Pacific Journal of Mathematics

ON THE DIFFERENCE BETWEEN THE EMPIRICAL HISTOGRAM AND THE NORMAL CURVE, FOR SUMS. II

PERSI W. DIACONIS AND DAVID AMIEL FREEDMAN

Vol. 100, No. 2 October 1982

PACIFIC JOURNAL OF MATHEMATICS
Vol. 100, No. 2, 1982

ON THE DIFFERENCE BETWEEN THE EMPIRICAL HISTOGRAM AND THE NORMAL CURVE, FOR SUMS: PART II

PERSI DIACONIS AND DAVID FREEDMAN

1. Introduction. Let $X_1, X_2, \cdots$, be independent, identically distributed random variables. Suppose the X_i are integer-valued and have span one:

(1.1) g.c.d.$\{j - k: j, k \in S > 0\} = 1$, where $j \in S$ iff $P\{X_1 = j\} > 0$.

Suppose too

$$E(X_1^4) < \infty . \tag{1.2}$$

Let

$$\mu = E(X_1) , \qquad \sigma^2 = \operatorname{Var} X_1 , \qquad \mu_3 = E[(X_1 - \mu)^3] . \tag{1.3}$$

Let $S_n = X_1 + \cdots + X_n$. Take k independent copies of S_n, and let N_{nj} be the number of these sums which are equal to j. Up to scaling, the counts N_{nj} correspond to the empirical histogram for the k sums.

Of course,

$$E(N_{nj}) = kp_{nj} , \quad \text{where} \quad p_{nj} = P(S_n = j) .$$

In a previous paper [2] we studied the behavior of

$$\max_j (N_{nj} - kp_{nj}) , \tag{1.4}$$

corresponding to the maximum deviation between the empirical histogram and its expected value. In this paper we will study the maximum deviation between the empirical histogram and an approximation to the expected value, based on the normal curve.

In more detail, the probabilities p_{nj} can be well approximated by

$$\tilde{p}_{nj} = \frac{1}{\sigma\sqrt{2\pi n}} \exp\left(-\frac{1}{2}t_{nj}^2\right) \tag{1.5}$$

where

$$t_{nj} = (j - n\mu)/(\sigma\sqrt{n}) .$$

The asymptotic behavior of the maximum deviation

$$\max_j (N_{nj} - k\tilde{p}_{nj}) \tag{1.6}$$

is the topic of this paper.

Our main results give the asymptotic distribution of the location and size of this maximum deviation. When the number of repetitions k is "small", sampling error dominates and the maximum deviation is asymptotically the same as the maximum in (1.4). When the number of repetitions k is "large", the bias term enters. In both cases the maximum deviation is taken on at a unique location with probability approaching one. The location and size of the maximum are asymptotically independent; and suitably normalized the location has a limiting normal distribution while the size has a limiting extreme value distribution. The results are more carefully described in §2.

Of course, the empirical histogram could be approximated by the normal curve directly. In this case too, the asymptotic behavior of the maximum deviation can be analyzed by methods very similar to the ones presented here, but we do not pursue the details. Similar remarks apply to the frequency polygon derived from the empirical histogram, and to Edgeworth expansions for p_{nj}.

2. The normal approximation.

Clearly,

$$N_{nj} - k\tilde{p}_{nj} = (N_{nj} - kp_{nj}) + k(p_{nj} - \tilde{p}_{nj}) . \tag{2.1}$$

The first term on the right represents sampling error; the second, basis. Suppose that

$$k/[n^{1/2}(\log n)^3] \longrightarrow \infty \quad \text{as} \quad n \longrightarrow \infty . \tag{2.2}$$

This condition insures that the histogram converges uniformly to the expected histogram p_{nj}. See [3] for further discussion. Suppose too

$$\mu_3 = E[(X_1 - \mu)^3] \neq 0 , \quad \text{where} \quad \mu = E(X_1) . \tag{2.3}$$

The results of this section can be summarized as follows.

If $k \ll n^{3/2}$, bias is negligible, so $\max_j (N_{nj} - k\tilde{p}_{nj})$ shows the same asymptotic behavior as $\max_j (N_{nj} - kp_{nj})$. This maximum has been carefully analyzed in [2].

If $k \gg n^{3/2} \log n$, sampling error is negligible. The maximum is analyzed in §3.

If k is between $n^{3/2}$ and $n^{3/2} \log n$ in order of magnitude, sampling error and bias both contribute to $\max_j (N_{nj} - k\tilde{p}_{nj})$. The asymptotic behavior of $\max_j (N_{nj} - k\tilde{p}_{nj})$ will be described in this section.

If $\mu_3 = 0$, the critical rates for k change: we do not pursue this. Likewise, if (2.2) fails, the asymptotics change: large deviation corrections become relevant. We do not pursue this either. Finally, if the fourth-moment condition is dropped, new behavior is possible:

see §5 of [2] for a related discussion. We begin with case $k = O(n^{3/2} \log n)$, and use Theorem (1.24) of [4].

The following notation will be helpful, although it seems tedious indeed. In view of (1.1-1.2), we have from the Edgeworth expansion

$$\sigma\sqrt{2\pi n}\,(p_{nj} - \tilde{p}_{nj}) = \frac{c}{\sqrt{n}} H_3(t_{nj}) \exp\left(-\frac{1}{2}t_{nj}^2\right) + O\left(\frac{1}{n}\right) \tag{2.4}$$

where

$$t_{nj} = (j - n\mu)/(\sigma\sqrt{n})$$
$$H_3(t) = t^3 - 3t$$
$$c = \frac{1}{6}\mu_3/\sigma^3$$
$$\mu = E(X_1)\,, \qquad \sigma^2 = \operatorname{Var} X_1\,, \qquad \mu_3 = E[(X_1 - \mu)^3]\,.$$

The "0" is uniform in j. Let

$$\ell = \sqrt{\frac{k}{\sigma\sqrt{2\pi n}}}\,. \tag{2.5}$$

Then

$$(N_{nj} - k\tilde{p}_{nj})/\ell = \alpha_{nj} Z_{nj} + \beta_{nj}[2 \log (\sigma\sqrt{n})]^{1/2} \tag{2.6}$$

where

$$Z_{nj} = (N_{nj} - kp_{nj})/\sqrt{kp_{nj}} \tag{2.7}$$

$$\alpha_{nj} = (\sigma\sqrt{2\pi n}\; p_{nj})^{1/2} \tag{2.8}$$

$$\beta_{nj} = [2 \log (\sigma\sqrt{n})]^{-1/2} \cdot \ell \cdot \sigma\sqrt{2\pi n}\,(p_{nj} - \tilde{p}_{nj})\,. \tag{2.9}$$

By (2.4), β_{nj} can be approximated as

$$\beta_{nj} = \gamma_n \beta(t_{nj}) + O(\gamma_n/\sqrt{n}) \tag{2.10}$$

where

$$\gamma_n = [2 \log (\sigma\sqrt{n})]^{-1/2} \cdot n^{-3/4} \cdot k^{1/2} \tag{2.11}$$

$$\beta(t) = \sigma^{-1/2}(2\pi)^{-1/4} c H_3(t) \exp\left(-\frac{1}{2}t^2\right)\,. \tag{2.12}$$

Let

$$\alpha(t) = \exp\left(-\frac{1}{4}t^2\right) \tag{2.13}$$

$$w_n(x) = (\log n - 2 \log\log n + x + \log 4\sigma^2)^{1/2} \tag{2.14}$$

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(-\frac{1}{2}u^2\right) du\,. \tag{2.15}$$

The main result of this section is (2.17), which disposes of the case $k = O(n^{3/2} \log n)$. We next give the precise conditions for this result to hold.

Condition for (2.16). Suppose (1.1–1.3) and (2.2). Do not assume (2.3). Define γ_n by (2.11). Suppose $\gamma_n \to \gamma$ finite as $n \to \infty$. Note that $\gamma = 0$ is allowed. Suppose, as will be the case for most γ's, that the function $\alpha + \gamma\beta$ has a unique global maximum, say at t_∞; and that $\alpha''(t_\infty) + \gamma\beta''(t_\infty) < 0$. Abbreviate

$$\rho^2 = -[\alpha''(t_\infty) + \gamma\beta''(t_\infty)]/\alpha(t_\infty) > 0 \,. \tag{2.16}$$

As is easily seen, for n sufficiently large, $\alpha + \gamma_n\beta$ has a unique global maximum, say at t_n; and $t_n \to t_\infty$.

(2.17) PROPOSITION. *Suppose the conditions given above: in particular,*

$$n^{1/2}(\log n)^3 \ll k = O(n^{3/2} \log n) \,.$$

With probability approaching one, $M_n = \max_j (N_{nj} - k\tilde{p}_{nj})$ is assumed at a unique index L_n. Furthermore, the chance that

$$\rho[2 \log (\sigma\sqrt{n})]^{1/2} \cdot \left[\frac{1}{\sigma\sqrt{n}}(L_n - n\mu) - t_n\right] < y$$

and

$$M_n/\mathscr{L} < \alpha(t_n)w_n(x) + \alpha_n\beta(t_n)[2 \log (\sigma\sqrt{n})]^{1/2}$$

converges to

$$\Phi(y) \exp\left\{-\frac{1}{2\rho}e^{-x/2}\right\} \,.$$

Proof. Let I be a long (but finite) closed interval, which contains t_∞ as an interior point. If the j in $\max_j (N_{nj} - k\tilde{p}_{nj})$ is restricted so that $t_{nj} \in I$, the conclusions of the proposition follow from Theorem (1.24) of [4], taking $\varepsilon_n = 1/(\sigma\sqrt{n})$ and $c_n = n\mu$ and $\alpha_n = \alpha$ and $\beta_n = \gamma_n\beta$, so $\beta_\infty = \gamma\beta$. Conditions (1.1–23) of [4] are satisfied by our assumptions and the Edgeworth expansion (2.4).

It only remains to show that if I is long enough, the j's with $t_{nj} \notin I$ make essentially no contribution to the maximum: compare (2.34) of [4]. Indeed, the maximum over I has been proved to be of order

$$\rho \cdot [\alpha(t_\infty) + \gamma\beta(t_\infty)] \cdot \sqrt{\log n} \,.$$

Now t_∞ is the location of the global maximum of $\alpha(\cdot) + \gamma\beta(\cdot)$, which is necessarily positive, for $\alpha(0) + \gamma\beta(0) > 0$. By (4.1-3) of [2], for I long enough, the maximum of $N_{nj} - kp_{nj}$ over j with $t_{nj} \notin I$ is, with probability near one, only a small multiple of $\ell \cdot \sqrt{\log n}$. Likewise, by (2.4) of the present paper, the maximum of $k(p_{nj} - \tilde{p}_{nj})$ over j with $t_{nj} \notin I$ is bounded above by

$$cn^{-1/2}\ell^2\left[\sup_{t \notin I} H_3(t) \exp\left(-\frac{1}{2}t^2\right)\right] + O(n^{-1}\ell^2) .$$

In the remainder term, $n^{-1}\ell^2 = o[\ell(\log n)^{1/2}]$. In the lead term, $n^{-1}\ell^2 = O[\ell \cdot (\log n)^{1/2}]$, and the sup is small for I long. It was at this point that the growth condition $k = O(n^{3/2} \log n)$ became critical. □

(2.18) COROLLARY. *Suppose* (1.1-2). *Suppose* $k/n^{1/2}(\log n)^3 \to \infty$ *as* $n \to \infty$, *but* $k/n^{3/2} \to 0$. *Then, the asymptotic joint distribution of the location* L_n *and size* M_n *of* $\max_j (N_{nj} - k\tilde{p}_{nj})$ *coincides with that of* $\max_j (N_{nj} - kp_{nj})$, *as determined in* [2].

Proof. Clearly, α has its maximum of 1 at $t_\infty = 0$, where it is locally quadratic:

$$\alpha(t) = 1 - \frac{1}{4}t^2 + O(t^4) \quad \text{as} \quad t \longrightarrow 0 .$$

Furthermore,

$$\beta(t) = bt + O(t^3) \quad \text{as} \quad t \longrightarrow 0 ,$$

where b depends on σ and μ_3; it may vanish. Recall that t_n is the location of the maximum of $\alpha + \gamma_n\beta$. Recall γ_n from (2.11) and verify that $\gamma_n \to 0$.

Some easy calculus shows that

$$t_n = O(\gamma_n) = o[(\log \sigma\sqrt{n})^{-1/2}]$$

so t_n may be dropped from the normalization of L_n in (2.11). Likewise,

$$\gamma_n\beta(t_n) = O(\gamma_n^2) = o(1/\log n)$$

so the term

$$\gamma_n\beta(t_n)[2 \log (\sigma\sqrt{n})]^{1/2} = o[(\log n)^{-1/2}]$$

can be dropped from the normalization of M_n. Finally,

$$\alpha(t_n) = 1 - O(\gamma_n^2)$$

so $\alpha(t_n)$ can be dropped from the normalization. □

3. **The bias term.** We now consider the case $k/[n^{3/2}\log n] \to \infty$, when the bias term in (2.1) dominates. Assumption (2.3) is back in force. Recall β from (2.12). Note that $c = \mu_3/[6\sigma^3] \neq 0$; without real loss of generality, suppose $c > 0$.

The graph of β is sketched in FIGURE 1 below. This function is anti-symmetric about 0, where it vanishes. Likewise, it vanishes at $\pm\infty$. It has four critical points, at the roots of $x^4 - 6x^2 + 3 = 0$. The global max occurs at

$$t^* = -(3 - \sqrt{6})^{1/2} \doteq -0.74 . \tag{3.1}$$

The second derivative of β does not vanish at any of the four critical points.

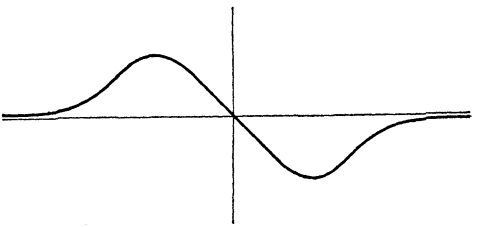

FIGURE 1. The graph of β

(3.2) PROPOSITION. *Suppose* (1.1-2), *and* $k \gg n^{3/2}\log n$. *Suppose* $\mu_3 > 0$, *the case* $\mu_3 < 0$ *being symmetric. Then*

$$\sqrt{n}\,\max_j (N_{nj} - k\tilde{p}_{nj})/\beta(t^*)\ell^2 \longrightarrow \sigma^{1/2}(2\pi)^{1/4}$$

in probability. Furthermore, for any δ *positive, with probability approaching one, the* max *is taken on only for* j's *with*

$$|t_{nj} - t^*| < \delta .$$

Proof. Refer back to (2.1). As Theorem (1.8) of [2] demonstrates, the sampling error term in (2.1) is of order $\ell \cdot (\log n)^{1/2}$. On the other hand, the Edgeworth expansion (2.4) shows the bias term to be of order

$$n^{-1/2}\ell^2\sigma^{1/2}(2\pi)^{1/4}\beta(t_{nj}) + O(n^{-1}\ell^2) . \tag{3.3}$$

If t_{nj} is close to t^*, then $\beta(t_{nj})$ is close to $\beta^* > 0$, and $n^{-1/2}\ell^2$ dominates the sampling-error order $\ell \cdot (\log n)^{1/2}$. Plainly, $n^{-1/2}\ell^2$ also dominates the remainder $n^{-1}\ell^2$. If on the other hand t_{nj} is bounded away

from t^*, then $\beta(t_{nj})$ is bounded below by $\beta(t^*)$, so (3.3) is too small to influence the max. □

Suppose

$$\mu_3 > 0 \quad \text{and} \quad n^{3/2} \log n \ll k \ll n^{5/2}/\log n \,. \tag{3.4}$$

Then the asymptotic distribution of $\max_j (N_{nj} - k\tilde{p}_{nj})$ can still be deduced from [4, (1.24)], as we show next. The case $\mu_3 < 0$ is symmetric, but $\mu_3 = 0$ is different. If k is of order $n^{5/2}/\log n$ or larger, more terms in the Edgeworth expansion (2.4) for p_{nj} become relevant. If k is of order $n^{7/2}$ or larger, the assymptotic distribution becomes degenerate. We do not pursue these issues here: See §3 of [1] for a related discussion.

Before stating Proposition (3.19) formally, we indicate the heuristics. Proposition (3.2) suggests that

$$\max_j (N_{nj} - k\tilde{p}_{nj})/\mathscr{L}$$

is essentially

$$\sigma^{1/2}(2\pi)^{1/4}\beta(t^*)\mathscr{L}/\sqrt{n} = \gamma_n\beta(t^*)[2 \log (\sigma\sqrt{n})]^{1/2} \,.$$

We consider the difference. The idea is to use [4, (1.24)] again, but on a new scale and with new functions α and β. The starting point is (2.6). In particular,

$$\mathscr{L}^{-1}(N_{nj} - k\tilde{p}_{nj}) - \gamma_n\beta(t^*)[2 \log (\sigma\sqrt{n})]^{1/2} = \alpha_{nj}Z_{nj} + \gamma_{nj} \tag{3.5}$$

where

$$\gamma_{nj} = [\beta_{nj} - \gamma_n\beta(t^*)][2 \log (\sigma\sqrt{n})]^{1/2} \,. \tag{3.6}$$

Now β is locally quadratic at t^*, and in effect we expand γ_{nj} around t^*. Informally, by (2.9-11),

$$\beta_{nj} \doteq \gamma_n\beta(t_{nj})$$

so

$$\begin{aligned}
\gamma_{nj} &\doteq [\beta(t_{nj}) - \beta(t^*)] \cdot \gamma_n \cdot [2 \log (\sigma\sqrt{n})]^{1/2} \\
&\doteq \frac{1}{2}\beta''(t^*)(t_{nj} - t^*)^2 \cdot k^{1/2}/n^{3/4} \\
&= \frac{1}{2}\beta''(t^*)\sigma^{-2}(j - n\mu - \sigma\sqrt{n}\, t^*)^2 \cdot k^{1/2}/n^{7/4} \,.
\end{aligned}$$

Parenthetically, this heuristic can be made rigorous if $k \gg n^{3/2}(\log n)^5$, but is a bit too aggresive with smaller k's.

The center c_n called for in [4, (1.23)] is now defined as follows:

$$c_n = n\mu + \sigma\sqrt{n}\, t^* . \tag{3.7}$$

The scale factor ε_n should satisfy

$$\varepsilon_n^2 \sqrt{2 \log \frac{1}{\varepsilon_n}} \doteq k^{1/2}/n^{7/4} .$$

We set

$$m = n^{7/8}/k^{1/4} \quad \text{and} \quad \varepsilon_n = m^{-1}(2 \log m)^{-1/4} . \tag{3.8}$$

Thus,

$$\gamma_{nj} \doteq \frac{1}{2}\beta''(t^*)\sigma^{-2}\theta_{nj}^2 \sqrt{2 \log \frac{1}{\varepsilon_n}}$$

where we write

$$\theta_{nj} = \varepsilon_n(j - c_n) . \tag{3.9}$$

This is to avoid confusion with the $t_{nj} = (j - n\mu)/(\sigma\sqrt{n})$ used earlier. More formally, to make contact with [4] we view k and hence m as functions of n. We make the definitions (3.6–3.9). Let I be a long (finite) closed interval with 0 as an interior point. Let

$$\tilde{\beta}_{nj} = \gamma_{nj} \cdot \left[2 \log \frac{1}{\varepsilon_n}\right]^{-1/2} . \tag{3.10}$$

We propose to study the maximum of

$$\begin{aligned} &\sigma^{-1}(N_{nj} - k\tilde{p}_{nj}) - \gamma_n \beta(t^*)[2 \log (\sigma\sqrt{n})]^{1/2} \\ &= \alpha_{nj} Z_{nj} + \tilde{\beta}_{nj} \cdot \left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2} \end{aligned} \tag{3.11}$$

over j's with $\theta_{nj} \in I$, using [4]. For the function α_n of [4, (1.4)] we take

$$\alpha_n(\theta) = \alpha(t^* + \sigma^{-1}\delta_n\theta) \tag{3.12}$$

where α was defined in (2.13) and

$$\delta_n^4 = k^{-1}n^{3/2}(2 \log m) . \tag{3.13}$$

Thus, α_n is defined over the whole real line. Clearly,

$$t_{nj} = t^* + \sigma^{-1}\delta_n\theta_{nj} . \tag{3.14}$$

So $\alpha_n(\theta_{nj}) = \alpha(t_{nj})$. Note, however, that t_{nj} is centered at t^* while θ_{nj} is centered at 0. Further, $\delta_n \to 0$ because $k \gg n^{3/2} \log n$ by (3.4).

Thus, $t_{nj} \to 0$ uniformly over j with $\theta_{nj} \in I$.

For the function β_n in [4, (1.5)] we take

$$\beta_n(\theta) = \left(\log m / \log \frac{1}{\varepsilon_n}\right)^{1/2} \cdot \delta_n^{-2}[\beta(t^* + \sigma^{-1}\delta_n\theta) - \beta(t^*)] \tag{3.15}$$

where β was defined in (2.12).

Before proceeding, it is helpful to note

$$\frac{2}{8}\log n < \log m < \frac{7}{8}\log n \quad \text{and} \quad \log \frac{1}{\varepsilon_n} \approx \log m \tag{3.16}$$

by the growth condition (3.4). We claim

$$\alpha_{nj} = \alpha_n(\theta_{nj}) + o(1/\log m) \tag{3.17}$$

as $n \to \infty$, uniformly in j with t_{nj} confined to a compact interval. This is routine to verify, estimating α_{nj} from (2.8) and the Edgeworth expansion (2.4), the $\tilde{p}_{nj}$ having been defined in (1.5). More explicitly,

$$\begin{aligned} \alpha_{nj}^2 &= \sigma\sqrt{2\pi n}\ p_{nj} \\ &= \alpha(t_{nj})^2 + O(1/\sqrt{n})\ . \end{aligned}$$

Now $\alpha(t_{nj})$ cannot run away to zero, and

$$\begin{aligned} \alpha_{nj} &= \alpha(t_{nj}) + O(1/\sqrt{n}) \\ &= \alpha_n(\theta_{nj}) + o(1/\log m)\ , \end{aligned}$$

because $\sqrt{n} \gg \log n \geqq 8/7 \log m$ by (3.16). This completes the proof of (3.17). If θ_{nj} is confined to I, then $t_{nj} = t^* + O(\delta_n)$, and $\delta_n \to 0$ because $k \gg n^{3/2}\log n$. So (3.17) establishes condition [4, (1.4)].

Condition [4, (1.5)] can be verified by tedious algebra: this is where the growth condition $k \ll n^{5/2}/\log n$ comes in. We need a little more:

$$\begin{aligned} &\tilde{\beta}_{nj} = \beta_n(\theta_{nj}) + o(1/\log m)\ , \\ &\text{as} \quad n \longrightarrow \infty\ , \quad \text{uniformly in } j\ . \end{aligned} \tag{3.18}$$

Indeed,

$$\begin{aligned} \tilde{\beta}_{nj} &= \gamma_{nj} \cdot \left[2\log \frac{1}{\varepsilon_n}\right]^{-1/2} && \text{by (3.8)} \\ &= \left[\log(\sigma\sqrt{n})/\log \frac{1}{\varepsilon_n}\right]^{1/2} \cdot [\beta_{nj} - \gamma_n\beta(t^*)] && \text{by (3.6)} \\ &= \gamma_n \cdot \left[\log \sigma\sqrt{n}/\log \frac{1}{\varepsilon_n}\right]^{1/2} \cdot [\beta(t_{nj}) - \beta(t^*) + O(1/\sqrt{n})] && \text{by (2.10)} \\ &= n^{-3/4}k^{1/2}\left[2\log \frac{1}{\varepsilon_n}\right]^{-1/2} \cdot [\beta(t_{nj}) - \beta(t^*) + O(1/\sqrt{n})] && \text{by (2.11)} \end{aligned}$$

$$= n^{-3/4}k^{1/2}\left[2\log\frac{1}{\varepsilon_n}\right]^{-1/2}\cdot[\beta(t_{nj}) - \beta(t^*)] + o(1/\log m)$$

because $k \ll n^{5/2}/\log n$ and $\log m \sim \log n$

$$= n^{-3/4}k^{1/2}\delta_n^2\left[2\log\frac{1}{\varepsilon_n}\right]^{-1/2}\delta_n^{-2}[\beta(t_{nj}) - \beta(t^*)] + o(1/\log n)$$

$$= \left[\log m/\log\frac{1}{\varepsilon_n}\right]^{1/2}\delta_n^{-2}[\beta(t_{nj}) - \beta(t^*)] + o(1/\log m) \qquad \text{by (2.31)}$$

$$= \beta_n(\theta_{nj}) + o(1/\log m) \qquad \text{by (2.32–2.33)}.$$

This completes the argument for (2.36), i.e., condition [4, (1.5)].

Condition [4, (1.6)] is clear. For [4, (1.7)], let $\alpha_\infty(\theta) = \alpha(t^*)$ for all θ, so $\alpha_n \to \alpha_\infty$ because α is continuous and $\delta_n \to 0$. Likewise, let

$$\beta_\infty(\theta) = \frac{1}{2}\beta''(t^*)\sigma^{-2}\theta^2 .$$

That $\beta_n \to \beta_\infty$ can be verified by expanding β in a Taylor series around t^*, the location of its maximum; $\beta'(t^*) = 0$ and $\beta''(t^*) < 0$. Condition [3, (1.8)] is clear, at least for large n. We write θ_n for the location of the maximum of $\alpha_n + \beta_n$, and note that $\theta_\infty = 0$. By calculus, $\theta_n = 0(\delta_n)$, and $\delta_n \to 0$. The remaining conditions for [4, (1.24)] are all verified easily. In conformity with [4, (1.15)], let

$$\tilde{\rho}^2 = -[\alpha_\infty''(0) + \beta_\infty''(0)]/\alpha_\infty(0)$$
$$= -\beta''(t^*)/[\alpha(t^*)\sigma^2] .$$

(3.19) PROPOSITION. *Assume* (1.1–2) *and* (3.4). *In the notation given above, with probability approaching one,* $M_n = \max_j(N_{nj} - k\tilde{p}_{nj})$ *is assumed at a unique index* L_n. *Furthermore, the chance that*

$$\tilde{\rho}\sqrt{2\log\frac{1}{\varepsilon_n}}[\varepsilon_n(L_n - c_n) - \theta_n] < y \tag{3.20}$$

and

$$\begin{aligned} &\sqrt{k^{-1}}M_n - \gamma_n\beta(t^*)[2\log(\sigma\sqrt{n})]^{1/2} \\ &\quad < \alpha_n(\theta_n)\left[2\log\frac{1}{\varepsilon_n} - 2\log\log\frac{1}{\varepsilon_n} + x\right]^{1/2} + \beta_n(\theta_n)\left[2\log\frac{1}{\varepsilon_n}\right]^{1/2} \end{aligned} \tag{3.21}$$

converges to

$$\Phi(y)\cdot\exp\left\{-\frac{1}{2\tilde{\rho}}e^{-1/2x}\right\} .$$

Proof. If the max is taken only over j witn $\theta_{nj} \in I$, the con-

clusion is immediate from [4, (1.24)]. The right side of (3.21) is essentially

$$[\alpha_n(\theta_n) + \beta_n(\theta_n)] \cdot \left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2}$$

and $\alpha_n(\theta_n) + \beta_n(\theta_n) > \alpha_n(0) + \beta_n(0) = \alpha(t^*) > 0$.

Consider the j's with

$$\theta_{nj} \in I \quad \text{but} \quad |t_{nj} - t^*| < \delta \,. \tag{3.22}$$

We have to argue that such j's do not matter, i.e., the max over such j's is of smaller order than $\alpha(t^*) \cdot [2 \log 1/\varepsilon_n]^{1/2}$. Apply [4, (3.1)] to the Z_{nj}, but use the original scaling, i.e., take the ε_n in [4, (3.1)] to be $1/(\sigma\sqrt{n})$. With overwhelming probability

$$\begin{aligned} \max_j Z_{nj} &< 2[2 \log (\sigma\sqrt{n})]^{1/2} \\ &< 4\left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2} \end{aligned}$$

by (3.16). Hence

$$\begin{aligned} &\max_j \left\{\alpha_{nj} Z_{nj} + \tilde{\beta}_{nj}\left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2}\right\} \\ &\quad \leqq \left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2} \cdot [4 \max_j \alpha_{nj} + \max_j \tilde{\beta}_{nj}] \\ &\quad \leqq \left[2 \log \frac{1}{\varepsilon_n}\right]^{1/2} \cdot [4 \max_j \alpha_n(\theta_{nj}) + \max_j \beta_n(\theta_{nj}) + o(1/\log m)] \end{aligned}$$

by (3.17–3.18), where the max is taken over the j's satisfying (3.12). Now $\alpha \leqq 1$ by its definition (2.13), so the definition (2.30) of α_n shows

$$\max_j \alpha_n(\theta_{nj}) \leqq 1 \,.$$

Next, t^* is the location of the global maximum of β and β is locally quadratic at t^*, so for δ small, for $0 < \delta' < \delta$,

$$\begin{aligned} \max_u \{\beta(t^* + u) &: \delta' \leqq u \leqq \delta\} = \beta(t^* + \delta') \\ \max_u \{\beta(t^* - u) &: \delta' \leqq u \leqq \delta\} = \beta(t^* - \delta') \,. \end{aligned}$$

Let $I = [-\theta_0, \theta_0]$. Then, by (3.15),

$$\max_j \beta_n(\theta_{nj}) \leqq \max [\beta_n(\theta_0), \beta_n(-\theta_0)] \,.$$

So

$$\limsup_{n\to\infty} \max_j \beta_n(\theta_{nj}) \leqq \frac{1}{2}\beta''(t^*)\delta^{-2}\theta_0^2 \,.$$

Recall $\beta''(t^*) < 0$; choose θ_0 so large that

$$\lambda = 4 + \frac{1}{2}\beta''(t^*)\sigma^{-2}\theta_0^2 < 0 .$$

With overwhelming probability,

$$\max_j \left\{\alpha_{nj}Z_{nj} + \tilde{\beta}_{nj}\left[2\log\frac{1}{\varepsilon_n}\right]^{1/2}\right\} < \lambda\left[2\log\frac{1}{\varepsilon_n}\right]^{1/2} + o(1)$$

where $\lambda < 0$ and the max is taken over j's satisfying (3.22). Such j's do not matter. Finally, j's with $|t_{nj} - t^*| \geqq \delta$ do not matter, by (3.2). □

Note. $\gamma_n \to \infty$ by its definition (2.11) and the growth condition $k \gg n^{3/2}\log n$, and $\beta(t^*) > 0$, so the term subtracted from $\mathscr{L}^{-1}M_n$ on the left side of (3.21) is of order

$$\gamma_n\beta(t^*)(\log n)^{1/2}$$

with $\gamma_n \to \infty$. The terms on the right side are of order

$$\alpha(t^*)(2\log m)^{1/2}$$

and $\log m \sim \log n$. Thus, the term on the left dominates, in agreement with (3.2).

(3.23) COROLLARY. *Suppose* (3.4), *and in addition* $k \gg n^{3/2}(\log n)^3$. *Then the scaling in* (3.19) *can be simplified: the chance that*

$$\tilde{\rho}(2\log m)^{1/4}m^{-1}(L_n - n\mu - \sigma\sqrt{n}\,t^*) < y$$

and

$$\mathscr{L}^{-1}M_n - \gamma_n\beta(t^*)[2\log(\sigma\sqrt{n})]^{1/2} < \alpha(t^*)\left[2\log m - \frac{3}{2}\log\log m + \frac{1}{4}\log 2 + x\right]^{1/2}$$

converges to

$$\Phi(y)\cdot\exp\left\{-\frac{1}{2\tilde{\rho}}e^{-1/2x}\right\} .$$

Proof. Recall that θ_n is the location of the maximum of $\alpha_n + \beta_n$, so $\theta_n = O(\delta_n)$ as defined in (3.13). With our new condition on k, we have $\delta_n = o[1/(\log m)^{1/2}]$. So θ_n can be dropped in (3.20). Likewise, in (3.21),

$$\alpha_n(\theta_n) = \alpha(t^*) + o[1/(\log m)^{1/2}]$$

and

$$\beta_n(\theta_n) = o[1/(\log m)] \,. \qquad \square$$

REFERENCES

1. P. Diaconis and D. Freedman, *On the mode of an empirical histogram*, Pacific J. Math., **9** (1982).
2. ———, *On the difference between the empirical and expected histograms for sums*, Pacific J. Math., **100** (1982a), 287-327.
3. D. Freedman, *A central limit theorem for empirical histograms*, Z. Wahrscheinlichkeitstheorie Verw. Gebiete, **41** (1977), 1-11.
4. ———, *On the maximum of scaled multinomial variables*, Pacific J. Math., **100** (1982b), 329-358.

Received January 7, 1981. Research of the first author partially supported by NSF Grant MSC-77-16974, and the research of the second author was partially supported by NSF Grant MCS-77-01665.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720

PACIFIC JOURNAL OF MATHEMATICS

EDITORS

DONALD BABBITT (Managing Editor)
University of California
Los Angeles, CA 90024

HUGO ROSSI
University of Utah
Salt Lake City, UT 84112

C. C. MOORE and ARTHUR AGUS
University of California
Berkeley, CA 94720

J. DUGUNDJI
Department of Mathematics
University of Southern California
Los Angeles, CA 90007

R. FINN and J. MILGRAM
Stanford University
Stanford, CA 94305

ASSOCIATE EDITORS

R. ARENS E. F. BECKENBACH B. H. NEUMANN F. WOLF K. YOSHIDA

SUPPORTING INSTITUTIONS

UNIVERSITY OF ARIZONA
UNIVERSITY OF BRITISH COLUMBIA
CALIFORNIA INSTITUTE OF TECHNOLOGY
UNIVERSITY OF CALIFORNIA
MONTANA STATE UNIVERSITY
UNIVERSITY OF NEVADA, RENO
NEW MEXICO STATE UNIVERSITY
OREGON STATE UNIVERSITY
UNIVERSITY OF OREGON
UNIVERSITY OF SOUTHERN CALIFORNIA
STANFORD UNIVERSITY
UNIVERSITY OF AAWAII
UNIVERSITY OF TOKYO
UNIVERSITY OF UTAH
WASHINGTON STATE UNIVERSITY
UNIVERSITY OF WASHINGTON

The Supporting Institutions listed above contribute to the cost of publication of this Journal, but they are not owners or publishers and have no responsibility for its content or policies,

Mathematical parers intended for publication in the *Pacific Journal of Mathematics* should be in typed form or offset-reproduced, (not dittoed), double spaced with large margins. Please do not use built up fractions in the text of the manuscript. However, you may use them in the displayed equations. Underline Greek letters in red, German in green, and script in blue. The first paragraph or two must be capable of being used separately as a synopsis of the entire paper. Please propose a heading for the odd unmbered pages of less than 35 characters. Manuscripts, in triplicate, may be sent to any one of the editors. Please classify according to the scheme of Math. Reviews, Index to Vol. **39**. Supply name and address of author to whom proofs should be sent. All other communications should be addressed to the managing editor, or Elaine Barth, University of California, Los Angeles, California, 90024.

50 reprints to each author are provided free for each article, only if page charges have been substantially paid. Additional copies may be obtained at cost in multiples of 50.

The *Pacific Journal of Mathematics* is issued monthly as of January 1966, Regular subscription rate: $114.00 a year (6 Vol., 12 issues). Special rate: $57.00 a year to individual members of supporting institution.

Subscriptions, orders for numbers issued in the last three calendar years, and changes of address shoud be sent to Pacific Journal of Mathematics, P.O. Box 969, Carmel Valley, CA 93924, U.S.A. Old back numbers obtainable from Kraus Periodicals Co., Route 100, Millwood, NY 10546.

PUBLISHED BY PACIFIC JOURNAL OF MATHEMATICS, A NON-PROFIT CORPORATION
Printed at Kokusai Bunken Insatsusha (International Academic Printing Co., Ltd.).
8-8, 3-chome, Takadanobaba, Shinjuku-ku, Tokyo 160, Japan.

Pacific Journal of Mathematics
Vol. 100, No. 2 October, 1982