Probabilités et Statistiques pour le S3SV. Notes de cours

Elisabeth Gassiat et Wendelin Werner

Université Paris-Sud

Table des matières

1 Introduction.

L'objectif du cours est de mettre en place des procédures statistiques permettant de répondre à des questions concernant des phénomènes comportant une part aléatoire. Typiquement, les questions auxquelles on proposera des réponses sont du type:

- Pour telle espèce, la composition du génome est-elle homogène dans les différentes bases ? Est-elle la même pour deux espèces données?
- Dans les confitures de fraise de la marque XXX, quel est le taux de sucre?
- A la naissance, et pour telle espèce, les individus mâles sont-ils plus lourds ou plus légers que les individus femelles?
- Quelle est la proportion de graines qui germeront dans telle production de graines?

Pour répondre à ces questions, on doit faire des "expériences", dont les résultats ne pourront donner une réponse exacte et certaine, puisque ces résultats ont une variabilité intrinsèque. Les résultats de ces expériences sont des **variables aléatoires**. Pour répondre aux questions, il faut leur donner une formulation simple, qui porte en général sur un nombre "idéalisé" de la variable aléatoire: sa valeur moyenne, sa dispersion, etc...

Choisir un **modèle** c'est choisir le type de loi de probabilité qui va décrire les variables aléatoires de l'expérience. Ensuite, il faut aussi choisir le ou les paramètre(s) de cette loi en utilisant au mieux les informations que l'on a à sa disposition.

Dans ce cours, on commencera par rappeler les lois de variables aléatoires les plus couramment utilisées dans des modélisations simples; les paramètres descriptifs utiles; on rappellera les calculs de probabilité déjà vus l'année précédente; on abordera ensuite les tests et intervalles de confiance, et on étudiera un certain nombre de procédures statistiques permettant de résoudre quelques problèmes typiques.

2 Rappels de probabilité.

2.1 Lois de variables aléatoires

Une variable aléatoire X est un nombre qui est le résultat d'une expérience à l'issue incertaine. On aime bien quantifier le résultat d'une expérience à l'aide d'un nombre: on peut regarder des sommes et des moyennes etc.

2.1.1 Variables discrètes

Lorsque X prend ses valeurs dans un ensemble E qui est fini ou dénombrable alors on dit que la variable X est discrète. La loi de X est alors la donnée des nombres p_x où x parcourt l'ensemble E. Pour chaque résultat possible x, le nombre p_x représente la probabilité pour que $X = x$. Ceci est noté $p_x = P(X = x)$. Par exemple lorsque X est le résultat du lancer d'un dé à six faces, alors $E = \{1, 2, 3, 4, 5, 6\}$, et pour tout x dans E, $p_x = 1/6$. On a toujours $p_x \in [0, 1]$.

Si on somme toutes les valeurs p_x où x varie dans E, on obtient 1. On note ceci de la manière suivante:

$$\sum_{x\ :\ x\in E} p_x = 1.$$

On exprime parfois p_x en pourcentage. Lorsque $p_x = 0.2$ on dit que la probabilité est 20% (on multiplie p_x par 100 pour obtenir le pourcentage. Par exemple $p_x = 0.001$ signifie 0.1%).

On dit que l'on a équiprobabilité lorsque E est fini et lorsque tous les p_x sont égaux. Dans le cas du lancer du dé, on a par exemple équiprobabilité. En fait, puisque $\sum_{x\in E} p_x = 1$, si on a équiprobabilité, alors tous les p_x sont égaux à l'inverse du nombre d'éléments de E.

Exemples de lois discrètes:

- La loi de Bernoulli de paramètre $p \in [0,1]$: $E = \{0,1\}$, $p_0 = 1-p$, $p_1 = p$.
- La loi binomiale $B(n,p)$: $E = \{0,1,2,\ldots,n\}$ et pour tout $j \in E$, $p_j = C_n^j p^j (1-p)^{n-j}$.
- La loi géométrique de paramètre q: $E = \{1,2,\ldots\}$ et pour tout $n \geq 1$, $p_n = q(1-q)^{n-1}$.
- La loi de Poisson de paramètre λ: $E = \{0,1,2,\ldots\}$ et pour tout $n \geq 0$, $p_n = e^{-\lambda}\lambda^n/n!$

Parfois, on cherche à calculer la probabilité pour que X appartienne à un certain sous-ensemble A de E. Pour cela, il suffit d'additionner les p_x pour tous les x dans A. On note:

$$P(X \in A) = \sum_{x : x\in A} p_x.$$

Notons que l'on a bien $P(X \in E) = 1$.

2.1.2 Variables à densité

Si on peut mesurer X avec une précision infinie (par exemple X est un instant, une distance etc), comment caractériser sa loi? Par exemple, X est le résultat d'un lancer de javelot. Si on mesure au mètre près $P(X \in [20,21]) = 9\%$. Si on mesure au décimètre près, $P(X \in [20.0, 20.1]) = 1\%$. Si on mesure au centimètre près, la probabilité pour que $x \in [20.00, 20.01]$ sera encore environ 10 fois plus petite. En fait lorsque δ est très petit, on voit apparaître une relation de proportionnalité entre $P(X \in [20, 20+\delta])$ et δ. Le coefficient de proportionnalité est la densité de la loi de X en $x = 20$.

On dit qu'une variable aléatoire X a pour densité la fonction f, si pour tout réel x,

$$P(X \in [x, x+dx]) = f(x)dx.$$

$f(x)$ est le coefficient de proportionnalité en x.

On calcule alors $P(X \in A)$ en intégrant f:

$$P(X \in A) = \int_{x\in A} f(x)dx.$$

Notons que f est une fonction positive telle que

$$\int_{x\in\mathbb{R}} f(x)dx = P(X \in \mathbb{R}) = 1.$$

Exemples de lois à densité:

- La loi uniforme: C'est l'analogue de l'équiprobabilité. La densité est constante sur tout un intervalle. Par exemple, la loi uniforme sur l'intervalle $[a,b]$ a pour densité $1/(b-a)$ sur $[a,b]$ (et une densité nulle en dehors de cet intervalle). Le coefficient $1/(b-a)$ assure que $\int_a^b dx/(b-a) = 1$.
- La loi exponentielle: La densité d'une loi exponentielle de paramètre λ, est $\lambda e^{-\lambda x}$ lorsque $x > 0$ (et zéro lorsque $x \leq 0$).
- La loi Gaussienne centrée réduite $\mathcal{N}(0,1)$. On dit que X suit la loi $\mathcal{N}(0,1)$ si X a pour densité

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

  Le terme $1/(\sqrt{2\pi})$ assure que l'intégrale de la densité sur $]-\infty,\infty[$ vaut 1.
- La loi Gaussienne $\mathcal{N}(m,\sigma^2)$. Lorsque X suit la loi $\mathcal{N}(0,1)$, alors σX suit la loi $\mathcal{N}(0,\sigma^2)$. De même $m+\sigma X$ suit la loi $\mathcal{N}(m,\sigma^2)$. La densité de la loi $\mathcal{N}(m,\sigma^2)$ est

$$\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2}).$$

2.2 Espérance, variance

2.2.1 Espérance

L'espérance d'une variable aléatoire est la moyenne a priori des valeurs prises par X (pondérées par leurs probabilité). Elle est notée $E(X)$ ou m_X. Dans le cas discret:

$$E(X) = \sum_{x:x\in E} x p_x.$$

Par exemple, pour le dé à six faces,

$$E(X) = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = \frac{7}{2}.$$

Dans le cas ou X a pour densité f, on a de même,

$$E(X) = \int x f(x) dx.$$

Exemples:

- Si X a pour loi la loi de Poisson de paramètre λ, $E(X) = \lambda$.
- Si X a pour loi la loi uniforme sur $[a,b]$, $E(X) = (a+b)/2$.
- Si X a pour loi la loi exponentielle de paramètre λ, alors $E(X) = 1/\lambda$.
- Si X a pour loi $\mathcal{N}(m,\sigma^2)$ alors $E(X) = m$.

On peut noter que l'espérance d'une somme de variables aléatoires est égale à la somme des espérances de chacune de ces variables.

2.2.2 Variance, écart-type

On cherche à définir un nombre qui donne une indication sur l'ordre de grandeur de l'écart entre une réalisation de X et sa moyenne a priori m_X. Pour cela, on définit la variance de X, notée $\text{var}(X)$ ou σ_X^2, par

$$\text{var}(X) = E((X - m_X)^2).$$

C'est la valeur moyenne a priori des carrés des écarts à m_X. L'écart-type σ_X est la racine carrée de la variance de X. Le nombre σ_X donne une idée de l'ordre de grandeur de $X - m_X$.

Par exemple, si $m = 20$ et $\sigma = 0.1$ on aura typiquement des valeurs du type 20.1, 20.02, 19.9, 19.95. Lorsque $m = 20$ et $\sigma = 10$, alors on a plutôt des valeurs du type 8, 18, 30, 25 etc.

Lorsque l'on additionne à une variable aléatoire X un nombre réel a fixé, on ne change pas sa variance. Par contre, on ajoute a à son espérance.

Lorsque l'on multiplie une variable aléatoire X par un nombre positif fixé λ, alors on multiplie son espérance aussi par λ et on multiplie sa variance par λ^2. Par exemple

$$E(2X) = 2E(X) \text{ et } \text{var}(2X) = 4\text{var}(X).$$

Ceci est une conséquence immédiate de la définition de l'espérance et de la variance.

En particulier, si X suit la loi $\mathcal{N}(m, \sigma^2)$ alors X a pour espérance m et pour variance σ^2. On appelle parfois la loi $\mathcal{N}(m, \sigma^2)$, la loi normale de moyenne m et de variance σ^2.

Notons que l'on a la relation suivante, parfois pratique pour calculer la variance:

$$\text{var}(X) = E(X^2) - m_X^2.$$

En effet, en développant $(X - m_X)^2$ dans la définition de la variance, on voit que

$$\text{var}(X) = E(X^2) - 2m_X E(X) + m_X^2 = E(X^2) - 2m_X^2 + m_X^2.$$

2.3 Inégalité de Markov, inégalité de Bienaymé-Chebychev

Soit X une variable aléatoire discrète, avec (pour tout $x \in E$), $P(X = x) = p_x$. Rappelons que l'espérance de X vaut

$$m_X = E(X) = \sum_{x \,:\, x \in E} x p_x.$$

Supposons que $P(X < 0) = 0$, c'est à dire que X est positive. Alors, pour tout nombre x_0 fixé, on a

$$E(X) \geq \sum_{x : x \geq x_0} x p_x \geq x_0 \sum_{x : x \geq x_0} p_x = x_0 P(X \geq x_0).$$

Proposition 1 (Inégalité de Markov) *Pour toute variable aléatoire positive X, et pour tout nombre positif x_0,*

$$P(X \geq x_0) \leq \frac{E(X)}{x_0}.$$

En gros, si l'espérance de X n'est pas très grande et si X est positive, alors la probabilité pour que X soit grand n'est pas trop grande.

Supposons maintenant que Y est une variable aléatoire d'espérance m_Y. On pose alors $X = (Y - m_Y)^2$. C'est une variable aléatoire positive (puisque c'est un carré). De plus, l'espérance de X est la variance de Y. L'inégalité de Markov pour X s'écrit alors en termes de Y pour $x = y^2$ comme suit:

$$y^2 P((Y - m_Y)^2 \geq y^2) \leq \text{var}(Y).$$

Proposition 2 (Inégalité de Bienaymé-Chebychev) *Pour tout $y > 0$,*

$$P(|Y - m_Y| \geq y) \leq \frac{\text{var}(Y)}{y^2}.$$

Quand la variance de Y n'est pas grande, Y ne peut pas être trop éloigné de son espérance avec une grande probabilité.

2.4 Indépendance

On dit que les variables aléatoires discrètes $X_1, \ldots, X_n$ sont indépendantes si pour tous $x_1, \ldots, x_n$, on a

$$\begin{aligned} & P(X_1 = x_1 \text{ et } X_2 = x_2 \text{ et } \ldots \text{ et } X_n = x_n) \\ = \quad & P(X_1 = x_1) \times P(X_2 = x_2) \times \cdots \times P(X_n = x_n). \end{aligned}$$

Par exemple, lorsque l'on lance un dé à six faces deux fois de suite et que l'on observe X_1 puis X_2 on a

$$P(X_1 = x_1 \text{ et } X_2 = x_2) = \frac{1}{36}$$

pour tous x_1 et x_2 dans $\{1, 2, \ldots, 6\}$.

De même, on dit que les variables à densité $X_1, \ldots, X_n$ sont indépendantes si pour tous $x_1, \ldots, x_n$,

$$\begin{aligned} & P(X_1 \in [x_1, x_1 + dx_1] \text{ et } \ldots \text{ et } X_n \in [x_n, x_n + dx_n]) \\ = \quad & P(X_1 \in [x_1, x_1 + dx_1]) \times \cdots \times P(X_n \in [x_n, x_n + dx_n]). \end{aligned}$$

On peut remarquer que si Y et Z sont deux variables aléatoires indépendantes alors

$$E(YZ) = E(Y)E(Z).$$

En effet,

$$\begin{aligned} E(YZ) &= \sum_{y,z} yzP(Y=y, Z=z) \\ &= \sum_{y,z} yzP(Y=y)P(Z=z) \qquad \text{(indépendance)} \\ &= \sum_{y} yP(Y=y) \sum_{z} zP(Z=z) \qquad \text{(factorisation)} \\ &= E(Y)E(Z) \end{aligned}$$

Attention! Il se peut que $E(YZ) = E(Y)E(Z)$ même si Y et Z ne sont pas indépendantes.

2.5 Sommes de variables aléatoires indépendantes

On suppose que $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes. On définit $S = X_1 + \cdots + X_n$ la somme des variables $X_1, \ldots, X_n$. On a déjà vu que l'espérance m_S de S était égale à la somme des espérances $m_{X_1} + \cdots + m_{X_n}$. Lorsque les variables sont indépendantes, on a en fait que la variance de la somme est la somme des variances. En d'autres termes:

Proposition 3 *Si $X_1, \cdots, X_n$ sont n variables aléatoires indépendantes,*

$$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n).$$

Pour le démontrer pour $n = 2$, on pose $Y = X_1 - m_{X_1}$ et $Z = X_2 - m_{X_2}$. Ce sont deux variables aléatoires indépendantes de moyenne nulle de sorte que $E(YZ) = 0$. Alors,

$$\text{var}(X_1 + X_2) = E((Y+Z)^2) = E(Y^2) + 2E(YZ) + E(Z^2) = \text{var}(X_1) + \text{var}(X_2).$$

Attention: si les variables ne sont pas indépendantes, la variance de leur somme n'est en général pas la somme des variances.

2.6 Moyennes empiriques

On suppose maintenant que $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes de même loi. On appelle m leur espérance commune, et σ^2 leur variance commune. On pose

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}.$$

$\overline{X}$ est une variable aléatoire, qui est la moyenne de $X_1, \ldots X_n$. On l'appelle moyenne empirique. D'après ce qui précède

$$E(\overline{X}) = \frac{1}{n} E(X_1 + \cdots + X_n) = \frac{1}{n}(m + \cdots + m) = m.$$

De même, comme les variables sont indépendantes,

$$\mathrm{var}(\overline{X}) = \frac{1}{n^2}\mathrm{var}(X_1 + \cdots + X_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

On écrit alors l'inégalité de Bienaymé-Chebychev pour $\overline{X}$:

Proposition 4 (Loi des grands nombres) *Pour tout $x > 0$,*

$$P(|\overline{X} - m| > x) \leq \frac{\sigma^2}{nx^2}.$$

On note que lorsque n augmente vers l'infini, alors le terme de droite tend vers zéro: Donc la probabilité pour que $|\overline{X} - m| > x$ tend vers zéro.

Lorsque n devient grand, $\overline{X}$ a tendance a être de plus en plus près de m.

Il est important de ne pas confondre $\overline{X}$, la moyenne empirique, avec m, l'espérance des X_i, que l'on appelle parfois leur moyenne. m est un nombre fixé (la moyenne "a priori" des valeurs que peut prendre la variable), alors que $\overline{X}$ est une variable aléatoire.

2.7 Sommes de Gaussiennes indépendantes, sommes de variables de Poisson indépendantes

Les lois gaussiennes et les lois de Poisson possèdent toutes deux une propriété remarquable:

Proposition 5

- *Si $X_1, \ldots, X_n$ sont n variables gaussiennes indépendantes alors la loi de $X_1 + \cdots + X_n$ est encore une loi gaussienne.*
- *De même, si $X_1, \ldots, X_n$ sont n variables de Poisson indépendantes alors la loi de $X_1 + \cdots + X_n$ est encore une loi de Poisson.*

Rappelons que l'espérance de $S = X_1 + \cdots + X_n$ est la somme des espérances, et que la variance de S est la somme des variances lorsque $X_1, \ldots, X_n$ sont indépendantes.

Donc, si $X_1, \cdots, X_n$ sont des variables aléatoires indépendantes de loi de Poisson d'espérances $\lambda_1, \ldots, \lambda_n$ alors la loi de $S = X_1 + \cdots + X_n$ est une loi de Poisson $\mathcal{P}(\lambda)$ d'espérance $\lambda = \lambda_1 + \cdots + \lambda_n$. Ceci peut se montrer de la manière

suivante, dans le cas où $n = 2$:

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{l=0}^{k} P(X_1 = l \text{ et } X_2 = k - l) \\ &= \sum_{l=0}^{k} P(X_1 = l) P(X_2 = k - l) \\ &= \sum_{l=0}^{k} e^{-\lambda_1} \frac{\lambda_1^l}{l!} e^{-\lambda_2} \frac{\lambda_2^{k-l}}{(k-l)!} \\ &= \frac{e^{-\lambda_1 - \lambda_2}}{k!} \sum_{l=0}^{k} \frac{k!}{l!(k-l)!} \lambda_1^l \lambda_2^{k-l} \\ &= e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!}. \end{aligned}$$

Lorsque $X_1, \cdots, X_n$ sont des variables gaussiennes indépendantes de lois respectives $\mathcal{N}(m_1, \sigma_1^2), \ldots, \mathcal{N}(m_n, \sigma_n^2)$ alors la loi de S est $\mathcal{N}(m_1 + \cdots + m_n, \sigma_1^2 + \cdots + \sigma_n^2)$.

Dans le cas particulier important où $m_1 = m_2 = \cdots = m_n$ et où $\sigma_1^2 = \cdots = \sigma_n^2$, c'est à dire que $X_1, \cdots, X_n$ sont gaussiennes indépendantes de même loi, alors:

- La loi de S est $\mathcal{N}(nm, n\sigma^2)$
- La loi de $\overline{X} = S/n$ est $\mathcal{N}(m, \sigma^2/n)$
- La loi de $\sqrt{n}\frac{\overline{X} - m}{\sigma} = \frac{S - nm}{\sigma\sqrt{n}}$ est $\mathcal{N}(0, 1)$.

2.8 Approximations de sommes de variables aléatoires

Plus généralement, on dit que $X_1, \ldots, X_n$ est un n-échantillon d'une loi $\mathcal{L}$ lorsque $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes de même loi $\mathcal{L}$. En d'autres termes, ce sont les résultats d'une même expérience recommencée n fois de manière indépendante.

2.8.1 Approximations de la Binomiale

Supposons maintenant que $X_1, \ldots, X_n$ est un n-échantillon de la loi de Bernoulli de paramètre p. En d'autres termes, chacun des X_i vaut 1 avec probabilité p et 0 avec probabilité $1 - p$.

Notons que l'espérance de X_1 est p et que la variance de X_1 est $E(X_1^2) - E(X_1)^2 = p(1 - p)$.

Alors, lorsque n est grand, et suivant la valeur de p, on peut approximer la loi de $S = X_1 + \cdots + X_n$ en utilisant une loi de Poisson ou en utilisant une loi gaussienne. Rappelons que la loi de S est alors une loi binomiale $\mathcal{B}(n, p)$. Cependant, lorsque n est grand, la loi binomiale fait intervenir des termes énormes (n!) difficiles à manipuler. On utilise alors des approximations de la loi binomiale.

Proposition 6 (Approximation par une loi de Poisson) *Lorsque le nombre n est grand et $\lambda = np$ est petit (en pratique, n supérieur à 30 et np inférieur à 5), alors la loi de S est proche d'une loi de Poisson de paramètre λ.*

Notons que l'espérance de la loi de Poisson est λ qui est bien identique à l'espérance de S. Prenons un exemple: $p = 0.001$ et $n = 1000$. Que vaut $P(S = 3)$? On approxime la loi de S par une loi de Poisson de paramètre $1000 \times 0.001 = 1$. Alors:

$$P(S = 3) \sim e^{-1}1^3/3! = 0.061 = 6.1\%.$$

Proposition 7 (Approximation par une Gaussienne) *Lorsque n est grand, et lorsque np et $n(1-p)$ ne sont pas trop petits (en pratique: tous deux supérieurs à 10), alors la loi de*

$$\frac{S - np}{\sqrt{np(1-p)}}$$

est proche d'une loi Gaussienne $\mathcal{N}(0,1)$.

Notons que $E(S) = np$ et donc que l'espérance de $(S - np)/\sqrt{np(1-p)}$ est nulle (tout comme celle de la loi $\mathcal{N}(0,1)$. De même, la variance de S est égale à n fois la variance de X_1 (comme somme de n variables indépendantes de même loi que X_1). Donc $\text{var}(S) = np(1-p)$. La variance de $(S - np)/\sqrt{np(1-p)}$ est donc bien 1, tout comme la variance de la loi $\mathcal{N}(0,1)$.

En pratique, dire que la loi de $(S - np)/\sqrt{np(1-p)}$ est proche d'une loi $\mathcal{N}(0,1)$, revient à dire que pour tout $a < b$ fixés, la probabilité

$$P\left(\frac{S - nm}{\sqrt{np(1-p)}} \in [a,b]\right)$$

peut être approximée par

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

Par exemple, supposons que $p = 0.5$ et que $n = 10000$. Ceci se produit par exemple lorsque l'on joue 10000 fois à pile ou face. $S = X_1 + \cdots + X_n$ représente alors le nombre de fois où est apparu 'face'. Que vaut alors $P(S \in [4900, 5100])$? Pour déterminer celà, on note que $S \in [4900, 5100]$ est équivalent à $(S - 5000)/50 \in [-2, 2]$. Mais la loi de $(S - 5000)/50$ est proche d'une loi gaussienne $\mathcal{N}(0,1)$. On regarde dans la table la probabilité pour qu'une telle gaussienne soit plus grande que 2: Cette probabilité vaut 0.0228. La probabilité pour qu'elle soit inférieure à -2 est également 0.0228. Donc

$$P(S \in [4900, 5100]) \sim 0.9544 = 1 - 2P(N > 2).$$

On peut remarquer ici que l'utilisation de l'inégalité de Chebychev conduit à $P(S \in [4900, 5100]) \geq 0.75$, ce qui est moins précis.

Quand on peut les employer, les approximations par des lois de Poisson ou gaussiennes donnent des informations plus précises que l'inégalité de Chebychev.

2.8.2 Cas général: Théorème central limite

En fait, l'approximation par une variable aléatoire gaussienne est licite pour n'importe quelle loi:

Proposition 8 (Théorème central limite) *Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi $\mathcal{L}$ de moyenne m et de variance σ^2. On pose $S = X_1 + \cdots + X_n$. Alors, lorsque n tend vers l'infini, la loi de*

$$\frac{S - nm}{\sigma\sqrt{n}}$$

converge vers la loi $\mathcal{N}(0,1)$.

2.8.3 Méthodes empiriques

Il faut noter l'importance du théorème central limite: Il permet non seulement de dire que lorsque n est grand alors S/n est très proche de la moyenne m, mais il donne également une information précise sur l'écart entre la valeur observée de S/n et m.

Par exemple si l'on sait a priori que $\sigma^2 = 1$ et que $n = 100$, et que l'approximation est licite, alors la probabilité pour que $\overline{X} > m + \sigma \times 1.96/\sqrt{n} = m + 0.196$ est proche de 2.5%. En particulier, si m est inconnue, on voit que la probabilité pour que m soit plus petite que la valeur observée $\overline{X}$ moins 0.196 est de l'ordre de 2.5%.

Plus généralement, si l'on cherche à obtenir des informations sur une quantité (inconnue) qui peut s'écrire comme l'espérance d'une certaine variable reliée à l'expérience, une moyenne empirique bien choisie sera le bon outil statistique. Autrement dit, si l'expérience conduit à observer les variables aléatoires $X_1, \ldots, X_n$, et que l'on cherche à obtenir des informations sur la valeur (inconnue) $\theta = E(\phi(X_1))$ où ϕ est une fonction donnée, alors on peut définir les variables $Y_i = \phi(X_i)$ et leur appliquer le théorème central limite pour voir que la moyenne empirique

$$\overline{Y} = \frac{Y_1 + \ldots + Y_n}{n} = \frac{\phi(X_1) + \ldots + \phi(X_n)}{n}$$

a une loi proche de $\mathcal{N}(\theta, \sigma^2/n)$ lorsque n est grand pour σ^2 égal à la variance de Y_1.

3 Tests d'hypothèses et intervalles de confiance.

3.1 Principe des tests

Jusqu'à présent, nous avons supposé que nous connaissions la loi d'une ou plusieurs variables aléatoires, et nous avons montré comment, en utilisant cette information, on pouvait calculer des probabilités d'événements.

En pratique, on observe souvent des résultats de variables aléatoires dont on ne connait pas la loi. On cherche en fait justement, à partir des observations des résultats à obtenir des informations sur la loi. Compte tenu du fait que ces observations sont le résultat d'expériences aléatoires, les conclusions que l'on peut en tirer ne sont jamais des certitudes.

Commençons par un exemple simple. Un fabricant annonce que parmi les nombreux produits qu'il vend, moins de 10% sont défectueux. On en choisit par exemple 10 au hasard et on constate qu'ils ont tous un défaut. On a envie d'en conclure que le fabricant a menti. Evidement, il se peut que le fabricant n'ait pas menti et qu'on a eu une malchance terrible de tomber dix fois de suite sur un produit défectueux. Cependant, on sent bien que cela est fort peu probable, et on peut déclarer avec relative certitude qu'il a menti.

Maintenant, supposons que parmi les 10 produits testés, 4 s'avèrent défectueux. On est à ce moment moins confiant lorsque l'on affirme que le fabricant a menti. Le principe des tests statistiques que nous allons décrire dans la suite du cours est d'utiliser les calculs de probabilités que nous avons présenté jusqu'à maintenant pour pouvoir déterminer exactement à partir de quel nombre de produits défectueux on peut affirmer (avec une confiance de 95% disons) que le producteur a menti.

Le même type de raisonnenent s'applique par exemple lorsque l'on se pose les questions suivantes:

1. Quelle est la proportion de graines qui germeront dans telle production de graines ?
2. Ou: dans les confitures de fraise de la marque XXX, quel est le taux de sucre ?

On sème des graines et on observe si elles germent, ou on ouvre quelques pots de confiture, on analyse le contenu et on pèse la quantité de sucre. On souhaite répondre aux questions à l'aide des expériences.

Etudions plus en détail la question 2. On peut se demander: le taux de sucre est-il de 50% ? Pour cela, on fait n analyses, qui donnent pour résultats $x_1, \ldots, x_n$. Bien entendu, tous ces résultats sont différents, à cause d'une part des erreurs de mesure, d'autre part du fait que la fabrication de confiture contient une part de variabilité. On va alors supposer que les résultats des expériences sont les valeurs observées de variables gaussiennes indépendantes $X_1, \ldots, X_n$ de loi $\mathcal{N}(m, \sigma^2)$. Choisir la forme gaussienne de la loi, c'est choisir un modèle. Le paramètre m est celui sur lequel la question se pose. Le paramètre σ indique la précision que l'on pense avoir sur la mesure. On peut alors reformuler la question: "est-ce que m vaut 50?"

Pour cela, on va chercher une variable $Z(X_1, \ldots, X_n)$ donnant une bonne approximation de m; ici, la moyenne observée (ou moyenne empirique) $Z = \overline{X}$ convient. On répondra selon la valeur observée de $Z = \overline{X}$: Si Z est proche de 50, on dira "oui, le taux est 50". Si la valeur observée est éloignée de 50, on dira "non, le taux n'est pas 50". Autrement dit, on choisira un domaine I de valeurs proches de 50 et on répondra "oui" si $Z \in I$, et "non" si $Z \notin I$. Evidemment, Z peut prendre toutes les valeurs possibles, donc la réponse que l'on donne n'est pas certaine: on appelle **niveau du test** la probabilité de répondre "non" alors

que la réponse vraie est "oui": c'est **l'erreur de première espèce.**

Il y a une autre manière de se tromper: si on répond "oui" alors que la réponse vraie est "non": c'est **l'erreur de deuxième espèce.**

Le niveau du test est souvent noté α. On choisit l'intervalle I de telle sorte que si la vraie réponse est "oui", alors la probabilité pour que Z tombe dans I est $1-\alpha$. A ce moment-là, si la réponse vraie est "oui", la probabilité pour se tromper est bien α.

La façon dont on choisit le domaine I dépend aussi de la façon dont on craint de se tromper: si l'on craint que le taux soit plus grand que 50, on décidera de répondre "non" seulement si Z est plus grand qu'une valeur seuil fixée.

Résumons encore le principe général: On observe des résultats de variables aléatoires dont on ne connait pas la loi. On cherche, à partir des observations des résultats à obtenir des informations sur la loi. Compte tenu du fait que ces observations sont le résultat d'expériences aléatoires, les conclusions que l'on peut en tirer ne sont jamais des certitudes.

Supposons par exemple que l'on observe le résultat d'une expérience aléatoire de loi inconnue $\mathcal{L}$: $Z = z_{obs}$. On souhaite à partir de cette observation décider si une loi $\mathcal{L}_0$ est plausible et compatible avec cette observation. Pour cela, l'idée est de déterminer un intervalle I de valeurs tel que $Z \in I$ avec une grande probabilité (si la loi de Z est $\mathcal{L}_0$). On peut le faire en utilisant les techniques développées jusqu'à maintenant. Si la valeur observée z_{obs} tombe dans cet ensemble de valeurs plausibles pour $\mathcal{L}_0$, alors on peut accepter l'hypothèse que la loi est $\mathcal{L}_0$. Sinon, cela signifie que l'observation est tombée dans l'ensemble des valeurs très peu plausibles pour $\mathcal{L}_0$, et on rejette l'hypothèse suivant laquelle la loi est $\mathcal{L}_0$.

Plus précisément, pour faire un test, il faut:

* Une hypothèse prioritaire notée H_0 sur la loi de Z: "La loi $\mathcal{L}$ de Z est exactement la loi $\mathcal{L}_0$" où $\mathcal{L}_0$ est une loi explicitement décrite.

* Une hypothèse de remplacement H_1, qui sera la conclusion si on décide que l'hypothèse H_0 est fausse. H_1 peut être une hypothèse précise sur la loi de Z (par exemple "la loi $\mathcal{L}$ est précisément la loi $\mathcal{L}_1$) ou une hypothèse un peu floue ("la loi $\mathcal{L}$ n'est pas la loi $\mathcal{L}_0$").

* Le niveau α du test, qui est un petit nombre (souvent $1\%, 2.5\%, 5\%$).

Pour effectuer le test de H_0 contre H_1 au niveau α, on procède comme suit:

Tout d'abord, on regarde si les valeurs de Z dans le cas où l'hypothèse H_1 est réalisée, ont tendance à être plus grandes ou plus petites que si l'hypothèse H_0 est réalisée (c'est en fait la seule information sur l'hypothèse H_1 que l'on utilise). Supposons par exemple que sous H_1, les valeurs de Z ont tendance à être plus grandes que sous H_0.

On va chercher un nombre z_0 dans la table des valeurs pour la loi $\mathcal{L}_0$ (ou par le calcul), tel que si Z suit la loi $\mathcal{L}_0$ alors

$$P(Z > z_0) = \alpha.$$

En d'autres termes, avec probabilité $1-\alpha$ et si Z a pour loi $\mathcal{L}_0$, $Z \leq z_0$. Les observations plus grandes que z_0 sont exceptionnelles si Z suit la loi $\mathcal{L}_0$. On

observe $Z = z_{obs}$. Si la valeur de $z_{obs} > z_0$ (c'est à dire si elle est exceptionnellement grande pour la loi $\mathcal{L}_0$) alors la conclusion du test sera que H_0 est fausse et que H_1 est réalisée. Si la valeur observée z_{obs} est plus petite que z_0, on garde l'hypothèse H_0.

Schéma de la construction d'un test:

1. Déterminer dans quel cadre on se situe, c'est à dire préciser le modèle.
2. Au vu de la question posée, déterminer quelles seront les hypothèses (H_0) et (H_1) du test. Attention: il faut choisir pour H_0 l'hypothèse sous laquelle on sait faire des calculs.
3. Déterminer une statistique (une fonction des échantillons observés) dont on connaît la loi sous (H_0) et dont la loi sous (H_0) est différente de la loi sous (H_1).
4. A partir de la loi de la statistique sous (H_0) et sous (H_1), on établit une règle qui permet de décider quand on rejette et quand on accepte (H_0). On en déduit la forme de la région de rejet de (H_0).
5. On calcule enfin le ou les seuil(s) de la région de rejet grâce à la loi de la statistique sous (H_0).
6. On calcule la valeur observée de la statistique et on conclut.

Maintenant, le choix a priori d'un niveau est relativement arbitraire. Lorsqu'ayant fixé un niveau, on répond par "oui" ou "non" après avoir comparé la valeur observée avec la valeur seuil, on n'indique pas si la valeur observée est proche ou loin de la valeur seuil. Or cette information a un sens: si la valeur observée est loin de la valeur seuil, c'est que le résultat est hautement probable sous H_0. On préfère alors souvent répondre à l'aide du niveau de signification (ou niveau observé).
On appelle **niveau de signification (ou niveau observé)** d'un test la valeur α_{obs} à partir de laquelle, étant donnée l'observation, on rejetterait H_0. Autrement dit, du test,

- Pout $\alpha < \alpha_{obs}$, on accepterait H_0,
- Pour $\alpha > \alpha_{obs}$, on rejetterait H_0.

3.2 Cas des variables gaussiennes de variance connue

En pratique, on teste la loi à partir de beaucoup de réalisations. Par exemple, on dispose du résultat de n variables aléatoires $X_1, \ldots, X_n$ indépendantes, de même loi $\mathcal{L}$ inconnue. On veut alors tester des hypothèses sur la loi $\mathcal{L}$. Pour cela, il faut alors choisir une variable de test Z qui est une fonction des n observations $X_1, \ldots, X_n$. Souvent, on choisit

$$Z = X_1 + X_2 + \cdots + X_n$$

ou alors

$$Z = \frac{X_1 + \cdots + X_n}{n}.$$

On fait une hypothèse sur la loi des X_i: Par exemple, H_0:"la loi $\mathcal{L}$ est la loi $\mathcal{N}(m_0, 1)$", pour une valeur donnée de m_0. Alors, d'après les chapitres précédents, on est capable d'en déduire la loi de Z. Dans notre exemple, si H_0 est vraie, alors la loi de

$$Z = \frac{X_1 + \cdots + X_n}{n}$$

est la loi $\mathcal{N}(m_0, 1/n)$. On teste alors cette hypothèse sur la loi de Z à partir de l'observation de Z.

Supposons que l'on souhaite tester au niveau 5% l'hypothèse H_0: "$m = 50$" contre H_1: "$m \neq 50$". Si H_0 est réalisée, la loi de $Z = (X_1 + \cdots + X_{10000})/10000$ est $\mathcal{N}(50, 1/10000)$. En particulier, la loi de $100(Z - 50)$ est alors $\mathcal{N}(0, 1)$. On fait un test bilatère qui élimine les trop grandes ou trop petites observations de Z. On a

$$P(100(Z - 50) \in [-1.96, 1.96]) = 95\%.$$

On garde donc H_0 dès que $Z \in [50 - 0.0196, 50 + 0.0196]$ et on rejette H_0 sinon. Imaginons que la valeur observée de Z est $Z = 50.03$, alors on rejette l'hypothèse H_0 car la moyenne observée est anormalement grande.

Si l'on avait souhaité tester au niveau 5% l'hypothèse H_0: "$m = 50$" contre H_1: "$m > 50$", on aurait choisi de rejeter H_0 seulement si Z est plus grand qu'une certaine valeur seuil z. On fait alors un test unilatère qui élimine les trop grandes observations de Z. Pour calculer z, si on se fixe un niveau α, on veut que si Z suit la loi $\mathcal{N}(50, 1/10000)$ on ait

$$P(Z \geq z) = 5\%.$$

Pour cela il faut se ramener à la loi gaussienne centrée réduite en écrivant

$$P(100(Z - 50) \geq 100(z - 50)) = 5\%.$$

On lit dans la table $100(z - 50) = 1.64$, et on rejette donc H_0 dès que $Z \geq 50 + 0.0164$

3.3 Cas des sommes de variables de Bernoulli

Reprenons le premier exemple. On se demande si la probabilité, pour une graine, de germer est de 80%; on sème n graines et on regarde si elle germent. On va choisir le modèle suivant: on dira $x_i = 1$ si la i–ème graine a germé, et $x_i = 0$ sinon. Autrement dit, on suppose que les valeurs $x_1, \ldots, x_n$ sont le résultat de l'observation de variables indépendantes $X_1, \ldots, X_n$ de loi de Bernoulli $\mathcal{B}(p)$, et la question que l'on se pose porte sur la valeur de p.

Pour cela, on peut utiliser les résultats sur l'approximation de la loi Binomiale en choisissant $Z = X_1 + \cdots + X_n$. En effet, si on observe un n-échantillon $X_1, \ldots, X_n$ d'une variable de Bernoulli (avec n grand) de paramètre p et que l'on effectue une hypothèse sur la valeur de p, alors d'après les chapitres précédents, on connait la loi approchée de $Z = X_1 + \cdots + X_n$ (une loi de Poisson ou une loi normale), et l'on peut alors comparer la valeur observée aux valeurs théoriques pour cette loi.

Par exemple, on veut tester H_0: "$p = p_0$" (pour une valeur fixée p_0) contre H_1: "$p < p_0$". Pour la germination des graines, on se demande si la probabilité, pour une graine donnée, de germer, est 0.8, avec l'idée que l'on veut détecter si c'est moins de 0.8.

On observe que sur 100 graines semées, 78 d'entre elles germent.

Sous l'hypothèse H_0, la loi de Z est approximativement gaussienne $\mathcal{N}(80, 16)$, et sous H_1, Z a tendance à prendre des valeurs plus petites. On choisit donc de rejeter H_0 si $Z \leq z$ pour une valeur z à déterminer en fonction du niveau α choisi. Par exemple pour $\alpha = 5\%$, on écrit que sous H_0,

$$P\left(\frac{Z-80}{4} \leq -1.64\right) = 0.05,$$

ce qui conduit à fixer $z = 80 - 4 \cdot 1.64$. Si $Z_{obs} = 78$, on accepte H_0.

3.4 Puissance d'un test

Il arrive que l'hypothèse de rechange H_1 soit une hypothèse précise sur la loi $\mathcal{L}$, du style: "La loi $\mathcal{L}$ est exactement la loi $\mathcal{L}_1$" où $\mathcal{L}_1$ est une loi explicite (par exemple la loi $\mathcal{N}(1, 2)$).

En d'autres termes, on doit décider si la loi de Z est $\mathcal{L}_0$ ou $\mathcal{L}_1$. Le test de H_0 contre H_1 consiste à choisir $\mathcal{L}_0$ si la valeur observée est "habituelle" pour $\mathcal{L}_0$, et à choisir $\mathcal{L}_1$ seulement si la valeur observée est extrêmement inhabituelle pour $\mathcal{L}_0$. Typiquement, on trouve un intervalle I (qui peut être $(\infty, z_0]$, ou (z_0, ∞) ou $[z_1, z_2]$ suivant le type - unilatère ou bilatère- de test) à partir de la table de valeurs pour $\mathcal{L}_0$ tel que si $z_{obs} \in I$ alors on choisit H_0 et si $z_{obs} \notin I$ alors on choisit H_1.

Pour mesurer si le test détecte effectivement H_1 dans le cas où la loi est en fait $\mathcal{L}_1$, on définit la puissance du test comme étant la probabilité, en supposant que H_1 est vraie, pour que la conclusion du test soit effectivement H_1. En d'autres termes, la puissance du test est $P_{H_1}(Z \notin I)$.

C'est une façon de mesurer si l'on arrive à détecter la différence entre les lois $\mathcal{L}_0$ et $\mathcal{L}_1$.

3.5 Intervalles de confiance

Souvent, on sait que la loi de la variable Z est d'un certain type (loi normale, loi de Poisson), mais on ne privilégie pas une loi particulière que l'on cherche à tester. On a donc une famille $\mathcal{L}_\theta$ de lois possibles: par exemple $\mathcal{L}_\theta$ serait la loi de Poisson de paramètre θ. On souhaite alors "connaitre" la valeur de θ. Evidemment on ne peut connaitre cette valeur avec certitude, il s'agit donc de donner une fourchette dans laquelle se trouve θ avec forte probabilité. Construire cette fourchette, c'est construire un intervalle de confiance. On se donne un niveau de confiance $1 - \alpha$ (la forte probabilité avec laquelle on souhaite que la fourchette contienne θ), et l'on construit un intervalle $I = [A, B]$ tel que la probabilité que θ soit dans I soit au moins $1-\alpha$. Evidemment, A et B dépendent de l'expérience. D'ailleurs, si c'étaient des nombres fixes, comme θ est aussi un

nombre fixe, parler de la probabilité que θ soit dans I n'aurait pas de sens. Autrement dit,

Définition: Un intervalle de confiance I de niveau de confiance $1-\alpha$ pour un paramètre θ est un intervalle $I=[A,B]$ dont les bornes A et B sont des variables aléatoires qui dépendent de l'expérience, donc dont on peut calculer des valeurs observées à l'aide des observations) , et qui vérifie: pour toute valeur possible de θ,

$$P_\theta\left(\theta \in I\right) \geq 1-\alpha.$$

Il faut noter que, comme un test, un intervalle de confiance est une procédure, avec laquelle on calcule une valeur observée; si on refait l'expérience, la valeur observée change, mais pas la procédure : ce n'est pas l'intervalle de confiance qui change, c'est sa valeur observée.

Une méthode de construction de l'intervalle de confiance est la suivante. On se donne un niveau de test α et on regarde l'ensemble des valeurs de θ telles que, si l'on teste l'hypothèse H_0: "$\mathcal{L}=\mathcal{L}_\theta$" contre H_1:"$\mathcal{L}\neq\mathcal{L}_\theta$" au niveau α, alors on garde l'hypothèse H_0.

Exemple: On sait que la variable aléatoire Z suit une loi gaussienne $\mathcal{N}(m,1)$ de variance 1, mais on ne connait pas son espérance m. On observe $z_{obs}=3$. On cherche à déterminer un intervalle de confiance avec 95% de confiance pour m. Si $m=m_0$, alors

$$P(Z\in[m_0-1.96,m_0+1.96])=95\%.$$

La règle de décision du test de H_0:"$m=m_0$" contre H_1:"$m\neq m_0$" est donc: si $Z\in[m_0-1.96,m_0+1.96]$, alors on décide $m=m_0$. Autrement dit, si $m_0\in[Z-1.96,Z+1.96]$. Un intervalle de confiance pour m au niveau de confiance $1-\alpha$ est donc $I=[Z-1.96,Z+1.96]$. Sa valeur observée est $I_{obs}=[1.04,4.96]$.

4 Test d'ajustement du χ^2

Le but de ce chapitre est de présenter un test permettant de tester des hypothèses sur des lois discrètes à partir du résultat d'un grand nombre d'expériences indépendantes. Dans le cas de sommes de variables aléatoires de Bernoulli, le résultat de chaque expérience X_i pouvait prendre deux valeurs: 0 ou 1. Souvent, les résultats peuvent prendre 3 valeurs, ou plus. Ici, on s'intéressera à ces cas.

4.1 La loi du χ^2

On définit maintenant une loi de probabilité qui nous servira pour tester des hypothèses sur des lois discrètes. Suposons que $Y_1,\ldots,Y_d$ sont des variables indépendantes ayant pour loi la loi normale centrée réduite $\mathcal{N}(0,1)$. On pose alors

$$Z=Y_1^2+\cdots+Y_d^2.$$

La loi de Z s'appelle la loi du χ^2 à d degrés de liberté. On la note χ^2_d. Notons que Z est toujours positive (c'est une somme de carrés).

On peut montrer que sa densité est (pour tout $x \geq 0$),

$$c_d x^{(d-2)/2} e^{-x/2}$$

(c_d est une constante choisie pour que $\int_0^\infty c_d x^{(d-2)/2} e^{-x/2} dx = 1$). Cette densité décroit donc rapidement lorsque x croît vers l'infini. En pratique, pour calculer des probabilités, si $d \leq 30$, on utilise les tables de valeurs numériques. Si $d \geq 30$, on peut utiliser l'approximation par une loi normale. En effet, on peut appliquer le Théorème central limite: Z/d est une moyenne empirique, et il suffit de calculer l'espérance et la variance des Y_i pour appliquer le théorème. On a:

$$E(Y_1^2) = 1, \quad \mathrm{Var}(Y_1^2) = 2,$$

donc le théorème central limite donne, pour d grand

$$\frac{Z - d}{\sqrt{2d}} \sim \mathcal{N}(0, 1).$$

4.2 Le modèle étudié

On suppose que l'on a un n-échantillon $X_1, \ldots, X_n$ d'une loi discrète $\mathcal{L}$. On note $a_1, \ldots, a_d$ les valeurs possibles pour X_1. La loi $\mathcal{L}$ est donc caractérisée par la donnée de $P(X_1 = a_1)$, $P(X_1 = a_2)$, ..., $P(X_1 = a_d)$. On souhaite tester une hypothèse sur cette loi à partir des observations de $X_1, \ldots, X_n$.

Soit l'hypothèse H_0 suivante sur la loi $\mathcal{L}$:

$$P(X_1 = a_1) = p_1, \ldots, P(X_1 = a_d) = p_d$$

où $p_1, \ldots, p_d$ sont des nombres donnés explicites avec $p_1 + \ldots + p_d = 1$.

Pour tester H_0, on veut tirer de l'expérience des informations sur les probabilités $P(X_1 = a_1)$, $P(X_1 = a_2)$, ..., $P(X_1 = a_d)$. Pour $i = 1, \ldots, d$, on note N_i le nombre d'obervations a_i parmi les n expériences. Il est naturel de comparer chacun des p_i avec la moyenne observée N_i/n.

Par exemple, on fait un sondage d'intentions de vote parmi n personnes. Il y a d candidats notés $a_1, \ldots, a_d$ aux élections. Parmi les n personnes, N_1 déclarent vouloir voter pour a_1, N_2 pour a_2, etc. Notons que $N_1 + \ldots N_d = n$.

4.3 Le théorème limite

On pose $n_1 = np_1$, $n_2 = np_2$, ..., $n_d = np_d$. Si l'hypothèse H_0 est vérifiée n_1 représente la moyenne théorique pour N_1, c'est à dire l'espérance de la variable aléatoire qui compte le nombre de personnes parmi les n qui vont voter a_1 . Par exemple si $n = 200$ et $p_1 = 0.5$, alors on s'attend en moyenne à avoir $n_1 = 100$ observations pour N_1.

Proposition 9 *Lorsque $n_1 \geq 5, \ldots, n_d \geq 5$, on peut approximer la loi de*

$$Z = \frac{(N_1 - n_1)^2}{n_1} + \cdots + \frac{(N_d - n_d)^2}{n_d}$$

par la loi χ^2_{d-1}.

Il est important de noter que l'on utilise la loi à $d-1$ degrés de liberté (et non d degrés de liberté). De plus, lorsque la loi $\mathcal{L}$ des X_i n'est pas donnée par $p_1, \ldots p_d$ alors les N_i auront tendance à s'éloigner des n_i et donc Z aura tendance à être plus grande.

4.4 Le test

On fait le test de l'hypothèse H_0 contre l'hypothèse H_1 : "l'hypothèse H_0 est fausse", au niveau α. On procède comme suit, en utilisant Z comme variable de test:

- On calcule les moyennes théoriques $n_1, \ldots, n_d$ et on vérifie que $n_1 \geq 5$, ..., $n_d \geq 5$.
- Sous l'hypothèse H_0, Z a tendance à être plus petite que sous H_1, on fait donc un test unilatère qui élimine H_0 pour les trop grandes valeurs observées de Z.
- Dans la table de la loi de χ^2_{d-1}, on cherche z_0 tel que si Z suit χ^2_{d-1}, alors

$$P(Z > z_0) = \alpha.$$

- La règle de décision est: Si $z_{obs} > z_0$, on élimine H_0 au profit de H_1. Si $z_{obs} \leq z_0$, on garde l'hypothèse H_0.
- On calcule la valeur observée

$$z_{obs} = \frac{(N_1 - n_1)^2}{n_1} + \cdots + \frac{(N_d - n_d)^2}{n_d}$$

  et on conclut.

Un petit exemple numérique: On fait un sondage d'intentions de vote auprès de 200 personnes, il y a trois candidats, et on veut tester au niveau 5% l'hypothèse $p_1 = 0.5$, $p_2 = 0.25$ et $p_3 = 0.25$. On observe $N_1 = 120$, $N_2 = 40$, $N_3 = 40$.

Les moyennes théoriques sont $n_1 = 100, n_2 = n_3 = 50$. Elles sont bien supérieures à 5. On observe pour Z:

$$z_{obs} = \frac{20^2}{100} + \frac{10^2}{50} + \frac{10^2}{50} = 4 + 2 + 2 = 8.$$

Dans la table, on trouve que pour la loi χ^2_2,

$$P(Z > 5.99) = 5\%.$$

Ici, on observe $z_{obs} > 5.99$, donc on élimine l'hypothèse H_0. Les écarts entre les valeurs observées et les moyennes théoriques sont trop grands.

4.5 Aménagements

Que faire lorsque pour l'une ou plusieurs des valeurs de i, $n_i < 5$? Ceci ne doit pas être un obstacle pour faire un test qui élimine H_0 si les observations sont trop éloignées des moyennes théoriques. Ce que l'on fait est que l'on regroupe les observations correspondantes aux 'petits n_i' comme une seule et même réponse (qui a une moyenne théorique supérieure à 5), puis on fait le test décrit précédemment. Par exemple, si $d = 5$ et l'on trouve $n_1 = 50$, $n_2 = 25$, $n_3 = 18$, $n_4 = 3$, $n_5 = 4$, on regroupe les réponses 4 et 5. On a alors

$$n_1 = 50,\ n_2 = 25,\ n_3 = 18,\ n_{4 \text{ ou } 5} = 7.$$

On fait alors le test précédent (avec $d = 4$ puisqu'il n'y a alors que 4 choix possibles).

Un autre problème est parfois que l'hypothèse sur H_0 ne donne pas la loi $\mathcal{L}$ de manière explicite et qu'il manque un paramètre pour la déterminer. Par exemple, H_0: "La loi $\mathcal{L}$ est une loi de Poisson". Ici, on ne connait pas le paramètre de la loi de Poisson. Alors, on estime le ou les paramètres manquants (supposons qu'il y en a r) par une méthode ad hoc (la plus simple possible), puis on fait le test avec ces valeurs du paramètre. Mais, à la fin, au lieu d'utiliser la table du χ^2_{d-1}, on utilise celle du χ^2_{d-1-r}. Il faut enlever un degré de liberté par paramètre estimé.

Exemple numérique: Sur 100 jours, on observe $N_0 = 21$ jours sans accident sur la $N118$. Il y a $N_1 = 30$ jours avec 1 accident, 29 jours avec 2 accidents, 14 jours avec 3 accidents et 6 jours avec 4 accidents. On souhaite tester au niveau 5% l'hypothèse H_0: "Le nombre d'accidents dans une journée suit une loi de Poisson".

On commence par estimer le paramètre λ de la loi par la moyenne observée (rappelons que l'espérance d'une variable de Poisson de paramètre λ est λ). On observe

$$\lambda = 1.54 \text{ et donc } e^{-\lambda} = 0.21.$$

Pour une variable de Poisson de paramètre λ, les probabilités sont

$$p_0 = 0.21,\ p_1 = 0.33,\ p_2 = 0.26,\ p_3 = 0.13,\ p_{4 \text{ ou plus}} = 0.07.$$

Ici, on regroupe toutes les observations supérieures à 4 afin d'avoir une moyenne théorique supérieure à 5. On a donc ici

$$n_0 = 21, n_1 = 33, n_2 = 26, n_3 = 13, n_{4 \text{ ou plus}} = 7$$

On observe

$$z_{obs} = 0/21 + 3^2/33 + 3^2/26 + 1^2/13 + 1^2/7 = 0.83.$$

On regarde dans la table de la loi $\chi_{5-1-1} = \chi^2_3$ que

$$P(Z > 7.8) = 5\%.$$

Comme $0.83 < 7.8$, on accepte l'hypothèse H_0.

5 Modèles Gaussiens.

5.1 Exemple introductif.

Imaginons que l'on s'intéresse à la quantité de matière grasse présente dans un produit donné. Par une méthode particulière, on peut mesurer cette quantité de matière grasse, et si l'on répète n fois l'expérience, on obtient n résultats $x_1, \ldots, x_n$. Les résultats ne sont pas tous identiques: la variabilité (faible...) étant due à la méthode de mesure. Deux questions se posent alors: quelle est cette variabilité? Et quelle est, finalement, la quantité de matière grasse présente dans le produit ?

On peut choisir pour ces questions la modélisation gaussienne, c'est à dire supposer que les résultats des n expériences sont les valeurs observées de n variables aléatoires $X_1, \ldots, X_n$ indépendantes de loi gaussienne $\mathcal{L} = \mathcal{N}(m, \sigma^2)$. Ce type de modélisation est licite lorsque l'on pense que les variations observées sont dues à l'accumulation de beaucoup de facteurs indépendants (cf. le théorème central limite). On veut alors répondre à des questions sur les paramètres m (représentant la quantité "réelle" de matière grasse) et σ (représentant la variabilité de la mesure), c'est à dire construire des tests ou des intervalles de confiance. Les situations sont les suivantes:

- m est inconnu et σ^2 est connue (la variabilité de la méthode a été établie par ailleurs): cette situation a déjà été vue en section 3;
- m est connu et σ^2 est inconnue (la quantité de matière grasse est connue pour ce produit, on veut connaitre la variabilité d'une nouvelle méthode de mesure);
- m est inconnu et σ^2 est inconnue (on ne connait ni la quantité de matière grasse, ni la variabilité de la méthode de mesure);

On s'intéressera aussi à la situation où l'on cherche à comparer deux quantités. Par exemple, on considère deux produits dont on veut comparer les taux de matière grasse. On fait une série de mesures pour le premier produit, que l'on modélise comme étant les valeurs observées d'un n-échantillon d'une loi gaussienne $\mathcal{L}_1 = \mathcal{N}(m_1, \sigma^2)$, et une série de mesures pour le deuxième produit, que l'on modélise comme étant les valeurs observées d'un m-échantillon d'une loi gaussienne $\mathcal{L}_2 = \mathcal{N}(m_2, \sigma^2)$. Si l'on procède avec la même méthode de mesure, on peut considérer que les deux lois ont la même variance inconnue σ^2. Dans ce cas, on cherchera à comparer m_1 et m_2.

5.2 Moyenne connue, variance inconnue.

5.2.1 La variable de test

On dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathcal{L} = \mathcal{N}(m_0, \sigma^2)$, où m_0 est connue et σ^2 est inconnue. Puisque

$$\sigma^2 = E\left((X_1 - m_0)^2\right),$$

on peut penser (méthode empirique, section 2.8.3) avoir une idée de σ^2 en utilisant la moyenne empirique des écarts quadratiques

$$V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_0)^2. \tag{1}$$

pour lesquels on a $E(V^2) = \sigma^2$.

On remarque que

$$\frac{nV^2}{\sigma^2} = \left(\frac{X_1 - m_0}{\sigma}\right)^2 + \left(\frac{X_2 - m_0}{\sigma}\right)^2 + \ldots + \left(\frac{X_n - m_0}{\sigma}\right)^2,$$

et que les $(X_i - m_0)/\sigma$ sont indépendantes de loi $\mathcal{N}(0,1)$, donc:

Proposition 10 *nV^2/σ^2 suit la loi χ^2_n.*

Cette proposition est l'outil de base pour construire tests et intervalles de confiance sur σ lorsque m_0 est connue. Pour tester l'hypothèse "la valeur de σ^2 est 1", on regardera si la valeur observée de $nV^2/1$ est plausible pour la loi χ^2_n.

5.2.2 Tests sur σ

Soit à tester H_0 : "$\sigma = \sigma_0$" contre H_1 : "$\sigma > \sigma_0$". On choisit comme variable de test $Z = nV^2/\sigma_0^2$, qui a tendance à être plus grande sous H_1 que sous H_0. On décide donc de rejeter H_0 quand Z sera plus grand qu'une valeur seuil z_0. Puisque sous H_0, Z suit une loi χ^2_n, on lit z_0 dans la table du χ^2 à n degrés de liberté, de telle sorte que si on choisit un niveau α,

$$P_{H_0}(Z > z_0) = \alpha.$$

On calcule ensuite la valeur observée de Z, et on la compare à z_0: si elle est plus grande, on rejette H_0; sinon, on accepte H_0.

Exemple: on veut connaitre la variabilité de la mesure de taux de matière grasse. Pour cela, on effectue des mesures (en grammes pour 100 grammes) d'un produit dont on sait qu'il contient 60.9 grammes par kilogramme de matière grasse. On obtient les résultats suivants: 60.85; 61.05; 60.95; 60.80; 60.90. On veut tester H_0 : "$\sigma = 0.1$" contre H_1 : "$\sigma > 0.1$". Avec $n = 5$ et au niveau $\alpha = 5\%$, on lit dans la table χ^2_5: $z_0 = 11.070$. On calcule

$$\begin{aligned} Z_{obs} &= \frac{1}{0.01} \times \big((60.85 - 60.9)^2 + (61.05 - 60.9)^2 \\ &\quad + (60.95 - 60.9)^2 + (60.80 - 60.9)^2 + (60.90 - 60.9)^2\big) \\ &= 3.75 \end{aligned}$$

Comme $z_0 > 3.75$, on accepte H_0.

5.2.3 Intervalles de confiance pour σ

L'objectif ici est de donner un intervalle pour lequel on puisse affirmer, avec une probabilité connue, que la variabilité de la mesure, c'est à dire σ, y appartient.

Rappelons qu'on peut construire un intervalle de confiance pour σ au niveau $1-\alpha$ comme l'ensemble des valeurs de σ_0 pour lequel on accepte H_0 : "$\sigma = \sigma_0$" contre H_1 : "$\sigma \neq \sigma_0$" au niveau α.

Soit donc σ_0 une valeur, et soit à tester H_0 : "$\sigma = \sigma_0$" contre H_1 : "$\sigma \neq \sigma_0$" au niveau α. Comme dans le test précédent, on choisit comme variable de test $Z = nV^2/\sigma_0^2$, qui se décale à droite ou à gauche sous H_1, et qui suit une loi χ^2_n sous H_0. On décide donc de rejeter H_0 quand Z sera plus petit qu'une valeur seuil z_0 ou plus grand qu'une autre valeur seuil z_1. Puisque sous H_0 Z suit une loi χ^2_n, on lit z_0 et z_1 dans la table du χ^2 à n degrés de liberté, de telle sorte que si on choisit un niveau α,

$$P_{H_0}\left(Z < z_0 \text{ ou } Z > z_1\right) = \alpha.$$

Ceci peut se réécrire: si σ_0 est la valeur de σ,

$$P_{\sigma_0}\left(z_0 \leq \frac{nV^2}{\sigma_0^2} \leq z_1\right) = 1-\alpha. \tag{2}$$

On accepte H_0 si $z_0 \leq Z \leq z_1$, autrement dit si $z_0 \leq \frac{nV^2}{\sigma_0^2} \leq z_1$. On voit que le système d'inéquation $z_0 \leq \frac{nV^2}{\sigma_0^2} \leq z_1$ est équivalent à $\frac{nV^2}{z_1} \leq \sigma_0^2 \leq \frac{nV^2}{z_0}$, on accepte donc H_0 si $\sqrt{\frac{nV^2}{z_1}} \leq \sigma_0 \leq \sqrt{\frac{nV^2}{z_0}}$. Un intervalle de confiance pour σ au niveau α est donc

$$I = \left[\sqrt{\frac{nV^2}{z_1}}; \sqrt{\frac{nV^2}{z_0}}\right].$$

On remarque au passage que (2) se réécrit en: pour toute valeur de σ_0, on a

$$P_{\sigma_0}\left(\sqrt{\frac{nV^2}{z_1}} \leq \sigma_0 \leq \sqrt{\frac{nV^2}{z_0}}\right) = 1-\alpha. \tag{3}$$

Ceci correspond à la définition alternative d'un intervalle de confiance donnée en section 3.5. Bien remarquer que dans la définition, un intervalle de confiance est un **intervalle aléatoire**, pour lequel on calcule une **valeur observée** I_{obs}.

Exemple: au niveau de confiance 95%, pour $n = 5$ on lit $z_0 = 0.831$ et $z_1 = 12.833$. On a l'intervalle de confiance $I = [\sqrt{\frac{5V^2}{12.833}}; \sqrt{\frac{5V^2}{0.831}}]$. La valeur observée est $I_{obs} = [0.054; 0.213]$.

5.3 Moyenne inconnue, variance inconnue

5.3.1 La variable de test pour la variance

On se place maintenant dans le cas où l'on dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathcal{L} = \mathcal{N}(m, \sigma^2)$, où m est inconnue et σ^2 est inconnue. En utilisant l'idée de "moyenne empirique", on peut penser avoir une idée de m à l'aide de

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{4}$$

qui a l'avantage d'être centrée en m, mais dont la loi $\mathcal{N}(m, \sigma^2/n)$ dépend du paramètre inconnu σ^2. Pour avoir une idée de σ^2, on peut penser remplacer dans V^2 le paramètre inconnu m par $\overline{X}$. On définit

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2. \tag{5}$$

En divisant par $n-1$ (et non pas n), un calcul relativement simple montre qu'on obtient un estimateur sans biais de σ^2, c'est à dire que

$$E(S^2) = \sigma^2.$$

En fait, pour tester ou construire des intervalles de confiance pour σ, on utilise le résultat théorique suivant:

Proposition 11 *$(n-1)S^2/\sigma^2$ suit la loi χ^2_{n-1}.*

Pour construire des tests ou des intervalles de confiance pour σ quand m est inconnue, on procède donc comme dans le cas où m est connue, en remplaçant V^2 par S^2 et n par $n-1$.

5.3.2 La loi de Student (et la variable de test pour l'espérance).

On définit maintenant une loi de probabilité qui nous servira pour "libérer" la loi de $\overline{X}$ du paramètre inconnu σ. Suposons que U soit une variable de loi normale centrée réduite $\mathcal{N}(0,1)$, et V une variable indépendante de U et de loi du χ^2 à d degrés de liberté. On pose alors

$$Z = \frac{U}{\sqrt{V/d}}.$$

La loi de Z s'appelle la loi de Student à d degrés de liberté. On la note $\mathcal{T}_d$. Notons que la loi de Z est symétrique (comme la loi de U): la densité de $\mathcal{T}_d$ est paire. Lorsque d tend vers l'infini, la loi $\mathcal{T}_d$ tend vers la loi normale centrée réduite $\mathcal{N}(0,1)$. En pratique, pour calculer des probabilités, si $d \leq 30$, on utilise les tables de valeurs numériques. Si $d \geq 30$, on peut utiliser l'approximation par la loi normale centrée réduite.

Si, avec notre échantillon gaussien, on définit

$$U = \frac{\sqrt{n}(\overline{X} - m)}{\sigma} \text{ et } V = \frac{(n-1)S^2}{\sigma^2},$$

U suit la loi normale centrée réduite $\mathcal{N}(0,1)$ et V suit la loi du χ^2 à $n-1$ degrés de liberté. Si ces deux variables sont indépendantes, alors $Z = \frac{U}{\sqrt{V/n-1}}$ suit la loi de Student à $n-1$ degrés de liberté. En fait, c'est le cas. On remarque que dans Z, σ se simplifie en haut et en bas de la barre de fraction. Le théorème suivant est l'outil de base pour construire tests et intervalles de confiance pour m lorsque σ est inconnu.

Proposition 12 *Les variables aléatoires $\overline{X}$ et S^2 sont indépendantes. La variable $\frac{\sqrt{n}(\overline{X}-m)}{S}$ suit la loi $\mathcal{T}_{n-1}$.*

En résumé, on peut donc faire des tests séparément sur les valeurs de m et de σ^2. Pour m on utilisera la proposition ci-dessus, pour σ^2, on employera la procédure décrite précédemment (avec la loi du χ^2).

5.3.3 Tests sur m

Supposons que, pour ce produit que l'on croyait contenir 60.90% de matière grasse, on se demande si c'est bien le cas, ou si en fait le taux n'est pas plus grand. Il s'agit là de tester H_0: "$m = m_0$" contre H_1: "$m > m_0$" pour une valeur fixée m_0. On choisit comme variable de test $Z = \sqrt{n}(\overline{X} - m_0)/S$, qui a tendance à être plus grande sous H_1 que sous H_0. On décide donc de rejeter H_0 quand Z sera plus grand qu'une valeur seuil z_0. Puisque, sous H_0, Z suit la loi $\mathcal{T}_{n-1}$, on lit z_0 dans la table de Student à $n-1$ degrés de liberté, de telle sorte que si on choisit un niveau α,

$$P_{H_0}(Z > z_0) = \alpha.$$

On calcule ensuite la valeur observée de Z, et on la compare à z_0: si elle est plus grande, on rejette H_0; sinon, on accepte H_0.

Exemple: Avec les données précédentes, on veut tester H_0 : "$m = 60.9$" contre H_1 : "$m > 60.9$". Avec $n = 5$ et au niveau $\alpha = 5\%$, on lit dans la table $\mathcal{T}_4$ $z_0 = 2.132$. On calcule $\overline{X}_{obs} = 60.91$, puis $4S^2_{obs} = 0.037$, d'où $Z_{obs} = \frac{\sqrt{5}\cdot 0.01}{\sqrt{0.037/4}}$, soit $Z_{obs} = 0.232$, donc on accepte H_0.

5.3.4 Intervalles de confiance pour m

On va construire un intervalle de confiance pour m au niveau de confiance $1-\alpha$, en utilisant la deuxième définition équivalente que l'on vient de voir. On vérifiera ensuite (pour cette fois-ci, ensuite on admettra que les deux méthodes sont équivalentes), en construisant l'intervalle par la première méthode que l'on obtient le même résultat.

Pour toute valeur m, on sait, d'après la proposition, que $\sqrt{n}(\overline{X}-m)/S$ suit la loi $\mathcal{T}_{n-1}$. En lisant dans la table de Student à $n-1$ degrés de liberté, et sous $1-\alpha/2$, puisque la loi de Student est symétrique, on trouve z_0 tel que

$$P_m\left(-z_0 \leq \frac{\sqrt{n}(\overline{X}-m)}{S} \leq z_0\right) = 1-\alpha. \tag{6}$$

On voit que le système d'inéquations $-z_0 \leq \sqrt{n}(\overline{X}-m)/S \leq z_0$ est équivalent à $\overline{X} - z_0 \cdot S/\sqrt{n} \leq m \leq \overline{X} + z_0 \cdot S/\sqrt{n}$, et qu'on a donc, pour toute valeur de m,

$$P_m\left(\overline{X} - \frac{z_0 \cdot S}{\sqrt{n}} \leq m \leq \overline{X} + \frac{z_0 \cdot S}{\sqrt{n}}\right) = 1-\alpha,$$

autrement dit que

$$I = \left[\overline{X} - \frac{z_0 \cdot S}{\sqrt{n}}; \overline{X} + \frac{z_0 \cdot S}{\sqrt{n}}\right] \tag{7}$$

est un intervalle de confiance pour m au niveau de confiance $1-\alpha$.

Regardons maintenant l'autre méthode. Soit donc m_0 une valeur, et soit à tester H_0 : "$m = m_0$" contre H_1 : "$m \neq m_0$" au niveau α. Comme dans le test précédent, on choisit comme variable de test $Z = \sqrt{n}(\overline{X}-m_0)/S$, qui se décale à droite ou à gauche sous H_1, et qui suit une loi $\mathcal{T}_{n-1}$ sous H_0. On décide donc de rejeter H_0 quand Z sera plus petit qu'une valeur seuil ou plus grand qu'une autre valeur seuil. Puisque sous H_0 Z suit la loi $\mathcal{T}_{n-1}$, qui est symétrique, on lit dans la table de Student à $n-1$ degrés de liberté, de telle sorte que si on choisit un niveau α,

$$P_{H_0}(Z < -z_0 \text{ ou } Z > z_0) = \alpha.$$

On accepte H_0 si $-z_0 \leq Z \leq z_0$, autrement dit si $\overline{X} - z_0 \cdot S/\sqrt{n} \leq m_0 \leq \overline{X} + z_0 \cdot S/\sqrt{n}$. On voit donc que l'ensemble des valeurs m_0 pour lesquelles on accepte H_0 est bien l'intervalle I donné en (7).

Exemple: reprenons nos données pour calculer un intervalle de confiance pour le taux de matière grasse au niveau 95%. Dans la table de la loi de Student à 4 degrés de liberté, on lit $z_0 = 2.776$. Un intervalle de confiance pour le taux de matière grasse avec 5 mesures est donc $I = [\overline{X} - 2.776 \cdot S/\sqrt{5}; \overline{X} + 2.776 \cdot S/\sqrt{5}]$. La valeur observée est $I_{obs} = [60.79; 61.03]$.

5.4 Comparaison de deux moyennes.

On dispose de deux échantillons: un n-échantillon $X_1, \ldots, X_n$ de loi $\mathcal{L} = \mathcal{N}(m_1, \sigma_1^2)$, et un m-échantillon $Y_1, \ldots, Y_m$ de loi $\mathcal{L} = \mathcal{N}(m_2, \sigma_2^2)$. Tous les paramètres, m_1, m_2, σ_1 et σ_2 sont inconnus. On fait l'hypothèse suivante:

Hypothèse. Les deux échantillons ont même variance, c'est à dire $\sigma_1 = \sigma_2$.

On notera maintenant σ cette valeur commune. Dans certains cas, cette hypothèse est valide de manière évidente: c'est le cas de l'exemple proposé, où la loi gaussienne modélise une erreur de mesure, et où, pour les deux produits, le taux de matière grasse est mesuré avec le même procédé expérimental. Dans d'autres cas, c'est une hypothèse qu'il faut valider (on n'apprendra pas cette année comment la valider, mais il faut le faire en situation réelle!). Prenons un autre exemple. Imaginons que l'on souhaite savoir si, pour une (grande) population donnée, le poids à la naissance d'un nouveau-né (né à terme) est en moyenne le même pour un garçon et pour une fille. On pèse alors n nouveaux-nés garçons (nés à terme) et m nouveaux-nés filles (nées à terme). On peut modéliser le résultat des n premières pesées comme étant les réalisations de n variables gaussiennes indépendantes de loi $\mathcal{L} = \mathcal{N}(m_1, \sigma^2)$, et le résultat des m deuxièmes pesées comme étant les réalisations de m variables gaussiennes indépendantes de loi $\mathcal{L} = \mathcal{N}(m_2, \sigma^2)$. Ici, les significations des m_i et de σ ne sont pas les mêmes que dans l'exemple précédent: m_1 représente la moyenne des poids des nouveaux-nés garçons (nés à terme) dans **toute** la population considérée, et σ représente la variabilité du poids d'un nouveau-né garçon (né à terme) dans cette population. De même pour m_2 et $\sigma_2 = \sigma$. Supposer $\sigma = \sigma_1 = \sigma_2$, ici, veut dire que la variabilité du poids d'un nouveau-né (né à terme) dans toute la population est la même pour les garçons et pour les filles.

Pour avoir une idée de m_1, on peut penser utiliser $\overline{X} = (\sum_{i=1}^{n} X_i)/n$ qui a pour loi $\mathcal{N}(m_1, \sigma^2/n)$, et pour avoir une idée de m_2, on peut penser utiliser $\overline{Y} = (\sum_{i=1}^{m} Y_i)/m$ qui suit la loi $\mathcal{N}(m_2, \sigma^2/m)$. Pour pouvoir les utiliser, il faut se "libérer" du paramètre inconnu σ^2.

Pour avoir une idée de σ^2, on peut penser utiliser

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

puisque l'on sait que $(n-1)S_X^2/\sigma^2$ suit la loi χ^2_{n-1}. Mais de la même manière, on pourrait aussi penser utiliser

$$S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2,$$

puisque les X_i et les Y_j ont même variance. Pour utiliser toute l'information disponible, il est plus efficace d'utiliser les deux. Posons

$$V = (n-1)S_X^2 + (m-1)S_Y^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + \sum_{i=1}^{m} (Y_i - \overline{Y})^2.$$

Quand on additionne deux variables indépendantes de loi de χ^2, on obtient une variable de loi de χ^2 dont le nombre de degrés de liberté est la somme des degrés de liberté de chacune des variables additionnée (revoir la définition de la loi du χ^2 pour s'en convaincre). Les X_i et les Y_j étant des variables indépendantes, S_X^2 et S_Y^2 sont des variables indépendantes, et donc

Proposition 13 *La variable V/σ^2 suit la loi du χ^2 à $n+m-2$ degrés de liberté.*

Maintenant, pour comparer m_1 et m_2, on peut utiliser la variable $\overline{X}-\overline{Y}$, qui suit la loi $\mathcal{N}(m_1-m_2, \sigma^2(\frac{1}{n}+\frac{1}{m}))$. Si l'on pose

$$U = \frac{(\overline{X}-\overline{Y}-(m_1-m_2))}{\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

alors U/σ suit la loi normale centrée réduite $\mathcal{N}(0,1)$. On obtient alors le théorème suivant, qui est l'outil de base pour comparer les moyennes de deux échantillons gaussiens de même variance inconnue:

Proposition 14 *La variable $U\sqrt{n+m-2}/\sqrt{V}$ suit la loi $\mathcal{T}_{n+m-2}$.*

Soit par exemple à tester H_0: "$m_1 = m_2$" contre $m_1 > m_2$. On choisit comme variable de test

$$Z = \frac{(\overline{X}-\overline{Y})}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{V}{n+m-2}}},$$

qui a tendance à être concentrée autour de 0 sous H_0, et à être plus grande sous H_1 que sous H_0. On décide donc de rejeter H_0 quand Z sera plus grand qu'une valeur seuil z_0. Puisque sous H_0 Z suit la loi $\mathcal{T}_{n+m-2}$, on lit z_0 dans la table de Student à $n+m-2$ degrés de liberté, de telle sorte que si on choisit un niveau α,

$$P_{H_0}(Z > z_0) = \alpha.$$

On calcule ensuite la valeur observée de Z, et on la compare à z_0: si elle est plus grande, on rejette H_0; sinon, on accepte H_0.

Si on veut construire un intervalle de confiance pour $m_1 - m_2$ au niveau de confiance $1-\alpha$, on écrit que pour toutes valeurs m_1 et m_2, on sait que $U\sqrt{n+m-2}/\sqrt{V}$ suit la loi $\mathcal{T}_{n+m-2}$. En lisant dans la table de Student à $n+m-2$ degrés de liberté, et sous $1-\alpha/2$, puisque la loi de Student est symétrique, on trouve z_0 tel que

$$P_{m_1,m_2}\left(-z_0 \leq \frac{U\sqrt{n+m-2}}{\sqrt{V}} \leq z_0\right) = 1-\alpha. \tag{8}$$

En écrivant le système d'inéquations de manière équivalente, si l'on pose

$$I = \left[\overline{X}-\overline{Y}-z_0\sqrt{\frac{(1/n+1/m)V}{n+m-2}}, \overline{X}-\overline{Y}+z_0\sqrt{\frac{(1/n+1/m)V}{n+m-2}}\right]$$

alors pour toutes valeurs de m_1 et m_2,

$$P_{m_1,m_2}((m_1-m_2)\in I) = 1-\alpha,$$

autrement dit, I est un intervalle de confiance pour $m_1 - m_2$ au niveau de confiance $1-\alpha$.

6 Couples de variables aléatoires.

Un couple de variables aléatoires est une variable aléatoire (X, Y) à valeur dans un ensemble à deux dimensions (avec une abscisse et une ordonnée), c'est à dire un couple de deux nombres, résultat d'une expérience à l'issue incertaine. Par exemple, une urne contient trois boules numérotées 1,2 et 3. On tire successivement et sans remise deux boules de l'urne. Soit X le numéro obtenu au premier tirage, et Y le résultat obtenu au deuxième tirage. (X, Y) est alors un couple de variables aléatoires discrètes. Ou bien on tire au hasard un individu dans une grande population, et on note sa taille X et son poids Y. (X, Y) est alors un couple de variables aléatoires continues.

6.1 Loi d'un couple de variables aléatoires

Comme dans le cas d'une seule variable aléatoire, on va s'intéresser à décrire la loi du couple (X, Y). Nous ferons cette description dans deux cas différents.

6.1.1 Cas discret

C'est le cas où l'ensemble des valeurs que peut prendre (X, Y) est un ensemble fini ou dénombrable E de valeurs. La loi de (X, Y) est alors la donnée des nombres $p_{(x,y)}$ où (x, y) parcourt l'ensemble E. Pour chaque résultat possible (x, y), le nombre $p_{(x,y)}$ représente la probabilité pour que $X = x$ et $Y = y$. On a

$$p_{(x,y)} \in [0, 1] \text{ et } \sum_{(x,y)\in E} p_{(x,y)} = 1.$$

Remarquer que ceci est simplement la réécriture de ce qui a été vu pour une variable aléatoire discrète, dans le cas où cette variable s'écrit comme un couple de variables: il n'y a ici **rien de réellement nouveau.**

La loi permet de calculer les probabilités d'événements relatifs à (X, Y): si A est une partie de E,

$$P\left((X, Y) \in A\right) = \sum_{(x,y)\in A} p_{(x,y)}.$$

Si E est fini, on peut résumer la loi de (X, Y) dans un tableau.

Exemple de l'urne. Pour calculer la probabilité dobtenir un résultat (x, y) particulier, on peut utiliser la formule des probabilités composées (**rappel**):

$$P\left(X = x \text{ et } Y = y\right) = P\left(X = x\right) P\left(Y = y | X = x\right).$$

La loi de (X, Y) est donnée dans le tableau ci-dessous:

x - y	1	2	3
1	0	1/6	1/6
2	1/6	0	1/6
3	1/6	1/6	0

6.1.2 Cas continu

C'est le cas où l'ensemble des valeurs que peut prendre (X, Y) est une partie de $\mathbb{R}^2$. La loi de (X, Y) peut alors être identifiée par sa densité de probabilité $f(x, y)$. Comme dans le cas d'une variable X, on a:

$$f(x, y) \geq 0 \text{ et } \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy = 1.$$

La loi permet de calculer les probabilités d'événements relatifs à (X, Y): si A est une partie de $\mathbb{R}^2$,

$$P\left((X, Y) \in A\right) = \int \int_{(x,y)\in A} f(x, y) dx dy.$$

Exemple: Soit (X, Y) un couple de densité
$f(x, y) = \begin{cases} 6xy^2 & \text{si } 0 \leq x \leq 1 \text{ et } 0 \leq y \leq 1 \\ 0 & \text{sinon} \end{cases}$
On a bien $f(x, y) \geq 0$ pour tous x et y, et

$$\begin{aligned} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 6xy^2 dx dy \\ &= \int_0^1 6x \left[\frac{y^3}{3}\right]_0^1 dx \\ &= \int_0^1 2x dx \\ &= 1. \end{aligned}$$

Si on veut par exemple calculer $P(X < Y)$:

$$\begin{aligned} P(X < Y) &= \int \int_{x<y} f(x, y) dx dy \\ &= \int_0^1 y^2 dy \int_0^y 6x dx \\ &= \int_0^1 y^2 dy \left[3x^2\right]_0^y \\ &= \int_0^1 3y^4 dy \\ &= 3/5. \end{aligned}$$

6.2 Lois marginales.

Si (X, Y) est un couple de variables aléatoires, X est une variable aléatoire, ainsi que Y. On appelle **loi marginale** de X la loi de la variable aléatoire X. De même, on appelle **loi marginale** de Y la loi de la variable aléatoire Y. Du coup, pour distinguer, on parlera de **loi jointe** pour désigner la loi du couple (X, Y).

6.2.1 Cas discret

Calculer la loi marginale de X, c'est calculer, pour toutes les valeurs x de X possibles, la quantité $P(X = x)$. On a:

$$P(X = x) = \sum_y P(X = x \text{ et } Y = y)$$

et donc la loi marginale de X est donnée par les nombres $q_x = P(X = x)$:

$$q_x = \sum_y p_{(x,y)}.$$

De même, la loi marginale de Y est donnée par les nombres $r_y = P(Y = y)$:

$$r_y = \sum_x p_{(x,y)}.$$

On remarque que si la loi de (X, Y) est donnée par un tableau avec les x en colonne et les y en ligne, calculer la loi marginale de X c'est faire la somme des lignes, et calculer la loi marginale de Y c'est faire la somme des colonnes. On peut alors donner le résultat dans le même tableau.
Exemple de l'urne (continué):

x - y	1	2	3	q_x
1	0	1/6	1/6	1/3
2	1/6	0	1/6	1/3
3	1/6	1/6	0	1/3
r_y	1/3	1/3	1/3	

6.2.2 Cas continu

La loi marginale de X peut alors être identifiée par sa densité de probabilité $g(x)$. Par analogie avec le cas discret, on a:

$$g(x) = \int_{-\infty}^{+\infty} f(x, y)dy$$

On remarque que g est bien une densité de probabilité: $g(x) \geq 0$ pour tout x, et

$$\int_{-\infty}^{+\infty} g(x)dx = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dydx = 1.$$

De même, la loi marginale de Y peut être identifiée par sa densité de probabilité $h(y)$, et l'on a:

$$h(y) = \int_{-\infty}^{+\infty} f(x, y)dx.$$

Exemple (continué): la densité marginale de X est

$$g(x) = \begin{cases} \int_0^1 6xy^2dy = 6x[\frac{y^3}{3}]_0^1 = 2x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{sinon} \end{cases}$$

La densité marginale de Y est

$$h(y) = \begin{cases} \int_0^1 6xy^2 dx = 6y^2[\frac{x^2}{2}]_0^1 = 3y^2 & \text{si } 0 \leq y \leq 1 \\ 0 & \text{sinon} \end{cases}$$

6.3 Calculs d'espérances.

Pour calculer l'espérance d'une fonction de (X, Y) on procède comme on a déjà vu.

6.3.1 Cas discret.

Soit k une fonction de E dans $\mathbb{R}$. Alors

$$E\left(k(X,Y)\right) = \sum_{(x,y)\in E} k(x,y)p_{(x,y)}.$$

6.3.2 Cas continu.

Soit k une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$. Alors

$$E\left(k(X,Y)\right) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k(x,y)f(x,y)dxdy.$$

6.4 Covariance, corrélation.

La loi jointe du couple (X, Y) décrit complètement comment les variables X et Y sont liées. On définit deux nombres, la covariance entre X et Y et la corrélation entre X et Y, qui seront des "indicateurs de liaison" des deux variables.

Définition. La covariance entre X et Y est le nombre

$$\mathrm{Cov}(X,Y) = E\left[(X - E(X))\,(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

On peut remarquer que si l'on prend $Y = X$, on obtient

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X),$$

et que si l'on prend $Y = -X$, on obtient

$$\mathrm{Cov}(X,-X) = -\mathrm{Var}(X).$$

Remarque: pour calculer $E(X)$, on peut utiliser la formule avec la loi jointe, ou celle avec la loi marginale (de même pour $E(Y)$).

Exemples (continués): Pour l'exemple du cas discret, on obtient $E(X) = E(Y) = 2$ et $E(XY) = 11/3$, et donc $\mathrm{Cov}(X,Y) = 11/3 - 4 = -1/3$. Pour l'exemple du cas continu, on obtient $E(X) = 2/3$, $E(Y) = 3/4$ et $E(XY) = 1/2$, et donc $\mathrm{Cov}(X,Y) = 0$.

Définition. La corrélation entre X et Y est le nombre

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

On peut remarquer que si Y est une fonction affine de X de coefficient directeur positif, c'est à dire si il existe deux réels a et b, avec $a > 0$, tels que $Y = aX + b$, on obtient

$$\text{Corr}(X, aX + b) = 1.$$

De même si Y est une fonction affine de X de coefficient directeur négatif, c'est à dire si il existe deux réels a et b, avec $a > 0$, tels que $Y = -aX + b$, on obtient

$$\text{Corr}(X, -aX + b) = -1.$$

En fait, ces deux cas sont extrêmes, et l'on a:

Proposition 15 *On a toujours*

$$-1 \leq Corr(X,Y) \leq 1.$$

$Corr(X,Y) = 1$ si et seulement si Y est une fonction affine de X de coefficient directeur positif, et $Corr(X,Y) = -1$ si et seulement si Y est une fonction affine de X de coefficient directeur négatif.

On peut interpréter le coefficient de corrélation comme un indice de liaison affine entre deux variables. Si ce coefficient est positif, celà signifie que le fait que X soit plutôt plus grand que d'ordinaire à tendance à indiquer que Y est aussi plutôt grand. Mais ceci n'est qu'une "indication". **Seule la loi jointe décrit complètement le lien entre les deux variables:** on peut avoir un coefficient de corrélation nul alors que les variables ne sont pas indépendantes.

6.5 Critère d'indépendance

On sait que deux variables discrètes X et Y sont indépendantes si et seulement si pour tous x et y on a $P(X = x \text{ et } Y = y) = P(X = x)P(Y = y)$. Autrement dit, si pour tous x et y on a $p_{(x,y)} = q_x r_y$. On peut résumer cela en:

Proposition 16 *Les variables X et Y sont indépendantes si et seulement si la loi jointe de (X,Y) est égale au produit des lois marginales de X et Y.*

Dans le cas de variables continues, on peut aussi retenir ce critère, qui dit que X et Y sont indépendantes si et seulement si pour tous x et y on a $f(x,y) = g(x)h(y)$.

On remarque que, comme l'espérance d'un produit de variables indépendantes est égale au produit des espérance, si X et Y sont indépendantes on a $E(XY) = E(X)E(Y)$, et donc

Proposition 17 *Si X et Y sont indépendantes, $Cov(X,Y) = 0$. Attention, la réciproque est fausse: on peut avoir $Cov(X,Y) = 0$ alors que les variables X et Y ne sont pas indépendantes.*

Exemples (continués). Dans l'exemple discret, les variables X et Y ne sont pas indépendantes car par exemple

$$P(X = 1 \text{ et } Y = 1) = 0 \neq P(X = 1)P(Y = 1).$$

Dans l'exemple continu, X et Y sont indépendantes car on a bien $f(x, y) = g(x)h(y)$.

6.6 Loi conditionnelle

On a une autre façon de décrire le lien entre deux variables aléatoires: les lois condtionnelles.

6.6.1 Cas discret

La loi conditionnelle de X conditionnellement à $Y = y$ est la loi de probabilité $(q_x^{Y=y})_x$ donnée par: pour tout x,

$$q_x^{Y=y} = \frac{p_{x,y}}{r_y}.$$

C'est le quotient de la loi jointe par la loi marginale de Y. On note que

$$q_x^{Y=y} = P(X = x|Y = y).$$

On vérifie que c'est bien une probabilité: pour tout x, $0 \leq q_x^{Y=y} \leq 1$, et $\sum_x q_x^{Y=y} = 1$.

6.6.2 Cas continu

La loi conditionnelle de X conditionnellement à $Y = y$ est la loi de densité $g^{Y=y}(x)$ donnée par: pour tout x,

$$g^{Y=y}(x) = \frac{f(x, y)}{h(y)}.$$

C'est le quotient de la densité jointe par la densité marginale de Y. On vérifie que c'est bien une densité de probabilité: pour tout x, $0 \leq g^{Y=y}(x)$, et

$$\int_0^\infty g^{Y=y}(x)dx = 1.$$

Remarque: les variables X et Y sont indépendantes si et seulement si pour tout y, la loi de X conditionnelle à $Y = y$ ne dépend pas de y, autrement dit si et seulement si pour tout y, la loi de X conditionnelle à $Y = y$ est la loi marginale de X.

6.7 Test du χ^2 d'indépendance

On considère $(X_1, Y_1), \ldots, (X_n, Y_n)$ un n-échantillon d'une loi d'un couple de variables (X, Y). On veut tester l'hypothèse H_0: " les variables X et Y sont indépendantes" contre H_1: " les variables X et Y ne sont pas indépendantes".
On se place dans le cas où X peut prendre k valeurs différentes $x_1, \ldots, x_k$, et Y peut prendre l valeurs différentes $y_1, \ldots, y_l$, de sorte que (X, Y) peut prendre kl valeurs différentes. Si l'on note $N_{i,j}$ le nombre de fois, dans l'échantillon, où (X, Y) prend la valeur (x_i, y_j), il est naturel d'estimer $p_{(x_i,y_j)}$ par $N_{i,j}/n$.
Si l'on note N_i le nombre de fois, dans l'échantillon, où X prend la valeur x_i, on a

$$N_i = \sum_{j=1}^{l} N_{i,j},$$

et il est naturel d'estimer $P(X = x_i)$ par N_i/n. De même, si l'on note N^j le nombre de fois, dans l'échantillon, où Y prend la valeur y_j, on a $N^j = \sum_{i=1}^{k} N_{i,j}$, et il est naturel d'estimer $P(Y = y_j)$ par N^j/n. Si l'on suppose X et Y indépendantes, il est alors naturel d'estimer $p_{(x_i,y_j)}$ par $(N_i/n)(N^j/n)$.

C'est sur cette idée qu'est construit le test du χ^2 d'indépendance. Le résultat de base pour ce test est:

Proposition 18 *Lorsque n est grand, la loi de*

$$Z = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\frac{N_i N^j}{n} - N_{i,j})^2}{\frac{N_i N^j}{n}}$$

sous H_0 est approximativement la loi du χ^2 à $(k-1)(l-1)$ degrés de liberté. De plus, sous H_1, Z tend vers l'infini avec n.

Pour appliquer le résultat en pratique, on vérifiera que les $N_{i,j}$ sont supérieurs à 5. Il est clair que l'on décidera de rejeter H_0 pour de trop grandes valeurs de Z, autrement dit la région de rejet est $Z > z_0$, et que, pour un niveau α fixé, on lit z_0 dans la table du $\chi^2_{(k-1)(l-1)}$. On calcule ensuite la valeur observée de Z, et on la compare à z_0.

6.8 Test du χ^2 d'homogénéité.

On considère maintenant deux échantillons, $X_1, \ldots, X_n$, un n-échantillon d'une loi d'une variable X, et $Y_1, \ldots, Y_m$, un m-échantillon d'une loi d'une variable Y . On veut tester l'hypothèse H_0: " les variables X et Y ont même loi" contre H_1: " les variables X et Y n'ont pas même loi".
On se place dans le cas où X et Y peuvent prendre k valeurs différentes (les valeurs possibles pour X et pour Y sont les mêmes) $a_1, \ldots, a_k$. On supposera que les X_i et les Y_j sont **indépendants**. Si l'on note N_i le nombre de fois, dans le premier échantillon, où X prend la valeur a_i, il est naturel d'estimer $P(X = a_i)$ par N_i/n. Si l'on note M_i le nombre de fois, dans l'échantillon, où Y prend la valeur a_i, il est naturel d'estimer $P(Y = a_i)$ par M_i/n.

Si maintenant on suppose que X et Y ont même loi, alors on peut compter dans les deux échantillons le nombre de fois où a_i est observé, pour estimer $P(X = a_i)$ (qui est aussi $P(Y = a_i)$) par $(N_i + M_i)/(n + m)$.
C'est sur cette idée qu'est construit le test du χ^2 d'homogénéité. Le théorème de base pour ce test est:

Proposition 19 *Lorsque n et m sont grands, la loi de*

$$Z = \sum_{i=1}^{k} n\frac{(\frac{N_i+M_i}{n+m} - \frac{N_i}{n})^2}{\frac{N_i+M_i}{n+m}} + m\frac{(\frac{N_i+M_i}{n+m} - \frac{M_i}{m})^2}{\frac{N_i+M_i}{n+m}}$$

sous H_0 est approximativement la loi du χ^2 à $(k-1)$ degrés de liberté.
Sous H_1, Z tend vers l'infini avec n.

Pour appliquer ces résultats, on vérifie en pratique que les N_i et les M_i sont supérieurs à 5.

Il est clair que l'on décidera de rejeter H_0 pour de trop grandes valeurs de Z, autrement dit la région de rejet est $Z > z_0$, et que, pour un niveau α fixé, on lit z_0 dans la table du $\chi^2_{(k-1)}$. On calcule ensuite la valeur observée de Z, et on la compare à z_0.

7 Récapitulatif des différents tests

7.1 Construction générale d'un test: plan de rédaction.

Pour construire un test, il faut suivre les étapes indiquées:

1. **Le modèle.** A partir des données de l'exercice, déterminer dans quel cadre on se situe, c'est à dire préciser le modèle:
   Cadre 1 : On considère un échantillon d'une loi d'une variable aléatoire X ou (X, Y) et on veut faire un test portant sur cette loi (par exemple sur des paramètres inconnus).
   Cadre 2 : On considère deux échantillons, dont un d'une loi d'une variable aléatoire X et l'autre d'une variable Y, où l'on suppose que X et Y sont indépendantes. On veut faire un test comparant les lois de X et Y.
2. **Détermination des hypothèses (H_0) et (H_1).** Au vu de la question posée dans l'exercice, déterminer quelles seront les hypothèses (H_0) et (H_1) du test. Attention: il faut choisir pour H_0 l'hypothèse sous laquelle on sait faire des calculs.
3. **Choix de la statistique de test.** En fonction du cadre de l'exercice et des hypothèses (H_0) et (H_1), on détermine une statistique (une fonction des échantillons observés) dont on connaît la loi sous (H_0) et dont la loi sous (H_0) est différente de la loi sous (H_1).
4. **Etablissement d'une règle de décision.** A partir de la loi de la statistique sous (H_0) et sous (H_1), on établit une règle qui permet de décider quand on rejette et quand on accepte (H_0). On en déduit la forme de la région de rejet de (H_0).

5. **Calcul exact de la région de rejet** On calcule enfin le ou les seuil(s) de la région de rejet grâce à la loi de la statistique sous (H_0).
6. **Décision.** On calcule la valeur observée de la statistique et on conclut.

7.2 Cas d'un seul échantillon

1. **Test sur la probabilité d'un évènement**

   Cadre : On fait n expériences aléatoires indépendantes et on regarde si un événement A est réalisé, autrement dit, on observe un n-échantillon $X_1, \dots, X_n$ d'une loi de Bernoulli de paramètre $p(A)$.

   Hypothèses : On veut tester (H_0): $p(A) = p_0$ contre (H_1): $p(A) \neq p_0$ ou $p(A) < p_0$ ou $p(A) > p_0$.

   Statistique : On considère la statistique S égale au nombre de réalisations de l'événement A dans les n expériences, c'est-à-dire $S = \sum_{i=1}^{n} X_i$.

   Loi sous (H_0) et sous (H_1) : Sous l'hypothèse (H_0), on a $S \sim \mathcal{B}(n, p_0)$ et, sous (H_1), si $p(A) = p$, $S \sim \mathcal{B}(n, p)$. Si n est grand, on approxime la loi binomiale $\mathcal{B}(n, p_0)$ par une loi de poisson $\mathcal{P}(np_0)$ ou une loi gaussienne $\mathcal{N}(np_0, np_0(1 - p_0))$.

2. **Test d'ajustement du χ^2**

   Cadre : Soit X une variable aléatoire discrète à valeurs dans un ensemble $\mathcal{E}$. On observe des valeurs $x_1, \dots, x_n$ d'un n-échantillon de la loi de X, $X_1, \dots, X_n$.

   Hypothèses : On veut tester (H_0): $P(X = i) = p_i$ pour tout i dans $\mathcal{E}$ (les p_i étant donnés) contre (H_1): il existe i dans $\mathcal{E}$ tel que $P(X = i) \neq p_i$.

   Statistique : On divise l'ensemble $\mathcal{E}$ en k classes $\mathcal{E}_1, \dots, \mathcal{E}_k$ de façon à avoir
   $n \sum_{i \in \mathcal{E}_j} p_i \geq 5$ pour tout $j \leq k$, et on utilise alors la statistique du Chi-deux définie par :

   $$Z = \sum_{j=1}^{k} \frac{(N_j - n_j)^2}{n_j},$$

   où :
   - N_j est le nombre d'éléments de $\{X_1, \dots, X_n\}$ qui sont dans la classe $\mathcal{E}_j$,
   - $n_j = n \sum_{i \in \mathcal{E}_j} p_i$ (les n_j sont les répartitions sous (H_0)).

   Loi sous (H_0) et sous (H_1) : Sous l'hypothèse (H_0), $Z \approx \chi^2_{k-1}$. Sous (H_1), $Z \to +\infty$ si $n \to +\infty$.

   Règle de décision : On rejette (H_0) lorsque Z prend de grandes valeurs.

3. **Test d'indépendance du χ^2**

   Cadre : On considère un couple de variables aléatoires (X, Y), où X est à valeurs dans $\{a_1, \dots, a_k\}$ et Y dans $\{b_1, \dots, b_l\}$. On observe un n-échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ de ce couple.

Hypothèses : On veut tester (H_0): X et Y sont indépendantes contre (H_1): X et Y ne sont pas indépendantes.

Statistique :

$$Z = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\frac{N_i N^j}{n} - N_{i,j})^2}{\frac{N_i N^j}{n}},$$

où :

- $N_{i,j}$ est le nombre d'éléments de $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ qui prennent la valeur (a_i, b_j),
- N_i est le nombre d'éléments de $\{X_1, \ldots, X_n\}$ qui prennent la valeur a_i,
- N^j est le nombre d'éléments de $\{Y_1, \ldots, Y_n\}$ qui prennent la valeur b_j,

Loi sous (H_0) *et sous* (H_1)*:* Sous l'hypothèse (H_0), $Z \approx \chi^2_{(k-1)(l-1)}$. Sous (H_1), $Z \to +\infty$ si $n \to +\infty$.

Règle de décision : On rejette (H_0) lorsque Z prend de grandes valeurs.

4. **Modèle gaussien**

Cadre : On dispose d'un n-échantillon $X_1, \ldots, X_n$ de loi $\mathcal{N}(m, \sigma^2)$.

(a) **Moyenne inconnue, variance connue**

Hypothèses : On veut tester (H_0): $m = m_0$ contre (H_1): $m \neq m_0$ ou $m < m_0$ ou $m > m_0$.

Statistique : $X^* = \frac{\overline{X} - m_0}{\sqrt{\frac{\sigma^2}{n}}}$.

Loi sous (H_0) *et sous* (H_1) *:* Sous l'hypothèse (H_0), on a $X^* \sim \mathcal{N}(0,1)$. Sous (H_1), si m est l'espérance des X_i, $(\overline{X} - m)/\sqrt{\frac{\sigma^2}{n}} \sim \mathcal{N}(0,1)$.

(b) **Moyenne connue, variance inconnue**

Hypothèses : On veut tester (H_0): $\sigma = \sigma_0$ contre (H_1): $\sigma \neq \sigma_0$ ou $\sigma < \sigma_0$ ou $\sigma > \sigma_0$.

Statistique : $\Sigma^2 = \sum_{i=1}^{n} (X_i - m)^2$.

Loi sous (H_0) *et sous* (H_1) *:* Sous l'hypothèse (H_0), on a $\frac{\Sigma^2}{\sigma_0^2} \approx \chi^2_n$. Sous (H_1), si σ est la variance, $\Sigma^2/\sigma^2 \approx \chi^2_n$.

(c) **Moyenne inconnue, variance inconnue**

c1) Test sur la moyenne:

Hypothèses : On veut tester (H_0): $m = m_0$ contre (H_1): $m \neq m_0$ ou $m < m_0$ ou $m > m_0$.

Statistique : $T = \sqrt{n} \frac{\overline{X} - m_0}{S}$, avec $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$.

Loi sous (H_0) *et sous* (H_1) *:* Sous l'hypothèse (H_0), on a $T \approx \mathcal{T}_{n-1}$. Sous (H_1), si m est l'espérance, $\sqrt{n}(\overline{X} - m)/S \approx \mathcal{T}_{n-1}$

c2) Test sur la variance:

Hypothèses : On veut tester (H_0): $\sigma = \sigma_0$ contre (H_1): $\sigma \neq \sigma_0$ ou $\sigma < \sigma_0$ ou $\sigma > \sigma_0$.

Statistique : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Loi sous (H_0) *et sous* (H_1) *:* Sous l'hypothèse (H_0), on a $(n-1)\frac{S^2}{\sigma_0^2} \approx \chi^2_{n-1}$. Sous (H_1), si σ est la variance, $(n-1)\frac{S^2}{\sigma^2} \approx \chi^2_{n-1}$.

7.3 Cas de deux échantillons

1. **Test d'homogénéité du χ^2**

Cadre : On considère un couple de variables aléatoires (X, Y), où X et Y sont à valeurs dans $\{a_1, \ldots, a_k\}$. On suppose que X et Y sont indépendantes. On observe un n-échantillon $X_1, \ldots, X_n$ de X et un m-échantillon $Y_1, \ldots, Y_m$ de Y.

Hypothèses : On veut tester (H_0): X et Y suivent la même loi contre (H_1): X et Y ne suivent pas la même loi.

Statistique :

$$Z = \sum_{i=1}^{k} n\frac{(\frac{N_i+M_i}{n+m} - \frac{N_i}{n})^2}{\frac{N_i+M_i}{n+m}} + m\frac{(\frac{N_i+M_i}{n+m} - \frac{M_i}{m})^2}{\frac{N_i+M_i}{n+m}},$$

où :
- N_i est le nombre d'éléments de $\{X_1, \ldots, X_n\}$ qui prennent la valeur a_i,
- M_i est le nombre d'éléments de $\{Y_1, \ldots, Y_m\}$ qui prennent la valeur a_i,

Loi sous (H_0) *et sous* (H_1)*:* Sous l'hypothèse (H_0), $Z \approx \chi^2_{k-1}$. Sous (H_1), $Z \to +\infty$ si $n \to +\infty$.

Règle de décision : On rejette (H_0) lorsque Z prend de grandes valeurs.

2. **Comparaison de moyennes en modèle gaussien**

Cadre : On dispose d'un n_1-échantillon $X_1, \ldots, X_{n_1}$ de loi $\mathcal{N}(m_1, \sigma_1^2)$ et d'un n_2-échantillon $Y_1, \ldots, Y_{n_2}$ de loi $\mathcal{N}(m_2, \sigma_2^2)$. On suppose les X_i et les Y_j indépendants.

Hypothèses : On veut tester (H_0): $m_1 = m_2$ contre (H_1): $m_1 \neq m_2$ ou $m_1 < m_2$ ou $m_1 > m_2$.

a) Variances connues :

Statistique :

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Loi sous (H_0) *:* Sous l'hypothèse (H_0), on a $Z \sim \mathcal{N}(0, 1)$.

b) Variances inconnues mais égales : On suppose en plus que $\sigma_1 = \sigma_2$.

Statistique :

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

où $S^2 = \frac{(n_1-1)S_X^2 + (n_2-1)S_Y^2}{n_1+n_2-2}$.

Loi sous (H_0) *:* Sous l'hypothèse (H_0), on a $T \approx \mathcal{T}_{n_1+n_2-2}$.