A Better Tool for the Cognitive Scientist's Toolbox: Randomization Statistics

Michael D. Byrne[1]
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332-0170
byrne@cc.gatech.edu

Abstract

Cognitive Science has typically proceeded with two major forms of research: model-building and experimentation. Traditional parametric statistics are normally used in the analysis of experiments, yet the assumptions required for parametric tests are almost never met in Cognitive Science. The purpose of this paper is twofold: to present a viable alternative to traditional parametric statistics—the randomization test—and to demonstrate that this method of statistical testing is particularly suited to research in Cognitive Science.

Introduction

One of the oldest methods of investigating the phenomenology of human cognition is the experiment, usually conducted in the laboratory. Experiments, however, are only useful to the extent that they can demonstrate reliable results.

The reliability of experimental results has, for some time, been assessed through the use of inferential statistics. While this is conceptually a sound process, the actual methods usually employed to do this have recently been subjected to increasing scrutiny and suspicion. Traditional statistical techniques for the analysis of experiments rely on *parametric* tests of statistical reliability. These tests make assumptions about the underlying form of the data and the method used to collect the data.

Probably the most common statistical method used to analyze experiments is the analysis of variance, or ANOVA. In the case of two-sample tests, a simpler equivalent, the t-test, is normally employed. The ANOVA F-test (taken as the typical parametric test) makes a number of assumptions which are *rarely* met in Cognitive Science research. Following is a list of the assumptions and the nature of the violation typically encountered.

Assumptions of Parametric Tests

Random Sampling

One of the most fundamental assumptions made in parametric inferential statistics is that of random sampling. The hypotheses in parametric tests concern population parameters (usually means), where estimators of those parameters are found by *randomly sampling* from a population. In essence, a t-test tests a hypothesis like $\mu_1 = \mu_2$. The terms μ_1 and μ_2 only have meaning in the context of random sampling from some population. Indeed, the mathematics underlying the t-test is based on the estimation of a "standard error of the mean," which refers to the standard deviation of the theoretical *sampling distribution* of the mean. If random sampling is not employed, references to this distribution make little sense.

In Cognitive Science, as in almost all experimental research, random sampling is not only not done, but is almost totally impractical. Experimenters do not generate exhaustive lists of their populations and generate random numbers to select people—samples of subjects are almost always convenience samples, such as "those subjects who sign up for the experiment." Random sampling is occasionally carried out by survey researchers, but rarely by experimenters.

Normally Distributed Data

Parametric methods assume that the distribution from which the random sample is drawn has a distribution that

[1]This work was supported by a graduate fellowship from the National Science Foundation.

is, or closely approximates, the "normal" bell-shaped, symmetric curve. The null probability distributions to which parametric test statistics are compared are mathematically derived from the normal probability distribution.

Fortunately, sampling from a population, even if that population does not conform to the normal distribution, yields normal distributions for the parameter estimates as the number of subjects grows larger. Thus, large random samples insure that even if the underlying distribution in the population is not normal, the sampling distributions of the parameters will be. Unfortunately, most Cognitive Science experiments fail to meet this criteria on several levels.

First, data in Cognitive Science are typically not distributed normally. Reaction times and error rates, for example, are almost always skewed distributions because they by definition cannot have large left-hand tails (negative scores are not possible, while there is no upper bound on positive scores). Real data sets are also prone to outliers, to which parametric tests are not particularly robust.

This alone does not constitute a fundamental problem if the sample size is large and is based on random sampling. While samples of 200 or so are not uncommon in some branches of behavioral research, samples of as many as 50 are large for Cognitive Science.

Homogeneity of Variance

Parametric tests on means (such as the ANOVA) assume equality among the variances of the groups from which the samples being compared are drawn. That is, if three groups are being compared, then the variances around all three means must be equal in order for the statistical test to be valid. In normal distributions, the mean and the variance of a distribution are independent, but this is not true in other distributions. In most real data, for example, groups with higher means tend to have higher variances as well. Cognitive Science data is no exception.

In repeated-measures or within-subjects experiments, a stronger form of this assumption, called sphericity, is required. While the details of the sphericity assumption are complex, the basic concept is that the repeated-measures ANOVA makes critical assumptions about the nature of the data distribution. Even relatively minor violations of sphericity can have a serious impact on the validity of the ANOVA. Unfortunately, it is often difficult to accurately determine if the sphericity assumption is met (Hays, 1988), so *any* within-subjects experiment analyzed with an ANOVA (which are not uncommon in Cognitive Science) is a possible cause for concern.

Random Assignment

One other assumption that is almost universal in experimental work is that of random assignment. That is, experimental units, typically subjects, have equal probability of being assigned to each level of the independent variable(s). The classic example of random assignment is in a two-group design wherein the experimenter flips a coin for each subject to determine the group into which that subject is placed.

Random assignment is not so much a statistical assumption as it is a common and necessary practice to insure the internal validity of an experiment (Campbell & Stanley, 1963). That is, random assignment is generally necessary in order to help insure that any differences that are observed can be attributed to the experimental manipulation and not to subject differences. Meeting this assumption is not difficult, and is the norm in laboratory work, including that in Cognitive Science.

Since it is nearly impossible to guarantee that experiments in Cognitive Science will meet any of the assumptions normally associated with statistical testing and experimentation save for random assignment, what is clearly necessary is a method for performing statistical tests that makes no assumptions about the data other than random assignment. Fortunately, such a method exists, though it is only recently that it has become practical. This method is called the "randomization" approach to statistics.

Randomization Tests

While the practical development of randomization tests is a relatively recent phenomenon, the basic principles were developed almost 60 years ago (Fisher, 1935). The basic tenet underlying randomization testing is simply this: if, under the null hypothesis of no effect of grouping, random assignment was used, every possible arrangement of the data is equally likely. Thus, much like the simple binomial test, it is possible to empirically generate a null distribution without making any further assumptions about the data. An example will help illustrate.

Taking a paradigm that should be familiar to all Cognitive Science readers, consider an experiment comparing two isomorphs of the Tower of Hanoi in which the dependent measure is the number of minutes the subjects take to solve the problem they are given. Group A receives the standard TOH, while Group B receives a more difficult isomorph, such as the "Monster Change" isomorph. The results of running nine subjects are as follows:

Group A: 12, 7, 4, 3
Group B: 8, 10, 12, 15, 22

The difference between means is 6.9.

If subjects were randomly assigned, there are

$$\binom{9}{4} = \frac{9!}{4!\,5!} = 126$$

possible unique arrangements of the data, and, if the null hypothesis of no association between the groups is true, all of these arrangements are equally likely. Yet some of these arrangements are clearly more "extreme" in some sense than others. Consider

Group A: 3, 4, 7, 8
Group B: 10, 12, 15, 22

and

Group A: 3, 22, 8, 10
Group B: 4, 7, 12, 15, 12.

The difference between the means in the first case is 6.3 minutes, while in the second it is a mere 0.75. Yet under the null hypothesis, these are equally probably outcomes. Given these nine scores, what is the probability that they would *randomly* fall into this most extreme arrangement? It is 1/126, which is approximately 0.008. Moving up a level of abstraction, what is the probability of observing a result that is as extreme or more than the result that was actually obtained? That is, What is the probability that the results observed could have been observed by chance if the null hypothesis is true?

As with the binomial test, one simply figures out the probability of each outcome that is as extreme or more than the obtained data. Doing this requires permuting the data and obtaining some index of difference (in this case, the difference between the means) for every unique permutation of the data so that those permutations that are as or more extreme than the observed date can be identified. The number of permutations that meet the extremity criteria is then divided by the number of possible permutations, directly yielding a probability value. This can then be compared to the nominal alpha level (conventionally, .05), and, if less that this value, one can conclude that the data are not independent of the grouping—that is, that there was an effect of the manipulation.

In the example presented, assuming a one-tailed test, there are five permutations (including the one observed) that meet the criteria. Thus, the p-value for this experiment would be 5/126, or about .04. In this case, this is almost exactly the same p-value that one would obtain with a traditional t-test. Parametric tests and randomization tests do not always agree, however. For example, if the most extreme observation in the data (22) is changed *in the direction of greater difference*, as little as five minutes (to 27), the t-test is no longer significant! On the other hand, the randomization test is not affected at all by this change.

How is this possible? A moment's reflection should make this clear. Under randomization, the more extreme data point will change the *size* of the differences observed in each permutation, but will not change the ordering. In the case of the t-test, however, even though the difference in means used to compute the t statistic is larger (7.9 as opposed to 6.9), the estimate of the standard error of the mean (the denominator of the t statistic) is inflated as well (3.29 vs. 4.20), and the p-value of this test is just slightly higher than .05, and would lead the researcher to fail to reject the null hypothesis. Parametric tests are sensitive to extreme data points, particularly when sample sizes are small.

Evaluating Randomization Tests

There are several relevant questions beyond simple appropriateness—there are practical considerations as well. The remainder of this paper will address several issues relevant to the use of randomization tests in behavioral research.

Flexibility

While the logic underlying the randomization test is quite straightforward for two independent samples, it is not immediately obvious that such a method generalizes to more complex designs. While the logic is somewhat more complex, the same basic techniques can be applied to arbitrarily complex factorial designs, repeated-measures designs, correlations, trend analysis and so on (Edgington, 1987). There is still some disagreement on how to treat interactions in complex designs, but a seemingly sound method has been developed by ter Braak (1992).

More complex multivariate statistical methods, such as factor analysis, path analysis, and the like, do not yet have randomization counterparts. However, the use of such statistical techniques is rare in Cognitive Science.

Acceptability

Despite the apparent sensibility of randomization testing, one important consideration is the general acceptance level associated with such a technique. Could one, for instance, publish a paper having used such a technique?

While randomization tests are not yet a common practice in behavioral research, they are discussed in such conservative statistical references as Siegel & Castellan (1988). Randomization tests are beginning to appear in introductory statistics texts (May, Masson, & Hunter, 1990) and have been discussed seriously in statistical

psychology journals for over a decade (e.g. Still & White, 1981).

Agreement

Certainly, one consideration is whether or not the adoption of a new statistical technique will greatly change the expectations of the researcher in terms of things like statistical power. That is, will these tests generally behave as well or better than the tests currently in use?

The answer to this question is a qualified yes. When the distributional assumptions of the ANOVA are met, the randomization test and the ANOVA generally agree with one another (e.g. Bradbury, 1987). This leads to the rationale used by some proponents of parametric tests—why use less standard randomization tests when the results generally agree?

The critical issue is that the results of the two procedures do not always agree, as demonstrated in the example above. It is generally difficult to predict exactly what the behavior of both parametric and randomization methods will be with different kinds of distributions. Violation of distributional assumptions tends to result in less power with parametric statistics, particularly with smaller sample sizes. Randomization tests appear to be more robust to such violations. (For an excellent brief review of parametric vs. nonparametric methods, see Hunter & May, 1993).

Pedagogy

Another important issue that arises in the use of any tool is the ease with which it can be learned/taught. It should be relatively clear that it is possible to learn the basic concepts of randomization testing rapidly, as the example used earlier should illustrate. It may, in fact, be easier to learn randomization tests, as one does not have to first master concepts like sampling distributions, variance pooling, and the like, which are typically prerequisites to understanding even simple t-tests. Some instructors (e.g. Peterson, 1991) maintain that instruction in randomization concepts focuses students' attention on issues of statistical inference and away from the more mundane memorizing formulae and such.

Availability

The case for using randomization tests, particularly in the kind of experiments typically done by Cognitive Scientists, is a strong one, perhaps "too good to be true." If randomization tests are the best thing to do, why isn't everyone already using them?

There are several answers to this question. First, the full development of randomization techniques (especially for analyzing interactions) is a relatively recent phenomenon, of which too few practitioners are aware. People simply do not know what the tests are or how they are available.

Second, most researchers use standard statistical packages such as SAS, SPSS, BMDP, SYSTAT, etc. These packages have yet to incorporate randomization tests, so performing such tests requires the use of some other program, or alternately, programming the tests by hand. There are statistics programs that include randomization tests, such as NPSTAT (May, Masson, & Hunter, 1989), StatXact (Mehta & Patel, 1991), RT (Manly, 1991), and CANOCO (ter Braak, 1988).

Programming randomization tests by hand is actually not all that difficult, and some texts even include code to make this easier. Siegel & Castellan (1988) includes code for the two-sample case, and Edgington (1987) includes extensive amounts of code for a variety of randomization tests.

One consideration when doing randomization tests is that of computational power. The number of permutations grows explosively as the number of data points increases. For example, there are over four million permutations in a design with three groups and nine cases per group. Increasing the design to ten cases per group raises this number to over 30 million. Obviously, computing all the possible permutations would be impractical in such cases. However, randomly sampling from the space of possible permutations yields an approximate test that is still valid, but for power considerations the largest possible sample should be used (see Edgington, 1987, for a complete explanation for this method of approximation). 10,000 permutations is typically considered more than adequate, but this is still time-consuming.

Computational considerations provide another reason why randomization tests are particularly appropriate to research, or rather, researchers in Cognitive Science. Many Cognitive Science researchers are both competent programmers and have access to high-speed workstations, which typically provide excellent floating-point performance.

Conclusions

Since most empirical research in Cognitive Science is based on experimentation, and most experiments violate one or more of the assumptions of traditional parametric statistical tests, the Cognitive Science community should be sensitive to issues of statistical methodology.

Randomization tests provide a viable, practical alternative to parametric tests, and thus it is recommended that Cognitive Science research adopt, or at least carefully consider, the use of randomization tests.

Acknowledgments

I would like to thank Richard May for pointing me to many of these references, Vicky Coon and Ángel Cabrera for proofreading, and Chris Hertzog for both commentary on an earlier draft and getting me started on this topic.

References

Bradbury, I. 1987. Analysis of variance versus randomization tests—a comparison. *British Journal of Mathematical and Statistical Psychology, 40,* 177-187.

Campbell, D. T., & Stanley, J. C. 1963. *Experimental and quasi-experimental designs for research.* Boston: Houghton Mifflin.

Edgington, E. S. 1987. *Randomization tests* (2nd ed.). New York: Marcell Dekker.

Fisher, R. A. 1935. *The design of experiments.* Edinburgh: Oliver & Boyd.

Hays, W. L. 1988. *Statistics* (4th ed.). New York: Holt, Rinehart, & Winston.

Hunter, M. A., & May, R. B. 1993. Some myths concerning parametric and nonparametric tests. *Canadian Psychology, 34(4).* Forthcoming.

Manly, B. F. J. 1991. *RT: A program for randomization testing, version 1.01.* University of Otago, New Zeland.

May, R. B., Masson, M. E. J., & Hunter, M. A. 1989. Randomization tests: Viable alternatives to normal curve tests. *Behavior Research Methods, Instruments, & Computers, 21,* 482-483.

May, R. B., Masson, M. E. J., & Hunter, M. A. 1990. *Application of statistics in behavioral research.* New York: Harper & Row.

Mehta, C., & Patel, N. 1991. *StatXact, version 2.0.* Cambridge, MA: Cytel Corporation.

Peterson, I. 1991. Pick a sample. *Science News, 140,* 56-58.

Siegel, S. & Castellan, N. J. 1988. *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Still, A. W., & White, A. P. 1981. The approximate randomization test as an alternative to the *F* test in analysis of variance. *British Journal of Mathematical and Statistical Psychology, 34,* 243-252.

ter Braak, C. J. F. 1988. *CANOCO—FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis, version 2.1.* Report LWA-88888-02. Wageningen: Agricultural Mathematics Group.

ter Braak, C. J. F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In K-H. Jöckel, G. Rothe, & W. Sendler (Eds.), *Bootstrapping and related resampling techniques* (pp. 79-86). Berlin: Springer Verlag.