Review of statistics

Michel Bierlaire

michel.bierlaire@epfl.ch

Transport and Mobility Laboratory

Probability distributions

A probability density function on a set S of outcomes must

- be non negative for all outcomes in S,
- sum up or integrate to 1.

Example:

$$f(x) = \frac{x}{4} + \frac{7x^3}{2}, \text{ with } 0 \leq x \leq 1,$$

is a PDF.

Is it useful in practice?

Probability distributions

A PDF should model probabilistic behavior of real-world phenomena.

- Normal distribution
- Poisson distribution
- Gamma distributions
- Extreme Value distributions
- ...

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Normal distribution

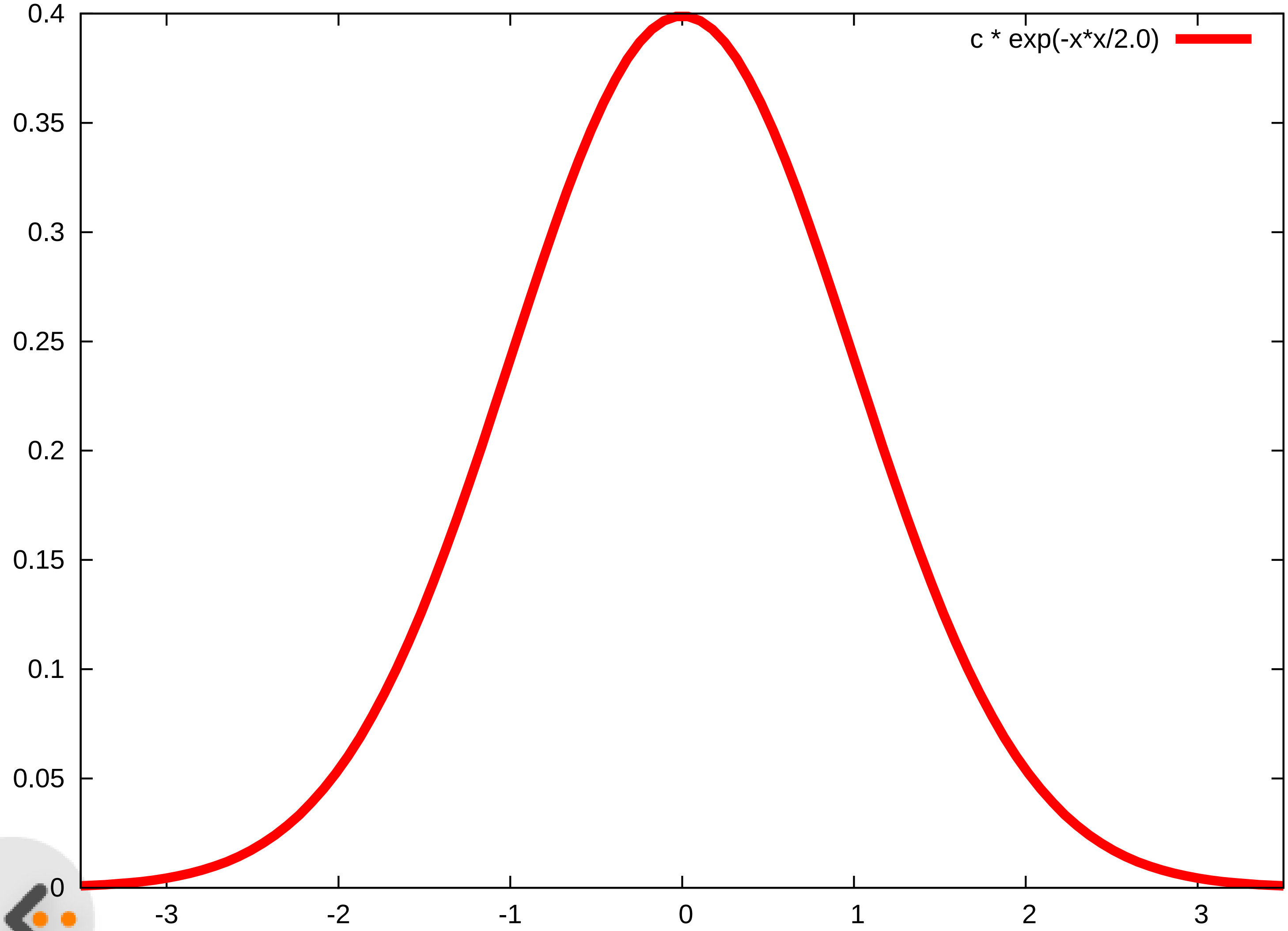
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Motivation: Central Limit Theorem

- $X_1, X_2, \ldots$ infinite sequence of i.i.d random variables, with finite mean μ and finite variance σ^2.
- For any number a and b

$$\lim_{n\to\infty} P\left(a \leq \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

Normal distribution



Normal distribution

Cumulative Distribution Function (CDF)

$$P(X \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

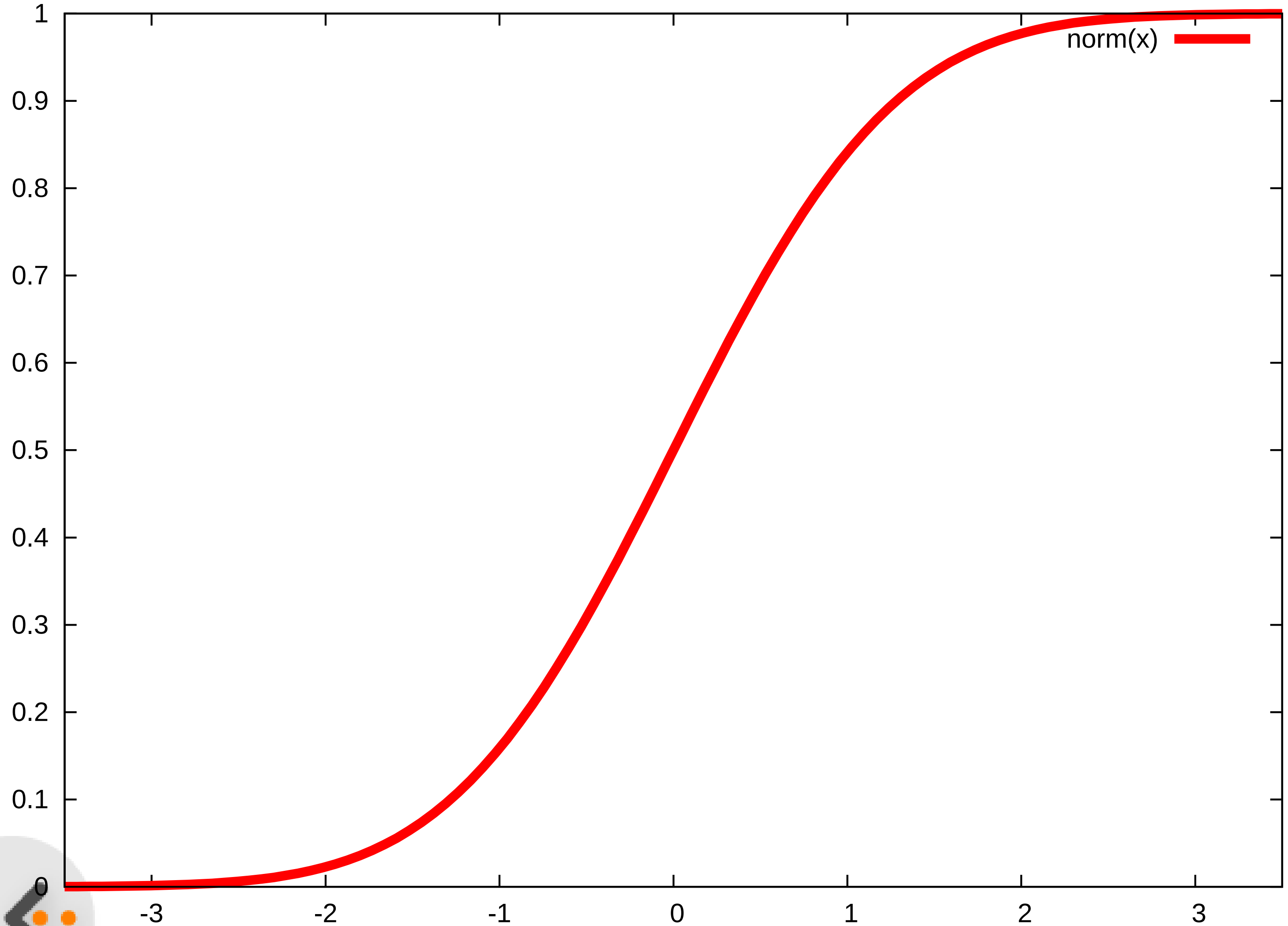
No closed form formula

Notation:

$$X \sim N(0, 1)$$

- $f_X(x)$ is the PDF
- $F_X(x)$ is the CDF

Normal distribution



Normal distribution

$$X \sim N(\mu, \sigma^2)$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}.$$

$$Y \sim N(0, 1)$$

$$Y = \frac{X - \mu}{\sigma}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Normal distribution

- Linear combinations of normal r.v.:
  - X_i, $i = 1, \ldots, n$
  - $X_i \sim N(\mu_i, \sigma_i^2)$
  - X_i independent
  - Then, if $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, n$

$$\sum_{i=1}^{n} \alpha_i X_i \sim N\left(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2\right)$$

Normal distribution

- Linear transformation of a normal r.v.
  - $X \sim N(\mu, \sigma^2)$
  - $\alpha, \beta \in \mathbb{R}$
  - Then,

$$\alpha + \beta X \sim N\left(\alpha + \beta\mu, \beta^2\sigma^2\right)$$

- Parameter estimation

Parameter	Estimator	Method/properties
μ	$\bar{x}$	Unbiased, maximum likelihood
σ^2	$\frac{n}{n-1}s^2$	Unbiased
σ^2	s^2	Maximum likelihood

Extreme value distribution

- $X_1, \ldots, X_n$ i.i.d.
- $f_{X_i}(x) = f(x)$, $F_{X_i}(x) = F(x)$, $i = 1, \ldots, n$
- $X'_n = \max(X_1, \ldots, X_n)$
- Applications:
  - rainfall
  - floods
  - earthquakes
  - air pollution
  - ...

Extreme value distribution

Emil
Julius
Gumbel

1891–1966

- father of extreme value theory
- politically involved left-wing pacifist in Germany,
- strongly against right wing's campaign of organized assassination (1919)
- first German professor to be expelled from university under the pressure of the Nazis
- in 1932 he left Heidelberg to Paris, where he met Borel and Fréchet.
- in 1940, he had to escape to New-York, where he continued his fight against Nazism by helping the US secret service.

Extreme value distribution

- $X'_n = \max(X_1, \ldots, X_n)$
- $F_{X'_n} = F(x)^n$. Indeed

$$P(X'_n \leq x) = P(X_1 \leq x)P(X_2 \leq x)\ldots P(X_n \leq x)$$

- Warning: if $n \to \infty$

$$\lim_{n\to\infty} F_{X'_n}(x) = \begin{cases} 1 & \text{if } F(x) = 1 \\ 0 & \text{if } F(x) < 1 \end{cases}$$

Degenerate distribution

Extreme value distribution

- We want a limiting distribution which is non degenerate
- Limiting distribution of some sequence of transformed "reduced" values
- For instance $a_n X'_n + b_n$
- a_n, b_n do not depend on x
- CDF of limiting distribution: $G(x)$
- Let's identify desired properties

Extreme value distribution

$$
\begin{array}{ccc|l}
X_1 & \ldots & X_n & \max(X_1, \ldots, X_n) \\
X_{n+1} & \ldots & X_{2n} & \max(X_{n+1}, \ldots, X_{2n}) \\
\vdots & & \vdots & \\
X_{(i-1)n+1} & \ldots & X_{in} & \max(X_{(i-1)n+1}, \ldots, X_{in}) \\
\vdots & & \vdots & \\
X_{(N-1)n+1} & \ldots & X_{Nn} & \max(X_{(N-1)n+1}, \ldots, X_{Nn})
\end{array}
$$

Two ways of seeing $\max(X_1, \ldots, X_{Nn})$ when $n \to \infty$

1. As a max of many X_i, the CDF should look like $G(a_N x + b_N)$
2. The CDF of the max of each row is $G(x)$
3. So the CDF of the max of all rows is $G(x)^N$.

Extreme value distribution

Stability postulate (Fréchet, 1927):

$$G(x)^N = G(a_N x + b_N)$$

We consider here the case $a_N = 1$ to obtain the so-called “type I extreme value distribution”

$$G(x)^N = G(x + b_N)$$

We have also

$$\begin{aligned} G(x)^{MN} &= G(x+b_N)^M &= G(x+b_N+b_M) \\ G(x)^{MN} &= G(x+b_{MN}) & \end{aligned}$$

Extreme value distribution

Therefore

$$G(x + b_N + b_M) = G(x + b_{MN})$$

that is

$$b_N + b_M = b_{MN}$$

so that b_N must be of the form

$$b_N = -\sigma' \ln N,$$

and the stability postulate becomes

$$G(x)^N = G(x - \sigma' \ln N)$$

Let's take the logarithm twice

Extreme value distribution

$$G(x)^N = G(x - \sigma' \ln N)$$

$$N \ln G(x) = \ln G(x - \sigma' \ln N)$$

Warning: G is a CDF, so $G(x) \leq 1$ and $\ln G(x) \leq 0$, $\forall x$

$$-N \ln G(x) = -\ln G(x - \sigma' \ln N)$$

$$\ln N + \ln(-\ln G(x)) = \ln(-\ln G(x - \sigma' \ln N))$$

Define $h(x) = \ln(-\ln G(x))$ to obtain

$$\ln N + h(x) = h(x - \sigma' \ln N)$$

h is affine.

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Extreme value distribution

$$\begin{aligned}
\ln N + h(x) &= h(x - \sigma' \ln N) \\
h(x) &= \alpha x + \beta \\
h(0) &= \beta \\
\ln N + \alpha x + \beta &= \alpha(x - \sigma' \ln N) + \beta \\
\alpha &= -\tfrac{1}{\sigma'}
\end{aligned}$$

Therefore

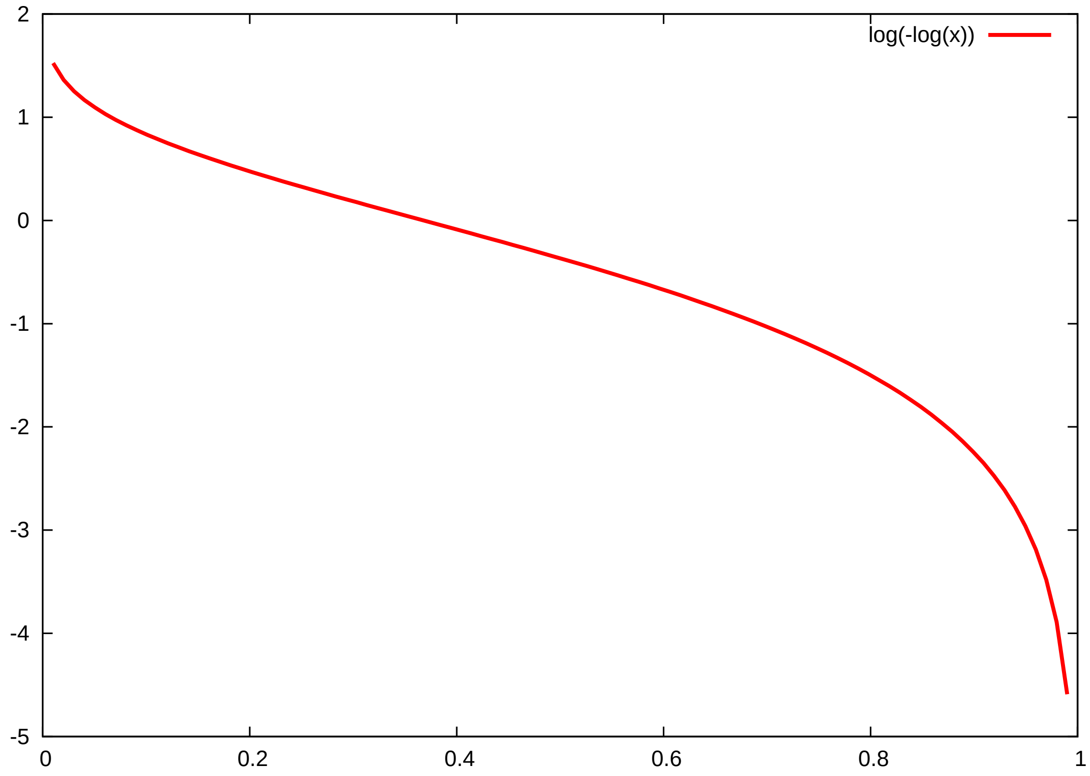
$$h(x) = h(0) - \frac{x}{\sigma'}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Extreme value distribution

G is increasing in x (CDF), so h is decreasing in x



Therefore, $\sigma' > 0$

Extreme value distribution

$$h(x) = \ln(-\ln G(x)) = h(0) - \frac{x}{\sigma'}$$

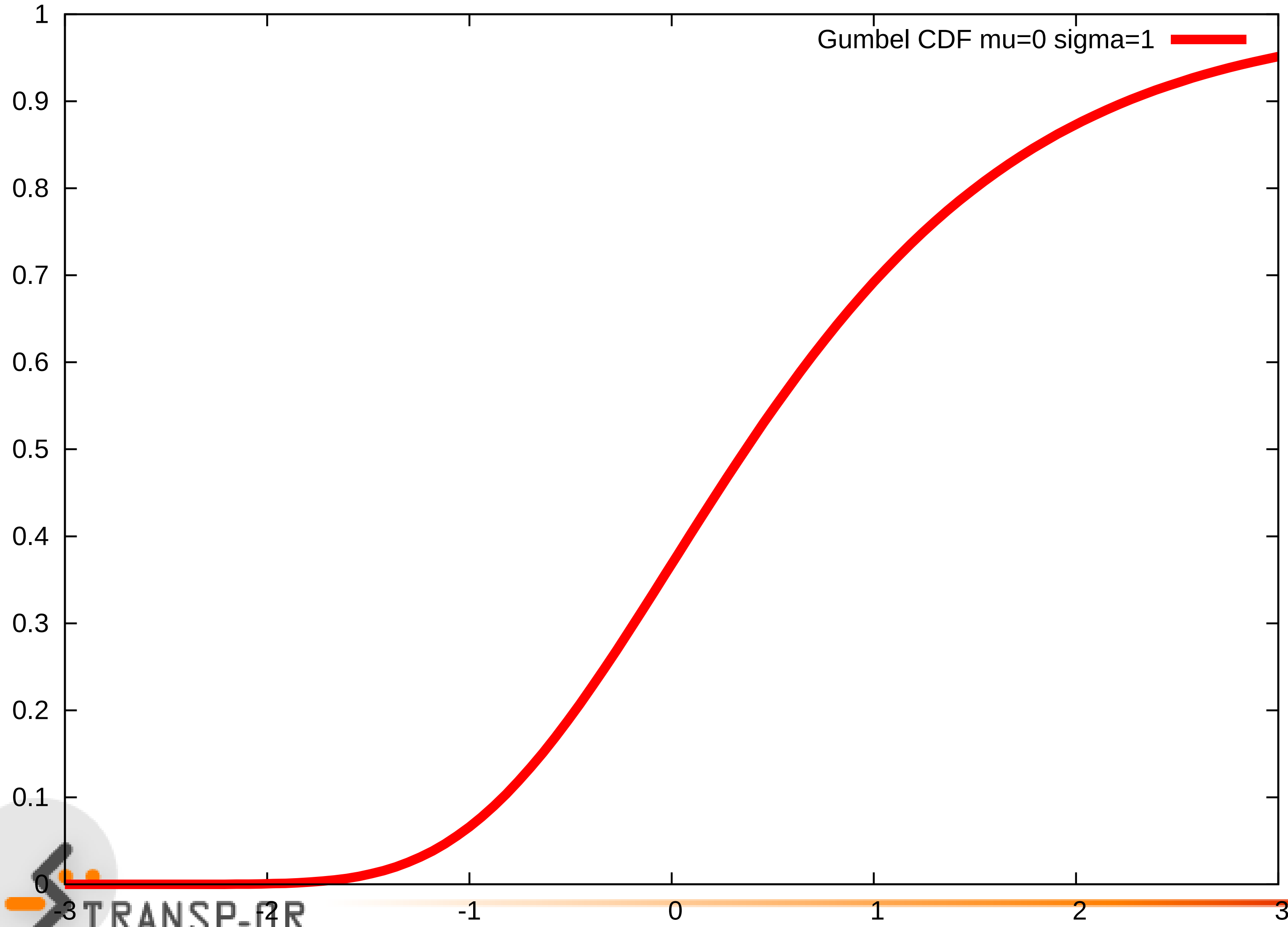
$$-\ln G(x) = \text{exp}\left(h(0) - \frac{x}{\sigma'}\right) = \text{exp}\left(-\frac{x - \sigma' h(0)}{\sigma'}\right)$$

$$G(x) = \text{exp}\left(-\text{exp}\left(-\frac{x - \sigma' h(0)}{\sigma'}\right)\right)$$

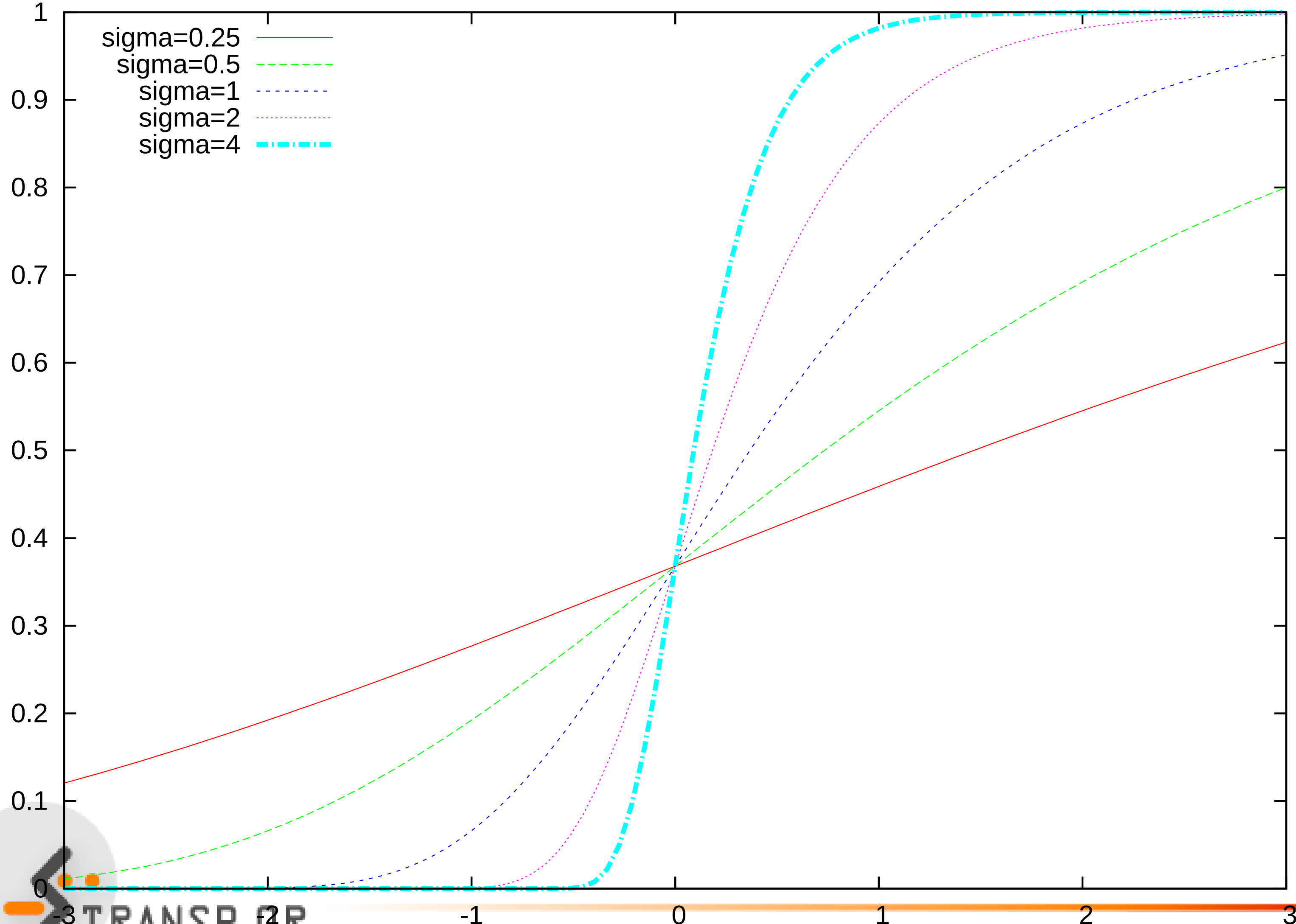
Let $\sigma = 1/\sigma'$ and $\mu = \sigma' h(0) = \ln(-\ln G(0))/\sigma$

$$G(x) = \text{exp}\left(-\text{exp}\left(-\sigma(x-\mu)\right)\right)$$

Extreme value distribution



Extreme value distribution



Extreme value distribution

Type I Extreme Value Distribution or Gumbel Distribution

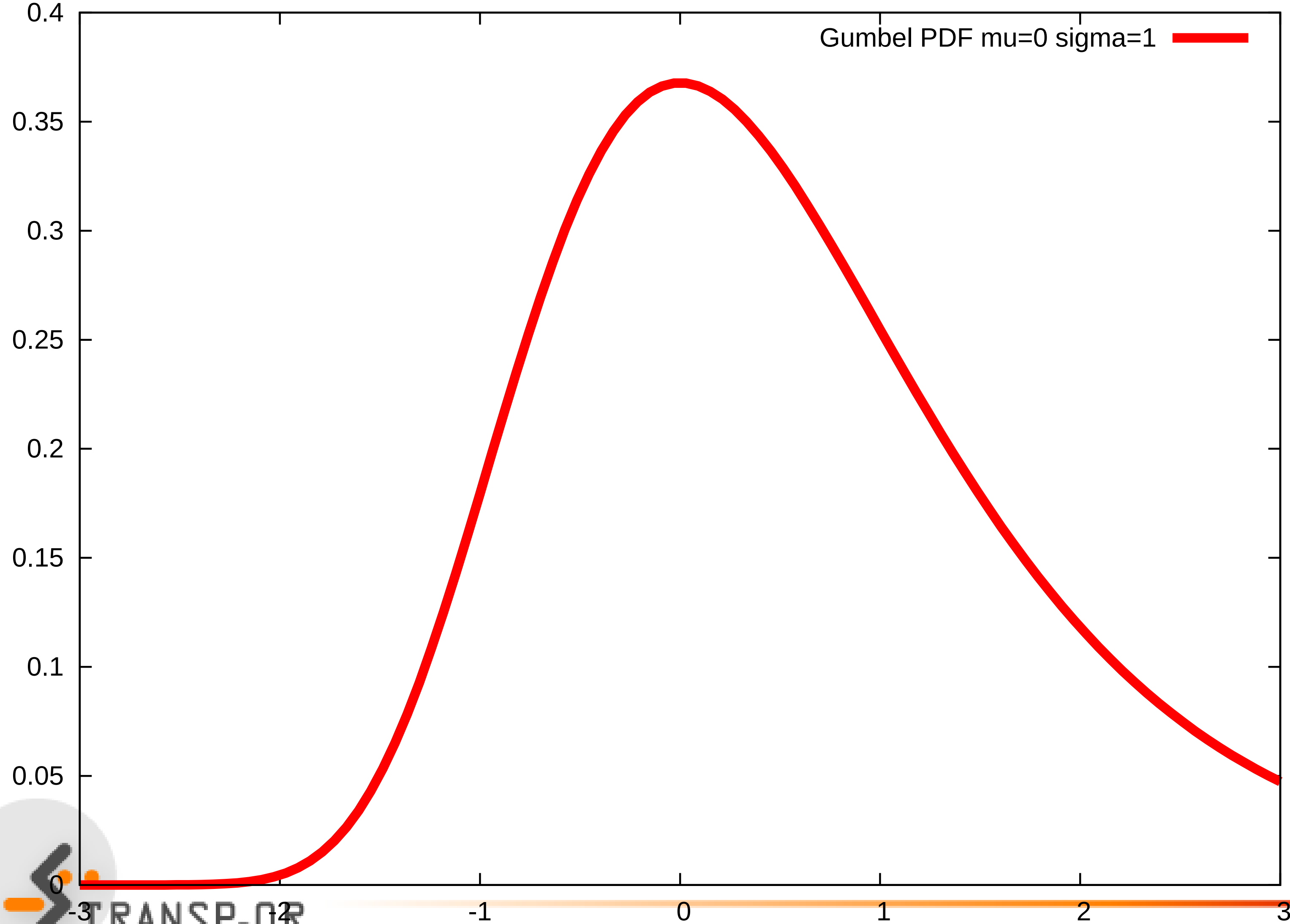
- $X \sim EV(\mu, \sigma)$
- Location parameter: μ
- Scale parameter: $\sigma > 0$
- CDF: closed form

$$F_X(x) = \text{exp}\left(-e^{-\sigma(x-\mu)}\right)$$

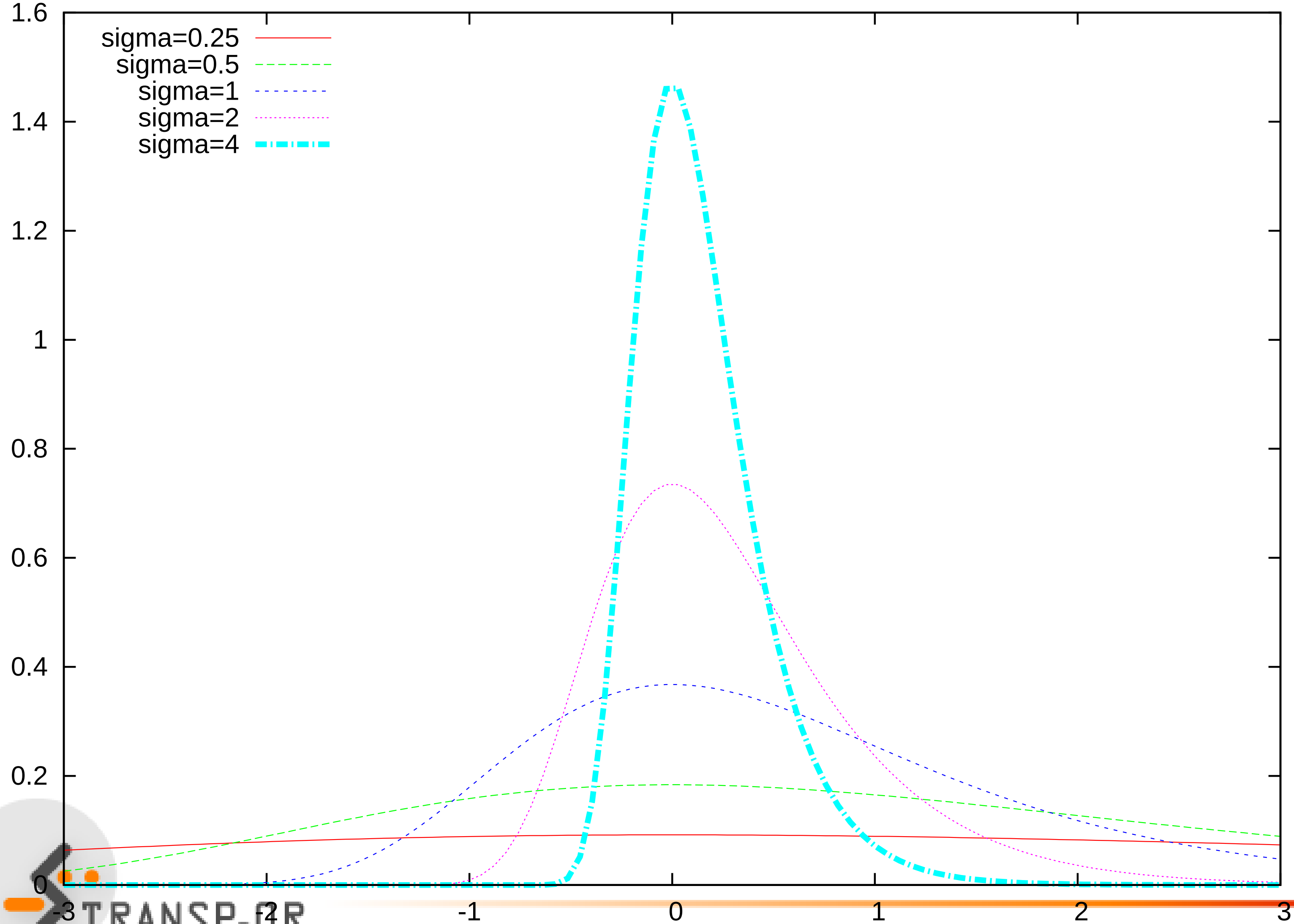
- PDF

$$f_X(x) = \sigma e^{-\sigma(x-\mu)} \text{exp}\left(-e^{-\sigma(x-\mu)}\right)$$

Extreme value distribution



Extreme value distribution



TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Extreme value distribution

Properties

- Mode: μ
- Mean: $\mu + \gamma/\sigma$ where γ is Euler's constant

$$\gamma = -\int_0^{+\infty} e^{-x} \ln x dx = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n \right) \approx 0.57721566$$

- Variance: $\pi^2/6\sigma^2$

Extreme value distribution

Properties (ctd)

- Let $X \sim EV(\mu, \sigma)$, $\alpha > 0$ and $\beta \in \mathbb{R}$. Then

$$\alpha X + \beta \sim EV(\alpha\mu + \beta, \sigma/\alpha)$$

- Let $X_1 \sim EV(\mu_1, \sigma)$ and $X_2 \sim EV(\mu_2, \sigma)$

$$X = X_1 - X_2 \sim \text{Logistic}(\mu_2 - \mu_1, \sigma)$$

that is

$$F_X(x) = \frac{1}{1 + \exp(-\sigma(x - (\mu_2 - \mu_1)))}$$

Extreme value distribution

Properties (ctd)

- Let $X_1 \sim EV(\mu_1, \sigma)$ and $X_2 \sim EV(\mu_2, \sigma)$

$$X = \max(X_1, X_2) \sim EV\left(\frac{1}{\sigma}\ln(e^{\sigma\mu_1} + e^{\sigma\mu_2}), \sigma\right)$$

- Let $X_i \sim EV(\mu_i, \sigma)$, $i = 1, \ldots, n$

$$X = \max(X_1, \ldots, X_n) \sim EV\left(\frac{1}{\sigma}\ln\sum_{i=1}^{n} e^{\sigma\mu_i}, \sigma\right)$$

- The sum of two EV r.v. is not an EV r.v.

Estimation

- Families of models with parameters
- Estimation: approximate parameters from a random sample
- Estimator: random variable
- Classical methods: maximum likelihood, method of moments (least squares)

Estimation

Likelihood function

Let $x_1, \ldots, x_n$ *be a realization of a random sample* $X_1, \ldots, X_n$ *from* $f_X(x; \theta)$*, where* $\theta \in \mathbb{R}^p$ *is a vector of unknown parameters. The function* $L : \mathbb{R}^p \to [0, 1]$

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

provides the likelihood *of the sample as a function of* θ*.*

Estimation

Maximum likelihood estimate

Let $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ from $f_X(x;\theta)$, where $\theta \in \mathbb{R}^p$ is a vector of unknown parameters. If $\hat{\theta}$ is such that

$$L(\hat{\theta}) \geq L(\theta)$$

for all possible values of θ, then $\hat{\theta}$ is called the maximum likelihood estimate *for θ.*

Note: it is computationally easier to maximize

$$\ln L(\theta) = \ln \prod_{i=1}^{n} f_X(x_i;\theta) = \sum_{i=1}^{n} \ln f_X(x_i;\theta)$$

where $\ln L : \mathbb{R}^p \to]-\infty, 0]$

Properties of estimators

Unbiasedness

Let $X_1, \ldots, X_n$ *be a random sample from* $f_X(x;\theta)$*. An estimator* $\hat{\theta}$ *is said to be* unbiased *if*

$$E(\hat{\theta}) = \theta.$$

Properties of estimators

Efficiency (scalar)

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for $\theta \in \mathbb{R}$. If

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

then $\hat{\theta}_1$ is more efficient *than $\hat{\theta}_2$.*

Efficiency (vector)

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for $\theta \in \mathbb{R}^p$. If the matrix

$$\mathrm{Var}(\hat{\theta}_2) - \mathrm{Var}(\hat{\theta}_1)$$

is positive definite, then $\hat{\theta}_1$ is more efficient *than $\hat{\theta}_2$. We note*

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

Properties of estimators

Cramer-Rao bound (scalar)

Let $X_1, \ldots, X_n$ be a random sample from $f_X(x;\theta)$, and $\hat{\theta}$ an unbiased estimator of $\theta \in \mathbb{R}$. Under appropriate assumptions,

$$\begin{aligned}\operatorname{Var}(\hat{\theta}) &\geq \left(-nE\left[\frac{\partial^2 \ln f_X(x;\theta)}{\partial\theta^2}\right]\right)^{-1} \\ &= \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta^2}\right]\right)^{-1}\end{aligned}$$

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Properties of estimators

Cramer-Rao bound (vector)

Let $X_1, \ldots, X_n$ *be a random sample from* $f_X(x;\theta)$, *and* $\hat{\theta}$ *an unbiased estimator of* $\theta \in \mathbb{R}^p$. *Under appropriate assumptions,*

$$\mathrm{Var}(\hat{\theta}) \geq -E[\nabla^2 \ln L(\theta)]^{-1}$$

that is

$$\mathrm{Var}(\hat{\theta}) + E[\nabla^2 \ln L(\theta)]^{-1}$$

is positive definite. The matrix

$$-E[\nabla^2 \ln L(\theta)]$$

is called the information matrix.

Asymptotic properties of estimators

Consistency

An estimator $\hat{\theta}_n$ is said to be consistent for θ if it converges in probability to θ, that is $\forall \varepsilon > 0$,

$$\lim_{n \to 0} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1.$$

Asymptotic properties of estimators

Under fairly general assumptions, maximum likelihood estimators are

- consistent
- asymptotically normal
- asymptotically efficient (asymptotic variance = Cramer-Rao bound)

Warning: large sample properties

Estimator of the asymptotic variance for ML

- Cramer-Rao Bound with the estimated parameters

$$\hat{V} = -\nabla^2 \ln L(\hat{\theta})^{-1}$$

- Berndt, Hall, Hall & Haussman (BHHH) estimator

$$\hat{V} = \left(\sum_{i=1}^{n} \hat{g}_i \hat{g}_i^T \right)^{-1}$$

where

$$\hat{g}_i = \frac{\partial \ln f_X(x_i;\theta)}{\partial \theta}$$

Hypothesis test: t-test

Is the estimated parameter $\hat{\theta}$ significantly different from a given value θ^*?

- $H_0 : \hat{\theta} = \theta^*$
- $H_1 : \hat{\theta} \neq \theta^*$

Under H_0, if $\hat{\theta}$ is normally distributed with known variance σ^2

$$\frac{\hat{\theta} - \theta^*}{\sigma} \sim N(0, 1).$$

Therefore

$$P(-1.96 \leq \frac{\hat{\theta} - \theta^*}{\sigma} \leq 1.96) = 0.95 = 1 - 0.05$$

Hypothesis tests

$$P(-1.96 \leq \frac{\hat{\theta}-\theta^*}{\sigma} \leq 1.96) = 0.95 = 1 - 0.05$$

H_0 can be rejected at the 5% level if

$$\left|\frac{\hat{\theta}-\theta^*}{\sigma}\right| \geq 1.96.$$

- If $\hat{\theta}$ asymptotically normal
- If variance unknown
- A t test should be used with n degrees of freedom.
- When $n \geq 30$, the Student t distribution is well approximated by a $N(0,1)$

Hypothesis tests

- Let $X_1, \ldots, X_n$ be a random sample from $f_X(x;\theta)$, $\theta \in \mathbb{R}^p$
- $\hat{\theta}_U \in \mathbb{R}^p$ is the maximum likelihood estimator.
- $\hat{\theta}_R \in \mathbb{R}^q$, $q < p$, is the ML estimator of a restricted model.
  - e.g. $\theta_1 = \theta_2 = \ldots = \theta_p$
- H_0 : the restrictions are correct
- Under H_0,

$$-2(\ln L(\theta_R) - \ln L(\theta_U)) = -2 \ln \frac{L(\theta_R)}{L(\theta_U)} \sim \chi^2(p-q)$$