18.650
Statistics for Applications

Chapter 2: Parametric Inference

The rationale behind statistical modeling

- Let $X_1, \ldots, X_n$ be n independent copies of X.
- The goal of statistics is to learn the distribution of X.
- If $X \in \{0, 1\}$, easy! It's Ber(p) and we only have to learn the parameter p of the Bernoulli distribution.
- Can be more complicated. For example, here is a (partial) dataset with number of siblings (including self) that were collected from college students a few years back: 2, 3, 2, 4, 1, 3, 1, 1, 1, 1, 1, 2, 2, 3, 2, 2, 2, 3, 2, 1, 3, 1, 2, 3, ...
- We could make no assumption and try to learn the pmf:

x	1	2	3	4	5	6	≥ 7
$\mathbb{P}(X = x)$	p_1	p_2	p_3	p_4	p_5	p_6	$\sum_{i \geq 7} p_i$

  That's 7 parameters to learn.
- Or we could assume that $X \sim$ Poiss(λ). That's 1 parameter to learn!

Why we are doing Statistical modelling?

in Practice, if we have data, if we observe a Bunch of Points, There is many way's we can think of this list of number's. Like discrete, or continuous, or ≥ 1, .etc..

what statistical modelling is doing is to try to compress this information that could actually describe in a very naive way.

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} P$ (Same distribution)

The Goal of statistical modelling is to understand what is the distribution (P) we actually have here?

* Now we can start making Assumption's on this distribution P

  Ex: P is Pdf, Smooth (Assumption)

  these Assumption's make our Life Simpler, Because By making Successive Assumption's we are reducing the D.O.f of this space of distribution's

Ex: Assumption : if $x_i \in \{1, 0\} \Rightarrow x_i \sim Ber(p)$

↓

we just need to find p

Ex: Assumption : if $x_i \in \mathbb{N} \Rightarrow x_i \sim Possion(\lambda)$

we need to find λ

Assumption: $x_i \in \{1, 2, 3, 4, 5, 6, \geq 7\}$

$\Rightarrow$ discrete distribution

$\Rightarrow$ $x_i \sim$ PMf

x	1	2	3	4	5	6	≥ 7
$P(X=x)$	$P(x=1)$ p_1	$P(x=2)$ p_2	$P(x=3)$ p_3	. p_4	. p_5	. p_6	p_7

we are trying to estimate these values

The Ultimate goal of statistic is to say what distribution our data come from? Because that's basically the best we are going to be able to do.

Modelling 101: The purpose of modelling is to restrict the space of possible Distributions to a subspace that's actually plausible, but much simple for us to estimate.

Statistical model (1)

Formal definition

Let the observed outcome of a statistical experiment be a sample $X_1, \ldots, X_n$ of n i.i.d. random variables in some measurable space E (usually $E \subseteq \mathbb{R}$) and denote by $\mathbb{P}$ their common distribution. A *statistical model* associated to that statistical experiment is a pair

$$(E, (\mathbb{P}_\theta)_{\theta \in \Theta}),$$

where:

- E is *sample space*; The set in which X lives
- $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is a family of probability measures on E;
- Θ is any set, called *parameter set*. The set in which θ lives.

Statistical Experiment:

It's a pair $\left(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta}\right)$

$\mathbb{P}_\theta$: Probability distribution

$\Downarrow$

$\theta = (\mu, \sigma^2) \Rightarrow N(\mu, \sigma^2)$ (2 dim)

$\theta = \{p_1, p_2 \dots, p_6, p_{>7}\}$ (7 dim)

$\text{Pois}(\theta)_{\theta \in \mathbb{R}}$

What's important here is once they gave us θ we know exactly all the probabilities associated with random variable. we know its distribution perfectly.

* The purpose of the statistical model is to once we estimate the parameter, we actually know exactly what distribution it has.

Statistical model (2)

- Usually, we will assume that the statistical model is *well specified*, i.e., defined such that $\mathbb{P} = \mathbb{P}_\theta$, for some $\theta \in \Theta$.

- This particular θ is called the true parameter, and is unknown: The aim of the statistical experiment is to *estimate* θ, or check it's properties when they have a special meaning ($\theta > 2$?, $\theta \neq 1/2$?, ...)

- For now, we will always assume that $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$: The model is called *parametric*.

Two things

① The purpose of the statistical model is once we Estimate the parameters, we Exactly know what distribution it has

② we could potentially have several parameter's that give us the same distribution, (that would still be fine, because we could estimate one guy, or we could estimate the other guy, we would still recover underlying distribution of our data

The problem is that this could create really annoying, theoretical problem's, like things don't work, the algorithm's wont work. etc..

what we typically Assume is the model is identified.

Vocabulary:

Well specified: for our observation's X

there exists $\exists \theta \in \Theta$ such that X follows $\mathbb{P}_\theta \Rightarrow X \sim \mathbb{P}_\theta$

(This is a strong Assumption) because

" All Model's Are wrong But some of them are useful "

$\Rightarrow$ mean's that may its not true that this Possion distribution we Assumed for # Sibling's for college student is not Perfectly correct.

$\Rightarrow$ when we make this Assumption, we are actually Assuming that the data really comes from Poisson distribution.

$X \sim \mathbb{P}_\theta$ θ^* or θ = true Parameter

The Aim of this statistical experiment is to Estimate θ, so that once we actually Plug in θ in $\mathbb{P}_\theta$, we would know that the Prob that my r.v takes any value.

Statistical model (3)

Examples

1. For n Bernoulli trials:
$$\left(\{0,1\}, (\mathsf{Ber}(p))_{p\in(0,1)}\right).$$
2. If $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Exp}(\lambda)$, for some unknown $\lambda > 0$:
$$\left(\mathbb{R}_+^*, (\mathsf{Exp}(\lambda))_{\lambda>0}\right).$$
3. If $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{Poiss}(\lambda)$, for some unknown $\lambda > 0$:
$$\left(\mathbb{N}, (\mathsf{Poiss}(\lambda))_{\lambda>0}\right).$$
4. If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$:
$$\left(\mathbb{R}, \left(\mathcal{N}(\mu, \sigma^2)\right)_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+^*}\right).$$

what we are trying to do as a statistician is to inject as much knowledge about the question and about the problem that we can so that the data has to do the minimal job and hencefourth, we actually need less data.

PARAmetric model (Parametric statistic)

$\theta \in \mathbb{R}^d$: wher d Number of Parameter's

NON Parametric Model: infinite number of Parameter's to estimate.

θ Cannot be represented by vector's

For Ex: function

Model $P_f = \{$ distribution of $\cdot f \quad f \geq 0, \int f = 1 \}$

Some statistical models:

① Bernoulli trails $\Rightarrow (\{0,1\}, \{Ber(p)\}_{p \in (0,1)})$

② Exponential $\Rightarrow ([0,+\infty), \{Exp(\lambda)\}_{\lambda > 0})$

③ Poisson $\Rightarrow (\mathbb{N}, \{Poiss(\lambda)\}_{\lambda > 0})$

④ Sibling's model $\Rightarrow (\{1,2,3 \dots 7\}, \{P(x=k) = p_k\}_{k=1 \text{ to } 7})$

⑤ Gaussian $\left(\mathbb{R}, \{N(\mu, \sigma^2)\}_{\substack{\mu \in \mathbb{R} \\ \sigma^2 \in \mathbb{R}^+}}\right)$

$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$

Identification

$\theta \longmapsto \mathbb{P}_\theta$ (injective)

The parameter θ is called *identified* iff the map $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, i.e.,

$$\theta = \theta' \Rightarrow \mathbb{P}_\theta = \mathbb{P}_{\theta'}.$$

$\theta \neq \theta' \Longrightarrow \mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$

Examples i.e. $\mathbb{P}_\theta = \mathbb{P}_{\theta'} \Longrightarrow \theta = \theta'$

1. In all four previous examples, the parameter was identified.

2. If $X_i = \mathbb{1}_{Y_i \geq 0}$, where $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, are unobserved: μ and σ^2 are not identified (but $\theta = \mu/\sigma$ is).

Parameter estimation (1)

Idea: Given an observed sample $X_1, \dots, X_n$ and a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, one wants to *estimate* the parameter θ.

Definitions

- *Statistic*: Any measurable[1] function of the sample, e.g., $\bar{X}_n, \max_i X_i$, $X_1 + \log(1 + |X_n|)$, sample variance, etc...
- *Estimator* of θ: Any statistic whose expression does not depend on θ.
- An estimator $\hat{\theta}_n$ of θ is *weakly* (resp. *strongly*) *consistent* iff

$$\hat{\theta}_n \xrightarrow[n \to \infty]{\mathbb{P} \text{ (resp. } a.s.)} \theta \quad \text{(w.r.t. } \mathbb{P}_\theta).$$

[1] Rule of thumb: if you can compute it exactly once given data, it is measurable. You may have some issues with things that are implicitly defined such as $\sup$ or $\inf$ but not in this class

we need some measure of performance of a given parameter, we need to be able to evaluate if eyeballing the problem is worse than actually collecting a large amount of data. To able to answer these question's we need to answer what accuracy mean's

Estimator : (Statistic)

An Estimator is a measurable function of data

$$\hat{\Theta} = \hat{\Theta}(X_1, X_2, \ldots X_n)$$

An Estimator is an Statistic which does not depend on Θ.

An Estimator is said to be consistent, if, when we collect more and more data, $\hat{\Theta}$ is getting closer and closer to true parameter.

A Estimator $\hat{\Theta}(X_1, X_2, \ldots X_n)$ is called

weakly convergent if it converge in probability.

Strongly convergent if it is Almost surely

Parameter estimation (2)

- *Bias* of an estimator $\hat{\theta}_n$ of θ:

$$\mathbb{E}\left[\hat{\theta}_n\right] - \theta.$$

- *Risk* (or *quadratic risk*) of an estimator $\hat{\theta}_n$:

$$\mathbb{E}\left[|\hat{\theta}_n - \theta|^2\right].$$

Remark: If $\Theta \subseteq \mathbb{R}$,

$$\text{"Quadratic risk} = \text{bias}^2 + \text{variance"}.$$

BIAS:- of a Estimator

$$\mathbb{E}[\hat{\theta}] - \theta$$

if $\text{Bia} = \mathbb{E}[\hat{\theta}] - \theta = 0 \Rightarrow \hat{\theta}$ is unBaised.

What is it mean to be unBaised?

* May Be for this Particular round of data we collected $(x_1, x_2, \ldots x_n)$, The Estimate $\hat{\theta}(x_1, x_2, \ldots x_n)$ we are pretty far from the the true Estimator, But if I we do this experiment over and over again, on average all the values of my Estimator I got, this would be my true Parameter.

* If I were to repeat this experiment, in average we would get this thing right.

Ex: Estimator $\overline{x_n}$

$$\mathbb{E}[\overline{x_n}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{n}{n}\mathbb{E}[x]$$

$$= \mathbb{E}[x]$$

$$\Rightarrow \mathbb{E}[\overline{x_n}] = \mathbb{E}[x] = p$$

Another Estimator : X_1

$$\hat{\theta}(x_1, x_2, \ldots x_n) = x_1$$

$$E[x_1] = p \quad \text{(unbailed)}$$

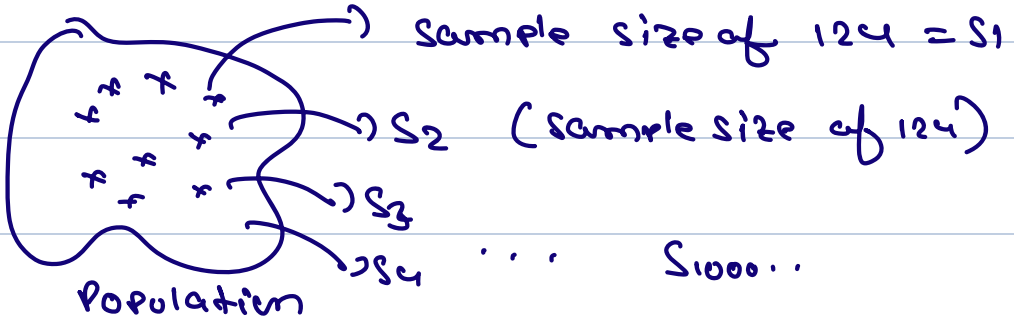
This should probably illustrate to us that Bias is not something that really is telling you the entire picture.

Ex: take only one observation (same Bias)

Bias is telling us where are we in average, But surely not telling us what fluctuations we are getting.

when we start having fluctuations coming into the picture, we actually have to look at risk and quadratic risk of a Estimator

* $x_1, x_2, \ldots, x_n \overset{iid}{\sim} Ber(p)$ assume $n = 124$



If I were to repeat this, 1000 times, every 1000 of those times I collect 124 data point's and then I do it again & again & again.

then in average the Number I should get should be closer to θ^* $\Rightarrow \mathbb{E}[\hat{\theta}] = \underbrace{\theta^*}_{\text{same}} = \theta$

as : $\mathbb{E}[\hat{\theta}_n] - \theta$

Quadratic Risk : or L_2 risk of $\hat{\theta}$

$$\mathbb{E}\left[|\hat{\theta} - \theta|^2\right]$$

if $\mathbb{E}\left[|\hat{\theta}-\theta|^2\right] \xrightarrow[n\to\infty]{} 0$, my Estimator $\hat{\theta}$ is

weakly consitants.

$$\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}]+\mathbb{E}[\hat{\theta}]-\theta)^2\right]$$

$$= \mathbb{E}\Big[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2 + (\mathbb{E}[\hat{\theta}]-\theta)^2 + 2(\hat{\theta}-\mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}]-\theta)\Big]$$

$$= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}[\hat{\theta}])^2\right] + (\mathbb{E}[\hat{\theta}]-\theta)^2$$

$$= \text{var}(\hat{\theta}) + \text{bais}^2$$

$$\Rightarrow \text{Quadratic risk} = \text{var}(\hat{\theta}) + \text{bais}^2$$

There is usally an inherant trade off b/w getting a small Bias and small variance. if we reduce one too much then the other one gonna increase.

Example:

$$x_1, x_2, \ldots, x_n \overset{iid}{\sim} \text{Ber}(\theta)$$

① $\hat{\theta} = \bar{X}_n \qquad \mathbb{E}[\bar{X}_n] = \mathbb{E}[X]$

$\Rightarrow$ Bias $= 0$ (unBaised)

$$\text{var}(\hat{\theta}) = \text{var}(\bar{X}_n) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2$$

$$= \frac{\sigma^2}{n} = \frac{\theta(1-\theta)}{n}$$

Quadratic risk $= \text{Var}(\hat{\theta}) + \text{Bias}^2$

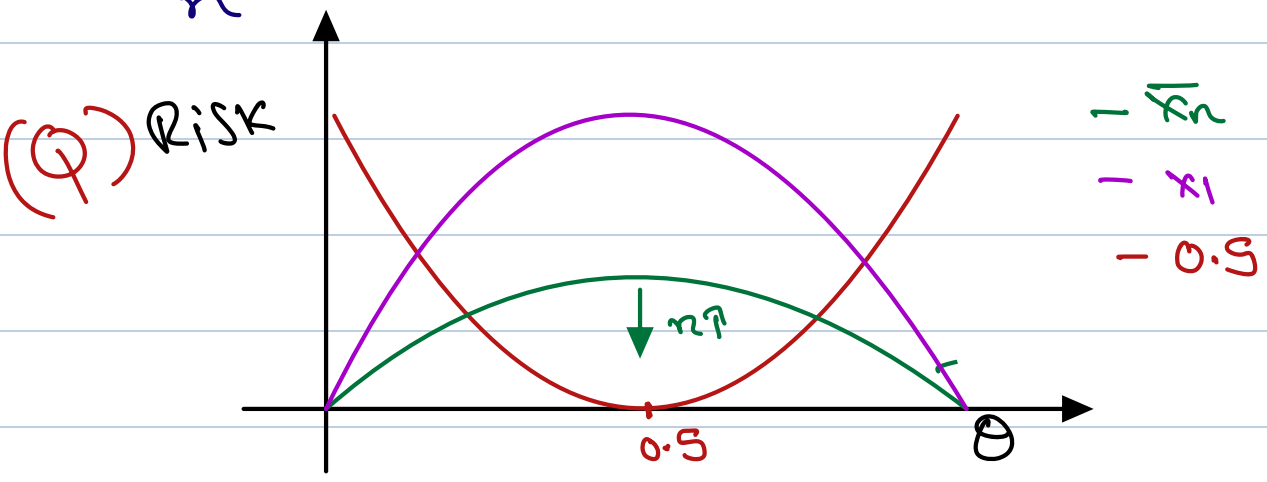
$$= \frac{\theta(1-\theta)}{n} + 0$$

$$= \frac{\theta(1-\theta)}{n}$$

This is just summarizing the performance of an Estimator in an r.v

Three Estimator's for θ

① $\bar{x}_n$	② 0.5	③ x_1
Bias $= 0$	Bias $= (0.5 - \theta)$	Bias $= 0$
Var $= \frac{\theta(1-\theta)}{n}$	Var $= 0$	Var $= \theta(1-\theta)$
risk $= \frac{\theta(1-\theta)}{n}$	$(0.5-\theta)^2$	$\theta(1-\theta)$



In NON-Parametric estimation, all we do is Bias / variance trade off's all the time.

Confidence intervals (1)

Let $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ be a statistical model based on observations $X_1, \dots, X_n$, and assume $\Theta \subseteq \mathbb{R}$.

Definition

Let $\alpha \in (0, 1)$.

- *Confidence interval (C.I.) of level* $1-\alpha$ for θ: Any random (i.e., depending on $X_1, \dots, X_n$) interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:

$$\mathbb{P}_\theta\left[\mathcal{I} \ni \theta\right] \geq 1-\alpha, \qquad \forall \theta \in \Theta.$$

- *C.I. of asymptotic level* $1-\alpha$ for θ: Any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that:

$$\lim_{n\to\infty} \mathbb{P}_\theta\left[\mathcal{I} \ni \theta\right] \geq 1-\alpha, \qquad \forall \theta \in \Theta.$$

* A Confidence interval is a random interval, The Bound's of the interval depends on random data.
* $\bar{X}$ is what the random thing that make fluctuate the confidence interval
* Now we have interval and have its Boundaries, Boundaries are not allowed to depend on unknown Parameter (Otherwise its not a Confidence interval) Just Like an Estimator, that depends on unknown Parameter, is not a Estimator.
* Confidence interval has to be Something that we can Compute once I Collect data.

$$P_\theta [I \ni \theta] \geq 1-\alpha \quad \forall \theta \in \Theta$$

Probability that I contain's θ

$\theta \in I$ (wrong), $I \ni \theta$ (correct)

$\Rightarrow$ The randomness is in I, θ is the true unknown value (deterministic)

The confidence interval is still a random variable

* Now, if we start plugging in numbers instead of r.v $(X_1, X_2, \ldots X_n)$, Ex: Pot 1,0,0,1,1,0 ..., in this case the random interval is actually for Ex: $[0.42, 0.65]$,
* The Probability that $\theta \in [0.42, 0.65]$ is not $1-\alpha$, i.e $\mathbb{P}_\theta[[0.42, 0.65] \ni \theta] \neq 1-\alpha$, It is either 0 (if θ not in there), 1 if its there.

Confidence intervals (2)

Example: Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Ber}(p)$, for some unknown $p \in (0,1)$.

- LLN: The sample average $\bar{X}_n$ is a strongly consistent estimator of p.
- Let $q_{\alpha/2}$ be the $(1-\frac{\alpha}{2})$-quantile of $\mathcal{N}(0,1)$ and

$$\mathcal{I} = \left[\bar{X}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \bar{X}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \right].$$

- CLT: $\lim_{n\to\infty} \mathbb{P}_p [\mathcal{I} \ni p] = 1-\alpha, \quad \forall p \in (0,1).$
- Problem: $\mathcal{I}$ depends on p !

Confidence intervals (3)

Two solutions:

- Replace $p(1-p)$ with $1/4$ in $\mathcal{I}$ (since $p(1-p) \leq 1/4$).
- Replace p with $\bar{X}_n$ in $\mathcal{I}$ and use Slutsky's theorem.

CLT:

$$\frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}} \xrightarrow[n\to\infty]{d} N(0,1)$$

$$\Rightarrow \quad \bar{X} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Using CLT, the distribution of $\frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}}$ approximately equal to $N(0,1)$

$$\Rightarrow \lim_{n\to\infty} P_p\left(\left| \frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}} \right| \le q_{\alpha/2} \right) = 1-\alpha$$

$$\Rightarrow \lim_{n\to\infty} P_p\left(-q_{\alpha/2} \le \frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}} \le q_{\alpha/2} \right) = 1-\alpha$$

$$\Rightarrow \lim_{n\to\infty} P_p\left(p - \sqrt{\frac{p(1-p)}{n}} \cdot q_{\alpha/2} \le \bar{X} \le p + \sqrt{\frac{p(1-p)}{n}} \cdot q_{\alpha/2} \right) = 1-\alpha$$

18.650 / 18.6501 Statistics for Applications

Fall 2016