Statistical Tests - General Remarks

Statistics can be used for lots of purposes. Broadly speaking we can distinguish two kinds of statistical statements. One sort of statement is some sort of description. For example, we might note that the average German folksong is 23.4 notes in length or that nearly a third of Western classical music is written in triple meter. A more interesting statement might be: Of all European countries, Italian music makes the least use of the minor mode. All of these are called *descriptive statistics*. Descriptive statistics can be useful, but they have a certain "so what?" quality to them. They simply describe something about the world.

There is a second type of statistics, called *inferential statistics*. These statistics are used to test hypotheses or conjectures. They are the most important types of statistics in modern empirical research.

Confidence Level and Statistical Significance

Recall the concept of the *confidence level*. The confidence level is the line-in-the-sand drawn by the researcher that indicates when the research admits defeat. It is expressed as a percentage, e.g. 95% confidence level. Related to the confidence level is the *significance level*. The significance level is reciprocally related to the confidence level. It is expressed as a probability rather than a percentage. A 95% confidence level corresponds to a .05 significance level. A 99% confidence level corresponds to a .01 significance level. The significance level is symbolically represented by the value *alpha* (α). Formally, the significance level is the probability of rejecting a true null hypothesis.

Statistical Tests

Over the past century or so, statisticians have devised a number of statistical tests. Examples of statistical procedures include Student's t-test, Wilcoxon signed rank test, the chi-square test, Kruskal-Wallis test, and the Kolmogorov-Smirnov test. There are also various statistical values that can be computed, such as Pearson's *r,* Kendall's *tau,* and Spearman's *rho*.

Why, we might ask, are there so many statistical tests? The answer is that each test is specialized for a different data situation. Three conditions are especially important: (1) the type of measurement scale, (2) the distribution of the data, and (3) the data-gathering conditions.

In the first instance, recall that there are different measurement *scales*: nominal, ordinal, interval and ratio. For example, Pearson's *r* is a good way to calculate correlations, but only for interval and ratio scales. If the data are ordinal (1st, 2nd,

3rd, etc.) then one should use Spearman's *rho* instead.

A second consideration is the *distribution* of the data. There are many different kinds of distributions. For example, many types of data are normally distributed. That is, the data exhibit a familiar bell-shaped curve where most of the values reside in near the center. Sometimes the data exhibit a uniform distribution, with each value equally likely (as in the case of numbers on a roulette wheel). Another kind of distribution is the bimodal distribution. Here, the data exhibit two peaks rather than one. The pitch of people's voices exhibits a bimodal distribution: the lower peak corresponds to the pitch of the average male voice, and a higher peak corresponds to the pitch of the average female voice.

Data may be conform to any of a large number of distributions, such as a logarithmic distribution, a chi distribution, or a binominal distribution. When the distribution is known, there are various mathematical tricks that allow statisticians to infer various properties of the data. Most of these tricks have been developed for the normal distribution, so when the data is normally distributed, it is possible to use more powerful statistical tools. However, in many research situations, we don't have any idea how the data are distributed.

Statistical tests differ in a number of other ways. For example, some are used to compare ratios, some are used to compare means, some are used to compare variances, etc. In addition, there are different statistical tests that are employed depending on the manner in which the data are collected. For example, different tests may be required depending on whether the experiment employs *within-subjects* or *between-subjects* design. For each combination of different conditions, there is an appropriate statistical test.

Each statistical test is said to employ a "model." Each model makes various assumptions about the data. For example, all statistical tests assume that the data are independent. In order for the test to be legitimate, the data must conform to the assumptions (or "model") for that test. (Once again, non-parametric tests make the fewest assumptions.)

Parametric vs. Non-parametric Tests

As noted, when the distribution of the data is known, various mathematical tricks can be used and statistical procedures can be used that extract more information out of the data. Statistical tests based on known distributions are referred to as *parametric tests*.

It is common, however, for the researcher to have no knowledge about the distribution of the observed data. When the distribution is unknown, we employ so-called *non-parametric tests*. Non-parametric tests cannot squeeze as much information out of a data set as parametric tests. However, non-parametric tests are appreciated because they can be used for a much wider range of types of data. *Non-parametric tests* are conservative tests that make fewer assumptions about the data.

Examples of non-parametric tests include the chi-square test, the Siegel-Tukey test, Spearman's rank correlation coefficient, Wald-Wolfowitz runs test, and the Wilcoxon signed-rank test.

Non-parametric tests make fewer assumptions about the data. When using a non-parametric test, there is less chance of violating one of the assumptions of the statistical model behind the test. Consequently, the researcher is less likely to get into trouble when using a non-parametric test. If there is some uncertainty about which parametric test to use for a given set of data, the researcher may elect to use a non-parametric test instead. Especially if the results prove statistical significant, the non-parametric test is clearly adequate. However, using the appropriate parametric test is more likely to produce a significant result — compared with a non-parametric test.

Statistical Significance

A statistical test allows us to calculate the likelihood of accepting the null hypothesis in light of the collected data. This probability is referred to simply as *p*. The calculation of *p* is the most important part of a statistical test.

In most research, the researcher is hoping that the research hypothesis is true. This means that the researcher hopes that the null hypothesis can be rejected. Accordingly, we are typically looking for a small probability in favor of the null hypothesis — that is, we are looking for a small value for *p*.

A result is said to be "statistically significant" when the value of *p* is less than the significance level, alpha. For example, if our significance level (α) is .05 (corresponding to a 95% confidence level), then a *p* less than .05 means the results are "statistically significant." If our significance level (α) is .01, then *p* must be less than .01 in order for the results to be statistically significant.

Degrees of Freedom

In research, we commonly use a sample to infer some property of a population. Recall that the *law of large numbers* states that as the sample size increases, the sample is more likely to provide a more accurate estimate of the true population value. The greater the number of observations, and the greater the number of ways values are free to vary, the greater the likelihood that our results are accurate. In statistics, the number of ways that values are free to vary is referred to as the *degrees of freedom* (abbreviated *df*). Degrees of freedom are related to, but not the same as, the number of observations. Different statistical tests interpret degrees of freedom differently. In all cases, the degrees of freedom plays a pivotal role in calculating *p*.

Reporting

When reporting the results of a statistical test, two forms of reporting can be distinguished: *formal* and *informal*. A formal statement might look like this:

> In comparing the tempo of minor-mode works with major-mode works, we carried out a matched-pairs *t*-test, whose results permit rejection of the null hypothesis at the prior established 95% confidence level (t=14.93; df=88; p=.0003). Consequently, the observations are consistent with the alternative hypothesis, that the tempo of minor-mode works is slower than major-mode works.

This formal statement (1) identifies the type of statistical test carried out (matched-pairs *t*-test), (2) reminds readers of the confidence level (95%) — which would have been identified earlier in the article, prior to any data collection, (3) reports the value of the statistic, in this case *t* (t=14.93), (4) identifies the degrees of freedom, abbreviated *df* (88), (5) reports the calculated value of *p* (.0003), (6) reminds the reader that the statistical test is a test of the null hypothesis (that there is no difference in tempo between major- and minor-mode works), and (7) having rejected the null hypothesis, we accept the alternative (or experimental) hypothesis, namely, that the tempo of minor-mode works is slower than major-mode works.

This sort of formal reporting is almost never used in contemporary empirical research publications. Instead, researchers prefer a shortened, less formal, way of reporting statistical tests. Here are a couple of examples:

> We found a significant correlation between size and shape (r=0.83; df=80; p<0.0001).

> There appears to be no relationship between stem length and notehead slant (Kendall's T=0.98; df=142; p=.73).

Sometimes, the researchers will remind the reader of the *presumed* a priori confidence level.

> The results are consistent with the hypothesis at the 95% confidence level (χ^2=4.9; df=2; p=0.02).

This might seem gratuitous because it is rare that any research report will state *a priori* the selected confidence level before collecting their data. The 95% confidence level has become so commonplace in research that it is assumed to be THE minimum criterion for statistical significance. In other words, common research practice has tended to "fix" the line-in-the-sand. Statisticians remind us that there is nothing magical about the 95% confidence level. A researcher is free to choose whatever confidence level they deem appropriate (according to the moral repercussions of making a Type I or Type II error). However, the 95% confidence level has simple become a sort of "social norm" for research.

More About Reporting

Nearly all statistical tests make use of the following: (1) confidence level, (2) significance level (α), (3) calculation of an appropriate statistic, (4) calculation of *p*, (5) degrees of freedom, and (6) effect size.

The word "significant" should not be mistaken for "important." Something can be statistically significant, but not especially important. In the past, only a minority of research papers reported the effect size.

By way of illustration, there is a small but statistically significant association between living under a high-voltage electrical transmission tower and death from leukemia. However, the effect is miniscule. In the United States, a person is roughly 10,000 times more likely to die from an automobile accident.

Examples of statistics related to effect size include: R-squared (in correlation), (for chi-Square, phi); Cohen's d,

In statistics related to Pearson's correlation coefficient, we are interested in five interrelated values: *r, df,* α*, p* and r^2. We interpret all five values with respect to each other:

1. What is the correlation (*r*)? Is the correlation big or small, positive or negative? Correlation values range between -1 and +1. We cannot take the correlation coefficient at face value: *r* values may be *real* or *spurious*. We need to interpret *r* with respect to other statistical values.

2. What are the degrees of freedom (*df*)? Are there lots of observations, or just a few? Note that degrees of freedom are related to (though not exactly the same as) the number of observations. Degrees of freedom range between zero and infinity.

3. We use the correlation value (*r*) in conjunction with the degrees of freedom (*df*) to calculate the probability (*p*) that the relationship might arise by chance. *p* values range between 0 and 1.

4. The significance level (α - alpha) is where the researcher draws a line in the sand and says "If the probability of observing this relationship by chance is above this value, then I'm going to accept that the correlation could simply arise by chance, and therefore I will admit defeat — that my hypothesis is wrong." Remember that the significance level must be set before you calculate *p*. ("*It's not research if you don't invite failure*.")

   Notice that we never test our experimental hypothesis directly. Instead, we test whether the results look like they could arise simply by chance. ("*Aim not to be right, but to be not not right*.")

   If the *p* value is bigger than the *a priori* significance level, then the correlation value (*r*) is considered *spurious* — we are simply looking at "noise" (i.e.,

natural variation). However, if the *p* value is smaller than the *a priori* significance level, then the correlation value has a high likelihood of being *real*.

5. If our correlation is considered statistically significant, we can now calculate the effect size (r^2). The r-squared value tells us how much of the variability in one variable can be attributed to the other variable.

Seeking Help

Acquiring an appropriate level of knowledge about statistics is one of the empirical researcher's foremost challenges. Statistics is a rich and powerful field of knowledge. At times, statistics can seem magical — pulling useful information out of observations that otherwise appear to be a confusing mess. For the novice, the whole enterprise can seem intimidating. As you gain experience, you will become more comfortable with statistics. There are excellent software packages that will do all of the complicated calculations. But you will need advice about the most appropriate statistical test to use, given your project. The best advice for the novice researcher is to seek help:

Slogan: *Make friends with a statistician*.

Even experienced researchers appreciate the opportunity to chat with a statistician. Most important, talk with a statistician *before* you collect any data. Describe what you are planning to do and listen carefully to the advice. Statistical consultants are thrilled when people come and talk with them before the data are collected. Take advantage of their expertise.