# Plug-in estimator and the delta method

Olga Klopp

In previous course we have seen:

- Statistical Experience, statistical model, sampling
- Empirical distribution function :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\big(X_i \le x\big), \quad x \in \mathbb{R}$$

and some **asymptotic** properties:

$$\widehat{F}_n(x) \xrightarrow{a.s.} F(x), \qquad \left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{a.s.} 0$$

their convergence rate:

$$\sqrt{n}\big(\widehat{F}_n(x) - F(x)\big) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x))),$$
$$\sqrt{n}\left\|\widehat{F}_n(x) - F(x)\right\|_\infty \xrightarrow{d} K$$

- **non-asymptotic** properties thanks to Chebyshev.

## 1 Estimation of functionals in the sampling model

- **Goal:** estimate a characteristic scalar $T(F)$ of an unknown probability distribution function $F$ from a random sample of size $n$ $X_1, \dots, X_n \overset{iid}{\sim} X \sim F$

$$\text{data:} \quad X_1, \dots, X_n \overset{i.i.d}{\sim} F \quad \rightsquigarrow \quad \text{problem:} \quad \text{estimate} \quad T(F)$$

- Examples:
  - Already seen: value in one fixed point $T(F) = F(x) = \mathbb{P}(X \le x) = \mathbb{E}\left[I(X \le x)\right]$
  - Regular Functional:

$$T(F) = h\left(\int_{\mathbb{R}} g(x)dF(x)\right) = h\big(\mathbb{E}\, g(X)\big)$$

where $g, h : \mathbb{R} \to \mathbb{R}$ are regular and $X \sim F$

- Examples of regular functions:
  - Mean : $T(F) = m(F) = \int_{\mathbb{R}} x dF(x) = \mathbb{E}(X)$.
  - Variance :

$$T(F) = \sigma^2(F) = \int_{\mathbb{R}} \big(x - m(F)\big)^2 dF(x) = \mathbb{E}\left(X - \mathbb{E}\, X\right)^2$$

It measures how far a set of realizations of a r.v. are spread out from their average value.

– Skewness :

$$T(F) = \alpha(F) = \frac{\int_{\mathbb{R}} \big(x - m(F)\big)^3 dF(x)}{\sigma^3(F)} = \frac{\mathbb{E}(X - \mathbb{E}\,X)^3}{\sigma^3(F)}$$

It is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. For example, for a unimodal distribution, negative skew indicates that the tail on the left side of the probability density function is longer or fatter than the right side (it does not distinguish these two kinds of shape).

– Kurtosis :

$$T(F) = \kappa(F) = \frac{\int_{\mathbb{R}} \big(x - m(F)\big)^4 dF(x)}{\sigma^4(F)} = \frac{\mathbb{E}\,\big(X - \mathbb{E}\,X\big)^4}{\sigma^4(F)}$$

In a similar way to the concept of skewness, kurtosis is a descriptor of the shape of a probability distribution. Greater peakedness and heavy tailedness often tend to be seen when kurtosis is higher.

### 1.1 Examples of non-regular functionals

**Definition 1.** *Let $X$ be a r.v. with cdf $F$ and $0 < p < 1$. We call* **$p$-quantile** *of $X$:*

$$q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

A quantile is the proportion of cases we find below a certain value. It is the application of the quantile function (the inverse function of the cumulative distribution function)

- when $F$ is **continuous and strictly increasing** the **$p$-quantile** is the only solution of

$$F(q_p) = p \qquad (\text{ that is } \; q_p = F^{-1}(p)\;).$$

- the **median** $= \text{med}(F) = q_{1/2}(F)$ is the value such that the random variable that is equally likely to fall above or below it. If a distribution is symmetric, then the median is the mean (so long as the latter exists). But, in general, the median and the mean can differ.

- les **quartiles** $= \{q_{1/4}(F), \text{med}(F), q_{3/4}(F)\}$

## 2 Plug-in estimator

A very simple idea: to evaluate the same functionals at the empirical distribution based on a sample:

**Definition 2.** *The plug-in estimator of $T(F)$ is the estimator $T(\widehat{F}_n)$.*

- when $T(F) = h\big(\mathbb{E}\,g(X)\big)$ then the *plug-in* estimator of $T(F)$ is:

$$\boxed{T(\widehat{F}_n) = h\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\Big)}$$

- when $T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$, then the *plug-in* estimator is the empirical quantile :

$$\boxed{T(\widehat{F}_n) = \inf\{x \in \mathbb{R} : \widehat{F}_n(x) \geq p\}}$$

E.g., the median is the value separating the higher half of a data sample from the lower half: $\{1, 3, 3, \mathbf{6}, 7, 8, 9\}$

**Exercise** Prove $\mathbb{E}(\widehat{F}_n(x)) = F(x)$ and $\text{Var}(\widehat{F}_n(x)) = \dfrac{F(x)(1 - F(x))}{n}$.

## 2.1 Asymptotic performance of the plug-in estimator for the estimation of regular functionals of the form $T(F) = h\big(\mathbb{E}\, g(X)\big)$

Convergence (consistency) : if $g, h : \mathbb{R} \to \mathbb{R}$, $h$ continuous and $\mathbb{E}\,|g(X)| < \infty$, then $T(\widehat{F}_n) \overset{\text{a.s.}}{\to} T(F)$ (Law of large numbers + continuous map theorem).

Convergence rate (asymptotic normality) :

1. CLT :
$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}\, g(X)\Big) \xrightarrow{d} \mathcal{N}\big(0, \mathrm{Var}\big[g(X)\big]\big)$$
where $\mathrm{Var}\big[g(X)\big] = \mathbb{E}\left[(g(X) - \mathbb{E}\, g(X))^2\right]$

2. We have a result of the type $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$. How to transfer this result to $\sqrt{n}(h(Z_n) - h(c_1)) \xrightarrow{d}$ ?

**Theorem 1** (**Delta Method**). *Let $(Z_n)$ be a sequence of r.v. and $V$ a r.v. such that*
$$a_n(Z_n - c_0) \xrightarrow{d} V$$
*where $(a_n)$ is a sequence of real positives numbers tending to $+\infty$ and $c_0$ a constant. Let $h : \mathbb{R} \to \mathbb{R}$ be a function* **continues and differentiable at** $c_0$. *Then*
$$\boxed{a_n\big(h(Z_n) - h(c_0)\big) \xrightarrow{d} h'(c_0)V}$$

1. if $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$ and $h$ differentiable at $c_1$ then
$$\boxed{\sqrt{n}\big(h(Z_n) - h(c_1)\big) \xrightarrow{d} \mathcal{N}\big(0, c_2[h'(c_1)]^2\big)}$$

2. if $V \sim \mathcal{N}(\mu, v)$ and $a \in \mathbb{R}$ then $aV \sim \mathcal{N}(a\mu, a^2 v)$.

3. the central idea of the proof of the Delta method is a the first order Taylor approximation formula for $h$ at $c_0$ : when $n \to \infty$
$$a_n(h(Z_n) - h(c_0)) \approx h'(c_0)\big[a_n(Z_n - c_0)\big] \approx h'(c_0)V.$$

**Proposition 1** (**Asymptotic normality of the estimator plug-in in the case of regular functional**).

*If $\mathbb{E}[g(X)^2] < +\infty$ and $h$ is a continuous and differentiable function in $\mathbb{E}\, g(X)$, then*
$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$
*where* $v(F) = h'\big(\mathbb{E}\left[g(X)\right]\big)^2 \mathrm{Var}\big[g(X)\big]$.

To build a confidence interval, we would like replace $v(F)$ by $v(\widehat{F}_n)$ : when $h$ is $\mathcal{C}^1$, we can show that $v(\widehat{F}_n) \xrightarrow{\mathbb{P}} v(F)$ and, using Slutsky's theorem,
$$\boxed{\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{v(\widehat{F}_n)^{1/2}} \xrightarrow{d} \mathcal{N}\big(0, 1\big)}$$

## 2.2 Another application of the Delta method: variance stabilization

Standard statistical techniques often assume that data are normally distributed, with constant variance not depending on the mean of the data. In some applications these assumptions are violated, e.g. gene-expression microarray data have a complicated error structure, with a variance that changes with the mean in a non-linear fashion. Data that violate these assumptions can often be brought in line with the assumptions by application of a transformation.

In applied statistics, a **variance-stabilizing transformation** is a data transformation that is specifically chosen to simplify data analysis. The aim behind the choice of a variance-stabilizing transformation is to find a simple function $f$ to apply to values $x$ in a data set to create new values $y = (x)$ such that the variability of the values $y$ is not related to their mean value. Variance-stabilizing transformations are well known for certain parametric families of distributions, such as the Poisson and the binomial distribution, some types of data analysis proceed more empirically

- Let $X_1, \ldots, X_n$ be a random sample of size $n$ from **Exponential distribution** with parameter $\theta \in [0, 1]$. The exponential distribution is the probability distribution that describes the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. In addition to being used for the analysis of Poisson point processes it is found in various other applications.
  - The probability density function (pdf) of an exponential distribution is $f(\theta, x) = \theta \exp(-\theta x) I(x > 0)$, with mean $\mathbb{E}_\theta X = 1/\theta$ and variance $\mathrm{Var}_\theta X = 1/\theta^2$
  - CLT : $\sqrt{n}\big(\bar{X}_n - 1/\theta\big) \xrightarrow{d} \mathcal{N}(0, 1/\theta^2)$
  - **Pb.:** Asymptotic variance depends on the unknown parameter $\theta$
  - Delta method: if $g$ is $\mathcal{C}^1$ then :
$$\sqrt{n}\big(g(\bar{X}_n) - g(1/\theta)\big) \xrightarrow{d} \mathcal{N}\big(0, (g'(1/\theta))^2/\theta^2\big)$$
  - in particular for $g(\theta) = \log(\theta)$, we have
$$\sqrt{n}\big(g(\bar{X}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}(0, 1)$$
- Let $X_1, \ldots, X_n$ be a random sample of size $n$ from **Bernoulli distribution** with parameter $\theta \in [0, 1]$.
  - CLT : $\sqrt{n}\big(\bar{X}_n - \theta\big) \xrightarrow{d} \mathcal{N}(0, \theta(1-\theta))$
  - Asymptotic variance depends on the unknown parameter $\theta$
  - Delta Method: if $g$ is $\mathcal{C}^1$ then :
$$\sqrt{n}\big(g(\bar{X}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\big(0, (g'(\theta))^2\theta(1-\theta)\big)$$
  - in particular, for $g(\theta) = 2arcsin(\sqrt{\theta})$, we have
$$\sqrt{n}\big(g(\bar{X}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}(0, 1)$$

  Recall: $\arcsin : [-1, 1] \to [-\pi/2, \pi/2]$ and $\arcsin'(x) = \dfrac{1}{\sqrt{1-x^2}}$.

**Example R:** variance-stabilizing transformation

# 3 Empirical Quantiles and Applications

"Theoretical" $p$-Quantile:
$$T(F) = q_p(F) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$$

Empirical $p$-quantile:
$$T(\widehat{F}_n) = \widehat{q}_{n,p} = \inf\{x \in \mathbb{R} : \widehat{F}_n(x) \geq q\}$$

Question : What are the statistical properties of the procedure of the estimation of $q_p(F)$ using $\widehat{q}_{n,p}$? (Pb.: we are no longer in the regular case!)

## 3.1 Empirical Quantiles: Explicit expression using order statistics

**Definition 3.** *Let $X_1, \dots, X_n$ be a random sample of size $n$. We call* **order statistics** *the $n$ statistics $X_{(1)}, \dots, X_{(n)}$ which are constructed in the following way:*

$$X_{(1)} \leq \cdots \leq X_{(n)}$$

1. for $p$-quantile $(0 < p < 1)$:

$$\boxed{\widehat{q}_{n,p} = X_{(k)} = X_{(\lceil np \rceil)} \text{ where } \quad \frac{k-1}{n} < p \leq \frac{k}{n}}$$

2. in particular, the empirical median verifies:

$$\boxed{\widehat{q}_{n,1/2} = \text{med}(\widehat{F}_n) = X_{(\lceil n/2 \rceil)}} \text{ where } \lceil t \rceil = \min(n \in \mathbb{N} : n \geq t)$$

## 3.2 The boxplot: a synthetic representation of the dispersion of the data

Boxplots are really useful ways to display the data. At the centre of the plot is the median, which is surrounded by a box the top and bottom of which are the limits within which the middle 50% of observations fall (the interquartile range). Sticking out of the top and bottom of the box are two whiskers that extend to one and a half times the interquartile range. Like histograms, they also tell us whether the distribution is symmetrical or skewed. If the whiskers are the same length then the distribution is symmetrical (the range of the top and bottom 25% of scores is the same); however, if the top or bottom whisker is much longer than the opposite whisker then the distribution is asymmetrical. Data beyond whiskers are considered as *outliers*. Whiskers:

$$X_* = \min\{X_i : |X_i - \hat{q}_{n,1/4}| \leq 1,5\mathcal{I}_n\},$$
$$X^* = \max\{X_i : |X_i - \hat{q}_{n,3/4}| \leq 1,5\mathcal{I}_n\}.$$

Interquartile range:

$$\mathcal{I}_n = \hat{q}_{n,3/4} - \hat{q}_{n,1/4}.$$

**R-example: Boxplot**

## 3.3 The Q-Q plot

*Q - Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other*

Assume that we are given a random sample of size $n$: $X_1, \dots, X_n$ and a cdf $F_{ref}$, we want to test if the following hypothesis is true:

$$(H_0) \qquad \text{“The } X_i \text{ are distributed according to } F_{ref}\text{”}$$

To ”accept or refuse visually” this hypothesis, one can draw the qq-plot: it's the point cloud

$$\boxed{\left(q_{i/n}(F_{ref}), \widehat{q}_{n,i/n}\right)_{i=1}^n = \left(q_{i/n}(F_{ref}), X_{(i)}\right)_{i=1}^n}$$

1. if the point cloud approximately lies on the line $y = x$, then the hypothesis is accepted (we also draw the line $y = x$ on a qq-plot)

2. if the points approximately lie on a straight line then the assumption is true up to a transformation (centering and scaling); usually, we normalize the data

**R-example: Q-Q plot**

### 3.4 Convergence of empirical quantiles

**Theorem 2.** *Let $X$ be a r.v. (we denote by $F$ its cdf) with probability density function $f_X$. We assume that $f_X$ is strictly positive a.s. on the interval $I \subset \mathbb{R}$ and zero outside this interval. Let $0 < p < 1$. We have*

$$\boxed{\widehat{q}_{n,p} \xrightarrow{a.s.} q_p(F) = q_p}$$

*If in addition the density function $f_X$ of $X$ is continuous at $q_p$ then $\widehat{q}_{n,p}$ is asymptotically Gaussian:*

$$\boxed{\sqrt{n}(\widehat{q}_{n,p} - q_p) \xrightarrow{d} \mathcal{N}\Big(0, \frac{p(1-p)}{f_X(q_p)^2}\Big)}$$

**Asymptotic variance** of $\widehat{q}_{n,p}$ is given by

$$\boxed{\frac{p(1-p)}{f_X(q_p)^2}}$$

The quantity $f_X(q_p)$ is unknown.

- As $\widehat{q}_{n,p}$ is **strongly consistent** (converges a.s.) and $f_X$ is continuous in $q_p$,

  $$f_X(\widehat{q}_{n,p}) \xrightarrow{a.s.} f_X(q_p)$$

  Using Slustky we can replace $q_p$ by $\widehat{q}_{n,p}$:

  $$\frac{\sqrt{n} f_X(\widehat{q}_{n,p})}{\sqrt{p(1-p)}}(\widehat{q}_{n,p} - q_p) \xrightarrow{d} \mathcal{N}(0,1)$$

- But $f_X(\widehat{q}_{n,p})$ is also unknown! (density estimation problem)

# 4 Limitations of the plug-in approach

The estimation of $T(F)$ using $T(\widehat{F}_n)$is not always possible:

- Example : if $F$ has a continuous density $f$ and we want to estimate it for a given $x_0$:

  $$T(F) = f(x_0) = F'(x_0),$$

  we **can not take** as an estimator $\widehat{F}_n'(x_0)$ as $\widehat{F}_n$ is piecewise.

The estimation of $T(F)$ using $T(\widehat{F}_n)$ is not always desirable :

- Often we have additional information: $F$ belongs to a particular subclass of distributions (the model) and there are more suitable choices than the plug-in estimator (see following courses).