Guide pour une leçon de modélisation stochastique. Fonctions de répartition empiriques. Tests de Kolmogorov-Smirnov. Estimation des quantiles

Sana Louhichi

De l'observation à la théorie..
De la pomme tombée à la loi de la gravitation universelle..

Guide pour une leçon de modélisation stochastique.

Fonctions de répartition empiriques. Tests de Kolmogorov-Smirnov. Estimation des quantiles.

Sana Louhichi,
E.mail : sana.louhichi@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK Grenoble, France.

Table des matières

Avant-propos **v**

Organigrammes **vi**
0.1 Modélisation . . . vi
0.2 Modélisation Statistique . . . vi
0.3 Modélisation Probabiliste . . . vii
0.4 Modélisation Probabiliste VS Modélisation Statistique . . . viii

1 Mémento **1**
1.1 Fonctions de répartition empiriques . . . 1
1.1.1 Mots clés . . . 1
1.1.2 Synthèse . . . 1
1.2 Quantiles et Quantiles empiriques . . . 5
1.2.1 Mots clés . . . 5
1.2.2 Synthèse . . . 5
1.3 Tests de Kolmogorov-Smirnov . . . 7
1.3.1 Mots clés . . . 7
1.3.2 Tests d'ajustement à une loi donnée . . . 7
1.3.3 Tests d'ajustement à une famille de lois donnée . . . 10
1.3.4 Tests de comparaison de deux échantillons . . . 10

2 Problématiques et Modélisation statistique **13**
2.1 Objectifs . . . 13
2.2 Problématique I. . . . 13
2.3 Problématique II. . . . 16
2.4 Problématique III. . . . 16
2.5 Problématique VI. . . . 19

3 Exercices (plutôt théoriques) **21**

4 Exercices (plutôt appliqués) **23**

BIBLIOGRAPHIE **25**

Index **25**

Table des figures **27**

Notations **29**

Avant-propos

Ce manuscrit est sous la forme d'un guide plutôt que d'un cours détaillé classique. Il s'adresse à toute personne ayant une formation en mathématiques et souhaitant approfondir ses connaissances en probabilités et statistique en rapport avec la modélisation.

Le chapitre zéro présente trois organigrammes schématisant des différentes notions de modélisation et mettant l'accent sur les « allers-retours » entre l'observation et la théorie associée.

Le premier chapitre est un mémento qui rappelle les résultats théoriques utiles, et parfois indispensables, à maîtriser pour la leçon de modélisation en question. Les résultats sont énoncés sous la forme d'un résumé et donc sans démonstration mais souvent accompagnés par des illustrations graphiques. L'auteur intéressé trouvera sans doute les démonstrations des résultats de son intérêt. Des mots clés sont détaillés pour chaque section du mémento, orientant ainsi vers le cœur du sujet. Des questions et des exercices sont posés au fur et à mesure de la progression du texte. Le mémento sera utile pour la résolution des problèmes de modélisation du second chapitre.

Le second chapitre concerne des problèmes qui nécessitent de la modélisation statistique pour leurs résolutions. On a, à la disposition, des données observées et une question en rapport avec la population dont ces données sont tirées. Cette population ne peut pas être observée en entier. Il s'agit donc de résoudre (même partiellement) un problème lié à une population entière en se basant sur des observations partielles de cette population. On est amené à construire un modèle statistique, i.e., une formule mathématique aléatoire, par exemple définir des variables aléatoires dont des réalisations possibles sont les données observées et qui sont à la disposition. On est souvent amené aussi à ajouter des hypothèses qu'on appellera des hypothèses de modélisation. Ces hypothèses seront parfois indispensables pour la résolution théorique (mathématique) du problème posé. Il faudrait juste s'assurer qu'elles sont cohérentes avec l'observation. Une fois que le modèle statistique et les hypothèses de modélisations sont posés, la tâche serait de développer la théorie utile pour sa résolution, ramenant ainsi le problème de la vie réelle à une question théorique. Dès que l'on dispose des réponses ou des éléments de réponses à cette question théorique, on revient au point de départ (i.e. aux données et à la problématique concrète du départ) et on rediscute la solution théorique apportée.

Les simulations sont un outil puissant pour la modélisation statistique. Elles permettent de générer des données simulées, de réaliser des analyses statistiques sur ces données, et d'obtenir des informations sur la performance du modèle statistique. Cela permet ainsi de mieux comprendre le modèle, de visualiser les résultats, de valider ou d'infirmer ou de quantifier les idées intuitives.

Ce guide se termine par des exercices théoriques permettant de mieux comprendre le cours et aussi par des exercices dont les résolutions nécessitent un logiciel.

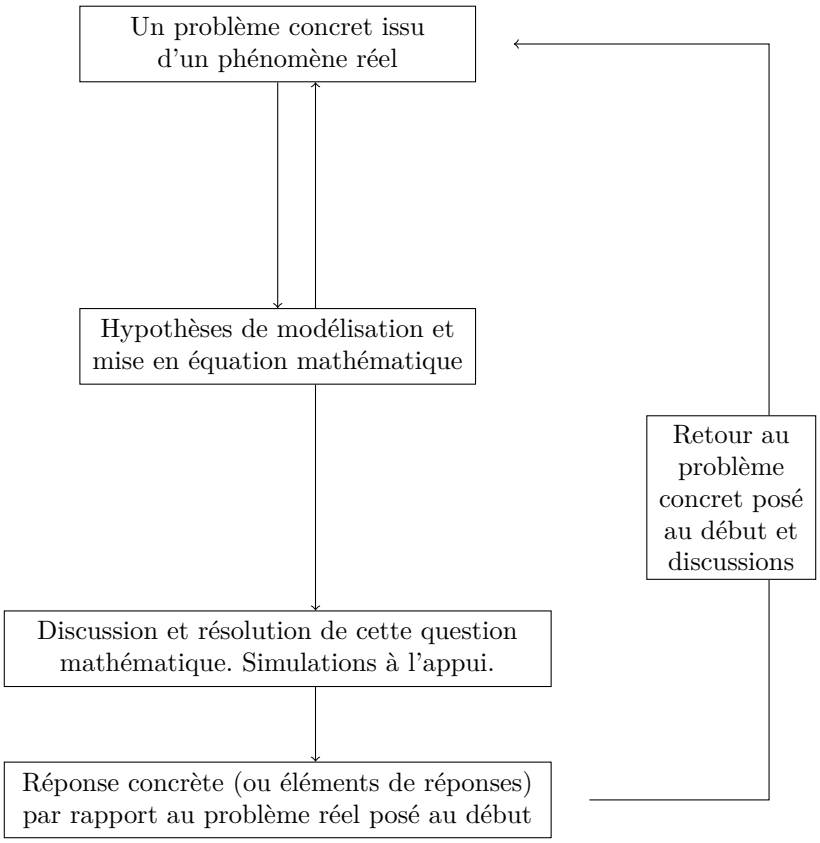
Les références, à ce sujet, sont nombreuses. Une liste non exhaustive est donnée à la fin de ce manuscrit.

L'autrice.

Organigrammes. Le but de ces trois organigrammes ci-dessous est d'expliquer schématiquement en quoi consiste la modélisation déterministe ou stochastique (statistique ou probabiliste). Dans tous ces cas de modélisations, on démarre d'un problème concret qu'on reformule mathématiquement. Des « allers-retours » du problème concret vers la théorie et vice versa sont nécessaires afin de vérifier, tout particulièrement, la validité des hypothèses de modélisation posées ainsi que le développement de la théorie. La littérature la-dessus est abondante, voir par exemple : ici ou ici.

0.1 Modélisation

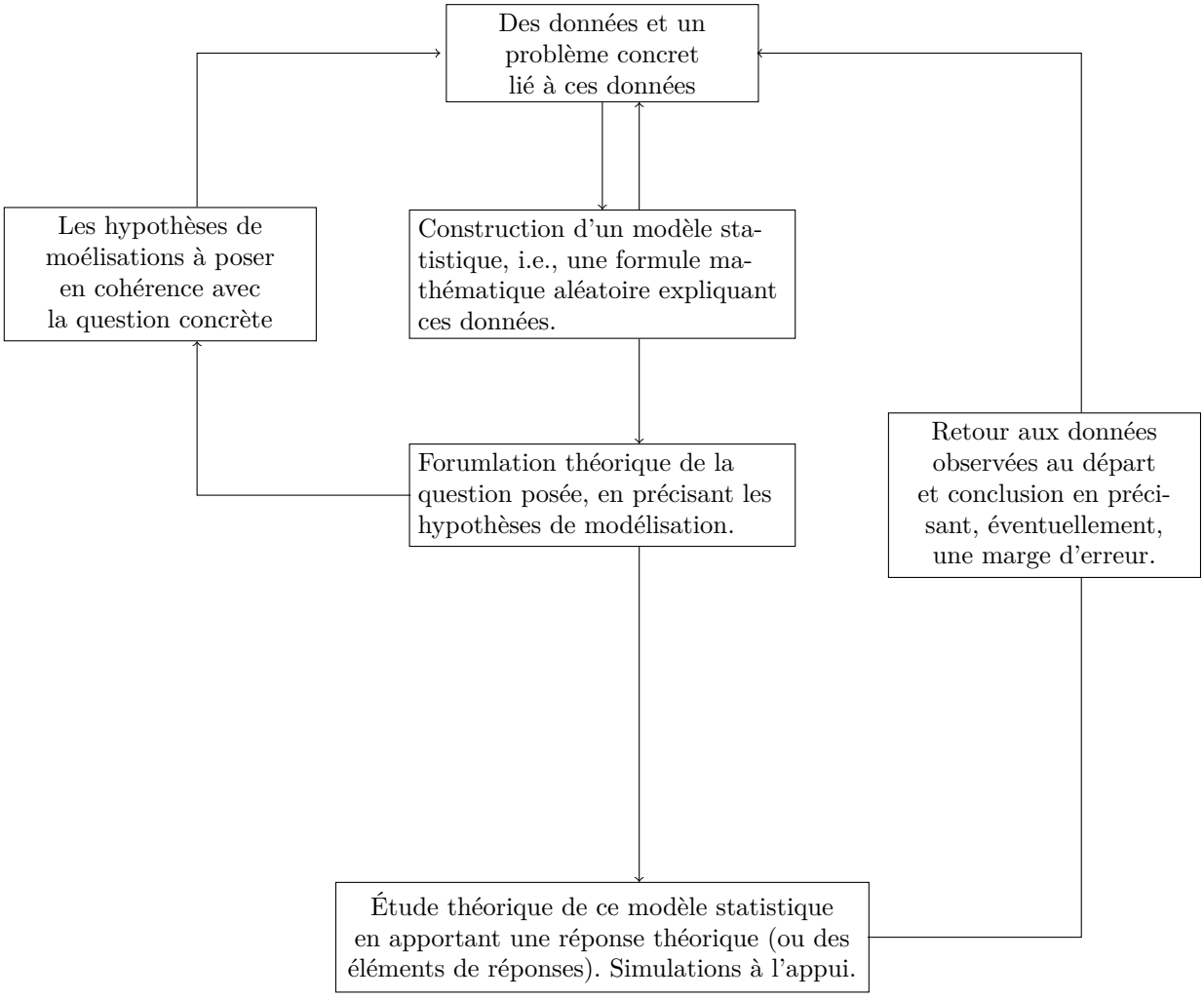
La modélisation est une représentation simplifiée d'un phénomène réel afin de mieux le comprendre et de l'analyser. Ce qui permet de répondre à une question concrète liée à ce phénomène. Cette modélisation peut-être déterministe ou stochastique (« stochastique » pour dire, en rapport avec la description des phénomènes aléatoires et incertains).



0.2 Modélisation Statistique

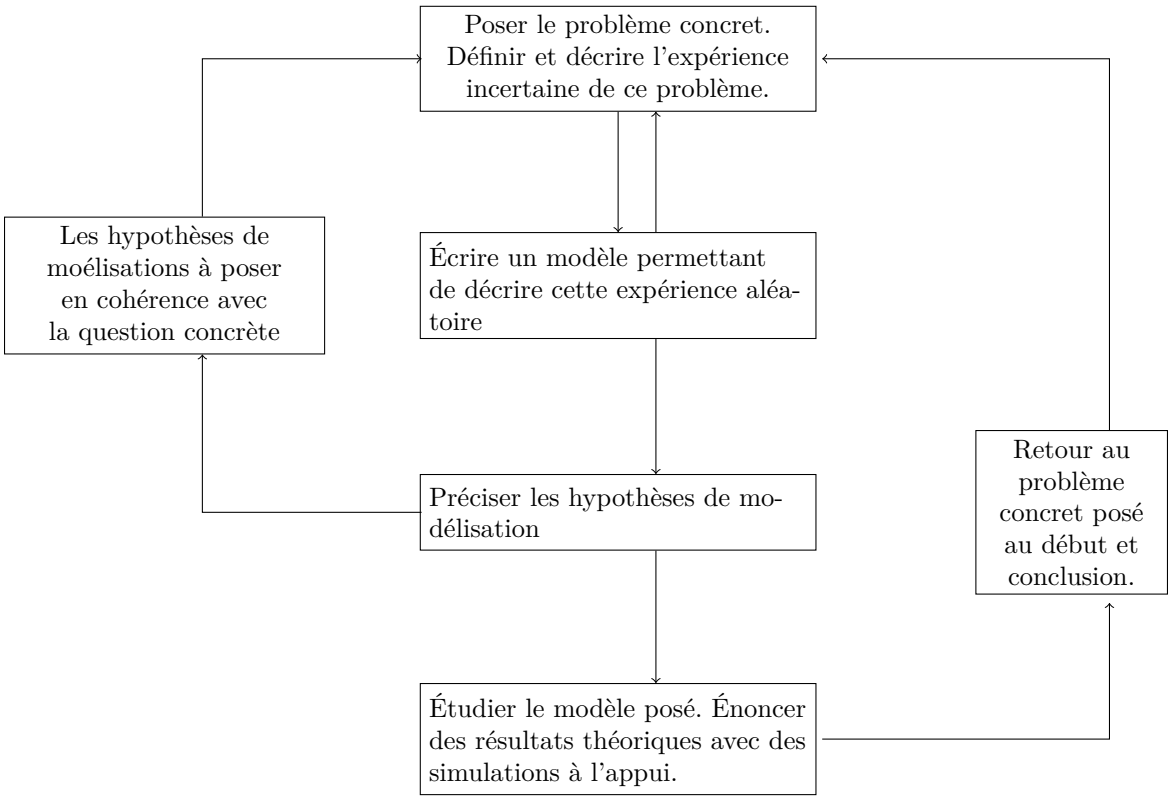
La modélisation statistique est un problème de modélisation stochastique qui utilise des techniques de la statistique mathématique afin d'expliquer un ensemble des données observées. Ces techniques permettent de développer un modèle statistique qui permettra d'expliquer et de prédire. Elles permettent aussi de construire des estimateurs, d'étudier leurs propriétés, de construire des intervalles

de confiance, de faire des tests statistiques, de prendre une décision (avec une marge d'erreur) en se basant sur ces données observées



0.3 Modélisation Probabiliste

La modélisation probabiliste est aussi une modélisation stochastique. Elle utilise des concepts de probabilité afin de représenter des phénomènes aléatoires et incertains. Un modèle probabiliste peut démarrer d'une expérience aléatoire. Son étude peut utiliser des variables aléatoires, des distributions de probabilités, des processus stochastiques, des chaînes de Markov...



0.4 Modélisation Probabiliste VS Modélisation Statistique

La modélisation probabiliste fournit un cadre théorique général pour traiter l'incertain, alors que la modélisation statistique, pouvant être vue comme un cas spécifique de la modélisation probabiliste, est basée sur l'analyse de données observées (expérimentaux). Elle utilise des méthodes probabilistes pour comprendre la population dont est tirée l'échantillon des données observées, tandis que la modélisation probabiliste peut également être utilisée pour décrire des expériences aléatoires même en l'absence de données observées.

Chapitre 1

Mémento

1.1 Fonctions de répartition empiriques

1.1.1 Mots clés

Distribution empirique, définition de la fonction de répartition empirique, ses propriétés, son graphe, statistique d'ordre, ex aequo, convergence presque sûre et en loi de la fonction de répartition empirique, Théorème de Glivenko-Cantelli, loi de Kolmogorov-Smirnov (et le théorème associé : Théorème 1.2 ci-dessous).

1.1.2 Synthèse

Soient $X_1, \cdots, X_n$ n variables aléatoires réelles (définies sur un même espace probabilisé, condition que l'on supposera le long de ce guide) indépendantes et identiquement distribuées (i.i.d.) de fonction de répartition commune F, i.e., $F(x) = \mathbb{P}(X_1 \leq x)$. On dira que $(X_1, \cdots, X_n)$ est un échantillon aléatoire ou tout simplement échantillon.

Définition 1.1. *On appelle fonction de répartition empirique associée à l'échantillon* $(X_1, \cdots, X_n)$*, la fonction aléatoire* F_n *définie pour* $x \in \mathbb{R}$ *par*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}.$$

— Le mot « empirique » fait appel à l'expérience, aux observations et non à la théorie.
— La fonction de répartition F est une quantité « théorique » son calcul est lié à la loi d'une variable aléatoire et non aux observations.

Démontrer la proposition suivante.

Proposition 1.1. *On a, pour* x *fixé,*
(a) $\mathbb{E}(F_n(x)) = F(x)$, $\mathrm{Var}(F_n(x)) = \frac{F(x)(1-F(x))}{n}$.
(b) $nF_n(x)$ *suit la loi Binomiale de paramètres* n *et* $p = F(x)$.
(c) $(F_n(x))_n$ *converge presque sûrement (ps) vers* $F(x)$. *[Voir Figure 1.1.]*

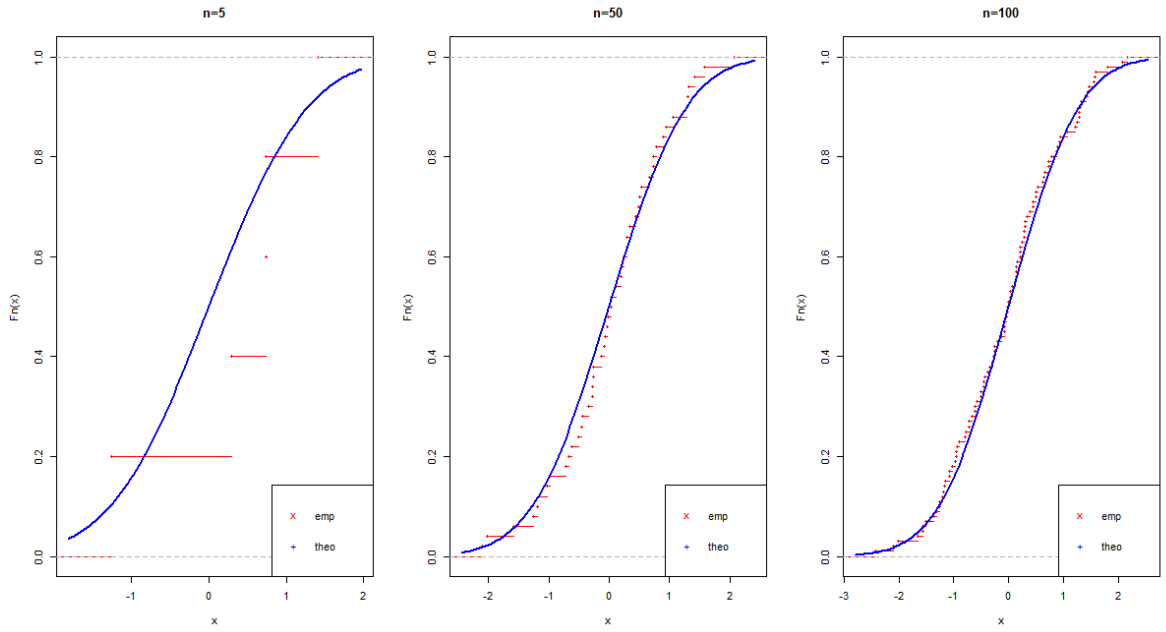


FIGURE 1.1 – Illustrations de la convergence ps de la fonction de répartition empirique F_n (em) vers la fonction de répartition F (theo).

(d) $\sqrt{n}(F_n(x) - F(x))$ *converge en loi, lorsque* n *tend vers l'infini, vers la loi normale centrée et de variance* $F(x)(1 - F(x))$. *En d'autres termes, la loi de* $F_n(x)$, *peut être approchée, lorsque* n *est suffisamment grand, par la loi normale*

$$\mathcal{N}\left(F(x), \frac{F(x)(1-F(x))}{n}\right),$$

[Voir Figure 1.2.]

Pour la représentation du graphe de la fonction aléatoire F_n, on a besoin de rappeler la notion de statistique d'ordre.

Définition 1.2. *La statistique d'ordre associée à l'échantillon* $(X_1, \cdots, X_n)$ *est le vecteur aléatoire* $(X_{(1)}, \cdots, X_{(n)})$ *vérifiant*
- $\{X_{(1)}, \cdots, X_{(n)}\} = \{X_1, \cdots, X_n\}$ *ps,*
- $X_{(1)} \leq \cdots \leq X_{(n)}$ *ps.*

Proposition 1.2. *La fonction* $x \mapsto F_n(x)$ *est, ps,*
- *croissante de 0 à 1, constante par palier*
- *continue à droite ayant une limite à gauche (càdlàg).*

La démonstration de la Proposition 1.2 est accessible avec des connaissances basiques en probabilités.

Exercice 1.1.
On suppose que F est continue. Montrer, en utilisant l'exercice 3.1, que ps

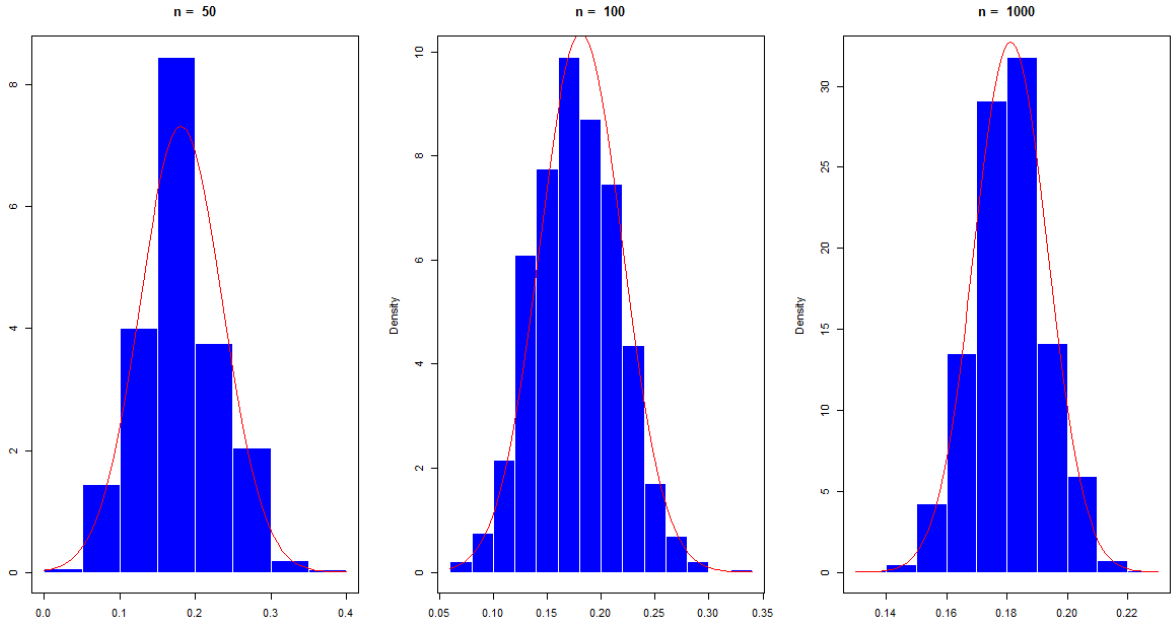


FIGURE 1.2 – Illustration de la normalité asymptotique de $F_n(x)$ pour x fixé (ici $x = 0.4$).

1. F_n est constante sur les intervalles suivants : $]-\infty, X_{(1)}[$, $[X_{(1)}, X_{(2)}[, \cdots, [X_{(n-1)}, X_{(n)}[$, $[X_{(n)}, +\infty[$
2. Que vaut $F_n(X_{(j)}) - F_n(X_{(j-1)})$, pour $1 \leq j \leq n$?.

Par la suite, on énoncera des résultats de convergence de nature non-paramétrique. Le théorème suivant établit la convergence uniforme ps de F_n vers F, $\|\cdot\|_\infty$ désigne la norme infinie, i.e., $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$, pour une fonction f donnée.

Théorème 1.1. *Théorème de Glivenko-Cantelli.*
Soient $X_1, \cdots, X_n$ n variables aléatoires réelles i.i.d. de fonction de répartition commune F. Alors, ps,

$$\lim_{n\to\infty} \|F_n - F\|_\infty = 0.$$

[Voir Figure 1.3]

La démonstration du théorème de Glivenko-Cantelli est accessible avec des connaissances basiques en probabilité.

Définition 1.3. *La quantité $\|F_n - F\|_\infty$ est appelée la distance de Kolmogorov-Smirnov.*

Question 1.1. *Soit G une autre fonction de répartition différente de F. Que peut-on dire, asymptotiquement et ps, de $\|F_n - G\|_\infty$? Justifier votre réponse.*

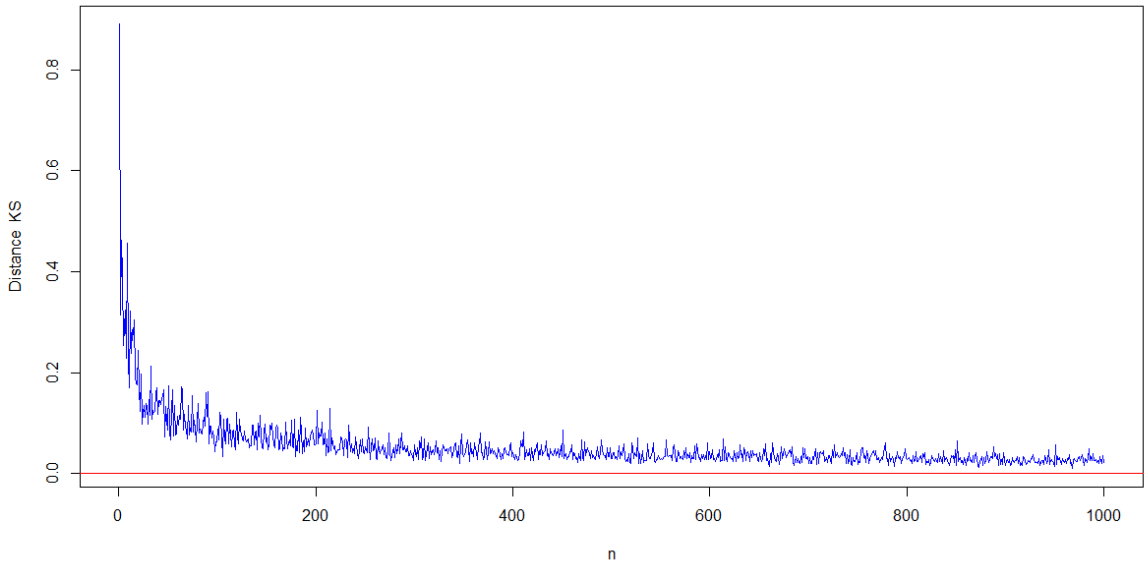


FIGURE 1.3 – Illustration de la convergence p.s. et uniforme de $\|F_n - F\|_\infty$ vers 0.

On admet l'inégalité suivante.

Proposition 1.3. *Inégalité DKW (Dvoretzky-Kiefer-Wolfowitz) On a, pour tout $\epsilon > 0$,*

$$\mathbb{P}(\|F_n - F\|_\infty \geq \epsilon) \leq 2e^{-n\epsilon^2}$$

Question 1.2. *Utiliser l'inégalité DKW pour construire une bande de confiance asymptotique pour F. Illustrer la graphiquement sur des données simulées.*

Montrer la proposition suivante.

Proposition 1.4. *Si F est continue alors*

$$\|F_n - F\|_\infty = \max_{1\leq j\leq n} \max\left(\frac{j}{n} - F(X_{(j)}), F(X_{(j)}) - \frac{j-1}{n}\right)$$

Question 1.3. *En quoi la Proposition 1.4 peut-elle être utile ?*

On admet le théorème suivant

Théorème 1.2. *Théorème de Kolmogorov-Smirnov. Soient $X_1, \cdots, X_n$ n. v.a. i.i.d. de fonction de répartition F* **continue**. *Alors*

$$T_n := \sqrt{n} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

converge en loi vers la loi de Kolmogorov-Smirnov, dont la fonction de répartition est donnée, pour tout $x > 0$, par

$$F_{KS}(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2x^2k^2}. \tag{1.1}$$

Voir Figures 1.4 et 1.5.

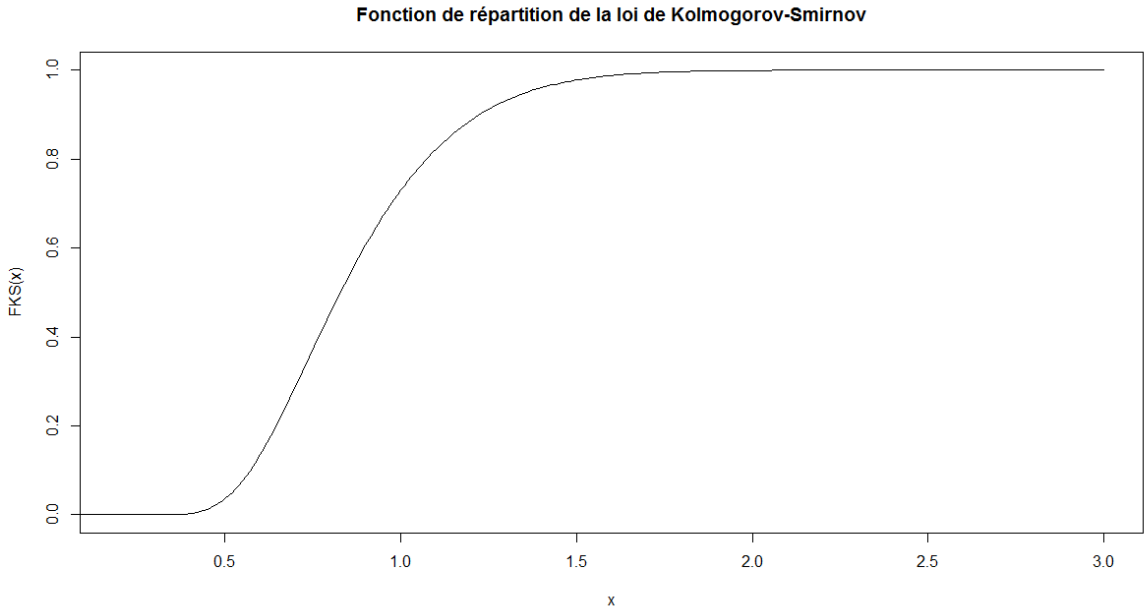


FIGURE 1.4 – Fonction de répartition de la loi de Kolmogorov-Smirnov.

La loi asymptotique de $\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|$ est donc une loi libre de la loi des observations $X_1,\cdots,X_n$ (on dira par la suite qu'une loi est libre si cette loi ne dépend pas des observations ni de leur loi commune) . Ses quantiles sont calculables en développant un programme informatique.

1.2 Quantiles et Quantiles empiriques

1.2.1 Mots clés

Quantile d'une variable aléatoire, fonction quantile, quantiles empiriques, Médiane, Médiane empirique, diagramme QQ-plot, statistique d'ordre.

1.2.2 Synthèse

Définition 1.4. *Soit X une variable aléatoire de fonction de répartition F. La fonction quantile Q de X est la fonction $Q:]0,1[\to\mathbb{R}$ définie pour $u\in]0,1[$ par,*

$$Q(u)=\inf\{x\in\mathbb{R},\ F(x)\geq u\}. \tag{2.2}$$

La fonction réciproque généralisée d'une fonction càdlàg et croissante g, notée $g^{(-1)}$, est définie par,

$$g^{(-1)}(u)=\inf\{x\in\mathbb{R},\ g(x)\geq u\}$$

La fonction quantile d'une v.a. X est donc la fonction réciproque généralisée de sa fonction de répartition F.

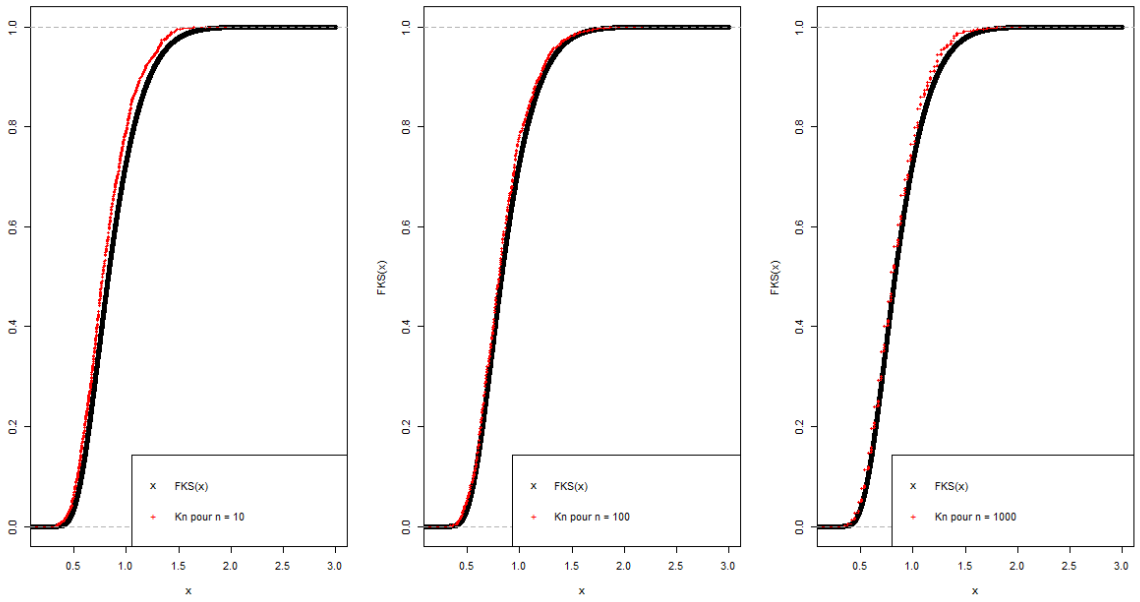


FIGURE 1.5 – Illustration de la convergence en loi de $T_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)|$ vers la loi de Kolmogorov-Smirnov.

Exercice 1.2.

1. On suppose que F est continue et strictement croissante. A quoi correspond, dans ce cas, la fonction Q ?
2. Calculer la fonction quantile d'une variable aléatoire X dans les deux cas suivants :
   — X est une v.a. de loi de Bernoulli.
   — X est une v.a. de loi exponentielle.

Soient $X_1, \cdots, X_n$ une suite de v.a. i.i.d. de fonction de répartition F. La question est de construire un estimateur de la fonction quantile Q (définie par (2.2)). La méthode de substitution (the plug-in principle, en anglais) amène à la définition suivante :

Définition 1.5. *La fonction quantile empirique de $X_1 \cdots, X_n$ est définie par,*

$$Q_n(u) = \inf\{x \in \mathbb{R},\ F_n(x) \geq u\}. \tag{2.3}$$

Exercice 1.3.
Donner une expression de $Q_n(u)$ à l'aide de la statistique d'ordre associée à $X_1, \cdots, X_n$ (on pourra se faire aider par la représentation graphique de F_n).

Les deux résultats de convergence suivants sont utiles.

Proposition 1.5. *Soit $u \in]0,1[$ fixé. On suppose que F est strictement croissante au voisinage de $Q(u)$. Alors $Q_n(u)$ converge ps, lorsque n tend vers l'infini, vers $Q(u)$.*

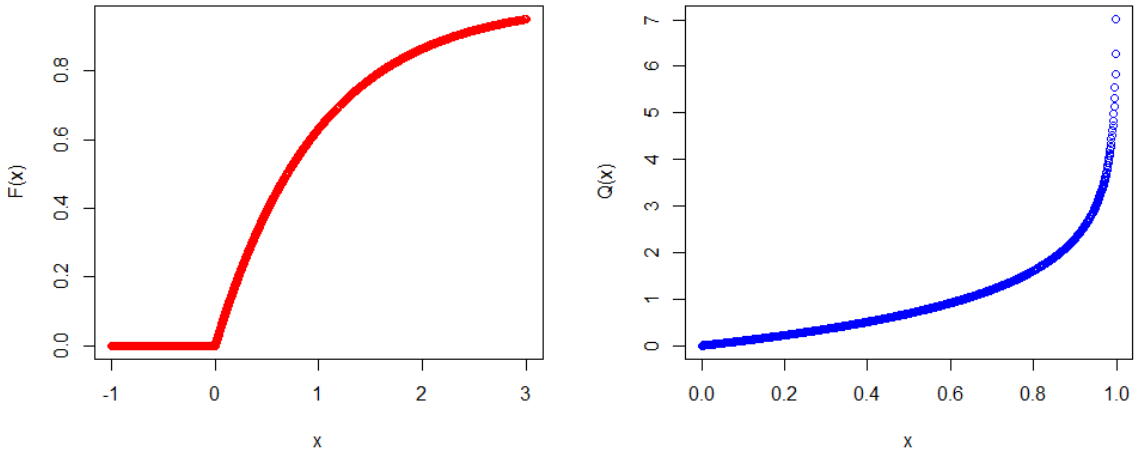


FIGURE 1.6 – Fonction de répartition et fonction quantile d'une loi exponentielle.

Proposition 1.6. *On suppose que F admet une dérivée continue f strictement positive. On a, dans ce cas, pour tout $u \in]0,1[$*

$$\sqrt{n}(Q_n(u) - Q(u))$$

converge en loi, lorsque n tend vers l'infini, vers la loi normale centrée et de variance

$$\frac{u(1-u)}{f^2(Q(u))}.$$

[Voir Figure 1.8]

Exercice 1.4.
Soit $X_1, \cdots, X_n$ une suite de v.a. i.i.d. de fonction de répartition F et de densité f continue et strictement positive. Construire un intervalle de confiance asymptotique pour la médiane de la loi de X_1 (quitte à ajouter une condition, laquelle ?)

1.3 Tests de Kolmogorov-Smirnov

1.3.1 Mots clés

Théorème de Glivenko-Cantelli, loi de Kolmogorov-Smirnov, p-valeur, test d'ajustement à une loi donnée, test d'ajustement à une famille de lois, test d'homogénéité, région de rejet, loi continue, statistique du test, loi libre sous H_0, hypothèses statistiques (hypothèse nulle et hypothèse alternative), test asymptotique.

1.3.2 Tests d'ajustement à une loi donnée

Soient $X_1, \cdots, X_n$ une suite de v.a. i.i.d de fonction de répartition F **continue** et inconnue. Soit F_0 une autre fonction de répartition continue et connue. La question est de voir si on peut « accepter »

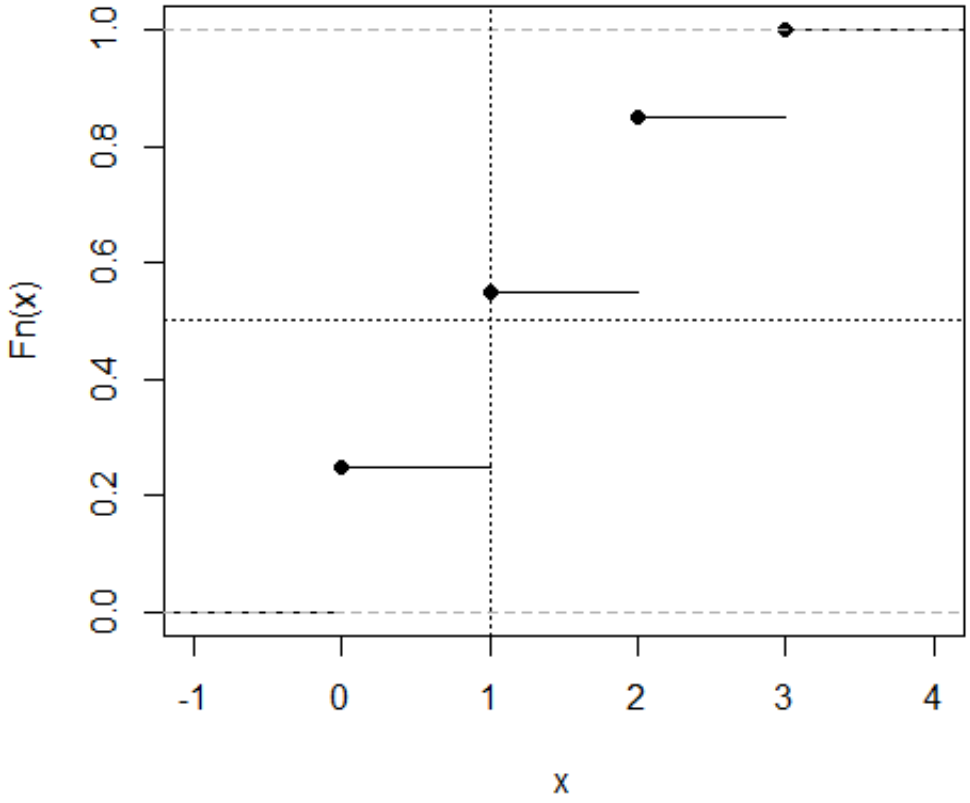


FIGURE 1.7 – Ici $Q_n(0.5) = 1$.

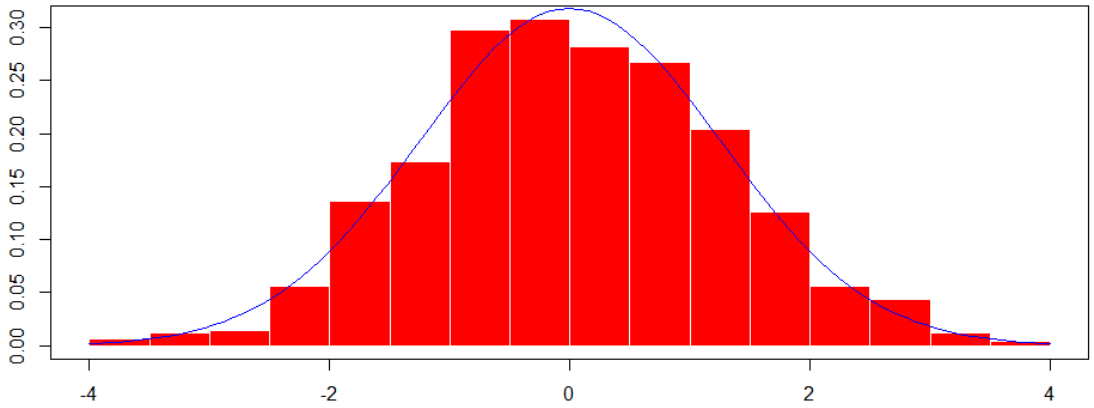


FIGURE 1.8 – Histogramme de la distribution de $\sqrt{n}(Q_n(0.5)-Q(0.5))$ et courbe de la densité d'une loi normale centrée et de variance $\frac{0.25}{f^2(Q(0.5))}$ (cas d'un échantillon simulé de loi normale).

(avec une marge d'erreur contrôlée) la décision que X_1 a pour fonction de répartition F_0. La question est donc d'étudier le test dont l'hypothèse nulle est :

$$H_0 : F = F_0$$

contre l'alternative,

$$H_1 : F \neq F_0.$$

Comme dans tout problème de test, la première étape est de préciser la statistique de test ainsi que sa loi (ou sa loi asymptotique) sous l'hypothèse de base H_0.
Le théorème 1.2 est très utile ici : sous H_0, la statistique du test

$$T_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \tag{3.4}$$

converge en loi lorsque n tend vers l'infini vers la loi de Kolmogorov Smirnov dont la fonction de répartition est donnée par (1.1) et dont les quantiles sont calculables. Soit $\alpha \in]0,1[$ l'erreur de première espèce qu'on suppose fixée. On rejettera l'hypothèse H_0 en faveur de H_1 si la valeur observée de T_n, soit t_n, vérifie

$$t_n \geq q_{ks}(1-\alpha), \tag{3.5}$$

$q_{ks}(1-\alpha)$ étant le quantile d'ordre $1-\alpha$ de la loi de Kolmogorov-Smirnov.

On peut aussi calculer la p-valeur ce test qui est

$$1 - F_{ks}(t_n), \tag{3.6}$$

F_{ks} étant la fonction de répartition de la loi de Kolmogorov-Smirnov définie dans (1.1).

Exercice 1.5.

1. Comment se comporte asymptotiquement T_n (définit par (3.4)) sous l'hypothèse H_1 ?
2. Montrer (3.5) c'est-à-dire que la zone de rejet de l'hypothèse H_0 est l'ensemble des observations $(x_1, \cdots, x_n)$ de $(X_1, \cdots, X_n)$ pour lesquelles la valeur observée, t_n de T_n vérifie (3.5).
3. Monter l'expression de la p-valeur donnée dans (3.6).
4. Quelles sont les instructions du logiciel que vous utilisez qui permettent de calculer $F_{ks}(t_n)$, $q_{ks}(1-\alpha)$.
5. Écrire un programme permettant de calculer t_n.

1.3.3 Tests d'ajustement à une famille de lois donnée

Soient $X_1, \cdots, X_n$ une suite de v.a. i.i.d de fonction de répartition F continue et inconnue. Soit F_λ une autre fonction de répartition continue et partiellement connue. On connaît, cependant, la famille des lois :

$$\mathcal{F}_\Lambda = \{F_\lambda, \lambda \in \Lambda\}$$

Par exemple on voudrait savoir si les variables aléatoires $X_1, \cdots, X_n$ sont distribuées selon la loi exponentielle (ou selon une loi normale) : on connaît la famille des lois qui est paramétrique mais on ne connaît pas les paramètres de ces lois. La question est donc, d'étudier le test dont l'hypothèse nulle est

$$H_0 : F \in \mathcal{F}_\Lambda$$

contre l'alternative,

$$H_1 : F \notin \mathcal{F}_\Lambda.$$

Les étapes pour résoudre ce problème de test sont les suivantes :

— Estimer λ par la méthode de maximum de vraisemblance. Soit λ_n cet estimateur.

— Étudier la statistique

$$\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_{\lambda_n}(x)|,$$

étudier sa convergence en loi, donner sa loi limite et voir si elle est libre sous H_0. Les quantiles de cette loi limite sont-ils calculables ? Si c'est le cas, la résolution se fait comme dans le cas du test d'ajustement à une loi donnée.

1.3.4 Tests de comparaison de deux échantillons

Soient $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_m)$ deux échantillons indépendants de fonctions de répartitions respectives F et G continues toutes les deux. On souhaite savoir si on peut considérer que les deux échantillons ont la même loi ? On peut donc tester

$$H_0 : F = G$$

contre

$$H_1 : F \neq G$$

Une statistique naturelle, pour ce test, est construite à partir des fonctions de répartition empiriques de ces deux échantillons : il s'agit de

$$\|F_n - G_m\|_\infty := \sup_{x \in \mathbb{R}} \|F_n(x) - G_m(x)\|,$$

avec

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}, \;\; G_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}_{Y_i \leq x}$$

On peut utiliser l'exercice suivant afin de résoudre ce problème de test de comparaison.

Exercice 1.6.

1. Montrer que $\|F_n - G_m\|_\infty$ est de loi libre sous H_0 (on peut se ramener à $n+m$ v.a. i.i.d. de loi uniforme sur $[0,1]$)
2. Résoudre ce problème de test de comparaison.

Chapitre 2

Problématiques et Modélisation statistique

2.1 Objectifs

La modélisation statistique consiste à étudier « théoriquement » des données observées qu'on supposera qu'elles sont des réalisations d'une ou des grandeurs aléatoires. On élabore un ensemble d'outils mathématiques afin de décrire, d'expliquer et de prédire une valeur non observée, tout en se basant sur ces données observées.

2.2 Problématique I.

Une entreprise vous a fourni des durées de vie, ci-dessous, de 100 composants issus de sa fabrication [1], l'unité étant une unité de temps.

```
0.68354198 0.01747703 2.28914465 4.12769753 3.20651889 0.55603291
0.72594578 0.65967872 0.85472120 1.31901392 0.70319142 2.10390758
0.73311302 1.45554978 0.16978426 3.07302725 0.41826852 1.02086487
0.46214164 0.06673642 3.14362218 2.24085115 0.81826526 4.73186359
1.39878137 6.83703802 0.95410121 8.37648244 6.15104049 0.52912209
1.01936533 6.80575392 0.51865541 2.24316593 1.23256563 0.68665697
7.73302720 0.75815587 0.74473017 0.40689020 1.13674565 1.48749318
3.16062872 0.63598313 0.68044854 4.13000516 0.74040332 2.24300981
1.43958422 1.91786137 0.97777445 1.49839379 0.01635843 1.34176269
1.98618814 5.82135190 0.06340729 1.58049354 0.71499477 6.59145007
0.42259102 1.63187575 0.61942846 3.39832393 1.48000414 5.20007941
2.54197566 3.12972540 5.18653246 4.41163001 3.78628957 0.54192978
7.22248123 6.39603209 4.94948807 1.26011930 1.10834842 7.45714511
0.09068439 2.29438904 0.50793927 0.25344299 1.44342465 0.37647426
2.55817818 0.24257523 1.17689111 0.41953704 6.88343358 0.76528913
3.19515322 3.98177254 5.47047076 9.55920181 0.61015641 2.85344153
6.95408382 0.57734417 0.78708007 0.63173289
```

1. Il s'agit des données simulées

L'entreprise affirme que la durée de vie moyenne de sa fabrication est de 2 unité de temps. L'entreprise vous confie ces données avec pour objectif de les analyser afin d'avoir le maximum d'informations sur ses fabrications futures (qui seront toutes dans des mêmes conditions de fabrications que celles actuelles). Par exemple, être capable

(Q1) d'avoir une idée sur la proportion de composants encore fonctionnels à n'importe quel instant,

(Q2) ou d'avoir une idée sur l'instant au-delà duquel au moins 50% des composants sont encore fonctionnels.

(AI) Étude descriptive

Les données ci-dessus constituent un échantillon de données, à ne pas confondre avec un échantillon aléatoire, qu'on notera $x_1, \cdots, x_{100}$, $x_1 = 0.68354198, \cdots, x_{100} = 0.63173289$. Toute manipulation de ces 100 observations demeure une étude empirique (i.e. expérimentale, s'appuyant sur les observations). Vu les questions que l'entreprise se pose, on commence par traçer la courbe de la fonction de répartirion empirique associée à ces 100 observations $x_1, \cdots, x_{100}$, il s'agit du graphe de la fonction :

$$f_{100} : x \mapsto \frac{1}{100} \sum_{i=1}^{100} \mathbb{I}_{x_i \leq x}$$

Exercice 2.1.

1. Tracer (en utilisant le langage informatique de votre choix) la courbe de f_{100}. [Pour indication, voir Figure 2.1].
2. Comment en déduire l'instant au déla duquel la moitié des composants sont fonctionnels sur cet échantillon de données ?

Ca va de soi que cette étude est empirique et est basée sur ces 100 observations. Mais l'objectif de l'entreprise est de pouvoir déduire des conclusions, avec beaucoup de chances, sur toute sa production (qui est assurée dans les mêmes conditions que ces 100 observations) et non pas que sur ces 100 observations. D'où l'importance de l'étape suivante.

(BI) Modélisation et Inférence statistique

Hypothèses de modélisation

On suppose que les 100 observations ci-dessus sont des réalisations de 100 variables aléatoires $X_1, \cdots, X_{100}$, que nous supposerons i.i.d. Il s'agit de notre première hypothèse de modélisation. On rappelle qu'une hypothèse de modélisation est une supposition formulée afin d'expliquer un phénomène observé et de construire le modèle mathématique associé. Plus les hypothèses de modélisation correspondent à la réalité (en particulier en les confrontant aux données observées associées), plus le modèle est susceptible d'être précis et les conclusions tirées seront plus pertinentes. La validité des hypothèses de modélisation est donc cruciale pour les performances d'un modèle. Lorsque les hypothèses de modélisation ne s'adaptent pas bien aux données réelles, cela peut entraîner des erreurs dans les prédictions ou dans les conclusions. Par conséquent, il est essentiel de valider et d'améliorer constamment les hypothèses de modélisation.

Question 2.1. *Cette première hypothèse de modélisation (à savoir, les données observées sont des réalisations de v.a. i.i.d.) vous semble-t-elle raisonnable ? Justifier votre réponse.*

Soit donc $X_1, \cdots, X_n$ n variable aléatoires indépendantes de fonction de répartition commune F inconnue (et dont on a observé les 100 réalisations ci-dessus). On Suppose aussi que F est continue.

Question 2.2. *Cette deuxième hypothèse de modélisation portant sur la continuité de F vous semble-t-elle raisonnable en vue des* 100 *observations ci-dessus ? Justifier votre réponse.*

(CI) Analyse statistique inférentielle

L'objectif de l'analyse inférentielle est de faire une étude dont le but est de pouvoir généraliser les conclusions tirées sur l'échantillon de données (ici les 100 observations ci-dessus) à l'échelle de la population ce qui correspond ici à toute la fabrication de cette entreprise (produits fabriqués dans des mêmes conditions, tout en s'assurant que nos hypothèses de modélisation demeurent réalistes).

On démarre de notre hypothèse de modélisation : $X_1, \cdots, X_n$ n v.a. i.i.d. de fonction de répartition F inconnue qu'on suppose continue.

Un estimateur naturel de F est la fonction de répartition empirique F_n de cet échantillon aléatoire $(X_1, \cdots, X_n)$. On rappelle que, pour $x \in \mathbb{R}$,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{X_i \leq x},$$

qui est un estimateur de $F(x)$ et dont une réalisation pour $n = 100$ est la fonction f_{100} ci-haut introduite (pour x fixé, $f_{100}(x)$ est une estimation de $F_{100}(x)$).

Exercice 2.2.
En utilisant l'inégalité DKW, construire et calculer une bande de confiance pour F au niveau 0.95.

Le graphe de f_{100} (voir Figure 2.1) laisse croire que F peut être la fonction de répartition d'une loi exponentielle de paramètre λ, que l'on notera par $\mathcal{E}(\lambda)$. Comme λ est l'inverse de l'espérance de la loi $\mathcal{E}(\lambda)$ et comme le fabricant affirme que la durée de vie moyenne est de 2, on prendra sans mettre en doute l'affirmation du constructeur $\lambda = 1/2$. Soit donc F_0 la fonction de répartition de la loi $\mathcal{E}(0.5)$. On souhaiterait affirmer ou infirmer (avec une certaine marge d'erreur) la conclusion que les données observées sont issues de la loi $\mathcal{E}(0.5)$. On pose donc le test statistique suivant :

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0 \tag{2.1}$$

qu'on se propose de résoudre. Il s'agit d'un test d'ajustement à une loi donnée. Comme F est supposée continue, [2] On peut utiliser le test de Kolmogorov-Smirnov.

Question 2.3. *Préciser les différences entre F, F_0 et F_n ?.*

Exercice 2.3.

1. Quelle est la statistique du test décrit dans (2.1) ? Vérifier qu'elle est libre ? Comment peut-on la calculer sur les données ?
2. Résoudre ce problème de test en précisant sa p-valeur ?
3. En quoi l'hypothèse de modélisation est utile pour cette étude ?
4. Conclure tout en répondant à la question (Q1) de l'entreprise.

2. Dans le cas discret, les tests d'adéquations de khi-deux peuvent être utilisés.

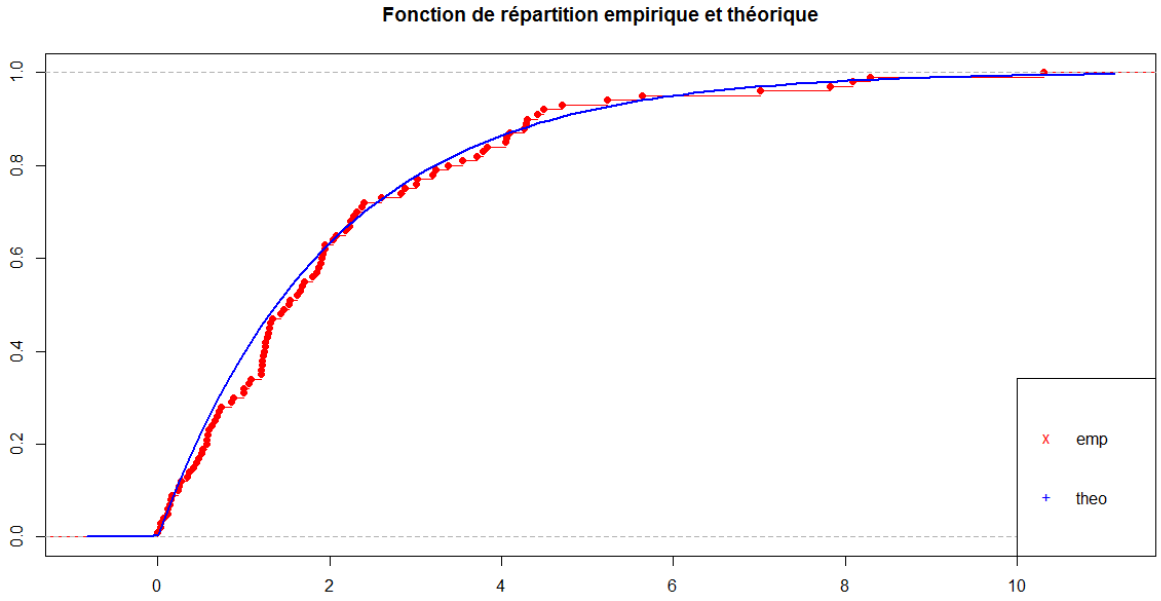


FIGURE 2.1 – Graphe de f_{100} en rouge

2.3 Problématique II.

On reprend la problématique I mais en mettant en doute l'affirmation du fabricant sur la durée de vie moyenne de son produit. On ne supposera plus que c'est 2 mais plutôt une autre valeur inconnue. Il s'agit ici de faire un test d'ajustement à une famille de lois qui est la famille paramétrique de la loi exponentielle, de fonction de répartition

$$F_\lambda(x) = (1 - \exp(-\lambda x))\mathbb{I}_{x>0},$$

pour $\lambda > 0$. Soit donc le test,

$$H_0 : F \in \{F_\lambda,\ \lambda > 0\} \quad \text{contre} \quad H_1 : F \notin \{F_\lambda,\ \lambda > 0\} \tag{3.2}$$

Exercice 2.4.

1. Quelle est la statistique du test décrit dans (3.2) ? Vérifier qu'elle est libre sous H_0. Comment peut-on la calculer sur les données ?
2. Résoudre ce problème de test en précisant sa p-valeur ?
3. Conclure tout en répondant à la question (Q1) de l'entreprise.

2.4 Problématique III.

Une entreprise concurrente affirme que ses composants ont plus de performances que ceux fabriqués par la première entreprise. Elle fournit les 150 observations, ci-dessous, correspondants aux durées de vie de 150 composants issus de sa fabrication. Les données sont exprimées dans une même unité de temps.

```
 5.500572685  3.629419532  3.210846633  2.573115200  0.955138974
 1.138221238  1.986082289  0.928477189  7.621581792  1.615409225
 3.666974686  1.291723177  4.813959047  9.607209520  0.063010045
 2.136582317 14.139939013 4.842176705  5.156777099  2.224679714
16.530562244  5.113496671  1.382457823  7.923361130  0.555608666
 6.861744336  1.205689467  4.253182385 10.635679267  6.979759449
 2.114261663  1.364466386  5.184430338  8.376332931  4.770425800
11.884361639  0.977662448  1.192509571  0.191949673  2.072348150
 0.441168034  4.463605441 12.235093438  0.020248058  7.090439863
 3.545986269  2.381299727  8.280886322  2.150365733  0.284614541
 0.392262649  3.094586540  2.465039719  3.602775588  4.530802749
 2.285899062  3.721908919  0.082389496  0.872212087  4.763199467
 0.037777852  0.132963309  7.949266165  3.851722455  1.500339672
 8.291128948  4.509009012  3.476394738  5.195761178  1.135951445
14.790169462  7.089815805  0.902702626  6.843184964  0.673886089
11.598430659  5.083418399  3.192542136  2.372019222  4.681674275
 0.581912899  5.000172350  1.782066109  8.737255103  2.466807645
 8.636634789  3.271948365  2.105900194  6.735459141  6.555102497
 0.606414475  1.606718119  2.566719053  1.738496067  5.416493594
 0.251510061 24.828956543  2.806184295  6.512502886  0.564252755
 2.435787091  3.008054040  5.264472376  4.858323902  5.564861036
 4.344310080  0.006596938  0.256606142  5.851465560 12.597033291
 0.636218116  9.112586341  8.392572000  2.422233233 10.373495914
15.424963794 0.434714849  0.115228076  1.703677222  0.969835890
 2.161658874  0.473787708 11.014415659  6.688485272  2.130372038
15.390809438 1.170186588  6.856871009  1.618636355  6.235859349
 2.269207072  5.872250039  1.082619295  3.106471402  4.661225878
 0.423852941  1.427732101  5.441074412  3.319499659  1.621479272
 3.674349280 14.415418885 11.395503827 11.624368945  0.749001054
 0.781485433  1.235254641  5.939705122  9.121810145 15.411781213
```

On vous fournit ces données afin de voir :

(Q3) s'il y a une différence significative entre ces deux jeux de données
(Q4) si on peut considérer que les composants fabriqués par la seconde entreprise sont plus performants que ceux construits par la première entreprise (comme l'affirmait son fabricant).

(AIII) Étude descriptive

On notera par $y_1, \cdots, y_{150}$ ces 150 observations.

Exercice 2.5.

1. Tracer la fonction $g_{150} : x \mapsto \frac{1}{150} \sum_{i=1}^{150} \mathbb{I}_{y_i \leq x}$
2. Représenter sur le même graphe les fonctions g_{150} et f_{100}. Que peut-on observer ? [voir, pour un élément de réponse, Figure 2.2]

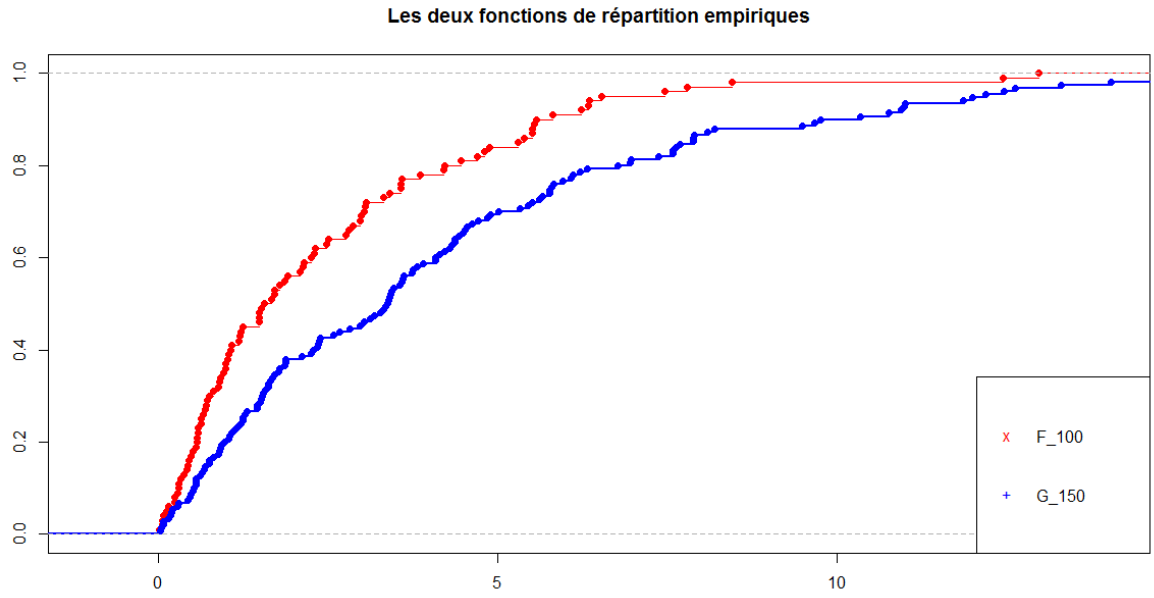


FIGURE 2.2 –

L'observation de ces deux échantillons de données ci-dessus ainsi que l'étude empirique, portant sur les deux graphes des fonctions de répartitions empiriques f_{100} et g_{150} (Figure 2.2), laissent croire que les durées de vies des composants de la 2ème entreprise sont plus longues que ceux de la première entreprise. Cette étude n'est basée que sur l'observation de deux échantillons de données, ce qu'on peut affirmer ne concerne que ces deux échantillons de données et non pas toutes les populations (i.e. toutes les fabrications de ces deux entreprises). D'où l'importance des deux étapes suivantes.

(BIII) Modélisation et Inférence statistique

Hypothèses de modélisation

On supposera que les données de la seconde entreprise sont des réalisations de 150 variables aléatoires $Y_1, \cdots, Y_{150}$ i.i.d. de fonction de répartition G inconnue qu'on supposera, ici aussi, continue. Nous sommes donc en présence de deux échantillons aléatoires $X_1, \cdots, X_n$ et $Y_1, \cdots, Y_m$. On supposera que ces deux échantillons aléatoires sont indépendants.

Question 2.4. *Ces hypothèses de modélisation (en particulier la continuité de G et l'indépendance de deux échantillons) vous semblent-elles raisonnables ?*

(CIII) Analyse statistique inférentielle

On a donc deux échantillons aléatoires $(X_1, \cdots, X_n)$ et $(Y_1, \cdots, Y_m)$ vérifiant :

— ces 2 échantillons sont indépendants
— les v.a. $X_1, \cdots, X_n$ sont i.i.d. Il en est de même pour $Y_1, \cdots, Y_m$
— X_1 a pour fonction de répartition F continue et inconnue

— Y_1 a pour fonction de répartition G continue et inconnue.

Pour répondre à la question (Q_3), on fera un test de kolmogorov-Smirnov de comparaison de deux échantillons indépendants, soit

$$H_0 : F = G \quad \text{contre} \quad H_1 : F \neq G \tag{4.3}$$

Exercice 2.6.

1. Quelle est la statistique du test décrit dans (4.3) ? Vérifier qu'elle est libre sous H_0. Comment peut-on la calculer sur les données ?
2. Résoudre ce problème de test en précisant sa p-valeur ?
3. En quoi les hypothèses de modélisation sont-elles utiles pour cette étude théorique ?
4. Conclure tout en répondant à la question (Q3) de l'entreprise.

L'objectif maintenant est de savoir si la 2ème entreprise est plus performante que la première.

Exercice 2.7.

1. Montrer que si $X_1 \leq Y_1$ ps alors $G(x) \leq F(x)$ pour tout $x \in \mathbb{R}$.
2. Que peut-on déduire quant aux performances de ces deux entreprises si $G \leq F$ (i.e. $G(x) \leq F(x)$ pour tout $x \in \mathbb{R}$) ?

On se propose maintenant d'étudier le test de comparaison unilatéral suivant :

$$H_0 : F = G \quad \text{contre} \quad H_1 : G < F \tag{4.4}$$

Exercice 2.8.

1. Quelle est la statistique du test décrit dans (4.4) ? Vérifier qu'elle est libre sous H_0. Comment peut-on la calculer sur les données ?
2. Résoudre ce problème de test en précisant sa p-valeur ?
3. Conclure tout en répondant à la question (Q4) de l'entreprise.

2.5 Problématique VI.

On considère dans cette section, les données de la première entreprise et on souhaite répondre à la question (Q2) ou plus généralement à la question suivante :

(Q5) Estimer le premier instant au-delà duquel au maximum une proportion de $x\%$ de la fabrication est encore fonctionnelle (la valeur de x étant donnée).

L'instant $t_{100,x}$ (sur les 100 observations fournies par la première entreprise) vérifie :

$$t_{100,x} = \inf\left\{t > 0,\ \ f_{100}(t) \geq 1 - \frac{x}{100}\right\}$$

L'instant $t_{100,x}$ est donc un quantile empirique d'ordre $1 - \frac{x}{100}$ et le premier instant qu'on cherche à estimer n'est autre que

$$Q(1 - \frac{x}{100})\ (= F^{(-1)}(1 - \frac{x}{100})),$$

pour $x \in]0, 100[$.

Exercice 2.9.
On fixe $x = 30$.

1. Donner une estimation ponctuelle de $Q(0.7)$ (en se basant sur les 100 observations).
2. Construire un intervalle de confiance pour $Q(0.7)$ au niveau 0.95.
3. Comparer avec les quantiles d'une loi exponentielle dont on précisera le paramètre (on pourra représenter graphiquement le diagramme Quantile-Quantile (QQ plot)).

Chapitre 3

Exercices (plutôt théoriques)

Exercice 3.1.
Soit $X_1, \cdots, X_n$ n v.a. réelles i.i.d. de fonction de répartition F continue. Montrer que,

$$\mathbb{P}(\exists\ i \neq j,\ \ X_i = X_j) = 0.$$

En déduire que, ps,

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)}.$$

Exercice 3.2.
Montrer le théorème de Glivenko-Cantelli.

Exercice 3.3.
Soit $(X_1, \cdots, X_n)$ un échantillon de v.a. i.i.d. de fonction de répartition F. Soit F_n la fonction de répartition empirique de cet échantillon. Soit F_0 une fonction de répartition donnée. On rappelle que,

$$\|F_n - F_0\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

1. Montrer que si F_0 est continue alors

$$\|F_n - F_0\|_\infty = \max_{1 \leq j \leq n} \max\{j/n - F_0(X_{(j)}), F_0(X_{(j)}) - (j-1)/n\}.$$

   Que se passe-t-il si F_0 n'est pas continue ?.
2. Montrer en utilisant le théorème de Glivenko-Cantelli que si $F \neq F_0$ alors presque sûrement

$$\liminf_{n \to \infty} \|F_n - F_0\|_\infty > 0.$$

3. Soit $(U_1, \cdots, U_n)$ un échantillon de loi uniforme sur $[0, 1]$. Montrer que $(X_1, \cdots, X_n)$ et $(F^{(-1)}(U_1), \cdots, F^{(-1)}(U_n))$ ont la même loi.
4. Montrer qu'on a, ps, $(F^{(-1)}(U_i) \leq x)$ est équivalent à $U_i \leq F(x)$.
5. En déduire que si $F = F_0$ alors la loi de $\|F_n - F_0\|_\infty$ ne dépend de F_0 qu'à travers $F_0(\mathbb{R})$.
6. Montrer que si F_0 est continue et si $F = F_0$ alors $\|F_n - F_0\|_\infty$ a la même loi que $\|G_n - F_U\|_\infty$ (F_U est la fonction de répartition de la loi uniforme sur $[0, 1]$, G_n est la fonction de répartition empirique de l'échantillon $U_1, \cdots, U_n$)
7. On souhaite tester $H_0 : F = F_0$ contre $H_1 : F \neq F_0$.

— Préciser une région de rejet de H_0 pour laquelle l'erreur de première espèce est inférieur ou égale à α. On dira que le test est de **niveau** α.
— Vérifier que si F_0 est continue alors ce test est même de **taille** α, c'est-à-dire il existe une région de rejet de l'hypothèse H_0 pour laquelle l'erreur de première espèce est α.

Exercice 3.4.
Soit X une variable aléatoire de fonction de répartition inconnue F, et soit $(X_1, \cdots, X_n)$ un n-échantillon de la loi de X. Pour tout $\theta > 0$, on note par F_θ la fonction de répartition définie par

$$F_\theta(x) = (1 - \exp(-x/\theta))I_{x>0}.$$

1. On suppose que $F \in \mathcal{F}$ où $\mathcal{F} = \{F_\theta, \theta \in \mathbb{R}\}$. Déterminer l'estimateur $\hat{\theta}_n$ du maximum de vraisemblance de θ.
2. On note par F_n la fonction de répartition empirique de $(X_1, \cdots, X_n)$. Montrer que la loi de $\sup_{t\in\mathbb{R}} |F_n(t) - F_{\hat{\theta}_n}(t)|$ est libre lorsque $F \in \mathcal{F}$. En déduire un test de l'hypothèse $H_0 : F \in \mathcal{F}$ contre $H_1 : F \notin \mathcal{F}$.

Exercice 3.5.
Soient $(X_1, \cdots, X_n)$ un n-échantillon d'une loi de fonction de répartition F continue et $(Y_1, \cdots, Y_m)$ un m-échantillon de fonction de répartition G continue, et est indépendant de $(X_1, \cdots, X_n)$. On note par F_n la fonction de répartition empirique de $(X_1, \cdots, X_n)$, par G_m la fonction de répartition empirique de $(Y_1, \cdots, Y_m)$ et on définit

$$\|F_n - G_m\|_\infty = \sup_{t\in\mathbb{R}} |F_n(t) - G_m(t)|.$$

On s'intéresse aux hypothèses $H_0 : F = G$ contre $H_1 : F \neq G$.

1. Montrer que si les v.a. $(X_i)_{1\leq i\leq n}$ et $(Y_j)_{1\leq j\leq m}$ ont la même loi, alors la loi de $\|F_n - G_m\|_\infty$ est libre des lois des observations.
2. En utilisant le fait que la loi de $\|F_n - G_m\|_\infty$ est connue sous H_0, construire un test de H_0 contre H_1.

Chapitre 4

Exercices (plutôt appliqués)

Exercice 4.1.
Illustrer graphiquement (par simulation) les résultats de convergence de la Proposition 1.1.

Exercice 4.2.
Écrire des programmes avec le logiciel de votre choix permettant d'avoir les illustrations graphiques du Mémento sur des données simulées.

Exercice 4.3.
On rappelle le théorème de Kolmogorov-Smirnov : soient $X_1, \cdots, X_n$ n. v.a. i.i.d. de fonction de répartition F_0 continue. Alors

$$T_n = \sqrt{n} \sup_x |F_n(x) - F_0(x)|$$

converge en loi vers la loi de Kolmogorov-Smirnov, dont la fonction de répartition est donnée, pour tous $x > 0$, par

$$F_{KS}(x) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2x^2k^2}$$

1. Tracer la fonction de répartition de la loi de Kolmogorov-Smirnov et illustrer cette convergence en loi.
2. Vérifier par simulation que la loi asymptotique de T_n ne dépend pas de la loi de X_1.

Exercice 4.4.
Illustrer le théorème de Glivenko-Cantelli sur des données simulées.

Exercice 4.5.
On a testé un échantillon de 5 appareils et noté leurs durées de vie en heures :

Appareil	1	2	3	4	5
Durée de vie	133	169	8	122	58

On voudrait savoir si on peut modéliser la loi de la variable durée de vie par une loi exponentielle.

1. Estimer le paramètre λ de la loi exponentielle (de densité $\lambda e^{\lambda x} I_{x \geq 0}$).
2. Formuler les hypothèses du test (H_0 et H_1).
3. Comparer la distribution observée à la distribution théorique au moyen d'un test de Kolmogorov-Smirnov. Calculer la p-valeur du test. Que peut-on en conclure ?

Bibliographie

[1] WikiStat cliquer ici

[2] All of Nonparametric Statistics. Larry Wasserman. Springer (2006).

[3] Statistique mathématique en action, Master et Agrégation externe de mathématiques (2e édition). Vincent Rivoirard, Gilles Stoltz Vuibert (2012).

[4] Pour une introduction à quelques logiciels voir la page web de Sophie Lemaire
cliquer ici

Index

Échantillon aléatoire, 1, 15, 18
Échantillon de données, 14

Empirique, 1, 14

Fonction de répartition empirique, 15
Fonction réciproque généralisée, 5

Hypothèse de modélisation, v, 14, 18

Inférence, 14, 15, 18
Inégalité DKW, 4, 15

Libre (loi), 5

Modèle statistique, v
Modélisation, vi, 14, 18
Modélisation probabiliste, vii
Modélisation statistique, vi, 13

Quantile, 5
Quantile empirique, 6

Simulations, v
Statistique d'ordre, 2

Table des figures

1.1 Illustrations de la convergence ps de la fonction de répartition empirique F_n (em) vers la fonction de répartition F (theo). 2
1.2 Illustration de la normalité asymptotique de $F_n(x)$ pour x fixé (ici $x = 0.4$). 3
1.3 Illustration de la convergence p.s. et uniforme de $\|F_n - F\|_\infty$ vers 0. 4
1.4 Fonction de répartition de la loi de Kolmogorov-Smirnov. 5
1.5 Illustration de la convergence en loi de $T_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ vers la loi de Kolmogorov-Smirnov. 6
1.6 Fonction de répartition et fonction quantile d'une loi exponentielle. 7
1.7 Ici $Q_n(0.5) = 1$. 8
1.8 Histogramme de la distribution de $\sqrt{n}(Q_n(0.5) - Q(0.5))$ et courbe de la densité d'une loi normale centrée et de variance $\frac{0.25}{f^2(Q(0.5))}$ (cas d'un échantillon simulé de loi normale). 9

2.1 Graphe de f_{100} en rouge 16
2.2 18

Notations

càdlàg continue à droite ayant une limite à gauche 2
i.e. c'est-à-dire 1
i.i.d. indépendantes et identiquement distribuées 1
ps presque sûrement 1
v.a. variable aléatoire 6