PSY 201: Statistics in Psychology

Lecture 37
Catch all
Tell the truth!

Greg Francis

Purdue University

Fall 2023

HONESTY

- It is important that you are honest about what happened in an experiment and in its analysis
  - To yourself
  - To other researchers

HONESTY

- “The first principle is that you must not fool yourself – and you are the easiest person to fool.” Richard Feynman
- Let’s look at several ways you can fool yourself
  - Optional stopping
  - Pilot studies
  - HARKing

OPTIONAL STOPPING

- Hypothesis testing is a *procedure* that controls the Type I error rate
- It works when we know the sampling distribution for the null hypothesis
- The sampling distribution depends on the sample size, so we have to know that
    - Seems trivial, just see how many subjects you have
    - But no.
    - The sample size that matters is how many subjects you *would* run if you repeated the experiment many times
- Surprisingly, many people do not know how many subjects they would run if the experiment were repeated

OPTIONAL STOPPING

- Suppose you run a between subjects test of means with $n_1 = n_2 = 25$

  $$t = 2.0$$

  which gives

  $$p = 0.0512$$

- Some people call this a "marginally significant" result, meaning it is close to the $\alpha = 0.05$ criterion
  - This is nonsense, what you have is a non-significant result
  - You do not get to conclude anything
- But you might think that the results are suggestive, so you run 10 more subjects in each group. Now, with $n_1 = n_2 = 35$, you get

  $$t = 2.2$$

  $$p = 0.0312$$

- What is the Type I error rate for this kind of procedure?

OPTIONAL STOPPING

- The Type I error rate has to be bigger than $\alpha = 0.05$ because your first test had that error rate
- The second test (with added subjects) has some unknown additional Type I error rate
- What would you have done if the second test produced?:

$$t = 1.99$$

$$p = 0.0506$$

- If you would have run even more subjects, then you need to consider those steps as also being part of the Type I error rate of your *procedure* (even if they were not actually done in your particular situation)

OPTIONAL STOPPING

- In fact, you have to know what you would do for every (infinitely many) possible situations
- If you are willing to keep adding subjects until you get a significant result, your Type I error rate is 1.0!
- The best way out of this problem is to *fix* a sample size and stick to it. This is best done with a good power analysis before gathering any data.
- If that is not possible, then honestly describe what was done. Describe each test and explain why subjects were added.

PILOT STUDIES

- When investigating a new topic, it is common to run multiple experiments while identifying what to measure and how to do it
- For example, suppose you want to study the effect of eating bananas on recall of words
- There are lots of variables to consider
  - How many bananas?
  - How long after eating do you study words?
  - How long after eating do you test?
  - What kind of words do you use?
- You can explore hundreds of these variables to find a combination that shows an effect
- It might be tempting to use statistical significance to decide whether a study "works"
- Don't do it!

PILOT STUDIES

- What sometimes happens is people interpret the difference between significant and non-significant results as indicating *methodological* differences:
  - One banana does not improve memory (warning: accepting the null!)
  - Two bananas does improve memory
- But both of these studies also involve random sampling
  - A null effect might produce a Type I error
  - A real effect might not produce a significant result (Type II error)
- It is dishonest to run tests like these and not report the results, even if you think you can explain why a study "failed"

PILOT STUDIES

- If you only report successful studies (publication bias), it becomes impossible for other scientists to interpret the Type I error rate of your results
- They do not know if you are reporting the only result you tested for
- Or if you are reporting one study out of dozens of others that did not work
- Your best bet is to figure out how to run a good study and then do it once.
    - Easier said than done
    - You might spend years figuring out how to run a good study

HARKing

- Hypothesizing After the Results are Known
- You sometimes learn a *lot* after looking at your data
- Sometimes scientists look at the data then identify hypotheses that match the results
- Sometimes scientists then "pretend" that they predicted the outcome and write up their paper accordingly
- Don't do this. It is fraud.
- Just be honest and explain that you learned from your findings.

HARKing

- Be careful about what you learn, though
- You might run a study on eating bananas and word memory and sift through a large set of subjects to find a subset that shows an effect
  - Age: Young, Middle, Old
  - Sex: male, female
  - Socioeconomic status: quartile
  - Religious affiliation: Christian, Muslim, Jewish, Atheist, Buddhists, Other
- Maybe you find:
  - A significant improvement for Young, male, 25th percentile SES, Buddhists
  - A significant decrement for Young, female, 75th percentile SES, Christians
  - A significant increment for Young, female, 10th percentile SES, Atheists
  - A significant decrement for Old, female, 25th percentile SES, Jews

HARKing

- Sometimes these kinds of conclusions feel like *aha!* moments, where you suddenly have deep insight into what is going on
- Based on other research, you realize all the increments are for people with "high sense of self" while the decrements are for people with "low sense of self" (I'm just making these terms up)
- You are very possibly seeing "signal" in pure "noise"
- What you are doing is exploratory work
- It is (maybe) good for coming up with ideas, but you cannot use one set of data both to identify ideas and to test them
- In a follow-up test you need to measure whatever other variables you think really matter (e.g., "sense of self")

IN THE WILD

- Scientists do these kinds of “questionable research practices” all the time
  - Often unintentionally
  - They just do not know any better
- This is why you hear so much conflicting advice on some topics
  - Chocolate is good for you / chocolate is bad for you
  - A glass of wine a day is good for you / no it’s not
  - Take statins to improve your health / they seem to do nothing
- This is why you sometimes see nonsense published in journals
  - People can get information from the future
  - Eating breakfast makes a woman more likely to have a boy baby
  - Women find men wearing red shirts to be more attractive

STATISTICS LIMITS

- Hypothesis testing (and statistics in general) is not synonymous with science
- Science is about identifying mechanisms to explain why things happen the way they do
- Hypothesis testing (at best) prevents misinterpretations of signal for noise, but that is not enough to identify mechanisms
- At best, statistics is a check on interpreting noise as if it were signal
- At worst, statistics is a way of “validating” noise as if it were a signal
- In some sense, the best science does not require statistics

CONCLUSIONS

- Good science is difficult to do well
- There are lots of ways to “cheat” hypothesis testing
- People actually do cheat
- Be skeptical about published work
- Use some common sense!

NEXT TIME

- Review for the final exam