Part 1

Dear Biomech-L readers,
Recently I posted a question regarding concerns with statistical analyses in studies with small sample sizes. The question generated a lot of interest - I received over 50 replies, not including those who were just requesting that I post a summary of replies. The sheer number of responses reflects how powerful the listserve is as a resource for information.
I tried to post the summary of responses in a single email, but I failed because the message was greater than the maximal size allowed. Therefore, I have broken the replies into two separate emails. Part 1 has a summary of all replies, and Part 2 has the actual replies by all persons.
The original post was:
Dear Readers,

I am hoping that some of you who have expertise in the area of statistics and scientific journal review can help me with the following concern.

Recently I have submitted papers to peer-review journals that describe the results of investigations performed on 'small sample sizes'. Obviously, small is a relative term. For the sake of this discussion, my samples sizes have been greater than 4 and less than 10 persons.

Multiple times I have received reviewer comments that the sample size was too small, which limited my results. What concerns me is that most of my investigations involve a repeated measures design, during which subjects are tested in two or three environments, with the objective being to determine the affect the environment has on my measure. I typically use a paired t-test or repeated measures ANOVA with a post hoc evaluation if a significant main effect is determined.

In these papers, if I were to fail to reject the null hypothesis, then I can understand the concern about power and sample size. However, in many of my papers, I have determined a significant effect relative to an a priori $p<.05$. In this case, isn't the concern for a small sample size negated, since we did reject the null and thus did not commit a Type II error?

If I am mistaken in my reasoning, I would appreciate any advice on how to avoid this issue in the future (other than to increase sample size, which for me is often not possible and the reason why I have a small n in the first place)? If I am not mistaken, then how should I handle this issue with reviewers? I am concerned they are simply making an assumption that the sample size is too small based on a subjective judgment.

I bring this issue to the group because in our area of science, limited sample size is often an issue.

Responses were generalized into 4 main areas:
*

Correct Assumption - the finding of a significant difference indicates that there were a sufficient amount of samples. (7 replies)
* Concerns about meeting the assumptions for an ANOVA (Type 1 error) (16 replies)
* Application to the population - potential that the small sample size may not be representative of the population that you wish to infer your results to. (5 replies)
* Provide Effect Size calculations with confidence intervals to give an indication of the magnitude of the difference (2-3 replies)
* Perform a priori power analysis to justify the small sample size (11 replies)
* General Comments - (9 replies)

Overall, there appears to be some disagreement among the readership about the question. While there were many more replies suggesting alternatives and reasons why there are problems with small sample sizes, 15% of the responses suggest that if a statistical difference was found, then adequate power existed.
I believe that the comments regarding potential violations of normal distribution, sphericity, etc.., in the sample make perfect sense, and those tests should be performed and included in a justification of why the small sample is adequate. If the violations require use of non-parametric methods, then the appropriate tests should be conducted.
Regarding the external validity of the results, readers had good suggestions and concerns about how generalizable the data from a small sample can be to a larger population. It appears that the researcher

needs to address this as a limitation or address this issue in the project report to control for this factor. This concern is a case-by-case issue. For example, if your sample has very few subjects who are all males between the ages of 22-26, it may not be appropriate to generalize the results to the entire world population. On the other hand, it may be appropriate to generalize the results to persons with very similar characteristics to the chosen sample.

The use of effect size calculations to accompany and/or replace parametric and nonparametric tests is a very good idea, and while my graduate statistical training did not emphasize the use of effect size calculations, it seems that it is a good rule of thumb to include these calculations in all papers regardless of the sample size. Since tests of means will only tell if a differences exists, but not the magnitude of the difference, effect size and confidence interval computations give the reader greater insight into the size of the difference. Used without tests of means, effect sizes can still be used to evaluate differences between means without strict statistical testing. Unless I am mistaken, this approach allows the researcher to talk about the effect of a treatment while allowing the reader to draw their own conclusions.

Comments about the a priori power analysis also make sense, and power analyses should be performed prior to any investigation as a normal course of planning. However, the issue at hand is not that too few subjects were tested due to poor planning, but rather that a limited amount of subjects could be tested due to constraints outside of the investigator's control. The a priori power analysis results might be useful in defending the results should the small sample actually have sufficient power. However, in many cases, this is not the case, which is the issue we are concerned about.

General comments included recommendations for various freeware that can be downloaded and used in analyses, and recommendations for 'rule of thumb' subject sizes.

Overall Lesson Leaned

Based on the collective wisdom of the responses from the group, I feel the following needs to be taken into account when working with small sample sizes:

*

Perform an a priori power analysis to determine just how small your sample is compared to the size necessary to have sufficient power

* Test your data to determine if they meet the assumptions necessary for the parametric test of choice

* Use effect size and confidence interval calculations to determine the size of any measured effect regardless of the other statistical tests used
* Be sure to justify why a small sample was used and to discuss limitations to generalizability and increased chances of Type 1 error

The attached summary is rather long, but I suggest that you read through all of the responses as they are all of value. I organized them in the order presented in the summary.

John

John De Witt, Ph.D., C.S.C.S.
Exercise Physiology Laboratory Lead / Biomechanist
Exercise Physiology Laboratory
NASA - Johnson Space Center
john.k.dewitt@nasa.gov
281-483-8939 / 281-483-4181 (fax)

Part 2

Continuation of the summary of the replies to the question posed by John De Witt regarding statistical power are sample sizes. Part 2 includes responses in the categories of 'CORRECT IN THE APPROACH' and 'CONCERNS ABOUT MEETING ANOVA ASSUMPTIONS'

SUMMARY OF RESPONSES

CORRECT IN THE APPROACH

John,

I agree with you that these reviewers are wrong. Sample size 5 can be enough, if you do an experiment in 5 subjects and the effect of an intervention or treatment is always in the same direction, a simple non-parametric test will tell you that there is a probability of 1 in 2 to the power 5 that this occurs by chance. This is less than 0.05 which is considered enough. I agree with you that in such cases, the sample size turns out to be sufficient (but only after getting the results).

The example also shows nicely that a sample size of 4 is always too small.

Good comment on the null hypothesis. If you were to hypothesize that a treatment has no effect, and you get one effect in 2 subjects and the opposite effect in 3 (i.e. exactly what you would expect if there was no effect), the same non-parametric statistic will tell you that this can happen by chance with a probability much higher than 0.05, so there the sample size is not enough for the conclusion that there is no effect.

Ton van den Bogert

Hi, John. I would think by definition that if you find a significant effect, the sample size is not too small. From all statistics books I have read, performing a power study after the fact (as reviewers have suggested to me in the past) is essentially meaningless. If you fail to find a significant effect, then one reason can be that the study is underpowered. Then I guess it could make sense to do a post-hoc power study to see how big a sample size you would need. But if you find a significant effect, I think your sample size is big enough.
Just my two cents,
Dana Carpenter

Statistical power increases with sample size therefore if an effect is weak you need a larger sample to discover a difference. If the effect is strong, you don't need a large sample. Since you got a significant difference with a small sample than the effect was strong. Therefore, I totally agree with you that you did NOT need a larger sample. It would have been a waste of time and money.
Gordon Robertson

Dear John,
I agree with you. If you found a statistically significant effect then you had enough power. The only problem may be for results that were not statistically significant. Perhaps if you included the effect size (required by some journals) and power you may be able to address the reviewers' concerns.
Regards,
Danny Russell

Hi John
I think that you are right on.
If you collect data from substantially more people or other animals than is needed, I would argue that it is unethical.

The repeated measures design is quite powerful and not all reviewers recognize that.
Perhaps you should include your apriori power calculation in the methods section of the paper.
rk
Rodger Kram, Ph.D.

Hi John,
I think you are exactly right. If you are rejecting the null hypothesis, type II error is not a concern. I assume you have appealed to the editor
and referred the reviewer to a statistics text book? Obviously this can be
a touchy issue and should be phrased appropriately, but the bottom line is that you are correct.

_____
Samuel R. Ward, PT, PhD

Hi John,
I emailed an article from JBJS that you might find interesting; although the author discusses the improper use of a post-hoc power analysis in cases where the results are not significant, I think that much of it is relevant to your question. As long as you have performed a power calculation using a clinically meaningful difference, then your study should have adequate power. If the results are significant, then I don't quite understand why there is even a question about power; perhaps you could reference this article in response to your reviewers. I am by no means an expert in statistics, but I hope this helps.
Regards,
Chris Deuel, Ph.D

CONCERNS ABOUT MEETING ANOVA ASSUMPTIONS

Hi, John
Good to hear you from Biomech-L.
I had same situation previously on repeated measure studies. I hope these articles would give some help to you. I like Overall_Doyle 1994 article.
Normally according to your pilot study, we will get proposed statistical power. And using this result of statistical power, you need to search the value inside of table depending on your number of treatment. However, most of article only suggested one-way repeated measure situation. I have not seen two-way repeated measure case for sample size determination. Thus you have to guess the appropriate number of

treatment in your study.
It's not easy job.
There is alternative way in stat if your data collection was done.
Since ANOVA and repeated measure ANOVA used a model (assumption of normality for ANOVA and sphericity for repeated measures ANOVA), there is error in reporting when sample size was small. In case of huge violation of assumption, the results are meaningless sometimes. Because it is very hard to meet this assumption with small sample.
Thus some people suggested non-parametric stat method. Some non-parametric stat method is assumption-free method. And it's idea is same as paired t-test. Wilcoxon T test or Wilcoxon test is the non-parametric equivalent of paired t test.
I hope it would work for you.
Have a good weekend!!
YK - Young-Kwan Kim

Hi John,
My name is Tamar and I'm a Biostatistics PhD student. I got your email through Biomch-L.
There are probably people who can be more helpful, but in case you didn't hear from them---
There are a few problems with small samples. Foe once, you would need a pretty big effect to reach statistical significance.
Another important issue is that of asymptotic tests. Many statistical tests rely on asymptotic distributions and thus with a small sample size, the may not be valid. For example, the variance of the parameters estimates and their confidence intervals that are output in regular software packages are not reliable.

What I would recommend you to try is "exact methods"- usually permutation based. (You can look up this link for example for an explanation http://v8doc.sas.com/sashtml/stat/chap28/sect28.htm)
The basic idea in permutation tests are - you take the measurements you have and put aside your summary statistics. You permute the results between the groups or environments and have a new summary statistics. You repeat many times. At the end you compare your "real" statistics with the distribution of the permuted ones and see how likely it is to receive this statistics if the measurements were taken randomly.

Another question- did you compare the results of an experiment on a subject between two environments and then took the difference? or did you "ignore" the fact that you have the same subject doing both experiments in two environments? (because, depends on your design and question, I'm not sure it helps you in terms of type 1 and type 2 error rates that you have two or three experiments done by the same person in

different environments).

I hope I helped you somehow

- Tamar Sofer.

John,
It sounds to me that you are technically correct in your statistics; however, while a significant result by definition means your power is sufficient (for that variable), there is a danger in having too few samples. Consider if you only ran a single subject, found significance and claimed success. I am not up to date on guidelines that define the minimum number of samples to ensure your sample is representative. Maybe you could do an analysis where you add or remove samples and compare the increasing samples until they demonstrate convergence? Perhaps you could make the argument that you are in practice testing the population and not just a representative sample? I'm not sure if that would be easier to pass by a reviewer or not. Otherwise you might try the nonparametric methods. I'm not convinced that they would be more correct than your repeated measures ANOVA, but you may find reviewers are more accepting.

Good luck

Bryan Kirking

Dear Dr. Dewitt,

In my opinion, your argument is sound. With your sample size, you were able to determine that $p < 0.05$, which is a well accepted criteria for statistical significance. Certainly, with a larger sample size you might be able to measure smaller p value if the trend continues, but unless you think that $p < 0.05$ is insufficient, adding more subjects wouldn't help that concern. I also agree that a power analysis is important when you do not detect a statistically significant difference, in order to determine how confident you are that there is no difference. However, with a small sample size there is the concern as to whether the data is really normally distributed-- to address this concern, I recommend performing a non parametric test (e.g. Wilcoxon Signed Rank Test in place of or in addition to a paired t test).

In our research we also often have repeated measures designs. I find that people are very confused as to how we can detect small differences with large variations between subjects. Often I find it

useful to present the difference in a variable (calculated for each subject) due to the treatments-- in this case, the standard deviation bars shrink considerably and you can easily visualize the difference between treatments.

Thanks for your time.

Sincerely,

Lou DeFrate, Ph.D.

John - I take your example of a repeated measures design where each subject tests two or more treatments. If the small number of subjects is truly a random sample from a homogeneous population where the differences in the response measure between any two treatments is normally distributed and that all these differences have the same population variance, then yes, the t-test or multiple comparisons after ANOVA are all valid. The problem lies with the above assumptions - both with how the subjects are sampled and with respect to the distribution of the response variable. With very small sample sizes, moderate violations of these assumptions can lead to completely erroneous results. With larger sample sizes, ANOVA and t-tests are much more robust to such violations.

As a reviewer or reader of the journal, I would be extremely skeptical of trying to make inference about a population of astronauts from what you observed on 4 or 5 self-selected subjects (who are probably not even astronauts!). Also, all you are doing when you do ANOVA or t-tests is compare means - says nothing about variance or (for example) what percent of potential astronuats might be helped by this countermeasure. You could not hope to estimate this percentage with any relaibility with such small sample sizes.

You could point out these limitations in your manuscript - but whether the journal would accept this sort of disclaimer, I wouldn't know.

Al - Al Feiveson

I'll echo Al's comments, and provide my "2-cents" as well, as I've experienced similar struggles with small-n studies, and as a Biostatistician have had to really contemplate the appropriateness of hypothesis testing in general, and then specifically using traditional

t-test or ANOVA techniques. And I've taught a fair amount on this subject too, so forgive my verbose email!

One requirement that we all have with our small-n studies is to begin our inquiry into the data with a very deliberate, nearly obnoxious test of model assumptions. This is especially relevant because, as Al indicated, what we learned early in grad.-school about statistics, when we were all "newbies" to the math and art of it, was that "ANOVA is robust to violations of assumptions," but what we sometimes forget from those lessons is that this is only true for violations of SOME of the assumptions, and then only when there is sufficient n to be able to rely on the law of large numbers. We don't live in that world here at NASA most of the time, ao we MUST pay special attention to our assumption testing.

With repeated measures designs in the ANOVA context, that means NOT-ONLY that our data are normally distributed, but also that it meets the assumption of sphericity, and homogeneity of variance (if there are any between-subjects factors). These latter two are critical, and there are statistical tests to determine whether you've met the assumptions. With big-n studies, even if your data aren't normally distributed, studies have shown that as long as you meet the latter two assumptions, usually ANOVA performs adequately.. but again, we don't have big-n, so we can't rely on that being true for us. And, if you haven't met the assumptions, then you probably should be using other techniques, or at least considering alternative adjustments for violations, and/or data transformations.

In the instances where you DO MEET all of the statistical assumptions (possibly after some data transformations), then it would be beneficial in your publication efforts that you BEGIN your statistics/results section by clarifying what tests you have performed to test them, and that your data meet the assumptions. That way you convince the reviewers/readers that your data are appropriately analyzed with your techniques. Then proceed...

As for post-hoc tests with repeated measures, that's another area where researchers sometimes get confused, and a statistically savvy reviewer will pick up on it. The term "post-hoc" tests is typically reserve for comparisons between GROUPS, not between times/within groups, so I need to clarify that we're not talking about something like Tukey's post-hoc adjustments, but instead something like Bonferonni, or other flavors appropriate for within-group comparisons. There again, you're starting to add fuel to the fire for the skeptic, so you should be considerate of the skeptics and choose more CONSERVATIVE options for these tests. Bonferonni is a good choice because it's commonly used and known to be

conservative. Another approach might be to utilize a-priori contrasts, which hold alphs to .05 (or whatever) for k-1 comparisons, where k= the number of levels of your repeated measures factor. Sometimes researchers want to "make all pairwise comparisons" because they haven't clearly thought out what they REALLY want to do, so they avoid a-priori contrasts because it's limited to k-1 comparisons. This is an unfortunate situation, because often times we aren't truly interested in ALL pairwise comparisons, but instead a scientifically meaningful set of comparisons. For example, if you're interested in comparing pre-flight observations to multiple observations taken during and post flight, then simple effects contrasts comparing all values to pre would solve the problem without getting excessive on your Bonferonni adjustments for all possible pairs (do you really care if R+12 is different from R+14?). There are other commonly uses contrasts too... maybe you want to model the PATTERN of change to determine if it fits different polynomial functions (common in biological science). A few other choices exist, but my point is that maybe you can avoid some criticism by narrowing in on the comparisons that ARE important, and thus increase your potential for determining significance (because you haven't adjusted critical alpha so much), AND address your scientific inquiry more appropriately. Icing on the cake--less critical reviews.

Having said all of this... you might also consider newer statistical techniques commonly referred to as mixed-modeling, multi-level modeling, hierarchical modeling, or growth modeling. These techniques are fairly recent extension of ANOVA/Regression, and they are better capable of handling some of the data problems that we experience with repeated-measures ANOVA. They won't solve all of your post-hoc comparison problems, but they are generally more appropriate for all longitudinal research (big or small n). If your data meet the assumptions for these tests, then you're better off starting here instead of ANOVA. (FYI: I've seen NIH make references in their presentations to new investigators stating something to the effect that if you propose repeated measures ANOVA for longitudinal research instead of MLM, it's a flat out reason for rejection.) Remember... you need to meet assumptions of WHATEVER test you employ, so you might as well shoot for the best test first!

There are good applied text books out there on these techniques if you are interested. Software is a little tricky sometimes, but SAS does an excellent job if you're already a SAS user. STATA is equally good and easier to use if you're starting from scratch. R software is commonly used too but I've no experience with it. Avoid SPSS for these techniques...

Rob

_____
Robert J. Ploutz-Snyder, PhD

Hello
I am by far an expert in statistics however, like you I often work on small samples. I generally use non-parametric tests based on the idea that t-tests and ANOVAs usually require n>30 and certainly are based on the assumption of a normal distribution. Apparently tests of normality also require n>30 therefore in very small samples, it is not possible to test for normality.
I don't know it that will help you any...
Johanna
Johanna Robertson

Dr. Dewitt, I do not consider myself a stats expert, however I do teach basic stats to our orthopaedic residents and at College of Charleston where I am a professor in exercise science.
While I understand your concern regarding sample size I would encourage you to use a non-parametric test on your samples.
With a repeated measure use Friedman's and with 3 or more seperate groups used Kruskal Wallis.
While these are less powerful tests and actually ANOVA and t-tests are more robust this may reduce the referee concern and comments you are seeing on your reviews.
With paired samples used Mann Whitney and/or Wilcoxon depending whether the samples are paired or independent.
My contribution may be what you already know.
If not, hope this is helpful.
Regards.
Bill Barfield, Ph.D., FACSM

Pax!

In fact I was just calculating sample size effects for our own paper.
The required sample size depends entirely on the purpose of the study, desired strength of the results, estimated size of the effect, and other background assumptions. For instance if I want to establish a correlation between two variables with $p < 0.05$ and $r >= 27$ % then n about 50 is required. But if one eg wants to test whether an exercise program affects weight then 5 persons may suffice. In this case, if the alternative hypothesis is no change, then an observed weight decrease

for all 5 has a p = 1/2^5 = 1/32 (simple sign test). (Of course without a control we cannot conclude it was the program alone that resulted in the change ...). The point is to be able to show with probabilistic arguments that the probability that the result is due to *pure chance* is less than 1:20. Non-parametric tests (such as Wilcoxon) can be useful here since they do not require assumptions about the distributions. The sign test though does not reflect the size of the change in the example so one might use the variable z = x_before - x_after (or a scaled variant [x_before - x_after]/x_before if appropriate) instead. If the hypothesis is that no systematic changes has occurred, one may assume that z is normally distributed around 0, and use the sample variance Vx_before + Vx_after to estimate that of z. This finally lets us calculate p = Prob(Z < z) (prob that decrease is due to chance). One possibility today is to use computer simulations to calculate probabilities, they can make strong arguments showing *concretely* the odds against the result being due to chance. Ok i got to hurry to the office.

Regards Frank Borg

John Dewitt,
It is my understanding that ANOVA and t-test have an underlying normality assumption; with such a small sample size you cannot show that your samples are in fact normally distributed. In other words the a priori p value is meaningless if the assumptions of the testing tool are violated. I recommend a non-parametric tests like Mann-Whitney or Kruskal-Wallis tests, they are not as powerful but that is the trade off.

By the way, the rule of thumb given to me to test normality was a minimum sample size of 12. This was quickly followed by, "I have been sworn to secrecy as to from where that number came."

Cheers,
Rob Richards

Have you looked into the aspect of non-parametric v. parametric statistical results? That addresses 'small' sample sizes...I am not an expert but I had a similar issue in the past....good luck
Nicole Jacobs

Hello,
Just my humble opinion, but given that Type I error can still be committed, even with small samples, the reviewers concerns are valid.

A solution would be to assure your reviewers that you have tested your data for assumptions of normalcy (eg., skewness, variance), and given the repeated measures design, that sphericity is not an issue. Even with small sample sizes, Type I error is a potential risk because of the influence of skewness associated with a small cluster of observations at one end...also, with small samples, providing all your data in a table might alleviate concerns - allow the reader to see the variance, etc.

Thanks,

Daniel
Daniel Cipriani, PT, PhD

Dear John,
There are several problems with small samples:

1. Statistical power: you analysed this problem correctly, in my eyes.

2. The verification of normal distribution, with is one of the prerequisites to apply parametric tests, such as Student's "t" and ANOVA (to a lesser extent for the latter). The problem is that the tests that are available to test normality of distribution are not appropriate for small samples. Thus, a major criticism is the choice of statistical tests. Non-parametric tests, such as a Wilcoxon matched pairs test, would be more appropriate.

3. Power to generalise observations from the sample to the target population. Of course, if the sample is small, the chance is large that the sample is not representative enough of the population.

I hope that this is of use

With kindest regards

Veronique
_____
Prof. V. Feipel, PhD

John,

Just a quick note . . . I 'll try to add more later:

I use paired testing methods in my studies as well due to sample size/ cost constraints. In your case, I'm obviously not familiar with the specifics; however, two possibilities come to mind:

1 If your subjects are not randomly drawn from the target population, the results are biased regardless of sample size and/or the statistical outcome.

For Example: Measuring weight from a population of athletes would not represent the general US population. Assessing usability based on 4 20 year olds can not be extrapolated to older adults.

It seems obvious, however, I have seen several studies where internal subjects were used to assess product usability. The results, of course, did not match the general population becuase of the subjects' familiarity with the control design.

2. Is Normaility assumed? The complexity of the measure's distribution may not be captured with such a small sample size. If the distribution is sufficiently skewed, nonparametric techniques would be requred.

Scott A. Ziolek

Dear Dr. De Witt,

My understanding is that most traditional parametric statistical analyses (e.g., t-test) are based on the assumption that the data are normally distributed. When this assumption is true, the p-value, based on the area under the normal distribution bell curve, is meaningful. However, with only 4-10 data points, it is very difficult to assess if the data fit the assumption of normal distribution. In this case, using a traditional statistical analysis may be problematic, even though your results are "significant". Based on my previous conversation with some statisticians, they usually recommend having at least 30 data points in order to assess the distribution of the data, which could be very challenging in biomechanical research. Recently, there are some modern statistical analyses that doesn't require normal distribution assumption. You can try them to see if the results match the results of the paired-t test. If it does, then it may provide a strong support to convince the reviewers.

Ching

****

Liang-Ching Tsai, MS, PT

Part 3

Continuation of the summary of the replies to the question posed by John De Witt regarding statistical power are sample sizes. Part 3 includes responses in the categories of 'APPLICATION TO THE POPULATION', 'PROVIDE EFFECT SIZE CALCULATIONS WITH CONFIDENCE INTERVALS TO GIVE AN INDICATION OF THE MAGNITUDE OF THE DIFFERENCE', 'PERFORM A PRIORI POWER ANALYSIS TO JUSTIFY THE SMALL SAMPLE SIZE', and 'GENERAL COMMENTS'

SUMMARY OF RESPONSES

APPLICATION TO THE POPULATION - POTENTIAL THAT THE SMALL SAMPLE SIZE MAY NOT BE REPRESENTATIVE OF THE POPULATION THAT YOU WISH TO INFER YOUR RESULTS TO

Hello Dr. Dewitt,

I feel your pain! I also use repeated measure studies with limitations on how large I can get my sample size, and testing populations known to have high variability.

The smaller your sample size, the greater the chance that your sample may not be representative of the whole population. That can cause reviewers to be skeptical of your results even if they do reach statistical significances:
The first people to volunteer to participate tend to be very active, and highly motivated people (or that their parents are highly motivated if testing minors, etc.).
Individual differences and outliers can have a much stronger effect on the overall study results than in larger samples. You could get positive results, but ones that may not apply to the whole population depending on how representative your sample is.

For an extreme example, last I checked, there were 9 people known to have a complete loss of their sense of proprioception. One of them, Ian Waterman, actively participates in research. By testing Ian, we may guess that this condition affects people in their 20s but that it is possible for people to relearn how to control their movements through the use of vision alone. However, even though my sample is more than 10% of the entire population, the things we learn from studying a small sample may not apply to the rest of that population: most of the rest of that population developed this condition when they were 50 or more years of age, and Ian is the only one who has relearned how to walk. He is an outlier, his high level of functioning makes it less daunting for him to get to the labs for research studies.

In terms of planning studies, multi-site research can help get larger samples. Depending on the population, it may also help if you can get portable data collection equipment as many people are willing to let researchers come to their home rather than drive several hours for a study.

In terms of publishing, providing the observed power statistics when you send in your results can help address researchers concerns--statisticians are typically satisfied if power is above .7, though lower is acceptable in some fields.

Some journals will flat out not accept papers if the sample size is considered too small, so checking if the journal has published some papers in similar populations with small sample size before deciding where to submit the paper can help. The more focused the journal is on your particular research area, the more likely reviewers are to understand the inherent limitations on sample sizes, along with the benefits of your repeated measures design.

If you have provided information on group means & stds, consider providing data on each individual's performance. When reviewers see that 9 out of the 10 people in the sample benefited from a particular environment, it helps assure them that your significant results aren't just due to an outlier or two. Depending on what kind of measurements you are collecting, MANOVAs may be able to increase your power, especially if you have multiple dependent variables.

Many reviewers will simply be satisfied if you make certain to address the limits in sample size in the discussion, and suggest other researchers try to replicate your results before giving a lot of weight to the recommendations you suggest based on your results.
Others may prefer that the paper be presented as a 'pilot study'
rather than something more definitive, but would still be perfectly willing to publish that pilot study and others may still be perfectly willing to cite it.

Hope this helps,

--
Genna Mulvey, Ph.D.

John,

Excellent questions. I believe there is quite a problem in the scientific literature based on the inconsistent and incorrect use of statistical analysis techniques. Too many researchers were poorly trained in statistics and do not attempt to remedy that or try to keep up with advancements in the field. The push for many journals to have standards for statistical analysis and recent papers on statistical analysis support your questions and my concerns. I hope you had a chance to read the papers out earlier this year on statistical analysis by Will Hopkins et al. in MSSE (Feb 09) and by myself in Sports Biomechanics (March 09).

Here are my opinions on your questions:

Rejecting the null hypothesis does not negate all concerns about a small sample size. Remember, the size and quality of the sample (representative of the population) are of critical importance because the assumed purpose of statistical tests as decision makers for the effect of the treatment/independent variable on THE POPULATION. Too many modern scientific writers forget this fact and mix up the internal and external validity issues in writing up their reports. Often modern writers talk about the statistical test they do on the sample evidence, and unconsciously switch to external validity and talk about the results in general. Often they compound their mistake

of overgenearlizing from a small convenience sample to the population of similar subjects, to OVERgeneralizing to all subjects and musculoskeletal systems! To make matters worse, it doesn't matter what the statistical test says if there is just one error in experimental control or bias introduced from non-randomization in the sample.

My advice is to focus the discussion with the reviewer/readers on the justification for the sample size and limit the discussion of your results to your sample. All statistical tests, somewhere, are a subjective decision rule. You have to subjectively set the alpha level and expected difference/association subjectively. Focus on why you need the sample to be small. Unfortunately, there are not a lot of well-know sample size calculation formulae for the repeated measures designs you described.

In some cases, what looks like a small sample is actually almost the whole population. John, I believe that if you studied 10 astronauts in a study it would certainly represent a large percentage of astronauts on the planet. If the study is a preliminary exploration of an issue, it certainly makes sense to use a small sample and limit the explanation/discussion of the results to the sample and encourage further study in a larger sample or other populations to verify the results. I don't believe it is very effective to argue that biomechanical hard and costly to collect and calculate, so a sample size of 10 is common in biomechanics. I think you are just as likely to run into a small sample bias in reviewers from almost any journal (biomechanics or not).

Editor and reviewer bias/standards about sample sizes are hard to overcome, but if you focus on the statitical and research design issues and limit your discussion of the results you have put yourself in the best position to win the argument.

Some of the best work is often published in second tier journals because of bias an other injustice in so-call top tier journals. Remember that you will have the last laugh when your paper published in this less prestigious journal creates numerous citations and contributes to the decline in prestige of the journal with reviewers that did not focus on substantive issues.

_____

Duane Knudson, Ph.D.

Dear Dr. Dewitt,

I have encountered the same problem a few times recently.

Assuming sample size was the only thing that has been criticized, statistically speaking, in your manuscript, I would argue the way we understand the term "Significant difference".

Some reviewers suggest that a p-value represent whether your observed difference really exist or you saw the different by chance. A greater sample size will reduce the chance of this situation. Stats textbook tell us otherwise. A p-value tell us the chance that your observed difference represent the population where your data were sampled from. There is a difference in your samples if you see one. But you are risking a false positive or type I error. Meaning there may not be a difference in the population even you see difference in your sample. In your case, I sense that you have sampled the entire population! You are not even trying to use your data to represent a larger population, you have exhausted, or close to, the population. Therefore, there is a difference if you see one. Theoretically, you don't even need to run t-test or paired t-test, if that is the case.

Tell the reviewers that you have exhausted the population, if I am correct here.

"Statistics helps you generating an educated guess, if you don't know the truth." A friend of mine once said, "You don't need statistics if you know the truth."

Cheers,

Li
Li Li, Ph.D.

Dr. Dewitt,
Generally the comment on small sample size is that the results may not be generalizable. Certainly small sample size can decrease statistical power, but small sample sizes also tend to violate normality assumptions in the distribution, so the statistical inferences become less trustworthy. In particular, your sample mean values may not accurately represent population mean values. Therefore it is hard to infer if the population mean really changes during the 2 or 3 conditions in your experiment. Thus this is a type I error issue.

In terms of explaining the work, this seems to be a limitation that needs to be addressed, but not necessarily correctable. You could say that "This result needs to be confirmed/replicated in other similar experiments." A more statistically explicit way to deal with the issue may be the use of Bayesian inference, but this is not very widely used, and may be more difficult to explain than "the small sample size is a limitation."

I hope that helps.

Sincerely,
Hyun Gu Kang, PhD

John,

Please post all the replies you get so that we all can learn from them.
Obviously, if you find significance the effect was large enough. The concern is in the actual power of the sample size. One might surmise that the generalizability of the results are limited due to the small sample size. In other words, even though power appears to be sufficient since the null hypothesis was rejected, we may have a question of whether or not the sample studied accurately reflects the population under study (assumptions of your statistical procedures are normally distributed curves and with a small sample size it is difficult to assess whether or not the curve truly is normally distributed). If you look at your confidence intervals they may be rather wide even though you found significance. It would seem to me, and I am not a statistician, that ways around this might be to compare your confidence intervals with those of similarly done studies that have been published and use this data in your discussion when you address the small sample size as a limitation, explain clearly why you had to have such a small sample size and/or call the manuscript a pilot.

I'm not sure if this helps at all.

Regards,

Ken

PROVIDE EFFECT SIZE CALCULATIONS WITH CONFIDENCE INTERVALS TO GIVE AN INDICATION OF THE MAGNITUDE OF THE DIFFERENCE

Hi John,
If you look at where the statistical techniques that we base all of our "significant" results on came from, they all assume decisively normal / Gaussian distributions, which is impossible to verify with any confidence for any sample that isn't on the order of hundreds or thousands of samples at least, and the holy "$p < 0.05$" threshold is completely arbitrary and has no basis in physiology.

I'm not sure exactly how to address the issues you mentioned, but one statistic I like a lot is the effect size. It was developed by a guy named Cohen:

Cohen J (1990). Statistical Power Analysis for Behavioral Sciences.
New Jersey: Erlbaum.

The effect size (ES) is simply the ratio of the difference in group or condition means to the pooled standard deviation. Cohen provides some guidelines for interpreting ES as a measure of the "biological significance" of the results. For example if $ES > 0.8$, that implies a result with "strong" biological significance that would be even more significant at the p-value level if you had more subjects in the study. I like to use it as an argument against folks who complain about small sample sizes, even though I see that argument as trivial and pointless in our field since everyone's sample size is small.

Hope all is well,
Ross
Ross Miller [rosshm@mac.com]

PERFORM A PRIORI POWER ANALYSIS TO JUSTIFY THE SMALL SAMPLE SIZE

John,
It appears to me you are doing things right for a data analysis that is focused on p-value based interpretations. I assume you have adjusted the critical p value if you were doing multiple comparisons in the single paper. Other than that, my only suggestion is, if you are thinking about doing a post-hoc power test to prove to the reviewers that there was sufficient power even with your small sample size, then the following article arguing against that approach could be considered:
.Hoenig, J., & Heisey, D. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician, 55, 19-24

I would be interested in seeing other comments you get.

****

Gordon Chalmers, Ph.D.

Dear John,

Since it's a must to work with small sample sizes you are to do three things:

1- Make sure that no one in your field has done similar research using bigger sample sizes.

2- Report in your paper the motive and/or reasons to do so:
it should make sense if it is the nature of your research and all the researchers in your area have reported similar cases and have used small sample sizes as well.

3- When representing your results, report the power of your statistical analysis associated with the significance levels you obtained for tests. (you should be able to calculate this value easily using your statistical software package **SPSS, SAS, or Minitab**). Power of 80% or more should justify your results.

_____

Best Regards

_____

Tamer Khalaf, PhD

Dear John,

I am not an expert, but I would say you are largely right. As you say, the sample size you need depends on your effect size. If whatever you're looking at has a large effect, you only need a small sample. The classical reference here is the work by Jacob Cohen - for instance "Cohen J. A power primer. Psychological Bulletin 112(1992):155-159" which you can find on the web through Google.

Nevertheless, it would still be good (definitely for a paper but I suppose also for your application to an ethical committee?) to perform a power analysis. You can download free software (G*Power) to do that from

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3

Since you probably have a good idea of the effect size you try to find, and have good data on standard deviations, you should be able to do sample size calculations which you can include in future papers.

One thing where the reviewers might have a point is the generalizability of your findings. After all, with only four volunteers they might have some special characteristic? Therefore you might want to describe clearly in the methods section of your paper how you selected your volunteers, and in the discussion section you should try to point out why you think your research findings on four people are more widely valid.

Hope this helps and best regards,

Jan Herman

Hi John,
As an author I certainly understand your frustration, but as a peer reviewer this is one of the ways we gage how rigorous the science is. No doubt some reviewers are more picky than others, but there's probably room for compromise on both sides: it really is the author's job to demonstrate that their experiments have been designed to maximize the value of the results.

Small samples are a fact of life in our field, and your approach of using repeated measures designs is one way to deal with this. However, taking the extra step of doing a "power" or "sample size" analysis is not terribly onerous, and can help you determine how many reps/conditions you need with a given sample to acheive a certain power.

All you need is a computer program that does power calculations. There are many out there, but my favorite is G*Power, created at University of Duesseldorf. You can get it free at this website:
http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
Among the F-tests you will find various repeated measures designs to chose from.

Another suggestion, if you want to gain a better understanding of the "nuts & bolts"
behind power analysis is to get a copy of Cohen's "Statistical Power Analysis for the Behavioral Sciences" (Lawrence Earlbaum Associates, NJ). It's basically the bible of statistical power. It is also convenient that G*Power bases its calculations largely on Cohen's work, so having both at hand is very helpful.

Cheers,
Chris

Chris A. McGibbon, PhD

John,

All I can give you is my 2 cents. First, a little statistical knowledge is dangerous and many reviewers fancy themselves experts.

If you conducted pre-experimental sample size estimation for a power of 0.80, you are on fairly safe ground. You also also have to demonstrate that the data meet the assumptions of the statistical test (most people neglect this aspect of data analysis). A significant result then means something. I do not understand what the problem is.

If you conducted pre-experimental sample size estiamtion for sample size estimation at a power of 0.80, and the data meet the assumptions of the statistical test, and you fail to demonstrate significant differences, then there really was no significant differences. Post-experimental power analysis will always say the you need more subjects, and numerous papers exist that explain why post experimental power analysis is a fallacious form of data analysis. This is where most people get "dinged".

You will have to justify your results based on pre-experimental sample size estimation techniques and that the data meet the assumption of the statistical tests. If the reviewer continues to give you a hard time, they probably do not understand statistics enough to recognize the right answer and you are in trouble. Sorry ;-)

Best Wishes,

David. David Gabriel [dgabriel@brocku.ca]

Hi John,

I don't know exactly what the reviewers have written but I have gotten similar comments before. Interestingly enough, I received similar comments on computer simulations of N=1, which didn't make sense at all.

It sounds like that your statistics is solid. What you could do is to add a power analysis to your manuscript justifying the small sample size. While this wouldn't change anything on the subsequent analysis it would take the concerns off the reviewer's minds before they even think about questioning it.

Good luck,
Michael Liebschner - Michael Liebschner [mal1@bcm.tmc.edu]

Hi John,

Like I said, I even got the comment when I submitted a manuscript on a numerical simulation. Since the model will always return the same answer it really didn't make any sense to add another model.

You also made an interesting point that I sometimes question on work done by my colleagues. In your case your small sample size resulted in rejecting your null hypothesis, which is good. My collaborators sometimes use a small sample size because the experiment is cumbersome and time consuming.
However, they ended up having to accept the null hypothesis. When I asked them to do a power analysis with the data at hand to see if it would make sense to just add a couple more samples I just given the comment that their previous analyses justified the sample size. This brings out an interesting question, what is more important, your statistics or the actual research findings? In my opinion, statistics is only one way to explain that data and the actual data are more important. If you research data suggest that you should go back and include a few more samples to get statistical significance than you should have the obligation to do so.

Best wishes,
Michael - Michael Liebschner [mal1@bcm.tmc.edu]

John,
One suggestion might be to perform a brief pilot study prior to engaging in the full blown event. Using your pilot data you can calculate an effect size and then using some statistical software (i.e. G*Power) you could calculate a

sample size appropriate for the effect size at a given level of power.

In your manuscripts, in order to limit the discussion on whether or not your sample size is sufficient you could discuss the findings of the pilot study and your sample size determination/power analysis.

Hope this helps.

Jason Scibek, PhD, ATC

Dr. De Witt,
You may be able to respond by performing an a priori power analysis with something like GPower software that will tell you the number of subjects required for the study based on pilot or others' work. I believe this will certainly improve your study and the significance of the results, rather than choosing what might be considered by the statistically bent reviewers to be a random number of collected subjects. I would certainly be interested to hear what others on the listserv say though, and most importantly how the reviewers have responded to your pleas.

Sincerely,

Toran MacLeod

I ran into the same problem in trying to get my dissertation approved through my college dean. It wasn't until I ran a power analysis and showed that I could have sufficient power in a repeated measures design with 8 participants that he signed off on the prospectus. You often can't go back and collect data on more individuals, but if you include an a priori power analysis in your methods section, that may answer the reviewer's concerns.
Gary Christopher

GENERAL COMMENTS

Dear John,
I work in the field of sport psychology and movement science and often have the same problem.
Recently we started to use a priori calculations on sample size, power, etc. using the GPOWER Software. It can be used for free.
You can find it here: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
It is used a priori, for instance to calculate sample size, given alpha, power and effect size.
If is used post-hoc to calculate achieved power, given alpha, sample size and effect size.
If you can estimate effect sizes from current research and use a priori calculation to justify your design, this may help in your argumentation. Even if you calculate achieved power in case you have an effect or not, this will strengthen your argumentation and get rid of any subjective judgment on sample sizes. A further resource is: Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NY: Lawrence Erlbaum Associates...but maybe you already know this.
Bests,
Tom. - Heinen, Thomas [T.Heinen@dshs-koeln.de]

Hi john, I too have the same issue. I have done a series of longitudinal studies on the motion mechanics of pregnant women. They were tested 5 times. I also have a control group so I have a really good understanding of what is a change due to pregnancy and what is repeat testing. Getting pregnant women into a lab when they A) have morning sickness and B) likely to have not even told their friend and family they are pregnant because they are less than 12 weeks is really really hard. I still get papers knocked back because of small n.(n=9 maternal in this case). My frustration is mounting. Given the very small numbers of papers on gait of pregnant women in existence I find it astoundinsg that the papers keep get being knocked back because of sample size. Perhaps it is no wonder there is a small number of published studies. There may well be several studies out there that have never managed to get past this issue with the journals.
Please put your summary of replies in biomech - l as i think there is a probably a number of us who are working with small n for a very good reason but it is still valuable information to get out.
Wendy - Wendy Gilleard [wendy.gilleard@scu.edu.au]

Hi John,

Unfortunately this is a very 'easy' and common criticism of studies: that the sample size is too small. Often, I feel, the criticism is made without much thought. In my opinion, there are three essential issues regarding sample size selection:

1) The nature of the population. For young, healthy, heterogeneous individuals, a sample size of around 10 seems fine for most biomechanics studies; some would say that once the sample size in the double digits it's acceptable! For a clinical/patient population, a larger sample size would be required to be able to generalize the results to that clinical population.

2) Study design. If your study involves any kind of between-group comparison then more than 10 would be needed (depending on the anticipated effect size), but it doesn't sound like this is the case for your research.

3) Available subject pool. It sounds like this is a limitation for you.
This is often also a limitation in clinical research where there are just not a lot of patients available that meet certain criteria. As long as this limitation is explained in the paper, it should be acceptable.

Finally, as long as you justify your sample size, I believe it would be harsh to allow a seemingly small sample to prevent a paper from being published.

I hope this helps!

Avril - Avril Mansfield [avril.mansfield@utoronto.ca]

Hi John,

I believe the rule of thumb for comparing the means of two groups is that you need 16 per group to detect an effect size of 1 (which is actually rather large). While not a formal reference, I believe this link helps to explain that concept: http://www.childrensmercy.org/stats/size/quick.asp.
For my own a priori calculations, I usually refer to Chow's "Sample Size Calculations in Clinical Research" (http://www.amazon.com/Sample-Calculations-Clinical-Research-Biostatistics/dp/0824709705). This book contains formulae for some common study designs in clinical research. If that doesn't contain the design I'm concerned about, I will try searching statistics journals for relevant articles.

In reading over my response, I noted a typo. Of course, I meant with a young, healthy, *homogeneous* population, a sample of ~10 is sufficient.

I look forward to reading the summary of responses.

Avril - Avril Mansfield [avril.mansfield@utoronto.ca]

Small sample size is indeed a concern when using standard statistical measures. Ty using Effect Size statistics and magnitude based inferences, this will overcome your problem.
There are recent publications by Batterham and W. Hopkins (one in MSSE and one in IJSPP) on this issue. Also look at Will Hopkins site sportscience.org for further reading and helpful spreadsheets.

Kind regards,

Paul Montgomery

Hey John,

the described problem seems to be very common in many research areas that only have a small sample size available. To understand the concerns the reviewers have, it is necessary to clarify the meaning of "significance" in this context.

First of all a statistical analysis normally is performed to estimate an effect, observed in a representative sample and to transfer that observation to a bigger population. For this reason, the experimental sample should fulfil some conditions related to the research question / hypothesis.

To estimate an effect from a sample, it should be normally distributed, have a certain standard deviation etc... If you only want to draw conclusions to a very small population that presumably has a small standard deviation and a limited number of varying factors that (in the best case) could even be described or controlled, a very small sample size may be sufficient. In the discussion it needs to become clear, why this small sample size should be sufficient. - Some times, one should even ask the question, if a statistical analysis is needed / necessary in these cases.

If you want to draw conclusions for a bigger population that has more and some uncontrolled factors that may also effect the measured variables, you need a bigger sample size. The sample should have the same standard deviation that would be expected for the whole population. Normally this is only provided by a certain minimum

sample size. Otherwise the SD of the sample and of the population is not homogeneous.

In addition, the sample size depends on the number of factors you want / need to measure. The more factors you have, the bigger your sample size needs to be. In the best case, you only have one factor (all other conditions are constant / or very well controlled for all measurements).

Furthermore the sample size depends on the magnitude of the detectable contrast you want to measure. This again is directly related to the standard deviation of the sample or (as described above) of the presumable SD of the whole population.

A "significant" difference for a measured parameter depends on two facts.
One is the sample size, the other is the magnitude of the difference related to the standard deviation between the two samples for the parameter. - In a paired test the sample sizes should be the same (as every subject performs every test). If the sample size is very small, the standard deviation within one test may be very small as well. This needs not to represent the "real"
standard deviation you might get, if you are measuring more subjects or looking at the whole population. If now the difference between the first and the second test is somewhat bigger than the small standard deviation, you may see a "significant" difference between the tests. This may be true for the small sample size, but does not necessarily needs to be true for a bigger sample size or even a whole population.

This leads back to the introduction. The sample size and subject selection must be related to the research question and the size of the population for what the estimations / conclusions should be made.
Dr. Lars Janshen [lars.janshen@hu-berlin.de]

John,

My guess is the reviewers are concerned that you have a Type I error, rather than Type II:
Null hypothesis was true and you said it wasn't.
The findings are outside your alpha level (extremes).
Problem is you may lead others astray to dead end.

This could still happen in the case you have suggested. One of the ways to decrease the reviewers' concerns would be to provide them with an effect size (basically how much different are the conditions). SPSS and some other programs will calculate this for you when you select the appropriate option; however it is not available under T-tests, only with ANOVA's. This isn't to say you can't calculate it by hand after running a T-test. Then you can say whether the effect is small medium or large.

Effect sizes using partial eta2 (ηp2) were also obtained for each dependent variable using the formula: ηp2 = SSeffect / (SSeffect - SSerror), where SSeffect = effect variance and SSerror = error variance. Interpretation of effect size was done using a scale for effect size classification based on F-values for effect size and were converted to ηp2 using the formula: F = (ηp2 / (1 - ηp2))0.5. Consequently, the scale for classification of ηp2 was: 0.04 = trivial, 0.041 to 0.249 = small, 0.25 to 0.549 = medium, 0.55 to 0.799 = large, and .0.8 = very large [Comyns TM, Harrison AJ, Hennessy L, Jensen RL. The determination of the optimal complex training resistive load in male rugby players. Sport Biomech 6: 59-70, 2007]. I hope the equations come out...

Essentially it gives you an indication of how different is different, and not just the probability of difference. This is because you could find a significant difference, but it doesn't really mean much because the values are still so close together. Hope that helps.

Randall Jensen, PhD, FACSM, CSCS

Small sample size might be just an outlier.. even if you reject , you might be just evaluating the outliers. Although it shows that you didnt do type II error, it doesnt justify that you did it right either. You might be just working on outliers of a big sample. still doing the right but only for outliers as your sample is not big enough.
Hope this gives a different point of view to your understanding.
Regards,
Senay Mihcin

John De Witt, Ph.D., C.S.C.S.
Exercise Physiology Laboratory Lead / Biomechanist
Exercise Physiology Laboratory
NASA - Johnson Space Center
john.k.dewitt@nasa.gov
281-483-8939 / 281-483-4181 (fax)