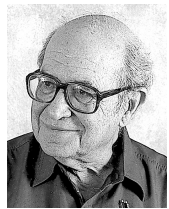
Review of Jacob Cohen's The Earth is Round ($p < .05$)

Will Quinlan

October 7, 2024

An Introduction to Cohen's Paper

- Scientific advancement has been hindered by null hypothesis testing (NHST)
- Researchers often misinterpret p-values restulting in mistakes in what can be concluded
- Critique the "ritualization" of NHST
- Focus on real-world relevance
- Argues for better scientific practices



On Null Hypothesis Significance Testing

- Definition: Method of statistical inference by which an experimental factor is tested against a hypothesis of no effect or no relationship based on a given observation
  - In other words: Null vs Alternate
- Decisions are determined by the signifiance threshold
- NHST dominates science
- If one presents a significant result of a known fact like "The Earth is round," nothing is gained

A Ritualization of Reasoning ($p < .05$)

- Cohen argues against the rigid reliance on p-values
- Standard p-value of .05
- Dichotomous reject-accept decision
- "The primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one . . . believes the hypothesis . . . being tested" - Bill Rozeboom (1960)

Misinterpretation of p-values

- Consider a Scenario
  - A scientist believes a certain rare disease does not exist in a population
  - H_0: P = 0
  - He draws a random sample of 30 cases and finds that 1 person has the disease
  - $P_S = 1/30 = .033$
- How should H_0 be tested? Chi-square? How about a Fischer exact test? Is there enough power?
- Cohen among many other scientists believe that academia would complain unless this result had a p-value attached to it

So, what is wrong?

- We want NHST to tell us, "Given these data, what is the probability that H_0 is true?"
- What NHST actually tells us, "Given that H_0 is true, what is the probability of these (or more extreme) data?" before asking your question
- Academics have pointed out the failures of NHST for many years, even before Cohen's paper
  - NHST was described "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" - Meehl in 1967
  - Joseph Berkson attacked NHST in 1938
  - Lancelot Hogben's book-length critique appeared in 1957

The Permanent Illusion

- Many believe the level of significance at which H_0 is rejected is the probability that it is correct
- Lets dive into why this thinking is wrong
- Syllogistic Reasoning according to Aristotle
  - If the null hypothesis is correct, then this datum (D) can not occur
  - D has, however, occurred
  - Therefore, the null hypothesis is false
- This is an example of a modus tollens syllogism, denying the antecedent by denying the consequent
  - If P, then Q
  - Not Q (Q is false)
  - Therefore, not P (P is false)

Probabilistic Syllogism

- NHST makes the modus tollens syllogism probabilistic as follows:
  - If the null hypothesis is correct, then these data are highly unlikely
  - These data have occurred
  - Therefore, the null hypothesis is highly unlikely
- This makes what was a valid modus tollens syllogism formally invalid
  - If P, then Q is likely
  - Not Q
  - Therefore, P is unlikely

Examples of Syllogisms

- Some Examples:
  - If a person is a Martian, then he is not a member of Congress
  - This person is a member of Congress
  - Therefore, he is not a Martian
- How about this one?
  - If a person is an American, then he is not a member of Congress
  - This person is a member of Congress
  - Therefore, he is not an American

Examples of Syllogisms cont.

- What about now?
  - If a person is an American, then he is probably not a member of Congress
  - This person is a member of Congress
  - Therefore, he is probably not an American
- Notice how these two syllogisms are formally the same
  - If H_0 is true, then this result (statistical significance) would probably not occur
  - This result has occurred
  - Therefore, H_0 is probably not true

$\mathrm{P}(\mathrm{D}|H_0) \neq P(\mathrm{H}_0|\mathrm{D})$

- When testing H_0, one is finding the probability data (D) could have arisen if H_0 were true, $\mathrm{P}(\mathrm{D}|H_0)$
- If this probability is small, one can conclude that H_0 is true and D is unlikely
- What about the reverse probability $\mathrm{P}(H_0|\mathrm{D})$?
  - When rejecting H_0, one wants to conclude H_0 is unlikely
  - This probability is only available through Bayes theorem where we need to know $\mathrm{P}(H_0)$
  - Bayesian statisticians use a prior probability or distribution of probabilities to deal with this problem, but does it hold up?

A Look at Psychiatric Diagnoses

- Incidence of schizophrenia in adults is 2%
- A proposed screening test is estimated to have 95% accuracy ($\mathrm{P}(\text{normality}|H_0) \approx 0.95$)
- The screening test is supposed to have 97% accuracy in declaring normality ($\mathrm{P}(\text{schizophrenia}|H_1) > 0.97$)
- Thus, we have a test that is highly sensitive and highly specific
  - H_0 = The case is normal
  - H_1 = The case is schizophrenic
  - D = The test result (the data) is positive for schizophrenia

A Look at Psychiatric Diagnoses cont.

- $P(D|H_0) < .05$ seems like what we want, but it is not
- We want $P(H_0|D)$ which equals .60, not the .05 we may have believed

$$P(H_0|D) = \frac{P(H_0) \cdot P(\text{test wrong}|H_0)}{P(H_0) \cdot P(\text{test wrong}|H_0) + P(H_1) \cdot P(\text{test correct}|H_1)}$$

- Schizophrenic

Result	Normal	Schiz	Total
Negative Test (Normal)	949	1	950
Positive Test (Schiz)	30	20	50
Total	979	21	1,000

Replication

- The error previously demonstrated can also be applied to replication of tests
- If there was a successful rejection of H_0 many believe replications will also result in the rejection of H_0
- Many believe that a p of .99 means 99% of the time a result will replicate
- Typical level of power for medium effect sizes of .50
  - The chances are in three replications only one in eight would result in significant results

More Syllogisms

- We have just seen many failures in logic by researchers such as if H_0 is rejected, then the theorey is established. Invalid syllogism and example below:
  - If it rains (A), then the ground will be wet (B).
  - The ground is wet (B).
  - Therefore, it rained (A).
- However, even if a valid modus tollens syllogism in used, misinterpretations are still made
  - When H_0 is rejected, it can be because of a variety of auxillary theories, and not what precipitated the research
  - Although it is convenient, accept-reject decisions; although convenient, are not how science is done

The Nil Hypothesis

- Some propositions consider what is occuring within a population (i.e. the proportion of males in this population is .75)
- Cohen refers to the H_0 when the effect size $= 0$ as the nil hypothesis
- Effecet size is effectively the practical significance of a test
- However, universally H_0 is taken to mean nil (zero), Ex:
  - H_0 is the proportion of males in a population if .50
  - H_0 is the raters reliability is 0
- These are cases when H_0 is almost universally rejected
- Cohen states that the Nil Hypothesis is always false
- Tukey wrote, "It is foolish to ask 'Are the effects of A and B different?' They are always different—for some decimal place"

The Nil Hypothesis cont.

- Even if a test is statistically significant with $p < .000001$, there can be ambient correlation noise among often arbitrarily paired variables
- Given the fact that the nil hypothesis is always false, the rate of Type I errors is 0%, not 5%, and that only Type II errors can be made, which run typically at about 50%
- The sample effect size necessary for significance is notably larger than tha actual population effect size
- The use of a Bonferoni adjustment is then adjusting for non-existent alpha error which in turn overestimates the population effect size
- Again, typically little is learned from A is larger than B ($p < .01$)

Quantifying Effects

- Confidence intervals and effect sizes are necessary for scientific advancement
- Scientists cannot only record pulling a rubberband makes it longer, how much longer is necessary
- Correlation coefficients depend on the selection of the population, regression coefficients do not
- Correlations do not provide insight to causal strength, within group variation can change a correlations strength
- Cohen's d and f — like correlations, these values are impacted by the variability across the population
- The context of a regression matters greatly as seen in Cohen's height and IQ
  - Cohen had a statistically significant correlation between height and IQ
  - This translated to a regression coefficient that meant to raise a childs iq by 30 required enough growth hormone to raise his or her height by 14 feet

What to do?

- 1. There is no magic test to replace NHST
- 2. More emphasis needs to be placed on detective work with data as opposed to sanctification
  - Cohen calls for a shift toward understanding and improving data rather than making mechanical accept-reject decisions
- 3. More emphasis on confidence intervals and effect sizes
  - Confidence intervals aren't frequently used because they expose large variability
- 4. Move away from the point "nil hypothesis" to "good-enough" range null hypotheses which will make NHST and power more useful

Thank you all for listening!