Statistics 3858 : Statistical Models, Parameter Space and Identifiability

In an experiment or observational study we have data $X_1, X_2, \ldots, X_n$. These we view as a observations of random variables with some joint distribution.

Definition 1 *A statistical model is a family of distributions $\mathcal{F}$ such that for any possible n, a given distribution $f \in \mathcal{F}$ gives a joint distribution of $X_1, X_2, \ldots, X_n$.*

Note that f above may be either a joint pdf, pmf or cdf, depending on the context. Every $f \in \mathcal{F}$ must specify the (joint) distribution of X_i, $i = 1, \ldots, n$. Sometimes we use a subscript n, that is f_n, to indicate the dependence on the sample size n.

For a given sample size n, let f_n be the joint pdf of the random variables $X_i, i = 1, 2, \ldots, n$. Suppose the X_i's are iid with marginal pdf f. Then the joint pdf is of the form

$$f_n(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \ . \tag{1}$$

There is of course the analogue for iid discrete r.v.s. Notice also in the iid case the statistical model can also be viewed or described by the simpler one dimensional marginal distribution. That is by specifying the one dimensional marginal, say f, the one specifies the n dimensional joint pdf $f_n \in \mathcal{F}$.

Notice, in this iid case, a single one dimensional marginal f will specify the who sequence $f_1, f_2, f_3, \ldots$ of the n dimensional marginals of $X_1 \sim f_1$, $(X_1, X_2) \sim f_2$, $(X_1, X_2, X_3) \sim f_3$, ... In order to be technically more precise we would need to consider $\mathcal{F}$ as a set of sets, one of these subsets for each collection of distributions $\{f_1, f_2, f_3 \ldots\}$ that gives the lower n dimensional marginals of $X_1, X_2, X_3, \ldots$.

In this iid case we can simplify the description of the family $\mathcal{F}$ to the corresponding family of marginal distributions f. For example if X_i's are iid normal, then the marginal distributions belong to

$$\{f(\cdot; \theta) : \theta = (\mu, \sigma^2), \mu \in R, \sigma^2 \in R^+\} \ .$$

Shortly we will consider parameter spaces and so will not consider the formulation of a statistical model in a more precise form. For our purposes it is the specification of the joint distribution of $X_1, \ldots, X_n$ for all relevant sample sizes n.

In many dependent random variables cases we can also obtain their joint distribution. For example consider the so called autoregressive order one process, AR(1). It is defined iteratively as

$$X_{i+1} = \beta X_i + \epsilon_{i+1} \tag{2}$$

Specifically suppose that the r.v.s ϵ_i are iid $N(0, \sigma^2)$ and independent of the random variables up to time index less that i. Let f be the $N(0, \sigma^2)$ pdf. Then the conditional distribution of X_1 given that $X_0 = x_0$ is

$$f_{X_1|X_0=x_0}(x) = f(x - \beta x_0)$$

Similarly we have the conditional distribution of X_{t+1} given $X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t$, which by the Markov property is the same as the conditional distribution of X_{t+1} given $X_t = x_t$, given by

$$f_{X_{t+1}|X_t=x_t}(x) = f(x - \beta x_t)$$

This then gives the joint conditional pdf f_n of $X_1, X_2, \ldots, X_n$ conditioned on $X_0 = x_0$ as

$$f_n(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i - \beta x_{i-1})$$

In this case the joint distribution for any n is equivalent to knowing the initial condition x_0, the parameter β and the (marginal) distribution f of the random innovations ϵ. For example one may then talk about a normal autoregressive model with initial condition x_0 as a short hand for the statistical model (2) where ϵ_t are iid $N(0, \sigma^2)$. Notice there are two additional *parameters* β and σ^2.

One can also obtain for example the conditional distribution of $X_1, X_2, \ldots, X_n$ conditioned on $X_0 = x_0$ for a Markov chain. For this statistical model, in the case of a time homogenous Markov chain, one needs to specify the transition matrix P.

If one knows which distribution $f \in \mathcal{F}$ is the true distribution, then based on this one can make any probability statement, that is make statistical predictions of a future observation.

In the case of iid r.v.s one can then make predictive statements such as calculating $P(X_{n+1} \in [a, b])$ for any interval $[a, b]$, or statements such as calculating $\mathrm{E}(X_{n+1})$.

In the case of an AR(1) process, this means that one must know β, σ^2 and then one has the conditional distribution of X_{n+1} conditioned on $X_n = x_n$, the observed value. This allows one to then calculate the conditional expected value of the next observation

$$\mathrm{E}(X_{n+1}|X_n = x_n) = \beta x_n$$

or calculating the conditional probability of the next observation falling into a specified interval $P(X_{n+1} \in [a, b]|X_n = x_n)$.

In the AR(1) with normal innovations we can find the joint pdf of $X_1, \ldots, X_n$ given $X_0 = x_0$,

$$f_n(x_1, \ldots, x_n|x_0) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \beta x_{i-1})^2}{2\sigma^2}} \right\} .$$

The student should try to see why we have the formula above.

In the case of iid data, one can give the marginal distribution, and related statements from this, about the distribution of a future or new observation. This allows one to give the so called prediction interval of a new or future observation.

Without the idea of a statistical model, when we have observations $x_1, x_2, \ldots, x_n$ we would not be able to predict anything about a new or future observation.

The statistical question is the following : based on data $X_1, X_2, \ldots, X_n$ how can one estimate or guess which if any distribution in $\mathcal{F}$ is the correct one?

To work towards answering this question, we now consider parametric families, and simplify the question to estimation of these parameters.

Θ will be a set of numbers of d-tuples, that is $\Theta \subset R^d$. This will be called a parameter space for the family $\mathcal{F}$ is there is a one to one correspondence between Θ and $\mathcal{F}$.

Example 1 : Suppose X_i, $i = 1, \ldots, n$ are iid normal. The family of normal densities on R is then in a 1 to 1 correspondence with

$$\Theta = \{\theta = (\mu, \sigma) : \mu \in R, \sigma > 0\} = R \times R^+$$

Example 2 : Suppose X_i, $i = 1, \ldots, n$ are iid exponential. The family of exponential densities on R is then in a 1 to 1 correspondence with

$$\Theta = \{\theta : \theta > 0\} = (0, \infty) = R^+$$

Example 3 : Suppose X_i, $i = 1, \ldots, n$ are iid Gamma. The family of Gamma densities on R is then in a 1 to 1 correspondence with

$$\Theta = \{\theta = (\alpha, \lambda) : \alpha > 0, \lambda > 0\} = (0, \infty) \times (0, \infty) = R^+ \times R^+$$

Example 4 : Suppose X_i, $i = 1, \ldots, n$ are iid Binomial, size m. The family of Binomial pmf-s is then in a 1 to 1 correspondence with

$$\Theta = \{\theta : 0 \leq \theta \leq 1\} = [0, 1]$$

We may also further restrict in some examples the set of parameters to be not 0 or 1, that is

$$\Theta = \{\theta : 0 < \theta < 1\} = (0, 1)$$

This is because for example if $\theta = 0$, then with probability 1, all the random variables take on the value 0, which is not very interesting as a random process.

Another comment to note is that m is not a parameter in the same sense as θ, since m is chosen by the experimenter, whereas θ is typically not known, and hence the experiment is performed to estimate θ. The number m is called an ancillary parameter, and in not the object of our inference procedures.

Example 5 : Suppose X_i, $i = 1, \ldots, n$ are iid Poisson. The family of Poisson pmf-s is then in a 1 to 1 correspondence with

$$\Theta = \{\theta : \theta > 0\} = (0, \infty) = R^+$$

Example 6 : Consider the AR(1) process (2) with normal innovations. The joint distribution of f_n f_n of $X_1, X_2, \ldots, X_n$ conditioned on $X_0 = x_0$ is then in a one to one correspondence with

$$\Theta = \{\theta = (\beta, \sigma^2) : -1 < \beta < 1, \sigma^2 > 0\} = (-1, 1) \times R^+ \ .$$

There is a restriction for the parameter β, that is $|\beta| < 1$. This is discussed in a time series course and not further here.

Example 6 : Some distributions have infinite dimensional parameters. Consider a probability distribution on the non-negative integers. If is of the form

$$f(k) = a_k, k = 0, 1, 2, \ldots$$

and $f(k) = 0$ for all other numbers k, and where $a_k \geq 0$, $\sum_{k=0}^{\infty} a_k = 1$. In order to specify such a distribution on must specify infinitely many numbers a_k, $k = 0, 1, 2, \ldots$. Thus the family of distributions $\mathcal{F}$ on the non-negative integers requires a parameter space of dimension ∞.

Notice that for this family of distributions there are many special finite dimensional subfamilies of distributions, for example the Poisson family.

Example 7 : Consider the family of probability distributions on the set set $A = \{1, 2, 3, \ldots, m\}$ for a given positive integer $m \geq 2$. To specify a probability distribution on A we must specify numbers a_j such that

$$a_j \geq 0 \ , \text{ and } \sum_{j=1}^{m} a_j = 1$$

Thus we must specify $m - 1$ numbers, say $a_1, a_2, \ldots, a_m$ which are greater than or equal to 0 and sum to a value less than 1. The last value a_m is then determined as

$$a_m = 1 - (a_1 + a_2 + \ldots + a_{m-1})$$

Thus a parameter space to specify this family of distributions is of dimension $m-1$. A natural parameter space is

$$\Theta = \left\{ (a_1, a_2, \ldots, a_m) \ : \ a_j \geq 0, j = 1, 2, \ldots, m \text{ and } \sum_{j=1}^{m} a_j = 1 \right\}$$

This set is called the simplex of order m.

Definition 2 *We say a parameter is identifiable if for any two members of the family of distributions that are equal (as functions in their arguments), then the corresponding parameters are equal.*

Example 4 : An element of this family of distributions is a function which maps $k \in \{0, 1, \ldots, m\}$ to a real number by the formula

$$f(k;\theta) = \binom{m}{k}(1-\theta)^{m-k}\theta^k$$

and maps all other arguments to 0. Suppose that we take two such functions, say f_1 and f_2 determined by θ_1 and θ_2 respectively, and that $f_1 = f_2$. Thus for each $k \in \{0, 1, \ldots, m\}$ we have $f_1(k) = f_2(k)$, that is

$$\binom{m}{k}(1-\theta_1)^{m-k}\theta_1^k = \binom{m}{k}(1-\theta_2)^{m-k}\theta_2^k$$

In particular by taking $k = m$ we have

$$\theta_1^m = \theta_2^m$$

and hence $\theta_1 = \theta_2$. Thus the parameter θ is identifiable in the Binomial model.

The student should verify that the parameter is identifiable in a Poisson model.

Example 1 : Consider the normal model. Here a member of this family of distributions is of the form

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ , \ -\infty < x < \infty$$

where $\theta = (\mu, \sigma^2)$. For θ_1 and θ_2 consider the two corresponding normal pdf-s f_1 and f_2 respectively. Suppose $f_1 = f_2$, that is for all arguments x we have $f_1(x) = f_2(x)$. Can we conclude that $\theta_1 = \theta_2$?

$$\begin{aligned} f_1(x) = f_2(x) \quad &\Leftrightarrow \quad \frac{1}{\sqrt{\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{\sqrt{\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \\ &\Leftrightarrow \quad e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}} = \frac{\sqrt{\sigma_1^2}}{\sqrt{\sigma_2^2}} \\ &\Leftrightarrow \quad \frac{-(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2} = \frac{1}{2}\left(\log(\sigma_1^2) - \log(\sigma_2^2)\right) \end{aligned}$$

Since this last line holds for all x, the LHS is a polynomial of degree 2 in x and the RHS is constant with respect to x, then the coefficients of x and x_2 must be equal to 0.

An alternative method is to note that LHS is differentiable with respect to the argument x, and hence the second derivative of LHS must equal the second derivative of RHS which is 0. We then conclude

$$\text{coefficient of } x^2 := -\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} = 0$$

and hence $\sigma_1^2 = \sigma_2^2$. Next

$$\text{coefficient of } x := \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} = 0$$

and we conclude $\mu_1 = \mu_2$.

Aside : If we used a parameter space $\Theta = \{(\mu, \sigma) : \mu \in R, \sigma^2 > 0\}$, the parameter would not be identifiable, since both (μ, σ) and $(\mu, -\sigma)$ would give the same normal pdf.

Aside : In the binomial example (Example 4) we could not use the polynomial or differentiability property as the pmf is not even continuous in real arguments, and in particular not differentiable. To see this the student should sketch the function $f(x;\theta)$, with $\theta = \frac{1}{2}$, where

$$f(x;\theta) = \begin{cases} \binom{m}{x}(1-\theta)^{m-x}\theta^x & \text{if } x = 0, 1, \ldots, m \\ 0 & \text{otherwise} \end{cases}$$

In addition we could not use the polynomial coefficients property for the function $f(x;\theta)$ to identify θ, because f is not a polynomial in real argument x.

The student should now verify that the Gamma model is identifiable. Can you use differentiability with respect to the argument x in this case?

The student should verify that the bivariate normal model is identifiable. Notice the pdf is a function of two arguments. Can you use properties of polynomials or differentiability to verify parameter identifiability?

Are there cases where the parameter is not identifiable? When the model and parameters are properly formulated the answer in general is no, but sometimes there are additional constraints. For the iid models we consider in this course the parameters will always be identifiable. For time series model there are sometimes additional restrictions or constraints on the parameter space (consider the AR(1) Gaussian model discussed above where it is required that $-1 < \beta < 1$). In some multivariate time series models it has taken some years in order to determine that the parameters are identifiable. Another model known as a competing risks model gives a model of in which there are several diseases, each with a potential *death* time $\tau_1, \ldots, \tau_K$, and the patient then dies at the minimum of these potential death times, $T = \min(\tau_1, \ldots, \tau_K)$. When a patient dies the cause $1, \ldots, K$ is not known or at least the other potential death times are not observed. The distribution of T does not allow one to identify the parameters of the distributions of the other r.v.s $\tau_1, \ldots, \tau_K$, except in the case that the $\tau_1, \ldots, \tau_K$ are independent.

In the area of experimental design there is also non-identifiability. It is for this reason there are often additional parameter restrictions, for example sums of certain parameters equal to 0, and the issue of confounding.