To p, or not to p?

Dr James Abdey[†]

[†]Department of Statistics
London School of Economics and Political Science

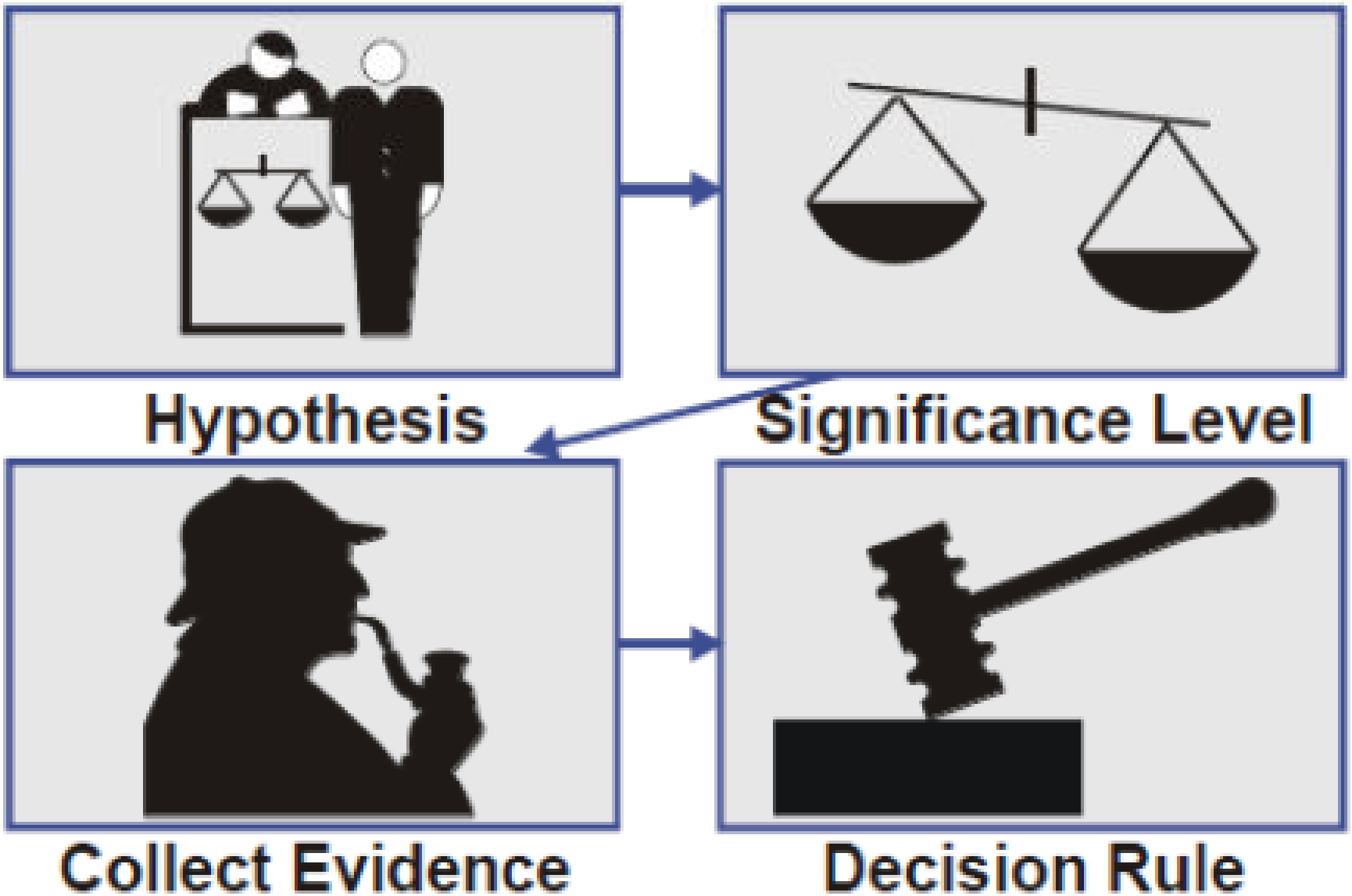
Overview

***P*-values** have become a staple measure used in hypothesis testing, an interesting application of which is the field of **forensic statistics**.

Advances in technology and forensic data collection allow statistics to be applied to the law – specifically the probabilistic assessment of a defendant's guilt based on data evidence.

This talk considers the use of p-values to determine the **likelihood of guilt** while attempting to appease the different 'schools' of hypothesis testing which advocate alternative approaches to testing – the battle between the **Frequentists** and the **Bayesians**.

The judicial process



Hypothesis testing: judicial analogy

In a criminal court, you put defendants on trial because you suspect they are guilty of a crime, but how does the trial proceed?

Determine the **null and alternative hypotheses**. The alternative hypothesis is your initial research hypothesis (the defendant is guilty). The null hypothesis is the logical opposite (the defendant is not guilty).

Select a **significance level** as the amount of evidence needed to convict. In court, the evidence must prove guilt **'beyond a reasonable doubt'**.

Collect evidence. Use a **decision rule** to make a judgement. If the evidence is:

- sufficiently strong, reject the null hypothesis
- not strong enough, fail to reject the null hypothesis. Note that failing to prove guilt does not prove that the defendant is innocent!

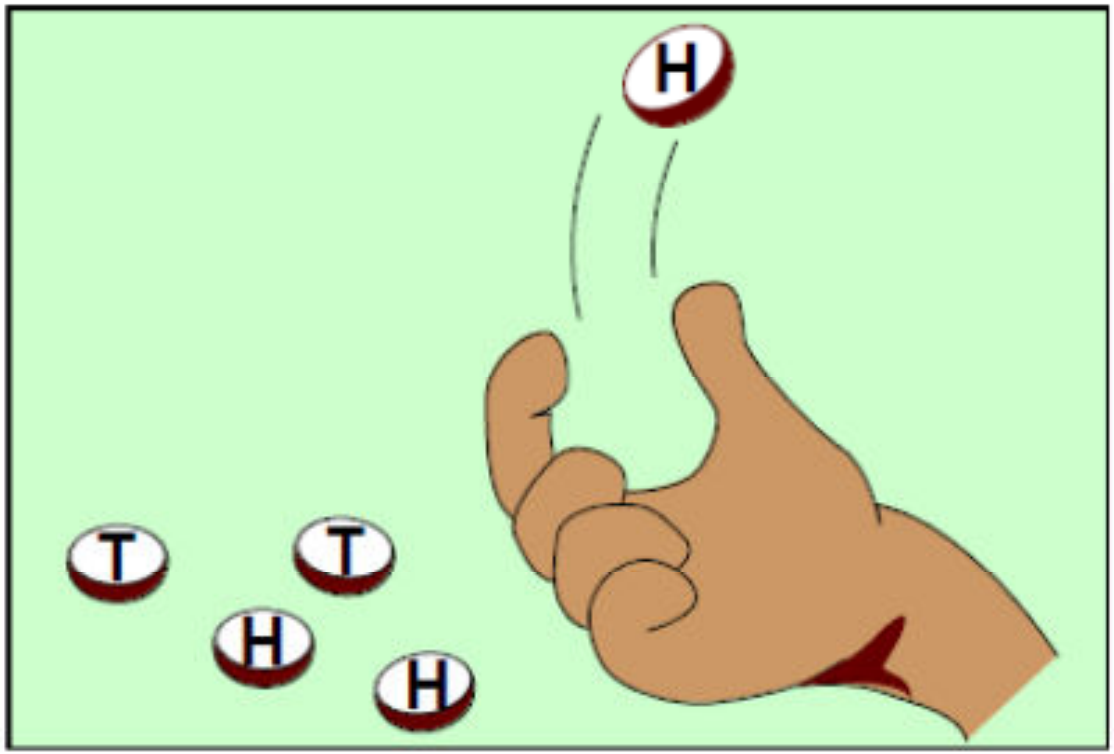
Statistical hypothesis testing follows this same logical path.

Coin example

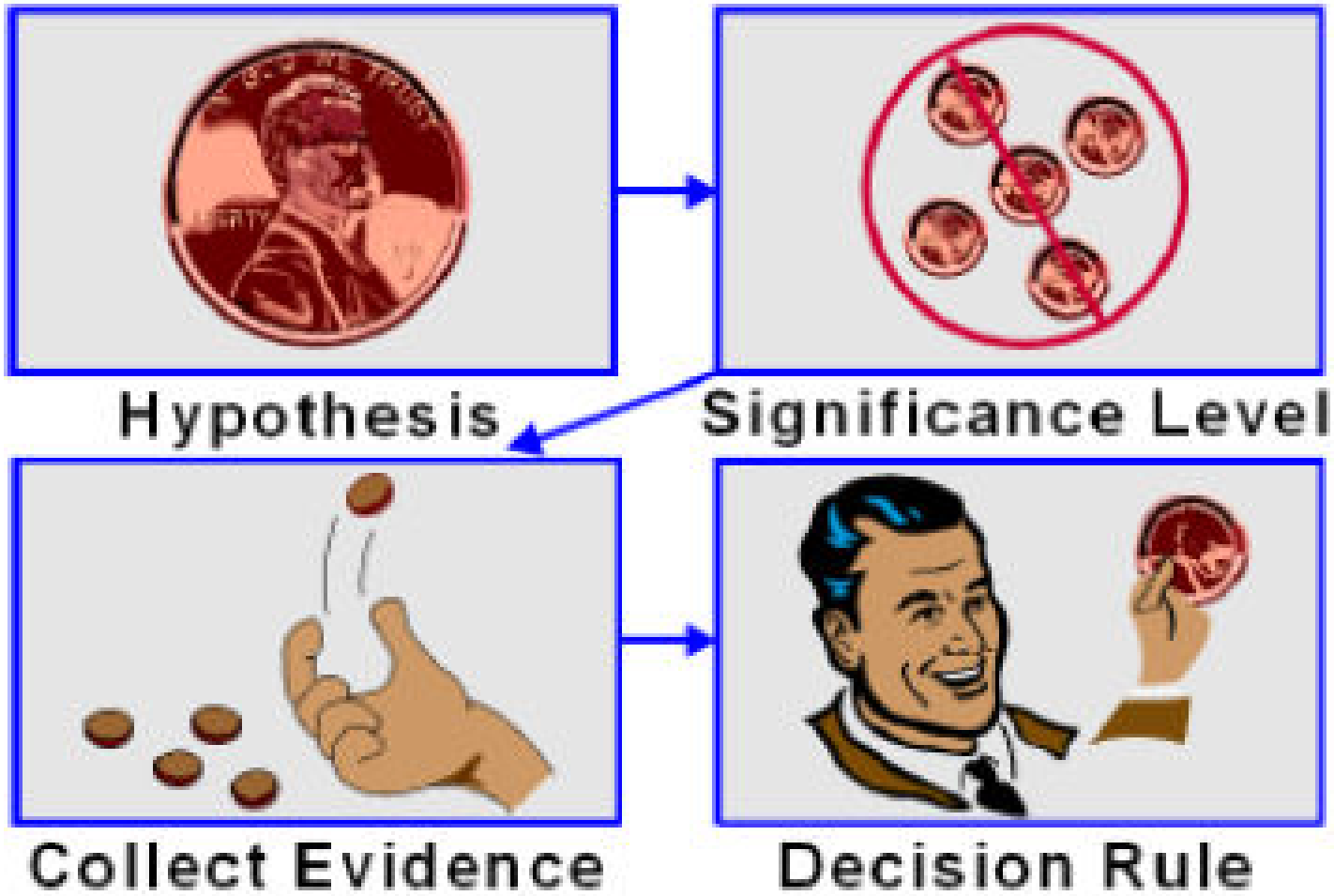
Suppose you want to know whether a coin is fair.

You cannot flip it forever, so you decide to take a sample.

Flip it five times and count the number of heads and number of tails.



Hypothesis testing: coin analogy



Coin example: test whether a coin is fair

You suspect that the coin is not fair but recall the legal example and begin by assuming the coin is fair.

You select a significance level. If you observe five heads in a row or five tails in a row, you conclude the coin is not fair; otherwise, you decide there is not enough evidence to show the coin is not fair.

You flip the coin five times and count the number of heads and number of tails.

You evaluate the data using your decision rule and make a decision that there is:

- enough evidence to reject the assumption that the coin is fair
- not enough evidence to reject the assumption that the coin is fair.

Types of errors

You used a decision rule to make a decision, but was the decision correct?

	ACTUAL	
DECISION	H_0 Is True	H_0 Is False
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

Types of errors

Recall that you start by assuming that the coin is fair.

The **probability of a Type I error**, often denoted α, is the probability that you reject the null hypothesis when it is true. It is also called the significance level of a test.

- In the legal example, it is the probability that you conclude the person is guilty when he or she is innocent.
- In the coin example, it is the probability that you conclude the coin is not fair when it is fair.

Types of errors

The **probability of a Type II error**, often denoted β, is the probability that you fail to reject the null hypothesis when it is false.

- In the legal example, it is the probability that you fail to find the person guilty when he or she is guilty.
- In the coin example, it is the probability that you fail to find the coin is not fair when it is not fair.

The **power** of a statistical test is equal to $1 - \beta$ where β is the Type II error rate. This is the probability that you correctly reject the null hypothesis.

The decision rule

It is important to clarify that:

- α, the probability of a **Type I error**, is specified by the experimenter *before* collecting data
- the ***p*-value** is calculated from the collected data.

In most statistical hypothesis tests, you compare α and the associated *p*-value to make a decision.

Remember, α is set ahead of time based on the circumstances of the experiment.

The level of α is chosen based on the cost of making a Type I error. It is also a function of your knowledge of the data and theoretical considerations.

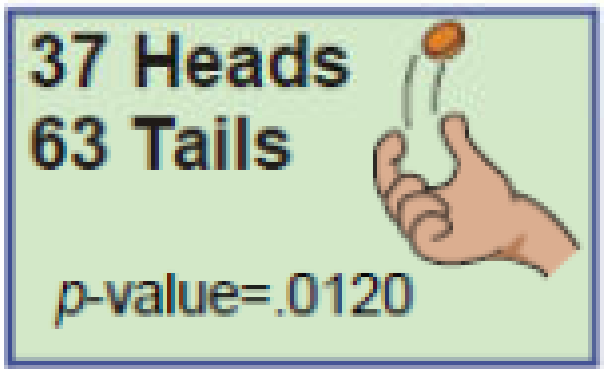
The decision rule

In general, you compare α and the *p*-value in order to:

- **reject** the null hypothesis if the *p*-value $< \alpha$
- **fail to reject** the null hypothesis if the *p*-value $\geq \alpha$.

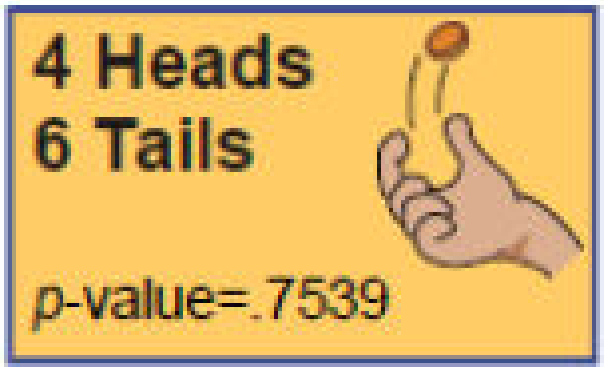
Effect size influence

Flip a coin 100 times and decide whether it is fair.



Sample size influence

Flip a coin and get 40% heads and decide whether it is fair.



The Frequentists and the Bayesians

When applying statistical methods to evaluate forensic evidence, the **likelihood ratio** is favoured due to its assessment of the **plausibility of both propositions** (e.g. 'guilty' versus 'not guilty', or 'Mr X committed the crime' versus 'Another person committed the crime').

This term is then used to update the prior odds to give the posterior odds in a Bayesian framework.

We refer to this likelihood ratio as the **value of evidence**.

The Frequentists and the Bayesians

In contrast, the well-known p-value considers the strength of evidence in support, or otherwise, of some null hypothesis, H_0.

However, it is possible to take into account the behaviour of the p-value under both H_0 and the alternative hypothesis, H_1, by **deriving the density functions of the p-value under both hypotheses**.

By taking the **ratio of these density functions**, a likelihood ratio statistic can be constructed which retains the spirit of the likelihood ratio described above (by considering the plausibility of both hypotheses), while using a readily computable p-value.

It can be shown that this approach leads to **consistently interpreted evidence**, hence avoiding the usual contradictions in interpretation.

P-value distributions

Bayesian criticism of the p-value stems from its failure to consider H_1, although we will advocate consideration of the p-value distribution under H_1 as well as under H_0.

Consider a test statistic, X_n, with a sampling distribution dependent on the sample size, n. Suppose X_n is used to test $H_0 : \theta = a$ against $H_1 : \theta = b$, where $a < b$. Let X_n have the left-continuous distribution function $F_{X_n}(x_n) = P(X_n \leq x_n)$ under H_0, for realised x_n. The p-value statistic, i.e. significance probability, is then a random variable, P, whose realised value, p, is calculated for continuous test statistics as:

$$p = P(X_n > x_n) = 1 - F_{X_n}(x_n) = \bar{F}_{X_n}(x_n) \qquad (1)$$

hence p-values are one-to-one transformations of the random variable X_n, so are themselves random variables. **Advantages of the p-value include its simplicity, i.e. a real number restricted to the unit interval, and also its universal application across test statistics with any distribution under H_0 due to the transformation in (1).**

Distribution under the null hypothesis

Let p be the realised p-value with $F_P(p \,|\, \mathrm{H}_0)$ being the corresponding distribution function under H_0, then using (1):

$$\begin{aligned} F_P(p \,|\, \mathrm{H}_0) = P(P \leq p \,|\, \mathrm{H}_0) &= P(1 - F_{X_n}(x_n) \leq p \,|\, \mathrm{H}_0) \\ &= 1 - F_{X_n}(F_{X_n}^{-1}(1-p)) \\ &= 1 - (1-p) \\ &= p \end{aligned}$$

for $p \in [0, 1]$. Hence this gives a density function $f_p(p \,|\, \mathrm{H}_0) = 1$, consistent with a **uniform density over the unit interval, i.e. $P \in U(0, 1)$**, independent of the test statistic distribution, sample size and effect size.

Therefore, under H_0 it is impossible to distinguish p-values obtained from small and large samples as well as between tests engineered to have high power and those less powerful.

Distribution under the alternative hypothesis

Denoting the left-continuous distribution function of X_n under the alternative hypothesis by $G_{X_n}(x_n)$, then the p-value distribution function under H_1, $F_P(p \,|\, \mathrm{H}_1)$, becomes:

$$F_P(p \,|\, \mathrm{H}_1) = P(P \leq p \,|\, \mathrm{H}_1) = P(1 - F_{X_n}(x_n) \leq p \,|\, \mathrm{H}_1) = 1 - G_{X_n}(F_{X_n}^{-1}(1-p))$$

which clearly depends on the test statistic's distribution under both hypotheses.

The corresponding density function is merely the likelihood ratio for the hypothesis test since, by application of the chain rule:

$$f_P(p \,|\, \mathrm{H}_1) = \frac{\partial}{\partial p} F_P(p \,|\, \mathrm{H}_1) = \frac{g_{X_n}(F_{X_n}^{-1}(1-p))}{f_{X_n}(F_{X_n}^{-1}(1-p))}$$

where the densities of X_n under the null and alternative hypotheses are $f_{X_n}(x_n)$ and $g_{X_n}(x_n)$, respectively.

P-value likelihood ratio

Considering the behaviour of the p-value under both H_0 and H_1 it is easy to construct a likelihood ratio based on a conventional p-value:

$$LR_P = \frac{f_P(p \mid \mathrm{H}_0)}{f_P(p \mid \mathrm{H}_1)} = (f_P(p \mid \mathrm{H}_1))^{-1}$$

since $f_P(p \mid \mathrm{H}_0) = 1$ due to the uniformity of p-values under H_0. LR_P can also be viewed as the ratio of *test statistic* densities under H_0 and H_1 since:

$$(f_P(p \mid \mathrm{H}_1))^{-1} = \frac{f_{X_n}(F_{X_n}^{-1}(1-p))}{g_{X_n}(F_{X_n}^{-1}(1-p))} = \frac{f_{X_n}(x_n)}{g_{X_n}(x_n)}.$$

Compare this with the forensic-based likelihood ratio:

$$\frac{P(E \mid \mathrm{H}_0)}{P(E \mid \mathrm{H}_1)}$$

where E denotes 'evidence', and H_0 and H_1 denote the prosecutor and defence propositions, respectively; that is the ratio of the *data* densities under H_0 and H_1.

Example

Suppose trace evidence in the form of glass fragments is recovered from a suspect's clothing. Assume initially that the suspect provides an explanation for the origin of the fragment, say source A. So we have:

- H_0: the fragment of glass on the suspect's clothing came from the window at the crime scene;
- H_1: the fragment of glass on the suspect's clothing came from source A.

Let θ be the mean refractive index of the source of the fragment found on the suspect's clothing. Under the prosecutor's proposition, H_0, suppose $\theta = \theta_0$ where θ_0 is the mean refractive index of the window at the crime scene. Under the defence's proposition, H_1, then $\theta = \theta_1$ where θ_1 is the mean refractive index of source A. Of course, if $\theta_0 = \theta_1$ then the two candidate sources are indistinguishable. However, suppose $\theta_0 \neq \theta_1$ and assume that the variance of measurements within a source is constant across all sources, i.e. σ^2.

Example

In the Frequentist world we have:

$$\mathrm{H}_0 : \theta = \theta_0 \qquad \text{vs.} \qquad \mathrm{H}_1 : \theta = \theta_1.$$

Under H_0, for n recovered glass fragments, the test statistic:

$$\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma} \sim N(0, 1)$$

where the random variable $\bar{X}$ denotes the mean measurement of the fragments.

Without loss of generality, suppose $\theta_0 < \theta_1$, for sample measurement $\bar{x}$ the p-value is:

$$p = 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \theta_0)}{\sigma}\right).$$

Example

The p-value likelihood ratio statistic is, noting the effect size is $\delta = (\theta_1 - \theta_0)/\sigma$:

$$LR_P = \frac{\phi(Z_p)}{\phi(Z_p - \sqrt{n}\delta)} = \frac{\phi\left(\frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma}\right)}{\phi\left(\frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma} - \sqrt{n}\delta\right)} = \frac{\phi\left(\frac{\sqrt{n}(\bar{x}-\theta_0)}{\sigma}\right)}{\phi\left(\frac{\sqrt{n}(\bar{x}-\theta_1)}{\sigma}\right)}$$

which is simply:

$$\frac{(2\pi)^{-1/2}\exp\{-n(\bar{x}-\theta_0)^2/2\sigma^2\}}{(2\pi)^{-1/2}\exp\{-n(\bar{x}-\theta_1)^2/2\sigma^2\}} = \exp\left\{\frac{n((\bar{x}-\theta_1)^2 - (\bar{x}-\theta_0)^2)}{2\sigma^2}\right\}.$$

Example

However, although LR_P is calculated using the p-value, p, this is identical to the Bayesian-derived likelihood ratio.

To see this, the Bayesian hypotheses are:

$$\mathrm{H}_0 : \bar{X} \,|\, \mathrm{H}_0 \sim N(\theta_0, \sigma^2/n) \qquad \text{vs.} \qquad \mathrm{H}_1 : \bar{X} \,|\, \mathrm{H}_1 \sim N(\theta_1, \sigma^2/n).$$

Hence the Bayesian likelihood ratio is simply:

$$\begin{aligned} LR_B = \frac{f_{\bar{X}}(\bar{x} \,|\, \theta_0,\ \mathrm{H}_0)}{f_{\bar{X}}(\bar{x} \,|\, \theta_1,\ \mathrm{H}_1)} &= \frac{(2\pi\sigma^2/n)^{-1/2} \exp\{-n(\bar{x}-\theta_0)^2/2\sigma^2\}}{(2\pi\sigma^2/n)^{-1/2} \exp\{-n(\bar{x}-\theta_1)^2/2\sigma^2\}} \\ &= \exp\left\{\frac{n((\bar{x}-\theta_1)^2 - (\bar{x}-\theta_0)^2)}{2\sigma^2}\right\}. \end{aligned}$$

Example

So, both approaches to inference are equivalent ($LR_B = LR_P$).

The advantage with the likelihood ratio, LR_P, is that p-values are ubiquitous, routinely appearing in the output of numerous statistical packages and form a part of the scientific community's research vocabulary.

By accounting for the distribution of the p-value under both hypotheses, as opposed to just considering H_0, a more suitable evaluation of the strength of the evidence using the p-value can be obtained.

Example

Suppose $\theta_0 = 1.518458$, $\sigma = 4 \times 10^{-5}$,then the Bayesian likelihood ratio is approximately:

$$LR_B = 100\sqrt{n}\exp\left\{-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right\} = 100\sqrt{n}\exp\{-z_n^2/2\}$$

where $z_n = \sqrt{n}(\bar{x}-\theta_0)/\sigma$, the standardised distance between $\bar{x}$ and θ_0.

Similarly the p-value likelihood ratio is approximately:

$$LR_P = 100\sqrt{n}\exp\left\{-Z_p^2/2\right\}.$$

Here, for a two-sided test, Z_p is defined as the $(1-p/2)$th percentile and so $z_n = Z_p$ resulting in $LR_B = LR_P$, as demonstrated above.

Example

The following table reports, for this numerical example, the likelihood ratio values for a variety of sample sizes for three standardised distances between $\bar{x}$ and θ_0 ($z_n = 2$, 2.6 and 3.3).

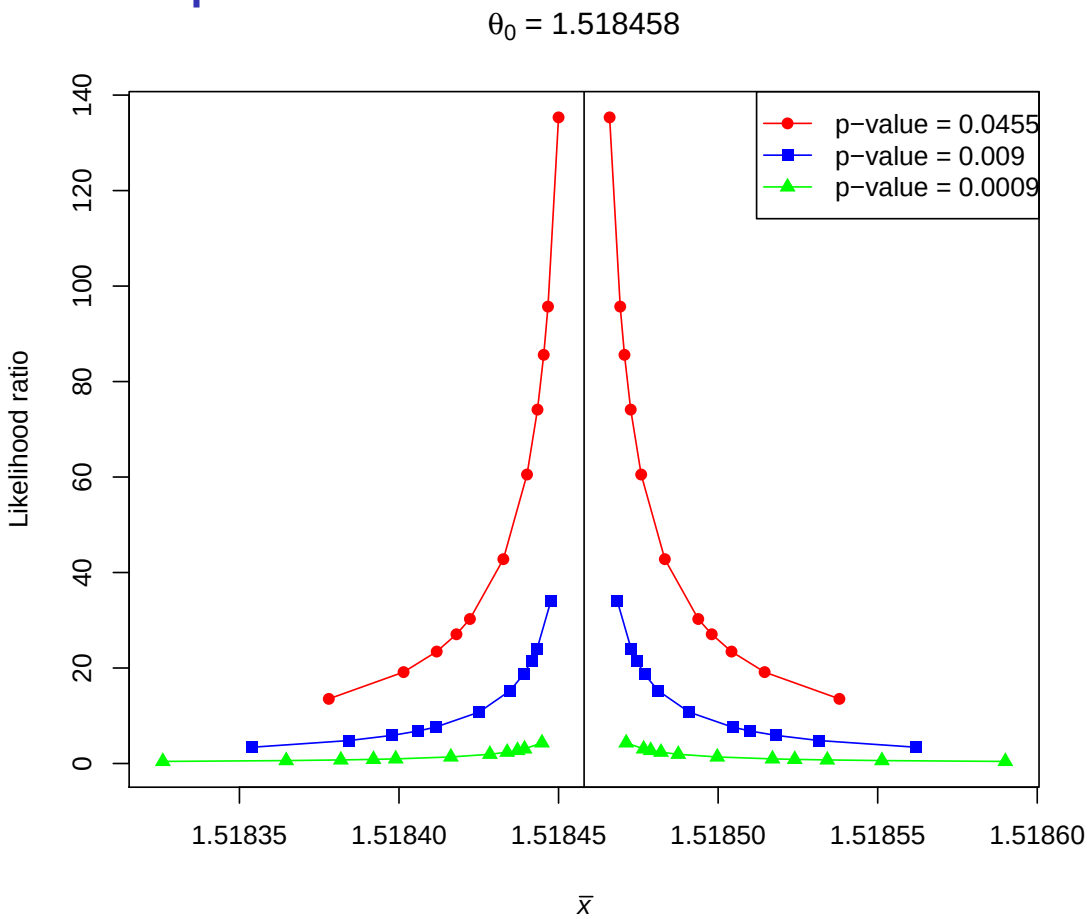
The subsequent figure displays these results.

These z_n values correspond with p-values of 0.0455, 0.009 and 0.0009 respectively such that classical hypothesis testing would reject the null hypothesis at the 5%, 1% and 0.1% significance levels respectively, although once the behaviour of the p-value under the alternative hypothesis is taken into account, it is seen that only for $n \leq 5$ in the case of $z_n = 3.3$ is the alternative hypothesis the more plausible one (the likelihood ratio is less than 1), despite 'significant' p-values throughout.

Example

n	$z_n = 2$ (p-value $= 0.0455$)	$z_n = 2.6$ (p-value $= 0.009$)	$z_n = 3.3$ (p-value $= 0.0009$)
1	13.53	3.40	0.43
2	19.14	4.82	0.61
3	23.44	5.90	0.75
4	27.07	6.81	0.86
5	30.26	7.61	0.97
10	42.80	10.77	1.37
20	60.52	15.23	1.93
30	74.13	18.65	2.36
40	85.59	21.53	2.73
50	95.70	24.08	3.05
100	135.34	34.05	4.32

Example



Likelihood ratio profiles for three standardised distances between $\bar{x}$ and θ_0: $z_n = 2$ (p-value $= 0.0455$), $z_n = 2.6$ (p-value $= 0.009$) and $z_n = 3.3$ (p-value $= 0.0009$). Each plotted point corresponds to a different sample size: $n = 100$, 50, 40, 30, 20, 10, 5, 4, 3, 2 and 1 as each profile moves further away from $\theta_0 = 1.518458$.

Conclusions

This talk has presented a **likelihood ratio approach to evidence evaluation** which leads to consistent conclusions for both Bayesians and Frequentists.

The approaches yield **identical inferential conclusions** when testing a population mean parameter under both simple and composite forms of the alternative hypothesis.

Hence a small (significant) p-value can still indicate strong support for a null hypothesis once the alternative hypothesis is taken into account.

These results highlight that results obtained using p-values are consistent with those obtained using Bayesian approaches, provided the distribution of the p-value under both hypotheses is taken into consideration when drawing inferential conclusions.