Tests statistiques d'hypothèse (2)

Frédéric Sur
Mines Nancy

11 mai 2026

On a vu dans le cours précédent des tests statistiques d'hypothèse dont la mise en œuvre nécessite que certaines propriétés sur la loi des échantillons soient satisfaites. Dans le présent cours, nous discutons plusieurs tests permettant d'étudier les échantillons : comparaison d'échantillons ou étude de leur distribution. De très nombreux tests sont disponibles dans la littérature, nous n'en évoquerons que quelques-uns pour donner un panel des possibilités qui s'offrent au statisticien. L'essentiel est d'avoir compris le principe de fonctionnement des tests, afin d'être capable de mettre en œuvre un test qui ne serait pas listé ici, mais qu'on pourrait trouver dans la littérature ou qui pourrait être proposé par une IA générative.

1 Tests de comparaison de deux échantillons

Dans cette section, la question est de savoir si deux échantillons peuvent être discernés d'un point de vue statistique. Autrement dit, on veut savoir si les deux échantillons ont été prélevés de manière indépendante dans la même population homogène.

Exemple

Les notes à l'examen du groupe A1 et du groupe A2 sont-elles comparables ? Ou alors la différence est-elle significative ?

1.1 Comparaison de deux échantillons gaussiens indépendants

On suppose disposer de deux échantillons gaussiens indépendants de moyennes m_1 et m_2, de variances σ_1^2 et σ_2^2, et de tailles n_1 et n_2, respectivement.

Les hypothèses sont :

— $\mathscr{H}_0 : m_1 = m_2$ et $\sigma_1 = \sigma_2$

— $\mathscr{H}_1 : m_1 \neq m_2$ ou $\sigma_1 \neq \sigma_2$

Pour une raison qu'on va voir, on commence par tester l'égalité des variances.

On sait que si S_1^2 et S_2^2 désignent les estimateurs non corrigés de la variance des deux échantillons, alors les statistiques $n_1 S_1^2/\sigma_1^2$ et $n_2 S_2^2/\sigma_2^2$ sont distribuées selon des lois du χ^2 à respectivement $n_1 - 1$ et $n_2 - 1$ degrés de liberté (d.d.l.) [1].

Introduisons la statistique de test :

$$F = \frac{n_1 S_1^2/((n_1-1)\sigma_1^2)}{n_2 S_2^2/((n_2-1)\sigma_2^2)}$$

Comme S_1 et S_2 sont indépendants (comme les deux échantillons), cette statistique suit une **loi de Fisher-Snedecor** à $(n_1 - 1, n_2 - 1)$ degrés de liberté. Voir l'annexe A pour la définition et la table de cette loi.

1. Si $(X_1, \ldots, X_n)$ est un n-échantillon de loi $\mathscr{N}(\mu, \sigma^2)$ et S^2 est la variance empirique (non corrigée), alors nS^2/σ^2 suit la loi du χ^2 à $n-1$ d.d.l. Voir l'exercice 3 de la séance 2 de la première partie du cours.

Sous l'hypothèse $\mathscr{H}_0$ ($\sigma_1 = \sigma_2$), F devient :

$$F = \frac{n_1 S_1^2/(n_1-1)}{n_2 S_2^2/(n_2-1)} = \frac{S_1^{*2}}{S_2^{*2}}$$

où S_1^* et S_2^* désignent les estimateurs corrigés de la variance. Si on met la plus grande des deux quantités au numérateur, le rapport est supérieur à 1, et proche de 1 lorsque $\sigma_1 = \sigma_2$.

On cherche ainsi une région critique de la forme $[k, +\infty[$ avec $k > 1$ tel que $P(F > k|\mathscr{H}_0) = \alpha$.

Si les variances sont jugées égales ($\sigma_1 = \sigma_2 = \sigma$, car l'hypothèse $\mathscr{H}_0$ n'est pas rejetée par le test précédent), on passe au test des espérances.

Comme la somme de deux variables indépendantes suivant une loi du χ^2 suit encore une loi du χ^2 dont le d.d.l. est la somme des d.d.l. des deux variables, $\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ suit une loi du χ^2 à $n_1 + n_2 - 2$ d.d.l.

Par ailleurs, les moyennes empiriques $\overline{X_1}$ et $\overline{X_2}$ sont distribuées selon des lois gaussiennes de moyennes m_1 et m_2 et de variances σ^2/n_1 et σ^2/n_2, respectivement. Ainsi, $\overline{X_1} - \overline{X_2}$ suit une loi gaussienne de moyenne $m_1 - m_2$ et de variance $\sigma^2(1/n_1 + 1/n_2)$ (car $\overline{X_1}$ et $\overline{X_2}$ sont indépendantes).

Ici, σ est inconnu. On introduit donc :

$$T = \frac{(\overline{X_1} - \overline{X_2} - (m_1 - m_2))/(\sigma\sqrt{1/n_1 + 1/n_2})}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)/(\sigma^2(n_1 + n_2 - 2))}} = \frac{\overline{X_1} - \overline{X_2} - (m_1 - m_2)}{\sqrt{(n_1 S_1^2 + n_2 S_2^2)(1/n_1 + 1/n_2)}}\sqrt{n_1 + n_2 - 2}$$

qui suit une loi de Student à $n_1 + n_2 - 2$ d.d.l. (c'est la conséquence de la définition de cette loi, voir la proposition 12 du chapitre 1 de la première partie du cours).

Sous $\mathscr{H}_0 : m_1 = m_2$, on fait un test bilatéral sur la statistique de test T.

Remarque 1 : robustesse du test de Student. Si n_1 et n_2 sont « grands » (supérieurs à 20, ou 30 selon les auteurs...), alors le test d'égalité des moyennes peut tout de même être utilisé si les échantillons ne sont pas gaussiens et même si $\sigma_1 \neq \sigma_2$.

Remarque 2 : dans l'exercice 3 de la séance précédente (données appariées), on ne peut pas utiliser ces tests car les deux populations ne sont pas indépendantes.

1.2 Comparaison de plusieurs échantillons décrits par une variable qualitative

Une variable aléatoire **qualitative** (par opposition aux variables aléatoires **quantitatives** vues jusqu'à présent) est une variable prenant un nombre fini de valeurs. Les valeurs possibles sont appelées les **modalités**.

On suppose disposer de k échantillons d'une variable aléatoire à r modalités dont on observe une réalisation. Si on note n_{ij} le nombre d'individus de l'échantillon i qui possèdent la modalité j, on peut représenter les données sous la forme d'un tableau :

	modalité 1	modalité 2	...	modalité r	total
échantillon 1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
échantillon 2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
$\vdots$					
échantillon k	n_{k1}	n_{k2}		n_{kr}	$n_{k.}$
total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

L'hypothèse que l'on teste est $\mathscr{H}_0$: « les échantillons proviennent de la même distribution »

Notons qu'il s'agit d'un **test non-paramétrique**.

Exemple

On se demande si trois traitements présentent la même efficacité. On forme trois groupes ($k = 3$

échantillons) de malades traités chacun avec un traitement. Chaque malade a, après traitement, des symptômes modérés ou faibles ($r = 2$ modalités). Les trois groupes sont-ils discernables d'un point de vue statistique ? (l'alternative étant qu'on ne peut pas exclure qu'ils sont distribués de la même manière, ce qui suggère que les traitements ont la même efficacité, ou la même inefficacité).

Sous $\mathscr{H}_0$, notons $p_1, p_2, \dots p_r$ les probabilités des r modalités (les mêmes probabilités pour tous les échantillons donc). Les p_j ne sont pas connus, on les estime par $n_{.j}/n$. Sous cette hypothèse, l'effectif moyen attendu pour l'échantillon i et modalité j est $n_{i.}p_{ij} \simeq n_{i.}n_{.j}/n$. On note m_{ij} cette dernière quantité.

On admet alors que sous $\mathscr{H}_0$,

$$d^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

est réalisation d'une variable aléatoire D^2 qui converge en loi lorsque $n \to +\infty$ vers une variable aléatoire de loi du χ^2 à $(k-1)(r-1)$ d.d.l. C'est une conséquence du théorème central limite dans sa version multivariée (ce qui explique pourquoi apparaît une loi du χ^2). Sous $\mathscr{H}_1$, D^2 tend vers l'infini presque sûrement.

Comme d^2 est petit sous $\mathscr{H}_0$ (et grand sous $\mathscr{H}_1$, cela veut dire que si les effectifs sont « grands », on peut rejeter $\mathscr{H}_0$ sous un test unilatéral à droite selon la loi du χ^2 à $(k-1)(r-1)$ d.d.l.

Il s'agit du **test d'homogénéité du χ^2**. Nous avons déjà vu et nous verrons d'autres tests du χ^2.

Remarque : ce test peut être appliqué à des échantillons décrits par une variable quantitative, en regroupant les valeurs. Par exemple, si X est une variable aléatoire mesurant la température, on peut former les modalités « froid », « tempéré », « chaud » en fixant des seuils de température.

2 Tests d'ajustement à une loi

Il s'agit à présent de savoir si un échantillon est distribué selon une loi donnée. On a vu par exemple que de nombreux tests supposent que l'échantillon est gaussien. La première étape, avant même la mise en œuvre du test, serait donc de vérifier cette hypothèse de gaussianité.

2.1 Méthodes qualitatives empiriques

Avant de mettre en œuvre des tests statistiques, un certain nombre de méthodes empiriques permettent de vérifier que l'hypothèse sur la distribution des données est réaliste. C'est la première chose à faire dans toute étude statistique, d'autant plus sur un logiciel où ces informations sont facilement accessibles.

On peut commencer par tracer l'histogramme des données, et voir par exemple si elles ont la forme de la distribution attendue, comme une « courbe en cloche » pour une distribution gaussienne. Ensuite, on calcule des moments empiriques. Par exemple, pour une loi de Poisson on sait qu'espérance (moment d'ordre 1) et variance (moment d'ordre 2 centré) doivent être égales, et pour une loi gaussienne le coefficient d'asymétrie (*skewness*, moment d'ordre 3 centré réduit) est nul et le coefficient d'aplatissement (*kurtosis*, moment d'ordre 4 centré réduit) est 3. Il existe également des ajustements graphiques que l'on ne détaillera pas ici : une transformation des observations permet de les représenter dans un domaine où elles sont réparties sur une droite. Par exemple, les logiciels fournissent généralement un *QQ-plot* qui représentent les observations sur la *droite de Henry* si elles sont bien distribuées de manière gaussienne.

On aura compris que ces méthodes donnent des indications mais ne permettent pas de trancher dans un sens ou dans l'autre, de par leur nature qualitative.

2.2 Tests statistiques

Parmi les nombreux **tests d'ajustement** de la littérature (on parle aussi de **tests d'adéquation** ou de **tests de conformité**), nous en discutons deux.

2.2.1 Test d'ajustement du χ^2

On considère une variable aléatoire X qualitative ou quantitative mais discrétisée comme dans la section 1.2 : X prend des valeurs dans un ensemble fini de taille k, chaque valeur ayant une probabilité p_j ($1 \leq j \leq k$).

On dispose d'observations : n_j dans la classe j ($1 \leqslant j \leqslant k$). La taille de l'échantillon est $n = \sum_{j=1}^{k} n_j$. Si les observations sont la réalisation d'un n-échantillon de X (hypothèse $\mathscr{H}_0$), l'effectif moyen attendu dans chaque classe j est $m_j = np_j$

On calcule alors :

$$d^2 = \sum_{j=1}^{k} \frac{(n_j - m_j)^2}{m_j}$$

On admet que, sous $\mathscr{H}_0$, d^2 est réalisation d'une variable aléatoire D^2 qui converge en loi vers une variable aléatoire de loi du χ^2 à $k-1$ degrés de liberté, et sous $\mathscr{H}_1$ (le n-échantillon n'est pas distribué comme X), D^2 converge vers l'infini p.s.

Il s'agit du **test d'ajustement du χ^2**. Attention, il ne doit pas être confondu avec le test d'homogénéité du χ^2 (décrit en section 1.2) malgré les similitudes.

Comme d^2 est petit sous $\mathscr{H}_0$, on peut rejeter $\mathscr{H}_0$ à l'aide d'un test unilatéral à droite selon cette loi, si n est « suffisamment grand ». En pratique, on demande à ce que np_i soit supérieur à 5 (cela dépend des auteurs). Si ce n'est pas le cas, on peut regrouper certaines des modalités de X.

Lorsque les probabilités p_j sont issues d'une loi paramétrique dont le (ou les) paramètre(s) est estimé sur les observations (par exemple : on veut tester l'ajustement à une variable gaussienne discrétisée, il faut alors estimer moyenne et variance), il faut diminuer le nombre de degrés de libertés et considérer, si p est le nombre de paramètres estimés, $k-1-p$ d.d.l.

2.2.2 Test d'ajustement de Kolmogorov-Smirnov

Soit X une variable aléatoire de fonction de répartition F et $(X_1, \dots X_n)$ un n-échantillon distribué comme X.

La fonction de répartition empirique de cet échantillon est définie pour tout $x \in \mathbb{R}$ par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant x}$$

où $\mathbf{1}_A$ désigne l'indicatrice de l'événement A.

$F_n(x)$ désigne donc la proportion d'éléments du n-échantillon qui sont inférieurs à x. Il s'agit d'une variable aléatoire car $F_n(x)$ dépend du n-échantillon.

On admet la convergence suivante :

$$P\left(\sup_x |F_n(x) - F(x)| > \frac{c}{\sqrt{n}}\right) \to \alpha(c)$$

pour tout $c > 0$, où $\alpha(c)$ est une fonction que l'on sait calculer explicitement (la formule est un peu compliquée, on ne la précise pas ici), qui est décroissante par rapport à c. Le point remarquable est que $\alpha(c)$ ne dépend pas de F (et donc de la distribution de X). Voir aussi la **distance de Kolmogorov-Smirnov**, définition 28 du chapitre 2 de la première partie du cours).

On en déduit donc un test statistique unilatéral à droite si n est « assez grand », basé sur la statistique $\sqrt{n}D_n$ où $D_n = \sup_x |F_n(x) - F(x)|$. Par exemple, $\alpha(c) = 0,05$ si $c = 1,358$: on rejette donc « le n-échantillon est distribué comme X » ($\mathscr{H}_0$) au risque 5% si $D_n > 1,358/\sqrt{n}$. Par ailleurs, $\alpha(c) = 0,01$ si $c = 1,629$, cette valeur est utile pour un test au risque de 1%.

Il s'agit du **test de Kolmogorov-Smirnov**. Il s'étend à la comparaison de deux fonctions empiriques. Il permet alors de faire un test d'homogénéité.

Remarque : attention, si la fonction de répartition F fait intervenir des paramètres estimés empiriquement sur l'échantillon, en toute rigueur le test n'est plus valable. Par exemple, si on suppose X de loi gaussienne dont les paramètres sont estimés sur l'échantillon, alors on utilise plutôt le test de Lilliefors ou le test de Shapiro-Wilk.

3 Tests d'indépendance

Dans de nombreuses situations, on est intéressé par savoir si deux variables aléatoires sont statistiquement indépendantes.

Exemple

Soit X une variable aléatoire modélisant la catégorie socio-professionnelle (selon 7 modalités) et Y l'orientation politique (selon 5 modalités). On fait un sondage qui fournit des effectifs pour chaque couple de catégorie socio-professionnelle et orientation politique. Ces données peuvent-elles permettre de rejeter ou accepter une hypothèse d'indépendance entre X et Y ?

Exemple

Soit X et Y modélisant le taux de deux hormones dans le sang. Après avoir recueilli des échantillons sur différentes personnes, on aimerait décider si X et Y sont indépendantes.

Nous donnons ici quelques tests utiles dans différentes situations (variables qualitatives ou quantitatives comme dans les exemples précédents).

On considère dans la suite deux variables X et Y, et l'hypothèse $\mathscr{H}_0$: « les deux variables sont indépendantes ».

3.1 Variables quantitatives

Les deux tests les plus courants sont des tests sur des coefficients de corrélation.

3.1.1 Cas gaussien : coefficient de corrélation de Pearson

Lorsque X et Y sont des variables gaussiennes, l'indépendance est équivalente à la nullité du **coefficient de corrélation de Pearson** (c'est celui que vous connaissez) entre X et Y.

Rappelons que ce coefficient de corrélation est :

$$\rho_P(X,Y) = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)\mathrm{var(Y)}}$$

À partir de deux n-échantillons $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_n)$, on calcule le coefficient de corrélation empirique :

$$R = \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2 \sum_{i=1}^n (Y_i-\overline{Y})^2}}$$

Il est possible d'exprimer la loi de R pour construire un test.

Néanmoins, il est plus simple d'effectuer la **transformée de Fisher** de R qui consiste à calculer :

$$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

Lorsque n est « grand » ($n > 25$), la variable aléatoire Z est distribuée selon une loi normale de moyenne $\frac{1}{2}\ln((1+\rho_P)/(1-\rho_P))$ et de variance $1/(n-3)$ qui ne dépend pas de ρ_P.

On construit ainsi un test sous l'hypothèse $\mathscr{H}_0 : \rho_P = 0$.

3.1.2 Cas non gaussien : coefficient de corrélation de Spearman

Lorsque les variables ne sont pas gaussiennes, le test précédent n'est pas utilisable. On introduit le **coefficient de corrélation de Spearman**.

On considère le n-échantillon $(U_1,\dots,U_n)$ établi en classant dans l'ordre croissant le n-échantillon $(X_1,\dots,X_n)$ de manière à ce que U_i est le rang de la variable X_i après classement (on suppose qu'il n'y a pas d'ex-aequo). Même chose pour $(V_1,\dots,V_n)$ avec $(Y_1,\dots,Y_n)$.

Le coefficient de corrélation de Spearman de X et Y est le coefficient de corrélation de Pearson des rangs U et V :

$$r_S(X,Y)=\frac{\text{cov}(U,V)}{\text{var}(U)\text{var(V)}}$$

L'intérêt de r_S est qu'il est invariant par transformation monotone croissante des variables (une telle transformation ne change pas les rangs).

Exemple

Sur une réalisation : si $(x_1,x_2,x_3)=(3,4;2,1;5,3)$, alors $(u_1,u_2,u_3)=(2,1,3)$.

D'une part : $\overline{U}=\overline{V}=(n+1)/2$, car les n-échantillons U_i et V_i sont des listes d'indices obtenues par permutation de $\{1,2,\dots,n\}$., et $\sum_{k=1}^n k=n(n+1)/2$.

D'autre part : $S_U^2=S_V^2=(n^2-1)/12$ après développement de $\sum_{k=1}^n(k-(n+1)/2))^2$

On en déduit l'expression du coefficient de corrélation de Spearman empirique :

$$\begin{aligned}R_S(X,Y) &= \frac{\sum_{i=1}^n(U_i-\overline{V})(V_i-\overline{V})}{\sqrt{\sum_{i=1}^n(U_i-\overline{U}^2\sum_{i=1}^n(V_i-\overline{V^2}}}\\ &= \frac{\sum_{i=1}^n U_iV_i/n-(n+1)^2/4}{(n^2-1)/12}\\ &= 1-\frac{6}{n(n^2-1)}\sum_{i=1}^n D_i^2\end{aligned}$$

avec $D_i=U_i-V_i$, une fois tous les calculs faits.

On voit (en réfléchissant un peu…) que :

— si $R_S=1$, alors pour tout i, $D_i=0$: les classements sont identiques (il y a une relation croissante entre X et Y) ;

— si $R_S=-1$, alors pour tout i, les classements sont l'inverse l'un de l'autre (il y a une relation décroissante entre X et Y) ;

— si $R_S=0$, alors pour tout i, $\sum D_i^2=n(n^2-1)/6$. C'est le cas qui nous intéresse.
Tout d'abord, $\sum D_i^2=\sum(U_i-V_i)^2=2\sum i^2-2\sum U_iV_i=n(n+1)(2n+1)/3-2\sum U_iV_i$ Maintenant, si U et V sont indépendants, $E(\sum U_iV_i)=\sum E(U_i)E(V_i)=n(n+1)^2/4$, et donc : $E(\sum D_i^2)=n(n^2-1)/6$ après calcul. Ainsi $R_S=0$ correspond au cas où les différences entre les rangs sont « statistiquement banales » sous hypothèse $\mathscr{H}_0$.

Pour les « grandes valeurs » de n ($n>30$), sous hypothèse d'indépendance de X et Y, R_S est distribué selon une loi normale de moyenne nulle et de variance $1/(n-1)$, ce qui permet de construire un test d'hypothèse.

3.2 Variables qualitatives

Soient deux variables qualitatives X et Y prenant respectivement q et r modalités notées respectivement x_i ($1\le i\le q$) et y_j ($1\le j\le r$). Les observations permettent de remplir le tableau à double-entrée (dit *tableau de contingence*) suivant :

	y_1	y_2	$\dots$	y_r	total
x_1	n_{11}	n_{12}		n_{1r}	$n_{1.}$
x_2	n_{21}	n_{22}		n_{2r}	$n_{2.}$
$\vdots$					
x_q	n_{q1}	n_{q2}		n_{qr}	$n_{q.}$
total	$n_{.1}$	$n_{.2}$		$n_{.r}$	n

Sous hypothèse d'indépendance, on a $p(X = x_i, Y = y_j) = p(X = x_i)p(Y = y_j)$. Sous cette hypothèse, l'effectif moyen attendu pour $(X = x_i, Y = y_j)$ est $m_{ij} = Np(X = x_i)p(Y = y_j) \simeq N(n_{i.}/N)(n.j/N) = n_{i.}n.j/N$.

On calcule alors :

$$d^2 = \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

qui a des petites valeurs si X et Y sont indépendants. On admet que, sous $\mathscr{H}_0$ (indépendance de X et Y), d^2 est réalisation d'une variable aléatoire D^2 qui converge en loi vers une variable aléatoire de loi du χ^2 à $(q-1)(r-1)$ degrés de liberté ; sous $\mathscr{H}_1$ (non-indépendance), D^2 tend vers l'infini p.s.

Cela nous permet de construire un test unilatéral à droite.

Il s'agit du **test d'indépendance du χ^2** (différent donc, des tests du χ^2 introduits en sections 1.2 et 2.2.1).

4 Limite des tests statistiques d'hypothèse à l'heure des données massives

Dans de nombreuses applications pratiques il est « facile » de disposer de grandes quantités de données, et par ailleurs les ressources de calcul disponibles de plus en plus facilement permettent de les traiter. On est dans le cas où la taille n des échantillons est très grande. Le problème est que dans ce cas, tout devient significatif. Aucune paire de variables aléatoires ne peut être déclarée comme indépendante. Par exemple, en sections 3.1.1 ou 3.1.2, la variance de Z ou R_S varie en $1/n$, il faut donc que la corrélation soit à un niveau irréalistement faible pour que l'indépendance ne soit pas rejetée. De la même manière, les ajustements sont toujours rejetés. En effet, le même argument sur la variance tient pour le test de Kolmogorov-Smirnov en section 2.2.2. Cette situation est logique : déclarer qu'un phénomène suit une certaine loi est toujours un choix de modélisation forcément imparfait : souvenez-vous que tous les modèles sont faux, mais certains sont utiles. Il est donc logique que disposer d'un grand nombre d'observations permette d'invalider ce modèle simplificateur, qui garde sans doute un intérêt par ailleurs. Le problème du p-hacking évoqué à la séance précédente est un autre écueil que l'on peut rencontrer lorsqu'un grand nombre de tests est possible sur ces données massives.

Pour cette raison, nous ne nous poserons pas de problèmes de significativité statistique dans le cours *Introduction à l'apprentissage automatique* en deuxième année. La théorie des tests statistiques d'hypothèse fait néanmoins partie de la « caisse à outils » de l'ingénieur, et doit avoir été comprise avant d'envisager d'aller plus loin. Par ailleurs, il existe de nombreux cas pratiques dans lesquels les échantillons ont une taille limitée (parce que les données sont difficiles ou onéreuses à acquérir) : la statistique classique reste alors très utile.

5 Exercices

5.1 Deux procédés de fabrication

Dans une usine, des tiges d'acier sont produites sur deux chaînes de production distinctes. La longueur des tiges est affectée d'une certaine variabilité.

On prélève six tiges produites sur la première chaîne, leur longueur est mesurée à :

$$12,8; 12,9; 10,9; 11,5; 11,7; 10,5$$

On prélève huit tiges produites sur la seconde chaîne, leur longueur est mesurée à :

$$12,7; 10,5; 11,3; 10,8; 10,7; 11,6; 11,8; 10,9$$

Y a-t-il une différence significative dans la production des deux chaînes? On supposera que la longueur des tiges est distribuée selon une loi normale.

Indication : sur la première chaîne, la moyenne des longueurs des tiges est $11,717$ et leur variance corrigée est de $0,9657$; sur la deuxième chaîne, la moyenne des longueurs des tiges est $11,288$ et leur variance corrigée est de $0,5241$.

5.2 Générateur aléatoire

Les générateurs aléatoires disponibles dans les langages de programmation sont en fait pseudo-aléatoires. Ils sont définis à l'aide d'une suite déterministe. Il est donc important de tester les propriétés qu'ils sont censés suivre.

On a généré 20 valeurs entre 0 et 1, présentées dans l'ordre croissant ici pour une meilleure lisibilité :
0,008 ; 0,012 ; 0,027 ; 0,087 ; 0,207 ; 0,327 ; 0,392 ; 0,396 ; 0,433 ; 0,437 ; 0,453 ; 0,475 ; 0,590 ; 0,643 ; 0,645 ; 0,681 ; 0,736 ; 0,737 ; 0,823 ; 0,928.

1) Ces valeurs sont-elles cohérentes avec le fait que le générateur est censé produire des nombres aléatoires répartis de manière uniforme sur $[0,1]$? Vous utiliserez un test de Kolmogorov-Smirnov et utiliserez le graphique ci-après (à compléter).

2) Même question avec un test du χ^2 d'adéquation en séparant les données en répartissant les données dans les intervalles $[0,1/2]$, $[1/2,3/4]$, $[3/4,1]$.

5.3 Snack ou non ?

Un exploitant de cinéma a compté si ses clients achetaient des snacks ou pas, selon le type de film qu'ils venaient voir. Le tableau suivant rassemble les résultats :

Type de Film	Snack	Pas de Snack	**Total Ligne**
Action	50	75	**125**
Comédie	125	175	**300**
Famile	90	30	**120**
Horreur	45	10	**55**
Total Colonne	**310**	**290**	**600**

Y a-t-il un lien entre l'achat de snacks et le type de film ?

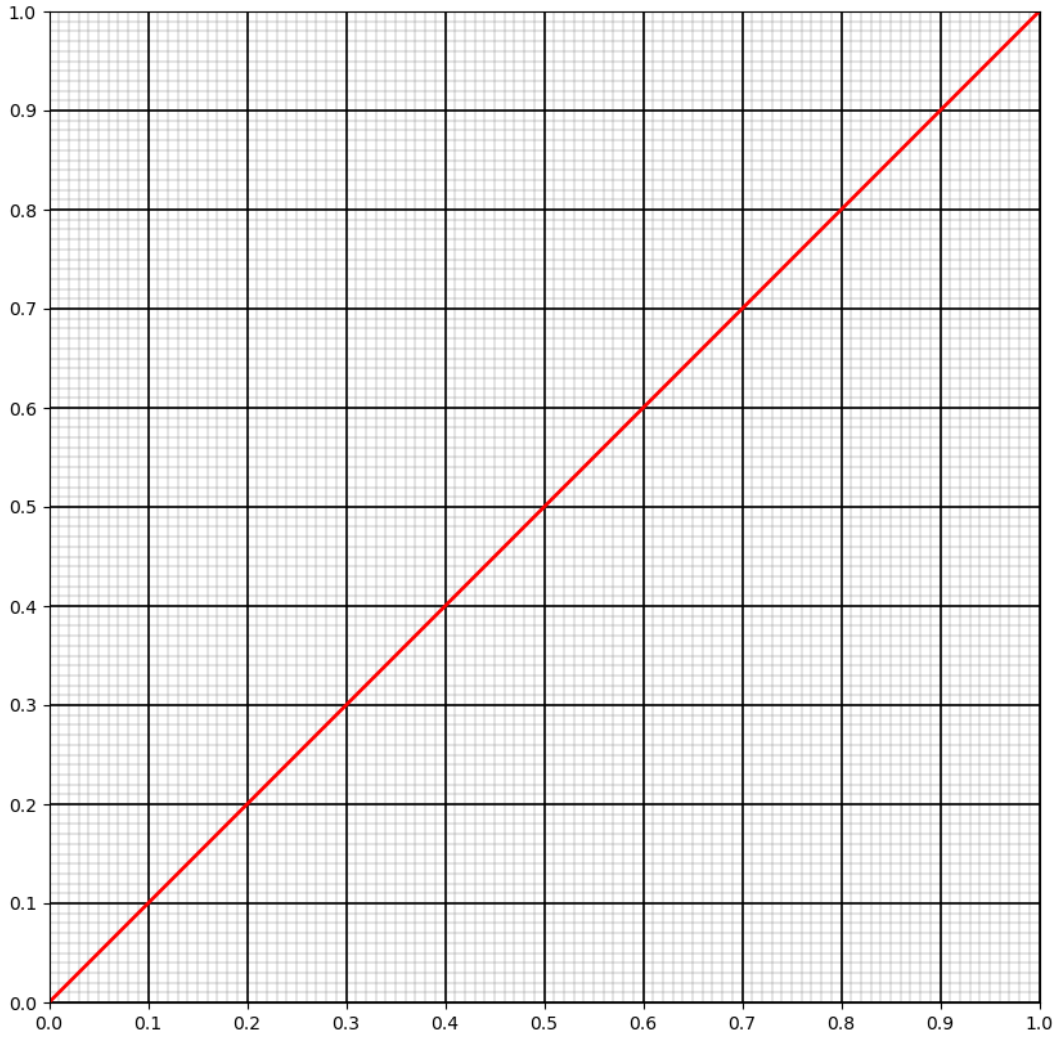

FIGURE 1 – Graphique pour l'exercice 2.

5.4 Test des signes (1)

Selon J. Hemerlijk[2] : *"The sign test is probably the oldest test in existence. It has been applied already in 1710 by John Arbuthnot"*.

On dispose d'un échantillon de paires de valeurs $(X_1,\ldots,X_n)$ et $(Y_1,\ldots,Y_n)$. On forme les différences $D_i = Y_i - X_i$ pour tout $1 \leqslant i \leqslant n$.

Le test des signes consiste à examiner le signe des différences D_i. L'idée est que s'il y a autant de signes + que de signes −, alors il n'y a pas de différence significative entre les X_i et les Y_i. Il s'agit d'un test d'appariement plus général que celui de l'exercice 3 de la séance précédente car il ne suppose pas que les D_i soient distribués normalement. Par ailleurs, il tient uniquement compte de la comparaison entre X_i et Y_i et pas de la *valeur* de la différence, ce qui est intéressant dans les situations où la valeur de la différence des variables aléatoires n'a pas de réelle signification physique.

2. J. Hemerlik. A theorem on the sign test when ties are present. Proceedings Nederl. Akademie van Wetenschappen Series A 55 (1952) p. 322.

Notons π la probabilité d'observer $D > 0$.

1) Soit S le nombre de valeurs D_i telles que $D_i > 0$ (signe +). Quelle est la loi de S ?

2) Proposez une hypothèse nulle $\mathscr{H}_0$ traduisant l'absence de différence significative entre X et Y, et déterminez la p-valeur du test bilatéral sur $\mathscr{H}_0$.

Remarque : pour simplifier, on suppose qu'il n'y pas de différences nulles. Les différences nulles sont discutées dans l'exercice suivant, et sont l'objet de l'article de Hemerlijk.

3) Appliquez le test aux données de l'exercice 3 de la séance précédente, rappelées ci-dessous :

Individu	1	2	3	4	5	6	7	8	9	10
Avant	61,2	66,3	80,4	70,5	73,0	78,1	57,9	71,4	74,2	75,3
Après	62,0	64,3	81,5	68,3	72,9	77,8	63,2	70,3	75,1	75,0

Indication : $\sum_{k=0}^{4} \binom{10}{k} = 386$.

5.5 Test des signes (2)

Une étude clinique évalue l'efficacité d'une application de méditation sur le niveau de stress perçu. On mesure le score de stress (sur une échelle de 0 à 100) chez $n = 50$ utilisateurs avant et après un mois de pratique.

Les résultats observés pour les différences $D_i = \text{Score}_{\text{après}} - \text{Score}_{\text{avant}}$ sont les suivants :

— 30 utilisateurs ont vu leur score baisser ($D_i < 0$).

— 10 utilisateurs ont vu leur score augmenter ($D_i > 0$).

— 10 utilisateurs n'ont observé aucun changement ($D_i = 0$).

On souhaite tester si l'application a un effet ou pas, à l'aide du test des signes.

1) Les valeurs nulles ($D_i = 0$) posent un problème. Il y a deux manière simple de traiter les zéros : méthode d'exclusion (suppression des observations pour lesquelles la différence est nulle) ou méthode de redistribution (répartition équitable entre les signes + et −) des zéros. Laquelle de ces deux approches a sans doute plus tendance à privilégier l'absence d'effet ? (on attend une justification qualitative basée sur le comportement attendu de la puissance du test)

2) En utilisant la méthode d'exclusion ($n' = 40$ utilisateurs), on souhaite tester $\mathscr{H}_0 : \pi = 0,5$ contre l'alternative bilatérale $\mathscr{H}_1 : \pi \neq 0,5$.

On sait que que si X suit une loi binomiale, alors $X' = (X - E(X))/\sigma_X$ converge en loi vers la loi normale centrée réduite.

1. Rappelez l'espérance $E(S)$ et la variance $\text{Var}(S)$ du nombre de signes positifs S sous $\mathscr{H}_0$.
2. Calculez la valeur de la statistique de test centrée réduite $Z = \frac{S - E(S)}{\sqrt{\text{Var}(S)}}$.

3) Au seuil de signification $\alpha = 5\%$, concluez quant à l'efficacité de l'application de méditation. Calculez également la p-valeur approximative associée à ce test. Comparez à la p-valeur que vous auriez obtenu en utilisant la méthode de redistribution.

A Loi de Fisher-Snedecor

Si U_1 et U_2 sont de variables aléatoires indépendantes suivant des lois du χ^2 à d_1 et d_2 degrés de liberté, alors

$$F = \frac{U_1/d_1}{U_2/d_2}$$

est une variable aléatoire qui suit une **loi de Fisher-Snedecor** à d_1 et d_2 degrés de liberté (on dit aussi loi de Fisher, ou loi de Snedecor). C'est une loi classique dont on connaît un certain nombre de propriétés. Voir par exemple `https://fr.wikipedia.org/wiki/Loi_de_Fisher`.

La table suivante donne, pour une variable aléatoire F distribuée selon une loi de *Fisher-Snedecor* (aussi appelée loi de Fisher) à d_1 et d_2 degrés de liberté, la valeur de F_{crit} tel que $P(X > F_{\text{crit}}) = 0,05$. Cette valeur est utile pour un test à droite au risque $\alpha = 5\%$ illustré par la figure ci-dessous.

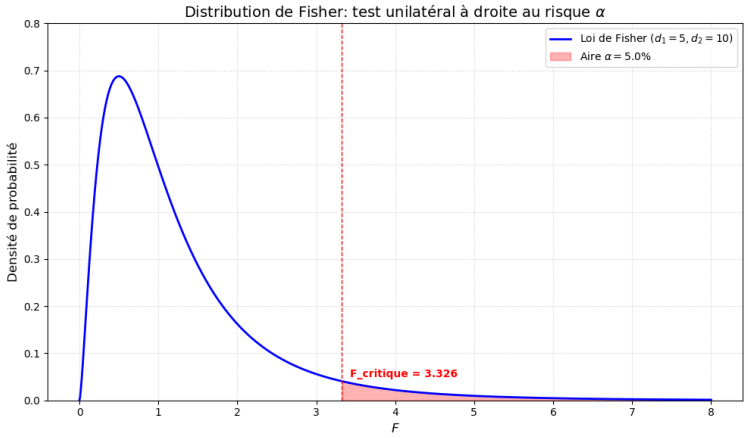

TABLE 1 – Table de la loi de Fisher-Snedecor ($\alpha = 0.05$)

$d_2 \backslash d_1$	1	2	3	4	5	6	7
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103

$d_2 \backslash d_1$	8	9	10	20	30	50	100
1	238.883	240.543	241.882	248.013	250.095	251.774	253.041
2	19.371	19.385	19.396	19.446	19.462	19.476	19.486
3	8.845	8.812	8.786	8.660	8.617	8.581	8.554
4	6.041	5.999	5.964	5.803	5.746	5.699	5.664
5	4.818	4.772	4.735	4.558	4.496	4.444	4.405
6	4.147	4.099	4.060	3.874	3.808	3.754	3.712
7	3.726	3.677	3.637	3.445	3.376	3.319	3.275
8	3.438	3.388	3.347	3.150	3.079	3.020	2.975
9	3.230	3.179	3.137	2.936	2.864	2.803	2.756
10	3.072	3.020	2.978	2.774	2.700	2.637	2.588
20	2.447	2.393	2.348	2.124	2.039	1.966	1.907
30	2.266	2.211	2.165	1.932	1.841	1.761	1.695
50	2.130	2.073	2.026	1.784	1.687	1.599	1.525
100	2.032	1.975	1.927	1.676	1.573	1.477	1.392