Chapitre II: Statistique inférentielle

Introduction Générale

Hasard . . . Probabilité . . . Modélisation . . .

- Concepts fondateurs de la théorie des probabilités difficiles et délicats à manier : les mathématiques s'y mélangent à la philosophie, aux problèmes de société, à la théologie parfois, et surtout à la vie quotidienne, de laquelle ils sont tirés et où ils prennent leur sens.
- Confusion de langage (chance/probabilité, loi des séries/loi des grands nombres, prédire/prévoir, impossible, probable, . . .)
- Connotation plus ou moins magique ou irrationnelle, (le hasard fait bien les choses, je n'ai jamais de chance, . . .) Difficulté de penser le collectif en lieu et place de l'individu

Introduction Générale

Le Hasard

"Le hasard n'est que le nom donné à notre ignorance". ***Emile Borel***, Le hasard, 1938.

Introduction Générale

Histoire ancienne

- La notion de probabilité, dans sa forme la plus simple, remonte à l'origine des jeux de hasard. On joue aux dés depuis des milliers d'années.
- Les cartes à jouer étaient déjà anciennes en Asie et au Moyen Orient lorsqu'elles apparurent en Europe au 14e siècle. De nombreux jeux, plus ou moins complexes, utilisent les cartes ou les dés et établir des stratégies pour ces jeux exigeait de se questionner sur les chances de chacun de gagner, ou sur la probabilité de certains événements.
- La notion de probabilité restait aussi rudimentaire au début, il suffisait de savoir quelles sont les chances de tirer un double six ou encore de piger une carte de pique.

Introduction Générale

Historique

L'homme a toujours pris des risques (régulés) pour gagner de l'argent et le faire fructifier :

- 1750 av JC chez les mésopotamiens : Le Code d'Hammurabi prévoit une régulation des taux autorisés pour les prêts, avec un maximum de 20 % pour les prêts d'argent !
- 4ieme siècle avant JC, chez les grecs : Les intérêts des prêts à la grosse aventure sont généralement limités à 10-12% pour un aller simple et 20-30% pour un aller-retour !
- 2nd siècle ap JC chez les romains : La table d'Ulpianus spécifiait les valeurs des rentes à vies...

Introduction Générale

Probabilité

- Le véritable début de la théorie des probabilités date de la correspondance entre Pierre de Fermat et Blaise Pascal en 1654.
- En XVIII-XIX siècle, beaucoup de scientifiques de tous ordre ont apporté leur contribution au développement de cette science PASCAL, HUYGENS, BERNOULLI, MOIVRE, LAPLACE, GAUSS, MENDEL, PEARSON, FISCHER etc.

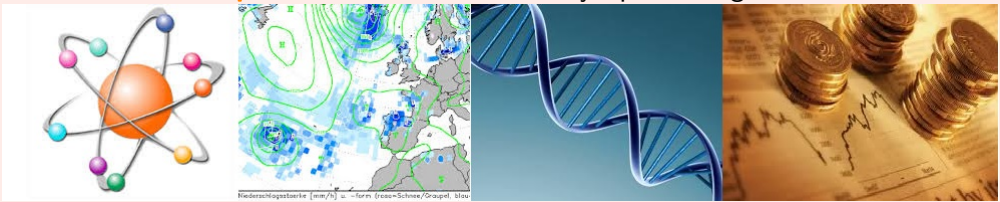
Le terme **Probabilité** a été utilisé au Moyen Âge en jurisprudence. Il est issu du latin « probare » qui signifie "prouver" ainsi désigne l'appréciation des éléments de preuves lors d'un jugement tels que les preuves, les indices ou les témoignages. En 1361, le calcul des probabilités est alors une « science dont le but est de déterminer la vraisemblance d'un évènement ». Une opinion est alors probable si elle « a une apparence de vérité ».

Le mathématicien Marin Mersenne utilise ces termes en 1637.

Introduction Générale

Divers domaines d'application

- **Modélisation de phénomènes aléatoires** Physique, biologie, finances...



- **Economie, assurance, finance** : particulièrement utilisée en mathématiques financières notamment avec le modèle Black-Scholes pour les prix d'options d'achats ou de ventes à des dates et prix donnés.
- **Sciences de l'information :** Claude Shannon et Norbert Wiener définissent, indépendamment, en 1948 une nouvelle définition de la quantité d'information dans le domaine de l'ingénierie des communications grâce à des méthodes probabilistes.

Notions de Probabilités

Historiquement, la notion de probabilité s'est dégagée à partir d'exemples simples empruntés aux jeux de hasard (le mot hasard vient de l'arabe *az-zahr* : le dé).
Nous allons introduire cette notion en l'associant à un exemple : le jeu de dé.

- Une "**expérience aléatoire**" ou "**épreuve aléatoire**" est une expérience due au hasard, c'est à dire dont on ne peut pas prévoir à l'avance le résultat, mais dont on connaît toutes les issues possibles (Exemple : L'expérience est le jet d'un dé cubique ordinaire. Le résultat de l'expérience est le nombre indiqué sur la face supérieure du dé).
- Les résultats d'une telle expérience sont appelés "**éventualités**" ou "**événements élémentaires**" ou "**issues**"

Notions de Probabilités

- L'ensemble des éventualités est appelé "**univers**" et est souvent noté U ou Ω (Exemple : $\Omega = \{1, 2, 3, 4, 5, 6\}$).
- Un **événement** est une partie de Ω, c'est-à-dire un **sous ensemble** de l'univers, ou encore un ensemble d'éventualités (Exemple : L'événement *obtenir un nombre pair* est le sous-ensemble $A = \{2, 4, 6\}$ de Ω).
- On dit que l'événement A est réalisé si le résultat de l'expérience appartient à A (Exemple : Si la face supérieure du dé indique 5, A n'est pas réalisé. Si elle indique 4, A est réalisé.)
- Si un événement ne contient qu'un seul élément, on dit que c'est un événement élémentaire (Exemple : $B = \{1\}$ est un des 6 événements élémentaires de Ω).

Langage des événements

Soit A et B deux événements liés à une expérience aléatoire dont l'univers est noté Ω.

- L'**événement contraire de A dans** Ω est l'événement qui contient les éléments de Ω qui ne sont pas dans A. C'est le **complémentaire de A dans** Ω et il est noté $\overline{A}$.
- L'**événement «A et B»** est l'événement qui contient tous les éléments de Ω qui sont à la fois dans A et B. Cet événement est noté $A \cap B$.
- L'**événement «A ou B»** est l'événement qui contient tous les éléments de Ω qui sont soit dans A soit dans B. Cet événement est noté $A \cup B$.
- On dit que les événements A et B sont **incompatibles** ou **disjoints** lorsqu'ils n'ont pas d'éléments en commun, c'est à dire lorsque $A \cap B = \emptyset$.

Probabilité d'un événement aléatoire

Définition 1

Si l'univers Ω est constitué de n événements élémentaires $\{e_i\}$, une *mesure de probabilité* sur Ω consiste à se donner n nombres $P_i \in [0,1]$, les probabilités des événements élémentaires, tels que

$$\sum_{i=1}^{n} P_i = 1.$$

Si l'événement A est la réunion disjointe de k événements élémentaires $\{e_i\}$, avec $0 < k < n$, la probabilité de A vaut, par définition,

$$P(A) = P(\bigcup_{i=1}^{k} \{e_i\}) = \sum_{i=1}^{k} P(e_i) = \sum_{i=1}^{k} P_i.$$

Par suite, $0 \leqslant P(A) \leqslant 1$.

Cardinalité

.

Cardinal

Le cardinal d'un ensemble fini E, noté $Card(E)$, est le nombre d'éléments de E.
L'ensemble des parties de l'ensemble E est notée $\mathcal{P}(E)$.

Example 1.

Exemple 1 Si $E = \{a, b, c\}$, nous avons $Card(E) = 3$ et

$$\mathcal{P}(E) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Cardinalité

Propriétés des cardinaux :

• Soit E un ensemble fini. Toute partie A de E est finie et $Card(A) \leqslant Card(E)$. Une partie A de E est égale à E si et seulement si le cardinal de A est égal à celui de E.

• Soient A et B deux parties d'un ensemble fini E et A^c le complémentaire de A dans E.

1. $Card(A \cup B) = Card(A) + Card(B) - Card(A \cap B)$.
2. $Card(A \backslash B) = Card(A) - Card(A \cap B)$.
3. $Card(A^c) = Card(E) - Card(A)$.
4. Si A et B sont disjointes, alors $Card(A \cup B) = Card(A) + Card(B)$.

• Soient E et F deux ensembles finis. Nous avons :

1. $Card(E \times F) = Card(E) \times Card(F)$.
2. $Card(\mathcal{P}(E)) = 2^{Card(E)}$.

Probabilité d'un événement aléatoire

La signification concrète de la probabilité d'un événement A est la suivante. Dans une expérience aléatoire, plus $P(A)$ est proche de 1, plus A a de chances d'être réalisé ; plus $P(A)$ est proche de 0, moins il a de chances d'être réalisé.

Probabilité *uniforme* ou *équiprobabilité* :

Tous les P_i valent $1/n$. La probabilité d'un sous-ensemble à k éléments vaut alors

$$P(A) = \frac{k}{n} = \frac{\mathrm{Card}(A)}{\mathrm{Card}(\Omega)}$$

.

On exprime aussi cette propriété par la formule

$$P(A) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}.$$

Probabilité d'un événement aléatoire

Propriétés :

Si A et B sont incompatibles, i.e., si leur intersection $A \cap B$ est vide, alors

$$P(A \cup B) = P(A) + P(B).$$

On appelle $\emptyset$ l'évènement impossible, puisqu'il n'est jamais réalisé. Sa probabilité vaut $P(\emptyset) = 0$.
On note $\bar{A}$ l'évènement contraire de A. C'est le complémentaire de A dans Ω. Sa probabilité vaut

$$P(\bar{A}) = 1 - P(A).$$

Proposition 2 ***(Théorème des probabilités totales)***.

Si A et B sont deux sous-ensembles de Ω,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Preuve :
$P(A) = P(A \setminus B) + P(A \cap B)$ car $A \setminus B$ et $A \cap B$ sont incompatibles. De même $P(B) = P(B \setminus A) + P(A \cap B)$ car $B \setminus A$ et $A \cap B$ sont incompatibles. De plus, $P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B)$, car $A \setminus B$, $B \setminus A$ et $A \cap B$ sont incompatibles. En additionnant, il vient $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probabilité d'un événement aléatoire

Proposition 2 ***(Théorème des probabilités totales)***.

Si A et B sont deux sous-ensembles de Ω,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Probabilité conjointe

La probabilité que deux événements A et B se réalisent est appelée probabilité conjointe de A et B, notée $P(A \cap B)$ et s'énonçant probabilité de A et B. Le calcul de cette probabilité s'effectue de manière différente selon que A et B sont dépendants ou indépendants, c'est-à-dire selon que la réalisation de l'un influence ou non celle de l'autre.

Événements indépendants

Je lance un dé rouge et un dé vert et je cherche la probabilité d'obtenir un total de 2. Je dois donc obtenir 1 avec chacun des deux dés. La probabilité d'obtenir 1 avec le dé rouge est $1/6$ et demeurera $1/6$ quelque soit le résultat du dé vert. Les deux événements "obtenir 1 avec le dé rouge" et "obtenir 1 avec le dé vert" sont indépendants.

Probabilité conjointe

Proposition

Si deux événements sont indépendants, la probabilité qu'ils se réalisent tous les deux est égale au produit de leurs probabilités respectives. On peut donc écrire :

$$P(A \cap B) = P(A) \times P(B).$$

Dans notre exemple : $P(\text{total} = 2) = P(\text{dé vert} = 1) \times P(\text{dé rouge} = 1) = 1/36$.

Remarque

Les tirages avec remise constituent une bonne illustration d'événements indépendants.

Probabilitée conditionnelle-Indépendance

Si deux événements sont dépendants plutôt qu'indépendants, comment calculer la probabilité que les deux se réalisent, puisque la probabilité de réalisation de l'un dépend de la réalisation de l'autre ? Il nous faut connaître pour cela le degré de dépendance des deux événements qui est indiqué par la notion de probabilité conditionnelle.

Définition 3

Soient A et B deux événements, A étant supposé de probabilité non nulle. On appelle *probabilité conditionnelle* de B par rapport à A, la probabilité de réalisation de l'événement B sachant que A est réalisé. On la note

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

$P(B|A)$ se lit *p de B si A* ou *p de B sachant A*.

Théorème 4 *(Théorème des probabilités composées ou règle de la multiplication).*

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B).$$

En voici une généralisation. Soit $A_1, \ldots, A_k$ un système complet d'évènements. Alors

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(A_j)P(B|A_j).$$

Théorème 5 *(Formule de Bayes)*

Soit $A_1, \ldots, A_k$ un système complet d'évènements. Soit E un évènement de probabilité non nulle. Alors

$$P(A_j|E) = \frac{P(A_j \cap E)}{P(E)} = \frac{P(A_j)P(E|A_j)}{\sum_{i=1}^{k} P(A_i)P(E|A_i)}.$$

Probabilitée conditionnelle-Indépendance

Exercice 1

L'entreprise possède actuellement deux chaines de production, l'une pour des drones à deux hélices et l'autre pour des drones à quatre hélices. Il arrive que les batteries des drones fabriqués aient un défaut et dans ce cas, on dira que les drones sont défectueux. On souhaite avoir une idée du pourcentage de drones défectueux sur l'ensemble de la production on prélève 500 drones dans la production de l'entreprise et on obtient les résultats suivants :

1. 300 drones possèdent deux hélices
2. Parmi les drones à deux hélices, 2% sont défectueux
3. Parmi les drones à quartes hélices, 96% ne présentent aucun défaut.

Probabilitée conditionnelle-Indépendance

Exercice 1

Un drone est choisi au hasard parmi les 500 drones prélevés. On considère les événements suivants :
H : "le drone possède deux hélices".
D : "Le drone est défectueux".

- Donner la valeur des probabilités $P(A)$, $P_A(\bar{D})$, $P_{\bar{A}}(\bar{D})$.
- Représenter la situation par un arbre pondéré de probabilité.
- Calculer la probabilité que la drone possède deux hélices et soit défectueuse.
- Montrer que la probabilité qu'un drone pris au hasard soit défectueuse et égale à $0,028$.

Probabilitée conditionnelle-Indépendance

Exercice 2

Un fabriquant de téléphone portables se fournit en microprocesseurs au près de deux entreprises A et B, l'entreprise A fournit 55% des microprocesseurs, le reste étant fourni par l'entreprise B, il s'avère que 1% des microprocesseurs provenant de l'entreprise A et 1,5% des microprocesseurs provenant de l'entreprise B sont défectueuses.
On prélève au hasard un microprocesseurs dans le stock du fabricant. Tous les microprocesseurs ont la même probabilité d'être prèlever.
On considère les événements suivants :
A "Le microprocesseur provient de l'entreprise A "
D "Le microprocesseur est défectueuse "

1. Déduire d'après les informations figurant dans l'énoncé les probabilités $P(A)$ et $P_A(D)$.
2. Représenter la situation par un arbre de probabilité pondéré.
3. Calculer $P(A \cap D)$ et $P(\bar{A} \cap D)$.

Probabilitée conditionnelle-Indépendance

Exercice 2

1. Justifier que la probabilité de prélever un microprocesseur défectueuse est 0,01225.
2. Calculer la probabilité que le microprocesseur provienne de l'entreprise B sachant qu'il est défectueuse, arandi le résultat à 10^{-3}.

Exercice 3 *(Formule de Bayes)*

Une entreprise utilise trois types d'ampoules électriques notés T_1, T_2 et T_3 dans les proportions 60% , 30% , 10%. Les probabilités de bon fonctionnement de ces trois types pour un temps donné s'élèvent à 0,9, 0,8 et 0,5 respectivement. Quelle est la probabilité qu'une ampoule tombée en panne soit du type T_1 ?

Solution Si on introduit les événements
$B = \{$une ampoule choisie au hasard tombe en panne $\}$
$A_k = \{$ une ampoule est du type $T_k\}$, on a

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{3} P(B|A_i)P(A_i)} = \frac{0,1 \times 0,6}{0,1 \times 0,6 + 0,2 \times 0,3 + 0,5 \times 0,1} = \frac{6}{17}.$$

Exercice 2 *(Formule de Bayes)*

Une urne contient 5 boules noires et 3 boules blanches. Quelle est la probabilité d'extraire 2 boules blanches en 2 tirages ?

Solution : Tirage sans remise

Appelons B_1, l'événement : obtenir une boule blanche au premier tirage.
Appelons B_2, l'événement : obtenir une boule blanche au deuxième tirage.
La probabilité cherchée $P(B_1 \cap B_2)$ est égale à $P(B_1) \times P(B_2|B_1)$. Or $P(B_1)$ vaut $3/8$ et $P(B_2|B_1)$ est égale à $2/7$ puisque lorsqu'une boule blanche est sortie au premier tirage, il ne reste plus que 7 boules au total, dont 2 seulement sont blanches. On conclut que $P(B_1 \cap B_2) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$.

Variables aléatoires

Exemple 1

On jette deux fois une pièce de monnaie non truquée, et on s'intéresse au nombre de fois que le côté “face" a été obtenu. Pour calculer les probabilités des divers résultats, on introduira une variable X qui désignera le nombre de “face" obtenu. X peut prendre les valeurs 0,1,2.

Exemple 2

On lance une fléchette vers une cible circulaire de rayon égal à 50 cm et on s'intéresse à la distance entre la fléchette et le centre de la cible. On introduira ici une variable X, distance entre l'impact et le centre de la cible, qui peut prendre n'importe quelle valeur entre 0 et 50.

Variables aléatoires

Dans ces deux cas, X prend des valeurs réelles qui dépendent du résultat de l'expérience aléatoire. Les valeurs prises par X sont donc aléatoires. X est appelée variable aléatoire.

Définition 1

Soit un univers Ω associé à une expérience aléatoire, sur lequel on a défini une mesure de probabilité. Une *variable aléatoire* X est une application de l'ensemble des événements élémentaires de l'univers Ω vers $\mathbb{R}$ (vérifiant quelques conditions mathématiques non explicitées ici).

Une variable aléatoire est une variable (en fait une fonction !) qui associe des valeurs numériques à des événements aléatoires.

Variables aléatoires

Par convention, une variable aléatoire sera représentée par une lettre majuscule X alors que les valeurs particulières qu'elle peut prendre seront désignées par des lettres minuscules $x_1, x_2, \ldots, x_i, \ldots, x_n$.
Les deux variables aléatoires définies dans les exemples 1 et 2 sont de natures différentes. La première est discrète, la seconde continue.

Variables aléatoires discrètes

Définition 2

Une *variable aléatoire discrète* est une variable aléatoire qui ne prend que des valeurs entières, en nombre fini ou dénombrable.

Pour apprécier pleinement une variable aléatoire, il est important de connaître quelles valeurs reviennent le plus fréquemment et quelles sont celles qui apparaissent plus rarement. Plus précisément, on cherche les probabilités associées aux différentes valeurs de la variable

Variables aléatoires discrètes

Définition 3

Associer à chacune des valeurs possibles de la variable aléatoire la probabilité qui lui correspond, c'est définir la ***loi de probabilité*** ou la *distribution de probabilité* de la variable aléatoire.
Pour calculer la probabilité que la variable X soit égale à x, valeur possible pour X, on cherche tous les événements élémentaires e_i pour lesquels $X(e_i) = x$, et on a

$$P(X = x) = \sum_{i=1}^{k} P(\{e_i\}),$$

si $X = x$ sur les événements élémentaires $e_1, e_2, \ldots, e_k$.

La ***fonction de densité discrète*** f est la fonction de $\mathbb{R}$ dans $[0, 1]$, qui à tout nombre réel x_i associe $f(x_i) = P(X = x_i)$. On a bien sûr $\sum_i f(x_i) = 1$.

Fonction de répartition

En statistique descriptive, on a introduit la notion de fréquences cumulées croissantes. Son équivalent dans la théorie des probabilités est la fonction de répartition.

Définition 4

La *fonction de répartition* d'une variable aléatoire X indique pour chaque valeur réelle x la probabilité que X prenne une valeur au plus égale à x. C'est la somme des probabilités des valeurs de X jusqu'à x. On la note F.

$$\forall x \in \mathbb{R}, \quad F(x) = P(X \leqslant x) = \sum_{x_i \leqslant x} P(X = x_i).$$

La fonction de répartition est toujours croissante, comprise entre 0 et 1 et se révélera un instrument très utile dans les travaux théoriques.

Espérance mathématique d'une distribution de probabilité

Définition 5

1. Soit X une variable aléatoire discrète qui prend un nombre fini de valeurs $x_1, x_2, \ldots, x_n$ et dont la loi de probabilité est $f : f(x_i) = P(X = x_i)$. L'*espérance mathématique* de X, notée $E(X)$, est définie par

$$E(X) = \sum_{i=1}^{n} x_i f(x_i).$$

Variance d'une distribution de probabilités

Définition 6

- On appelle *variance* de la variable aléatoire X la valeur moyenne des carrés des écarts à la moyenne,

$$Var(X) = E\left((X - E(X))^2\right).$$

Le calcul de la variance se simplifie en utilisant l'expression :

$$Var(X) = E(X^2) - E(X)^2.$$

Fonction de répartition

Définition 7

- On appelle *écart-type* de la variable aléatoire X la racine carrée de sa variance.

$$\sigma(X) = \sqrt{Var(X)}.$$

Dans le cas d'une variable aléatoire discrète finie,

$$Var(X) = \sum_{i=1}^{n} (x_i - E(X))^2 f(x_i) = \left(\sum_{i=1}^{n} x_i^2 f(x_i) \right) - E(X)^2.$$

Variables aléatoires discrètes

Exemple 1

On lance 2 fois une pièce de monnaie équilibrée. Soit X une variable aléatoire discrète telle que $X =$ nombre de côtés "face" qui peut prendre les valeurs 0,1,2.

1. Déterminer la loi de probabilité de la variable X.

Solution

La variable $X =$ nombre de côtés "face" peut prendre les valeurs 0,1,2.

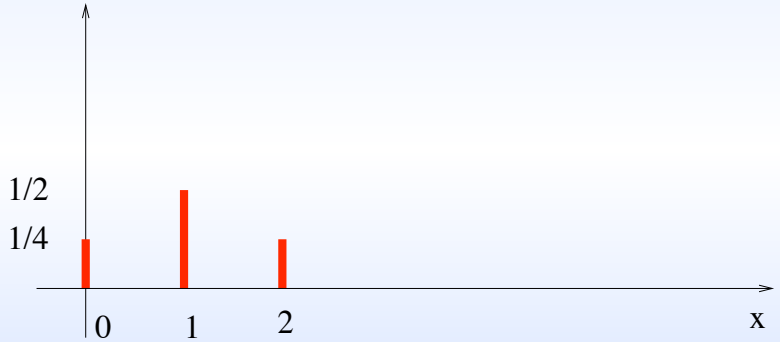
$$
\begin{aligned}
f(0) &= P(X=0) = P((\text{pile},\text{pile})) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4};\\
f(1) &= P(X=1) = P((\text{pile},\text{face})) + P((\text{face},\text{pile})) = \frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2} = \frac{1}{2};\\
f(2) &= P(X=2) = P((\text{face},\text{face})) = \frac{1}{2}\times\frac{1}{2} = \frac{1}{4};\\
f(x) &= 0 \text{ si } x \notin \{0,1,2\}.
\end{aligned}
$$

On présente sa distribution de probabilité dans un tableau.

x	0	1	2	total
$f(x) = P(X=x)$	1/4	1/2	1/4	1

Représentation graphique de la distribution de probabilité

Elle s'effectue à l'aide d'un diagramme en bâtons où l'on porte en abscisses les valeurs prises par la variable aléatoire et en ordonnées les valeurs des probabilités correspondantes.
Dans l'exemple du jet de pièces :



Variables aléatoires discrètes

Exemple 2

On lance 3 fois une pièce de monnaie équilibrée. Soit X une variable aléatoire discrète telle que $X =$ nombre de côtés "face" qui peut prendre les valeurs 0,1,2,3

1. Déterminer la loi de probabilité de la variable X.
2. Calculer sa fonction de répartition
3. calculer l'espérance, la variance et l'écart-type

Solution

1- La variable X = nombre de côtés "face" peut prendre les valeurs 0,1,2,3

$$
\begin{aligned}
f(0) &= P(X=0) = P((\text{pile},\text{pile},\text{pile})) = \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{8};\\
f(1) &= P(X=1) = P((\text{pile},\text{pile},\text{face})) + P((\text{face},\text{pile},\text{pile})) + P((\text{pile},\text{face},\text{pile}))\\
&= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{8}+\frac{1}{8}+\frac{1}{8} = \frac{3}{8};\\
f(2) &= P(X=2) = P((\text{face},\text{face},\text{pile})) + P((\text{pile},\text{face},\text{face})) + P((\text{face},\text{pile},\text{face}))\\
&= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{8}+\frac{1}{8}+\frac{1}{8} = \frac{3}{8};\\
f(3) &= P(X=3) = P((\text{face},\text{face},\text{face})) = \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2} = \frac{1}{8};\\
f(x) &= 0 \text{ si } x \notin \{0,1,2,3\}.
\end{aligned}
$$

Solution

2- On présente sa distribution de probabilité dans un tableau.

x	0	1	2	3	total
$f(x) = P(X = x)$	$1/8$	$3/8$	$3/8$	$1/8$	1
$F(x) = P(X \leqslant x)$	$1/8$	$4/8$	$7/8$	$8/8$	

3-

x	0	1	2	3	total
$f(x)$	$1/8$	$3/8$	$3/8$	$1/8$	1
$P_i x_i$	0	$3/8$	$6/8$	$3/8$	$12/8$
$P_i x_i^2$	0	$3/8$	$12/8$	$9/8$	$24/8$

Variables aléatoires discètes

Solution

Donc $E(X) = \sum_{i=1}^{n} p_i x_i = 3/2 = 1,5$
$Var(X) = (\sum_{i=1}^{n} p_i x_i^2) - E(X)^2 = 3 - 9/4 = 3/4.$
$\sigma(X) = \sqrt{Var(X)} = \sqrt{3/4}$

Fonction de répartition

Exemple 3

Loi de probabilité de la v.a. discrète finie X égale à la somme des points marqués lors du lancer de deux dés non truqués :

x_i	2	3	4	5	6	7	8	9	10	11	12	total
$f(x_i)$	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36	1

On calcule aisément la fonction de répartition à partir de la connaissance des couples $(x_i; f(x_i))$,

$$F(x) = \begin{cases} 0 & \text{si } x < x_1, \\ \displaystyle\sum_{j=1}^{i} f(x_i) & \text{pour } 1 \leqslant i \leqslant k-1, \\ 1 & \text{si } x \geqslant x_k. \end{cases}$$

Solution

1- La variable $X =$ égale à la somme des points marqués lors du lancer de deux dés non truqués peut prendre les valeurs 2,3,4,5,6,7,8,9,10,11,12

$$
\begin{aligned}
f(2) &= P(X=2) = P((1,1)) = \frac{1}{6}\times\frac{1}{6} = \frac{1}{36};\\
f(3) &= P(X=3) = P((1,2)) + P((2,1)) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} = \frac{1}{36} + \frac{1}{36} = \frac{1}{18};\\
f(4) &= P(X=4) = P((2,2)) + P((3,1)) + P((1,3)) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6}\\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12};\\
f(5) &= P(X=5) = P((1,4)) + P((4,1)) + P((2,3)) + P((3,2))\\
&= \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9};
\end{aligned}
$$

Solution

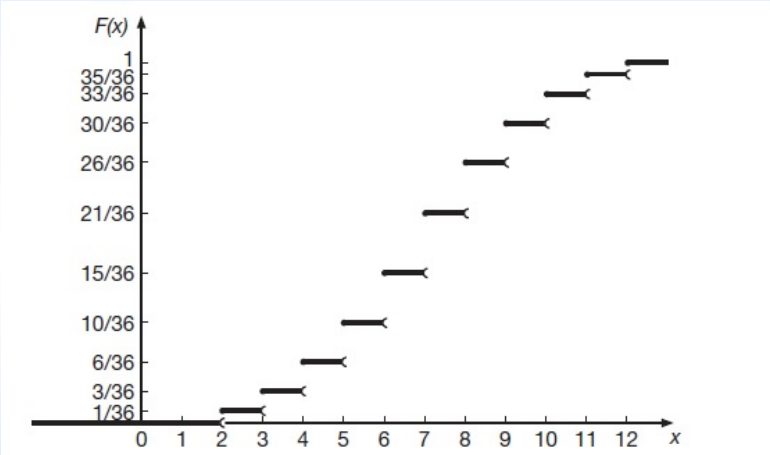
$$
\begin{aligned}
f(6) &= P(X=6) = P((1,5)) + P((5,1)) + P((2,4)) + P((4,2)) + P((3,3)) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}; \\
f(7) &= P(X=7) = P((1,6)) + P((6,1)) + P((5,2)) + P((2,5)) + P((3,4)) + P((4,3)) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{6}; \\
f(8) &= P(X=8) = P((6,2)) + P((2,6)) + P((3,5)) + P((5,3)) + P((4,4)) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{5}{36}; \\
f(9) &= P(X=9) = P((5,4)) + P((5,4)) + P((6,3)) + P((3,6)) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9};
\end{aligned}
$$

Solution

$$
\begin{aligned}
f(10) &= P(X=10) = P((5,5)) + P((6,4)) + P((4,6)) \\
&= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}; \\
f(11) &= P(X=11) = P((6,5)) + P((5,6)) \\
&= \frac{1}{36} + \frac{1}{36} = \frac{1}{18}; \\
f(12) &= P(X=8) = P(6,6) = \frac{1}{36}
\end{aligned}
$$

Fonction de répartition

Fonction de répartition



Variables aléatoires continues

Définition

Une variable aléatoire est dite ***continue*** si elle peut prendre toutes les valeurs d'un intervalle fini ou infini.

Fonction de densité de probabilité

La **fonction de densité de probabilité** f pour une variable aléatoire continue. Elle a les propriétés suivantes :

1. f **est une fonction toujours positive**.
2. $P(a \leqslant X \leqslant b) = \int_a^b f(x)\,dx$,
3. $\int_{\mathbb{R}} f(x)\,dx = 1$.

Variables aléatoires continues

Fonction de Répartition

De même que pour les variables aléatoires discrètes, on peut définir la fonction de répartition F de la variable continue X qui permet de connaître la probabilité que X soit inférieure à une valeur donnée :

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)\, dt.$$

Proposition

1. F est continue et croissante sur $\mathbb{R}$.
2. $\forall x \in \mathbb{R}, \quad F'(x) = f(x)$.
3. $\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1$.
4. $P(a \leqslant X \leqslant b) = P(X \leqslant b) - P(X \leqslant a) = F(b) - F(a)$.
5. $P(X > x) = 1 - F(x)$.
6. $P(a \leqslant X \leqslant b) = P(a < X \leqslant b) = P(a \leqslant X < b) = P(a < X < b) = F(b) - F(a)$.

Espérance mathématique d'une distribution de probabilité

Définition

1. Si la variable aléatoire X est continue et a pour fonction de densité de probabilité f, *son espérance mathématique* est

$$E(X) = \int_{\mathbb{R}} xf(x)\,dx,$$

pourvu que la fonction $x \mapsto xf(x)$ soit intégrable sur $\mathbb{R}$.

Variance d'une distribution de probabilités

Définition

- Dans le cas d'une variable aléatoire continue,

$$Var(X) = \int_{\mathbb{R}} (x - E(X))^2 f(x)\, dx = \left(\int_{\mathbb{R}} x^2 f(x)\, dx \right) - E(X)^2.$$

- On appelle *écart-type* de la variable aléatoire X la racine carrée de sa variance.

$$\sigma(X) = \sqrt{Var(X)}.$$

Variables aléatoires continues

Exemple 1

Soit f la fonction définie sur $\mathbb{R}$ par $f(x) = ke^{-x}$ si $x \geqslant 0$, $f(x) = 0$ sinon.

1. Déterminer k pour que f soit la fonction de densité de probabilité d'une variable aléatoire X.
2. Déterminer la fonction de répartition de la variable X.
3. Calculer $P(1 < X < 2)$.

Solution

1. f doit être une fonction positive, donc il nous faut impérativement trouver pour k une valeur positive. Une fonction de densité de probabilité doit vérifier $\int_{\mathbb{R}} f(x)\,dx = 1$, donc $\int_0^{+\infty} ke^{-x}\,dx = 1$. Il en résulte que $k = 1$.
2. Par définition la fonction de répartition de X est la fonction F définie par

$$F(x) = \begin{cases} \displaystyle\int_0^x e^{-t}\,dt = 1 - e^{-x} & \text{si } x > 0, \\ 0 & \text{sinon.} \end{cases}$$

3.

$$P(1 < X < 2) = \int_1^2 e^{-x}\,dx = e^{-1} - e^{-2} \sim 0.23.$$

Exemple 2

1. $f(x) = hx^2$ avec $x \in [0,4]$

   Pour que f soit une densité il faut que $\int_0^4 f(x)\,dx = 1$

   On a $\int_0^4 f(x)\,dx = \int_0^4 hx^2\,dx = h\int_0^4 x^2\,dx = 1 \Rightarrow h\Big[\frac{x^3}{3}\Big]_0^4 = 1,$

   ce qui donne $h\Big(\frac{4^3}{3} - 0^3/3\Big) = 1 \Rightarrow h = 3/64.$

   Donc $f(x) = \frac{3}{64}x^2$ est une densitée de probabilité pour $x \in [0,4]$.

2. $E(x) = \int_0^4 xf(x)dx = \int_0^4 x\frac{3}{64}x^2dx = \frac{3}{64}\int_0^4 x^3dx = 3/64\Big[\frac{x^4}{4}\Big]_0^4 =$

   $\frac{3}{64}\Big(\frac{4^4}{4} - 0^4/4\Big) = \frac{3}{64} \times 64 = 3.$

Variables aléatoires continues

Exemple

1. $V(x) = \int_0^4 (x - E(x))^2 f(x)\, dx = \int_0^4 x^2 f(x)\, dx - E(x)^2 = \int_0^4 x^2 \cdot \frac{3}{64} x^2\, dx - E(x)^2 = \frac{3}{64} \int_0^4 x^4\, dx - E(x)^2 = \frac{3}{64} \left[\frac{x^5}{5}\right]_0^4 - E(x)^2 = \frac{3}{64} \frac{4^5}{5} - 3^2 = \frac{48}{5} - 9 = 0.6.$
2. $\sigma(X) = \sqrt{Var(X)} = \sqrt{0.6} = 0.77$

Loi d'une fonction de variable aléatoire

Si φ est une fonction définie sur $\mathbb{R}$ à valeurs dans $\mathbb{R}$, l'application $\varphi \circ X$, notée $Y = \varphi(X)$ est une variable aléatoire dont on peut déterminer la fonction de répartition et donc la loi de probabilité à partir de celle de X.

1) Changement de variable $Y = aX + b$.

Les paramètres a ($a \neq 0$) et b sont des nombres réels. Connaissant la fonction de répartition de X, on peut calculer la fonction de répartition F_Y de la v.a. Y :

- Pour $a > 0$:

$$F_Y(y) = P(Y \leqslant y) = P(aX + b \leqslant y) = P\Big(X \leqslant \frac{y-b}{a}\Big) = F_X\Big(\frac{y-b}{a}\Big).$$

- Pour $a < 0$:

$$F_Y(y) = P(Y \leqslant y) = P\Big(X \geqslant \frac{y-b}{a}\Big) = \begin{cases} 1 - F_X\Big(\frac{y-b}{a}\Big) & \text{si } X \text{ est une V.A.C,} \\ 1 - P\Big(X < \frac{y-b}{a}\Big) & \text{si } X \text{ est une V.A.D.} \end{cases}$$

Lorsque la variable aléatoire X est continue, on obtient la fonction de densité f_Y par dérivation de la fonction F_Y.

Loi d'une fonction de variable aléatoire

2) Autres types de fonctions ($Y = \varphi(X)$).

• Si φ est bijective (donc monotone),

φ croissante : $F_Y(y) = P(Y \leqslant y) = P(X \leqslant \varphi^{-1}(y)) = F_X(\varphi^{-1}(y))$

φ décroissante :

$$F_Y(y) = P(Y \leqslant y) = P(X \geqslant \varphi^{-1}(y)) = \begin{cases} 1 - F_X(\varphi^{-1}(y)) & \text{si } X \text{ est une v.a.c ,} \\ 1 - P\Big(X < \varphi^{-1}(y)\Big) & \text{si } X \text{ est une v.a.d.} \end{cases}$$

Si X est une v.a. continue et si la fonction φ est dérivable, on obtient la fonction de densité f_Y par dérivation de la fonction F_Y.

Loi d'une fonction de variable aléatoire

Exemple :

Soit une v.a. continue X, on peut calculer les fonctions de répartition et de densité de $Y = exp(X)$, la fonction exponentielle étant croissante :

$$F_Y(y) = \begin{cases} 0 & \text{si } y < 0, \\ F_X(\ln(y)) & \text{pour } y > 0. \end{cases} \Rightarrow f_Y(y) = \begin{cases} 0 & \text{si } y < 0, \\ \frac{1}{y} f_X(\ln(y)) & \text{pour } y > 0. \end{cases}$$

• φ **quelconque**

Le principe consiste toujours à identifier la fonction de répartition F_Y en recherchant l'antécédent pour X de l'événement $\{Y \leqslant y = \varphi(x)\}$.

Par exemple, pour $Y = X^2$:

$$F_Y(y) = \begin{cases} 0 \quad \text{si } y < 0, , \\ P(-\sqrt{y} \leqslant X \leqslant +\sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \quad \text{pour } y \geqslant 0. \end{cases}$$

Propriétés de l'espérance mathématique et de la variance

Propriétés :

Changement d'origine	Changement d'échelle	Transformation affine
$E(X+c) = E(X) + c$	$E(aX) = aE(X)$	$E(aX+c) = aE(X) + c$
$Var(X+c) = Var(X)$	$Var(aX) = a^2 Var(X)$	$Var(aX+c) = a^2 Var(X)$
$\sigma(X+c) = \sigma(X)$	$\sigma(aX) = \lvert a\rvert\sigma(X)$	$\sigma(aX+c) = \lvert a\rvert\sigma(X)$

Définition :

- Une variable aléatoire X est dite *centrée* si son espérance mathématique est nulle.
- Une variable aléatoire X est dite *réduite* si son écart-type est égal à 1.
- Une variable aléatoire centrée réduite est dite *standardisée*.

A n'importe quelle variable aléatoire X, on peut associer la variable standardisée

$$Z = \frac{X - E(X)}{\sigma(X)}.$$

Principales distributions de probabilités

De nombreuses situations pratiques peuvent être modélisées à l'aide de variables aléatoires qui sont régies par des lois spécifiques. Il importe donc d'étudier ces modèles probabilistes qui pourront nous permettre par la suite d'analyser les fluctuations de certains phénomènes en évaluant, par exemple, les probabilités que tel événement ou tel résultat soit observé.

La connaissance de ces lois théoriques possède plusieurs avantages sur le plan pratique :

- Les observations d'un phénomène particulier peuvent être remplacées par l'expression analytique de la loi où figure un nombre restreint de paramètres (1 ou 2, rarement plus).
- La loi théorique agit comme modèle (idéalisation) et permet ainsi de réduire les irrégularités de la distribution empirique. Ces irrégularités sont souvent inexplicables et proviennent de fluctuations d'échantillonnage, d'imprécision d'appareils de mesure ou de tout autre facteur incontrôlé ou incontrôlable.
- Des tables de probabilités ont été élaborées pour les lois les plus importantes. Elles simplifient considérablement les calculs.

Lois Discrètes

Loi uniforme discrète

Elle modélise des situations d'équiprobabilités.
On dit qu'une variable aléatoire X suit une loi uniforme discrèete lorsqu'elle prend ses valeurs dans $\{1, ..., n\}$ avec des probabilités élémentaires identiques. Puisque la somme des ces dernières doit valoir 1, on en déduit qu'elles doivent toutes être égales à $1/n$:

$$\forall k = 1, ..., n \quad P(X = k) = \frac{1}{n}.$$

Paramètres de la distribution

On calcule aisément :

$$E(X) = \sum_{k=1}^{n} \frac{k}{n} = \frac{n+1}{2}.$$

$$V(X) = E(X^2) - E(X)^2 = \sum_{k=1}^{n} \frac{k^2}{n} - \frac{(n+1)^2}{4} = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

Exemple :

Soit $X =$ résultat d'un jet de dé à six faces non-truqué.
Les $n = 6$ modalités possibles, $x_1 = 1$; $x_2 = 2$; $x_3 = 3$; $x_4 = 4$; $x_5 = 5$; $x_6 = 6$, ont toutes pour probabilité élémentaire $1/6$:

$$\forall k = 1, ..., 6 \quad P(X = k) = \frac{1}{6}.$$

$$E(X) = \frac{7}{2}, \qquad V(X) = \frac{35}{12}.$$

Loi Bernoulli

Définition :

Une variable aléatoire discrète qui ne prend que les valeurs 1 et 0 avec les probabilités respectives p et $q = 1 - p$ est appelée variable de Bernoulli.

Exemple :

Une urne contient deux boules rouges et trois boules vertes. On tire une boule de l'urne. La variable aléatoire $X =$ nombre de boules rouges tirées est une variable de Bernoulli. On a : $P(X = 1) = 2/5 = p$, $P(X = 0) = 3/5 = q$.

Loi de probabilités

x	0	1
$f(x) = P(X = x)$	q	p

Loi Bernoulli

Paramètres de la distribution :

On calcule

$$\begin{aligned} E(X) &= 0.q + 1.p = p, \\ V(X) &= E(X^2) - E(X)^2 = (0^2 q + 1^2 p) - p^2 = p - p^2 = pq, \end{aligned}$$

$E(X) = p$	$V(X) = pq$	$\sigma(X) = \sqrt{pq}$

Loi Binomiale

a) On effectue une épreuve de Bernoulli. Elle n'a donc que deux issues : le succès avec une probabilité p ou l'échec avec une probabilité q.
b) On répète n fois cette épreuve. Les n épreuves sont indépendantes entre elles, ce qui signifie que la probabilité de réalisation de l'événement "succès" est la même à chaque épreuve et est toujours égale à p. Dans cette situation, on s'intéresse à la variable $X =$ "nombre de succès au cours des n épreuves".

Distribution de probabilités

On dit que la variable aléatoire X suit une *loi binomiale de paramètres n et p*. On note $X \hookrightarrow B(n,p)$. Avec $P(X=k) = \binom{k}{n} p^k q^{n-k}$.

Remarque : L'adjectif binomial vient du fait que lorsqu'on somme toutes ces probabilités, on retrouve le développement du binôme de Newton,

$$\sum_{k=0}^{n} \binom{k}{n} p^k q^{n-k} = (p+q)^n = 1.$$

Loi Binomiale

Exemple :

Dans un exercice militaire, un soldat a le droit de tirer sur une cible mobile 10 fois, si la probabilité d'atteindre cette cible est $0,7$, quelle est la probabilité que ce soldat atteint la cible au moins 2 fois.

Cette expérience aléatoire consiste a répéter la même expérience (tirer sur une cible) 10 fois de suite. c'est donc une expérience binomiale. Soit X la variable aléatoire qui modélise cette expérience, on a $X \hookrightarrow B(10, 0,7)$ et on cherche $P(X \geqslant 2)$. Donc

$$\begin{aligned} P(X \geqslant 2) = 1 - P(X < 2) \quad &= 1 - \Big(P(X = 0) + P(X = 1)\Big) \\ &= 1 - \Big(\tbinom{0}{10}(1 - 0,7)^{10} + \tbinom{1}{10}(0,7)(1 - 0,7)^{9}\Big) = 0,856. \end{aligned}$$

Loi Binomiale

Paramètres descriptifs de la Distribution

Nous savons que $X = X_1 + \cdots + X_n$ avec $E(X_i) = p$ pour $i = 1, 2, \ldots, n$, donc $E(X) = E(X_1) + \cdots + E(X_n) = np$.
Les variables X_i sont indépendantes et $Var(X_i) = pq$ pour $i = 1, 2, \ldots, n$, donc $Var(X) = Var(X_1) + \cdots + Var(X_n) = npq$.

$E(X) = np$	$V(X) = npq$	$\sigma(X) = \sqrt{npq}$

Somme de deux variables binomiales :
Si X_1 et X_2 sont des variables *indépendantes* qui suivent des lois binomiales $B(n_1, p)$ et $B(n_2, p)$ respectivement, alors $X_1 + X_2$ suit une loi binomiale $B(n_1 + n_2, p)$.

Loi Géométrique

a) On effectue une épreuve de Bernoulli. Elle n'a donc que deux issues : le succès avec une probabilité p ou l'échec avec une probabilité $q = 1 - p$.
b) On répète l'épreuve jusqu'à l'apparition du premier succès.
c) Toutes les épreuves sont indépendantes entre elles.
Dans cette situation, on s'intéresse à la variable $X =$ "nombre de fois qu'il faut répéter l'épreuve pour obtenir le premier succès".
On est donc dans les mêmes hypothèses que pour la loi binomiale, mais le nombre d'épreuves n'est pas fixé à l'avance. On s'arrête au premier succès.

Distribution de probabilités

L'ensemble des valeurs prises par X est $1, 2, 3, \ldots$. On cherche la probabilité d'avoir recours à n épreuves pour obtenir le premier succès.

$$P(X = n) = q^{n-1}p.$$

On dit que la variable aléatoire X suit une *loi géométrique de paramètre* p. On note $X \hookrightarrow G(p)$.

Loi Géométrique

Remarque :

L'appellation *géométrique* vient du fait qu'en sommant toutes les probabilités, on obtient une série géométrique. En effet,

$$\sum_{n=1}^{+\infty} q^{n-1}p = \frac{p}{1-q} = 1.$$

Loi Géométrique

Paramètres descriptifs de la Distribution

$E(X) = 1/p$	$Var(X) = q/p^2$	$\sigma(X) = \sqrt{q}/p$

Remarque

On peut interpréter l'expression de l'espérance de façon intuitive. En effet en n épreuves, on s'attend à obtenir np succès et par conséquent, le nombre moyen d'épreuves entre deux succès devrait être $\frac{n}{np} = \frac{1}{p}$.

Loi de Poisson

Définition :

On peut considérer la loi de Poisson de paramètre λ comme la loi limite d'une loi binomiale $B(n, \lambda/n)$ lorsque n tend vers l'infini, le produit des paramètres $n.\lambda/n$ restant toujours constant égal à λ.

On écrit $X \hookrightarrow P(\lambda)$.

Proposition :

La loi de Poisson de paramètre λ est donnée par

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Loi de Poisson

Paramètres descriptifs de la distribution :

$E(X) = \lambda$	$Var(X) = \lambda$	$\sigma(X) = \sqrt{\lambda}$

Approximation de la loi binomiale par la loi de Poisson

On approche la loi $B(n, p)$ par la loi $P(np)$ dès que $n > 20$, $p \leqslant 0.1$ et $np \leqslant 5$.

RÈGLE IMPORTANTE. Lorsqu'on approche une loi par une autre, on choisit le ou les paramètres de la loi approchante de manière que l'espérance (et la variance lorsqu'on a suffisamment de paramètres) de la loi approchante soit égale à l'espérance (et la variance) de la loi approchée.

Loi de Poisson

Somme de deux lois de Poisson :

On peut considérer la loi de Poisson de paramètre λ comme la loi limite d'une loi binomiale $B(n, \lambda/n)$ lorsque n tend vers l'infini, le produit des paramètres $n.\lambda/n$ restant toujours constant égal à λ.

On écrit $X \hookrightarrow P(\lambda)$.

Si X_1 et X_2 sont des variables aléatoires *indépendantes* qui suivent des lois de Poisson de paramètres respectifs λ_1 et λ_2, alors $X_1 + X_2$ suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.

Les principales lois de probabilité : Loi de poisson

Loi de poisson :Exercices

-Toutes les études antérieures montrent qu'il existe une probabilité de $0,005$ pour qu'une personne soit atteinte d'une maladie.
-Des analyses ont été effectuées sur un échantillon de 400 personnes.
-On définit la variable X : « le nombre de personnes déclarées positives à la maladie »
TAF

1. Quelle est la loi suivie par la variable X
2. Présenter la loi de probabilité de X
3. Calculer $E(x)$ et $E(Y)$

Les principales lois de probabilité : Loi de poisson

Loi de poisson :Solution

1- La probabilité de l'événement est faible

On dit que la variable aléatoire X suit une loi de poisson paramètres m On note $X \hookrightarrow P(X = x) = \dfrac{m^x e^{-m}}{x!}$.

On a $n = 400$ et $p = 0.005$ alors $m = E(x) = np = 400 \times 0.005 = 2$

donc $X \hookrightarrow P(2)$

2- $P(X = 0) = \dfrac{2^0 e^{-2}}{0!} = 0.1353$

$P(X = 1) = \dfrac{2^1 e^{-2}}{1!} = 0,2707$

$P(X = 2) = \dfrac{2^2 e^{-2}}{2!} = 0,2707$

Les principales lois de probabilité : Loi de poisson

Loi de poisson :Solution

$$P(X=3) = \frac{2^3 e^{-2}}{3!} = 0,1804$$

$$P(X=4) = \frac{2^4 e^{-2}}{4!} = 0,0902$$

3- $\boxed{E(X) = m = 2} \quad \boxed{Var(X) = m = 2} \quad \boxed{\sigma(X) = \sqrt{2}}$

Les principales lois de probabilité :

Loi de poisson :Solution

$$P(X=3)=\frac{2^3e^{-2}}{3!}=0,1804$$

$$P(X=4)=\frac{2^4e^{-2}}{4!}=0,0902$$

3- $\boxed{E(X)=m=2}$ $\boxed{Var(X)=m=2}$ $\boxed{\sigma(X)=\sqrt{2}}$

Les principales lois de probabilité : Loi continue

Loi uniforme continue

Soit a et b deux réels tels que $a < b$.La fonction f définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } x \in [a;b] \\ 0 & \text{sinon} \end{cases}$$ est une densité de probabilité.

Définition :

Soit a et b deux réels tels que $a < b$, et X une variable aléatoire.
On dit que X suit **la loi uniforme sur** $[a;b]$ lorsque X suit la loi à densité continue f définie sur $\mathbb{R}$ par

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } x \in [a;b] \\ 0 & \text{sinon} \end{cases}$$

Les principales lois de probabilité : Loi continue

Définition :

On note $X \hookrightarrow X(a;b)$. Sa fonction de répartition est donnée par :

$$F(x) = \begin{cases} 0 & \text{pour } x < a, \\ \dfrac{x-a}{b-a} & \text{pour } x \in [a;b] \\ 1 & \text{pour } x > b. \end{cases}$$

Proposition :

Soit X une variable aléatoire suivant la loi uniforme sur $[a;b]$. Alors l'espérance $E(X)$ de X est :

$$E(X) = \int_a^b xf(x)\,dx = \frac{a+b}{2}.$$

La variance de la loi uniforme continue vaut : $V(X) = \dfrac{(b-a)^2}{12}$.

Les principales lois de probabilité : Loi Exponentielle

Introduction :

On se place dans le cas d'un phénomène d'attente et on s'intéresse à la variable aléatoire qui représente le temps d'attente pour la réalisation d'un événement ou le temps d'attente entre la réalisation de deux événements successifs. Si on se place dans le cas où l'intensité α du processus de Poisson est constante, ce temps d'attente suit une loi exponentielle de paramètre α.

Exemple. Lorsque l'événement attendu est la mort d'un individu (ou la panne d'un équipement), α s'appelle le taux de mortalité (ou le taux de panne). Dire qu'il a une valeur constante, c'est supposer qu'il n'y a pas de vieillissement (ou pas d'usure s'il s'agit d'un équipement), la mort ou la panne intervenant de façon purement accidentelle.

Les principales lois de probabilité : Loi Exponentielle

Définition :

Soit α un nombre strictement positif. On dit qu'une variable aléatoire continue X suit une loi exponentielle de paramètre α si sa fonction de densité est

$$f(x) = \begin{cases} \alpha e^{-\alpha x} \text{ pour } x \geqslant 0, \\ 0 \quad \text{sinon} \end{cases}$$

On note $X \hookrightarrow Exp(\alpha)$. Sa fonction de répartition est donnée par :

$$F(x) = \begin{cases} 0 \text{ pour } x < 0, \\ 1 - e^{-\alpha x} \text{ pour } x \geqslant 0, \\ 1 \quad \text{pour } x = +\infty \end{cases}$$

Notons que, $P(X > x) = 1 - F(x) = e^{-\alpha x}$ pour $x \geqslant 0$.

Les principales lois de probabilité : Loi Exponentielle

Proposition :

Soit X une variable aléatoire suivant une loi exponentielle de paramètre α. Alors l'espérance $E(X)$ de X est :

$$E(X) = \frac{1}{\alpha}.$$

Et la variance vaut :

$$V(X) = \frac{1}{\alpha^2}.$$

Exercice

La durée de vie T en année, d'un appareil avant la première panne suit une loi exponentielle de paramètre α. D'après une étude, la probabilité que cet appareil tombe en panne pour la première fois avant la fin de la première année est 0,2. D'après cette étude, déterminer la valeur de α à 10^{-2} près.

Les principales lois de probabilité : Loi Normale

Définition :

Une variable aléatoire continue suit une loi normale si l'expression de sa fonction de densité de probabilités est de la forme :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

La loi dépend des deux réels m et σ appelés paramètres de la loi normale. On la note $\mathcal{N}(m, \sigma)$.

Remarque :

1. Une fonction de densité de probabilité étant toujours positive, le paramètre σ est donc un réel strictement positif.
2. On démontre que f est bien une fonction de densité de probabilité car $\int_{\mathbb{R}} f(x)\,dx = 1$. Pour le démontrer on utilise que $\int_{\mathbb{R}} e^{-x^2/2}\,dx = \sqrt{2\pi}$ (c'est l'intégrale de Gauss).

Les principales lois de probabilité : Loi Normale

Proposition :

$$E(X) = m, \quad Var(X) = \sigma^2, \quad \sigma(X) = \sigma.$$

On peut faire le calcul directement, à partir de l'intégrale de Gauss.

Somme de deux variables normales :

Soient X_1 et X_2 deux variables indépendantes. Si X_1 suit $\mathcal{N}(m_1, \sigma_1)$ et X_2 suit $\mathcal{N}(m_2, \sigma_2)$, alors $X_1 + X_2$ suit $\mathcal{N}(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.

Loi Normale :Propriétés de la distribution normale

Loi normale centrée réduite ou loi normale standardisée :

Nous avons qu'à toute variable aléatoire X, on pouvait associer une variable dite standardisée $\dfrac{X - E(X)}{\sigma(X)}$ d'espérance nulle et de variance unité (vaut 0).

On montre assez facilement que si on effectue cette transformation sur une variable suivant une loi normale, la variable standardisée suit encore une loi normale mais cette fois-ci de paramètres 0 et 1. La loi standardisée est appelée loi normale centrée réduite, et notée $\mathcal{N}(0,1)$. Donc si X suit $\mathcal{N}(m,\sigma)$, on pose $T = \dfrac{X - m}{\sigma}$ et T suit $\mathcal{N}(0,1)$.

On peut résumer la correspondance de la façon suivante :

$X \rightharpoonup \mathcal{N}(m,\sigma)$ $E(X) = m$ $Var(X) = \sigma^2$	$T = \frac{X-m}{\sigma}$	$T \rightharpoonup \mathcal{N}(0,1)$ $E(T) = 0$ $Var(T) = 1$

Loi Normale :Propriétés de la distribution normale

Loi normale centrée réduite ou loi normale standardisée :

La loi $\mathcal{N}(0,1)$ est tabulée à l'aide la fonction de répartition des valeurs positives. Elle donne les valeurs de $\Phi(t) = P(0 \leqslant T \leqslant t) = \int_0^t \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$ pour $t > 0$. Ce nombre représente l'aire sous la courbe représentative de la distribution et au dessus de l'intervalle $[0, t]$. Pour cette raison la table de la loi normale est appelée table d'aires (elle est aussi appelée **Loi de Laplace-Gauss**). Cette table ne dépend d'aucun paramètre, mais permet cependant de déterminer les probabilités de n'importe quelle distribution normale !

Loi Normale :Propriétés de la distribution normale

Loi normale centrée réduite : Comment utiliser la table d'aires ?

La première colonne de la table indique les unités et les dixièmes des valeurs de T alors que les centièmes des valeurs de T se lisent sur la ligne supérieure de la table. La valeur trouvée à l'intersection de la ligne et de la colonne adéquates donne l'aire cherchée.

a) Je cherche la valeur de A l'intersection de la ligne "0.5" et de la colonne "0.00", je lis 0.1915.

b) Je cherche la valeur de $P(-0.5 \leqslant T \leqslant 0)$. J'utilise la symétrie de la courbe par rapport à l'axe des ordonnées et j'en conclus que
$P(-0.5 \leqslant T \leqslant 0) = P(0 \leqslant T \leqslant 0.5) = 0.1915$. Et que pensez-vous de la valeur de $P(-0.5 < T < 0)$?

c) Je cherche la valeur de $P(-2.24 \leqslant T \leqslant 1.12)$. L'aire cherchée correspond à la somme suivante

$$P(-2.24 \leqslant T \leqslant 1.12) = P(-2.24 \leqslant T \leqslant 0) + P(0 < T \leqslant 1.12)$$

$$= 0.4875 + 0.3686 = 0.8561.$$

Loi Normale :Propriétés de la distribution normale

Loi normale centrée réduite : Comment utiliser la table d'aires ?

d) Je cherche la valeur de $P(1 \leqslant T \leqslant 2)$. L'aire cherchée correspond à la différence suivante

$$P(1 \leqslant T \leqslant 2) = P(0 \leqslant T \leqslant 2) - P(0 \leqslant T \leqslant 1) = 0.4772 - 0.3413 = 0.1359.$$

e) Je cherche la valeur t de T telle que $P(0 \leqslant T \leqslant t) = 0.4750$. C'est le problème inverse de celui des exemples précédents. Il s'agit de localiser dans la table l'aire donnée et de déterminer la valeur de T correspondante. Je trouve $t = 1.96$.

Loi Normale :Propriétés de la distribution normale

Approximation de la loi binomiale par la loi normale

On approche la loi $\mathcal{B}(n,p)$ par la loi $\mathcal{N}(np,\sqrt{npq})$ dès que $\begin{cases} n \geqslant 30 \\ np \geqslant 15 \\ nq \geqslant 15 \end{cases}$

Remarque

Remplacer une loi binomiale par une loi normale simplifie considérablement les calculs. En effet les tables de la loi binomiale dépendent de deux paramètres et les valeurs de n dans ces tables sont limitées supérieurement par 20. La loi normale, elle, après standardisation ne dépend d'aucun paramètre .

Loi Normale :Propriétés de la distribution normale

Approximation de la loi de Poisson par la loi normale

On démontre qu'on peut aussi approcher la loi de Poisson par la loi normale pour les grandes valeurs du paramètre de la loi de Poisson. La seule qui puisse convenir est celle qui a même espérance et même variance. On approche donc la loi $\mathcal{P}(\lambda)$ par la loi $\mathcal{N}(\lambda, \sqrt{\lambda})$. En pratique, cela s'applique dès que $\lambda \geqslant 16$.

On approche la loi $\mathcal{P}(\lambda)$ par la loi $\mathcal{N}(\lambda, \sqrt{\lambda})$ dès que $\lambda \geqslant 16$

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Exercice 1 : Supposons qu'une tentative pour obtenir une communication téléphonique échoue (par exemple, parce que la ligne est occupée) avec la probabilité 0.25 et réussisse avec la probabilité 0.75. On suppose que les tentatives sont indépendantes les unes des autres. Quelle est la probabilité d'obtenir la communication si l'on peut effectuer trois tentatives au maximum ?

Exercice 2 : Un fabricant de pièces de machine prétend qu'au plus 10% de ses pièces sont défectueuses. Un acheteur a besoin de 120 pièces. Pour disposer d'un nombre suffisant de bonnes pièces, il en commande 140. Si l'affirmation du fabricant est valable, quelle est la probabilité que l'acheteur reçoive au moins 120 bonnes pièces ?

Exercice 3 : Les statistiques antérieures d'une compagnie d'assurances permettent de prévoir qu'elle recevra en moyenne 300 réclamations durant l'année en cours. Quelle est la probabilité que la compagnie reçoive plus de 350 réclamations pendant l'année en cours ?

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Exercice 4 : Le nombre moyen de clients qui se présentent à la caisse d'un supermarché sur un intervalle de 5 minutes est de 10. Quelle est la probabilité qu'aucun client ne se présente à la caisse dans un intervalle de deux minutes (deux méthodes possibles) ?

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution de l'Exercice 1 : Nous nous intéressons à la variable $X =$ « nombre de tentatives nécessaires pour obtenir la communication », ce que l'on peut considérer comme le nombre d'essais à faire pour obtenir le premier succès. X suit une loi géométrique de paramètre $p = 0.75$.
On cherche à déterminer $P(X \leqslant 3) = P(X = 1) + P(X = 2) + P(X = 3)$.

- On peut obtenir la communication au 1er essai. On a pour cela une probabilité $P(X = 1) = q^0 p^1 = p = 0.75$.
- On peut obtenir la communication au 2ème essai. On a pour cela une probabilité $P(X = 2) = q^1 \times p = 0.25 \times 0.75 = 0.1875$.
- On peut obtenir la communication au 3ème essai. On a pour cela une probabilité $P(X = 3) = q^2 p = 0.25^2 \times 0.75 = 0.0469$.

Finalement la probabilité d'obtenir la communication en trois essais maximum est $0.75 + 0.1875 + 0.0469 = 0.9844$ soit 98.5 %.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution de l'Exercice 2 :

Appelons X la variable aléatoire correspondant au "nombre de bonnes pièces dans le lot de 140 pièces".

X prend ses valeurs entre 0 et 140. De plus pour chaque pièce, on n'a que deux éventualités : elle est bonne ou elle est défectueuse. La probabilité qu'une pièce soit défectueuse est 0.1. Par conséquent elle est bonne avec la probabilité 0.9. On est donc dans une situation type : X suit la loi binomiale $\mathcal{B}(140, 0.9)$ de paramètres $n = 140$ et $p = 0.9$.

On veut déterminer la probabilité que l'acheteur reçoive au moins 120 bonnes pièces sur les 140, soit $X \geqslant 120$. A priori, il nous faudrait calculer la somme des probabilités $P(X = 120) + P(X = 121) + \cdots + P(X = 140)$, ce qui serait épouvantablement long. On approxime donc la loi binomiale par une loi tabulée.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution de l'Exercice 2 :

Comme $n \geqslant 30$, $np = 126 \geqslant 15$ et $nq = 14$, on pourra approcher la loi binomiale par une loi normale. On choisit la loi normale qui a la même espérance et le même écart-type.
Donc X qui suit la loi $\mathcal{B}(140, 0.9)$ sera approchée par Y qui suit la loi $\mathcal{N}(126, 3.55)$.
Pour remplacer une loi discrète par une loi continue, il est préférable d'utiliser la correction de continuité,
On se ramène enfin à la loi normale centrée réduite. On pose $T = \frac{Y-126}{3.55}$, et

$$\begin{aligned} P(Y > 120) &= P(T > \frac{120-126}{3.55}) = P(T > -1.69) \\ &= P(T < 1.69) = 0.5 + \Phi(1.69) = 0.5 + 0,4545 = 0,96. \end{aligned}$$

Conclusion : l'acheteur a 96 chances sur 100 de recevoir 120 bonnes pièces sur les 140 achetées.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution de l'Exercice 3 :

La variable X qui nous intéresse est le "nombre de réclamations reçues pendant une année". Il s'agit du nombre de réalisations d'un événement pendant un intervalle de temps donné. X suit donc une loi de Poisson. Le nombre moyen de réalisations dans une année est 300. Cette valeur moyenne est aussi le paramètre de la loi de Poisson. Donc X suit la loi $\mathcal{P}(300)$.

On cherche à déterminer $P(X > 350)$. Il n'y a pas de table de la loi de Poisson pour cette valeur du paramètre. Il nous faut donc approcher X qui suit la loi de Poisson $\mathcal{P}(300)$ par Y qui suit la loi normale de même espérance et de même écart-type, c'est-à-dire $\mathcal{N}(300, \sqrt{300})$.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution de l'Exercice 3 :

On se ramène finalement à la loi normale centrée réduite. On pose $T = \frac{X-300}{\sqrt{300}}$.

$$P(X > 350) = P(T > \frac{350 - 300}{\sqrt{300}}) = P(T > 2.89) = 0,5 - \Phi(2.89) = 0.0019.$$

La compagnie d'assurances a donc 0,19% de chances de recevoir plus de 350 réclamations en un an.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution n ^{0}1 de l'Exercice 4 :

Considérons la variable aléatoire $X =$ "nombre de clients se présentant à la caisse dans un intervalle de deux minutes". Nous reconnaissons une situation type et la variable X suit une loi de Poisson. Vu qu'en moyenne 10 clients se présentent en 5 mn, l'intensité α du processus est de 2 clients par minute, $\alpha = 2$. Or le paramètre de la loi de Poisson est αt_0, t_0 étant ici 2 minutes. D'où $\lambda = 4$.

On cherche à calculer $P(X = 0)$. D'après la formule du cours,

$P(X = 0) = e^{-\lambda} = e^{-4} = 0.018$.

Loi Normale :Propriétés de la distribution normale

Quelques exercices types

Solution n^{0}2 de l'Exercice 4 :

Considérons à présent la question sous un autre angle en s'intéressant au temps d'attente Y entre deux clients. Le cours nous dit que la loi suivie par une telle variable est une loi exponentielle. Son paramètre α est l'intensité du processus de Poisson soit ici $\alpha = 2$. Y suit donc la loi $Exp(2)$.

Sa fonction de densité est $2e^{-2x}$ pour $x > 0$ exprimé en minutes. On en déduit que

$$P(Y \geqslant 2) = \int_2^{+\infty} 2e^{-2x}\, dx = [-e^{-2x}]_2^{+\infty} = e^{-4} = 0.018.$$

La distribution du χ^2 de Pearson

Elle a été découverte en 1905 par le mathématicien britannique Karl Pearson (1857-1936) qui travailla également sur les problèmes de régression avec le généticien Sir Francis Galton. Cette distribution (qui se prononce khideux) est très importante pour tester l'ajustement d'une loi théorique à une distribution expérimentale (test du χ^2) et pour déterminer la loi de la variance d'un échantillon.

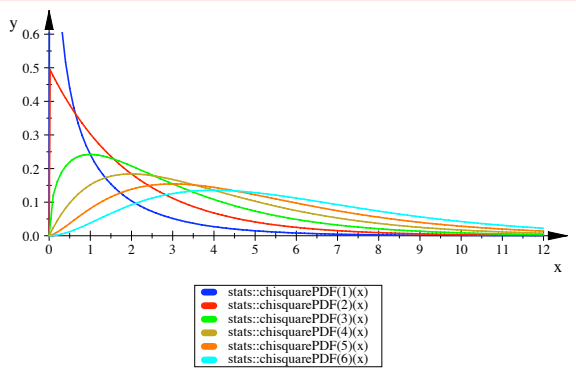
Définition

Si $X_1, X_2, \ldots, X_n$ sont n variables aléatoires indépendantes qui suivent toute la loi normale centrée réduite, alors la quantité $X = X_1^2 + X_2^2 + \cdots + X_n^2$ est une variable aléatoire distribuée selon la loi du χ^2 à n degrés de liberté.
On note $X \rightharpoonup \chi_n^2$.

La distribution du χ^2 de Pearson

Forme de la distribution

La distribution du χ^2 est continue à valeurs positives et présente un étalement sur le côté supérieur. Elle ne dépend que du nombre de degrés de liberté n.
Ci-dessous, densité de χ^2_n pour $n = 1, \ldots, 6$.



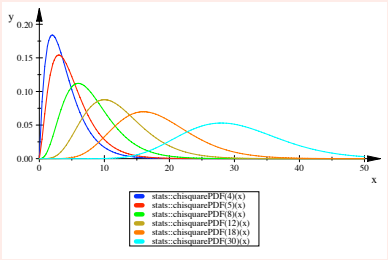
La distribution du χ^2 de Pearson

Paramètres descriptifs

$E(X) = n$, $V(X) = 2n$.

Approximation par une loi normale

A mesure que n augmente, la loi du χ^2 tend vers la loi normale, comme on peut le constater sur le graphique ci-dessous.



Densité de χ^2_n pour $n = 4, 5, 8, 12, 18, 30$.

La distribution du χ^2 de Pearson

En pratique, on peut considérer que pour $n \geqslant 30$, on peut remplacer la loi du χ^2 à n degrés de liberté par la loi normale $\mathcal{N}(n, \sqrt{2n})$.

La distribution de Fisher-Snedecor

Cette distribution fut découverte par l'anglais Fisher en 1924 puis tabulée par Snédecor en 1934. Elle interviendra lors des comparaisons des variances de deux échantillons (test d'hypothèse F).

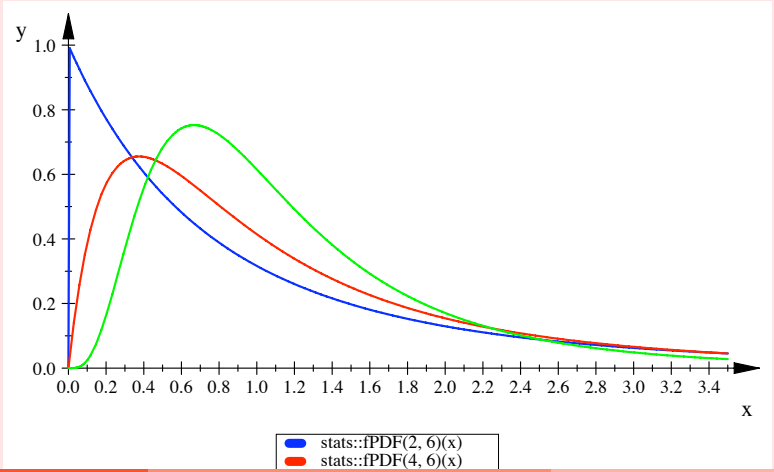
Définition

Si X_1 et X_2 sont deux variables aléatoires indépendantes qui suivent toutes les deux une loi de khi-deux de degrés de liberté respectifs n_1 et n_2, alors la quantité $F = \frac{X_1/n_1}{X_2/n_2}$ est une variable aléatoire qui suit la loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté.
On note $F \rightharpoonup F_{n_1,n_2}$. Cette variable ne prend que des valeurs positives.

La distribution de Fisher-Snedecor

Forme de la distribution

A mesure que les valeurs n_1 et n_2 augmentent, la loi de Fisher tend vers une loi normale.



La distribution de Student

Student est le pseudonyme de V.S Gosset, 1908.

Définition

Soient X et Y deux variables aléatoires indépendantes, la première étant distribuée selon une loi normale centrée réduite $\mathcal{N}(0,1)$ et la deuxième selon une loi de khi-deux à n degrés de liberté. La quantité $\mathrm{T} = \frac{X\sqrt{n}}{\sqrt{Y}}$ est une variable aléatoire qui suit une *loi de Student à n degrés de liberté*.
On écrit $T \rightharpoonup T_n$.

La distribution de Student

Paramètres descriptifs

On a $E(T_n) = 0$ si $n > 1$ et $Var(T_n) = \frac{n}{n-2}$ si $n > 2$.

Approximation par une loi normale

A mesure que n augmente, la distribution de Student à n degrés de liberté se rapproche de plus en plus de celle de celle de la loi normale centrée réduite.
En pratique : si $T \rightharpoonup T_n$ pour $n \geqslant 30$, on pourra écrire que $T \rightharpoonup \mathcal{N}(0,1)$.

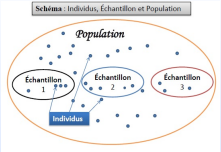
Échantillonnage

Avant de présenter les méthodes d'échantillonnage et la méthode de détermination de l'échantillon, il est essentiel de maîtriser deux notions de base : celle de population et celle d'échantillon

Définitions

Population : Ensemble que l'on observe et qui sera soumis à une analyse statistique (Par exemple les étudiants de l'ENCG, la population féminine, les fonctionnaires,..). Chaque élément de cet ensemble est un **Individu** ou **Unité statistique**.

Échantillon C'est un sous ensemble de la population considérée. Le nombre d'individus dans l'échantillon est la **taille** de l'échantillon.



Échantillonnage

Définition générale

"La théorie mathématique des probabilités suppose que, pour connaître les événements qui peuvent survenir dans une population donnée, il n'est possible d'étudier ou d'interroger qu'une petite partie de celle-ci, à condition de respecter des règles rigoureuses de sélection de cette fraction de population. Seules garanties de sa représentativité."

Hélène Yvonne Meynaud et Denis Duclos dans De l'échantillonnage à la remise du produit.

- **L'échantillonnage est un procédé qui permet de définir un échantillon dans un travail d'enquête. Il s'agit d'étudier une partie sélectionnée pour établir des conclusions applicables à un tout.**
- **En d'autres termes, l'échantillonnage est une sélection précise de personnes ciblées pour réaliser un entretien, un focus group, un sondage ou un questionnaire.**

Échantillonnage

Deux types d'échantillons peuvent être distingués : les échantillons non-probabilistes et les échantillons probabilistes

Les échantillons non-probabilistes

Les sujets ou les objets sont choisis selon une procédure pour laquelle la sélection n'est pas aléatoire. Ce type d'échantillon pose plusieurs problèmes inférentiels.

Les échantillons probabilistes

Dans ce cas, les sujets ou les objets sont choisis selon une procédure où la sélection est aléatoire. Deux règles sont à respecter dans les procédures d'échantillonnage :

1. La base d'échantillonnage doit inclure toutes les entités i.e. les sujets ou les objets ou les unités statistique à partir desquels le choix des entités sera fait.
2. Les entités doivent être sélectionnées par une procédure d'échantillonnage indépendante et aléatoire à l'aide par exemple d'une table de nombres aléatoires.

Échantillonnage

Le statisticien décrit habituellement ces ensembles à l'aide de mesures telles que le nombre d'unités, la moyenne, l'écart-type et le pourcentage.

- Les mesures que l'on utilise pour décrire une population sont des paramètres. Un paramètre est une caractéristique de la population.
- Les mesures que l'on utilise pour décrire un échantillon sont appelées des statistiques. Une statistique est une caractéristique de l'échantillon.

Remarques :

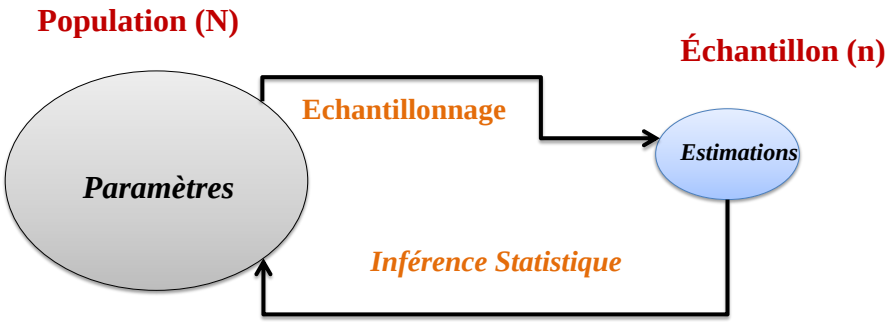
Nous allons voir dans ce chapitre et dans le suivant comment les résultats obtenus sur un échantillon peuvent être utilisés pour décrire la population. On verra en particulier que les statistiques sont utilisées pour estimer les paramètres.

Échantillonnage

Afin de ne pas confondre les statistiques et les paramètres, on utilise des notations différentes, comme le présente le tableau récapitulatif suivant.

	Population	**Échantillon**
Définition	C'est l'ensemble des unités considérées par le statisticien.	C'est un sous-ensemble de la population choisie pour étude.
Caractéristiques	Ce sont les paramètres	Ce sont les statistiques
Notations	N = taille de la population (si elle est finie)	n = taille de l'échantillon
Si on étudie un caractère quantitatif	moyenne de la population $m = \frac{1}{N}\sum_{i=1}^{N} x_i$	moyenne de l'échantillon $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
	écart-type de la population $\sigma_{pop} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$	écart-type de l'échantillon $\sigma_{ech} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Si on étudie un caractère qualitatif	proportion dans la population p	proportion dans l'échantillon f

Échantillonnage



Échantillonnage

Avantages de l'échantillonnage

- Coût moindre.
- Gain de temps.
- C'est la seule méthode qui donne des résultats dans le cas d'un test destructif.

Méthodes d'échantillonnage

- Échantillonnage sur la base du jugement (par exemple, dans les campagnes électorales certains districts électoraux sont des indicateurs fiables de l'opinion publique).
- Échantillonnage aléatoire simple. Tous les échantillons possibles de même taille ont la même probabilité d'être choisis et tous les éléments de la population ont une chance égale de faire partie de l'échantillon (On utilise souvent une table de nombres aléatoires pour s'assurer que le choix des éléments s'effectue vraiment au hasard).

Échantillonnage

Inconvénient de l'échantillonnage

L'échantillonnage a pour but de fournir suffisamment d'informations pour pouvoir faire des déductions sur les caractéristiques de la population. Mais bien entendu, les résultats obtenus d'un échantillon à l'autre vont être en général différents et différents également de la valeur de la caractéristique correspondante dans la population. On dit qu'il y a des fluctuations d'échantillonnage. Comment, dans ce cas, peut-on tirer des conclusions valables ? En déterminant les lois de probabilités qui régissent ces fluctuations. C'est l'objet de ce chapitre.

La variable aléatoire : moyenne d'échantillon

Introduction : Position du problème

Si nous prélevons un échantillon de taille n dans une population donnée, la moyenne de l'échantillon nous donnera une idée approximative de la moyenne de la population. Seulement si nous prélevons un autre échantillon de même taille, nous obtiendrons une autre moyenne d'échantillon. Sur l'ensemble des échantillons possibles, on constatera que certains ont une moyenne proche de la moyenne de la population et que d'autres ont une moyenne qui s'en écarte davantage.

La variable aléatoire : moyenne d'échantillon

Comment traiter le problème ?

Un échantillon de taille n (appelé aussi un n-échantillon), obtenu par échantillonnage aléatoire, va être considéré comme le résultat d'une expérience aléatoire. A chaque échantillon de taille n on peut associer la valeur moyenne des éléments de l'échantillon. On a donc défini une variable aléatoire qui à chaque n-échantillon associe sa moyenne échantillonnale. On la note $\bar{X}$. Cette variable aléatoire possède bien entendu :

- Une distribution de probabilité.
- Une valeur moyenne (la moyenne des moyennes d'échantillons, vous suivez toujours ?).
- Un écart-type.

Remarque :

Le but de ce paragraphe est de déterminer ces trois éléments.

Avant de continuer, essayons de comprendre sur un exemple ce qui se passe.

La variable aléatoire : moyenne d'échantillon

Exemple 1

Une population est constituée de 5 étudiants en statistique (le faible effectif n'est pas dû à un manque d'intérêt pour la matière de la part des étudiants mais au désir de ne pas multiplier inutilement les calculs qui vont suivre !). Leur professeur s'intéresse au temps hebdomadaire consacré à l'étude des statistiques par chaque étudiant.

On a obtenu les résultats suivants.

Étudiant	Temps d'étude (en heures)
A	7
B	3
C	6
D	10
E	4
Total	30

La variable aléatoire : moyenne d'échantillon

La moyenne de la population est $m = 30/5 = 6$.

Si le professeur choisit un échantillon de taille 3, quelles sont les différentes valeurs possibles pour la moyenne de son échantillon ? Quelle relation existe-t-il entre cette moyenne d'échantillon et la véritable moyenne 6 de la population ?

Toutes les possibilités sont regroupées dans le tableau ci-dessous.

Numéro de l'échantillon	Échantillon	Valeurs du temps d'étude dans cet échantillon	Moyennes de l' échantillon
1	A, B, C	7,3,6	5.33
2	A, B, D	7,3,10	6.67
3	A, B, E	7,3,4	4.67
4	A, C, D	7,6,10	7.67
5	A, C, E	7,6,4	5.67
6	A, D, E	7,10,4	7.00
7	B, C, D	3,6,10	6.33
8	B, C, E	3,6,4	4.33
9	B, D, E	3,10,4	5.67
10	C, D, E	6,10,4	6.67
Total			60.00

La variable aléatoire : moyenne d'échantillon

On constate que :

- Il y a 10 échantillons ($C_5^3 = 10$).
- La moyenne des échantillons varie entre 4.33 et 7.67, ce qui signifie que la distribution des moyennes d'échantillon est moins dispersée que la distribution des temps d'étude des étudiants, située entre 3 et 10.
- Il est possible que deux échantillons aient la même moyenne. Dans cet exemple, aucun n'a la moyenne de la population ($m = 6$).
- La moyenne des moyennes d'échantillon est $E(\bar{X}) = 60/10 = 6$.

En fait, nous allons voir que le fait que l'espérance de $\bar{X}$ (c'est-à-dire la moyenne des moyennes d'échantillon) est égale à la moyenne de la population n'est pas vérifié seulement dans notre exemple. C'est une loi générale.

La variable aléatoire : moyenne d'échantillon

Bien, me direz-vous, mais pourquoi faire tout cela ? Dans la réalité, on ne choisit qu'un seul échantillon. Alors comment le professeur de statistique qui ne connaît qu'une seule moyenne d'échantillon pourra-t-il déduire quelque chose sur la moyenne de la population ? Tout simplement en examinant "jusqu'à quel point" la moyenne d'un échantillon unique s'approche de la moyenne de la population. Pour cela, il lui faut la distribution théorique de la variable aléatoire $\bar{X}$ ainsi que l'écart-type de cette distribution.

Étude de la variable : moyenne d'échantillon

Définition de la variable

On considère une population dont les éléments possèdent un caractère mesurable qui est la réalisation d'une variable aléatoire X qui suit une loi de probabilité d'espérance m et d'écart-type σ_{pop}. On suppose que la population est infinie ou si elle est finie que l'échantillonnage se fait avec remise.

- On prélève un échantillon aléatoire de taille n et on mesure les valeurs de X sur chaque élément de l'échantillon. On obtient une suite de valeurs $x_1, x_2, \ldots, x_n$.
- Si on prélève un deuxième échantillon toujours de taille n, la suite des valeurs obtenues est $x'_1, x'_2, \ldots, x'_n$, puis $x''_1, x''_2, \ldots, x''_n$... etc... pour des échantillons supplémentaires.

Étude de la variable : moyenne d'échantillon

Définition de la variable

x_1, x_1', x_1'',... peuvent être considérées comme les valeurs d'une variable aléatoire X_1 qui suit la loi de X. De même, x_2, x_2', x_2'',... peuvent être considérées comme les valeurs d'une variable aléatoire X_2 qui suit aussi la loi de X, ... et x_n, x_n', x_n'',... celles d'une variable aléatoire X_n qui suit encore et toujours la même loi, celle de X.

- X_1 pourrait se nommer "valeur du premier élément d'un échantillon". X_2 pourrait se nommer "valeur du deuxième élément d'un échantillon". ....X_n pourrait se nommer "valeur du n-ième élément d'un échantillon".
- L'hypothèse d'une population infinie ou d'un échantillonnage avec remise nous permet d'affirmer que ces n variables aléatoires sont indépendantes.

Étude de la variable : moyenne d'échantillon

Rappel sur les notations

Par convention, on note toujours les variables aléatoires à l'aide de lettres majuscules (X_i) et les valeurs qu'elles prennent dans une réalisation à l'aide de lettres minuscules (x_i). Si les valeurs prises par X dans un échantillon sont $x_1, x_2, \ldots, x_n$, la moyenne $\bar{x}$ de l'échantillon est donnée par $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Cette valeur n'est rien d'autre que la valeur prise dans cet échantillon de la variable aléatoire $\frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Définition

On définit donc la *variable aléatoire* moyenne d'échantillon $\bar{X}$ par

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Paramètres descriptifs de la distribution

Espérance et Variance

- $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{nm}{n} = m$, car les variables suivent toutes la même loi d'espérance m.
- $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma_{pop}^2}{n^2} = \frac{\sigma_{pop}^2}{n}$, car les variables suivent toutes la même loi de variance et sont indépendantes.

Proposition

$$E(\bar{X}) = m, \quad Var(\bar{X}) = \frac{\sigma_{pop}^2}{n}.$$

Paramètres descriptifs de la distribution

Remarque

1. Nous venons de démontrer ce que nous avions constaté sur notre exemple : la moyenne de la distribution d'échantillonnage des moyennes est égale à la moyenne de la population.
2. On constate que plus n croît, plus $Var(\bar{X})$ décroît.

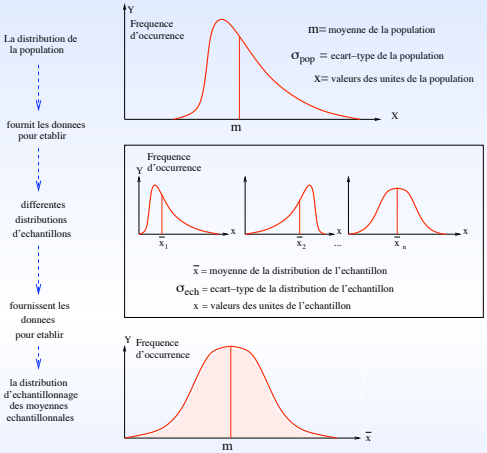
Dans l'exemple d'introduction, nous avions en effet constaté que la distribution des moyennes d'échantillon était moins dispersée que la distribution initiale. En effet, à mesure que la taille de l'échantillon augmente, nous avons accès à une plus grande quantité d'informations pour estimer la moyenne de la population. Par conséquent, la différence probable entre la vraie valeur de la moyenne de la population et la moyenne échantillonnale diminue. L'étendue des valeurs possibles de la moyenne échantillonnale diminue et le degré de dispersion de la distribution aussi. $\sigma(\bar{X})$ est aussi appelé l'*erreur-type* de la moyenne.

Paramètres descriptifs de la distribution

On peut schématiser le passage de la distribution de la variable aléatoire X à celle de la variable aléatoire $\bar{X}$ en passant par les différents échantillons par le graphique ci-après. Mais connaître les paramètres descriptifs de la distribution de $\bar{X}$ ne suffit pas. Il faut connaître aussi sa distribution de probabilité. On se demande alors : dépend elle

1. de la distribution de X ?
2. de la taille n de l'échantillon ?

Paramètres descriptifs de la distribution



m= moyenne de la distribution d'echantillonnage des moyennes

$\sigma_{pop}/\sqrt{n}$ = ecart–type de la distribution d'echantillonnage des moyennes

$\bar{x}$ = valeurs des moyennes de tous les echantillons possibles

Distribution de la moyenne d'échantillon

Nous allons distinguer deux cas : celui des grands échantillons ($n \geqslant 30$) et celui des petits échantillons ($n < 30$).

Cas des grands échantillons : $n \geqslant 30$

On peut appliquer le théorème centrale-limite.

1. Nous sommes en présence de n variables aléatoires indépendantes.
2. Elles suivent la même loi d'espérance m et de variance σ^2_{pop}, donc aucune n'est prépondérante.

Conclusion. Lorsque n devient très grand, la distribution de $S = X_1 + \cdots + X_n$ se rapproche de celle de la loi normale d'espérance nm et de variance $n\sigma^2_{pop}$, S suit approximativement $\mathcal{N}(nm, n\sigma^2_{pop})$.

Distribution de la moyenne d'échantillon

Par conséquent, pour n assez grand, la distribution de $\bar{X} = S/n$ se rapproche de celle de la loi normale d'espérance m et de variance σ_{pop}^2/n c'est-à-dire $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$. On peut donc considérer que $\dfrac{\bar{X} - m}{\sigma_{pop}/\sqrt{n}}$ suit la loi $\mathcal{N}(0, 1)$.

Proposition

Si $n \geqslant 30$, $\bar{X}$ suit approximativement $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$.

Distribution de la moyenne d'échantillon

Remarque

- En pratique, on considère que cela est vrai à partir de $n \geqslant 30$ et que lorsque la forme de la distribution de X est pratiquement symétrique, $n \geqslant 15$ est convenable.
- Ce théorème est très puissant car il n'impose aucune restriction sur la distribution de X dans la population.
- Si la variance est inconnue, un grand échantillon ($n \geqslant 30$) permet de déduire une valeur fiable pour σ^2_{pop} en calculant la variance de l'échantillon σ^2_{ech} et en posant

$$\sigma^2_{pop} = \frac{n}{n-1}\sigma^2_{ech} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Distribution de la moyenne d'échantillon : Cas des petits échantillons : $n < 30$

Nous nous plaçons alors exclusivement dans le cas où X suit une loi normale dans la population.

Cas où σ_{pop} est connu

X suit une loi normale $\mathcal{N}(m, \sigma_{pop})$ donc les variables X_i suivent toutes la même loi $\mathcal{N}(m, \sigma_{pop})$. De plus elles sont indépendantes. $S = X_1 + \cdots + X_n$ a une distribution normale et la variable $\bar{X} = S/n$ suit aussi une loi normale, la loi $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}})$. Donc $\dfrac{\bar{X} - m}{\sigma_{pop}/\sqrt{n}}$ suit la loi $\mathcal{N}(0, 1)$.

$$\text{Si} \begin{cases} n < 30 \\ \sigma_{pop} \text{ connu} \end{cases}, \bar{X} \text{ suit } \mathcal{N}(m, \tfrac{\sigma_{pop}}{\sqrt{n}}).$$

Distribution de la moyenne d'échantillon

Exemple

Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Il semble plausible de supposer que les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne $m = 150$ et de variance $\sigma^2_{pop} = 100$. On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?

Distribution de la moyenne d'échantillon

Solution : Test d'aptitude

On considère la variable aléatoire $\bar{X}$ moyenne d'échantillon pour les échantillons de taille $n = 25$. On cherche à déterminer $P(146 < \bar{X} < 154)$.
Pour cela, il nous faut connaître la loi suivie par $\bar{X}$. Examinons la situation. Nous sommes en présence d'un petit échantillon ($n < 30$) et heureusement dans le cas où la variable X (résultat au test d'aptitude) suit une loi normale. De plus, σ_{pop} est connu. Donc $\bar{X}$ suit $\mathcal{N}(m, \frac{\sigma_{pop}}{\sqrt{n}}) = \mathcal{N}(150, 10/5)$. On en déduit que $T = \frac{\bar{X}-150}{2}$ suit $\mathcal{N}(0, 1)$.
La table donne

$$\begin{aligned} P(146 < \bar{X} < 154) &= P(\frac{146-150}{2} < T < \frac{154-150}{2}) = P(-2 < T < 2) \\ &= 2P(T < 2) - 1 = 2 \times 0,9772 - 1 = 0.9544. \end{aligned}$$

Distribution de la variable proportion d'échantillon

Il arrive fréquemment que nous ayions à estimer dans une population une proportion p d'individus possédant un caractère qualitatif donné.
Bien sûr, cette proportion p sera estimée à l'aide des résultats obtenus sur un n-échantillon. La proportion f obtenue dans un n-échantillon est la valeur observée d'une variable aléatoire F, fréquence d'apparition de ce caractère dans un échantillon de taille n, appelée proportion d'échantillon. On se pose une troisième fois la question. La moyenne des fréquences d'observation du caractère sur l'ensemble de tous les échantillons de taille n est-elle égale à la proportion p de la population ?

Paramètres descriptifs de la distribution de F

F est la fréquence d'apparition du caractère dans un échantillon de taille n. Donc $F = X/n$ où X est le nombre de fois où le caractère apparaît dans le n-échantillon. Par définition X suit $\mathcal{B}(n, p)$. Donc $E(X) = np$ et $Var(X) = npq$.
Il en résulte que

$$E(F) = p \quad \text{et} \quad Var(F) = \frac{pq}{n}.$$

Conséquences.

1. La réponse à la question que nous nous posions est oui : l'espérance de la fréquence d'échan-tillon est égale à la probabilité théorique d'apparition dans la population.
2. Lorsque la taille de l'échantillon augmente, la variance de F diminue, ce qui est logique : plus on a d'informations, plus il est probable que la proportion observée dans l'échantillon soit proche de la proportion de la population.

Distribution de la proportion d'échantillon dans le cas des grands échantillons

On sait que si $n \geqslant 30$, $np \geqslant 15$ et $nq \geqslant 15$, on peut approcher la loi binomiale par la loi normale de même espérance et de même écart-type. Donc F suit approximativement $\mathcal{N}(p, \sqrt{\frac{pq}{n}})$, et la variable $T = \dfrac{F - p}{\sqrt{\frac{pq}{n}}}$ suit alors approximativement la loi $\mathcal{N}(0, 1)$.

Exemple

Selon une étude sur le comportement du consommateur, 25% d'entre eux sont influencés par la marque, lors de l'achat d'un bien. Si on interroge 100 consommateurs pris au hasard, quelle est la probabilité pour qu'au moins 35 d'entre eux se déclarent influencés par la marque ?

Distribution de la proportion d'échantillon dans le cas des grands échantillons

Solution : Influence de la marque

Appelons F la variable aléatoire :"proportion d'échantillon dans un échantillon de taille 100". Il s'agit ici de la proportion de consommateurs dans l'échantillon qui se déclarent influencés par la marque. On cherche à calculer $P(F > 0.35)$.
Il nous faut donc déterminer la loi de F. Or $np = 100 \times 0.25 = 25$ et
$nq = 100 \times 0.75 = 75$. Ces deux quantités étant supérieures à 15, on peut considérer que F suit $\mathcal{N}(p, \sqrt{\frac{pq}{n}}) = \mathcal{N}(0.25, 0.0433)$.
On utilise la variable $T = \frac{F-0.25}{0.0433}$ qui suit la loi $\mathcal{N}(0,1)$. Il vient

$$P(F > 0.35) = P(T > 2.31) = 1 - P(T < 2.31) = 1 - 0.9896 = 0.0104.$$

Distribution de la proportion d'échantillon dans le cas des grands échantillons

Conclusion : Il y a environ une chance sur 100 pour que plus de 35 consommateurs dans un 100 - échantillon se disent influencés par la marque lorsque l'ensemble de la population contient 25% de tels consommateurs.

En pratique, il est peu fréquent de connaître p : on doit plutôt l'estimer à partir d'un échantillon. Comment faire ? C'est ce que nous traiterons dans le prochain chapitre

Estimation : Introduction

Dans de nombreux domaines (scientifiques, économiques, épidémiologiques...), on a besoin de connaître certaines caractéristiques d'une population. Mais, en règle générale, on ne peut pas les évaluer facilement du fait de l'effectif trop important des populations concernées. La solution consiste alors à estimer le paramètre cherché à partir de celui observé sur un échantillon plus petit.
L'idée de décrire une population à partir d'un **échantillon** réduit, à l'aide d'un "multiplicateur", n'a été imaginée que dans la seconde moitié du XVIIIème siècle, notamment par l'école arithmétique politique anglaise. Elle engendra une véritable révolution : l'observation d'échantillons permettait d'éviter des recensements d'une lourdeur et d'un prix exorbitants. Toutefois, on s'aperçut rapidement que les résultats manquaient d'exactitude. Nous savons maintenant pourquoi : on ne prenait en considération ni la *représentativité* de l'échantillon, ni les *fluctuations* d'échantillonnage. C'est là que le hasard intervient.

Estimation : Introduction

La première précaution à prendre est donc d'obtenir un échantillon représentatif. Nous pourrons en obtenir un par tirage au sort (voir le chapitre précédent sur l'échantillonnage aléatoire simple) :**le hasard** participe donc au travail du statisticien qui l'utilise pour pouvoir le maîtriser ! Mais, même tiré au sort, un échantillon n'est pas l'image exacte de la population, en raison des fluctuations d'échantillonnage. Lorsque, par exemple, on tire au sort des échantillons dans un urne contenant 20 % de boules blanches, on obtient des échantillons où la proportion de boules blanches fluctue autour de 20 %. Ces fluctuations sont imprévisibles : le hasard peut produire n'importe quel écart par rapport à la proportion de la population (20 %). Cependant, on s'en doute, tous les écarts ne sont pas également vraisemblables : les très grands écarts sont très peu probables. Au moyen du calcul des probabilités, le statisticien définit un intervalle autour du taux observé, intervalle qui contient probablement le vrai taux : c'est "**l'intervalle de confiance**" ou, plus couramment, la "fourchette".

Estimation : Introduction

Si l'on ne peut connaître le vrai taux par échantillonnage, peut-on au moins le situer avec certitude dans la fourchette ? Non. Le hasard étant capable de tous les caprices, on ne peut raisonner qu'en termes de probabilités, et la fourchette n'a de signification qu'assortie d'un certain risque d'erreur. On adopte souvent un risque de 5 % : cinq fois sur cent, le taux mesuré sur l'échantillon n'est pas le bon, le vrai taux étant en dehors de la fourchette. On peut diminuer le risque d'erreur mais alors la fourchette grandit et perd de son intérêt. Bien entendu, il existe une infinité de fourchettes, une pour chaque risque d'erreur adopté. On doit trouver un compromis entre le risque acceptable et le souci de précision.

Estimation : Introduction

Dans ce cours, nous allons apprendre à estimer à l'aide d'un échantillon :

- Dans le cas d'un caractère quantitatif la moyenne m et l'écart-type d'une population.
- Dans le cas d'un caractère qualitatif, la proportion p de la population.

Ces estimations peuvent s'exprimer par une seule valeur (estimation ponctuelle), soit par un intervalle (estimation par intervalle de confiance). Bien sûr, comme l'échantillon ne donne qu'une information partielle, ces estimations seront accompagnées d'une certaine marge d'erreur.

Estimation : L'estimation ponctuelle

Définition

Estimer un paramètre, c'est en chercher une valeur approchée en se basant sur les résultats obtenus dans un échantillon. Lorsqu'un paramètre est estimé par un seul nombre, déduit des résultats de l'échantillon, ce nombre est appelé *estimation ponctuelle* du paramètre.
L'estimation ponctuelle se fait à l'aide d'un *estimateur*, qui est une variable aléatoire d'échantillon. L'estimation est la valeur que prend la variable aléatoire dans l'échantillon observé.

Estimation : L'estimation ponctuelle

Estimation ponctuelle

Estimer un paramètre, par exemple : une moyenne, une variance, une proportion, etc..., c'est chercher une valeur approchée en se basant sur les résultats d'un échantillon. Lorsqu'un paramètre est estimé par un seul nombre déduit des résultats de l'échantillon, ce nombre est appelé une estimation ponctuelle du paramètre.

- $\bar{X}$ est un estimateur de la moyenne m :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- S^2 est un estimateur de la variance σ^2_{pop} :

$$S^2 = \frac{n}{n-1}\sigma^2_{ech}.$$

- f est un estimateur de la proportion p.

Estimation par intervalle de confiance

Estimation par IC

Les estimations ponctuelles, bien qu'utiles, ne fournissent aucune information concernant la précision des estimations, c'est-à-dire qu'elles ne tiennent pas compte de l'erreur possible dans l'estimation due aux fluctuations d'échantillonnage. La théorie des intervalles de confiance (IC) consiste à construire, autour de l'estimation ponctuelle, un intervalle qui aura une grande probabilité $(1 - \alpha)$ de contenir la vraie valeur du paramètre.

Estimation par intervalle de confiance

Estimation par IC

1) Estimation par IC de la moyenne m de la population

• 1^{er} **cas :** Lorsque la taille de l'échantillon est grande ($n \geqslant 30$) et l'écart type de la population de X est **connue**, on obtient un intervalle de confiance pour m au coefficient de confiance $(1-\alpha)$ de la forme :

$$I_\alpha = [\bar{x} - t\frac{\sigma_{pop}}{\sqrt{n}}; \bar{x} + t\frac{\sigma_{pop}}{\sqrt{n}}] \quad i.e., \quad P\Big(m \in I_\alpha\Big) = (1-\alpha).$$

Avec $2\Phi(t) = 1-\alpha$, où Φ est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$.
n est la taille de l'échantillon
α est le seuil de risque .

Estimation par intervalle de confiance

Estimation par IC

Ceci est aussi vrai pour de petits échantillons lorsque la variable aléatoire X suit une loi normale et que l'écart type de X est connue.
Lorsque la taille de l'échantillon est grande ($n \geqslant 30$) et l'écart type de la population de X est **inconnue**, on obtient un intervalle de confiance pour m au coefficient de confiance $(1-\alpha)$ de la forme :

$$I_\alpha = [\bar{x} - t\frac{S}{\sqrt{n}}; \bar{x} + t\frac{S}{\sqrt{n}}].$$

avec S est l'estimateur ponctuel de σ_{pop}

Estimation par intervalle de confiance

Exemple 1

On a observé la taille de $n = 200$ hommes marocains adultes. Après calcul , on a obtenu une moyenne de $\bar{x} = 168$ cm. Si on suppose que la variance connue vaut $\sigma^2 = 1$. Donnez un intervalle de confiance à 95% de la vraie moyenne de la population.

Corrigé :

Puisque $\alpha = 0.05$ (5%)(car 1-$\alpha = 0,95$),
alors $2\Phi(t) = 0.95$, donc $\Phi(t) = 0,475 = \Phi(1,96)$ (D'aprés le tableau de la loi normal)
par suite $t = 1.96$.
Finalement $I_{5\%} = [\bar{x} - t\frac{\sigma_{pop}}{\sqrt{n}}; \bar{x} + t\frac{\sigma_{pop}}{\sqrt{n}}] = [167.86; 168.14]$, i.e.,
$P(m \in [167.86; 168.14]) = 0.95$.

Estimation par intervalle de confiance

Estimation par IC

• 2^{me} **cas :** Lorsque la taille de l'échantillon est petite ($n < 30$) et X suit une loi normale de l'écart type **inconnue**, alors $\bar{X}$ suit une loi de student avec $n-1$ degré de liberté on obtient un intervalle de confiance pour m au coefficient de confiance $(1-\alpha)$ de la forme :

$$I_\alpha = [\bar{x} - z\frac{s}{\sqrt{n}}; \bar{x} + z\frac{s}{\sqrt{n}}]$$

z se déduit da la table student comme suit :

$$P(T > z) = \frac{\alpha}{2}.$$

Estimation par intervalle de confiance

Exemple 2

Un reporter pour un journal étudiant est en train de rédiger un article sur le coût du logement près du campus. Un échantillon de 10 appartements (trois et demi) dans un rayon de 1 km de l'université a permis d'estimer le coût moyen du loyer mensuel à 350 par mois et un écart type de 30. Quel est l'intervalle de confiance de 95% pour la moyenne des loyers mensuels ? Supposons que les loyers suivent une loi normale.

Corrigé :

pour un coefficient de confiance de $0,95$, on a $\alpha = 0,05$, et $\frac{\alpha}{2} = 0,025$. On a $n - 1 = 10 - 1 = 9$ degrés de liberté, alors la table de la distribution Student nous donne $z = 2,262$. Finalement $I_{5\%} = I_\alpha = [\bar{x} - z\frac{s}{\sqrt{n}}; \bar{x} + z\frac{s}{\sqrt{n}}] = [328.54; 371.46]$. i.e., nous sommes confiants à 95% que la moyenne des loyers mensuels (le vrai paramètre de la population m), se trouve entre 328.54 et 371.46.

Détermination de la taille de l'échantillon :

Quelle est la taille n de l'échantillon qui permettrait d'affirmer qu'en utilisant un estimateur ponctuel, l'erreur commise pour un coefficient de confiance $(1-\alpha)$ serait moindre que la marge d'erreur E ?
Si par exemple on fixe :

$$E = t\frac{\sigma}{\sqrt{n}},$$

l'erreur maximale commise pour un coefficient de confiance $(1-\alpha)$, alors la taille de l'échantillon sera :

$$n = \left[\frac{t.\sigma}{E}\right]^2.$$

Exemple

Si on fixe une marge d'erreur $E = 500$, alors pour un écart type de $\sigma = 5000$ et pour $t = 1.96$, on trouve $n = 384$. i.e., on a besoin d'un échantillon de taille $n = 384$ pour arriver à une précision de ± 500 à un seuil de confiance de 95%.

Estimation par IC de la proportion p de la population

Soit $2\Phi(t) = 1 - \alpha$. Lorsque n est grand ($n \geqslant 30$) et si f est la proportion échantillonnale alors un intervalle de confiance pour la proportion p inconnue de la population au coefficient de confiance $(1 - \alpha)$ de la forme :

$$I_\alpha = \left[f - t\sqrt{\frac{f(1-f)}{n}}; f + t\sqrt{\frac{f(1-f)}{n}} \right].$$

Sachant qu'on a remplacé p par son estimateur ponctuel. t est la solution de l'équation $2\Phi(t) = 1 - \alpha$ et n est la taille de l'échantillon.

Estimation par IC de la proportion p de la population

Exemple

SPI est une compagnie qui se spécialise dans les sondages politiques, à l'aide de sondages téléphoniques, les interviewers demandent aux citoyens pour qui ils voteraient si les élections avaient lieu aujourd'hui. Récemment, SPI a trouvé que 220 votants sur 500 voterait pour un candidat particulier. SPI veut estimer l'intervalle de confiance à 95% pour la proportion des votants qui sont en faveur de ce candidat.

Corrigé :

On a $n = 500$, $f = 220/500 = 0.44$ et $t = 1.96$ donc

$I_{5\%} = \left[f - t\sqrt{\frac{f(1-f)}{n}}; f + t\sqrt{\frac{f(1-f)}{n}} \right] = [0.3965; 0.4835]$, i.e., SPI est confiant à 95% que la proportion des votants qui favoriseront ce candidat est entre 0.3965 et 0.4835.