

## Hypothesis Testing using Maximum Likelihood

Lecture #13: Tuesday, 15 February 2005  
Lecturer: Prof. Charles Elkan  
Scribe: Jan Schellenberger  
Reviewer: Gary Hon

### 1 General Framework and definitions

Hypothesis testing using maximum likelihood is a general framework for creating tests to decide whether a hypothesis is supported by the data. The general setup is the same as before.  $P_\theta$  is a family of distributions with some unknown parameter  $\theta \in \Theta$ .  $x = (x_1, x_2, \dots, x_n)$  is an iid training set where  $x_i \sim P(x|\theta)$ . Let  $\Omega$  be a subset of  $\Theta$ . This set is The Null Hypothesis ( $H_0$ ). We wish to define a test to judge whether  $x$  came from  $\Omega$  or some other  $\theta \in \Theta$ .

*definition* A *test* is a function  $t : x \rightarrow \{\text{reject}, \text{accept}\}$

*definition* The *power function* defines the probability of rejecting the Null Hypothesis at a given  $\theta$ .  $\alpha : \Theta \rightarrow \mathcal{R}$

$$\alpha(\theta) = P(\text{reject}|\theta) = P(t(x) = \text{reject}|x \sim P_\theta)$$

To be a 'good' test we want  $\alpha(\theta)$  to be low for  $\theta \in \Omega$  and high for  $\theta \notin \Omega$ . We want to reject  $H_0$  if it is false, but accept it if it is true with high probability.

*definition* The *size* of a test is an upper bound on the probability of rejecting the Null Hypothesis given that it is true:  $\sup_{\theta \in \Omega} \alpha(\theta)$

*definition* The *significance level* of a test is the desired size of a test. This value (often .05) represents the maximum probability of rejecting  $H_0$  given that it is true.

Notes:

- Sometimes  $H_0$  is merely a point  $\Omega = \{\theta_0\}$ .
- In general, the size is defined as the sup rather than the max over all  $\theta$  because  $\Omega$  can be an open set.
- If  $x$  has discrete values only, it may not be possible to set the size exactly equal to the significance.

There are many tests that can be developed in this framework, however we focus on a likelihood ratio test. Define  $\hat{\theta}$  as the maximum likelihood estimator (MLE) over all  $\Theta$ ; mathematically,  $\hat{\theta} = \text{MLE of } \theta \in \Theta$ . Define  $\theta_0$  as the MLE of  $\theta$  over  $\Omega$ . Intuitively we

want to reject  $H_0$  if  $P(x|\hat{\theta})$  is much greater than  $P(x|H_0)$ . If  $H_0$  is true then we expect  $\theta_0$  to be close to  $\hat{\theta}$ .

To express this more concisely we define the likelihood ratio (LR).

$$\lambda(x) = \frac{P(x|\hat{\theta})}{P(x|\theta_0)} = \frac{\max_{\theta \in \Theta} P(x|\theta)}{\max_{\theta \in \Omega} P(x|\theta)}$$

$\lambda(x) \geq 1$  because  $P(x|\hat{\theta}) > P(x|\theta_0)$

*definition* The Likelihood Ratio Test is: reject  $H_0$  if and only if  $\lambda(x) \geq k$  for some threshold  $k$ .

Notes:

- If we have a sufficient statistic  $t(x)$  then  $\lambda(x)$  is a function of  $t(x)$ .  $\lambda(x) = f(t(x))$  for some function  $f$ . Often  $\lambda(x)$  is an increasing function of  $t(x)$  and then the condition  $\lambda(x) \geq k$  becomes  $t(x) \geq k'$ .
- $k$  can be determined by looking at the size of the test and setting it equal (or less than) the significance level.
- Casella and Berger define the opposite ratio:  $\Omega$  over  $\Theta$ , so their  $\lambda(x) \leq 1$  by definition.

**Example 1.1** (Deriving the T-Test). The set  $x$  is made from two sets of  $n$  normally distributed values having the same variance. We want to test whether these two sets have the same mean. Symbolically,  $x = (y, z)$  where  $y = (y_1, y_2, \dots, y_n)$  and  $z = (z_1, z_2, \dots, z_n)$ . The data are distributed as:

$$\begin{aligned} y_i &\sim N(\mu_1, \sigma^2) \\ z_i &\sim N(\mu_2, \sigma^2) \end{aligned}$$

The Null Hypothesis is that the means are the same,  $H_0 : \mu_1 = \mu_2$ . The parameter space is  $\Theta = (\mu_1, \mu_2, \sigma^2 | \sigma^2 > 0)$ . The reduced Null Hypothesis parameter space is  $\Omega = (\mu_1, \mu_2, \sigma^2 | \mu_1 = \mu_2, \sigma^2 > 0)$ .

Because these are normal distributions, the joint probability is:

$$P(x|\theta) = \frac{1}{(2\pi)^n \sigma^{2n}} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^n (y_i - \mu_1)^2 + \sum_{i=1}^n (z_i - \mu_2)^2 \right) \right]$$

In order to maximize this over all  $\theta \in \Theta$  set  $\hat{\mu}_1 = \bar{y}$ ,  $\hat{\mu}_2 = \bar{z}$ ,  $\hat{\sigma}^2 = \frac{1}{2n} (\sum_{i=1}^n (y_i - \bar{y})^2 + \sum_{i=1}^n (z_i - \bar{z})^2)$ . Notice here that the MLE of the variance is used rather than the unbiased estimator.

In the restricted case of the Null Hypothesis where  $\mu_1 = \mu_2$  the MLEs are

$$\begin{aligned} \hat{\mu} &= \mu_1 = \mu_2 = \frac{\bar{y} + \bar{z}}{2} \\ \sigma_{MLE}^2 &= \hat{\sigma}^2 = \frac{1}{2n} \left[ \sum (y_i - \hat{\mu})^2 + \sum (z_i - \hat{\mu})^2 \right] \end{aligned}$$

Note that  $\dot{\sigma}^2 \geq \hat{\sigma}^2$ . Now we calculate  $\lambda(x)$ .

$$\begin{aligned}\lambda(x) &= \lambda(y_1, y_2 \dots y_n, z_1, z_2, \dots, z_n) \\ &= \frac{P(x|\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)}{P(x|\hat{\mu}, \hat{\sigma}^2)} \\ P(x|\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) &= \frac{1}{(2\pi)^n \hat{\sigma}^{2n}} \exp\left[-\frac{1}{2} \cdot 2n\right] \\ P(x|\hat{\mu}, \hat{\sigma}^2) &= \frac{1}{(2\pi)^n \hat{\sigma}^{2n}} \exp[-n] \\ \lambda(x) &= \left(\frac{\dot{\sigma}^2}{\hat{\sigma}^2}\right)^n\end{aligned}$$

Note that this quantity does not depend on either  $\mu_1$  or  $\mu_2$ . If  $\mu_1 = \mu_2$  under  $H_0$  we would expect  $\dot{\sigma}^2 = \hat{\sigma}^2$  and therefore the likelihood ratio would be close to 1. The likelihood test is  $\left(\frac{\dot{\sigma}^2}{\hat{\sigma}^2}\right)^n > k$  or  $\frac{\dot{\sigma}^2}{\hat{\sigma}^2} > \sqrt[n]{k}$ . As  $n \rightarrow \infty$ , the threshold gets close to 1. It becomes increasingly unlikely that  $\dot{\sigma}^2$  and  $\hat{\sigma}^2$  are far apart, so even small deviations from  $\dot{\sigma}^2$  will result in rejection of the Null Hypothesis.

It can be shown that  $\dot{\sigma}^2 = \hat{\sigma}^2 + \frac{1}{4}(\bar{y} - \bar{z})^2$ , which allows us to write:

$$\lambda(x) = \left(1 + \frac{(\bar{y} - \bar{z})^2}{4\hat{\sigma}^2}\right)^n$$

From this,  $\lambda(x) > k$  if and only if  $\frac{|\bar{y} - \bar{z}|}{\hat{\sigma}} > k'$ . In order to figure out what  $k'$  is we compute  $\sup_{\theta \in \Omega} P\left(\frac{|\bar{y} - \bar{z}|}{\hat{\sigma}} > k'\right) = \alpha$

This test is known as the *Student t-test* and the distribution it follows is called the *t-distribution*. The actual distribution is difficult work out and is different for each  $n$ . Fortunately there is a general theorem that lets us estimate the distribution of the likelihood ratio for large  $n$ .

*Theorem - Wilk's Theorem (1938)* Suppose  $(x_1, x_2, \dots, x_n)$  are iid samples chosen from a known pdf  $f(x|\theta)$  and

$$\begin{aligned}H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0\end{aligned}$$

As  $n \rightarrow \infty$ ,  $2 \log(\lambda(x_1, x_2, \dots, x_n)) \rightarrow \chi_1^2$ , where  $\chi_1^2$  is the Chi-squared distribution of 1 degree of freedom.

This theorem states that for  $x$  distributed according to any pdf, if the Null Hypothesis is just a single value, then the likelihood ratio approaches a Chi-squared distribution. It can be used to set the threshold for a Likelihood Ratio test.