Classical Hypothesis Testing

Conor Mayo-Wilson

Philosophy of Statistics
June 17th, 2014

Review

Today:

- Models of experiments: Classical/Frequentist vs. Bayesian
- Classical/frequentist hypothesis tests and criticisms

Common Model of an Experiment

Common Model of Experiments

- Θ - Set of experimental setups.
  - E.g., Number of red balls in an urn
  - E.g., Biases of coin
- Ω - Set of experimental outcomes.
  - E.g., Which ball is selected
  - E.g., Sequences of coin tosses (which might have, for example, different angular velocities)
- A sequence of random variables $X_n : \Omega \to \mathbb{R}$ representing the observable parts of the experimental outcome:
  - E.g., The color of the selected ball
  - E.g., Sequences of Heads or Tails

Common Model of Experiments

- A subset $A \subseteq \mathcal{P}(\Theta)$ representing permissible estimates or actions.
  - E.g., Single elements of Θ = Point Estimates
  - E.g., Intervals of $\Theta = \mathbb{R}$ = Interval Estimates
  - E.g., Elements of a collection of disjoint subsets of Θ = Hypotheses
- Estimators $\hat{\theta}_n : \mathbb{R}^n \to A$ take observations as input and return a permissible estimate or actions.
  - E.g., The sample mean $\hat{\theta}(x) = \overline{x}$ is a point estimator.
  - E.g., Interval estimators
  - E.g., Hypothesis Tests
- A loss/utility function $L : A \times \Theta \to \mathbb{R}$ representing the payoff $L(a, \theta)$ of action a if the true experimental setup is θ.

The central difference between the two models concerns what events can be assigned probability.

Bayesian Model

Bayesian: A single probability distribution P is defined over both experimental setups Θ and their outcomes Ω.

- The distribution P is often taken to represent a single agent's **degrees of belief**.
- Nothing in the mathematics prohibits interpreting P as an evidential probability or as a propensity (if experimental setups Θ are also produced by some chance device).

Bayesian Model

Classical: For each experimental setup $\theta \in \Theta$, there is a probability distribution P_θ over **only the outcomes** Ω.

- P_θ is often interpreted as saying something about long-term frequencies.
- For reasons I've explained, I think it is better to interpret P_θ as a propensity, but any objective view of probability is consistent with the mathematics.

The Central Difference

Question: Why does the difference matter for methodology?

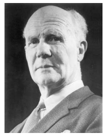
Answer:

- Bayesians can calculate a posterior distribution $P(\theta|X_1, \ldots X_n)$ given the data, and minimize expected loss relative to it.
- Because there is no well-defined prior probability of θ in the classical statistician's framework, the posterior distribution is also not well-defined.
  - So classical statisticians have a number of criteria used to assess the reliability of different estimators.

Classical Hypothesis Tests

Two Traditions

vs.



"Classical" hypothesis testing is really two different sets of techniques.

- Fisher
- Neyman-Pearson

Fisher vs. Neyman-Pearson

	Fisher	Neyman-Pearson
Alternative hypothesis?	No	Yes
Key Concepts	P-value	Size and Power
Normative Upshot	Evidential	Behavioral
One-shot?	Single Experiment	Long Run

To review the differences between two traditions, here's a toy example:

Hypothesis Testing - Toy Example

- Suppose a coin factory produces two types of coins, one with bias $\frac{1}{4}$ and the other with bias $\frac{3}{4}$.
  - $\Theta = \{\frac{1}{4}, \frac{3}{4}\}$.
- Suppose your **null-hypothesis** Θ_0 is that the coin has bias $\frac{1}{4}$.

Hypothesis Testing - Toy Example

Suppose you have flipped the coin 52 times, and observed 26 heads.

- $\Omega = \{0,1\}^{52}$.
- $X_n : \Omega \to \mathbb{R}$ is the map

$$\omega = \langle \omega_1, \omega_2, \ldots \omega_{52} \rangle \mapsto \omega_n.$$

- Here's what you know about the experimental outcome $\omega \in \Omega$:

$$\sum_{n \leq 52} X_n(\omega) = 26$$

Fisherian Testing

- For each data sequence $x \in \mathbb{R}^{52}$, there is a set of "more extreme values" E_x.
  - E.g., If x has n_x heads, then E_x might be all data sequences containing at least n_x many heads.
- In general, the P-value of an observed outcome x is defined as:

$$\sup_{\theta \in \Theta_0} P_\theta(\langle X_1, \ldots, X_{52} \rangle \in E_x)$$

where Θ_0 is the null hypothesis.

Fisherian Testing

- In the example, if E_x is all sequences involving more heads those observed, then the P-value is

$$P_{\frac{1}{4}}(\sum_{n\leq 52} X_n \geq 26) = 0.00009021965471400772$$

Test Statistics

- Typically, P-values are computed using some test statistic.
- A test statistic is a series of functions $T_n : \mathbb{R}^n \to \mathbb{R}^k$ (for some k) that "summarize" the observed data:
  - Sample mean: $T_n(X_1(\omega), \ldots, X_n(\omega)) = \frac{1}{n}\sum_{j\leq n} X_j(\omega)$
  - Sample Variance:

    $$T_n(X_1(\omega), \ldots, X_n(\omega)) = \frac{1}{n}\sum_{j\leq n}(X_j - \overline{X})^2$$

    where $\overline{X}$ is the sample mean.
  - The Sample Itself!: $T_n(X_1(\omega), \ldots, X_n(\omega)) = \langle X_1(\omega), \ldots, X_n(\omega)\rangle$

Fisherian Testing

- Fisher interprets a low P-value as strong evidence against the null hypothesis.
  - Notice the "evidential" interpretation.
  - Notice the evidence is **against** the null hypothesis, not for some alternative.
- Reject the null hypothesis if the P-value is low.

Fisherian Testing

Objection 1: The choice of null hypothesis is arbitrary

- In the example, the sample may seem like strong evidence against the null, but if the coin factory produces only the two types of coins, it's equally strong evidence against $\Theta_1 = \{\frac{3}{4}\}$.

Fisherian Testing

Objection 2: The choice of test-statistic is arbitrary.

- Why the number of heads? Why not the number of heads on even tosses?

Fisherian Testing

Objection 3: The choice of the set of extreme values is arbitrary.

- Suppose 13 heads rather than 26 heads had been observed in the example. Would Fisher reject the null hypothesis using "too good to be true" that caused him to question Mendel?

Fisherian Testing

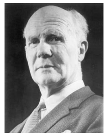
Objection 4: Different P-values can be obtained from the same evidence, and hence, the P-value cannot be a measure of the strength of evidence again the null hypothesis.

Fisherian Testing

Responses:

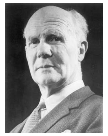
- We'll discuss some responses next week.
- Neyman and Pearson tests, however, already address these three objections (at least in special cases) . . .

Neyman and Pearson



- Neyman and Pearson argue that a null-hypothesis Θ_0 ought to always be tested against an alternative Θ_1.
- Consequently, they aim to minimize two types of error:
  - **Type I Error:** Rejecting the null when it's true.
  - **Type II Error:** Accepting the null when it's false.

Neyman and Pearson



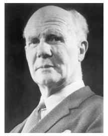
- Suppose you employ a test that rejects the null hypothesis if the data belongs to a set R called the rejection region.
- The size of a test is the greatest chance of committing a Type I error:

$$\sup_{\theta \in \Theta_0} P_\theta(\mathbf{X} \in R)$$

- The Power is the greatest chance of committing a Type II error:

$$\sup_{\theta \in \Theta_1} P_\theta(\mathbf{X} \notin R)$$

Neyman and Pearson



Consider the example.

- Suppose your rejection region R is the set of coin flips containing more heads than tails:
- The size of this test is

$$P_{\frac{1}{4}}(\sum_{n \leq 52} X_n \geq 27)$$

- The Power of this test is:

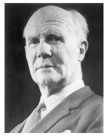
$$P_{\frac{3}{4}}(\sum_{n \leq 52} X_n < 27)$$

Neyman and Pearson

- Clearly, there is a tradeoff between size and power.
- You can minimize the chance of Type I error by always retaining the null hypothesis.
- You can minimize the chance of Type II error by always rejecting the null.

How do Neyman and Pearson navigate this tradeoff?

Neyman and Pearson



- First, fix the size α of the test (customary is .05).
- Then find a rejection region R that maximizes the power if the size is α.

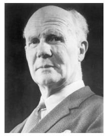
Objections

Question: Aren't Neyman and Pearson subject to the same objections as Fisher?

Answer: Not always.

Objections

vs.



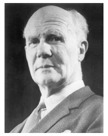
Objection 1: The choice of the null hypothesis is arbitrary.
Answer: Recall the difference between Fisher and Neyman and Pearson's interpretations of hypothesis tests . . .

Neyman and Pearson

> *But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behaviour": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts ; if $x > x_0$, reject H, if $x \leq x_0$,, accept H.* **Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x_0$, or false when $x > x_0$.** *But it may often be proved that if we behave according to such a rule, then in the long run me shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false.*

Objections

vs.

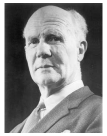


Objection 1: The choice of the null hypothesis is arbitrary.
Answer: The null hypothesis is chosen with respect to one's long-term goals.

- Since one hypothesis may be more important than another, it may be more important to minimize Type I error in the long run.
- So even if the evidence is symmetric between two hypotheses, the choice of the null is not arbitrary.

Objections

vs.



Objection 2: The choice of the rejection region is arbitrary.
Answer: For Neyman and Pearson, in the example, the choice of rejection region in the example is uniquely determined if one fixes the size α of the test.

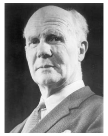
Namely, it is the set of observable data sequences $\mathbf{X}$ such that:

$$\frac{P_{\frac{1}{4}}(\mathbf{X})}{P_{\frac{3}{4}}(\mathbf{X})} \leq k_\alpha$$

where k_α is a constant depending upon α.

Objections

vs.

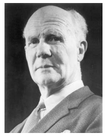


Objection 3: The choice of the test statistic is arbitrary.
Answer: For Neyman and Pearson, the test statistic should be sufficient.

More on this in a second . . .

Objections

vs.



Objection 4: The *P*-value is not a measure of evidential strength.
Answer: That's right. The size of the test, which is closely related to the *P*-value for Fisher, is a measure of long run correctness.

Today

Next Class:

- Why some take objections to hypothesis testing motivate use of interval estimates (esp. confidence intervals).
- The Likelihood Principle: Why its proof attacks classical methodology in general, and not just hypothesis tests.