Maximum Likelihood Methods
(Hogg Chapter Six)

STAT 406-01: Mathematical Statistics II *

Spring Semester 2016

Contents

Thursday 18 February 2016
– Read Section 6.1 of Hogg

0 Administrata

Effects of snow day: homeworks due on Thursday through Spring Break. First prelim exam now Thursday, March 3. Shift

back to normal after Spring Break.

1 Maximum Likelihood Estimation

1.1 Maximum Likelihood Estimates

We now return to methods of classical, or frequentist statistics, which are phrased not in terms of probabilities which quantify our confidence in general propositions with unknown truth values, but exclusively in terms of the frequency of outcomes in hypothetical repetitions of experiments. One of the fundamental tools is the sampling distribution, which we'll write as a joint pdf for a random vector $\mathbf{X}$: $f_{\mathbf{X}}(\mathbf{x})$. If the model has a parameter θ, we'll write this as $f_{\mathbf{X}}(\mathbf{x};\theta)$, and when we consider this as a function of θ, we'll write it $L(\theta;\mathbf{x})$ or sometimes $L(\theta)$. And we will often find it convenient to work with the logarithm $\ell(\theta) = \ln L(\theta)$.

Recall from last semester that one possible choice of an estimate for θ given an observed data vector $\mathbf{x}$ is the maximum likelihood estimate

$$\widehat{\theta}(\mathbf{x}) = \text{argmax}_\theta\, L(\theta;\mathbf{x}) = \text{argmax}_\theta\, \ell(\theta;\mathbf{x}) \tag{1.1}$$

If we take the functional form of $\widehat{\theta}(\mathbf{x})$ and insert the random vector $\mathbf{X}$ into the function, we get a statistic $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ which is known as the *maximum likelihood estimator*. Note that this is a random variable not in the formal sense which we invoked to assign a Bayesian probability distribution to an unknown parameter Θ, but rather a function of the random vector $\mathbf{X}$ whose value will be different in different random trials, even with a fixed value for the parameter θ. (Of course, if we're using $\widehat{\theta}$ as an estimator, the true value of the parameter θ will be unknown.)

A useful special case is where the random vector $\mathbf{X}$ is a sample of size n drawn from some distribution $f(x;\theta)$. Then the likelihood function will be

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) \tag{1.2}$$

and the log-likelihood will be

$$\ell(\theta;\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i;\theta) \tag{1.3}$$

1.1.1 Motivation

Since we know that in the Bayesian framework, the posterior pdf for the parameter θ given the observed data $\mathbf{x}$ is

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_\Theta(\theta)\, f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\ , \tag{1.4}$$

we see that if the prior pdf $f_\Theta(\theta)$ is a constant, the posterior will be proportional to $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = L(\theta;\mathbf{x})$, and therefore the $\widehat{\theta}(\mathbf{x})$ which maximizes the likelihood will also maxmize the posterior psd $f_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$.

Hogg also invokes two theorems which indicate that the MLE matches up with the true unknown value of the parameter in the limit that the size n of a sample goes to infinity:

- **Theorem 6.1.1**: Suppose θ_0 is the true value of a parameter θ, and $\mathbf{X}$ is a sample of size n from a distribution $f(x;\theta)$. Given certain regularity conditions, most importantly[1] that θ_0 is not at the boundary of the parameter space for θ, the true parameter maximizes the likelihood in the limit $n \to \infty$, i.e.,

$$\lim_{n\to\infty} P\left(L(\theta_0;\mathbf{X}) > L(\theta_0;\mathbf{X})\right) \qquad \text{for } \theta \neq \theta_0 \tag{1.5}$$

[1] We also need different θ values to give different pdfs for X, and that the support space for X is the same for all θ.

- **Theorem 6.1.3**: Under the same regularity conditions, the maximum likelihood estimator, if it exists, converges in probability to θ_0:

$$\widehat{\theta}_n \xrightarrow{P} \theta_0 \tag{1.6}$$

The first theorem says that the true value becomes the maximum likelihood value in the limit of an infinitely large sample, while the second says that the mle becomes the true value in that limit.

1.1.2 Example: Signal Amplitude

See Hogg for several examples of maximum-likelihood solutions associated with random samples. To complement those, we're going to consider a couple of cases where the random vector is not quite a random sample, but still drawn from a distribution with a parameter θ.

First, consider a simplified example from signal processing. Suppose you have a set of data points $\{x_i\}$ which represent a signal with a known shape, described by $\{h_i\}$ and an unknown amplitude θ, on top of which there is some Gaussian noise with variances $\{\sigma_i^2\}$. I.e., the distribution associated with each data value is $X_i \sim N(\theta h_1, \sigma_i^2)$. This means the likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i - \theta h_i)^2}{2\sigma_i^2}\right) \tag{1.7}$$

and the log-likelihood is

$$\ell(\theta) = -\sum_{i=1}^{n} \frac{(x_i - \theta h_i)^2}{2\sigma_i^2} + \text{const} \tag{1.8}$$

The derivative is

$$\ell(\theta) = \sum_{i=1}^{n} \frac{h_i(x_i - \theta h_i)}{\sigma_i^2} = \sum_{i=1}^{n} \frac{h_i x_i}{\sigma_i^2} - \theta \sum_{i=1}^{n} \frac{h_i^2}{\sigma_i^2} \tag{1.9}$$

so the maximum likelihood estimate is

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} h_i x_i/\sigma_i^2}{\sum_{j=1}^{n} h_j^2/\sigma_j^2} \tag{1.10}$$

We can also write this in matrix form (which would allow generalization to covariances between data points via a non-diagonal variance-covariance matrix $\mathbf{\Sigma}$)

$$\widehat{\theta} = \left(\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{h}\right)^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} \tag{1.11}$$

This makes the signal model, with maximum-likelihood amplitude,

$$\mathbf{h}\widehat{\theta} = \mathbf{h}\left(\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{h}\right)^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} = \mathbf{\Sigma}^{1/2}\mathbf{P}\mathbf{\Sigma}^{-1/2}\mathbf{x} \tag{1.12}$$

where

$$\mathbf{P} = \mathbf{\Sigma}^{-1/2}\mathbf{h}\left(\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{h}\right)^{-1}\mathbf{h}^{\mathrm{T}}\mathbf{\Sigma}^{-1/2} \tag{1.13}$$

is a projection matrix.

Note that in a more complicated problem, we might have a superposition of different templates, each with its own amplitude[2], so the signal would be $\mathbf{X} \sim N_n\left(\sum_{i=1}^{k}\theta_i\mathbf{h}_i, \mathbf{\Sigma}\right)$

1.1.3 Example: Bradley-Terry Model

The Bradley-Terry model[3] is designed to model "paired comparisons" between objects (teams or players in games, foods in taste tests, etc). Each object has a strength π_i, $0 < \pi_i < \infty$, and in a comparison between two objects, object i will be chosen over object j with a probability $\frac{\pi_i}{\pi_i+\pi_j}$. Consider a restricted

[2] E.g., for continuous gravitational waves, you have two polarizations, each with amplitude and phase, so four parameters.

[3] Zermelo, *Mathematische Zeitschrift* **29**, 436 (1929); Bradley and Terry *Biometrika* **39**, 324 (1952)

version of the model where an object with unknown strength θ is compared to a series of n objects[4] with known strengths π_1, π_2, ..., π_n. Let X_i be a Bernoulli random variable which is 1 if the object wins the ith comparison and 0 if it loses, so that

$$X_i \sim b\left(1, \frac{\theta}{\theta + \pi_i}\right) \tag{1.14}$$

Then the likelihood is

$$L(\theta; \mathbf{x}) = f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} \pi_i^{1-x_i}}{\theta + \pi_i} \tag{1.15}$$

and the log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \left[x_i \ln \theta + (1 - x_i) \ln \pi_i - \ln(\theta + \pi_i)\right] \tag{1.16}$$

and

$$\ell'(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - \sum_{i=1}^{n} \frac{1}{\theta + \pi_i} \tag{1.17}$$

If we define $y = \sum_{i=1}^{n} x_i$ to be the total number of comparisons won, the mle for θ satisfies

$$y = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{\widehat{\theta}}{\widehat{\theta} + \pi_i} \tag{1.18}$$

The right hand side is the average number of comparison wins you'd predict, and the mle is just the strength which makes this equal to the actual number observed.

(1.18) can't be solved in closed form, one can find the mle numerically by iterating[5]

$$\widehat{\theta} = \frac{y}{\sum_{i=1}^{n} (\widehat{\theta} + \pi_i)^{-1}} \tag{1.19}$$

[4]Note that, since the strengths are assumed to be known, some of these comparisons may actually be repeated comparisons with the same object.

[5]Ford, *American Mathematical Monthly* **64**, 28 (1957)

1.1.4 Reparameterization

A key property of maximum-likelihood estimates is captured in Hogg's

- **Theorem 6.1.2**: If you change variables in the likelihood from θ to some other $\eta = g(\theta)$, and work out the mle for η, it is $\widehat{\eta} = g(\widehat{\theta})$.

The proof is almost trivial, so rather than repeat it, we'll demonstrate the property in our Bradley-Terry example. Let $\lambda = \ln \theta$ so that $-\infty < \lambda < \infty$. The calculation of the log likelihood proceeds as before, and we can substitute $\theta = e^{\lambda}$ into (1.16) to get

$$\ell(\lambda) = \sum_{i=1}^{n} \left[x_i \lambda + (1 - x_i) \ln \pi_i - \ln(e^{\lambda} + \pi_i)\right] \tag{1.20}$$

The derivative is

$$\ell'(\lambda) = \sum_{i=1}^{n} \left(x_i - \frac{e^{\lambda}}{e^{\lambda} + \pi_i}\right) \tag{1.21}$$

so the maximum-likelihood equation is

$$y = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{e^{\widehat{\lambda}}}{e^{\widehat{\lambda}} + \pi_i} \tag{1.22}$$

which is just the same as (1.18) with $\widehat{\theta}$ replaced by $e^{\widehat{\lambda}}$.

Tuesday 23 February 2016 – Read Section 6.2 of Hogg

1.2 Fisher Information and The Cramér-Rao Bound

We now turn to a consideration of the properties of estimators and other statistics. Suppose we have a random vector **X** which

is a random sample from a distribution $f(x;\theta)$ which includes a parameter θ. We know the joint sampling distribution is

$$f_{\mathbf{X}|\theta}(x_1,\ldots,x_n|\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{1.23}$$

If we have a statistic $Y = u(\mathbf{X})$, this is a random variable. The mean and variance of Y are not random variables, but they depend on θ because the sampling distribution does:

$$\mu_Y(\theta) = E\left(Y|\theta\right) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u(x_1,\ldots,x_n)\prod_{i=1}^{n} f(x_i;\theta)\,dx_i \tag{1.24}$$

and

$$\begin{aligned}\mathrm{Var}(Y|\theta) &= E\left(\left(Y-\mu_Y(\theta)\right)^2|\theta\right)\\ &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[u(\mathbf{x})-\mu_Y(\theta)\right]^2 f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\,d^n x\end{aligned} \tag{1.25}$$

1.2.1 Fisher Information

The first such function we want to consider is the derivative with respect to the parameter θ of the log-likelihood $\ell(\theta;x) = f(x;\theta)$. We will focus first on the case of a single random variable before considering a random sample. We will also assume an even broader set of regularity conditions than last time, specifically that we're allowed to interchange derivatives and integrals.

The quantity of interest is

$$u(x) = \ell'(\theta,x) = \frac{\partial}{\partial\theta}\ln f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta} \tag{1.26}$$

and is one measure of how much the probability distribution changes when we change the parameter. Note in this case the function $u(x)$ depends explicitly on the parameter, in addition to the θ dependence of the distribution for the random variable X. $u(X) = \ell'(\theta,X)$ is a statistic whose expectation value is

$$E\left(\ell'(\theta,X)\right) = \int_{\mathcal{S}}\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}\,f(x;\theta)\,dx = \int_{\mathcal{S}}\frac{\partial f(x;\theta)}{\partial\theta}\,dx \tag{1.27}$$

where we explicitly restrict the integral to the support space $x\in\mathcal{S}$ so that $f(x;\theta)>0$. Now the regularity conditions (especially the fact that $\mathcal{S}$ is the same for any allowed value of θ) allow us to pull the derivative outside the integral so we have

$$E\left(\ell'(\theta,X)\right) = \frac{d}{d\theta}\int_{\mathcal{S}} f(x;\theta)\,dx = \frac{d}{d\theta}(1) = 0 \tag{1.28}$$

where we have used the fact that $f(x;\theta)$ is a normalized pdf in x. So note that while $\ell'(\widehat{\theta}(x),x) = 0$, i.e., the derivative of the log-likelihood is zero at the maximum-likelihood value of θ for a given x, the expectation value $E\left(\ell'(\theta,X)\right)$ vanishes for any θ. This allows us to write the variance as

$$\begin{aligned}\mathrm{Var}\left(\ell'(\theta,X)\right) &= E\left(\left[\ell'(\theta,X)\right]^2\right)\\ &= \int_{\mathcal{S}}\left(\frac{\partial\ln f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)\,dx \equiv I(\theta)\end{aligned} \tag{1.29}$$

$I(\theta)$ is a property of the distribution known as the *Fisher information*. You've encountered it on the homework in the context of the Jeffreys prior $f_\Theta(\theta)\propto\sqrt{I(\theta)}$.

Note that if we go back and differentiate (1.28) we get

$$\begin{aligned}0 &= \frac{d}{d\theta}E\left(\ell'(\theta,X)\right) = \frac{d}{d\theta}\int_{\mathcal{S}}\frac{\partial\ln f(x;\theta)}{\partial\theta}\,f(x;\theta)\,dx\\ &= \int_{\mathcal{S}}\left(\frac{\partial^2\ln f(x;\theta)}{\partial\theta^2}\,f(x;\theta) + \frac{\partial\ln f(x;\theta)}{\partial\theta}\,\frac{\partial f(x;\theta)}{\partial\theta}\,dx\right)dx\\ &= \int_{\mathcal{S}}\frac{\partial^2\ln f(x;\theta)}{\partial\theta^2}\,f(x;\theta)\,dx + \int_{\mathcal{S}}\left(\frac{\partial\ln f(x;\theta)}{\partial\theta}\right)^2 f(x;\theta)\,dx\end{aligned} \tag{1.30}$$

This means there's an equivalent (and sometimes easier to calculate way to write the Fisher information as

$$I(\theta) = -\int_{\mathcal{S}} \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\, f(x;\theta)\, dx = -E\left(\ell''(\theta; X)\right) \tag{1.31}$$

Example: Fisher information for the mean of a normal distribution To give a very simple example, consider the normal distribution $N(\theta, \sigma^2)$. The likelihood is

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\theta))}{2\sigma^2}\right) \tag{1.32}$$

so the log-likelihood is

$$\ln f(x;\theta) = -\frac{(x-\theta)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \tag{1.33}$$

whose derivatives are

$$\frac{\partial}{\partial\theta}\ln f(x;\theta) = \frac{x-\theta}{\sigma^2} \tag{1.34a}$$

$$\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta) = -\frac{1}{\sigma^2} \tag{1.34b}$$

So the Fisher information is $I(\theta) = \frac{1}{\sigma^2}$. (Note in passing that the expectation of the first derviative is $\frac{1}{\sigma^2}(X-\theta)$, whose expectation value is indeed zero.)

Fisher information for a sample of size n We can now consider the Fisher information associated with a random sample of size n, given that

$$I(\theta) = -E\left(\left.\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right|\theta\right) \tag{1.35}$$

In fact, since the joint sampling distribution, and therefore the likelihood for the sample, is

$$f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta)\ , \tag{1.36}$$

the log likelihood will be

$$\ln f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)\ , \tag{1.37}$$

and therefore the Fisher information is

$$-E\left(\left.\frac{\partial^2 \ln f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)}{\partial\theta^2}\right|\theta\right) = -\sum_{i=1}^{n} E\left(\left.\frac{\partial^2 \ln f(X_i;\theta)}{\partial\theta^2}\right|\theta\right) = nI(\theta) \tag{1.38}$$

where we've used the fact that both the derivative and the expectation value are linear operations, and that the random variables $\{X_i\}$ are iid.

Note that the Fisher information for a sample of size n from a normal distribution is $\frac{n}{\sigma^2}$, which is one over the variance of the sample mean.

1.2.2 The Cramér-Rao Bound

Now we return to consideration of a general statistic $Y = u(\mathbf{X})$, with a special eye towards one being used as an estimator of the parameter θ. Recalling the definition of the expectation value

$$\mu_Y(\theta) = E(Y) = \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} u(x_1,\ldots,x_n)\, f_{\mathbf{X}|\Theta}(\mathbf{x},\theta)\, d^n x \tag{1.39}$$

and variance

$$\begin{aligned}\mathrm{Var}(Y|\theta) &= E\left([Y-\mu_Y(\theta)]^2|\theta\right) \\ &= \int_{\mathcal{S}} \cdots \int_{\mathcal{S}} [u(\mathbf{x})-\mu_Y(\theta)]^2\, f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\, d^n x\end{aligned} \tag{1.40}$$

we notice a similarity to the Fisher information for the sample

$$nI(\theta) = E\left(\left(\frac{\partial \ln f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta)}{\partial\theta}\right)^2 \middle| \theta\right) \tag{1.41}$$

Both are of the form $E\left([a(\mathbf{X})]^2|\theta\right) = \langle a|a\rangle$ where we have defined the *inner product*

$$\langle a|b\rangle = E\left(a(\mathbf{X})\, b(\mathbf{X})|\theta\right) = \int_{\mathcal{S}} a(x)\, b(x)\, f(x;\theta)\, dx \tag{1.42}$$

This behaves in many ways like a dot product $\vec{a}\cdot\vec{b}$, and in particular satisfies the *Cauchy-Schwarz inequality*[6]

$$\langle a|a\rangle\, \langle b|b\rangle \geq |\langle a|b\rangle|^2 \tag{1.43}$$

This means that

$$\mathrm{Var}(Y|\theta)\, nI(\theta) \geq E\left((Y-\mu_Y(\theta))\frac{\partial \ln f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta)}{\partial\theta}\middle|\theta\right)$$

$$= E\left(u(\mathbf{X})\left(\frac{\partial \ln f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta)}{\partial\theta}\right)\middle|\theta\right) - \mu_Y(\theta) \underbrace{E\left(\frac{\partial \ln f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta)}{\partial\theta}\middle|\theta\right)}_{0}$$

$$= \int_{\mathcal{S}}\cdots\int_{\mathcal{S}} u(x_1,\ldots,x_n)\, \frac{1}{\cancel{f_{\mathbf{X}|\Theta}(\mathbf{x},\theta)}}\, \frac{\partial f_{\mathbf{X}|\Theta}(\mathbf{X}|\theta)}{\partial\theta}\, \cancel{f_{\mathbf{X}|\Theta}(\mathbf{x},\theta)}\, d^n x$$

$$= \frac{d}{d\theta}\int_{\mathcal{S}}\cdots\int_{\mathcal{S}} u(x_1,\ldots,x_n)\, f_{\mathbf{X}|\Theta}(\mathbf{x},\theta)\, d^n x = \frac{d}{d\theta}\mu_Y(\theta) \tag{1.44}$$

where the last step holds as long as $u(\mathbf{x})$ doesn't depend explicitly on θ. The inequality can be solved for the variance of the statistic Y to give the *Cramér-Rao bound*:

$$\mathrm{Var}(Y|\theta) \geq \frac{\mu_Y(\theta)}{nI(\theta)} \tag{1.45}$$

[6] In vector algebra, this is a consequence of the law of cosines $\vec{a}\cdot\vec{b} = |\vec{a}|\left|\vec{b}\right|\cos\phi(\vec{a},\vec{b})$

In the special case where Y is an unbiased estimator of θ, so that $\mu_Y(\theta)$, it simplifies to

$$\mathrm{Var}(Y|\theta) \geq \frac{1}{nI(\theta)} \qquad \text{if } E\left(Y|\theta\right) = \theta \tag{1.46}$$

I.e., the variance of an unbiased estimator can be no less than the reciprocal of the Fisher information of the sample.

To return to our simple example of a Gaussian sampling distribution $N(\theta,\sigma^2)$, let $Y = \overline{X}$, the sample mean, which we know is an unbiased estimator of the distribution mean θ. Then the Cramér-Rao bound tells us

$$\mathrm{Var}(\overline{X}|\theta) \geq \frac{1}{nI(\theta)} = \frac{\sigma^2}{n} \tag{1.47}$$

Of course, we know that the variance of the sample mean *is* σ^2/n, which means that in this case the bound is saturated, and the sample mean is the lowest-variance unbiased estimator of the distribution mean for a normal distribution with known variance.

When the Cramér-Rao bound is saturated, i.e., $\mathrm{Var}(Y|\theta) = \frac{1}{nI(\theta)}$ for an unbiased estimator, we say that Y is an *efficient estimator* of θ. Otherwise, we can define the ratio of the Cramér-Rao bound to the actual variance as the *efficiency* of the estimator.

Another selling point of maximum likelihood estimates is their relationship to efficiency. A MLE is not always an efficient estimator, but one can prove using Taylor series that it becomes so in the limit that the sample size goes to infinity. Another important theorem (the proof of which we omit for now) is that, subject to the usual regularity conditions, the distribution for the MLE converges to a Gaussian:

$$\sqrt{n}\left(\widehat{\theta}(X) - \theta\right) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right) \tag{1.48}$$

Tuesday 1 March 2016
– Read Section 6.3 of Hogg

2 Maximum-Likelihood Tests

So far we've used maximum likelihood methods to consider ways to estimate the parameter θ of a distribution $f(x;\theta)$. We now consider tests designed to compare a null hypothesis $\mathcal{H}_0$ which specifies $\theta = \theta_0$ and $\mathcal{H}_1$, which allows for arbitrary θ within the allowed parameter space. (Note that since these are classical hypothesis tests, we don't require $\mathcal{H}_1$ to specify a prior probability distribution for θ.) In each case, we will construct a test statistic from a sample of size n, drawn from $f(x;\theta)$, which is asymptotically chi-square distributed in the limit $n \to \infty$.

2.1 Likelihood Ratio Test

Define as usual the likelihood

$$L(\theta;\mathbf{x}) = f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{2.1}$$

and its logarithm

$$\ell(\theta;\mathbf{x}) = \ln f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta) \tag{2.2}$$

Recall that for Bayesian hypothesis testing, the natural quantity to construct was the *Bayes Factor*

$$\mathcal{B}_{01} = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{L(\theta_0;\mathbf{x})}{\int_{\theta\in\Omega} L(\theta;\mathbf{x})\, f_\Theta(\theta|\mathcal{H}_1)\, d\theta} \tag{2.3}$$

in the frequentist case we don't define a prior $f_\Theta(\theta|\mathcal{H}_1)$, so instead, of all the θ values at which to evaluate the likelihood, we choose the one which maximizes it[7], and look at the *likelihood ratio*

$$\Lambda(\mathbf{x}) = \frac{L(\theta_0;\mathbf{x})}{L(\widehat{\theta}(\mathbf{x});\mathbf{x})} \tag{2.4}$$

Note that in the frequentist picture $L(\theta;\mathbf{X})$ and $\Lambda(\mathbf{X})$ are statistics, i.e., random variables constructed from the sample $\mathbf{X}$.

2.1.1 Example: Gaussian Distribution

Consider a sample of size n drawn from a $N(\theta,\sigma^2)$ distribution, for which we know the mle is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n}$ and the likelihood is

$$\begin{aligned} L(\theta;\overline{x}) &= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &\propto \exp\left(-\frac{n}{2\sigma^2}(\theta-\overline{x})^2\right) \end{aligned} \tag{2.5}$$

and the likelihood ratio is

$$\Lambda(\mathbf{x}) = \exp\left(-\frac{n}{2\sigma^2}(\theta_0-\overline{x})^2\right) \tag{2.6}$$

Since $\overline{X} \sim N(\theta,\sigma^2/n)$, we have

$$-2\ln\Lambda(\mathbf{X}) = \frac{(\overline{X}-\theta_0)^2}{\sigma^2/n} \sim \chi^2(1) \qquad \text{assuming } \mathcal{H}_0 \tag{2.7}$$

Now, this won't be exactly true in general, but the various asymptotic theorems show that in general, for a regular underlying distribution, it's true in the asymptotic sense:

$$\chi^2_L = -2\ln\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2(1) \qquad \text{assuming } \mathcal{H}_0 \tag{2.8}$$

[7] There are various arguments which make this a reasonable choice, most revolving around the fact that for large samples, the likelihood function is approximately Gaussian with a peak value of $L(\widehat{\theta}(\mathbf{x});\mathbf{x})$.

For finite n, we assume this is approximately true, and compare $-2\ln\Lambda(\mathbf{X})$ to the $100\times(1-\alpha)$th percentile of $\chi^2(1)$ distribution to get a test with false alarm probability α.

2.2 Wald Test

The statistic $-2\ln\Lambda(\mathbf{X})$ constructed from the likelihood ratio is only one of several statistics which are asymptotically χ^2. Another possibility is to consider the result that

$$\sqrt{n}\left(\widehat{\theta}(X)-\theta_0\right)\xrightarrow{D}N\left(0,\frac{1}{I(\theta_0)}\right) \tag{2.9}$$

and construct the statistic

$$\chi^2_W=\frac{n\left(\widehat{\theta}(X)-\theta_0\right)^2}{1/I(\widehat{\theta}(X))}=nI(\widehat{\theta}(X))\left(\widehat{\theta}(X)-\theta_0\right)^2 \tag{2.10}$$

Comparing this to the percentiles of a $\chi^2(1)$ is known as the Wald test.

2.2.1 Example: Gaussian Distribution

The distribution of the statistic χ^2_W should converge to $\chi^2(1)$ as $n\to\infty$. We can see what the exact distribution is for a given choice of the underlying sampling distribution. Assuming again a $N(\theta,\sigma^2)$ distribution for $f(x;\theta)$, we have Fisher information $I(\theta)=\frac{1}{\sigma}^2$, maximum likelihood estimator $\widehat{\theta}(X)=\overline{X}$, and Wald test statistic

$$\chi^2_W=\frac{n\left(\overline{X}-\theta_0\right)^2}{\sigma^2} \tag{2.11}$$

which is of the same form as (2.7), which we know to be exactly $\chi^2(1)$ for any n.

2.3 Rao Scores Test

The final likelihood-based test involves the "score"

$$\frac{\partial\ln f(X_i;\theta)}{\partial\theta} \tag{2.12}$$

associated with each random variable in the sample, whose sum is the derivative of the log-likelihood

$$\ell'(\theta;\mathbf{X})=\frac{\partial\ln f(X_i;\theta)}{\partial\theta} \tag{2.13}$$

We know the expectation value is

$$E\left(\ell'(\theta;\mathbf{X})\right)=\sum_{i=1}^{n}E\left(\frac{\partial\ln f(X_i;\theta)}{\partial\theta}\right)=\sum_{i=1}^{n}0=0 \tag{2.14}$$

and the variance is

$$\mathrm{Var}\left(\ell'(\theta;\mathbf{X})\right)=\sum_{i=1}^{n}\mathrm{Var}\left(\frac{\partial\ln f(X_i;\theta)}{\partial\theta}\right)=\sum_{i=1}^{n}I(\theta)=nI(\theta) \tag{2.15}$$

so we construct a test statistic which, in the limit that the sum of the scores is normally distributed, becomes a $\chi^2(1)$:

$$\chi^2_R=\frac{\left(\ell'(\theta_0;\mathbf{X})\right)^2}{nI(\theta_0)} \tag{2.16}$$

Comparing this to the percentiles of a $\chi^2(1)$ is known as the Rao scores test. It has the advantage over the other two tests that we don't need to work out the maximum likelihood estimator to apply it.

2.3.1 Example: Gaussian Distribution

If the underlying distribution is $N(\theta,\sigma^2)$, so that

$$\ln f(x;\theta)=-\frac{1}{2}\ln 2\pi\sigma^2-\frac{(x-\theta)^2}{2\sigma^2} \tag{2.17}$$

and

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = \frac{x-\theta}{\sigma^2} \tag{2.18}$$

which makes the derivative of the log-likeihood

$$\ell'(\theta;\mathbf{X}) = \sum_{i=1}^{n} \frac{X_i - \theta}{\sigma^2} = n\,\frac{\overline{X}-\theta}{\sigma^2} \tag{2.19}$$

which makes the Rao scores statistic

$$\chi^2_R = n^2\,\frac{(\overline{X}-\theta_0)^2}{\sigma^4}\,\frac{\sigma^2}{n} = \frac{(\overline{X}-\theta_0)^2}{\sigma^2/n} \tag{2.20}$$

which is, again, exactly $\chi^2(1)$ distributed in this case.

Wednesday 2 March 2016 – Review for Prelim Exam One

The exam covers materials from the first four class-weeks of the term, i.e., Hogg sections 11.1-11.3 and 6.1-6.2 (and associated topics covered in class through February 23), and problem sets 1-4.

Thursday 3 March 2016 – First Prelim Exam

Tuesday 8 March 2016 – Read Section 6.4 of Hogg

3 Multi-Parameter Methods

We now consider the case where the probability distribution $f(x;\boldsymbol{\theta})$ depends not just on a single parameter, but p parameters $\{\theta_1,\theta_2,\ldots,\theta_p\} = \{\theta_\alpha\} \equiv \boldsymbol{\theta}$. The likelihood for a sample of size n is then

$$L(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}) \tag{3.1}$$

and the log-likelihood is

$$\ell(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i;\boldsymbol{\theta}) \tag{3.2}$$

3.1 Maximum Likelihood Estimation

We're still interested in the set of parameter values $\{\widehat{\theta}_1,\ldots,\widehat{\theta}_p\}$ which maximize the likelihood (or, equivalently, its logarithm). But to maximize a function of p parameters, we need to set all of the partial derivatives to zero, which means there are now p maximum likelihood equations

$$\frac{\partial \ell(\boldsymbol{\theta};\mathbf{x})}{\partial \theta_\alpha} = 0 \qquad \alpha = 1,2,\ldots,p \tag{3.3}$$

We can in principle solve these p equations for the p unknowns $\{\widehat{\theta}_\alpha\}$.

For example, consider a sample of size n drawn from a $N(\theta_1,\theta_2)$ distribution with pdf

$$f(x;\theta_1,\theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right) \tag{3.4}$$

so that

$$\ln f(x;\theta_1,\theta_2) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\theta_2 - \frac{(x-\theta_1)^2}{2\theta_2} \tag{3.5}$$

The partial derivatives are

$$\frac{\partial \ln f(x;\theta_1,\theta_2)}{\partial \theta_1} = \frac{x-\theta_1}{\theta_2} \tag{3.6a}$$

$$\frac{\partial \ln f(x;\theta_1,\theta_2)}{\partial \theta_2} = -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2} \tag{3.6b}$$

Taking the partial derivatives of (3.2) and using linearity then gives us

$$\frac{\partial \ell(\mathbf{x};\theta_1,\theta_2)}{\partial \theta_1} = \frac{(\sum_{i=1}^n x_i) - n\theta_1}{\theta_2} \tag{3.7a}$$

$$\frac{\partial \ell(\mathbf{x};\theta_1,\theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\sum_{i=1}^n (x_i-\theta_1)^2}{2\theta_2^2} \tag{3.7b}$$

We need to set both of these to zero to find $\widehat{\theta}_1$ and $\widehat{\theta}_2$, i.e., the coupled maximum-likelihood equations are

$$\frac{(\sum_{i=1}^n x_i) - n\widehat{\theta}_1}{\widehat{\theta}_2} = 0 \tag{3.8a}$$

$$-\frac{n}{2\widehat{\theta}_2} + \frac{\sum_{i=1}^n (x_i-\widehat{\theta}_1)^2}{2\widehat{\theta}_2^2} = 0 \tag{3.8b}$$

We can solve the first equation for

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x} \tag{3.9}$$

and then substitute that into the second to get

$$\frac{n}{2\widehat{\theta}_2} = \frac{\sum_{i=1}^n (x_i-\overline{x})^2}{2\widehat{\theta}_2^2} \tag{3.10}$$

i.e.,

$$\widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2 = \frac{n-1}{n}\, s \tag{3.11}$$

where $s = \frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2$ is the usual (unbiased) sample variance.

3.2 Fisher Information Matrix

If we consider the "score" statistic for a single observation

$$\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \theta_\alpha} \tag{3.12}$$

we see that it's now actually a random vector with one component per parameter. It is still true that

$$E\left(\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \theta_\alpha}\right) = \int_{\mathcal{S}} \frac{1}{\cancel{f(x;\boldsymbol{\theta})}} \frac{\partial f(x;\boldsymbol{\theta})}{\partial \theta_\alpha} \cancel{f(x;\boldsymbol{\theta})}\, dx = 0 \tag{3.13}$$

(You can show this by taking $\frac{\partial}{\partial \theta_\alpha}$ of the normalization integral for $f(x;\boldsymbol{\theta})$.) Since this is a random vector, we can construct its variance-covariance matrix $\mathbf{I}(\boldsymbol{\theta})$ with elements

$$I_{\alpha\beta}(\boldsymbol{\theta}) = E\left(\left[\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \theta_\alpha}\right]\left[\frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial \theta_\beta}\right]\right) \tag{3.14}$$

As before, we can differentiate (3.13) with respect to θ_β to get the identity

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_\beta}\int_{\mathcal{S}} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial \theta_\alpha} f(x;\boldsymbol{\theta})\, dx \\ &= \int_{\mathcal{S}} \frac{\partial^2 \ln f(x;\boldsymbol{\theta})}{\partial \theta_\alpha\, \partial \theta_\beta} f(x;\boldsymbol{\theta})\, dx + \int_{\mathcal{S}} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial \theta_\alpha} \frac{\partial f(x;\boldsymbol{\theta})}{\partial \theta_\beta}\, dx \\ &= E\left(\frac{\partial^2 \ln f(X;\boldsymbol{\theta})}{\partial \theta_\alpha\, \partial \theta_\beta}\right) + \int_{\mathcal{S}} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial \theta_\alpha} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial \theta_\beta} f(x;\boldsymbol{\theta})\, dx \end{aligned} \tag{3.15}$$

i.e., an equivalent way of writing the Fisher information matrix is

$$I_{\alpha\beta}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln f(X;\boldsymbol{\theta})}{\partial \theta_\alpha\, \partial \theta_\beta}\right) \tag{3.16}$$

Note that by definition and construction, the Fisher matrix is symmetric.

3.2.1 Example: normal distribution

If we return to the case of a $N(\theta_1, \theta_2)$ distribution, taking derivatives of (3.6) gives us

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial \theta_1^2} = -\frac{1}{\theta_2} \tag{3.17a}$$

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial \theta_2^2} = \frac{1}{2\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3} \tag{3.17b}$$

$$\frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial \theta_2\, \partial \theta_1} = \frac{\partial^2 \ln f(x;\theta_1,\theta_2)}{\partial \theta_1\, \partial \theta_2} = -\frac{x-\theta_1}{2\theta_2^2} \tag{3.17c}$$

If we recall $E(X) = \theta_1$ and $\mathrm{Var}(X) = E((X-\theta_1)^2) = \theta_2$, we get the Fisher information matrix elements $I_{\alpha\beta}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln f(X;\theta_1,\theta_2)}{\partial\theta_\alpha\, \partial\theta_\beta}\right)$:

$$I_{11}(\boldsymbol{\theta}) = \frac{1}{\theta_2} \tag{3.18a}$$

$$I_{22}(\boldsymbol{\theta}) = -\frac{1}{2\theta_2^2} + \frac{\theta}{\theta_2^3} = \frac{1}{2\theta_2^2} \tag{3.18b}$$

$$I_{12}(\boldsymbol{\theta}) = I_{21} = 0 \tag{3.18c}$$

i.e.,

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix} \tag{3.19}$$

3.2.2 Fisher Information Matrix for a Random Sample

As in the single-parameter case, linearity means that the Fisher information matrix for a sample of size n is just n times the FIM for the distribution:

$$-E\left(\frac{\partial^2 \ell(\boldsymbol{\theta};\mathbf{X})}{\partial\theta_\alpha\, \partial\theta_\beta}\right) = -\sum_{i=1}^{n} E\left(\frac{\partial^2 \ln f(X;\boldsymbol{\theta})}{\partial\theta_\alpha\, \partial\theta_\beta}\right) = nI_{\alpha\beta}(\boldsymbol{\theta}) \tag{3.20}$$

3.2.3 Error Estimation

Recall that for a model with a single parameter θ, the Cramér-Rao bound stated that an unbiased estimator of θ had minimum variance of $\frac{1}{nI(\theta)}$. Also, the maximum-likelihood estimator $\widehat{\theta}$ from a sample of size n saturated that bound in the limit $n \to \infty$, and in fact its distribution converged to a Gaussian with the minimum variance specified by the bound:

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{D} N\left(0, \frac{1}{I(\theta)}\right) \tag{3.21}$$

In the multi-parameter case, we have a maximum-likelihood estimator $\widehat{\boldsymbol{\theta}}$ which is a random vector with p components $\{\widehat{\theta}_\alpha | \alpha = 1, \ldots, p\}$. The corresponding limiting distribution is a multivariate normal:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{D} N_p\left(0, \mathbf{I}(\boldsymbol{\theta})^{-1}\right) \tag{3.22}$$

The matrix inverse $\mathbf{I}(\boldsymbol{\theta})^{-1}$ of the Fisher information matrix is the variance-covariance matrix of the limiting distribution.

It is also this inverse matrix which provides the multiparameter version of the Cramér-Rao bound. If $T_\alpha(\mathbf{X})$ is an unbiased estimator of one parameter θ_α, the lower bound on its variance is

$$\mathrm{Var}\left(T_\alpha(\mathbf{X})\right) \geq \frac{1}{n}\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{\alpha\alpha} \tag{3.23}$$

I.e., the diagonal elements of the inverse Fisher matrix provide lower bounds on the variance of unbiased estimators of the parameters. In cases like (3.19), where the Fisher matrix is diagonal, its inverse is also diagoal, with $\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{\alpha\alpha}$ being the same

as $1/I(\boldsymbol{\theta})_{\alpha\alpha}$. But in general,[8]

$$\left[\mathbf{I}^{-1}(\boldsymbol{\theta})\right]_{\alpha\alpha} \geq 1/I(\boldsymbol{\theta})_{\alpha\alpha} = \frac{1}{I(\theta_\alpha)} \tag{3.24}$$

This is thus a stronger bound than one would get by assuming all of the "nuisance" parameters to be known and applying the usual Cramér-Rao bound to the parameter of interest.

Thursday 10 March 2016 – Read Section 6.5 of Hogg

3.3 Maximum Likelihood Tests

We return now to the question of hypothesis testing when the parameter space associated with the distribution is multi-dimensional. Hogg considers this in a somewhat limited context where the null hypothesis $\mathcal{H}_0$ involves picking values for one or more parameters or combinations of parameters and leaving the rest unspecified. Formally this is written as $\mathcal{H}_0 : \boldsymbol{\theta} \in \omega$ where ω is some lower-dimensional subspace of the parameter space Ω. The alternative hypothesis is $\mathcal{H}_1 : [\boldsymbol{\theta} \in \Omega] \wedge [\boldsymbol{\theta} \notin \omega]$ The number of parameters in the full parameter space is p, while ω is defined by specifing values for q independent functions of the parameters, where $0 < q \leq p$. This means that ω is a $p-q$-dimensional space. The case considered before was $p = q = 1$ so that ω was a zero-dimensional point in the one-dimensional parameter space Ω.

Since $\mathcal{H}_0$ is also a composite hypothesis with different possible parameter values, if we were constructing a Bayes factor, we'd also need to specify a prior distribution for $\boldsymbol{\theta} \in \omega$ associated with $\mathcal{H}_0$:

$$\mathcal{B}_{01} = \frac{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_0)}{f_{\mathbf{X}}(\mathbf{x}|\mathcal{H}_1)} = \frac{\int_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_0)\, d^{p-q}\theta}{\int_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta};\mathbf{x})\, f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\mathcal{H}_1)\, d^{p}\theta} \tag{3.25}$$

In classical statistics we don't have a prior distribution associated with either hypothesis, so we let them each "put their best foot forward" by picking the parameter values which maximize the likelihood, which we refer to as $\widehat{\boldsymbol{\theta}}(\mathbf{x}) = \text{argmax}_{\boldsymbol{\theta}\in\Omega}\, L(\boldsymbol{\theta};\mathbf{x})$ and $\widehat{\boldsymbol{\theta}}_0(\mathbf{x}) = \text{argmax}_{\boldsymbol{\theta}\in\omega}\, L(\boldsymbol{\theta};\mathbf{x})$. The likelihood ratio for use in tests is thus

$$\Lambda(\mathbf{x}) = \frac{L(\widehat{\boldsymbol{\theta}}_0(\mathbf{x});\mathbf{x})}{L(\widehat{\boldsymbol{\theta}}(\mathbf{x});\mathbf{x})} = \frac{\max_{\boldsymbol{\theta}\in\omega} L(\boldsymbol{\theta};\mathbf{x})}{\max_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta};\mathbf{x})} \tag{3.26}$$

3.3.1 Example: Mean of Normal Distribution with Unknown Variance

As an example of the method, consider a sample of size n drawn from a $N(\theta_1, \theta_2)$ i.e., a normal distribution where the mean and the variance are both unknown parameters. The likelihood is

$$L(\boldsymbol{\theta},\mathbf{x}) = (2\pi\theta_2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i-\theta_1)^2}{2\theta_2}\right) \tag{3.27}$$

$\mathcal{H}_0$ is $\theta_1 = \mu_0$, $0 < \theta_2 < \infty$, while $\mathcal{H}_1$ is $-\infty < \theta_1 < \infty$, $0 < \theta_2 < \infty$. $\mathcal{H}_0$ is effectively a one-parameter model which has mles of

$$\widehat{\theta}_{01} = \mu_0 \tag{3.28a}$$

$$\widehat{\theta}_{02} = \frac{1}{n}\sum_{i=0}^{n}(x_i-\mu_0) = \widehat{\sigma}_0^2 \tag{3.28b}$$

[8] This is easy to see in the case $p = 2$, where $\mathbf{I} = \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}$ and $\mathbf{I}^{-1} = \frac{1}{I_{11}I_{22}-I_{12}^2}\begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix}$ so that $\frac{1}{[\mathbf{I}^{-1}]_{11}} = I_{11} - \frac{I_{12}^2}{I_{22}} \leq I_{11}$.

while the MLE for $\mathcal{H}_1$ was found last time to be

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \tag{3.29a}$$

$$\widehat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \widehat{\sigma}^2 \tag{3.29b}$$

We see that the maximized likelihoods are

$$L(\widehat{\boldsymbol{\theta}}(\mathbf{x}); \mathbf{x}) = (2\pi\widehat{\sigma}^2)^{-n/2} e^{-n/2} \tag{3.30a}$$

$$L(\widehat{\boldsymbol{\theta}}_0(\mathbf{x}); \mathbf{x}) = (2\pi\widehat{\sigma}_0^2)^{-n/2} e^{-n/2} \tag{3.30b}$$

So the likelihood ratio is

$$\Lambda(\mathbf{X}) = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}\right)^{n/2} = \left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \mu_0)^2}\right)^{n/2} \tag{3.31}$$

If we write

$$\begin{aligned}(X_i - \mu_0)^2 &= [(X_i - \overline{X}) + (\overline{X} - \mu_0)]^2 \\ &= (X_i - \overline{X})^2 + (\overline{X} - \mu_0)^2 + 2(X_i - \overline{X})(\overline{X} - \mu_0)\end{aligned} \tag{3.32}$$

we can see that

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2 + 2\underbrace{\left(\sum_{i=1}^{n} X_i - n\overline{X}\right)}_{0}(\overline{X} - \mu_0) \tag{3.33}$$

so

$$\Lambda(\mathbf{X}) = \left(1 + \frac{n(\overline{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)^{-n/2} \tag{3.34}$$

This means that thresholding on the value of $\Lambda(\mathbf{X})$ is the same as thresholding on the second term in parentheses, or equivalently on

$$T = \frac{\overline{X} - \mu_0}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2/n}} \tag{3.35}$$

But we've constructed this last statistic so that, if $\mathcal{H}_0$ holds and the $\{X_i\}$ are independent $N(\mu_0, \theta_2)$ random variables, Student's theorem tells is this is t-distributed with $n-1$ degrees of freedom. So for a test at significance level α we reject $\mathcal{H}_0$ if $T > t_{n-1,\frac{\alpha}{2}}$ or $T < -t_{n-1,\frac{\alpha}{2}}$, since

$$\Lambda(\mathbf{X}) = \left(1 + \frac{T^2}{n-1}\right)^{-n/2} \tag{3.36}$$

So we reject $\mathcal{H}_0$ if

$$\Lambda(\mathbf{X}) < \left(1 + \frac{t_{n-1,\frac{\alpha}{2}}^2}{n-1}\right)^{-n/2} \tag{3.37}$$

3.3.2 Example: Mean of Multivariate Normal Distribution with Unit Variance

We know that the behavior of maximum likelihood tests is simplest when the sample is drawn from a normal distribution with known variance whose mean is the parameter in question. Now we have multiple parameters, so the straightforward extension is to have the sample drawn from a multivariate normal distribution whose means are the parameters. So the sample now consists of n independent random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$, each with p elements and drawn from a multivariate normal distribution. For simplicity, we'll use the identity matrix as the variance-covariance matrix, so $\mathbf{X}_i \sim N_p(\boldsymbol{\theta}, \mathbf{1}_{p\times p})$. We'll define the null hypothesis $\mathcal{H}_0$ to be that the first q of the $\{\theta_\alpha\}$ are zero. The

likelihood function is

$$L(\boldsymbol{\theta};\mathbf{x}) = (2\pi)^{-n/2}\exp\left(h - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{x}_i-\boldsymbol{\theta})\right)$$
$$= (2\pi)^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\sum_{\alpha=1}^{p}([\mathbf{x}_i]_\alpha-\theta_\alpha)^2\right) \tag{3.38}$$
$$\propto \exp\left(-\frac{n}{2}\sum_{\alpha=1}^{p}(\theta_\alpha-\overline{x}_\alpha)^2\right)$$

where

$$\overline{x}_\alpha = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{x}_i]_\alpha \tag{3.39}$$

Thus the maximum likelihood solution under $\mathcal{H}_1$ is $\widehat{\theta}_\alpha = \overline{x}_\alpha$, whike under $\mathcal{H}_0$ it is

$$[\widehat{\boldsymbol{\theta}}_0]_\alpha = \begin{cases} 0 & \alpha = 1,\ldots,q \\ \overline{x}_\alpha & \alpha = q+1,\ldots,p \end{cases} \tag{3.40}$$

and the likelihood ratio is

$$\Lambda(\{\mathbf{x}_i\}) = \frac{L(\widehat{\boldsymbol{\theta}}_0;\mathbf{x})}{L(\widehat{\boldsymbol{\theta}};\mathbf{x})} = \exp\left(-\frac{n}{2}\sum_{\alpha=1}^{q}\overline{x}_\alpha^2\right) \tag{3.41}$$

Under $\mathcal{H}_0$, the $\{\overline{X}_\alpha\}$ are independent and $\overline{X}_\alpha \sim N(0,\frac{1}{n})$ for $\alpha = 1,\ldots,q$. Thus, if the null hypothesis is true,

$$-2\ln\Lambda(\{\mathbf{X}_i\}) = \sum_{\alpha=1}^{q}\frac{\overline{X}_\alpha{}^2}{1/n} \sim \chi^2(q) \tag{3.42}$$

3.3.3 Large-Sample Limit

This last result also holds generically as a limiting distribution for large samples $n \to \infty$:

$$-2\ln\Lambda(\{\mathbf{X}_i\}) \xrightarrow{D} \chi^2(q) \tag{3.43}$$