

Introduction to Statistical Inference

Textbook Notes

Summer, 2021

Contents

1	Estimation	1
1.1	Models and Likelihood	1
1.2	Statistics, Estimands, Estimators, and Estimates	3
1.3	Method of Moments	3
2	Quality of Estimation	4
2.1	Loss Functions	4
2.2	Bias-Variance Tradeoff	4
2.3	Consistency of Estimators	4
3	Maximum Likelihood Estimation	5
3.1	Properties of the MLE	6
3.2	Kullback-Leibler Divergence	7
3.3	Fisher Information	7
4	Bayesian Statistical Inference	8
4.1	Prior to Posterior	8

PROFESSOR:
Joe Blitzstein

SCRIBE:
Matthew Nazari*

1 Estimation

1.1 Models and Likelihood

Definition 1.1 (statistical model). A statistical model is a family of probability distributions indexed by a parameter $\theta \in \Theta$. The *parameter space* Θ is the set of all allowable parameter values.

def. statistical model

The family of all normal distributions, $\{\mathcal{N}(\mu, \sigma^2) : \mu, \sigma \in \mathbb{R}, \sigma > 0\}$, is a 2-dimensional parametric model where $\Theta = \{\theta : \theta \in \mathbb{R} \times \mathbb{R}^+\}$. But models don't have to be named distributions.

Example 1.1 Suppose we have a population of cats and dogs, and we are studying the weight Y of a random animal from the population. Let p be the proportion of cats in the population. Suppose the weight of a random cat and a random dog is $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ respectively. Then by the LOTP, the CDF of Y is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(Y \leq y \mid \text{cat}) P(\text{cat}) + P(Y \leq y \mid \text{dog}) P(\text{dog}) \\ &= p \Phi\left(\frac{y - \mu_1}{\sigma_1}\right) + (1 - p) \Phi\left(\frac{y - \mu_2}{\sigma_2}\right). \end{aligned}$$

The CDF of Y depends on the parameter $\theta = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)$, so we write it as $F_Y(y \mid \theta)$. Our model, therefore, is the collection of all CDFs $F_Y(y \mid \theta)$ indexed by $\theta \in \Theta = \{\theta : \theta \in [0, 1] \times (\mathbb{R} \times \mathbb{R}^+)^2\}$. This is a 5-dimensional parametric model.

*matthewnazari@college (email)

We say our *data* is the realization y_1, \dots, y_n of the random variables Y_1, \dots, Y_n . These random variables \mathbf{Y} have some unknown, but true, distribution $F_Y(\cdot | \theta)$ parametrized by θ . Often times it is plausible to assume Y_1, \dots, Y_n are i.i.d: $Y_j \stackrel{\text{i.i.d.}}{\sim} F_Y(y_j | \theta)$. Then,

$$F_Y(\mathbf{y} | \theta) = \prod_{j=1}^n F_{Y_j}(y_j | \theta).$$

Definition 1.2 (likelihood function). Suppose we observe \mathbf{y} to be the value of \mathbf{Y} . The function $L(\theta; \mathbf{y}) = f_Y(\mathbf{y} | \theta)$ is the *likelihood function*. It is a function of θ and fixed with \mathbf{y} . The idea here is that we can now compare various candidate distributions on how consistent they are with the observed data \mathbf{y} . If $L(\theta_1; \mathbf{y}) > L(\theta_2; \mathbf{y})$, then the data seems more consistent with θ_1 than θ_2 .

def. likelihood function

We cannot conclude that one value for θ is more likely than another yet since we have not imposed a probability distribution on θ .

- In the *frequentist perspective*, θ is regarded as fixed but unknown and it does not have a distribution. One way to approximate θ , then, is to use maximum likelihood estimation (MLE) which says estimate θ using

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{y}).$$

- In the *Bayesian perspective*, θ does have a distribution. That is, θ is an r.v. rather than being fixed but unknown. This allows us to make probability statements about various θ values when accompanied with further knowledge like prior information. We have a prior density $\pi(\theta)$ for θ and obtain the posterior density with Bayes' Rule:

$$\pi(\theta | \mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y} | \theta)}{f(\mathbf{y})} \propto \pi(\theta)f(\mathbf{y} | \theta) = L(\theta; \mathbf{y})\pi(\theta).$$

This means *the posterior is proportional to the likelihood times prior*.

Example 1.2 Let $Y \sim \text{Bin}(n, p)$ with $n = 3$ and p unknown. Suppose we observe that $Y = 2$. The likelihood function is then $L(p) = \binom{3}{2}p^2(1-p)$. We don't care about constant scaling since rescaling has no effect on MLE and likelihood ratios. Thus, $L(p) = p^2(1-p)$ and $L(p) = 12p^2(1-p)$ are equally valid ways to express the likelihood function. Note that in a Bayesian approach with the prior $p \sim \text{Unif}(0, 1)$, the posterior is the likelihood.

Definition 1.3 (log-likelihood function). It is common and mathematically convenient to work with the *log-likelihood function* $l(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y})$ instead of the likelihood function. If the Y_j are i.i.d then

def. log-likelihood function

$$l(\theta; \mathbf{y}) = \sum_{j=1}^n \log f_{Y_j}(y_j | \theta).$$

Theorem 1.1 (invariance property of likelihood). The likelihood function is unchanged under reparameterization: namely, consider a likelihood function $L(\theta; \mathbf{y})$ and let $\psi = g(\theta)$ be a reparameterization, where g is injective. Then

$$L(\psi; \mathbf{y}) = L(\theta; \mathbf{y}).$$

1.2 Statistics, Estimands, Estimators, and Estimates

Definition 1.4 (statistic). A *statistic* $T(\mathbf{Y})$ is a function of Y_1, \dots, Y_n (and possibly other known quantities). The function T cannot depend on unknown parameters.

def. statistic

Example 1.3 Let our data be the realization of the random variables Y_1, \dots, Y_n where $Y_j \stackrel{\text{i.i.d.}}{\sim} F_{Y_j}(y_j | \theta)$. The sample mean is a very common statistic:

$$T(\mathbf{Y}) = \bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j.$$

Notice that we can compute \bar{Y} by knowing Y_1, \dots, Y_n , but \bar{Y} still does not depend on the unknown parameter θ . The distribution of \bar{Y} , however, can and does depend on θ .

Definition 1.5 (estimand). An *estimand* is an object that we wish to learn about the data.

def. estimand

Example 1.4 For each of n individuals we observe two variables. The resulting data are the i.i.d. pairs $(X_1, Y_1), \dots, (X_n, Y_n)$. Examples of estimands include $E[Y_1]$, $\text{Var}(Y_1)$, the CDF $F_{Y_1}(y)$, $\text{Corr}(X_1, Y_1)$, and the probability $P(\tilde{X} < \tilde{Y})$.

Definition 1.6 (estimator). A statistic $\hat{\theta} = T(\mathbf{Y})$ constructed with the intention to approximate an estimand θ is called an *estimator*. All estimators are statistics and therefore r.v.s.

def. estimator

Definition 1.7 (bias). The *bias* of an estimator $\hat{\theta}$ measures on average how far off an estimator is from the estimand:

def. bias

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

If $\hat{\theta}$ is *unbiased*, then $\text{bias}(\hat{\theta}) = 0$. Sometimes, unbiased estimators are far more useful and reasonable than biased estimators.

Definition 1.8 (standard error). The *standard error* of an estimator $\hat{\theta}$ is its standard deviation:

def. standard error

$$\text{SE}(\hat{\theta}) = \text{SD}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

Definition 1.9 (estimate). An *estimate* is a realization of an estimator. If our data \mathbf{y} are a realization of \mathbf{Y} and $T(\mathbf{Y})$ is an estimator for some estimand θ , then $T(\mathbf{y})$ is an estimate of θ .

def. estimate

1.3 Method of Moments

Definition 1.10 (sample moment). For a statistical model that Y_1, \dots, Y_n are i.i.d., then the k -th *sample moment*

def. sample moment

$$M_k = \frac{1}{n} \sum_{j=1}^n Y_j^k$$

is an estimator $\hat{\theta} = M_k$ of the k -th moment $\theta = E[Y_1^k]$. The k -th sample moment M_k has properties (if $\text{Var}(Y_1^k) < \infty$) $\text{bias}(M_k) = 0$ and $\text{Var}(M_k) = \text{Var}(Y_1^k)/n$.

The idea of *method of moments* (MoM) is to express the estimand in terms of moments and then directly replace the estimand with the estimator and the moments by the

sample moments.

2 Quality of Estimation

2.1 Loss Functions

Definition 2.1 (loss function). A loss function $L(\theta, \hat{\theta})$ is the loss associated with using the estimator $\hat{\theta}$ when the true parameter value is θ . We require $L(\theta, \hat{\theta}) \geq 0$ and $L(\theta, \theta) = 0$. The expected loss is $E[L(\theta, \hat{\theta})]$.

def. loss function

Definition 2.2 (mean square error, MSE). The loss function $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$ is called the squared error loss and its expected value is the mean square error (MSE):

def. mean square error, MSE

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2].$$

The square root of the MSE is sometimes used and is called the root mean square error (RMSE).

Definition 2.3 (mean absolute error, MAE). The loss function $L(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$ is called the absolute error loss and its expected value is the mean absolute error (MAE):

def. mean absolute error, MAE

$$\text{MAE}(\hat{\theta}) = E[|\hat{\theta} - \theta|].$$

Squared error loss punishes large errors more severely than absolute error loss. Mathematically, square error loss is easier to work with because its derivative is continuous. Because of this, the square error loss is much more widely used.

Definition 2.4 (0-1 loss). The loss function $L(\theta, \hat{\theta}) = I(\hat{\theta} \neq \theta)$ is called the 0-1 loss and its expected value is $P(\hat{\theta} \neq \theta)$. This loss function is useful when the estimator is only ever right or wrong.

def. 0-1 loss

2.2 Bias-Variance Tradeoff

Ideally, we want our estimator $\hat{\theta}$ to have a low bias and a low variance. However, often times there is a trade off between these two in estimation. This situation is ubiquitous in machine learning and data science.

Theorem 2.1 The mean square error of an estimator $\hat{\theta}$ is the variance plus the square of the bias:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2.$$

Example 2.1 For i.i.d. data Y_1, \dots, Y_n with mean μ and variance σ^2 , then \bar{Y} is unbiased and $\text{MSE}(\bar{Y}) = \text{Var}(\bar{Y}) = \sigma^2/n$. This means $\text{MSE}(\bar{Y}) \rightarrow 0$ as $n \rightarrow \infty$.

2.3 Consistency of Estimators

Definition 2.5 (consistent). An estimator $\hat{\theta}$ is consistent if it converges in probability to θ as the sample size $n \rightarrow \infty$. Namely, for every $\epsilon > 0$ we have

def. consistent

$$P(|\hat{\theta} - \theta| \geq \epsilon) \rightarrow 0$$

as $n \rightarrow \infty$. In shorthand, this is expressed as $\hat{\theta} \xrightarrow{p} \theta$.

Example 2.2 Let Y_1, \dots, Y_n be i.i.d. There are infinitely many consistent estimators for $E[Y_1]$. Most simply, \bar{Y} is consistent by the strong law of large numbers. While a very inefficient use of the data, discarding odd numbered data and taking the mean of all even data is also consistent:

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} Y_{2k}.$$

Adding a second term to any consistent estimator that converges to 0 as n approaches ∞ gives us a ridiculous yet consistent estimator like

$$\bar{Y} + \frac{10^{100}}{\sqrt{n}}.$$

Theorem 2.2 If $\text{MSE}(\hat{\theta}) \rightarrow 0$ for some estimator $\hat{\theta}$ as the sample size $n \rightarrow \infty$, then $\hat{\theta}$ is consistent.

Theorem 2.3 If g is a continuous function and $\hat{\theta} \xrightarrow{p} \theta$, then $g(\hat{\theta}) \xrightarrow{p} g(\theta)$.

Example 2.3 If $S^2 \xrightarrow{p} \text{Var}(Y_1) > 0$, then $S \xrightarrow{p} \sqrt{\text{Var}(Y_1)} = \text{SD}(Y_1)$ and $S^{-1} \xrightarrow{p} 1/\text{SD}(Y_1)$.

3 Maximum Likelihood Estimation

Definition 3.1 (maximum likelihood estimate). The *maximum likelihood estimate* of parameter θ is the value $\hat{\theta}$ that maximizes the likelihood function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{y}).$$

def. maximum
likelihood estimate

This parameter value is the most consistent with the data since it makes the observed data as probable as possible. The corresponding estimator is called the *maximum likelihood estimator*:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta; \mathbf{Y}).$$

Example 3.1 Consider a model $Y \sim \text{Bin}(N, p)$ with n known and $\theta = p$ unknown. This is only a sample size of 1. Dropping the binomial coefficient since it acts as a constant,

$$l(p; y) = \log L(p; y) = \log(p^y(1-p)^{n-y}) = y \log p + (n-y) \log(1-p).$$

To find the MLE \hat{p} , we set the derivative of $l(p; y)$ to 0 and rearrange for \hat{p} :

$$\begin{aligned} l'(p; y) &= \frac{\partial l(p; y)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p}, \\ \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} &= 0 \implies \hat{p} = \frac{y}{n}. \end{aligned}$$

With the second derivative test, we know that $p = \hat{p}$ is a local maximum. Moreover, the log-likelihood is globally concave since $l''(p; y) = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} < 0$, meaning \hat{p} is a global maximum and thus the unique MLE. To assess the MLE, we can compute that \hat{p} is

unbiased and decreases to 0 at a rate of \sqrt{n} as $n \rightarrow \infty$:

$$\begin{aligned}\text{bias}(\hat{p}) &= E[\hat{p}] - p = \frac{E[Y]}{n} - p = \frac{np}{n} - p = 0, \\ \text{SE}(\hat{p}) &= \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{\text{Var}(Y)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \frac{c}{\sqrt{n}}\end{aligned}$$

3.1 Properties of the MLE

Remember that the MLE is just one of many estimators. While other estimators may be preferable over the MLE, the MLE tends to work well in theory and practice. And in using the MLE, we get many useful properties.

Theorem 3.1 (invariance of MLE). Let $\hat{\theta}$ be the MLE of θ , and let g be an injection. Then the MLE of $g(\theta)$ is $g(\hat{\theta})$. This follows directly from the invariance property of likelihood.

Definition 3.2 (invariance of MLE for non-injective transformations). If $\hat{\theta}$ is the MLE of θ and g is not an injection, then we define the MLE of $g(\theta)$ to be $g(\hat{\theta})$ since the invariance of the MLE is so convenient.

def. invariance of MLE for non-injective transformations

Example 3.2 Let $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with both parameters unknown. We can parametrize the model in terms of the mean and the standard deviation, $\theta = (\mu, \sigma)$, and let this be our estimand. (We can parametrize the model however we want. The MLE of μ and σ will be the same because of the invariance property of likelihood). We calculate that (removing the constant factor of the normal PDF)

$$\begin{aligned}l(\mu, \sigma; \mathbf{Y}) &= \log L(\mu, \sigma; \mathbf{Y}) = \log \prod_{j=1}^n \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_j - \mu)^2\right) \\ &= \log\left(\frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^n (Y_j - \mu)^2\right)\right) \\ &= -\frac{1}{2\sigma^2} \left(\sum_{j=1}^n (Y_j - \bar{Y})^2 + n(\bar{Y} - \mu)^2\right) - n \log \sigma.\end{aligned}$$

Maximizing this function using multivariable calculus gives

$$\hat{\theta} = \left(\bar{Y}, \sqrt{\frac{\sum_{j=1}^n (Y_j - \bar{Y})^2}{n}}\right)$$

as the MLE of θ . Notice that $\hat{\mu}$ is just the sample mean and $\hat{\sigma}$ is the sample standard deviation. By the invariance of MLE, the MLE of σ^2 is simply $\hat{\sigma}^2$. Extending the example, suppose that Y_j is the height of the j th individual in feet and our new estimand is the probability of someone being more than six feet tall:

$$\theta = P(Y_j > 6) = P(Y_1 > 6) = 1 - P(Y_1 \leq 6) = 1 - \Phi\left(\frac{6 - \mu}{\sigma}\right).$$

To get the MLE of θ , invariance explains that we simply make the parameters don hats:

$$\hat{\theta} = 1 - \Phi\left(\frac{6 - \hat{\mu}}{\hat{\sigma}}\right).$$

3.2 Kullback-Leibler Divergence

In order to eliminate confusion, we summarize the usage of three different thetas:

- θ^* is the true value or estimand generating \mathbf{Y} through $F_{\mathbf{Y}}(\mathbf{y} \mid \theta^*)$,
- θ is the argument in the likelihood $L(\theta; \mathbf{Y})$ and log-likelihood $l(\theta; \mathbf{Y})$, and
- $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} l(\theta; \mathbf{Y})$ is an estimator of θ^* .

It is useful to measure the expected log-likelihood of data evaluated at θ :

$$\mathbb{E}[l(\theta; \mathbf{Y})] = \int l(\theta, \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y} \mid \theta^*) d\mathbf{y}.$$

This means high likelihoods of data tend to pop up for values of θ near θ^* . We can show that this expectation is globally maximized at θ^* even in models where the data are not i.i.d.

Definition 3.3 (Kullback-Leibler divergence). The Kullback-Leibler divergence (or distance) from $F_{\mathbf{Y}}(\mathbf{y} \mid \theta^*)$ to $F_{\mathbf{Y}}(\mathbf{y} \mid \theta)$ is

def. Kullback-Leibler divergence

$$K(\theta^*, \theta) = \mathbb{E} \left[\log \frac{L(\theta^*; \mathbf{Y})}{L(\theta; \mathbf{Y})} \right] = \mathbb{E}[l(\theta^*; \mathbf{Y}) - l(\theta; \mathbf{Y})]$$

computed under the distribution $\mathbf{Y} \sim F_{\mathbf{Y}}(\mathbf{y} \mid \theta^*)$. The Kullback-Leibler divergence measures the impact on expected log-likelihood if we use an approximate distribution in place of the true distribution.

Theorem 3.2 The Kullback-Leibler divergence is non-negative and globally minimized at 0 when $\theta = \theta^*$: For any $\theta \neq \theta^*$, $K(\theta^*, \theta) > 0$. If $\theta = \theta^*$, $K(\theta^*, \theta) = 0$.

This result means that $\mathbb{E}[l(\theta; \mathbf{Y})]$ is maximized at $\theta = \theta^*$, hence shedding light on why MLE is useful. The MLE $\hat{\theta}$ maximizes the *observed* log-likelihood function, while θ^* maximizes the *expected* log-likelihood function. Using the MLE exploits that the log-likelihood will tend to have its peak near θ^* .

3.3 Fisher Information

Definition 3.4 (score function). The score function is

def. score function

$$s(\theta; \mathbf{y}) = \frac{\partial l(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{L(\theta; \mathbf{y})} \frac{\partial L(\theta; \mathbf{y})}{\partial \theta}.$$

We've seen the score function already; solving $s(\theta; \mathbf{y}) = 0$ gives the MLE. Note that the score function can also be fixed for the true value θ^* and vary randomly with the data \mathbf{Y} : $s(\theta^*; \mathbf{Y})$. (This, however, is not a statistic since it depends on the unknown true value).

Theorem 3.3 Under some regularity conditions, $\mathbb{E}[s(\theta^*; \mathbf{Y})] = 0$ and $\operatorname{Var}(s(\theta^*; \mathbf{Y})) = -\mathbb{E}[s'(\theta^*; \mathbf{Y})]$.

Definition 3.5 (Fisher information). The Fisher information in the sample for a parameter θ is

def. Fisher information

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \operatorname{Var}(s(\theta; \mathbf{Y})) = \mathbb{E}[s(\theta; \mathbf{Y})^2] = -\mathbb{E}[s'(\theta; \mathbf{Y})].$$

where we compute the assumption that the true parameter is θ .

If the log-likelihood has a stronger peak at the MLE, the data seems very informative. The Fisher information captures the amount of this curvature: it is the average value of the curvature of the log-likelihood averaged over all possible datasets. In some sense, it is the average amount of information that the data has to offer about the parameter.

4 Bayesian Statistical Inference

The *Bayesian* approach to statistics gives the parameter θ of the model a distribution to reflect our uncertainty about it. We say “there is a 30% chance of rain tomorrow” even though this is not true. It will only rain or not rain, we just don’t know which. If we knew enough about meteorology and had enough computational power, we can exactly determine whether it will rain or not tomorrow. This does not reflect reality, so we use the knowledge we can obtain to determine a best guess to whether it will rain or not. The same goes with θ : even though it is fixed but unknown, treating it like a random variable reflects our uncertainty about it.

4.1 Prior to Posterior

Theorem 4.1 Consider a parametric model $f(\mathbf{y} | \theta)$ for data \mathbf{y} , and let $\pi(\theta)$ be the prior density on the parameter θ . Then the posterior density for θ is proportional to the likelihood times the prior:

$$\pi(\theta | \mathbf{y}) = \frac{L(\theta; \mathbf{y})\pi(\theta)}{f(\mathbf{y})} \propto L(\theta; \mathbf{y})\pi(\theta).$$