RProbSup: CALCULATING THE PROBABILITY OF SUPERIORITY

Author:
Olivia A. Ortelli

Faculty Sponsor:
John Ruscio
Department of Psychology

ABSTRACT

Researchers are encouraged to report statistics in their studies, including measures of effect size and confidence intervals (CIs). The probability of superiority (*A*) has many appealing characteristics, such as being robust to parametric assumptions, easy to understand, and generalizable. While *A* was originally created to compare scores from two discrete groups, extensions of the statistic have been created to allow researchers to use *A* for correlated samples, as well as for more than two conditions. Former research on methods for creating confidence intervals for *A* suggests the use of bootstrap methods, which can be used for each variation of calculating *A*. The use of bootstrapping methods for calculating CIs and the various extensions of *A* have been brought together to create a package, *RProbSup*, which allows users to easily calculate different variations of *A*, its standard error (SE), and CI. This paper illustrates how to use *RProbSup* and highlights the benefits of using these statistical methods.

INTRODUCTION

The American Psychological Association advises researchers to report measures of effect size in their studies (APA, 2010) to indicate the magnitude of difference between groups. To choose an appropriate effect size, researchers must consider the scale, size, variability, and normality of their data, as well as the effect size's ability to clearly communicate its meaning to the reader (Vargha & Delaney, 2000). Many commonly used measures of effect size for between-group comparisons, including the standardized mean difference (*d*) and the point bi-serial correlation (r_{pb}), require parametric assumptions to be satisfied. In real world studies, it may be challenging or impossible to conform to these assumptions, such as if one's data is discrete, weighted, or distributed asymmetrically.

McGraw and Wong (1992) advocated for the use of the common language effect size estimator (*CL*) as the solution to finding the most effective, accurate, and easy to understand effect size estimator when comparing scores across two groups. Vargha and Delaney (2000) built on McGraw and Wong's (1992) foundation with recommendations to use the probability of superiority, *A*, instead. *A* is described as a nonparametric generalization of *CL* that is robust to the parametric assumptions and other factors.

Ruscio (2008) described the benefits and limitations of commonly used effect size measures including the standardized mean difference (*d*), point bi-serial correlation (r_{pb}), common language effect size indicator (*CL*), and measure of stochastic superiority, otherwise referred to as the probability of superiority (*A*). Each effect size serves different functions. For example, *d* compares two populations' means on the dependent variable while r_{pb} compares two populations' correlations on the dependent variable. *A*, a probability-based measure of effect size, is similar to *CL* in that they both measure the probability that a randomly chosen member of one group will perform better than a randomly chosen member of the other group. However, *A* accounts for ties by awarding 0.5 credit where *CL* does not (Vargha & Delaney, 2000). Ruscio (2008) promotes the use of the *A* statistic in the field of psychology due to its many merits including its ease of understanding, ability to be generalized, usefulness with ordinal data, and robustness to parametric assumptions, base rates, and outliers.

OVERVIEW OF THE PROBABILITY OF SUPERIORITY

The most commonly used effect size measures require some type of parametric assumptions to be met. Parametric assumptions include normally distributed data and homogeneity of variances, and these are not always approximated well with real-world data. The probability of superiority, otherwise known as the *A* statistic, is especially helpful due to its ability to be used without making parametric assumptions, allowing it to appropriately be used in many contexts and situations. Ruscio and Gera (2013) explain how discrete data would violate parametric assumptions, thus using *d* or r_{pb} would be inappropriate, whereas using *A* would be appropriate and effective. Therefore, while other measures of effect size require continuous data, *A* is applicable for both continuous and discrete data. *A* can also be useful when one has weighted data (Ruscio & Gera, 2013).

To calculate *A*, one can use the formula provided by Vargha and Delaney (2000): $A = P(X_1>X_2) + 0.5P(X_1=X_2)$, where X_1 and X_2 are vectors of scores for two groups and *P* represents calculating the probability, as per standard probability theory notation. The formula can be read as the probability that group one will score higher than group two, accommodating ties as half credit. For example, in a medical study between a treatment group and a control group, an *A* statistic of 0.823 would indicate the treatment group is 82.3% more likely to score higher on the health assessment than the control group. The bounds of the *A* statistic are 0 and 1. If there exists no difference between the scores of two groups, this is equivalent to saying *A* = 0.5, which Vargha and Delaney (2000) refer to as stochastic equality. Alternatively, if there is a 100% chance that one group would score higher than the other group, *A* = 1. Calculating *A* is equivalent to finding the area under the receiver operating characteristic (ROC) curve (Ruscio & Mullen, 2012). Therefore, when referring to the area under the ROC curve (AUC), one is still referring to the probability that a randomly selected member of one group scores higher than a randomly selected member of another group.

While *A* was originally introduced to measure the probability of superiority for two discrete groups, variants of the formula have been established to extend the use of the statistic to other research designs. One variation of *A* allows one to measure the probability of superiority for two correlated samples, rather than for two groups. Vargha and Delaney (2000) manipulated the original formula so to measure *A* for two correlated samples one can use the formula

$$A = [P(X_1>X_2) + 0.5P(X_1=X_2)]/n,$$

where X_1 and X_2 are vectors of scores for two measures, rather than two groups, and *n* is the total number of participants.

Vargha and Delaney (2000) introduced two additional variations of *A*: the average absolute deviation (*AAD)* and the average absolute pairwise deviation (*AAPD*). *AAD* calculates *A* for research designs with more than two groups (Vargha & Delaney, 2000). *AAPD* is another variation introduced by Vargha and Delaney (2000) to use when a research design has more than two groups, but the researchers want to analyze pairs of groups compared to all others, rather than each group independently. Ruscio and Gera (2013) discuss two additional variations of *A*: A_{ik} and A_{ord}. A_{ik} is used for singling out one group compared to all others while A_{ord} is used to determine the extent to which scores seem to be rank-ordered (Ruscio and Gera, 2013). *AAD*, *AAPD*, A_{ik}, and A_{ord} can all be applied for research designs utilizing groups or correlated samples. These four extensions of *A* allow it to be used for various research designs and in a multitude of fields.

The American Psychological Association (APA) refers to reporting confidence intervals (CIs) as "the best reporting strategy" due to CIs' abilities to reveal information about both location and precision (APA, 2010, p. 34). Bootstrapping allows researchers to generate their own empirical sampling distribution by randomly drawing from the given sample the specified number of cases, with replacement (Ruscio & Mullen, 2012). Ruscio and Mullen (2012) explain thoroughly the benefits of using bootstrap methods to construct a standard error (SE) or CI for *A* rather than analytical approaches. The existing analytical approaches to constructing CIs treat the sampling distributions as if they are symmetric in shape, which would only be appropriate when *A* is 0.50 (Ruscio & Mullen, 2012). Because bootstrapping methods generate their own sampling distributions, these methods can construct asymmetric CIs which can be more accurate. Also, because analytic CIs are usually constructed as a point estimate ± 1.96 times the estimated SE, they might extend into theoretically impossible values (e.g., values < 0 or > 1 for the *A* statistic). Bootstrap CIs cannot extend into impossible values. For other variants of the *A* statistic (*AAD*, *AAPD*, A_{IK}, and A_{ord}) there are no analytical approaches to constructing SEs and CIs. Thus, accurate bootstrapping methods are even more significant for reporting statistics because, as of our current knowledge, it is the only way to do so.

OVERVIEW OF THE RprobSup.R PACKAGE

To calculate *A*, its SE, and construct a CI for it, Ruscio (2012) created two sets of programs, *A.R* (2012) and *Bootstrap CI for A.R* (2012), which are described in Ruscio and Gera (2013) and Ruscio and Mullen (2012), respectively. Both sets of programs have been available, free of charge, on Ruscio's professional website (ruscio.pages.tcnj.edu). The purpose of this project is combining the code into an R package, *RProbSup*. A benefit of this project is that distributing code through an R package rather than a professional website is accessible in a more conventional manner, and still free of charge, through the Comprehensive R Archive Network (*CRAN*).

RProbSup STRUCTURE

RProbSup consists of twenty-one functions, only one of which should be called directly by the user. Users will provide their data through a matrix. For a between-subjects design, the matrix must consist of cases (rows) by scores (column 1) and group codes (column 2). For a within-subjects design, the matrix must consist of scores with each sample in its own column. Next, the user will specify whether one is calculating the *A* statistic for two or more groups (between-subjects design) or for two or more correlated samples (within-subjects design) by calling 1 or 2 for the argument "design," respectively. There exist five variations of the *A* statistic: the fast calculation, the average absolute deviation (A_{AAD}), the average absolute pairwise deviation (A_{AAPD}), and when users want to single-out a group (A_{ik}) or use ordinal data (A_{ord}). A user can specify which variation they are using in the argument "statistic" by selecting 1, 2, 3, 4, or 5, respectively. *RProbSup* displays the *A* statistic and, using bootstrap methods, its estimated SE and a CI. The default bootstrap method for each function is the bootstrap bias-corrected and accelerated (BCA) method, represented as 1 in the argument "ci.method"; however, users can also specify to use the bootstrap percentile (BP) method, by calling 2 for the argument "ci.method."

CALCULATING THE A STATISTIC, STANDARD ERROR, AND CONFIDENCE INTERVALS USING *'RProbSup'*

FAST CALCULATION

The A Statistic (A). A1() and A2() calculate *A* and its SE and construct a CI for the *A* statistic for two groups and two correlated samples, respectively. A user should not call A1() nor A2() directly, but instead use A() and specify which variation is to be used in the arguments "design" and "statistic." To calculate the *A* statistic, SE, and CI for two groups, one must first create a matrix consisting of the scores for groups (or samples, if using a between-subjects design) one and two to two columns; call this matrix data. The first example indicates the calculation of the *A* statistic, SE, and CI for two groups by directing the function A()to call A1() by calling 1 for the argument "design." The second example demonstrates the calculation of the *A* statistic, SE, and CI for two correlated samples by directing the function A()to call A2() by calling 2 for the argument "design." In both cases, the user must call 1, representing the fast calculation, for the argument "statistic." In the first example, y1 and y2 represent the scores for groups one and two, whereas, in the second example y1 and y2 represent the scores for samples one and two. For users who may be new to using *R*, note cbind() combines data by columns, c() combines the arguments, and rep() replicates the first argument the number of times specified in the second argument. Understanding these basic *R* functions will be useful in understanding how A() functions.

```
> y1 <- c(6, 7, 8, 7, 9, 6, 5, 4, 7, 8, 7, 6, 9, 5, 4)
> y2 <- c(7, 5, 6, 7, 6, 4, 3, 5, 4, 5, 4, 5, 7, 4, 5)
> data <- cbind(c(y1, y2), c(rep(1, length(y1)), rep(2, length(y2))))
> A(data, 1, 1)
   A:  0.747
  SE:  0.085
```

```
95% CI:  0.551 to 0.889
```

An *A* value of 0.747 for two or more group using the fast calculation indicates that on average, there was a 74.7% chance that a randomly selected score from the first group was higher than a randomly selected score from the second group.

```
> data <- cbind(y1, y2)
> A(data, 2, 1)
     A:  0.767
    SE:  0.101
95% CI:  0.5 to 0.9
```

An *A* value of 0.767 for two or more correlated samples using the fast calculation indicates that on average, there was a 76.7% chance that a randomly chosen participant's score for the first measure was higher than their score for the second measure. Notice that with the same data, the fast calculation for two groups compared to two correlated samples results in a different *A* value, SE, and CI. The difference in these values can be explained by how *A* is calculated differently for correlated samples; *A* is calculated by comparing scores of two measures and dividing by the total number of participants. Because bootstrapping is used to create the SE and CI, a different approach to calculating *A* results in the creation of a new sampling distribution, thus changing the SE and CI.

VARIATIONS OF A

Average absolute deviation (AAD). One generalization of *A* introduced by Vargha and Delaney (2000) is the average absolute deviation (*AAD*). One can study the extent of the differences among more than two groups by determining whether the scores in one group are different that the union of the scores of every other group by using *AAD*. An *A* statistic is calculated for every group compared to all others, and the series of the calculated *A* statistic helps calculate the *AAD*, which represents the estimation of stochastic homogeneity for the sample (Vargha & Delaney, 2000). To calculate the average absolute deviation, one can use the formula:

$$\frac{\sum_{i=1}^{k} |A_{ik} - .50|}{k} + 0.50,$$

where the number of groups is represented by *k*, the comparison group is represented by *i*, and each *A* statistic calculated for each *i* group compared to union of all others is represented as A_{ik} (Ruscio & Gera, 2013). To further understand the derivation of the formula for *AAD*, see Vargha and Delaney (2000). To calculate the *AAD* for two or more groups or two or more correlated samples, users will call A(data, 1, 2) or A(data, 2, 2), respectively, in *RProbSup*. Users will need a matrix of cases by scores and group codes or a matrix of scores with each sample in its own column, respectively.

The following examples will demonstrate how to use *RProbSup* to calculate the *AAD* for two or more groups or samples and can be used as a guideline for the subsequent variations of *RProbSup*, as well. The final basic function users will want to become familiar with is rnorm(), which generates the specified number of random values from the normal distribution (with default mean of 0 and standard deviation of 1).

```
> set.seed(1)
> x1 <- rnorm(25)
> x2 <- x1 - rnorm(25, mean = 1)
> x3 <- x2 - rnorm(25, mean = 1)
> x.bs <- cbind(c(x1, x2, x3), c(rep(1, 25), rep(2, 25), rep(3, 25)))
> A(x.bs, 1, 2)
     A:  0.721
    SE:  0.029
95% CI:  0.621 to 0.759
```

An *A* value of 0.721 for the *AAD* for two or more groups indicates that on average, there was a 72.1% chance that a randomly selected score in one group was higher than a randomly selected score from the union of all other groups in the study.

```
> x.ws <- cbind(x1, x2, x3)
> A(x.ws, 2, 2)
   A:  0.793
  SE:  0.023
95% CI:  0.713 to 0.82
```

An *A* value of 0.793 for the *AAD* for two or more correlated samples indicates that on average, there was a 79.3% chance that a randomly selected participant's score for one measure was higher than a randomly selected score from the union of all other measures for that participant.

Average absolute pairwise deviation (AAPD). Vargha and Delaney (2000) also introduced the average absolute pairwise deviation (*AAPD*). In this generalization, the *A* statistic is calculated for each pair of groups, rather than each group individually as in *AAD*. A similar process of *AAD* follows in that the *A* statistics gathered from each pair of groups is aggregated into a series which represents *AAPD*, representing pairwise stochastic homogeneity (Vargha & Delaney, 2000). To calculate the average absolute pairwise deviation, one can use the formula:

$$\frac{\sum_{i=1}^{k-1}\ \sum_{j=i+1}^{k}|A_{ij}-.50|}{\frac{k(k-1)}{2}}+0.50,$$

where each *A* statistic calculated for each *i* group compared each *j* group is represented as A_{ij} (Ruscio & Gera, 2013). To further understand the derivation of the formula for *AAPD*, see Vargha and Delaney (2000). To calculate the *AAPD* in *RProbSup*, users need a matrix of cases by scores and group codes.

```
> A(x.bs, 1, 3)
   A:  0.807
  SE:  0.041
95% CI:  0.707 to 0.873
```

An *A* value of 0.807 for the *AAPD* for two or more groups indicates that on average, there was an 80.7% chance that a score selected from a randomly chosen pair of scores was higher than a score from the union of all other pairs of scores.

```
> A(x.ws, 2, 3)
   A:  0.907
  SE:  0.034
95% CI:  0.787 to 0.947
```

An *A* value of 0.907 for the *AAPD* for two or more correlated samples indicates that on average, there was a 90.7% chance that a randomly selected pair of scores for a participant was higher than a score from the union of all other pair of scores across the different measures for the same participant.

IK. Ruscio and Gera (2013) discuss two additional variants of *A*. One variant, A_{ik}, is applicable when a user wants to single out one group and compare it to all others, differing from *AAD* as the union of the groups is treated as a pool of one sample (Ruscio & Gera, 2013). To calculate this variant of *A*, one can use the formula, $Pr(Y_i>Y_{\sim i})$, where *~i* represents the newly considered sample pool of each group without the comparison group, *i* (Ruscio & Gera, 2013). To calculate A_{ik} in *RProbSup*, users need a matrix of cases by scores and group codes as well as specify the number of reference groups to compare to all others (default is 1).

```
> A(x.bs, 1, 4)
   A:  0.832
  SE:  0.049
95% CI:  0.704 to 0.906
```

An *A* value of 0.832 while using the variation A_{ik} for two or more groups indicates that on average, there was an 83.2% chance that a score from one group was higher than a randomly selected score from any other group.

```
> A(x.ws, 2, 4)
```

```
    A:  0.94
   SE:  0.032
95% CI:  0.8 to 0.98
```

An *A* value of 0.94 while using the variation A_{ik} for two or more correlated samples indicates that on average, there was a 94% chance that a randomly selected participant's score for one measure was higher than a score from any other measure.

Ord. The final variation of the probability of superiority that users can calculate in *RProbSup* is useful in determining the "extent to which scores tend to be rank-ordered among groups" (Ruscio & Gera, 2013, p. 215). The *A* statistic is calculated for each pair of adjacent scores for all scores then the summation of the various *A* statistics is used in calculating the ordinal comparison among the groups (Ruscio & Gera, 2013). To calculate the ordinal variation of *A* one can use the following formula:

$[Pr(Y_1>Y_2)+Pr(Y_2>Y_3)+\ldots+Pr(Y_{k-1}>Y_k)]/(k-1)$.

To calculate A_{ord} in *RProbSup*, users need a matrix of cases by scores and group codes.

```
> A(x.bs, 1, 5)
    A:  0.762
   SE:  0.041
95% CI:  0.67 to 0.834
```

An *A* value of 0.762 while using the variation A_{ord} for two or more groups indicates that on average, across all sequential comparisons, there was an 76.2% chance that a score in one group was higher than the score in the subsequent group.

```
> A(x.ws, 2, 5)
    A:  0.88
   SE:  0.041
95% CI:  0.76 to 0.94
```

An *A* value of 0.88 while using the variation A_{ord} for two or more correlated samples indicates that on average, across all sequential comparisons, there was an 88% chance that a randomly selected participant's score in one measure was higher than the score for the subsequent measure.

CREATING CONFIDENCE INTERVALS USING BOOTSTRAP METHODS

Ruscio and Mullen (2012) recommend using the bootstrap percentile (BP) method, as well as the bootstrap bias-corrected and accelerated (BCA) method. Both the BP and BCA methods are empirical methods for constructing CIs. Rather than estimating the parameters, a sample is treated as an unbiased estimate of the population (Ruscio & Mullen, 2012). A large number of bootstrap samples are created by drawing new samples of equal size, with replacement, from the original data. Next, one performs the desired analysis on each bootstrap sample. The process is repeated many times, providing a large number of values to be compiled into an empirical sampling distribution. In particular, the BP method identifies values at the 2.5th percentile as well as the 97.5th percentile of the newly create empirical sampling distribution to construct the limits of a 95% CI (Ruscio & Mullen, 2012). The BCA method is similar to the BP method. However, the BCA method accounts for skewness in the distribution and adjusts these limits accordingly (Ruscio & Mullen, 2012).

When compared to other empirical and analytical methods, the BCA method was found to be superior due to its robustness to various population characteristics such as distribution shape and unequal variances (Ruscio & Mullen, 2012). The BCA method constructed the highest mean percentage of coverage for the CI, resembling the 95% CI that many statisticians are familiar with (Ruscio & Mullen, 2012). That is, the CIs produced by the BCA method contained the real *A* value for the population 94.4% of the time, indicating a very accurate CI. Because the BCA method is the most robust method and most accurately constructs a CI for the *A* statistic, Ruscio and Mullen (2012) recommend using the BCA method to construct CIs and calculate SE for the *A* statistic.

Better to understand the difference between the BCA and BP methods, let us consider the following sample data from Ruscio and Mullen (2012):

y1 = (6, 7, 8, 7, 9, 6, 5, 4, 7, 8, 7, 6, 9, 5, 4)
y2 = (4, 3, 5, 3, 6, 2, 2, 1, 6, 7, 4, 3, 2, 4, 3)

Let y1 represent the scores for group one and let y2 represent the scores for group two. When one specifies for the program to use the percentile (BP) method (represented by 2), *RProbSup* will display the following results:

```
> data <- cbind(c(y1, y2), c(rep(1, length(y1)), rep(2, length(y2))))
> A(data, 1, 1, ci.method = 2)
    A:  0.884
   SE:  0.058
95% CI:  0.756 to 0.976
Constructed using percentile method with B = 1999 bootstrap samples
```

When calculating the *A* statistic for the two groups using the default method for constructing CIs, the BCA method (represented by 1), *RProbSup* will display the following results:

```
> A(data, 1, 1)
    A:  0.884
   SE:  0.058
95% CI:  0.718 to 0.964
Constructed using BCA method with B = 1999 bootstrap samples
```

Notice how using the BCA method results in a different CI with adjusted endpoints than the CI calculated when using the BP method. The difference between the CIs can be explained by the fact that the BCA method accounts for the skewness in the data (Ruscio & Mullen, 2012). Notice, specifying which bootstrapping method to use adjusts the CI and does not affect the calculation of *A* nor the SE. Ruscio and Mullen (2012) concluded in their simulation study that the BCA method performed the best when compared to the BP method and other empirical and analytical methods for constructing CIs for *A*.

CONCLUDING REMARKS AND FUTURE DIRECTIONS

The American Psychological Association requires researchers to report measures of effect size to accompany their statistical tests, including SEs and CIs (APA, 2010). It is common practice for researchers to use Cohen's *d* or the r_{pb} to compare scores for two groups even when parametric assumptions may be violated. Instead, the probability of superiority may be a more appropriate measure.

McGraw and Wong (1992) highlighted the benefits of using the common language effect size estimator (*CL*) due to its ease of generalizing its results and explaining its significance to those not trained in statistics. Vargha and Delaney (2000) introduced a measure of stochastic superiority (*A*). Ruscio (2008) reviewed the benefits of using the probability of superiority, Ruscio and Gera (2013) highlight the various uses of the *A* statistic in different contexts, and Ruscio and Mullen (2012) show how to estimate its *SE* and construct *CI*s for it.

The programs created to accompany the work of Ruscio and Gera (2012) and Ruscio and Mullen (2012) have been condensed to be represented in one single *R* package, titled *RProbSup* (2018) (available at https://CRAN.R-project.org/package=RProbSup). Users can specify which variation of *A* the user would like the *A* statistic calculated for, as well as be given the SE and CI for this statistic using bootstrapping methods. It is our hope that investigators can now easily and effectively take advantage of the code in a user-friendly manner.

REFERENCES

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361-365. doi: 10.1037/0033-2909.111.2.361

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19-30. doi: 10.1037/1082-989X.13.1.19

Ruscio, J. (2018). RProbSup: Calculates Probability of Superiority. R package version 1.0. https://CRAN.R-project.org/package=RProbSup

Ruscio, J., & Gera, B. L. (2013). Generalizations and extensions of the probability of superiority effect size estimator. *Multivariate Behavioral Research, 48*(2), 208-219. doi: 10.1080/00273171.2012.738184

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201-223. doi: 10.1080/00273171.2012.658329

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the *CL* common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132. doi: 10.3102/10769986025002101