Hypothesis Testing

The simplest setting is as follows.
We have a sequence of observations $X_1, X_2, \ldots, X_n$ which are iid & are generated from one of two known distributions :

F_0, under the null hypothesis H_0

F_1, under the alternative hypothesis H_1

We have to decide which of these two hypotheses is true.

Examples :

1. A transmitter in a communication system transmits one of two possible symbols. The received signal is corrupted by additive Gaussian noise, of zero mean & known variance, & the receiver has to decode the signal, i.e., decide which of the two symbols was transmitted.

2. A patient is admitted to hospital with flu, and the doctor has to decide which of two different strains is responsible for the infection. The data is a gene sequence of the flu virus in the patient, which may exhibit some mutations from the original sequence of the strain. The original sequences & the mutation probabilities are known.

3. A website has to determine whether a login attempt has been generated by a human or a bot. It sets a series of tasks, which a human ~~has~~ performs correctly with probability p each, and a bot with probability q each; p and q are known.

Formulating the Problem

H_0 : hypothesis that X_i are iid $\sim F_0$

H_1 : —"— F_1

Given data $X_1, X_2, \ldots, X_n$, decide which of the hypotheses is true.

H_0 is called the null hypothesis & typically identified with normal behaviour / status quo, while H_1 is called the alternative hypothesis & identified with a change of interest.

There are two kinds of error that can be made. We call the corresponding error probabilities :

False alarm probability

$$\alpha = \mathbb{P}(\text{decide } H_1 \mid H_0 \text{ is true})$$

Detection failure

$$\beta = \mathbb{P}(\text{decide } H_0 \mid H_1 \text{ is true})$$

Clearly, any hypothesis test has to trade off between these two error probabilities.

~~Optimal Tests~~

The false alarm probability, α, can be made 0 by always choosing H_0. Conversely, the detection failure probability can be minimised by always deciding H_1.
These trivial tests show that there is an inevitable tension, or tradeoff, between minimising the two error probabilities.

Hence, a natural problem formulation is to impose a threshold, a maximum acceptable value, on one type of error, and seek to minimise the other type.
An alternative, though essentially equivalent, formulation, is to seek to minimise some linear combination of the error probabilities

Some terminology

α is called the significance level of the test, & ~~β~~ $1-\beta$ is called its power.
(More precisely, the significance level is the maximum acceptable value of α that a test is designed to guarantee.)

Optimal Tests

How do we determine, for a given α^*, a test that minimises β while ensuring that $\alpha \leq \alpha^*$?

Let f_0 & f_1 denote the densities of the distributions F_0 & F_1.

A key role is played by the likelihood ratio

$$\frac{L_1(x_1, \ldots, x_n)}{L_0(x_1, \ldots, x_n)} = \prod_{k=1}^{n} \frac{f_1(x_i)}{f_0(x_i)}$$

Now, what is a statistical test? It is a decision rule which specifies, for any possible set of observations, which of the hypotheses will be chosen. In other words, it is a function

$$T : \mathcal{X}^n \to \{0, 1\}$$

Here, $\mathcal{X}$ is the set in which the observations lie. Typically, $\mathcal{X} = \mathbb{R}$ (the real numbers) or $\mathbb{Z}$ (the integers) or vectors of some fixed length d with real or integer elements.

Q: What is the best choice of function, T?

Optimal Tests (cont.)

Given a test, i.e. a function, T, it is easy to define the two error probabilities.

Define $T^{-1}(0) = \{(x_1, x_2, \cdots, x_n) : T(x_1, \cdots, x_n) = 0\}$ to be the set of observations which result in H_0 being chosen. Define $T^{-1}(1)$ similarly. Then,

$$\alpha = \int_{T^{-1}(1)} f_0(x_1) f_0(x_2) \cdots f_0(x_n)\, dx_1 dx_2 \cdots dx_n$$

$$1-\beta = \int_{T^{-1}(1)} f_1(x_1) f_1(x_2) \cdots f_1(x_n)\, dx_1 dx_2 \cdots dx_n$$

$$= \int_{T^{-1}(1)} \frac{L_1(x_1, \cdots, x_n)}{L_0(x_1, \cdots x_n)} f_0(x_1) \cdots f_0(x_n)\, dx_1 \cdots dx_n$$

Hopefully, the following is now intuitive.

Neyman-Pearson Lemma (Optimality of Likelihood Ratio Tests)

Among all tests that guarantee $\alpha \leq \alpha^*$, for ~~any~~ fixed α^*, there is a test of the form

$$\text{Reject } H_0 \text{ if } \frac{L_1(\text{data})}{L_0(\text{data})} > \tau$$

that minimises β.

τ is a threshold that depends on α^*.

Optimal Tests (cont.)

Technical Remark : We glossed over a technical detail in stating the Neyman-Pearson Lemma. For it to hold for all α^*, we need to allow randomised tests / decision rules. These are tests in which, after observing the data, we choose one of the two competing hypotheses with probabilities that depend on the data. If these probabilities can only take values 0 or 1, then it is a hard / deterministic decision rule.

The reason we need these is that, especially for a discrete observation space $\mathcal{X}$, not all values of α may be achievable. Maybe any (hard) decision rule can only achieve false alarm probabilities $\alpha = 0.1, 0.2, 0.3$ or 0.4.

If $\alpha^* = 0.25$, so we are allowed false alarm probabilities bigger than 0.2, then no single deterministic test might be optimal. But one that randomises suitably between those with $\alpha = 0.2$ & $\alpha = 0.3$ might be!

This is a fairly minor technical detail, but the theory is much cleaner if we allow randomised tests

More Remarks

~~Given a test $T: \mathcal{X}^n \to \{0,1\}$, not necessarily~~

1. We assumed that the observations $X_1, X_2, \ldots$ are iid, both because this is the most common scenario in applications, and because it makes it easy to calculate likelihoods. However, this assumption is not required for the Neyman-Pearson Lemma, which is valid irrespective of dependencies in the data.

2. Given a test $T: \mathcal{X}^n \to \{0,1\}$, we can, in principle, calculate α & β by integration. In practice, this can be quite complicated to do exactly. But we can get good approximations for large n. For suitable choices of the threshold τ, both α & β decay exponentially to zero. See Cover & Thomas, Information Theory, for more on this.

3. Hypothesis testing is used in many different fields, which often have their own terminology. $1-\alpha$ is sometimes called the specificity of a test, and $1-\beta$ its sensitivity or recall.

Examples

b. It has been suggested that, outside congested periods, vehicle flow rates on a motorway could be used as an automatic indicator of an accident having occurred.
Under normal conditions, it has been calibrated that vehicles pass a detector after iid exponentially distributed times with a mean of 4s. After an accident, this increases to a mean of 20s.
We want to be able to detect accidents within 5 minutes of their occurrence.
Design a test with a false alarm probability no bigger than 10^{-3}, which is based on measuring 15 inter-vehicle times $X_1, X_2, \ldots, X_{15}$ at the detector.
What is the probability that the test fails to detect an accident when one has occurred?

Example (cont.)

Define $\lambda_0 = \frac{1}{4}$, $\lambda_1 = \frac{1}{20}$

H_0 : $X_1, \ldots, X_{15}$ iid $\sim$ Exp(λ_0)

H_1 : —— ,, —— Exp(λ_1)

$$\therefore \quad \frac{L_1(x_1, \ldots, x_{15})}{L_0(x_1, \ldots, x_{15})} = \frac{\lambda_1^{15} e^{-\lambda_1 x_1} \cdots e^{-\lambda_1 x_{15}}}{\lambda_0^{15} e^{-\lambda_0 x_1} \cdots e^{-\lambda_0 x_{15}}}$$

$$= \left(\frac{\lambda_1}{\lambda_0}\right)^{15} \exp\left(-\sum_{k=1}^{15} (\lambda_1 - \lambda_0) x_k\right)$$

The Neyman-Pearson lemma tells us that there is an optimal test of the form:

Reject H_0 if $L_1/L_0 > \tau$, i.e., if

$$\left(\frac{\lambda_1}{\lambda_0}\right)^{15} \exp\left(-\sum_{k=1}^{15} (\lambda_1 - \lambda_0) x_k\right) > \tau$$

$$\Leftrightarrow \quad 15 \ln\left(\frac{\lambda_1}{\lambda_0}\right) - (\lambda_1 - \lambda_0) \sum_{k=1}^{15} x_k > \ln \tau$$

$$\Leftrightarrow \quad (\lambda_0 - \lambda_1) \sum_{k=1}^{15} x_k > \ln \tau + 15 \ln\left(\frac{\lambda_0}{\lambda_1}\right)$$

$$\Leftrightarrow \quad \frac{1}{15} \sum_{k=1}^{15} x_k > \frac{1}{(\lambda_0 - \lambda_1)} \left(\ln\left(\frac{\lambda_0}{\lambda_1}\right) + \frac{1}{15} \ln \tau\right)$$

This is the general form of the test. We now need to choose τ to meet the specified constraint on the false alarm probability.

Example (cont.)

For a fixed value of τ, how do we compute the false alarm probability?

Recall that this is $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$.

Now, if H_0 is true, then we know that $X_1, X_2, \ldots, X_{15}$ are iid $\sim \text{Exp}(\frac{1}{4})$, and we want to compute

$$\alpha = \mathbb{P}\left(\frac{1}{15} \sum_{k=1}^{15} X_i > \frac{1}{\lambda_0 - \lambda_1} \left(\ln\left(\frac{\lambda_0}{\lambda_1}\right) + \frac{1}{15} \ln(\tau) \right) \right)$$

Define $\xi = \frac{1}{15} \sum_{k=1}^{15} X_i$.

In this example, we know the exact distribution of the test statistic ξ. It is a Gamma$(15, 15\lambda_0)$ distribution.

(15: shape parameter; $15\lambda_0$: scale parameter)

We say Y has a Gamma(n, λ) distribution if it has density

$$f_Y(y) = \begin{cases} \lambda(\lambda y)^{n-1} e^{-\lambda y}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

We can use the known distribution of ξ to explicitly calculate α as a function of τ, & then choose τ so that $\alpha = 10^{-3}$.

Remarks

1. The likelihood ratio $\frac{L_1(x_1, \dots, x_n)}{L_0(x_1, \dots, x_n)}$ depended on the data only via its sample mean

$$\xi = \frac{1}{n} \sum_{k=1}^{n} x_i .$$

This is because the sample mean is a sufficient statistic for the exponential distribution.

2. In this example, a sufficient statistic was 1-dimensional (took values in $\mathbb{R}$). If our data were normally distributed with unknown mean & variance, then the sample mean & sample variance would together constitute a sufficient statistic, which would thus be 2-dimensional.
This is still a big reduction in dimension compared to the raw data

3. In order to compute probabilities of the two types of errors, we need to know the distribution of the sufficient statistic under each of the distributions, F_0 & F_1.
Often, this can only be computed numerically or by simulation

More examples

1. The number of requests received by a website over a 1-minute period has a Gaussian distribution with mean 100 & variance 100 under normal conditions. However, when it is subject to a botnet attack, the mean increases to 500, while the variance stays the same. The distribution remains Gaussian.

   Design a test which will detect such an attack within 20 minutes of its start, and whose false alarm probability (over a 20-minute period) is no bigger than 10^{-4}

2. Repeat when the variance also changes to 500

Solutions

Note that the exact values of the means & variances are not important, nor whether one or both of them change. What is important is that they are known, under both H_0 & H_1.

We will give a general solution that covers both cases

Working out the solution

H_0 : $X_1, X_2, \dots, X_{20}$ iid $\sim N(\mu_0, \sigma_0^2)$

H_1 : —"— $N(\mu_1, \sigma_1^2)$

Compute the likelihood ratio :

$$\frac{L_1(X_1, \dots, X_{20})}{L_0(X_1, \dots X_{20})} = \left(\frac{\sigma_0}{\sigma_1}\right)^{20} \prod_{k=1}^{20} \frac{e^{-\frac{(X_k-\mu_0)^2}{2\sigma_0^2}}}{e^{-(X_k-\mu_1)^2/2\sigma_1^2}}$$

$$\therefore \log \frac{L_1}{L_0} = 20 \log \frac{\sigma_0}{\sigma_1} + \sum_{k=1}^{20} \left[\frac{(X_k-\mu_1)^2}{2\sigma_1^2} - \frac{(X_k-\mu_0)^2}{2\sigma_0^2} \right]$$

By the Neyman-Pearson lemma, the optimal test is of the form :

Reject H_0 if $\log \frac{L_1}{L_0} > \tau$

The problem is to determine τ so that

$$\mathbb{P}\left(\sum_{k=1}^{20} \left[\frac{(X_k-\mu_1)^2}{2\sigma_1^2} - \frac{(X_k-\mu_0)^2}{2\sigma_0^2} \right] > \tau + 20 \log \frac{\sigma_1}{\sigma_0} \,\middle|\, H_0 \right)$$

$$\leq 10^{-4}.$$

If $\sigma_0 = \sigma_1$, the expression is much easier to simplify, but conceptually there is no difference.