A general view of hypothesis testing

Christos Dimitrakakis

July 23, 2009

Abstract

A general overview of hypothesis testing is given. The Bayesian and distribution-free framework to multiple hypothesis testing and to null hypothesis testing are discussed. Some practical algorithms are introduced, together with associated performance bounds.

1 Introduction

Hypothesis testing is a decision making problem, where we are asked to decide between hypotheses. As is usual with decision making problems, each choice reasults in a loss. The aim of the decision maker is to minimise this loss.

More specifically, imagine that we have a choice from a set of decisions $H = \{ h_i : i = 1, \dots, m \}$. In addition, consider that the state of the world is $\theta \in \Theta$, where Θ is the set of all possible world states. We are given a loss function $\ell : H \times \Theta \to \mathbb{R}$. We wish to choose $h \in H$ such that the expected loss is minimised

$$\mathbf{E}(\ell|h) \triangleq \int_\Theta \ell(h, \theta) p(\theta) \, \mathrm{d}\theta, \tag{1.1}$$

where $p(\theta)$ is a suitable density over Θ, which we for the moment shall leave unspecified.

We distinguish two types of decision making problems. We shall call the first, where H is given, multiple hypothesis testing. In the second type of problem we are only given a single hypothesis, h_0, referred to hereafter as the *null* hypothesis. In this setting, we must construct an alternative hypothesis to compare against. We shall call this type of problem null hypothesis testing.

2 Multiple hypothesis testing

Let some measurable space $(\mathcal{X}, \mathfrak{B})$ and $\Gamma = \{ p(\cdot|\theta) : \theta \in \Theta \}$ be a set of densities on $(\mathcal{X}, \mathfrak{B})$, indexed by θ and equipped with distance $D(\cdot \| \cdot)$ on Θ. In the sequel we use $\mathbf{P}_\theta$ and $\mathbf{E}_\theta$ to denote probabilities and expectations for the density $p(\cdot|\theta)$.

We are given a sequence $x^n = x_1, \dots, x_n$, with $x_i \in \mathcal{X}$ and a collection of sets $\{ \Theta_i : i = 1, \dots, m \}$, with $\Theta_i \subset \Theta$ for all $i \neq j$. Given a prior probability $\xi(\Theta_i)$ describing how certain we are, before seeing any data, that the particular

set Θ_i contains the true θ from which the data is generated, we wish to estimate the posterior probability $\xi(\Theta_i|x^n)$. This is our belief that the data has been generated by some $\theta \in \Theta_i$ after we have seen all the data. It is relatively easy to calculate the resulting belief via conditional probabilities:

$$\xi(\Theta_i|x^n) = \frac{p_w(x^n|\Theta_i)\xi(\Theta_i)}{\sum_j p_w(x^n|\Theta_j)\xi(\Theta_j)}, \tag{2.1}$$

where $p_w(x^n|\Theta_i)$ is the data likelihood under $\theta \in \Theta_i$ and w is a prior density over Θ_i. Usually there is more than one element θ in Θ_i for which $w(\theta) > 0$. Then the likelihood implies a marginalization:

$$p_w(x^n|\Theta_i) = \int_{\Theta_i} p(x^n|\theta)w(\theta|\Theta_i)\,\mathrm{d}\theta. \tag{2.2}$$

After we have calculated the posterior over Θ it is a simple matter to calculate the expected loss under our belief ξ:

$$\mathbf{E}_\xi(\ell|h) \triangleq \int_\Theta \ell(h,\theta)\xi(\theta)\,\mathrm{d}\theta, \qquad \xi(\theta) = \sum_i w(\theta|\Theta_i)\xi(\Theta_i). \tag{2.3}$$

Then we merely must take h minimising the expected loss.

2.1 Selecting between mutually exclusive hypotheses

A very common case in hypothesis testing is each hypothesis h_i corresponds to guessing that $\theta \in \Theta_i$, while the subsets are mutually exclusive: $\Theta_i \cap \Theta_j = \emptyset$ for all $i \neq j$ and the function is the zero-one loss for guessing the correct subset Θ_i:

$$\ell(h_i,\theta) = \begin{cases} 0, & \text{if } \theta \in \Theta_i \\ 1, & \text{if } \theta \notin \Theta_i. \end{cases}$$

A second case is when the subsets form a sequence $\emptyset \subset \Theta_1 \subset \cdots \subset \Theta_m \subset \Theta$. This is a classic problem in the model selection literature and thus we shall not deal with it here.

3 Null hypothesis testing

Let some measurable space $(\mathcal{X}, \mathfrak{B})$ and $\Gamma = \{\, p(\cdot|\theta) \,:\, \theta \in \Theta \,\}$ be a set of densities on $(\mathcal{X}, \mathfrak{B})$, indexed by θ and equipped with distance $D(\cdot\|\cdot)$ on Θ. In the sequel we use $\mathbf{P}_\theta$ and $\mathbf{E}_\theta$ to denote probabilities and expectations for the density $p(\cdot|\theta)$.

We are given a sequence $x^n = x_1, \ldots, x_n$, with $x_i \in \mathcal{X}$ and a subset $\Theta_0 \subset \Theta$. We wish to test the hypothesis that $x \sim \theta$, from some $\theta \in \Theta_0$, which we call the *null* hypothesis. In addition, we define the ϵ-null subset of Θ:

$$\Theta_\epsilon \triangleq \{\, \theta' \in \Theta \,:\, \rho(\theta', \Theta_0) \leq \epsilon \,\}, \qquad \rho(\theta', \Theta_0) \triangleq \inf_{\theta \in \Theta_0} D(\theta\|\theta'), \tag{3.1}$$

which is the ϵ-extended set of the null set Θ_0. It holds that

$$\emptyset \subset \Theta_0 \subset \Theta_\epsilon \subset \Theta. \tag{3.2}$$

This is illustrated in Figure 1.

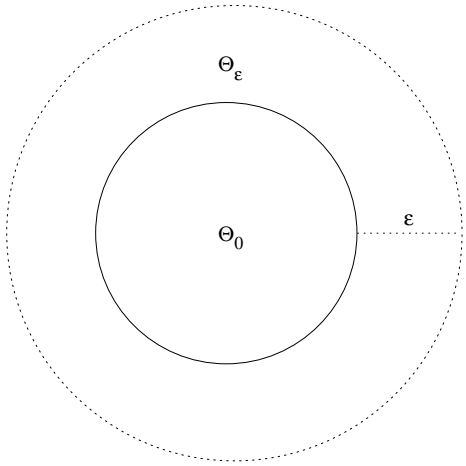

Figure 1: The null hypothesis is that the distribution θ lies in $\Theta_0 \subset \Theta$. The alternative hypothesis is that the distribution θ lies outside the ϵ-null set Θ_ϵ

Our aim is to use the observations x^n to guess a set $\Theta' \subset \Theta$ that is ϵ-close to θ, i.e. such that $\rho(\theta, \Theta') < \epsilon$, with high probability. We limit the decision problem to two choices: between h_0, where we decide that $\theta \in \Theta_0$ and h_1, where we decide that $\theta \notin \Theta_\epsilon$. We wish to bound the probability of deciding h_0 when h_1 is true by δ_1 and conversely, the probability of deciding h_0 when h_1 is true by δ_0. When $\theta \in \Theta_\epsilon \setminus \Theta_0$, the set we choose is always ϵ-close to θ by construction. When θ is in that indeterminate region, we cannot make any probabilistic guarantees about our guess.

Finally, if we suffer a loss ℓ_i with probability at most δ_i, then we can bound our expected loss by

$$\mathbf{E}\,\ell \leq \sum_i \delta_i \ell_i. \tag{3.3}$$

In the case where the loss is 0 if we are within ϵ, and 1 if we make the wrong guess, then the total loss is simply bounded by $\delta_1 + \delta_2$.

3.1 Testing a mean

As a motivational example, consider mean estimation. Let $\mu \triangleq \mathbf{E}_{\theta^*} x$ and $\hat{x} \triangleq \frac{1}{n}\sum_{i=1}^n x_i$. We form a point null hypothesis, the set $\Theta_0 = \{\theta^*\}$. Define $D(\hat{x}\|\mu) \triangleq |\hat{x} - \mu|$. Then, from Hoeffding's inequality (A.2) we have:

$$\mathbf{P}_{\theta^*}(|\hat{x} - \mu| > t) < 2\exp(-2nt^2). \tag{3.4}$$

Let $\phi \in \Theta$ such that $|\mu - \mathbf{E}_\phi\, x| \geq \epsilon$. Then

$$\begin{aligned}\mathbf{P}_\phi\left(|\hat{x} - \mu| \leq t \;\middle|\; |\mu - \mathbf{E}_\phi\, x| \geq \epsilon\right) &= \mathbf{P}_\phi\left(\hat{x} - \mu \leq t \vee \mu - \hat{x} \leq t \;\middle|\; |\mu - \mathbf{E}_\phi\, x| \geq \epsilon\right) \\ &\leq \mathbf{P}_\phi(\hat{x} \leq \mu + t |\, \mathbf{E}_\phi\, x > \mu + \epsilon) + \mathbf{P}_\phi(\hat{x} \geq \mu - t |\, \mathbf{E}_\phi\, x < \mu - \epsilon).\end{aligned}$$

The worst case for the first term, is $\mathbf{E}_\phi\, x = \mu + \epsilon$ thus

$$\begin{aligned}\mathbf{P}_\phi(\hat{x} \leq \mu + t |\, \mathbf{E}_\phi\, x > \mu + \epsilon) &\leq \mathbf{P}_\phi(\hat{x} \leq \mu + t |\, \mathbf{E}_\phi\, x = \mu + \epsilon) && (3.5)\\ &\leq \exp\left(-2n(t-\epsilon)^2\right), && (3.6)\end{aligned}$$

via Hoeffding. It is easy to see that second term also gives rise to the same inequality, which allows us to write:

$$\mathbf{P}_\phi\left(|\hat{x} - \mu| \leq t \;\middle|\; |\mu - \mathbf{E}_\phi\, x| \geq \epsilon\right) \leq 2\exp\left(-2n(\epsilon - t)^2\right). \tag{3.7}$$

If we now take the enlargement ϵ to be $2t$, the above becomes bounded by $2\exp\left(-2nt^2\right)$ as well. Thus, both (3.7) and (3.4) are bounded by the same quantity.

Thus, let us choose some probability δ to bound (3.4). This leads in choosing the confidence bound

$$t = \sqrt{\frac{\log(2/\delta)}{2n}}.$$

We employ this confidence bound in Algorithm 1, which accepts the null hypothesis if the difference between the estimated mean and the mean of θ^* is within the confidence bound.

Algorithm 1 Hypothesis mean accept

```
procedure MEAN TEST(μ, x^n, δ)
    x̂ = (1/n) Σ_{i=1}^n x_i
    ε = sqrt((2 log 2/δ) / n)
    if |x̂ − μ| ≤ ε/2 then
        Decide θ ∈ Θ_0.
    else
        Decide θ ∉ Θ_ε.
    end if
end procedure
```

Lemma 3.1. *Algorithm 1 decides for a set Θ', with the property that $\rho(\theta, \Theta') < \sqrt{\frac{\log 2/\delta}{2n}}$, with probability at least $1 - \delta$.*

The proof of this lemma follows immediately from the above discussion.

3.2 Testing a distribution with distribution-free methods

Let us now turn to the problem of estimating a distribution from a sample $x^n \in \mathcal{X}^n$. Using this sample we form the empirical distribution by creating a partition B_n of $\mathcal{X}$, of size k, such that, each set X in the partition B_n contains at least $\frac{n}{k}$ observations. More precisely,

$$B_n \triangleq \left\{ X_j \subset \mathcal{X},\ j = 1, \dots, k : \bigcup_{i=1}^{k} X_i = \mathcal{X} \wedge X_i \cap X_j = \emptyset\ \forall i \neq j \right\}, \quad (3.8)$$

and if we denote the number of samples contained in each subset X_j by $s_j \triangleq \sum_{i=1}^{n} \mathbb{I}\{x_i \in X_j\}$, then

$$s_j \geq \left\lfloor \frac{n}{k} \right\rfloor, \qquad \forall X_j \in B_n. \quad (3.9)$$

This is the only property we require of B_n, so the exact details of selecting the partition are not relevant. Then, the empirical measure $\mathbf{P}_\psi$ arising from the partition B_n is

$$\mathbf{P}_\psi(X_j) \triangleq \frac{s_j}{n}. \quad (3.10)$$

Furthermore, the true measure $\mathbf{P}_\theta$ for this partition is

$$\mathbf{P}_\theta(X_j) \triangleq \mathbf{P}_\theta(x \in X_j). \quad (3.11)$$

Then we define the distance

$$D_k(\psi \| \theta) \triangleq \sum_{j=1}^{k} |\, \mathbf{P}_\psi(X_j) - \mathbf{P}_\theta(X_j)|. \quad (3.12)$$

Using this distance, we can employ Algorithm 2 to decide whether the observed data are from a distribution in Θ_0, or from some distribution that is not more than ϵ-close to Θ_0.

Lemma 3.2. *Algorithm 2 decides for a set Θ', with the property that $\rho(\theta, \Theta') < \sqrt{\frac{8}{n}[k \log 2 + \log 1/\delta]}$, with probability at least $1 - \delta$.*

Proof. Let $\epsilon \triangleq \sqrt{\frac{8}{n}[k \log 2 + \log 1/\delta]}$. There are three, mutually exclusive cases. That $\theta \in \Theta_0$, that $\theta \in \Theta_\epsilon \setminus \Theta_0$ and that $\Theta \in \Theta \setminus \Theta_\epsilon$. Let h_0 be the decision that $\theta \in \Theta_0$ and h_1 that $\theta \notin \Theta_\epsilon$.

If $\theta \in \Theta_\epsilon \setminus \Theta_0$, then we are always ϵ-close to θ no matter what we decide. Thus, in that case the lemma is trivially satisfied.

If $\theta \in \Theta_0$, then we are ϵ-close to θ only if we decide h_0. Thus, we must bound the probability that we decide h_1. We only decide h_1 when $\rho(\psi, \Theta_0) > \epsilon/2$, but since $\rho(\psi, \Theta_0) \leq D_k(\psi \| \theta)$ for any $\theta \in \Theta_0$ by definition:

$$\mathbf{P}_\theta(h_1) \leq \mathbf{P}_\theta(D_k(\psi \| \theta) > \epsilon/2) = \mathbf{P}_\theta(\|\psi - \theta\|_1 > \epsilon/2) \quad (3.13)$$

$$< (2^k - 2) \exp(-\frac{n}{8}\epsilon^2) < 2^k \exp(-\frac{n}{8}\epsilon^2), \quad (3.14)$$

Algorithm 2 Histogram test

procedure HISTOGRAM TEST(Θ_0, x^n, δ)
  $k = \sqrt{n}$.
  Let $\mathbf{P}_\psi$ be the empirical measure on an $\frac{1}{k}$-net from x^n.
  $\epsilon = \sqrt{\frac{8}{n}[k \log 2 + \log 1/\delta]}$.
  if $\rho(\psi, \Theta_0) \leq \epsilon/2$ **then**
    Decide $\theta \in \Theta_0$.
  else
    Decide $\theta \notin \Theta_\epsilon$.
  end if
end procedure

from Weissman's inequality (A.5). Substituting ϵ, we obtain

$$\mathbf{P}_\theta(h_1) < 2^k \exp(-k \log 2 - \log 1/\delta) = \delta. \tag{3.15}$$

Conversely, if $\theta \notin \Theta_\epsilon$, then we are only ϵ-close to θ if we decide h_1. Thus, we must bound the probability that we decide h_0. This occurs if $\rho(\psi, \Theta_0) \leq \epsilon/2$, thus, for any $\theta \in \Theta_0$:

$$\mathbf{P}_\theta(h_1) \leq \mathbf{P}_\theta(\rho(\psi, \theta) \leq \epsilon/2) = \mathbf{P}_\theta(\|\psi - \theta\|_1 \leq \epsilon/2) \tag{3.16}$$

$$\leq (2^k - 2) \exp(-\frac{n}{8}\epsilon^2) < 2^k \exp(-\frac{n}{8}\epsilon^2), \tag{3.17}$$

from Weissman's inequality as before. Substituting ϵ, we obtain $\mathbf{P}_\theta(h_1) < \delta$ once more. □

The empirical measure estimated by Algortihm 2 is only ϵ-close to the true distribution θ with respect to the defined $n^{-1/2}$-net. However, it is possible to relate this distance $D_k(\psi\|\theta)$ to the L_1 distance over $\mathcal{X}$

$$D(\psi\|\theta) \triangleq \int_{\mathcal{X}} |p(x|\psi) - p(x|\theta)| \, \mathrm{d}\mu. \tag{3.18}$$

This requires that we are more careful with the choice of the partition B_n.

Lemma 3.3. *If $\exists L, M > 0$ such that $p(x|\theta)$ satisfies*

$$|p(x|\theta) - p(x'|\theta)| < L\|x - x'\|, \qquad \mu\left(\{\, x \,:\, p(x|\theta) < t \,\}\right), < Mt \tag{3.19}$$

for all $t > 0$, while the empirical measure ψ over the partition B_n satisfies

$$D_k(\psi\|\theta) < \epsilon, \tag{3.20}$$

then

$$D(\psi\|\theta) \triangleq \int_{\mathcal{X}} |p(x|\psi) - p(x|\theta)| \, \mathrm{d}\mu \leq L|\mathcal{X}|\epsilon. \tag{3.21}$$

Proof.

$$D(\psi\|\theta) = \int_{\mathcal{X}} |p(x|\psi) - p(x|\theta)| \,\mathrm{d}\mu(x) = \sum_{X\in B_n} \int_X |p(x|\psi) - p(x|\theta)| \,\mathrm{d}\mu(x) \tag{3.22}$$

$$\leq \sum_{X\in B_n} \int_X \epsilon \,\mathrm{d}\mu(X) \leq \epsilon \sum_{X\in B_n} L|X| = L|\mathcal{X}|\epsilon. \tag{3.23}$$

□

A Auxilliary results

A.1 Hoeffding inequality

The general Hoeffding inequality states that for any sequence of $x_i \in [b_i, b_i + h_i]$, then

$$\mathbf{P}\left(\sum_{i=1}^n |x_i - \mathbf{E}\, x_i| \geq n\epsilon\right) \leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n h_i^2}\right) \tag{A.1}$$

A specific form of the Hoeffding inequality for $x \in [0,1]$, $\hat{x} \triangleq \frac{1}{n}\sum_{i=1}^n x_i$, is often useful:

$$\mathbf{P}\left(\hat{x} - \mathbf{E}\, x \geq \epsilon\right) \leq \exp\left(-2n\epsilon^2\right), \tag{A.2}$$

where the two-sided form can be recovered via a union bound.

We can use the Hoeffding inequality to get a bound on the first order deviation between $\hat{p}$ and p, two multinomial distributions of order m.

Corollary A.1. *If $\hat{p} = \frac{1}{n}\sum_{i=1}^n p(i)$ and $p(i) \in \mathbb{R}^m$ are observations from some multinomial distribution with m outcomes, then*

$$\mathbf{P}(\|\hat{p} - p\|_1 \geq \epsilon) \leq 2m\exp\left[-2n\left(\frac{\epsilon}{m}\right)^2\right]. \tag{A.3}$$

Proof. Note that

$$\|\hat{p} - p\|_1 = \|\sum_i \hat{p}(i) - p(i)\| \leq \sum_i |\hat{p}(i) - p(i)|. \tag{A.4}$$

If A implies B then $\mathbf{P}(A) \leq \mathbf{P}(B)$. [1] This, since from (A.4), $\|\hat{p} - p\|_1 \geq \epsilon$ implies $\sum_i |\hat{p}(i) - p(i)| \geq \epsilon$,

$$\begin{aligned}\mathbf{P}\left(\|\hat{p} - p\|_1 \geq \epsilon\right) &\leq \mathbf{P}\left(\sum_{i=1}^m |\hat{p}(i) - p(i)| \geq \epsilon\right) \leq \mathbf{P}\left(\bigvee_{i=1}^m |\hat{p}(i) - p(i)| \geq \epsilon/m\right) \\ &\leq \sum_{i=1}^m \mathbf{P}\left(|\hat{p}(i) - p(i)| \geq \epsilon/m\right) \leq 2m\exp\left[-2n\left(\frac{\epsilon}{m}\right)^2\right].\end{aligned}$$

□

[1] This is easy to prove. Note that $\mathbf{P}(B) = \mathbf{P}(B|A)\,\mathbf{P}(A) + \mathbf{P}(B|\bar{A})\,\mathbf{P}(\bar{A})$. Since A implies B, $\mathbf{P}(B|A) = 1$, thus $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B|\bar{A})\,\mathbf{P}(\bar{A}) \geq \mathbf{P}(A)$ since $\mathbf{P}(B|\bar{A}), \mathbf{P}(\bar{A}) \geq 0$

A.2 Weissman inequality

For $p \in [0,1]^m$, a multinomial distribution with m outcomes, $\hat{p} \triangleq \frac{1}{n}\sum_{i=1}^{n} p_i$:

$$\mathbf{P}\left(\|\hat{p} - p\|_1 \geq \epsilon\right) \leq (2^m - 2)\exp\left(-\frac{n}{2}\epsilon^2\right). \tag{A.5}$$

This inequality was proved by Weissman et al. [1].

This inequality might look slightly worse than (A.3) due to the exponential terms. Howeverm this is not true for the case of interest. If the Weissman bound is smaller than 1, then the Hoeffding bound is larger than the Weissman bound.

References

[1] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M.J. Weinberger. Inequalities for the L_1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.