Advanced Statistics II: Non Parametric Tests

Aurélien Garivier

ParisTech

February 27, 2011

Outline

A First Motivation: Model Validation

In a Gaussian linear model

$$\forall i = 1, \ldots, n \qquad Y_i = \sum_{j=1}^{p} \theta_j X_{i,j} + \sigma \epsilon_i$$

it is assumed that the ϵ_i are i.i.d. $\mathcal{N}(0,1)$

$\Longrightarrow$ Can we *check* that this is the case ?

Two aspects:

- identically distributed (residual vs fitted value)
- gaussian distribution : Q-Q plots

Key Remarks

Prop: if X has a continuous Cumulative Distribution Function (CDF) $F : t \mapsto P(x \leq t)$, then $F(X) \sim \mathcal{U}[0,1]$.

Consequence: if $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(0,1)$, then the distribution of $F(X_1), \dots, F(X_n)$ are i.i.d. $\mathcal{U}[0,1]$ (they are *free* of F).

Definition: The *order statistics* of an n-uple $(U_1, \dots, U_n)$ is the n-uple $(U_{(1)}, \dots, U_{(n)})$ such that

$$\{U_1, \dots, U_n\} = \{U_{(1)}, \dots, U_{(n)}\} \quad \text{and} \quad U_{(1)} \leq \dots \leq U_{(n)}$$

Prop: If $U_{(1)}, \dots, U_{(n)}$ is the order statistics of i.i.d $\mathcal{U}[0,1]$ random variables, then

$$\mathbb{E}[U_{(i)}] = \frac{i}{n+1}$$

Free statistic

Definition A statistic $S = S(X_1, \ldots, X_n)$ is *free* if its distribution depends only on n and not on the distribution of $(X_1, \ldots, X_n)$.

Free statistics are useful to build (non-parametric) tests

The permutation sorting a sample is a free statistic (uniformly distributed).

Q-Q plots

Prop: Let $X_1, \ldots, X_n$ be i.i.d. random variables, with cdf F, and let G be a cdf. Consider thepoints

$$\left\{ \left(F^{-1}\left(\frac{i}{n+1}\right), X_{(i)} \right), 1 \leq i \leq n \right\}$$

- if $F = G$, then the points are approximately lying on the first diagonal.
- If $F \neq G$, then (at least some of) these points deviate from the first diagonal.

Remark: in practice, for Gaussian variables one may use $F^{-1}((i - 0.375)/(n + 0.25))$ for better performance.

Kolmogorov-Smirnov Statistic

Defintion The *empirical distribution function* F_n of $X_1, \dots, X_n$ is the mapping $\mathbb{R} \to [0, 1]$ defined by

$$F_n(t) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_i \le t\}} .$$

It is the CDF of the empirical measure $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$.

Definition The *Kolmogorov-Smirnov Statistic* between the sample $X_1, \dots, X_n$ and the CDF G is

$$D_n(Z_{1:n}, G) = \sup_{t \in \mathbb{R}} |G(t) - F_n(t)|$$

Properties of the K-S statistic

Prop: The KS statstic can be computed by:

$$D_n(X_{1:n}, G) = \max_{1\leq i\leq n} \max\left\{\left|G(X_{(i)}) - \frac{i-1}{n}\right|, \left|G(X_{(i)}) - \frac{i}{n}\right|\right\}$$

Prop: if F is the CDF of X_i, then $D_n(X_{1:n}, F)$ is a *free* statistic, and it has the distribution of

$$\sup_{u\in[0,1]}\left|u - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{U_i\leq u\}}\right|$$

where the U_i are i.i.d. $\mathcal{U}[0,1]$.

Glivenko-Cantelli Theorem and K-S limiting distribution

Theorem: If F denotes the CDF of X_i, then

$$D_n(X_{1:n}, F) \to 0 \quad a.s.$$

as n goes to infinity. If $F \neq G$, then $D_n(X_{1:n}, G)$ remains lower-bounded as n goes to infinity.

Theorem: As n goes to infinity,

$$P\left(D_n(X_{1:n}, F) > \frac{c}{\sqrt{n}}\right) \to 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2r^2c^2} .$$

The RHS is 5% when $c = 1.36$.

Comparison of two samples

Let $X_1, \ldots, Y_m$ and $Y_1, \ldots, Y_n$ be two samples, with CDF respectively F and G, and with empirical CDF F_m and G_n.

Definition The *Kolmogorov-Smirnov Statistic* between the sample $X_1, \ldots, X_n$ the sample $Y_1, \ldots, Y_n$

$$D_{m,n}(X_{1:m}, Y_{1:n}) = \sup_t |G_n(t) - F_m(t)|$$

Prop: Asymptotic behavior of the K-S statistic:

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{m+n}} D_{m,n}(X_{1:m}, Y_{1:n}) > \frac{c}{\sqrt{n}}\right) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\mathrm{e}^{-2r^2c^2}$$

Outline

Fitting a distribution

Rank Tests for the comparison of two samples

- Two unrelated samples: Mann-Whitney signed-rank test
- Two related samples: Wilcoxon signed-rank test

Bootstrap

Unrelated samples

Assume that $(X_i)_{1 \le i \le m}$ and $(Y_j)_{1 \le j \le n}$ are two *independent* samples with CDF, respectively, F and G.

The goal is to test $H_0 : F = G$ against $H_1 : P(Y_j > X_i) \neq 1/2$.

The alternative hypothese is more restrictive than for the K-S test.

The idea is that, under H_0, the X_i and Y_i are "tangled" while, under H_1, the X_i tend to be smaller (or larger) than the Y_j.

Mann-Whitney Statistic

The Mann-Whitney Statistic is defined as

$$U_{m,n} = \sum_{\substack{i=1..m \\ j=1..n}} \mathbf{1}_{\{Y_j > X_i\}}$$

Prop: Under H_0, $U_{m,n}$ is a free statistic,

$$\mathbb{E}[U_{m,n}] = \frac{mn}{2} \text{ and } \mathrm{Var}(U_{m,n}) = \frac{mn(m+n+1)}{12} \ .$$

Besides, when $m, n \to \infty$,

$$\zeta_{m,n} = \frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \rightharpoonup \mathcal{N}(0,1) \ ,$$

while, under H_1, $|\zeta_{m,n}| \to \infty$.

Mann-Whitney Test

Testing procedures are derived as usual. The Mann-Whitney test can also be used with unilateral alternatives

For effective computation, observe that $U_{m,n}$ can be obtained as follows:

- sort all the elements of $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ in increasing order
- define $R_{m,n}$ to be the sum of the ranks of all the elements $\{Y_1, \ldots, Y_n\}$
- then

$$U_{m,n} = R_{m,n} - \frac{n(n+1)}{2} \, .$$

Student, K-S or Mann-Whitney ?

Student's test is the most powerful, but it has strong requirements (normality of the two samples, equality of the variances).

The Mann-Whitney test is non-parametric and more robust. In the normal case, its relative efficiency wrt. Student's test is about 96%. Besides, it can be used on ordinal data.

The K-S test has a more general alternative hypothese. However, it is less powerful.

$\Longrightarrow$ If normality cannot be assumed, and if its alternative hypothese is sufficiently discriminating, use Mann-Whitney

Warning: in R, this test is implemented under name `wilcox.test`.

Related samples

Let $(X_i, Y_i), 1 \leq i \leq n$ be independent, identically distributed pairs of real-valued random variables.

We assume that the CDF of the X_i is F, while the CDF of the Y_i is $t \mapsto F(t-\theta)$ for some real θ.

In other words, Y_i has the same distribution as $X_i + \theta$.

Goal: we want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$.

Example: evolution of the blood pressure after administration of a drug, double correction of a test

Wilcoxon statistic

Definition: for $1 \leq i \leq n$, let $Z_i = X_i - Y_i$. The *Wilcoxon statistic* W_+ is defined by

$$W_n^+ = \sum_{k=1}^{n} k \mathbb{1}_{\{Z_{[k]}>0\}} ,$$

where $Z_{[k]}$ is such that $|Z_{[1]}| \leq |Z_{[2]}| \leq \cdots \leq |Z_{[n]}|$.

Prop: if the distribution of Z is symetric and if $P(Z = 0) = 0$, then the sign $\mathrm{sgn}(Z)$ and the absolute value $|Z|$ are independent.

Prop: under H_0, the variables $\mathbb{1}_{\{Z_{[k]}>0\}}$ are i.i.d. $\mathcal{B}(1/2)$ and

$$\mathbb{E}[W_n^+] = \frac{n(n+1)}{4}, \quad \mathrm{Var}[W_n^+] = \frac{n\,(n+1)\,(2n+1)}{24} .$$

Limiting distribution

Prop: Under H_0, as n goes to infinity,

$$\zeta_n = \frac{W_n^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

converge to the standard normal distribution $\mathcal{N}(0, 1)$.

Prop: Under H_1, ζ_n goes to $-\infty$ (resp. $+\infty$) if $\theta > 0$ (resp. $\theta < 0$).

Remark: the test needs not be bilateral, for example if $H_1 = \theta < 0$ the null hypothese is rejected when W_n^+ is too large.

Outline

"Pulling yourself up by your own bootstraps"

What to do when the classical testing or estimating procedure can't be trusted? When the distribution is strongly non-gaussian? When the amount of data is not sufficient to assume normality?

Idea: create new data *from* the data available !

Principle: resampling

- Given a sample $X_1, \ldots, X_n$ from a distribution P, let P_n be the empirical distribution
- **plug-in**: to estimate a functional $T(P)$, estimate $T(P_n)$!
- Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of $T(P)$
- for k from 1 to N ($N \gg 1$), repeat
  1. **Resampling:** sample $\tilde{X}_1^k, \ldots, \tilde{X}_n^k$ from P_n i.e. from $\{X_1, \ldots, X_n\}$ *with replacement*
  2. compute an estimator $\hat{\theta}_k = \hat{\theta}_k\left(\tilde{X}_1, \ldots, \tilde{X}_n\right)$ of $T(P_n)$.
- **Bootstrap idea**: the empirical distribution of the $\left(\hat{\theta}_k\right)_k$ is close to the distribution of $\hat{\theta}$

Berry-Esseen Theorem

Theorem

Let $X_1, \ldots, X_n$ be iid with $\mathbb{E}[X_i] = 0, \mathbb{E}[X_i^2] = \sigma^2$ and $\mathbb{E}\left[|X_i|^3\right] = \rho < \infty$. If $F^{(n)}$ is the distribution of $(X_1 + \cdots + X_n)/\sigma\sqrt{n}$ and $\mathcal{N}$ is the CDF of the standard normal distribution, then

$$\left|F^{(n)}(x) - \mathcal{N}(x)\right| \leq \frac{3\rho}{\sigma^3\sqrt{n}}$$

Properties of the Bootstrap distribution

- **shape:** because it approximates the sampling distribution, the bootstrap distribution can be used to check normality of the latter.
- **spread:** the standard deviation of the sampling distribution $\mathrm{Var}[P_n]^{1/2}$ is approximately the standard error of the statistic $\hat{\theta}_n$.
- **center:** the bias of the bootstrap distribution mean $T(P_n)$ from the value of the statistic on the sample $\hat{\theta}$ is the same as the bias of $\hat{\theta}_n$ from $T(P)$

Bootstrap Confidence Intervals

- **Bootstrap t-confidence interval:** instead of $\hat{\sigma}/\sqrt{n}$, use the standard deviation of the bootstrap distribution to estimate the deviation of the sampling distribution.
  Requests that its shape is nearly gaussian.
- **Bootstrap percentile confidence interval:** keep as a α-confidence interval the central $1-\alpha$ values of $\left(\hat{\theta}_k\right)_k$
- Example: confidence interval for the mean, regression.

Comparing two groups

Given independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$,

1. Draw a resample of size n of the first sample $X_1, \ldots, X_n$, and a separate resample of size m of the second sample $Y_1, \ldots, Y_m$.
2. Compute the statistic that compares the two groups, such as the difference between the two sample means
3. Repeat the first two steps 10000 times
4. Construct the bootstrap distribution of the statistic.
5. Inspect the shape, bias, and bootstrap error (i.e., the standard deviation of the bootstrap distribution).