Lindleys Paradox

Sebastian.Fay

August 2023

History

The Problem of different outcomes depending on the approach, namely bayesian and frequentist, one is using was first discussed in Harold Jeffreys textbook in 1939 (Jeffreys, 1939).
It was given the name Lindleys paradox after Dennis Lindley called it a paradox in a 1957 paper (Lindley, 1957).

Recap: Hypothesis testing in general

This section follows narrowly Henze (2016)
In hypothesis testing one deals with an unknown distribution P of a random experiment.
In most hypothesis tests we assume one knows what family of distributions P could be an element of. For example we know that the data is gaussian, but we don't know the parameter μ and σ.
The goal is to give a reasonable estimation. This includes a measure of the probability of the estimation being wrong.
To do this, we have to formulate this mathematically:

$$P \in \{P_\theta | \theta \in \Theta\}$$

Where θ is the parameter(s) one want's to estimate. To extend our gaussian example this would be $\theta = (\mu, \sigma) \in \mathbb{R}^2$.
Now one defines the hypotheses.

$$\Theta_0 \subseteq \Theta$$
$$\Theta_A \subseteq \Theta \setminus \Theta_0$$

Often Θ_0 is only one specific element. For example when one assumes an underlying Ber(p) distribution for the data and one wants to test if $p \neq 0,5$, one could set $\Theta_0 = 0,5$.
We define our hypothesis now as follows:

$$H_0 := \{P_\theta | \theta \in \Theta_0\}$$
$$H_A := \{P_\theta | \theta \in \Theta_A\}$$

Where H_0 is often called the null hypothesis and H_A the alternative.
The idea of hypothesis testing now is that the data we acquired by running the experiment multiple times may make some θ more plausible than others.

Recap: Frequentist Approach

This section still follows narrowly Henze (2016)
The frequentist approach is probabily the one you are most familiar with, considering you learned it in school. Starting where we left of just now we formulate "we use the data to estimate which θ is more plausible" mathematically:
We define a test as

$$\phi : \mathbb{R}^n \to \{0, 1\}$$

We will put n (almost always one assumes i.i.d.) outcomes of n experiments in the test and get a decision 0 or 1. Where we define 1 as rejection.
Thereby it follows from our definitions of H_0 and H_A that:

$$\{x \in \mathbb{R}^n | \phi(x) = 1\}$$

If our observation is one of these x, H_0 is rejected so it's called the critical area (you can see it as a dashed line in the plot).

$$\{x \in \mathbb{R}^n | \phi(x) = 0\}$$

If our observation is one of these x, H_0 is not rejected.

One quickly sees that you could now have 2 kinds of errors.
Error of first order (the one we will control) and the error of the second order.
For the error of first order Θ_0 should be big (you almost never reject) and for the error of the second order Θ_0 should be small (you almost always reject). So we come to the conclusion that we can only control one error.
We decide to control the error of the first order.
Mathematically:

$$P_\theta(\phi(x) = 1) \leq \alpha \forall \theta \in \Theta_0$$

In other words the error of rejecting H_0 if H_0 is true is less than α. We call α the significance level (you may change it in the plot).

p-Value

For this Definition see also *P-Wert* (2023)
the p-Value is defined as

$$\inf\{\alpha | \phi_\alpha(x) = 1\}$$

Intuitively this means: how small must my significance niveau be, so that we reject H_0 for the given observation.
Thereby you can see the p-Value in the answer of the frequentist approach.

Intro: Bayesian Approach

The things discussed here you can find at *Lindley's paradox* (2023)
As the name suggests this approach uses Bayes Theorem.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In Bayesian Hypothesis testing we are interested in $P(H_0|k)$. In other words how likely do we think H_0 is after having the experiment outcome k. We calculate this by using Bayes Theorem. For example:

$$P(H_0|k) = \frac{P(k|H_0)\pi(H_0)}{P(k|H_A)\pi(H_A) + P(k|H_0)\pi(H_0)}$$

The Paradox

This section still follows an example from narrowly *Lindley's paradox* (2023)
Now I finally present you with an example (that is certainly not copied from the wikipedia article *Lindley's paradox* (2023))
Let's say in a city we observed 49,581 boys and 48,870 girls.
We are interested in $P(boys = 1)$.

Frequentist

1. We assume our data to be binomially distributed
2. We Set our Hypothesis $H_0 : p = 0.5$ and $H_A : p \neq 0.5$
3. We choose our $\alpha = 0.05$

The frequentist would now perform a 2 sided test. We will however use a one sided one, but this doesn't change the outcome:

Because the data is very large we approximate the data via a gaussian distribution, with

$$\mu = np = 98,451 \cdot 0.5 = 49,225.5$$
$$\sigma^2 = np(1-p) = 98,451 \cdot 0.5^2 = 24,612,75$$
$$P(X > x|\mu = 49,225.5) = \int_{x=49,225.5}^{98,451} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$
$$= \int_{x=49,225.5}^{98,451} \frac{1}{\sqrt{2\pi 24,612,75}} \cdot e^{-\frac{(u-49,225.5)^2}{224,612,75}} du \approx 0.0117$$

two sided it would be 0.0235 so the p-Value is certainly < 0.05

Therefore we reject H_0

Bayesian

1. We set our prior probabilities $\pi(H_0) = \pi(H_A) = 0.5$ (equally likely).
2. For $H_0 : p = 0.5$ we assume a binomial distribution and a Uniform distribution of p under H_A (all p are equally likely)

$$P(H_0|k) = \frac{P(k|H_0)\pi(H_0)}{P(k|H_A)\pi(H_A) + P(k|H_0)\pi(H_0)}$$
$$P(k|H_0) = \binom{n}{k} \cdot 0.5^k \cdot (1-0.5)^{n-k} \approx 1.95 \cdot 10^{-4}$$
$$P(k|H_A) = \int_0^1 \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} dp \approx 1.02 \cdot 10^{-5}$$
$$P(H_0|k) \approx 0.95$$

Thereby H_0 is much more likely under the data than H_A.
Consequently H_0 is not rejected.

Discussion Paradox

This discussion still follows narrowly *Lindley's paradox* (2023)

Truly a Paradox?

Now, does this tell us that our hypothesis testing is fundamentally flawed and only utter chaos rules?
The short answer is: No
The careful reader may have noticed me subtlety using different wording when talking about the hypothesis depending on the approach used.
The reason we get different answers is, because we answered different questions. The Frequentist doesn't care what H_A looks like. The question underlying here is: "does H_0 explain the data well enough?". Whereas the Bayesian approach defines the H_A and asks: "is H_0 more likely than H_A?".

The "Paradox" may be phrased as: "H_0 may not describe the data well enough but still better than H_A." So the paradox was never a paradox to begin with.

"Resolving" the Paradox

But this still leaves a bad aftertaste. We know that this is no paradox, but this doesn't change the fact that both approaches are reasonable and we still want an answer for our test.
Proposed methods are (*Lindley's paradox*, 2023):
Fitting the significance level to the data size. According to the law of large numbers as the data size gets bigger there are less fluctations. Naaman (Naaman, 2016) proposed to then lower the significance level to control the false positives (you may try that yourself).

Using an uninformative prior (*Prior probability*, 2023). To be honest, matching a gaussian distribution against an uniform one is a bit "unfair" when testing binomially distributed data. The idea of an uninformative prior is to use an objective prior. By using a gaussian prior we assumed more than we did for the Uniform one (you may also switch this in the app).

Extentions

In the plot you may change parameters that also change the outcomes.
The following calculations are for one sided test. In the plot 2 sided is used. This is because in the plot we want to move p in the full range, but the formulas I gave you earlier are for one sided tests, so you could calculate the following values with the formulas given earlier.

Varying sample (with the same ratio)

For a sample size of 100 persons under the same ratio you would have 50.36 boys. Because 0.36 boys doesn't make sense we round it to 50. Under these parameters one would get:
p-Value: $0.47087 > 0.05$
Bayes: $P(H_0|k) = 0.899$
A little deviation is hard to statistically reject with a small sample, so the p-Value becomes much bigger. The certainty for both approaches is generally higher the larger the sample is.

Varying ratio

Assume we have a sample of 100 people with 55 boys. One would calculate then:
p-Value:$0.398 > 0.05$
Bayes: $P(H_0|k) = 0.83$
Both approaches are hypothesis tests so it makes sense, that increasing the distance between the H_0 prediction and the actual numbers reduces the probability of H_0 being true.

Conclusion

We learned multiple things. Firstly, that naming in academia can be cruel, secondly maybe a little bit of hypothesis testing and lastly and most important: that the Lindley Paradox isn't a real paradox. It is simply a disagreement, on whether a hypothesis should be rejected or not, by two different approaches. Both approaches are valid and agree most of the time. The disagreement follows from the fact, that each approach answers a slightly different question. We also discussed ways to reduce the disagreement in the approaches. Armed with this knowledge I hope you can try to test out Lindleys Paradox a bit more and wish you a good time with the application as a whole.

References

Henze, P. D. N. (2016, January). *Video: Hypothesis testing.* Retrieved from `\url{https://www.youtube.com/watch?v=4YwYYAwimLk&list=PLfk0Dfh13pBPH8ABVC9WBiOb5Vh4WdE7M&index=16}` (Accessed: August 2, 2023)

Jeffreys, H. (1939). *Theory of probability.* `https://mathscinet.ams.org/mathscinet/relay-station?mr=0000924`. (Accessed: August 21, 2023)

Lindley, D. (1957). *A statistical paradox.* https://academic.oup.com/biomet/article-abstract/44/1-2/187/269805?redirectedFrom=fulltext&login=false. (Accessed: August 21, 2023)

Lindley's paradox. (2023, August). Retrieved from \url{https://en.wikipedia.org/wiki/Lindley%27s_paradox} (Accessed: August 2, 2023)

Naaman, M. (2016). *Almost sure hypothesis testing and a resolution of the jeffreys-lindley paradox.* https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-10/issue-1/Almost-sure-hypothesis-testing-and-a-resolution-of-the-Jeffreys/10.1214/16-EJS1146.full. (Accessed: August 21, 2023)

Prior probability. (2023, August). Retrieved from \url{https://en.wikipedia.org/wiki/Prior_probability} (Accessed: August 2, 2023)

P-Wert. (2023, August). Retrieved from \url{https://de.wikipedia.org/wiki/P-Wert} (Accessed: August 2, 2023)