Note 19. Binary hypothesis testing

Alex Fu

Fall 2022

1 Introduction to inference

The first two modules of probability and random processes have lived in the realm of **descriptive** probability so far — extracting useful information from a given *model*, a mathematical description of some part of reality. For instance, we found probabilities given a sample space, conditional probabilities given an event, expectations given a random variable, moments given a distribution, class properties given a chain, times given a process, and so forth.

But rarely or almost never will we have a given model in practice, such as a full probability space, distribution, or Markov chain. Even if we did, no model is a perfect description of reality — the only unsimplified model is reality itself. What we *have* access to is partial, imperfect information: a small, finite number of samples; empirical frequencies of observed outcomes in collected data; measurements marred with uncertainty, errors, noise, biases, perturbations, or more.

The more realistic *inverse* problem — determining a model given limited information — is the subject of **inferential** probability. Its common goals may seem quite familiar to you: estimation, approximation, prediction, learning, training, classification, decision, regression, analysis, etc.

Our basic setup for inference is as follows. The random variable X describes some *hidden*, *latent*, or *underlying* true **state** of the world, whose exact value or distribution is unknown to us. The **observation** Y is partially determined by X, following a given **model** $Y \mid X$, and partially by some other randomness.

We wish to "reverse" the model $Y \mid X$ to obtain "$X \mid Y$" somehow. Bayes' rule tells us that

$$p_{X|Y} = \frac{p_{Y|X} \cdot p_X}{p_Y},$$

but p_X is unknown! Thus our goal is to infer $\hat{X}$, a function of Y so that $\hat{X} \mid Y$ is almost $X \mid Y$. The inferred $\hat{X}$ is usually *optimal* in the sense of minimizing a cost, loss, or objective function, such as a probability $\mathbb{P}(X \neq \hat{X})$, or a distance $\|X - \hat{X}\|$, often mean squared error.

2 Definitions

Definition 1 (Setup for binary hypothesis testing).

Let $X \in \{0,1\}$ represent a choice between two **hypotheses**: the two distributions of the *null hypothesis* H_0 and the *alternative hypothesis* H_1. A single **observation** $Y \in \mathbb{R}$ is given, where $Y \sim H_0$ if $X = 0$ and $Y \sim H_1$ if $X = 1$.

We want to infer an optimal **test** or **decision rule** $\hat{X} = r(Y)$, which assigns 0 or 1 to every $y \in \mathbb{R}$ for which distribution y is more likely to have been drawn from. The assignments of $r(y) = 0$ or $r(y) = 1$ are *failing to reject* or *rejecting the null hypothesis* respectively.

The most common probabilities associated with hypothesis testing are

- The **probability of type I error**, **false alarm** (PFA), or **false positive** is the probability of incorrectly rejecting the null hypothesis,

$$\alpha := \mathbb{P}(\hat{X} = 1 \mid X = 0) = \mathbb{P}_{H_0}(\hat{X} = 1).$$

- The **significance level** $\alpha^* \in [0,1]$ is a preset upper bound on the PFA, often 0.05. It should be lower when false alarms, such as false diagnoses of cancer, are more "costly."

- The **probability of type II error**, **miss rate**, or **false negative rate** is the probability of failing to rejecting an incorrect null hypothesis,

$$\beta := \mathbb{P}(\hat{X} = 0 \mid X = 1) = \mathbb{P}_{H_1}(\hat{X} = 0).$$

- The **probability of correct detection** (PCD) or **power** of a test is the probability of correcting rejecting an incorrect null hypothesis,

$$1 - \beta = \mathbb{P}(\hat{X} = 1 \mid X = 1) = \mathbb{P}_{H_1}(\hat{X} = 1).$$

Our optimization problem is to find the test which maximizes PCD given a constrained PFA:

$$r^* = \operatorname*{argmax}_{r \colon \mathbb{R} \to \{0,1\}} \mathbb{P}(r(Y) = 1 \mid X = 1) \quad \text{s.t.} \quad \mathbb{P}(r(Y) = 1 \mid X = 0) \le \alpha^*$$

Definition 2 (Rejection region; acceptance region).

An equivalent characterization of a decision rule r is in terms of the **rejection region**

$$R := \{y \in \mathbb{R} : r(y) = 1\},$$

the values of Y for which the test rejects the null hypothesis. Its complement, the **acceptance region** $A = R^c$, works equally well.

Example 1 (Motivating examples for terminology).

The *null* hypothesis has its name because it is typically the hypothesis of *no effect*, also called a *negative* result: a lack of cancer, fire, or defect. The *alternative* hypothesis often describes a *positive* result, which may be undesirable: the presence of cancer, fire, or defect. The test or alarm tries to detect a positive result, and either hits or misses.

We work conservatively, as if the null hypothesis is true, until there is strong enough evidence to *reject* the null, usually a *significant* positive result, for which $\mathbb{P}_{H_1}(\text{result})$ is far more likely than $\mathbb{P}_{H_0}(\text{result})$. We will never "accept" the null hypothesis as true, only *fail to reject* it due to a lack of significant evidence pointing to the alternative hypothesis.

An important class of hypotheses involves a bit of confusing notation (to us). Let Θ be a space of parameters, and let θ^* be the true parameter by which the observation $X \sim \mathbb{P}(X = x; \theta^*)$ is drawn. Then the hypotheses are $H_0 \colon \theta \in \Theta_0$ and $H_1 \colon \theta \in \Theta_1$, where Θ_0 and Θ_1 partition Θ. We work with **simple** hypotheses, where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, so $H_X \colon \theta = \theta_X$.

An example is visualized below. Here, the two simple hypotheses are $H_0 \colon \mu = -1$ and $H_1 \colon \mu = 1$. The PFA and PCD of an arbitrary rejection region are the highlighted areas under the distributions of H_0 and H_1 respectively.

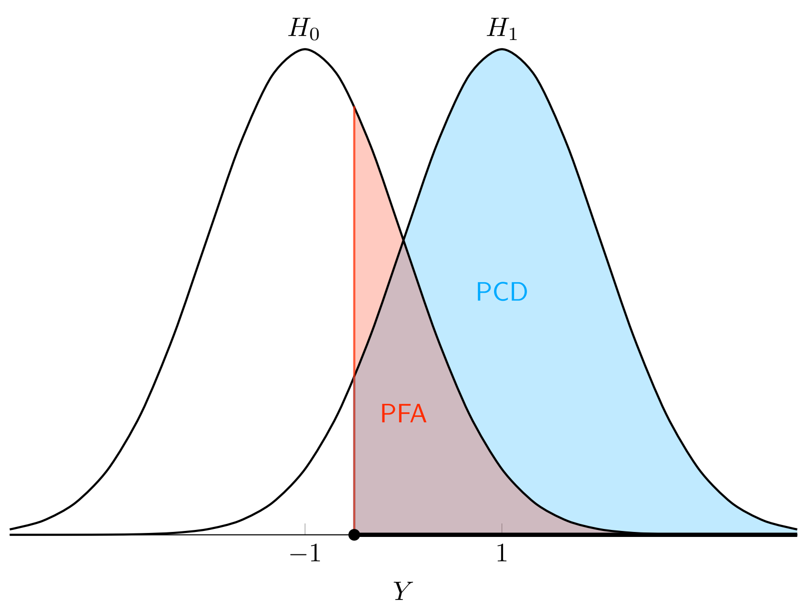

Figure 1: A binary hypothesis test.

We find that the region of overlap characterizes the conflict of binary hypothesis testing: greater rejection of the null $\hat{X} = 1$ increases PCD, but at the same time increases PFA. So a good test should selectively or *greedily* reject the null, favoring observations y that increase the PCD far more than they increase the PFA, which motivates the *likelihood ratio*.

3 The Neyman–Pearson likelihood ratio test

Definition 3 (Likelihood ratio).

The **likelihood ratio** of the observation $Y \in \mathbb{R}$ is the function

$$L(y) := \frac{\mathbb{P}_{H_1}(y)}{\mathbb{P}_{H_0}(y)} = \frac{f_{Y|X}(y \mid 1)}{f_{Y|X}(y \mid 0)},$$

the ratio between the probability that the value y is sampled from H_1 to the probability that y is sampled from H_0.

A natural starting point for $\hat{X}$ is the MLE: we reject the null at values for which $f_{Y|X}(y \mid 1) > f_{Y|X}(y \mid 0)$, or $L(y) > 1$. But a key problem is that we need *fine control* over $\alpha \leq \alpha^*$, as α^* is any significance level in $[0, 1]$. As the PCD increases with the PFA, we can always try to achieve the maximum $\alpha = \alpha^*$ without any loss of generality.

So, we want the possible values of α of the test to range over $[0, 1]$, but the MLE does not allow us this level of control. Instead, we can consider a **threshold test**:

$$r(y) = \begin{cases} 1 & \text{if } L(y) > \lambda \\ 0 & \text{otherwise,} \end{cases}$$

which has a threshold parameter $\lambda \in \mathbb{R}$ we can set depending on α^*. For instance, the MAP of X is a threshold test: if π is a prior distribution on X, then

$$\hat{X}_{\text{MAP}} = \mathbb{1}\{\mathbb{P}_{H_1}(y) \cdot \pi(1) > \mathbb{P}_{H_0}(y) \cdot \pi(0)\} = \mathbb{1}\left\{L(Y) > \frac{\pi(0)}{\pi(1)}\right\}.$$

In general, it seems that a threshold test allows us to freely set α as

$$\alpha = \mathbb{P}(r(Y) = 1 \mid X = 0) = \mathbb{P}(L(Y) > \lambda \mid X = 0).$$

But one problem arises when $L(Y)$ is discrete: even as $\lambda \in \mathbb{R}$ varies smoothly, the corresponding values of α will jump up and down discretely. For instance, consider the trivial example $H_0 = H_1$ and $\alpha^* = 0.5$, in which $L(Y) = 1$ and $\alpha \in \{0, 1\}$. This motivates *randomization at the threshold*:

$$r(y) = \begin{cases} 1 & \text{if } L(y) > \lambda \\ \text{Bernoulli}(\gamma) & \text{if } L(y) = \lambda \\ 0 & \text{if } L(y) < \lambda \end{cases}$$

for some **randomization constant** $\gamma \in [0, 1]$. The choices of $\gamma = 0$ or 1 bring us back to the simple threshold test, but let us see why randomization works more generally.

Consider the typical problematic scenario:

$$\mathbb{P}_{H_0}(L(Y) > \lambda) < \alpha^* < \mathbb{P}_{H_0}(L(Y) \geq \lambda),$$

so no choice of λ in a simple threshold test will allow us to set $\alpha = \alpha^*$ and maximize PCD. So, let us first approximate α^* as closely as we can without exceeding it:

$$\begin{aligned} \lambda^* &:= \operatorname*{argmax}_{\lambda \in \mathbb{R}} \alpha(\lambda) \quad \text{s.t.} \quad \alpha(\lambda) \le \alpha^* \\ &= \inf \{\lambda : \alpha(\lambda) \le \alpha^*\}. \end{aligned}$$

For convenience, we write $\alpha(\lambda) := \mathbb{P}_{H_0}(L(Y) > \lambda)$ for the value of α given by the choice of λ. If $\alpha(\lambda^*) = \alpha^*$ already, then we are done! Otherwise, we have found *the* threshold λ^* at which

$$\alpha(\lambda^*) + 0 \cdot \mathbb{P}_{H_0}(L(Y) = \lambda^*) < \alpha^* \le \alpha(\lambda^*) + 1 \cdot \mathbb{P}_{H_0}(L(Y) = \lambda^*),$$

the same scenario we started with. α^* must fall in one of these intervals, and now we can use randomization to interpolate to "fill the gap" between $\alpha(\lambda^*)$ and α^*:

$$\gamma = \frac{\alpha^* - \alpha(\lambda^*)}{\mathbb{P}_{H_0}(L(Y) = \lambda^*)} \in (0, 1].$$

Let us verify that we have achieved our initial goal: to be able to set $\alpha = \alpha^*$ for *any* $\alpha^* \in [0, 1]$. If we find the threshold $\lambda = \lambda^*$ and randomization constant γ as above, then

$$\begin{aligned} \alpha &= \mathbb{P}_{H_0}(r(Y) = 1) \\ &= \mathbb{P}_{H_0}(L(Y) > \lambda) + \mathbb{P}_{H_0}(L(Y) = \lambda, \text{Bernoulli}(\gamma) = 1) \\ &= \mathbb{P}_{H_0}(L(Y) > \lambda) + \gamma \cdot \mathbb{P}_{H_0}(L(Y) = \lambda) \\ &= \alpha^*. \end{aligned}$$

For good measure, we can also find the rejection region of the test as

$$R = \{y : L(y) > \lambda\} \cup \{y : L(y) = \lambda \wedge \text{Bernoulli}(\gamma) = 1\}.$$

What was the point of the above? Well, we have just derived the optimal hypothesis test.

Theorem 1 (Neyman–Pearson lemma).

The Neyman–Pearson likelihood ratio test is the uniformly most powerful test among all tests with significance level at most α^*. That is, the solution to

$$r^* = \operatorname*{argmax}_{r \colon \mathbb{R} \to \{0,1\}} \mathbb{P}(r(Y) = 1 \mid X = 1) \quad \text{s.t.} \quad \mathbb{P}(r(Y) = 1 \mid X = 0) \le \alpha^*$$

is a threshold test with randomization,

$$r^*(y) = \begin{cases} 1 & \text{if } L(y) > \lambda \\ \text{Bernoulli}(\gamma) & \text{if } L(y) = \lambda \\ 0 & \text{if } L(y) < \lambda \end{cases}$$

for some threshold $\lambda \in \mathbb{R}$ and randomization constant $\gamma \in [0, 1]$.

In other words, let $\alpha = \text{PFA}(r^*)$ and $1-\beta = \text{PCD}(r^*)$. If r' has rejection region R' and

$$\begin{aligned} \alpha' &= \text{PFA}(r') = \mathbb{P}_{H_0}(Y \in R') \leq \alpha \\ 1-\beta' &= \text{PCD}(r') = \mathbb{P}_{H_1}(Y \in R'), \end{aligned}$$

then $1-\beta' \leq 1-\beta$. Furthermore, r^* is the *unique* optimal test with PFA α, so that $\alpha' < \alpha$ implies $1-\beta' < 1-\beta$, or r' is strictly less powerful.

Proof. Let R be the rejection region of r^*. We wish to show that

$$\mathbb{P}_{H_1}(Y \in R) \geq \mathbb{P}_{H_1}(Y \in R'),$$

or equivalently, after subtracting the probability of the common region $Y \in R \cap R'$,

$$\int_{R \setminus R'} L(y) \cdot \mathbb{P}_{H_0}(y)\, dy \geq \int_{R' \setminus R} L(y) \cdot \mathbb{P}_{H_0}(y)\, dy.$$

We know that $L(y) \geq \lambda$ on R and $L(y) < \lambda$ on R^c by the definition of R. Moreover,

$$\int_R \mathbb{P}_{H_0}(y)\, dy = \alpha \geq \alpha' = \int_{R'} \mathbb{P}_{H_0}(y)\, dy.$$

Then we are done after subtracting $\mathbb{P}_{H_0}(R \cap R')$ from both α and α'.

$$\mathbb{P}_{H_1}(R \setminus R') \geq \lambda \cdot \mathbb{P}_{H_0}(R \setminus R') \geq \lambda \cdot \mathbb{P}_{H_0}(R' \setminus R) \geq \mathbb{P}_{H_1}(R' \setminus R).$$

Now suppose that r' is another optimal test with $\alpha' = \alpha$, so then $\beta' = \beta$, and

$$\int_{\mathbb{R}} (r^*(y) - r'(y)) \cdot (\mathbb{P}_{H_1}(y) - \lambda \cdot \mathbb{P}_{H_0}(y))\, dy = (\beta - \lambda \cdot \alpha) - (\beta' - \lambda \cdot \alpha') = 0.$$

By the definition of r^*, the integrand is nonnegative, so $r^*(y) - r'(y) \neq 0$ is only possible on the event $\{L(Y) = \lambda\} = \{y : \mathbb{P}_{H_1}(y) - \lambda \cdot \mathbb{P}_{H_0}(y) = 0\}$. If it has zero probability, then $r^* = r'$ a.s.; otherwise, the randomization constants must also agree, so $r^* = r'$ a.s., proving uniqueness.

Lastly, a technical note: Bernoulli(γ) is a random variable defined only on the event $\{L(Y) = \lambda\}$ and independent of X, so that the chain rule behaves as expected.

$$\begin{aligned} \mathbb{P}_{H_0}(L(Y) = \lambda, \text{Bernoulli}(\gamma) = 1) &= \mathbb{P}_{H_0}(L(Y) = \lambda) \cdot \mathbb{P}_{H_0}(\text{Bernoulli}(\gamma) = 1 \mid L(Y) = \lambda) \\ &= \mathbb{P}_{H_0}(L(Y) = \lambda) \cdot \gamma. \end{aligned}$$

So with randomization, the rejection region R is not necessarily uniquely determined. $\square$

The constraint $\mathbb{P}_{H_0}(L(Y) > \lambda) + \gamma \cdot \mathbb{P}_{H_0}(L(Y) = \lambda) \leq \alpha^*$ now appears to depend on two unknowns, λ and γ, but the derivation above also gives us a useful procedure to determine both parameters from one inequality by first setting $\gamma = 0$.

1. Find the likelihood ratio L.
2. Find the threshold λ without randomization.
3. Find the randomization constant γ if the PFA is still less than α^*.

4 Examples

$L(y)$ is often difficult to analyze, so we want to find a simpler equivalent condition to $L(y) > \lambda$. We can do so when $L(y)$ is monotonic: $\{L(Y) > \lambda\}$ is equivalent to $\{Y > t\}$, or $\{Y < t\}$, whose probability is known from the distribution of Y given in the hypotheses.

Example 2 (Normal hypotheses).

Let $Y \sim \mathcal{N}(X, \sigma^2)$, and let $\alpha^* \in [0, 1]$. Then, let us first find the likelihood ratio:

$$L(y) = \frac{f_{Y|X}(y \mid 1)}{f_{Y|X}(y \mid 0)} = \exp\left(-\frac{(x-1)^2}{2\sigma^2} + \frac{x^2}{2\sigma^2}\right) = \exp\left(\frac{2x-1}{2\sigma^2}\right).$$

We now observe that $L(y)$ is monotonically increasing: as H_1 is "to the right of" H_0, a larger observed value gives a higher likelihood of $Y \sim H_1$. If the two hypotheses were swapped, $L(y)$ would instead be monotonically decreasing. In any case, the likelihood ratio test becomes

$$r^*(Y) = \mathbb{1}\{Y > t\}$$

for some threshold $t \in \mathbb{R}$. There is no randomization in the continuous case, as $\mathbb{P}(Y = t) = 0$ for every $t \in \mathbb{R}$. Then the optimization problem becomes

$$\operatorname*{argmax}_{t \in \mathbb{R}} 1 - \Phi\left(\frac{t-1}{\sigma}\right) \quad \text{s.t.} \quad 1 - \Phi\left(\frac{t-0}{\sigma}\right) = \alpha^*$$

after expanding $\alpha = \mathbb{P}_{H_0}(Y > t)$ and $1 - \beta = \mathbb{P}_{H_1}(Y > t)$. We are done after finding

$$t = \sigma \cdot \Phi^{-1}(1 - \alpha^*).$$

Randomization is quite likely to appear in the case of discrete hypotheses.

Example 3 (Categorical hypotheses).

Let H_0 and H_1 be the categorical distributions $(0.2, 0.3, 0.5)$ and $(0.8, 0.1, 0.1)$ respectively, for the probabilities of drawing a red, green, or blue marble out of a jar, and let $\alpha^* = 0.25$. We observe that the likelihood ratio is a discrete function of Y:

$$L(y) = \begin{cases} 4 & \text{if } y = \text{red} \\ \frac{1}{3} & \text{if } y = \text{green} \\ \frac{1}{5} & \text{if } y = \text{blue.} \end{cases}$$

We also observe that for discrete likelihood ratios, the choice of the threshold λ is without loss of generality from the values taken on by $L(y)$. For instance, $\lambda = 2$ defines almost the same rule as $\lambda = \frac{1}{3}$, except without randomization, the finer control we need.

Then, setting $\lambda = \frac{1}{3}$ and $\lambda = \frac{1}{5}$ result in the respective PFAs of

$$\mathbb{P}_{H_0}(Y = \text{red}) = 0.2$$
$$\mathbb{P}_{H_0}(Y \in \{\text{red}, \text{green}\}) = 0.5.$$

As $0.2 < \alpha^* < 0.5$, we take $\lambda = \frac{1}{3}$ and introduce randomization. We need

$$\mathbb{P}_{H_0}(Y = \text{red}) + \gamma \cdot \mathbb{P}_{H_0}(Y = \text{green}) = \alpha^*,$$

from which we get $\gamma = \frac{1}{6}$. To see the effect of randomization, we can find the PCD with and without randomization:

$$\mathbb{P}_{H_1}(Y = \text{red}) + \gamma \cdot \mathbb{P}_{H_1}(Y = \text{green}) = 0.8 + \frac{1}{6} \cdot 0.1$$
$$\mathbb{P}_{H_1}(Y = \text{red}) = 0.8.$$

The final example of tuning γ is also an example of a test optimal in the Neyman–Pearson sense, but nonoptimal in that it needlessly increases the PFA without increasing the PCD, in the case of $\alpha^* > \frac{1}{2}$. This demonstrates a problem with setting $\alpha = \alpha^*$ in general: if the hypotheses have no overlap, then a clean boundary with PFA $= 0$ and PCD $= 1$ is the clear better choice over an overreactive Neyman–Pearson test that sets PFA $= \alpha^*$ and PCD $= 1$.

Example 4 (Tuning the randomization constant).

Let $H_0 \colon Y \sim \text{Uniform}([-1, 1])$ and $H_1 \colon Y \sim \text{Uniform}([0, 2])$, and let α^* be given.

$$L(y) = \frac{\mathbb{1}\{0 \leq y \leq 2\}}{\mathbb{1}\{-1 \leq y \leq 1\}}$$

takes values in $\{0, 1, \infty\}$, in particular on $[-1, 0)$, $[0, 1]$, and $(1, 2]$ respectively.

a. In the case of $\lambda = 0$, we find $\frac{1}{2} + \frac{1}{2}\gamma = \alpha^*$.

b. In the case of $\lambda = 1$, we find $\frac{1}{2}\gamma = \alpha^*$.

c. In the case of $\lambda = \infty$, we find the (mostly unsatisfiable) equation $0 = \alpha^*$.

So, the Neyman–Pearson test sets $\lambda = 1$ and $\gamma = 2\alpha^*$ for $0 \leq \alpha^* \leq \frac{1}{2}$, and sets $\lambda = 0$ and $\gamma = 2(\alpha^* - \frac{1}{2})$ for $\frac{1}{2} \leq \alpha^* \leq 1$. By the remark above, the threshold $\lambda = 0$ creates extraneous false alarms, which means that significance levels $\alpha^* > \frac{1}{2}$ do not make sense here.

We can describe this Neyman–Pearson test equivalently and more explicitly by its acceptance or rejection of the null for each of $Y \in [-1, 0)$, $Y \in [0, 1]$, and $Y \in (1, 2]$.

We have far from given a complete introduction to hypothesis testing, which is studied further in statistics. We encourage the reader to look further into one-tailed tests, two-tailed tests, and Bayesian testing if interested. For now, we leave with the following challenge: generalize the Neyman–Pearson test to n i.i.d. observations $Y_1, \ldots, Y_n$.