Statistical Thinking Comment Policy

Comments are welcomed. Be informed, informative, respectful, and criticize ideas, not people. Please read our Comment Policy before commenting.

Comments on this entire site are premoderated (only moderators can see this message). Change site settings. ✕

69 Comments Statistical Thinking Disqus' Privacy Policy Frank Harrell

Favorite Tweet Share Sort by Best

Join the discussion…

Joseph • 4 years ago

As an early career clinician-scientist this blog is exactly what I have been looking for a while.

I have been guilty of dichotimization and categorisation in my own research. I have to say it is primarily because of how institutionalized the medical world views research. Our leaders dictate this way of thinking as do the statisticians who are paid to help with this type of research.
My own logical deduction kept me questioning the amount of information loss by categorization. I have to say that there is simply not enough information or support out there for clinician-scientists who want to change this way of thinking. The average knowledge of a doctor on even basic statistical methodology is shocking. There needs to be a foundational change in the way that we educate doctors if this is ever to change. For what its worth,I also think the peer review system is far too broken and contributes to the ongoing problem. Many of the correct methods of doing statistics would be rejected by peer reviewers for being too 'novel' or 'unproven.'

The BBR document is great on touching on some of these issues.

As a researcher whose research is primarily on biomarkers I do have a lot of questions - which clearly cannot be answered all here. One of the things I have wanted to do is to step away from the problems of stepwise selection in cox regression and also categorisation. Is multiple linear regression the way to start? How does one incorporate binary variables (mutation status) with continuous variables in a model? Page 404 in the BBR seems to touch on this. What resources apart from your own do you suggest as a go to for a clinician with some statistical knowledge and a reasonable working knowledge of R?

2 ^ | ⌄ • Reply • Share ›

Frank Harrell Mod ➦ Joseph • 4 years ago

Great questions Joseph, and I'm glad you are worried about these things. For some of the issues, asking questions on http://datamethods.org is recommended. As for resources, my book Regression Modeling Strategies and its course notes available from http://fharrell.com/links as well as Ewout Steyerberg's text Clinical Prediction Models are worth some study. Also see these papers for tutorials and case studies: http://hbiostat.org/papers/.... My R rms package implements many of the methods in my book, and especially makes it easy to relax the linearity assumption for predictors.

^ | ⌄ • Reply • Share ›

Jorge Teixeira • a year ago

Dear Frank, noobie question, but the 7 assumptions apply also to anova of repeated measures?

Or just to t-test or anova of the deltas? Since these are equivalent to ANOVA RM, I presume so... but just to double check!

Thanks!

^ | ⌄ • Reply • Share ›

Frank Harrell Mod ➦ Jorge Teixeira • a year ago

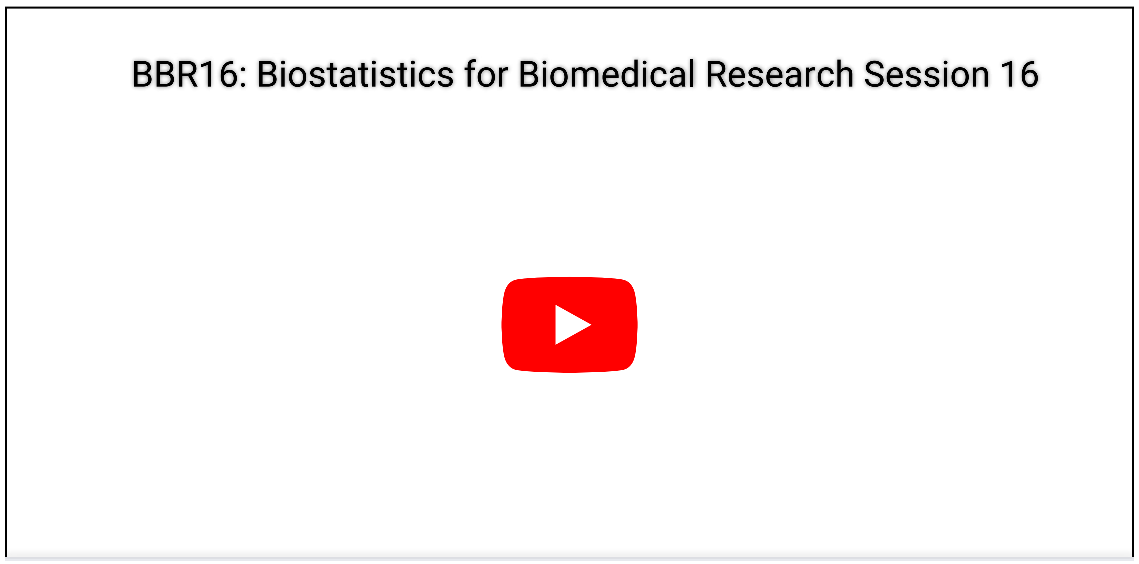
The assumptions apply when you compute a change from baseline. If you are using a good longitudinal model with flexible adjustment for baseline then there are fewer assumptions.

^ | ⌄ • Reply • Share ›

Jorge Teixeira ➦ Frank Harrell • a year ago

Thanks Frank.

1) Can you confirm that we can assess pre-post changes with mixed-effects with sort of ANCOVA structure (instead the time:group interaction), as you suggested in the end of this video



see more

^ | ⌄ • Reply • Share ›

Frank Harrell Mod ➦ Jorge Teixeira • a year ago

This isn't the best place for general questions. Use datamethods.org (if health related) or stats.stackexchange.com. But a short answer is that with a longitudinal model that is adjusted for baseline you can compute all kinds of interesting contrasts including ones that are functions of baseline. So you estimate change from a specific baseline value without subtracting baseline before doing the model.

^ | ⌄ • Reply • Share ›

Jinn-Yuh Guh • 4 years ago

Change scores are used frequently in clinical trials in nephrology, e.g. “doubling of serum creatinine”, “30% or 40% decrease of estimated GFR”, etc. Are they invalid outcome measures? Note that serum creatinine levels are used as an enrollment criteria in most clinical trials in nephrology.

^ | ⌄ • Reply • Share ›

Frank Harrell Mod ➦ Jinn-Yuh Guh • 4 years ago

I go into detail about the problems with serum creatinine in http://fharrell.com/doc/bbr... Section 14.4.2. Delta SCr and delta eGFR have multiple problems, including non-monotonic relationships with hard clinical endpoints. The "doubling" and "40% decrease" ideas were completely pulled out of thin air and not backed up by any data. The best way I know to summarize the problem is this: Two patients who have doubling of SCr but started at different summarize the problem is this: Two patients who have doubling of SCr but started at different baseline SCr will have much different outcomes. Therefore "doubling of SCr" means different things to different patients. Therefore it's not an adequate summary of two SCRs.

• Reply • Share ›

Jinn-Yuh Guh → Frank Harrell • 4 years ago • edited

These definitions of outcomes are dichotomous to help the clinicians to make dichotomous clinical decisions such as “acute kidney injury” or “chronic kidney disease stage 3”, etc. “Dichotomania” is bad for statistics, but it is useful for clinical practice. What do you suggest for a better outcome measure in clinical trials to help the clinicians make a dichotomous decision?

• Reply • Share ›

Frank Harrell Mod → Jinn-Yuh Guh • 4 years ago

I have written about this in the Diagnosis chapter of http://fharrell.com/doc/bbr.... You have described decision making in a way that clinicians do not actually practice. If you did practice that way, not only could you be easily replaced by AI, AI would make better decisions than you. And as argued in the Annals of Internal Medicine commentary "Against Diagnosis" labels such as the ones you listed can even be counterproductive. For clinical practice. I would show (as I've tweeted earlier) a heatmap relating baseline and current SCr to risk of a bad clinical outcome, and make decisions risk based. Making a decision on the basis of a ratio of SCr is easily shown to not be risk based, so will be suboptimal. As far as outcome measures in RCTs are concerned, if you were prepared to make a dichotomization using SCr, you must really like SCr, so just do a standard longitudinal analysis of serial SCr measurements and keep everything continuous.

• Reply • Share ›

Frank Harrell • 5 years ago

You have the point of view of many classical ("orthodox") statisticians and I respect that there are many different views on the subject from smart people. At this point it's probably best to move on, just concluding that you and I have different views of statistics.

• Reply • Share ›

Frank Harrell • 5 years ago

Steven you are quite mistaken in your interpretation of ORBITA and your belief that power calculations with made up standard deviation estimates remain valid after the study is quite wrong. A likelihoodist or Bayesian would say "the data are everything". (the Bayesian would also need to add "along with the prior"). Type I and II error are relevant before the data are in, i.e., are relevant in the initial planning of fixed sample size studies, and are not relevant to a single observed dataset once the data arrive.

Your simulations miss the point. Rand Wilcox and John Tukey have shown that invisible increases in the tail density from Gaussian make confidence interval coverage go awry, and in the particular case of the one-sample problem from the lognormal distribution, the confidence non-coverage probabilities in both tails are quite far from 0.025 for a putative 0.95 confidence interval even with n=50,000. Details about that were in another stackexchange post.

• Reply • Share ›

Steven McKinney • 5 years ago

As for the StackExchange question you reference, the R code provided shows comparison of two means from a lognormal density. I see no discussion of confidence intervals therein at all.

I used that code to generate 50,000 random draws of 1,000 observations from a lognormal. If I calculate the mean minus the expected value and divide by the specified standard deviation for each of the 50,000 draws and plot a density estimate, a standard normal distribution fits almost exactly on top of the density of the 50,000 standardized means. So for averages of lognormal data at a sample size of 1,000 the CLT is certainly demonstrably accurate. The example presented by the original poster of that code involves 50,000 repetitions of two samples each consisting of 50 observations.

The StackExchange discussion includes a plot showing a density curve for the simulated data t statistics with a T distribution density curve overlaid. On code line 60 there is a logical bug, dt(x, 8) so a T distribution density with 8 degrees of freedom is shown in the image instead of one with 98 d.f. If I correct the bug, the T density tails lie nicely on top of the simulation data density curve. So even for two samples of size 50 the t statistic is behaving close to advertised rates. This discussion of the behaviour of the chi-square denominator requiring hundreds of thousands of observations before it shows its CLT convergence is a red herring. The problem concerns average estimates for the mean of lognormal data, not variance estimates of the second moment of a lognormal. The CLT says

see more

• Reply • Share ›

Steven McKinney • 5 years ago

An a-priori power calculation, involving a pre-specified effect size of scientific relevance, are key ingredients in allowing proper inference after study data are collected and analyzed. There is indeed a formal way to embed the power into the interpretation of the results. References for a sound inferential procedure can be found for example in "Why perform a priori sample size calculation?" (DOI: 10.1503/cjs.018012), Deborah Mayo's "Error and Inference" book, page 263, and van Belle et al.'s "Biostatistics", page 20, 135, 146-147. If I have misunderstood the underpinnings of frequentist statistical inference, then many other statisticians and philosophers far more talented than I have misunderstood. Power is not irrelevant once a study is done, the a-priori power analysis is our best estimate of the state of affairs before the study is run and remains our best estimate, as power for a study can not be calculated using the data from the study. There are plenty of discussions of the inappropriateness of post-hoc power calculations - calculating power for a study, using the data from the study, e.g. "Some Practical Guidelines for Effective Sample Size Determination" (DOI: 10.1198/000313001317098149) Of course the standard deviation calculated from the study data will differ from the standard deviation calculated from the a-priori data used in the power calculation. Estimates have "fuzz". That does not render prior data irrelevant and wrong. How could you use a prior distribution in a Bayesian analysis under that tenet - your posterior will differ from your prior, so the prior must have been quite wrong.

Reply • Share ›

Frank Harrell • 5 years ago

Well put Mike. Deep down I think it is as my colleague Drew Levy says 'cognitive laziness' and its cousin is racism and racial profiling. Dichotomization saves time and cranial oxygen consumption.

Reply • Share ›

Mike Babyak • 5 years ago

I do think that categorization is a vestige from the days of hand calculation. I often point out that statisticians at the time understood that it was not ideal and even wrote about it--there was even a correction factor for the loss of power. At some point, that concern was forgotten.

That categorization continues is, I think, a fascinating sociology of science study. An incredible amount of rationalization goes on in the face of compelling evidence. I've given papers to colleagues, held workshops, talks, begged and pleaded one-on-one, all with apparently little effect. Fortunately, my students have taken in more of the message--and they know I'm watching!

Reply • Share ›

Frank Harrell • 5 years ago

To add one example demonstrating why power is irrelevant after the study is completed, consider a case where a certain standard deviation has been assumed in doing the power calculation, and after the study it is found that the standard deviation is actually larger by a factor of 1.3. The original power calculation is not only not relevant now, it is also quite wrong.

Reply • Share ›

Frank Harrell • 5 years ago

Steven I'm sorry to say that you have misunderstood the underpinnings of frequentist statistical inference. There is no formal way, nor need, to embed the power into the interpretation of the results. The evidence provided by the study in the frequentist context is contained in the 0.95 confidence interval, and all values within that interval would be accepted at the 0.05 significance level if tested with that value as a null value. And even though the sample mean will converge to the true population mean as n goes to infinity, that has nothing to do with claims about its precision in a finite sample. The confidence interval does that. I stand by my statement that the editors/reviewers/authors made an unjustified 'evidence for absence' claim. And Bayesian methods are dramatically different than frequentist in this setting. Bayesian methods will not allow the 'evidence for absence' error, as when you compute the probability of similarity or non-inferiority, that probability will not be large enough to support those assertions. The prior distribution for this could be a normal distribution that even gives the benefit of the doubt to similarity, and you'd still find the posterior probability is not high enough. When interpreting the results, you can almost pretend the observed mean is not there. By the way, a few simple simulations where one draws data from the lognormal distribution will show the irrelevance of the central limit theorem. A sample of size 50,000 is not large enough for the CLT to provide accuracy confidence intervals when the data have a lognormal distribution. See for example https://stats.stackexchange...

Reply • Share ›

Steven McKinney • 5 years ago

This study provided no evidence of a difference greater than the minimum difference of medical relevance. The statement "PCI did not increase exercise time by more than the effect of a placebo procedure" is a valid claim at the type II error rate for which this study was devised. I don't know what it means that an a-priori power analysis is not relevant after the fact. The a-priori power analysis is precisely the reason that we can make a declarative statement about a large p-value after the precisely the reason that we can make a declarative statement about a large p-value after the experimental results are in. The power analysis was not performed with the data from the trial.

In the presence of an a-priori power analysis, this is the proper interpretation of what a large p-value means. This study did not just run a test and comment willy-nilly on a resultant p-value. This study pre-specified a difference of medical relevance and performed a power analysis using similar previously available data to ensure that the type II error rate was known. These are the steps all too frequently not performed in other studies, that then leave such other studies unable to say anything about the state of affairs in the presence of large p-values, as the error rate of claims is then unknown. Without an a-priori power analysis, interpreting a large p-value as demonstrating no difference is a misinterpretation. This study did not misinterpret.

I do put faith in means, the law of large numbers and central limit theorem suggest that such faith is well placed, and we can know the rate at which our faith is misplaced. I understand very well that

see more

∧ | ∨ • Reply • Share ›

Frank Harrell • 5 years ago

You made some excellent points Steven, and I think the ORBITA study was carried out in excellent fashion. I wrote http://www.fharrell.com/201... to temper my original criticism of the article. However there are two things I really disagree with in your otherwise excellent comments: The authors, reviewers, and journal editor did in fact commit an "absence of evidence is not evidence of absence" error because the abstract of their article contains the phrase "PCI did not increase exercise time by more than the effect of a placebo procedure." This is a major misinterpretation of what a large p-value means (which Fisher, the inventor of p-values said means "get more data"). Had the article stopped with the confidence interval and not reported the p-value, all would have been well (other than the problem of computing change form baseline in exercise time). This misinterpretation of p-values (a very common misinterpretation in Lancet, NEJM, and JAMA) caused a firestorm among cardiologists which could have been avoided Secondly, consider your statement "measured a difference of 16.6 seconds between the two groups, well below the minimum difference ..." The problem with your reasoning is that you are putting faith in the point estimate of 16.6s when in fact that estimate has a "fuzz" or margin of error. For this purpose you need to dwell on the most favorable confidence limit, which exceeds the 30 sec. point. You also forgot that power is not relevant after the fact, so the fact that 4 out of 5 replicate studies would have found a difference >= 30s is completely irrelevant at this point. What we have now is the confidence interval from the study, and an entire Bayesian posterior distribution would have even been better to compute.

∧ | ∨ • Reply • Share ›

Steven McKinney • 5 years ago

The high profile paper published in the Lancet (the ORBITA trial) shows a very good understanding of statistical issues. The authors recognized that no truly blinded study of this medical manoeuvre had ever been done. Years of anecdotal publications litter the literature, enough to convince many who do not understand statistical issues that this trial would be unethical. What is unethical is to continue to promote ill-founded medical manoeuvres based on poorly done studies.

These authors worked hard to convince others of the errors in their thinking, and arranged for a proper blinded clinical trial. They registered their trial plan beforehand, with ClinicalTrials.gov and pre-published with the Lancet. They identified an outcome of no medical relevance (30 second difference between the two groups) and performed a power calculation using then-available data which showed that 100 cases per treatment group would provide 80% power to detect such a difference.

The authors, reviewers, and editor did not allow a classic absence of evidence is not evidence of absence error to be made. In the presence of a power analysis showing adequate sample size to detect any difference larger than that of no medical relevance, a large p-value does provide sound statistical evidence that a difference of medical relevance is likely not present, i.e. that the null hypothesis is the relevant hypothesis to accept, at the stated type II error rate. This study measured a difference of 16.6 seconds between the two groups, well below the minimum difference of medical

see more

∧ | ∨ • Reply • Share ›

Otavio Ranzani • 5 years ago

Thank you Prof Frank Harrell!

∧ | ∨ • Reply • Share ›

Frank Harrell • 5 years ago

This sort of double difference is based on a misunderstanding of what a parallel group randomized trial really is. It assumes many things including linear effect of baseline and slope of baseline of 1.0, no regression to the mean, no measure errors, etc. Simple ANCOVA would be far superior here, and would open up to robust semiparametric methods. And any time you quote a change from baseline in a paper when the baseline value is what qualified the patient to be enrolled, you need to have a second baseline to use in difference calculations.

⌃ | ⌄ • Reply • Share ›

Otavio Ranzani • 5 years ago

Prof. Frank Harrell, thank you for all your immeasurable contribution to the stats knowledge.

The "change in scores" is elsewhere, but it is quite intriguing how the authors analysed their primary outcome (a cytokine) in this RCT in JAMA today (22/08/2017). They analysed it as a mean difference (between groups), but of a change from baseline (apologies if I could not be clear), i.e.: delta from d14-d1 in group A = T; delta from d14-d1 in group B = Z, they compared the difference T - Z. Including their sample size and SAP were based on this "difference from a change".
Could you comment on it? Could we relate this analysis with an ANCOVA?

Many Thanks,

http://jamanetwork.com/jour...

⌃ | ⌄ • Reply • Share ›

Otavio Ranzani • 5 years ago

This comment has been removed by the author.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

Fantastic points. Biomarker research in general is of the same low quality as nutritional epidemiology research, and I am constantly amazed at how a research will create a potentially good prognostic or diagnostic marker then proceed to destroy it by categorization as if she actually hated the marker.

Also concerning the blog article hitting on the "milder side" I think that is true, especially when I think of the many reviews I do for medical journals. If you think the published papers are bad, the ones that don't get published are often astonishingly bad. It is often as if the authors are allergic to statisticians and don't want them anywhere near.

⌃ | ⌄ • Reply • Share ›

Unknown • 5 years ago

I could not agree more with the examples. However, they are if anything on the milder side of the scale. For example, there are plenty papers (like this one http://dx.doi.org/10.2337/d... that look at a continuous marker that leads to disease diagnosis, plot it over time relative to when it exceeds the diagnostic threshold and then claim that this proves an acceleration of the underlying disease just before people get diagnosed. I have stopped counting how many of those have by now come out in supposedly good journals... But try to get a properly done methodological paper published that discusses the problem with this kind of analysis (7 years and counting). I guess that one is still somewhat subtle. But what about combining several biased assessments and one not overly biased one with the claim that the last one fixes all the bias? You'd think that people would not make that obvious a mistake, but I have also seen a few papers like that (e.g. https://doi.org/10.2337/dia... although it is perhaps not immediately obvious that the type I error rate one produces is as high as 99%).

⌃ | ⌄ • Reply • Share ›

Gunter Kuhnle • 5 years ago

Thanks for your blog and your comments - they have given me a lot to think about (and confirmed some of my suspicions about my field).

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

My personal observation is that users of such methods tend to be afraid of algebra, and they think that categorizing a continuous variable makes fewer assumptions than just fitting a straight line (which they are comfortable with). So they tend to avoid flexible, powerful approaches such as regression splines (the ease of using splines with software packages nowadays notwithstanding). In fact categorization makes far more assumptions: the relationship is piecewise flat and the interval boundaries are correctly specified, plus there being discontinuities in nature at these boundaries (which we never see).

⌃ | ⌄ • Reply • Share ›

Gunter Kuhnle • 5 years ago

Might I ask: were quantiles used to make analyses easier at a time of limited computing power, and there is no need to do this anymore? At least on my data, using the approaches you suggest are no more difficult or time consuming and comparing the results is very interesting.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

We don't do the right thing because it's easy. We do it because it's right.

⌃ | ⌄ • Reply • Share ›

Gunter Kuhnle • 5 years ago

It's difficult to find many papers in nutritional epidemiology that don't use quantiles (at least in the field I'm working in), and so it is the method most people use. I agree that it has to be changed, but I don't think it is easy.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

It is not considered to be the standard method; it is just commonly used. It is so invalid and misleading that things MUST change. Quantile groups are also impossible to interpret clinically, as I've written about in my RMS course notes (see the Links page on this blog). If researchers want to do good science and want the research to be reproducible, statistical approaches MUST change.

⌃ | ⌄ • Reply • Share ›

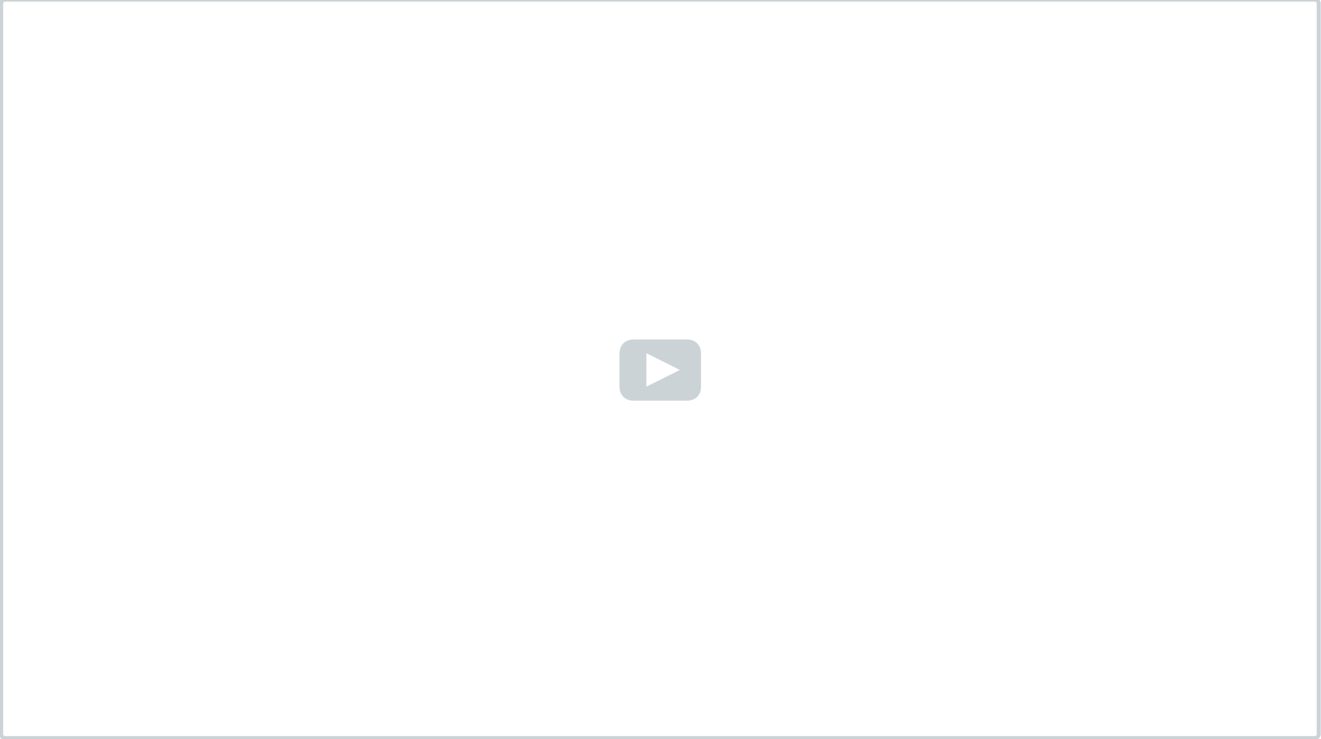
Gunter Kuhnle • 5 years ago

Thank you - that is very interesting. Given that categorisation is now considered to be the standard method, do you think this will change?

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

Errors in measurements are definitely a challenge. But what also has a large effect on the quality of research in nutritional epidemiology is the low statistical standards that have been set. The use of categorization to deal with measurement errors is a common misconception. Grouping actually makes things worse by loss of information and because of the fact that measurement errors put a subject in the wrong group, which is a maximal error. This demonstration helps illustrate these points:



The work of Ray Carroll is an excellent source for formal analysis of nutritional data accounting for measurement error.

⌃ | ⌄ • Reply • Share ›

Gunter Kuhnle • 5 years ago

Thanks for your comment on this! I think one reason for categorising appears to be the unreliability in dietary assessment methods and the difficulty of addressing this in analyses. Study participants can be classified into quantiles - and this is reasonably reliable; using absolute intake data (or scores) would be much more difficult.

How would you address uncertainty in exposure assessment here? There are not many publications who seem to do this (at least in nutritional epidemiology).

Best wishes,

Gunter

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

Good question Kris. Smoothing splines require special methods but regression splines don't. When the variable of interest is a subject-level continuous variable, the way you create terms for a regression spline is completely agnostic to other aspects of the model (random effects, semiparametric model, etc.). So I would add the nonlinear terms as main effects and as treatment interactions, whether you use regression splines or fractional polynomials.

⌃ | ⌄ • Reply • Share ›

Kris • 5 years ago

Hi Frank, I'm finding these critiques very useful and can see some immediate changes that I can make in my own work.

I had a query relating to the application of smoothing splines for subgroup analyses. I would like to do try what you've described under 'improper subgrouping' heading and estimate a non-linear interaction between age and treatment effect (which for this outcome and population is fairly likely). The difficulty is that this is from a cluster randomised trial which means the primary analysis uses a multilevel model. I have looked but can't find a smoothing splines procedure compatible with hierarchical data.

My first instinct is to ignore the clustering and find an appropriate form with a fractional polynomial analysis, then go back and add the continuous variable in this form to the main (multi-level) model and estimate the interaction with the treatment. Do you think that there is a better way to approach this?

Thanks, Kris.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

Jack I believe you are exactly correct. I have seen statisticians switch to 'precision medicine' and invent solutions that are inappropriate for clinical use (e.g., they require non-available information) and I have seen many more statisticians do decent work but not question their medical leaders about either clinical practice or analysis strategy. We are afflicted with two problems: timidity and wanting to profit from the funding that NIH and other agencies make available with too little methodologic peer review.

⌃ | ⌄ • Reply • Share ›

Jack Wilkinson • 5 years ago

Thanks for taking the time to reply Frank. I feel like we (methodologists) have a bit of a conflict of interest when it comes to these medical fads. We get grants to develop 'new' precision medicine methodology, so have an interest in promoting the hype.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

'Stratified medicine' AKA 'precision medicine' AKA 'personalized medicine' is overhyped and based on misunderstandings. Stephen Senn has written eloquently about this. There are some real differential treatment effects out there (especially in cancer and bacterial infections) but true interactions with treatment are more rare than many believe. Clinicians need to first understand the simple concept of risk magnification that always exists (e.g., sicker patients have more absolute treatment benefit, no matter what is the cause of the sickness) and why that is different from differential treatment effectiveness. Any field that changes names every couple of years is bound to be at least partly BS. See my Links page here to get to the BBR notes that go into great detail in the analysis of covariance chapter.

⌃ | ⌄ • Reply • Share ›

Jack Wilkinson • 5 years ago

Frank, if real interactions are not very common, what does that mean for stratified medicine, in your view?

By 'stratified medicine' I mean interactions with baseline characteristics, rather than 'responder analysis', which is obviously silly for reasons you've touched on in your discussion of change.

Thanks.

⌃ | ⌄ • Reply • Share ›

Frank Harrell • 5 years ago

Spot on Gary. The avoidance of statistical expertise is one of the most irritating aspects of this. And researchers don't realize the downstream problems caused by their poor analysis, e.g., requiring more biomarkers to be measured because the information content of any one biomarker is minimized by dichotomization They also fail to realize that cutpoints must mathematically be functions of the other predictors to lead to optimum decisions. I'm going to edit the text to point to my summary of problems of categorization on our Author Checklist. I need to add your excellent paper to the list of references there too.

⌃ | ⌄ • Reply • Share ›

Gary Collins • 5 years ago

I dont think many of the dichotomisers understand what they're doing, they've collected data, only to throw a portion of it away by dichotomising continuous measurements (and then they'll throw more data away, by randomly splitting their data into development and validation data). The consequences of dichotomising is there for all to see, a substantial loss of predictive accuracy - both a drop in discrimination and poor calibration (see http://bit.ly/2pfbZiw). Plenty of systematic reviews show this is poor behaviour is widespread (http://bit.ly/2qkjR2x, http://bit.ly/2ptKRIN). Prediction models are typically developed without statistical input, and peer review (again often a statistician is not reviewing) is clearly failing to pick this up.

BW
Gary

^ | ˅ • Reply • Share ›

Frank Harrell • 5 years ago

Great comments David and you raise a lot of good points. Stratum-specific treatment effects are inappropriate for many reasons. Interaction effect estimation is what is needed, and such assessment must be done in the presence of aggressive main effect adjustment. I have detailed this in my BBR notes - see http://www.fharrell.com/p/b... section 13.6. I slightly disagree with your statement 'is largely disregarded'. In my experience real interactions on an appropriate relative scale are not very common.

^ | ˅ • Reply • Share ›

David McAllister • 5 years ago

Hi Frank,
What do you think about the interpretation of reported treatment-covariate interactions?
This is a big interest of mine as I am doing a 4-year project looking at heterogeneity of treatment effects (on various scales) according to the presence or absence of comorbidities.
I think that while the dangers of sub-group reporting are appreciated by most, the potential that some treatment have lower efficacy (on some relative scale) is largely disregarded.
Moreover, there is a tendency to report stratum-specific hazard ratios along with some NHST P-values for the interaction which is often close to one, which many readers take as evidence for no interaction. When I have used the published data to estimate confidence intervals for interactions, the lower and upper limits have been consistent with massive interactions in either direction.
best wishes,
David

^ | ˅ • Reply • Share ›

Frank Harrell • 5 years ago

You are right that totality of evidence needs to be considered. Frequentist methods don't help very much in that regard. I am concentrated on the primary endpoing for simplicity. I disagree that a 35% effect is realistic for powering a study. When investigators choose a low-power binary endpoint they have the obligation of finding the money to make that endpoint 'work' by enrolling a much larger number of patients. Relative efficacy of 15% or 25% are commonly used in power calculations, and we need at least 0.9 power, not 0.8. Many papers have been written showing that for common disorders, relative effects smaller than 15% will still result in large public health benefits. And your reasoning to justify type II errors being 20 times less important than type I is not compelling. You've only justified that type II errors may be allowed to be larger than type I errors. I feel that too many resources are wasted by studies being launched on "a wing and a prayer."

^ | ˅ • Reply • Share ›

Steven McKinney • 5 years ago

"A big part of the issue in this particular result is that the investigators thought that type I error was 20 times more important than type II error. How does that make any sense?"

Type I error: We declare the drug to be effective when in fact it is not.

Type II error: We declare the drug to be not effective when in fact it is.

The onus is on the drug developers to demonstrate a marked effect for the new drug. Declaring a sugar pill to be something marvellous is expensive. How much money are we spending right now on this drug? How much does a course of this drug cost? If it is not doing anything, we are wasting money that could be better spent on other treatments with more efficacious outcome. We are also wasting money and resources treating people for side effects for a drug that isn't offering much help. Adding more placebos to our formulary is bloating our health care budget unnecessarily. It yields more profits for drug companies, which they love, but yields a health care system that is a cancer on the national economy. So at a time when health care costs are out of control, we need tough tests of treatment effectiveness. We need to trim out any treatment that isn't showing a large benefit.

Declaring a drug to be not effective when in fact it is is of course tragic. The cost then is in quality of

see more

^ | ˅ • Reply • Share ›

Frank Harrell • 5 years ago

Steven you are right that this study is better done than many studies and it is good to see 'negative' studies published (only the more expensive multi-center clinical trials tend to be published when 'negative'). But why did you omit the confidence interval of [0.60, 1.54] from your comment? That is 'negative']. But why did you omit the confidence interval of [0.66, 1.54] from your comment? That is the most important piece of frequentist evidence reported in the paper. The confidence interval tells all. We know little more after the study than we did before the study about the relative efficacy. Notice that the trial was designed to detect a whopping 35% reduction and the lower confidence interval was only a 34% reduction. A big part of the issue in this particular result is that the investigators thought that type I error was 20 times more important than type II error. How does that make any sense? After a study is completed, the error rates are not relevant and the data are. Note also that we are not 80% confident that the drug is not doing much. Not only are the data consistent with a 30% reduction in odds of a primary event, but the power is not relevant in the calculation of this probability. What would be needed for you to make that statement is a Bayesian posterior probability of non-efficacy (one minus probability of efficacy). On the non-relevance of error rates see papers by Blume, Royall, etc. One statistician (I wish I remember to whom this should be attributed) gave this analogy: A frequentist judge is one who brags that in the history of her court she only convicts 0.03 of innocent defendants. Judges are supposed to maximize the probability of making correct decisions for individual defendants. Long-run operating characteristics are not useful in interpreting results once they are in. A side issue is that confidence intervals, though having a formal definition that will seem to be non-useful to most, have the nice property that that are equally relevant no matter whether the p-value is 'significant' or not.

^ | v • Reply • Share ›

Steven McKinney • 5 years ago

The study was registered at ClinicalTrials.gov
https://clinicaltrials.gov/...

The study pre-specified a difference of medical relevance - 35% reduction in odds ratio - and type I and type II error rates (1% and 20% respectively).

http://www.sciencedirect.co...

"Statistical power and sample size considerations

The sample size is based on an assumed composite primary end point event rate (death, MI, dialysis, mechanical assist) of 32% for placebo, a 35% relative reduction for levosimendan (20.8% event rate at 30 days), a significance level of .01, and 80% power. A total sample size of 760 should provide 201 events."

They report measured outcomes estimates, confidence intervals, in addition to p-values, as recommended in the ASA statement on p-values.

see more

^ | v • Reply • Share ›

Load more comments