

Computing a Test Statistic for Scale Parameters Based on a Jackknife Procedure

Jean G. Orelie, Research Triangle Institute, Research Triangle Park, NC

Abstract:

It is well known that the F test for equality of variance is not robust when the assumption of normality does not hold. When normality of the underlying distributions cannot be assumed, one can test for scale parameters using nonparametric methods. However, nonparametric tests for scale parameters have some disadvantages. First, the computations can be laborious; second, there is low efficiency compared to the F test when the distributions are normal. In this paper, using the IML® module in the SAS® System, we implement a test for scale parameters for two independent samples based on a Jackknife procedure proposed by R. G. Miller in 1968. The advantage of this test based on the Jackknife is that it does not require any assumptions for the location parameters. When the distributions are normal, this test has an asymptotic relative efficiency equal to 1 with respect to the F test.

Introduction

Given two independent samples, the classic approach to comparing their variances is to use an F test. Since the F test is not robust to nonnormality, an alternative is to resort to a nonparametric test in the event normality cannot be assumed. One disadvantage of nonparametric tests is the difficulty involved in the computations; however, some of these tests to compare scale parameters can be implemented in the SAS system. In particular, we will show how to implement a test that uses a Jackknife procedure in the IML module.

This Jackknife test was proposed by R. G. Miller[1]. The Miller Jackknife test is an asymptotically distribution free test that can be used to compare the scale parameters from two independent samples. This test offers many advantages over other nonparametric tests. Some nonparametric tests, for example, require that the medians be equal, an assumption that can be difficult to meet. The Jackknife test, on the other hand, does not require any assumption regarding the location parameters of the two samples. Another disadvantage of the most used tests for scale parameters is that they are not asymptotically efficient with respect to the F test when the two samples come from normal distribution. In contrast, the Miller

Jackknife test has an asymptotic efficiency equal to one when the underlying distributions are normal.

In this paper, we provide a SAS macro that handles the laborious computations involved in the Miller Jackknife test for any sample size. Since the Jackknife test is only asymptotically distribution free, we recommend using this macro for moderate to large sample sizes.

Assumptions of the Miller Jackknife Test

Assume we have m observations $X_1, \dots, X_m$ and n observations $Y_1, \dots, Y_n$. The X_i s and Y_j s are mutually independent. For the X_i s, μ_1 and σ_1 are respectively the location and scale parameters. The Y_j s have for location and scale parameters μ_2 and σ_2 . We also assume that $\frac{x_i - \mu_1}{\sigma_1}$ and $\frac{y_j - \mu_2}{\sigma_2}$ have the same continuous distribution ψ with a finite fourth moment.

This latter assumption may seem to be burdensome but it is required for the test statistic to converge asymptotically to a normal $N(0, 1)$ distribution. The proof of that claim is not in the scope of this paper. For further details the reader is referred to Miller[1].

$$\text{Let } \gamma^2 = \frac{\sigma_2^2}{\sigma_1^2}. \text{ We want to test:} \quad H_0: \gamma^2 = 1 \quad (1)$$

Procedure

Since we have two distributions on which we have no information, we can use a Jackknife procedure to obtain a robust estimate of a function that involves parameters from the two distributions. For two independent samples X and Y , the idea is to randomly divide the X 's into m' subgroups of size k_1 and the Y 's into n' subgroups of size k_2 . We then take natural logarithms of sample

variance of X's or Y's each time removing a subgroup of observations. Let S_i be the natural log of the sample variance of the X's after removing the i^{th} subgroup $i=1...m'$, T_j the natural log of the sample variance of the Y's after removing the j^{th} subgroup, S_o the natural log of the sample variance of all the X's and T_o the natural log of the sample variance of all the Y's. We can obtain a point estimate of the natural logarithm of the variance of the Xs by computing the distances

$A_i = m' S_o - (m' - 1) S_i$. The mean of these distances, $\bar{A}$, is an estimate of the natural log of the variance of X's. In a similar manner, we can compute $\bar{B}$, a point estimate of the natural log of the variance of the Y's. Note that $\bar{B} - \bar{A}$ estimates:

$$\ln(\sigma_2^2) - \ln(\sigma_1^2) = \ln(\sigma_2^2/\sigma_1^2) = \ln(\gamma^2). \quad (2)$$

We can remove the arbitrariness in the computations by taking $k_1 = k_2 = 1$. This means that to compute S_i or T_j we simply remove the i^{th} element of the X's or the j^{th} element for Y's. This choice of $k_1 = k_2 = 1$ is optimal to obtain a point estimate for the natural log of the ratio of the variances. The eleven steps below specify the algorithm to compute the test statistic for $k_1 = k_2 = 1$.

1. Compute S_i , the natural log of the variance of the X_i s with the i^{th} observation removed
2. Compute T_j , the natural log of the variance of the Y_j s with the j^{th} observation removed
3. Compute S_o , the natural log of all the X_i s
4. Compute T_o , the natural log of all the Y_j s
5. Compute $A_i = m S_o - (m - 1) S_i$
6. Compute $B_j = n T_o - (n - 1) T_j$

7. Compute $\bar{A}$ = sample mean of the A_i s. Note that $\bar{A}$ is an estimate of $\ln(\sigma_1^2)$
8. Compute $\bar{B}$ = sample mean of the B_j s. Note that $\bar{B}$ is an estimate of $\ln(\sigma_2^2)$

9. Compute
$$V_1 = \frac{\text{Sample Variance of the } A_i\text{s}}{m}$$

10. Compute
$$V_2 = \frac{\text{Sample Variance of the } B_j\text{s}}{n}$$

11. Compute the test statistic $Q = \frac{\bar{B} - \bar{A}}{\sqrt{V_1 + V_2}}$. Note that $Var(\bar{B} - \bar{A}) = Var(\bar{B}) + Var(\bar{A})$, since

$Cov(\bar{B}, \bar{A}) = 0$ because of the independence of X and Y .

Under the null hypothesis H_0 , Q has an asymptotic $N(0,1)$ distribution. For a desired level α two sided test, we reject H_0 if

$$|Q| > Z \text{ where } Prob(Z > z) = \alpha/2. \text{ Thus a}$$

100(1 - α)% confidence interval for

$$\ln(\gamma^2) = \ln(\sigma_2^2/\sigma_1^2) \text{ is } (\ln(\gamma_L^2), \ln(\gamma_U^2)) \text{ where}$$

$$\ln(\gamma_L^2) = [\bar{B} - \bar{A}] - Z_{(\alpha/2)} * \sqrt{(V_1 + V_2)} \text{ and}$$

$$\ln(\gamma_U^2) = [\bar{B} - \bar{A}] + Z_{(\alpha/2)} * \sqrt{(V_1 + V_2)}. \text{ Hence, we can}$$

derive an approximate 100(1 - α)% confidence interval for γ^2 to be (γ_L^2, γ_U^2) where

$$\gamma_L^2 = \exp\left([\bar{B} - \bar{A}] - Z_{(\alpha/2)} * \sqrt{(V_1 + V_2)}\right) \text{ and .}$$

$$\gamma_U^2 = \exp\left([\bar{B} - \bar{A}] + Z_{(\alpha/2)} * \sqrt{(V_1 + V_2)}\right)$$

Description of SAS Code

The SAS code that implements the algorithm of the previous section is included in Appendix A. The user must input a total of five macro variables. The macro variable DATA is for the input data set (either temporary or permanent) that contains the variables on which the test will be run. Macro variables X and Y are used for the analysis variables (which is assigned to X or Y is irrelevant). Finally, macro variables M and N indicate the number of observations in X and Y respectively.

The computations are carried out in the IML module of the SAS system. To apply the Jackknife, we treated each analysis variable as a vector. Consider one of the variables, X, with M observations. To create the M subsets from X that comes by removing an observation each time, we consider three cases. First we can remove the first element of that vector X, in which case we take the remaining second to m^{th} elements to form our subset. Second, we can remove the m^{th} element of the vector X, forming our subset by taking the first to the $(m-1)^{th}$ element. Finally, we can remove the i^{th} element that is neither the first nor the last element. To construct a subset for this latter case, we proceed in two phases. We first create a vector that consist of the first to the $(i-1)^{th}$ elements of X and a second vector that consists of taking the $(i+1)^{th}$ element to the m^{th} element of X. By concatenating these 2 vectors, we can then obtain a subset for the case where the i^{th} element to be removed is neither 1 or M. The computations involving the second analysis variable are handled in a similar manner.

The macro computes the test statistic and provide the p-value for a two sided test. By inverting the test statistic, we also give a 95% confidence interval for the ratio of the two scale parameters. Narrower or wider confidence intervals can be obtained easily by slightly modifying the macro. Since this test is only asymptotically distribution free, it is important to be careful with small sample sizes. With a small sample size, a t distribution could be used to compare the Q statistic. This macro assumes that there is an adequate sample size.

Example

Miller[1] and Shorack[2] carried out Monte Carlo simulations in which they showed that the Miller Jackknife test is more powerful than other nonparametric tests for scale parameters such as Ansari-Bradley, Moses

or Levene test. It is not the purpose of this paper to replicate these comparisons. To demonstrate what results can be expected overall with the Jackknife, we provide two examples. Using the SAS system, we generated Normal and Gamma random variables of various sample sizes. We compared a normal $N(3, 2)$ and a normal $N(7, 3)$ to test $\gamma=1$ for a normal distribution when the actual value of $\gamma=1.5$. We also compared a normal $N(5, 8)$ and a normal $N(11, 4)$ to test $\gamma=1$ when the actual value of $\gamma=2$. For the Gamma distributions, the ratios were the same. The results of these computations are provided in Table 1 and 2.

Table 1. Normal Distribution

Sample Size	$\gamma=1.5$		$\gamma=2$	
	Jackknife Estimate of Ratio	Pvalue	Jackknife Estimate of Ratio	Pvalue
30	1.41	0.350	1.80	0.170
50	1.50	0.140	2.07	0.037
100	1.45	0.100	2.13	< 0.001
200	1.61	0.001	2.14	< 0.001
500	1.50	< 0.001	2.00	< 0.001
1000	1.50	< 0.001	2.01	< 0.001

Table 2. Gamma Distribution

Sample Size	$\gamma=1.5$		$\gamma=2$	
	Jackknife Estimate of Ratio	Pvalue	Jackknife Estimate of Ratio	Pvalue
30	1.33	0.500	1.65	0.242
50	1.37	0.297	1.44	0.218
100	1.95	0.010	1.77	0.003
200	1.65	0.003	2.11	< 0.001
500	1.44	0.002	2.07	< 0.001
1000	1.43	< 0.001	2.01	< 0.001

Conclusion

The Miller Jackknife test provides a reasonable alternative to the classic approach for comparing the variance of two independent samples when one is not willing to make any assumptions regarding the underlying distributions of the two samples. Given that the test is only asymptotically distribution free, we recommend using it for adequate sample sizes. The macro we have written removes the computing burden that could be the primary reason for not using this test.

References

- [1] Miller, Rupert G., Jr.(1968). Jackknifing Variances, *Annals Mathematical Statistics*, Vol. 39, 567-582
- [2] Shorack, Galen R.(1969). Testing and estimating ratios of scale parameters, *Journal American Statistical Association*, Vol. 64, 999-1013.
- [3] Hollander, Myles and Wolfe, Douglas A(1973). Nonparametric Statistical Methods. John Wiley & Sons.

Acknowledgements

The author wishes to thank the following colleagues for their remarks and comments:

Audrey Bearden-Branch
David Kellog
Larry Myers
Akhil Vaish

A special thanks to Dr. Jacqueline Dietz, of North Carolina State University who also provided very helpful comments.

SAS, SAS/IML are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Jean G. Orelie, P. O. Box 12194, Research Triangle Park, NC, 27709. (919)541-5826. Orelie@rti.org.

Appendix A: SAS CODE FOR THE MACRO

```

/*****
*****

```

```

M:\SASFILES\MILLER.MAC

```

```

THIS PROGRAM COMPUTES THE MILLER
JACKKNIFE STATISTIC. IT IMPLEMENTS THE
ALGORITHM DEVELOPPED BY R. G. MILLER.

```

```

*****
*****/

```

```

/*****
*****

```

```

VARIABLES NEEDED:

```

```

FOR THIS MACRO TO WORK, YOU NEED TO
DEFINE THE FOLLOWING MACRO VARIABLES

```

```

DATA: YOUR DATA SET NAME;

```

```

XVAR: THE NAME OF YOUR X VARIABLE.

```

```

YVAR: THE NAME OF YOUR Y VARIABLE

```

```

M: THE LENGTH OF THE X VECTOR THAT IS
THE NUMBER OF OBSERVATIONS FOR X

```

```

N: THE LENGTH OF THE Y VECTOR THAT IS
THE NUMBER OF OBSERVATIONS FOR Y

```

```

*****
*****/

```

```

LIBNAME JGO 'C:\SASFILES\DATA';

```

```

%MACRO MILLER;

```

```

DATA MILLRDAT;

```

```

SET &DATA;

```

```

X=&XVAR;

```

```

Y=&YVAR;

```

```

PROC IML;

```

```

USE MILLRDAT;

```

```

READ ALL VAR {X Y};

```

```

/* COMPUTE NATURAL LOG OF SAMPLE
VARIANCE OF Xs */

```

```

XBAR=X[,.];

```

```

SUMSQX=X[,##];

```

```

VARX=(1/%EVAL(&M-1))*(SUMSQX-&M*XBAR**
2);

```

```

S=LOG(VARX);

```

```

/* COMPUTE NATURAL LOG OF SAMPLE
VARIANCE OF Ys */

```

```

YBAR=Y[,.];

```

```

SUMSQY=Y[,##];

```

```

VARY=(1/%EVAL(&N-1))*(SUMSQY-&N*YBAR**2
);
T=LOG(VARY);

```

```

/*CREATE THE M SUBSETS BY REMOVING AN
ELEMENT I, FROM X*/

```

```

DO I=1 TO &M;

```

```

IF I=1 THEN

```

```

X1=X[2:&M,];

```

```

IF I > 1 & I < &M THEN DO;

```

```

X11=X[1:i-1,];

```

```

X21=X[I+1:&M,];

```

```

X1=X11//X21;

```

```

END;

```

```

IF I=&M THEN

```

```

X1=X[1:&M-1,];

```

```

/* THE NEXT 4 LINES WILL COMPUTE THE  $S_i$  */

```

```

XBARI=X1[,.];

```

```

SUMSQXI=X1[,##];

```

```

VARXI=(1/%EVAL(&M-2))*(SUMSQXI-%EVAL(&M
-1)*XBARI**2);

```

```

SI=LOG(VARXI);

```

```

/* COMPUTE THE  $A_i$  */

```

```

AI=%EVAL(&M)*S-%EVAL(&M-1)*Si;

```

```

A=A//AI;

```

```

END;

```

```

/* COMPUTE  $\bar{A}$  AND  $V_1$  */

```

```

ABAR=A[,.];

```

```

ASUMSQ=A[,##];

```

```

V1=((1/%EVAL(&M-1))*(ASUMSQ-&M*ABAR**2))/
&M;

```

```

/* CREATE THE N SUBSETS BY REMOVING AN
ELEMENT J, FROM Y*/

```

```

DO J=1 TO &N;

```

```

IF J=1 THEN

```

```

YJ=Y[2:&N,];

```

```

IF J > 1 & J < &N THEN DO;

```

```

Y1J=Y[1:J-1,];

```

```

Y2J=Y[J+1:&N,];

```

```

YJ=Y1J//Y2J;

```

```

END;

```

```

IF J=&N THEN

```

```

YJ=Y[1:&N-1,];

```

STATISTICS

/* THE NEXT 4 LINES WILL COMPUTE THE T_j */

```
YBARJ=YJ[:,:];
SUMSQYJ=YJ[:,##];
VARYJ=(1/%EVAL(&N-2))*(SUMSQYJ-%EVAL(&N-1)*YBARJ**2);
TJ=LOG(VARYJ);
```

/* COMPUTE THE B_j */

```
BJ=%EVAL(&N)*T-%EVAL(&N-1)*Tj;
B=B/BJ;
END;
```

/* COMPUTE $\bar{B}$ AND V_2 */

```
BBAR=B[:,:];
BSUMSQ=B[:,##];
V2=((1/%EVAL(&N-1))*(BSUMSQ-&N*BBAR**2))/&N;
```

/* COMPUTE TEST STATISTIC */

```
Q=(BBAR-ABAR)/(SQRT(V1+V2));
```

/* COMPUTE PVALUES */

```
PVALUE2S=2*(1-PROBNORM(Q));
```

/* COMPUTE POINT ESTIMATES AND
CONFIDENCE INTERVALS */

```
LCI=EXP((BBAR-ABAR)-1.96*SQRT(V1+V2));
UCI=EXP((BBAR-ABAR)+1.96*SQRT(V1+V2));
JACKVARX=EXP(ABAR);
JACKVARY=EXP(BBAR);
LOGRATIO=BBAR-ABAR;
RATIO=EXP(LOGRATIO);
```

/* PRINT TEST STATISTICS, PVALUES... */

```
FILE PRINT;
PUT "JACKKNIFE ESTIMATE OF LOG(RATIO): "
LOGRATIO;
PUT "JACKKNIFE ESTIMATE OF RATIO: " RATIO;
PUT "JACKKNIFE ESTIMATE OF LOG(VARX),
ABAR: " ABAR;
PUT "JACKKNIFE ESTIMATE OF LOG(VARY),
BBAR: " BBAR;
PUT "JACKKNIFE ESTIMATE OF VARX : "
JACKVARX;
PUT "JACKKNIFE ESTIMATE OF VARY: "
JACKVARY;
```

```
PUT "V1, ESTIMATE OF VARIANCE OF ABAR:" " "
V1;
PUT "V2, ESTIMATE OF VARIANCE OF BBAR:" "
"V2;
PUT "TEST STATISTIC:" " " Q;
PUT "TWO SIDED PVALUE:" " " PVALUE2S;
PUT "95% CONFIDENCE INTERVAL:" " " (" LCI ",
" UCI");
PUT "SAMPLE SIZE FOR X: " &M;
PUT "SAMPLE SIZE FOR Y: " &N;
PUT "VARX, SAMPLE VARIANCE OF THE Xs: "
VARX;
PUT "VARY, SAMPLE VARIANCE OF THE Ys: "
VARY;
```

QUIT;

%MEND MILLER;