Statistical Analysis Q&A

Purpose

This document is intended for researchers intending to conduct clinical trials with human subjects. The goal is to provide a step-by-step resource for beginning the statistical plan of a clinical trial. Although some information may be rudimentary, this document details important steps in their order of approach. This document is non-comprehensive, and intended to serve as a general guide.

Contents

Statistical Considerations

The basic considerations for statistical plans are: the type of statistical analysis, defining clinically meaningful importance, the number of patients required, whether you will be testing for significance, desired level of significance or confidence, and power analysis.

1. **I'm conducting a pilot study, what do I need to know?**
    a. Sample size, alpha level, when/how to conduct a power analysis and appropriate statistical analyses, level of desired significance, and if possible, an estimate of the actual or estimated results, variance, and effect size.

    b. Most of you are probably conducting a pilot study prior to a larger clinical trial – examining a treatment/device that hasn't been used before, or is now being used for new purposes. These studies are frequently used as a portion of grant submissions. Sample size (*n*) or the number of participants you'll need, is important. Although it's important to conduct power analyses for larger studies to determine the likelihood of correctly indicating that an effect is or isn't there, pilot studies are too small to conduct tests of significance, but instead are used to inform future work. Still, pilot studies that are too small mean that future research would have little to go off of – and would instead become another, larger pilot study, pushing back the process of determining whether a treatment/device works. The tradeoff is that small studies are more feasible to conduct and in the past have sometimes meant publications (that standard is now changing). However, there's a growing push to justify sample size, and a broader awareness that many poor practices have been used in the past – meaning, amongst others, that utilizing a particular sample size because it's been done before, is less feasible. Importantly, justification of sample size is often expected by review panels for grant funding.

    c. All pilot studies (and full clinical trials) should report descriptive data such as proportions, along with a general recommendation for including confidence intervals. But, the sample

size for a pilot study should still be justified even though no inferential statistics will be conducted. Essentially, there are three options:

i. Follow a general guideline to rationalize a small sample size (e.g., an n of 12 per group [3]), but be unable to conduct any inferential statistics or draw stronger conclusions than a better estimate of whether an effect could exist in future studies. Clinically important differences, although not statistical differences, may be reported, but should not be interpreted as evidence that the trial will work.

ii. Calculate a power analysis for the desired statistical analysis, knowing some important values from a range of information, including population means, variance, expected size of the effect, etc. This is typically unfeasible for a pilot study, given that it is examining work that has not been done before.

iii. Calculate a power analysis (see Section 5) for an 80% confidence interval. In this situation, you are testing for *clinical importance*, not statistical significance. The difficulty with this approach is that unless the difference between groups is expected to be large (in which case, sample size is pleasantly small), the *n* becomes somewhat less feasible (this is the same problem as when power analysis is involved, however). The benefits of this approach are that this trial serves as a useful basis for future studies (and well-informed future power analyses), is more likely to become published, provides a greater added value,

iv. and, should the calculated sample size be sufficiently large, according to a subsequent power analysis, inferential statistics may be conducted to test for a statistically significant difference (although, these results should be interpreted cautiously, since the power analysis is unlikely to completely reach the typically desired level).

    1. The methods for conducting a confidence interval power analysis are to determine what clinically meaningful improvement would look like, how many participants are necessary to detect that the meaningful improvement has not occurred (using the goal you are improving from, such as average scores, with standard deviations), and what percentage of the sample size for the subsequent main study would be required (i.e., about 9% that of the anticipated main study). Working through those steps:

v. **What would searching for a clinically meaningful improvement look like?**
    1. Suppose an average quality of life rating was 8 on a scale of 10, with a standard deviation of 2. A goal of improving this rating by 1 point would be a reduction of ¼ of a standard deviation. Using an alpha of .05 and power of 80% for a one-sided confidence interval, the calculation is for 9% of the *n* for the intended main study. Calculating a power analysis for the lowest *n* that excludes .25 from the upper confidence limit results in an *n* of 45.6, rounded to 46 for an even *n* of 23 per group (the main study *n* would be 504). Attrition rates should be estimated based upon prior work or clinical experience, and added to the calculated *n*.

    2. ***Importantly, lack of statistical significance does not mean lack of a clinically meaningful result.*** Thus, a pilot study or clinical trial can have

important findings without reaching a desired level of significance. Ideally both clinical and statistical significance should be reported, although studies in the past have often only reported statistical significance. The benefit of defining and reporting clinically meaningful results is that your findings are more likely to be considered meaningful by those reviewing your data, so it really does strengthen the condition of your study with minimal additional effort.

vi. Finally, *IF* you have the correct type of data or situation for conducting an odds ratio, relative risk, or robust variations of more common analyses (e.g., Fisher's Exact Test, Mann-Whitney U, Kolmogarov-Smirnov, Wilcoxon, Kruskal-Wallis), you do *NOT* need to worry about sample size! Each of the non-ratio/risk analyses is a variation of the chi-squared test, t-test, or ANOVA, and you should still calculate a power analysis to demonstrate that those typical statistics cannot used – and report this rationale as justification for these alternative analyses, but your statistics will be accurate even if your *n* is small (e.g., <5 per group). You should also report a rationale for using the small sample size that made it impossible to conduct the standard inferential analyses – e.g., an extremely small available population or recruitment success. Just note that for each of these analyses, either power or the ability to detect differences is generally very low under small sample sizes (e.g., < 12), making it difficult to find significant effects. For a detailed overview of when these statistics are appropriate, see http://www.annemergmed.com/article/S0196-0644(05)82571-5/pdf [1].

2. **Choosing a type of analysis: What is appropriate for my data?**
   a. There are two basic types of analyses: descriptive and inferential. All data analyses should include descriptives. Descriptives are exactly that – descriptions of your data, using proportions, variance (standard deviation or standard error), count data, etc. Inferential statistics are those commonly known as significance testing, and typically are used to make predictions about how the treatment/device would perform on other patients in the future. Inferential statistics are not used in pilot studies. Importantly when beginning to consider the appropriate analyses, the Consolidated Standards of Reporting Trials (CONSORT) checklist (http://www.consort-statement.org/ provides a broad overview of what information clinical trials should report.

3. **Which type of inferential statistics should I be considering?**
   a. The most common types of analyses for clinical trials are t-tests, confidence intervals, correlations, chi-squared tests, different types of ANOVAs or regression, and either the odds ratio or risk ratio. Non-inferiority or superiority trials are also frequently used, but are plagued by a large degree of complexity and debate over correct practices. Choosing which analysis is appropriate for your data isn't simple, particularly since there's been a recent push identifying incorrect statistics in older publications – i.e., that basing chosen statistics on what's been done in the past may not be either accurate or acceptable – and towards selecting more appropriate analyses, which are not necessarily as intuitive. Possibly one of the most useful "newly recommended" techniques is the odds ratio. This method provides no p-value or demonstration of a "significant" effect, but does provide a confidence interval for how likely the results are to be a true effect – and most importantly, unlike many other techniques, it can be used for small sample sizes. Other extremely useful tests are "exact" tests – e.g., Fisher's Exact Test. This particular test is a variation of the chi-squared test,

used to determine whether a difference between conditions exists for count data. Importantly, these tests provide a fairly "exact" depiction of your results, and are not harmed by small sample sizes – a reason many otherwise appropriate analyses would fail to be appropriate. For a comprehensive explanation of the most common analyses, as well as which one is most appropriate for your data, can be found at the following link: https://s3.amazonaws.com/academia.edu.documents/34639004/10.1111_jocn.12343_1_.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1510696827&Signature=pQl4TVDXY4HZMjL%2FUe0m%2FpaLMfk%3D&response-content-disposition=inline%3B%20filename%3DSelecting_the_most_appropriate_inferenti.pdf [2].

4. **What are significance levels, and what am I looking for?**
    a. Significance levels determine how likely you are to incorrectly claim that there is a statistical difference between groups. Typically this is a value of alpha = .05, or a 5% chance of incorrectly claiming an effect. This value will sometimes be more stringent, particularly in cases where it is very important not to make a false claim, in which case the value is 2.5%. In situations where one-tailed or one-direction t-tests are run, 2.5% is also generally used. In clinical trials, a value of 10% is sometimes used in situations where a tradeoff can be made between falsely claiming a statistical difference when none exists, and the opportunity to identify a particularly important or difficult-to-identify finding that might otherwise be overlooked. In general, not much thought needs to go into selecting a significance level – the value of .05 is chosen and does not need to be justified, whereas less stringent criteria must be.

5. **What is a power analysis, and how is it related to the number of patients needed for my trial?**
    a. First, power analysis is an important part of research design that hasn't always been reported in the past, but is becoming increasingly expected of researchers. It determines how many patients or participants are necessary to detect an effect of a certain size (i.e., how much better/worse one treatment group is), with a pre-defined level of confidence. For common power values of 80% or even 95%, should a difference between conditions exist, you would detect this difference 80-95% of the time. As the power level increases, so does the required *n*, but we can be more confident in those results. For clinical trials, a power level of 80% is generally viewed as acceptable.

    b. Importantly, the power analysis itself depends on the type of statistical analysis intended to be run for the primary study hypothesis. Thus, *before* determining how many participants should be run, it's necessary to know *how* the collected data will be examined.

    c. Power analyses can be conducted knowing a variety of information for these calculations, but commonly are based on: the type of statistical analysis, the chosen level of power, significance level, and if possible, any average outcomes from prior work.

    d. After collecting your data, you can go back and "fact-check" your power to ensure it reaches the same power your sample size calculation intended to achieve.

e. *ALL* power analyses should be conducted based on the most important outcome of interest (i.e., although multiple outcomes may be measured, your goal would be to maximize your chances of determining the significance or non-significance of the most important effect).

6. When can I avoid using a power analysis?

i. It is only appropriate to exclude a power analysis when no significance testing is to be done. That is, when only descriptive data (e.g., averages and variance) are reported. This circumstance occurs when too few patients are involved in the trial to test for significance – a situation that can be identified by conducting that same power analysis, inconveniently.

7. What is an effect size, and what role does it play in my results?

a. Effect sizes describe the size of your findings' import, and can occasionally be important even when a result is not significant. For example, if your work comes close to, but has not quite reached the p = .05 level, the effect size may demonstrate there is still an important difference or contribution, even though it isn't statistically significant. Conversely, if your result is significant, an effect size may indicate either that the significance is a poor indicator of your results, or further emphasize the strength of your results. To summarize, effect sizes prevent the overestimation or underestimation of an effect. In a "new statistics" movement, the emphasis on significant p-values has been reduced, considering them as supplementary, and instead placing the focus on confidence intervals and effect sizes. Some journals have even begun to require that p-values not be included. However, I would not ignore p-values just yet – just keep this idea in mind: that statistics are rapidly changing from what we've known in the past. Effect sizes are good to report! Some journals have begun to require them, others have not yet caught onto the movement but are likely to in the future.

8. If my study is *not* a pilot study, but instead a full clinical trial, what should I be doing?

a. Exactly everything described above, except inferential statistics and significance testing will always be required, and power analyses become extremely important rather than "just" important. Reporting effect sizes or 95% confidence intervals in additional to all other statistics should again strongly be considered, even more so than in a pilot study. Finally, instead of utilizing 9% of the required sample size (for a pilot trial, described above), you will need 100% of the *n* from your power analysis, which will typically become a substantial number of participants.

References

1. Gaddas GM, Gaddis, ML. Introduction to biostatistics: Part 5, statistical inference techniques for hypothesis testing with nonparametric data. Biostatistics. 1990; 19:1056-1059.

2. Bettany-Saltikov J, Whittaker VJ. Selecting the most appropriate inferential statistical test for your quantitative research study. Journal of Clinical Nursing. 2014; 23:1520-1531.

3. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharmaceut. Statist. 2005; 4: 287-291.