Estimation of Gap Distribution

L. H. Crow** and I. N. Shimi*

Florida State University

FSU Statistics Report M209

September, 1971

The Florida State University
Department of Statistics
Tallahassee, Florida

*Prepared under contract No. DOT-FH-11-7661 with the U. S. Department of Transportation, Federal Highway Administration. The opinions, findings and conclusions expressed in this publication are those of the authors and not necessarily those of the Federal Highway.

**Supported in part by a National Institute of Health fellowship grant (5-T01-GM-913).
1. **INTRODUCTION.** A large number of articles have recently been published dealing with the estimation of the distribution of gaps between cars. All these articles use the usual statistical procedures when the total number of gap measurements is fixed and the total length of time needed to take the observations is random. However, these procedures are not practical, especially when we are dealing with light traffic, since the length of time needed for the accumulation of the required number of observations might be too long to be practical. Also, this type of procedure is not desirable when the length of time needed for the observations or the segment of roadway over which the gaps develop must not exceed a certain length. For example, when certain project deadlines are to be met, or when equipment or personnel used in collecting the observations can only be used for some specified length of time. A more important and serious objection to the usually used procedures is the following. The underlying assumption for the validity of the estimators obtained in the usual procedures is that we are dealing with a space-time homogeneous traffic flows, otherwise the estimation procedure is meaningless. The space-time homogeneity of the traffic flow can be safely assumed over short periods of time and over bounded segments of roadway. So, unless we limit the time of observations and the segment of roadway under consideration to bounded lengths, we might not have the required homogeneity of the traffic flow. The present article deals with the estimation of the gap distribution, for a wide class of distributions, with the above remarks in mind. The techniques developed later
use traffic data gathered by two different and currently employed methods. One method uses the Bureau of Public Roads Traffic Analyzer, the other uses the Aerial Photographic data.

2. **Sampling Techniques.**
   
   (1) Gaps within fixed periods of time.

   One might be interested in estimating the distribution of gaps between cars passing a fixed point during a given period of time, \( T \) say. The traffic data for this case is gathered using the Traffic Analyzer, which computes passage time and speeds as each car passes a given fixed point. The sampling scheme used is the following. We start the observations with the passage of a car over the sensing device of the Traffic Analyzer. Keep taking observations on the analyzer for a predetermined length of time, \( T \) say. Wait for a certain period of time to elapse from the end of the first observation period to guarantee the independence of the gaps in the next observation period from the gaps in the first observation period. After this guarantee period, we start the second observation period with the passage of a car over the sensing device of the analyzer. The second period also lasts for a time interval of length \( T \) from its start. We repeat these guarantee and observation periods a number of times, \( n \) say. Assuming that the length of the observation periods is not too long to make the space-time homogeneous flow assumption invalid, we will show in the next section how to use data collected by this method to estimate the gap-distribution for the class of distributions under consideration. In each observation period we have a random
number \( K_r, r = 1, 2, \ldots, n \), of cars passing the fixed point of the analyzer. Let \( X_{jr} \) be the gap after the \( j \)-th car in the \( r \)-th observation period. Then \( K_r \) is the first integer for which \( \sum_{j=1}^{K_r} X_{jr} \geq T \).

(ii) Gaps within a fixed interval of roadway.

One might be interested in estimating the distribution of gaps between cars that are traveling within a given segment of roadway of length \( T \). The traffic data for this case can be gathered using the Aerial Photographic method. Assume that traffic flows over the one-lane segment \((a,b)\), whose length is \( T \), in the direction \( a \) to \( b \). We take an aerial photograph of the configuration of the cars on this segment when the leading car hits the end point \( b \). The number of cars

\[
\begin{array}{ccccccccc}
\times & \times & \times & \times & \times & \times & \times & \times & \times \\
\downarrow & & & & & & & & \\
b & X_{1r} & X_{2r} & \cdots & X_{K_r-1,r} & a
\end{array}
\]

in each photograph will be random. We take a number of these aerial photographs, \( n \) say, with a sufficient period between photographs to allow all the cars in the preceding frame to have left the segment \((a,b)\) before we take the next frame. This period between photographs guarantees the independence of observations in any photograph from the observations in the others. Again, let \( K_r, r = 1, \ldots, n \), denote the random number of cars in the \( r \)-th photograph and let \( X_{jr} \) be the gap behind the \( j \)-th car in the \( r \)-th photograph. Then \( K_r \) will be the first integer such that \( \sum_{j=1}^{K_r} X_{jr} \geq T \).
Throughout this paper it is assumed that the $X_{jr}$, $j \geq 1$, $r = 1, \ldots, n$, are independent, identically distributed random variables. Also, we assume that the c.d.f. $F(\cdot)$ of the $X_{jr}$'s has the following properties: 

$F(x) = 1 - \exp(-\lambda g(x))$, $0 \leq x \leq T$, where $\lambda > 0$, $g(\cdot)$ is known and is a strictly increasing function with $g(0) = 0$ and derivative $g'(x)$, $0 < x \leq T$. Note that nothing is assumed about the c.d.f. for $x > T$. This parametric class of distributions is obviously relevant to gaps between cars. It includes for example the exponential (when $g(x) = x$), the Weibull (when $g(x) = x^\alpha$, $\alpha > 0$) and the extreme-value distributions (when $g(x) = e^x - 1$).

In this paper the maximum likelihood estimate (MLE) $\hat{\lambda}_n$ of $\lambda$ will be given and shown to be strongly consistent, asymptotically normally distributed and asymptotically efficient, as $n \to \infty$. In the above two sampling plans (i) and (ii), we shall call the interval of time of length $T$ and the interval of roadway of length $T$ the observation interval. This unifies the argument for both sampling plans.

We would like to point out a very tempting pitfall one might be led to if not careful. Suppose we have determined the MLE for $\lambda$ when $g(x) = x$, i.e., when the gap distribution is the exponential. Now suppose one does not have gaps that are exponentially distributed but random variables which may be transformed to exponential random variables, i.e., when $X$ has c.d.f. $F(x) = 1 - \exp(-\lambda g(x))$, then $Y = g(x)$ will have the exponential c.d.f. $1 - \exp - \lambda y$. In this case, one might be tempted to make this transformation on the observations and use the results for
the exponential case. However, if the exponential results were used then the length of the observation interval will not necessarily be T, which violates the purpose of fixing the length of the observation interval as mentioned above. To see this, suppose that the c.d.f. of the gaps between cars is \( F(x) = 1 - \exp(-\lambda g(x)) \), \( x > 0, \lambda > 0, g(x) \) is strictly increasing function, differentiable for \( x > 0 \), \( g(0) = 0 \) and \( g(\infty) = \infty \). Then the random variables \( W_{ir} = g(X_{ir}) \), \( i \geq 1, \ldots, n \), have an exponential c.d.f. with mean \( 1/\lambda \). To use the exponential theory, one must choose a constant \( C \) such that in each of the \( n \) observations intervals, the random number of items put on test, \( K_r, r = 1, \ldots, n \), is the first integer such that \( \sum_{i=1}^{K_r} W_{ir} \geq C \). Hence the length of the observation interval is the \( r \)-th observation period is

\[
\sum_{i=0}^{K_r-1} X_{ir} + g^{-1}(C - \sum_{i=0}^{K_r-1} g(X_{ir})) , \text{ (where } X_{0r} = 0 \text{ for } r = 1, \ldots, n) .
\]

If \( C \) is to be chosen such that the \( r \)-th observation interval is \( T \), then for \( K_r = 1 \), the length of the observation interval is \( g^{-1}(C) = T \). Thus, \( g(T) \) is the only candidate for the value of \( C \). Now, if

\[
\sum_{i=0}^{K_r-1} X_{ir} + g^{-1}(g(T) - \sum_{i=0}^{K_r-1} g(X_{ir})) = T \text{ for } K_r > 1, \text{ this would imply that } g(T) - \sum_{i=0}^{K_r-1} g(X_{ir}) = g(T - \sum_{i=0}^{K_r-1} X_{ir})
\]
which is, in general, not true for non-linear g. Thus, when one makes such a transformation, the observation interval may be less than T or greater than T and, also, the length of the observation interval may vary in each of the n observation intervals. The work presented in this paper makes it possible to avoid this pitfall.

3. MAXIMUM LIKELIHOOD ESTIMATE.

Let \( X_1, X_2, \ldots \) be independent, identically distributed positive random variables with distribution function \( F(x), x > 0 \). Then, with probability one, there exists an integer \( N \) such that \( \sum_{i=0}^{N} X_i \geq T \). If \( K \) is the stopping variable defined to be the first integer such that \( \sum_{j=1}^{K} X_j \geq T \), then \( K \) will be finite with probability one. Let

\[
A_k = \{(x_1, \ldots, x_k) : \sum_{j=1}^{k-1} x_j \leq T\} \quad \text{and} \quad B_k = \{(x_1, x_2, \ldots, x_k) : \sum_{j=1}^{k} x_j \geq T\}.
\]

We adopt the convention that sums of the form \( \sum_{j=1}^{0} \) are equal to 0 and products of the form \( \prod_{j=1}^{0} \) are equal to 1. Since \( K < \infty \) w.p.1 it follows that

\[
1 = \sum_{k=1}^{\infty} \int_{A_k} \cdots \int_{B_k} \prod_{j=k}^{K} \text{dF}(x_j)
\]

\[
= \sum_{k=1}^{\infty} \int_{A_k} \cdots \int_{\left[1 - F\left(T - \sum_{i=1}^{k-1} x_i\right)\right]} \prod_{j=1}^{k-1} \text{dF}(x_j).
\]

Hence, for the sampling plan considered in this paper, the likelihood \( L \) for the \( r^{th} \) observation interval is given by

\[
L_r = \prod_{j=1}^{K_r-1} \lambda g'(X_{jr}) \exp(-\lambda g(X_{jr})) \cdot \exp(-\lambda g(T - \sum_{i=1}^{j-1} X_{ir})).
\]
and the likelihood for the n independent observation intervals is
\[ L = \prod_{r=1}^{n} L_r. \]
Let
\[ Y_{ir} = X_{ir}, \quad i = 1, \ldots, K_r - 1, \quad r = 1, \ldots, n, \] and
\[ Y_{Kr} = T - \sum_{j=1}^{K_r-1} X_{jr}, \quad r = 1, \ldots, n. \]
Then
\[ L = \prod_{r=1}^{n} \lambda^{K_r - 1} \prod_{j=1}^{K_r - 1} g'(Y_{jr}) \exp\{-\lambda \sum_{i=1}^{K_r} g(Y_{ir})\}. \tag{3.2} \]
Thus, we have the following

**Lemma 3.1**

The MLE of \( \lambda \) based on the n independent experiments is
\[ \hat{\lambda}_n = \frac{\sum_{r=1}^{n} (K_r - 1)}{\sum_{r=1}^{n} \sum_{i=1}^{K_r} g(Y_{ir})}. \tag{3.3} \]

**Proof:** The proof follows from (3.2).

4. **Strong consistency of \( \hat{\lambda}_n \)**

In this section it is shown that the MLE \( \hat{\lambda}_n \), given by (3.3) converges a.s. to \( \lambda \). We will first prove a couple of results needed in the proof of the main result of this section.

**Lemma 4.1**

If \( F_1 \) is any c.d.f. such that \( F(x) = F_1(x), \ 0 < x \leq T \), then
\[ P[K = k|F] = P[K = k|F_1], \text{ for all } k \geq 1. \]
Proof:

The proof follows from equation (3.1).

Let $E$ denote the expectation operator. The next result will be used to prove the strong consistency of $\hat{\lambda}_n$.

Theorem 4.2

$$\frac{1}{\lambda} E(K-1) = E\left( \sum_{i=1}^{K} g(Y_{i1}) \right) \quad (4.1)$$

Proof:

By Lemma 4.1 $E(K)$ does not depend on $F(x)$ for $x > T$. Hence, we may extend $g(x)$, for $x > T$, in any manner we wish to keep $F(\cdot)$ a c.d.f., and $E(K)$ will remain unchanged. We, therefore, choose $g(x)$, for $x > T$, to be $g(x) = g(T) + (x-T)$. That is, we are now considering $F(x) = 1 - \exp(-\lambda g(x))$, $0 < x$, $g(\cdot)$ is a strictly increasing function, $g(\infty) = \infty$, and $g'(x)$ exists for $x > 0$. Thus $g(X_1)$ has an exponential distribution with mean $1/\lambda$. By Wald's Lemma (1944),

$$\frac{1}{\lambda} E(K) = E\left( \sum_{i=1}^{K} g(X_i) \right). \quad (4.2)$$

Now, $E(g(X_K)) = E\left(E(g(X_K) \mid \sum_{i=1}^{K-1} X_i) \right)$

$$= E\left(E(g(X_K) \mid X_K > T - \sum_{i=1}^{K-1} X_i) \right)$$

$$= E\left(\begin{array}{c} E(g(X) \mid g(X) \geq g(T - \sum_{i=1}^{K-1} X_i)) \end{array} \right)$$

where $g(X)$ is a random variable with c.d.f. $1 - \exp(-\lambda g(x))$, $g(x) > 0$. Thus
\[ E(G(X_{k})) = E\left(\frac{1}{\lambda} + g(T - \sum_{i=1}^{K-1} X_i)\right) \]
\[ = \frac{1}{\lambda} + E(g(T - \sum_{i=1}^{K-1} X_i)) \quad (4.3) \]

Equations (4.2) and (4.3) imply (4.1).

The main result of this section is the following

**Theorem 4.3**

The MLE, \( \hat{\lambda}_n \), given by (3.3), is a strongly consistent estimator of \( \lambda \).

**Proof:**

By the strong law of large numbers \( \sum_{r=1}^{n} (K_r-1)/n \to E(K_{1}-1) \) a.s.,
and \( \sum_{r=1}^{n} \sum_{i=1}^{K_r} g(Y_{ir})/n \to E\left( \sum_{i=1}^{K_1} g(Y_{i1}) \right) \) a.s.

Hence \( \hat{\lambda}_n = \frac{\sum_{r=1}^{n} (K_r-1)/n}{\sum_{r=1}^{n} \sum_{i=1}^{K_r} g(Y_{ir})/n} \to \frac{E(K_{1}-1)}{E\left( \sum_{i=1}^{K_1} g(Y_{i1}) \right)} \) a.s.

The result follows from (4.1).

5. **ASYMPTOTIC NORMALITY OF \( \hat{\lambda}_n \).**

In this section the asymptotic normality of the MLE \( \hat{\lambda}_n \) is shown for two different, but asymptotically equivalent, normalizing sequences.

We begin with
Theorem 5.1

The asymptotic distribution of \( \frac{\lambda_n - \lambda}{\sqrt{D/n}} \) is Normal \((0,1)\), where

\[
D = \frac{1}{K_1} \mathbb{E}^2 \left( \sum_{j=1}^{K_1} g(Y_{1j}) \right)
\]

\[
\mathbb{E}^2 \left( \sum_{j=1}^{K_1} g(Y_{1j}) \right)
\]

Var \left[ (K_1-1) - \lambda \sum_{j=1}^{K_1} g(Y_{1j}) \right]

(5.1)

Proof

Let \( M_r = \sum_{j=1}^{K_r} g(Y_{rj}) \), \( V_r = (K_r - 1) \), \( M(n) = \sum_{r=1}^{n} \frac{M_r}{n} \), \( V(n) = \sum_{r=1}^{n} \frac{V_r}{n} \),

and let \( Z_r \) be the two-dimensional random vector, \( Z_r = (M_r, V_r) \), \( r = 1, \ldots, n \).

Also, let \( H(a,b) \) be the function of the two variables \( a,b \), \( H(a,b) = a/b \).

Now, \((M(n), V(n))\) is the first moment vector corresponding to the sample \( Z_1, Z_2, \ldots, Z_n \). By Cramér (1946, pages 353, 367), \( H(V(n), M(n)) = \lambda_n \) is asymptotically normal with asymptotic mean \( \frac{E(V(n))}{E(M(n))} = \lambda \) by (4.1), and asymptotic variance

\[
\frac{\text{Var}[V(n)]}{E^2[M(n)]} - 2\text{cov}(V(n), M(n)) \frac{E[V(n)]}{E^3[M(n)]} + \text{Var}[M(n)] \frac{E^2[V(n)]}{E^4[M(n)]}
\]

\[
= \frac{1}{K_1} \left( \text{Var}(K_1 - 1) - 2\lambda \text{cov}(K_1 - 1, \sum_{j=1}^{K_1} g(Y_{1j})) \right)
\]

\[
\times n \mathbb{E}^2 \left( \sum_{j=1}^{K_1} g(Y_{1j}) \right)
\]

\[
+ \lambda^2 \text{Var} \left( \sum_{j=1}^{K_1} g(Y_{1j}) \right)
\]

using (4.1). Hence the result.
Theorem 5.2

\[ E[K_{1-1}] = \text{Var}[K_{1-1} - \lambda \sum_{j=1}^{K} g(Y_{j1})] \] (5.2)

Proof:

Let \( K = K_{1-1}, X_i = X_{i1}, i = 1, \ldots, K - 1 \), and \( Y_i = Y_{i1}, i = 1, \ldots, K \).

Also, let \( f(x) = \lambda g'(x) \exp(-\lambda g(x)), 0 < x \leq T, \) and

\[
p(x_1, \ldots, x_{k-1} | \lambda) = \prod_{j=1}^{k-1} f(x_j) [1 - P(T - \sum_{i=1}^{k-1} x_i)].
\]

It is easy to verify, similar to Cramér (1946, p. 502), that

\[
E \left( \frac{d}{d\lambda} \log p(x_1, \ldots, x_{k-1} | \lambda) \right)^2 = -E \left( \frac{d^2}{d\lambda^2} \log p(x_1, \ldots, x_{k-1} | \lambda) \right) \] (5.3)

The left hand side of (5.3) is equal to \( E \left( \frac{K-1}{\lambda} - \sum_{j=1}^{K} g(Y_j) \right)^2 \). But using (4.1) we have

\[
\text{Var} \left( \frac{K-1}{\lambda} - \sum_{j=1}^{K} g(Y_j) \right) = E \left( \frac{K-1}{\lambda} - \sum_{j=1}^{K} g(Y_j) \right)^2. \] (5.4)

The right hand side of (5.3) is equal to \( \frac{1}{\lambda^2} E(K-1) \). Hence, (5.2) follows.

Using (4.1) and (5.2) it follows that

\[
D = \lambda^2 / E(K_{1-1}). \] (5.5)

From this we have the following

Corollary 5.3

The asymptotic distribution of \( \frac{\hat{\lambda} - \lambda}{\sqrt{\frac{\lambda^2}{2nE(K_{1-1})}} \sqrt{nB(K_{1-1})}} \) is normal \((0,1)\).

By the strong law of large numbers Corollary 5.3 gives (see Cramér, 1946, p. 254)
Corollary 5.4

The asymptotic distribution of
\[ \sqrt{\frac{n}{\lambda}} \left( \frac{\hat{\lambda} - \lambda}{\sqrt{\sum_{r=1}^{K_r-1} \lambda^2}} \right) \]

is Normal (0, 1).

6. ASYMPTOTIC EFFICIENCY OF \( \hat{\lambda}_n \).

Let \( h(X_1, \ldots, X_{K-1}) \) be an estimate of \( \lambda \), where \( X_i = X_{i1}, \ldots, X_i, i = 1, \ldots, K-1, K = K_1. \)

Then

Theorem 6.1

\[ D(1 + \frac{d}{d\lambda} B(\lambda))^2 \leq \text{Var}(h(X_1, \ldots, X_{K-1})), \text{where } B(\lambda) = E(h(X_1, \ldots, X_{K-1})|\lambda) - \lambda, \]

and \( D \) is given by (5.5).

Proof:

It is straightforward to show that

\[ [1 + \frac{d}{d\lambda} B(\lambda)]^2 \leq \text{Var}(h(X_1, \ldots, X_{K-1})) \]

\[ \times \text{Var}\left( \frac{d}{d\lambda} \log p(X_1, \ldots, X_{K-1}|\lambda) \right)^2 \]

using (5.4), (5.2) and (5.5) yields the result.

This implies that if \( h \) is an unbiased estimator of \( \lambda \) based on the outcomes of \( n \) observation intervals, then

\[ \frac{D}{n} \leq \text{Var}(h_1). \quad (6.1) \]

From this we have the following
Theorem 6.2

\( \hat{\lambda}_n \) is an asymptotically efficient estimator of \( \lambda \).

Proof:

Our concept of efficiency is the same as the concept given by BAN estimators for fixed sample size. (see Rao (1968), p. 284).

The result then follows from Theorem 5.1 and inequality (6.1).
REFERENCES

