Note that u counts the number of each type of transition that occurred in the sequences $X_1, X_2, \ldots, X_n$.

For i.i.d. samples, we typically observe many replicates from the same underlying model. When we use a time series or spatial model, the usual way data arrives is as a single sample from that model, whose length or size might not be a priori specified. For these models to be useful in practice, we need them to be specified with a not very large set of parameters, so that as the data grows (i.e., as the sequence gets longer) we have a hope of being able to estimate the parameters. Of course, it might be the situation that for a time-series or spatial model we have not just one sample, but i.i.d. data. For instance, in Chapter 1 during our discussion of the maximum likelihood estimation in a very short Markov chain, we analyzed the case where we received many data sets of the chain of length 3, where each was an i.i.d. sample from the same underlying distribution.

5.3. Parameter Estimation

Given a parametric statistical model and some data, a typical problem in statistics is to estimate some of, or all of, the parameters of the model based on the data. At this point we do not necessarily assume that the model accurately models the data. The problem of testing whether or not a model actually fits the data is the subject of the next section, on hypothesis testing.

Ideally, we would like a procedure which, as more and more data arrives, if the underlying distribution that generated the data comes from the model, the parameter estimate converges to the true underlying parameter. Such an estimator is called a consistent estimator.

Definition 5.3.1. Let $\mathcal{M}_\Theta$ be a parametric statistical model with parameter space Θ. A *parameter* of a statistical model is a function $s : \Theta \to \mathbb{R}$. An *estimator* of s is a function from the data space D to $\mathbb{R}$, $\hat{s} : D \to \mathbb{R}$. The estimator $\hat{s}$ is *consistent* if $\hat{s} \longrightarrow_p s$ as the sample size tends to infinity.

Among the simplest examples of estimators are the plug-in estimators. As the name suggests, a plug-in estimator is obtained by plugging in values obtained from the data to estimate parameters.

Example 5.3.2. As a simple example, consider the case of a binomial random variable, with $r+1$ states, $0, 1, \ldots, r$. The model consists of all distributions of the form

$$\left\{ \left(\theta^r, \binom{r}{1}\theta^{r-1}(1-\theta), \ldots, (1-\theta)^r \right) : \theta \in [0,1] \right\}.$$

Under i.i.d. sampling, data consists of n repeated draws $X^{(1)}, \ldots, X^{(n)}$ from an underlying distribution p_θ in this model. The data is summarized by a

vector of counts $u = (u_0, \ldots, u_r)$, where $u_i = \#\{j : X^{(j)} = i\}$. We would like to estimate the parameter θ from the data of counts u. The value $p_\theta(0) = \theta^r$, hence, if we had a consistent estimator of $p_\theta(0)$, we could obtain a plug-in estimate for θ by extracting the rth root. For example, the formula

$$\sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} 1_{x=0}(X^{(i)})} = \sqrt[r]{\frac{u_0}{n}}$$

gives a consistent plug-in estimator of the parameter θ.

Intuitively, the plug-in estimator from Example 5.3.2 is unlikely to be a very useful estimator, since it only uses very little information from the data to obtain an estimate of the parameter θ. When choosing a consistent plug-in estimator, we would generally like to use one whose variance rapidly tends to zero as $n \to \infty$. The estimator from Example 5.3.2 has high variance and so is an inefficient estimator of the parameter θ.

Another natural choice for an estimator is a *method of moments* estimator. The idea of the method of moments is to choose the probability distribution in the model whose moments match the empirical moments of the data.

Definition 5.3.3. Given a random vector $X \in \mathbb{R}^m$ and an integer vector $\alpha \in \mathbb{N}^m$, the *αth moment* is

$$\mu_\alpha = \mathrm{E}[X_1^{\alpha_1} \cdots X_m^{\alpha_m}].$$

Given i.i.d. data $X^{(1)}, \ldots, X^{(n)}$, their αth *empirical moment* is the estimate

$$\hat{\mu}_\alpha = \frac{1}{n}\sum_{i=1}^{m} (X_1^{(i)})^{\alpha_1} \cdots (X_m^{(i)})^{\alpha_m}.$$

So in method of moments estimation, we find formulas for some of the moments of the random vector $X \sim p_\theta \in \mathcal{M}_\Theta$ in terms of the parameter vector θ. If we calculate enough such moments, we can find a probability distribution in the model whose moments match the empirical moments. For many statistical models, formulas for the moments in terms of the parameters are given by polynomial or rational formulas in terms of the parameters. So finding the method of moments estimator will turn into the problem of solving a system of polynomial equations.

Example 5.3.4 (Binomial random variable)**.** Consider the example of the model of a binomial random variable from Exercise 5.3.2. Given a binomial random variable $X \sim \mathrm{Bin}(\theta, r)$, the first moment $\mathrm{E}[X] = r\theta$. The empirical first moment of $X^{(1)}, \ldots, X^{(n)}$ is the sample mean $\bar{X}$. Hence the method of moments estimate for θ in the binomial model is

$$\hat{\theta} = \tfrac{1}{r}\bar{X}.$$

The method of moments estimators can often lead to interesting algebraic problems [**AFS16**]. One potential drawback to the method of moments estimators is that the empirical higher moments tend to have high variability, so there can be a lot of noise in the estimates.

Among the many possible estimators of a parameter, one of the most frequently used is the maximum likelihood estimator (MLE). The MLE is one of the most commonly used estimators in practice, both for its intuitive appeal and for useful theoretical properties associated with it. In particular, it is usually a consistent estimator of the parameters and, with certain smoothness assumptions on the model, it is asymptotically normally distributed. We will return to these properties in Chapter 7.

Definition 5.3.5. Let D be data from some model with parameter space Θ. The *likelihood function*

$$L(\theta \mid D) := p_\theta(D)$$

in the case of discrete data and

$$L(\theta \mid D) := f_\theta(D)$$

in the case of continuous data. Here $p_\theta(D)$ is the probability of observing the data given the parameter θ in the discrete case, and $f_\theta(D)$ is the density function evaluated at the data in the continuous case. The *maximum likelihood estimate* (MLE) $\hat{\theta}$ is the maximizer of the likelihood function:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta \mid D).$$

Note that we consider the likelihood function as a function of θ with the data D fixed. This contrasts the interpretation of the probability distribution where the parameter is considered fixed and the random variable is the unknown (random) quantity.

In the case of i.i.d. sampling, so $D = X^{(1)}, \ldots, X^{(n)}$, the likelihood function factorizes as

$$L(\theta \mid D) = L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{i=1}^{n} L(\theta \mid X^{(i)}).$$

In the case of discrete data, this likelihood function is thus only a function of the vector of counts u, so that

$$L(\theta \mid X^{(1)}, \ldots, X^{(n)}) = \prod_{j} p_\theta(j)^{u_j}.$$

In the common setting in which we treat the vector of counts itself as the data, we need to multiply this quantity by an appropriate multinomial coefficient

$$L(\theta \mid u) = \binom{n}{u} \prod_j p_\theta(j)^{u_j},$$

which does not change the maximum likelihood estimate but will change the value of the likelihood function when evaluated at the maximizer.

It is common to replace the likelihood function with the *log-likelihood function*, which is defined as

$$\ell(\theta \mid D) = \log L(\theta \mid D).$$

In the case of i.i.d. data, this has the advantage of turning a product into a sum. Since the logarithm is a monotone function both the likelihood and log-likelihood have the same maximizer, which is the maximum likelihood estimate.

Example 5.3.6 (Maximum likelihood of a binomial random variable)**.** Consider the model of a binomial random variable with r trials. The probability $p_\theta(i) = \binom{r}{i}\theta^i(1-\theta)^{r-i}$. Given a vector of counts u, the log-likelihood function is

$$\begin{aligned} \ell(\theta, u) &= C + \sum_{i=0}^{r} u_i \log(\theta^i(1-\theta)^{r-i}) \\ &= C + \sum_{i=0}^{r} \left(iu_i \log\theta + (r-i)u_i \log(1-\theta)\right), \end{aligned}$$

where C is a constant involving logarithms of binomial coefficients but does not depend on the parameter θ. To calculate the maximum likelihood estimate, we differentiate the log-likelihood function with respect to θ and set it equal to zero, arriving at:

$$\frac{\sum_{i=0}^{r} iu_i}{\theta} - \frac{\sum_{i=0}^{r}(r-i)u_i}{1-\theta} = 0.$$

Hence, the maximum likelihood estimator, $\hat{\theta}$, is given by

$$\hat{\theta} = \frac{\sum_{i=0}^{r} iu_i}{rn}.$$

Note that $\frac{1}{n}\sum_{i=0}^{r} iu_i$ is the sample mean $\bar{X}$, so the maximum likelihood estimate of θ is the same as the method of moments estimate of θ.

The gradient of the log-likelihood function is called the *score function*. Since the gradient of a function is zero at a global maximum of a differentiable function, the equations obtained by setting the score function to zero are called the *score equations* or the critical equations. In many cases,

these equations are algebraic and the algebraic nature of the equations will be explored in later chapters.

For some of the most standard statistical models, there are well-known closed formulas for the maximum likelihood estimates of parameters.

Proposition 5.3.7. *For a multivariate normal random variable, the maximum likelihood estimates for the mean and covariance matrix are*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}, \qquad \hat{\Sigma} = \frac{1}{n}(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T.$$

Proof. The log-likelihood function has the form

$$\log(\mu, \Sigma \mid D) = -\frac{1}{2}\sum_{i=1}^{n}\left(m\log(2\pi) + \log|\Sigma| + (X^{(i)} - \mu)^T\Sigma^{-1}(X^{(i)} - \mu)\right).$$

The *trace trick* is useful for rewriting this log-likelihood as

$$\log(\mu, \Sigma \mid D) = -\frac{1}{2}\left(nm\log(2\pi) + n\log|\Sigma| + \operatorname{tr}(\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)\Sigma^{-1})\right).$$

Differentiating with respect to μ and setting equal to zero yields the equation

$$-\frac{1}{2}\sum_{i=1}^{n}\Sigma^{-1}(X^{(i)} - \mu) = 0.$$

From this we deduce that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$, the sample mean.

To find the maximum likelihood estimate for the covariance matrix we substitute $K = \Sigma^{-1}$ and differentiate the log-likelihood with respect to an entry of K. One makes use of the classical adjoint formula for the inverse of a matrix to see that

$$\frac{\partial}{\partial k_{ij}}\log|K| = (1 + \delta_{ij})\sigma_{ij},$$

where δ_{ij} is the Dirac delta function. Similarly,

$$\frac{\partial}{\partial k_{ij}}\operatorname{tr}(\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)K) = n(1 + \delta_{ij})s_{ij},$$

where $S = \frac{1}{n}\sum_{i=1}^{n}((X^{(i)} - \mu)(X^{(i)} - \mu)^T)$ is the sample covariance matrix. Putting these pieces together with our solution that the maximum likelihood estimate of μ is $\hat{\mu}$ gives that $\hat{\Sigma} = \frac{1}{n}(X^{(i)} - \hat{\mu})(X^{(i)} - \hat{\mu})^T$. □

Proposition 5.3.8. *Let $\mathcal{M}_{1\perp\!\!\!\perp 2}$ be the model of independence of two discrete random variables, with r_1 and r_2 states respectively. Let $u \in \mathbb{N}^{r_1 \times r_2}$ be the table of counts for this model obtained from i.i.d. samples from the model. Let $u_{i_1+} = \sum_{i_2} u_{i_1 i_2}$ and $u_{+i_2} = \sum_{i_1} u_{i_1 i_2}$ be the table marginals, and $n = \sum_{i_1,i_2} u_{i_1 i_2}$ the sample size. Then the maximum likelihood estimate for a distribution $p \in \mathcal{M}_{1\perp\!\!\!\perp 2}$ given the data u is*

$$\hat{p}_{i_1 i_2} = \frac{u_{i_1+} u_{+i_2}}{n^2}.$$

Proof. A distribution $p \in \Delta_{\mathcal{R}}$ belongs to the independence model if and only if we can write $p_{i_1 i_2} = \alpha_{i_1} \beta_{i_2}$ for some $\alpha \in \Delta_{r_1 - 1}$ and $\beta \in \Delta_{r_2 - 1}$. We solve the likelihood equations in terms of α and β and use them to find $\hat{p}$. Given a table of counts u, the log-likelihood function for a discrete random variable has the form

$$\begin{aligned} \ell(\alpha, \beta \mid u) &= \sum_{i_1, i_2 \in \mathcal{R}} u_{i_1 i_2} \log p_{i_1 i_2} \\ &= \sum_{i_1, i_2 \in \mathcal{R}} u_{i_1 i_2} \log \alpha_{i_1} \beta_{i_2} \\ &= \sum_{i_1 \in [r_1]} u_{i_1+} \log \alpha_{i_1} + \sum_{i_2 \in [r_2]} u_{+i_2} \log \beta_{i_2}. \end{aligned}$$

From the last line, we see that we have two separate optimization problems that are independent of each other: maximizing with respect to α and maximizing with respect to β. Remembering that $\alpha_{r_1} = 1 - \sum_{i_1=1}^{r_1 - 1} \alpha_{i_1}$ and computing partial derivatives to optimize shows that $\hat{\alpha}_{i_1} = \frac{u_{i_1+}}{n}$. Similarly, $\hat{\beta}_{i_2} = \frac{u_{+i_2}}{n}$. □

Unlike the three preceding examples, most statistical models do not possess closed form expressions for their maximum likelihood estimates. The algebraic geometry of solving the critical equations will be discussed in later chapters, as will some numerical hill-climbing methods for approximating solutions.

5.4. Hypothesis Testing

A hypothesis test is a procedure given data for deciding whether or not a statistical hypothesis might be true. Typically, statistical hypotheses are phrased in terms of statistical models: for example, does the unknown distribution, about which we have collected i.i.d. samples, belong to a given model, or not. Note that statisticians are conservative so we rarely say that we accept the null hypothesis after performing a hypothesis test, only that