Jackknife Estimator of Species Richness with S-PLUS

Christina D. Smith  
Kansas State University

Jeffrey S. Pontius  
Kansas State University

Abstract

An estimate of the number of species, $S$, usually called species richness by ecologists, in an area is one of the basic statistics used to ascertain biological diversity. Traditionally ecologists have used the number of species observed in a sample, $S_0$, to estimate $S$, realizing that $S_0$ is a lower bound for $S$. One alternative to $S_0$ is to use a nonparametric procedure such as jackknife resampling. For species richness, a closed form of the jackknife estimator is available. Typically statistical software contains only the traditional iterative form of the jackknife estimator. The purpose of this article is to propose an S-PLUS function for calculating the noniterative first order jackknife estimator of species richness and some associated plots and statistics.

Keywords: nonparametric estimation, species counts.

1. Introduction

Estimating the true number of species in an area, $S$, usually called species richness by ecologists, is one of the basic statistics used to ascertain biological diversity. To estimate species richness one would naturally consider the observed count of species, $S_0$, from a given sample. However, it is clear that $S_0$ is a lower bound for the true number of species. For $S_0$ to accurately estimate $S$ the researcher must actually observe every species. If the researcher can only sample a few plots from the area, then $S_0$ is likely to be smaller than $S$. Even if a census of the area is done it is likely that some species will be missed because of human error, environmental fluctuations that effect observations, or very small species detection probabilities.

2. Jackknife estimation

In the late 1970s statisticians and ecologists began to avidly look for alternative procedures for estimating $S$. The estimators considered included frequentist, Bayesian and nonparametric
philosophies, and sampling from finite and infinite populations (Mingoti and Meeden 1992; Bunge and Fitzpatrick 1993).

One alternative, presented by Smith and van Belle (1984), to using $S_0$ as an estimator of species richness is to use a nonparametric procedure such as jackknife resampling. The jackknife is useful because it is known to reduce bias and, for estimates of species richness, it has a closed form. Another useful characteristic of the jackknife estimator of species richness is that the estimator is based on the presence or absence of a species in a given plot rather than on the abundance of the species. To use the jackknife estimator for species richness, data must be collected at $n$ locations (e.g., plots) in the designated area for which $S$ is to be estimated.

The basic idea behind the first order jackknife estimator of $S$ is to base it on the amount of unique species information that is contained in each observation. Following Smith and van Belle (1984)

1. Remove one of the observations, say, $i$, where $i \in \{1, 2, \ldots, n\}$ denotes the labels of the sample units.
2. Compute an estimate of $S$, $\hat{S}_{-i}$, on all observations excluding $i$.
3. Compute the pseudovalue $\hat{S}_i = n\hat{S} - (n - 1)\hat{S}_{-i}$, where $\hat{S}$ is an estimate of $S$ based on all $n$ observations.
4. Repeat 1. through 3. for each $i$, $i = 1, 2, \ldots, n$.
5. The first-order jackknife estimator of $S$ is

$$J_n = \frac{1}{n} \sum \hat{S}_i.$$

Note that in step 1 two observations could be removed, and in fact, as many as $n - 1$ observations could be removed to obtain higher order jackknife estimators (Smith and van Belle 1984).

A closed form solution to the jackknife algorithm is available. Here the jackknife estimator depends on the number of unique species in the removed observations (e.g. plots). The closed form of the first order jackknife estimator of species richness, as given by Smith and van Belle (1984), is

$$J_n(S) = S_0 + \frac{n - 1}{n} \sum_{i=1}^{n} r_i,$$

where $S_0$ is the observed species count over all plots, $r_i$ is the number of species that are found only in plot $i$, and $n$ is the number of plots. Note that when all species are observed on at least two plots, $J_n(S) = S_0$ because $r_i = 0$ for all $i = 1, 2, \ldots, n$. When there is more variability between observations the $r_i$‘s and $J_n(S)$ become larger.

An estimator of the variance of $J_n(S)$ is given by

$$\widehat{\var} [J_n(S)] = \frac{n - 1}{n} \sum_{i=1}^{n} \left( r_i - \frac{1}{n} \sum_{i=1}^{n} r_i \right)^2.$$
This is a measure of the average deviation of the $r_i$’s from the observed mean of the $r_i$’s. Our 
\texttt{S-PLUS} function reports the standard error of $J_n(S)$, $\sqrt{\text{VAR}[J_n(S)]}$.

### 3. Performance of the jackknife estimator

A few researchers have evaluated the performance of $J_n(S)$ including Smith and van Belle (1984), Palmer (1990), and Hellmann and Fowler (1999). Smith and van Belle (1984) evaluated $J_n(S)$ under the assumption that the abundance of a given species has a Poisson distribution. They showed that $J_n(S)$ is less biased than $S_0$ and that the expected bias approaches zero as the species density (number of species per plot) increases.

Palmer (1990) evaluated $J_n(S)$ based on samples taken from hardwood stands in North Carolina. A census was taken at 30 locations to obtain the “true” species richness, then samples were taken from plots on the 30 locations. Palmer used the mean deviation to show that $J_n(S)$ has less bias than $S_0$, and used the mean squared deviation to show that $J_n(S)$ has less variability (more precision) than $S_0$.

Hellmann and Fowler (1999) considered the bias, precision, and accuracy of $J_n(S)$ based on samples from five different forested locations in Michigan. Each location contained 160 plots. The five locations had different types of tree growth and ranged from 5 total species to 25 total species. The 160 plots at each location were considered to be the population of plots, and samples of different sizes were taken from each population.

Hellmann and Fowler’s results indicated that $J_n(S)$ is less biased than $S_0$ when less than 60% of the “population” is sampled and that $J_n(S)$ is typically less precise than $S_0$ but it is usually more accurate than $S_0$. Note that Hellmann and Fowler measured precision by $\text{VAR}[J_n(S)]$ and measured accuracy by $\text{MSE}[J_n(S)]$. They also pointed out that the characteristics of $J_n(S)$, as well as $S_0$, depend on the sample size.

Note that Palmer (1990) had different results regarding precision than Hellmann and Fowler (1999). This may be because they were looking at different data sets. However, they are not exactly clear about their definitions so it is possible that they observed different results because they were looking at different characteristics or using different estimates of precision.

### 4. Algorithm for calculating the jackknife estimate

\texttt{S-PLUS} does contain an iterative jackknife procedure, but as mentioned previously, a closed form jackknife estimator exists for estimating $S$. The following outlines an algorithm for an \texttt{S-PLUS} function which calculates a noniterative first order jackknife estimate of species richness for each of several sampling periods (e.g., years). The difficult task is in identifying the number of unique species in each plot.

1. The data set should have the following headings: “Period” for identifying the sampling period (e.g. years), “Plot” for each unique sampling location, and “Species” for the actual species observed in each period on each plot.

2. Identify the number of periods and the number of plots in the data set for future use.

3. Create a matrix that contains the number of unique species on each plot for each year.
(a) Create a storage matrix with the number of rows equal to the total number of plots, and with three columns for sampling period, plot, and count.

(b) Identify the plots listed for each period, and compute the total number of species for each period.

(c) Identify the number of species not on plot $i$ and subtract number of species not on plot $i$ from the total number of species. This is the number of unique species on plot $i$, $r_i$.

(d) Fill the storage matrix with the information obtained in steps (b) and (c) and assign header names: “Period”, “Plot”, “Count”.

4. Create a matrix of jackknife estimates.

(a) Create a storage matrix with the number of rows equal to the number of sampling periods in the data set, and with columns for sampling period, the observed count, the jackknife estimate, the standard error of the jackknife estimate, and the number of plots for the given period.

(b) Identify the plots listed for each year and compute the total number of species for each sampling period.

(c) Calculate $\sum_{i=1}^{n} r_i$ and $\sum_{i=1}^{n} \left( r_i - \frac{1}{n} \sum_{i=1}^{n} r_i \right)^2$ for each period from the data set created in step 3.

(d) Multiply by the appropriate constants to get the estimate of the first order jackknife and the estimate of its variance.

(e) Fill the storage matrix and assign header names: ”Period”, ”Number of Plots”, ”Observed”, ”Jackknife Estimate”, ”Standard Error”.

5. S-PLUS functions

The basic structure of our S-PLUS function follows the previously stated algorithm with some additional plots and statistics. This function also calculates the 95% standard error of the first order jackknife estimate and calculates a standard normal confidence interval. Note that the use of a normal confidence interval is appropriate if the number of plots is large, say, $n \geq 30$. Our function produces a table identifying the number of plots in which unique species occur for each sampling period, a set of notched box plots (McGill, Tukey, and Larsen 1978) of the number of species on each plot for each sampling period, and a dot plot (Cleveland 1984) of the observed counts and first order jackknife estimates.

The following is our S-PLUS code. Note that the function, `jack.fun`, calls the functions `species.boxplot` and `jackone.plot`, which are also displayed here. The functions were coded in S-PLUS 5.1 (Insightful Corporation 1999) for Unix operating systems (see, for example, Krause and Olson (2000)). The function has also run successfully in S-PLUS 6.2 on Windows XP. Note that the function will calculate $J_n(S)$ and $S_0$ in R if the plotting sections are removed.

The function `jack.fun` has five arguments. The first argument, `mydata`, should be replaced with the name of the data frame containing the data set you wish to use. An example of
how the data set (data frame) should look is given in Table 1. The default alpha level for the second argument is 0.05 but it can be changed using alpha. The argument unique.species creates a table identifying the number of plots containing possible numbers of unique species, the default is true. If no table is desired set unique.species = F. The argument box.plot calls a function which creates box plots of the number of species per plot for each period. If the box plots are not desired set box.plot = F. Finally, the argument est.plot calls a function which creates a dot plot of the observed counts and the first order jackknife estimates. If the plot is not desired set est.plot = F. Note that jack.fun will always output a table of the observed count and the jackknife estimate for each period (see example in Section 6).

The following is the S-PLUS code for our original function for jackknife estimation of species richness, jack.fun.

```r
jack.fun <- function(mydata, alpha=0.05, unique.species=T, box.plot=T, est.plot=T) {
  ##install data set##
  attach(mydata)
  on.exit(detach(mydata))

  ##get number of periods##
  periods <- length(unique(Period))

  ##get total number of plots over all periods##
  num.row <- 0
  for(y in unique(Period)) {
    plots <- length(unique(Plot[Period==y]))
    num.row <- num.row + plots
  }
```

Table 1: Example of data for use in jack.fun
##create matrix of unique number of species for each plot for each period##

```r
jack.data <- matrix(NA, nrow=num.row, ncol=3)
index.row <- 0
for(y in unique(Period)) {
  plot.index <- sort(unique(Plot[Period==y]))
  all.species <- length(unique(Species[Period==y]))
  for(p in plot.index) {
    index.row <- index.row + 1
    jack.data[index.row,1] <- y
    jack.data[index.row,2] <- p
    jack.data[index.row,3] <- all.species
       - length(unique(Species[(Period==y)&!(Plot==p)]))
  }
}
headers <- c("Period", "Plot", "Count")
dimnames(jack.data) <- list(NULL, headers)
jack.data <- data.frame(jack.data)
##create table for number of unique species in each period##
if (unique.species) {
  species.unique <- table(jack.data$Period, jack.data$Count)
}
##create matrix of jackknife estimates of species richness for each Period##
jackone.est <- matrix(NA, nrow=periods, ncol=7)
index.row <- 1
for (y in unique(jack.data$Period)) {
  all.species <- length(unique(Species[Period==y]))
  plot.count <- length(unique(Plot[Period==y]))
  jacksum <- 0
  for (p in sort(unique(jack.data$Plot[jack.data$Period==y]))) {
    jacksum <- jacksum+jack.data$Count[(jack.data$Period==y)&(jack.data$Plot==p)]
  }
  jackvar <- 0
  for (p in sort(unique(jack.data$Plot[jack.data$Period==y]))) {
    jackvar <- jackvar + (jack.data$Count[(jack.data$Period==y)&(jack.data$Plot==p)] -
                          (jacksum/plot.count)) *
                          (jack.data$Count[(jack.data$Period==y)&(jack.data$Plot==p)] -
                          (jacksum/plot.count))
  }
  jackone.estimate <- all.species + ((plot.count-1)/plot.count)*jacksum
  jackone.variance <- ((plot.count-1)/plot.count)*jackvar
  jackone.est[index.row,1] <- y
  jackone.est[index.row,2] <- plot.count
  jackone.est[index.row,3] <- all.species
  jackone.est[index.row,4] <- jackone.estimate
  jackone.est[index.row,5] <- sqrt(jackone.variance)
```

The following is S-PLUS code for the function `species.boxplot` which is called by the function `jack.fun` for creating a variable width notched box plots for the number of species per plot for each year. For an example of `species.boxplot` output see Figure 1.

```splus
species.boxplot <- function(data) {
  ##get total number of plots over all periods##
  sum <- 0
  for(y in unique(Period)) {
    plots <- length(unique(Plot[Period==y]))
    sum <- sum + plots
  }

  ##create matrix for number of species on each plot in each period##
  species.data <- matrix(NA, nrow=sum, ncol=3)
  index.row <- 0
  for(y in unique(Period)) {
    plot.index <- unique(Plot[Period==y])
    for(p in plot.index) {
      index.row <- index.row + 1
      species.count <- length(unique(Species[(Period==y) & (Plot==p)]))
      species.data[index.row,1] <- y
      species.data[index.row,2] <- p
      species.data[index.row,3] <- species.count
    }
  }

  headers <- c("Period", "Plot", "Count")
  dimnames(species.data) <- list(NULL, headers)
  species.plot <- boxplot(split(species.data$Count, species.data$Period),
                           varwidth = T, notch = T, main = "Species Counts per Plot", xlab = "Period",
                           ylab = "Number of Species per Plot")
}
```
The following is S-PLUS code for the function jackone.plot which is called by the function jack.fun for creating a dot plot of the jackknife estimates produced by jack.fun. For an example of jackone.plot output see Figure 2.

```r
jackone.plot <- function(jackone.est) {
    # create and attach table from jackknife procedure output#
    periods <- matrix(c(jackone.est[,1], jackone.est[,1]), ncol=1)
    periods.length <- length(jackone.est[,1])
    type.obs <- matrix(NA, nrow=periods.length, ncol=1)
    type.obs[,1] <- 0
    type.jkl <- matrix(NA, nrow=periods.length, ncol=1)
    type.jkl[,1] <- 1
    type <- as.matrix(rbind(type.obs, type.jkl))
    est <- matrix(c(jackone.est[,3], jackone.est[,4]), ncol=1)
    jackone.table <- cbind(periods, type, est)
    headers <- c("period", "type", "est")
    dimnames(jackone.table) <- list(NULL, headers)
    jackone.table <- as.data.frame(jackone.table)

    ## define trellis parameters##
    trellis.device(motif)
    strip.col <- trellis.par.get("strip.background")
    strip.col$col <- 0
    trellis.par.set("strip.background", strip.col)
    newdot.line <- trellis.par.get("dot.line")
    newdot.line$lty <- 2
    trellis.par.set("dot.line", newdot.line)
    newdot.line.col <- trellis.par.get("dot.line")
    newdot.line.col$col <- 1
    trellis.par.set("dot.line", newdot.line.col)
    newdot.line.lwd <- trellis.par.get("dot.line")
    newdot.line.lwd$wd <- 1
    trellis.par.set("dot.line", newdot.line.lwd)
    super.symbols <- trellis.par.get("superpose.symbol")
    super.symbols$pch <- c(1,2,0,3,4,5,6)
    trellis.par.set("superpose.symbol", super.symbols)
    new.symbol.col <- trellis.par.get("superpose.symbol")
    new.symbol.col$col <- c(rep(1,7))
    trellis.par.set("superpose.symbol", new.symbol.col)
    new.symbol.size <- trellis.par.get("superpose.symbol")
    new.symbol.size$cex <- c(rep(3,7))
    trellis.par.set("superpose.symbol", new.symbol.size)

    ## create plot of observed and jackknife estimates##
    estimates.plot <- dotplot(jackone.table[,1] ~ jackone.table[,3], 
                              groups = jackone.table[,2],
```
main = "Species Richness Estimates",
xlab = "species richness estimates",
key = list(text = list(c("Observed Count","First-order Jackknife")),
points = Rows(trellis.par.get("superpose.symbol"), 1:2)),
strip = function(...) { strip.default(...,style=1) },
panel = function(x,y,...) {
  dot.line <- trellis.par.get("dot.line")
  abline(h=unique(y), lwd=dot.line$lwd, lty=dot.line$lty,
         col=dot.line$col)
  panel.superpose(x,y,...)
}

print(estimates.plot)
}

6. Example

We provide an example of species richness estimates based on data collected for the Land
Condition Trend Analysis (LCTA) project at Fort Riley, KS. The LCTA project monitors
the environment at the fort and collects information on soil, vegetation, birds and mammals.
The data for birds has been collected from 1991 through 2002 on approximately 60 plots per
year (sampling period). The data include the year, the plot, and the species found on each
plot for each year.

Table 2 contains a list of the number of plots in which unique species occur. Notice that for
our data most plots contain zero unique species. In 1991, only 8 plots contained one unique
species and no plots contained 2 or 3 unique species. Table 3 displays the year, the number of

| Year | 0 | 1 | 2 | 3 |
|------|---|---|---|---|
| 1991 | 51| 8 | 0 | 0 |
| 1992 | 55| 4 | 0 | 1 |
| 1993 | 50| 4 | 0 | 0 |
| 1994 | 47| 7 | 2 | 0 |
| 1995 | 50| 7 | 1 | 0 |
| 1996 | 54| 5 | 1 | 0 |
| 1997 | 52| 4 | 1 | 1 |
| 1998 | 49| 10| 1 | 0 |
| 1999 | 51| 7 | 0 | 0 |
| 2000 | 50| 6 | 2 | 0 |
| 2001 | 48| 9 | 2 | 0 |
| 2002 | 51| 5 | 2 | 0 |

Table 2: Number of plots in which unique species occur
Figure 1: Box plots of the number of species per plot per period

Figure 2: Plot of species richness estimates per period
Table 3: Species richness estimates

| Year | Plots | Observed Count | Jackknife Estimate | Standard Error | Lower Limit | Upper Limit |
|------|-------|----------------|--------------------|---------------|-------------|-------------|
| 1991 | 59    | 44             | 51.86              | 2.61          | 46.75       | 56.97       |
| 1992 | 60    | 45             | 51.88              | 3.46          | 45.10       | 58.67       |
| 1993 | 54    | 36             | 39.93              | 1.91          | 36.19       | 43.66       |
| 1994 | 56    | 40             | 50.80              | 3.55          | 43.84       | 57.76       |
| 1995 | 58    | 40             | 48.84              | 3.07          | 42.82       | 54.87       |
| 1996 | 60    | 41             | 47.88              | 2.84          | 42.32       | 53.44       |
| 1997 | 58    | 42             | 50.84              | 3.92          | 43.17       | 58.52       |
| 1998 | 60    | 42             | 53.80              | 3.38          | 47.18       | 60.42       |
| 1999 | 58    | 37             | 43.88              | 2.46          | 39.06       | 48.70       |
| 2000 | 58    | 41             | 50.83              | 3.47          | 44.02       | 57.64       |
| 2001 | 59    | 41             | 53.78              | 3.73          | 46.47       | 61.09       |
| 2002 | 58    | 43             | 51.84              | 3.38          | 45.23       | 58.46       |

plots sampled each year, the observed count, $S_0$, the jackknife estimate, $J_n(S)$, the standard error of $J_n(S)$, and the lower and upper limits of the 95% standard normal confidence interval for $S$ based on $J_n(S)$. At first glance the standard errors for $J_n(S)$ may seem unreasonably small. However, as noted, the numbers of unique species are very small which drives down the variances of $J_n(S)$. Figure 1 contains a variable width notched box plots for the number of species per plot for each year. Figure 2 is a dot plot of $S_0$ and $J_n(S)$ for each year.

7. Summary

We have demonstrated the need for a function which calculates first order jackknife estimates for species richness and how to implement such a function. Note that a function could be written for any order jackknife procedure. However, the calculations for jackknife variance quickly become difficult. Also, based on our experience, the second order jackknife procedure does not give estimates that are much different from the first order jackknife estimates.

Acknowledgments

Partial funding was provided under RWO37, Land Condition Trend Analysis on Fort Riley, KS, by the United States Geological Service.

References

Bunge J, Fitzpatrick M (1993). “Estimating the Number of Species: A Review.” Journal of the American Statistical Association, 88, 364–373.

Cleveland WS (1984). Visualizing Data. Hobart Press, Summit, NJ.
Hellmann J, Fowler G (1999). “Bias, Precision, and Accuracy of Four Measures of Species Richness.” *Ecological Applications*, 9, 824–834.

Insightful Corporation (1999). *S-PLUS (Version 5.1)*. Seattle, WA. URL http://www.insightful.com/.

Krause A, Olson M (2000). *The Basics of S and S-PLUS*. Springer-Verlag, New York, NY, 2nd edition.

McGill R, Tukey JW, Larsen WA (1978). “Variations of Box Plots.” *The American Statistician*, 32, 12–16.

Mingoti S, Meeden G (1992). “Estimating the Total Number of Distinct Species Using Presence and Absence Data.” *Biometrics*, 48, 863–875.

Palmer M (1990). “The Estimation of Species Richness by Extrapolation.” *Ecology*, 71, 1195–1198.

Smith E, van Belle G (1984). “Nonparametric Estimation of Species Richness.” *Biometrics*, 40, 119–129.

**Affiliation:**

Christina D. Smith
Kansas State University
Department of Statistics
101 Dickens Hall
Manhattan, Kansas, United States of America
E-mail: cdsmith@ksu.edu

Jeffrey S. Pontius
Kansas State University
Department of Statistics
101 Dickens Hall
Manhattan, Kansas, United States of America
E-mail: pontius@stat.ksu.edu

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 15, Issue 3 Submitted: 2004-09-30
January 2006 Accepted: 2005-12-05