To p or not to p?

Reflections on p-value Statements

Statistically Speaking Series

Mary J. Kwasny, ScD

Northwestern Medicine | Northwestern University Feinberg School of Medicine

BUT FIRST…

Q: Where does one go to find a statistician?
(A: a **bar** *chart*)

Northwestern Medicine | Northwestern University Feinberg School of Medicine

Biostatistics Resources

Northwestern Medicine | NUCATS Clinical and Translational Sciences Institute

Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Stanley Manne Children's Research Institute

Ann & Robert H. Lurie Children's Hospital of Chicago

Shirley Ryan Abilitylab

Biostatistics Collaboration Center (BCC)

- Supports **non-cancer** research at NU
- Provides investigators an initial 1-2 hour consultation subsidized by the FSM Office of Research

Biostatistics Research Core (BRC)

- Supports **Lurie Children's Hospital affiliates**
- **Stanley Manne Research Institute at Lurie Children's**

Quantitative Data Sciences Core (QDSC)

- Supports all **cancer-related** research at NU
- Provides free support to all Cancer Center members subsidized by RHLCCC
- Grant

Northwestern University Data Analysis and Coordinating Center (NUDACC)

- Supports prospective, multicenter research
- Spans the full life cycle of research
- Grant

Northwestern Medicine | Northwestern University Feinberg School of Medicine

Contact Information

Non-cancer

- Biostatistics Collaboration Center (BCC)
  - Website: http://www.feinberg.northwestern.edu/sites/bcc/index.html
  - Email: **bcc@northwestern.edu**
  - Phone: 312.503.2288

Cancer

- Quantitative Data Sciences Core (QDSC)
  - Website: https://www.cancer.northwestern.edu/research/shared-resources/quantitative-data-sciences.html
  - Email: **qdsc_rhlccc@northwestern.edu**
  - Phone: 312.503.2288

Lurie Children's

- Biostatistics Research Core (BRC)
  - Website: https://www.luriechildrens.org/en/research/research-areas/clinical-research/biostatistics-research-core/
  - Email: **mereed@luriechildrens.org**
  - Phone: 773.755.6328

Data Coordinating

- Northwestern University Data Analysis and Coordinating Center (NUDACC)
  - Website: https://www.feinberg.northwestern.edu/sites/nudacc/
  - Email: **nudacc@northwestern.edu**

Northwestern Medicine | Northwestern University Feinberg School of Medicine

Special Thanks to

- Ron Wasserstein, Executive Director of the American Statistical Association
  - Who shared a fabulous slide deck and gave time to talk through some of my thoughts…
- The many statisticians (including RW) who were involved in the ASA Statement of P-values
- The many more statisticians (including RW) who were involved in The American Statistician's special issue on p-values and significance levels both as editors, writers, etc..
- None of whom were given any time to look at this presentation and find any misinterpretations or mistakes on my part…

DISCLAIMER

The views expressed in this presentation are my own and do not represent the opinions of any entity whatsoever with which I have been, am now, or will be affiliated.

Northwestern Medicine | Northwestern University Feinberg School of Medicine

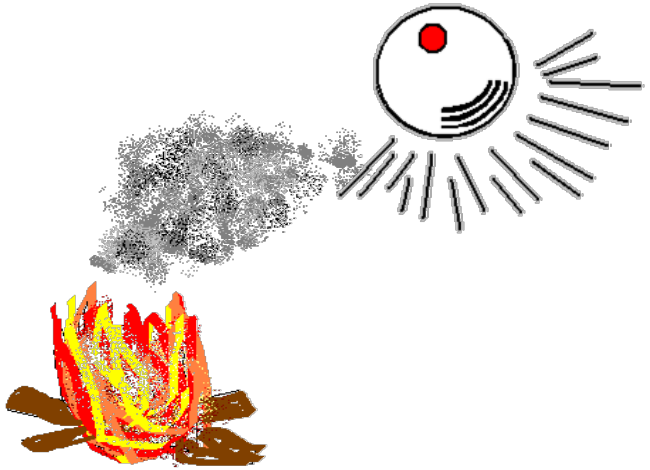
Outline

- Basic information about p-values
- The origins of “Statistical Significance”
- Over-estimating significance and Misinterpreting p-values
- Historical perspectives of the “controversy”
- 6 principles of ASA statement
- Reactions
- 6 good practices to consider

Where did p-values come from?

A quick look at hypothesis testing, and the history of p-values…

"testing" [as we know it] quick review

		Reality	
		No difference in Groups/Treatment *(Ho true)* All clear!	There is some Group Effect *(Ho not true)* FIRE!!
Test Result	Reject Ho ***(p < 0.05)*** ALARM!	Type I Error (α) α= 0.05 (5%)	Power 0.80 (80%)
	Fail to reject Ho ***(p > 0.05)*** All clear!	Confidence 0.95 (95%)	Type II Error (β) 0.20 (20%)

Note: COLUMN probabilities add to 1 –Testing conditions on a certain Reality!

So, what's a p-value?

- We know some stuff
- We want to know some more
- We design an experiment to help us
- We collect data from the experiment
- We summarize the data from an experiment into a number we call a "statistic"
- Compute a probability from that statistic – that's the p-value
- If the p-value is small enough, we call it "statistically significant"
- What is small enough? Pre-set threshold (what rate do you want a fire alarm to go off if there is not a fire?); typically, 0.05.

Nuances

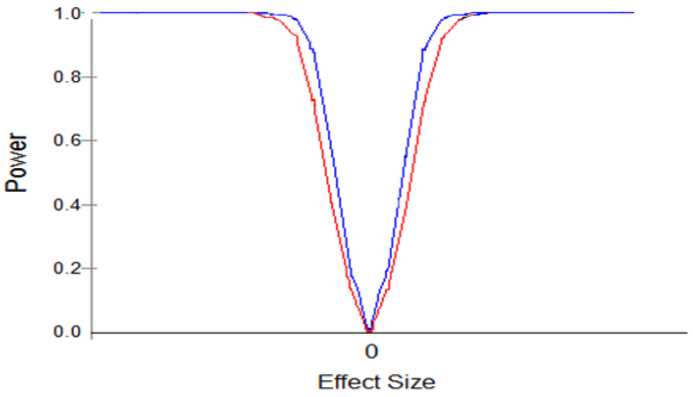
- p-values are random variables that assume values between 0, 1.
- "statistical significance" is determined once the p-value is smaller than a pre-specified "type I error rate" (historically 0.05)
- Many studies (mostly RCTs) have been designed (calculated N) to detect a certain effect size (e.g. a clinically meaningful difference in groups), at a certain POWER based on the variables and treatments under consideration, the sampling scheme and study design. (e.g. how big a fire?)
- All of this is good (I think), but hypothesis testing (and statistical significance) seems to have gotten WAY out of control.
- Some excess has been mitigated by things like CONSORT guidelines.

Small effect of treatment

Large sample needed

Large effect of treatment

Small sample needed



And to actually COMPUTE a p-value…

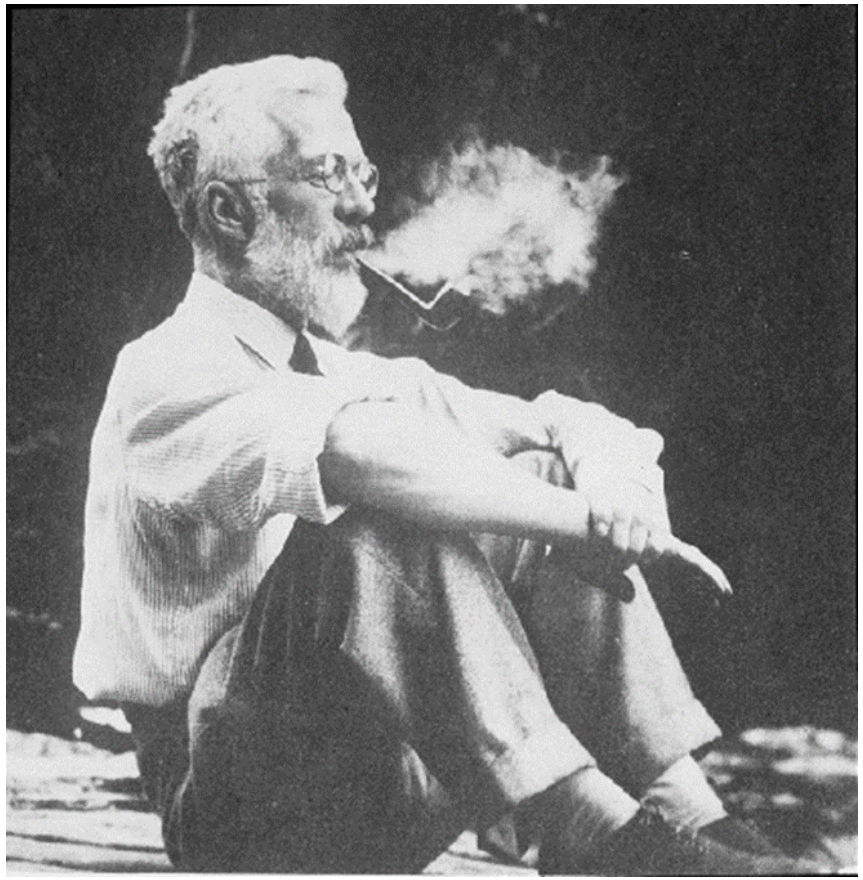
- We assumed a **bunch** of stuff
- One key assumption: No difference in Groups/Treatment (i.e. our NULL hypothesis)
- But, failure of any assumption affects the p-value

- Historically, p-values were calculated as early as 1710. (John Arbuthnot studied boys and girls born over 82 years to determine boys were more likely to be born than girls)
- The name "p-value" was first used by Karl Pearson in 1902.
- But "significance"… well, that didn't start until 1925.

Where did statistical significance come from?

And why we all drank the cool aid…

Fisher and Statistical Significance



sig·nif·i·cant

/sigˈnifikənt/

adjective

1. **sufficiently great or important to be worthy of attention; noteworthy.**
   "a significant increase in sales"
   synonyms: notable, noteworthy, worthy of attention, remarkable, important, of importance, of consequence, signal; More
2. **having a particular meaning; indicative of something.**
   "in times of stress her dreams seemed to her especially significant"

To Fisher, "Statistical Significance" meant that the result was worth further scrutiny.

"Just a Theory": 7 Misused Science Words. Scientific American LiveScience April 2, 2013.

#6. Another word that sets scientists' teeth on edge is "significant."

"That's a huge weasel word. Does it mean statistically significant, or does it mean important?" said Michael O'Brien, the dean of the College of Arts and Science at the University of Missouri.

In statistics, something is significant if a difference is unlikely to be due to random chance. But that may not translate into a meaningful difference, in, say, headache symptoms or IQ.

Yes, but ironically… not quite. Things that happen 5% of the time, are not necessarily "unlikely"… (something that happens 5% of the time… kinda might happen 1.5 times per month.. or 18 times a year… but I digress…) Dr. O'Brien does have a point, though.. statistical significance does not necessarily imply clinical relevance.

Hypothetical example…

- The National Vital Statistics System stated the rate of preterm births was 10% in 2018.
- Imagine a study that wants to show a decrease in preterm births through [insert some amazing intervention].
- What rate would be an clinically meaningful decrease? (obviously the lower the rate the better!) – when going to a statistician to design the study, this is typically one of the first questions we ask…

What rate would be an clinically meaningful decrease?

- Here you could consider what other interventions have accomplished, or the likelihood of how impactful your intervention could be, balanced by the difficulty of adherence/compliance… there may be other things that make this "relevant"
- Let's say the PI really thought he could get it down to 8%… (vs 10% in a control arm), … possibly as that would impact the bottom line financials at the hospital, or some other reason.
- A POWER calculation showed that a sample size of 6426 (3213/group) would have 80% power to detect that difference.
- (i.e. if in fact the REALITY was that the population of women who get this intervention had a mean rate of 8% and the population of women who did not get the intervention had a mean rate of10%… sampling variability accounted for… the investigator would have an 80% chance to see a "statistically significant" result – blah, blah, blah.)

What happened?

- PI was funded (this is my imagination), and randomized a sample of 6426 to intervention or usual care (3213/group). No drop out (vivid imagination)… and 273 intervention and 321 usual care had preterm births.

- Preterm birth rate in intervention was 8.5%, rate in usual care was 10%. This difference was statistically significant (p=0.043). Authors write up paper. Everyone is happy. Intervention widely implemented.

What *really* happened?

- Question: Are interventions ever widely implemented *without* further evaluation (or should those results be seen with scrutiny)?

- What if, in reports to the DSMB, it was noted that among the full-term births, the rate of c-sections in the intervention group was 34.5% and 32% in usual care. This was also statistically significant (p=0.042). Should the intervention be pulled?

- While it is important (vital) to monitor interventions for [long term] adverse effects – could we consider the c-section result "worthy of further scrutiny" – also, this was *not* a planned comparison. If the study was half of the size that it was, these same rates would NOT be statistically different (p=0.169). Was there a biologic reason that this intervention may have increased the likelihood of c-sections?

What's the problem?

- This $p<0.05$ "bright line thinking" is not in the best interest of science.
- While having decision tools is helpful, we should also rely on our own (and others') experiences to determine what has clinical relevance… and to determine if any result merits further scrutiny.
- Think about bright line thinking in a completely ridiculous situation…

Imagine if Tinder had an algorithm…

- My experimental results are interesting. I should spend more time with them, maybe repeat the experiment. I may be on to something, but it will take time to be sure.
- You tiny, beautiful p-value. You are the result that I want to spent the rest of my life with. Let's publish and get grants together. I love you!

Another fundamental problem

- Most "early" statisticians were mathematicians… (I was a mathematician).
- One form of proof we learn *and some of us loved* was PROOF BY CONTRADICTION.

- Basically, you posit something, follow a natural course using mathematical laws, and eventually come to a contradiction (like 2=1 or something). Since everything else you did was legit, the thing you posited must be wrong. This is more or less what hypothesis testing emulates. But, a lot of stuff gets assumed (in a null hypothesis).. and rather than a stark contradiction, you aim to get a small probability… so, it is likely that your hypothesis is wrong. (As previously discussed, we calculate the probability of the data given the null hypothesis)

Unfortunately, most of us want the reverse –

- the probability that the null is false, given the data.
- More unfortunately, these are often COMPLETELY different.

		Reality	
		No difference in Groups/Treatment *(Ho true)*	There is some Group Effect *(Ho not true)*
Test Result	Reject Ho *(p < 0.05)*	Type I Error (α) α= 0.05 (5%)	Power 0.80 (80%)
	Fail to reject Ho *(p > 0.05)*	Confidence 0.95 (95%)	Type II Error (β) 0.20 (20%)

Conditional on what!?

- Most MDs are familiar with the ideas of Sensitivity vs Positive Predictive Value in disease screening.
- Sensitivity has been calculated in a lab, using people scientists KNOW to have a certain condition or disease.
- **P(test +|disease +)=sensitivity**. (like fire alarm quality checks)
- (Similarly, specificity=Pr(test -|disease -) was calculated on a bunch of people scientist know to NOT have the disease)
- Positive predictive value is what you can use to *reassess* if a patient has a certain disease knowing they tested positive for it.

Pr(disease + |test +)=Positive predictive value.

Conditional on what (example)!?

- Tests for sickle cell anemia have a sensitivity of 98% and specificity of 99%. Prevalence of Sickle Cell Disease in Hispanic Americans is about 1/16300=0.00006
- So, the sensitivity= **Pr(Test+|Disease+)=0.98**, but if a Hispanic American tested positive, the probability they had the disease, or the positive predicted value is
- $\mathbf{Pr(D+|T+)=}\frac{Pr(T+|D+)Pr(D+)}{Pr(T+|D+)Pr(D+)+Pr(T+|D-)Pr(D-)}\mathbf{=0.006.}$
- Hypothesis testing is privy to the same type of confusion.

Pr(data|null hypotheses)≠ Pr(null hypothesis| data)

Why has this gotten out of control?

- p-values are SO easy to calculate these days.
- We LIKE having an objective threshold to make decisions.
  - In many ways it is easier to not have to decide if a 34.5% rate is "higher" than a 32% rate from a clinical standpoint. While we know there is variability, etc., etc. we are still (generally speaking) a species that just doesn't seem to fundamentally have a feeling for risks and rates.
- BIG data caused a big stir about FDR (false discovery rates), and more emphasis was put on adjusting p-values for multiple comparisons, etc., etc…
- And/Or.. Maybe it's just that we are noticing it more…

Historical context.. (well 1925-1985)

- As mentioned 1925 was the date “statistical significance” was more or less coined.
- Between 1933 and about 1955 Fisher v. Neyman & Pearson (and Hill & Doll)… debated.
- 1937, William Gosset (a.k.a. Student) told Pearson fixed values of significance (i.e. $p<0.05$) was “nearly valueless”
- Leonard Savage noted in 1954, “statistical significance tells us what to say but not what to do”
- Beale (1972) “What’s so significant about 0.05” – American Psychologist
- Greenwald (1975) – Consequences of prejudice against the null hypothesis – Psychological Bulletin
- Pratt (1976), quoted by Yoccuz (1991) "The author believes that tests provide a poor model of most real problems, usually so poor that their objectivity is tangential and often too poor to be useful." Cox (1977) "As a second example, consider significance tests. They are also widely overused and misused."
- Preece (1982) "In analysis, overemphasis on significance testing continues."
- Parkhurst (1985) "Failing to reject a null hypothesis is distinctly different from proving a null hypothesis; the difference in these interpretations is not merely a semantic point. Rather, the two interpretations can lead to quite different biological conclusions….“

Recent historical context.. (well 1986-2000)

- Cox (1986) "The continued very extensive use of significance tests is alarming.“ … and …"It has been widely felt, probably for thirty years and more, that significance tests are overemphasized and often misused and that more emphasis should be put on estimation and prediction.”
- Tukey (1991) "Are the effects of A and B different? They are always different---for some decimal place."
- Cohen (1994). The earth is round ($p<0.01$)
- Robinson (1997) – Reflections on statistical and substantive significance
- Rigsby (1999) Getting past the statistical referee: Moving away from p-values and towards interval estimation
- Approximately 400 references (this number could be quite low) now exist in the quantitative literature than warn of the limitations of hypothesis testing. Harlow et al. (1997) provide a recent [sic] edited book entitled, (Lawrence Erlbaum Associates, Publishers, What If There Were No Significance Tests? London)

When I started to realize what was going on…

- Ziliak and McCloskey. The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives (2008) *University of Michigan Press.*
- Goodman 2008 "A Dirty Dozen: Twelve p-value Misconceptions" *Seminars in Hematology*.
- Seife 2011 "The Mind-Reading Salmon: the true meaning of statistical significance" *Scientific American*
- Feb 2014: George Cobb launched the idea that the ASA should come up with a policy statement on p-values.
- March 2015: Basic and Applied Social Psychology declared hypothesis testing "invalid" and banned p-values
  - October 2015: P-value panel meets at ASA office… March 2016: P-value statement published
  - October 2017: Symposium on Statistical Inference
  - March 2019: Special issue TAS was published.

Statement, symposium, and special issue…

"The expectation is that the symposium and special issue of TAS will lead to a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical science, and science itself, into a new age."

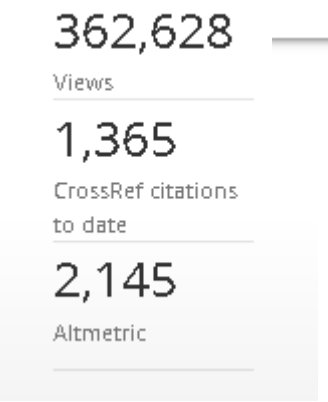
The p-value panel

- Naomi Altman
- Jim Berger
- Yoav Benjamini
- Don Berry
- Brad Carlin
- John Carlin
- George Cobb
- Marie Davidian
- Steve Fienberg
- Andrew Gelman
- Steve Goodman
- Sander Greenland
- Guido Imbens
- John Ioannidis
- Valen Johnson
- Michael Lavine
- Michael Lew
- Rod Little
- Deborah Mayo
- Chuck McCulloch
- Michele Millar
- Sally Morton
- Regina Nuzzo
- Hilary Parker
- Kenneth Rothman
- Don Rubin
- Stephen Senn
- Uri Simonsohn
- Dalene Stangl
- Philip Stark
- Steve Ziliak

I heard there was a lot of fighting, and discussion and … well, has anyone here gotten different advice from just 2 different statisticians?!? Or perhaps a statistician and a statistical reviewer?? (imagine 32!)

The ASA Statement – six principles

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.



The TAS special issue…

The big change

"Moving to a World Beyond p<0.05"
https://amstat.tandfonline.com/doi/full/10.1080/00031305.2019.1583913#.XYjKQ25FxPY

"Scientists rise up against statistical significance"
https://www.nature.com/articles/d41586-019-00857-9

Saying farewell to "statistically significant"

Why get rid of “statistical significance?”

- “Significance” doesn’t mean what people thinks it means
- Bright line thinking can lead to bad science and bizarre behavior
- Decades of debate (complaining) have done little to nothing
- Last, but not least, publication bias has potentially eliminated lines of research that could have been promising…or at least has not helped forward scientific thinking and theory…

To be Clear

- This recommendation is NOT to abandon p-values
- “A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.”
- “Despite the limitations of p-values (as noted in Principles 5 and 6 of the ASA statement), however, we are not recommending that the calculation and use of continuous p-values be discontinued.”

.

We know… change won’ t be easy..

“The basic explanation is neither philosophical nor scientific, but sociologic; everyone uses them. It’s the same reason we can use money. When everyone believes in something’s value, we can use it for real things; money for food, and p-values for knowledge claims, publication, funding, and promotion. It doesn’t matter if the p-value doesn’t mean what people think it means; it becomes valuable because of what it buys.”

- Goodman 2019 (TAS)

So, now what?

Where do we go from here?

What are we supposed to do?

- The ASA (panel, and authors of the many articles in the TAS) suggest the **ATOM** Principle.
- **A**ccept uncertainty - To accept uncertainty requires that we treat statistical results as being much more incomplete and uncertain than is currently the norm
- Be **T**houghtful -look ahead to prospective outcomes. What magnitudes of differences, odds ratios, or other effect sizes are practically important?
- Be **O**pen - Remember that one study is rarely enough. Is there *really* any "groundbreaking new study?" Encourage pre-registration of studies, transparent and complete research practices. (in some cases, can you get code/data to be shared?)
- Be **M**odest – Encourage others to reproduce work (not just analyses!)

What have other Journals opted to do or share?

- At least two major journals (JAMA and NEJM) have had editorials by statisticians to discuss these new guidelines *by statisticians who sat on the panel and/or wrote articles in the TAS issue*. (remember what I said about statisticians not necessarily always agreeing?)

- JAMA headline:

The Importance of Predefined Rules and Prespecified Statistical Analyses Do Not Abandon Significance

- NEJM headline

EDITORIAL

New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D., Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D., Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H

JAMA (Ioannidis, 2019)

- "Banning statistical significance while retaining p-values will not improve numeracy and may foster statistical confusion and create problematic issues with study interpretation, a state of statistical anarchy…
- Uniformity… makes it easier to compare like with like and avoid having some… effects be more privileged that others in unwarranted ways.
- Without clear rules… science and policy may rely less on data and evidence and more of subjective opinions…"

- **Other panel members' reaction:** That was not the message.
- The Statement is not suggesting anything about entering into a statistical anarchy. Simply, that Statistical Significance does not have the *gravitas* that has come to be associated with it.
- .. p-values do not compare "like with like" either..

NEJM (2019)

- "The new guidelines discuss … replace P values with estimates of effects or association and 95% confidence intervals when neither the protocol nor the statistical analysis plan has specified methods used to adjust for multiplicity….
- The use of P values to summarize evidence in a study requires, on the one hand, thresholds that have a strong theoretical and empirical justification and, on the other hand, proper attention to the error that can result from uncritical interpretation of multiple inferences….
- A well designed randomized or observational study will have a primary hypothesis and pre-specified method of analysis, and the significance level from that analysis is a reliable indicator of the extent…"

- **Other panel members' reaction:** Not happy with this either
- While as a practicing statistician I agree with some of this (especially with well designed RCTs, and NOT reporting p-values for secondary outcomes), the authors extend this to observational studies and make some blanket statements that really are not true. (e.g. is there theoretical and empirical justification for a type I error of 5%, or is it simply historical and convenient?).

Back to JAMA…

- NEJM had guidelines to authors, so JAMA had to step it up, and produce guidelines themselves, but basically didn't say much new…
  - They acknowledge the debate.
  - Make sure you are clear about clinical importance vs. statistical significance
  - Report effect sizes
  - "The results for the pre-specified primary outcome take priority and precedence over all other outcomes, should be the focus of the submitted manuscript, and should be reported in detail."
  - And so on…

Change is HARD!

- In latest presentations, the ATOM recommendation has expanded a bit to ATOMIC recommendations.
- IC = Institutional Change.
- Institutional Change is REALLY HARD.

- Statisticians have not necessarily helped. But I think that part of that is the lack of understanding of the issues, and I'm encouraged that there are more and more webinars, and talks at Statistical Conferences to get the discussion moving.

Moving beyond statistical significance good practices to consider (1)

ASA Panel and TAS recommendations

- Lead with effect estimates and associated measures of uncertainty, such as interval estimates.
- Focus on the substantive implications of these estimates. Do the interval bounds have qualitatively different practical consequences? Do not focus on whether the interval includes the null value (0 for a difference, 1 for a ratio).
- Discuss your estimates in the context of prior evidence, plausibility of causal mechanism, study design and data quality, etc.

Moving beyond statistical significance good practices to consider (2)

ASA Panel and TAS recommendations

- Present p-values as continuous—not categorized—values, and do so for several hypothesized effects (for example, the null value and the minimum practically important effect size). Consider plotting the entire p-value function, an option in some software packages. [*I'm not entirely sure about that…*]
- Recognize that p-values are uncertain. Interpret them as descriptive measures of compatibility between your data and your statistical model, which includes not only a test hypothesis but also many other assumptions, such that the failure of any one assumption could account for a low p-value.
- Most importantly, avoid exaggerated claims in either direction about the importance of an effect or the lack thereof.

In closing…

- P-values are not inherently bad… and only some statisticians (Bayesians, mostly) are talking about "getting rid of them" (they have also proposed various different metrics to use, but really, all have the same issues that p-values have…so I did not get into those today!) but keep in mind they condition on a "reality" that may or may not be true.

- We know that this will take time… change is hard… and there will be a lot of false starts… and a lot of extreme opinions – keep thinking ATOMIC!!!

- Statisticians are trying to make institutional changes to improve science. We are ready to fight reviewers/editors who want to keep the *status quo*, and we have a growing list of citations behind us. Please give us a chance!!

As this was compiled over break…

Statistician delivery:

P=0.041

What is it?

It’s statistically significant!!

But what is it?

It’s a major award!!

Statistically Speaking: Upcoming Lectures

We hope to see you again!

Wednesday, March 18

Biostat Basics: Some Practical Things to Know
Nina Srdanovic, MS, Statistical Analyst, Division of Biostatistics, Department of Preventive Medicine

Monday, May 11

Logistic Regression: Odds & Ends
Lauren Balmert, PhD, Assistant Professor, Division of Biostatistics, Department of Preventive Medicine

All lectures will be held from Noon to 1 pm in Baldwin Auditorium, Robert H. Lurie Medical Research Center, 303 E. Superior St.

http://www.feinberg.northwestern.edu/sites/bcc/education/lecture/2019.html