n

Thomas R. Knapp
2012
©

Preface

As the title suggests, this paper is about the number of "things" (n) that are appropriate for different aspects of quantitative research, especially in education and in nursing (the two fields I know well). The topics range from "What size sample should you have?" to "How many statistical inferences should you make for a given study? Each section (there are 15 of them) starts out with a question, an answer is provided, and then one or more reasons are given in defense of that answer. Most of those answers and reasons are based upon the best that there is in the methodological research literature; I have provided two references for each section. You might find some of them to be old, but old is not necessarily bad and is often very good. [I've also sneaked in some of my own personal opinions about the appropriate n for various situations.]

I could have entitled this paper "N" rather than "n". Statisticians generally prefer to use "N" for the number of things in an entire population and "n" for the number of things in a sample drawn from a population. For the purpose of this paper I prefer "n" throughout.

Enjoy!

Table of Contents

Page Question

Section I

Question: What size sample should you have?

Answer: It depends

Why?

There are many matters to take into consideration. For example,

1. how many observations you can afford to take with the resources you have.
2. whether you only want to summarize the data in hand or whether you want to make an inference from a sample to a population from which the sample was drawn.
3. how far off you can afford to be IF you want to make a sample-to-population inference.
4. the degree of homogeneity in the population from which the sample was drawn.
5. the reliability and validity of your measuring instrument(s).

Let's take those matters one at a time.

1. Suppose you wanted to determine the relationship between height and weight for the population of third graders in your classroom. [Yes, that is a perfectly defensible population.] Even struggling school districts and financially-strapped school principals can probably afford to purchase a stadiometer/scale combination that is found in just about every doctor's office. (There might already be one in the school nurse's office.) If you have a class of, say, 20 students, and you have a couple of hours to devote to the project, you can have each pupil stand on the scale, lower the rod at the top of the scale onto his(her) head, read off his(her) height and weight, and write them down on a piece of paper. Do that for each pupil, make a scatter diagram for the data, and calculate the Pearson product-moment correlation coefficient, using one of the many formulas for it (see Section XI of this monograph) or have some sort of computer software do the plotting and the calculating for you. There. End of project. Not very costly. Also of limited interest to anyone other than yourself, those pupils, and perhaps their parents, but that's a separate matter from the determination of the sample size, which in this case is the size of the population itself.

[An aside: Do you think you should make two separate scatter diagrams: one for the boys (suppose there are 10 of them) and one for the girls (suppose there are 10 of them also)? Or would one that includes the data for both boys and girls suffice? Think about that.]

But if you wanted to determine that relationship for the entire school, or the entire school district, or the entire state, etc., unless you had a huge grant of some sort the cost of carrying out the study would undoubtedly be beyond your financial capability, so you would likely resort to sampling (see next paragraph).

2. Rather than using all of the pupils in your third-grade classroom you might want to take a random sample of your elementary school, determine the relationship between height and weight for that sample, and infer that the relationship for the entire school should be about the same, plus or minus some amount. The size of that sample would depend upon how accurate you would like that inference to be (see next paragraph).

3. The accuracy of a sample-to-population inference depends upon the so-called "margin of error", which in turn depends upon the sample size. If you want to be very accurate, you must draw a very large sample. If you don't mind being far off, a very small sample size would suffice, even as few as three people. If you choose only one pupil you can't calculate a Pearson r. If you choose two pupils the correlation must be +1, -1, or indeterminate. [Do you know why?] If you were interested in estimating some other quantity, say the mean height in a population, you might even be able to get away with n = 1 (see next paragraph).

4. If you happen to know that everyone in the population is exactly alike with respect to a particular characteristic, e.g., age in years at time of entering kindergarten, and that is a characteristic in which you are interested, it would not only be possible to estimate its mean with an n of 1 but it would be wasteful of resources to take any larger sample size than 1. (There is a whole sub-specialty in psychological research called single case analysis where n is always equal to 1, but various analyses are carried out within person.)

5. The more reliable the measuring instrument, the smaller the margin of error, all other things being equal. And the more valid the measuring instrument, the more you are convinced that you're measuring the right quantity.

For a discussion of the justification for sample size in general, see Hayat (in press). For a discussion of "sample size" (number of occasions) for single case analysis, see Kratochwill and Levin (2010).

Section II

Question: How many observations in a nested design are comparable to how many observations in a non-nested design?

Answer: This many: $mk/[1 + \rho(m-1)]$, where m is the number of observations within cluster, k is the number of clusters, and ρ is the within-cluster correlation.

Why?

It all has to do with the independence (or lack of same) of the observations. In a nested design, e.g., one in which participants are "run" in clusters, the observations for participants within each cluster are likely to be more similar to one another than the observations for participants in different clusters. Using the above formula for a simple case of two clusters with seven observations within each cluster and a within-cluster correlation (a measure of the amount of dependence among the observations) of .50, we get 9.33 (call it 9). That is the "effective sample size" for the nested design, compared to a larger sample size of 2(7)= 14 if the observations for all of the participants had been independent in a non-nested design. If you would carry out a traditional significance test or construct a traditional confidence interval for those data using the n of 14, you would be under-estimating the amount of sampling error and therefore be more likely to make a Type I error or have a confidence interval that is too tight.

For discussions of the general problem and for further examples, see Killip, Mahfoud, and Pearce (2004) and Knapp (2007a).

Section III

Question: How many treatment groups should you have in a true experiment (randomized controlled trial)?

Answer: Usually just two.

Why?

The focus of most true experiments is on the difference in the effectiveness of two treatments: one experimental, one control. Some people argue that you should have at least three groups: one experimental group, another experimental group, and a placebo group (no treatment at all); or that you should have a factorial design for which the effects of two or more variables could be tested in a single study (see following paragraph). Having three treatment conditions rather than two can get complicated in several respects. First of all, three treatments are at least 1.5 times harder to manage than two treatments. Secondly, the analysis is both more difficult and rather controversial. [Do you have to conduct an over-all comparison and then one or more two-group comparisons? If so, which groups do you compare against which other groups? How do you interpret the results? Do you have to correct for multiple comparisons?] Third, what should the placebo group get(do) while the two experimental groups are receiving their respective treatments?

The article by Green, Liu, and O'Sullivan (2002) nicely summarizes the problems entailed in using factorial designs, but see Freidlin, Korn, Gray, et al. (2008) for a counter-argument to theirs.

Section IV

Question: How many points should a Likert-type attitude scale have?

Answer: It doesn't really matter.

Why?

Several researchers have studied this problem. The consensus of their findings (see, for example, Matell & Jacoby, 1971) is that neither the reliability nor the validity of a measuring instrument is affected very much if you have two (the minimum), five (the usual; also the number that Rensis Likert used in his original 1932 article in the Archives of Psychology), or whatever number of categories for the scale. That is actually counter-intuitive, because you would expect the greater the number of categories, the more sensitive the scale. The more important consideration is the verbal equivalent associated with each of the points. For example, is "sometimes" more often or less often than "occasionally"?

It also doesn't seem to matter whether the number of scale points is odd or even. Some investigators like to use an odd number of scale points so that a respondent can make a neutral choice in the middle of the scale. Others object to that, insisting that each respondent should make a choice on either the agree side or the disagree side, and not be permitted to "cop out".

It is the analysis of the data for Likert-type scales that is the most controversial. Some people (like me) claim that you should not use means, standard deviations, and Pearson r's for such scales. Others see nothing wrong in so doing. The latter position is clearly the predominant one in both the education and the nursing literature. But see Marcus-Roberts and Roberts (1987) for a convincing argument that the predominant position is wrong.

Section V

Question: How many items should you have on a cognitive test?

Answer: As many as practically feasible.

Why?

The reliability of a test is positively related to the number of items (the greater the number of items, the higher the reliability, all else being equal, especially the validity of the instrument). And if your test is multiple-choice, that same claim also holds for the number of options per item (the greater the number of options, the higher the reliability). Ebel (1969, 1972) explains both of these matters nicely. Common sense, however, dictates that you can't have huge numbers of items or choices, because that would lead to fatigue, boredom, cheating, or resistance on the part of the test-takers.

The standard error of measurement, which is a scale-bound indicator of the reliability of a measuring instrument for an individual person, can actually be approximated by using the formula .43 times the square root of n, where n is the number of items on the test. (You can look it up, as Casey Stengel used to say.) It is then possible to establish a confidence interval around his(her) obtained score and get some approximate idea of what his(her) "true score" is. The "true score" is what the person would have gotten if the test were perfectly reliable, i.e., what he(she) "deserved" to get.

Section VI

Question: How many equivalent forms should you have for a measuring instrument?

Answer: At least two (for "paper-and-pencil" tests).

Why?

If you want to determine the reliability of a measuring instrument, it is best to have more than one form so you can see if the instrument is consistent in the scores that it produces. You could administer a single form on two separate occasions and compare the results, but participants might "parrot back" all or many of the same responses on the two occasions, leading to an over-estimate of its "real" reliability. Most researchers don't administer two forms once each OR the same form twice. They just administer one form once and determine the item-to-item internal consistency (usually by Cronbach's alpha), but that is more an indicator of an instrument's homogeneity or the coherence of its items than its reliability (see Kelley, 1942).

The well-known Scholastic Aptitude Test (SAT) has more than two equivalent forms, mainly because the test has been developed to sample a large amount of content and because the developers are so concerned about cheating. The various forms are subject to severe security precautions.

If the instrument is a physical measuring instrument such as a stadiometer (for measuring height) there is no need for two or more equivalent forms (the stadiometer can't "parrot back" a height). But you must have two or more forms, by definition, if you're carrying out a method-comparison study. (See Bland & Altman, 2010 for several examples of that kind of study, which is "sort of" like reliability but the instruments are not equivalent or parallel, in the usual sense of those terms.)

Section VII

Question: How many judges should you have for an inter-rater reliability study?

Answer: One or more

Why?

It depends upon whether you're interested in within-judge agreement or between-judges agreement. If the former (which is actually referred to as intra-rater reliability) you might only care about the consistency of ratings given by a particular judge, in which case he(she) would have to rate the same individuals on at least two occasions. It is also necessary to specify whether it is the determination of the degree of absolute agreement which is of concern or whether the determination of the degree of relative agreement is sufficient. (If there are four persons to be rated and the judge gives them ratings of 1,3,5, and 7 at Time 1 and gives them ratings of 2,4,6, and 8, respectively, at Time 2, the absolute agreement is zero but the relative agreement is perfect.)

If between-judges agreement is of primary concern, the seriousness of the rating situation should determine whether there are only two judges or more than two. It is usually more "fair" to determine the consensus of ratings made by several judges rather than just the average ratings for two judges, but it is a heck of a lot more work!

Two sources for interesting discussions of the number of judges and how to determine their agreement are Stemler (2004) and LeBreton and Senter (2008). LeBreton and Senter use a 20-questions format (not unlike the format used in this monograph).

Section VIII

Question: How many factors are interpretable in a factor analysis?

Answer: As many as there are eigenvalues greater than or equal to one for the correlation matrix.

Why?

If you have n variables, the "worst" (most heterogeneous solution) that can happen is that all of the eigenvalues of the correlation matrix are equal to one; i.e., each variable is its own factor. In the opposite extreme, the "best" (most homogeneous solution) that can happen is that there is just one big eigenvalue equal to n and therefore indicative of a unidimensional construct. (See Kaiser, 1960 for his "little jiffy" approach.) Although he has some concerns about the eigenvalues-greater-than-one "rule", Cliff (1988) provides a nice summary of its justification.

[If you don't know what an eigenvalue (sometimes called a latent root or a characteristic root) is, you can find out by reading the two sources cited in the preceding paragraph or by looking it up in a multivariate statistics textbook.]

The answer holds for both exploratory factor analysis and confirmatory factor analysis. [I've never understood the need for confirmatory factor analysis. Why bother to hypothesize how many factors there are and then find out how smart you are? Why not just carry out an exploratory factor analysis and find out how many there are?]

Section IX

Question: What is the minimum number of observations that will produce a unique mode for a variable?

Answer: Three

Why?

Suppose you have one observation, a. If so, you don't have a variable. Suppose you have two observations, a and b. If they're both the same, you again don't have a variable. If they're different, neither can be the mode. Now suppose you have three observations. If they're all the same, no variable. If they're all different, no mode. If two of them, say a and b, are the same but the third, c, is different from either of them, then a = b= the mode.

The mode doesn't come up very often, but it should. A manufacturer of men's overalls can only produce a small number of different sizes of overalls, and if optimization of profits is of primary concern (what else?) the modal size of men in general is what is important to know, so that more overalls of that size are produced than any other. A more scientific example arises in the case of something like blood type. Lacking specific information of the blood types of people in a given community, the American Red Cross would probably want to keep a greater supply of the modal blood type (which is O positive) than any other. Note that for blood type neither the mean nor the median would be relevant (or able to be computed), because the measurement of blood type employs a nominal scale.

There is an excellent Australian source on the internet called Statistics S1.1: Working with data, which is accessible free of charge and which is the finest source I've ever seen for understanding what a mode is and how it differs from a median and a mean. (It is intended for school children and uses British currency and British spelling, but I hope that doesn't bother you.) And if you ever get interested in a set of data that has two modes (bimodality), I wrote a very long article about that (Knapp, 2007b).

Section X

Question: How many categories should you have in a two-way contingency table ("cross-tab")?

Answer: However many it takes (but be careful).

Why?

Two-way contingency tables are used to investigate the relationship between two nominal variables, such as sex and political affiliation, race and blood type, and the like. For the sex-by-political affiliation example, sex requires two categories (male and female) and political affiliation requires at least two (Democrat and Republican) and as many as six or more (if you include Independent, Libertarian, Green, None, and others). For race you might have the usual four (Caucasian, African-American, Hispanic, Asian-American) or even more, if you need to add Native-American and others. Blood type requires eight (A positive, A negative, B positive, B negative, AB positive, AB negative, O positive, and O negative). But if you have a relatively small sample you might have to "collapse" a, say, 4x8 table, into a smaller table, and depending upon how you do the collapsing you can get quite different answers for the relationship between the row variable and the column variable. Collapsing of categories is all too common in quantitative research, and should only be used when absolutely necessary. (See Cohen, 1983, and Owen & Froman, 2005.)

Section XI

Question: How many ways are there to calculate a Pearson product-moment correlation coefficient?

Answer: At least 14

Why?

In a classic article several years ago, my friend Joe Rodgers and his colleague Alan Nicewander (1988) showed there were 13 mathematically equivalent ways. I subsequently wrote to Joe and told him about a 14th way (Knapp, 1979). There might be even more.

It of course doesn't really matter which formula is used, as long as it's one of the 14 and the calculations are carried out properly, by computer program, by hand, or whatever. Far more important is whether a Pearson r is appropriate in the first place. It is strictly an indicator of the direction and degree of LINEAR relationship between two variables. Researchers should always plot the data before carrying out the calculation. If the plot "looks" linear or if the data "pass" a test of linearity, fine. If not, a data transformation is called for.

Section XII

Question: How many significance tests should be carried out for baseline data in a true experiment (randomized controlled trial)?

Answer: None.

Why?

If the participants have been randomly assigned to treatment conditions, there is no need for testing baseline differences. (See Senn, 1994, and Assmann, Pocock, Enos, & Kasten, 2000.) The significance test or the confidence interval takes into account any differences that might have occurred by chance. And there are at least two additional problems: (1) how do you determine what variables on which such tests should be performed?; and (2) what do you do if you find a statistically significant difference for a particular variable? [Use it as a covariate? That's bad science. Covariates should be chosen based upon theoretical expectations and before seeing any data.]

Section XIII

Question: How many independent (predictor) variables should you have in a multiple regression analysis?

Answer: Since it's "multiple" you need at least two, but don't have too many.

Why?

Some sources suggest a "rule of thumb" of a 10:1 ratio of number of independent variables to number of participants, but it all depends upon how good a "fit" you require. If you choose to base the determination of the number of independent variables upon a power analysis, you could use Cohen's (1992) table "in reverse"; i.e., you specify the level of significance, the desired power, and your sample size; and then read off the number of independent variables that would give you the power you want.

Knapp and Campbell-Heider (1989) provide a summary of the various guidelines that have been promulgated for the number of participants vs. the number of variables for a variety of multivariate analyses, including multiple regression analysis.

Section XIV

Question: How many degrees of freedom are there for a given statistic and its sampling distribution?

Answer: It's usually something minus one.

Why?

The concept of number of degrees of freedom (df) is probably the most mysterious concept in all of statistics. Its definitions range all the way from "the number of unconstrained observations" to "something you need to know in order to use a table in the back of a statistics book". Let's take a couple of examples:

1. For a sample mean, if you know what all of the observations are except for one, and you also know what the mean is, then that one observation must be such that the sum of it and the others is equal to n times the mean, where n is the total number of observations. Therefore, the number of degrees of freedom associated with a sample mean is n-1. And if you use a table of the t sampling distribution to construct a confidence interval for the unknown population mean you find the value of t for n-1 degrees of freedom that you should lay off on the high side and on the low side of the sample mean.

2. For a 2x2 contingency table ("cross-tab"), if you want to test the significance of the difference between two independent proportions or percentages, if you know one of the cell frequencies, and you know the two row (r) totals and the two column (c) totals, the other three cell frequencies are not free to vary, so there is only one degree of freedom associated with that table, calculated by multiplying (r-1) by (c-1), which in this case = (2-1)x(2-1) = 1X 1 =1. You then calculate the value of chi-square by using the traditional formula and refer that value to a table of the chi-square sampling distribution for df=1 to find out if your result is or is not statistically significant with respect to your chosen alpha level.

Statistical consultants occasionally tell clients that "you lost one degree of freedom for this" or "you lost two degrees of freedom for that", which usually conveys nothing to the clients, since the clients don't know they have any to start with!

For all you ever need to know (and then some) about degrees of freedom, see the article by Helen Walker (1940...yes, 1940), but see also Ron Dotsch's Degrees of Freedom Tutorial (accessible free of charge on the internet) if you are particularly interested in the concept as it applies to the analysis of variance and the associated F tests.

Section XV

Question: How many statistical inferences should you make for a given study?

Answer: No more than one.

Why?

If you have data for an entire population (no matter what its size) or if you have data for a "convenience", non-random sample, no statistical inferences are warranted. If you do have data for a random sample and you want to make an inference from the sample to the population, one hypothesis test or one confidence interval (but not both, please) is fine. If you have a complicated study involving several groups and/or several variables, and if you carry out more than one statistical inference, it is usually incumbent upon you to make some sort of statistical "correction" (e.g., Bonferroni) or your error probabilities are greater than you assume a priori. Best to avoid the problem entirely by concentrating on a single parameter and its sample estimate.

If you are familiar with the research literature in whatever your particular discipline might be, you might find some of the claims in the previous paragraph to be unreasonable. Just about every research report is loaded with more than one p-value and/or more than one confidence interval, isn't it, no matter whether for a population, a convenience sample, or a random sample? Yes, that's true, but it doesn't make it right.

For opposing views on the matter of adjusting the alpha level when making more than one significance test, see the article by O'Keefe (2003) and the response by Hewes (2003). Neither of them argues for having no more than one statistical inference per study, but I do.

References

Assmann, S., Pocock, S.J., Enos, L.E., & Kasten, L.E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet, 355 (9209), 1064-1069.

Bland, J.M., & Altman, D.G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. International Journal of Nursing Studies, 47, 931–936.

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. Psychological Bulletin, 103, 276-279.

Cohen, J. (1983). The cost of dichotomization. Applied Psychological Measurement, 7, 249-253.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112 (1), 155-159.

Dotsch, R. (n.d.) Degrees of Freedom Tutorial. Accessible on the internet.

Ebel, R.L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Ebel, R.L. (1972). Why a longer test is usually more reliable. Educational and Psychological Measurement, 32, 249-253.

Freidlin, B., Korn, E.L., Gray, T., et al. (2008). Multi-arm clinical trials of new agents: Some design considerations. Clinical Cancer Research, 14, 4368-4371.

Green, S., Liu, P-Y, & O'Sullivan, J. (2002). Factorial design considerations. Journal of Clinical Oncology, 20, 3424-3430.

Hayat, M.J. (in press). Understanding sample size determination in nursing research. Western Journal of Nursing Research.

Hewes, D.E. (2003). Methods as tools. Human Communication Research, 29 (3), 448-454.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

Kelley, T.L. (1942). The reliability coefficient. Psychometrika, 7 (2), 75-83.

Killip, S., Mahfoud, Z., & Pearce, K. (2004) What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Annals of Family Medicine, 2, 204-208.

Knapp, T.R. (1979). Using incidence sampling to estimate covariances. Journal of Educational Statistics, 4, 41-58.

Knapp, T.R. (2007a). Effective sample size: A crucial concept. In S.S. Sawilowsky (Ed.), Real data analysis (Chapter 2, pp. 21-29). Charlotte, NC: Information Age Publishing.

Knapp, T.R. (2007b). Bimodality revisited. Journal of Modern Applied Statistical Methods, 6 (1), 8-20.

Knapp, T.R., & Campbell-Heider, N. (1989). Numbers of observations and variables in multivariate analyses. Western Journal of Nursing Research, 11, 634-641.

LeBreton, J.M., & Senter, J.L. (2008) Answers to 20 questions about interrater reliability and interrater agreement. Organizational Research Methods 11, 815-852.

Kratochwill, T.R., & Levin, J.R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. Psychological Methods, 15 (2), 124-144.

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. Journal of Educational Statistics. 12, 383-394.

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert Scale items? Educational and Psychological Measurement, 31, 657-674.

O'Keefe, D.J. (2003). Against familywise alpha adjustment. Human Communication Research, 29 (3), 431-447.

Owen, S.V., & Froman, R.D. (2005). Why carve up your continuous data? Research in Nursing & Health, 28, 496-503.

Rodgers, J.L., & Nicewander, W.A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42 (1), 59-66.

Senn, S. (1994). Testing for baseline balance in clinical trials. Statistics in Medicine, 13, 1715-1726.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9 (4).

Statistics S 1.1. (n.d.) Working with data. Accessible on the internet.

Walker, H.W. (1940). Degrees of freedom. Journal of Educational Psychology, 31 (4), 253-269.