \\ & \quad = \sigma^{-1} E \mid I(Y_k' + 1, \alpha_k) - I(Y_k', \alpha_k) + 1 - \gamma \mid \\ & \quad \le \sigma^{-1}(\alpha_k^{-1} + E \mid Y_k' - \alpha_k \mid (\alpha_k^{-1} + 2/(Y_k' + 1)) + \mid 1 - \gamma \mid) \end{aligned} \tag{3.7}$$

from Morris (1975, equation (5.31)). Now

$$\begin{aligned} E\,|\,Y_k' - \alpha_k\,|/(Y_k' + 1) &\le E(1/(Y_k' + 1)) + E\,|\,1 - \alpha_k/(Y_k' + 1)\,| \\ &= \alpha_k^{-1}(1 - e^{-\alpha_k}) + \alpha_k^{-1}(E\,|\,Y_k' - \alpha_k\,| - \alpha_k e^{-\alpha_k}) \\ &= O(\alpha_k^{-1/2}), \end{aligned}$$

which together with (3.5), (3.6) and (3.7) gives (1.6).

APPENDIX

PROOF OF LEMMA 1. We have

$$\begin{aligned} &|\,g_k(n^{-1/2}v,\, n^{-1/2}t) - 1\,| \\ &\quad \le \tfrac{1}{2}\, n^{-1}E(vY_k + tZ_k)^2 \le n^{-1}C_1v^2 + \varepsilon t^2 \le \tfrac{1}{9} + \varepsilon t^2 \end{aligned} \tag{A1}$$

for $n > n_\varepsilon$, using (1.1) and $\ell_1^{-1} \le n^{1/2}$; summing the first inequality over k also gives

$$\textstyle\sum_k |\,g_k(n^{-1/2}v,\, n^{-1/2}t) - 1\,| \le \tfrac{1}{2}\,(v^2 + t^2). \tag{A2}$$

If ε is small enough to make the final bound in (A1) less than $\frac{1}{4}$ for all k then using the inequality $|\log(1 + z) - z| < |z|^2$, $|z| < \frac{1}{2}$ in (A1) and (A2), we find that for $n > n_\varepsilon$,

$$\begin{aligned} &\textstyle\sum_k |\log g_k(n^{-1/2}v,\, n^{-1/2}t) - (g_k(n^{-1/2}v,\, n^{-1/2}t) - 1)| \\ &\quad \le \min(\tfrac{1}{2}\,\varepsilon(v^2 + t^2)^2,\, \tfrac{1}{8}(v^2 + t^2)). \end{aligned} \tag{A3}$$

Next,

$$\begin{aligned} &|\textstyle\sum_k (g_k(n^{-1/2}v,\, n^{-1/2}t) - 1 + \tfrac{1}{2}\, n^{-1}v^2EY_k^2 + \tfrac{1}{2}\, n^{-1}t^2EZ_k^2)| \\ &\quad = |\textstyle\sum_k E(\exp(ivn^{-1/2}Y_k + itn^{-1/2}Z_k) - 1 - i(vn^{-1/2}Y_k + tn^{-1/2}Z_k) \\ &\qquad + \tfrac{1}{2}\,(vn^{-1/2}Y_k + tn^{-1/2}Z_k)^2)| \\ &\quad \le \tfrac{1}{6}\textstyle\sum_k E\,|\,n^{-1/2}vY_k + n^{-1/2}tZ_k\,|^3 I(|\,n^{-1/2}Z_k\,| < \varepsilon) \\ &\qquad + \textstyle\sum_k E(n^{-1/2}vY_k + n^{-1/2}tZ_k)^2\, I(|\,n^{-1/2}Z_k\,| \ge \varepsilon) \\ &\quad \le \tfrac{2}{3}\textstyle\sum_k E\,|\,n^{-1/2}vY_k\,|^3 + \tfrac{2}{3}\,\varepsilon\,|\,t\,|\textstyle\sum_k E(n^{-1/2}tZ_k)^2 \\ &\qquad + 2v^2\textstyle\sum_k n^{-1}EY_k^2 I(|\,n^{-1/2}Z_k\,| \ge \varepsilon) \\ &\qquad + 2t^2\textstyle\sum_k n^{-1}EZ_k^2 I(|\,n^{-1/2}Z_k\,| \ge \varepsilon). \end{aligned} \tag{A4}$$

Now

$$\begin{aligned} \textstyle\sum_k n^{-1}EY_k^2 I(|\,n^{-1/2}Z_k\,| \ge \varepsilon) &\le \textstyle\sum_k n^{-1}EY_k^2\,|\,n^{-1/2}Z_k/\varepsilon\,|^{2/3} \\ &\le (\textstyle\sum_k n^{-1}E\,|\,Y_k\,|^3)^{2/3}(\textstyle\sum_k n^{-1}E(n^{-1/2}Z_k/\varepsilon)^2)^{1/3} \\ &= (\ell_1/\varepsilon)^{2/3}, \end{aligned}$$

which together with (A3) and (A4) gives for $n > n_\varepsilon$

$$|\sum_k \log(g_k(n^{-1/2}v, n^{-1/2}t) + \tfrac{1}{2}v^2 + \tfrac{1}{2}t^2)| \leq \min(\varepsilon P(v, t), \tfrac{5}{12}(v^2 + t^2))$$

and the lemma follows using $|e^z - 1| \leq |z| e^{|z|}$.

REFERENCES

[1] BARTLETT, M. S. (1938). The characteristic function of a conditional statistic. *J. London Math. Soc.* **13** 62–67.

[2] ENGLUND, G. (1981). A remainder term estimate for the normal approximation in classical occupancy. *Ann. Probab.* **9** 684–692.

[3] HOLST, L. (1972). Asymptotic normality and efficiency for certain goodness of fit tests. *Biometrika* **59** 137–145.

[4] LOÈVE, M. (1955). *Probability Theory.* Van Nostrand, New York.

[5] 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. Statist.* **13** 306–317.

[6] MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188.

[7] QUINE, M. P. (1980). Three limit theorems for scores based on occupancy numbers. *Ann. Probab.* **8** 148–156.

[8] QUINE, M. P. (1983). A Berry-Esseen bound for scores based on occupancy numbers. *Probability and Mathematical Statistics: Essays in honour of Carl-Gustav Esseen* 140–153. Uppsala University.

[9] QUINE, M. P. and ROBINSON, J. (1982). A Berry-Esseen bound for an occupancy problem. *Ann. Probab.* **10** 663–671.

[10] RÉNYI, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Publ. Math. Inst. Hung. Acad. Sci.* **7** 203–214.

[11] STECK, G. P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** No. 12, 237–284.

DEPARTMENT OF MATHEMATICAL STATISTICS
UNIVERSITY OF SYDNEY
N.S.W. 2006
AUSTRALIA