DOCUMENT RESUME

ED 395 956	TM 025 092
AUTHOR	Gillaspy, James Arthur, Jr.
TITLE	Evaluating Result Replicability: Better Alternatives to Significance Tests.
PUB DATE	27 Jan 96
NOTE	24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 27, 1996).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Discriminant Analysis; *Evaluation Methods; *Groups; Heuristics; *Research Methodology; *Statistical Significance; *Test Use
IDENTIFIERS	Bootstrap Methods; Cross Validation; *Descriptive Discriminant Analysis; Jackknifing Technique; *Research Replication

ABSTRACT

Three procedures for evaluating the replicability of descriptive discriminant analysis (DDA) results are discussed. The techniques include cross-validation, the jackknife, and the bootstrap. Discriminant analysis is a multivariate technique used when group membership or classification is the focus of the analysis. DDA is used to describe major differences among groups. Several procedures are available to researchers for evaluating the replicability of DDA results. The first is cross-validation, which involves randomly dividing the data into two roughly equal groups and deriving discriminant function coefficients for each. The discriminant function coefficients are switched and used to predict the group membership of the other sample. The bootstrap involves copying a data set over and over again into a megafile and then repetitively drawing different samples with different combinations of subjects. The jackknife differs in that different subsets or groups of subjects are repetitively dropped from the original data set. An example illustrates application of the jackknife. Appendix A presents the heuristic data set for the example, and Appendix B contains Statistical Package for the Social Sciences commands for the jackknife. (Contains 2 tables and 26 references.) (SLD)

Running Head: RESULT REPLICABILITY

Evaluating Result Replicability: Better Alternatives to Significance Tests

James Arthur Gillaspy, Jr.

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, January 27, 1996.

Abstract

The present paper provides an introduction to three procedures for evaluating the replicability or invariance of descriptive discriminant analysis (DDA) results. The techniques discussed include cross-validation, the jackknife (Tukey, 1958), and the bootstrap (Diaconis & Efron, 1987; Lunneborg, 1990). A heuristic example of the application of the jackknife to DDA is presented.

Evaluating Result Replicability: Better Alternatives to Significance Tests

Thompson (1994c) stated that "Science is the business of isolating relationships that (re)occur under stated conditions, so that knowledge is created and can be cumulated" (p. 6). In other words, knowledge is advanced only if research results replicate. Regrettably, the behavioral sciences are dominated by an analytic model, statistical significance testing, that does not address the issue of result replicability (Carver, 1978; Cohen, 1990, 1994). Indeed, though it is widely known that statistical significance is largely a product of sample size rather than result importance or generalizability, statistical significance is still often confused with replicability (Carver, 1978; Thompson, 1989).

Recently, Cohen (1990; 1994) and Thompson (1993, 1994a, 1994b) provided strong arguments that statistical significance testing does not evaluate result replicability, and thus does not tell us what we (as scientists) want to know about our findings. As Stevens (1986, p. 58) explained, a result which is sample specificity lacks "generalizability,...is of limited scientific value" regardless of statistical significance or practical significance. Thus, researchers should be cautious of placing too much confidence in findings until the replicability of the results is demonstrated. The most interesting results are those that regardless of statistical significance, are generalizable from a sample to a population (Fish, 1986).

Perhaps, however, the tide is beginning to turn. The latest APA style manual encouraged researchers to go beyond the use of statistical significance tests (APA, 1994, p. 18). American Psychologist has also periodically featured an article discouraging the use of statistical significance testing (Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1988).

In the quest for other more helpful analytic models. Thompson (1994) emphatically urged researchers that, "the replicability of results must be empirically investigated, either through actual replication of the study, or by using methods such as cross-validation, the jackknife, or the bootstrap" (p. 1). These three methods are all different statistical techniques known as invariance or replicability analyses. Basically, invariance analyses are a set of procedures designed to determine how stable statistical results are likely to be across different samples. These procedures typically involve performed the analysis on different subgroups of the original data set and comparing these different results. Invariance analyses are conducted after data has been analyzed with more traditional statistical analysis.

The present paper provides an introduction to three procedures for evaluating the replicability of descriptive discriminant analysis (DDA) results. The techniques discussed include cross-validation, the jackknife (Tukey, 1958), and the bootstrap (Diaconis & Efron, 1987; Lunneborg, 1990). A heuristic example of the application of the jackknife is presented. To aid in comprehension, the basic concepts and interpretation of DDA is also reviewed.

Discriminant Analysis

Discriminant analysis is a multivariate technique used when group membership or classification is the focus of the analysis. Discriminant analysis employs factor analytic methods to extract orthogonal synthetic factors or functions. These discriminant functions are extracted in such a way to maximize the differences between groups (Klecka, 1980; Stevens, 1988).

There are two distinct types of discriminant analysis depending on the purpose of the data analysis, Predictive Discriminant Analysis (PDA) and Descriptive Discriminant Analysis (DDA). As the name implies, PDA is used to make predictions about the classification of individuals in two or more groups. In the case of PDA, the researcher has data from individuals from known groups and uses the discriminant function to make predictions about other individuals whose group membership is not known. An example might be to classify a sample of psychotherapists' based on their theoretical orientation. The discriminant function used to make the classifications could then be used to predict the theoretical orientation of other samples of therapists.

Descriptive discriminant analysis, on the other hand, is used to describe the "major differences among...groups" (Stevens, 1972, p. 501). The reason to use DDA is to identify variables that characterize the differences between two or more groups (Afifi & Clark, 1984). In the therapists' theoretical orientation example, a DDA could be used to yield information concerning the variables that distinguish the therapists' models of psychotherapy. These variables might include attitudes about human nature, the change process, unconscious processes, and the therapeutic relationship.

Invariance Procedures

Several procedures for investigating the generalizability of DDA results are available to researchers. The first such approach is referred to as cross-validation. Cross-validation involves randomly dividing the data into two roughly equal groups, for example, Subsample I and Subsample II (Afifi & Clark, 1984). Discriminant function coefficients are derived for each of the subsamples and then switched so that the discriminant function coefficients for Subsample I are used to predict the group membership of the Subsample II and vice versa. If the group membership is accurately predicted, the researcher can place more confidence in the discriminant function.

An advantage of cross-validation is that it requires only a single sample of data. A potential shortcoming of cross-validation, however, is that dividing the original data further reduces sample size. This can be problematic when the sample is already small (Daniel, 1989).

The Bootstrap

A second procedure for evaluating result replicability is called the bootstrap (Diaconis & Efron, 1987; Lunneborg, 1990). The logic of the bootstrap analysis as explained by Thompson (1989) "involves copying a data set over and over again into a megafile and then repetitively drawing different samples with different combinations of subjects...to determine how sampling influences results." (p. 3) Thompson (1992) has developed a computer program to perform bootstrap following DDA called DISCSTRA.

Applied to DDA, the bootstrap empirically evaluates the replicability of the Function Coefficients, Structure Coefficients, and Group Centroids by analyzing many, many different configurations of subjects, including configurations in which a subject may be represented several times or not at all, through the process of resampling with replacement from a "mega-file."

The first step in conducting a bootstrap analysis of DDA is to create a mega-file of data by copying the data set hundreds or thousands of times. Next, hundreds or thousands of resamples (with replacement) are taken from the mega-file. Resamplings should be the same _n_ as original sample. A discriminant analysis is then run on each one of the resamples using Procrustean rotation to the actual function matrix of the original sample to ensure that the function coefficients are consistent across resamples.

One should note that DDA uses factor analytic techniques to extract the functions that maximize the differences between groups. Functions are orthogonal synthetic combinations of the original variables. Therefore, the problem with doing bootstrap is whether the same factors emerge in the same order across resamplings. DISCSTRA (Thompson, 1992) solves this problem by performing a Procrustean rotation of the function matrix for each resamplings into a position of best fit with the function matrix of the original data set. The means and standard errors (standard deviations) of the function coefficients, the structure coefficients, and the group centroids of all the resamplings are then computed and compared to make judgements about sample specificity. It is suggested that if the parameter to standard error ratio is 2 to 1, then one can have more confidence that the DDA results are reasonably replicable.

The Jackknife

The jackknife, developed by Tukey (1958), differs from the bootstrap in that different subsets or groups of subjects are repetitively dropped out of the original data set. The statistics of interest are calculated for each truncated data set, and the results are averaged. The term "jackknife" was conferred on this technique because, like a scout's jackknife, the procedure was considered "a rough-and-ready instrument capable of being utilized in all contingencies and emergencies" (Miller, 1964, p. 1594).

Applied to descriptive discriminant analysis, the jackknife empirically evaluates the Function Coefficients, as well as the Structure Coefficients and Group Centroids by analyzing the data with different individual subjects or subsets of individual subjects dropped out of the original data set.

Steps to Run Jackknife for Discriminant Analysis

(a) Run a Discriminate Analysis on the entire sample data yielding Function Coefficients, Structure Coefficients, and Group Centroids;

(b) Divide the original sample (*N*) into (*k*) subsets of equal size (n). Each subset can be as small as 1 and as large as the largest multiplicative factor of *N*;

(c) Delete (in repetition) each subset from the original sample and run DDA on each truncated data set. It should be noted that smaller subsets are preferable because they produce more repetitions, and thus make it easier to detect outliers and thus have more confidence in the results;

(d) Calculate "pseudovalues" from each truncated data set using the original function coefficients, the truncated function coefficients, and k;

(1) Pseudovalues = $J_i\Theta' = k(\Theta) - (k-1)\Theta$

where Θ' = Function coefficient on original data

k = number of subsets

Θ_i = Function coefficient of truncated data set

(e) Average the pseudovalues to produce the "Jackknifed Coefficients"

(2) Jackknifed Coefficient = $J(\Theta') = \sum J_i\Theta/k$

(f) Interpret the jackknifed coefficients.

There are several different methods for evaluating the results of the jackknife. However, Thompson (1984) cautioned that no guidelines for interpreting replicability results have been established, and that therefore the researchers must exercise their own judgment.

The t-statistic can be used to evaluate jackknife results (Tukey, 1958), because jackknifed coefficients have been found to be normally distributed. The following is the formula for calculating t with jackknifed coefficients:

(3) $\underline{t} = \dfrac{\text{Jackknifed Coefficient}}{\text{Standard error of the means of the pseudovalues}}$

where df = (k-1).

The t calculated is then compared to a critical value from a t table to determine statistical significance. Understandably, Fish (1986) argued that it was illogical to evaluate a replicability result with statistical significance testing, an analytic model that cannot address the issue of generalizability.

Two alternatives to statistical significance are the estimated parameter to standard error ratio and the use of confidence intervals. The parameter to standard error ratio involves comparing the ratio of the jackknifed coefficients with the standard error of the pseudovalues. A 2 to 1 ratio of parameter to standard error suggests result stability across samples.

Confidence intervals can also be construct around the jackknifed coefficients at a probability that is meaningful to the researcher. The formula for computing confidence levels for jackknifed coefficients is:

(4) Confidence Interval = $J(\Theta') + z \bullet SD$.

If the confidence interval contains the original function coefficient then you can expect your results to replicate.

An Example of the Jackknife Applied to DDA Results

For the present example, a DDA was conducted on a heuristic data set developed by Fish (1988) and Taylor (1991). This hypothetical data consisted of 64 cases, with 2 predictor variables (X and Y) and 4 subgroups. The cases were randomly assigned to 8 (*k*) groups of 8 subjects each for the purpose of conducting the jackknife. The entire data set and the SPSS commands for the discriminant analysis and the repetitions of DDA on each truncated data set (the method used to conduct the jackknife using SPSS) are included in Appendices A and B, respectively.

Table 1 presents the discriminant function coefficients yielded from the entire sample and the 8 truncated data subsets. Performing the discriminant analysis using the entire sample produced discriminant function coefficients for Function I of 1.55689 and -1.20080 and for Function II of .01167 and .99103 for Variables X and Y, respectively.

INSERT TABLE 1 ABOUT HERE

Using Equation 1, the discriminant function coefficients produced by each repetition were used to compute pseudovalues for each of the 8 truncated data sets. The jackknifed discriminant function coefficients were then calculated by averaging the 8 pseudovalues for each discriminant function. For Function I, the jackknifed coefficients were, for X and Y respectively, 1.52475 and -1.16834 and for Function II, .00944 and 1.01644. Finally, the original discriminant function coefficients were compared to the jackknifed coefficients with a *t*-test. The pseudovalues, the jackknifed coefficients, the standard error of the mean of the pseudovalues, and t calculated are reported in Table 2.

INSERT TABLE 2 ABOUT HERE

Discussion

As mentioned earlier there are several methods for interpreting the results of a jackknife analysis. By just "eyeballing" the jackknifed coefficients and the function coefficients from the entire data set, one can see that the values are very similar, suggesting result stability. This interpretation is generally supported by the ratios of the jackknifed coefficients to the standard error of the means, as well as confidence interval. All ratios are 2-to-1 or greater except for Variable X on Function II (.035). In addition, all the jackknifed coefficients fall within a 95% confidence interval, though Variable X on Function II approaches zero.

Taken together the jackknife analysis seem to indicate that the sample results are likely to replicate with future samples. However, the jackknifed coefficients for Variable X on Function II are very small and vary considerable (SD = .269) compared to the other variables. Thus, while the results will most likely replicate in future research, Variable X will have limited predictive utility.

Summary

In light of the inadequacy of statistical significance testing to address the real issue of science, result replicability, it should be obvious that alternative analytic methods are necessary. The present paper has introduced three procedures for evaluating the replicability of descriptive discriminant analysis (DDA) results. The techniques discussed included cross-validation, the bootstrap, and the jackknife. A heuristic example was used to demonstrate the basic concepts and the application of the jackknife to DDA.

References

Afifi, A. A., & Clark, V. (1984). Computer-aided multivariate analysis. Belmont, CA: Lifetime Learning Publications.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Carver, R. C. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.

Cooil, B., Winer, R. S., & Rados, D. L. (1987). Cross-validation for prediction. Journal of Marketing Research, 24, 271-279.

Crask, M. R., & Perreault, W. D. (1977). Validation of discriminant analysis in marketing research. Journal of Marketing Research, 14, 60-68.

Daniel, L. G. (1989, January). Use of the jackknife statistic to establish the external validity of discriminant analysis results. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston. (ERIC Document Reproduction Service No. ED 305 382)

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 248(5), 116-130.

Fish, L. (1986, November). The importance of invariance procedures as against tests of statistical significance. Paper presented at the annual meeting of the Southwest Educational Research Association, Memphis. (ERIC Document Reproduction Service No. ED 278 707)

Klecka, W. R. (1980). Discriminant analysis. Newbury Park, CA: Sage Publications.

Kupfermid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Miller, R. T. (1964). A trustworthy jackknife. The Annal of Mathematical Statistics, 35, 1594-1605.

Lunneborg, C. E. (1990). Review of Computer intensive methods for testing hypotheses. Educational and Psychological Measurement, 50, 441-455.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the DESD, and althernative indices. American Psychologist, 46, 1086-1087.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Stevens, J. P. (1972). Four methods of anlyzing between variation for the k-group MANOVA problem. Multivariate Behavioral Research, 7, 499-522.

Stevens, J. P. (1992). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Newbury Park, CA: Sage.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-6.

Thompson, B. (1992). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminat analysis functin and structure coefficients and group centroids. Educational and Psychological Measurement, 52, 905-911.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61(4), 361-377.

Thompson, B. (1994a). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B. (1994b). The pivoltal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62(2), 157-176.

Thompson, B. (1994c). The concept of statistical significance testing. *ERIC/AE Digest Measurement Update*, *4*(1), 5-6.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29*, 614.

Table 1
Standardized Discriminant Function Coefficients for the Jackknife Subsets

Group Deleted	Function I X	Function I Y	Function II X	Function II Y
none	1.55689	-1.20080	.01167	.99103
1	1.56334	-1.20857	.02623	.97970
2	1.54426	-1.16118	-.02421	1.01833
3	1.46279	-1.09874	.04494	.96672
4	1.71091	-1.50572	.19439	.83535
5	1.61345	-1.24890	-.02809	1.02189
6	1.52628	-1.05436	-.15854	1.11490
7	1.57848	-1.29292	.11018	.91218
8	1.49277	-1.06311	-.06899	1.05020

Table 2
Computed Pseudovalues, Jackknifed Coefficients, and t-values

Group Deleted	Function I X	Function I Y	Function II X	Function II Y
1	1.51174	-1.07641	-.09025	1.07034
2	1.64530	-1.47814	.26283	.79993
3	2.21559	-1.91522	-.22122	1.16120
4	.47875	.93364	-1.26737	2.08079
5	1.16097	-.86410	.28999	.77501
6	1.77116	-2.22588	1.20314	.12394
7	1.40576	-.55596	-.67790	1.54298
8	2.00576	-2.16463	.57629	.57684
Jackknifed Coefficient	1.52475	-1.16834	.00944	1.01644
SD	.19	.39	.269	.213
t_{calc}	8.025*	-3.158*	.035	4.772*
$T_{crit\ .05\ level}$	2.365	2.365	2.365	2.365

*indicates coefficient stability

Appendix A
Heuristic Data Set for Jackknife with DDA

01	1	4	2	5
02	1	5	3	8
03	1	4	4	2
04	1	4	5	3
05	1	3	4	4
06	1	6	5	6
07	1	5	6	7
08	1	7	5	2
09	1	6	6	1
10	1	8	6	8
11	1	7	6	1
12	1	9	7	5
13	1	8	7	4
14	1	8	8	3
15	1	9	8	7
16	1	9	9	6
17	2	1	2	8
18	2	3	3	4
19	2	3	5	3
20	2	3	5	6
21	2	2	5	5
22	2	4	6	4
23	2	4	5	2
24	2	5	6	5
25	2	6	6	6
26	2	6	6	1
27	2	6	7	7
28	2	7	7	8
29	2	7	7	2
30	2	8	9	3
31	2	8	9	7
32	2	9	9	1
33	3	4	1	8
34	3	4	2	6
35	3	3	2	3
36	3	2	4	5
37	3	5	3	2
38	3	7	4	1
39	3	4	5	7
40	3	5	4	5
41	3	7	5	8
42	3	9	5	6
43	3	6	5	4
44	3	5	6	1
45	3	7	6	7

46	3	9	7	3
47	3	8	6	5
48	3	8	5	2
49	4	1	7	4
50	4	1	2	3
51	4	1	1	2
52	4	2	2	8
53	4	2	3	3
54	4	2	3	1
55	4	3	2	7
56	4	3	3	4
57	4	3	4	7
58	4	4	5	6
59	4	4	4	5
60	4	4	5	4
61	4	4	6	2
62	4	5	6	1
63	4	5	7	8
64	4	5	7	6

Appendix B
SPSS Commands for Jackknife with Discriminant Analysis

```
LIST
  VARIABLES=id group x y subset
  /CASES= BY 1
  /FORMAT= WRAP UNNUMBERED .
DISCRIMINANT
  /GROUPS=group(1 4)
  /VARIABLES=x y
  /ANALYSIS ALL
  /PRIORS EQUAL
  /CLASSIFY=NONMISSING POOLED .
USE ALL.
COMPUTE filter_$=(subset ~= 1).
VARIABLE LABEL filter_$ 'subset ~= 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 2).
VARIABLE LABEL filter_$ 'subset ~= 2 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 3).
VARIABLE LABEL filter_$ 'subset ~= 3 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 4).
VARIABLE LABEL filter_$ 'subset ~= 4 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
```

```
USE ALL.
COMPUTE filter_$=(subset ~= 5).
VARIABLE LABEL filter_$ 'subset ~= 5 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 6).
VARIABLE LABEL filter_$ 'subset ~= 6 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 7).
VARIABLE LABEL filter_$ 'subset ~= 7 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
USE ALL.
COMPUTE filter_$=(subset ~= 8).
VARIABLE LABEL filter_$ 'subset ~= 8 (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
DISCRIMINANT
  /GROUPS=GROUP(1 4)
  /VARIABLES=X Y
  /STATISTICS .
```