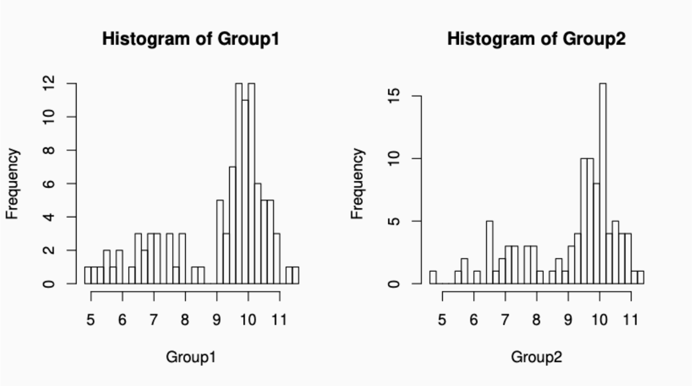

Figure 10.4: Concentration in antibodies for both groups.

- for two-sample tests of homogeneity, the QQ-plot of the corresponding empirical CDFs is also informative.

Examples of applications are given on Figure 10.5, which uses the R command `qqplot()`.

More generally, the QQ-plot of $F$ and $G$ lies on a straight line (not necessarily the diagonal) if and only if there exists a linear relation between $X$ and $Y$. More precisely, the QQ-plot has equation $y = ax + b$ (with $a > 0$) if and only if $Y$ has the same law as $aX + b$.

### 10.3.2 Kolmogorov–Smirnov test

Let us denote by $F_1$ and $F_2$ the respective CDFs of the samples $\mathbf{X}_{1,n_1}$ and $\mathbf{X}_{2,n_2}$. The null and alternative hypotheses for homogeneity tests are

$$H_0 = \{F_1 = F_2\}, \qquad H_1 = \{F_1 \neq F_2\}.$$

The Kolmogorov–Smirnov test for these hypotheses is a variation of the Kolmogorov test studied in the previous section. It is based on the Kolmogorov–Smirnov statistic

$$\xi_{n_1,n_2} = \sup_{x \in \mathbb{R}} \left| \widehat{F}_{1,n_1}(x) - \widehat{F}_{2,n_2}(x) \right|,$$

which can be computed with similar arguments as those detailed in Remark 10.2.4. The test is based on the following result.

**Lemma 10.3.2** (Freeness of the Kolmogorov–Smirnov statistic)**.** *Assume that $F_1$ and $F_2$ are continuous. Under $H_0$, the statistic $\xi_{n_1,n_2}$ is free: its law only depends on $n_1$ and $n_2$.*

**Exercise 10.3.3.** Prove Lemma 10.3.2.

The law of $\xi_{n_1,n_2}$ under $H_0$ is called the *Kolmogorov–Smirnov distribution with parameters $n_1$ and $n_2$*, its quantile of order $r$ is denoted by $x_{n_1,n_2,r}$.

**Corollary 10.3.4** (Kolmogorov–Smirnov test)**.** *The test rejecting $H_0$ when $\xi_{n_1,n_2} \geq x_{n_1,n_2,1-\alpha}$ is consistent and has level $\alpha$.*

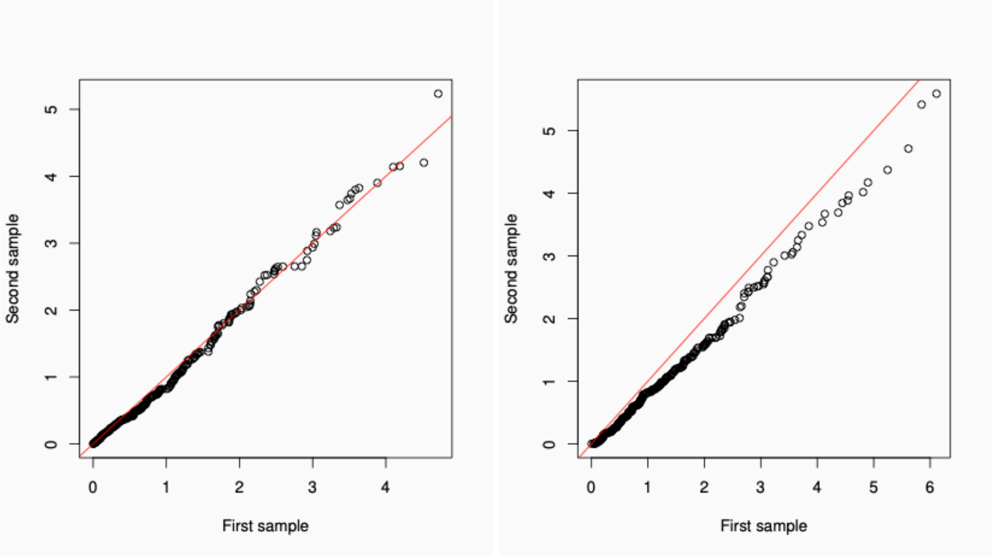

Figure 10.5: Two-sample QQ-plots. On the left-hand figure, the two samples are $\mathcal{E}(1)$-distributed; on the right-hand figure, the samples have respective distribution $\mathcal{E}(1)$ and $\Gamma(1.3, 1)$. On both figures, the diagonal $x = y$ is added in red.

The application of this test to the data of Example 10.3.1, thanks to the R command `ks.test()`, yields a $p$-value of 0.7, which allows not to reject $H_0$. Based on Donsker's Theorem, it is also possible to design an asymptotic version of the Kolmogorov–Smirnov test. Another popular nonparametric test of homogeneity is the *Mann–Whitney $U$-test* [7, Example 12.7, p. 66]. Wilcoxon's test, which addresses *matched* samples, is studied in Exercise 10.A.4.

## 10.3 The Kolmogorov–Smirnov test of homogeneity

In this section, we consider nonparametric tests of homogeneity, aiming at checking whether two independent samples $\mathbf{X}_{1,n_1} = (X_{1,1}, \ldots, X_{1,n_1})$ and $\mathbf{X}_{2,n_2} = (X_{2,1}, \ldots, X_{2,n_2})$ have the same distribution, without making any parametric assumption on this distribution.

**Example 10.3.1** (Efficiency of a vaccine)**.** *In order to study the efficiency of a vaccine, a group of* $200$ *people is split into two groups of* $n_1 = n_2 = 100$ *people. The first group is treated with the vaccine while the other group receives a placebo. One week after, the concentration of antibodies in the patients' blood is measured for both groups. The corresponding histograms are plotted on Figure 10.4. Clearly, the distribution of the concentration in antibodies is not Gaussian. Therefore Student's and Fisher's tests from Lecture 8 are not appropriate, and a nonparametric homogeneity test must be employed.*

We first present the QQ-plot, which is a heuristic tool allowing to assess the closeness of two empirical distributions. We then describe the principle of the Kolmogorov–Smirnov test, which relies on ideas which are similar to Kolmogorov one-sample tests.

### 10.3.1 QQ-plot

Consider two real-valued random variables $X$ and $Y$, with respective CDFs $F$ and $G$. The QQ-plot (for *Quantile-Quantile*) of $F$ and $G$ is the parametric curve $u \in (0,1) \mapsto (F^{-1}(u), G^{-1}(u))$, where $F^{-1}$ and $G^{-1}$ are the respective pseudo-inverses of $F$ and $G$. It provides a visual representation of 'how different' $F$ and $G$ are. In particular, it is supported by the diagonal $\{x = y\}$ if and only if $F = G$. This property can be used with empirical CDFs in the following contexts:

- for the goodness-of-fit test for a sample $X_1, \ldots, X_n$ and null hypothesis $\{F = F_0\}$, the QQ-plot of $F_0$ and $\widehat{F}_n$ allows to visually determine whether both distributions are close or not;

### 10.2.3 Goodness-of-fit to a family of distributions: the Lilliefors correction

Let $\mathcal{P}_0$ be a subset of the space of probability measures (with a continuous CDF) on $\mathbb{R}$. Similarly to the $\chi_2$ test in Section 9.3, the Kolmogorov test may generally be adapted to the set of hypotheses

$$H_0 = \{P \in \mathcal{P}_0\}, \qquad H_1 = \{P \notin \mathcal{P}_0\}.$$

We shall study the specific case (once again, similar to that of Section 9.3) where $\mathcal{P}_0$ is a *parametric family*, which thus writes

$$\mathcal{P}_0 = \{P_{0,\theta}, \theta \in \Theta\}, \qquad \text{with } \Theta \subset \mathbb{R}^q.$$

Following the lines of Section 9.3, a natural approach consists in:

(i) finding a consistent estimator $\widehat{\theta}_n$ of $\theta$ under $H_0$;

(ii) compare the distance between the empirical CDF $\widehat{F}_n$ of the sample, and the CDF $F_{0,\widehat{\theta}_n}$ of the probability measure in $\mathcal{P}_0$ corresponding to the estimated value $\widehat{\theta}_n$ of $\theta$.

In some cases, it may then be proved that the law of the statistic

$$\zeta'_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_{0,\widehat{\theta}_n}(x) \right|$$

is *free* under $H_0$, that is to say that it depends on $n$ (and on the model $\mathcal{P}_0$), but non on the underlying value $\theta$ of the parameter. In order to check this property, it is useful to mimic the proof of Proposition 10.2.6 and to express both $\widehat{F}_n(x)$ and $\widehat{\theta}_n$ in terms of independent uniform random variables $U_1, \ldots, U_n$, see Example 10.2.13 below.

Just like the estimation of $\theta$ by $\widehat{\theta}_n$ changes the number of degrees of freedom of the limiting $\chi_2$ distribution in the context of Section 9.3 (see Proposition 9.3.2), **the law of $\zeta'_n$ under $H_0$ will generally not be the Kolmogorov law from Section 10.2.2**, but a certain probability measure, depending on the model $\mathcal{P}_0$, whose quantiles $z'_{n,r}$ need to be computed, often by numerical simulation.

This procedure is called the *Lilliefors correction*. It was initially designed to test whether a sample is Gaussian or not[7]. Here, we detail the example of the Exponential model, and refer to Exercise 10.A.3 for the Gaussian model. Another popular Gaussian fit test, the Shapiro–Wilk test, is presented in Section 10.4.

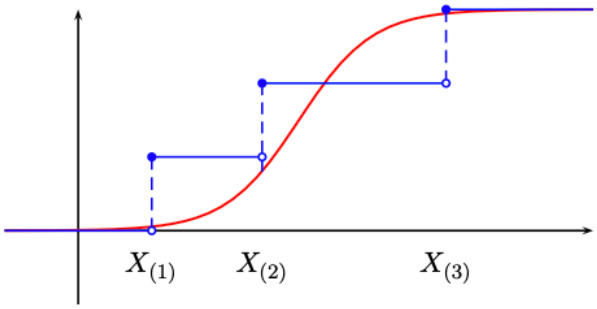

Figure 10.3: The maximum of $|\widehat{F}_n - F|$ is reached at the point $X_{(2)}$.

### 10.2.2 The nonasymptotic Kolmogorov test

Unlike the case of Pearson's statistic for finite state space models, the Kolmogorov statistic enjoys a peculiar property allowing to derive nonasymptotic tests: it is *free*[6] under $H_0$.

**Proposition 10.2.6** (Freeness of Kolmogorov's statistic)**.** *If $P_0$ has a continuous CDF $F_0$ on $\mathbb{R}$, the law of Kolmogorov's statistic $\zeta_n$ under $H_0$ only depends on $n$ and not on $F_0$.*

The proof of Proposition 10.2.6 is postponed below. It relies on the notion of *pseudo-inverse* of a CDF.

**Definition 10.2.7** (Pseudo-inverse)**.** *Let $F : \mathbb{R} \to [0, 1]$ be a CDF. The* pseudo-inverse *$F^{-1} : [0, 1] \to [-\infty, +\infty]$ is defined by*

$$\forall u \in [0, 1], \qquad F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\},$$

*where we take the conventions that* $\inf \mathbb{R} = -\infty$ *and* $\inf \varnothing = +\infty$.

**Lemma 10.2.9** (Properties of the pseudo-inverse)**.** *Let $F : \mathbb{R} \to [0, 1]$ be a CDF.*

*(i) For all $u \in (0, 1)$, $x \in \mathbb{R}$, $F^{-1}(u) \leq x$ if and only if $u \leq F(x)$.*

*(ii) Let $U$ be a uniform random variable on $[0, 1]$. Then $F$ is the CDF of the random variable $F^{-1}(U)$.*

**Definition 10.2.10** (Kolmogorov's law)**.** *Let $(U_i)_{i \geq 1}$ be a sequence of independent random variables uniformly distributed on $[0, 1]$. For all $n \geq 1$, the law of the random variable*

$$Z_n = \sup_{u \in (0,1)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{U_i \leq u\}} - u \right|$$

*is called the* Kolmogorov law *with parameter $n$.*

As a conclusion, we may thus derive a nonasymptotic test, based on the statistic $\zeta_n$. In R, this test is performed with the command `ks.test()`.

**Corollary 10.2.11** (Nonasymptotic Kolmogorov test)**.** *Under the assumptions of Proposition 10.2.6, the test with rejection region*

$$W_n = \{\zeta_n \geq z_{n,1-\alpha}\}$$

*where $z_{n,r}$ is the quantile of order $r$ of Kolmogorov's law with parameter $n$, has level $\alpha$.*

**Corollary 10.1.4** (Asymptotic of the supremum). *Under the assumptions of Theorem 10.1.3, the random variable*

$$g_n = \sqrt{n} \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right|$$

*converges in distribution to the random variable $g_\infty = \sup_{x \in \mathbb{R}} |\beta(F(x))|$. When $F$ is* continuous, *then*

$$g_\infty = \sup_{t \in [0,1]} |\beta(t)|,$$

*and its CDF writes*

$$\forall y > 0, \qquad \mathbb{P}(g_\infty \leq y) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp\left(-2k^2 y^2\right).$$

The explicit computation of the CDF of $g_\infty$ is due to Kolmogorov[4].

## 10.2 The Kolmogorov tests for goodness-of-fit

### 10.2.1 The asymptotic Kolmogorov test

We now fix a probability measure $P_0$ on $\mathbb{R}$, with CDF $F_0$, and address the test of the hypotheses

$$H_0 = \{F = F_0\}, \qquad H_1 = \{F \neq F_0\}.$$

**Definition 10.2.1** (Kolmogorov's statistic). *The Kolmogorov statistic is the statistic*

$$\zeta_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|.$$

Its asymptotic behaviour is described by the Glivenko–Cantelli and Donsker Theorems, which provide a natural asymptotic test.

**Proposition 10.2.2** (Asymptotic Kolmogorov test). *If $P_0$ has a continuous CDF $F_0$ on $\mathbb{R}$, the test with rejection region*

$$W_n = \{\sqrt{n}\zeta_n \geq a\},$$

*where $a > 0$ is defined by the relation*

$$\sum_{k=-\infty}^{+\infty} (-1)^k \exp\left(-2k^2 a^2\right) = 1 - \alpha,$$

*is consistent and has asymptotic level $\alpha$.*

**Remark 10.2.4** (Computation of the Kolmogorov statistic). *To implement Kolmogorov's test, one needs to compute the value of the statistic $\zeta_n$, which a priori requires to evaluate $F_0(x)$ at all points $x \in \mathbb{R}$. However, the monotonicity of $F_0$ together with the fact that $\widehat{F}_n$ is piecewise constant show that the supremum is necessarily reached at one of the $n$ points $X_1, \ldots, X_n$ (see Figure 10.3), so that*

$$\zeta_n = \max_{1 \leq k \leq n} \max \left\{ \left| \frac{k-1}{n} - F_0(X_{(k)}) \right|, \left| \frac{k}{n} - F_0(X_{(k)}) \right| \right\},$$

*where $X_{(1)} \leq \cdots \leq X_{(n)}$ denotes the nondecreasing reordering of $X_1, \ldots, X_n$. Therefore, only $n$ evaluations of $F_0$ are necessary.*

## 9.1 Empirical distribution in the finite setting

In this section, we consider iid random variables $X_1, \ldots, X_n$ which take their values in a finite state space $E$ with cardinality $m$. The basic idea for $\chi_2$ tests consists in approximating $P$ with the *empirical measure* of the sample, which is the probability measure $\widehat{P}_n$ on $E$ with probability mass function $(\widehat{p}_{n,x})_{x\in E}$ defined by

$$\forall x \in E, \qquad \widehat{p}_{n,x} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i = x\}}.$$

It is clear that $\widehat{P}_n$ is an unbiased and strongly consistent estimator of $P$. Therefore, if one fixes a probability measure $P_0$ on $E$ and wants to construct a goodness-of-fit test for the hypotheses

$$H_0 = \{P = P_0\}, \qquad H_1 = \{P \neq P_0\},$$

a natural idea consists in choosing a distance $d$ on the space of probability measures on $E$ (seen as a subset of $\mathbb{R}^m$) and taking a rejection region of the form

$$W_n = \{d(\widehat{P}_n, P_0) \geq a_n\}$$

for some threshold $a_n$ to be chosen appropriately. In this section, we introduce and study a specific distance-like function $d$, the $\chi_2$ *distance*, which has the advantage to make the computation of the threshold $a_n$ *independent from* $P_0$, a property which is recurrent in the construction of nonparametric tests.

$$\forall x \in \mathbb{R}, \qquad \widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}.$$

$$\lim_{n\to+\infty} \widehat{F}_n(x) = \mathbb{E}[\mathbb{1}_{\{X_1 \leq x\}}] = F(x), \qquad \text{almost surely.}$$

**Theorem 10.1.1** (Glivenko–Cantelli Theorem). *Let $F$ be a CDF on $\mathbb{R}$ and $(X_i)_{i\geq 1}$ be a family of independent random variables with CDF $F$. We have*

$$\lim_{n\to+\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| = 0, \qquad \textit{almost surely.}$$

**Definition 10.1.2** (Brownian bridge). *The* Brownian bridge *is the unique (in law) random process $(\beta(t))_{t\in[0,1]}$ such that:*

**Theorem 10.1.3** (Donsker Theorem). *Let $F$ be a CDF on $\mathbb{R}$ and $(X_i)_{i\geq 1}$ be a family of independent random variables with CDF $F$. The random function*

$$G_n : x \in \mathbb{R} \mapsto \sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)$$

*converges in distribution (in an appropriate functional space[3]) to the random function*

$$G : x \in \mathbb{R} \mapsto \beta(F(x)),$$

*where $\beta : [0,1] \to \mathbb{R}$ is the Brownian bridge.*