Statistiques

IUT Biotechnologie 2ème année

Jean-Christophe Breton

Université de La Rochelle

Octobre-Novembre 2008

version du 04 octobre 2008

Table des matières

1 Lois de probabilité usuelles **1**
- 1.1 Dénombrement . . . 1
- 1.2 Lois discrètes classiques . . . 3
  - 1.2.1 Lois de v.a. finies déjà connues . . . 3
  - 1.2.2 Lois Géométriques . . . 4
  - 1.2.3 Loi de Poisson . . . 4
- 1.3 Lois à densité classiques . . . 5
  - 1.3.1 Loi uniforme . . . 5
  - 1.3.2 Lois exponentielles . . . 6

2 Loi normale (ou gaussienne) **7**
- 2.1 Définition . . . 7
- 2.2 Règle de calcul de probabilités . . . 8
- 2.3 Table de la loi $\mathcal{N}(0,1)$ . . . 9
- 2.4 Approximation par la loi normale . . . 9
- 2.5 Lois dérivées de la loi normale . . . 10
  - 2.5.1 Loi du khi-deux . . . 10
  - 2.5.2 Loi de Student . . . 10

3 Estimation statistique **11**
- 3.1 Introduction . . . 11
- 3.2 Loi d'échantillonage . . . 12
  - 3.2.1 Pour des moyennes . . . 12
  - 3.2.2 Pour des fréquences . . . 12
- 3.3 Estimation ponctuelle . . . 12
  - 3.3.1 Définition . . . 12
  - 3.3.2 Estimation de la moyenne et de la variance . . . 13
- 3.4 Intervalles de confiance . . . 13
  - 3.4.1 Principe . . . 13
  - 3.4.2 Calcul d'un IC . . . 14
  - 3.4.3 Un exemple d'application . . . 17

4 Tests d'hypothèses **18**
4.1 Introduction 18
4.2 Test sur la moyenne 21
4.3 Test sur la variance dans le cas gaussien 22
4.4 Test sur une proportion 22
4.5 Tests de comparaison 23
4.5.1 Comparaison de deux moyennes 23
4.5.2 Comparaison de deux proportions 24
4.6 Les Tests du χ^2 25
4.6.1 Principe 25
4.6.2 Exemples 25

Chapitre 1

Lois de probabilité usuelles

1.1 Dénombrement

Considérons un ensemble $\Omega = \{\omega_1, \ldots, \omega_n\}$ de cardinal n.

Permutation

Le nombre de permutations d'un ensemble est le nombre de manières d'ordonner ses éléments. Le nombre de permutations de Ω est $n! = 1 \times 2 \times 3 \times \cdots \times n$.

En effet, il s'agit de trouver tous les reordonnements de $\{\omega_1, \ldots, \omega_n\}$. On a d'abord n choix pour le premier terme, puis $n-1$ pour le deuxième puis $n-2$ puis ... puis 2 choix pour l'avant dernier et enfin plus qu'un seul pour le dernier. Il y a donc $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = n!$.

Exercice. Faire la preuve pour $n = 3$ et trouver les $3! = 6$ permutations de $\{A, B, C\}$.

Exemple. Un professeur doit faire passer dans la journée 5 étudiants pour un oral de rattrapage. Il a $5! = 120$ manières de choisir l'ordre de passage.

Tirage avec remise

Tirage de p objets (avec remise) dans un ensemble de cardinal n.

Pour chaque tirage, il y a n objets possibles à tirer, il y a donc en tout $n \times \cdots \times n = n^p$ tirages possibles (avec remise) dans un ensemble de cardinal n.

Exemple. Un professeur note chaque étudiant d'une classe de 30 étudiants par une note entière de 0 à 20. Le nombre de résultats possibles est le nombre de manières de choisir de façon indépendante 30 éléments de l'ensemble $\{0, 1, \ldots, 20\}$ de cardinal 21. Il y a donc 21^{30} résultats possibles pour l'ensemble de la classe.

Arrangement (tirages ordonnés sans remise)

On appelle tirage sans remise de p éléments dans un ensemble Ω de cardinal n, tout tirage successif de p éléments de Ω, chaque élément ne pouvant être tiré plus d'une fois.

Bien évidemment, pour qu'un tel tirage puisse exister, il faut avoir $p \leq n$.

Le nombre de tirages sans remise est

$$n(n-1)\dots(n-p+1) = \frac{n!}{(n-p)!}$$

Remarque 1.1.1 Le nombre $n!/(n-p)!$ s'appelle le nombre d'arrangements, on le note A_n^p. Lorsque $n = p$, on retrouve le nombre de permutations, puisqu'on tire tous les éléments de Ω et qu'en fait, on les a reordonnés.

Exemple. 3500 personnes se présentent au concours de l'agrégation de Mathématiques. 300 places sont mises au concours. Combien y-a-t-il de palmarès possibles (en supposant qu'il n'y ait pas d'*ex-aequo*) ?

Réponse : $3500 \times 3499 \times \cdots \times 3202 \times 3201 = \dfrac{3500!}{3200!}$.

Combinaison (tirages désordonnés sans remise)

C'est aussi le nombre de parties d'un ensemble Ω possédant p éléments.

C'est exactement le nombre de manières de choisir p objets dans un ensemble de n objets, l'ordre n'ayant pas d'importance.

On sait qu'il y a $n!/(n-p)!$ tirages de p objets lorsque l'on tient compte de l'ordre. Or un tirage (désordonné) donné (où l'ordre n'est pas pris en compte) représente $p!$ tirages où l'ordre est pris en compte (car il y a $p!$ permutations de l'ensemble des p objets du tirage). Il y a donc $p!$ fois plus de tirages de p objets lorsque l'on tient compte de l'ordre. Finalement, le nombre de tirages (sans tenir compte de l'ordre) est

$$\frac{n!}{p!(n-p)!}.$$

Exemple. Dénombrer le nombre de tirages sans remise de 2 éléments parmi 4 avec ordre puis sans ordre.

Exemple. 3500 personnes se présentent au concours de l'agrégation de Mathématiques. 300 places sont mises au concours. Combien y-a-t-il de promotions possibles ?

Réponse : C_{3500}^{300}. Ici, Ω est l'ensemble des candidats et il s'agit de choisir 300 d'entre eux. On s'intéresse aux différentes promotions possibles, prises dans leur ensemble, sans tenir compte du classement de la promotion.

• Rappelons d'abord la définition des coefficients binomiaux et la formule du binome de Newton :

$$C_n^k = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} \quad 0 \leq k \leq n, \qquad (a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}.$$

C_n^k s'interprète comme le nombre d'échantillons de taille k dans une population de taille n. Par exemple, si dans une urne de n boules distinctes, on en tire k, il y a C_n^k tirages différents possibles.

Rappelons les propriétés immédiates suivantes pour tout $n \in \mathbb{N}^*$ et $k \leq n$:

$$C_n^k = C_n^{n-k}, \quad C_n^n = C_n^0 = 1, \quad C_n^{n-1} = C_n^1 = n$$
$$C_n^{k-1} + C_n^k = C_{n+1}^k \quad \text{(triangle de Pascal)}.$$

1.2 Lois discrètes classiques

L'espérance $E[X]$ d'une v.a. X donne sa valeur moyenne. Sa variance $\mathrm{Var}(X) = E[X^2] - E[X]^2$ donne une indication sur sa dispersion autour de sa valeur moyenne. Son écart-type est $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

1.2.1 Lois de v.a. finies déjà connues

Loi de Bernoulli de paramètre p notée $b(p)$. Une v.a. X suit une loi de Bernoulli de paramètre $p \in [0,1]$ si elle ne prend que les deux valeurs 0 et 1 avec :

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p := q.$$

Son espérance est $E[X] = 0 \times (1-p) + 1 \times p = p$. Sa variance est $\mathrm{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1-p)$.

Une v.a. $X \simeq b(p)$ modélise si le succès ou l'échec d'une expérience qui a une probabilité p de succès. $X = 1$ en cas de succès. $X = 0$ en cas d'échec.

Exemple. Pile ou face avec $p = 1/2$ si la pièce est équilibrée, $p \neq 1/2$ si elle est truquée.

Loi equirépartie sur un ensemble fini $\{x_1, \ldots, x_n\}$ notée $\mathcal{E}\{x_1, \ldots, x_n\}$. Une v.a. X prenant un nombre fini de valeurs $x_1, \ldots, x_n$ suit une loi equirépartie quand

$$\mathbb{P}_X(\{x_i\}) = \frac{1}{n}, \quad 1 \leq i \leq n.$$

Son espérance est $E[X] = \dfrac{x_1 + \cdots + x_n}{n}$.

Exemple. Jet d'un dé (équilibré).

Loi binomiale de paramètres n, p notée $\mathcal{B}(n,p)$. Une v.a. suit une loi binomiale de paramètres $n \in \mathbb{N}^*$ et $p \in [0,1]$ si elle prend ses valeurs possibles parmi $\{0, 1, 2, \ldots, n\}$ et pour tout $k = 0, 1, \ldots, n$, on a

$$\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k} \tag{1.1}$$

où $C_n^k = \dfrac{n!}{k!(n-k)!}$ est le coefficient binomial.

Son espérance est $E[X] = np$. Sa variance est $\mathrm{Var}(X) = np(1-p)$.

Une v.a. $X \simeq B(n,p)$ modélise le nombre de succès dans une suite de n expériences indépendantes où il y a une probabilité p de succès à chacune.

Ainsi, $\mathbb{P}(X = k)$ est la probabilité d'avoir exactement k succès en n épreuves. On en déduit l'explication suivante des différents facteurs de (1.1) :

- p^k est la probabilité des k succès (par indépendance des tirages),
- $(1-p)^{n-k}$ est la probabilité des $n-k$ échecs (pour avoir **exactement** k succès, il faut bien que les $n-k$ autres épreuves soient des échecs),
- et C_n^k pour tenir compte de tous les choix possibles des k épreuves réussies sur les n réalisées.

Intéressons nous maintenant aux lois des v.a. discrètes prenant un nombre infini de valeurs.

1.2.2 Lois Géométriques

Définition 1.2.1 *Une v.a. X suit la loi géométrique de paramètre $p \in]0,1[$ notée $\mathcal{G}(p)$ si elle prend des valeurs entières positives non nulles et*

$$\mathbb{P}(X=k) = (1-p)^{k-1}p, \quad k \in \mathbb{N}^*.$$

Son espérance est $E[X] = 1/p$. Sa variance est $\mathrm{Var}(X) = 1/p^2$.

Une v.a. $X \simeq \mathcal{G}(p)$ modélise le rang du premier succès dans une suite infinie d'épreuve indépendante où à chacune il y a une probabilité p de succès.

1.2.3 Loi de Poisson

Cette loi intervient dans les processus aléatoires dont les éventualités sont faiblement probables et survenant indépendamment les unes des autes : cas de phénomènes accidentels, d'anomalies diverses, de problèmes d'encombrement (files d'attente), de rupture de stocks, etc.

Définition 1.2.2 *On dit qu'une v.a. discrète X suit une loi de Poisson de paramètre $\lambda > 0$ si elle prend des valeurs entières postives ou nulles et*

$$\forall k \in \mathbb{N}, \quad \mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

La loi de Poisson de paramètre $\lambda > 0$ est notée $\mathcal{P}(\lambda)$.

Son espérance est $E[X] = \lambda$. Sa variance est $\mathrm{Var}(X) = \lambda$.

Approximation de la loi binomiale par la loi de Poisson

En liaison avec les lois binomiales, on dispose de la règle pratique suivante :

Règle. *Lorsque n est « grand » et np est « petit », on peut remplacer la loi binomiale $\mathcal{B}(n,p)$ par la loi de Poisson $\mathcal{P}(\lambda)$ où $\lambda = np$.*

En général, on considère que lorsque n est de l'ordre de quelques centaines et np est de l'ordre de quelques unités, l'approximation de $\mathcal{B}(n,p)$ par $\mathcal{P}(np)$ est assez bonne.

Intérêt : si n est grand, le calcul des coefficients binomiaux C_n^k est fastidieux, voire impossible. En approchant par la loi de Poisson, le calcul devient assez simple.

Exemple : Le président d'un bureau de vote est né un 1er avril. Il décide de noter le nombre de personnes ayant leur anniversaire le même jour que lui parmi les 500 premiers votants.

La situation peut être assimilée à une suite de 500 épreuves indépendantes répétées avec une probabilité $p = 1/365$ de succès (on néglige les effets des années bissextiles, sinon il faudrait plutôt prendre $p = 4/(3 \times 365 + 366)$). Notons X la variable aléatoire qui compte le nombre de succès. X suit une loi $\mathcal{B}(500, p)$, ainsi :

$$\mathbb{P}(X = k) = C_{500}^k p^k (1-p)^{500-k}.$$

Comme 500 est « grand » et $np = 500/365 \simeq 1,37$, la règle ci-dessus permet l'approximation par la loi $\mathcal{P}(\lambda)$ avec $\lambda = 500/365$. Voici une comparaison numérique pour les petites valeurs de k :

k	0	1	2	3	4	5
$\mathbb{P}(X = k)$	$0,2537$	$0,3484$	$0,2388$	$0,1089$	$0,0372$	$0,0101$
$\dfrac{e^{-\lambda}\lambda^k}{k!}$	$0,2541$	$0,3481$	$0,2385$	$0,1089$	$0,0373$	$0,0102$

On constate effectivement que les valeurs approchées sont très proches des valeurs réelles.

1.3 Lois à densité classiques

Une loi est à densité (de densité f) si les probabilités s'expriment comme des intégrales :

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(t)dt.$$

1.3.1 Loi uniforme

Cette loi modélise un phénomène uniforme sur un intervalle donné.

Définition 1.3.1 *La v.a. X suit une loi uniforme sur l'intervalle borné $[a, b]$ si elle a une densité f constante sur cet intervalle et nulle en dehors. Elle est notée $\mathcal{U}([a, b])$. Sa densité est alors*

$$f(t) = \begin{cases} 1/(b-a) & si \quad t \in [a, b], \\ 0 & si \quad t \notin [a, b]. \end{cases}$$

Cette loi est l'équivalent continue de la loi discrète equirépartie.

Son espérance est $E[X] = \frac{b-a}{2}$. Sa variance est $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Le résultat suivant permet d'éviter des calculs fastidieux pour la probabilité uniforme d'un intervalle.

Proposition 1.3.1 *Si X est une v.a. de loi uniforme sur $[a, b]$ alors pour tout intervalle I de $\mathbb{R}$:*

$$\mathbb{P}(X \in I) = \frac{l([a, b] \cap I)}{l([a, b])}$$

où $l(J)$ désigne la longueur de l'intervalle J ($l([a, b]) = b - a$).

1.3.2 Lois exponentielles

Définition 1.3.2 *Soit α un réel strictement positif. La v.a. X suit une loi exponentielle de paramètre α, notée $\mathcal{E}(\alpha)$, si elle admet pour densité :*

$$f(t) = \alpha e^{-\alpha t} \mathbf{1}_{[0,+\infty[}(t).$$

Son espérance est $E[X] = 1/\alpha$. Sa variance est $\mathrm{Var}(X) = 1/\alpha^2$.

En pratique, à la place de la fonction de répartition, on utilise souvent la fonction de survie G d'une v.a. de loi exponentielle

$$G_X(x) = \mathbb{P}(X > x) = 1 - F_X(x) = \begin{cases} 1 & \text{si } x \leq 0, \\ e^{-\alpha x} & \text{si } x \geq 0. \end{cases}$$

Les lois exponentielles sont souvent utilisées pour modéliser des temps d'attente ou des durées de vie. Par exemple, les temps d'attente à partir de maintenant du prochain tremblement de terre, de la prochaine panne d'un appareil, de la prochaine désintégration dans un réacteur nucléaire suivent des lois exponentielles. Le paramètre α désigne alors l'inverse du temps d'attente moyen.

Chapitre 2

Loi normale (ou gaussienne)

C'est une loi très importante pour plusieurs raisons :
- Elle apparaît dans de nombreux problèmes courants (pour les modéliser),
- Bien souvent, on peut approcher une loi par une loi normale.
- De plus, on dispose de la table de ses valeurs à laquelle on se réferre pour des calculs approchés.

Synonymes pour cette loi : loi gaussienne, loi de Gauss.

2.1 Définition

Définition 2.1.1 *La loi normale standard* $\mathcal{N}(0,1)$ *est celle de densité* $f_{0,1}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Son espérance est $E[X] = 0$. Sa variance est $\mathrm{Var}(X) = 1$.

Définition 2.1.2 *On dit que la v.a.* X *suit une loi normale* $\mathcal{N}(m, \sigma^2)$ *si elle a pour densité la fonction*

$$f_{m,\sigma}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{(t-m)^2}{2\sigma^2}\big).$$

Son espérance est $E[X] = m$. Sa variance est $\mathrm{Var}(X) = \sigma^2$.

Remarque 2.1.1 Cette loi est fondamentale en théorie des probabilités et en statistique : c'est la loi limite de la moyenne dans une suite infinie d'épreuves répétées indépendantes. En pratique elle sert à modéliser les effets additifs de petits phénomènes aléatoires indépendants répétés souvent.

Règles pour les lois normales.

• *Si* $X \simeq \mathcal{N}(m, \sigma^2)$ *et* $a \in \mathbb{R}$ *alors* $aX \simeq \mathcal{N}(am, a\sigma^2)$.

• *Quand on somme des v.a. gaussiennes indépendantes de loi* $\mathcal{N}(m_1, \sigma_1^2)$ *et* $\mathcal{N}(m_2, \sigma_2^2)$, *on obtient une v.a. gaussienne avec pour paramètres la somme des paramètres* $\mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.

$$X_1 \simeq \mathcal{N}(m_1, \sigma_1^2) \perp\!\!\!\perp X_2 \simeq \mathcal{N}(m_2, \sigma_2^2) \implies X_1 + X_2 \simeq \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2).$$

Plus généralement quand $X_1, \ldots, X_n$ sont n v.a. indépendante de lois $\mathcal{N}(m, \sigma^2)$, alors

$$\frac{X_1 + \cdots + X_n}{n} \simeq \mathcal{N}(m, \frac{\sigma^2}{n}).$$

Notez encore qu'on peut facilement passer d'une loi normale à la loi standard.

Proposition 2.1.1 *Si la v.a. X suit une loi $\mathcal{N}(m, \sigma^2)$, alors $Y := \dfrac{X - m}{\sigma}$ suit la loi $\mathcal{N}(0, 1)$.*

La v.a. Y s'appelle la v.a. centrée réduite associée à X. En fait, pour faire des calculs effectifs de probabilité, grâce à ce résultat, on commencera systématiquement par se ramener d'une loi normale quelconque $\mathcal{N}(m, \sigma^2)$ à la loi normale standard $\mathcal{N}(0, 1)$. On pourra alors utiliser la table des valeurs pour cette loi.

Démonstration : Calculons pour $a < b$ quelconques $\mathbb{P}(a \leq Y \leq b)$:

$$\begin{aligned} \mathbb{P}\Big(a \leq \frac{X - m}{\sigma} \leq b\Big) &= \mathbb{P}(\sigma a + m \leq X \leq \sigma b + m) \\ &= \int_{\sigma a + m}^{\sigma b + m} \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{(t-m)^2}{2\sigma^2}\big)\, dt. \end{aligned}$$

Il suffit alors de faire le changement de variable $s = (t - m)/\sigma$ pour obtenir

$$\forall a \in \mathbb{R},\ \forall b > a, \quad \mathbb{P}(a \leq Y \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{s^2}{2}\big)\, ds,$$

c'est à dire Y suit la loi $\mathcal{N}(0, 1)$. □

2.2 Règle de calcul de probabilités

Dans l'utilisation de la table de la loi normale standard $\mathcal{N}(0, 1)$, on aura des calculs de probabilités à faire. On les fera avec les règles suivantes :

$$\begin{aligned} \mathbb{P}(X = a) &= 0 \\ \mathbb{P}(X < a) &= \mathbb{P}(X \leq a) \\ \mathbb{P}(X > a) &= 1 - \mathbb{P}(X \leq a) \\ \mathbb{P}(X \leq -a) &= \mathbb{P}(X \geq a) = 1 - \mathbb{P}(X < a) \\ \mathbb{P}(-a \leq X \leq a) &= 2\mathbb{P}(X \leq a) - 1. \end{aligned}$$

Les trois premières règles sont vraies pour toute v.a. X à densité (car pour ces lois les points sont négligeables). Les deux dernières sont vraie pour toute loi symétrique (càd avec densité paire : $f(-t) = f(t)$, comme la loi normale ou (cf. après) la loi de Student mais pas la loi du χ^2).

2.3 Table de la loi $\mathcal{N}(0,1)$

La table de la loi $\mathcal{N}(0,1)$ permet deux choses pour une v.a. $X_0 \simeq \mathcal{N}(0,1)$:

1. Connaissant la valeur de $t \geq 0$, trouver la valeur de $\mathbb{P}(X_0 \leq t)$,
2. Connaissant la valeur de d'une probabilité $\mathbb{P}(X_0 \leq t)$, trouver la valeur de $t \geq 0$ correspondant.

Objectif. En général, on souhaite calculer des probabilités du type

$$\mathbb{P}(X > t),\ \mathbb{P}(X < t),\ \mathbb{P}(|X| > t),\ \mathbb{P}(s < X < t)$$

lorsque X suit une loi normale $\mathcal{N}(m, \sigma^2)$ pas nécessairement centrée réduite.

Étape 1 Reexprimer les probabilités à calculer avec la v.a. centrée réduite $X_0 = \frac{X-m}{\sigma}$.

Étape 2 Via les règles ci-dessus, se ramener à des probabilités du type $\mathbb{P}(X_0 \leq t_0)$ pour certains $t_0 \geq 0$.

Étape 3 Utiliser la table de la loi normale standard.

Exercice 1. Si $X \simeq \mathcal{N}(3, 0.25)$, calculer $\mathbb{P}(X > 3.5)$.

Méthode. D'abord on centre et on réduit, pour obtenir une v.a. $X_0 \simeq \mathcal{N}(0,1)$,

$$X_0 = \frac{X-3}{\sqrt{0.25}}.$$

On remarque ensuite l'égalité d'événements suivante :

$$\{X > 3.5\} = \{X_0 > 1\}\ ;$$

enfin, on cherche $\mathbb{P}(X_0 < 1)$ à partir de la table de la loi $\mathcal{N}(0,1)$. On trouve

$$\mathbb{P}(X > 3.5) = 0.16.$$

Exercice 2. Si $X \sim \mathcal{N}(3; 0,25)$ et si $\mathbb{P}(X > t) = 0,6$, calculer t (et trouver $t = 2,875$).

2.4 Approximation par la loi normale

Un résultat général de probabilité (le théorème central limite, TCL) justifie l'approximation de certaines lois par des lois normales. On utilisera par la suite les deux approximations de loi suivantes :

Loi de X	Loi approchée de X	conditions requises
$\mathcal{B}(n,p)$	$\mathcal{N}(np, np(1-p))$	$n \geq 30$, $np \geq 10$, $n(1-p) \geq 10$
$\mathcal{P}(\lambda)$	$\mathcal{N}(\lambda, \lambda)$	$\lambda \geq 10$

Correction de continuité. Lorsque l'on approche une loi discrète par une loi à densité, il convient de faire une correction de continuité que l'on peut résumer avec la formule suivante : pour toute les valeurs x_i de X,

$$\mathbb{P}_{\text{discrète}}(X = x_i) \simeq \mathbb{P}_{\text{à densité}}(x_i - 0.5 \leq X \leq x_i + 0.5).$$

Cette formule s'interprète bien graphiquement.

2.5 Lois dérivées de la loi normale

Parfois d'autres lois que la loi normale sont utiles dans les approximations (cf. les calculs d'intervalle de confiance, de test). Ce sont les lois de Student et du χ^2 (lire khi-deux). Ces lois dépendent d'un paramètre n entier, appelé degré de liberté (d.d.l.).

De même que pour la loi normale $\mathcal{N}(0,1)$, on disposera de tables pour ces lois. Les mêmes règles de calcul que pour la loi normale s'appliqueront pour reexprimer les probabilités qu'on cherchera en des probabilités disponibles dans ces tables.

2.5.1 Loi du khi-deux

Soient $X_1, \ldots, X_n$ des v.a. indépendantes de même loi $\mathcal{N}(0,1)$. Posons $\chi^2 = \sum_{i=1}^{n} X_i^2$. Par définition, la v.a. χ^2 suit une loi du khi-deux à n degrés de liberté (abréviation d.d.l.). On note cette loi $\chi^2(n)$.

Propriétés.
- $\chi^2 \geq 0$, cette loi n'est donc pas symétrique,
- χ^2 admet une densité (difficile à retenir),
- $E[\chi^2] = n$ et $\text{Var}(\chi^2) = 2n$,
- Pour $n \geq 30$, $\sqrt{2\chi^2} - \sqrt{2n-1}$ suit approximativement une loi $\mathcal{N}(0,1)$.

2.5.2 Loi de Student

Elle se définit à partir d'une loi $\mathcal{N}(0,1)$ et d'une loi $\chi^2(n)$. Soient X et χ^2 deux v.a. indépendantes telles que $X \simeq \mathcal{N}(0,1)$ et $Y \simeq \chi^2(n)$. Posons $T = \frac{X}{\sqrt{\chi^2}}\sqrt{n}$. Par définition, la v.a. T suit une loi de student à n degrés de liberté. On note cette loi $\mathcal{T}(n)$.

Propriétés.
- $\mathcal{T}(n)$ admet une densité paire, cette loi est donc symétrique,
- $E[T] = 0$ et $\text{Var}(T) = n/(n-2)$ si $n > 2$,
- Pour $n > 30$, $\mathcal{T}(n)$ peut être approchée par $\mathcal{N}(0,1)$.

Chapitre 3

Estimation statistique

3.1 Introduction

L'objectif de l'estimation statistique est le suivant : évaluer certaines grandeurs associées à une population à partir d'observations faites sur un échantillon. Bien souvent, ces grandeurs sont des moyennes ou des variances. On prendra soin de distinguer ces grandeurs théoriques (inconnues et à estimer) de celles observées sur un échantillon.

Exemples de problèmes :
- Quelle est la fréquence (probabilité) de survenue d'un certain cancer chez les souris ?
- Quelle est la glycémie moyenne d'un patient ?
- Quelle est l'écart moyen de la glycémie d'un patient autour de sa glycémie moyenne ?

On apporte deux types de réponses à ces questions : à partir d'un échantillon,

1. On « calcule »une valeur qui semble être la meilleure possible : on parle d'estimation ponctuelle,
2. On « calcule »un intervalle de valeurs possibles : c'est la notion d'intervalle de confiance.

On se placera toujours dans la situation suivante :
- Un échantillon ω est obtenu par tirages avec remise de n individus dans la population de référence,
- Les valeurs observées $x_1, \ldots, x_n$ d'une grandeur (ex : poids) sur un échantillon ω ne dépendront donc pas les unes des autres (ce ne serait pas le cas avec des tirages sans remise).

Un échantillon est la donnée de n va. $X_1, \ldots, X_n$ de même loi.
Une observation correspond à une réalisation $\omega \in \Omega$ du hasard. On a alors

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Si on change d'observation, cela correspond à changer la réalisation du hasard en $\omega' \in \Omega$ et on a d'autres valeurs observées sur l'échantillon :

$$x'_1 = X_1(\omega'), \ldots, x'_n = X_n(\omega').$$

On modélisera donc cette situation par un ensemble fondamental

$$\Omega = \{ \text{ échantillons } \omega \text{ de taille } n \text{ avec remise } \}$$

et des variables aléatoires $X_1, \ldots, X_n$ indépendantes (car tirages avec remise) et de même loi (car on observe la même grandeur). On a ainsi pour un échantillon ω donné, des valeurs observées $X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$.

3.2 Loi d'échantillonage

3.2.1 Pour des moyennes

Soit une population d'effectif total N connu. On considère un échantillon d'effectif n. Un élément quelconque X de l'échantillon suit la loi d'échantillonage de taille n et de moyenne $\bar{X}$. Quand n devient grand ($n \geq 30$), la loi d'échantillonnage peut être approchée par la loi normale $\mathcal{N}(\bar{X}, \sigma^2/n)$ où σ^2 est supposée connue.

Exemple. Dans une population, l'écart-type de la taille est 5 cm. Si sur 200 personnes, la taille moyenne observée est $\bar{X} = 175$ cm, alors la taille X d'un individu quelconque issu de cette population suit la loi d'échantillonage $\mathcal{N}(175; 0,125)$ (car $\sigma^2/n = 5^2/200$).

3.2.2 Pour des fréquences

On étudie une population de taille N (connu) et un caractère X à deux éventualités (échec ou succès) avec probabilité p. On sait (cf. loi de Bernoulli) que $E[X] = p$ et $\mathrm{Var}(X) = p(1-p)$.

Si on prélève un échantillon de taille n, le nombre de succès X_n est compté par une loi binomiale $\mathcal{B}(n,p)$ avec $E[X_n] = np$ et $\mathrm{Var}(X_n) = np(1-p)$.

Quand n est grand ($n \geq 30$), la loi de la fréquence X_n/n des succès s'approxime par

$$\mathcal{N}\left(p, \frac{p(1-p)}{n}\right).$$

Exemple. Considérons une population où 10% des gens développent une certaine allergie. Dans un échantillon de 200 personnes de cette population, le nombre d'allergiques suit la loi binomiale $\mathcal{B}(200; 0,1)$. On l'approxime la loi de la fréquence par la loi normale $\mathcal{N}(0,1; 9.10^{-4})$.

3.3 Estimation ponctuelle

3.3.1 Définition

On cherche à estimer une valeur θ inconnue liée à un certain phénomène aléatoire, en général, la moyenne μ ou la variance σ^2 ou encore l'écart-type σ de la loi du phénomène.

Pour ce faire, on dispose d'observations indépendantes du phénomènes, càd de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi (celle du phénomène). On parle d'un échantillon. On définit à partir de l'échantillon une nouvelle variable aléatoire notée T dont les valeurs seront proches de celle de la grandeur θ à estimer. Cette nouvelle variable aléatoire T sera appelée **estimateur** de θ.

Il peut y avoir plusieurs estimateurs pour une même grandeur θ, certains meilleurs que d'autres.

Exemple. $\theta = \mu =$ moyenne des poids des nouveaux nés en France. Ici, on prendra comme estimateur T la variable aléatoire donnée par la moyenne (arithmétique) observée sur un échantillon de 10 nouveaux nés. On note cet estimateur en général $\bar{X}$:

$$\bar{X} = \frac{X_1 + \cdots + X_{10}}{10}.$$

La valeur de $\bar{X}$ calculée sur cet échantillon noté $\bar{x} = \bar{X}(\omega)$ sera appelée **estimation** de μ.

3.3.2 Estimation de la moyenne et de la variance

Étant donné un échantillon $X_1, \ldots, X_n$ d'un caractère X inconnu, on admet que
- le meilleur estimateur de la moyenne $\mu = E[X]$ du caractère X est

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

- le meilleur estimateur de la variance $\sigma^2 = \mathrm{Var}(X)$ du caractère X est la variance empirique corrigée

$$\begin{aligned} S_c^2 &= \frac{n}{n-1}\left(\frac{1}{n}(\sum_{i=1}^{n} X_i^2) - \bar{X}^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{aligned}$$

Dans le cas particulier où le caractère X suit une loi de Bernoulli $b(p)$, comme la moyenne μ est égale à la proportion p, c'est une estimation de proportion (ou de fréquence) qu'on fait quand on estime sa moyenne $E[X] = p$.

3.4 Intervalles de confiance

3.4.1 Principe

Un estimateur permet de calculer une valeur sur un échantillon qui devrait être proche du paramètre θ sans pour autant savoir si cette valeur est totalement fiable.

C'est pourquoi on a introduit la notion d'intervalle de confiance : c'est un intervalle dans lequel se trouve θ avec une probabilité grande $1-\alpha$ (où α est un risque qu'on se fixe, en général, petit). On peut en théorie choisir $1-\alpha$ aussi proche de 1 que l'on veut, mais alors l'intervalle de confiance grandit et devient imprécis. Il s'agit donc d'un compromis entre précision (intervalle peu étendu) et sûreté (α petit).

La probabilité $1-\alpha$ est appelée **niveau de confiance** et α le **risque (de 1ère espèce)**, càd la probabilité que l'intervalle proposé (qu'on notera IC, pour intervalle de confiance) ne contienne pas la valeur à estimer θ.

Problème : comment trouver un intervalle de confiance ? L'idée est de trouver une variable aléatoire U de loi connue qui serait une fonction des observations aléatoires $X_1, \ldots, X_n$ et de θ, le paramètre à estimer.

Exemple. Supposons que $X_1, \ldots, X_n$ suivent une loi $\mathcal{N}(\mu, 1)$ et que l'on cherche un intervalle de confiance pour μ avec un niveau de confiance de 0.95. On a déjà vu que $\bar{X} \sim \mathcal{N}(\mu, 1/n)$. On connaît donc la loi de

$$U = \frac{\bar{X} - \mu}{1/\sqrt{n}} \sim \mathcal{N}(0,1).$$

On remarque alors que la condition $\mu \in [\bar{X} - t/\sqrt{n}, \bar{X} + t/\sqrt{n}]$ équivaut à $|U| \leq t$ dont la probabilité doit être de 0.95. La table de $\mathcal{N}(0,1)$ permet alors de trouver t tel que

$$\mathbb{P}(\mu \in [\bar{X} - t/\sqrt{n}, \bar{X} + t/\sqrt{n}]) = \mathbb{P}(|U| \leq t) = 0.95.$$

D'après les propriété de la loi normale (symétrie), t doit vérifier :

$$\mathbb{P}(U \leq t) = 1 - 0.05/2 = 0.975.$$

On trouve dans la table de $\mathcal{N}(0,1)$ la valeur $t = 1.96$. L'intervalle de confiance cherché pour un échantillon ω donné de taille n est donc

$$[\bar{X}(\omega) - t/\sqrt{n}, \bar{X}(\omega) + t/\sqrt{n}] = [\bar{x} - 1.96/\sqrt{n}, \bar{x} + 1.96/\sqrt{n}].$$

Si par exemple pour notre échantillon de taille $n = 100$, on a $\bar{x} = 2$, alors on a $IC = [1.894, 2.196]$.

3.4.2 Calcul d'un IC

On suppose que les observations $x_1, \ldots, x_n$ sont issues de n v.a. indépendantes $X_1, \ldots, X_n$ de même loi $\mathcal{N}(\mu, \sigma^2)$.

Si la loi n'est pas gaussienne, on suppose alors que la taille de l'échantillon est grande ($n \geq 30$ en pratique), le théorème central limite (TCL) permet de faire des approximations par des lois normales, ce qui donnera des intervalles de confiance approximatifs mais suffisant en pratique.

On fera donc systématiquement comme si les échantillons sont gaussiens lorsque sa taille est élevé.

On va chercher les expressions des intervalles de confiance au niveau de confiance $1-\alpha$ pour la moyenne μ noté $IC_{1-\alpha}(\mu)$ et pour la variance σ^2 noté $IC_{1-\alpha}(\sigma^2)$.

Calcul de $IC_{1-\alpha}(\mu)$ lorsque σ^2 est connu

Étant donné $\bar{X}$, l'estimateur ponctuel de μ calculé sur l'échantillon, l'intervalle de confiance pour μ cherché se calcule à partir d'un échantillon ω donné de taille n par

$$IC_{1-\alpha}(\mu) = \left[\bar{X}(\omega) - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X}(\omega) + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

où t_α est donné par

$$\mathbb{P}(|U| \leq t_\alpha) = 1 - \alpha \Leftrightarrow \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2$$

dans la table de la loi $\mathcal{N}(0,1)$ de la v.a. U.

On remarquera que si l'on change d'échantillon ω, la moyenne observée $\bar{X}(\omega)$ change et l'intervalle de confiance $IC_{1-\alpha}(\mu)$ change aussi.

Calcul de $IC_{1-\alpha}(\mu)$ lorsque σ^2 est inconnu

Dans cette situation l'expression précédente de l'intervalle de confiance ne peut être calculée car σ^2 n'est plus connu.

Idée : remplacer σ^2 par son estimateur

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

et faire comme avant sauf qu'il faut remplacer la loi normale $\mathcal{N}(0,1)$ par la loi de Student $\mathcal{T}(n-1)$. L'intervalle de confiance pour μ se calcule à partir d'un échantillon ω donné de taille n par

$$IC_{1-\alpha}(\mu) = \left[\bar{X}(\omega) - t_\alpha \frac{S_c}{\sqrt{n}}, \bar{X}(\omega) + t_\alpha \frac{S_c}{\sqrt{n}}\right] \tag{3.1}$$

où t_α est donné par

$$\mathbb{P}(|U| \leq t_\alpha) = 1 - \alpha \Leftrightarrow \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2$$

dans la table de **Student** $\mathcal{T}(n-1)$ de la v.a. U.

Quand n est grand ($n \geq 30$), on peut considérer que la loi de Student est proche de la normale et prendre t_α dans la table de la loi normale.

Cas d'une proportion

Cela correspond à chercher la moyenne p d'une loi de Bernoulli $b(p)$ dont on ne connait pas la variance σ^2 (et pour cause pour une telle loi $\sigma^2 = p(1-p)$).

Bien sûr, on n'est pas dans le cadre d'une loi normale (puisque la loi est $b(p)$), il faut alors supposer l'échantillon assez grand et la loi de référence redevient la loi normale (par le TCL).

Dans le cas d'un intervalle de confiance pour une proportion p inconnue, (3.1) devient

$$IC_{1-\alpha}(p) = \left[f - t_\alpha \sqrt{\frac{f(1-f)}{n}}, f + t_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$$

où $f = \bar{X}(\omega)$ est la fréquence observée du caractère considéré sur l'échantillon étudié (c'est donc l'estimateur sur l'échantillon de l'inconnue p) et t_α est toujours donné par $\mathbb{P}(|U| \leq t_\alpha) = \alpha$ ou $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2$ dans la table de la loi normale $\mathcal{N}(0,1)$ de la v.a. U.

Les conditions requises pour une bonne approximation par la loi normale sont $n \geq 30$, $nf \geq 10$, $n(1-f) \geq 10$.

Calcul de $IC_{1-\alpha}(\sigma^2)$ lorsque μ est connue

L'intervalle de confiance de la variance σ^2 se calcule à partir de l'échantillon de taille n par

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{\sum_{i=1}^n (X_i(\omega) - \mu)^2}{b}, \frac{\sum_{i=1}^n (X_i(\omega) - \mu)^2}{a} \right]$$

où a et b sont à trouver dans la table de la loi $\chi^2(n)$ de la v.a. U par

$$\mathbb{P}(U \leq a) = \alpha/2 \text{ et } \mathbb{P}(U \leq b) = 1 - \alpha/2.$$

Calcul de $IC_{1-\alpha}(\sigma^2)$ lorsque μ est inconnue

À nouveau, comme μ est inconnue, l'idée est de la remplacer par son estimation $\bar{X}$. L'intervalle de confiance de la variance σ^2 se calcule alors à partir de l'échantillon de taille n par

$$IC_{1-\alpha}(\sigma^2) = \left[\frac{nS^2(\omega)}{b}, \frac{nS^2(\omega)}{a} \right],$$

où $S^2(\omega) = \dfrac{\sum_{i=1}^n (X_i(\omega) - \bar{X}(\omega))^2}{n}$ et où les réels a et b sont à déterminer dans la table de la loi $\chi^2(n-1)$ de la v.a. U par

$$\mathbb{P}(U \leq a) = \alpha/2 \text{ et } \mathbb{P}(U \leq b) = 1 - \alpha/2.$$

3.4.3 Un exemple d'application

On suppose que le taux de cholestérol X d'un individu choisi au hasard dans une population donnée suit une loi normale. Sur un échantillon ω de 100 individus, on constate la moyenne des taux observés est $\bar{x} = \bar{X}(\omega) = 1.55$(gr pour mille). On constate aussi une variance corrigée $s_c^2 = S_c^2(\omega) = 0.25$. Donner un intervalle de confiance pour la moyenne μ au niveau de confiance 0.95.

Réponse :

$$IC_{0.95}(\mu) = \left[\bar{X} - t_\alpha \frac{S_c}{\sqrt{n}}, \bar{X} + t_\alpha \frac{S_c}{\sqrt{n}}\right],$$

où t_α est donné par $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2 = 0,975$ dans la table de Student $\mathcal{T}(99)$. à l'aide de la table de la loi de $U \simeq \chi^2(99)$.

La table de Student la plus proche dont on dispose est celle $\mathcal{T}(100)$ pour laquelle $t_{0,05} = 1,984$. On en déduit

$$\begin{aligned} IC_{0.95}(\mu) &= \left[1,55 - 1,984\frac{0,25}{10}\ ;\ 1,55 - 1,984\frac{0,25}{10}\right] \\ &= [1,504, 2,046] \end{aligned}$$

Chapitre 4

Tests d'hypothèses

4.1 Introduction

Il y a deux grands types de tests : les tests paramétriques et les tests non paramétriques (exemple : test du χ^2). Un test non paramétrique teste une propriété (indépendance ou pas, homgénéité ou pas ...). Un test paramétrique consiste à vérifier si une caractéristique d'une population, que l'on notera θ, satisfait une hypothèse que l'on pose a priori, appelée hypothèse nulle H_0. Il s'agit donc de tester un paramètre. Elle est en général de la forme H_0 : « $\theta = \theta_0$ »ou H_0 : « $\theta \leq \theta_0$ »ou encore H_0 : « $\theta \geq \theta_0$ ». Comme pour les intervalles de confiance, on a besoin pour cela d'un échantillon dont les valeurs sont celles prises par n v.a. $X_1, \ldots, X_n$ indépendantes de même loi. Voici la procédure générale d'un test que l'on illustrera avec l'exemple suivant :

Énoncé. Le temps de réaction X d'une souris à un certain test suit une loi normale de moyenne 19 minutes. On désire expérimenter un certain produit que l'on administre à 8 de ces souris. On obtient les temps de réaction (en minutes) suivants :

$$15, 14, 21, 12, 17, 12, 19, 18.$$

Le produit réduit-il le temps de réaction moyen ?

1. Formuler H_0 et l'hypothèse alternative H_1, par exemple H_0 : « $\mu = 19$ » contre H_1 : « $\mu < 19$ » ;
2. Choisir un risque α représentant la probabilité de rejeter H_0 à tort (exemple $\alpha = 0.05$). C'est le risque de 1ère espèce.

   Il y un autre risque β dit de 2ème espèce représentant la probabilité d'accepter H_0 à tort. C'est un risque que l'on contrôle assez mal (mais en général, on préfère contrôler le risque de 1ère espèce car il est lié au rejet de H_0 et la notion de rejet semble plus définitive alors que l'acceptation (ou le non rejet) de H_0 peut toujours être confirmé ou infirmé par un autre test).
3. Choisir une v.a. U (appelée statistique, en pratique elle est donnée) dépendant de $X_1, \ldots, X_n$ et du paramètre $\theta = \mu$ (ici) à tester dont la loi est connue sous H_0 (càd

lorsque H_0 est vraie). Par exemple, si X suit une loi normale de moyenne μ, alors

$$U = \frac{\hat{X} - \mu}{S_c/\sqrt{n}} \simeq \mathcal{T}(n-1) \quad \text{(sous } H_0\text{)}.$$

où

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \quad \text{et} \quad S_c = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}.$$

Ici, lorsque H_0 est vraie, on a $\mu = 19$ et

$$U = \frac{\overline{X} - 19}{S_c/\sqrt{8}} \sim \mathcal{T}(7).$$

La loi de U dépend des hypothèses : si H_1 est vraie, U aura tendance à prendre des valeurs négatives alors que sous H_0, U aura tendance à prendre des valeurs autour de 0.

4. Déterminer une zone de rejet de H_0 notée R_α vérifiant :

$$\text{Sous } H_0, \quad \mathbb{P}(U \in R_\alpha) = \alpha.$$

On choisit la forme de la zone de rejet R_α en examinant le comportement de U sous H_1 : dans notre exemple on prendra $R_\alpha =]-\infty, t_\alpha[$ avec $t_\alpha < 0$ que l'on détermine donc par

$$\mathbb{P}(U < t_\alpha) = \alpha = 0.05$$

ce qui donne (d'après la table de $\mathcal{T}(7)$) : $t_\alpha = -1.895$ et $R_\alpha =]-\infty, -1.895[$.

5. Utiliser la règle de décision suivante : calculer la valeur u de U observée sur l'échantillon et regarder
   – si $u \in R_\alpha$, on rejette H_0,
   – si $u \notin R_\alpha$, on ne rejette pas H_0.

   Dans notre exemple, on a $\bar{x} = 16$ et $s_c = 3.29$, d'où

$$u = \frac{16 - 19}{3.29/\sqrt{8}} = -2.58$$

   valeur qui est dans $R_\alpha =]-\infty, -1.895[$. On rejette donc H_0 (avec un risque d'erreur de 5%).

6. Si H_0 est rejetée, c'est que le produit réduit effectivement le temps moyen de réaction.

Le choix de H_0 est parfois dicté par le bon sens ; par exemple imaginons un diagnostic pour une maladie grave :

– Décider que le patient est malade à tort entraîne des traitements désagréables.
– Décider que le patient n'est pas malade à tort entraîne des conséquences plus graves.

Dans ce cas mieux vaut poser : H_0 « le patient est malade »puisque l'on peut contrôler le risque de 1ère espèce, qui correspond à l'erreur la plus lourde de conséquences.

Remarque :

• **Acceptation/non rejet.**

En général, un test négatif amène à rejeter une hypothèse mais un test positif n'amène jamais à accepter d'emblée une hypothèse. Dans le meilleur des cas, on ne rejettera pas l'hypothèse d'emblée.

Concrètement, cela se comprend de la façon suivante : imaginez que vous perdiez un bouton de chemise. Si vous en trouvez un par hasard, vous pouvez faire l'hypothèse H_0 : « le bouton trouvé est mon bouton perdu ». Vous pouvez faire des tests (taille, couleur, forme, nombre de trous, etc). Si l'un de ces tests est négatif, alors vous rejeterez l'hypothèse H_0. Mais tous les tests positifs ne pourront jamais prouver que l'hypothèse H_0 est vraie (au maximum, ils créeront une présomption de vérité pour H_0 mais aucune certitude).

• **Tests unilatéral et bilatéral.** Le test est bilatéral lorsque l'hypothèse alternative H_1 est symétique (par exemple « $\mu \neq \mu_0$ »), il est unilatéral sinon (par exemple H_1' : « $\mu > \mu_0$ »ou H_1'' : « $\mu < \mu_0$ »)

Dans le cas bilatéral, pour un seuil d'erreur de α, la zone de rejet R_α devra vérifier $\mathbb{P}(|U| \geq t_\alpha) = \alpha$ et (donc par symétrie) $\mathbb{P}(U \leq -t_\alpha) = \mathbb{P}(U \geq t_\alpha) = \alpha/2$.

Dans le cas unilatéral (toujours pour un seuil d'erreur α), la zone de rejet R_α devra vérifier

- $\mathbb{P}(U \geq t_\alpha) = \alpha$ dans le cas H_1' : « $\mu > \mu_0$ »
- ou $\mathbb{P}(U \leq -t_\alpha) = \alpha$ dans le cas H_1'' : « $\mu < \mu_0$ ».

Il faut donc veiller à savoir si on fait un test unilatéral ou bilatéral pour voir si c'est $\alpha/2$ ou α qui est à rechercher dans la table correspondante.

• **Risque exact.** Le choix préalable du risque α est facultatif. Si l'on n'a pas choisi de risque d'erreur α, on peut quand même pratiquer le test et calculer la valeur u observée sur l'échantillon de la statistique de test U. On peut alors chercher dans la table correspondante la valeur de α telle que u et t_α soient numériquement proches. On appelle cette valeur α_{reel} de α le **risque exact** pour une décision de rejet.

La procédure de décision est alors la suivante :

- si α_{reel} est grand ($\simeq$ 10% ou plus), il y a un risque notable à rejeter H_0. On proposera alors le non-rejet de H_0.
- Si α_{reel} est intermédiaire (entre $0,5\%$ et 10%), on se contentera d'indiquer le risque exact d'un rejet de H_0.
- Si α_{reel} est petit (moins de $0,5\%$), on propose le rejet de H_0.

Dans l'exemple précédent, avec $t_\alpha = 2,58$, on trouve dans la table de $\mathcal{T}(7)$ que le α correspondant est entre $0,05$ et $0,02$. Le risque réel α_{reel} est donc compris entre 2% et 5%.

4.2 Test sur la moyenne

On suppose qu'on a un échantillon gaussien ou alors que sa taille est suffisamment grande pour qu'on puisse l'approcher par une loi gaussienne.

On suppose donc que la variable considérée suit une loi $\mathcal{N}(\mu, \sigma^2)$ et on s'intéresse à la moyenne théorique μ, supposée inconnue. Certaines circonstances amènent à formuler la question suivante :

La moyenne théorique μ est-elle égale à une certaine valeur μ_0 ?

Pour cela, on désire faire le test suivant : H_0 : « $\mu = \mu_0$ » contre H_1 : « $\mu \neq \mu_0$ ».

Supposons σ^2 connue. Dans ce cas, on considère la statistique

$$U = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \simeq \mathcal{N}(0,1) \quad (\text{sous } H_0)$$

où $\overline{X}$ est l'estimation ponctuelle de μ sur l'échantillon. On définit une zone rejet R_α de la forme

$$R_\alpha =]-\infty, -t_\alpha[\cup]t_\alpha, +\infty[$$

où le nombre t_α est donné par la table $\mathcal{N}(0,1)$ de la v.a. U. avec

$$\mathbb{P}(|U| > t_\alpha) = \alpha \quad \text{càd} \quad \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2. \tag{4.1}$$

Noter que t_α est lié avec le risque de 1ère espèce.

Si on choisit $\alpha = 0.05$, on a $t_\alpha = 1.96$ d'après la table $\mathcal{N}(0,1)$. Et si choisit $\alpha = 0.1$, on a $t_\alpha = 1.645$.

Il reste alors à calculer la valeur u de U à partir de l'échantillon et à se décider en fonction de $u \in R_\alpha$ ou non.

$$\begin{cases} \text{Si } u \in R_\alpha, \text{ alors rejette } H_0 \text{ avec un risque d'erreur de } \alpha\% \\ \text{Si } u \notin R_\alpha, \text{ alors on ne rejette pas } H_0 \text{ avec un risque d'erreur de } \alpha\% \end{cases}$$

Supposons σ^2 inconnue. Dans ce cas, on considère la statistique

$$U = \frac{\overline{X} - \mu_0}{S_c/\sqrt{n}} \simeq \mathcal{T}(n-1) \quad (\text{sous } H_0)$$

où $S_c^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

La procédure est la même que précedemment mais avec t_α dans (4.1) à chercher dans la table de $\mathcal{T}(n-1)$ de la v.a. $U \simeq \mathcal{T}(n-1)$.

Remarques.

- Ces deux tests sont encore valables dans le cas non gaussien si l'échantillon est assez grand (de taille $n \geq 30$).
- Si on teste l'hypothèse alternative H_1' : « $\mu > \mu_0$ », il faut prendre $R_\alpha = [t_\alpha, +\infty[$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

– Si on teste l'hypothèse alternative H_1'' : « $\mu < \mu_0$ », il faut prendre $R_\alpha =]-\infty, -t_\alpha]$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

4.3 Test sur la variance dans le cas gaussien

On suppose que la variable considérée suit une loi $\mathcal{N}(\mu, \sigma^2)$ et on s'intéresse à la variance théorique σ^2, supposée inconnue. Certaines circonstances mènent à formuler la question suivante :

La variance théorique σ^2 est-elle égale à une certaine valeur σ_0^2 ?

On définit le test suivant dans le cas où μ est inconnu, avec un risque α : l'hypothèse à tester est H_0 : « $\sigma^2 = \sigma_0^2$ »contre H_1 : « $\sigma^2 \neq \sigma_0^2$ ».

On considère la statistique

$$U = \frac{(n-1)S_c^2}{\sigma_0^2} \simeq \chi^2(n-1) \quad \text{(sous } H_0\text{)}$$

où $S_c^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. On définit la zone de rejet R_α par

$$R_\alpha = [0, a_\alpha[\cup]b_\alpha, +\infty[$$

où a_α et b_α sont donnés par la table de $\chi^2(n-1)$ pour la v.a. U. avec les équations

$$\mathbb{P}(U \leq a_\alpha) = \mathbb{P}(U \leq b_\alpha) = \alpha/2.$$

Enfin, on calcule $u = U(\omega)$ et on regarde si $u \in R_\alpha$ ou si $u \notin R_\alpha$ pour conclure.

Remarques.
– Ce test est encore valable dans le cas non gaussien si l'échantillon est assez grand (de taille $n \geq 30$).
– Si on teste l'hypothèse alternative H_1' : « $\sigma^2 > \sigma_0^2$ », il faut prendre la zone de rejet $R_\alpha = [b_\alpha, +\infty[$ avec $\mathbb{P}(U \leq b_\alpha) = 1 - \alpha$.
– Si on teste l'hypothèse alternative H_1'' : « $\sigma^2 < \sigma_0^2$ », il faut prendre la zone de rejet $R_\alpha =]0, a_\alpha]$ avec $\mathbb{P}(U \leq a_\alpha) = \alpha$.

4.4 Test sur une proportion

On teste ici la proportion théorique (vraie et inconnue) p d'individus possédant une certaine caractéristique C, dans une population donnée. On souhaite le comparer à une proportion p_0 de référence. Dans cette situation, on observe sur chaque individu d'un échantillon de taille n la présence ou l'absence de la caractéristique C.

Si on observe n_1 fois le caractère étudié, on va estimer p par $f = \frac{n_1}{n}$.

Lorsque $n \geq 30$, $nf \geq 10$, $n(1-f) \geq 10$, on peut considèrer le test : H_0 : « $p = p_0$ »contre H_1 : « $p \neq p_0$ »avec la statistique de test

$$U = \frac{\sqrt{n}(f - p_0)}{\sqrt{f(1-f)}} \simeq \mathcal{N}(0,1) \quad \text{(sous } H_0\text{)}.$$

On définit la zone de rejet R_α de la forme

$$R_\alpha =]-\infty, -t_\alpha[\cup]t_\alpha, +\infty[$$

où t_α est donné dans la table de $\mathcal{N}(0,1)$ pour la v.a. U. par l'équation

$$\mathbb{P}(|U| > t_\alpha) = \alpha \quad \text{càd} \quad \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2.$$

Enfin, on calcule $u = U(\omega)$ et on regarde si $u \in R_\alpha$ ou si $u \notin R_\alpha$.

Si on teste l'hypothèse alternative H_1 : « $\mu > \mu_0$ », il faut prendre $R_\alpha = [t_\alpha, +\infty[$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

Remarques.

Si on teste l'hypothèse alternative H_1 : « $p > p_0$ », il faut prendre $R_\alpha = [t_\alpha, +\infty[$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

Si on teste l'hypothèse alternative H_1 : « $p < p_0$ », il faut prendre $R_\alpha =]-\infty, -t_\alpha]$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

4.5 Tests de comparaison

4.5.1 Comparaison de deux moyennes

On considère deux populations sur lesquelles sont définies deux caractères numériques X et Y distribués selon des lois de moyennes μ_1 et μ_2 et de même variance σ^2 (inconnue). On souhaite tester s'il y a une différence significative entre les moyennes des deux populations. L'hypothèse nulle à tester est H_0 : « $\mu_1 = \mu_2$ »contre H_1 : « $\mu_1 \neq \mu_2$ ».

On dispose d'un échantillon de taille n_1 pour X et de taille n_2 pour Y. On introduit préliminairement

$$S_c^2 = \frac{(n_1 - 1)S_{c,1}^2 + (n_2 - 1)S_{c,2}^2}{n_1 + n_2 - 2}$$

avec les variances empiriques corrigées de X : $S_{c,1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$ et de Y : $S_{c,2}^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$. Puis soit

$$S_c^* = S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

On considère la statistique

$$U = \frac{\bar{X}_1 - \bar{X}_2}{S_c^*} \simeq \mathcal{T}(n_1 + n_2 - 2) \quad \text{(sous } H_0\text{)}$$

où $\bar{X}_1$ et $\bar{X}_2$ sont les estimations ponctuelles de μ_1 et μ_2. On définit la zone de rejet R_α par

$$R_\alpha =]-\infty, -t_\alpha] \cup [t_\alpha, +\infty[$$

où t_α est à déterminer dans la table de Student $\mathcal{T}(n_1 + n_2 - 2)$ de la v.a. U avec

$$\mathbb{P}(|U| > t_\alpha) = \alpha \quad \text{càd} \quad \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2.$$

On conclut en calculant u à partir des échantillons de X et de Y et en testant si $u \in R_\alpha$ ou pas.

Remarques.

Si on teste l'hypothèse alternative H_1 : « $\mu_1 > \mu_2$ », il faut prendre $R_\alpha = [t_\alpha, +\infty[$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

Si on teste l'hypothèse alternative H_1 : « $\mu_1 < \mu_2$ », il faut prendre $R_\alpha =]-\infty, -t_\alpha]$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

4.5.2 Comparaison de deux proportions

On compare deux proportions inconnues p_1 et p_2. On souhaite tester si ce sont les mêmes. L'hypothèse nulle à tester est H_0 : « $p_1 = p_2$ »contre H_1 : « $p_1 \neq p_2$ ».

On dispose de deux séries d'observations, de taille n_1 pour p_1 qu'on estime par f_1 et de taille n_2 pour p_2 qu'on estime par f_2. Soit

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

et $S^* = \sqrt{f(1-f)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. On considère la statistique

$$U = \frac{f_1 - f_2}{S^*} \simeq \mathcal{N}(0,1) \quad (\text{sous } H_0).$$

On définit la zone de rejet R_α par

$$R_\alpha =]-\infty, -t_\alpha] \cup [t_\alpha, +\infty[$$

où t_α est à déterminer dans la table normale $\mathcal{N}(0,1)$ de la v.a. U avec

$$\mathbb{P}(|U| > t_\alpha) = \alpha \quad \text{càd} \quad \mathbb{P}(U \leq t_\alpha) = 1 - \alpha/2.$$

On conclut en calculant u à partir des deux séries observées et en testant si $u \in R_\alpha$ ou pas.

Remarques.

Si on teste l'hypothèse alternative H_1 : « $p_1 > p_2$ », il faut prendre $R_\alpha = [t_\alpha, +\infty[$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

Si on teste l'hypothèse alternative H_1 : « $p_1 < p_2$ », il faut prendre $R_\alpha =]-\infty, -t_\alpha]$ avec $\mathbb{P}(U \leq t_\alpha) = 1 - \alpha$.

4.6 Les Tests du χ^2

4.6.1 Principe

On peut distinguer trois types de test du χ^2 :
- le test du χ^2 d'adéquation (H_0 : « le caractère X suit-il une loi particulière ? »),
- le test du χ^2 d'homogénéité (H_0 : « le caractère X suit-il la même loi dans deux populations données ? ») ,
- le test du χ^2 d'indépendance (H_0 : « les caractères X et Y sont-ils indépendants ? »).

Ces trois tests ont un principe commun qui est le suivant : on répartit les observations dans k classes dont les effectifs sont notés $n_1 = N_1(\omega), \ldots, n_k = N_k(\omega)$. L'hypothèse H_0 permet de calculer les effectifs théoriques, notés $n_{1,\text{th}}, \ldots, n_{k,\text{th}}$. On rejette H_0 si les effectifs observés sont trop différents des effectifs théoriques.

Pour cela on utilise la statistique de test

$$U = \sum_{i=1}^{k} \frac{(N_i - n_{i,\text{th}})^2}{n_{i,\text{th}}}.$$

Cette statistique suit la loi $U \simeq \chi^2(k-1-m)$ où k est le nombre de classes et m est le nombre de paramètres estimés nécessaires au calcul des effectifs théoriques.

Il faut s'assurer que les effectifs théoriques sont plus grands que 5 et donc faire des regroupements de classes si besoin est.
À partir de là, on calcule la zone de rejet unilatérale $R_\alpha =]t_\alpha, +\infty[$ au risque α en déterminant t_α dans la table de $\chi^2(k-1-m)$ par

$$\mathbb{P}(U > t_\alpha) = \alpha.$$

La règle de décision est la suivante :
- Si $u = \sum_{i=1}^{k} \frac{(n_i - n_{i,\text{th}})^2}{n_{i,\text{th}}}$ appartient à R_α, on rejette H_0,
- Si $u = \sum_{i=1}^{k} \frac{(n_i - n_{i,\text{th}})^2}{n_{i,\text{th}}}$ n'appartient pas à R_α, on accepte H_0.

Remarque :
- Contrairement aux autres tests, les tests du χ^2 n'exigent pas de formuler l'hypothèse alternative H_1, qui correspond à la négation de H_0.
- Les effectifs théoriques doivent être supérieurs à 5. Si ce n'est pas le cas, il faut regrouper des classes.
- Dans la statistique $U \simeq \chi^2(k-1-m)$, on manipule des effectifs et non des pourcentages.

4.6.2 Exemples

Exemple 1. Un croisement entre roses rouges et blanches a donné en seconde génération des roses rouges, roses et blanches. Sur un échantillon de taille 600, on a trouvé

les résultats suivants :

couleur	effectifs
rouges	141
roses	315
blanches	144

Peut-on affirmer que les résultats sont conformes aux lois de Mendel ?

Il s'agit donc de tester

$H_0 : p_{\text{rouges}} = 0.25$, $p_{\text{roses}} = 0.5$, $p_{\text{blanches}} = 0.25$ au risque disons $\alpha = 0.05$.

On dresse alors le tableau suivant :

couleur	effectifs observés n_i	effectifs théoriques $n_{i,\text{th}}$
rouges	141	$0.25 \times 600 = 150$
roses	315	$0.5 \times 600 = 300$
blanches	144	$0.25 \times 600 = 150$

Ici on a $k = 3$ classes et $m = 0$ (aucun paramètre à estimer pour pouvoir calculer les effectifs théoriques) donc $k - 1 - m = 2$; on calcule ensuite $R_\alpha =]t_\alpha, +\infty[$ à l'aide de la table de $\chi^2(2)$ et on obtient $t = 5.991$. Enfin, on calcule

$$u = U(\omega) = \frac{(141 - 150)^2}{150} + \frac{(315 - 300)^2}{300} + \frac{(144 - 150)^2}{150} = 1.53 \notin R_\alpha.$$

On propose le non rejet de l'hypothèse : on ne peut pas dire que les observations contredisent la loi de Mendel.

Exemple 2. On observe le nombre X d'accidents journaliers sur une période de 50 jours dans une certaine ville. On obtient :

Nombre d'accidents	Nombre de jours
0	21
1	18
2	7
3	3
4	1

On constate que $\bar{x} = 0.9$ et $s^2 = 0.97$. Peut-on affirmer que X suit une loi de Poisson ? (risque $\alpha = 0.05$)

H_0 : X suit une loi de Poisson de paramètre 0.9

On dresse donc le tableau suivant :

Nombre d'accidents	Nombre de jours	Nombre de jours théorique
0	21	$50 \times \mathrm{e}^{-0.9} = 20.330$
1	18	$50 \times \mathrm{e}^{-0.9} \times 0.9 = 18.295$
au moins 2	11	$50 \times (1 - \mathrm{e}^{-0.9}(1 + 0.9)) = 11.376$

On a regroupé les 3 dernières classes pour avoir un effectif théorique ≥ 5 dans la dernière classe. Dans cet exemple 2, on a $k = 3$ classes et $m = 1$ paramètre estimé à savoir $\lambda \simeq \bar{x}$ nécessaire au calcul des effectifs théoriques ; donc $k-1-m = 1$ est le nombre de d.d.l. de U. On calcule ensuite $R_\alpha =]t, +\infty[$ à l'aide de la table de $\chi^2(1)$ et on obtient $t = 3.841$. Enfin, on calcule

$$u = U(\omega) = \frac{(21 - 20.33)^2}{20.33} + \frac{(18 - 18.295)^2}{18.295} + \frac{(11 - 11.376)^2}{11.376} = 0.039 \notin R_\alpha.$$

On ne rejette pas H_0 au risque d'erreur 0.05.