A model for the statistical analysis of a roving creel census: Preliminary report.

**ROBBIN DRAFT SECTION OF A PAPER BY GROSSLEIN, McCaughan and ROBSON**

**Relationships between frequency distributions of pre-interview time and total trip time in a roving creel census of fishermen.**

We assume that the probability of being interviewed by the roving enumerator is proportional to the total length \( T \) of the fishing trip,

\[
P(\text{interviewed}|T) = kT,\]

and that if an interview does occur it is "equally likely" to occur at any point during the fishing trip. Thus, if \( X \) denotes pre-interview time then

\[
f_X(x|T, \text{interviewed}) = \begin{cases} \frac{1}{T} & \text{for } 0 \leq x \leq T \\ 0 & \text{otherwise} \end{cases}
\]

The conditional frequency distribution of total trip length for interviewed fishermen, \( f_T(t|\text{interviewed}) \), is then related to the unconditional distribution for all fishermen, \( f_T(t) \), by

\[
f_T(t|\text{interviewed}) = \frac{tf_T(t)}{\int_0^\infty tf_T(t)} = \frac{tf_T(t)}{\mu_T}
\]

and the distribution of pre-interview time for interviewed fishermen is related to \( f_T(t) \) by
(2) \[
    f_X(x|\text{interviewed}) = \frac{\int \xi(t)dt}{\int \tau(t)dt} = \frac{1-P_x(x)}{\mu_T}.
\]

From (1), assumption (1) is seen to imply that

\[
    E(T|\text{interviewed}) = \frac{E(T^2)}{E(T)}
\]

and

\[
    \text{var}(T|\text{interviewed}) = \frac{E(T^3)}{E(T)} - \left[ \frac{E(T^2)}{E(T)} \right]^2.
\]

Similarly from (2), assumptions (1) and (ii) imply

\[
    E(X|\text{interviewed}) = \frac{1}{2} \frac{E(T^2)}{E(T)} = \frac{1}{2} E(T|\text{interviewed})
\]

and

\[
    \text{var}(X|\text{interviewed}) = \frac{1}{3} \frac{E(T^3)}{E(T)} - \frac{1}{4} \left[ \frac{E(T^2)}{E(T)} \right]^2
\]

\[
= \frac{1}{12} \frac{E(T^3)}{E(T)} + \frac{1}{4} \text{var}(T|\text{interviewed}).
\]

Statistical tests of assumptions (1) and (ii) can be made if in addition to the roving interviewer's sample there is available an independent sample of completed-trip interviews, as might be obtained at public launching sites, marinas, etc. We suppose that m parties have been interviewed in this manner, giving the observations $T_1, \ldots, T_m$ on completed trip time, and that D of these parties had
been visited by the roving interviewer in the course of their fishing trip.

Let \( \bar{T}_D \) denote the mean completed trip time for these \( D \) parties, then if assumption (1) is true the test statistic

\[
Z_T = \sqrt{D} \left( \bar{T}_D - \frac{1}{m} \sum_{i=1}^{m} T_i \right) \sqrt{ \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{T_i} \right)^2 - \frac{1}{m} \left( \frac{1}{\sum_{i=1}^{m} T_i} \right)^2 }
\]

is approximately normally distributed with mean 0 and variance 1. We note that if \( m \) is large relative to \( D \) then the sampling model is essentially pps (probability proportionate to size) sampling with replacement from the population \( \{T_1, \cdots, T_m\} \), and for this model we would have the exact results

\[
E(\bar{T}_D | T_1, \cdots, T_m, D) = \frac{1}{D} \sum_{i=1}^{m} T_i
\]

\[
\text{var}(\bar{T}_D | T_1, \cdots, T_m, D) = \frac{1}{D} \left[ \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{T_i} \right)^2 - \left( \frac{1}{\sum_{i=1}^{m} T_i} \right)^2 \right]
\]

The same argument provides a test of the second moment,
\[ Z_{T^2} = \frac{\sum_{1}^{D} T_i^2 - D \frac{\sum_{1}^{D} T_i^3}{\sum_{1}^{D} T_i}}{\sqrt{D \left( \frac{\sum_{1}^{D} T_i^3}{\sum_{1}^{D} T_i} - \frac{1}{3} \right)}} \]

though \( Z_{T^2} \) is not statistically independent of \( Z_T \). When several such samples are available then \( \sum Z_{T_1}^2 \) and \( \sum Z_{T_2}^2 \) are approximately chi-square variables with degrees of freedom equal to the number of samples.

Tests of assumption (ii) may likewise be constructed as functions of the completed trip times \( T_1, \ldots, T_D \) and the pre-interview times \( x_1, \ldots, x_D \) for the \( D \) parties contacted by the roving interviewer. Letting \( u = x/T \) we see by assumption (ii) that \( u_1, \ldots, u_D \) are independent random variables uniformly distributed on the interval \((0,1)\). Under this hypothesis the test statistics

\[ Z_{1u} = \frac{1}{\sqrt{D/12}} \sum_{1}^{D} (u_1 - \frac{1}{2}) \quad \text{(testing whether the mean of } u \text{ is } \frac{1}{2}) \]

\[ Z_{2u} = \frac{1}{\sqrt{4D/5}} \left( \frac{D - 12 \sum (u_1 - \frac{1}{2})^2}{\sum (u_1 - \frac{1}{2})^2} \right) \quad \text{(testing whether } \sigma_u^2 = 1/12) \]

\[ Z_{3u} = \frac{1}{\sqrt{\frac{D}{12}} \sum\left( T_1 - \frac{\sum_{1}^{D} T_i}{D} \right)^2} \quad \text{(testing for correlation between } u \text{ and } T) \]
\[ z_{4u} = \frac{D}{\sqrt{\sum_{i=1}^{n} T_i^2 - T_D^2}} \] (testing for correlation between \( \sigma_u^2 \) and \( T \))

are approximately independent and identically distributed as standard normal deviates. The sum of squares, \( z_{1u}^2 + z_{2u}^2 + z_{3u}^2 + z_{4u}^2 \), is therefore approximately distributed as chi-square with 4 degrees of freedom. Unlike the preceding tests of assumption (i), the \( Z_u \) statistics may be computed from combined samples, such as all pairs \((X_i, T_i)\) measured during the one or more years of creel censusing a fishery.

**SHOW THESE TESTS ON ONEIDA LAKE DATA**

If both assumptions (i) and (ii) are satisfied then an unbiased point estimator of \( k \) is

\[ k = \frac{D}{\sum_{i=1}^{m} T_i} \]

with conditional variance

\[ \text{var}(\hat{k}|T_1, \ldots, T_m) = \sum_{i=1}^{m} \frac{T_i^2 (1 - kT_i)}{(\sum_{i=1}^{m} T_i)^2}. \]

The maximum likelihood estimator \( k^* \) can be obtained iteratively as the smallest root of

\[ \frac{D}{k^*} = \sum_{i=D+1}^{m} \frac{T_i}{1-k^*T_i} \]

with conditional variance of approximately
\[
\text{var}(k^*|T_1, \ldots, T_m) = \frac{1}{\sum_{i=1}^{m} \frac{T_i}{k(N-kT_i)}}
\]

A valid but conservative interval estimator of \(k\) is given by

\[
\frac{p_L(D|m)}{T_m} \leq k \leq \frac{p_u(D|m)}{T_m}
\]

where \(p_L(D|m)\) and \(p_u(D|m)\) are lower and upper binomial confidence limits associated with \(D\) "successes" in \(m\) trials. These limits are conservative (the actual confidence level exceeds the nominal confidence level) if there is variation in \(T_1, \ldots, T_m\); otherwise the actual confidence level is the same as the nominal confidence level. Note also that if \(T_1, \ldots, T_m\) are chosen to be equal then \(k^* = k^\ast\), and \(\text{var}(k|T_1 = \ldots = T_m = T) = \text{var}(k^*|T_1 = \ldots = T_m = T)\) is minimized by choosing \(T\) as large as possible.

Similar reasoning applied to the results of the roving interviewer produces

\[
\hat{N}-m
\]

a point estimator of the total effort \(N\) of the remaining \(N-m\) parties present in the fishery during the period of sampling. If the roving interviewer obtains \(n-D\) interviews from this group, with observations \(X_1, \ldots, X_{n-D}\) then

\[
E(n-D|T_1, \ldots, T_{N-m}) = k \sum_{i=1}^{N-m} T_i
\]

so \((n-D)/k\) estimates \(N\), or
The observations $D$ and $n - D$, coming from two disjoint segments of the population, are statistically independent and

$$\varnothing(D) = k \mathbb{E}^2 - k^2 \mathbb{E}^2 \frac{N}{1} \frac{N}{1}$$

$$\varnothing(n - D) = k \mathbb{E}^2 - k^2 \mathbb{E}^2 \frac{N}{1} \frac{N}{1}$$

Substituting $k$ for $k$ in $\varnothing(D)$ provides an estimate of this variance, and to obtain an estimate of $\varnothing(n - D)$ we note that the pps sampling model gives

$$\begin{align*}
\mathbb{E}(n - D) & = \frac{1}{2} \left( \frac{N}{m} \sum T^2 \mathbb{E}^2 \frac{1}{1} \frac{N}{1} \right) \\
\mathbb{E}(n - D) & = \frac{1}{2} \frac{k}{N-m} \mathbb{E}^2 \frac{N}{1} \frac{N}{1}
\end{align*}$$

and so

$$\begin{align*}
\mathbb{E}(n - D - 2k \sum X_i | T_1, \ldots, T_N) & = \frac{1}{2} \frac{k}{N-m} \mathbb{E}^2 \frac{N}{1} \frac{N}{1}
\end{align*}$$

Utilizing the fact that $(n - D) \mathbb{E}^2 - D \mathbb{E}^2 \frac{N}{1} \frac{N}{1}$ is approximately normally distributed with zero mean and with variance $(2\mathbb{E})^2 \varnothing(n - D) + (\sum \mathbb{E})^2 \varnothing(D)$,

$$\begin{align*}
\mathbb{E}(n - D - 2k \sum X_i | T_1, \ldots, T_N) & = \frac{1}{2} \frac{k}{N-m} \mathbb{E}^2 \frac{N}{1} \frac{N}{1}
\end{align*}$$

we obtain approximate confidence limits on $\sum T$ as the solutions to the quadratic
These results apply over a sampling period for which the proportionality factor \( k \) remains constant. The homogeneity of \( k \) over \( v \) sampling periods may be tested by

\[
\frac{1}{\sum_{i=1}^{v} \frac{D_i}{m_i}} \sum_{i=1}^{v} \frac{m_i}{\sum_{j=1}^{\hat{m}} T_{ij} - k \sum_{j=1}^{\hat{m}} T_{ij}^2} - \left( \frac{\sum_{i=1}^{v} \frac{D_i}{m_i}}{\sum_{i=1}^{v} \sum_{j=1}^{\hat{m}} T_{ij}^2 - k \sum_{i=1}^{v} \sum_{j=1}^{\hat{m}} T_{ij}} \right)^2 = \chi^2_{v-1}
\]

when

\[
\hat{k} = \frac{\sum_{i=1}^{v} \frac{D_i}{m_i}}{\sum_{i=1}^{v} \sum_{j=1}^{\hat{m}} T_{ij}}
\]

Fitting a probability density function to the observed frequency distribution of pre-interview time

We first note the interesting fact that if the monotonically decreasing function (2) is a simple exponential density function then the distribution of pre-interview time for interviewed fishermen is identical with the distribution
of total trip time for all fishermen -- in fact, these two distributions are identical if and only if they are simple exponential densities. Clearly, if $f_T(t)$ is exponential,

$$f_T(t) = \frac{1}{\mu_T} e^{-\frac{t}{\mu_T}},$$

then also

$$\frac{1-F(x)}{1/\mu_T} = \frac{1}{\mu_T} e^{-\frac{x}{\mu_T}}.$$

which, from (X), is $f_X(x|\text{interviewed})$. Likewise, if $f_X(x|\text{interviewed})$ is exponential,

$$f_X(x|\text{interviewed}) = \frac{1}{\mu_X} e^{-\frac{x}{\mu_X}},$$

then applying (X) at $x = 0$ ($F_T(0) = 0$) we find $\mu_X = \mu_T$ and hence at $x$ we get

$$e^{-\frac{x}{\mu_X}} = 1 - F_T(x)$$

or

$$F_T(x) = \frac{1}{\mu_X} e^{-\frac{x}{\mu_X}}.$$

To verify that these two distributions are identical only in the exponential case we assume

$$f_X(x|\text{interviewed}) = f_T(x).$$
From (x) we then get
\[
\frac{\gamma}{\mu_T} = \frac{f_T(x)}{1 - F_T(x)} = - \lambda \log [1 - F_T(x)]
\]
and integrating from 0 to \( t \) \((F_T(0) = 0)\) gives
\[
\frac{t}{\mu_T} = - \log [1 - F_T(t)]
\]
or
\[
F_T(t) = 1 - e^{-\frac{t}{\mu_T}}.
\]

The main interest value of this result lies not in its applicability to the creel census situation but rather in its inapplicability. An exponential distribution of completed fishing trip time is an absurdity, implying that fishermen are memoryless; however, as seen from equation (2), the frequency distribution of pre-interview time always shows shape characteristics of an exponential density function particularly when graphed in the form of a histogram. We now see that this appearance is deceiving, that fitting an exponential function to the distribution of pre-interview times would be an absurdity.

To determine a type of function which could reasonably be used to fit our observed frequency distribution of pre-interview time we need some indication of the functional form of the distribution of completed trip time. The \( 17 \frac{1}{2} \) completed
trip times for boat livery fishermen, though not representative of all fishermen. 

We suggest that an appropriate family of density functions to choose from is the gamma family,

\[ f(t) = \alpha^\beta t^{\beta-1} e^{-\alpha t} / \Gamma(\beta). \]

Figure 7 shows the frequency distribution of these 174 observations along with the gamma density function having the same mean and variance,

\[
\begin{align*}
\sum_{i=1}^{13} f_i &= 174, \\
\sum_{i=1}^{13} \left(1 - \frac{1}{2}\right)f_i &= 6.144 = \frac{\beta}{\alpha}, \\
\sum_{i=1}^{13} \left(1 - 6.644\right)^2 f_i &= 4.91 = \frac{\beta^2}{\alpha^2}
\end{align*}
\]

or

\[
\begin{align*}
\alpha &= 1.251, \\
\beta &= 7.689.
\end{align*}
\]

The fit appears reasonably good and we have therefore assumed that the same function, with different parameter values, approximates the frequency distribution of completed trip times for the entire fishery. The \( \alpha \) and \( \beta \) appropriate to this larger population are determined from a sample of 1350 pre-interview times for which

\[
\begin{align*}
\sum f_i &= 1350, \\
\sum_{i=1}^{1350} \left(1 - \frac{1}{2}\right)f_i &= 2.434 = \frac{\beta + 1}{\alpha^2}, \\
\sum_{i=1}^{1350} \left(1 - \frac{1}{2}\right)^2 f_i &= 10.334 = \frac{(\beta + 1)(\beta + 2)}{3\alpha^2}
\end{align*}
\]

or

\[
\begin{align*}
\alpha &= .666, \\
\beta &= 2.344.
\end{align*}
\]
Since the purpose of these calculations is only to develop "reasonable" input for simulation we lose nothing by rounding off $\beta$ to the nearest integer, $\beta = 2$, which is more convenient for simulation, and recalculating $\alpha$ from

$$2.43^\alpha = \frac{2 + 1}{2}, \quad 10.33^\alpha = \frac{(2 + 1)(2 + 2)}{2!}, \quad \alpha = 0.62$$

The resulting density function for pre-interview time

$$f_x(z | \text{interviewed}) = .31(1 + .62z) e^{-0.62z}$$

is shown in Figure ?? along with the corresponding observed frequency distribution and the underlying density for completed trip time

$$f_t(t) = (.62)^2 t e^{-0.62t}.$$