

Introduction to Statistics

IST, the slides, set 1

Introduction to Estimation Theory

Richard D. Gill

Dept. of Mathematics
Leiden University

September 18, 2009

Statistical Models

Data, Parameters: Maths

Data, Parameters: Less Maths

The i.i.d. Case

Statistics and Estimators

Statistics and Estimators are functions of the data

Estimation Theory: Good Estimation

Statistics and Cookery

Parameters and Parametrizations

The Chef Discusses Cookery

Statistical Models: Data, Parameters: Maths

A statistical model is a probability model $(\Omega, \mathcal{F}, P_\theta)$ in which the probability measure P_θ is indexed by an unknown *parameter* $\theta \in \Theta$, together with a specification of the observed *data* \mathbf{X} – a random variable on the probability space, with values \mathbf{x} in some outcome space $(\mathcal{X}, \mathcal{B})$

Almost all readers can safely forget about the specifications of the collections of “events” \mathcal{F} and \mathcal{B}

But for completeness, \mathcal{F} is a nice collection of nice subsets A of the sample space Ω , and \mathcal{B} is a nice collection of nice subsets B of the outcome space \mathcal{X} . These are simply those sets $A \subseteq \Omega$ and $B \subseteq \mathcal{X}$ for which it makes sense to talk about $P_\theta(A)$ and $P_\theta(\mathbf{X} \in B) = P_\theta(\{\omega : \mathbf{X}(\omega) \in B\})$, respectively. Students of measure theory will know what I mean by “nice”, and will also realise that we can usually safely forget about these qualifications.

So I'll do this slide all over again (“next slide please!”)

Statistical Models: Data, Parameters: Less Maths

A statistical model is a probability space Ω on which the probability measure P_θ is indexed by an unknown parameter $\theta \in \Theta$, together with a specification of the observed data \mathbf{X} – a random element on the probability space with values \mathbf{x} in some outcome space \mathcal{X}

Statistical Models: Data, Parameters: Less Maths (cont.)

The idea behind this is that we *observe* the values taken by a collection of random variables defined on this probability space, collected into a random vector or matrix (or whatever) \mathbf{X} . Sometimes \mathbf{x} is called *the data*, sometimes *the dataset*

More precisely (and actually, without loss of generality), \mathbf{X} is a list, or a list of lists, or . . . , of ordinary random variables, i.e., random variables taking values in the set of real numbers $\mathbb{R} = (-\infty, \infty)$, or extended real numbers $\overline{\mathbb{R}} = [-\infty, +\infty]$, or in some subset thereof. Examples are: the nonnegative reals $[0, \infty)$; the integers \mathbb{Z} , the nonnegative integers including $+\infty$, $\overline{\mathbb{N}}$; the numbers $\{0, 1, \dots, 10\}$

There is sometimes disagreement as to whether the *natural numbers* include 0 or not

Statistical Models: Data, Parameters: Less Maths (cont.)

Both \mathbf{X} and a possible realization thereof, \mathbf{x} , are often called *the data*. Most often \mathbf{X} is a real random vector or a real random matrix, i.e., a vector or a matrix of ordinary (possibly extended real valued) random variables defined on the sample space (Ω, \mathcal{F}) . Sometimes it is just one ordinary real random variable! Sometimes it is something of *much* more complicated structure or shape

We see a realization $\mathbf{X}(\omega) = \mathbf{x}$, but we don't know which $\theta \in \Theta$ underlies the probability measure P_θ on Ω which actually “directed” the *Goddess Fortuna* in her choice of $\omega \in \Omega$. We don't even get to see – or at least, for the moment we are forgetting we saw it – the actual choice ω , otherwise we would have called ω the observed data

I am here assuming that the model is *true* or *correct*. We can also study the case that the actual probability measure which directed Fortuna is not one of the P_θ 's but some other probability measure, say Q

Statistical Models: Data, Parameters: Less Maths (cont.)

Supposing \mathbf{X} takes values \mathbf{x} in some space \mathcal{X} with σ -algebra \mathcal{B} of measurable sets $B \subseteq \mathcal{X}$, the only relevant things for us are the induced probability distributions $P_\theta^{\mathbf{X}}$ of \mathbf{X} , defined by $P_\theta^{\mathbf{X}}(B) = P_\theta(\mathbf{X} \in B)$, $B \subseteq \mathcal{X}$, $\theta \in \Theta$

Thus we are interested in the induced indexed family of probability spaces $(\mathcal{X}, \mathcal{B}, (P_\theta^{\mathbf{X}} : \theta \in \Theta))$, where we get to observe which realized dataset, a point $\mathbf{x} \in \mathcal{X}$, has been chosen by Fortuna, whose choice was made under the law of the probability measures or distributions $P_\theta^{\mathbf{X}}$. We don't know which θ , through the distribution $P_\theta^{\mathbf{X}}$, actually directed Fortuna in her choice of the actually observed \mathbf{x}

In many cases, these probability **distributions** are continuous distributions on $\mathcal{X} = \mathbb{R}^n$ for some n , i.e., have probability **densities** $f(\mathbf{x}; \theta)$. In many other cases, they are discrete, i.e., concentrated on some (at most) countable set of outcomes, and therefore have probability **mass functions** $p(\mathbf{x}; \theta)$

Statistical Models: Data, Parameters: Less Maths (cont.)

We will usually not have to distinguish between these two cases. In the sequel we'll often write p or f , whether we are talking about densities, mass functions, or both ...

Indeed, from the point of view of measure theory, both an ordinary multivariate probability density, and an ordinary discrete probability mass function, are *densities* of the corresponding probability measure with respect to some reference measure (ordinary n -dimensional *hypervolume measure*, and ordinary *counting measure* respectively).

**A statistical model is therefore often specified by merely saying:
“Suppose the data $\mathbf{X} = \mathbf{x}$ has density $f(\mathbf{x}; \theta)$ on \mathcal{X} , for some $\theta \in \Theta$ ”.**

Or we might say (“next slide please”)

Statistical Models: Data, Parameters: In a Nutshell

A statistical model for observed data $\mathbf{X} = \mathbf{x}$,
with unknown parameter θ is defined by

$$\mathbf{X} \sim f(\mathbf{x}; \theta), \quad \mathbf{x} \in \mathcal{X}; \quad \text{for some } \theta \in \Theta$$

Example: One observation from an unknown normal distribution

$$Y \sim \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}\right), \quad y \in \mathbb{R}; \quad (\mu, \sigma) \in (-\infty, \infty) \times (0, \infty)$$

The iid Case

If $\mathbf{X} = (X_1, \dots, X_n)$ where the X_i are independent and identically distributed random variables (or vectors or ...) then everything depends on the *sample size* n and the family of probability distributions of just one of the X_i 's, which (apparently) depends on some parameter $\theta \in \Theta$.

Often these probability distributions have a density or mass-function $f(\cdot; \theta)$. We write X to stand for any of the X_i and x for a possible value of X . Our data $\mathbf{X} = \mathbf{x}$ is of the form $\mathbf{x} = (x_1, \dots, x_n)$ and \mathbf{X} has a (joint) density or mass-function $\prod_{i=1}^n f_X(x_i; \theta)$.

Since f often stands for a generic probability density, and since now there are an awful lot of probability densities in the picture, I have started to append an index to f , in order to indicate whose density I'm talking about.

Statistical Models in the iid case

A statistical model often comes down to consideration of a family of probability densities (or mass functions) of data \mathbf{x} , indexed by a parameter θ , $f(\mathbf{x}; \theta)$.

In the i.i.d. case, $\mathbf{x} = (x_1, \dots, x_n)$, and

$$f(\mathbf{x}; \theta) = f_{\mathbf{X}}(\mathbf{x}; \theta) = \prod_{i=1}^n f_{X_i}(x_i; \theta) = \prod_{i=1}^n f_X(x_i; \theta).$$

Statistics and estimators:

Estimands, estimators, estimates

A *statistic* T is a function of the data \mathbf{X} , taking values in some measurable space \mathcal{T} .

Thus $T = T(X) = T(X(\omega))$; i.e., we can think of T both as a function defined on the outcome space of \mathbf{X} ; and as a function defined on Ω ; and hence we can think of T both as a random variable (or vector or ...) defined on $(\mathcal{X}, \mathcal{B}, P_\theta^{\mathbf{X}})$ or on $(\Omega, \mathcal{F}, P_\theta)$.

An *estimator* of θ is a statistic $\hat{\theta}$ taking values in Θ .

A realized value of $\hat{\theta}$, which is of course just some point in Θ , might often be denoted by the same symbol; but it has a different name: it is then called an *estimate*. What we are estimating, θ , is called the *estimand*.

Estimation Theory: The Art of Good Estimation

A main task of estimation theory is to tell us how to find good estimators. But what is a good estimator?

We suppose that the data is generated by one particular but *unknown* value of θ , often denoted θ_0

Thus $\hat{\theta}$ is a random variable taking values in Θ , and having a probability distribution depending on θ_0

An estimator is thus a *recipe* for converting data \mathbf{x} into an estimate $\hat{\theta}$

A good estimator ends up lying close to θ_0 with large probability; *whatever* θ_0 might actually be

Statistics and Cookery

As is well known, if you want to tell lies you hire a statistician to cook the numbers for you

Estimation theory is about recipes for recipes. In other words, meta-cookery (meta-lying)

Anyone can use a cookery book. Only a good theoretical statistician can keep on inventing good new recipes; especially when for some reason or other she has to use a new kind of ingredients. Top statisticians are like five star chefs. This class is a beginners' class in top-chef-school

The proof of the pudding is in the eating. We investigate whether $\hat{\theta}$ is usually close to θ_0 , whatever θ_0 might be, or not. If it is, then this statistic does not lie. Still, if another estimator is usually closer, then the first, though telling the truth, is not telling the whole truth. The best estimator tells the truth, only the truth, and the whole truth, as much as is mathematically possible, so help me Goddess Fortuna.

Statistics and Cookery (cont.)

Unfortunately, you can please some of the people all of the time, or all of the people some of the time, but not all of the people all of the time. Tastes differ. Some estimators are better when θ_0 is somewhere over here, but worse when actually θ_0 is over there. There is almost never an estimator which is absolutely best, independent of θ_0 . We have to make some kind of compromise. By definition, there is no best compromise. It's a matter of taste, of negotiation.

Parameters and Parametrizations

There are parameters and parameters

Often we are interested in estimating some component of a vector parameter θ , or indeed, any other function of θ . If $\phi = \phi(\theta)$ is this so-called *parameter of interest*, then we often call ϕ itself “a parameter”; and a statistic $\hat{\phi}$ which takes values in the same space as ϕ can be called “an estimator of ϕ ”

An important point is that we can often usefully think of estimating some feature of P_θ , i.e., some function of θ , without first figuring out “how to estimate θ ”. If however $\hat{\theta}$ is an estimator of θ and $\phi = \phi(\theta)$ is the parameter of interest, then $\hat{\phi} = \phi(\hat{\theta})$ is called the *plug-in* estimator of ϕ

Not all good estimators of a parameter ϕ are plug-in estimators. Some good estimators are plug-in estimators based on lousy estimators of θ

Interest parameters and nuisance parameters

If we can write $\theta = (\phi, \psi)$ where we are only interested in ϕ then we call ϕ the interest parameter and ψ the nuisance parameter

Parameters and Parametrizations

Replacing θ by some one-to-one function of θ is called a *reparametrization* of the model. The model is really the same, we just happen to describe it differently. At the most abstract level, a statistical model is just a particular family of probability distributions of \mathbf{X} . These distributions themselves form a parametrization of the model. Parameters of interest are interesting functions (some might say: functionals) of the unknown probability distribution. Nuisance parameters are everything else which needs to be specified in order to fix the probability distribution. Unidentified parameters are functions of θ which can't be expressed as functions of $P_{\theta}^{\mathbf{X}}$. Identified parameters can be.

The Chef Discusses Cookery

Back to basics

This is where the underlying probability space Ω turns up again. We often define models by describing probability mechanisms which lead to the observed data, but which depend on plenty of unobserved things. In fact we are often most interested in those unobserved things, or rather, features of their distribution. If we “did the wrong experiment” however, the parameters of interest won’t be functions of the distribution of the data, though they did appear in the definition of the “underlying probability model of the phenomenon of interest”.

The Chef Discusses Cookery

A statistician's life is interesting

Often in statistics there are two parts of the underlying probability model: the part corresponding to the phenomenon of interest, and the part corresponding to what we get to see of it in our experiment, our sample, whatever... Thus our experimental design, sampling design, observational frame, whatever ..., can introduce more randomness and more unknown parameters into the distribution of the data. It can make our life more easy or make our life more difficult, especially depending on whether we are smart or dumb (if we are lucky enough to have any choice in the matter)

This is certainly the thing that makes a statistician's life *interesting*

An ancient Chinese curse is: “may you live in interesting times”

The Chef Discusses Cookery

A mathematician's apology

Pedantic mathematicians object to our habit of using the same symbol to stand for completely different things at the same time, e.g., both for a function and for a possible value of that function. We also often use “illegal” (lazy physicists’) notation in which the *name* for an object determines its *nature*; in which the same symbol is used for different objects but where the *context* in which it appears, hopefully determines which “variety” is meant.

We realise that this might be a subjective criterion. I shall be even more wicked still: attach whichever subscripts or superscripts to objects I think are important to emphasize, and leave them off again after a few slides when we all realise what we are talking about. *Arguments* will change into *indices* (sub- or superscript, for no particular reason except local aesthetics), and later disappear altogether

One has the choice of either writing out everything explicitly which means a huge burden of complex notation, which might actually obscure the intended meaning, or of using various kinds of unofficial shorthands, which allow the insiders to quickly see what is meant, but can be impossible for outsiders to decipher.

The Chef Discusses Cookery

A statistician's exhortation

In this respect, statistics is much like physics: we abuse mathematics as much as we use it. One person's laziness or carelessness is another person's efficiency or clarity. Explicitness can obscure meaning, just as much as implicitness can obscure meaning. We always have to make a compromise.

At the end of the day, we just have to get used to the language which after a long cultural development is at this moment the most convenient compromise for some subset of "insiders". In a hundred years' time it will no longer be a good compromise any more.

C'est la vie!