# Feuille de travaux pratiques - Scilab #5

Ce document a pour but de rappeler quelques éléments de base concernant l'implémentation des tests statistiques dans Scilab. On traite ainsi les cas des tests du $\chi^2$ d'adéquation et d'indépendance ainsi que le test de Kolmogorov–Smirnov. Les exercices à traiter en priorité sont indiqués en rouge.

## 1 Sur les tests du $\chi^2$

### 1.1 Test d'adéquation à une loi

Comme son nom l'indique, ce test a pour but de décider si un vecteur d'observations est ou non une réalisation d'un échantillon de variables aléatoires de loi prescrite. On observe ainsi une réalisation $(x_1, \dots, x_n)$ d'un vecteur aléatoire $(X_1, \dots, X_n)$ dont les entrées sont supposées indépendantes et identiquement distribuées à valeurs dans un ensemble fini $A = \{a_1, \dots, a_k\}$ et de loi inconnue $p = (p_1, \dots, p_k)$ où $p_j := \mathbb{P}_p(X_1 = a_j)$ pour $j \in \{1, \dots, k\}$.

On suppose par ailleurs donnée une loi a priori $p^0 = (p_1^0, \dots, p_k^0)$. On souhaite tester l'hypothèse nulle $H_0 = \{p = p^0\}$ contre l'hypothèse alternative $H_1 = \{p \neq p^0\}$. Pour $j \in \{1, \dots, k\}$, on note $\widehat{p}_j := \frac{1}{n} \sum_{i=1}^n 1_{X_i = a_j}$ la fréquence empirique de $a_j$. L'idée qui est à la base du test est bien sûr que le vecteur $\widehat{p}$ est plus proche de $p^0$ sous l'hypothèse nulle $H_0$ que sous l'hypothèse alternative $H_1$. Afin de quantifier la "proximité", on utilise la pseudo-distance du $\chi^2$ :

$$D_n = D_n(p^0, \widehat{p}) := n \times \sum_{j=1}^k \frac{(\widehat{p}_j - p_j^0)^2}{p_j^0},$$

dont le comportement asymptotique est le suivant :

**Proposition 1** *Sous l'hypothèse $H_0$, la suite $D_n$ converge en loi vers $Z \sim \chi^2(k-1)$ lorsque $n$ tend vers l'infini. Sous l'hypothèse alternative $H_1$, la suite $D_n$ tend presque sûrement vers l'infini avec $n$.*

Étant donné un niveau $\alpha$ (par exemple $\alpha = 5\%$) et un réel $\eta_\alpha$ tel que $\mathbb{P}(Z > \eta_\alpha) = \alpha$, la zone de rejet $W_n = \{D_n > \eta_\alpha\}$ fournit alors un test de niveau asymptotique $\alpha$ de $H_0 = \{p = p^0\}$ contre $H_1 = \{p \neq p^0\}$. Le seuil $\eta_\alpha$ peut-être déterminé avec la fonction `cdfchi` de `Scilab`.

**Exercice 1** On suppose donnés une mesure de probabilité $p^0$ de support fini $A$, un vecteur de données $(x_i)_{1 \leq i \leq n} \in A^n$ et un seuil $0 < \alpha < 1$.

1. Écrire un programme qui calcule le vecteur des fréquences empiriques $\widehat{p}$ associé à $(x_i)_{1 \leq i \leq n}$.
2. Écrire ensuite un programme qui calcule la statistique $D_n(p^0, \widehat{p})$.
3. En déduire un programme qui prend en entrées $p$, $(x_i)_{1 \leq i \leq n}$ et $\alpha$ et qui en sortie donne le résultat du test du $\chi^2$ d'adéquation de niveau $\alpha$.

**Remarque 1** *En pratique, on considère que l'approximation en loi par* $\chi^2(k-1)$ *est valide sous* $H_0$ *si* $n \times \min_{1\leq j\leq k} p_j^0 \geq 5$. *Si cette condition n'est pas satisfaite, on peut regrouper les valeurs de* $a_j$ *pour lesquelles* $p_j^0$ *est trop faible et augmenter ainsi le minimum.*

**Exercice 1** En faisant appel deux cent fois consécutives à un générateur d'entiers pseudo aléatoires, avec un niveau de confiance de 99%, décider si le générateur fournit des données équiréparties dans les entiers de 0 à 9.

## 1.2 Test d'indépendance

Nous rappelons maintenant la procédure du test d'indépendance du $\chi^2$. La problématique est la suivante : on dispose d'un échantillon d'une loi à deux composantes $Z = (X, Y)$ et l'on souhaite déterminer si les variables $X$ et $Y$ sont indépendantes ou non. Considérons donc $n$ données $(z_1, \ldots, z_n) = ((x_1, y_1), \ldots, (x_n, y_n))$ dont on suppose qu'elles sont les réalisations indépendantes et identiquement distribuées de variables aléatoires $(Z_1, \ldots, Z_n) = ((X_1, Y_1), \ldots, (X_n, Y_n))$ à valeurs dans des ensembles finis :

$$X_i \in \{A_1, \ldots, A_k\}, \quad Y_i \in \{B_1, \ldots, B_\ell\}, \quad \forall\, 1 \leq i \leq n.$$

On note $p = (p_{jl}, 1 \leq j \leq k, 1 \leq l \leq \ell)$ la loi du couple $Z = (X, Y)$, c'est-à-dire :

$$p_{jl} = \mathbb{P}(Z = (A_j, B_l)) = \mathbb{P}(X = A_j, Y = B_l).$$

On introduit les fréquences empiriques

$$\widehat{p}_{jl} = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=A_j, Y_i=B_l}, \quad \widehat{q}_j = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=A_j}, \quad \widehat{r}_l = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i=B_l}.$$

L'asymptotique de la statistique de test

$$D_n := n \sum_{j,l} \frac{(\widehat{p}_{jl} - \widehat{q}_j \widehat{r}_l)^2}{\widehat{q}_j \widehat{r}_l}$$

est la suivante :

**Proposition 2** *Sous l'hypothèse* $H_0$, $D_n$ *converge en loi vers* $Z \sim \chi^2((k-1)(\ell-1))$ *lorsque* $n$ *tend vers l'infini. Sous l'hypothèse alternative* $H_1$, $D_n$ *tend presque sûrement vers l'infini avec* $n$.

À nouveau, étant donnés un niveau $\alpha$ et un réel $\eta_\alpha$ tel que $\mathbb{P}(Z \geq \eta_\alpha) = \alpha$, la zone de rejet $W_n = \{D_n > \eta_\alpha\}$ fournit un test de niveau asymptotique $\alpha$ de $H_0 = \{X \text{ et } Y \text{ indépendantes}\}$ contre $H_1 = \{X \text{ et } Y \text{ non indépendantes}\}$.

**Exercice 2** Supposons donnés un vecteur $(x_i, y_i)_{1\leq i\leq n}$ et un seuil $0 < \alpha < 1$.

1. Écrire un programme qui calcule les fréquences empiriques $\widehat{p}$, $\widehat{r}$ et $\widehat{r}$.
2. Écrire ensuite un programme qui calcule la statistique $D_n$.
3. En déduire un programme qui prend en entrées le vecteur $(x_i)_{1\leq i\leq n}$ et le seuil $\alpha$ et qui en sortie donne le résultat du test du $\chi^2$ d'indépendance de niveau $\alpha$.

**Exercice 2** On désire étudier la répartition des naissances suivant le type du jour dans la semaine (jours ouvrables ou week-end) et suivant le mode d'accouchement (naturel ou par césarienne). Les données proviennent du "National Vital Statistics Report" et concernent les naissances aux USA en 1997.

| Naissances | Naturelles | César. | Total |
|---|---|---|---|
| J.O. | 2331536 | 663540 | 2995076 |
| W.E. | 715085 | 135493 | 850578 |
| Total | 3046621 | 799033 | 3845654 |

| Naissances | Naturelles | César. | Total |
|---|---|---|---|
| J.O. | 60.6% | 17.3% | 77.9% |
| W.E. | 18.6% | 3.5% | 22.1% |
| Total | 79.2% | 20.8% | 100.0% |

Tester au niveau 0.1% = 0.001 l'hypothèse d'indépendance entre le type du jour de naissance (jour ouvrable ou week-end) et le mode d'accouchement (naturel ou césarienne).

# 2 Tests non paramétriques

Dans toute cette section $\mu$ désigne une mesure de probabilité sur $\mathbb{R}$ et $F$ la fonction de répartition associée. On considère $(X_1, \ldots, X_n)$ un $n-$échantillon de loi $\mu$ et on note $(X_{(1)}, \ldots, X_{(n)})$ la statistique d'ordre associée.

## 2.1 Fonction de répartition empirique

La fonction de répartition empirique $F_n$ associée à l'échantillon $(X_1, \ldots, X_n)$ est définie pour tout $x \in \mathbb{R}$ par

$$F_n(x) := \frac{\#\{1 \leq k \leq n, X_k \leq x\}}{n},$$

ou encore

$$F_n(x) = \frac{k}{n} \text{ si } X_{(k)} \leq x < X_{(k+1)}.$$

Le théorème de Glivenko–Cantelli assure que $||F_n - F||_\infty$ tend presque sûrement vers zéro lorsque $n$ tend vers l'infini, d'autre part, le théorème de Kolmogorov-Smirnov assure que, si $F$ est continue, $\sqrt{n}\,||F_n - F||_\infty$ converge en loi lorsque $n$ tend vers l'infini vers une variable $K$ dont la loi est appelée loi de Kolmogorov :

$$\lim_{n\to+\infty} \mathbb{P}\left(\sqrt{n}\,||F_n - F||_\infty \leq x\right) = \mathbb{P}(K \leq x) = 1 - 2\sum_{k=1}^{+\infty}(-1)^{k-1}e^{-2k^2x^2}.$$

La fonction de répartition de la loi de Kolmogorov n'est pas préprogrammée dans le noyau de `Scilab`, mais elle est accessible par la commande `pks` de la boîte à outils `Stixbox`. On peut charger cette boîte en allant dans `Scilab -> Applications -> Gestionnaire de modules` puis en choisissant le répertoire `Analyse de données et statistiques-> Stixbox` et en cliquant sur `Installer`. Pour vérifier que la boîte est bien installée, copiez-collez, puis exécutez le code ci-dessous :

```
x=linspace(0,2);y=pks(x);
plot(x,y,'b');
xtitle('Fonction de répartition de la loi de Kolmogorov');
```

**Exercice 3** L'objet de cet exercice est d'illustrer le théorème de Glivenko–Cantelli et le théorème de Kolmogorov–Smirnov. Vous choisirez la fonction de répartition $F$ parmi les fonctions de répartitions continues déjà implémentées dans `Scilab`, voir `cdfnor` ou `cdfchi` par exemple.

1. Illustrer le fait que, pour tout $x \in \mathbb{R}$, la suite $(F_n(x))_n$ converge presque sûrement vers $F(x)$ lorsque $n$ tend vers l'infini, en représentant sur une même graphique, la fonction $F$ et plusieurs fonctions de répartition empiriques.
2. Montrer que

$$||F_n - F||_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \max_{1\le i\le n} \max\left( \left| \frac{i}{n} - F(X_{(i)}) \right|, \left| \frac{i-1}{n} - F(X_{(i)}) \right| \right).$$

3. Illustrer les théorèmes de Glivenko–Cantelli et de Kolmogorov–Smirnov.

## 2.2 Test d'adéquation de Kolmogorov–Smirnov

Grâce au théorème de Kolmogorov–Smirnov, on peut facilement mettre en oeuvre un test pour déterminer si un vecteur de données est ou non une réalisation d'un échantillon de loi prescrite. Si la loi en question a pour fonction de répartition $F$, on calcule la fonction de répartition empirique $F_n$ associée aux données ainsi que la statistique

$$D_n = \sqrt{n}\, ||F_n - F||_\infty .$$

À un seuil $0 < \alpha < 1$, on associe le nombre $c_\alpha$ tel que $\mathbb{P}(K \ge c_\alpha) = \alpha$. On a par exemple

| $\alpha$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|---|
| $c_\alpha$ | 1.22 | 1.36 | 1.48 | 1.63 | 1.73 | 1.95 |

La zone de rejet $W_n = \{D_n > c_\alpha\}$ fournit alors un test de niveau asymptotique $\alpha$ de l'hypothèse $H_0$ "les données sont des réalisation i.i.d. de loi de fonction de répartition $F$ contre l'hypothèse alternative $H_1$.

**Exercice 3** Au seuil de confiance 95%, décidez si les données `Z` téléchargeables ici sont des réalisations i.i.d. d'une variable exponentielle de paramètre 1.

## 2.3 Comparaison d'échantillon, test de Smirnov

Soient $(X_1, \dots, X_n)$ et $(Y_1, \dots, Y_m)$ deux échantillons et $F_n$ et $G_m$ les fonctions de répartitions empiriques associées. On cherche à tester l'hypothèse $H_0$ "ces deux échantillons proviennent d'une même loi continue" contre l'hypothèse alternative $H_1$. Pour cela, on considère la statistique

$$K_{n,m} := \sqrt{\frac{mn}{m+n}} \times \sup_{x\in\mathbb{R}} |F_n(x) - G_m(x)| .$$

Sous l'hypothèse $H_0$, $K_{m,n}$ converge en loi vers une variable de loi de Kolmogorov lorsque $m$ et $n$ tendent vers l'infini. La zone de rejet associée au test est donc du type $\{K_{m,n} \ge c_\alpha\}$ où $\mathbb{P}(K \ge c_\alpha) = \alpha$. Ce test est directement implémenté dans `Scilab` via le module `Stixbox` sous le nom de commande `kstwo`.