To pool or not to pool: That is the confusion

Thomas R. Knapp

Prologue

Isn't the English language strange? Consider the word "pool". I go swimming in a pool. I shoot pool at the local billiards parlor. I obtain the services of someone in the secretarial pool to type a manuscript for me. I participate in a pool to try to predict the winners of football games. I join a car pool to save on gasoline. You and I pool our resources.

And now here I am talking about whether or not to pool data?! With 26 letters in our alphabet I wouldn't think we'd need to use the word "pool" in so many different ways. (The Hawaiian alphabet has only 12 letters...the five vowels and seven consonants H,K,L,M,N,P,and W; they just string lots of the same letters together to make new words.)

What is the meaning of the term "pooling data"?

There are several situations in which the term "pooling data" arises. Here are most of them:

1. Pooling variances

Let's start with the most familiar context for pooling data (at least to students in introductory courses in statistics), viz., the pooling of sample variances in a t test of the significance of the difference between two independent sample means. The null hypothesis to be tested is that the means of two populations are equal (the populations from which the respective samples have been randomly sampled). We almost never know what the population variances are (if we did we'd undoubtedly also know what the populations means are, and there would be no need to test the hypothesis), but we often assume that they are equal, so we need to have some way of estimating from the sample data the variance that the two populations have in common. I won't bore you with the formula (you can look it up in almost any statistics texxtbook), but it involves, not surprisingly, the two sample variances and the two sample sizes. You should also test the "poolability" of the sample variances before doing the pooling, by using Bartlett's test or Levene's test, but almost nobody does; neither test has much power.

[Note: There is another t test for which you don't assume the population variances to be equal, and there's no pooling. It's variously called the Welch-

Satterthwaite test or the Behrens-Fisher test. It is the default t test in Minitab. If you want the pooled test you have to explicitly request it.]

2. Pooling within-group regression slopes

One of the assumptions for the appropriate use of the analysis of covariance (ANCOVA) for two independent samples is that the regression of Y (the dependent variable) on X (the covariate) is the same in the two populations that have been sampled. If a test of the significance of the difference between the two within-group slopes is "passed" (the null hypothesis of equality of slopes is not rejected), those sample slopes can be pooled together for the adjustment of the means on the dependent variable. If that test is "failed" (the null hypothesis of equality of slopes is rejected) the traditional ANCOVA is not appropriate and the Johnson-Neyman technique (Johnson & Neyman, 1936) must be used in its place.

3. Pooling raw data across two (or more) subgroups

This is the kind of pooling people often do without thinking through the ramifications. For example, suppose you were interested in the relationship between height and weight for adults, and you had a random sample of 50 males and a random sample of 50 females. Should you pool the data for the two sexes and calculate one correlation coefficient, or should you get two correlation coefficients (one for the males and one for the females)? Does it matter?

The answer to the first question is a resounding "no" to the pooling. The answer to the second question is a resounding "yes". Here's why. In almost every population of adults the males are both taller and heavier than the females, on the average. If you pool the data and create a scatter plot, it will be longer and skinnier than the scatterplots for the two sexes treated separately, thereby producing a spuriously high correlation between height and weight. Try it. You'll see what I mean. And read the section in David Howell's (2007) statistics textbook (page 265) regarding this problem. He provides an example of real data for a sample of 92 college students (57 males, 35 females) in which the correlation between height and weight is .60 for the males, .49 for the females, and .78 for the two sexes pooled together.

4. Pooling raw data across research sites

This is the kind of pooling that goes on all the time (often unnoticed) in randomized clinical trials. The typical researcher often runs into practical difficulties in obtaining a sufficient number of subjects at a single site and "pads" the sample size by gathering data from two or more sites. In the analysis he(she) almost never tests the treatment-by-site interaction, which might "be there" and would constrain the generalizability of the findings.

5. Pooling data across time

There is a subtle version of this kind of pooling and a not-so-subtle version. Researchers often want to combine data for various years or minutes or whatever, for each unit of analysis (a person, a school, a hospital, etc.), usually by averaging, in order to get a better indicator of a "typical" measurement. They (the researchers) usually explain why and how they do that, so that's the not-so-subtle version. The subtle version is less common but more dangerous. Here the mistake is occasionally made of treating the Time 2 data for the same people as though they were different people from the Time 1 people. The sample size accordingly looks to be larger than it is, and the "correlatedness" of the data at the two points in time is ignored, often to the detriment of a less sensitive analysis. (Compare, for example, data that should be treated using McNemar's test for correlated samples with data that are appropriately handled by the traditional chi-square test of the independence of two categorical variables.)

6. Pooling data across scale categories

This is commonly known as "collapsing" and is frequently done with Likert-type scales. Instead of distinguishing between those who say "strongly agree" from those who say "agree'", the data for those two scale points are combined into one over-all "agree" designation. Likewise for "strongly disagree" and "disagree". This can result in a loss of information, so it should be used as a last resort.

7. Pooling "scores" on different variables

There are two different ways that data can be pooled across variables. The first way is straightforward and easy. Suppose you were interested in the trend of average (mean) monthly temperatures for a particular year in a particular city. For some months you have temperatures in degrees Fahrenheit and for other months you have temperatures in degrees Celsius. (Why that might have happened is not relevant here.) No problem. You can convert the Celsius temperatures to Fahrenheit by the formula $F = (9/5)C + 32$; or you can convert the Fahrenheit temperatures to Celsius by using the formula $C = (5/9)(F - 32)$.

The second way is complicated and not easy. Suppose you were interested in determining the relationship between mathematical aptitude and mathematical achievement for the students in your particular secondary school, but some of the students had taken the Smith Aptitude Test and other students had taken the Jones Aptitude Test. The problem is to estimate what score on the Smith test is equivalent to what score on the Jones test. This problem can be at least approximately solved if there is a normative group of students who have taken both the Smith test and the Jones test, you have access to such data, and you have for each test the percentile equivalent to each raw score on each test. For each student in your school who took Smith you use this "equipercentile method" to estimate what he(she) "might have gotten" on Jones. Assign to him(her) the

Jones raw score equivalent to the percentile rank that such persons obtained on Smith. Got it? Whew!

8. Pooling data from the individual level to the group level

This is usually referred to as "data aggregation". Suppose you were interested in the relationship between secondary school teachers' numbers of years of experience and the mathematical achievement of their students. You can't use the individual student as the unit of analysis, because each student doesn't have a different teacher (except in certain tutoring or home-school situations). But you can, and should, pool the mathematical achievement scores across students in their respective classrooms in order to get the correlation between teacher years of experience and student mathematical achievement.

9. Pooling cross-sectional data to approximate panel data

Cross-sectional data are relatively easy to obtain. Panel (longitudinal) data are not. Why? The principal reason is that the latter requires that the same people are measured on each of the occasions of interest, and life is such that people often refuse to participate on every occasion or they are unable to participate on every occasion (some even die). And you might not even want to measure the same people time after time, because they might get bored with the task and just "parrot back" their responses, thereby artificially inflating the correlations between time points.

What has been suggested is to take a random sample of the population at Time 1, a different random sample at Time 2,...etc. and compare the findings across time. You lose the usual sensitivity provided by having repeated measurements on the same people, but you gain some practical advantages.

There is a more complicated approach called a cross-sectional-sequential design, whereby random samples are taken from two or more cohorts at various time points. Here is an example (see Table 1, below) taken from an article that Chris Kovach and I wrote several years ago (Kovach & Knapp, 1989, p. 26). You get data for five different ages (60, 62, 64, 66, and 68) for a three-year study (1988, 1990, 1992). Nice, huh?

TABLE 1 A Cross-Sectional-Sequential Design

COHORT					
1924			1988	1990	1992
1926		1988	1990	1992	
1928	1988	1990	1992		
AGE	60	62	64	66	68

Figures in the table are the years in which data would be collected for each cohort listed.

10. Pooling findings across similar studies

This very popular approach is technically called "meta-analysis" (the term is due to Glass, 1976), but it should be called "meta-synthesis" (some people do use that term), because it involves the combining of results, not the breaking-down of results. I facetiously refer to it occasionally as "a statistical review of related literature", because it has come to replace almost all narrative reviews in certain disciplines. I avoid it like the plague; it's much too hard to cope with the problems involved. For example, what studies (published only? published and unpublished?) do you include? How do you determine their "poolability"? What statistical analysis(es) do you employ in combining the results?

Summary

So, should you pool or not? Or, putting it somewhat differently, when should you pool and when should you not? The answer depends upon the following considerations, in approximately decreasing order of importance:

1. The research question(s). Some things are obvious. For example, if you are concerned with the question "What is the relationship between height and weight for adult females?" you wouldn't want to toss in any height&weight data for adult males. But you might want to pool the data for Black adult females with the data for White adult females, or the data for older adult females with the data for younger adult females. It would be best to test the poolability before you do so, but if your sample is a simple random sample drawn from a well-defined population of adult females you might not know or care who's Black and who's White. On the other hand, you might have to pool if you don't have an adequate number of both Blacks and Whites to warrant a separate analysis for each.

2. Sample size. Reference was made in the previous paragraph to the situation where there is an inadequate number of observations in each of two (or more) subgroups, which would usually necessitate pooling (hopefully poolable entities).

3. Convenience, common sense, necessity

In order to carry out an independent sample t test when you assume equal population variances, you must pool. If you want to pool across subgroups, be careful; you probably don't want to do so, as the height and weight example (see above) illustrates. When collapsing Likert-type scale categories you might not have enough raw frequencies (like none?) for each scale point, which would prompt you to want to pool. For data aggregation you pool data at a lower level to produce data at a higher level. And for meta-analysis you must pool; that's what meta-analysis is all about.

A final caution

Just as "acceptance" of a null hypothesis does not mean it is necessarily true, "acceptance" in a poolability test does not mean that poolability is necessarily justified.

References

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Howell, D.C. (2007). Statistical methods for psychology (6th ed.). Belmont, CA: Thomson.

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their applications to some educational problems. Statistical Research Memoirs, 1, 57-93.

Kovach, C.R., & Knapp, T.R. (1989). Age, cohort, and time-period confounds in research on aging. Journal of Gerontological Nursing, 15 (3), 11-15.