Tests statistiques, la suite ...

Cathy Maugis-Rabusseau (INSA Toulouse / IMT)

2020-2021

Contents

Préface

Ce polycopié s'inspire de polycopiés antérieurs faits par des collègues du GMM, qu'ils en soient ici remerciés.

Chapter 1

Rappels sur les tests

Dans ce chapitre, le vocabulaire de base de la théorie des tests est rappelé. L'ensemble des tests paramétriques vus en 3ème année ne seront pas rappelés ici, une feuille de TD leur sera dédiée pour révision.

1.1 Rappels généraux sur les tests statistiques

1.1.1 Hypothèse nulle et hypothèse alternative

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et X une v.a de $(\Omega, \mathcal{A})$ dans $(E, \mathcal{E})$. On se donne un modèle statistique, c'est-à-dire une famille de probabilité sur $(E, \mathcal{E})$: $\{P_\theta,\ \theta \in \Theta\}$. On considère un n-échantillon $\mathcal{X} = (X_1, \dots, X_n)$ dont la loi est supposée appartenir à $\{P_\theta,\ \theta \in \Theta\}$.

Se poser un problème de test consiste tout d'abord à définir deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, appelées hypothèse nulle et hypothèse alternative respectivement. On considère donc deux sous-ensembles disjoints Θ_0 et Θ_1 de Θ et on dit que l'on teste

$$\mathcal{H}_0 : \theta \in \Theta_0 \text{ contre } \mathcal{H}_1 : \theta \in \Theta_1.$$

A partir de l'échantillon $\mathcal{X}$, on souhaite alors construire une règle de décision (région de rejet) pour décider entre ces deux hypothèses.

Rappelons que les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ ne jouent pas un rôle symétrique. L'hypothèse nulle est l'hypothèse que l'on privilégie car elle est présumée vraie tant que l'échantillon observé ne conduit pas à la rejeter au profit de l'hypothèse alternative. On parlera d'**hypothèse simple** lorsque le sous-ensemble associé est un singleton et d'**hypothèse composite** sinon.

1.1.2 Tests statistiques

Definition 1.1. Un test statistique consiste en une partition de Ω en deux ensembles : l'ensemble $\mathcal{R}$ des valeurs possibles de l'échantillon qui conduisent au rejet de $\mathcal{H}_0$ au profit de $\mathcal{H}_1$, appelée **région de rejet** (ou région critique) du test, et son complémentaire.

Definition 1.2. On appelle **fonction de test** de région de rejet $\mathcal{R}$ la statistique

$$\phi(x) = \mathbb{1}_{x\in\mathcal{R}}.$$

Autrement dit, si $\phi(x) = 1$, on rejette $\mathcal{H}_0$, et si $\phi(x) = 0$, on ne rejette pas $\mathcal{H}_0$.

1.1.3 Erreur de première espèce et p-valeur

Definition 1.3. Etant donné un test de $\mathcal{H}_0$ contre $\mathcal{H}_1$ de région de rejet $\mathcal{R}$, la fonction **erreur de première espèce** est définie pour tout $\theta_0 \in \Theta_0$ par

$$\underline{\alpha}(\theta_0) = \mathbb{P}_{\theta_0}(\mathcal{X} \in \mathcal{R}).$$

La **taille du test** correspond à l'erreur de première espèce maximale

$$\alpha^\star = \sup_{\theta_0\in\Theta_0} \mathbb{P}_{\theta_0}(\mathcal{X} \in \mathcal{R}).$$

Definition 1.4. Soit $\alpha \in [0, 1]$ et soit un test de région de rejet $\mathcal{R}$ pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$. On dit que ce test est

- **de niveau** α s'il est de taille au plus α ($\alpha^\star \leq \alpha$)
- de **niveau exactement** α s'il est de taille α ($\alpha^\star = \alpha$)
- de **niveau asymptotique** α si $\alpha^\star \underset{n\to+\infty}{\longrightarrow} \alpha$

Definition 1.5. Supposons avoir construit pour tout $\alpha \in]0, 1[$ un test de niveau α de $\mathcal{H}_0$ contre $\mathcal{H}_1$ de région de rejet $\mathcal{R}_\alpha$. On appelle **p-valeur** de la famille de tests le plus petit seuil à partir duquel on rejette $\mathcal{H}_0$ à partir de l'échantillon observé $\mathcal{X}^{obs}$

$$p(\mathcal{X}^{obs}) = \inf\{\alpha \in]0, 1[;\ \mathcal{X}^{obs} \in \mathcal{R}_\alpha\}.$$

1.1.4 Erreur de seconde espèce et puissance

Definition 1.6. Soit un test de région de rejet $\mathcal{R}$ pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$. La fonction **erreur de seconde espèce** de ce test est définie pour tout $\theta_1 \in \Theta_1$ par

$$\underline{\beta}(\theta_1) = \mathbb{P}_{\theta_1}(\mathcal{X} \notin \mathcal{R})$$

et l'erreur de seconde espèce maximale vaut

$$\beta^\star = \sup_{\theta_1\in\Theta_1} \mathbb{P}_{\theta_1}(\mathcal{X} \notin \mathcal{R}).$$

Definition 1.7. On appelle **fonction puissance** du test basé sur la région de rejet $\mathcal{R}$ l'application

$$\pi : \theta_1 \in \Theta_1 \mapsto \mathbb{P}_{\theta_1}(\mathcal{X} \in \mathcal{R}) = 1 - \beta(\theta_1) \in [0,1].$$

Parmi les tests de même niveau on préfère toujours celui qui est le plus puissant.

Definition 1.8. On dira que le test basé sur la région de rejet $\mathcal{R}$ est meilleur que celui basé sur la région de rejet $\mathcal{R}'$ s'ils sont tous les deux de niveau α et que

$$\forall \theta \in \Theta_1,\ \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}) \geq \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}').$$

Definition 1.9. On dit que le test basé sur la région de rejet $\mathcal{R}_\alpha$ est **uniformément plus puissant (UPP)** au niveau α si :

1. $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}_\alpha) \leq \alpha$.
2. Pour toute région de rejet $\mathcal{R}'_\alpha$ telle que $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}'_\alpha) \leq \alpha$, on a

$$\forall \theta \in \Theta_1,\ \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}_\alpha) \geq \mathbb{P}_\theta(\mathcal{X} \in \mathcal{R}'_\alpha).$$

1.2 Tests paramétriques (MIC3)

Dans l'UF de statistique de MIC3, les tests suivants ont été étudiés :

- Avec un échantillon gaussien,
  - Test de conformité de la moyenne avec variance connue
  - Test de conformité de la moyenne avec variance inconnue
  - Test de conformité de la variance
- Avec un échantillon non gaussien,
  - Test de conformité de la moyenne
- Avec deux échantillons gaussiens
  - Test de comparaison des deux moyennes
  - Test de comparaison des deux variances

Pour la construction de tous ces tests, on suppose que la loi de(s) échantillon(s) appartient à un modèle paramétrique c'est-à-dire une famille de lois donnée décrite par un nombre fini de paramètres. On parle alors de **tests paramétriques**, mais en général, cette hypothèse est difficilement vérifiée en pratique. On parle de test non-paramétrique quand il est valable quelque soit la loi de l'échantillon. Dans la suite, nous allons étudier quelques tests non-paramétriques pour répondre à différents objectifs : test d'ajustement, test d'indépendance de deux échantillons, test d'homogénéité, …

Chapter 2

Tests basés sur la fonction de répartition empirique et sur les rangs

Les slides associés à ce chapitre sont disponibles ici

2.1 Rappels

2.1.1 Fonction de répartition et quantiles

Soit X une v.a.r de fonction de répartition F. On rappelle que, pour tout $t \in \mathbb{R}$,

$$F(t) = \mathbb{P}(X \leq t).$$

Definition 2.1. Soit F une fonction de répartition. On définit la **fonction quantile (ou inverse généralisée)** F^{-1} de F par

$$\forall p \in [0,1],\ F^{-1}(p) = \inf\{t \in \mathbb{R};\ F(t) \geq p\}.$$

Remark. Si F est une bijection, F^{-1} est la bijection réciproque.

Exercise 2.1. Calculez F^{-1} pour la loi de Bernoulli de paramètre θ.

Proposition 2.1. *Soit $t \in \mathbb{R}$ et $p_0 \in]0,1[$.*

1. *F est croissante, continue à droite, $\lim_{t\to-\infty} F(t) = 0$, $\lim_{t\to+\infty} F(t) = 1$.*
2. *$\{t \in \mathbb{R};\ F(t) \geq p_0\} = [F^{-1}(p_0), +\infty[$*
3. *F^{-1} est croissante*
4. *$F \circ F^{-1}(p_0) \geq p_0$ avec égalité si $p_0 \in F(\mathbb{R})$.*

5. $F(t) \geq p_0 \Leftrightarrow t \geq F^{-1}(p_0)$
6. *Si* $U \sim \mathcal{U}([0,1])$, $F^{-1}(U)$ *a pour fonction de répartition* F.

2.1.2 Fonction de répartition empirique

Soit $X_1, X_2, ... X_n$ une suite de variables aléatoires réelles i.i.d. de fonction de répartition F.

Definition 2.2. On appelle **fonction de répartition empirique** associée au n-échantillon $(X_1, X_2, ... X_n)$ la fonction

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}.$$

Remark. La fonction de répartition empirique de l'échantillon $(X_1, ..., X_n)$ s'exprime également à partir des statistiques d'ordre $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_{(i)} \leq t}$$

ce qui permet de tracer facilement son graphique (voir Figure 2.1).

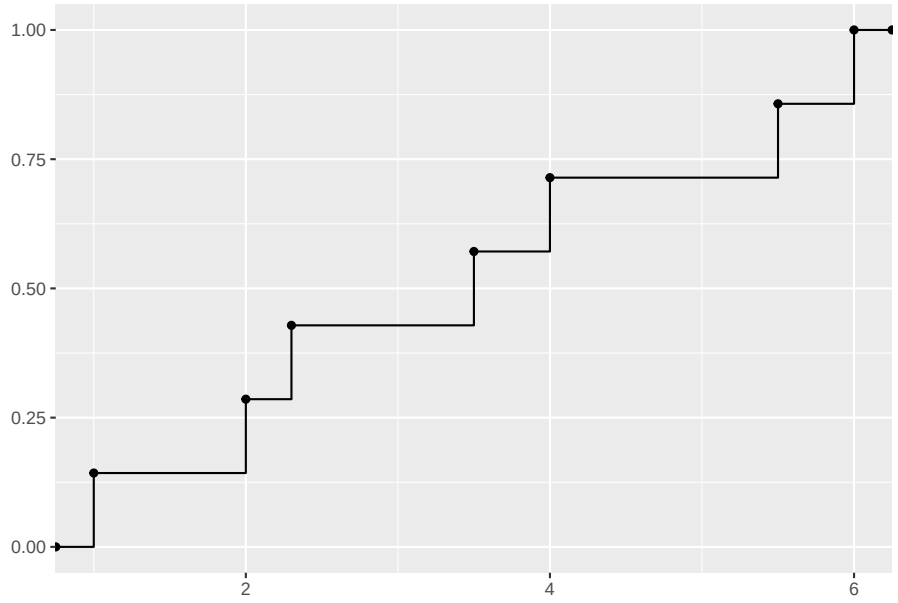


Figure 2.1: Fonction de répartition empirique pour l'échantillon observé *(2, 3.5, 1, 4, 2.3, 6, 5.5)*

Rappelons quelques propriétés de la fonction de répartition empirique.

Proposition 2.2. *Propriétés de la fonction $\hat{F}_n(.)$*

- *$\hat{F}_n$ est croissante, continue à droite, $\lim_{t\to-\infty} \hat{F}_n(t) = 0$, $\lim_{t\to+\infty} \hat{F}_n(t) = 1$.*
- *Pour tout $t \in \mathbb{R}$, $n\hat{F}_n(t)$ suit une loi binomiale de paramètre $(n, F(t))$.*
- *Pour tout $t \in \mathbb{R}$, $\hat{F}_n(t)$ est un estimateur sans biais de $F(t)$.*
- *Pour tout $t \in \mathbb{R}$*

$$Var(\hat{F}_n(t)) = \frac{F(t)(1-F(t))}{n} \underset{n\to+\infty}{\longrightarrow} 0.$$

- *Pour tout $t \in \mathbb{R}$, $\hat{F}_n(t) \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} F(t)$*
- *On déduit du TLC que pour tout $t \in \mathbb{R}$ tel que $F(t)(1-F(t)) \neq 0$,*

$$\sqrt{n}(\hat{F}_n(t) - F(t)) \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0, F(t)(1-F(t))).$$

- *Glivenko-Cantelli (admis)*

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| \underset{n\to+\infty}{\overset{p.s}{\longrightarrow}} 0.$$

2.2 Test de Kolmogorov de comparaison ou d'adéquation

Soit $X_1, \dots, X_n$ des v.a.r i.i.d. de même loi que X, de fonction de répartition F supposée continue. On se donne une fonction de répartition F_0 supposée continue sur $\mathbb{R}$ et Y_0 une v.a.r de fonction de répartition F_0.

On souhaite construire un test de $\mathcal{H}_0$: "X et Y_0 ont la même loi $(F = F_0)$" contre :

- $\mathcal{H}_1$: X et Y_0 ne suivent pas la même loi $(F \neq F_0)$
- $\mathcal{H}_1^+$: X a tendance à prendre des valeurs plus petites que Y_0 $(F \geq F_0)$
- $\mathcal{H}_1^-$: X a tendance à prendre des valeurs plus grandes que Y_0 $(F \leq F_0)$

Example 2.1. On mesure les durées de vie de 20 ampoules d'un même type. Les résultats, en heures, sont : 673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916.
Est-ce que l'on peut affirmer, au risque 5%, que la durée de vie d'une ampoule de ce type ne suit pas la loi exponentielle $\mathcal{E}(1/1500)$?
On modélise donc la durée de vie de la $i$ème ampoule par X_i, F est sa fonction de répartition inconnue et F_0 est la fonction de répartition de la loi $\mathcal{E}(1/1500)$.

L'idée du test de Kolmogorov est d'estimer la fonction de répartition inconnue F par la fonction de répartition empirique $\hat{F}_n$ de l'échantillon $(X_1, \dots, X_n)$ et de comparer cette fonction de répartition empirique avec la fonction de répartition donnée F_0.

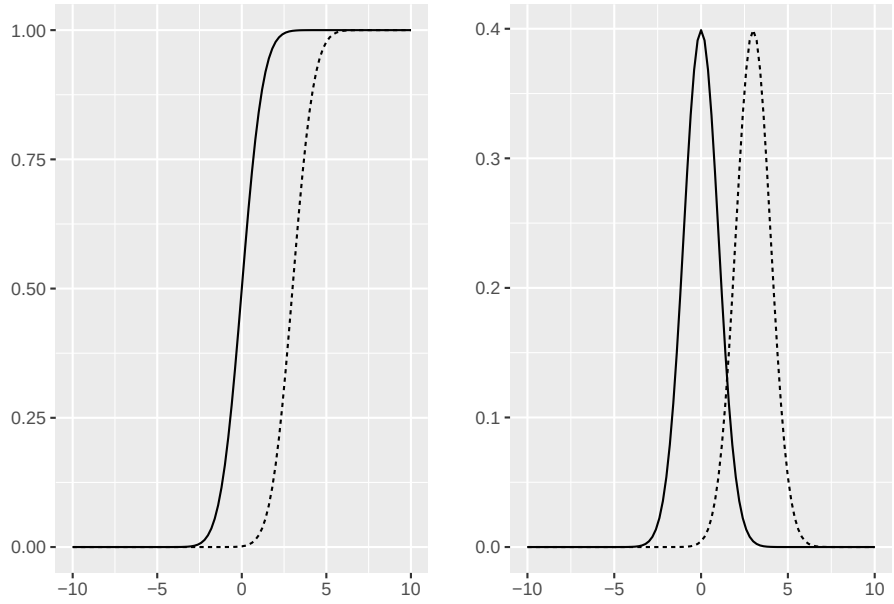


Figure 2.2: Fonction de répartition (à gauche) et densité (à droite) pour la loi N(0,1) et N(3,1)

Definition 2.3. Pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$, le **test de Kolmogorov** est fondé sur la statistique de test

$$D_n = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F_0(t)|.$$

La région de rejet au niveau α est alors de la forme $\mathcal{R}_\alpha = \{D_n \geq d_{n,1-\alpha}\}$.

Proposition 2.3. *Propriétés de la statistique D_n*

- *La loi de D_n sous l'hypothèse H_0 ($F = F_0$) est indépendante de F_0.*
- *Comme $\hat{F}_n$ est une fonction en escalier et que F_0 est croissante, l'écart maximal entre $\hat{F}_n$ et F_0 est atteint en l'un des sauts de $\hat{F}_n$. Ainsi, avec $X_{(1)} \leq ... \leq X_{(n)}$ l'échantillon ordonné, $X_{(0)} = -\infty$ et $X_{(n+1)} = +\infty$, on obtient que*

$$D_n = \max_{i=0,\dots,n} \left\{ \max\left(\left| \frac{i}{n} - F_0(X_{(i)}) \right| ; \left| \frac{i}{n} - F_0(X_{(i+1)}) \right| \right) \right\},$$

ce qui permet de calculer facilement D_n.

Remark. La loi de D_n sous H_0 est tabulée. On trouve dans les tables les quantiles $d_{n,1-\alpha}$ tels que

$$\mathbb{P}_{H_0}(D_n \geq d_{n,1-\alpha}) \leq \alpha,$$

(en étant le plus proche possible de α). Ces tables sont obtenues à partir de simulations de D_n, sous l'hypothèse que les X_i sont i.i.d. de loi uniforme sur $[0,1]$ ($F_0 = \mathbb{1}_{[0,1]}$). Si la loi de D_n dépendait de F_0, il faudrait construire une table pour chaque loi F_0.

Proposition 2.4. *De la même façon, pour tester*

- $\mathcal{H}_0 : F = F_0$ *contre* $\mathcal{H}^+ : F \geq F_0$, *on utilise*

$$D_n^+ = \sup_{t\in\mathbb{R}}(\hat{F}_n(t) - F_0(t))$$

  et la région de rejet de niveau α *est de la forme* $\mathcal{R}_\alpha = \{D_n^+ > d_{n,1-\alpha}^+\}$.
- $\mathcal{H}_0 : F = F_0$ *contre* $\mathcal{H}^- : F \leq F_0$, *on utilise*

$$D_n^- = \sup_{t\in\mathbb{R}}(F_0(t) - \hat{F}_n(t))$$

  et la région de rejet de niveau α *est de la forme* $\mathcal{R}_\alpha = \{D_n^- > d_{n,1-\alpha}^-\}$.

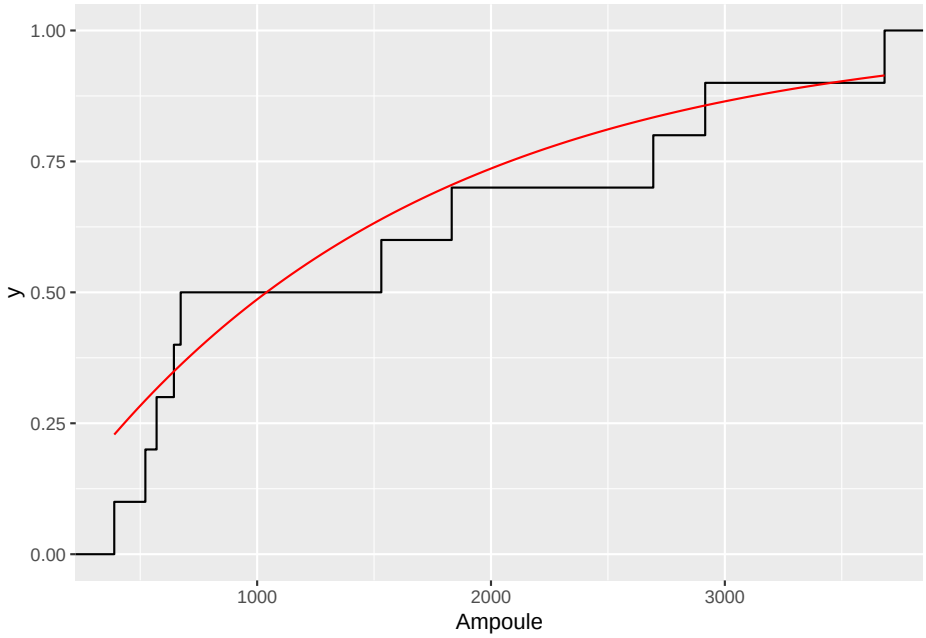
Proposition 2.5. *(admise)*

$\forall\lambda > 0,\ \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n^+ \geq \lambda) \underset{n\to+\infty}{\longrightarrow} \exp(-2\lambda^2)$	*Smirnov (1942)*
$\forall\lambda > 0, \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n \geq \lambda) \underset{n\to+\infty}{\longrightarrow} 2\sum_{k=1}^{\infty}(-1)^{k+1}\exp(-2k^2\lambda^2)$	*Kolmogorov (1933)*
$\forall\lambda > 0,\ \mathbb{P}_{\mathcal{H}_0}(\sqrt{n}D_n \geq \lambda) \leq 2\exp(-2\lambda^2)$	*Massart (1990)*

Example 2.2. Revenons à notre exemple sur la durée de vie des ampoules. La fonction de répartition empirique et la fonction de répartition de la loi $\mathcal{E}(1/1500)$ sont représentées sur la Figure **??**. On met en place un test de Kolmogorov pour tester

$$\mathcal{H}_0 : F = F_0 \text{ contre } \mathcal{H}_1 : F \neq F_0,$$

où F_0 est la fonction de répartition de la loi exponentielle $\mathcal{E}(1/1500)$, à l'aide de la fonction `ks.test`. La p-valeur valant 0.597, on ne rejette pas l'hypothèse nulle au niveau 5%.




```
Ampoule = c(673, 389, 1832, 570, 522, 2694, 3683, 644, 1531, 2916)
ks.test(Ampoule,pexp,1/1500,alternative="two.sided")
```

```
    One-sample Kolmogorov-Smirnov test

data:  Ampoule
D = 0.22843, p-value = 0.597
alternative hypothesis: two-sided
```

2.3 Tests de comparaison de deux échantillons

On considère deux échantillons indépendants

- $X_1, \dots, X_n$ i.i.d. de fonction de répartition F
- $Y_1, \dots, Y_m$ i.i.d. de fonction de répartition G.

On note $N = n + m$.

Dans le cas de deux échantillons gaussiens (F correspond à une loi normale $\mathcal{N}(m_0, \sigma^2)$ et G à la loi $\mathcal{N}(m_1, \sigma^2)$), on peut utiliser un test de Student pour tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \neq G$. Nous nous plaçons ici dans un cadre non paramétrique, les lois des variables X_i et Y_j ne sont pas supposées connues.

2.3.1 Tests de Kolmogorov-Smirnov

Dans cette section, on souhaite tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \neq G$. On note $\hat{F}_n$ la fonction de répartition empirique de l'échantillon $(X_1, \dots, X_n)$ et $\hat{G}_m$ celle de l'échantillon $(Y_1, \dots, Y_m)$.

Definition 2.4. Le **test de Kolmogorov-Smirnov** est défini par la statistique de test

$$D_{n,m} = \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \hat{G}_m(t)|.$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{D_{n,m} \geq d_{n,m,1-\alpha}\}$.

Proposition 2.6. *Si F est continue, la loi de $D_{n,m}$ sous l'hypothèse nulle $F = G$ est indépendante de F. Cette loi est tabulée.*

Remark. Pour faire un test unilatéral ($\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \geq G$), on utilise la statistique de test

$$D^+_{n,m} = \sup_{t \in \mathbb{R}} (\hat{F}_n(t) - \hat{G}_m(t)).$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{D^+_{n,m} \geq d^+_{n,m,1-\alpha}\}$.

Example 2.3. On souhaite comparer deux médicaments pour soulager la douleur post-opératoire. On a observé 16 patients, dont 8 ont pris le médicament A habituel, et les 8 autres un médicament B expérimental. Dans le tableau suivant sont reportés les temps (en heures) entre la prise du médicament et la sensation de soulagement.

médicament A	médicament B
6.8	4.4
3.1	2.5
5.8	2.8
4.5	2.1
3.3	6.6
4.7	1.5
4.2	4.8
4.9	2.3

Les fonctions de répartition empiriques des deux échantillons sont représentées en Figure 2.3.

Si on veut tester une différence d'efficacité entre les deux médicaments

$$\mathcal{H}_0 : F_A = F_B \text{ contre } \mathcal{H}_1 : F_B \neq F_A$$

```
mA = c(6.8,3.1,5.8,4.5,3.3,4.7,4.2,4.9)
mB = c(4.4,2.5,2.8,2.1,6.6,1.5,4.8,2.3)
ks.test(mB, mA, alternative="two.sided")
```

```
    Two-sample Kolmogorov-Smirnov test

data:  mB and mA
D = 0.625, p-value = 0.08702
alternative hypothesis: two-sided
```

Si on veut tester si le médicament B est plus efficace que le médicament A :

$$\mathcal{H}_0 : F_A = F_B \text{ contre } \mathcal{H}_1 : F_B \geq F_A$$

```
ks.test(mB, mA, alternative="greater")
```

```
    Two-sample Kolmogorov-Smirnov test

data:  mB and mA
D^+ = 0.625, p-value = 0.04394
alternative hypothesis: the CDF of x lies above that of y
```

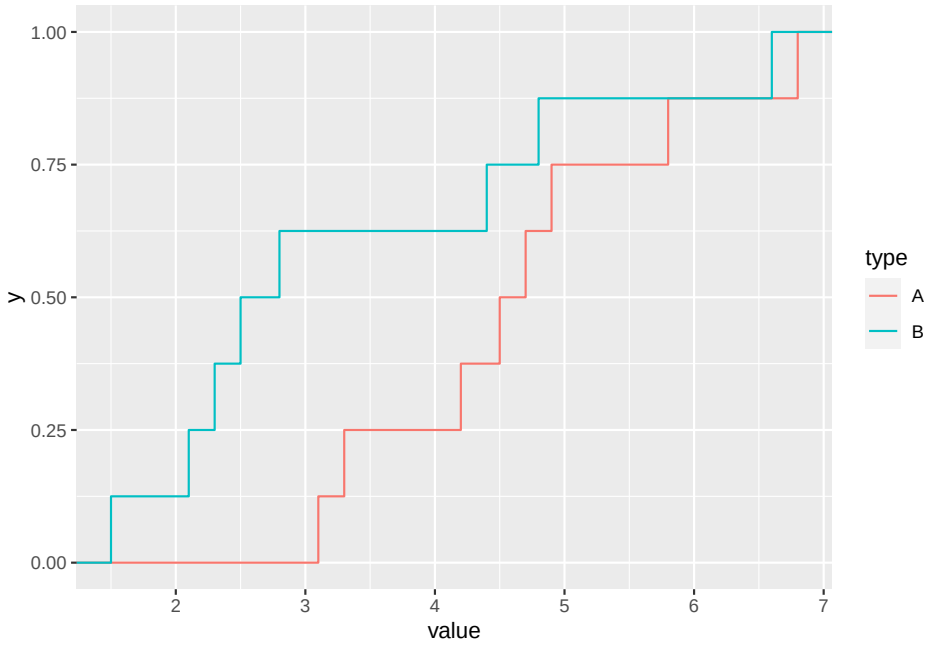


Figure 2.3: Fonction de répartition empirique pour le médicament A en rouge et le médicament B en bleu.

2.3.2 Test de Wilcoxon- Mann-Whitney

On va s'intéresser dans cette section au test de Mann-Whitney et celui de Wilcoxon (qui sont en fait équivalents) basés sur les rangs. Pour simplifier la présentation, nous allons supposer dans un premier temps qu'il n'y a pas d'ex-aequo dans les deux échantillons :

- les X_i sont tous distincts
- les Y_j sont tous distincts
- les X_i sont distincts des Y_j ($\forall i \neq j, X_i \neq Y_j$).

On reviendra sur le cas des ex-aequo en section 2.3.2.3.

2.3.2.1 Test de Mann-Whitney

Supposons que l'on souhaite tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \geq G$. On suppose que F et G sont continues.

Le principe du test de Mann-Whitney consiste à déterminer le nombre de couples (X_i, Y_j) pour lesquels $Y_j > X_i$. Sous $\mathcal{H}_1$, pour tout t, $\mathbb{P}(Y \leq t) \leq P(X \leq t)$ (avec parfois l'inégalité stricte), par conséquent pour tout t, $\mathbb{P}(Y > t) \geq P(X > t)$ et le nombre de couples (X_i, Y_j) pour lesquels $Y_j > X_i$ prend des valeurs plus grandes sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$.

Proposition 2.7. *On appelle* ***test de Mann-Whitney*** *pour $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \geq G$ le test défini à partir de la statistique*

$$MW_{X<Y} = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{1}_{X_i<Y_j}.$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{MW_{X<Y} \geq u_{(n,m),1-\alpha}\}$.

Remark. La loi de $MW_{X<Y}$ sous $\mathcal{H}_0$ peut être établie par récurrence (cf Caperaa and van Cutsem (1988), p 126). On note

$$p_{n,m}(k) = \mathbb{P}_{\mathcal{H}_0}(MW_{X<Y} = k) \text{ pour } k = 0, 1, \dots, mn$$

$$p_{n,0}(k) = p_{0,m}(k) = 1 \text{ pour } k = 0; = 0 \text{ pour } k \neq 0.$$

Alors pour tout k,

$$(n+m)p_{n,m}(k) = np_{n-1,m}(k) + mp_{n,m-1}(k-n).$$

Cette formule de récurrence permet de calculer la loi de $MW_{X<Y}$ sous $\mathcal{H}_0$.

On peut aussi utiliser un résultat asymptotique.

Theorem 2.1 ((Hajek (1968)) (admis))**.** *Sous $\mathcal{H}_0$,*

$$\frac{MW_{X<Y} - \mathbb{E}_{\mathcal{H}_0}[MW_{X<Y}]}{\sqrt{Var_{\mathcal{H}_0}(MW_{X<Y})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ quand } n \to +\infty, n/(n+m) \to \lambda \in]0,1[.$$

On utilise ce résultat en pratique si $n, m \geq 8$. On a, sous l'hypothèse nulle $\mathcal{H}_0$ que

$$\mathbb{E}_{\mathcal{H}_0}[MW_{X<Y}] = \frac{mn}{2} \text{ et } \mathrm{Var}_{\mathcal{H}_0}(MW_{X<Y}) = mn\left(\frac{n+m+1}{12}\right).$$

On peut faire le raisonnement similaire pour tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \leq G$. Dans ce cas le nombre de couples (X_i, Y_j) pour lesquels $Y_j < X_i$ prend des valeurs plus grandes sous $\mathcal{H}_1$ que sous $\mathcal{H}_0$.

Proposition 2.8. *On appelle* ***test de Mann-Whitney*** *pour $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \leq G$ le test défini à partir de la statistique*

$$MW_{X>Y} = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{1}_{X_i > Y_j}.$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{MW_{X>Y} \geq \tilde{u}_{(n,m),1-\alpha}\}$.

La statistique $MW_{X>Y}$ vérifie des propriétés similaires à celles de $MW_{X<Y}$ vues précédemment.

Enfin dans le cas d'un test bilatéral de $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \neq G$, on combine les deux tests précédents :

Proposition 2.9. *On appelle* ***test de Mann-Whitney*** *pour $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \neq G$ le test défini à partir de la statistique*

$$MW_{X,Y} = max(MW_{X<Y}, MW_{X>Y}).$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{MW_{X,Y} \geq v_{(n,m),1-\alpha}\}$.

2.3.2.2 Test de Wilcoxon

Revenons au test de $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \geq G$. Il existe une autre forme équivalente du test de Mann-Whitney, appelée test de la somme des rangs de Wilcoxon.

Soit $Z = (Z_1, \ldots, Z_n, Z_{n+1}, \ldots, Z_N) = (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ l'échantillon complet. On définit $(R_1, \ldots, R_m)$ où R_j est le rang de Y_j dans l'échantillon complet ordonné $Z_{(.)}$:

$$R_j = \sum_{k=1}^{N} \mathbb{1}_{Z_k < Y_j} + 1.$$

Proposition 2.10. *La statistique de Wilcoxon consiste à calculer la somme des rangs des individus du deuxième échantillon :*

$$W_Y = \sum_{j=1}^{m} R_j.$$

Comme on a la relation

$$MW_{X<Y} = W_Y - \frac{m(m+1)}{2},$$

les deux statistiques conduisent au même test.

De façon similaire, on peut construire la statistique de test de Wilcoxon W_X (somme des rangs des X_i dans Z) liée à la statistique de test $MW_{X>Y}$ pour tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \leq G$.

On peut remarquer que

$$W_X + W_Y = \sum_{k=1}^{n+m} k = \frac{N(N+1)}{2}.$$

Example 2.4. On reprend l'exemple des médicaments. On veut tester si le médicament B est plus efficace que le A ($\mathcal{H}_0 : F_A = F_B$ contre $\mathcal{H}_1 : F_B \geq F_A$). On a alors l'échantillon complet ordonné observé

$$\begin{aligned} z_{(.)} &= (1.5, 2.1, 2.3, 2.5, 2.8, 3.1, 3.3, 4.2, 4.4, 4.5, 4.7, 4.8, 4.9, 5.8, 6.6, 6.8) \\ &= (mB_6, mB_4, mB_8, mB_2, mB_3, mA_2, mA_5, mA_7, mB_1, mA_4, mA_6, mB_7, mA_8, mA_3, mB_5, mA_1) \end{aligned}$$

Les rangs observés pour les valeurs de B valent donc

$$R_1 = 9, R_2 = 4, R_3 = 5, R_4 = 2, R_5 = 15, R_6 = 1, R_7 = 12, R_8 = 3$$

ce qui donne $W_B = 51$ et $W_A = (16 \times 17)/2 - 51 = 85$.

On a également que

$$\begin{aligned} MW_{B<A} &= \sum_{i=1}^{8}\sum_{j=1}^{8} \mathbb{1}_{B_i<A_j} \\ &= 5+5+5+6+6+7+7+8 = 49 = W_A - (8\times 9)/2 \end{aligned}$$

et

$$\begin{aligned} MW_{B>A} &= \sum_{i=1}^{8}\sum_{j=1}^{8} \mathbb{1}_{B_i>A_j} \\ &= 3+3+3+2+2+1+1+0 = 15 = W_B - (8\times 9)/2. \end{aligned}$$

```
wilcox.test(mB,mA,alternative="less")
```

```
    Wilcoxon rank sum exact test

data:  mB and mA
W = 15, p-value = 0.04149
alternative hypothesis: true location shift is less than 0
```

2.3.2.3 Traitement des ex-aequos

Nous avons supposé les lois continues, donc la probabilité d'avoir des ex-aequos est nulle. En pratique, soit parce que les lois ne sont pas continues, soit parce qu'on a des mesures arrondies, on peut avoir des ex-aequos. Dans ce cas, on peut considérer les statistiques de test de Mann-Whitney suivantes :

$$\tilde{MW}_{X<Y} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\mathbb{1}_{X_i<Y_j} + \frac{1}{2}\mathbb{1}_{X_i=Y_j}\right\}$$

et

$$\tilde{MW}_{X>Y} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\mathbb{1}_{X_i>Y_j} + \frac{1}{2}\mathbb{1}_{X_i=Y_j}\right\}$$

respectivement. On peut remarquer que $\tilde{MW}_{X<Y} + \tilde{MW}_{X>Y} = nm$.

Pour le test de Wilcoxon, on utilise les rangs moyens : le rang de tous les éléments d'un groupe d'ex-aequos est la moyenne des rangs des éléments du groupe. On corrige ainsi les R_j définis précédemment.

Example 2.5. On considère les valeurs observées suivantes pour les deux échantillons :

$$\underline{x} = (5, 3, 6, 8, 1, 6) \text{ avec } n = 6 \text{ et } \underline{y} = (5, 7, 9, 5, 2) \text{ avec } m = 5.$$

On obtient alors le tableau des valeurs ordonnées et rangs suivant :

$x_{(.)}$	1		3		5		6	6		8	
$y_{(.)}$		2		5		5			7		9
$\tilde{R}_i$	1		3		5		7.5	7.5		10	
$\tilde{R}_j$			2		5		5		9		11

Ainsi

$$\left(\tilde{MW}_{X<Y}\right)^{obs} = 1 + (2 + \frac{1}{2}) + (2 + \frac{1}{2}) + 5 + 6 = 17,$$

$$\left(\tilde{W}_Y\right)^{obs} = \sum_{j=1}^{5}\tilde{R}_j = 2 + 5 + 5 + 9 + 11 = 32$$

et on retrouve bien que $\left(\tilde{M}W_{X<Y}\right)^{obs} = \tilde{W}_Y^{obs} - \frac{5\times 6}{2}$.

2.3.3 Test de la médiane

On veut tester $\mathcal{H}_0$: $F = G$ contre $\mathcal{H}_1$: $F \geq G$ et on suppose que F et G sont continues. Le principe du test de la médiane consiste à déterminer le nombre de variables du deuxième échantillon qui sont supérieures à la médiane de l'ensemble des observations.

Definition 2.5. Le **test de la médiane** est défini à partir de la statistique

$$M_{X,Y} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{R_j > \frac{N+1}{2}}.$$

La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{M_{X,Y} \geq m_{n,m,1-\alpha}\}$.

Example 2.6. Test de localisation.

$X_1, \dots, X_n$ sont i.i.d. de fonction de répartition F et $Y_1, \dots, Y_m$ sont i.i.d. de fonction de répartition $G = F(.-\mu)$. Par exemple, on étudie la pression artérielle de patients soumis à un traitement contre l'hypertension (Y_j), et on les compare à des patients non traités (X_i). Supposons qu'après traitement, la loi de la pression artérielle est translatée de μ. Le traitement est efficace si $\mu < 0$, il est inefficace si $\mu = 0$.

Loi de $M_{X,Y}$ sous $\mathcal{H}_0$:

- Si N pair,

  $$\forall k \in \left\{\max(0, m - \frac{N}{2}), \dots, \min(m, \frac{N}{2})\right\}, \mathbb{P}_{\mathcal{H}_0}(mM_{X,Y} = k) = \frac{C_m^k C_{N-m}^{N/2-k}}{C_N^{N/2}}.$$

  Donc $nM_{X,Y}$ suit une loi hypergéométrique $\mathcal{H}(N, \frac{N}{2}, m)$.

On en déduit que $\mathbb{E}_{\mathcal{H}_0}[M_{X,Y}] = \frac{1}{2}$ et $\mathrm{Var}_{\mathcal{H}_0}(M_{X,Y}) = \frac{n}{4m(N-1)}$.

- Si N est impair,

  $$\forall k \in \left\{\max(0, m - \frac{N+1}{2}), \dots, \min(m, \frac{N-1}{2})\right\}, \mathbb{P}_{\mathcal{H}_0}(mM_{X,Y} = k) = \frac{C_m^k C_{N-m}^{\frac{N-1}{2}-k}}{C_N^{\frac{N-1}{2}}}.$$

  Donc $nM_{X,Y}$ suit une loi hypergéométrique $\mathcal{H}(N, \frac{N-1}{2}, m)$.

On en déduit que $\mathbb{E}_{\mathcal{H}_0}[M_{X,Y}] = \frac{N-1}{2N}$ et $\mathrm{Var}_{\mathcal{H}_0}(M_{X,Y}) = \frac{n(N+1)}{4mN^2}$.

La connaissance de la loi de $M_{X,Y}$ sous $\mathcal{H}_0$ permet de déterminer la région de rejet du test. Pour $n, m \geq 30$, on peut approximer la loi de $M_{X,Y}$ sous $\mathcal{H}_0$ par la loi $\mathcal{N}(\mathbb{E}_{\mathcal{H}_0}[M_{X,Y}], \mathrm{Var}_{\mathcal{H}_0}(M_{X,Y}))$.

2.4 Tests de normalité

Dans cette section, on considère un n-échantillon $(X_1, \dots, X_n)$ de fonction de répartition F. On note $\hat{F}_n$ la fonction de répartition empirique associée à cet échantillon.

Pour illustrer les différentes méthodes, nous allons considérer les trois échantillons de taille $n = 200$ suivants :

- *Ech1* : un échantillon simulé selon la loi $\mathcal{N}(2, 1)$
- *Ech2* : un échantillon simulé selon la loi uniforme sur l'intervalle $[2, 4]$
- *Ech3* : un échantillon simulé selon la loi de Cauchy de paramètre 1

```
n=200
Ech1=rnorm(n,2,1)
Ech2=runif(n,min=2,max=4)
Ech3=rcauchy(n)
```

2.4.1 Méthode graphique : droite de Henry

La méthode de la droite de Henry, aussi appelée "Normal Probability Plot" ou"Q-Q Plot", consiste à représenter les points $(X_{(i)}, \Phi^{-1} \circ \hat{F}_n(X_{(i)}))$, où $X_{(1)} \leq \dots \leq X_{(n)}$ est l'échantillon ordonné et Φ représente la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Notons que $\hat{F}_n(X_{(i)}) = i/n$. Sous l'hypothèse que les X_i sont i.i.d. de loi normale, les points $(X_{(i)}, \Phi^{-1} \circ \hat{F}_n(X_{(i)}))$ sont pratiquement alignés. Le Q-Q plot pour les trois échantillons simulés est donné en Figure 2.4.

2.4.2 Test de normalité de Kolmogorov-Smirnov (de Lilliefors)

On souhaite tester l'hypothèse $\mathcal{H}_0$: "les X_i suivent une loi normale", contre l'hypothèse $\mathcal{H}_1$: "les X_i ne suivent pas une loi normale". On note

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \; S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Definition 2.6. Le **test de normalité de Kolmogorov** est fondé sur la statistique de test

$$D\mathcal{N}_n = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \Phi(t; \bar{X}, S^2)|$$

où $\Phi(.; \bar{X}, S^2)$ est la fonction de répartition de la loi normale $\mathcal{N}(\bar{X}, S^2)$. La région de rejet au niveau α est de la forme $\mathcal{R}_\alpha = \{D\mathcal{N}_n > d_{n,1-\alpha}\}$.

Proposition 2.11. *Sous l'hypothèse $\mathcal{H}_0$, i.e les X_i suivent une loi normale $\mathcal{N}(\mu, \sigma^2)$, la loi de $D\mathcal{N}_n$ ne dépend pas des paramètres inconnus (μ, σ^2). Il*

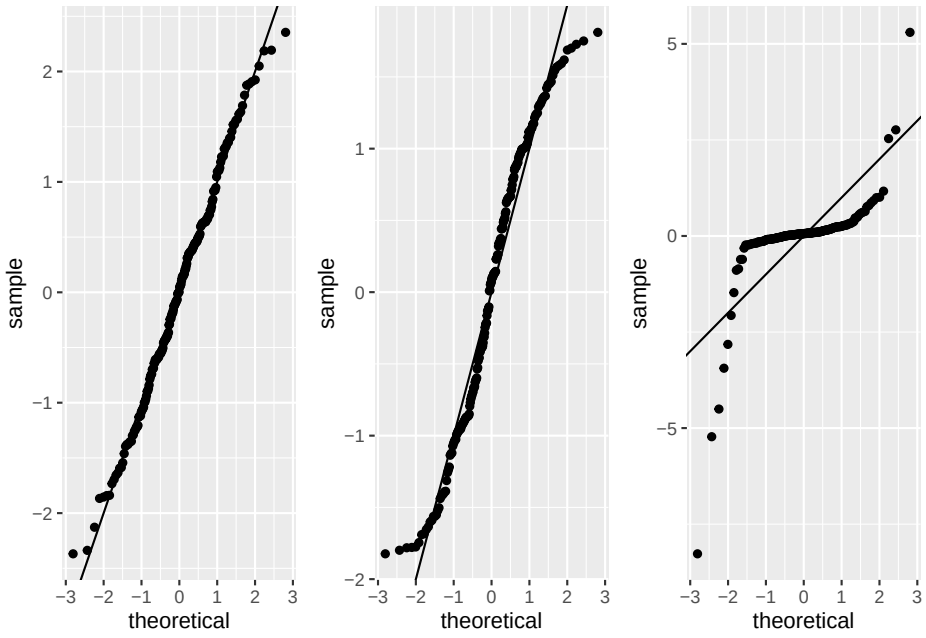


Figure 2.4: Q-Q plot pour les 3 échantillons (N(2,1) à gauche, U([2,4]) au centre, et C(1) à droite)

s'agit de la loi de

$$KS\mathcal{N}_n = \sup_{t\in\mathbb{R}} \left|\hat{\Phi}_n(t) - \Phi\left(\frac{t-\hat{\mu}_{\mathcal{Z}}}{S_{\mathcal{Z}}}\right)\right|$$

où $\mathcal{Z} = (Z_1, \dots, Z_n)$ *i.i.d de loi* $\mathcal{N}(0,1)$, $\hat{\Phi}_n$ *la fonction de répartition empirique de* $\mathcal{Z}$, $\hat{\mu}_{\mathcal{Z}}$ *la moyenne empirique et* $S^2_{\mathcal{Z}}$ *la variance empirique de* $\mathcal{Z}$.

La loi de $D\mathcal{N}_n$ est tabulée (on peut par exemple la simuler avec $\mu = 0$ et $\sigma^2 = 1$ pour en estimer les quantiles).

On applique le test de normalité de Kolmogorov-Smirnov sur les trois échantillons simulés :

```
library(nortest)
lillie.test(Ech1)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  Ech1
D = 0.041892, p-value = 0.532
```

```
lillie.test(Ech2)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  Ech2
D = 0.083624, p-value = 0.001691
```

```
lillie.test(Ech3)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  Ech3
D = 0.34735, p-value < 2.2e-16
```

2.4.3 Test de Shapiro-Wilk

Il s'agit d'un test basé sur les L-statistiques (combinaison linéaire des statistiques d'ordre), qui se base sur une comparaison de la variance empirique avec un estimateur de la variance des X_i qui a de bonnes propriétés sous l'hypothèse de normalité.

2.4.3.1 Estimation de la moyenne et de la variance à l'aide des statistiques d'ordre pour des lois symétriques

Soit $X_1, \dots, X_n$ i.i.d. On note $\mu = \mathbb{E}[X_i]$ et $\sigma^2 = \mathrm{Var}(X_i)$. La loi de $Y_i = (X_i - \mu)/\sigma$ est supposée symétrique (ce qui signifie que $-Y_i$ a même loi que Y_i). On note $(X_{(1)}, \dots, X_{(n)})$ l'échantillon des X_i ordonné : $X_{(1)} \leq \dots \leq X_{(n)}$. On note $(Y_{(1)}, \dots, Y_{(n)})$ l'échantillon des Y_i ordonné. On a

$$Y_{(i)} = (X_{(i)} - \mu)/\sigma.$$

Pour $i, j \in \{1, \dots, n\}$, on note $\alpha_i = \mathbb{E}[Y_{(i)}]$ et $B_{i,j} = \mathrm{Cov}(Y_{(i)}, Y_{(j)})$. On a alors

$$X_{(i)} = \mu + \alpha_i \sigma + \varepsilon_i,$$

avec $\mathbb{E}[\varepsilon_i] = 0$. Les ε_i ne sont pas indépendantes. La matrice de variance-covariance du vecteur $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ est $\sigma^2 B$. On note 1_n et α les vecteurs de $\mathbb{R}^n$ définis par

$$1_n = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_n \end{pmatrix}.$$

On note A la matrice de taille $(n, 2)$ définie par $A = (1_n, \alpha)$. Enfin, on note $X_{(.)} = (X_{(1)}, ..., X_{(n)})'$ et $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)'$. On a la relation

$$X_{(.)} = A \begin{pmatrix} \mu \\ \sigma \end{pmatrix} + \varepsilon.$$

L'estimateur des moindres carrés pondérés de (μ, σ) est obtenu en minimisant en les paramètres (μ, σ) le critère :

$$\left(X_{(.)} - A \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \right)' B^{-1} \left(X_{(.)} - A \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \right).$$

On obtient comme solution de ce système

$$\begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} = (A' B^{-1} A)^{-1} A' B^{-1} X_{(.)}.$$

(cf cours sur le modèle linéaire)

$$A' B^{-1} A = \begin{pmatrix} 1_n' B^{-1} 1_n & 1_n' B^{-1} \alpha \\ \alpha' B^{-1} 1_n & \alpha' B^{-1} \alpha \end{pmatrix}.$$

Lemma 2.1. *Lorsque la loi des Y_i est symétrique, $1_n' B^{-1} \alpha = 0$, la matrice $A' B^{-1} A$ est donc diagonale.*

Il en résulte que

$$\hat{\mu}_n = \frac{1_n' B^{-1} X_{(.)}}{1_n' B^{-1} 1_n}, \quad \hat{\sigma}_n = \frac{\alpha' B^{-1} X_{(.)}}{\alpha' B^{-1} \alpha}.$$

On peut montrer que, si la loi des Y_i n'est pas symétrique, $\hat{\sigma}_n$ sous-estime σ.

2.4.3.2 Procédure de test

Definition 2.7. Soit $Y_1, ... Y_n$ i.i.d. de loi $\mathcal{N}(0, 1)$ et $Y_{(1)} \leq ... \leq Y_{(n)}$ l'échantillon ordonné. Soit $\alpha = (\mathbb{E}[Y_{(1)}], ... \mathbb{E}[Y_{(n)}])'$. Soit B la matrice de covariance du vecteur $(Y_{(1)}, ..., Y_{(n)})$. Le **test de Shapiro-Wilk** pour tester l'hypothèse de normalité des X_i est basé sur la statistique de test :

$$SW_n = \frac{\hat{\sigma}_n^2 (\alpha' B^{-1} \alpha)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2 (\alpha' B^{-2} \alpha)}.$$

On peut l'écrire sous la forme

$$SW_n = \frac{\left(\sum_{i=1}^n a_i X_{(i)} \right)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2},$$

avec

$$(a_1, ..., a_n) = \frac{\alpha' B^{-1}}{(\alpha' B^{-1} B^{-1} \alpha)^{1/2}}.$$

La région de rejet est de la forme $\{SW_n \leq c_{n,\alpha}\}$.

Les a_i sont tabulés, ce qui permet de calculer facilement SW_n, les quantiles $c_{n,\alpha}$ sont également tabulés. On peut interpréter SW_n comme une mesure de corrélation (au carré) entre les données ordonnées et les statistiques d'ordre d'une loi normale.

Le test de Shapiro-Wilk est ici appliqué sur les trois échantillons simulés :

```
shapiro.test(Ech1)
```

```
    Shapiro-Wilk normality test

data:  Ech1
W = 0.99268, p-value = 0.4201
```

```
shapiro.test(Ech2)
```

```
    Shapiro-Wilk normality test

data:  Ech2
W = 0.95826, p-value = 1.289e-05
```

```
shapiro.test(Ech3)
```

```
    Shapiro-Wilk normality test

data:  Ech3
W = 0.45775, p-value < 2.2e-16
```

Chapter 3

Tests du khi-deux

Les slides associés à ce chapitre sont disponibles ici

La famille des tests du khi-deux regroupe des tests d'objectifs variés (ajustement, indépendance, homogénéité, ...) mais qui ont en commun de mesurer l'écart à l'hypothèse nulle via une "divergence du khi-deux" et dont les statistiques de test associées suivent asymptotiquement une loi du khi-deux. Les tests du khi-deux sont valables pour l'étude de données qualitatives (ou discrètes) à support fini. Cependant, en pratique, ces tests sont aussi appliqués à des données discrètes à support infini ou continues après regroupement en classes.

3.1 Test d'ajustement du khi-deux

3.1.1 Objectif et principe du test

Soit X une variable aléatoire qualitative ou quantitative discrète à $K > 1$ modalités $\{a_1, \dots, a_K\}$, de loi $\pi = (\pi_1, \dots, \pi_K)$ inconnue, où

$$\pi_k = \mathbb{P}(X = a_k) > 0, \ \forall k \in \{1, \dots, K\}.$$

On dispose d'un n-échantillon $(X_1, \dots X_n)$ de même loi que X. On se donne par ailleurs une loi de probabilité $\mathcal{L}_0$ sur $\{a_1, \dots, a_K\}$ connue caractérisée par $\mathbf{p}^0 = (p_1^0, \dots, p_K^0)$ tels que

$$\forall k, \ p_k^0 \in]0, 1[\text{ et } \sum_{k=1}^{K} p_k^0 = 1.$$

On souhaite tester : $\mathcal{H}_0 : X \sim \mathcal{L}^0$ contre $\mathcal{H}_1 : X$ ne suit pas la loi $\mathcal{L}^0$. On peut donc retraduire ces hypothèses de test par

$$\mathcal{H}_0 : \forall k, \ \pi_k = p_k^0 \text{ contre } \mathcal{H}_1 : \exists k, \ \pi_k \neq p_k^0.$$

Une idée naturelle consiste à estimer la loi de probabilité π de X à l'aide de l'échantillon $(X_1, \dots, X_n)$ et de comparer cet estimateur avec la loi $\mathbf{p}^0$. On va donc noter $N_k = \sum_{i=1}^n \mathbb{1}_{X_i = a_k}$ le nombre de fois que l'on obtient la valeur a_k dans l'échantillon et on estime π_k par $\hat{\pi}_k = \frac{N_k}{n}$. On considère alors la statistique

$$T_n = n \sum_{k=1}^K \frac{\left(\hat{\pi}_k - p_k^0\right)^2}{p_k^0} = \sum_{k=1}^K \frac{\left(N_k - np_k^0\right)^2}{np_k^0}.$$

Cette statistique est appelée divergence du khi-deux entre les lois $\hat{\pi}$ et $\mathbf{p}^0$. Elle mesure la "distance" entre les proportions observées et les proportions théoriques sous $\mathcal{H}_0$. Mais ce n'est pas une distance car il n'y a pas la propriété de symétrie.

3.1.2 Lien avec la loi multinomiale

Proposition 3.1. *La v.a. $N = (N_1, \dots, N_K)'$ suit une loi multinomiale $\mathcal{M}(n, \pi)$ sur $\mathbb{N}^K$, c'est-à-dire que pour tout $(n_1, \dots, n_K) \in \mathbb{N}^K$ on a*

$$\mathbb{P}(N_1 = n_1, \dots, N_K = n_K) = \begin{cases} \frac{n!}{n_1! \dots n_K!} \pi_1^{n_1} \dots \pi_K^{n_K} & \text{si } \sum_{k=1}^K n_k = n \\ 0 & \text{sinon.} \end{cases}$$

Ainsi les hypothèses du test peuvent se traduire par

$$\mathcal{H}_0 : N \sim \mathcal{M}(n, \mathbf{p}^0) \text{ contre } \mathcal{H}_1 : N \text{ ne suit pas } \mathcal{M}(n, \mathbf{p}^0).$$

Proposition 3.2. *Soit $\sqrt{\pi} = \left(\sqrt{\pi_1}, \dots, \sqrt{\pi_K}\right)'$. On a*

$$Y_n = \left(\frac{N_1 - n\pi_1}{\sqrt{n\pi_1}}, \dots, \frac{N_K - n\pi_K}{\sqrt{n\pi_K}} \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}_K(0, \Gamma),$$

où $\Gamma = I_K - (\sqrt{\pi})(\sqrt{\pi})'$ est la matrice de projection orthogonale sur Vect$(\sqrt{\pi})^\perp$.

Sous l'hypothèse que $X_1, \dots, X_n$ sont i.i.d. de loi $\pi = (\pi_1, \dots, \pi_K)$,

$$Z_n = \sum_{k=1}^K \frac{\left(N_k - n\pi_k\right)^2}{n\pi_k} \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2(K-1).$$

3.1.3 Procédure de test

On est maintenant en mesure de définir la procédure de test.

Proposition 3.3. *Le test d'ajustement du χ^2 consiste à rejeter l'hypothèse nulle $\pi = \mathbf{p}^0$ au niveau α si*

$$T_n = \sum_{k=1}^K \frac{\left(N_k - np_k^0\right)^2}{np_k^0} > x_{K-1,1-\alpha}$$

où $x_{K-1,1-\alpha}$ est le $1-\alpha$ quantile d'un χ^2 à $K-1$ degrés de liberté. Ce test est de niveau asymptotique α.

On considère en pratique que l'approximation de la loi de T_n sous l'hypothèse nulle par une loi du χ^2 à $K-1$ degrés de liberté est bonne dès lors que $np_k^0 \geq 5$ pour tout k. Lorsque ce n'est pas le cas, on regroupe des classes jusqu'à ce que ces conditions soient vérifiées. Mais lorsqu'on regroupe des modalités, la région de rejet change car la loi limite dépend du nombres de modalités.

Que peut-on dire de la puissance du test?

On peut remarquer que

$$\frac{T_n}{n} \geq \|N/n - \mathbf{p}^0\|^2 \xrightarrow{p.s} \|\pi - \mathbf{p}^0\|^2$$

par la loi des grands nombres et donc $T_n \xrightarrow{p.s} +\infty$. La puissance du test tend donc vers 1 quand n tend vers $+\infty$.

3.1.4 Exemple de Mendel

Chez les pois, le caractère couleur est codé par un gène présentant deux formes allèles J et v correspondant aux couleurs jaune et vert. Le jaune est dominant et le vert récessif. Le caractère de forme, rond ou ridé, est porté par un autre gène à deux allèles R (dominant) et r (récessif). On croise 2 populations (pures) de pois: l'une jaune et ronde, l'autre verte et ridée. Selon la prédiction de Mendel, au bout de 2 croisements, la proportion de pois

- [JR] jaunes et ronds est 9/16
- [Jr] jaunes et ridés est 3/16
- [vR] verts et ronds est 3/16
- [vr] verts et ridés est 1/16

Dans ses expériences, Mendel a obtenu les résultats suivants : $N_{JR} = 315$, $N_{Jr} = 101$, $N_{vR} = 108$, $N_{vr} = 32$. Ici $K = 4$ et l'on obtient que $(T_n)^{obs} = 0.47$ et $x_{3,0.95} = 7.82$. On accepte donc très largement l'hypothèse de Mendel.

```
Nk=c(315,101,108,32)
n = sum(Nk)
ptheo = c(9,3,3,1)/16
chisq.test(Nk,p=ptheo)
```

```
    Chi-squared test for given probabilities

data:  Nk
X-squared = 0.47002, df = 3, p-value = 0.9254
```

```
sum(((Nk - (n*ptheo))^2) / (n*ptheo))
```

```
[1] 0.470024
```

```
qchisq(0.95,3)
```

```
[1] 7.814728
```

3.2 Test du χ^2 d'adéquation à une famille de lois

3.2.1 Principe du test

Soit Θ un ouvert de $\mathbb{R}^d$ avec $1 \leq d < K$. Etant donnée une famille de lois de probabilités $(\mathcal{L}(\theta))_{\theta\in\Theta}$ définies sur $\{a_1, \dots, a_K\}$, on veut tester

$$\mathcal{H}_0 : \exists\theta \in \Theta,\ X \sim \mathcal{L}(\theta)$$

contre

$$\mathcal{H}_1 : \text{la loi de } X \text{ n'appartient pas à } (\mathcal{L}(\theta))_{\theta\in\Theta}.$$

Les lois $(\mathcal{L}(\theta))_{\theta\in\Theta}$ sont caractérisées par les vecteurs de probabilités sur $\{a_1, \dots, a_K\}$

$$\mathcal{P}(\Theta) = \{\mathbf{p}(\theta) = (p_1(\theta), \dots, p_K(\theta));\ \theta \in \Theta\}.$$

On souhaite donc tester

$$\mathcal{H}_0 : \pi \in \mathcal{P}(\Theta) \text{ contre } \mathcal{H}_1 : \pi \notin \mathcal{P}(\Theta).$$

L'idée du test est de remplacer $\mathbf{p}^0$ dans T_n par la loi de $\mathcal{P}(\Theta)$ "la plus proche" de π au vu des données, c'est-à-dire la loi $\mathbf{p}(\hat{\theta})$ où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance pour le paramètre θ basé sur l'échantillon $(X_1, \dots, X_n)$ sous $\mathcal{H}_0$. On considère donc la statistique suivante :

$$\hat{T}_n = n\sum_{k=1}^{K} \frac{\left(\hat{\pi}_k - p_k(\hat{\theta})\right)^2}{p_k(\hat{\theta})} = \sum_{k=1}^{K} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})}.$$

Or, on a le résultat (admis) suivant :

Theorem 3.1. *Supposons que*

- *Pour tout $k = 1, \dots, K$, $\theta \mapsto p_k(\theta)$ est $\mathcal{C}^2$ sur Θ et vérifie pour tout $\theta \in \Theta$, $p_k(\theta) \neq 0$.*
- *Pour tout $\theta \in \Theta$, les vecteurs $v_i = (\partial_i p_1(\theta), \dots, \partial_i p_K(\theta)))'$ pour $i = 1, \dots, d$ forment une famille libre de $\mathbb{R}^K$ (bonne paramétrisation).*
- *Pour tout θ, si $X_1, \dots, X_n$ sont i.i.d. de loi $\mathbf{p}(\theta)$ alors l'estimateur du maximum de vraisemblance $\hat{\theta}$ est consistant vers θ.*

Sous ces conditions, si $X_1, \dots, X_n$ sont i.i.d. de loi $\mathbf{p}(\theta)$ alors

$$\hat{T}_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2(K - d - 1).$$

On construit le test du χ^2 d'adéquation à $\mathcal{P}(\Theta)$ de la manière suivante:

on rejette l'hypothèse nulle si

$$\hat{T}_n = \sum_{k=1}^{K} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})} > x_{K-d-1,1-\alpha}.$$

La p-valeur vaut

$$p((\hat{T}_n)^{obs}) = \mathbb{P}_{\mathcal{H}_0}(\hat{T}_n \geq (\hat{T}_n)^{obs}) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(\chi^2(K-d-1) \geq (\hat{T}_n)^{obs}).$$

Sous l'alternative:

$$\frac{\hat{T}_n}{n} \geq \mathrm{d}^2(\mathrm{N/n}, \mathcal{P}(\,)) \xrightarrow{\text{p.s}} \mathrm{d}^2(\pi, \mathcal{P}(\,)),$$

et donc la puissance tend vers 1 dès que $\mathrm{d}^2(\pi, \mathcal{P}(\,)) > 0$.

Remark. Le nombre de degrés de liberté de la loi asymptotique est donné par "nombre de modalités - 1 - nombre de paramètres à estimer sous $\mathcal{H}_0$."

3.2.2 Exemple

Pour 10000 fratries de 4 enfants, on a relevé le nombre de garçons :

Nb de garçons (k)	0	1	2	3	4
Effectifs (N_k)	572	2329	3758	2632	709

On décide de modéliser les naissances en supposant qu'elles sont indépendantes et que la probabilité d'avoir un garçon vaut $\theta \in]0,1[$. On note X_i le nombre de garçons dans la $i$ème fratrie. On souhaite donc tester

$$\mathcal{H}_0 : X_i \sim \mathcal{B}in(4,\theta) \text{ contre } \mathcal{H}_1 : X_i \text{ ne suit pas une loi } \mathcal{B}in(4,\theta), \theta \in]0,1[.$$

Sous $\mathcal{H}_0$, l'estimateur du maximum de vraisemblance pour θ est donné par $\hat{\theta} = \frac{\overline{X}_n}{4}$. On peut alors calculer $\mathbf{p}(\hat{\theta}) = (p_0(\hat{\theta}), \dots, p_4(\hat{\theta}))$ avec $p_k(\hat{\theta}) = \mathbb{P}(U = k)$ pour $U \sim \mathcal{B}in(4,\hat{\theta})$. Sous $\mathcal{H}_0$, la statistique de test

$$\hat{T}_n = \sum_{k=0}^{4} \frac{\left(N_k - np_k(\hat{\theta})\right)^2}{np_k(\hat{\theta})} \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \chi^2(5-1-1) = \chi^2(3).$$

```
classes = c(0,1,2,3,4)
Nk = c(572,2329,3758,2632,709)
n = sum(Nk)
pihat = Nk / n
thetahat = sum(Nk * classes) / (n*4)
```

```
ptheo = dbinom(0:4,4,thetahat)
Tobs = sum(((Nk - (n*ptheo))^2) / (n*ptheo))
print(Tobs)
```

```
[1] 0.9882779
```

```
val = chisq.test(Nk,p=ptheo)  # Attention aux degrés de liberté !
print(val)
```

```
    Chi-squared test for given probabilities

data:  Nk
X-squared = 0.98828, df = 4, p-value = 0.9116
```

```
pval = 1 - pchisq(val$statistic,3)
print(pval)
```

```
X-squared
0.8040883
```

3.3 Test du χ^2 d'indépendance

3.3.1 Principe du test

Soient X et Y deux variables aléatoires admettant un nombre fini de modalités, $\{a_1, \dots, a_K\}$ et $\{b_1, \dots, b_L\}$ respectivement. On considère n couples aléatoires $(X_1, Y_1), \dots, (X_n, Y_n)$ indépendants et de même loi que (X, Y). On souhaite tester

$\mathcal{H}_0$: X et Y sont indépendantes contre $\mathcal{H}_1$: X et Y ne sont pas indépendantes.

On va ici donner une idée de la construction de la statistique de test. On rappelle tout d'abord que la loi du couple (X, Y) est caractérisée par les probabilités

$$\mathbb{P}(X = a_k, Y = b_l) \text{ pour tout } k = 1, \dots, K,\ l = 1, \dots, L.$$

On est sous $\mathcal{H}_0$ quand $\forall (k,l),\ \mathbb{P}(X = a_k, Y = b_l) = \mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$ et sous $\mathcal{H}_1$ si $\exists (k,l),\ \mathbb{P}(X = a_k, Y = b_l) \neq \mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$.

On introduit les variables aléatoires suivantes :

- $N_{k,l} = \sum_{i=1}^n \mathbb{1}_{X_i = a_k, Y_i = b_l}$
- $N_{k,.} = \sum_{i=1}^n \mathbb{1}_{X_i = a_k} = \sum_{l=1}^L N_{k,l}$
- $N_{.,l} = \sum_{i=1}^n \mathbb{1}_{Y_i = b_l} = \sum_{k=1}^K N_{k,l}$

On peut alors estimer $\mathbb{P}(X = a_k, Y = b_l)$ par $N_{k,l}/n$ et $\mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$ par $N_{k,.}N_{.,l}/n^2$. En reprenant le même raisonnement que dans les sections

précédentes, on obtient la statistique de test suivante :

$$I_n = n \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(\frac{N_{k,l}}{n} - \frac{N_{k,.}N_{.,l}}{n^2}\right)^2}{\frac{N_{k,.}N_{.,l}}{n^2}} = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,.}N_{.,l}}{n}\right)^2}{\frac{N_{k,.}N_{.,l}}{n}}.$$

3.3.2 Procédure de test

On suppose que $\forall k,\ \mathbb{P}(X = a_k) > 0$ et $\forall l,\ \mathbb{P}(Y = b_l) > 0$. Alors, sous $\mathcal{H}_0$,

$$I_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2((K-1)(L-1)).$$

Remark. Pour retrouver rapidement le nombre de degrés de liberté, il faut remarquer que le nombre de modalités du couple (X, Y) vaut KL. Et sous $\mathcal{H}_0$, comme $\forall (k,l),\ \mathbb{P}(X = a_k, Y = b_l) = \mathbb{P}(X = a_k)\mathbb{P}(Y = b_l)$, il faut estimer les $\mathbb{P}(X = a_k)$ pour $k = 1, \dots, K-1$ et les $\mathbb{P}(Y = b_l)$ pour $l = 1, \dots, L-1$. Ainsi le nombre de degrés de liberté est $(KL-1) - [(K-1)+(L-1)] = (K-1)(L-1)$.

Proposition 3.4. *Soit $\alpha \in]0,1[$. Le test de région de rejet*

$$\mathcal{R}_\alpha = \{I_n > x_{(K-1)(L-1),1-\alpha}\}$$

est un test de niveau asymptotique α pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$.

3.3.3 Exemple

Une enquête a été réalisée auprès d'un échantillon de 250 personnes au sujet de l'abaissement à 16 ans du droit de vote. Les réponses ont été classées suivant le niveau d'instruction des personnes interrogées :

Niveau d'instruction	Pour	Contre	$N_{k.}$
Brevet	10	15	25
Bac	20	85	105
Bac +2 et plus	20	100	120
$N_{.l}$	50	200	250

Peut-on affirmer, au risque d'erreur de 5%, qu'il existe une relation entre l'opinion d'une personne sur cette question et son niveau d'instruction ?

```
contingence = matrix(c(10,20,20,15,85,100),ncol=2)
chisq.test(contingence)
```

```
    Pearson's Chi-squared test
```

```
data:  contingence
X-squared = 7.1429, df = 2, p-value = 0.02812
```

3.4 Test d'homogénéité

3.4.1 Principe du test

Soit $E_1, \dots, E_L$ L-échantillons indépendants de lois discrètes sur le même support $\{a_1, \dots, a_K\}$. On note π_l la loi discrète de l'échantillon $E_l = (X_{l,1}, \dots, X_{l,n_l})$ de taille n_l. On veut tester

$$\mathcal{H}_0 : \text{ les échantillons sont issus de la même loi } (\pi_1 = \dots = \pi_L)$$

contre

$$\mathcal{H}_1 : \text{ les échantillons ne sont pas issus de la même loi } (\exists j \neq l,\ \pi_j \neq \pi_l)$$

3.4.2 Procédure de test

On note $N_{k,l} = \sum_{i=1}^{n_l} \mathbb{1}_{X_{l,i}=a_k}$, $N_{k,\cdot} = \sum_{l=1}^{L} N_{k,l}$ et $N_{\cdot,l} = \sum_{k=1}^{K} N_{k,l} = n_\ell$. On considère alors la statistique de test :

$$J_n = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(N_{k,l} - \frac{N_{k,\cdot} N_{\cdot,l}}{n}\right)^2}{\frac{N_{k,\cdot} N_{\cdot,l}}{n}}.$$

On suppose que $\forall k, \forall l, \pi_{l,k} = \mathbb{P}(X_{l,i} = a_k) > 0$. Alors, sous $\mathcal{H}_0$,

$$J_n \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2((K-1)(L-1)).$$

Proposition 3.5. *Soit $\alpha \in]0,1[$. Le test de région de rejet*

$$\mathcal{R}_\alpha = \{J_n > x_{(K-1)(L-1),1-\alpha}\}$$

est un test de niveau asymptotique α pour tester $\mathcal{H}_0$ contre $\mathcal{H}_1$.

3.4.3 Exemple

Dans cet exemple, on souhaite savoir si le taux de participation à un club sportif des élèves de deux collèges A et B est identique ou pas. On a donc deux échantillons $E_1 = (X_{1,1}, \dots, X_{1,n_1})$ et $E_2 = (X_{2,1}, \dots, X_{2,n_2})$ avec

$$X_{l,i} = \text{ participation du } i\text{ème élève du collège } l \in \{a_1, a_2\} = \{\text{"}oui\text{"}, \text{"}non\text{"}\}.$$

On veut tester

$$\mathcal{H}_0 : \text{ les 2 populations sont homogènes (même taux de participation)}$$

contre

$$\mathcal{H}_1 : \text{ les 2 populations ne sont pas homogènes.}$$

On a les effectifs observés suivants :

Partic. / Ech	collège A	collège B	$N_{k,.}$
oui	12	26	38
non	38	34	72
$N_{.,l}$	50	60	$n = 110$

et les effectifs théoriques sont :

Partic. / Ech	collège A	collège B
oui	17,27	20,73
non	32,73	39,27

La statistique de test observée vaut donc

$$(J_n)^{obs} = \frac{(12 - 17,27)^2}{17,27} + ... + \frac{(34 - 39,27)^2}{39,27} = 4,504 > x_{1,0.95} = 3.84$$

donc on rejette $\mathcal{H}_0$, le taux de participation à un club sportif est différent entre les deux collèges.

Bibliography

Caperaa, P. and van Cutsem, B. (1988). *Méthodes et modèles en statistique non paramétrique : exposé fondamental.* Dunod.