7. Schools of statistics

Contents

(All variables are real and one-dimensional unless otherwise specified.)

1. Is my coin fair?

Scientists often need to justify a claim based on experimental data. This involves **testing**, which is subtly different from but closely related to **modelling**. Take a coin as an example.

- Testing: is it fair?
- Modelling: how fair is it?

The former is a yes-no question, whereas the latter numerically asks for the coin's property. Still, both questions requires **statistics** or **inferential statistics** to be specific, as opposed to descriptive statistics.

No wonder is the subject related to probability—but how? Statistics becomes equally confusing when one really tries to interpret the numbers it gives. In fact, the **philosophically uncertain** nature of probability leads to, at least, three **practically distinct** paradigms in statistics.

The inconvenient consequence is that while all scientists inevitably do some statistics, few of them know what they are actually doing.

1.1 Overview

Frequentism first gave rise to *Ronald Fisher*'s **significance test**. Fisher pioneered frequentist statistics for his research in biology, but he is commonly attributed as the father of modern statistics. He proposed measuring a hypothesis's **statistical significance**, aka ***p*-value**.

Based on Fisher's idea, mathematicians *Jerzy Neyman* and *Egon Pearson* formulated their **hypothesis test**, which they claimed to generalize Fisher's theory. However, Fisher was irritated and strongly opposed their work. Neyman later also invented **confidence interval**, which further highlights the frequentist nature of probability.

In contrast, the idea of Bayesianism remains relatively intact throughout history: one should update his **prior** belief to a **posterior** belief according to an evidence-based **Bayes factor**, then his posterior belief constructs a **credible interval**. Because of its apparent subjectivity, scientists preferred frequentism until a physicist *Harold Jeffreys* (who influenced *Edwin Jaynes*) put forth its usage in science.

1.2 Formulation

Let us compare their theories with a coin. We toss a coin for $N = 100$ times. It gives a head for $h^* = 60$ times and a tail for $t^* = N - h^* = 40$ times. Let r be the coin's true probability of head. We know that the number of heads H follows a binomial distribution $B(h; N, r)$.

$$P(H = h \mid N, r) = B(h; N, r) = \binom{N}{h} r^h (1 - r)^{N-h}$$

Now we ask ourselves two questions.

- Testing: is the coin fair? (Yes or no.)
- Modelling: what is r? (Number.)

For testing, we will first set up a **null hypothesis** $H_0 : r = 0.5$, which is true if the coin is fair. We will assume H_0 is true, then we will assess the data and see if they **nullify** (i.e. overthrow) the hypothesis. For modelling, we hope to get a reasonable range of r.

2. Fisher's significant coin

2.1 Testing

Fisher uses a **significance test** that measures H_0's **significance** p, aka ***p*-value**. It is the conditional probability of the observation on some hypothesis.

$$p \cong P(\mathrm{observation} \mid \mathrm{hypothesis})$$

It is rigorously defined as the conditional probability to observe **the observed data or more "extreme" data** given that H_0 is correct. The meaning of "extreme" depends on context; here, it means measuring $h > h^*$. We should consider data more extreme than our observation as well because we have observed $h = h^* = 60$ at some point if the final observation is, say, $h = 80$ instead.

$$\therefore p = 2P(H \geq h^* \mid H_0)$$

The leading factor 2 accounts for the fact that it is as improbable to get many tails as to get many heads. (Check statistics textbooks for the difference between **one-tailed** and **two-tailed** events.)

Fisher compares p with a **significance level** α, which is specified before tossing the coin. If $p < \alpha$, Fisher rejects H_0 with the following arguments.

- If H_0 is true, p is the **frequency** of observation after many experiments.
- If p is small, we will not get the observation easily.
- But now we get the observation, so the assumption H_0 is unlikely true.

A common choice is $\alpha = 0.05$, so Fisher on average wrongly rejects H_0 once every $1/\alpha = 20$ tests he does. On the other hand, Fisher does not support H_0 even if $p > \alpha$; he merely does not reject it because the evidence does not instruct him to do so.

$$p = 2 \sum_{h=h^*}^{N} \binom{N}{h} 0.5^h 0.5^{N-h} \approx 0.115$$

As $p > \alpha$, Fisher does not reject H_0 and believes that getting $h^* = 60$ heads in $N = 100$ tosses does not hint at a biased coin.

2.2 Modelling

Whether he rejects H_0 or not, Fisher does not say anything about the value of r.

3. Neyman's confident coin

3.1 Testing

Neyman constructs a **hypothesis test**. Unlike Fisher, Neyman explicitly compares the null hypothesis H_0 with an **alternative hypothesis** like $H_1 : r \neq 0.5$. But similar to Fisher, Neyman also specifies a significance level $\alpha = 0.05$ before tossing the coin.

Neyman then defines a **rejection region** $R_{\text{rej}} \equiv [0, h_1) \cup (h_2, N]$. He rejects H_0 if h^* falls into R_{rej}. He hopes to restrict the frequency of type I error, i.e. believing in H_1 when H_0 is true, with α.

$$\sum_{h \in R_{\text{rej}}} P(H = h \mid H_0) = \alpha$$

As one equation does not fix two unknowns, we have to impose an extra constraint. Two common practices are

- the constraint of **median** $P(H < h_1 \mid H_0) = P(H > h_2 \mid H_0)$ and
- the constraint of **mode** $P(H = h_1 \mid H_0) = P(H = h_2 \mid H_0)$.

Because a binomial distribution is vertically symmetric, the two constraints are equivalent and both give $h_2 = N - h_1$. As H is discrete, we cannot solve the equation with $\alpha = 0.05$ exactly. The closest solution $R_{\text{rej}} \equiv [0, 41) \cup (59, N]$ corresponds to $\alpha \approx 0.0569$. Because $h^* = 60 > 59$, Neyman rejects H_0 in favour of H_1 with $\alpha \approx 0.0569$.

After that, Neyman computes the frequency β of type II error, i.e. believing in H_0 when H_1 is true. He defines this frequency as the test's **power**.

$$\beta = 1 - \sum_{h \in R_{\text{rej}}} P(H = h \mid H_1)$$

In this case, $\beta(r) = \sum_{h=h_1}^{h_2} B(h; N, r)$. What does this mean, though? Suppose Neyman repeats the tossing experiment with the same coin for many times. He rejects H_0 in a test if its number of heads falls into $R_{\text{rej}} = [0, 41) \cup (59, N]$.

- If H_0 is true, Neyman correctly believes in H_0 in **at least** $1 - \alpha \approx 94.3\%$ of all tests.
- If H_1 is true, Neyman correctly believes in H_1 in **at least** $100\,[1 - \beta(r)]\,\%$ of all tests, e.g. $r = 0.7$ gives 98.8% and $r = 0.6$ gives 54.3%. As $r \to 0.5$, the accuracy drops because the coin is merely a bit biased, making Neyman hardly able to distinguish it from a fair coin.

Neyman-Pearson lemma. H_1 need not be $\neg H_0$. One may use hypotheses like $H_1' : r < 0.5$, $H_1'' : r > 0.5$, or $H_1''' : r = 0.6$. It is however good to use a **simple hypothesis**, which specifies its distribution completely; for example, only H_1''' is simple here. When the hypotheses in question are both simple, the **Neyman-Pearson lemma** helps determine R_{rej}.

Consider $\begin{cases} H_0 : r = 0.5 \\ H_1 : r = 0.6 \end{cases}$ with distributions $\begin{cases} f_0(h) = B(h; N, 0.5) \\ f_1(h) = B(h; N, 0.6) \end{cases}$. Their **likelihood ratio** is $\Lambda(h) \equiv f_0(h) / f_1(h)$. The lemma states that if

$$P[\Lambda(h) \leq \eta \mid H_0] = \alpha\,,$$

then a test with $R_{\text{rej}} = \{h \mid \Lambda(h) \leq \eta\}$ produces a type II error the **least frequently** among all tests done at the same significance level α; in other words, the test is the **most powerful**. While the value of η is not important, the important consequence is that the condition $\Lambda(h) \leq \eta$ implies $h \geq h_c$.

$$\Lambda(h) = \left(\frac{2}{3}\right)^h \left(\frac{5}{4}\right)^N \leq \eta \quad \Rightarrow \quad h \geq h_c \equiv \frac{\log \eta + N \log(4/5)}{\log(2/3)}$$

With $\alpha = 0.05$ predefined, Neyman gets $h_c = 59$ by solving $P(h \geq h_c \mid H_0) = \alpha \approx 0.0443$. The corresponding probability of type II error is $P(h < h_c \mid H_1) = \beta \approx 0.378$. The lemma guarantees it to be the **minimum** among all tests with $\alpha \approx 0.0433$.

Because $h^* = 60 > h_c = 59$, Neyman rejects H_0 in favour of H_1. After applying the same rejection region on many tests,

- if H_0 is true, Neyman correctly believes in H_0 in at least $1 - \alpha \approx 95.6\%$ of the tests;
- if H_1 is true, Neyman correctly believes in H_1 in at least $1 - \beta \approx 62.3\%$ of the tests.

As you may recognize, this is in fact the **Neyman-Pearson detector** that we previously learnt (despite a slightly different formulation).

3.2 Modelling

Neyman argues with a **confidence interval**. First, he estimates its true value with $\hat{r} = h^*/N = 0.6$. If each toss earns one point with a head and zero points with a tail, r and $\hat{r}$ are the population and sample means of points. According to the **central limit theorem**, the random variable

$$Z = \frac{\hat{r} - r}{\hat{\sigma}/\sqrt{N}}$$

becomes **standard normal** ocne $N \to \infty$, where $\hat{\sigma} = \sqrt{\hat{r}\,(1-\hat{r})}$ is the sample standard deviation of points. Given some significance level α, Neyman solves $P(-z \leq Z \leq z) = 1 - \alpha$ for a cutoff value z, with which he solves $\dfrac{\hat{r} - r}{\hat{\sigma}/\sqrt{N}} = \pm z$ for the range of r.

$$r \in \left[\hat{r} - \frac{\hat{\sigma} z}{\sqrt{N}}, \hat{r} + \frac{\hat{\sigma} z}{\sqrt{N}}\right]$$

For example, $\alpha = 0.05$ gives $z \approx 1.96$ and thus $r \in [0.504, 0.696]$. Neyman calls this the $1 - \alpha = 95\%$ **confidence interval** of r.

Neyman **does not** think that $P(r \in [0.504, 0.696]) = 0.95$ because r is **certainly** inside or outside the **fixed** interval, i.e. the probability is either zero or one. On the contrary, Neyman believes that after many experiments, the **variable** interval $\left[\hat{r} - 1.96\dfrac{\hat{\sigma}}{\sqrt{N}}, \hat{r} + 1.96\dfrac{\hat{\sigma}}{\sqrt{N}}\right]$ **covers** (i.e. contains) r in 95% of all tests. The particular interval $[0.504, 0.696]$ may belong to the 95% of "good" intervals or the remaining 5% of "bad" intervals.

4. Jeffreys's credible coin

4.1 Testing

Jeffreys also compares the null hypothesis H_0 with an explicit alternative hypothesis H_1. Since Bayes' theorem states that

$$P(H_i \mid H = h^*) = \frac{P(H = h^* \mid H_i) P(H_i)}{\sum_j P(H = h^* \mid H_j) P(H_j)},$$

Jeffreys defines the **posterior odds** of H_0 against H_1 with

$$\frac{P(H_0 \mid H = h^*)}{P(H_1 \mid H = h^*)} = \underbrace{\frac{P(H = h^* \mid H_0)}{P(H = h^* \mid H_1)}}_{K} \frac{P(H_0)}{P(H_1)}.$$

On the right hand side, the first ratio is defined as the **Bayes factor** K, and the second ratio is defined the **prior odds** of H_0 against H_1. If nothing about H_0 or H_1 is known before an experiment, Jeffreys argues that one should not be biased towards any one. Therefore, the prior

odds is assumed to be one, making the posterior odds equal to K. Now consider $\begin{cases} H_0 : r = 0.5 \\ H_1 : r = 0.6 \end{cases}$.

$$K = \frac{B(h^*; N, 0.5)}{B(h^*; N, 0.6)} \approx 0.134$$

As $K < 1$ implies a higher posterior probability of H_1, Jeffreys rejects H_0 in favour of H_1. Originally, he believes in the hypotheses equally, but after seeing $h^* = 60$ heads in $N = 100$ tosses, his belief in H_1 becomes $1/K \approx 7$ times stronger than in H_0.

Prior distribution. If the alternative hypothesis is not simple, e.g. $H_1 : r \neq 0.5$, we need to assume its prior distribution $f(r \mid H_1)$. The null hypothesis $H_0 : r = 0.5$ in some sense has $f(r \mid H_0) = \delta(r - 0.5)$, for which $r = 0.5$ is infinitely more believable than other values.

$$P(H = h^* \mid H_1) = \int_0^1 B(h^*; N, r) f(r \mid H_1) \mathrm{d}r$$

The simplest choice is a uniform prior, meaning that all possible values of r are equally likely. (When nothing is known a priori, it is the most objective accoding to the principle of maximum entropy.)

$$K = \frac{0.5^{h^*} 0.5^{t^*}}{\int_0^1 r^{h^*} (1 - r)^{t^*} \mathrm{d}r} \approx 1.10$$

Because $K > 1$, Jeffreys does not reject H_0 in favour of H_1. While he originally does not prefer any hypothesis, he believes in H_0 slightly more after the experiment, i.e. he believes that r follows a Dirac distribution more than that it follows a uniform distribution.

The **arbitrary** choice of prior alters K. For example, if Jeffreys uses $f(r \mid H_1) = 6r(1 - r)$, which peaks at $r = 0.5$, the Bayes factor becomes $K \approx 0.767$ and rejects H_0 in favour of H_1. Instead of stubbornly believing in "r must be 0.5", Jeffreys now believes that r may attain some other values instead.

4.2 Modelling

Jeffreys calculates the posterior distribution of r with Bayes' theorem. Let $f(r)$ be its prior distribution.

$$f(r \mid H = h^*) = \frac{r^{h^*} (1 - r)^{t^*} f(r)}{\int_0^1 \rho^{h^*} (1 - \rho)^{t^*} f(\rho) \mathrm{d}\rho}$$

The distribution contains all inferred information of r, but it is conventionally expressed as a $100\,(1 - \alpha)\,\%$ **credible interval** $[r_1, r_2]$ so that it is comparable to Neyman's confidence interval.

$$\int_{r_1}^{r_2} f(r \mid H = h^*) \mathrm{d}r = 1 - \alpha$$

The equation does not uniquely define the interval. It is commonly fixed with three constraints:

- $P(r < r_1 \mid H = h^*) = P(r > r_2 \mid H = h^*)$, defining the **central interval**;
- $f(r = r_1 \mid H = h^*) = f(r = r_2 \mid H = h^*)$, defining the **smallest interval**; and
- $\mu - r_1 = r_2 - \mu$ so that the mean $\mu \equiv \int_0^1 r f(r \mid H = h^*) \mathrm{d}r$ lies at the interval's centre.

With a uniform prior and $\alpha = 0.05$, the three intervals are $[0.502, 0.691]$, $[0.503, 0.692]$, and $[0.504, 0.692]$. They all differ slightly from Neyman's $[0.504, 0.696]$. Although they sound similar, a credible interval and a confidence interval treat the true probability of head r with a fundamentally distinct philosophy.

- It is **probabilistic** for r to fall into a 95% credible interval. You believe in this event with 95% of belief.
- It is **deterministic** for r to fall into a 95% confidence interval—either yes or no. You know that 95% of the intervals cover r after many experiments, though.

5. Mutual disagreements

Fisher, Neyman, and Jeffreys mutually disagree with each other.

Against Fisher. Fisher tests a hypothesis without stating an alternative. He thinks that a hypothesis is **intrinsically** good or bad, whereas Neyman and Jeffreys argue that a hypothesis can only be **relatively** better or worse than another hypothesis. They challenge Fisher: what if H_0 is rejected? Furthermore, Jeffreys criticizes $p = 2P(H \geq h^* \mid H_0)$ for depending on $H > h^*$, which does not happen at all—how could a possibly true hypothesis be overthrown by some unobserved events?

Against Neyman. Neyman's probabilities of type I and type II errors make sense only after **a long run**, then how should one interpret a **single** test's result? Furthermore, although $R_{\text{rej}} = [0, 41) \cup (59, 100]$ rejects both $h^* = 60$ and $h^* = 90$, Neyman cannot systematically tell in which test he feels more confident. Fisher and Jeffreys thus criticize Neyman for **failing to respond** to evidence. Fisher also criticizes Neyman's philosophy fundamentally: rejecting H_0 should never imply supporting H_1.

Against Jeffreys. Jeffreys needs to specify prior probabilities, which Fisher and Neyman criticize for being completely **subjective**. They also dislike interpreting probability as "degree of belief" in science: isn't science supposed to be objective and independent from experimenters, is it?

(Interestingly though, it is proved that a credible interval coincides with a confidence interval in certain cases.)

So which approach should a general scientist follow? Basically any one as long as it is applied consistently. Unfortunately, the incompatible philosophies are often mixed up as something awkward, e.g. advocating H_1 because of the tiny *p*-value of H_0. This is nothing different from buying the moon is made of white chocolate because it cannot produce certain data if it is made of cheese.

$$P(H_1 \mid \text{observation}) \neq 1 - P(\text{observation} \mid H_0)$$

If you are interested, read [***Could Fisher, Jeffreys and Neyman Have Agreed on Testing?***](https://projecteuclid.org/euclid.ss/1056397485) (Berger, 2003).