Shiken: JALT Testing & Evaluation SIG Newsletter. 8(2) August 2004 (p. 2 – 5)

Some Preliminary Thoughts on Statistics and Background Information on SPSS (Part 1)

by H.P.L. Molloy and Tim Newfields

This article introduces a few basic statistical concepts for novice researchers. The subsequent two articles will highlight several additional fundamental statistical notions and mention how those are calculated using the SPSS Statically Software Package.

The role of statistics

Statistics can function in two basic ways. It can help us to identify patterns or factors related to constructs. Secondly and importantly, it can help us to not see patterns and recognize some events simply as random phenomena.

Human beings are very good at seeing patterns and this is useful in many respects. However, we are a little too good at seeing patterns for research purposes because we tend to see patterns in essentially random collections of things. The Rorschach test is one example of this tendency, as is the way we perceive constellations among randomly distributed stars, or discern familiar shapes among passing clouds.

Statistics can help us understand which underlying patterns might exist in any given set of raw data, and which apparent patterns are probably due to random chance. For example, imagine you had the test scores for two groups of students. Examining those raw scores and average scores, it may seem obvious that one group was "better". A t-test, however, may reveal this is actually not the case. Face interpretations of raw data are often misleading: this is why researchers rely on statistics to discern fundamental patterns.

Our friend the null hypothesis

Many statistical procedures and tests are firmly rooted in the assumptions of post-positivism. When using such procedures, what we are generally trying to do is to falsify a null hypothesis.

The null hypothesis says "there is no difference" between the items under scrutiny in a study. The null hypothesis is always given in negative terms: there is always a "no" or "not" in it. A factor in a study is considered significant only if there is enough statistical weight to actually reject a specific null hypothesis. The de facto starting position for any empirical research is a negative null hypothesis.

One might ask, "Why don't we just test our alternative hypothesis?" and begin with an assumption there is a difference in a research variable? Novice researchers might feel tempted to start with an alternative hypothesis rather than the null hypothesis. However, the post-positivist position on epistemology, the study of knowledge, contends convincingly that we can never prove positive hypotheses: we can only falsify hypotheses.

". . . the post-positivist position on epistemology, the study of knowledge, contends convincingly that we can never prove positive hypotheses: we can only falsify hypotheses."

Let us consider an example. Suppose you had the idea: "All of the cars in Tokyo are less than ten years old." How would you test this? We can walk though the streets and look at all the cars. Suppose we found 700 cars, all of them under ten years old. Have we proved our hypothesis? No: the next car might be over than ten years old. In fact, no matter how many cars we check we have to keep on looking, because it is possible that the next car could be over a decade old.

If we adopt the null hypothesis, however, things are different. The null hypothesis would be "None of the cars in Tokyo are over ten years old." We could use the same test procedure to investigate this. However, the first time we found a car over ten years old, the study would be finished. Since we have falsified the hypothesis, our work is done. Therefore it is important to understand that empirical research begin with a null hypotheses, then systematically explore ways to possibly falsify it.

Alpha values, beta values, and power

Alpha values

Statistics is a way of systematically investigating the difference(s) between groups of numbers. The information we are interested in is usually translated into numbers in some way or another, no matter what form it was in originally. Our aspiration is that any numerical difference should also reflect a difference in the real world. However, when gathering data, all kinds of inaccuracies (or "error") come in. For example, in a test of grammar ability, some students might not pay attention. Others might simply guess at the answers. Still others might answer some questions in novel or unexpected ways which are not actually "wrong" - just unexpected. All of these inaccuracies are regarded as random variations conceptually. Statistical tests should attempt to discern whether there is any difference in the raw data caused by random variations (or "errors") rather than by the effect which is being explored.

This is where the "alpha" value or "*p*" value comes in. (SPSS often calls it "**Significance level**" and abbreviates it as "**SIG**." This choice was unfortunate, given the usual synonymy "importance." Statistically significant does not mean important; it only means "probably not random.") The alpha value shows us what are the chances that a given difference in numbers is due to random error rather than the effect which is being explored.

Let's consider a case in point. Suppose you have done a t-test between the scores by two different groups of students. Group A had higher scores than group B. The results of the t-test was significant at $p < 0.05$. What does this mean? It means that there is a less-than-five-percent chance that the difference was due random variation (or "error").

> *"Type 1 errors . . . occur when the null hypothesis is rejected when it shouldn't be."*

The *p* value is also often mentioned in relation to Type 1 errors, which occur when the null hypothesis is rejected when it shouldn't be. If SPSS tells you that the " **SIG**." for your test is 0.06, for example, that means that you have a 6-in-100 chance of rejecting the null hypothesis when there was really no difference.

Alpha or *p* values are usually set at 0.05 or 0.01 in our field, but that's only a convention. If you can justify it, you're perfectly free to choose any alpha or *p* value. However, unless you have good reasons to use a weird *p*

value, it's best to stick to 0.05 or 0.01, as those are what people are used to.

Alpha, *p*, or significance, values should be chosen in the planning stages of a study - before collecting any data. You should have good reasons for not using 0.05 or 0.01, but also should also have reasons for choosing either 0.05 or 0.01. Is your study one that is only descriptive or is it exploratory, done in preparation for more rigorous research? You're probably safe with 0.05, as no big decisions depend on the chances of making an error. Is your study one that will affect people's lives (such as one involving entrance examinations)? You might want to use the more conservative 0.01 value. In medical research, values of 0.001 or 0.0001 are often chosen since people may die because of a mistaken decision.

Again, the alpha value you choose determines how great the chance will be that you are willing to accept making a mistake. It does not have anything to do with how strong the effect you find is: if you find that your study yields of significance value of 0.06 and you chose 0.05 as the acceptable chance of making an error, you have not found something "almost significant." You have found nothing. Move on.

What if no statistical difference is found? Does that mean that the null hypothesis is accepted? Yes. Does that mean that there really isn't any difference? Not necessarily.

Beta values and power

If you find no difference, it could be that you didn't look hard enough. Perhaps you would have found a difference if you had looked at a larger number of students.

The situation is analogous to looking, for example, for something lost in your home. Imagine you have misplaced your house key. You look through your kitchen, going into cabinets, checking the table carefully, and scanning the floor. You don't find the key. The chances are pretty good, perhaps we could say 19 out of 20, that the key is not in the kitchen. There is only a 1 in 20 (or the familiar 0.05) chance that the key is in fact in the kitchen. Can you now say that the key is irretrievably lost? No. You haven't checked the rest of your home. There may still be a pretty good chance that the key is elsewhere. If you don't check the rest of the house, you don't have such a good chance of finding the key.

Search the rest of the house can be conceptualized of in terms of beta values. The beta value shows the complement of the alpha value: it shows you the chance of mistakenly deciding that things are the same when they are really not.

This is often spoken of as Type 2 error, or the chance of retaining the null hypothesis when it is not true. The beta value most often comes into play SPSS under the name of power, or statistical power. Power is calculated by subtracting beta from 1, so that if you have a beta value of 0.26, the power value will be 0.76. Power, in the house key analogy, can be thought of as "how hard you looked." It describes how likely you were to find the key, given the search that you did.

> *"[A] Type 2 error . . . [is] the chance of retaining the null hypothesis when it is not true."*

Generally, a statistical power of 0.80 (that is, a beta value of 0.20) is considered adequate. What can you do if your power is too low? There are five principal strategies to increase power for any particular statistical procedure, only the last of which strikes me as being both ethical and practicable for most foreign language teacher-researchers.

The first strategy is to use a more liberal alpha value: a value of 0.05 gives greater power than an alpha value of 0.01, if all other things are equal. As mentioned above, however, it is a bad (and unethical) idea to change your alpha value after you've begun your study.

The second strategy is to do your work again, using a stronger treatment. This option only works with experimental or quasi-experimental studies. If you have been studying how planning effects performance on a test and find no effect (that is, do not get a significant result), you may find one if you do the study again, giving participants more planning time. This means starting your study again from scratch.

The third way involves decreasing the variability of the numbers you are studying. Perhaps something you didn't think of was affecting the numbers you work with: if you can somehow account for this, you may end up with less-varied scores and so more power. For example, perhaps you designed a study to see which of two ways of learning new lexical items is more effective. You may realize, after collecting the data study, that linguistic proficiency may affect the ability to memorize new lexical items. f you manage to get hold of TOEIC® scores for your participants, it may be possible to statistically account for proficiency and so increase the power of your study. This is not always practicable, however.

The fourth way to increase power is to make your measurement instruments more accurate. Like other strategies, making measurement instruments, your tests or whatever you are using to measure, more accurate reduces error. This can be thought of as making it easier to see differences. It is always a good idea to use as accurate an instrument as you can (hence the desirability of piloting), but if you have already collected data it is not always possible.

The fifth, and easiest, way to increase statistical power is to increase the sample size. Get more participants, test more students, or interview more people: this will have the effect of decreasing the variability of the numbers. Moreover, it is usually easier to do than rerunning your study or redesigning your statistical analysis. Hence this is often the easiest way to increase statistical power. (Though it is possible to have a study with too much statistical power, in our situations this is unlikely. For example, if you managed to do a study involving every living soul in Japan and divided them into two groups, you would be likely to find a statistically significant, non-random difference between the groups - almost regardless of two groups you used. Whether the difference is important or not is another question.)

Summary

This article has briefly mentioned the role of statistics, the importance of the null hypothesis, and some background information on alpha and beta values and the concept of statistical power. The next article in this series will mention the difference between counting, ordering, and measuring, various types of distributions, and the distinction between parametric and non-parametric statistics. The final article in this series will explore some concepts in descriptive statistics and recommendations for using SPSS.

HTML: http://jalt.org/test/mn_1.htm / PDF: http://jalt.org/test/PDF/MolNew1.pdf