The Likelihood Principle

Conor Mayo-Wilson

Philosophy of Statistics
June 30th, 2014

Review

Last Month:

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

Today

Today

- 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.

To review the differences between two traditions, here's an 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 ten-thousand times, and observed 2575 heads.

- $\Omega = \{0,1\}^{10^4}$.
- $X_n : \Omega \to \mathbb{R}$ is the map

$$\omega = \langle \omega_1, \omega_2, \ldots \omega_{10^4} \rangle \mapsto \omega_n.$$

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

$$\sum_{n \leq 10^4} X_n(\omega) = 2575$$

Objections to Classical Hypothesis Testing

Objection 1: "Statistically significant" does not mean practically significant.

Example

Example:

- In the toy example, the probability of obtaining 2575 or more heads is about 4%, which is significant at the .05 level.
- However, unless you own a casino, you probably do not care if your coin has bias .2575 or bias .25.

Mayo and Spanos' Answer

vs.

Answer: Test the null $\mu = \mu_0$ vs. the alternative $\mu > \mu_1$, where μ_1 is a "practically significant" deviation.

Mayo and Spanos [2011]

Objection 2: With a large enough sample, the smallest, insignificant deviation from a point null will be rejected.

Example

Example:

- In the toy example, suppose, unbeknownst to you, the coin factory produces coins with bias $\frac{1}{4} + \epsilon$ for some very small $\epsilon > 0$.
- You wouldn't care if you had one such coin or a coin with bias precisely $\frac{1}{4}$.
- You are flipping one such coin.
- Then, under any reasonable test, the probability of rejecting the null hypothesis $\Theta_0 = \{\frac{1}{4}\}$ at any significance level $\alpha > 0$ approaches one as the number of flips approaches infinity.

Mayo and Spanos' Answer

vs.

Answer:

- If the null hypothesis is false, it should be rejected. That doesn't entail accepting the alternative hypothesis.
- **My Note:** The "change the hypothesis" response also could be used here: Test the null $\mu \in (\mu_0 - \epsilon, \mu_0 + \epsilon)$.

Mayo and Spanos [2011]

Objection 3: The choice of the null hypothesis is arbitrary.

Example

Example:

- If your null hypothesis is $\Theta_0 = \{\frac{1}{4}\}$, then the probability of obtaining 2575 or more heads is about 4%, which is significant at the .05 level. So you reject the null.
- If your null hypothesis is $\Theta_0 = \{\frac{3}{4}\}$, of obtaining 2575 or fewer heads is approximately zero, which is significant at the .05 level. So you reject the null.

Mayo and Spanos' Answer

vs.

Answer: The choice of the null hypothesis can only cause erroneous inferences if one fallaciously infers that evidence again the null hypothesis is evidence for some specific alternative.

Mayo and Spanos [2011, pp. 175]

Objection 4: Statistically insignificant results are taken as evidence that the null hypothesis is true.

Mayo and Spanos' Answer

vs.

Answer:

- That's a mistake of practitioners; not of the methodology.
- In certain settings, we can quantify how "large" a variation from the null hypothesis is compatible with failure to reject. So "statistically insignificant" results can have some informative value.

Mayo and Spanos [2011]

Interval Estimation

Competitive State Summary

State	EV	O	R	Forecast		90% Prediction Interval	%
NM	5	183	355	Obama +8.9	+/- 6.6		98%
ME-2	1	184	354	Obama +8.2	+/- 8.9		94%
OR	7	191	347	Obama +7.9	+/- 6.3		98%
MI	16	207	331	Obama +7.2	+/- 5.4		98%
MN	10	217	321	Obama +7.0	+/- 5.4		98%
PA	20	237	301	Obama +5.1	+/- 5.0		96%
WI	10	247	291	Obama +4.0	+/- 5.6		88%
NV	6	253	285	Obama +3.5	+/- 5.6		85%
IA	6	259	279	Obama +2.9	+/- 6.3		78%
NH	4	263	275	Obama +2.8	+/- 6.8		75%
OH	18	281	257	Obama +2.6	+/- 5.2		80%
CO	9	290	248	Obama +1.1	+/- 5.8		63%
VA	13	303	235	Obama +0.8	+/- 4.8		61%
FL	29	332	206	Romney +0.7	+/- 5.3		59%
NC	15	347	191	Romney +2.6	+/- 5.0		81%
NE-2	1	348	190	Romney +6.5	+/- 6.8		94%
AZ	11	359	179	Romney +7.8	+/- 6.7		97%
MO	10	369	169	Romney +9.0	+/- 5.5		99%
GA	16	385	153	Romney +9.7	+/- 5.5		100%
MT	3	388	150	Romney +9.7	+/- 7.4		98%

Alternative Solution: Abandon hypothesis testing. Use interval estimates.

- Interval estimates quantify the effect size, not just statistical significance.
- They give a range of plausible values, not simply the verdict "reject" or "retain."

The Likelihood Principle

Likelihood Principle: Two data sequences (from possibly different experiments) with the same likelihood function provide the same evidence.

Likelihood Principle: Two data sequences (from possibly different experiments) with the same likelihood function provide the same evidence.

Comparing Evidence

- Sometimes we may wish to compare the evidence provided by two experiments.
- For example, *ceteris paribus*, if we wish to say that one study provides better evidence of climate change than another.
- Can we formalize the idea of "stronger", "similar", and "identical" evidence?

Formalizing Evidence

- Birnbaum [1962] is completely agnostic about how to represent evidence.
- Let Exp_Θ be the class of experiments whose underlying set of states of the world are equal to Θ.
- Pieces of evidence about Θ are represented by elements of a set EVIDENCE_Θ.
- There is a function $Ev : Exp_\Theta \times \mathbb{R}^n \to \text{EVIDENCE}_\Theta$ with
  - **Input:** An experiment $\mathcal{E}$ and a data sequence $x \in \mathbb{R}^n$ from that experiment.
  - **Output:** A piece of evidence $Ev(\mathcal{E}, x) \in \text{EVIDENCE}_\Theta$.

Likelihood Principle: Two data sequences (from possibly different experiments) with the same likelihood function provide the same evidence.

Two Experiments

Example:

- **Experiment 1:** Flip a coin of unknown bias twelve times.
- **Experiment 2:** Flip a coin of unknown bias until three heads are observed.

Two Experiments

Consider two classical experiments with the same set of states of the worlds Θ:

- $\mathcal{E}_1 = \langle \Theta, \Omega_1, \{P_\theta\}_{\theta \in \Theta}, \langle X_n \rangle_{n \in \mathbb{N}} \rangle$
- $\mathcal{E}_2 = \langle \Theta, \Omega_2, \{Q_\theta\}_{\theta \in \Theta}, \langle Y_n \rangle_{n \in \mathbb{N}} \rangle$

Two Experiments

The experiments may differ in any number of ways:

- They may have different sets of experimental outcomes
  - I.e., $\Omega_1 \neq \Omega_2$
- Different sequences of observations may be possible.
  - I.e., $\langle X_n \rangle_{n \in \mathbb{N}} \neq \langle Y_n \rangle_{n \in \mathbb{N}}$
- The probabilities of the same data sequence may differ, even if the world θ is held fixed.
  - I.e., $P_\theta \neq Q_\theta$

Likelihood Principle: Two data sequences (from possibly different experiments) with the same likelihood function provide the same evidence.

Same Likelihood Function

- Let $x = \langle X_1(\omega), \ldots X_m(\omega) \rangle$ be an outcome of the first experiment.
- Let $y = \langle Y_1(\omega), \ldots Y_n(\omega) \rangle$ be an outcome of the second experiment.
- Say x and y determine the same likelihood function if there is a constant $c_{x,y}$ such that

$$p_\theta(x) = c_{x,y} \cdot q_\theta(y)$$

for all $\theta \in \Theta$

Two Experiments

Example:

- **Experiment 1:** Flip a coin of unknown bias twelve times.
- **Experiment 2:** Flip a coin of unknown bias until three heads are observed.

Two Experiments: Same Data:

Outcome: Let $x = y$ be a data sequence consisting of 12 flips, with 9 tails, and a heads on the last flip.

Two Experiments: Same Likelihood Function

Example:

- **Experiment 1:** Flip a coin of unknown bias twelve times.
  - Regardless of the bias $\theta \in \Theta$ of the coin:

$$P_\theta(x) = \binom{12}{3} \cdot \theta^3(1-\theta)^9$$

- **Experiment 2:** Flip a coin of unknown bias until three heads are observed.
  - Regardless of the bias $\theta \in \Theta$ of the coin:

$$Q_\theta(x) = \binom{11}{2} \cdot \theta^3(1-\theta)^9$$

Note: Here, $c_{x,x} = \frac{\binom{12}{3}}{\binom{11}{2}} = \frac{220}{55} = 4$.

Two Experiments: Same Evidence

Example: According to the Likelihood Principle, the sequence of coin flips provides the same evidence about the bias of the coin, regardless of which of the two experiments is conducted.

Two Experiments: Same Evidence

- Some find this intuitive: your data is your data!
- Others don't: The fact that you could have flipped to coin for a longer or shorter time in the second experiment should affect the inferences you draw.

Useless counterfactuals

Here are some cases in which classical techniques make use of counterfactual dependencies that seem irrelevant.

LP and Classical Statistics

Upshot: Classical statistical methods violate the LP. Why?

Two Experiments: Same Likelihood Function

Example: Suppose two experimenters test the null hypothesis $\Theta_0 = \{\frac{1}{2}\}$ at the .05 level in the following two experiments.

- **Experiment 1:** Flip a coin of unknown bias twelve times.
- **Experiment 2:** Flip a coin of unknown bias until three heads are observed.

Same Likelihood Function, Different Verdicts

Experiment 1: Recall if the coin has bias θ, then the probability of k many heads is:

$$\binom{12}{k} \cdot \theta^k (1-\theta)^{12-k}$$

When $\theta = \frac{1}{2}$:

$$\theta^k (1-\theta)^{12-k} = (\frac{1}{2})^{12}$$

So the probability of three or fewer heads is:

$$P_\theta(x) = \left(\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right) \cdot \frac{1}{2^{12}}$$

Same Likelihood Function, Different Verdicts

Experiment 1: If you do the calculation, you'll find:

$$P_\theta(x) \approx 7.3\%.$$

So you would not reject the null hypothesis at the .05 level.

Same Likelihood Function, Different Verdicts

Experiment 2: Flip a coin of unknown bias until three heads are observed.

The probability of needing at least k many throws is:

$$\binom{k}{2} \cdot \theta^3 (1-\theta)^{k-3}$$

When $\theta = \frac{1}{2}$, this probability is equal to the following:

$$\binom{k}{2} \cdot \frac{1}{2^k}$$

So the probability of needing 12 or more tosses is:

$$1 - \left(\binom{11}{2} \frac{1}{2^{11}} + \binom{10}{2} \frac{1}{2^{10}} + \ldots + \binom{2}{2} \frac{1}{2^3} \right)$$

Same Likelihood Function, Different Verdicts

Experiment 2: If you do the calculation, you'll find this probability is about 3.3%. So you would reject the null hypothesis at the .05 level.

Same Likelihood Function, Different Verdicts

Moral: Classical hypothesis testing violates the LP.

Because of the close relationship between confidence intervals and hypothesis tests, so will confidence intervals.

So what?

Birnbaum's Theorem: LP is entailed by two principles nearly universally endorse by classical statisticians: (i) the conditionality principle, and (ii) the sufficiency principle.

Mixed Experiments

- Suppose you want to measure the charge of an electron.
- There are two experiments that you might conduct, and you think both are equally reliable.
- You flip a coin to decide which experiment to conduct.
- Randomly choosing among a collection of experiments is called a mixed experiment.

Conditionality Principle

Conditionality Principle: If you conduct a mixed experiment in which you observe x after conducting the component experiment $\mathcal{E}$, then your evidence is the same as if you had observed x after conducting $\mathcal{E}$ without randomizing among

Conditionality Principle

- Making this precise requires giving a formal definition of a mixed experiment, which I won't do.
- Important:
  - The randomizing device should not provide additional information about Θ:
    - I.e., The probability that the randomizing device takes particular values is not a function of $\theta \in \Theta$.
    - E.g., Don't flip the coin of unknown bias to choose which of the two experiments above to conduct. The extra coin flip gives you additional information!
  - Birnbaum is likewise agnostic about how to formalize the notion of evidence here.

Sufficiency Principle

Sufficiency: All evidence from an experiment concerning Θ in the experiment is available in a sufficient statistic.

Sufficiency Principle

- Note: A sufficient statistic may provide no evidence for any $\theta \in \Theta$. For instance, suppose you flip a coin to determine whether or not it will rain one year from today in Munich.
- **Major Lemma:** The set of likelihoods is a sufficient statistic, i.e., the function:

$$T : x \mapsto \langle P_\theta(x) \rangle_{\theta \in \Theta}$$

Proof of Birnbaum's Theorem:

- Let x and y be data from experiments $\mathcal{E}_1$ and $\mathcal{E}_2$ determining the same likelihood function, with constant $c_{x,y} \geq 0$.
- We want to show $Ev(\mathcal{E}_1, x) = Ev(\mathcal{E}_2, y)$.

Proof of Birnbaum's Theorem:

- Consider the mixed experiment $\mathcal{E}^*$ obtained by flipping a coin with bias $\frac{1}{c_{x,y}+1}$ to determine whether to conduct $\mathcal{E}_1$ or $\mathcal{E}_2$.
- Consider the probability of observing $\langle Heads, x \rangle$ and $\langle Tails, y \rangle$ in the mixed experiment.

Proof of Birnbaum's Theorem: In the mixed experiment $\mathcal{E}^*$:

$$\begin{aligned} p^*_\theta(\langle Heads, x \rangle) &= \frac{1}{c_{x,y}+1} \cdot p_\theta(x) \\ &= \frac{1}{c_{x,y}+1} \cdot c_{x,y} \cdot q_\theta(y) \\ &= (1 - \frac{1}{c_{x,y}+1}) \cdot q_\theta(y) \\ &= p *_\theta (\langle Tails, y \rangle) \end{aligned}$$

Proof of Birnbaum's Theorem:

- SUFFICIENCY and the lemma $\Rightarrow$

$$Ev(\mathcal{E}^*, \langle H, x\rangle) = Ev(\mathcal{E}^*, \langle T, y\rangle)$$

- CONDITIONALITY $\Rightarrow$

$$Ev(\mathcal{E}^*, \langle H, x\rangle) = Ev(\mathcal{E}_1, x)$$

- CONDITIONALITY $\Rightarrow$

$$Ev(\mathcal{E}^*, \langle T, y\rangle) = Ev(\mathcal{E}_2, y)$$

- Transitivity of equality then entails

$$Ev(\mathcal{E}_1, x) = Ev(\mathcal{E}_2, y)$$

which is what we wanted to show.

References I

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Mayo, D. and Spanos, A. (2011). Error statistics. In Forster, M. and Bandyopadhyay, P. S., editors, *Philosophy of statistics*, number 7 in Handbook of the Philosophy of Science, pages 153—198.