Estimation of the Number of Terms in a Sum

by

George P. McCabe, Jr.

Department of Statistics

Division of Mathematical Sciences

Mimeograph Series No. 264

September, 1971

ABSTRACT

Estimation of the Number of Terms in a Sum

Fixed sample size and sequential procedures are investigated for estimating the parameter n from observations on a sequence of i.i.d. random variables each of which is the sum of n i.i.d. random variables with known mean and variance.

Estimation of the Number of Terms in a Sum

by

George P. McCabe, Jr.

Purdue University

1. Introduction. Let $\{X_{ij}: i = 1,\ldots, n; j = 1,\ldots\}$ be a doubly indexed set of i.i.d. random variables with non-zero mean μ and finite variance σ^2. Suppose that one can observe a fixed or variable number of terms of the sequence $(Y_j: j = 1,\ldots)$ where $Y_j = X_{1j} + \ldots + X_{nj}$. Given that μ and σ^2 are known, both fixed sample size and sequential procedures for estimating the unknown parameter n are considered.

Feldman and Fox (1968) studied the problem of estimating the parameter n of a binomial distribution on the basis of a fixed number of observations. This corresponds to the case in which the X_{ij} are Bernoulli random variables with known p. They also investigated estimators of the parameter μ of a normal $N(\mu,\mu)$ distribution and gave references to several related problems.

The case in which the X_{ij} are Poisson with mean one corresponds to the problem of estimating a Poisson parameter when it is assumed to be an integer. McCabe (a) studied this problem and some generalizations to an exponential family (b).

The procedures proposed in this paper are robust in the sense that their properties can be evaluated in terms of the known parameters μ and σ^2 only. In general, if the distribution is completely known, then sharper results can be obtained. As an example, the normal case is treated in detail.

Since each of the Y's can be divided by the known mean μ, we can and do assume without loss of generality that the Y's are i.i.d. with mean n

and variance $n\sigma^2$ where σ^2 is known and n is to be estimated.

2. Fixed sample size estimate. Let $n.i.(z)$ denote the nearest integer to z. For convenience in what follows, we adopt the convention that $n.i.\ (i + 1/2) = i$ if i is an integer and $n.i.\ (z) = 1$ if $y \leq 3/2$. Let k be the size of the sample.

Motivated by the fact that $\overline{Y}_k = (Y_1 + \ldots + Y_k)/k$ is an unbiased estimate of n, we consider the estimate

$$\hat{n} = n.i.\ (\overline{Y}_k). \tag{2.1}$$

If the underlying distribution of the X_{ij} is continuous and symmetric then n will also be unbiased; but in general, this will not be the case.

Let $P(n)$ denote the probability that $\hat{n}$ is unequal to n when n is the true value of the parameter. Then

$$P(n) = P_n((k/n)^{1/2}|\overline{Y}_k - n|/\sigma \geq (k/n)^{1/2}/2\sigma) \leq 4\sigma^2 n/k \tag{2.2}$$

by Chebyshev's inequality. In the normal case,

$$P(n) < (8n\sigma^2/\Pi k)^{1/2} \exp(-k/8n\sigma^2) \quad . \tag{2.3}$$

Following Feldman and Fox (1968) we define an estimate $\hat{\theta}$ to be α-consistent for the parameter θ if for any $\varepsilon > 0$,

$$P_\theta(|\hat{\theta} - \theta|/\theta^\alpha > \varepsilon) \to 0 \quad \text{as} \quad \theta \to \infty$$

Let $\alpha > 1/2$ and $\varepsilon > 0$ be fixed. Since

$$|\hat{n} - n| \geq |\overline{Y}_k - n| - |\hat{n} - \overline{Y}_k| \geq |\overline{Y}_k - n| - 1/2 \quad ,$$

it follows that

$$P_n(|\hat{n} - n|/n^{\alpha} > \varepsilon) \leq P((k/n)^{1/2}|\bar{Y}_k - n|/\sigma > (k/n)^{1/2}(1/2 + \varepsilon n^{\alpha})/\sigma)$$

$$\leq n\sigma^2/k(1/4 + \varepsilon n^{\alpha} + n^{2\alpha}) \to 0$$

as $n \to \infty$. Therefore, for $\alpha > 1/2$, $\hat{n}$ is α-consistent for n.

Since the sample variance $s^2 = (k-1)^{-1}\Sigma(Y_j - \bar{Y}_k)^2$ is an unbiased estimate of $n\sigma^2$, one could consider using n.i. (s^2/σ^2) as an alternative to $\hat{n}$. For large values of n, however, the variance of $\bar{Y}$ will be less than the variance of s^2/σ^2. In the normal case, $\text{var}(s^2/\sigma^2)$ equals $2n^2/(k-1)$ which is greater than $n\sigma^2/k$, the variance of $\bar{Y}$, whenever $n \geq \sigma^2(k-1)/2k$.

It should be noted that even for the normal case, the estimate $\hat{n}$ is not optimal. In fact, it is not even a function of the sufficient statistics. However, if we consider n to be a continuous parameter, the Cramer-Rao bound for the variance of an unbiased estimate is

$$n\sigma^2/k(1 + \sigma^2/2n).$$

When n is large, this expression is approximately the variance of $\bar{Y}_k$, whereas if σ^2 is large, it is close to the variance of s^2/σ^2.

Clearly any reasonable analysis of this problem must specify the relation between the quantities k and $n\sigma^2$. Feldman and Fox (1968) were interested in k small relative to $n\sigma^2$, whereas in the present work, we will focus on the case where k is large relative to $n\sigma^2$. In this context, the problem may be viewed as a simultaneous test of the countable set of hypotheses:

$$\{H_n: \; EY_j = n, \; n = 1,2,\ldots\} \quad .$$

For k large relative to $n\sigma^2$, the discreteness of the parameter space becomes more important and the Cramer-Rao bound does not give an adequate picture of the structure of the problem. Using the results of Chapman and Robbins (1951), we can obtain a more informative bound on the variance of an

unbiased estimate of n for the normal case as follows:

For any unbiased estimate $\tilde{n}$ of n, let $\sigma_n^2(\tilde{n})$ denote its variance. Then

$$\sigma_n^2(\tilde{n}) \geq 1/\inf_h h^{-2} E_n(f_{n+h} - f_n)^2/f_n^2$$

where f_n and E_n denote the density of $Y_1, \ldots, Y_k$ and expectation respectively when n is the true value of the parameter. It is easy to show that

$$\sigma_n^2(\tilde{n}) \geq \left(\left(\frac{n}{(n^2-1)^{1/2}} \exp\left(\frac{1}{(n+1)\sigma^2}\right)\right)^k - 1\right)^{-1} \quad . \tag{2.4}$$

Although the Chapman-Robbins bound is not necessarily attainable, it does indicate the possibility of an estimate having variance which decreases exponentially with the sample size. This property can be demonstrated for $\hat{n}$ in the normal case as follows:

Let n be fixed and let σ_n^2 denote the variance of $\hat{n}$. Then

$$\sigma_n^2 = \sum_{m=-n+1}^{-1} m^2 P_n(\hat{n}=n+m) + \sum_{m=1}^{\infty} m^2 P(\hat{n}=n+m)$$

$$< 2 \sum_{m=1}^{\infty} m^2 P_n(n+m-1/2 < \overline{Y}_k \leq n+m+1/2)$$

$$< 2 \sum_{m=1}^{\infty} \int_{m-1/2}^{m+1/2} (X + 1/2)^2 \, dN(X;\, 0, n\sigma^2/k) \quad ,$$

where

$$dN(X;u,\sigma^2) = (2\Pi\sigma^2)^{-1/2} \exp(-(X-u)^2/2\sigma^2)dX \quad .$$

Combining the integral terms, performing the transformation

$$Y = X(k/4n\sigma^2)^{1/2} \quad ,$$

and letting

$$a = (k/4n\sigma^2)^{1/2}$$

yields

$$(25) \qquad \sigma_n^2 < 2 \int_a^\infty ((n\sigma^2/k)^{1/2} Y + 1/2)^2 \, dN(Y;0,1) \, .$$

Now, for $a > 0$,

$$\int_a^\infty \exp(-Y^2/2)dY < a^{-1} \exp(-a^2/2) \quad ,$$

and

$$\int_a^\infty Y \exp(-Y^2/2)dY = \exp(-a^2/2) \quad .$$

Also, since

$$\int_a^\infty (Y^2-1) \exp(-Y^2/2)dY = a \exp(-a^2/2),$$

it follows that

$$\int_a^\infty Y^2 \exp(-Y^2/2)dY < (a + a^{-1}) \exp(-a^2/2) \quad .$$

Substituting these bounds into (2.5) gives

$$(2.6) \qquad \sigma_n^2 < (8\sigma^2 n/\Pi k)^{1/2} (1+n\sigma^2/k) \exp(-k/8n\sigma^2) \, .$$

3. Sequential Rules. In this section, two classes of rules are investigated. A sequential rule consists of a stopping variable N and a terminal decision function $\hat{n}$. For each rule in the first class, the bound on the error probabilities is calculated without reference to the original distribution except through the known parameter σ^2. Thus, given an arbitrary

preassigned bound ε, one can construct a rule which has the property that the probability of error is less than or equal to ε for all n and for all possible distributions of the X_{ij} having the same value of σ^2. The second class of rules exploits the properties of the normal distribution and indicates the type of results that can be obtained if the underlying distribution is known.

Class I:

Let $\varepsilon > 0$ be given and define

$$K = 2[16\sigma^2/3\varepsilon] + 2$$

where $[\cdot]$ denotes the greatest integer function. Define

$$N_1 = \inf \{k \geq K: k \geq K \bar{Y}_k\} \tag{3.1}$$

and

$$\hat{n}_1 = \text{n.i.} \ (N_1 K^{-1}) \ . \tag{3.2}$$

Theorem 3.1. For N_1 and $\hat{n}_1$ defined above,

$$E_n N_1 < \infty \quad \text{for all } n \tag{3.3 (a)}$$

and

$$\sup_n P_n(\hat{n}_1 \neq n) \leq \varepsilon \ . \tag{3.4 (b)}$$

Proof. Let n be fixed. Clearly, for $k > Kn$,

$$P_n(N_1 > k) \leq P_n(k < K \bar{Y}_k)$$
$$= P_n((k/n\sigma^2K^2)^{1/2}(k-Kn) < (k/n\sigma^2)^{1/2}(\bar{Y}_k - n)) \ .$$

So,

$$P_n(N_1 > k) \leq n\sigma^2K^2/k(k-Kn)^2 \tag{3.5}$$

by Chebyshev's inequality. Now,

$$E_nN_1 = \sum_{k=0}^{\infty} P_n(N_1 > k)$$

$$\leq Kn + 1 + \sum_{k>Kn}^{\infty} P_n(N_1 > k)$$

$$\leq Kn + 1 + n\sigma^2K^2 \sum_{k>Kn}^{\infty} k^{-1}(k-Kn)^{-2}$$

by (3.5). Since the sum is convergent, part (a) follows.

To calculate the error bound, we first observe that

$$P_n(\hat{n}_1 \neq n) = P_n(N_1K^{-1} \leq n - \tfrac{1}{2}) + P_n(N_1K^{-1} > n + \tfrac{1}{2}) \quad . \tag{3.6}$$

Now,

$$P_n(N_1K^{-1} \leq n - \tfrac{1}{2}) = P_n(k \geq K\bar{Y}_k \text{ for some } k = K,\ldots, K(n-\tfrac{1}{2}))$$

$$= P_n(Z_1 + \ldots + Z_k \geq (nkK-k^2)/K \text{ for some } k=K,\ldots,K(n-\tfrac{1}{2})).$$

where

$$Z_i = n - Y_i \quad .$$

Note that K and $K(n-\frac{1}{2})$ are integers by the definition of K. Clearly the Z_i are i.i.d. with mean zero and variance $n\sigma^2$. Since

$$\min_{K\leq k\leq K(n-\frac{1}{2})} \{K^{-1}(nkK-k^2)\} = K(n-\tfrac{1}{2})/2 \quad ,$$

it follows that

$$P_n(N_1K^{-1} < n - \tfrac{1}{2}) \leq P_n(Z_1+\ldots+Z_k \geq K(n-\tfrac{1}{2})/2 \text{ for some } k=K,\ldots,K(n-\tfrac{1}{2})).$$

Now, applying the Kolmogorov inequality, we obtain

$$(3.7) \qquad P_n(N_1K^{-1} < n - \tfrac{1}{2}) \leq 4n\sigma^2/K(n-\tfrac{1}{2}) \quad \text{for all} \quad n .$$

On the other hand,

$$P_n(N_1K^{-1} > n + \tfrac{1}{2}) \leq P_n(\overline{Y}_m < n + \tfrac{1}{2})$$

where

$$m = K(n + \tfrac{1}{2}) \quad .$$

This can be bounded by the Chebyshev inequality to give

$$(3.8) \qquad P_n(N_1K^{-1} > n + \tfrac{1}{2}) \leq 4\sigma^2 n/K(n + \tfrac{1}{2}) \quad \text{for all} \quad n.$$

Substituting (3.7) and (3.8) into (3.6) yields

$$\begin{aligned} P_n(\hat{n}_1 \neq n) &\leq 8\sigma^2n^2/K(n^2-1/4) \\ &\leq 32\sigma^2/3K \\ &\leq \varepsilon \quad \text{for all} \quad n. \end{aligned}$$

Obviously the rules in this first class are rather crude and the methods used are most elementary in nature. Nonetheless, this class constitutes a solution to the problem of estimating n with a uniformly small bound on the probability of error.

The rules in this class can be easily modified so that N can only take on the values $K(j + \frac{1}{2})$, $j = 1,2,\ldots$. This would eliminate the necessity of calculating $\overline{Y}$ at each stage and the observations could be taken in groups of size K.

Class II: We now assume that the Y_j are normal with mean n and variance $n\sigma^2$. For any $\varepsilon > 0$, let

$$A = \log(1 + 2/\varepsilon).$$

Define

$$N_2 = \inf\{k \geq 1: |\bar{Y}_k - n| \leq \frac{1}{2} - A(n+1)\sigma^2/k \text{ for some } n\} \tag{3.9}$$

and let

$$\hat{n}_2 = \text{n.i.}\ (\bar{Y}_N) \ . \tag{3.10}$$

Let

$$k_n = 2A(n+1)\sigma^2 \ .$$

Theorem 3.2. If the Y_j are i.i.d. normal with mean n and variance $n\sigma^2$, then for N_2 and $\hat{n}_2$ defined above,

$$\text{(a)} \quad P_n(N_2 < \infty) = 1 \tag{3.11}$$

and

$$\text{(b)} \quad \sup_n P_n(\hat{n}_2 \neq n) \leq \varepsilon \ . \tag{3.12}$$

Proof. Let n be fixed and let $k = r\,k_n$ where $r > 1$. Then,

$$\begin{aligned} P_n(N_2 > k) &\leq P_n(|\bar{Y}_k - n| > \tfrac{1}{2}(\tfrac{r-1}{r})) \\ &= P_n((k/n\sigma^2)^{1/2}|\bar{Y}_k - n| > \tfrac{1}{2}(\tfrac{r-1}{r})(k/n\sigma^2)^{1/2}) \\ &\leq 4r\, n\sigma^2/k_n(r-1)^2 \ . \end{aligned}$$

Taking limits as k (or equivalently r) goes to infinity gives (a).

Let $A_{k,m}$ be the set of values of $Y_1, \ldots, Y_k$ such that $N_2 = k$ and $\hat{n}_2 = m$. Note $\hat{n}_2 = m$ implies $k \geq Km$. Now,

$$(3.13)\qquad P_n(\hat{n}_2 \neq n) = \sum_{m\neq n} P_n(\hat{n}_2 = m)$$

$$= \sum_{m\neq n} \sum_{k\geq k_m} \int_{A_{k,m}} f_n(Y_1,\ldots,Y_k)\,dY_1\ldots dY_k$$

$$= \sum_{m\neq n} \sum_{k\geq k_m} \int_{A_{k,m}} (2\Pi n\sigma^2)^{-1/2} \exp\Big(-\frac{1}{2n\sigma^2}\sum_{i=1}^{k}(Y_i-n)^2\Big)\,dY_1\ldots dY_k$$

Observe that

$$\sum_{i=1}^{k}(Y_i-n)^2 - \sum_{i=1}^{k}(Y_i-m)^2 = k(n-m)(n+m-2\overline{Y}_k) \quad .$$

Now, for $n > m$,

$$\overline{Y}_k \leq m + \frac{1}{2} - A(m+1)\sigma^2/k \qquad \text{on} \qquad A_{k,m} \quad .$$

Hence,

$$\sum_{i=1}^{k}(Y_i-n)^2 - \sum_{i=1}^{k}(Y_i-m)^2 \geq k(n-m)(n-m-1+k_m/k)$$

$$= (n-m)(k(n-m-1) + k_m) \quad .$$

But since $k \geq k_m$ on $A_{k,m}$, it follows that

$$\sum_{i=1}^{k}(Y_i-n)^2 - \sum_{i=1}^{k}(Y_i-m)^2 \geq k_m(n-m)^2 \quad .$$

In a similar manner, it can be shown that

$$\sum_{i=1}^{k}(Y_i-n)^2 - \sum_{i=1}^{k}(Y_i-m)^2 \geq k_m(n-m)^2$$

on $A_{k,m}$ for $n < m$.

Therefore, for any $m \neq n$,

$$(3.14)\qquad \exp\Big(-\Big(\sum_{i=1}^{k}(Y_i-n)^2-\sum_{i=1}^{k}(Y_i-m)^2\Big)/2n\sigma^2\Big)$$

$$\leq \exp(-k_m(n-m)^2/2n\sigma^2)$$

$$= \exp(-A(n-m)^2(m+1)/n)$$

on the set $A_{k,m}$.

Using (3.13) and (3.14), we obtain

$$P_n(\hat{n}_2\neq n) \leq \sum_{m\neq n}\exp\Big(-\frac{A(n-m)^2(m+1)}{n}\Big)\sum_{k\geq k_m}\int_{A_{k,m}}(2\Pi n\sigma^2)^{-\frac{1}{2}}\exp\Big(\frac{-\sum_{i=1}^{k}(Y_i-m)^2}{2n\sigma^2}\Big)dY_1\ldots dY_k \ .$$

Now, since the function to be integrated is a density for each m, it follows that

$$\sum_{k\geq k_m}\int_{A_{k,m}}(2\Pi n\sigma^2)^{-\frac{1}{2}}\exp\Big(\frac{-\sum_{i=1}^{k}(Y_i-n)^2}{2n\sigma^2}\Big)\,dY_1\ldots dY_k \leq 1 \ .$$

Hence,

$$P_n(\hat{n}_2\neq n) \leq \sum_{m\neq n}\exp(-A(n-m)^2(m+1)/n) \ .$$

Since

$$(n-m)(m+1)/n \geq 1 \qquad \text{for} \qquad m \leq n-1$$

and

$$(m-n)(m+1)/n \geq 1 \qquad \text{for} \qquad m \geq n+1 \ ,$$

it follows that

$$P_n(\hat{n}_2\neq n) \leq \sum_{m\neq n}\exp(-A|n-m|)$$

$$< 2/(e^A-1)$$

$$\leq \varepsilon \ .$$

Although the assumption of normality was used to calculate the error bounds, the stopping rule can be investigated in greater generality.

Theorem 3.3. If the Y_j are i.i.d. with mean n and variance $n\sigma^2$, then for any $\delta > 0$

$$\limsup_{A \to \infty} P_n(N_2 k_n^{-1} \leq 1 + \delta) = 1 \tag{3.15}$$

where

$$k_n = 2A(n+1)\sigma^2 \quad .$$

Proof. Let $\delta > 0$ be fixed and let $k = k_n(1+\delta)$. For convenience, it will be assumed that k is an integer. Now,

$$\begin{aligned} P_n(N_2 > k) &\leq P((\sigma^2 n/k)^{-1/2} |\overline{Y}_k - n| > (k/n)^{-1/2} \delta/2\sigma(1+\delta)) \\ &\leq 4n\sigma^2(1+\delta)^2/k\delta^2 \\ &= 2n(1+\delta)/A(n+1)\delta^2 \quad . \end{aligned}$$

Taking limits as $A \to \infty$ gives the desired result.

With the imposition of moment conditions which are trivially satisfied in the normal case, stronger results can be obtained.

Let $c_n = 2\sigma^2(n+1)$. Then $k_n = c_n A$.

Lemma. If $EY_j^4 < \infty$ (or equivalently, if $EX_{ij}^4 < \infty$) then for any $c' \geq c > c_n$, there exists a constant M, which depends on c and n but not on c' or A, such that

$$P_n(N_2 > i) \leq Mi^{-2} \quad , \tag{3.16}$$

where

$$i = c'A \text{ is an integer.}$$

Proof. First note that

$$P_n(N_2 > i) \leq P_n(|\overline{Y}_i - n| > 1/2 - A(n+1)\sigma^2/i)$$

$$\leq P_n(|\overline{Y}_i - n| > B)$$

where

$$B = 1/2 - (n+1)\sigma^2/k \quad .$$

Furthermore,

$$P(N_2 > i) \leq B^{-4} E_n(\overline{Y}_i - n)^4 \tag{3.17}$$

by the Markov inequality. Let Y be a random variable with the same distribution as each of the Y_j. Then,

$$E_n(\overline{Y}_i - n)^4 = i^{-4}(i(E_n(Y-n)^4 - 3\sigma^4) + 3i^2\sigma^4)$$

$$\leq i^{-2} \max(3\sigma^4, E_n(Y-n)^4) \quad .$$

By assumption, $E_n(Y-n)^4 < \infty$. Therefore, letting

$$M = B^{-4} \max(3\sigma^4, E_n(Y-n)^4)$$

and substituting into (3.17) gives the desired result.

Theorem 3.3. If $EY_j^4 < \infty$ then

$$\limsup_{A \to \infty} k_n^{-1} E_n(N_2) = 1 \tag{3.18}$$

Proof. Let n be fixed and let $c > c_n$. Then,

$$E_n N_2 = \sum_{i=0}^{\infty} P_n(N_2 > i)$$

$$\leq cA + \sum_{i \geq cA} P_n(N_2 > i) \quad .$$

Moreover,

$$E_n N_2 \leq cA + \sum_{i \geq cA} Mi^{-2}$$

by the previous lemma. Obviously, the above sum is convergent and hence goes to zero as $A \to \infty$. Since c was arbitrary subject only to $c > c_n$, (3.18) follows.

Although a small uniform bound for the error probabilities cannot be constructed for fixed sample size rules, a rough comparison of the fixed and sequential schemes in the normal case can be made.

Let n be fixed. For k large, the error probability for the fixed sample size procedure (2.1) is approximately

$$(3.19) \qquad (8n/\Pi k)^{\frac{1}{2}} \sigma \exp(-k/8n\sigma^2) \quad .$$

For A large, the sequential procedure requires approximately $k_n = 2A(n+1)\sigma^2$ observations and the error bound is

$$(3.20) \qquad 2/(e^A - 1) \quad .$$

Equating (3.19) and (3.20) and assuming that A and k are large, we find that k is approximately $8An\sigma^2$. Thus, for n large, the sequential plan requires on the average only 1/4 as many observations as the fixed sample size procedure.

REFERENCES

[1] Chapman, D. G. and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. Ann. Math. Statist. 22, 581-586.

[2] Feldman, D. and Fox, M. (1968). Estimation of the parameter n in the binomial distribution. J. Amer. Statist. Assoc. 63, 150-158.

[3] McCabe, G. (a) Sequential estimation of a Poisson integer mean. (to be published in Ann. Math. Statist.).

[4] McCabe, G. (b) Sequential estimation of a restricted mean parameter of an exponential family. (submitted for publication).