STATISTICAL RESEARCH REPORT No 4
Institute of Mathematics 1973
University of Oslo

UNIFORMLY MINIMUM VARIANCE (UMVU) ESTIMATORS BASED ON SAMPLES FROM A RIGHT TRUNCATED AND RIGHT ACCUMULATED GEOMETRIC DISTRIBUTION

by

Erik N. Torgersen

ABSTRACT

UNIFORMLY MINIMUM VARIANCE (UMVU) ESTIMATORS BASED ON SAMPLES FROM A RIGHT TRUNCATED AND RIGHT ACCUMULATED GEOMETRIC DISTRIBUTION

by

Erik N. Torgersen

University of Oslo

Consider a sample, $X_1,\ldots,X_n$, of size n from the distribution:

$$P(X = x) = p^x q \; ; \; x = 0,1,\ldots,K-1$$

and

$$P(X = K) = p^K$$

where $p = 1-q$ is an unknown parameter.

It is shown that, provided $n \geqq 2$ and $K \geqq 2$, this experiment does not admit a complete and sufficient statistic. We provide answers to the following problems:

Which functions of $(X_1,\ldots,X_n)$ minimizes at a given value of p , the variance among all unbiased estimators of their expectations?

Which functions of $(X_1,\ldots,X_n)$ are UMVU estimators of their expectations?

Which functions of p have UMVU estimators?

Suppose a function of p does have an UMVU estimator. How do we find it?

How must n be chosen so that a given function of p has an UMVU estimator based on n observations?

Introduction

Consider a sample, $X_1, \dots, X_n$, of size n from the distribution:

$$P(X = x) = p^x q \; ; \; x = 0, 1, \dots, K-1$$

and

$$P(X = K) = p^K$$

where $p = 1-q$ is an unknown parameter.

It is shown that, provided $n \geqq 2$ and $K \geqq 2$, this experiment does not admit a complete and sufficient statistic. We provide answers to the following problems:

Which functions of $(X_1, \dots, X_n)$ minimizes at a given value of p , the variance among all unbiased estimators of their expectations?

Which functions of $(X_1, \dots, X_n)$ are UMVU estimators of their expectations?

Which functions of p have UMVU estimators?

Suppose a function of p does have an UMVU a estimator. How do we find it?

How must n be chosen so that a given function of p has an UMVU estimator based on n observations?

The basic results on UMVU estimation in Lehmann and Scheffé [1950 ; Completeness, similar regions and unbiased estimation. Sankhyā 10, 305-340] and in Bahadur [1957; On unbiased estimates of uniformly minimum variance. Sankhyā 18, 211-224] are used constantly and without explicit references.

Uniformly minimum variance unbiased estimators based on samples from a right truncated and right accumulated geometric distribution.

Consider a factory producing items of a certain kind. The items are produced, one after another, on a machine. The output, from this machine is, provided failure does not occurs, K items a day. Failure, however, implies repair which makes further production impossible on the day of occurence. Repair will always be completed before the next workday, i.e. we assume that the machine is in working condition at the start of every workday. Finally it will be assumed that the conditional probability of failure during the production of the x-th item given that no previous failure has occured is $q = 1-p$, $x = 1,2,\dots,K$.

Let X denote the total number of items produced on a given day. Then - by the above assumptions -

$$P(X = x) = p^x q; \quad x = 0,1,\dots,K-1$$

and

$$P(X = K) = p^K$$

We will assume below that $p \in]0,1[$ is totally unknown. This assumption is, however, not at all essential - any infinite subset of $[0,1]$ may replace $]0,1[$ as parameter set in most of the following considerations.

Let us first consider the experiment $\mathcal{E}$ obtained by observing the number X of items produced on a single day. Clearly our experiment is a sub experiment of the complete experiment obtained by observing a geometrically distributed random variable Y such that $P(Y = y) = p^y q$; $y = 0,1,2,\ldots$. It follows that $\mathcal{E}$ is complete and it is easily seen that a function g of p has an unbiased (and hence an UMVU) estimator if and only if g is a polynomial of degree $\leqq K$. (Note that X is binomial when $K = 1$.)

It will be convenient to write the distribution of X as:

$$P(X = x) = p^K q^{a(x)} p^{t(x) - Ka(x)}; \quad x = 0,1,\ldots,K$$

where $a(x) = 1$ or 0 as $x < K$ or $x = K$ and $t(x) = xa(x)$.

Consider now the outputs $X_1,\ldots,X_n$ on n workdays with the same number K. Assuming independence the experiment obtained is $\mathcal{E}^n$ and the joint distribution of $X_1,\ldots,X_n$ is given by:

$$P(X_1=x_1,\ldots,X_n=x_n) = p^{nK} q^{a_n(x_1,\ldots,x_n)} p^{t_n(x_1,\ldots,x_n) - Ka_n(x_1,\ldots,x_n)}$$

where $a_n(x_1,\ldots,x_n) = a(x_1)+\ldots+a(x_n)$ and $t_n(x_1,\ldots,x_n) = t(x_1)+\ldots+t(x_n)$.

It follows that the joint distribution is exponential and it is easily seen that $A_n = a_n(X)$ and $T_n = b_n(X)$, together, constitutes a minimal sufficient statistic. The yoint distribution of A_n, T_n is given by:

$$P(A_n = a, T_n = t) = \binom{n}{a} C_K(a,t) p^{nK} q^a p^{t-Ka};$$

$$t = 0,1,\ldots,(K-1)a, a = 0,1,\ldots,n$$

where

$$C_K(a,t) = \#\{(x_1,\ldots,x_a) : 0 \leq x_i \leq K-1;\ i = 1,\ldots,a \quad x_1+\ldots+x_a = t\}$$

when $t = 0,1,\ldots,(K-1)a, \ a = 1,2,\ldots,n$

and

$$C_K(0,0) = 1$$

Note that $C_K(a,t) \geqq 1$ whenever it is defined.

In the case "$K = 1$" this reduces to the experiment consisting of n independent binomial trials, each having "success"probability q. In this case A_n is a complete sufficient statistic and we know that a function g of p has an unbiased estimator if and only if it is a polynomial of degree $\leqq n$.

It remains to discuss the case $K \geqq 2$.

It will be shown that, provided $n \geqq 2$, the pair (A_n, T_n) is not complete, and consequently - by minimal sufficiency - that the model does not admit any complete and sufficient statistic.

Let us first determine the set of statistics $\delta(A_n, T_n)$ which minimizes the variance at a given value $p_0 \in]0,1[$ within the class of all unbiased estimators of $E_p\delta(A_n, T_n)$. As a preliminary result we establish:

Proposition G1.

A statistic $\delta(A_n, T_n)$ minimizes the variance at p_0 within the class of all unbiased estimators of $E_p\delta(A_n T_n)$ if and only if any function $\gamma(a,t) : t = 0,1,\ldots,(K-1)a, \ a = 1,\ldots,n$ satisfying the identity:

$$\sum_{t,a} \gamma(a,t) p^{nK} q^a p^{t-Ka} \underset{p}{\equiv} 0$$

also satisfies:

$$\sum_{t,a} \gamma(a,t)\delta(a,t)p_o^{nK} q_o^{a} p_o^{t-Ka} = 0$$

where $q_o = 1-p_o$.

<u>Proof</u>: By the proof of theorem in Lehmann and Stein [ ], $\delta(A_n T_n)$ minimizes the variance at p_o within the class of unbiased estimators of $E_p\delta(A_n T_n)$ if and only if $\delta(A_n, T_n)$ is uncorrelated at p_o with any unbiased estimator of zero - and by some rechristening of constants - this is precisely the criterion above. □

It will be more convenient to rewrite this criterion as:

<u>Proposition G.2.</u>

$\delta(A_n, T_n)$ minimizes the variance at p_o within the class of all unbiased estimators of $E_p\delta(A_n, T_n)$ if and only if any function $\gamma(a,t)$: $t = 0,1,\ldots,(K-1)a$, $a = 0,\ldots,n$ satisfying the identity:

$$\sum_{a,t} \gamma(a,t)z^{(K-1)a-t}(1+z)^a \equiv 0$$

satisfies:

$$\sum_{a,t} \gamma(a,t)\delta(a,t)z_o^{(K-1)a-t}(1+z_o)^a = 0$$

where $z_o = -\dfrac{1}{p_o}$.

Proof: Follows easily by substituting $z = -\frac{1}{p}$ in the identities appearing in proposition G.1. □

For each pair (a,t) where $t = 0,1,\ldots,(K-1)a$ and $a = 0,1,\ldots,n$ put $f_{a,t}(z) = z^{(K-1)a-t}(1+z)^a$. Then:

$$a \leqq \deg f_{a,t}(z) \leqq K_a - t \leqq K_a \leqq K_n$$

Now $(1+z)^a = (1+z)^{a-1}+z(1+z)^{a-1}$; $a \geqq 1$ so that:

$$f_{a,t}(z) = z^{(K-1)a-t}(1+z)^{a-1}+z^{(k-1)a-t+1}(1+z)^{a-1}$$

$$= f_{a-1,t'}(z)+f_{a-1,t'-1}(z)$$

$t' = t-(K-1)$ provided $a \geqq 1$ and $t \geqq K$. (i.e. $a \geqq 2$ and $t \geqq K$.)

The set of all real valued functions $\gamma(a,t)$: $t = 0,1,\ldots,(K-1)a$, $a = 0,\ldots,n$ may be identified with R^N where

$$N = 1+\sum_{a=1}^{n} [(K-1)a+1]$$

$$= n+1+(K-1)\frac{n(n+1)}{2}$$

Let Γ denote the subspace of R^N consisting of all functions $\gamma(a,t)$; $t = 0,1,\ldots,(K-1)a$, $a = 0,\ldots,n$ satisfying the identity:

$$\sum_{a,t} \gamma(a,t)f_{a,t}(z) \underset{z}{\equiv} 0$$

To each point (a,t) such that $a \geqq 1$ and $t \geqq K$ corresponds an element $\xi_{a,t}$ of Γ defined by:

$$\xi_{a,t}(a,t) = 1 = -\xi_{a,t}(a-1,t-K+1) = -\xi_{a,t}(a-1,t-K)$$

$$\xi_{a,t}(a',t') = 0 \quad \text{when} \quad (a',t') \notin \{(a,t),(a-1,t-K+1),(a-1,t-K)\}$$

These functions are obviously linearily independent. Hence:

$$\dim \Gamma \geqq \sum_{a=2}^{n} [(K-1)a-(K-1)] = (K-1)\,\frac{n(n-1)}{2}$$

On the other hand $\Gamma = V^{\perp}$ where V consists of all functions $f_{a,t}(z)$; $t = 0,1,\ldots,(K-1)a$, $a = 1,\ldots,n$. Hence

$$\dim \Gamma = N - \dim V$$

By the rowrank = column rank theorem $\dim V$ is the dimension of the linear space spanned by the polynomials $f_{a,t}$; $t = 0,1,\ldots,(K-1)a$, $a = 0,\ldots,n$. Hence - since the polynomials $1,z,\ldots,z^{nK}$ are linear combinations of the polynomials $f_{a,t}$; $t = 0,\ldots,(K-1)a$, $a = 0,\ldots,n$ - $\dim V = nK+1$. It follows that:

$$\dim \Gamma = N-nK-1 = (K-1)\,\frac{n(n-1)}{2}$$

This imply that the functions $\xi_{a,t}$: $t \geqq K$, $a \geqq 2$ spans Γ.

<u>Theorem G.3.</u>

$\delta(A_n,T_n)$ minimizes the variance at $p_o = 1-q_o$ within the class of all unbiased estimators of $E_p\delta(A_n,T_n)$ if and only if

$$(\S) \qquad \delta(a,t) = (1-\frac{1}{q_o})\delta(a-1,t-K+1)+\frac{1}{q_o}\delta(a-1,t-K)$$

when $t \geqq K$ and $a \geqq 2$.

Remark 1. The space of estimators $\delta(A_n,T_n)$ which minimizes the variance at p_o within the class of all unbiased estimators of $E_p\delta(A_n,T_n)$ is - by (§) - isomorphic to the $Kn+1$ dimensional space of real valued functions on $\{(a,t) : a = t = 0$ or $1 \leqq a \leqq n$ and $0 \leqq t < K\}$. This is as it should be since a real valued function g of p has an unbiased estimator (and among them there is a unique one minimizing the variance at p_o) if and only if g is a polynomial of degree $\leqq nK$.

Remark 2. By (§) - any statistic $\delta(A_n,T_n)$ which minimizes the variance at two values of p within the class of all unbiased estimators of $E_p\,\delta(A_n,T_n)$ is an UMVU estimator.

Proof of the theorem. By the considerations above $\delta(A_n,T_n)$ minimizes the variance at p_o within the class of unbiased estimators of $E_p\,\delta(A_n,T_n)$ if and only if:

$$\sum_{a,t} \xi_{a_o,t_o}(a,t)\,\delta(a,t) f_{a,t}(z_o) = 0$$

where $z_o = -\frac{1}{p_o}$, when $t_o \geqq K$ and $a_o \geqq 2$; i.e. if and only if $[\delta(a_o,t_o)-\delta(a_o-1,t_o-K+1)]p_o = \delta(a_o,t_o)-\delta(a_o-1,t_o-K)$; $t \geq K$, $a \geqq 2$. □

We have also obtained the following characterizations of the unbiased estimators of zero.

Theorem G.4.

A statistic $\eta(A_n, T_n)$ is an unbiased estimator of zero if and only if the function $(a,t) \longrightarrow (-1)^{Ka-t}\eta(a,t)\binom{n}{a}C_K(a,t)$ belongs to Γ. It follows that the functions
$(a,t) \longrightarrow (-1)^{Ka-t}\binom{n}{a}^{-1}C_K(a,t)^{-1}\xi_{a_0,t_0}(a,t) : a_0 \geq 2 \quad t_0 \geq K$
spans the space of unbiased estimators of zero.

Corollary G.5.

The space of unbiased estimators of zero considered as a sub space of R^N has dimension = dim $\Gamma = \frac{n(n-1)(K-1)}{2}$. In particular $\mathcal{E}^n$ does not admit a complete and sufficient statistic when $K \geq 2$ and $n \geq 2$.

The algebra of UMVU estimators is now easily obtained:

Theorem G.6.

$\delta(A_n, T_n)$ is an UMVU estimator of its expectation if and only if $\delta(1,0) = \delta(a,t)$ when $1 \leq a < n$ or ($a = n$ and $t \geq K$).

Proof: Follows from remark 2 after theorem G.3. □

Corollary G.7.

$\delta(A_n, T_n)$ is an UMVU estimator of its expectation if and only if it is measurable w.r.t. the set algebra generated by the $K+1$ events:

$$A_n = T_n = 0 \text{ and } (A_n, T_n) = (n, t) \text{ where } t < K.$$

Corollary G.8.

A function g of p has an UMVU estimator if and only if it is of the form:

$$(\dagger) \qquad g(p) = dp^{nK} + q^n[d_0 + d_1 p + \dots + d_{K-1} p^{K-1}] + d_K$$

where $d, d_0, \dots, d_K$ are constants. The UMVU estimator of g given by $(\dagger)$ is

$$dI_{[A_n = T_n = 0]} + d_0 I_{[A_n = n, T_n = 0]} + d_1 I_{[A_n = n, T_n = 1]} + \dots$$
$$+ d_{K-1} I_{[A_n = n, T_n = K-1]} + d_K.$$

Corollary G.9.

Let U_n denote the space of expectations of UMVU estimators based on n observations. Suppose $n > m \geqq 1$. Then $U_n \cap U_m$ consists of all functions of the form constant, $+q^n \varphi(q)$ where φ is a polynomial of degree $\leqq K - 1 - (n-m)$. Finite intersections of sets U_n may be reduced to this case since:

$$\bigcap_{i=1}^{r} U_{n_i} = U_n \cap U_m \text{ where } n = \max_i n_i \text{ and}$$

$m = \min_i n_i$.

---

the $K+1$ events:

$$A_n = T_n = 0 \quad \text{and} \quad (A_n, T_n) = (n,t) \quad \text{where} \quad t < K.$$

Corollary G.8.

A function g of p has an UMVU estimator if and only if it is of the form:

$$(\dagger) \qquad g(p) = dp^{nK} + q^n[d_0 + d_1 p + \dots + d_{K-1} p^{K-1}] + d_K$$

where $d, d_0, \dots, d_K$ are constants. The UMVU estimator of g given by $(\dagger)$ is

$$dI_{[A_n=T_n=0]} + d_0 I_{[A_n=n, T_n=0]} + d_1 I_{[A_n=n, T_n=1]} + \dots$$
$$+ d_{K-1} I_{[A_n=n, T_n=K-1]} + d_K.$$

Corollary G.9.

Let U_n denote the space of expectations of UMVU estimators based on n observations. Suppose $n > m \geq 1$. Then $U_n \cap U_m$ consists of all functions of the form constant $+ q^n \varphi(q)$ where φ is a polynomial of degree $\leq K-1-(n-m)$. Finite intersections of sets U_n may be reduced to this case since:

$$\bigcap_{i=1}^{r} U_{n_i} = U_n \cap U_m \quad \text{where} \quad n = \max_i n_i \quad \text{and}$$

$m = \min_i n_i$.

Remark: Let g be a non constant function of p which has an UMVU estimator based on n observations for some n. Then we may write:

$$g(p) = cp^{nK}+q^{n}[c_0+\dots+c_{K-1}q^{K-1}]+c_K$$

Suppose first that $n \geqq 2$. If $c \neq 0$ then g does not have a UMVU estimator based on m observations for any $m \neq n$. If $c = 0$, then g have a UMVU estimator based on m observations if and only if $n-[K-1-\max\{i : 0 \leqq i \leqq K-1, c_i \neq 0\}] \leqq m \leqq n+\min\{i : 0 \leqq i \leqq K-1, c_i \neq 0\}$.

Suppose next that $n = 1$. Then we may rewrite g as:

$$g(p) = q[c_0'+\dots+c_{K-1}'q^{K-1}]+c_K$$

Hence g has a UMVU estimator based on m observations if and only if

$$m \leqq \min\{i : 0 \leqq i \leqq K-1, c_i' \neq 0\}$$

It follows that the set of positive integers m such that g has an UMVU estimator based on m observations is an interval with at most $K-1$ points.

<u>Proof</u>: By corollary G.8, U_n is the space of all functions g of the form:

$$g(p) = cp^{nK}+q^n[c_o+c_1q+\dots+c_{K-1}q^{K-1}]+c_K$$

where c $c_o,\dots,c_K$ are constants. Suppose $g \in U_n \cap U_m$ where $n > m \geqq 1$. Then there are constants d $d_o,\dots,d_{K-1}$ and d_K so that

$$g(p) = dp^{mK}+q^m[d_o+d_1q+\dots+d_{K-1}q^{K-1}]+d_K.$$

Hence - since $\deg g \leqq mK < nK$ - $c = 0$. It follows that q^m is a factor in the polynomial $dp^m+d_K-c_K$. Hence $0 = d1^{mK}+d_K-c_K = d+d_K-c_K$, so that q^m is a factor in $d(p^{mK}-1)$. Consider first the case $m \geqq 2$. Then $d = 0$ so that $q^{n-m}[c_o+c\ q+\dots+c_{K-1}q^{K-1}] \equiv [d_o+d_1q+\dots+d_{K-1}q^{K-1}]$. This imply that $c_i = 0$ when $i > K-1-(n-m)$. Suppose next that $m = 1$. It follows from the identity:

$$q^n[c_o+\dots+c_{K-1}q^{K-1}]+c_K \equiv dp^K+q[d_o+\dots+d_{K-1}q^{K-1}]+d_K$$

that $c_i = 0$ when $i > K-n = K-1-(n-m)$.

We have - so far - shown that any $g \in U_n \cap U_m$ is of the form:

$$g(p) = \text{constant} + q^n\varphi(q) \quad \text{where}$$

φ is a polynomial of degree $\leqq K-1-(n-m)$, and it is easily seen that - conversely - any g of this form is in $U_n \cap U_m$.

Finally let $n > m > l \geqq 1$ and consider a function g in $U_n \cap U_l$, i.e. g is of the form

$$g(p) = \text{constant} + q^n \varphi(q) \quad \text{where}$$

φ is a polynomial of degree $\leqq K-1-(n-1)$. Then $q^{n-m}\varphi(q)$ is a polynomial of degree $\leqq K-1-(m-1)$, i.e. $g \in U_m$. Hence $U_n \cap U_m \cap U_l = U_n \cap U_l$, and this proves the last statement. $\square$