Resampling Data:
Using a Statistical Jackknife

S. Sawyer — Washington University — March 11, 2005

1. Why Resample? Suppose that we want to estimate a parameter θ that depends on a random quantity sample $X = (X_1, X_2, \ldots, X_n)$ in a complicated way. For example, θ might be the sample variance of X or the log sample variance. If the X_i are vector valued, θ could be the Pearson correlation coefficient.

Assume that we have an estimator $\phi_n(X_1, X_2, \ldots, X_n)$ of θ but do not know the probability distribution of $\phi_n(X)$ given θ. This means that we cannot estimate the error involved in estimating θ by $\phi_n(X_1, \ldots, X_n)$, and that we cannot tell if we can conclude $\theta \neq 0$ from an observed $\phi_n(X) \neq 0$, no matter how large.

More generally, can we get a confidence interval for θ depending only on the observed $X_1, X_2, \ldots, X_n$, or test $H_0 : \theta = \theta_0$ just using the data $X_1, X_2, \ldots, X_n$?

Methods that try to estimate the bias and variability of an estimator $\phi_n(X_1, X_2, \ldots, X_n)$ by using the values of $\phi_n(X)$ on subsamples from $X_1, X_2, \ldots, X_n$ are called *resampling* methods. Two common resampling methods are the *jackknife*, which is discussed below, and the *bootstrap*.

The jackknife was invented by Quenouille in 1949 for the more limited purpose of correcting possible bias in $\phi_n(X_1, X_2, \ldots, X_n)$ for small n. Tukey in 1958 noticed that the procedure could be used to construct reasonably reliable confidence intervals for a wide variety of estimators $\phi_n(X)$, and so might be viewed as being as useful to a statistician as a regular jackknife would be to an outdoorsman. Bootstrap methods were invented by Bradley Efron around 1979. These are computationally more intensive (although easier to program) and give more accurate results in some cases.

2. The Jackknife Recipe. Let $\phi_n(X) = \phi_n(X_1, \ldots, X_n)$ be an estimator defined for samples $X = (X_1, X_2, \ldots, X_n)$. The i^{th} *pseudovalue* of $\phi_n(X)$ is

$$ps_i(X) = n\phi_n(X_1, X_2, \ldots, X_n) - (n-1)\phi_{n-1}(\,(X_1, \ldots, \ldots, X_n)_{[i]}) \qquad (1)$$

In (1), $X_{[i]}$ means the sample $X = (X_1, X_2, \ldots, X_n)$ with the i^{th} value X_i *deleted* from the sample, so that $X_{[i]}$ is a sample of size $n-1$. Note

$$ps_i(X) = \phi_n(X) + (n-1)\big(\phi_n(X) - \phi_{n-1}(X_{[i]})\big)$$

so that $ps_i(X)$ can be viewed as a bias-corrected version of $\phi_n(X)$ determined by the trend in the estimators $\phi_n(X)$ from $\phi_{n-1}(X_{[i]})$ to $\phi_n(X)$.

The basic jackknife recipe is to treat the pseudovalues $ps_i(X)$ as if they were independent random variables with mean θ. One can then obtain confidence intervals and carry out statistical tests using the Central Limit Theorem. Specifically, let

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) \quad \text{and} \quad V_{ps}(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(ps_i(X) - ps(X)\big)^2 \tag{2}$$

be the mean and sample variance of the pseudovalues. The sample mean $ps(X)$ was Quenouille's (1949) bias-corrected version of $\phi_n(X)$. The jackknife 95% confidence interval for θ is

$$\left(ps(X) - 1.960\sqrt{\frac{1}{n}V_{ps}(X)}\,,\quad ps(X) + 1.960\sqrt{\frac{1}{n}V_{ps}(X)}\right) \tag{3}$$

Similarly, one can define a jackknife P-value for the hypothesis $H_0 : \theta = \theta_0$ by comparing

$$Z \;=\; \frac{\sqrt{n}\,\big(ps(X) - \theta_0\big)}{\sqrt{V_{ps}(X)}} \;=\; \frac{ps(X) - \theta_0}{\sqrt{(1/n)V_{ps}(X)}} \tag{4}$$

with a standard normal variable.

Remark: Technically speaking, the pseudovalues in (1) are for what is called the *delete-one* jackknife. There is also a more general *delete-k* or *block* jackknife that we discuss below.

3. Examples (1) If $\phi_n(X) = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}$ is the sample mean for $\theta = E(X_i)$, then the pseudovalues

$$ps_i(X) \;=\; n\overline{X} \;-\; (n-1)\overline{X_{[i]}} \;=\; X_i$$

are the same as the original values. Thus

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) = \overline{X} \qquad \text{and} \qquad V_{ps}(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2 \tag{5}$$

are the usual sample mean and variance.

(2) If $\phi_n(X) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$ is the sample variance, then, after some algebra, the pseudovalues of $\phi_n(X)$ are

$$ps_i(X) \;=\; \frac{n}{n-2}(X_i - \overline{X})^2 \;-\; \frac{1}{(n-1)(n-2)}\sum_{j=1}^{n}(X_j - \overline{X})^2 \tag{6}$$

The mean of the pseudovalues

$$ps(X) \;=\; \frac{1}{n}\sum_{i=1}^{n} ps_i(X) \;=\; \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

is the same as $\phi_n(X)$ in this case also.

(3) If $\phi_n(X) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$ is the sample variance with $1/(n-1)$ replaced by $1/n$, then the pseudovalues of $\phi_n(X)$ are

$$ps_i(X) \;=\; \frac{n}{n-1}(X_i - \overline{X})^2 \tag{7}$$

This implies that

$$ps(X) = \frac{1}{n}\sum_{i=1}^{n} ps_i(X) = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \overline{X}\big)^2$$

is the usual sample variance. Note that $E\big(\phi_n(X)\big) = \frac{n-1}{n}\sigma^2$ for $\sigma^2 = \mathrm{Var}(X)$ while $E\big(ps(X)\big) = \sigma^2$, so that $ps(X)$ is a bias-corrected version of $\phi_n(X)$.

4. A Simple Example. Suppose that we have four observations $\{1, 2, 3, 4\}$ with $\phi_4(X) = \overline{X}$. Thus $\phi_4(X) = (1/4)\sum_{i=1}^{4} X_i = (1/4)(1+2+3+4) = 2.5$.

The four delete-one values are $\phi_3(X_{[1]}) = (1/3)(2+3+4) = 3.0$, $\phi_3(X_{[2]}) = (1/3)(1+3+4) = 2.67$, $\phi_3(X_{[3]}) = (1/3)(1+2+4) = 2.33$, and $\phi_3(X_{[4]}) = (1/3)(1+2+3) = 2.00$.

The four pseudovalues are $ps_1(X) = 4\phi_4(X) - 3\phi_3(X_{[1]}) = 4(2.50) - 3(3.0) = 10 - 9 = 1.0$, $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[2]}) = 4(2.50) - 3(2.67) = 10 - 8 = 2.0$, $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[3]}) = 4(2.50) - 3(2.33) = 10 - 7 = 3.0$, and $ps_2(X) = 4\phi_4(X) - 3\phi_3(X_{[4]}) = 4(2.50) - 3(2.00) = 10 - 6 = 4.0$. Thus the four pseudovalues are the same as the original observations, as they should be for $\phi_n(X) = \overline{X}$.

5. Another Example. Suppose that we have 16 observations

17.23	13.93	15.78	14.91	18.21	14.28	18.83	13.45
18.71	18.81	11.29	13.39	11.57	10.94	15.52	15.25

and that we are interested in estimating the variance σ^2 of the data and in finding a 95% confidence interval for σ^2. In order to minimize any possible effect of outliers, we apply the jackknife to the log sample variance

$$\phi_n(X_1, \ldots, X_n) \;=\; \log(s^2) \;=\; \log\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

instead of to s^2 directly. For these 16 observations, $s^2 = 7.171$ and $\phi_n(X) = \log(s^2) = 1.9701$.

The delete-one values $\phi_{n-1}(X_{[i]})$ on the 16 subsamples with $n-1 = 15$ are

1.994	2.025	2.035	2.039	1.940	2.032	1.893	2.011
1.903	1.895	1.881	2.009	1.905	1.848	2.038	2.039

The corresponding pseudovalues $ps_i(X) = n\phi_n(X) - (n-1)\phi_{n-1}(X_{[i]})$ are

1.605	1.151	0.998	0.942	2.416	1.043	3.122	1.362
2.972	3.097	3.308	1.393	2.951	3.806	0.958	0.937

The mean of the pseudovalues is 2.00389, which is a little larger than the initial estimate $\phi_n(X) = 1.9701$. The sample variance of the 16 pseudovalues is 1.091. The jackknife 95% confidence interval for the log variance $\log(\sigma^2)$ is $(1.492, 2.516)$.

The 16 values in this case were drawn from a probability distribution whose true variance is 5.0, with $\log(5.0) = 1.609$, which is well within the 95% jackknife confidence interval.

6. The Delete-*k* or Block Jacknife. If n is large, the pseudovalues $ps_i(X)$ in (1) may be too close together, and the variance $V_{ps}(X)$ may be mostly sampling error. In that case, we can define a *block jackknife* instead of the *delete-one* jackknife defined above by proceeding as follow. Assume $n = n_b k$, where k will be the block size and n_b is the number of blocks. Define

$$ps_i(X) = n_b\phi_n(X_1, X_2, \ldots, X_n) - (n_b - 1)\phi_{n-k}(\,(X_1, \ldots, \ldots, X_n)_{[i]}\,) \quad (8)$$

instead of (1), where now $1 \le i \le n_b$ and $X_{[i]}$ means the sample $X = (X_1, X_2, \ldots, X_n)$ with the i^{th} *block* of k values — that is, with indices j in the range $ik + 1 \le j \le ik + k$ — removed.

For example, if $n = 200$, we might set $k = 20$ and $n_b = 10$. Each of the $n_b = 10$ pseudovalues (8) would be defined in terms of ϕ_{200} on the full sample and ϕ_{180} on a subsample of size 180. We then treat the 10 pseudovalues (8) as a sample of 10 independent values with mean θ and proceed as before.

7. A Warning and Another Example: The jackknife should NOT be applied if the estimator $\phi_n(X)$ is too discontinuous as a function of the X_i, or if $\phi_n(X)$ depends on one or a few values in X.

For example, suppose that $\phi_n(X)$ is the *sample median* of $X = (X_1, X_2, \ldots, X_n)$ where $X_1, \ldots, X_n$ are distinct. Then

Exercise: Suppose that $n = 2m$ is even. Prove that

(a) There exists two numbers a, b depending on $X_1, \dots, X_n$ such that each value $\phi_{n-1}(X_{[i]})$ is either a or b but

(b) the bias-corrected mean $ps(X) = \phi_n(X)$.

How do these results change if $n = 2m + 1$ is odd?

8. Coverage Frequencies for Jackknife Confidence Intervals. As a test of the jackknife confidence interval (3), we generate 10,000 samples of size $n = 20$ from the probability distribution $12x(1-x)^2$ on the unit interval $(0, 1)$. (This is a beta density with parameters $\alpha = 2$ and $\beta = 3$.)

For each sample of size 20, we compute the jackknife 95% confidence (3) for the variance, and count the number of samples out of 10,000 for which the jackknife 95% confidence interval contains the true variance, which is 0.048 in this case.

The *coverage probability* of a (putative) confidence interval is the probability that it actually contains the true value. If it is supposed to be a 95% confidence interval, then the coverage probability should be as close to 0.95 as possible. If the coverage probability is higher, then the confidence intervals are too conservative. If the coverage probability is lower, then they are too small and we may be misled.

We do the same calculation for $\phi_n(X)$ replaced by the logarithm of the variance, and also for the jackknife 99% confidence interval (3) with 1.96 replaced by 2.576.

As a third test, we generate 10,000 samples of $n = 20$ pairs of standard normal variables (X_i, Y_i) with a theoretical Pearson correlation coefficient of $\rho = 0.50$, and apply the same procedures for the sample (Pearson) correlation coefficient.

Some typical jackknife confidence intervals in these cases are

	Beta: variance	Beta: log variance	Normal: ρ
95%:	(0.0221, 0.0655)	(-4.2420, -2.9155)	(0.1976, 0.8387)
99%:	(0.0153, 0.0723)	(-4.4505, -2.7071)	(0.0968, 0.9395)

The estimated coverage probabilities were

95%:	0.9025	0.9390	0.9017
99%:	0.9503	0.9779	0.9556

Thus the confidence intervals tend to be a bit small, but are approximately correct. Psuedovalues for the logarithm of variance are better behaved than pseudovalues for the variance itself, presumably because the effect of large sample variances is smaller.

We might expect the results to be worse if the sample were smaller or if the distribution was more heavy tailed, for example having an exponential instead of a beta distribution, but this remains to be checked.