Statistics for Applications

Chapter 2: Parametric Inference

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 $(E, \mathcal{F})$ (usually $E \subseteq \mathbb{R}$) and denote by $\mathbb{P}$ their common distribution. A *statistical model* associated to that statistical experiment is a triplet

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

where:

- $(E, \mathcal{F})$ is the measurable space of the observations;
- $(\mathbb{P}_\theta)_{\theta \in \Theta}$ is a family of probability measures on $(E, \mathcal{F})$;
- Θ is any set, called *parameter set*.

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

Statistical model (3)

Examples

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

Identification

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

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

Examples

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

2. If $X_i = \mathbb{1}_{U_i \geq 0}$, where $U_1, \ldots, U_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, are unobserved: μ and σ^2 are not identified (but μ/σ is).

Parameter estimation (1)

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

Definitions

- *Statistic*: Any measurable 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

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

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 $\Theta \subseteq \mathbb{R}$,

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

Confidence intervals (1)

Let $(E, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ be a statistical model based on observations $X_1, \ldots, X_n$, 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 $X_1, \ldots, X_n$) 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:

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

Confidence intervals (2)

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

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

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

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