Lecture Notes 13
36-705

Today we will discuss *point estimation.* Given $X_1, \ldots, X_n \sim p(X;\theta)$ we would like to construct an *estimator* $\widehat{\theta}(X_1, \ldots, X_n)$ of $\theta = (\theta_1, \ldots, \theta_k)$. An *estimator*

$$\widehat{\theta} = \widehat{\theta}_n = w(X_1, \ldots, X_n)$$

is any function of the data. Keep in mind that the parameter is a fixed, unknown constant. The estimator is a random variable.

In the next few lectures we will discuss ways to construct estimators and then we dicuss how to compare or evaluate them. The questions we are trying to answer are:

1. Are there general purpose methods to come up with estimators of θ?
2. Given two (or more) estimators is there a general framework in which we can compare estimators?
3. How do we analyze complex estimators (say estimators that are not simple averages)?

An "estimator" refers to a random variable (a statistic, a function of the sample) and an "estimate" refers to its realized value. We have already studied estimation in a relatively simple context: estimating the mean. When θ is not a mean then we need to think a bit harder to decide how to estimate it. We will focus on general purpose methods for estimation.

For now, we will discuss three methods of constructing estimators:

1. The Method of Moments (MOM)
2. Maximum likelihood (MLE)
3. Bayes estimators.

Some Terminology. Throughout these notes, we will use the following terminology:

1. $\mathbb{E}_\theta(\widehat{\theta}) = \int \cdots \int \widehat{\theta}(x_1, \ldots, x_n) p(x_1;\theta) \cdots p(x_n;\theta) dx_1 \cdots dx_n$.
2. Bias: $\mathbb{E}_\theta(\widehat{\theta}) - \theta$.
3. The distribution of $\widehat{\theta}_n$ is called its *sampling distribution.*
4. The standard deviation of $\widehat{\theta}_n$ is called the *standard error* denoted by $\mathrm{se}(\widehat{\theta}_n)$.
5. $\widehat{\theta}_n$ is *consistent* if $\widehat{\theta}_n \xrightarrow{p} \theta$. Later we will see that if bias $\to 0$ and $\mathrm{Var}(\widehat{\theta}_n) \to 0$ as $n \to \infty$ then $\widehat{\theta}_n$ is consistent.

1 The Method of Moments

Suppose that $\theta = (\theta_1, \ldots, \theta_k)$. Define

$$\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^n X_i, & \mu_1(\theta) &= \mathbb{E}(X_i) = \int x p_\theta(x) dx \\
m_2 &= \frac{1}{n}\sum_{i=1}^n X_i^2, & \mu_2(\theta) &= \mathbb{E}(X_i^2) = \int x^2 p_\theta(x) dx \\
&\vdots & &\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^n X_i^k, & \mu_k(\theta) &= \mathbb{E}(X_i^k) = \int x^k p_\theta(x) dx.
\end{aligned}$$

Let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ solve:

$$m_j = \mu_j(\widehat{\theta}), \quad j = 1, \ldots, k.$$

In other words, we equate the first k sample moments with the first k theoretical moments. This defines k equations with k unknowns.

Example 1 *$N(\beta, \sigma^2)$ with $\theta = (\beta, \sigma^2)$. Then $\mu_1 = \beta$ and $\mu_2 = \sigma^2 + \beta^2$. Equate:*

$$\frac{1}{n}\sum_{i=1}^n X_i = \widehat{\beta}, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = \widehat{\sigma}^2 + \widehat{\beta}^2$$

to get

$$\widehat{\beta} = \overline{X}_n, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Example 2 *Suppose*

$$X_1, \ldots, X_n \sim \text{Binomial}(k, p)$$

where both k and p are unknown. We get

$$kp = \overline{X}_n, \quad \frac{1}{n}\sum_{i=1}^n X_i^2 = kp(1-p) + k^2p^2$$

giving

$$\widehat{p} = \frac{\overline{X}_n}{k}, \quad \widehat{k} = \frac{\overline{X}_n^2}{\overline{X}_n - \frac{1}{n}\sum_i (X_i - \overline{X}_n)^2}.$$

The method of moments was popular many years ago because it is often easy to compute. Lately, it has attracted attention again. For example, there is a large literature on estimating mixtures of Gaussians using the method of moments.

2 Maximum Likelihood

The most popular method for estimating parameters is maximum likelihood. One of the reasons is that, under certain conditions, the maximum likelihood estimator is optimal. We'll discuss optimality later.

The maximum likelihood estimator (mle) $\widehat{\theta}$ is defined as the maximizer of

$$\mathcal{L}(\theta) = p(X_1, \ldots, X_n; \theta) \overset{iid}{=} \prod_i p(X_i; \theta).$$

This is the same as maximizing the log-likelihood

$$\ell(\theta) = \log \mathcal{L}(\theta).$$

Often it suffices to solve

$$\frac{\partial \ell(\theta)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, k.$$

Example 3 *Binomial.* $\mathcal{L}(p) = \prod_i p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$ *where* $S = \sum_i X_i$. *So*

$$\ell(p) = S \log p + (n - S)\log(1-p)$$

and $\widehat{p} = \overline{X}_n$.

Example 4 $X_1, \ldots, X_n \sim N(\mu, 1)$.

$$\mathcal{L}(\mu) \propto \prod_i e^{-(X_i-\mu)^2/2} \propto e^{-n(\overline{X}_n-\mu)^2}, \quad \ell(\mu) = -\frac{n}{2}(\overline{X}_n - \mu)^2$$

and $\widehat{\mu} = \overline{X}_n$. *For* $N(\mu, \sigma^2)$ *we have*

$$\mathcal{L}(\mu, \sigma^2) \propto \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right\}$$

and

$$\ell(\mu, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Set

$$\frac{\partial \ell}{\partial \mu} = 0, \quad \frac{\partial \ell}{\partial \sigma^2} = 0$$

to get

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Example 5 *Let* $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. *Then*

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} I(\theta > X_{(n)})$$

and so $\widehat{\theta} = X_{(n)}$.

What is the method of moments estimator above? How would you compare the two estimators?

2.1 MLE and MoM for exponential families

The log-likelihood in an exponential family is concave and given by

$$\ell(\theta; x_1, \ldots, x_n) \propto \left[\sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(x_j) - nA(\theta)\right],$$

so we can simply take the derivative with respect to θ and set this equal to 0. Using the facts we have seen in the last lecture about the derivative of A, we can see that this amounts to solving the following system of equations for θ:

$$\mathbb{E}_{p(X;\theta)}[T_i(X)] = \frac{1}{n}\sum_{j=1}^{n} T_i(x_j) \quad \text{for} \quad i \in \{1, \ldots, s\}.$$

So the maximum likelihood estimator simply picks the parameters θ to match the empirical expectations of the sufficient statistics to the expected value of the sufficient statistics under the distribution.

Usually we cannot compute this estimator in closed form so we use an iterative algorithm (like gradient ascent) to maximize the likelihood. However, you should remember that exponential families have concave likelihoods so this is usually tractable. For exponential families as we can see above the method of moments coincides with the MLE (if we chose the sufficient statistics to direct which moments to compute).

Suppose that $\theta = (\eta, \xi)$. The *profile likelihood* for η is defined by

$$\mathcal{L}(\eta) = \sup_{\xi} \mathcal{L}(\eta, \xi).$$

To find the mle of η we can proceed in two ways. We could find the overall mle $\widehat{\theta} = (\widehat{\eta}, \widehat{\xi})$. The mle for η is just the first coordinate of $(\widehat{\eta}, \widehat{\xi})$. Alternatively, we could find the maximizer of the profile likelihood. These give the same answer. Do you see why?

2.2 Equivariance and the profile likelihood

The mle is *equivariant.* if $\eta = g(\theta)$ then $\widehat{\eta} = g(\widehat{\theta})$. Suppose g is invertible so $\eta = g(\theta)$ and $\theta = g^{-1}(\eta)$. Define $\mathcal{L}^*(\eta) = \mathcal{L}(\theta)$ where $\theta = g^{-1}(\eta)$. So, for any η,

$$\mathcal{L}^*(\widehat{\eta}) = \mathcal{L}(\widehat{\theta}) \geq \mathcal{L}(\theta) = \mathcal{L}^*(\eta)$$

and hence $\widehat{\eta} = g(\widehat{\theta})$ maximizes $\mathcal{L}^*(\eta)$. For non invertible functions this is still true if we define $\mathcal{L}^*(\eta)$ to be the profile likelihood.

Example 6 *Binomial. The mle is* $\widehat{p} = \overline{X}_n$. *Let* $\psi = \log(p/(1-p))$. *Then* $\widehat{\psi} = \log(\widehat{p}/(1-\widehat{p}))$.

3 Bayes Estimator

To define the Bayes estimator, we begin by treating θ as a random variable. This point requires much discussion (which we will have later). For now, just tentatively think of θ as random. We start with a *prior distribution* $p(\theta)$ on θ. Note that

$$p(x_1, \ldots, x_n|\theta)p(\theta) = p(x_1, \ldots, x_n, \theta).$$

Now compute the *posterior distribution* by Bayes' theorem:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}$$

where

$$p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta.$$

This can be written as

$$p(\theta|x_1, \ldots, x_n) \propto \mathcal{L}(\theta)p(\theta) = \text{Likelihood} \ \times \ \text{prior}.$$

Now compute a point estimator from the posterior. For example:

$$\widehat{\theta} = \mathbb{E}(\theta|x_1, \ldots, x_n) = \int \theta p(\theta|x_1, \ldots, x_n)d\theta = \frac{\int \theta p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta}.$$

Example 7 *Let* $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$*. Let the prior be* $\theta \sim \text{Beta}(\alpha, \beta)$*. Hence*

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

and

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

Set $Y = \sum_i X_i$*. Then*

$$p(\theta|X) \propto \underbrace{\theta^Y (1-\theta)^{n-Y}}_{\text{likelihood}} \times \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\text{prior}} \propto \theta^{Y+\alpha-1}(1-\theta)^{n-Y+\beta-1}.$$

Therefore, $\theta|X \sim \text{Beta}(Y+\alpha, n-Y+\beta)$*. The Bayes estimator is*

$$\widetilde{\theta} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{\theta}_{mle} + \lambda\, \overline{\theta}$$

where

$$\overline{\theta} = \frac{\alpha}{\alpha+\beta}, \qquad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

This is an example of a conjugate prior.

Example 8 *Let* $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ *with* σ^2 *known. Let* $\mu \sim N(m, \tau^2)$*. Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \overline{X}_n + \frac{\frac{\sigma^2}{n}}{\tau^2 + \frac{\sigma^2}{n}} m$$

and

$$\text{Var}(\mu|X) = \frac{\sigma^2 \tau^2 / n}{\tau^2 + \frac{\sigma^2}{n}}.$$

4 MSE

Now we discuss the evaluation of estimators. The mean squared error (MSE) is

$$\mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \int \cdots \int (\widehat{\theta}(x_1, \ldots, x_n) - \theta)^2 p(x_1; \theta) \cdots p(x_n; \theta) dx_1 \ldots dx_n.$$

The bias is

$$B = \mathbb{E}_\theta(\widehat{\theta}) - \theta$$

and the variance is

$$V = \text{Var}_\theta(\widehat{\theta}).$$

Theorem 9 *We have*

$$MSE = B^2 + V.$$

Proof: Let $m = \mathbb{E}_\theta(\widehat{\theta})$. Then

$$\begin{aligned} MSE &= \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathbb{E}_\theta(\widehat{\theta} - m + m - \theta)^2 \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 + 2\mathbb{E}_\theta(\widehat{\theta} - m)(m - \theta) \\ &= \mathbb{E}_\theta(\widehat{\theta} - m)^2 + (m - \theta)^2 = V + B^2. \end{aligned}$$

■

An estimator is *unbiased* if the bias is 0. In that case, the MSE = Variance. There is often a tradeoff between bias and variance. So low bias can imply high variance and vice versa.

Example 10 *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then*

$$\mathbb{E}(\overline{X}) = \mu, \quad \mathbb{E}(S^2) = \sigma^2.$$

The MSE's are

$$\mathbb{E}(\overline{X} - \mu)^2 = \frac{\sigma^2}{n}, \quad \mathbb{E}(S^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1}.$$

It is worth thinking about how one defines the MSE when θ is multivariate (as in the example above), and what the analogous bias-variance decomposition is.

We would like to choose an estimator with small MSE. However, the MSE is a function of θ. Later, we shall discuss minimax estimators, that use the maximum of the MSE over θ as a way to compare estimators.

5 Unbiased estimators, Fisher Information, Cramér-Rao

In the olden days, many people focused on unbiased estimators. More modern treatments do not often emphasize this point of view since there are many known examples where a small amount of bias can result in large reductions in variance.

With that said, there are still pieces of this classical theory that are useful. One of the important pieces is the Cramér-Rao bound which provides a lower bound on the variance of an unbiased estimator. In many problems, this bound will provide some at least heuristic guidelines into the difficulty of an estimation problem. Later on in the course we will talk about other ways of proving lower bounds that do not restrict attention to unbiased estimators (i.e. we will discuss what are called minimax lower bounds).

5.1 Fisher Information

The log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log p(X_i;\theta).$$

We can also define the gradient of this function, which is called the *score function*:

$$s(\theta) = s(\theta; X_1, \ldots, X_n) = \nabla_\theta \ell(\theta) = \sum_{i=1}^{n} \nabla_\theta \log p(X_i;\theta).$$

This gradient is a d-dimensional vector. The Fisher Information matrix is the expected outer product of the score, i.e.:

$$I(\theta) = \mathbb{E}[s(\theta)s(\theta)^T].$$

The Fisher information matrix is a $d \times d$ matrix. Lets take a quick look at a couple of examples:

Example 1: Suppose that $X \sim \text{Ber}(p)$, then the log-likelihood is given by:

$$\ell(p) = X\log(p) + (1-X)\log(1-p),$$

and the score is:

$$s(p) = \frac{X}{p} - \frac{1-X}{1-p} = \frac{X-p}{p(1-p)}.$$

We can then compute the Fisher information:

$$I(p) = \frac{1}{p^2(1-p)^2}\mathbb{E}\left[(X-p)^2\right] = \frac{1}{p(1-p)}.$$

Example 2: Suppose that $X \sim N(\mu, \sigma^2)$ where σ is known, then the log-likelihood is given by:

$$\ell(\mu) = -\frac{1}{2\sigma^2}(X-\mu)^2,$$

so that the score is:

$$s(\mu) = \frac{X-\mu}{\sigma^2},$$

and the Fisher information is:

$$I(\mu) = \mathbb{E}\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

An important property that we will use is that the score function has mean zero, i.e.

$$\mathbb{E}_\theta[s(\theta)] = \int \cdots \int s(\theta; x_1, \ldots, x_n) p_\theta(x_1, \ldots, x_n) dx_1 \cdots dx_n = 0.$$

Proof: Notice that,

$$\begin{aligned}
\mathbb{E}_\theta[s(\theta)] &= \sum_{i=1}^n \int \nabla_\theta \log p(x_i; \theta) \ np(x_1, \ldots, x_n; \theta) dx_1 dx_2 \ldots dx_n \\
&= \sum_{i=1}^n \int \nabla_\theta \log p(x_i; \theta) \ \ p(x_i; \theta) dx_i \\
&= n \int \nabla_\theta \log p(x_1; \theta) p(x_1; \theta) dx_1,
\end{aligned}$$

using the i.i.d. assumption several times. Under some regularity conditions we can switch the derivative and integral so we obtain,

$$\begin{aligned}
\int \nabla_\theta \log p(x_1; \theta) \ p(x_1; \theta) dx_1 &= \int \frac{\nabla_\theta p(x_1; \theta)}{p(x_1; \theta)} p(x_1; \theta) dx_1 \\
&= \nabla_\theta \int p(x_1; \theta) dx_1 = \nabla_\theta 1 = 0. \ \square
\end{aligned}$$

One consequence of this property is that we can interpret the Fisher information matrix as the covariance matrix of the score, i.e.

$$I(\theta) = \mathbb{E}\Big[(s(\theta) - \mathbb{E}(s(\theta)))(s(\theta) - \mathbb{E}(s(\theta)))^T\Big].$$

You should check that the following holds:

$$I(\theta) = nI_1(\theta)$$

where $I_1(\theta)$ is the Fisher information based on one observation.

Lemma 11 *The Fisher information satisfies*

$$I_1(\theta) = -\mathbb{E}\left[\nabla_\theta^2 \log p(X; \theta)\right].$$

Proof. To see this observe that,

$$\begin{aligned}
\nabla_\theta^2 \log p(X; \theta) &= \nabla_\theta \frac{\nabla_\theta p(X; \theta)}{p(X; \theta)} \\
&= \frac{\nabla_\theta^2 p(X; \theta)}{p(X; \theta)} - \frac{(\nabla_\theta p(X; \theta) \nabla_\theta p(X; \theta)^T)}{p(X; \theta)^2} \\
&= \frac{\nabla_\theta^2 p(X; \theta)}{p(X; \theta)} - s(\theta) s(\theta)^T.
\end{aligned}$$

Now, notice that,

$$\mathbb{E}\left[\frac{\nabla_\theta^2 p(X;\theta)}{p(X;\theta)}\right] = \int \nabla_\theta^2 p(X;\theta) = \nabla_\theta^2 \int p(X;\theta) = \nabla_\theta^2 1 = 0,$$

which yields the result. □

The Fisher information is measuring the expected curvature of the log-likelihood function around the point θ. As we will see in future lectures if the log-likelihood is more curved (i.e. $I(\theta)$ is appropriately "large") then θ is easier to estimate.

Example 3: Let $X_1, \ldots, X_n \sim \text{Ber}(p)$. Then

$$I(p) = \frac{n}{p(1-p)}.$$

Example 4: For exponential families we have seen that the log-likelihood is given as:

$$\ell(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{s} \theta_i \sum_{j=1}^{n} T_i(X_j) - nA(\theta),$$

so the Hessian is:

$$I(\theta) = n\nabla_\theta^2 A(\theta) = n\mathbb{E}\Bigg[(T(X) - \mathbb{E}[T(X)])(T(X) - \mathbb{E}[T(X)])^T\Bigg],$$

so the Fisher information matrix is n times the Hessian of the log-partition function or alternatively it is the covariance matrix of the vector of sufficient statistics.

5.2 Cramér-Rao Bound

Let us briefly consider again the Bernoulli example. We observe $X_1, \ldots, X_n \sim \text{Ber}(p)$ and estimate $\widehat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. The estimator is unbiased and has variance $p(1-p)/n$ which is precisely the inverse of the Fisher information. This turns out to be a fairly general phenomenon. Indeed, the Cramér-Rao bound assures us that this estimator is unimprovable in a certain sense. We focus first on the univariate case (when $\theta \in \mathbb{R}$) and then consider the multivariate extension.

Cramér-Rao Bound: Suppose that $X_1, \ldots, X_n \sim p(X;\theta)$ and that $\widehat{\theta}$ is an unbiased estimator of θ. Then

$$\text{Var}(\widehat{\theta}) \geq \frac{1}{nI_1(\theta)}.$$

Proof: Consider that,

$$\mathrm{cov}(\widehat{\theta}, s(\theta)) = \mathbb{E}((\widehat{\theta} - \theta)s(\theta)) = \mathbb{E}(\widehat{\theta}s(\theta)),$$

since $\mathbb{E}[s(\theta)] = 0$. Furthermore,

$$\begin{aligned}
\mathbb{E}(\widehat{\theta}s(\theta)) &= \int \widehat{\theta}(x_1, \ldots, x_n) \nabla_\theta \log p(x_1, \ldots, x_n; \theta) p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n \\
&= \int \widehat{\theta}(x_1, \ldots, x_n) \frac{\nabla_\theta p(x_1, \ldots, x_n; \theta)}{p(x_1, \ldots, x_n; \theta)} p(x_1, \ldots, x_n; \theta) dx_1 \ldots dx_n \\
&= \int \widehat{\theta}(x_1, \ldots, n) \ \nabla_\theta p(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n \\
&= \nabla_\theta \int \widehat{\theta}(x_1, \ldots, n) \ p(x_1, \ldots, x_n; \theta) dx_1 \cdots dx_n = \nabla_\theta \ \theta = 1.
\end{aligned}$$

Notice that for any fixed ζ we can write:

$$\mathrm{Var}(\widehat{\theta} - \zeta s(\theta)) = \mathrm{Var}(\widehat{\theta}) + \zeta^2 \mathrm{Var}(s(\theta)) - 2\zeta \mathrm{cov}(\widehat{\theta}, s(\theta)) = \mathrm{Var}(\widehat{\theta}) + \zeta^2 n I_1(\theta) - 2\zeta.$$

Using the fact that variances are positive we can write:

$$\mathrm{Var}(\widehat{\theta}) \geq 2\zeta - \zeta^2 n I_1(\theta)$$

Take $\zeta = 1/(nI_1(\theta))$ to obtain the Cramér-Rao bound.

Multivariate Generalization: The Cramér-Rao bound can be derived for a multivariate parameter θ in a very similar fashion. For any two positive semi-definite matrices A and B write

$$A \succeq B$$

to mean that

$$v^T A v \geq v^T B v$$

for any vector v. The multivariate Cramer-Rao bound is

$$\mathrm{Var}(\widehat{\theta}) \succeq I(\theta)^{-1} = \frac{1}{n} I_1(\theta)^{-1}.$$

Examples: In both the Gaussian and Bernoulli models, as a consequence of the Cramér-Rao bound we can conclude that the MLE is the best unbiased estimator.

6 Beyond unbiased estimators: Decision theory

The central idea in decision theory is that we want to minimize our *expected* loss. Let $X_1, \ldots, X_n \sim p(X; \theta)$, with $\theta \in \Theta$. We choose a loss function $L(a, \theta)$ which measures how far a point a is from θ. Some common loss functions are:

1. **Squared loss:** $L(a,\theta) = (a-\theta)^2$.
2. **Absolute loss:** $L(a,\theta) = |a-\theta|$.

There are however many other loss functions. For instance, we sometimes consider losses like:

$$L(a,\theta) = \frac{(a-\theta)^2}{|\theta|+1},$$

which penalizes errors in estimation more for small values of θ than for large values. We can similarly design a loss function that penalizes errors more strongly for large values of θ.

Another important point is that there are cases when we do not really care about estimating the parameter well but rather just the distribution $p(x;\theta)$. This is true when we care about prediction in regression or in density estimation. In this case we could define the loss between θ and a in terms of the distributions $p(x;\theta)$ and $p(x;a)$. One canonical example is:

Kullback-Leibler loss:

$$L(a,\theta) = \text{KL}(p(x;\theta), p(x;a)) = \int p(x;\theta) \log\left(\frac{p(x;\theta)}{p(x;a)}\right) dx.$$

Once we have a loss function, and an estimator, we can assess the estimator via its expected loss. This expected loss is called the *risk* of the estimator. Suppose we consider an estimator $\widehat{\theta} = \widehat{\theta}(X_1,\ldots,X_n)$; the risk is

$$R(\theta,\widehat{\theta}) = \mathbb{E}_\theta L(\widehat{\theta},\theta).$$

Ideally, we would like to find an estimator $\widehat{\theta}$ such that for any other estimator θ' we have that:

$$R(\theta,\widehat{\theta}(X)) \leq R(\theta,\theta')$$

for all values θ. Such estimators will most often not exist – why not? So we will need to find another way to define an optimal estimator.