Marc-Henry Beaucamps

L3
M1-M2

Comprendre la statistique inférentielle

Fondements, méthodes et interprétation



ellipses

Références sciences

Comprendre la statistique inférentielle

Fondements, méthodes et interprétation

Marc-Henry Beaucamps

ellipses

Collection Références sciences

dirigée par Paul de Laboulaye
paul.delaboulaye@editions-ellipses.fr

Retrouvez tous les livres de la collection et des extraits sur www.editions-ellipses.fr

À Coline, Ombeline, Norah, …

ISBN 9782340-121690
Dépôt légal : novembre 2026
©Ellipses Édition Marketing S.A.
8/10 rue la Quintinie 75015 Paris

DANGER LE PHOTOCOPILLAGE TUE LE LIVRE

Le Code de la propriété intellectuelle et artistique n'autorisant, aux termes des alinéas 2 et 3 de l'article L. 122-5, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective » et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite » (alinéa 1er de l'article L. 122-4).
Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

www.editions-ellipses.fr

AVANT-PROPOS

Dans un contexte où l'on est amené à expérimenter de nouvelles méthodes de production, tester l'efficacité d'un traitement médical, comparer des procédés industriels, mesurer la satisfaction de clients, suivre un taux de pollution ou prévoir des ventes, il devient nécessaire de travailler à partir d'échantillons. En pratique, il est rarement possible de mesurer l'ensemble d'une population, pour des raisons de coût, de temps ou de faisabilité.

Cet ouvrage met l'accent sur les méthodes d'estimation, leur compréhension conceptuelle, leurs conditions d'application, leurs limites ainsi que sur l'analyse raisonnée des résultats. Estimation ponctuelle, intervalles de confiance, tests d'hypothèses, risques de première et de deuxième espèce, puissance statistique, méthodes paramétriques et non paramétriques, analyse de la variance, tests du Khi-deux et régression linéaire simple y sont présentés de manière progressive, en privilégiant un usage scientifiquement fondé de l'inférence statistique.

Avec une seule observation, il est impossible de tirer des conclusions générales fiables. Supposons par exemple qu'une enquête de satisfaction soit réalisée auprès d'un seul client afin d'évaluer le niveau de satisfaction de l'ensemble de la clientèle d'un magasin : une telle conclusion apparaîtrait naturellement peu crédible. Si l'on interroge dix personnes, l'estimation devient déjà plus informative ; avec cent personnes, elle gagne encore en précision. Toutefois, un autre échantillon de cent individus conduirait probablement à un résultat légèrement différent. Toute estimation comporte donc une part d'incertitude qu'il convient de mesurer et d'encadrer. À ce titre, la statistique inférentielle fournit un cadre pour raisonner en situation d'incertitude.

Dans cette perspective, cet ouvrage s'inscrit dans une démarche d'accompagnement et d'appropriation des méthodes de statistique inférentielle. Il cherche à en développer une compréhension à la fois intuitive et méthodologiquement fondée, tout en favorisant une articulation étroite entre théorie, exemples concrets et applications pratiques. De nombreux exemples appliqués et exercices corrigés s'appuient sur de petits jeux de données pouvant être exploités sous Excel, afin de permettre au lecteur de reproduire les calculs et de mieux comprendre les méthodes présentées.

L'objectif n'est pas seulement de présenter des outils de calcul, mais également de fournir les clés nécessaires à leur utilisation critique. Car un résultat statistique n'est

jamais une certitude : il constitue un raisonnement probabiliste fondé sur des données observées et assorti d'un risque d'erreur qu'il convient de connaître et de maîtriser[1].

[1] Si j'ai seul la responsabilité de cet ouvrage, je dois ici remercier Bernard Girard, normalien, d'avoir bien voulu m'apporter avec la grande humilité qui l'habite, ses connaissances en probabilité et en statistique.

Je remercie également les Éditions Ellipses pour la confiance qu'elles ont accordée à ce projet. Mes remerciements vont tout particulièrement à Monsieur de Laboulaye pour ses remarques, ses suggestions et son accompagnement attentif tout au long de la préparation de cet ouvrage.

TABLE DES MATIÈRES

CHAPITRE 1 - CADRE DE LA STATISTIQUE INFÉRENTIELLE 11

I - QU'EST-CE QUE LA STATISTIQUE INFÉRENTIELLE ? 12

II - ÉLÉMENTS DE BASE DE LA STATISTIQUE INFÉRENTIELLE ... 15

A - PROBABILITÉS ET ÉVÉNEMENTS 15

B - VARIABLE ALÉATOIRE (V.A.) 21

C - DEGRÉS DE LIBERTÉ 26

D - MESURES STATISTIQUES D'UNE VARIABLE ALÉATOIRE 27

E - LOIS DE PROBABILITÉ 32

F - LOI DES GRANDS NOMBRES 39

G - ÉCHANTILLON ALÉATOIRE SIMPLE 41

H - THÉORÈME CENTRAL LIMITE (TCL) 43

I - CONVERGENCE EN LOI 50

J - HISTOGRAMME ET FONCTION DE RÉPARTITION 53

K - CORRECTION DE CONTINUITÉ (C.C.) 61

L - APPROXIMATION PAR LA LOI NORMALE 62

III - RÉSUMÉ DU CHAPITRE 1 68

IV - EXERCICES DU CHAPITRE 1 70

CHAPITRE 2 - ESTIMATIONS 77

I - ESTIMATEUR 77

A - PROPRIÉTÉS D'UN BON ESTIMATEUR 78

B - ESTIMATION PAR LE MOMENT 78

C - ESTIMATION PAR LE MAXIMUM DE VRAISEMBLANCE (EMV) 79

II - INTERVALLE DE CONFIANCE DE L'ESTIMATEUR 87

A - INTERVALLE DE CONFIANCE DE LA MOYENNE 88

B - CAS DES PETITS ÉCHANTILLONS ($n < 30$) 94

C - INTERVALLE DE CONFIANCE D'UNE PROPORTION....97
D - INTERVALLE DE CONFIANCE DE L'ÉCART TYPE....102
E - INTERVALLE DE CONFIANCE UNILATÉRAL....107
III - ESTIMATION DE LA DIFFÉRENCE ENTRE DEUX MOYENNES....108
A - NATURE DE LA COMPARAISON DE DEUX MOYENNES....108
B - COMPARAISON DE MOYENNES DE DEUX ÉCHANTILLONS INDÉPENDANTS....111
C - COMPARIASON DE MOYENNES D'ÉCHANTILLONS APPARIÉS....115
IV - CONTRAINTES DES TESTS PARAMÉTRIQUES....117
A - NORMALITÉ DE LA DISTRIBUTION....119
B - ÉGALITE DES VARIANCES....122
C - AUTRE MÉTHODE DE TEST D'ÉGALITÉ DES VARIANCES (σ CONNU)....125
V - ESTIMATION DE LA DIFFÉRENCE ENTRE DEUX PROPORTIONS....128
VI - RÉSUMÉ DU CHAPITRE 2....129
VII - EXERCICES DU CHAPITRE 2....131

CHAPITRE 3 - TESTS D'HYPOTHÈSES & P-VALUE....135
I - TESTS D'HYPOTHÈSES....135
II - HYPOTHÈSE NULLE (H0) ET HYPOTHÈSE ALTERNATIVE (H1)....136
III - SENS DU TEST : TEST BILATÉRAL OU TEST UNILATÉRAL....138
A - TEST UNILATÉRAL....139
B - TEST BILATÉRAL....139
IV - P-VALUE....141
A - RÈGLES DE DÉCISION STATISTIQUE....142
B - COMPARAISON D'UNE PROPORTION INCONNUE π À UNE VALEUR HYPOTHÉTIQUE $\pi 0$....144
C - COMPARAISON DE DEUX MOYENNES....146
D - COMPARAISON D'UN ÉCART TYPE À UNE VALEUR HYPOTHÉTIQUE....148
E - TEST DE CONFORMITÉ....150
F - COMPARAISON DES COEFFICIENTS DE RÉGRESSION....151

V - VALEUR CRITIQUE (V.C.) 157
VI - ERREUR DE DEUXIÈME ESPÈCE : LE RISQUE BÊTA.....159
A - RISQUE DE PREMIÈRE ESPÈCE α 159
B - RISQUE DE DEUXIÈME ESPÈCE β 160
C - PUISSANCE DU TEST 161
D - CALCUL DU RISQUE BÊTA 162
VII - RÉSUMÉ DU CHAPITRE 3.....166
VIII - EXERCICES DU CHAPITRE 3 167

CHAPITRE 4 - ANALYSE DE VARIANCE (ANOVA).....173
I - FONDEMENTS DE L'ANOVA.....173
A - NATURE DE L'ANALYSE DE VARIANCE 173
B - LA STATISTIQUE DE L'ANOVA SUIT UNE LOI DE FISHER 175
C - CONTRAINTES DE L'ANOVA 176
II - ANALYSE DE LA VARIANCE À UN FACTEUR 178
A - TABLEAU ANOVA 183
B - ANOVA SUR ÉCHANTILLONS DE TAILLES INÉGALES 187
III - COMPARAISONS MULTIPLES DE MOYENNES (CMM) 192
A - MÉTHODE DE TUKEY 193
B - TEST AVEC XLSTAT 196
C - MÉTHODE DE BONFERRONI 198
IV - LIEN ENTRE LA RÉGRESSION ET L'ANOVA 203
A - VARIABLES QUALITATIVES.....203
B - SIGNIFICATIVITÉ STATISTIQUE DES COEFFICIENTS DE RÉGRESSION 204
C - CALCUL DES COEFFICIENTS À MAIN LEVÉE 204
D - CALCUL DU MODÈLE AVEC EXCEL 207
V - ANOVA À DEUX FACTEURS 213
A - ANOVA À DEUX FACTEURS CROISÉS SANS RÉPÉTITION.....214
B –ÉCRITURE SYMBOLIQUE D'UN MODÈLE ANOVA
À DEUX FACTEURS CROISÉS SANS RÉPÉTITION.....218
C - ANOVA À DEUX FACTEURS CROISÉS AVEC RÉPÉTITIONS 220

VI - COMPARAISON DE PLUSIEURS VARIANCES TEST DE HARTLEY 231

VII - RÉSUMÉ DU CHAPITRE 4 233

VIII - EXERCICES DU CHAPITRE 4 235

CHAPITRE 5 - TEST DU KHI-DEUX 241

I - TEST DU KHI-DEUX : PRÉSENTATION 243

A - HYPOTHÈSE NULLE D'INDÉPENDANCE 243

B - CONDITIONS D'APPLICATION DU TEST 246

C - LE TEST DU KHI-DEUX EN PRATIQUE 247

II - EXEMPLES D'APPLICATION 248

A - TEST D'INDEPENDANCE DU KHI-DEUX 248

B - TEST POUR MULTINOMIALES 252

C - TEST D'HOMOGÉNÉITÈ 255

D - TEST D'AJUSTEMENT À UNE LOI DE PROBABILITÉ 258

III - V DE CRAMER 262

IV - CAS DES PETITS EFFECTIFS THÉORIQUES 264

A - CORRECTION DE YATES 264

B - TEST EXACT DE FISHER 266

V - RÉSUMÉ DU CHAPITRE 5 272

VI - EXERCICES DU CHAPITRE 5 273

CHAPITRE 6 - TESTS NON PARAMÉTRIQUES 279

I - TEST DES SIGNES 280

A - TEST DES SIGNES SUR LA MÉDIANE 280

B - INTERVALLE DE CONFIANCE POUR LA MÉDIANE 284

C - COMPARAISON DE DEUX ÉCHANTILLONS APPARIÉS AVEC LA MÉDIANE DES DIFFÉRENCES 288

D - APPROXIMATION NORMALE DES GRANDS ÉCHANTILLONS (N > 30) 290

II - TESTS SUR LES RANGS 295

A - PRINCIPE DES TESTS SUR LES RANGS 295

B - TEST SUR LES RANGS DE WILCOXON SUR DEUX ÉCHANTILLONS INDÉPENDANTS 296
C - TEST DE MANN-WHITNEY (M-W) 301
D - ÉCHANTILLONS APPARIÉS - TEST DES RANGS SIGNÉS DE WILCOXON 306
E - KRUSKAL WALLIS (K-W) 310
III - RÉSUMÉ DU CHAPITRE 6 316
IV - EXERCICES DU CHAPITRE 6 317

ANNEXES 321
A - THÉORÈME CENTRAL LIMITE 321
B - INTERVALLE DE CONFIANCE DE L'ÉCART TYPE 323
C - QUELQUES PROPRIÉTÉS DE LA VARIANCE 325

SCHÉMAS DÉCISIONNELS : CHOIX DES TESTS 329

LEXIQUE 333

NOTATIONS 337

TABLES STATISTIQUES 341

GRILLES D'INTERPRÉTATION 369

SOLUTIONS DES EXERCICES 373

ERREURS FRÉQUENTES - INTERPRÉTATIONS ERRONÉES 399
A - ERREURS FRÉQUENTES 399
B - INTERPRÉTATIONS ERRONÉES 402
C - PIÈGES MÉTHODOLOGIQUES 403
D - CONDITIONS À VÉRIFIER AVANT CHAQUE TEST 405

RÉFÉRENCES 407

INDEX 411

CHAPITRE 1 - CADRE DE LA STATISTIQUE INFÉRENTIELLE

Au cours de la pandémie de la Covid-19, de nombreux laboratoires ont développé des vaccins. Pour recevoir l'agrément de mise sur le marché par l'Agence Européenne des Médicaments, ils devaient produire une évaluation de leur vaccin. Ainsi Pfizer/BioNTech a mené une étude d'efficacité reposant sur la comparaison d'un échantillon test et d'un échantillon témoin en novembre 2020. Dans ces deux échantillons, il a recensé un total de 170 malades : 8 malades provenaient du groupe test vacciné et 162 du groupe témoin, soit une efficacité du vaccin proche de 94,5%[1]. On ne pouvait pas bien sûr tester tous les malades, pas plus qu'on ne pouvait fonder notre décision sur un seul patient ; il fallait donc bâtir un échantillon aléatoire[2] suffisamment représentatif de la population étudiée. Mais en raison du caractère aléatoire de l'échantillon, les résultats comportent une part d'incertitude. C'est précisément ici qu'intervient la statistique inférentielle qui permet de contrôler les aléas liés à l'échantillonnage. On fait alors appel aux lois de probabilité pour modéliser l'aléa et quantifier l'incertitude associée aux résultats obtenus. Ainsi statistique et probabilités se complètent pour fournir des estimations et formuler des conclusions fiables sur un paramètre inconnu de la population en maîtrisant le risque d'erreur.

D'une façon générale, le besoin d'information est croissant dans de nombreux domaines : politique, marketing, contrôle qualité, analyse sensorielle, etc. Par exemple, en 1965 on comptait quatorze sondages électoraux pendant la campagne présidentielle ; à l'occasion de la campagne présidentielle de 2017[3] on en recensait environ 500 ; et lors de la campagne présidentielle de 2022, il avait déjà été réalisé plus de 100 sondages sept mois avant le premier tour (*1er tour : 10 avril),* auquel s'ajoutait un Rolling quotidien IFOP - Fiducial 2022 pour Paris Match, LCI & Sud Radio[4].

---

[1] Les données proviennent du journal *L'Express*, en collaboration de l'*AFP*, du 23 novembre 2020. https://www.lexpress.fr/actualite/societe/sante/covid-19-quatre-questions-sur-l-efficacite-des-differents-vaccins_2139278.html

[2] Le tirage au sort a été réalisé, à l'intérieur d'une population sélectionnée.

[3] « Sondages électoraux : Peut-on leur faire confiance ? » Les Cafés de la Statistique du 9 mai 2017 - *SFdS*.

[4] Le rolling *IFOP* est un sondage quotidien réalisé de façon continue : son échantillon d'environ 500 personnes interrogées, représentatif de la population française, était renouvelé par tiers tous les jours. Ce type de sondage permet de suivre les évolutions de l'opinion en continu.

Par ailleurs il existe de nombreuses situations qui nous incitent à travailler à partir d'échantillons :

- Celles où il est impossible de faire autrement. Un géologue qui souhaite connaître la composition d'un dépôt de limon au fond d'un étang ne va pas analyser tous les sédiments ; un fournisseur ne va pas griller toute sa production d'ampoules pour en mesurer la durée de vie ; pas plus qu'une personne ne peut donner tout son sang pour réaliser une analyse. Il en va de même lorsque les populations sont infinies ou assimilées à des populations infinies (processus continus, profils pédologiques, trafic routier, production halieutique, etc.). Dans toutes ces situations, on recourt à l'échantillonnage.
- Lorsque les populations sont difficiles à dénombrer, comme dans le cas du nombre de logements commencés et construits : ces projets peuvent être annulés puis repris, ce qui peut entraîner des doubles comptes. Une situation analogue se rencontre lorsqu'il faut dénombrer le nombre d'arbres dans une forêt, ou le nombre de poissons dans un océan.
- En présence d'un changement. Par exemple, la réalisation du Grand Paris modifie les systèmes territoriaux ; pour en évaluer l'impact, il est nécessaire de réaliser des enquêtes afin de mieux appréhender la qualité de vie recherchée par les habitants, les besoins en transports, et les orientations stratégiques à adopter pour le développement des PME.
- Dans certains cas, les données issues d'enquête peuvent s'avérer plus fiables que les données administratives, notamment lorsque les définitions statistiques de ces dernières sont susceptibles d'évoluer avec le temps [1].
- Et puis tout simplement lorsque les données sont inexistantes, il est impossible d'interroger toute la population, car ce serait trop long et trop coûteux.

Dans ces situations, on exploite souvent des échantillons aléatoires. Les résultats obtenus comportent certes une part d'erreur due à la variabilité de l'échantillonnage et à divers facteurs externes non maîtrisables (comme les erreurs d'observation ou la non réponse), mais ils fournissent généralement des estimations fiables des paramètres souvent inconnus dans la population.

I - QU'EST-CE QUE LA STATISTIQUE INFÉRENTIELLE ?

Plusieurs personnes tirées au hasard goûtent une sauce. Quelle que soit leur appréciation — qu'ils la jugent bonne ou mauvaise — leur jugement repose sur une

[1] En 2007 un débat s'était instauré sur les statistiques du chômage ; d'un côté on disposait des données administratives de *l'ANPE*, de l'autre côté il y avait les données d'enquête de *l'INSEE* ; c'est cette dernière qui offrit la meilleure précision.

forme d'inférence. Ils attribuent à la sauce les caractéristiques perçues à partir de quelques gorgées goûtées.

La statistique inférentielle consiste à extrapoler les résultats obtenus à partir d'un échantillon à l'ensemble de la population. (1-1)

La Figure 1-1 illustre ce point :

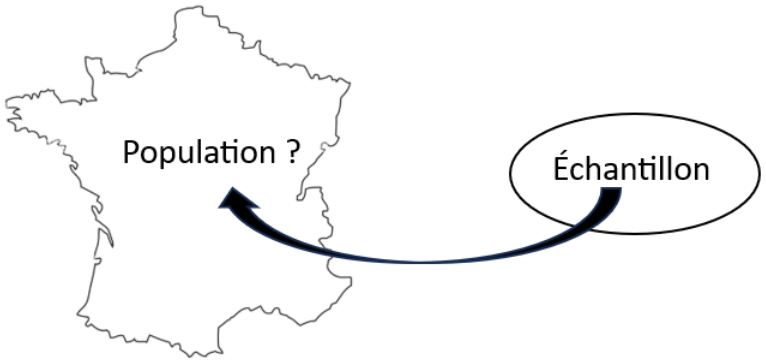

FIGURE 1-1 : Extrapolation des résultats d'échantillon à l'ensemble de la population.

Dans cet exemple, il existe des facteurs non contrôlés comme les préférences personnelles, les sensibilités gustatives, l'état émotionnel, etc. Pour ces raisons le résultat ne reflètera pas exactement les propriétés globales de toute la sauce. Tous ces facteurs contribuent à la variabilité des appréciations. C'est pourquoi on utilise les lois de probabilités afin de mesurer l'incertitude liée aux facteurs exogènes intervenant dans la prise de décision.

Élargissons ce propos à trois exemples :

1) Une entreprise lance un nouveau produit ; elle souhaite connaître les intentions d'achat des clients. Elle interroge aléatoirement 100 personnes ; parmi elles, 17 envisagent de l'acheter, soit 17%. Il s'agit d'une estimation ponctuelle, qui repose sur des clients tirés au hasard ; un autre échantillon aurait produit un résultat différent. Pour contrôler cette incertitude, on construit une région de confiance ayant un niveau de confiance fixé. Ainsi, avec un niveau de confiance de 95 %, les données observées conduisent à estimer que la vraie proportion se situe entre 9,6 % et 24,3 %.
2) Les jeunes (18-24 ans) se disent plus préoccupés par le réchauffement climatique que leurs aînés (65 ans et plus) ; leurs aînés estiment qu'ils le sont tout autant. On clarifie cette polémique ; on réalise un sondage auprès de 400 jeunes et de 600 aînés : 93% des jeunes se disent préoccupés, contre

80% pour leurs aînés. Avec un niveau de confiance de 95%, on estime que la préoccupation des jeunes se situe entre 9 et 17 points au-dessus à celle de leurs aînés.

3) Lors d'un contrôle qualité, sur 100 pièces échantillonnées 2 sont défectueuses. Le client tolère au plus 1% de pièces défectueuses. Refuse-t-il pour autant la livraison ? Pas nécessairement. En raison des fluctuations d'échantillonnage, une proportion observée de 2 % peut rester compatible avec une proportion réelle de 1 %, avec seuil de risque fixé.

Ces exemples illustrent la place croissante des échantillons dans l'analyse des phénomènes et soulignent l'importance de la statistique inférentielle pour contrôler l'incertitude associée aux résultats issus des données d'échantillon.

STATISTIQUE INFÉRENTIELLE VS STATISTIQUE DESCRIPTIVE

En cela la statistique inférentielle se distingue de la statistique descriptive. La statistique descriptive répond à la question : « Que montrent les données que j'ai observées ? », tandis que la statistique inférentielle répond à une autre question : « Que puis-je dire sur la population à partir de ces données, et avec quel risque d'erreur ? ».

C'est précisément ce passage de l'échantillon à la population qui rend nécessaire le recours aux probabilités.

L'objectif de la statistique descriptive est de résumer et d'organiser les données d'un échantillon au moyen d'indicateurs (moyenne, dispersion), de tableaux et de graphiques. Son rôle est essentiellement exploratoire : elle décrit les données observées mais ne permet pas de généraliser les résultats à la population, comme le fait la statistique inférentielle qui raisonne en termes de probabilité.

Une différence de moyennes ou une forme de distribution qui semble différente, n'ont aucun fondement scientifique sans un cadre probabiliste permettant de contrôler le risque d'erreur. Sans les probabilités, la statistique resterait descriptive : elle constaterait, mais ne pourrait ni conclure, ni décider, ni généraliser. Or conclure, décider et généraliser comportent toujours une part d'incertitude, que la statistique inférentielle permet précisément de maîtriser.

Supposons que nous lancions 1000 fois trois dés et que nous additionnions les points obtenus. La statistique descriptive dénombrera le nombre d'occurrences de chaque somme observée, tandis que la statistique inférentielle cherchera à généraliser cette répartition à l'ensemble des lancers possibles. Les sommes observées se répartiront selon une distribution approximativement symétrique. On pourra alors estimer qu'il existe environ 5 % de chances d'obtenir une somme strictement supérieure à 15.

Cette conclusion statistique repose sur un modèle de probabilité, qui permet de quantifier l'incertitude et de contrôler le risque d'erreur (cf. infra § L2).

II - ÉLÉMENTS DE BASE DE LA STATISTIQUE INFÉRENTIELLE

A - PROBABILITÉS ET ÉVÉNEMENTS

Tirons aléatoirement 3 pièces dans la production ; chaque pièce peut être bonne (B) ou défectueuse (D). Le nombre de résultats possibles est égal à $2^3 = 8$, soit 8 cas possibles. Supposons qu'on veuille connaître la probabilité de l'événement « avoir au plus 1 pièce défectueuse » ; c'est avoir aucune pièce défectueuse ou une pièce défectueuse. Quatre résultats satisfont cet évènement : BBB (aucune pièce défectueuse), BBD (la dernière pièce est défectueuse), BDB (la deuxième pièce est défectueuse, et DBB (la première pièce est défectueuse).

La probabilité p de cet événement est alors donnée par :

$$p = \frac{nombre\ de\ cas\ favorables}{nombre\ de\ cas\ possibles} = \frac{4}{8} = 0{,}5.$$

Ainsi :

Un événement est un sous ensemble des résultats possibles. (1-2)

On a donc une chance sur deux que l'événement « avoir au plus une pièce défectueuse » se produise dans un échantillon de 3 pièces. C'est la probabilité que nous aurions obtenue si on avait effectué un très grand nombre de tirages de trois pièces. Nous aurions aussi pu nous intéresser aux événements : « avoir aucune pièce défectueuse », « avoir plus d'une pièce défectueuse », etc.

Quel que soit l'événement, la probabilité qu'il se réalise, résulte d'un très grand nombre d'expériences aléatoires et indépendantes conduites dans des conditions similaires d'expérimentation sur une même population.

Ce très grand nombre de répétitions aléatoires assure la stabilité des résultats qui deviennent très proches de la réalité. Si on lance aléatoirement 10 fois une pièce de monnaie non truquée, il y a peu de chances d'obtenir 5 faces et 5 piles et pourtant la probabilité d'avoir pile ou face est $\pi = 0{,}5$. Par contre si on lance un million de fois la pièce, il est fort probable que nous obtenions des résultats voisins de 499988 Piles et 500012 Faces, soit une probabilité p très proche de 0,5 d'obtenir pile ou face. Si la probabilité diffère, on pourra conclure que la pièce est truquée.

Ce qui explique cette valeur 0,5, c'est l'indépendance des tirages aléatoires successifs ; il n'existe pas de mécanismes qui modifient l'égalité des chances.

Ainsi une probabilité est la valeur limite d'une fréquence relative simple d'un évènement lorsque le nombre de répétitions n devient très grand. (1-3)

En termes symboliques, la probabilité p d'un évènement s'écrit :

$$p = \lim_{n\to\infty}\left(\frac{freq}{n}\right) \qquad 0 \leq p \leq 1 \tag{1-4}$$

La somme des probabilités de tous les évènements élémentaires d'une expérience est égale à 1.

On s'attend donc à ce que, en moyenne l'événement se réalise un certain nombre de fois ; si on lance n fois une pièce ($n \to \infty$), avec la probabilité $\pi = 0{,}5$ d'obtenir Face, on s'attend en moyenne à observer $n\pi$ fois l'évènement « obtenir Face ».

A1 - PROBABILITÉS CONDITIONNELLES ET INDÉPENDANCE

La notion d'indépendance est essentielle en calcul des probabilités. La notion de conditionnement, l'est également : si une information change, les probabilités changent ; on parle dans ce cas de probabilité conditionnelle. En revanche si l'information n'a aucune influence sur la probabilité de l'événement étudié, on considérera que les deux événements sont indépendants.

PROBABILITÉS CONDITIONNELLES

Parler de probabilité conditionnelle signifie :

On restreint une probabilité à une condition.

(1-5)

Entre un dé truqué et un dé équilibré, les conditions expérimentales diffèrent ; ces deux situations obéissent à deux distributions de probabilité distinctes, car les mécanismes sous-jacents ne sont pas les mêmes.

EXEMPLE 1-1 : OBTENIR DEUX PIÈCES SIMULTANÉMENT BONNES

Dans une unité de production, le contrôleur qualité souhaite connaître la probabilité d'obtenir simultanément deux bonnes pièces dans un échantillon aléatoire de dix pièces, sachant qu'en moyenne trois pièces sur dix sont défectueuses.

- Probabilité que la 1re pièce P1 soit bonne = 7/10, s'écrit :

$$P(P1) = \frac{7}{10}.$$

- La probabilité que la 2e pièce P2 soit bonne sachant que la première pièce est bonne est donnée par :

$$P(P2/P1) = 6/9.$$

La probabilité que la 2e pièce soit bonne, est restreinte à la condition que la 1re pièce soit bonne.

Ainsi la probabilité que la 1re et la 2e pièce soient bonnes est égale à 0,467, soit le produit des deux probabilités (7/10 * 6/9) ; on a 46,7% de chances de tirer simultanément 2 pièces bonnes dans un échantillon de 10 pièces qui comporte 3 pièces defectueuses.

On déduit à partir de cet exemple que la probabilité d'obtenir simultanément deux bonnes pièces est :

$$P(P1 \cap P2) = P(P1) * P(P2/P1). \quad (1\text{-}6)$$

La probabilité conditionnelle que la 2e pièce soit bonne sachant que la 1re pièce est bonne est donnée par :

$$P(P2/P1) = \frac{P(P1 \cap P2)}{P(P1)}. \quad (1\text{-}7)$$

INDÉPENDANCE

On peut introduire la notion d'indépendance par cette définition :

Deux évènements sont indépendants si la réalisation de l'un n'apporte aucune information sur la réalisation de l'autre. (1-8)

Illustrons cette définition avec le lancer de dé : Si les lancers sont effectués dans des conditions identiques, chaque lancer peut être considéré comme un événement aléatoire indépendant des autres.

Cette notion d'indépendance est importante. Elle signifie que chaque observation apporte une information (cf. infra § H1).

Ainsi, la probabilité de deux événements indépendants $A1$ et $A2$ s'écrit :

$$P(A1 \cap A2) = P(PA1) * P(A2). \quad (1\text{-}9)$$

Il en découle que la probabilité d'événements indépendants est égale au produit des probabilités :

$$P(\cap A_i) = P(A_1) * \ldots * P(A_n). \quad (1\text{-}10)$$

La Figure 1-2 illustre la notion d'indépendance des observations :

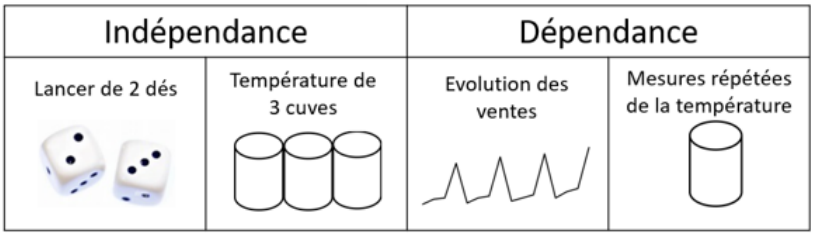

FIGURE 1-2 : Exemples de dépendance et d'indépendance des observations.

- Le lancer de dés non truqués est aléatoire ; le résultat d'un lancer ne dépend pas des précédents.
- La mesure effectuée sur chacune des trois cuves ne dépend pas des mesures réalisées sur les autres cuves ; les observations sont donc indépendantes.
- En revanche les données temporelles de ventes peuvent être influencées par la saison ou la météo ; le chiffre d'affaires d'un trimestre peut dépendre de celui du précédent trimestre ; dans ces situations il y a dépendance.
- De même que les températures répétées d'une cuve sont corrélées ; chaque mesure dépend de la précédente ; les observations sont donc liées.

Ainsi lorsque les observations sont dépendantes, elles apportent des informations très similaires voire redondantes. La quantité d'information utile contenue dans l'échantillon s'en trouve réduite. Si cette dépendance est ignorée, les résultats des estimations et des tests statistiques peuvent être faussés.

EXEMPLE 1-2 : LA QUALITÉ DÉPEND-T-ELLE DU SITE ?

Les pièces sont fabriquées sur deux sites S1 et S2. Les pièces sont soit bonnes, soit elles ont besoin d'être retouchées. La distribution conjointe des probabilités associées aux sites et à l'état des pièces est la suivante :

Sites /Pièces	Bonnes B)	Retouchées (M)	Total
S1	0,467	0,2	0,667
S2	0,233	0,1	0,333
Total	0,7	0,3	1

On souhaite savoir si la proportion de pièces à retoucher est indépendante du site.

RÉPONSE :

Nous déduisons à partir du tableau ci-dessus :

- La probabilité qu'une pièce soit retouchée :

$$P(M) = 0{,}3.$$

- La probabilité qu'une pièce soit produite sur le site S1 :

$$P(S1) = 0{,}667.$$

La probabilité que la pièce soit retouchée sachant qu'elle est produite sur le site S1 :

$$P(M/S1) = \frac{P(M \cap S1)}{P(S1)} = \frac{0{,}2}{0{,}667} = 0{,}3. \qquad \text{cf. (1-7)}$$

On constate que :

$$P(M) = P(M/S1).$$

Ces deux probabilités sont égales.

Ceci montre que la qualité de la production est indépendante du site ; le site et la qualité des pièces sont deux variables statistiquement indépendantes. La probabilité conditionnelle est égale à la probabilité marginale.

D'après les résultats précédents, on peut écrire :

$$P(M) = \frac{P(M \cap S1)}{P(S1)}.$$

On déduit l'expression de l'indépendance :

$$P(M \cap S1 = P(M) * P(S1). \qquad \text{cf. (1-9)}$$

Ainsi, la probabilité qu'une pièce soit produite sur le site S1 et qu'elle nécessite une retouche est égale au produit des deux probabilités.

Sur le site S2 les calculs conduisent à la même conclusion :

$$P(M) = P\left(\frac{M}{S2}\right) = 0{,}3.$$

$$avec\ P(M/S2) = \frac{P(M \cap S2}{P(S2)} = \frac{0{,}1}{0{,}333} = 0{,}3.$$

A2 - PROBABILITÉ D'UN ÉVÉNEMENT

EXEMPLE 1-3 : POIDS D'UN ENFANT DE 3 ANS

On a relevé les poids d'un très grand nombre d'enfants âgés de 0 à 3 ans. La lecture graphique (Figure 1-3) permet de prédire qu'à 3 ans, un enfant pèsera en moyenne entre 11kg et 17 kg, avec un risque d'erreur de 6% qu'il ait un poids inférieur à 11kg ou supérieur à 17kg.

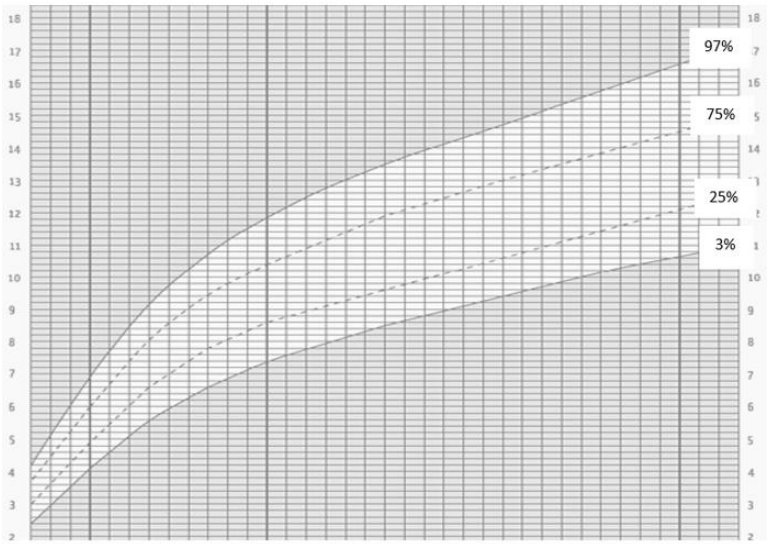

FIGURE 1-3 : Distribution des poids des enfants de 0-3 ans.
Tracé établi à partir de l'étude séquentielle française de la croissance CIE-INSERM (M Sempé).
http://www.guidegrossesse.com/examens-medicaux/poids-bebe.htm

A3 - AXIOMES DE PROBABILITÉS

Les probabilités vérifient deux axiomes :

1. La somme des probabilités des évènements élémentaires est égale à 1, c'est-à-dire P(Ω) = 1, où Ω représente l'ensemble des issues possibles. Ainsi la probabilité d'un événement A_i est comprise entre 0 et 1 :

$$0 \leq P(A_i) \leq 1.$$

2. La probabilité de l'union d'événements deux à deux disjoints est égale à la somme de leurs probabilités :

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P(A_1) + \cdots + P(A_n).$$

On dira que deux événements sont disjoints ou incompatibles si leur intersection est vide :

$$P(A_1 \cap A_2) = \emptyset.$$

Si on considére les événements A_1 « avoir moins d'une pièce défectueuse » et A_2 « avoir plus de 2 pièces défectueuses ». Ces deux événements sont disjoints car leur intersection est un espace vide.

EXEMPLE 1-4 : RÉPARTITION DES VENTES

Nous disposons de la répartition des ventes annuelles dans un magasin :

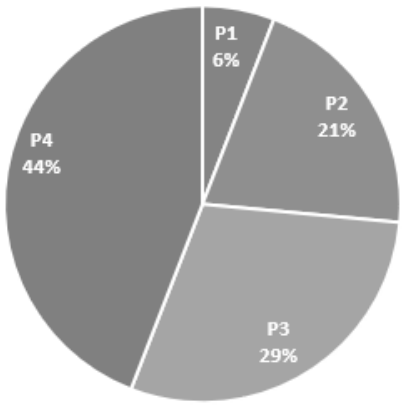

FIGURE 1-4 : Diagramme circulaire - répartition des ventes par produits.

Nous pouvons obtenir une information sur la probabilité de ces ventes.

Sur la Figure 1-4, on constate que le produit P1 cumule 6% des ventes annuelles ; P1 occupe 6% de la surface. La probabilité qu'un client achète le produit P1 est égale à $\pi = 0{,}06$. Le total des ventes occupe 100% de la surface et correspond à une probabilité totale π = 1. Ceci vérifie l'axiome 1 selon lequel les probabilités des ventes varient entre 0 et 1.

Par ailleurs le produit P1 représente 6% des ventes, le produit P2, 21% des ventes. Ces produits P1 et P2 sont disjoints. La probabilité d'acheter le produit P1 ou le produit P2 est égale à 0,27 (0,06 + 0,21), soit la somme des deux probabilités ; l'intersection $P1 \cap P2$ est vide. Ceci vérifie l'axiome 2.

B - VARIABLE ALÉATOIRE (V.A.)

Un service de production s'intéresse au montant des achats de ses clients. Il s'agit d'une variable aléatoire car ce montant peut prendre différentes valeurs qui dépend

du client observé. Elle est dite aléatoire, car le montant des achats d'un client n'est pas connu à l'avance.

Sylvie Méléard[1] définit la notion d'aléatoire comme suit : « Un phénomène est dit aléatoire si, reproduit maintes fois dans des conditions identiques, il se déroule chaque fois différemment de telle sorte que le résultat de l'expérience change d'une fois sur l'autre de manière imprévisible ».

Si on tire un échantillon aléatoire de n clients, à chaque client est associé un montant d'achat, inconnu à l'avance, que l'on modélise par une variable aléatoire X_i $(i = 1, \dots, n)$. Comme les clients sont prélevés au hasard, les variables $X_1, \dots, X_n$ ont la même distribution de probabilités que la variable X qui modélise le montant des achats d'un client pris au hasard dans la population des clients du magasin.

À chaque variable aléatoire X_i correspond un montant x_i réellement dépensé par le i^e client. On obtient ainsi l'échantillon des montants observés des n clients :

$$x_1, x_2, \dots, x_n$$

Par exemple, si on tire au hasard 8 clients, nous avons 8 variables aléatoires $X_1, \dots, X_8$, associées à 8 montants d'achats :

$$x_1, x_2, \dots, x_8$$

On l'illustre dans le Tableau suivant :

X_i	1	2	3	4	5	6	7	8
x_i	20	63	45	52	19	28	34	41

On déduit que la moyenne et l'écart type calculés sur cet échantillon sont eux-mêmes des variables aléatoires : ils résultent d'une combinaison de variables aléatoires associées aux observations.

En effet ces statistiques dépendent des valeurs prises par les variables aléatoires individuelles de l'échantillon, qui varient aléatoirement d'un échantillon à un autre.

Ainsi la moyenne observée sur un échantillon de 8 clients est définie par :

$$\bar{x} = \frac{1}{8}\sum_{1}^{8} x_i.$$

[1] Définition donnée par Sylvie Méléard - *Aléatoire - Introduction à la théorie et au calcul des probabilités* - Editions de l'Ecole Polytechnique - 2011 - Définition 1.3.1.

EXEMPLE 1-5 : CONTRÔLE QUALITÉ DES LOTS

On prélève aléatoirement 100 lots. Pour chacun d'eux, on observe le nombre de pièces défectueuses. Ce nombre est modélisé par une variable aléatoire X_i $(i = 1, \dots, 100)$. Les valeurs observées sont notées $x_1, \dots, x_{100}$.

Toutes ces variables X_i ont la même distribution que la variable aléatoire X, qui représente le nombre de pièces défectueuses dans un lot pris au hasard.

On suppose que X prend les valeurs $\{0,1,2,3\}$. La Figure 1-5 modélise la variable aléatoire X.

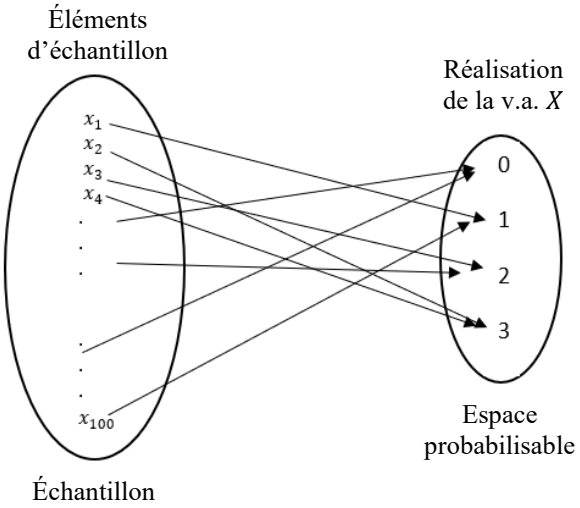

FIGURE 1-5 : Les valeurs possibles de la variable aléatoire $X = \{0, 1, 2, 3\}$.

Sur la Figure 1-5, les valeurs $\{x_1, \dots, x_{100}\}$, sont les réalisations observées des variables aléatoires X_1 , …, X_{100} . Une variable aléatoire associe une valeur numérique à chaque issue d'une expérience aléatoire. C'est pourquoi elle est définie comme une fonction de l'espace échantillon vers l'ensemble des nombres réels.

Un échantillon aléatoire constitué des valeurs observées $\{x_1, x_2, \dots, x_n\}$ permet d'estimer empiriquement les probabilités associées aux différentes valeurs de la variable aléatoire X (dans le cas discret).

La distribution des fréquences relatives fournit ainsi une approximation de la loi de probabilité de la variable X (cf. infra § E). On peut ainsi estimer la probabilité associée à chaque valeur de X.

On en déduit la définition d'une variable aléatoire :

> **Une variable aléatoire *X* est une fonction qui associe une valeur numérique à chaque issue d'une expérience aléatoire.** (1-11)

B1 - NATURE DES VARIABLES ALÉATOIRES

La nature de la variable aléatoire détermine le type de loi de probabilité que l'on utilise pour modéliser le caractère étudié dans la population.[1]

Sur la Figure 1-6, on distingue deux grands types de variables : les variables quantitatives et les variables qualitatives. Chacun de ces deux types présente deux sous-types - respectivement les variables discrètes et les variables continues, les variables nominales pures (sans ordre entre les catégories) et les variables ordinales (avec un ordre entre les catégories).

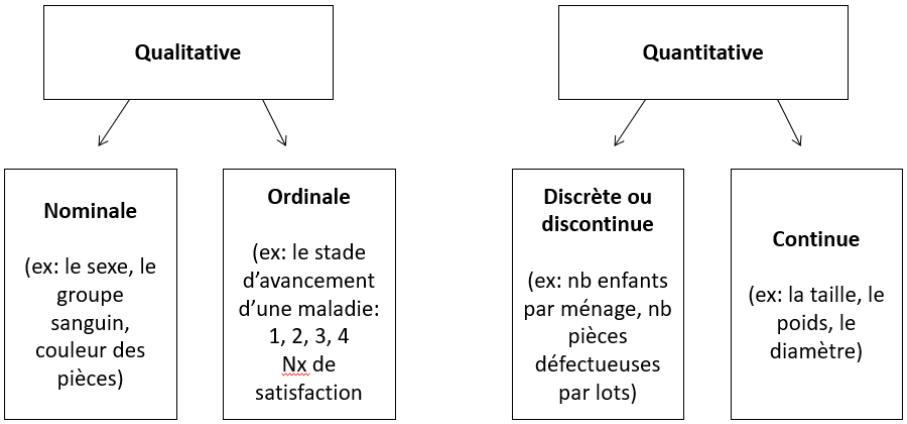

FIGURE 1-6 : Nature des variables aléatoires.

À propos des variables quantitatives, elles peuvent donc être discrète ou continue :

- Si la variable est discrète, elle prend un nombre fini ou dénombrable de valeurs : plusieurs individus peuvent avoir la même valeur. On modélise sa

[1] La population ne désigne pas uniquement un ensemble de personnes. Elle peut aussi correspondre, par exemple, aux entreprises d'une région, aux établissements d'enseignement supérieur en France, aux pièces produites sur une chaîne de fabrication ou encore aux exploitations agricoles d'un territoire.

Une population est constituée d'individus statistiques, c'est-à-dire d'unités d'observation sur lesquelles on recueille des données, telles que la taille des entreprises, le poids des pièces, le rendement des exploitations, etc. Il s'agit de l'ensemble complet des individus sur lesquels porte l'étude.

distribution à l'aide d'une loi de probabilité discrète (comme la loi binomiale ou la loi de Poisson).

- Si la variable est continue, elle peut prendre une infinité de valeurs dans un intervalle réel : en théorie aucun individu n'a la même valeur[1]. On la modélise avec une loi de probabilité continue (comme la loi normale, ou la loi exponentielle).

REMARQUES :

R1 : Lorsqu'une variable quantitative discrète prend un grand nombre de valeurs et que celles-ci sont regroupées en classes, on peut l'approcher par un modèle continu.

Par exemple, la population des communes françaises est une variable quantitative discrète, car elle ne peut prendre que des valeurs entières. Toutefois, en pratique, cette variable est approximée par une variable continue en regroupant les valeurs observées en classes d'amplitude.

Le nombre optimal de classes[2] peut être déterminé par la règle de Sturges en fonction de la taille de l'échantillon n, selon la formule :

$$nombre\ de\ classes = 1 + \frac{10 * log_{10}(n)}{3}.$$

Il est possible de déterminer le nombre de classes en prenant la racine carrée de n $\left(\sqrt{n}\right)$. Cette méthode est simple et efficace pour une première approche, mais elle est moins adaptée aux petits échantillons et moins rigoureuse que la règle de Sturges.

R2 : La différence entre une variable qualitative ordinale et une variable quantitative discrète, tient à l'échelle de mesure. Par exemple dans le cas d'une variable discrète, la différence entre 1 et 2 unités d'un objet est la même que celle entre 3 et 4 unités ; les écarts sont constants et mesurables. En revanche pour une variable qualitative ordinale, la différence d'intensité d'une maladie entre les niveaux 1 et 2 n'est pas nécessairement équivalente à la différence entre les niveaux 3 et 4 ; les écarts ne sont pas constants.

---

[1] Par définition une variable aléatoire continue prend une infinité de valeurs non dénombrable dans un intervalle, si bien qu'en théorie, il est quasiment impossible que deux individus aient exactement la même valeur numérique.

Dire qu'un ensemble est infini non dénombrable signifie qu'il contient une infinité de valeurs différentes, et on ne peut pas les énumérer. Au contraire une variable discrète est une variable aléatoire qui peut prendre un nombre fini ou dénombrable de valeurs distinctes.

[2] On fait varier souvent le nombre de classes entre 5 et 15. On ne crée pas un très grand nombre de classes pour rendre lisible la série de données.

C - DEGRÉS DE LIBERTÉ

Les degrés de liberté correspondent au nombre de valeurs qu'un ensemble de données peut prendre librement, en tenant compte des contraintes imposées (comme une somme fixée ou une moyenne connue) ; ce sont des valeurs qui peuvent varier indépendamment des autres.

On peut l'illustrer ainsi : dans une série échantillonnée de taille n, on peut choisir librement les $(n-1)$ premières valeurs ; ces $(n-1)$ valeurs sont libres ; en revanche la dernière valeur sera contrainte par la somme fixée de la série ou par sa moyenne. En raison de cette contrainte on perd un degré de liberté.

Les degrés de liberté c'est la liberté de saisir n'importe quelle valeur, telle qu'aucune d'entre elles n'est calculable à partir des autres. (1-12)

EXEMPLE 1-6 : NOMBRE DE DEGRÉS DE LIBERTÉ (DDL)

Lors du Tour de France, une entreprise prévoit de distribuer 1 000 articles promotionnels dans 5 villes étapes.

L'entreprise aura la liberté de distribuer librement les articles dans 4 villes étapes, sauf dans la dernière ville dont le nombre sera contraint par le total fixé (1 000).

Par exemple :

Villes étapes	Valeurs
V1	10
V2	500
V3	136
V4	48
V5	***306***
TOTAL	1000

Elle perd donc un degré de liberté ; ainsi le nombre d'articles promotionnels distribués dans la ville $V5$ est exprimée en fonction du nombre total d'articles promotionnels :

$$V5 = 1000 - (V1 + V2 + V3 + V4).$$

De sorte que $V5$ n'apporte aucune nouvelle information. La série des 5 villes a donc 4 degrés de liberté.

D - MESURES STATISTIQUES D'UNE VARIABLE ALÉATOIRE

Les principales mesures statistiques incluent la moyenne, l'écart type et la médiane. Lorsqu'elles sont calculées à partir d'un échantillon, elles sont notées respectivement $\bar{x}, s, \tilde{x}$; elles constituent des statistiques d'échantillon. En revanche lorsqu'elles concernent toute la population, elles sont notées μ, σ et v ; ce sont des paramètres constants de la population. Ainsi une moyenne observée dans un échantillon, notée $\bar{x}$, est une estimation ponctuelle de la moyenne réelle μ dans la population.

D1 - MOYENNE $\bar{x}$ ET ESPÉRANCE E(X)

La moyenne empirique $\bar{x}$, est calculée à partir des données observées dans un échantillon aléatoire.

L'espérance $E(X)$ est la valeur théorique que l'on s'attend à obtenir, en moyenne, si l'on répète un grand nombre de fois une même expérience aléatoire. Dans le jeu, on parle d'espérance de gain ; en médecine, on évoque l'espérance de vie.

Si on place une somme de 100 000 € sur un support financier, dont le rendement moyen est de 4,5%, on pourra espérer un gain de 4 500 euros ; cette valeur correspond à l'espérance du gain, notée $E(X) = 4\,500$. On parle d'espérance car cette valeur tient compte des probabilités des différents gains possibles.

Ainsi on déduit implicitement la formule de l'espérance pour une variable discrète :

$$E(X) = \sum_{i}^{n} x_i * p_i. \qquad (1\text{-}13)$$

soit la moyenne des valeurs pondérées par les probabilités des gains possibles. Ici les x_i représentent les valeurs prises par la variable aléatoire discrète X, et les p_i leurs probabilités associées, avec $p_i = P(X = x_i)$.

Pour une variable aléatoire continue, l'espérance est donnée par :

$$E(X) = \int_{-\infty}^{+\infty} x * f(x)\, dx. \qquad (1\text{-}14)$$

où $f(x)$ est la densité de probabilité de la variable aléatoire X.

EXEMPLE 1-7 : ESPÉRANCE D'UN DÉ

On prend un dé non truqué. La variable aléatoire X est définie par les 6 faces du dé :

$$X = \{1; 2; 3; 4; 5; 6\}.$$

Chaque face a la même chance d'apparaître, avec une probabilité $\pi = 1/6$.

Si l'on lance un très grand nombre de fois un dé équilibré, les résultats se répartiront presque équitablement entre les six faces, chaque face ayant une probabilité de 1/6.

La moyenne espérée à laquelle on peut s'attendre après un très grand nombre de tirages, s'approchera de la valeur 3,5 ; cette moyenne est alors dite espérance de X, notée $E(X)$:

Faces x_i	**1**	**2**	**3**	**4**	**5**	**6**
Probabilités p_i	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3{,}5.$$

cf. (1-13)

Cet exemple montre que l'espérance est étroitement liée à la moyenne. C'est pourquoi l'espérance théorique $E(X)$ est souvent notée μ, lorsqu'on fait référence à une population théorique :

$$E(X) = \mu.$$

D2 - VARIANCE ET ÉCART TYPE

L'écart type σ est un indicateur de dispersion autour de la moyenne. Il donne une information sur la variabilité des données ; il contribue à préciser l'estimation de la moyenne ou de la proportion. En finance, on parlera de volatilité, en production d'instabilité des processus, en marketing de la variabilité des comportements, etc. C'est la racine carrée de la variance.

La variance est aussi une mesure de la dispersion des valeurs autour de la moyenne, obtenue en calculant la moyenne des carrés d'écart à la moyenne. Lorsqu'on dispose des données complètes d'une population de taille N, la variance de la population est notée σ^2 ; c'est une constante qui est obtenue par :

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2. \qquad (1\text{-}15)$$

où μ représente la moyenne de la population, et x_i la valeur de l'individu i.

En prenant la racine carrée de la variance, on obtient l'écart type, noté σ :

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}. \qquad (1\text{-}16)$$

L'écart type s'exprime dans la même unité de mesure que les données, tandis que la variance est exprimée dans l'unité au carré des données, ce qui rend difficile son interprétation. On peut additionner ou soustraire des variances lorsque les variables sont indépendantes[1], mais on ne peut pas additionner ou soustraire des écarts types.

Lorsque l'on calcule l'écart type à partir d'un échantillon, on utilise la moyenne $\bar{x}$ de l'échantillon ; on perd donc un degré de liberté (cf. supra § C) ; il ne reste que $n-1$ valeurs libres pour estimer la variance.

Ainsi, il n'y a que $n-1$ écarts réellement indépendants. Ceci explique la présence de $n-1$ au dénominateur dans la formule de la variance.

On déduit l'écart type, noté s, appelé écart type corrigé ou écart type estimé. La formule de l'écart type d'échantillon s est de la forme :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (1\text{-}17)$$

En revanche, lorsqu'on calcule la variance σ^2 de la population cf. (1-16), la somme des carrés, $(x_i - \mu)^2$, est divisée par N car la variance de la population ne dépend pas d'une hypothèse d'indépendance. On ne fait pas d'estimation ; on connaît μ exactement ; toutes les valeurs de la population sont connues et considérées comme fixes[2]. Il n'y a donc aucune incertitude d'échantillonnage et aucun degré de liberté à corriger. Il n'est donc pas nécessaire de la corriger pour un biais d'estimation.

EXEMPLE 1-8 : DISPERSION DES SUPERFICIES AGRICOLES

Considérons une commune comportant 5 exploitations agricoles ($N = 5$), dont les superficies de ces exploitations exprimées en dizaines d'hectares sont les suivantes :

$$1\,;\ 4\,;\ 3\,;\ 2\,;\ 5.$$

---

[1] On retrouve ici trois propriétés fondamentales de la variance lorsque les variables aléatoires X et Y sont indépendantes (plus généralement non corrélées i.e. $Cov(X,Y) = 0$) :

$$V(X+Y) = V(X) + V(Y) \quad car\ \ Cov(X+Y) = Cov(X,Y) = 0$$
$$V(X-Y) = V(X) + V(Y) \quad car\ \ Cov(X-Y) = -Cov(X,Y) = 0$$
$$V(aX) = a^2 V(X) \quad (a \in \mathbb{R})$$

[2] On utilise le terme « fixe » car on est sur une population fixe et non aléatoire. Le terme « indépendant » est réservé aux variables aléatoires.

a. Calculer l'écart type σ :
b. S'il s'agissait d'un échantillon de taille $n = 5$, quelle serait la valeur de l'écart type ?

RÉPONSES :

a.

Superficies en dizaine d'ha x_i	**Ecarts à la moyenne** $(x_i - \mu)$	**(Ecarts à la moyenne)²** $(x_i - \mu)^2$
1	-2	4
4	1	1
3	0	0
2	-1	1
5	2	4
Moyenne $\mu = 3$	**Total** $\sum = 0$	**Total** $\sum = 10$

Variance de la population :

$$\sigma^2 = \frac{10}{5} = 2. \qquad \text{cf. (1-15)}$$

Ecart type de la population :

$$\sigma = \sqrt{2} \approx 1{,}41. \qquad \text{cf. (1-16)}$$

Ainsi, la superficie moyenne des exploitations de la population est de 3 dizaines d'hectares, avec un écart type de 1,41 dizaine d'hectares. Exprimées en hectares, ces valeurs correspondent à une moyenne de 30 ha et à un écart type de 14,1 ha.

Cette dispersion autour de la moyenne est plutôt forte[1]. Elle est évaluée avec le coefficient de variation, notée CV :

$$CV = \frac{14{,}1}{30} \approx 47\%.$$

Que constate-t-on dans le tableau à propos du calcul de l'écart type ?

La somme des écarts des valeurs à la moyenne est nulle ; si l'on arrêtait précisément les calculs à cet endroit, la somme des dispersions serait nulle, et identique d'une série à une autre. Bien sûr ce n'est pas possible.

[1] Coefficient de variation est une mesure relative de la dispersion des données qui correspond au rapport entre l'écart type et la moyenne. À titre indicatif, une grille d'interprétation figure en annexe.

On élève donc ces écarts au carré qui deviennent positifs ; on les divise par N, le nombre d'individus dans la population. La moyenne des écarts au carré donne la variance. On déduit l'écart type égal à 14,1 ha.

b.

La moyenne d'échantillon n'est plus une constante, mais devient une variable aléatoire puisqu'elle dépend des valeurs de l'échantillon. La somme des carrés des écarts est divisée par $(n-1)$ puisque le dernier écart est contraint par la somme. On déduit alors l'écart type d'échantillon s en dizaine d'hectares :

$$s = \sqrt{\frac{1}{5-1} * 10} \approx 1{,}58. \qquad \text{cf. (1-17)}$$

D3 - MÉDIANE

La médiane est la valeur qui partage une série de données en deux groupes de même effectif : la moitié des observations est située en dessous de la médiane, l'autre moitié au-dessus. C'est un quantile[1]. Lorsqu'on tire une valeur au hasard dans la population, elle a autant de chances d'être au-dessus qu'au-dessous de la médiane, soit $\pi = 0{,}5$.

Pour déterminer la médiane, on classe les données par ordre croissant : la médiane est la valeur située au centre de la série.

Contrairement à la moyenne, la médiane est peu sensible aux valeurs extrêmes ; elle est plus robuste. Elle est souvent préférée à la moyenne lorsque la distribution est asymétrique ou contient des valeurs aberrantes.

EXEMPLE 1-9 : TEMPS DES PAUSES CAFÉ

On interroge aléatoirement 9 personnes sur leur temps de pause. On obtient les données suivantes :

11', 15', 14', 10', 9', 13', 12', 16', 8'.

a. Déterminer la médiane

RÉPONSE :

On trie les données dans un ordre croissant :

8', 9', 10', 11', **12',** 13', 14', 15', 16'.

[1] Un quantile est une valeur qui partage une distribution en parties contenant chacune une même proportion d'observations.

Le rang de la médiane avec $n = 9$ est donné par :

$$rang = \frac{n+1}{2} = 5^e\ valeur.$$

Médiane $\tilde{x} = 12'$. Le temps médian des pauses est de 12'. Cette valeur partage la série en deux parties comportant chacune 50% des observations.

E - LOIS DE PROBABILITÉ

Une loi de probabilité modélise une expérience aléatoire. Lorsque l'on répète cette expérience un grand nombre de fois dans les mêmes conditions, les fréquences observées tendent à se stabiliser autour des probabilités théoriques.

Une loi de probabilité est une distribution de probabilités qui associe une probabilité à chacune des valeurs possibles d'une variable aléatoire. (1-18)

Prenons le cas d'une variable aléatoire continue (v.a.c.), sur laquelle on réalise un très grand nombre de mesures ($n \to \infty$). Pour étudier la loi de probabilité associée, on regroupe les valeurs observées en classes de même amplitude, puis on construit un histogramme des densités de fréquence relative qui fournit une approximation de la densité de probabilité.

La densité de fréquence relative est obtenue avec la formule :

$$\text{densité de fréquence relative} = \frac{\text{Fréquence relative de la classe}}{\text{amplitude de la classe}}. \quad (1\text{-}19)$$

La surface de chaque barre est proportionnelle à la fréquence relative de la classe ; la largeur des barres correspond à l'amplitude de la classe, tandis que la hauteur des barres représente la densité de fréquence.

Lorsque le nombre de mesures devient très grand et les classes deviennent très fines, l'histogramme tend à se confondre avec une courbe lisse qui converge vers la fonction de densité de probabilité de la variable aléatoire.

Il résulte que :

Une loi de probabilité décrit la répartition des probabilités associées aux résultats possibles d'une expérience aléatoire, et permet ainsi de modéliser l'incertitude. (1-20)

Notion d'incertitude

Avant la réalisation d'une expérience aléatoire, le résultat est incertain. La loi de probabilité permet de répondre à la question : *quelle est la plausibilité relative de ce résultat ?*

Considérons le temps d'attente à un guichet. On ne connaît pas à l'avance ce temps, qui dépend de nombreux facteurs : affluence, vitesse de traitement, incidents éventuels, etc. L'incertitude porte sur la valeur future de ce temps d'attente.

La loi de probabilité décrit cette incertitude : les attentes courtes sont plus plausibles, les attentes longues sont possibles mais rares. La loi permet de calculer une durée moyenne, des probabilités de dépassement d'un seuil donné, etc.

E1 - LOI NORMALE

La loi normale est une loi continue parmi d'autres, mais elle est très répandue en statistique. Elle permet de modéliser approximativement de nombreux phénomènes aléatoires observés dans la nature, en géologie, en sociologie, en économie, en médecine, en sciences sociales, etc.

Le terme « normal » vient du caractère universel de la loi où les valeurs proches de la moyenne sont les plus fréquentes. On dira qu'une valeur est normale lorsqu'elle se situe autour de la moyenne attendue pour la population observée. Par exemple, la taille d'un enfant est considérée comme normale si elle est proche de la moyenne pour son âge ; la température est normale si elle correspond aux normales saisonnières. Ces valeurs centrales reflètent ce qui est habituel au sein d'une population ; elles caractérisent statistiquement les phénomènes étudiés.

En raison des tirages aléatoires, les individus diffèrent d'un échantillon à l'autre, avec des valeurs qui varient ; ceci est à la base de l'erreur d'échantillonnage. Ces erreurs sont positives ou négatives, et leur distribution suit souvent une loi normale avec des valeurs négatives aussi probables que les valeurs positives. Les erreurs proviennent de facteurs non maîtrisables. Imaginons un artilleur qui tire avec un canon sur un objectif donné. Ses tirs seront déviés de l'objectif à cause de l'usure du canon, des irrégularités du projectile et de l'irrégularité des vents[1]. Il est donc normal d'avoir de petites erreurs, il n'est pas normal d'avoir des erreurs trop grandes.

Dans une population, les erreurs correspondent aux écarts entre les valeurs observées de la variable aléatoire X_i et la moyenne μ de la population. Il résulte que la loi normale est souvent utilisée pour modéliser une distribution des erreurs.

Ces erreurs proviennent de l'échantillonnage, avec des échantillons qui varient d'un tirage aléatoire à l'autre. (1-21)

[1] Exemple extrait de Jean Bass, *Éléments de calcul des probabilités*, Masson et Cie, 1974.

Les erreurs se répartissent donc symétriquement autour de la moyenne, en forme de cloche ; la courbe s'aplatit dans les queues de la distribution, c'est-à-dire que la densité diminue progressivement à mesure que l'on s'éloigne de la moyenne, s'approchant de zéro dans les queues de la distribution.

Supposons que l'on veuille contrôler le poids des sacs de riz produits sur une machine. Les poids varient sous l'effet de facteurs aléatoires tels que l'usure de l'outil de production, l'expérience des opérateurs, les conditions environnementales (température, humidité). L'historique des poids indique que les erreurs sont distribuées selon une loi normale, notée $\mathcal{N}(\mu_1, \sigma_1)$ avec autant d'erreurs de poids négatives que d'erreurs de poids positives. Si nous avions mesuré les poids des sacs sur une autre machine, la distribution des erreurs aurait suivi une autre loi normale $\mathcal{N}(\mu_2, \sigma_2)$, indépendante de la première.

Cette loi joue également un rôle central dans de nombreux tests paramétriques, qui ne sont valides que si l'hypothèse de normalité est vérifiée. C'est le cas, par exemple, des tests de comparaison de moyennes (test de Student), des tests d'égalité des variances (test de Fisher), de l'analyse des résidus dans un modèle de régression ou encore du calcul de coefficients de corrélation.

Caractéristiques de la loi normale

La loi normale dépend de deux paramètres : son espérance, notée μ, et son écart type, noté σ. Elle vérifie alors trois propriétés remarquables relatives à des intervalles centrés sur la moyenne, illustrées sur la Figure 1-7 :

- Environ 68% des observations se trouvent dans l'intervalle $[\mu - 1\sigma;\ \mu + 1\sigma]$, à une distance d'un écart type autour de la moyenne.
- Environ 95% des observations se trouvent dans l'intervalle $[\mu - 2\sigma;\ \mu + 2\sigma]$, à une distance de deux écarts types autour de la moyenne ; la valeur précise lue dans la Table est 1,96σ.
- Environ 99,73% des observations se trouvent dans l'intervalle $[\mu - 3\sigma;\ \mu + 3\sigma]$, à une distance de trois écarts types autour de la moyenne.

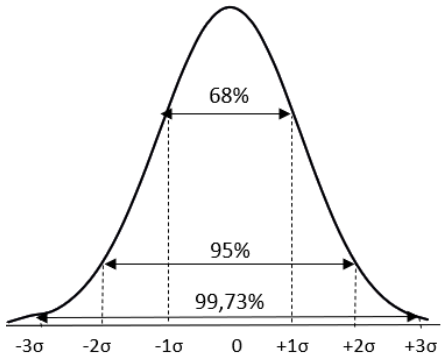

FIGURE 1-7: Distribution de densité de probabilité de la loi normale.

La courbe qui décrit la forme de la loi normale sur la Figure 1-7 correspond à la fonction de densité de probabilité, notée $f(x)$, associée à la variable aléatoire X définie sur l'intervalle $]-\infty\,;\,+\infty[$. Elle est donnée par la formule suivante :

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

où $f(x)$ représente la densité associée à une valeur réelle x.

Fonction de densité de la loi normale centrée réduite

La fonction de densité la plus utilisée est celle de la loi normale centrée réduite $\mathcal{N}(0;1)$, de moyenne 0 et d'écart type 1. On transforme la variable aléatoire X qui suit une loi normale $\mathcal{N}(\mu;\sigma)$, en une variable aléatoire centrée réduite Z :

$$Z = \frac{X-\mu}{\sigma}.$$

Cette variable Z suit la loi normale centrée réduite $\mathcal{N}(0;1)$; elle a pour fonction de densité de probabilité :

$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.} \qquad (1\text{-}22)$$

Probabilités lues dans la Table II de la loi normale

Les probabilités lues dans la Table II de la loi normale, sont des probabilités cumulées. Elles fournissent la valeur de :

$$P(Z > z).$$

La table donne $P(Z > z)$ et non pas $P(Z < z)$, car $P(Z > z)$ correspond directement aux probabilités en queue de la distribution. Or c'est le cœur de la statistique inférentielle de s'intéresser aux valeurs extrêmes. Ces probabilités sont utilisées pour déterminer les zones de rejet dans les tests d'hypothèses, et pour calculer les p-values, c'est-à-dire pour mesurer la compatibilité des données observées avec l'hypothèse testée. Elles facilitent ainsi la lecture directe des probabilités utilisées dans les tests statistiques.

D'autres Tables de la loi normale existent sur internet ou dans les livres, mais elles peuvent se présenter différemment ; par exemple on lira $P(Z \leq z)$. Supposons que nous ayons obtenu une statistique centrée réduite $z = 1{,}28$; alors dans la Table II de la loi normale centrée réduite du livre ci-jointe on lira la probabilité $P(Z > z) = 0{,}1$, tandis que si la Table sur internet communique $P(Z \leq z)$, on lira une probabilité $p = 0{,}9$.

E2 - LOI BINOMIALE

La loi binomiale est une loi de probabilité discrète associée à une variable aléatoire discrète, appelée variable binomiale. Une variable binomiale est la somme de n variables de Bernoulli[1] indépendantes, chacune pouvant prendre deux valeurs : 1 (succès) avec une probabilité π, ou 0 (échec) avec une probabilité $1-\pi$. Cette loi est notée $\mathcal{B}(n;\pi)$, où n est le nombre d'essais et π la probabilité de succès pour chaque essai. En termes littéraux, on dira : Quelle est la probabilité qu'un événement A_i se produise exactement k fois parmi n essais indépendants, sachant que la probabilité de succès de l'événement A_i est π ?

La probabilité associée est donnée par la formule de la loi binomiale :

$$P(X=k)=\binom{n}{k}*\pi^k(1-\pi)^{n-k}. \tag{1-23}$$

π^k : la probabilité d'avoir k succès.
$(1-\pi)^{n-k}$: la probabilité que les autres essais soient des échecs.
$\binom{n}{k}$: est le coefficient binomial ; il indique le nombre de façons d'avoir k succès parmi n essais. Il se calcule ainsi :

$$\binom{n}{k}=\frac{n!}{k!\,(n-k)!}. \tag{1-24}$$

Rappelons que :

$$n!=n*(n-1)*(n-2)*\ldots.*1.$$

Par exemple $3!=3*2*1=6$.

EXEMPLE 1-10 : FIDÉLISER *k* CLIENTS

Une entreprise souhaite déterminer la probabilité de fidéliser k clients parmi trois nouveaux clients $(n=3)$, sachant qu'en moyenne elle a 10% de chances de fidéliser un nouveau client $(\pi=0{,}10)$.

a. Quelle est la loi de la variable aléatoire X « Nombre de clients fidélisés » ?
b. Construire l'arbre de probabilité.

---

[1] Une variable de Bernoulli a deux issues possibles : succès ou échec, bon ou mauvais, 0 ou 1 ; elle a pour espérance $E(X)=\pi$ et pour variance $V(X)=\pi(1-\pi)$. Ainsi la somme de n variables de Bernoulli indépendantes suit une loi binomiale ; elle a pour espérance $E(X)=n\pi$ et pour variance $V(X)=n\pi(1-\pi)$.

c. Calculer la probabilité de fidéliser un client sur trois clients.
d. Déterminer la loi de probabilité de la variable aléatoire X.

RÉPONSES :

a.

La variable aléatoire $X =$ Nombre de clients fidélisés parmi 3 nouveaux clients.

Les 3 essais sont supposés indépendants.

La v.a. X peut prendre les valeurs $X = \{0 ; 1 ; 2 ; 3\}$.

La probabilité de succès : $\pi = 0{,}1$; la probabilité d'échec : $1 - \pi = 0{,}9$.

X suit une loi binomiale $\mathcal{B}(3; 0{,}1)$.

b.

Il existe 8 événements possibles, c'est-à-dire 2 choix possibles à chaque nouveau client, avec trois nouveaux clients successifs, soit $2^3 = 8$ (cf. Figure 1-8).

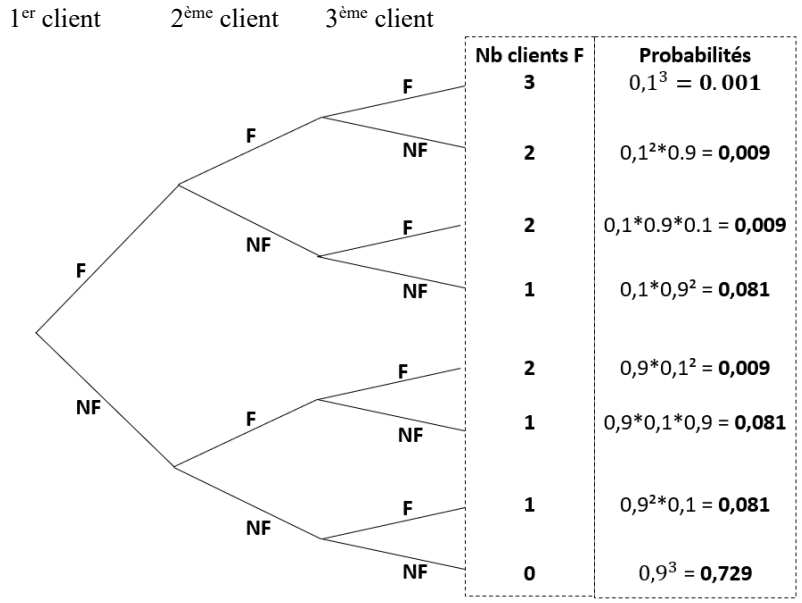

FIGURE 1-8 : Arbre de probabilités.

Lecture :

- le premier client est Fidélisé (F), ou Non Fidélisé (NF).
- le deuxième client est soumis à la même alternative : Fidélisé ou Non Fidélisé.

Si le premier client est fidélisé, le deuxième peut être fidélisé ou non fidélisé. Si le premier client n'est pas fidélisé, le deuxième peut également être fidélisé ou non fidélisé.

- la même lecture est faite pour le troisième client.

On déduit que la probabilité d'un événement est égale au produit des probabilités (cf. (1-9)). Supposons que nous obtenions l'évènement $\{F; NF; NF\}$, la probabilité de cet événement est égale à $0{,}1 * 0{,}9 * 0{,}9 = 0{,}081$.

c.

En observant l'arbre de probabilité (Figure 1-8), nous distinguons trois possibilités de fidéliser un client parmi trois clients :

$$\{F; NF; NF\}\ ou\ \{NF; F; NF\}\ ou\ \{NF; NF; F\}.$$

Chacun de ces trois évènements a une probabilité égale à 0,081. La probabilité $P(X = 1)$ est égale au cumul des probabilités des trois évènements d'avoir un client Fidélisé, soit :

$$3 * 0{,}81 = 0{,}243.$$

La probabilité d'obtenir 1 client fidélisé parmi 3 est donnée par la formule binomiale (1-23) :

$$P(X = 1) = \binom{3}{1} * 0{,}1^1(1 - 0{,}1)^{3-1} = 0{,}243.$$

avec le coefficient binomial :

$$\binom{3}{1} = 3 \text{ façons de choisir un client Fidélisé parmi 3.} \quad \text{cf. (1-24)}$$

d.

Pour connaître la loi de la variable aléatoire « Nombre de clients fidélisés », on calcule les probabilités associées à chaque valeur de la variable $X = \{0; 1; 2; 3\}$.

Nous obtenons ainsi la loi de la variable aléatoire X (cf. Tableau 1-1), dont la somme des probabilités est égale à 1 :

TABLEAU 1-1 : Loi de la v.a. *X*

Nb clients F	**Probabilités**
0	0,729
1	0,243
2	0,027
3	0,001
TOTAL	1

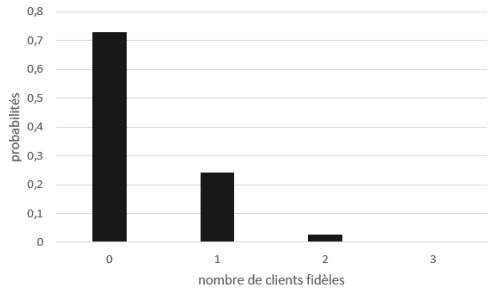

FIGURE 1-9 : Loi de probabilité de la v.a. « Nombre de clients fidélisés ».

La Figure 1-9 présente la loi de probabilité du nombre de clients fidélisés parmi trois nouveaux clients ($n = 3$), sachant que la probabilité de fidéliser un nouveau client est de 10 % ($\pi = 0{,}10$).

La Table IV de la loi binomiale présente les probabilités associées aux différentes valeurs de k, pour des valeurs spécifiques de n et π.

Excel :

Ces probabilités s'obtiennent directement en appelant la fonction Excel *loi.binomiale.n* dans la barre de commande *fx*.

Quatre arguments sont renseignés : le nombre de succès k; la taille de l'échantillon n ; la probabilité de succès π ; VRAI/FAUX ; soit :

LOI.BINOMIALE.N(1;3;0,1;FAUX) = 0,243

FAUX signifie que l'on cherche la probabilité simple P(X=1). Si nous souhaitions obtenir la probabilité cumulée P(X<2) = P(X=0) + P(X=1), l'argument deviendrait VRAI. La probabilité VRAI d'avoir moins de 2 clients fidélisés parmi 3 nouveaux clients sachant $\pi = 0{,}1$ renvoie la valeur :

=LOI.BINOMIALE.N(1;3;0,1;VRAI) = 0,972

Excel calcule la probabilité $P(X \leq 1)$.

F - LOI DES GRANDS NOMBRES[1]

La loi des grands nombres (LGN) est considérée comme une loi fondamentale en statistique. Elle signifie qu'avec un grand nombre d'expériences, la fréquence relative des évènements observés converge vers une valeur limite, sa probabilité.

[1] Cette loi a été énoncée pour la première fois par Jacques Bernoulli au XVIIIe siècle.

Illustrons ce mécanisme. On lance une pièce non truquée n fois. Les lancers sont indépendants, aléatoires et réalisés dans les mêmes conditions d'expérimentation.

Le Tableau 1-2 présente les résultats successifs (par exemple, le nombre Face) à mesure que le nombre de lancers augmente.

La Figure 1-10 montre une courbe qui se stabilise autour d'une valeur limite $\pi = 0{,}5$ lorsque le nombre de lancers n devient très grand.

Dans le Tableau 1-2, X_n représente la variable aléatoire associée aux n lancers. La réalisation x_n correspond au résultat du n^e lancer (Pile ou Face). Les résultats sont ensuite cumulés : par exemple au 3e lancer on observe 1 Face, soit une fréquence relative de p=0,33.

Sur la Figure 1-10, l'axe des abscisses représente le nombre de lancers, et l'axe des ordonnées la fréquence cumulée des « Face ». Lorsque n est faible, les fluctuations sont importantes, puis à mesure que le nombre de lancers augmente, la fréquence converge vers cette valeur limite $\pi = 0{,}5$.

TABLEAU 1-2 : Résultats de n lancers

X_n	x_n	cumul succès	P(X=cumul succès)
x_1	0	0	0
x_2	0	0	0
x_3	1	1	0,33
	1	2	0,5
	1	3	0,4
	...		
x_n	1	$n/2$	0,5

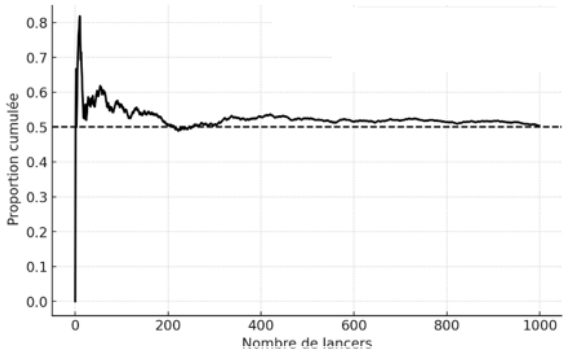

FIGURE 1-10 : Convergence vers la valeur limite $\pi = 0{,}5$ lorsque le nombre de lancers est grand.

Sur cet exemple simple, on voit le mécanisme par lequel s'opère la loi des grands nombres : lorsque le nombre n de répétitions devient grand, la fréquence relative cumulée se stabilise et converge vers cette valeur limite qui correspond à la probabilité théorique.

Ceci dit la loi des grands nombres n'est pas une loi de probabilité au sens strict ; elle ne permet pas de prédire un évènement spécifique.

Une autre illustration de la loi des grands nombres est présentée dans l'exemple 1-7. On a supposé un très grand nombre de lancers de dé, de sorte que la moyenne des résultats observés converge vers une valeur limite égale à 3,5, c'est à dire l'espérance de la variable aléatoire.