P-Values are Rarely Used in Forensic Science That is (not) too Bad

Alicia Carriquiry
Iowa State University

May 6, 2020

Innocent Until Proven Guilty

What appears to be the perfect set-up for hypothesis testing, turns out to be a rocky road indeed.

Outline

- **Part I: Evaluation of Evidence: where *p*-values should NOT be used**
  - The legal setting as a hypothesis test
  - Motivating example
  - Problems and some alternatives.
- **Part II: Making predictions: where *p*-values would be useful**
  - *p*-values as sentinel statistics for algorithm fairness.

Hypothesis testing in the legal context

- Drawing a parallel between hypothesis testing and the question asked in trial is tempting:
  - Before evidence is introduced, defendant is presumed innocent.
  - Evidence is introduced during trial.
  - Jurors update their presumption based on the *weight* of the evidence.
- If G denotes "guilt" and $\bar{G}$ denotes "not guilt", we formulate the test:

$$\mathrm{H}_0 : \bar{G} \text{ versus } \mathrm{H}_A : G,$$

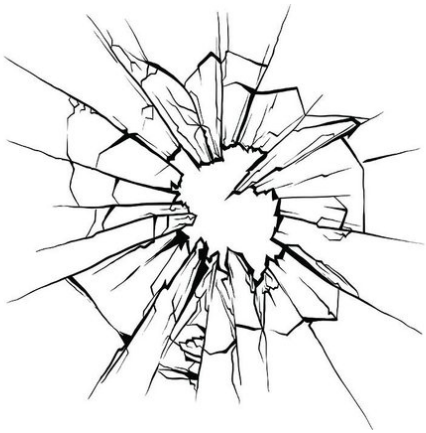
and base the decision on the evidence E presented during trial.

Errors

- Two types of errors
  - type I: reject the null hypothesis when it is true (false positive)
  - type II: fail to reject the null when it is false (false negative)
- Type I error often considered more serious: we only want to reject the null if strong evidence against it
- In the context of the justice system
  - type I error is to decide guilty when person is innocent
  - type II error is to decide innocent when person is guilty

Really?

- It is not that simple.
- A crime is committed and evidence is found at the crime scene. E.g., a finger print and some blood.
- A suspect is charged with committing the crime.
- A possible forensic question: **is the suspect the source of the evidence found at the crime scene**?
- We focus on this "simpler" question of source.



Motivation: Glass fragments as evidence

- A window was broken in the commission of a crime.
- Glass fragments are recovered from the defendant's clothing.
- Two populations of glass fragments to compare:
  - The fragments from the known source (broken window), and
  - The fragments with questioned source (from the suspect).
- Question of interest is whether these populations differ in important ways (are distinguishable)

ASTM 2927-16

- ASTM 2927-16: Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons
  - Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition)..... The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass."

ASTM 2927-16

- 11. Calculation and Interpretation of Results

  11.1. The procedure below shall be followed to conduct a forensic glass comparison when using the recommended match criteria:

  11.1.1. For the Known source fragments, using a minimum of 9 measurements (from at least 3 fragments, if possible), calculate the mean for each element.

  11.1.2. Calculate the standard deviation for each element. This is the Measured SD.

  11.1.3. Calculate a value equal to at least 3% of the mean for each element. This is the Minimum SD.

  11.1.4. Calculate a match interval for each element with a lower limit equal to the mean minus 4 times the SD (Measured or Minimum, whichever is greater) and an upper limit equal to the mean plus 4 times the SD (Measured or Minimum, whichever is greater).

  11.1.5. For each Recovered fragment, using as many measurements as practical, calculate the mean concentration for each element.

  11.1.6. For each element, compare the mean concentration in the Recovered fragment to the match interval for the corresponding element from the Known fragments.

  11.1.7. If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not "match" and the glass samples are considered distinguishable.

- This is a statistical inference procedure

Hypothesis testing

- For a single element, we can formalize the ASTM approach as follows:
  1. Let μ_c be mean chemical concentration of one element of population from which window fragments are sampled. Sample estimate is $\bar{x}_c$.
  2. Let μ_s be the mean of the population from which suspect's fragments were sampled, estimated as $\bar{x}_s$.
  3. Test $H_0 : \mu_c = \mu_s$ versus the alternative $H_A : \mu_c \neq \mu_s$.
- ASTM establishes that decision is H_0 if the difference $\bar{x}_c - \bar{x}_s \leq 4SD$.
- We know how to do this, but should we?

Test statistics and p-values

- The difference between the two means is quantified using some statistic.
- The result of test is usually summarized by a *p*-value measuring the strength of the statistical evidence against the null hypothesis.
- Definition: a ***p*-value** gives the probability that we would get data like the data we have observed in the sample (or something even more extreme) given that the null hypothesis is true.
- Importantly, the *p*-value only addresses the null hypothesis. It does not speak to the likelihood of the alternative hypothesis being true.

Limitations

- In the legal context, the use of hypothesis testing and p-values is fraught with problems:
  1. The null hypothesis is backwards: defendant is presumed to be the source of the evidence until we prove otherwise.
  2. The p-value tells us nothing about the probabilities of false positives or negatives.
  3. Hypothesis are not treated symmetrically.
  4. Even if we fail to reject H_0, we cannot conclude same source.
  5. We are not answering the question of interest.

Hypotheses are backwards

- If H_0 corresponds to "same source", then the presumption is that defendant is source of evidence.
- Consequently:
  - Burden of proof is on defense: present strong evidence to contradict null.
  - The noisier the measurements, the harder to reject the null.
- In ASTM method, "acceptance region" is mean difference $\pm$ 4SD. Mean difference must be quite large to reject the null.
- The larger the SD, the wider the region where we fail to reject H_0: more measurement noise leads to higher probability of "same source" conclusion.

p-value is useless

- The p-value just tells us whether our test statistic is unexpectedly "extreme" when the null is true.
- That tells us nothing about the probability of the null being true!
- Interpretation of p-values in Court (and elsewhere) is often wrong: tiny p-value means that the probability that H_0 is true is also tiny.
- p-values are confused with the probability of incorrectly concluding different source.
- Failing to reject the null hypothesis does not mean that null is true.

Asymmetric treatment

- The two hypothesis are not treated symmetrically.
- To reject the null hypothesis, the evidence against it must be convincing.
- If we subscribe to the notion that it is better to let a guilty person go than to incarcerate an innocent person, then:
  - Make it easier to reject the null hypothesis
  - Increase the type I error and decrease the type II error.

A match may not be probative

- This is an important point: even if we find that two items are *indistinguishable*, that does not imply that they have a common source.
- The match may have occurred by chance!
- The p-value is absolutely silent about this.

The Two-Stage Approach

- One common statistical approach solves the forensic problem in two stages
- Stage 1 (Similarity)
  - determine if the crime scene and suspect objects agree on one or more characteristics of interest (typically using a hypothesis/significance test)
  - two samples "are indistinguishable", "can't be distinguished", "match"
- Stage 2 (Identification)
  - assess the significance of this agreement by finding the likelihood of such agreement occurring by chance

Note that....

- Stage 1 - Using a binary decision
  - A binary decision (to reject the null hypothesis or not) requires the selection of a cutoff or threshold (e.g., .05 p-value or 4-sigma interval)
  - Choice of threshold impacts the error rates associated with the test
    - a low "threshold" makes it easy to reject ... risks a type I error which rejects a true match
    - a high "threshold" makes it easy to accept the null ... risks a type II error which declares non-matching populations as indistinguishable and could thus incriminate incorrectly
  - To estimate the random match probability, we need information about the *relevant population*, e.g., background frequencies of alleles, or distribuiton of chemical concentrations in "similar" glass.

An alternative approach

- Clearly, we should evaluate the likelihood of observing the evidence under two competing hypothesis:
  - Defendant is source of evidence: H_p
  - Defendent is not source of evidence: H_d.
- These roughly correspond to Stages I and II in the two-Stage approach.
- Both steps can be summarized into a **Likelihood Ratio (LR)** statistic:

$$LR = \frac{Pr(E|H_p)}{Pr(E|H_d)}.$$

Still not enough

- In Court, we are interested in $Pr(H_p|E)$ and $Pr(H_d|E)$.
- **Prosecutor's fallacy**: when $Pr(E|H_p)$ is confused with $Pr(H_p|E)$.
- To estimate $Pr(H_p|E)$ and $Pr(H_d|E)$, need to rely on Bayes' Theorem.
- No such thing as a free lunch:
  - Need to define a *prior*, which in the legal context corresponds to the background frequency of E.
  - Run into the thorny issue of choice of *reference class*.

When *p*-values ARE useful

- In recent years, algorithms to predict behavior have become popular in the criminal justice system.
- Can we predict whether a person who is released on bail will show up at trial? Will a parolee re-offend?
- Classification-type algorithms, called *Risk Assessment Tools* are all the rage:
  - Train the algorithm using labeled data and a collection of features.
  - Use the algorithm to predict outcome on new items given their features.
- Many examples, most algorithms proprietary.

Are algorithms "fair"?

- Most algorithms do not use race as a feature, but do use variables such as criminal history.

An algorithm used in Pennsylvania

- Not a black box.
- Extensive validation studies:
  - Does algorithm have low probability of wrongly classifying a person as high or low risk?
  - Is accuracy independent of personal attributes e.g., race, gender, age?
- Validation study took over three years, about 27,000 cases.
- Metrics used: accuracy, sensitivity, specificity, area under the curve.
- Typically not used: hypothesis testing.

Probability of re-offending

- During the study period:
  - Among 19,629 White offenders, 4,651 (or 24%) re-offended.
  - Among 7,405 Black offenders, 2,063 (or 28%) re-offended.
- This is "ground truth".
- How well did the algorithm predict the proportion who might re-offend (high risk persons) and the proportion who would be unlikely to re-offend (low risk persons)?
- Were those predictions equally accurate for Blacks and Whites?

Among those who DID NOT re-offend

	Low risk predicted by algorithm		
Race	Yes (correct)	No	Total
White	13385	1656	15041
Black	4517	825	5342

- Among Whites, 11% (1656/15041) were **incorrectly** classified as high-risk.
- Among Blacks, 15% (825/5342) were **incorrectly** classified as high-risk.
- Difference is worth exploring:

$$\chi^2_{1df} = 72.5, \qquad p < 0.0001.$$

Among those who DID re-offend

	High risk predicted by algorithm		
Race	Yes (correct)	No	Total
White	2215	2436	4651
Black	1204	859	2063

- Among Whites, 52% were **incorrectly** given the benefit of the doubt.
- Among Blacks, 42% were **incorrectly** given the benefit of the doubt.
- Difference is highly significant.

THANKS FOR LISTENING

alicia@iastate.edu