Bell Laboratories

subject: The Infinitesimal Jackknife

date: June 30, 1972
from: Louis A. Jaeckel
MM 72-1215-11

Memorandum for File

1 Introduction

The jackknife is a general-purpose tool for estimating the variance, and reducing the bias, of a wide variety of estimates. The basic idea of the jackknife is to omit one observation and recompute the estimate using the remaining observations. This is done for each observation, and the results are then combined to produce a variance estimate or bias correction. If we assign each observation a weight, then omitting an observation is the same as giving it a weight of zero. Instead of doing that, we shall give the observation only slightly less weight than the others, and consider the limiting case as this deficiency in the weight approaches zero. We call the result of this process the "infinitesimal jackknife", or *IJK* for short. We thus have a set of procedures analogous to the ordinary jackknife, or *OJK*.

It turns out that for many estimates the two kinds of jackknife have the same asymptotic behavior. However, to obtain asymptotic results, it seems more natural to work with the *IJK*. Furthermore, the *IJK* point of view gives us deeper insight into the nature of the jackknife process, and thereby helps us to see how to apply the jackknife idea in more complex situations.

We begin in Section 2 with the one-sample case, where we have n i.i.d. random variables and a statistic T, a function which is symmetric in its arguments. We define the *IJK* in this case, and in Section 3 we give some examples. We then discuss its asymptotic properties in Section 4, under the assumption that the distribution of the observations is discrete, with a certain form. We show that if T satisfies certain conditions, then both the *IJK* and the *OJK* give consistent estimates of the asymptotic variance and "asymptotic bias" of T. Then in Section 5 we consider some of the problems which arise when the observations lie in a more general space. In Section 6 we discuss some extensions and open questions, and finally, in Section 7 we consider a more general model, which includes the case of linear regression.

2 The one-sample case

Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution F. Although we shall sometimes consider real-valued observations, the X_i could be more general random variables in most of this work. We want to estimate some unknown real parameter θ of F. We want to think of θ and the procedure for estimating it as being somehow independent of n, the sample size. That is, for each n we must have a function $\hat{\theta}$ for estimating θ, symmetric in its n arguments, such that all of these functions "do the same thing" to the observations. A broad class of such estimates may be characterized as follows:

Suppose we may write $\theta = T(F)$, where T is a real-valued functional defined on some appropriate set of probability distributions including F and a sufficiently rich set of probability distributions "near" F. Suppose further that for each n, θ is estimated by $\hat{\theta} = T(\hat{F})$, where $\hat{F}$ is the empirical probability distribution (usually denoted by F_n). That is, $\hat{F}$ assigns probability $1/n$ to each of the X_i. We write $\hat{F}$ to de-emphasize the dependence on sample size and to indicate that $\hat{F}$ is to be thought of as an estimate of F. We now have a sequence of estimates whose definitions are not explicitly dependent on n. Since $\hat{F}$ approaches F in some sense as n increases, $\hat{\theta} = T(\hat{F})$ will approach $\theta = T(F)$, if T is well behaved. We shall assume that $\hat{\theta}$ is consistent; that is, that $\hat{\theta}$ converges to θ in probability. We shall see that this is a natural framework for our problem. Such a framework has been used by von Mises (1947) and others.

We begin by defining the *OJK*. Let $\hat{\theta}$ be the estimate of θ based on $X_1, \ldots, X_n$. If we omit X_i and estimate θ from the $n-1$ remaining observations, we call this estimate the i^{th} *pseudoestimate*, and we write it as $\hat{\theta}_{(i)}$. We also define $p_{(i)} = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$. This is known as the i^{th} *pseudovalue*. Although much of the literature deals primarily with the pseudovalues, we shall concentrate on the pseudoestimates. If we average over i we have the (ordinary) jackknifed estimate

$$p_{(\cdot)} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \ . \qquad \frac{1}{n}\sum_{i=1}^{n} p_{(i)} = p_{(\cdot)}$$

We observe that

$$p_{(i)} - p_{(\cdot)} = -(n-1)\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right) \ .$$

The jackknife does two things for us: It gives a variance estimate for $\hat{\theta}$, and it reduces the bias in $\hat{\theta}$. We estimate the variance of $\hat{\theta}$ (or of $p_{(\cdot)}$) by

$$\hat{V} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(p_{(i)} - p_{(\cdot)}\right)^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2 \ .$$

We reduce the bias by using $p_{(\cdot)}$ as our estimate, rather than $\hat{\theta}$. See Miller (1964). We can also think of this as estimating and subtracting off the bias term of order n^{-1} in $\hat{\theta}$. That is,

the estimated bias is

$$\hat{B} = \hat{\theta} - p_{(\cdot)} = (n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right) \ ,$$

and the jackknifed estimate is

$$\hat{\theta} - \hat{B} \ .$$

We now define the *IJK*. If we attach a weight to each observation, then we can think of omitting an observation as the same as giving that observation a weight of zero. Instead of doing that, as the *OJK* does, we shall give the observation only slightly less weight than the others, ~~and consider the limiting case as the deficiency in the weight approaches zero.~~ We assign weights $w_1, \ldots, w_n$ to $X_1, \ldots, X_n$, respectively. We shall see below that it is not necessary that $\sum w_i = 1$. We assume T is defined for discrete probability distributions, by which we mean distributions which concentrate all of their mass on a finite number of points. In particular, we wish to evaluate T at distributions which assign arbitrary probabilities $w_1, \ldots, w_n$ to the points $X_1, \ldots, X_n$, respectively. We can then write T as a function of $2n$ variables:

$$T(X_1, \ldots, X_n; w_1, \ldots, w_n) \ . \tag{1}$$

If all of the w_i are $1/n$, we have

$$\hat{\theta} = T(\hat{F}) = T\left(X_1, \ldots, X_n; \tfrac{1}{n}, \ldots, \tfrac{1}{n}\right) \ .$$

To simplify some of the algebra, we extend the definition of T as follows. If G is any probability distribution for which T is defined, and c is a positive constant, then we let $T(cG) = T(G)$. Thus, if we are considering discrete distributions with possible values $z_i,\ i = 1, \ldots, I$, we can assign to each z_i the weight w_i without requiring that $\sum w_i = 1$, and T will be defined for these weights. The following lemma will be useful.

Lemma 1. Let G be a discrete distribution, defined by $P(X = z_i) = g_i$, $i = 1, \ldots, I$. We assume the z_i are distinct. Let w_i be a variable weight, attached to the value z_i. Let

$$D_i^G = \left.\frac{\partial T}{\partial w_i}\right|_{w_j = g_j,\ j=1,\ldots,I}$$

and

$$D_{ji}^G = \left.\frac{\partial^2 T}{\partial w_j \partial w_i}\right|_{w_k = g_k,\ k=1,\ldots,I} \ .$$

Then, if the respective derivatives exist,

$$\sum_{i=1}^{I} g_i D_i^G = 0 \qquad \text{and} \qquad \sum_{i=1}^{I}\sum_{j=1}^{I} g_i g_j D_{ji}^G = 0 \ .$$

Proof. Since $T(cG) = T(G)$, if we let $w_i = cg_i$, we have

$$0 = \frac{dT(cg_1, \ldots, cg_I)}{dc} = \sum_i \frac{\partial T}{\partial w_i} \frac{dw_i}{dc} = \sum_i g_i \frac{\partial T}{\partial w_i} .$$

Evaluating at $c = 1$, we obtain the first result. If we differentiate the expression above once again with respect to c, we obtain

$$\sum_i g_i \sum_j g_j \frac{\partial^2 T}{\partial w_j \partial w_i} = 0 .$$

Evaluating at $c = 1$, we obtain the second result.

If we reduce w_i by ϵ and leave the other weights at $1/n$, we have

$$\hat{\theta}_{(i)}(\epsilon) = T\left(X_1, \ldots, X_n; \frac{1}{n}, \ldots, \frac{1}{n} - \epsilon, \frac{1}{n}, \ldots\right) .$$

If $\epsilon = 1/n$, then $w_i = 0$ and we have $\hat{\theta}_{(i)}$ as defined earlier. Assume we can differentiate T with respect to w_i. (That is, we let $\epsilon \to 0$). Let

$$\hat{D}_i = \left.\frac{\partial T}{\partial w_i}\right|_{x_j = X_j,\, w_j = \frac{1}{n},\, j=1,\ldots,n} .$$

Similarly, let

$$\hat{D}_{ii} = \left.\frac{\partial^2 T}{\partial w_i^2}\right|_{x_j = X_j,\, w_j = \frac{1}{n},\, j=1,\ldots,n} .$$

Since T is defined even if the sum of the weights is not one, we can perform these differentiations. Note that we are differentiating with respect to the weight w_i, rather than the value X_i. We can now form the Taylor series expansion

$$\begin{aligned} \hat{\theta}_{(i)}(\epsilon) - \hat{\theta} &= T\left(\ldots, \frac{1}{n} - \epsilon, \ldots\right) - T\left(\ldots, \frac{1}{n}, \ldots\right) \\ &= -\epsilon \hat{D}_i + \frac{\epsilon^2}{2} \hat{D}_{ii} - \ldots . \end{aligned} \tag{2}$$

We can define a variance estimate $\hat{V}(\epsilon)$ by

$$n^2 \epsilon^2 \hat{V}(\epsilon) = (1 - \epsilon) \sum \left[\hat{\theta}_{(i)}(\epsilon) - \hat{\theta}_{(\cdot)}(\epsilon)\right]^2 .$$

If $\epsilon = 1/n$, this is the ordinary jackknife $\hat{V}$ defined above. Using the shortened expansion

$$\hat{\theta}_{(i)}(\epsilon) \cong \hat{\theta} - \epsilon \hat{D}_i \,,$$

we see that

$$\hat{\theta}_{(\cdot)}(\epsilon) \cong \hat{\theta} - \frac{\epsilon}{n} \sum \hat{D}_i = \hat{\theta} \,,$$

because by Lemma 1,

$$\sum \frac{1}{n} \hat{D}_i = 0 \,.$$

So

$$\hat{\theta}_{(i)}(\epsilon) - \hat{\theta}_{(\cdot)}(\epsilon) \cong -\epsilon \hat{D}_i$$

and

$$n^2 \epsilon^2 \hat{V}(\epsilon) \cong (1-\epsilon) \sum \epsilon^2 \hat{D}_i^2 \,.$$

Letting $\epsilon \to 0$, we have

$$n\hat{V}(0) = \frac{1}{n} \sum \hat{D}_i^2 \,.$$

$\hat{V}(0)$ is the *IJK* variance estimate for $\hat{\theta}$.

We can define a bias estimate $\hat{B}(\epsilon)$ by

$$n^2 \epsilon^2 \hat{B}(\epsilon) = n(1-\epsilon) \left(\hat{\theta}_{(\cdot)}(\epsilon) - \hat{\theta} \right) \,.$$

If $\epsilon = 1/n$, this is the $\hat{B}$ defined above. We write

$$\hat{\theta}_{(i)}(\epsilon) \cong \hat{\theta} - \epsilon \hat{D}_i + \frac{\epsilon^2}{2} \hat{D}_{ii} \,.$$

Now, since $\sum \hat{D}_i = 0$,

$$\hat{\theta}_{(\cdot)}(\epsilon) \cong \hat{\theta} + \frac{\epsilon^2}{2n} \sum_{i=1}^{n} \hat{D}_{ii} \,,$$

so

$$n^2 \epsilon^2 \hat{B}(\epsilon) \cong n(1-\epsilon) \left(\frac{\epsilon^2}{2n} \sum \hat{D}_{ii} \right) \,.$$

Letting $\epsilon \to 0$, we have

$$n\hat{B}(0) = \frac{1}{2n} \sum \hat{D}_{ii} \,.$$

$\hat{B}(0)$ is the *IJK* estimate of the bias in $\hat{\theta}$. To be precise, it is an estimate of the component of the bias which is of order n^{-1}. The jackknifed estimate is then

$$\hat{\theta} - \hat{B}(0) \,.$$

3 Some examples.

If we want to apply the *IJK* to a given estimate, we must first write the estimate as a function not only of the n observations, but also of the n weights. In order to do this we must think of the estimate as a functional to be evaluated not only at $\hat{F}$, but also at discrete distributions with arbitrary weights. In other words, we must remove the explicit dependence on n usually present in the definition of an estimate. The new definition must also satisfy $T(cG) = T(G)$.

To illustrate the variance and the bias aspects of the *IJK*, we consider the sample variance, which we define as

$$\hat{\theta} = T = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 .$$

We must write this as a function of the X_i and weights w_i. We begin with the sample mean, which we rewrite as

$$M(X_1, \ldots, X_n; w_1, \ldots, w_n) = \frac{\sum w_i X_i}{\sum w_i} .$$

Note that $M(\hat{F}) = \int x d\hat{F}(x) = \bar{X}$ and $M(F) = \int x dF(x) = E(X)$. Similarly, we can write T as

$$T(X_1, \ldots, X_n; w_1, \ldots, w_n) = \frac{1}{\sum w_i} \sum w_i \left(X_i - M \right)^2 .$$

We also have $T(\hat{F}) = \int \left[x - M(\hat{F}) \right]^2 d\hat{F}(x) = \hat{\theta}$ and $T(F) = \int \left[x - M(F) \right]^2 dF(x) = \mathrm{var} X$.

We now differentiate T. We obtain

$$\frac{\partial T}{\partial w_k} = \frac{1}{\sum w_i} \left\{ \sum w_i \, 2(X_i - M) \left(-\frac{\partial M}{\partial w_k} \right) + (X_k - M)^2 \right\} - \frac{1}{\left(\sum w_i \right)^2} \sum w_i \left(X_i - M \right)^2 .$$

Since $\sum w_i X_i = M \sum w_i$, the first expression in the brackets vanishes, and we have

$$\frac{\partial T}{\partial w_k} = \frac{1}{\sum w_i} \left[(X_k - M)^2 - T \right] .$$

Hence,

$$\hat{D}_k = \left(X_k - \bar{X} \right)^2 - \frac{1}{n} \sum_i \left(X_i - \bar{X} \right)^2 ,$$

and our estimate of the variance of $\hat{\theta}$ is

$$\hat{V}(0) = \frac{1}{n^2} \sum_k \left[\left(X_k - \bar{X} \right)^2 - \frac{1}{n} \sum_i \left(X_i - \bar{X} \right)^2 \right]^2 .$$

If we differentiate again, we obtain

$$\frac{\partial^2 T}{\partial w_k^2} = \frac{1}{\sum w_i}\left\{2\left(X_k - M\right)\left(-\frac{\partial M}{\partial w_k}\right) - \frac{\partial T}{\partial w_k}\right\} - \frac{1}{\left(\sum w_i\right)^2}\left[\left(X_k - M\right)^2 - T\right] \ .$$

Since

$$\frac{\partial M}{\partial w_k} = \frac{1}{\sum w_i}\left(X_k - M\right) \ ,$$

we find that

$$\hat{D}_{kk} = -4\left(X_k - \bar{X}\right)^2 + \frac{2}{n}\sum_i \left(X_i - \bar{X}\right)^2 \ .$$

If we fudge the definition of $\hat{B}(0)$ a little, by replacing n by $n-1$ in the denominator, we have

$$\hat{B}(0) = \frac{1}{2n(n-1)}\sum \hat{D}_{kk} = -\frac{1}{n(n-1)}\sum \left(X_i - \bar{X}\right)^2 \ .$$

The jackknifed estimate is then

$$\begin{aligned} \hat{\theta} - \hat{B}(0) &= \left(\frac{1}{n} + \frac{1}{n(n-1)}\right)\sum \left(X_i - \bar{X}\right)^2 \\ &= \frac{1}{n-1}\sum \left(X_i - \bar{X}\right)^2 \ , \end{aligned}$$

which we know to be exactly unbiased. If we had used our original definition of $\hat{B}(0)$ instead, the bias would not have been eliminated completely. But the remaining bias would only have been of order n^{-2}; the bias of order n^{-1} would have been removed, which is all the *IJK* is intended to do. Actually, the multiplicative constants used in the definitions of $\hat{V}(0)$ and $\hat{B}(0)$ were chosen somewhat arbitrarily. Further work is needed to determine the best constants to use.

We give one more example of the *IJK* variance estimate. Suppose $\hat{\theta}$ is defined to be the root of the equation

$$0 = \frac{1}{n}\sum_{i=1}^{n} h\left(X_i, \hat{\theta}\right) \ ,$$

where $h(x,\theta)$ is some given function. If $h(x,\theta) = \frac{\partial}{\partial\theta}\log f(x,\theta)$, where $f(x,\theta)$ is the density of the X_i, then $\hat{\theta}$ is the maximum likelihood estimate of θ. However, this type of estimate may occur in other contexts; see for example Huber (1964). We define T, using weights, as the root of

$$0 = \sum w_i\, h(X_i, T) \ . \tag{3}$$

We can write $\hat{\theta} = T(\hat{F})$, since $\hat{\theta}$ is defined by $0 = \int h(x, \hat{\theta}) d\hat{F}(x)$. The parameter to be estimated is $\theta = T(F)$, where θ satisfies $0 = \int h(x, \theta) dF(x)$.

If we differentiate in (3) and let $h_2(x, \theta) = \frac{\partial}{\partial \theta} h(x, \theta)$, we obtain

$$0 = \sum w_i \, h_2(X_i, T) \frac{\partial T}{\partial w_k} + h(X_k, T) \, ,$$

so

$$\frac{\partial T}{\partial w_k} = - \frac{h(X_k, T)}{\sum w_i \, h_2(X_i, T)}$$

and

$$\hat{D}_k = - \frac{h(X_k, T)}{\frac{1}{n} \sum h_2(X_i, T)} \, .$$

Therefore,

$$n \hat{V}(0) = \frac{\frac{1}{n} \sum h^2(X_k, T)}{\left\{ \frac{1}{n} \sum h_2(X_i, T) \right\}^2} \, .$$

Under appropriate conditions this quantity will converge to the asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta)$. See the asymptotic variance formulas given in Huber (1964), and in Brillinger (1964), in which the *OJK* is applied to the maximum likelihood estimate.

4 Asymptotic behavior of the IJK and OJK for discrete F.

In this section we shall assume that F is discrete, with the form given below. This case will be easier to deal with than the general case, which we shall discuss in the next section. The basic concepts involved will emerge in our treatmemt of the discrete case.

We assume that under F, X can take on the M distinct values z_j, each with probability $1/M$. We describe a random sample of size n from F as follows. For each j, $j = 1, \ldots, M$, we define a random variable $\hat{W}_j$ by

$$\hat{W}_j = \frac{1}{n} \{ \text{number of times } z_j \text{ occurs in the sample} \} \, .$$

The vector $(n\hat{W}_1, \ldots, n\hat{W}_M)$ has a multinomial distribution, and $\sum \hat{W}_j = 1$. It is easy to verify that

$$E \, \hat{W}_j = \frac{1}{M} \, ,$$

$$\operatorname{Var} \hat{W}_j = \frac{1}{nM} \left(1 - \frac{1}{M} \right) \, ,$$

and

$$\text{Cov}\left(\hat{W}_j, \hat{W}_k\right) = -\frac{1}{nM^2} \; .$$

If we write in vector notation

$$\mathbf{w} = (W_1, \ldots, W_M) \, ,$$

then we can write

$$\mathbf{w}_0 = \left(\frac{1}{M}, \ldots, \frac{1}{M}\right)$$

and

$$\hat{\mathbf{w}} = \left(\hat{W}_1, \ldots, \hat{W}_M\right) \; .$$

As $n \to \infty$, the vector

$$\sqrt{n}\,(\hat{\mathbf{w}} - \mathbf{w}_0) = \left(\sqrt{n}\left(\hat{W}_1 - \frac{1}{M}\right), \ldots, \sqrt{n}\left(\hat{W}_M - \frac{1}{M}\right)\right)$$

has a multivariate normal limiting distribution, for which all means are zero, all variances are $\frac{1}{M}(1 - \frac{1}{M})$, and all covariances are $-\frac{1}{M^2}$.

We assume T is defined for discrete distributions which assign arbitrary non-negative weights W_j to the z_j. Letting $\mathbf{z} = (z_1, \ldots, z_M)$, we can write the estimate as

$$\hat{\theta} = T(\hat{F}) = T(\mathbf{z}, \hat{\mathbf{w}}) \; . \tag{4}$$

We also have

$$\theta = T(F) = T(\mathbf{z}, \mathbf{w}_0) \; .$$

Since the z_j are fixed, we shall think of $T(\mathbf{z}, \mathbf{w})$ as a function of the M variables W_j. Although the $\hat{W}_j$ actually take on only certain rational values, we shall consider the W_j to be continuous variables, with $\sum W_j$ not necessarily equal to one, so that we can differentiate T with respect to these variables. As before, we assume that $T(cG) = T(G)$. Note that although (4) and (1) appear similar in form, in that each describes T as a function of values and weights, there is a very important difference between them. In (1) the X_i are random variables and the w_i are constants, whereas in (4) it is the z_j that are fixed, while the $\hat{W}_j$, are the random variables.

Differentiating T with respect to the weights, we have

$$D_j = \left.\frac{\partial T}{\partial W_j}\right|_{\mathbf{w}=\mathbf{w}_0}$$

and

$$D_{jk} = \left.\frac{\partial^2 T}{\partial W_j \partial W_k}\right|_{\mathbf{w}=\mathbf{w}_0} \; .$$

Note the difference between the D_j and D_{jj} defined here, and the $\hat{D}_i$ and $\hat{D}_{ii}$ defined in Section 2. D_j and D_{jj} are derivatives of T evaluated at F, whereas $\hat{D}_i$ and $\hat{D}_{ii}$ are derivatives evaluated at $\hat{F}$. By Lemma 1,

$$\sum_{j=1}^{M} D_j = 0 \qquad \text{and} \qquad \sum_{j=1}^{M}\sum_{k=1}^{M} D_{jk} = 0 \; .$$

We use these derivatives to form the following Taylor series expansion:

$$\begin{aligned} T(\hat{F}) - T(F) &= T(\mathbf{z}, \hat{\mathbf{w}}) - T(\mathbf{z}, \mathbf{w}_0) \qquad (5) \\ &= \sum_{j=1}^{M} \left(\hat{W}_j - \frac{1}{M}\right) D_j + \frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{M} \left(\hat{W}_j - \frac{1}{M}\right)\left(\hat{W}_k - \frac{1}{M}\right) D_{jk} + \dots \\ &= S_1 + \frac{1}{2} S_2 + \dots \; , \end{aligned}$$

where S_1 and S_2 are the respective sums above.

We begin with the variance of $T(\hat{F})$. If T is well behaved, we may expect that the following approximation is valid for large n:

$$T(\hat{F}) - T(F) \cong S_1 \; .$$

By the definition of S_1 we see that $ES_1 = 0$ and

$$\begin{aligned} \mathrm{Var} S_1 &= \sum_j \sum_k \mathrm{Cov}\left(\hat{W}_j - \frac{1}{M}, \hat{W}_k - \frac{1}{M}\right) D_j D_k \\ &= \sum_j \frac{1}{nM}\left(1 - \frac{1}{M}\right) D_j^2 + \sum\sum_{j \neq k} \left(-\frac{1}{nM^2}\right) D_j D_k \\ &= \frac{1}{nM}\sum_j D_j^2 - \frac{1}{nM^2}\sum_j \sum_k D_j D_k \; . \end{aligned}$$

Since the last sum is $\left(\sum D_j\right)^2 = 0$, we have

$$\mathrm{Var}\left(\sqrt{n}\, S_1\right) = \frac{1}{M}\sum_j D_j^2 = V \; .$$

We assume $V > 0$. Since, as we remarked earlier, $\sqrt{n}\,(\hat{\mathbf{w}} - \mathbf{w}_0)$ is asymptotically multivariate normal, we see that $\sqrt{n}\, S_1$ is asymptotically normal $(0, V)$. If the approximation above may be used, then $\sqrt{n}\,\left[T(\hat{F}) - T(F)\right]$ has this same limiting distribution.

We now consider the *IJK* variance estimate. Suppose the value of X_i is z_j. Then we can write $X_i = z_{j(i)}$; that is, $j(i)$ is a function of i, determined by the sample. We replace the $\hat{D}_i$ of Section 2 by $\hat{D}_{j(i)}$, where $\hat{D}_j$ is defined here as

$$\hat{D}_j = \left.\frac{\partial T}{\partial W_j}\right|_{\mathbf{w}=\hat{\mathbf{w}}} .$$

The two terms above are the same because they are both the derivative of T with respect to the weight attached to X_i. Using this terminology and the definition of $\hat{W}_j$, we can write

$$n\hat{V}(0) = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_{j(i)}^2 = \sum_{j=1}^{M} \hat{W}_j \hat{D}_j^2 .$$

Since $\hat{\mathbf{w}}$ is near $\mathbf{w}_0$ for large n, we may expect $\hat{D}_j$ to be near D_j. If so, then the last sum above will be near V. In other words, $n\hat{V}(0)$ is an approximation to the asymptotic variance of $\sqrt{n}T(\hat{F})$. Although the *IJK* thus gives us an estimate of the asymptotic variance rather than the actual finite sample size variance, for most applications we may regard $\hat{V}(0)$ as an estimate of the actual variance of $T(\hat{F})$. We shall see that the *OJK* variance estimate, which is a sum of squares of differences instead of derivatives, also gives us an estimate of V. We now state these results formally as a theorem.

Theorem 1. *Suppose for each* $j = 1, \ldots, M$, $\frac{\partial T}{\partial W_j}$ *exists and is a continuous function of* $\mathbf{w}$ *for all* M*-vectors* $\mathbf{w}$ *in some convex neighborhood* U *of* $\mathbf{w}_0$*. Then, as* $n \to \infty$*:*

(i) $\sqrt{n}\left[T(\hat{F}) - T(F)\right]$ *is asymptotically normal with mean zero and variance*

$$V = \frac{1}{M}\sum_{j=1}^{M} D_j^2 .$$

(ii) n *times the IJK variance estimate converges in probability to* V*. That is,*

$$n\hat{V}(0) = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_i^2 \xrightarrow{P} V .$$

(iii) n *times the OJK variance estimate converges in probability to* V*. That is,*

$$n\hat{V} = (n-1)\sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2 \xrightarrow{P} V .$$

Proof. Note that U is a neighborhood in the full M-dimensional space of vectors $\mathbf{w}$, rather than in the set of vectors for which $\sum W_j = 1$. Since we assumed $T(c\mathbf{w}) = T(\mathbf{w})$, we can work with T and its derivatives without regard to whether this sum is one.

(i) We shall show that

$$\sqrt{n}\left[T(\hat{F}) - T(F)\right] - \sqrt{n}\sum_{j=1}^{M}\left(\hat{W}_j - \frac{1}{M}\right)D_j \xrightarrow{P} 0\ .$$

This will imply that the two terms above have the same limiting distribution. We saw above that this sum is asymptotically normal $(0, V)$.

Consider the line segment from $\mathbf{w}_0$ to $\hat{\mathbf{w}}$. We parametrize this segment as

$$\hat{\mathbf{w}}(t) = \mathbf{w}_0 + t\,(\hat{\mathbf{w}} - \mathbf{w}_0)$$

with $0 \le t \le 1$, and we consider $T[\hat{\mathbf{w}}(t)]$ as a function of t. Since $\hat{\mathbf{w}} \xrightarrow{P} \mathbf{w}_0$, we know that $\hat{\mathbf{w}}$, and hence the entire line segment, lie in U with probability approaching one. If we assume $\hat{\mathbf{w}}$ lies in U, then by the mean value theorem there is a $\hat{t}$ between 0 and 1 such that

$$\begin{aligned} T(\hat{F}) - T(F) &= T[\hat{\mathbf{w}}(1)] - T[\hat{\mathbf{w}}(0)] \\ &= \left.\frac{dT[\hat{\mathbf{w}}(t)]}{dt}\right|_{t=\hat{t}} \\ &= \sum_{j=1}^{M}\left(\hat{W}_j - \frac{1}{M}\right)\hat{D}_j(\hat{t})\,, \end{aligned}$$

where

$$\hat{D}_j(\hat{t}) = \left.\frac{\partial T}{\partial W_j}\right|_{\mathbf{w}=\hat{\mathbf{w}}(\hat{t})}\ .$$

(Note that $\hat{t}$ is a random variable.) Hence,

$$\sqrt{n}\left[T(\hat{F}) - T(F)\right] - \sqrt{n}\sum_{j}\left(\hat{W}_j - \frac{1}{M}\right)D_j = \sum_{j}\sqrt{n}\left(\hat{W}_j - \frac{1}{M}\right)\left(\hat{D}_j(\hat{t}) - D_j\right)\ .$$

Since $\hat{\mathbf{w}} \xrightarrow{P} \mathbf{w}_0$ and $0 \le \hat{t} \le 1$, we have $\hat{\mathbf{w}}(\hat{t}) \xrightarrow{P} \mathbf{w}_0$. So, since $P(\hat{\mathbf{w}} \in U) \to 1$ and $\frac{\partial T}{\partial W_j}$ is continuous for $\mathbf{w}$ in U, we have $\hat{D}_j(\hat{t}) \xrightarrow{P} D_j$ for each j. Since $\sqrt{n}\left(\hat{W}_j - \frac{1}{M}\right)$ is bounded in probability for each j, it follows that the sum on the right side above converges to zero in probability as $n \to \infty$, and (i) is proved.

(ii) We saw earlier that we could write

$$n\hat{V}(0) = \sum_{j}\hat{W}_j\hat{D}_j^2\ .$$

For the reasons given above in the proof of (i), we have $\hat{D}_j \xrightarrow{P} D_j$, and hence $\hat{D}_j^2 \xrightarrow{P} D_j^2$, for all j. Since $\hat{W}_j \xrightarrow{P} 1/M$, we see that

$$n\hat{V}(0) \xrightarrow{P} V \ .$$

We defer the proof of (iii) to the next section, where we shall show that it follows from a more general result.

We turn now to the bias in $T(\hat{F})$. We assume for the moment that the following approximation based on (5) is valid for large n:

$$T(\hat{F}) - T(F) \cong S_1 + \frac{1}{2}S_2 \ .$$

We take the expected value of the right side above. Since $ES_1 = 0$ and $\sum\sum D_{jk} = 0$, we have

$$\begin{aligned} E\left(S_1 + \frac{1}{2}S_2\right) &= \frac{1}{2}\sum_j\sum_k E\left(\hat{W}_j - \frac{1}{M}\right)\left(\hat{W}_k - \frac{1}{M}\right)D_{jk} \\ &= \frac{1}{2}\left\{\sum_j \frac{1}{nM}\left(1 - \frac{1}{M}\right)D_{jj} + \sum\sum_{j\neq k}\left(-\frac{1}{nM^2}\right)D_{jk}\right\} \\ &= \frac{1}{2}\left\{\frac{1}{nM}\sum_j D_{jj} - \frac{1}{nM^2}\sum_j\sum_k D_{jk}\right\} \\ &= \frac{1}{n}\frac{1}{2M}\sum_j D_{jj} \ . \end{aligned}$$

We call

$$B = \frac{1}{2M}\sum_j D_{jj}$$

the "asymptotic bias" of $T(\hat{F})$. If we can use the above approximation, we see that

$$ET(\hat{F}) \cong T(F) + \frac{B}{n} \ .$$

If we can estimate B/n and subtract this estimate from $T(\hat{F})$, we will have an estimate of $T(F)$ with a bias of order less than n^{-1}. We rewrite the *IJK* bias estimate as we did the variance estimate by replacing $\hat{D}_{ii}$ by $\hat{D}_{j(i),j(i)}$, where

$$\hat{D}_{jj} = \left.\frac{\partial^2 T}{\partial W_j^2}\right|_{\mathbf{w}=\hat{\mathbf{w}}} \ .$$

We can then write

$$n\hat{B}(0) = \frac{1}{2n}\sum_{i=1}^{n} \hat{D}_{j(i),j(i)} = \frac{1}{2}\sum_{j=1}^{M} \hat{W}_j \hat{D}_{jj} \ .$$

As with the variance estimate, we may expect the last sum above to be an estimate of B. We shall see that the *OJK* bias estimate is also an estimate of B.

Theorem 2. *Suppose for each* $j, k = 1, \ldots, M$, $\frac{\partial T}{\partial W_j}$ *and* $\frac{\partial^2 T}{\partial W_j \partial W_k}$ *exist and are continuous functions of* $\mathbf{w}$ *for all* M*-vectors* $\mathbf{w}$ *in some convex neighborhood* U *of* $\mathbf{w}_0$*. Then, as* $n \to \infty$*:*

(i)

$$n\left\{T(\hat{F}) - T(F) - \sum_{j=1}^{M}\left(\hat{W}_j - \frac{1}{M}\right) D_j\right\}$$

has a limiting distribution whose expectation is

$$B = \frac{1}{2M}\sum_{j=1}^{M} D_{jj} \ .$$

(ii) n *times the IJK bias estimate converges in probability to* B*. That is,*

$$n\hat{B}(0) = \frac{1}{2n}\sum_{i=1}^{n} \hat{D}_{ii} \xrightarrow{P} B \ .$$

(iii) n *times the OJK bias estimate converges in probability to* B*. That is,*

$$n\hat{B} = n(n-1)\left(\hat{\theta}_{(\cdot)} - \hat{\theta}\right) \xrightarrow{P} B \ .$$

Proof. (i) We parametrize the line segment from $\mathbf{w}_0$ to $\hat{\mathbf{w}}$ as before:

$$\hat{\mathbf{w}}(t) = \mathbf{w}_0 + t\left(\hat{\mathbf{w}} - \mathbf{w}_0\right) \ ,$$

with $0 \le t \le 1$. We then expand in a Taylor series:

$$\begin{aligned} T(\hat{F}) - T(F) &= T\left[\hat{\mathbf{w}}(1)\right] - T\left[\hat{\mathbf{w}}(0)\right] \\ &= \left.\frac{dT}{dt}\right|_{t=0} + \frac{1}{2}\left.\frac{d^2T}{dt^2}\right|_{t=\hat{t}} \\ &= \sum_{j=1}^{M}\left(\hat{W}_j - \frac{1}{M}\right) D_j + \frac{1}{2}\sum_{j=1}^{M}\sum_{k=1}^{M}\left(\hat{W}_j - \frac{1}{M}\right)\left(\hat{W}_k - \frac{1}{M}\right)\hat{D}_{jk}(\hat{t}) \ , \end{aligned}$$

where $0 \le \hat{t} \le 1$ and

$$\hat{D}_{jk}(\hat{t}) = \left. \frac{\partial^2 T}{\partial W_j \partial W_k} \right|_{\mathbf{w}=\hat{\mathbf{w}}(\hat{t})} .$$

Hence

$$n \left\{ T(\hat{F}) - T(F) - \sum_j \left(\hat{W}_j - \frac{1}{M} \right) D_j \right\} - \frac{n}{2} \sum_j \sum_k \left(\hat{W}_j - \frac{1}{M} \right) \left(\hat{W}_k - \frac{1}{M} \right) D_{jk} \qquad (6)$$

$$= \frac{n}{2} \sum_j \sum_k \left(\hat{W}_j - \frac{1}{M} \right) \left(\hat{W}_k - \frac{1}{M} \right) \left(\hat{D}_{jk}(\hat{t}) - D_{jk} \right) .$$

Since $\hat{\mathbf{w}} \xrightarrow{P} \mathbf{w}_0$, $P(\hat{\mathbf{w}} \in U) \to 1$, $0 \le \hat{t} \le 1$, and the derivatives of T are continuous for $\mathbf{w}$ in U, we have $\hat{D}_{jk}(\hat{t}) \xrightarrow{P} D_{jk}$. Therefore, since for each j and k, $n \left(\hat{W}_j - \frac{1}{M} \right) \left(\hat{W}_k - \frac{1}{M} \right)$ is bounded in probability, the sum on the right side of (6) converges to zero in probability as $n \to \infty$.

The second expression on the left side of (6) has a limiting distribution because it is a fixed combination of a fixed number of random variables which have a limiting joint distribution. It follows that the first expression on the left side of (6) has the same limiting distribution. By a computation similar to the computation of $\frac{1}{2} E S_2$ above, we find that the expectation of this limiting distribution is B, and (i) is proved.

Note that we are dealing here with the expectation of the limiting distribution, rather than with the limit of the expectations. However, we may in general use B as an approximation to the expectation of the first expression in (6).

(ii) For the reasons given above in the proof of (i) we have $\hat{D}_{jj} \xrightarrow{P} D_{jj}$ and $\hat{W}_j \xrightarrow{P} 1/M$. Therefore, if we rewrite $n\hat{B}(0)$ as we did earlier, we have

$$n\hat{B}(0) = \frac{1}{2} \sum_{j=1}^{M} \hat{W}_j \hat{D}_{jj} \xrightarrow{P} B .$$

We defer the proof of (iii) to the next section.

We remark that under the conditions of Theorem 2, $\sqrt{n} T(\hat{F})$, $\sqrt{n} \left[T(\hat{F}) - \hat{B}(0) \right]$, and $\sqrt{n} \left[T(\hat{F}) - \hat{B} \right]$ all have asymptotic variance V. Therefore our variance estimates may be used to estimate the asymptotic variance of the original estimate, or of either form of jackknifed estimate.

5 Asymptotic behavior for general F.

If F is not discrete, some complications arise. We need first a generalization of the concept of differentiating T with respect to the weight assigned to a possible value of X. We must then consider the relationship of the derivative of T evaluated at a distribution near F to the derivative of T evaluated at F. We shall follow the approach of von Mises (1947), who considered these problems in his work on differentiable statistical functions. We would like to have general theorems analogous to our Theorems 1 and 2 of Section 4. However, general proofs of the asymptotic properties of $T(\hat{F})$ would involve us in difficulties which are beyond the scope of this paper, so we shall give in each case what amounts to an outline of a proof. Then, for a given T and F, it will often be possible to fill in the missing steps for that particular example. We do, however, give fairly general proofs of the consistency of the jackknife estimates of variance and bias.

We define a class $\mathcal{F}$ of finite measures as follows. Let S be the set of possible values for X. S can be a fairly arbitrary set; it need not be a set of real numbers. For each x in S, let δ_x be the probability measure which assigns measure one to the point x. Let $\mathcal{F}$ be the set of all linear combinations of F and an arbitrary finite number of the δ_x measures. Let $\mathcal{F}^+$ be the set of positive measures in $\mathcal{F}$, not including the zero measure. We assume T is defined for probability measures in $\mathcal{F}^+$. We extend T to all of $\mathcal{F}^+$ by letting $T(cG) = T(G)$ for all $c > 0$. Note that $\mathcal{F}^+$ is convex and contains all possible $\hat{F}$.

We now define the derivative of T, essentially following von Mises (1947). We say T is differentiable at G in $\mathcal{F}^+$ if there exists a function $T'(G, x)$, defined at all x in S, with the following property: Let H be any member of $\mathcal{F}$ such that $G + tH$ is in $\mathcal{F}^+$ for all t in some interval $0 \le t \le t_H$, $t_H > 0$, so that $T(G + tH)$ is defined for t in this interval. Then, for any such H, $T'(G, x)$ satisfies

$$\begin{aligned} \left.\frac{dT(G+tH)}{dt}\right|_{t=0} &= \lim_{t\to 0} \frac{1}{t}\{T(G+tH) - T(G)\} \qquad (7) \\ &= \int T'(G,x)\, dH(x)\,. \end{aligned}$$

If we let $H = G$ in (7), we see that since $T(cG) = T(G)$,

$$\int T'(G,x)\, dG(x) = 0\,, \qquad (8)$$

as in Lemma 1. Now, if we let $H = \delta_{x_0} - G$ in (7) and apply (8), we find

$$\begin{aligned} \lim_{t\to 0} \frac{1}{t}\{T[G + t(\delta_{x_0} - G)] - T(G)\} &= \int T'(G,x)\, d(\delta_{x_0} - G)(x) \\ &= T'(G, x_0)\,. \end{aligned}$$

T' can often be found by a direct application of this equation. Hampel (1968) has defined $T'(G,x)$ in this way and has called it the "influence curve", since it expresses the influence on T of adding a small mass to G at x. We observe that the vector of derivatives $(D_1,\ldots,D_M)$ in Section 4 is the discrete analog of the function $T'(F,x)$.

Equation (7) gives the derivative of $T(G+tH)$ at $t=0$. It will be useful to have a similar expression for the derivative at an arbitrary t_0, $0<t_0<t_H$. If we let $G_{t_0}=G+t_0H$ and $u=t-t_0$, we can write $G+tH=(G+t_0H)+(t-t_0)H=G_{t_0}+uH$. Then, assuming T is differentiable at G_{t_0}, we have

$$\left.\frac{dT(G+tH)}{dt}\right|_{t=t_0}=\left.\frac{dT\left(G_{t_0}+uH\right)}{du}\right|_{u=0}=\int T'\left(G_{t_0},x\right)dH(x)\;.$$

We now assume that T is differentiable, in the sense defined above, at all G in some convex neighborhood of F in $\mathcal{F}^+$, such that $\hat{F}$ lies in the neighborhood with probability approaching one. Following the reasoning in the proof of Theorem 1, we parametrize the segment from F to $\hat{F}$ by

$$\hat{F}(t)=F+t(\hat{F}-F)$$

for $0\le t\le 1$. Then, if $\hat{F}$ lies in the neighborhood, we can write

$$\begin{aligned}T(\hat{F})-T(F) &= T\left[\hat{F}(1)\right]-T\left[\hat{F}(0)\right]\\ &= \left.\frac{dT[\hat{F}(t)]}{dt}\right|_{t=\hat{t}}=\int T'(\hat{F}(\hat{t}),x)d(\hat{F}-F)(x)\,,\end{aligned}$$

for some $0\le\hat{t}\le 1$. For large n, $\hat{F}$ is near F, so we would expect $T'\left(\hat{F}(\hat{t}),x\right)$ to be near $T'(F,x)$ in some sense. We write

$$\sqrt{n}\left[T(\hat{F})-T(F)-\int T'(F,x)d(\hat{F}-F)(x)\right]$$

$$=\sqrt{n}\int\left[T'(\hat{F}(\hat{t}),x)-T'(F,x)\right]d(\hat{F}-F)(x)\;.$$

We shall assume that the integral on the right side converges to zero in probability. This is the missing step which can often be verified for a particular example. (von Mises (1947) used a somewhat different method in his Theorem 1, p. 327.) By (8), we have

$$\sqrt{n}\int T'(F,x)d(\hat{F}-F)(x)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T'(F,X_i)\;.$$

Now this sum is a sum of i.i.d. random variables. Using (8) again, we have $ET'(F,X) = 0$. If we let

$$V = \int \left[T'(F,x)\right]^2 dF(x)$$

and assume that $0 < V < \infty$, we see that the sum above is asymptotically normal $(0, V)$. It follows from our assumptions that $\sqrt{n}\,[T(\hat{F}) - T(F)]$ has this same limiting distribution.

We would now like to estimate V. If we think of $\hat{F}$ as an estimate of F, and $T'(\hat{F}, x)$ as an estimate of $T'(F,x)$, then a natural estimate of V is

$$\int \left[T'(\hat{F},x)\right]^2 d\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \left[T'(\hat{F}, X_i)\right]^2 .$$

But this is exactly n times the *IJK* variance estimate, because in the notation of this section we can write

$$\begin{aligned} \hat{D}_i &= \lim_{t\to 0} \frac{1}{t}\left\{T(\hat{F} + t\delta_{X_i}) - T(\hat{F})\right\} \\ &= \int T'(\hat{F},x)\, d\delta_{X_i}(x) = T'(\hat{F}, X_i) . \end{aligned}$$

So the estimate above is $n^{-1}\sum \hat{D}_i^2 = n\hat{V}(0)$. If we knew the shape of $T'(F,x)$, the influence curve, in advance, we could use this knowledge to estimate V. But since F is unknown, and $T'(F,x)$ depends on both T and F, we generally do not know T' in advance. The key to the jackknife procedure is that it provides us with an estimate of T' when F is unknown.

We now give conditions under which the *IJK* and *OJK* variance estimates converge in probability to V. The theorem is formulated in a fairly abstract way.

Lemma 2. *Let $Z_1, \ldots, Z_n$ be i.i.d. non-negative random variables with $EZ < \infty$, and let $\bar{z}_n$ be their average. Then for any $z_0 > EZ$,*

$$P\left[\bar{z}_n < z_0\right] \to 1 \qquad as \qquad n \to \infty .$$

Proof. By the weak law of large numbers, $\bar{z}_n \xrightarrow{P} EZ$.
Therefore, $P[|\bar{z}_n - EZ| < z_0 - EZ] \to 1$. The result follows.

Definition. Let $\hat{F}_{(i)}$ be the discrete measure which assigns weight $1/n$ to each X_j, $j \neq i$. That is, $\hat{F}_{(i)} = \hat{F} - n^{-1}\delta_{X_i}$.

Theorem 3. *Suppose T is differentiable at F, and $V = \int T'(F,x)^2\, dF(x) < \infty$. Let $X_1, \ldots, X_n$ be i.i.d. with distribution F. Suppose that for all $\epsilon > 0$ there is a convex set U in $\mathcal{F}^+$ and a function $A(x)$ such that T is differentiable at all G in U and*

(a1) $P[\hat{F} \in U] \to 1 \quad as \quad n \to \infty,$

(a2) $P[\hat{F}_{(i)} \in U,\, i = 1, \ldots, n] \to 1 \quad as \quad n \to \infty,$

(b) *For all* $G \in U$ *and* $x \in S$*:*

$$|T'(G,x) - T'(F,x)| \le A(x),$$

and (c) $\int A^2(x) dF(x) \le \epsilon$.

Then as $n \to \infty$ *we have for the IJK*

$$n\hat{V}(0) \xrightarrow{P} V \ ,$$

and for the OJK

$$n\hat{V} \xrightarrow{P} V \ .$$

In many cases we can let $A(x) = \sqrt{\epsilon}$ for all x. An example where a non-constant $A(x)$ is needed is the sample variance, which we discussed in Section 3. The set U can often be defined in terms of some measure of distance in $\mathcal{F}^+$.

We need the following lemma.

Lemma 3. *Under the assumptions of the theorem, for all* $\epsilon > 0$*, there is a convex set* U *in* $\mathcal{F}^+$ *and a function* $C(x)$ *such that (a1) and (a2) above hold,* T *is differentiable at all* G *in* U*, and*

(d) *For all* $G \in U$ *and* $x \in S$*:*

$$|T'(G,x)^2 - T'(F,x)^2| \le C(x),$$

and (e) $\int C(x)\, dF(x) \le \frac{\epsilon}{2}$.

Proof. For a given ϵ, choose ϵ_1 so that $2\sqrt{\epsilon_1 V} + \epsilon_1 \le \epsilon/2$. Let U and A be the set and the function provided by the assumptions of the theorem for ϵ_1. Define $C(x)$ by $C(x) = 2A(x)|T'(F,x)| + A^2(x)$. Then for all $G \in U$ and $x \in S$, if we write $R(x) = T'(G,x) - T'(F,x)$, we have

$$T'(G,x) = T'(F,x) + R(x)$$

and

$$T'(G,x)^2 - T'(F,x)^2 = 2R(x)T'(F,x) + R^2(x) \ .$$

By (b),

$$\left|T'(G,x)^2 - T'(F,x)^2\right| \le 2A(x)\,|T'(F,x)| + A^2(x) = C(x)\,.$$

By (c),

$$\begin{aligned}\int C(x)\,dF(x) &= 2\int A(x)\,|T'(F,x)|\,dF(x) + \int A^2(x)\,dF(x) \\ &\le 2\left\{\int A^2(x)\,dF(x)\right\}^{1/2}\cdot\left\{\int T'(F,x)^2\,dF(x)\right\}^{1/2} + \epsilon_1 \\ &\le 2\sqrt{\epsilon_1 V} + \epsilon_1 \le \frac{\epsilon}{2}\,.\end{aligned}$$

Proof of the theorem. For the *IJK* variance estimate, we have $n\hat{V}(0) = n^{-1}\sum T'(\hat{F},X_i)^2$. Fix $\epsilon > 0$. We obtain U and $C(x)$ from Lemma 3. If we assume $\hat{F}\in U$, we have

$$\begin{aligned}\Delta &= \left|\frac{1}{n}\sum T'(\hat{F},X_i)^2 - \frac{1}{n}\sum T'(F,X_i)^2\right| \\ &\le \frac{1}{n}\sum\left|T'(\hat{F},X_i)^2 - T'(F,X_i)^2\right| \\ &\le \frac{1}{n}\sum C(X_i)\,.\end{aligned}$$

By (a1), (e) and Lemma 2, we have $P[\Delta<\epsilon]\to 1$ as $n\to\infty$. Therefore, $\Delta\xrightarrow{P}0$, so $n\hat{V}(0)$ and $n^{-1}\sum T'(F,X_i)^2$ have the same limiting distribution. Since the latter expression is an average of i.i.d. random variables, it converges in probability to its expected value, which is V. Therefore, $n\hat{V}(0)\xrightarrow{P}V$.

We now consider the *OJK* variance estimate. If we let $\bar{T} = n^{-1}\sum T(\hat{F}_{(i)})$, then we can write

$$\begin{aligned}n\hat{V} &= (n-1)\sum_{i=1}^{n}\left[T(\hat{F}_{(i)}) - \bar{T}\right]^2 \\ &= (n-1)\sum\left[T(\hat{F}_{(i)}) - T(\hat{F})\right]^2 - n(n-1)\left[\bar{T} - T(\hat{F})\right]^2\,.\end{aligned}$$

We consider the two components above separately. For each i, we parametrize the segment from $\hat{F}$ to $\hat{F}_{(i)}$ as

$$\hat{F}_{(i)}(t) = \hat{F} + t\left(\hat{F}_{(i)} - \hat{F}\right) = \hat{F} - \frac{t}{n}\,\delta_{X_i}\,,$$

where $0\le t\le 1$. Now fix $\epsilon>0$. We obtain U and $C(x)$ from Lemma 3. If we assume $\hat{F}$ and all $\hat{F}_{(i)}$, and hence all $\hat{F}_{(i)}(t)$, are in U, then for some $0\le\hat{t}_i\le 1$,

$$T(\hat{F}_{(i)}) - T(\hat{F}) = \left.\frac{dT\left[\hat{F}_{(i)}(t)\right]}{dt}\right|_{t=\hat{t}_i}$$

$$
\begin{aligned}
&= -\frac{1}{n}\int T'(\hat{F}_{(i)}(\hat{t}_i), x)d\delta_{X_i}(x) \\
&= -\frac{1}{n}T'(\hat{F}_{(i)}(\hat{t}_i), X_i) \ .
\end{aligned}
$$

So

$$
n\sum\left[T(\hat{F}_{(i)}) - T(\hat{F})\right]^2 = \frac{1}{n}\sum T'(\hat{F}_{(i)}(\hat{t}_i), X_i)^2 \ .
$$

If we let

$$
\Delta = \left|\frac{1}{n}\sum T'(\hat{F}_{(i)}(\hat{t}_i), X_i)^2 - \frac{1}{n}\sum T'(F, X_i)^2\right| \ ,
$$

then by (a2) and the same argument used above for the *IJK*, we see that $\Delta \xrightarrow{P} 0$, so that the two sums in its definition have the same limiting distribution. It follows that

$$
(n-1)\sum\left[T(\hat{F}_{(i)}) - T(\hat{F})\right]^2 \xrightarrow{P} V \ .
$$

We can now see the close relationship between the *IJK* and the *OJK*. We observe that the differences $T(\hat{F}_{(i)}) - T(\hat{F})$ in the sum above, times $-n$, are approximations to the respective derivatives $T'(\hat{F}, X_i)$, which occur in the *IJK* variance estimate.

Finally, we show that the second component of $n\hat{V}$ above converges to zero in probability. If we fix $\epsilon > 0$, we obtain U and $A(x)$ from the assumptions of the theorem. If we assume $\hat{F}$ and all $\hat{F}_{(i)}$ are in U, then, as before, we have

$$
T(\hat{F}_{(i)}) - T(\hat{F}) = -\frac{1}{n}T'(\hat{F}_{(i)}(\hat{t}_i), X_i) \ ,
$$

so we can write

$$
n\left[\bar{T} - T(\hat{F})\right] = \sum_{i=1}^{n}\left[T(\hat{F}_{(i)}) - T(\hat{F})\right] = -\frac{1}{n}\sum T'(\hat{F}_{(i)}(\hat{t}_i), X_i) \ .
$$

Then

$$
\begin{aligned}
\Delta &= \left|\frac{1}{n}\sum T'(\hat{F}_{(i)}(\hat{t}_i), X_i) - \frac{1}{n}\sum T'(F, X_i)\right| \\
&\leq \frac{1}{n}\sum\left|T'(\hat{F}_{(i)}(\hat{t}_i), X_i) - T'(F, X_i)\right| \\
&\leq \frac{1}{n}\sum A(X_i) \ .
\end{aligned}
$$

Since $[EA(X)]^2 \le EA^2(X) \le \epsilon$, it follows from Lemma 2 that $P[\Delta < 2\sqrt{\epsilon}] \to 1$ as $n \to \infty$. Hence, $\Delta \xrightarrow{P} 0$, so the two sums in its definition have the same limiting distribution. By (8), $ET'(F, X) = 0$, so each of these sums converges to zero in probability. It follows that

$$n(n-1)\left[\bar{T} - T(\hat{F})\right]^2 \xrightarrow{P} 0 \ ,$$

and the proof is complete.

We return now to the proof of part (iii) of Theorem 1. We show that under the assumptions of that theorem, the conditions of Theorem 3 are satisfied. It is clear from (7) that $T'(F, x)$ in this case is just the vector $(D_1, \ldots, D_M)$, so T is differentiable at F. For a given ϵ, let $A(x) = \sqrt{\epsilon}$ for all x. Then, since S is a finite set and the $\frac{\partial T}{\partial W_j}$ are assumed continuous, condition (b) is satisfied for all $\mathbf{w}$ in some convex open set U_ϵ containing $\mathbf{w}_0$. It is clear from Section 4 that conditions (a1) and (a2) are satisfied. We can therefore apply Theorem 3 to complete the proof of Theorem 1.

We now turn to the bias in $T(\hat{F})$. Our treatment of the bias will in many ways parallel our treatment of the variance. We say T is twice differentiable at G in $\mathcal{F}^+$ if it is differentiable there, and if there exists a function $T''(G, x, y)$ defined at all x, y in S such that for any admissible H (as described earlier),

$$\left.\frac{d^2 T(G + tH)}{dt^2}\right|_{t=0} = \int\int T''(G, x, y)\, dH(x) dH(y) \ .$$

We assume $T''(G, x, y) = T''(G, y, x)$. If we let $H = G$, we find, as in (8),

$$\int\int T''(G, x, y)\, dG(x) dG(y) = 0 \ . \tag{9}$$

To find the asymptotic bias, we assume that T is twice differentiable, in the sense defined above, at all G in some convex neighborhood of F in $\mathcal{F}^+$, such that $\hat{F}$ lies in the neighborhood with probability approaching one. We again parametrize the segment from F to $\hat{F}$ by $\hat{F}(t) = F + t(\hat{F} - F)$, $0 \le t \le 1$. Then if $\hat{F}$ lies in the neighborhood, we can expand in a Taylor series, as in Theorem 2:

$$\begin{aligned}
T(\hat{F}) - T(F) &= T[\hat{F}(1)] - T[\hat{F}(0)] \\
&= \left.\frac{dT}{dt}\right|_{t=0} + \frac{1}{2}\left.\frac{d^2 T}{dt^2}\right|_{t=\hat{t}} \\
&= \int T'(F, x)\, d(\hat{F} - F)(x) \\
&\qquad + \frac{1}{2}\int\int T''(\hat{F}(\hat{t}), x, y)\, d(\hat{F} - F)(x) d(\hat{F} - F)(y)
\end{aligned}$$

for some $0 \le \hat{t} \le 1$. For large n, $\hat{F}$ is near F, so we would expect $T''(\hat{F}(\hat{t}), x, y)$ to be near $T''(F, x, y)$ in some sense. We write

$$n\left\{T(\hat{F}) - T(F) - \int T'(F,x)\, d(\hat{F}-F)(x)\right\} \tag{10}$$

$$-\frac{n}{2}\int\int T''(F,x,y)\, d(\hat{F}-F)(x)\, d(\hat{F}-F)(y)$$

$$= \frac{n}{2}\int\int \left[T''(\hat{F}(\hat{t}),x,y) - T''(F,x,y)\right]\, d(\hat{F}-F)(x)\, d(\hat{F}-F)(y)\ .$$

We shall assume that the integral on the right side above converges to zero in probability. As before, this is the missing step which can often be verified for a particular example. If this assumption is true, then if either of the two expressions on the left side of (10) has a limiting distribution, then both have the same limiting distribution.

We consider the second of these expressions. If we write

$$n(\hat{F}-F) = \sum_{i=1}^{n} (\delta_{X_i} - F)\ ,$$

we have, using (9),

$$n^2 \int\int T''(F,x,y)\, d(\hat{F}-F)(x)\, d(\hat{F}-F)(y)$$

$$\begin{aligned}
&= \sum_i \sum_j \int\int T''(F,x,y)\, d\left(\delta_{X_i} - F\right)(x)\, d\left(\delta_{X_j} - F\right)(y) \\
&= \sum_i \sum_j \left\{T''(F,X_i,X_j) - \int T''(F,X_i,y)\, dF(y) - \int T''(F,x,X_j)\, dF(x)\right\} \\
&= \sum_i \sum_j h(X_i,X_j)\ ,
\end{aligned}$$

where h is defined by the expression in brackets above. If we let

$$U = \binom{n}{2}^{-1} \sum\sum_{i<j} h(X_i,X_j)\ ,$$

then U is a U-statistic of second order. We assume that $Eh^2(X,Y) < \infty$. We observe that by the definition of h, $E[h(X,Y)|Y=y] = 0$. It follows from the work of Hoeffding (1948)

that nU has a limiting distribution as $n \to \infty$. Since $Eh(X,Y) = 0$, the expected value of the limiting distribution of nU is zero. We must also consider terms of the form $h(X,X)$. Let

$$B = \frac{1}{2} \int T''(F,x,x)\, dF(x) \; .$$

We assume $\int |T''(F,x,x)|\, dF(x) < \infty$. Then $\frac{1}{2} Eh(X,X) = B$, and hence,

$$Z = \frac{1}{2n} \sum h(X_i, X_i) \xrightarrow{P} B \; .$$

The second expression on the left side of (10) now becomes

$$\frac{1}{2n} \sum_i \sum_j h(X_i, X_j) = Z + \frac{n-1}{2} U \; .$$

We conclude from the discussion above that this quantity has a limiting distribution whose expected value is B. By (8), we have

$$E\left\{ \int T'(F,x)\, d(\hat{F} - F)(x) \right\} = 0,$$

so, as we remarked in Section 4, we have by (10) the following approximation for large n:

$$ET(\hat{F}) \cong T(F) + \frac{B}{n} \; .$$

We call B the "asymptotic bias" of $T(\hat{F})$.

We would like to estimate B, so that we can remove the bias term of order n^{-1} from $T(\hat{F})$. If we regard $T''(\hat{F},x,x)$ as an estimate of $T''(F,x,x)$, a natural estimate of B is

$$\frac{1}{2} \int T''(\hat{F},x,x)\, d\hat{F}(x) = \frac{1}{2n} \sum_{i=1}^{n} T''(\hat{F}, X_i, X_i) \; .$$

But this is just n times the *IJK* bias estimate, because in the notation of this section,

$$\begin{aligned} \hat{D}_{ii} &= \left. \frac{d^2 T(\hat{F} + t\delta_{X_i})}{dt^2} \right|_{t=0} \\ &= \int\int T''(\hat{F},x,y)\, d\delta_{X_i}(x)\, d\delta_{X_i}(y) = T''(\hat{F}, X_i, X_i) \; . \end{aligned}$$

So the estimate above is $n\hat{B}(0)$.

We now give conditions, similar to those in Theorem 3, under which the *IJK* and *OJK* bias estimates converge to B.

Theorem 4. *Suppose T is twice differentiable at F, and $\int |T''(F,x,x)|\, dF(x) < \infty$. Let $X_1, \ldots, X_n$ be i.i.d. with distribution F. Suppose that for all $\epsilon > 0$ there is a convex set U in $\mathcal{F}^+$ and a function $A(x)$ such that T is twice differentiable at all G in U and*

(a1) $P[\hat{F} \in U] \to 1 \quad as \quad n \to \infty$,

(a2) $P[\hat{F}_{(i)} \in U,\, i = 1, \ldots, n] \to 1 \quad as \quad n \to \infty$,

(b) *For all $G \in U$ and $x \in S$:*

$$|T''(G,x,x) - T''(F,x,x)| \le A(x)\ ,$$

and (c)

$$\int A(x)\, dF(x) \le \epsilon\ .$$

Then, if $B = \frac{1}{2} \int T''(F,x,x)\, dF(x)$, as $n \to \infty$ we have for the IJK,

$$n\hat{B}(0) \xrightarrow{P} B$$

and for the OJK

$$n\hat{B} \xrightarrow{P} B\ .$$

Proof. For the *IJK* bias estimate, we have $n\hat{B}(0) = (2n)^{-1} \sum T''(\hat{F}, X_i, X_i)$. If we fix $\epsilon > 0$, we obtain U and $A(x)$ from the assumptions above. If we assume $\hat{F} \in U$, we have

$$\begin{aligned} \Delta &= \left| \frac{1}{2n} \sum T''(\hat{F}, X_i, X_i) - \frac{1}{2n} \sum T''(F, X_i, X_i) \right| \\ &\le \frac{1}{2n} \sum \left| T''(\hat{F}, X_i, X_i) - T''(F, X_i, X_i) \right| \\ &\le \frac{1}{2n} \sum A(X_i)\ . \end{aligned}$$

By (a1), (c) and Lemma 2, $P[\Delta < \epsilon] \to 1$ as $n \to \infty$. Therefore $\Delta \xrightarrow{P} 0$, so $n\hat{B}(0)$ and $(2n)^{-1} \sum T''(F, X_i, X_i)$ have the same limiting distribution. Since the latter expression is an average of i.i.d. random variables, it converges in probability to B. Therefore, $n\hat{B}(0) \xrightarrow{P} B$.

We now consider the *OJK* bias estimate. We write

$$\beta = \frac{n^2}{n-1}\, \hat{B} = n^2 \left[\bar{T} - T(\hat{F}) \right] = n \sum \left[T(\hat{F}_{(i)}) - T(\hat{F}) \right]$$

where $\bar{T} = n^{-1} \sum T(\hat{F}_{(i)})$. We fix $\epsilon > 0$ and obtain U and $A(x)$ from the assumptions above. For each i, we parametrize the segment from $\hat{F}$ to $\hat{F}_{(i)}$ as

$$\hat{F}_{(i)}(t) = \hat{F} + t(\hat{F}_{(i)} - \hat{F}) = \hat{F} - \frac{t}{n}\,\delta_{X_i} \ ,$$

where $0 \le t \le 1$. If we assume $\hat{F}$ and all $\hat{F}_{(i)}$, and hence all $\hat{F}_{(i)}(t)$, are in U, then for some $0 \le \hat{t}_i \le 1$,

$$\begin{aligned} T(\hat{F}_{(i)}) - T(\hat{F}) &= \left.\frac{dT[\hat{F}_{(i)}(t)]}{dt}\right|_{t=0} + \frac{1}{2}\left.\frac{d^2T[\hat{F}_{(i)}(t)]}{dt^2}\right|_{t=\hat{t}_i} \\ &= -\frac{1}{n}\int T'(\hat{F}, x)\, d\delta_{X_i}(x) \\ &\quad + \frac{1}{2n^2}\int\int T''(\hat{F}_{(i)}(\hat{t}_i), x, y)\, d\delta_{X_i}(x)\, d\delta_{X_i}(y) \\ &= -\frac{1}{n}T'(\hat{F}, X_i) + \frac{1}{2n^2}T''(\hat{F}_{(i)}(\hat{t}_i), X_i, X_i) \ . \end{aligned}$$

Since $\sum T'(\hat{F}, X_i) = 0$, we can write

$$\beta = \frac{1}{2n}\sum T''(\hat{F}_{(i)}(\hat{t}_i), X_i, X_i) \ .$$

If we let

$$\Delta = \left|\frac{1}{2n}\sum T''(\hat{F}_{(i)}(\hat{t}_i), X_i, X_i) - \frac{1}{2n}\sum T''(F, X_i, X_i)\right| \ ,$$

then by (a2) and the same argument used above for the *IJK*, we see that $\Delta \xrightarrow{P} 0$, so that the two sums in its definition have the same limiting distribution. Therefore, $\beta \xrightarrow{P} B$, from which it follows that $n\hat{B} \xrightarrow{P} B$ also, and the proof is complete.

We can apply Theorem 4 to complete the proof of Theorem 2 by an argument analogous to that given above for applying Theorem 3 to Theorem 1. We remark, as we did for the discrete case, that under the conditions of Theorems 3 and 4, if the original estimate has asymptotic variance V, then both forms of jackknifed estimate do also.

6 Some extensions and open questions.

A number of questions are raised by the jackknife procedures. For example: Which is better, the *IJK* or the *OJK*? Since for large n they are nearly the same, the choice of which to use would probably depend on which is easier to compute. Does the bias correction increase

or decrease the mean squared error of the estimate? Does the variance estimate (or bias estimate) have a bias in it? Could that bias be removed by jackknifing? Could we estimate the variance of the variance estimate this way? Such an estimate would be of interest because it could be converted into an approximate "degrees of freedom", with which we could studentize the original estimate or form a confidence interval. Some recent work has appeared on higher order bias reduction. See Gray, Watkins and Adams (1972). Analogous *IJK* procedures could presumably be defined by considering higher order derivatives. If $T(\hat{F})$ is asymptotically normal, the *IJK* or *OJK* variance estimate gives us a normal approximation to the distribution of $T(\hat{F})$ for sample size n. It may be possible to improve on this approximation by considering higher order terms in the Taylor expansions of Sections 4 and 5.

In Section 2 we defined the *IJK* by analogy with the *OJK*. Then, in Sections 4 and 5, when we derived the asymptotic variance and bias of $T(\hat{F})$, we saw that the *IJK* estimates of these quantities were in a sense their natural estimates. So if we want to apply the *IJK* method to other problems, it appears that what we should do is first derive the asymptotic variance (or bias, or other quantity) of the estimate by considering derivatives as in Sections 4 and 5, and then try to find an estimate of the quantity thus derived. For example, consider the following situation:

Suppose we estimate more than one parameter from the sample, and we want to estimate their joint moments. For example, suppose we have two statistics T and U, which we write in the notation of Section 4 as

$$T(\hat{F}) = T(F) + \sum_i \left(\hat{W}_i - \frac{1}{M}\right) D_i^T + \dots$$

and

$$U(\hat{F}) = U(F) + \sum_j \left(\hat{W}_j - \frac{1}{M}\right) D_j^U + \dots ,$$

and suppose we want to estimate their covariance. We write

$$\begin{aligned} & E\left\{[T(\hat{F}) - T(F)][U(\hat{F}) - U(F)]\right\} \\ &\cong\ E\left\{\sum_i \left(\hat{W}_i - \frac{1}{M}\right) D_i^T \cdot \sum_j \left(\hat{W}_j - \frac{1}{M}\right) D_j^U\right\} \\ &=\ E\left\{\sum_j \left(\hat{W}_j - \frac{1}{M}\right)^2 D_j^T D_j^U + \sum\sum_{i\neq j} \left(\hat{W}_i - \frac{1}{M}\right)\left(\hat{W}_j - \frac{1}{M}\right) D_i^T D_j^U\right\} \\ &=\ \frac{1}{nM} \sum D_j^T D_j^U \ . \end{aligned}$$

We can estimate this quantity by

$$\frac{1}{n^2}\sum_{i=1}^{n}\hat{D}_i^T\hat{D}_i^U\ ,$$

if T and U are well behaved. So the *IJK* gives us an estimate of the covariance of the two statistics.

Miller (1964) showed that certain functions of the sample mean can be jackknifed. This result was extended to functions of U-statistics by Arvesen (1969). We can ask the following more general question: Given a statistic T which can be jackknifed, and a sufficiently smooth function h, can we jackknife $U = h(T)$? We shall consider this question for the *IJK* variance estimate. We shall see that the result of jackknifing U is the same as for an alternative method, which we now describe. If we expand $h[T(\hat{F})]$ in a Taylor series

$$U(\hat{F}) = h[T(\hat{F})] \cong h[T(F)] + [T(\hat{F}) - T(F)]\cdot h'[T(F)]\,,$$

we see that

$$\operatorname{Var} U(\hat{F}) \cong \operatorname{Var} T(\hat{F})\cdot\{h'[T(F)]\}^2\ .$$

So if $\hat{V}^T(0)$ is the *IJK* estimate of $\operatorname{Var} T(\hat{F})$, an estimate of $\operatorname{Var} U(\hat{F})$ would be

$$\hat{V}^T(0)\cdot\left\{h'[T(\hat{F})]\right\}^2\ .$$

If we apply the *IJK* directly to $U(\hat{F})$, we find, using the notation of Section 4,

$$\frac{\partial U}{\partial w_i} = \frac{dh(T)}{dT}\cdot\frac{\partial T}{\partial w_i}$$

and

$$\hat{D}_i^U = \left.\frac{\partial U}{\partial w_i}\right|_{w_j=\frac{1}{n},\,j=1,\ldots,n} = h'[T(\hat{F})]\cdot\hat{D}_i^T\ ,$$

so that

$$\begin{aligned}\hat{V}^U(0) = \frac{1}{n^2}\sum\left(\hat{D}_i^U\right)^2 &= \frac{1}{n^2}\sum\left(\hat{D}_i^T\right)^2\cdot\{h'[T(\hat{F})]\}^2\\ &= \hat{V}^T(0)\cdot\{h'[T(\hat{F})]\}^2\ .\end{aligned}$$

This is the same as the estimate obtained by the Taylor series method above.

A similar computation can be done for the estimated bias in $h[T(\hat{F})]$.

In this paper we have restricted ourselves to models with independent observations. But if we look back at the Taylor expansion (5), we see that all we really need to know are the expectations, variances, and covariances of the $\hat{W}_j$. So we may be able to apply the *IJK* in some models where we do not have independence, so long as we have sufficient information about the $\hat{W}_j$.

7 A more general model.

We now consider a more general situation, in which the X_i are not identically distributed and T is not symmetric in its arguments. We follow the procedure mentioned in Section 6; that is, we approximate the variance using derivatives, and then we find an estimate for that variance expression. A particular example, which we shall consider in detail, is the problem of estimating the variance of the estimated slope of a regression line.

Suppose the $X_i, i = 1, \ldots, n$ are independent, and the distribution of X_i is F_i. We assume that each F_i is discrete, and that for each i,

$$P(X_i = z_{ij}) = \frac{1}{M}, \qquad j = 1, \ldots, M \ .$$

We define the random variable $\hat{W}_{ij}$ by

$$\hat{W}_{ij} = \begin{cases} 1 & \text{if } X_i = z_{ij} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E\hat{W}_{ij} = \frac{1}{M}, \quad \operatorname{Var} \hat{W}_{ij} = \frac{1}{M}\left(1 - \frac{1}{M}\right) \ ,$$

$$\operatorname{Cov}(\hat{W}_{ij}, \hat{W}_{i\ell}) = -\frac{1}{M^2} \quad \text{for} \quad j \neq \ell \, ,$$

and

$$\operatorname{Cov}(\hat{W}_{ij}, \hat{W}_{k\ell}) = 0 \quad \text{for} \quad i \neq k \ .$$

To make the notation here more like that of Section 4, we write

$$F = (F_1, \ldots, F_n),$$

and

$$\hat{F} = (\hat{F}_1, \ldots, \hat{F}_n),$$

where $\hat{F}_i$ is simply the distribution which assigns probability one to X_i. We can then think of the estimate $T(\hat{F})$ as a function of the Mn random variables $\{\hat{W}_{ij}\}$. The quantity to be estimated is then $T(F)$. We assume $T(cG) = T(G)$.

We now write down the Taylor expansion, through first derivatives only. We have

$$T(\hat{F}) = T(F) + \sum_{i=1}^{n} \sum_{j=1}^{M} \left(\hat{W}_{ij} - \frac{1}{M}\right) D_{ij} + \ldots,$$

where

$$D_{ij} = \left.\frac{\partial T}{\partial W_{ij}}\right|_{W_{k\ell}=\frac{1}{M},\, k=1,\ldots,n,\, \ell=1,\ldots,M} .$$

We assume that the model is well behaved, in the sense that for each n we have a model defined for that n, such that as $n \to \infty$ this sequence of models converges to some "asymptotic model". Then if T is well behaved, the asymptotic variance of T may be found from the expansion above. We have

$$\begin{aligned}
\operatorname{Var} T(\hat{F}) &\cong \sum_i \sum_j \sum_k \sum_\ell \operatorname{Cov}(\hat{W}_{ij}, \hat{W}_{k\ell}) D_{ij} D_{k\ell} \qquad (11)\\
&= \sum_i \left\{ \sum_j \operatorname{Var} \hat{W}_{ij} \cdot D_{ij}^2 + \sum_{j \neq \ell} \sum \operatorname{Cov}(\hat{W}_{ij}, \hat{W}_{i\ell}) D_{ij} D_{i\ell} \right\}\\
&= \sum_i \left\{ \frac{1}{M} \sum_j D_{ij}^2 - \frac{1}{M^2} \sum_j \sum_\ell D_{ij} D_{i\ell} \right\}\\
&= \sum_i \left\{ \frac{1}{M} \sum_j D_{ij}^2 - \left(\frac{1}{M} \sum_j D_{ij} \right)^2 \right\} .
\end{aligned}$$

Note that $\sum_j D_{ij}$ is not necessarily zero.

To estimate the variance, we must somehow estimate the above expression from the observations, as we did in the i.i.d. case. A little thought shows that this is not as simple as before. In the i.i.d. case, $\hat{F}$ was an estimate of F. But here we have a different F_i for each observation, and the best estimate we have of it is $\hat{F}_i$, which we defined above to have all of its weight at X_i. It seems that we can surmount this problem only if there are some additional restrictions imposed on the model. To illustrate this difficulty, we consider estimating the slope of a regression line.

Let t_i, $i = 1, \ldots, n$ be a set of n distinct numbers such that

$$\sum t_i = 0 \qquad \text{and} \qquad \sum t_i^2 = 1 . \qquad (12)$$

Let X_i, $i = 1, \ldots, n$ be independent, with discrete distribution F_i as described above, and suppose that

$$EX_i = \alpha + \beta t_i, \quad i = 1, \ldots, n ,$$

for some unknown α and β. We estimate β by

$$\hat{\beta} = T(\hat{F}) = \frac{\sum (t_i - \bar{t})(X_i - \bar{X})}{\sum (t_i - \bar{t})^2} .$$

We state the definition in this general form because when we vary the weights, as we do below when we differentiate, then (12) will not hold in general. Using our notation, we shall see that we can write

$$\begin{aligned} T(F) &= \sum_{i=1}^{n} t_i \left\{ \frac{1}{M} \sum_{j=1}^{M} z_{ij} \right\} \\ &= \sum t_i(\alpha + \beta t_i) = \beta \ . \end{aligned}$$

So we are estimating $T(F)$ by $T(\hat{F})$, as in the i.i.d. case.

We now derive $\operatorname{Var} \hat{\beta}$, which is the quantity we wish to estimate. Let $\sigma_i^2 = \operatorname{Var} X_i$. Then

$$\operatorname{Var} \hat{\beta} = \sum_i t_i^2 \sigma_i^2 \ .$$

If we find the variance using (11), we should obtain approximately the same result. We begin by writing T as a function of the Mn weights, W_{ij}. We write

$$\bar{t} = \frac{\sum_i t_i \sum_j W_{ij}}{\sum_i \sum_j W_{ij}} \qquad \text{and} \qquad \bar{X} = \frac{\sum_i \sum_j z_{ij} W_{ij}}{\sum_i \sum_j W_{ij}} \ .$$

We then have

$$T(\{z_{ij}\}, \{W_{ij}\}) = \frac{\sum_i (t_i - \bar{t}) \sum_j (z_{ij} - \bar{X}) W_{ij}}{\sum_i (t_i - \bar{t})^2 \sum_j W_{ij}} \tag{13}$$

$$= \frac{(\sum \sum W_{ij}) (\sum t_i \sum z_{ij} W_{ij}) - (\sum t_i \sum W_{ij}) (\sum \sum z_{ij} W_{ij})}{(\sum \sum W_{ij}) (\sum t_i^2 \sum W_{ij}) - (\sum t_i \sum W_{ij})^2} = \frac{N}{D} \ .$$

If all of the W_{ij} are $1/M$, we have $T = \beta$, the true value of the parameter. We now differentiate.

$$\begin{aligned} \frac{\partial T}{\partial W_{k\ell}} = \frac{1}{D^2} \Bigg\{ D \Bigg[& \sum t_i \sum z_{ij} W_{ij} + \Big(\sum \sum W_{ij} \Big) t_k z_{k\ell} \\ & - t_k \sum \sum z_{ij} W_{ij} - \Big(\sum t_i \sum W_{ij} \Big) z_{k\ell} \Bigg] \\ & - N \Big[\sum t_i^2 \sum W_{ij} + \Big(\sum \sum W_{ij} \Big) t_k^2 - 2 \Big(\sum t_i \sum W_{ij} \Big) t_k \Big] \Bigg\} \end{aligned}$$

Letting $W_{ij} = \frac{1}{M}$ for all i and j, we find $N = n\beta$ and $D = n$, so

$$\begin{aligned} D_{k\ell} &= \frac{1}{n^2} \left\{ n \left[\beta + n t_k z_{k\ell} - n\alpha t_k \right] - n\beta \left[1 + n t_k^2 \right] \right\} \\ &= t_k \left(z_{k\ell} - \alpha - \beta t_k \right) \ . \end{aligned}$$

We can now find the variance given by (11). We have

$$\begin{aligned}\operatorname{Var}\hat{\beta} &\cong \sum_i \left\{ \frac{t_i^2}{M} \sum_j (z_{ij} - \alpha - \beta t_i)^2 - \left[\frac{t_i}{M} \sum_j (z_{ij} - \alpha - \beta t_i) \right]^2 \right\} \\ &= \sum_i t_i^2 \sigma_i^2 \,,\end{aligned}$$

exactly the same quantity that we found directly above.

We would like to estimate this quantity from the observations. So we think of attaching a weight v_i to each observation and taking derivatives with respect to them, as in the i.i.d. case. If we write $\hat{\beta}$ using these n weights, we have

$$\begin{aligned}\hat{\beta}(\{X_i\},\{v_i\}) &= \frac{\sum (t_i - \bar{t})(X_i - \bar{X}) v_i}{\sum (t_i - \bar{t})^2 v_i} \\ &= \frac{\left(\sum v_i\right)\left(\sum t_i X_i v_i\right) - \left(\sum t_i v_i\right)\left(\sum X_i v_i\right)}{\left(\sum v_i\right)\left(\sum t_i^2 v_i\right) - \left(\sum t_i v_i\right)^2} = \frac{N}{D} \,,\end{aligned} \tag{14}$$

where

$$\bar{t} = \frac{\sum t_i v_i}{\sum v_i} \qquad \text{and} \qquad \bar{X} = \frac{\sum X_i v_i}{\sum v_i} \,.$$

Note the difference between this expression and (13), where we had Mn weights. (Actually, (14) is a special case of (13); if we define the W_{ij} in (13) to be v_i if $X_i = z_{ij}$ and 0 otherwise, we have (14).) We can now differentiate (14) as we did for (13). We have

$$\begin{aligned}\frac{\partial \hat{\beta}}{\partial v_k} = \frac{1}{D^2} \Big\{ D \Big[\sum t_i X_i v_i + \Big(\sum v_i\Big) t_k X_k - t_k \sum X_i v_i - \Big(\sum t_i v_i\Big) X_k \Big] \\ - N \Big[\sum t_i^2 v_i + \Big(\sum v_i\Big) t_k^2 - 2 \Big(\sum t_i v_i\Big) t_k \Big] \Big\} \,.\end{aligned}$$

Letting $v_i = 1$ for all i, and letting $\hat{\alpha} = \frac{1}{n} \sum X_i$, we find $N = n\hat{\beta}$ and $D = n$, so

$$\begin{aligned}\hat{D}_k &= \frac{1}{n^2} \left\{ n \left[\hat{\beta} + n t_k X_k - n t_k \hat{\alpha} \right] - n\hat{\beta} \left[1 + n t_k^2 \right] \right\} \\ &= t_k \left(X_k - \hat{\alpha} - \hat{\beta} t_k \right) \,.\end{aligned}$$

Since we assumed that $EX_i = \alpha + \beta t_i$, we have $\sum_j D_{ij} = 0$ in (11), so the quantity we must estimate is

$$V = \sum_i \left\{ \frac{1}{M} \sum_j D_{ij}^2 \right\}$$

$$= \sum_i t_i^2 \cdot \frac{1}{M} \sum_j (z_{ij} - \alpha - \beta t_i)^2 \ .$$

Our estimate of the variance is

$$\hat{V} = \sum_i \hat{D}_i^2 = \sum_i t_i^2 (X_i - \hat{\alpha} - \hat{\beta} t_i)^2 \ .$$

This is the *IJK* estimate of the variance of $\hat{\beta}$. If we had not assumed $\sum_j D_{ij} = 0$, we could not have estimated (11), unless there were more than one observation for each t_i so we could form a "within groups" measure of variance. However, if the assumed model did not hold, we could presumably use $\hat{V}$ as an estimate of the mean square error of $\hat{\beta}$, as in classical regression analysis. We see that $\hat{V}$ is not the same as the classical estimate

$$S^2 = \frac{1}{n-2} \sum (X_i - \hat{\alpha} - \hat{\beta} t_i)^2 \ ,$$

in deriving which all the σ_i^2 are assumed equal. $\hat{V}$ is more general, however, since it applies for arbitrary σ_i^2. If all of the σ_i^2 are equal, we see that for large n and well behaved t_i, t_i^2 and $(X_i - \hat{\alpha} - \hat{\beta} t_i)^2$ are nearly uncorrelated, so we have

$$\frac{1}{n} \sum t_i^2 (X_i - \hat{\alpha} - \hat{\beta} t_i)^2 \cong \left(\frac{1}{n} \sum t_i^2 \right) \cdot \frac{1}{n} \sum (X_i - \hat{\alpha} - \hat{\beta} t_i)^2 \ ,$$

so that $\hat{V} \cong S^2$.

We can now see the problem referred to earlier. We are estimating V, a sum containing Mn terms, by $\hat{V}$, a sum of n terms. In the i.i.d. case, each $\hat{D}_i^2$ could be regarded as an estimate of the corresponding D_i^2, whereas here each individual $\hat{D}_i^2$ may be nowhere near $\frac{1}{M} \sum_j D_{ij}^2$. Thus we must impose conditions on the model to insure that in the aggregate the $\hat{D}_i^2$ may serve as estimates of the $\frac{1}{M} \sum_j D_{ij}^2$, so that $\hat{V}$ will be a good estimate of V. For example, if F_i changes in some gradual, regular way as i varies, then we may have a kind of redundancy in the model which would allow us to estimate V by $\hat{V}$. In the regression example above, if we assume that the t_i are well behaved in the sense that no small subset of the t_i contribute an unduly large amount to $\sum t_i^2$, and if we make a similar assumption about the σ_i^2, then $\hat{V}$ should be a reasonably good estimate of V.

If we want to derive asymptotic results, such as that $\hat{V}$ converges to the same limit that V converges to as $n \to \infty$, then we not only need conditions similar to those described in Sections 4 and 5, but we also need to impose on the model conditions of the kind just discussed.

We could also define an *OJK* variance estimate analogous to the *IJK* estimate above. As we remarked in Section 5, the differences formed by recomputing the estimate with an

observation omitted may be thought of as approximations to the $\hat{D}_i$, and thus may be used to form an *OJK* variance estimate.

Acknowledgment. The author is indebted to Dr. Colin L. Mallows for reading earlier versions of this paper and making many valuable suggestions.

MH-1215-LAJ-rb LOUIS A. JAECKEL

References

Arvesen, J. (1969). "Jackknifing U-statistics." *Ann. Math. Statist.* **40**, 2076-2100.

Brillinger, D.R. (1964). "The asymptotic behavior of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates." *Rev. Int. Statist. Inst.* **32**, 202-206.

Gray, H.L., Watkins, T.A. and Adams, J.E. (1972). "On the jackknife statistic, its extensions, and its relation to e_n-transformations." *Ann. Math. Statist.* **43**, 1-30.

Hampel, F.R. (1968). "Contributions to the theory of robust estimation." Ph.D. dissertation, Univ. of California, Berkeley.

Hoeffding, W. (1948). "A class of statistics with asymptotically normal distribution." *Ann. Math. Statist.* **19**, 293-325.

Huber, P.J. (1964). "Robust estimation of a location parameter." *Ann. Math. Statist.* **35**, 73-101.

Miller, R. (1964). "A trustworthy jackknife." *Ann. Math. Statist.* **35**, 1594-1605.

von Mises, R. (1947). "On the asymptotic distribution of differentiable statistical functions." *Ann. Math. Statist.* **18**, 309-348.