CORRESPONDANCE ON THE ISSUE OF MEDIAN SPLIT IN STRESS RESEARCH

Dear Colleagues,

You will find below an email correspondence between Sonia Lupien and two expert members of the Stress Centre extending a discussion that took place at the first of the Journal Club's held by the Centre for Studies on Human Stress. Joseph Rochford PhD of the Douglas Hospital Research Centre and graduate statistics professor led the Journal Club that dealt with the appropriateness of performing a Median Split on data in the behavioral sciences, and most particularly in the field of stress research. Jeremy Miles PhD is a fellow of the Royal Statistical Society that Sonia approached to have his view on the matter. The discussion was based on the article "**Breaking up is hard to do: The Heartbreak of Dichotomizing Continuous Data**" by David L. Streiner (2002) that appeared in the *Canadian Journal of Psychiatry 47:* 262-267.

Enjoy!
Tania Elaine Schramek
Coordinator, Centre for Studies on Human Stress

Sonia Lupien
Dear Jeremy,

I am contacting you today to ask for your advice-comment on a Stats paper. We have Journal clubs in which we discuss new ideas in the literature. We had one today on the attached paper, which basically tells us that it is a major mistake to perform a median split in our studies. I wanted to know what my statistician friend would think about this.

I don't know, but I thought that having a short written comment from you on the paper could be put on the website of the Centre for Studies on Human Stress (www.douglas.qc.ca/stress) which could then lead to discussion between our members!

Let me know what you think and if you want to take me up on my offer!

Cheers
Sonia

Jeremy Miles

Sonia, sure, sounds like fun. Any time anyone wants me to ramble on about statistics, I'm happy. Below is a text I have written to answer your question and to discuss the paper.

There is something of a division between psychologists, who use statistical methods in their research, and statisticians, who think about evaluate, and developed those methods in the first place. It's rare for psychologists to read articles in statistics journals, and perhaps rarer for statisticians to read articles in psychology journals. (And when the statisticians do read the articles, if they object, there is little they can do about it. This is in contrast with some medical journals, for example the British Medical Journal, which allows 'rapid responses' to articles on its website. For examples of two of mine see:
http://www.bmj.com/cgi/eletters/329/7477/1259#87226
http://www.bmj.com/cgi/eletters/330/7481/17#91392

Psychologists tend to learn statistics by reading books written by other psychologists. Few psychologists ever read '*Statistical methods for research workers*' (and I'll admit to having a copy, and only referring to it a couple of times) – instead, in the past, they read '*Fundamental statistics in psychology and education'*, by Guilford, who had read Fisher, but had misinterpreted parts of it (or at least had altered the emphases; e.g. Kramer & Gigerenzer, 2005). Other psychologists then read Guilford, wrote textbooks, and so on, and the errors remained – unnoticed, and although this is changing in recent texts, there is still much misinterpretation of the meaning of a probability value (see, e.g. Haller & Kraus, 2002), it's correct in Fisher, but wrong in many other places. (I wonder if Fisher didn't take his time to emphasize the point, because he was just too clever to realize how hard it was.)

There are many examples of things that psychologists who use statistics worry about, which statisticians don't. Here are some of them:

- Worrying about homogeneity of variance in the t-test. The short answer is: don't worry about it, just don't assume it. (The t-test has two versions, one of which assumes it, one of which doesn't. The version that doesn't assume it is hard to do by hand, and if we are teaching students to do statistics by hand, this is hard. But no one who does a t-test 'professionally' does it by hand, but the historical problem stays with us [Zimmerman, 2004]).
- Normal distribution assumptions cause all sorts of worry – unnecessarily - if you are worried about it, bootstrap it. Almost any statistic can be used non-parametrically by bootstrapping. Programs favored by statisticians (S, R, Stata, SAS) make bootstrapping easy. SPSS, favored by psychologists, doesn't. If your sample size is moderate to large, the normal distribution doesn't matter that much anyway. There is a joke in statistical circles that mathematical statisticians think that normal distributions are important, because applied statisticians find them everywhere. Applied statisticians think that normal distributions are important, because the mathematical statisticians say that they should find them there. It's not a very funny joke (but jokes that statisticians make rarely are). In fact if you look, normal distributions are found almost nowhere (Micceri, 1989). People still check for normal distribution, and they check using significance tests – if your normality test is not significant, it just means that your sample wasn't very large.

- A final example of this historical effect is that of ANOVA and regression. If you don't have a computer, multiple regression is hard work. Really hard work. I was once told a story by a statistician who was an undergraduate in the 1950s. His class group of about 8 students decided to carry out a regression analysis, using 5 predictor variables. It took them all afternoon to do the calculations – that's about 4 person days, and these were statistics students, not psychologists. This meant that regression was hard work. One of Fisher's many insights was that in the special case where predictors are uncorrelated, you can do regression in a much easier way, by partitioning sums of squares. Of course, predictors are rarely uncorrelated, but they are if the predictors are categorical and you can assign individuals to those categories – that is, if you have done an experiment. Regression and analysis of variance are different ways of thinking about exactly the same thing, but the way that psychologists are taught makes them believe that they are different. Statistics packages almost never do ANOVA any more – they do regression, and then they tweak the results to make it look like ANOVA.

Now, where were we? Oh yes, back to the point. Dichotomizing. Psychologists tend to learn statistics by learning experimental methods, such as t-tests and ANOVA first, and then regression. No psychology undergraduate course that I know of teaches interactions as a form of multiple regression, but they are, and one can think of them like that. Because of psychologists' comfort with ANOVA, they tend to want to twist their data into a format which allows the use of ANOVA, rather than regression. Maxwell & Delaney (1993) wrote, 13 years ago, "*For many years, behavioral statisticians have chided psychological researchers for artificially dichotomizing continuous variables*", and go on to cite McNemar, from 1969.

As the target article (Streiner 2002) states in its conclusion about dichotomizing – "*Don't*". I'd extend this a little and use a quote that I use when I review papers that have employed dichotomization, from MacCallum, Zhang, Preacher, & Rucker (2002) "*The use of dichotomization in practice is described, and justifications that are offered for such usage are examined. The authors present the case that dichotomization is rarely defensible and often will yield misleading results*." (But I'm going to be using the Streiner quote in the future.)

Dichotomizing is equivalent, in power terms, of discarding about one-third of your data. If those data were expensive, time consuming, or difficult to collect, this is obviously foolish. You might have saved 1/3 of your time, analyzed the data appropriately, and then had the same power to detect an effect. However, occasionally data are cheap, or even free. They are given to you, to analyze, and the sample is large. On these occasions, why not just do it the 'easy' way, and dichotomize? The answer is provided by Maxwell & Delaney (1993), who showed that, as is obvious, you lose power to detect main effects. As is less obvious, you also increase the probability of detecting interactions, when there are no interactions present in the population. This means that our true type I error rate is somewhere above our nominal type I error rate – that is to say, we think we're using 0.05 as a cut-off to determine what's statistically significant, but we're not, we're using a higher value. And we don't know how much higher. And if

there's one thing that statisticians (being a conservative lot) dislike more than having lower power, it's having a higher type I error rate.

As a final point, I'd say that analyzing data properly is hard. There are many different ways to do it, and it's impossible for anyone to be familiar with all of them (a statistician colleague of mine once said that part of being a good statistician is knowing who might be able to solve a problem). However, designing studies and collecting data is hard too (possibly harder). There's no reason that any one person should be able to do both of those things.

Sonia Lupien
OK....you agree with Streiner.....that is TOO bad!

Jeremy Miles
I suspected you might not like that. Sorry.

Sonia Lupien
I say 'too bad' because I am still not sure about this. First, most of the stats data show that median split (MS) will increase the probability of Type II, so in my mind, if I still find an effect with a MS, then I'm in business, right??? (I can sense your sadistic smile at preparing your answer to this one!).....

Jeremy Miles
Yes, that's true. But if you are doing interactions, it increases Type I.

Sonia Lupien
Second, in behavioral sciences, we usually measure more than one variable and between you and I, regression analyses are hard to interpret. Now, let's say that I have 5 dependent variables and I am interested in one in particular (for example, depressive symptomatology). Then, I could median split the subjects on their score on the depressive questionnaire I used and see whether they differ on all other 4 variables that I gathered. If I do a regression, I will only know that 1 is related to 2, that is related to 3 etc, but in my mind (and I may be wrong here...), I won't see that small, subtle differences between groups....this is why and when I find MS useful.

Jeremy Miles
But if you think there will be effects/differences, surely they're not dependent variables any more? I'm not exactly sure what your hypothesis is.

Sonia Lupien
My last comment for you for our email conversations, is that if MS is so bad, then can I do a cluster analysis and split my groups according to the cluster and check their scores on my other variables?

Jeremy Miles
Probably not. (I'm very negative today, sorry). Cluster analysis is a bit evil too. The

problem with cluster analysis is that it will find you stuff, when there's nothing there but nonsense. Much better than cluster analysis is latent class analysis. This is similar to cluster analysis, but you get a p-value for the number of clusters, and it does the whole model at one - if you think something predicts cluster membership then you can put it in the model at the same time.

Sonia Lupien
In summary, do we really really have to be so rigid in order to understand our results????

Jeremy Miles
No - not to understand them, but to do them. We should use the best approach that we have to understanding our data, and then we should use the clearest approach that we have to explain the results. So, do a horribly complicated analysis, and then to explain what's going on, use a median split (or a median split analogy).

There's easier/better ways to understand interactions in regression out there. For e.g. Curran and Bauer wrote a paper in Multivariate Behavioral Research on this (Possibly not a journal you read every day) which I wrote about in my particularly exciting regression blog: http://www.jeremymiles.co.uk/regressionbook/2006/03/interpreting-interactions.html

Joseph Rochford

Jeremy raises many interesting and challenging ideas in his commentary, some of which are directly germane to the question of the practice of dichotomization, other less so, but nonetheless notable. I agree with all of his conclusions, albeit not always for the same reasons. I should like, therefore, with this correspondence, to add my "two cents".

On the difference between psychologists and statisticians

First, a confession: Unlike Jeremy, I am not a Fellow of the Royal Statistical Society, although I am addicted to crunching numbers (much more intellectually stimulating then cross-word puzzles, and it is my preferred way of trying to ward off early dementia). I am a psychologist by training, but not a member of the American Psychology Association (I am probably eligible, but to paraphrase Groucho Marx: "I am not sure I would want to join a group that would consider me as a potential member!). That being stated, I am an observer of behaviour, of both living things and inanimate numbers, and feel I can comment on the distinction between psychologists and statisticians. In fact, I think the respective populations should be redefined: statisticians and any experimentalist who thinks he/she is required to perform some form of an inferential statistical test.

My experience with trained statisticians has been mostly with the mathematical species, and less so with the applied species. The problem with this breed is that they are very good theoreticians, and not especially pragmatic. They develop many wonderful statistical tests, and then go to great lengths to show how they are not applicable for

"real-life data" Most experimentalists are more concrete, they do stats because they have to. As a Ph.D. candidate once stated to me at his thesis defence: "*I do statistics because I am required to. However, if I really want to know if the means of two groups are different, I look at the figure and check for overlapping error bars*." My aim here is not to defend the "eyeball" method of determining statistical significance, but rather to point out that experimentalists often see statistics as a necessary evil. Sort of like writing grant applications, if you don't do it, you can't play in your lab.

In the course of many discourses, I have often heard the statistician object: "*Your data cannot be analyzed parametrically, because they clearly violate the assumptions of every parametric test*." Experimentalists reply with the standard rebuttal: "*Perhaps, but most parametric tests are "robust" to violations of their assumptions, and, in addition, nonparametric alternatives are either not available or lack power*". There are 3 problems (at least) with this contention. First, "robust" is undefined, or at least most experimentalists cannot cite any evidence to quantify how much the violated assumption will impact on the results of the statistical tests (the simulations are out there, folks, we just don't bother to look them up). Second, there are nonparametric tests that can be applied for the purposes of (1) simple regression, (2) multiple regression, and (3) multifactorial experiments. I spend 2 hours, for instance, introducing my students on how one can adapt the Kruskal Wallis and or Friedman tests to two-way, factorial designs (be they 2 between, 2 within or mixed). Third, no one bothers to assess relative power *a posteriori* (we might do so *a priori*, but only to appease the one statistician or quantitative expert who may be sitting on the grant committee we are applying to). If a test yields a significant difference, power is irrelevant, if not, we generally don't bother to check, after the fact, whether our failure to reject the null is a reasonable decision, or if we simply were pretending to drive a formula 1 automobile, whereas in fact we were riding a bicycle.

Historical legacies and laziness

Experimentalists are practitioners: "These data MUST be analyzed, so let me use the "gold standard" test. By gold standard, read "parametric", because a random sample of articles from any scientific journal in any year will reveal that the great majority of statistics reported in the results section come from parametric tests. Additionally, as identified by Jeremy, there is an historical precedent to the use of parametric tests and the avoidance of power analysis: In prehistoric times, calculating nonparametric tests manually was labour-intensive, and don't get me started about power analysis! Prior to the hand-held calculator (remember those days?), we had tables of squares and square roots, we had the slide rule (or the abacus for the more eccentric and manually adept), making it relatively easy to compute sums of squares. Where are the tables that permit rank-ordering?

These issues are especially acute when we consider complicated experimental or correlational designs (i.e., two or more factors, multiple predictor or dependent variables). If you have nothing better to do with your time: Go back and look at the kinds of experiments that were done in the late 50's and early 60's. Many years ago, I

performed an informal (and non-scientific) survey from the *Journal of Comparative and Physiological Psychology* (in those days, this was "the impact factor" journal in the area of animal learning and physiological psychology). What struck me was that the great majority of experiments consisted of either two group or one factor designs. I suspect that the prospect of trying to analyze a multifactorial experiment manually dissuaded most scientists from attempting it!

A second factor is laziness and/or ignorance. Even today, with the advent of the lap-top computer, we are hindered at times by the dearth of available (or user-friendly, or affordable) software packages that can analyze multifactorial designs nonparametrically, and the problem with performing a power analysis on a multifactorial design is, as a colleague once stated to me, "you have to know what you are doing". Although I show my students how one can use nonparametric tests to analyze two-way designs, I don't assess them on it (i.e., there is no quiz or exam question forcing them to do it). I like to convince myself that I don't assess this skill because it is computationally intensive, and I don't want to force my students into conducting a test they likely will not use in the future. However, in the light of cool, hard reflection, I realize I am not that nice a guy, I'm just too lazy to do the computations myself (as an aside, I feel compelled to note that I do force my students into applying different data transformations, and then assess which one is the best. I take great satisfaction in seeing that most students conclude most transformations don't really make that much of a difference to the data. So, there is an important lesson learned!).

I mentioned above that experimentalists are practitioners (with apologies to any quantitative experimentalists who might be reading this) and I really believe this to be the case. That being said, as in medicine, there are good practitioners and bad practitioners. The bad practitioners are the ones who either can not (because of a lack of knowledge) or will not (because of lethargy) take the time to develop a reasoned justification for a test. M.D.'s don't get the diagnosis and treatment right 100% of the time, but I for one am much more willing to be treated by a physician who can state: Given your symptoms I had a choice between disease X and Y. I choose X because of… Accordingly, I choose treatment X, because it is the most effective." In addition, I, for one, no longer accept the following response from a grad student: I performed this test on my data because that is the one that my software package computes."

ANOVA, Regression and Dichotomization

Enough rambling and sermonizing, I'm supposed to talk about dichotomizing. So, let me begin, in a round-about way.

Jeremy makes the observation that multiple regression (MR) and ANOVA are perceived as being different, whereas they are in fact the same. They are the same, as Jeremy (and Fisher before him) notes, if the predictors are categorical, and you have conducted an experiment. My statistics professor in graduate school, Dr. Alistair McLean, was a diminutive man with a booming Scottish accent, the contrast between his stature and his enunciation no doubt contributed to the popularity of his declarations, one of which was

(and please try to read this with as heavy a Scottish accent as you can muster): "*Analysis of variance is simply a special case of multiple regression*." Ever wonder why programs like SPSS refer to the "general linear model" when they output the results of an ANOVA?

Whereas in some instances, MR = ANOVA, to paraphrase George Orwell, "*ANOVA and MR are created equal, but MR is more equal then ANOVA*". Just because they can do the same thing in one instance does not mean they do so in every case. As Jeremy alluded to, psychologists are generally exposed to MR **after** having received at least two (and sometimes three) exposures to t-tests, F-tests and the like. Nothing wrong with this, necessarily, although it did leave me with one false impression (this impression was mine, and mine alone, but I suspect it maybe shared by many others): If we tweak predictor variables from continuous to categorical, we are converting a correlational design to an experimental design. Now that I am older and (hopefully) wiser, I fully appreciate the logical error in this quasi-syllogism.

Why am I being so forthright about my intellectual deficiencies? Because dichotomization creates categories! And categorization implies experimentation! Let's assume we have measured two dependent variables, we take one dependent variable, perform a median split (or some other dichotomizing or "trichotomizing" technique) and use it to form "groups". We then compute the means of these "groups" on a second dependent variable and assess whether they are different using a t- or F-test (depending on the number of groups formed). Well, we were taught that t-tests and ANOVAs are used for experimental designs, right? So, the implication is that by dichotomizing, and by subjecting the dichotomy to ANOVA, we are in fact treating what is in reality a correlational design as if it were an experimental design. OOOPPPS, time to revisit *Research Methods 101*, and to re-examine the issue of the limitations of correlational designs in the context of determining cause-effect relationships.. In experiments you "DO DIFFERENT THINGS" to different groups of subjects, or at different times in the same subjects. That is what allows us to conclude that what you did differently may be causal. Spitting subjects into different categories as a function of some arbitrary distinction does not constitute "DOING DIFFERENT THINGS". It is the equivalent of putting our apples into an apple basket and our oranges into an orange basket. All we have done is put them into different baskets. And, whereas I can recall my high-school math teacher insisting that it was never a good idea to mix apples with oranges, I also recall that this guideline was never followed by a justification (other then the theory of "ordered sets", and ordered sets are not statistics).

In short, when we dichotomize, we sort but do not manipulate. Dichotomization may constitute differential categorization, but you are doing the same THING to the entire data set. Some tangential ideas on the above statement:

- For those of you more comfortable with technical jargon, dichotomizing on the basis of one variable means you are treating a dependent variable as an independent variable. This presumes that you know the direction of the cause-

effect relationship, which is always an assumption that needs to be made carefully, and even in those instances, can be quite the "slippery slope".

- Nothing wrong with dichotomizing to form groups and then using these groups, as defined, to do different things to (i.e., to manipulate differentially). Here, instead of assigning subjects randomly, you assign through dichotomization. So, for instance, if you want to assess whether income influences stress, define the median (Stats Canada to the rescue!), use this to split a subjects into high and low income, then take half of each of these groups, stress em' and measure cort levels, and use the other half as controls (i.e., don't stress 'em). One caveat: standard ANOVA assumes randomization, dichotomization may invalidate this assumption. It is for this reason that statisticians have developed ANOVA specifically for such "blocked designs".

- What's so special about the median? The median defines the 50-50 split. But as Streiner notes so eloquently, it can create differences where none really exist. If median IQ is 100, and I want to hire a research assistant with "average" intelligence (because I don't want a total incompetent, nor can I afford to hire a "wiz kid"), I'm not going to discriminate between candidates with IQ scores of 98 and 102! This is akin to the difference between "statistical significance" and "practical significance". Test enough subjects, and you may find that the mean body height of children raised in "impoverished" environments is significantly different from those raised in "enriched" environments. However, I doubt very much (or at least hope) that any politician would feel compelled to invest monies in any new programs to correct this "problem" if the means were respectively, 166 and 171 centimetres (a difference of about 2 inches, for those of you more familiar with traditional measurement).

- Dichotomizing can also eliminate differences where they do exist. If I can afford to hire a wiz-kid, I certainly want to discriminate between a candidate with an I.Q. score of 150 and one with an I.Q. score of 102.

One last point that Jeremy makes which I believe is very, very important, and deserves emphasis. It is true that dichotomizing reduces the power to observe a significant main effect. Streiner's first example illustrates this point, and in general, this is what he is referring to when claiming that dichotomizing reduces the "signal to noise" ratio. So the retort could be: "If I still get a significant main effect with reduced power, I can be even more confident that my effect is real." OK, but this assumes you are more willing to make a Type II error relative to a Type I error. Also, it throws your whole power estimate (assuming you have done one) up in the air. Fact is, statisticians have developed variants of power analysis that allow you to see what happens when you manipulate the respective probabilities of Type I and II errors.

In addition, Jeremy takes the argument to its logical conclusion, for two or more factor designs, at least. Dichotomization may reduce your power to observe main effects, but it increases the chances of seeing a significant interaction. Why? Here is one way to think

about it: When you have a big main effect, it “steals” variability from your estimate of the interaction sum of squares. Total variability is fixed, so if you give a large piece of the birthday cake to the birthday boy, not as much cake is left over to divvy up among the guests. If you classify your birthday boy as obese, and restrict his access to the cake, you have more cake left over for the guests. Most of us are (or at least should be) aware of the fact that the phenomena we study as life scientists are multiply caused, and in complicated ways. But to manipulate your data in a way that maximizes your chances of seeing an interaction does not seem like fair play.

Come to think of it, speaking of fair play, isn’t highlighting the differences between statisticians and experimentalists a form of dichotomization? Isn’t it funny, then, that a statistician and an experimentalist both feel they should be put into the “anti-dichotomisation” basket?

Jeremy Miles

The response is great, I particularly liked the last bit about the birthday cake - I'll be using that next time I need to explain it (if that's OK). (And I also like the fact that we agree).

Joseph Rochford

Also glad we agree. Sorry, Sonia, but sometimes life does throw us little curveballs. Feel free to use the cake analogy.

Jeremy Miles

I think I got into statistics from the necessary evil (as I've explained before - Miles, 2006, available here: http://www.jeremymiles.co.uk/mestuff/publications/46.pdf). Every time I wanted to test an interesting hypothesis, it turned out that the stats were hard.

I'm impressed that Joe's students listen while he explains adaptations of Friedman and Kruskal Wallis tests; mine never would, he's obviously more compelling than me.

Joseph Rochford

Second, I said that I present variations of K-W and Friedman to my students; I never claimed that they listen. In fact, subjective impressions are that this would be one situation in which it would be safe to dichotomize: snorers and eye-closers. If I added "attentive listeners" and "bright-eyed faces" to the categorization, I’d definitely have to do a Fisher's on the data, in that I would have 2 cells with frequencies less than 5. This would still be the case if I collapsed the data over the 12 years I have been teaching stats.

If Sonia can refer to me as a stats freak, I can continue to refer to you as a statistician. We now need to classify her...

Jeremy:

Sonia can take it us disagreeing with her - she's tough.

Sonia's last note

OK guys….you win. But I am certain that when I am alone at home, the curtains are down, no one is watching….I will open my laptop and do a median split on my new dataset, just to see….just to understand what the data are telling me…and then, when something comes out of this, I will call my two great stat-friends and tell them about what I found and ask them what stat method I should use to show what the median split told me!