Advanced Statistics : Non parametric statistics and Hypothesis Testing

A4 ESILV
Tutorial 4

Exercise 1. Population Growth in Alabama

We measure the population growth (in %) in 9 rural and 7 urban counties in Alabama.

Rural counties	1.1	−21.7	−16.3	−11.3	−10.4	−7.0	−2.0	1.9	6.2
Urban counties	−2.4	9.9	14.2	18.4	20.1	1.2	70.4		

1. Which statistical test(s) can be applied to these data ?
2. We want to determine wether the population growth in rural counties is stochastically smaller that that in urban counties. Perform the appropriate test(s) at significance level $\alpha = 5\%$.
   Instruction: Notice that $\mathbb{P}(D^+_{9,7} \geq 0,571) \leq 0.05$ *where* $D^+_{9,7}$ *is the KS statistic with* $n_1 = 9$ *and* $n_2 = 7$ *and* $\mathbb{P}(W_X \leq 61) \leq 0.05$ *where* W_X *is the statistic of the Wilcoxon sum-rank test.*

Exercise 2. Lilliefors test for Exponential Distribution

Starting from the Kolmogorov−Smirnov test, we want to construct a test that allows verifying whether the random variables $(X_i)_{1\leq i\leq n}$ follow an exponential distribution with parameter $\frac{1}{\lambda}$, where $\lambda > 0$. We denote by F_λ the cumulative distribution function (CDF) of the exponential law with parameter $\frac{1}{\lambda}$, and we define

$$D_n = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_{\hat{\lambda}_n}(x)|, \qquad \text{with } \hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

$\hat{F}_n$ is the empirical c.d.f. of the random variables $(X_i)_{1\leq i\leq n}$. Assume that the random variables $(X_1, \ldots, X_n)$ follow an exponential distribution with parameter $\frac{1}{\lambda}$. Before constructing a table for the law of D_n under this hypothesis, we must verify that the law of D_n does not depend on λ. Let $Y_i = X_i/\lambda$ for $1 \leq i \leq n$.

1. Show that the random variables $(Y_i)_{1\leq i\leq n}$ follow an exponential distribution $\mathcal{E}(1)$.
2. Express D_n as a function of $(Y_i)_{1\leq i\leq n}$.
3. Conclude.

Exercise 3. Anderson−Darling Test

Let $(X_1, \ldots, X_n)$ be an n-sample from a distribution with cumulative distribution function F_0. Let $\hat{F}_n$ be the empirical distribution function associated with this sample. We define the test statistic

$$T_n = n\int_{\mathbb{R}} (\hat{F}_n(x) - F_0(x))^2 \, \psi(F_0(x)) \, dF_0(x) \quad \text{with} \quad \psi(x) = \frac{1}{x(1-x)}.$$

1. Compute $\mathbb{E}\left[\frac{(\hat{F}_n(x)-F_0(x))^2}{F_0(x)(1-F_0(x))}\right]$ and deduce $\mathbb{E}_0(T_n)$.
2. Show that the distribution of T_n does not depend on F_0.

3. Let $U_{(i)} = F_0(X_{(i)})$ for $1 \leq i \leq n$. Show that
$$T_n = -n + \sum_{i=1}^{n} \frac{-2i+1}{n} \ln(U_{(i)}) - \frac{2(n-i)+1}{n} \ln(1 - U_{(i)}).$$

Exercise 4. Choice of the Test

For each of the following cases, determine whether the situation can be reduced to the case of a single-sample observation, i.e: that is, whether the samples are paired or unpaired, and indicate which statistical test(s) between the KS, Wilcoxon rank-sum also called Wilcoxon−Mann−Whitney, Lilliefors, Anderson−Darling, Wilcoxon signed rank, signed rank, Chi-square goodness-of-fit test could be used.

1. Two rain gauges are placed in the same location, and we measure the amount of water collected over about ten rainy days. We want to know whether the two gauges provide comparable data.

2. We compare the results of 20 students on the midterm and the final exam. We want to know whether students performed better on the midterm than on the final exam.

3. We want to know whether a die is fair or biased.

4. We want to know whether the number of errors on a 1000-words page follows a Poisson distribution with parameter $\lambda = 4$.

5. Two groups of students take the same test. We want to know whether the first group performed better than the second.

6. We compare the data collected by two rain gauges, one in Rennes and the other in Paris, over the course of a year. The rainy days are not the same, but we would like to know whether, over the year, rainfall is the same in both cities.

7. We measure the height of students in the same Master's program. We want to know whether these measurements follow a normal distribution (parameters unknown).

Exercise 5. Kernel density estimation

Let $(X_1, \ldots, X_n)$ a n-sample (i.i.d. random variables) with unknown density f that we want to estimate. Let $K : \mathbb{R} \to \mathbb{R}$ an integrable function such that $\int_{\mathbb{R}} K(u)du = 1$. We assume that the function K is positive function bounded by M. An estimator of f is given by
$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$
The parameter $h \in \mathbb{R}_+^*$ is called the bandwidth of the kernel.

1. Show that $\hat{f}_h$ is a probability density over $\mathbb{R}$.

2. Let $x \in \mathbb{R}$ be fixed then show that $\hat{f}_h(x)$ is an unbiased estimator of
$$f_h(x) = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{u-x}{h}\right) f(u)du.$$

3. Let h and x be fixed, then show that $\hat{f}_h(x)$ converge almost surely to $f_h(x)$.

4. We defined the integrated quadratic risk at point x_0 by $\mathbb{E}_f[(\hat{f}_h(x_0) - f(x_0))^2]$. Show that it can be decomposed by
$$\mathbb{E}_f[(\hat{f}_h(x_0) - f(x_0))^2] = b^2(x_0) + v(x_0),$$
where $b(x_0) = f_h(x_0) - f(x_0)$ and $v(x_0) = \mathbb{E}_f[(\hat{f}_h(x_0) - f_h(x_0))^2]$.

5. We assume that $\int_{\mathbb{R}} K(u)^2 du < +\infty$. Show that
$$v(x_0) \leq \frac{C_1}{nh} \quad \text{where } C_1 = M \int_{\mathbb{R}} K^2(u) du.$$
Instruction: we could write $\hat{f}_h(x_0) - f_h(x_0) = \frac{1}{nh} \sum_{i=1}^{n} Y_i(x_0)$ *where* $Y_i(x_0)$ *are i.i.d. and centered random variables.*

6. We assume that f is α-hölder function with $\alpha \in]0, 1]$ and $\int_{\mathbb{R}} |u|^\alpha |K(u)| du < +\infty$. Show that
$$|b(x_0)| \leq C_2 h^\alpha \quad \text{where } C_2 = L \int_{\mathbb{R}} |u|^\alpha |K(u)| du.$$

7. Deduce an optimal choice of the bandwidth h_n^* and the convergence rate of the integrated quadratic risk.