Lecture 03: Parametric Statistical Models

03.01. Motivation

03.02. Objectives

03.03. The goals of statistics

Trinity of statistical inference:

- **Estimation**: A single number
- **Confidence intervals**: The bars around this number
- **Hypothesis testing**: Yes/no answer (all we care about)

03.04. Statistical modelling

Statistical model: The mathematical framework to be able to answer the “trinity of statistical inference” questions.

The goal of statistics is to learn the distribution of X.

In some cases the type of observation will dictate the distribution, and there is no modelling framework, e.g. when obs are binary (i.e. {0,1}s) the only “possible” distribution is Bernulli. Otherwise we need to make reasonable assumptions on the underlying distribution.

In case of discrete observations, we could try to estimate the PMF directly, of each observation. But that would require many observations, as I have to estimate many parameters. I can instead assume an underlying distribution (like the Poisson for the number of sibling example) and then pool all my observations to contribute to learning one parameter rather than dividing them, and each part contributes to estimate a different parameter.

The more parameters you have to learn, the more observations you're going to need to do that. The rule of thumb is if you need to learn two parameters, you need twice as much observations.

Aside to the parameter, we will also be able to test if the data arise from a given distribution, whatever its parameters are.

03.05. Statistical model

A model just means something which is like slightly simpler than what reality actually is, but hopefully captures most of it.

There isn't one correct model. The task of a statistician is to use reasonable assumptions to find a tractable model that gives *useful* approximation to a given data set.

Note that *useful* will depend on the question that is asked for, the research question.

Statistical experiment: collecting the observations of the random variable of interest (we will assume the observations are i.i.d.). Something that generates data.

Sample: the set, collection of the observations

We can now construct a **statistical model** $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$ associated with this experiment, where

- E is a sample space for X, i.e. a set that contains all possible outcomes of X;
- $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ is a family of probability distributions on E;
- Θ is a parameter set, i.e. a set consisting of some possible values of θ, the parameter space.

E, the space on which you know your random variables live, is typically $\in \mathbb{R}$. But you should be as precise as possible, for example positive integers is more precise than R. We associate a random variable with its smallest possible sample space

In statistics (as opposed to probability theory), the term sample space, denoted in this course by E, is always the range of the random variable X_i, so that $X_i(\omega) \in E$ for every realization of uncertainty $\omega \in \Omega$. Because statistics deals with observations, we practically never think or talk about the domain of the random variables that model our data sets.

$\mathbb{P}$ is the "true" distribution family. We have to chose it based on the kind of obs we have, e.g. integer values we *can* use Poisson, but we could have used others.

Basically the statistical model is made by the *pair*

A statistical model is a pair (sample-space, probability-family). But it's really three things:

- It's the sample space;
- It's the form of the distribution as a function of theta;
- And it's also the candidate thetas that I want you to have.

[0,1] denotes the closed interval between 0 and 1. In contrast, {0,1} denotes the set with two elements, 0 and 1.

03.06. Types of Statistical Models

The model is really something that's telling you how you want to think of your distributions as possibly being, like we said, for example, Poisson.

We will *assume* that the model is well specified:

Well specified model: there exists a Θ such that P is equal to P_θ. I.e. P is the class of models that I am considering.

A *misspecified* model is one P, the true prob distribution generating my data, is not in $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$: Either is an other family of distribution of the real parameter θ is not within the Θ we are assuming.

You can study statistical models in both cases, but in this class, we're going to focus on the case where the model is well specified.

This particular theta, the one that actually corresponds to the true distribution you're looking for is called the true parameter. It's unknown, and the goal of statistical inference will be to estimate it, maybe form some error bars around it, and finally check some properties of it, like being different than zero, or different than 0.5 (e.g. does people have a preference for a particular direction when they turn they head for kissing?), or… whatever is the experiment about.

Notation: θ is the parameter for an abstract probability model, but for specific examples its name depend on the distribution: in Bernulli we refer to it as p, in Poisson and exponential it is λ, in Gaussian is μ, etc…

While the sample space is where my random variables, my observations, live, the *parameter space* is where my parameter lives.

The sample space **shoudn't** depend on any parameter.

Parameteric model: we assume $\Theta \in R^d$ with $d \geq 1$ but finite, that is, is a vector of dimensions d.

Non-parametric model: The parameter space Θ is of infinite dimensions. It doesn't mean there is not parameter. It requires more work than a parametric model.

- We'll touch a bit of it when we'll discuss regression, but most of the time we'll deal with parametric models.

Semi-parametric model: We decompose Θ in $\Theta = \Theta_1 X \Theta_2$, where Θ_1 is a finite dimensionalal and Θ_2 is infinite dimesional.

- We care only to the finite-dimensional parameter and treate the inifinite dimensional one as a *nuissance*.
- We will not touch them in this class
- ex. $\Theta = p * f$, where p is the Bernoulli probability (the one-dimensional parametric set made by all numbers between 0 and 1) and f is a function, and function typically lives in an infinite dimensional space. Note that it doesn't matter if p is countable, non-contable or infinite. What it matters is that it is finite *dimensions* (one in this case).
- Even if you care only of the finite parameter, you still need to take into account the infinite-dimension parameter: typically, what you do is you just look at the worst possible function that nature could have given you, and you try to learn this model based on this.
- They arise in survival analysis and econometrics. We'll see one example, the Cox proportional hazard model.

03.07. Examples of Parametric Models

(but actually also non-parameteric models are treated)

Notation:

- N = Natural numbers (all positive integers starting from 1)
- Z = Integers (all integers, positive and negative)
- Q = Rational numbers (numbers that can be written as ratio)
- R = Real numbers (including irrational numbers as √2, π or e)

We'll write some statistical model.

Bernulli trials (e.g. Kiss example):

- $E = \{0, 1\}$
- Candidate probability distribution family is: is Ber§
- $\Theta =]0, 1[$

Poisson (e.g. number of siblings):

- $E = N$
- Candidate probability distribution family is: is Poisson(λ)
- $\Theta = R^+$

Gaussian, where I don't know both the mean and the variance

- The parameter is now the *pair* (mean, variance)
- $E = R$
- Candidate probability distribution family is: is $Gauss(\mu, \sigma^2)$
- $\Theta = RXR^+$ (note it's a cross, cartesian product between two spaces)

d-dimensional Gaussian $\mathcal{N}_d(\mu, I)$

- The variance is known and equal to the Identity matrix (1s on the principal diagonal and 0s otherwise)
- $E = R^d$

- Candidate probability distribution family is: is $N_d(\mu, I)$
- $\Theta = R^d$
- We'll see it in multivariate regression and PCA
- In this case the variance is *not* a parameter. If it wasn't given it would have been the space of all matrices that are valid covariance matrices (i.e. positive definite matrices).

Non-parametric model

- you put some assumptions on the underlining distribution.
- PDF: it's a integrable function that integrates to 1 and it is non-negative
    - if I don't tell you anything more than that, I'm really not telling you anything: there's no modeling happening. I'm just saying my random variable, I model as well, a random variable.
    - but if I start telling you something specific about what this pdf looks like, without going all the way to saying it's a pdf of a Gaussian with a known parameter, but something that say something about its shape, for example, I am doing a non-parametric model

A common shape that people use is the unimodal. What does unimodal look like? Well, unimodal is a function that increases, but I don't know how this happen, on $(-\infty, a)$, and then decreases (again, I don't specify how) on $(a, +\infty)$.

It's a large chunk of possible probability distribution (e.g. the Gaussian is one), but is a model that's enough for me to actually learn something from data.

$X_N \in R$ **have an unknown unimodal pdf** f:

- $E = R$
- Θ = {unimodal pdfs} (the set of unimodal distributions)
- Candidate probability distribution family is: is $P_\theta = P_f$ - the PDFs with unimodal pdf f

$X_N \in [0, 1]$ **with an unknown invertible cdf** F:

- $E = [0, 1]$
- Θ = {all invertable CDF whose pdf integrates in [0,1] to 1}
- Candidate probability distribution family is: is $P_\theta = P_F$ - the invertable CDF whose pdf integrates in [0,1] to 1

03.08. Exercises on Statistical models

X is an exponential random variable, Y is an indicator variable if X > 5. Y is a **censored version** of the Exponential random variable X: we cannot directly observe X, but we are able to gather some information about it (in this case, whether or not X is larger than 5.)

03.09. Further examples

Linear regression model

$Y_i = \beta^T X_i + \epsilon_i$

with $\epsilon_i \sim N(0,1)$ and $X_i \sim N_d(0, I_d)$

We will come back to it

In the linear regression model, each observation is not a number or a vector. It's a pair, where the first guy is a vector in R^d, and the second guy is a vector in R. The first one is say, the input variable. The second one is the output variable.

For example the observation Y_i is a cardiovascular risk index and X_i is a vector of the individual characteristics that influence the risk, like weight, age, sex…

I want to go from a random vector in our R^d to a number in R. And so the way I'm going to do it is in the simplest possible way. I'm going to actually make it a linear function, something that just takes my x_d, my x_i and just takes the inner product with some unknown vector $beta$.

This is the simplest possible way I can go from this guy to this guy.

And then I'm going to add some Gaussian noise, for example. It's unlikely that the model is going to be right. So I'm going to allow for some Gaussian noise. Now in this case, the parameter I don't know is beta in R^d and my x_i's for example, I'm going to assume that they have some multivariate Gaussian distribution, just like the one we saw before, except that I'm going to tell you ahead of time that means even 0 in R^d. So this is the standard linear Gaussian linear regression model.

- $E = R^d \times R$, i.e. R^{d+1} (all possible input observations in d dimensions plus the output)
- Θ = R^d$ (the vector of unknown betas)

We are not imposing anything more on Θ, but a huge amount of work in linear regression is to actually impose extra structural constraints on beta, as to srink Θ .

P_θ is actually the distribution of the pair (x,y). That is the distribution of X and the conditional distribution of Y given X

For the CDF(X,Y) I simply multiply them: $N_d(o, I_d) * N(x^T\beta, 1)$

(the conditional distribution of Y given X has the average $x^T\beta$ - this is just a number in the conditional form, plus the ϵ r.v. with variance = 1, this is why it has variance 1, as it is the sum of a r.v. with a constant).

Cox proportional Hazard model

Again couples (X_i, Y_i) where X_i is a vector and Y_i is a scalar. The conditional distribution of Y given X has CDF F of the form

$$F(t) = 1 - e^{-\int_0^t h(u) e^{\beta^T x} du}$$

where h is a unknown non-negative nuissance function and $\beta \in R^d$ are the parameter of interest.

This is a model that arises a lot in survival analysis, returning the probability that people live longer than some time t.

P(X>t) = 1-cdf(t)

What you're trying to do is to see which of those guys–which of those variables that you're observing – actually impacts their survival. And that's what the Cox Proportional Hazard model is trying to understand. Positively, negativly or they are irrelevant.

03.10. Identifiability

Injectivity

The notation $f : S \to T$ denotes that f is a function, also called a map, defined on all of a set S and whose outputs lie in a set T. A function $f : S \to T$ is injective if for all $(x, y) \in S$, $f(x) = f(y)$ implies that $x = y$.

Alternatively: a function is injective if we can uniquely recover some input x based on an output $f(x)$.

A parabola is *not* injective, as it has two x that corresponds to the same y, so you can not invert, i.e.e recover x given y.

Here we're not actually talking about a model. WE are talking about a parameter. The parameter is said to be **identifiable** if and only if the map $\theta \in \Theta$ maps through P_θ is injective.

$$\theta \neq \theta' \implies P_\theta \neq P_{\theta'}$$

or, equivalently,

$$P_\theta = P_{\theta'} \implies \theta = \theta'$$

With two different parameters I will obtain two different distributions.

Clearly, that's a desirable property: if the parameter one was giving me the same distribution, meaning the same observations as the parameter two, there's no way, even with an infinite number of observations, that I could actually decide whether my parameter is one or two.

All previous examples were identifiable.

If X_i is an indicator function for $Y_i \geq 0$ and $Y \sim N(\mu, \sigma^2)$, observing just X, μ and σ^2 are not identifiable, but $\theta = \mu/sigma$ is.

X just tell me if the normal is positive or not. It is Bernulli with prob p equal to $N(\mu, \sigma^2) > 0$, i.e. $1 - \Phi(\frac{0-\mu}{\sigma})$. So, $\frac{\mu}{\sigma} = -\Phi^{-1}(1-p)$.

if I have a Bernoulli, all I'm going to be able to get is one single number. That's the only thing that determines the Bernoulli. I cannot just split into two numbers just magically.

The model that I have is reasonable?

- Is it well-specified?
- Is it identifiable?

The formal mathematical way to prove that a model is not identifiable is to find two sets of parameters that results the same distribution.

If I were to actually look at the parameter set, now, I start to have a choice. I could say, oh, I want μ, σ^2 to belong to $R \times R^+$. Or I could say, I just want $\mu \ \sigma$ to belong to R. And we know that if I write the first one, I end up with an identifiable model. But if I take the second one, I don't. So the first one is not the one I want.

03.11. Identifiability exercises