"Who can act? Critical assumptions at the foundations of statistical analysis"
Explaining differences among means – What can that mean?

Peter J. Taylor
Programs in Science, Technology & Values and Critical & Creative Thinking
University of Massachusetts, Boston, MA 02125, USA
617 287 7636; 617 287 7656 (fax); peter.taylor@umb.edu

Draft 1 July 10

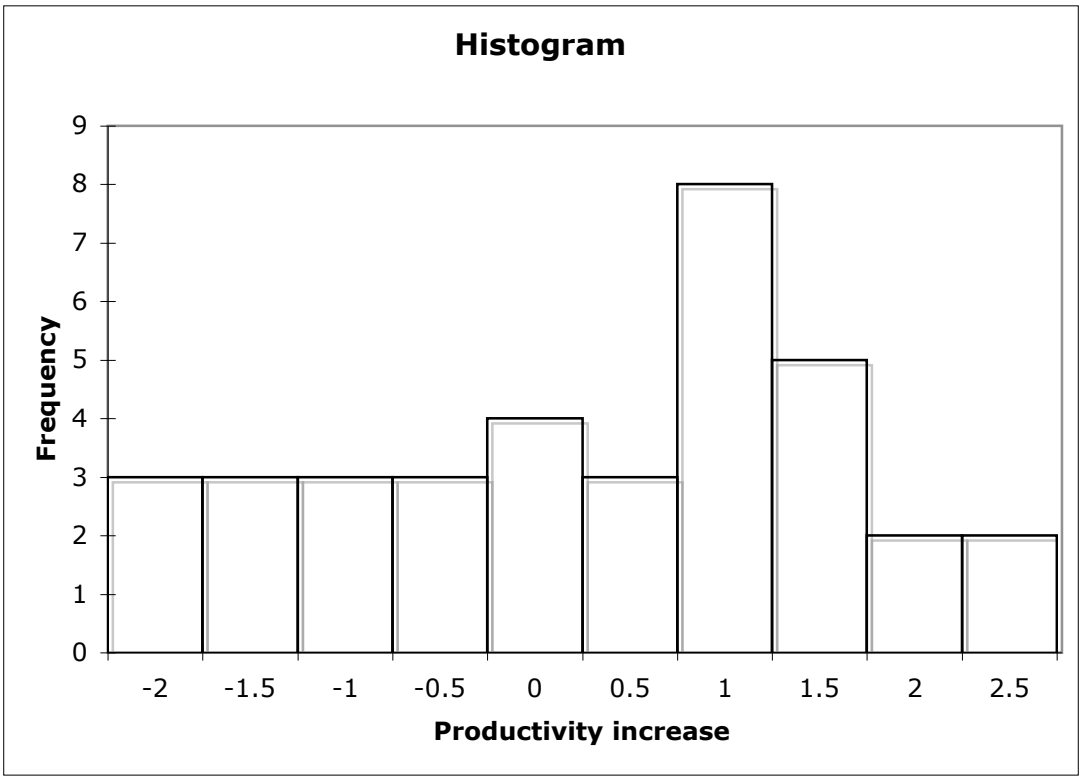
While preparing to teach a course on epidemiology for non-specialists I made a websearch for a simple teaching example on the t-test for comparing the means (averages) of two groups for some measurement. The first example I found compared the mean productivity for two groups of workers, one group of 40 workers averaging 4.8 (in some unspecified units) with a standard deviation of 1.2 and the other group of 45 averaging 5.2 a standard deviation of 2.4 (Figure 1, data generated by the author to match the example). Thinking about this example led me to articulate the sequence of thoughts and questions that follow about the foundations of statistical analysis. In particular, my inquiry explores contrasts between: the statistical emphasis on averages or types around which there is variation or noise; variation as a mixture of types; the dynamics (or heterogeneous mix of dynamics) that generated the data analyzed; and participatory restructuring of these dynamics in the future. A key issue is who is assumed to be able to take action—who are the "agents"—and who are the subjects that follow directions given by others.

1. The t-test assesses the difference between the means, here 0.4, in relation to the spread of measurements around the means and the "sample" size of the two groups. Statistical analysis deems the difference to be less "significant" the larger the spread (captured by the standard deviation) and the smaller the sample. The idea of the statistical analysis is that, even if the groups were actually drawn from the same population, their means could be different by chance. That chance is higher when the spread is larger and the samples smaller. In the example above, the t-test says the chance of a difference of 0.4 is about 0.16. (We'll look at the assumptions

behind this estimation in due course; see #qqff.) With the chance well above 0.05 statisticians advise us not to conclude that the two groups of workers are drawn from different populations, that is, populations with different mean productivity.

2. Suppose, however, that the means were 4.6 and 5.4 for the same sample size and standard deviations (Figure 2). The chance of seeing group means that different when they actually came from the same population is now only around .025. So what could be done with that result? Note first something I didn't mention above: the second group of workers had music playing; the first did not. All other things that might differ—e.g., age, sex, kind of work—had equivalent mixes in both groups. (Sometimes this is described as "all other things being equal," but the workers do not have to be equal in all respects other than the music. To be precise, they vary within groups, but there is no systematic difference in the range of their characteristics or conditions other than the music.) The obvious thing then to be done with the result is that employers conclude that playing music is a good thing for productivity in their firm and, respectively, adopt or continue this practice.

3. There is something else I didn't yet mention: in the original example there was actually only one workplace—the first group in the example is made up of workers measured on one day; the second group is made up of workers measured on a later day when the music was playing. The different size of the groups is simply related to different numbers of missing measurements on the two days. We could, therefore, look at the change in productivity for individual workers who were measured on both days. Suppose that we go back to the first example and find that this change averaged 0.5 with a standard deviation of 1.3 for the 36 workers measured on both days (Figure 3). The chance of a mean difference of this size if the workers actually came from the same population—that is, if music playing had no systematic effect on individuals' productivity, whether good or bad—is 0.01. (Notice that this is a smaller chance than 0.16. The lower value is to be expected when the example is actually one in which the same individuals are measured twice.) Given that the mean difference is positive, again the obvious thing to do is for the employer to play the music.

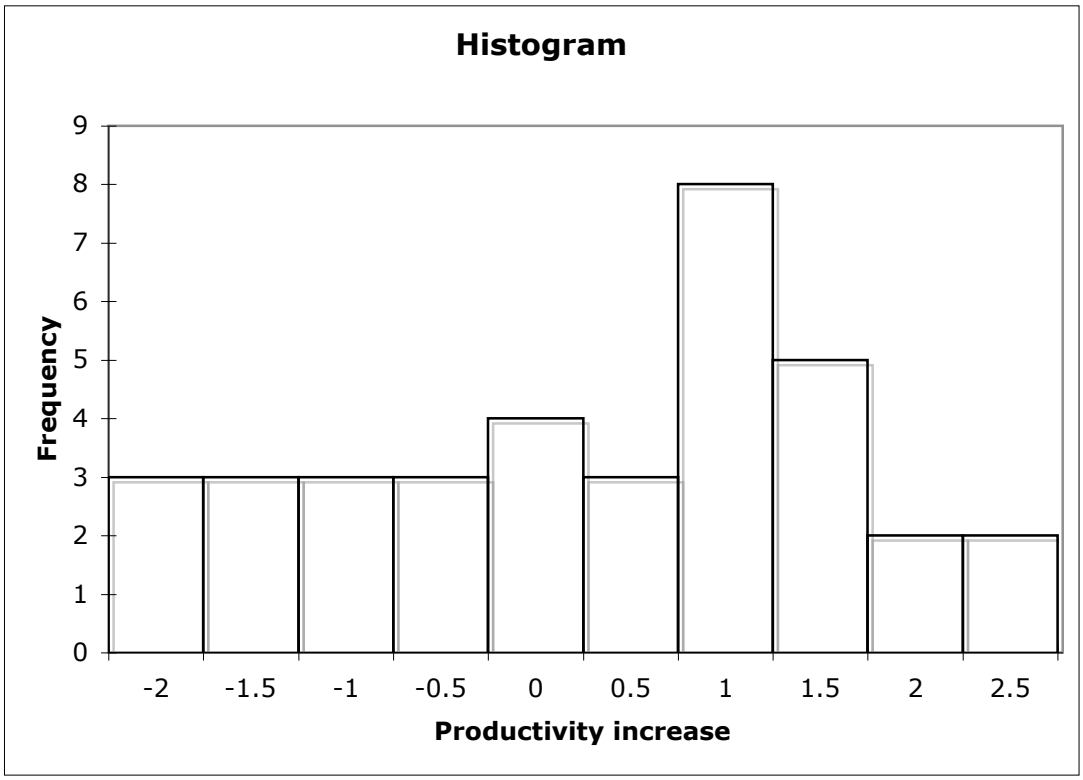


4. Yet, given that the mean difference is 0.5 and the standard deviation is 1.5, there must be many individuals who show a negative difference, that is, whose productivity declined when music was playing. In fact, this was the case for 12 of the 36 (see Figure 3). Should they oppose the playing of music, even though they are in the minority? If they do, should the employer ignore their opposition given that the firm's average individual productivity increases? Does the employer have to power to ignore any opposition? If so, the employer's power to switch on the music comes at the expense of one third of the workforce. In effect, the employer treats them as part of a music-enhances-productivity population, even though they don't fit this type.

5. The employer, faced with competition from other firms and cognizant of obligations to shareholders, might justify playing music by pointing to the increase in average productivity of the workers, which translates into an increase in overall productivity of the firm. There are, however, other paths to higher overall productivity that the employer could consider. The employer might start by asking individuals in the minority why their productivity decreased

when the music played. Suppose it turned out that the tasks of those whose productivity decreased required greater concentration than the tasks of their fellow workers, or that the music chosen is not to their liking. The employer might then rearrange the workplace so that music was not played in areas where workers had to concentrate hard. Or, using headphones linked to airplane-style audio-systems, individual workers might choose from a selection of musical styles. Once the employer starts consulting individual workers, the employer might go on to ask individuals whose productivity increase was well above the mean increase to explain why. It might turn out, for example, that the music countered the tedium of their work and made them less likely to take extended bathroom breaks. By learning about the different individuals, the employer is able, in effect, to dividing the range of individuals into a set of types in relation to working when music is playing. Actions taken by the employer can then be customized accordingly. Such actions might even lead to a higher overall productivity for the firm than switching on music for all. Of course, switching on music for all is simpler and probably less expensive, but it is a matter of empirical investigation whether the firm's net profit would increase more through the customized changes or the simpler one-size-for-all action.

Qq plot of diverse individuals superimposed on histogram



6. There are other things to consider about the one-size-for-all action by the employer. It keeps our focus on productivity in relation to playing music or not, and thereby keeps attention away from the dynamics (or mechanisms or causal connections) through which factors in addition to music influence productivity. We are left to hope that whatever the dynamics are, the addition of music does not lead to any long-term shifts in them. In other words, whatever dynamics generated the data we analyze, we assume that these same dynamics continue into the future even after playing music is added to them. Perhaps, however, a number of workers, including even some who like music, react negatively to the employer exerting the power to pipe in music, worrying, say, that this opens the door to advertizing, anti-union messages, and so on. Moreover, to some extent, a similar assumption about the continuation of past dynamics underlies the customized actions. For example, if headphones were used so as to allow choice of music, would the quality of intra-office communication continue as before? However, there is one difference between the one-size-for-all and customized actions. The latter, by acknowledging the range of circumstances underlying the increases and decreases in individuals' productivity, opens the door to further attention to the dynamics through which factors in

addition to music influence productivity. Of course, much more data is needed to investigate these dynamics and the employer might judge as unwarranted the cost of collecting and analysing the data and acting on any results.

Qq plot of underlying dynamics, current & future

7. Imagine, however, an employer who consults workers, acknowledges the range of circumstances influencing productivity, and worries about whether past dynamics continue even after an intervention (here: switching on music) into them. These steps open the door to the employer mobilizing the workers in a participatory planning process. Skilful facilitators can lead participants through processes that elicit diverse items of knowledge about the current circumstances, generate novel proposals for improvement, and ensure that the participants are invested in collaborating to bring the resulting plans to fruition. If this collaborative change happens, it would matter less whether the past dynamics continued as before because the workers would have become agents in the ongoing assessment and reorganization of their work lives. Moreover, improvement in productivity could result from plans unrelated to the initial issue about having music played. Of course, this scenario assumes that the employer and workers can all be brought together and kept interacting despite differences and tensions until plans are developed in which all are invested.

Qq schema of facilitated participatory planning

8. Could any generalizable lessons be learned from the participatory planning approach to the music-in-the-workplace issue? Suppose that a suite of actions emerged that resulted in increased productivity and profit for the firm. Given that the actions were pursued together, it might be hard to draw direct associations between specific actions and the improved productivity. As such, although the experience in one firm might inspire and stimulate employers and workers in other firms, what happened in the first firm might not provide support for direct adoption of specific measures in other firms. In any case, interested firms would have to pursue their own well-facilitated participatory planning to ensure that their own workers became invested in any changes.

9. How much of a problem is it that the results of participatory planning cannot be extrapolated with confidence from one situation to another? The answer depends on how we envisage people in other situations taking up thecomparisons between groups. A data analyst for the first firm might report an increase in the mean that is unlikely to have occurred by chance and then hope other employers to decide to start playing music in their firms on the strength of the reported increase in the mean for this first firm. In hoping for such an outcome the data analyst would be accepting the employer's power to switch on music and to ignore those whose productivity is adversely affected (#4). At the same time, the data analyst would be discounting the frustrating and all-too-common experience of policy-makers not adopting the policies indicated by the results of data analysis. If we wanted to address such frustration we might see the value of participatory planning: by involving more people in discussion or debate it makes it harder for policy-makers to brush aside the results of the data analysis. Of course, participatory planning also takes us out of the realm of any straightforward extrapolation of results from one firm to another.

10. How straightforward is it in practice to extrapolate from a comparison in one situation to another situation, for example, to take the productivity changes associated with music in our example (#3) as an indication that productivity changes associated with music would occur in other firms. An assumption required for such extrapolation is that there is no systematic difference from the first situation to the other in the range of the characteristics or conditions that might modulate the range of effects of music on changing individuals' productivities. This assumption is difficult to establish without knowing what the range of relevant characteristics or conditions are. To see how readily the assumption might break down, let's go back to the first example and examine the workers who were absent on one of the two days. If we omit them from the analysis (as we did in #3 and #4), we are assuming that workers who were measured both days are no different from the workers measured one time only. Something I didn't mention is that the latter were present both days, but their measurements were not submitted until after the analysis was done. Looking at their measurements, it turns out that 9 of these 14 had negative increases in productivity and their average was just above zero. This makes us wonder whether the range of characteristics of the workers who submitted their measurements differs

from that of the workers included in the initial analysis. (Note: this revelation does not qualitatively alter the earlier discussion from #3 onwards. The mean increase for all 50 workers was 0.4 and the chance of a difference of this size if the workers actually came from the same population is about 0.04. This is still small enough to make continuing the music the "obvious thing to do.") We might have thought in advance that the later submitters would not differ from the others. Now that it seems that they do, our hypothesis-generating brains might get to work. For example, we can wonder if late submitters are workers who resist top-down workplace management and, as such, associate submission of productivity information sheets with acquiescence to management. It is hard to know without more investigation. In any case, if the assumption of no systematic difference breaks down in the case of workers at the one workplace, it must be even less reliable when we extrapolate from one workplace to another.

11. Suppose we doubt the no systematic difference assumption in #10. We can divide the data into slices based on characteristics other than productivity that we might have measured and do separate extrapolations for each slice. We might, say, determine the change in productivity with music for only the men in both workplaces, then determine it for only the women. We might determine it for only workers under 30, then for only workers over 30. And so on. The smaller the sample in each slice, the less significant a given difference is deemed to be according to statistical analysis (see #1). Although the principle of slicing still holds, there must be diminishing returns to repeated slicing.

12. What would it mean for a participatory planning approach to take the slicing into account? I have never heard of this happening, but two responses to this hypothetical scenario suggest themselves. In the first, the participants divide into small groups according to the characteristics being used to slice the data. The participatory planning process is undertaken for each slice separately and the plans implemented separately by each slice. A coordinating committee would probably be needed to bring the plans into line with each other and try to resolve any conflicts. In the second response, all the participants would be informed of all the results of the slice-specific data analyses and keep them in mind as the participatory planning process proceeds. (The data analyses might have shown, for example, that all individuals in the under-30 slice

show positive productivity increases but for the over-30 slice the change is more often negative than positive.)

13. The issue of extrapolating from a comparison in one situation to other situations leads me to return to the original t-test—the one that we were considering before learning that the groups of workers were actually the same people measured twice (without music playing, then with music). The t-tests in #1 and #2 compared groups of workers from two workplaces, one with and the other without music playing. For this comparison to be meaningful, we have to assume that there is no systematic difference from the first set of workers to the other set in the range of the characteristics or conditions that might modulate the range of effects of music—or lack of music—on individuals' productivities. This is very similar to the point in #10, but here the issue is not extrapolation to other situations of changes in productivity seen for a group of workers in one situation, but establishing a change in productivity by comparing two different groups of workers. If we are in doubt about the no systematic difference assumption, we can again divide the data into slices based on characteristics other than productivity that we might have measured and do separate comparisons for each slice.

14. What would participatory planning mean for separate groups of workers? Again this is a hypothetical scenario, but a number of responses suggest themselves. As a first response, the two groups of workers undertake the participatory planning process separately, although one group is told that playing of music is an option that can (or should?) be considered. Perhaps, at the end of the processes, each group could be informed of the plans formulated by the other group and revise their plans if they are so moved. As a second response, the participants in each group would be informed of the considerations and conclusions from the other group at each step in the participatory planning process and be able to take these into account as they proceed. A variant of this response would be needed if one group undertakes the process before the other, in which case only the later group's process can be informed by the other group's. In all of these responses, slice-specific planning or analyses could be brought into play (see #12).

15. Participatory planning is conceivable for groups of workers in firms, but how far could the scenario of using participatory planning be stretched? It is not hard to imagine extending it to a

group of humans that do not know each other and do not come together in defined assemblages, say, single mothers or secondary school science teachers (where the issue is no longer playing music in the workplace). We would bring together representatives from the group for the participatory planning process. Admittedly, facilitation of participatory planning becomes more difficult in practice when the participants are not from one community and have less in common (e.g., in upbringing, work, lifestyle, language, and so on). Moreover, because only a fraction of the group would be involved, the process would be less likely (using the words from #7) to "ensure that the [group members] are invested in collaborating to bring the resulting plans to fruition." The representatives could, however, be encouraged to take this as a key issue to factor into their deliberations. Of course, doubts may arise about how representative the representatives are with respect to the range of relevant characteristics or conditions, especially since these conditions may not be fully known (see #10).

16. The participatory planning process is harder to envisage for trials involving groups of non-humans, such as crop plants of a certain variety or machine tools of a certain design. It is not impossible, however: we could assemble spokespeople for the non-humans, choosing them so as to span the range of expertise and interests relevant to the contrast in question, which presumably would not be the playing of music but, say, application of fertilizer to the plants. Plant physiologists, farm-workers, plant geneticists, pest management specialists, agricultural extension agents, accountants, and so on, may have a range of insights that could enter a participatory planning process aimed at increased productivity of the crop but entertained more options than yes-or-no for fertilizer. In two key respects, however, the workaround is limited. First, unlike the workers in the firm who experience no music then music, no individual plant experiences both sets of conditions. The analogy then is to the case in which there are two firms each with a separate set of workers and where in one, music is not played; in the other, it is. Second, the individual plants from any one variety are often very similar, if not identical. Sometimes plant breeders pay attention to the variation and choose individuals with desirable traits to be the parents of the next generation, but often the variation among the plants of any given variety under any one set of conditions is treated as noise. In any case, it is easy to imagine farmers who are content to use their power to switch on the fertilizer as long as the mean

yield in the trials is higher and to ignore the possibility that some plants in a variety yield low under fertilizer application.

Let us put participatory planning to the side for a moment and delay consideration of how we might expose the dynamics that generated the data. Qq could this be an extension of the PP?

17. Suppose that the playing of music was not a simple yes or no situation, but one in which workers at the firm on the second day were played music to different degrees, e.g., for different lengths of time or at different volumes. The obvious adjustment to the analyses would be to divide the group according to the degree of music and compare the changes in productivity among the slices. A generalization of the t-test, namely, the 1-way analysis of variance or "ANOVA" is used for such comparisons. The analysis does not assume that the mean productivity change for the slices increases with degree of music playing. It is possible for slices with an intermediate degree of music to show the most positive mean productivity change. The considerations raised in #10 about no systematic differences, in that case among the workers who submitted their analysis on time and those that didn't, applies here to workers played different degrees of music. As before, workers could be measured at two different firms, where music was only played at the second firm but this time played there to different degrees. The differences in mean productivity between the first firm and each slice in the second firm could be compared. Workers could also be measured at several different firms, at each one music being played to a different degree, and the mean productivities compared. The considerations in #13 about no systematic differences in relevant characteristics other than the degree of music being played apply to the last two cases.

18. Up to this point, the cases have all been experiments. The employer(s) deliberately established whether music is played and the degree of music being played to groups of workers. With a seemingly simple shift the last case in #17 could become an "observational" case. That is, suppose the firms in which we are interested happen to have music playing to different degrees even though no employer established this deliberately with a view to changing productivity. As in the last case in #17, the mean productivities could be compared. A notable difference, however, is that the observational case could be extended to factors or variables that,

unlike, the playing of music, could not be experimentally altered, e.g., skin color. Qq PP means something different…

19. Whether experimental or observational, the cases in #17 and 18 admit a variant, namely, the degree of music played for any slice within a firm or for each firm could be non-uniform. The analyses would then be comparing productivities in relation to the mean degree of music being played for the slice or firm.

20. Whether the comparison is in relation to discrete degrees of music being played or mean degrees (#19), it is possible to look for trends in productivity across degrees of music playing. If the degree of music played for any slice within a firm or for each firm is non-uniform (#19), these analyses can use the actual degree of music playing as long as that has been recorded.

How can we act?

qqshift back to diversity & to underlying heterogeneity

factor /variable

2 -> Not music but gender.

Calculation of chances (assumes imaginary probabilistic dynamics)

comparison b/w two situations mimics one place, 2 measures
-> can we stretch the notion of part planning? (see #7 [adjusted so it leads into this point] and ..

extend t-test to AOV and slicing to regression… add principal components

issue of slicing on many variables & underlying variables (in dynamical relations)

find notes on multiple slicing..

—or not implementing the policy the way it was practiced in the initial trial situation that was analyzed.

underlying factors

2 -> How do we know all other things "equal."
3 Missing values might be different
one workplace situation as model for milti sites, but are all other thinsg equal
complaints about group average policy not working..
paired vs. unpaired is bad example for web, but interesting given that most analysis is unpaired but simulates paired.
Fraction of people being treated as if they were of different ttype