Phenotype Screening
CORPORATION
enabling discovery

2016 Hot Topic Workshop : Final Agenda

Small-Sample-Size Statistics in Agriculture; How to Maximize Business Value

Doctor, It Hurts When I *p*

Ronald L. Wasserstein, Executive Director, ASA

November 3, 2016

The Talk

- They think they know all about it already, because they learned about it from others like them.
- It is not nearly as interesting as they thought it would be.
- They've stopped listening before you've stopped talking.
- Chances are, they now understand it even less.

Does “screen time” affect sleep habits of school age children?

SLEEP HEALTH
JOURNAL OF THE NATIONAL SLEEP FOUNDATION

Interactive vs passive screen time and nighttime sleep duration among school-aged children

Jennifer Yland, BA Candidate, Stanford Guan, MPH, Erin Emanuele, MPH, Lauren Hale, PhD

Received: February 17, 2015; Received in revised form: June 22, 2015; Accepted: June 24, 2015; Published Online: August 13, 2015

DOI: http://dx.doi.org/10.1016/j.sleh.2015.06.007

The researchers had hypotheses, based on previous research

- “We hypothesized that use of any form of electronic media would be negatively associated with sleep duration.”
- “Furthermore, we expected that the strength of the association would vary based on the level of interactivity of the screen type.”
- “More specifically, we hypothesized that interactive forms of screen time, such as computer use and video gaming, would be associated with shorter bedtime sleep duration compared to passive forms of screen time, such as watching television.”

Why were they interested?

- Lack of sleep (insufficient sleep duration) increases risk of poor academic performance as well as certain adverse health outcomes

- Is there a relationship between weekday nighttime sleep duration and screen exposure (television, chatting, video games)?

Who were the subjects?

“We used age 9 data from an ethnically diverse national birth cohort study, the Fragile Families and Child Wellbeing Study, to assess the association between screen time and sleep duration among 9-year-olds, using screen time data reported by both the child (n = 3269) and by the child's primary caregiver (n = 2770).”

Fragile Families and Child Wellbeing Study

- “The FFCW is a longitudinal cohort study that has followed approximately 5000 children, born between 1998 and 2000, since birth. Data were collected in 20 cities with populations of at least 200,000 across the United States. The sample was designed to include a high number of unmarried parents and racial minorities, along with a high proportion of low socioeconomic status.”

		NONE	HALF AN HOUR OR LESS PER WEEKDAY	MORE THAN HALF AN HOUR BUT LESS THAN AN HOUR PER WEEKDAY	1 -2 HOURS PER WEEKDAY	MORE THAN 2 HOURS PER WEEKDAY	REF	DK
D1A.	Hang out with friends? Do you spend no time at all, spend half an hour or less per weekday, more than half an hour but less than an hour per weekday, 1-2 hours per weekday, or more than 2 hours per weekday?	0	1	2	3	4	-1	-2
D1B.	Hang out with family members?.....	0	1	2	3	4	-1	-2
D1C.	Do household chores or help at home?	0	1	2	3	4	-1	-2
D1D.	Spend time on the computer doing school work?.....	0	1	2	3	4	-1	-2
D1E.	Spend time on the computer chatting or instant messaging with friends?	0	1	2	3	4	-1	-2
D1F.	Spend time on the computer or TV playing computer games?	0	1	2	3	4	-1	-2
D1G.	Spend time watching TV and movies?	0	1	2	3	4	-1	-2
D1H.	Attend practice or lessons or an after-school Program?.....	0	1	2	3	4	-1	-2

I12 How many hours of sleep a night does {CHILD} usually get during the week?

|___|___|
ENTER NUMBER OF HOURS A NIGHT
OR
REFUSED..... -1
DON'T KNOW..... -2

I3. Now think for a moment about a typical weekday for your family, including daytime and evening hours. How much time would you say {CHILD} spends watching television or watching videos on TV, either in your home or somewhere else?

IF LESS THAN 1 HOUR PER WEEKDAY, CODE AS ZERO.

PROBE: Do not count time {he/she} spends playing video games on TV.

|___|___|
ENTER HOURS PER WEEKDAY
OR
REFUSED ..... -1

DON'T KNOW ..... -2

What did the researchers find?

- Children who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day ($P<.001$) by about 11 minutes.
- Children who spent more than 2 hours per day of chatting on the computer had shorter sleep duration than those who chatted less than 2 hours/day ($P<.05$) by about 16 minutes.
- The researchers did not find a significant association between playing videogames/working on the computer for more than 2 hours per day and weekday nighttime sleep duration

When the researchers adjusted for other factors

- Children who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day ($P<.05$) by about 6 minutes.
- No other significant associations found.

This is a fairly typical type of study

- Typical scientifically
- Typical statistically
- Atypical communication

Unfortunately, it makes all-too-typical mistakes

To understand these mistakes, let’s describe the null hypothesis significance testing procedure (NHSTP).

What is the null hypothesis significance testing procedure?

- Question(s) posed
- Data collected

What is the null hypothesis significance testing procedure (NHSTP)?

- Evidence from the data regarding the research question is summarized in a specific way:
  - Compute a “statistic” that measures the question of interest.
  - Compute the probability that statistic would be as “large” as it is or even larger UNDER THE ASSUMPTION that there is no effect (in this case, of TV watching on sleep duration).
  - This assumption of no effect is called the “null hypotheses.”
  - The probability computed is called the “p-value.”

What is the null hypothesis significance testing procedure (NHSTP)?

- If the p-value is “small enough,” the researcher concludes there is a “significant effect.”
- “Small enough” has come very commonly to mean $P<.05$.

Certain assumptions must be made to compute a p-value

- An underlying statistical model
- Many things related to that model (randomness, representativeness, missing data, and so on)
- The null hypothesis

In terms of this example:

- The null hypothesis (informally stated) is that sleep duration is not associated with screen time (of various types).
- That is, when we calculate the p-value, we assume the answer to our question is no, there is no association between screen time and sleep duration.
- The p-value is calculated based on the assumption that there is no effect.
- *The p-value is calculated based on the assumption that there is no effect.*
- **The p-value is calculated based on the assumption that there is no effect.**

What's the logic?

- If the p-value is small, this means that it is relatively unlikely that we would have seen the data we saw if all the assumptions were true.
- So, we either had bad luck (random error), or one or more of the assumptions may not be true.
- One of those assumptions, the assumption of no effect, is commonly THE assumption that is thought to be untrue.

In the example:

- Children who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day ($P<.001$) by about 11 minutes.

In the example:

- This means that, if **all** of the assumptions are correct, including the null hypothesis, there is less than a 1 in 1000 chance that the researchers would have observed the result they did or one even larger. (The result they observed is an average difference of about 11 minutes from one group to the other.)

In the example:

- A 1 in 1000 chance is not very likely
- So it is not likely that, if all of the assumptions are correct, we would have observed the outcome we observed (11 minutes difference in sleep time) or one even larger.
- Therefore, we should evaluate these assumptions, including the null hypothesis

R.A. Fisher called such results “significant”

To Fisher, this meant that the result was worth further scrutiny

- Unfortunately, the word “significant” is loaded with meaning
- Statisticians and others draw the distinction between “statistical significance” and “practical 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"

What people tend to conclude in these situations? (What will the blogs say?)

- Research shows that children who watch TV more during the weekday sleep less than those who don't.
- And from there it is a short walk to "TV is not good for kids and should be limited" or "TV is causing poor performance in school because it makes kids sleep less."
- Authors' conclusion in abstract: "No specific type or use of screen time resulted in significantly shorter sleep duration than another, suggesting that caution should be advised against excessive use of all screens." – In other words, though not demonstrated in the study, all screen usage is suspect.

There is no “p-value transitivity property”

- They argue (in effect):
  - TV = chatting = video games in this study
  - TV results in less sleep in this study
  - Therefore, we should watch out for all the other things, too.
- But the study does not and cannot prove the first assertion!

“If there is enough evidence that one effect is significant, but not enough evidence for the second being significant, that doesn’t mean that the two effects are different from each other. Analogously, **if you can prove that one suspect was present at a crime scene, but can’t prove the other was, that doesn’t mean that you have proved that the two suspects were in different places.”** (emphasis mine)

(http://mindhacks.com/2015/10/03/statistical-fallacy-impairs-post-publication-mood/)

What is scientifically appropriate to conclude?

- The children **in this study** who watched more than 2 hours/day of TV had shorter sleep duration compared with those who watched less than 2 hours/day by about 11 minutes.
- If all of our assumptions, including those about the representativeness of the sample, are correct, the study suggests that nine year old children from this population who watch more than 2 hours/day of TV….

In the sleep research, even if all of our assumptions are correct…

- Does 11 minutes less sleep really matter? Why?
- Furthermore, the “11 minutes” measure is an estimate that has variance – we learn nothing about that variance from the way the data summary is reported (i.e., via a p-value)

And what if THIS had happened:

- Suppose the study showed that children who watched 2 or more hours of TV slept on average 90 minutes per night less than those who did not, but the p-value was 0.09.
- Is this result “insignificant”?

Why did the ASA issue a “statement on p-values and statistical significance?”

Let’s be clear. Nothing in the ASA statement is new.

Statisticians and others have been sounding the alarm about these matters for decades, to little avail.

(Wasserstein and Lazar, 2016)

- "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.

- Cox, D.R. 1986. Some general aspects of the theory of statistics. *International Statistical Review* 54: 117-126.
- A world of quotes illustrating the long history of concern about this can be viewed at David F. Parkhurst, School of Public and Environmental Affairs, Indiana University
- http://www.indiana.edu/~stigtsts/quotsagn.html

Why did the ASA issue a “statement on p-values and statistical significance?”

FEATURE HUMANS & SOCIETY, NUMBERS

Odds Are, It's Wrong

Science fails to face the shortcomings of statistics

BY TOM SIEGFRIED 2:40PM, MARCH 12, 2010

Magazine issue: Vol. 177 #7, March 27, 2010, p. 26

ScienceNews

MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

A journal went so far as to ban p-values

CONTEXT NUMBERS

P value ban: small step for a journal, giant leap for science

Editors reject flawed system of null hypothesis testing

BY TOM SIEGFRIED 3:18PM, MARCH 17, 2015

ASA statement articulates 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.

Biggest takeaway message from the ASA statement – bright line thinking is bad for science

"(S)cientists have embraced and even avidly pursued meaningless differences solely because they are statistically significant, and have ignored important effects because they failed to pass the screen of statistical significance…It is a safe bet that people have suffered or died because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action." (Rothman)

p equal or nearly equal to 0.06

- almost significant
- almost attained significance
- almost significant tendency
- almost became significant
- almost but not quite significant
- almost statistically significant
- almost reached statistical significance
- just barely below the level of significance
- just beyond significance
- "... surely, God loves the .06 nearly as much as the .05." (Rosnell and Rosenthal 1989)

p equal or nearly equal to 0.08

- a certain trend toward significance
- a definite trend
- a slight tendency toward significance
- a strong trend toward significance
- a trend close to significance
- an expected trend
- approached our criteria of significance
- approaching borderline significance
- approaching, although not reaching, significance

And, God forbid, p close to but not less than 0.05

- hovered at nearly a significant level (p=0.058)
- hovers on the brink of significance (p=0.055)
- just about significant (p=0.051)
- just above the margin of significance (p=0.053)
- just at the conventional level of significance (p=0.05001)
- just barely statistically significant (p=0.054)
- just borderline significant (p=0.058)
- just escaped significance (p=0.057)
- just failed significance (p=0.057)

Thanks to Matthew Hankins for these quotes

- https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

A fundamental problem

We want P(H|D) but p-values give P(D|H)

The problem illustrated (Carver 1978)

What is the probability of obtaining a dead person (D) given that the person was hanged (H); that is, in symbol form, what is p(D|H)?

Obviously, it will be very high, perhaps .97 or higher.

The problem illustrated (Carver 1978)

Now, let us reverse the question: What is the probability that a person has been hanged (H) given that the person is dead (D); that is, what is p(H|D)?

This time the probability will undoubtedly be very low, perhaps .01 or lower.

The problem illustrated (Carver 1978)

No one would be likely to make the mistake of substituting the first estimate (.97) for the second (.01); that is, to accept .97 as the probability that a person has been hanged given that the person is dead.

Carver, R.P. 1978. The case against statistical testing. *Harvard Educational Review* 48: 378-399.

Inference is hard work.

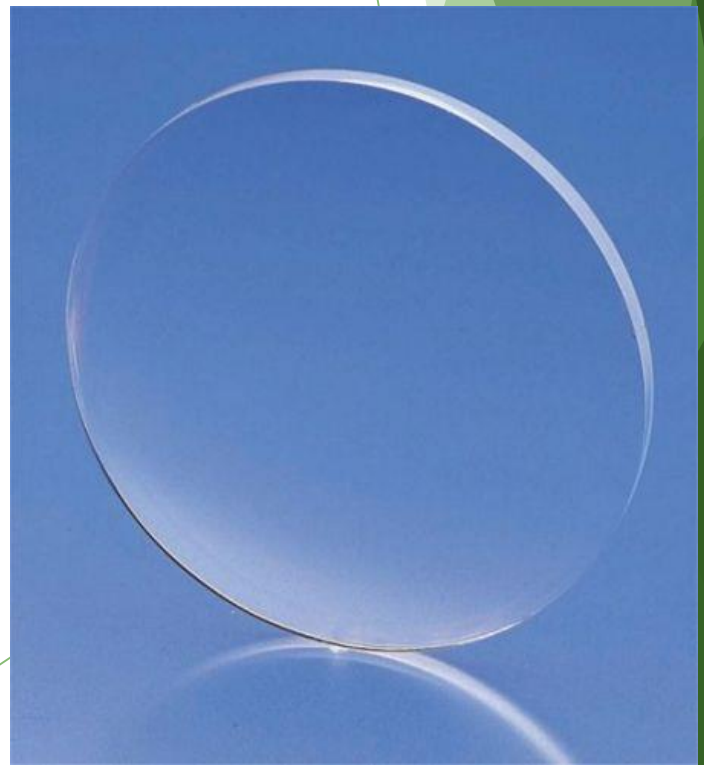
- Simplistic (“cookbook”) rules and procedures are not a substitute for this hard work.
- Cookbook + artificial threshold for significance = appearance of objectivity

In a world where $p<0.05$ carried no meaning…

- What would you have to do to get your paper published, your research grant funded, your drug approved, your policy or business recommendation accepted?

You’d have to be convincing!

You will also have to be transparent



Wrapping up…

P-values themselves are not the problem, but…

- They are hard to explain
- They are easy to misunderstand
- They don’t directly address the question of interest
- When mixed with bright line thinking, they lead to bad science.

Does the ASA statement go far enough?

- The ASA statement does not go as far as it should go.
- However, it goes as far as it could go.

Haiku

Little p-value
what are you trying to say
of significance?

-Steve Ziliak

ron@amstat.org

@RonWasserstein