Stat 427/627 (Statistical Machine Learning, Baron)

Notes on Jackknife

Jackknife is an effective resampling method that was proposed by Morris Quenouille to estimate and reduce the bias of parameter estimates.

1.1 Jackknife resampling

Resampling refers to sampling from the original sample S with certain weights. The original weights are $(1/n)$ for each units in the sample, and the original *empirical distribution* is

$$\hat{F} = \begin{cases} \text{mass } \dfrac{1}{n} \text{ at each observation } x_i, \ i \in S \\ 0 \text{ elsewhere} \end{cases}$$

Resampling schemes assign different weights. Jackknife re-assigns weights as follows,

$$\hat{F}_{JK} = \begin{cases} \text{mass } \dfrac{1}{n-1} \text{ at each observation } x_i, \ i \in S, i \neq j \\ 0 \text{ elsewhere, including } x_j \end{cases}$$

That is, Jackknife removes unit j from the sample, and the new jackknife sample is

$$S_{(j)} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n).$$

1.2 Jackknife estimator

Suppose we estimate parameter θ with an estimator $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. The bias of $\hat{\theta}$ is

$$\text{Bias}(\hat{\theta}) = \mathbf{E}_\theta(\hat{\theta}) - \theta.$$

- How to estimate $\text{Bias}(\hat{\theta})$?
- How to reduce $|\text{Bias}(\hat{\theta})|$?
- If the bias is not zero, how to find an estimator with a smaller bias?

For almost all reasonable and practical estimates, $\text{Bias}(\hat{\theta}) \to 0$, as $n \to \infty$. Then, it is reasonable to assume a power series of the type

$$\mathbf{E}_\theta(\hat{\theta}) = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} + \ldots,$$

with some coefficients $\{a_k\}$.

1.2.1 Delete one

Based on a Jackknife sample $S_{(j)}$, we compute the Jackknife version of the estimator,

$$\hat{\theta}_{(j)} = \hat{\theta}(S_{(j)}) = \hat{\theta}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n),$$

whose expected value admits representation

$$\mathbf{E}_\theta(\hat{\theta}_{(j)}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots.$$

1.2.2 Average

For the sake of a smaller variance, let us average all such estimates and define

$$\hat{\theta}_{(\bullet)} = \frac{1}{n} \sum_{j=1}^{n} \hat{\theta}_{(j)}.$$

This averaged estimator has the same expected value as each $\hat{\theta}_{(j)}$,

$$\mathbf{E}_\theta(\hat{\theta}_{(\bullet)}) = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots.$$

1.2.3 Combine $\hat{\theta}_{(\bullet)}$ with $\hat{\theta}$

Now it is easy to combine the averaged Jackknife estimator $\hat{\theta}_{(\bullet)}$ with the original $\hat{\theta}$, to kill the main term in the bias of $\hat{\theta}$. Consider

$$\begin{aligned} \mathbf{E}_\theta \left\{ n\hat{\theta} - (n-1)\hat{\theta}_{(\bullet)} \right\} &= \{n\theta - (n-1)\theta\} + \{a_1 - a_1\} + \left\{ \frac{a_2}{n} - \frac{a_2}{n-1} \right\} + \ldots \\ &= \theta + \frac{a_2}{n(n-1)} + \ldots \\ &= \theta + \frac{a_2}{n^2} + O(n^{-3}). \end{aligned} \tag{1}$$

1.2.4 The Jackknife estimator

The Jackknife estimator of θ is

$$\hat{\theta}_{JK} = n\hat{\theta} - (n-1)\hat{\theta}_{(\bullet)}.$$

According to (1), its bias is of order $O(n^{-2})$ instead of $O(n^{-1})$, and thus, we have achieved our goal of bias reduction,

$$\text{Bias}(\hat{\theta}_{JK}) = \frac{a_2}{n^2} + O(n^{-3})$$

1.2.5 Estimation of the bias

In general, estimation of the bias is a tricky problem because we observed an estimator $\hat{\theta}$ only once, we cannot compute the average of such estimators, and it is not clear how this average differs from the true parameter θ. Now we can use the Jackknife method to estimate the bias.

We know that $\hat{\theta}_{JK}$ is "almost" unbiased, therefore, the difference between the original estimator $\hat{\theta}$ and $\hat{\theta}_{JK}$ is a good estimator of Bias$(\hat{\theta})$,

$$\widehat{\text{Bias}}(\hat{\theta}) = \hat{\theta} - \hat{\theta}_{JK} = (n-1)(\hat{\theta}_{(\bullet)} - \hat{\theta}).$$

1.2.6 Example - sample variance

As an example, consider an MLE version of the sample variance

$$\hat{\theta} = \frac{\sum_1^n (x_i - \bar{x})^2}{n} = \frac{\sum_1^n x_i^2}{n} - \bar{x}^2,$$

which is the maximum likelihood estimator of the population variance $\theta = \sigma^2 = \text{Var}X$ under the Normal distribution.

This estimator is biased. As we know, the unbiased version of the sample variance is

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}.$$

Apply the Jackknife method to the biased estimator $\hat{\theta}$.

First, delete unit j and compute

$$\begin{aligned}
\hat{\theta}_{(j)} &= \frac{\sum_{i\neq j} x_i^2}{n-1} - \bar{x}_{(j)}^2 \\
&= \frac{\sum_1^n x_i^2 - x_j^2}{n-1} - \frac{\left(\sum_1^n x_i - x_j\right)^2}{(n-1)^2} \\
&= \frac{\sum_1^n x_i^2 - x_j^2}{n-1} - \frac{\left(\sum_1^n x_i\right)^2 + x_j^2 - 2x_j \sum_1^n x_i}{(n-1)^2}
\end{aligned}$$

Then, average all $\hat{\theta}_{(j)}$,

$$\begin{aligned}
\hat{\theta}_{(\bullet)} &= \frac{1}{n}\sum_1^n \hat{\theta}_{(j)} \\
&= \frac{\sum_1^n x_i^2 - \frac{1}{n}\sum_1^n x_j^2}{n-1} - \frac{\left(\sum_1^n x_i\right)^2 + \frac{1}{n}\sum_1^n x_j^2 - 2\frac{1}{n}\sum_1^n x_j \sum_1^n x_i}{(n-1)^2} \\
&= \left(\frac{1}{n} - \frac{1}{n(n-1)^2}\right)\sum_1^n x_i^2 - \frac{n-2}{n(n-1)^2}\left(\sum_1^n x_j\right)^2.
\end{aligned}$$

Final step - obtain the Jackknife estimator

$$\begin{aligned}
\hat{\theta}_{JK} &= n\hat{\theta} - (n-1)\hat{\theta}_{(\bullet)} \\
&= \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) - \left(\frac{n-1}{n} - \frac{1}{n(n-1)}\right)\sum x_i^2 + \frac{n-2}{n(n-1)}\left(\sum x_j\right)^2 \\
&= \left(1 - \frac{n-1}{n} + \frac{1}{n(n-1)}\right)\sum x_i^2 - \left(\frac{1}{n} - \frac{n-2}{n(n-1)}\right)\left(\sum x_j\right)^2 \\
&= \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = s^2
\end{aligned}$$

The Jackknife method immediately converted the biased version of the sample variance into the unbiased version!

This was anticipated. Expected value of $\hat{\theta}$ is actually

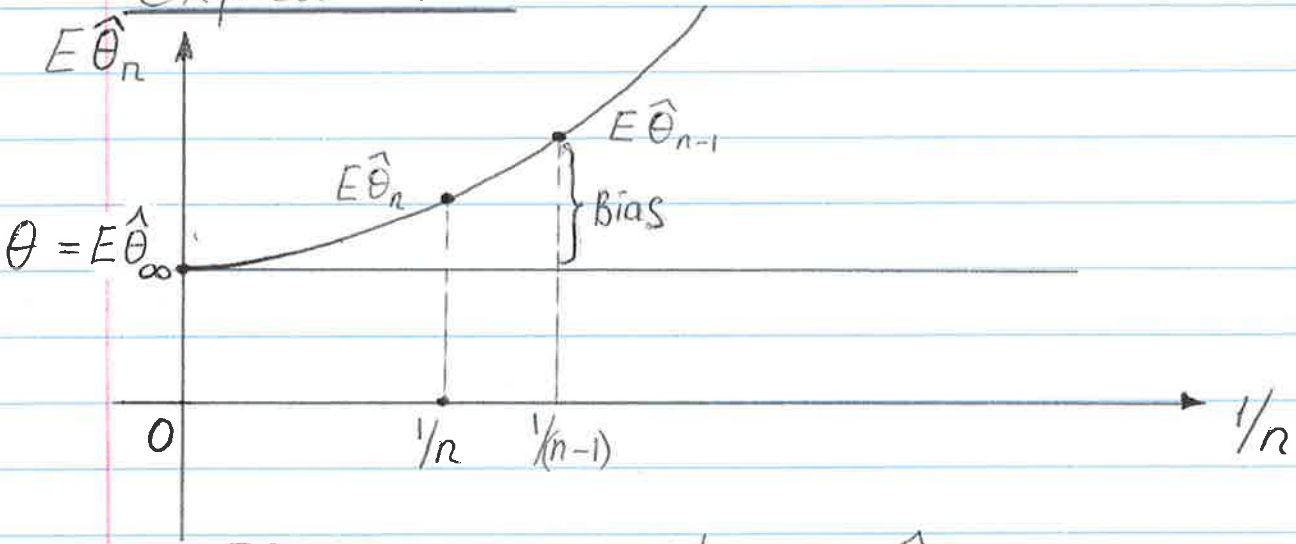
$$\mathbf{E}_\theta(\hat{\theta}) = \theta - \frac{\theta}{n},$$

with coefficients $a_1 = -\theta$ and $a_j = 0$ for all $j \geq 2$. Jackknife removes the (a_1/n) term of the power series. Since there are no other terms in this case, Jackknife removed the entire bias.

1.3 References

- (First mention of Jackknife) M. H. Quenouille. Approximate tests of correlation in time series. *J. of the Royal Statistical Society* B 11: 68–84 (1949)
- (Main reference) M. H. Quenouille. Notes on bias in estimation. *Biometrika* 43 (3-4): 353-360 (1956)
- R. G. Miller. The Jackknife - a review. *Biometrika* 61 (1): 1–15 (1974)

Geometric Explanation



If $n = \infty$, then $\frac{1}{n} = 0$, $\hat{\theta} = \theta$.

Up to 2^{nd} order terms,

Bias

$$\frac{E\hat{\theta}_n - \theta}{E\hat{\theta}_{n-1} - E\hat{\theta}_n} \approx \frac{1/n - 0}{1/n-1 - 1/n} = n - 1$$

$$\Rightarrow \text{Bias}(\hat{\theta}_n) \approx (n-1)(E\hat{\theta}_{n-1} - E\hat{\theta}_n)$$

$$\Rightarrow \widehat{\text{Bias}}(\hat{\theta}_n) = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_n)$$

$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_1^n \hat{\theta}_{(i)}$ estimates $E\hat{\theta}_{n-1}$

Estimates $E\hat{\theta}_n$

Extention

Double-Jackknife, etc.

Similarly, get $\hat{\theta}_{(jk)} = \hat{\theta}(S \setminus \{i_j, i_k\})$

Two units are deleted.

$$\hat{\theta}_{(\cdot\cdot)} = \binom{n}{2}^{-1} \sum_{(j,k)} \hat{\theta}_{(jk)}$$

What is the best combination

$$\hat{\theta}^{(2)}_{Jack} = u\,\hat{\theta} + v\,\hat{\theta}_{(\cdot)} + w\,\hat{\theta}_{(\cdot\cdot)} \quad ?$$

Derivation

$$E\hat{\theta} = \theta + \frac{a_1}{n} + \frac{a_2}{n} + \ldots \qquad (u)$$

$$E\hat{\theta}_{(\cdot)} = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots \qquad (v)$$

$$E\hat{\theta}_{(\cdot\cdot)} = \theta + \frac{a_1}{n-2} + \frac{a_2}{(n-2)^2} \qquad (w)$$

Solve the system:

$$\begin{cases} u + v + w = 1 \\ \dfrac{u}{n} + \dfrac{v}{n-1} + \dfrac{w}{n-2} = 0 \\ \dfrac{u}{n^2} + \dfrac{v}{(n-1)^2} + \dfrac{w}{(n-2)^2} = 0 \end{cases} \qquad (*)$$

Then $E\left(u\hat{\theta} + v\hat{\theta}_{(\cdot)} + w\hat{\theta}_{(\cdot\cdot)}\right)$

$$= (u+v+w)\theta + \left(\frac{u}{n} + \frac{v}{n-1} + \frac{w}{n-2}\right) a_1$$

$$+ \left(\frac{u}{n^2} + \frac{v}{(n-1)^2} + \frac{w}{(n-2)^2}\right) a_2 + O(n^{-3})$$

$$= \theta + O(n^{-3}).$$

$$\Rightarrow \quad \mathrm{Bias}\left(\hat{\theta}^{(2)}_{Jack}\right) = O(n^{-3}), \quad n \to \infty.$$

To solve $(*)$, find

$$\Delta = \begin{vmatrix} 1 & 1 & 1 \\ \frac{1}{n} & \frac{1}{n-1} & \frac{1}{n-2} \\ \frac{1}{n^2} & \frac{1}{(n-1)^2} & \frac{1}{(n-2)^2} \end{vmatrix} = \left(\frac{1}{n-2} - \frac{1}{n-1}\right)\left(\frac{1}{n-1} - \frac{1}{n}\right)\left(\frac{1}{n-2} - \frac{1}{n}\right)$$

$$\Delta_1 = \begin{vmatrix} \frac{1}{n-1} & \frac{1}{n-2} \\ \frac{1}{(n-1)^2} & \frac{1}{(n-2)^2} \end{vmatrix} = \frac{1}{(n-1)(n-2)}\left(\frac{1}{n-2} - \frac{1}{n-1}\right), \text{ etc.}$$

$$u = \frac{\Delta_1}{\Delta} = \frac{1/(n-1)(n-2)}{\left(\frac{1}{n-2} - \frac{1}{n}\right)\left(\frac{1}{n-1} - \frac{1}{n}\right)} = n^2/2$$

$$v = \frac{\Delta_2}{\Delta} = -(n-1)^2, \qquad w = \frac{\Delta_3}{\Delta} = \frac{(n-2)^2}{2}.$$

Thus,
a double-Jackknife estimator of θ is

$$\hat{\theta}^{(2)}_{Jack} = u\,\hat{\theta} + v\,\hat{\theta}_{(\cdot)} + w\,\hat{\theta}_{(\cdot\cdot)}$$

$$= \frac{n^2}{2}\hat{\theta} - (n-1)^2\,\hat{\theta}_{(\cdot)} + \frac{(n-2)^2}{2}\hat{\theta}_{(\cdot\cdot)}$$

Also, we can estimate the bias of $\hat{\theta}_{Jack}$:

$$\widehat{Bias}\left(\hat{\theta}_{Jack}\right) = \hat{\theta}_{Jack} - \hat{\theta}^{(2)}_{Jack}$$

So, should we keep reducing the bias by $\hat{\theta}^{(3)}_{Jack}$, $\hat{\theta}^{(4)}_{Jack}$, ... ?

Note on Bias Reduction.

Based on $S = (y_1 \dots y_n)$, estimate $\varphi(\theta)$, $\theta \in \Theta$ is a parameter.

Suppose there is no unbiased estimator of $\varphi(\theta)$, but

$$\exists \{T_k(S)\}_{k=1}^{\infty} \ : \ E_\theta T_k(S) \to \varphi(\theta) \quad \forall \theta \in \Theta$$

i.e. T_k has a bias which $\to 0$.

Then, one <u>should not</u> reduce the bias till $|Bias| < \varepsilon$!

$\boxed{\text{Theorem}}$ (Doss, Sethuraman, 1989).

(1) Let all probability measures P_θ be absolutely continuous with respect to each other, $\theta \in \Theta$.

(2) $\forall \theta_1, \theta_2 \in \Theta$, let $E_{\theta_2}\left(\frac{dP_{\theta_1}}{dP_{\theta_2}}\right)^2 < \infty$

i.e. $\left(dP_{\theta_1}/dP_{\theta_2}\right) \in L^2(\Omega, \mathcal{F}, P_{\theta_2})$

(3) $\exists \{T_k(S)\} \ : \ E_\theta T_k \to \varphi(\theta) \quad \forall \theta$.

Then : $Var_\theta(T_k) \xrightarrow[k\to\infty]{} \infty \quad \forall \theta \in \Theta$.

Proof:

If $Var_\theta(T_k) \not\to \infty$, then $\exists \theta \in \Theta$

and $\exists$ subsequence $\{T_{k^*}\} \subseteq \{T_k\}$: $Var_\theta(T_{k^*}) \leq C$
(bounded)

$\Rightarrow \{T_{k^*}\}$ is bounded in $L^2(\Omega, \mathcal{F}, P_\theta)$.

$\Rightarrow$ $\forall$ sequence in a bounded set has a convergent subsequence $\{T_{k^{**}}\}$.

$T_{k^{**}} \rightarrow T$. This T will be an unbiase
estimator

$\Rightarrow \langle T_{k^{**}}, f \rangle_{L^2} \longrightarrow \langle T, f \rangle_{L^2}$
for $\forall f \in L^2(\Omega, \mathcal{F}, P_\theta)$.

Let $f = \frac{dP_{\theta_1}}{dP_\theta} \in L^2(\Omega, \mathcal{F}, P_\theta)$ for an arbitrary θ_1.

$$\begin{tikzcd}
\Rightarrow \langle T_{k^{**}}, \frac{dP_{\theta_1}}{dP_\theta} \rangle \arrow[d] & = \int T_{k^{**}} \frac{dP_{\theta_1}}{dP_\theta} dP_\theta = E_{\theta_1} T_{k^{**}} \arrow[d, "\text{by the conditions}"] \\
\langle T, \frac{dP_{\theta_1}}{dP_\theta} \rangle = E_{\theta_1} T & \varphi(\theta_1)
\end{tikzcd}$$

$\Rightarrow E_{\theta_1} T = \varphi(\theta_1) \quad \forall \theta_1 \in \Theta$.

$\Rightarrow$ found an unbiased estimator !!!

Contradiction $\Rightarrow Var_{\theta_1} T_k \to \infty$. QED

The Annals of Statistics
1989, Vol. 17, No. 1, 440–442

THE PRICE OF BIAS REDUCTION WHEN THERE IS NO UNBIASED ESTIMATE

BY HANI DOSS[1] AND JAYARAM SETHURAMAN[2]

Florida State University

Let ϕ be a parameter for which there is no unbiased estimator. This note shows that for an arbitrary sequence of estimators $T^{(k)}$, if the biases of $T^{(k)}$ tend to 0 then their variances must tend to ∞.

1. Introduction. Let $X = (X_1, \dots, X_n)$ have distribution P_θ, where the unknown parameter varies in Θ. Suppose that we need to estimate a real valued function $\phi(\theta)$ of the parameter. Let $\hat{\phi} = \hat{\phi}(X)$ be a biased estimator of ϕ. There exist several procedures for reducing the bias of $\hat{\phi}$: jackknifing, bootstrapping [see Efron (1982)] and other procedures based on expansions of $E_\theta(\hat{\phi})$ [see Cox and Hinkley (1974), Section 8.4]. These procedures may not eliminate the bias completely, and one often hears the following suggestion. Let $\hat{\phi}^{(1)}$ be obtained from $\hat{\phi}$ by one of these bias-reduction procedures. If $\hat{\phi}^{(1)}$ is still biased, repeat the bias-reduction procedure and obtain $\hat{\phi}^{(2)}$, $\hat{\phi}^{(3)}$, etc., until a desired amount of reduction in bias is obtained or the bias is removed completely. Such "higher-order bias corrections" are described for instance in the review paper of Miller (1974) in connection with the jackknife. There are examples where no unbiased estimator of ϕ exists but there exists a sequence of estimators $\hat{\phi}, \hat{\phi}^{(1)}, \hat{\phi}^{(2)}, \dots,$ whose biases converge to 0 (see Section 2).

The purpose of this note is to show (Theorem 1) that when no unbiased estimator of ϕ exists, then reducing the bias to 0 necessarily forces the variance of the estimators to tend to ∞. This theorem therefore gives qualitative support to the widely held view that bias reduction is by itself not a desirable property, but becomes desirable only if it can be demonstrated that it is accompanied by a reduction in mean squared error.

2. Main result and remarks. Let $(\mathscr{X}, \mathscr{S})$ be a measurable space and $(P_\theta, \theta \in \Theta)$ be a family of probability measures on $(\mathscr{X}, \mathscr{S})$. Let ϕ be a real valued function defined on Θ. The bias of an estimator $T = T(X)$ is defined by $\beta_T(\theta) = E_\theta(T(X)) - \phi(\theta)$, assuming that $E_\theta(T(X))$ exists.

THEOREM 1. *Suppose that*

A1. $P_{\theta_1} \ll P_{\theta_2}$ *for all* θ_1, θ_2 *in* Θ,
A2. $\int (dP_{\theta_1}/dP_{\theta_2})^2 \, dP_{\theta_2} < \infty$ *for all* θ_1, θ_2 *in* Θ

Received December 1987.
[1] Research supported by Air Force Office of Scientific Research Grant 88-0040.
[2] Research supported by U.S. Army Research Office Grant DAAL03-86-K-0094.
AMS 1980 *subject classifications.* Primary 62F11; secondary 62A99.
Key words and phrases. Bias reduction, jackknife estimate of bias, bootstrap estimate of bias.

and that $\{T_k\}_{k=1}^{\infty}$ *is a sequence of estimators for which*

$$\beta_{T_k}(\theta) \to 0 \quad \textit{for all } \theta \textit{ in } \Theta. \tag{1}$$

If there does not exist an unbiased estimator of ϕ *then*

$$\operatorname{Var}_\theta(T_k) \to \infty \quad \textit{as } k \to \infty, \textit{ for all } \theta \in \Theta. \tag{2}$$

PROOF. Suppose that (2) is not true. Then there exists a θ_0 in Θ and a subsequence $\{k^*\}$ of $\{k\}$ such that $\operatorname{Var}_{\theta_0}(T_{k^*})$ is bounded. Now, consider the usual Hilbert space $H_{\theta_0} = L^2(\mathscr{X}, \mathscr{S}, P_{\theta_0})$ of all functions that are square-integrable with respect to P_{θ_0}. Notice that $\{T_{k^*}\}$ is a norm-bounded set in H_{θ_0}. From the sequential weak-compactness of norm-bounded sets, there exists a T in H_{θ_0} and a subsequence $\{k^{**}\}$ of $\{k^*\}$ such that $T_{k^{**}} \to T$ weakly in H_{θ_0} along the subsequence $\{k^{**}\}$, i.e.,

$$\int T_{k^{**}} f \, dP_{\theta_0} \to \int Tf \, dP_{\theta_0} \quad \text{for every function } f \text{ in } H_{\theta_0}.$$

In particular, setting $f = dP_\theta / dP_{\theta_0}$, we get

$$E_\theta(T_{k^{**}}) \to E_\theta(T),$$

along the subsequence $\{k^{**}\}$, for all θ in Θ. From (1), it now follows that $E_\theta(T) = \phi(\theta)$, that is T is unbiased for ϕ, which contradicts one of our assumptions. Hence (2) holds and the proof is complete. □

There are many examples of situations to which this theorem applies. One class can be obtained from the idea of the following example. Consider the family of Poisson distributions with parameter λ with $\lambda > 0$. It is well known that there exists no unbiased estimator of $1/\lambda$ and that all polynomials in λ are unbiasedly estimable. From (a slight modification of) the Stone–Weierstrass theorem, there exists a sequence of polynomials in λ which converge to $1/\lambda$ for each λ. Thus there exists a sequence of estimators which are unbiased for these polynomials in λ and whose biases in estimating $1/\lambda$ converge to 0. A simple calculation shows that $\int (dP_{\lambda_1}/dP_{\lambda_2})^2 \, dP_{\lambda_2} = \exp(\lambda_2 - 2\lambda_1 + \lambda_1^2/\lambda_2)$. Thus Theorem 1 applies to this case and the variances of these estimators must tend to ∞.

It may appear that Theorem 1 does not apply to estimates based on the jackknife, since the "delete-one" jackknife can be formed only a finite number of times. However, a situation with an infinite sequence of estimators based on the jackknife arises in the following example, based on an idea of Gaver and Hoel (1970). Suppose that the data consists of a Poisson process $\{N(t);\, t \in [0,1]\}$ with rate λ. In connection with the biased maximum likelihood estimator $\hat{\phi} = e^{-\lambda N(1)}$ of $e^{-\lambda}$, Gaver and Hoel suggest splitting the interval $[0,1]$ into n nonoverlapping intervals each of length $1/n$, and letting N_i be the number of events in the ith interval. These are independent and identically distributed and one can therefore form the delete-one jackknife as usual. This yields, for each n, an estimate $\hat{\phi}_{(n)}$ and they show that as $n \to \infty$, $\hat{\phi}_{(n)}$ converges to an estimate $\hat{\phi}^{(1)}$ which depends on the Poisson process only through the sufficient statistic $N(1)$. This procedure can be repeated indefinitely in principle, giving a sequence of estimators $\{\hat{\phi}^{(k)}\}_{k=1}^{\infty}$.

REFERENCES

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

GAVER, D. P., JR. and HOEL, D. G. (1970). Comparison of certain small-sample Poisson probability estimates. *Technometrics* **12** 835–850.

MILLER, R. G. (1974). The jackknife—A review. *Biometrika* **61** 1–15.

DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306-3033