

- 1 What is statistical inference ?
- 2 Point estimation
 - MSE of estimators
 - Method of Moments
- 3 Fitting probability distributions
 - Nonparametric estimation of the CDF
 - Fitting a parametric model

Statistical inference

The statistical approach

- 1 Probability : **given a generating process**, what are the properties of the outcomes ?
Ex : if you toss a coin 200 times, what is the probability of observing 110 heads ?
- 2 Statistical inference : **given the outcomes (sample, data)**, what can we say about the process that generated the data ?
Ex : if you toss a coin 200 times and you observe 110 heads, is the coin toss fair ?

↪ use data to infer the generative distribution : parameter estimation and hypothesis tests

↪ *inference* : an assumption you make about the law of probability based on the information you have = **the sample**

Statistical inference

The statistical model

X = number of heads (in 200 coin tosses)

- Probability : $X \sim \text{Binomial}\left(200, \frac{1}{2}\right)$
- Statistics : $X \sim \text{Binomial}(200, p)$ with $0 \leq p \leq 1$: how to estimate p ? is hypothesis $p = \frac{1}{2}$ true?

↔ What information about X ? the observation of a *sample* :

HTHHTTTHTHTT.....HH, H = head and T = tail

↔ Or equivalently, $X_i = 1$ if we get a H or $X_i = 0$ if it is a T :

$$X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, \dots, X_{200} = 1$$

Statistical model :

- X_1, X_2, \dots, X_{200} is a sample of independent and identically distributed (i.i.d.) variables
- their common distribution is $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$ (Bernoulli) and p is an unknown parameter.

Statistical inference

Simple random sampling

Let X be a character studied in a population of size N . We observe a sample X_1, \dots, X_n of n individuals in the population.

2 sampling schemes :

- simple random sampling *without replacement* : more natural but the X_i s are not independent
- simple random sampling *with replacement* : more mathematically convenient

\Leftrightarrow if the population size is infinite (or N large), the 2 samplings are the same in practice (weak dependence of the X_i s)

\Leftrightarrow if the population is finite (N small), sample mean and variance are computed with a *finite population correction*.

Reference : To go further, chapter 7 *Survey sampling* Rice's book

Parametric models

Given a sample X_1, \dots, X_n i.i.d. with CDF F , how do we infer F ?

Definitions :

- A **statistical model** is a set of probability distributions (CDF, PDF or PMF).
- A **parametric model** can be parameterized by a finite number of real parameters $\theta \in \Theta \subset \mathbb{R}^P$.

Examples :

- $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, one-dimensional parametric model : $\theta = p$
- $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$, 2-parameter model

$$\left\{ pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \mu \in \mathbb{R}, \sigma > 0 \right\}$$

$$\theta = (\mu, \sigma)$$

- If we assume $F \in \{ \text{all CDF's} \}$, the model is **nonparametric** (not finite dimensional)

Definition : A point estimator $\hat{\theta}$ of a parameter θ is a function of the sample : $\hat{\theta} = g(X_1, \dots, X_n)$ where g does not depend on θ .

By convention, we often denote an estimator by $\hat{\theta}$ or T .

θ is a fixed unknown quantity. $\hat{\theta}$ is a *random variable*.

Example : $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$

- $T_1 = X_1$
- $T_2 = \frac{X_1 + X_2}{2}$
- $T_3 = \bar{X} = \frac{X_1 + \dots + X_n}{n}$
- $T_4 = \frac{\min(X_i) + \max(X_i)}{2}$

Variety of possible estimates : how to choose which one to use ?

How do we assess the quality of $\hat{\theta}$?

- 1 Bias of an estimator : $\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$
 $\hat{\theta}$ is **unbiased** if $\mathbb{E}(\hat{\theta}) = \theta, \forall \theta \in \Theta$: $\hat{\theta}$'s law is centered on the true parameter value.

- 2 Mean squared error $\text{MSE} = \mathbb{E} \left[(\hat{\theta} - \theta)^2 \right]$

It measures the concentration of $\hat{\theta}$'s law about the true parameter value.

- 3 An estimator is usually a function of the sample size : $\hat{\theta} = \hat{\theta}_n$. It should converge to θ as we collect more and more data.

Definition : $\hat{\theta}_n$ is **consistent** if $\text{MSE}(\hat{\theta}_n) \rightarrow 0$ as n tends to ∞ .

Estimators

Sampling distribution

- The distribution of $\hat{\theta}$ is called the **sampling distribution**.
- The square root of the variance of $\hat{\theta}$ is called the **standard error** of the estimator

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

$\text{se}(\hat{\theta})$ depends on θ , it is an unknown quantity that we usually can estimate.

Example : $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, $\hat{p} = \bar{X} = \frac{X_1 + \dots + X_n}{n}$

- $E(\bar{X}) = E(X_i) = p$
- $\text{MSE} = E(\bar{X} - p)^2 = \text{Var}(\bar{X}) = \frac{\text{Var}(X_i)}{n} = \frac{p(1-p)}{n} \rightarrow 0$
- The estimated standard error of p is $\frac{\bar{X}(1-\bar{X})}{n}$

Proposition

The MSE can be written as $MSE = \text{bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$

Proof :

$$\begin{aligned}(\hat{\theta} - \theta)^2 &= (\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2 \\&= \left(\hat{\theta} - \mathbb{E}(\hat{\theta}) \right)^2 + 2 \left(\hat{\theta} - \mathbb{E}(\hat{\theta}) \right) \left(\mathbb{E}(\hat{\theta}) - \theta \right) + \left(\mathbb{E}(\hat{\theta}) - \theta \right)^2 \\&= \left(\hat{\theta} - \mathbb{E}(\hat{\theta}) \right)^2 + 2 \left(\hat{\theta} - \mathbb{E}(\hat{\theta}) \right) \text{bias}(\hat{\theta}) + \text{bias}^2(\hat{\theta})\end{aligned}$$

Then take the expectation on each side and note that $\mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta})) = 0$ and $\text{Var}(\hat{\theta}) = \mathbb{E} \left((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 \right)$.

Estimators

Sample variance

We can compare estimators by comparing their MSE.

Example : $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$, 2 estimators of σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{and} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \sigma^2/n$ (*show it*)
- $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$ (*show it*)
But $E[(X_i - \mu)^2] = \text{Var}(X_i) = \sigma^2$ and $E[(\bar{X} - \mu)^2] = \text{Var}(\bar{X}) = \sigma^2/n$. Thus,
 $E(S^2) = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2$.
 S^2 is an unbiased estimate of σ^2
- $\text{Var}(S^2) = 2\sigma^4/(n-1)$ (*admitted*) so

$$\text{MSE}(S^2) = 0^2 + 2\sigma^4/(n-1) = \frac{2}{n-1}\sigma^4$$

Estimators

Sample variance

- $\hat{\sigma}^2 = \frac{n-1}{n} S^2$; thus, $E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2$
- $\text{Var}(\hat{\sigma}^2) = \frac{(n-1)^2}{n^2} \text{Var}(S^2) = 2 \frac{n-1}{n^2} \sigma^4$; thus :

$$\text{MSE}(\hat{\sigma}^2) = 2 \frac{n-1}{n^2} \sigma^4 + \frac{1}{n^2} \sigma^4 = \frac{2n-1}{n^2} \sigma^4$$

In conclusion, as $\frac{2n-1}{n^2} < \frac{2}{n-1}$,

$$\text{MSE}(\hat{\sigma}^2) < \text{MSE}(S^2)$$

$\hat{\sigma}^2$ is better than S^2 in the sense of MSE criterion.

Estimators

Method of Moments

The method of moments provide estimators that are easy to compute but not optimal.

The *k*th moment of a probability law is $\mu_k = E(X^k)$ and the *k*th sample moment is $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$.

If θ_1 and θ_2 can be expressed as $\theta_1 = h_1(\mu_1, \mu_2)$ and $\theta_2 = h_2(\mu_1, \mu_2)$, then the method of moments estimates are

$$\hat{\theta}_1 = h_1(\hat{\mu}_1, \hat{\mu}_2) \quad \text{and} \quad \hat{\theta}_2 = h_2(\hat{\mu}_1, \hat{\mu}_2)$$

Examples :

- The sample mean $\bar{X} = \hat{\mu}_1$ is the moment estimate of $E(X)$.
- The sample variance $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$ is the moment estimate of $\sigma^2 = \mu_2 - \mu_1^2$. (*show it*)

Let X_1, \dots, X_n be an i.i.d. sample from a uniform law on $[0, \theta]$, $\theta > 0$. The density function is

$$f(x) = \frac{1}{\theta} \text{ for } 0 \leq x \leq \theta, \quad 0 \text{ otherwise.}$$

- 1 Compute the expectation and the variance of the X_i s
- 2 Give the estimator $\hat{\theta}$ of θ obtained from the method of moments.
- 3 Find the bias, standard error and MSE of $\hat{\theta}$.

Nonparametric estimation of the CDF

ecdf

Let $X_1, \dots, X_n \sim F$ be an i.i.d. sample.

Definition. The **empirical cumulative distribution function** (ecdf) is the CDF that put mass $1/n$ at each data point X_i :

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \leq x}$$

$\hat{F}(x)$ is a consistent estimate of $F(x)$.

- $E(\hat{F}(x)) = F(x)$
- $\text{Var}(\hat{F}(x)) = \frac{F(x)(1-F(x))}{n}$
- $\text{MSE} = \frac{F(x)(1-F(x))}{n} \rightarrow 0$

Nonparametric estimation of the CDF

Sample quantiles

Let the CDF F be strictly increasing.

- For $0 < p < 1$, the p th quantile is $F^{-1}(p)$.
- The empirical p th quantile or sample quantile is the quantile of the ecdf : $\hat{F}^{-1}(p)$.

As \hat{F}^{-1} is not invertible, we define $F^{-1}(p) = \inf\{x, \hat{F}(x) \geq p\}$.

Remark : the sample median is the sample quantile of order $1/2$.

Nonparametric estimation of the CDF

QQ-plot

Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the ordered observations of the sample. \widehat{F} is the step function from 0 to 1 with value $\widehat{F}(x) = \frac{k}{n}$ if $X_{(k)} \leq x < X_{(k+1)}$; if $X_{(k)} = X_{(k+1)}$, the jump at $X_{(k)}$ is $2/n$.

- If np is not an integer, there is one value $j = 1, \dots, n$ such that $(j-1)/n < p < j/n$: $\widehat{F}^{-1}(p) = X_{(j)}$.
- If np is an integer j , then we can define the p th sample quantile as $\frac{X_{(j)} + X_{(j+1)}}{2}$.

QQ-plot : assessing the fit of the data to a model F

\hookrightarrow plot $X_{(j)}$ vs $F^{-1}\left(\frac{j}{n+1}\right), j = 1, \dots, n$.

Fitting a parametric model

An example

The gamma density function depends on two parameters, α and λ

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x \leq \infty$$

The family of gamma distributions provides a flexible set of densities for nonnegative random variables.

The first two moments of the gamma distribution are

$$\mu_1 = \frac{\alpha}{\lambda}, \quad \mu_2 = \frac{\alpha(\alpha + 1)}{\lambda^2}$$

The method of moment estimates are

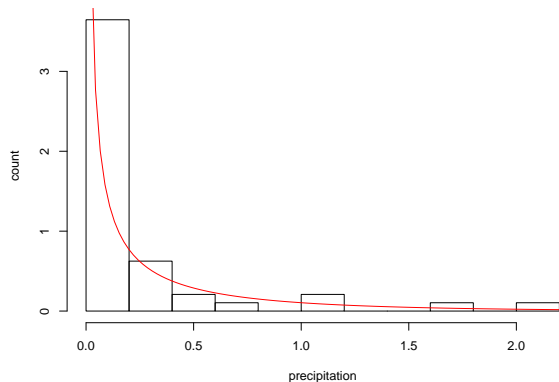
$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2}, \quad \hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$.

Fitting a parametric model

An example

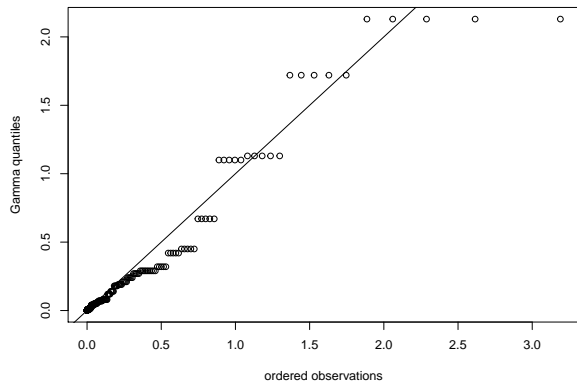
We consider the fit of the amounts of precipitation during 227 storms in Illinois from 1960 to 1964 to a gamma distribution (Le Cam and Neyman 1967).



Fitting a parametric model

An example

Gamma quantile-quantile plot of Illinois' data : sample quantiles vs estimated gamma quantiles.



- *All of statistics* Wasserman : sections 6.1, 6.2, 6.3.1, 7.1, 9.1, 9.2