Statistical inference, parametric models

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
4. Maximum likelihood estimation

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 ?

$\hookrightarrow$ use data to infer the generative distribution : parameter estimation and hypothesis tests
$\hookrightarrow$ *inference* : an assumption you make about the law of probability based on the information you have = the sample

X = number of heads (in 200 coin tosses)

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

$\hookrightarrow$ What information about X ? the observation of a *sample* :

$$HTHHTTTHTHTT......HH, \quad H = \text{head and } T = \text{tail}$$

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

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

Statistical model :

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

Parametric models

Given a sample $X_1, \ldots, X_n$ 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 :

- $X_1, \ldots, X_n \sim Bernoulli(p)$, one-dimensional parametric model : $\theta = p$
- $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$, 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 $F \in \{$ all CDF's $\}$, the model is **nonparametric** (not finite dimensional)

Estimators

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

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

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

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

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

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

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

1. Bias of an estimator : $\text{bias}(\widehat{\theta}) = \mathrm{E}(\widehat{\theta}) - \theta$
   $\widehat{\theta}$ is **unbiased** if $\mathrm{E}(\widehat{\theta}) = \theta$, $\forall\theta \in \Theta$: $\widehat{\theta}$'s law is centered on the true parameter value.
2. Mean squared error $\text{MSE} = \mathrm{E}\left[(\widehat{\theta} - \theta)^2\right]$
   It measures the concentration of $\widehat{\theta}$'s law about the true parameter value.
3. An estimator is usually a function of the sample size : $\widehat{\theta} = \widehat{\theta}_n$. It should converge to θ as we collect more and more data.
   Definition : $\widehat{\theta}_n$ is **consistent** if $\text{MSE}(\widehat{\theta}_n) \longrightarrow 0$ as n tends to ∞.

- The distribution of $\widehat{\theta}$ is called the sampling distribution.
- The square root of the variance of $\widehat{\theta}$ is called the standard error of the estimator
$$\mathrm{se}(\widehat{\theta}) = \sqrt{\mathrm{Var}(\widehat{\theta})}.$$
$\mathrm{se}(\widehat{\theta})$ depends on θ, it is an unknown quantity that we usually can estimate.

Example : $X_1, \ldots, X_n \sim \mathit{Bernoulli}(p)$, $\widehat{p} = \overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$

- $\mathrm{E}(\overline{X}) = \mathrm{E}(X_i) = p$
- $\mathrm{MSE} = \mathrm{E}(\overline{X} - p)^2 = \mathrm{Var}(\overline{X}) = \dfrac{\mathrm{Var}(X_i)}{n} = \dfrac{p(1-p)}{n} \longrightarrow 0$
- The estimated standard error of p is $\sqrt{\dfrac{\overline{X}(1-\overline{X})}{n}}$

Proposition

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

Proof :

$$(\widehat{\theta} - \theta)^2 \;=\; (\widehat{\theta} - \mathrm{E}(\widehat{\theta}) + \mathrm{E}(\widehat{\theta}) - \theta)^2$$

Then develop and take the expectation on each side.

We can compare estimators by comparing their MSE.

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

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

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

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

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

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

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

$$\mathsf{MSE}(\widehat{\sigma}^2) < \mathsf{MSE}(S^2)$$

.

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

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

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

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

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

Examples :

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

Exercise

Let $X_1, \ldots, X_n$ 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_is
2. Give the estimator $\widehat{\theta}$ of θ obtained from the method of moments.
3. Find the bias, standard error and MSE of $\widehat{\theta}$.

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

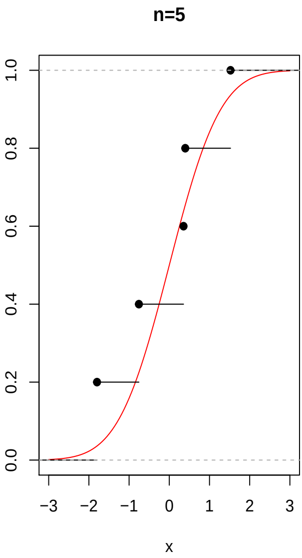
Definition. The empirical cumulative distribution function (ecdf) is the CDF that puts mass $1/n$ at each data point X_i :

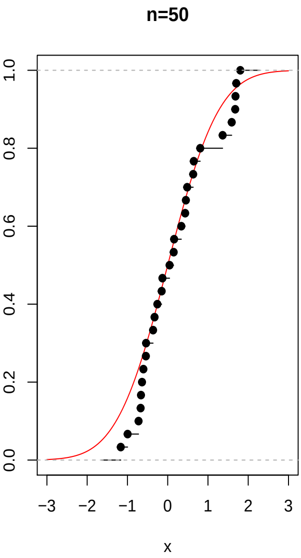
$$\widehat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x}$$

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

- $\mathrm{E}\left(\widehat{F}(x)\right) = F(x)$
- $\mathrm{Var}\left(\widehat{F}(x)\right) = \dfrac{F(x)(1-F(x))}{n}$
- MSE $= \frac{F(x)(1-F(x))}{n} \longrightarrow 0$

ecdf





Let the CDF F be strictly increasing.

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

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

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

Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ 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 $j = 1, \ldots, n$ 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 pth 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 $F^{-1}\left(\frac{j}{n+1}\right)$,$j = 1, \ldots, n$.

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

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

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

The first two moments of the gamma distribution are

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

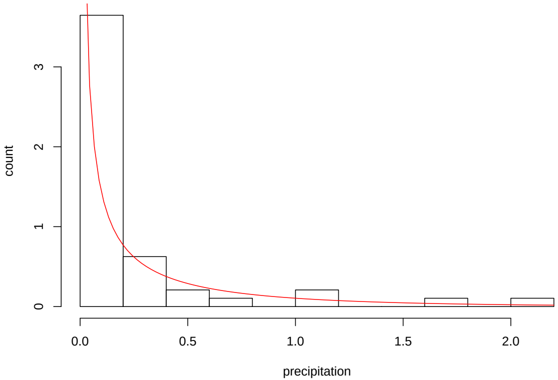
The method of moments estimates are

$$\widehat{\lambda} = \frac{\overline{X}}{\widehat{\sigma}^2}, \quad \widehat{\alpha} = \frac{\overline{X}^2}{\widehat{\sigma}^2}$$

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

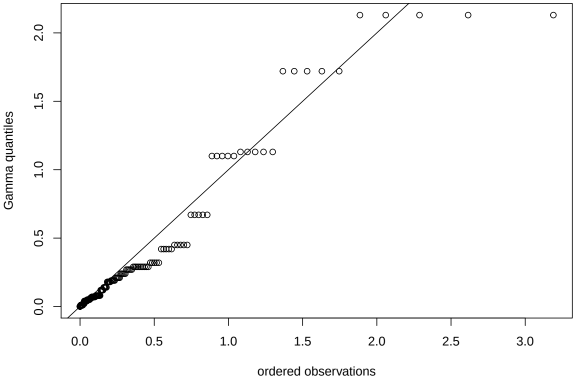
Fitting a parametric model
An example

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.



Maximum Likelihhod Estimation

a method for parametric models

Let θ be a parameter of the sample distribution (possibly a vector of parameters).

Notation :

- f_θ is the PDF of X_i if X_i is a continuous random variable.
- f_θ is the PMF of X_i if X_i is a discrete random variable.

Definition

The function of θ $L_{X_1,\ldots,X_n}(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$ is called the likelihood function of the sample $(X_1, \ldots, X_n)$.
The maximum likelihood estimator (MLE) is the value of θ that maximizes $L_{X_1,\ldots,X_n}(\theta)$.

Properties of the MLE

Mathematically easier to maximize $\log L(\theta)$
$\hookrightarrow$ the MLE $\widehat{\theta}_{ML}$ maximizes the log-likelihood

- $\widehat{\theta}_{ML}$ is a random variable
- *Consistency* : $\widehat{\theta}_{ML}$ converges to the true value θ (the MSE tends to 0)
- if $\widehat{\theta}_{ML}$ is the MLE of θ, $g(\widehat{\theta}_{ML})$ is the MLE of $g(\theta)$ (*equivariance*)
- *optimality* : $\widehat{\theta}_{ML}$ has the smallest variance, at least for large samples

Computation of the MLE

If $\log L$ is a differentiable function

- Take the derivative(s) of $\log L(\theta)$ and set it equal to 0 (*likelihood equations*)
- Verify that the solution is indeed a global maximum of the log-likelihood

- $(Y_1, \dots, Y_n) \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

$$\log\{L(\theta)\} = -\frac{n}{2}\{\log(\sigma^2) + \log(2\pi)\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2$$

$$\widehat{\mu}_{MV} = \overline{Y}, \qquad \widehat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Computation of the MLE

- $(Y_1, \ldots, Y_n) \sim \mathrm{Gamma}(\alpha, \lambda)$, $\theta = (\alpha, \lambda)$

$$\log\{L(\theta)\} = n\alpha \log(\lambda) + (\alpha - 1) \sum_i \log X_i - \lambda \sum_i X_i - n \log \Gamma(\alpha)$$

$\widehat{\lambda} = \dfrac{\widehat{\alpha}}{\overline{X}}$ and

$$n \log \widehat{\alpha} - n \log \overline{X} + \sum_i \log X_i - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} = 0$$

If the likelihood equation can not be solved in closed form, use an iterative optimization algorithm (implemented in python or R for example).

iterative optimization algorithms

When using these algorithms, you need

- to start the iterative procedure
- to check that the algorithm has converged
- to check that the solution is correct (is it a global maximum ? change the initial value and restart the algorithm)

Reference

- *Mathematical statistics and data analysis*, Rice : sections 8.1 to 8.5.1