1 Point Estimation

Under the hypothesis that the sample was generated from some parametric statistical model, a natural way to understand the underlying population is by estimating the parameters of the statistical model.

One can of course wonder what happens if the model is wrong? Or where do models really come from? In some rare cases, we actually know enough about our data to hypothesise a reasonable model. Most often however, when we specify a model, we do so hoping that it can provide a useful approximation to the data generation mechanism. The George Box quote is worth remembering in this context: “all models are wrong, but some are useful.”

1.1 The Method of Moments

Suppose that \( \theta = (\theta_1, \ldots, \theta_k) \) so that there are \( k \) unknown parameters. We can estimate \( \theta \) by matching \( k \) moments. Let

\[
m_1 = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad m_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2, \ldots, \quad m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k.
\]

Let \( \mu_i = \int x^i p_\theta(x) dx \) denote the \( i \)th population moment. This depends on \( \theta \) so we write it as \( \mu_i(\theta) \). The method-of-moments prescribes estimating the parameters: \( \theta_1, \ldots, \theta_k \) by solving the system of equations:

\[
m_1 = \mu_1(\theta_1, \ldots, \theta_k) \\
\vdots \\
m_k = \mu_k(\theta_1, \ldots, \theta_k).
\]

Example 1: If \( X_1, \ldots, X_n \sim N(\theta, \sigma^2) \), we would solve:

\[
\frac{1}{n} \sum_{i=1}^{n} X_i = \theta, \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \theta^2 + \sigma^2,
\]
to obtain the estimators:

\[ \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n \]

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 \]

\[ = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \equiv s_n^2. \]

Example 2: Suppose \( X_1, \ldots, X_n \sim \text{Bin}(k, p) \) where \( k \) and \( p \) are both unknown. Now

\[ \mu_1 = kp \]
\[ \mu_2 = kp(1 - p) + k^2 p^2. \]

Solving, we get

\[ \bar{X} = kp, \quad \frac{1}{n} \sum_{i=1}^{n} X_i^2 = kp(1 - p) + k^2 p^2 \]

which gives

\[ \hat{p} = \frac{\bar{X} - \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}{\bar{X}}, \]
\[ \hat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}. \]

1.2 Maximum Likelihood Estimation

The most popular technique to derive estimators is via the principle of maximum likelihood. Suppose that \( X_1, \ldots, X_n \sim p_\theta \) where \( p_\theta \) denotes either the pmf or pdf.

The likelihood function is defined by:

\[ L(\theta) \equiv L(\theta; X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i). \]

The log-likelihood function is

\[ \ell(\theta) \equiv \ell(\theta; X_1, \ldots, X_n) = \log L(\theta). \]
The **maximum likelihood estimator**, or mle — denoted by \( \hat{\theta} \) or \( \hat{\theta}_n \) — is the value of the \( \theta \) that maximizes \( L(\theta) \). Note that \( \hat{\theta} \) also maximizes \( \ell(\theta) \). We write

\[
\hat{\theta} = \arg\max L(\theta) = \arg\max_{\theta} \ell(\theta).
\]

Keep in mind that \( \hat{\theta} \) is a function of the data. Sometimes we will write \( \hat{\theta} \) as \( \hat{\theta}(X_1, \ldots, \theta_n) \) to emphasize this point.

Later, we shall see that the mle has many optimality properties in certain settings. Finding the mle might not be easy. Sometimes we need to resort to numerical techniques.

The typical way to compute the MLE (suppose that we have \( k \) unknown parameters) is to either analytically or numerically solve the system of equations:

\[
\frac{\partial}{\partial \theta_i} \ell(\theta) = 0 \quad i = 1, \ldots, k.
\]

**Note:** We can throw away any constants not depending on \( \theta \) in the likelihood function when we find the mle. This does not affect the location of the maximizer.

**Example 1:** Suppose \( X_1, \ldots, X_n \sim N(\theta, 1) \), then the likelihood function is given as:

\[
L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i - \theta)^2}{2}\right) \propto e^{-n(\theta - \overline{X}_n)^2/2}
\]

and

\[
\ell(\theta) = -\frac{n(\theta - \overline{X}_n)^2}{2}.
\]

We get \( \hat{\theta}_n = \overline{X}_n \). Since \( \ell''(\hat{\theta}) < 0 \) this is indeed a maximum.

**Example 2:** Suppose that \( X_1, \ldots, X_n \sim \text{Ber}(p) \), then the log-likelihood is given by

\[
\ell(p) \propto \sum_{i=1}^{n} X_i \log p + (1 - X_i) \log (1 - p)
\]

\[= n\overline{X} \log p + n(1 - \overline{X}) \log (1 - p),\]

which is maximized at \( \hat{p} = \overline{X} \).

**Invariance of the MLE.** The mle is invariant to transformations. This means that the mle of \( r(\theta) \) is \( ru(\hat{\theta}) \) for any function \( r \). We will not prove this but it is a very useful fact. We will discuss other properties of the MLE in future lectures.
2 Bayes Estimators

The third general method to derive estimators is the Bayes estimator. We treat $\theta$ as a random variable and assign it a distribution $p(\theta)$ called the prior distribution.

**This opens up a bunch of philosophical questions that we will deal with later in the course.**

Now we can use Bayes theorem to get the distribution of $\theta$ given $X_1, \ldots, X_n$, which is called the posterior distribution:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(\theta, x_1, \ldots, x_n)}{p(x_1, \ldots, x_n)} = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{\int p(x_1, \ldots, x_n|\theta)p(\theta)d\theta} = \frac{L(\theta)p(\theta)}{\int L(\theta)p(\theta)d\theta} \propto L(\theta)p(\theta).$$

Finally, we can use the mean of $p(\theta|x_1, \ldots, x_n)$ as an estimator:

$$\hat{\theta} = \int \theta p(\theta|X_1, \ldots, X_n)d\theta.$$

We call this, the Bayes estimator. We could also use the median or mode of the posterior.

**Example:** Suppose $X_1, \ldots, X_n \sim \text{Ber}(\theta)$. We will first need to define the Beta distribution: $\theta$ has a Beta distribution with parameters $\alpha$ and $\beta$ if its density on $[0, 1]$ is

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

We write $\theta \sim \text{Beta}(\alpha, \beta)$. The mean of the Beta distribution is: $\alpha/(\alpha + \beta)$. Let $S = \sum_i nX_i$.

The posterior distribution is

$$p(\theta|X_1, \ldots, X_n) \propto L(\theta)p(\theta) \propto \theta^S(1-\theta)^{n-S}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{S+\alpha-1}(1-\theta)^{n-S+\beta-1}.$$

Thus, the posterior distribution is $\text{Beta}(S + \alpha, n - S + \beta)$. We write

$$\theta|X_1, \ldots, X_n \sim \text{Beta}(S + \alpha, n - S + \beta).$$

The mean is $(S + \alpha)/(n + \alpha + \beta)$. Thus,

$$\hat{\theta}_n = \frac{S + \alpha}{n + \alpha + \beta}.$$

A common choice is $\alpha = \beta = 1$ (so that the prior for $\theta$ is uniform). In that case:

$$\hat{\theta} = \frac{n\bar{X} + 1}{n + 2} = \frac{n}{n + 2}\bar{X} + \frac{2}{n + 2}\frac{1}{2} = w\bar{X} + (1-w)^1_{2},$$
which can be viewed as a convex combination of the MLE and the prior mean 1/2. Note that, when \( n \) is large, \( \hat{\theta} \approx \text{X}_n \), which is the mle.

**Example 2:** Suppose that \( X_1, \ldots, X_n \) drawn from \( N(\theta, \sigma^2) \). Assume that \( \sigma^2 \) is known. Let’s use the prior \( \theta \sim N(\mu, \tau^2) \). It can be shown that the posterior is \( N(a, b^2) \) where

\[
a = \frac{n\tau^2}{\sigma^2 + n\tau^2} \text{X}_n + \frac{\sigma^2}{\sigma^2 + n\tau^2} \mu, \quad b^2 = \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}.
\]

**Exercise:** Prove this.

The Bayes estimator is thus

\[
\hat{\mu} = \frac{n\tau^2}{\sigma^2 + n\tau^2} \text{X}_n + \frac{\sigma^2}{\sigma^2 + n\tau^2} \mu = w \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) + (1 - w) \mu
\]

where \( w = \frac{n\tau^2}{\sigma^2 + n\tau^2} \). When \( n \) is large, \( w \approx 1 \) and \( \hat{\mu} \approx \text{X}_n \).