

Statistiques et analyses de données

Antoine Géré

Année 2026 - 2027¹



Ces notes sont en cours d'élaboration. Si vous avez la moindre question ou remarque ne pas hésiter à contacter par mail : a.gere@istom.fr.

*Ces cinq visages incarnent la statistique telle qu'elle se pratique aujourd'hui : non plus une affaire de formules abstraites, mais un outil pour comprendre et transformer le monde réel. **Hans Rosling** a montré, avec Gapminder, comment de simples données de santé et de développement peuvent renverser nos idées reçues sur les pays du Sud. **Esther Duflo**, prix Nobel d'économie, a fait de l'expérimentation rigoureuse, les essais randomisés, une méthode pour savoir ce qui marche vraiment contre la pauvreté. **Hadley Wickham** a bâti les outils (le tidyverse, en R) avec lesquels toute une génération apprend aujourd'hui à explorer et visualiser les données. **Cathy O'Neil** rappelle l'envers du décor : quand les algorithmes décident à notre place, ils peuvent aussi amplifier les injustices, d'où la nécessité d'un regard critique. Enfin, **David Spiegelhalter** s'attache à rendre l'incertitude et le risque compréhensibles, pour que chacun puisse lire un chiffre sans se laisser tromper. Ensemble, ils dessinent une discipline vivante, au service de la santé, de l'environnement et du développement.*

Résumé

coming.

Table des matières

1	Statistiques descriptives	2
1.1	Statistiques à une variable quantitative	2
1.2	Statistiques à une variable qualitative	3
1.3	Statistiques à deux variables qualitatives	3
1.4	Statistiques à deux variables quantitatives	3
1.5	Statistiques à une variable qualitative et une variable quantitative	3

¹version du 16 septembre 2026

2	Probabilités	3
2.1	Notion de probabilité	4
2.1.1	Espace des possibles et événements	4
2.1.2	Probabilité	5
2.1.3	Dénombrement	6
2.1.4	Conditionnement	8
2.1.5	Indépendance et répétitions	10
2.2	Variables aléatoires	11
2.2.1	Loi et fonction de répartition	12
2.2.2	Espérance, variance, écart-type	13
2.3	Lois usuelles	15
2.3.1	Lois discrètes	15
2.3.2	Lois continues	17
2.3.3	Synthèse : formulaire des lois usuelles	19
2.4	Théorèmes limites : de l'aléa à l'échantillon	20
2.5	Exercices d'application (corrigés)	21
3	Échantillonnage : de la population à l'échantillon	23
4	Estimations	23
4.1	Estimation ponctuelle : construire et choisir un estimateur	23
4.2	Estimation par intervalle de confiance	23
5	Tests d'hypothèse	23
5.1	Le degré de liberté pour les tests du χ^2 d'ajustement	24
5.1.1	Situation usuelle	24
5.1.2	Situation générale	24
5.2	Le degré de liberté pour les tests du χ^2 d'indépendance	25
5.2.1	La règle mathématique	25
5.2.2	Illustration	25
5.3	Le degré de liberté pour les tests du χ^2 d'homogénéité	26
5.4	Le degré de liberté pour l'ANOVA à 1 Facteur	26
5.4.1	Les degrés de liberté Inter-groupes, ddl_{Inter}	27
5.4.2	Les degrés de liberté Intra-groupes, ddl_{Intra}	27
5.4.3	Bilan	27
6	Analyse en Composantes Principales	28
7	Analyse factorielle des correspondances	28
8	Analyse des correspondances multiples	28
9	Classification Ascendante Hiérarchique	28
10	Analyse Factorielle Multiple	28
11	Modélisation	29

1 Statistiques descriptives

1.1 Statistiques à une variable quantitative

- Effectifs
- Fréquences
- Tableaux synthétiques

- Caractéristiques de position (mode(s), moyenne, médiane)
- Caractéristiques de dispersion (étendue, variance et écart-type, quartiles, déciles, écart interdéciles)
- Représentations graphiques
- Indicateur de concentration (masse, médiale, ...)

1.2 Statistiques à une variable qualitative

- Données brutes
- Tableaux synthétiques
- Représentations graphiques

1.3 Statistiques à deux variables qualitatives

coming.

1.4 Statistiques à deux variables quantitatives

coming.

1.5 Statistiques à une variable qualitative et une variable quantitative

coming.

2 Probabilités

Un ingénieur agronome décide toujours dans l'incertain. Le rendement change d'une parcelle à l'autre, la maladie frappe certains pieds et pas d'autres, la pluie de la prochaine saison reste inconnue. Face à cette variabilité, deux attitudes : la subir, ou la *modéliser*. Là où la statistique descriptive résume un échantillon déjà observé, les probabilités fournissent le langage mathématique du hasard : elles permettent de raisonner de l'échantillon vers la population et de *chiffrer le risque de se tromper*. C'est ce passage — de « j'ai compté 36 pieds de manioc atteints sur 200 » à « la prévalence réelle est comprise entre 13% et 23% » — qui fait tout l'intérêt du chapitre.

Ce chapitre pose les fondations de la statistique inférentielle, en trois mouvements. Il construit d'abord le *modèle probabiliste* — expérience aléatoire, événements, probabilité, conditionnement et formule de Bayes. Il introduit ensuite les *variables aléatoires* et leurs paramètres — loi, espérance, variance — puis les *lois usuelles*, discrètes (Bernoulli, binomiale, Poisson) et continues (uniforme, exponentielle, normale), avec lesquelles se modélisent un comptage de ravageurs, un taux de germination ou un rendement. Il s'achève sur les *théorèmes limites*, loi des grands nombres et théorème central limite, qui expliquent l'omniprésence de la loi normale et préparent l'estimation. Les notions sont illustrées par des exemples issus des agricultures des pays du Sud et accompagnées de leur mise en œuvre sous R.

Un pont avec le chapitre précédent. La statistique descriptive décrivait *ce qui a été observé* ; les probabilités donnent le langage de *ce qui peut être observé*. Le passage se fait presque terme à terme : en remplaçant la fréquence f_i par la probabilité p_i , la moyenne empirique devient l'**espérance**, la variance empirique la **variance théorique**, et les fréquences cumulées la **fonction de répartition**. Décrire un échantillon et modéliser une population sont ainsi les deux faces d'un même geste.

2.1 Notion de probabilité

2.1.1 Espace des possibles et événements

Définition 1 (Expérience aléatoire, univers).

Une **expérience aléatoire** peut être répétée, a plusieurs résultats possibles, et son résultat est imprévisible. L'**univers** (ou ensemble fondamental) Ω est l'ensemble de tous les résultats possibles (**événements élémentaires** ω).

Remarque (Nature de l'univers).

Selon l'expérience, Ω peut être :

- *fini* — l'état d'un **pied de manioc** vis-à-vis de la mosaïque africaine : $\Omega = \{\text{sain, malade}\}$ (2 issues) ;
- *infini dénombrable* — le **nombre de pièges à phéromones** posés avant la première capture d'une mouche des fruits : $\Omega = \{1, 2, 3, \dots\}$;
- *infini indénombrable* — le **rendement d'une parcelle de mil** (en t/ha), qui peut prendre n'importe quelle valeur d'un intervalle continu : $\Omega = [0, +\infty[$.

Un univers fini ou dénombrable est dit **discret**, un univers indénombrable est dit **continu**. Cette distinction gouverne toute la suite : variables aléatoires *discrètes* (loi donnée par des probabilités $P(X = x_i)$) contre variables *continues* (loi donnée par une densité).

Définition 2 (Événement).

Un **événement** A est une partie de Ω ; il est *réalisé* lorsqu'un des ω qui le composent l'est. Ω est l'événement **certain**, $\emptyset$ l'événement **impossible**.

Exemple 1 (Pieds de sorgho atteints par le charbon).

On prélève trois pieds de sorgho et l'on note ω le nombre de pieds atteints par le charbon : $\Omega = \{0, 1, 2, 3\}$. L'événement « au moins un pied atteint » est le sous-ensemble $A = \{1, 2, 3\}$, réalisé dès qu'on observe 1, 2 ou 3 pieds malades. Son contraire « aucun pied atteint » est $\bar{A} = \{0\}$, et « au plus deux pieds atteints » correspond à $\{0, 1, 2\}$.

Remarque (Vocabulaire).

Le vocabulaire des événements traduit celui des ensembles.

Notation	Ensembliste	Probabiliste
Ω	ensemble plein	événement certain
$\emptyset$	ensemble vide	événement impossible
$\omega \in A$	$\omega \in A$	ω réalise A
$A \subset B$	A inclus dans B	A implique B
$A \cup B$	réunion	A ou B
$A \cap B$	intersection	A et B
$\bar{A}$	complémentaire	événement contraire
$A \cap B = \emptyset$	disjoints	incompatibles

Propriété (Lois de De Morgan).

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \quad \text{et} \quad \overline{A \cap B} = \bar{A} \cup \bar{B}.$$

Exemple 2 (Double bio-agression sur un pied de maïs).

Soit $A =$ « le pied est attaqué par la chenille légionnaire » et $B =$ « le pied est attaqué par le foreur de tige ». Un pied *indemne* n'est attaqué ni par l'un ni par l'autre : c'est l'événement $\bar{A} \cap \bar{B}$, que la première loi de De Morgan identifie à $\overline{A \cup B}$ (le contraire de « attaqué par au moins un des deux »). De même, « le pied n'est pas attaqué par les deux à la fois » s'écrit $\overline{A \cap B} = \bar{A} \cup \bar{B}$ (« épargné par la légionnaire ou par le foreur »).

Définition 3 (Incompatibilité, système complet).

A et B sont **incompatibles** si $A \cap B = \emptyset$. Une famille (A_i) non vides, deux à deux incompatibles et de réunion Ω , forme un **système complet d'événements** (une partition de Ω).

Exemple 3 (État sanitaire d'un pied).

On observe un pied de manioc. Soit $A = \ll \text{atteint par la mosaïque} \gg$, $\bar{A} = \ll \text{sain} \gg$. $\{A, \bar{A}\}$ est un système complet. Sur trois pieds, « au moins un atteint » = « aucun atteint », ce qui simplifie souvent les calculs.

2.1.2 Probabilité

Définition 4 (Probabilité).

Une **probabilité** sur Ω est une application P qui, à tout événement A , associe un réel $P(A)$ vérifiant les trois axiomes :

1. **positivité** : $P(A) \geq 0$;
2. **normalisation** : $P(\Omega) = 1$;
3. **additivité** : pour toute suite d'événements (A_i) deux à deux incompatibles,

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i).$$

Sur un univers discret $\Omega = \{\omega_1, \omega_2, \dots\}$, il suffit de se donner les probabilités élémentaires $p_i = P(\omega_i) \geq 0$ telles que $\sum_i p_i = 1$; la probabilité d'un événement en découle par sommation :

$$P(A) = \sum_{\omega_i \in A} p_i.$$

Exemple 4 (État sanitaire d'un pied de mil).

On classe un pied de mil en trois états, qui forment un système complet d'événements : sain ($P = 0,7$), parasité par le striga ($P = 0,2$), attaqué par un foreur de tige ($P = 0,1$). On vérifie $0,7 + 0,2 + 0,1 = 1$. La probabilité qu'un pied soit attaqué (par l'un ou l'autre bio-agresseur), ces deux événements étant incompatibles, vaut $P = 0,2 + 0,1 = 0,3$.

Proposition 1 (Propriétés élémentaires).

$P(\bar{A}) = 1 - P(A)$; $P(\emptyset) = 0$; $P(A) \leq 1$; $A \subset B \Rightarrow P(A) \leq P(B)$; $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Démonstration.

$A, \bar{A}$ incompatibles de réunion Ω : $1 = P(A) + P(\bar{A})$. Pour la réunion, $A \cup B = A \cup (B \cap \bar{A})$ et $B = (B \cap A) \cup (B \cap \bar{A})$ (unions d'incompatibles) donnent $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Exemple 5 (Pratiques sur les parcelles d'une exploitation).

On tire au hasard une parcelle d'une exploitation. Soit I l'événement « la parcelle est irriguée » et F « la parcelle est fertilisée ». On connaît $P(I) = 0,4$, $P(F) = 0,5$ et $P(I \cap F) = 0,3$ (irriguée et fertilisée). Les propriétés élémentaires permettent d'en déduire, sans autre information :

- la probabilité qu'une parcelle ne soit *pas* irriguée, par passage au **contraire** : $P(\bar{I}) = 1 - P(I) = 0,6$;
- la probabilité qu'elle soit irriguée *ou* fertilisée, par la formule de la **réunion** : $P(I \cup F) = P(I) + P(F) - P(I \cap F) = 0,4 + 0,5 - 0,3 = 0,6$;
- la probabilité qu'elle ne reçoive *ni* irrigation *ni* fertilisation, contraire de la réunion (loi de De Morgan) : $P(\bar{I} \cap \bar{F}) = 1 - P(I \cup F) = 0,4$.

On vérifie au passage la **monotonie** : l'événement « irriguée et fertilisée » ($P = 0,3$) est inclus dans « fertilisée » ($P = 0,5$), et l'on a bien $0,3 \leq 0,5$.

Remarque (Probabilité et fréquence).

Sur un grand nombre d'observations, la fréquence relative tend vers la probabilité (loi des grands nombres) : c'est le fondement de l'estimation.

Exemple 6 (Croisement mendélien en sélection variétale).

On croise deux lignées de sorgho hétérozygotes Aa pour un caractère à dominance stricte (a récessif). La théorie prévoit 3/4 de phénotypes dominants, 1/4 de récessifs. Sur le terrain :

Phénotype	$n = 50$		$n = 1000$	
	effectif	fréq.	effectif	fréq.
dominant	36	0,72	755	0,755
récessif	14	0,28	245	0,245

La fréquence observée s'approche de la probabilité théorique 0,75 quand l'effectif croît.

Définition 5 (Probabilité uniforme).

Si Ω est fini et les événements élémentaires **équiprobables**, alors $P(\omega) = 1 / \text{card}(\Omega)$ et $P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} = \frac{\text{cas favorables}}{\text{cas possibles}}$: le calcul se ramène à du **dénombrement**.

Exemple 7 (Tirage d'une parcelle dans un essai).

Un essai comporte 12 parcelles élémentaires de même surface, dont 3 reçoivent le traitement témoin. On tire une parcelle au hasard (situation d'équiprobabilité). La probabilité de tomber sur une parcelle témoin est $P = \frac{\text{card}(A)}{\text{card}(\Omega)} = \frac{3}{12} = \frac{1}{4}$.

Travaux pratiques 1 (Stabilisation des fréquences).

Prévalence de la mosaïque du manioc $p = 0,2$.

```
p <- 0.2
obs <- sample(c(1, 0), 5000, replace = TRUE, prob = c(p, 1 - p))
freq <- cumsum(obs) / seq_along(obs)
plot(freq, type = "l", ylim = c(0, 0.4),
      xlab = "nombre de pieds observes", ylab = "frequence de pieds atteints")
abline(h = p, col = "red", lty = 2)
```

2.1.3 Dénombrement

Sous l'hypothèse d'équiprobabilité, calculer une probabilité revient à *compter* : combien de cas favorables, combien de cas possibles. Le dénombrement fournit les formules de ce comptage. Tout repose sur une même question — « de combien de façons peut-on tirer k éléments parmi n ? » — dont la réponse dépend de deux choix : l'ordre des éléments compte-t-il, et un même élément peut-il être tiré plusieurs fois (remise) ?

Propriété (Principe multiplicatif).

Si une opération se décompose en p étapes successives offrant respectivement $n_1, n_2, \dots, n_p$ choix, indépendants les uns des autres, alors le nombre total de résultats possibles est le produit $n_1 \times n_2 \times \dots \times n_p$.

Exemple 8 (Composer un itinéraire technique).

Un essai croise 3 variétés, 4 doses d'azote et 2 dates de semis. Le nombre d'itinéraires techniques distincts à comparer vaut, par le principe multiplicatif, $3 \times 4 \times 2 = 24$. C'est ce même principe qui engendre toutes les formules qui suivent : le n^k d'un tirage ordonné avec remise, comme le produit $n(n-1) \dots (n-k+1)$ d'un arrangement.

Tirage de k parmi n	avec remise	sans remise
ordonné (l'ordre compte)	n^k	$A_n^k = \frac{n!}{(n-k)!}$
non ordonné (l'ordre est ignoré)	$\binom{n+k-1}{k}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Les deux cas *sans remise* sont de loin les plus fréquents en agronomie (on ne prélève pas deux fois le même pied, la même parcelle, la même semence).

Définition 6 (Factorielle et permutations).

La **factorielle** de n est $n! = n \times (n-1) \times \dots \times 2 \times 1$, avec la convention $0! = 1$. Le nombre de façons d'**ordonner** (permuter) n éléments distincts vaut $n!$.

Exemple 9 (Calendrier d'opérations culturales).

On planifie l'ordre de 4 opérations sur une parcelle — labour, semis, sarclage, récolte. Chacune occupant une position distincte, il existe $4! = 24$ enchaînements possibles.

Définition 7 (Arrangements).

Un **arrangement** est un tirage *ordonné* et *sans remise* de k éléments parmi n . Leur nombre est

$$A_n^k = \frac{n!}{(n-k)!} = n \times (n-1) \times \cdots \times (n-k+1).$$

Exemple 10 (Ordre de semis — arrangement).

Sur une parcelle conduite *en relais* — c'est-à-dire où une culture est semée avant la récolte de la précédente, de sorte que les cycles se chevauchent dans le temps sur la même parcelle — on fixe l'*ordre* de semis de 3 cultures choisies parmi {mil, sorgho, niébé, arachide, coton} ($n = 5$). L'ordre comptant et sans répétition, il y a $A_5^3 = 5 \times 4 \times 3 = 60$ séquences possibles. Par exemple : (mil $\rightarrow$ niébé $\rightarrow$ sorgho), (arachide $\rightarrow$ mil $\rightarrow$ coton) ou (coton $\rightarrow$ sorgho $\rightarrow$ niébé). Comme l'ordre importe, (mil $\rightarrow$ niébé $\rightarrow$ sorgho) et (sorgho $\rightarrow$ niébé $\rightarrow$ mil) comptent pour deux séquences distinctes.

Définition 8 (Combinaisons).

Une **combinaison** est un tirage *non ordonné* et *sans remise* de k éléments parmi n (un sous-ensemble de taille k). Leur nombre est le **coefficient binomial**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{A_n^k}{k!}.$$

Exemple 11 (Choix de variétés — combinaison).

Pour un essai variétal, on retient 3 variétés de riz NERICA parmi les 8 disponibles, sans se soucier de l'ordre : il y a $\binom{8}{3} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$ choix possibles.

Propriété (Coefficients binomiaux).

- **cas particuliers** : $\binom{n}{0} = \binom{n}{n} = 1$ et $\binom{n}{1} = n$;
- **symétrie** : $\binom{n}{k} = \binom{n}{n-k}$;
- **relation de Pascal** : $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$;
- **somme** : $\sum_{k=0}^n \binom{n}{k} = 2^n$ (nombre total de sous-ensembles d'un ensemble à n éléments) ;
- **formule du binôme** : $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$ — socle de la loi binomiale.

Ces coefficients se lisent directement dans le **triangle de Pascal**, où chaque terme est la somme des deux situés au-dessus de lui (relation de Pascal) ; la ligne n donne les $\binom{n}{k}$ pour $k = 0, \dots, n$:

$n \backslash k$	0	1	2	3	4	5	...
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

$$\binom{n-1}{k-1} + \binom{n-1}{k} = \binom{n}{k}$$

Exemple 12 (Calculer un coefficient binomial).

Les propriétés évitent de manipuler de grandes factorielles. Par symétrie, $\binom{8}{6} = \binom{8}{2} = \frac{8 \times 7}{2} = 28$. Et la ligne $n = 5$ du triangle de Pascal se lit directement : $\binom{5}{0}, \dots, \binom{5}{5} = 1, 5, 10, 10, 5, 1$.

Exemple 13 (Probabilité par dénombrement).

Un lot contient 20 sachets de semences, dont 5 sont périmés et 15 encore bons. Pour un contrôle, on en prélève 4 *simultanément* — donc sans remise, et l'ordre du prélèvement n'a pas d'importance. Quelle est la probabilité d'y trouver *exactement un* sachet périmé ?

Cas possibles. Un résultat est un choix de 4 sachets parmi 20, sans ordre : ce sont des **combinaisons**, toutes équiprobables. Leur nombre est

$$\text{card}(\Omega) = \binom{20}{4} = \frac{20 \times 19 \times 18 \times 17}{4 \times 3 \times 2 \times 1} = 4845.$$

Cas favorables. Obtenir exactement un périmé, c'est en choisir 1 parmi les 5 périmés et compléter avec 3 sachets parmi les 15 bons ; par le principe multiplicatif,

$$\binom{5}{1} \times \binom{15}{3} = 5 \times 455 = 2275.$$

Probabilité. Sous équiprobabilité, c'est le rapport des deux :

$$P(\text{exactement un périmé}) = \frac{\text{cas favorables}}{\text{cas possibles}} = \frac{2275}{4845} \approx 0,47.$$

On dénombre ici avec des combinaisons parce que le prélèvement est simultané : seul le *contenu* du lot de 4 compte, pas l'ordre de sortie. Ce schéma — un tirage sans remise dans une population partagée en deux catégories (périmés / bons) — est exactement celui de la *loi hypergéométrique*, étudiée plus loin.

```
Code 1 (Dénombrement sous R).      factorial(5)      # 5! = 120
choose(8, 3)      # nombre de combinaisons C(8,3) = 56
prod(5:3)         # arrangement A_5^3 = 5*4*3 = 60
```

2.1.4 Conditionnement

Définition 9 (Probabilité conditionnelle).

Pour $P(B) > 0$, la probabilité de A sachant B est $P(A | B) = \frac{P(A \cap B)}{P(B)}$. L'application $P(\cdot | B)$ est une probabilité (elle en a toutes les propriétés).

Exemple 14 (L'information modifie la probabilité).

Sur 200 pieds de manioc observés, 60 se trouvent en bordure de parcelle, et parmi ces derniers 18 sont infectés par la mosaïque. La probabilité qu'un pied soit infecté *sachant* qu'il est en bordure est

$$P(I \mid \text{bordure}) = \frac{P(I \cap \text{bordure})}{P(\text{bordure})} = \frac{18/200}{60/200} = \frac{18}{60} = 0,30,$$

à comparer à la prévalence sur l'ensemble du champ : conditionner par « bordure » change la probabilité.

Remarque (Pièges classiques).

$A \mid B$ n'est *pas* un événement (il n'existe pas d'« événement conditionnel »); en général $P(A \mid B) \neq P(B \mid A)$; ne pas confondre $P(A \mid B)$ et $P(A \cap B)$.

Proposition 2 (Probabilités composées).

Pour deux événements A et B de probabilités non nulles,

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

Plus généralement, pour n événements (règle des chaînes) :

$$P(A_1 \cap \dots \cap A_n) = P(A_1) P(A_2 \mid A_1) P(A_3 \mid A_1 \cap A_2) \dots P(A_n \mid A_1 \cap \dots \cap A_{n-1}).$$

Exemple 15 (Deux semences traitées, tirage sans remise).

Un sachet contient 6 semences traitées et 4 non traitées. On en tire deux successivement, sans remise. En notant T_1, T_2 les événements « la 1^{re} (resp. 2^e) semence est traitée », la formule des probabilités composées donne

$$P(T_1 \cap T_2) = P(T_1) P(T_2 \mid T_1) = \frac{6}{10} \times \frac{5}{9} = \frac{30}{90} = \frac{1}{3}.$$

Après avoir tiré une traitée, il n'en reste que 5 sur 9 : c'est tout le sens de la probabilité conditionnelle $P(T_2 \mid T_1)$.

Exemple 16 (Système complet : un terroir par type de sol).

Les parcelles d'un terroir se répartissent en trois types de sol, exclusifs et recouvrant l'ensemble : sableux (A_1), argileux (A_2), limoneux (A_3), avec $P(A_1) = 0,5$, $P(A_2) = 0,3$, $P(A_3) = 0,2$. Comme les A_i sont deux à deux incompatibles et que $0,5 + 0,3 + 0,2 = 1$, la famille $\{A_1, A_2, A_3\}$ est une partition de l'univers des parcelles.

Proposition 3 (Probabilités totales).

Soit (A_i) un système complet d'événements (partition de Ω , cf. définition 3) avec $P(A_i) > 0$. Alors, pour tout événement B ,

$$P(B) = \sum_i P(B \mid A_i) P(A_i).$$

Démonstration.

$B = \bigcup_i (B \cap A_i)$, réunion d'incompatibles, donc $P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i) P(A_i)$.

Exemple 17 (Rendement moyen sur un terroir hétérogène).

On reprend la partition par type de sol ci-dessus. La probabilité qu'une parcelle donne un *bon rendement* dépend du sol : $P(B \mid A_1) = 0,3$ (sableux), $P(B \mid A_2) = 0,5$ (argileux), $P(B \mid A_3) = 0,7$ (limoneux). Pour une parcelle prise au hasard, la formule des probabilités totales donne

$$P(B) = 0,3 \times 0,5 + 0,5 \times 0,3 + 0,7 \times 0,2 = 0,15 + 0,15 + 0,14 = 0,44.$$

Remarque (Deux méthodes de calcul).

Deux outils très efficaces : l'**arbre de probabilités** (au 1^{er} niveau les probabilités simples $P(A_i)$, au 2^e niveau les conditionnelles $P(B \mid A_i)$; on multiplie le long des branches) et le **tableau croisé** des probabilités conjointes $P(A_i \cap B)$, dont on complète les marges.

Proposition 4 (Formule de Bayes).

Sous les mêmes hypothèses, pour $P(B) > 0$ et tout indice j ,

$$P(A_j \mid B) = \frac{P(A_j) P(B \mid A_j)}{\sum_i P(A_i) P(B \mid A_i)}.$$

Elle permet de « remonter » de l'effet observé B vers sa cause A_j .

Remarque (Tests de diagnostic).

Pour un test de détection (d'un virus, d'un ravageur) : la **sensibilité** est $P(+ | \text{malade})$, la **spécificité** $P(- | \text{sain})$; la **valeur prédictive positive** $VPP = P(\text{malade} | +)$ et la **valeur prédictive négative** $VPN = P(\text{sain} | -)$ s'obtiennent par Bayes. La VPP dépend fortement de la prévalence.

Exemple 18 (Test de détection du virus de la mosaïque du manioc).

Sensibilité $P(+ | M) = 0,95$, spécificité $P(- | \bar{M}) = 0,90$, prévalence $P(M) = 0,05$. Alors

$$VPP = P(M | +) = \frac{0,95 \cdot 0,05}{0,95 \cdot 0,05 + 0,10 \cdot 0,95} \approx 0,33.$$

Un pied déclaré positif n'a qu'une chance sur trois d'être réellement infecté : prudence avant tout arrachage systématique.

2.1.5 Indépendance et répétitions

Définition 10 (Indépendance).

Soit deux événements A et B d'un même univers Ω , de probabilités non nulles. On dit que A et B sont **indépendants** lorsque $P(A | B) = P(A)$ — autrement dit, savoir que B est réalisé ne change pas la probabilité de A . Cette condition équivaut à $P(A \cap B) = P(A)P(B)$ et à $P(B | A) = P(B)$.

Exemple 19 (Germination de deux espèces).

On sème une graine de maïs et une graine de niébé issues de lots distincts : la levée de l'une n'influe pas sur celle de l'autre. En notant G_M (« le maïs lève », $P = 0,9$) et G_N (« le niébé lève », $P = 0,8$), l'indépendance donne directement la probabilité que les deux lèvent :

$$P(G_M \cap G_N) = P(G_M)P(G_N) = 0,9 \times 0,8 = 0,72.$$

Proposition 5.

Si A et B sont indépendants, alors A et $\bar{B}$, $\bar{A}$ et B , $\bar{A}$ et $\bar{B}$ le sont aussi.

Exemple 20 (Passage au contraire).

Avec G_M ($P = 0,9$) et G_N ($P = 0,8$) indépendants, G_M et $\bar{G}_N$ le sont aussi. La probabilité que le maïs lève mais pas le niébé est donc

$$P(G_M \cap \bar{G}_N) = P(G_M)P(\bar{G}_N) = 0,9 \times 0,2 = 0,18.$$

Remarque.

Incompatible $\neq$ indépendant. Deux événements incompatibles de probabilités non nulles ne sont jamais indépendants : $A \cap B = \emptyset \Rightarrow P(A \cap B) = 0 \neq P(A)P(B)$.

Définition 11 (Indépendance mutuelle).

Soit n événements $A_1, \dots, A_n$. Ils sont :

- **indépendants deux à deux** si

$$P(A_i \cap A_j) = P(A_i)P(A_j) \quad \text{pour tout } i \neq j;$$

- **mutuellement indépendants** si, pour toute sous-famille $J \subset \{1, \dots, n\}$,

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j).$$

La seconde condition est *plus forte* : pour trois événements A, B, C , elle réclame non seulement les trois égalités deux à deux, mais aussi $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Exemple 21 (Trois pieds prélevés indépendamment).

Dans un champ de prévalence $p = 0,1$, on prélève 3 pieds indépendamment. Les événements M_1, M_2, M_3 (« le pied i est malade ») sont mutuellement indépendants : la probabilité de toute intersection est le produit des probabilités. Ainsi

$$P(M_1 \cap M_2 \cap M_3) = p^3 = 0,001, \quad P(M_1 \cap M_2) = p^2 = 0,01.$$

Proposition 6 (Épreuves indépendantes).

Répéter N fois, indépendamment, une même expérience (Ω, P) produit l'univers Ω^N muni de $P^N((\omega_1, \dots, \omega_N)) = P(\omega_1) \cdots P(\omega_N)$.

Exemple 22 (Cinq semis successifs).

On répète 5 fois, indépendamment, le semis d'une graine de taux de levée $p = 0,8$. L'univers est $\{\text{lève, ne lève pas}\}^5$, et la probabilité que les 5 graines lèvent est

$$P = 0,8^5 \approx 0,33.$$

C'est ce cadre d'épreuves indépendantes répétées qui conduira au schéma de Bernoulli et à la loi binomiale.

Remarque (Vers l'échantillon i.i.d.).

Un **échantillon** $X_1, \dots, X_n$ est exactement n épreuves indépendantes d'une même expérience : des variables *indépendantes et identiquement distribuées* (i.i.d.). Le chapitre d'estimation en dépend entièrement.

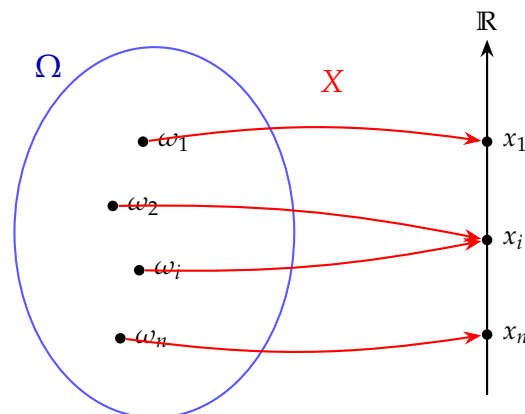
2.2 Variables aléatoires

Définition 12 (Variable aléatoire).

Une **variable aléatoire** (v.a.) est une application

$$X : \Omega \longrightarrow \mathbb{R}$$

qui associe à chaque résultat ω de l'expérience un nombre réel $X(\omega)$. La valeur $x = X(\omega)$ effectivement obtenue est une **réalisation** de X . Une v.a. est **discrète** lorsqu'elle prend un nombre fini ou dénombrable de valeurs, et **continue** lorsqu'elle en prend un continuum.



À chaque résultat $\omega \in \Omega$, la variable aléatoire X associe une valeur $X(\omega) \in \mathbb{R}$; plusieurs résultats peuvent partager la même image (ici ω_2 et ω_i ont pour image x_i).

Exemple 23 (Nombre de pieds parasités).

On prélève 3 pieds de mil ; un résultat est un triplet $\omega = (r_1, r_2, r_3)$ donnant l'état de chaque pied, *sain* (S) ou *parasité* (P) par le striga. On pose $X =$ « nombre de pieds parasités », à valeurs dans $X(\Omega) = \{0, 1, 2, 3\}$. Par exemple :

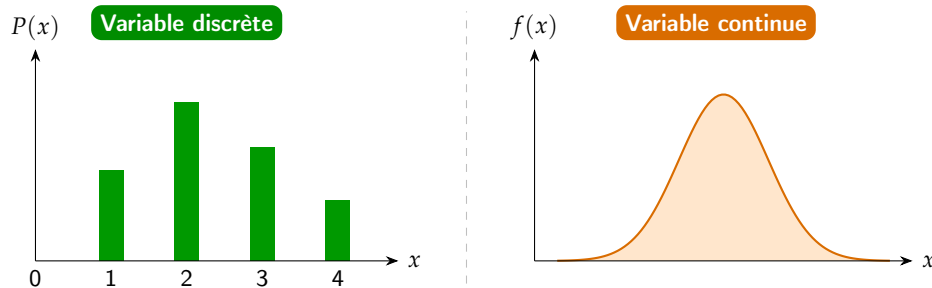
$$X(S, S, S) = 0, \quad X(P, S, S) = 1, \quad X(P, S, P) = 2, \quad X(P, P, P) = 3.$$

Les trois configurations à un seul pied parasité — (P, S, S) , (S, P, S) , (S, S, P) — ont toutes pour image 1 : X n'est pas injective. C'est une v.a. **discrète** ; le *rendement* de la parcelle (en t/ha) serait, lui, une v.a. **continue**.

Remarque.

On note les variables aléatoires par des majuscules $(X, Y, \dots)$ et leurs réalisations par des minuscules $(x, y, \dots)$. Toute combinaison de v.a. est encore une v.a. : si X et Y en sont, alors $X + Y$, XY et λX ($\lambda \in \mathbb{R}$) le sont aussi.

2.2.1 Loi et fonction de répartition



Une variable **discrète** prend des valeurs dénombrables $(0, 1, 2, \dots)$: sa loi se lit sur un diagramme en bâtons $P(x)$. Une variable **continue** peut prendre toute valeur d'un intervalle : sa loi est décrite par une densité $f(x)$, dont l'aire sous la courbe vaut 1.

Définition 13 (Loi d'une v.a. discrète).

La **loi** d'une v.a. discrète X , à valeurs $x_1, x_2, \dots$, est la donnée des probabilités

$$p_i = P(X = x_i), \quad \text{avec } p_i \geq 0 \quad \text{et} \quad \sum_i p_i = 1.$$

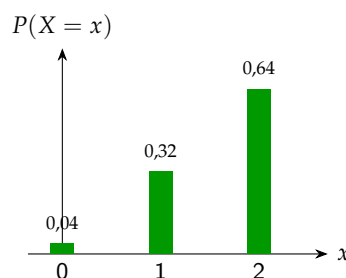
On la représente par un diagramme en bâtons.

Exemple 24 (Loi du nombre de graines germées).

On sème 2 graines de taux de germination $p = 0,8$, indépendamment. La v.a. $X = \ll \text{nombre de graines germées} \gg$ prend les valeurs $\{0, 1, 2\}$, de loi

$$P(X = 0) = 0,2^2 = 0,04, \quad P(X = 1) = 2 \times 0,8 \times 0,2 = 0,32, \quad P(X = 2) = 0,8^2 = 0,64,$$

et l'on vérifie bien $0,04 + 0,32 + 0,64 = 1$.



Définition 14 (Fonction de répartition).

La **fonction de répartition** de X est

$$F(t) = P(X \leq t), \quad t \in \mathbb{R}.$$

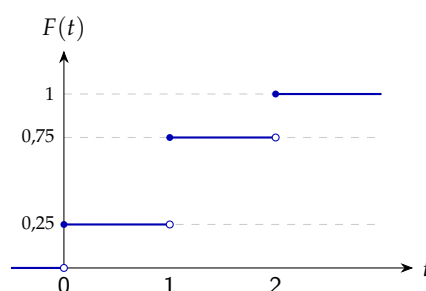
Elle est croissante, avec $\lim_{t \rightarrow -\infty} F(t) = 0$ et $\lim_{t \rightarrow +\infty} F(t) = 1$. Pour une v.a. discrète, F est en escalier et $P(X = x_i) = F(x_i) - F(x_{i-1})$; pour une v.a. continue, F est continue.

Exemple 25 (Fonction de répartition d'un croisement).

Dans une descendance de 2 plantes issues d'un croisement, $X = \ll \text{nombre de plantes au phénotype récessif (sensible)} \gg$

» prend les valeurs $\{0, 1, 2\}$ de loi $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. Sa fonction de répartition est en escalier :

$$F(t) = \begin{cases} 0 & \text{si } t < 0, \\ 0,25 & \text{si } 0 \leq t < 1, \\ 0,75 & \text{si } 1 \leq t < 2, \\ 1 & \text{si } t \geq 2. \end{cases}$$



Définition 15 (Densité d'une v.a. continue).

Une v.a. continue X admet une **densité** f , vérifiant

$$f(x) \geq 0 \quad \text{et} \quad \int_{\mathbb{R}} f(x) dx = 1.$$

Les probabilités se lisent alors comme des *aires* sous la courbe :

$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a), \quad F(x) = \int_{-\infty}^x f(t) dt.$$

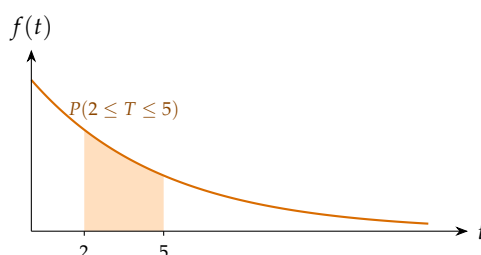
En particulier $P(X = x) = 0$: seuls les intervalles portent une probabilité non nulle.

Exemple 26 (Densité : durée avant germination).

La durée T (en jours) avant la germination d'une graine se modélise souvent par la densité $f(t) = \lambda e^{-\lambda t}$ pour $t \geq 0$ (loi exponentielle), ici avec $\lambda = 0,2$. La probabilité que la graine germe entre le 2^e et le 5^e jour est

$$P(2 \leq T \leq 5) = \int_2^5 0,2 e^{-0,2t} dt = e^{-0,4} - e^{-1} \approx 0,30,$$

tandis que $P(T = 3) = 0$: pour une v.a. continue, seuls les intervalles portent une probabilité non nulle.



2.2.2 Espérance, variance, écart-type

Définition 16 (Espérance).

L'**espérance** (ou moyenne théorique) de X est

$$E(X) = \sum_i x_i p_i \quad (\text{cas discret}), \quad E(X) = \int_{\mathbb{R}} x f(x) dx \quad (\text{cas continu}).$$

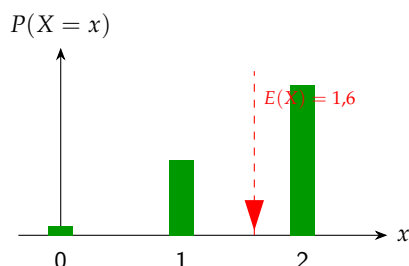
Plus généralement, pour une fonction g , $E(g(X)) = \sum_i g(x_i) p_i$ (resp. $\int g(x) f(x) dx$).

Exemple 27 (Espérance du nombre de graines germées).

Pour la v.a. X (germination de 2 graines, $p = 0,8$) de loi $0,04 / 0,32 / 0,64$,

$$E(X) = 0 \times 0,04 + 1 \times 0,32 + 2 \times 0,64 = 1,6.$$

En moyenne, 1,6 graine sur 2 lève; on retrouve $E(X) = np$. L'espérance est le *centre de gravité* de la loi (triangle rouge ci-dessous).



Proposition 7 (Linéarité de l'espérance).

Pour tous réels a, b et toutes v.a. X, Y :

$$E(aX + b) = a E(X) + b, \quad E(X + Y) = E(X) + E(Y).$$

Exemple 28 (Transformation et somme).

Avec $E(X) = 1,6$: si $Y = 3X + 1$, alors $E(Y) = 3 \times 1,6 + 1 = 5,8$; et pour deux parcelles indépendantes de même loi, le total $X_1 + X_2$ a pour espérance $E(X_1) + E(X_2) = 3,2$.

Définition 17 (Variance et écart-type).

La **variance** mesure la dispersion de X autour de son espérance :

$$V(X) = E[(X - E(X))^2] = E(X^2) - (E(X))^2 \quad (\text{König-Huygens}).$$

L'**écart-type** $\sigma(X) = \sqrt{V(X)}$ s'exprime dans la même unité que X .

Exemple 29 (Variance du nombre de graines germées).

Avec $E(X) = 1,6$ et $E(X^2) = 0^2 \cdot 0,04 + 1^2 \cdot 0,32 + 2^2 \cdot 0,64 = 2,88$,

$$V(X) = 2,88 - 1,6^2 = 0,32, \quad \sigma(X) = \sqrt{0,32} \approx 0,57,$$

en accord avec $V(X) = np(1-p) = 2 \times 0,8 \times 0,2$.

Proposition 8 (Propriétés de la variance).

Pour tous réels a, b :

$$V(aX + b) = a^2 V(X), \quad V(X + Y) = V(X) + V(Y) + 2 \text{cov}(X, Y).$$

Si X et Y sont indépendantes, alors $\text{cov}(X, Y) = 0$ (la réciproque est fausse) et $V(X + Y) = V(X) + V(Y)$; plus généralement, pour des v.a. indépendantes, $V(\sum_i X_i) = \sum_i V(X_i)$.

Exemple 30 (Effet d'une transformation et d'une somme).

Pour $Y = 3X + 1$: $V(Y) = 3^2 V(X) = 9 \times 0,32 = 2,88$ (l'ajout de la constante 1 ne change rien). Pour deux parcelles indépendantes, $V(X_1 + X_2) = 0,32 + 0,32 = 0,64$.

Définition 18 (Variable centrée réduite).

La **variable centrée réduite** associée à X (d'écart-type non nul) est

$$Z = \frac{X - E(X)}{\sigma(X)}, \quad \text{et alors} \quad E(Z) = 0, \quad V(Z) = 1.$$

Démonstration.

$$E(Z) = \frac{1}{\sigma} (E(X) - E(X)) = 0; V(Z) = \frac{1}{\sigma^2} V(X - E(X)) = \frac{1}{\sigma^2} \sigma^2 = 1.$$

Exemple 31 (Centrer-réduire).

Pour X (germination), $Z = \frac{X - 1,6}{0,57}$. La réalisation $X = 2$ correspond à $z = \frac{2 - 1,6}{0,57} \approx 0,70$: elle se situe à environ 0,70 écart-type au-dessus de la moyenne.

Remarque.

La propriété $V(\sum_i X_i) = \sum_i V(X_i)$ (indépendance) donne $V(\bar{X}_n) = \sigma^2/n$: c'est elle qui explique qu'une moyenne sur davantage de parcelles est plus précise.

2.3 Lois usuelles

2.3.1 Lois discrètes

Presque toutes les lois discrètes usuelles se construisent à partir d'une même brique, l'**épreuve de Bernoulli** : une expérience à deux issues, le *succès* (probabilité p) et l'*échec* (probabilité $1 - p$). Selon que l'on compte les succès sur un nombre fixé d'épreuves, le rang du premier succès, ou des événements rares, on obtient les lois binomiale, géométrique et de Poisson.

Définition 19 (Loi de Bernoulli).

La **loi de Bernoulli** $\mathcal{B}(p)$ ($p \in [0, 1]$) est celle de la v.a. X valant 1 en cas de succès et 0 en cas d'échec :

$$P(X = 1) = p, \quad P(X = 0) = 1 - p, \quad E(X) = p, \quad V(X) = p(1 - p).$$

Exemple 32.

Un pied de mil est parasité par le striga (succès, $p = 0,2$) ou sain. Son indicatrice suit $\mathcal{B}(0,2)$, d'espérance 0,2 et de variance $0,2 \times 0,8 = 0,16$.

Définition 20 (Loi binomiale).

En répétant n épreuves de Bernoulli *indépendantes* de même paramètre p , la v.a. X comptant le nombre de succès suit la **loi binomiale** $\mathcal{B}(n, p)$:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \{0, \dots, n\}, \quad E(X) = np, \quad V(X) = np(1 - p).$$

Démonstration.

On écrit $X = \sum_{i=1}^n X_i$, somme de n Bernoulli indépendantes : la linéarité donne $E(X) = np$, et l'indépendance $V(X) = \sum_i V(X_i) = np(1 - p)$.

Exemple 33 (Germination d'un lot).

Sur $n = 20$ semences de sorgho de taux de germination $p = 0,8$, le nombre de graines levées suit $\mathcal{B}(20; 0,8)$: on attend $E(X) = 16$ levées, avec $V(X) = 20 \times 0,8 \times 0,2 = 3,2$ (sous R : `dbinom`, `pbinom`).

Définition 21 (Loi hypergéométrique).

Pour un tirage *sans remise* de n éléments dans une population de taille N dont K sont des « succès », le nombre X de succès tirés suit la **loi hypergéométrique** :

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad E(X) = n \frac{K}{N}.$$

Lorsque N est grand devant n , elle est proche de $\mathcal{B}(n, K/N)$.

Exemple 34 (Contrôle d'un lot de semences).

Un sac de $N = 100$ graines d'arachide contient $K = 15$ graines contaminées par l'aflatoxine ; sur $n = 10$ prélevées sans remise, le nombre de contaminées suit cette loi (sous R : `dhyper`).

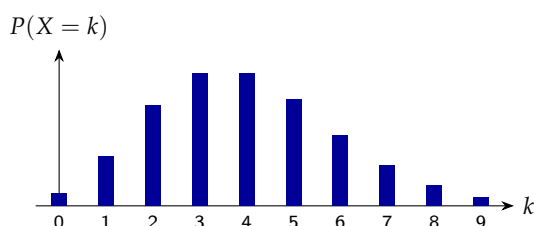
Définition 22 (Loi de Poisson).

La **loi de Poisson** $\mathcal{P}(\lambda)$ ($\lambda > 0$) modélise le nombre d'événements rares survenant dans un intervalle de temps, d'espace ou de surface :

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

Exemple 35 (Comptages au champ).

Le nombre de mouches des fruits capturées par piège dans un verger de manguiers, de moyenne $\lambda = 4$ par jour, suit $\mathcal{P}(4)$: par exemple $P(X = 0) = e^{-4} \approx 0,018$ (aucune capture).



Proposition 9 (Loi des événements rares).

La loi de Poisson est la limite de la binomiale quand n est grand et p petit : si $X_n \sim \mathcal{B}(n, \lambda/n)$, alors $P(X_n = k) \rightarrow e^{-\lambda} \lambda^k / k!$. En pratique, on remplace $\mathcal{B}(n, p)$ par $\mathcal{P}(np)$ dès que $n \geq 50$ et $\lambda = np \leq 5$.

Exemple 36 (De la binomiale à la Poisson).

Sur $n = 200$ pieds de manioc, la mosaïque touche chacun avec une faible probabilité $p = 0,02$. Le nombre d'atteints suit $\mathcal{B}(200; 0,02)$, bien approchée par $\mathcal{P}(\lambda)$ avec $\lambda = np = 4$: sans calcul de factorielles de grands nombres, $P(X = 0) \approx e^{-4} = 0,018$ et $P(X = 3) \approx e^{-4} \frac{4^3}{3!} = 0,195$.

Définition 23 (Loi géométrique).

En répétant indéfiniment une épreuve de Bernoulli de paramètre $p \in]0, 1[$, la v.a. X égale au **rang du premier succès** suit la **loi géométrique** $\mathcal{G}(p)$:

$$P(X = k) = (1 - p)^{k-1} p, \quad k \geq 1, \quad E(X) = \frac{1}{p}, \quad V(X) = \frac{1 - p}{p^2}.$$

Exemple 37.

Nombre de pièges à phéromones posés avant la première capture d'un ravageur, chaque piège capturant avec probabilité $p = 0,3$: en moyenne $E(X) = 1/0,3 \approx 3,3$ pièges.

Proposition 10 (Absence de mémoire).

$P(X > k) = (1 - p)^k$ et, pour tous entiers $k, \ell \geq 0$, $P(X > k + \ell \mid X > k) = P(X > \ell)$: la loi géométrique est « sans mémoire ».

Exemple 38 (Sans mémoire).

Chaque piège capture avec $p = 0,3$. Sachant que les deux premiers pièges n'ont rien pris, la probabilité qu'il en faille encore plus de 3 est la même qu'au départ : $P(X > 3) = (1 - p)^3 = 0,7^3 \approx 0,34$. Les échecs passés n'« usent » pas la chance future.

Loi	$P(X = k)$	$E(X)$	$V(X)$	Exemple (Sud)
Bernoulli $\mathcal{B}(p)$	$p^k (1 - p)^{1-k}$	p	$p(1 - p)$	pied parasité ou non
Binomiale $\mathcal{B}(n, p)$	$\binom{n}{k} p^k (1 - p)^{n-k}$	np	$np(1 - p)$	graines germées sur n
Poisson $\mathcal{P}(\lambda)$	$e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ	captures par piège
Géométrique $\mathcal{G}(p)$	$(1 - p)^{k-1} p$	$\frac{1}{p}$	$\frac{1 - p}{p^2}$	1 ^{er} piège qui capture

2.3.2 Lois continues

Une v.a. continue est décrite par sa densité f (d'aire totale 1) : les probabilités sont des *aires* sous la courbe. Trois lois sont fondamentales — uniforme, exponentielle et surtout normale — auxquelles s'ajoutent les lois qui en dérivent et servent à l'estimation.

Loi uniforme.

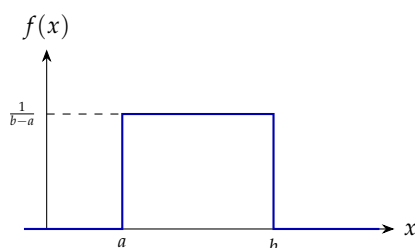
Définition 24 (Loi uniforme).

$X \sim \mathcal{U}([a, b])$ a une densité *constante* sur $[a, b]$:

$$f(x) = \frac{1}{b-a} \text{ sur } [a, b] \text{ (0 ailleurs),} \quad F(x) = \frac{x-a}{b-a} \text{ sur } [a, b],$$

$$E(X) = \frac{a+b}{2}, \quad V(X) = \frac{(b-a)^2}{12}.$$

Toutes les valeurs de l'intervalle sont « également probables ».



Exemple 39.

Le générateur aléatoire de R (`runif`) tire selon $\mathcal{U}([0, 1])$; on l'utilise notamment pour *randomiser* l'affectation des traitements aux parcelles d'un essai.

Loi exponentielle.

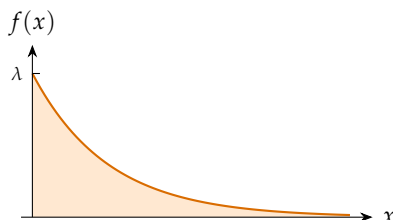
Définition 25 (Loi exponentielle).

$X \sim \mathcal{E}(\lambda)$ ($\lambda > 0$) a pour densité et fonction de répartition

$$f(x) = \lambda e^{-\lambda x} \text{ (} x \geq 0 \text{),} \quad F(x) = 1 - e^{-\lambda x} \text{ (} x \geq 0 \text{),}$$

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}.$$

C'est l'équivalent continu de la loi géométrique : elle modélise le *temps d'attente* jusqu'au premier événement d'un processus de Poisson.



Proposition 11 (Absence de mémoire).

Pour tous $s, t \geq 0$, $P(X > s+t \mid X > s) = P(X > t)$.

Démonstration.

Comme $P(X > t) = e^{-\lambda t}$, $\frac{P(X > s+t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$.

Exemple 40.

Temps d'attente entre deux averses en saison des pluies au Sahel, ou durée de bon fonctionnement d'une motopompe d'irrigation avant panne. Si les pannes surviennent au rythme $\lambda = 0,1$ par jour, la durée moyenne sans panne est $E(X) = 10$ jours.

Exemple 41 (Illustration de l'absence de mémoire).

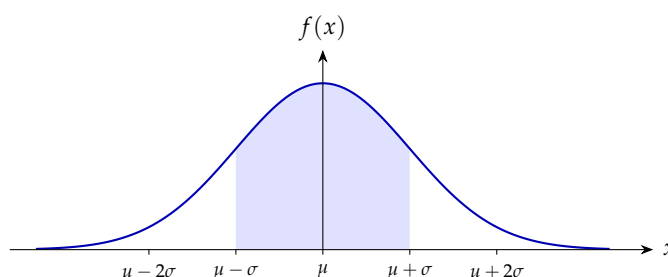
Pour cette motopompe ($\lambda = 0,1$), sachant qu'elle a déjà fonctionné 30 jours sans panne, la probabilité qu'elle tienne encore au moins 10 jours est $P(X > 10) = e^{-1} \approx 0,37$ — exactement comme si elle était neuve.

Loi normale.

Définition 26 (Loi normale (Laplace–Gauss)).

$X \sim \mathcal{N}(\mu, \sigma^2)$ a pour densité

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad E(X) = \mu, \quad V(X) = \sigma^2.$$



Exemple 42 (Rendement de mil).

Le rendement de mil entre exploitations d'un terroir suit $\mathcal{N}(900; 150^2)$ (en kg/ha). La proportion d'exploitations sous 700 kg/ha est $P(X < 700) = \Phi\left(\frac{700-900}{150}\right) = \Phi(-1,33) \approx 0,09$.

Propriété.

La densité est symétrique par rapport à $x = \mu$; son maximum $\frac{1}{\sigma\sqrt{2\pi}}$ est atteint en μ ; elle présente deux points d'inflexion en $\mu \pm \sigma$, et est d'autant plus resserrée que σ est petit.

Définition 27 (Loi normale centrée réduite).

La variable centrée réduite $Z = \frac{X - \mu}{\sigma}$ suit $\mathcal{N}(0, 1)$, de fonction de répartition notée Φ (tabulée). On a

$$\Phi(-z) = 1 - \Phi(z), \quad P(Z > z) = 1 - \Phi(z), \quad P(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

Exemple 43 (Lecture de la table).

Pour $X \sim \mathcal{N}(30, 3^2)$: $P(X \leq 33) = \Phi\left(\frac{33-30}{3}\right) = \Phi(1) = 0,8413$, et $P(27 \leq X \leq 33) = 2\Phi(1) - 1 = 0,683$.

Propriété (Règle des 68–95–99,7).

$$P(\mu \pm \sigma) \approx 0,683, \quad P(\mu \pm 2\sigma) \approx 0,954, \quad P(\mu \pm 3\sigma) \approx 0,997,$$

d'où l'intervalle central à 95% : $\mu \pm 1,96\sigma$ ($qnorm(0.975) = 1,96$).

Propriété (Combinaison de normales).

Toute combinaison affine $aX + b$ d'une normale est normale, et une somme de normales indépendantes est normale :

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \text{ indép.} \implies \sum_i a_i X_i \sim \mathcal{N}\left(\sum_i a_i \mu_i, \sum_i a_i^2 \sigma_i^2\right).$$

En particulier $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ pour un échantillon gaussien.

Proposition 12 (Approximation de la binomiale).

Pour n grand, la binomiale s'approche par une normale de mêmes moyenne et variance :

$$\mathcal{B}(n, p) \approx \mathcal{N}(np, np(1-p)) \quad \text{si } n \geq 20, np \geq 10, n(1-p) \geq 10.$$

Exemple 44 (Plants hors calibre).

Sur $n = 100$ plants de cacao dont 10% sont hors calibre, $X \sim \mathcal{B}(100; 0,1)$. Comme $np = 10$ et $n(1-p) = 90$, on approche par $\mathcal{N}(10; 9)$ (soit $\sigma = 3$), d'où $P(X \geq 16) \approx P(Z \geq \frac{16-10}{3}) = 1 - \Phi(2) \approx 0,023$.

Code 2 (Loi normale sous R).

```
# Rendement de mil : mu = 900 kg/ha, sigma = 150
pnorm(700, 900, 150) # P(rendement < 700)
qnorm(0.975, 900, 150) # borne haute d'un intervalle central a 95%
```

Lois dérivées de la normale.

Définition 28 (Chi-deux, Student, Fisher).

- **Chi-deux** χ_k^2 : somme des carrés de k normales standard indépendantes ; $E = k$, $V = 2k$ (sert à l'IC d'une variance).
- **Student** $t_k = \frac{Y}{\sqrt{Z/k}}$ ($Y \sim \mathcal{N}(0,1)$, $Z \sim \chi_k^2$ indépendantes) : symétrique, plus étalée que la normale, et $t_k \rightarrow \mathcal{N}(0,1)$ quand $k \rightarrow \infty$ (IC d'une moyenne à σ inconnu).
- **Fisher** F_{k_1, k_2} : rapport de deux χ^2 normalisés (comparaison de variances, ANOVA).

Exemple 45 (Où elles apparaissent).

Sur un essai à $n = 10$ parcelles de rendements gaussiens : l'incertitude sur la *variance* des rendements fait intervenir un χ_9^2 ; la comparaison du rendement *moyen* à une norme, à σ inconnu, se lit sur un t_9 ; et la comparaison des variances de *deux* traitements, sur une loi de Fisher. Ces trois lois reviendront aux chapitres d'estimation et de tests.

2.3.3 Synthèse : formulaire des lois usuelles

Les tableaux ci-dessous rassemblent, pour chaque loi, sa fonction de masse ou sa densité, son espérance et sa variance, puis son *cadre d'utilisation* — la situation qu'elle modélise.

Lois discrètes			
Loi	$P(X = k)$	$E(X)$	$V(X)$
Bernoulli $\mathcal{B}(p)$	$p^k(1-p)^{1-k}$, $k \in \{0, 1\}$	p	$p(1-p)$
Binomiale $\mathcal{B}(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ
Géométrique $\mathcal{G}(p)$	$(1-p)^{k-1} p$, $k \geq 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hypergéom. $\mathcal{H}(N, K, n)$	$\frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$	$n \frac{K}{N}$	$n \frac{K}{N} \left(1 - \frac{K}{N}\right) \frac{N-n}{N-1}$

Lois continues

Loi	densité $f(x)$	$E(X)$	$V(X)$
Uniforme $\mathcal{U}([a, b])$	$\frac{1}{b-a}$ sur $[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle $\mathcal{E}(\lambda)$	$\lambda e^{-\lambda x}$ ($x \geq 0$)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normale $\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2
Chi-deux χ_k^2	—	k	$2k$
Student t_k	—	0 ($k > 1$)	$\frac{k}{k-2}$ ($k > 2$)
Fisher F_{k_1, k_2}	—	$\frac{k_2}{k_2-2}$ ($k_2 > 2$)	—

Cadre d'utilisation

Loi	Situation modélisée
Bernoulli $\mathcal{B}(p)$	une épreuve à deux issues, succès / échec : présence d'un caractère, réussite d'un test.
Binomiale $\mathcal{B}(n, p)$	nombre de succès sur n épreuves indépendantes de même probabilité (tirage avec remise).
Poisson $\mathcal{P}(\lambda)$	nombre d'événements <i>rare</i> s dans un intervalle de temps, d'espace ou de surface (comptages).
Géométrique $\mathcal{G}(p)$	rang du <i>premier</i> succès ; temps d'attente discret.
Hypergéom. $\mathcal{H}(N, K, n)$	nombre de succès lors d'un tirage <i>sans remise</i> dans une population finie (contrôle de lot).
Uniforme $\mathcal{U}([a, b])$	valeur tirée « au hasard », toutes également probables sur $[a, b]$: randomisation, simulation.
Exponentielle $\mathcal{E}(\lambda)$	temps d'attente entre deux événements d'un processus de Poisson ; durée de vie sans mémoire.
Normale $\mathcal{N}(\mu, \sigma^2)$	grandeur continue résultant de nombreuses causes (mesures, rendements) ; loi limite (TCL).
Chi-deux χ_k^2	somme de carrés de normales ; inférence sur une <i>variance</i> .
Student t_k	moyenne d'un <i>petit</i> échantillon à écart-type inconnu.
Fisher F_{k_1, k_2}	comparaison de deux variances (ANOVA).

Remarque (Sommes de v.a. indépendantes).

$\mathcal{B}(m, p) + \mathcal{B}(n, p) = \mathcal{B}(m+n, p)$; $\mathcal{P}(\lambda) + \mathcal{P}(\mu) = \mathcal{P}(\lambda + \mu)$; $\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. La somme des carrés de k variables $\mathcal{N}(0, 1)$ indépendantes est un χ_k^2 .

Remarque (Approximations (seuils indicatifs)).

$$\mathcal{H}(N, K, n) \approx \mathcal{B}(n, \frac{K}{N}) \quad (N \geq 10n), \quad \mathcal{B}(n, p) \approx \mathcal{P}(np) \quad (n \geq 30, p \leq 0,1),$$

$$\mathcal{B}(n, p) \approx \mathcal{N}(np, np(1-p)) \quad (np \geq 5, n(1-p) \geq 5), \quad \mathcal{P}(\lambda) \approx \mathcal{N}(\lambda, \lambda) \quad (\lambda \geq 20).$$

2.4 Théorèmes limites : de l'aléa à l'échantillon

Proposition 13 (Markov, puis Bienaymé–Tchebychev).

Pour $X \geq 0$ et $a > 0$: $P(X \geq a) \leq E(X)/a$. Pour X d'espérance μ et variance σ^2 , et $h > 0$: $P(|X - \mu| \geq h) \leq \sigma^2/h^2$.

Démonstration.

Bienaymé–Tchebychev s'obtient en appliquant Markov à $(X - \mu)^2 \geq 0$ au seuil h^2 : $P((X - \mu)^2 \geq h^2) \leq E[(X - \mu)^2]/h^2 = \sigma^2/h^2$.

Théorème 1 (Loi des grands nombres).

Pour un échantillon i.i.d. d'espérance μ , la moyenne empirique $\bar{X}_n$ converge vers μ quand $n \rightarrow \infty$: le rendement moyen mesuré sur de nombreuses parcelles approche le vrai rendement moyen du terroir.

Théorème 2 (Théorème central limite).

Pour un échantillon i.i.d. d'espérance μ et de variance σ^2 , pour n grand ($n \geq 30$ en pratique),

$$S_n = \sum_i X_i \approx \mathcal{N}(n\mu, n\sigma^2), \quad \bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

quelle que soit la loi mère. Le TCL explique l'omniprésence de la loi normale.

Propriété (Approximations).

$\mathcal{B}(n, p) \rightarrow \mathcal{P}(np)$ si $n \geq 30$, $p \leq 0,1$; $\mathcal{B}(n, p) \rightarrow \mathcal{N}(np, np(1-p))$ si $n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$;
 $\mathcal{P}(\lambda) \rightarrow \mathcal{N}(\lambda, \lambda)$ si $\lambda \geq 20$.

Travaux pratiques 2 (Illustrer LGN et TCL).

```
set.seed(1)
pop <- function(n) rexp(n, rate = 1/12) # loi mere asymetrique, moy 12 t/ha
moyennes <- replicate(2000, mean(pop(40))) # moyenne de n = 40 parcelles
hist(moyennes, breaks = 30, freq = FALSE,
     main = "Distribution de la moyenne (n = 40)")
curve(dnorm(x, 12, 12/sqrt(40)), add = TRUE, col = "red")
```

Loi	Exemple (Sud)	$E(X)$	$V(X)$
Bernoulli $\mathcal{B}(p)$	pied de mil parasité ou non	p	$p(1-p)$
Binomiale $\mathcal{B}(n, p)$	graines de sorgho germées	np	$np(1-p)$
Hypergéom.	lot de semences (sans remise)	$n \frac{K}{N}$	—
Poisson $\mathcal{P}(\lambda)$	captures de mouches des fruits	λ	λ
Géométrique $\mathcal{G}(p)$	1 ^{er} piège qui capture	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Uniforme $\mathcal{U}(a, b)$	tirage « au hasard »	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle $\mathcal{E}(\lambda)$	attente entre deux averses	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normale $\mathcal{N}(\mu, \sigma^2)$	rendement de mil	μ	σ^2

2.5 Exercices d'application (corrigés)

Ces exercices reprennent, dans l'ordre du chapitre, les notions étudiées : dénombrement, conditionnement et Bayes, indépendance, lois discrètes et continues, puis approximations. Les corrections suivent chaque énoncé.

Exercice 1 (Dénombrement).

Pour un essai comparatif, on dispose de 6 variétés de riz.

- (1) De combien de façons peut-on en retenir 3 (sans tenir compte de l'ordre) ?
- (2) Si l'on fixe en plus l'ordre de semis de ces 3 variétés sur une parcelle en relais, combien de plans obtient-on ?

Correction.

- (1) $\binom{6}{3} = 20$ (combinaison).
- (2) $A_6^3 = 6 \times 5 \times 4 = 120$ (arrangement).

Exercice 2 (Probabilité et lois de De Morgan).

Dans un terroir, 45% des exploitations pratiquent l'agriculture de conservation (C), 30% utilisent des engrais minéraux (E), 10% font les deux. Quelle est la probabilité qu'une exploitation ne fasse ni l'un ni l'autre ?

Correction. Par De Morgan, $\bar{C} \cap \bar{E} = \overline{C \cup E}$, donc $P(\bar{C} \cap \bar{E}) = 1 - P(C \cup E) = 1 - (0,45 + 0,30 - 0,10) = 0,35$.

Exercice 3 (Test de diagnostic (Bayes)).

Un test rapide de détection de l'aflatoxine sur un lot d'arachide a une sensibilité $P(+ | C) = 0,90$ et une spécificité $P(- | \bar{C}) = 0,85$; la prévalence de lots contaminés est $P(C) = 0,10$. Un lot est déclaré positif : quelle est la probabilité qu'il soit réellement contaminé (VPP) ?

Correction. Par la formule de Bayes,

$$VPP = P(C | +) = \frac{0,90 \times 0,10}{0,90 \times 0,10 + 0,15 \times 0,90} = \frac{0,09}{0,225} = 0,40.$$

Malgré un bon test, moins d'un lot positif sur deux est réellement contaminé (effet de la faible prévalence).

Exercice 4 (Indépendance).

On sème une graine de maïs (levée avec probabilité 0,9) et une graine de niébé (levée avec probabilité 0,8), indépendamment. Probabilité que

- (1) les deux graines lèvent ;
- (2) au moins une lève ?

Correction.

- (1) $0,9 \times 0,8 = 0,72$.
- (2) $1 - (0,1 \times 0,2) = 1 - 0,02 = 0,98$.

Exercice 5 (Loi binomiale).

Le striga parasite chaque pied de mil avec probabilité $p = 0,2$, indépendamment. Sur $n = 15$ pieds, soit X le nombre de pieds parasités.

- (1) Loi, espérance et variance de X ?
- (2) $P(X = 0)$ et $P(X \geq 1)$?

Correction.

- (1) $X \sim \mathcal{B}(15; 0,2)$, $E(X) = 3$, $V(X) = 15 \times 0,2 \times 0,8 = 2,4$.
- (2) $P(X = 0) = 0,8^{15} \approx 0,035$, donc $P(X \geq 1) \approx 0,965$.

Exercice 6 (Loi de Poisson).

Un piège capture en moyenne $\lambda = 3$ mouches des fruits par jour.

- (1) $P(\text{aucune capture})$?
- (2) $P(\text{au moins 2 captures})$?
- (3) Ce comptage résulte d'un très grand nombre de mouches ayant chacune une faible probabilité d'être piégée : quelle loi discrète cela justifie-t-il ?

Correction.

- (1) $P(X = 0) = e^{-3} \approx 0,050$.
- (2) $P(X \geq 2) = 1 - e^{-3}(1 + 3) = 1 - 4e^{-3} \approx 0,80$.
- (3) la loi de Poisson (limite de la binomiale pour des événements rares).

Exercice 7 (Loi géométrique).

Chaque piège capture un ravageur avec probabilité $p = 0,3$, indépendamment. Soit X le rang du premier piège qui capture.

- (1) $P(X = 3)$?
- (2) $P(X > 4)$?
- (3) Nombre moyen de pièges nécessaires ?

Correction.

- (1) $P(X = 3) = 0,7^2 \times 0,3 = 0,147$.
- (2) $P(X > 4) = 0,7^4 \approx 0,240$.
- (3) $E(X) = 1/0,3 \approx 3,3$ pièges.

Exercice 8 (Loi exponentielle).

La durée de bon fonctionnement d'une motopompe suit une loi exponentielle de paramètre $\lambda = 0,1$ par jour.

- (1) Durée moyenne sans panne ?
- (2) $P(\text{elle dépasse 15 jours})$?
- (3) Sachant qu'elle a déjà tenu 20 jours, $P(\text{qu'elle tienne encore 15 jours})$?

Correction.

- (1) $E(X) = 1/\lambda = 10$ jours.
- (2) $P(X > 15) = e^{-1,5} \approx 0,22$.
- (3) Par absence de mémoire, la réponse est la même qu'en (2) : $e^{-1,5} \approx 0,22$.

Exercice 9 (Loi normale et approximation).

Le poids d'un régime de banane plantain suit $\mathcal{N}(18; 3^2)$ (en kg).

- (1) $P(\text{poids} > 21)$?
- (2) Quel poids dépassent les 10% de régimes les plus lourds ?
- (3) Sur une plantation, 12% des plants sont chétifs; sur $n = 150$ plants, en approchant la binomiale par une normale, estimer $P(\text{plus de 25 chétifs})$.

Correction.

- (1) $P(X > 21) = 1 - \Phi\left(\frac{21-18}{3}\right) = 1 - \Phi(1) = 0,159$.
- (2) $x = 18 + 1,28 \times 3 \approx 21,8$ kg ($\text{qnorm}(0.90)$).
- (3) $Y \sim \mathcal{B}(150; 0,12)$, $np = 18$, $n(1-p) = 132$, $\sigma = \sqrt{18 \times 0,88} \approx 3,98$; d'où $P(Y > 25) \approx P(Z > \frac{25-18}{3,98}) = 1 - \Phi(1,76) \approx 0,039$.

3 Échantillonnage : de la population à l'échantillon

coming.

4 Estimations

4.1 Estimation ponctuelle : construire et choisir un estimateur

coming.

4.2 Estimation par intervalle de confiance

coming.

5 Tests d'hypothèse

coming.

5.1 Le degré de liberté pour les tests du χ^2 d'ajustement

Dans un test du χ^2 d'ajustement, le **degré de liberté** représente le nombre de modalités dont l'effectif peut varier librement avant que l'ensemble du système ne soit mathématiquement bloqué par les contraintes de l'expérimentation ou de l'enquête.

5.1.1 Situation usuelle

Imaginons un agronome travaillant avec une coopérative de producteurs au Burkina Faso pour s'adapter à la baisse des précipitations. Il met en place un essai comparatif avec 3 variétés de sorgho ($k = 3$) :

- *Variété A* : Locale traditionnelle (cycle long).
- *Variété B* : Améliorée précoce (cycle court).
- *Variété C* : Hybride résistante au stress hydrique.

L'enveloppe budgétaire allouée par le projet permet de financer l'installation de parcelles de démonstration chez 30 paysans au total ($N = 30$). L'agronome doit répartir ces parcelles :

- Pour la variété A, il a un choix total : il décide de l'installer chez 12 paysans.
- Pour la variété B, il conserve une marge de manœuvre : il choisit de l'attribuer à 10 paysans.
- Pour la variété C ... **il n'a plus aucune liberté de choix**. Le total étant fixé à 30, la variété C est obligatoirement affectée aux $30 - (12 + 10) = 8$ paysans restants.

Règle du $k - 1$

Sur les k modalités initiales, l'expérimentateur n'a pu choisir librement l'effectif que pour $k - 1$ d'entre elles. La dernière est mathématiquement verrouillée par l'effectif total N .

$$ddl = k - 1$$

5.1.2 Situation générale

Dans un test d'ajustement du χ^2 , la formule standard des degrés de liberté est

$$ddl = k - 1$$

où k est le nombre de modalités comparées et 1 représente la contrainte du total général fixe. Cependant, cette relation suppose que la loi théorique de référence est parfaitement connue à l'avance.

Dans des situations plus complexes, l'agronome veut tester si ses données de terrain s'ajustent à une loi mathématique théorique (loi de Poisson, loi Normale, loi Binomiale) dont **il ne connaît pas les paramètres exacts au départ**.

Pour construire son tableau d'effectifs théoriques, il est alors obligé d'utiliser ses propres données observées pour **estimer** ces paramètres manquants. Chaque paramètre estimé (m) ajoute une contrainte mathématique au système et supprime une « liberté » supplémentaire. La relation générale devient alors :

$$ddl = k - 1 - m \quad (1)$$

Exemple concret

Un entomologiste compte le nombre de larves d'insectes par pied de canne à sucre sur un échantillon de 200 plants. Il classe ses observations en $k = 5$ catégories : 0 larve, 1 larve, 2 larves, 3 larves, et 4 larves ou plus. Il souhaite vérifier si la répartition de ces parasites suit une **loi de Poisson**.

Le problème : Pour calculer les effectifs théoriques de la loi de Poisson, la formule mathématique exige de connaître la moyenne théorique (λ) du nombre de larves par pied. L'entomologiste ne la connaît pas.

L'estimation : Il utilise donc son échantillon de terrain pour calculer la moyenne observée. Imaginons qu'il trouve une moyenne de 1,2 larve par pied. Il utilise cette valeur unique comme estimation de λ .

- Il y a $k = 5$ catégories d'infestation.
- On retire 1 degré de liberté pour la contrainte du total des 200 plants.
- On retire 1 degré de liberté car on a extrait la moyenne des données pour utiliser le modèle théorique.

Le calcul final des degrés de liberté pour ce test est donc :

$$ddl = k - 1 - m = 5 - 1 - 1 = 3$$

5.2 Le degré de liberté pour les tests du χ^2 d'indépendance

Contrairement au test d'ajustement qui étudie une seule variable, le **test du χ^2 d'indépendance** analyse la relation entre **deux variables qualitatives**. Les données sont alors croisées dans un tableau à double entrée appelé **tableau de contingence**.

5.2.1 La règle mathématique

Pour un tableau de contingence comportant L lignes et C colonnes, la relation mathématique des degrés de liberté est :

$$ddl = (L - 1) \times (C - 1) \quad (2)$$

Cette relation représente le nombre de cases intérieures du tableau que l'on peut remplir de manière totalement indépendante avant que les totaux des lignes et des colonnes ne verrouillent automatiquement le reste des valeurs.

5.2.2 Illustration

Contexte

Un agronome réalise une enquête auprès de **100 caféiculteurs** au Honduras ($N = 100$). Il cherche à savoir si le **type de fertilisation** appliqué aux caféiers (Variable 1, en lignes) influence la **qualité commerciale du café** récolté (Variable 2, en colonnes).

Le tableau est structuré avec :

- **3 lignes** ($L = 3$) : Compost seul, Engrais chimique, Pratique mixte.
- **3 colonnes** ($C = 3$) : Standard, Premium, Spécialité.

Avant de commencer le calcul du χ^2 , l'agronome calcule les totaux pour chaque ligne et chaque colonne à partir des données de son enquête. Ces totaux (appelés **les marges**) deviennent des **contraintes fixes et inviolables** pour la suite du calcul.

Remplissons le tableau ligne par ligne pour observer le moment exact où le système se bloque mathématiquement :

- **Sur la ligne « Compost » (Ligne 1)** : L'agronome commence à remplir les cases.
 - Il est **libre** de saisir une valeur pour la qualité *Standard* (ex : 10),
 - puis pour la qualité *Premium* (ex : 15).

- En revanche, la troisième case (*Spécialité*) est **bloquée** : le total de la ligne devant impérativement faire 40, sa valeur est obligatoirement $40 - (10 + 15) = 15$.

- **Sur la ligne « Chimique » (Ligne 2) :**

- Il est à nouveau **libre** d'attribuer des valeurs pour les deux premières colonnes (ex : 20 et 5).
- La troisième case est immédiatement **bloquée** par la contrainte de ligne ($35 - (20 + 5) = 10$).

- **Sur la ligne « Mixte » (Ligne 3) : Toutes les cases sont déjà bloquées.** En effet, les totaux de chaque colonne ont été fixés au départ.

- Pour la colonne *Standard*, le total visé est 45. Ayant déjà posé 10 et 20 sur les lignes précédentes, la dernière case vaut obligatoirement $45 - (10 + 20) = 15$.
- Il en va de même pour toutes les autres colonnes de cette ligne.

Voici la cartographie du tableau de contingence de l'enquête. Les cases vertes représentent les zones d'expression libre du hasard ou de l'agronome. Les cases rouges représentent les cellules mathématiquement contraintes par les marges.

Fertilisation \ Qualité	Standard	Premium	Spécialité	Total Lignes (Fixe)
Compost	10 (Libre)	15 (Libre)	15 (Bloqué)	40
Chimique	20 (Libre)	5 (Libre)	10 (Bloqué)	35
Pratique mixte	15 (Bloqué)	10 (Bloqué)	0 (Bloqué)	25
Total Colonnes (Fixe)	45	30	25	100 (N)

La zone de liberté forme un rectangle parfait de $2 \times 2 = 4$ cases en haut à gauche du tableau. C'est exactement le résultat fourni par notre relation :

$$ddl = (L - 1) \times (C - 1) = (3 - 1) \times (3 - 1) = 2 \times 2 = 4$$

5.3 Le degré de liberté pour les tests du χ^2 d'homogénéité

Le **test du χ^2 d'homogénéité** utilise la même configuration technique (tableau de contingence à double entrée) que le test d'indépendance. La seule différence réside dans le protocole de collecte : au lieu de tirer un seul grand échantillon, l'expérimentateur fixe à l'avance la taille de plusieurs sous-échantillons (groupes) pour vérifier si leurs comportements sont similaires.

Pour un tableau comportant L lignes (les groupes comparés) et C colonnes (les variables observées), la formule des degrés de liberté reste :

$$ddl = (L - 1) \times (C - 1) \quad (3)$$

5.4 Le degré de liberté pour l'ANOVA à 1 Facteur

Dans le cadre d'une **ANOVA à 1 facteur** (Analyse de Variance), la notion de degré de liberté (ddl) s'applique différemment par rapport au test du χ^2 . L'ANOVA n'analyse pas des fréquences, mais cherche à comparer deux types de variabilités (variances) qui s'affrontent :

1. La variabilité **Inter-groupes** (due à l'effet du facteur).
2. La variabilité **Intra-groupes** (due à l'erreur résiduelle).

Pour effectuer cette comparaison, le système sépare mathématiquement les degrés de liberté totaux en deux sous-catégories distinctes.

Contexte pour l'illustration

Un agronome teste l'effet de la taille des branches sur le rendement de cacaoyers au Guatemala. Il dispose d'un échantillon total de $N = 12$ **parcelles homogènes**, réparties équitablement en $k = 3$ **groupes** expérimentaux (Taille Mensuelle, Taille Trimestrielle, et un Témoin sans taille). Chaque groupe possède donc un effectif de $n = 4$ parcelles de répétition.

5.4.1 Les degrés de liberté Inter-groupes, ddl_{Inter}

Les ddl_{Inter} sont directement associés à la **variabilité entre les groupes**. Ils mesurent combien de moyennes de traitements peuvent varier librement avant de verrouiller le système.

La relation est définie par :

$$ddl_{\text{Inter}} = k - 1 \quad (4)$$

La contrainte réside dans le fait que la moyenne des k groupes doit obligatoirement être ancrée autour de la **moyenne générale** globale de l'essai.

- La moyenne observée du Groupe A (Taille mensuelle) est libre (ex : 10 kg).
- La moyenne observée du Groupe B (Taille trimestrielle) est également libre (ex : 14 kg).
- La moyenne du Groupe C (Témoin)... **est automatiquement bloquée**. Sa valeur est mathématiquement forcée pour respecter la valeur de la moyenne générale.

Avec 3 groupes à comparer, le système ne dispose que de $3 - 1 = 2$ informations indépendantes.

5.4.2 Les degrés de liberté Intra-groupes, ddl_{Intra}

Les ddl_{Intra} (ou degrés de liberté de l'erreur résiduels) mesurent la **variabilité au sein de chaque groupe**. Ils comptabilisent la flexibilité des répétitions individuelles soumises à un *même* traitement.

La relation s'écrit :

$$ddl_{\text{Intra}} = N - k \quad (5)$$

Chaque groupe de l'essai fonctionne comme un sous-système fermé ayant sa propre contrainte : sa moyenne interne.

- Dans le Groupe A (comprenant $n = 4$ parcelles), la moyenne interne est fixée à 10 kg. Les 3 premières parcelles peuvent afficher des rendements aléatoires dictés par la nature (ex : 11, 9 et 12 kg). La 4^e parcelle perd toute liberté : elle est mathématiquement forcée (à 8 kg) pour que la moyenne du groupe reste stable. Chaque sous-groupe perd ainsi exactement 1 degré de liberté.
- Liberté résiduelle dans le Groupe A : $4 - 1 = 3$ parcelles libres.
- Liberté résiduelle dans le Groupe B : $4 - 1 = 3$ parcelles libres.
- Liberté résiduelle dans le Groupe C : $4 - 1 = 3$ parcelles libres.

Sur l'ensemble du terrain, le hasard intra-groupe a pu s'exprimer sur $3 + 3 + 3 = 9$ parcelles indépendantes. L'application confirme le modèle : $N - k = 12 - 3 = 9$.

5.4.3 Bilan

L'ANOVA réalise une décomposition de la liberté globale du système :

$$ddl_{\text{Total}} = ddl_{\text{Inter}} + ddl_{\text{Intra}} \implies 11 = 2 + 9$$

Tableau de structure d'une ANOVA à 1 facteur

Source de variation	Somme des Carrés	Degrés de liberté (<i>ddl</i>)	Carré Moyen (<i>CM</i>)	Statistique <i>F</i>
Inter-groupes (Facteur)	SCE_{Inter}	$k - 1 = 2$	$CM_{\text{Inter}} = \frac{SCE_{\text{Inter}}}{2}$	$F_{\text{obs}} = \frac{CM_{\text{Inter}}}{CM_{\text{Intra}}}$
Intra-groupes (Erreur)	SCE_{Intra}	$N - k = 9$	$CM_{\text{Intra}} = \frac{SCE_{\text{Intra}}}{9}$	
Totale	SCE_{Totale}	$N - 1 = 11$		

Une Somme des Carrés (*SCE*) augmente artificiellement avec le nombre d'observations. Pour obtenir une variance pure et comparable (appelée **Carré Moyen**), le statisticien doit impérativement diviser chaque *SCE* par son propre nombre de degrés de liberté.

La statistique finale F_{obs} est le rapport de ces deux variances épurées.

6 Analyse en Composantes Principales

ACP
coming.

7 Analyse factorielle des correspondances

AFC
coming.

8 Analyse des correspondances multiples

ACM
coming.

9 Classification Ascendante Hiérarchique

CAH
coming.

10 Analyse Factorielle Multiple

AFM
coming.

11 Modélisation

coming.

Modélisation par la Régression Linéaire Multiple (RLM)

Modèle Logit (Régression logistique)



Étudiant(e) (Nom, Prénom) :

Promotion, groupe :

Email :