

Décision et prévision statistiques

thierry verdel

2016

v1 - 21/07/2016

Ce document reprend en grande partie le polycopié réalisé jusqu'en 1999 par Claude Chambon qui enseignait la statistique à l'Ecole des Mines de Nancy. Je lui dois beaucoup des années pendant lesquelles nous avons travaillé ensemble. Il est décédé le 23 septembre 2010.

Note : ce document a été entièrement réalisé (figures comprises) et mis en page avec Mathematica™

Sommaire

Chapitre 1. Probabilités et variables aléatoires	1
1. Probabilités	3
1.1. Événements	3
1.2. Algèbre des événements	4
1.3. Axiomes de Kolmogorov	4
1.4. Événements équiprobables	4
1.5. Théorème des probabilités totales	5
2. Probabilités conditionnelles	5
2.1. Axiome des probabilités conditionnelles	5
2.2. Théorème des probabilités composées	6
2.3. Événements indépendants	6
2.4. Le théorème de Bayes	7
3. Variables aléatoires discrètes	7
3.1. Définition	7
3.2. Loi binomiale	8
3.3. Loi hypergéométrique	9
3.4. Loi de Poisson	10
3.5. Processus de Poisson	10
3.6. Espérance mathématique d'une variable aléatoire discrète	11
3.6.1. Loi de Bernoulli	11
3.6.2. Loi binomiale	11
3.6.3. Loi de Poisson	12
3.6.4. Processus de Poisson	12
Exercices du chapitre 1	13

Chapitre 2. La loi normale	15
1. Variables aléatoires continues	17
1.1. Loi de probabilité	17
1.2. La loi uniforme	17
1.3. La loi exponentielle	18
1.4. Loi de probabilité à deux dimensions	18
1.5. Indépendance de deux variables aléatoires	19
1.6. Fonctions de variables aléatoires	19
2. Espérance mathématique	20
2.1. Espérance et moments d'une variable aléatoire	20
2.2. Variance et écart-type	20
2.3. Inégalité de Bienaymé-Tchebychev	20
2.4. Linéarité de l'opérateur Espérance	21
2.5. Calculs de variances de variables aléatoires	21
2.5.1. Loi uniforme	22
2.5.2. Variable de Bernoulli	22
2.5.3. Loi de Poisson	22
2.5.4. Loi exponentielle	23
2.6. Indépendance et covariance de deux variables aléatoires	23
2.7. Variance d'une somme de variables aléatoires indépendantes	24
3. Loi normale	24
3.1. Définition et propriétés	24
3.2. Calcul de la moyenne et de la variance	25
3.3. La loi normale réduite	26
3.4. Fonctions linéaires de variables normales	27

3.5. Théorème central limite	27
3.5.1. La loi normale comme modèle	28
3.5.2. Limite de la loi binomiale	28
3.6. La loi log-normale	28
Exercices du chapitre 2	29
 Chapitre 3. Le contrôle statistique	31
1. Distribution d'échantillonnage	33
1.1. Population	33
1.2. Echantillon	33
1.3. Caractéristiques des échantillons	33
1.3.1. Caractéristiques de tendance centrale	34
1.3.2. Caractéristiques de dispersion	34
1.4. Distributions d'échantillonnage	35
1.5. Loi de la moyenne d'un échantillon prélevé dans une population normale	36
2. Contrôle Statistique	36
2.1. Contrôle de fabrication et contrôle de réception	36
2.2. Contrôle de réception	37
2.3. Contrôle en cours de fabrication	38
2.3.1. Intervalle de tolérance et déchets	38
2.3.2. Règle de contrôle	39
2.3.3. Efficacité d'un contrôle	39
2.3.4. Risque β	40
2.3.5. Choix de la taille de l'échantillon à prélever	40
2.4. Contrôle progressif	41
Exercices du chapitre 3	43
 Chapitre 4. L'estimation statistique	45
1. Estimateur et intervalle de confiance	47
1.1. La loi des grands nombres	47
1.2. Estimation et estimateur	47
1.3. Estimateur et convergence en probabilité	48
1.4. Intervalle de confiance d'une estimation	49
1.5. Estimation d'une proportion	50
1.6. Estimation d'une moyenne	50
1.7. Estimation d'une variance	51
2. Intervalle de confiance de la variance inconnue d'une population normale	52
2.1. Loi du χ^2	52
2.1.1. Sommes de variables suivant des lois du χ^2	53
2.1.2. Moyenne et variance d'une variable qui suit une loi du χ^2	53
2.2. Loi de la variance d'un échantillon extrait d'une population normale dont σ est connu	54
2.3. Intervalle de confiance de la variance inconnue d'une population normale	55
3. Intervalle de confiance de la moyenne inconnue d'une population normale de σ inconnu	55
3.1. Loi de Student	55
3.2. Loi de la moyenne d'un échantillon extrait d'une population normale de σ inconnu	56
3.3. Intervalle de confiance de la moyenne inconnue d'une population normale de σ inconnu	56
4. Estimation du maximum de vraisemblance	57
4.1. Estimation du paramètre d'une loi de Poisson	58
4.2. Estimation du paramètre d'une loi exponentielle	58
4.3. Estimation des paramètres d'une loi normale	59

Exercices du chapitre 4	61
Chapitre 5. Comparaisons statistiques	63
1. Tests d'hypothèse	65
1.1. Théorie de Neyman et Pearson	65
1.2. Détermination de la région d'acceptation	66
1.3. Test sur une proportion	66
1.4. Test sur une moyenne	67
1.5. Cas d'hypothèses composites	68
2. Tests usuels de comparaison à un standard	69
2.1. Rappel des lois outils usuelles	69
2.1.1. Loi normale centrée réduite	69
2.1.2. Loi du χ^2	69
2.1.3. Loi de Student	69
2.2. Comparaison de la moyenne d'une population normale de σ^2 connue à une valeur donnée μ_0	70
2.3. Comparaison de la variance d'une population normale à une valeur donnée σ_0^2	71
2.4. Comparaison de la moyenne d'une population normale de σ^2 inconnue à une valeur donnée μ_0 ..	71
2.5. Test des appariements	71
3. Comparaison sur échantillons de deux populations normales	72
3.1. Comparaison des variances de deux populations normales	72
3.2. Estimation de σ^2	73
3.3. Comparaison des moyennes de deux populations normales	73
3.4. Estimation de la différence des moyennes des populations	74
Exercices du chapitre 5	75
Chapitre 6. Faits et modèles	77
1. Distributions statistiques	79
1.1. Mise en ordre des observations	79
1.2. Représentations graphiques des distributions	79
1.2.1. Histogramme	79
1.2.2. Diagramme des fréquences cumulées	80
1.2.3. Boîte à moustaches	80
1.3. Caractéristiques des distributions	80
2. Fréquences et probabilités	81
2.1. Retour sur la loi des grands nombres	81
2.2. Nombre de mesures à effectuer pour une précision donnée	81
2.3. Estimation d'une proportion et intervalle de confiance	82
2.4. Comparaison de deux proportions	82
2.5. Métrique du χ^2	82
3. Techniques de raccordement entre distributions statistiques et lois de probabilité	84
3.1. Loi de référence	84
3.2. Détermination du type de la loi de référence	85
3.2.1. Raccordement à une loi normale	85
3.2.2. Raccordement à une loi log-normale	85
3.3. Estimation des paramètres de la loi de référence	85
3.4. Vérification de la légitimité d'un raccordement effectué	86
4. Tests non paramétriques	87
4.1. Test de comparaison de plusieurs populations qualitatives	87
4.2. Test de la médiane	87
4.3. Test des signes	88

4.4. Test d'indépendance entre deux variables qualitatives	88
Exercices du chapitre 6	91
 Chapitre 7. La régression linéaire	97
1. La droite des moindres carrés	99
1.1. Nuage des individus	99
1.2. Caractérisation de la droite de régression	100
1.3. Analyse de la variance	101
1.4. Coefficient de corrélation	101
2. Propriétés statistiques de la droite des moindres carrés	102
2.1. Le modèle de la régression linéaire	102
2.2. Propriétés de a et b	103
2.3. Estimation de σ^2	104
3. La prévision statistique	104
3.1. Objectifs	104
3.2. Hypothèse de normalité	105
3.3. Test d'indépendance des variables	105
3.4. Test de nullité de l'ordonnée à l'origine	106
3.5. Intervalles de confiance pour une valeur donnée x de X	106
3.5.1. Intervalle de confiance d'un point de la droite $y = \alpha x + \beta$	106
3.5.2. Intervalle de confiance d'une observation	107
4. Comparaison de deux régressions	108
4.1. Comparaison des variances	108
4.2. Comparaison des pentes	109
4.3. Comparaison des ordonnées à l'origine	114
Exercices du chapitre 7	111
 Chapitre 8. L'expérimentation statistique	115
1. Analyse de la variance à un facteur	117
1.1. Recherche de l'influence d'un facteur	117
1.2. La relation d'analyse de la variance	117
1.3. Le modèle	118
1.4. Test d'analyse de la variance	118
1.5. Calcul pratique	118
1.6. Test de linéarité d'une régression	119
2. Etude de l'influence de deux facteurs	120
2.1. Plan factoriel	120
2.2. Modèle additif et modèle avec interaction	121
2.3. Relation d'analyse de la variance	122
2.4. Les tests d'analyse de la variance	122
2.4.1. Test de l'interaction	122
2.4.2. Test de l'influence d'un facteur	123
2.4.3. Exécution des calculs	123
2.5. Analyse de la variance sans répétitions	123
Exercices du chapitre 8	125
 Tables numériques	129
Nombres au Hasard	131
Loi Binomiale	133
Loi de Poisson	137

Loi normale centrée réduite	141
Loi de Student-Fischer	143
Loi du χ^2	145
Loi de Snédécor (F-Ratio distribution)	147
Index	149

Probabilités et variables aléatoires

L'objectif du chapitre est d'utiliser le concept de probabilité pour construire un certain nombre de modèles pouvant rendre compte de situations concrètes, où l'application de lois déterministes est impossible parce que les phénomènes sont très compliqués, ou les facteurs trop nombreux.

1 Probabilités

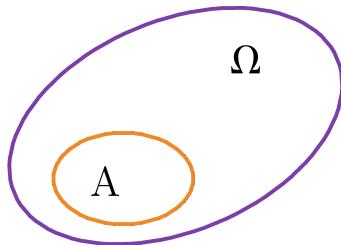
1.1 Événements

L'expérimentateur se trouve souvent dans la situation suivante : il peut prévoir quels sont les résultats possibles de son expérience, mais pas celui qui se réalisera. Plus précisément, il peut déterminer l'ensemble des possibilités, qui sera désigné par Ω ; mais celle de ces possibilités qui se réalisera effectivement lui est inconnue avant que l'expérience soit faite.

Si, par exemple, l'expérience consiste à lancer un dé à six faces, on peut définir l'ensemble des résultats possibles $\Omega = \{1, 2, 3, 4, 5, 6\}$ mais on ne sait pas à l'avance celui qui sera obtenu après le jet du dé.

L'ensemble Ω des possibilités étant défini, on appelle *événement* toute partie de cet ensemble.

Dans l'exemple précédent, l'événement A : « le résultat du jet est un chiffre impair » est constitué par les possibilités suivantes : $A = \{1, 3, 5\}$.



On dit alors que l'événement A est *réalisé* lorsque le résultat effectivement obtenu est l'une des possibilités appartenant à cet événement, un élément de A : *1 ou 3 ou 5*.

On appelle *événement élémentaire*, une partie de l'ensemble Ω des possibilités qui ne contient qu'un seul élément. Il y a donc autant d'événements élémentaires que de parties de cardinal égal à 1 dans Ω .

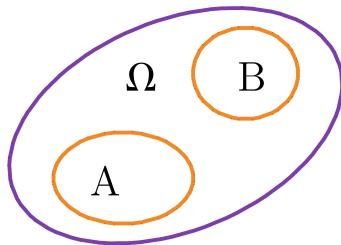
On appelle *événement impossible*, un événement qui ne contient aucun des éléments de Ω . Il lui correspond la partie vide \emptyset de Ω .

On appelle, par contre, *événement certain*, l'ensemble Ω de toutes les possibilités. Il lui correspond la partie pleine de Ω .

On appelle enfin, *événements incompatibles*, deux parties disjointes de Ω . Lançant par exemple un dé à six faces, les deux événements :

- A : le résultat est un chiffre pair
- B : le résultat est un chiffre impair

sont incompatibles puisque $A = \{1, 3, 5\}$ et $B = \{2, 4, 6\}$ n'ont aucun élément commun.



1.2 Algèbre des événements

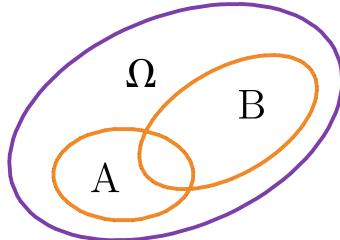
Chaque événement étant une partie de l'ensemble Ω des possibilités, l'ensemble des événements est l'ensemble des parties de Ω . Pour n possibilités, il y a donc 2^n événements $\left(\sum_{k=0}^n \binom{n}{k} = 2^n\right)$ y compris l'événement impossible et l'événement certain.

Ces événements s'organisent les uns par rapport aux autres par la relation d'inclusion : $A \subset B$ lorsque tout élément de A appartient à B . On dit alors que l'événement A *implique* l'événement B ; chaque fois que A est réalisé B l'est aussi.

Les opérations ensemblistes sur les événements sont d'usage courant. C'est ainsi qu'à chaque événement A on peut faire correspondre l'événement complémentaire \bar{A} , constitué de la partie de Ω complémentaire de A .

On peut aussi faire correspondre à deux événements A et B :

- leur réunion $A \cup B$ qui s'interprète comme l'événement dans lequel A ou B est réalisé,
- leur intersection $A \cap B$ comme l'événement dans lequel A et B sont tous les deux réalisés.



Lançant par exemple un dé, soient les événements :

- A : le résultat du jet est un chiffre pair,
- B : le résultat est un multiple de 3.

On peut définir les événements suivants :

- \bar{A} : le résultat est un chiffre impair,
- $A \cup B$: le résultat est 2 ou 3 ou 4 ou 6,
- $A \cap B$: le résultat est 6.

1.3 Axiomes de Kolmogorov

Faire correspondre une probabilité à chaque événement X , c'est-à-dire à chaque partie X de l'ensemble Ω des possibilités, c'est définir une application de l'ensemble $P(\Omega)$ des parties de Ω , dans l'ensemble des nombres réels, qui satisfasse les trois conditions suivantes (souvent appelées axiomes de Kolmogorov) :

- *positivité* : la probabilité d'un événement est un nombre positif ou nul :

$$\forall X \in \Omega, p(X) \geq 0$$

- *échelle* : la probabilité d'un événement impossible est nulle, celle d'un événement certain est égale à 1 :

$$p(\emptyset) = 0, \quad p(\Omega) = 1$$

- *additivité* : l'union de deux événements incompatibles, donc tels que $A \cap B = \emptyset$, a pour probabilité la somme des probabilités de ces événements :

$$p(A \cup B) = p(A) + p(B)$$

Il en résulte immédiatement une relation très utile : la somme des probabilités de deux événements complémentaires est égale à 1 :

$$p(A) + p(\bar{A}) = 1$$

1.4 Événements équiprobables

L'une des conséquences de la condition d'additivité est que la probabilité d'un événement quelconque est égale à la somme des probabilités des événements élémentaires e_i qui le constituent, puisque deux parties réduites à un seul élément sont disjointes :

$$p(A) = \sum_{e_i \in A} p(e_i)$$

Il en résulte que la connaissance des probabilités des événements élémentaires détermine entièrement les probabilités sur un ensemble Ω de possibilités.

Un cas particulier important est celui où les événements élémentaires sont *équiprobables* :

$$p(e_1) = \dots = p(e_i) = \dots = p(e_n)$$

Comme on a $\sum_{e_i \in \Omega} p(e_i) = 1$, la probabilité de chacun des n événements élémentaires e_i est égale à $\frac{1}{n}$, et la probabilité d'un événement A de Ω est alors égale au quotient de son cardinal $|A|$ par celui de Ω :

$$p(A) = \frac{|A|}{|\Omega|}$$

On retrouve ici la définition bien connue : *la probabilité est égale au nombre de cas favorables divisé par le nombre de cas possibles*. Mais il faudrait ajouter, pour qu'elle soit correcte : ... *sous la condition stricte que les cas soient équiprobables*. Et ce ne saurait plus dès lors constituer une définition.

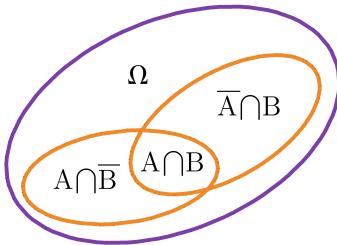
1.5 Théorème des probabilités totales

Considérons deux événements A et B non disjoints, et calculons la probabilité $p(A \cup B)$ de leur réunion. L'axiome d'additivité permet d'écrire que :

$$p(A \cup B) = p(A \cap \bar{B}) + p(A \cap B) + p(\bar{A} \cap B)$$

avec $p(A) = p(A \cap \bar{B}) + p(A \cap B)$

et $p(B) = p(\bar{A} \cap B) + p(A \cap B)$



D'où la relation souvent appelée théorème des probabilités totales :

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

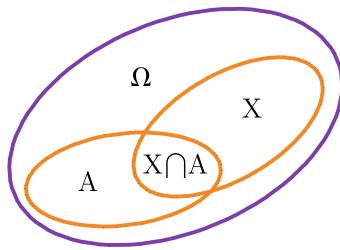
que l'on peut énoncer ainsi : si un événement peut se produire *soit* par l'arrivée d'un événement A , *soit* par l'arrivée d'un événement B , sa probabilité est égale à la somme des probabilités de A et de B moins la probabilité pour que A et B se produisent ensemble.

2 Probabilités conditionnelles

2.1 Axiome des probabilités conditionnelles

Dans ce qui précède, l'ensemble Ω des possibilités était donné une fois pour toutes. En fait, dès que certains des événements possibles se réalisent, l'ensemble des possibilités se trouve modifié.

Si l'événement A se réalise, les événements possibles deviennent en effet l'ensemble des parties de A , et non plus l'ensemble des parties de Ω . Dans ce cas, pour tout événement X de l'ensemble Ω , seule la partie $X \cap A$ reste alors un événement possible.



La réalisation d'un événement A modifie donc l'ensemble des possibilités, et chaque événement X devient $X \cap A$. Elle modifie aussi les probabilités.

On désignera par $p(X|A)$ la probabilité de l'événement X si A est réalisé. On l'appelle *probabilité conditionnelle de X sachant que A est réalisé*, et par définition :

$$p(X|A) = \frac{p(X \cap A)}{p(A)}$$

Il n'existe donc pas de probabilité conditionnelle lorsque la probabilité de A est nulle.

2.2 Théorème des probabilités composées

La définition même des probabilités conditionnelles permet d'écrire que :

$$p(A \cap B) = p(A) \times p(B|A)$$

et aussi que :

$$p(A \cap B) = p(B) \times p(A|B)$$

C'est le théorème des probabilités composées, que l'on peut énoncer ainsi : *si un événement résulte du concours de deux événements, sa probabilité est égale à celle de l'un d'eux multipliée par la probabilité conditionnelle de l'autre sachant que le premier est réalisé.*

Soit, par exemple, à calculer la probabilité pour que, tirant successivement deux cartes d'un jeu de 32 cartes, ces deux cartes soient des valets. Appelons A et B les deux événements suivants :

- A : la première carte est un valet (A désigne tous les tirages possibles dont la 1ère carte est un valet)
- B : la deuxième carte est un valet (B désigne tous les tirages possibles dont la 2e carte est un valet)

La probabilité cherchée est $p(A \cap B)$ qui est aussi égale à $p(A) \times p(B|A)$.

Lors du premier tirage, il y a 32 cartes et 4 valets dans le jeu, d'où $p(A) = \frac{4}{32}$.

Lors du second tirage, il reste 31 cartes et seulement 3 valets, puisque l'événement A est réalisé, d'où $p(B|A) = \frac{3}{31}$.

Le résultat est donc : $p(A \cap B) = \frac{4}{32} \times \frac{3}{31} = \frac{3}{248} \simeq 0.012$.

2.3 Événements indépendants

Par définition, deux événements sont indépendants si la probabilité de l'un n'est pas modifiée lorsque l'autre est réalisé. On a donc, par exemple :

$$p(A|B) = p(A)$$

Il en résulte que :

$$p(A \cap B) = p(A) \times p(B)$$

et la réciproque est évidente.

On peut donc énoncer que : *la condition nécessaire et suffisante pour que deux événements soient indépendants, est que la probabilité de leur intersection soit égale au produit de leurs probabilités.*

Les deux événements A et B de l'exemple précédent n'étaient pas indépendants. Mais si, par contre, on tire la deuxième carte après remise de la première dans le jeu, les résultats des deux tirages deviennent indépendants et $p(A \cap B) = p(A) \times p(B) = \frac{4}{32} \times \frac{4}{32} = \frac{1}{64} \simeq 0.0156$.

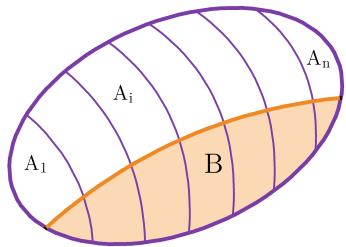
Il est essentiel de bien réaliser la différence entre événements *incompatibles* : $p(A \cap B) = 0$, et événements *indépendants* : $p(A \cap B) = p(A) \times p(B)$.

2.4 Le théorème de Bayes

Considérons un événement B dont la réalisation dépend de l'intervention de l'une des causes $A_1, \dots, A_i, \dots, A_n$.

Soit $p(B|A_i)$ la probabilité conditionnelle de B , si c'est la cause A_i qui intervient.

Et soit $p(A_i)$ la probabilité d'intervention de A_i , appelée probabilité *a priori* de A_i .



Le théorème de Bayes, appelé aussi *théorème de la probabilité des causes*, calcule la probabilité $p(A_i|B)$ qui est la probabilité pour que la cause A_i ait entraîné la réalisation de B . Cette dernière probabilité est appelée la probabilité *a posteriori* de A_i .

Ce théorème, qui a plus de 2 siècles et qui était tombé en désuétude, a repris de l'intérêt vers la fin des années 1970. Il est maintenant utilisé dans de nombreux domaines, comme en reconnaissance des formes, en sûreté industrielle ou encore pour le traitement des courriels non sollicités (spams).

Par définition des probabilités conditionnelles, on peut écrire :

$$p(A_i \cap B) = p(A_i) \times p(B|A_i) = p(B) \times p(A_i|B) \implies p(A_i|B) = \frac{p(A_i) \times p(B|A_i)}{p(B)}$$

Par ailleurs, le théorème des probabilités totales permet d'écrire :

$$p(B) = \sum_{k=1}^n p(A_k \cap B) = \sum_{k=1}^n p(A_k) \times p(B|A_k)$$

D'où le théorème :

$$p(A_i|B) = \frac{p(A_i) \times p(B|A_i)}{\sum_{k=1}^n p(A_k) \times p(B|A_k)}$$

3 Variables aléatoires discrètes

3.1 Définition

À un ensemble Ω d'événements élémentaires $\{e_1, \dots, e_i, \dots, e_n\}$, faisons correspondre un nombre X prenant l'une des valeurs $x_1, \dots, x_i, \dots, x_n$, lorsque l'événement élémentaire correspondant se réalise. Le nombre X est appelé *variable aléatoire*.

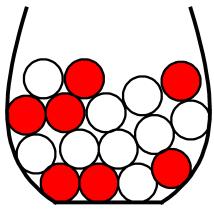
Une variable aléatoire est définie si l'on connaît les probabilités $p(x_1), \dots, p(x_i), \dots, p(x_n)$ correspondant aux différentes valeurs possibles de X .

Ces probabilités sont évidemment telles que $p(x_1) + \dots + p(x_i) + \dots + p(x_n) = 1$.

La correspondance $\{x_i, p(x_i)\}$ est appelée *loi de probabilité* (ou *distribution de probabilité*) de la variable aléatoire X .

Si les valeurs de X , en nombre fini ou infini, sont discrètes, la variable aléatoire est dite *discrète*.

Prenons l'exemple d'une urne contenant des boules blanches et des boules rouges, ces dernières en proportion ϖ .



On tire au hasard une boule dans l'urne et on considère les deux événements élémentaires :

- e_1 : la boule est blanche,
- e_2 : la boule est rouge.

Attachons à e_1 et e_2 un nombre X qui prend la valeur 0 si e_1 est réalisé et la valeur 1 si c'est e_2 .

Ce nombre est une variable aléatoire discrète qui est appelée *variable de Bernoulli*, dont la loi de probabilité est donnée par les deux probabilités :

$$p(X = 1) = \varpi \text{ et } p(X = 0) = (1 - \varpi)$$

qui ont bien une somme égale à 1.

3.2 Loi binomiale

La loi binomiale découle, elle aussi, de la prise en considération du modèle de l'urne. Soit une urne contenant des boules rouges et des boules blanches, les premières en proportion ϖ , on convient de prélever n boules et on s'intéresse à la variable aléatoire K : nombre de boules rouges parmi les n boules tirées. On se demande alors quelle est la probabilité $p(k)$ pour que K soit égale à un nombre donné k .

Il faut cependant préciser qu'il y a deux façons de procéder au tirage des n boules :

- *tirage non exhaustif* : on prélève une boule, puis une autre après remise de la première dans l'urne, puis une troisième après remise de la seconde, etc. Les prélèvements successifs sont donc indépendants puisque la composition de l'urne est la même avant chaque prélèvement ;

- *tirage exhaustif* : on prélève chacune des n boules sans remise des précédentes. Les prélèvements ne sont plus indépendants, la composition de l'urne étant modifiée après chaque prélèvement, et cela d'autant plus que la taille n de l'échantillon prélevé est élevée devant celle de la population des boules contenues dans l'urne.

Au premier mode de tirage est attachée la *loi binomiale* ; au second, la *loi hypergéométrique* qui sera envisagée à la section suivante.

Dans le cas du tirage non exhaustif, à chaque boule rouge correspond la probabilité ϖ d'être prélevée, à chaque boule blanche la probabilité $(1 - \varpi)$. Soit alors le résultat suivant : B, R, R, B, ..., R, B, qui correspond à k boules rouges et $(n - k)$ boules blanches. Les probabilités liées au tirage de chacune des boules sont respectivement égales à $(1 - \varpi), \varpi, \varpi, (1 - \varpi), \dots, \varpi, (1 - \varpi)$, du fait de la non-exhaustivité des tirages successifs. Par application du théorème des probabilités composées, la probabilité de l'ensemble de l'échantillon s'écrit alors $\varpi^k(1 - \varpi)^{n-k}$.

Cela étant posé, il faut noter que le résultat obtenu se réfère arbitrairement à un certain ordre d'arrivée des boules rouges et blanches. Or, quand on se propose de calculer la probabilité d'avoir k boules rouges parmi les n boules extraites, l'ordre d'arrivée est indifférent. Autrement dit, on n'est pas intéressé par la probabilité d'une combinaison particulière de k boules rouges parmi les n boules, mais par la probabilité de l'ensemble des $\binom{n}{k}$ combinaisons possibles. Par application du théorème des probabilités totales, cette probabilité s'écrit alors :

$$p(k) = \binom{n}{k} \varpi^k (1 - \varpi)^{n-k} \text{ encore parfois notée } C_n^k \varpi^k (1 - \varpi)^{n-k}$$

On a évidemment $\sum_{k=0}^n p(k) = 1$ puisqu'on peut écrire $\sum_{k=0}^n \binom{n}{k} \varpi^k (1 - \varpi)^{n-k} = (\varpi + (1 - \varpi))^n = 1$, chaque probabilité étant l'un des termes successifs du développement du binôme. C'est d'ailleurs à ce fait qu'est due l'appellation de loi binomiale.

La loi binomiale dépend de deux paramètres n et ϖ . Des tables numériques permettent d'obtenir les probabilités $p(k)$ pour différentes valeurs de n et ϖ ou, plus souvent, les probabilités cumulées :

$$P(k) = p(0) + p(1) + \dots + p(k) = \sum_{i=0}^k \binom{n}{i} \varpi^i (1 - \varpi)^{n-i}$$

L'intérêt pratique de la loi binomiale est très grand. Au lieu de parler d'une urne contenant une certaine proportion ϖ de boules rouges, il suffit en effet de parler d'une population contenant une certaine proportion ϖ d'individus présentant une certaine qualité ou ayant un certain avis, pour constater que ce modèle théorique permet de définir la probabilité du nombre d'individus ayant cette qualité ou cet avis, et susceptibles de figurer dans un échantillon de n individus tirés au hasard dans la population en question. La loi binomiale joue ainsi un rôle important dans un grand nombre de problèmes de jugement sur échantillon : sondages d'opinion ou contrôle du nombre de pièces défectueuses dans une fabrication.

Notons à ce stade que nous avons utilisé la majuscule P pour désigner une somme de probabilités. Il lui correspond la fonction de répartition de la variable K . Par contre, nous avons utilisé la minuscule p pour désigner une probabilité simple qui correspond à la densité de probabilité de la variable K . Nous reverrons ce vocabulaire dans le prochain chapitre.

Avec les notations précédentes, une variable aléatoire K qui suit une loi binomiale sera notée $K \sim \mathcal{B}(n, \varpi)$.

3.3 Loi hypergéométrique

Au tirage exhaustif correspond la loi hypergéométrique. Dans ce cas, la composition de l'urne est modifiée après chaque tirage. Il convient donc de préciser la composition initiale de l'urne de la façon suivante :

- N : nombre total de boules dans l'urne,
- $R = N\varpi$: nombre total de boules rouges,
- $N - R = N(1 - \varpi)$: nombre total de boules blanches.

Tirer n boules dont k rouges revient à tirer k boules parmi les R rouges et $(n - k)$ boules parmi les $(N - R)$ blanches.

Si nous individualisons chaque boule, le nombre d'échantillons de n boules que l'on peut tirer de l'urne est égal à $\binom{N}{n}$, le nombre d'échantillons de k boules rouges prélevées parmi les R rouges est égal à $\binom{R}{k}$ et celui des échantillons de $(n - k)$ boules blanches prélevées parmi les $(N - R)$ blanches est égal à $\binom{N - R}{n - k}$. Le nombre d'échantillons contenant k boules rouges et $(n - k)$ boules blanches est donc égal à $\binom{R}{k} \binom{N - R}{n - k}$. Tous ces échantillons ayant la même probabilité $\frac{1}{\binom{N}{n}}$ d'être extraits, la probabilité $p(k)$ cherchée est donc égale à :

$$p(k) = \frac{\binom{R}{k} \binom{N - R}{n - k}}{\binom{N}{n}}$$

Une variable aléatoire K qui suit une loi hypergéométrique sera notée $K \sim \mathcal{H}(N, n, R)$ avec les notations précédentes.

3.4 Loi de Poisson

À chaque couple de valeurs n et ϖ correspond, dans le cas d'un tirage non exhaustif, une loi binomiale. Pour des raisons de commodité de calcul, les statisticiens se sont efforcés de trouver des lois approchées plus facile à utiliser. La loi de Poisson est l'une d'elles. Elle correspond aux hypothèses : n grand, ϖ petit, le produit $n\varpi = \lambda$ étant fini. On peut écrire dans ces conditions :

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

Si l'on fait tendre n vers l'infini, $\frac{n(n-1)\cdots(n-k+1)}{n^k} \rightarrow 1$, $(1 - \frac{\lambda}{n})^k \rightarrow 1$ et on peut montrer que $(1 - \frac{\lambda}{n})^n \rightarrow e^{-\lambda}$. À la limite, on obtient alors la loi de Poisson définie par les probabilités :

$$p(k) = e^{-\lambda} \times \frac{\lambda^k}{k!}$$

Avec *Mathematica* :

$$\begin{aligned} & \left\{ \text{Limit}\left[\frac{n!}{(n-k)! n^k}, n \rightarrow \infty \right], \text{Limit}\left[\left(1 - \frac{\lambda}{n}\right)^n, n \rightarrow \infty \right] \right\} \\ & \{1, e^{-\lambda}\} \end{aligned}$$

Elle dépend du seul paramètre λ et il existe des tables qui donnent, pour différentes valeurs de λ , les probabilités cumulées correspondantes :

$$P(k) = \sum_{i=0}^k e^{-\lambda} \times \frac{\lambda^i}{i!}$$

La loi de Poisson présente donc l'intérêt de simplifier les calculs, puisqu'une seule table poissonnienne se substitue à un grand nombre de tables binomiales. En pratique et en première analyse, on peut utiliser l'approximation de Poisson quand $n \geq 50$ et $\varpi \leq 0.1$. Ceci lui confère un champ d'application très large, en particulier dans l'échantillonnage industriel où les proportions de déchets sont heureusement faibles.

Une variable K suivant une loi de Poisson de paramètre λ sera notée $K \sim \mathcal{P}(\lambda)$

3.5 Processus de Poisson

Mais en réalité l'importance de la loi de Poisson dépasse de beaucoup ce seul cadre. Elle peut être obtenue de façon toute différente et très intéressante du point de vue des applications pratiques.

Considérons une suite d'événements tels que :

- les événements sont indépendants ;
- la probabilité d'apparition d'un événement pendant un intervalle de temps Δt est proportionnelle à Δt , égale à $\lambda \Delta t$, et la probabilité pour qu'il se produise 2 événements pendant Δt est du second ordre par rapport à la première ;
- le phénomène est stationnaire, c'est-à-dire que ses caractéristiques ne dépendent pas de l'origine du temps d'observation.

On dit alors qu'on a affaire à un processus de Poisson dont λ est le *taux d'arrivée*.

Considérons la situation du processus à l'instant $t + \Delta t$ et supposons qu'il y a eu k événements enregistrés jusqu'à cet instant. Cela ne peut provenir que des deux situations suivantes :

nombre d'événements enregistrés jusqu'à l'instant	
t	$t + \Delta t$
$k - 1$	k
k	k

La probabilité $p_k(t + \Delta t)$ de réalisation de k événements jusqu'à l'instant $t + \Delta t$ est donc :

$$p_k(t + \Delta t) = \lambda \Delta t p_{k-1}(t) + (1 - \lambda \Delta t) p_k(t)$$

puisque $\lambda \Delta t$ est la probabilité de réalisation d'un événement pendant un intervalle de temps Δt et que $(1 - \lambda \Delta t)$ est la probabilité du contraire.

Cette relation n'est vraie que pour les valeurs non nulles de k : si aucun événement ne s'est réalisé avant $t + \Delta t$, c'est qu'aucun événement n'était réalisé avant t . On a donc :

$$p_0(t + \Delta t) = (1 - \lambda \Delta t) p_0(t)$$

que l'on peut écrire :

$$\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t)$$

En faisant tendre Δt vers 0, on obtient : $p'_0(t) = -\lambda p_0(t)$ et, par conséquent : $p_0(t) = e^{-\lambda t}$ puisque $p_0(t)$ est égale à 1 à l'instant $t = 0$.

Avec *Mathematica* :

$$\begin{aligned} \text{DSolve}[\{p_0'[t] == -\lambda p_0[t], p_0[0] == 1\}, p_0[t], t] \\ \{ \{p_0[t] \rightarrow e^{-t \lambda}\} \} \end{aligned}$$

La relation qui correspond aux valeurs non nulles de k s'écrit, de la même façon :

$$\frac{p_k(t + \Delta t) - p_k(t)}{\Delta t} = p'_k(t) = \lambda p_{k-1}(t) - \lambda p_k(t)$$

et permet d'obtenir, par récurrence :

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

On reconnaît la loi de Poisson de paramètre λt .

Le champ d'application du processus de Poisson est très vaste. L'expérience montre en effet qu'il permet de modéliser de nombreux phénomènes. Il s'agira, par exemple, du nombre de particules émises par du radium, du nombre de défaillances d'une certaine machine, du nombre de clients qui se présentent à un guichet, du nombre d'appels à un standard téléphonique, du nombre de ruptures de fil dans une tréfilerie, ou du nombre de secousses sismiques ressenties dans la ville de Mexico.

3.6 Espérance mathématique d'une variable aléatoire discrète

Etant donnée une variable aléatoire notée X , susceptible de prendre les valeurs $x_1, \dots, x_i, \dots, x_n$ avec les probabilités $p(x_1), \dots, p(x_i), \dots, p(x_n)$. On appelle *espérance mathématique* de la variable X , la quantité :

$$\mathbb{E}(X) = \sum_{i=1}^n x_i p(x_i)$$

L'appellation d'*espérance mathématique*, due à Blaise Pascal, sera justifiée ultérieurement par la loi des grands nombres, qui énonce que la moyenne d'une série d'observations d'une variable tend vers l'espérance de cette variable, quand le nombre d'observations augmente indéfiniment.

En appliquant cette définition aux variables qui viennent d'être étudiées, on trouve les valeurs suivantes.

3.6.1 Loi de Bernoulli

$$\mathbb{E}(X) = 1 \times \varpi + 0 \times (1 - \varpi) = \varpi$$

3.6.2 Loi binomiale

$$\mathbb{E}(K) = \sum_{k=1}^n k \binom{n}{k} \varpi^k (1 - \varpi)^{n-k}$$

expression que l'on peut encore écrire :

$$\mathbb{E}(K) = \sum_{k=1}^n n \varpi \times \left[\frac{(n-1) \dots (n-(k-1))}{(k-1)!} \varpi^{k-1} (1 - \varpi)^{(n-1)-(k-1)} \right]$$

et où l'on reconnaît :

$$\mathbb{E}(K) = n \varpi (\varpi + (1 - \varpi))^{n-1} = n \varpi$$

3.6.3 Loi de Poisson

Un calcul analogue donne :

$$\mathbb{E}(K) = \lambda$$

3.6.4 Processus de Poisson

$$\mathbb{E}(K) = \lambda t$$

On comprend alors pourquoi λ , qui est le nombre moyen d'événements par unité de temps, s'appelle le taux d'arrivée.

Exercices du chapitre 1

Exercice 1

On tire, de façon exhaustive, deux cartes d'un jeu de 52. Quelle est la probabilité que l'une au moins soit un valet.

Exercice 2

Monsieur et Madame Dupont ont deux enfants.

- Ils vous informent que l'un des deux est une fille. Quelle est la probabilité pour que l'autre soit un garçon ?
- Ils précisent que l'aîné de leurs deux enfants est une fille. Quelle est la probabilité pour que le plus jeune enfant soit un garçon ?

Exercice 3

Un jeu télévisé propose aux candidats de choisir une porte parmi 3. Derrière l'une d'elle se trouve un cadeau de grande valeur. Après avoir fait son choix, le présentateur fait ouvrir une des deux portes non choisies derrière laquelle il n'y a rien et demande alors au candidat s'il veut changer de porte. Est-il avantageux pour le candidat de changer son choix initial ?

Exercice 4

Un lot d'articles doit être testé par prélèvement, aux fins de réception par l'acheteur. Au regard de critères purement techniques, un article peut être classé, après essai, en bon ou mauvais.

Le test est réglé de la façon suivante. On prend deux articles au hasard dans le lot. S'ils sont bons, tous les deux, le lot est accepté. Si les deux sont mauvais, le lot est refusé. Si l'un est mauvais, l'autre bon, on tire de nouveau deux articles au hasard. Si ces deux derniers articles sont bons, le lot est accepté. Sinon, le lot est définitivement refusé.

Calculer en fonction de la proportion ϖ d'articles défectueux dans le lot (proportion réelle, mais inconnue de la personne qui fait le test) la probabilité p d'acceptation du lot. On admettra que les tirages sont faits de façon non exhaustive ou, ce qui revient au même, que la taille du lot est suffisamment grande pour que la proportion ϖ reste inchangée quel que soit le résultat d'un tirage. Tracer la courbe $p(\varpi)$, appelée « courbe d'efficacité ».

Exercice 5

La probabilité d'atteindre un certain objectif avec un certain canon est égale à 5%. Quel est le nombre minimum de coups à tirer pour avoir une probabilité de 99% d'atteindre au moins une fois l'objectif ?

Exercice 6

Un entrepreneur de transports possède deux autocars qu'il peut louer chaque jour pour la journée. Le nombre de demandes présentées par jour est distribué approximativement suivant une loi de Poisson de moyenne égale à 1,5.

a) Calculer :

- la proportion des jours pour lesquels aucune demande n'est présentée,
 - la proportion des jours pour lesquels les demandes ne peuvent pas être entièrement satisfaites.
- b) Si les deux véhicules sont utilisés de façon à faire le même nombre de demandes, quelle est la proportion des jours pour lesquels un autocar donné n'est pas en service ?
- c) Quelle est la proportion du nombre total des demandes qui est refusée ?

Exercice 7

Des candidats n'ayant pas assez soigné la présentation matérielle de leur copie, le correcteur, plutôt que de chercher à lire de tels brouillons, décide de noter au hasard en accordant une égale probabilité à toutes les notes entières possibles de 0 à 20. Quelle est la loi de probabilité que suit la meilleure note du groupe ?

Exercice 8

L'assemblage de l'aile d'un avion nécessite 2500 rivets. La probabilité pour que l'un des rivets utilisés soit défectueux est égale à 0,002. Quelles sont les probabilités pour que, sur l'aile :

- a) Il n'y ait aucun rivet défectueux.
- b) Il y en ait au moins 10.

Exercice 9

On estime que 50% des gens répondent à un questionnaire immédiatement et que 60% de ceux qui ne répondent pas immédiatement répondent après un rappel. Un questionnaire est envoyé à 40 personnes et une lettre de rappel à ceux qui ne répondent pas immédiatement. Quelle est la probabilité d'avoir finalement au moins 30 réponses au questionnaire après rappel ?

Exercice 10

Une épidémie s'est déclarée dans une petite ville. D'après l'ensemble des symptômes, l'origine est soit un streptocoque, soit un virus. L'origine par streptocoque est plus dangereuse et requiert des mesures particulières, mais les statistiques montrent qu'elle est 4 fois moins probable que l'origine par virus.

Les analyses de laboratoire permettent de déceler la présence éventuelle de streptocoque, mais les techniques utilisées présentent des risques d'erreur : le streptocoque a environ 3 chances sur 10 de n'être pas décelé et, lorsqu'il n'est pas présent chez le malade, 1 chance sur 10 d'être décelé dans les préparations, par suite d'erreur ou de contamination.

- a) Les analyses pratiquées sur 5 malades ayant donné pour la présence de streptocoque : {oui, non, oui, non, oui}, quel diagnostic porter sur l'origine de l'épidémie ?
- b) Souligner les points qui, dans ce résultat, semblent en valoir la peine.

Exercice 11

Pour le stockage d'un produit extrêmement toxique, on a installé en parallèle 3 détecteurs de fuite, de même type, autour de la cuve de stockage. On sait que ce type de détecteur se déclenche intempestivement dans 10% des cas et ne détecte pas la fuite dans 2% des cas. Par ailleurs, des calculs ont montré que la probabilité d'une fuite dans cette unité de stockage était égale à 5%. Quelle est la probabilité d'une fuite si 2 détecteurs ont réagi ?

La loi normale

L'objectif du chapitre est de présenter la loi normale qui est le modèle probabiliste le plus utilisé pour décrire de très nombreux phénomènes observés dans la pratique.

Une grande attention devra être accordée aux concepts, essentiels en statistique, d'espérance mathématique et de variance et aux opérations qui leurs sont attachées.

1 Variables aléatoires continues

1.1 Loi de probabilité

Rappelons que si, à un ensemble de possibilités Ω , nous attachons un nombre X prenant les valeurs $x_1, \dots, x_i, \dots, x_n$, lorsque se produit l'un des événements $e_1, \dots, e_i, \dots, e_n$, on dit que X est une variable aléatoire. La loi de probabilité de cette variable est définie lorsqu'on connaît les probabilités :

$$p(x_1), \dots, p(x_i), \dots, p(x_n)$$

correspondant aux valeurs possibles de X . Ces probabilités sont obligatoirement telles que :

$$p(x_1) + \dots + p(x_i) + \dots + p(x_n) = 1$$

et la correspondance $\{x_i, p(x_i)\}$ est appelée *loi ou distribution de probabilité* de la variable aléatoire X .

Si les valeurs possibles de X sont réparties de façon continue sur un intervalle fini ou infini, X est une *variable aléatoire continue*. Une telle variable est définie si l'on connaît la probabilité que X prenne une valeur dans tout intervalle $[x, x+h]$. On se donne pour cela la *fonction de répartition* de X :

$$P(x) = \text{Prob} \{X \leq x\} = \mathbb{P}(X \leq x)$$

qui permet de calculer, pour tout intervalle :

$$\mathbb{P}(x < X \leq x+h) = P(x+h) - P(x)$$

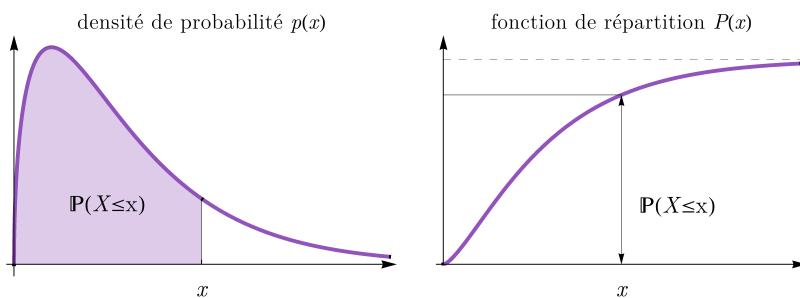
Un cas particulier important, auquel nous nous attacherons exclusivement dans ce qui suit, est celui où la fonction de répartition est continue et peut être mise sous la forme :

$$P(x) = \int_{-\infty}^x p(u) du$$

où $p(x)$ s'appelle la *densité de probabilité* de X , appellation qui résulte du fait que :

$$p(x) = \lim_{\Delta x \rightarrow 0} \frac{\mathbb{P}(x < X \leq x+\Delta x)}{\Delta x}.$$

La densité de probabilité $p(x)$ ou la fonction de répartition $P(x)$ définissent la loi de probabilité d'une variable aléatoire continue X . Elles donnent lieu aux représentations graphiques suivantes :



Il est important de bien noter que, conformément aux axiomes qui définissent les probabilités :

$$\int_{-\infty}^{+\infty} p(x) dx = 1 \text{ et } 0 \leq P(x) \leq 1$$

La densité de probabilité et la fonction de répartition sont fréquemment notées respectivement $f(x)$ et $F(x)$.

1.2 La loi uniforme

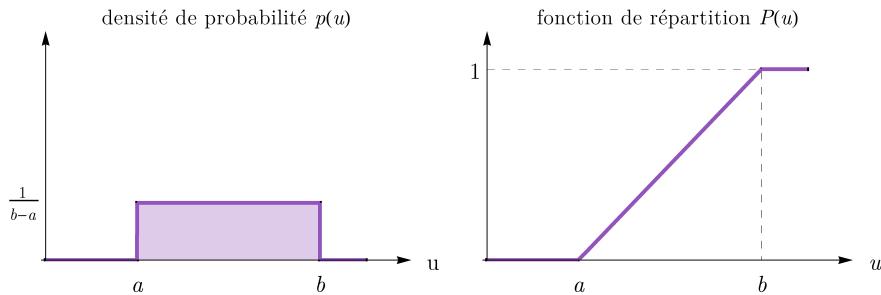
La variable aléatoire U est distribuée *uniformément* sur l'intervalle $[a, b]$ si sa densité de probabilité est constante sur cet intervalle et égale à :

$$p(u) = \frac{1}{b-a}$$

et si sa fonction de répartition a , par conséquent, l'équation suivante :

$$P(u) = \frac{u-a}{b-a}$$

Elle donne donc lieu aux représentations suivantes :



1.3 La loi exponentielle

Nous avons montré dans le chapitre précédent que si la probabilité d'apparition d'un événement pendant un intervalle de temps Δt était égale à $\lambda \Delta t$, la probabilité pour qu'il se produise k fois pendant un intervalle de temps t , était donnée par la loi de Poisson de paramètre λt :

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Considérons maintenant les intervalles de temps T qui s'écoulent entre les évènements successifs d'un processus de Poisson et $F(t)$ la fonction de répartition de T . On a donc :

$$F(t) = \mathbb{P}(T \leq t) = 1 - \mathbb{P}(T > t)$$

Or cette dernière probabilité est égale à la probabilité pour qu'il ne se produise aucun événement jusqu'à l'instant t , c'est-à-dire $p_0(t)$. Par conséquent, on peut écrire :

$$F(t) = 1 - e^{-\lambda t}$$

C'est la loi exponentielle dont la densité de probabilité s'écrit $p(t) = \lambda e^{-\lambda t}$. Elle peut constituer un modèle intéressant pour les durées de vie aléatoires de certains matériels (tubes électroniques par exemple) : à chaque instant t de la vie du matériel, la probabilité de défaillance pendant l'intervalle de temps Δt qui suit, est indépendante de t et égale à $\lambda \Delta t$.

1.4 Loi de probabilité à deux dimensions

Si à un événement aléatoire sont attachés deux nombres X et Y , ces deux nombres définissent un *vecteur aléatoire à deux dimensions*. La loi de probabilité d'un tel vecteur peut être définie par la fonction de répartition :

$$P(x, y) = \mathbb{P}(X \leq x, Y \leq y)$$

où la virgule se lit « et ». Si cette fonction est dérivable en x et y , on peut définir la densité de probabilité $p(x, y)$ qui est telle que :

$$p(x, y) dx dy = \mathbb{P}(x < X \leq x + dx, y < Y \leq y + dy)$$

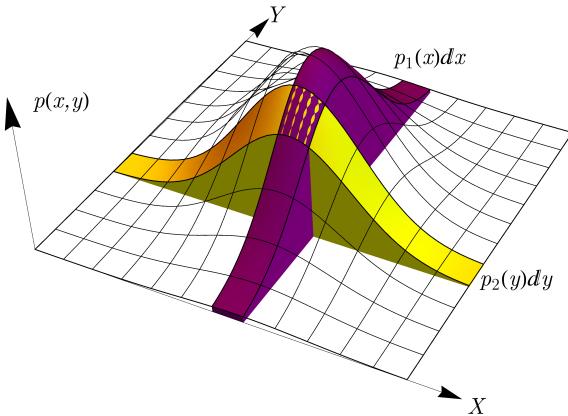
Géométriquement $p(x, y) dx dy$ peut s'interpréter comme la probabilité pour que l'extrémité du vecteur aléatoire (X, Y) se trouve dans une aire $ds = dx dy$ autour du point (x, y) .

S'intéressant à la variable X , par exemple, indépendamment de la variable Y , on obtient la *distribution marginale* de X en calculant sa densité de probabilité $p_1(x)$:

$$p_1(x) dx = \mathbb{P}(x < X \leq x + dx) = dx \int_{-\infty}^{+\infty} p(x, y) dy$$

De même la distribution marginale de Y est définie par la densité de probabilité $p_2(y)$:

$$p_2(y) dy = \mathbb{P}(y < Y \leq y + dy) = dy \int_{-\infty}^{+\infty} p(x, y) dx$$



1.5 Indépendance de deux variables aléatoires

Par définition, deux variables aléatoires sont *indépendantes* si la probabilité pour que la valeur de l'une d'elle tombe dans un intervalle donné, ne dépend pas de la valeur prise par l'autre. La probabilité conditionnelle de X , par exemple, est donc indépendante de Y . Or elle s'écrit, avec les notations ci-dessus et grâce au théorème des probabilités composées :

$$\mathbb{P}(x \leq X < x + dx / y \leq Y < y + dy) = \frac{p(x,y) dx dy}{p_2(y) dy} = \frac{p(x,y) dx}{p_2(y)}$$

Si les deux variables sont indépendantes, cette expression doit dépendre de x seulement et l'on doit donc pouvoir faire disparaître $p_2(y)$ par simplification. De la même façon, l'expression :

$$\mathbb{P}(y \leq Y < y + dy / x \leq X < x + dx) = \frac{p(x,y) dy}{p_1(x)}$$

doit dépendre de y seulement et l'on doit pouvoir la simplifier pour faire disparaître $p_1(x)$. Il en résulte que $p(x, y)$ doit être égal au produit des densités de probabilité marginales de X et de Y :

$$p(x, y) = p_1(x) p_2(y)$$

Cette condition est évidemment *nécessaire et suffisante*. Elle se généralise pour un nombre quelconque de variables aléatoires indépendantes dans leur ensemble.

1.6 Fonctions de variables aléatoires

Soient une fonction f et une variable aléatoire X dont la densité de probabilité est notée $p(x)$. La variable aléatoire $f(X)$, *fonction de la variable aléatoire X*, est définie de la façon suivante : $f(X)$ prend la valeur $f(x)$, lorsque X prend la valeur x .

On remarquera qu'une variable $f(X)$ peut prendre la même valeur pour deux valeurs différentes x_i et x_j de X . La probabilité de la valeur $f(x_i)$ ou $f(x_j)$ est alors égale à $p(x_i) + p(x_j)$.

On peut aussi définir des fonctions de plusieurs variables aléatoires. La *somme de deux variables aléatoires*, notamment, se définit de la façon suivante : étant donné deux variables aléatoires X et Y , leur somme est la variable aléatoire $(X + Y)$ qui prend la valeur $(x + y)$ lorsque X prend la valeur x et Y prend la valeur y . Là encore, une même valeur de la somme peut être obtenue pour deux couples différents de valeurs de X et Y . Penser, par exemple, à la variable *somme de deux dés jetés*.

2 Espérance mathématique

2.1 Espérance et moments d'une variable aléatoire

Etant données une variable aléatoire X dont la densité de probabilité est $p(x)$ et une fonction f , on désigne par le terme d'espérance mathématique de la variable aléatoire $f(X)$, notée $\mathbb{E}[f(X)]$, l'expression :

$$\mathbb{E}[f(X)] = \int_{-\infty}^{+\infty} f(x) p(x) dx$$

C'est donc un *opérateur* qui transforme la variable aléatoire $f(X)$ en un nombre. Appliqué à la variable X elle-même, l'opérateur donne sa moyenne μ :

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{+\infty} x p(x) dx$$

Dans le cas d'une variable aléatoire discrète $x_1, \dots, x_i, \dots, x_n$, l'expression précédente devient :

$$\mu = \mathbb{E}[f(X)] = \sum_{i=1}^n f(x_i) p(x_i)$$

et toutes les propriétés de l'opérateur \mathbb{E} que nous démontrerons par la suite, pour une variable continue, s'étendent sans difficulté au cas d'une variable discrète.

Lorsque $f(X)$ est une puissance de X , l'expression $\mathbb{E}(X^k)$ est appelée *moment* d'ordre k de la variable aléatoire X . La moyenne $\mu = \mathbb{E}(X)$ est ainsi le moment d'ordre 1 de la variable X . Elle s'interprète comme l'abscisse du centre de gravité de la distribution de probabilité et c'est, à ce titre, une caractéristique de tendance centrale de la distribution : les valeurs d'une variable aléatoire se répartissent autour de sa moyenne.

2.2 Variance et écart-type

Il peut alors s'avérer intéressant de rapporter une variable aléatoire à sa moyenne, autrement dit de la *centrer*. On obtient alors le *moment centré* d'ordre k :

$$\mathbb{E}[(X - \mu)^k] = \int_{-\infty}^{+\infty} (x - \mu)^k p(x) dx$$

Le moment centré d'ordre 2 est appelé *variance* et il revêt une importance toute particulière. Nous le noterons $\mathbb{V}(X)$ ou très souvent σ^2 :

$$\sigma^2 = \mathbb{V}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$$

Par analogie avec la mécanique, on peut dire que μ est le centre de gravité et que σ^2 est le moment d'inertie de la distribution.

Sa racine carrée σ est appelée l'*écart-type*, qui s'interprète comme une mesure de la dispersion d'une variable aléatoire : plus les valeurs de la variable sont susceptibles de s'éloigner de sa moyenne, plus son écart-type est grand. C'est ce que montre l'inégalité de Bienaymé-Tchebychev.

2.3 Inégalité de Bienaymé-Tchebychev

Soit X une variable aléatoire de moyenne μ et d'écart-type σ , mais à ceci près quelconque. La probabilité pour qu'elle prenne une valeur à l'extérieur d'un intervalle $[\mu - a, \mu + a]$, où a est un nombre positif, est donnée par l'intégrale :

$$\mathbb{P}(|X - \mu| > a) = \int_{|x-\mu|>a} p(x) dx$$

Pour les valeurs x extérieures à l'intervalle $[\mu - a, \mu + a]$, $(x - \mu)^2$ est supérieur à a^2 , on en déduit :

$$\mathbb{P}(|X - \mu| > a) < \frac{1}{a^2} \int_{|x-\mu|>a} (x - \mu)^2 p(x) dx$$

On majore encore l'intégrale en intégrant de $-\infty$ à $+\infty$:

$$\mathbb{P}(|X - \mu| > a) < \frac{1}{a^2} \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx$$

C'est-à-dire :

$$\mathbb{P}(|X - \mu| > a) < \frac{\sigma^2}{a^2}$$

C'est l'inégalité de Bienaymé-Tchebychev qui s'interprète ainsi : plus σ est petit, plus les grandes valeurs de $|X - \mu|$ sont improbables, donc moins la variable X est dispersée autour de sa moyenne μ .

Si l'on pose $a = t\sigma$, l'inégalité s'écrit :

$$\mathbb{P}(|X - \mu| > t\sigma) < \frac{1}{t^2}$$

et exprime que la probabilité pour que X s'éloigne de sa moyenne de plus que t fois son écart-type σ , est inférieure à $\frac{1}{t^2}$. Il y a, en particulier, moins de 1 chance sur 4 pour que X s'éloigne de μ de plus de 2 fois σ et il est très improbable (probabilité inférieure à $\frac{1}{100}$) que X s'éloigne de μ de plus de 10 fois σ .

Ces résultats sont très généraux puisqu'ils ne nécessitent aucune hypothèse sur la forme de la loi de probabilité de X . On peut, bien sûr, les rendre beaucoup plus précis et les serrer davantage dans chaque cas particulier où la loi de probabilité de X est connue.

2.4 Linéarité de l'opérateur Espérance

On montre très facilement que, a et b étant deux constantes, on a :

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Soient maintenant deux variables X et Y définies par leur densité de probabilité $p(x, y)$. L'espérance mathématique de leur somme s'écrit :

$$\begin{aligned}\mathbb{E}(X + Y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y) p(x, y) dx dy \\ \mathbb{E}(X + Y) &= \int_{-\infty}^{+\infty} x dx \int_{-\infty}^{+\infty} p(x, y) dy + \int_{-\infty}^{+\infty} y dy \int_{-\infty}^{+\infty} p(x, y) dx \\ \mathbb{E}(X + Y) &= \int_{-\infty}^{+\infty} x p_1(x) dx + \int_{-\infty}^{+\infty} y p_2(y) dy \\ \mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y)\end{aligned}$$

Cette propriété s'applique, bien entendu, à la somme d'un nombre quelconque de variables.

On notera, d'autre part, que la démonstration n'a requis aucune hypothèse sur l'indépendance des variables aléatoires considérées.

2.5 Calculs de variances de variables aléatoires

Rappelons qu'étant donnée une variable aléatoire de moyenne μ , sa variance s'écrit :

$$\mathbb{V}(X) = \mathbb{E}[(X - \mu)^2]$$

De la forme même de la variance, il résulte que :

- la variance d'une constante est égale à 0,
- dans un changement d'échelle de rapport k , la variance est multipliée par k^2 .

Si, d'autre part, on développe l'expression de la variance, et qu'on utilise les propriétés de linéarité de l'espérance, il vient :

$$\mathbb{V}(X) = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu \mathbb{E}(X) + \mu^2$$

et, puisque $\mathbb{E}(X) = \mu$:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu^2$$

Cette relation sera très souvent utilisée. On la retiendra facilement en écrivant que :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

et en énonçant que : *la variance est égale à l'espérance du carré moins le carré de l'espérance.* Appliquons la à quelques-unes des variables aléatoires déjà rencontrées.

2.5.1 Loi uniforme

Soit la variable U distribuée uniformément sur l'intervalle $[0, a]$. Elle a pour densité de probabilité $\frac{1}{a}$, donc pour moyenne :

$$\mathbb{E}(U) = \int_0^a x \frac{1}{a} dx = \frac{a}{2}$$

pour moment d'ordre 2 :

$$\mathbb{E}(U^2) = \int_0^a x^2 \frac{1}{a} dx = \frac{a^2}{3}$$

et, enfin, pour variance :

$$\mathbb{V}(U) = \mathbb{E}(U^2) - \mathbb{E}(U)^2 = \frac{a^2}{12}$$

Avec *Mathematica* :

```
dist = UniformDistribution[{0, a}]; {Mean[dist], Variance[dist]}
{a/2, a^2/12}
```

2.5.2 Variable de Bernoulli

Les moments d'ordres 1 et 2 sont égaux :

$$\mathbb{E}(X) = \mathbb{E}(X^2) = 1 \times \varpi + 0 \times (1 - \varpi) = \varpi$$

La variance est donc égale à :

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \varpi - \varpi^2 = \varpi(1 - \varpi)$$

Avec *Mathematica* :

```
dist = BernoulliDistribution[\varpi]; {Mean[dist], Variance[dist]}
{\varpi, (1 - \varpi) \varpi}
```

2.5.3 Loi de Poisson

La loi de Poisson est définie par la densité de probabilité :

$$p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

et nous avons déjà calculé sa moyenne :

$$\mathbb{E}(K) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda \sum_{k=0}^{\infty} p(k-1) = \lambda$$

Pour trouver sa variance, calculons d'abord et selon le même principe :

$$\mathbb{E}[K(K-1)] = \sum_{k=0}^{\infty} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=2}^{\infty} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = \lambda^2 \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2 \sum_{k=0}^{\infty} p(k-2) = \lambda^2$$

On a, par conséquent :

$$\mathbb{E}[K(K-1)] = \mathbb{E}(K^2) - \mathbb{E}(K) = \lambda^2$$

d'où l'on déduit que :

$$\mathbb{E}(K^2) = \lambda^2 + \lambda$$

et enfin que :

$$\mathbb{V}(K) = \mathbb{E}(K^2) - \mathbb{E}(K)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

La loi de Poisson est donc telle que moyenne et variance soient toutes les deux égales à λ .

Avec *Mathematica* :

```
dist = PoissonDistribution[λ]; {Mean[dist], Variance[dist]}

{λ, λ}
```

2.5.4 Loi exponentielle

La densité de probabilité est :

$$p(t) = \lambda e^{-\lambda t}$$

En intégrant par parties, on trouve pour la moyenne :

$$\mathbb{E}(T) = \int_0^\infty t \lambda e^{-\lambda t} dt = \frac{1}{\lambda}$$

puis, pour le moment d'ordre 2 :

$$\mathbb{E}(T^2) = \int_0^\infty t^2 \lambda e^{-\lambda t} dt = \frac{2}{\lambda^2}$$

ce qui montre que l'écart-type est égal à la moyenne :

$$\mathbb{V}(T) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Avec *Mathematica* :

```
Integrate[t λ e^{-λ t}, {t, 0, ∞}, Assumptions → {λ > 0}]

1
λ

Integrate[t^2 λ e^{-λ t}, {t, 0, ∞}, Assumptions → {λ > 0}]

2
λ^2

dist = ExponentialDistribution[λ]; {Mean[dist], Variance[dist]}

{1/λ, 1/λ^2}
```

2.6 Indépendance et covariance de deux variables aléatoires

Soient X et Y deux variables aléatoires définies par leur densité de probabilité $p(x, y)$. L'espérance mathématique du produit de ces deux variables est, par définition :

$$\mathbb{E}(X Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x y p(x, y) dx dy$$

Si les deux variables aléatoires sont *indépendantes* :

$$p(x, y) = p_1(x) p_2(y)$$

et $\mathbb{E}(X Y) = \int_{-\infty}^{+\infty} x p_1(x) dx \int_{-\infty}^{+\infty} y p_2(y) dy = \mathbb{E}(X) \mathbb{E}(Y)$

Il s'agit là d'un théorème très important qui peut s'énoncer de la façon suivante : *si deux variables aléatoires sont indépendantes, l'espérance de leur produit est égale au produit de leurs espérances.*

Ce théorème justifie la définition d'une quantité appelée la *covariance* de X et Y :

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

qui a la propriété suivante : *la covariance de deux variables aléatoires indépendantes est nulle.*

Il faut faire attention au fait que la réciproque n'est pas vraie, en général.

2.7 Variance d'une somme de variables aléatoires indépendantes

Soient X et Y deux variables aléatoires. La variance de leur somme (ou de leur différence) peut s'écrire (espérance du carré moins carré de l'espérance) :

$$\mathbb{V}(X \pm Y) = \mathbb{E}[(X \pm Y)^2] - \mathbb{E}(X \pm Y)^2$$

Le premier terme du second membre se développe en :

$$\mathbb{E}[(X \pm Y)^2] = \mathbb{E}(X^2) \pm 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$$

et le second terme en :

$$\mathbb{E}(X \pm Y)^2 = \mathbb{E}(X)^2 \pm 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2$$

On a donc, en réorganisant les termes :

$$\mathbb{V}(X \pm Y) = (\mathbb{E}(X^2) - \mathbb{E}(X)^2) \pm 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) + (\mathbb{E}(Y^2) - \mathbb{E}(Y)^2)$$

où l'on reconnaît :

$$\mathbb{V}(X \pm Y) = \mathbb{V}(X) \pm 2\text{Cov}(X, Y) + \mathbb{V}(Y)$$

Si, maintenant, les deux variables X et Y sont *indépendantes*, leur covariance est nulle, et :

$$\mathbb{V}(X \pm Y) = \mathbb{V}(X) + \mathbb{V}(Y)$$

propriété qui s'énonce ainsi : la variance de la somme ou de la différence de deux variables aléatoires indépendantes est égale à la *somme* de leurs variances.

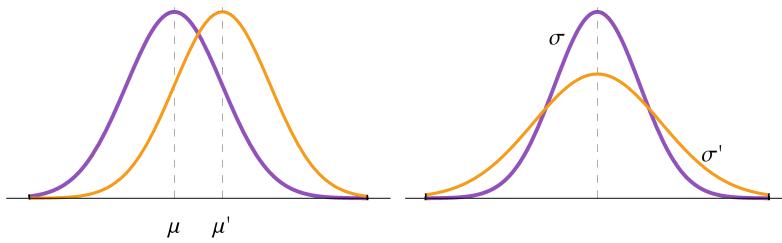
3 Loi normale

3.1 Définition et propriétés

La loi de probabilité la plus utilisée en statistique est la loi normale, encore appelée *loi de Gauss* ou *de Laplace-Gauss*. Une variable aléatoire X suit une loi normale si sa densité de probabilité a pour équation :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Son graphe est la fameuse *courbe en cloche* dont le tracé dépend des deux paramètres μ et σ qui ne sont autres que la *moyenne* et l'*écart-type* de X comme on le montrera au paragraphe suivant.



La distribution est symétrique par rapport à μ qui caractérise donc la tendance centrale. Quant à σ , il caractérise la dispersion de la distribution. Plus il est grand, plus la distribution est étalée de part et d'autre de μ . Les points d'inflexion se trouvent à $(\mu - \sigma)$ et $(\mu + \sigma)$, avec :

$$\mathbb{P}(|X - \mu| \leq \sigma) = 68.26 \%$$

Il sera bon de conserver en mémoire deux autres valeurs intéressantes :

$$\mathbb{P}(|X - \mu| \geq 1.96 \sigma) = 5 \%$$

$$\mathbb{P}(|X - \mu| \geq 2.58 \sigma) = 1 \%$$

que l'on pourra aussi comparer à celles que donnait l'approximation de *Bienaymé-Tchebychev* pour une loi de probabilité quelconque.

Le fait qu'une variable X suive une loi normale de paramètres μ et σ se note généralement $X \sim \mathcal{N}(\mu, \sigma^2)$, loi normale de moyenne μ et variance σ^2 .

3.2 Calcul de la moyenne et de la variance

Nous allons montrer que la moyenne et la variance d'une variable qui suit une loi normale de paramètres μ et σ sont respectivement égales à μ et σ^2 . La démonstration est uniquement calculatoire et pourra être omise.

Par définition, la moyenne est égale à :

$$\mathbb{E}(X) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Pour calculer cette intégrale, faisons le changement de variable, classique pour les calculs sur la loi normale :

$$u = \frac{x-\mu}{\sigma}$$

Il vient alors :

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\mu + \sigma u) e^{-\frac{u^2}{2}} du \\ \mathbb{E}(X) &= \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u e^{-\frac{u^2}{2}} du \end{aligned}$$

Or :

$$\int_{-\infty}^{+\infty} u e^{-\frac{u^2}{2}} du = 0 \text{ et } \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du = \sqrt{2\pi}$$

Ce dernier résultat se montre en posant :

$$I = \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du$$

d'où :

$$I^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+y^2)} dx dy = 2\pi$$

que l'on calcule facilement en passant en coordonnées polaires. Finalement, on arrive à :

$$\mathbb{E}(X) = \mu$$

De la même façon :

$$\mathbb{E}[(X - \mu)^2] = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} (x - \mu)^2 e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

qui devient, après changement de variable :

$$\mathbb{E}[(X - \mu)^2] = \frac{\sigma^3}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}} du$$

L'intégrale se calcule par parties et vaut $\sqrt{2\pi}$. Il vient alors :

$$\mathbb{E}[(X - \mu)^2] = \sigma^2$$

Avec Mathematica :

$$\left\{ \int_{-\infty}^{+\infty} u e^{-\frac{u^2}{2}} du, \int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}} du, \int_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}} du, \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x^2+y^2)} dy dx \right\}$$

$$\{0, \sqrt{2\pi}, \sqrt{2\pi}, 2\pi\}$$

$$\frac{1}{\sigma \sqrt{2\pi}} \text{Integrate}[x e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \{x, -\infty, +\infty\}, \text{Assumptions} \rightarrow \{\sigma > 0\}]$$

$$\mu$$

$$\frac{1}{\sigma \sqrt{2\pi}} \text{Integrate}[(x - \mu)^2 e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \{x, -\infty, +\infty\}, \text{Assumptions} \rightarrow \{\sigma > 0\}]$$

$$\sigma^2$$

3.3 La loi normale centrée réduite

Etant donnée une variable X suivant une loi normale de moyenne μ et d'écart-type σ , sa densité de probabilité est telle que :

$$p(x) dx = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Faisant le changement de variable $U = \frac{X-\mu}{\sigma}$, on peut écrire $du = \frac{dx}{\sigma}$, d'où :

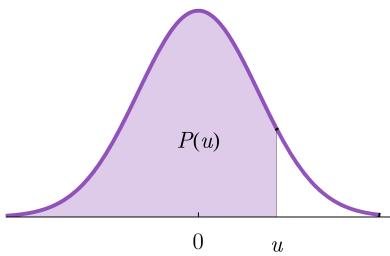
$$p(u) du = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

La nouvelle variable suit donc une loi normale qui est dite *centrée* ($\mathbb{E}(U) = 0$) et *réduite* ($\mathbb{V}(U) = 1$) et qui ne fait intervenir aucun paramètre.

Cette loi, appelée *loi normale centrée réduite*, est tabulée. Les tables donnent généralement, pour différentes valeurs de u , les probabilités :

$$\mathbb{P}(-\infty < U \leq u) = P(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{t^2}{2}} dt$$

qui, compte tenu de la symétrie de la loi et de l'additivité des probabilités pour des intervalles disjoints, permettent de connaître la probabilité attachée à n'importe quel intervalle.



On en déduit facilement la probabilité pour qu'une variable X suivant une loi normale quelconque de

moyenne μ et d'écart-type σ , soit comprise dans un intervalle donné $[x_1, x_2]$:

$$\mathbb{P}(x_1 \leq X \leq x_2) = \mathbb{P}\left(\frac{x_1 - \mu}{\sigma} \leq U \leq \frac{x_2 - \mu}{\sigma}\right)$$

3.4 Fonctions linéaires de variables normales

Etant données deux variables X_1 et X_2 normales et indépendantes et deux constantes a_1 et a_2 , on peut montrer que la fonction linéaire $(a_1 X_1 + a_2 X_2)$ suit elle-même une loi normale. Pour spécifier cette loi, il faut connaître sa moyenne et sa variance. Les propriétés générales (c'est-à-dire indépendantes des lois de probabilité en jeu) des opérateurs espérance et variance permettent d'établir que :

$$\mathbb{E}(a_1 X_1 + a_2 X_2) = a_1 \mathbb{E}(X_1) + a_2 \mathbb{E}(X_2)$$

et que :

$$\mathbb{V}(a_1 X_1 + a_2 X_2) = a_1^2 \mathbb{V}(X_1) + a_2^2 \mathbb{V}(X_2)$$

Cette propriété s'étend évidemment à une fonction linéaire d'un nombre *quelconque* de variables indépendantes suivant des lois normales.

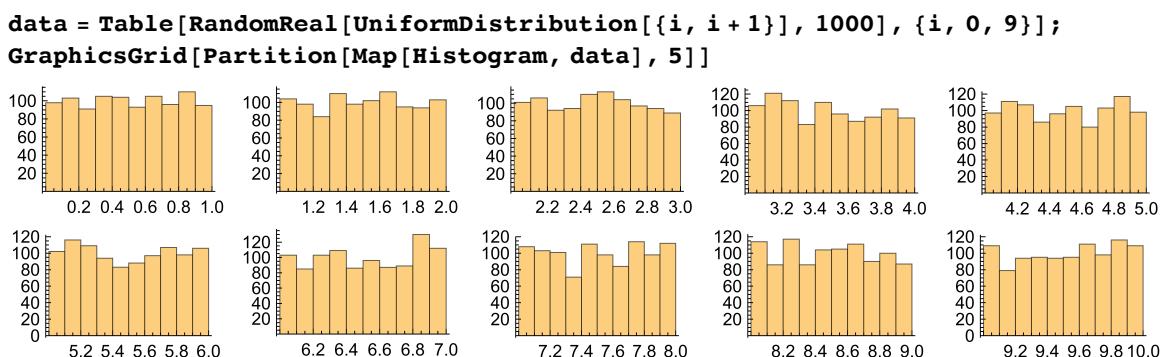
3.5 Théorème central limite

Ce théorème établit une propriété encore plus générale que la précédente et qui va justifier l'importance considérable de la loi normale, à la fois comme *modèle* pour décrire des situations pratiques, mais aussi comme *outil* théorique. Il s'énonce ainsi :

Soit $X_1, \dots, X_i, \dots, X_n$, une suite de n variables aléatoires *indépendantes*, de moyennes respectives $\mu_1, \dots, \mu_i, \dots, \mu_n$, de variances respectives $\sigma_1^2, \dots, \sigma_i^2, \dots, \sigma_n^2$ et de *lois de probabilité quelconques*, leur somme suit une loi qui, lorsque n augmente, tend vers une *loi normale* de moyenne $\mu = \sum_{i=1}^n \mu_i$ et de variance $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Les seules conditions restrictives sont que les variances soient finies, qu'aucune ne soit prépondérante devant les autres, ainsi que la condition dite de *Lindeberg* qui est presque toujours vérifiée en pratique.

On peut vérifier expérimentalement ce théorème, ici avec *Mathematica*, en sommant par exemple 10 lois uniformes distinctes dont la figure suivante montre des histogrammes en fréquences absolues obtenues sur un échantillon aléatoire de 1000 valeurs.

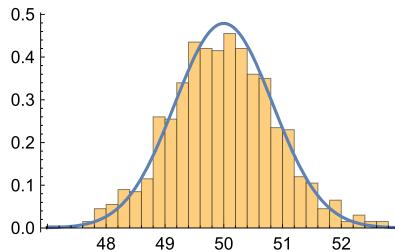


Calculons la moyenne et de la variance de la somme des lois :

```
Sum[Mean[UniformDistribution[{i, i+1}]], {i, 0, 9}]
50
Sum[Variance[UniformDistribution[{i, i+1}]], {i, 0, 9}]
5
6
```

Et comparons la distribution de la somme des lois uniformes avec celle d'une loi normale :

```
Show[{Histogram[Total[data], Automatic, "PDF"],
  Plot[PDF[NormalDistribution[50, 5/6], x], {x, 47, 53}]}]
```



3.5.1 La loi normale comme modèle

Prenons l'exemple du fonctionnement d'un tour. Le réglage du tour a pour but d'obtenir des pièces présentant une cote bien définie ; mais on sait que de multiples causes perturbatrices agissent au cours de l'usinage d'une pièce : vibrations, usures, variations de courant... Or si les causes perturbatrices sont nombreuses, si leurs effets interviennent de façon *additive*, enfin si la dispersion provoquée par chacune d'elles reste faible par rapport à la dispersion totale, alors le théorème central limite signifie que l'on doit observer une fluctuation globale très voisine de la loi normale. Et, comme ce mécanisme d'intervention de causes perturbatrices est très répandu dans la nature, il en résulte que la loi normale occupe en statistique une place privilégiée.

3.5.2 Limite de la loi binomiale

On a défini une variable de Bernoulli comme une variable qui prend la valeur 1 avec la probabilité ϖ , et la valeur 0 avec la probabilité $(1 - \varpi)$, et montré que sa moyenne est égale à ϖ et sa variance à $\varpi(1 - \varpi)$. Or, on peut considérer une variable binomiale comme la somme K de n variables de Bernoulli. Il résulte du théorème central limite que, si n est suffisamment grand (en pratique à partir de $n = 50$), la loi binomiale peut être approximée par une *loi normale* de moyenne $n\varpi$ et de variance $n\varpi(1 - \varpi)$. C'est pourquoi les tables de la loi binomiale s'arrêtent généralement à $n = 50$.

3.6 La loi log-normale

On appelle *loi log-normale* une loi qu'une transformation du type $\log(X)$ ramène à une loi normale. Cette loi est très répandue, notamment dans le domaine des sciences naturelles (géologie, biologie, ...) et des sciences humaines (psychologie, économie, ...). Chaque fois que les causes perturbatrices à l'origine des fluctuations observées répondent aux conditions du théorème central limite mais sont *multiplicatives*, on enregistre des distributions modélisables par des lois log-normales.

Exercices du chapitre 2

Exercice 1

Le profilé laminé à partir d'un lingot est découpé en billettes de 8 mètres de longueur. L'extrémité du profilé correspondant au laminage de la tête du lingot présente un défaut sur une certaine longueur X qui suit une loi normale de moyenne égale à 15 mètres et d'écart-type égal à 5 mètres.

- Pour tenter d'éliminer la longueur atteinte, on déclasse systématiquement les trois billettes de tête. Quel est le risque pour que la quatrième billette présente encore un défaut ?
- Calculer le nombre de billettes à déclasser pour que la première billette retenue soit propre avec un risque inférieur à 1 %.

Exercice 2

On usine des axes sur un tour automatique. L'intervalle de tolérance sur le diamètre de ces axes est [355, 365] en centièmes. Les diamètres des axes usinés sur la machine sont distribués suivant une loi normale de moyenne correspondant au réglage de la machine et d'écart-type 2 centièmes. Montrer que le meilleur réglage est 360. Quelle est alors la proportion de déchets ? Quelle est la proportion de déchets si la machine se dérègle et que la moyenne devient égale à 358 ?

Exercice 3

La durée de vie d'un certain type de lampes à incandescence suit une loi normale de moyenne égale à 160 heures. Les spécifications impliquent que 80 % au moins de la production tombe entre 120 et 200 heures. Quel est le plus grand écart-type acceptable pour que la fabrication réponde aux spécifications ?

Exercice 4

Il existe deux méthodes pour doser le phosphore dans l'acier d'une coulée :

- méthode C : dosage par voie chimique,
- méthode S : dosage par méthode spectrographique.

Pour toutes les coulées effectuées pendant une certaine période, on a réalisé les deux mesures et calculé la variance $\sigma^2(X)$ des mesures obtenues par la méthode C, la variance $\sigma^2(Y)$ des mesures obtenues par la méthode S, et la variance $\sigma^2(X - Y)$ des différences entre les mesures relatives à une même coulée. On a trouvé : $\sigma^2(X) = 20$, $\sigma^2(Y) = 35$ et $\sigma^2(X - Y) = 50$.

Quelle est la variance de la vraie teneur en phosphore pendant la période considérée ? Quelles sont les variances caractérisant les erreurs de mesure commises par les méthodes C et S ? Quelle est la covariance entre les deux mesures ?

On admettra que le résultat d'une mesure est la somme de la vraie teneur en phosphore et d'une erreur indépendante de cette teneur (modèle de l'erreur "absolue").

Exercice 5

Une étude sur la consommation électrique des abonnés EDF en heure de pointe montre qu'elle est bien représentée par une loi normale. L'étude montre par ailleurs que 22,1% des abonnés ont une consommation inférieure à 4 kW et 6,2% une consommation supérieure à 7 kW.

- 1) En déduire la moyenne et la variance de la loi de consommation électrique des abonnés.
- 2) Quelle puissance minimale doit fournir EDF pour satisfaire au moins 99 % de la demande dans un secteur qui comprend 100 abonnés ?

Exercice 6

Une étude a montré que, pour un haut-fourneau donné, lorsqu'on vise un poids de coulée μ_Q , on obtient un poids de fonte distribué suivant une loi normale de moyenne justement égale au poids visé μ_Q et d'écart-type σ_Q tel que $\sigma_Q = \gamma \mu_Q$ où γ est le coefficient de variation du H.F. étudié. Les moyens de transport dont on dispose entre le H.F. et l'aciérie consistent en n poches identiques. Lorsqu'on cherche à remplir au maximum une poche, le poids de fonte contenue suit une loi normale de moyenne μ_P et d'écart-type σ_P .

Calculer le poids de coulée qu'il faut viser pour qu'au risque α près, on ne doive pas reboucher le H.F. non vide (autrement dit, on doit vider le H.F. avant de le reboucher). Application numérique : $\gamma = 0.01$, $n = 2$, $\alpha = 10\%$, $\mu_P = 30 T$ et $\sigma_P = 2 T$.

Exercice 7

Un problème grave est posé en Lorraine par les effondrements d'anciennes exploitations souterraines du minerai de fer, qui menacent des villages construits à leur aplomb. Une des méthodes d'exploitation les plus courantes consistait à laisser en place, dans la couche exploitée, des piliers destinés à soutenir les terrains sus-jacents. Dans l'une de ces anciennes exploitations, les mesures ont montré que les contraintes dans les piliers suivaient une loi normale de moyenne égale à 10 bars et d'écart-type égal à 3 bars, que la résistance à la compression du minerai suivait une loi normale de moyenne égale à 50 bars et d'écart-type égal à 15 bars. On peut par ailleurs raisonnablement accepter l'indépendance entre contrainte et résistance. Quel est la probabilité de ruine d'un pilier ?

Exercice 8

Des pièces de monnaie ont un poids distribué suivant une loi normale de moyenne 10 g et d'écart-type 0.15 g. Pour les compter, on les pèse par lots d'environ un kilo. Quel est le risque de faire une erreur d'une unité ?

Exercice 9

Dans une tréfilerie, l'atelier pour l'émaillage du fil de cuivre comporte 4 machines. Chaque machine est alimentée à partir d'un stock reconstitué chaque jour. Sachant que la production journalière d'une machine suit une loi normale de moyenne 10 tonnes et d'écart-type 2 tonnes, calculer le stock pour que le risque de pénurie soit égal à 1 % seulement.

Que devient le stock total qui serait nécessaire si l'on décidait d'organiser l'atelier de telle sorte que les 4 machines puissent être alimentées à partir d'un stock commun ?

Le contrôle statistique

Le contrôle de la qualité constitue sans doute l'un des postes les plus importants de l'entreprise moderne. Les méthodes de fabrication à la chaîne exigent une parfaite interchangeabilité entre les innombrables pièces fabriquées en série. Un tel résultat

ne peut être atteint que si les spécifications imposées sont rigoureusement respectées. A ce premier objectif de qualité s'en ajoute un second qui semble s'opposer au précédent, celui de l'économie. La qualité voudrait qu'on se livre à un examen minutieux de la fabrication ; l'économie exige qu'on réduise au maximum tous les frais, en particulier ceux du contrôle qui peuvent constituer une part très importante du prix de revient. Cela fait apparaître la nécessité de substituer à une

inspection à 100% des pièces fabriquées, un contrôle par échantillonnage, ou contrôle statistique, qui devient d'ailleurs inévitable lorsqu'on doit procéder à des essais destructifs (sur la résistance des matériaux, par exemple). Il importe, dès lors, de rechercher des modes de contrôle qui permettent à la fois de prélever un nombre de pièces aussi faible que possible, et de déterminer aussi bien que possible la qualité d'un lot.

1 Distribution d'échantillonnage

1.1 Population

En statistique, on appelle *population* un ensemble d'éléments caractérisés par un critère qui permet de les identifier sans ambiguïté. Chacun des éléments est appelé *individu*. Ces appellations sont liées aux origines démographiques de la statistique. On parlera, par exemple, de la population des pièces usinées sur une machine pendant telle période en s'intéressant, non pas aux individus en tant que tels, mais à une ou plusieurs de leurs caractéristiques.

Chacune des caractéristiques sur laquelle on décide de faire porter l'observation est appelée *variable* (ou caractère). Il importe de faire ici une distinction entre deux types de variables et par conséquent, deux catégories de populations :

- populations à caractéristiques *qualitatives*. Pour des pièces cylindriques par exemple, on distinguera entre pièces satisfaisantes et pièces défectueuses. On se trouve dans ce cas chaque fois qu'un contrôle s'effectue par calibre, c'est-à-dire que l'on distingue, par exemple, les pièces dont le diamètre appartient ou n'appartient pas à un certain intervalle admissible appelé intervalle de tolérance.
- populations à caractéristiques *quantitatives* (ou mesurables). Dans l'exemple des pièces, on peut mesurer le diamètre et classer les pièces suivant les valeurs qu'il peut prendre.

La notion de population n'est pas toujours très facile à définir précisément. Si on considère, par exemple, la production journalière d'une machine, on peut, a priori, parler de la population des pièces produites. Mais, au cours de la journée, la machine a pu se dérégler, un technicien a peut-être procédé à un réglage, etc. Il n'est donc pas évident que l'ensemble de la production journalière constitue une population unique et bien homogène.

Un autre type de difficulté peut se présenter pour définir la population. Supposons par exemple, que la variable étudiée soit la résistance du béton dans un barrage. L'individu, c'est-à-dire l'élément sur lequel on effectue la mesure, est une éprouvette découpée et usinée suivant un standard. Pour définir la population, il faut alors se référer à l'ensemble infini de toutes les éprouvettes susceptibles d'être réalisées dans les mêmes conditions à partir des coulées étudiées.

1.2 Echantillon

Une partie essentielle de la statistique consiste à porter des jugements sur une population à partir d'échantillons ; c'est ce qu'on appelle *l'inférence statistique*. Un *échantillon* est un ensemble d'individus prélevés, suivant un procédé bien défini, dans l'ensemble plus important constitué par la population. Le nombre d'individus prélevés, souvent noté n , s'appelle la *taille* de l'échantillon.

Toutes les statistiques établies sur les échantillons impliqueront que ces derniers sont *représentatifs* de la population dont ils proviennent. C'est le cas s'ils ont été prélevés *au hasard*, tous les individus de la population ayant la même probabilité de faire partie de l'échantillon prélevé. En pratique, l'obtention d'échantillons au hasard présente certaines difficultés qui peuvent être levées si l'on peut numérotter chaque individu de la population et qu'on utilise une table de nombres au hasard.

Dans toute la suite, nous admettrons que les échantillons sont prélevés de façon non exhaustive ou bien que la taille de la population est suffisamment importante devant celle de l'échantillon pour que l'on puisse se ramener à ce cas.

1.3 Caractéristiques des échantillons

Nous désignerons par $x_1, x_2, \dots, x_i, \dots, x_n$, les valeurs prises par une variable X pour chacun des individus constituant l'échantillon ; ce que l'on appelle une série d'observations. De telles séries peuvent être caractérisées par un certain nombre de valeurs typiques que nous allons définir.

1.3.1 Caractéristiques de tendance centrale

On appelle caractéristique de tendance centrale, une fonction des observations dont la valeur est comprise entre les valeurs extrêmes de la série et qui donne une mesure du milieu ou du centre de l'ensemble des observations.

La plus couramment utilisée est la *moyenne arithmétique* :

$$m = \frac{1}{n} \sum_{i=1}^n x_i, \text{ souvent notée aussi } \bar{x}$$

qui est très facile à calculer et possède d'importantes propriétés théoriques, par ailleurs assez faciles à établir. Toutefois, la moyenne possède l'inconvénient d'être très sensible au retrait ou à l'ajout d'une observation « aberrante ». On dit alors que c'est une statistique peu *robuste*.

Une caractéristique de tendance centrale plus robuste est la *médiane* dont les propriétés théoriques sont par contre plus compliquées à manipuler que pour la moyenne. Lorsqu'on a classé les observations dans l'ordre des grandeurs croissantes, la médiane est la valeur de l'observation qui se trouve au rang $\frac{n+1}{2}$, si n est impair. Si n est pair ($n = 2p$), c'est le milieu de l'intervalle $[x_p, x_{p+1}]$.

Plus généralement on définit les *quantiles*. Le premier *décile*, par exemple, est la valeur telle qu'il y ait 10 % des observations plus petites qu'elle, tandis que le neuvième *décile* est la valeur telle qu'il y ait 10 % des observations plus grandes qu'elle. De même, les premier et troisième *quartiles* ont respectivement 25 % et 75 % des observations plus petites qu'eux. La médiane est donc aussi le deuxième quartile. Plus précisément, on définira le p -quantile ($0 < p < 1$) d'une variable aléatoire X comme étant la valeur x_p telles que $\mathbb{P}(X \leq x_p) \geq p$.

1.3.2 Caractéristiques de dispersion

On appelle caractéristique de dispersion, une fonction des observations dont la valeur rend compte de l'*étalement* des valeurs observées autour de leur tendance centrale.

On appelle ainsi *étendue* w d'une série d'observations, l'écart entre la plus grande et la plus petite valeur de la série :

$$w = x_{\max} - x_{\min}$$

dont le principal avantage est la simplicité de calcul, mais qui est, par contre, très peu robuste.

On appelle *variance* s^2 d'une série d'observations, la quantité qui est définie par la relation :

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$$

et qui mesure l'écart quadratique moyen entre les observations et leur moyenne. Elle joue un rôle très important dans toute la statistique mathématique. Pour la calculer, on notera l'incontournable relation :

$$s^2 = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2 m \sum_{i=1}^n x_i + n m^2 \right) = \frac{1}{n} \left(\sum_{i=1}^n x_i^2 - 2 m n m + n m^2 \right) = \frac{1}{n} \sum_{i=1}^n x_i^2 - m^2$$

que l'on mémorisera facilement en retenant que *la variance est égale à la moyenne des carrés moins le carré de la moyenne*.

La racine carrée s de la variance est appelée l'*écart-type* :

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - m)^2}$$

Variance et écart-type sont assez peu robustes. Plus favorable de ce point de vue, on peut calculer l'écart absolu moyen e des observations à leur moyenne :

$$e = \frac{1}{n} \sum_{i=1}^n |x_i - m|$$

Enfin, une caractéristique de dispersion extrêmement robuste est la distance *interquartile*, écart entre le troisième quartile Q_3 et le premier quartile Q_1 :

$$I_Q = Q_3 - Q_1$$

On notera dans ce qui précède que nous avons utilisé des lettres minuscules latines (m et s notamment) pour caractériser les échantillons. Les lettres grecques (μ et σ par exemple) seront réservées aux populations.

1.4 Distributions d'échantillonnage

On appelle *population de référence ou modèle*, une population définie par une loi de probabilité $P(x)$ pouvant être considérée comme à l'origine des résultats observés, c'est-à-dire telle que la probabilité pour que la variable étudiée prenne une valeur dans un certain intervalle $[x, x + h]$, pour un individu prélevé au hasard, soit :

$$\mathbb{P}(x < X \leq x + h) = P(x + h) - P(x)$$

Cela étant, considérons une population de référence définie par la loi de probabilité d'une certaine variable aléatoire X . Si nous prélevons *au hasard* et de façon non exhaustive, un échantillon de taille n , nous observons n valeurs :

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

dont on peut calculer la moyenne m et la variance s^2 .

Mais un autre échantillon de taille n , prélevé au hasard dans la même population, conduirait à d'autres valeurs :

$$x'_1, x'_2, \dots, x'_i, \dots, x'_n$$

puis m' et s'^2 , a priori différentes à cause des fluctuations de l'échantillonnage. On pourrait ainsi répéter ces mesures et ces calculs sur un grand nombre d'échantillons différents.

Les nombres x_1, x'_1, \dots peuvent alors être considérés comme des réalisations d'une certaine variable aléatoire X_1 , les nombres x_2, x'_2, \dots comme des réalisations d'une variable aléatoire X_2 , et plus généralement, les nombres x_i, x'_i, \dots comme des réalisations d'une variable aléatoire X_i .

Les variables aléatoires $X_1, X_2, \dots, X_i, \dots$ sont *indépendantes* (si l'échantillonnage est non exhaustif) et ont même loi de probabilité : celle de la population de référence.

Les valeurs m, m', \dots peuvent alors être considérées comme des réalisations d'une variable aléatoire que nous noterons M , fonction des variables aléatoires $X_1, X_2, \dots, X_i, \dots, X_n$:

$$M = \frac{1}{n} \sum_{i=1}^n X_i, \text{ souvent aussi notée } \bar{X}.$$

La distribution de M est appelée *distribution d'échantillonnage*. Il en est de même de celle de la variable aléatoire :

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2$$

La détermination des *lois d'échantillonnage* est fondamentale en statistique. Nous avons déjà étudié une de ces lois : la loi binomiale qu'on rencontre chaque fois que l'on étudie un lot de pièces contenant une proportion ϖ de pièces défectueuses et qu'on y prélève, de façon non exhaustive, un échantillon de n pièces. Le nombre K de pièces défectueuses trouvées dans l'échantillon est une variable aléatoire qui peut prendre les valeurs $0, \dots, k, \dots, n$, avec les probabilités :

$$\mathbb{P}(K = k) = \binom{n}{k} \varpi^k (1 - \varpi)^{n-k}$$

La suite fournira d'autres exemples de lois d'échantillonnage. Dans ce chapitre, nous nous limiterons à un autre cas particulier concernant cette fois une population à caractéristique quantitative.

1.5 Loi de la moyenne d'un échantillon prélevé dans une population normale

Dans la pratique, on rencontre très fréquemment des populations à caractéristiques quantitatives qui peuvent être *raccordées* à des populations de référence définies par une loi de probabilité normale :

$$\mathbb{P}(X \leq x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}(\frac{u-\mu}{\sigma})^2} du$$

On dit alors que la population est normale de moyenne μ et d'écart-type σ .

Considérons la variable aléatoire M , moyenne d'un échantillon prélevé au hasard dans une telle population :

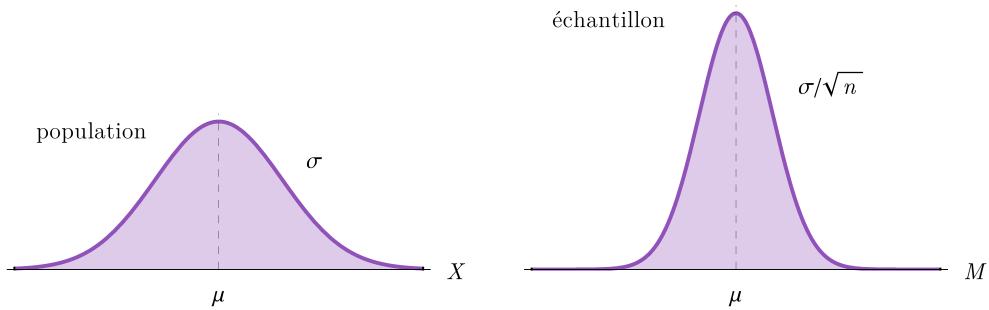
$$M = \frac{X_1 + X_2 + \dots + X_n}{n}$$

On a établi qu'une somme de variables indépendantes suivant des lois normales, suivait elle-même une loi normale de moyenne égale à la somme des moyennes des variables, et de variance égale à la somme de leurs variances. Il en résulte que M suit une loi normale de moyenne :

$$\mathbb{E}(M) = \frac{\mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n)}{n} = \mu$$

et de variance :

$$\mathbb{V}(M) = \frac{\mathbb{V}(X_1) + \mathbb{V}(X_2) + \dots + \mathbb{V}(X_n)}{n^2} = \frac{\sigma^2}{n}$$



On peut aussi écrire ce résultat très important en introduisant la loi normale centrée réduite :

$$\mathbb{P}\left(\left|M - \mu\right| > u \frac{\sigma}{\sqrt{n}}\right) = \frac{2}{\sqrt{2\pi}} \int_u^{+\infty} e^{-\frac{t^2}{2}} dt$$

2 Contrôle Statistique

2.1 Contrôle de fabrication et contrôle de réception

Nous appellerons qualité ϖ d'un lot, la proportion de pièces défectueuses qu'il contient. La qualité d'une pièce peut, quant à elle, être déterminée, suivant les cas, de deux manières :

- par un *contrôle qualitatif* où la pièce est passée au calibre et où l'on vérifie si elle correspond ou non aux tolérances imposées par le cahier des charges ;
- par un *contrôle quantitatif* où la pièce est mesurée et où l'on vérifie si la valeur obtenue est comprise dans l'intervalle de tolérance.

On distingue encore deux types de contrôle :

- le *contrôle de fabrication*, effectué à chaque stade de l'élaboration d'un produit pour limiter la proportion de rebuts ;
- le *contrôle de réception*, effectué par le client qui achète une certaine quantité de ce produit et veut vérifier que sa qualité correspond bien à ce qu'il attend.

Ces méthodes de contrôle reposent sur les mêmes bases. Nous nous limiterons donc à l'étude du contrôle de réception dans le cas qualitatif et du contrôle de fabrication dans le cas quantitatif.

2.2 Contrôle de réception

Dans ce cas, l'objectif du contrôle est de porter un jugement sur la proportion ϖ inconnue de pièces défectueuses contenues dans un lot. En général, deux parties se trouvent en présence, un fournisseur et un client, dont les exigences peuvent se traduire de la façon suivante :

- le *fournisseur* ne voudrait pas qu'on lui refuse un lot contenant ϖ_1 ou moins de déchets ;
- le *client* ne voudrait pas accepter un lot s'il contient ϖ_0 ou plus de déchets.

Il faut remarquer dès ce stade que ϖ_0 et ϖ_1 sont définis à partir de considérations techniques et non statistiques. Le fournisseur peut déterminer ϖ_1 en fonction de son matériel, de son prix de vente, etc. Le client détermine ϖ_0 en fonction des conséquences économiques ou techniques qu'entraîne pour lui la présence de pièces défectueuses dans les lots réceptionnés.

Définir une *règle de contrôle* va donc consister à fixer la taille n de l'échantillon à prélever et une valeur limite c telles que :

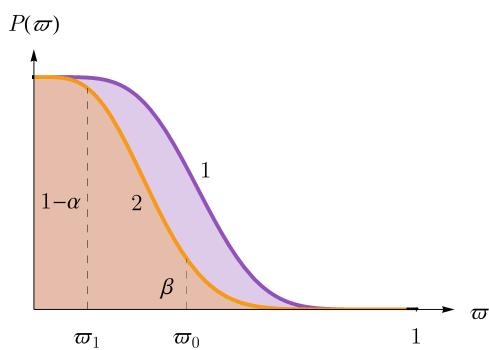
- si on trouve moins de c pièces défectueuses parmi les n pièces prélevées, on accepte le lot ;
- si on en trouve plus de c , on le refuse.

Examinons les conséquences d'une telle règle. Connaissant n et c , il est possible, par référence à la loi binomiale, de calculer la probabilité $P(\varpi)$ d'accepter un lot où la proportion de pièces défectueuses serait égale à une valeur ϖ donnée quelconque :

$$P(\varpi) = \sum_{k=0}^c \binom{n}{k} \varpi^k (1 - \varpi)^{n-k}$$

$P(\varpi)$ varie avec ϖ . La courbe représentative de la fonction $P(\varpi)$ est appelée la *courbe d'efficacité* de la règle de contrôle.

Elle a la forme indiquée sur la figure suivante. On constate que la règle relative à la courbe 2 est plus sévère que celle qui correspond à la courbe 1, la probabilité d'accepter un lot de qualité donnée ϖ étant plus faible : on dit qu'elle est plus *efficace*.



On constate d'autre-part, qu'à une règle donnée correspond :

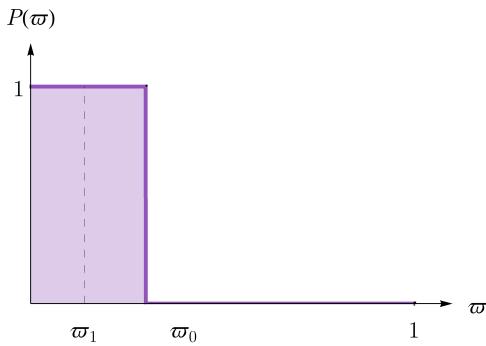
- une probabilité α de refuser un lot présentant une proportion ϖ_1 de déchets ; c'est le *risque du fournisseur* ;
- une probabilité β d'accepter un lot présentant une proportion ϖ_0 de déchets ; c'est le *risque du client*.

Il en résulte qu'il est possible, inversement, de déterminer la règle du contrôle, c'est-à-dire n et c , telle que la courbe d'efficacité passe par les deux points $\{\varpi_0, \beta\}$ et $\{\varpi_1, 1 - \alpha\}$, à partir des deux relations :

$$\begin{aligned}\beta &= \sum_{k=0}^c \binom{n}{k} \varpi_0^k (1 - \varpi_0)^{n-k} \\ 1 - \alpha &= \sum_{k=0}^c \binom{n}{k} \varpi_1^k (1 - \varpi_1)^{n-k}\end{aligned}$$

Il est clair qu'aussi petits soient-ils choisis, les risques α et β devront être consentis par le fournisseur et le client.

A la limite, la courbe d'efficacité qui conduirait à accepter avec certitude tout lot contenant une proportion de déchets inférieure à ϖ_1 et à refuser avec certitude tout lot contenant une proportion de déchets supérieure à ϖ_0 , devrait avoir la forme indiquée ci-après. Seule une inspection à 100 % permettrait de l'obtenir.



2.3 Contrôle en cours de fabrication

2.3.1 Intervalle de tolérance et déchets

Considérons les pièces usinées sur un tour automatique. Si l'on s'intéresse au diamètre de ces pièces, on constate que, pour un certain réglage de la machine, c'est une variable aléatoire que l'on peut mettre sous la forme :

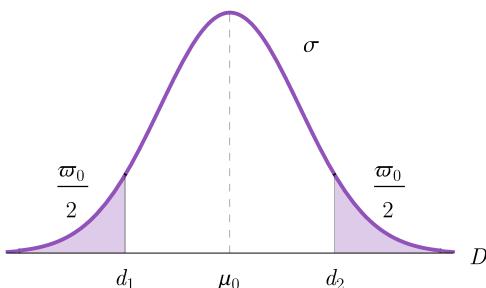
$$D = \mu + \sigma U$$

où μ est un nombre certain, U une variable normale centrée et réduite, et σ correspond à la variabilité de la machine et en constitue une caractéristique intrinsèque pour un certain niveau d'usure. La machine sera réglée au mieux si μ est égal à la cote théorique μ_0 prévue pour la pièce.

Mais, même si le réglage représente ce qu'on peut obtenir de mieux sur la machine dans l'état où elle se trouve, il n'empêchera pas pourtant la production d'une proportion ϖ_0 de pièces défectueuses, c'est-à-dire dont le diamètre sort des limites de tolérance $[d_1, d_2]$:

$$\varpi_0 = 2 \mathbb{P}(U > \frac{d_2 - \mu_0}{\sigma})$$

On constate que ϖ_0 est d'autant plus grand que σ l'est devant l'intervalle de tolérance.



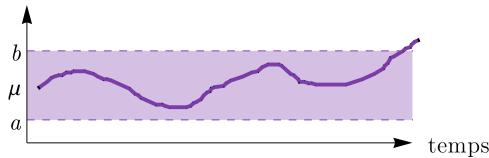
2.3.2 Règle de contrôle

Le principe du contrôle en cours de fabrication est de prélever à intervalles réguliers n pièces consécutives, de calculer leur moyenne et de vérifier qu'elle tombe dans un certain intervalle :

$$[a, b] = \left[\mu_0 - u_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \mu_0 + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

où $u_{\alpha/2}$ désigne la valeur d'une variable normale centrée réduite U telle que :

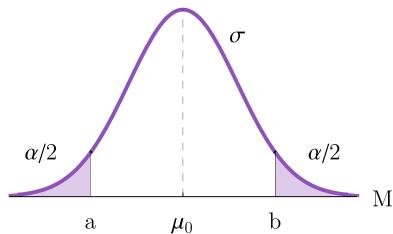
$$\alpha = 2 \mathbb{P}(U > u_{\alpha/2})$$



2.3.3 Efficacité d'un contrôle

Si la moyenne calculée tombe dans l'intervalle $[a, b]$, on laisse la fabrication se poursuivre ; sinon, on règle la machine. α représente donc le *risque de procéder à un réglage de la machine alors qu'elle n'est pas déréglée* :

$$\alpha = \mathbb{P}(M \notin [a, b] | \mu = \mu_0)$$



Supposons maintenant que la machine se dérègle et que la moyenne de la fabrication devienne $\mu \neq \mu_0$. On peut calculer la probabilité :

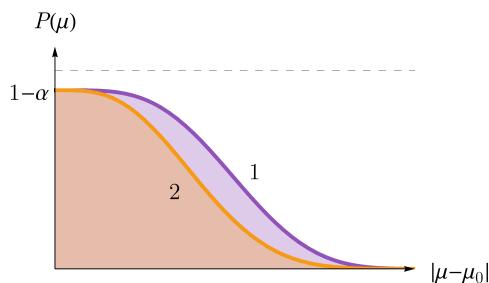
$$P(\mu) = \mathbb{P}(M \in [a, b] | \mu)$$

qui désigne le *risque de laisser la fabrication se poursuivre alors que la machine est déréglée*.

On a évidemment :

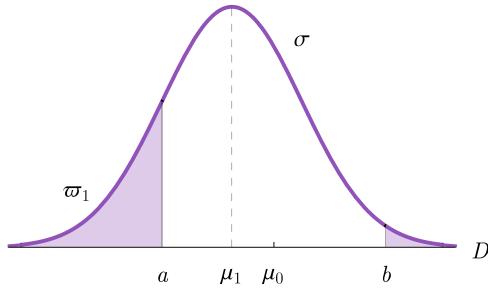
$$P(\mu_0) = 1 - \alpha$$

La courbe représentative de la fonction $P(\mu)$ est appelée courbe d'efficacité du contrôle. Pour un risque α donné, l'efficacité croît avec n : on a d'autant plus de chances de détecter un déréglage donné $|\mu - \mu_0|$ que les échantillons prélevés sont plus importants.



2.3.4 Risque β

Considérons maintenant la valeur μ_1 de μ qui correspond à une proportion ϖ_1 de déchets, que l'on considère comme inadmissible parce que trop importante, et qui est déterminée à partir de considérations techniques et (ou) économiques.



Remarquons que la valeur μ_1' , symétrique de μ_1 par rapport à μ_0 (milieu de l'intervalle $[a, b]$), correspond au même déréglage $|\mu_1 - \mu_0|$ et conduit à la même proportion ϖ_1 de déchets.

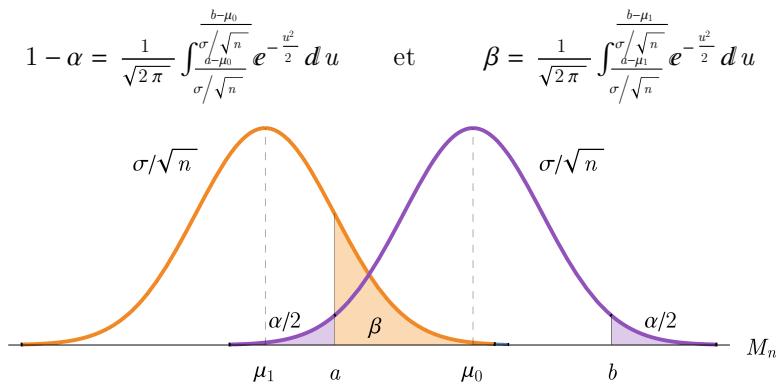
Ayant ainsi déterminé μ_1 , on note généralement β la probabilité $P(\mu_1)$ qui est le *risque de ne pas régler la machine alors que son déréglage est inadmissible* :

$$\beta = \mathbb{P}(M \in [a, b] | \mu = \mu_1)$$

2.3.5 Choix de la taille de l'échantillon à prélever

Il peut dès lors s'avérer intéressant, au lieu de fixer au départ n et α , de se fixer α et β et d'en déduire la règle de contrôle, c'est-à-dire n et $[a, b]$ de telle sorte que la courbe d'efficacité passe par les deux points $(\mu_0, 1 - \alpha)$ et (μ_1, β) .

La figure suivante illustre les deux relations qui permettent de mener facilement les calculs :



Le calcul pratique peut être notablement simplifié en tenant compte du fait que la probabilité :

$$\frac{1}{\sqrt{2\pi}} \int_{\frac{b-\mu_1}{\sigma/\sqrt{n}}}^{+\infty} e^{-\frac{u^2}{2}} du$$

est presque toujours négligeable et que l'on peut écrire, par conséquent :

$$\beta \approx \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\mu_1}{\sigma/\sqrt{n}}}^{+\infty} e^{-\frac{u^2}{2}} du$$

2.4 Contrôle progressif

Les règles de contrôle que nous venons d'envisager consistent, dans le cas qualitatif par exemple, à :

- prélever dans le lot un échantillon de n pièces que l'on examine une à une,
- accepter le lot si on trouve c ou moins de c pièces défectueuses, sinon le refuser.

Il peut être judicieux d'envisager d'autres règles de contrôle, par exemple telles que l'on accepte le lot si le résultat d'un premier échantillon est bon, qu'on le refuse si ce résultat est mauvais, et qu'on examine un deuxième échantillon si ce résultat est moyen, de manière à confirmer d'une façon ou d'une autre la première opinion qu'on avait pu former. Une telle règle peut s'énoncer ainsi :

- prélever un échantillon de n_1 pièces ;
- si on trouve c_1 ou moins de c_1 pièces défectueuses, accepter le lot ; si on en trouve c_2 ou plus de c_2 , refuser le lot ; si on en trouve entre c_1 et c_2 , prélever un nouvel échantillon de n_2 pièces ;
- si le nombre total de pièces défectueuses trouvées dans les deux échantillons est inférieur ou égal à c_3 , accepter le lot ; le refuser dans le cas contraire.

Il s'agit là de ce qu'on appelle un *plan d'échantillonnage double*. On peut utiliser aussi des plans d'échantillonnage multiples où l'on prélève plus de deux échantillons. L'avantage des plans multiples réside dans le fait que, pour une efficacité donnée, ils conduisent à un nombre moyen de pièces prélevées inférieur au nombre de pièces prélevées dans le plan simple équivalent.

Les plans multiples conduisent, à la limite, aux *plans progressifs*. Dans ce cas, l'effectif de l'échantillon n'est pas précisé à l'avance. On prélève les pièces une à une et, à chaque stade, on calcule le nombre total k_n de pièces défectueuses. Suivant la valeur de k_n , on accepte le lot, on le refuse ou on prélève une $(n+1)$ ème pièce.

Exercices du chapitre 3

Exercice 1

Pour contrôler un lot important d'articles, on adopte la règle suivante :

- on prélève un article au hasard ; s'il est mauvais on refuse le lot ; s'il est bon, on prélève un 2e article.
- si le 2e article est mauvais, on refuse le lot ; s'il est bon, on prélève un 3e article.
- si le 3e article est mauvais, on refuse le lot ; s'il est bon, on accepte le lot.

a) Calculer la probabilité de refuser le lot et l'espérance du nombre d'articles prélevés, en fonction de la proportion ϖ d'articles défectueux dans le lot.

b) Comparer ces résultats à ceux que l'on aurait obtenus en prélevant directement trois articles, et en refusant le lot si l'un d'eux au moins est mauvais.

Exercice 2

Une entreprise réceptionne périodiquement des lots de pièces d'un certain type destinées à entrer dans des assemblages. On suppose que les conditions d'utilisation de ces pièces rendent souhaitable le rejet d'un lot contenant une proportion de déchets supérieure à 8 %.

a) On décide d'adopter la règle suivante : on prélève 100 pièces que l'on examine une à une, et on compte le nombre k de pièces défectueuses ; si $k \leq 3$, on accepte le lot ; si $k > 3$ on le refuse. En admettant l'approximation de Poisson, quelle est la limite supérieure du risque d'accepter un lot contenant plus de 8 % de déchets ?

b) Le fournisseur souhaiterait ne pas se voir refuser des lots jugés convenables lorsqu'ils contiennent moins de 3% de déchets. Quel risque maximum encourre-t-il avec la règle précédente ?

c) Que deviennent les risques précédents si l'on prélève 200 pièces et que l'on fixe comme limite d'acceptation $k \leq 9$?

d) On adopte un plan d'échantillonnage multiple défini de la façon suivante :

- on prélève 100 pièces. Si $k_1 \leq 3$, on accepte le lot; si $k_1 > 9$, on le refuse ;
- si $3 < k_1 \leq 9$, on prélève à nouveau 100 pièces. Si $k_1 + k_2 \leq 9$, on accepte le lot ; sinon on le refuse.

Comparer son efficacité par rapport au précédent. Quelle est l'espérance mathématique du nombre de pièces prélevées ?

Exercice 3

Un fabricant produit des axes cylindriques dont le diamètre est distribué suivant une loi normale d'écart-type égal à 8. Les limites de tolérance pour le diamètre sont [90, 130]. Pour réceptionner un lot important de ces pièces, on décide d'adopter soit un procédé quantitatif A, soit un procédé qualitatif B. Comparer les deux procédés.

Procédé A : On prélève n pièces, on mesure leurs diamètres et on calcule la moyenne m des n diamètres, puis on adopte la règle suivante :

- si $a \leq m \leq b$, on accepte le lot ;
- si $m < a$ ou $m > b$, on le refuse.

On détermine n , a et b en s'imposant les conditions :

- si la proportion de pièces défectueuses dans le lot (diamètre extérieur à l'intervalle [90, 130]) est supérieure à 5 %, on veut avoir 96 % de chance au moins de refuser le lot ;
- si cette proportion est inférieure à 2 %, on veut avoir 15 % de chance au plus de refuser le lot.

Procédé B : On prélève 300 pièces, on les passe au calibre et on compte le nombre k de pièces défectueuses, puis on adopte la règle suivante :

- si $k \leq 8$, on accepte le lot ;
- si $k > 8$, on le refuse.

Exercice 4

Dans un atelier de peinture, l'air est analysé toutes les heures. Le seuil maximal admissible pour un certain polluant est fixé à 7,7 ppm. Si la teneur du polluant est distribuée suivant une loi normale de moyenne 7,6 et d'écart-type 0,04 et si l'erreur de mesure suit une loi normale de moyenne 0 et d'écart-type 0,03 calculer les probabilités pour que :

- a) Une mesure dépasse 7,7.
- b) La moyenne de 2 mesures dépasse 7,7.

Exercice 5

Pour réceptionner un lot d'acier, on se propose d'examiner sa résistance à la traction. Pour cela, on prélève dans le lot des éprouvettes cylindriques qui sont soumises à un effort de traction jusqu'à rupture. En admettant que l'ensemble des éprouvettes que l'on pourrait extraire donnerait lieu, en ce qui concerne leur résistance, à une distribution normale d'écart-type égal à 20 kg/mm^2 , on se propose d'élaborer une règle de contrôle satisfaisant aux conditions suivantes. On prélèvera un nombre fixé n d'éprouvettes et on déterminera la moyenne m de leurs résistances. Le lot sera accepté (ou refusé) selon que m sera supérieure (ou inférieure) à une certaine valeur x .

Déterminer n et x de telle façon que :

- si la résistance moyenne du lot est égale (ou inférieure) à 80 kg/mm^2 , la probabilité β qu'il soit accepté est égale (ou inférieure) à 0.001 ;
- si la résistance moyenne du lot est égale (ou supérieure) à 100 kg/mm^2 , la probabilité α qu'il soit refusé est égale (ou inférieure) à 0.01.

Quelle est, en fonction de la résistance moyenne du lot, la probabilité pour que le lot soit accepté après un tel jugement sur échantillon ? Quelle signification concrète peut-on donner à α et β du point de vue de l'acheteur et de celui du fournisseur ?

Exercice 6

Un lot d'un certain type de fusées a été stocké pendant deux ans. Les spécifications de ce lot indiquent une portée en moyenne égale à 2000 mètres avec un écart-type de 100 mètres. On veut contrôler si le lot est encore en état. On fixe pour cela que :

- la probabilité de réformer le lot si la portée moyenne a diminué de 100 mètres doit être égale à 90% ;
- la probabilité de le réformer si elle n'a pas changé doit être égale à 5% seulement.

- a) Combien de fusées faut-il contrôler ?
- b) Si la moyenne de l'échantillon contrôlé est égale à 1930 mètres, doit-on réformer le lot ?

L'estimation statistique

Le problème traité dans ce chapitre est le suivant : on se trouve en présence d'un échantillon et l'on cherche à déterminer explicitement la loi de probabilité définissant la population de référence dont ces observations peuvent être considérées comme issues. Nous admettrons spécifiée la forme analytique de la loi de probabilité que suivent les observations. Dans ces conditions, on se trouve conduit à estimer les paramètres $\theta_1, \theta_2, \dots$ de la loi de probabilité $p(x; \theta_1, \theta_2, \dots)$ à partir de l'échantillon observé x_1, x_2, \dots, x_n , c'est à dire à tirer de cet échantillon une information concernant la valeur des paramètres inconnus. Il s'agit de plus de pouvoir émettre un jugement sur la qualité de cette information

1 Estimateur et intervalle de confiance

1.1 La loi des grands nombres

Considérons une suite de variables aléatoires $X_1, \dots, X_i, \dots, X_n$ indépendantes, et ayant toutes la même loi de probabilité qu'une variable aléatoire X . La loi de probabilité de X peut être *quelconque* de moyenne μ et de variance σ^2 .

Soit $M = \frac{1}{n}(X_1 + \dots + X_i + \dots + X_n)$ la moyenne arithmétique des variables $X_1, \dots, X_i, \dots, X_n$.

Nous avons calculé au chapitre précédent sa moyenne $\mathbb{E}(M) = \mu$ et sa variance $\mathbb{V}(M) = \frac{\sigma^2}{n}$.

Soit ε un nombre choisi arbitrairement aussi petit que l'on veut. En utilisant l'inégalité de Bienaymé-Tchebychev (voir chapitre 2), on peut écrire que :

$$\mathbb{P}(|M - \mu| > \varepsilon) < \frac{\sigma^2}{n\varepsilon^2}$$

et, en posant $\frac{\sigma^2}{n\varepsilon^2} = \eta$, on voit qu'étant donnés ε et η aussi petits qu'on le veut, il est possible de trouver un nombre $N = \frac{\sigma^2}{\eta\varepsilon^2}$ tel que $n \geq N$ entraîne :

$$\mathbb{P}(|M - \mu| > \varepsilon) < \eta$$

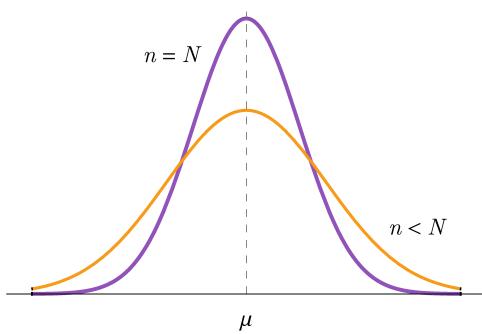
C'est la loi des grands nombres que l'on peut encore énoncer ainsi : quand n augmente, la moyenne M converge en probabilité vers l'espérance mathématique de X . Il est important de bien prendre conscience de la différence entre la notion de convergence classique et celle, toute nouvelle, de convergence en probabilité.

Dire que M tend au sens classique vers μ , ce serait dire qu'on peut déterminer n tel qu'étant donné ε aussi petit qu'on veut : $|M - \mu| \leq \varepsilon$.

Dire qu'il y a convergence en probabilité, c'est dire que l'événement $\{|M - \mu| \leq \varepsilon\}$ n'est pas certain, mais que sa probabilité peut être rendue aussi voisine de 1 qu'on le veut, à condition que n soit suffisamment grand :

$$\mathbb{P}(|M - \mu| \leq \varepsilon) \geq 1 - \eta$$

On peut l'illustrer comme suit : quand n augmente, la distribution de M se resserre autour de μ .



1.2 Estimation et estimateur

Pour simplifier la suite de l'exposé, nous supposerons que la loi de référence à déterminer dépend d'un seul paramètre θ .

Le problème se réduit donc à déterminer une fonction des observations $\theta^*(x_1, x_2, \dots, x_n)$, aussi voisine que possible de la valeur vraie θ qui est inconnue. On dit alors que la statistique θ^* est une *estimation* de θ .

On peut utiliser, pour résoudre ce problème, la notion d'estimateur. Etant donné une variable aléatoire $T(X_1, X_2, \dots, X_n)$ fonction des variables aléatoires X_1, X_2, \dots, X_n , on dit qu'elle constitue un *estimateur* de θ si :

- son espérance mathématique tend vers θ quand n augmente indéfiniment : $\mathbb{E}(T) \xrightarrow[n \rightarrow \infty]{} \theta$
- sa variance tend vers 0 quand n augmente indéfiniment : $\mathbb{E}[T - \mathbb{E}(T)]^2 \xrightarrow[n \rightarrow \infty]{} 0$

Dans le cas particulier où $\mathbb{E}(T) = \theta$ quel que soit n , l'estimateur T est dit *sans biais*.

1.3 Estimateur et convergence en probabilité

Pour éclairer la compréhension de la définition d'un estimateur, on peut la rapprocher de celle de la convergence en probabilité.

Un estimateur T est dit convergent, si T converge en probabilité vers θ , quand n augmente indéfiniment, c'est-à-dire si, étant donnés deux nombres ε et η aussi petits qu'on le veut, il est possible de déterminer un nombre N tel que $n > N$ entraîne :

$$\mathbb{P}(|T - \theta| > \varepsilon) < \eta$$

On a alors l'important résultat suivant : *un estimateur T dont l'espérance mathématique tend vers θ et dont la variance tend vers 0, quand n augmente indéfiniment, est convergent*. La démonstration qui suit peut être omise.

D'après l'inégalité de Bienaymé-Tchebychev, on peut écrire :

$$\mathbb{P}\left(|T - \mathbb{E}(T)| > \frac{\varepsilon}{2}\right) < 4 \frac{\mathbb{V}(T)}{\varepsilon^2}$$

$$\text{ou } \mathbb{P}\left(|T - \mathbb{E}(T)| \leq \frac{\varepsilon}{2}\right) \geq 1 - 4 \frac{\mathbb{V}(T)}{\varepsilon^2}$$

Mais $\mathbb{E}(T)$ converge vers θ au sens ordinaire de l'analyse, on peut donc trouver un nombre N_1 tel que, pour $n > N_1$:

$$|\mathbb{E}(T) - \theta| \leq \frac{\varepsilon}{2}$$

On peut alors écrire, pour un tel n :

$$\mathbb{P}(|T - \mathbb{E}(T)| + |\mathbb{E}(T) - \theta| \leq \varepsilon) \geq 1 - 4 \frac{\mathbb{V}(T)}{\varepsilon^2}$$

Notant maintenant que :

$$|T - \mathbb{E}(T)| + |\mathbb{E}(T) - \theta| \geq |T - \mathbb{E}(T) + \mathbb{E}(T) - \theta| = |T - \theta|$$

et d'une manière générale que :

$$Y \geq X \implies \mathbb{P}(Y \leq \varepsilon) \leq \mathbb{P}(X \leq \varepsilon)$$

on en déduit que :

$$\mathbb{P}(|T - \theta| \leq \varepsilon) \geq 1 - 4 \frac{\mathbb{V}(T)}{\varepsilon^2}$$

Or, puisque $\mathbb{V}(T)$ tend vers 0 quand n augmente indéfiniment, on peut également trouver un nombre N_2 tel que :

$$\mathbb{V}(T) \leq \frac{\eta \varepsilon^2}{4} \text{ pour } n > N_2$$

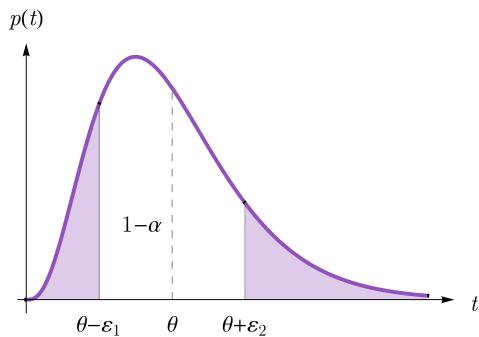
Par suite, pour n supérieur au plus grand des deux nombres N_1 et N_2 , on aura donc, comme on le voulait :

$$\mathbb{P}(|T - \theta| > \varepsilon) < \eta$$

Ce critère permet de définir l'*efficacité* d'un estimateur : *un estimateur est d'autant plus efficace que sa variance est plus petite*.

1.4 Intervalle de confiance d'une estimation

Il reste maintenant à définir la précision de l'estimation. Considérons, pour cela, la distribution de la variable aléatoire T . Si nous convenons de considérer comme négligeable un certain seuil de probabilité α , nous pouvons déterminer un intervalle $[\theta - \varepsilon_1, \theta + \varepsilon_2]$ tel qu'il lui corresponde la probabilité $(1 - \alpha)$.



Il résulte de la définition même de cet intervalle que l'on a la probabilité $(1 - \alpha)$ d'observer l'évènement $\{\theta - \varepsilon_1 \leq T \leq \theta + \varepsilon_2\}$.

Cela étant, chaque fois que cette double inégalité est vérifiée, c'est-à-dire dans la proportion $(1 - \alpha)$ des cas, la double inégalité :

$$\theta - \varepsilon_2 \leq \theta \leq \theta + \varepsilon_1$$

est, elle aussi, vérifiée. L'intervalle :

$$[\theta - \varepsilon_2, \theta + \varepsilon_1]$$

est ainsi un intervalle aléatoire auquel peut être associée la probabilité $(1 - \alpha)$ de recouvrir la vraie valeur inconnue de θ :

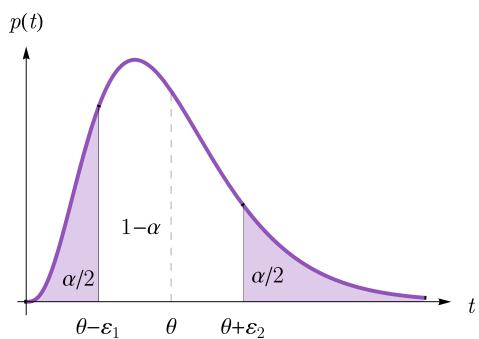
$$\mathbb{P}(\theta - \varepsilon_2 \leq \theta \leq \theta + \varepsilon_1) = 1 - \alpha$$

Si maintenant, nous observons les résultats de l'échantillon effectivement prélevé et calculons la valeur θ^* de T pour cet échantillon, l'intervalle :

$$[\theta^* - \varepsilon_2, \theta^* + \varepsilon_1]$$

est appelé *intervalle de confiance* de l'estimation de θ , au seuil de probabilité $(1 - \alpha)$.

Remarquons qu'il y a une infinité de façons de répartir la probabilité α , dont l'une correspond à un intervalle minimal, mais qui n'est pas toujours facile à déterminer en pratique. C'est pourquoi on convient généralement de répartir α par moitié de part et d'autre de l'intervalle. Cette répartition donne lieu à l'intervalle minimum dans le cas particulier où la densité de probabilité est symétrique et décroît pour des valeurs qui s'éloignent de θ .



1.5 Estimation d'une proportion

Considérons une population qui contient une proportion ϖ inconnue de pièces défectueuses. La variable aléatoire K , nombre de pièces défectueuses dans un échantillon de taille n , suit une loi binomiale de moyenne $n\varpi$ et de variance $n\varpi(1-\varpi)$. Si nous considérons maintenant la variable fréquence $\frac{K}{n}$, elle a pour moyenne ϖ et pour variance $\frac{\varpi(1-\varpi)}{n}$. $\frac{K}{n}$ a donc les propriétés d'un estimateur sans biais de ϖ ($\mathbb{E}\left(\frac{K}{n}\right) = \varpi$) et convergent ($\mathbb{V}\left(\frac{K}{n}\right) \xrightarrow{n \rightarrow \infty} 0$).

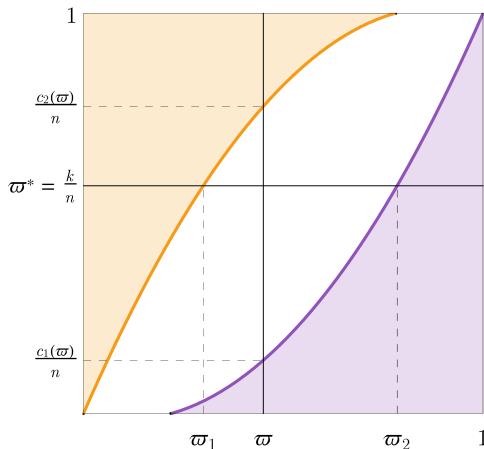
Pour estimer ϖ , il suffit donc, ayant prélevé un échantillon de taille n dans la population, de calculer le nombre k de pièces défectueuses puis :

$$\varpi^* = \frac{k}{n}$$

Pour déterminer l'intervalle de confiance, au risque α , de cette estimation, soient $c_1(\varpi)$ et $c_2(\varpi)$ les nombres tels que pour chaque valeur possible de ϖ :

$$\frac{\alpha}{2} \leq \sum_{k=0}^{c_1(\varpi)} \binom{n}{k} \varpi^k (1-\varpi)^{n-k} \quad \text{et} \quad \frac{\alpha}{2} \leq \sum_{k=c_2(\varpi)}^{\infty} \binom{n}{k} \varpi^k (1-\varpi)^{n-k}$$

Il est alors possible de construire le graphe ci-après en portant ϖ en abscisse et $\frac{c_1(\varpi)}{n}$ ou $\frac{c_2(\varpi)}{n}$ en ordonnée. La surface comprise entre les deux courbes est ainsi le lieu des valeurs possibles de $\frac{k}{n}$ pour l'ensemble des valeurs possibles de ϖ , lorsqu'on néglige le seuil de probabilité α .



Lisant maintenant le graphique suivant l'horizontale d'ordonnée $\varpi^* = \frac{k}{n}$, on peut immédiatement obtenir l'intervalle de confiance $[\varpi_1, \varpi_2]$ correspondant à cette estimation.

1.6 Estimation d'une moyenne

Etant donnée une population de moyenne μ inconnue et de variance σ^2 connue, soit M la variable aléatoire moyenne d'un échantillon de taille n . On a montré que $\mathbb{E}(M) = \mu$ et $\mathbb{V}(M) = \frac{\sigma^2}{n}$. M constitue donc un estimateur sans biais (et convergent) de μ . Par conséquent, ayant prélevé un échantillon, sa moyenne m est une estimation de μ :

$$m = \mu^*$$

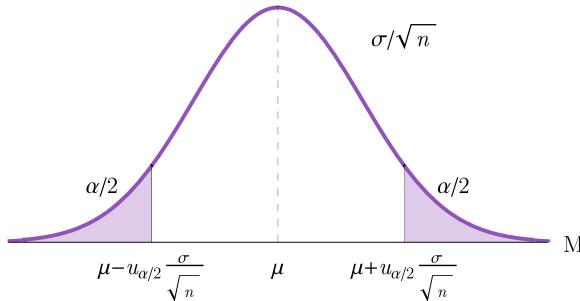
Si ce résultat est absolument général et qu'il est, en particulier, indépendant de la forme analytique de la loi de probabilité suivie par les observations, la détermination d'un intervalle de confiance nécessite la connaissance de cette forme. Admettons qu'il s'agisse d'une loi normale, de variance σ^2 connue.

Il en résulte que M suit aussi une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Etant donné un seuil de probabilité α , on peut alors écrire :

$$\mathbb{P}\left(\mu - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq M \leq \mu + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

où $u_{\alpha/2}$ est lu dans la table de la loi normale centrée réduite de telle façon que :

$$\mathbb{P}(|U| \geq u_{\alpha/2}) = \alpha$$



On en déduit l'intervalle de confiance au risque α de μ :

$$m - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq m + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Bien souvent la variance σ^2 n'est pas connue. Nous verrons, dans la suite du chapitre, comment procéder dans ce cas.

Notons que le *théorème central limite* permet de généraliser ces résultats à une loi de probabilité quelconque, mais à condition que n soit suffisamment grand (quelques dizaines en pratique).

1.7 Estimation d'une variance

Soit une population quelconque de variance inconnue. Soit S^2 la variable aléatoire « *variance d'un échantillon de taille n* », qui s'écrit :

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2$$

où M désigne la *variable aléatoire moyenne d'un échantillon de taille n* . On peut encore écrire :

$$S^2 = \frac{1}{n} \sum_{i=1}^n [(X_i - \mu) - (M - \mu)]^2$$

et, en développant le carré (moyenne des carrés moins carré de la moyenne) :

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (M - \mu)^2$$

Calculons alors $\mathbb{E}(S^2)$. Compte tenu de la linéarité de l'opérateur Espérance, il vient :

$$\mathbb{E}(S^2) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] - \mathbb{E}[(M - \mu)^2]$$

et, en notant que $\mathbb{E}[(X_i - \mu)^2]$ et $\mathbb{E}[(M - \mu)^2]$ sont respectivement les variances de X_i et de M :

$$\mathbb{E}(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2$$

On pourrait montrer, d'autre part, mais les calculs sont assez laborieux, que :

$$\mathbb{V}(S^2) = \left(\frac{n-1}{n}\right)^2 \times \frac{\mu_4 - \sigma^4}{n} + 2 \frac{n-1}{n^3} \sigma^4$$

où μ_4 désigne le moment centré d'ordre 4, $\mathbb{E}[(X_i - \mu)^4]$.

Il en résulte que S^2 est un estimateur convergent de σ^2 , mais qu'il n'est pas sans biais. Le passage à un estimateur sans biais ne présente toutefois aucune difficulté. Il suffit de considérer la quantité : $\frac{n}{n-1} S^2$ dont l'espérance mathématique est :

$$\mathbb{E}\left(\frac{n}{n-1} S^2\right) = \frac{n}{n-1} \mathbb{E}(S^2) = \sigma^2$$

Nous noterons, par la suite σ^{*2} cette estimation sans biais de σ^2 :

$$\sigma^{*2} = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1}$$

qui est calculée directement par presque toutes les calculettes effectuant des calculs statistiques.

Le calcul de l'intervalle de confiance, dans le cas d'une population normale, nécessite la connaissance préalable d'une nouvelle loi d'échantillonnage que nous allons établir maintenant.

2 Intervalle de confiance de la variance inconnue d'une population normale

2.1 Loi du χ^2

Soient U_1, U_2, \dots, U_ν , ν variables aléatoires *indépendantes* qui suivent des lois *normales réduites*. Posons :

$$Z^2 = U_1^2 + U_2^2 + \dots + U_\nu^2$$

La variable Z^2 suit alors une loi appelée *loi du χ^2 (khi deux)* à ν degrés de liberté ce que l'on note $Z^2 \sim \chi^2(\nu)$. Nous allons établir la forme analytique de cette loi qui joue un rôle essentiel en statistique et dont il faut retenir la définition, mais le calcul qui suit n'est, quant à lui, pas essentiel.

Nous nous proposons de déterminer la loi de probabilité de la variable Z^2 ; mais nous allons d'abord déterminer la loi de probabilité de la variable Z , c'est-à-dire que nous allons calculer la probabilité pour que Z se trouve compris dans l'intervalle $[c, c+dc]$.

La probabilité pour qu'on ait à la fois :

$$u_1 \leq U_1 < u_1 + du_1, u_2 \leq U_2 < u_2 + du_2, \dots \text{ et } u_\nu \leq U_\nu < u_\nu + du_\nu$$

s'écrit, U_1, U_2, \dots, U_ν étant *indépendantes* :

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{u_1^2}{2}} du_1 \times \dots \times \frac{1}{\sqrt{2\pi}} e^{-\frac{u_\nu^2}{2}} du_\nu = \frac{1}{(2\pi)^{\nu/2}} e^{-\frac{1}{2}(u_1^2 + \dots + u_\nu^2)} du_1 du_2 \dots du_\nu$$

Ce résultat peut s'interpréter géométriquement de manière très simple en considérant, dans un espace à ν dimensions, le point P de coordonnées (U_1, U_2, \dots, U_ν) . La probabilité pour que P tombe à l'intérieur d'un certain volume dv défini autour du point P est :

$$\frac{1}{(2\pi)^{\nu/2}} e^{-\frac{1}{2}\overline{OP}^2} dv$$

Dans ces conditions, la probabilité pour que la variable χ_ν soit comprise entre les deux valeurs c et $c+dc$ est égale à la probabilité pour que P tombe dans la région comprise entre les deux sphères de centre O et de rayons c et $c+dc$. Or la densité de probabilité est constante dans cette région et vaut :

$$\frac{1}{(2\pi)^{\nu/2}} e^{-\frac{c^2}{2}}$$

D'autre part, le volume de la sphère de centre O et de rayon c est de la forme $k c^\nu$ où k est une certaine constante, si bien que le volume compris entre les deux sphères de rayons c et $c+dc$ est égal à $dv = k \nu c^{\nu-1} dc$. On obtient alors finalement :

$$\mathbb{P}(c \leq Z < c+dc) = \frac{k \nu}{(2\pi)^{\nu/2}} c^{\nu-1} e^{-\frac{c^2}{2}} dc$$

Pour obtenir maintenant la loi de probabilité de la variable χ_ν^2 , faisons le changement de variable $x = c^2$. On obtient :

$$\mathbb{P}(x \leq Z^2 < x + dx) = \frac{k\nu}{2(2\pi)^{\nu/2}} x^{(\frac{\nu}{2}-1)} e^{-\frac{x}{2}} dx$$

qui définit la loi de probabilité cherchée, à la constante k près.

Pour calculer cette constante, on peut écrire que $\mathbb{P}(Z^2 \geq 0) = 1$, soit :

$$\frac{k\nu}{2(2\pi)^{\nu/2}} \int_0^\infty x^{(\frac{\nu}{2}-1)} e^{-\frac{x}{2}} dx = 1$$

d'où finalement l'expression de la loi du χ^2 :

$$\mathbb{P}(x \leq Z^2 < x + dx) = \frac{x^{(\nu/2-1)} e^{-x/2}}{\int_0^\infty x^{(\nu/2-1)} e^{-x/2} dx} dx$$

Notons qu'on définit en mathématiques les fonctions Γ (gamma) dont les équations sont de la forme : $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$, où a est une constante positive. L'intégrale au dénominateur de l'expression de la loi de probabilité du χ^2 peut alors s'écrire : $\int_0^\infty x^{(\frac{\nu}{2}-1)} e^{-\frac{x}{2}} dx = 2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)$, et on peut la calculer à partir des valeurs des fonctions Γ .

Il existe des tables de la loi du χ^2 qui donnent généralement, pour chaque valeur du nombre ν de degrés de liberté, la valeur x telle que : $\mathbb{P}(Z^2 > x) = \alpha$, où α est une probabilité donnée. Il est par conséquent inutile de retenir la formule de la loi du χ^2 . Par contre, il est indispensable de connaître les propriétés suivantes de la loi du χ^2 .

2.1.1 Sommes de variables suivant des lois du χ^2

Si Z_1^2 et Z_2^2 sont deux variables *indépendantes* qui suivent des lois du χ^2 à respectivement ν_1 et ν_2 degrés de liberté, leur somme $Z_1^2 + Z_2^2$ suit une loi du χ^2 à $(\nu_1 + \nu_2)$ degrés de liberté. Cela résulte immédiatement de la définition de la loi du χ^2 .

2.1.2 Moyenne et variance d'une variable qui suit une loi du χ^2

Si Z^2 est une loi du χ^2 à ν degrés de liberté, alors :

$$\mathbb{E}(Z^2) = \nu \text{ et } \mathbb{V}(Z^2) = 2\nu$$

En effet, U étant une variable centrée réduite, on a $\mathbb{E}(U) = 0$ et $\mathbb{V}(U) = 1$. Or $\mathbb{V}(U) = \mathbb{E}(U^2) - \mathbb{E}(U)^2$. Il en résulte que $\mathbb{E}(U^2) = 1$.

D'autre part, l'espérance d'une somme est égale à la somme des espérances, d'où :

$$\mathbb{E}(Z^2) = \mathbb{E}(U_1^2) + \mathbb{E}(U_2^2) + \cdots + \mathbb{E}(U_\nu^2) = \nu$$

Pour calculer la variance, ayant affaire à une somme de variables indépendantes, on calcule :

$$\mathbb{V}(Z^2) = \mathbb{V}(U_1^2) + \mathbb{V}(U_2^2) + \cdots + \mathbb{V}(U_\nu^2) = \nu \mathbb{V}(U^2)$$

et sachant que la variance est égale à l'espérance du carré moins le carré de l'espérance :

$$\mathbb{V}(U^2) = \mathbb{E}(U^4) - \mathbb{E}(U^2)^2$$

Enfin, on montre facilement, en intégrant par parties, que :

$$\mathbb{E}(U^4) = \int_{-\infty}^{\infty} u^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = 3$$

d'où $\mathbb{V}(Z^2) = \nu(3 - 1) = 2\nu$

Ou avec *Mathematica* :

```
Expectation[{U^2, U^4}], U \[Distributed] NormalDistribution[0, 1]
```

```
{1, 3}
```

```

{e, e2} = Expectation[{z, z^2}, z ~ ChiSquareDistribution[v]]
{v, v (2 + v) }

Simplify[e2 - e^2] (* définition de la variance *)
2 v

```

2.2 Loi de la variance d'un échantillon extrait d'une population normale dont l'écart-type est connu

Nous allons montrer que, étant donné un échantillon de taille n qui est extrait d'une population *normale* de moyenne μ et de variance égale à σ^2 , la variable aléatoire :

$$\frac{n S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - M)^2}{\sigma^2}$$

suit une *loi du χ^2* à $(n - 1)$ degrés de liberté.

Nous montrerons d'autre part, que les variables M et $\sum_{i=1}^n (X_i - M)^2$ sont des variables *indépendantes*. Cette dernière propriété sera utilisée un peu plus loin.

La démonstration qui suit peut être omise.

Au cours de ce chapitre, nous avons montré que l'on peut écrire :

$$\sum_{i=1}^n (X_i - M)^2 = \sum_{i=1}^n X_i^2 - n M^2$$

Soit alors P le point de coordonnées (X_1, X_2, \dots, X_n) , et soient (Y_1, Y_2, \dots, Y_n) ses nouvelles coordonnées après un changement de coordonnées orthonormales, que nous allons choisir tel que la coordonnée Y_n soit justement égale à $\sqrt{n} M$. Dans ces conditions, $\sum_{i=1}^n X_i^2 - n M^2$ sera égal à $\sum_{j=1}^{n-1} Y_j^2$, c'est-à-dire à la somme de $(n - 1)$ variables, dont nous allons montrer qu'elles sont indépendantes et qu'elles ont même variance σ^2 .

Ces conditions sont réalisées si l'axe des Y_n est choisi passant par le vecteur unitaire dont toutes les coordonnées sont égales à $1 / \sqrt{n}$ dans l'ancien système d'axes. Les nouvelles coordonnées de P s'écrivent alors :

$$\begin{aligned}
Y_1 &= a_1^1 X_1 + \cdots + a_i^1 X_i + \cdots + a_n^1 X_n \\
&\dots \\
Y_j &= a_1^j X_1 + \cdots + a_i^j X_i + \cdots + a_n^j X_n \\
&\dots \\
Y_{n-1} &= a_1^{n-1} X_1 + \cdots + a_i^{n-1} X_i + \cdots + a_n^{n-1} X_n \\
Y_n &= \frac{1}{\sqrt{n}} X_1 + \cdots + \frac{1}{\sqrt{n}} X_i + \cdots + \frac{1}{\sqrt{n}} X_n
\end{aligned}$$

Les a_i^j sont déterminés d'une infinité de façons par les relations :

$$\begin{aligned}
\frac{a_1^j}{\sqrt{n}} + \cdots + \frac{a_i^j}{\sqrt{n}} + \cdots + \frac{a_n^j}{\sqrt{n}} &= \frac{1}{\sqrt{n}} \sum_i a_i^j = 0 \quad (\forall j, Y_j \perp Y_n \text{ soit } Y_j \cdot Y_n = 0) \\
a_1^j a_1^k + \cdots + a_i^j a_i^k + \cdots + a_n^j a_n^k &= 0 \quad (\forall j, k, Y_j \perp Y_k \text{ soit } Y_j \cdot Y_k = 0) \\
(a_1^j)^2 + \cdots + (a_i^j)^2 + \cdots + (a_n^j)^2 &= 1 \quad (\forall j, \|Y_j\| = 1 \text{ soit } Y_j \cdot Y_j = 1)
\end{aligned}$$

Notant que les X_i sont indépendantes entre elles et de même moyenne μ , on en déduit que :

$$\begin{aligned}
\mathbb{E}(Y_j) &= \mathbb{E}\left(\sum_i a_i^j X_i\right) = \sum_i a_i^j \mathbb{E}(X_i) = \sum_i a_i^j \times \mu = \mu \sum_i a_i^j = 0 \\
\mathbb{V}(Y_j) &= \mathbb{V}\left(\sum_i a_i^j X_i\right) = \sum_i a_i^{j2} \mathbb{V}(X_i) = \sum_i a_i^{j2} \sigma^2 = \sigma^2 \sum_i a_i^{j2} = \sigma^2
\end{aligned}$$

Les nouvelles variables Y_j suivent donc des lois normales (combinaisons linéaires de variables normales), centrées, de variance σ^2 . Elles sont indépendantes puisque orthonormales (on peut vérifier que $\mathbb{E}(Y_j Y_k) = 0 = \mathbb{E}(Y_j) \times \mathbb{E}(Y_k)$ quoique cela ne constitue pas une condition suffisante d'indépendance). On sait également que $\sum_{j=1}^n Y_j^2 = \sum_{i=1}^n X_i^2$ puisqu'il s'agit d'un changement de coordonnées orthonormales. Comme Y_n a été choisi de telle sorte que $Y_n^2 = n M^2$, la variable :

$$\frac{n S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - M)^2}{\sigma^2} = \frac{\sum_{i=1}^n X_i^2 - n M^2}{\sigma^2}$$

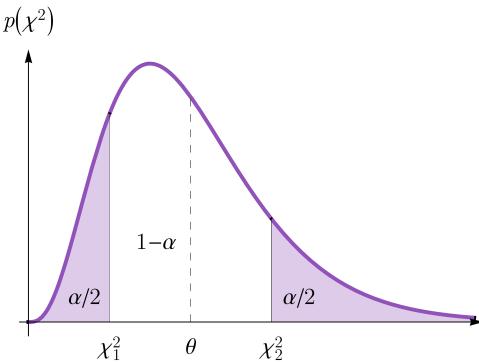
peut donc s'exprimer en fonction des carrés de $(n - 1)$ variables normales, réduites et indépendantes :

$$\frac{n S^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - M)^2}{\sigma^2} = \sum_{j=1}^{n-1} \left(\frac{Y_j}{\sigma} \right)^2$$

Elle suit donc une loi du χ^2 à $(n - 1)$ degrés de liberté. Et la variable Y_n étant indépendante des variables Y_1, \dots, Y_{n-1} , M et $\sum_{i=1}^n (X_i - M)^2$ sont nécessairement des variables indépendantes.

2.3 Intervalle de confiance de la variance inconnue d'une population normale

Soit une population normale de variance σ^2 inconnue. La variable aléatoire $\frac{n S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - M)^2}{\sigma^2}$ suit une loi du χ^2 à $(n - 1)$ degrés de liberté.



Se fixant un seuil de probabilité α , il est possible de déterminer l'intervalle $[\chi_1^2, \chi_2^2]$ tel que :

$$\chi_1^2 \leq \frac{n S^2}{\sigma^2} \leq \chi_2^2 \text{ avec la probabilité } (1 - \alpha).$$

On en déduit l'intervalle de confiance pour σ^2 , au risque α :

$$\frac{n S^2}{\chi_2^2} \leq \sigma^2 \leq \frac{n S^2}{\chi_1^2}$$

3 Intervalle de confiance de la moyenne inconnue d'une population normale d'écart-type inconnu

3.1 Loi de Student

Considérons $v + 1$ variables aléatoires *normales, réduites, indépendantes* entre elles. Désignons les par $U, U_1, \dots, U_i, \dots, U_v$. La variable :

$$T = \frac{U}{\sqrt{\frac{1}{v} \sum_{i=1}^v U_i^2}}$$

suit, par définition, une loi de Student à v degrés de liberté ce que nous noterons ainsi $T \sim \mathcal{T}(v)$.

En remarquant que $Z^2 = \sum_{i=1}^{\nu} U_i^2$ suit une loi du χ^2 à ν degrés de liberté, on peut encore écrire T sous la forme suivante :

$$T = \frac{U}{\sqrt{Z^2/\nu}}$$

où U et Z^2 sont des variables indépendantes qui suivent respectivement une loi $N(0, 1)$ et une loi $\chi^2(\nu)$.

Pour $\nu = 1$, la loi de Student s'identifie à une loi appelée la loi de Cauchy connue pour n'avoir ni moyenne, ni variance finies. On montre d'autre part que, lorsque $\nu \rightarrow \infty$, la loi de Student tend vers une loi normale centrée réduite. Mais, pour ν fini, elle est plus étalée que la loi normale, sa variance (pour $\nu > 2$) étant égale à $\frac{\nu}{\nu-2} > 1$.

Il existe des tables donnant, pour un nombre de degrés de liberté donné, et pour des seuils de probabilité α fixés les valeurs t telles que : $\mathbb{P}(|T| > t) = \alpha$.

3.2 Loi de la moyenne d'un échantillon extrait d'une population normale d'écart-type inconnu

En notant σ^* l'estimation sans biais de σ^2 :

$$\sigma^{*2} = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1}$$

nous allons montrer que la quantité :

$$t = \frac{m-\mu}{\sigma^*/\sqrt{n}} = \frac{m-\mu}{s/\sqrt{n-1}}$$

est une réalisation d'une variable de Student à $(n - 1)$ degrés de liberté.

En effet, la variable :

$$U = \frac{M-\mu}{\sigma/\sqrt{n}}$$

suit une loi normale centrée réduite puisque si les X_i suivent une loi normale de moyenne μ et d'écart-type σ , M suit une loi normale de moyenne μ et écart-type $\frac{\sigma}{\sqrt{n}}$.

D'autre part, la variable :

$$\frac{n S^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - M)^2}{\sigma^2}$$

suit une loi du χ^2 à $(n - 1)$ degrés de liberté. Et ces deux variables sont *indépendantes*.

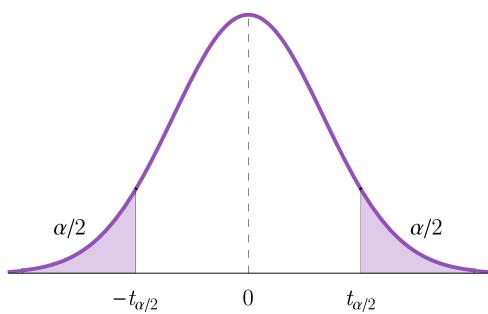
Donc, la variable :

$$T = \frac{U}{\sqrt{\frac{n S^2}{\sigma^2}/(n-1)}} = \frac{M-\mu}{\sqrt{\sigma^2 \frac{n S^2}{\sigma^2}/n(n-1)}} = \frac{M-\mu}{S/\sqrt{n-1}} = \frac{M-\mu}{S/\sqrt{n-1}}$$

suit une loi de Student à $(n - 1)$ degrés de liberté.

3.3 Intervalle de confiance de la moyenne inconnue d'une population normale d'écart-type inconnu

Soit $t_{\alpha/2}$ la valeur lue dans la table de Student à $(n - 1)$ degrés de liberté, correspondant au risque α réparti symétriquement.



Il en résulte que :

$$-t_{\alpha/2} \leq \frac{m-\mu}{s/\sqrt{n-1}} \leq +t_{\alpha/2}$$

Que l'on peut écrire également ainsi :

$$-t_{\alpha/2} \leq \frac{m-\mu}{\sigma^*/\sqrt{n}} \leq +t_{\alpha/2}$$

Et donc l'on déduit, l'intervalle de confiance de la moyenne inconnue μ au risque α :

$$m - t_{\alpha/2} \frac{\sigma^*}{\sqrt{n}} \leq \mu \leq m + t_{\alpha/2} \frac{\sigma^*}{\sqrt{n}}$$

expression dont la forme est la même que celle de l'intervalle de confiance établi dans le cas où l'écart-type σ est connu :

$$m - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq m + u_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

4 Estimation du maximum de vraisemblance

L'estimation du maximum de vraisemblance est une autre méthode statistique couramment utilisée pour inférer les paramètres de la loi de probabilité d'une population à partir d'un échantillon issu de cette population. Cette méthode est due à Ronald Fisher (1922). Elle consiste à estimer les paramètres recherchés de la loi de probabilité de la population par les valeurs de ces paramètres qui maximisent la probabilité d'obtenir l'échantillon observé.

Supposons que la population de référence est caractérisée par une variable aléatoire X ne dépendant que d'un seul paramètre θ et notons $p(x|\theta)$ sa densité de probabilité si la variable est continue ou la probabilité que $X = x$ si elle est discrète. On appelle *vraisemblance de θ au vu de l'échantillon x_1, x_2, \dots, x_n* , tiré au hasard, la fonction :

$$L(x_1, x_2, \dots, x_n | \theta) = L(\theta) = p(x_1 | \theta) \times p(x_2 | \theta) \times \dots \times p(x_n | \theta) = \prod_{i=1}^n p(x_i | \theta)$$

La méthode du maximum de vraisemblance consiste donc à rechercher la valeur de θ , que nous noterons θ^* , qui maximise cette fonction, ce que l'on note $\theta^* = \text{argmax}(L(\theta))$, c'est-à-dire qui rende maximale la probabilité d'obtenir l'échantillon observé. La recherche de cette valeur se ramène donc un problème d'optimisation. Si la fonction L est deux fois dérivable et admet un maximum global en θ^* , alors sa dérivée première s'annule en θ^* et sa dérivée seconde est négative :

$$L'(\theta^*) = 0 \text{ et } L''(\theta^*) \leq 0$$

Inversement, si on trouve une valeur de θ^* satisfaisant ces conditions, alors θ^* est un maximum local. Il faut alors vérifier s'il est aussi un maximum global.

La fonction L étant positive et le logarithme népérien une fonction croissante, on utilise souvent le logarithme népérien de la vraisemblance pour calculer θ^* car le calcul de la dérivée est alors souvent plus simple (le produit des densités étant ramené à une somme). On l'appelle la *log-vraisemblance*. Le paramètre θ^* recherché est donc la valeur de θ qui satisfait les conditions suivantes en plus de correspondre au maximum global :

$$\frac{\partial \ln(L(\theta))}{\partial \theta} = 0 \text{ et } \frac{\partial^2 \ln(L(\theta))}{\partial^2 \theta} \leq 0$$

A chaque échantillon tiré au hasard dans la population correspond une valeur de θ^* . θ^* est donc une réalisation d'une variable aléatoire. A la section 1.2, nous avons noté T cette variable et défini sous quelles conditions elle constituait un estimateur de θ .

On peut montrer que l'estimateur du maximum de vraisemblance est un estimateur convergent, asymptotiquement efficace (variance minimale) et asymptotiquement distribué selon une loi normale. Par contre il peut être biaisé.

4.1 Estimation du paramètre d'une loi de Poisson

Dans le cas d'un échantillon issu d'une population obéissant à une loi de Poisson, il s'agit d'estimer le paramètre λ de cette loi. La vraisemblance de λ au vu de l'échantillon (k_1, k_2, \dots, k_n) et son logarithme s'écrivent :

$$\begin{aligned} L(\lambda) &= p(k_1 | \lambda) \times p(k_2 | \lambda) \times \cdots \times p(k_n | \lambda) = \prod_{i=1}^n e^{-\lambda} \times \frac{\lambda^{k_i}}{k_i!} \\ \log(L(\lambda)) &= \sum_{i=1}^n \log\left(e^{-\lambda} \times \frac{\lambda^{k_i}}{k_i!}\right) = \sum_{i=1}^n \left(-\lambda + \log\left(\frac{\lambda^{k_i}}{k_i!}\right)\right) = -\lambda n + \sum_{i=1}^n \left(k_i \log(\lambda) - \log(k_i!)\right) \\ &= -\lambda n + \log(\lambda) \sum_{i=1}^n k_i - \sum_{i=1}^n \log(k_i!) \end{aligned}$$

On calcule et on annule la dérivée première par rapport à λ :

$$\frac{\partial \log(L(\lambda))}{\partial \lambda} = -n + \frac{\sum_{i=1}^n k_i}{\lambda} = 0 \implies \lambda^* = \frac{\sum_{i=1}^n k_i}{n}$$

On vérifie par ailleurs que :

$$\frac{\partial^2 \log(L(\lambda))}{\partial^2 \theta} = -\frac{\sum_{i=1}^n k_i}{\lambda^2} \leq 0$$

L'estimation la plus vraisemblable du paramètre λ est donc *la moyenne de l'échantillon*.

4.2 Estimation du paramètre d'une loi exponentielle

Dans le cas d'une loi exponentielle de paramètre λ , la log-vraisemblance de λ au vu de l'échantillon (t_1, t_2, \dots, t_n) s'écrit :

$$\log(L(\lambda)) = \log\left(\prod_{i=1}^n \lambda e^{-\lambda t_i}\right) = \sum_{i=1}^n \log(\lambda e^{-\lambda t_i}) = \sum_{i=1}^n (\log(\lambda) - \lambda t_i) = n \log(\lambda) - \lambda \sum_{i=1}^n t_i$$

On annule la dérivée première :

$$\frac{\partial \log(L(\lambda))}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n t_i = 0 \implies \lambda^* = \frac{n}{\sum_{i=1}^n t_i} = \frac{1}{\frac{\sum_{i=1}^n t_i}{n}} = 1 / \bar{t}$$

La dérivée seconde est toujours négative :

$$\frac{\partial^2 \log(L(\lambda))}{\partial^2 \lambda} = -\frac{n}{\lambda^2} \leq 0$$

On a donc bien affaire à un maximum global. L'estimation la plus vraisemblable du paramètre λ d'une loi exponentielle est donc *l'inverse de la moyenne de l'échantillon*.

4.4 Estimation des paramètres d'une loi normale

Intéressons nous maintenant à une population normale dont on cherche à estimer les paramètres μ et σ^2 à partir d'un échantillon. La vraisemblance et la log-vraisemblance de μ et de σ au vu de l'échantillon (x_1, x_2, \dots, x_n) s'écrivent :

$$L(\mu, \sigma) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$$

$$\log(L(\mu, \sigma)) = \frac{n}{2} \log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

On recherche le maximum de cette dernière fonction en résolvant le système d'équations suivant :

$$\begin{cases} \frac{\partial \log(L(\mu, \sigma))}{\partial \mu} = -\frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} = 0 \\ \frac{\partial \log(L(\mu, \sigma))}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^3} = 0 \end{cases}$$

On montre alors que la log-vraisemblance est maximale en :

$$\begin{cases} \mu^* = \frac{\sum_{i=1}^n x_i}{n} = m \\ \sigma^{*2} = \frac{\sum_{i=1}^n (x_i - m)^2}{n} \end{cases}$$

On retrouve l'estimateur de μ vu en section 1.6. Par contre l'estimateur de la variance ici obtenu est un estimateur biaisé comme vu à la section 1.7.

Exercices du chapitre 4

Exercice 1

S'assurer que l'on connaît parfaitement bien les réponses aux questions suivantes.

- a) Définition d'une variable normale réduite à partir d'une variable normale quelconque ?
- b) Définition de la loi du χ^2 ?
- c) Définition de la loi de Student ?
- d) Quelle est la loi qui fait intervenir la moyenne d'un échantillon et les paramètres de la loi normale de référence ?
- e) Comment estimer la variance de la loi de référence
 - si sa moyenne n'est pas connue,
 - si par contre elle est connue ?
- f) Quelle est la loi qui fait intervenir la variance de la loi de référence et son estimation ?
- g) Que devient la loi précisée en d) si l'écart-type de la loi de référence n'est pas connu et qu'il faille l'estimer ?

Exercice 2

On a mesuré la capacité (en microfarad) de 25 condensateurs et calculé la moyenne $m = 2.086$ et l'écart-type $s = 0.079$. Déterminer les intervalles de confiance de l'estimation de la moyenne μ de la population normale de référence, en choisissant un risque de 5 %, puis de 1 %. Est-il normal que le second soit plus grand que le premier ?

Exercice 3

L'airbag (ou coussin gonflable) est un système de sécurité de plus en plus souvent installé dans les automobiles. Son gonflement est assuré par un dispositif pyrotechnique dont les caractéristiques importantes sont la moyenne et l'écart-type du délai entre la mise à feu et l'explosion. Lors de l'étude d'un certain type de dispositif d'allumage, les résultats des mesures, effectuées sur 10 exemplaires, ont été (en millisecondes) : {28, 28, 31, 31, 33, 30, 31, 27, 32, 29}.

- a) Calculer, au risque 5%, l'intervalle de confiance de la moyenne du délai si on connaît l'écart-type de la population de référence et qu'il est égal à 2.
- b) Calculer ce même intervalle si on ne connaît pas l'écart-type de la population de référence.
- c) Calculer, au même risque, l'intervalle de confiance de la variance du délai, dont on déduira celui de l'écart-type dans le cas où on ne connaît pas l'écart-type de la population de référence.

Exercice 4

Les poids de pièces usinées en cuivre sont distribués normalement. Ayant prélevé 9 pièces, on a obtenu les poids suivants en grammes : 18.457 ; 18.434 ; 18.444 ; 18.461 ; 18.453 ; 18.447 ; 18.452 ; 18.440 ; 18.443. Calculer m et s^2 , puis les intervalles de confiance au risque 5 % de la moyenne μ et de la variance σ^2 inconnues.

Exercice 5

On dispose de 15 mesures du diamètre apparent vertical de la planète Vénus et calculé $m = 42.95$ (en secondes d'arc) et $s^2 = 0.212627$. On supposera ces mesures issues d'une distribution normale.

- 1) Calculer l'intervalle de confiance à 95% de la moyenne μ du vrai diamètre apparent vertical de la Vénus
- 2) Combien faudrait-il faire de mesures au minimum pour obtenir, au même niveau de confiance, une précision de ± 0.1 ?

Exercice 6

Un fabricant de piles électriques indique sur ses produits que la durée de vie moyenne de ses piles est de 200 heures. Une association de consommateurs prélève au hazard un échantillon de 100 piles et observe une durée de vie de 190 heures en moyenne avec un écart type de 30 heures. S'agit-il de publicité mensongère ?

Exercice 7

On a mesuré les durées de vie en heures de fonctionnement de 10 tubes électroniques du même type. On a obtenu les résultats ordonnés suivants : 26 ; 31 ; 34 ; 40 ; 49 ; 60 ; 72 ; 85 ; 123 ; 179. Trouver une estimation sans biais du taux de défaillance, en admettant le modèle du processus de Poisson.

Comparaisons statistiques

Nous présentons dans ce chapitre un raisonnement nouveau. Son inventeur, au début de ce siècle, avait pris le pseudonyme de Student. Le problème qui lui était posé était le suivant : l'engrais a-t-il une influence sur le rendement des cultures de pomme de terre ? Pour le résoudre, Student imagine de choisir 4 parcelles. Chacune d'elles est divisée en deux, et on la cultive en traitant l'une des moitiés choisie au hasard, avec de l'engrais et l'autre non. Après la récolte, on calcule les rendements et, pour une parcelle donnée, la différence de rendements entre les deux moitiés avec engrais et sans engrais. Les 4 différences obtenues sont : {11, 30, -6, 13}. Student convient de considérer ces valeurs comme des réalisations d'une variable aléatoire D . Il fait alors l'hypothèse que l'engrais n'a pas d'influence. Si cette hypothèse est vraie, la moyenne $\mathbb{E}(D)$ de la variable D est nulle. La démarche se poursuit par une sorte de raisonnement par l'absurde, en vérifiant si les valeurs observées peuvent être considérées comme compatibles ou non avec $\mathbb{E}(D) = 0$. Si elles sont incompatibles, l'hypothèse faite doit être remise en cause, et l'on peut conclure à l'influence de l'engrais ... Ce raisonnement, théorisé plus tard par Neyman et Pearson, est appelé le test d'hypothèse.

1 Tests d'hypothèse

1.1 Théorie de Neyman et Pearson

On suppose donnée une certaine variable aléatoire X dont la loi de probabilité dépend des hypothèses que l'on désire tester. Plus précisément, on suppose qu'il existe plusieurs hypothèses $\mathcal{H}_0, \mathcal{H}_1, \dots, \mathcal{H}_n$ parfaitement connues (qui peuvent être en nombre fini ou non, dénombrable ou non) et que la loi de probabilité dépend de l'hypothèse vraie. Le test va permettre de porter un jugement sur l'hypothèse faite et d'évaluer le degré de validité du jugement, cela à partir de la valeur prise par X .

Nous étudierons d'abord le cas où l'on fait deux hypothèses simples \mathcal{H}_0 et \mathcal{H}_1 . Une hypothèse est dite simple si elle définit complètement et d'une manière unique la loi de probabilité de X ; sinon, elle est dite composite. C'est ainsi, par exemple, qu'en présence d'un lot de pièces distinguées en *convenables* et *défectueuses*, les deux hypothèses :

- \mathcal{H}_0 : le lot contient 5 % de déchets
- \mathcal{H}_1 : le lot contient 10 % de déchets

sont des hypothèses simples puisque chacune d'elles définit entièrement le lot. Tandis que les deux hypothèses :

- \mathcal{H}_0 : le lot contient 5 % ou moins de 5 % de déchets
- \mathcal{H}_1 : le lot contient plus de 5 % de déchets

sont des hypothèses composites puisque ni l'une ni l'autre ne définit entièrement le lot.

Supposons donc qu'il existe deux hypothèses simples \mathcal{H}_0 et \mathcal{H}_1 couvrant l'ensemble des possibilités ; cela veut dire que l'une ou l'autre des deux hypothèses \mathcal{H}_0 et \mathcal{H}_1 est réalisée nécessairement. Dans ce cas, il est possible d'émettre l'un des deux jugements :

- \mathcal{H}_0 est vraie, donc \mathcal{H}_1 est fausse
- \mathcal{H}_1 est vraie, donc \mathcal{H}_0 est fausse

On peut symboliser cet ensemble par le tableau ci-dessous où figurent en première ligne les états possibles et en première colonne les jugements portés. Le tableau contient les conséquences des différentes combinaisons.

		état réalisé	
		\mathcal{H}_0 est réalisée	\mathcal{H}_1 est réalisée
jugement porté	\mathcal{H}_0 est vraie	jugement correct	jugement faux
	\mathcal{H}_1 est vraie	jugement faux	jugement correct

Parmi les deux hypothèses \mathcal{H}_0 et \mathcal{H}_1 , il en existe en général une dont le rejet à tort a des conséquences plus fâcheuses que pour l'autre. Il est donc normal de ne pas traiter \mathcal{H}_0 et \mathcal{H}_1 de façon symétrique. Admettant alors que \mathcal{H}_0 représente une circonstance favorable et \mathcal{H}_1 une circonstance défavorable, on peut se tromper de deux manières :

- en considérant comme défavorable ce qui est favorable ; c'est l'*erreur de première espèce* ;
- en considérant comme favorable ce qui ne l'est pas ; c'est l'*erreur de deuxième espèce*.

C'est exactement en ces termes que se posait le problème du contrôle de réception, où ces deux types d'erreur correspondaient à des préoccupations toutes différentes : celle du fournisseur d'une part, et celle du client d'autre part.

Pour relier maintenant le jugement porté à l'observation de la variable X , on opère ainsi :

- on dit que \mathcal{H}_0 est vraie si la valeur observée de X , soit x , se trouve dans un certain domaine w , appelé région d'acceptation de l'hypothèse \mathcal{H}_0 ;
- on dit que \mathcal{H}_1 est vraie si la valeur observée appartient à \bar{w} , appelé région critique ou région de rejet.

Pour choisir le domaine w , on impose en général deux conditions :

- que la probabilité de commettre l'erreur de première espèce soit égale à un seuil déterminé α choisi a priori aussi faible qu'on le veut ;
- que la probabilité β de commettre l'erreur de deuxième espèce soit minimale.

Il importe de noter en effet que la première condition ne suffit pas, sauf cas très particulier, à définir w de façon unique.

Il est possible maintenant de compléter le tableau précédent en indiquant les règles de jugement et les probabilités pour qu'il soit correct ou faux :

		état réalisé	
		\mathcal{H}_0 est réalisée	\mathcal{H}_1 est réalisée
jugement porté	\mathcal{H}_0 est vraie ($X \in w$)	jugement correct ($1-\alpha$)	jugement faux (β)
	\mathcal{H}_1 est vraie ($X \notin w$)	jugement faux (α)	jugement correct ($1-\beta$)

Un tel mode de raisonnement est appelé test d'hypothèse. Le complément à l'unité de β , soit $(1-\beta)$ est appelé puissance du test : un test est d'autant plus puissant, pour un risque de première espèce fixé, que le risque de deuxième espèce est plus petit.

1.2 Détermination de la région d'acceptation

Si l'on note $p_0(x)$ et $p_1(x)$ les densités de probabilité de X , respectivement dans le cadre des hypothèses \mathcal{H}_0 et \mathcal{H}_1 , les deux conditions précédentes s'expriment par les deux équations suivantes :

$$\begin{aligned} \int_w p_0(x) dx &= 1 - \alpha \\ \int_w p_1(x) dx &= \beta \text{ minimum} \end{aligned}$$

Neyman et Pearson ont démontré qu'elles sont satisfaites s'il existe une constante positive λ et un domaine w tels que pour x appartenant à w :

$$p_1(x) < \lambda p_0(x) \quad (1)$$

sous la contrainte :

$$\int_w p_0(x) dx = 1 - \alpha \quad (2)$$

Appliquons ce résultat à deux exemples.

1.3 Test sur une proportion

Supposons qu'ayant prélevé un échantillon de n pièces dans un certain lot, on veuille tester l'hypothèse :

- \mathcal{H}_0 : la proportion de déchets est ϖ_0 , contre l'hypothèse
- \mathcal{H}_1 : la proportion de déchets est ϖ_1 .

Le nombre d'articles défectueux dans l'échantillon est une variable aléatoire définie par les probabilités $p_0(k)$ si \mathcal{H}_0 est vraie et $p_1(k)$ si c'est \mathcal{H}_1 :

$$p_0(k) = \binom{n}{k} \varpi_0^k (1 - \varpi_0)^{n-k}$$

$$p_1(k) = \binom{n}{k} \varpi_1^k (1 - \varpi_1)^{n-k}$$

La condition (1) s'écrit :

$$\binom{n}{k} \varpi_1^k (1 - \varpi_1)^{n-k} < \lambda \binom{n}{k} \varpi_0^k (1 - \varpi_0)^{n-k}$$

Et, après simplification et passage aux logarithmes, on obtient :

$$k \log\left(\frac{\varpi_0}{\varpi_1}\right) + (n-k) \log\left(\frac{1-\varpi_0}{1-\varpi_1}\right) + \log(\lambda) > 0,$$

Soit, pour $\varpi_1 > \varpi_0$:

$$k < \frac{n \log\left(\frac{1-\varpi_1}{1-\varpi_0}\right) - \log(\lambda)}{\log\left(\frac{\varpi_1}{\varpi_0}\right) - \log\left(\frac{1-\varpi_1}{1-\varpi_0}\right)} = k_s$$

L'inégalité se réduit donc à $k < k_s$

Pour déterminer k_s , il suffit d'utiliser la condition (2) qui s'écrit :

$$\sum_{k=0}^{k_s} \binom{n}{k} \varpi_0^k (1 - \varpi_0)^{n-k} = 1 - \alpha$$

On notera que la région d'acceptation ne dépend pas de la valeur ϖ_1 , c'est-à-dire de l'hypothèse \mathcal{H}_1 . Par contre, le risque de deuxième espèce en dépend puisque :

$$\beta = \sum_{k=0}^{k_s} \binom{n}{k} \varpi_1^k (1 - \varpi_1)^{n-k}$$

1.4 Test sur une moyenne

Soit un échantillon de taille n prélevé dans une population normale X d'écart-type σ connu, mais de moyenne μ inconnue. Considérons les hypothèses :

$$\mathcal{H}_0 : \mu = \mu_0$$

$$\mathcal{H}_1 : \mu = \mu_1$$

Sous l'hypothèse \mathcal{H}_0 , la densité de probabilité d'un tel échantillon (X_1, X_2, \dots, X_n) s'écrit :

$$p_0(x_1) \times p_0(x_2) \times \cdots \times p_0(x_n)$$

les variables X_i étant indépendantes et de même densité de probabilité que celle de X . Dès lors, X étant une loi normale, la région d'acceptation est définie par :

$$\frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2}} < \frac{\lambda}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2}}$$

En simplifiant par le facteur et en passant au logarithme on a :

$$-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_1)^2}{\sigma^2} < \log(\lambda) - \frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2}$$

Expression que l'on peut écrire aussi, après multiplication par $2\sigma^2$:

$$\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \mu_1)^2 < 2\sigma^2 \log(\lambda)$$

$$\sum_{i=1}^n x_i^2 - 2\mu_0 \sum_{i=1}^n x_i + n\mu_0^2 - (\sum_{i=1}^n x_i^2 - 2\mu_1 \sum_{i=1}^n x_i + n\mu_1^2) < 2\sigma^2 \log(\lambda)$$

Soit, en notant m la moyenne empirique $m = \frac{\sum_{i=1}^n x_i}{n}$:

$$-2\mu_0 n m + n\mu_0^2 + 2\mu_1 n m - n\mu_1^2 < 2\sigma^2 \log(\lambda)$$

$$2m(\mu_1 - \mu_0) - (\mu_1^2 - \mu_0^2) < \frac{2\sigma^2 \log(\lambda)}{n}$$

$$2m(\mu_1 - \mu_0) - (\mu_1 - \mu_0)(\mu_1 + \mu_0) < \frac{2\sigma^2 \log(\lambda)}{n}$$

Et en supposant que $\mu_1 > \mu_0$:

$$\begin{aligned} 2m - (\mu_1 + \mu_0) &< \frac{2\sigma^2 \log(\lambda)}{n(\mu_1 - \mu_0)} \\ m &< \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2 \log(\lambda)}{n(\mu_1 - \mu_0)} = m_s \end{aligned}$$

Pour définir m_s , il suffit alors d'écrire que :

$$\mathbb{P}(M > m_s | \mu = \mu_0) = \alpha$$

où M désigne la variable aléatoire moyenne d'un échantillon de taille n . Remarquons que, dans ce deuxième exemple aussi, la région d'acceptation ne dépend pas de l'hypothèse \mathcal{H}_1 .

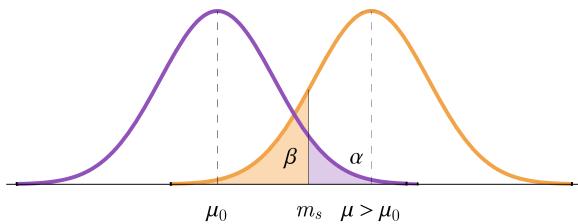
1.5 Cas d'hypothèses composites

En réalité, très souvent, le problème n'est pas de choisir entre deux hypothèses simples \mathcal{H}_0 et \mathcal{H}_1 , mais entre une hypothèse simple \mathcal{H}_0 et un ensemble plus ou moins vaste d'hypothèses $\mathcal{H}_1, \dots, \mathcal{H}_i, \dots, \mathcal{H}_n$, ou même à un ensemble continu d'hypothèses \mathcal{H} .

Dans ce cas, on peut se ramener au problème précédent en comparant successivement \mathcal{H}_0 à chacune des hypothèses de l'ensemble \mathcal{H} . Si, par exemple, on compare \mathcal{H}_0 à \mathcal{H}_i , la méthode exposée plus haut permet de trouver une région w_i telle que le risque de première espèce soit égal à α et que le risque de deuxième espèce β_i soit minimum. On obtient ainsi un ensemble de régions d'acceptation $w_1, \dots, w_i, \dots, w_n$ et, dans le cas général, on ne peut pas aller plus loin.

Mais il existe un cas particulier très intéressant, celui où les différentes régions w_i ont une partie commune w . Dans ce domaine w , le test utilisé est dit uniformément le plus puissant (en abréviation de l'anglais : UMP). En effet, lorsque X tombe dans w , on est sûr que le risque de première espèce est égal à α et que le risque de deuxième espèce est minimum, quelle que soit l'hypothèse \mathcal{H} vérifiée. Les deux exemples précédents constituent une illustration de ce cas, la région d'acceptation étant, comme nous l'avons souligné, indépendante de l'hypothèse \mathcal{H}_1 . Pas tout à fait cependant.

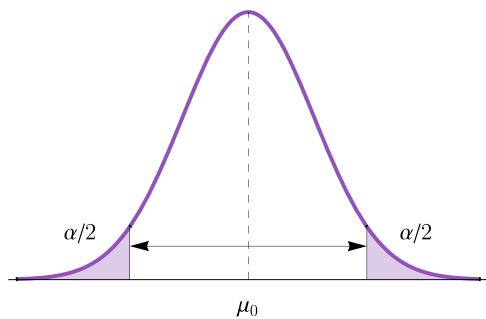
Notons, en effet, que nous avons supposé, respectivement dans chacun des deux exemples, que $\varpi_1 > \varpi_0$ et que $\mu_1 > \mu_0$. Et nous avons abouti alors à des régions d'acceptation de la forme $k < k_s$ et $m < m_s$ telles que le risque α soit bloqué à l'une des extrémités de la distribution de la variable étudiée.



Si donc il s'agit de comparer deux hypothèses de la forme : $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta > \theta_0$, on est conduit à ce qu'on appelle un test à droite, où le risque de première espèce est bloqué à droite.

Le test d'hypothèses de la forme $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta < \theta_0$, conduit à un test appelé test à gauche.

Dans le cas, enfin, d'hypothèses de la forme $\mathcal{H}_0 : \theta = \theta_0$ et $\mathcal{H}_1 : \theta \neq \theta_0$, il apparaît logique de répartir le risque α aux deux extrémités de la distribution. Le test est alors un test symétrique.



2 Tests usuels de comparaison à un standard

2.1 Rappel des lois outils usuelles

La détermination des régions d'acceptation nécessite la mise en oeuvre des lois de probabilité caractéristiques des échantillons prélevés dans des populations de référence spécifiées. D'où l'extrême importance d'une connaissance précise des lois de probabilité usuelles définies dans le chapitre précédent, mais que nous allons reprendre ici.

2.1.1 Loi normale centrée réduite

Etant donnée une variable X qui suit une loi normale de moyenne μ et d'écart-type σ , alors :

$$U = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Comme la variable $M = \frac{\sum_{i=1}^n X_i}{n}$, moyenne d'un échantillon de taille n prélevé dans une population $\mathcal{N}(\mu, \sigma^2)$ suit une loi normale de moyenne μ et d'écart-type $\frac{\sigma}{\sqrt{n}}$. Il en résulte que :

$$\frac{M-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

2.1.2 Loi du χ^2

Etant données ν variables U_1, U_2, \dots, U_ν indépendantes et suivant des lois normales centrées réduites, on sait que, par définition d'une loi du χ^2 :

$$Z_\nu = U_1^2 + U_2^2 + \dots + U_\nu^2 \sim \chi^2(\nu)$$

Il en résulte qu'étant donné un échantillon $(X_1, \dots, X_i, \dots, X_n)$, prélevé dans une population normale $\mathcal{N}(\mu, \sigma^2)$, on a :

$$Z = \sum_{i=1}^n \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$$

Appelant $S^2 = \frac{\sum_{i=1}^n (X_i-M)^2}{n}$ la variance de l'échantillon, alors :

$$Z = \frac{\sum_{i=1}^n (X_i-M)^2}{\sigma^2} = \frac{n S^2}{\sigma^2} \sim \chi^2(n-1)$$

2.1.3 Loi de Student

Etant données $(\nu+1)$ variables normales, centrées, réduites et indépendantes, notées U et U_i , on sait que :

$$T = \frac{U}{\sqrt{\frac{\sum_{i=1}^\nu U_i^2}{\nu}}} \sim \mathcal{T}(\nu)$$

Il en résulte qu'étant données M et S^2 la moyenne et la variance d'un échantillon de taille n prélevé dans une population $\mathcal{N}(\mu, \sigma^2)$:

$$T = \frac{\frac{M-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{nS^2}{\sigma^2}/(n-1)}} = \frac{M-\mu}{\sigma} \times \frac{\sqrt{n}}{\sqrt{n}} \times \frac{\sigma}{S/\sqrt{n-1}} = \frac{M-\mu}{S/\sqrt{n-1}} \sim \mathcal{T}(n-1)$$

On notera qu'une réalisation de cette variable fait apparaître l'estimation de la variance σ^2 :

$$t = \frac{m-\mu}{s/\sqrt{n-1}} = \frac{m-\mu}{\sigma^*/\sqrt{n}}$$

à rapprocher, dans sa forme, de l'expression donnée en section 2.1.1.

2.2 Comparaison de la moyenne d'une population normale de variance σ^2 connue à une valeur donnée μ_0

Nous allons procéder en 4 étapes.

1) Faisons l'hypothèse que la moyenne de la population est égale à μ_0 :

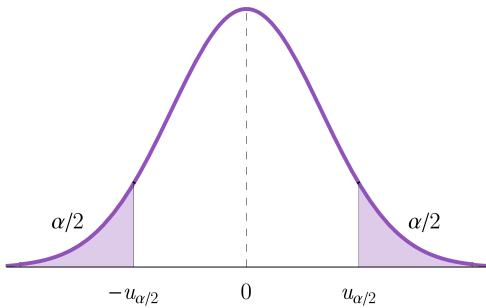
$\mathcal{H}_0 : \mu = \mu_0$, l'hypothèse alternative étant :

$\mathcal{H}_1 : \mu \neq \mu_0$.

2) Il en résulte que la moyenne M d'un échantillon de taille n suit une loi normale de moyenne μ_0 et de variance $\frac{\sigma^2}{n}$ et que, par conséquent :

$$U = \frac{M-\mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

3) Fixons nous un risque α que nous conviendrons de considérer comme négligeable. Il en résulte un certain intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$ dans lequel la variable U a une probabilité $(1 - \alpha)$ de tomber si l'hypothèse est exacte et, par conséquent, hors duquel U a une probabilité α petite de tomber. Négliger cette probabilité α , c'est considérer qu'il est impossible de trouver U en dehors de l'intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$, si l'hypothèse est vraie.



4) On calcule à partir des données de l'échantillon effectivement obtenu (x_1, \dots, x_n) la valeur u de U et on la situe par rapport à l'intervalle $[-u_{\alpha/2}, u_{\alpha/2}]$. On conclut alors de la façon suivante :

- si u tombe à l'extérieur de l'intervalle, on préfère rejeter l'hypothèse, en sachant toutefois qu'on assume le risque α de la rejeter à tort.

- si u tombe à l'intérieur de l'intervalle, cela ne signifie nullement, hélas, que l'hypothèse faite est vraie, mais seulement que les données recueillies *ne sont pas en contradiction avec cette hypothèse*. Autrement dit, on est dans l'incapacité de conclure ni en faveur, ni en défaveur de l'hypothèse. On verra que dans les applications pratiques, cela est généralement moins gênant qu'il n'y paraît, parce que c'est contre un rejet, fait à tort, de l'hypothèse qu'il faut se prémunir, la conservation de l'hypothèse correspondant au statu quo.

2.3 Comparaison de la variance d'une population normale à une valeur donnée σ_0^2

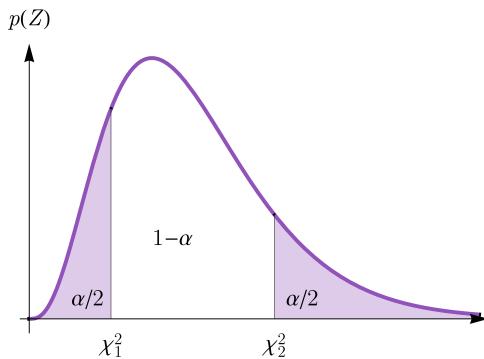
Faisant l'hypothèse :

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2$$

Alors :

$$Z = \frac{nS^2}{\sigma_0^2} = \frac{\sum_{i=1}^n (X_i - M)^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Il en résulte que, si l'hypothèse est vraie, $\frac{nS^2}{\sigma_0^2}$ a la probabilité $(1-\alpha)$ de tomber dans l'intervalle $[\chi_1^2, \chi_2^2]$ où χ_1^2 et χ_2^2 sont lus dans la table de la loi du χ^2 à $(n-1)$ degrés de liberté. Il suffit alors, comme précédemment, de calculer la valeur $\frac{nS^2}{\sigma_0^2}$ à partir des observations, de la placer par rapport à l'intervalle $[\chi_1^2, \chi_2^2]$ et enfin de conclure.



2.4 Comparaison de la moyenne d'une population normale (de variance inconnue) à une valeur donnée μ_0

Faisant l'hypothèse :

$$\mathcal{H}_0 : \mu = \mu_0,$$

Alors :

$$T = \frac{M - \mu_0}{S/\sqrt{n-1}} \sim \mathcal{T}(n-1)$$

Le test revient à placer la quantité :

$$t = \frac{m - \mu_0}{s/\sqrt{n-1}} \text{ ou } t = \frac{m - \mu_0}{\sigma^*/\sqrt{n}} \text{ (avec } \sigma^*{}^2 = \frac{nS^2}{n-1})$$

par rapport à l'intervalle $[-t_{\alpha/2}, t_{\alpha/2}]$ lu dans la table de Student à $(n-1)$ degrés de liberté.

2.5 Test des appariements

Nous avons présenté, dans l'introduction du chapitre, le dispositif expérimental qui consiste, disposant de n parcelles, à diviser chacune de ces parcelles en deux, et à cultiver chaque parcelle en soumettant l'une des moitiés à un certain traitement et l'autre moitié à un autre traitement. A chaque parcelle correspondront, en fin de culture, deux rendements *appariés*.

Imaginons un autre exemple, dans lequel on veuille confronter deux appareils de mesure et que, pour ce faire, on utilise n supports en procédant, sur chacun d'eux, à deux mesures à l'aide des deux appareils soumis à examen. Les deux mesures seront dites *appariées* et les résultats obtenus se présenteront, en définitive, comme suit :

mesures 1 : $x_1, x_2, \dots, x_i, \dots, x_n$
 mesures 2 : $y_1, y_2, \dots, y_i, \dots, y_n$

Soit d_i la différence $d_i = (y_i - x_i)$ et soient m_d et σ_d^* la moyenne et l'écart-type estimé des différences. On admet que les d_i sont des réalisations d'une variable D qui suit une loi normale. Le test de l'hypothèse $\mathcal{H}_0 : \mathbb{E}(D) = 0$ (pas d'influence du traitement ou pas de différence entre les appareils de mesures) est le test présenté au paragraphe précédent avec $\mu_0 = 0$.

3 Comparaison sur échantillons de deux populations normales

3.1 Comparaison des variances de deux populations normales

La comparaison de deux populations normales revient à se demander si elles ont *même moyenne* et *même variance* puisque ces deux paramètres suffisent à déterminer entièrement une distribution normale. Pour des raisons théoriques qui apparaîtront dans un paragraphe suivant, la comparaison des variances doit précéder celle des moyennes.

Soient n_1 et s_1^2 la taille et la variance de l'échantillon extrait de la première population, et soient n_2 et s_2^2 la taille et la variance de l'échantillon extrait de la deuxième population. Nous savons que les estimations sans biais des variances σ_1^2 et σ_2^2 des deux populations s'écrivent :

$$\sigma_1^{*2} = \frac{n_1 s_1^2}{n_1 - 1} \text{ et } \sigma_2^{*2} = \frac{n_2 s_2^2}{n_2 - 1}$$

Dans l'hypothèse d'égalité des variances des deux populations ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), ces deux estimations ne diffèrent qu'en raison des aléas de l'échantillonnage. Il en est de même de leur quotient $f = \frac{\sigma_1^{*2}}{\sigma_2^{*2}}$ qui ne diffère de 1 qu'à cause des aléas de l'échantillonnage.

Le statisticien Ronald Aylmer Fisher, biologiste et mathématicien britannique, auteur du test classique que nous allons présenter, a retenu cette forme et calculé la loi de probabilité de la variable :

$$F(\nu_1, \nu_2) = \frac{Z_1/\nu_1}{Z_2/\nu_2}$$

où Z_1 et Z_2 sont deux variables aléatoires indépendantes qui suivent des lois du χ^2 à respectivement ν_1 et ν_2 degrés de liberté.

Dans l'hypothèse d'égalité des variances des deux populations, si l'on désigne par S_1^2 et S_2^2 les variables, dont les variances des échantillons qui en sont extraits au hasard, sont des réalisations, $\frac{n_1 S_1^2}{\sigma^2}$ et $\frac{n_2 S_2^2}{\sigma^2}$ sont indépendantes et suivent des lois du χ^2 à respectivement $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté. Il en résulte, par définition de cette variable, que le quotient :

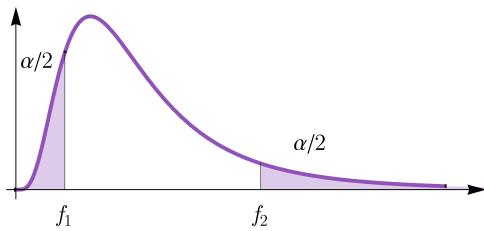
$$F = \frac{\frac{n_1 S_1^2}{\sigma^2}/(n_1 - 1)}{\frac{n_2 S_2^2}{\sigma^2}/(n_2 - 1)} = \frac{n_1 S_1^2/(n_1 - 1)}{n_2 S_2^2/(n_2 - 1)} \quad (\text{après simplification par } \sigma^2)$$

suit une loi de Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté, ce que l'on notera ainsi $F \sim \mathcal{F}(n_1 - 1, n_2 - 1)$. Par conséquent, si l'hypothèse d'égalité des variances est vérifiée, la quantité :

$$f = \frac{\sigma_1^{*2}}{\sigma_2^{*2}}$$

est une réalisation d'une telle loi de Snedecor.

Cette loi définie, la suite des opérations est maintenant bien connue. Se fixant un seuil de probabilité α négligeable, on lit dans la table de Snedecor à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté les valeurs f_1 et f_2 correspondant au dessin ci-dessous.



Telles qu'elles sont présentées, les tables de la loi de Snedecor portent, en tête de colonnes, le nombre de degrés de liberté du numérateur v_1 et, en tête de lignes, celui du dénominateur v_2 . Elles fournissent, à l'intersection de la colonne v_1 et de la ligne v_2 , la limite supérieure f_2 de l'intervalle d'acceptation. Elles fournissent donc, à l'intersection de la colonne v_2 et de la ligne v_1 , la valeur $1/f_1$ de l'intervalle d'acceptation.

Plus souvent appelé test de Fisher, ce test prend parfois l'appellation de test de Fisher-Snedecor ou test de Snedecor ou encore de F-test.

3.2 Estimation de σ^2

En admettant que le résultat du test précédent ne s'oppose pas à l'hypothèse d'égalité des variances, il peut être utile d'estimer la valeur commune σ^2 des variances des deux populations.

Puisque, dans l'hypothèse d'égalité des variances, $\frac{n_1 S_1^2}{\sigma^2}$ et $\frac{n_2 S_2^2}{\sigma^2}$ sont des variables indépendantes qui suivent des lois du χ^2 , respectivement à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté leur somme $\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ suit une loi du χ^2 à $(n_1 + n_2 - 2)$ degrés de liberté, dont la moyenne et la variance sont respectivement $(n_1 + n_2 - 2)$ et $2(n_1 + n_2 - 2)$.

Il en résulte que la variable $\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$ est un estimateur sans biais et convergent de σ^2 , puisque :

$$\mathbb{E}\left[\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\right] = \sigma^2 \text{ et } \mathbb{V}\left[\frac{n_1 S_1^2 + n_2 S_2^2}{(n_1 + n_2 - 2)}\right] = \frac{2\sigma^4}{n_1 + n_2 - 2} \rightarrow 0.$$

Par conséquent, la quantité :

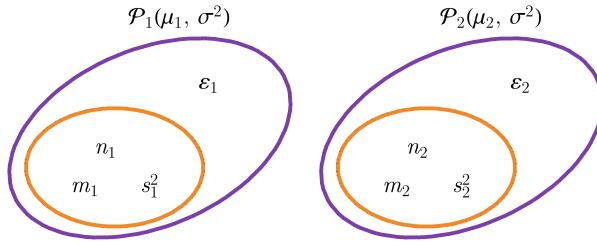
$$\sigma^{*2} = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

calculée à partir des observations, est une estimation sans biais de σ^2 , la variance commune aux deux populations.

3.3 Comparaison des moyennes de deux populations normales

Dans l'hypothèse de populations normales, une fois testée l'égalité des variances, il suffit de tester l'égalité des moyennes pour pouvoir considérer que les populations sont identiques. Les raisons théoriques qui conduisent à présenter la comparaison des variances avant celle des moyennes peuvent, à ce stade, être explicitées. En effet, le test de comparaison des variances ne faisait aucune hypothèse sur l'égalité des moyennes. Par contre, le test d'égalité des moyennes implique l'égalité des variances. Il est donc nécessaire de vérifier cette égalité avant de s'intéresser aux moyennes.

Cela étant, soient deux populations normales \mathcal{P}_1 et \mathcal{P}_2 de moyennes μ_1 et μ_2 , mais de même variance σ^2 . Soient n_1 et n_2 les tailles de deux échantillons \mathcal{E}_1 et \mathcal{E}_2 prélevés au hasard respectivement dans chacune de ces deux populations ; soient m_1 et m_2 leurs moyennes, et soient s_1^2 et s_2^2 leurs variances.



Dans ces conditions, il est permis de considérer que :

- m_1 est une réalisation d'une variable M_1 normale de moyenne μ_1 et de variance σ^2 / n_1 ,
- m_2 est une réalisation d'une variable M_2 normale de moyenne μ_2 et de variance σ^2 / n_2 ,
- s_1^2 et s_2^2 sont des réalisations des variables S_1^2 et S_2^2 telles que la variable $\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ suit une loi du χ^2 à $(n_1 + n_2 - 2)$ degrés de liberté et est indépendante de M_1 et M_2 .

Faisons maintenant l'hypothèse que $\mu_1 = \mu_2 = \mu$. Il en résulte que la variable $(M_1 - M_2)$ suit une loi normale de *moyenne nulle* et de variance égale à la somme des variances de M_1 et M_2 , c'est-à-dire à $\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$. Par conséquent :

$$U = \frac{M_1 - M_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

Pour éliminer la quantité σ inconnue, il suffit de considérer le quotient :

$$T = \frac{M_1 - M_2}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} / \sqrt{\frac{\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}}{n_1 + n_2 - 2}} = \frac{M_1 - M_2}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \mathcal{T}(n_1 + n_2 - 2)$$

Pour simplifier l'écriture, on peut tenir compte de ce que figure, au dénominateur, l'expression de l'estimateur sans biais de σ^2 . Par conséquent :

$$t = \frac{m_1 - m_2}{\sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

est une réalisation d'une loi de Student qu'il suffit, pour conclure, de placer par rapport à l'intervalle $[-t_{\alpha/2}, t_{\alpha/2}]$ correspondant au risque α choisi. Si t n'appartient pas à l'intervalle, on dit souvent que la différence entre les moyennes observées est *significative* au risque α sinon, qu'elle n'est *pas significative*.

3.4 Estimation de la différence des moyennes des populations

Si la différence observée entre les moyennes m_1 et m_2 des échantillons est *significative* (d'une différence entre les moyennes μ_1 et μ_2 des populations), il peut s'avérer utile d'estimer la différence $\Delta = \mu_1 - \mu_2$. La variable $(M_1 - M_2)$ est évidemment un estimateur sans biais de Δ . Quant à la détermination de l'intervalle de confiance, elle repose sur la prise en compte de la variable :

$$T = \frac{(M_1 - M_2) - \Delta}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim \mathcal{T}(n_1 + n_2 - 2)$$

On a, par conséquent, au risque α près :

$$(m_1 - m_2) - t_{\alpha/2} \sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \Delta \leq (m_1 - m_2) + t_{\alpha/2} \sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Exercices du chapitre 5

Exercice 1

On a prélevé, au hasard dans une population normale de moyenne μ et d'écart-type σ , un échantillon de taille $n = 10$. La moyenne et la variance calculées sur cet échantillon sont respectivement $m = 4$ et $s^2 = 6$.

- Calculer une estimation sans biais de σ et son intervalle de confiance au risque 5%.
- Tester l'hypothèse $\sigma = 2$ au risque 5%.
- En admettant σ connu égal à 2, tester l'hypothèse $\mu = 3$ au risque 5%.
- Tester, au risque 5%, l'hypothèse $\mu = 3$ sans faire aucune hypothèse sur la valeur de σ .
- Calculer une estimation sans biais de μ et son intervalle de confiance au risque 5% sans faire aucune hypothèse sur la valeur de σ .
- En admettant μ connu égal à 3, est-il possible d'envisager un test plus efficace que celui mis en oeuvre en b) pour tester l'hypothèse $\sigma = 2$?

Exercice 2

Pour comparer les rendements de deux variétés de blé A et B, on a ensemencé 10 couples de deux parcelles voisines, l'une en variété A, l'autre en variété B, les 10 couples étant répartis dans des localités différentes. On a obtenu les résultats suivants :

couple n°	1	2	3	4	5	6	7	8	9	10
récolte A	45	32	56	49	45	38	47	51	42	38
récolte B	47	34	52	51	48	44	45	56	46	44

Que peut-on conclure de ces résultats ?

Exercice 3

On donne ci-après les pourcentages de matière grasse dans un aliment, déterminés sur 10 échantillons par deux méthodes d'analyse différentes A et B.

échantillon n°	1	2	3	4	5	6	7	8	9	10
méthode A	24.6	25.3	25.3	25.6	25.6	25.9	26	27	27.3	27.7
méthode B	24.9	25.6	25.8	26.2	26.1	26.7	26.3	26.9	28.4	27.1

Comparer ces deux méthodes.

Exercice 4

On a prélevé au hasard un échantillon \mathcal{E}_1 de taille $n_1 = 10$ dans une population normale \mathcal{P}_1 de moyenne μ_1 et d'écart-type σ_1 . La moyenne et la variance calculées sur cet échantillon sont $m_1 = 4$ et $s_1^2 = 6$.

On préleve de même, dans une autre population normale \mathcal{P}_2 de moyenne μ_2 et d'écart-type σ_2 , un échantillon \mathcal{E}_2 de taille $n_2 = 15$, avec $m_2 = 7$ et $s_2^2 = 20$.

- Tester l'hypothèse $\sigma_2 = \sigma_1$, au risque 5%.
- Tester l'hypothèse $\sigma_2 = 2\sigma_1$, au risque 5%.
- En admettant que $\sigma_2 = 2\sigma_1$, calculer une estimation sans biais de σ_1 , à partir des deux échantillons, et son intervalle de confiance au risque 5%.
- Utiliser un test du χ^2 pour tester simultanément les hypothèses $\sigma_2 = 4$ et $\sigma_1 = 2$.

- e) En admettant que $\sigma_2 = 2\sigma_1 = 4$, tester, au risque 5%, l'hypothèse $\mu_2 = 2\mu_1$.
- f) Calculer une estimation de μ_1 à partir des deux échantillons, en admettant que $\mu_2 = 2\mu_1$ et son intervalle de confiance au risque 5%.

Exercice 5

Il y a des raisons de penser que l'épaisseur de la cire dont sont enduits des sacs en papier est plus irrégulière à l'extérieur qu'à l'intérieur. Pour le vérifier 75 mesures de l'épaisseur ont été faites et ont donné les résultats suivants :

- surface intérieure : $\sum x = 71.25$ et $\sum x^2 = 91$
- surface extérieure : $\sum y = 48.75$ et $\sum y^2 = 84$.

- a) Faire un test pour déterminer, au risque 5%, si la variabilité de l'épaisseur de la cire est plus grande à l'extérieur qu'à l'intérieur des sacs.
- b) Revenant à la loi de F , calculer l'intervalle de confiance à 95% du rapport des variances.

Exercice 6

Deux chaînes de fabrication produisent des transistors. Des relevés effectués pendant 10 jours ont donné les résultats suivants :

- ligne 1 : $m_x = 2800$ et $\sum(x - m_x)^2 = 103\,600$
- ligne 2 : $m_y = 2680$ et $\sum(y - m_y)^2 = 76\,400$

On admettra que les écarts-type σ_x et σ_y sont inconnus mais égaux.

- a) Peut-on conclure, au risque de 5%, à une différence entre les productions moyennes des deux lignes ?
- b) Quel est l'intervalle de confiance à 95% de la différence ?

Exercice 7

Le lancement d'un nouveau produit nécessitant une réorganisation complète d'un atelier (donc des dépenses importantes), une entreprise décide de faire une étude de marché sous la forme d'un questionnaire dont on retient en particulier l'information suivante : « la personne interrogée est ou n'est pas intéressée par le nouveau produit ». Soit p la proportion (réelle mais inconnue) des personnes intéressées par le nouveau produit. L'entreprise juge qu'il est raisonnable de lancer ce nouveau produit si plus de 50% des personnes sont réellement intéressées. A partir d'une enquête, elle décide donc de faire le test $\mathcal{H}_0 : p \leq 0.5$ contre $\mathcal{H}_1 : p > 0.5$

- 1) Que représentent les risques de 1ère et 2e espèce associés à ce test ?
- 2) Sur 100 personnes interrogées, combien doivent répondre qu'elles sont intéressées pour que l'entreprise se décide à investir pour le lancement d'un nouveau produit avec moins de 1 chance sur 100 de se tromper ?
- 3) Sur 100 personnes interrogées, 58 ont finalement répondu être intéressées par le nouveau produit. Quelle conclusion en tirer ?

Faits et modèles

Nous avons envisagé certaines lois de probabilité susceptibles de constituer des modèles pour les populations de références. Il s'agit maintenant, en présence d'observations, de choisir le modèle adapté et de vérifier que les observations disponibles s'y raccordent bien.

1 Distributions statistiques

1.1 Mise en ordre des observations

Ayant effectué des observations sur les n individus constituant un échantillon, la mise en ordre consiste à grouper ensemble les résultats identiques, c'est-à-dire à faire correspondre, aux valeurs observées de la variable prise en considération, les nombres d'individus ayant présenté ces valeurs. Le tableau obtenu définit ce qu'on appelle une *distribution statistique*.

Dans le cas d'une variable susceptible de prendre les valeurs discrètes $x_1, \dots, x_k, \dots, x_r$, les résultats se présentent sous la forme du tableau ci-dessous.

Valeurs	Effectifs
x_1	n_1
\vdots	\vdots
x_k	n_k
\vdots	\vdots
x_r	n_r

Lorsque la variable est continue, il est commode de procéder à des groupages en classes. Cela consiste à diviser l'intervalle de variation de la variable en classes :

$$\left] x_1 - \frac{h}{2}, x_1 + \frac{h}{2} \right], \left] x_1 + \frac{h}{2}, x_2 + \frac{h}{2} \right], \dots, \left] x_{r-1} - \frac{h}{2}, x_r + \frac{h}{2} \right]$$

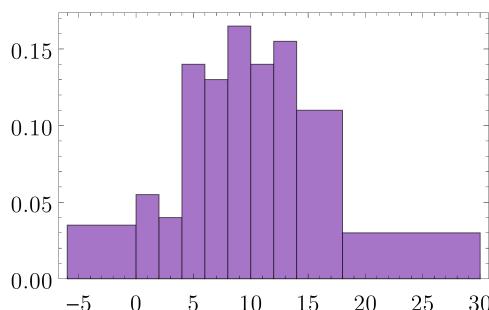
puis à grouper ensemble les valeurs observées qui tombent dans une même classe. Une telle opération fait perdre de l'information mais on peut montrer toutefois que la perte d'information est négligeable si l'on choisit l'intervalle de classe de façon à obtenir 10 à 15 classes.

1.2 Représentations graphiques des distributions

Nous allons décrire les trois représentations graphiques les plus utilisées, dans le cas d'une variable continue. La transposition au cas d'une variable discrète ne pose pas de problèmes.

1.2.1 Histogramme

Ayant gradué l'axe des abscisses suivant les intervalles retenus, on construit sur chaque intervalle un rectangle de *surface* proportionnelle à la fréquence (absolue ou relative) des observations qui lui correspondent.

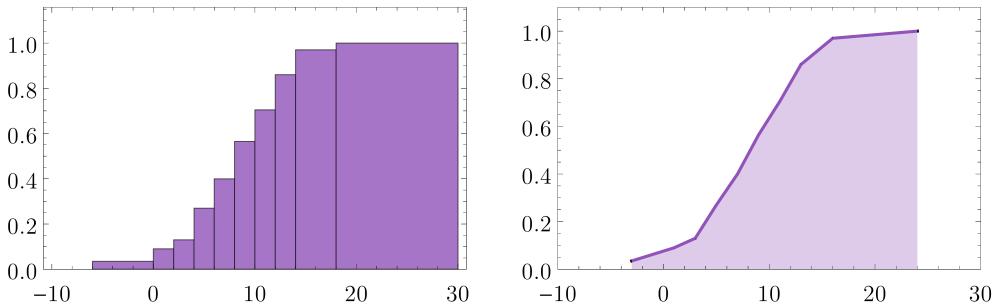


La surface est préférée à la hauteur pour éliminer l'influence de l'inégalité éventuelle des classes.

1.2.2 Diagramme des fréquences cumulées

C'est un graphe en escalier (ci-dessous à gauche) qui fait correspondre à chaque observation x , en abscisse, la fréquence, en ordonnée, des observations inférieures ou égales à x . On envisage le plus souvent les fréquences relatives, les fréquences cumulées sont alors comprises entre 0 et 1.

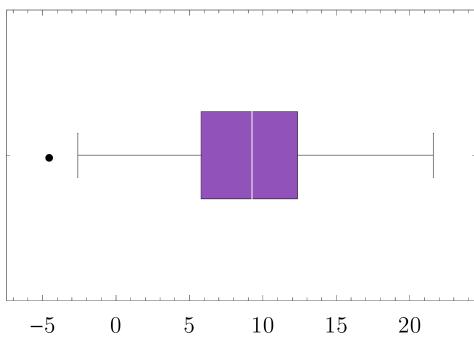
Une autre façon, plus simple, de le construire est de définir des classes, de faire correspondre à chaque frontière supérieure la fréquence cumulée correspondante et de joindre les points par des segments de droite (ci-dessous à droite). Cela revient à admettre une distribution uniforme à l'intérieur des classes.



1.2.3 Boîte à moustaches (box and whiskers plot)

Cette représentation a été imaginée par John Tukey en 1977, dans le même courant d'idées que celui des statistiques *robustes*. Elle consiste en une « boîte » dont les côtés verticaux correspondent aux *quartiles* de la distribution et le segment interne à la *médiane*.

De part et d'autre de la boîte, on définit deux « moustaches » de longueur au plus égale à 1,5 fois l'*étendue interquartile*. Quand une observation dépasse la longueur maximale des moustaches, elle est considérée comme « aberrante » et individualisée. S'il n'y a pas d'observation aberrante, la moustache s'arrête à l'observation immédiatement supérieure (ou inférieure). Pour certains auteurs, les moustaches s'étendent jusqu'aux valeurs extrêmes même si celles-ci dépassent l'intervalle défini plus haut ou alors jusqu'aux premiers et neuvièmes déciles ou encore au 5^e et 95^e centiles.



1.3 Caractéristiques des distributions

Rappelons que les caractéristiques essentielles sont :

- pour la tendance centrale, la moyenne m ,
- pour la dispersion, la variance s^2 .

Appelant f_k la fréquence relative $\frac{n_k}{n}$ pour la valeur x_k (ou pour la classe de centre x_k), on peut écrire que $m = \sum f_k x_k$ et $s^2 = \sum f_k (x_k - m)^2$.

S'il n'y a pas eu regroupement en classes, la fréquence de chaque valeur x_i est $\frac{1}{n}$ et on retrouve les formules habituelles $m = \frac{1}{n} \sum x_i$ et $s^2 = \frac{1}{n} \sum (x_i - m)^2$.

Il convient de noter la parenté évidente entre concepts statistiques et concepts probabilistes :

- à la notion de fréquence f_k , pour une distribution statistique, correspond celle de probabilité p_k , pour une loi de probabilité.
- à la notion de moyenne d'une distribution $m = \sum f_k x_k$ correspond la notion d'espérance mathématique $\mathbb{E}(X) = \sum p_k x_k$.
- enfin, à la notion de variance d'une distribution $s^2 = \sum f_k(x_k - m)^2$ correspond celle de variance d'une variable aléatoire $\mathbb{V}(X) = \sum p_k(x_k - \mu)^2$.

2 Fréquences et probabilités

2.1 Retour sur la loi des grands nombres

Considérons un ensemble de possibilités E et deux événements *complémentaires* A et \bar{A} . A est, par exemple, l'évènement : la variable X prend sa valeur dans un certain intervalle $[x, x']$. Soit ϖ la probabilité de A ; celle de \bar{A} est donc égale à $(1 - \varpi)$.

Faisons successivement n épreuves (expériences) identiques, et supposons qu'au cours de ces n épreuves, A se produise k fois et \bar{A} , par conséquent, $(n - k)$ fois. La fréquence relative de A est $f_n = \frac{k}{n}$. L'expérience montre que, si n est assez grand, f_n est voisin de ϖ . C'est la fameuse loi des grands nombres due à Bernoulli en 1713, et qui jette un pont entre fréquences et probabilités

Nous avons démontré, au chapitre 4 en partant de l'inégalité de Bienaymé-Tchebychev, qu'étant donnée une suite de variables aléatoires indépendantes et suivant la même loi de probabilité, leur moyenne convergeait *en probabilité* vers la moyenne de la loi. Appliquée à la moyenne de n variables de Bernoulli X_1, \dots, X_n , cela va permettre de démontrer ce qui précède.

En effet, soit $F_n = \frac{X_1 + \dots + X_n}{n}$ cette moyenne. Elle a pour espérance ϖ et pour variance $\frac{\varpi(1-\varpi)}{n}$ et l'inégalité de Bienaymé-Tchebychev permet d'écrire que :

$$\mathbb{P}(|F_n - \varpi| > \varepsilon) \leq \frac{\varpi(1-\varpi)}{n \varepsilon^2}$$

Ce résultat s'énonce ainsi : ε étant un nombre positif arbitraire, aussi petit que l'on veut, la probabilité pour que la fréquence relative F_n s'écarte de la probabilité ϖ d'une quantité supérieure à ε , tend vers 0 lorsque le nombre d'épreuves augmente indéfiniment.

D'un point de vue pratique, cela exprime qu'en faisant un nombre suffisant d'épreuves, il est possible d'avoir une « idée » aussi précise qu'on le veut de la probabilité ϖ qu'on ne connaît pas. Il suffit, par exemple, de faire un nombre assez grand de tirages dans une urne de composition inconnue (proportion ϖ de boules noires) pour que la fréquence observée des boules noires soit *presque sûrement* très voisine de la proportion ϖ . La loi des grands nombres constitue, à ce titre, la base de la statistique mathématique.

2.2 Nombre de mesures à effectuer pour une précision donnée

L'inégalité de Bienaymé-Tchebychev majore beaucoup la probabilité cherchée. Il en résulte que la valeur de n qu'il faut dépasser pour que cette probabilité n'excède pas un seuil fixé, est inutilement grande. Souhaitant, par exemple, estimer une proportion (ou une probabilité) ϖ voisine de 0.2 avec une précision égale à ± 0.01 et un risque de 5%, l'application de l'inégalité de Bienaymé-Tchebychev conduit à un nombre n très grand puisque :

$$0.05 = \mathbb{P}(|F_n - \varpi| > 0.01) \leq \frac{0.2 \times 0.8}{n(0.01)^2} \implies n \geq 32\,000$$

Il est préférable d'utiliser le théorème central limite qui établit que, si n est suffisamment grand, la fréquence relative obéit à une loi qui s'approche d'une loi normale de moyenne ϖ et de variance $\frac{\varpi(1-\varpi)}{n}$. D'où il résulte que :

$$U = \frac{F_n - \varpi}{\sqrt{\varpi(1-\varpi)/n}} \sim \mathcal{N}(0, 1)$$

Il s'ensuit que :

$$\mathbb{P}(|F_n - \varpi| > \varepsilon) \simeq 2 \mathbb{P}\left(U > \frac{\varepsilon}{\sqrt{\varpi(1-\varpi)/n}}\right)$$

qui conduit, avec les mêmes données que ci-dessus, à :

$$0.05 \simeq 2 \mathbb{P}\left(U > \frac{0.01}{\sqrt{0.2 \times 0.8/n}}\right)$$

et, après lecture dans la table de la loi normale réduite, à $n \geq 6146$.

2.3 Estimation d'une proportion et intervalle de confiance

Puisque $\mathbb{E}(F_n) = \varpi$ et que $\mathbb{V}(F_n) = \frac{\varpi(1-\varpi)}{n} \rightarrow 0$, F_n est un estimateur sans biais de ϖ .

D'autre part, l'approximation de la loi de F_n par une loi normale permet d'écrire, au risque α près, que :

$$|f_n - \varpi| \leq u_{\alpha/2} \sqrt{\frac{\varpi(1-\varpi)}{n}}, \text{ où } u_{\alpha/2} \text{ est lu dans une table de la loi normale centrée réduite.}$$

Si n est suffisamment grand, on peut approximer le deuxième membre de l'inégalité en remplaçant ϖ par son estimation f_n . D'où l'intervalle de confiance, au risque α :

$$f_n - u_{\alpha/2} \sqrt{\frac{f_n(1-f_n)}{n}} < \varpi < f_n + u_{\alpha/2} \sqrt{\frac{f_n(1-f_n)}{n}}$$

2.4 Comparaison de deux proportions

Soit ϖ_1 et ϖ_2 les proportions caractérisant deux populations et soit f_1 et f_2 les fréquences observées sur deux échantillons, de tailles respectives n_1 et n_2 , prélevés au hasard dans chacune de ces populations. Faisant l'hypothèse que $\varpi_1 = \varpi_2 = \varpi$, une démarche calquée sur celle mise en oeuvre pour la comparaison de deux moyennes permet d'établir que, en estimant ϖ par la quantité $\varpi^* = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$, le quotient :

$$u = \frac{f_1 - f_2}{\sqrt{\varpi^*(1-\varpi^*)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

est approximativement une réalisation d'une variable normale centrée réduite, si l'hypothèse est vraie. Pour la tester, il suffit alors de placer u par rapport à l'intervalle correspondant au risque choisi.

2.5 Métrique du χ^2

Soit $p(x)$ la densité de probabilité d'une variable aléatoire X . La probabilité pour que X prenne une valeur dans l'intervalle $]x_k - \frac{h}{2}, x_k + \frac{h}{2}]$ est égale à $\varpi_k = \int_{x_k - \frac{h}{2}}^{x_k + \frac{h}{2}} p(x) dx$, et la probabilité pour que X prenne une valeur en dehors de cet intervalle est $(1 - \varpi_k)$. Pour un échantillon de n observations de la variable X , le nombre de valeurs n_k tombant dans l'intervalle $]x_k - \frac{h}{2}, x_k + \frac{h}{2}]$ est une réalisation d'une variable N_k qui suit une loi binomiale de moyenne $n \varpi_k$ et de variance $n \varpi_k(1 - \varpi_k)$, et lorsque n augmente N_k converge en probabilité vers $n \varpi_k$.

Pour tenir compte de toutes les classes, soit $n_1, \dots, n_k, \dots, n_r$ les *effectifs observés* dans les classes et soit $\varpi_1, \dots, \varpi_k, \dots, \varpi_r$, les probabilités théoriques correspondant à la loi de référence. On définit ce qu'on appelle les *effectifs théoriques* dans les classes, qui sont les quantités $n\varpi_1, \dots, n\varpi_k, \dots, n\varpi_r$. Notons qu'effectifs observés et théoriques ont même somme n .

Effectifs observés	Effectifs théoriques
n_1	$n\varpi_1$
\vdots	\vdots
n_k	$n\varpi_k$
\vdots	\vdots
n_r	$n\varpi_r$
n	n

Pour mesurer la *distance* entre distribution observée des n_k et distribution théorique des $n\varpi_k$, on peut envisager plusieurs quantités.

L'une d'elle est la distance de Kolmogorov qui est la quantité :

$$D = \sup \{|n_k - n\varpi_k|\}$$

Mais on retient le plus souvent ce qu'on appelle la distance du χ^2 suivante :

$$d = \sum_{k=1}^r \frac{(n_k - n\varpi_k)^2}{n\varpi_k}$$

Nous allons montrer que *si l'échantillon provient bien d'une population définie par la loi de probabilité envisagée*, cette quantité est une réalisation d'une loi du χ^2 à $(r-1)$ degrés de liberté : nombre de classes moins 1.

La démonstration qui suit n'est pas essentielle. Pour simplifier les notations, nous la ferons dans le cas de trois classes, mais elle est facilement transposable au cas général.

Considérons les variables $X_1 = \frac{N_1 - n\varpi_1}{\sqrt{n\varpi_1}}$, $X_2 = \frac{N_2 - n\varpi_2}{\sqrt{n\varpi_2}}$, $X_3 = \frac{N_3 - n\varpi_3}{\sqrt{n\varpi_3}}$. Nous avons :

$$\mathbb{E}(X_i) = \frac{\mathbb{E}(N_i) - n\varpi_i}{\sqrt{n\varpi_i}} = \frac{n\varpi_i - n\varpi_i}{\sqrt{n\varpi_i}} = 0$$

$$\mathbb{V}(X_i) = \frac{\mathbb{V}(N_i)}{n\varpi_i} = \frac{n\varpi_i(1-\varpi_i)}{n\varpi_i} = 1 - \varpi_i$$

On en déduit aisément que :

$$\mathbb{E}(X_i^2) = \mathbb{V}(X_i) + \mathbb{E}(X_i)^2 = 1 - \varpi_i$$

A la limite, quand n augmente, ces variables sont donc des lois normales centrées et de variances respectives $(1 - \varpi_1)$, $(1 - \varpi_2)$, $(1 - \varpi_3)$. Mais ces lois ne sont pas indépendantes puisque $N_1 + N_2 + N_3 = n$. On en déduit que $\sqrt{n\varpi_1} X_1 + n\varpi_1 + \sqrt{n\varpi_2} X_2 + n\varpi_2 + \sqrt{n\varpi_3} X_3 + n\varpi_3 = n$ puis, du fait que $\varpi_1 + \varpi_2 + \varpi_3 = 1$, que :

$$\sqrt{\varpi_1} X_1 + \sqrt{\varpi_2} X_2 + \sqrt{\varpi_3} X_3 = 0.$$

Calculons maintenant les covariances en rappelant que $\text{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$, soit, dans notre cas, que $\text{Cov}(X_i, X_j) = \mathbb{E}(X_i X_j)$. Pour cela, multiplions l'égalité précédente par X_1 , puis X_2 , puis X_3 puis calculons les espérances. On obtient :

$$\sqrt{\varpi_1} \mathbb{E}(X_1^2) + \sqrt{\varpi_2} \mathbb{E}(X_1 X_2) + \sqrt{\varpi_3} \mathbb{E}(X_1 X_3) = 0$$

$$\sqrt{\varpi_1} \mathbb{E}(X_1 X_2) + \sqrt{\varpi_2} \mathbb{E}(X_2^2) + \sqrt{\varpi_3} \mathbb{E}(X_2 X_3) = 0$$

$$\sqrt{\varpi_1} \mathbb{E}(X_1 X_3) + \sqrt{\varpi_2} \mathbb{E}(X_2 X_3) + \sqrt{\varpi_3} \mathbb{E}(X_3^2) = 0$$

Ayant montré précédemment que $\mathbb{E}(X_i^2) = 1 - \varpi_i$, notre système s'écrit alors :

$$\begin{aligned}\sqrt{\varpi_1} (1 - \varpi_1) + \sqrt{\varpi_2} \mathbb{E}(X_1 X_2) + \sqrt{\varpi_3} \mathbb{E}(X_1 X_3) &= 0 \\ \sqrt{\varpi_1} \mathbb{E}(X_1 X_2) + \sqrt{\varpi_2} (1 - \varpi_2) + \sqrt{\varpi_3} \mathbb{E}(X_2 X_3) &= 0 \\ \sqrt{\varpi_1} \mathbb{E}(X_1 X_3) + \sqrt{\varpi_2} \mathbb{E}(X_2 X_3) + \sqrt{\varpi_3} (1 - \varpi_3) &= 0\end{aligned}$$

Après résolution, tenant compte de ce que $\varpi_1 + \varpi_2 + \varpi_3 = 1$, on trouve :

$$\begin{aligned}\mathbb{E}(X_1 X_2) &= -\sqrt{\varpi_1 \varpi_2} \\ \mathbb{E}(X_2 X_3) &= -\sqrt{\varpi_2 \varpi_3} \\ \mathbb{E}(X_1 X_3) &= -\sqrt{\varpi_1 \varpi_3}\end{aligned}$$

La suite de la démonstration m'a été proposée par Rémi Peyre, enseignant-chercheur à Mines Nancy. Nous allons utiliser le concept de vecteur gaussien qui, par définition, est un vecteur dont toute combinaison linéaire des coordonnées est une variable aléatoire gaussienne. Ainsi, d'après ce qui précède le vecteur (X_1, X_2, X_3) , traité comme un vecteur ligne, est un vecteur gaussien centré, de matrice de covariance :

$$M = \begin{pmatrix} 1 - \varpi_1 & -\sqrt{\varpi_1 \varpi_2} & -\sqrt{\varpi_1 \varpi_3} \\ -\sqrt{\varpi_1 \varpi_2} & 1 - \varpi_2 & -\sqrt{\varpi_2 \varpi_3} \\ -\sqrt{\varpi_1 \varpi_3} & -\sqrt{\varpi_2 \varpi_3} & 1 - \varpi_3 \end{pmatrix}$$

Soit O une matrice 3×3 orthogonale (son inverse est égale à sa transposée) dont le premier vecteur-ligne est $(\sqrt{\varpi_1}, \sqrt{\varpi_2}, \sqrt{\varpi_3})$. Une telle matrice existe bien vu que ce vecteur est normé et peut donc être complété en une base orthonormale. Il est alors clair que :

$$M = I_3 - \begin{pmatrix} \sqrt{\varpi_1 \varpi_1} & \sqrt{\varpi_1 \varpi_2} & \sqrt{\varpi_1 \varpi_3} \\ \sqrt{\varpi_1 \varpi_2} & \sqrt{\varpi_2 \varpi_2} & \sqrt{\varpi_2 \varpi_3} \\ \sqrt{\varpi_1 \varpi_3} & \sqrt{\varpi_2 \varpi_3} & \sqrt{\varpi_3 \varpi_3} \end{pmatrix} = O^T I_3 O - O^T \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} O = O^T \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} O$$

Notre but est de trouver la loi de $X_1^2 + X_2^2 + X_3^2$, qui peut aussi s'écrire, en notation matricielle :

$$X X^T = X O^T O X^T = Y Y^T = Y_1^2 + Y_2^2 + Y_3^2, \text{ où } Y = X O^T.$$

Mais puisque X est un vecteur gaussien centré, Y est un vecteur gaussien centré également, et sa matrice de covariance est (en vertu du lemme élémentaire disant que si X est un vecteur gaussien centré n -dimensionnel et P une matrice constante $n \times m$, alors $X P$ est un vecteur gaussien centré, m -dimensionnel, avec $\text{Cov}(X P) = P^T \text{Cov}(X) P$) :

$$\text{Cov}(Y) = \text{Cov}(X O^T) = O \text{Cov}(X) O^T = O O^T \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} O O^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

cela signifie, vu que Y est un vecteur gaussien centré, que Y_1 est nul et que Y_2 et Y_3 sont des vecteurs $\mathcal{N}(0, 1)$ indépendants. Ainsi $X_1^2 + X_2^2 + X_3^2 = Y_2^2 + Y_3^2 \sim \chi^2(2)$, par définition de ce qu'est la loi $\chi^2(2)$.

3 Techniques de raccordement entre distributions statistiques et lois de probabilité

3.1 Loi de référence

Le problème, sous la forme la plus générale, consiste à caractériser à partir des données le type de la loi de référence, puis à préciser cette loi par estimation des paramètres qui la définissent complètement. En pratique, cependant, on n'opère pas exactement ainsi. Les lois de référence s'identifiant le plus souvent aux lois de probabilité fondamentales (loi binomiale, normale, log-normale), il s'avère plus simple :

- de rapprocher la distribution examinée de la loi de probabilité à laquelle il semble intuitivement (ou pour des raisons théoriques) qu'elle doive se raccorder ;
- de vérifier ensuite la validité du rapprochement ainsi opéré.

Lorsque le raccordement à l'une des lois fondamentales s'avère injustifié, il y a lieu de faire appel à d'autres lois de référence, et il en existe un nombre considérable (loi gamma, loi beta, loi de Pareto, loi de Gumbel, loi de Weibull, ...), ou d'en créer éventuellement pour la circonstance.

3.2 Détermination du type de la loi de référence

Il n'y a pas de recette particulière pour déterminer le type de la loi de référence à laquelle on soupçonne la distribution observée de se rattacher. En général, on se laisse guider par des considérations logiques ou bien on tente des rapprochements qui semblent résulter de la forme des distributions observées.

Dans le cas de distributions relatives à des variables discrètes, le raccordement une loi binomiale ou de Poisson s'impose de prime abord.

Dans le cas de variables continues, le raccordement aux lois normale ou log-normale s'avère très souvent, mais pas toujours, légitime. Pour vérifier, avant tout calcul compliqué, que l'hypothèse de tels raccordements n'est pas a priori absurde, on dispose de moyens simples et rapides.

3.2.1 Raccordement à une loi normale

La loi normale est une loi symétrique. De plus, on a vu que l'intervalle $[\mu - u\sigma, \mu + u\sigma]$ comprend approximativement la probabilité : 50% pour $u = 2/3$, 68% pour $u = 1$, 95% pour $u = 2$ et presque 100% pour $u = 3$. Si donc une distribution observée en pratique est telle que les fréquences des observations comprises à l'intérieur de ces intervalles sont voisines de ces probabilités, il y a présomption de normalité.

On peut également vérifier cette présomption à l'aide d'une transformation connue sous l'appellation de *droite de Henry*. Soit $P(x)$ le graphe de la fonction de répartition d'une loi normale ; il a une forme en S. Il existe dans le commerce un papier dit « gausso-arithmétique » qui, par un changement d'échelle de l'axe des ordonnées, permet de réaliser une anamorphose de $P(x)$ qui le transforme en une droite. Dès lors, si l'on trace sur ce papier le diagramme des fréquences cumulées de la distribution observée et que ses points sont sensiblement alignés, on peut penser que le raccordement à une loi normale est légitime. Un tel graphique est réalisé par tous les logiciels de calculs statistiques. Une grille faisant cette anamorphose se trouve après la série d'exercices en fin de chapitre.

3.2.2 Raccordement à une loi log-normale

Pour reconnaître sommairement si une distribution observée est du type log-normal, il est également commode d'utiliser un graphique de Henry avec échelle des abcisses logarithmique (voir grille gausso-logarithmique en fin de chapitre).

3.3 Estimation des paramètres de la loi de référence

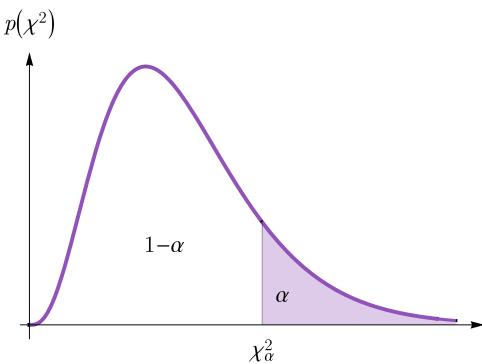
La loi de référence dépend le plus souvent d'un certain nombre de paramètres qu'il est nécessaire d'estimer pour la définir complètement. Une loi binomiale ou de Poisson est entièrement définie par la proportion π à laquelle elle correspond (n étant connu). Une loi normale est entièrement définie par sa moyenne μ et son écart-type σ . Il convient donc, à partir des données disponibles, d'estimer soit la proportion π , soit la moyenne μ et l'écart-type σ de la loi de référence binomiale, de Poisson ou normale, pour ne considérer que ces trois lois là.

3.4 Vérification de la légitimité d'un raccordement effectué

La comparaison des n_k observés et des $n\varpi_k$ théoriques met en évidence des différences plus ou moins fortes. Cela n'a rien d'étonnant puisque, dans l'hypothèse où le raccordement opéré est justifié, la distribution des $n\varpi_k$ n'est que la loi limite de la distribution des n_k . Il reste toutefois à savoir si les différences ainsi mises en évidence sont compatibles avec les seuls aléas de l'échantillonnage. Ce n'est en effet qu'à cette condition qu'on peut considérer le raccordement opéré comme légitime.

La vérification consiste à déterminer la loi d'une certaine fonction de l'ensemble des fluctuations entre effectifs observés et théoriques, *dans l'hypothèse où ces fluctuations ne sont effectivement dues qu'aux aléas de l'échantillonnage.*

Retenant la quantité $d = \sum_{k=1}^r \frac{(n_k - n\varpi_k)^2}{n\varpi_k}$, nous avons montré que, dans ces conditions, elle était une réalisation d'une loi du χ^2 . A un seuil de probabilité α faible pouvant être considéré comme négligeable correspond une valeur χ_α^2 telle que la probabilité d'observer $\chi^2 > \chi_\alpha^2$ soit justement égale à α .



Si la valeur χ^2 observée est supérieure à χ_α^2 , il paraît préférable de mettre en doute l'hypothèse de la légitimité du raccordement. Si, au contraire, χ^2 est inférieur à χ_α^2 , il n'y a pas de raison de mettre en doute cette hypothèse. Comme on s'y est déjà habitué, cela ne signifie malheureusement pas qu'elle soit vraie. Or ce que l'on souhaiterait généralement c'est confirmer la validité du modèle envisagé. L'aspect négatif du test statistique, dans le sens où il ne prend pas en compte le risque de conserver à tort l'hypothèse faise, est gênant dans ce cas précis.

Notez que l'on a effectué un test à droite, puisque ce sont des écarts importants entre effectifs observés et théoriques que l'on veut éventuellement détecter, donc une valeur grande du χ^2 . Ajoutons enfin deux remarques sur la mise en oeuvre du test.

La première est que, pour que la loi de la quantité χ^2 soit suffisamment voisine d'une loi du χ^2 , il faut non seulement que n soit assez grand, mais encore que les nombres $n\varpi_k$ ne soient pas trop petits : en pratique ils ne doivent pas être inférieurs à 5. Si certains d'entre eux sont trop petits, il est nécessaire de procéder à des *groupages de classes*.

La seconde remarque est que, le plus souvent, la loi de référence dépend d'un ou plusieurs paramètres inconnus. A ce moment là, les $n\varpi_k$ sont calculés non pas à partir des paramètres véritables de la loi, mais à partir des paramètres estimés. Ils sont donc eux-mêmes aléatoires. On démontre alors que le nombre de degrés de liberté de la loi du χ^2 à laquelle il faut se référer est égal à $(r - 1 - p)$, où p est le *nombre de paramètres estimés*.

4 Tests non paramétriques

Fort souvent on est amené à prendre en considération des variables dont on ignore la distribution. Il n'est alors plus possible de se référer aux tests de comparaison décrits dans le chapitre 5. Pour lever cette difficulté, on s'est donc préoccupé de définir des tests, dits *non paramétriques*, ne faisant aucune hypothèse sur la nature des populations mises en jeu.

Il existe une très grande variété de tels tests non paramétriques, mais nous nous limiterons à la présentation de ceux qui sont les plus utilisés et qui se trouvent reposer sur la prise en compte d'une même quantité χ^2 que le test ci-dessus du raccordement entre une distribution observée et une distribution théorique.

4.1 Test de comparaison de plusieurs populations qualitatives

Soit p populations $\mathcal{P}_1, \dots, \mathcal{P}_i, \dots, \mathcal{P}_p$ dont les individus sont distingués suivant r catégories $C_1, \dots, C_k, \dots, C_r$ qui peuvent être les modalités d'une variable qualitative (ou les classes d'une variable quantitative). Pour deux lots de pièces, par exemple, classées en bonnes ou mauvaises, on a $p = 2$ et $r = 2$.

On a prélevé un échantillon dans chacune de ces populations. Soient $n_1, \dots, n_j, \dots, n_p$, leurs tailles et soit n_{jk} le nombre d'individus qui proviennent de la population \mathcal{P}_j et qui appartiennent à la catégorie C_k .

	\mathcal{P}_1	...	\mathcal{P}_j	...	\mathcal{P}_p
C_1	n_{11}	...	n_{j1}	...	n_{p1}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
C_k	n_{1k}	...	n_{jk}	...	n_{pk}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
C_r	n_{1r}	...	n_{jr}	...	n_{pr}
	n_1	...	n_j	...	n_p

Si l'on fait l'hypothèse que les populations sont *identiques*, alors les probabilités d'appartenir à chacune des classes sont les mêmes pour toutes les populations, soit $\varpi_1, \dots, \varpi_k, \dots, \varpi_r$, et l'on peut définir des effectifs théoriques dans chaque classe et pour chaque population : $n_j \varpi_k$ pour la classe C_k de la population \mathcal{P}_j . Il semble naturel d'estimer la probabilité dans la classe C_k par :

$$\varpi_k^* = \frac{\sum_{j=1}^p n_{jk}}{\sum_{j=1}^p n_j}$$

et d'envisager la quantité :

$$d = \sum_{j=1}^p \sum_{k=1}^r \frac{(n_{jk} - n_j \varpi_k^*)^2}{n_j \varpi_k^*}$$

pour tester l'hypothèse d'identité des populations. On montre que, sous cette hypothèse, elle obéit à une loi du χ^2 à $(p(r - 1) - (r - 1)) = (p - 1)(r - 1)$ degrés de liberté.

4.2 Test de la médiane

Etant donnés les résultats fournis par deux échantillons de taille n_1 et n_2 :

échantillon 1 : x_1, x_2, \dots, x_{n_1}

échantillon 2 : y_1, y_2, \dots, y_{n_2}

Arrangeons l'ensemble de ces résultats selon une même suite croissante : $x_1, y_1, y_2, x_2, y_3, x_3, \dots$ et désignons la médiane de cette suite par M .

Après décompte des observations au dessus et en dessous de M , le tableau des données peut être résumé ainsi :

Observés	$> M$	$< M$	Total
Echantillon 1	n_{11}	n_{12}	n_1
Echantillon 2	n_{21}	n_{22}	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	n_1+n_2

Dans l'hypothèse où les deux populations sont identiques, la proportion théorique des observations au dessus et en dessous de la médiane est dans tous les cas $1/2$. Au tableau précédent correspond le tableau théorique ci-contre, et on est en définitive ramené à un test du χ^2 avec un degré de liberté.

Théoriques	$> M$	$< M$	Total
Echantillon 1	$\frac{n_1}{2}$	$\frac{n_1}{2}$	n_1
Echantillon 2	$\frac{n_2}{2}$	$\frac{n_2}{2}$	n_2
Total	$\frac{n_1+n_2}{2}$	$\frac{n_1+n_2}{2}$	n_1+n_2

4.3 Test des signes

Ce test s'applique à des observations appariées. Sur un même individu i on a effectué deux mesures y_i et x_i et on s'intéresse aux différences $d_i = y_i - x_i$. Dans le test classique on prenait en compte les valeurs de ces différences, mais dans le test des signes on ne retiendra que les signes, plus ou moins, de ces différences. Il y a donc perte d'information.

S'il n'y a pas de différence entre les mesures, la probabilité d'un signe plus est égale à celle d'un signe moins et égale à 0.5. S'il y a n individus dans l'échantillon, les effectifs théoriques sont égaux à $0.5 n$ et on est encore ramené à un test du χ^2 avec un degré de liberté sur la quantité :

$$\frac{(n_+ - 0.5 n)^2}{0.5 n} + \frac{(n_- - 0.5 n)^2}{0.5 n}$$

4.4 Test d'indépendance entre deux variables qualitatives

Soit $x_1, \dots, x_i, \dots, x_p$ et soit $y_1, \dots, y_j, \dots, y_q$ les modalités de deux variables qualitatives X et Y . Un échantillon de n individus sur lesquels ont été repérées les valeurs prises simultanément par les deux variables a donné les résultats ci-contre : n_{ij} est le nombre d'individus ayant présenté à la fois la valeur x_i de X et la valeur y_j de Y . $n_{i.}$ et $n_{.j}$ représentent respectivement le total de la ligne x_i et celui de la colonne y_j .

	y_1	...	y_j	...	y_q	Total
x_1	n_{11}	...	n_{1j}	...	n_{1q}	n_1
\vdots	\vdots		\vdots		\vdots	\vdots
x_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i.}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{.p}$
Total	$n_{.1}$...	$n_{.j}$...	$n_{.p}$	n

Soit les probabilités suivantes :

$$\varpi_{i.} = \mathbb{P}(X = x_i)$$

$$\varpi_{.j} = \mathbb{P}(Y = y_j)$$

$$\varpi_{ij} = \mathbb{P}(X = x_i \text{ et } Y = y_j)$$

Faisons l'hypothèse que les deux variables sont indépendantes. Il s'ensuit, d'après le théorème des probabilités composées, que :

$$\varpi_{ij} = \varpi_{i.} \varpi_{.j}$$

Estimons $\varpi_{i.}$ et $\varpi_{.j}$ respectivement par $\varpi_{i.*} = \frac{n_{i.}}{n}$ et $\varpi_{.*j} = \frac{n_{.j}}{n}$, donc ϖ_{ij} par $\varpi_{ij*} = \frac{n_{i.} n_{.j}}{n^2}$.

Sous l'hypothèse d'indépendance, l'effectif théorique correspondant à l'effectif observé n_{ij} est égal à $n \varpi_{ij}^* = \frac{n_i n_j}{n}$ et la quantité :

$$d = \sum_{i=1}^p \sum_{j=1}^q \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$$

est la réalisation d'une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

Exercices du chapitre 6

Exercice 1

Il y a dans un certain pays 25% d'illettrés. On considère les 400 prisonniers d'une prison de ce pays.

- a) Calculer, en admettant que l'échantillon constitué par les prisonniers est tiré au hasard dans la population du pays, la probabilité pour que la fréquence des prisonniers illettrés dépasse 35% ?
- b) Quelle valeur cette fréquence n'a-t-elle qu'une chance sur 100 de dépasser ?

Exercice 2

Un sondage d'opinion sur les intentions de vote d'un certain électoralat a porté sur un échantillon de 2500 électeurs représentatifs de cet électoralat et a fourni les résultats suivants : 1200 électeurs favorables au candidat A et 1300 favorables au candidat B.

- a) En admettant que la proportion ϖ inconnue de l'électoralat favorable au candidat B soit égale à 0.5 (ballottage), donner la fourchette dans laquelle la fréquence relative à un échantillon de 2500 électeurs aurait la probabilité 95% de tomber.
- b) Que penser d'un journal qui annoncerait, au vu des résultats du sondage, l'élection du candidat B ?
- c) Quelle devrait être la taille de l'échantillon pour qu'une différence de 4% entre les fréquences (48% favorables à A et 52% à B) permette de conclure, avec moins de une chance sur 100 de se tromper, à l'élection du candidat B ?

Exercice 3

Une enquête réalisée auprès de 4000 foyers d'une grande ville en 1990 a montré que 1944 d'entre eux possédaient une machine à laver la vaisselle. Une enquête réalisée dans la même ville en 1995 sur 5000 foyers a montré que 2587 possédaient un tel appareil. Peut-on admettre que le taux d'équipement a augmenté ?

Exercice 4

La direction du marketing d'une entreprise a fait procéder à une enquête auprès d'un échantillon de n consommateurs en leur soumettant deux modes de présentation A et B d'une même marchandise et en leur demandant de faire connaître leur préférence. Soient n_1 et n_2 les nombres de consommateurs ayant préféré respectivement les présentations A et B et soient ϖ_1 et ϖ_2 les proportions correspondantes dans la population. n_1 et n_2 sont des réalisations de variables aléatoires liées par la relation $N_1 + N_2 = n$. On est amené à se demander si la différence constatée $d = n_1 - n_2$ est significative d'une préférence réelle dans la population pour l'un des modes de présentation.

- a) d est une réalisation d'une variable D que l'on peut écrire $D = 2N_1 - n$. Calculer sa moyenne et sa variance.
- b) On se propose de tester l'hypothèse d'une différence nulle $\varpi_1 - \varpi_2 = 0$, dans la population échantillonnée. Quelle est l'estimation de la variance de D qui vous paraît devoir être utilisée ? A quelle condition doit satisfaire la différence d pour qu'elle puisse être considérée comme significative au niveau de signification 95% ?

Exercice 5

Sur un échantillon de 10000 bébés, 5136 sont des garçons et 4864 sont des filles. Peut-on admettre que les probabilités pour qu'un bébé soit un garçon ou une fille sont égales ?

Exercice 6

On effectue des croisements de poules blanches et noires. D'après les lois de Mendel, lorsqu'on effectue un tel croisement, on a 1 chance sur 4 d'obtenir une poule blanche, 1 chance sur 4 d'obtenir une poule noire et 1 chance sur 2 d'obtenir une poule bleue (hybride). Les 158 croisements effectués ont donné 43 poules blanches, 40 poules noires et 75 poules bleues. Ces résultats vous paraissent-ils compatibles avec les lois de Mendel ?

Exercice 7

On considère ci-dessous la distribution du dernier chiffre de 200 lectures de pesée. Peut-on craindre que celui qui effectuait les pesées a une préférence pour certains chiffres ? Peut-on justifier cette crainte ?

x	0	1	2	3	4	5	6	7	8	9	Total
$n(x)$	35	16	15	17	17	19	11	16	30	24	200

Exercice 8

La direction du personnel d'une usine veut déterminer si le nombre d'avis d'arrêt pour maladie dépend du jour de la semaine. Pendant la période étudiée, il y a eu 720 avis d'arrêt. Les cas de maladie du samedi après-midi où l'on ne travaille pas ne sont notifiés que le lundi. Chaque jour on notifie les arrêts survenus entre la veille 17 h et le jour 17 h, sauf le samedi où l'on s'arrête à 13 h et le lundi où sont notifiés les arrêts du samedi 13 h au lundi 17 h. Faisant l'hypothèse d'une répartition uniforme des arrêts, que peut-on conclure de cette enquête ?

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Total
Effectifs	189	100	127	115	102	87	720

Exercice 9

La fabrication de pièces dans un atelier de mécanique donne lieu à un certain pourcentage de pièces rebutées comme non utilisables. On a observé 100 lots différents de 100 pièces chacun qui ont donné les résultats suivants.

Rebuts par lot	0	1	2	3	4	5	6	7	8	9	10	11	Total
Nombre de lots	0	2	9	14	20	18	15	9	6	4	2	1	100

Quelle distribution théorique paraît devoir donner une bonne description de la distribution observée et pour quelle raison ? Tester l'accord entre les observations et le modèle.

Exercice 10

On a mesuré les duretés Rockwell sur 100 tôles minces en faisant 3 mesures au centre, dont on a pris la moyenne. Les résultats des mesures, une fois ordonnés et mis en classes, ont été rassemblés dans le tableau suivant.

Classes de dureté	53	54	55	56	57	58	59	60	61	62	63	64	Total
Nb d'observations	1	4	8	13	16	19	14	11	6	5	2	1	100

Que penser du raccordement à une loi normale ? Tracer la droite de Henry.

Exercice 11

On donne dans le tableau suivant la distribution des vitesses de passage de bobines de tôle lors de l'opération de décapage (décalaminage chimique et mécanique). Ces vitesses sont en m/mn.

Classes de vitesse	40	50	60	70	80	90	100	110	120	130	140	150	Total
Effectifs	3	4	19	38	62	72	77	56	46	16	5	2	400

Raccorder cette distribution à une loi normale. On donne $m \approx 95$ et $s \approx 20$

Exercice 12

On étudie conjointement la couleur des cheveux et celle des sourcils d'une certaine population. On les classe en deux catégories : clairs (blonds ou roux) et foncés (bruns ou noirs). On trouve les résultats suivants.

Cheveux Sourcils	Cheveux Clairs	Cheveux Foncés	Total
Sourcils Clairs	30 472	3238	33 710
Sourcils Foncés	3364	9468	12 832
Total	33 836	12 706	46 542

Y-a-t-il une dépendance entre la couleur des cheveux et celle des sourcils ?

Exercice 13

Dans une usine, on a remplacé la commande manuelle de quelques presses par une commande automatique. On désire voir si cette modification a une influence sur les accidents du travail. On a relevé, pendant une période donnée, le nombre d'ouvriers qui ont eu ou non des accidents et on les a classés suivant qu'ils travaillaient sur des presses à commande manuelle ou à commande automatique. On a obtenu les résultats suivants.

	Manuelle	Automatique	Total
Accidentés	25	23	48
Non accidentés	183	112	295
Total	208	135	343

La modification du type de commande a-t-elle une influence sur le nombre des accidents ?

Exercice 14

En vue d'étudier les effets du tabac sur l'artério-sclérose, on a classé 870 individus selon :

- leur degré d'artério-sclérose (grave, moyen, faible),
- leur consommation de tabac (non fumeurs, légers fumeurs, moyens fumeurs, grands fumeurs de cigarettes, fumeurs de pipe et cigares).

Analyser le tableau suivant qui a été obtenu.

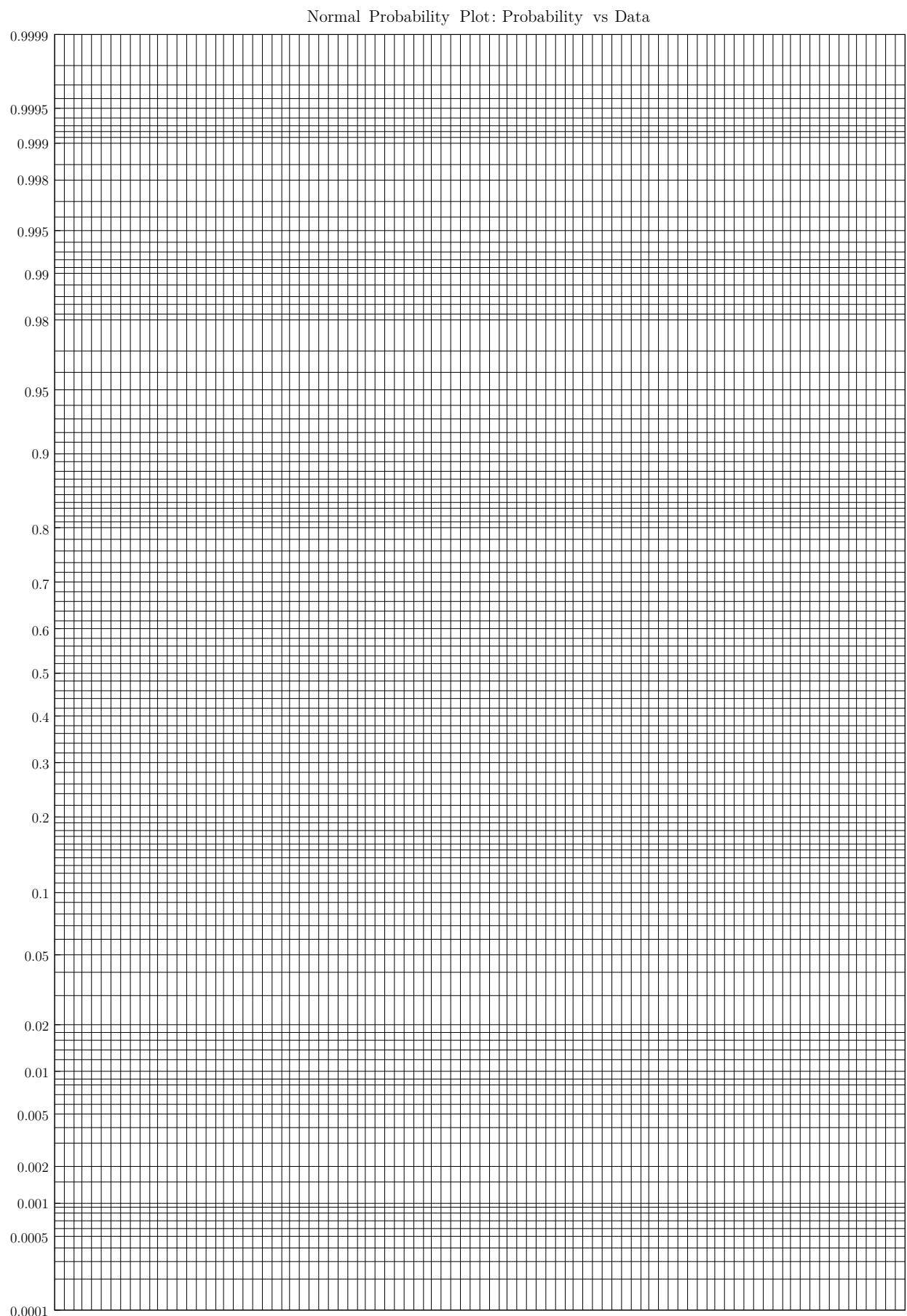
Sclérose	Non fumeurs	Légers	Moyens	Grands	Pipes et cigares
Grave	16	29	76	50	7
Moyen	97	96	180	122	42
Faible	48	27	32	27	21

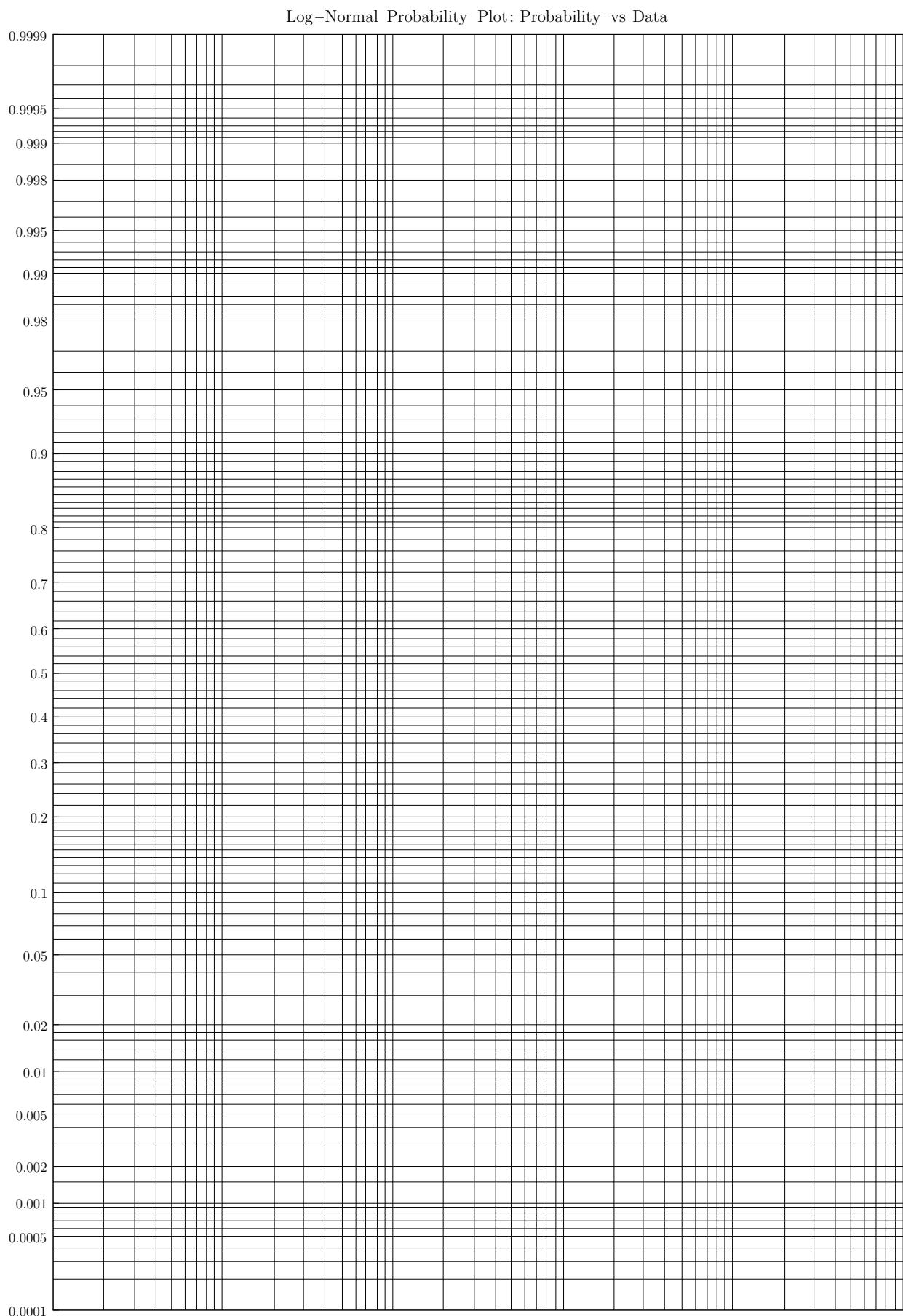
Exercice 15

Les conclusions d'une étude de la Gendarmerie sur 823 accidents graves (ayant provoqué la mort ou des blessures pour lesquelles une hospitalisation de plus de 8 jours a été nécessaire) étaient les suivantes : « *Le risque d'être tué pour un passager placé à l'avant est le même que celui encouru par le conducteur, mais ce risque est presque deux fois plus élevé que pour un passager placé à l'arrière. On peut donc parler de 'place du mort' pour la place occupée par le passager avant droit, mais uniquement si l'on compare le passager avant à l'ensemble des passagers, conducteur exclu. Du point de vue de la gravité des accidents, le passager avant est celui qui est le plus exposé* ».

Justifier ces conclusions à partir du tableau des résultats.

Place	Tués	Blessés graves	Blessés légers	Indemnités	Total
Conducteur	46	153	95	38	332
Place avant	34	150	60	20	264
Place arrière	16	81	96	34	227
Total	96	384	251	92	823





La régression linéaire

Les sciences exactes sont fondées sur la notion de relations répétables, qui peut s'énoncer ainsi : dans les mêmes conditions, les mêmes causes produisent les mêmes effets. Notant alors x la mesure des causes, et y celle des effets, la liaison entre y et x s'écrit suivant la relation fonctionnelle $y = f_c(x)$: à une valeur donnée de x correspond une valeur bien déterminée de y .

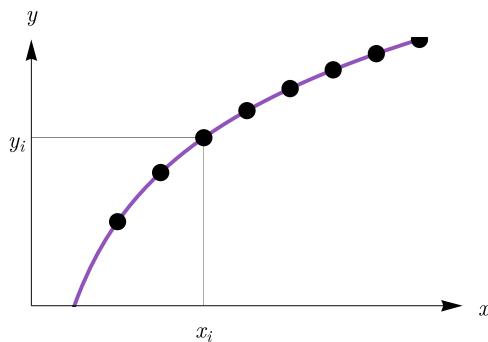
Or, pour de nombreux phénomènes (notamment industriels), une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité. Il en résulte que la reproductibilité des conditions, d'une expérience à une autre, ne peut être garantie. Partant de cette constatation, la statistique va permettre d'étendre la notion de relation fonctionnelle répétable, à celle de corrélation où la relation entre x et y est entachée d'une certaine dispersion due à la variabilité des conditions d'expérience : on écrira $y = f(x) + \varepsilon$, où ε est une variable aléatoire.

1 La droite des moindres carrés

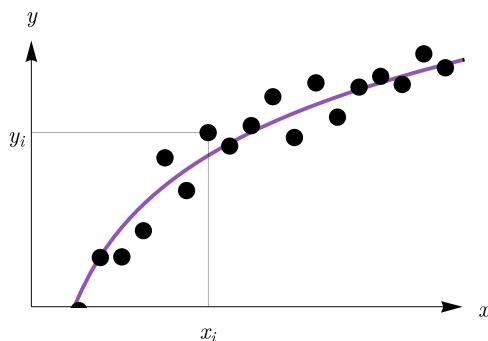
1.1 Nuage des individus

Le problème consiste à étudier l'influence d'une variable quantitative X sur une autre variable quantitative Y . La première est souvent appelée variable *explicative* (ou encore exogène) et la seconde est appelée variable *expliquée* (ou encore endogène). Pour cela, on a réalisé une expérimentation qui consiste à prélever un échantillon de n individus, et à mesurer sur chacun d'eux les valeurs prises par chacune des deux variables. En vue, par exemple, d'étudier l'influence de la teneur en carbone d'un acier sur sa résistance à la traction, on a procédé à la mesure de ces deux variables sur 100 éprouvettes. On dispose donc d'un échantillon de n couples d'observations (x_i, y_i) que l'on peut représenter sur un graphique, dans le plan \mathbb{R}^2 , où chaque point i , de coordonnées (x_i, y_i) , correspond à un couple d'observations. Plusieurs cas peuvent se présenter.

Les points s'alignent sur une courbe qui, dans l'hypothèse la plus simple est une droite. On dit que la relation entre Y et X est *fonctionnelle* : lorsque la valeur de X est donnée, celle de Y est déterminée sans ambiguïté. C'est le cas idéal qui, expérimentalement, n'est jamais réalisé de façon parfaite.



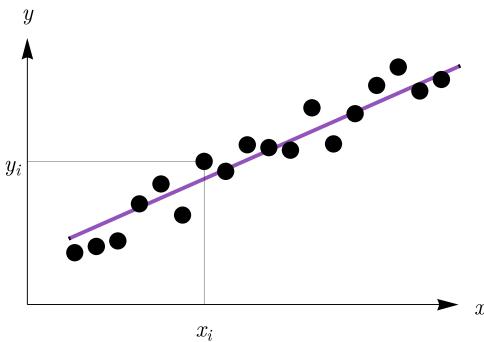
Les mesures sont en effet toujours entachées de quelque imprécision. Les points forment alors un nuage. Mais celui-ci présente une orientation qui suggère, par exemple, que lorsque X augmente, la valeur moyenne de Y augmente également comme dans l'illustration suivante.



Lorsque X est donné, Y n'est pas complètement déterminé : ses valeurs se dispersent autour d'une certaine valeur moyenne. Mais les valeurs moyennes décrivent, lorsque X varie, une courbe qui est appelée la *ligne de régression* de Y par rapport à X :

$$\mathbb{E}(Y|X = x) = f(x)$$

La liaison entre Y et X est alors appelée *stochastique* (ou statistique). Un cas particulièrement important est celui où le nuage se dispose suivant une forme allongée et exhibe une tendance sensiblement linéaire. C'est à ce cas de *régression linéaire* que nous allons nous attacher dans ce chapitre.



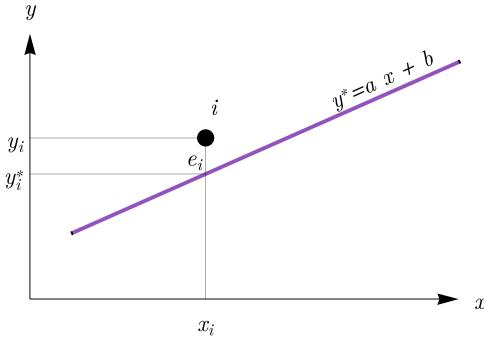
Cette condition de linéarité n'est pas aussi restrictive qu'il pourrait paraître : on peut toujours trouver une transformation mathématique qui permet de passer d'une ligne de régression d'équation quelconque à une droite. Si la tendance est, par exemple, de la forme $y = b x^a$, il suffira d'effectuer les changements de variable $y' = \log(y)$ et $x' = \log(x)$ pour retrouver une relation linéaire : $\log(y) = a \log(x) + \log(b)$.

1.2 Caractérisation de la droite de régression

Nous cherchons une droite $y^* = a x + b$ qui décrive au mieux la tendance du nuage observé. La démarche la plus couramment utilisée consiste à :

- faire l'hypothèse que, pour chaque individu i , on a : $y_i = a x_i + b + e_i$, où e_i est une certaine « erreur », appelée *résidu*, qui s'ajoute à la valeur $y_i^* = a x_i + b$ qui résulterait d'une relation linéaire entre Y et X ,
- à rechercher la droite $y^* = a x + b$, dite *droite des moindres carrés*, telle que la somme quadratique des résidus e_i soit minimale, c'est-à-dire que :

$$S = \sum_{i=1}^n e_i^2 \text{ soit minimale.}$$



Cette quantité, égale à $\sum_{i=1}^n (y_i - a x_i - b)^2$ quand on l'exprime en fonction a et b , est minimale si les dérivées partielles par rapport à a et b sont nulles :

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^n x_i (y_i - a x_i - b) = -2 \sum_{i=1}^n x_i e_i = 0 \quad (1)$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^n (y_i - a x_i - b) = -2 \sum_{i=1}^n e_i = 0 \quad (2)$$

En appelant \bar{y} et \bar{x} les moyennes des valeurs x_i et y_i :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ et } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

l'équation (2) permet d'obtenir la valeur de l'ordonnée à l'origine de la droite :

$$b = \bar{y} - a \bar{x}$$

qui signifie que la droite des moindres carrés passe par le point moyen du nuage, de coordonnées (\bar{x}, \bar{y}) .

Le système des deux équations (1) et (2) devient alors, en remplaçant b par sa valeur :

$$\sum_{i=1}^n x_i[(y_i - \bar{y}) - a(x_i - \bar{x})] = 0 \quad (3)$$

$$\sum_{i=1}^n [(y_i - \bar{y}) - a(x_i - \bar{x})] = 0 \quad (4)$$

Multipliant (4) par \bar{x} et retranchant de (3), il vient :

$$\sum_{i=1}^n (x_i - \bar{x}) [(y_i - \bar{y}) - a(x_i - \bar{x})] = 0 \quad (5)$$

d'où l'on déduit la valeur de a :

$$a = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

On peut noter que $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ est la covariance empirique de X et Y et que $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ est la variance empirique de X . Par conséquent, l'expression de a peut s'écrire :

$$a = \frac{\text{Cov}(X, Y)}{\text{V}(X)}$$

1.3 Analyse de la variance

Rapportant l'observation y_i à la moyenne \bar{y} des observations, on peut écrire :

$$(y_i - \bar{y}) = a(x_i - \bar{x}) + e_i$$

Dans cette expression, la quantité $a(x_i - \bar{x})$ représente ce qui est « expliqué » par X , et la quantité e_i est une erreur qu'on a appellée un *résidu*.

En élevant au carré et en sommant pour toutes les observations, il vient :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = a^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n e_i^2 \quad (6)$$

En effet, le double produit est nul puisqu'il peut s'écrire :

$$2 a \sum_{i=1}^n (x_i - \bar{x})[(y_i - \bar{y}) - a(x_i - \bar{x})]$$

forme sous laquelle on reconnaît le premier membre de l'équation (5) précédente.

La relation (6) est appelée *équation d'analyse de la variance*. En fait, il s'agit de sommes de carrés et non de variances et on l'écrit généralement sous la forme suivante :

$$\text{SCT} = \text{SCE} + \text{SCR}$$

où l'on peut retenir qu'elle décompose la somme des carrés *totale* SCT en une somme de carrés *expliquée* SCE et une somme de carrés *résiduelle* SCR.

1.4 Coefficient de corrélation

On définit alors le carré du coefficient de corrélation noté r^2 comme le ratio :

$$r^2 = \frac{\text{SCE}}{\text{SCT}}$$

Il représente la part relative de la variabilité totale de Y qui est expliquée par X :

$$\text{SCE} = r^2 \text{SCT}$$

Symétriquement, $(1 - r^2)$ représente la part résiduelle :

$$\text{SCR} = (1 - r^2) \text{SCT} \quad (7)$$

En explicitant SCE et SCT puis a , on peut écrire :

$$r^2 = \frac{a^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{[\text{Cov}(X, Y)]^2}{\text{V}(X) \text{V}(Y)}$$

Le coefficient de corrélation a le signe de la covariance :

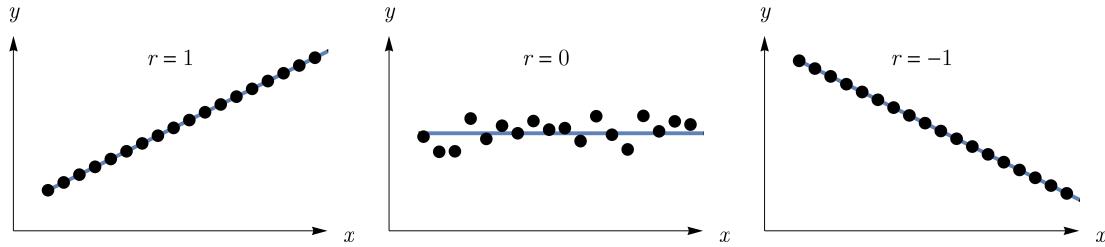
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{V}(X)} \sqrt{\text{V}(Y)}}$$

de telle façon que, si X et Y varient dans le même sens, r est positif ; sinon, il est négatif.

Il résulte de la relation (7) que le coefficient de corrélation est toujours compris entre -1 et 1, puisqu'une somme de carrés est nécessairement positive.

Le coefficient de corrélation présente les valeurs remarquables suivantes :

- si $|r| = 1$, il y a une relation fonctionnelle linéaire entre X et Y ;
- si $r = 0$, Y est indépendante de X : la covariance est nulle et la droite de régression est horizontale ;
- la liaison entre X et Y est d'autant plus intime que $|r|$ est voisin de 1 et d'autant plus faible que $|r| \rightarrow 0$.

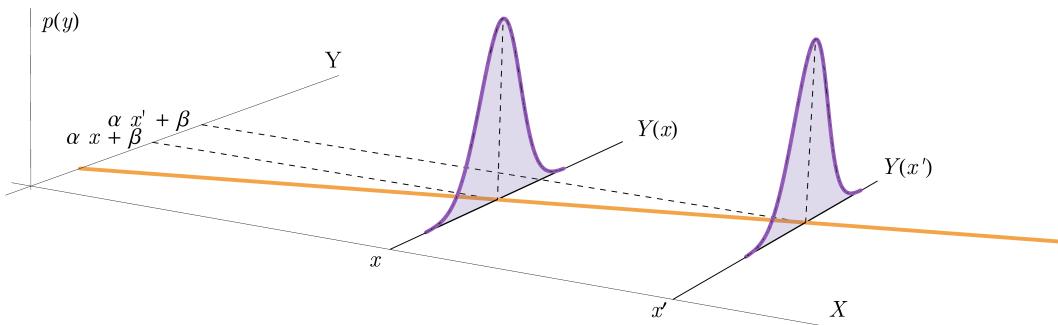


2 Propriétés statistiques de la droite des moindres carrés

2.1 Le modèle de la régression linéaire

Nous nous sommes jusqu'ici limités à décrire l'échantillon des valeurs observées (x_i, y_i) sans faire d'hypothèses sur la structure de la population dans laquelle il a été prélevé. Pour pouvoir pratiquer l'inférence, c'est-à-dire émettre des conclusions qui soient valables pour cette population, nous sommes obligés d'adopter un modèle de population.

Nous admettrons que, pour un individu i prélevé au hasard dans la population, x_i est *déterministe*, et que y_i est une réalisation d'une variable aléatoire $Y_i = \alpha x_i + \beta + \varepsilon_i$. Les paramètres α et β sont des quantités certaines mais inconnues qu'il faudra estimer.



Les quantités ε_i sont des variables aléatoires avec les propriétés suivantes :

- elles sont centrées : $\mathbb{E}(\varepsilon_i) = 0$,
- elles ont même variance : $\text{V}(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) - \mathbb{E}(\varepsilon_i)^2 = \mathbb{E}(\varepsilon_i^2) = \sigma^2$,
- elles sont indépendantes : $\text{Cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) - \mathbb{E}(\varepsilon_i) \mathbb{E}(\varepsilon_j) = \mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ si $i \neq j$.

Pour une valeur donnée x_i , on a :

$$\mathbb{E}[Y(x_i)] = \mathbb{E}(Y_i) = \alpha x_i + \beta$$

La ligne de régression est donc la droite d'équation $y = \alpha x + \beta$. La dispersion autour de cette droite correspond à un écart-type σ . Elle est indépendante de X .

Rappelons que nous avions écrit, à partir de la droite des moindres carrés, que :

$$y_i = a x_i + b + e_i$$

Sous les hypothèses ci-dessus, nous allons montrer que a et b sont des estimations sans biais de α et β et qu'il est possible d'estimer σ^2 à partir de $\text{SCR} = \sum_{i=1}^n e_i^2$.

2.2 Propriétés de a et b

Conformément au modèle adopté, a est à considérer comme une réalisation de la variable aléatoire :

$$A = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et b comme une réalisation de la variable aléatoire :

$$B = \bar{Y} - A \bar{x}.$$

A et B sont des *estimateurs sans biais et convergents* de α et β . Les calculs qui le montrent constituent un bon entraînement à la pratique des opérateurs *espérance* et *variance*.

En effet, tenant compte de ce que $Y_i = \alpha x_i + \beta + \varepsilon_i$, on peut mettre A sous la forme :

$$A = \alpha + \frac{\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \alpha + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i - \bar{\varepsilon} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \alpha + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

car $\sum_{i=1}^n (x_i - \bar{x}) = 0$ par définition de \bar{x} .

Et B sous la forme suivante :

$$B = \beta - \bar{x}(A - \alpha) + \bar{\varepsilon}$$

On en déduit tout d'abord les espérances mathématiques de A et B :

$$\begin{aligned} \mathbb{E}(A) &= \alpha + \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}(\varepsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \alpha & \text{car } \mathbb{E}(\varepsilon_i) = 0 \\ \mathbb{E}(B) &= \beta - \bar{x} \mathbb{E}(A - \alpha) + \mathbb{E}(\bar{\varepsilon}) = \beta \text{ car } \mathbb{E}(A - \alpha) = \mathbb{E}(A) - \alpha = 0 \text{ et } \mathbb{E}(\bar{\varepsilon}) = \mathbb{E}(\varepsilon_i) = 0 \end{aligned}$$

On peut calculer ensuite les variances de A et B :

$$\begin{aligned} \mathbb{V}(A) &= \frac{[\sum_{i=1}^n (x_i - \bar{x})^2] \mathbb{V}(\varepsilon_i)}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \mathbb{V}(B) &= \bar{x}^2 \mathbb{V}(A) + \mathbb{V}(\bar{\varepsilon}) - 2 \bar{x} \text{Cov}(\bar{\varepsilon}, A) = \frac{\bar{x}^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sigma^2}{n} = \sigma^2 \left(\frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} \right) \end{aligned}$$

puisque la covariance de $\bar{\varepsilon}$ et A est nulle. En effet :

$$\text{Cov}(\bar{\varepsilon}, A) = \mathbb{E}\left[\frac{1}{n} \sum_{j=1}^n \varepsilon_j \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] = \frac{1}{n} \sum_{j=1}^n \frac{\sum_{i=1}^n (x_i - \bar{x}) \mathbb{E}(\varepsilon_i \varepsilon_j)}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0$$

On peut enfin calculer la covariance de A et B :

$$\begin{aligned} \text{Cov}(A, B) &= \text{Cov}(A - \alpha, B - \beta) = \mathbb{E}[(A - \alpha)(B - \beta)] - 0 \\ &= \mathbb{E}[(A - \alpha)(\bar{\varepsilon} - \bar{x}(A - \alpha))] = \text{Cov}(\bar{\varepsilon}, A) - \bar{x} \mathbb{V}(A) = - \frac{\bar{x} \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

On constate ainsi que A et B sont des estimateurs de α et β : *sans biais* ($\mathbb{E}(A) = \alpha$, $\mathbb{E}(B) = \beta$) et *convergents* ($\mathbb{V}(A) \xrightarrow[n \rightarrow \infty]{\longrightarrow} 0$ et $\mathbb{V}(B) \xrightarrow[n \rightarrow \infty]{\longrightarrow} 0$) mais *qui ne sont pas indépendants* ($\text{Cov}(A, B) \neq 0$).

Par contre, A et \bar{Y} sont *indépendants* ($\text{Cov}(\bar{Y}, A) = 0$). Ce résultat sera exploité un peu plus loin.

2.3 Estimation de σ^2

Montrons maintenant que :

$$\sigma^{*2} = \frac{\text{SCR}}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{(1-r^2) \sum_{i=1}^n (y_i - \bar{y})^2}{n-2}$$

est une *estimation sans biais* de σ^2 .

Utilisant conjointement :

$$\begin{aligned} (y_i - \bar{y}) &= a(x_i - \bar{x}) + e_i \\ (y_i - \bar{y}) &= \alpha(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned}$$

on peut écrire que :

$$e_i = (\varepsilon_i - \bar{\varepsilon}) - (a - \alpha)(x_i - \bar{x})$$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n [(\varepsilon_i - \bar{\varepsilon}) - (a - \alpha)(x_i - \bar{x})]^2 = \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2 - (a - \alpha)^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

car, revenant à l'expression de A (dont a est une réalisation), le double produit peut s'écrire :

$$-2(a - \alpha) \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x}) = -2(a - \alpha)(a - \alpha) \sum_{i=1}^n (x_i - \bar{x})^2 = -2(a - \alpha)^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

On arrive alors à :

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n \varepsilon_i^2 - n \bar{\varepsilon}^2 - (a - \alpha)^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Calculons maintenant l'espérance :

$$\mathbb{E}\left(\sum_{i=1}^n e_i^2\right) = n \sigma^2 - n \frac{\sigma^2}{n} - \mathbb{E}((A - \alpha)^2) \sum_{i=1}^n (x_i - \bar{x})^2$$

Or :

$$\mathbb{E}((A - \alpha)^2) = \mathbb{V}(A - \alpha) + (\mathbb{E}(A - \alpha))^2 = \mathbb{V}(A - \alpha) + 0^2 = \mathbb{V}(A) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

On arrive donc à :

$$\mathbb{E}\left(\sum_{i=1}^n e_i^2\right) = n \sigma^2 - n \frac{\sigma^2}{n} - \sigma^2 = (n - 2) \sigma^2$$

et finalement :

$$\mathbb{E}(\sigma^{*2}) = \sigma^2$$

3 La prévision statistique

3.1 Objectifs

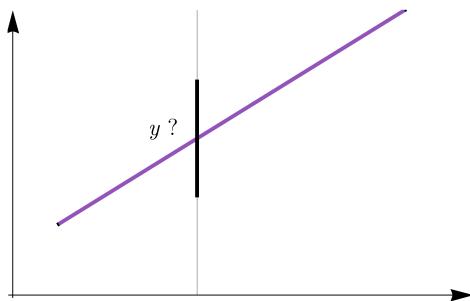
Dans une entreprise, on peut distinguer trois fonctions essentielles que nous allons brièvement illustrer par des exemples.

Décision : les performances d'un matériel dépendent de son âge. Au-dessous d'un certain seuil de performance, il convient de le réformer. Étant donné l'âge d'un matériel, il faudra décider de sa réforme ou de son maintien en activité.

Prévision : la consommation en matière première (ou en énergie) dépend de la quantité produite. Visant, pour une période future, une certaine production, quel stock de matière première faut-il prévoir ?

Contrôle : dans le même contexte, une certaine production ayant été assurée pour une certaine consommation, cette dernière est-elle « normale », faible, élevée ?

Ces trois problèmes se formulent finalement de la même façon. Pour une valeur donnée de X , quelle valeur attribuer à Y , et *avec quelle précision* ? D'un point de vue pratique, c'est l'objectif principal de ce qui suit.



Nous chercherons à apporter des réponses aux questions suivantes :

- la liaison entre les deux variables Y et X est-elle significative ? Autrement dit, peut-on ou non admettre que $\alpha = 0$?
- quels intervalles de confiance retenir pour les paramètres du modèle α et β ?
- pour une valeur donnée de X , comment estimer la valeur correspondante de Y ?

3.2 Hypothèse de normalité

La résolution de ces problèmes nécessite de compléter le modèle, en admettant pour les ε_i , outre les hypothèses précédentes (variance constante, indépendance), l'hypothèse de *normalité*.

Cette dernière hypothèse va permettre d'établir les lois de probabilité des estimateurs A et B , et celle d'un point quelconque $A x + B$. En effet, les x_i étant fixés, ces trois quantités sont des combinaisons linéaires des ε_i , donc suivent elles aussi des lois normales.

On peut montrer d'autre part que, sous l'hypothèse de normalité des ε_i , la quantité : $\frac{\sum_{i=1}^n \varepsilon_i^2}{\sigma^2}$ suit une loi du χ^2 à $(n - 2)$ degrés de liberté, et qu'elle est indépendante des quantités A et B .

3.3 Test d'indépendance des variables

La variable aléatoire A suit une loi normale dont nous avons montré en 2.2 que :

$$\mathbb{E}(A) = \alpha \text{ et } \mathbb{V}(A) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Faisons l'hypothèse qu'il n'y a pas de liaison entre les variables, c'est-à-dire que $\alpha = 0$. Il en découle que A suit une loi de moyenne nulle, donc que la quantité $\frac{A}{\sigma(A)}$ suit une loi normale centrée réduite.

Par suite, si on estime σ^2 par :

$$\sigma^{*2} = \frac{(1-r^2) \sum_{i=1}^n (y_i - \bar{y})^2}{n-2}$$

on a :

$$T = \frac{A}{\sigma^*/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim \mathcal{T}(n-2)$$

Il suffit de calculer sa valeur expérimentale :

$$t = \frac{a}{\sigma^*/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

et de la comparer au seuil lu dans la table correspondante pour le risque choisi.

Les mêmes propriétés permettent de calculer un intervalle de confiance pour α .

3.4 Test de nullité de l'ordonnée à l'origine

La variable aléatoire B suit une loi normale dont nous avons montré en 2.2 que :

$$\mathbb{E}(B) = \beta \text{ et } \mathbb{V}(B) = \left(\frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} \right) \sigma^2$$

En procédant comme précédemment, on peut tester si la droite de régression passe par l'origine ($\beta = 0$), ou calculer un intervalle de confiance pour l'estimation de β .

3.5 Intervalles de confiance pour une valeur donnée x de X

Nous allons envisager successivement les intervalles de confiance :

- d'un point de la droite de régression $y(x) = \alpha x + \beta$, c'est-à-dire de la *valeur moyenne des observations* pour une valeur x donnée,
- puis d'un point du nuage $Y(x)$, c'est-à-dire de la *valeur d'une observation* pour une valeur x donnée.

Il est absolument nécessaire de bien prendre conscience de la différence fondamentale entre ces deux problèmes, dont les applications sont nombreuses et importantes : il s'agit dans le second cas de l'intervalle de confiance de $Y(x)$ alors que dans le premier cas, il s'agit de l'intervalle de confiance de $\mathbb{E}[Y(x)]$.

3.5.1 Intervalle de confiance d'un point de la droite $y = \alpha x + \beta$

L'équation de la droite de régression s'écrit :

$$y(x) = \alpha x + \beta.$$

Celle de la droite des moindres carrés s'écrit :

$$y^*(x) = a x + b$$

que l'on considère comme une réalisation de la variable aléatoire

$$Y(x) = A x + B.$$

D'après les résultats établis en 2.2, $\mathbb{E}[Y(x)] = y(x)$. Donc le point $y^*(x)$ de la droite des moindres carrés est une estimation du point correspondant de la droite de régression $y(x)$.

Pour calculer maintenant la variance de $Y(x)$, on l'écrit sous la forme :

$$Y(x) = A(x - \bar{x}) + \bar{Y}$$

qui sera commode puisqu'on a vu que A et \bar{Y} sont indépendantes. D'où la variance :

$$\mathbb{V}[Y(x)] = (x - \bar{x})^2 \mathbb{V}(A) + \mathbb{V}(\bar{Y}) = \left(\frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} \right) \sigma^2 = \left(\frac{(x - \bar{x})^2}{n s^2} + \frac{1}{n} \right) \sigma^2$$

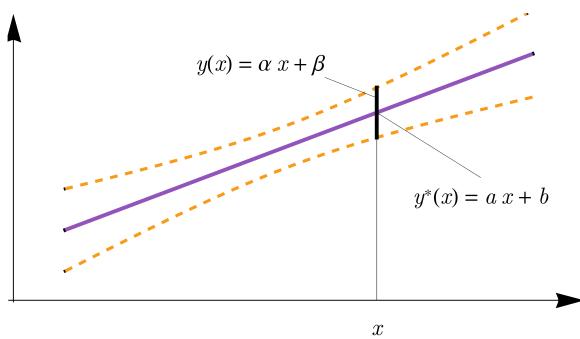
La quantité $\frac{Y(x) - y(x)}{\sigma[Y(x)]}$ suit une loi normale centrée réduite. Et en estimant σ par σ^* , le quotient :

$$T = \frac{Y(x) - y(x)}{\sigma^* \sqrt{\frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n}}} \sim \mathcal{T}(n - 2)$$

Cette propriété permet de trouver un intervalle de confiance pour $y(x)$. Pour un risque α donné, on a :

$$y_\alpha(x) \in a x + b \pm t_{\alpha/2} \sigma^* \sqrt{\frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n}}$$

Lorsque x varie, les limites ainsi obtenues décrivent une hyperbole dont les branches délimitent l'intervalle de confiance de $y(x)$, c'est-à-dire celle de l'espérance de $Y(x)$.



3.5.2 Intervalle de confiance d'une observation

Un échantillon de n points a permis de déterminer les estimations a , b et σ^* . Nous cherchons à faire des prévisions sur l'ordonnée y_{n+1} d'un $(n+1)$ ème point d'abscisse x_{n+1} donnée. Cela revient à estimer y_{n+1} et à déterminer son intervalle de confiance.

Pour estimer y_{n+1} on prendra :

$$y^*(x_{n+1}) = a x_{n+1} + b$$

qui est une estimation sans biais de y_{n+1} . En effet :

y_{n+1} est une réalisation de $Y_{n+1} = y(x_{n+1}) + \epsilon_{n+1}$ avec $y(x_{n+1}) = \alpha x_{n+1} + \beta$

$y^*(x_{n+1})$ est une réalisation de $Y(x_{n+1}) = A x_{n+1} + B$

et, d'après 2.2, on a bien :

$$\mathbb{E}(Y_{n+1}) = \mathbb{E}(\alpha x_{n+1} + \beta + \epsilon_{n+1}) = \alpha x_{n+1} + \beta$$

$$\mathbb{E}[Y(x_{n+1})] = \mathbb{E}[A x_{n+1} + B] = \mathbb{E}(A) x_{n+1} + \mathbb{E}(B) = \alpha x_{n+1} + \beta$$

Pour déterminer maintenant la précision de cette estimation, il faut caractériser l'erreur de prévision $Y_{n+1} - Y(x_{n+1})$ en calculant sa variance. Ecrivons pour cela que :

$$\begin{aligned} Y_{n+1} - Y(x_{n+1}) &= (Y_{n+1} - y(x_{n+1})) - (Y(x_{n+1}) - y(x_{n+1})) \\ &= \epsilon_{n+1} - (Y(x_{n+1}) - y(x_{n+1})) \end{aligned}$$

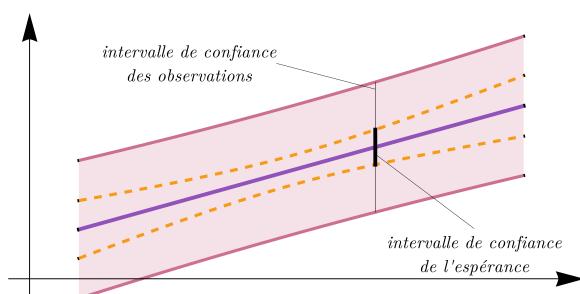
Les deux quantités ϵ_{n+1} et $Y(x_{n+1})$ sont indépendantes puisque la seconde ne fait intervenir que les n premières observations, alors que la première concerne la $(n+1)$ ème observation. Et, par conséquent, les variances s'ajoutent :

$$\mathbb{V}[Y_{n+1} - Y(x_{n+1})] = \mathbb{V}(\epsilon_{n+1}) + \mathbb{V}[Y(x_{n+1})] = \sigma^2 + \left(\frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n} \right) \sigma^2$$

Il en résulte que :

$$T = \frac{Y_{n+1} - Y(x_{n+1})}{\sigma^* \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim \mathcal{T}(n-2)$$

ce qui permet de trouver l'intervalle de confiance cherché.



On remarque, et c'est normal, que plus x_{n+1} est éloigné de \bar{x} , plus cet intervalle est grand. Il serait, de toute façon, illusoire et dangereux de prétendre faire des prévisions de $Y(x)$ pour des valeurs de x qui se trouveraient en dehors de l'intervalle de variation des données expérimentales ayant permis de calculer les relations sur lesquelles reposent ces prévisions.

En fait, on simplifie le plus souvent l'expression de l'intervalle de confiance d'une observation en notant que, si x_{n+1} n'est pas trop éloigné de \bar{x} , la quantité $(x_{n+1} - \bar{x})^2$ est généralement négligeable devant la quantité $\sum_{i=1}^n (x_i - \bar{x})^2$, et en admettant que n est suffisamment grand pour que l'on puisse négliger $\frac{1}{n}$ devant 1.

Dans ces conditions, la plage de confiance des observations, au risque α , est comprise entre les deux droites parallèles :

$$y = a x + b \pm t_{\alpha/2} \sigma^*$$

4 Comparaison de deux régressions

Soit deux groupes d'individus, sur lesquels ont été mesurées les valeurs de deux variables Y et X : n_1 individus pour le premier groupe, et n_2 pour le second.

Groupe 1		Groupe 2	
Y	X	Y	X
y_{11}	x_{11}	y_{12}	x_{12}
\vdots	\vdots	\vdots	\vdots
y_{i1}	x_{i1}	$y_{i'2}$	$x_{i'2}$
\vdots	\vdots	\vdots	\vdots
y_{n_1}	x_{n_1}	$y_{n_2} 2$	$x_{n_2} 2$

Désignons la droite des moindres carrés correspondant au premier groupe par :

$$y_1^* = a_1 x + b_1$$

et sa variance résiduelle estimée par σ_1^{*2} .

Désignons la droite des moindres carrés correspondant au second groupe par :

$$y_2^* = a_2 x + b_2$$

et sa variance résiduelle estimée par σ_2^{*2} .

La comparaison va porter successivement sur les *variances*, puis sur les *pentes* et, enfin, sur les *ordonnées à l'origine*. Les tests correspondants étant calqués sur ceux qui ont été mis en oeuvre pour la comparaison de deux populations, nous nous limiterons à leur principe.

4.1 Comparaison des variances

Le test à appliquer est celui de Snedecor, au quotient :

$$f = \frac{\sigma_1^{*2}}{\sigma_2^{*2}}$$

avec $(n_1 - 2)$ degrés de liberté au numérateur, et $(n_2 - 2)$ degrés de liberté au dénominateur.

Si l'hypothèse d'égalité des variances est acceptable ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), on peut adopter comme estimation de la variance commune la quantité :

$$\sigma^{*2} = \frac{(n_1-2)\sigma_1^{*2}+(n_2-2)\sigma_2^{*2}}{(n_1-2)+(n_2-2)}$$

4.2 Comparaison des pentes

a_1 est une réalisation de la variable aléatoire A_1 de moyenne α_1 et de variance $\frac{\sigma^2}{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2}$.

a_2 est une réalisation de la variable aléatoire A_2 de moyenne α_2 et de variance $\frac{\sigma^2}{\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2}$.

Sous l'hypothèse $\alpha_1 = \alpha_2 = \alpha$ la variable aléatoire $(A_1 - A_2)$ suit une loi normale de moyenne nulle et de variance :

$$\mathbb{V}(A_1 - A_2) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2} + \frac{1}{\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2} \right)$$

Donc :

$$T = \frac{A_1 - A_2}{\sigma^* \sqrt{\frac{1}{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2} + \frac{1}{\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2}}} \sim \mathcal{T}(n_1 + n_2 - 4)$$

ce qui permet de tester l'égalité des pentes.

4.3 Comparaison des ordonnées à l'origine

La même démarche que ci-dessus permet d'établir que, sous l'hypothèse $\beta_1 = \beta_2 = \beta$:

$$T = \frac{B_1 - B_2}{\sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{\bar{x}_1^2}{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2} + \frac{\bar{x}_2^2}{\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2}}} \sim \mathcal{T}(n_1 + n_2 - 4)$$

ce qui permet de tester l'égalité des ordonnées à l'origine.

Exercices du chapitre 7

Exercice 1

Le tableau ci-après donne les résultats d'un certain nombre de déterminations de la distance nécessaire (y en mètres) à l'arrêt par freinage d'une automobile lancée à différentes vitesses (x en km/h). Une étude graphique montre que la courbe représentant y en fonction de x est manifestement concave vers les y positifs, mais que si l'on utilise x^2 au lieu de x , la liaison apparaît sensiblement linéaire. Peut-on justifier ce fait par une loi physique ?

Admettant la validité de ce type de liaison entre y et x^2 , on suppose de plus que la vitesse x peut être déterminée avec une grande précision et que les écarts constatés sont dûs à des fluctuations aléatoires de y autour d'une vraie valeur correspondant à une liaison linéaire représentée par l'équation $y = \alpha x^2 + \beta$.

Vitesse (x)	33	49	65	33	79	49	93
Distance (y)	5.3	14.45	20.26	6.5	38.45	11.23	50.42
x^2	1089	2401	4225	1089	6241	2401	8649

$$\Sigma y = 146.61 \quad \Sigma x^2 = 26,095$$

$$\Sigma y^2 = 4836.3019 \quad \Sigma x^4 = 145\,507\,351 \quad \Sigma x^2 y = 836\,155.41$$

- a) Quelle est la meilleure estimation de α et β ? Quelle hypothèse supplémentaire suppose cette estimation?
- b) Déterminer les limites de confiance à 95% pour les estimations de α et β .
- c) Considérant le cas d'une voiture dont la vitesse est de 85 km/h, estimer la valeur moyenne correspondante de y . En donner une limite supérieure au seuil de confiance 99%.
- d) On suppose que pour une voiture se déplaçant à 85 km/h, on observe une distance de freinage $y = 55$ mètres. Cette valeur peut-elle être considérée comme étant, à des fluctuations aléatoires admissibles près, d'accord avec l'équation d'estimation trouvée ?

Exercice 2

On a déterminé sur une série de 16 coulées Thomas la température y du bain d'acier liquide à la fin de l'opération (à l'aide d'un pyromètre à immersion) et la température x du centre de la flamme (à l'aide d'un pyromètre à flamme) juste avant le rabattement du convertisseur. Le tableau ci-dessous donne les résultats obtenus. Les températures sont exprimées en degrés centigrades.

Bain (y)	1610	1590	1600	1600	1593	1570	1608	1580	1592	1608	1612	1606	1595
Flamme (x)	1504	1490	1505	1495	1490	1475	1508	1480	1482	1510	1520	1510	1492
Bain (y)	1590	1597	1618										
Flamme (x)	1485	1495	1515										

- a) Vérifier graphiquement que la régression de y en x peut être considérée comme linéaire.
- b) Estimer l'équation de la droite de régression de y en x et l'écart-type de y lié par x . Avec quelle précision la température de la flamme permet-elle de connaître la température du bain d'acier dans les conditions des essais?
- c) Peut-on considérer que la différence entre y et x ne dépend pas de x ?

$$\Sigma x = 23\,956 \quad \Sigma y = 25\,569$$

$$\Sigma x^2 = 35\,870\,818 \quad \Sigma y^2 = 40\,863\,179 \quad \Sigma x y = 38\,285\,523$$

Exercice 3

Les données ci-dessous sont relatives à des mesures de la limite élastique (y) et de la résistance à la traction (x) en MPa d'alliages d'or destinés à des prothèses dentaires.

x	1148	1638	1678	1292	1422	1285	1152	1357	867	1158	1082	907	752	1115	1307
y	724	1293	1296	925	1078	948	893	1077	550	870	669	517	495	692	1014

x	1528	1357	1405	1127	1073	1308	812	1260	1008	875
y	1282	1007	978	849	670	953	497	798	657	580

On donne les résultats de calculs suivants :

$$m_x = 1196.52 \quad m_y = 852.48$$

$$\Sigma (x - m_x)^2 = 1\,450\,472.24 \quad \Sigma (y - m_y)^2 = 1\,451\,542.24 \quad \Sigma (x - m_x)(y - m_y) = 1\,406\,707.76$$

- En admettant que $\mathbb{E}(Y|x) = \alpha x + \beta$, estimer α et β par la méthode des moindres carrés.
- Calculer les intervalles de confiance à 95% de α et β .
- Estimer la valeur moyenne de la limite élastique pour une résistance égale à 1290 MPa et calculer son intervalle de confiance à 90%.
- Calculer l'intervalle de confiance à 90% pour la limite élastique correspondant à une résistance de 1290 MPa.

Exercice 4

Les données ci-dessous sont relatives à l'étalonnage d'une méthode gravimétrique pour le dosage de la chaux en présence de magnésium. La variable en x est la teneur vraie et la variable en y est la teneur mesurée (en mg).

Vraie (x)	20.	22.5	25.	28.5	31.	35.5	33.5	37.	38.	40.
Mesurée (y)	19.8	22.8	24.5	27.3	31.	35.	35.1	37.1	38.5	39.

- En admettant que $\mathbb{E}(Y|x) = \alpha x + \beta$, estimer α et β par la méthode des moindres carrés.
- Caractériser la précision de la méthode gravimétrique.
- Tester l'hypothèse $\alpha = 1$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.
- Tester l'hypothèse $\beta = 0$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.
- Bâtir et mettre en oeuvre un test permettant de tester simultanément que $\alpha = 1$ et que $\beta = 0$, la probabilité d'accepter l'hypothèse si elle est vraie étant encore égale à 90%.

$$\Sigma x = 311.0 \quad \Sigma y = 310.1$$

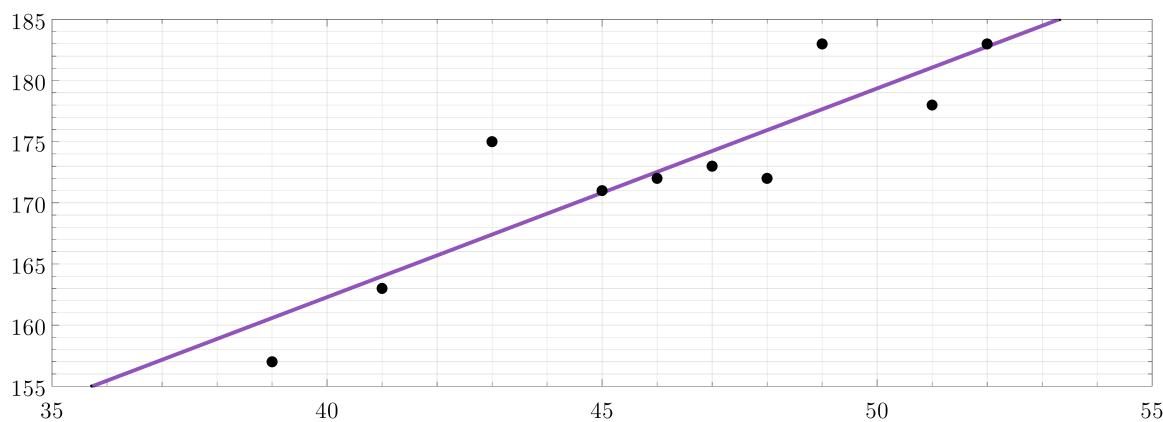
$$\Sigma x^2 = 10\,100.00 \quad \Sigma y^2 = 10\,055.09 \quad \Sigma xy = 10\,074.80$$

Exercice 5

Quand des anthropologues étudient des ossements humains, l'un des points importants est de déterminer la taille des individus. Comme les squelettes sont souvent incomplets, on estime cette taille à partir de mesures sur des petits os. Dans un article intitulé « *The Estimation of Adult Stature from Metacarpal Bone Length* », une équipe de chercheurs a ainsi présenté une méthode permettant d'estimer la taille d'un individu en fonction de la longueur des métacarpes, les os de la paume de main, validée sur les données suivantes où x est la longueur de l'os metacarpal du pouce et y la taille de l'individu.

x (mm)	45	51	39	41	52	48	49	46	43	47
y (mm)	171	178	157	163	183	172	183	172	175	173

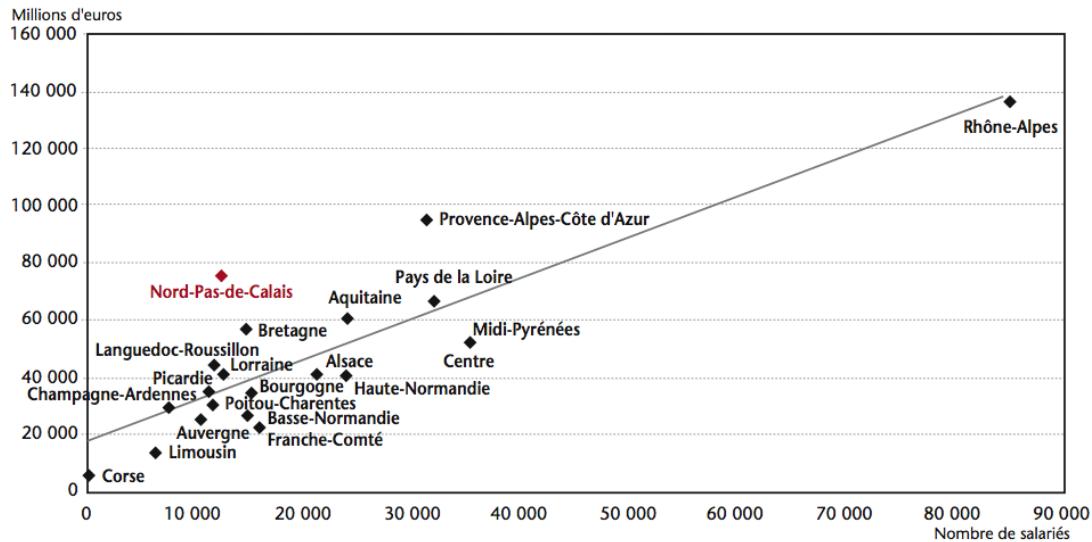
On a représenté ci-après les données et la droite des moindres carrés reliant y à x .



- 1) Calculer les coefficients de la droite des moindres carrés. Vérifiez avec le graphique.
- 2) Pour quel risque minimal, peut-on considérer que la relation entre x et y est significative ?
- 3) Donner l'intervalle de confiance à 95% de la hauteur moyenne des individus dont l'os metacarpal du pouce serait long de 50 mm
- 4) Des éléments anthropologiques complémentaires ont permis d'estimer à 1m90 la taille d'un individu dont l'os metacarpal du pouce est de 50 mm. Que penser de cet individu ?

Exercice 6

La figure suivante indique, pour les 21 régions françaises de province et de métropole, le PIB (y) par région en fonction du nombre d'emplois (x) dans la haute technologie, pour l'année 2000 (source : INSEE Nord-Pas-de-Calais). Le nuage de points, de forme allongée, suggère l'existence d'une relation linéaire (figurée par la droite des moindres carrées) entre ces deux variables.



On donne par ailleurs les résultats intermédiaires suivants :

$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum x \cdot y$
431 200	992 600	15 078 020 000	64 038 160 000	29 144 300 000

- Calculer les coefficients a et b , estimations des paramètres α et β de la relation linéaire ($\alpha x + \beta$) qu'on cherche à mettre en évidence.
- La relation obtenue est elle significative au risque 5% ?
- Pour 12000 emplois de haute technologie, quelle est l'espérance mathématique du PIB et son intervalle de confiance à 95 % ?
- Dans cette étude, la région Nord-Pas-de-Calais affiche un PIB de 76 Milliards d'euros pour environ 12000 emplois de haute technologie. Que pensez de cette région par rapport aux autres ?
- La région Nord-Pas-de-Calais ainsi que la région Provence-Alpes-Côte d'Azur sont en effet assez éloignées du modèle obtenu. Selon vous, quelles raisons structurelles propres à ces régions pourraient expliquer cet écart ?
- Quel défaut présente le modèle de régression choisi ici et comment aurait-on pu le corriger ?

L'expérimentation statistique

Imaginons le cas suivant : un fabricant d'ampoules électriques ayant le choix entre 4 types de filaments se propose d'étudier l'influence de la nature du filament sur la durée de vie des ampoules fabriquées. Pour ce faire, il va faire fabriquer 4 échantillons d'ampoules identiques, sauf en ce qui concerne le filament, faire brûler les ampoules jusqu'à extinction, puis comparer les résultats obtenus. La technique statistique permettant cette comparaison est appelée l'analyse de la variance. Elle se présente comme une technique d'analyse de l'influence d'une variable qualitative appelée facteur (ici, le facteur « filament ») sur une variable quantitative (ici, la durée de vie des lampes). L'objectif du chapitre est de présenter cette technique dans le cas de l'influence d'un facteur, puis de plusieurs facteurs.

1 Analyse de la variance à un facteur

1.1 Recherche de l'influence d'un facteur

Nous noterons A le facteur et appellerons $A_1, \dots, A_j, \dots, A_p$ ses p modalités. Le problème est l'étude de l'influence du facteur A sur la variable quantitative Y . L'expérimentation disponible a consisté à réaliser, pour chaque modalité A_j du facteur, un certain nombre n_j de mesures de la variable Y étudiée de sorte de disposer d'un tableau comme le suivant, où $\bar{y}_1, \dots, \bar{y}_i, \dots, \bar{y}_p$ sont les moyennes des colonnes :

A_1	...	A_i	...	A_p
y_{11}		y_{i1}		y_{p1}
\vdots		\vdots		\vdots
\vdots		y_{ij}		\vdots
$y_{1 n_1}$		\vdots		\vdots
		y_{in_i}		y_{pn_p}
\bar{y}_1	...	\bar{y}_i	...	\bar{y}_p

La technique dite « d'analyse de la variance » que nous allons présenter constitue une extension du test de comparaison de moyennes que nous avons vu au chapitre 5, dans le cas de plus de 2 populations normales.

1.2 La relation d'analyse de la variance

Appelons \bar{y} la moyenne générale des mesures :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij} \text{ avec } n = \sum_{i=1}^p n_i.$$

Effectuons alors la décomposition :

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$$

En élevant au carré et en sommant, le double produit est nul. En effet :

$$2 \sum_{i=1}^p \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i) = 2 \sum_{i=1}^p (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

par définition des moyennes \bar{y}_i .

On obtient donc :

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

relation appelée d'analyse de la variance, qui décompose la somme des carrés totale :

$$\text{SCT} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

en une somme des carrés mesurant la variabilité *intercolonnes* (c'est-à-dire l'influence du facteur) :

$$\text{SCA} = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2$$

et une somme des carrés mesurant la variabilité *intracolonnes* (somme des carrés résiduelle) :

$$\text{SCR} = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Notons la grande généralité de cette relation puisqu'elle a été établie sans faire aucune hypothèse sur les données. Cependant, la structure de la relation de base : $(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$ revient à admettre implicitement l'*additivité* de l'influence du facteur $(\bar{y}_i - \bar{y})$ et d'un résidu $(y_{ij} - \bar{y}_i)$.

On peut associer des degrés de liberté à ces sommes :

- pour SCT : $(n - 1)$ degrés de liberté (nombre de valeurs - 1)
- pour SCA : $(p - 1)$ degrés de liberté (nombres de modalités du facteur - 1)
- pour SCR : $(n - p)$ par différence des valeurs précédentes $(n - 1) - (p - 1) = (n - p)$

1.3 Le modèle

Pour permettre l'inférence statistique, il est nécessaire de poser un certain nombre d'hypothèses. Le modèle de base de l'analyse de la variance s'écrit :

$$Y_i = \mu_i + \varepsilon_i = \mu + \alpha_i + \varepsilon_i$$

Les α_i sont des quantités inconnues, mais certaines, qui mesurent l'influence du facteur A . Pour lever leur indétermination à une constante près, on a l'habitude de poser :

$$\sum_{i=1}^p n_i \alpha_i = 0$$

Les ε_i représentent les fluctuations aléatoires correspondant aux erreurs de mesure ou à l'influence des facteurs non contrôlés. Nous poserons qu'il n'y a pas d'erreur systématique, ou qu'elle est contenue dans μ , donc que $E(\varepsilon_i) = 0$.

Les hypothèses suivantes stipulent que les ε_i :

- sont indépendants : $Cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$,
- ont même variance (homoscédasticité) : $\forall i, V(\varepsilon_i) = \sigma^2$ (homoscédasticité)
- suivent des lois normales.

Parmi ces hypothèses, la plus restrictive est certainement la seconde d'après laquelle l'erreur sur la variable Y est indépendante de la valeur prise par Y , c'est-à-dire notamment, qu'elle n'est pas de type multiplicatif. Pour vérifier si elle est légitime, on dispose de plusieurs tests dont le plus connu est celui de *Bartlett* mais ce dernier est très sensible à l'hypothèse de normalité.

1.4 Test d'analyse de la variance

Faisons l'hypothèse \mathcal{H}_0 que le facteur A n'a pas d'influence sur la variable Y . Cela signifie que les Y_i ont toutes la même moyenne μ et donc que $\alpha_1 = \dots = \alpha_i = \dots = \alpha_p = 0$. Sous \mathcal{H}_0 , on peut alors montrer que la quantité $\frac{SCA}{\sigma^2}$ obéit à une loi du χ^2 à $(p - 1)$ degrés de liberté. Autrement dit que $\frac{SCA}{p-1}$ est une estimation de σ^2 . Comme d'autre part, la quantité $\frac{SCR}{\sigma^2}$ obéit également à une loi du χ^2 à $(n - p)$ degrés de liberté, la quantité $\frac{SCR}{n-p}$ est aussi une estimation de σ^2 . Il en résulte que :

$$f = \frac{SCA}{p-1} / \frac{SCR}{n-p}$$

suit une loi de Snedecor à $(p - 1)$ et $(n - p)$ degrés de liberté et qu'à ce titre, sa valeur est proche de 1.

Si la valeur f calculée est supérieure au seuil f_α lu dans la table de Snedecor, on pourra rejeter l'hypothèse \mathcal{H}_0 et conclure en faveur de l'hypothèse alternative $\mathcal{H}_1 : \exists i, j, \alpha_i \neq \alpha_j$ qui traduit une influence du facteur A . Si elle est inférieure, l'information disponible ne permet pas de conclure à une influence du facteur A . Il importera d'effectuer un *test à droite*. En effet, les faibles valeurs de f correspondent à des différences faibles entre les moyennes \bar{y}_i des colonnes, alors que le test vise à mettre en évidence des différences fortes.

1.5 Calcul pratique

On calcule :

$$SCT = \sum_{i=1}^p \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^p \sum_{j=1}^{n_i} y_{ij}^2 - n \bar{y}^2$$

$$SCA = \sum_{i=1}^p n_i (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^p n_i \bar{y}_i^2 - n \bar{y}^2$$

et enfin par différence :

$$\text{SCR} = \text{SCT} - \text{SCA}$$

On constitue alors le tableau suivant :

Variation	SC	Degrés de liberté	f calculé	F Snedecor
Facteur	SCA	$p-1$	$\frac{\text{SCA}/(p-1)}{\text{SCR}/(n-p)}$	F_α
Résiduelle	SCR	$n-p$		
Totale	SCT	$n-1$		

1.6 Test de linéarité d'une régression

Ce test concerne les problèmes de régression qui ont fait l'objet du chapitre 7, mais il est plus facile de le présenter si les résultats de l'analyse de la variance sont connus.

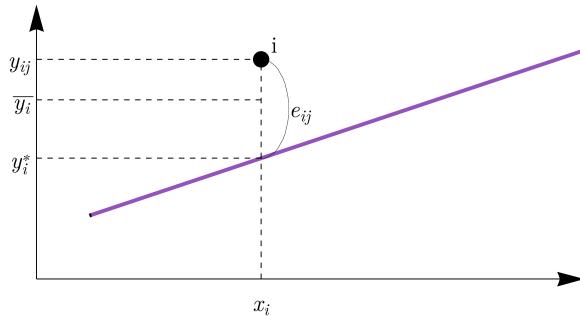
Nous avons supposé dans ce chapitre que la ligne de régression $\mathbb{E}[Y(x)] = f(x)$ était une droite.

Si l'expérimentation a été menée de telle sorte que, pour chaque valeur de la variable explicative X , on dispose de q mesures de la variable expliquée Y , il est possible de tester la linéarité de la régression.

En fait, il n'est pas nécessaire qu'il y ait le même nombre q de mesures pour chaque valeur de X , mais seulement qu'il y en ait plusieurs. Nous nous placerons toutefois ici dans ce cas particulier. On dispose donc du tableau des observations ci-dessous, qui a la même structure que celui d'une analyse de la variance.

x_1	...	x_i	...	x_p
y_{11}		y_{i1}		y_{p1}
\vdots		\vdots		\vdots
\vdots		y_{ij}		\vdots
\vdots		\vdots		\vdots
y_{1q}		y_{iq}		y_{pq}
$\overline{y_1}$...	$\overline{y_i}$...	$\overline{y_p}$

Le principe du test de linéarité consiste à s'assurer que les moyennes $\overline{y_i}$ ne sont pas « trop éloignées » de la droite de régression.



Le déroulement en est le suivant. Soit :

$$y_i^* = a x_i + b$$

le point d'abscisse x_i de la droite des moindres carrés et soit :

$$e_{ij} = (y_{ij} - y_i^*)$$

le résidu correspondant à l'observation y_{ij} . On peut décomposer e_{ij} en :

$$e_{ij} = (\overline{y_i} - y_i^*) + (y_{ij} - \overline{y_i}).$$

En élevant au carré et en sommant sur i et j , le double produit est nul et on obtient :

$$\sum_{i=1}^p \sum_{j=1}^q (y_{ij} - y_i^*)^2 = q \sum_{i=1}^p (\bar{y}_i - y_i^*)^2 + \sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_i)^2$$

Soit S_1^2 le premier terme de la décomposition relatif aux variations de y expliquées par celles de x :

$$S_1^2 = q \sum_{i=1}^p (\bar{y}_i - y_i^*)^2$$

qui est appelé le *défaut d'ajustement*.

Et soit S_2^2 le second terme, qui concerne les variations résiduelles :

$$S_2^2 = \sum_{i=1}^p \sum_{j=1}^q (y_{ij} - \bar{y}_i)^2$$

Si on a bien affaire à une *régression linéaire*, $\frac{S_1^2}{\sigma^2}$ suit une loi du χ^2 à $(p-2)$ degrés de liberté. Comme d'autre part, $\frac{S_2^2}{\sigma^2}$ suit une loi du χ^2 à $p(q-1)$ degrés de liberté et que S_1^2 et S_2^2 sont indépendantes (puisque chaque \bar{y}_i est indépendant de $\sum_{j=1}^q (y_{ij} - \bar{y}_i)^2$), il en résulte que le quotient :

$$f = \frac{\frac{S_1^2}{\sigma^2}}{p-2} / \frac{\frac{S_2^2}{\sigma^2}}{p(q-1)}$$

suit une loi de Snedecor à $(p-2)$ et $p(q-1)$ degrés de liberté. Cette propriété permet de tester la linéarité. Ici encore, c'est un test à droite qu'il faut faire, puisque ce que l'on veut éventuellement montrer c'est une valeur élevée du défaut d'ajustement qui montrerait que les variations de y sont principalement expliquées par les variations de x .

2 Etude de l'influence de deux facteurs

Imaginons que le fabricant d'ampoules évoqué plus haut, se préoccupe d'étudier l'influence, sur la durée de vie des ampoules, non seulement du type de filament utilisé, mais également de la nature du gaz de remplissage.

Il pourrait évidemment faire, d'une part, une première étude « filament » en utilisant l'analyse de la variance à un facteur, puis procéder, d'autre part, à une étude « gaz » en tous points analogue. Cela fait, il lui resterait à rapprocher les résultats de ces deux études pour se faire une idée de l'influence des deux facteurs étudiés. Mais en procédant de la sorte, il postulera implicitement l'additivité des influences « filament » et « gaz », ce qui n'est pas acquis.

L'analyse de la variance à deux facteurs va permettre de traiter globalement le problème, et de mettre en évidence, éventuellement, ce qu'il est convenu d'appeler les *interactions* des facteurs étudiés.

2.1 Plan factoriel

Soit, d'une façon générale, A et B les deux facteurs dont on se propose d'étudier l'influence sur une variable quantitative Y . Nous appellerons $A_1, \dots, A_i, \dots, A_p$ les p modalités du facteur A , et $B_1, \dots, B_j, \dots, B_q$ les q modalités du facteur B . La mise en oeuvre de l'analyse de la variance à deux facteurs nécessite de disposer d'au moins une mesure de Y pour toute combinaison (A_i, B_j) des modalités des facteurs.

Nous admettrons que l'expérimentation a permis de réaliser r répétitions, c'est-à-dire r mesures pour chacune des pq combinaisons des modalités des facteurs. Le cas où il n'y a pas de répétitions ($r = 1$) fera l'objet d'un paragraphe particulier.

Les essais sont donc menés de façon à obtenir le tableau de mesures ci-dessous, une des difficultés de l'expérimentation étant d'éviter les mesures manquantes.

	A_1	...	A_i	...	A_p
B_1	y_{111} ... y_{11r}	...	y_{i11} ... y_{i1r}	...	y_{p11} ... y_{p1r}
\vdots	\vdots		\vdots		\vdots
B_j	y_{1j1} ... y_{1jr}	...	y_{ij1} ... y_{ijk} ... y_{ijr}	...	y_{pj1} ... y_{pjr}
\vdots	\vdots		\vdots		\vdots
B_q	y_{1q1} ... y_{1qr}	...	y_{iq1} ... y_{iqr}	...	y_{pq1} ... y_{pqr}

Le plan d'expérience ainsi réalisé est appelé *plan factoriel*. Il est dit *équilibré* parce qu'il y a le même nombre de mesures dans chaque case du tableau. Il existe d'autres plans d'expérience équilibrés qui évitent le principal inconvénient du plan factoriel, qui est d'être très coûteux du point de vue du nombre de mesures à effectuer.

2.2 Modèle additif et modèle avec interaction

Le modèle le plus général, en admettant l'additivité des erreurs ε_{ij} , est le suivant :

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}$$

En explicitant μ_{ij} , un modèle couramment utilisé est le modèle additif :

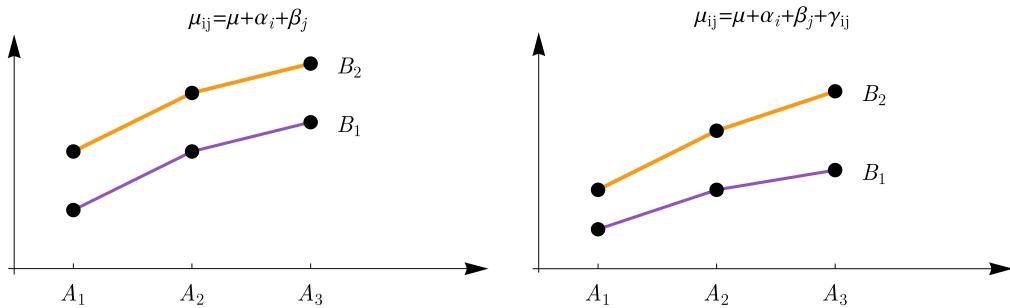
$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

On suppose ainsi qu'il y a *additivité* des effets : l'action conjuguée des modalités A_i et B_j est la somme des actions isolées de A d'une part et de B d'autre part.

Si l'on ne suppose pas réalisée cette hypothèse restrictive d'additivité, on adopte le modèle avec *interaction* :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Il n'y a plus additivité des effets car, aux actions directes de A et B , s'ajoute le terme γ_{ij} qui traduit un effet supplémentaire dû à la conjonction des modalités A_i et B_j .



On dit que $(\alpha_1, \dots, \alpha_p)$ et $(\beta_1, \dots, \beta_q)$ sont les actions des facteurs A et B , tandis que $(\gamma_{11}, \dots, \gamma_{pq})$ sont les interactions du couple (A, B) . On peut encore dire que le modèle avec interaction traduit le fait que l'action du facteur A , par exemple, dépend des modalités du facteur B , comme l'illustrent les figures ci-dessus.

Pour lever l'indétermination de μ , on pose les relations suivantes :

$$\sum_{i=1}^p \alpha_i = 0 ; \sum_{j=1}^q \beta_j = 0 ; \sum_{i=1}^p \gamma_{ij} = 0 \text{ pour tout } j \text{ et } \sum_{j=1}^q \gamma_{ij} = 0 \text{ pour tout } i$$

2.3 Relation d'analyse de la variance

Appelons :

\bar{y}_i la moyenne d'une colonne du tableau des mesures : $\bar{y}_i = \frac{1}{q r} \sum_{j k} y_{ijk}$

\bar{y}_j la moyenne d'une ligne du tableau : $\bar{y}_j = \frac{1}{p r} \sum_{i k} y_{ijk}$

\bar{y}_{ij} la moyenne d'une case du tableau : $\bar{y}_{ij} = \frac{1}{r} \sum_k y_{ijk}$

\bar{y} la moyenne générale des mesures : $\bar{y} = \frac{1}{p q r} \sum_{i j k} y_{ijk}$

Effectuons alors la décomposition :

$$(y_{ijk} - \bar{y}) = (\bar{y}_i - \bar{y}) + (\bar{y}_j - \bar{y}) + [(\bar{y}_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y})] + (y_{ijk} - \bar{y}_{ij})$$

En élévant au carré et en sommant, les doubles produits s'annulent par définition des différentes moyennes, à la condition stricte que le tableau soit complet, c'est-à-dire qu'il n'y ait aucune mesure manquante. On obtient par conséquent :

$$\sum_{i j k} (y_{ijk} - \bar{y})^2 = q r \sum_i (\bar{y}_i - \bar{y})^2 + p r \sum_j (\bar{y}_j - \bar{y})^2 + r \sum_{i j} [(\bar{y}_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y})]^2 + \sum_{i j k} (y_{ijk} - \bar{y}_{ij})^2$$

que nous noterons symboliquement :

$$SCT = SCA + SCB + SCAB + SCR$$

C'est la relation d'analyse de la variance. Elle permet de décomposer la somme des carrés totale en quatre sommes. Les deux premières correspondent respectivement aux *actions* de *A* et de *B*. La troisième correspond à l'*interaction* de *A* et *B*. La dernière est la somme des carrés *résiduelle*.

2.4 Les tests d'analyse de la variance

Admettons, comme dans le cas d'un seul facteur, que les ε_{ij} sont des variables aléatoires centrées, de même variance σ^2 , indépendantes, et qu'elles suivent des lois normales. Il est alors possible d'effectuer une inférence statistique à partir des observations, et de tester :

- la présence d'une interaction,
- l'influence d'un facteur.

2.4.1 Test de l'interaction

Faisons l'hypothèse qu'il n'y a pas d'interaction des facteurs *A* et *B*, c'est-à-dire que :

$$\forall i, j : \gamma_{ij} = 0$$

On montre que, s'il en est ainsi, la quantité $\frac{SCAB}{\sigma^2}$ suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

Comme d'autre part, la quantité $\frac{SCR}{\sigma^2}$ suit une loi du χ^2 à $(n-p-q) = p q (r-1)$ degrés de liberté, il en résulte que le quotient :

$$f_{AB} = \frac{SCAB}{(p-1)(q-1)} / \frac{SCR}{p q (r-1)}$$

suit une loi de Snedecor à $(p-1)(q-1)$ et $p q (r-1)$ degrés de liberté, s'il n'y a pas d'interaction.

2.4.2 Test de l'influence d'un facteur

Faisons l'hypothèse que le facteur A , par exemple, n'a pas d'influence sur la variable Y . On montre alors que la quantité $\frac{SCA}{\sigma^2}$ suit une loi du χ^2 à $(p - 1)$ degrés de liberté.

Par conséquent, la quantité :

$$f_A = \frac{SCA}{(p-1)} / \frac{SCR}{p q(r-1)}$$

suit une loi de Snedecor à $(p - 1)$ et $p q(r - 1)$ degrés de liberté.

2.4.3 Exécution des calculs

On calcule SCA, SCB, SCAB et SCR par les formules suivantes :

$$SCA = q r \sum_i \bar{y}_i^2 - p q r \bar{y}^2$$

$$SCB = p r \sum_j \bar{y}_j^2 - p q r \bar{y}^2$$

$$SCAB = r \sum_{ij} \bar{y}_{ij}^2 - p q r \bar{y}^2 - SCA - SCB$$

$$SCT = \sum_{ijk} y_{ijk}^2 - p q r \bar{y}^2$$

Puis SCR s'obtient par différence :

$$SCR = SCT - SCA - SCB - SCAB$$

On dresse enfin le tableau :

SC	DL	f calculé	F Snedecor
SCA	$p-1$	$f_A = \frac{SCA/(p-1)}{SCR/pq(r-1)}$	F_A
SCB	$q-1$	$f_B = \frac{SCB/(q-1)}{SCR/pq(r-1)}$	F_B
SCAB	$(p-1)(q-1)$	$f_{AB} = \frac{SCAB/(p-1)(q-1)}{SCR/pq(r-1)}$	F_{AB}
SCR	$pq(r-1)$		
SCT	$pqr-1$		

2.5 Analyse de la variance sans répétitions

Supposons qu'on n'ait réalisé qu'une seule mesure y_{ij} pour chaque couple de modalités (A_i, B_j) , conformément au tableau ci-dessous.

	A_1	...	A_i	...	A_p
B_1	y_{11}	...	y_{i1}	...	y_{p1}
\vdots	\vdots		\vdots		\vdots
B_j	y_{1j}	...	y_{ij}	...	y_{pj}
\vdots	\vdots		\vdots		\vdots
B_q	y_{1q}	...	y_{iq}	...	y_{pq}

L'équation d'analyse de la variance s'écrit alors :

$$\sum_{ij} (y_{ij} - \bar{y})^2 = q \sum_i (\bar{y}_i - \bar{y})^2 + p \sum_j (\bar{y}_j - \bar{y})^2 + \sum_{ij} [(y_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y})]^2$$

soit, avec les notations habituelles :

$$SCT = SCA + SCB + SCAB$$

Il devient impossible de tester l'interaction, puisqu'on ne dispose plus d'une quantité telle que SCR permettant, par division, d'éliminer σ^2 et d'obtenir une loi de Snedecor. Il est donc nécessaire dans ce cas de faire l'hypothèse (impossible à vérifier) qu'il n'y a pas d'interaction. On doit donc adopter le modèle additif :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Sous cette condition, et quelles que soient les actions des facteurs A et B , on montre, comme dans le cas général, que $\frac{\text{SCAB}}{\sigma^2}$ suit une loi du χ^2 à $(p-1)(q-1)$ degrés de liberté.

Dès lors, pour tester l'influence de A , par exemple, faisons l'hypothèse que les α_i sont tous nuls. Elle entraîne que $\frac{\text{SCA}}{\sigma^2}$ suit une loi du χ^2 à $(p-1)$ degrés de liberté et, par conséquent, que la quantité :

$$f_A = \frac{\text{SCA}}{(p-1)} / \frac{\text{SCAB}}{(p-1)(q-1)}$$

est la réalisation d'une loi de Snedecor à $(p-1)$ et $(p-1)(q-1)$ degrés de liberté.

Exercices du chapitre 8

Exercice 1

Les données suivantes représentent l'effet du temps T (en heures), sur la perte H (en ppm) de l'hydrogène contenu dans des échantillons d'acier à 20 degrés centigrades.

T	1	2	6	17	30
$\log(T)$	0	0.69	1.79	2.83	3.4
H	7.7	7.5	6.2	5.7	4.2
	8.5	8.1	6.8	5.3	4.6

- a) Le facteur temps a-t-il une influence sur la perte en hydrogène ?
- b) Peut-on admettre que la relation entre H et T est de la forme $H = \alpha \log(T) + \beta$?

Avec $y = H$ et $x = \log(T)$, on donne :

$$\begin{aligned} 2 \sum_i x_i &= 17.42 & 2 \sum_i x_i^2 &= 46.4982 \\ \sum_{i,j} y_{ij} &= 64.6 & \sum_{i,j} y_{ij}^2 &= 437.46 \\ 2 \sum_i \bar{y}_i^2 &= 436.62 & \sum_{i,j} x_i y_{ij} &= 95.084 \end{aligned}$$

Exercice 2

Un laboratoire utilise 4 thermomètres de façon interchangeable pour faire des mesures de température. Pour étudier si les résultats diffèrent suivant les thermomètres, ces derniers ont été placés dans un récipient maintenu à température constante. Trois lectures ont été faites avec chaque thermomètre. Les résultats en degrés centigrades ont été les suivants :

Thermo 1	Thermo 2	Thermo 3	Thermo 4
0.9	0.3	-0.6	0
1.2	-0.2	-1.	0.4
0.8	0.1	-0.7	0.5

Que peut-on en conclure ? On donne : $\sum_{i,j} y_{ij} = 1.7$; $\sum_{i,j} y_{ij}^2 = 5.29$; $3 \sum_i \bar{y}_i^2 = 4.85$.

Exercice 3

On fait passer un test de connaissance à 16 personnes novices ou expertes du domaine en objet. Le test est calibré pour donner une loi normale. On obtient les résultats suivants :

	hommes				femmes			
	Novices	15	16	17	18	26	27	28
Experts	20	21	22	23	31	32	33	34

avec $\sum y = 392$ et $\sum y^2 = 10\,208$.

Peut-on conclure à une influence du genre et du niveau d'expertise sur les résultats aux tests ?

Exercice 4

Un ciment est caractérisé par sa résistance à la compression (mesurée sur des prismes de béton fabriqués à partir d'un échantillon du ciment). Pour étudier la variabilité des productions journalières d'un four à ciment, on a réalisé deux mesures sur des échantillons prélevés chaque jour pendant une période de 10 jours. On a obtenu les résultats ci-après.

	A (i)									
Jour	1	2	3	4	5	6	7	8	9	10
Mesure 1	290	294	281	273	271	270	274	266	264	272
Mesure 2	298	273	272	285	318	311	290	315	259	300

$$\sum_{i,j} y_{ij} = 5676 \quad \sum_{i,j} y_{ij}^2 = 1\,616\,672 \quad \sum_i \bar{y}_i^2 = 806\,315$$

Que conclure de ces résultats ?

Exercice 5

Les produits appelés « cru », à l'entrée d'un four de cimenterie, sont ajustés en tenant compte du titre en carbonates. On veut savoir avec quelle précision est connu ce titre dans le cadre de l'usine. Le laboratoire dispose de cinq chimistes. Il y a deux batteries de dosage.

L'expérimentation a été menée dans le but de mettre en évidence une influence possible de l'opérateur ou de la batterie sur le résultat d'un dosage. Le tableau suivant indique les résultats pour 10 crus, chacun d'eux ayant fait l'objet d'une analyse par chaque opérateur et sur chaque batterie. Dans chaque case du tableau, le premier nombre correspond à la batterie n°1 et le second à la batterie n°2.

Opérateur	Cru									
	1	2	3	4	5	6	7	8	9	10
B	77.25	79.25	78.50	87.50	82.75	84.25	80.80	77.65	78.80	85.90
	77.10	79.45	78.55	87.30	82.90	84.05	81.05	77.70	78.55	86.05
C	77.35	79.60	78.80	87.4	83.05	84.25	81.10	78.00	78.70	86.20
	77.30	79.70	78.75	87.45	83.05	84.20	81.00	78.00	78.75	86.00
G	77.15	79.40	78.60	87.50	83.15	84.45	81.20	78.00	78.60	86.10
	77.40	79.75	78.75	87.55	83.15	84.45	81.20	78.05	78.70	86.45
N	77.05	79.10	78.40	87.10	82.95	84.25	80.90	77.85	78.40	85.90
	77.15	79.40	78.45	87.05	83.15	84.15	81.05	77.90	79.05	86.30
P	77.15	79.40	78.60	87.20	82.95	84.15	81.10	77.95	78.90	86.20
	77.25	79.55	78.45	87.30	82.90	84.05	81.00	77.85	78.75	86.05

Analyser les résultats. On donne les sommes de carrés correspondants aux actions des facteurs et à leurs interactions doubles et triple :

SCT	1169.92	SCAB	0.3611
SCA	1167.78	SCAC	0.1969
SCB	0.4374	SCBC	0.5224
SCC	0.0016	SCABC	0.6141

A : Cru
B : Opérateur
C : Batterie

Exercice 6

On veut connaître l'influence des facteurs de fabrication sur la qualité d'un ciment et, plus particulièrement, l'influence sur la cuisson des trois facteurs : qualité de l'alumine, température de cuisson, temps de palier de cuisson.

Les critères de cuisson sont : la teneur en alumine libre et la teneur en chaux libre des produits à la sortie du four.

Lors des essais effectués, six crus ont été testés (cru industriel témoin, cru AH₃, cru Guilini, cru A6000, cru A8000, cru A9000), à trois températures différentes (1350°, 1400°, 1450°). A chaque température, les temps de palier étaient de 10, 20 et 30 mn. On a obtenu les résultats suivants :

Températures		1350°		1400°		1450°	
Temps	Crus	Al ₂ O ₃	CaO libre	Al ₂ O ₃	CaO libre	Al ₂ O ₃	CaO libre
10 mn	Témoin	11.20	1.10	6.10	0.40	3.55	0.10
	AH ₃	10.75	0.75	4.30	0.20	3.65	0.05
	Guilini	8.60	0.25	5.45	0.10	4.50	0.05
	A6000	10.25	1.55	5.45	0.25	1.75	0.05
	A8000	7.40	2.55	3.20	0.85	0.85	0.05
	A9000	4.70	0.95	2.20	0.20	1.70	0.07
20 mn	Témoin	9.50	0.65	5.00	0.20	1.55	0.10
	AH ₃	14.10	1.20	3.55	0.20	2.35	0.15
	Guilini	7.60	0.15	5.30	0.20	4.00	0.15
	A6000	9.40	1.30	5.60	0.25	2.45	0.10
	A8000	5.70	1.50	2.05	0.85	0.95	0.07
	A9000	3.20	0.45	1.55	0.20	0.50	0.08
30 mn	Témoin	8.25	0.70	5.15	0.15	2.65	0.00
	AH ₃	15.55	1.80	2.20	0.10	2.75	0.08
	Guilini	7.80	0.20	4.65	0.05	3.00	0.06
	A6000	7.75	0.65	4.70	0.30	2.60	0.05
	A8000	3.55	1.05	2.05	0.55	0.80	0.05
	A9000	3.00	0.35	1.60	0.10	0.65	0.08

Analyser ces résultats. On donne les éléments de calcul :

Somme	Alumine	Chaux
SCA	345.95	7.56
SCB	145.99	2.59
SCC	8.22	0.28
SCAB	87.60	2.47
SCAC	0.95	0.32
SCBC	7.11	0.75
SCABC	23.47	1.29

A : Température (3 modalités)

B : Cru (6 modalités)

C : Temps de palier (3 modalités)

Tables numériques

Les tables qui suivent ont été obtenues avec le logiciel Mathematica™. Elles permettent d'obtenir, sans calcul, les quantités utiles à la résolution de la plupart des problèmes de statistique.

Nombres au Hasard

40840	59407	23813	51798	77383	76786	80760	36192	83109	05831	67819	77840	51342	42336
25113	09335	99962	37540	07171	80470	91142	76829	30521	27392	42447	42705	57863	92309
19511	91273	54426	66783	71576	14646	04230	13996	85155	97856	42004	72213	85576	83611
65459	73251	66417	40155	22827	02659	45045	06068	61533	54624	07034	67071	83735	67892
32277	02009	09224	49674	89731	35382	60844	83661	98924	70767	96252	30957	38157	78995
70524	43539	33072	05072	96867	62153	28252	85214	15324	99731	96380	92053	06145	66622
91813	46012	45757	23928	58661	18723	72863	82719	09873	41918	26396	64493	39643	18849
26195	81110	50291	94493	96461	36986	92568	82468	65001	53031	56221	41537	54722	52439
16172	99566	01889	15658	77862	84130	83809	83833	89378	75825	15236	87400	74159	46191
59472	26044	59862	51313	43517	56198	19228	18806	19648	13325	82325	11644	70299	22966
48274	18139	59792	13366	83517	76830	55294	10304	51646	81055	49629	48551	09807	50798
72769	75057	84650	38918	89311	02434	99296	35605	76754	70342	57794	75449	87137	38134
48134	90926	40991	92198	39295	08890	84300	49495	29571	90697	47603	79466	42697	54382
88979	65869	07791	51518	98801	61923	62851	10417	89495	88379	48994	02367	70955	04423
10545	52322	34141	65328	29461	35825	53746	14255	97454	56117	94407	13374	16271	18780
38310	98230	29803	29581	86515	62025	90591	99173	40062	73328	77749	38432	19550	39884
95826	71102	18264	06524	81361	37119	54241	63989	95612	03211	33165	16511	77958	32634
89254	57523	29831	75432	53354	51556	77972	23839	93209	87358	85460	42105	11642	38218
22040	39913	74266	43643	03079	02414	10337	30157	05077	35863	64666	88997	13123	95418
21817	19263	24930	12732	26709	02396	42839	86775	63724	89686	99621	35564	94347	71370
12613	01989	08312	89318	43276	38949	06961	49371	81437	68994	87079	41197	42080	55701
21796	88513	09092	15441	11088	24754	70931	25047	59892	08106	36971	08650	95797	88766
48491	84711	44711	45059	04381	70177	75328	52937	95813	10083	46508	41253	60284	47217
01550	17480	75819	50937	11796	52343	40478	65417	73057	22578	77223	72354	47132	41527
51915	60054	58730	87599	74294	71192	19973	31575	67008	78823	04584	14372	75608	15446
44700	55006	23359	41084	26515	76964	95846	71325	39996	80998	09176	73893	19242	20292
95876	87497	97883	38743	06630	53281	21770	09778	72662	69124	79753	70815	79083	61419
68021	32495	40025	72656	83834	62676	75137	59389	23077	57070	14050	71506	10197	85740
49503	56667	98977	32887	46833	74266	36412	69399	58861	20449	73134	97131	27531	82726
16120	45822	17976	73159	30298	98341	72067	85062	56112	70708	54530	77835	31295	24416
50764	24053	72168	56384	14660	15942	93523	84986	03831	52727	78922	15882	22495	53836
76495	46901	87360	52010	62256	58859	91823	27294	08532	39145	06521	35998	68669	00714
60272	32764	34507	87421	93667	26512	48055	94515	79302	14306	91569	05658	33219	58570
01848	94995	63862	96008	71226	93271	23677	58647	82861	46202	14482	75696	97826	65611
21412	50664	60369	38070	31213	28581	74427	06245	52821	04943	53694	37347	58031	73488
13494	67332	92798	79052	40510	49549	90496	82663	43707	71966	39822	43820	86494	02575
25449	46336	93529	98835	08127	74018	70697	05506	05898	29118	31120	82752	25568	86299
93178	37015	53250	26468	94274	59613	26612	64124	71149	33704	73437	40821	51188	33462
92805	35636	32476	22063	64062	74167	87303	37700	82530	33428	33061	56853	43944	55256
71814	88492	35907	74118	64727	29519	70093	58917	21647	81974	20682	78840	46263	90902
47580	85469	12736	78259	09720	31633	13194	48718	57736	79269	48826	53378	69878	94558
40991	17271	92306	34090	42963	91423	99386	32273	14121	16067	13722	51552	04563	09102
01294	11328	93910	83939	28974	19822	17671	73909	72966	79598	09311	81834	47192	94287
81582	18086	92908	42069	74414	39356	18345	53803	32090	78031	04708	34326	86573	69697
31404	97747	53791	23284	35185	80045	79959	78157	25956	66926	44017	76860	72758	75635
08171	34589	74654	76726	11763	72104	73972	16678	15891	16478	49080	02794	82250	61219
22638	93636	72456	60669	27699	88376	35456	28179	90840	35321	53194	14992	27844	18166
89516	23878	00864	79351	28035	14482	59907	28262	54034	31674	26018	76875	51151	66477
47556	86442	00554	41459	41412	53460	72762	19301	63266	23057	17095	13513	95348	70216
65624	07525	94251	08505	22022	44039	08569	34514	12226	76388	40431	57841	79064	14568

Loi Binomiale

$$\text{Fonction de répartition } P_N(k) = \sum_{i=0}^k \binom{N}{i} \varpi^i (1-\varpi)^{N-i}$$

Taille de l'échantillon : $N = 5$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.9510	0.9039	0.8587	0.8154	0.7738	0.7339	0.6957	0.6591	0.6240	0.5905	0.3277	0.1681	0.0778	0.0312
1	0.9990	0.9962	0.9915	0.9852	0.9774	0.9681	0.9575	0.9456	0.9326	0.9185	0.7373	0.5282	0.3370	0.1875
2	1	0.9999	0.9997	0.9994	0.9988	0.9980	0.9969	0.9955	0.9937	0.9914	0.9421	0.8369	0.6826	0.5000
3	1	1	1	1	1	0.9999	0.9999	0.9998	0.9997	0.9995	0.9933	0.9692	0.9130	0.8125
4	1	1	1	1	1	1	1	1	1	1	0.9997	0.9976	0.9898	0.9688
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 10$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.9044	0.8171	0.7374	0.6648	0.5987	0.5386	0.4840	0.4344	0.3894	0.3487	0.1074	0.0282	0.0060	0.0010
1	0.9957	0.9838	0.9655	0.9418	0.9139	0.8824	0.8483	0.8121	0.7746	0.7361	0.3758	0.1493	0.0464	0.0107
2	0.9999	0.9991	0.9972	0.9938	0.9885	0.9812	0.9717	0.9599	0.9460	0.9298	0.6778	0.3828	0.1673	0.0547
3	1	1	0.9999	0.9996	0.9990	0.9980	0.9964	0.9942	0.9912	0.9872	0.8791	0.6496	0.3823	0.1719
4	1	1	1	1	0.9999	0.9998	0.9997	0.9994	0.9990	0.9984	0.9672	0.8497	0.6331	0.3770
5	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9936	0.9527	0.8338	0.6230
6	1	1	1	1	1	1	1	1	1	0.9991	0.9894	0.9452	0.8281	
7	1	1	1	1	1	1	1	1	1	0.9999	0.9984	0.9877	0.9453	
8	1	1	1	1	1	1	1	1	1	1	0.9999	0.9983	0.9893	
9	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9990	
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 15$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.8601	0.7386	0.6333	0.5421	0.4633	0.3953	0.3367	0.2863	0.2430	0.2059	0.0352	0.0047	0.0005	-
1	0.9904	0.9647	0.9270	0.8809	0.8290	0.7738	0.7168	0.6597	0.6035	0.5490	0.1671	0.0353	0.0052	0.0005
2	0.9996	0.9970	0.9906	0.9797	0.9638	0.9429	0.9171	0.8870	0.8531	0.8159	0.3980	0.1268	0.0271	0.0037
3	1	0.9998	0.9992	0.9976	0.9945	0.9896	0.9825	0.9727	0.9601	0.9444	0.6482	0.2969	0.0905	0.0176
4	1	1	0.9999	0.9998	0.9994	0.9986	0.9972	0.9950	0.9918	0.9873	0.8358	0.5155	0.2173	0.0592
5	1	1	1	1	0.9999	0.9999	0.9997	0.9993	0.9987	0.9978	0.9389	0.7216	0.4032	0.1509
6	1	1	1	1	1	1	1	0.9999	0.9998	0.9997	0.9819	0.8689	0.6098	0.3036
7	1	1	1	1	1	1	1	1	1	1	0.9958	0.9500	0.7869	0.5000
8	1	1	1	1	1	1	1	1	1	1	0.9992	0.9848	0.9050	0.6964
9	1	1	1	1	1	1	1	1	1	1	0.9999	0.9963	0.9662	0.8491
10	1	1	1	1	1	1	1	1	1	1	0.9993	0.9907	0.9408	
11	1	1	1	1	1	1	1	1	1	1	0.9999	0.9981	0.9824	
12	1	1	1	1	1	1	1	1	1	1	0.9997	0.9963		
13	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9995
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 20$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.8179	0.6676	0.5438	0.4420	0.3585	0.2901	0.2342	0.1887	0.1516	0.1216	0.0115	0.0008	—	—
1	0.9831	0.9401	0.8802	0.8103	0.7358	0.6605	0.5869	0.5169	0.4516	0.3917	0.0692	0.0076	0.0005	—
2	0.9990	0.9929	0.9790	0.9561	0.9245	0.8850	0.8390	0.7879	0.7334	0.6769	0.2061	0.0355	0.0036	0.0002
3	1	0.9994	0.9973	0.9926	0.9841	0.9710	0.9529	0.9294	0.9007	0.8670	0.4114	0.1071	0.0160	0.0013
4	1	1	0.9997	0.9990	0.9974	0.9944	0.9893	0.9817	0.9710	0.9568	0.6296	0.2375	0.0510	0.0059
5	1	1	1	0.9999	0.9997	0.9991	0.9981	0.9962	0.9932	0.9887	0.8042	0.4164	0.1256	0.0207
6	1	1	1	1	1	0.9999	0.9997	0.9994	0.9987	0.9976	0.9133	0.6080	0.2500	0.0577
7	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9679	0.7723	0.4159	0.1316	
8	1	1	1	1	1	1	1	1	1	0.9999	0.9900	0.8867	0.5956	0.2517
9	1	1	1	1	1	1	1	1	1	1	0.9974	0.9520	0.7553	0.4119
10	1	1	1	1	1	1	1	1	1	1	0.9994	0.9829	0.8725	0.5881
11	1	1	1	1	1	1	1	1	1	1	0.9999	0.9949	0.9435	0.7483
12	1	1	1	1	1	1	1	1	1	1	1	0.9987	0.9790	0.8684
13	1	1	1	1	1	1	1	1	1	1	1	0.9997	0.9935	0.9423
14	1	1	1	1	1	1	1	1	1	1	1	1	0.9984	0.9793
15	1	1	1	1	1	1	1	1	1	1	1	1	0.9997	0.9941
16	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9987
17	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9998
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 30$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.7397	0.5455	0.4010	0.2939	0.2146	0.1563	0.1134	0.0820	0.0591	0.0424	0.0012	—	—	—
1	0.9639	0.8795	0.7731	0.6612	0.5535	0.4555	0.3694	0.2958	0.2343	0.1837	0.0105	0.0003	—	—
2	0.9967	0.9783	0.9399	0.8831	0.8122	0.7324	0.6487	0.5654	0.4855	0.4114	0.0442	0.0021	—	—
3	0.9998	0.9971	0.9881	0.9694	0.9392	0.8974	0.8450	0.7842	0.7175	0.6474	0.1227	0.0093	0.0003	—
4	1	0.9997	0.9982	0.9937	0.9844	0.9685	0.9447	0.9126	0.8723	0.8245	0.2552	0.0302	0.0015	—
5	1	1	0.9998	0.9989	0.9967	0.9921	0.9838	0.9707	0.9519	0.9268	0.4275	0.0766	0.0057	0.0002
6	1	1	1	0.9999	0.9994	0.9983	0.9960	0.9918	0.9848	0.9742	0.6070	0.1595	0.0172	0.0007
7	1	1	1	1	0.9999	0.9997	0.9992	0.9980	0.9959	0.9922	0.7608	0.2814	0.0435	0.0026
8	1	1	1	1	1	1	0.9999	0.9996	0.9990	0.9980	0.8713	0.4315	0.0940	0.0081
9	1	1	1	1	1	1	1	0.9999	0.9998	0.9995	0.9389	0.5888	0.1763	0.0214
10	1	1	1	1	1	1	1	1	1	0.9999	0.9744	0.7304	0.2915	0.0494
11	1	1	1	1	1	1	1	1	1	1	0.9905	0.8407	0.4311	0.1002
12	1	1	1	1	1	1	1	1	1	1	0.9969	0.9155	0.5785	0.1808
13	1	1	1	1	1	1	1	1	1	1	0.9991	0.9599	0.7145	0.2923
14	1	1	1	1	1	1	1	1	1	1	0.9998	0.9831	0.8246	0.4278
15	1	1	1	1	1	1	1	1	1	1	0.9999	0.9936	0.9029	0.5722
16	1	1	1	1	1	1	1	1	1	1	1	0.9979	0.9519	0.7077
17	1	1	1	1	1	1	1	1	1	1	1	0.9994	0.9788	0.8192
18	1	1	1	1	1	1	1	1	1	1	1	0.9998	0.9917	0.8998
19	1	1	1	1	1	1	1	1	1	1	1	1	0.9971	0.9506
20	1	1	1	1	1	1	1	1	1	1	1	1	0.9991	0.9786
21	1	1	1	1	1	1	1	1	1	1	1	1	0.9998	0.9919
22	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9974
23	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9993
24	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9998
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 40$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.6690	0.4457	0.2957	0.1954	0.1285	0.0842	0.0549	0.0356	0.0230	0.0148	0.0001	—	—	—
1	0.9393	0.8095	0.6615	0.5210	0.3991	0.2990	0.2201	0.1594	0.1140	0.0805	0.0015	—	—	—
2	0.9925	0.9543	0.8822	0.7855	0.6767	0.5665	0.4625	0.3694	0.2894	0.2228	0.0079	0.0001	—	—
3	0.9993	0.9918	0.9686	0.9252	0.8619	0.7827	0.6937	0.6007	0.5092	0.4231	0.0285	0.0006	—	—
4	1	0.9988	0.9933	0.9790	0.9520	0.9104	0.8546	0.7868	0.7103	0.6290	0.0759	0.0026	—	—
5	1	0.9999	0.9988	0.9951	0.9861	0.9691	0.9419	0.9033	0.8535	0.7937	0.1613	0.0086	0.0001	—
6	1	1	0.9998	0.9990	0.9966	0.9909	0.9801	0.9624	0.9361	0.9005	0.2859	0.0238	0.0006	—
7	1	1	1	0.9998	0.9993	0.9977	0.9942	0.9873	0.9758	0.9581	0.4371	0.0553	0.0021	—
8	1	1	1	1	0.9999	0.9995	0.9985	0.9963	0.9919	0.9845	0.5931	0.1110	0.0061	0.0001
9	1	1	1	1	1	0.9999	0.9997	0.9990	0.9976	0.9949	0.7318	0.1959	0.0156	0.0003
10	1	1	1	1	1	1	0.9999	0.9998	0.9994	0.9985	0.8392	0.3087	0.0352	0.0011
11	1	1	1	1	1	1	1	1	0.9999	0.9996	0.9125	0.4406	0.0709	0.0032
12	1	1	1	1	1	1	1	1	1	0.9999	0.9568	0.5772	0.1285	0.0083
13	1	1	1	1	1	1	1	1	1	1	0.9806	0.7032	0.2112	0.0192
14	1	1	1	1	1	1	1	1	1	1	0.9921	0.8074	0.3174	0.0403
15	1	1	1	1	1	1	1	1	1	1	0.9971	0.8849	0.4402	0.0769
16	1	1	1	1	1	1	1	1	1	1	0.9990	0.9367	0.5681	0.1341
17	1	1	1	1	1	1	1	1	1	1	0.9997	0.9680	0.6885	0.2148
18	1	1	1	1	1	1	1	1	1	1	0.9999	0.9852	0.7911	0.3179
19	1	1	1	1	1	1	1	1	1	1	1	0.9937	0.8702	0.4373
20	1	1	1	1	1	1	1	1	1	1	1	0.9976	0.9256	0.5627
21	1	1	1	1	1	1	1	1	1	1	1	0.9991	0.9608	0.6821
22	1	1	1	1	1	1	1	1	1	1	1	0.9997	0.9811	0.7852
23	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9917	0.8659
24	1	1	1	1	1	1	1	1	1	1	1	1	0.9966	0.9231
25	1	1	1	1	1	1	1	1	1	1	1	1	0.9988	0.9597
26	1	1	1	1	1	1	1	1	1	1	1	1	0.9996	0.9808
27	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9917
28	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9968
29	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9989
30	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9997
31	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9999
32	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Taille de l'échantillon : $N = 50$

k	$\varpi=1\%$	$\varpi=2\%$	$\varpi=3\%$	$\varpi=4\%$	$\varpi=5\%$	$\varpi=6\%$	$\varpi=7\%$	$\varpi=8\%$	$\varpi=9\%$	$\varpi=10\%$	$\varpi=20\%$	$\varpi=30\%$	$\varpi=40\%$	$\varpi=50\%$
0	0.6050	0.3642	0.2181	0.1299	0.0769	0.0453	0.0266	0.0155	0.0090	0.0052	—	—	—	—
1	0.9106	0.7358	0.5553	0.4005	0.2794	0.1900	0.1265	0.0827	0.0532	0.0338	0.0002	—	—	—
2	0.9862	0.9216	0.8108	0.6767	0.5405	0.4162	0.3108	0.2260	0.1605	0.1117	0.0013	—	—	—
3	0.9984	0.9822	0.9372	0.8609	0.7604	0.6473	0.5327	0.4253	0.3303	0.2503	0.0057	—	—	—
4	0.9999	0.9968	0.9832	0.9510	0.8964	0.8206	0.7290	0.6290	0.5277	0.4312	0.0185	0.0002	—	—
5	1	0.9995	0.9963	0.9856	0.9622	0.9224	0.8650	0.7919	0.7072	0.6161	0.0480	0.0007	—	—
6	1	0.9999	0.9993	0.9964	0.9882	0.9711	0.9417	0.8981	0.8404	0.7702	0.1034	0.0025	—	—
7	1	1	0.9999	0.9992	0.9968	0.9906	0.9780	0.9562	0.9232	0.8779	0.1904	0.0073	0.0001	—
8	1	1	1	0.9999	0.9992	0.9973	0.9927	0.9833	0.9672	0.9421	0.3073	0.0183	0.0002	—
9	1	1	1	1	0.9998	0.9993	0.9978	0.9944	0.9875	0.9755	0.4437	0.0402	0.0008	—
10	1	1	1	1	1	0.9998	0.9994	0.9983	0.9957	0.9906	0.5836	0.0789	0.0022	—
11	1	1	1	1	1	1	0.9999	0.9995	0.9987	0.9968	0.7107	0.1390	0.0057	—
12	1	1	1	1	1	1	1	0.9999	0.9996	0.9990	0.8139	0.2229	0.0133	0.0002
13	1	1	1	1	1	1	1	1	0.9999	0.9997	0.8894	0.3279	0.0280	0.0005
14	1	1	1	1	1	1	1	1	0.9999	0.9393	0.4468	0.0540	0.0013	—
15	1	1	1	1	1	1	1	1	1	0.9692	0.5692	0.0955	0.0033	—
16	1	1	1	1	1	1	1	1	1	0.9856	0.6839	0.1561	0.0077	—
17	1	1	1	1	1	1	1	1	1	0.9937	0.7822	0.2369	0.0164	—
18	1	1	1	1	1	1	1	1	1	0.9975	0.8594	0.3356	0.0325	—
19	1	1	1	1	1	1	1	1	1	0.9991	0.9152	0.4465	0.0595	—
20	1	1	1	1	1	1	1	1	1	0.9997	0.9522	0.5610	0.1013	—
21	1	1	1	1	1	1	1	1	1	0.9999	0.9749	0.6701	0.1611	—
22	1	1	1	1	1	1	1	1	1	1	0.9877	0.7660	0.2399	—
23	1	1	1	1	1	1	1	1	1	1	0.9944	0.8438	0.3359	—
24	1	1	1	1	1	1	1	1	1	1	0.9976	0.9022	0.4439	—
25	1	1	1	1	1	1	1	1	1	1	0.9991	0.9427	0.5561	—
26	1	1	1	1	1	1	1	1	1	1	0.9997	0.9686	0.6641	—
27	1	1	1	1	1	1	1	1	1	1	0.9999	0.9840	0.7601	—
28	1	1	1	1	1	1	1	1	1	1	1	0.9924	0.8389	—
29	1	1	1	1	1	1	1	1	1	1	1	0.9966	0.8987	—
30	1	1	1	1	1	1	1	1	1	1	1	0.9986	0.9405	—
31	1	1	1	1	1	1	1	1	1	1	1	0.9995	0.9675	—
32	1	1	1	1	1	1	1	1	1	1	1	0.9998	0.9836	—
33	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9923
34	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9967
35	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9987
36	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9995
37	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9998
38	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Loi de Poisson

Fonction de répartition $P(k) = \sum_{i=0}^k e^{-\lambda} \frac{\lambda^i}{i!}$

k	$\lambda=0.1$	$\lambda=0.2$	$\lambda=0.3$	$\lambda=0.4$	$\lambda=0.5$	$\lambda=0.6$	$\lambda=0.7$	$\lambda=0.8$	$\lambda=0.9$	$\lambda=1.$	$\lambda=1.1$	$\lambda=1.2$	$\lambda=1.3$	$\lambda=1.4$
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.3329	0.3012	0.2725	0.2466
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.6990	0.6626	0.6268	0.5918
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.9004	0.8795	0.8571	0.8335
3	1	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.9743	0.9662	0.9569	0.9463
4	1	1	1	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9946	0.9923	0.9893	0.9857
5	1	1	1	1	1	1	0.9999	0.9998	0.9997	0.9994	0.9990	0.9985	0.9978	0.9968
6	1	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9997	0.9996	0.9994
7	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9999

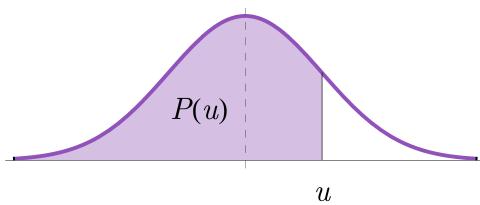
k	$\lambda=1.5$	$\lambda=2.$	$\lambda=2.5$	$\lambda=3.$	$\lambda=3.5$	$\lambda=4.$	$\lambda=4.5$	$\lambda=5.$	$\lambda=5.5$	$\lambda=6.$	$\lambda=6.5$	$\lambda=7.$	$\lambda=7.5$	$\lambda=8.$
0	0.2231	0.1353	0.0821	0.0498	0.0302	0.0183	0.0111	0.0067	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003
1	0.5578	0.4060	0.2873	0.1991	0.1359	0.0916	0.0611	0.0404	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030
2	0.8088	0.6767	0.5438	0.4232	0.3208	0.2381	0.1736	0.1247	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138
3	0.9344	0.8571	0.7576	0.6472	0.5366	0.4335	0.3423	0.2650	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424
4	0.9814	0.9473	0.8912	0.8153	0.7254	0.6288	0.5321	0.4405	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996
5	0.9955	0.9834	0.9580	0.9161	0.8576	0.7851	0.7029	0.6160	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912
6	0.9991	0.9955	0.9858	0.9665	0.9347	0.8893	0.8311	0.7622	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134
7	0.9998	0.9989	0.9958	0.9881	0.9733	0.9489	0.9134	0.8666	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530
8	1	0.9998	0.9989	0.9962	0.9901	0.9786	0.9597	0.9319	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925
9	1	1	0.9997	0.9989	0.9967	0.9919	0.9829	0.9682	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166
10	1	1	0.9999	0.9997	0.9990	0.9972	0.9933	0.9863	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159
11	1	1	1	0.9999	0.9997	0.9991	0.9976	0.9945	0.9890	0.9799	0.9661	0.9467	0.9208	0.8881
12	1	1	1	1	0.9999	0.9997	0.9992	0.9980	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362
13	1	1	1	1	1	0.9999	0.9997	0.9993	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658
14	1	1	1	1	1	1	0.9999	0.9998	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827
15	1	1	1	1	1	1	1	0.9999	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918
16	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963
17	1	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9992	0.9984
18	1	1	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9997	0.9993
19	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9997
20	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9999
21	1	1	1	1	1	1	1	1	1	1	1	1	1	1

k	$\lambda=8.5$	$\lambda=9.$	$\lambda=9.5$	$\lambda=10.$	$\lambda=10.5$	$\lambda=11.$	$\lambda=11.5$	$\lambda=12.$	$\lambda=12.5$	$\lambda=13.$	$\lambda=13.5$	$\lambda=14.$	$\lambda=14.5$	$\lambda=15.$
0	0.0002	0.0001	0.0001	—	—	—	—	—	—	—	—	—	—	—
1	0.0019	0.0012	0.0008	0.0005	0.0003	0.0002	0.0001	0.0001	0.0001	—	—	—	—	—
2	0.0093	0.0062	0.0042	0.0028	0.0018	0.0012	0.0008	0.0005	0.0003	0.0002	0.0001	0.0001	0.0001	—
3	0.0301	0.0212	0.0149	0.0103	0.0071	0.0049	0.0034	0.0023	0.0016	0.0011	0.0007	0.0005	0.0003	0.0002
4	0.0744	0.0550	0.0403	0.0293	0.0211	0.0151	0.0107	0.0076	0.0053	0.0037	0.0026	0.0018	0.0012	0.0009
5	0.1496	0.1157	0.0885	0.0671	0.0504	0.0375	0.0277	0.0203	0.0148	0.0107	0.0077	0.0055	0.0039	0.0028
6	0.2562	0.2068	0.1649	0.1301	0.1016	0.0786	0.0603	0.0458	0.0346	0.0259	0.0193	0.0142	0.0105	0.0076
7	0.3856	0.3239	0.2687	0.2202	0.1785	0.1432	0.1137	0.0895	0.0698	0.0540	0.0415	0.0316	0.0239	0.0180
8	0.5231	0.4557	0.3918	0.3328	0.2794	0.2320	0.1906	0.1550	0.1249	0.0998	0.0790	0.0621	0.0484	0.0374
9	0.6530	0.5874	0.5218	0.4579	0.3971	0.3405	0.2888	0.2424	0.2014	0.1658	0.1353	0.1094	0.0878	0.0699
10	0.7634	0.7060	0.6453	0.5830	0.5207	0.4599	0.4017	0.3472	0.2971	0.2517	0.2112	0.1757	0.1449	0.1185
11	0.8487	0.8030	0.7520	0.6968	0.6387	0.5793	0.5198	0.4616	0.4058	0.3532	0.3045	0.2600	0.2201	0.1848
12	0.9091	0.8758	0.8364	0.7916	0.7420	0.6887	0.6329	0.5760	0.5190	0.4631	0.4093	0.3585	0.3111	0.2676
13	0.9486	0.9261	0.8981	0.8645	0.8253	0.7813	0.7330	0.6815	0.6278	0.5730	0.5182	0.4644	0.4125	0.3632
14	0.9726	0.9585	0.9400	0.9165	0.8879	0.8540	0.8153	0.7720	0.7250	0.6751	0.6233	0.5704	0.5176	0.4657
15	0.9862	0.9780	0.9665	0.9513	0.9317	0.9074	0.8783	0.8444	0.8060	0.7636	0.7178	0.6694	0.6192	0.5681
16	0.9934	0.9889	0.9823	0.9730	0.9604	0.9441	0.9236	0.8987	0.8693	0.8355	0.7975	0.7559	0.7112	0.6641
17	0.9970	0.9947	0.9911	0.9857	0.9781	0.9678	0.9542	0.9370	0.9158	0.8905	0.8609	0.8272	0.7897	0.7489
18	0.9987	0.9976	0.9957	0.9928	0.9885	0.9823	0.9738	0.9626	0.9481	0.9302	0.9084	0.8826	0.8530	0.8195
19	0.9995	0.9989	0.9980	0.9965	0.9942	0.9907	0.9857	0.9787	0.9694	0.9573	0.9421	0.9235	0.9012	0.8752
20	0.9998	0.9996	0.9991	0.9984	0.9972	0.9953	0.9925	0.9884	0.9827	0.9750	0.9649	0.9521	0.9362	0.9170
21	0.9999	0.9998	0.9996	0.9993	0.9987	0.9977	0.9962	0.9939	0.9906	0.9859	0.9796	0.9712	0.9604	0.9469
22	1	0.9999	0.9999	0.9997	0.9994	0.9990	0.9982	0.9970	0.9951	0.9924	0.9885	0.9833	0.9763	0.9673
23	1	1	0.9999	0.9999	0.9998	0.9995	0.9992	0.9985	0.9975	0.9960	0.9938	0.9907	0.9863	0.9805
24	1	1	1	1	0.9999	0.9998	0.9996	0.9993	0.9988	0.9980	0.9968	0.9950	0.9924	0.9888
25	1	1	1	1	1	0.9999	0.9998	0.9997	0.9994	0.9990	0.9984	0.9974	0.9959	0.9938
26	1	1	1	1	1	0.9999	0.9999	0.9997	0.9995	0.9992	0.9987	0.9979	0.9967	0.9967
27	1	1	1	1	1	1	0.9999	0.9999	0.9998	0.9996	0.9994	0.9989	0.9983	0.9983
28	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9997	0.9995	0.9991	0.9991
29	1	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9998	0.9996	0.9996
30	1	1	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9998	0.9998
31	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9999
32	1	1	1	1	1	1	1	1	1	1	1	1	1	1

k	$\lambda=16$	$\lambda=17$	$\lambda=18$	$\lambda=19$	$\lambda=20$	$\lambda=21$	$\lambda=22$	$\lambda=23$	$\lambda=24$	$\lambda=25$	$\lambda=26$	$\lambda=27$	$\lambda=28$	$\lambda=29$
3	0.0001	—	—	—	—	—	—	—	—	—	—	—	—	—
4	0.0004	0.0002	0.0001	—	—	—	—	—	—	—	—	—	—	—
5	0.0014	0.0007	0.0003	0.0002	0.0001	—	—	—	—	—	—	—	—	—
6	0.0040	0.0021	0.0010	0.0005	0.0003	0.0001	0.0001	—	—	—	—	—	—	—
7	0.0100	0.0054	0.0029	0.0015	0.0008	0.0004	0.0002	0.0001	—	—	—	—	—	—
8	0.0220	0.0126	0.0071	0.0039	0.0021	0.0011	0.0006	0.0003	0.0002	0.0001	—	—	—	—
9	0.0433	0.0261	0.0154	0.0089	0.0050	0.0028	0.0015	0.0008	0.0004	0.0002	0.0001	0.0001	—	—
10	0.0774	0.0491	0.0304	0.0183	0.0108	0.0063	0.0035	0.0020	0.0011	0.0006	0.0003	0.0002	0.0001	—
11	0.1270	0.0847	0.0549	0.0347	0.0214	0.0129	0.0076	0.0044	0.0025	0.0014	0.0008	0.0004	0.0002	0.0001
12	0.1931	0.1350	0.0917	0.0606	0.0390	0.0245	0.0151	0.0091	0.0054	0.0031	0.0018	0.0010	0.0006	0.0003
13	0.2745	0.2009	0.1426	0.0984	0.0661	0.0434	0.0278	0.0174	0.0107	0.0065	0.0038	0.0022	0.0013	0.0007
14	0.3675	0.2808	0.2081	0.1497	0.1049	0.0716	0.0477	0.0311	0.0198	0.0124	0.0076	0.0046	0.0027	0.0016
15	0.4667	0.3715	0.2867	0.2148	0.1565	0.1111	0.0769	0.0520	0.0344	0.0223	0.0142	0.0088	0.0054	0.0033
16	0.5660	0.4677	0.3751	0.2920	0.2211	0.1629	0.1170	0.0821	0.0563	0.0377	0.0248	0.0160	0.0101	0.0063
17	0.6593	0.5640	0.4686	0.3784	0.2970	0.2270	0.1690	0.1228	0.0871	0.0605	0.0411	0.0274	0.0179	0.0115
18	0.7423	0.6550	0.5622	0.4695	0.3814	0.3017	0.2325	0.1748	0.1283	0.0920	0.0646	0.0445	0.0300	0.0199
19	0.8122	0.7363	0.6509	0.5606	0.4703	0.3843	0.3060	0.2377	0.1803	0.1336	0.0968	0.0687	0.0478	0.0326
20	0.8682	0.8055	0.7307	0.6472	0.5591	0.4710	0.3869	0.3101	0.2426	0.1855	0.1387	0.1015	0.0727	0.0511
21	0.9108	0.8615	0.7991	0.7255	0.6437	0.5577	0.4716	0.3894	0.3139	0.2473	0.1905	0.1436	0.1060	0.0767
22	0.9418	0.9047	0.8551	0.7931	0.7206	0.6405	0.5564	0.4723	0.3917	0.3175	0.2517	0.1952	0.1483	0.1104
23	0.9633	0.9367	0.8989	0.8490	0.7875	0.7160	0.6374	0.5551	0.4728	0.3939	0.3209	0.2559	0.1998	0.1529
24	0.9777	0.9594	0.9317	0.8933	0.8432	0.7822	0.7117	0.6346	0.5540	0.4734	0.3959	0.3242	0.2599	0.2042
25	0.9869	0.9748	0.9554	0.9269	0.8878	0.8377	0.7771	0.7077	0.6319	0.5529	0.4739	0.3979	0.3272	0.2637
26	0.9925	0.9848	0.9718	0.9514	0.9221	0.8826	0.8324	0.7723	0.7038	0.6294	0.5519	0.4744	0.3997	0.3301
27	0.9959	0.9912	0.9827	0.9687	0.9475	0.9175	0.8775	0.8274	0.7677	0.7002	0.6270	0.5509	0.4749	0.4014
28	0.9978	0.9950	0.9897	0.9805	0.9657	0.9436	0.9129	0.8726	0.8225	0.7634	0.6967	0.6247	0.5500	0.4753
29	0.9989	0.9973	0.9941	0.9882	0.9782	0.9626	0.9398	0.9085	0.8679	0.8179	0.7593	0.6935	0.6226	0.5492
30	0.9994	0.9986	0.9967	0.9930	0.9865	0.9758	0.9595	0.9360	0.9042	0.8633	0.8134	0.7553	0.6903	0.6206
31	0.9997	0.9993	0.9982	0.9960	0.9919	0.9848	0.9735	0.9564	0.9322	0.8999	0.8589	0.8092	0.7515	0.6874
32	0.9999	0.9996	0.9990	0.9978	0.9953	0.9907	0.9831	0.9711	0.9533	0.9285	0.8958	0.8546	0.8051	0.7479
33	0.9999	0.9998	0.9995	0.9988	0.9973	0.9945	0.9895	0.9813	0.9686	0.9502	0.9249	0.8918	0.8505	0.8011
34	1	0.9999	0.9998	0.9994	0.9985	0.9968	0.9936	0.9882	0.9794	0.9662	0.9472	0.9213	0.8879	0.8465
35	1	1	0.9999	0.9997	0.9992	0.9982	0.9962	0.9927	0.9868	0.9775	0.9637	0.9441	0.9178	0.8841
36	1	1	0.9999	0.9998	0.9996	0.9990	0.9978	0.9956	0.9918	0.9854	0.9756	0.9612	0.9411	0.9144
37	1	1	1	0.9999	0.9998	0.9995	0.9988	0.9974	0.9950	0.9908	0.9840	0.9737	0.9587	0.9381
38	1	1	1	1	0.9999	0.9997	0.9993	0.9985	0.9970	0.9943	0.9897	0.9825	0.9717	0.9562
39	1	1	1	1	0.9999	0.9999	0.9996	0.9992	0.9983	0.9966	0.9936	0.9887	0.9810	0.9697
40	1	1	1	1	1	0.9999	0.9998	0.9996	0.9990	0.9980	0.9961	0.9928	0.9875	0.9795
41	1	1	1	1	1	1	0.9999	0.9998	0.9995	0.9988	0.9976	0.9955	0.9920	0.9864
42	1	1	1	1	1	1	1	0.9999	0.9997	0.9993	0.9986	0.9973	0.9950	0.9911
43	1	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9992	0.9984	0.9969	0.9944
44	1	1	1	1	1	1	1	1	0.9999	0.9997	0.9993	0.9986	0.9973	0.9950
45	1	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9995	0.9989	0.9978
46	1	1	1	1	1	1	1	1	1	0.9999	0.9999	0.9997	0.9994	0.9987
47	1	1	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9996	0.9992
48	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9998	0.9996
49	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9998
50	1	1	1	1	1	1	1	1	1	1	1	1	0.9999	0.9999
51	1	1	1	1	1	1	1	1	1	1	1	1	1	0.9999
52	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Loi normale centrée réduite

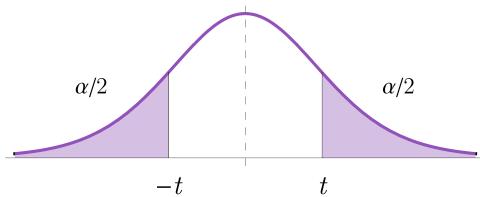
Fonction de répartition $P(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{x^2}{2}} dx$ (en couleur sur la représentation suivante)



u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

Loi de Student

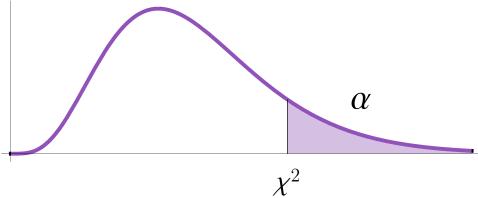
Valeurs (absolues) du t ayant la probabilité α d'être dépassées



		α												
		0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
ν	1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	636.6
	2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.60
3	0.137	0.277	0.424	0.584	0.765	0.979	1.250	1.638	2.353	3.182	4.541	5.841	12.92	
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610	
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869	
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959	
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.500	5.408	
8	0.130	0.262	0.400	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.897	3.355	5.041	
9	0.129	0.261	0.398	0.544	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781	
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.813	2.228	2.764	3.169	4.587	
11	0.129	0.260	0.396	0.540	0.697	0.875	1.088	1.363	1.796	2.201	2.718	3.106	4.437	
12	0.128	0.259	0.395	0.539	0.696	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318	
13	0.128	0.259	0.394	0.538	0.694	0.870	1.080	1.350	1.771	2.160	2.650	3.012	4.221	
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.625	2.977	4.141	
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.603	2.947	4.073	
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.584	2.921	4.015	
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965	
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922	
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.540	2.861	3.883	
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850	
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819	
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792	
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.320	1.714	2.069	2.500	2.807	3.768	
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745	
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725	
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707	
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690	
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674	
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659	
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646	
40	0.127	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.705	3.551	
80	0.126	0.254	0.387	0.527	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.416	
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.374	
∞	0.126	0.253	0.385	0.524	0.675	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291	

Loi du χ^2

Valeurs du $\chi^2(v)$ ayant la probabilité α d'être dépassées



		α																	
		0.995	0.99	0.975	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.025	0.01	0.005	0.001		
v	1	—	0.0002	0.0010	0.0039	0.0158	0.0642	0.149	0.455	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83		
	2	0.0100	0.0201	0.0506	0.103	0.211	0.446	0.713	1.39	2.41	3.22	4.61	5.99	7.38	9.21	10.60	13.82		
	3	0.0717	0.115	0.216	0.352	0.584	1.01	1.42	2.37	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27		
	4	0.207	0.297	0.484	0.711	1.06	1.65	2.19	3.36	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47		
	5	0.412	0.554	0.831	1.15	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.52		
	6	0.676	0.872	1.24	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46		
	7	0.989	1.24	1.69	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32		
	8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	17.53	20.09	21.96	26.12		
	9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88		
	10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59		
	11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	21.92	24.73	26.76	31.26		
	12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91		
	13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53		
	14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12		
	15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70		
	16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25		
	17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79		
	18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31		
	19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82		
	20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31		
	21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80		
	22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27		
	23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73		
	24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18		
	25	10.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62		
	26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05		
	27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	43.19	46.96	49.64	55.48		
	28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89		
	29	13.12	14.26	16.05	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30		
	30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70		
	31	14.46	15.66	17.54	19.28	21.43	24.26	26.44	30.34	34.60	37.36	41.42	44.99	48.23	52.19	55.00	61.10		
	32	15.13	16.36	18.29	20.07	22.27	25.15	27.37	31.34	35.66	38.47	42.58	46.19	49.48	53.49	56.33	62.49		
	33	15.82	17.07	19.05	20.87	23.11	26.04	28.31	32.34	36.73	39.57	43.75	47.40	50.73	54.78	57.65	63.87		
	34	16.50	17.79	19.81	21.66	23.95	26.94	29.24	33.34	37.80	40.68	44.90	48.60	51.97	56.06	58.96	65.25		
	35	17.19	18.51	20.57	22.47	24.80	27.84	30.18	34.34	38.86	41.78	46.06	49.80	53.20	57.34	60.27	66.62		
	36	17.89	19.23	21.34	23.27	25.64	28.74	31.12	35.34	39.92	42.88	47.21	51.00	54.44	58.62	61.58	67.99		
	37	18.59	19.96	22.11	24.07	26.49	29.64	32.05	36.34	40.98	43.98	48.36	52.19	55.67	59.89	62.88	69.35		
	38	19.29	20.69	22.88	24.88	27.34	30.54	32.99	37.34	42.05	45.08	49.51	53.38	56.90	61.16	64.18	70.70		
	39	20.00	21.43	23.65	25.70	28.20	31.44	33.93	38.34	43.11	46.17	50.66	54.57	58.12	62.43	65.48	72.05		
	40	20.71	22.16	24.43	26.51	29.05	32.35	34.87	39.34	44.16	47.27	51.81	55.76	59.34	63.69	66.77	73.40		

Index

- Analyse de la variance, 101, 117-120, 122-123
- Bayes, 7
- Bienaymé-Tchebychev, 20-21, 25, 47-48, 81
- Boîte à moustaches, 80
- Coefficient de corrélation, 101-102
- Comparaison de deux proportions, 82
- Comparaison de deux régressions, 108
- Comparaison de moyennes, 72-73, 76, 117
- Comparaison de pentes, 109
- Comparaison de variances, 72-73, 108
- Comparaison d'ordonnées à l'origine, 109
- Comparaisons statistiques, 63
- Contrôle, 31, 36-39, 41
- Contrôle de fabrication, 36, 38
- Contrôle de réception, 36-37, 65
- Contrôle progressif, 41
- Convergence en probabilité, 47-48
- Covariance, 23-24, 101-102
- Diagramme des fréquences cumulées, 80, 85
- Distance du χ^2 , 83
- Droite des moindres carrés, 99-100, 102-103, 106, 108, 119
- Ecart-type, 20
- Echantillon, 33
- Echantillonnage, 33, 35, 41, 52, 72, 85
- Efficacité, 37-41, 48
- Equiprobables, 4-5
- Espérance mathématique, 11, 15, 20-21, 23, 47-48, 51, 81, 103
- Estimateur, 47-48, 50-51, 58-59, 73-74, 82, 103, 105
- Estimation, 45, 47, 49, 50-52, 56-59, 70, 72-74, 82, 84-85, 103-104, 106-108, 118
- Estimation d'une moyenne, 50
- Estimation d'une proportion, 50, 82
- Estimation d'une variance, 51, 73
- Événements, 3-8, 10-12, 17
- Événements indépendants, 6, 10
- Expérimentation statistique, 115
- Histogramme, 27, 79
- Hypothèses composites, 65, 68
- Indépendance, 19, 21, 23, 32, 55, 88-89, 105
- Influence de facteurs, 115, 118, 120, 124, 124
- Intervalle de confiance, 47, 49-52, 55-57, 74, 82, 105-108
- Intervalle de confiance d'une moyenne, 50
- Intervalle de confiance d'une variance, 51
- Kolmogorov, 4, 83
- Loi binomiale, 8-11, 28, 35, 37, 50, 84, 85, 133
- Loi de Bernoulli, 8, 11, 22, 28, 81

- Loi de la Moyenne, 36, 56
- Loi de Poisson, 10-12, 18, 22-23, 58, 85, 137
- Loi de Snedecor, 72-73, 108, 118, 120-124, 147
- Loi de Student, 55-56, 69, 71, 74, 77, 143
- Loi des grands nombres, 11, 47, 81
- Loi du χ^2 , 52-56, 69, 71, 73, 74, 83, 86-87, 89, 118, 120, 122-124, 145
- Loi exponentielle, 18, 23, 58
- Loi hypergéométrique, 8-9
- Loi log-normale, 28, 85
- Loi normale, 15, 24-28, 36, 50-51, 56, 58-59, 69-70, 72, 74, 82, 85, 105-106, 109, 141
- Loi normale centrée réduite, 26, 36, 51, 56, 69, 82, 105-106, 141
- Loi uniforme, 17, 22

- Maximum de vraisemblance, 57

- Neyman et Pearson, 63, 65-66
- Nombres au hasard, 33, 131

- Plan factoriel, 120-121
- Population, 33
- Prévision statistique, 104
- Probabilités composées, 6, 8, 19, 88
- Probabilités conditionnelles, 5-7
- Processus de Poisson, 6, 11-12, 18

- Régression linéaire, 97
- Risque α , 37-38, 66
- Risque β , 37-38, 66

- Taille d'un échantillon, 40
- Test d'analyse de la variance, 118
- Test de la médiane, 87
- Test de linéarité d'une régression, 119
- Test de nullité de l'ordonnée à l'origine, 106
- Test des appariements, 71
- Test des signes, 88
- Test d'indépendance, 88, 105
- Test d'influence d'un facteur, 123
- Test d'interaction, 122
- Test sur une moyenne, 67, 70-71, 73
- Test sur une proportion, 66
- Test sur une variance, 71-72
- Tests d'hypothèse, 65
- Tests non paramétriques, 87
- Théorème central limite, 27-28, 51, 82

- Variables continues, 17, 85
- Variables discrètes, 7, 85
- Variance, 20-25, 27-28, 34-36
- Vraisemblance, 57-59

Ce cours de statistique s'articule en 8 chapitres exposant de manière progressive les méthodes essentielles de la statistique dont l'ingénieur a souvent besoin. Après une introduction aux notions de probabilité et de variable aléatoire dans le chapitre 1, nous présentons, au chapitre 2, la loi normale sur laquelle s'appuieront la plupart des raisonnements qui suivront. Le contrôle statistique, qui est une activité essentielle de l'entreprise fait l'objet du 3e chapitre. Le chapitre 4 est consacré à l'estimation statistique (ponctuelle et par intervalles de confiance) des paramètres des lois d'appartenance des données échantillonnées. La comparaison de ces estimations à des valeurs supposées connues, de même que la comparaison de populations entre elles font l'objet du chapitre 5. On présente dans le chapitre 6 la technique de l'ajustement d'un ensemble de données à un modèle. La régression linéaire fait l'objet du chapitre 7. On y insiste notamment sur les hypothèses qui sous-tendent ce modèle et sur ses applications à la prévision et au contrôle. Enfin, nous traitons de l'expérimentation statistique, c'est-à-dire de l'étude de l'influence de facteurs sur un processus complexe par l'utilisation des plans d'expérience et l'analyse de la variance.

Les méthodes ainsi présentées constituent la base de la statistique dite inférentielle, celle grâce à laquelle, l'ingénieur pourra contrôler, décider et prévoir, à partir d'échantillons de données.

Les fondements théoriques sont rapidement donnés, parfois simplement énoncés, et on pourra les approfondir avec l'abondante littérature disponible, notamment sur internet. C'est volontairement que nous nous consacrons à l'essentiel, c'est-à-dire à l'apprentissage des méthodes et modes de raisonnement de la statistique de très grande importance dans de nombreuses activités humaines.

Pour les lecteurs extérieurs à Mines Nancy, je précise que ce document ne comporte pas d'exemples (ou très peu) parce que c'est en cours et en TD que les applications sont traitées.