Table des matières

Première partie

Notions de probabilité

Chapitre 1

Notions de probabilité

L'objet de toute étude statistique est de formuler des lois valables pour un ensemble d'êtres ou d'éléments auquel on donne le nom de **population**.

1.1 Définitions

Définition 1.1 *On appelle population (notée Ω) un ensemble d'éléments ou individus (notés ω), possédant au moins une caractéristique commune et exclusive, permettant de l'identifier et de le distinguer sans ambiguïté de tout autre.*

Exemples :

1. Ω_1 ="Ensemble des étudiants de la maîtrise de biologie cellulaire à St Charles" ;
2. Ω_2 = "Ensemble des poissons d'un lac" ;
3. Ω_3 = "Ensemble des familles d'un pays" ;
4. Ω_4 = "Ensemble de cellules".

Ω_1 constitue bien une population car l'élément constitutif de la population est bien identifié (une personne), et ses caractéristiques communes et exclusives : il est étudiant ET en maîtrise de biologie cellulaire ET à St Charles.

Remarque : Un individu n'est pas forcément un individu au sens usuel. Pour Ω_1, l'individu ω est une personne. Pour Ω_3, l'individu (ω) est une famille.

Définition 1.2 *On appelle variable aléatoire (notée T; X; Y ... Z) une application qui à un individu ω associe une valeur numérique appelée réalisation de la variable aléatoire (notée $T(\omega)=t$; $X(\omega)=x$; $Y(\omega)=y$ ... $Z(\omega)=z$).*

Exemples : L'observation des cellules permet de définir la variable aléatoire "état de la cellule" :

$$\begin{array}{ccll} \Omega_4 & \longrightarrow & \{0,1\} & \\ \omega_i & \longmapsto & C(\omega_i) & = 1 \text{ si la cellule i est cancéreuse} ; \\ & & & = 0 \text{ sinon.} \end{array}$$

Un jet de dés permet de définir la variable aléatoire “numéro de la face tirée” :

$$\begin{array}{ccccl} \Omega_1 & \longrightarrow & \{1,2,3,4,5,6\} & & \\ \omega_i & \longmapsto & X(\omega_i) & = x_i = & \text{numéro de la face tirée} \end{array}$$

L'observation du groupe sanguin d'un individu dans la population humaine permet de définir la variable aléatoire “groupe sanguin” :

$$\begin{array}{ccccl} \Omega_2 & \longrightarrow & \{1,2,3,4\} & & \\ \omega_i & \longmapsto & X(\omega_i) & = 1 & \text{si le groupe sanguin est de type A} \\ & & & = 2 & \text{si le groupe sanguin est de type B} \\ & & & = 3 & \text{si le groupe sanguin est de type AB} \\ & & & = 4 & \text{si le groupe sanguin est de type O} \end{array}$$

L'observation de la taille d'un poisson permet de définir la variable aléatoire “taille” :

$$\begin{array}{ccccl} \Omega_3 & \longrightarrow & \{\mathbb{R}^+\} & & \\ \omega_i & \longmapsto & X(\omega_i) & = x_i = & \text{taille du poisson} \end{array}$$

Définition 1.3 *On appelle espace fondamental l'ensemble E constitué de tous les résultats possibles de la variable aléatoire.*

Exemple : Pour la variable aléatoire “jet de dés”, $E = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$; pour la variable aléatoire “groupe sanguin”, $E = \{1,\ 2,\ 3,\ 4\}$; pour la variable aléatoire “taille du poisson”, $E = \mathbb{R}^+$.

Définition 1.4 *On appelle événement -noté A, B, ...- une proposition concernant les résultats de la variable aléatoire, ce qui est équivalent à un sous ensemble de l'espace fondamental E. On notera $\mathcal{E}$ l'ensemble des événements associé à E.*

Exemple 1 L'événement “obtenir un résultat pair” pour la variable aléatoire “jet de dés” est $A1 = \{2,\ 4,\ 6\}$; l'événement “obtenir un résultat impair” pour la variable aléatoire “jet de dés” est $A2 = \{1,\ 3,\ 5\}$; l'événement “avoir un 1" pour la variable aléatoire ”jet de dés" est $A3 = \{1\}$, l'événement “obtenir un 7” pour la variable aléatoire “jet de dés” est $\emptyset$. Ces quatre ensembles font partie de $\mathcal{E} = \{A1,\ A2,\ A3,\ \emptyset,\ ...,\ E\}$.

Exemple 2 L'événement “ne pas être du groupe sanguin A” pour la variable aléatoire “groupe sanguin” est $A1 = \{2,\ 3,\ 4\}$; L'événement “être donneur universel” pour la variable aléatoire “groupe sanguin” est $A1 = \{3\}$.

Exemple 3 L'événement “être compris entre 10 et 12 cm” pour la variable aléatoire “taille du poisson” est $A1 = [10,\ 12]$; l'événement “être strictement supérieur à 12” pour la variable aléatoire “taille du poisson” est $A2 =]12,\ +\infty[$.

La première caractéristique d'une variable aléatoire, c'est qu'on ne connaît pas avant sa réalisation, la valeur qu'elle prend. Puisqu'on ne peut connaître la valeur prise par la variable aléatoire considérée, on va s'intéresser aux chances ou probabilités de tout événement. “Connaître” la variable aléatoire ce n'est pas connaître à l'avance les résultats, c'est connaître quelle est la chance ou probabilité de réalisation de tout événement.

1.2 Définition d'une probabilité

1.2.1 Espace probabilisable

Définition 1.5 *Soit Ω un ensemble, on appelle espace probabilisable la donnée d'un couple $(\Omega, \mathcal{A})$ où $\mathcal{A}$ est une collection de sous ensembles de Ω vérifiant :*

1. $\Omega \in \mathcal{A}$
2. *Si* $A \in \mathcal{A}$ *alors* $A^c \in \mathcal{A}$
3. *Pour toute suite* $(A_n)_{n\in\mathbb{N}}$ *d'éléments de* $\mathcal{A}$, $\cup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

$\mathcal{A}$ est dite tribu associée à Ω.

Exemple : Soit $E = \{1, 2, 3, 4, 5, 6\}$ de l'exemple "jet de dé". Si on définit $\mathcal{E} = \mathcal{P}(E)$ =ensemble des parties de E, alors $(E, \mathcal{E})$ constitue un espace probabilisable. $\mathcal{E}$ correspond alors à l'ensemble des événements qu'il est possible d'associer à la variable aléatoire "jet de dé".

1.2.2 Probabilité

Définition

Définition 1.6 *Soit $(\Omega, \mathcal{A})$ un espace probabilisable. On appelle probabilité P associée à $(\Omega, \mathcal{A})$ l'application :*

$$\begin{array}{rccll} P: & \mathcal{A} & \longrightarrow & [0,\ 1] & \\ & A & \longmapsto & P(A) & \text{telle que :} \end{array}$$

- $P(\Omega) = 1$;
- *Pour toute suite* $(A_i)_i$ *d'évenements deux à deux disjoints (ie* $A_i \cap A_j,\ \forall i \neq j$*) alors* $P(\cup_i A_i) = \sum_i P(A_i)$

Définition 1.7 *Le triplet $(\Omega, \mathcal{A}, P)$ est dit espace probabilisé.*

Propriété 1.1 *Soient A et B deux événements quelconques, alors on a toujours :*

$$\begin{array}{lcl} P(\emptyset) & = & 0 \\ P(A^c) & = & 1 - P(A) \\ P(A \cup B) & = & P(A) + P(B) - P(A \cap B) \\ A \subseteq B & \Rightarrow & P(A) \leq P(B). \end{array}$$

Propriété 1.2 *Quand l'espace fondamental est fini et si ses éléments sont équiprobables alors la probabilité d'obtenir un événement A est notée $P(A)$, avec $P(A) = \dfrac{CardA}{CardE}$. Le cardinal de A désigne le nombre d'éléments de A.*

Exemples :

1. Soit un dé, quelle est la probabilité d'obtenir un nombre pair ?
   $E = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ et $A = \{\ 2,\ 4,\ 6\}$. Donc $CardE = 6$ et $CardA = 3$. La probabilité d'obtenir un nombre pair est donc $\mathrm{P}(A) = \frac{CardA}{CardE} = 1/2$.

2. Soit un dé, quelle est la probabilité d'obtenir un nombre supérieur ou égal à 3 ?
   $E = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$ et $B = \{\ 3,\ 4,\ 5,\ 6\}$. Donc $CardE$= 6 et $CardB$= 4. La probabilité d'obtenir un nombre supérieur ou égal à 3 est donc P(B)=$\frac{CardB}{CardE} = 2/3$.

3. Soit un dé, quelle est la probabilité d'obtenir un nombre pair et supérieur ou égal à 3 ?
   Pour cela nous calculons la probabilité de l'intersection des événements A et B :
   $A \cap B = \{\ 4,\ 6\}$ d'où $\mathrm{P}(A \cap B) = \dfrac{CardA \cap B}{CardE} = 1/3$.

Loi de probabilité d'une variable aléatoire

Définition 1.8 *Soit X une variable aléatoire réelle, E son espace fondamental muni de $\mathcal{E}$. On definit P_X la probabilité associée à X par : $P_X(A) = P\{\omega\ /\ X(\omega) \in A\} = P(X \in A)$. P_X est dite distribution théorique ou encore loi de probabilité de X.*

Exemple : Soit Ω="Ensemble des étudiants de la maîtrise" et $\mathcal{A}$= "Ensemble de tous les groupes d'étudiants de maîtrise qu'il est possible de constituer". On définit alors sur $(\Omega, \mathcal{A})$ la probabilité :

$$\begin{array}{rccl} P & (\Omega, \mathcal{A}) & \longrightarrow & [0,1] \\ & A & \longmapsto & P(A) = \dfrac{cardA}{card\Omega} \end{array}$$

Soit X la variable "groupe sanguin" :

$$\begin{array}{ccccll} \Omega & \longrightarrow & (\{1,2,3,4\},\ \mathcal{E} & & & \\ \omega & \longmapsto & X(\omega) & = 1 & \text{si le groupe sanguin est de type A} \\ & & & = 2 & \text{si le groupe sanguin est de type B} \\ & & & = 3 & \text{si le groupe sanguin est de type AB} \\ & & & = 4 & \text{si le groupe sanguin est de type O} \end{array}$$

Soit l'événement "être donneur universel"= $A = \{4\}$. Alors on a :

$$P_X(A) = P_X\{4\} = P\{\omega\ /\ X(\omega) = 4\} = \frac{\{omega\ /\ X(\omega) = 4\}}{card\Omega}$$

Probabilité conditionnelle

Définition 1.9 *Si A et B sont deux événements (avec $P(A) \neq 0$), on appelle probabilité conditionnelle de B quand A est réalisé, le nombre réel défini par $P(B/A) \doteq \dfrac{P(A \cap B)}{P(A)}$.*

Exemple : Jet de dé $\begin{array}{ccc} (\Omega, \mathcal{A}, P) & \longrightarrow & (E = \{1,2,3,4,5,6\},\ \mathcal{E},\ P_X) \\ \omega & \longmapsto & X(\omega) \end{array}$

Considérons les deux événements : A="Obtenir un résultat pair" et B="Obtenir un résultat supérieur ou égal à 4".

$P_X(B/A) \doteq \dfrac{P_X(A \cap B)}{P_X(A)} = \dfrac{P_X\{4,\ 6\}}{P_X\{2,\ 4,\ 6\}}$. Nous obtenons donc $\dfrac{1/3}{1/2} = 2/3$.

1.3 Indépendance

1.3.1 Indépendance de deux événements

Définition 1.10 *Deux événements A et B sont dits indépendants si et seulement si $P(A \cap B) = P(A) \times P(B)$.*

Remarques :

1. Si A et B sont indépendants alors $P(B/A) = P(B)$
2. "indépendance" signifie que la réalisation de l'événement A ne modifie pas les chances ou la probabilité de réalisation de l'événement B
3. Pour montrer la dépendance de deux événements A et B, il suffit de vérifier que $P(A \cap B) \neq P(A) \times P(B)$

Exemple :

1. Considérons l'exemple précédent (obtenir un résultat pair et supérieur ou égal à 4). La probabilité de l'intersection des événements A et B ($A \cap B = \{\ 4,\ 6\}$) est égale à : $\mathrm{P}(A \cap B) = \dfrac{CardA \cap B}{CardE} = 1/3$.
   Si nous calculons $P(A) \times P(B)$, nous obtenons $P(A) \times P(B) = 1/2 \times 1/2 = 1/4$ donc $P(A \cap B) \neq P(A) \times P(B)$. Les événements A et B ne sont pas indépendants.
2. Reprenons l'exemple précédent en considérant B_1 l'événement "obtenir un résultat supérieur ou égal à 3". La probabilité de l'intersection des événements A et B_1 ($A \cap B_1 = \{\ 4,\ 6\}$) est égale à : $\mathrm{P}(A \cap B_1) = \dfrac{CardA \cap B_1}{CardE} = 1/3$.
   Si nous calculons $P(A) \times P(B_1)$, nous obtenons $P(A) \times P(B_1) = 1/2 \times 2/3 = 1/3$ donc $P(A \cap B_1) = P(A) \times P(B_1)$. Les événements A et B_1 sont donc indépendants.

1.3.2 Indépendance de deux variables aléatoires

Définition 1.11 *Soient X et Y deux variables aléatoires réelles définies sur la même population. X et Y sont dites variables aléatoires indépendantes si et seulement si $\forall(a_1,\ a_2,\ b_1,\ b_2)\ /\ a_1 < b_1 \wedge a_2 < b_2$, $P\{\omega / a_1 < X(\omega) < b_1 \wedge a_2 < Y(\omega) < b_2\} = P_X[a1,\ b1] \times P_Y[a2,\ b2]$.*

Exemple : Considérons un jeu de 32 cartes (ω), les variables X "forme de la carte" et Y "couleur de la carte" sont indépendantes.

$$
\begin{array}{lcll}
\Omega & \longrightarrow & \{1,2,3,4,5,6,7,8\} & \\
\omega_i & \longmapsto & X(\omega_i) & = x_i = 1 \quad \text{si la carte est un “7"} \\
 & & & = x_i = 2 \quad \text{si la carte est un “8"} \\
 & & & = x_i = 3 \quad \text{si la carte est un “9"} \\
 & & & = x_i = 4 \quad \text{si la carte est un “10"} \\
 & & & = x_i = 5 \quad \text{si la carte est un “Valet"} \\
 & & & = x_i = 6 \quad \text{si la carte est une “Dame"} \\
 & & & = x_i = 7 \quad \text{si la carte est un “Roi"} \\
 & & & = x_i = 8 \quad \text{si la carte est un “As"} \\
\Omega & \longrightarrow & \{0,1\} & \\
\omega_i & \longmapsto & Y(\omega_i) & = y_i = 0 \quad \text{si la carte est rouge} \\
 & & & = y_i = 1 \quad \text{si la carte est noire}
\end{array}
$$

Montrons par exemple que les deux événements C= “obtenir une carte superieur ou égale au valet” et D= “obtenir une carte roug” sont indépendants.

$P(C \cap D) = P\{\omega / 5 \leq X(\omega) \leq 8 \wedge Y(\omega) = 0\} = P_X[5,8] \times P_Y\{0\}$.

$Card\ C = 16$
$Card\ D = 16$
$Card\ E = 32$
$Card\ (C \cap D) = Card\{\omega / 5 \leq X(\omega) \leq 8 \wedge Y(\omega) = 0\} = 8$

La probabilité de l'intersection des événements C et D est :

$$P(C \cap D) = \frac{Card\{\ Valet\heartsuit,\ Valet\diamondsuit,\ Dame\heartsuit,\ Dame\diamondsuit,\ Roi\diamondsuit,\ Roi\heartsuit,\ As\diamondsuit,\ As\heartsuit\}}{CardE} = 8/32 = 1/4.$$

Si nous calculons $P(C) \times P(D)$, nous obtenons $P(C) \times P(D) = 16/32 \times 16/32 = 1/4$ donc $P(C \cap D) = P(C) \times P(D)$. Les événements C et D sont donc bien indépendants.

Interprétation : Intuitivement, deux variables aléatoires sont indépendantes si la distribution de l'une ne dépend pas des valeurs de l'autre. Par exemple, le poids et la tension arterielle seraient indépendantes si la distribution de la tension artérielle était la même quelque soit le poids. Pratiquement, cela signifie que si l'on regroupait les sujets d'une population en sous populations de sujets de même poids, la distribution de la tension artérielle serait la même dans toutes ces populations. Ce n'est pas le cas, ce qui signifie que ces variables aléatoires ne sont pas indépendantes.

1.4 Fonction de répartition d'une variable aléatoire

Définition 1.12 .

$$
\begin{array}{lcll}
Soit\ X\ (\Omega, \mathcal{A}, P) & \longrightarrow & (\mathbb{R},\ \mathcal{E},\ P_X) & \\
\omega & \longmapsto & X(\omega) & \textit{une variable aléatoire.}
\end{array}
$$

On appelle fonction de répartition de X noté $F_X(t)$, la probabilité pour que X soit inférieur à t, que l'on note : $P(X \leq t) = P_X\{\infty, t\}$. La fonction de répartition est donc la fonction :

$$\begin{array}{rcl} F_X : \mathbb{R} & \longrightarrow & [0,\ 1] \\ t & \longmapsto & F_X(t) = P(X \leq t) = P_X\{\infty, t\}. \end{array}$$

Remarque :

1. F est une fonction croissante
2. $\lim_{t \to -\infty} F_X(t) = P_X(\emptyset) = 0$
3. $\lim_{t \to +\infty} F_X(t) = P_X(E) = 1$

Nous allons dans le prochain chapitre étudier certaines lois de probabilités classiques, d'abord pour les variables aléatoires discrètes puis pour les variables aléatoires continues.

Chapitre 2

Lois de probabilité

Nous allons dans ce chapitre étudier certaines lois de probabilité ou distributions classiques. D'abord pour les variables aléatoires discrètes (*i.e.* dont E est fini ou dénombrable) puis pour les variables continues (*i.e.* quand E n'est pas dénombrable, par exemple E=$\mathbb{R}$).

2.1 Lois de probabilité discrètes

2.1.1 Généralités

Définition 2.1 *Une variable aléatoire est dite discrète lorsqu'elle prend un nombre fini ou dénombrable de valeurs.*

Remarque : Un ensemble est dénombrable lorsqu'il peut être mis en bijection avec $\mathbb{N}$. On peut se représenter un ensemble dénombrable, comme un ensemble dont on peut énumerer les valeurs.

Exemples :

"Jet de dé" : $X_1 : \Omega \longrightarrow E = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$
$\omega \longmapsto X(\omega) =$ numéro de la face.

"Groupe sanguin" : $X_2 : \Omega \longrightarrow E = \{1,\ 2,\ 3,\ 4\}$
$\omega \longmapsto X(\omega) = 1$ si le groupe sanguin est de type A
$= 2$ si le groupe sanguin est de type B
$= 3$ si le groupe sanguin est de type AB
$= 4$ si le groupe sanguin est de type O

"Pile ou face" : $X_3 : \Omega \longrightarrow E = \{0,\ 1\}$
$\omega \longmapsto X(\omega) = 0$ si c'est pile ;
$= 1$ si c'est face.

"Proies" : $X_4 : \Omega \longrightarrow E = \mathbb{N}$
$\omega \longmapsto X(\omega) =$ nombre de proies dans l'estomac du poisson ω

Propriété 2.1 *Pour une variable aléatoire discrète, tout événement peut être décrit comme la réunion d'éléments unitaires de l'espace fondamental E.*

Exemples :

$$\begin{aligned} A_1 &= \text{“obtenir un résultat pair”} \\ &= \{2,\ 4,\ 6\} \\ &= \{2\} \cup \{4\} \cup \{6\} \end{aligned}$$

$$\begin{aligned} A_2 &= \text{“Ne pas être donneur universel”} \\ &= \{1,\ 2,\ 4\} \\ &= \{1\} \cup \{2\} \cup \{4\} \end{aligned}$$

$$\begin{aligned} A_2 &= \text{“Avoir plus de deux proies”} \\ &= \{3,\ 4,\ 5, \ldots\} \\ &= \{3\} \cup \{4\} \cup \{5\} \cup \ldots \\ &= U_{i>2}\{i\} \end{aligned}$$

Définition 2.2 *La loi de probabilité ou distribution d'une variable aléatoire discrète dont les valeurs possibles sont $E = \{c_j,\ j = 0, 1, \ldots\}$ est donnée par l'ensemble des probabilités $\{p_j,\ j = 0, 1, \ldots\}$ de ses événements élémentaires, de sorte que $\sum_j p_j = 1$ et telle que :*
$P_X\{c_j\} = P(X = c_j) = P\{\omega / X(\omega) = c_j\} = p_j$.

Exemples : X_1= “jet de dé” $\{p_1, p_2, p_3, p_4, p_5, p_6\} = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$; X_3= “Pile ou face” $\{p_0, p_1\} = \{\frac{1}{2}, \frac{1}{2}\}$; X_3= “Pile ou face truqué” $\{p_0, p_1\} = \{\frac{1}{4}, \frac{3}{4}\}$

Remarques :

1. On est alors capable de calculer la probabilité associée à tout événement.

   Par exemple
   $$\begin{aligned} & P_{X_1}\{\text{ obtenir un chiffre pair}\} \\ = \ & P_{X_1}\{2,\ 4,\ 6\} \\ = \ & P_{X_1}(\{2\} \cup \{\ 4\} \cup \{\ 6\}) \\ = \ & P_{X_1}\{2\} + P_{X_1}\{4\} + P_{X_1}\{6\} \\ = \ & \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ = \ & \frac{1}{2} \end{aligned}$$
   Donc, si dispose pour une variable aléatoire discrète X de l'ensemble des probabilités associées à ses éléments unitaires, alors on connait parfaitement X au sens défini au chapitre 1 (*i.e.* on sait calculer la probabilité de tout événement concernant X)
2. On désigne souvent par loi discrète la distribution de probabilité d'une variable aléatoire discrète.

Propriété 2.2 *Deux variables aléatoires dicrètes X et Y sont indépendantes si et seulement si les événements $(X = c_i)$ et $(Y = c_j)$ sont indépendants, c'est à dire si $\forall (i, j) p_{ij} = p_i \times p_j$.*

2.1.2 Moments d'une distribution de probabilité discrète

Définition 2.3 *Nous appelons moyenne théorique ou espérance d'une variable aléatoire discrète le nombre lorsqu'il existe et a un sens :* $\mu = \mu(X) = \sum_{j \geq 0} p_j \times c_j$.

Exemple : Pour un jet de dé à 6 faces $\mu = \mu(X) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3,5$

Remarques :

1. Calculer la moyenne n'a pas toujours de sens. C'est le cas lorsque l'on considère les variables sexe, pile ou face, couleur des yeux.
2. La moyenne μ est un paramètre qui donne la valeur centrale de la distribution (valeur centrale n'est pas synonyme de milieu de l'étendue ou plage de variation). μ est une constante si et seulement si on connaît les probabilité p_j. Si on ne connaît pas les p_j (ex : dé truqué), μ existe mais n'est pas connu. On verra par la suite comment on peut en obtenir une estimation ainsi qu'une marge d'incertitude associée à cette estimation.

Définition 2.4 *Nous appelons variance théorique d'une variable aléatoire discrète, le nombre (lorsqu'il existe et a un sens) défini par :*

$$\sigma^2 = \sigma^2(X) = \mu(X - \mu(X))^2 = \sum_{j \geq 0} p_j (c_j - \mu(X))^2$$

Exemple : Quelle est la variance théorique de variable aléatoire "jet de dé" ?

$$\sigma^2(X) = \frac{1}{6}(1 - 3,5)^2 + \frac{1}{6}(1 - 3,5)^2 + \ldots + \frac{1}{6}(1 - 3,5)^2 = 2,916$$

Remarques :

1. $\sigma(X)$ est dit écart-type (par exemple, $\sigma(X) = 1,7$ pour le jet de dé). L'avantage est que l'unité est alors la même que celle de X.
2. Variance et écart-type sont des indices caractérisant la dispersion des valeurs de la variable aléatoire. Plus la variance est élevée, plus la dispersion est grande. Quelle serait la variance d'un dé truqué de la manière suivante : $(p_1, p_2, \ldots, p_6) = (0, 0, 0, 0, 0, 1)$? (réponse : 0).
3. $\sigma^2(X) = \mu(X^2) - \mu(X)^2$.

2.1.3 Distributions de probabilités discrètes usuelles

Loi de Bernoulli

La variable aléatoire discrète la plus simple est celle qui prend deux valeurs que l'on code généralement 0 et 1. Par exemple "être malade/ ne pas être malade", "pile / face", "Homme / Femme"

Définition 2.5 *Une variable aléatoire discrète qui prend deux valeurs 0 ou 1 suit ou est distribuée selon une loi de Bernoulli. On note* $X \hookrightarrow \mathcal{B}er(p)$ *ou* $\mathcal{L}(X) = \mathcal{B}er(p)$.

Remarque : Connaître la distribution de probabilité d'une loi discrète, c'est connaître les probabilités de ses éléments unitaires. La distribution d'une loi de Bernoulli est définie par

$$\begin{cases} P(X=0) = p = \text{probabilité de ne pas être malade} \\ P(X=1) = 1-p = \text{probabilité d'être malade} \end{cases}$$

Propriété 2.3 .

Si $X \hookrightarrow \mathcal{B}er(p)$, *alors :* $\mu(X) = p$

$\sigma^2(X) = p(1-p) = pq$

Remarques : $\sigma^2(X)$ est maximale pour $p = \frac{1}{2}$. La plus grande dispersion qui correspond à l'incertitude maximale sur les valeurs de X est atteinte pour $p = \frac{1}{2}$.

Loi Binomiale

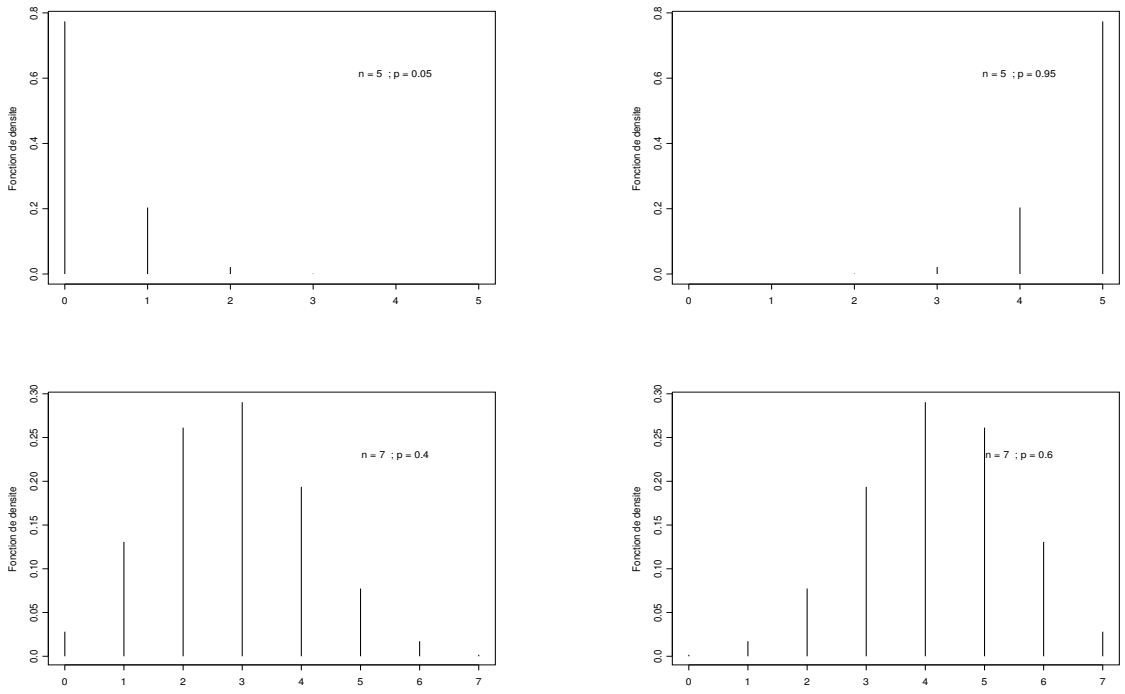


FIG. 2.1 – Densités de lois binomiales pour différentes valeurs des paramètres n et p.

On considère la variable aléatoire X="être malade" X suit une loi de Bernoulli $\mathcal{B}er(p)$. On considère maintenant un groupe de n personnes et on compte le nombre de personnes malades :

$$Y = \sum_{i=1}^{n} X_i(\omega_i) \text{ avec } X_i(\omega) = \begin{array}{ll} 1 & \text{si malade} \\ 0 & \text{sinon} \end{array}$$

Y est une variable aléatoire pouvant être définie ainsi :

$$\begin{array}{ccc} \Omega_1 \times \ldots \times \Omega_n & \longrightarrow & \{0,\ 1,\ \ldots n\} \\ \omega = (\omega_1,\ \ldots,\ \omega_n) & \longmapsto & Y(\omega) = \sum_{i=1}^{n} X_i(\omega_i) \end{array}$$

Définition 2.6 *Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes suivant une distribution $\mathcal{B}er(p)$. Alors $Y = \sum_{i=1}^{n} X_i$ suit une loi Binomiale de paramètres p et n. On note $Y \hookrightarrow \mathcal{B}(n,\ p)$ ou $\mathcal{L}(Y) = \mathcal{B}(n,\ p)$.*

Propriété 2.4 *Si Y suit une loi Binomiale de paramètres p et n, alors $\forall k \in \{0,\ \ldots,\ n\}$, $P(Y = k) = \mathcal{C}_n^k\ p^k\ (1-p)^{n-k}$ avec $\mathcal{C}_n^k = \dfrac{n!}{k!\ (n-k)!}$*

- Cas où $n = 2$ (*i.e.* On considère un groupe de deux personnes)
$$\begin{aligned} P(Y=0) &= P(X_1 = 0 \wedge X_2 = 0) = P(X_1 = 0) \times P(X_2 = 0) = q^2 \\ P(Y=2) &= P(X_1 = 1 \wedge X_2 = 1) = P(X_1 = 1) \times P(X_2 = 1) = p^2 \\ P(Y=1) &= P((X_1 = 1 \wedge X_2 = 0) \vee (X_1 = 0 \wedge X_2 = 1)) \\ &= P(X_1 = 1) \times P(X_2 = 0) + P(X_1 = 0) \times P(X_2 = 1) \\ &= pq + qp \\ &= 2pq \end{aligned}$$
- Cas général : Les résultats précédents se généralisent au cas où le groupe est composé de n personnes. Si l'échantillon comprend k malades (donc $n-k$ non malades), la probabilité correspondante est $p^k\ (1-p)^{n-k}$. Mais il faut tenir compte du fait que les k malades ne sont pas forcémént les k premiers sujets du groupe. Donc on va multiplier $p^k\ (1-p)^{n-k}$ par le nombre possible de groupes de n personnes composés de k malades et $(n-k)$ non malades. Ce nombre vaut : $\mathcal{C}_n^k = \dfrac{n!}{k!\ (n-k)!}$. Ainsi, $P(Y = k) = \mathcal{C}_n^k\ p^k\ (1-p)^{n-k}$.

Remarque :

a) Pour $\mathcal{B}(n = 2,\ p)$, on a bien $\mathcal{C}_2^1 = \dfrac{2!}{1!\ (2-1)!} = 2$

b) $\sum_{k=0}^{n} P(Y = k) = 1$ donc $\sum_{k=0}^{n} \dfrac{n!}{k!\ (n-k)!} p^k\ (1-p)^{n-k} = 1$

c) La loi binomiale dépend de deux paramètres p et n.

d) Les variables aléatoires X_i doivent être indépendantes et suivre la même loi de Bernoulli $\mathcal{B}er(p)$. Cela signifie que la proportion p doit rester la même *i.e.* que les tirages se font avec remise (cas "pile ou face") ou que la population peut être assimilée à l'infini de sorte que p n'est pas modifié quand on extrait un individu (cas "malade / non malade").

Propriété 2.5 .

a) Si $Y \hookrightarrow \mathcal{B}(n,\ p)$ alors $\mu(Y) = np$ et $\sigma^2(Y) = npq$

b) Si $X \hookrightarrow \mathcal{B}(n_1,\ p)$ et $Y \hookrightarrow \mathcal{B}(n_2,\ p)$ sont indépendantes, alors $X + Y = \mathcal{B}(n_1 + n_2,\ p)$

Loi de Poisson

Ce type de distribution intervient lorsqu'on effectue un comptage du nombre de réalisations d'un événement donné dans un intervalle de temps fixé. Nombre de particules emises par une source radioactive en une seconde, nombre d'avions se présentant sur une piste d'atterrissage en une heure. Nombre de proies (non digérées) contenues dans un estomac.

Définition 2.7 *Soit X une variable aléatoire à valeurs dans $\mathbb{N}$:*

$$\begin{array}{rcl} X : (\Omega, \mathcal{A}, P) & \longrightarrow & \mathbb{N} \\ \omega & \longmapsto & X(\omega) \end{array}$$

X suit une loi de Poisson de Paramètre λ, notée $\mathcal{P}(\lambda)$, si et seulement si

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

Propriété 2.6 *Si $X \hookrightarrow \mathcal{P}(\lambda)$ alors $\mu = \lambda$ et $\sigma^2 = \lambda$.*

Propriété 2.7 *Si X_1 et X_2 sont deux variables aléatoires indépendantes telles que $\mathcal{L}(X_1) = \mathcal{P}(\lambda_1)$ et $\mathcal{L}(X_2) = \mathcal{P}(\lambda_2)$, alors $\mathcal{L}(X_1 + X_2) = \mathcal{P}(\lambda_1 + \lambda_2)$*

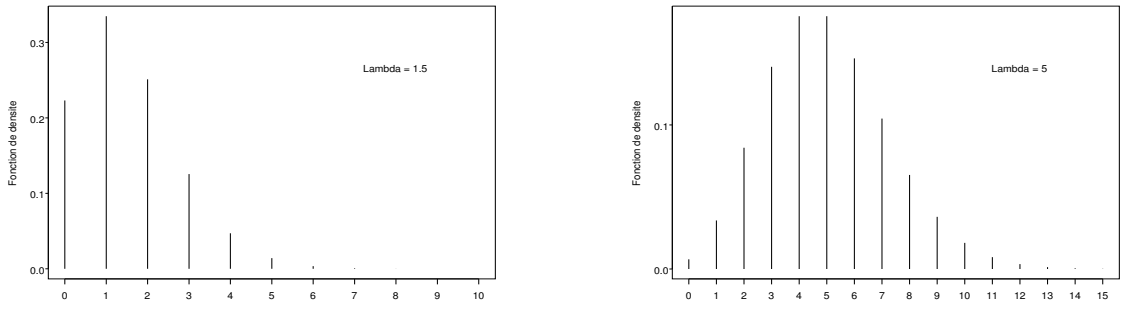


FIG. 2.2 – Densités de lois de Poisson pour différentes valeurs de λ.

Loi géométrique

Loi binomiale négative

Loi multinomiale

2.1.4 Approximation de la loi binomiale par la loi de Poisson

2.2 Lois de probabilité continues

2.2.1 Généralités

Définition 2.8 *Une variable aléatoire réelle est dite continue si elle est à valeurs dans un sous ensemble non dénombrable de $\mathbb{R}$.*

Exemple : On mesure la taille des poissons d'un lac :

$$\begin{array}{rcl} T : (\Omega, \mathcal{A}, P) & \longrightarrow & \mathbb{R}^+ \\ \omega & \longmapsto & T(\omega) = \text{ taille du poisson } \omega \end{array}$$

Propriété 2.8 *Tout événement peut être défini comme un intervalle ou une réunion d'intervalles de* $\mathbb{R}$.

Exemple : $A =$ "*avoir une taille* ≤ 12 *cm*" $= [12, +\infty)$; $B =$ "*avoir une taille* < 10 *cm ou* > 20 *cm*" $= [0, 10[\cup]20, +\infty)$

Conséquence : La loi de Probabilité d'une variabilité continue est connue dès qu'on connaît, pour tout intervalle]a,b] la probabilité pour que X soit comprise entre a et b :
$P_X]a, b] = P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a)$.

Remarque : $P_X\{a\} = P_X]a, a] = P(a < X \leq a) = P(X \leq a) - P(X \leq a) = F_X(a) - F_X(a) = 0$.
Donc, pour une variable aléatoire continue, la probabilité d'observer une valeur réelle donnée a est nulle, quelque soit $a \in \mathbb{R}$. Par exemple, lorsque l'on dit qu'un poisson mesure 23 cm, il sagit d'un abus de langage, la seule affirmation correcte est que l'observation se situe entre $23 - \varepsilon$ et $23 + \varepsilon$, avec $\varepsilon > 0$.

2.2.2 Densité de probabilité

Définition 2.9 *Soit X une variable aléatoire continue de fonction de répartion F_X. S'il existe une fonction numerique de la variable réelle f telle que :*

$$P(a < X < b) = F_X(b) - F_X(a) = \int_a^b f(t)dt,$$

On dit que f est la densité de probabilité de X et on la note f_X.

Conséquence : Pour une telle variable, la probabilité se présente comme une aire sous la courbe de f_X, l'aire totale étant égale à 1.

2.2.3 Moments d'une distribution continue

Définition 2.10 *Nous appelons moyenne théorique μ d'une variable aléatoire continue X, le nombre lorsqu'il existe :*

$$\mu = \mu(X) = \int_{-\infty}^{+\infty} t \times f_X(t)dt.$$

Remarque : C'est l'analogue du cas discret $\mu(X) = \sum_{i \geq 0} c_i \times p_i$.

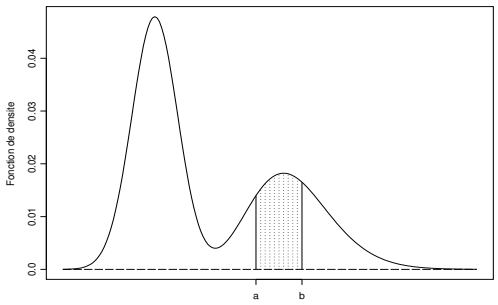


FIG. 2.3 – Densité de probabilité d'une variable aléatoire continue.

Définition 2.11 *Nous appelons variance théorique σ^2 d'une variable aléatoire continue X, le nombre lorsqu'il existe :*

$$\sigma^2 = \sigma^2(X) = \int_{-\infty}^{+\infty} (t - \mu(X))^2 \times f_X(t)dt$$

Remarque : C'est l'analogue du cas discret $\sigma^2(X) = \sum_{i \geq 0} (c_i - \mu(X))^2 \times p_i$.

2.2.4 Lois de probabilité de variables aléatoires continues réelles

Loi de Laplace-Gauss ou loi normale

Pour les variables aléatoires continues, c'est la loi la plus utilisée en statistique.

Définition 2.12 *Une variable aléatoire continue X suit une loi normale de paramètres μ et $\sigma^2 \geq 0$ si elle admet comme densité de probabilité :*

$$f_X(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu(X)}{\sigma}\right)^2}$$

On note $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$.

Remarque : La notation $\mathcal{N}(\mu, \sigma)$ existe également. Aussi, pour enlever toute ambiguïté, on précisera laquelle des deux notations est employée. Par exemple, $X \hookrightarrow \mathcal{N}(\mu, \sigma^2 = 9)$. ou $X \hookrightarrow \mathcal{N}(\mu, \sigma = 3)$.

Propriété 2.9 *Si $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$, alors : $\mu(X) = \mu$ et $\sigma^2(X) = \sigma^2$.*

Propriété 2.10 *Si $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$, alors $Y = aX + b$ suit une loi normale de moyenne $\mu(Y) = \mu(aX + b) = a\mu(X) + b$; et de variance $\sigma^2(Y) = \sigma^2(aX + b) = a^2\ \sigma^2(X)$.*

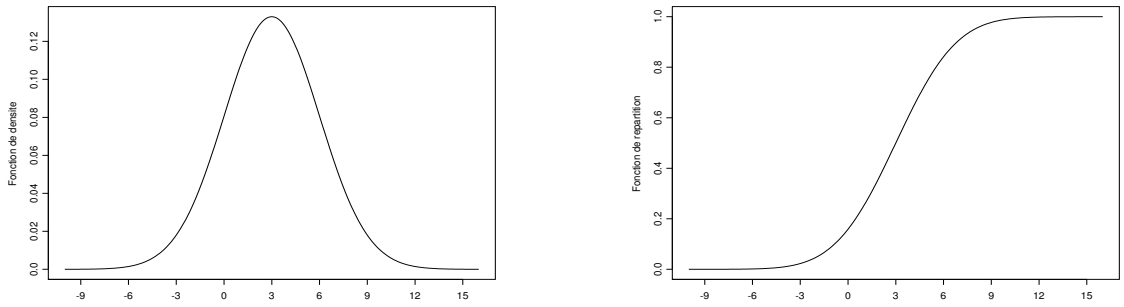


FIG. 2.4 – Fonctions de densité et de répartition pour une loi $\mathcal{N}(\mu = 3, \sigma^2 = 9)$.

Propriété 2.11 *Si* $X_1 \hookrightarrow \mathcal{N}(\mu_1, \sigma_1^2)$ *et* $X_2 \hookrightarrow \mathcal{N}(\mu_2, \sigma_2^2)$ *sont indépendantes, alors* $Y = X_1 + X_2$ *suit une loi normale :*

- *de moyenne* $\mu(Y) = \mu(X_1 + X_2) = \mu(X_1) + \mu(X_2)$ *;*
- *et de variance* $\sigma^2(Y) = \sigma^2(X_1 + X_2) = \sigma^2(X_1) + \sigma^2(X_2)$.

Remarques :

1. $f_X(x)$est est une fonction symétrique par rapport à la moyenne μ et admet deux points d'inflexion d'abcisse $\mu - \sigma$ et $\mu + \sigma$.
2. La loi normale dépend de deux paramètres μ et σ^2. Il existe donc une infinité de lois normales :

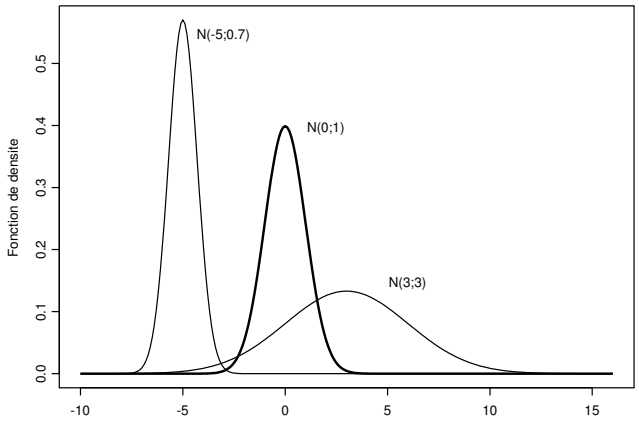


FIG. 2.5 – Fonction de densité pour trois lois normales.

Cependant, à partir d'une loi normale quelconque, $X \hookrightarrow \mathcal{N}(\mu, \sigma^2)$, on peut se ramener à $\mathcal{Z} \hookrightarrow \mathcal{N}(0, 1)$. En effet $\mathcal{Z} = \dfrac{X - \mu}{\sigma}$ suit alors d'après la propriété (2.11) une loi normale

de moyenne $\mu(\mathcal{Z}) = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$

3. Le calcul des probabilités associées à la loi normale n'est pratiquement pas possible avec des moyens simples. En effet, la primitive de :

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu(X)}{\sigma}\right)^2}$$

n'existe pas. On se sert alors de tables construite pour la loi $\mathcal{N}(0,1)$ dont nous verrons l'utilisation en T.D.

4. Lorsque l'on doit calculer une probabilité relative à une loi normale quelconque $\mathcal{N}(\mu, \sigma^2)$ on utilise le changement de variable indiqué en 2.
exemple :
$P_X([a,b]) = P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} < \mathcal{Z} < \frac{b-\mu}{\sigma}\right)$. Ainsi la table de la loi normale $\mathcal{N}(0,1)$ suffit à calculer les probabilités relatives à n'importe quelle loi normale.

5. **Vocabulaire :** La loi $\mathcal{N}(0,1)$ est dite loi normale standard ou encore loi normale centrée réduite.

6. Importance de la loi normale
   - La loi normale est certainement la loi de probabilité la plus utilisée. Comme on le verra par la suite, elle intervient souvent comme loi limite vers laquelle convergent certains modèles : par exemple la moyenne de n mesures indépendantes et de même loi a une distribution qui se rapproche de plus en plus d'une distribution normale lorsque n augmente.
   - Introduite comme "Loi normale des erreurs", d'où son nom, la loi normale semble décrire assez bien la distribution de certains caractères biométriques, par exemple la taille d'un individu choisi au hasard dans une population donnée.
   - Elle est à l'origine du développement de modèles probabilistes : les modèles gaussiens (Loi du Chi-Deux, loi de Student, loi de Fisher-Snedecor).
   - Enfin, elle peut être utilisée comme approximation des lois discrètes binomiale et de Poisson.

Lois déduites de la loi normale

On déduit de la loi normale standard trois types de variables aléatoires réelles continues qui qui sont d'un grand usage en statistique.

La loi du Chi-Deux (χ^2)

Définition 2.13 *Soient (X_i) une suite de variables aléatoires continues indépendantes normales $/\mathcal{L}(X_i) = \mathcal{N}(0,1)$ Alors $\mathcal{Z} = \sum_{i=1}^{n} X_i^2$ suit une loi de chi-deux à n degrés de liberté. On note $\mathcal{Z} \hookrightarrow \chi_n^2$*

Remarque : n est appelé degrés de liberté de la loi de χ^2.

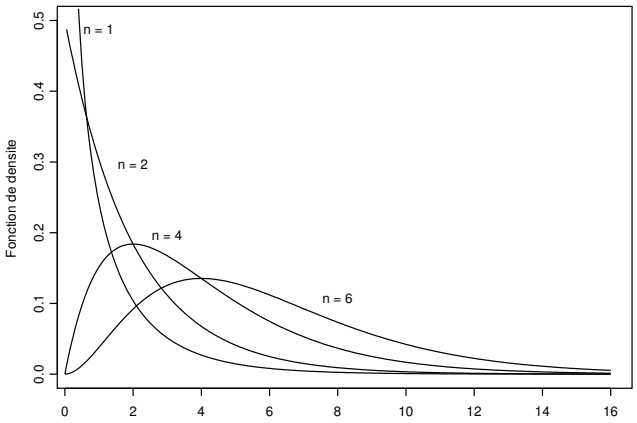


FIG. 2.6 – Fonction de densité de la loi du χ^2_n pour plusieurs valeurs de n.

Propriété 2.12 .

1) $\mu(\mathcal{Z}) = n$ *et* $\sigma^2(\mathcal{Z}) = 2n$

2) Considérons $\mathcal{Z}_1$ *et* $\mathcal{Z}_2$ *deux variables aléatoires continues indépendantes. Si* $\mathcal{Z}_1 \hookrightarrow \chi^2_{n_1}$ *et* $\mathcal{Z}_2 \hookrightarrow \chi^2_{n_2}$ *, alors* $\mathcal{Z}_1 + \mathcal{Z}_2 \hookrightarrow \chi^2_{n_1+n_2}$

La loi de Student t_n

Définition 2.14 *Soient* $X \hookrightarrow \mathcal{N}(0,1)$ *et* $Y \hookrightarrow \chi^2_n$*indépendantes. Alors* $T = \frac{X}{\sqrt{\frac{Y}{n}}}$ *on note* $\mathcal{L}(T) = t_n$.

Propriété 2.13 *Si* $\mathcal{L}(T) = t_n$*, alors* $\mu(T)$ *existe si et seulement si* $n > 1$ *et* $\sigma^2(T)$ *existe si est seulement si* $n > 2$*, on a alors* $\mu(T) = 0$ *et* $\sigma^2(T) = \dfrac{n}{n-2}$.

La loi de Fisher-Snedecor $\mathcal{F}_{k_1,k_2}$

Définition 2.15 *Soient* X *et* Y *deux variables aléatoires continuessuivant* $\chi^2_{k_1}$ *et* $\chi^2_{k_2}$ *indépendantes. Alors* $\mathcal{F} = \frac{X/k_1}{Y/k_2}$ *suit une loi de Fisher-Snedecor à* k_1 *et* k_2 *de degrés de liberté. On note* $\mathcal{F} \hookrightarrow \mathcal{F}_{k_1,k_2}$

Propriété 2.14 *Si* $\mathcal{L}(\mathcal{F}) = \mathcal{F}_{k_1;k_2}$ *alors* $\mu(\mathcal{F})$ *n'existe que si* $n_2 > 2$ *et* $\sigma^2(\mathcal{F})$ *n'existe que si* $n_2 > 4$.

alors on a $\mu(\mathcal{F}) = \dfrac{n_2}{n_2 - 2}$ *et* $\sigma^2(\mathcal{F}) = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)}$

Propriété 2.15 *Si* $\mathcal{L}(T) = t_n$ *alors* $\mathcal{L}(T^2) = \mathcal{F}_{1,n}$

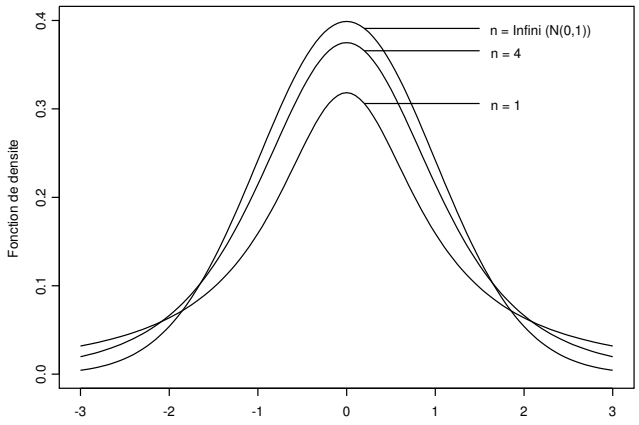


FIG. 2.7 – Fonction de densité de la loi de Student t_n pour plusieurs valeurs de n.

Remarque : De la propriété précédente, on en déduit que si $\mathcal{L}(\mathcal{F}) = \mathcal{F}_{1,n}$ et $\mathcal{L}(T) = t_n$ alors $\forall \alpha \in [0,1]$

$$P(\mathcal{F} \leq \mathcal{F}_{1-\alpha}) = 1 - \alpha \Leftrightarrow P(-\sqrt{\mathcal{F}_{1-\alpha}} \leq T \leq \sqrt{\mathcal{F}_{1-\alpha}}) = 1 - \alpha$$

Approximations par la loi normale

Loi binomiale par loi normale

Loi de Poisson par loi normale

2.3 Opérations sur les variables aléatoires

Propriété 2.16 *Soient X et Y deux variables aléatoires continues, a et b deux réels fixés alors :*

(i) $\mu(aX + b) = a\mu(X) + b$
(ii) $\mu(X + Y) = \mu(X) + \mu(Y)$
(iii) $\sigma^2(aX + b) = a^2\sigma^2(X)$
(iv) Si X et Y sont indépendantes : $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$

C'est grâce à cette propriété qu'on a obtenu celles concernant la somme de plusieurs bernoulli, de binomiales et lois normales.

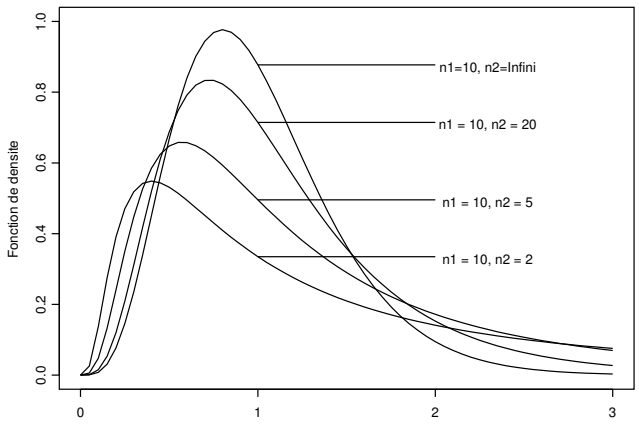


FIG. 2.8 – Fonction de densité de la loi de Fisher-Snedecor $\mathcal{F}_{n_1,n_2}$ pour plusieurs valeurs des paramètres.

Deuxième partie

Introduction à la statistique

Chapitre 5

Statistique descriptive à une dimension

On se place désormais dans le cas le plus courant où on ne peut pas observer toute la population (ex : ensemble des foyers de la population française, poissons d'un lac,...). Les observations porteront sur une partie de la population, le but étant d'inférer à partir de ces observations des résultats valables au niveau de la population (ex : quelle est la part des foyers monoparentaux dans la population française ? Peut on dire que la taille moyenne des poissons du lac est supérieur à 20 cm,...). Ce sera l'objet de chapitres ultérieurs. Avant ces études, nous allons voir maintenant un certain nombre d'outils (indices, graphiques) permettant de décrire n observations qui ont été faites d'une variable aléatoire.

5.1 Définitions

Définition 5.1 *On appelle* **échantillon** *un sous-ensemble de la population* Ω*. Un échantillon est dit représentatif si chaque élément de la population a une probabilité connue et non nulle d'en faire partie.*

Exemple : Pour Ω="Ensemble des étudiants de maîtrise de biologie cellulaire à St Charles", un échantillon peut être obtenu en prenant les 30 premiers étudiants par ordre alphabétique. Cet échantillon est non représentatif car tous les autres étudiants ont alors une probabilité nulle d'en faire partie. On se propose de tirer au sort 30 étudiants (sur les 90). Cet échantillon est représentatif car tous les étudiants ont une probabilité non nulle (égale à $\frac{1}{90}$) d'en faire partie.

On considère les relevés ou observations d'une variable aléatoire réelle sur un échantillon de n individus extraits d'une population de référence.

Définition 5.2 *On appelle* **n-échantillon** *l'ensemble des n valeurs numériques* $(x_1, \ldots, x_n)$ *observées sur un échantillon de n individus.*

Exemple 5.1 On a compté le nombre de loges capsulaires sur 1905 coquelicots.
Quelle est la population ? Réponse : L'ensemble des individus constituant l'espèce "coquelicots".
Quel est l'échantillon ? Réponse : c'est l'ensemble des 1905 coquelicots.

Quel est le n-échantillon ? Réponse : c'est l'ensemble des 1905 valeurs observées de la variable "nombre de loges capsulaires" : $\{16, 3, 17, 2, ..., 13, 9, 11\}$.

Comment est définie la variable aléatoire réelle "nombre de loges capsulaires" ?

$$\begin{array}{rccl} L: & (\Omega, \mathcal{A}, P) & \longrightarrow & \mathbb{N} \\ & \omega & \longmapsto & L(\omega) = \text{"nombre de loges capsulaires"}. \end{array}$$

Exemple 5.2 On a pesé 100 poissons provenant d'un lac.
Quelle est la population ? Réponse : L'ensemble des poissons du lac.
Quel est l'échantillon ? Réponse : c'est l'ensemble des 100 poissons.
Quel est le n-échantillon ? Réponse : c'est l'ensemble des 100 valeurs observées de la variable "poids du poisson" $\{180, 200, 150, 180, ..., 320, 195\}$.

$$\begin{array}{rccl} Pds: & (\Omega, \mathcal{A}, P) & \longrightarrow & \mathbb{R}^+ \\ & \omega & \longmapsto & Pds(\omega) = \text{"Poids du poisson"} \end{array}$$

Définition 5.3 *On distingue quatre types de variables selon leurs valeurs possibles :*

- *Les* **variables qualitatives nominales** *sont celles dont les valeurs sont des attributs sans ordre naturel (ex. sexe, csp).*
- *Les* **variables qualitatives ordinales** *sont celles dont les valeurs sont des attributs avec un ordre naturel (ex. activité faible, moyenne, forte).*
- *Les* **variables quantitatives discrètes** *sont celles dont les valeurs sont le résultat d'un dénombrement (ex. nombre d'enfants dans une famille, pouls, nombre d'oiseaux nichant dans une falaise).*
- *Les* **variables quantitatives continues** *sont celles dont les valeurs sont des mesures (ex. taille, poids).*

Remarque :

1. Les variables aléatoires qualitatives nominales, ordinales et quantitatives discrètes sont des variables discrètes. Les variables quantitatives continues sont des variables continues.
2. Le type d'une variable dépend de sa nature, mais aussi de la manière dont elle a été mesurée et/ou codée et/ou du choix de la personne qui mène l'étude.

   Exemples :

   L'âge est par nature une variable aléatoire quantitative continue. Mais on l'exprime généralement en années (valeurs discrètes) donc à ce titre elle pourra être considérée comme une variable quantitative discrète.

   Un nombre d'œufs d'une espèce de poisson est par nature une variable aléatoire quantitative discrète. Mais en général, on ne compte pas les œufs, on estime leur nombre en divisant par exemple le poids total par le poids moyen d'un œuf. Elle pourra alors être considérée comme une variable quantitative continue.

5.2 Distributions statistiques empiriques

Définition 5.4 *On appelle distribution statistique empirique d'un n-échantillon la donnée d'un regroupement* $(c_1, n_1), (c_2, n_2) \ldots, (c_k, n_k)$ *construit de la manière suivante :*

1) Pour les variables discrètes, c_j est la jième valeur possible de la variable et n_j le nombre de fois où cette valeur a été observée (c'est l'effectif). Les valeurs c_j sont également appelées modalités de la variable aléatoire.
2) Pour les variables continues, on construit une partition de l'ensemble des valeurs en k intervalles : $c_1 =]a_0, a_1], c_2 =]a_1, a_2], c_k =]a_{k-1}, a_k]$, et on compte le nombre de valeurs observées dans chaque intervalle. Les valeurs c_j sont appelées classes.

Définition 5.5 *Pour une distribution statistique empirique donnée, on appelle fréquence empirique de la modalité ou de la classe c_j le nombre $f_j = \frac{n_j}{n}$ où $n = \sum_{j=1}^{k} n_j$*

Exemples :

1. Pour l'exemple 5.1 la distribution statistique empirique est :

Nombre de loges (C_j)	6	7	8	9	10	11	12	13	14	15
Nombre de coquelicots (n_j)	3	11	38	106	152	238	305	315	302	234
Fréquence (f_j)	0,02	0,06	0,2	0,056	0,08	0,125	0,16	0,165	0,159	0,123

Nombre de loges (C_j)	16	17	18	19	20
Nombre de coquelicots (n_j)	128	50	19	3	1
Fréquence (f_j)	0,067	0,026	0,01	0,02	0,01

2. Pour l'exemple 5.2 la distribution statistique empirique est :

Classe (C_j)	]180, 200]	]200, 220]	]220, 240]	]240, 250]	]250, 260]	]260, 280]	]280, 300]	]300, 320]
Nbre poissons (n_j)	6	8	15	23	16	16	10	6
Fréquence (f_j)	0,06	0,08	0,15	0,23	0,16	0,16	0,1	0,06

Remarques :

1. Dans le cas d'une variable continue, la distribution statistique empirique dépend du choix des classes. Il n'existe pas de choix optimal du nombre de classes k. Il est conseillé de manière à obtenir une distribution régulière qui représente au mieux la distribution des données, de choisir des classes de même amplitude et d'utiliser pour leur nombre la règle de Sturges $k = 1 + \frac{10ln(n)}{3ln(10)} = 7,6$ (dans notre exemple)
2. On parle de ditribution statistique empirique ou de fréquences empiriques parce qu'elles sont liées aux n-échantillon considéré. Si on considérait un autre échantillonage on n'observerai pas la même distribution empirique, la différence étant due aux fluctuations de l'échantillonage.

3. Lien entre distribution empirique et distribution théorique (ou loi de probabilité).
   Sous certaines conditions d'échantillonage, on montre que (loi faible des grands nombres) :

$$\lim_{n\to+\infty} f_j = p_j \quad \text{(cas discret)}$$

$$\lim_{n\to+\infty} f_j = P_X]a_{j-1},\ a_j] \quad \text{(cas continu)}$$

5.3 Paramètres de position d'un n-échantillon

Ils sont utilisés pour résumer la position centrale d'une distribution à l'aide d'un indice.

Définition 5.6 *La moyenne empirique d'un n-échantillon est le paramètre de position noté* $\bar{x}$ *et défini par* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *pour un échantillon, et par* $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j \times c_j = \sum_{j=1}^{k} \frac{n_j}{n} \times c_j$

Pour l'exemple 5.1 nous avons :

$$\bar{x} = \frac{16+3+17+\ldots+2+15}{1905} = \frac{3\times 6+11\times 7\ldots+1\times 20}{1905} = 12,756.$$

Définition 5.7 *La médiane empirique me(x)d'un n-échantillon est l'indice de position dont la valeur est telle que* 50% *des observations lui sont supérieures.Si l'on classe les observations par ordre croissant des* x_i *alors :*

- *Si n est impair* $me(x) = x_{(\frac{n+1}{2})}$
- *Si n est impair* $me(x) = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$

Pour notre exemple 1, n est impair, la médiane est la $\frac{1905+1}{2} = 953^{\text{ème}}$ observation lorsqu'elles sont classées par ordre croissant, ici $me(x) = 13$.

Définition 5.8 *Le mode ou classe modale d'une distribution statistique est le paramètre de position* $mo(x)$ *dont la valeur est celle de la modalité ou la classe la plus fréquente.*

Exemple 5.1 $mo(x) = 13$.

Remarques :

1. Moyennes et médianes ne s'appliquent qu'aux variables quantitatives ou ordonnées (médiane). Pour une variable qualitative il faut attribuer une valeur numérique (un codage) à chacune de ses modalités. La valeur obtenue pour la moyenne ou la médiane dépend alors du codage choisi. Lorsque ce choix est arbitraire (ex : catégories socio-professionnelles, code postal), la valeur de moyenne ou de médiane n'a aucune interprétation.
2. Si on considère plusieurs échantillons d'une même variable aléatoire, on peut pour chacun d'eux calculer leur moyenne $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots$ Ces valeurs n'ont aucune raison d'être égales, à cause des fluctuations de l'échantillonnage. Ainsi $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots$ peuvent être considérées comme des réalisations d'une variable aléatoire moyenne définie ainsi :

$$\begin{array}{rcl}(\Omega \times \ldots \times \Omega,\ \mathcal{A},\ P) & \longrightarrow & \mathbb{R} \\ \omega = (\omega_1, \ldots, \omega_n) & \longmapsto & \bar{X}(\omega) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X(\omega_i)\end{array}$$

3. Sous certaines conditions, il est prouvé que :

$$\lim_{n \to +\infty} \bar{x} = \mu \ \text{ (loi forte des grands nombres)}$$

On peut interpréter ce résultat en disant que lorsque l'on prend la population tout entière ($n \to +\infty$), on retrouve bien l'expression de la moyenne théorique.

5.4 Paramètres de dispersion d'un n-échantillon

Ils ont utilisés pour résumer la dispersion des observations d'une variable aléatoire.

Définition 5.9 *la variance empirique d'un n-échantillon est le paramètre de dispersion, noté* $s^2(x)$ *qui vaut :* $s^2(x) = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2 = \dfrac{1}{n-1}\sum_{j=1}^{k} n_j \times (c_j - \bar{x})^2 = \sum_{j=1}^{k} \dfrac{n_j}{n-1} \times (c_j - \bar{x})^2$.
On appelle écart-type empirique $s(x)$.

Pour l'exemple 5.1, $s^2(x) = \frac{1}{1904}[3(6 - 12.76)^2 + \ldots] = 2.23^2 \simeq 5$

Remarque :

1. L'unité dans laquelle est exprimée la variance est celle de la variable au carré, c'est pourquoi on lui préfère l'écart-type qui a la même unité que la variable concernée.
2. Sous certaines conditions, il est prouvé que

$$\lim_{n \to +\infty} s^2(x) = \sigma^2(X) \ \text{ (loi forte des grands nombres)}$$

Il existe d'autres indices de dispersion :

Définition 5.10 *On appelle étendue* $e(x) = max(x_i) - min(x_i)$

Définition 5.11 *On appelle coefficient de variation l'indice* $cv(x) = \dfrac{s(x)}{\bar{x}} \times 100$

Remarque : Le coefficient de variation permet d'exprimer la variabilité en terme relatif alors que la variance (ou l'écart-type) l'exprime en terme absolue. Il ne dépend pas de l'unité choisie et permet donc de comparer les variabilités de deux variables qui ne sont pas exprimées dans la même unité.

Exemple : On a mesuré la taille et le poids de 10 poissons. Les deux n-échantillons sont :
$x = \{23, 20, 17, 15, 30, 25, 24, 27, 22, 19\}$ pour la taille en centimètres.
$y = \{250, 220, 150, 180, 350, 250, 200, 240, 200, 100\}$ pour le poids exprimé en gramme.
On a $\bar{x} = 22.2$ cm, $s(x) = 4.59$ cm, $cv(x) = 20.6\%$
$\bar{y} = 214$ cm, $s(x) = 67.03$ cm, $cv(x) = 31.2\%$.

5.5 Représentations graphiques

On considère la distribution empirique d'un n-échantillon : $\{(c_1, n_1), (c_2, n_2) \dots (c_k, n_k)\}$, et sa distribution des fréquences empiriques $\{\frac{n_1}{n}, \frac{n_2}{n} \dots \frac{n_k}{n}\}$

Définition 5.12 *Cas discret*
On appelle graphique des fréquences le graphe où on reporte sur l'axe des abscisses les modalités c_j et où on trace au dessus de chacune d'elles un segment vertical de hauteur proportionelle à la fréquence empirique.

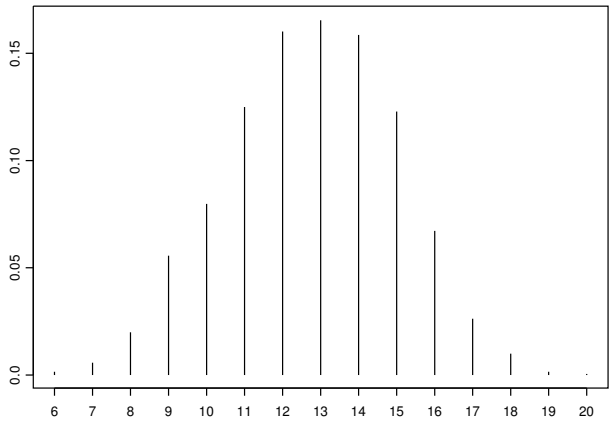


FIG. 5.1 – Diagramme des fréquences de la variable aléatoire "Nombre de loges capsulaires"

Définition 5.13 *Cas continu $c_j =]a_{j-1}, a_j]$*
On appelle histogramme des fréquences le graphe où on reporte sur l'axe des abscisse les classes et au dessus de chacune d'elles un rectangle dont l'aire est égale ou proportionnelle à la fréquence associée.

Remarque 1 : l'unité de mesure d'un histogramme est une aire. Lorsque toutes les classes ont même largeur ; alors seulement l'aire est proportionnelle à la hauteur et il est possible de définir une unité sur l'axe des ordonnées.

Remarque 2 : Lien entre fréquence empirique et fréquence théorique (ou probabilité). Pour un n-échantillon, on mesure la probabilité de l'intervalle $[a, b]$ par la surface du rectangle : $P_X[a, b] \simeq h \times \Delta x$.

Si on peut observer un plus grand nombre d'individus, on peut exprimer cette probabilité sous la forme : $P_X[a, b] \simeq h_1\Delta x_1 + h_2\Delta x_2 + \ldots = \sum_i h_i \Delta x_i$ où $h_i\Delta x_i = \Delta F = F_{i+1} - F_i$

Quand on fait tendre Δx vers 0, l'équation $\Delta F_i = h_i \Delta x_i$ devient $dF = hdx$ et $h = \frac{dF}{dx} = f(x)$ est la densité de probabilité de X.

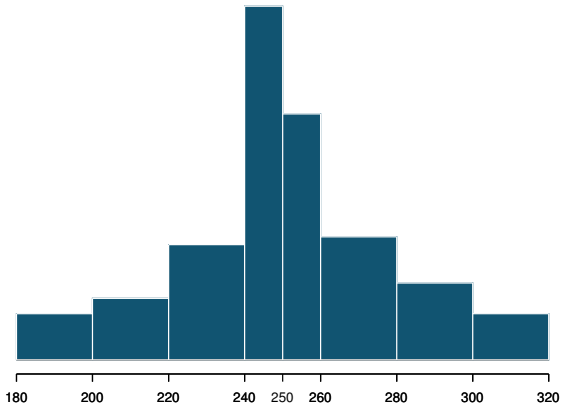


FIG. 5.2 – Histogramme de la variable aléatoire “poids des poissons”

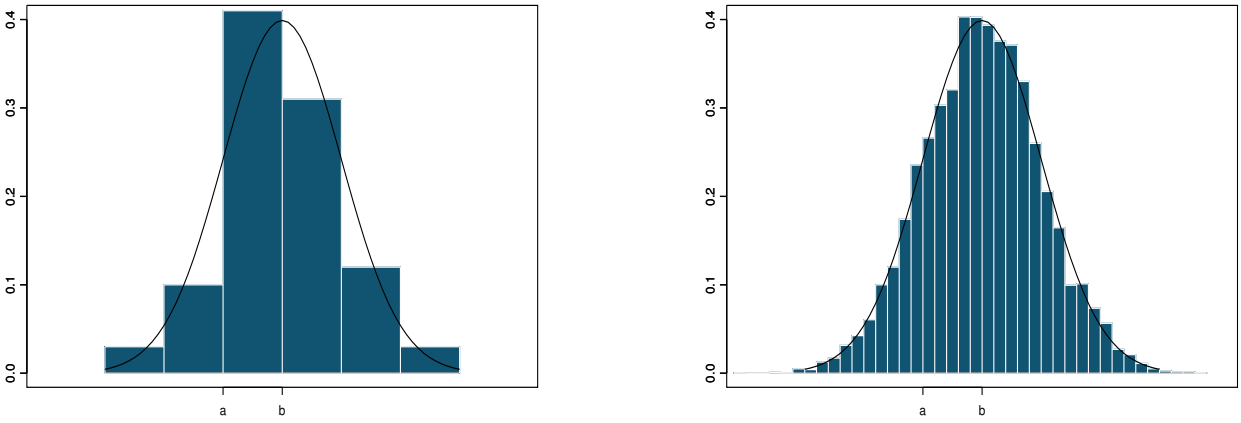


FIG. 5.3 – “Convergence” d’un histogramme vers une fonction de densité

Chapitre 6

Fluctuations d'échantillonnage

taratta Lorsqu'on veut évaluer pour un ensemble d'individus (*i.e.* une population) la moyenne μ d'une variable aléatoire quantitative (taille, poids, nombre de loges capsulaire ...) ou la proportion p d'individus de la population présentant un caractère donnée dans le cas d'une variable qualitative (proportion de malades, proportion de billes bleues ...), on ne dispose que d'un échantillon de la population. Les valeurs de moyenne $\overline{x}$ ou de proportion p_{obs} que l'on peut calculer à partir d'un échantillon ne sont pas égales à la moyenne μ exacte ou à la proportion p exacte de la population. Elles fluctuent autour des valeurs exactes et ce sont leurs fluctuations que l'on se propose d'étudier ici.

Autrement dit, quelles sont les lois de probabilité de :

$$\begin{array}{rccl} \overline{X}: & (\Omega \times \ldots \times \Omega, \ \mathcal{A}, \ \mathcal{P}) & \longrightarrow & \mathbb{R} \\ & \omega = (\omega_1, \ldots, \omega_n) & \longmapsto & \overline{X}(\omega) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X(\omega_i) = \overline{x} \end{array}$$

$$\begin{array}{rccl} \overline{P}: & (\Omega \times \ldots \times \Omega, \ \mathcal{A}, \ \mathcal{P}) & \longrightarrow & [0,1] \\ & \omega = (\omega_1, \ldots, \omega_n) & \longmapsto & \overline{P}(\omega) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} P(\omega_i) = \overline{p} \end{array}$$

Les résultats présentés dans ce chapitre découlent d'un théorème appelé théorème central limite :

Théorème 6.1 *Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires réelles indépendantes de même loi possédant une moyenne μ et une variance σ^2. Alors quelle que soit la loi des X_i,*

$$\lim_{n\to+\infty} \mathcal{L}\Big(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}\Big) = \mathcal{N}(0,1). \tag{6.1}$$

Remarque 6.1 L'égalité (6.1) signifie qu'en tout point t de l'intervalle de variation où F est continue, $\lim_{n\to+\infty} F_n(t) = F(t)$, où on note F_n la fonction de répartion de la variable aléatoire $\dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ et F la fonction de répartition de la loi $\mathcal{N}(0,1)$.

6.1 Fluctuation d'échantillonnage d'une moyenne

6.1.1 Distribution d'une moyenne

Propriété 6.1 *Soit $X_1, \dots X_n$, n variables aléatoires indépendantes de même distribution et possédant une moyenne μ et une variance σ^2. Alors la variable aléatoire $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ suit asymptotiquement une distribution normale de moyenne $\mu' = \mu$ et de variance $\sigma'^2 = \frac{\sigma^2}{n}$*

Remarque 6.2 .

1. Asymptotiquement signifie que si n est assez "grand", alors la loi de $\overline{X}$ peut être approximée par une normale $\mathcal{N}(\mu, \sigma/\sqrt{n})$.
2. Cette approximation est d'autant meilleure que
   (a) La distibution de X est proche d'une loi normale (symétrique, ...)
   (b) n est "grand" : nous admettrons dans les problèmes à traiter que la moyenne suit une loi normale dès que n atteint 30, lorsqu'il s'agit d'une variable quantitative continue.
3. La démonstration de la propriété précédente est basée sur le théorème central limite. Si n est assez "grand", alors à n fixé :

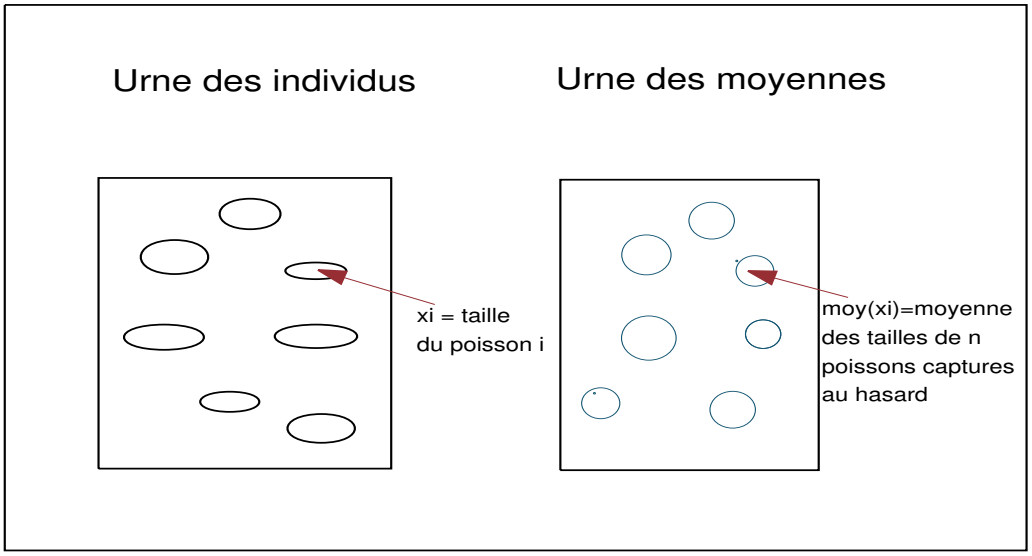
$$\begin{array}{rcl} \mathcal{L}\Big(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}\Big) & \simeq & \mathcal{N}(0,1) \\ \iff \quad \mathcal{L}\Big(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\Big) & \simeq & \mathcal{N}(0,1) \\ \iff \quad \mathcal{L}(\overline{X}-\mu) & \simeq & \mathcal{N}(0, \sigma/\sqrt{n}) \\ \iff \quad \mathcal{L}(\overline{X}) & \simeq & \mathcal{N}(\mu, \sigma/\sqrt{n}) \end{array}$$

6.1.2 Interprétation

On considère la variable aléatoire taille pour les poissons d'un lac. On peut symboliquement représenter la population sous la forme d'une urne composée de boules, chaque boule représentant un poisson. Sur chaque boule est inscrite la taille du poisson.

La moyenne de l'ensemble des boules, c'est à dire la moyenne de l'ensemble des tailles des poissons est la moyenne théorique $\mu(T)$. La variance de l'ensemble des tailles des poissons est la variance théorique $\sigma^2(T)$. Cette urne correspond à la variable T, elle est appelée urne des individus.

On peut associer à l'urne des individus une seconde urne dite urne des moyennes où chaque boule représente un échantillon de n poissons. Sur chaque boule est inscrite la moyenne $\overline{X}$ de cet échantillon. Cette seconde urne correspond à la variable aléatoire $\overline{X}$, sa moyenne théorique est $\mu'(\overline{X}) = \mu(T)$, et sa variance $\sigma'^2(\overline{X}) = \frac{\sigma^2(T)}{n}$.



$$X : \quad (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}, P_X) \qquad\qquad \overline{X} : \quad \Omega \times \ldots \times \Omega = \Omega^{\otimes n} \longrightarrow (\mathbb{R}, P_{\overline{X}})$$
$$\omega \longmapsto T(\omega) \qquad\qquad \omega = (\omega_1, \ldots, \omega_n) \longmapsto \overline{X}(\omega) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Moyenne théorique : μ

Variance théorique : σ^2

Distribution : **?**

Moyenne théorique : $\mu' = \mu$

Variance théorique : $\sigma'^2 = \dfrac{\sigma^2}{n}$

Distribution : approximativement normale dès que n est grand ($n > 30$)

6.1.3 Intervalle de fluctuation d'une moyenne

On considère un lac dont les poids des poissons (urne des individus) ont une moyenne $\mu = 100\ g$ et une variance $\sigma^2 = 2500\ g^2$ ($\sigma = 50\ g$). On en déduit que la variable aléatoire "poids moyen de n=100 poissons" est distribuée asymptotiquement selon une distribution :

$$\mathcal{N}(\mu' = 100, \sigma'^2 = \frac{2500}{100} = 25),$$

d'aprés la propriété 6.1 et parce que la condition $n \geq 30$ est respectée.

On peut alors décrire les fluctuations de $\overline{X}$. Par exemple quelle est la probabilité que la moyenne de 100 poissons pris au hasard appartienne à l'intervalle [90,110] ?

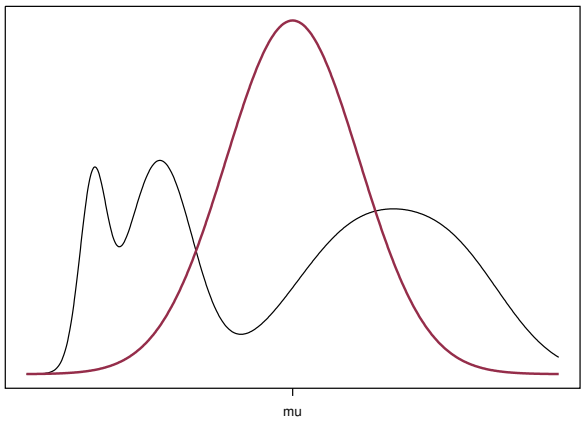


FIG. 6.1 – Distibution de l'urne des individus, et de l'urne des moyennes (en gras)

$P_X([90,110]) = P(90 < \overline{X} < 110) = P\Big(\frac{90-100}{5} < \frac{\overline{X}-100}{5} < \frac{110-100}{5}\Big) = P(-2 < \mathcal{Z} < 2) \simeq 0.95$.

Formule générale

On peut ansi, si on se fixe un risque α, construire un intervalle de fluctuation centré sur la moyenne(ou intervalle de pari) auquel la moyenne $\overline{X}$ d'un échantillon a une probabilité $1-\alpha$ d'appartenir.

$$\begin{aligned} & P(\mu - a < \overline{X} < \mu + a) &=& \ 1-\alpha \\ \iff \quad & P\Big(\frac{\mu - a - \mu}{\sigma/\sqrt{n}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{\mu + a - \mu}{\sigma/\sqrt{n}}\Big) &=& \ 1-\alpha \\ \iff \quad & P\Big(-z_\alpha < \mathcal{Z} < z_\alpha\Big) &=& \ 1-\alpha \\ \iff \quad & P\Big(|\mathcal{Z}| > z_\alpha\Big) &=& \ 1-\alpha \end{aligned}$$

On en déduit que $z_\alpha = \dfrac{a}{\sigma/\sqrt{n}} \iff a = z_\alpha \times \dfrac{\sigma}{\sqrt{n}}$

d'où l'intervalle : $[\mu - z_\alpha \times \dfrac{\sigma}{\sqrt{n}}, \mu + z_\alpha \times \dfrac{\sigma}{\sqrt{n}}]$

Remarques :

1. L'intervalle de fluctuation précédent peut être considéré comme un intervalle de pari où ce dernier consisterait à parier que la moyenne d'un échantillon appartienne à l'intervalle. Le risque associé à ce pari est que la moyenne n'appartienne pas à l'intervalle, il est égal à α.
2. L'intervalle de fluctuation d'une moyenne sera d'autant plus petit que σ est faible (c.a.d que la dispersion des valeurs individuelles est faible), que n est grand (c.a.d que la taille de l'échantillon est importante), et que α est important.

6.2 Fluctuations d'échantillonnage d'une proportion

6.2.1 Distribution d'une proportion

Propriété 6.2 *Soient $X_1 \dots X_n$ variables aléatoires indépendantes suivant chacune une Bernoulli de paramètre p $\mathcal{B}er(p)$ Alors la variable aléatoire "proportion du caractère pour un n-échantillon" :*

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i$$

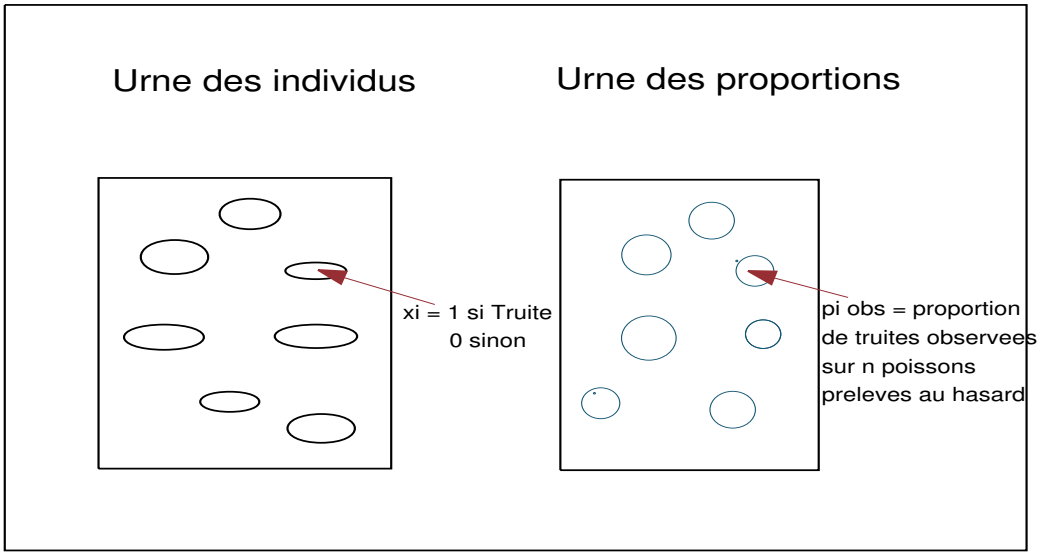
est asymptotiquement distribuée suivant une distribution normale $\mathcal{N}(\mu' = p, \sigma'^2 = \frac{pq}{n})$

Remarque :

1. Il suffit d'appliquer la propriété 6.1 à la variable aléatoire $\frac{1}{n}\sum_{i=1}^{n} X_i$ en remarquant que $\mu(X) = p$ et que $\sigma^2(X) = pq$.
2. Il semblerait logique d'appeler "grands échantillons" ce dont l'effectif dépasse une certaine valeur n, il se trouve qu'une définition ne peut convenir ici : pour une urne contenant 50% (p=0.5) de boules noires l'approximation est correcte dès que $n \geq 10$, tandis que la même approximation ne sera obtenue pour une urne ne contenant que 1% ($p = 0.01$) de boules noires l'approximation est correcte dès que $n \geq 500$. Ainsi, la définition de "grand échantillon" dépend aussi de p. On considèrera l'aproximation valable dès que np et nq ≥ 5.
3. np représente l'effectif des boules noires (c.a.d la modalité 1 des X_i) si la composition de l'échantillon est identique à celle de la population.

6.2.2 Interprétation de la propriété

On considère la variable aléatoire "espèce" pour les poissons d'un lac peuplé de truites et d'autres espèces. On peut symboliquement représenter la population sous la forme d'une urne composées de boules, chaque boule représentant un poisson. Sur chaque boule est inscrit 1 s'il s'agit d'une truite et 0 s'il s'agit d'une autre espèce. La moyenne de l'ensemble des boules est la moyenne théorique de la Bernoulli ($\mu = p$). La variance de l'ensemble des boules est la variance théorique de Bernoulli $\sigma^2 = pq$. On appellera cette urne : urne des individus. On peut associer à l'urne des individus une seconde urne dite urne des proportions où chaque boules va représenter un échantillon de n poissons et sur laquelle sera écrite la proportion de truite de l'échantillon. Cette seconde urne correspond à la variable aléatoire T, sa moyenne théorique est $\mu' = p$ et sa variance $\sigma'^2 = \frac{pq}{n}$.



$$X : \begin{array}{ccc} (\Omega, \mathcal{A}, P) & \longrightarrow & (\mathbb{R}, P_X) \\ \omega & \longmapsto & X(\omega) \end{array}$$

Moyenne théorique : $\mu = p$
Variance théorique : $\sigma^2 = pq$
Distribution : Ber

$$T : \begin{array}{ccc} \Omega \times \ldots \times \Omega = \Omega^{\otimes n} & \longrightarrow & (\{0,1\}, P_T) \\ \omega = (\omega_1, \ldots, \omega_n) & \longmapsto & T(\omega) = \frac{1}{n} \sum_{i=1}^{n} X_i \end{array}$$

Moyenne théorique : $\mu' = p$
Variance théorique : $\sigma'^2 = \dfrac{pq}{n}$
Distribution : approximativement normale dès que np et nq sont grands ($np > 5$ et $nq > 5$)

6.2.3 Intervalle de fluctuation d'un pourcentage

On considère un lac (urne des individus) peuplé à 30% de truites ($p = 0.3$). Quelle est la probabilité pour que la proportion de truites observée sur un échantillon de 50 poissons soit comprise entre 0.2 et 0.4 ?

L'urne des proportions est distribuée approximativement suivant une loi normale de moyenne $\mu' = p = 0.3$ et de variance $\sigma'^2 = pq/n = 0.0042$ soit $\sigma' = 0.065$.

$$
\begin{aligned}
P(0.2 < T < 0.4) &= P\Big(\frac{0.2-0.3}{0.065} < Z = \frac{T-0.3}{0.065} < \frac{0.4-0.3}{0.065}\Big) \\
\iff P(0.2 < T < 0.4) &= P\Big(\frac{-0.1}{0.065} < Z < \frac{0.1}{0.065}\Big) \\
\iff P(0.2 < T < 0.4) &= P(-1.543 < Z < 1.543) \\
\iff P(0.2 < T < 0.4) &= 2 \times P(Z < 1.543) - 1 \\
\iff P(0.2 < T < 0.4) &= 2 \times 0.94 - 1 = 0.88
\end{aligned}
$$

Donc le “risque" que la proportion observée de truites sur un échantillon de 50 individus, soit hors de l'intervalle $[0.2\,;\,0.4]$ est de 12%.

Formule générale

On peut ainsi, si on se fixe un risque α, construire un intervalle de fluctuation (ou intervalle de pari) auquelle la proportion observée d'un échantillon de n individus a une probabilité $1-\alpha$ d'y appartenir :

$$
\begin{aligned}
P(p-a < T < p+a) &= 1-\alpha \\
\iff P\Big(\frac{p-a-p}{\sqrt{\frac{pq}{n}}} < \frac{T-p}{\sqrt{\frac{pq}{n}}} < \frac{p+a-p}{\sqrt{\frac{pq}{n}}}\Big) &= 1-\alpha \\
\iff P\Big(-z_\alpha < \mathcal{Z} < z_\alpha\Big) &= 1-\alpha \\
\iff P\Big(|\mathcal{Z}| > z_\alpha\Big) &= 1-\alpha
\end{aligned}
$$

L'intervalle est donc $\Big[p - z_\alpha\sqrt{\frac{pq}{n}}, p + z_\alpha\sqrt{\frac{pq}{n}}\Big]$.

Chapitre 7

L'estimation

7.1 Position du problème

Nous désirons connaître, dans une population donnée la moyenne μ d'une variable aléatoire continue X. Au lieu de rechercher la valeur exacte de μ par l'examen exhaustif des individus de la population, on tire au sort un échantillon de taille n et on veut à partir de l'observation de $\bar{x}$ induire des renseignements sur μ.

Nous désirons connaître, dans une population, la proportion p de sujets présentant un certain caractère (i.e. la moyenne p d'une variable aléatoire de Bernoulli $Ber(p)$). Pour cela, on tire un échantillon de taille n et on veut induire de la proportion observée p_{obs} des renseignements sur p.

7.2 Estimation ponctuelle

Définition 7.1 *Un estimateur T d'un paramètre θ est une variable aléatoire dont la valeur - notée $\widehat{\theta}$ - prise à l'issue d'une expérience constitue l'estimation de θ :*

$$\begin{array}{rccc} T: & (\Omega, \mathcal{A}, P) & \longrightarrow & E \\ & \omega & \longmapsto & T(\omega) = \widehat{\theta} \end{array}$$

Exemple : Soit X une variables aléatoires réellescontinue de moyenne μ. μ est un paramètre, *i.e.* une valeur réelle fixe et inconnue. Alors :

$$\begin{array}{rccc} \bar{X}: & (\Omega^{(}\otimes, \mathcal{A}_{\setminus}, P) & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i = \bar{x} = \widehat{\mu} \end{array}$$

constitue un estimateur de μ et $\widehat{\mu} = \bar{x}$ constitue l'estimation de pour un échantillon donné de taille n.

Définition 7.2 *On appelle écart quadratique moyen de T relativement à θ la quantité :*

$$E\Big[(T-\theta)^2\Big].$$

Propriété 7.1 *On montre que :*

$$E\Big[(T-\theta)^2\Big] = var(T) + E^2(T-\Theta)$$

Définition 7.3 *Biais d'un estimateur*

On appelle biais d'un estimateur T du paramètre θ : $E(T-\theta) = E(T) - \theta$
- *Si $E(T-\theta) = 0 \iff E(T) = \theta$, T est dit sans biais*
- *Si $E(T-\theta) \neq 0 \iff E(T) \neq \theta$, T est dit biaisé*

7.2.1 Précision d'une moyenne

Définition 7.4 *Estimation ponctuelle d'une moyenne.*
Soit X une variable aléatoire continue. On appelle estimateur de la moyenne $\mu(X) = \mu$ la variable aléatoire :

$$\begin{array}{rccl} \bar{X}: & \Omega \times \Omega \times \ldots \Omega & \longrightarrow & \mathbb{R} \\ & (\omega_1, \ldots, \omega_n) & \longmapsto & \bar{X}(\omega) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X(\omega_i) = \bar{x} = \widehat{\mu} \end{array}$$

L'observation $\bar{x}$ de $\bar{X}$ pour un n-échantillon est dite estimation ponctuelle de μ.

Exemple : On observe le poids de $n = 100$ poissons. Leur poids moyen vaut 112 g ($\bar{x} = 112$ g). 112 g est une estimation ponctuelle de μ.

Remarque : L'estimation "ponctuelle" ne peut pas donner entièrement satisfaction puisqu'en raison des fluctuations d'échantillonnage elle fluctue d'un échantillon à l'autre. Il est donc nécessaire de connaître son degré de précision *i.e.* d'associer à $\widehat{\mu}$ un intervalle I dans lequel on a la quasi-certitude de cerner la valeur exacte μ.

Définition 7.5 *On appelle* **intervalle de confiance** *de niveau α pour l'estimation $\widehat{\mu}$, l'intervalle centré sur $\widehat{\mu}$ auquel μ a une probabilité $1-\alpha$ d'appartenir.*

Propriété 7.2 *L'intervalle de confiance associé à l'estimation $\bar{x}$ de l'espérance μ est, pour un échantillon de taille n, et pour un niveau α l'intervalle :*

(i) $IC_{1-\alpha} = \Big[\bar{x} - z_{alpha/2}\dfrac{\sigma}{\sqrt{n}},\ \bar{x} + z_{alpha/2}\dfrac{\sigma}{\sqrt{n}}\Big]$ *si σ est connue.*

(ii) $IC_{1-\alpha} = \Big[\bar{x} - z_{alpha/2}\dfrac{s(x)}{\sqrt{n}},\ \bar{x} + z_{alpha/2}\dfrac{s(x)}{\sqrt{n}}\Big]$ *si σ est inconnue.*

où $z_{alpha/2}$ / $P(|Z| < z_{alpha/2}) = 1-\alpha$.

"Démonstration"

On se sert de la connaissance de l'approximations de la distribution de $\bar{X}$. Au chapitre précédent, nous avons vu que $\mathcal{L}(\bar{X}) \simeq N(\mu, \dfrac{\sigma^2}{n}) \simeq N(\mu, \dfrac{s^2(x)}{n})$ lorsque σ^2 est inconnue.

On peut alors construire un intervalle de fluctuation -ou de pari- ed μ au risque α :

$$\Big[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ;\ \mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$$

On a donc : $P\Big(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) = 1-\alpha \iff P\Big(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big) = 1-\alpha$

Cela peut être compris ainsi :

Si on se fixe un risque α, on a une probabilité $1-\alpha$ que notre estimation $\bar{x}$ appartienne à l'intervalle de pari 7.2.1. On va donc définir l'intervalle de confiance comme étant égal à l'ensemble des valeurs de μ pour lesquelles l'intervalle de pari contient $\bar{x}$:

- μ_i ? La plus petite valeur -notée μ_i- acceptable pour μ est alors la valeur pour laquelle $\bar{x}$ serait limite droite de l'intervalle de pari : $\mu_i = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- μ_s ? La plus grande valeur -notée μ_s- acceptable pour μ est alors la valeur pour laquelle $\bar{x}$ serait limite gauche de l'intervalle de pari : $\mu_s = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

Remarque 1 : L'intervalle de confiance (IC_α) est différent de l'intervalle de fluctuation (IF_α) :

- L'intervalle de fluctuation $\Big[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$ est un intervalle fixe qui porte sur une valeur $\bar{x}$ aléatoire d'un échantillon.
- L'intervalle de confiance $\Big[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big]$ est un intervalle aléatoire qui porte sur une valeur μ fixe et inconnue.

Remarque 2 : L'intervalle de confiance (i) dépend de la valeur σ de l'écart-type $\sigma(X)$ pour la population. σ est en général inconnue, mais si on considère un "grand" échantillon, l'écart-type empirique $s(x)$ observé sur l'échantillon constitue une estimatiuon assez précise de σ. On peut donc dans (i) remplacer σ par $s(x)$ pour obtenir (ii).

Exemple : n=100, $\bar{x} = 112g$, $s(x) = 50$.

Si on se fixe un risque $\alpha = 0.05$, on en déduit l'intervalle de confiance de la moyenne μ des poids de l'ensemble des poissons du lac :

$$\Big[\mu_i = 112 - 1.96 \times \frac{50}{\sqrt{100}}\ ;\ \mu_s = 112 + 1.96 \times \frac{50}{\sqrt{100}}\Big] \simeq [102\ ;\ 122]$$

Remarque 2 : Précision de l'intervalle de confiance
Pour un riique α donné, la précision d'un intervalle de confiance est d'autant meilleure que l'intervalle de confiance est petit, c'est à dire que :

(i) Le caractère étudié est moins variable.

Par exemple, si $s(x) = 10$, on obtient :
$IC_{0.05} = \Big[\mu_i = 112 - 1.96 \times \frac{10}{\sqrt{100}}\ ;\ \mu_s = 112 + 1.96 \times \frac{10}{\sqrt{100}}\Big] \simeq [110\ ;\ 114]$.

(ii) La taille de l'échantillon est grande.

Par exemple si les résultats portent sur $n = 100000$ individus, l'intervalle de confiance de l'exemple précédent devient :

$$\left[\mu_i = 112 - 1.96 \times \frac{50}{\sqrt{10000}} \ ; \ \mu_s = 112 + 1.96 \times \frac{50}{\sqrt{10000}}\right] \simeq [110 \ ; \ 114]$$

$$= \quad \left[\mu_i = 112 - 1.96 \times \frac{50}{100} \ ; \ \mu_s = 112 + 1.96 \times \frac{50}{100}\right]$$

$$= \quad [111 \ ; \ 113]$$

Donc pour obtenir un intervalle de confiance 10 fois plus précis, il faut 100 fois plus de sujets ou individus.

7.2.2 Nombre de sujets nécessaires

On peut calculer le nombre de sujets nécessaires pour obtenir une précision fixée à l'avance.

Exemple : Combien faut-il de poissons pour que l'intervalle de confiance de la moyenne soit : $\bar{x} +$ ou $- 3g$ pour un coefficient de sécurité égal à 95 % ?

L'intervalle de confiance au risque $\alpha = 5\%$ vaut $\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \ ; \ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$. Donc $\mu_s - \mu_i = 2 \times 1.96\frac{\sigma}{\sqrt{n}} \simeq 4\frac{\sigma}{\sqrt{n}} = 6 \iff n = \left(\frac{4\sigma}{6}\right)^2$. Si σ est connu (et est égal à 50) alors nous en déduisons qu'il faut n=$\left(\frac{4 \times 50}{6}\right)^2 = (33.33)^2 \simeq 1111$ poissons.

Formule générale

Si on se fixe un intervalle de confiance de largeur i et un risque α

Chapitre 8

Principe des tests statistiques

Nous allons d'abord présenter le principe des tests statistiques au travers du test de comparaison d'une moyenne à une moyenne théorique. Nous verrons dans la suite du chapitre les tests de comparaison d'un pourcentage à un pourcentage théorique, de comparaison de deux moyennes, de comparaison de deux pourcentages.

8.1 Comparaison d'une moyenne à une moyenne théorique

8.1.1 Position du problème

On étudie un lac d'altitude peuplé d'une espèce de poissons. On connaît la taille moyenne de cette espèce dans des conditions "normales" ou "théoriques", elle vaut 110 g. On désire savoir si la moyenne des poissons du lac est différente ou pas de la moyenne théorique.

Afin de décider, on a prélevé dans le lac 100 poissons au hasard. Ces poissons ont été pesés. On a observé pour cet échantillon une moyenne empirique $\bar{x} = 112 \ \ g$ et une variance enpirique $s^2(x) = 2500 \ \ g^2$ soit $s(x) = 50 \ \ g$. On se demande donc si ces observations sont compatibles avec l'hypothèse : "la moyenne μ des poids des poissons du lac est egale à $110g$ ".

$$\begin{array}{rccl} \text{Notons } X : & (\Omega, \mathcal{A}, P) & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & X(\omega) = \text{Poids du poisson } \omega \end{array}$$

$\mathcal{L}(X)$ est inconnue, $\mu(X)$ est inconnue. Autrement dit, on veut au vu des observations choisir ou décider entre deux hypothèses :

$(H_0) \ \ \mu = 110g \ \ (\mu = \mu_0)$

$(H_1) \ \ \mu \neq 110g \ \ (\mu \neq \mu_0)$

La différence qu'on observe entre $\bar{x}$ et μ_0 : $\bar{x} - \mu_0 = 112g - 110g = 2g$ peut avoir deux origines :
$\longrightarrow$ Soit (H_0) est vraie et $\bar{x} - \mu_0$ correspond aux fluctuations de l'échantillonnage
$\longrightarrow$ Soit (H_0) est fausse et donc la moyenne μ est vraiment différente de μ_0
Les deux cas sont possibles. Il est donc impossible de répondre avec certitude à la question posée. On va donc choisir un des deux cas, à ce choix étant associé un risque de se tromper.

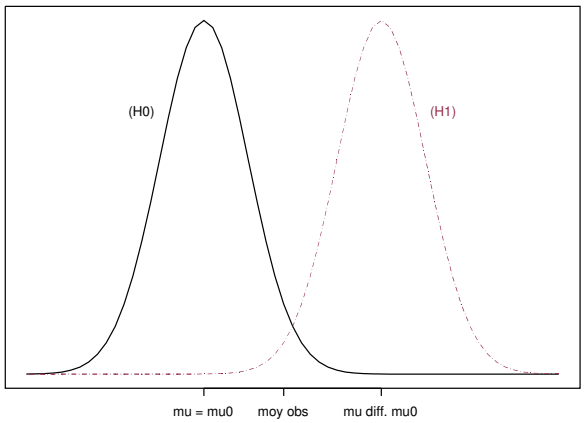


FIG. 8.1 – Distribution de la variable aléatoire moyenne sous (H0) -trait continu-, et sous (H1) - pointillés-

8.1.2 Procédure de décision

Construire un test de l'hypothèse nulle H_0 contre l'hypothèse alternative H_1, c'est établir une règle de décision permettant de choisir entre H_0 et H_1. Si H_0 est vraie, alors le poids des poissons (la variable aléatoire X) est distribuée suivant une distribution inconnue de moyenne $\mu_0 = 110g$ et de variance σ^2 inconnue. Donc d'aprés le chapitre 6 la variable aléatoire moyenne $\bar{X}$ est distribuée approximativement suivant une loi normale de moyenne $\mu_0 = 110g$ et de variance $\dfrac{\sigma^2}{n}$.

On peut alors construire un intervalle de fluctuation (ou de pari) de $\bar{X}$ au risque α fixé. Prenons par exemple $\alpha = 0.05$:

$$\begin{aligned} & \left[\mu_0 - 1.96\frac{\sigma}{\sqrt{n}}, \quad \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}\right] \\ = \quad & \left[110 - 1.96\frac{50}{10}, 110 + 1.96\frac{50}{10}\right] \qquad \text{en remplaçant } \sigma \text{ par } s(x) \\ \simeq \quad & [100, 120] \end{aligned}$$

Donc, si (H_0) est vraie, observer une valeur $\bar{x}$ hors de l'intervalle de pari ne se produit que dans $\alpha = 5\%$ des cas. C'est un événement qui est dit "rare". On constuit alors la règle de décision du test de la manière suivante :

$\longrightarrow$ Si $\bar{x}$ appartient à l'intervalle de fluctuation, $\bar{x}$ appartient à un événement considéré comme "non rare" si (H_0) est vraie. On décide alors de ne pas rejeter (H_0) au risque β de se tromper.
$\longrightarrow$ Si $\bar{x}$ n'appartient pas à l'intervalle de fluctuation, $\bar{x}$ appartient à un événement considéré comme "rare" si (H_0) est vraie. On décide de rejeter (H_0) au risque α de se tromper.

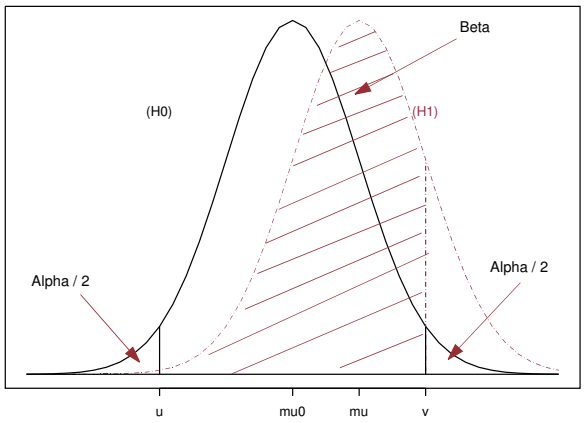


FIG. 8.2 – Erreurs de première espèce (α) et de seconde espèce (β) dans le cas d'un test bilatéral.

Conditions : Il faut que la distribution de X soit normale ou $n \geq 30$.

Exemple : $\bar{x} = 112g$ appartient à l'intervalle [100,120]. Donc, en appliquant la règle de décision précédente, on conclut au non rejet de (H_0) au risque $\alpha = 0.05$ de se tromper. (i.e. si (H_0) est vraie, et si on considère 100 échantillons, dans environ 5% des cas la moyenne $\bar{x}$ observée sortira de l'intervalle de fluctuation et donc on rejettera (H_0) à tort)

8.1.3 Interprétation

1. Pour l'exemple précédent, la conclusion du test correspond au fait que la différence observée $\bar{x} - \mu_0 = 112 - 110 = 2g$ ne peut pas, sauf à prendre un risque $\alpha \geq 0.05$, être attribuée à autre chose qu'aux fluctuations de l'échantillonnage. Cela ne veut pas dire que la vraie moyenne μ vaut $110g$ c'est pourquoi on dit "on ne rejette pas (H_0)" plutôt que "on accepte (H_0)".
2. (H_0) est l'hypothèse privilégiée. C'est celle que l'on conserve si le résultat n'est pas clair. On dit que le test est conservatif car il conserve (H_0) sauf si les données conduisent à le rejeter. En ce sens, on peut établir une analogie entre un test d'hypothèse et un procès : tout supect est présumé innocent et l'accusation doit apporter la preuve de sa culpabilité avant que la justice ne décide de le condamner. Cette preuve doit de plus s'appuyer sur des éléments matériels, comme le test qui utilise les données pour rejeter l'hypothèse. Quand on accepte (H_0), on ne prouve pas qu'elle est vraie, on accepte de conserver (H_0) parce que l'on a pas pu accumuler suffisamment de preuve contre elle. Accepter (H_0) c'est acquitter faute de preuve !
3. Un test est un mécanisme qui permet de trancher entre deux hypothèses (H_0 et H_1) au vu des résultats d'un échantillon. Un seule de ces deux hypothèses est vraie. Il existe donc 4 cas possibles schématisés dans le tableau ci-dessous avec les probabilités correspondantes.

Vérité → / Décision ↓	H_0	H_1
H_0	$1-\alpha$	β
H_1	α	$1-\beta$

Il existe ainsi deux types d'erreurs suivant la décision qui est prise :
- L'erreur de première espèce qui consiste à décider (H_1) alors que (H_0) est vraie. Sa probabilité est notée $P(H_1/H_0) = \alpha$. C'est la probabilité de condamner un innocent.
- L'erreur de seconde espèce qui consiste à décider de ne pas rejeter (H_0) alors que (H_1) est vraie. Sa probabilité est notée $P(H_0/H_1) = \beta$. C'est la probabilité d'acquitter un coupable.

4. Les risques α et β ne sont pas indépendants. C'est le principe du vendeur d'oranges. Imaginons un vendeur qui vend des oranges avec pépins (bon marché) et des oranges sans pépin (plus chères) : son risque est de vendre peu cher des oranges sans pépin alors que le risque du consommateur est d'acheter cher des oranges avec pépins.

   Dans le cas d'un test, α est fixé par l'utilisateur (c'est le risque associé à l'intervalle de fluctuation). β dépendra lui de α (plus α est petit plus β est grand), de n (plus n est grand plus β est petit), et de μ (plus l'écart $\mu - \mu_0$ est important plus β est petit).

5. Le risque β dépend de la vraie valeur de μ. Calculons son expression dans le cas du test de comparaison d'une moyenne observée à une moyenne théorique.

$$\begin{aligned}
\beta &= P_{H_1}(\mu_0 - u_{\frac{\alpha}{2}}, \mu_0 + u_{\frac{\alpha}{2}}) \\
&= P_{H_1}\Big(\mu_0 - u_{\frac{\alpha}{2}} < \bar{X} < \mu_0 + u_{\frac{\alpha}{2}}\Big) \\
&= P\Big(\frac{\mu_0 - u_{\frac{\alpha}{2}} - \mu}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{\mu_0 + u_{\frac{\alpha}{2}} - \mu}{\sigma/\sqrt{n}}\Big)
\end{aligned}$$

   Dans le cas de notre exemple, quel est le risque associé à notre décision si la vaie valeur de μ est égale à 112g ?

$$\begin{aligned}
\beta &= P\Big(\frac{110 - 10 - 112}{50/10} < \frac{\bar{X} - 110}{50/10} < \frac{110 + 10 - 112}{50/10}\Big) \\
&= P\Big(\frac{-12}{5} < \mathcal{Z} < \frac{8}{5}\Big) \\
&= P(-2.4 < \mathcal{Z} < 1.6) \\
&= P(\mathcal{Z} < 1.6) - 1 + P(\mathcal{Z} < 2.4) \\
&= 0.945 - 1 + 0.992 \\
&= 0.937
\end{aligned}$$

   Ainsi, si la vraie valeur de μ est $\mu = 112$ g, on a $93,7$ % de chances de conclure à (H_0) *i.e.* de se tromper !

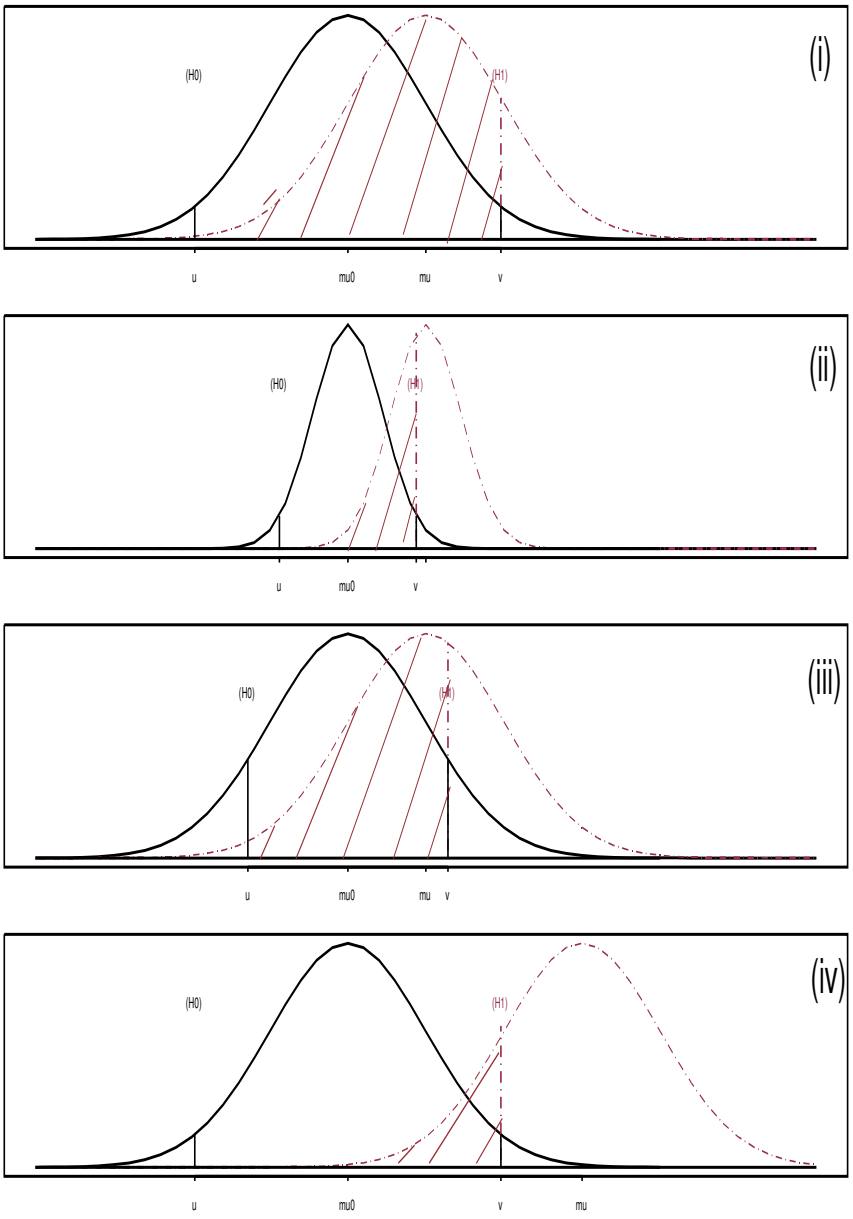


FIG. 8.3 – Evolution du risque de seconde espèce β : (i) test de référence ; (ii) On augmente la taille n de l'échantillon ; (iii) On augmente le risque de première espèce α ; (iv) L'écart $\mu - \mu_0$ augmente.

6. **Puissance d'un test**
   On appelle puissance d'un test la quantité $1-\beta$. Elle mesure la performance du test c.a.d la probabilité de mettre en évidence une hypothèse H_1 qui existe. La puissance d'un test peut être comparée à une loupe : si on perçoit un signe on peut affirmer son existence ; si on ne le perçoit pas peut être serait-il perceptible avec une loupe plus puissante ?

8.2 Comparaison d'une proportion à une proportion théorique

8.2.1 Position du problème

Un lac est peuplé de truites et d'autres espèces. On note p la proportion (inconnue) de truites.

On peut alors considérer la variable aléatoire :

$$\begin{array}{rccll} X: & \Omega & \longrightarrow & \{0,1\} & \\ & \omega & \mapsto & X(\omega) & = 1 \text{ si c'est une truite} \\ & & & & = 0 \text{ si c'est un omble} \end{array}$$

La variable aléatoire X suit une loi de Bernoulli $Ber(p)$

On veut répondre à la question suivante : " La proportion p de truites du lac est-elle supérieure à 0,5 ?"

Autrement dit, on doit donc décider entre :

(H_0) $p = 0.5$ $(p = p_0)$ et

(H_1) $p > 0.5$ $(p > p_0)$

Pour répondre à la question, on va pêcher n=100 poissons et on va calculer la proportion empirique $\hat{p}$ (ou p_{obs}) de truites.

8.2.2 Procédure de décision

Si (H_0) est vraie, alors la distribution de la variable aléatoire $\bar{X}$ (ou de l'urne des proportions) suit approximativement (si np_0>5 et nq_0>5) une loi normale de moyenne $\mu = p_0$ et de variance $\sigma^2 = \dfrac{p_0 \times q_0}{n}$: $\mathcal{N}(\mu = 0.5,\ \sigma = 0.05)$.

Si on se fixe un risque α, par exemple α=0.05, on peut contruire un intervalle de fluctuation des proportions auquel la proportion observée sur un échantillon de taille n a un risque α de ne pas appartenir si (H_0) est vraie :

$$\begin{array}{rcll} IF_\alpha & = & \Big(-\infty, p_0 + z_\alpha \sqrt{\dfrac{p_0 \times q_0}{n}} \Big] & \text{où} \quad z_\alpha \ / \ P(Z < z_\alpha) = 1 - \alpha \\ & = & \Big(-\infty, 0.5 + 1.64 \times \sqrt{\dfrac{0.5 \times 0.5}{100}} \Big] & \\ & = & (-\infty,\ 0.535] & \end{array}$$

Remarque : On n'a pas pris un intervalle symétrique car comme (H_1) est $p > 0.5$, si on prenait un intervalle symétrique la zone de rejet située "à gauche" ne correspondrait pas à l'hypothèse (H_1). Donc, on fait passer "tout le risque " à droite.

FIG. 8.4 – Erreurs de première espèce (α) et de seconde espèce (β) dans le cas d'un test unilatéral.

Donc si (H_0) est vraie, 95% des échantillons de 100 poissons auront une proportion observée de truite comprise entre $(-\infty,\ 0.535]$.

On construit alors la règle de décision suivante :

- Si p_{obs} appartient à l'intervalle de fluctuation IF_α alors on ne rejette pas l'hypothèse (H_0) au risque β de se tromper.
- Si p_{obs} n'appartient pas à l'intervalle de fluctuation IF_α alors on rejette l'hypothèse (H_0) au risque α de se tromper.

8.2.3 Exemple

On observe pour l'échantillon considéré $p_{obs} = 0.43$. Nous décidons donc de ne pas rejeter (H_0)(au vu des observations) au risque β de se tromper.

8.2.4 Conditions d'application

Elles sont liées à l'approximation de la distribution de $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, (i.e. de l'urne proportion) par une loi normale : $n_0 p > 5$ et $n_0 q > 5$. Où p_0 est le pourcentage théorique i.e. sous (H_0).

8.3 Test bilatéral et test unilatéral

Les tests de comparaison consistent à choisir entre :

$$(H_0) \quad \mu = \mu_0 \quad (\text{ resp. } p = p_0)$$
$$\text{et} \quad (H_1)$$

L'hypothèse (H_1) peut être de deux types :

(i) (H_1) $\mu \neq \mu_0$, (c'est l'exemple des moyennes qui a été vu précédemment)

(ii) (H_1) $\mu > \mu_0$ ou $(\mu > \mu_0$, (c'est l'exemple des proportions qui a été vu précédemment)

(i) Dans ce cas le test est dit **bilatéral** car la zone de rejet du test est composée de deux demi-droites : $\left(-\infty, \mu_0 - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right]$ et $\left(\mu_0 + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, -\infty \right]$. Par exemple, si $\alpha = 0.05$, alors $z_{alpha/2} = 1.96$.

(ii) Dans ce cas le test est dit **unilatéral** car la zone de rejet du test n'est constituée que par une demi-droite : $\left(\mu_0 + z_{\alpha} \times \frac{\sigma}{\sqrt{n}}, -\infty \right]$. Par exemple, si $\alpha = 0.05$, alors $z_{alpha} = 1.64$.

On est conduit à choisir (H_1) sous la forme $p_0 < p$ dans deux types de circonstances :

1. $p < p_0$ ne paraît pas possible, sur la base de ce que l'on connaît de la question. Par exemple, comparaison de la proportion p de gens malades lorsqu'ils sont soumis à un toxique à la proportion p_0 de malades sans toxique.
2. $p < p_0$, tous en étant possible, n'est pas interessant. C'est le cas lorsqu'on étudie un nouveau traitement qui ne sera jugé efficace que s'il est meilleur que le traitement de référence.

Chapitre 9

Le test du χ^2

9.1 Comparaison d'une répartition observée à une répartition théorique- Test du χ^2

9.1.1 Position du problème

Exemple 1 : On a effectué le croisement de basalmines blanches à grandes fleurs (B/G) avec des basalmines rouges à petites fleurs (r/p). En première génération, toutes les fleurs sont blanches à grandes fleurs. On obtient en deuxième génération quatre catégories avec les effectifs suivants :

	B/G	B/p	r/G	r/p
Effectifs	1790	547	548	213

Soit la variable aléatoire $C =$ "couleur de la fleur en F_2"

$$\begin{array}{rccll} C: & \Omega & \longrightarrow & \{1,2,3,4\} & \\ & \omega & \longmapsto & X(\omega) & = 1 \text{ si B/G.} \\ & & & & = 2 \text{ si B/p.} \\ & & & & = 3 \text{ si r/G.} \\ & & & & = 4 \text{ si r/p.} \end{array}$$

Peut on dire que la distribution phénotypes est conforme à la répartition mendelienne *i.e.*

$$\mathcal{L}(C) = \left(p_1 = \frac{9}{16};\ p_2 = \frac{3}{16};\ p_3 = \frac{3}{16};\ p_4 = \frac{1}{16}\right) \ ?$$

Exemple 2 : On a mesuré la taille de 217 oiseaux, en millimètre. Les observations sont regroupées dans le tableau suivant :

Classes	<685	685-705	705-725	725-745	>745
Effectifs	3	31	74	67	42

Peut-on dire que la distribution des tailles est conforme à une loi normale ?

9.1.2 Définition du χ^2

Soit une urne contenant k catégories de boules différentes en proportion $p_1, p_2 ... p_k$. Comme $\sum_{i=1}^{k} p_i = 1$, la composition de l'urne est parfaitement définie par $(k-1)$ de ces proportions.

Définition 9.1 *$k-1$ est dit nombre de degrés de libertés (ddl). C'est le nombre de proportions permettant de déterminer parfaitement la composition de l'urne.*

Par exemple, l'urne théorique de l'exemple 1 correspondant à la théorie mendelienne à les propriété suivantes :

$$p_1 = \frac{9}{16};\ p_2 = \frac{3}{16};\ p_3 = \frac{3}{16}$$
$$p_4 = 1 - \frac{(9+3+3)}{16} = \frac{1}{16}$$

Si on extrait par tirage au sort $n = 3098$ boules, et si la répartition de l'échantillon était la même que celle de l'urne théorique on observerait :

$$\begin{array}{llll} C_1 = & np_1 = & 1742.625 & \text{B/G} \\ C_2 = & np_2 = & 580.875 & \text{B/p} \\ C_3 = & np_3 = & 580.875 & \text{r/G} \\ C_4 = & np_4 = & 193.625 & \text{r/p} \end{array}$$

Il existe donc un "écart" entre la composition de l'échantillon et la composition théorique. On se propose de préciser les fluctuations d'échantillonage de cet écart. Il est d'abord nécessaire de caratériser cet écart *i.e.* de choisir un indice exprimant "l'écart" entre la composition théorique et la composition observée :

Phénotype	B/G	B/p	r/G	r/p	Total
p_i	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	1
Effectifs observés O_i	1790	547	548	213	3098
Effectifs théoriques C_i	1742.625	580.875	580.875	193.625	3098
$\frac{(O_i - C_i)^2}{C_i}$	1.288	1.975	1.86	1.939	7.06

Indice ?
- Somme des écarts ? Non, car elle est nulle.
- Somme des valeurs absolues ? Non, car ne se prête pas facilement au calcul des probabilités.
- Somme des carrés des écarts ? Il donne le même poids à tous les écarts, qu'ils se rapportent à de petits ou à des grands effectifs calculés.

On choisit donc comme indice : $KH = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2} + ... + \frac{(O_k - C_k)^2}{C_k} = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$

Pour l'exemple 1, nous obtenons :

$$KH_{obs} = \frac{(1742.625 - 1790)^2}{1742.625} + \frac{(580.875 - 547)^2}{580.875} + \frac{(580.875 - 548)^2}{580.875} + \frac{(193.625 - 213)^2}{193.625} = 7.06$$

Avantage : Si l'échantillon provient effectivement de l'urne théorique, alors on connaît la distribution de KH donc ses fluctuations d'échantillonnage.

Propriété 9.1 *Si l'échantillon provient de la distribution théorique (p_1.. p_k), alors* $KH = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$ *suit approximativement une loi de χ^2 (chi-deux) à $k-1$ ddl. On note $\mathcal{L}(KH) = \chi^2_{k-1}$.*

Remarques :

1. L'avantage principal de la variable KH est que sa distribution ne dépend pas de la composition de l'urne, mais ne dépend que du nombre de catégories ou modalités.
2. Condition de validité de la propriété. La propriété précédente est valide lorsque n tend vers $+\infty$ et n_i tend vers $+\infty$ quelque soit i. Concrètement on considère l'approximation comme correcte dès que les effectifs calculés $C_i = np_i \geq 5$.

9.1.3 Comparaison d'une répartition théorique à une répartition observée

Reprenons l'exemple 1. On se demande s'il est plausible au vu de l'échantillon que la distribution des couleurs soit conforme à la distribution théorique mendelienne. Autrement dit on doit décider entre :

$$(H_0) : \quad (p_1 \ ; \ p_2 \ ; \ p_3 \ ; \ p_4) \quad = \left(\frac{9}{16} \ ; \ \frac{3}{16} \ ; \ \frac{3}{16} \ ; \ \frac{1}{16}\right)$$

$$(H_1) : \quad (p_1 \ ; \ p_2 \ ; \ p_3 \ ; \ p_4) \quad \neq \left(\frac{9}{16} \ ; \ \frac{3}{16} \ ; \ \frac{3}{16} \ ; \ \frac{1}{16}\right)$$

Si l'hypothèse (H_0) est vérifiée, alors d'aprés la propriété précédente $KH = \sum_{i=1}^{4} \frac{(O_i - C_i)^2}{C_i}$ suit approximativement une distribution du χ^2 à $4 - 1 = 3$ ddl. Les conditions d'approximation de la distribution sont ici vérifiées (tous les $C_i \geq 5$). On peut alors construire la règle de décision du test en se fixant un risque de première espèce $\alpha = 0.05$ et en définissant $x_{th,3}$ comme la valeur telle que $P_{(H_0)}(KH \geq x_{th}) = 0.05$:

- Si $KH_{obs} \geq x_{th}$, on rejette (H_0) au risque α de se tromper.
- Si $KH_{obs} < x_{th}$, on ne rejette pas (H_0) au risque β (inconnu ici) de se tromper.

Dans notre cas le $KH_{obs} = 7,06$ et $x_{th} = 7,81$. on conclut donc que l'on ne rejette pas (H_0) au risque β de se tromper, c'est à dire qu'on ne peut, au vu de notre échantillon, rejeter l'hypothèse que la distribution des phénotypes obéit à la loi de mendel.

9.1.4 Comparaison d'une répartition observée à une répartition théorique dépendante de un ou plusieurs paramètres

On a mesuré la taille de 217 oiseaux en mm (exemple 2). Les observations sont regroupées en classes dans le tableau suivant :

Classes	<685	685-705	705-725	725-745	>745	Total
Effectifs observés O_i	3	31	74	67	42	217
Effectifs théoriques $C_i = np_i$	5.22	26.18	64.84	73.2	47.55	217
$\frac{(O_i - C_i)^2}{C_i}$	0.94	0.89	1.29	0.53	0.65	4.29

La moyenne des tailles de l'échantillon des 217 oiseaux est la moyenne $\bar{x} = 728.1$ et l'écart type $s(x) = 21.8$. Ces valeurs ont été calculées avant de mettre en classe les observations.

(H_0) : la distribution des tailles est une loi normale

(H_1) : la distribution des tailles n'est pas une loi normale

La différence avec le cas précédent (exemple 1) c'est que (H_0) correspond à une loi (normale, dans notre exemple) dont les paramètres μ et σ sont inconnus.

Pour pouvoir utiliser la statistique du χ^2 on doit calculer, sous l'hypothèse (H_0), les nombres p_i correspondants aux probabilités des différentes classes considérées :

$$p_1 = P(T < 685) = P\left(\frac{T-\mu}{\sigma} < \frac{685-\mu}{\sigma}\right) = P\left(Z < \frac{685-\mu}{\sigma}\right) = P(Z < -1.977) = 0.024$$

$$p_2 = P(685 < T < 705) = P\left(\frac{685-\mu}{\sigma} < Z < \frac{705-\mu}{\sigma}\right) = P(-1.977 < Z < -1.06) = 0.121$$

$$p_3 = P(705 < T < 725) = P\left(\frac{705-\mu}{\sigma} < Z < \frac{725-\mu}{\sigma}\right) = P(-1.06 < Z < -0.14) = 0.299$$

$$p_4 = P(725 < T < 745) = P\left(\frac{725-\mu}{\sigma} < Z < \frac{745-\mu}{\sigma}\right) = P(-0.14 < Z < 0.775) = 0.337$$

$$p_5 = P(T > 745) = 1 - P\left(Z < \frac{745-\mu}{\sigma}\right)) = 1 - 0.781 = 0.219$$

On choisit la loi normale la plus plroche de nos observations i.e. $\mathrm{N}(\mu = 728.1; \sigma = 21.8)$.

On peut alors completer le tableau précédent.

Propriété 9.2 *Si on estime p paramètres, alors si* (H_0) *est vraie,* $KH = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$ *suit approximativement une distribution dite de* χ^2 *à* $k-1-p$ *degrés de libertés.*

Conditions d'approximation : n tend $+\infty$ et ni tend $+\infty$ quelque soit i. Concrètement, on conidère les conditions d'approximation remplis dès que $C_i = np_i > 5$

On peut alors d'après la propriété précédente dire que si (H_0) est vérifié alors la statistique $KH = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$ suit approximativement une loi de χ^2 à $5-1-2 = 2$ degrés de libertés.
Les conditions d'approximations sont vérifiées (tous les $C_i \geq 5$). On peut alors construire la règle de décision du test. On se fixe pour cela un risque d'erreur de première espèce $\alpha = P(H_1/H_0) = 0.05$ et on définit $x_{th,2}$ comme la valeur telle que : $P_{(H_0)}(KH \geq x_{th}) = 0.05$ ici $x_{th} = 5.99$. Alors :

- Si $KH_{obs} \geq x_{th}$, on rejette (H_0) au risque α de se tromper.
- Si $KH_{obs} < x_{th}$, on ne rejette pas (H_0) au risque β (inconnu ici) de se tromper.

Dans notre cas le $KH_{obs} = 4.296$ et $x_{th} = 5.99$. on conclut donc que l'on ne rejette pas (H_0) au risque β de se tromper.

9.2 Comparaison de plusieurs répartitions observées : test du χ^2 d'indépendance

On étudie trois lacs (Lac A, B,C) peuplés tous trois de quatre espèces, omble chevalier, truite commune, samon de fontaine (SF), truite arc-en-ciel (TAC).

Question : "la répartition des quatre espèces est elle équivalente dans les trois lacs" ou encore "la répartition des espèces est indépendante du lac" ? Pour cela, on va prélever dans chaque lac un échantillon. On observe la répartition suivante :

Classes	Omble	Truite	SF	TAC	Total
Lac A	17	30	12	14	73
Lac B	10	37	3	20	70
Lac C	22	19	7	14	62
Total	49	86	22	48	205

9.2.1 Définition du χ^2

Le tableau précédent est dit tableau de contingence de deux variables. Dans notre exemple il s'agit des variables aléatoires lac (3 modalités i=A, B ou C) et espèce (5 modalités).
On note :

p_{ij} = P(Lac="i" et Espèce="j")

pi = P(Lac="i"), quelques soient les espèces

pj = P(Espèce="j") quels que soient les lacs.

Répondre à la question posée va consister à choisir entre les deux hypothèses :

(H_0) $p_{ij} = p_i \times p_j$ i.e. les variables espèce et lac sont indépendantes

(H_1) $p_{ij} \neq p_i \times p_j$ i.e. les variables espèce et lac ne sont pas indépendantes

Sous l'hypothèse (H_0), les effectifs théoriques (tableau de contingence théorique) sont alors égaux à :

$$C_{ij} = n \times p_i \times p_j = n \times \frac{n_{i.}}{n} \times \frac{n_{.j}}{n} = \frac{n_{i.} \times n_{.j}}{n}$$

où on remplace p_i et p_j par leurs estimations $\widehat{p_i} = \frac{n_{i.}}{n}$ et $\widehat{p_j} = \frac{n_{.j}}{n}$.

On en déduit le tableau des effectifs théoriques (valeurs entre parenthèses) :

Classes	Omble	Truite	SF	TAC	Total
Lac A	17 (17.45)	30 (30.62)	12 (7.8)	14 (17.1)	73
Lac B	10 (16.73)	37 (29.37)	3 (7.51)	20 (16.4)	70
Lac C	22 (14.82)	19 (26)	7 (6.65)	14 (14.52)	62
Total	49	86	22	48	205

Nous en déduisons que sous (H_0) :

$$KH_{obs} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - C_{ij})^2}{C_ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.} \times n_{.j}}{n})^2}{\frac{n_{i.} \times n_{.j}}{n}}$$

suit une distribution du χ^2 à $I \times J - 1 - (I-1) - (J-1) = (I-1) \times (J-1)$ ddl $= 6$ ddl.

Remarque : $(I-1) + (J-1)$ est le nombre de paramètres estimés : (I-1) valeurs de p_i (si on en connaît $(I-1)$ on connaît la dernière), et $(J-1)$ valeurs de p_j.

On en déduit la règle de décision du test en se fixant un risque de première espèce $\alpha = 0.05$ et où x_{th} est la valeur telle que : $P(\chi^2_6 \geq x_{th}) = 0.05$ $(x_{th} = 12.59)$:

- Si $\chi^2_{obs} \geq x_{th}$ alors on rejette (H_0) au risque $\alpha = 0.05$ de se tromper
- Si $\chi^2_{obs} < x_{th}$ alors on ne rejette pas (H_0) au risque β (inconnu ici) de se tromper.

Dans notre exemple, $\chi^2_{obs} = 16.04$ est supérieure à x_{th}. On conclut donc au rejet de (H_0) au risque $\alpha = 0.05$ de se tromper.

Chapitre 10

Comparaison de plusieurs moyennes : analyse de la variance

10.1 Comparaison de deux variances

10.1.1 Exemple

On désire comparer les précisions d'une machine fabriquant des comprimés lorsqu'elle produit de "'petits" comprimés (type A) et de "grands comprimés" (type B). Pour cela, on prélève pour chaque type de production un échantillon de comprimés que l'on pèse. Les résultats obtenus sont les suivants :

Type	Taille de l'échantillon	$\bar{x}$	$s^2(x)$
A	$n_A = 7$	$8,2$	$5,1$
B	$n_B = 11$	$15,4$	$6,2$

On considère donc : $X_A : \begin{array}{ccc} \Omega_A & \longrightarrow & \mathbb{R} \\ \omega & \longmapsto & X_A(\omega) = \text{Poids du comprimé } \omega \end{array}$

On considère donc : $X_B : \begin{array}{ccc} \Omega_B & \longrightarrow & \mathbb{R} \\ \omega & \longmapsto & X_B(\omega) = \text{Poids du comprimé } \omega \end{array}$

On va construire un test permettant de tester :

$$\begin{array}{rccc} & (H_0) & \sigma^2(X_A) & = & \sigma^2(X_B) \\ \text{contre} & (H_1) & \sigma^2(X_A) & \neq & \sigma^2(X_B) \end{array}$$

Propriété 10.1 *Si les variables aléatoires X_A et X_B suivent une loi normale et sont indépendantes alors : $Y_A = s_A^2 \dfrac{n_A - 1}{\sigma_A^2}$ et $Y_B = s_B^2 \dfrac{n_B - 1}{\sigma_B^2}$ suivent respectivement des lois de $\chi^2_{n_A-1}$ et $\chi^2_{n_B-1}$ et sont indépendantes.*

Conséquence : D'après la définition de la loi de Fisher-Snedecor, alors $F = \dfrac{Y_A}{n_A - 1} \Big/ \dfrac{Y_B}{n_B - 1} = \dfrac{s_A^2}{\sigma_A^2} \times \dfrac{s_B^2}{\sigma_B^2}$ est distribuée selon une loi F_{n_A-1,n_B-1}.

10.1.2 Construction du test

Sous (H_0), $F = \frac{s_A^2}{s_B^2}$ est distribuée selon une loi F_{n_A-1,n_B-1}. Si on se fixe un risque de première espèce α, on peut construire un intervalle de fluctuation de F sous H_0 au risque α : $IF_{1-\alpha} = [f_i,\ f_s]$, où f_i / $P(F_{n_A-1,n_B-1} > f_s) = \alpha/2$ et f_s / $P(F_{n_A-1,n_B-1} < f_i) = \alpha/2$.

GRAPHE

On peut alors construire la règle de décision du test :

- Si $F_{obs} \in IF_{1-\alpha}$ alors on ne rejette pas (H_0) au risque β.
- Si $F_{obs} \notin IF_{1-\alpha}$ alors on rejette (H_0) au risque α.

Exemple Sous (H_0), le rapport $F = \frac{s_A^2}{s_B^2}$ suit une distribution $F_{6,1}$. Si on se fixe un risque $\alpha = 0.05$ alors l'intervalle de fluctuation de F est $IF_{0.95} = [0.18; 4.07]$. La valeur observée de la statistique est $F_{obs} = \frac{5.1}{6.2} = 0.82 \in IF_{0.95}$

D'après la règle de décision du test on conclut au non rejet de l'hypothèse d'homogénéité des variances au risque β de se tromper.

Remarques :

1- Les conditions de validité du test sont la normalité de X_A et X_B et leur indépendance.

2- La valeur f_s de la borne supérieure de l'intervalle de fluctuation a été obtenue à partir d'une table (dont nous verrons l'utilisation en TD) donnant f_s / $P(F_{n_A-1,n_B-1} > f_s) = \alpha'$ (en choisissant $\alpha' = \frac{\alpha}{2}$). On peut se servir de cette table pour calculer la valeur de f_i de la manière suivante :

$$\begin{aligned} & P(F_{n_A-1,n_B-1} < f_i) &=& \alpha/2 \\ \iff & P(\frac{s_A^2}{s_B^2} < f_i) &=& \alpha/2 \\ \iff & P(\frac{s_B^2}{s_A^2} > \frac{1}{f_i}) &=& \alpha/2 \\ \iff & P(F_{n_B-1,n_A-1} > \frac{1}{f_i}) &=& \alpha/2 \end{aligned}$$

Il suffit donc de lire dans la table la valeur de $\frac{1}{f_i}$ et d'en déduire f_i.

3- Le test précédent est dit **bilatéral** parce que l'hypothèse alternative étant $\sigma^2(X_A) \neq \sigma^2(X_B)$, la zone de rejet est décomposée en deux parties correspondant respectivement à :

$P(F > F_S) < \alpha/2$ $(i.e.$zone où$s^2(X_A) >> s^2(X_B))$

et $P(F < F_i) < \alpha/2$ $(i.e.$zone où$s^2(X_B) >> s^2(X_A))$

Ainsi, on rejettera (H_0) lorsque $s^2(X_A) >> s^2(X_B)$ ou $s^2(X_B) >> s^2(X_A)$.

On peut aussi considérer comme hypothèse alternative :

$$(H_1)\ \ \sigma^2(X_A) \geq \sigma^2(X_B)$$

On définira alors comme zone de rejet du test l'intervalle : $IF_{1-\alpha} = [f_s, +\infty)$ tel que $P(F_{n_A-1,n_B-1} > f_s) < \alpha$. Dans ce cas le test est dit unilatéral : on ne s'interesse plus qu'à mettre en évidence le cas où $\sigma^2(X_A) \geq \sigma^2(X_B)$, donc à répondre à la question " La machine est-elle moins précise quand elle produit les gros comprimés" alors que dans le cas unilatéral on cherchait à répondre à la question : "Les productions correspondent-elles à deux précisions différentes ?"

10.2 Analyse de la variance à un facteur

10.2.1 Position du problème

On désire comparer les rendements de quatre variétés de blé (A, B, C et D). Pour cela, on a planté chaque variété dans 6 parcelles et on a observé les rendements obtenus :

Type de variété	Rendement moyen (quintaux/ha)	s^2
A	51	81
B	52.1	85
C	69.3	103
D	57.7	94

On peut définir X_A (resp. X_B, X_C, et X_D) :

$$\begin{array}{rccl} X_A : & \Omega_A & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & X_A(\omega) = \text{Rdt obtenu pour la parcelle} \omega \\ & & & \text{ensemencée par la variété A} \end{array}$$

On désire comparer les moyennes des variables rendements, autrement dit décider entre les deux hypothèses :

$$\begin{array}{ll} (H_0) & \mu_A = \mu_B = \mu_C = \mu_D \\ (H_1) & \text{Il y a au moins une différence entre deux des quatre moyennes} \end{array}$$

où on note $\mu_A = \mu(X_A)$; $\mu_B = \mu(X_B)$; $\mu_C = \mu(X_C)$ et $\mu_D = \mu(X_D)$.

10.2.2 Principe de l'analyse de la variance

Si (H_1) est vraie, les moyennes μ_A, μ_B, μ_C et μ_D ne sont pas égales. Lorsqu'on regroupe les quatre populations (*ie* les rendements des quatre variétés), la moyenne générale est μ et la variance totale est σ_T^2. σ_T^2 sera d'autant plus grande que les moyennes μ_A, μ_B, μ_C et μ_D sont dispersées, c'est à dire que les différences entre elles sont importantes.

A l'inverse, si (H_0) est vraie, les moyennes μ_A, μ_B, μ_C et μ_D sont égales et σ_T^2 est égale à σ^2 qui est la variance au sein de chacune des populations, *ie* la variance de chacune des variables aléatoires X_A, X_B, X_C et X_D.

Le principe du test de comparaison des moyennes va alors être de décomposer la variabilité totale σ_T^2 en une variabilité intragroupe (dispersions individuelles à l'intérieur des groupes) à une variabilité intergroupe (dispersion des moyennes des groupes), puis à comparer ces variabilité.

- Sont-elles égales ? Alors la dispersion des moyennes n'est pas plus importante que celle des observations individuelles, donc les moyennes sont considérées comme identiques.
- La variabilité inter est-elle supérieure à la variabilité intra ? Alors la dispersion des moyennes est plus importante que celle des observations individuelles, donc les moyennes sont considérées comme différentes.

On transforme ainsi le problème initial (comparer les moyennes) en une comparaison de variances : c'est le principe de l'analyse de la variance.

10.2.3 Analyse de la variance à un facteur à effets fixes

Décomposition de la variabilité des observations

La variabilité des observations des k ($k = 4$ dans notre exemple) échantillons est mesurée par la somme des carrés des écarts totale :

$$SC_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

où k désigne le nombre de population et n_j l'effectif de l'échantillon de la population j.
On peut décomposer SC_T ainsi :

$$\begin{aligned} SC_T &= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\underbrace{x_{ij} - \bar{x_j}}_{\text{variabilité intra-groupe}} + \underbrace{\bar{x_j} - \bar{x}}_{\text{variabilité inter-groupe}})^2 \\ &= \underbrace{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x_j})^2}_{\text{SCR}} + \underbrace{\sum_{j=1}^{k} n_j (\bar{x_j} - \bar{x})^2}_{\text{SC}_F} \end{aligned}$$

SC_F ne dépend que de la dispersion des moyennes $\bar{x_j}$ des groupes. Elle est appellée somme des carrés des écarts inter-groupes ou factorielle. La variance empirique correspondante (variance inter-groupe) est alors $s_F^2 = \dfrac{SC_F}{k-1}$.

SC_R ne dépend que de la dispersion des valeurs au sein de chaque échantillon, elle est appellée somme des carrés des écarts intra-groupes ou résiduelle car c'est "ce qui reste" de la somme de carrés totale une fois enlevée la variation entre les groupes. La variance empirique correspondante est $s_{\text{R}}^2 = \dfrac{SC_{\text{R}}}{n-k}$.

Valeurs théoriques des sommes des carrés des écarts

Propriété 10.2 *Si les variables aléatoires X ont même variance σ^2 dans chaque population, alors :*

- $\mu(SCE_T) = (n-1)\ \sigma^2 + \sum_{j=1} k\ (\mu_j - \mu)^2$
- $\mu(SCE_R) = (n-k)\ \sigma^2$

- $\mu(SCE_F) = (k-1)\ \sigma^2\ +\ \sum_{j=1}^{k} n_j(\mu_j - \mu)^2$

En appliquant les résultats de la propriété précédente, on en déduit que :

- $\mu(s_R^2) = \sigma^2$
- $\mu(s_F^2) = \sigma^2 + \dfrac{1}{k-1}\sum_{j=1}^{k} n_j(\mu_j - \mu)^2$

Ainsi, la variance entre groupes constitue une estimation non biaisée de σ^2 seulement si (H_0) est vraie (car alors $\mu_j = \mu$). Sinon, $\mu(s_F^2)\ >\ \sigma^2$.

On peut alors exprimer les hypothèses testées de façon différentes, mais équivalentes :

(H_0) $\mu_1 = \mu_2 = \ldots = \mu_k$	$\Longleftrightarrow$	(H_0) $\sigma_F^2 = \sigma_R^2$
(H_1) Il y a au moins une différence entre les kmoyennes.		(H_1) $\sigma_F^2 > \sigma_R^2$

Réalisation du test

Propriété 10.3 *Si les k variables aléatoires X_A, X_B,... indépendantes sont distribuées normalement et ont même variance σ^2 alors :*

$\dfrac{SC_R}{\sigma^2}$ *suit une distribution χ^2_{n-k}. Si de plus (H_0) est vraie, alors :*

(i) $\dfrac{SC_F}{\sigma^2}$ *suit une distribution χ^2_{k-1}*

(ii) On montre par ailleurs que $\dfrac{SC_R}{\sigma^2}$ *est indépendante de* $\dfrac{SC_F}{\sigma^2}$.

On en déduit donc que si (H_0) est vraie alors :

$$F = \frac{SC_F/(k-1)}{SC_R/(n-k)} \text{ est distribuée selon une distribution } F_{k-1,n-k}$$

Si on se fixe un risque α, on peut construire un intervalle de fluctuation de F au risque α : $[0, f_s]$ où f_s / $P(F_{k-1,n-k} > f_s)\ <\ \alpha$.

On peut alors définir la règle de décision du test :

- Si $F_{obs}\ <\ f_s$ alors on ne rejette pas l'hypothèse d'égalité des moyennes au risque β de se tromper.
- Si $F_{obs}\ >\ f_s$, alors on rejette l(hypothèse d'égalité des moyennes au risque α de se tromper.

Exemple (suite)

Nous avons $k = 4$ et $n = 24$. Sous (H_0), $F = \frac{SC_F/3}{SC_R/20}$ est distribuée selon une distribution $F_{3,20}$. Si on se fixe un risque de première espèce $\alpha = 0.05$, alors nous en déduisons l'intervalle de fluctuation au risque α : $[0\ ;\ f_s] = [0\ ;\ 3,86]$ où $f_s\ /\ P(F_{3,20} > f_s) < \alpha$.
Calculons maintenant F_{obs} :

$$\begin{aligned} SC_F &= \sum_{j=1}^{k} n_j\ (\bar{x_j} - \bar{x})^2 &&= \sum_{j=1}^{k} n_j\ \bar{x_j}^2\ -\ n\bar{x}^2 \\ &&&= 6\Big[51^2 + 52.1^2 + 69.3^2 + 57.7^2\Big] - 24 \times (57.525)^2 \\ &&&= 1264.125 \end{aligned}$$

$$\begin{aligned} SC_R &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x_j})^2 &&= \sum_{j=1}^{k}(n_j - 1) \times s_j^2 \quad \text{car } s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x_j})^2 \\ &&&= 5 \times \Big[81 + 85 + 103 + 94\Big] \\ &&&= 1815. \end{aligned}$$

Donc $F_{obs} = 4.64\ >\ f_s$. D'après la règle de décision du test, on conclut au rejet de l'hypothèse (H_0) au risque $\alpha = 0.05$ de se tromper.

Présentation des résultats d'une analyse de la variance

Les résultats précédents peuvent être présentés sous la forme d'une table dite table d'analyse de la variance :

Source de variation	SC	ddl	CM	F
Entre groupes (factorielle)	SC_F	k-1	$s^2 = \frac{SC_F}{k-1}$	$F_{obs} = \frac{s^2}{s_R^2}$
Résiduelle	SC_R	$n-k$	$s_R^2 = \frac{SC_R}{n-k}$	
Totale	SC_T	n-1		

Pour notre exemple, le tableau d'analyse de la variance est :

Source de variation	SC	ddl	CM	F
Variété	1264.125	3	421.375	4.64
Résiduelle	1815	20	90.75	
Totale	3079.125	23		