Statistical Power and Effect Size in Information Retrieval Experiments

Michael J. Nelson
Faculty of Information and Media Studies
University of Western Ontario
mnelson@julian.uwo.ca

Statistical tests are used in information retrieval to test various hypotheses such as which indexing method is better or which retrieval system is better. Sometimes when using these statistical tests there is not enough evidence to reject the null hypothesis. Then either we have correctly discovered a true null hypothesis or made a type II error (probability denoted by b) and falsely accepted a null hypothesis. The power of a statistical test, denoted by 1-b, is the probability of rejecting a false null hypothesis as we should. The main difficulty is that this procedure allows us to control the significance level directly but the power may be not be very large and the power is not often calculated by standard statistical packages. One of the reasons this is difficult is that if we have a composite hypothesis there is actually a power curve which depends on the actual value of the population parameter being estimated. One of the parameters which is very important for making judgements and recommendations for design is the effect size. This is a dimensionless parameter based on the difference of means divided by the variance (at least for a test based on the difference on two means). The calculation of effect size for different experimental designs is reviewed and applied to information retrieval tests. The other question the researcher needs to answer is "How big of a difference makes a real life significant difference?". In information retrieval the response variable is often recall or precision so we must answer the question in terms of how large a difference in precision or recall between two searches is useful or practically significant.

Introduction

Although there has been a lot of work done on particular evaluation measures and there have been many retrieval tests performed over the years, it has been difficult to find statistical hypotheses and their corresponding tests which can be applied in information retrieval experiments. Hypotheses are used in information retrieval to test such things as which indexing method is better or which retrieval system is better.

Early on, the SMART system used the t-test, sign test and Wilcoxon rank test (Williamson, Williamson and Lesk 1971; Salton and McGill 1983) to

compare measures such as average precision and normalized recall scores between two retrieval tests. Tague-Sutcliffe (1992) advocated the use of analysis of variance (ANOVA) techniques, but also mentioned the Mann-Whitney test when the assumptions of ANOVA are not met. In all of these tests the researchers are assuming a comparison of two or more sets of retrievals (queries run against a database) with some characteristic of the retrievals being different, either indexing, searching, retrieval system used, etc. The evaluation measure in most cases was some form of precision, recall or other measures derived from the basic retrieved, not retrieved, relevant, not relevant numbers. Even when a different evaluation measure is being used, such as Saracevic's crossproduct odds ratio, it turns out that the test statistic has a t-distribution (Saracevic et al. 1988; Saracevic and Kantor 1988a; Saracevic and Kantor 1988b).

The purpose of this paper is to show the importance of the power of a statistical test. Also the relationship of various concepts connected with power such as sensitivity analysis and effect size will be explored in the context of testing the effectiveness of information retrieval systems.

Statistical Hypothesis Testing

To appreciate the concepts involved in power analysis, it is necessary to have a clear understanding of traditional hypothesis testing. This can be found in many basic statistics textbooks (such as Loether and McTavish 1980) but is reviewed here. It is assumed that one of two hypotheses, the null or the alternative, must hold. The hypotheses are stated like the following example:

> H_0: The two sets of retrievals give the same results.
>
> H_A: The two retrieval tests are not equal (give different results).

A level of significance is chosen, often 0.01 or 0.05. This is the probability that the null hypothesis H_0 is rejected when in fact it is true which is called a Type I error. One minus this probability is called the confidence; this is the probability that a null hypothesis is not rejected when it should not. If one fails to reject the null hypothesis and says the two retrieval tests are the same when in fact there is a difference, then one has committed a Type II error, whose probability is denoted by b. The

power of a statistical test, denoted by 1-b, is the probability of rejecting a false null hypothesis as we should. This can be summarized by a table. (Table 1)

Table 1

	Decision Made	
"Truth"	H_0	H_A
H_0	Pr(Correct Null) = 1 - α = confidence	Pr(False positive) = α = Pr(Type I error)
H_A	Pr(False Negative) = β =Pr(Type II error)	Pr(Correct Positive) = 1 - β = Power

The best experimental designs will try to minimize the probability both Type I and Type II errors. In most designs however, the emphasis is on choosing the a level and very little consideration is given to the probability of a Type II error or the power of the test except indirectly by controlling the sample size.

Variables Affecting Power

For the two sample t test, under the assumptions of independent normal errors with equal variances, there are five factors affecting power. They are α, sample size, ratio of sample sizes for the two groups, difference of means, and error variance (in the population). If equal group sizes are chosen then power is a maximum for this variable and we are left with four variables. Sample size and α can be controlled by the researcher. The larger the sample size, the greater the power and the larger α, the greater the power. Of course increasing a then increases the probability of a Type I error, so the two types of error are inversely related, all other things being equal. This leaves the two population parameters which are generally not known. They may be estimated from the sample in a post-hoc analysis but this does not help researchers to design a good experiment a priori. For more complex designs, such as ANOVA, the same results hold except the means and variances of the different groups are needed. Generally more advanced designs have controls on the variance and so tend to be more powerful. Since mean and variance depend on the scale of the measure being used, it is very

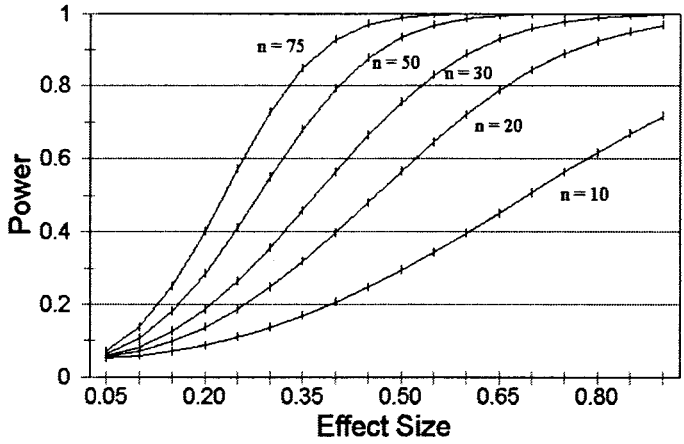
useful to introduce a scale invariant measure for power analysis called the *effect size*. The effect size parameter depends on the particular test being used, but for a two sample t-test it is calculated by dividing the difference of the population means by the common standard deviation (one of the assumptions is equal variance). Thus if equal group sizes are chosen, the variables are reduced to α, sample size and effect size. Effect size is thus related to the actual difference in the populations of two groups that is being measured by sampling. Unfortunately, the effect size is very difficult to know or even estimate in most cases.

For example suppose there are two systems using different ranking algorithms on the same database performing searches on twenty different queries. This is a paired t-test as the same query is run on both systems. Then if the mean difference between the systems as measured by precision is 0.1 and the variance of this difference is 0.09 (so standard deviation is 0.3) then the test statistic is 1.58 and the critical value of t at the $\alpha = 0.05$ level is 2.09 so the null hypothesis cannot be rejected. What is the power of this test? Trying to estimate the power of such a test post hoc can be done by using tables (Cohen 1988, 48). For this case the effect size is 0.1 divided by 0.3 which is 0.3333. The tables in Cohen give a power of 0.30. An alternative way to calculate power is to use one of the several computer programs available. For a lengthy review of software for power analysis see Thomas and Krebs (1997). The software used in this paper is called Gpower by Faul and Erdfelder (Gpower 1992) which is available for free over the Internet. Using Gpower the more accurate figure for the power is 0.2935. Whichever way the power is calculated in this case, one can see that given there is such a difference between the two systems there is only a 0.3 probability that this test will detect such a difference. The probability of a Type II error is then 1.0 - 0.3 = 0.7. Not a very good experimental design!

The next natural question is "What sample size do I need to get a reasonable power?". One way to accomplish this is to choose a reasonable power level, say 0.80, and to calculate the sample size that will give you this power keeping the effect size and a level the same. Using Gpower gives a sample size of 74. A more comprehensive view is graph the power against the effect size for different

sample sizes as in Figure 1. This is sometime called a sensitivity analysis because the variables involved are not linearly related and it helps to see how the three variables interact. For example, if n = 50 the interval from about 0.2 to 0.4 the power changes very rapidly, i.e. power is very sensitive to changes in the effect size. Another way to state this is to say that for some effect sizes the power can change rapidly with increases in the sample size.

Figure 1. Power as a Function of Effect Size
alpha = 0.05, Paired t-test.



Examples from the Literature or Post-hoc Analysis

What is a reasonable effect size for information retrieval experiments? When Cohen (1988, 26) discusses various statistical tests he recommends "small", "medium" and "large" effect sizes for that test. Although these are somewhat arbitrary, they are based on experience in social science research. For the independent t-test Cohen suggests small = 0.2, medium = 0.5 and large = 0.8. For the paired t-test these values are divided by $\sqrt{2}$ which gives use small = 0.14 medium = 0.35 and large = 0.57. So the example in the previous section is a medium effect size according to Cohen.

To calculate some effect sizes from the literature, data from a description by Williamson, Williamson and Lesk (1971) will be used.

They compared a stemming search with a thesaurus search on forty-two queries by using a paired t-test. The output measures used included log precision, normalized recall and precision at various recall levels. The mean of the differences and the standard deviations are reported so the effect size can be calculated (Table 2). Notice that most values in the "medium" range of effect size according to Cohen. Of course a different experiment comparing different search methods may have different effect sizes.

Table 2

Mean of differences	Standard deviation of differences	Effect size
0.0407	0.0788	0.517
0.0413	0.0771	0.536
0.0526	0.1348	0.39
0.0343	0.0739	0.464
0.0135	0.0434	0.311
0.0199	0.0583	0.341
0.0325	0.0758	0.429
0.0361	0.0805	0.448
0.0458	0.0994	0.461
0.0879	0.1483	0.593
0.0499	0.1569	0.318
0.0401	0.1462	0.274
0.0521	0.1288	0.405
0.053	0.1299	0.408

More complex research designs generally give better power. In particular, consider the repeated measures design used by Tague-Sutcliffe (1995) to analyze the TREC-3 results. This design also benefits from a large sample size of fifty queries and forty-two retrieval runs of these queries against the same database for a total of 2100 observations for the ad-hoc queries. Using Gpower section for "other F-tests" the power of this test to detect Cohen's medium effect size for ANOVA is very close to one. Even for small effects the power is 0.94. This means that even small differences between systems or between queries will be detected by the test. Tague-Sutcliffe (1995)

in fact showed the results are very significant, but that when a post-hoc Sheffé test is conducted there are very large groups of systems with no significant differences. Since the ANOVA test is very powerful this result is very reliable.

Non-Parametric Tests

Although the examples used have been parametric tests which have the assumptions of normality and equality of variances, it is still possible to carry out a power analysis when a corresponding non-parametric test such as the Wilcoxon test or Friedman test are applied. Generally, if the data does not satisfy the basic assumptions the non-parametric alternative often gives a more powerful result. This is beyond the scope of this paper but for those who are interested see the paper by Singer, Lovie and Lovie (1986).

Conclusion

One of the more difficult decisions when doing a power analysis is what effect size the research should be designed to detect. For information retrieval this comes down to deciding how big a difference in precision (or other output) measure is necessary in order to make a real difference in retrieval techniques or systems. In other words, how much of a difference in precision is important to the end user or searcher? This is not an easy question to decide as it probably depends on a number of factors like the purpose of the search, size of output, etc. It should be pointed out that this is a dilemma in most social science research: real world significance versus statistical significance. The advantage of power analysis is that it tells the researcher exactly what differences the statistical test is likely to detect.

References

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 3 ed. New York: Academic Press.

Loether, H. J., and D. G. McTavish. 1980. *Descriptive and Inferential Statistics: An Introduction.* 2 ed. Boston: Allyn and Bacon.

Muller, K. E., and V. A. Benignus. 1992. "Increasing scientific power with statistical power." *Neurotoxicology and Teratology* 14: 211-9.

GPOWER: A Priori, Post-Hoc, and Compromise Power Analyses for MS-DOS [Computer Program]. 1992. Bonn, FRG: Bonn University, Dept. of Psychology.

Salton, G., and M. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

Saracevic, T., and P. Kantor. 1988a. "A study of information seeking and retrieving. II. Users, questions and effectiveness." *Journal of the American Society for Information Science* 39(3): 177-96.

Saracevic, T., and P. Kantor. 1988b. "A study of information seeking and retrieving. III. Searchers, searches and overlap." *Journal of the American Society for Information Science* 39(3): 197-216.

Saracevic, T., P. Kantor, A. Y. Chamis, and D. Trivision. 1988. "A study of information seeking and retrieving. I. Background and methodology." *Journal of the American Society for Information Science* 39(3): 161- 76.

Singer, B., A. D. Lovie, and P. Lovie. 1986. Sample size and power. In *New Developments in Statistics for Psychology and the Social Sciences*, ed. A.D. Lovie, 129-142. London: The British Psychological Society and Methuen.

Tague-Sutcliffe, J. 1992. "The pragmatics of information retrieval experimentation, revisited." *Information Processing and Management* 28(4): 467-90.

Tague-Sutcliffe, Jean. 1995. "A statistical analysis of the TREC-3 data." In *Overview of the Third Test Retrieval Conference (TREC-3)*, ed. Donna Harman, 385- 398. Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-225.

Thomas, L., and C. J. Krebs. 1997. "A review of statistical power analysis software." *Bulletin of the Ecological Society of America* 78(2): 128- 39.

Williamson, D., R. Williamson, and M. Lesk. 1971. "The Cornell implementation of the SMART System." In *The SMART Retrieval System: Experiments in Automatic Document Processing*, ed. G. Salton, 12-54. Englewood Cliffs, N.J.: Prentice-Hall