Frequentist parameter estimation

Kevin P. Murphy

Last updated October 1, 2007

1 Introduction

Whereas probability is concerned with describing the relative likelihoods of generating various kinds of data, statistics is concerned with the opposite problem: inferring the causes that generated the observed data. Indeed, statistics used to be known as **inverse probability**.

As mentioned earlier, there are two interpretations of probability: the frequentist interpretation in terms of long term frequencies of future events, and the Bayesian interpretation, in terms of modeling (subjective) uncertainty given the current data. This in turn gives rise to two approaches to statistics. The **frequentist approach** to statistics is the most widely used, and hence is sometimes called the **orthodox approach** or **classical approach**. We will give a very brief introduction here. For more details, consult one of the many excellent textbooks on this topic, such as [Ric95] or [Was04].

2 Point estimation

Point estimation refers to computing a single "best guess" of some quantity of interest from data. The quantity could be a parameter in a parametric model (such as the mean of a Gaussian), or a regression function, or a prediction of a future value of some random variable. We assume there is some "true" value for this quantity, which is fixed but unknown, call it θ. Our goal is to construct an **estimator**, which is some function g that takes sample data, $\mathcal{D} = (X_1, \ldots, X_N)$, and returns a point estimate $\hat{\theta}_N$:

$$\hat{\theta}_N = g(X_1, \ldots, X_N) \tag{1}$$

Since $\hat{\theta}_N$ depends on the particular observed data, $\hat{\theta} = \hat{\theta}(\mathcal{D})$, it is a random variable. (Often we omit the subscript N and just write $\hat{\theta}$.)

An example of an estimator is the **sample mean** of the data:

$$\hat{\mu} = \overline{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2}$$

Another is the **sample variance**:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x}) \tag{3}$$

A third example is the empirical fraction of heads in a sequence of heads (1s) and tails (0s):

$$\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{4}$$

where $X_n \sim Be(\theta)$.

There are many possible estimators, so how do we know which to use? Below we describe some of the most desirable properties of an estimator.

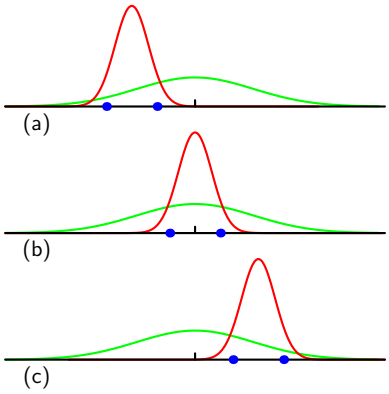

Figure 1: Graphical illustration of why $\hat{\sigma}^2_{ML}$ is biased: it underestimates the true variance because it measures spread around the empirical mean $\hat{\mu}_{ML}$ instead of around the true mean. Source: [Bis06] Figure 1.15.

2.1 Unbiased estimators

We define the **bias** of an estimator as

$$\text{bias}(\hat{\theta}_N) = E_{\mathcal{D}\sim\theta}(\hat{\theta}_N - \theta) \tag{5}$$

where the expectation is over data sets $\mathcal{D}$ drawn from a distribution with parameter θ. We say that $\hat{\theta}_N$ is **unbiased** if $E_\theta(\hat{\theta}_N) = \theta$. It is easy to see that $\hat{\mu}$ is an unbiased estimator:

$$E\hat{\mu} = E\frac{1}{N}\sum_{n=1}^{N} X_n = \frac{1}{N}\sum_n E[X_n] = \frac{1}{N}N\mu \tag{6}$$

However, one can show (exercise) that

$$E\hat{\sigma}^2 = \frac{N-1}{N}\sigma^2 \tag{7}$$

Therefore it is common to use the following unbiased estimator of the variance:

$$\hat{\sigma}^2_{N-1} = \frac{N}{N-1}\hat{\sigma}^2 \tag{8}$$

In Matlab, `var(X)` returns $\hat{\sigma}^2_{N-1}$ whereas `var(X,1)` returns $\hat{\sigma}^2$.

One might ask: why is σ^2_{ML} biased? Intuitively, we "used up" one "degree of freedom" in estimating μ_{ML}, so we underestimate σ. (If we used μ instead of μ_{ML} when computing σ^2_{ML}, the result would be unbiased.) See Figure 1.

2.2 Bias-variance tradeoff

Being unbiased seems like a good thing, but it turns out that a little bit of bias can be useful so long as it reduces the variance of the estimator. In particular, suppose our goal is to minimize the **mean squared error** (MSE)

$$MSE = E_\theta(\hat{\theta}_N - \theta)^2 \tag{9}$$

It turns out that when minimizing MSE, there is a **bias-variance tradeoff**.

Theorem 2.1. *The MSE can be written as*

$$MSE = bias^2(\hat{\theta}) + Var_\theta(\hat{\theta}) \tag{10}$$

Proof. Let $\overline{\theta} = E_{\mathcal{D}\sim\theta}(\hat{\theta}(\mathcal{D}))$. Then

$$\begin{aligned} E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \theta)^2 &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta} + \overline{\theta} - \theta)^2 && (11)\\ &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + 2(\overline{\theta} - \theta)E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta}) + (\overline{\theta} - \theta)^2 && (12)\\ &= E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta})^2 + (\overline{\theta} - \theta)^2 && (13)\\ &= V(\hat{\theta}) + \text{bias}^2(\hat{\theta}) && (14) \end{aligned}$$

where we have used the fact that $E_{\mathcal{D}}(\hat{\theta}(\mathcal{D}) - \overline{\theta}) = \overline{\theta} - \overline{\theta} = 0$. □

Later we will see that simple models are often biased (because they cannot represent the truth) but have low variance, whereas more complex models have lower bias but higher variance. **Bagging** is a simple way to reduce the variance of an estimator without increasing the bias: simply take weighted combinations of estimatros fit on different subsets of the data, chosen randomly with replacement.

2.3 Consistent estimators

Having low MSE is good, but is not enough. We would also like our estimator to converge to the true value as we collect more data. We call such an estimator **consistent**. Formally, $\hat{\theta}$ is consistent if $\hat{\theta}_N$ converges in probability to θ as $N \to \infty$. $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\sigma}^2_{N-1}$ are all consistent.

3 The method of moments

The k'th moment of a distribution is

$$\mu_k = E[X^k] \tag{15}$$

The k'th **sample moment** is

$$\hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{N} X_n^k \tag{16}$$

The **method of moments** is simply to equate $\mu_k = \hat{\mu_k}$ for the first few moments (the minimum number necessary) and then to solve for θ. Although these estimators are not optimal (in a sense to be defined later), they are consistent, and are simple to compute, so they can be used to initialize other methods that require iterative numerical routines. Below we give some examples.

3.1 Bernoulli

Since $\mu_1 = E(X) = \theta$, and $\hat{\mu}_1 = \frac{1}{N}\sum_{n=1}^{N} X_n$, we have

$$\hat{\theta} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{17}$$

3.2 Univariate Gaussian

The first and second moments are

$$\begin{aligned} \mu_1 &= E[X] = \mu && (18)\\ \mu_2 &= E[X^2] = \mu^2 + \sigma^2 && (19) \end{aligned}$$

So $\sigma^2 = \mu_2 - \mu_1^2$. The corresponding estimates from the sample moments are

$$\begin{aligned} \hat{\mu} &= \overline{X} && (20)\\ \hat{\sigma}^2 &= \frac{1}{N}\sum_{n=1}^{N} x_n^2 - \overline{X}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{X})^2 && (21) \end{aligned}$$

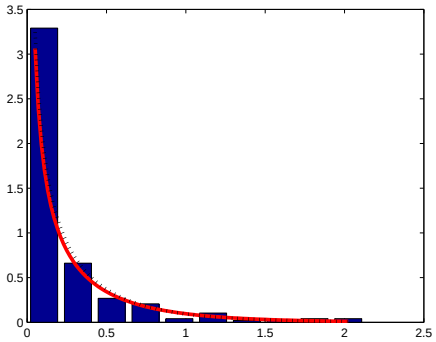

Figure 2: An empirical pdf of some rainfall data, with two Gamma distributions superimposes. Solid red line = method of moment. dotted black line = MLE. Figure generated by `rainfallDemo`.

3.3 Gamma distribution

Let $X \sim Ga(a, b)$. The first and second moments are

$$\mu_1 = \frac{a}{b} \tag{22}$$

$$\mu_2 = \frac{a(a+1)}{b^2} \tag{23}$$

To apply the method of moments, we must express a and b in terms of μ_1 and μ_2. From the second equation

$$\mu_2 = \mu_1^2 + \frac{\mu_1}{b} \tag{24}$$

or

$$b = \frac{\mu_1}{\mu_2 - \mu_1^2} \tag{25}$$

Also

$$a = b\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2} \tag{26}$$

Since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$, we have

$$\hat{b} = \frac{\overline{x}}{\hat{\sigma}^2}, \hat{a} = \frac{\overline{x}^2}{\hat{\sigma}^2} \tag{27}$$

As an example, let us consider a data set from the book by [Ric95, p250]. The data records the average amount of rain (in inches) in southern Illinois during each storm over the years 1960 - 1964. If we plot its empirical pdf as a histogram, we get the result in Figure 2. This is well fit by a Gamma distribution, as shown by the superimposed lines. The solid line is the pdf with parameters estimated by the method of moments, and the dotted line is the pdf with parameteres estimated by maximum likelihood. Obviously the fit is very similar, even though the parameters are slightly different numerically:

$$\hat{a}_{mom} = 0.3763, \hat{b}_{mom} = 1.6768, \hat{a}_{mle} = 0.4408, \hat{b}_{mle} = 1.9644 \tag{28}$$

This was implemented using `rainfallDemo`.

4 Maximum likelihood estimates

A **maximum likelihood estimate** (MLE) is a setting of the parameters θ that makes the data as likely as possible:

$$\hat{\theta}_{mle} = \arg\max_{\theta} p(D|\theta) \tag{29}$$

Since the data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is iid, the likelihood factorizes

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{30}$$

It is often more convenient to work with log probabilities; this will not change $\arg\max L(\boldsymbol{\theta})$, since log is a monotonic function. Hence we define the **log likelihood** as $\ell(\boldsymbol{\theta}) = \log p(\mathcal{D}|\boldsymbol{\theta})$. For iid data this becomes

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{31}$$

The mle then maximizes $\ell(\boldsymbol{\theta})$.

MLE enjoys various theoretical properties, such as being consistent and **asymptotically efficient**, which means that (roughly speaking) the MLE has the smallest variance of all well-behaved estimators (see [Was04, p126] for details). Therefore we will use this technique quite widely. We consider several examples below that we will use later.

4.1 Univariate Gaussians

Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$p(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{32}$$

$$\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \tag{33}$$

To find the maximum, we set the partial derivatives to 0 and solve. Starting with the mean, we have

$$\frac{\partial \ell}{\partial \mu} = -\frac{2}{2\sigma^2}\sum_{n}(x_n - \mu) = 0 \tag{34}$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=n}^{N} x_n \tag{35}$$

which is just the empirical mean. Similarly,

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{1}{2}\sigma^{-4}\sum_{n}(x_n - \hat{\mu}) - \frac{N}{2\sigma^2} = 0 \tag{36}$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2 \tag{37}$$

$$= \frac{1}{N}\left[\sum_{n} x_n^2 + \sum_{n}\hat{\mu}^2 - 2\sum_{n} x_n\hat{\mu}\right] \tag{38}$$

$$= \frac{1}{N}[\sum_{n} x_n^2 + N\hat{\mu}^2 - 2N\hat{\mu}^2] = \frac{1}{N}[\sum_{n} x_n^2 + N(\frac{1}{N}\sum_{n} x_n)^2 - 2N(\frac{1}{N}\sum_{n} x_n)^2] \tag{39}$$

$$= \frac{1}{N}\sum_{n} x_n^2 - (\frac{1}{N}\sum_{n} x_n)^2 - = \frac{1}{N}\sum_{n} x_n^2 - (\hat{\mu})^2 \tag{40}$$

since $\sum_n x_n = N\hat{\mu}$. This is just the empirical variance.

4.2 Bernoullis

Let $X \in \{0, 1\}$. Given $\mathcal{D} = (x_1, \ldots, x_N)$, the likelihood is

$$p(D|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \tag{41}$$

$$= \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} \tag{42}$$

$$= \theta^{N_1}(1-\theta)^{N_2} \tag{43}$$

where $N_1 = \sum_i x_i$ is the number of heads and $N_2 = \sum_i (1 - x_i)$ is the number of tails The log-likelihood is

$$L(\theta) = \log p(D|\theta) = N_1 \log \theta + N_2 \log(1-\theta) \tag{44}$$

Solving for $\frac{dL}{d\theta} = 0$ yields

$$\theta_{ML} = \frac{N_1}{N} \tag{45}$$

the empirical fraction of heads.

Suppose we have seen 3 tails out of 3 trials. Then we predict that the probability of heads is zero:

$$\theta_{ML} = \frac{N_1}{N_1 + N_2} = \frac{0}{0+3} \tag{46}$$

This is an example of the **sparse data problem**: if we fail to see something in the training set, we predict that it can never happen in the future. Later we will consider Bayesian and MAP point estimates, which avoid this problem.

4.3 Multinomials

The log-likelihood is

$$\ell(\theta; D) = \log p(D|\theta) = \textstyle\sum_k N_k \log \theta_k \tag{47}$$

We need to maximize this subject to the constraint $\sum_k \theta_k = 1$, so we use a **Lagrange multiplier**. The constrained cost function becomes

$$\tilde{\ell} = \sum_k N_k \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right) \tag{48}$$

Taking derivatives wrt θ_k yields

$$\frac{\partial \tilde{\ell}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0 \tag{49}$$

Taking derivatives wrt λ yields the original constraint:

$$\frac{\partial \tilde{\ell}}{\partial \lambda} = \left(1 - \sum_k \theta_k\right) = 0 \tag{50}$$

Using this sum-to-one constraint we have

$$N_k = \lambda \theta_k \tag{51}$$

$$\sum_k N_k = \lambda \sum_k \theta_k \tag{52}$$

$$N = \lambda \tag{53}$$

$$\hat{\theta}_k = \frac{N_k}{N} \tag{54}$$

Hence $\hat{\theta}_k$ is the fraction of times k occurs. If we did not observe $X = k$ in the training data, we set $\hat{\theta}_k = 0$, so we have the same sparse data problem as in the Bernoulli case (in fact it is worse, since K may be large, so it is quite likely that we didn't see some of the symbols, especially if our data set is small).

4.4 Gamma distribution

The pdf for $Ga(a, b)$ is

$$Ga(x|a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} \tag{55}$$

So the log likelihood is

$$\begin{aligned} \ell(a, b) &= \sum_n [a \log b + (a-1) \log x_n - b x_n - \log \Gamma(a)] &(56)\\ &= N a \log b + (a-1) \sum_n \log x_n - b \sum_n x_n - N \log \Gamma(a) &(57)\end{aligned}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial a} &= N \log b + \sum_n \log x_n - N \frac{\Gamma'(a)}{\Gamma(a)} &(58)\\ \frac{\partial \ell}{\partial b} &= \frac{Na}{b} - \sum_n x_n &(59)\end{aligned}$$

where

$$\frac{\partial}{\partial a} \log \Gamma(a) \stackrel{\text{def}}{=} \psi(a) = \frac{\Gamma'(a)}{\Gamma(a)} \tag{60}$$

is the **digamma** function (which in matlab is called `psi`). Setting $\frac{\partial \ell}{\partial b} = 0$ we fund

$$\hat{b} = \frac{N \hat{a}}{\sum_n x_n} = \frac{\hat{a}}{\overline{x}} \tag{61}$$

But when we substitute this in to $\frac{\partial \ell}{\partial a} = 0$ we get a nonlinear equation for a:

$$0 = N \log \hat{a} - N \log \overline{x} + \sum_n \log x_n - N \psi(a) \tag{62}$$

This equation cannot be solved in closed form; an iterative method for finding the roots (such as Newton's method) must be used. We can start this process from the method of moments estimate.

In Matlab, if you type `type(which('gamfit'))`, you can look at its source code, and you will find that it estimates a by calling `fzero` with the following function:

$$lkeqn(a) = -\frac{1}{N} \sum_n \log X_n - \log \overline{X} - \log(a) + \psi(a) \tag{63}$$

It then substitutes $\hat{a}$ into Equation 61.

4.5 Why maximum likelihood?

Recall that the KL divergence between "true" distribution p and approximation q is defined as

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \text{const} - \sum_x p(x) \log q(x) \tag{64}$$

where the constant term only depends on p (which is fixed) and not q (which needs to be estimated). If we drop the constant term, the result is called the **cross entropy**. Now suppose p is the **empirical distribution**, which puts a probability atom on the observed training data and zero mass everywhere else:

$$p_{emp}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \tag{65}$$

Now we get

$$KL(p_{emp}||q) = \text{const} - \sum_x p(x) \log q(x) = \text{const} - \frac{1}{n} \sum_i \log q(x_i) \tag{66}$$

This is just the average negative log likelihood of q on the training set. So minimizing KL divergence to the empirical distribution is equivalent to maximizing likelihood.

5 Sampling distributions

In addition to a point estimate, it is useful to have some measure of uncertainty. Note that the frequentist notion of uncertainty is quite different from the Bayesian. In the frequentist view, uncertainty means: how much would my estimate change if I had different data? This is called the **sampling distribution** of the estimator. In the Bayesian view, uncertainty means: how much do I believe my estimate given the current data? This is called the **posterior distribution** of the parameter. In otherwords, the frequentist is concerned with $E_{\mathcal{D}}[\hat{\theta}|\mathcal{D}]$ (and its spread), whereas the Bayesian is concerned with $E_\theta[\theta|\mathcal{D}]$ (and its spread). Sometimes these give the same answers, but not always.

The distribution of $\hat{\theta}$ is called the sampling distribution. Its standard deviation is called the **standard error**:

$$se(\hat{\theta}) = \sqrt{\text{Var } (\hat{\theta})} \tag{67}$$

Often the standard error depends on the unknown distribution. In such cases, we often estimate it; we denote such estimates by $\hat{se}$.

Later we will see that many sampling distributions are approximately Gaussian as the sample size goes to infinity. More precisely, we say an estimator is **asymptotically Normal** if

$$\frac{\hat{\theta}_N - \theta}{se} \rightsquigarrow \mathcal{N}(0, 1) \tag{68}$$

(where $\rightsquigarrow$ here means converges in distribution).

5.1 Confidence intervals

A $1 - \alpha$ **confidence interval** is an interval $C_n = (a, b)$ where a and b are functions of the data $X_{1:N}$ such that

$$P_\theta(\theta \in C_N) \geq 1 - \alpha \tag{69}$$

In other words, (a, b) traps θ with probability $1 - \alpha$. Often people use 95% confidence intervals, which corresponds to $\alpha = 0.05$.

If the estimator is asymptotically normal, we can use properties of the Gaussian distribution to compute an approximate confidence interval.

Theorem 5.1. *Suppose that $\hat{\theta}_N \approx \mathcal{N}(\theta, \hat{se}^2)$. Let Φ be the cdf of a standard Normal, $Z \sim \mathcal{N}(0, 1)$, and let $z_{\alpha/2} = \Phi^{-1}(1 - (\frac{\alpha}{2}))$, so that $P(Z > z_{\alpha/2}) = \alpha/2$. By symmetry of the Gaussian, $P(Z < -z_{\alpha/2}) = \alpha/2$, so $P(-z_{\alpha/2} < Z < z_{\alpha/a}) = 1 - \alpha$. (see Figure 3). Let*

$$C_N = (\hat{\theta}_N - z_{\alpha/2}\hat{se}, \hat{\theta}_N + z_{\alpha/2}\hat{se}) \tag{70}$$

Then

$$P(\theta \in C_N) \rightarrow 1 - \alpha \tag{71}$$

Proof. Let $Z_N = (\hat{\theta} - \theta)/\hat{se}$. By assumption, $Z_n \rightsquigarrow Z$. Hence

$$\begin{aligned} P(\theta \in C_N) &= P(\hat{\theta}_N - z_{\alpha/2}\hat{se} < \theta < \hat{\theta}_N + z_{\alpha/2}\hat{se}) && (72)\\ &= P(z_{\alpha/2}\hat{se} < \frac{\hat{\theta_N} - \theta}{\hat{se}} < z_{\alpha/2}\hat{se}) && (73)\\ &\rightarrow P(z_{\alpha/2}\hat{se} < Z < z_{\alpha/2}\hat{se}) && (74)\\ &= 1 - \alpha && (75) \end{aligned}$$

□

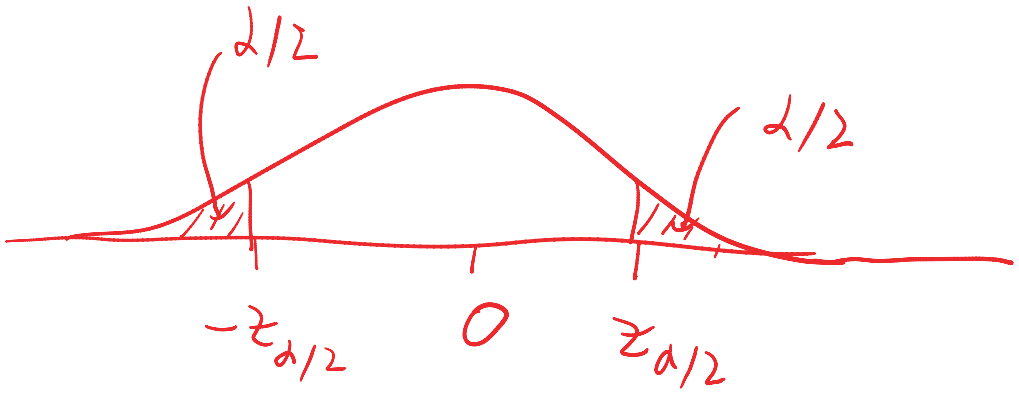

Figure 3: A $\mathcal{N}(0,1)$ distribution with the $z_{\alpha/2}$ cutoff points shown. The central non shaded area contains $1-\alpha$ of the probability mass. If $\alpha = 0.05$, then $z_{\alpha/2} = 1.96 \approx 2$.

For 95% confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$, which is why people often express their results as $\hat{\theta}_N \pm 2\hat{se}$.

5.2 The counter-intuitive nature of confidence intervals

Note that saying that "$\hat{\theta}$ has a 95% confidence interval of $[a,b]$" does *not* mean $P(\hat{\theta} \in [a,b]|\mathcal{D}) = 0.95$. Rather, it means

$$p_{D' \sim P(\cdot|\theta)}(\hat{\theta} \in [a(D'), b(D')]) = 0.95 \tag{76}$$

i.e., if we were to repeat the experiment, then 95% of the time, the true parameter would be in the $[a,b]$ interval. To see that these are not the same statement, consider this example (from [Mac03, p465]). Suppose we draw two integers from

$$p(x|\theta) = \begin{cases} 0.5 & \text{if } x = \theta \\ 0.5 & \text{if } x = \theta + 1 \\ 0 & \text{otherwise} \end{cases} \tag{77}$$

If $\theta = 39$, we would expect the following outcomes each with prob 0.25:

$$(39,39), (39,40), (40,39), (40,40) \tag{78}$$

Let $m = min(x_1, x_2)$ and define a CI as

$$[a(D), b(D)] = [m, m], \tag{79}$$

For the above samples this yields

$$[39,39], \quad [39,39], \quad [39,39], \quad [40,40] \tag{80}$$

which is clearly a 75% CI. However, if $D = (39,39)$ then $p(\theta = 39|D) = P(\theta = 38|D) = 0.5$. And if $D = (39,40)$ then $p(\theta = 39|D) = 1.0$. Thus even if we know $\theta = 39$, we only have 75% "confidence" in this fact. Later we will see that Bayesian **credible intervals** give the more intuitively correct answer.

5.3 Sampling distribution for Bernoulli MLE

Consider estimating the parameter of a Bernoulli using $\hat{\theta} = \frac{1}{N}\sum_n X_n$. Since $X_n \sim Be(\theta)$, we have $S = \sum_n X_n \sim Binom(N, \theta)$. So the sampling distribution is

$$p(\hat{\theta}) = p(S = N\hat{\theta}) = \text{Binom}(N\hat{\theta}|N, \theta) \tag{81}$$

We can compute the mean and variance of this distribution as follows.

$$E\hat{\theta} = \frac{1}{N}E[S] = \frac{1}{N}N\theta = \theta \tag{82}$$

so we see this is an unbiased estimator. Also

$$\begin{aligned} \text{Var}\ \hat{\theta} &= \text{Var}\ [\frac{1}{N}\sum_n X_n] && (83)\\ &= \frac{1}{N^2}\sum_n \text{Var}\ [X_n] && (84)\\ &= \frac{1}{N^2}\sum_n \theta(1-\theta) && (85)\\ &= \frac{\theta(1-\theta)}{N} && (86) \end{aligned}$$

So

$$se = \sqrt{\theta(1-\theta)/N} \tag{87}$$

and

$$\hat{se} = \sqrt{\hat{\theta}(1-\hat{\theta})/N} \tag{88}$$

We can compute an exact confidence interval using quantiles of the Binomial distribution. However, for reasonably small N, the Binomial is well approximated by a Gaussian, so $\hat{\theta}_N \approx \mathcal{N}(\theta, \hat{se}^2)$. So an approximate $1-\alpha$ confidence interval is

$$\hat{\theta}_N \pm z_{\alpha/2}\hat{se} \tag{89}$$

5.4 Large sample theory for the MLE

Computing the exact sampling distribution of an estimator can often be difficult. Fortunately, it can be shown that, for certain models[1], as the sample size tends to infinity, the sampling distribution becomes Gaussian. We say the estimator is **asymptotically Normal**.

Define the **score function** as the derivative of the log likelihood

$$s(X,\theta) = \frac{\partial \log p(X|\theta)}{\partial \theta} \tag{90}$$

Define the **Fisher information** to be

$$\begin{aligned} I_N(\theta) &= \text{Var}\ \left(\sum_{n=1}^{N} s(X_n,\theta)\right) && (91)\\ &= \sum_n \text{Var}\ (s(X_n,\theta)) && (92) \end{aligned}$$

Intuitively, this measures the stability of the MLE wrt variations in the data set (recall that X is random and θ is fixed). It can be shown that $I_N(\theta) = NI_1(\theta)$. We will write $I(\theta)$ for $I_1(\theta)$. We can rewrite this in terms of the second derivative, which measures the curvature of the likelihood.

Theorem 5.2. *Under suitable smoothness assumptions on p, we have*

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log p(X|\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2}\log p(X|\theta)\right] \tag{93}$$

We can now state our main theorem.

Theorem 5.3. *Under appropriate regularity conditions, the MLE is asymptotically Normal.*

$$\hat{\theta}_N \sim \mathcal{N}(\theta, se^2) \tag{94}$$

[1] Intuively, the requirement is that each parameter in the model get to "see" an infinite amount of data.

where

$$se \approx \sqrt{1/I_N(\theta)} \tag{95}$$

is the standard error of $\hat{\theta}_N$*. Furthemore, this result still holds if we use the estimated standard error*

$$\hat{se} \approx \sqrt{1/I_N(\hat{\theta})} \tag{96}$$

The proof can be found in standard textbooks, such as [Ric95]. The basic idea is to use a Taylor series expansion of ℓ' around θ.

The intuition behind this result is as follows. The **asymptotic variance** is given by

$$\frac{1}{NI(\theta)} = -\frac{1}{E\ell''(\theta)} \tag{97}$$

so when the curvature at the MLE $|\ell''(\hat{\theta})|$, is large, then the variance is low, whereas if the curvature is nearly flat, the variance is high. (Note that $\ell''(\hat{\theta}) < 0$ since $\hat{\theta}$ is a maximum of the log likelihood.) Intuitively, the curvature is large if the parameter is "well determined".[2]

For example, consider $X_n \sim Be(\theta)$. The MLE is $\hat{\theta} = \frac{1}{N}\sum_n X_n$ and the log likelihood is

$$\log p(x|\theta) = x \log \theta + (1-x)\log(1-\theta) \tag{98}$$

so the score function is

$$s(X,\theta) = \frac{X}{\theta} - \frac{1-X}{1-\theta} \tag{99}$$

and

$$-s'(X,\theta) = \frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2} \tag{100}$$

Hence

$$I(\theta) = E_\theta(-s'(X,\theta)) = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} \tag{101}$$

So

$$\hat{se} = \frac{1}{\sqrt{I_N(\hat{\theta}_N)}} = \frac{1}{\sqrt{NI(\hat{\theta}_N)}} = \left(\frac{\hat{\theta}(1-\hat{\theta})}{N}\right)^{\frac{1}{2}} \tag{102}$$

which is the same as the result we derived in Equation 88.

In the multivariate case, the **Fisher information matrix** is defined as

$$I_N(\theta) = -\begin{pmatrix} E_\theta H_{11} & \cdots & E_\theta H_{1p} \\ \ddots & & \\ E_\theta H_{p1} & \cdots & E_\theta H_{pp} \end{pmatrix} \tag{103}$$

where

$$H_{jk} = \frac{\partial^2 \ell_N}{\partial\theta_j \partial\theta_k} \tag{104}$$

is the Hessian of the log likelihood. The multivariate version of Theorem 5.3 is

$$\hat{\theta} \sim \mathcal{N}(\theta, I_N^{-1}(\theta)) \tag{105}$$

References

[Bis06] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Mac03] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Ric95] J. Rice. *Mathematical statistics and data analysis*. Duxbury, 1995. 2nd edition.

[Was04] L. Wasserman. *All of statistics. A concise course in statistical inference*. Springer, 2004.

[2] From the Bayesian standpoint, the equivalent statement is that the parameter is well determined if the posterior uncertainty is small. (Sharply peaked Gaussians have lower entropy than flat ones.) This is arguably a much more natural intepretation, since it talks about our uncertainty about θ, rather than variance induced by changing the data.