In Neil Salkind (Ed.), *Encyclopedia of Research Design.*
Thousand Oaks, CA: Sage. 2010

Jackknife

Hervé Abdi · Lynne J. Williams

1 Introduction

The jackknife or "leave one out" procedure is a cross-validation technique first developed by Quenouille to estimate the bias of an estimator. John Tukey then expanded the use of the jackknife to include variance estimation and tailored the name of jackknife because like a jackknife—a pocket knife akin to a Swiss army knife and typically used by boy scouts—this technique can be used as a "quick and dirty" replacement tool for a lot of more sophisticated and specific tools. Curiously, despite its remarkable influence on the statistical community, the seminal work of Tukey is available only from an abstract (which does not even mention the name of jackknife) and from an almost impossible to find unpublished note (although some of this note found its way into Tukey's complete work).

The jackknife estimation of a parameter is an iterative process. First the parameter is estimated from the whole sample. Then each element is, in turn, dropped from the sample and the parameter of interest is estimated from this smaller sample. This estimation is called a *partial estimate* (or also a *jackknife replication*). A *pseudo-value* is then computed as the difference between the whole sample es-

Hervé Abdi
The University of Texas at Dallas

Lynne J. Williams
The University of Toronto Scarborough

Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

timate and the partial estimate. These pseudo-values reduce the (linear) bias of the partial estimate (because the bias is eliminated by the subtraction between the two estimates). The pseudo-values are then used in lieu of the original values to estimate the parameter of interest and their standard deviation is used to estimate the parameter standard error which can then be used for null hypothesis testing and for computing confidence intervals. The jackknife is strongly related to the bootstrap (*i.e.,* the jackknife is often a linear approximation of the bootstrap) which is currently the main technique for computational estimation of population parameters.

As a potential source of confusion, a somewhat different (but related) method, also called jackknife is used to evaluate the quality of the prediction of computational models built to *predict* the value of dependent variable(s) from a set of independent variable(s). Such models can originate, for example, from neural networks, machine learning, genetic algorithms, statistical learning models, or any other multivariate analysis technique. These models typically use a very large number of parameters (frequently *more* parameters than observations) and are therefore highly prone to over-fitting (*i.e.,* to be able to perfectly predict the data within the sample because of the large number of parameters, but being poorly able to predict *new* observations). In general, these models are too complex to be analyzed by current analytical techniques and therefore the effect of over-fitting is difficult to evaluate directly. The jackknife can be used to estimate the actual predictive power of such models by predicting the dependent variable values of each observation as if this observation were a new observation. In order to do so, the predicted value(s) of each observation is (are) obtained from the model built on the sample of observations minus the observation to be predicted. The jackknife, in this context, is a procedure which is used to obtain an unbiased prediction (*i.e.,* a random effect) and to minimize the risk of over-fitting.

2 Definitions and Notations

The goal of the jackknife is to estimate a parameter of a population of interest from a random sample of data from this population. The parameter is denoted θ, its estimate from a sample is denoted T, and its jackknife estimate is denoted T^*. The sample of N observations (which can be univariate or multivariate) is a set denoted $\{X_1, \ldots, X_n, \ldots, X_N\}$. The sample estimate of the parameter is a function of the observations in the sample. Formally:

$$T = f\left(X_1, \ldots, X_n, \ldots, X_N\right) \ . \tag{1}$$

An estimation of the population parameter obtained *without* the nth observation, is called the n-th *partial prediction*, and is denoted T_{-n}. Formally:

$$T_{-n} = f\left(X_1, \ldots, X_{n-1}, X_{n+1} \ldots, X_N\right) \ . \tag{2}$$

A *pseudo-value* estimation of the nth observation is denoted T_n^*, it is computed as the difference between the parameter estimation obtained from the whole sample and the parameter estimation obtained without the nth observation. Formally:

$$T_n^* = NT - (N-1)T_{-n} \ . \tag{3}$$

The jackknife estimate of θ, denoted T^*, is obtained as the mean of the pseudo-values. Formally:

$$T^* = \overline{T}_\bullet^* = \frac{1}{N}\sum_n^N T_n^* \ , \tag{4}$$

where $\overline{T}_\bullet^*$ is the mean of the pseudo-values. The variance of the pseudo-values is denoted $\widehat{\sigma}^2_{T_n^*}$ and is obtained with the usual formula:

$$\widehat{\sigma}^2_{T_n^*} = \frac{\sum\left(T_n^* - \overline{T}_\bullet^*\right)^2}{N-1} \ . \tag{5}$$

Tukey conjectured that the T_n^*s could be considered as independent random variables. Therefore the standard error of the parameter estimates, denoted $\widehat{\sigma}^2_{T^*}$, could be obtained from the variance of the pseudo-values from the usual formula for the standard error of the mean as:

$$\widehat{\sigma}_{T^*} = \sqrt{\frac{\widehat{\sigma}^2_{T_n^*}}{N}} = \sqrt{\frac{\sum\left(T_n^* - \overline{T}_\bullet^*\right)^2}{N(N-1)}} \ . \tag{6}$$

This standard error can then be used to compute confidence intervals for the estimation of the parameter. Under the independence assumption, this estimation is distributed as a Student's t distribution with $(N-1)$ degrees of freedom. Specifically a $(1-\alpha)$ confidence interval can be computed as

$$T^* \pm t_{\alpha,\nu}\widehat{\sigma}_{T^*} \tag{7}$$

with $t_{\alpha,\nu}$ being the α-level critical value of a Student's t distribution with $\nu = N-1$ degrees of freedom.

2.1 Jackknife without pseudo-values

Pseudo-values are important to understand the inner working of the jackknife, but they are not computationally efficient. Alternative formulas using only the partial estimates can be used in lieu of the pseudo-values. Specifically, if $\overline{T}_\bullet$ denotes the mean of the partial estimates and $\widehat{\sigma}_{T_{-n}}$ their standard deviation, then T^* (*cf.* Equation 4) can be computed as

$$T^* = NT - (N-1)\overline{T}_\bullet \tag{8}$$

and $\widehat{\sigma}_{T^*}$ (*cf.* Equation 6) can be computed as

$$\widehat{\sigma}_{T^*} = \sqrt{\frac{N-1}{N}\sum\left(T_{-n} - \overline{T}_\bullet\right)^2} = (N-1)\frac{\widehat{\sigma}_{T_{-n}}}{\sqrt{N}} \tag{9}$$

2.2 Assumptions of the Jackknife

Although the jackknife makes no assumptions about the shape of the underlying probability distribution, it requires that the observations are independent of each other. Technically, the observations are assumed to be independent and identically distributed (*i.e.,* in statistical jargon: "i.i.d."). This means that the jackknife is not, in general, an appropriate tool for time series data. When the independence assumption is violated, the jackknife underestimates the variance in the data-set which makes the data look more reliable than they actually are.

Because the jackknife eliminates the bias by subtraction (which is a linear operation), it works correctly only for statistics which are *linear* functions of the parameters or the data, and whose distribution is continuous or at least "smooth enough" to be considered as such. In some cases, linearity can be achieved by transforming the statistics (*e.g.,* using a Fisher Z-transform for correlations, or a logarithm transform for standard deviations), but some non-linear or non-continuous statistics, such as the median, will give very poor results with the jackknife no matter what transformation is used.

2.3 Bias estimation

The jackknife was originally developed by Quenouille as a nonparametric way to estimate and reduce the bias of an estimator of a population parameter. The bias of an estimator is defined as the difference between the expected value of this estimator and the true value of the population parameter. So formally, the bias,

denoted $\mathcal{B}$, of an estimation T of the parameter θ is defined as

$$\mathcal{B} = \mathbf{E}\{T\} - \theta \ , \tag{10}$$

with $\mathbf{E}\{T\}$ being the expected value of T.

The jackknife estimate of the bias is computed by replacing the expected value of the estimator (*i.e.,* $\mathbf{E}\{T\}$) by the biased estimator (*i.e.,* T) and by replacing the parameter (*i.e.,* θ) by the "unbiased" jackknife estimator (*i.e.,* T^*). Specifically, the jackknife estimator of the bias, denoted $\mathcal{B}_{\text{jack}}$ is computed as:

$$\mathcal{B}_{\text{jack}} = T - T^* \ . \tag{11}$$

2.4 Generalizing the performance of predictive models

Recall that the name "jackknife" refers to two related, but different techniques (and this is sometimes a source of confusion). The first technique, presented above, estimates population parameters and their standard error. The second technique evaluates the generalization performance of predictive models. In these models, predictor variables are used to predict the values of dependent variable(s). In this context, the problem is to estimate the quality of the prediction for *new* observations. Technically speaking, the goal is to estimate the performance of the predictive model as a *random effect* model. The problem of estimating the random effect performance for predictive models is becoming a crucial problem in domains such as, for example, bio-informatics and neuroimaging (see, *e.g.,* Kriegeskorte *et al.*, 2009; Vul *et al.*, 2009) because the data sets used in these domains are typically comprised of a very large number of variables (often a much larger number of variables than observations—A configuration called the "small N, large P" problem). This large number of variables makes statistical models notoriously prone to over-fitting.

In this context, the goal of the jackknife is to estimate how a model would perform when applied to *new* observations. This is done by dropping in turn each observation and fitting the model for the remaining set of observations. The model is then used to predict the left-out observation. With this procedure, each observation has been predicted as a new observation.

In some cases a jackknife can perform both functions, thereby generalizing the predictive model as well as finding the unbiased estimate of the parameters of the model.

3 Example: Linear regression

Suppose that we performed a study examining children's speech rate as a function of their age. The children's age (denoted X) was used as a predictor of their speech rate (denoted Y). Dividing the number of words said by the time needed to say them gave the speech rate (expressed in words per minute) of each child. The results of this (fictitious) experiment are shown in Table 1.

We will use these data to illustrate how the jackknife can be used to 1) estimate the regression parameters and their bias and 2) evaluate the generalization performance of the regression model. As a preliminary step, the data are analyzed by a standard regression analysis and we found that the regression equation is equal to:

$$\widehat{Y} = a + bX = 90 + 1.25X \ . \tag{12}$$

The predicted values are given in Table 1. This regression model corresponds to a coefficient of correlation of $r = .8333$ (*i.e.,* the correlation between the Y-s and the $\widehat{Y}$-s is equal to .8333).

3.1 Estimation of regression parameters and bias

In this section, we will use the jackknife to estimate the intercept, the slope, and the value of the coefficient of correlation for the regression.

We drop each observation in turn and compute, for the slope and the intercept, the partial estimates (denoted b_{-n} and a_{-n}) and pseudo-values (denoted b_n^* and a_n^*). So, for example, when we drop the first observation, we use the observations

Table 1: Data from a study examining children's speech rate as a function of age. The independent variable is the age of the child (X). The dependent variable is the speech rate of the child in words per minutes (Y). The values of $\widehat{Y}$ are obtained as $\widehat{Y} = 90 + 1.25X$. X_n is the value of the independent variable, Y_n is the value of the dependent variable, $\widehat{Y}_n$ is the predicted value of the dependent variable predicted from the regression, $\widehat{Y}_n^*$ is the predicted value of the dependent variable predicted from the jackknife derived unbiased estimates, $\widehat{Y}_{\text{jack}}$ is the predicted values of the dependent variable when each value is predicted from the corresponding jackknife partial estimates.

Obs	X_n	Y_n	$\widehat{Y}_n$	$\widehat{Y}_n^*$	$\widehat{Y}_{\text{jack},n}$
1	4	91	95.0000	94.9986	97.3158
2	5	96	96.2500	96.1223	96.3468
3	6	103	97.5000	97.2460	95.9787
4	9	99	101.2500	100.6172	101.7411
5	9	103	101.2500	100.6172	100.8680
6	15	108	108.7500	107.3596	111.3962

2 through 6 to compute the regression equation with the partial estimates of the slope and intercept as (*cf.* Equation 2):

$$\widehat{Y}_{-1} = a_{-1} + b_{-1}X = 93.5789 + 0.9342X \ . \tag{13}$$

From these partial estimates, we compute a pseudo-value by adapting Equation 3 to the regression context. This gives the following jackknife pseudo values for the nth observation:

$$a_n^* = Na - (N-1)a_{-n} \quad \text{and} \quad b_n^* = Nb - (N-1)b_{-n} \ , \tag{14}$$

and for the first observation, this equation becomes:

$$a_1^* = 6 \times 1.25 - 5 \times 0.9342 = 2.8289 \quad \text{and} \quad b_1^* = 6 \times 90 - 5 \times 93.5789 = 72.1053 \ . \tag{15}$$

Table 2 gives the partial estimates and pseudo values for the intercept and slope of the regression. From this table we can find that the jackknife estimates of the regression will give the following equation for the prediction of the dependent variable (the prediction using the jackknife estimates is denoted $\widehat{Y}_n^*$):

$$\widehat{Y}_n^* = a^* + b^*X = 90.5037 + 1.1237X \ . \tag{16}$$

The predicted values using the jackknife estimates are given in Table 1. It is worth noting that, for regression, the jackknife parameters are *linear* functions of the standard estimates. This implies that the values of $\widehat{Y}_n^*$ can be perfectly predicted from the values of $\widehat{Y}_n$. Specifically,

$$\widehat{Y}_n^* = \left(a^* - a\frac{b^*}{b}\right) + \frac{b^*}{b}\widehat{Y}_n \ . \tag{17}$$

Therefore the correlation between the $\widehat{Y}_n^*$ and the $\widehat{Y}_n$ is equal to one, this, in turn, implies that the correlation between the original data and the predicted values is the same for both $\widehat{Y}$ and $\widehat{Y}_n^*$.

The estimation for the coefficient of correlation is slightly more complex because, as mentioned earlier, the jackknife does not perform well with non-linear statistics such as correlation. So, the values of r are transformed using Fisher Z-transform (see Abdi *et al.*, 2009) *prior* to jackknifing. The jackknife estimate is computed on these Z-transformed values, and the final value of the estimate of r is obtained by using the inverse of the Fisher Z-transform (using r rather than the transformed Z values would lead to a gross over-estimation of the correlation). Table 2 gives the partial estimates for the correlation, the Z-transformed values,

Table 2: Partial estimates and pseudo-values for the regression example of Table 1

Obs	Partial Estimates				Pseudo-Values		
	a_{-n}	b_{-n}	r_{-n}	Z_{-n}	a_n^*	b_n^*	Z^*
1	93.5789	0.9342	.8005	1.1001	72.1053	2.8289	1.6932
2	90.1618	1.2370	.8115	1.1313	89.1908	1.3150	1.5370
3	87.4255	1.4255	.9504	1.8354	102.8723	0.3723	−1.9835
4	90.1827	1.2843	.8526	1.2655	89.0863	1.0787	0.8661
5	89.8579	1.2234	.8349	1.2040	90.7107	1.3832	1.1739
6	88.1887	1.5472	.7012	0.8697	99.0566	−0.2358	2.8450
Mean	$\overline{a}_\bullet$	$\overline{b}_\bullet$	—	$\overline{Z}_\bullet$	a^*	b^*	Z^*
	89.8993	1.2753	—	1.2343	90.5037	1.1237	1.0219
					Jackknife Estimates		
SD	$\widehat{\sigma}_{a_{-n}}$	$\widehat{\sigma}_{b_{-n}}$	—	$\widehat{\sigma}_{Z_{-n}}$	$\widehat{\sigma}_{a_n^*}$	$\widehat{\sigma}_{b_n^*}$	$\widehat{\sigma}_{Z_n^*}$
	2.1324	0.2084	—	0.3240	10.6622	1.0418	1.6198
					Jackknife Standard Deviations		
SE	—	—	—	—	$\widehat{\sigma}_{a^*} = \frac{\widehat{\sigma}_{a_n^*}}{\sqrt{N}}$	$\widehat{\sigma}_{b^*} = \frac{\widehat{\sigma}_{b_n^*}}{\sqrt{N}}$	$\widehat{\sigma}_{Z^*} = \frac{\widehat{\sigma}_{Z_n^*}}{\sqrt{N}}$
	—	—	—	—	4.3528	0.4253	.6613
					Jackknife Standard Error		

and the Z-transformed pseudo-values. From Table 2, we find that the jackknife estimate of the Z-transformed coefficient of correlation is equal to $Z^* = 1.019$ which, when transformed back to a correlation, gives a value of the jackknife estimate for the correlation of $r^* = .7707$. Incidently, this value is very close to the value obtained with another classic alternative population unbiased estimate called the shrunken r, which is denoted $\widetilde{r}$, and computed as

$$\widetilde{r} = \sqrt{1 - \left[(1 - r^2)\frac{(N-1)}{(N-2)}\right]} = \sqrt{1 - \left[(1 - .8333^2)\frac{5}{4}\right]} = .7862 \ . \tag{18}$$

Confidence intervals are computed using Equation 7. For example, taking into account that the $\alpha = .05$ critical value for a Student's t distribution for $\nu = 5$ degrees of freedom is equal to $t_{\alpha,\nu} = 2.57$, the confidence interval for the intercept

is equal to:

$$a^* \pm t_{\alpha,\nu}\widehat{\sigma}_{a^*} = 90.5037 \pm 2.57 \times \frac{10.6622}{\sqrt{6}} = 90.5037 \pm 2.57 \times 4.3528 = 90.5037 \pm 11.1868 \ . \tag{19}$$

The bias of the estimate is computed from Equation 11. For example, the bias of the estimation of the coefficient of correlation is equal to:

$$\mathcal{B}_{\text{jack}}(r) = r - r^* = .8333 - .7707 = .0627 \ . \tag{20}$$

The bias is positive and this shows (as expected) that the coefficient of correlation over-estimates the magnitude of the population correlation.

3.2 Estimate of the generalization performance of the regression

In order to estimate the generalization performance of the regression, we need to evaluate the performance of the model on *new* data. These data are supposed to be randomly selected from the same population as the data used to build the model. The jackknife strategy here is to predict each observation as a new observation, this implies that each observation is predicted from its *partial estimates* of the prediction parameter. Specifically, if we denote by $\widehat{Y}_{\text{jack},n}$ the jackknife predicted value of the nth observation, the jackknife regression equation becomes:

$$\widehat{Y}_{\text{jack},n} = a_{-n} + b_{-n}X_n \ . \tag{21}$$

So, for example, the first observation is predicted from the regression model built with observations 2 to 6, this gives the following predicting equation for $\widehat{Y}_{\text{jack},1}$ (*cf.* Tables 1 and 2):

$$\widehat{Y}_{\text{jack},1} = a_{-1} + b_{-1}X_1 = 93.5789 + 0.9342 \times 4 = 97.3158 \ . \tag{22}$$

The jackknife predicted values are listed in Table 1. The quality of the prediction of these jackknife values can be evaluated, once again, by computing a coefficient of correlation between the predicted values (*i.e.,* the $\widehat{Y}_{\text{jack},n}$) and the actual values (*i.e.,* the Y_n). This correlation, denoted r_{jack}, for this example is equal to $r_{\text{jack}} = .6825$. It is worth noting that, in general, the coefficient r_{jack} is *not* equal to the jackknife estimate of the correlation r^* (which, recall, is in our example equal to $r^* = .7707$).

4 see also

Bias; Bootstrapping; Coefficients of correlation, alienation and determination; Pearson product-moment correlation; R^2 (R-squared); Regression; Reliability; Standard error of estimate

Further readings

1. Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental Design and analysis for psychology*. Oxford: Oxford University Press.
2. Bissel (1975). The jackknife—toy, tool or two-edged weapon? *The Statistician*, **24**, 79–100.
3. Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American,* May, 116-130.
4. Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
5. Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* **37**, 36-48.
6. Gentle, J.E. (1998). *Elements of computational statistics*. New York: Springer.
   Kriegeskorte, K., Simmons, W.K., Bellgowan, P.S.F., & Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, **12**, 535–540.
7. Krzanowski, W.J., & Radley, D. (1989). Nonparametric confidence and tolerance regions in canonical variate analysis. *Biometrics*, **45**, 1163–1173.
8. Hinkley, D.V., (1983). Jackknife methods. In, Johnshon, N.L., Kotz, S., & Read, C.B. (Eds). *Encyclopedia of Statistical Sciences (Volume 4)*. New York: Wiley. pp 280–287.
9. Manly, B.F.J. (1997). *Randomization, Bootstrap, and Monte Carlo methods in biology (2nd Edition)*. New York: Chapman & Hall.
10. Miller, R.G. (1974). The jackknife: a review. *Biometrika*, **61**, 1–17.
11. Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.
12. Shao, J., & Tu, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer Verlag.
13. Tukey, J.W. (1958). Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics*, **29**, 614.
14. Tukey, J.W., (1986). The future of processes of data analysis. In *The Collected Works of John W. Tukey (Volume IV)*. New York: Wadsworth. pp 517–549.
15. Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives in Psychological Sciences*, **4**, 274–290.