DOCTOR, IT HURTS WHEN I *P*

Ronald L. Wasserstein
Executive Director
American Statistical Association
October 19, 2017

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."

Why were they interested?

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

So the researchers ask…

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

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 online** had shorter sleep duration than those who chatted less than 2 hours/day (P<.05) by about 16 minutes.

What did the NOT researchers find?

No significant association between playing videogames/working on the computer for more than 2 hours per day and weekday nighttime sleep duration

This is a fairly typical type of study

- Typical scientifically
- Typical statistically
- Atypical communication

This is a good study in many ways

But it makes all-too-typical mistakes

(over)simplified version of results

- TV watching over two hours/day reduces sleep duration by 11 minutes (on average)
- Chatting over two hours/day reduces it by 16 minutes
- But… video games have no significant impact on sleep duration
- By the way: This amount of sleep loss is less than that reported in other studies

Let's think practically about these results

- Does 11 minutes less sleep per night make a lot of difference? (The literature should be able to tell us)
- Shouldn't we have considered what a "significant" sleep loss would be ahead of time?
- Sixteen minutes of sleep loss is more serious than 11 minutes (though still maybe not serious).
- But the results for 11 minutes had $p<.001$ while for 16 minutes only $p<.05$.

And what if instead we had seen…

- Children who watch more than 2 hours/day of TV slept on average 90 minutes less than those who watch less than 2/hours per day ($p<.08$)

- Let's back up now to get a handle on what p-values are, and are not

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

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.
- (Nutshell version): If all assumptions are true, including the assumption that there is no difference in sleep duration, then there is only a small chance we would observe the data we did. So, some assumption is probably untrue.

In the example:

- In symbols, we could say P(D|H) < .001
- In English, the probability that we would observe our data if (or given that) the null hypothesis is true is less than 1 in 1000

But that's not really what we want to know

We want to know from the data whether the hypothesis is true or not

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.

There is also a problem of practice

- It is common practice to assign a threshold to p-values
- “If the p-value is small, say, less than 0.05, then we have observed something important.”
- Results that meet this threshold are called “statistically significant” (or just “significant)

But this leads to bad research behavior

- What happens if my p-value is ALMOST but not quite less than 0.05?
- Who wants to be insignificant?

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/

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, but not good science
- P-values don’t directly answer the question you usually want to know

RON@AMSTAT.ORG

@RON_WASSERSTEIN