3.4. **Tests d'ajustement.** On dispose de la réalisation d'un échantillon $(X_1, ..., X_n)$ d'une v.a.r. X et on souhaite déterminer la loi de X. La première étape consiste à "deviner" une loi possible pour X, en regardant l'histogramme des fréquences constitué par la réalisation de notre échantillon par exemple. On construit alors un test pour savoir si X suit ou non la loi que l'on a devinée, mettons $\mathcal{L}$; autrement dit, on pose

$$H_0 \ : \ [X \text{ suit la loi } \mathcal{L}], \quad H_1 \ : \ [X \text{ ne suit pas la loi } \mathcal{L}].$$

3.4.1. *Test d'ajustement du chi-deux.*

Soit $X(\Omega)$ l'ensemble des valeurs prises par X sous l'hypothèse H_0. On choisit une partition $C_1, ..., C_J$ de $X(\Omega)$, chaque C_j, pour $j \in \{1, ..., J\}$, étant appelé une *classe*. On définit alors les variables aléatoires $N_1, ..., N_J$, *effectifs empiriques des classes*, comme les nombres de v.a. de l'échantillon appartenant aux classes C_1,...,C_J respectivement. On peut donc calculer ces effectifs par les formules : pour tout $j \in \{1, ..., J\}$,

$$N_j = \sum_{i=1}^{n} \mathbf{1}_{C_j}(X_i)$$

ou bien, de façon équivalente,

$$N_j = \text{Card } \{i \in \{1, ..., n\} / \ X_i \in C_j\}.$$

On note n_j les réalisations des observées sur l'échantillon des N_j.

On note $p_j = \mathbf{P}[X \in C_j]$ pour tout $j \in \{1, ..., J\}$ la proportion *théorique* de résultat que l'on doit trouver dans la classe j. On appelle alors *effectif théorique de la classe* C_j la quantité np_j.

Remarque. Le J-uplet $(N_1, ..., N_J)$ suit une loi multinomiale de paramètres $(n, p_1, ..., p_J)$. En particulier, chaque N_j suit une loi binomiale de paramètres (n, p_j). L'effectif théorique de la classe j est l'effectif théorique *moyen* de la classe, soit $\mathbf{E}[N_j] = np_j$. □

On a alors le résultat suivant :

Théorème 3.2. Soit D^2 la variable aléatoire définie par

$$D^2 = \sum_{j=1}^{J} \frac{(N_j - np_j)^2}{np_j}.$$

Alors, lorsque la taille de l'échantillon n tend vers l'infini, D^2 converge en loi vers une variable du chi-deux à $J - 1$ degrés de liberté.

Le *test d'ajustement du chi-deux* est le suivant : on admet que D^2 suit approximativement une loi du Chi-deux à $J-1$ degrés de liberté. On fixe le risque de première espèce α petit. La région de rejet de H_0 est choisie de la forme $[D^2 > c]$ où c est à déterminer en fonction de α dans une table de la loi du Chi-deux à $J - 1$ degrés de liberté. Pour appliquer le test, il suffit alors de calculer la réalisation de la variable D^2 que l'on obtient avec nos données et de constater si elle se trouve ou non dans la région de rejet de H_0.

Remarque. L'approximation que l'on fait en supposant que D^2 suit une loi du Chi-deux à $J - 1$ degrés de liberté n'est admise que si les effectifs empiriques et théoriques des classes "ne sont pas trop petits". Le seuil fixé dépend largement des auteurs de traités de statistique. Nous demanderons que les réalisations des N_j et que les np_j soient supérieurs à 5. Dans le cas contraire, on modifiera les classes en les regroupant pour obtenir ces conditions. □

3.4.2. *Test d'ajustement du Chi-deux avec estimation de paramètres.* Très souvent, on veut pouvoir ne spécifier, dans l'hypothèse H_0, que la loi de X et non les paramètres de cette loi, que l'on ignore a priori et que l'on ne peut qu'estimer. Le test vise donc à choisir entre les hypothèses

$$H_0 : [X \rightsquigarrow \mathcal{L}(t)], \quad H_1 : \overline{H}_0,$$

où t est la réalisation observée d'un estimateur T du paramètre $\theta \in \mathbb{R}^p$ de la loi $\mathcal{L}$. On procède exactement commme ci-dessus, en utlisant cette fois-ci le théorème :

Théorème 3.3. Soit D^2 la variable aléatoire définie par

$$D^2 = \sum_{j=1}^{J} \frac{(N_j - np_j)^2}{np_j}.$$

Alors, lorsque la taille de l'échantillon n tend vers l'infini, D^2 converge en loi vers une variable du chi-deux à $J - 1 - p$ degrés de liberté (p est le nombre de paramètres estimés).

3.4.3. *Test d'ajustement de Kolmogorov-Smirnov.* Il s'agit d'un test d'ajustement à une loi entièrement spécifiée. On l'utilise uniquement pour faire des ajustements à des lois de fonction de répartition continue. Il repose sur la convergence de la fonction de répartition empirique d'un échantillon vers la fonction de répartition de la variable parente vue au chapitre 1.

Théorème 3.4. Sous les hypothèses du théorème 1.1, la vitesse de convergence de F_n^* vers F_X est précisée par

$$\lim_{n \to +\infty} \mathbf{P}\left[\sqrt{n} \sup_{x \in \mathbb{R}} |F_n^*(x) - F_X(x)| \leq y\right] = K(y) = \sum_{k \in \mathbb{Z}} (-1)^k e^{-2k^2 y^2}.$$

Pour le *test d'ajustement de Kolmogorov-Smirnov*, on procède comme suit : on veut tester l'hypothèse selon laquelle X suit une loi $\mathcal{L}$ de fonction de répartition F. On teste toujours

$$H_0 \ : \ [X \text{ suit la loi } \mathcal{L}], \quad H_1 \ : \ [X \text{ ne suit pas la loi } \mathcal{L}].$$

On fixe un seuil de risque α petit. Sous l'hypothèse H_0, la variable aléatoire

$$\sqrt{n} D_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)|$$

admet approximativement comme fonction de répartion la fonction K. On cherche une région de rejet de la forme

$$[\sqrt{n} D_n > y]$$

en déterminant y en fonction de α dans une table. On procède ensuite au test en calculant la réalisation de $\sqrt{n} D_n$ obtenue avec nos données.

Remarque. Il existe d'autres tests d'ajustement. Citons simplement le test de Cramer-Von Mises qui s'applique sous les mêmes conditions que celui de Kolmogorov-Smirnov et qui repose également sur la convergence de la fonction de répartition empirique de l'échantillon vers la fonction de répartition de la variable parente.

3.5. **Tests de comparaison entre échantillons indépendants.** On dispose de m échantillons indépendants d'une certaine variable X et on désire savoir si ces échantillons proviennent d'une même population. On peut reformuler le problème de la façon suivante : chaque échantillon est une suite finie i.i.d. d'une variable parente X^k et le problème est de savoir si les X^k ont même loi.

3.5.1. *Tests paramétriques pour la comparaison de deux échantillons.* Lorsque l'on a deux échantillons indépendants ($m = 2$), on peut déjà essayer de savoir si les variables parentes ont même espérance et même variance.

On suppose que les variables parentes des deux échantillons, X^1 et X^2, admettent des moments d'ordre 2 et on note

$$m_k = \mathbf{E}\left[X^k\right] \text{ et } \sigma_k^2 = \mathbf{Var}\left[X^k\right] \text{ pour } k = 1, 2.$$

On note $(X_1^1, ..., X_{n_1}^1)$ l'échantillon de la v.a. X^1 et $(X_1^2, ..., X_{n_2}^2)$ celui de la v.a. X^2.

Cas des échantillons gaussiens. On suppose que X^1 et X^2 suivent des lois gaussiennes. On commence par comparer les variances et, si elles ne sont pas significativement différentes, on comparera les moyennes sous l'hypothèse que les variances sont égales.

Pour comparer les variances, on utilise le résultat suivant :

Théorème 3.5. Soient X et Y deux v.a. indépendantes suivant respectivement des loi du Chi-deux à n et p degrés de liberté. Alors

$$F = \frac{X/n}{Y/p}$$

suit une loi de Fisher-Snedecor à (n, p) degrés de liberté notée $F_{n,p}$.

On applique ce résultat en procédant au *test de Fisher-Snedecor* suivant : on pose

$$H_0 \ : \ [\sigma_1 = \sigma_2], \quad H_1 \ : \ [\sigma_1 \neq \sigma_2];$$

on choisit un seuil de risque α petit. Si on note S_1^2 et S_2^2 les variances empiriques (sans biais) des échantillons, et on définit F par

$$F = S_1^2/S_2^2 \text{ si } s_1^2/s_2^2 \geq 1 \text{ et } F = S_2^2/S_1^2 \text{ sinon,}$$

où s_k^2 est la réalisation de la variable S_k^2 pour $k = 1, 2$. On cherche alors la région de rejet sous la forme

$$W = \{(x_1^1, ..., x_{n_1}^1) \in \mathbb{R}^{n_1}, (x_1^2, ..., x_{n_2}^2) \in \mathbb{R}^{n_2} / \ f > f_\alpha\}$$

où f_α est à déterminer en fonction de α dans une table de la loi de Fisher-Snedecor, et f est la réalisation de F (i.e. $f = \max(s_1^2/s_2^2, s_2^2/s_1^2)$).

Si le test précédent n'a pas conduit à rejeter l'hypothèse $\sigma_1 = \sigma_2$, on procède au test de Student sur les moyennes comme suit : on suppose $\sigma = \sigma_1 = \sigma_2$ (inconnue) et on teste les hypothèses :

$$H_0 \ : \ [m_1 = m_2], \quad H_1 \ : \ [m_1 \neq m_2]$$

en choisissant un seuil de risque α petit.

On note $\overline{X}_k$ la moyenne empirique de l'échantillon k. Sous l'hypothèse H_0,

$$T = \frac{(\overline{X}_1 - \overline{X}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}\sqrt{1/n_1 + 1/n_2}} \rightsquigarrow T_{n_1+n_2-2}.$$

On choisit lors la région de rejet de la forme

$$W = \{(x_1^1, ..., x_{n_1}^1) \in \mathbb{R}^{n_1}, (x_1^2, ..., x_{n_2}^2) \in \mathbb{R}^{n_2} / \ |t| > t_{\alpha/2}\}$$

avec t réalisation de T.

Remarque. Lorsque $\sigma_1 \neq \sigma_2$ et que les échantillons sont suffisamment grands (quelques dizaines d'observations), on peut encore appliquer le test de Student.

Cas des échantillons non gaussiens. Dans ce cas, le test de Fisher-Snedecor ne peut plus s'appliquer, mais on peut encore appliquer le test de Student si les échantillons sont assez grands, que les variances soient égales ou non. Le test de Student est un test robuste.

3.5.2. *Test non paramétrique de comparaison de deux échantillons ou plus : le test du Chi-deux.* On dispose de m échantillons de v.a. X^1,...,X^m. Comme pour le test d'ajustement du Chi-deux, on partage en J classes l'ensemble des valeurs prises par ces variables aléatoires. Pour tout $k \in \{1, ..., m\}$ et pour tout $j \in \{1, ..., J\}$, on note N_{kj} le nombre de réalisations de l'échantillon k qui sont dans la classe C_j. On pose

$$N_{.j} = \sum_{k=1}^{m} N_{kj} \text{ l'effectif empirique de la classe } j \in \{1, ..., J\}.$$

On notera également

$$N_{k.} = \sum_{j=1}^{J} N_{kj} = n_k \text{ la taille de l'echantillon } k \in \{1, ..., m\}$$

et

$$N = \sum_{k=1}^{m} \sum_{j=1}^{J} N_{kj} = n \text{ le nombre total d'observations.}$$

On pose enfin

$$D_0^2 = \sum_{k=1}^{m} \sum_{j=1}^{J} \frac{(N_{kj} - N_{k.}N_{.j}/N)^2}{N_{k.}N_{.j}/N}.$$

On peut montrer que, sous l'hypothèse (H_0) que les échantillons proviennent d'une même population et sont indépendants, D_0^2 suit approximativement une loi du Chi-deux à $(m-1)(J-1)$ degrés de liberté. On procède alors comme dans le test d'ajustement du Chi-deux avec les hypothèses H_0 et $H_1 = \overline{H}_0$.

Sous l'hypothèse H_0 selon laquelle les générateurs sont identiques et indépendants, on sait que D_0^2 suit une loi du chi-deux à $3 \times 2 = 6$ degrés de liberté et on applique le test du chi-deux habituel avec un risque de première espèce α à choisir.

Remarque. Ce test s'applique également pour établir l'indépendance de deux variables aléatoires. En effet, supposons que l'on observe un échantillon $((X_1, Y_1), ..., (X_n, Y_n))$ d'un couple de v.a. (X, Y). L'ensemble des valeurs prises par X est partitionné en les classes C_i, $i = 1$ à I et celui des valeurs prises par Y en les classes C^j, $j = 1$ à J. Si on note $C_{ij} = C_i \times C^j$ alors

$$\{C_{ij},\ i = 1..I,\ j = 1..J\}$$

forme une partition de l'ensemble des valeurs prises par (X, Y). On note alors N_{ij} l'effectif de la classe C_{ij}.

On pose

$$H_0 : [X \text{ et } Y \text{ sont indépendantes}], \quad H_1 : \overline{H}_0.$$

Sous l'hypothèse H_0, la variable

$$D_0^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n} \rightsquigarrow \chi^2_{(I-1)(J-1)}$$

et on procède à un test du chi-deux.