8.4 Lecture 11 Friday 02/09/01

Homework and Labs. see the logictics section

This Friday there will be a midterm exam.

9 Maximum Likelihood Estimation

Definition:
The maximum likelihood estimate (mle) of θ is that value of θ that maximises $lik(\theta)$: it is the value that makes the observed data the "most probable".

9.1 Maximum Likelihood of Multinomial Cell Probabilities

$X_1, X_2, \ldots, X_m$ are counts in cells/ boxes 1 up to m, each box has a different probability (think of the boxes being bigger or smaller) and we fix the number of balls that fall to be n:$X_1 + X_2 + \cdots + X_m = n$. The probability of each box is p_i, with also a constraint: $p_1 + p_2 + \cdots + p_m = 1$, this is a case in which the $X_i's$ are NOT independent, the joint probability of a vector $x_1, x_2, \ldots, x_m$ is called the multinomial ,and has the form:

$$formhere$$

Each box taken separately against all the other boxes is a binomial, this is an extension thereof. (look at page 72)

We study the log-likelihood of this :

$$l(p_1, p_2, ..., p_m) = logn! - \sum_{i=1}^{m} logx_i! + \sum_{i=1}^{m} x_i logp_i$$

However we can't just go ahead and maximise this we have to take the constraint into account so we have to use the Lagrange multipliers again.

We use

$$L(p_1, p_2, ..., p_m, \lambda) = l(p_1, p_2, ..., p_m) + \lambda(1 - \sum_{i}^{m} p_i)$$

By posing all the derivatives to be 0, we get the most natural estimate

$$\hat{p_i} = \frac{x_i}{n}$$

Hardy-Weinberg Remember this is a trinomial with three boxes: the probabilities are parametrized by:

$$(1-\theta)^2 \qquad 2\theta(1-\theta) \qquad \theta^2$$

$$l'(\theta) = -\frac{2X_1 + X_2}{1-\theta} + \frac{2X_3 + X_2}{\theta}$$

$$l''(\theta) = -\frac{2X_1 + X_2}{(1-\theta)^2} + \frac{2X_3 + X_2}{\theta^2}$$

Each of the counts is binomially distributed with probabilities as described above so that:

$$\begin{aligned} E(X_1) &= n(1-\theta)^2 \\ E(X_2) &= 2n\theta(1-\theta) \\ E(X_3) &= n\theta^2 \end{aligned}$$

9.1.1 Using the Bootstrap to build Confidence Intervals

We call the unknown parameter θ_o and our estimate $\hat{\theta}$. Suppose that we had an ideal (unrealistic) situation in which we knew the distribution of $\hat{\theta} - \theta_o$, we will be interested especially in its quantiles : denote the α quantile by $\underline{\delta}$ and the $1-\alpha$ quantile by $\bar{\delta}$. By definition we have: $P(\hat{\theta} - \theta_o \leq \underline{\delta}) = \frac{\alpha}{2}$ $P(\hat{\theta} - \theta_o \leq \bar{\delta}) = 1\text{-}\alpha_{\frac{-}{2}}$

A note about the Bootstrap :

1. Would n't we have got the same answer without centering with regards to $\hat{\theta}$ (as it happended in the example-unfortunate I admit)

   What we called and were the ideal quantiles from which we build the confidence interval : $[\hat{\theta}-, \hat{\theta}-]$.Estmated by : $[\hat{\theta}-^*, \hat{\theta}-^*]$ where * for instance is the $1-\alpha$th quantile of $\hat{\theta}^* - \hat{\theta}$, the $\hat{\theta}$'s do NOT cancel out. We can show why.

2. Would these intervals be the same if we took the distribution of $\hat{\theta}^*$ to mimick simply that of the $\hat{\theta}$'s ?

   NO !

3. Where is the big picture ?

   You get what you pay for in assumptions- we'll see alot more of that.

   No clear cut answer as yet in a general framework, there are neater ones, called decision theory, we'll see later.

9.2 Large Sample Theory for MLE

MLE estimates are consistent under reasonable conditions. We will give ideas about how this is proved without the technical subtleties.

0Under appropriate smoothness conditions on f (the density), the mle from an iid sample is consistent.

Proof:
The extimate maximises $l(\theta)$ so it also maximises

$$\frac{1}{n}l(\theta) = \frac{1}{n}\sum_{i=1}^{n} logf(X_i|\theta) \longrightarrow Elogf(X|\theta) = \int logf(X|\theta)f(X|\theta_o)dx$$

Here we will have to admit that the θ that maximises $l(\theta)$ is close to the one that maximises $Elogf(X|\theta)$, so if we differentiate with regards to θ we get :

$$\int logf(X|\theta)f(X|\theta_o)dx = \int \frac{(x|\theta)}{f(x|th)}f(x|\theta_o)dx$$

and we see that for the particular value $\theta = \theta_o$ this is

$$\int f(X|\theta_o)dx = (1) = 0$$

so θ_o IS a stationary point.

We have interchanged derivationa and integration, to be allowed to do this there are smoothness conditions to impose on f.

Define the quantity

$$I(\theta) = E[logf(x|\theta)]^2$$

under appropraite smoothness conditions on f, $I(\theta)$ can be written :

$$I(\theta) = -E[logf(x|\theta)]$$

Proof:
Double derivation of the quantity who is a density gives the result.(with exchanges between order of diff and integration).

0Under smoothness conditions on f, the probability distribution of

$$\sqrt{nI(\theta_o)}(\hat{\theta}_n - \theta_o) \Longrightarrow (0, 1)$$

This says that the mle is asymptotically unbiased and that it's variance is inversely proportional to n and $I(\theta_o)$.

Proof: (Just formal heuristics)

Taylor expansion of $l'(\hat{\theta})$ which is itself 0:

$$\begin{aligned} 0 &= l'(\hat{\theta}) \approx l'(\theta_o) + (\hat{\theta} - \theta_o)l''(\theta_o) \\ (\hat{\theta} - \theta_o) &\approx -\frac{l'(\theta_o)}{l''(\theta_o)}(*) \\ n^{\frac{1}{2}}(\hat{\theta} - \theta_o) &\approx - \end{aligned}$$

We gave two theorems and an important lemma defining a good way of computing what is known as Fisher's information:

$$I(\theta_o) = E(l'(\theta_o)^2) = -E(l''(\theta_o))$$

The first theorem said that mle is consistent, the second that we have convergence in distribution of

$$\sqrt{nI(\theta_o)}(\hat{\theta}_n - \theta_o) \Longrightarrow (0, 1)$$

In particular the "asymptotics variance" of the mle $\hat{\theta}$ is

These theorems are very useful and in particular they allow the construction of:

9.3 Confidence Intervals

- exact methods
- approximate methods based on the theorems above
- bootstrap methods

9.3.1 Exact Methods

Poisson Example We saw that both the mm and mle estimate for the unique parameter needed to specify a Poisson r.v. is the sample mean: $\hat{\lambda} = \bar{X}$. We could use the fact that it is known that the sum of n Poisson(λ) rvs is Poisson($n\lambda$), this depends on the unknown λ but plugging in and using tables is possible. (This is an exact approach) : We will take another approach:

9.3.2 Approximate Methods

We will suppose that the sample size is big enough for us to use the theorems and thus need $I(\lambda)$. By definition we have :

$$I(\lambda) = E[\frac{\partial}{\partial\lambda} log f(x|\lambda)]^2$$

for the Poisson, the log-liklihood was :

$$log f(x|\lambda) = x log\lambda - \lambda - log x!$$

so that :

$$I(\lambda) = E(\frac{X}{\lambda} - 1)^2$$