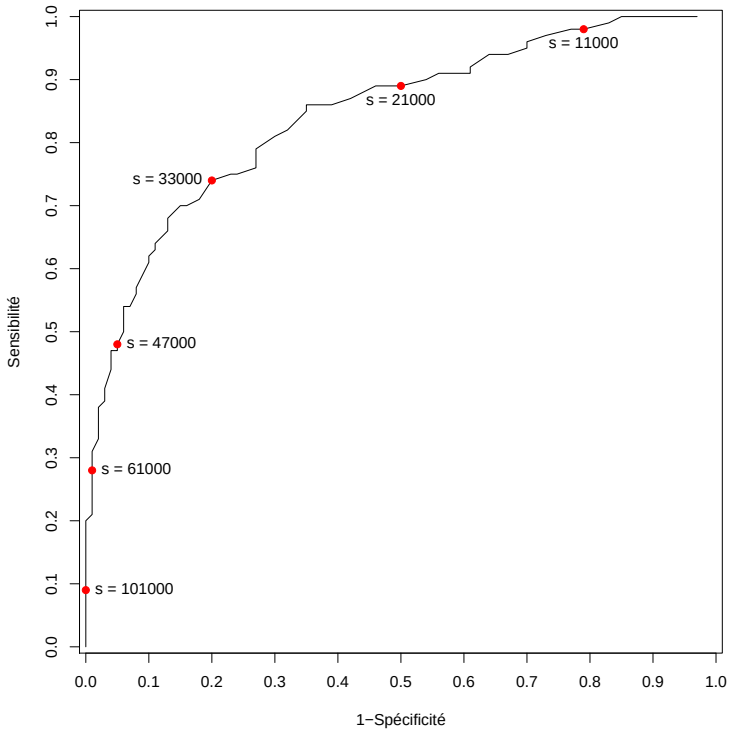
Université Paris-Sud 11
Faculté de Médecine

M1 de Santé Publique
DU SSV

Modélisation



Année 2018 – 2019

H. Perdry
herve.perdry@gmail.com

> Plût au ciel que le lecteur, enhardi et devenu momentanément féroce comme ce qu'il lit, trouve, sans se désorienter, son chemin abrupt et sauvage, à travers les marécages désolés de ces pages sombres et pleines de poison; car, à moins qu'il n'apporte dans sa lecture une logique rigoureuse et une tension d'esprit égale au moins à sa défiance, les émanations mortelles de ce livre imbiberont son âme comme l'eau le sucre. Il n'est pas bon que tout le monde lise les pages qui vont suivre ; quelques-uns seuls savoureront ce fruit amer sans danger.
>
> LAUTRÉAMONT, Les Chants de Maldoror.

Qu'est-ce que ce cours ?

Ce cours s'adresse aux étudiants de Master 1 Santé Publique, second semestre.

Il suppose qu'on a suivi un cours d'introduction aux probabilités et statistiques. Ce cours est cependant consacré aux concepts de base de la théorie, avec une insistance particulière sur la façon dont on construit les tests. Les principales nouveautés par rapport au cours du premier semestre sont : la méthode du Delta, les courbes ROC, et une rapide introduction à l'anova.

Le style est délibérément un peu plus mathématisant que dans la plupart des cours introductifs de biostatistiques. On suppose que passée la fraîcheur du premier contact, le lecteur et la lectrice sauront s'enhardir et trouver leur chemin dans ces pages.

Comme document complémentaire, nous conseillons principalement le document de cours de proba/stas du premier semestre (cours de M. Broët). Si vous n'avez pas suivi ce cours, vous pourrez vous en procurer facilement une copie auprès de vos camarades.

Ceux qui veulent utiliser un livre (mais cela ne devrait en rien être nécessaire) peuvent regarder :

- Bernard Prum, *Modèle linéaire (Comparaison de groupes et régression)*, éditions Inserm.
- Simone Bénazeth et coll., *Biomathématiques : Pharmacie, Médecine 1e et 2e années,* éditions Masson.
- Michel Lejeune, *Statistique (La théorie et ses applications)*, Springer-Verlag.

Notations

Les exercices ou questions signalés d'une étoile sont plus difficiles. Quand il y a deux étoiles, ils ne sont destinés qu'à ceux qui sont les plus à l'aise en mathématiques.

La notation log désigne le logarithme népérien, que la norme ISO 80000 recommande pourtant de noter ln. Si on avait besoin du logarithme décimal on utiliserait la notation $\log_{10}$ – rappelons que $\log_{10}(x) = \frac{\log(x)}{\log(10)}$. Les fonctions trigonométriques telles que sin prennent des angles exprimés en radians et non en degrés, et arcsin renvoie des radians et non des degrés.

Méthode de travail recommandée

Les étudiants inscrits sur place sont invités à venir en cours et en TD, et à préparer les TD.

Pour les étudiants par correspondance, la méthode de travail recommandée est simplement de :

- lire les chapitres ; (est-il besoin de préciser qu'on ne lit pas un cours comme un roman, mais le stylo à la main, en prenant des notes ?)
- quand les exemples contiennent des calculs, les refaire soi-même ;
- faire les exercices.

Vous recevrez mi-mars un devoir à la maison, qui entrera pour 20% dans la composition de la note de première session (pour la deuxième session, seul l'examen final sera pris en compte).

N'hésitez pas à utiliser le forum mis à votre disposition pour discuter des points problématiques entre vous : `bit.ly/yReB79` (identifiant : `forum_SSV` mot de passe : `student99`). Si vous y voyez une question à laquelle vous pensez savoir répondre, n'hésitez pas à le faire ! Un forum sert à s'entraider. Si des questions précises restent sans réponse, les modérateurs interviendront pour y répondre le mieux possible. Merci de faire preuve de politesse à leur égard.

Chapitre 1

Expériences aléatoires

1.1 Qu'est-ce qu'une expérience aléatoire ?

Les expériences aléatoires nous sont familières : jouer à « pile ou face », lancer un dé (à 6 faces ou plus), mélanger un jeu de cartes et distribuer ces cartes, jouer au loto, choisir un morceau de papier portant un nom dans un chapeau, tirer à la courte paille qui sera mangé ; mais aussi : sortir de cours et aller attendre le bus à côté de la fac (combien de temps va-t-on l'attendre ? combien de personnes à bord ?) ; et encore : choisir une personne « au hasard » dans une population donnée (par exemple, une des personnes assises dans la salle de cours : quelle est la couleur de ses yeux ? son prénom ? sa tension artérielle ? son petit déjeuner préféré ? son sexe ? sa stature ?) ; etc.

Ce qui rend ces expériences *aléatoires* différentes d'expériences *déterministes* c'est qu'il est impossible de prévoir leur résultat à l'avance. Il est cependant possible de décrire leurs résultats, et d'en tirer des prévisions (on attend le bus en moyenne 10 minutes, environ 8 personnes sur 10 ont les yeux bruns, etc).

Il est difficile de définir le hasard ; c'est un problème philosophique sur lequel il n'est pas forcément nécessaire de s'attarder ; on peut agir (calculer, faire des tests du χ^2, etc) sans comprendre ; comme le dit Maurice Biraud dans *Un taxi pour Tobrouk,* « deux intellectuels assis vont moins loin qu'une brute qui marche ».

1.1.1 « Sources de hasard »

Il est pourtant intéressant d'y réfléchir : d'où vient ce qu'on appelle le hasard ?

Paramètres non observés

Considérons un système déterministe dont on n'observe pas toutes les composantes, par exemple un montage électrique comportant en série deux interrupteurs A et B et une ampoule.

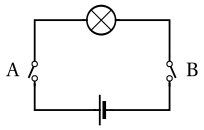


Figure 1. Montage à deux interrupteurs

Si un expérimentateur qui s'apprête à fermer l'interrupteur A ne peut pas observer l'interrupteur B, il ne sait pas si la lampe va s'allumer ou non.

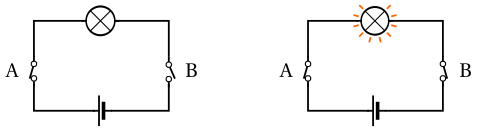


Figure 2. Résultat de l'expérience selon l'état de B

S'il peut, par exemple, répéter l'expérience un grand nombre de fois (l'état de l'interrupteur B pouvant changer pour une raison inconnue entre les expériences), il pourra par exemple calculer la proportion d'expériences où la lampe s'allume, et être amené à considérer que la lampe s'allume avec une certaine probabilité, égale à cette fréquence. La situation idéale serait que l'expérimentateur dispose d'un grand nombre de montages identiques, et qu'il n'utilise chacun qu'une fois, ce qui nous dispense de nous demander pourquoi et comment l'interrupteur B change d'état d'une expérience à l'autre.

On peut proposer un exemple analogue dans le domaine médical : on administre une forte dose de codéine à un patient (on ferme l'interrupteur A) ; suivant son génotype au gène *CYP2D6* (qu'on n'a pas observé au préalable : c'est l'interrupteur B), qui code une enzyme impliquée dans le métabolisme des xénobiotiques (et donc, des médicaments), ce patient est ou n'est pas « acétyleur ultra-rapide » (s'il est acétyleur ultra-rapide, il métabolisera *très rapidement* une partie de la codéine en morphine) ; dans ce dernier cas, il risque une intoxication sévère (c'est l'ampoule allumée).

Dans ce cas, bien que les mécanismes en jeu soient déterministes, le hasard et les probabilités servent à modéliser les incertitudes qu'on a sur certains paramètres.

Rencontre de séries causales indépendantes

Cournot (un intellectuel assis ?) a proposé la définition suivante du hasard, souvent citée : « le hasard est la rencontre de deux séries causales indépendantes ». On peut tenter d'illustrer cette idée avec l'exemple du temps d'attente à l'arrêt de bus d'un étudiant qui sort de cours : première série causale, celle qui détermine l'heure à laquelle le cours prend fin (fatigue de l'enseignant, concentration des étudiants, bon approvisionnement en craie ou en feutres, etc) ; deuxième série causale, celle qui détermine l'heure de passage du bus (densité de la circulation, nombre de passagers qui montent et descendent aux arrêts qui précèdent, etc). Le résultat est, pour une personne qui n'observe que la première série causale, l'impossibilité de prévoir avec précision son temps d'attente à l'arrêt de bus.

Séries causales trop complexes pour être analysées

Cet exemple nous aiguille vers d'autres considérations. Dans une série causale complexe, comme celle de la circulation d'un bus, connaître tous les paramètres avec assez de précision pour prédire le résultat de l'expérience est impossible : Monsieur A s'arrête un instant devant une vitrine de libraire ; ceci a pour conséquence de le faire arriver un peu en retard à l'arrêt de bus : il court, et le chauffeur complaisant l'attend ; le léger retard pris par le bus a pour conséquence de lui faire rater un feu vert. Il attend au feu rouge pendant 100 secondes, il y a deux personnes de plus à faire monter à l'arrêt suivant que s'il n'avait pas attendu, etc. D'autre part ce qui a retenu l'attention de Monsieur A dans la vitrine du libraire est une édition originale d'*Hereditary Genius* de Francis Galton, achetée la veille par la librairie au bibliophile Monsieur B et placée dans la vitrine par le libraire quelques minutes avant le passage de Monsieur A ; etc. Les séries causales s'entrochoquent et sont manifestement trop complexes pour être appréhendées ! Cette « amplification » des petites actions évoque de ce que les mathématiciens appellent *chaos déterministe*, une « source de hasard » qui peut être ainsi rapprochée de la définition de Cournot.

Chaos déterministe

Dans certains systèmes déterministes, la moindre pertubation en cours d'expérience, la moindre incertitude quant aux conditions initiales, peut avoir pour conséquence l'impossibilité de prévoir l'issue de l'expérience. Un exemple simple est donné par la *planche de Galton*, un jeu de fête foraine où une bille descend à travers un réseau de clous disposés en quinconce. Quand la bille tombe à la

verticale d'un clou, elle va rester un bref instant en équilibre sur ce clou, puis tombera vers la gauche ou vers la droite. Comment ce « choix » s'effectue-t-il ? La mécanique est une théorie parfaitement déterministe ; pourtant, lors d'expériences successives réalisées avec une planche de Galton, on obtiendra des résultats différents ; ces différences s'expliquent par de petites différences de conditions initiales (on n'a pas lâché la bille exactement de la même façon) et par la présence de petites perturbations (vibrations, courants d'air) qui modifient le déroulement de l'expérience.

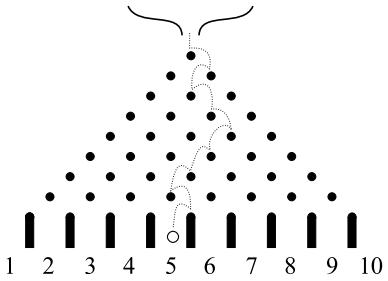


Figure 3. Planche de Galton

Dans un système de ce type, il est pratiquement impossible, même en affinant les mesures, de lever l'indétermination, de prédire le résultat de l'expérience, le système fonctionnant comme un « amplificateur d'incertitude ». Cette impossibilité est à opposer au cas du montage électrique proposé plus haut, où une observation simple suffirait à lever l'indétermination (état de l'interrupteur B, génotype du patient en *CYP2D6*).

Dans le domaine médical, on pourra penser au cas de jumeaux monozygotes dont l'un a un diabète de type I (une maladie auto-immune), l'autre non. Leur système immunitaire suit, au cours de leur vie, une trajectoire qu'on peut comparer à celle de la bille qui traverse la planche de Galton ; les bifurcations ont, dans un cas, abouti à un diabète, pas dans l'autre.

Le hasard en mécanique quantique

Jusqu'à présent, nous avons parlé de hasard là où il n'y a qu'imprévisibilité, due à notre ignorance (rémédiable ou non). Néanmoins, en physique moderne, plus précisément en mécanique quantique (et jusqu'à nouvel ordre), le hasard est une réalité. Placez un bloc de matière radioactive (du granite ou du bois, par exemple) sous un détecteur. Comptez le nombre N de particules détectées pendant un temps t ; mesurez le temps T écoulé entre deux tels évéments ; pour la physique moderne, ceci est bien le produit du hasard, quelque chose de fondamentalement indéterminé (même si des lois de probabilité pour les valeurs de N et T existent) qui ne dépend pas de variables cachées ou de pertubations extérieures mal observées. Un physicien proposera volontiers l'*expérience mentale* suivante : placez un atome d'hydrogène dans un état excité dans un univers vide (sic) ; au bout d'un temps T, l'atome retombe dans son état fondamental en émettant un photon ; la valeur de T n'est pas déterminée (elle suit une loi de probabilité exponentielle, cf 5.2.2).

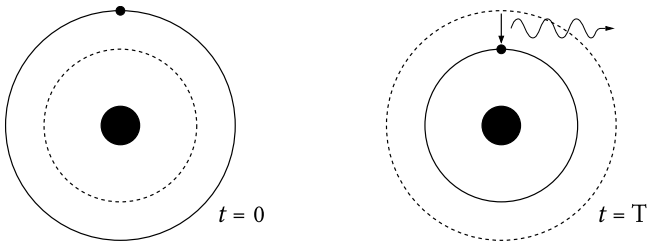


Figure 4. Atome d'hydrogène se désexcitant en émettant un photon

Même la position des particules est incertaine, ou plutôt la donnée simultanée de la position x et de la vitesse v d'une particule de masse m, sont probabilistes ; l'écart-type σ_x de la position et l'écart-type σ_v de la vitesse sont liés par la relation d'incertitude de Heisenberg, $\sigma_x \sigma_v \geqslant \frac{\hbar}{2m}$, où $\hbar = 1{,}055 \cdot 10^{-34}$ est la constante de Dirac.

Le hasard dans les sciences de la vie

Qu'en est-il du hasard en biologie, en médecine ? L'exposition à certains facteurs environnementaux, la survenue d'une mutation dans une lignée cellulaire, l'inactivation d'un chromosome X sur deux dans les cellules d'une femme, l'activation ou l'inactivation de certains gènes... tout cela relève bien du hasard, que ce soit à cause de paramètres non observés (exposition à un facteur environnemental), ou de séries causales trop complexes pour être analysées, ou même de « hasard vrai », certains des événements de la vie de la cellule faisant intervenir le mouvement brownien des molécules plongées dans le cytoplasme.

Quelques pistes bibliographiques :

Elowitz MC et coll (2002). Stochastic gene expression in a single cell. Science.
Kupiec JJ et coll (2009). Le hasard au cœur de la cellule. Éditions Syllepse.

1.2 Jeux de hasard

Les jeux de hasard sont des exemples familiers d'expériences aléatoires ; ils sont à l'origine du développement des probabilités, suivis de près par les études démographiques, l'actuariat, l'économie ; Cardan et Gallilée sont apparemment parmi les premiers à s'être intéressés à la question avant l'essort de la théorie sous l'impulsion de Fermat et de Pascal.

L'histoire commence par des questions posées par le Chevalier de Méré à Pascal : nous allons en donner un aperçu afin d'illustrer les concepts de base de la théorie.

1.2.1 Le premier problème du Chevalier de Méré

Pascal eut, avant de devenir mystique et dévot, une jeunesse frivole. Parmi ses amis, il comptait un joueur assidu, le Chevalier de Méré, qui lui posa un problème qu'on peut formuler ainsi :

> « Pourquoi est-il avantageux de parier qu'on va « sortir » un six en lançant quatre fois le dé, alors qu'il ne l'est pas de parier qu'on va sortir un double six en lançant vingt-quatre fois deux dés ? »

Le Chevalier de Méré avait en effet « remarqué », à la suite de nombreuses parties, que le premier pari était avantageux mais que le second ne l'était pas. Ce résultat *expérimental* lui semblait paradoxal. Il faisait le raisonnement suivant : tirer un six arrive une fois sur six, et tirer un double six arrive une fois sur trente-six, soit six fois moins souvent ; en lançant les deux dés vingt-quatre fois au lieu de quatre fois, soit six fois plus de lancers, les deux paris doivent être équivalents.

Le raisonnement du chevalier a beau être erroné, sa façon de poser le problème est la bonne : *quelle est la proportion de parties gagnantes quand on joue un grand nombre de parties ?* Sa façon d'essayer de répondre à la question est bonne malgré ses erreurs : il s'agit de raisonner et calculer à partir du *fait* suivant – fait donné a priori, bien qu'il soit vérifiable par l'expérience : *quand on lance un dé un grand nombre de fois, la proportion de lancers où on obtient un « six » est de* $1/6$ *(un sur six).*

Dans une formulation à peine plus mathématisée, la question est : *quelle est la probabilité de gagner une partie, sachant que la probabilité de tirer un six est* $\frac{1}{6}$ *?*

On voit apparaître ici une définition possible pour la notion de *probabilité* : la probabilité d'un résultat possible est sa *fréquence* quand on observe un grand nombre d'expériences aléatoires ; dans le cas où on considère des événements « équipossibles », comme ici l'apparition d'une quelconque des six faces d'un dé parfaitement cubique, c'est également le rapport du nombre d'événements favorables au nombre d'événements possibles : ici, un événement favorable (la face « six » est au-dessus) pour six événements possibles (les six faces du dé).

Enfin, la méthode expérimentale pratiquée par le chevalier est déjà du domaine des *statistiques* : pour estimer une probabilité inconnue, pratiquer un grand nombre d'expériences et en relever les résultats ; la probabilité cherchée est approximativement le rapport du nombre d'expériences favorables (ici, de parties gagnantes) au nombre total d'expériences pratiquées (c'est exactement ce qu'on appellera plus tard la *méthode de Monte-Carlo*).

La réponse à la question du Chevalier sera donnée d'abord en section 2.3.3 puis en exercice.

1.3 Description d'expériences aléatoires répétées

Nous allons ici décrire brièvement les procédures usuelles utilisées pour décrire de façon compacte les résultats d'une série d'expériences aléatoires – on parle d'un *échantillon*.

Il convient de faire la différence entre :

— les mesures quantitatives discrètes, par exemple : le nombre d'enfants dans une famille ;
— les mesures quantitatives continues (la stature ou le poids d'un sujet) ;
— les mesures qualitatives, comme le sexe d'un individu, sa ville de naissance, ou la couleur de ses yeux.

Certaines mesures peuvent être mixtes, par exemple un taux d'anticorps mesurés chez des individus exposés à un agent infectieux : chez certains patients cette mesure sera nulle, chez d'autres, elle sera positive. On ne peut pas traiter cette mesure comme une mesure purement continue, il conviendra de rapporter d'une part la proportion d'individus chez lesquels la mesure est nulle, et d'autre part, de décrire la distribution des mesures strictement positives.

1.3.1 Une mesure qualitative

On rapportera des tables d'effectifs, ou des proportions (en précisant la taille de l'échantillon), ou les deux à la fois, comme ceci :

Marrons	Verts	Bleus	Total
90 (60%)	40 (27%)	20 (13%)	150

Table 1. Effectifs (proportions) des différentes couleurs d'yeux

1.3.2 Une mesure quantitative discrète

Pour une mesure discrète on pourra réaliser une table comme pour une mesure qualitative. Un diagramme « en bâton » peut également être pertinent.

0	1	2	3	4	5	6
16 (8%)	47 (23.5%)	51 (25.5%)	35 (17.5%)	29 (14.5%)	18 (9%)	4 (2%)

Table 2. Nombre d'enfants par couple (total 200 couples)

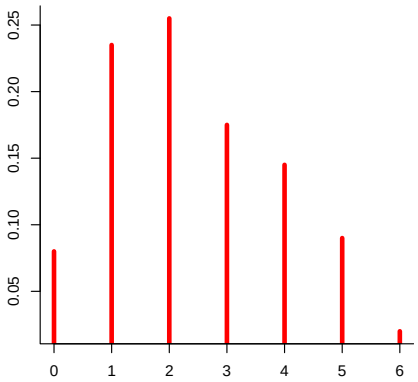


Figure 5. Proportions de couples avec 0 à 6 enfants (estimé sur un total de 200 couples)

On rapportera également classiquement des mesures de localisation et de dispersion.

Si n est la taille de l'échantillon, si on note $x_1, \dots, x_n$ les mesures, la moyenne est

$$\overline{x} = \frac{1}{n}(x_1 + \cdots + x_n).$$

C'est une mesure de localisation.

Dans le cas d'une variable discrète, on peut la retrouver sans peine à partir d'une table comme la table 2 : ici c'est $\overline{x} = \frac{1}{200}(16 \times 0 + 47 \times 1 + \cdots + 6 \times 4) = 2{,}42$.

L'*écart absolu moyen* est une mesure de dispersion de l'échantillon autour de sa moyenne. C'est la moyenne des écarts absolus, c'est-à-dire des valeurs absolues des écarts à la moyenne :

$$\boxed{\mathrm{e}_a = \frac{1}{n}\left(\left|x_1 - \overline{x}\right| + \cdots + \left|x_n - \overline{x}\right|\right).}$$

La *variance* et l'*écart-type* sont d'autres mesures de dispersion de l'échantillon. On définit la variance comme la moyenne des carrés des écarts à $\overline{x}$ (ou moyenne des écarts quadratiques) :

$$\tilde{\mathrm{S}}^2 = \frac{1}{n}\left(\left(x_1 - \overline{x}\right)^2 + \cdots + \left(x_n - \overline{x}\right)^2\right).$$

Son calcul est facilité par la formule décentrée suivante :

$$\boxed{\tilde{\mathrm{S}}^2 = \frac{1}{n}\left(x_1^2 + \cdots + x_n^2\right) - \overline{x}^2.}$$

En effet,

$$\begin{aligned}
\tilde{\mathrm{S}}^2 &= \frac{1}{n}\left(\left(x_1 - \overline{x}\right)^2 + \cdots + \left(x_n - \overline{x}\right)^2\right) \\
&= \frac{1}{n}\left(\left(x_1 - \overline{x}\right)x_1 + \cdots + \left(x_n - \overline{x}\right)x_n\right) - \frac{1}{n}\left(\left(x_1 - \overline{x}\right)\overline{x} + \cdots + \left(x_n - \overline{x}\right)\overline{x}\right) \\
&= \frac{1}{n}\left(\left(x_1^2 + \cdots + x_n^2\right) - \overline{x}(x_1 + \cdots + x_n)\right) + \frac{1}{n}\left(\left(x_1 - \overline{x}\right) + \cdots + \left(x_n - \overline{x}\right)\right)\overline{x} \\
&= \frac{1}{n}\left(\left(x_1^2 + \cdots + x_n^2\right) - \overline{x} \times n\overline{x}\right) + 0 \\
&= \frac{1}{n}\left(x_1^2 + \cdots + x_n^2\right) - \overline{x}^2.
\end{aligned}$$

L'écart-type est la racine carrée de la variance. Il est plus directement interprétable, comme l'écart absolu moyen.

Ces mesures sont faciles à calculer à partir d'une table récapitulative comme la table 2. On trouve un écart absolu moyen égal à 1,27, une variance égale à 2,26 et un écart-type égal à 1,50.

D'autres mesures de localisation et de dispersion peuvent être utilisées : la médiane, l'écart interquartiles, etc. Certaines d'entre elles sont introduites au paragraphe suivant.

1.3.3 Une mesure quantitative continue

Supposons qu'on a mesuré la stature de 200 individus. Une table comme celle qui suit est un peu encombrante – et quand la taille des échantillons augmente, cette solution ne peut plus être retenue :

145	153	156	158	159	161	163	164	166	168	170	172	174	177	179	181	185
146	153	156	158	159	161	163	165	166	168	170	172	174	177	179	182	186
149	153	156	158	159	161	163	165	166	168	171	173	174	177	179	182	186
150	153	156	158	159	162	163	165	166	168	171	173	174	177	180	183	187
151	153	157	158	160	162	163	165	166	168	171	173	175	177	180	183	188
151	153	157	158	160	162	163	165	167	169	171	173	175	178	180	183	188
151	154	157	158	160	162	163	166	167	169	171	173	175	178	180	183	188
151	155	157	159	160	162	163	166	167	169	171	173	176	178	180	184	190
152	155	157	159	161	162	164	166	167	169	171	174	176	178	181	184	
153	155	157	159	161	162	164	166	167	170	171	174	177	178	181	185	
153	155	158	159	161	163	164	166	167	170	171	174	177	178	181	185	
153	156	158	159	161	163	164	166	168	170	171	174	177	178	181	185	

Table 3. 200 statures (en centimètres) classées par ordre croissant

Représentations de la distribution : histogrammes et fonction de répartition

Une solution est de « discrétiser » les mesures, en créant des classes, par exemple de largeur 5 cm (notons que les valeurs étaient déjà discrétisées, en classes de 1 cm – la précision de la mesure).

[145,150]	(150,155]	(155,160]	(160,165]	(165,170]	(170,175]	(175,180]	(180,185]	(185,190]
4	19	33	34	32	29	25	17	7

Table 4. répartition de 200 statures en 9 classes de largeur 5 centimètres

Une telle table est mieux représentée par un histogramme. Les rectangles de l'histogramme peuvent avoir une hauteur qui donne l'effectif dans chaque classe (axe de gauche sur la figure), ou bien cet effectif divisé par (l'effectif total × la largeur de la classe), de sorte que la surface totale des rectangles vaut 1 (axe de droite). Dans ce cas, l'histogramme est une approximation de la densité de la variable.

Figure 6. Histogrammes de la stature (effectif ou densité)

Cette dernière solution présente également l'avantage de permettre de faire des classes de largeurs variables, ce qui n'est pas possible quand on choisit de représenter les effectifs. Le polygone des fréquences s'obtient en joignant les milieux des côtés supérieurs des rectangles de l'histogramme :

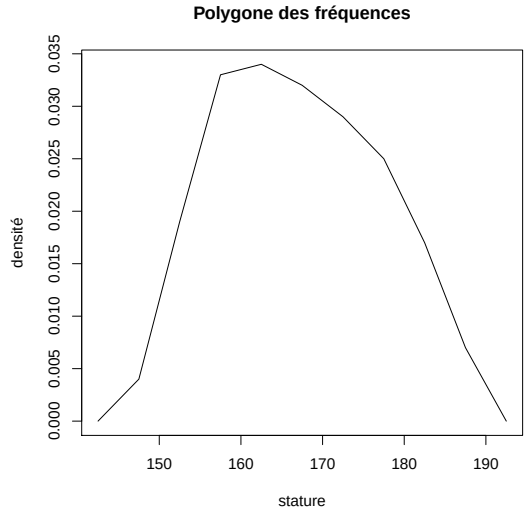


Figure 7. Polygone des fréquences

La fonction de répartition empirique n'est pas souvent représentée. Elle n'est pourtant pas dénuée d'intérêt : il s'agit de la fonction $F_n(x)$ qui donne la proportion de mesures inférieures ou égales à x. Avec les données de la table 3, on a par exemple 4 statures $\leqslant 151$ d'où $F(150) = \frac{4}{200} = 0{,}02$, etc. C'est une fonction « en escalier » dont les marches ont une hauteur multiple de $\frac{1}{n}$. La figure suivante donne la fonction $F_n(x)$ pour notre échantillon.

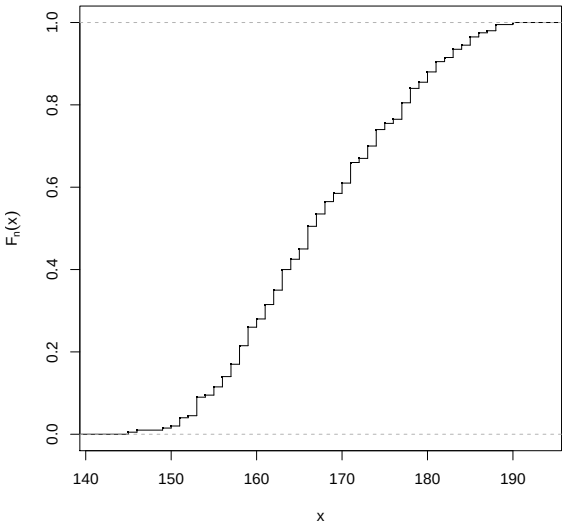


Figure 8. Fonction de répartition empirique

Localisation et dispersion

On rapporte bien entendu la moyenne (ici, 167,46 cm), la variance (100,4 cm^2) et l'écart-type (10,0 cm), ou l'écart absolu moyen (8,4 cm). Notez que la variance est bien une mesure en cm^2 – elle est difficile à interpréter directement, contrairement à l'écart-type ou à l'écart absolu moyen.

Les *quantiles* de l'échantillon, liés à la fonction de répartition empirique, sont souvent rapportés. La *médiane* est le quantile de niveau 0,5, c'est-à-dire la valeur $x_{0,5}$ telle que $F_n(x_{0,5}) = 0{,}5$: on a la moitié des données $\leqslant x_{0,5}$ (on arrondira à l'entier supérieur si besoin). Si les mesures sont classées par ordre croissant, c'est la mesure de rang $0{,}5 \times n$: dans notre cas c'est la mesure de rang 100, $x_{0,5} = 166$ cm

(la présence de mesures égales dans l'échantillon fait qu'on a en fait $F_n(166) = 0{,}505$). La médiane est une mesure de localisation, comme la moyenne (ici la moyenne est légèrement supérieure)

Les autres quantiles sont définis de façon analogue : le quantile de niveau α est x_α tel que $F_n(x_\alpha) = \alpha$: une proportion α des mesures est $\leqslant x_\alpha$. C'est la mesure de rang αn (on arrondit à l'entier supérieur si besoin – mais d'autres règles plus complexes existent, qui correspondent à diverses façons de « lisser » la fonction de répartition empirique). On rapporte fréquemment le *premier* et le *troisième quartiles*, qui sont respectivement $x_{0,25}$ et $x_{0,75}$. Dans notre exemple le premier quartile est $x_{0,25} = 159$, le troisième quartile est $x_{0,75} = 175$.

L'*écart inter-quartile* est une mesure de dispersion populaire ; on le notera IQR (comme *Inter Quartile Range)*. Il est définit par

$$\boxed{\text{IQR} = x_{0,75} - x_{0,25}.}$$

Dans notre exemple, on a IQR = 16 cm – notons que l'écart-type est 10,0 cm, et l'écart absolu moyen 8,4 cm.

Boîtes à moustaches

La *boîte à moustaches* ou *boîte de Tukey* est une façon très compacte de représenter une distribution, qui repose sur les quantiles. On dessine une boîte dont les bords vont du premier quartile au troisième quartile ; dans la boîte, un trait montre la position de la médiane. De part et d'autre de la boîte, des moustaches montrent l'étendue des données – plusieurs conventions concurrentes sont utilisées pour les moustaches (il conviendrait normalement de préciser quelle convention on utilise quand on dessine de telles boîtes). La convention la plus fréquente est que les moustaches finissent aux dernières mesures à une distance $< 1{,}5 \times \text{IQR}$ du bord de la boîte. Si il y a des données qui ne sont pas dans l'intervalle figuré par les moustaches, on les considère comme des mesures « exceptionnelles » *(outliers),* et on les représente par des points.

Dans notre exemple, la boîte va de $x_{0,25} = 159$ à $x_{0,75} = 175$, avec une bande en $x_{0,5} = 166$; la longueur critique d'$1{,}5 \times \text{IQR}$ est 24 cm, et on voit que la mesure la plus petite, 145, n'est pas à une distance plus grande du bord inférieure de la boîte en $x_{0,25} = 159$; c'est donc l'extrémité de la première moustache ; de même, la mesure la plus grande, 190, est l'extrémité de la seconde moustache, et il n'y pas de mesures exceptionnelles.

Figure 9. Boîte à moustaches des statures

Coefficients d'asymétrie et d'applatissement

On utilisera également souvent les termes anglais *skewness* et anglo-grec *kurtosis*.

Pour $k \geqslant 2$, définissons le k^{e} moment centré par

$$m_k = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^k.$$

Le 2^{e} moment centré est donc la variance $\tilde{S}^2$, qui mesure la dispersion des mesures.

Si les mesures sont distribuées de façon parfaitement symétrique autour de leur moyenne $\overline{x}$, la somme des $(x_i - \overline{x})^3$ sera nulle ; une valeur positive correspond à un excès de valeurs x_i à droite de $\overline{x}$, alors qu'une valeur négative correspond à un excès de valeurs x_i à gauche de $\overline{x}$. Le coefficient

d'asymétrie s'obtient en « normalisant » le 3^{e} moment comme suit :

$$\gamma_1 = \frac{m_3}{m_2^{3/2}} = \sqrt{n}\frac{\sum_i \left(x_i - \overline{x}\right)^3}{\left(\sum_i \left(x_i - \overline{x}\right)^2\right)^{3/2}}.$$

Un des effets de cette normalisation est de faire que γ_1 est sans unité : par exemple, si les x_i sont des mesures en centimètres, les moments m_3 et m_2 sont respectivement en cm^3 et cm^2 et γ_1 est sans unité.

Le même raisonnement conduit à définir le coefficient d'applatissement

$$\kappa = \frac{m_4}{m_2^2} = n\frac{\sum_i \left(x_i - \overline{x}\right)^4}{\left(\sum_i \left(x_i - \overline{x}\right)^2\right)^2}.$$

Les données de la table 3 donnent des valeurs $\gamma_1 = 0{,}16$ et $\kappa = 2{,}2$.

La figure suivante illustre les valeurs de γ_1 et de κ pour diverses distributions. La loi normale (cf chapitre 6) correspond au cas $\gamma = 0$ et $\kappa = 3$.

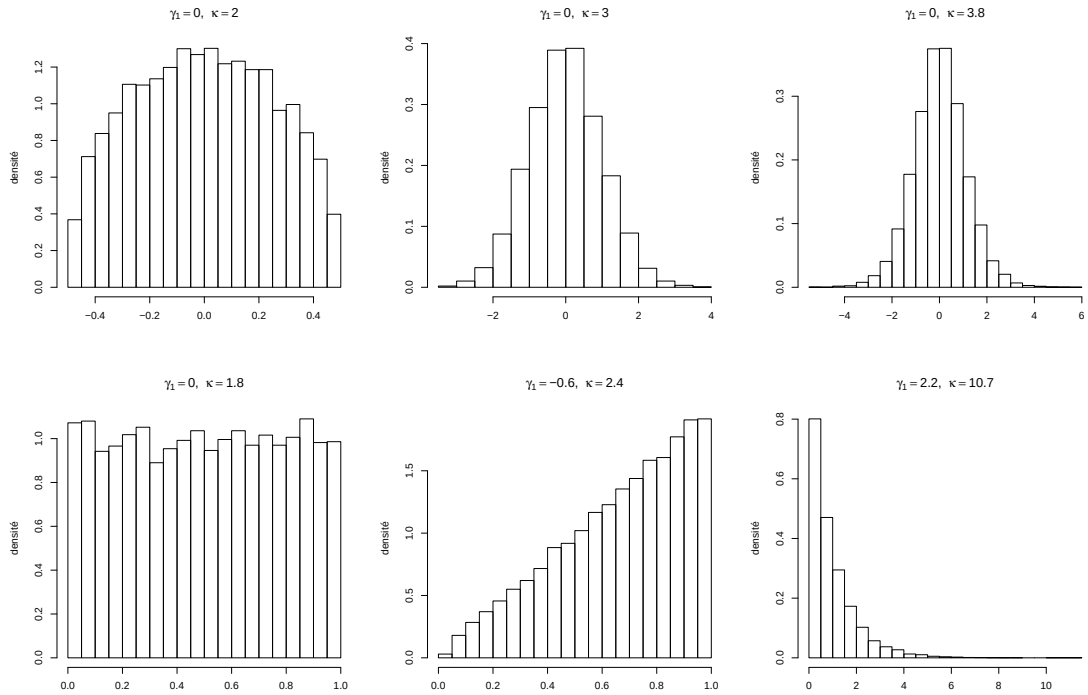


Figure 10. Exemples de coefficients d'asymétrie et d'applatissement

1.3.4 Deux mesures qualitatives ou discrètes

Quand on effectue deux mesures sur une même expérience aléatoire (par exemple sur un même individu tiré au hasard), on peut dresser une *table de contingence* comme celle-ci :

	Marrons	Verts	Bleus
Hommes	44	24	10
Femmes	46	16	10

Table 5. Effectifs des différentes couleurs d'yeux pour chacun des deux sexes

La distribution de chacune des deux variables peut être retrouvée en faisant la somme des colonnes ou des lignes de la table (on parle de distribution marginale).

1.3.5 Deux mesures continues

On peut bien sûr dresser une table de contingence après avoir discrétisé les deux mesures.

	[45,55]	(55,65]	(65,75]	(75,85]	(85,95]	(95,105]	(105,115]	(115,125]
[145,150]	1	–	–	–	–	–	–	–
(150,155]	10	2	4	–	–	–	–	–
(155,160]	11	9	7	4	–	–	–	–
(160,165]	9	8	13	2	–	–	–	–
(165,170]	5	11	9	4	3	–	–	–
(170,175]	5	8	5	7	3	1	–	–
(175,180]	1	5	6	6	4	2	1	–
(180,185]	–	3	4	6	2	1	1	–
(185,190]	–	1	–	2	3	–	–	1

Table 6. Stature et poids de 200 individus

La représentation graphique pertinente est le nuage de points. Dans le cas où les données sont arrondies (par exemple, à l'entier le plus le proche), des points peuvent être confondus. Une des solutions possibles est d'ajouter un peu de « bruit » à chacun des points pour qu'ils ne soient plus superposés – opération connue sous le nom de *jittering*.

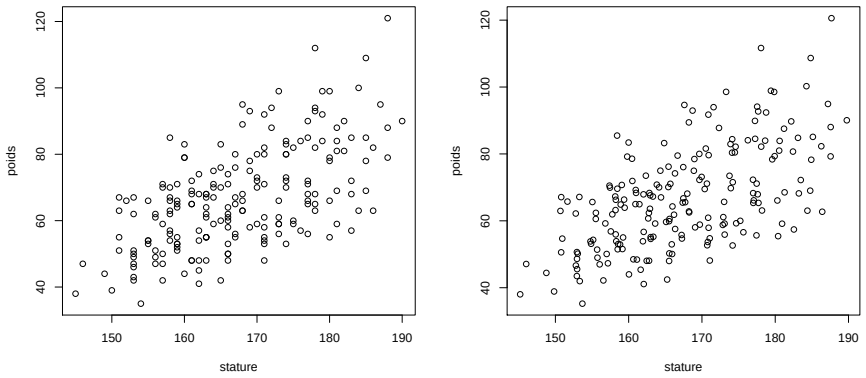


Figure 11. Stature et poids de 200 individus (sans *jittering* et avec)

On voit sur notre exemple qu'il y a une relation entre le poids et la stature : plus un individu est grand, plus son poids tend à être élevé. Pour quantifier cette relation, on calculera la *covariance* et la *corrélation*.

La covariance entre les mesures x_i et y_i pour $i = 1,\ldots,n$ est la moyenne des produits des écarts (algébriques) :

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right).$$

Si $\sigma_{xy} > 0$, les deux variables tendent à dévier du même côté de leur moyenne respective, et si $\sigma_{xy} < 0$ elles tendent à dévier dans des sens opposés.

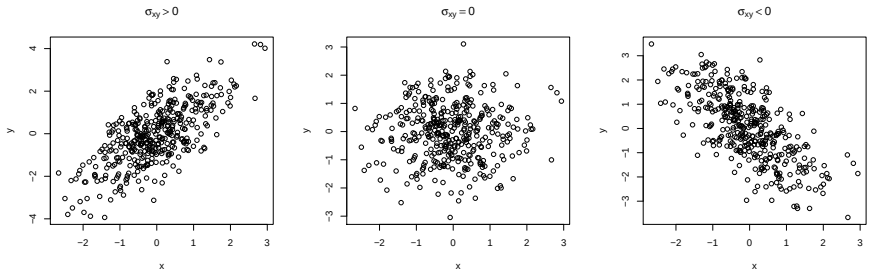


Figure 12. Signe de σ_{xy}

La corrélation est la covariance divisée par le produit des écart-types σ_x et σ_y :

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Elle a comme premier avantage d'être sans unité ; et comme deuxième avantage d'être comprise entre -1 et 1.

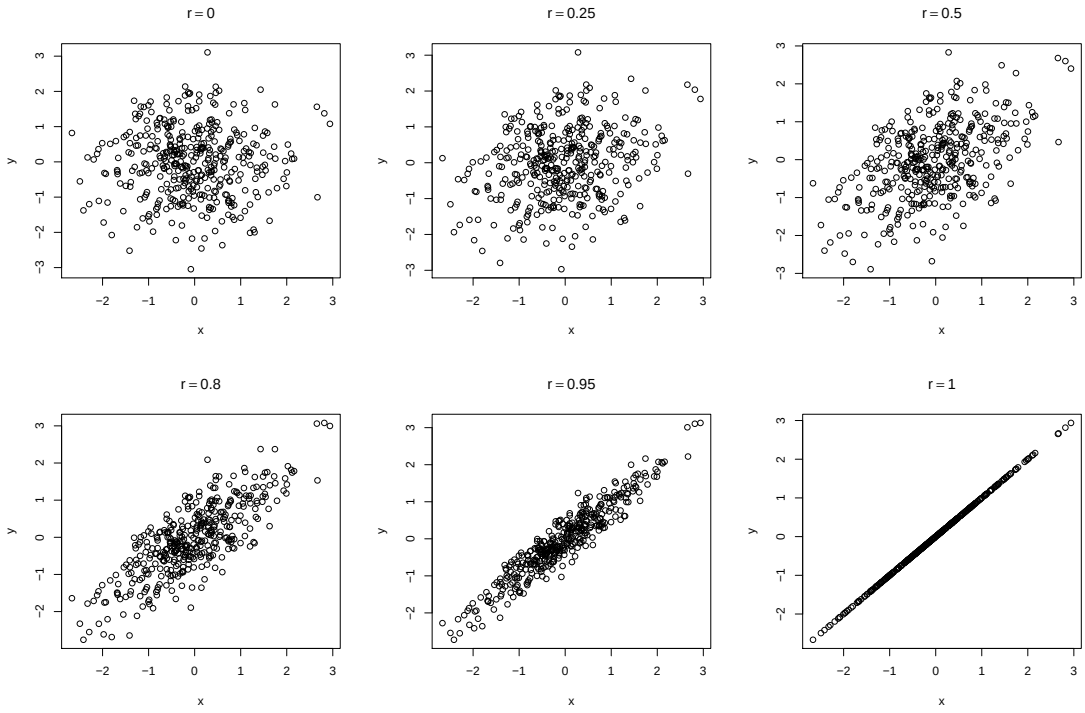


Figure 13. Valeurs de r

1.3.6 Une mesure qualitative ou discrète, une mesure continue

On pourra décrire la distribution de la mesure continue pour chacun des niveaux de la mesure qualitative (ou chaque valeur de la mesure discrète). Une solution très populaire est de dessiner des boîtes à moustaches côte à côte.

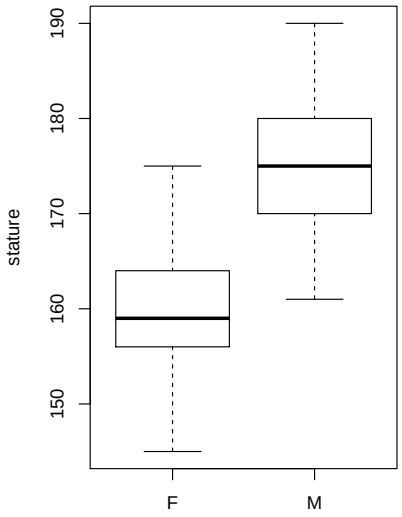


Figure 14. Boîtes à moustache pour la stature, en fonction du sexe

1.4 Exercices

Exercice 1

On a mesuré chez 60 individus la concentration plasmatique de BZ (bithure de zytron). Les mesures sont rassemblées dans la table ci-dessous.

0,14	0,49	0,86	1,57
0,20	0,53	0,89	1,58
0,23	0,58	0,90	1,66
0,27	0,61	0,96	1,71
0,27	0,62	1,01	1,78
0,31	0,62	1,11	1,96
0,34	0,69	1,14	2,08
0,35	0,73	1,16	2,08
0,36	0,73	1,21	2,22
0,37	0,74	1,25	2,75
0,38	0,75	1,32	3,28
0,38	0,76	1,37	4,84
0,40	0,77	1,41	7,44
0,46	0,78	1,44	7,54
0,46	0,82	1,53	7,94

Table 7. 60 concentrations plasmatiques en millimoles par litre

On donne également $\sum_i x_i = 83{,}13$, $\sum_i x_i^2 = 278{,}27$, $\sum_i x_i^3 = 1602{,}04$.

1. Calculez la moyenne, la médiane, l'écart-type et l'écart interquartile.

2. Tracez un histogramme et une boîte à moustaches. La distribution paraît-elle symétrique ?

3. En utilisant la formule

$$m_3 = \frac{1}{n}\sum_i x_i^3 - 3m_2\overline{x} - \overline{x}^3$$

calculez le coefficient d'asymétrie γ_1.

(On ne demande pas de démontrer la formule, mais les plus hardies pourront s'y risquer).

Chapitre 2

Probabilités

2.1 Le formalisme des espaces probabilisés

Pour modéliser la notion d'expérience aléatoire, nous introduisons d'abord la notion abstraite d'espace probabilisé Ω ; cette notion vous sera très probablement (!) nécessaire si vous voulez lire un jour des ouvrages de proba/stats qui sortent un peu du style « les meilleures recettes de cuisine de Tatie Suzon » (par exemple, l'excellent *Statistiques : la théorie et des applications* de Michel Lejeune, chez Springer-Verlag). Cependant, dans les chapitres qui suivent, cette notion ne sera presque pas utilisée.

Les mathématiciens, pour satisfaire aux besoins de rigueur propres à leur discipline, ont besoin de notions bien plus détaillées que celles que nous présenterons : le présent chapitre ne prétend pas se substituer à un cours classique de théorie de la mesure.

2.1.1 Espaces probabilisés

Définition 1 *Soit Ω l'ensemble des résultats possible d'une expérience aléatoire. On notera la plupart du temps un résultat isolé $\omega \in \Omega$.*

Les événements *sont des parties de Ω.*

Deux événements A *et* B *sont* incompatibles *si* $A \cap B = \emptyset$.

Une probabilité sur Ω est une fonction $\mathbb{P}$ de l'ensemble des parties (ou de certaines parties) de Ω à valeurs dans $[0,1]$, qui vérifie

— $\mathbb{P}(\Omega) = 1$

— *Si* A_1, A_2, *…, sont deux à deux incompatibles, alors*

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \sum_{k=1}^{+\infty} \mathbb{P}(A_k).$$

Un ensemble Ω muni d'une telle probabilité $\mathbb{P}$ est appelé un *espace probabilisé.*

Définition 2 *Deux événements* A *et* B *sont dit indépendants si*

$$\mathbb{P}(A \textit{ et } B) = \mathbb{P}(A) \times \mathbb{P}(B).$$

Plus généralement, des événements $A_1, \ldots, A_n$ *sont dit indépendants si*

$$\mathbb{P}(A_1 \textit{ et } A_2 \textit{ et } \cdots \textit{ et } A_n) = \mathbb{P}(A_1) \times \mathbb{P}(A_2) \times \cdots \times \mathbb{P}(A_n).$$

Dans un souci de lisibilité, on notera parfois (A ou B) l'union des événements A et B, c'est-à-dire $A \cup B$; de manière analoque (A et B) $= A \cap B$ est l'intersection de A et B, et (non A) $= \overline{A} = \Omega \setminus A$ est le complémentaire de A.

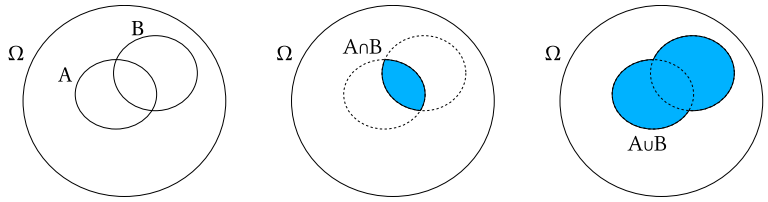


Figure 15. Intersection et union de A et B

Toutes les propriétés des probabilités se déduisent de ce qui précède. On peut par exemple en déduire les règles de calcul suivantes :

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$,
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$,
- $\mathbb{P}\left(\overline{A}\right) = 1 - \mathbb{P}(A)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$.

Ou encore, en notation « logique » :

- non (A ou B) = (non A) et (non B),
- non (A et B) = (non A) ou (non B),
- $\mathbb{P}$(non A) $= 1 - \mathbb{P}(A)$,
- $\mathbb{P}$(A ou B) $= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}$(A et B).
- $\mathbb{P}$(A ou B ou C) $= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}$(A et B) $- \mathbb{P}$(A et C) $- \mathbb{P}$(B et C) $+ \mathbb{P}$(A et B et C).

Les deux exemples qui suivent illustrent la relation entre théorie des probabilités, mesure de surface, ou dénombrement.

2.1.2 Exemple : la cible

On tire sur une cible Ω de surface finie ; on suppose que tous les points de la cible sont également susceptibles d'être atteints.

Une partie A de Ω s'interprète naturellement comme l'événement « atteindre un point de A ».

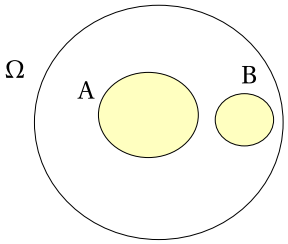


Figure 16. Une cible Ω et deux événements incompatibles $A, B \subset \Omega$

La probabilité de l'événement A est le rapport des surfaces de A et de Ω :

$$\mathbb{P}(A) = \frac{S(A)}{S(\Omega)}.$$

Si A et B sont des événements incompatibles, c'est-à-dire des parties disjointes de la cible, on a $S(A \cup B) = S(A) + S(B)$ et donc $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

On peut donner dans ce modèle un exemple d'événements indépendants : on prend pour A la moitié inférieure de la cible, et pour B la moitié droite. Alors $A \cap B$ est le quart inférieur droit, et on a bien $\mathbb{P}(A) = \frac{1}{2}$, $\mathbb{P}(B) = \frac{1}{2}$, et $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \frac{1}{4}$.

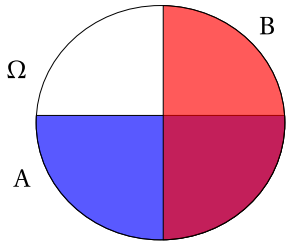


Figure 17. Deux événements indépendants.

Cet exemple illustre bien ce qu'on entend par « indépendance » de A et B : ici, savoir que le point atteint sur la cible est dans A ne donne aucune information sur le fait qu'il est ou non dans B.

On peut également utiliser ce modèle pour comprendre la formule

$$\mathbb{P}(A\cup B\cup C)=\mathbb{P}(A)+\mathbb{P}(B)+\mathbb{P}(C)-\mathbb{P}(A\cap B)-\mathbb{P}(A\cap C)-\mathbb{P}(B\cap C)+\mathbb{P}(A\cap B\cap C)$$

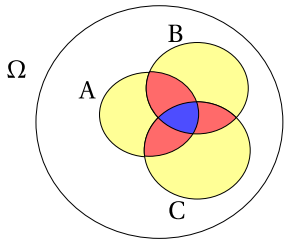


Figure 18. Calcul de $\mathbb{P}(A\cup B\cup C)$

En faisant la somme $S(A)+S(B)+S(C)$, on a compte deux fois les régions coloriées en rouge et trois fois les régions bleues. Il faut donc soustraire $S(A\cap B)$, $S(A\cap C)$, et $S(B\cap C)$, mais on a alors retiré trois fois la région bleue, il faut donc pour finir additionner $S(A\cap B\cap C)$.

Le premier à avoir introduit de tels modèles en probabilités est le naturaliste Buffon, avec le jeu du franc-carreau. Son exemple le plus célèbre, celui de l'« aiguille de Buffon », est dans la même veine.

2.1.3 Exemple : tirage uniforme dans un ensemble fini

Considérons un ensemble Ω fini ; on tire au hasard un des éléments de Ω, tous les éléments étant également susceptibles d'être choisis.

On note $|A|$ (*cardinal* de A) le nombre d'éléments d'une partie A de Ω. La probabilité de l'événement A (tirer un des éléments de la partie A) est naturellement

$$\mathbb{P}(A)=\frac{|A|}{|\Omega|}.$$

On a bien $\mathbb{P}(\Omega)=1$.

On parle alors de probabilité uniforme ; tous les éléments de Ω sont considérés comme équiprobables, et la probabilité d'un événement A s'obtient par dénombrement des éléments de A. On résume généralement cela par la formule

$$\boxed{\mathbb{P}(A)=\frac{|A|}{|\Omega|}=\frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}.}$$

Si des ensembles A_1, A_2, … sont deux à deux disjoints, alors le cardinal de leur union s'obtient en sommant leurs cardinaux respectifs :

$$|A_1\cup A_2\cup\cdots|=\sum_k |A_k|,$$

ce qui implique la propriété fondamentale des espaces probabilisés

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \sum_k \mathbb{P}(A_k).$$

On peut également illustrer sur cet exemple le calcul de $\mathbb{P}(A \cup B \cup C)$.

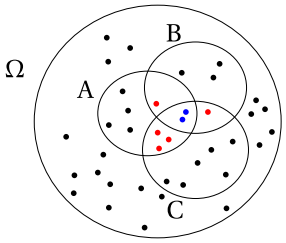


Figure 19. Calcul de $\mathbb{P}(A \cup B \cup C)$ dans le cas du dénombrement

2.1.4 Exemple : un coup de dé

Considérons l'expérience simple du lancer d'un dé dans une « piste » circulaire.

On se contentera de poser $\Omega = \{1,2,3,4,5,6\}$, en résumant un lancer par le chiffre porté par la face supérieure du dé quand il a fini de rouler ; on a alors $\mathbb{P}(\omega) = \frac{1}{6}$ pour tout ω. On a également, par exemple,

$$\mathbb{P}(\{1,3,5\}) = \mathbb{P}(1) + \mathbb{P}(3) + \mathbb{P}(5) = \frac{1}{2}.$$

Cette formalisation suffit amplement à répondre à toutes les questions élémentaires sur le jet de dés.

2.2 Probabilités conditionnelles

Définition 3 *Si* A *est un événement de probabilité non nulle, la probabilité conditionnelle de* B *sachant* A *est par définition*

$$\boxed{\mathbb{P}(B|A) = \frac{\mathbb{P}(A \textit{ et } B)}{\mathbb{P}(A)}.}$$

Notons que si A est un événement de probabilité non nulle, A et B sont indépendants ssi $\mathbb{P}(B|A) = \mathbb{P}(B)$. De même A et B sont incompatibles ssi $\mathbb{P}(B|A) = 0$.

2.2.1 Exemple de la cible

On peut illustrer cette définition avec l'exemple de la cible.

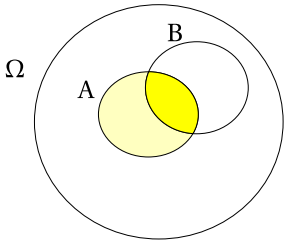


Figure 20. Calcul de $\mathbb{P}(B|A)$

Sachant que le point atteint est dans A, la probabilité qu'il soit dans B est le rapport des surfaces de $A \cap B$ et de A, donc

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

2.2.2 Formule des probabilités totales

Commençons par un cas particulier de la formule des probabilités totales :

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}\left(B|\overline{A}\right)\mathbb{P}\left(\overline{A}\right).$$

La preuve est simple : on écrit

$$B = (B \text{ et } A) \text{ ou } (B \text{ et non } A),$$

ou encore en notation ensembliste

$$B = (B \cap A) \cup \left(B \cap \overline{A}\right).$$

Les événements $(B \cap A)$ et $\left(B \cap \overline{A}\right)$ étant incompatibles, on a

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(B \cap A) + \mathbb{P}\left(B \cap \overline{A}\right) \\ &= \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}\left(B|\overline{A}\right)\mathbb{P}\left(\overline{A}\right).\end{aligned}$$

On prouve exactement de la même façon le résultat général qui suit :

Lemme 1 (Formule des probabilités totales) *Soit* $A_1, \ldots, A_n$ *une famille d'événements incompatibles telle que* $\Omega = A_1$ *ou* $\cdots$ *ou* A_n*. On a*

$$\boxed{\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \cdots + \mathbb{P}(B|A_n)\mathbb{P}(A_n).}$$

2.2.3 Formule de Bayes

La formule de Bayes permet de relier la probabilité de A sachant B à la probabilité de B sachant A.

Lemme 2 *Soient* A *et* B *deux évéments de probabilité non nulle. On a*

$$\boxed{\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}\left(B|\overline{A}\right)\mathbb{P}\left(\overline{A}\right)}.\end{aligned}}$$

La preuve de la première formule se déduit immédiatement des égalités

$$\begin{aligned}\mathbb{P}(A \text{ et } B) &= \mathbb{P}(A|B)\mathbb{P}(B), \\ &= \mathbb{P}(B|A)\mathbb{P}(A).\end{aligned}$$

Pour la seconde formule, on a simplement exprimé $\mathbb{P}(B)$ à l'aide de la formule des probabilités totales.

Exemple 1 Dans une population donnée, on estime l'incidence de la tuberculose à 1 cas pour 10000 personnes-années et la prévalence du VIH à 0,4%. Chez 15% des nouveaux cas de tuberculose, on diagnostique une infection au VIH.

Soit TB l'événement : « déclarer la tuberculose pendant une période d'un an » et VIH l'événement : « être VIH+ ». Les données de l'énoncé se traduisent en termes de probabilités par $\mathbb{P}(TB) = 10^{-4}$, $\mathbb{P}(VIH) = 4 \cdot 10^{-3}$, et $\mathbb{P}(VIH \,|\, TB) = 15 \cdot 10^{-2}$.

On peut maintenant utiliser la formule de Bayes pour estimer l'incidence de la tuberculose chez les personnes VIH+ :

$$\begin{aligned}\mathbb{P}(TB \,|\, VIH) &= \frac{\mathbb{P}(VIH \,|\, TB)\mathbb{P}(TB)}{\mathbb{P}(VIH)} \\ &= \frac{15 \cdot 10^{-2} \times 10^{-4}}{4 \cdot 10^{-3}} \\ &= 3{,}75 \cdot 10^{-3}\end{aligned}$$

soit 3,75 cas par 1 000 personnes-années. □

2.3 Quelques exemples de calculs de probabilités élémentaires

Nous montrons ici comment quelques calculs simples peuvent s'écrire dans le formalisme des espaces probabilisés. En pratique on pourra s'en passer, l'idée est ici d'apprivoiser l'utilisation d'un ensemble d'expériences Ω.

2.3.1 Lancer un dé

Si on note A_k (pour $k = 1,\dots,6$) l'événement

$$A_k = \text{on a obtenu } k$$

soit, si $\Omega = \{1,2,3,4,5,6\}$,

$$A_k = \{k\}$$

on a $\mathbb{P}(A_k) = \frac{1}{6}$ pour $k = 1,\dots,6$. On a donc par exemple

$$\mathbb{P}(\text{X impair}) = \mathbb{P}(A_1 \cup A_3 \cup A_5) = \mathbb{P}(A_1) + \mathbb{P}(A_3) + \mathbb{P}(A_5) = \frac{1}{2}.$$

2.3.2 Lancer un dé deux fois de suite

Pour décrire une telle expérience composée de deux expériences élémentaires, il faut un ensemble Ω plus grand, un ensemble dont chaque élément correspond à deux lancers de dés. On prendra par exemple

$$\Omega = \{1,\dots,6\}^2 = \{(\omega_1,\omega_2) : \omega_1,\omega_2 \in \{1,\dots,6\}\}.$$

l'ensemble des couples d'éléments de $\{1,2,3,4,5,6\}$. Ainsi chaque élément de Ω, c'est-à-dire chaque expérience considérée, est constitué de deux « expériences élémentaires ».

On calcule la probabilité d'un élément de Ω en écrivant

$$\mathbb{P}((\omega_1,\omega_2)) = \left(\frac{1}{6}\right)^2 = \frac{1}{36},$$

ce qui revient à écrire que tous les couples sont équiprobables ; c'est simplement la traduction du fait que les lancers sont indépendants.

On peut visualiser Ω par une table comme celle-ci, où chaque case correspond à un élément de Ω. Toutes les cases sont équiprobables, il y a $6^2 = 36$ cases toutes de probabilité $\frac{1}{36}$.

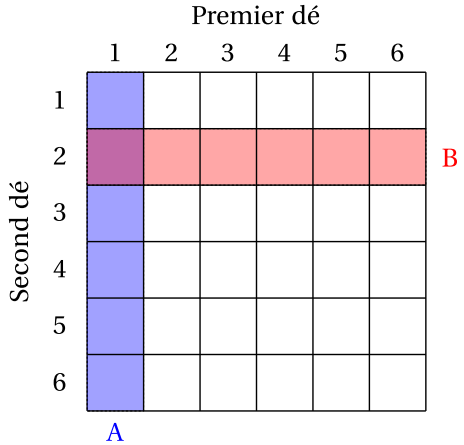


Figure 21. Lancers de deux dés représentés par une table

La probabilité d'un événement A s'obtient en comptant le nombre d'éléments ω dans A. On a ainsi

$$\mathbb{P}(A) = \frac{1}{6^2} \times \text{nombre d'éléments de A}$$

soit la formule classique

$$\mathbb{P} = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}.$$

Si on prend par exemple

$$A = \{\text{premier dé donne 1}\} = \{(1, \omega_2)\},$$

et

$$B = \{\text{second dé donne 2}\} = \{(\omega_1, 2)\},$$

A a 6 éléments (les cases coloriées en bleu dans la table ci-dessus) et on a $\mathbb{P}(A) = \frac{1}{6^2} \times 6 = \frac{1}{6}$. De même B a 6 élements (cases rouges) et $\mathbb{P}(B) = \frac{1}{6}$, ce qui est bien le résultat attendu !
D'autre part l'intersection de A et B a un seul élement :

$$A \cap B = \{(1,2)\},$$

et $\mathbb{P}(A \cap B) = \mathbb{P}(A \text{ et } B) = \frac{1}{6^2} = \frac{1}{36} = \mathbb{P}(A)\mathbb{P}(B)$, ce qui est également ce qu'on attend pour deux lancers indépendants.

2.3.3 Avec le Chevalier de Méré : lancer un dé quatre fois

De façon analogue à ce qu'on vient de faire pour deux dés, notre ensemble d'expérience est

$$\Omega = \{1, \ldots, 6\}^4 = \{(\omega_1, \omega_2, \omega_3, \omega_4) : \omega_1, \ldots, \omega_4 \in \{1, \ldots, 6\}\},$$

avec la probabilité

$$\mathbb{P}((\omega_1, \omega_2, \omega_3, \omega_4)) = \left(\frac{1}{6}\right)^4,$$

Considérons la première expérience du Chevalier de Méré (voir section **??**) : quelle est la probabilité d'obtenir un six parmi les quatre lancers ? Soit C l'événement « on a obtenu (au moins) un six parmi les quatre lancers » :

$$C = \{(\omega_1, \omega_2, \omega_3, \omega_4) : \omega_1 = 6 \text{ ou } \omega_2 = 6 \text{ ou } \omega_3 = 6 \text{ ou } \omega_4 = 6\}.$$

Il faut compter les éléments de C, ce qui n'est pas simple. Il est plus simple de calculer la probabilité de $\overline{C}$, l'événement complémentaitre : « ne faire aucun six sur les 4 lancers », ce qui revient à compter le nombre d'élément de

$$\overline{C} = \{(\omega_1, \omega_2, \omega_3, \omega_4) : \omega_1 \neq 6 \text{ et } \omega_2 \neq 6 \text{ et } \omega_3 \neq 6 \text{ et } \omega_4 \neq 6\}.$$

Chaque ω_i peut prendre 5 valeurs, donc le nombre d'éléments de $\overline{C}$ est 5^4, et on a $\mathbb{P}\left(\overline{C}\right) = \left(\frac{5}{6}\right)^4$.
Il était également possible de définir les événements D_1, D_2, D_3 et D_4, définis par D_i = « on n'a pas fait 6 au i^elancer ». On a $\mathbb{P}(D_i) = \frac{5}{6}$, et $\overline{C} = D_1 \cap D_2 \cap D_3 \cap D_4$ d'où, par indépendance des D_i, $\mathbb{P}(\overline{C}) = \mathbb{P}(D_1) \times \cdots \times \mathbb{P}(D_4) = \left(\frac{5}{6}\right)^4$.
On calcule enfin

$$\mathbb{P}(C) = 1 - \mathbb{P}(\overline{C}) = 1 - \left(\frac{5}{6}\right)^4 \simeq 0{,}518.$$

Le premier pari considéré par le chevalier est avantageux, comme il l'avait constaté. L'étude du second pari est laissée en exercice.

La façon *cool* de rédiger ça

En pratique, on se contente de dire : les expériences étant indépendantes, la probabilité de ne faire six à aucun des 4 tirages est

$$\left(\frac{5}{6}\right)^4,$$

et la probabilité d'avoir fait au moins une fois six est donc

$$1-\left(\frac{5}{6}\right)^4.$$

2.4 Lancer un nombre infini de dés ?

Le silence éternel de ces espaces infinis m'effraie.
Blaise Pascal.

Attention, ce paragraphe peut occasionner de légers vertiges à certains sujets sensibles. Il pourra aisément être omis en première lecture.

Le formalisme mathématique présenté prend tout son sens quand il s'agit de penser une expérience aussi vertigineuse que celle annoncée : lancer un nombre infini de dés ! Pourquoi faire ? Parce qu'on se pose des questions comme celle-ci : si on fait la moyenne

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

des valeurs $X_1, \ldots, X_n$ obtenues par n lancers de dés indépendants, quelle est la limite de $\overline{X}_n$ quand n tend vers l'infini ?

Comment faire ? De même qu'il était possible de résumer quatre lancers de dés en une seule expérience $\omega = (\omega_1, \ldots, \omega_4)$, le formalisme permet de résumer un nombre infini de lancers en une seule expérience $\omega = (\omega_1, \omega_2, \ldots)$. L'ensemble de ces expériences est noté $\Omega^{\mathbb{N}}$:

$$\Omega^{\mathbb{N}} = \{(\omega_1, \omega_2, \ldots) : \omega_i = 1, 2, \ldots, 6.\}.$$

Cela permet d'écrire $\overline{X}_n(\omega)$ à la place de $\overline{X}_n$; toute l'incertitude est reportée sur le choix d'un seul élément $\omega \in \Omega^{\mathbb{N}}$. On a $X_i(\omega) = \omega_i$ et $\overline{X}_n(\omega) = \frac{1}{n}(\omega_1 + \cdots + \omega_n)$.

On verra plus tard (cf 4.5.4) que la valeur naturelle à attendre pour la limite de $\overline{X}_n$ quand n tend vers l'infini est la valeur moyenne (l'espérance) d'un lancer de dé, soit

$$\lim_{n\to\infty} \overline{X}_n(\omega) = \frac{1}{6} \times 1 + \cdots + \frac{1}{6} \times 6 = 3{,}5.$$

Cela n'est évidemment pas le cas pour tous les ω envisageables, par exemple ça n'est pas le cas pour $\omega^{(1)} = (1, 1, 1, \ldots)$ (que des 1 !). On a alors pour tout n, $\overline{X}_n\left(\omega^{(1)}\right) = 1$, et la limite vaut 1.

Cependant considérons l'ensemble des ω pour lesquels la limite des $\overline{X}_n(\omega)$ vaut 3,5 :

$$A = \left\{\omega \in \Omega^{\mathbb{N}} : \lim_{n\to\infty} \overline{X}_n(\omega) = 3{,}5\right\}.$$

La loi *forte* des grands nombres assure que $\mathbb{P}(A) = 1$! C'est-à-dire que les ω « pathologiques » pour lesquels la limite n'est pas 3,5 existent (les moyennes calculées pourraient osciller indéfiniement) mais la probabilité de tirer un ω « non-pathologique » pour lequel la limite vaut 3,5 est égale à 1 : il est *certain* qu'on tirera $\omega \in A$. On est certainement dans un monde « non-pathologique », un monde où la moyenne des X_i tend vers 3,5.

Commentaire imagé

L'intérêt de cet ensemble « compliqué » $\Omega^{\mathbb{N}}$ est de pouvoir imaginer qu'on a lancé un nombre infini de dés d'un seul coup, au lieu d'imaginer qu'on répète les lancers jusqu'à l'infini.

Si on a un seau qui contient 60 dés, qu'on le renverse au sol, et qu'on compte les 1, les 2, etc, on n'a pas exactement 10 résultats de chaque ; un petit écart à la proportion $\frac{1}{6}$ est possible. L'écart sera moins grand si on a un seau de 600 dés ; et si on dispose d'un seau contenant un nombre infini de dés (et d'une pièce infinie où le renverser) la loi des grands nombres nous dit qu'avec une probabilité égale à 1 on observera exactement la proportion $\frac{1}{6}$.

2.5 Techniques de dénombrement

Le calcul d'une probabilité, dans le cas de tirages équiprobables, par la formule

$$\frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}},$$

demande de savoir dénombrer (compter !) les cas. Nous en avons vu quelques exemples dans le cas des lancers de dés. Nous allons rapidement donner quelques techniques utiles.

Le modèle favori des probabilistes est l'urne ; l'urne du probabiliste est un récipient contenant des boules qu'on tire à l'aveuglette. Ces boules sont de couleurs variées, éventuellement numérotées. À chaque tirage, toutes les boules présentes dans l'urne ont la même chance d'être choisies.

2.5.1 Permutations de k objets

On considère k boules numérotées de 1 à k, placées dans une urne. On les en sort une à une, et on s'intéresse à l'ordre dans lequel elles sortent. L'ensemble Ω des résultats possibles, appelé *ensemble des permutations de* $\{1,\dots,k\}$ est donc

$$\Omega = \left\{(\omega_1,\dots,\omega_k) : \text{si } i \neq j,\ \omega_i \neq \omega_j \text{ et chaque } \omega_i \in \{1,\dots,k\}\right\}.$$

Tous ces résultats sont bien sûr équiprobables. Quelle est la probabilité d'obtenir l'un d'entre eux, par exemple $(1,2,\dots,k)$? Il faut compter les élements de Ω.

Commençons par des cas simples : avec $k = 1$, il y a une façon de faire, avec $k = 2$, deux résultats sont possibles : $(1,2)$ et $(2,1)$; avec $k = 3$ les six possibilités sont

$$\begin{array}{lll} (1,2,3) & (2,1,3) & (3,1,2) \\ (1,3,2) & (2,3,1) & (3,2,1) \end{array}$$

On voit se dessiner le schéma général ; k façons de choisir le premier élément ; pour chacun de ces choix, $k-1$ choix pour le second élements, puis $k-2$ pour le troisième... jusqu'au dernier élément où un seul choix est possible.

Le nombre de permutations est donc

$$\boxed{k! = k \times (k-1) \times \cdots \times 2 \times 1.}$$

On lit $k!$: « k factorielle » ou « factorielle de k ».

2.5.2 Choix de k objets parmi n : les coefficients binomiaux

On considère n boules numérotées de 1 à n placées dans une urne ; on en sort k une à une. On s'intéresse au résultat de ce tirage.

Quand l'ordre importe

Quand l'ordre importe, l'ensemble Ω des résultats possibles est

$$\Omega = \left\{(\omega_1, \ldots, \omega_k) : \text{si } i \neq j,\ \omega_i \neq \omega_j \text{ et chaque } \omega_i \in \{1, \ldots, n\}\right\}.$$

On l'appelle *ensemble des arrangements de k élements de* $\{1, \cdots, n\}$.

Il y a n façons de choisir ω_1, puis pour chacun de ces choix $n-1$ façons de choisir ω_2... jusqu'au dernier élément ω_k pour lequel $n-k+1$ choix sont possibles. Le nombre d'arrangements de k éléments pris parmi n est donc

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

Quand l'ordre n'importe pas

Intéressons-nous aux résultats des tirages quand l'ordre n'importe pas.

L'ensemble des résultats possibles, appelé « combinaisons de k éléments », est

$$\Omega = \left\{\{\omega_1, \ldots, \omega_k\} : \text{si } i \neq j,\ \omega_i \neq \omega_j \text{ et chaque } \omega_i \in \{1, \ldots, n\}\right\}.$$

(Par convention, $\{\omega_1, \ldots, \omega_k\}$ note un ensemble à k éléments, et l'ordre des éléments dans un ensemble n'importe pas.)

Chaque combinaison peut être ordonnée de $k!$ façon possibles pour obtenir un arrangement de k éléments ; on sait qu'il y a $\frac{n!}{(n-k)!}$ arrangements, on en déduit que le nombre de combinaisons de k éléments parmi n est

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

C'est le nombre de façons de choisir k éléments parmi n, on lit généralement $\binom{n}{k}$ par « k parmi n ». On remarque que

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

ce qui s'explique très logiquement : choisir les k boules que l'on retire de l'urne est équivalent à choisir les $n-k$ boules qu'on y laisse.

Le binôme de Newton et le triangle de Pascal

La *formule du binôme de Newton* est

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Pour cette raison, les $\binom{n}{k}$ sont appelés *coefficients binomiaux*.

Le triangle de Pascal est un moyen simple de calculer les coefficients binomiaux en utilisant la relation $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$.

	$k=0$	1	2	3	4	5
$n=0$	1					
1	1	1				
2	1	2	1			
3	1	3	3	1		
4	1	4	6	4	1	
5	1	5	10	10	5	1

Table 8. Le triangle de Pascal

Chaque élément du triangle est formé par la somme de l'élément qui est juste au-dessus de lui et de celui qui est au-dessus et une case à gauche.

Pour calculer un coefficient binomial isolé, on aura soin de n'effectuer les multiplications qu'après avoir simplifié la fraction :

$$\begin{aligned}\binom{11}{4} &= \frac{11!}{7!\,4!} \\ &= \frac{2\times 3\times\cdots\times 11}{(2\times 3\times 4)\times(2\times 3\times\cdots\times 7)} \\ &= \frac{8\times 9\times 10\times 11}{2\times 3\times 4} \\ &= 3\times 10\times 11 \\ &= 330.\end{aligned}$$

2.6 Et les statistiques dans tout ça ?

Les probabilités sont la règle du jeu auquel jouent les statisticiens.

Si on nous donne un dé à 6 faces, on peut légitimement faire *l'hypothèse* qu'il s'agit d'un *dé honnête*, c'est-à-dire que chaque face a une probabilité $\frac{1}{6}$ de sortir. On peut alors réaliser des expériences aléatoires (lancer le dés, relever le résultat) pour *tester cette hypothèse* (par exemple en faisant un test du χ^2).

On peut également essayer *d'estimer* au mieux la probabilité de chaque face, tout simplement en réalisant un grand nombre d'expérience et en estimant les probabilités inconnues par les fréquences observées pour chacune des six faces. Ces probabilitées estimées pourraient éventuellement également servir à *prédire* au mieux les valeurs des futurs lancers.

Ces problèmes statistiques, *test d'hypothèse*, *estimation de paramètres inconnus*, et *prédiction*, se comprennent, se formalisent dans le cadre de la théorie des probabilités.

2.7 Exercices

Exercice 1 Inégalité de Boole On considère des évéments $A_1, A_2, \ldots, A_n$. Montrez graphiquement l'inégalité suivante :

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

Quand a-t-on égalité ?

Exercice 2 On considère k boules numérotées de 1 à k, placées dans une urne. On les en sort une à une ; quelle est la probabilité que la boule portant le numéro 1 soit tirée avant la boule portant le numéro 2 ?

Exercice 3 Les paris du Chevalier de Méré

1. On lance quatre fois de suite un dé à six faces. Quelle est la probabilité d'obtenir au moins une fois un six ?

2. On lance vingt-quatre fois de suite deux dés à six faces. Quelle est la probabilité d'obtenir au moins une fois un double-six ?

Exercice 4 1. On tire une carte dans un jeu de 32 cartes. Calculer la probabilité de :
- tirer un as ;
- tirer un as noir ;
- tirer un as ou une carte noire.

2. On tire deux cartes l'une après l'autre, sans remise, dans un jeu de 32 cartes. Calculer la probabilité de
   — tirer un as, puis un roi ;
   — tirer une dame, puis un cœur.
3. Même chose pour un tirage avec remise.
4. On tire deux cartes à la fois dans un jeu de 32 cartes. Calculer la probabilité de
   — tirer un as et un roi ;
   — tirer une dame et un cœur.

Exercice 5 On lance deux dés à 6 faces. Calculez la probabilité des événements suivants.

$$\begin{aligned} A &= \{\text{Le premier dé donne un résultat pair}\} \\ B &= \{\text{Le second dé donne un résultat impair}\} \\ C &= \{\text{Les deux dés donnent un résultat de même parité}\} \end{aligned}$$

Ces événements sont-ils indépendants ? Sont-ils deux à deux indépendants ?

Exercice 6 On considère une urne qui contient 5 boules rouges, 5 boules noires, 4 boules vertes. On en tire 4 boules au hasard et sans remise.

1. Quelle est la probabilité d'avoir tiré exactement 2 boules rouges et 2 boules vertes ?
2. Quelle est la probabilité d'avoir tiré 2 boules rouges, une boule noire, une boule verte ?
3. Quelle est la probabilité que les 4 boules tirées aient la même couleur ?

Exercice 7 Trois garçons et quatre filles vont au cinéma. Ils s'installent sur une rangée de sept sièges. En supposant qu'ils s'asseoient au hasard, quelle est la probabilité pour que chaque garçon soit entouré de deux filles ?

Exercice 8 Paradoxe des anniversaires

1. Quelle la probabilité p_N pour que, dans un groupe de N étudiants, au moins deux étudiants fêtent leur anniversaire le même jour ? (On négligera les années bissextiles)
2.* En utilisant la formule approchée

$$\prod_{i=1}^{N}(1-x_i) \simeq \exp\left(-\sum_{i=1}^{N} x_i\right)$$

valable quand les x_i sont petits, calculer p_N pour $N = 30$.

Exercice 9 On considère un caractère mendélien, gouverné par un gène autosomal diallélique A/a où A domine a. Soit p la fréquence de l'allèle A et $q = 1-p$ la fréquence de l'allèle a. On suppose que les deux allèles portés par un individu sont indépendants.

1. Montrer qu'alors la probabilité qu'un individu ait pour génotype AA (respectivement Aa, aa) est p^2 (respectivement $2pq$, q^2).
2. Un individu a le phénotype dominant. Quel est la probabilité qu'il soit porteur d'un allèle a ?
3. [Question pénible !] Même question en supposant que ses parents ont le phénotype dominant.

Indication : Noter P_0, P_1 *les événements « père AA », « père Aa », et similairement* M_0, M_1 *pour le genotype de la mère,* E_0, E_1 *pour le génotype de l'enfant. La probabilité demandée est*

$$\mathbb{P}(E_1 | (E_0 \cup E_1) \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)).$$

Exercice 10 Monsieur Labayat a deux enfants dont un (au moins) est un garçon. Quelle est la probabilité que l'autre soit une fille ?

Exercice 11 (Problème de Monty Hall, d'après Wikipédia) *

Dans la phase finale d'un jeu télévisé, un candidat est placé devant trois portes fermées. Derrière l'une d'elles se trouve une voiture et derrière chacune des deux autres se trouve une chèvre.

Le candidat doit tout d'abord désigner une porte. Le présentateur du jeu ouvre alors une des deux portes restantes, derrière laquelle se trouve une chèvre (il sait où est la voiture).

Le candidat a alors le droit ou bien d'ouvrir la porte qu'il a choisie initialement, ou bien de modifier son choix et d'ouvrir la porte qui reste.

Quel est la meilleure stratégie ?

Exercice 12 Il existe des dés à 20 faces (qui sont des icosaèdres réguliers). Les faces sont numérotées de 1 à 20 ; on considérera des dés honnêtes, c'est-à-dire que tous les tirages possibles sont équiprobables.

1. On lance un dé à 20 faces ; sachant que le résultat du tirage est un nombre pair, quelle est la probabilité que ce résultat soit 2 ?

2. On lance deux dés à 20 faces ; quelle est la probabilité qu'un des deux au moins donne le chiffre 1 ?

Exercice 13 On lance un dé à 20 faces k fois. Si on a tiré un 1 au moins une fois, on gagné un petit gateau au chocolat.

1. Quel est la probabilité d'avoir un gateau quand $k = 1$? Quand $k = 2$?

2. Même question pour k quelconque. Y a-t-il une valeur de k au-delà de laquelle on a 90% de chance d'avoir un gateau ? Quelle est la valeur limite de cette probabilité quand k tend vers l'infini ?

Chapitre 3

Variables aléatoires

On considère dans ce chapitre un espace probabilisé $(\Omega, \mathbb{P})$ (cf définition en 2.1.1). Nous allons aller un peu plus avant en introduisant la notion de variable aléatoire, qui est simplement une mesure réalisée sur une expérience aléatoire.

Si Ω est l'ensemble des résultats possibles d'une expérience aléatoire, une mesure $X(\omega)$ réalisée sur $\omega \in \Omega$ est naturellement une fonction de Ω dans $\mathbb{R}$.

3.1 Variables aléatoires : définition générale

Définition 4 *Une* variable aléatoire *est une fonction* X *d'un ensemble* Ω *probabilisé, à valeurs dans* $\mathbb{R}$.

Définition 5 *La* loi *de* X *est la donnée, pour tous* $a \leqslant b \in \mathbb{R}$, *de la probabilité* $\mathbb{P}(a < X \leqslant b)$.

On admettra qu'il suffit de connaître les probabilités de $\mathbb{P}(a < X \leqslant b)$ pour tous les réels $a \leqslant b$, pour pouvoir calculer $\mathbb{P}(X \in A)$ pour toute partie A de $\mathbb{R}$. Ceci permet, quand on travaille avec des variables aléatoires, d'« oublier » l'espace probabilité Ω : la loi des variables suffit.

Remarque (transformation des données) : Si X est une variable aléatoire et $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction, alors $Y = \phi(X)$ est une variable aléatoire.

3.2 Variables aléatoires discrètes

3.2.1 Définition

On dit qu'une variable aléatoire X est *discrète* si elle ne peut prendre qu'un nombre dénombrable ou fini de valeurs $x_1, x_2, \ldots$.

Dans le cas dénombrable, en pratique X prendra souvent ses valeurs dans $\mathbb{N}$, mais si par exemple X peut prendre les valeurs 0, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, etc, X sera une variable discrète.

La loi de X est entièrement déterminée par la fonction de masse $\mathbb{P}(X = x)$ quand x décrit l'ensemble des valeurs possibles de X. On a pour tout A

$$\boxed{\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x).}$$

Exemple 2 On prend $\Omega = \{P, F\}$ et on assigne à P et F la probabilité $\frac{1}{2}$. On pose $X(P) = 0$ et $X(F) = 1$. X est une variable aléatoire discrète qui peut prendre les valeurs 0 et 1. La v.a. $Y = X - \frac{1}{2}$ est une variable aléatoire discrète qui peut prendre les valeurs $-\frac{1}{2}$ et $\frac{1}{2}$. □

Exemple 3 Soit A un ensemble fini (de nombres réels). On note |A| le cardinal de A (le nombre d'éléments de A).

Une variable X suit la loi uniforme sur A si pour tout $x \in A$, on a $\mathbb{P}(X = x) = \frac{1}{|A|}$. C'est une loi discrète. □

Exemple 4 On considère $\Omega = \{1,2,3,4,5,6\}$ (cf 2.1.4) avec $\mathbb{P}(\omega) = \frac{1}{6}$ pour $\omega \in \Omega$. La variable aléatoire qui donne le résultat d'un lancer de dés peut être définie simplement par $X(\omega) = \omega \in \mathbb{R}$. Sa loi est

$$\mathbb{P}(X = x) = \frac{1}{6} \text{ pour } x \in \{1,\dots,6\}.$$

Dans un cas de ce type, considérer l'ensemble Ω est inutile, on aurait pu directement donner la loi de X sans parler de l'espace probabilisé sous-jacent, ce qui ferait de cet exemple un cas particulier de l'exemple 3. □

Exemple 5 On définit une variable aléatoire par sa loi comme suit :

$$\mathbb{P}(X_E = k) = c \times \frac{1}{k^2} \text{ pour } k \in \mathbb{N}^\star$$

où c est une constante choisie pour que $\sum_{k\in\mathbb{N}^\star} \mathbb{P}(X_E = k) = 1$ (on sait grâce à Euler qu'il faut prendre $c = \frac{6}{\pi^2}$).

Cette variable prend un nombre dénombrable de valeurs : on peut les énumérer une à une (ce sont les valeurs 1, 2, etc). □

3.2.2 Espérance d'une variable aléatoire discrète

L'espérance de X est

$$\boxed{E(X) = \sum_x x\,\mathbb{P}(X = x)}$$

si cette quantité est définie (la somme se fait sur toutes les valeurs x possibles).

C'est donc la moyenne des valeurs x que peut prendre X, pondérées par la probabilité $\mathbb{P}(X = x)$ que cette valeur soit prise ; c'est la valeur moyenne de X, ou encore la *valeur attendue* de X (espérer vient du latin, *sperare* : « attendre quelque chose comme devant se réaliser » ; en anglais, on parle d'*expected value*).

Plus généralement, si ϕ est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, l'espérance de $\phi(X)$ est

$$\boxed{E(\phi(X)) = \sum_x \phi(x)\,\mathbb{P}(X = x).}$$

Le résultat suivant est important.

Proposition 3 *Si* X *et* Y *sont deux v.a. et a et b sont des constantes, on a*

$$\boxed{\begin{array}{rcl} E(aX + b) & = & aE(X) + b \\ E(aX + bY) & = & aE(X) + bE(Y). \end{array}}$$

Ce résultat est connu sous le nom de linéarité de l'espérance.

La preuve sera faite en exercice.

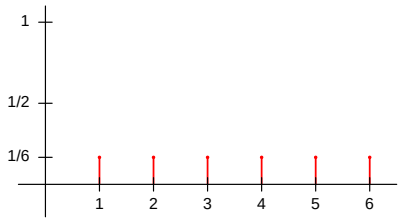


Figure 22. Représentation de la loi de X

x	1	2	3	4	5	6
$\mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 9. Table de la loi de X

Exemple 6 Poursuivons l'exemple 4. La loi de X est donnée par

$$\mathbb{P}(X = x) = \frac{1}{6} \text{ pour } x \in \{1,\dots,6\}.$$

Son espérance est $E(X) = \sum_{x=1}^{6} x \times \frac{1}{6} = \frac{21}{6} = \frac{7}{2}$.
Soit $Y = |X-3|$. Son espérance est $E(Y) = \sum_{x=1}^{6} |x-3| \times \frac{1}{6} = \frac{1}{6}(2+1+0+1+2+3) = \frac{9}{6} = \frac{3}{2}$. On peut compléter cet exemple en tabulant la loi de Y, ce qui explicite un calcul qui apparaît en filigranne dans le calcul de E(Y) ci-dessus : on a par exemple, $\mathbb{P}(Y = 0) = \mathbb{P}(X = 3) = \frac{1}{6}$, $\mathbb{P}(Y = 1) = \mathbb{P}(X \in \{2,4\}) = \mathbb{P}(X = 2) + \mathbb{P}(X = 4) = \frac{2}{6}$, etc.

x	0	1	2	3
$\mathbb{P}(X = x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

Table 10. Table de la loi de Y

□

Exemple 7 Reprenons la loi définie à l'exemple 5, par

$$\mathbb{P}(X = k) = c \times \frac{1}{k^2}$$

avec $c = \frac{\pi^2}{6}$. On peut montrer que

$$\sum_{k=1}^{n} k\mathbb{P}(X = k) \underset{n\to+\infty}{\longrightarrow} +\infty,$$

donc X n'a pas une espérance finie. □

3.3 Variables aléatoires continues à densité

3.3.1 Définition

Définition 6 *Soit* X *une variable aléatoire. On dit que* X *est une variable aléatoire continue de densité* f *si pour tous réels* $a \leqslant b$ *on a*

$$\boxed{\mathbb{P}(a \leqslant \mathrm{X} \leqslant b) = \int_a^b f(x)\,\mathrm{d}x.}$$

Dans ce cas on a pour tout $x \in \mathbb{R}$, $f(x) \geqslant 0$. *On a également* $\displaystyle\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$.

On a également, pour tout a, $\mathbb{P}(\mathrm{X} = a) = 0$ — attention à ne pas tomber dans le piège d'écrire $\mathbb{P}(\mathrm{X} = a) = f(a)$!

On a par contre l'approximation suivante :

$$\mathbb{P}\left(a - \frac{1}{2}h \leqslant x \leqslant a + \frac{1}{2}h\right) = \int_{a-\frac{1}{2}h}^{a+\frac{1}{2}h} f(x)\,\mathrm{d}x \simeq h \cdot f(a).$$

Cette approximation est illustrée figure 23 : l'intégrale considérée est l'aire de la figure dessinée en bleu, qui est approximativement celle d'un rectangle de largeur h et de hauteur $f(a)$. Plus h est petit, plus cette approximation est de bonne qualité.

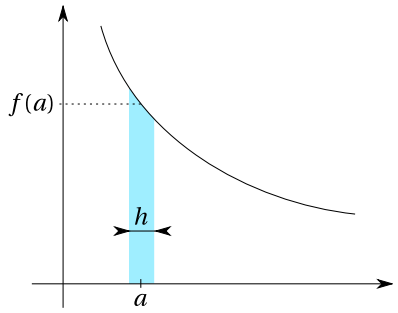


Figure 23. Probabilité d'être entre $a - \frac{1}{2}h$ et $a + \frac{1}{2}h$ quand h petit

En pratique, on considère souvent des variables continues pour modéliser une mesure discrète. Ainsi, en tout rigueur, la taille d'un individu tiré au hasard dans une population *finie* est une variable discrète (d'une part parce que la population est finie, d'autre part parce que la mesure est arrondie : une taille est mesurée en général au centimètre près...). Cependant si la population considérée est assez grande, si la mesure est assez précise, on sera amené à la modéliser par une loi continue (par exemple, sous certaines conditions, par une gaussienne, ou par un mélange de gaussiennes).

3.3.2 Espérance d'une variable aléatoire continue

L'espérance de X est

$$\boxed{\mathrm{E}(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx,}$$

si cette quantité est définie. Cette formule est très semblable à celle utilisée pour l'espérance d'une variable discrète : la somme discrète est simplement remplacée par une intégrale, une « somme continue ».

Plus généralement, si ϕ est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, l'espérance de $\phi(\mathrm{X})$ est

$$\boxed{\mathrm{E}(\phi(\mathrm{X})) = \int_{-\infty}^{+\infty} \phi(x) f(x) dx.}$$

La linéarité de l'espérance (proposition 3) reste vraie.

Exemple 8 La loi uniforme sur $[a,b]$, notée $\mathscr{U}([a,b])$, est la loi de densité

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

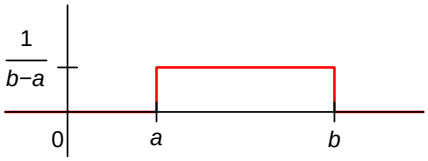


Figure 24. Densité de $\mathscr{U}([a,b])$

Calculons son espérance. On a

$$\begin{aligned} \mathrm{E(X)} &= \int_a^b \frac{x}{b-a}\,\mathrm{d}x \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b \\ &= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2.} \end{aligned}$$

□

3.4 Un exemple de variable aléatoire « mixte »

Croire que toute variable aléatoire est nécessairement discrète ou continue à densité serait un raccourci trop rapide. On peut facilement imaginer des situations intermédiaires. Considérons par exemple les expériences aléatoires suivantes : à chaque expérience, on fait un tirage à pile ou face, et un tirage aléatoire suivant une loi uniforme $\mathscr{U}([0,1])$. On peut écrire chaque expérience sous la forme $\omega = (\omega_1, \omega_2)$ avec $\omega_1 \in \{\mathrm{P,F}\}$ et $\omega_2 \in [1,2]$.

On définit une variable aléatoire X comme suit : pour tout $\omega = (\omega_1,\omega_2)$, si $\omega_1 = \mathrm{P}$ (pile) on pose $\mathrm{X}(\omega) = 0$, sinon on pose $\mathrm{X}(\omega) = \omega_2$.

On a ainsi $\mathbb{P}(\mathrm{X}=0) = \frac{1}{2}$ et, pour $a,b \in]0,1]$, $\mathbb{P}(a \leqslant \mathrm{X} \leqslant b) = \frac{1}{2}(b-a)$.

On peut imaginer une multitude de variables « mixtes » de ce type ; elles peuvent être la source de problèmes méthodologiques si on leur applique des méthodes conçues pour les variables continues.

Exemple 9 On s'intéresse au temps de survie T des patients après une opération. Certains meurent au bloc opératoire, on a donc $\mathbb{P}(\mathrm{T}=0) > 0$; pour les autres, T sera naturellement modélisé par une loi continue. □

Exemple 10 On mesure un taux d'anticorps A chez des individus exposés à un agent infectieux. Certains n'en ont pas du tout, d'où à nouveau une « masse positive » en 0 : $\mathbb{P}(\mathrm{A}=0) > 0$; et les valeurs positives de ce taux seront modélisés par une loi continue. □

3.5 Variance d'une variable aléatoire

Définition 7 *Soit* X *une variable aléatoire (continue ou discrète). Notons* $\mu = \mathrm{E(X)}$. *La variance de* X *est*

$$\boxed{\mathrm{var(X)} = \mathrm{E}\left((\mathrm{X}-\mu)^2\right)}$$

(si cette quantité est définie). Sa racine carrée est l'écart-type *de* X. *On note souvent l'écart-type* σ *et la variance* σ^2.

Un point de vocabulaire : La quantité $X-E(X)$ est *l'écart algébrique (à l'espérance, ou encore : à la valeur attendue)* ; son carré, $(X-E(X))^2$ est *l'écart quadratique*. La variance est donc la valeur moyenne, ou attendue, de l'écart quadratique. On considère l'écart quadratique plutôt que l'écart algébrique car on ne veut pas se soucier du signe — en fait, la valeur moyenne de l'écart algébrique est nulle ! Pour interpréter la valeur de la variance, on est conduit à prendre sa racine carré, l'écart-type. Notez que si X est une mesure en mètres, alors var(X) s'exprime en mètres carrés, et l'écart-type $\sqrt{\text{var}(X)}$ est en mètres.

Il serait sans doute également intéressant de considérer la valeur moyenne de l'écart absolu :

$$E\left(\left|X-E(X)\right|\right)$$

Pour justifier le choix de l'écart quadratique, nous nous contenterons de mentionner que la variance est beaucoup plus facile à calculer, grâce en particulier à la « formule de décentrement » suivante :

Proposition 4 *On a*

$$\boxed{\text{var}(X) = E\left(X^2\right) - E(X)^2.}$$

Prouvons ce résultat. On note $\mu = E(X)$; on a :

$$\begin{aligned} E\left((X-\mu)^2\right) &= E\left(X^2 - 2\mu X + \mu^2\right) \\ &= E\left(X^2\right) - 2\mu E(X) + \mu^2 \\ &= E\left(X^2\right) - 2\mu \times \mu + \mu^2 \\ &= E\left(X^2\right) - \mu^2 \\ &= E\left(X^2\right) - E(X)^2. \end{aligned}$$

On a utilisé la linéarité de l'espérance dans ce calcul.

Proposition 5 *Si* X *est une variable aléatoire et* a *et* b *sont des constantes, on a*

$$\boxed{\text{var}(aX+b) = a^2\,\text{var}(X).}$$

Exemple 11 Poursuivons les exemples 4 et 6. On a calculé $E(X) = \frac{7}{2}$. L'espérance de son carré est $E(X^2) = \sum_{x=1}^{6} x^2 \times \frac{1}{6} = \frac{91}{6}$. Sa variance est donc $\text{var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$. □

Exemple 12 Poursuivons l'exemple 8 : la loi uniforme sur $[a,b]$. On a calculé $E(X) = \frac{a+b}{2}$. Calculons l'espérance de X^2. On a

$$\begin{aligned} E\left(X^2\right) &= \int_a^b \frac{x^2}{b-a}\,dx \\ &= \left[\frac{x^3}{3(b-a)}\right]_a^b \\ &= \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}. \end{aligned}$$

Sa variance est donc

$$\begin{aligned} \text{var}(X) &= E\left(X^2\right) - E(X)^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

□

3.6 Fonction de répartition et quantiles

3.6.1 Fonction de répartition

Définition 8 *La* fonction de répartition *d'une variable aléatoire* X *est*

$$\boxed{\mathrm{F}(t) = \mathbb{P}(\mathrm{X} \leqslant t).}$$

Si X est une variable aléatoire continue à densité, la fonction F est croissante et continue ; elle est dérivable en tout point où f est continue et $\mathrm{F}' = f$. On a $\lim\limits_{t\to-\infty} \mathrm{F}(t) = 0$ et $\lim\limits_{t\to+\infty} \mathrm{F}(t) = 1$.

Il est utile de remarquer que $\boxed{\mathbb{P}(a < \mathrm{X} \leqslant b) = \mathrm{F}(b) - \mathrm{F}(a).}$

Exemple 13 Poursuivons les exemples 8 et 12. La fonction de répartition de la loi uniforme sur $[a,b]$ est

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

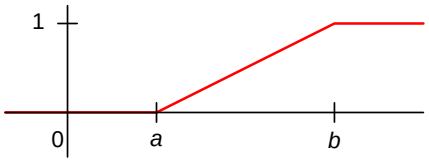


Figure 25. Fonction de répartition de $\mathscr{U}([a,b])$

□

3.6.2 Quantiles

Dans un souci de simplicité, nous ne donnons la définition des quantiles que dans le cas d'une variable continue à densité dont la fonction de répartition est strictement croissante et continue.

Définition 9 *Le quantile d'ordre* $\alpha \in]0,1[$ *d'une variable continue à densité* X *est le réel* x_α *tel que*

$$\boxed{\mathrm{F}(x_\alpha) = \mathbb{P}(\mathrm{X} \leqslant x_\alpha) = \alpha.}$$

Dans le cas général (variables discrètes ou même certaines variables continues à densité), il peut arriver qu'il n'existe pas de réel x tel que $\mathrm{F}(x) = \alpha$, ou qu'il en existe plusieurs ; on peut adapter la définition en prendant pour x_α la borne inférieure de $\{x : \mathrm{F}(x) \geqslant \alpha\}$. Nous ne nous préoccuperons pas davantage de ce cas de figure ici.

La médiane est le quantile d'ordre 0,5. Les premiers, deuxièmes et troisièmes quartiles sont les quantiles d'ordre 0,25, 0,5 et 0,75. Le n^{e} centile est le quantile d'ordre $n/100$.

Exemple 14 Poursuivons les exemples 8, 12 et 13. Le quantile d'ordre 0,5, c'est-à-dire la médiane, est $\frac{a+b}{2}$. Pour $\alpha \in]0,1[$ quelconque, le quantile d'ordre α est $q_\alpha = a + \alpha(b-a)$. □

3.7 Intervalle de pari

On utilise aussi la terminologie « intervalle de fluctuation ».

Définition 10 *On considère une variable aléatoire* X *de loi* $\mathscr{L}$. *Un intervalle de pari de niveau* $\gamma = 1-\alpha$ *(ou de risque* α*) pour* X *est un intervalle* $[a,b]$ *dans lequel* X *tombe avec une probabilité* γ *:*

$$\mathbb{P}(a \leqslant X \leqslant b) = \gamma.$$

On prendra typiquement comme valeur pour γ, $\gamma = 0{,}95$ *(ou* 95%).

Exemple 15 On considère la loi uniforme sur l'intervalle [0,1]. Tout intervalle $[a,b] \subseteq [0,1]$ tel que $b-a = 0{,}95$ est un intervalle de pari de niveau 0,95 :

$$\mathbb{P}(0{,}05 \leqslant X \leqslant 1) = \mathbb{P}(0 \leqslant X \leqslant 0{,}95) = \mathbb{P}(0{,}025 \leqslant X \leqslant 0{,}975) = 0{,}95.$$

□

De façon générale, si $0 \leqslant a < b \leqslant 1$, si on note q_a et q_b les quantiles de niveau a et b de la loi de X, on a

$$\boxed{\mathbb{P}(q_a \leqslant X \leqslant q_b) = b-a.}$$

Il est courant de construire un intervalle de pari au niveau $\gamma = 1-\alpha$ en excluant les valeurs extrêmes de la distribution, c'est-à-dire celles qui sont en-dessous du quantile $q_{\alpha/2}$ ou au-dessus du quantile $q_{1-\alpha/2}$:

$$\boxed{\mathbb{P}(q_{\alpha/2} \leqslant X \leqslant q_{1-\alpha/2}) = 1-\alpha.}$$

3.8 Exemples de lois discrètes : lois binomiale et hypergéometrique

Un type d'urne qui fournit un modèle très affectionné des probabilistes est l'urne bicolore : elle contient des boules noires et des boules rouges.

3.8.1 Tirage avec remise : loi binomiale

Au chapitre précédent, on s'est intéressé à des tirages *sans remise*, c'est-à-dire que toute boule retirée de l'urne est définitivement mise de côté. Dans le modèle du tirage *avec remise*, on tire une boule, on note sa couleur, et on la remet dans l'urne avant le tirage suivant. On peut ainsi réaliser autant de tirages que l'on souhaite.

On note p la proportion de boules rouges. On réalise n tirages et on note X le nombre de boules rouges tirées. On veut connaître la loi de X, c'est-à-dire la valeur de $\mathbb{P}(X = k)$ pour $k = 0,\dots,n$; c'est la *loi binomiale :*

Définition 11 *La loi binomiale de paramètres n et p est la loi discrète de distribution*

$$\mathbb{P}(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{si } k \in [0,\dots,n] \\ 0 & \text{sinon.} \end{cases}$$

On la note $\mathscr{Bin}(n,p)$.

On trouve ce résultat en considérant tout d'abord le nombre de façons d'avoir k boules rouges parmi les n boules tirées : c'est $\binom{n}{k}$; chacun de ces tirages contient k boules rouges (chacune arrive avec probabilité p) et $n-k$ boules noires (chacune avec proba $1-p$), donc chacun de ces tirages a une probabilité $p^k(1-p)^{n-k}$.

3.8.2 Tirage sans remise : loi hypergéométrique

On revient au tirage sans remise. On considère une urne contenant N boules dont M boules rouges et N − M boules noires. On tire au hasard (sans remise) n boules dans l'urne ; on note X le nombre de boules rouges tirées.

On suppose que $n \leqslant \mathrm{M}$ et $n \leqslant \mathrm{N} - \mathrm{M}$, de sorte que X peut prendre les valeurs 0 à n. Alors X suit une *loi hypergéométrique* :

Définition 12 *La loi hypergéométrique de paramètres* N, M *et* n *est la loi discrète de distribution*

$$\mathbb{P}(\mathrm{X} = k) = \frac{\binom{\mathrm{M}}{k}\binom{\mathrm{N}-\mathrm{M}}{n-k}}{\binom{\mathrm{N}}{n}}.$$

On la note $\mathscr{H}(\mathrm{N}, \mathrm{M}, n)$.

On trouve ce résultat en faisant le rapport du nombre de tirages avec k boules rouges au nombre de tirages possibles. Le nombre de tirage possibles est $\binom{\mathrm{N}}{n}$. Chaque tirage avec k boules rouges se fait en choisissant k boules parmi les M boules rouges, et $n-k$ boules parmi les N − M boules noires : soit en tout $\binom{\mathrm{M}}{k}\binom{\mathrm{N}-\mathrm{M}}{n-k}$ façons de choisir.

Quand M et N − M sont grands devant n, le fait de faire n tirages modifie peu la proportion des deux types dans la population. On peut donc, dans ce cas, approximer la loi hypergéométrique par une loi binomiale $\mathscr{B}in\left(n, p = \frac{\mathrm{M}}{\mathrm{N}}\right)$. Par contre, si n est du même ordre de grandeur que M, on ne peut pas faire l'approximation.

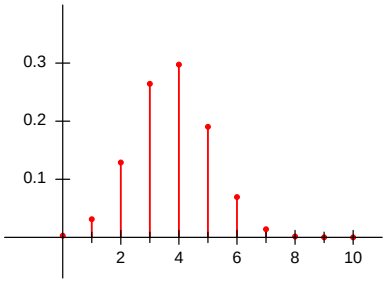


Figure 26. Fonction de masse de $\mathscr{H}(20, 12, 10)$

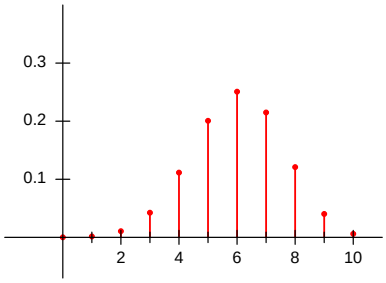


Figure 27. Fonction de masse de $\mathscr{B}in\left(n = 10, p = \frac{12}{20}\right)$

3.9 Exercices

Exercice 1 Soient X et Y des variables aléatoires discrètes, a et b des constantes. Démontrer que

$$\mathrm{E}(a\mathrm{X} + b) = a\mathrm{E}(\mathrm{X}) + b,$$

$$\mathrm{E}(\mathrm{X} + \mathrm{Y}) = \mathrm{E}(\mathrm{X}) + \mathrm{E}(\mathrm{Y}),$$

$$\mathrm{var}(a\mathrm{X} + b) = a^2\,\mathrm{var}(\mathrm{X})$$

Qu'en est-il de $\mathrm{var}(\mathrm{X} + \mathrm{Y})$?

Exercice 2 Soit X une variable aléatoire discrète définie par sa fonction de masse :

k	0	1	2
$\mathbb{P}(\mathrm{X} = k)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Calculer $\mathrm{E}(\mathrm{X})$, $\mathrm{E}\left(\mathrm{X}^2\right)$, $\mathrm{E}\left(2^{\mathrm{X}}\right)$, $\mathrm{var}(\mathrm{X})$.

Exercice 3 Soit X une variable aléatoire continue définie par sa densité :

$$f(x) = \begin{cases} 2x & \text{si } x \in [0,1] \\ 0 & \text{sinon} \end{cases}$$

1. Vérifier qu'il s'agit bien d'une fonction de densité.
2. Calculer E(X), E(X^2), E$\left(e^{X^2}\right)$, var(X).

Exercice 4 Soit X une variable aléatoire continue définie par sa densité :

$$f(x) = \begin{cases} 3x^2 & \text{si } x \in [0,1] \\ 0 & \text{sinon} \end{cases}$$

1. Vérifier qu'il s'agit bien d'une fonction de densité.
2. Calculer E(X), E(X^2), var(X).

Exercice 5 Déterminer la loi de la variable aléatoire correspondant à la somme de deux lancers d'un dé à six faces.

Exercice 6 Le paradoxe du Grand-Duc de Toscane

Le Grand-Duc de Toscane avait constaté que lorsqu'on lance trois dés à 6 faces, la somme 10 était obtenue plus souvent que la somme 9. Il y voyait un paradoxe, car il y a autant de façon d'obtenir 10 comme somme de trois dés, que d'obtenir 9 :

$$\begin{array}{rcccccccccccc} 10 & = & 6+3+1 & = & 6+2+2 & = & 5+4+1 & = & 5+3+2 & = & 4+4+2 & = & 4+3+3 \\ 9 & = & 6+2+1 & = & 5+3+1 & = & 5+2+2 & = & 4+4+1 & = & 4+3+2 & = & 3+3+3 \end{array}$$

Substituez vous à Galilée, et déterminez la loi de la variable aléatoire correspondant à la somme de trois lancers d'un dé à six faces (ce calcul aurait valu à Galilée de se voir offrir la chaire de mathématiques de l'Université de Pise). Pouvez-vous expliquer au Grand-Duc où est la faille de son raisonnement ?

Exercice 7 On jette un dé à 20 faces. On note X le résultat du tirage ; on a $X \in \{1,\ldots,20\}$.

On définit $X_1 \in \{0,1\}$ par

$$X_1 = \begin{cases} 0 & \text{si X pair} \\ 1 & \text{si X impair.} \end{cases}$$

On définit $X_2 \in \{0,1,2\}$ par

$$X_2 = \begin{cases} 0 & \text{si X divisible par 3} \\ 1 & \text{si X} - 1 \text{ divisible par 3} \\ 2 & \text{si X} - 2 \text{ divisible par 3,} \end{cases}$$

c'est-à-dire que $X_2 = 0$ quand on a tiré $X = 3,6,\ldots,18$,
$X_2 = 1$ quand on a tiré $X = 1,4,\ldots,19$,
et enfin : $X_2 = 2$ quand on a tiré $X = 2,5,\ldots,20$.

1. Quelles sont les valeurs des probabilités $\mathbb{P}(X_1 = 0)$ et $\mathbb{P}(X_1 = 1)$?
2. Quelles sont les valeurs des probabilités $\mathbb{P}(X_2 = 0)$, $\mathbb{P}(X_2 = 1)$ et $\mathbb{P}(X_2 = 2)$?
3. Calculer toutes les probabilités $P(X_1 = k \text{ et } X_2 = \ell)$ pour $k = 0,1$ et $\ell = 0,1,2$.
4. Les variables X_1 et X_2 sont-elles indépendantes ?

Exercice 8 En France le jeu du loto réalise trois fois par semaine un tirage sans remise de 5 balles dans une urne contenant 49 balles numérotées de 1 à 49.

1. Quel est le nombre de tirages possibles ?
2. Quel est le nombre de tirages contenant le numéro 1 ? Le numéro 2 mais pas le numéro 1 ?
3. On note X le plus petit numéro tiré. Quelle est la probabilité que X = 1 ? Que X = 2 ?
4. Donner la loi de X.

Chapitre 4

Variables aléatoires simultanées

Dans ce chapitre nous considérerons la situation où plusieurs mesures sont réalisées sur une même expérience aléatoire – on pourra penser à l'expérience aléatoire qui consiste à choisir un patient, dont on mesurera la stature, la masse, la cholestérolémie, etc.

4.1 Indépendance de variables aléatoires

Définition 13 *Des variables aléatoires* X,Y *sont indépendantes si pour tout* A,B $\subset \mathbb{R}$ *les événements* X $\in$ A *et* Y $\in$ B *sont indépendants.*

$$\mathbb{P}(\text{X} \in \text{A } \textit{et } \text{Y} \in \text{B}) = \mathbb{P}(\text{X} \in \text{A})\mathbb{P}(\text{Y} \in \text{B})$$

Autrement dit, les événements définis par les valeurs de X sont indépendants des événements définis par les valeurs de Y.

Cette définition s'étend naturellement au cas de n variables aléatoires.

4.2 Loi d'un couple de variables aléatoires

On considère (X,Y) un couple de variables aléatoires, c'est-à-dire une fonction d'un ensemble probabilisé Ω à valeur dans $\mathbb{R}^2$.

Par exemple, si Ω est l'ensemble de la population humaine, on peut prendre pour X(ω) la taille (arrondie au centimètre) d'un individu ω et pour Y(ω) sa masse (arrondie au kilo). Donner la loi de X d'une part et celle de Y d'autre part ne suffit pas à rendre compte de l'expérience. Par exemple, connaître $\mathbb{P}(\text{X} = 190)$ et $\mathbb{P}(\text{Y} = 82)$ ne suffit pas à calculer $\mathbb{P}(\text{X} = 190 \text{ et } \text{Y} = 82)$. Il faut donc donner la loi du couple (X,Y).

4.2.1 Description par les lois marginales et conditionnelles

Le moyen le plus simple de décrire la loi de (X,Y) est de donner d'une part la loi de X, d'autre part, pour chaque valeur x que X peut potentiellement prendre, donner la loi de Y conditionnellement à X = x ; c'est-à-dire la loi de Y dans les expériences aléatoires où X = x.

Bien sûr on pourrait procéder dans l'ordre inverse, et donner la loi de Y, puis pour toutes les valeurs y prises par Y, donner la loi de X conditionnellement à Y = y.

L'exemple suivant ne fait intervenir que des lois discrètes.

Exemple 16 Soit X la loi d'un dé à 6 faces : $\mathbb{P}(\text{X} = x) = \frac{1}{6}$ pour $x = 1, \ldots, 6$, et Y de valeurs possibles 0 et 1 et de loi donnée par $\mathbb{P}(\text{Y} = 0|\text{Xpair}) = \frac{1}{4}, \mathbb{P}(\text{Y} = 0|\text{Ximpair}) = \frac{1}{2}$.

Ceci permet calculer la probabilité de tout évènement, par exemple A = {(1,0), (2,1), (3,1)} :

$$\begin{aligned}\mathbb{P}((X,Y)\in A) &= \sum_{(x,y)\in A} \mathbb{P}(X=x)\mathbb{P}(Y=y|X=x)\\ &= \mathbb{P}(X=1)\mathbb{P}(Y=0|X=1)+\mathbb{P}(X=2)\mathbb{P}(Y=1|X=2)+\mathbb{P}(X=3)\mathbb{P}(Y=1|X=3)\\ &= \frac{1}{6}\times\frac{1}{2}+\frac{1}{6}\times\frac{3}{4}+\frac{1}{6}\times\frac{1}{2}\\ &= \frac{7}{24}\end{aligned}$$

□

Celui-ci ne fait intervenir que des lois à densité :

Exemple 17 Soit $X\sim\mathscr{U}([0,1])$ (sa densité $f(x)$ vaut 1 sur [0,1] et 0 en dehors) et Y de loi $\mathscr{U}([0,x])$ conditionnellement à $X=x$: $Y|X=x\sim\mathscr{U}([0,x])$ (sa densité conditionnelle $g(y|X=x)$ vaut $\frac{1}{x}$ sur $[0,x]$ et 0 en dehors).

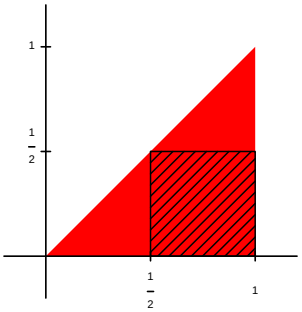


Figure 28. En rouge la région des valeurs possibles de X et Y. Le carré hachuré est A.

Ici aussi on peut calculer la probabilité d'un événement, par exemple si A est le carré $\left[\frac{1}{2},1\right]\times\left[0,\frac{1}{2}\right]$,

$$\begin{aligned}\mathbb{P}((X,Y)\in A) &= \int\int_{(x,y)\in A} f(x)g(y|X=x)\,\mathrm{d}y\,\mathrm{d}x\\ &= \int_{x=\frac{1}{2}}^{1} f(x)\int_{y=0}^{\frac{1}{2}} g(y|X=x)\,\mathrm{d}y\,\mathrm{d}x\\ &= \int_{x=\frac{1}{2}}^{1}\int_{y=0}^{\frac{1}{2}} \frac{1}{x}\,\mathrm{d}y\,\mathrm{d}x\\ &= \int_{x=\frac{1}{2}}^{1} \frac{1}{2}\times\frac{1}{x}\,\mathrm{d}x\\ &= \frac{1}{2}[\log x]_{\frac{1}{2}}^{1}\,\mathrm{d}x\\ &= -\frac{1}{2}\log\frac{1}{2} = \frac{1}{2}\log 2.\end{aligned}$$

□

Et ce troisième exemple fait intervenir les deux types de lois.

Exemple 18 Soit X qui suit une loi donnée par $\mathbb{P}(X=0)=\mathbb{P}(X=1)=\frac{1}{2}$, et Y décrit par deux lois conditionnelles $Y|X=0\sim\mathscr{U}([0,1])$ et $Y|X=1\sim\mathscr{U}([1,2])$.

On peut par exemple calculer

$$\begin{aligned}\mathbb{P}(X=0,Y\in[0;0{,}25]) &= \mathbb{P}(X=0)\mathbb{P}(Y\in[0;0{,}25]|X=0)\\ &= 0{,}5\times 0{,}25 = 0{,}125\end{aligned}$$

□

4.2.2 Description par la loi jointe

Le moyen le plus général de décrire la loi de (X,Y) est de donner sa loi jointe. On peut à cette fin utiliser la fonction de répartition conjointe :

Définition 14 *La* fonction de répartition conjointe *de* (X,Y) *est*

$$\boxed{F_{X,Y}(x,y) = \mathbb{P}(X \leqslant x \text{ et } Y \leqslant y).}$$

Loi discrète

Quand la paire (X,Y) est discrète, la fonction de masse conjointe est l'analogue de la fonction de masse.

Si on la connaît, on connaît la loi du couple (X,Y).

Définition 15 *Dans le cas discret, la* fonction de masse conjointe *de* (X,Y) *est*

$$\mathbb{P}(X = x \text{ et } Y = y).$$

On a la relation suivante entre fonction de répartition et fonction de masse :

$$\boxed{\mathbb{P}(X \leqslant x, Y \leqslant y) = \sum_{u \leqslant x} \sum_{v \leqslant y} \mathbb{P}(X = u, Y = v)}$$

Exemple 19 On reprend l'exemple où $\mathbb{P}(X = x) = \frac{1}{6}$ pour $x = 1, \ldots, 6$, et Y de valeurs possibles 0 et 1 avec $\mathbb{P}(Y = 0|\text{Xpair}) = \frac{1}{4}, \mathbb{P}(Y = 0|\text{Ximpair}) = \frac{1}{2}$.
Toutes les valeurs de $\mathbb{P}(X = x, Y = y)$ se calculent avec la formule $\mathbb{P}(X = x)\mathbb{P}(Y = y|X = x)$. □

Loi à densité

Définition 16 *Dans le cas continu à densité, la* densité conjointe *est une fonction* $\phi(x,y)$ *telle que*

$$\boxed{F_{X,Y}(x,y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} \phi(u,v)\,\mathrm{d}v\,\mathrm{d}u.}$$

De façon plus générale, pour $A \subset \mathbb{R}^2$,

$$\boxed{\mathbb{P}((X,Y) \in A) = \iint_{(u,v)\in A} \phi(u,v)\,\mathrm{d}v\,\mathrm{d}u.}$$

La densité conjointe est la dérivée de la fonction de répartition par rapport aux deux variables :

$$\boxed{\phi(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)}$$

Exemple 20 Considérons à nouveau $X \sim \mathcal{U}([0,1])$ (sa densité $f(x)$ vaut 1 sur [0,1] et 0 en dehors) et Y de loi $\mathcal{U}([0,x])$ conditionnellement à $X = x$: $Y|X = x \sim \mathcal{U}([0,x])$ (sa densité conditionnelle $g(y|X = x)$ vaut $\frac{1}{x}$ sur $[0,x]$ et 0 en dehors). La densité conjointe de (X,Y) est

$$\phi(x,y) = f(x)g(y|X = x) = \frac{1}{x}$$

sur le triangle rouge de la figure 28, et 0 en-dehors. □

Difficultés techniques dans le cas « mixte »

Dans le cas où une des variables est discrète et l'autre continue, il n'y a pas de fonction de masse jointe ni de densité jointe (pas au sens où on l'a défini ci-dessus). C'est le cas de l'exemple 18. Il y a des solutions plus ou moins simples à mettre en œuvre, mais le plus simple paraît d'en rester à la description par les lois marginales et conditionnelles, qui doit permettre en pratique de répondre à toutes les questions utiles.

L'exemple 18 peut paraître artificiel ; mais cette situation est souvent rencontrée en pratique, par exemple dans le cas où on s'intéresse au sexe et à la stature d'un patient ; on prend $X = 0$ pour un homme et $X = 1$ pour une femme, et Y sa stature. On pourra donner la loi de X (elle dépend de la proportion d'hommes et de femmes dans la population échantillonée), et la loi de Y conditionnellement à X (on utilisera probablement une loi normale de paramètres distincts pour chacun des deux sexes).

4.3 Retrouver les lois marginales et conditionnelles

Nous avons vu dans les exemples comment retrouver la fonction de masse conjointe ou la densité de masse conjointe à partir des lois marginales et conditionnelles. Voici comment on peut réaliser l'opération inverse.

4.3.1 Lois marginales

La variable X est une variable aléatoire dont la loi peut se déduire de la loi de (X,Y) : dans le cas discret, sa fonction de masse est

$$\boxed{\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x \text{ et } Y = y)}$$

et dans le cas continu, sa densité est

$$\boxed{f(x) = \int_y \phi(x,y)\,\mathrm{d}y.}$$

4.3.2 Lois conditionnelles

Dans le cas discret, on obtient la fonction de masse conditionnelle de Y par

$$\boxed{\mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x \text{ et } Y = y)}{\mathbb{P}(X = x)}}$$

Dans le cas continu, on obtient la densité conditionnelle de Y par

$$\boxed{g(y|X = x) = \frac{\phi(x,y)}{f(x)}.}$$

4.4 Covariance, corrélation

4.4.1 Covariance

Définition 17 *La* covariance *de* X *et* Y *est*

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}\big((X-\mathrm{E}(X))\cdot(Y-\mathrm{E}(Y))\big)\\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y).\end{aligned}$$

On a les règles de calcul suivantes :

$$\begin{aligned}\operatorname{cov}(X,X) &= \operatorname{var}(X)\\ \operatorname{cov}(X,Y) &= \operatorname{cov}(Y,X)\\ \operatorname{cov}(aX+b,Y) &= a\operatorname{cov}(X,Y)\\ \operatorname{cov}(X_1+X_2,Y) &= \operatorname{cov}(X_1,Y)+\operatorname{cov}(X_2,Y).\end{aligned}$$

On a en outre la propriété suivante :

Proposition 6 *La variance de la somme* X + Y *est*

$$\operatorname{var}(X+Y) = \operatorname{var}(X)+2\operatorname{cov}(X,Y)+\operatorname{var}(Y).$$

4.4.2 Corrélation linéaire

Définition 18 *Le* coefficient de corrélation linéaire *ou plus simplement* coefficient de corrélation *de* X *et* Y *est*

$$r_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

On a toujours $r_{X,Y} \in [-1,1]$. On a les règles de calcul suivantes :

$$\begin{aligned}\operatorname{corr}(X,X) &= 1\\ \operatorname{corr}(X,Y) &= \operatorname{corr}(Y,X)\\ \operatorname{corr}(aX+b,cY+d) &= \operatorname{corr}(X,Y) \quad \text{si } a,c \text{ ont même signe}\\ &= -\operatorname{corr}(X,Y) \text{ sinon}\end{aligned}$$

Cette dernière propriété montre que quand X et Y sont des mesures ayant une unité (longueur, température...) le coefficient de corrélation est invariant par les changements d'unités classiques (par exemple la conversion de mètres en millimètres ou en yards, de degrés Celsius en Fahrenheit, etc).

4.5 Variables aléatoires indépendantes

4.5.1 Distribution

Dans ce cas, dans le cas discret, on a

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

et dans le cas continu

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

4.5.2 Covariance de deux variables aléatoires indépendantes

Si X et Y sont indépendantes, on montre facilement que $E(XY) = E(X)E(Y)$; on a alors $\text{cov}(X,Y) = 0$; la réciproque n'est pas vraie. On en tire la propriété suivante :

Proposition 7 *Si* X *et* Y *sont indépendantes, la variance de la somme* $X+Y$ *est*

$$\boxed{\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).}$$

Cette propriété se généralise au cas n variables aléatoires indépendantes.

4.5.3 Variables aléatoires indépendantes de même loi

Il est classique en statistiques est de considérer des variables aléatoires $X_1, X_2, \ldots, X_n, \ldots$ indépendantes et de même loi. Cela correspond à des mesures répétées sur des expériences aléatoires indépendantes mais « identiques ».
Dans ce cas, si on note σ^2 la variance des X_i, le résultat précédent se généralise facilement en

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2.$$

On a d'autre part, par la linéarité de l'espérance, $E(X_1 + \cdots + X_n) = n\mu$ où μ est l'espérance de X_i.
On note

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

la moyenne empirique des X_i (c'est bien une variable aléatoire ! si on refait n expériences et qu'on calcule la moyenne empirique des n mesures, on trouvera généralement une autre valeur). On déduit de ce qui précède l'espérance et la variance de $\overline{X}_n$:

$$\boxed{E\left(\overline{X}_n\right) = \mu \text{ et } \text{var}\left(\overline{X}_n\right) = \frac{1}{n}\sigma^2.}$$

On voit que plus n est grand, moins $\overline{X}_n$ tend à être dispersé autour de μ.

4.5.4 La loi forte des grands nombres

Cette loi justifie l'idée que l'espérance est bien, comme l'intuition le veut, la moyenne obtenue quand on répète une expérience à l'infini.

Théorème 8 (Loi forte des grands nombres) *Soient* $X_1, X_2, \ldots$ *des variables aléatoires indépendantes et de même loi, admettant une espérance* μ. *On note* $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ *(la moyenne empirique des* X_i*). Alors, avec probabilité égale à 1,* $\overline{X}_n$ *tend vers* μ *quand* n *tend vers* ∞.

Notule mathématisante On peut donner un sens rigoureux à la phrase « $\overline{X}_n$ tend vers μ quand n tend vers ∞ ». Il faut pour cela considérer un ensemble probabilisé Ω assez grand pour que toutes les variables aléatoires indépendantes X_i soient définies en même temps sur Ω ; cela implique que chaque élément $\omega \in \Omega$ est en fait constitué d'une infinité d'expériences aléatoires indépendantes, chaque X_i prenant une mesure sur une de ces expériences. Alors le résultat de convergence ci-dessus peut-être exprimé ainsi : l'événement A défini par

$$A = \left\{\omega \in \Omega : \overline{X}_n \underset{n\to\infty}{\longrightarrow} \mu\right\}$$

est de probabilité $\mathbb{P}(A) = 1$. Voyez également le paragraphe 2.4.

4.5.5 La loi faible des grands nombres

Le théorème suivant est un peu moins fort que la loi forte des grands nombres, mais il est plus facile à montrer.

Théorème 9 (Loi faible des grands nombres) *Soient* $X_1, X_2, \ldots$ *des variables aléatoires indépendantes et de même loi, admettant une espérance* μ *et une variance* σ^2 *finie. On fixe un petit réel* $\varepsilon > 0$*. Alors la probabilité que* $\overline{X}_n$ *soit entre* $\mu - \varepsilon$ *et* $\mu + \varepsilon$ *tend vers* 1 *quand* n *tend vers* ∞ *:*

$$\mathbb{P}\left(\mu - \varepsilon < \overline{X}_n < \mu + \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 1.$$

La preuve de ce théorème reste plus compliquée que ce qui est demandé par ailleurs dans ce cours, mais nous l'incluons pour satisfaire la curiosité de certains d'entre vous.

Ce résultat se prouve en utilisant l'inégalité de Tchebychev :

Théorème 10 (Inégalité de Tchebychev) *Soit* X *une variable aléatoire admettant une espérance* μ *et une variance* σ^2 *finie. Soit* $c > 0$*. Alors*

$$\mathbb{P}(\mu - c\sigma < X < \mu + c\sigma) \geqslant 1 - \frac{1}{c^2}.$$

Preuve de l'inégalité de Tchebychev On pose $Y = \frac{X-\mu}{c\sigma}$, de sorte que $\mu - c\sigma < X < \mu + c\sigma$ est équivalent à $-1 < Y < 1$, c'est-à-dire $|Y| < 1$.

Notons que $E(Y) = 0$ et $E\left(Y^2\right) = \text{var}(Y) = \frac{1}{c^2}$. Si U est la variable aléatoire qui vaut 1 si $|Y| \geqslant 1$ et 0 sinon, on a

$$\mathbb{P}(|Y| \geqslant 1) = E(U).$$

D'autre part $0 \leqslant U \leqslant Y^2$, d'où $E(U) \leqslant E(Y^2) = \frac{1}{c^2}$, et donc

$$\mathbb{P}(|Y| < 1) = 1 - \mathbb{P}(|Y| \geqslant 1) = 1 - E(U) \geqslant 1 - \frac{1}{c^2}.$$

Preuve de la loi faible des grands nombres On applique l'inégalité de Tchebychev à $\overline{X}_n$ qui a pour espérance μ et pour écart-type $\sigma/\sqrt{n}$. En prenant $c = \frac{\varepsilon}{\sigma/\sqrt{n}}$, on obtient

$$\mathbb{P}\left(\mu - \varepsilon < \overline{X}_n < \mu + \varepsilon\right) \geqslant 1 - \frac{1}{c^2} = 1 - \frac{\sigma^2/n}{\varepsilon^2},$$

qui tend vers 1 quand n tend vers $+\infty$.

4.5.6 Somme de variables aléatoires indépendantes

Proposition 11 *Soient* X *et* Y *deux variables aléatoires discrètes indépendantes, de fonctions de masse* p_1 *et* p_2*. On pose* $Z = X + Y$*. On a*

$$\begin{aligned}\mathbb{P}(Z = z) &= \sum_{x,y\,:\,z=x+y} \mathbb{P}(X = x \text{ et } Y = y)\\ \mathbb{P}(Z = z) &= \sum_x \mathbb{P}(X = x \text{ et } Y = z - x)\\ &= \sum_x \mathbb{P}(X = x) \cdot \mathbb{P}(Y = z - x)\end{aligned}$$

Ce calcul est appelé « produit de convolution » des lois de X *et de* Y*.*

x	0	1	2
$\mathbb{P}(X=k)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\mathbb{P}(Y=k)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

Table 11. Lois de X et Y

Exemple 21 Considérons X et Y deux variables indépendantes de lois récapitulées ci-dessous. Soit $Z = X + Y$. Les valeurs que Z peut prendre sont les entiers de 0 à 4. On a :

$$\begin{aligned}
\mathbb{P}(Z=0) &= \mathbb{P}(X=0 \text{ et } Y=0)\\
&= \mathbb{P}(X=0)\mathbb{P}(Y=0)\\
&= \tfrac{1}{6}\\
\mathbb{P}(Z=1) &= \mathbb{P}(X=0 \text{ et } Y=1 \text{ ou } X=1 \text{ et } Y=0)\\
&= \mathbb{P}(X=0)\mathbb{P}(Y=1)+\mathbb{P}(X=1)\mathbb{P}(Y=0)\\
&= \tfrac{5}{18}\\
\mathbb{P}(Z=2) &= \mathbb{P}(X=0)\mathbb{P}(Y=2)+\mathbb{P}(X=1)\mathbb{P}(Y=1)+\mathbb{P}(X=0)\mathbb{P}(Y=2)\\
&= \tfrac{1}{3}\\
\mathbb{P}(Z=3) &= \mathbb{P}(X=1)\mathbb{P}(Y=2)+\mathbb{P}(X=2)\mathbb{P}(Y=1)\\
&= \tfrac{1}{6}\\
\mathbb{P}(Z=4) &= \mathbb{P}(X=2)\mathbb{P}(Y=2)\\
&= \tfrac{1}{18}.
\end{aligned}$$

On vérifie qu'on a bien $\sum_z \mathbb{P}(Z=z) = 1$. □

Dans le cas où X et Y sont des variables aléatoires continues indépendantes de densité f et g, la densité de $Z = X + Y$ est le produit de convolution $h = f * g$ défini par

$$\boxed{h(z) = \int_{-\infty}^{+\infty} f(t)\cdot g(z-t)\,\mathrm{d}t.}$$

4.5.7 Produit et quotient de variables aléatoires indépendantes

On ne considère ici que le cas de variables continues à densité.

Proposition 12 *Soient* X *et* Y *des variables aléatoires continues indépendantes de densité f et g. Leur produit* XY *est une variable aléatoire de densité*

$$\boxed{h(x) = \int_{-\infty}^{+\infty} f\left(\frac{x}{t}\right)\cdot g(t)\cdot\frac{1}{|t|}\,\mathrm{d}t.}$$

Leur quotient X/Y *est une variable aléatoire de densité*

$$\boxed{h(x) = \int_{-\infty}^{+\infty} f(xt)\cdot g(t)\cdot|t|\,\mathrm{d}t.}$$

Exemple 22 Produit de deux variables uniformes Soient U et V deux variables indépendantes de loi $\mathscr{U}([0,1])$. Notons $f(t)$ leur densité commune.

Leur produit UV suit une loi de densité $h(x)$. On a clairement $h(x)=0$ quand $x \notin [0,1]$. Pour $x \in [0,1]$ on a

$$\begin{aligned} h(x) &= \int_{-\infty}^{+\infty} f\left(\frac{x}{t}\right) \cdot f(t) \cdot \frac{1}{|t|}\,\mathrm{d}t \\ &= \int_0^1 f\left(\frac{x}{t}\right) \frac{1}{t}\,\mathrm{d}t \\ &= \int_x^1 \frac{1}{t}\,\mathrm{d}t \\ &= -\log(x). \end{aligned}$$

On peut vérifier que son espérance est 1/4. Le graphe de la densité est représenté figure 29. □

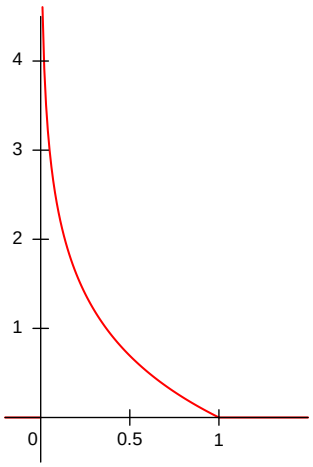


Figure 29. Densité de UV

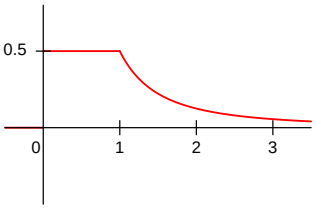


Figure 30. Densité de U/V

Exemple 23 Quotient de deux variables uniformes Soient U et V comme ci-dessus.

Leur quotient U/V suit une loi de densité $h(x)$. On a $h(x)=0$ quand $x<0$. Pour $x \geqslant 0$ on a

$$\begin{aligned} h(x) &= \int_{-\infty}^{+\infty} f(xt) \cdot f(t)|t|\,\mathrm{d}t \\ &= \int_0^1 f(xt)\,t\,\mathrm{d}t \\ &= \begin{cases} \displaystyle\int_0^1 t\,\mathrm{d}t = \frac{1}{2} & \text{si } x \leqslant 1 \\ \displaystyle\int_0^{\frac{1}{x}} t\,\mathrm{d}t = \frac{1}{2x^2} & \text{si } x \geqslant 1 \end{cases} \end{aligned}$$

Son espérance est infinie. Le graphe de la densité est représenté figure 30. □

4.6 Exercices

Exercice 1 Soit un couple de variables aléatoires (X,Y) dont la loi est donnée par le tableau ci-dessous.

$y=$	1	2
$x=1$	$\frac{2}{18}$	$\frac{4}{18}$
3	$\frac{1}{18}$	$\frac{2}{18}$
5	$\frac{3}{18}$	$\frac{6}{18}$

Valeurs de $\mathbb{P}(X=x,Y=y)$.

1. Déterminer la loi de X et celle de Y ; calculer l'espérance et la variance de chacune de ces variables.
2. Calculer cov(X,Y).
3. Les variables X et Y sont-elles indépendantes ?

Exercice 2 On considère deux variables aléatoires X et Y de loi décrite par

x	1	2	3
$\mathbb{P}(X=x)$	0,40	0,30	0,30

et

$y=$	1	2	3
$\mathbb{P}(Y=y\|X=1)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
$\mathbb{P}(Y=y\|X=2)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
$\mathbb{P}(Y=y\|X=3)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

1. Donner la loi jointe de (X,Y), en donnant pour tous les (x,y) possibles la valeur de $\mathbb{P}(X=x,Y=y)$.
2. Donner la loi de X conditionnellement à $Y=1$.
3. Calculer E(X), var(X), E(Y), var(Y).
4. Calculer le coefficient de corrélation cor(X,Y).

Exercice 3 Soit un couple de variables aléatoires (X,Y) dont la densité est

$$f(x,y)=\begin{cases} 6xy^2 & \text{si } x,y\in[0,1] \\ 0 & \text{sinon} \end{cases}$$

1. Vérifier que f est bien une densité.
2. Déterminer la densité de X et celle de Y ; calculer l'espérance de chacune de ces variables. Calculer cov(X,Y).
4. Calculer la densité de X conditionnellement à $Y=y$. Les variables X et Y sont-elles indépendantes ?

Exercice 4 Soit un couple de variables aléatoires (X,Y) dont la densité est

$$f(x,y)=\begin{cases} x+y & \text{si } x,y\in[0,1] \\ 0 & \text{sinon} \end{cases}$$

1. Vérifier que f est bien une densité.
2. Déterminer la densité de X et celle de Y ; calculer l'espérance et la variance de chacune de ces variables. Calculer cov(X,Y).
3. Calculer la densité de X conditionnellement à $Y=y$. Les variables X et Y sont-elles indépendantes ?

Exercice 5 Soient X et Y des variables aléatoires indépendantes de densité f et g et de fonction de répartition F et G. Soient U = min(X,Y) et V = max(X,Y).

1. Montrer que $(U \leqslant t) = (X \leq t \text{ ou } Y \leq t)$, et $(V \leqslant t) = (X \leq t \text{ et } Y \leq t)$.

2. Exprimer les densités de U et V en fonction de f, F, g et G.

3. Application pour X et Y suivant une loi $\mathscr{U}([0,1])$.

Exercice 6 Déterminer la densité de v.a. Z = X+Y quand X et Y sont des v.a. indépendantes suivant : 1. une loi uniforme $\mathscr{U}([0,1])$; 2. une loi exponentielle $\mathscr{E}(\lambda)$.

Chapitre 5

Processus de Bernoulli et processus de Poisson

Nous introduisons ici plusieurs lois fréquemment rencontrées, en les plaçant dans le cadre où elles apparaissent naturellement : un processus stochastique en temps discret, le processus de Bernoulli, et un processus stochastique en temps continu, le processus de Poisson.

5.1 Processus de Bernoulli

Une *expérience* ou *épreuve de Bernoulli* est une expérience aléatoire ayant deux résultats possibles, le *succès* ou l'*échec*. On note p la probabilité de succès et $q = 1 - p$ la probabilité d'échec.

On définit une variable aléatoire X sur l'ensemble des expériences possibles par $X(\omega) = 1$ si l'expérience ω est un succès, et $X(\omega) = 0$ si c'est un échec. X suit la loi de Bernoulli de paramètre p.

5.1.1 Loi de Bernoulli

Définition 19 *La loi de Bernoulli de paramètre p est la loi discrète de distribution*

$$\mathbb{P}(X = k) = \begin{cases} p & si\ k = 1 \\ 1 - p & si\ k = 0 \\ 0 & sinon. \end{cases}$$

On la note $\mathscr{B}(p)$.

Lemme 13 *L'espérance et la variance d'une variable $X \sim \mathscr{B}(p)$ sont $E(X) = p$ et $\mathrm{var}(X) = p(1-p)$.*

Le calcul est laissé en exercice.

Le *processus de Bernoulli* consiste à renouveler une expérience de Bernoulli de loi $\mathscr{B}(p)$, un nombre (potentiellement) infini de fois. On suppose que les expériences successives sont indépendantes.

On peut formaliser ceci en considérant une suite de variables aléatoires indépendantes, $(X_1, X_2, \ldots)$, où chaque X_i suit une loi $\mathscr{B}(p)$.

Exemple 24 Le modèle des tirages successifs avec remise dans une urne bicolore (cf section 3.8.1) donne un exemple de processus de Bernoulli : on définit le succès comme « tirer une balle rouge », et p comme la proportion de balles rouges dans l'urne. □

Exemple 25 On enchaîne les parties de « pile ou face ». C'est un processus de Bernoulli avec $p = \frac{1}{2}$. □

Exemple 26 On joue un numéro simple à la roulette. La valeur de p est $1/37$; les expériences successives sont indépendantes. On s'intéresse à la probabilité de gagner 2 fois en 10 parties, de ne pas avoir gagné une fois en 100 parties, de gagner pour la première fois avant la 37^{e} partie, etc. □

Exemple 27 On considère des essais cliniques successifs, sur une grande série de patients. De rares patients, en proportion p, ont une réaction allergique importante à la molécule testée. On s'intéresse à la probabilité d'avoir observé 2 réactions importantes après 100 essais, etc. □

5.1.2 Loi binomiale

On s'intéresse au nombre de succès observés après n expériences de Bernoulli, c'est-à-dire à $X = X_1 + \cdots + X_n$. La loi de X est, nous l'avons vu en 3.8.1, la loi binomiale de paramètres n et p, dont nous redonnons la défintion.

Définition 20 *La loi binomiale de paramètres n et p est la loi discrète de distribution*

$$\mathbb{P}(X = k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{si } k \in [0,\dots,n] \\ 0 & \text{sinon.} \end{cases}$$

On la note $\mathscr{B}in(n,p)$.

Lemme 14 *L'espérance et la variance d'une variable $X \sim \mathscr{B}in(n,p)$ sont $E(X) = np$ et $\mathrm{var}(X) = np(1-p)$.*

Ce résultat découle de la linéarité de l'espérance, et de l'additivité de la variance pour des variables indépendantes (cf section 4.5.3).

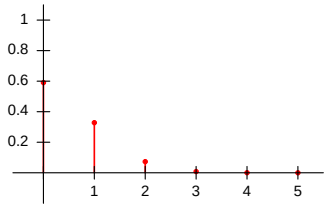


Figure 31. Fonction de masse de $\mathscr{B}in(5,0,1)$

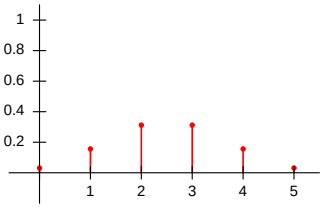


Figure 32. Fonction de masse de $\mathscr{B}in(5,0,5)$

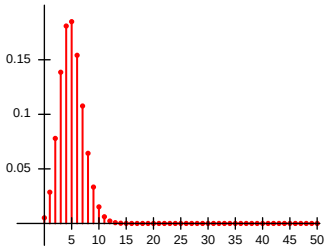


Figure 33. Fonction de masse de $\mathscr{B}in(50,0,1)$

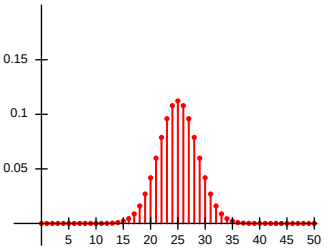


Figure 34. Fonction de masse de $\mathscr{B}in(50,0,5)$

5.1.3 Loi géométrique

On renouvelle une expérience de Bernoulli de loi $\mathscr{B}(p)$. On s'intéresse à X : le rang du premier succès.

Définition 21 *La loi géométrique de paramètre p est la loi discrète de distribution*

$$\mathbb{P}(X=k)=\begin{cases}(1-p)^{k-1}p & \text{si } k\geqslant 1\\ 0 & \text{sinon.}\end{cases}$$

Dans $\mathbb{P}(X=k)=(1-p)^{k-1}p$, le terme $(1-p)^{k-1}$ correspond à la probabilité d'enchaîner $k-1$ échecs, le terme p à la probabilité d'un succès à la k-ème expérience.

Lemme 15 *L'espérance et la variance d'une variable* X *de loi géométrique de paramètre p sont* $\mathrm{E}(X)=\frac{1}{p}$ *et* $\mathrm{var}(X)=\frac{1-p}{p^2}$. *Sa fonction de répartition est* $\mathrm{F}(k)=1-(1-p)^k$.

En effet on calcule

$$\begin{aligned}\mathrm{F}(k)&=\mathrm{P}(X\leqslant k)\\ &=\sum_{i=1}^{k}(1-p)^{i-1}p\\ &=\left(\frac{1-(1-p)^k}{p}\right)p\\ &=1-(1-p)^k.\end{aligned}$$

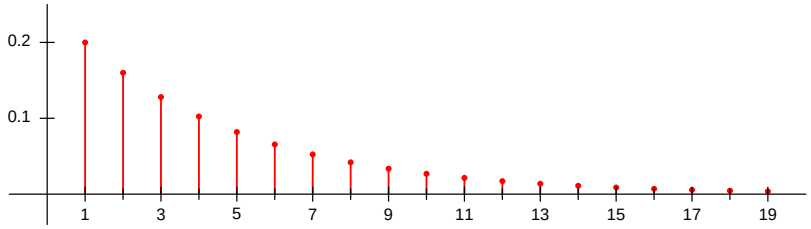


Figure 35. Fonction de masse de la loi géométrique de paramètre $p=0{,}2$

Propriété d'oubli

On a $\mathbb{P}(X>k)=1-\mathbb{P}(X\leqslant k)=(1-p)^k$ (c'est la probabilité de ne pas rencontrer de succès pendant les k premières expériences).
Ainsi, $\mathbb{P}(X>k+\ell)=\mathbb{P}(X>k)\mathbb{P}(X>\ell)$, ou encore :

$$\begin{aligned}\mathbb{P}(X>k+\ell|X>k) &= \frac{\mathbb{P}(X>k+\ell \text{ et } X>k)}{\mathbb{P}(X>k)}\\ &= \frac{\mathbb{P}(X>k+\ell)}{\mathbb{P}(X>k)}\\ &= \frac{(1-p)^{k+\ell}}{(1-p)^k}\\ &= (1-p)^{\ell}\\ &= \mathbb{P}(X>\ell)\end{aligned}$$

Cette probabilité conditionnelle peut s'interprèter ainsi : sachant qu'on a déjà fait k expériences sans rencontrer un succès, la probabilité de ne pas rencontrer de succès pendant les ℓ expériences qui suivent. Les expériences successives étant indépendantes, le fait d'avoir déjà attendu k expériences ne change pas la probabilité de devoir attendre ℓ expériences ou plus avant de rencontrer un succès ! Le processus est dit « sans mémoire ».

5.1.4 Loi binomiale négative

Attention ! Selon les ouvrages la définition de la loi binomiale négative connaît de multiples variantes (voir remarque finale).

Cette fois X est le rang du r-ème succès.

Définition 22 *La loi binomiale négative de paramètres r et p est la loi discrète de distribution*

$$\mathrm{P}(\mathrm{X}=k)=\begin{cases}\binom{k-1}{r-1}(1-p)^{k-r}p^r & \text{si } k\geqslant r\\ 0 & \text{sinon.}\end{cases}$$

Si $r=1$, on retrouve la loi géométrique. Dans le cas général X est la somme de r variables indépendantes suivant une loi géométrique de paramètres p. On en déduit son espérance et sa variance.

Lemme 16 *L'espérance et la variance d'une variable* X *de loi binomiale négative de paramètres r et p sont* $\mathrm{E}(\mathrm{X})=\frac{r}{p}$ *et* $\mathrm{var}(\mathrm{X})=\frac{r(1-p)}{p^2}$.

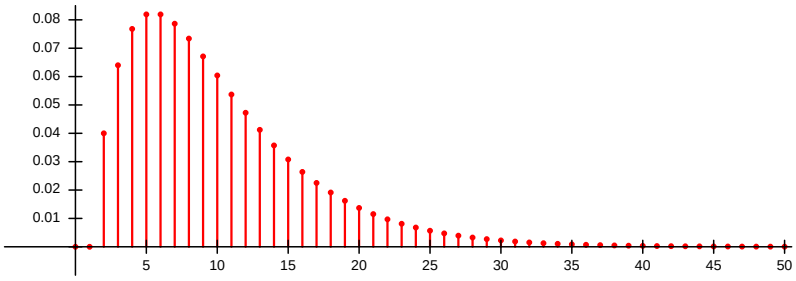


Figure 36. Fonction de masse de la loi binomiale négative de paramètres $p=0{,}2$, $r=2$

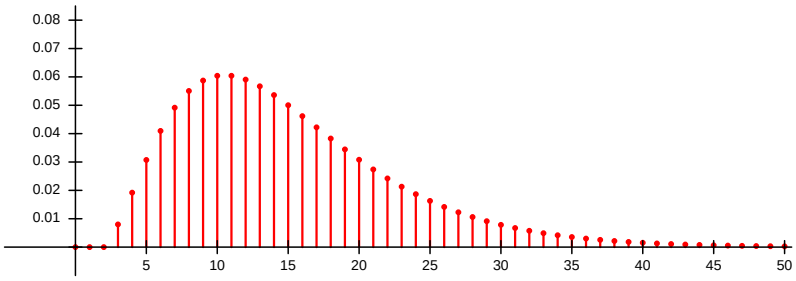


Figure 37. Fonction de masse de la loi binomiale négative de paramètres $p=0{,}2$, $r=3$

Remarque Beaucoup d'auteurs préfèrent définir la loi binomiale négative comme le nombre d'échecs enregistrés avant un succès, c'est-à-dire comme la loi d'une variable $\mathrm{Y}=\mathrm{X}-r$ où X suit la loi ci-dessus. L'avantage est que le support de la loi est alors $\mathbb{N}$ (le support est l'ensemble des k où $\mathbb{P}(\mathrm{X}=k)\neq 0$). Il existe encore d'autres variantes.

Exemple 28 Un test sérologique a une spécificité de 97%, c'est-à-dire qu'il est négatif chez 97% des sujets sains. Si 20 sujets sains sont testés chaque jour, quelle est la probabilité d'observer 1 test positif ou plus ? Et d'observer exactement un test positif ?

Quelle est l'espérance du nombre de sujets sains à tester avant d'observer le premier test positif ? La probabilité que le premier test positif survienne chez le septième sujet sain testé un jour donné ? Sachant que les 6 premiers sujets testés sont négatifs, quelle est la probabilité que le septième soit positif ?

Le nombre de tests positifs X suit une loi binomiale $\mathcal{B}in(n = 20, p = 0{,}03)$. On a $\mathbb{P}(X \geqslant 1) = 1 - P(X = 0) = 1 - (1-p)^n = 0{,}46$. On a $\mathbb{P}(X = 1) = np(1-p)^{n-1} = 0{,}34$.

Le nombre Y de tests réalisés avant le premier test positif suit une loi géométrique d'espérance $1/p = 33{,}3$. On a $\mathbb{P}(Y = 7) = (1-p)^6 p = 0{,}025$. Bien entendu, la probabilité que le septième sujet (sain) soit positif au test est 0,03, et elle ne dépend pas du résultat des tests précédents. □

5.2 Processus de Poisson

Le processus de Poisson est un processus de comptage en temps continu (le processus de Bernoulli est un processus en temps discret).

À nouveau on s'intéresse à la survenue d'événements indépendants, qui ont lieu avec une certaine probabilité ; mais à présent, au lieu d'observer des succès parmi une série d'expériences aléatoires, les événements arrivent à des instants $T_1 < T_2 < \cdots$ (qui sont des variables aléatoires).

Un exemple classique est l'observation de voitures au bord d'une route (une départementale plutôt que le boulevard périphérique parisien aux heures de pointe : on veut modéliser des événements *indépendants !)*. On va observer en moyenne 20 voitures par heure (par exemple) ; si on s'intéresse à l'intervalle de temps entre deux voitures, il sera en moyenne de 60/20 = 3 minutes.

Comment les observations (nombre de voitures, intervalle de temps entre deux voitures) vont-elles varier au fil des heures ?

Figure 38. Processus de Poisson : événements en temps continu

Plus formellement, on s'intéresse au nombre total d'occurrences d'un événement après qu'un temps total t se soit écoulé, sous les hypothèses suivantes :

— deux événements n'arrivent jamais simultanément ;
— la loi du nombre d'occurrences dans un intervalle de temps $[t_0, t_0 + \Delta_t]$ ne dépend que de Δ_t ; on dit que le processus n'a pas de mémoire : ce qui arrive entre t_0 et $t_0 + \Delta_t$ ne dépend pas de l'instant t_0 où on commence le comptage.

On appelle *taux du processus de Poisson* le nombre moyen λ_0 d'occurrences par unité de temps. Par exemple, $\lambda_0 = 20$ voitures par heure.

On peut alors montrer que le nombre d'occurrences pendant une durée Δ_t suit une *loi de Poisson* de paramètre $\lambda = \lambda_0 \times \Delta_t$, et que l'intervalle entre deux événements suit une *loi exponentielle* de paramètre $\lambda = \lambda_0$. Ces deux lois sont définies ci-dessous.

5.2.1 Loi de Poisson

Définition 23 *La loi de Poisson de paramètre* λ *est la loi discrète de distribution*

$$\mathbb{P}(X = k) = \begin{cases} \dfrac{\lambda^k}{k!} e^{-\lambda} & \text{si } k \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$$

On la note $\mathscr{P}(\lambda)$.

C'est la loi du nombre d'événements par unité de temps dans un processus de Poisson de taux λ. Avec $\lambda = \lambda_0 \times \Delta_t$, c'est la loi du nombre d'occurrences dans un processus de Poisson de taux λ_0 pendant une durée Δ_t. C'est bien une distribution de probabilité, car

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}.$$

Son espérance et sa variance sont égales :

Lemme 17 *L'espérance et la variance d'une variable aléatoire* $X \sim \mathscr{P}(\lambda)$ *sont* $E(X) = \text{var}(X) = \lambda$.

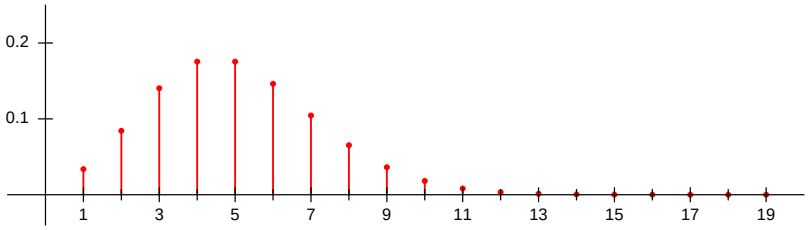


Figure 39. Fonction de masse de $\mathscr{P}(5)$

5.2.2 Loi exponentielle

Définition 24 *La loi exponentielle de paramètre* λ *est la loi continue de densité*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$$

On la note $\mathscr{E}(\lambda)$.

C'est la loi du temps qui s'écoule avant l'occurrence d'un événement dans un processus de Poisson de taux λ.

Lemme 18 *Soit* T *une variable aléatoire de loi* $\mathscr{E}(\lambda)$.
Son espérance est $E(T) = \frac{1}{\lambda}$ *; sa variance* $\text{var}(T) = \frac{1}{\lambda^2}$.
Sa fonction de répartition est

$$\begin{aligned} F(t) &= \mathbb{P}(T \leqslant t) \\ &= \begin{cases} 1 - e^{-\lambda t} & t \geqslant 0 \\ 0 & t < 0 \end{cases} \end{aligned}$$

Si on pose $T_1 = aT$, *alors* T_1 *suit une loi exponentielle* $T_1 \sim \mathscr{E}\left(\frac{\lambda}{a}\right)$.

La dernière propriété énoncée dans le lemme correspond à un changement d'unités de temps : par exemple si T est exprimé en heures, avec $\lambda = 20$ événements par heures, $E(T) = \frac{1}{\lambda} = 1/20$ est, en heures, le temps moyen entre deux événements ; prenons $a = 60$: $T_1 = aT$ est le temps entre deux événements exprimé en minutes — il suit toujours une loi exponentielle, d'espérance $E(T_1) = \frac{a}{\lambda} = \frac{60}{20} = 3$ minutes.

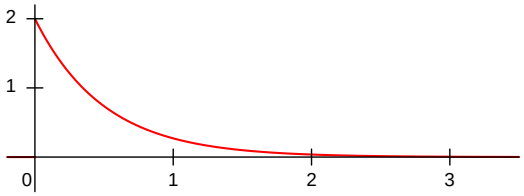


Figure 40. Densité de $\mathscr{E}(2)$

Propriété d'oubli

On vérifiera en exercice que si T suit une loi exponentielle,

$$\mathbb{P}(\mathrm{T} > t+s) = \mathbb{P}(\mathrm{T} > t)\mathbb{P}(\mathrm{T} > s),$$

et donc

$$\mathbb{P}(\mathrm{T} > t+s | \mathrm{T} > t) = \mathrm{P}(\mathrm{T} > s).$$

Cette propriété, analogue à celle déjà mentionnée pour la loi géométrique, est là encore liée au fait que le processus est sans mémoire. Ainsi, le temps T_1 écoulé entre deux occurrences, le temps T_2 écoulé un instant t_0 arbitraire et la première occurrence qui suit, etc, suivent tous une loi $\mathscr{E}(\lambda)$.

5.2.3 Loi Gamma

Le temps T avant la r^e observation est une somme $\mathrm{T} = \mathrm{T}_1 + \cdots + \mathrm{T}_r$ de r variables aléatoires indépendantes de loi $\mathscr{E}(\lambda)$. On montre que la loi de T est la loi Gamma de paramètres r et λ :

Définition 25 *La loi Gamma de paramètres r et λ est la loi continue de densité*

$$f(x) = \begin{cases} \dfrac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} & \text{si } x \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$$

On la note $\Gamma(r,\lambda)$.

De $\mathrm{T} = \mathrm{T}_1 + \cdots + \mathrm{T}_r$ on déduit que son espérance est $\dfrac{r}{\lambda}$ et sa variance $\dfrac{r}{\lambda^2}$ (les T_i étant indépendantes).

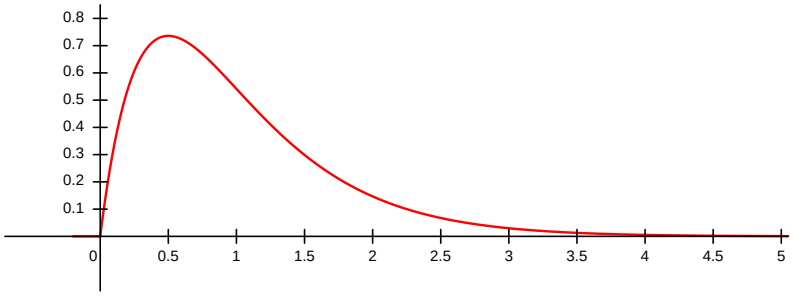


Figure 41. Densité de la loi Gamma de paramètres $r = 2$, $\lambda = 2$

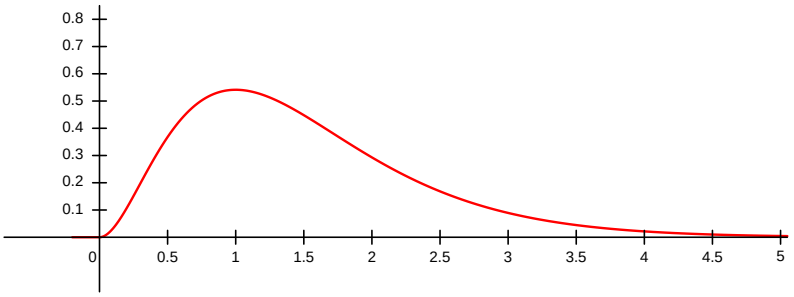


Figure 42. Densité de la loi Gamma de paramètres $r = 3$, $\lambda = 2$

Exemple 29 Le rétinoblastome est une tumeur occulaire qui se développe habituellement avant l'âge de 5 ans. Il a une prévalence d'environ 5 pour 100 000. Dans les pays industrialisés son pronostic est très bon, avec environ 90% de guérison, ce qui permet d'observer le nombre total de tumeurs chez un nombre non négligeable d'individus.

Il en existe une forme familiale à transmission dominante (environ 40%) des cas. Dans cette forme, on a très souvent des tumeurs multiples, et le nombre total de tumeurs constaté s'ajuste bien à une loi de Poisson de paramètre $\lambda \approx 3$.

Cette analyse statistique a conduit Alfred Knudson à suggérer le mécanisme suivant : la tumeur se développe dans des cellules ayant deux allèles mutés d'un gène « suppresseur de tumeur ».

Chez un individu ayant germinalement deux allèles sains, il faut deux mutations somatiques dans la même lignée cellulaire pour que ce cancer se produise.

Par contre, chez un individu hétérozygote pour la mutation (ayant une mutation germinale), une mutation somatique suffit. En supposant que la probabilité d'une mutation somatique est constante au cours du temps, l'apparition des tumeurs est un processus de Poisson, et leur nombre total doit bien suivre une loi de Poisson.

On a bien, dans ce cas, une transmission dominante. Ce modèle présente l'avantage d'expliquer les données observée.

Le modèle de Knudson est un des premiers modèles d'oncogénèse à avoir été proposé. Le gène *RB1* est le premier gène suppresseur de tumeur qui a été identifié. Voir entre autres :
Knudson AG (1971). Mutation and cancer : statistical study of retinoblastoma. PNAS
Knudson AG (1993). Antioncogenes and human cancer, PNAS. □

Question : si le modèle de Knudson est exact, quelle est la probabilité qu'un enfant porteur de la mutation germinale ne développe aucune tumeur ? C'est $\mathbb{P}(X = 0)$ où $X \sim \mathscr{P}(3)$, et donc $e^{-3} \equiv 0{,}05$.

5.2.4 Superposition de processus de Poisson

Reprenons l'exemple du passage des voitures sur une route de campagne. Il correspond à un processus de Poisson de paramètre λ_1. Un autre observateur s'intéresse aux camions, dont l'apparition suit un processus de Poisson de paramètre λ_2. Enfin, un troisième observateur s'intéresse au passage des tous les véhicules, voitures et camions confondus. Il s'agit également d'un processus de Poisson, de paramètre $\lambda = \lambda_1 + \lambda_2$.

Ce résultat doit sembler intuitif : il y a en moyenne λ_1 voitures par unité de temps, λ_2 camions, et donc $\lambda_1 + \lambda_2$ véhicules.

Plus généralement, la superposition de n processus de Poisson de paramètres $\lambda_1, \lambda_2, \ldots, \lambda_n$ est un processus de Poisson de paramètre $\lambda = \lambda_1 + \cdots + \lambda_n$.

Un exemple « plus médical » peut être donné par l'apparition des tumeurs, qui peuvent être séparées selon leur localisation, leur latéralité, leur type histologique, etc, ou comptées globalement.

Le résultat annoncé est résumé par le théorème suivant. Notez que si T_1 et T_2 sont les temps d'attente respectifs avant la première observation dans les processus 1 et 2, alors $T = \min(T_1, T_2)$ est

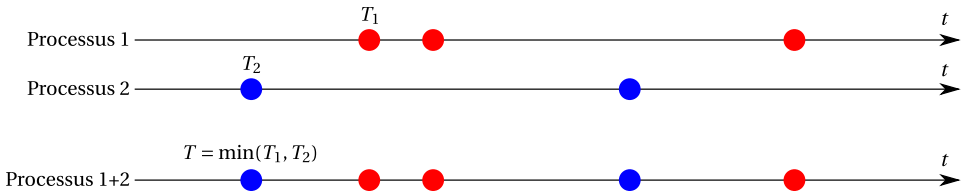


Figure 43. Superposition de processus de Poisson

le temps d'attente avant la première observation dans le processus obtenu par superposition des processus 1 et 2. Ceci se généralise à n processus.

Théorème 19 *Soient* $\lambda_1,\dots,\lambda_n \in \mathbb{R}^{>0}$ *et* $\lambda = \lambda_1 + \cdots + \lambda_n$. *a Soient* $X_1,\dots,X_n$ *des variables aléatoires indépendantes de lois respectives* $\mathscr{P}(\lambda_1),\dots,\mathscr{P}(\lambda_n)$. *Alors* $X = X_1 + \cdots + X_n$ *suit une loi* $\mathscr{P}(\lambda)$.

Soient $T_1,\dots,T_n$ *des variables aléatoires indépendantes de lois respectives* $\mathscr{E}(\lambda_1),\dots,\mathscr{E}(\lambda_n)$. *Alors* $T = \min(T_1,\dots,T_n)$ *suit une loi* $\mathscr{E}(\lambda)$.

5.2.5 Le processus de Poisson comme limite d'un processus de Bernoulli

Considérons un processus de Poisson de taux λ_0.

On peut diviser tout intervalle de temps de durée t en petits intervalles de très petite durée δ_t, assez petits pour que deux événements n'arrivent jamais durant le même petit intervalle (on pourra penser que δ_t est infiniment petit). L'intervalle de durée t est divisé en $n = t/\delta_t$ petits intervalles.

On considère le processus de Bernoulli qui consiste, à chaque petit intervalle de temps de durée δ_t, à regarder si un événement s'est produit ou non.

Notons p la probabilité qu'un événement arrive pendant la durée δ_t. La valeur de p va dépendre de δ_t : plus δ_t est petit, plus p doit être petit.

Exemple 30 Dans le cas des voitures, on pourra prendre $\delta_t = 1$ seconde ; les expériences successives consistent à vérifier, à chaque seconde, si une nouvelle voiture apparaît à l'horizon. Si $\lambda = 20$ voitures à l'heure, soit $\frac{20}{3600} = \frac{1}{180}$ voitures à la seconde (en espérance), la probabilité de voir apparaître une voiture dans un intervalle de temps d'une seconde est $p = \frac{1}{180}$. □

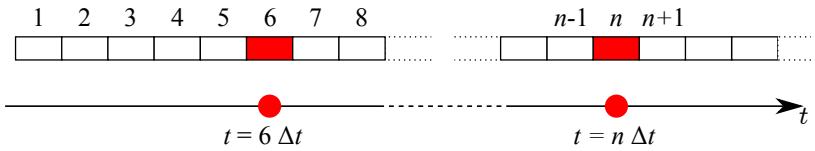


Figure 44. Le processus de Poisson comme limite d'un processus de Bernoulli

On considère un intervalle de temps de durée δ_t. Dans le processus de Poisson, le nombre d'événements attendus pendant cet intervalle suit une loi $\mathscr{P}(\lambda_0\delta_t)$, d'espérance $\lambda_0\delta_t$. Du point de vue du processus de Bernoulli, il suit une loi de Bernoulli $\mathscr{B}(p)$, d'espérance p. Pour que les espérances soient égales, il faut avoir

$$p = \lambda_0\delta_t.$$

On considère n expériences « instantanées » successives : cela correspond à une durée $n\delta_t$. Du point de vue du processus de Bernoulli, le nombre d'événements observés suit une loi binomiale $\mathscr{B}in(n,p)$; du point de vue Poisson, il suit une loi $\mathscr{P}(\lambda = \lambda_0 n\delta_t = np)$.

Ce raisonnement mène au principe suivant :

Proposition 20 *Si n est grand et p petit, (concrètement, disons $n > 50$ et $p < 0{,}1$), la loi binomiale $\mathscr{B}in(n,p)$ peut être approchée par la loi de Poisson $\mathscr{P}(\lambda = np)$.*

On peut également s'intéresser au nombre X d'instants écoulés avant le premier succès. Il suit une loi géométrique de paramètre p. D'autre part, dans le processus de Poisson, X correspond à une durée $T = X\delta_t$, qui suit une loi exponentielle $\mathscr{E}(\lambda_0)$. Les propriétés de la loi exponentielle impliquent que la loi de $\frac{1}{\delta_t}T$ est une loi $\mathscr{E}(\lambda_0\delta_t = p)$ (cf paragraphe 5.2.2).

Ceci mène au principe suivant :

Proposition 21 *Si p est petit, la loi géométrique de paramètre p peut être approchée par la loi exponentielle $\mathscr{E}(p)$.*

Ceci peut se vérifier directement sur les fonctions de répartition. On a vu que la fonction de répartition de la loi géométrique est, pour $k > 0$, $P(X \leqslant k) = 1 - (1-p)^k$. Pour p petit, on a $\log(1-p) \simeq -p$, d'où

$$\begin{aligned} P(X \leqslant k) &= 1 - (1-p)^k \\ &= 1 - e^{k\log(1-p)} \\ &\simeq 1 - e^{-kp} \end{aligned}$$

qui est bien la fonction de répartion de la loi exponentielle. On peut vérifier que plus p est petit, meilleure est l'approximation.

5.3 Exercices

Exercice 1 Soit $X \sim \mathscr{E}(\lambda)$. Pour $t, u \geqslant 0$, déterminer la probalité conditionnelle

$$\mathbb{P}(X \geqslant t + u | X \geqslant t).$$

Comment interpréter ce résultat ?

Exercice 2 Soient X et Y des variables aléatoires indépendantes, avec $X \sim \mathscr{E}(\lambda_1)$ et $Y \sim \mathscr{E}(\lambda_2)$. Soit $Z = \min(X,Y)$. Déterminer la loi de Z.

Exercice 3 Soit X une variable de loi $\mathscr{E}(\lambda)$ et $a > 0$. Déterminer la loi de $Y = aX$.

Exercice 4 Conditionnement sur la somme de deux lois de Poisson Soient X et Y des variables aléatoires indépendantes, avec $X \sim \mathscr{P}(\lambda_1)$ et $Y \sim \mathscr{P}(\lambda_2)$. Soit $Z = X + Y$. On sait que $Z \sim \mathscr{P}(\lambda = \lambda_1 + \lambda_2)$.

Montrez que pour $k \leqslant \ell$, $\mathbb{P}(X = k | Z = \ell) = \binom{\ell}{k} p^k (1-p)^{\ell-k}$ avec $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

Que rappelle cette probabilité ? Est-ce que ce résultat peut s'expliquer par un raisonnement basé sur l'intuition plutôt que le calcul ?

On considère le cas d'une forme héréditaire de cancer comme le rétinoblastome, qui touche indifféremment deux organes symmétriques (les deux yeux) selon un processus de Poisson. Parmi les patients ayant développé 3 tumeurs au total, quelle est la proportion attendue de patients ayant développé les 3 tumeurs au même œil ?

Exercice 5 Modèle de Knudson

1. Dans le modèle de Knudson, le nombre total de tumeurs chez un individu atteint d'une forme héréditaire de rétinoblastome suit une loi de Poisson de paramètre λ_K.

On s'intéresse au nombre de mutations qui apparaissent dans un seul œil. Montrer qu'il suit une loi de Poisson de paramètre $\lambda = \frac{\lambda_K}{2}$.

2. Considérant une population de 100 individus porteurs de la mutation prédisposante, calculez le nombre attendu d'individus développant à un œil 0, 1, … , 5 tumeurs ou plus, pour $\lambda_K = 2$ et $\lambda_K = 3$.

Dans chacun des deux cas, parmi les patients ayant développé au moins une tumeur, quelle est la proportion de ceux qui ont développé 1, … , 5 tumeurs ou plus ?

3. Comparez avec les données suivantes, extraites de l'article de Knudson (1971). Knudson est parvenu à compter les tumeurs dans un des deux yeux de 14 cas bilatéraux ; il y ajoute les tumeurs comptées par Stallard (1962).

tumeurs	patients	proportion
1	35	53%
2	17	26%
3	9	14%
4	4	6%
5	1	2%
Total	66	

5. Les porteurs de la mutation prédisposante qui n'ont pas développé de rétinoblastome ne sont pas comptabilisés, ce qui empêche d'estimer λ par $\overline{x}$, le nombre moyen de tumeur par œil chez les patients ayant développé la maladie. On admettra (cf exercice suivant) qu'on peut estimer λ par $\overline{x} - e^{-0,9(\overline{x}-1)}$; calculez cette valeur et réalisez un test de conformité.

Exercice 6 La seconde question du Chevalier de Méré

1. Le Chevalier de Méré et un de ses camarades de jeu décident de parier 64 pistoles sur un jeu en trois manches gagnantes. Chaque manche est un jeu équitable, c'est-à-dire qu'elle est gagnée ou perdue avec probabilité $\frac{1}{2}$. Le premier à gagner trois manches emporte la totalité de la somme.

Après trois manches, le chevalier a gagné deux manches et en a perdu une ; appelés par leurs affaires, les deux joueurs ne peuvent achever la partie, et doivent se séparer. Comment partager équitablement les 64 pistoles ? Ou, en vocabulaire moderne, quelle est l'espérance de gain de chacun des joueurs ?

2. Même question pour un jeu en N manches, après que le premier joueur ait gagné k manches et le second ℓ manches, $k, \ell < \mathrm{N}$.

Exercice 7 À l'hôpital de Morzy on reçoit en moyenne un cas de bilharziose par semaine. L'arrivée de ces patients est supposée suivre un processus de Poisson.

1. Soit X la variable aléatoire définie par X = nombre de cas reçus en une semaine donnée. Quelle est la loi de X ?

2. Calculer les probabilités suivantes : $\mathbb{P}(\mathrm{X}=0)$, $\mathbb{P}(\mathrm{X}=1)$, $\mathbb{P}(\mathrm{X}=2)$ et $\mathbb{P}(\mathrm{X}\geq 3)$.

3. Soit Y la variable aléatoire définie par Y = nombre de cas reçus en une journée. Quelle est la loi de Y ? Calculer $\mathbb{P}(\mathrm{Y}=0)$ et $\mathrm{P}(\mathrm{Y}=1)$.

4. Soit T le temps qui s'écoule entre l'arrivée de deux patients. Quelle est la loi de T ? (Préciser l'unité de temps choisie)

5. Quelle est la probabilité que T > 1 semaine ? Et que T > 2 semaines ?

Exercice 8 On a relevé les intervalles entre les tremblements de terre de magnitude $\geqslant 5$ en Californie de 1812 à 2014 (on élimine les répliques ayant eu lieu moins de deux semaines après un premier tremblement de terre). On obtient un total de 64 intervalles de temps, mesurés en mois. Ils sont tabulés ci-dessous, par ordre croissant de durée.

1	3	9	15	21	30	42	77
2	4	10	16	24	30	44	86
2	4	11	17	25	32	47	88
2	5	11	18	25	34	48	106
2	5	11	18	26	37	56	134
2	6	12	19	27	37	58	222
3	6	12	20	28	38	63	226
3	9	14	21	29	40	72	310

Table 12. 64 intervalles de temps en mois (source : Wikipédia)

La somme de ces valeurs vaut $\sum_{i=1}^{64} d_i = 2455$.

1. Donner la moyenne, la médiane, le premier et le troisième quartile de ces intervalles de temps.

On suppose dans la suite de l'exercice qu'on peut modéliser la durée entre deux tremblements de terre par une loi exponentielle $\mathscr{E}(\lambda)$.

2. Quelle est l'espérance de cette loi ? Déduisez-en une estimation $\widehat{\lambda}$ de la valeur de λ.

3. En utilisant cette estimation de λ, estimez la médiane, le premier et le troisième quartile de cette distribution.

4. Quelle est la probabilité qu'un nouveau tremblement de terre ait lieu dans les 12 prochains mois ? Dans les 60 prochains mois ?

5. Quelle est (en fonction de λ) la variance de $\overline{d} = \frac{1}{64}\sum_i d_i$? Déduisez-en un intervalle de confiance à 95% pour $1/\lambda$, puis pour λ et pour la médiane de la distribution.

Chapitre 6

Loi de Gauss

6.1 Loi normale ou de Gauss

C'est une loi omniprésente, la suite du cours permettra d'expliquer pourquoi.

Définition 26 *La loi normale, dite aussi loi de Gauss, ou loi gaussienne, de paramètres* μ *et* σ^2 *est la loi continue de densité*

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

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

Son espérance est μ et sa variance σ^2. La loi $\mathcal{N}(0,1)$ est dit *loi normale standard* ou *loi normale centrée réduite*. Si Z suit une loi normale centrée réduite, alors $X = \sigma Z + \mu$ suit une loi $\mathcal{N}(\mu, \sigma^2)$. Réciproquement, si X suit une loi $\mathcal{N}(\mu, \sigma^2)$ alors $Z = \dfrac{X-\mu}{\sigma}$ suit une loi normale centrée réduite.

De façon plus générale, si $X \sim \mathcal{N}(\mu, \sigma^2)$ alors $(aX + b) \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

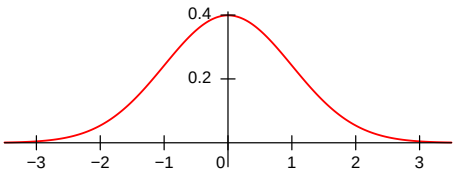


Figure 45. Densité de la loi normale centrée réduite

On a également l'importante propriété suivante :

> Si $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ sont **indépendantes**, alors $Z_1 + Z_2$ est normale, de loi $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Sa fonction de répartition ne peut pas s'exprimer avec des fonctions usuelles, on utilisera donc une table – ou un ordinateur. La densité étant symmétrique, on a ; pour $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z > a) = \mathbb{P}(Z < -a)$.

On en déduit que si $0 < \alpha < 1$, le quantile de niveau α est l'opposé du quantile de niveau $1-\alpha$: $z_\alpha = z_{1-\alpha}$.

Le quantile de niveau 0,975 de la loi $\mathcal{N}(0,1)$ est généralement connu par cœur : c'est $z_{0,975} = 1{,}96$. On en déduit l'*intervalle de pari*

$$\mathbb{P}(-1{,}96 \leqslant Z \leqslant 1{,}96) = 0{,}95.$$

6.2 Loi normale multivariée

Le résultat suivant est presque vrai : « si deux variables aléatoires X et Y suivent une loi normale et si $\mathrm{cov}(X,Y)=0$, alors X et Y sont indépendantes. »

On conçoit l'importance d'un tel résultat. Tel qu'énoncé ci-dessus il est faux; il manque une hypothèse : « (X,Y) suit une loi normale multivariée ». Il est donc difficile de faire l'économie des définitions suivantes :

Définition 27 *Soient* X *et* Y *des variables aléatoires. La loi de* (X,Y) *est une loi de Gauss bivariée si la loi marginale de* X *est gaussienne* $\mathcal{N}(\mu_X,\sigma_X^2)$*, et la loi de* Y *conditionnellement à* $X=x$ *est une loi normale, de variance indépendante de* x*, et d'espérance qui dépend linéairement de* x *:*

$$E(Y|X=x)=\alpha+\beta x,\quad \mathrm{var}(Y|X=x)=\tau.$$

On peut montrer qu'alors la loi marginale de Y *est une loi normale* $\mathcal{N}(\mu_Y,\sigma_Y^2)$*. Les coefficients* α *et* β *verifient*

$$\beta=r\frac{\sigma_Y}{\sigma_X}\ \text{ et }\ \alpha=\mu_y-r\frac{\sigma_Y}{\sigma_X}\mu_x,$$

où est r *le coefficient de corrélation entre* X *et* Y*. On a donc* $\mathrm{cov}(X,Y)=\sigma_{XY}=r\sigma_X\sigma_Y$*. On a d'autre part* $\tau=\mathrm{var}(Y|X=x)=(1-r^2)\sigma_Y^2$.

Ces formules donnent une interprétation simple et importante du coefficience de corrélation dans le cas gaussien multivarié : en espérance, Y s'écarte de r écart-types σ_Y de sa moyenne quand X s'écarte d'un écart-type σ_X de sa moyenne. Ou encore, r est le coefficient de proportionalité entre l'espérance de l'écart de Y à μ_Y, exprimé en écarts-types σ_Y, et l'écart de X à μ_X, exprimé en écarts-types σ_X. On a en effet

$$E(Y|X=x)=\mu_Y+r\frac{x-\mu_X}{\sigma_X}\sigma_Y$$

et

$$E\left(\frac{Y-\mu_Y}{\sigma_Y}\,\middle|\,X=x\right)=r\frac{x-\mu_X}{\sigma_X}.$$

On peut maintenant donner une définition générale de la loi gaussienne multivariée.

Définition 28 *Soient* $X_1,\dots,X_n$ *des variables aléatoires. On dit que* $(X_1,\dots,X_n)$ *suit une loi de Gauss multivariée si* $(X_1,\dots,X_{n-1})$ *suit une loi de Gauss multivariée et si la loi de* X_n *conditionnellement à* $X_1=x_1,\dots,X_{n-1}=x_{n-1}$ *est une loi normale, de variance indépendante de* $x_1,\dots,x_n$*, et d'espérance qui dépend linéairement de* $x_1,\dots,x_n$ *:*

$$E(X_n|X_1=x_1,\dots,X_{n-1}=x_{n-1})=\alpha+\beta_1x_1+\cdots+\beta_{n-1}x_{n-1},\quad \mathrm{var}(Y|X=x)=\tau.$$

La loi marginale de X_n *est normale.*

Cette loi est entièrement caractérisée par l'espérance $(\mu_1,\dots,\mu_n)$ *et de chacune des composantes, par les valeurs des variances* $\sigma_i^2=\mathrm{var}(X_i)$ *et des covariances* $\sigma_{ij}=\mathrm{cov}(X_i,X_j)$.

Il y a, dans le cas général comme dans le cas bivarié, des relations entre les coefficients $\alpha,\beta_1,\dots,\beta_{n-1}$ et les corrélations entre les composantes, mais elles sont plus techniques à écrire.

Proposition 22 *La loi de* $(X_1,\dots,X_n)$ *est une loi de Gauss multivariée, si et seulement si toute combinaison linéaire* $Z=a_1X_1+\cdots+a_nX_n$ *suit une loi normale.*

En particulier, chaque X_i est normale. On en déduit également que si les X_i sont indépendantes et si chaque X_i est normale, alors $(X_1,\dots,X_n)$ suit une loi normale multivariée.

On peut maintenant écrire la proposition suivante :

Proposition 23 *Si* (X,Y) *est suit une loi normale bivariée et si* $\mathrm{cov}(X,Y)=0$*, alors* X *et* Y *sont indépendantes.*

En pratique, on est souvent amené à considérer sans plus de commentaires que si des variables X et Y sont issues de mesures biologiques distinctes, et que X et Y sont normales, alors (X,Y) suit une loi normale bivariée ; et que l'indépendance de X et Y est donc équivalente à $\text{cov}(X,Y) = 0$. Il faudrait cependant dans une situation réelle vérifier au moins visuellement qu'on a bien un nuage de points qui a une forme elliptique : voyez les figures 46, 47, 48.

6.2.1 La densité dans le cas $n = 2$

Considérons le cas $n = 2$ et (X,Y) qui suit une loi normale bivariée, caractérisé par μ_X et μ_Y, les espérances de x et y, par σ_W^2, σ_Y^2, les variances de X et Y, et par $\sigma_{XY} = \text{cov}(X,Y) = r\sigma_X\sigma_Y$.
La densité conjointe de X et Y est

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-r^2}} \exp\left(\frac{1}{1-r^2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2r\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

Les figures 46 et 47 aident à comprendre la signification de ces paramètres, en montrant 500 points tirés au hasard d'une part, et la densité en niveaux de couleurs d'autre part, pour certaines valeurs de σ_X^2, σ_Y^2 et r. Remarquez en particulier que la forme elliptique des courbes de niveau de la densité.
On peut également écrire la densité dans le cas $\mu_X = \mu_Y = 0$ et $\sigma_X = \sigma_Y = 1$:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left(\frac{1}{1-r^2}\left(x^2 - 2rxy + y^2\right)\right).$$

Le cas général peut s'en déduire. On note qu'il est facile, si $r = 0$, de vérifier que $f(x,y) = \phi(x)\psi(y)$ avec ϕ et ψ les densités marginales de la X et Y (c'est-à-dire d'une loi normale standard) : on retrouve le résultat énoncé à la proposition 23.

6.2.2 Un exemple de variables gaussiennes dont la loi jointe n'est pas gaussienne

Il est facile de produire des exemples de variables X et Y qui sont gaussiennes mais dont la loi jointe n'est pas gaussienne ; ou autrement dit (X,Y) n'est pas gaussienne. Nous ne rentrerons pas dans les détails de la construction de cet exmple, contentons-nous de la figure 48, qui montrent un exemple d'une telle loi (500 points tirés au hasard d'une part, et sa densité en niveaux de couleurs d'autre part). La densité forme un fuseau et non plus une ellipse, bien que les lois marginales restent gaussiennes. On voit en particulier que la variance de Y conditionnellement à $X = x$ dépend de x (elle semble d'autant plus élevée que x est grand).

Cet exemple montre qu'il faut faire preuve de prudence : même si X et Y sont gaussiennes, on n'a pas forcément (X,Y) gaussienne ; en particulier on ne peut pas appliquer la proposition 23.

6.3 La loi du χ^2

On peut aussi écrire *loi du khi carré*, ou *loi du khi deux*.

Définition 29 *Soient $Z_1, \ldots, Z_d$ des variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$. La variable aléatoire $Y = \sum_{i=1}^d Z_i^2$ suit une loi continue à densité, appelée loi du χ^2 à d degrés de libertés, notée $\chi^2(d)$.*

Nous n'écrirons pas explicitement sa densité. Notons cependant que si d est pair, la loi $\chi^2(d)$ coincide avec la loi $\Gamma\left(\frac{d}{2}, \frac{1}{2}\right)$ (cf paragraphe 5.2.3).

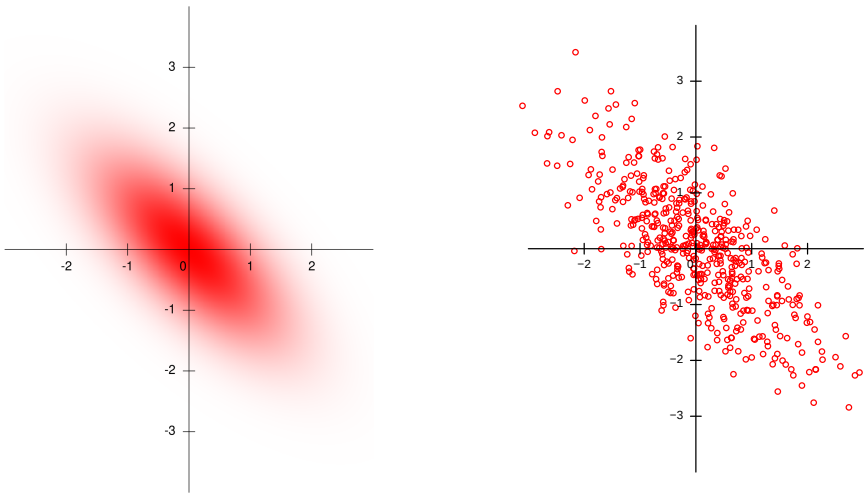


Figure 46. Densité et 500 points au hasard, pour (X,Y) normale bivariée avec $\sigma_X^2 = \sigma_Y^2 = 1$ et $r = -0{,}7$.

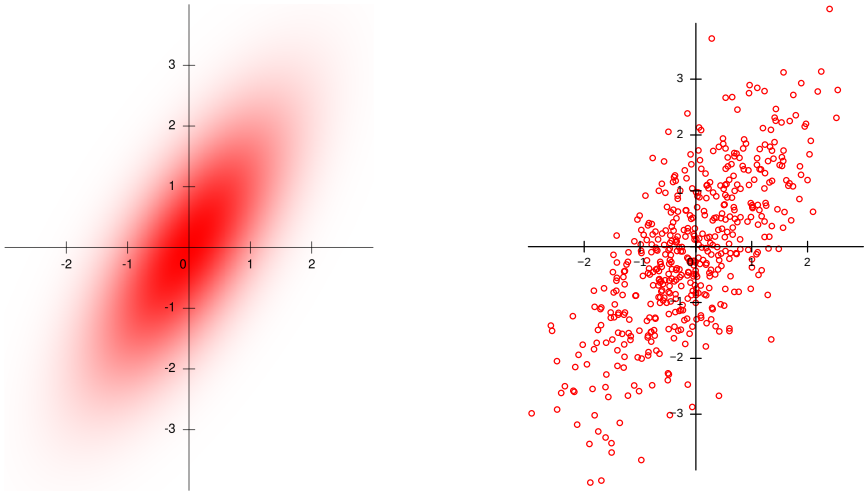


Figure 47. Densité et 500 points au hasard, pour (X,Y) normale bivariée avec $\sigma_X^2 = 1$, $\sigma_Y^2 = 2$ et $r = 0{,}7$.

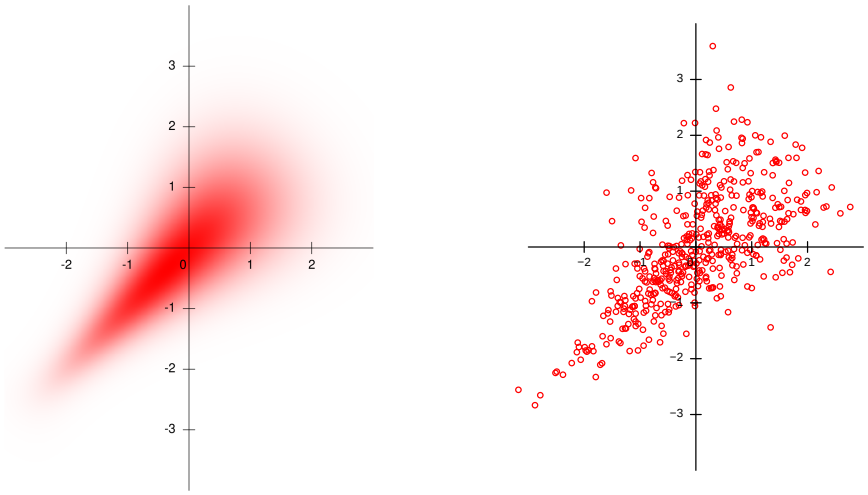


Figure 48. Densité et 500 points au hasard, pour (X,Y) non gaussien, X et Y étant gaussiens.

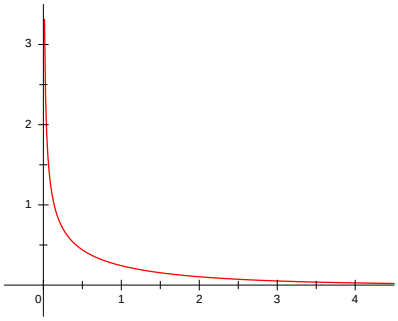


Figure 49. Densité du $\chi^2(1)$

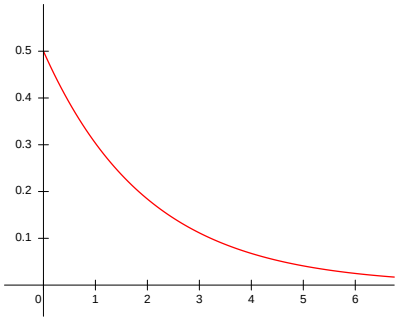


Figure 50. Densité du $\chi^2(2)$

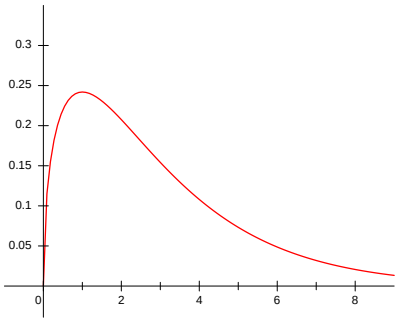


Figure 51. Densité du $\chi^2(3)$

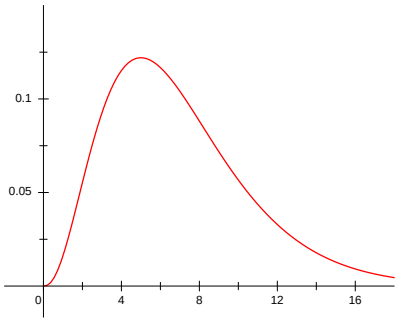


Figure 52. Densité du $\chi^2(7)$

De $Z_i \sim \mathcal{N}(0,1)$ on déduit que $E(Z_i^2) = \text{var}(Z_i) - E(Z_i)^2 = 1$. D'autre part on peut montrer que $\text{var}(Z_i^2) = 2$. On en déduit l'espérance et la variance de la loi $\chi^2(d)$:

Proposition 24 *Soit* $Y \sim \chi^2(d)$. *Alors* $E(Y) = d$ *et* $\text{var}(Y) = 2d$.

De la définition, on déduit également la proposition suivante.

Proposition 25 *Soient deux variables aléatoires indépendantes* $Y_1 \sim \chi^2(d_1)$ *et* $Y_2 \sim \chi^2(d_2)$. *Alors* $Y = Y_1 + Y_2$ *suit une loi* $\chi^2(d = d_1 + d_2)$.

Nous admettons la « réciproque » suivante.

Proposition 26 *Si* Y_1 *et* Y_2 *sont indépendantes, si* $Y_1 \sim \chi^2(d_1)$ *et* $Y = Y_1 + Y_2$ *suit une loi* $\chi^2(d)$ *alors* $Y_2 \sim \chi^2(d_2 = d - d_1)$.

6.4 Le théorème central de la limite

Ce théorème a d'abord été démontré par de Moivre au XVIII$^{\text{e}}$ siècle dans le cas d'une somme de lois de Bernoulli $\mathscr{B}\left(p = \frac{1}{2}\right)$. Il a été étendu par Laplace au cas où p est quelconque, et démontré sous sa forme actuelle par Lyapounov au début du XX$^{\text{e}}$ siècle. Son nom de « central » lui a été donné par Pólya (en allemand : *zentraler Grenzwertsatz*), en référence à la place « centrale » qu'il occupe en probabilités et statistiques ; c'est donc bien le théorème qui est central, pas la limite...

Attention à ne pas appeler improprement ce résultat « loi des grands nombres » (cf section 4.5.4). Notez que dans le théorème suivant il est nécessaire que les variables aléatoires aient une variance finie (ce n'était pas le cas dans le théorème 8).

Théorème 27 (Théorème central de la limite) *Soient* $X_1, X_2, \ldots$ *des variables aléatoires indépendantes et de même loi, admettant une espérance* μ *et une variance* σ^2.

On pose $S_n = X_1 + \cdots + X_n$. *Quand* n *tend vers l'infini, la loi de* $Y_n = \dfrac{S_n - n\mu}{\sqrt{n\sigma^2}}$ *s'approche de la loi normale centrée réduite* $\mathcal{N}(0,1)$.

Notez que $\mathrm{E}(\mathrm{S}_n) = n\mu$ et $\mathrm{var}(\mathrm{S}_n) = n\sigma^2$; on en déduit que pour tout n, $\mathrm{Y}_n = \dfrac{\mathrm{S}_n - \mu}{\sigma\sqrt{n\sigma^2}}$ est d'espérance nulle et de variance 1.

Notule mathématisante La formulation « la loi de Y_n s'approche de la loi normale $\mathcal{N}(0,1)$ » est imprécise. On peut la définir rigoureusement : on dit que Y_n *converge en loi* vers Z si pour tout $a \leqslant b$,

$$\mathbb{P}(a \leqslant \mathrm{Y}_n \leqslant b) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(a \leqslant \mathrm{Z} \leqslant b).$$

C'est en ce sens que, dans le théorème 27, on dit que la loi de Y_n s'approche de la loi normale.

En pratique, cela signifie que pour n « assez grand » on peut approcher la loi de S_n par la loi normale qui a même espérance et même variance, $\mathcal{N}(n\mu, n\sigma^2)$.

Corollaire 28 (Cas de la moyenne empirique) *La loi de* $\overline{\mathrm{X}}_n = \frac{1}{n}(\mathrm{X}_1 + \cdots + \mathrm{X}_n)$ *peut être approchée, dès que n assez grand, par une loi normale* $\mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$.

Corollaire 29 (Cas de la binomiale) *La loi binomiale* $\mathscr{B}in(n,p)$ *est la somme de n variables de Bernoulli indépendantes de paramètre p. On peut donc approcher cette loi binomiale par une loi normale* $\mathcal{N}(np, np(1-p))$ *« dès que n est assez grand ».*

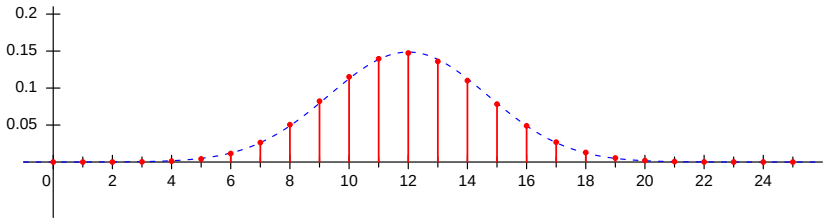


Figure 53. La fonction de masse de la loi $\mathscr{B}in(n=30, p=0{,}4)$ et la densité de $\mathcal{N}(np=12, np(1-p)=7{,}2)$

Parmi les lois que nous avons examinées dans le chapitre précédent, la loi binomiale négative et la loi Gamma sont des exemples de sommes de variables indépendantes de même loi. On peut donc, quand leur paramètre r est assez grand, les approcher par une loi normale de même espérance et de même variance.

Plus important sans doute, la loi du χ^2 : quand le nombre de degrés de liberté d est assez grand, la loi $\chi^2(d)$ peut être approchée par une normale $\mathcal{N}(\mu = d, \sigma^2 = 2d)$.

6.4.1 Quand n est-il assez grand ?

Nous sommes volontairement resté dans le flou sur ce point. On donne souvent des critères du genre « n est assez grand si $n \geqslant 30$ » mais c'est peu pertinent : cela dépend en fait de la loi des X_i. Le résultat suivant permet de quantifier l'erreur commise.

Théorème 30 (Théorème de Berry-Esséen) *Soient* $\mathrm{X}_1, \mathrm{X}_2, \ldots$ *des variables aléatoires indépendantes et de même loi, admettant une espérance* μ *et une variance* σ^2. *Soit* $\rho = \mathrm{E}\left(|\mathrm{X} - \mu|^3\right)$.

On pose $\mathrm{S}_n = \mathrm{X}_1 + \cdots + \mathrm{X}_n$ *et* $\mathrm{Y}_n = \dfrac{\mathrm{S}_n - n\mu}{\sqrt{n\sigma^2}}$. *On note* F_n *la fonction de répartition de* Y_n *et* Φ *la fonction de répartition de la loi normale centrée réduite* $\mathcal{N}(0,1)$. *On a, pour tout x,*

$$|\mathrm{F}_n(x) - \Phi(x)| \leqslant 0{,}48\frac{\rho}{\sigma^3\sqrt{n}}$$

Pour comprendre à quoi sert ce résultat, penchons-nous sur le cas de la binomiale. On sait que l'espérance de la loi $\mathscr{B}(p)$ est $\mu = p$, sa variance $\sigma^2 = p(1-p)$; la valeur de $\rho = \mathrm{E}\left(|\mathrm{X} - p|^3\right)$ est facile à calculer, c'est $\rho = p(1-p)\left(p^2 + (1-p)^2\right)$. Ainsi l'erreur commise est bornée par

$$0{,}48\frac{\rho}{\sigma^3\sqrt{n}} = 0{,}48\frac{p(1-p)\left(p^2+(1-p)^2\right)}{(p(1-p))^{3/2}\sqrt{n}} = 0{,}48\frac{p^2+(1-p)^2}{\sqrt{p(1-p)n}}$$

On a $p^2+(1-p)^2 \leqslant 1$, donc l'erreur est bornée par $0{,}48/\sqrt{p(1-p)n}$.

Pour donner un critère pour « n assez grand », il convient donc de le baser sur la valeur de $np(1-p)$; on est souvent assez peu exigeant, en prenant par exemple $np(1-p) > 10$ ou même 5. On rencontre aussi des critères qui donnent des résultats semblables, comme $np > 10$ et $n(1-p) > 10$.

6.4.2 Cas des lois discrètes : la correction de continuité

Dans le cas des lois discrètes, l'application d'une *correction de continuité* peut améliorer significativement les résultats obtenus par une application du théorème de la limite centrale.

Prenons un exemple : soit X une variable de loi $\mathscr{B}in(n=30, p=0{,}4)$. On a $\mathrm{E}(\mathrm{X}) = np = 12$ et $\mathrm{var}(\mathrm{X}) = np(1-p) = 7{,}2$. On peut approcher la loi de X par la loi normale $\mathscr{N}(\mu = 12, \sigma^2 = 7{,}2)$. Si on pose $\mathrm{Z} = \frac{\mathrm{X}-12}{\sqrt{7{,}2}}$, Z suit une loi normale centrée réduite.

Le calcul exact de $\mathbb{P}(10 \leqslant \mathrm{X} \leqslant 14)$ est possible mais fastidieux :

$$\begin{aligned}\mathbb{P}(10 \leqslant \mathrm{X} \leqslant 14) &= \binom{30}{10}0{,}4^{10}0{,}6^{20} + \binom{30}{11}0{,}4^{11}0{,}6^{19} + \cdots + \binom{30}{14}0{,}4^{14}0{,}6^{16} \\ &= 0{,}6483.\end{aligned}$$

En utilisant l'approximation normale, on procède classiquement comme suit :

$$\begin{aligned}\mathbb{P}(10 \leqslant \mathrm{X} \leqslant 14) &= \mathbb{P}\left(\frac{10-12}{\sqrt{7{,}2}} < \mathrm{Z} < \frac{14-12}{\sqrt{7{,}2}}\right) \\ &= \mathbb{P}(-0{,}75 < \mathrm{Z} < 0{,}75) \\ &= \mathrm{F}(0{,}75) - \mathrm{F}(-0{,}75) \\ &= 2 \times \mathrm{F}(0{,}75) - 1 \\ &= 0{,}5467.\end{aligned}$$

On voit que la qualité de l'approximation laisse à désirer. Remarquons d'autre part que si on avait calculé $\mathbb{P}(10 < \mathrm{X} < 14)$ de la même façon, on aurait trouvé le même résultat, la variable Z étant continue. Le calcul de $\mathbb{P}(\mathrm{X} = 10)$ par la même méthode conduirait à une valeur nulle, toujours à cause de la continuité de Z.

Le moyen de s'en sortir est d'écrire $\mathbb{P}(a \leqslant \mathrm{X} \leqslant b) = \mathbb{P}(a-0{,}5 \leqslant \mathrm{X} \leqslant b+0{,}5)$ (pour a, b entiers).

Le calcul devient :

$$\begin{aligned}\mathbb{P}(10 \leqslant \mathrm{X} \leqslant 14) &= \mathbb{P}(9{,}5 < \mathrm{X} < 14{,}5) \\ &= \mathbb{P}(-0{,}93 < \mathrm{Z} < 0{,}93) \\ &= 0{,}6476.\end{aligned}$$

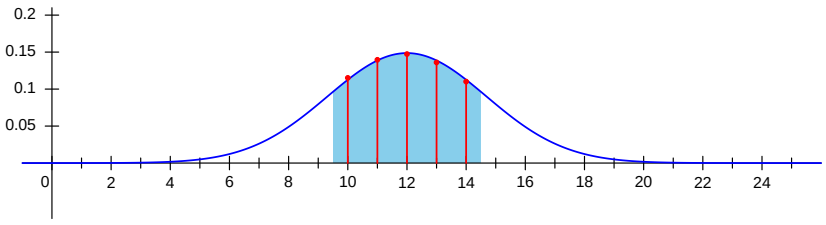


Figure 54. Calcul de $\mathbb{P}(\mathrm{X}=12)+\cdots+\mathbb{P}(\mathrm{X}=14)$ avec correction de continuité

Cette fois l'erreur est inférieure à un millième. On peut également calculer $\mathbb{P}(\mathrm{X}=10)$:

$$\begin{aligned}\mathbb{P}(\mathrm{X}=10) &= \mathbb{P}(9{,}5 < \mathrm{X} < 10{,}5) \\ &= \mathbb{P}(-0{,}93 < \mathrm{Z} < -0{,}56) \\ &= 0{,}112.\end{aligned}$$

La valeur exacte est $\binom{30}{10}0{,}4^{10}0{,}6^{20} = 0{,}115$.

Notons qu'on peut également approcher directement $\mathbb{P}(10)$ par la densité de la loi $\mathcal{N}(12,7{,}2)$ calculée en $x = 10$: $\frac{1}{\sqrt{2\pi \cdot 7{,}2}} e^{-\frac{(10-12)^2}{2 \cdot 7{,}12}} = 0{,}113$: voir la remarque finale du paragraphe 3.3.1 sur l'approximation de $\mathbb{P}(x - h/2 \leqslant \mathrm{X} \leqslant x + h/2)$ par $h \times f(x)$; ici on a $h = 1$.

En pratique, la correction de continuité est conseillée quand n n'est pas très grand, ou quand l'intervalle est petit (en particulier, quand il est réduit à un point comme dans ce dernier exemple).

6.5 Exercices

Exercice 1 On considère une variable aléatoire X qui suit une loi normale $\mathcal{N}(\mu = 5, \sigma^2 = 36)$.

On pose

$$\mathrm{Y}_1 = \mathrm{X} - 5 \qquad \mathrm{Y}_2 = \tfrac{1}{36}(\mathrm{X} - 5)$$

$$\mathrm{Y}_3 = \tfrac{1}{6}(\mathrm{X} - 5) \qquad \mathrm{Y}_4 = \tfrac{1}{36}(\mathrm{X} - 5)^2$$

1. Quelles sont les lois de Y_1, Y_2, Y_3 et Y_4 ?

2. Est-ce que X et Y_1 sont indépendantes ?

3. Quelle est la loi de $\mathrm{X} + \mathrm{Y}_1$?

Exercice 2 On considère X et Y, deux variables indépendantes de loi $\mathcal{N}(0,1)$. On pose $\mathrm{Z} = a\mathrm{X} + b\mathrm{Y}$.

1. Quelle est la covariance de X et Z ?

2. Quelle est la variance de Z ?

3. On fixe $a = 0{,}8$. Quelle valeur doit prendre b pour que var(Z) = 1 ? Quelle est alors la valeur du coefficient de corrélation corr(X,Z) ?

Exercice 3 On considère X et Y deux variables aléatoires suivant une loi normale, d'espérances respectives $\mu_\mathrm{X} = 1{,}8$, $\mu_\mathrm{Y} = 1$, et de variances respectives $\sigma^2_\mathrm{X} = 2$ et $\sigma^2_\mathrm{Y} = 2{,}3$; le coefficient de corrélation de X et Y est $r = 0{,}7$.

On suppose en outre que (X,Y) suit une loi normale bivariée.

1. Calculez cov(X,Y).

2. On pose $\mathrm{Z} = \mathrm{Y} - \mathrm{X}$. Quelle est l'espérance de Z ? Et sa variance ?

3. Calculez la probabilité que $\mathrm{Y} \geqslant \mathrm{X}$.

Exercice 4 La taille de la population mâle adulte de Klow (en Syldavie) suit une loi normale de moyenne $\mu = 175$ cm et d'écart-type $\sigma = 10$ cm.

1. Quelle proportion de la population a une taille supérieure ou égale à 175 cm ?

2. Quelle est la probabilité d'observer dans cette population un individu de taille *exactement égale* à 175 cm ? Et un individu de taille comprise entre 174,5 et 175,5 cm ?

3. Quelle est la probabilité d'observer un individu de taille supérieure à 180 cm ? Et de taille supérieure à 179,5 cm ?

4. Pour quelle valeur de a a-t-on la taille de 95% de la population entre $175 - a$ et $175 + a$?

Exercice 5

On considère deux variables aléatoires X et Y ; on suppose que (X,Y) suit une loi normale bivariée. On a E(X) = 2, E(Y) = 3, var(X) = 1, var(Y) = 2 et cov(X,Y) = 1.

1. On pose $\mathrm{Z} = \mathrm{Y} + a\mathrm{X}$. Calculez, en fonction de a, l'espérance et la variance de Z.

2. Calculez cov(Z,X), cov(Z,Y), et cov(Z, X − Y).

3. Pour quelle valeur de a a-t-on cov(X,Z) = 0 ? Dans ce cas, X et Z sont-elles indépendantes ?

Exercice 6 Une méthode de dosage pour un biomarqueur donne une mesure X qui suit une loi normale $\mathcal{N}(\mu, \sigma^2 = 0{,}16)$ où μ est la « vraie » valeur de la concentration du biomarqueur. On supposera dans la suite que $\mu = 4$.

1. Quel est l'intervalle de fluctuation (ou de pari) à 95% pour X ?

2. Pour augmenter la précision dans la mesure de μ, on décide de faire n dosages $X_1, \ldots, X_n$ et de prendre leur moyenne $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. On suppose que ces n mesures sont indépendantes.

Quel est l'intervalle de fluctuation si $n = 4$?

3. Quelle valeur de n faut-il choisir pour avoir un intervalle de fluctuation pour $\overline{X}_n$ inclu dans l'intervalle $[3{,}8; 4{,}2]$?

Exercice 7

Un biomarqueur B_1 suit une distribution normale $\mathcal{N}(\mu = 4, \sigma^2 = 0{,}64)$ chez les individus sains. Chez les individus infectés par le parasite *Pica pica*, il suit une distribution $\mathcal{N}(\mu = 6{,}5, \sigma^2 = 0{,}64)$.

En présence de symptômes évocateurs d'une infection, on propose d'administrer un traitement antiparasitaire quand $B_1 > s$ avec $s = 5$. Si $B_1 < s$ on met le patient en observation quelques jours avant de commencer le traitement si les symptômes persistent.

Partie I *(la partie II est indépendante de la partie I)*

1. Quelle est la spécificité de la procédure, c'est-à-dire la probabilité pour que $B_1 < 5$ chez un individu sain ?

2. Quelle est sa sensibilité, c'est-à-dire la probabilité qu'un individu infecté (et symptomatique) reçoive immédiatement le traitement ?

3. Quelle valeur faut-il choisir pour s pour que la sensibilité soit égale à 0,99 ? Quelle est alors la spécificité ?

Partie II

On dispose d'un deuxième biomarqueur B_2 ; on suppose que (B_1, B_2) suit une loi normale bivariée (autrement dit, c'est un vecteur gaussien).

On donne, pour la population saine, la loi marginale $B_2 \sim \mathcal{N}(\mu = 2, \sigma^2 = 0{,}16)$ et la corrélation $\text{cor}(B_1, B_2) = -0{,}625$.

4. Que vaut la covariance $\text{cov}(B_1, B_2)$?

5. On pose $B = B_1 + 1{,}6 \cdot B_2$. Quelle est la loi de B ?

6. Dans la population des individus infectés on a $B_1 \sim \mathcal{N}(\mu = 6{,}5, \sigma^2 = 0{,}64)$, $B_2 \sim \mathcal{N}(\mu = 2{,}45, \sigma^2 = 0{,}16)$, et $\text{cor}(B_1, B_2) = -0{,}625$. Quelle est la loi de B (défini à la question précédente) dans cette population ?

7. Calculer $\mathbb{P}(B > 9)$ dans la population saine et dans la population infectée.

Exercice 8

On admettra que la taille des hommes adultes dans la population suit une loi normale de moyenne $\mu = 175$ cm et d'écart-type $\sigma = 7$ cm, alors que celle des femmes a pour moyenne $\mu = 165$ cm, et écart-type 7 cm.

1. Quelle est la probabilité pour un homme que sa taille soit comprise entre 170 et 180 cm ? Et pour une femme ?

2. En supposant un sex ratio de 1 : 1 dans la population, quelle est la probabilité que la taille d'un individu pris au hasard soit comprise entre 170 et 180 cm ?

3. Quelle est la taille médiane de la population ? (vérifiez le résultat par un calcul).

4. Quelles sont l'espérance et la variance de la taille dans la population ?

Pour répondre à cette question, on pourra utiliser la propriété suivante :

$$\mathrm{E}(f(X)) = \mathrm{E}(f(X)|\text{homme})\mathbb{P}(\text{homme}) + \mathrm{E}(f(X)|\text{femme})\mathbb{P}(\text{femme})$$

où X est la taille et f une fonction quelconque.

5. On a relevé la taille d'un individu sur un registre, **mesurée au centimètre prêt :** 180 cm. Le sexe de l'individu n'est pas mentionné. Quel est la probabilité qu'il s'agisse d'un homme ?

Exercice 9 ***

On rappelle que si $X_1, \ldots, X_n$ sont indépendantes de loi de Poisson respectives $\mathcal{P}(\lambda_1), \ldots, \mathcal{P}(\lambda_n)$ alors $X = X_1 + \cdots + X_n$ suit une loi $\mathcal{P}(\lambda = \lambda_1 + \cdots + \lambda_n)$.

1. Soit $X \sim \mathscr{P}(\lambda)$ et $n \in \mathbb{N}$. Montrer qu'on peut écrire X comme somme n variables aléatoires indépendantes. Peut-on pour autant en déduire que la loi de X est bien approchée par une loi normale ?

2. Si $Y \sim \mathscr{P}(\alpha)$, on peut montrer que $\rho = E(|Y-\alpha|^3) < \alpha + 6\alpha^2 + 8\alpha^3$. Montrer que $\rho < 2\alpha$ si α est assez petit.

3. En déduire que l'erreur commise en approchant la loi de $X \sim \mathscr{P}(\lambda)$ par une loi normale tend vers 0 quand λ tend vers l'infini.

Chapitre 7

Méthode du Delta

7.1 Fonctions d'une variable aléatoire

Si X est une variable aléatoire, et $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction, alors $Y = \phi(X)$ est une variable aléatoire. On rappelle que son espérance peut s'obtenir, si X est une variable discrète, par

$$\boxed{E(Y) = \sum_x \phi(x)\mathbb{P}(X = x)} \tag{7.1}$$

et si X est une variable continue de densité $f(x)$, par

$$\boxed{E(Y) = \int_{-\infty}^{+\infty} \phi(x) f(x) \mathrm{d}x.} \tag{7.2}$$

7.1.1 Loi de $Y = \phi(X)$

Cas discret

Dans le cas discret, on peut retrouver la loi de Y à partir de celle X, en écrivant

$$P(Y = y) = \sum_{x:\phi(x)\,y} \mathbb{P}(X = x).$$

Quand la loi de X est tabulée c'est particulièrement élémentaire.

Cas continu à densité

Soit X une variable aléatoire de densité f et de fonction de répartition F. Soit ϕ une fonction de $\mathbb{R}$ dans $\mathbb{R}$. On pose $Y = \phi(X)$. Pour trouver sa densité (en supposant qu'elle en ait une ! voyez l'exemple 36), il est en général plus simple de commencer par exprimer sa fonction de répartition G en fonction de F, puis de la dériver.

Exemple 31 Si $Y = aX + b$ avec $a > 0$, alors $G(t) = \mathbb{P}(aX + b \leqslant t) = \mathbb{P}\left(X \leqslant \dfrac{t-b}{a}\right) = F\left(\dfrac{t-b}{a}\right)$, et la densité de Y est $g(t) = \dfrac{1}{a} f\left(\dfrac{t-b}{a}\right)$. □

Exemple 32 Si $X \sim \mathcal{U}([0,1])$, l'exemple précédent permet de vérifier que $Y = (b-a)X + a$ suit une loi $\mathcal{U}([a,b])$. Si on connaît l'espérance de X : $E(X) = \frac{1}{2}$ et sa variance : $\text{var}(X) = \frac{1}{12}$, on retrouve facilement les résultats développés dans les exemples 8, 12, 13 et 14 :

$$\begin{aligned} E(X) &= (b-a)E(X) + a \\ &= \frac{1}{2}(a+b) \end{aligned}$$

et

$$\begin{aligned} \text{var}(X) &= (b-a)^2 \text{var}(X) \\ &= \frac{1}{12}(b-a)^2. \end{aligned}$$

□

Exemple 33 Soit T une v.a. de loi $\mathcal{E}(\lambda)$; sa fonction de répartition est $F(t) = 1 - e^{-\lambda t}$ pour $t \geqslant 0$ (cf section 5.2.2). On pose $T_1 = aT$ (a est une constante positive). On a alors

$$\begin{aligned} \mathbb{P}(T_1 \leqslant t) &= \mathbb{P}(T \leqslant \frac{1}{a} t) \\ &= 1 - e^{-\frac{\lambda}{a} t}. \end{aligned}$$

On reconnaît la fonction de répartition de la loi $\mathcal{E}\left(\frac{\lambda}{a}\right)$. □

Exemple 34 Soit X une v.a. de loi uniforme sur $[-1,1]$, et $Y = X^2$. Pour $t < 0$ on a $G(t) = P(Y \leqslant t) = 0$. Pour $t \geqslant 0$, on a

$$\begin{aligned} G(t) &= \mathbb{P}(X^2 \leqslant t) \\ &= \mathbb{P}(-\sqrt{t} \leqslant X \leqslant \sqrt{t}) \\ &= \mathbb{P}(X \leqslant \sqrt{t}) - \mathbb{P}(X \leqslant -\sqrt{t}) \\ &= F(\sqrt{t}) - F(-\sqrt{t}) \\ &= \begin{cases} 1 & \text{si } t \geqslant 1 \\ \sqrt{t} & \text{si } t \leqslant 1 \end{cases} \end{aligned}$$

car

$$F(t) = \begin{cases} 0 & \text{si } t \leqslant -1 \\ \frac{1}{2}(t+1) & \text{si } -1 \leqslant t \leqslant 1 \\ 1 & \text{si } 1 \leqslant t \end{cases}$$

et donc, si $t \geqslant 1$, $F(\sqrt{t}) = 1$ et $F(-\sqrt{t}) = 0$, et si $t \leqslant 1$, $F(\sqrt{t}) - F(-\sqrt{t}) = \frac{1}{2}\left(\left(\sqrt{t}+1\right) - \left(-\sqrt{t}+1\right)\right) = \sqrt{t}$.
Enfin, la densité de Y, $g(t)$, est la dérivée de $G(t)$:

$$g(t) = \begin{cases} \dfrac{1}{2\sqrt{t}} & \text{si } 0 \leqslant t \leqslant 1 \\ 0 & \text{sinon} \end{cases}$$

□

Exemple 35 Soit X une v.a. de loi uniforme sur $[0,1]$, et $Y = -\log(1-X)$. Alors

$$\begin{aligned} G(t) &= \mathbb{P}(Y \leqslant t) = \mathbb{P}(\log(1-X) \geqslant -t) = \mathbb{P}\left(1 - X \geqslant e^{-t}\right) \\ &= \mathbb{P}\left(X \leqslant 1 - e^{-t}\right) = F\left(1 - e^{-t}\right) \end{aligned}$$

et donc $g(t) = G'(t) = f\left(1 - e^{-t}\right) e^{-t}$.
Si $t < 0$, $1 - e^{-t} < 0$ et donc $g(t) = 0$. Si $t \geqslant 0$, alors $1 - e^{-t} \in [0,1]$, et $g(t) = e^{-t}$. □

Exemple 36 Attention aux pièges!

Soit ϕ la fonction définie par

$$\phi(x) = \begin{cases} 0 & \text{si } x < 1 \\ x-1 & \text{si } x \geqslant 1. \end{cases}$$

Si $X \sim \mathscr{U}([0,2])$, alors $Y = \phi(X)$ n'admet pas de densité! En effet $\mathbb{P}(Y = 0) = \mathbb{P}(X < 1) = \frac{1}{2}$. D'autre part, pour $a,b \in]0,1]$, on $\mathbb{P}(a \leqslant Y \leqslant b) = \mathbb{P}(a+1 < X < b+1) = \frac{1}{2}(b-a)$. Nous avons déjà rencontré cette loi à la section 3.4. □

Exemple 37 Application au problème de la simulation informatique

Comment procèdent les ordinateurs pour simuler des valeurs d'une variable aléatoire X ayant une loi donnée, de fonction de répartition F?

La première étape est de simuler une variable U de loi discrète uniforme sur $[0,\dots,N-1]$, où N est un grand entier. Pour cette étape, il existe une multitude d'algorithmes plus ou moins satisfaisants.

On considère ensuite que $V = U/N$ est une variable aléatoire uniforme sur l'intervalle $[0,1]$ (quand N est très grand, c'est une bonne approximation, compte-tenu du fait que les ordinateurs ne manipulent de toutes façons que des nombres réels avec un nombre fini de chiffres après la virgule...).

Une fois qu'on dispose ainsi de valeurs aléatoires prises uniformément sur l'intervalle $[0,1]$, on peut les transformer pour obtenir n'importe quelle loi de fonction de répartition connue. En effet il suffit de poser $X = F^{-1}(V)$ où F^{-1} est la réciproque de F. On a alors

$$\mathbb{P}(X \leqslant t) = \mathbb{P}\big(F^{-1}(V) \leqslant t\big) = \mathbb{P}\big(V \leqslant F(t)\big) = F(t).$$

Ceci montre que X est une variable aléatoire de fonction de répartition F. □

7.2 Méthode du Delta : approximation de l'espérance et de la variance de $Y = \phi(X)$

Les formules exactes pour l'espérance et la variance de $Y = \phi(X)$ peuvent être difficiles à appliquer en pratique. Les approximations qui suivent sont plus faciles d'utilisation.

7.2.1 Approximations d'une fonction

Soit ϕ une fonction, et x_0 un point où ϕ et ses dérivées successives sont définies.

Près de x_0, c'est-à-dire en un point x tel que $h = x - x_0$ est petit, $\phi(x) = \phi(x_0 + h)$ n'est pas très différent de $\phi(x_0)$:

$$\phi(x) = \phi(x_0 + h) \simeq \phi(x_0).$$

On obtient une meilleure approximation en approchant le graphe de ϕ par sa tangente :

$$\phi(x_0 + h) \simeq \phi(x_0) + h\phi'(x_0),$$

ou encore

$$\phi(x) \simeq \phi(x_0) + (x - x_0)\phi'(x_0).$$

Une approximation encore plus précise est obtenue avec :

$$\phi(x) \simeq \phi(x_0) + (x - x_0)\phi'(x_0) + \frac{1}{2}(x - x_0)^2\phi''(x_0).$$

De façon générale, on a une approximation « d'ordre n » :

$$\phi(x) \simeq \phi(x_0) + (x - x_0)\phi'(x_0) + \frac{1}{2!}(x - x_0)^2\phi''(x_0) + \frac{1}{3!}(x - x_0)^3\phi'''(x_0) + \cdots + \frac{1}{n!}(x - x_0)^n f^n(x_0).$$

Exemple 38 La fonction $f(x) = \sqrt{x}$ a pour dérivées $f'(x) = \frac{1}{2\sqrt{x}}$ et $f''(x) = -\frac{1}{4}x^{-3/2}$. On a $f(16) = 4$, $f'(16) = \frac{1}{2}\frac{1}{4} = 0{,}125$ et $f''(1) = -\frac{1}{4}16^{3/2} = -\frac{1}{256} = 0{,}0039$; au voisinage de $x_0 = 16$, la fonction f peut s'approcher par

$$\sqrt{16+x} \simeq 4$$
$$\sqrt{16+x} \simeq 4 + 0{,}125x$$
$$\sqrt{16+x} \simeq 4 + 0{,}125x - 0{,}00195x^2$$

Avec $x = 1$, on obtient comme approximations pour $\sqrt{17}$: 4, puis 4,125, puis 4,12305 ; la vraie valeur est 4,12311. □

7.2.2 Méthode du Delta

Soient $\mu = E(X)$ et $\sigma^2 = \text{var}(X)$.
Si on considère « l'approximation d'ordre 1 » de $\phi(x)$ au voisinage de μ,

$$\phi(x) \simeq \phi(\mu) + (x-\mu)\phi'(\mu),$$

on a

$$\begin{aligned} E(Y) &\simeq E\big(\phi(\mu) + (X-\mu)\phi'(\mu)\big) \\ &= \phi(\mu) + (E(X)-\mu)\phi'(\mu) \end{aligned}$$

et donc

$$\boxed{E(Y) \simeq \phi(\mu).} \tag{7.3}$$

On a aussi

$$\begin{aligned} \text{var}(Y) &\simeq \text{var}\big(\phi(\mu) + (X-\mu)\phi'(\mu)\big) \\ &= \phi'(\mu)^2 \,\text{var}(X-\mu) \end{aligned}$$

et donc

$$\boxed{\text{var}(Y) \simeq \phi'(\mu)^2\sigma^2.} \tag{7.4}$$

On peut améliorer l'approximation de l'espérance en considèrant « l'approximation d'ordre 2 » de $\phi(x)$ au voisinage de μ :

$$\phi(x) \simeq \phi(\mu) + (x-\mu)\phi'(\mu) + \frac{1}{2}(x-\mu)^2\phi''(\mu).$$

On en déduit de la même façon

$$\boxed{E(Y) \simeq \phi(\mu) + \frac{1}{2}\phi''(\mu)\sigma^2.} \tag{7.5}$$

Il faut noter que les approximations de ϕ sont d'autant meilleures que x est proche de μ ; les approximations obtenues pour E(Y) et var(Y) seront donc d'autant meilleures que la variance de X est faible.

Exemple 39 Si X a espérance $E(X) = \mu = 16$ et variance $\text{var}(X) = \sigma^2 = 1$, alors $Y = \phi(X) = \sqrt{X}$ a pour espérance approximativement $\phi(16) + \frac{1}{2}\phi''(16) \times 1 = 4 - 0{,}00195 = 3{,}998$ et pour variance approximativement $\phi'(16)^2 \times 1 = 0{,}0156$ (cf exemple 38 pour les dérivées de ϕ). □

7.3 Méthode Delta : théorème limite

Le théorème central limite fournit un exemple d'une suite de variables $\overline{X}_n$ telle que la loi de

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

s'approche de la loi $\mathcal{N}(0,1)$ quand $n \to +\infty$ (autrement dit, la loi de $\overline{X}_n$ peut être approchée par la loi $\mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$ quand n assez grand).
En appliquant la méthode du Delta à cette suite de variables, on est certain que dès que n est assez grand, la variance de $\overline{X}_n$, $\text{var}\left(\overline{X}_n\right) = \frac{1}{n}\sigma^2$, devient assez petite pour que l'approximation soit bonne.

De plus, la loi de $\overline{X}_n$ étant approximativement normale, si la variance est assez petite pour que l'approximation linéaire de ϕ au voisinage de μ soit bonne, la loi de $\phi\left(\overline{X}_n\right)$ sera également approximativement normale.
Le théorème suivant formalise cette idée, en la généralisant un peu.

Théorème 31 (Méthode Delta) *Soient $Z_1, Z_2, \ldots$ une suite de variables aléatoires de variances $\sigma_1^2, \sigma_2^2, \ldots$ telles que $\sigma_n^2 \underset{n\to+\infty}{\longrightarrow} 0$. On suppose qu'il existe une constante μ telle que la loi de*

$$\frac{Z_n - \mu}{\sigma_n}$$

s'approche de la loi $\mathcal{N}(0,1)$ quand $n \to +\infty$ (autrement dit, la loi de Z_n peut être approchée par la loi $\mathcal{N}\left(\mu, \sigma_n^2\right)$ dès que n est assez grand).
Alors, si $\phi'(\mu) \neq 0$, la loi de

$$\frac{\phi(Z_n) - \phi(\mu)}{\phi'(\mu)\sigma_n}$$

s'approche de la loi $\mathcal{N}(0,1)$ (autrement dit, dès que n assez grand, la loi de $\phi(Z_n)$ peut être approchée par la loi $\mathcal{N}(\phi(\mu), \phi'(\mu)^2\sigma_n^2)$).

7.4 Applications

7.4.1 Cas d'une moyenne empirique

On revient au cas évoqué plus haut : soient $X_1, X_2, \ldots$ des variables aléatoires indépendantes et de même loi, admettant une espérance μ et une variance σ^2. On pose

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

La loi de

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

s'approche d'une loi $\mathcal{N}(0,1)$ (autrement dit, pour n grand on peut approcher la loi de $\overline{X}_n$ par une loi normale $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$).

Le théorème qui précède permet de conclure que (si $\phi'(\mu) \neq 0$), la loi de

$$\frac{\phi\left(\overline{X}_n\right) - \phi\left(\mu\right)}{\phi'\left(\mu\right)\frac{\sigma}{\sqrt{n}}}$$

s'approche de la loi normale centrée réduite $\mathcal{N}(0,1)$.

En pratique, on peut approcher la loi de $\phi\left(\overline{X}_n\right)$ par la loi normale $\mathcal{N}\left(\phi(\mu), \phi'(\mu)^2 \frac{\sigma^2}{n}\right)$.

7.4.2 Loi de Poisson

Pour tout $\lambda > 0$ on note U_λ une variable aléatoire de loi $\mathcal{P}(\lambda)$. On admettra que quand $\lambda \to +\infty$, la loi de

$$\frac{U_\lambda - \lambda}{\sqrt{\lambda}}$$

s'approche de la loi $\mathcal{N}(0,1)$ (cf exercice 9 chapitre 6).
On pose $Z_\lambda = \frac{1}{\lambda}U$ de sorte que $E(Z_\lambda) = 1$ et $\text{var}(Z_\lambda) = \frac{1}{\lambda} \underset{\lambda \to +\infty}{\longrightarrow} 0$. La loi de

$$\frac{Z_\lambda - 1}{\frac{1}{\sqrt{\lambda}}}$$

s'approche de la loi $\mathcal{N}(0,1)$ quand $\lambda \to +\infty$.
On prend $\phi(x) = \sqrt{x}$. Sa dérivée est $\phi'(x) = \frac{1}{2\sqrt{(x)}}$. On a $\phi(1) = 1$ et $\phi'(1) = \frac{1}{2}$. D'après le théorème 31 la loi de

$$\frac{\phi(Z_\lambda) - 1}{\frac{1}{2} \times \frac{1}{\sqrt{\lambda}}}$$

s'approche de la loi $\mathcal{N}(0,1)$ quand $\lambda \to +\infty$, ou autrement dit la loi de $\phi(Z_\lambda)$ peut être approchée par la loi $\mathcal{N}\left(1, \frac{1}{4\lambda}\right)$ dès que λ est assez grand..

On a $\sqrt{U_\lambda} = \sqrt{\lambda} Z_\lambda$, donc la loi de $\sqrt{U_\lambda}$ peut être approchée par la loi $\mathcal{N}\left(\sqrt{\lambda}, \frac{1}{4}\right)$ quand λ est assez grand.

7.5 Exercices

Exercice 1 (Loi de la p-valeur)

Soit X une variable aléatoire de densité f et de fonction de répartition F. On pose $P = 1 - F(X)$. Déterminer la loi de P. Commenter.

Exercice 2 Soit X une variable aléatoire continue de loi uniforme sur $[1-\alpha, 1+\alpha]$, avec $\alpha > 0$. Calculez son espérance et sa variance. Calculez l'espérance de e^X, d'abord de façon exacte puis de façon approchée avec la méthode Delta d'ordre 1 (équation 7.3) et avec la méthode Delta d'ordre 2 (équation 7.5) l'espérance de e^X. Comparez numériquement les valeurs obtenues pour quelques valeurs de α.

Exercice 3 On a un tableau de données donnant la concentration d'un médicament dans le sang de patients en fonction du temps après la prise. Une valeur moyenne et un écart-type ont été calculés.

t	0	1	2	3	4
moyenne	20,1	14,4	10,3	7,4	5,3
écart-type	1,0	1,8	2,3	3,4	4,1

Table 13. Moyennes des concentrations

On décide qu'il faut en fait s'intéresser au log de la concentration. Les données originales ne sont pas disponibles. On peut prendre le logarithme de la moyenne comme approximation de la moyenne des logarithmes, mais la méthode du Delta permet de faire mieux : utilisez la pour retrouver au mieux la valeur moyenne des logarithmes des concentrations.

Exercice 4 ***

On admettra que si ϕ est une fonction dérivable, on a pour tout δ

$$\phi(\mu+\delta) = \phi(\mu) + \delta\phi'(\mu + c_\delta),$$

où c_δ est entre 0 et δ.

1. Soit Z une variable aléatoire $\mathcal{N}(\mu,\sigma^2)$. Montrer qu'on peut écrire Z sous la forme $Z = \mu + \sigma\varepsilon$ avec $\varepsilon \sim \mathcal{N}(0,1)$.

2. Soit ϕ une fonction dérivable. Montrer que

$$\frac{\phi(Z) - \phi(\mu)}{\phi'(\mu)\sigma} = \varepsilon \times \left(\frac{\phi'(\mu + c)}{\phi'(\mu)}\right),$$

avec c entre 0 et $\sigma\varepsilon$ (la valeur de c dépend de ε).

3. En déduire que si $Z_1, Z_2, \ldots$ sont normales, d'espérance μ et de variances $\sigma_1^2, \sigma_2^2, \ldots$ telles que $\sigma_n^2 \underset{n\to+\infty}{\longrightarrow} 0$ alors quand n grand,

$$\frac{\phi(Z_n) - \phi(\mu)}{\phi'(\mu)\sigma_n}$$

s'approche d'un loi $\mathcal{N}(0,1)$. (c'est une version un peu simplifiée du théorème 7.3).

Exercice 5 Le *coefficient de variation* d'une variable aléatoire d'espérance μ et d'écart-type σ est $cv = \frac{\sigma}{\mu}$.

On mesure par spectrométrie de masse la concentration d'une molécule dans les urines de plusieurs individus. La mesure a une imprécision qu'on quantifie généralement par son coefficient de variation ; s'il est inférieur à 0,05, on considère que le protocole de mesure est correct.

On a des mesures comme celles-ci (exprimées en pg par mg de créatinine) :

Patient	Mesure 1	Mesure 2	Mesure 3
1	723,7	735,6	755,0
2	253,0	269,0	248,1
⋮	⋮	⋮	⋮

1. Estimer les coefficients de variation pour chacun des deux patients.

2. On note log le logarithme naturel. On considère une variable aléatoire X à valeurs positives ; on pose $Y = \log(X)$. En utilisant l'approximation donnée par la méthode du Delta :

$$\mathrm{var}\,(\Phi(X)) \simeq \Phi'\,(E(X))^2 \times \mathrm{var}(X),$$

montrez que l'écart-type de Y est proche du coefficient de variation de X.

3. Calculez les log des mesures rapportées dans la table ci-dessus, puis leur écart-type pour chacun des deux patients ; comparez avec les coefficients de variation calculés à la question 1. L'approximation est-elle bonne ?

4. En supposant que le logarithme des concentrations suit une loi normale, calculez un intervalle de confiance pour chacun des deux coefficients de variation.

Exercice 6 (log odds)

On considère n variables $X_1, \ldots, X_n$ indépendantes de loi $\mathcal{B}(p)$.

1. On pose $X = X_1 + \cdots + X_n$. Quelle est la loi de X ?

2. On pose $\widehat{p} = \frac{1}{n}X$. Montrer que $\widehat{p}$ est un estimateur non biaisé de p et calculer sa variance.

3. On s'intéresse à $\phi = \log\left(\frac{p}{1-p}\right)$, qu'on va naturellement estimer par $\widehat{\phi} = \log\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$. Utiliser la méthode du Delta pour montrer que sa variance est approximativement

$$\mathrm{var}\,\widehat{\phi} = \frac{1}{np(1-p)}$$

4. En déduire qu'on peut estimer la variance de $\widehat{\phi}$ par

$$\widehat{\mathrm{var}}\,\widehat{\phi} = \frac{1}{\mathrm{X}} + \frac{1}{n-\mathrm{X}}.$$

Exercice 7 Odds Ratio

L'odds ratio est une mesure usuelle du risque associé à un facteur de risque. Il est défini par

$$\mathrm{OR} = \frac{\mathbb{P}(\text{att}|\text{exp})/\mathbb{P}(\text{non att}|\text{exp})}{\mathbb{P}(\text{att}|\text{non exp})/\mathbb{P}(\text{non att}|\text{non exp})}$$

où « att » est l'abréviation d'« atteint » et « exp » est l'abréviation d'« exposé (au facteur de risque) ».

1. Montrer qu'on a

$$\mathrm{OR} = \frac{\mathbb{P}(\text{exp}|\text{att})/\mathbb{P}(\text{non exp}|\text{att})}{\mathbb{P}(\text{exp}|\text{non att})/\mathbb{P}(\text{non exp}|\text{non att})}.$$

2. Dans une enquête cas-témoins, on recrute n cas et m témoins, et on compte ceux d'entre eux ont été exposés. On obtient une table de données comme celle-ci :

	Exposés	Non exposés	Total
Cas	A	B	n
Témoins	C	D	m

En déduire un estimateur classique $\widehat{\mathrm{OR}}$ de l'OR.

3. Montrer que la variance de $\log\widehat{\mathrm{OR}}$ est égale à

$$\mathrm{var}\left(\log\left(\frac{\mathrm{A}}{\mathrm{B}}\right)\right) + \mathrm{var}\left(\log\left(\frac{\mathrm{C}}{\mathrm{D}}\right)\right).$$

En déduire (on utilisera l'exercice précédent) que la variance de $\log \mathrm{OR}$ est estimée par

$$\frac{1}{\mathrm{A}} + \frac{1}{\mathrm{B}} + \frac{1}{\mathrm{C}} + \frac{1}{\mathrm{D}}.$$

4. En supposant que la distribution de $\log\widehat{\mathrm{OR}}$ est approximativement normale, donner un intervalle de confiance approximatif à 95 % pour l'OR avec les observations suivantes :

	Exposés	Non exposés	Total
Cas	60	40	100
Témoins	45	55	100

5. Avec les mêmes données et en utilisant la variable OR, tester l'hypothèse H_0 : OR = 1 contre H_1 : OR $\neq 1$ (test bilatéral). Donner un degré de signification (p-valeur).

6. Comparer avec les résultats de la méthode de test usuelle pour l'effet du facteur (test du χ^2).

Chapitre 8

Estimations

8.1 Introduction

Le problème de l'estimation est le suivant : on observe un phénomène aléatoire de façon répétée, et on cherche à « estimer » certaines caractéristiques de sa loi : sa moyenne, sa variance, sa médiane, voire la fonction de répartition ou la densité de sa loi.

Le problème des tests d'hypothèse est relié au problème de l'estimation : on cherche à décider si les observations sont compatibles avec une hypothèse sur les caractéristiques de la loi : moyenne nulle, variance égale à une valeur fixée, loi normale, etc.

Pour modéliser n observations, on considèrera donc n variables aléatoires $X_1,\dots,X_n$ indépendantes de même loi $\mathscr{L}$. On appellera ces variables aléatoires un *échantillon* de taille n de la loi $\mathscr{L}$.

On distinguera les variables aléatoires $X_1,\dots,X_n$ des valeurs $x_1,\dots,x_n$ qui auront pu être observées lors d'une collecte de données (sondage, étude clinique, etc).

8.2 Premiers estimateurs

8.2.1 La moyenne empirique

Nous faisons ici le bilan de résultats qui ont déjà été énoncés dans les chapitres précédents.

Définition 30 *La* moyenne empirique *de l'échantillon* $X_1,\dots,X_n$ *est la variable aléatoire*

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

Ceci peut surprendre : quand on dispose d'un échantillon numérique $x_1,\dots,x_n$, la moyenne $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est un nombre, pas une variable aléatoire ! Cependant si on répète plusieurs fois l'expérience aléatoire consistant à collecter n valeurs $x_1,\dots,x_n$, on obtiendra une nouvelle valeur de $\overline{x}$ à chaque fois. C'est cette possibilité qui impose de considérer $\overline{X}_n$ comme une variable aléatoire, dont les différentes valeurs $\overline{x}$ calculés à chaque expérience sont des *instances*.

Il est donc important de s'intéressant à l'espérance et à la variance de $\overline{X}_n$. Si μ et σ^2 sont l'espérance et la variance de la loi $\mathscr{L}$ (la loi des X_i), l'espérance de $\overline{X}_n$ est

$$E\left(\overline{X}_n\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

De même, les X_i étant indépendantes, la variance de $\overline{X}_n$ est

$$\mathrm{var}\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n}\sigma^2.$$

Récapitulons :

Proposition 32 *Soient* μ *et* σ^2 *l'espérance et la variance de* $\mathscr{L}$. *L'espérance et la variance de* $\overline{X}_n$ *sont*

$$\boxed{E\left(\overline{X}_n\right) = \mu, \quad \text{var}\left(\overline{X}_n\right) = \frac{1}{n}\sigma^2}$$

Le résultat suivant est une conséquence du théorème de la limite centrale.

Proposition 33 *Quand n est assez grand la loi de* $\overline{X}_n$ *peut être approchée par la loi normale* $\mathscr{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

8.2.2 La variance empirique

Définition 31 *La* variance empirique non corrigée *de l'échantillon* $X_1, \ldots, X_n$ *est la variable aléatoire*

$$\boxed{\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.}$$

On se souvient de la formule $\text{var}(X) = E\left((X - E(X)^2\right) = E\left(X^2\right) - E(X)^2$. Une formule analogue est vraie pour la variance empirique non corrigée :

$$\begin{aligned}
\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 &= \sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right) \\
&= \sum_{i=1}^{n} X_i^2 - 2\underbrace{\left(\sum_{i=1}^{n} X_i\right)}_{n\overline{X}}\overline{X} + n\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.
\end{aligned}$$

Il suffit de diviser par n cette identité pour obtenir la *formule décentrée* suivante :

Proposition 34 *On a :*

$$\boxed{\begin{aligned}
\tilde{S}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2 \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right)
\end{aligned}}$$

Calculons l'espérance de $\tilde{S}^2$. Notons tout d'abord que

$$E\left(\left(\overline{X}\right)^2\right) = E\left(\overline{X}\right)^2 + \text{var}\left(\overline{X}\right) = \mu^2 + \frac{1}{n}\sigma^2$$

$$E\left(X_i^2\right) = E(X_i)^2 + \text{var}(X_i) = \mu^2 + \sigma^2$$

On utilise la formule décentrée :

$$\begin{aligned}
E\left(\tilde{S}^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E\left(\left(\overline{X}\right)^2\right) \\
&= \mu^2 + \sigma^2 - \left(\mu^2 + \frac{1}{n}\sigma^2\right) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Proposition 35 *L'espérance de* $\tilde{S}^2$ *est* $\frac{n-1}{n}\sigma^2$.

Ceci motive la définition suivante :

Définition 32 *La* variance empirique (corrigée) *de l'échantillon* $X_1,\dots,X_n$ *est la variable aléatoire*

$$\begin{aligned} S^2 &= \frac{n}{n-1}\tilde{S}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2\right) \end{aligned}$$

Proposition 36 *L'espérance de la variance empirique corrigée est* $E(S^2)=\sigma^2$.

Préciser sa variance est plus compliqué, voir plus bas.

8.3 Le modèle normal

Dans cette section, on suppose que les X_i sont indépendants, de loi $\mathcal{N}(\mu,\sigma^2)$. Dans ce cas on peut préciser la loi de $\overline{X}$ et S^2.

8.3.1 Loi de la moyenne empirique

La somme de variables indépendantes de loi normale suivant une loi normale, on voit facilement que $X_1+\cdots+X_n$ suit une loi $\mathcal{N}(n\mu,n\sigma^2)$, d'où $\overline{X}\sim\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$.

Proposition 37 *Si* $X_1,\dots,X_n$ *sont indépendantes de même loi* $\mathcal{N}(\mu,\sigma^2)$*, alors* $\overline{X}=\frac{1}{n}\sum_i X_i$ *est de loi* $\mathcal{N}\left(\mu,\frac{\sigma^2}{n}\right)$.

8.3.2 Loi de la variance empirique

Voyons maintenant quelle est la loi de $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.

Théorème 38 *Soient* $X_1,\dots,X_n$ *des variables indépendantes de loi* $\mathcal{N}(\mu,\sigma^2)$*. Alors* $\dfrac{n-1}{\sigma^2}S^2$ *suit une loi* $\chi^2(n-1)$*; on pourra utiliser la notation* $S^2\sim\frac{\sigma^2}{n-1}\chi^2(n-1)$*. En particulier, la variance de* S^2 *est* $\dfrac{2}{(n-1)}\sigma^4$.
En outre, $\overline{X}$ *et* S^2 *sont indépendantes.*

Le dernier point de ce théorème est très important, nous y reviendrons plus tard. Il n'est pas vrai quand la loi n'est pas normale ! On peut esquisser sa preuve : la covariance de $\overline{X}$ et de $X_i-\overline{X}$ est

$$\operatorname{cov}\left(\overline{X},X_i-\overline{X}\right)=\operatorname{cov}\left(\overline{X},X_i\right)-\operatorname{var}\left(\overline{X}\right)=\frac{1}{n}\sigma^2-\frac{1}{n}\sigma^2=0.$$

On en déduit que $\overline{X}$ est indépendant de chacun des $X_i-\overline{X}$, et à fortiori de la somme de leurs carrés.

8.4 Qualité d'un estimateur

Définition 33 *Soient* $X_1,\dots,X_n$ *des variables indépendantes de loi* $\mathscr{L}$*, et soit* θ *un paramètre (inconnu) de cette loi. Soit* $T = t(X_1,\dots,X_n)$ *un estimateur de* θ *(*T *est une variable aléatoire).*

On appelle biais *de* T *la valeur* $\text{biais}(T) = E(T-\theta) = E(T)-\theta$*. Si* $\text{biais}(T) = 0$*, on dit que* T *est sans biais.*

La variance de l'estimateur est bien sûr la variance de T.

L'erreur quadratique moyenne de l'estimateur est $\text{eqm}(T) = E\left((T-\theta)^2\right)$.

On peut décomposer l'erreur quadratique moyenne comme somme du carré du biais et de la variance :

$$\boxed{\text{eqm}(T) = \text{biais}(T)^2 + \text{var}(T).}$$

Démontrons ce dernier résultat. Dans le calcul ci-dessous, on note $\theta_e = E(T)$, de sorte que $\text{biais}(T) = \theta_e - \theta$.

$$\begin{aligned}
\text{eqm}(T) &= E\left((T-\theta)^2\right)\\
&= E\left(((T-\theta_e)+(\theta_e-\theta))^2\right)\\
&= E\left((T-\theta_e)^2 + 2(T-\theta_e)(\theta_e-\theta) + (\theta_e-\theta)^2\right)\\
&= E\left((T-\theta_e)^2\right) + 2E(T-\theta_e)\cdot(\theta_e-\theta) + E\left((\theta_e-\theta)^2\right)\\
&= \text{var}(T) + 2\left(E(T)-\theta_e\right)(\theta_e-\theta) + (\theta_e-\theta)^2\\
&= \text{var}(T) + \text{biais}(T)^2
\end{aligned}$$

En pratique, on privilégie souvent les estimateurs sans biais. Cependant de nombreux statisticiens préfèrent avoir l'écart quadratique moyen le plus faible possible, le but étant de minimiser l'écart entre la valeur observée et la valeur réelle (inconnue).

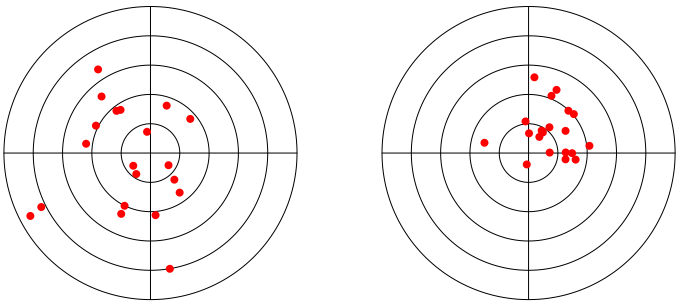


Figure 55. On compare souvent un estimateur avec un tireur. Vaut-il mieux un estimateur non biaisé mais de variance importante (carton de gauche) ou un un estimateur biaisé mais de faible variance (carton de droite) ?

Exemple 40 Avec $t(X_1,\dots,X_n) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n}$, T est la moyenne empirique, qui est estimateur de l'espérance de $\mathscr{L}$. D'après la proposition 32, cet estimateur est sans biais. □

Exemple 41 Le biais de $\tilde{S}^2$, la variance empirique non-corrigée, est $-\frac{1}{n}\sigma^2$. La variance empirique corrigée est sans biais. □

Exemple 42 Si les X_i suivent une loi normale $\mathscr{N}(\mu,\sigma^2)$, la variance de S^2 est $\dfrac{2}{(n-1)}\sigma^4$, et celle de $\tilde{S}^2$ est $\dfrac{2(n-1)}{n^2}\sigma^4$.

L'erreur quadratique moyenne de S^2 est égale à sa variance (le biais étant nul) :

$$\text{eqm}\left(S^2\right) = \frac{2}{(n-1)}\sigma^4,$$

et celle de $\tilde{S}^2$ est

$$\begin{aligned}\text{eqm}\left(\tilde{S}^2\right) &= \left(\frac{1}{n}\sigma^2\right)^2 + \frac{2(n-1)}{n^2}\sigma^4 \\ &= \frac{2n-1}{n^2}\sigma^4.\end{aligned}$$

On constate que $\text{eqm}\left(\tilde{S}^2\right) < \text{eqm}\left(S^2\right)$. Si on décide de choisir l'estimateur avec la plus petit erreur quadratique moyenne, il faut choisir $\tilde{S}^2$. La différence est cependant minime...

Remarquons finalement que le biais de $\tilde{S}^2$ tend vers 0 quand n augmente (il est *asymptotiquement sans biais*). □

8.5 Exemple : estimation d'une proportion

C'est le problème classique du sondage, ou de l'estimation de la proportion p d'une population répondant à un certain critère. On tire donc au hasard n individus dans la population (n étant fixé à l'avance), et on pose X égal au nombre d'individus répondant au critère. X est une variable aléatoire.

En toute rigueur, X suit une loi hypergéométrique (tirage sans remise), mais si la population considérée est grande, on peut approcher cette loi par la loi binomiale $\mathscr{B}in(n,p)$, où p est la proportion inconnue.

On l'estime par $\widehat{p} = \dfrac{X}{n}$. La loi de l'*estimateur* $\widehat{p}$ se déduit de celle de X. Il s'agit bien d'une variable aléatoire (si on répète le sondage plusieurs fois, on obtiendra des résultats différents) !

D'après le théorème de la limite centrale, quand n est assez grand la loi de X peut être approchée par une loi normale $\mathscr{N}(\mu = np, \sigma^2 = np(1-p))$; la loi de $\widehat{p}$ peut donc être approchée par une loi normale :

$$\boxed{\widehat{p} \sim \mathscr{N}\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right).}$$

On pouvait également arriver à cette conclusion en considérant $\widehat{p}$ comme la moyenne de n variables de Bernoulli indépendantes de loi $\mathscr{B}(p)$ et en appliquant le théorème de la limite centrée.

8.5.1 Transformation angulaire

Nous décrivons ici un artifice classique des « biostatistiques » : travailler sur $\arcsin\sqrt{\widehat{p}}$ au lieu de $\widehat{p}$.

On applique la méthode Delta pour obtenir les propriétés de cette transformation. Posons $\phi(x) = \arcsin\left(\sqrt{x}\right)$. La dérivée de ϕ est $\phi'(x) = \frac{1}{2\sqrt{x(1-x)}}$.

D'après le théorème 31, pour n assez grand, la loi de $\phi\left(\widehat{p}\right)$ est approximativement normale, d'espérance $\phi(p) = \arcsin\left(\sqrt{p}\right)$ et de variance

$$\left(\phi'(p)\right)^2 \frac{p(1-p)}{n} = \left(\frac{1}{2\sqrt{p(1-p)}}\right)^2 \frac{p(1-p)}{n} = \frac{1}{4n}.$$

Récapitulons :

$$\boxed{\arcsin\left(\sqrt{\widehat{p}}\right) \sim \mathscr{N}\left(\mu = \arcsin\left(\sqrt{p}\right), \sigma^2 = \frac{1}{4n}\right).}$$

L'intérêt premier de cette transformation est que la variance de la loi de $\arcsin\left(\sqrt{\widehat{p}}\right)$ est connue, ne dépendant plus du paramètre inconnu p. D'autre part, cette transformation améliore la qualité de l'approximation par la loi normale.

Cette méthode s'appelle *stabilisation de la variance*, elle peut être appliquée de façon générale à des estimateurs dont la variance dépend du paramètre à estimer.

8.6 Intervalle de confiance

On s'intéresse toujours à un paramètre θ d'une loi inconnue $\mathscr{L}$. Plutôt que d'en donner une estimation ponctuelle $\widehat{\theta}$, on peut préférer donner un ensemble de valeurs plausibles pour θ, sous la forme d'un intervalle.

Définition 34 *Soient* $X_1,\dots,X_n$ *des variables indépendantes de loi* $\mathscr{L}$*, et soit* θ *un paramètre de cette loi. Soient* T_1 *et* T_2 *qui se calculent à partir des* X_i *:* $T_1 = t_1(X_1,\dots,X_n)$ *et* $T_2 = t_2(X_1,\dots,X_n)$ *; on dit que* $[T_1,T_2]$ *est un* intervalle de confiance de niveau $\gamma = 1-\alpha$ *(ou au risque* α*) pour* θ *si on a*

$$\mathbb{P}(T_1 \leqslant \theta \leqslant T_2) \geqslant \gamma.$$

On prendra typiquement comme valeur pour γ*,* $\gamma = 0{,}95$ *(ou* 95%*).*

Remarque : Dans la définition ci-dessus aurait pu écrire simplement $\mathbb{P}(T_1 \leqslant \theta \leqslant T_2) = \gamma$; cependant dans le cas des lois discrètes il n'est pas toujours possible d'avoir exactement la valeur γ voulue ; on doit alors imposer d'avoir une probabilité égale *au moins à* γ.

On prendra typiquement $\gamma = 0{,}95$, et donc $\alpha = 0{,}05$. Dans certains cas on pourra prendre $T_1 = -\infty$, ou $T_2 = +\infty$, on parlera alors d'intervalle de confiance unilatéral.

Exemple 43 Supposons que $\mathscr{L}$ est une loi normale $\mathscr{N}(\mu,\sigma^2)$: on connaît sa variance σ^2 mais pas sa moyenne, qu'on cherche à estimer. Rappelons que le quantile d'ordre 0,975 de $\mathscr{N}(0,1)$ est $z_{0,975} = 1{,}96$, et que si $Z \sim \mathscr{N}(0,1)$ on a

$$\mathbb{P}(-1{,}96 < Z < 1{,}96) = 0{,}95.$$

L'estimateur naturel de μ est $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$; on sait qu'il suit une loi $\mathscr{N}\left(\mu, \frac{\sigma^2}{n}\right)$, et donc que

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathscr{N}(0,1).$$

On a donc

$$\mathbb{P}\left(-1{,}96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1{,}96\right) = 0{,}95,$$

puis

$$\mathbb{P}\left(-1{,}96\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < 1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95,$$

$$\mathbb{P}\left(-\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95,$$

$$\mathbb{P}\left(\overline{X}+1{,}96\frac{\sigma}{\sqrt{n}} > \mu > \overline{X}-1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95,$$

$$\mathbb{P}\left(\overline{X}-1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+1{,}96\frac{\sigma}{\sqrt{n}}\right) = 0{,}95.$$

L'intervalle étant symétrique par rapport à l'estimateur de μ, on parle d'« intervalle de confiance symétrique ». □

Exemple 44 On reprend l'exemple précédent. On lit dans une table de loi normale que si $Z \sim \mathcal{N}(0,1)$ alors $\mathbb{P}(Z < 1{,}64) = 0{,}95$. On a donc

$$\mathbb{P}\left(\overline{X} < \mu + 1{,}64\frac{\sigma}{\sqrt{n}}\right) = 0{,}95,$$

et

$$\mathbb{P}\left(\mu > \overline{X} - 1{,}64\frac{\sigma}{\sqrt{n}}\right) = 0{,}95.$$

C'est un exemple d'intervalle de confiance unilatéral. □

8.7 Exercices

Exercice 1

On a interrogé 900 personnes sur leurs intentions de votes ; 9 d'entre elles vont donner leur suffrage au candidat du Mouvement pour la Loi Normale (MLN).

1. On note p la proportion d'individus dans la population qui voteront MLN, et $\widehat{p} = 0{,}01$ l'estimation faite dans cet échantillon. Quelle est l'écart-type (approximatif) σ de $\arcsin\sqrt{\widehat{p}}$?

2. Montrer qu'avec une probabilité (approximativement) égale à 0,95, on a

$$\arcsin\sqrt{p} - 1{,}96\sigma \leqslant \arcsin\sqrt{\widehat{p}} \leqslant \arcsin\sqrt{p} + 1{,}96\sigma$$

3. En déduire un intervalle de confiance à 95% pour $\arcsin\sqrt{p}$, puis un intervalle de confiance pour p (attention à bien calculer en radians et non en degrés).

Exercice 2

On admettra le résultat suivant : si $X \sim \mathcal{P}(\lambda)$, quand λ est grand la loi de $U = 2\sqrt{X}$ est approximativement normale : $U \sim \mathcal{N}(2\sqrt{\lambda}, 1)$.

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes, de loi $\mathcal{P}(\lambda)$. On pose $S = X_1 + \cdots + X_n$ et $\overline{X} = \frac{1}{n}S$.

1. Quelle est l'espérance de $\overline{X} = \frac{1}{n}S$? Quelle est sa variance ?

2. Quelle est la loi de S ?

3. Montrer que

$$\mathbb{P}\left(2\sqrt{S} - 1{,}96 \leqslant 2\sqrt{n\lambda} \leqslant 2\sqrt{S} + 1{,}96\right) \simeq 0{,}95.$$

4. En déduire un intervalle de confiance à 95% pour λ.

5. Application : on observe $n = 10$ valeurs indépendantes, notées $x_1, \ldots, x_{10}$, tirées selon une loi de Poisson de paramètre λ inconnu. On donne $s = \sum_{i=1}^{10} x_i = 305$. Donnez un intervalle de confiance à 95% pour la valeur de λ.

Chapitre 9

Intervalles de confiance usuels

Nous rappelons ici quelques procédures classiques d'intervalles de confiance.

9.1 Intervalle de confiance sur une moyenne

On doit définir tout d'abord la loi t de Student.

Définition 35 *Soient* $Z \sim \mathcal{N}(0,1)$ *et* $Y \sim \chi^2(d)$*, indépendantes. La variable aléatoire* $X = \dfrac{Z}{\sqrt{\frac{Y}{d}}}$ *suit une loi continue à densité, appelée loi* t *de Student à* d *degrés de liberté, notée* $t(d)$.

Pour $d = 1$, la loi $t(d)$ n'a ni espérance, ni variance. Pour $d = 2$, son espérance est nulle, et elle n'a pas de variance. Pour $d \geqslant 3$, son espérance est nulle et sa variance vaut $\frac{d}{d-2}$.
La densité de la loi t évoque une normale « aplatie ». Quand d tend vers l'infini, la loi $t(d)$ s'approche de la loi normale standard.

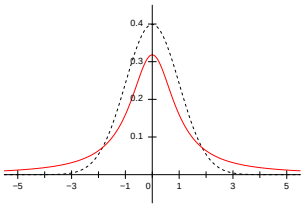


Figure 56. Densité de la loi $t(1)$ (loi normale en pointillés)

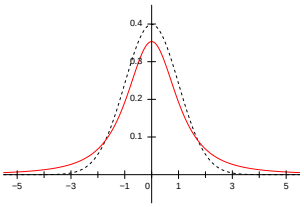


Figure 57. Densité de la loi $t(2)$

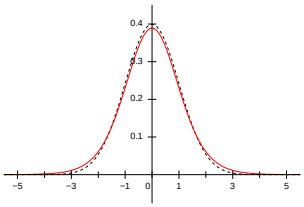


Figure 58. Densité de la loi $t(10)$

La définition de cette loi a été motivée par la proposition suivante.

Proposition 39 *Soient* $X_1, \ldots, X_n$ *des variables indépendantes de loi* $\mathcal{N}(\mu, \sigma^2)$*. On note* $\overline{X}$ *la moyenne empirique des* X_i*,* S^2 *leur variance empirique corrigée, et* S *la racine carrée de* S^2*. On pose*

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}}.$$

Alors $T \sim t(n-1)$.

⚠ Cette proposition est très importante. L'espérance de $\overline{X}$ est μ, et $\sqrt{\frac{S^2}{n}}$ est une estimation de sa variance : la variable T définie ci-dessus est donc une variable normale centrée (par sa moyenne, supposée connue) et réduite *par l'estimation de sa variance.*

En effet, posons $Z = \dfrac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0,1)$, et $Y = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$. On a déjà mentionné (théorème 38) que Y et Z sont indépendantes.

On a donc $\dfrac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t(n-1)$, et comme $\dfrac{Y}{n-1} = \dfrac{\frac{n-1}{\sigma^2} S^2}{n-1} = \dfrac{S^2}{\sigma^2}$ on a

$$\frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n} \times \frac{Y}{n-1}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n}}} = T.$$

9.1.1 Construction de l'intervalle de confiance

Soient $X_1, \ldots, X_n$ des variables indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On a vu dans l'exemple 43 comment construire un intervalle de confiance pour μ, quand σ^2 est connu.

Si on ne connaît pas σ^2 (ce paramètre est alors appelé *paramètre de nuisance* car sa présence est importune), les résultats précédents nous permettent de procéder de la même façon. On pose T comme dans la proposition ci-desssus ; $T \sim t(n-1)$.

Soit $t^{n-1}_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ de la loi $t(n-1)$. On a

$$\mathbb{P}\left(-t^{n-1}_{1-\alpha/2} < T < t^{n-1}_{1-\alpha/2}\right) = 1-\alpha,$$

et on en déduit la formule :

$$\boxed{\mathbb{P}\left(\overline{X} - t^{n-1}_{1-\alpha/2}\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t^{n-1}_{1-\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha.}$$

Pour la valeur classique $\alpha = 0{,}05$, on a $z_{0{,}975} = 1{,}96$.

9.1.2 Cas des grands échantillons et loi non-normale

Dans le cas normal, quand n est assez grand (disons, si $n \geqslant 30$), on peut approcher la loi $t(n-1)$ par une loi normale $\mathcal{N}(0,1)$, et $t^{n-1}_{1-\alpha/2}$ devient proche de $z_{1-\alpha/2}$.

Dans le cas général, si la loi $\mathscr{L}$ des variables $X_1, \ldots, X_n$ n'est pas normale, le théorème de la limite centrée assure que $\overline{X}$ est approximativement normale. Cependant, en général il n'y aucune raison que $\frac{n-1}{\sigma^2} S^2$ suive une loi de χ^2. L'utilisation des quantiles de la loi t est difficile à justifier dans ce cas. On se repose alors sur l'intervalle de confiance donné à l'exemple 43 :

$$\mathbb{P}\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right) \simeq 1-\alpha.$$

Comme on ne connaît pas σ, on le remplace par son estimation S, et on obtient un intervalle de confiance « asymptotique » :

$$\boxed{\mathbb{P}\left(\overline{X} - 1{,}96\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + 1{,}96\sqrt{\frac{S^2}{n}}\right) \simeq 1-\alpha.}$$

Cet intervalle de confiance n'est qu'approché. Plus n est grand, meilleure est l'approximation.

9.2 Intervalle de confiance sur une différence de moyennes

Nous allons nous restreindre au cas où on a deux échantillons indépendants issus de deux lois **de même variance :** un premier échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1,\sigma^2)$ et un second de taille n_2 issu d'une loi $\mathcal{N}(\mu_2,\sigma^2)$.

La moyenne empirique du premier échantillon est $\overline{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$, et celle du second échantillon $\overline{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2}{n_2}\right)$. On a donc

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1-\mu_2, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\sigma^2\right).$$

On va procéder comme précédemment, à une nuance près : le calcul de l'estimation de la variance. On dispose de deux estimations indépendantes : dans le premier échantillon, on calcule S_1^2 et on a $\frac{n_1-1}{\sigma^2}S_1^2 \sim \chi^2(n_1-1)$; dans le second échantillon, on calcule S_2^2 et on a $\frac{n_2-1}{\sigma^2}S_2^2 \sim \chi^2(n_2-1)$. On a donc (grâce à la proposition 25)

$$\frac{n_1-1}{\sigma^2}S_1^2 + \frac{n_2-1}{\sigma^2}S_2^2 = \frac{1}{\sigma^2}\left((n_1-1)S_1^2+(n_2-1)S_2^2\right) \sim \chi^2(n_1+n_2-2).$$

Cette dernière expression a pour espérance (n_1+n_2-2). Pour obtenir une estimation de σ^2 on prendra donc

$$S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

On peut maintenant centrer et réduire $\overline{X}_1-\overline{X}_2$ à l'aide de l'estimation de la variance, pour obtenir une loi t :

$$\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2).$$

On en tire l'intervalle de confiance classique :

$$\mathbb{P}\left((\overline{X}_1-\overline{X}_2) - t_{1-\alpha/2}^{n_1+n_2-2}S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < (\mu_1-\mu_2) < (\overline{X}_1-\overline{X}_2) + t_{1-\alpha/2}^{n_1+n_2-2}S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = 1-\alpha,$$

où $t_{1-\alpha/2}^d$ est le quantile d'ordre $1-\frac{1}{2}\alpha$ de la loi $t(d)$.

9.3 Intervalle de confiance sur la variance d'une loi normale

Avant toute chose, avertissons : ce qui suit n'est valable que pour une loi $\mathscr{L}$ normale. Si on supprime cette hypothèse, ces intervalles de confiance ne sont plus valables, même avec de grands échantillons. En effet $\frac{n-1}{\sigma^2}S^2$ ne suit pas en général une loi de χ^2 à $n-1$ degré de liberté, même si n est grand !

Soient $X_1,\ldots,X_n$ des variables indépendantes de loi $\mathcal{N}(\mu,\sigma^2)$.

On a

$$Y = \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1).$$

Notons x_a^{n-1} le quantile d'ordre a de la loi $\chi^2(n-1)$. On a

$$\mathbb{P}\left(x_{\alpha/2}^{n-1} < Y < x_{1-\alpha/2}^{n-1}\right) = 1-\alpha,$$

d'où on tire

$$\mathbb{P}\left(\frac{(n-1)S^2}{x_{1-\alpha/2}^{n-1}} < \sigma^2 < \frac{(n-1)S^2}{x_{\alpha/2}^{n-1}}\right) = 1-\alpha.$$

9.4 Rapport de variances de lois normales

On doit définir tout d'abord la loi F de Fisher-Snedecor.

Définition 36 *Soient* $Y_1 \sim \chi^2(d_1)$ *et* $Y_2 \sim \chi^2(d_2)$ *indépendantes. La variable aléatoire*

$$X = \frac{Y_1/d_1}{Y_2/d_2} = \frac{d_2 Y_1}{d_1 Y_2}$$

suit une loi continue à densité, appelée loi F *de Fisher-Snedecor à* d_1 *et* d_2 *degrés de liberté, notée* $F(d_1, d_2)$.

Notons que si $X \sim F(d_1, d_2)$ alors $\frac{1}{X}$ suit une loi $F(d_2, d_1)$. Si on note $F_\alpha^{d_1,d_2}$ le quantile d'ordre α de $F(d_1, d_2)$, il s'ensuit qu'on a également $F_\alpha^{d_1,d_2} = \frac{1}{F_{1-\alpha}^{d_2,d_1}}$.

L'espérance de la loi F n'existe que si $d_2 \geqslant 3$, elle vaut alors $\frac{d_2}{d_2-2}$. Sa variance n'existe que si $d_2 \geqslant 5$, elle vaut alors $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$.

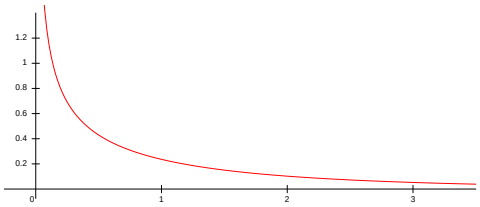


Figure 59. Densité de F(1,20)

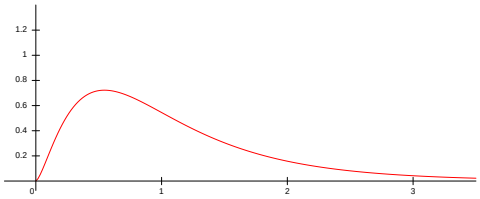


Figure 60. Densité de F(5,20)

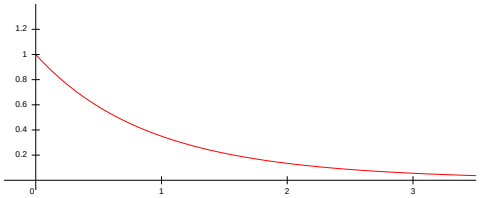


Figure 61. Densité de F(2,20)

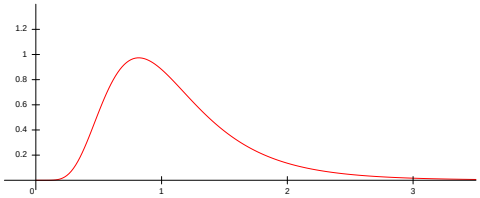


Figure 62. Densité de F(20,20)

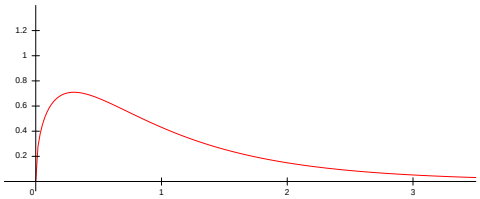


Figure 63. Densité de F(3,20)

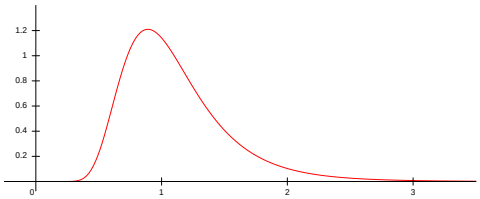


Figure 64. Densité de F(100,20)

La motivation de la définition de la loi F est dans les deux propositions suivantes.

Proposition 40 *On considère un premier échantillon de taille* n_1 *issu d'une loi* $\mathcal{N}(\mu_1, \sigma^2)$ *et un second de taille* n_2 *issu d'une loi* $\mathcal{N}(\mu_2, \sigma^2)$ *(moyennes éventuellement différentes, même variance).*
La variance empirique du premier échantillon est S_1^2*, celle du second est* S_2^2*. Alors*

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1).$$

En effet, on a $U = \frac{n_1-1}{\sigma^2}S_1^2 \sim \chi^2(n_1-1)$ et $V = \frac{n_2-1}{\sigma^2}S_2^2 \sim \chi^2(n_2-1)$, d'où $\frac{(n_2-1)U}{(n_1-1)V} \sim F(n_1-1, n_2-1)$, or $\frac{(n_2-1)U}{(n_1-1)V} = \frac{S_1^2}{S_2^2}$. On montre tout aussi aisément la proposition suivante.

Proposition 41 *On considère un premier échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et un second de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma_2^2)$ (moyennes et variances éventuellement différentes).*
La variance empirique du premier échantillon est S_1^2, celle du second est S_2^2. Alors

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

En reprenant les notations de la proposition qui précède, on a

$$\mathbb{P}(F_{\alpha/2}^{n_1-1,n_2-1} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{1-\alpha/2}^{n_1-1,n_2-1}) = 1-\alpha,$$

d'où l'intervalle de confiance au niveau $1-\alpha$ sur $\frac{\sigma_2^2}{\sigma_1^2}$:

$$\boxed{\mathbb{P}\left(\frac{S_2^2}{S_1^2}F_{\alpha/2}^{n_1-1,n_2-1} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2}F_{1-\alpha/2}^{n_1-1,n_2-1}\right) = 1-\alpha.}$$

9.5 Intervalle de confiance sur une proportion

C'est le cas classique du sondage. Soient $X_1, \ldots, X_n$ des variables indépendantes de loi $\mathcal{B}(p)$. On cherche à estimer la valeur de p.
On pose $S_n = X_1 + \cdots + X_n$. L'estimateur naturel de p est $\widehat{p} = \frac{S_n}{n}$. On a $S_n \sim \mathcal{B}in(n,p)$, et, si $np > 5$ et $n(1-p) > 5$ on peut approcher la loi de S_n par la loi normale $\mathcal{N}(\mu = np, \sigma^2 = np(1-p))$. On approche donc la loi de $\widehat{p}$ par $\mathcal{N}\left(\mu = p, \sigma^2 = \frac{p(1-p)}{n}\right)$.
On a donc

$$\mathbb{P}\left(z_{\alpha/2} < \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\alpha/2}\right) = 1-\alpha.$$

En tenant compte de $z_\alpha = -z_\alpha$, on a

$$\mathbb{P}\left(\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha.$$

Il reste des termes en $\sqrt{p(1-p)}$ dans les bornes, alors que p est ce qu'on cherche à exprimer ! On va simplement les remplacer par $\sqrt{\widehat{p}(1-\widehat{p})}$, ce qui n'est correct que pour les très grandes valeurs de n. On obtient l'intervalle de confiance classique

$$\boxed{\mathbb{P}\left(\widehat{p} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) = 1-\alpha.}$$

Notez la similarité avec l'intervalle de confiance sur une moyenne dans le cas non-normal, section 9.1.

Cet intervalle de confiance n'est pas de très bonne qualité, sauf pour les très grandes valeurs de n... il est cependant très largement utilisé.

L'usage veut qu'on vérifie a posteriori que np et $n(1-p)$ sont plus grands que 5 sur tout l'intervalle de confiance.

9.6 Intervalle de confiance sur la différence de deux proportions

On considère un échantillon de taille n_1 issu d'une loi $\mathscr{B}(p_1)$ et un de taille n_2 issu d'une loi $\mathscr{B}(p_2)$. On veut un intervalle de confiance sur $d = p_1 - p_2$.

On approche la loi de $\widehat{p}_1$ par $\mathscr{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$ et celle de $\widehat{p}_2$ par $\mathscr{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$. La loi de $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$ est donc approchée par

$$\mathscr{N}\left(\mu = p_1 - p_2 = d, \sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Comme précédemment on en tire l'intervalle de confiance

$$\mathbb{P}\left((\widehat{p}_1 - \widehat{p}_2) - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} < d < (\widehat{p}_1 - \widehat{p}_2) + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}\right) = 1-\alpha.$$

Ici encore, comparez avec la section 9.2.

9.7 Exercices

Exercice 1 En Syldavie ont lieu des élections au scrutin majoritaire au suffrage universel à deux tours. Deux candidats sont en lice pour le second tour : Alfred Afreu (candidat A) et Bénédicte Bouh (candidat B).

Les 2 millions d'électeurs Syldaves sont des ouvriers (40%), des paysans (40%) et des artisans (20%). Les ouvriers votent pour le candidat A à 20%, les paysans à 70%, et les artisans à 60%.

	Ouvriers	Paysans	Artisans
Part de la population	40%	40%	20%
Votes pour le candidat A	20%	70%	60%

Table sans numéro. Vote par catégorie sociale dans la population Syldave

Les instituts de sondage interrogent des électeurs au hasard afin de connaître leurs intentions de vote.

1. Toutes catégories confondues, quelle proportions de votes p_A doit obtenir le candidat A ?
2. L'institut Mentos procède en interrogeant 1000 électeurs au hasard. On note X le nombre d'entre eux qui déclarent voter pour le candidat A.
   (a) Quelle est la loi de X ? Calculez son espérance et sa variance.
   (b) Donnez un intervalle de pari à 95% pour l'estimation de p_A obtenue par cette méthode de sondage.
3. L'institut Ricqlès utilise la méthode des quotas : afin « d'améliorer la représentatitivé » de l'échantillon, sont interrogés 400 ouvriers, 400 paysans et 200 artisans. Donnez un intervalle de pari à 95% pour l'estimation de p_A obtenue par cette méthode de sondage.

Exercice 2 On considère un échantillon de $n = 20$ mesures $x_1, \ldots, x_{20}$ indépendantes, supposées suivre une loi normale $\mathscr{N}(\mu, \sigma^2)$ de moyenne μ et de variance σ^2 inconnues.

On a $\sum_i x_i = 22{,}4$ et $\sum_i x_i^2 = 80{,}9$.

1. Calculez des intervalles de confiance de niveaux 90% et 95% pour μ et σ^2.

2. On suppose la variance connue a priori : $\sigma^2 = 4$. Recalculez les intervalles de confiance pour μ.

Exercice 3 On considère deux échantillons de mesures :
- un premier échantillon avec $n_1 = 20$ mesures, suivant une loi $\mathcal{N}(\mu_1, \sigma_1^2)$; on donne $\sum x_i = 22{,}4$ et $\sum x_i^2 = 80{,}9$.
- un second échantillon avec $n_2 = 30$ mesures, suivant une loi $\mathcal{N}(\mu_2, \sigma_2^2)$; on donne $\sum x_i = 28{,}1$ et $\sum x_i^2 = 133{,}2$.

Calculez des intervalles de confiance au niveaux 90% et 95% pour $\frac{\sigma_1^2}{\sigma_2^2}$ et $\mu_1 - \mu_2$.

Exercice 4 On considère un échantillon de $n = 40$ mesures indépendantes. On donne $\sum x_i = 23$ et $\sum x_i^2 = 1882$.

1. Donner un intervalle de confiance au niveau 95% sur l'espérance de la loi commune aux mesures.

2. Même question en supposant que cette loi est normale.

Exercice 5

On pose $\Phi(p) = \arcsin(\sqrt{p})$ pour tout $p \in [0,1]$. On considère une proportion $\widehat{p} = \frac{x}{n}$ observée dans un échantillon de taille n. Un intervalle de confiance à 95% sur $\Phi(\widehat{p})$, *avec correction de continuité,* est donné par

$$\left[\Phi\left(\widehat{p} - \frac{1}{2n}\right) - 1{,}96\frac{1}{2\sqrt{n}}; \Phi\left(\widehat{p} + \frac{1}{2n}\right) + 1{,}96\frac{1}{2\sqrt{n}}\right]$$

Cet intervalle de confiance est utilisable dès que $n \geqslant 10$ indépendament de la valeur de $\widehat{p}$.[1]

Dans une étude de 1992, 159 prostituées de Glasgow ont accepté de donner un échantillon de salive qui a été utilisé pour tester leur séropositivité HIV. 4 échantillons étaient positifs.[2]

Estimer la prévalence du HIV dans cette population, et donner un intervalle de confiance.

[1] Pires AM, Amado C, Interval estimators for a binomial proportion : comparison of twenty methods. Revstat 6 (2008).

[2] McKeganey N et coll, Female streetworking prostitution ans HIV infection in Glasgow, BMJ 305 (1992).

Exercice 6 On considère une urne contenant n boules dont une seule boule noire, et $n-1$ boules blanches.

1. Lors d'un tirage dans cette urne, quelle est la probabilité p de tirer la boule noire ?

2. On réalise N tirages avec remise dans cette urne, et on compte X = le nombre de fois où on a tiré la boule noire. Quelle est la loi de X ?

3. Avec N = 5000 tirages, on a obtenu $x = 740$ fois la boule noire. Donnez un intervalle de confiance à 95% pour p.

4. Qu'en déduisez-vous pour n ?

Exercice 7 *La partie II peut être traitée indépendamment de la partie I. On peut traiter la partie III en admettant les résultats de la partie I.*

On considère $n+1$ variables aléatoires $X_1, \ldots, X_n, X_{n+1}$ indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$.

Partie I

1. Quelle est la loi de la moyenne empirique $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ des n premières variables ?

2. Quelle est la loi de $X_{n+1} - \overline{X}$?

3. On pose

$$Z = \frac{1}{\sigma}\sqrt{\frac{n}{n+1}}\left(X_{n+1} - \overline{X}\right).$$

Montrer que Z suit une loi normale centrée réduite.

4. On considère la variance empirique $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ des n premières variables. Quelle est la loi de $Y = \dfrac{n-1}{\sigma^2}S^2$?

5. On pose $T = \dfrac{Z}{\sqrt{\frac{Y}{n-1}}}$. Quelle est la loi de T ?

6. En déduire que si $n = 10$, on a

$$\mathbb{P}(-2{,}26 \leqslant T \leqslant 2{,}26) = 0{,}95.$$

Partie II

Le professeur Sokolodovenko a acheté un sachet de 10 pâtes de fruits à son confiseur. Il les pèse une par une, et obtient les masses x_1 à x_{10} suivantes (en grammes) : 19,3, 20,7, 21,5, 20,5, 20,6, 18,5, 21,2, 21,3, 20,7, 22,7. Pour faciliter les calculs, on donne $\sum x_i = 207{,}0$, et $\sum x_i^2 = 4297{,}00$.

On suppose que la masse d'une pâte de fruits prise au hasard chez le confiseur suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, les paramètres μ et σ^2 étant inconnus.

7. Calculer la moyenne empirique et la variance empirique des mesures du professeur, ainsi qu'une estimation de σ.

8. Donner un intervalle de confiance à 95 % sur la valeur de μ.

9. Donner un intervalle de confiance à 95 % sur la valeur de σ^2.

Partie III

Le professeur, qui a encore un peu faim, s'apprête a aller acheter une onzième pâte de fruits. Notons X_{11} la variable aléatoire définie par $X_{11} =$ masse de cette onzième pâte de fruits.

10. Le professeur, qui est un peu fantasque, désire avoir une idée de la masse de la pâte de fruits qu'il s'apprête à acheter. En utilisant le résultat de la question 6, montrer que

$$\mathbb{P}(17{,}95 \leqslant X_{11} \leqslant 23{,}45) = 0{,}95.$$

Chapitre 10

Tests d'hypothèses

10.1 Introduction

La situation typique dans laquelle on va se placer est la suivante : on dispose de données issues d'une expérience aléatoire (par exemple de mesures biologiques : la concentration d'un marqueur sanguin). On a vu qu'on sait estimer des paramètres caractéristiques de la loi sous-jacente aux variables mesurées, notament sa moyenne et sa variance. On va maintenant chercher à tester si ces estimations sont compatibles avec certaines hypothèses à priori sur les paramètres.

Dans les premières sections de ce chapitre, nous allons présenter les principales notions de la théorie des tests (hypothèse nulle et hypothèse alternative, risques). En fin de chapitre, en section 10.9, nous développerons un exemple : tester la nullité de la moyenne d'une loi normale de variance connue.

10.2 Le test comme procédure de décision

Au départ de la théorie des tests, il y a une expérience aléatoire (recueil de données), et une hypothèse H_0 qui est faite sur le modèle sous-jacent ; on veut tester cette hypothèse.

Ainsi par exemple, on observe $X_1, \ldots, X_n$ des variables indépendantes de même loi $\mathscr{L}$, et on fait une hypothèse sur $\mathscr{L}$. Le problème est souvent voisin de celui de l'estimation des paramètres de la loi $\mathscr{L}$: si θ est un paramètre (inconnu) de $\mathscr{L}$, on voudra tester une hypothèse de type $H_0 : \theta = \theta_0$.

On appelle H_0 « l'hypothèse nulle ». On est amené à formuler une « hypothèse alternative » H_1, qui (au moins dans les cas simples) sera très vague : cela sera par exemple $\theta \neq 0$. Dans certain cas, H_1 est à peine plus précise : ce sera par exemple $\theta > 0$, ce qui suffit à conduire à une procédure de test différente. Enfin, on peut avoir H_1 de la forme : $\theta = \theta_1$, ce qui est aussi précis que H_0, et change également la donne.

On fixe une règle de décision, qui permet pour toutes valeurs $x_1, \ldots, x_n$ prises par les variables aléatoires X_i, de trancher entre H_0 et H_1. En pratique, la règle est souvent de la forme suivante : on construit une statistique de test $T = t(X_1, \ldots, X_n)$ et on se fixe une règle sur les valeurs de T ; par exemple si T ne dépasse pas un certain seuil s, ou p lus généralement si T est dans un certain intervalle $[a, b]$, on tranche en faveur de l'hypothèse H_0, et sinon en faveur de H_1.

De façon plus générale, on définit une partie A de $\mathbb{R}$, et si $T \in A$ on tranche en faveur de H_0 ; si $T \in \overline{A}$ (le complémentaire de A) on tranche en faveur de H_1. Dans la suite, nous nous placerons dans ce cas de figure.

10.3 Erreurs et risques

Il y a deux façons de se tromper en faisant un test :

— *L'erreur de première espèce* consiste à trancher en faveur de H_1 à tort. Le *risque de première espèce* ou *risque* α est la probabilité de trancher pour H_1 alors que H_0 est vrai. C'est une probabilité conditionnelle :

$$\boxed{\alpha = \mathbb{P}(T \notin A | H_0)}$$

— *L'erreur de seconde espèce* consiste à trancher en faveur de H_0 à tort. Le *risque de seconde espèce* ou *risque* β est la probabilité de trancher pour H_0 alors que H_1 est vrai :

$$\boxed{\beta = \mathbb{P}(T \in A | H_1)}$$

On voudrait naturellement construire le test pour que ces risques soient les plus faibles possibles. Cependant, en général la loi de T sous l'hypothèse alternative H_1 n'est pas connue, et on ne peut pas calculer β. En revanche on connaît la loi de T sous H_0 et on peut calculer α. On est donc plus ou moins forcés de privilégier l'hypothèse nulle, de limiter le risque α, sans contrôle sur le risque β.

10.4 Asymétrie entre H_0 et H_1

L'asymétrie entre les deux hypothèses se traduit dans la conclusion : si le test tranche en faveur de H_0, on dira qu'« on accepte l'hypothèse nulle », si le test tranche en faveur de H_1, on dira qu'« on rejette l'hypothèse nulle ».

La partie $A \subset R$ qui définit le test (on accepte H_0 si $T \in A$, on rejette H_0 sinon) s'appelle la zone d'acceptation ; son complémentaire s'appelle la zone de rejet.

Il faut garder à l'esprit qu'on n'accepte H_0 que de façon provisoire ; il est possible par exemple que H_0 soit fausse mais que la taille de l'échantillon soit trop faible pour que le test soit efficace. Pour cette raison certains auteurs préfèrent la formulation « on ne rejette pas H_0 ».

Il convient cependant de remarquer que rejeter H_0 est en pratique tout aussi provisoire. Il ne manque pas de résultats publiés qui n'ont jamais pu être reproduits par la suite. Ainsi, en pratique, c'est comme dans toutes les sciences la *reproductibilité* des observations qui est la pierre de touche des résultats issus des études statistiques.

10.5 Degré de signification

On se place dans le contexte d'un test définit par une statistique de test T avec une règle de la forme « on rejette H_0 quand $T > s$ » (les tests usuels s'y ramènent sans peine, quitte à remplacer parfois la statistique de test par sa valeur absolue ou son carré...). Autrement dit, la zone d'acceptation A est $A = \{t \ : \ t \leqslant s\}$.

Le *degré de signification* d'un test, ou p-valeur (en anglais : *p-value*), est la probabilité que, H_0 étant supposée vraie, la statistique de test T prenne une valeur supérieure ou égale à la valeur observée $t = t(x_1, \ldots, x_n)$:

$$\boxed{p = \mathbb{P}(T > t)}$$

Un des intérêts du degré de signification est qu'on n'a pas besoin de calculer la valeur seuil s qui correspond à une valeur donnée pour le risque α du test : on peut choisir de façon équivalente de rejeter H_0 quand $p \leqslant \alpha$. D'autre part, le degré de signification a pris une (sans doute trop) grande importance dans l'esprit des utilisateurs des statistiques, et il est devenu à peu près impossible de ne pas donner p quand on fait un test ; plus p est petit, plus facilement vous emporterez la conviction de votre public.

10.6 Tests diagnostics : sensibilité, spécificité, etc.

Dans le contexte médical, où un test basé sur un ensemble de mesures biologiques est utilisé pour établir un diagnostic, on rencontrera souvent les notions de *spécificité* et *sensibilité*. Un tel test aura comme hypothèse nulle « l'individu est sain », et comme hypothèse alternative « l'individu est malade ». On dira que le test est positif si on a rejeté H_0 et négatif sinon.

La spécificité d'un test diagnostique est la proportion d'individus sains qui sont correctement diagnostiqués ; c'est donc la probabilité d'accepter H_0 quand H_0 est vraie ; c'est $\mathbb{P}(T \in A|H_0) = 1 - \mathbb{P}(T \notin A|H_0) = 1 - \alpha$. Ici A est la zone d'acceptation du test ; on note parfois $T \in A$ par T^- (test négatif) et $T \notin A$ par T^+ (test positif).

La sensibilité est la proportions d'individu atteints qui sont correctement diagnostiqués ; c'est la probabilité de rejeter H_0 quand H_1 est vraie ; c'est $\mathbb{P}(T \notin A|H_1) = 1 - \mathbb{P}(T \in A|H_1) = 1 - \beta$.

$$\boxed{\begin{aligned} \text{Spécifité} &= 1-\alpha \\ \text{Sensibilité} &= 1-\beta \end{aligned}}$$

En plaçant ainsi les tests statistiques dans le contexte médical, on comprend mieux les implications de l'habitude de se concentrer sur le seul risque α.

10.6.1 Valeurs prédictives

En pratique, il est également intéressant de connaître la *valeur prédictive positive* (VPP) et la *valeur prédictive négative* (VPN) d'un test, qui sont la probabilité d'être atteint (respectivement non atteint) si le test est positif (respectivement négatif).

En notant $\mathscr{P}$ la prévalence de la maladie dans la population testée, c'est-à-dire la probabilité d'être sous H_1 quand on effectue le test : $\mathscr{P} = \mathbb{P}(H_1)$, on a :

$$\begin{aligned} \text{VPP} &= \mathbb{P}(H_1|T \notin A) \\ &= \frac{\mathbb{P}(T \notin A|H_1)\mathbb{P}(H_1)}{\mathbb{P}(T \notin A)} \\ &= \frac{(1-\beta)\mathscr{P}}{\mathbb{P}(T \notin A|H_0)\mathbb{P}(H_0) + \mathbb{P}(T \notin A|H_1)\mathbb{P}(H_1)} \end{aligned}$$

$$\boxed{\text{VPP} = \frac{(1-\beta)\mathscr{P}}{\alpha(1-\mathscr{P}) + (1-\beta)\mathscr{P}}}$$

$$\begin{aligned} \text{VPN} &= \mathbb{P}(H_0|T \in A) \\ &= \frac{\mathbb{P}(T \in A|H_0)\mathbb{P}(H_0)}{\mathbb{P}(T \in A)} \\ &= \frac{(1-\alpha)(1-\mathscr{P})}{\mathbb{P}(T \in A|H_0)\mathbb{P}(H_0) + \mathbb{P}(T \in A|H_1)\mathbb{P}(H_1)} \end{aligned}$$

$$\boxed{\text{VPN} = \frac{(1-\alpha)(1-\mathscr{P})}{(1-\alpha)(1-\mathscr{P}) + \beta\mathscr{P}}}$$

On voit que les valeurs prédictives dépendent de la prévalence de la maladie. Il peut être nécessaire de distinguer la prévalence dans la population générale et de la prévalence dans la population à qui on fait passer le test (personnes demandant à être testées, ou à qui on propose un test sur la base d'indices divers).

Ici on a considéré la valeur prédictive globale du test, indépendamment de la valeur prise par la statistique de test T ; si on dispose de celle-ci, on peut calculer (toujours à l'aide de la prévalence) un risque a posteriori d'être atteint ou non (ou encore, d'être sous H_1 ou H_0).

10.6.2 Risque a posteriori

Une fois le test effectué l'idéal est de pouvoir assigner aux deux hypothèses rivales H_0 et H_1 des probabilités a posteriori. Quand la seule information disponible est de type $T \in A$ (ou T^-, test négatif) et $T \notin A$ (ou T^+, test positif), les valeurs prédictives positives et négatives remplissent ce rôle.

Mais quand on connaît la valeur observée de la statistique de test, on peut réaliser un calcul plus précis. Comme le calcul de la puissance (ou de la sensibilité), cela nécessite de connaître la loi des données sous H_1 ; et comme le calcul des valeurs prédictives, cela nécessite une information a priori sur $\mathbb{P}(H_0)$ et $\mathbb{P}(H_1)$ (ou prévalence).

Prenons un exemple relativement simple mais suffisament général pour illustrer la technique : une mesure biologique X suit une loi $\mathcal{N}(\mu_0, \sigma^2)$ sous H_0 (individu sain) et une loi $\mathcal{N}(\mu_1, \sigma^2)$ sous H_1.

On a observé chez un patient une valeur x de la variable aléatoire X. On aimerait pouvoir évaluer les probabilités conditionnelles $\mathbb{P}(H_0|X = x)$ et $\mathbb{P}(H_1|X = x)$; cependant X suivant une loi continue, l'événement $(X = x)$ est de probabilité nulle, et une telle probabilité conditionnelle n'a aucun sens !

Pour sortir de ce mauvais pas, nous allons « tricher » un peu en prenant en compte le fait que toutes nos mesures physiques sont « arrondies » : on ne mesure jamais $X = x$ mais en fait $X = x$ « à ε près », soit $X \in \left[x - \frac{\varepsilon}{2}; x + \frac{\varepsilon}{2}\right]$, que nous allons noter $X \approx x$. Comme nous allons le voir, la valeur de ε n'a pas d'importance, du moment que ε est petit ; dans la suite, ε est fixé.

Ceci permet d'évacuer le problème posé par la nullité de $\mathbb{P}(X = x)$: en notant f_0 (respectivement f_1) la densité de X sous H_0 (respectivement H_1), on a (pour ε petit) :

$$\mathbb{P}(X \approx x|H_0) \approx f_0(x)\varepsilon,$$
$$\mathbb{P}(X \approx x|H_1) \approx f_1(x)\varepsilon.$$

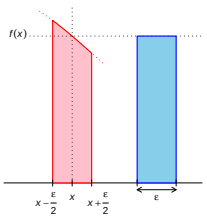


Figure 65. Justification de l'approximation de $\int_{x-\frac{\varepsilon}{2}}^{x+\frac{\varepsilon}{2}} f(t)\,\mathrm{d}t$ (en rouge) par $f(x)\varepsilon$ (en bleu).

On peut maintenant utiliser la formule de Bayes :

$$\mathbb{P}(H_1|X \approx x) = \frac{\mathbb{P}(X \approx x|H_1)\mathbb{P}(H_1)}{\mathbb{P}(X \approx x|H_0)\mathbb{P}(H_0) + \mathbb{P}(X \approx x|H_1)\mathbb{P}(H_1)},$$

et

$$\mathbb{P}(H_0|X \approx x) = \frac{\mathbb{P}(X \approx x|H_0)\mathbb{P}(H_0)}{\mathbb{P}(X \approx x|H_0)\mathbb{P}(H_0) + \mathbb{P}(X \approx x|H_1)\mathbb{P}(H_1)}.$$

On voit que ces quantités sont maintenant calculables. On simplifie l'écriture en s'intéressant direc-

tement à leur quotient :

$$\begin{aligned}\frac{\mathbb{P}(\mathrm{H}_1|\mathrm{X}\approx x)}{\mathbb{P}(\mathrm{H}_0|\mathrm{X}\approx x)} &= \frac{\mathbb{P}(\mathrm{H}_1)}{\mathbb{P}(\mathrm{H}_0)}\times\frac{\mathbb{P}(\mathrm{X}\approx x|\mathrm{H}_1)}{\mathbb{P}(\mathrm{X}\approx x|\mathrm{H}_0)}\\ &= \frac{\mathbb{P}(\mathrm{H}_1)}{\mathbb{P}(\mathrm{H}_0)}\times\frac{f_1(x)\varepsilon}{f_0(x)\varepsilon}\\ &= \frac{\mathbb{P}(\mathrm{H}_1)}{\mathbb{P}(\mathrm{H}_0)}\times\frac{f_1(x)}{f_0(x)}.\end{aligned}$$

La quantité ε disparaît du résultat final, ce qui rend cette formule valide indépendamment de la précision de la mesure, que nous avions introduit de façon arbitraire.

Pour finir on a

$$\boxed{\frac{\mathbb{P}(\mathrm{H}_1|\mathrm{X}=x)}{\mathbb{P}(\mathrm{H}_0|\mathrm{X}=x)} = \frac{\mathbb{P}(\mathrm{H}_1)}{\mathbb{P}(\mathrm{H}_0)}\times\frac{f_1(x)}{f_0(x)}.}$$

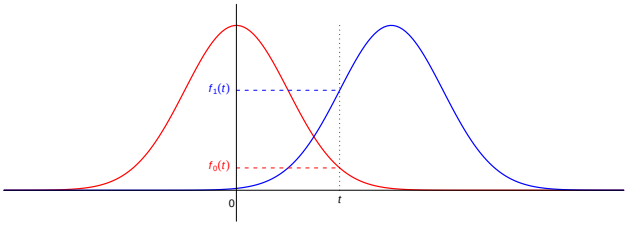


Figure 66. Les deux vraisemblances $f_0(t)$ et $f_1(t)$.

On convient généralement d'appeler la quantité $\frac{\mathbb{P}(\mathrm{H}_1)}{\mathbb{P}(\mathrm{H}_0)}$ *la cote* de l'événement H_1 (être atteint), ou encore en anglais *les odds* (terme le plus souvent utilisé en français également). La quantité $\frac{f_1(x)}{f_0(x)}$ est un *rapport de vraisemblance*. La formule ci-dessus peut se résumer ainsi : les odds a posteriori sont égaux aux odds a priori multipliés par le rapport de vraisemblance.

Les odds sont très utilisés en épidémiologie. Ils sont souvent notés comme un ratio « à l'ancienne » : par exemple 1 : 250 (1 contre 250). On peut retrouver facilement les probabilités $\mathbb{P}(\mathrm{H}_0)$ et $\mathbb{P}(\mathrm{H}_1)$ en se souvenant que leur somme vaut 1. Si par exemple les odds de H_1 contre H_0 valent 1 : 250, on a facilement $\mathrm{P}(\mathrm{H}_1) = \frac{1}{1+250}$ et $\mathbb{P}(\mathrm{H}_0) = \frac{250}{1+250}$.

10.7 Courbes ROC

Quand un test dépend d'un seuil s, la sensibilité et la spécificité du test, ou encore les risques α et β varient conjointement avec ce seuil. La courbe ROC est l'ensemble des points de coordonnées $(1-\text{Spécificité}, \text{Sensibilité}) = (\alpha, 1-\beta)$ quand ce seuil varie.

Vu la redondance du vocabulaire, on peut dire indifféremment que c'est le graphique de la sensibilité en fonction de (1-spécificité) ou de la puissance en fonction du risque α.

Les courbes ROC ont été inventée par les radaristes, pour évaluer les processus de filtrage destinés à débruiter le signal (un test est-il un processus de filtrage ?). ROC signifie *caractéristique opérante du receveur*, en anglais : *receiver operating characteristic*.

10.7.1 Utilité des courbes ROC

Les courbes ROC de deux tests permettent de comparer les tests de façon globale. Elles sont particulièrement indiquées pour comparer deux tests diagnostics qui utilisent des données biologiques différentes, mais elles peuvent aussi servir à la comparaison de tests statistiques qui utilisent les mêmes données.

Le test le plus simple à concevoir, et le moins efficace qui soit, consiste à rejeter l'hypothèse nulle avec une probabilité s, indépendamment des observations. La courbe ROC de ce test est la diagonale : pour une valeur de s donnée, on a $\alpha = 1 - \beta = s$.

Tout test raisonnable se placera donc au-dessus de la diagonale. Plus un test est efficace, plus sa courbe ROC est « bombée » et s'éloigne de la diagonale. L'« aire sous la courbe ROC », qui est comprise entre $\frac{1}{2}$ et 1, est utilisée pour résumer encore davantage l'efficacité d'un test.

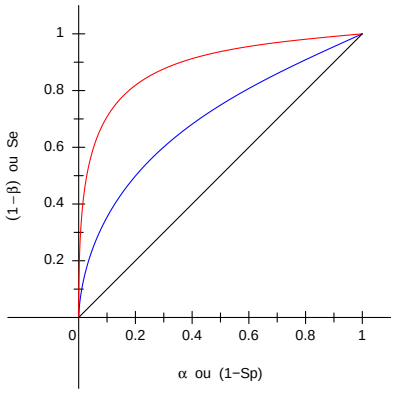


Figure 67. Le « test rouge » est plus efficace que le bleu.

10.8 Dualité entre tests et intervalles de confiance

Les tests classiques qui vont être présentés au chapitre suivant sont basés sur des calculs très similaires à ceux effectués pour les intervalles de confiance. On peut en fait en pratique passer généralement de l'un à l'autre.

10.8.1 De l'intervalle de confiance au test

Si on dispose d'un intervalle de confiance au niveau $\gamma = 1 - \alpha$, du type

$$\mathbb{P}(T_1 < \theta < T_2) \geqslant \gamma = 1 - \alpha,$$

on peut construire un test pour l'hypothèse nulle $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$, au risque α, en rejetant l'hypothèse nulle quand $\theta_0 \notin [T_1, T_2]$.

10.8.2 Du test à l'intervalle de confiance

Si on sait tester pour un risque α donné, l'hypothèse nulle $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$, on peut construire un intervalle de confiance pour θ, de niveau $\gamma = 1 - \alpha$, en y incluant toutes les valeurs de θ_0 pour lesquels le test ne rejette pas $H_0 : \theta = \theta_0$.

10.9 Un exemple détaillé

Nous allons analyser en détail un cas simple : un test sur la moyenne d'une loi normale de variance connue.

10.9.1 Problème et règles de décision

Soient $X_1, \ldots, X_n$ indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On suppose que la variance σ^2 est connue. On veut décider entre $H_0 : \mu = 0$ et $H_1 : \mu \neq 0$.

On peut imaginer une première règle, qui pourrait être posée naïvement sans connaissances préalables en statistiques : s'il y a plus de trois quarts ou moins d'un quart de X_i qui prennent des valeurs positives, la moyenne n'est pas nulle :

(R_1) Soit la statistique de test T égale au nombre de valeurs strictement positives parmi les X_i. On rejette H_0 quand $T < \frac{1}{4}n$ ou $T > \frac{3}{4}n$.

La moyenne empirique $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ est l'estimateur de μ. On définit une statistique de test $U = \frac{\sqrt{n}}{\sigma}\overline{X}$. On peut l'utiliser pour donner d'autres règles de décision :

(R_2) On rejette H_0 si $U > 1{,}64$.

(R_3) On rejette H_0 si $|U| > 1{,}96$.

Certaines de ces règles sont familières. Sont-elles toutes de « bonnes règles » ? Y en a-t-il de meilleures que d'autres ?

10.9.2 Risque α

Sous H_0, la loi des statistiques de test T et U est facile à déterminer : T suit une binomiale $\mathcal{B}in(n, \frac{1}{2})$, et U suit une normale $\mathcal{N}(0,1)$. On peut donc calculer le risque α associé à H_0 pour chacune des trois règles de rejet.

Pour la règle R_1 : $\alpha_1 = \mathbb{P}\left(T < \frac{1}{4}n\right) + \mathbb{P}\left(T > \frac{3}{4}n\right) = 1 - \mathbb{P}\left(\frac{1}{4}n \leqslant T \leqslant \frac{3}{4}n\right)$, avec $T \sim \mathcal{B}in\left(n, \frac{1}{2}\right)$, et donc

$$\alpha_1 = 1 - \sum_{\frac{1}{4}n \leqslant k \leqslant \frac{3}{4}n} \mathbb{P}(T = k)$$

On sait calculer $\mathbb{P}(T = k) = \binom{n}{k}\frac{1}{2^n}$ (sous H_0), donc on peut calculer α_1. C'est fastidieux à la main mais les ordinateurs font ça très bien. Voici quelques valeurs de α_1 (qui dépend de n) :

n	5	10	20	50
α_1	37,5%	10,9%	1,2%	0,03%

Table 14. α_1 en fonction de n

Alternativement, on peut approcher la loi de T par une loi normale pour calculer plus facilement des valeurs approchées de α_1.

Pour les règles R_2 et R_3 : la loi de U sous H_0 est une loi normale $\mathcal{N}(0,1)$ (c'est pour cette raison qu'on a défini U au lieu de donner des critères sur la valeur de $\overline{X}$), on a donc pour R_2, $\alpha_2 = \mathbb{P}(U > 1{,}64) = 0{,}05$ et pour R_3, $\alpha_3 = \mathbb{P}(|U| > 1{,}96) = 0{,}05$.

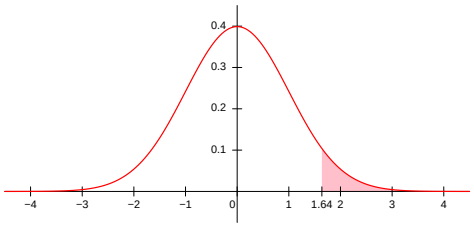


Figure 68. $\alpha_2 = \mathbb{P}(U > 1{,}64) = 5\%$

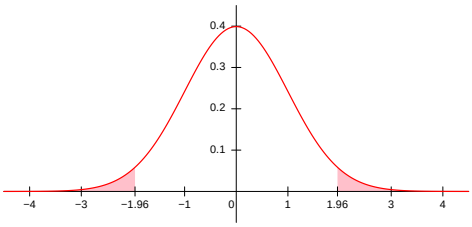


Figure 69. $\alpha_3 = \mathbb{P}(|U] > 1{,}96) = 5\%$

Ces deux tests basés sur U ont même risque α ; on peut d'ores et déjà prédire qu'ils n'auront pas même risque β sous une hypothèse H_1 précise donnée.

Le test de la règle R_2 détectera mieux que R_3 les « déviations à droite » de μ, c'est-à-dire les cas où la vraie valeur de μ est positive ; par contre si $\mu < 0$, le test R_2 perd toute puissance. Ce type de test est appelé « test unilatéral », il est utilisé quand les connaissances a priori sur les valeurs que μ est susceptibles de prendre permettent d'écarter les valeurs négatives.

10.9.3 Risque β

Si on se place sous H_1, on ne sait plus déterminer complètement la loi de T et U ; on sait qu'on a toujours une binomiale et une normale, mais elles dépendent d'un paramètre inconnu. Nous avons déjà commenté cette dissymétrie entre les hypothèses.

Si on précise la valeur de σ^2, et l'hypothèse alternative par $H_1 : \mu = \mu_1$, on peut calculer le risque β.

Pour la règle R_1 : sous H_1, on a $X_i \sim \mathcal{N}(\mu_1, \sigma^2)$. La probabilité que X_i soit positif est

$$\begin{aligned} p_1 &= \mathbb{P}(X_i > 0) \\ &= \mathbb{P}\left(\frac{X_i - \mu_1}{\sigma} > -\frac{\mu_1}{\sigma}\right) \\ &= \mathbb{P}(Z > -\Delta), \end{aligned}$$

où $\Delta = \frac{\mu_1}{\sigma}$ et $Z \sim \mathcal{N}(0,1)$.

La loi de T sous H_1 est $\mathcal{B}in(n, p_1)$. Le risque β est la probabilité que $\left(\frac{1}{4}n \leqslant T \leqslant \frac{3}{4}n\right)$, soit

$$\begin{aligned} \beta_1 &= \mathbb{P}\left(\frac{1}{4}n \leqslant T \leqslant \frac{3}{4}n\right) \\ &= \sum_{\frac{1}{4}n \leqslant k \leqslant \frac{3}{4}n} \mathbb{P}(T = k), \end{aligned}$$

avec $\mathbb{P}(T = k) = \binom{n}{k} p_1^k (1-p_1)^{n-k}$ sous H_1.

	n	5	10	20	50
($\Delta = 1$)	β_1	17,8%	20,3%	20,1%	4,5%
($\Delta = 1{,}5$)	β_1	3,9%	2,5%	0,9%	0,00%

Table 15. β_1 en fonction de n et Δ

On voit que si la taille de l'échantillon n est fixée, la loi de T sous H_1, et donc le risque β, ne dépendent que de Δ, qui est l'écart entre les valeurs de μ sous H_1 et sous H_0 rapporté à l'écart-type. On appelle Δ la *taille de l'effet*. Si par exemple $\Delta = 2$, on dira que la taille de l'effet est de 2 écart-types.

Ce phénomène est très général en statistique : pour tous les tests usuels, on peut définir une taille de l'effet, et la puissance dépendra unique de la taille de l'effet multipliée par la racine carrée de la taille de l'échantillon.

Pour la suite, on va se limiter à la règle R_3 (le test bilatéral) : la loi de $\overline{X}$ sous H_1 est $\mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right)$, et donc la loi de $U = \frac{\sqrt{n}}{\sigma}\overline{X}$ est $\mathcal{N}\left(\sqrt{n}\frac{\mu_1}{\sigma}, 1\right)$. Le risque β est la probabilité que U soit entre −1,96 et 1,96 ; en notant toujours $\Delta = \frac{\mu_1}{\sigma}$ on a $U \sim \mathcal{N}(\Delta\sqrt{n}, 1)$, et

$$\begin{aligned} \beta_3 &= \mathbb{P}(-1{,}96 \leqslant U \leqslant 1{,}96) \\ &= \mathbb{P}\left(-1{,}96 - \Delta\sqrt{n} \leqslant Z \leqslant 1{,}96 - \Delta\sqrt{n}\right) \\ &= F_Z(1{,}96 - \Delta\sqrt{n}) - F_Z(-1{,}96 - \Delta\sqrt{n}), \end{aligned}$$

où F_Z est la fonction de répartition de $Z = (U - \Delta\sqrt{n}) \sim \mathcal{N}(0,1)$. Le risque β ne dépend, là encore, que de la taille de l'échantillon et de la taille de l'effet (plus précisément, β ne dépend que de $\Delta\sqrt{n}$).

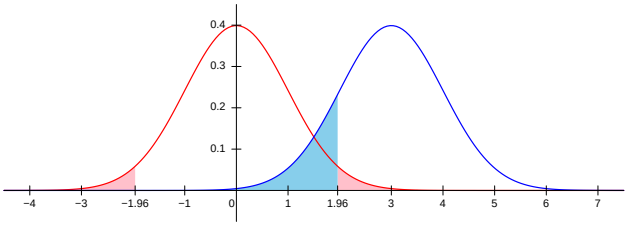


Figure 70. En rouge, la distribution de U sous H_0 et le risque α, en bleu la distribution de U sous H_1 et le risque β. Ici on a $\Delta\sqrt{n} = 3$.

On peut voir sur la figure que dès que $\Delta\sqrt{n}$ est assez grand (par exemple $\Delta\sqrt{n} > 1$), la quantité $F_Z(-1{,}96 - \Delta\sqrt{n}) = \mathbb{P}(Z \leqslant -1{,}96 - \Delta\sqrt{n}) = \mathbb{P}(U \leqslant -1{,}96)$ devient négligeable, et on a $\beta \approx \mathbb{P}(Z \leqslant 1{,}96 - \Delta\sqrt{n}) = \mathbb{P}(U \leqslant 1{,}96)$).

Attention, l'hypothèse $\Delta\sqrt{n}$ grand est vraiment nécessaire (pensez au cas $\Delta = 0$!).

	n	5	10	20	50
($\Delta = 1$)	β_3	39,1%	11,4%	0,6%	0,00%
($\Delta = 1{,}5$)	β_3	8,2%	0,27%	0,00%	0,00%

Table 16. β_3 en fonction de n et Δ

Notez que pour ce test, le risque α ne dépend pas de n, et vaut 0,05. Pour le test R_1, α varie avec n, ce qui rend la comparaison des risques β_1 et β_3 calculés ci-dessus impossible.

10.9.4 Calcul du degré de signification pour ces tests

Supposons ici qu'on a effectué $n = 10$ mesures d'une variable suivant une loi gaussienne de variance 1 ; on teste si sa moyenne est nulle.

-0,22	0,74	1,46	0,11	0,23
-0,72	1,26	1,57	0,85	0,60

Table 17. 10 mesures.

La valeur prise par la statistique de test T est $t = 8$: plus des trois-quarts des mesures sont positives, et la règle R_1 conduit à rejeter l'hypothèse nulle. Pour calculer le degré de signification, il faut changer légérement la formulation du test : on pose $T' = \left|T - \frac{1}{2}n\right|$, et la règle devient : on rejette H_0 quand $T' > \frac{1}{4}n$. Ici T′ prend la valeur $t' = 8 - 5 = 3$.

Dans notre exemple, T′ ne peut prendre que les valeurs entières de 0 à 5, et on a $\mathbb{P}(T' = k) = \mathbb{P}(T = 5 + k) + \mathbb{P}(T = 5 - k)$, ce qui peut être calculé sachant que $T \sim \mathscr{B}in\left(10, \frac{1}{2}\right)$ (sous H_0).

Le degré de signification du test est

$$\begin{aligned}\mathbb{P}(T' \geqslant 3) &= \mathbb{P}(T \geqslant 8) + \mathbb{P}(T \leqslant 2)\\ &= 0{,}109.\end{aligned}$$

Notons que c'est exactement le risque α de notre test, qui s'exprime en effet avec la nouvelle notation comme $\alpha = \mathbb{P}(T' > 2{,}5)$.

Passons aux règles R_2 et R_3. La valeur prise par la statistique de test U est $u = 1{,}86$. La règle R_2 conduit à rejeter H_0 ; le degré de signification est

$$\mathbb{P}(U \geqslant 1{,}86) = 0{,}0314.$$

La règle R_3 conduit à accepter H_0 ; le degré de signification est

$$\mathbb{P}(|U| \geqslant 1{,}86) = 2\mathbb{P}(U > 1{,}86) = 0{,}0628.$$

10.9.5 Courbe ROC

On va comparer ici les règle R_1 et R_3, légèrement modifiée pour pouvoir considérer un seuil de rejet s variable.

Le test R_1 s'exprime à partir de la statistique $T'' = \left|\frac{T - \frac{1}{2}n}{n}\right|$; on rejettera H_0 si $T'' > s$ (s varie entre 0 et $\frac{1}{2}$), ou encore si $\left(T < \frac{1}{2}n - sn\right)$ ou $\left(T > \frac{1}{2}n + sn\right)$. On a donc

$$\begin{aligned}\alpha_1 &= \mathbb{P}\left(T < \frac{1}{2}n - sn\right) + \mathbb{P}\left(T > \frac{1}{2}n + sn\right)\\ &= 1 - \mathbb{P}\left(\frac{1}{2}n - sn \leqslant T \leqslant \frac{1}{2}n + sn\right)\\ &= 1 - \sum_{\frac{1}{2}n - sn \leqslant k \leqslant \frac{1}{2}n + sn} \mathbb{P}(T = k)\end{aligned}$$

où $\mathbb{P}(T = k) = \binom{n}{k}\frac{1}{2^n}$ (sous H_0).

De même, on calcule le risque β en se plaçant sous H_1, où $T \sim \mathcal{B}in(n, p_1)$, p_1 étant toujours $p_1 = \mathbb{P}(Z > -\Delta)$, avec $\Delta = \frac{\mu_1}{\sigma}$ et $Z \sim \mathcal{N}(0,1)$. On a :

$$\begin{aligned}\beta_1 &= \mathbb{P}\left(\frac{1}{2}n - sn \leqslant T \leqslant \frac{1}{2}n + sn\right)\\ &= \sum_{\frac{1}{2}n - sn \leqslant k \leqslant \frac{1}{2}n + sn} \mathbb{P}(T = k)\end{aligned}$$

avec $\mathbb{P}(T = k) = \binom{n}{k} p_1^k (1 - p_1)^{n-k}$ sous H_1.

Ceci permet de tracer des courbes ROC pour le test R_1.

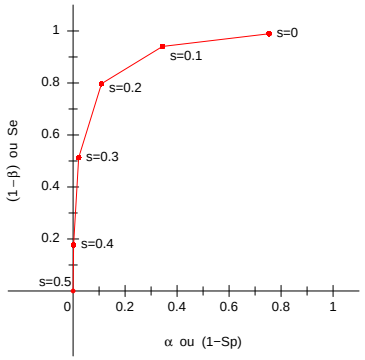


Figure 71. Courbe ROC du test R_1 avec $n = 10$ et $\Delta = 1$

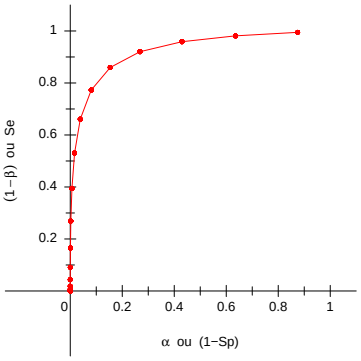


Figure 72. Courbe ROC du test R_1 avec $n = 40$ et $\Delta = 0{,}5$

La statistique de test étant discrète, (α, β) ne prend qu'un nombre fini de valeurs. Dans le graphique nous les avons reliées par des droites. Notez que la valeur de s qui donne lieu au test le plus strict (celui qui rejette H_0 le plus facilement) est $s = 0$: on n'accepte H_0 que si $T = \frac{1}{2}n$; le risque α pour cette valeur de s est $1 - \mathbb{P}\left(T = \frac{1}{2}n\right)$, qui est strictement plus petit que 1 quand n est pair.

Passons au test R_3. On rejette H_0 si $|U| > s$ (s varie entre 0 et $+\infty$). En notant F_Z la fonction de répartition de $\mathcal{N}(0,1)$, on a $\alpha_3 = 2(1 - F_Z(s))$ et, en procédant comme précédemment, $\beta_3 = F_Z\left(s - \Delta\sqrt{n}\right) - F_Z\left(-s - \Delta\sqrt{n}\right)$.

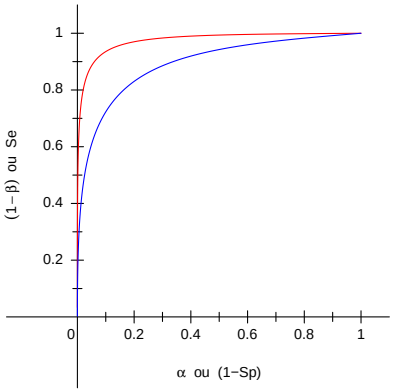


Figure 73. Courbes ROC du test R_3 avec $\Delta = 1$, $n = 5$ (bleu) $n = 10$ (rouge)

On peut maintenant comparer les tests R_1 et R_3, au moyen de leurs courbes ROC. On constate sur le graphe ci-dessous qu'à risque α comparable, R_3 a toujours un risque β plus faible (une puissance $1-\beta$ plus élévée).

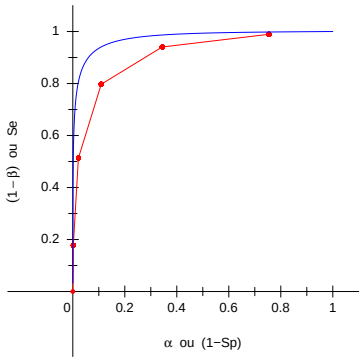


Figure 74. Courbes ROC des tests R_1 (rouge) et R_3 (bleu) avec $n = 10$ et $\Delta = 1$

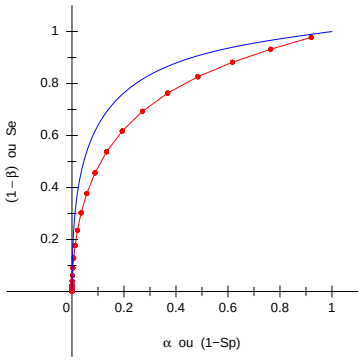


Figure 75. Courbes ROC des tests R_1 (rouge) et R_3 (bleu) avec $n = 100$ et $\Delta = 0{,}2$

Le test R_1 n'utilise pas toute l'information disponible : ni la valeur de σ^2, ni même la normalité des données. Il est donc normal qu'il soit moins puissant.

En revanche, on pourrait montrer qu'il est plus « robuste » à des écarts aux hypothèses faites : il continuera par exemple à donner de bons résultats même si la valeur de σ^2 utilisée est erronée. Il s'agit en fait d'un test de nullité de la médiane de la distribution.

10.10 Exercices

Exercice 1

On considère deux variables aléatoires X et Y obtenues après une transformation appropriée de deux mesures biologiques chez un même sujet.

On suppose que chez les sujets sains ces deux variables sont indépendantes et suivent une loi $\mathcal{N}(0,1)$.

On s'intéresse à la recherche d'états pathologiques qui peuvent se traduire par une valeur trop faible ou trop élevée de ces variables. On veut établir un test diagnostic.

1. Dans un premier temps, on décide de considérer qu'un sujet est dans un état pathologique quand X n'est pas entre $-1{,}96$ et $1{,}96$ ou quand Y n'est pas entre $-1{,}96$ et $1{,}96$, soit en notation mathématique quand

$$|X| > 1{,}96 \text{ ou } |Y| > 1{,}96.$$

a. Quelle est la probabilité α de classer à tort un sujet sain comme pathologique ?

b. Dessiner la zone d'acceptation de ce test.

2. On décide de modifier le test proposé, en utilisant toujours un test de la forme :

$$|X| > c \text{ ou } |Y| > c.$$

Pour quelle valeur de c a-t-on $\alpha = 0{,}05$?

3. On pose $Z = X^2 + Y^2$. Quelle est la loi de Z ?

4. Proposer un test de risque $\alpha = 0{,}05$ basé sur la valeur de Z.

5. Représenter sur un même dessin les zones d'acceptations des tests des questions 2 et 4.

Exercice 2

Le taux de gonadotrophine chorionique (hCG) mesuré dans le sang des femmes enceintes entre la quinzième et la dix-septième semaine de grossesse est un marqueur de risque de la trisomie 21 : plus le taux en est élevé, plus le risque de trisomie 21 est grand.

On note $F(x) = \mathbb{P}(\text{hCG} \leqslant x)$ pour une grossesse normale, et $G(x) = \mathbb{P}(\text{hCG} \leqslant x)$ pour une trisomie 21. On a estimé F et G à partir de 524 grossesses normales et 125 trisomies 21 : le tableau suivant donne ces valeurs estimées pour une série de valeurs de x (en UI/litre).

x	F(x)	G(x)	x	F(x)	G(x)	x	F(x)	G(x)
1000	0,03	0,00	32000	0,77	0,25	58990	0,99	0,70
2000	0,06	0,00	33000	0,80	0,26	61000	0,99	0,72
3000	0,10	0,00	34000	0,82	0,29	66000	0,99	0,74
4000	0,11	0,00	35000	0,84	0,30	69000	0,99	0,75
6000	0,13	0,00	36000	0,85	0,30	71000	0,99	0,76
7000	0,14	0,00	37000	0,87	0,32	74000	0,99	0,77
8000	0,15	0,00	38000	0,87	0,34	77000	0,99	0,78
9000	0,17	0,01	39000	0,89	0,36	78000	0,99	0,79
11000	0,21	0,02	39950	0,89	0,37	79000	1,00	0,80
12000	0,23	0,02	41000	0,90	0,38	80000	1,00	0,81
13000	0,27	0,03	41990	0,90	0,39	81000	1,00	0,82
14000	0,30	0,04	42000	0,91	0,41	83000	1,00	0,83
14960	0,30	0,05	43000	0,92	0,43	84000	1,00	0,84
15000	0,33	0,06	43010	0,92	0,44	88000	1,00	0,85
16000	0,36	0,06	44000	0,93	0,46	89000	1,00	0,86
17000	0,39	0,08	45000	0,94	0,46	90000	1,00	0,87
17500	0,39	0,09	46000	0,94	0,50	91000	1,00	0,88
18000	0,44	0,09	47000	0,95	0,52	97000	1,00	0,89
19000	0,46	0,10	47940	0,95	0,53	99000	1,00	0,90
21000	0,50	0,11	48000	0,96	0,53	101000	1,00	0,91
22000	0,54	0,11	49000	0,96	0,55	102000	1,00	0,92
23000	0,56	0,12	50000	0,96	0,56	102500	1,00	0,93
24000	0,58	0,13	52000	0,97	0,59	103000	1,00	0,94
25000	0,61	0,14	53000	0,97	0,60	107000	1,00	0,95
26000	0,65	0,14	54000	0,97	0,61	110000	1,00	0,96
26010	0,65	0,15	55000	0,98	0,62	115000	1,00	0,97
27000	0,68	0,18	56000	0,98	0,64	116000	1,00	0,98
28000	0,70	0,19	56950	0,98	0,65	130000	1,00	0,99
29000	0,73	0,21	57000	0,98	0,66	139000	1,00	1,00
30000	0,73	0,24	57500	0,98	0,67			
31000	0,76	0,25	58000	0,99	0,69			

Table 18. Fonctions de répartition de hCG chez les témoins et les cas
(données : Françoise Jauzein (INRP))

1. On veut réaliser un test se basant sur la seule mesure hCG : si hCG dépasse un certain seuil s, on prescrira une amniocentèse (un test permettant un dépistage sans ambiguïté de la trisomie).

Pour quel valeur de s a-t-on une spécificité de 95% ? Quelle est alors la sensibilité du test ?

2. Quels sont la spécifité et la sensibilité du test pour $s = 21000$? Et pour $s = 46000$?

3. Tracer la courbe ROC de ce test ; on ne placera que quelques points de façon précise, en indiquant lesquels.

Exercice 3 On admettra le résultat suivant : si $X \sim \mathscr{P}(\lambda)$, la loi de $U = 2\sqrt{X}$ est approximativement normale : $U \sim \mathscr{N}(2\sqrt{\lambda}, 1)$.

On suppose que le nombre de morts sur les routes en un mois suit une loi de Poisson $\mathscr{P}(\lambda)$, où le paramètre λ dépend du mois considéré.

1. En quoi cette modélisation paraît-elle raisonnable ou criticable ?

2. Certaines associations de lutte contre la violence routière affirment que la réforme du permis à points votée fin 2010 s'est traduite de façon immédiate par une augmentation du nombre de morts dans les premiers mois de 2011.

On donne dans la table qui suit les valeurs x_i et y_i du nombre de morts en 2010 et 2011 pour les mois de janvier à avril. Pour faciliter les calculs, on donne les valeurs des $u_i = 2\sqrt{x_i}$ et $v_i = 2\sqrt{y_i}$.

Mois	x_i (2010)	$u_i = 2\sqrt{x_i}$	y_i (2011)	$v_i = 2\sqrt{y_i}$
Janvier	273	33,05	331	36,39
Février	254	31,87	273	33,05
Mars	300	34,64	308	35,10
Avril	296	34,41	355	37,68

Table 19. Nombre de morts sur les routes pendant les 4 premiers mois de 2010 et 2011 (données : Observatoire national interministériel de la sécurité routière)

On veut tester l'hypothèse H_0 : « La mortalité n'a pas changé de façon significative en 2010 et 2011 », contre H_1 : « La mortalité a augmenté de façon significative ». Montrer que sous H_0, les valeurs de $\delta_i = v_i - u_i$ suivent une loi normale $\mathscr{N}(0,2)$. Réaliser le test et conclure.

Exercice 4 On dispose d'un échantillon de $n = 35$ mesures indépendantes $X_1, \ldots, X_{35}$ suivant une loi $\mathscr{N}(0, \sigma^2)$. La valeur $\mu = 0$ de l'espérance est connue avec certitude. On souhaite tester l'hypothèse nulle H_0 : « $\sigma^2 = 1$ » contre l'hypothèse alternative H_1 : « $\sigma^2 > 1$ » (il s'agit donc d'un test unilatéral).

1. On pose $Z = X_1^2 + \cdots + X_{35}^2$. Quelle est la loi de Z ?
2. En utilisant Z, donnez un critère de test de H_0 contre H_1 au risque $\alpha = 0{,}05$.
3. On observe une valeur $z = 30{,}1$. Donner une estimation de σ^2. Doit-on rejeter H_0 ?
4. On suppose que l'hypothèse H_1 est vraie, avec $\sigma^2 = 1{,}45$. Quelle est la valeur du risque β ?

Exercice 5 La morsure de l'araignée *Ferox apachea* (araignée A) est parfois mortelle si elle n'est pas soignée dans les heures qui suivent l'attaque. Cependant il existe une autre araignée qui lui ressemble beaucoup, *Ferox geronimo* (araignée B), et dont la morsure ne nécessite pas de traitement.

Le traitement de la morsure pouvant occasionner de graves effets secondaires, il est important de différencier les deux araignées. Pour distinguer ces araignées à coup sûr il faut observer leurs pièces buccales au microscope électronique, ce qui est rarement possible dans des délais raisonnables. Une autre possibilité est de compter les taches jaunes que ces araignées portent sur le dos, car les araignées A (dangereuses) en ont en général plus que les araignées B. La table suivante donne pour chaque espèce d'araignée, la proportion d'entre elles qui portent un nombre de taches donné (on supposera que les valeurs rapportées dans cette table sont exactes).

nb taches	0	1	2	3	4	5	6	7	8	9	10+
Araignée A	0,00	0,01	0,02	0,05	0,09	0,13	0,15	0,15	0,13	0,10	0,17
Araignée B	0,06	0,15	0,22	0,22	0,17	0,10	0,05	0,02	0,01	0,00	0,00

Ainsi par exemple 13% des araignées de l'espèce A portent 5 taches, et 10% des araignées de l'espèce B.

1. Vous êtes mordu par une araignée qui ne porte aucune tache. Acceptez-vous le traitement ? Pourquoi ?

2. Vous êtes mordu par une araignée qui porte plus de 9 taches. Acceptez-vous le traitement ? Pourquoi ?

3. Vous travaillez dans un dispensaire et vous voyez chaque mois plusieurs patients mordu par des araignées. Vous décidez d'établir une valeur seuil s (pour le nombre de taches) au-delà de laquelle vous traiterez systématiquement les patients. On traitera ce problème comme un test statistique dont l'hypothèse nulle est H_0 : « l'araignée est de type B », et l'hypothèse alternative est H_1 : « l'araignée est de type A ». On rejettera H_0 quand $N > s$ où N est le nombre de taches de l'araignée.

(a) Quel est le risque α du test quand $s = 5$? Quelle est sa puissance ?

(b) Même question pour $s = 3$, $s = 7$.

Chapitre 11

Tests usuels

Nous rappelons quelques procédures de test classiques. Les procédures sont généralement données pour des tests bilatéraux ; la généralisation au cas unilatéral n'est pas difficile.

11.1 Test sur une moyenne, cas gaussien

11.1.1 La procédure de test

Cette section est à mettre en regard de la section 9.1.

Soient $X_1, \dots, X_n$ des variables indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On teste l'hypothèse nulle $H_0 : \mu = \mu_0$ contre l'hypothèse alternative $H_1 : \mu \neq \mu_0$ (test bilatéral).

On pose

$$T = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}.$$

On a, sous H_0, $T \sim t(n-1)$ (proposition 39). Un test de risque α est obtenu en rejetant H_0 quand $|T| > t^{n-1}_{1-\alpha/2}$.

11.1.2 Calcul du risque β

Nous n'allons pas parler du calcul du risque β pour tous les tests présentés, mais nous saisissons ici l'occasion de présenter une loi « décentrée ». On comparera utilement la définition suivante avec la définition 35).

Définition 37 *Soient $\delta \in \mathbb{R}$, $Z \sim \mathcal{N}(0,1)$ et $Y \sim \chi^2(d)$. La variable aléatoire $X = \dfrac{Z+\delta}{\sqrt{\frac{Y}{d}}}$ suit une loi continue à densité, appelée loi t de Student décentrée à d degrés de liberté, de paramètre de non-centralité δ, notée $t(d, \delta)$.*

La proposition suivante est l'analogue de la proposition 39.

Proposition 42 *Soient $X_1, \dots, X_n$ des variables indépendantes de loi $\mathcal{N}(\mu_1, \sigma^2)$. On pose $\Delta = \frac{\mu_1 - \mu_0}{\sigma}$: c'est la « taille de l'effet », ici l'écart entre μ_1 et μ_0 exprimé en nombre d'écarts-type. Soit*

$$T = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}.$$

Alors $T \sim t\left(n-1, \Delta\sqrt{n}\right)$.

La loi de $\overline{X}$ est $\mathcal{N}\left(\mu_1, \frac{\sigma^2}{n}\right)$. On pose $Z = \dfrac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$; l'espérance de Z est $E(Z) = \frac{1}{\sqrt{\frac{\sigma^2}{n}}}\left(E\left(\overline{X}\right) - \mu_0\right) = \sqrt{n}\frac{\mu_1 - \mu_0}{\sigma} = \Delta\sqrt{n}$, et sa variance est $\frac{1}{\frac{\sigma^2}{n}} \operatorname{var}\left(\overline{X}\right) = 1$. La loi de Z est donc $\mathcal{N}\left(\Delta\sqrt{n}, 1\right)$. On pose $Y = \frac{n-1}{\sigma^2}S^2$; la loi de Y est $\chi^2(n-1)$.

On a donc $\dfrac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t(n-1, \Delta\sqrt{n})$, et comme dans la preuve de la proposition 39 on a

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}},$$

ce qui achève la preuve.

On peut maintenant calculer la puissance. On précise l'hypothèse alternative en posant $H_1 : \mu = \mu_1$. La proposition précédente montre que la loi de T sous H_1 est $t(n-1, \Delta\sqrt{n})$. Notons $F_{n-1,\Delta\sqrt{n}}$ sa fonction de répartition. Le risque β est la probabilité que T tombe dans la zone d'acceptation :

$$\begin{aligned}\beta &= \mathbb{P}(-t^{n-1}_{1-\alpha/2} \leqslant T \leqslant t^{n-1}_{1-\alpha/2}) \\ &= F_{n-1,\Delta\sqrt{n}}(t^{n-1}_{1-\alpha/2}) - F_{n-1,\Delta\sqrt{n}}(-t^{n-1}_{1-\alpha/2})\end{aligned}$$

Il existe des tables pour la fonction de répartition de la loi t décentrée, mais c'est un peu lourd à manipuler, et surtout cela fait figure d'antiquité : les logiciels de statistiques savent calculer ces valeurs, et nul ne passerait plus aujourd'hui par une table. Pour cet enseignement où malheureusement on n'a pas le loisir d'utiliser un ordinateur, nous en resterons donc là.

11.2 Test sur une moyenne, grands échantillons

Quand les échantillons sont grands, l'hypothèse de normalité devient inutile : on se repose sur le théorème central limite comme on l'a fait à la section 9.1.

On considère que pour n assez grand,

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$$

suit approximativement une loi normale centrée réduite. Un test de risque α est obtenu en rejetant H_0 quand $|Z| > z_{1-\alpha/2}$.

11.3 Comparaison de deux moyennes, cas gaussien

Cette section est à mettre en regard de la section 9.2. On considère un échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma^2)$ et un échantillon de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma^2)$ (les deux lois ont même variance σ^2). On va tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$ (test bilatéral).

Les moyennes empiriques des deux échantillons sont $\overline{X}_1$ et $\overline{X}_2$. On a

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma^2\right).$$

L'estimateur de σ^2 est

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

où S_1^2 et S_2^2 sont les estimateurs de la variances dans les deux échantillons.

Sous H_0, on a

$$T = \frac{(\overline{X}_1 - \overline{X}_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

Un test de risque α est obtenu en rejetant H_0 quand $|T| > t^{n-1}_{1-\alpha/2}$.

11.4 Comparaison de deux moyennes, grands échantillons

Dans le cas des grands échantillons, on n'a plus besoin que les distributions soient gaussiennes, et on on se repose sur le théorème de la limite centrale.

On considère un échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et un échantillon de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma_2^2)$ (les deux lois peuvent avoir des variances distinctes). On va tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$ (test bilatéral).

En supposant n_1 et n_2 assez grands (disons $n_1, n_2 \geqslant 30$), on sait que les moyennes empiriques des deux échantillons, $\overline{X}_1$ et $\overline{X}_2$ sont approximativement normales de variances $\frac{\sigma_1^2}{n_1}$ et $\frac{\sigma_2^2}{n_2}$. On a

$$\overline{X}_1 - \overline{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)\right).$$

On ne connait pas σ_1^2 et σ_2^2 ; on les remplace par leurs estimations S_1^2 et S_2^2, et on pose

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

On voit que la loi Z est proche de $\mathcal{N}(0,1)$.

Un test de risque α est obtenu en rejetant H_0 quand $|Z| > z_{1-\alpha/2}$.

11.5 Test sur la variance d'une loi normale

Cette section est à mettre en regard de la section 9.3. Soient $X_1, \ldots, X_n$ des variables indépendantes de loi $\mathcal{N}(\mu, \sigma^2)$. On va tester l'hypothèse nulle $H_0 : \sigma^2 = \sigma_0^2$ contre l'hypothèse alternative $H_1 : \sigma^2 \neq \sigma_0^2$ (test bilatéral).

Sous H_0 on a (théorème 38)

$$Y = \frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1).$$

On obtient un test de risque α en rejetant H_0 quand $Y < x^{n-1}_{\alpha/2}$ ou $Y > x^{n-1}_{1-\alpha/2}$.

11.6 Comparaison de deux variances

Cette section est à mettre en regard de la section 9.4. On considère un échantillon de taille n_1 issu d'une loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et un échantillon de taille n_2 issu d'une loi $\mathcal{N}(\mu_2, \sigma_2^2)$. On va tester l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$ (test bilatéral).

La variance empirique du premier échantillon est S_1^2, celle du second est S_2^2. Sous H_0 on a (proposition 40)

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1).$$

On obtient un test de risque α en rejetant H_0 quand

$$\frac{S_1^2}{S_2^2} < F_{\alpha/2}^{n_1-1,n_2-1} \text{ ou } \frac{S_1^2}{S_2^2} > F_{1-\alpha/2}^{n_1-1,n_2-1}.$$

En utilisant $F_\alpha^{d_1,d_2} = \dfrac{1}{F_{1-\alpha}^{d_2,d_1}}$, on peut transformer la première inégalité en

$$\frac{S_2^2}{S_1^2} > F_{1-\alpha/2}^{n_2-1,n_1-1},$$

ce qui facilite l'utilisation des tables.

11.6.1 Remarque sur la procédure « test F puis test t »

Étant donné deux échantillons gaussiens dont on veut comparer les moyennes, on pratique souvent la procédure suivante : on fait d'abord un test d'égalité des variances comme ci-dessus, et si cette hypothèse n'est pas rejetée on fait un test d'égalité des moyennes (section 11.3).

On a ainsi une procédure qui peut conclure de trois façons différentes : variances différentes, variances égales et moyennes différentes, variances égales et moyennes égales.

Le contrôle des risques de cette procédure n'est pas aisé. Supposons qu'on pratique les deux tests au seuil $\alpha = 5\%$.

Si les variances et les moyennes sont égales, le test F va rejeter l'égalité des variances dans 5% des cas ; le test t va à nouveau rejeter une partie des cas restant (en pratique environ 5%, bien que les deux tests ne soient pas a priori indépendants).

Si au contraire les variances sont différentes et les moyennes égales, le test F peut cependant ne pas rejeter l'hypothèse d'égalité des variances, et on pratique un test t dans des conditions douteuses.

Donnons quelques exemples, obtenus pour des échantillons de taille $n_1 = 10$ et $n_2 = 50$.

- Si $\sigma_1 = \sigma_2$ et $\mu_1 = \mu_2$, on a
  - Rejet de l'égalité des variances : environ 5% ;
  - Non-rejet des l'égalité des variances, rejet de l'égalité des moyennes : environ 5% ;
  - Non-rejet des l'égalité des variances, non-rejet de l'égalité des moyennes : environ 90%.
- Si $\sigma_1 = 1$, $\sigma_2 = 2$ et $\mu_1 = \mu_2$, on a
  - Rejet de l'égalité des variances : environ 19% ;
  - Non-rejet des l'égalité des variances, rejet de l'égalité des moyennes : environ 1,3% ;
  - Non-rejet des l'égalité des variances, non-rejet de l'égalité des moyennes : environ 79,7%.
- Si $\sigma_1 = 2$, $\sigma_2 = 1$ et $\mu_1 = \mu_2$, on a
  - Rejet de l'égalité des variances : environ 32,2% ;
  - Non-rejet des l'égalité des variances, rejet de l'égalité des moyennes : environ 8,5% ;
  - Non-rejet des l'égalité des variances, non-rejet de l'égalité des moyennes : environ 59,2%.

11.7 Test sur une proportion

Cette section est à mettre en regard de la section 9.5. Soient $X_1, \ldots, X_n$ des variables indépendantes de loi $\mathscr{B}(p)$. On va tester l'hypothèse nulle $H_0 : p = p_0$ contre l'hypothèse alternative $H_1 : p \neq p_0$ (test bilatéral).

L'estimateur de p est $\widehat{p} = \frac{1}{n}(X_1 + \cdots + X_n)$. Sous H_0 et pour n grand (disons, si $np_0 > 5$ et $n(1-p_0) > 5$), la loi de $\widehat{p}$ est approximativement normale :

$$\widehat{p} \sim \mathscr{N}\left(\mu = p_0, \sigma^2 = \frac{p_0(1-p_0)}{n}\right).$$

On obtient un test de risque α en rejetant H_0 quand

$$\left| \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{1-\alpha/2}.$$

11.8 Comparaison de deux proportions

Cette section est à mettre en regard de la section 9.6. On considère un échantillon de taille n_1 issu d'une loi $\mathscr{B}(p_1)$ et un de taille n_2 issu d'une loi $\mathscr{B}(p_2)$. On va tester l'hypothèse nulle $H_0 : p_1 = p_2$ contre l'hypothèse alternative $H_1 : p_1 \neq p_2$ (test bilatéral).

Sous H_0, en notant $p = p_1 = p_2$, si n est assez grand (disons si $n_1\widehat{p}_1(1-\widehat{p}_1) > 12$ et $n_2\widehat{p}_2(1-\widehat{p}_2) > 12$) on a

$$(\widehat{p}_1 - \widehat{p}_2) \sim \mathscr{N}\left(\mu = 0, \sigma^2 = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Il faut une estimation de $\widehat{p}$ pour estimer la variance ; on prend la proportion de succès observée dans la réunion des deux échantillons, qui peut s'exprimer comme

$$\widehat{p} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}.$$

On obtient un test de risque α en rejetant H_0 quand

$$\left| \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| > z_{1-\alpha/2}.$$

11.9 Séries appariées

On suppose qu'on a un échantillon de n variables gaussiennes $X_1, \ldots, X_n$, de même variance σ^2, chaque X_i étant d'espérance μ_i : $X_i \sim \mathscr{N}(\mu_i, \sigma^2)$. On a un deuxième échantillon $Y_1, \ldots, Y_n$ avec $Y_i \sim \mathscr{N}(\mu_i + \delta, \sigma^2)$. On veut tester l'hypothèse nulle $H_0 : \delta = 0$ contre l'hypothèse alternative $H_1 : \delta \neq 0$.

Une situation type qui conduit à ce modèle est celle où on compare, par exemple, deux méthodes de dosage d'anticorps, sur n échantillons distincts.

La solution classique est la suivante :

On pose $Z_i = Y_i - X_i$. Chaque Z_i suit une loi normale $\mathscr{N}(\delta, 2\sigma^2)$. On peut donc tester $\delta = 0$ au moyen d'un test t (section 11.1).

11.10 Tests du χ^2

Les tests du χ^2 sont omniprésents en biostatistiques. Nous les présentons ici sans justification, car la justification en est un peu complexe.

11.10.1 Test d'ajustement à une loi donnée

On suppose qu'on a N observations catégorielles, dans d catégories différentes. On note $O_1, \ldots, O_d$ le nombre d'observations qui tombent dans les catégories $1, \ldots, d$, avec $O_1 + \cdots + O_d = N$. On parle des effectifs observés pour les valeurs O_i.

On veut tester l'hypothèse nulle H_0 « la probabilité de la catégorie i est p_i », avec $p_1,\dots,p_d$ des probabilités données a priori (elle ne sont pas calculées à partir des observations).

On calcule les effectifs attendus $E_i = Np_i$. La statistique de test

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

suit, sous H_0, une loi $\chi^2(d-1)$. On obtient un test de risque α en rejetant H_0 quand $T > x_{1-\alpha}^{d-1}$, le quantile $1-\alpha$ de la loi $\chi^2(d-1)$.

La façon dont cette statistique T est construite est assez claire : on prend les carrés des écarts entre valeurs observées et attendues ; la division par E_i, la valeur attendue, a pour rôle de donner moins de poids à une déviation quand elle a lieu dans une catégorie où l'effectif attendu est grand : une erreur de 5 n'a pas la même importance dans une catégorie où on attend 10 observations que dans une catégorie où on en attend 100 !

Notons que si on pose $\widehat{p}_i = \frac{O_i}{N}$, la probabilité de la catégorie i estimée sur les observations, on a

$$T = N \times \sum_i \frac{(\widehat{p}_i - p_i)^2}{p_i},$$

ce qui donne une autre façon de voir cette statistique.

11.10.2 Test d'ajustement à une famille de lois

On considère toujours N observations dans d catégories différentes, récapitulées par des effectifs observées $O_1,\dots,O_d$ avec $\sum_i O_i = N$. On veut tester l'hypothèse nulle H_0 « il existe $\theta = (\theta_1,\dots,\theta_k)$ tel que la probabilité de la catégorie i est $p_i(\theta_1,\dots,\theta_k)$ », où les k paramètres de la loi $\theta_1,\dots,\theta_k$ sont à estimer à partir des observations $O_1,\dots,O_d$.

On estime $\widehat{\theta}_1,\dots,\widehat{\theta}_k$ à partir de $O_1,\dots,O_d$. On calcule les effectifs attendus $E_i = Np_i\left(\widehat{\theta}_1,\dots,\widehat{\theta}_k\right)$. La statistique de test

$$T = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

suit, sous H_0, une loi $\chi^2(d-k-1)$: on retire un degré de liberté par paramètre estimé (indépendant des autres paramètres). On obtient un test de risque α en rejetant H_0 quand $T > x_{1-\alpha}^{d-k-1}$, le quantile $1-\alpha$ de la loi $\chi^2(d-k-1)$.

Exemple 45 On a 100 observations (des entiers positifs) réparties selon la table suivante :

i	0	1	2	3
O_i	47	37	12	4

Table 20. 100 observations dans 4 catégories

On veut tester l'hypothèse selon laquelle ces observations suivent une loi de Poisson $\mathscr{P}(\lambda)$. On commence par estimer λ par la moyenne des observations $\widehat{\lambda} = 0{,}73$. On calcule ensuite $p_0 = e^{-\widehat{\lambda}} = 0{,}48$, $p_1 = \widehat{\lambda}e^{-\widehat{\lambda}} = 0{,}35$, $p_2 = \frac{1}{2}\widehat{\lambda}^2 e^{-\widehat{\lambda}} = 0{,}13$, et enfin, on regroupe ensemble toutes les observations potentiellement supérieures ou égale à 3 : $p_3 = 1 - p_0 - p_1 - p_2 = 0{,}04$. On a donc les effectifs attendus suivants :

i	0	1	2	3
O_i	48	35	13	4

Table 21. Les effectifs attendus

On calcule T = 0,21, qui est à comparer aux quantiles d'une loi de χ^2 à 2 degrés de liberté : on a ici $d = 4$ et $k = 1$ (un paramètre estimé, le paramètre λ de la loi de Poisson), donc $d - k - 1 = 2$. Le quantile de niveau 0,95 est égal à 5,99 : on ne rejette pas H_0. □

11.10.3 Test d'indépendance ou d'homogénéité

Ici les observations sont réparties en $d = a \times b$ catégories, organisées dans une table de contingence à a lignes et b colonnes. On les note O_{ij} pour $i = 1,\dots,a$, et $j = 1,\dots,b$.

Exemple 46 On a demandé à 100 étudiants quel était leur enseignant de stats préféré ; on a enregistré leurs réponses, ainsi que leur sexe. On obtient la table suivante :

Favori	Bob	Joe	Kate
Étudiant	12	22	30
Étudiante	16	8	12

Table 22. Les préférences des étudiants

On veut tester si la variable « sexe » est indépendante de la variable « enseignant favori », ou encore, ce qui est exactement la même chose, si la distribution de la variable « enseignant favori » est la même chez les étudiants que chez les étudiantes ; ou même, troisième façon de voir les choses, si la variable « sexe » est distribuée de façon identique dans les trois catégories définies par l'enseignant favori. □

L'hypothèse nulle à tester est la suivante : « la probabilité de la catégorie (i, j) est égale à $p_i q_j$ », où les p_i et les q_j sont à estimer à partir des observations. On les interprète ainsi : chaque $p_1,\dots,p_a$ est la la probabilité qu'une observation tombe dans la ligne $1,\dots,a$, et chaque $q_1,\dots,q_b$ la probabilité qu'elle tombe dans la colonne $1,\dots,b$.

On les estime à partir des effectifs marginaux :

$$\widehat{p}_i = \frac{1}{N}\sum_j O_{ij},$$
$$\widehat{q}_j = \frac{1}{N}\sum_i O_{ij}.$$

On peut alors calculer les effectifs attendus :

$$E_{ij} = N \times \widehat{p}_i \widehat{q}_j = \frac{1}{N}\left(\sum_j O_{ij}\right)\left(\sum_i O_{ij}\right)$$

et la statistique de test est $T = \sum_{ij}(O_{ij} - E_{ij})^2/E_{ij}$.

Le nombre de paramètres estimés pour les p_i est $a-1$: on estime $p_1,\dots,p_{a_1}$ et p_a s'en déduit ; même chose pour les q_j, on en a estimé $b-1$. Le nombre de degrés de libertés est donc

$$ab - (a - 1 + b - 1) - 1 = (a-1)(b-1).$$

Exemple 47 Terminons l'exemple précédent. Il y a plusieurs façons de mener le calcul des effectifs attendus, on en choisit une : il y a 64 garçons et 36 filles, donc on attend dans chaque catégorie des proportions 0,64 et 0,36 de chacun des deux sexes ; par exemple, parmi les 28 étudiants qui préfèrent Bob, on attend $0{,}64 \times 28 = 17{,}92$ garçons et $0{,}36 \times 28 = 10{,}08$ filles. De cette façon on trouve la table suivante pour les effectifs attendus :

Favori	Bob	Joe	Kate
Étudiant	17,92	19,20	26,88
Étudiante	10,08	10,80	15,12

Table 23. Les effectifs attendus

d'où on calcule

$$T = \frac{(12-17{,}92)^2}{17{,}92} + \frac{(22-19{,}20)^2}{19{,}20} + \cdots = 8{,}44.$$

Il faut comparer T aux quantiles d'une loi de $\chi^2(2)$: le quantile de niveau 0,95 est égal à 5,99, on rejette donc H_0 (si on fait le test au risque usuel $\alpha = 0{,}05$). □

11.10.4 Conditions de validité

Le fait que la loi de la statistique de test suive, sous H_0, une loi de χ^2, vient d'approximations de données entières par une loi normale ; comme à l'usuel, ces approximations sont d'autant meilleures que les effectifs sont importants.

On rencontre souvent les conditions de Cochran :

Conditions de Cochran.
On considère l'approximation par un χ^2 valide si tous les effectifs attendus E_i sont supérieurs où égaux à 5.

On peut être conduit à « fusionner » des catégories pour obtenir des effectifs plus important. Les conséquences d'une telle pratique ne sont pas claires : le choix de fusionner ou non des catégories étant fait en fonction des observations, *cela change la distribution de la statistique.*

Pearson a proposé une façon intéressante de faire des tests quand les effectifs attendus sont petits :

Test $N-1$ **de Pearson**
On utilise la statistique de test suivante :

$$T_1 = \frac{N-1}{N} T = \frac{N-1}{N} \sum_i \frac{(O_i - E_i)^2}{E_i}$$

et on la compare aux quantiles du χ^2 avec « le nombre habituel » de degrés de libertés. Ce test est valide dès que tous les effectifs attendus E_i sont supérieurs où égaux à 1.

Cette procédure donne de très bons résultats, et semble même dans certains cas préférable à la vérification des conditions de Cochran [1].

1. Ian Cambell, 2007. Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Statis. Med 26.

11.11 Exercices

Exercice 1 La durée de gestation des vaches est approximativement normale, de moyenne 280 jours avec un écart type de 6,75 jours.[1]

L'éleveur Martial Bottafoin note soigneusement la date d'insémination de ses reproductrices et la date de leur vélage. Il a pu ainsi calculer 50 durées de gestations x_i, exprimées en jours.

On a $\sum x_i = 13990$ et $\sum x_i^2 = 3917264$.

1. Calculer les estimations de la durée moyenne de gestation et de son écart type.

2. Ces estimations sont-elles compatibles avec les valeurs données ?

[1] Bougler et Derveaux, Étude des durées de gestation dans l'espèce bovine, (1969).

Exercice 2 Jesus Carnicero, le voisin de Martial Bottafoin, élève une race à viande (des Charolaises), réputée avoir une durée de gestation plus élevée que les laitiaires de Martial.

Pour 50 durées de gestations exprimées en jour, Jesus Carnicero calcule $\sum y_i = 14462$ et $\sum y_i^2 = 4185612$.

1. L'écart-type observé par Jesus Carnicero est-il compatible avec l'écart-type donné a priori ? Et avec l'écart-type observé par Martial Bottafoin ?

2. Comparez les durées moyennes de gestation chez les deux éleveurs.

Un quantile utile : $F_{0.975}^{49,49} = 1{,}76$.

Exercice 3 Dessiner l'allure de la zone d'acceptation du test F d'égalité entre deux variances estimées s_x^2 et s_y^2 au risque $\alpha = 5\%$ avec comme degrés de libertés : $d_1 = 100$ et $d_2 = 5$; $d_1 = d_2 = 100$; $d_1 = d_2 = 5$.

On mettra en abscisse la valeur de s_x^2 et en ordonnée celle de s_y^2.

Exercice 4 On estime une proportion p dans un échantillon de taille n par $\widehat{p} = \frac{x}{n}$ où x est le nombre d'observations répondant à un certain critère. On pose $\Phi(x) = \arcsin\left(\sqrt{x}\right)$.

On sait que la loi de $\widehat{p}$ est approximativement $\mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ et que la loi de $\Phi(\widehat{p})$ est approximativement $\mathcal{N}\left(\Phi(p), \frac{1}{4n}\right)$.

On cherche à estimer si une maladie donnée touche indifféremment les patients de deux sexes. Sur un échantillon de 40 patients, on observe 25 hommes.

1. Donner un intervalle de confiance sur la proportion p de patients mâles, d'abord par la méthode de Wald (approximation normale de $\widehat{p}$), puis en utilisant l'approximation normale de $\Phi(\widehat{p})$.

2. Tester si $p = \frac{1}{2}$, par les deux méthodes.

Exercice 5 * On considère toujours la transformation d'une proportion observée dans un échantillon de taille n, $\Phi(\widehat{p}) = \arcsin\left(\sqrt{\widehat{p}}\right)$ et son approximation normale.

On va calculer le risque β et la puissance du test bilatéral $H_0 : p = p_0$ contre $H_1 : p \neq p_0$ au risque $\alpha = 5\%$, basé sur $\Phi\left(\widehat{p}\right)$.

1. Montrer qu'on rejette H_0 quand

$$2\sqrt{n}\left|\Phi(\widehat{p}) - \Phi(\widehat{p}_0)\right| > 1{,}96.$$

2. On suppose que $p = p_1$. Montrer que la puissance du test ne dépend que de n et de $\Delta = 2\left(\Phi(p_1) - \Phi(p_0)\right)$.

3. Montrer que pour $\Delta > 0$ et $\Delta\sqrt{n}$ assez grand on peut approcher β par $\mathbb{P}\left(Z \leqslant 1{,}96 - \Delta\sqrt{n}\right)$, où $Z \sim \mathcal{N}(0,1)$. En déduire que $z_\beta \approx 1{,}96 - \Delta\sqrt{n}$.

4. On suppose qu'on a $p_0 = 0{,}5$ et $p_1 = 0{,}515$. À partir de quelle taille d'échantillon a-t-on une puissance de 80% ?

Chapitre 12

Analyse de la variance à un facteur

12.1 Introduction

On sait comparer deux moyennes mesurées dans deux groupes différents, par exemple dans un groupes de patients traités et dans un groupe de patients non traités : c'est le test t.

Comment faire quand on n'a plus deux groupes, mais p groupes, correspondant à p traitements distincts, et que la question posée est de savoir s'il y a une différence entre les traitements ? On peut faire des comparaisons deux à deux – il en faut $\frac{1}{2}p(p-1)$; cela pose plusieurs problèmes (tests multiples, puissance)... un test unique et « global » est préférable.

12.1.1 Le problème des tests multiples

Dans une étude statistique quelconque, si on fait plusieurs tests, se pose le problème du choix du niveau α auquel on fait ces tests ; et ceci, que les tests soient k comparaisons entre groupes, ou qu'il s'agisse de tests d'association entre une maladie donnée et k facteurs environnementaux différents.

Dans le cas des comparaisons entre groupes, les tests t envisagés ne sont pas indépendants, ce qui complique l'analyse. Nous allons nous cantonner dans un premier temps au cas de tests indépendants.

Supposons donc qu'on fasse k tests indépendants chacun au risque α. On note $H_0^1, \ldots, H_0^k$ les hypothèses nulles de chacun de ces tests. Supposons que toutes ces hypothèses nulles soient vraies ; la probabilité que le i^{e} test rejette (à tort) H_0^i est α ; mais la probabilité qu'après les k tests effectués, une des hypothèses nulles (ou plus) ait été rejetée à tort est bien supérieure !

Si on fait une analogie avec le jet d'un dé à 20 faces, la probabilité de tirer 1 est de $1/20 = 0{,}05$ quand on ne fait qu'un jet, mais si on fait 10 jets, la probabilité de tirer au moins un 1 est clairement plus importante.

Le calcul est simple : la probabilité de ne pas rejetter H_0^i est $(1-\alpha)$, donc la probabilité de ne rejetter aucun des H_0^i est $(1-\alpha)^k$; la probabilité d'en avoir rejetté au moins 1 est $1-(1-\alpha)^k$.

Correction de Šidák

Si on veut que le risque global soit α_0 (par exemple $\alpha_0 = 5\%$), il faut que $1-(1-\alpha)^k = \alpha_0$, c'est-à-dire qu'on fera chaque test au niveau

$$\alpha = 1-(1-\alpha_0)^{\frac{1}{k}}.$$

La table suivante donne les valeurs de α pour $\alpha_0 = 5\%$ et k variant de 1 à 12.

k	α	k	α	k	α
1	5,00%	5	1,02%	9	0,57%
2	2,53%	6	0,85%	10	0,51%
3	1,70%	7	0,73%	11	0,47%
4	1,27%	8	0,64%	12	0,43%

Correction de Bonferroni

Le lecteur vigilant aura remarqué que dans la table donnée ci-dessus, on a $\alpha \approx \frac{1}{k}5\%$. L'approximation $1-(1-\alpha)^k \approx k\alpha$ (pour α assez petit), permet de vérifier qu'en effet, on aura $\alpha_0 \approx k\alpha$ et qu'il faut donc prendre $\alpha = \frac{1}{k}\alpha_0$.

Si on veut que le risque global soit α_0, on fera chaque test au niveau $\alpha = \frac{1}{k}\alpha_0$.

Si les tests ne sont pas indépendants, cette correction reste valable : elle peut être dérivée de l'*inégalité de Boole* :

$$\mathrm{P}\left(\bigcup_{i=1}^{k} \mathrm{A}_i\right) \leq \sum_{i=1}^{k} \mathrm{P}\left(\mathrm{A}_i\right).$$

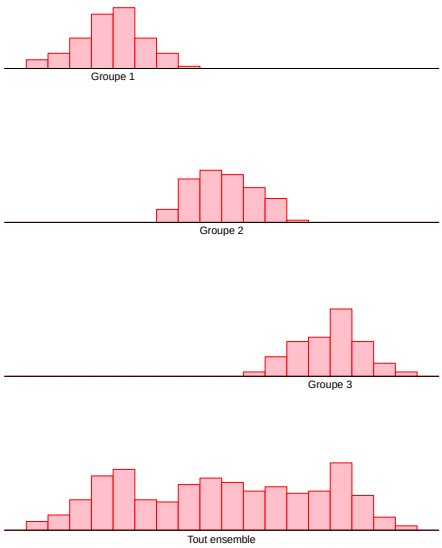
Si on prend pour A_i l'événement $(p_i < \alpha)$ (où p_i est le degré de signification du i^{e} test), on obtient donc que la probabilité de rejetter au moins une des hypothèses H_0^i est inférieur ou égale à $k\alpha$.

Avec $\alpha = \frac{1}{k}\alpha_0$, le risque global est plus petit que $k\alpha = \alpha_0$.

Pour revenir à notre problème, on pourrait choisir de comparer tous les groupes 2 à 2, en adaptant le niveau auquel est fait chaque test ; l'inconvénient serait une perte importante de puissance.

12.1.2 Utiliser la variance pour ne faire qu'un seul test

En imaginant un cas extrême, où les $p = 3$ groupes sont très distincts, comme illustré par les histogrammes ci-dessous :



On voit que dans chacun des groupes, les données sont assez peu étalées : la variance est faible ; si en revanche on considère toutes les données ensemble, l'étalement est beaucoup plus important : la variance est élevée.

L'analyse de variance est basée sur cette idée que si les groupes ont des moyennes distinctes, la variance totale est plus importante que la variance dans chaque groupe.

L'anglais *analysis of variance* a donné naissance à la très populaire abréviation *anova*, que nous utiliserons par la suite.

12.1.3 Plan du chapitre

Nous allons d'abord présenter le modèle de l'anova ; la différence entre une expérience planifiée et une expérience non planifiée sera soulignée avant la construction du test ; ensuite, nous considèrerons l'anova comme une méthode de modélisation, et nous verrons comment estimer les paramètres du modèle retenu. Dans les sections 12.11, 12.12 et 12.13, nous verrons comment comparer entre eux des modèles intermédiaires.

Enfin, dans la section 12.14.2, nous verrons un modèle différent qui conduit aux mêmes calculs que l'anova classique.

12.2 Le modèle de l'anova

On a n observations réparties en p groupes d'effectifs $n_1, \ldots, n_p$. Si tous les n_i ont la même valeur on dira que les données sont équiréparties.

Les observations sont notées X_{ij} avec $i = 1, \ldots, p$ et $j = 1, \ldots, n_i$: X_{ij} est la j^{e} observation du groupe i.

Dans le modèle de l'anova chaque X_{ij} suit une loi normale dont l'espérance dépend du groupe i : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. La variance de cette loi ne dépend pas du groupe ; cette hypothèse d'homogénéité des variances est appelée l'hypothèse d'*homoscédasticité* (du grec *skedasê* : éparpillement, dispersion).

On va tester l'hypothèse nulle $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$, contre H_1 : au moins deux des μ_i sont différents.

Pour construire le test, on se placera sous H_0 ; la variance sera estimée de deux façons différentes : une estimation en tenant compte des groupes, et l'autre sans en tenir compte. Sous H_1 l'estimation qui ne tient pas compte des groupes va donner une valeur plus élevée que celle qui en tient compte : on testera si elle est significativement plus élevée, ou non.

12.3 Exemples : expériences planifiées et non planifiées

On utilise souvent l'analyse de la variance dans deux situations distinctes : les expériences planifiées (études prospectives) et les expériences non-planifiées (études rétrospectives).

La théorie s'écrit en général pour le cas planifié, cependant en pratique l'utilisation de données issues d'expériences non planifiées est courante. Dans les deux cas, le modèle est le même ; seule l'interprétation du résultat des tests diffère.

12.3.1 Expériences planifiées

C'est le cas en particulier des essais thérapeutiques : on applique p traitements différents à n individus répartis en p groupes d'effectif $n_1, \ldots, n_p$. La répartition des individus dans chacun des groupes est faite par l'expérimentateur au moyen d'un tirage au sort. Cette opération est appelée la « randomisation » du test. Nous verrons plus tard son importance (paragraphe 12.10.3).

12.3.2 Un exemple

On compare l'effet de trois traitements sur le paludisme, en mesurant le temps de clairance parasitaire chez des patients symptomatiques, répartis de façon aléatoire en trois groupes.

On a les résultats suivants (en heures).

Traitement 1			Traitement 2			Traitement 3		
33	22	54	49	60	92	62	40	92
55	68	41	73	51	94	71	112	65
96	78	48	62	107	67	88	85	95
75	65	65	90			119		
$x_{1+} = 700$			$x_{2+} = 745$			$x_{3+} = 829$		

On peut calculer les moyennes de chaque groupe : $x_{1\bullet} = \frac{x_{1+}}{n_1} = \frac{700}{12} = 58{,}333$, $x_{2\bullet} = \frac{745}{10} = 74{,}5$, et $x_{3\bullet} = \frac{829}{10} = 82{,}9$.

On veut tester l'hypothèse nulle : le temps de clairance parasitaire est le même pour les trois traitements, contre l'hypothèse alternative : ces traitements ont des effets différents. Le test se fera au risque $\alpha = 0{,}05$.

12.4 Expériences non planifiées

C'est le cas en particulier de la recherche de facteurs qui modifient une variable d'intérêt : on recrute n individus dans la population, on observe (par un questionnaire sur leurs habitudes de vie, par génotypage si on cherche un facteur génétique , etc) à quel niveau se situe le facteur chez chacun d'eux.

12.4.1 Un exemple

Considérons un gène d'allèles B et b ; on veut tester l'hypothèse selon laquelle il module la concentration sanguine d'une certaine protéine. On a recruté une centaine d'individus, on les a génotypés pour le gène considéré et on a mesuré la concentration sanguine chez chacun d'eux.

Génotype BB					Génotype Bb					Génotype bb				
3,62	3,67	5,42	3,92	5,47	4,41	5,76	4,34	5,50	5,81	4,34	5,63	5,36	5,48	5,79
5,44	5,82	3,74	6,47	4,60	5,23	4,00	4,01	3,86	3,87	7,16	5,25	2,27	5,16	5,90
2,76	4,24				5,00	4,85	6,57	5,22	4,61	5,04	5,31	4,51	4,69	2,55
					5,82	4,88	4,38	7,21	5,01	6,78	4,09	6,91	1,40	6,81
					4,89	4,93	7,27	2,29	6,22	5,21	5,80	6,16	4,23	6,31
					4,93	6,19	4,83	6,89	4,90	5,51	4,14	7,47	3,86	4,47
					5,41	4,09	5,64	4,15	6,16	3,33	4,36	6,77	5,15	3,39
					5,47	6,67	5,10	6,09		3,97	5,39	4,02	3,14	5,63
										5,00	4,98	4,42	4,36	6,95
										3,23	4,71	5,12	5,38	
$n_1 = 12, x_{1+} = 55{,}17$					$n_2 = 39, x_{2+} = 202{,}46$					$n_3 = 49, x_{3+} = 242{,}89$				

Dans ce cas, les valeurs de n_1, n_2 et n_3 peuvent servir à estimer la fréquence des trois génotypes (pourvu que les individus soient représentatifs de la population générale).

On peut calculer la valeur moyenne dans chaque groupe : $x_{1\bullet} = 4{,}60$, $x_{2\bullet} = 5{,}19$ et $x_{3\bullet} = 4{,}96$. On veut tester l'hypothèse nulle : le génotype n'a pas d'influence sur la concentration, contre l'hypothèse alternative : il existe une différence entre les groupes.

Cet exemple sera traité en exercice.

La différence essentielle entre cette situation et l'expérience planifiée c'est que dans l'expérience planifiée le facteur n'est pas une variable aléatoire « incontrôlée » ; il est *fixé* par l'expérimentateur, ce qui permet la randomisation, impossible dans le cas de l'expérience non-planifiée.

12.5 Reformulation comme un modèle linéaire

Notre modèle est donc $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. En vue d'une plus grande généralité on peut le reformuler en posant

$$\boxed{X_{ij} = \mu + \alpha_i + E_{ij},}$$

où les E_{ij} sont indépendantes de loi $\mathcal{N}(0,\sigma^2)$.

En posant $\alpha'_i = \alpha_i + c$ et $\mu' = \mu - c$ on obtient un modèle équivalent $X_{ij} = \mu' + \alpha'_i + E_{ij}$: pour éviter cette multitude de modèles, on imposera $\sum_i n_i\alpha_i = 0$.

On a alors $\sum_i n_i\mu_i = \sum_i n_i(\mu+\alpha_i) = n\mu$ (on rappelle que $n = \sum_i n_i$). Dans la suite, nous utiliserons indifféremment les deux notations.

12.5.1 Notations

Pour alléger les calculs dans la suite, on notera la somme obtenue en faisant varier un indice en remplaçant cet indice par un signe « plus » :

$$X_{i+} = \sum_{j=1}^{n_i} X_{ij},$$

$$X_{++} = \sum_{ij} X_{ij} = \sum_{i=1}^{p} X_{i+}.$$

De même, la moyenne obtenue en faisant varier un indice est notée en remplaçant cet indice par un point :

$$X_{i\bullet} = \frac{1}{n_i} X_{i+},$$

$$X_{\bullet\bullet} = \frac{1}{n} X_{++} = \frac{1}{n}\sum_{i=1}^{p} n_i X_{i\bullet}.$$

Ainsi, $X_{i\bullet}$ est la moyenne empirique de l'échantillon issu du groupe i, et $X_{\bullet\bullet}$ est la moyenne empirique générale.

On note également les sommes de carrés des observations :

$$X^2_{i+} = \sum_{j=1}^{n_i} X^2_{ij}$$

$$X^2_{++} = \sum_{ij} X^2_{ij}$$

12.5.2 Estimation des paramètres du modèle linéaire

On a

$$X_{i+} = n_i\mu + n_i\alpha_i + E_{i+}$$

où $E_{i+} = \sum_j E_{ij}$ suit une loi $\mathcal{N}(0, n_i\sigma^2)$. On a donc

$$\begin{aligned} X_{++} &= \sum_i n_i\mu + \sum_i n_i\alpha_i + \sum_i E_{i+} \\ &= n\mu + E_{++} \end{aligned}$$

avec $E_{++} \sim \mathcal{N}(0, n\sigma^2)$. On a donc un estimateur pour μ,

$$\boxed{\widehat{\mu} = \frac{1}{n} X_{++} = X_{\bullet\bullet}}$$

de loi $\mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$.

On en déduit un estimateur pour α_i :

$$\boxed{\widehat{\alpha_i} = X_{i\bullet} - X_{\bullet\bullet}}$$

Il est de loi $\mathcal{N}\left(\alpha_i, \frac{1}{n_i}\sigma^2\right)$.

Notons que $X_{i\bullet}$ est un estimateur de $\mu_i = \mu + \alpha_i$; on appelle $X_{ij} - X_{i\bullet}$, qui est l'écart à la moyenne de groupe μ_i de la j-ème mesure du groupe i, *le résidu* ε_{ij}.

Reste l'estimation de σ^2. Les sommes de carrés de l'anova vont servir à ça.

12.6 Les sommes de carrés : totaux, factoriels, résiduels

On définit la somme des carrés des résidus du groupe i, ou somme des carrés des écarts à la moyenne, ou en abrégé *somme des carrés* par

$$\boxed{SC_i = \sum_{j=1}^{n_i} (X_{ij} - X_{i\bullet})^2 = \sum_{j=1}^{n_i} X_{ij}^2 - n_i (X_{i\bullet})^2 = X_{i+}^2 - \frac{1}{n_i} (X_{i+})^2}$$

La variance empirique (corrigée) du groupe i est le *carré moyen* $CM_i = \frac{1}{n_i - 1} SC_i$. On connaît la loi de la variance empirique, on en déduit que $SC_i \sim \sigma^2 \chi^2 (n_i - 1)$.

On sera amené plus tard à fusionner des groupes entre eux; la somme des carrés dans le groupe constitué par la fusion des groupes i et k sera par exemple notée $SC_{i,k}$.

12.6.1 La somme des carrés résiduels

La *somme des carrés résiduels* est définie par

$$\boxed{SCR = \sum_{i=1}^{p} SC_i.}$$

Une somme de $\chi^2(d_i)$ indépendants suivant une loi $\chi^2(d = \sum_i d_i)$, on a $SCR \sim \sigma^2 \chi^2 (n - p)$.

Le *carré moyen résiduel* : $CMR = \frac{1}{n-p} SCR$ est donc un estimateur sans biais de σ^2.

12.6.2 La somme des carrés totaux

La *somme des carrés totaux* est $SCT = SC_{1,\ldots,p}$: la somme des carrés calculée sur les données rassemblées en un seul groupe.

On a la formule de décentrage suivante :

$$\boxed{SCT = \sum_{ij} (X_{ij} - X_{\bullet\bullet})^2 = \sum_{ij} X_{ij}^2 - n (X_{\bullet\bullet})^2 = X_{++}^2 - \frac{1}{n} (X_{++})^2}$$

Sous H_0, SCT suit une loi $\sigma^2 \chi^2 (n - 1)$. Le *carré moyen total* : $CMT = \frac{1}{n-1} SCT$ est donc (sous H_0) un estimateur sans biais de σ^2.

12.6.3 La somme des carrés factoriels

On pose

$$\boxed{SCF = \sum_{i=1}^{p} n_i (X_{i\bullet} - X_{\bullet\bullet})^2 = \sum_{i=1}^{p} n_i (X_{i\bullet})^2 - \frac{1}{n} (X_{++})^2.}$$

On a également

$$\boxed{\mathrm{SCF} = \frac{1}{n}\sum_{i<k} n_i n_k (\mathrm{X}_{i\bullet} - \mathrm{X}_{k\bullet})^2.}$$

SCF est l'abréviation de *somme des carrés factoriels* : « factoriel » correspond à « facteur », un terme équivalent à « groupe ».

On remarque d'abord que SCF est égale à $\sum_i n_i (\widehat{\alpha_i})^2$.

On peut également remarquer que SCF est la quantité obtenue en remplaçant, dans SCT, chaque X_{ij} par la moyenne de groupe $\mathrm{X}_{i\bullet}$. C'est donc la somme des carrés totaux qu'on obtiendrait en « condensant » chaque groupe sur sa moyenne, en « oubliant » sa variance interne.

Considérons une variable aléatoire « fictive » A telle que $\mathbb{P}(\mathrm{A} = \mu_i) = p_i = \frac{n_i}{n}$ (la probabilité qu'une mesure prise au hasard parmi les X_{ij} tombe dans le groupe i).

L'espérance de A est $\frac{1}{n}\sum_i n_i \mu_i = \mu$ et sa variance

$$\frac{1}{n}\sum_i n_i (\mu_i - \mu)^2 = \frac{1}{n}\sum_i n_i \alpha_i^2,$$

en utilisant la notation $\mu_i = \mu + \alpha_i$.

On voit que SCF est évocateur de la variance empirique d'une telle variable.

On va voir plus bas que, sous H_0, $\frac{1}{\sigma^2}\mathrm{SCF}$ suit une loi $\chi^2(p-1)$.

Il en découle que le *carré moyen factoriel* : $\mathrm{CMF} = \frac{1}{p-1}\mathrm{SCF}$ est un estimateur sans biais de σ^2 (toujours sous H_0).

12.7 Le théorème fondamental de l'analyse de la variance

Théorème 43 *Avec les notations introduites ci-dessus, on a*

$$\boxed{\mathrm{SCT} = \mathrm{SCF} + \mathrm{SCR}.}$$

Les variables aléatoires SCF *et* SCR *sont indépendantes. On a* $\mathrm{SCR} \sim \sigma^2\chi^2(n-p)$*, et sous* H_0*,* $\mathrm{SCT} \sim \sigma^2\chi^2(n-1)$ *et* $\mathrm{SCF} \sim \sigma^2\chi^2(p-1)$.

Le quotient

$$\mathrm{F} = \frac{\mathrm{CMF}}{\mathrm{CMR}} = \frac{\mathrm{SCF}/(p-1)}{\mathrm{SCR}/(n-p)}$$

suit une loi F *de degrés de liberté* $p-1$ *et* $n-p$.

On obtient un test de risque α *en rejetant* H_0 *quand* $\mathrm{F} > \mathrm{F}_{1-\alpha}^{p-1,n-p}$.

Montrons d'abord le premier point. On a

$$\begin{aligned}(\mathrm{X}_{ij} - \mathrm{X}_{\bullet\bullet})^2 &= (\mathrm{X}_{ij} - \mathrm{X}_{i\bullet} + \mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})^2 \\ &= (\mathrm{X}_{ij} - \mathrm{X}_{i\bullet})^2 + 2(\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})(\mathrm{X}_{ij} - \mathrm{X}_{i\bullet}) + (\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})^2\end{aligned}$$

On fait la somme pour j variant de 1 à n_i :

$$\begin{aligned}\sum_{j=1}^{n_i}(\mathrm{X}_{ij} - \mathrm{X}_{\bullet\bullet})^2 &= \sum_{j=1}^{n_i}(\mathrm{X}_{ij} - \mathrm{X}_{i\bullet})^2 + 2(\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})\sum_{j=1}^{n_i}(\mathrm{X}_{ij} - \mathrm{X}_{i\bullet}) + n_i(\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})^2 \\ &= \sum_{j=1}^{n_i}(\mathrm{X}_{ij} - \mathrm{X}_{i\bullet})^2 + n_i(\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})^2 \\ &= \mathrm{SC}_i + n_i(\mathrm{X}_{i\bullet} - \mathrm{X}_{\bullet\bullet})^2,\end{aligned}$$

car $\sum_j (\mathrm{X}_{ij} - \mathrm{X}_{i\bullet}) = 0$.

On somme maintenant pour i variant de 1 à p :

$$\begin{aligned} \sum_{ij}(X_{ij}-X_{\bullet\bullet})^2 &= \sum_{i=1}^{p} SC_i + \sum_{i=1}^{p} n_i(X_{i\bullet}-X_{\bullet\bullet})^2 \\ SCT &= SCR + SCF\,. \end{aligned}$$

Nous admettrons l'indépendance de SCR et SCF. Nous avons déjà vu plus haut que $SCR \sim \sigma^2\chi^2(n-p)$, et sous H_0, $SCT \sim \sigma^2\chi^2(n-1)$; SCR et SCF étant indépendantes, il en découle que $SCF \sim \sigma^2\chi^2(p-1)$.

Les derniers points sont clairs ; il faut noter qu'on choisit d'effectuer un test unilatéral ; en effet si H_1 est vrai, SCF ne suit plus une loi $\sigma^2\chi^2(p-1)$, et par le phénomène d'augmentation de la variance que nous avons décrit au début du chapitre, SCF aura tendance à prendre des valeurs plus élevées (voir le paragraphe 12.8).

12.7.1 Remarque sur la construction du test

Avant de construire le test pièce par pièce, en donnant les idées qui en sont à l'origine, on pouvait s'attendre à une comparaison de CMT et CMR, la variance totale et la variance résiduelle ; mais ces deux variables aléatoires ne sont pas indépendantes, et leur quotient ne suit donc pas une loi F. On pourrait définir une nouvelle loi, spécifique à l'analyse de variance : la loi du quotient CMT/CMR est bien définie et pourrait être tabulée comme l'est la loi F.

Mais la loi F étant déjà connue, il est plus simple d'utiliser plutôt CMF/CMR. Bien sûr, les deux façons de faire donnent des tests équivalents.

On peut également voir les choses ainsi : puisque que $SCF = \sum_i (\widehat{\alpha_i})^2$, de grandes valeurs de SCF (et donc de F) plaident pour l'hypothèse alternative.

12.8 Compléments sur l'estimation des paramètres du modèles

On va en particulier se pencher sur les lois des sommes de carrés.

12.8.1 Sous l'hypothèse nulle

Sous H_0, le théorème 43 donne la loi de SCR, SCF et SCT ; on en déduit que

$$\text{sous } H_0,\ \begin{cases} E(CMR) &= \sigma^2 \\ E(CMF) &= \sigma^2 \\ E(CMT) &= \sigma^2 \end{cases}$$

Si on conserve l'hypothèse nulle, la moyenne globale est estimée par $\widehat{\mu} = X_{\bullet\bullet}$; la loi de $\widehat{\mu}$ est $\mathcal{N}(\mu, \frac{1}{n}\sigma^2)$. La variance σ^2 peut être estimée par CMT (avec $n-1$ degrés de liberté).

On a donc un intervalle de confiance pour μ, au niveau $1-\alpha$ (voir la section 9.1) :

$$\left[\widehat{\mu} \pm t^{n-1}_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}CMT}\right].$$

12.8.2 Sous l'hypothèse alternative

Nous n'écrirons pas les lois de SCF et SCT sous H_1 ; cependant en utilisant les formules de décentrage, on peut facilement calculer les valeurs attendues.

Par exemple, pour SCT :

$$SCT = \sum_{ij} X_{ij}^2 - \frac{1}{n}\left(\sum_{ij} X_{ij}\right)^2.$$

On a $\mathrm{E}\left(\mathrm{X}_{ij}^2\right) = \mathrm{var}(\mathrm{X}_{ij}) + \mathrm{E}(\mathrm{X}_{ij})^2 = \sigma^2 + \mu_i^2$; d'autre part la somme des X_{ij} suit une loi $\mathcal{N}(n_i\mu_i, n\sigma^2)$, d'où

$$\mathrm{E}\left(\left(\sum_{ij}\mathrm{X}_j\right)^2\right) = n\sigma^2 + \left(\sum_i n_i\mu_i\right)^2.$$

On en tire

$$\begin{aligned}
\mathrm{E(SCT)} &= \sum_{ij}\mathrm{E}\left(\mathrm{X}_{ij}^2\right) - \frac{1}{n}\left(n\sigma^2 + \left(\sum_i n_i\mu_i\right)^2\right) \\
&= n\sigma^2 + \sum_i n_i\mu_i^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2 + \sum_i n_i(\mu_i - \mu)^2 \\
&= (n-1)\sigma^2 + \sum_i n_i\alpha_i^2,
\end{aligned}$$

où on a utilisé $\mu_i = \mu + \alpha_i$, $\sum_i n_i\mu_i = n\mu$ et la formule de décentrage $\sum_i n_i(\mu_i - \mu)^2 = \sum_i n_i\mu_i^2 - n\mu^2$ (voir aussi paragraphe 12.6.3).

On a donc :

$$\text{sous } \mathrm{H}_1, \left\{\begin{array}{lcl} \mathrm{E(CMR)} &=& \sigma^2 \\ \mathrm{E(CMF)} &=& \sigma^2 + \frac{1}{p-1}\sum_i n_i\alpha_i^2 \\ \mathrm{E(CMT)} &=& \sigma^2 + \frac{1}{n-1}\sum_i n_i\alpha_i^2 \end{array}\right.$$

La valeur de $\sum_i n_i\alpha_i^2$ n'a pas d'intérêt en elle-même dans le cas planifié, puisqu'elle dépend du plan d'expérience. Elle sert simplement à quantifier l'écart à l'hypothèse nulle.

En cas de rejet de l'hypothèse nulle, l'expérimentateur sera intéressé par les estimations $\widehat{\mu}_i$ des moyennes de groupes, qui sont comme on l'a vu $\widehat{\mu}_i = \mathrm{X}_{i\bullet}$. La loi de $\widehat{\mu}_i$ est la loi normale $\mathcal{N}(\mu_i, \frac{1}{n_i}\sigma^2)$.

La variance σ^2 est estimée par $\widehat{\sigma}^2 = \mathrm{CMR} \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$. On a donc l'intervalle de confiance au risque α sur μ_i (voir la section 9.1) :

$$\left[\widehat{\mu}_i \pm t_{1-\frac{\alpha}{2}}^{n-p}\sqrt{\frac{1}{n_i}\mathrm{CMR}}\right].$$

12.9 Réalisation pratique du test

En pratique, on utilise un ordinateur et un logiciel de statistiques ! Cependant, un Crusoé statisticien perdu sur une île déserte, ou un étudiant condamné à faire ses exercices avec une simple calculette gagneront du temps en utilisant les formules de décentrage.

On note x_{ij} les observations réalisés. On rappelle qu'on note $x_{i+} = \sum_j x_{ij}$, la somme des observations du groupe i. On note également x_{i+}^2, la somme des carrés des observations du groupe i (et non le carré de x_{i+}), et

$$\begin{aligned}
x_{++} &= \sum_{ij} x_{ij} \\
x_{++}^2 &= \sum_{ij} x_{ij}^2.
\end{aligned}$$

On a alors :

$$\begin{aligned}
\mathrm{SC}_i &= x_{i+}^2 - \tfrac{1}{n_i}(x_{i+})^2, \\
\text{et}\quad \mathrm{SCT} &= x_{++}^2 - \tfrac{1}{n}(x_{++})^2.
\end{aligned}$$

On peut donc calculer $\mathrm{SCR} = \sum_i \mathrm{SC}_i$ et $\mathrm{SCF} = \mathrm{SCT} - \mathrm{SCR}$.

On présente traditionnellement l'analyse de la variance dans une table comme celle-ci :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	SCF	p - 1	CMF = SCF/(p-1)	F = CMF/CMR
résidus	SCR	n - p	CMR = SCR/(n-p)	
Total	SCT	n - 1	CMT = SCT/(n-1)	

Attention, on a bien SCT = SCF + SCR, et les degrés de liberté se somment également, mais on n'a pas CMT = CMF + CMR !

Il y a quelques termes de jargon dont on ne peut faire l'économie tant ils sont répandus : CMF est appelé *la variance factorielle* ou *variance inter-groupes* ou *variance inter* ; CMR est appelé *la variance résiduelle* ou *variance intra*. CMT (qui n'apparaît pas toujours dans la table) est *la variance totale*.

12.9.1 Solution de l'exemple du paludisme

Reprenons les données du paragraphe 12.3.2. On a $n_1 = 12$, $n_2 = 10$, $n_3 = 10$, et $n = 32$. On peut compléter le tableau de données en y ajoutant les x_{i+} et les x_{i+}^2 (qui sont en fait tout ce dont on a besoin pour faire l'anova) :

	Traitement 1			Traitement 2			Traitement 3		
	33	22	54	49	60	92	62	40	92
	55	68	41	73	51	94	71	112	65
	96	78	48	62	107	67	88	85	95
	75	65	65	90			119		
x_{i+}	700			745			829		
x_{i+}^2	45498			59113			73873		

On calcule également $x_{++} = 2274$ et $x_{++}^2 = 178484$.

On a ensuite $SC_1 = x_{1+}^2 - \frac{1}{n_1}(x_{2+})^2 = 45498 - \frac{1}{12}700^2 = 4664{,}7$, $SC_2 = 59113 - \frac{1}{10}745^2 = 3610{,}5$, et $SC_3 = 73873 - \frac{1}{10}829^2 = 5148{,}9$. La somme des carrés résiduels est SCR = 13424,1.

D'autre part $SCT = 178484 - \frac{1}{32}2274^2 = 16887{,}9$. On en déduit SCF = SCT − SCR = 3463,8.

On peut remplir la table d'analyse de la variance.

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	3 463,8	2	1 731,9	F = 3,74
résidus	13 424,1	29	462,9	
Total	16 887,9	31	544,8	

On compare la valeur de F calculée au quantile 0,95 de F(2,29) : il n'est pas dans la table mais on voit aisément à partir des degrés de libertés voisins qu'il est environ de 3,3. On rejette donc l'hypothèse nulle.

Nous avons déjà calculé la moyenne du temps de clairance parasitaire dans chacun des groupes ; sa variance est estimée par CMR = 462,9 (soit un écart-type estimé à 21,5).

12.10 Points divers

12.10.1 Tests sur les variances

Si l'hypothèse d'égalité des variances entre les groupes est douteuse, on peut souhaiter la mettre à l'épreuve avant de pratiquer le test de l'anova.

Une idée simple est de réaliser une série de tests F bilatéraux (cf 11.6). La statistique de test pour comparer les variances entre les groupes i et k est

$$\frac{CM_i}{CM_k} \sim F(n_i - 1, n_k - 1).$$

Cependant là encore se pose la question des tests multiples.

La bonne solution est d'utiliser une procédure de test « globale ». Il en existe plusieurs, disponibles dans les logiciels de statistiques (procédure de Bartlett, de Levene). Nous ne les aborderons pas dans ce cours.

12.10.2 Expérience planifiée et interprétation du résultat

Quand l'expérience est planifiée, un test positif peut s'interpréter comme une preuve d'une différence entre les différents groupes considérés, due aux différences entre les traitements appliqués dans chaque groupe.

Dans la pratique, il est cependant fréquent d'utiliser l'anova de façon rétrospective, comme dans l'exemple 12.4.1 Il faut être prudent dans l'interprétation du résultat positif d'un tel test : l'expérimentateur ne fixant pas arbitrairement le génotype, la randomisation est impossible ; rien ne dit que le locus génétique considéré est la cause de la différence observée entre les concentrations sanguines : il peut être corrélé à un autre locus génétique ou à un facteur environnemental. Cette critique est bien sûr valable pour toutes les expériences non planifiées.

12.10.3 Rôle de la « randomisation »

Imaginons que dans l'étude sur le paludisme, on ait décidé d'administrer le premier traitement aux 12 premiers patients acceptant de participer à l'étude, puis le second traitement aux 12 suivants, etc. Le résultat de cette démarche serait sujet à de multiples sources de bais : changement des conditions climatiques au cours du temps, ou des souches de *plasmodium* infectant les patients ; et, à ne pas négliger, impossibilité de mettre en place la démarche du *double aveugle* où ni le médecin ni le patient ne savent quel médicament est administré, ce qui permet d'éviter des effets placebo ou nocebo différents selon le traitement.

Une démarche tout aussi criticable serait que le responsable de l'étude décide du groupe où on place tel ou tel patient : consciemment ou non, il pourrait biaiser le résultat, par exemple en plaçant dans un des groupes plus de patients en bon état de santé générale que dans les deux autres.

C'est pourquoi en pratique la répartition des patients est faite par tirage au sort, opération appelée « randomisation ». Cette méthode permet le double aveugle, et les patients dans chaque groupe sont représentatifs de la population générale, ce qui permet d'estimer les effets moyens dans la population des patients.

On peut formaliser cette intuition. On suppose que chez un patient donné, la réponse au traitement i est de la forme

$$\mathrm{X}_i = \mu + \alpha_i + b + \mathrm{E},$$

où $\mathrm{E} \sim \mathcal{N}(0, \sigma^2)$, α_i est l'« effet traitement » et b est l'« effet patient ». Notons que jusqu'à présent, on a supposé l'absence d'« effet patient », c'est-à-dire que $b = 0$ chez tous les patients.

Maintenant, les patients sont tirés au hasard dans une population de patients ; on va supposer que quand on tire un patient au hasard, la valeur de la réponse b chez ce patient est une réalisation d'une loi normale $\mathcal{N}(0, \sigma_{\mathrm{P}}^2)$, où σ_{P}^2 est la variance de l'effet patient.

Chaque patient ne reçoit qu'un traitement. Le modèle devient

$$\mathrm{X}_{ij} = \mu + \alpha_i + \mathrm{B}_{ij} + \mathrm{E}_{ij},$$

avec $\mathrm{B}_{ij} \sim \mathcal{N}(0, \sigma_{\mathrm{P}}^2)$ et $\mathrm{E}_{ij} \sim \mathcal{N}(0, \sigma^2)$. Les E_{ij} et les B_{ij} sont des v.a. indépendantes ; on pose $\mathrm{E}'_{ij} = \mathrm{E}_{ij} + \mathrm{B}_{ij}$, on a

$$\mathrm{X}_{ij} = \mu + \alpha_i + \mathrm{E}'_{ij},$$

avec $\mathrm{E}'_{ij} \sim \mathcal{N}(0, \sigma_{\mathrm{P}}^2 + \sigma^2)$.

On voit que, grâce à la randomisation, la valeur moyenne de l'effet patient dans chaque groupe est nulle. La variance de l'effet patient fait partie de la variance résiduelle.

12.11 Contrastes

Les valeurs α_i des effets (estimables, sous la contrainte $\sum_i n_i\alpha_i = 0$, par $X_{i\bullet} - X_{\bullet\bullet}$) n'ont pas de sens individuellement : comme nous l'avons remarqué dans la section 12.5, elles sont définies à une constante additive près.

Par contre, les quantités $C_{ik} = \alpha_i - \alpha_k$ sont bien définies ; on appelle C_{ik} le contraste entre les traitements i et k. Elles peuvent faire l'objet de tests, d'estimation ponctuelle ou par intervalle de confiance.

On estime la valeur de C_{ik} par $\widehat{C}_{ik} = X_{i\bullet} - X_{k\bullet}$, qui suit une loi normale $\mathcal{N}\left(C_{ik}, \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_k}\right)\right)$.

On peut estimer la valeur de σ^2 par $\mathrm{CMR} \sim \frac{\sigma^2}{n-p}\chi^2(n-p)$; on a donc l'écart réduit entre l'estimation de C_{ik} et 0 :

$$T_{ik} = \frac{\widehat{C}_{ik}}{\sqrt{\mathrm{CMR}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}} \sim t(n-p).$$

On peut tester si $\alpha_i = \alpha_k$ au risque α en rejetant cette hypothèse dès que $|T_{ik}| > t^{n-p}_{1-\alpha/2}$.

D'autre part, on a l'intervalle de confiance au niveau α suivant (voir aussi la section 9.2) :

$$C_{ik} \in \left[(X_{i\bullet} - X_{k\bullet}) \pm t^{n-p}_{1-\alpha/2}\sqrt{\mathrm{CMR}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}\right].$$

On peut également faire des tests unilatéraux et donner des intervalles de confiance unilatéraux. Attention toutefois au niveau auquel on fait les tests : là encore se pose la question des tests multiples.

Enfin, il n'est pas difficile généraliser la méthode des contrastes à des quantités de la forme $\sum_i c_i\alpha_i$ avec $\sum_i c_i = 0$.

12.11.1 Application à l'exemple du paludisme

On a calculé $X_{1\bullet} = 58{,}333$, $X_{2\bullet} = 74{,}5$, et $X_{3\bullet} = 82{,}9$.

On estime donc $C_{12} = 58{,}333 - 74{,}5 = -16{,}167$, $C_{13} = 58{,}333 - 82{,}9 = -24{,}576$ et $C_{23} = 74{,}5 - 82{,}9 = -8{,}4$.

On a calculé CMR = 462,9 avec 29 degrés de libertés. Le quantile d'ordre 0,975 de $t(29)$ est 2,0452. On a les intervalles de confiance à 95% suivants :

$$C_{12} \in \left[-16{,}2 \pm 2{,}0452\sqrt{462{,}9\left(\tfrac{1}{12} + \tfrac{1}{10}\right)}\right] = [-35{,}04; 2{,}64]$$

$$C_{13} \in \left[-24{,}6 \pm 2{,}0452\sqrt{462{,}9\left(\tfrac{1}{12} + \tfrac{1}{10}\right)}\right] = [-43{,}44; -5{,}76]$$

$$C_{23} \in \left[-8{,}4 \pm 2{,}0452\sqrt{462{,}9\left(\tfrac{1}{10} + \tfrac{1}{10}\right)}\right] = [-28{,}07; 11{,}28]$$

12.12 Tests partiels

On peut vouloir tester l'égalité de certains effets seulement, par exemple si on a $p = 5$ groupes on peut s'intéresser à l'hypothèse $\mu_1 = \mu_3 = \mu_5$ ou de façon équivalente $\alpha_1 = \alpha_3 = \alpha_5$. On pourrait faire une anova en oubliant les groupes 2 et 4, mais l'estimation de la variance résiduelle est alors moins bonne. On a donc intérêt à garder le CMR pour l'estimation de la variance ; et il suffit de calculer un CMF qui correspondent à nos trois groupes d'intérêt.

Notons $I \subset \{1, \ldots, p\}$ l'ensemble des numéros des groupes pour lesquels on veut tester l'hypothèse $H_I : \alpha_i = \alpha_k,\ i, k \in I$, et p_I le nombre de groupes dans I ; dans notre exemple $I = \{1, 3, 5\}$ et $p_I = 3$. Le test se fait contre l'hypothèse H_1 : au moins deux des α_i ($i \in I$) sont différents.

12.12.1 Somme des carrés factoriels associée à I

Notons $n_{\mathrm{I}} = \sum_{i\in \mathrm{I}} n_i$ le nombre total d'individus dans les groupes en question ; dans notre exemple $n_{\mathrm{I}} = n_1 + n_3 + n_5$. On considère la somme de carrés factoriels associée à I :

$$\boxed{\mathrm{SCF_I} = \frac{1}{n_{\mathrm{I}}} \sum_{\substack{i<k \\ i,k\in \mathrm{I}}} n_i n_k (\mathrm{X}_{i\bullet} - \mathrm{X}_{k\bullet})^2 = \frac{1}{n_{\mathrm{I}}} \sum_{\substack{i<k \\ i,k\in \mathrm{I}}} n_i n_k \widehat{\mathrm{C}}_{i,k}^2.}$$

Notez que si $\mathrm{I} = \{1,\dots,p\}$, $\mathrm{SCF_I}$ est la somme des carrés factorielles classique (cf 12.6.3).

Dans notre exemple $\mathrm{SCF_I} = \frac{1}{n_1+n_3+n_5}\left(n_1 n_3 \widehat{\mathrm{C}}_{1,3}^2 + n_1 n_5 \widehat{\mathrm{C}}_{1,5}^2 + n_3 n_5 \widehat{\mathrm{C}}_{3,5}^2\right)$.

On peut également calculer $\mathrm{SCF_I}$ à partir de la somme de carrés du groupe obtenu par fusion des groupes i pour $i \in \mathrm{I}$: si on pose

$$\boxed{\mathrm{SC_I} = \sum_{i\in \mathrm{I}} \sum_{j=1}^{n_i} \mathrm{X}_{ij}^2 - \frac{1}{n_{\mathrm{I}}} \left(\sum_{i\in \mathrm{I}} \sum_{j=1}^{n_i} \mathrm{X}_{ij} \right)^2,}$$

on a

$$\boxed{\mathrm{SC_I} = \mathrm{SCF_I} + \sum_{i\in \mathrm{I}} \mathrm{SC}_i.}$$

Voir également la section 12.13.

12.12.2 Le test

Sous $\mathrm{H_I}$ on a $\mathrm{SCF_I} \sim \sigma^2 \chi^2(p_{\mathrm{I}} - 1)$. Le carré moyen factoriel $\mathrm{CMF_I} = \mathrm{SCF_I}/(p_{\mathrm{I}} - 1)$ estime σ^2 sans biais sous $\mathrm{H_I}$ et la statistique de test $\mathrm{F_I} = \frac{\mathrm{CMF_I}}{\mathrm{CMR}}$ suit une loi $\mathrm{F}(p_{\mathrm{I}} - 1, n - p)$ degrés de liberté.

On peut donc tester $\mathrm{H_I}$ au risque α en rejetant $\mathrm{H_I}$ quand $\mathrm{F_I} > \mathrm{F}_{1-\alpha}^{p_{\mathrm{I}}-1,n-p}$.

Dans le cas où $p_{\mathrm{I}} = 2$, la statistique $\mathrm{F_I}$ vérifie

$$\mathrm{F_I} = \left(\frac{\widehat{\mathrm{C}}_{ik}}{\sqrt{\mathrm{CMR}\left(\frac{1}{n_i} + \frac{1}{n_k}\right)}} \right)^2.$$

On retrouve le test t de la section précédente : le carré d'une loi $t(n-p)$ est bien un $\mathrm{F}(1, n-p)$.

12.12.3 Application à l'exemple du paludisme

Bien que les intervalles de confiance calculés à la section précédente laissent prévoir le résultat, illustrons la méthode en testant $\alpha_1 = \alpha_2$. On a donc $\mathrm{I} = \{1,2\}$, $n_1 = 12$, $n_2 = 10$, et $p_{\mathrm{I}} = 2$.

On calcule

$$\mathrm{SCF}_{1,2} = \frac{1}{12+10}\left(10 \times 12 \times (-16{,}167)^2\right) = 1425{,}7.$$

Le carré moyen est $\mathrm{CMF}_{1,2} = \frac{1425{,}7}{2-1} = 1425{,}7$. La statistique $\mathrm{F}_{1,2} = \frac{\mathrm{CMF_I}}{\mathrm{CMR}} = \frac{1425{,}7}{462{,}9} = 3{,}08$ est à comparer (pour un test au seuil 5%) au quantile 0,95 de la loi $\mathrm{F}(1,29)$ qui vaut environ 4,2 : on ne rejette pas l'hypothèse nulle à ce seuil.

12.13 Modèles emboîtés

Nous allons enfin décrire une procédure de comparaison de modèles plus générale que celle des contrastes (section 12.11), ou des tests partiels (section 12.12).

Les différentes hypothèses testées correspondent à divers modèles : pour l'anova globale, l'hypothèse nulle correspond au modèle (a) qui contraint $\mu_1 = \mu_2 = \cdots = \mu_p$, alors que l'hypothèse alternative correspond au modèle (b) qui laisse $\mu_1, \mu_2, \cdots, \mu_p$ varier librement. On voit que le modèle (a) peut être vu comme un cas (très) particulier du modèle (b), où les paramètres coincident ; on dira que (a) est emboîté dans (b).

Le modèle (b) « colle » toujours mieux aux données que le modèle (a), au prix d'un nombre de paramètres plus important ; le test de l'anova peut-être vu comme un moyen de décider si l'amélioration de l'adéquation aux données dans (b) est significative.

Nous avons vu comment tester dans la section 12.12 des hypothèses comme $\mu_1 = \mu_3 = \mu_5$; une telle hypothèse correspond à un modèle (c) qui contraint $\mu_1 = \mu_3 = \mu_5$ et laisse les autres μ_i varier librement. Le modèle (a) est emboîté dans (c) qui lui-même est emboîté dans (b).

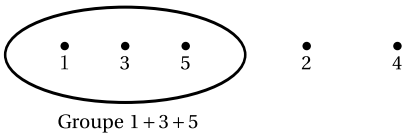


Figure 76. Dans le modèle (c), les groupes 1, 3 et 5 sont fusionnés en un seul groupe.

Pour ne pas alourdir l'écriture, nous utiliserons dans la suite le cas particulier de ce modèle, la généralisation à d'autres modèles étant facile.

Le test présenté en section 12.12 permet de comparer le modèle intermédiaire (c) au modèle le plus général, le modèle (b). Nous allons voir comment comparer deux modèles emboîtés quelconques.

12.13.1 Somme de carrés (résiduels) associée à un modèle

Chaque modèle définit des groupes de façons différentes : dans le modèle (a) on a un seul groupe, dans le modèle (b) on a p groupes, et dans le modèle (c) on a $p-2$ groupes.

Dans chaque modèle, on peut définir une somme de carrés résiduels, ou intra-groupe.

Dans le modèle (a), il n'y a qu'un groupe : la somme des carrés calculée dans ce groupe est simplement celle que nous avons appelée SCT.

Dans le modèle (b), il y a p groupes, et la somme des carrés résiduels est celle que nous avons appellée SCR.

Dans le modèle (c), il y a $p-2$ groupes. La somme des carrés résiduels se calcule en fusionnant les groupes qui sont confondus par le modèle.

Ainsi dans l'exemple où (c) est le modèle qui contraint $\mu_1 = \mu_3 = \mu_5$, et laisse μ_2 et μ_4 varier librement, on calcule la somme des carrés associée au groupe 1+3+5 :

$$\begin{aligned} \mathrm{SC}_{1,3,5} &= \sum_{i=1,3,5} \sum_{j=1}^{n_i} \mathrm{X}_{ij}^2 - \frac{1}{n_1+n_3+n_5} \left(\sum_{i=1,3,5} \sum_{j=1}^{n_i} \mathrm{X}_{ij} \right)^2 \\ &= \sum_{i=1,3,5} \mathrm{X}_{i+}^2 - \frac{1}{n_1+n_3+n_5} \left(\sum_{i=1,3,5} \mathrm{X}_{i+} \right)^2, \end{aligned}$$

qui correspond aux carrés des écarts à la moyenne dans ce groupe, et la somme des carrés résiduels associée au modèle est $\mathrm{SC}_{1,3,5} + \mathrm{SC}_2 + \mathrm{SC}_4$.

Remarquons que $\mathrm{SC}_{1,3,5} = \mathrm{SCF}_{1,3,5} + \mathrm{SC}_1 + \mathrm{SC}_3 + \mathrm{SC}_5$, ce qui peut servir à calculer $\mathrm{SCF}_{1,3,5}$, comme nous l'avons vu dans la section 12.12.

12.13.2 Ligne associée à un modèle

On peut dans un tableau d'anova insérer des lignes associées aux divers modèles envisagés. À un modèle H qui laisse varier librement k paramètres, on associe une ligne comportant une somme de carrés résiduels SCR_H, son degré de liberté $n-k$, et un carré moyen résiduel $\mathrm{CMR}_\mathrm{H} = \mathrm{SCR}_\mathrm{H}/(n-k)$ qui est un estimateur sans biais de σ^2 dans ce modèle.

Source	Somme des carrés	degrés de liberté	Carrés moyens
H	SCR_H	$n-k$	$\mathrm{CMR}_\mathrm{H} = \mathrm{SCR}_\mathrm{H}/(n-k)$

Dans le modèle (a) où $\mu_1 = \mu_2 = \mu_3$, il y a un paramètre qui varie librement, d'où $n-1$ degrés de liberté pour la somme des carrés SCR_a qui est SCT.

Dans le modèle (b) où les p paramètres $\mu_1, \ldots, \mu_p$ sont libres, on a $n-p$ degrés de liberté pour la somme des carrés SCR_b qui est SCR.

Dans le modèle (c) où $\mu_1 = \mu_3 = \mu_5$ et les autres μ_i sont libres, on a $p-2$ paramètres libres et $n-p+2$ degrés de libertés pour SCR_c.

12.13.3 Comparaison de modèles emboîtés

Si un modèle (d) est emboité dans un modèle (e), la somme de carrés SC_d est plus grande que la somme de carrés SC_e (plus il y a de groupes, plus la somme des carrés intra-groupes est petite).

On va chercher à tester si le modèle (e) « explique significativement mieux les données » que le modèle (d).

On fait un tableau d'analyse de la variance comme ceci :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
$\mathrm{H}_d - \mathrm{H}_e$	$\mathrm{SCF} = \mathrm{SCR}_d - \mathrm{SCR}_e$	$n = n_d - n_e$	$\mathrm{CMF} = \mathrm{SCF}/n$	$\mathrm{F} = \mathrm{CMF}/\mathrm{CMR}_e$
H_e	SCR_e	n_e	$\mathrm{CMR}_e = \mathrm{SCR}_e/n_e$	
H_d	SCR_d	n_d	$\mathrm{CMR}_d = \mathrm{SCR}_d/n_d$	

Les quantités SCF et n sont calculées par soustraction, le carré moyen doit être recalculé.

Si on se place sous l'hypothèse associée au modèle (d), on a $\mathrm{SCR}_d \sim \sigma^2\chi^2(n_d)$ et $\mathrm{SCR}_e \sim \sigma^2\chi^2(n_e)$, d'où $\mathrm{SCF} \sim \sigma^2\chi^2(n)$; on a alors $\mathrm{F} \sim \mathrm{F}(n, n_e)$, et on obtient un test au risque α pour l'hypothèse H_d en rejetant H_d si F dépasse le quantile $1-\alpha$ de la loi $\mathrm{F} \sim \mathrm{F}(n, n_e)$.

On vérifie que dans le cas des modèles emboîtés (a) et (b), c'est le test d'anova classique ; dans le cas des modèles emboîtés (c) et (b), c'est le test partiel présenté plus haut.

12.13.4 Application à l'exemple du paludisme

Nous reprenons le modèle (c) où $\mu_1 = \mu_2$: voir la section 12.12.3.

Nous avons déjà calculé la somme des carrés inter-groupes associée à ce modèle.

Le groupe constitué de la fusion des groupes 1 et 2 a 22 mesures, la somme des mesures est $x_{1+} + x_{2+} = 1445$ et la somme des carrés des mesures est $x_{1+}^2 + x_{2+}^2 = 104611$. On a donc $\mathrm{SC}_{1,2} = 104611 - \frac{1}{22}1445^2 = 9700{,}8$.

On remarque qu'on a bien $\mathrm{SCF}_{1,2} + \mathrm{SC}_1 + \mathrm{SC}_2 = 1425{,}7 + 4664{,}7 + 3610{,}5 = 9700{,}9 = \mathrm{SC}_{1,2}$ (aux erreurs d'arrondis près !).

La somme des carrés associée au modèle est $\mathrm{SC}_{1,2} + \mathrm{SC}_3 = 9700{,}8 + 5148{,}9 = 14849{,}7$.

Les somme de carrés associées aux trois modèles sont récapitulées dans le tableau suivant :

Source	Somme des carrés	degrés de liberté	Carrés moyens
H_b (3 par.)	13424,1	29	462,9
H_c (2 par.)	14849,7	30	495,0
H_a (1 par.)	16887,9	31	544,8

Comparaison de H_b et H_c

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
$H_c - H_b$	1425,6	1	1425,6	3,08
H_b	13424,1	29	462,9	
H_c	14849,7	30	495,0	

Le test est exactement le même que celui de la section 12.12.3, ce qui est une conséquence directe de l'égalité $SCF_{1,2} + SC_1 + SC_2 = SC_{1,2}$.

On ne rejette pas H_c au profit de H_b, c'est-à-dire que le modèles à trois paramètres μ_1, μ_2, μ_3 n'explique pas significativement mieux les observations que le modèle à deux paramètres $\mu_1 = \mu_2, \mu_3$.

Comparaison de H_c et H_a

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
$H_a - H_c$	2038,2	1	2038,2	4,11
H_c	14849,7	30	495,0	
H_a	16887,9	31	544,8	

La valeur 4,11 est à comparer au quantile 0,95 de F(1,30) qui vaut 4,171. On conclut là aussi (bien qu'on soit proche de la valeur critique !) que le modèle à deux paramètres $\mu_1 = \mu_2, \mu_3$ n'explique pas significativement mieux les observations que le modèle à un paramètre $\mu_1 = \mu_2 = \mu_3$.

On sait cependant que la comparaison de H_a et H_b, qu'on a déjà effectuée, conduit à considérer que le modèle à trois paramètres H_b explique significativement mieux les données que le modèle à un paramètre H_a !

Cet exemple illustre la difficulté qu'il peut y avoir à choisir un modèle après un test d'anova (en particulier, doit-on faire ces tests d'emboîtement au seuil $\alpha = 5\%$?). Il existe des procédures qui décident de façon globale quels groupes on peut fusionner ou non ; aucune ne fait vraiment l'unanimité. Nous ne les aborderons pas ici.

12.14 Le modèle aléatoire : test d'homogénéité

Nous nous plaçons ici dans un cadre différent du cadre présenté dans tout le chapitre, mais qui conduit à des calculs similaires.

Il s'agit d'un test d'homogénéité ; contrairement au cas des facteurs à « effets fixes » présentés jusqu'ici, qui sont bien déterminés et reproductibles, on a maintenant des « effets aléatoires » : on considère que l'effet a d'un groupe donné a été tiré au hasard (en fait, c'est le groupe qui a été tiré au hasard !). La valeur de a est bien sûr la même pour toutes les observations faites dans le groupe en question. Nous ne sommes plus intéressé par la valeur de a, mais par sa variabilité entre les groupes.

12.14.1 Exemple ostréicole

Comme illustration, nous proposons la situation suivante : une région ostréicole est touchée par une pollution d'algues toxiques. On cherche à savoir si la pollution est homogène, ou si il existe une variabilité sur la région. On choisi donc 5 ostréiculteurs **au hasard** dans la région, et on prélève 8

huitres chez chacun d'eux. On dose la concentration de toxine dans chacune de ces huitres. Il y a des dizaines d'ostréiculteurs dans cette régon : on ne cherche pas à connaître précisément la pollution chez les cinq ostréiculteurs sélectionnés, mais la variabilité à l'échelle de la région.

	Ostréiculteur 1		Ostréiculteur 2		Ostréiculteur 3		Ostréiculteur 4		Ostréiculteur 5	
	9,2	9,3	12,0	9,4	11,6	11,5	9,9	8,8	10,2	8,4
	8,7	8,9	10,3	11,0	11,6	11,9	8,2	8,4	10,6	11,1
	9,3	7,5	10,9	10,1	8,8	10,3	11,5	9,5	9,2	9,1
	8,8	8,1	14,0	11,0	11,2	10,2	9,0	10,0	8,0	9,4
x_{i+}	69,8		88,7		87,1		75,3		76	
x_{i+}^2	611,82		997,27		955,99		716,75		729,98	

12.14.2 Le modèle des effets aléatoires

On continue à utiliser la notations X_{ij} pour la j^e observation du groupe i, pour $i = 1,\dots,p$ et $j = 1,\dots,n_i$.

Le modèle est maintenant

$$\boxed{X_{ij} = \mu + A_i + E_{ij},}$$

où $A_i \sim \mathcal{N}(0,\sigma_A^2)$ est l'effet du groupe i, supposé tiré au hasard dans l'éventail des effets possibles, et $E_{ij} \sim \mathcal{N}(0,\sigma^2)$ est la variabilité intra-groupe.

L'hypothèse nulle est que l'effet est constant sur tous les groupes : $H_0 : \sigma_A^2 = 0$ et l'hypothèse alternative est qu'il y a une variabilité : $H_1 : \sigma_A^2 > 0$.

Sous H_0 on a $A_i = 0$ (variable gaussienne « dégénérée »), et $X_{ij} = \mu + E_{ij}$, donc $X_{ij} \sim \mathcal{N}(\mu,\sigma^2)$. C'est le même modèle que dans le cas des effets fixes : on peut utiliser la même statistique de test, dont la loi sous H_0 est connue. Il est par ailleurs intuitif que sous H_1, la statistique de test a tendance à être plus grande que sous H_0, comme dans le cas des effets fixes, ce qui justifie de réaliser exactement les mêmes tests.

12.14.3 Valeurs attendues sous H_1

Par contre, sous H_1, le modèle est très différent. Chaque X_{ij} est une loi normale centrée, de variance $\sigma_A^2 + \sigma^2$ (on parle d'analyse des composantes de la variance pour l'estimation de σ_A^2 et σ^2), mais les X_{ij} ne sont plus indépendants : deux mesures prises au sein du même groupe i ont covariance $\mathrm{cov}(X_{ij},X_{ij'}) = \sigma_A^2$.

En posant $n' = n\left(1 - \sum_{i=1}^p \left(\frac{n_i}{n}\right)^2\right)$, on a

$$\text{sous } H_1,\ \begin{cases} E(CMR) &= \sigma^2 \\ E(CMF) &= \sigma^2 + \frac{n'}{p-1}\sigma_A^2 \\ E(CMT) &= \sigma^2 + \frac{n'}{n-1}\sigma_A^2 \end{cases}$$

On peut utiliser ces résultats pour estimer $\sigma_A^2 = \frac{p-1}{n'}(CMF - CMR)$ quand cette quantité est positive, $\sigma_A^2 = 0$ sinon. Il est plus difficile de donner un intervalle de confiance sur σ_A^2 : il faudrait préciser la loi des estimations sous H_1.

12.14.4 Cas des données équiréparties

On se place ici dans le cas où on a le même nombre $n_i = r = \frac{n}{p}$ d'observations pour chaque niveau du facteur. On a en particulier $n' = n\frac{p-1}{p}$. On peut en outre préciser la loi de SCF sous H_1. En effet on a $X_{i+} = r\mu + rA_i + E_{i+}$, d'où

$$X_{i+} \sim \mathcal{N}(r\mu, r^2\sigma_A^2 + r\sigma^2),$$

puis

$$X_{i\bullet} \sim \mathcal{N}\left(\mu, \sigma_A^2 + \frac{1}{r}\sigma^2\right).$$

Les variables $X_{i\bullet}$ $(i = 1, \ldots, p)$ peuvent être vues comme des observations indépendantes de même loi $\mathcal{N}\left(\mu, \sigma_A^2 + \frac{1}{r}\sigma^2\right)$. Notons

$$\sigma_1^2 = \sigma_A^2 + \frac{1}{r}\sigma^2$$

leur variance commune. La moyenne empirique des $X_{i\bullet}$, $X_{\bullet\bullet} = \frac{1}{p}\sum_i X_{i\bullet}$ suit donc une loi normale :

$$X_{\bullet\bullet} \sim \mathcal{N}\left(\mu, \frac{1}{p}\sigma_1^2\right).$$

On estime naturellement σ_1^2 par

$$\frac{1}{p-1}\sum_{i=1}^{p}(X_{i\bullet} - X_{\bullet\bullet})^2 = \frac{1}{(p-1)r}\mathrm{SCF},$$

qui suit une loi $\frac{1}{p-1}\sigma_1^2\chi^2(p-1)$.

On a donc

$$\mathrm{SCF} = r\sum_{i=1}^{p}(X_{i\bullet} - X_{\bullet\bullet})^2 \sim \left(r\sigma_A^2 + \sigma^2\right)\chi^2(p-1).$$

L'estimation de σ_1^2 ci-dessus permet de donner un intervalle de confiance de niveau α sur μ :

$$\mathbb{P}\left(X_{\bullet\bullet} - t_{1-\alpha/2}^{p-1}\sqrt{\frac{\sigma_1^2}{p}} < \mu < X_{\bullet\bullet} + t_{1-\alpha/2}^{p-1}\sqrt{\frac{\sigma_1^2}{p}}\right) = 1-\alpha.$$

12.14.5 Application à l'exemple ostréicole

On calcule SCR = 40,28 et SCT = 73,57. La table d'analyse de la variance est

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	33,29	4	8,32	F = 7,23
résidus	40,28	35	1,15	
Total	73,57	39		

On compare la valeur de F au 95$^{\mathrm{e}}$ centile de la la loi F(4,35), qui vaut 2,64 : on rejette H_0.

On peut estimer la variance de la concentration moyenne de toxine dans les huitres des différents ostréiculteurs de la région par $\sigma_A^2 = \frac{p-1}{n'}(\mathrm{CMF} - \mathrm{CMR}) = \frac{4}{32}(8{,}32 - 1{,}15) = 0{,}89$. (ici $n' = 40 \times (1 - 1/5) = 32$), alors que la variance au sein d'une exploitation donnée est estimée par CMR = 1,15.

12.15 Exercices

Exercice 1 On considère un échantillon formé de 50 hommes et 50 femmes dont on a mesuré la stature. La stature moyenne empirique des hommes est de 176,00 cm, et celle des femmes de 168,00 cm.

1. Quelle est la stature moyenne calculée sur l'échantillon ?

2. La variance empirique calculée sur l'ensemble de l'échantillon vaut 73,50 cm^2. En supposant que la variance est la même chez les hommes et les femmes, quelle est l'estimation de la variance commune ?

Exercice 2 On reprend les données de l'exemple de la section 12.4.1, résumées dans le tableau ci-dessous.

Génotype BB	Génotype Bb	Génotype bb
$n_1 = 12$	$n_2 = 39$	$n_3 = 49$
$x_{1+} = 55{,}17$	$x_{2+} = 202{,}46$	$x_{3+} = 242{,}89$
$x_{1+}^2 = 267{,}3067$	$x_{2+}^2 = 1091{,}836$	$x_{3+}^2 = 1284{,}971$

Réaliser l'analyse de la variance pour tester l'égalité des concentrations mesurées dans chacun des trois groupes.

Exercice 3 Reprenons les données de l'exemple du paludisme, résumées ci-dessous

	Traitement 1	Traitement 2	Traitement 3
n_i	12	10	10
x_{i+}	700	745	829
x_{i+}^2	45498	59113	73873

Les trois niveaux de traitements correspondent en fait à trois posologies différentes : 10, 15 et 20 mg d'une même molécule.

On cherche à tester l'existence d'un effet-dose linéaire : le temps de clairance parasitaire évolue-t-il linéairement avec la dose de médicament ?

1. On utilise les notations du modèle linéaire : $X_{ij} = \mu + \alpha_i + E_{ij}$. Montrer que cette hypothèse est équivalente à $\alpha_1 - 2\alpha_2 + \alpha_3 = 0$.

2. Montrer que le contraste $C = \alpha_1 - 2\alpha_2 + \alpha_3$ peut être estimé par $\widehat{C} = X_{1\bullet} - 2X_{2\bullet} + X_{3\bullet}$. Quelle est la loi de $\widehat{C}$?

3. Calculer $\widehat{C}$ et un intervalle de confiance pour C. Conclure.

Exercice 4 On considère des données de temps de clairance parasitaire (TCP) obtenues par un même traitement anti-paludéen, sur des patients de quatre régions différentes : deux régions africaines (groupes 1 et 2) et deux régions asiatiques (groupes 3 et 4).

Les données sont résumées par l'effectif de chaque groupe, la somme des TCP observés dans chaque groupe ainsi que par la somme des carrés des TCP.

	n_i	x_{i+}	x_{i+}^2
Groupe 1	10	568	37 186
Groupe 2	10	621	42 965
Groupe 3	10	879	85 781
Groupe 4	10	732	55 390
Total	40	2800	221 322

1. Tester l'hypothèse selon laquelle le traitement a la même efficacité dans ces quatre régions.

2. Comparer les trois modèles emboîtés suivants :
   (a) Le modèle à un paramètre où le traitement a la même efficacité dans les quatres groupes ;
   (b) le modèle à deux paramètres où le traitement a la même efficacité dans les groupes africains, et la même efficacité dans les groupes asiatiques ;
   (c) le modèle à quatre paramètres, un pour chaque groupe.

Exercice 5 On administre quatre traitements différents à quatre groupes de patients ; leur efficacité est mesurée sur une échelle appropriée. Les résultats sont rassemblés dans la table qui suit :

Traitement	n_i	x_{i+}	x_{i+}^2
1	10	253,40	7173,56
2	20	502,50	13590,33
3	20	592,60	18670,08
4	15	512,70	18194,09

1. Y a-t-il une différence entre les groupes ?

2. Donner un intervalle de confiance à 90% pour la variance résiduelle.

3. Donner un intervalle de confiance à 90% pour la valeur moyenne de l'efficacité du traitement dans le groupe 2.

4. Tester l'hypothèse d'égalité des effets entre les groupes 2 et 3.

Exercice 6 (anova) On dose un taux d'anticorps chez un total de 30 patients répartis dans 3 hôpitaux (10 patients par hôpital). Voici les résultats bruts (en UI/l) :

Hôpital	Mesures									
A	13,1	18,6	18,4	15,9	10,9	16,1	16,8	16,4	13,9	12,7
B	13,1	13,9	14,6	12,4	17,1	16,4	17,4	17,0	15,0	9,8
C	16,1	17,8	12,9	13,5	14,1	14,5	13,0	11,2	17,6	13,6

Pour gagner du temps, on donne également la somme des mesures et la somme de leurs carrés dans les trois hôpitaux :

Hôpital	$\sum x_i$	$\sum x_i^2$
A	152,8	2392,66
B	146,7	2205,91
C	144,3	2122,93

On supposera que les mesures suivent une loi normale dans chacun des hôpitaux.

1. Quel est le taux d'anticorps moyen chez les patients de l'hôpital A ? Donnez un intervalle de confiance à 95%.

2. Même question pour les hôpitaux B et C.

3. On supposera que la variance des anticorps est la même dans les trois hôpitaux. En utilisant cette hypothèse, recalculez les trois intervalles de confiance précédents.

4. Y a-t-il une différence significative (au seuil $\alpha = 5\%$) entre les taux d'anticorps moyen dans les trois hôpitaux ?

Chapitre 13

Introduction à l'analyse de variance à deux facteurs

13.1 Introduction

Nous allons moins développer l'anova à deux facteurs que l'anova à un facteur ; en particulier nous ne parlerons pas ici de comparaison de modèles emboîtés bien que le principe soit toujours le même : on compare les sommes de carrés résiduelles de chacun des deux modèles au moyen d'un test F

Nous ne considérerons que des données recueillies par plan d'expérience, ce qui permet d'imposer des conditions aux effectifs de chaque catégorie. Ces conditions permettent de pouvoir estimer séparément les effets de chacun des facteurs. On parlera de *plan d'expérience équilibré*. Sans cette condition, tous les tests présentés dans ce chapitre ne sont plus valides.

Les études rétrospectives sont possibles, mais pour les traiter il faudrait se placer dans le cadre plus général du modèle linéaire. Leur interprétation est au moins aussi délicate, sinon plus, que dans le cas de l'anova 1.

Voici quelques applications envisagées de l'anova 2 avec plan d'expérience.

13.1.1 Deux traitements simultanés

On étudie l'effet de deux traitements administrés simultanément, à des niveaux variés, dans une étude randomisée en double aveugle. On s'intéresse à l'effet « individuel » de chaque traitement, et à l'interaction possible entre les traitements.

13.1.2 Un traitement (à plusieurs niveaux), et une variable de confusion

On suspecte que les hommes et les femmes réagissent de façon différente aux divers niveaux traitement donné. L'étude randomisée permet d'avoir à peu près le même sexe ratio dans les groupes associés au divers niveau de traitements. Cependant il est préférable de contrôler précisément ce sex ratio, ce qui permettra de mieux estimer l'effet moyen du traitement ; on pourra aussi tester l'hypothèse faite sur l'effet du sexe.

On considérera que les patients sont répartis en « blocs » selon leur sexe ; on parle d'effet bloc pour l'effet de ce facteur. Le premier facteur sera toujours appelé « traitement ».

13.1.3 Un traitement à plusieurs niveaux, une multitude de patients

De façon plus surprenante, on pourra également considérer chaque patient comme un bloc. On administrera à chaque patient plusieurs médicaments différents, à des moments différents ; ce type

de plan d'expérience est surtout envisageable dans le cas d'une maladie chronique.

Dans ce cas on ne s'intéresse plus aux valeurs précises de l'effet patient, ou effet bloc, pour les patients participant à l'étude, mais à la variabilité de l'effet dans la population des patients : comme dans l'exemple de la pollution ostréicole à la fin du chapitre précédent, on considèrera que l'effet patient est une variable aléatoire, et on s'intéressera à sa variance. On parle alors de modèle mixte.

C'est également une généralisation de la méthode des séries appariées (cf 11.9) :

13.2 Modèle à effets fixes

L'observation numéro k dans le groupe où le premier facteur (facteur A) est au niveau i et le second facteur (facteur B) est au niveau j est X_{ijk}, $i = 1,\dots,p$, $j = 1,\dots,q$, $k = 1,\dots,n_{ij}$.

Comme dans le chapitre précédent, on remplace un indice par un signe plus pour noter la sommation sur cet indice, et par un point pour noter la moyenne sur cet indice.

On note $n = n_{++}$: le nombre total d'observations. On aura

$$\begin{aligned}
n_{i+} &= \sum_{j=1}^{q} n_{ij}, & n_{+j} &= \sum_{i=1}^{p} n_{ij},\\
X_{ij+} &= \sum_{k=1}^{n_{ij}} X_{ijk}, & X_{ij\bullet} &= \tfrac{1}{n_{ij}} X_{ij+},\\
X_{i++} &= \sum_{j=1}^{q} X_{ij+}, & X_{i\bullet\bullet} &= \tfrac{1}{n_{i+}} X_{i++},\\
X_{+j+} &= \sum_{i=1}^{p} X_{ij+}, & X_{\bullet j\bullet} &= \tfrac{1}{n_{+j}} X_{+j+},\\
X_{+++} &= \sum_{ijk} X_{ijk}, & X_{\bullet\bullet\bullet} &= \tfrac{1}{n} X_{+++}.
\end{aligned}$$

On a des formules analogues pour les sommes de carrés des données, $X^2_{ij+} = \sum_k X^2_{ijk}$, etc.

On considèrera ici exclusivement le cas des données équilibrées : les données pour lesquelles

$$\boxed{n_{ij} = \frac{1}{n} n_{i+} n_{+j}.}$$

Cette condition peut s'interpréter comme ceci : si on présente les données dans un tableau, les lignes correspondant au facteur A et les colonnes au facteur B, dans toutes les lignes du tableau, les nombres d'observations sont dans des proportions identiques ; et dans toutes les colonnes du tableau, les nombres d'observations sont dans des proportions identiques.

Ainsi par exemple voici un plan d'expérience équilibré pour un facteur A à 4 niveaux et un facteur B à trois niveaux.

	Niveau B_1	Niveau B_2	Niveau B_3
Niveau A_1	$n_{11} = 2$	$n_{12} = 4$	$n_{13} = 10$
Niveau A_2	$n_{21} = 1$	$n_{22} = 2$	$n_{23} = 5$
Niveau A_3	$n_{31} = 3$	$n_{32} = 6$	$n_{33} = 15$
Niveau A_4	$n_{41} = 4$	$n_{42} = 8$	$n_{43} = 20$

Dans chaque ligne, les effectifs sont en proportion 1 : 2 : 5, c'est-à-dire qu'il suffit de multiplier les nombres 1, 2, et 5 par un même nombre pour obtenir les effectifs de la ligne ; par exemple, pour la ligne 1, on multiplie par 2 et on obtient 2, 4 et 10.

De même, dans chaque colonne, les effectifs sont en proportion 2 : 1 : 3 : 4.

13.2.1 Le modèle

Le modèle est

$$\boxed{X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk},}$$

où les E_{ijk} sont indépendantes de loi $\mathcal{N}(0,\sigma^2)$.

Comme dans la section 12.5, pour que l'écriture du modèle soit unique on impose des contraintes sur les constantes α_i, β_j, γ_{ij} :

$$\begin{array}{rcl@{\qquad}rcl} \displaystyle\sum_{i=1}^{p} n_{i+}\alpha_i &=& 0, & \displaystyle\sum_{j=1}^{q} n_{+j}\beta_j &=& 0, \\ \forall i,\ \displaystyle\sum_{j=1}^{q} n_{+j}\gamma_{ij} &=& 0, & \forall j,\ \displaystyle\sum_{i=1}^{p} n_{i+}\gamma_{ij} &=& 0. \end{array}$$

Les termes α_i représentent l'effet (principal) du facteur A, les β_j l'effet du facteur B, et les γ_{ij} représentent l'interaction entre les deux facteurs.

Les différentes hypothèses à tester sont

— $H_{0A} : \forall i, \alpha_i = 0$ (pas d'effet du facteur A) ;
— $H_{0B} : \forall j, \beta_j = 0$ (pas d'effet du facteur B) ;
— $H_{0AB} : \forall i,j, \gamma_{ij} = 0$ (pas d'interaction entre A et B).

13.2.2 Estimation des paramètres

On a

$$\begin{aligned} X_{ij+} &= \sum_k X_{ijk} \\ &= n_{ij}(\mu + \alpha_i + \beta_j + \gamma_{ij}) + E_{ij+} \\ &= \frac{1}{n} n_{i+} n_{+j} (\mu + \alpha_i + \beta_j + \gamma_{ij}) + E_{ij+} \qquad & E_{ij+} \sim \mathcal{N}(0, n_{ij}\sigma^2) \end{aligned}$$

$$\begin{aligned} X_{i++} &= \sum_j X_{ij+} \\ &= n_{i+}(\mu + \alpha_i) + \frac{n_{i+}}{n} \sum_j n_{+j}(\beta_j + \gamma_{ij}) + E_{i++} \\ &= n_{i+}(\mu + \alpha_i) + E_{i++} \qquad & E_{i++} \sim \mathcal{N}(0, n_{i+}\sigma^2) \end{aligned}$$

$$X_{+j+} = n_{+j}(\mu + \beta_j) + E_{+j+} \qquad E_{+j+} \sim \mathcal{N}(0, n_{+j}\sigma^2)$$

$$X_{+++} = n\mu + E_{+++} \qquad E_{+++} \sim \mathcal{N}(0, n\sigma^2)$$

On en déduit les estimateurs suivants :

$$\boxed{\begin{aligned} \widehat{\mu} &= X_{\bullet\bullet\bullet} \\ \widehat{\alpha_i} &= X_{i\bullet\bullet} - X_{\bullet\bullet\bullet} \\ \widehat{\beta_j} &= X_{\bullet j\bullet} - X_{\bullet\bullet\bullet} \\ \widehat{\gamma_{ij}} &= X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet} \end{aligned}}$$

13.2.3 Les sommes de carrés

Les sommes de carrés et les carrés moyens associés sont :

$$\begin{array}{rcl|rcl}
\mathrm{SCT} & = & \sum_{ijk}(\mathrm{X}_{ijk}-\mathrm{X}_{\bullet\bullet\bullet})^2 & \mathrm{CMT} & = & \dfrac{\mathrm{SCT}}{n-1} \\
\mathrm{SCF_A} & = & \sum_{i=1}^{p} n_{i+}(\mathrm{X}_{i\bullet\bullet}-\mathrm{X}_{\bullet\bullet\bullet})^2 & \mathrm{CMF_A} & = & \dfrac{\mathrm{SCF_A}}{p-1} \\
\mathrm{SCF_B} & = & \sum_{j=1}^{q} n_{+j}(\mathrm{X}_{\bullet j\bullet}-\mathrm{X}_{\bullet\bullet\bullet})^2 & \mathrm{CMF_B} & = & \dfrac{\mathrm{SCF_A}}{q-1} \\
\mathrm{SCF_{AB}} & = & \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(\mathrm{X}_{ij\bullet}-\mathrm{X}_{i\bullet\bullet}-\mathrm{X}_{\bullet j\bullet}+\mathrm{X}_{\bullet\bullet\bullet})^2 & \mathrm{CMF_{AB}} & = & \dfrac{\mathrm{SCF_{AB}}}{(p-1)(q-1)} \\
\mathrm{SCR} & = & \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n_{ij}}(\mathrm{X}_{ijk}-\mathrm{X}_{ij\bullet})^2 & \mathrm{CMR} & = & \dfrac{\mathrm{SCR}}{n-pq}.
\end{array}$$

Nous ne commenterons pas en détail ces définitions. On reconnaît dans la SCR la somme des résidus groupe par groupe, dans la SCT la somme des résidus obtenues en considérant un groupe unique, et dans les SCF les estimateurs des coefficients α_i, β_j et γ_{ij}, ce qui montre que de grandes valeusr de ces SCF plaide pour la non-nullité de ces coefficients.

Les sommes de carrés sont liées par la relation

$$\boxed{\mathrm{SCT} = \mathrm{SCF_A} + \mathrm{SCF_B} + \mathrm{SCF_{AB}} + \mathrm{SCR}.}$$

Leurs degrés de libertés sont liés par la relation

$$\boxed{n-1 = (p-1)+(q-1)+(p-1)(q-1)+(n-pq).}$$

13.2.4 Les lois

À partir des égalités suivantes :

$$\begin{array}{rcl}
\mathrm{X}_{i\bullet\bullet}-\mathrm{X}_{\bullet\bullet\bullet} & = & \alpha_i + \frac{1}{n_{i+}}\mathrm{E}_{i++} - \frac{1}{n}\mathrm{E}_{+++} \\
\mathrm{X}_{\bullet j\bullet}-\mathrm{X}_{\bullet\bullet\bullet} & = & \beta_j + \frac{1}{n_{+j}}\mathrm{E}_{+j+} - \frac{1}{n}\mathrm{E}_{+++} \\
\mathrm{X}_{ijk}-\mathrm{X}_{ij\bullet} & = & \mathrm{E}_{ijk} - \frac{1}{n_{ij}}\mathrm{E}_{ij+} \\
(\mathrm{X}_{ij\bullet}-\mathrm{X}_{i\bullet\bullet}-\mathrm{X}_{\bullet j\bullet}+\mathrm{X}_{\bullet\bullet\bullet}) & = & \gamma_{ij} + \left(\frac{1}{n_{ij}}\mathrm{E}_{ij+} - \frac{1}{n_{i+}}\mathrm{E}_{i++} - \frac{1}{n_{+j}}\mathrm{E}_{+j+} + \frac{1}{n}\mathrm{E}_{+++}\right)
\end{array}$$

on peut montrer que

$$\begin{array}{rcl}
\mathrm{E}(\mathrm{CMF_A}) & = & \frac{1}{p-1}\sum_i n_{i+}\alpha_i^2 + \sigma^2 \\
\mathrm{E}(\mathrm{CMF_B}) & = & \frac{1}{q-1}\sum_j n_{+j}\beta_j^2 + \sigma^2 \\
\mathrm{E}(\mathrm{CMF_{AB}}) & = & \frac{1}{(p-1)(q-1)}\sum_{ij} n_{ij}\gamma_{ij}^2 + \sigma^2 \\
\mathrm{E}(\mathrm{CMR}) & = & \sigma^2.
\end{array}$$

On a également

$$\begin{array}{rrcl}
\text{sous } \mathrm{H_{0A}}: & \mathrm{SCF_A} & \sim & \sigma^2\chi^2(p-1) \\
\text{sous } \mathrm{H_{0B}}: & \mathrm{SCF_B} & \sim & \sigma^2\chi^2(q-1) \\
\text{sous } \mathrm{H_{0AB}}: & \mathrm{SCF_{AB}} & \sim & \sigma^2\chi^2((p-1)(q-1)) \\
& \mathrm{SCR} & \sim & \sigma^2\chi^2(n-pq).
\end{array}$$

13.2.5 Les tests

On en déduit les statistiques de tests :

$$\begin{array}{rrcl} \text{sous } H_{0A}: & F_A & = \frac{CMF_A}{CMR} & \sim F(p-1,n-pq) \\ \text{sous } H_{0B}: & F_B & = \frac{CMF_B}{CMR} & \sim F(q-1,n-pq) \\ \text{sous } H_{0AB}: & F_{AB} & = \frac{CMF_{AB}}{CMR} & \sim F((p-1)(q-1),n-pq) \end{array}$$

Comme dans le cas de l'anova 1, on fera des tests unilatéraux à droite.

Enfin, tous ces tests sont indépendants, les variables SCR, SCF_A, SCF_B et SCF_{AB} étant indépendantes.

13.2.6 Cas particulier : une observation par cellule

Supposons qu'on a $n_{ij} = 1$ pour tout i, j ; on a $n = pq$. Si on suit les calculs fait jusqu'ici, on obtient SCR = 0, avec « 0 degré de liberté »! Les tests ne sont pas réalisables.

La solution est de postuler a priori que le modèle est additif :

$$X_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}.$$

Dans ce cas H_{0AB} est vrai : tous les γ_{ij} sont nuls, $SCF_{AB} \sim \sigma^2\chi^2((p-1)(q-1))$, et CMF_{AB} est une estimation de σ^2, avec $(p-1)(q-1) = pq-p-q+1 = n-p-q+1$ degrés de liberté. On pourra alors prendre SCF_{AB} comme somme de carrés résiduels (c'est bien la somme des carrés résiduels du modèle où les γ_{ij} sont nuls), et noter $SCR = SCF_{AB}$, $CMR = CMF_{AB}$.

Avec cette notation, on a la relation suivante : $SCT = SCF_A + SCF_B + SCR$.

On peut alors tester H_{0A} au moyen de $F_A = \frac{CMF_A}{CMR} \sim F(p-1,n-p-q+1)$ et H_{0B} au moyen de $F_B = \frac{CMF_B}{CMR} \sim F(q-1,n-p-q+1)$.

13.2.7 Réalisation pratique

On a les formules de décentrage suivantes :

$$\boxed{\begin{aligned} SCT &= X_{+++}^2 - \frac{1}{n}(X_{+++})^2 \\ SCF_A &= \sum_{i=1}^{p} \frac{1}{n_{i+}}(X_{i++})^2 - \frac{1}{n}(X_{+++})^2 \\ SCF_B &= \sum_{j=1}^{q} \frac{1}{n_{+j}}(X_{+j+})^2 - \frac{1}{n}(X_{+++})^2 \end{aligned}}$$

et d'autre part, $SCR = \sum_{ij} SC_{ij}$, où

$$\boxed{SC_{ij} = X_{ij+}^2 - \frac{1}{n_{ij}}(X_{ij+})^2}$$

Remarquons qu'on peut également calculer

$$\begin{array}{rcl} SCA_i & = & X_{i++}^2 - \frac{1}{n_{i+}}(X_{i++})^2, \\ SCB_j & = & X_{+j+}^2 - \frac{1}{n_{+j}}(X_{+j+})^2; \end{array}$$

on a ensuite $SCF_A = SCT - \sum_i SCA_i$ et $SCF_B = SCT - \sum_j SCB_j$.

On calcule pour finir $SCF_{AB} = SCT - SCF_A - SCF_B - SCR$.

On présentera les résultats dans une table d'analyse de la variance :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Traitement	SCF_A	$p-1$	$CMF_A = SCF_A/(p-1)$	$F = CMF_A/CMR$
Blocs	SCF_B	$q-1$	$CMF_B = SCF_B/(q-1)$	$F = CMF_B/CMR$
Interaction	SCF_{AB}	$(p-1)(q-1)$	$CMF_{AB} = SCF_{AB}/((p-1)(q-1))$	$F = CMF_{AB}/CMR$
Résidus	SCR	$n-pq$	$CMR = SCR/(n-pq)$	
Total	SCT	$n-1$		

13.2.8 Exemple (avec données répétées)

On applique un traitement à des patients hommes et femmes ; on soupçonne que l'effet de ce traitement dépend du sexe. Les données collectées sont les suivantes :

	Traitement A	Traitement B	Traitement C
Hommes	10,6 11,3 8,6 8,3	12,2 14,2 10,8 16,3	9,7 10,8 5,8 10,0 7,0 10,6
Femmes	9,1 12,6 15,4 14,2 13,2 10,6	10,9 13,3 12,5 12,5 9,5 14,0	9,5 13,9 12,6 12,0 12,4 7,6 10,5 11,3 10,2

Notez que le plan d'expérience est équilibré : dans chaque ligne les effectifs sont en proportion 2 : 2 : 3, et dans chaque colonne en proportion 2 : 3. Ces données peuvent être « récapitulées » par la donnée, pour chaque catégorie, de l'effectif, la somme des mesures, et la somme des carrés des mesures.

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$ $x_{11+} = 38{,}8$ $x^2_{11+} = 382{,}90$	$n_{12} = 4$ $x_{12+} = 53.5$ $x^2_{12+} = 732{,}81$	$n_{13} = 6$ $x_{13+} = 53.9$ $x^2_{13+} = 505{,}73$	$n_{1+} = 14$ $x_{1++} = 146{,}2$ $x^2_{1++} = 1621{,}44$
Femmes	$n_{21} = 6$ $x_{21+} = 75{,}1$ $x^2_{21+} = 966{,}97$	$n_{22} = 6$ $x_{22+} = 72{,}7$ $x^2_{22+} = 894{,}45$	$n_{23} = 9$ $x_{23+} = 100{,}0$ $x^2_{23+} = 1139{,}72$	$n_{2+} = 21$ $x_{2++} = 247{,}8$ $x^2_{2++} = 3001{,}14$
	$n_{+1} = 10$ $x_{+1+} = 113{,}9$ $x^2_{+1+} = 1349{,}87$	$n_{+2} = 10$ $x_{+2+} = 126{,}2$ $x^2_{+2+} = 1627{,}26$	$n_{+3} = 15$ $x_{+3+} = 153{,}9$ $x^2_{+3+} = 1645{,}45$	$n_{++} = 35$ $x_{+++} = 394{,}0$ $x^2_{+++} = 4622{,}58$

On peut présenter dans un tableau reprenant la structure du tableau de données le calcul des $SC_{ij} = x^2_{ij+} - \frac{1}{n_{ij}}(x_{ij+})^2$:

	Traitement A	Traitement B	Traitement C
Hommes	$SC_{11} = 6{,}54$	$SC_{12} = 17{,}25$	$SC_{13} = 21{,}53$
Femmes	$SC_{21} = 26{,}97$	$SC_{22} = 13{,}57$	$SC_{23} = 28{,}61$

On a par exemple calculé $SC_{11} = x_{11+}^2 - \frac{1}{n_{11}}(x_{11+})^2 = 382{,}90 - \frac{1}{4}38{,}8^2 = 6{,}54$.
On a ensuite $SCR = \sum_{ij} SC_{ij} = 114{,}46$;

On a d'autre part $SCT = x_{+++}^2 - \frac{1}{n}(x_{+++})^2 = 4622{,}58 - \frac{1}{35}394^2 = 187{,}27$.
On calcule ensuite SCF_A :

$$\begin{aligned} SCF_A &= \frac{1}{n_{1+}}(x_{1++})^2 + \frac{1}{n_{2+}}(x_{2++})^2 - \frac{1}{n}(x_{+++})^2 \\ &= \frac{1}{14}146{,}2^2 + \frac{1}{21}247{,}8^2 - \frac{1}{35}394^2 \\ &= 15{,}47 \end{aligned}$$

et de même pour $SCF_B = 33{,}66$.
On a $SCF_{AB} = SCT - SCF_A - SCF_B - SCR = 23{,}67$.

Solution alternative pour le calcul de SCF_A et SCF_B

On a signalé une deuxième méthode pour calculer ces carrés factoriels. Mettons la en application ici ; on calcule : $SCA_1 = x_{1++}^2 - \frac{1}{n_{1+}}(x_{1++})^2 = 1621{,}44 - \frac{1}{14}146{,}2^2 = 94{,}69$, $SCA_2 = 77{,}10$, puis $SCF_A = SCT - SCA_1 - SCA_2 = 15{,}47$;

De même $SCB_1 = 52{,}55$, $SCB_2 = 34{,}62$, $SCB_3 = 66{,}44$, puis $SCF_B = SCT - SCB_1 - SCB_2 - SCB_3 = 33{,}66$.
On vérifie que les résultats sont identiques.

Table d'anova

On peut remplir la table d'anova :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	15,47	1	15,47	F = 3,92
Traitement	33,66	2	16,83	F = 4,26
Interaction	23,67	2	11,83	F = 2,99
Résidus	114,46	29	3,95	
Total	187,27	34		

Après comparaison aux valeurs critiques : $F_{0,95}^{1,29} = 4{,}2$ et $F_{0,95}^{2,29} = 3{,}3$, on constate que seul l'effet du traitement est significatif.

Remarquons que si on ne tient pas compte du sexe des patients, dans ces données, une anova à un facteur donne la table suivante :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Traitement	33,66	2	16,83	F = 3,51
Résidus	153,61	32	4,80	
Total	187,27	34		

La valeur critique est $F_{0,95}^{1,32} = 3{,}3$: le test reste significatif, mais on est beaucoup plus près du seuil. Ainsi de façon générale, la prise en compte de facteurs supplémentaires augmente la puissance du test d'un facteur donné (ici le traitement), même si ces facteurs supplémentaires ont un effet trop faible pour être détecté.

13.3 Modèle mixte

On ne traite ici le modèle mixte que dans le cas des données équiréparties, c'est-à-dire que tous les n_{ij} sont égaux à r (le nombre de répétitions par cellule).

Le second facteur (le facteur bloc) sera considéré comme à effet aléatoire. C'est pertinent en particulier dans un plan d'expérience où le premier facteur est un traitement médical, et où les blocs sont les patients (considérés comme tiré au hasard dans une population de patients).

Les données sont donc des variables aléatoires X_{ijk}, $i = 1,\dots,p$, $j = 1,\dots,q$, $k = 1,\dots,r$.

On a $n_{i+} = qr$, $n_{+j} = pr$ et $n = n_{++} = pqr$.

13.3.1 Le modèle

Le modèle est

$$\boxed{X_{ijk} = \mu + \alpha_i + B_j + C_{ij} + E_{ijk},}$$

avec $\sum_i \alpha_i = 0$, $B_j \sim \mathcal{N}(0,\sigma_B^2)$, $C_{ij} \sim \mathcal{N}(0,\sigma_C^2)$ et $E_{ijk} \sim \mathcal{N}(0,\sigma^2)$, indépendantes.

Le modèle présenté ici est appelé « modèle non-contraint ». On aurait pu considérer le « modèle contraint » qui impose que la somme des termes d'interaction $B_{+j} = \sum_i B_{ij}$ dans un bloc donné est nulle ; dans ce cas les B_{ij} ne sont plus indépendants, il faut introduire des termes de covariance, ce qui complique la présentation du modèle.

Il y a un petit ambiguïté du fait que dans le modèle contraint, la statistique de test pour $\sigma_B^2 = 0$ est CMF_B/CMR ; dans le modèle non-contraint, c'est, comme nous allons le voir, CMF_B/CMF_{AB}. Quel est alors le bon test ? Nous y reviendrons.

Les hypothèses à tester sont

- $H_{0A} : \alpha_1 = \cdots = \alpha_p = 0$ (pas d'effet du facteur A : le traitement)
- $H_{0B} : \sigma_B^2 = 0$ (pas d'effet « principal » du facteur B : le bloc)
- $H_{0AB} : \sigma_C^2 = 0$ (pas d'interaction entre A et B)
- $H'_{0B} : \sigma_B^2 = \sigma_C^2 = 0$ (pas d'effet de B, pas d'interaction : le facteur A suffit à lui seul à expliquer les données observées).

13.3.2 Les sommes de carrés

Les données étant équiréparties, l'écriture des sommes de carrés se simplifie :

$$\begin{array}{rcl|rcl}
SCT & = & \displaystyle\sum_{ijk}(X_{ijk} - X_{\bullet\bullet\bullet})^2 & CMT & = & \dfrac{SCT}{pqr-1} \\
SCF_A & = & \displaystyle qr\sum_{i=1}^{p}(X_{i\bullet\bullet} - X_{\bullet\bullet\bullet})^2 & CMF_A & = & \dfrac{SCF_A}{p-1} \\
SCF_B & = & \displaystyle pr\sum_{j=1}^{q}(X_{\bullet j\bullet} - X_{\bullet\bullet\bullet})^2 & CMF_B & = & \dfrac{SCF_A}{q-1} \\
SCF_{AB} & = & \displaystyle r\sum_{i=1}^{p}\sum_{j=1}^{q}(X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet})^2 & CMF_{AB} & = & \dfrac{SCF_{AB}}{(p-1)(q-1)} \\
SCR & = & \displaystyle\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(X_{ijk} - X_{ij\bullet})^2 & CMR & = & \dfrac{SCR}{pq(r-1)}.
\end{array}$$

Les sommes de carrés sont toujours liées par la relation

$$SCT = SCF_A + SCF_B + SCF_{AB} + SCR.$$

et leurs degrés de liberté par

$$pqr - 1 = (p-1) + (q-1) + (p-1)(q-1) + pq(r-1).$$

13.3.3 Les lois

Notons qu'on a $B_+ \sim \mathcal{N}(0, q\sigma_B^2)$, $C_{i+} \sim \mathcal{N}(0, q\sigma_C^2)$ et $C_{++} \sim \mathcal{N}(0, pq\sigma_C^2)$.
On a également $E_{ij+} \sim \mathcal{N}(0, r\sigma^2)$ et $E_{i++} \sim \mathcal{N}(0, qr\sigma^2)$.
À partir des égalités suivantes :

$$\begin{aligned}
X_{i\bullet\bullet} &= \mu + \alpha_i + \tfrac{1}{q}B_+ + \tfrac{1}{q}C_{i+} + \tfrac{1}{qr}E_{i++} \\
X_{\bullet j\bullet} &= \mu + B_j + \tfrac{1}{p}C_{+j} + \tfrac{1}{pr}E_{+j+} \\
X_{\bullet\bullet\bullet} &= \mu + \tfrac{1}{q}B_+ + \tfrac{1}{pq}C_{++} + \tfrac{1}{pqr}E_{+++} \\
(X_{ij\bullet} - X_{i\bullet\bullet} - X_{\bullet j\bullet} + X_{\bullet\bullet\bullet}) &= \left(C_{ij} - \tfrac{1}{q}C_{i+} - \tfrac{1}{p}C_{+j} + \tfrac{1}{pq}C_{++}\right) + \tfrac{1}{r}\left(E_{ij+} - \tfrac{1}{q}E_{i++} - \tfrac{1}{p}E_{+j+} + \tfrac{1}{pq}E_{+++}\right)
\end{aligned}$$

on peut montrer que

$$\begin{aligned}
E(CMF_A) &= \tfrac{qr}{p-1}\textstyle\sum_{i=1}^{p}\alpha_i^2 + r\sigma_C^2 + \sigma^2 \\
E(CMF_B) &= pr\sigma_B^2 + r\sigma_C^2 + \sigma^2 \\
E(CMF_{AB}) &= r\sigma_C^2 + \sigma^2 \\
E(CMR) &= \sigma^2.
\end{aligned}$$

On a également

sous H_{0A} :	SCF_A	$\sim$	$(r\sigma_C^2 + \sigma^2)\chi^2(p-1)$
	SCF_B	$\sim$	$(pr\sigma_B^2 + r\sigma_C^2 + \sigma^2)\chi^2(q-1)$
sous H_{0B} :	SCF_B	$\sim$	$(r\sigma_C^2 + \sigma^2)\chi^2(q-1)$
sous H'_{0B} :	SCF_B	$\sim$	$\sigma^2\chi^2(q-1)$
	SCF_{AB}	$\sim$	$(r\sigma_C^2 + \sigma^2)\chi^2((p-1)(q-1))$
sous H_{0AB} :	SCF_{AB}	$\sim$	$\sigma^2\chi^2((p-1)(q-1))$
	SCR	$\sim$	$\sigma^2\chi^2(pq(r-1))$.

13.3.4 Les tests

On en déduit les statistiques de tests pour les diverses hypothèses nulles à tester :

$$\begin{aligned}
\text{sous } H_{0A}: \quad F_A &= \frac{CMF_A}{CMF_{AB}} \sim F(p-1,(p-1)(q-1)) \\
\text{sous } H_{0B}: \quad F_B &= \frac{CMF_B}{CMF_{AB}} \sim F(q-1,(p-1)(q-1)) \\
\text{sous } H'_{0B}: \quad F'_B &= \frac{CMF_B}{CMR} \sim F(q-1,pq(r-1)) \\
\text{sous } H_{0AB}: \quad F_{AB} &= \frac{CMF_{AB}}{CMR} \sim F((p-1)(q-1),pq(r-1))
\end{aligned}$$

On voit que la statistique de test de H'_{0B} est $\frac{CMF_B}{CMR}$, ce qui est la statistique de test pour l'absence d'effet de B dans le modèle contraint. Quand c'est possible, on testera de préférence H'_{0B} dans notre modèle, ce qui correspond à l'absence totale d'effet bloc (ni effet principal, ni interaction avec les traitements).

13.3.5 Cas particulier : une observation par cellule

Quand $r = 1$ (une seule observation par case), on ne peut plus tester l'absence d'interaction. Seules H_{0A} et H_{0B} sont testables. Comme dans le cas du modèle fixe, on peut choisir de « rebaptiser » SCF_{AB} en SCR et CMF_{AB} en CMR, de façon à avoir $SCT = SCF_A + SCF_B + SCR$.

Dans ce cas la distinction entre modèle à effets fixes et modèle mixte devient futile, les tests devenant identiques.

13.3.6 Réalisation pratique

Les formules de décentrage se simplifient un peu :

$$\begin{aligned}
\mathrm{SCT} &= \mathrm{X}_{+++}^2 - \frac{1}{n}(\mathrm{X}_{+++})^2 \\
\mathrm{SCF_A} &= \frac{1}{qr}\sum_{i=1}^{p}(\mathrm{X}_{i++})^2 - \frac{1}{n}(\mathrm{X}_{+++})^2 \\
\mathrm{SCF_B} &= \frac{1}{pr}\sum_{j=1}^{q}(\mathrm{X}_{+j+})^2 - \frac{1}{n}(\mathrm{X}_{+++})^2 \\
\mathrm{SCR} &= \sum_{ij}\mathrm{X}_{ijk}^2 - \frac{1}{r}(\mathrm{X}_{ij+})^2
\end{aligned}$$

Et on calcule $\mathrm{SCF_{AB}} = \mathrm{SCT} - \mathrm{SCF_A} - \mathrm{SCF_B} - \mathrm{SCR}$.

On présentera les résultats dans une table d'analyse de la variance :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Traitement	$\mathrm{SCF_A}$	$p-1$	$\mathrm{CMF_A} = \mathrm{SCF_A}/(p-1)$	$\mathrm{F} = \mathrm{CMF_A}/\mathrm{CMR}$
Blocs	$\mathrm{SCF_B}$	$q-1$	$\mathrm{CMF_B} = \mathrm{SCF_B}/(q-1)$	$\mathrm{F} = \mathrm{CMF_B}/\mathrm{CMR}$
Interaction	$\mathrm{SCF_{AB}}$	$(p-1)(q-1)$	$\mathrm{CMF_{AB}} = \mathrm{SCF_{AB}}/((p-1)(q-1))$	$\mathrm{F} = \mathrm{CMF_{AB}}/\mathrm{CMR}$
Résidus	SCR	$pq(r-1)$	$\mathrm{CMR} = \mathrm{SCR}/(pq(r-1))$	
Total	SCT	$n-1 = pqr-1$		

13.3.7 Exemple

Voici un exemple où on teste $p = 3$ traitements (effets fixes) sur $q = 12$ patients (effets aléatoires). On mesure l'efficacité des trois traitements selon une échelle donnée.

Dans la table ci-dessous on donne le résultat des mesures, ainsi que les sommes par lignes et par colonnes des mesures et de leurs carrés.

Patient	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	x_{i++}	x_{i++}^2
Trait. 1	5,2	5,5	7,1	5,9	5,0	7,1	6,4	7,4	6,5	7,0	6,5	6,0	75,6	483,14
Trait. 2	6,0	5,3	8,0	5,8	6,5	7,8	8,1	7,3	6,7	6,5	6,6	6,6	81,2	557,98
Trait. 3	6,5	6,5	8,1	6,0	6,5	7,6	7,5	7,5	6,6	7,1	6,6	6,7	83,2	581,04
x_{+j+}	17,7	17,3	23,2	17,7	18,0	22,5	22,0	22,2	19,8	20,6	19,7	19,3	240,0	
x_{+j+}^2	105,29	100,59	180,02	104,45	109,50	169,01	162,82	164,30	130,70	141,66	129,37	124,45		1622,12

Dans ce cas nous n'avons qu'une observation par cellule : on ne peut pas tester l'interaction. On suppose donc qu'il n'y en a pas (cf 13.3.5).

La somme des carrés factoriels associée au traitement est

$$\begin{aligned}
\mathrm{SCF_A} &= \tfrac{1}{12}\left((x_{1++})^2 + (x_{2++})^2 + (x_{3++})^2\right) - \tfrac{1}{n}(\mathrm{X}_{+++})^2 \\
&= \tfrac{1}{12}(75{,}6^2 + 81{,}2^2 + 83{,}2^2) - \tfrac{1}{36}240^2 \\
&= 2{,}587.
\end{aligned}$$

De même on calcule la SCF associée aux patients :

$$\begin{aligned}
\mathrm{SCF_B} &= \tfrac{1}{3}\left((x_{+1+})^2 + \cdots + (x_{+12+})^2\right) - \tfrac{1}{n}(\mathrm{X}_{+++})^2 \\
&= \tfrac{1}{3}(17{,}7^2 + 17{,}3^2 + 23{,}2^2 + \cdots) - \tfrac{1}{36}240^2 \\
&= 16{,}06.
\end{aligned}$$

La somme des carrés totaux est $SCT = x_{+++}^2 - \frac{1}{n}(x_{+++})^2 = 1622{,}16 - \frac{1}{36}240^2 = 22{,}16$.
La somme des carrés résiduels est obtenue dans ce cas par soustraction : $SCR = SCT - SCF_A - SCF_B = 3{,}513$.
On peut réaliser la table d'anova :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Traitement	2,587	2	1,293	F = 8,13
Patients	16,06	11	1,46	F = 9,18
Résidus	3,513	22	0,159	
Total	22,16	35		

Au seuil 5% la valeur critique pour l'effet traitement est (avec 2 et 22 ddl) vaut environ 3,44, et pour l'effet patient (avec 11 et 22 ddl) 2,25. On voit que les deux facteurs ont un effet.

13.4 Exercices

Exercice 1 Parmi les plans d'expérience suivants, lesquels sont équilibrés ?

	Niveau B_1	Niveau B_2	Niveau B_3
Niveau A_1	$n_{11} = 2$	$n_{12} = 2$	$n_{13} = 2$
Niveau A_2	$n_{21} = 3$	$n_{22} = 3$	$n_{23} = 3$
Niveau A_3	$n_{31} = 4$	$n_{32} = 4$	$n_{33} = 4$

Plan 1.

	Niveau B_1	Niveau B_2	Niveau B_3
Niveau A_1	$n_{11} = 2$	$n_{12} = 10$	$n_{13} = 4$
Niveau A_2	$n_{21} = 4$	$n_{22} = 20$	$n_{23} = 8$
Niveau A_3	$n_{31} = 3$	$n_{32} = 40$	$n_{33} = 6$

Plan 2.

	Niveau B_1	Niveau B_2	Niveau B_3
Niveau A_1	$n_{11} = 2$	$n_{12} = 4$	$n_{13} = 6$
Niveau A_2	$n_{21} = 10$	$n_{22} = 20$	$n_{23} = 30$

Plan 3.

Exercice 2 On applique un traitement à trois niveaux à des patients hommes et femmes, en tenant compte d'un éventuel effet du sexe des patients. Les mesures de l'efficacité du traitement sont rassemblées dans le tableau suivant.

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$ $x_{11+} = 53{,}5$ $x_{11+}^2 = 733{,}14$	$n_{12} = 4$ $x_{12+} = 56{,}2$ $x_{12+}^2 = 799{,}24$	$n_{13} = 6$ $x_{13+} = 79{,}6$ $x_{13+}^2 = 1076{,}82$	$n_{1+} = 14$ $x_{1++} = 189{,}3$ $x_{1++}^2 = 2609{,}20$
Femmes	$n_{21} = 6$ $x_{21+} = 92{,}2$ $x_{21+}^2 = 1428{,}60$	$n_{22} = 6$ $x_{22+} = 84{,}8$ $x_{22+}^2 = 1219{,}74$	$n_{23} = 9$ $x_{23+} = 147{,}2$ $x_{23+}^2 = 2433{,}50$	$n_{2+} = 21$ $x_{2++} = 324{,}2$ $x_{2++}^2 = 5081{,}84$
	$n_{+1} = 10$ $x_{+1+} = 145{,}7$ $x_{+1+}^2 = 2161{,}74$	$n_{+2} = 10$ $x_{+2+} = 141{,}0$ $x_{+2+}^2 = 2018{,}98$	$n_{+3} = 15$ $x_{+3+} = 226{,}8$ $x_{+3+}^2 = 3510{,}32$	$n_{++} = 35$ $x_{+++} = 513{,}5$ $x_{+++}^2 = 7691{,}04$

Tester l'effet du traitement, du sexe, et la présence d'une éventuelle interaction.

Exercice 3 On applique un traitement à trois niveaux à des patients hommes et femmes, en tenant compte d'un éventuel effet du sexe des patients. Les mesures de l'efficacité du traitement sont rassemblées dans le tableau suivant.

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 2$ $x_{11+} = 4{,}4$ $x^2_{11+} = 9{,}7$	$n_{12} = 3$ $x_{12+} = 13{,}7$ $x^2_{12+} = 65{,}15$	$n_{13} = 4$ $x_{13+} = 16{,}4$ $x^2_{13+} = 68{,}44$	$n_{1+} = 9$ $x_{1++} = 34{,}5$ $x^2_{1++} = 143{,}29$
Femmes	$n_{21} = 6$ $x_{21+} = 30{,}9$ $x^2_{21+} = 164{,}35$	$n_{22} = 9$ $x_{22+} = 51{,}2$ $x^2_{22+} = 301$	$n_{23} = 12$ $x_{23+} = 44$ $x^2_{23+} = 179{,}1$	$n_{2+} = 27$ $x_{2++} = 126{,}1$ $x^2_{2++} = 644{,}45$
	$n_{+1} = 8$ $x_{+1+} = 35{,}3$ $x^2_{+1+} = 174{,}05$	$n_{+2} = 12$ $x_{+2+} = 64{,}9$ $x^2_{+2+} = 366{,}15$	$n_{+3} = 16$ $x_{+3+} = 60{,}4$ $x^2_{+3+} = 247{,}54$	$n_{++} = 36$ $x_{+++} = 160{,}6$ $x^2_{+++} = 787{,}74$

Pour aider aux calculs de l'anova, on donne les résultats partiels suivants :

	Traitement A	Traitement B	Traitement C
Hommes			$SC_{13} = 1{,}2$
Femmes	$SC_{21} = 5{,}215$	$SC_{22} = 9{,}729$	$SC_{23} = 17{,}767$

1. Tester l'effet du traitement, du sexe, et la présence d'une éventuelle interaction.
2. Donner un intervalle de confiance à 95% sur l'efficacité du traitement A dans le groupe des femmes.

Exercice 4

On applique un traitement à trois niveaux à des patients hommes et femmes, en tenant compte d'un éventuel effet du sexe des patients. Les mesures de l'efficacité du traitement sont rassemblées dans le tableau suivant.

	Traitement A	Traitement B	Traitement C	
Hommes	$n_{11} = 4$ $x_{11+} = 19{,}5$ $x^2_{11+} = 97{,}59$	$n_{12} = 6$ $x_{12+} = 35{,}4$ $x^2_{12+} = 211{,}64$	$n_{13} = 6$ $x_{13+} = 47{,}5$ $x^2_{13+} = 378{,}87$	$n_{1+} = 16$ $x_{1++} = 102{,}4$ $x^2_{1++} = 688{,}1$
Femmes	$n_{21} = 6$ $x_{21+} = 24{,}7$ $x^2_{21+} = 105{,}91$	$n_{22} = 9$ $x_{22+} = 64{,}1$ $x^2_{22+} = 465{,}21$	$n_{23} = 9$ $x_{23+} = 68{,}1$ $x^2_{23+} = 521{,}45$	$n_{2+} = 24$ $x_{2++} = 156{,}9$ $x^2_{2++} = 1092{,}57$
	$n_{+1} = 10$ $x_{+1+} = 44{,}2$ $x^2_{+1+} = 203{,}5$	$n_{+2} = 15$ $x_{+2+} = 99{,}5$ $x^2_{+2+} = 676{,}85$	$n_{+3} = 15$ $x_{+3+} = 115{,}6$ $x^2_{+3+} = 900{,}32$	$n_{++} = 40$ $x_{+++} = 259{,}3$ $x^2_{+++} = 1780{,}67$

1. Le plan d'expérience est-il équilibré ?

2. Tester l'effet du traitement, du sexe, et la présence d'une éventuelle interaction, au seuil $\alpha = 0{,}05$. Pour faciliter le travail, on donne quelques valeurs des sommes de carrés SC_{ij}.

	Traitement A	Traitement B	Traitement C
Hommes	$SC_{11} = 2{,}528$		$SC_{13} = 2{,}828$
Femmes	$SC_{21} = 4{,}228$	$SC_{22} = 8{,}676$	

3. Donner un intervalle de confiance à 95% pour la différence d'effet entre les traitements B et C, chez les hommes d'une part, chez les femmes d'autre part.

Annexe A

Corrigé des exercices

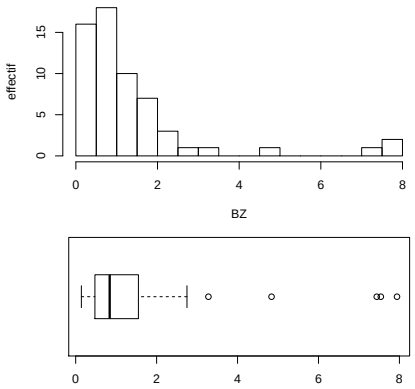
A.1 Exercices du chapitre 1

Corrigé de l'exercice 1

1. La moyenne est 1,39. On calcule la variance de l'échantillon $\tilde{S}^2 = 2{,}71$ d'où l'écart-type 1,64.

Les données sont triées et déjà séparées en quatre paquets dans la table : les quartiles sont 0,46 (premier quartile), 0,82 (médiane), 1,53 (troisième quartile) et l'IQR est IQR = 1,07.

2. On fait un histogramme avec par exemple des catégories limitées par les valeurs 0, 0,5, 1, etc. On l'aligne ici avec le boxplot, réalisé en suivant les conventions du cours.



La moustache supérieure s'arrête à 2,75, les 5 dernières mesures sont des mesures exceptionnelles (outliers).

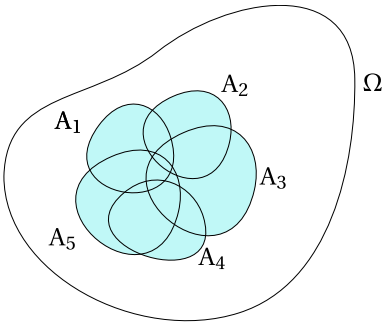
3. On a déjà calculé $m_2 = 2{,}71$. En utilisant la formule on trouve $m_3 = 12{,}71$ d'où on tire

$$\gamma_1 = 2{,}86.$$

Cette valeur correspond bien à une forte asymétrie à droite.

A.2 Exercices du chapitre 2

Corrigé de l'exercice 1 Il suffit du dessin suivant pour se convaincre de l'inégalité, en utilisant l'analogie entre probabilité et surface :



En effet quand on fait la somme des $\mathbb{P}(A_i)$, on compte plusieurs fois les morceaux qui sont dans plusieurs A_i.

On aura égalité quand les A_i sont deux à deux disjoints, c'est-à-dire quand il s'agit d'événements indépendants.

Corrigé de l'exercice 2 Première méthode : on procède par dénombrement : il y a $k!$ tirages possibles, ils sont équiprobables ; comptons le nombre de tirages favorables.

Il y a $k \times (k-1)$ positions différentes dans le tirage pour les boules 1 et 2, dont seulement la moitié est favorable (1 placé avant 2) ; pour chacune de ces $\frac{1}{2}k(k-1)$ positions des boules 1 et 2, il y a $(k-2)!$ permutations des $k-2$ boules restantes.

La probabilité cherchée est donc

$$p = \frac{\frac{1}{2}k(k-1) \times (k-2)!}{k!} = \frac{1}{2}$$

Deuxième méthode : on appelle Ω l'ensemble des tirages possibles, A l'ensemble des tirages favorables, et $B = \Omega \setminus A$. Il y autant de tirages dans A que dans B : on peut le voir en associant à un tirage $\omega \in B$ un tirage de $\omega' \in A$, obtenu est permutant les positions de 1 et 2. On a donc $|A| = |B|$, et $\Omega = |A| + |B| = 2|A|$. La probabilité cherchée est $p = \frac{|A|}{|\Omega|} = \frac{1}{2}$.

Corrigé de l'exercice 3 1. On calcule d'abord la probabilité de l'événement contraire $\overline{C}$: ne pas tirer une seule fois six.

À chaque tirage, on a une probabilité $\frac{5}{6}$ de tirer autre chose qu'un six. Les quatre tirages étant indépendants, on a

$$\mathbb{P}(\overline{C}) = \left(\frac{5}{6}\right)^4$$

et donc

$$\mathbb{P}(C) = 1 - \mathbb{P}(\overline{C}) = 1 - \left(\frac{5}{6}\right)^4 \simeq 0{,}51.$$

2. De même, pour le second pari, on trouve que la probabilité de gagner est

$$1 - \left(\frac{35}{36}\right)^{24} \simeq 0{,}49.$$

Le chevalier avait donc raison : ce second pari n'est pas avantageux, alors que le premier l'est.

Corrigé de l'exercice 4 1. On suppose tous les tirages équiprobables, et on applique la formule « nombre de tirages favorables / nombre de tirages possibles ».

$$\mathbb{P}(\text{as}) = \frac{4}{32} = \frac{1}{8}.$$

$$\mathbb{P}(\text{as noir}) = \frac{2}{32} = \frac{1}{16}.$$

Il y a 16 cartes noires, auxquelles il faut ajouter les deux as rouges, donc un total de 18 tirages favorables pour « as ou noir » :

$$\mathbb{P}(\text{as ou noir}) = \frac{18}{32} = \frac{9}{16}.$$

On pouvait également appliquer la formule $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$:

$$\begin{aligned} \mathbb{P}(\text{as ou noir}) &= \mathbb{P}(\text{as}) + \mathbb{P}(\text{noir}) - \mathbb{P}(\text{as noir}) \\ &= \frac{1}{8} + \frac{1}{2} - \frac{1}{16} = \frac{9}{16}. \end{aligned}$$

2. On tire un as avec probabilité $\frac{4}{32} = \frac{1}{8}$. Il reste pour le second tirage 31 cartes dont 4 rois, on tire donc un roi avec probabilité $\frac{4}{31}$. Pour finir, on a

$$\mathbb{P}(\text{as puis roi}) = \frac{1}{8} \times \frac{4}{31} = \frac{1}{62}.$$

Pour la question suivante, il faut distinguer le cas où on tire la dame de cœur, puis un cœur, ce qui arrive avec probabilité

$$\frac{1}{32} \times \frac{7}{31} = \frac{7}{31 \times 32},$$

du cas où on tire une des trois autres dames, puis un cœur, ce qui arrive avec probabilité

$$\frac{3}{32} \times \frac{8}{31} = \frac{24}{31 \times 32}.$$

Pour finir on a

$$\mathbb{P}(\text{dame puis cœur}) = \frac{7+24}{31 \times 32} = \frac{1}{32}.$$

3. Le cas avec remise est plus facile !

$$\mathbb{P}(\text{as puis roi}) = \frac{1}{8} \times \frac{4}{32} = \frac{1}{64}.$$

$$\mathbb{P}(\text{dame puis cœur}) = \frac{4}{32} \times \frac{8}{32} = \frac{1}{32}.$$

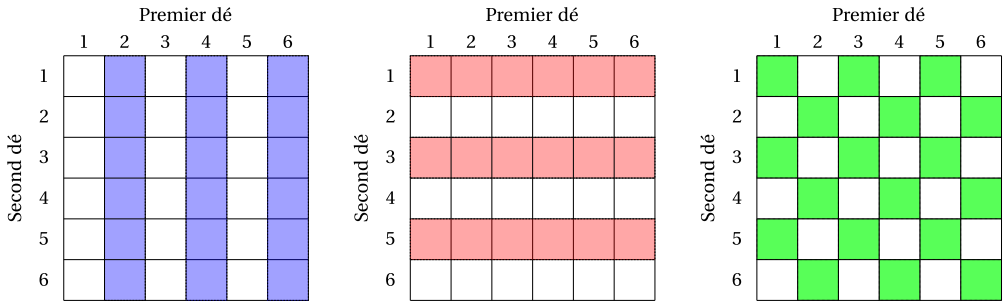
4. On se ramène à la question 2 en considérant que

$$\mathbb{P}(\text{as et roi}) = \mathbb{P}(\text{as puis roi}) + \mathbb{P}(\text{roi puis as}) = 2\mathbb{P}(\text{as puis roi}),$$

et donc $\mathbb{P}(\text{as et roi}) = \frac{1}{31}$, et de même $\mathbb{P}(\text{dame et cœur}) = \frac{1}{16}$.

Corrigé de l'exercice 5

Le corrigé se base sur la représentation de l'ensemble des résultats possibles sous la forme d'un tableau 6×6.



Les événements A, B et C sont ici représentés respectivement en bleu, rouge, et vert.

On voit que $|A| = |B| = [C| = 18$, d'où on déduit immédiatement que $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$.

En superposant les dessins, on voit que $|A \cap B| = |A \cap C| = |B \cap C| = 9$, d'où

$$\begin{array}{rcccl} \mathbb{P}(A \cap B) & = & \frac{1}{4} & = & \mathbb{P}(A)\mathbb{P}(B), \\ \mathbb{P}(A \cap C) & = & \frac{1}{4} & = & \mathbb{P}(A)\mathbb{P}(C), \\ \mathbb{P}(B \cap C) & = & \frac{1}{4} & = & \mathbb{P}(B)\mathbb{P}(C) \ : \end{array}$$

les événements sont deux à deux indépendants. Cependant, on a également $A \cap B \cap C = \emptyset$, d'où $\mathbb{P}(A \cap B \cap C) = 0$: les trois événements ne sont pas indépendants.

Corrigé de l'exercice 6

1. Le nombre de tirages possibles est le nombre de façons de choisir 4 boules parmi 14 :

$$\binom{14}{4} = \frac{14 \times 13 \times 12 \times 11}{4 \times 3 \times 2 \times 1} = 7 \times 13 \times 11 = 1001.$$

Le nombre de tirages avec 2 boules rouges et 2 boules vertes est égal au nombre de façons de choisir 2 boules parmi les 5 boules rouges, multiplié par le nombre de façons de choisir 2 boules parmi les 4 boules vertes :

$$\binom{5}{2} \times \binom{4}{2} = 10 \times 6 = 60.$$

La probabilité demandée est donc

$$\frac{60}{1001} \simeq 0{,}06.$$

2. De la même façon, le nombre de tirages avec 2 boules rouges, une boule noire, une boule verte est

$$\binom{5}{2} \times \binom{5}{1} \times \binom{4}{1} = 10 \times 5 \times 4 = 200.$$

La probabilité demandée est donc

$$\frac{200}{1001} \simeq 0{,}20.$$

3. Le nombre de tirages avec 4 boules rouges est

$$\binom{5}{4} = 5.$$

Il y a de même 5 tirages avec 4 boules noires ; et un seul tirage avec 4 boules vertes.

Le nombre de tirages monochromes est donc égal à $5 + 5 + 1 = 11$; la probabilité demandée est

$$\frac{11}{1001} \simeq 0{,}011.$$

Corrigé de l'exercice 7 Désignons les garçons par g_1, g_2, g_3 et les filles par f_1, f_2, f_3, f_4. L'ensemble Ω des résultats possibles est l'ensemble des permutations de $\{g_1, g_2, g_3, f_1, f_2, f_3, f_4\}$. Il a 7! éléments.

Les dispositions telles que chaque garçon soit entouré de deux filles sont de la forme

1	2	3	4	5	6	7
f	g	f	g	f	g	f

Il y a trois sièges (les sièges de numéros pairs) à faire occuper par les garçons et quatre (les sièges de numéro impair) par les filles ; et 3! façons de placer les garçons, 4! de placer les filles, soit en tout $3! \cdot 4!$ façons que tous s'asseoient et que chaque garçon soit entouré de deux filles.

La probabilité cherchée est donc

$$\frac{3!4!}{7!} = \frac{1 \cdot 2 \cdot 3 \cdot 1 \cdot 2 \cdot 3 \cdot 4}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7} = \frac{2 \cdot 3}{5 \cdot 6 \cdot 7} = \frac{1}{35}.$$

Corrigé de l'exercice 8 1. Calculons la probabilité de l'événement contraire : tous les étudiants sont nés un jour différent.

On numérote les étudiants de 1 à N de façon arbitraire, et les jours de l'année de 1 à 365. Les configurations possibles pour les jours de naissance sont les $\omega = (\omega_1,\ldots,\omega_N)$ où tous les ω_1 sont des nombres de 1 à 365. Il y en a 365^N possibles, toutes équiprobables.

Il faut dénombrer les ω pour lesquels il n'y a pas deux étudiants i, j avec $\omega_i = \omega_j$. Il y a 365 choix possibles pour ω_1, puis 364 pour ω_2 (on doit éviter le jour de naissance du premier étudiant), 363 pour ω_3, etc. Ainsi il y a au final

$$365 \times 364 \times \cdots \times (365 - N + 1)$$

Ainsi la probabilité que tous les étudiants soient nés un jour différent est

$$q_N = 365 \times 364 \times \cdots \times (365 - N + 1) \times \frac{1}{365^N} = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - N + 1}{365},$$

et on a $p_N = 1 - q_N$.

2. On peut réécrire

$$\begin{aligned} q_N &= (1-0) \times \left(1 - \frac{1}{365}\right) \times \cdots \times \left(1 - \frac{N-1}{365}\right) \\ &\simeq \exp\left(-\frac{1}{365} \sum_{i=0}^{N-1} i\right) \\ &= \exp\left(-\frac{N(N-1)}{2 \cdot 365}\right), \end{aligned}$$

et avec N = 30, on obtient $q_N \simeq \exp(-1{,}19) \simeq 0{,}30$, soit $p_N \simeq 0{,}70$.

Cette probabilité de 70% d'avoir deux étudiants fêtant leur anniversaire le même jour dans une classe de trente étudiants est beaucoup plus élevée que ce que l'intuition laisse généralement penser la première fois qu'on rencontre ce problème, d'où le nom de « paradoxe des anniversaires » pour ce problème classique.

Corrigé de l'exercice 9

1. Notons A_p l'événement « avoir reçu l'allèle A du gamète paternel » et A_m l'événement « avoir reçu l'allèle A du gamète maternel ». On défini de façon analogue a_p et a_m pour l'allèle a. On a $\mathbb{P}(A_p) = \mathbb{P}(A_m) = p$ et $\mathbb{P}(a_p) = \mathbb{P}(a_m) = q$.

Un individu est AA et A_p et A_m sont vrais, donc, par indépendance des allèles

$$\mathbb{P}(AA) = \mathbb{P}(A_p \text{ et } A_m) = \mathbb{P}(A_p)\mathbb{P}(A_m) = p^2.$$

De même $\mathbb{P}(aa) = q^2$ et

$$\mathbb{P}(Aa) = \mathbb{P}\big((A_p \text{ et } a_m) \text{ ou } (a_p \text{ et } A_m)\big) = 2pq.$$

2. L'individu a le phénotype dominant si il est AA ou Aa, donc la probabilité demandée est

$$\begin{aligned} \mathbb{P}(Aa | AA \text{ ou } Aa) &= \frac{\mathbb{P}(Aa)}{\mathbb{P}(AA \text{ ou } Aa)} \\ &= \frac{2pq}{1-q^2} = \frac{2q}{1+q}. \end{aligned}$$

Si q est petit (cas d'une maladie récessive rare), c'est $\simeq 2q$.

3. On calcule donc d'une part

$$\mathbb{P}(E_1 \cap (E_0 \cup E_1) \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) = \mathbb{P}(E_1 \cap (P_0 \cup P_1) \cap (M_0 \cup M_1))$$

et d'autre part

$$\mathbb{P}((E_0 \cup E_1) \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) = \mathbb{P}(E_0 \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) + \mathbb{P}(E_1 \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)).$$

Tout d'abord

$$\begin{aligned}\mathbb{P}(E_1 \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) &= \mathbb{P}(E_1 \cap P_0 \cap M_0) + \mathbb{P}(E_1 \cap P_0 \cap M_1) + \mathbb{P}(E_1 \cap P_1 \cap M_0) + \mathbb{P}(E_1 \cap P_1 \cap M_1)\\ &= \mathbb{P}(E_1|P_0 \cap M_0)\mathbb{P}(P_0 \cap M_0) + \mathbb{P}(E_1|P_0 \cap M_1)\mathbb{P}(P_0 \cap M_1)\\ &+ \mathbb{P}(E_1|P_1 \cap M_0)\mathbb{P}(P_1 \cap M_0) + \mathbb{P}(E_1|P_1 \cap M_1)\mathbb{P}(P_1 \cap M_1)\\ &= 0 + \frac{1}{2}\mathbb{P}(P_0)\mathbb{P}(M_1) + \frac{1}{2}\mathbb{P}(P_1)\mathbb{P}(M_0) + \frac{1}{2}\mathbb{P}(P_0)\mathbb{P}(M_1) + \frac{1}{2}\mathbb{P}(P_1)\mathbb{P}(M_1)\end{aligned}$$

et

$$\begin{aligned}\mathbb{P}(E_0 \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) &= \mathbb{P}(E_0|P_0 \cap M_0)\mathbb{P}(P_0 \cap M_0) + \mathbb{P}(E_0|P_0 \cap M_1)\mathbb{P}(P_0 \cap M_1)\\ &+ \mathbb{P}(E_0|P_1 \cap M_0)\mathbb{P}(P_1 \cap M_0) + \mathbb{P}(E_0|P_1 \cap M_1)\mathbb{P}(P_1 \cap M_1)\\ &= \mathbb{P}(P_0 \cap M_0) + \frac{1}{2}\mathbb{P}(P_0)\mathbb{P}(M_1) + \frac{1}{2}\mathbb{P}(P_1)\mathbb{P}(M_0) + \frac{1}{4}\mathbb{P}(P_1)\mathbb{P}(M_1).\end{aligned}$$

Donc

$$\mathbb{P}((E_0 \cup E_1) \cap (P_0 \cup P_1) \cap (M_0 \cup M_1)) = \mathbb{P}(P_0 \cap M_0) + \mathbb{P}(P_0)\mathbb{P}(M_1) + \mathbb{P}(P_1)\mathbb{P}(M_0) + \frac{3}{4}\mathbb{P}(P_1)\mathbb{P}(M_1),$$

et la probabilité recherchée est

$$\begin{aligned}&\frac{\frac{1}{2}\mathbb{P}(P_0)\mathbb{P}(M_1) + \frac{1}{2}\mathbb{P}(P_1)\mathbb{P}(M_0) + \frac{1}{2}\mathbb{P}(P_0)\mathbb{P}(M_1) + \frac{1}{2}\mathbb{P}(P_1)\mathbb{P}(M_1)}{\mathbb{P}(P_0 \cap M_0) + \mathbb{P}(P_0)\mathbb{P}(M_1) + \mathbb{P}(P_1)\mathbb{P}(M_0) + \frac{3}{4}\mathbb{P}(P_1)\mathbb{P}(M_1)}\\ =&\frac{\frac{1}{2}(p^2 \times 2pq + 2pq \times p^2 + 2pq \times 2pq)}{p^2 \times p^2 + p^2 \times 2pq + 2pq \times p^2 + \frac{3}{4}2pq \times 2pq}\\ =&\frac{2p^2q(p+q)}{p^2(p^2+4pq+3q^2)}\\ =&\frac{2q}{1+2q}.\end{aligned}$$

On voit que si q est petit, l'information apportée par le phénotype des parents est faible, car comme auparavant cette proba est $\simeq 2q$.

Corrigé de l'exercice 10 Notons F_0, F_1 et F_2 les événements « deux garçons », « un garçon, une fille » et « deux filles ».
On a $\mathbb{P}(F_0) = \mathbb{P}(F_2) = \frac{1}{4}$ et $\mathbb{P}(F_1) = \frac{1}{2}$. On calcule

$$\mathbb{P}(F_1|F_1 \text{ ou } F_2) = \frac{\mathbb{P}(F_1)}{\mathbb{P}(F_1 \text{ ou } F_0)} = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}} = \frac{2}{3}.$$

Corrigé de l'exercice 11 Dans la stratégie où on ne change pas de choix, on gagne quand on a choisit la bonne porte en première phase de jeu, avec une probabilité $\frac{1}{3}$.
Dans la stratégie où on change de choix :
– si on a choisi la bonne porte en premier lieu ($p = \frac{1}{3}$), on perd ;
– si on a choisi une des deux mauvaises portes en premier lieu ($p = \frac{2}{3}$), le présentateur ouvre celle des deux portes restantes qui cache une chèvre, et le changement de choix est alors profitable.
Cette stratégie est donc gagnante avec une probabilité $\frac{2}{3}$.

Corrigé de l'exercice 12 1. Il y a 10 tirages pairs possibles équiprobable, donc c'est $\frac{1}{10}$. On encore, en appelant X le résultat et avec la formule des probabilités conditionnelles :

$$\mathrm{P}(X = 2|X \text{ pair}) = \frac{\mathrm{P}(X = 2 \text{ et } X \text{ pair})}{\mathrm{P}(X \text{ pair})} = \frac{\mathrm{P}(X = 2)}{\mathrm{P}(X \text{ pair})} = \frac{1/20}{1/2} = \frac{1}{10}.$$

2. On appelle X et Y les résultats des deux tirages.

$$\begin{aligned} \mathrm{P}(\mathrm{X}=1 \text{ ou } \mathrm{Y}=1) &= 1-\mathrm{P}(\mathrm{X}\neq 1 \text{ et } \mathrm{Y}\neq 1) \\ &= 1-\mathrm{P}(\mathrm{X}\neq 1)\mathrm{P}(\mathrm{Y}\neq 1) \quad \text{car X et Y sont indépendantes} \\ &= 1-\left(\tfrac{19}{20}\right)^2 \\ &= 0{,}0975. \end{aligned}$$

Corrigé de l'exercice 13 1. Pour $k=1$, on a évidemment $p=\frac{1}{20}=0{,}05$. Pour $k=2$, on peut raisonner par dénombrement : il y a $20\times 20=400$ tirages possibles, dont $19\times 19=361$ tirages qui ne donnent pas droit à un gateau (faites un tableau des tirages si vous ne voyez pas pourquoi). On a donc $400-361=39$ tirages gagnants, et $p=\frac{39}{400}=0{,}0975$.

2. Il est plus facile de calculer la probabilité $q=1-p$ de ne pas avoir de gateau : c'est celle de n'avoir tiré 1 à aucun des k tirages, donc par indépendance des tirages $q=\left(\frac{19}{20}\right)^k$. On a donc $p=1-\left(\frac{19}{20}\right)^k$.

On aura $p>0{,}9$ si $\left(\frac{19}{29}\right)^k<0{,}1$. On passe aux logarithmes :

$$\begin{aligned} \left(\frac{19}{20}\right)^k &< 0{,}1 \\ k\log\left(\frac{19}{20}\right) &< \log 0{,}1 \\ -0{,}0513k &< -2{,}303 \\ k &> \frac{2{,}303}{0{,}0513}=44{,}8 \\ k &\geqslant 45 \end{aligned}$$

Quand k tend vers l'infini, $\left(\frac{19}{20}\right)^k$ tend vers 0 (car $\frac{19}{20}$ est plus petit que 1), et donc p tend vers 1.

A.3 Exercices du chapitre 3

Corrigé de l'exercice 1

$$\begin{aligned} \mathrm{E}(a\mathrm{X}+b) &= \sum_k (a\cdot k+b)\mathbb{P}(\mathrm{X}=k) \\ &= a\left(\sum_k k\mathbb{P}(\mathrm{X}=k)\right)+b\left(\sum_k \mathbb{P}(\mathrm{X}=k)\right) \\ &= a\mathrm{E}(\mathrm{X})+b. \end{aligned}$$

Pour l'espérance de X + Y on peut utiliser la formule faisant intervenir l'ensemble d'expériences Ω :

$$\begin{aligned} \mathrm{E}(\mathrm{X}+\mathrm{Y}) &= \sum_{\omega\in\Omega} \mathbb{P}(\omega)\big(\mathrm{X}(\omega)+\mathrm{Y}(\omega)\big) \\ &= \sum_{\omega\in\Omega} \mathbb{P}(\omega)\mathrm{X}(\omega)+\sum_{\omega\in\Omega} \mathbb{P}(\omega)\mathrm{Y}(\omega) \\ &= \mathrm{E}(\mathrm{X})+\mathrm{E}(\mathrm{Y}). \end{aligned}$$

On peut également introduire la fonction de masse conjointe de (X,Y) :

$$p(x,y)=\mathbb{P}(\mathrm{X}=x,\ \mathrm{Y}=y).$$

On a alors $\mathbb{P}(X = x) = \sum_y p(x,y)$ et $P(Y = y) = \sum_x p(x,y)$. On a également :

$$\begin{aligned} E(X+Y) &= \sum_{x,y}(x+y)p(x,y) \\ &= \sum_{x,y} xp(x,y) + yp(x,y) \\ &= \sum_x x\sum_y p(x,y) + \sum_y y\sum_x p(x,y) \\ &= \sum_x x\mathbb{P}(X=x) + \sum_y yP(Y=y) \\ &= E(X)+E(Y). \end{aligned}$$

Enfin,

$$\begin{aligned} E\left((aX+b)^2\right) &= E\left(a^2X^2+2abX+b^2\right) \\ &= a^2E\left(X^2\right)+2abE(X)+b^2 \end{aligned}$$

et

$$\begin{aligned} E(aX+b)^2 &= (aE(X)+b)^2 \\ &= a^2E(X)^2+2abE(X)+b^2 \end{aligned}$$

d'où

$$\mathrm{var}(aX+b) = a^2E\left(X^2\right) - a^2E(X)^2 = a^2\,\mathrm{var}(X).$$

D'autre part

$$E\left((X+Y)^2\right) = E(X^2)+2E(XY)+E(Y^2),$$

d'où $\mathrm{var}(X+Y) = \mathrm{var}(X)+2\mathrm{cov}(X,Y)+\mathrm{var}(Y)$.

Corrigé de l'exercice 2 On a

$$\begin{aligned} E(X) &= \sum_k k\mathbb{P}(X=k) \\ &= 0\times\frac{1}{4}+1\times\frac{1}{2}+2\times\frac{1}{4} \\ &= 1. \end{aligned}$$

De même on trouve $E\left(X^2\right) = \sum_k k^2\mathbb{P}(X=k) = \frac{3}{2}$ et $E\left(2^X\right) = \frac{9}{4}$. On a $\mathrm{var}(X) = E\left(X^2\right) - E(X)^2 = \frac{1}{2}$.

Corrigé de l'exercice 3

1. La fonction est positive, il suffit de vérifier que son intégrale vaut 1 :

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_0^1 2x\,\mathrm{d}x = 1.$$

2. On a

$$E(X) = \int_0^1 x\times 2x\,\mathrm{d}x = \frac{2}{3}.$$

$$E\left(X^2\right) = \int_0^1 2x^3\,\mathrm{d}x = \frac{1}{2}.$$

$$E\left(e^{X^2}\right) = \int_0^1 e^{x^2}\times 2x\,\mathrm{d}x = \left[e^{x^2}\right]_0^1 = e-1.$$

On calcule $\mathrm{var}(X) = E\left(X^2\right) - E(X)^2 = \frac{1}{18}$.

Corrigé de l'exercice 4

1. La fonction est positive, il suffit de vérifier que son intégrale vaut 1 :

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_0^1 3x^2\,\mathrm{d}x = 1.$$

2. On a

$$\mathrm{E(X)} = \int_0^1 x \times 3x^2\,\mathrm{d}x = \frac{3}{4}.$$

$$\mathrm{E}\left(\mathrm{X}^2\right) = \int_0^1 3x^4\,\mathrm{d}x = \frac{3}{5}.$$

On calcule $\mathrm{var(X)} = \mathrm{E}\left(\mathrm{X}^2\right) - \mathrm{E(X)}^2 = \frac{3}{80}$.

Corrigé de l'exercice 5 La somme des valeurs de deux dés pour chacun des 36 jets possibles et équiprobables sont énumérés dans le tableau ci-dessous.

Deuxième dé \ Premier dé	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Appelons S la somme des deux dés. Il suffit d'énumérer les cases pour trouver la loi de S :

$$\begin{array}{lll}
\mathbb{P}(\mathrm{S}=2) = \frac{1}{36} & \mathbb{P}(\mathrm{S}=6) = \frac{5}{36} & \mathbb{P}(\mathrm{S}=10) = \frac{3}{36} \\
\mathbb{P}(\mathrm{S}=3) = \frac{2}{36} & \mathbb{P}(\mathrm{S}=7) = \frac{6}{36} & \mathbb{P}(\mathrm{S}=11) = \frac{2}{36} \\
\mathbb{P}(\mathrm{S}=4) = \frac{3}{36} & \mathbb{P}(\mathrm{S}=8) = \frac{5}{36} & \mathbb{P}(\mathrm{S}=12) = \frac{1}{36} \\
\mathbb{P}(\mathrm{S}=5) = \frac{4}{36} & \mathbb{P}(\mathrm{S}=9) = \frac{4}{36} &
\end{array}$$

Si on appelle X et Y les deux variables indépendantes correspondant au jet de chacun des dés, définies par $\mathbb{P}(\mathrm{X}=k) = \mathbb{P}(\mathrm{Y}=k) = \frac{1}{6}$ pour $k = 1,\ldots,6$, on a $\mathrm{S} = \mathrm{X}+\mathrm{Y}$. On constate que la formule du produit de convolution :

$$\mathbb{P}(\mathrm{S}=m) = \sum_{k+\ell=m} \mathbb{P}(\mathrm{X}=k)\mathbb{P}(\mathrm{Y}=\ell)$$

n'est qu'un moyen élaboré de résumer cette procédure : par exemple pour $m = 4$

$$\mathbb{P}(\mathrm{S}=4) = \mathbb{P}(\mathrm{X}=1)\mathbb{P}(\mathrm{Y}=3) + \mathbb{P}(\mathrm{X}=2)\mathbb{P}(\mathrm{Y}=2) + \mathbb{P}(\mathrm{X}=3)\mathbb{P}(\mathrm{Y}=1) = \frac{3}{36}.$$

Corrigé de l'exercice 6 On procède de la même façon qu'à l'exercice précédent en prenant $\mathrm{T} = \mathrm{S}+\mathrm{X}$ avec S la somme de deux dés, dont la loi a été calculée à l'exercice précédent, et X (indépendante de S), le résultat d'un jet de dé. On peut refaire un tableau, mais attention : cette fois les cases ne sont plus équiprobables, les 11 valeurs possibles pour la somme de deux dés n'étant pas équiprobables.

Deux premiers dés

$p=$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
	2	3	4	5	6	7	8	9	10	11	12
1	3	4	5	6	7	8	9	10	11	12	13
2	4	5	6	7	8	9	10	11	12	13	14
3	5	6	7	8	9	10	11	12	13	14	15
4	6	7	8	9	10	11	12	13	14	15	16
5	7	8	9	10	11	12	13	14	15	16	17
6	8	9	10	11	12	13	14	15	16	17	18

Troisième dé

On aura donc par exemple $\mathrm{P}(\mathrm{T}=3)=\frac{1}{6}\times\frac{1}{36}=\frac{1}{216}$, puis $\mathrm{P}(\mathrm{T}=4)=\frac{1}{6}\times\frac{1}{36}+\frac{1}{6}\times\frac{2}{36}=\frac{3}{216}$, etc.

$$\begin{array}{llll}
\mathbb{P}(\mathrm{T}=3)=\frac{1}{216} & \mathbb{P}(\mathrm{T}=7)=\frac{15}{216} & \mathbb{P}(\mathrm{T}=11)=\frac{27}{216} & \mathbb{P}(\mathrm{T}=15)=\frac{10}{216}\\
\mathbb{P}(\mathrm{T}=4)=\frac{3}{216} & \mathbb{P}(\mathrm{T}=8)=\frac{21}{216} & \mathbb{P}(\mathrm{T}=12)=\frac{25}{216} & \mathbb{P}(\mathrm{T}=16)=\frac{6}{216}\\
\mathbb{P}(\mathrm{T}=5)=\frac{6}{216} & \mathbb{P}(\mathrm{T}=9)=\frac{25}{216} & \mathbb{P}(\mathrm{T}=13)=\frac{21}{216} & \mathbb{P}(\mathrm{T}=17)=\frac{3}{216}\\
\mathbb{P}(\mathrm{T}=6)=\frac{10}{216} & \mathbb{P}(\mathrm{T}=10)=\frac{27}{216} & \mathbb{P}(\mathrm{T}=14)=\frac{15}{216} & \mathbb{P}(\mathrm{T}=18)=\frac{1}{216}
\end{array}$$

À nouveau, on peut faire le lien avec la formule du produit de convolution.

L'erreur du Grand-Duc est de considérer que les 6 sommes sont équiprobables. Il y a par exemple, sur 216 lancers possibles, 6 façons d'obtenir le triplet $(6,2,1)$ dans un ordre quelconque, mais seulement 3 façons d'obtenir $(5,2,2)$, et une seule façon d'obtenir $(3,3,3)$. Avec ces considérations, en considérant les décompositions de 10 et 9 en somme de trois dés, que le Grand-Duc avait correctement trouvées, on peut vérifier qu'on a bien :

$$\begin{aligned}
\mathrm{P}(\mathrm{T}=10) &= \frac{6}{216}+\frac{3}{216}+\frac{6}{216}+\frac{6}{216}+\frac{3}{216}+\frac{3}{216} = \frac{27}{216}\\
\mathrm{P}(\mathrm{T}=9) &= \frac{6}{216}+\frac{6}{216}+\frac{3}{216}+\frac{3}{216}+\frac{6}{216}+\frac{1}{216} = \frac{25}{216}
\end{aligned}$$

Corrigé de l'exercice 7 1. On a $\mathbb{P}(\mathrm{X}_1=0)=\mathbb{P}(\mathrm{X}_1=1)=\frac{1}{2}$.

2. Par simple dénombrement, on a

$$\begin{aligned}
\mathbb{P}(\mathrm{X}_2=0) &= \mathbb{P}(\mathrm{X}_2=3)+\mathbb{P}(\mathrm{X}_2=6)+\cdots+\mathbb{P}(\mathrm{X}_2=18)\\
&= 6\times\tfrac{1}{20}=\tfrac{6}{20}.
\end{aligned}$$

De la même façon, $\mathbb{P}(\mathrm{X}_2=1)=\mathbb{P}(\mathrm{X}_2=2)=\frac{7}{20}$.

3. On a $\mathbb{P}(\mathrm{X}_1=0 \text{ et } \mathrm{X}_2=0)=\mathbb{P}(\mathrm{X}=6,12 \text{ ou } 18)=\frac{3}{20}$.

On procède de même pour toutes les combinaisons, pour trouver

$$\begin{array}{llll}
\mathbb{P}(\mathrm{X}_1=0 \text{ et } \mathrm{X}_2=0)=\frac{3}{20} & \mathbb{P}(\mathrm{X}_1=0 \text{ et } \mathrm{X}_2=1)=\frac{3}{20} & \mathbb{P}(\mathrm{X}_1=0 \text{ et } \mathrm{X}_2=2)=\frac{4}{20}\\
\mathbb{P}(\mathrm{X}_1=1 \text{ et } \mathrm{X}_2=0)=\frac{3}{20} & \mathbb{P}(\mathrm{X}_1=1 \text{ et } \mathrm{X}_2=1)=\frac{4}{20} & \mathbb{P}(\mathrm{X}_1=1 \text{ et } \mathrm{X}_2=2)=\frac{3}{20}
\end{array}$$

4. Les variables ne sont pas indépendantes, car par exemple $\mathbb{P}(\mathrm{X}_1=1 \text{ et } \mathrm{X}_2=1)\neq\mathbb{P}(\mathrm{X}_1=1)\mathbb{P}(\mathrm{X}_2=1)$.

Corrigé de l'exercice 8

1. Il y a

$$\begin{aligned}\binom{49}{5} &= \frac{49\times48\times47\times46\times45}{2\times3\times4\times5}\\ &= 49\times6\times47\times46\times3\\ &= 1\,906\,884\end{aligned}$$

tirages possibles.

2. Comptons les tirages contenant le numéro 1. Il faut choisir les 4 numéros restant parmi les numéros de 2 à 49, il y en a 48, ce qui donne

$$\begin{aligned}\binom{48}{4} &= \frac{48\times47\times46\times45}{2\times3\times4}\\ &= 2\times47\times46\times45\\ &= 194\,580\end{aligned}$$

tirages contenant le 1.

Comptons les tirages contenant le numéro 2 et pas le 1. Il faut choisir les 4 numéros restant parmi les numéros de 3 à 49, il y en a 47, ce qui donne

$$\begin{aligned}\binom{47}{4} &= \frac{47\times46\times45\times44}{2\times3\times4}\\ &= 47\times23\times15\times11\\ &= 178\,365\end{aligned}$$

tirages.

3. D'après ce qui précède, $\mathbb{P}(\mathrm{X}=1)=\frac{194\,580}{1\,906\,884}\simeq 0{,}1020$ et $\mathbb{P}(\mathrm{X}=2)=\frac{178\,365}{1\,906\,884}\simeq 0{,}0935$.

4. On généralise le raisonnement de la question 3 :

$$\mathbb{P}(\mathrm{X}=k)=\frac{\binom{49-k}{4}}{\binom{49}{5}}.$$

A.4 Exercices du chapitre 4

Corrigé de l'exercice 1 1. La loi de X s'obtient comme ceci :

$$\begin{aligned}\mathbb{P}(\mathrm{X}=1) &= \textstyle\sum_y \mathbb{P}(\mathrm{X}=1,\mathrm{Y}=y)=\frac{2}{18}+\frac{4}{18}=\frac{6}{18}=\frac{1}{3}\\ \mathbb{P}(\mathrm{X}=3) &= \tfrac{1}{18}+\tfrac{2}{16}=\tfrac{3}{18}=\tfrac{1}{6}\\ \mathbb{P}(\mathrm{X}=5) &= \tfrac{3}{18}+\tfrac{6}{18}=\tfrac{9}{18}=\tfrac{1}{2}\end{aligned}$$

On en déduit $\mathrm{E}(\mathrm{X})=\sum_x x\mathbb{P}(\mathrm{X}=x)=\frac{1}{6}(1\times2+3\times1+5\times3)=\frac{20}{6}=\frac{10}{3}$, et $\mathrm{E}\left(\mathrm{X}^2\right)=\sum_x x^2\mathbb{P}(\mathrm{X}=x)=\frac{43}{3}$. On a $\mathrm{var}(\mathrm{X})=\frac{43}{3}\left(\frac{10}{3}\right)^2=\frac{129-100}{9}=\frac{29}{9}$.

De même on a $\mathbb{P}(\mathrm{Y}=1)=\frac{1}{3}$ et $\mathbb{P}(\mathrm{Y}=2)=\frac{2}{3}$, d'où $\mathrm{E}(\mathrm{Y})=\frac{5}{3}$ et $\mathrm{var}(\mathrm{Y})=\frac{2}{9}$.

2. On a $\mathrm{E}(\mathrm{XY})=\sum_{x,y} x\cdot y\cdot\mathbb{P}(\mathrm{X}=x,\mathrm{Y}=y)=\frac{50}{9}$, d'où $\mathrm{cov}(\mathrm{X},\mathrm{Y})=\mathrm{E}(\mathrm{XY})-\mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y})=0$.

3. Les variables X et Y sont indépendantes : on vérifie facilement que pour tous x, y, on a $\mathbb{P}(\mathrm{X}=x,\mathrm{Y}=y)=\mathbb{P}(\mathrm{X}=x)\mathbb{P}(\mathrm{Y}=y)$.

Corrigé de l'exercice 2

1. En utilisant $\mathbb{P}(\mathrm{X}=x,\mathrm{Y}=y)=\mathbb{P}(\mathrm{X}=x)\mathbb{P}(\mathrm{Y}=y|\mathrm{X}=x)$, on obtient la table suivante.

$y=$	1	2	3
$x=1$	0,20	0,10	0,10
$x=2$	0,15	0,10	0,05
$x=3$	0,05	0,10	0,15

Table 24. Valeurs de $\mathbb{P}(\mathrm{X}=x,\mathrm{Y}=y)$

2. Il faut d'abors calculer la loi marginale de Y : on a $\mathbb{P}(Y=1)=0{,}4$, $\mathbb{P}(Y=2)=0{,}3$ et $\mathbb{P}(Y=3)=0{,}3$. On a alors $\mathbb{P}(X=1|Y=1)=0{,}20/0{,}40=0{,}5$, $\mathbb{P}(X=2|Y=1)=0{,}375$ et $\mathbb{P}(X=3|Y=1)=0{,}125$.

3. On calcule $E(X)=\sum_x x\mathbb{P}(X=x)=1\times 0{,}4+2\times 0{,}3+3\times 0{,}3=1{,}9$. De même on a $E(X^2)=4{,}3$ et on a $\mathrm{var}(X)=0{,}69$.

La loi de Y est la même que la loi de X, et on a donc $E(Y))=1{,}9$ et $\mathrm{var}(Y)=0{,}69$.

4. On commence par calculer la covariance. Tout d'abord on a $E(XY)=\sum_{x,y} xy\mathbb{P}(X=x,Y=y)=1\times 0{,}2+2\times 0{,}1+3\times 0{,}1+\cdots+9\times 0{,}15=3{,}8$, d'où $\mathrm{cov}(X,Y)=3{,}8-1{,}9\times 1{,}9=0{,}19$, et $\mathrm{cor}(X,Y)=0{,}19/\sqrt{0{,}69\times 0{,}69}\simeq 0{,}28$

Corrigé de l'exercice 3 1. La fonction f est partout positive. Il suffit de vérifier que son intégrale est égale à 1 :

$$\begin{aligned}\int_{x=-\infty}^{+\infty}\int_{y=-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_{x=0}^{1}\int_{y=0}^{1} 6xy^2\,\mathrm{d}x\,\mathrm{d}y\\ &= 6\int_{x=0}^{1}\left[\frac{1}{3}xy^3\right]_{y=0}^{1}\mathrm{d}x\\ &= 2\int_{x=0}^{1} x\,\mathrm{d}x=1.\end{aligned}$$

2. La densité de X est nulle en $x\notin[0,1]$. En $x\in[0,1]$, elle est obtenue par

$$\phi(x)=\int_{y=0}^{1} 6xy^2\,\mathrm{d}y=2x.$$

On en tire son espérance $E(X)=\frac{2}{3}$.

La densité de Y est de même, si $y\in[0,1]$, $\psi(y)=\int_{x=0}^{1}6xy^2\,\mathrm{d}x=3y^2$, et est nulle sinon. Son espérance est $E(Y)=\frac{3}{4}$.

On a

$$\begin{aligned}E(XY) &= \int_{x=0}^{1}\int_{y=0}^{1} xy\times 6xy^2\,\mathrm{d}x\,\mathrm{d}y\\ &= 6\int_{x=0}^{1} x^2\left[\frac{1}{4}y^4\right]_{y=0}^{1}\mathrm{d}x\\ &= \frac{3}{2}\int_{x=0}^{1} x^2\,\mathrm{d}x=\frac{1}{2}.\end{aligned}$$

Donc $\mathrm{cov}(X,Y)=E(XY)-E(X)E(Y)=0$.

3. La densité de X conditionnellement à $Y=y$ est (si $x,y\in[0,1]$)

$$\phi(x|Y=y)=\frac{f(x,y)}{\psi(y)}=2x=\phi(x).$$

On trouverait de même $\psi(y|X=x)=\psi(y)$. Les deux variables sont indépendantes; on a $f(x,y)=\phi(x)\psi(y)$.

Corrigé de l'exercice 4 1. La fonction f est partout positive. On vérifie que son intégrale est égale à 1 :

$$\begin{aligned}\int_{x=-\infty}^{+\infty}\int_{y=-\infty}^{+\infty} f(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_{x=0}^{1}\int_{y=0}^{1}(x+y)\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_{x=0}^{1}\left[xy+\frac{1}{2}y^2\right]_{y=0}^{1}\mathrm{d}x\\ &= \int_{x=0}^{1}\left(x+\frac{1}{2}\right)\mathrm{d}x=1.\end{aligned}$$

2. La densité de X est nulle pour $x\notin[0,1]$, et pour $x\in[0,1]$ on l'obtient par

$$\phi(x)=\int_{y=0}^{1}(x+y)\,\mathrm{d}y=x+\frac{1}{2}.$$

On calcule son espérance :

$$\mathrm{E}(\mathrm{X}) = \int_x x\phi(x)\,\mathrm{d}x = \int_0^1 \left(x^2 + \frac{1}{2}x\right)\mathrm{d}x = \frac{7}{12}.$$

d'autre part :

$$\mathrm{E}\left(\mathrm{X}^2\right) = \int_0^1 x^3 + \frac{1}{2}x^2\,\mathrm{d}x = \frac{5}{12}.$$

et $\mathrm{var}(\mathrm{X}) = \frac{11}{144}$.
La densité de Y est de même $\psi(y) = \int_{x=0}^1 (x+y)\,\mathrm{d}x = y + \frac{1}{2}$ pour $y \in [0,1]$: Y a même densité que X.
On a

$$\begin{aligned}\mathrm{E}(\mathrm{XY}) &= \int_{x=0}^1 \int_{y=0}^1 (x^2 y + x y^2)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{x=0}^1 \left[\frac{1}{2}x^2y^2 + \frac{1}{3}xy^3\right]_{y=0}^1 \\ &= \int_{x=0}^1 \left(\frac{1}{2}x^2 + \frac{1}{3}x\right) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.\end{aligned}$$

Et donc $\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y}) = \frac{1}{3} - \left(\frac{7}{12}\right)^2 = -\frac{1}{144}$.
3. La densité de X conditionnellement à $\mathrm{Y} = y$ est (pour $x, y \in [0,1]$)

$$\phi(x|\mathrm{Y} = y) = \frac{f(x,y)}{\psi(y)} = \frac{x+y}{y+\frac{1}{2}}.$$

Les deux variables ne sont pas indépendantes.

Corrigé de l'exercice 5 1. Découle immédiatement de la définition du min et du max.
2. On commence par la fonction de répartition de U :

$$\begin{aligned}\mathbb{P}(\mathrm{U} \leqslant t) &= \mathbb{P}(\mathrm{X} \leqslant t \text{ ou } \mathrm{Y} \leq t) \\ &= 1 - \mathbb{P}(\mathrm{X} > t \text{ et } \mathrm{Y} > t) \\ &= 1 - \mathbb{P}(\mathrm{X} > t)\,\mathbb{P}(\mathrm{Y} > t) \quad \text{(X, Y indépendantes)} \\ &= 1 - (1 - \mathrm{F}(t))(1 - \mathrm{G}(t)).\end{aligned}$$

On dérive pour obtenir la densité de U : $(1-\mathrm{F})g + (1-\mathrm{G})f$.
Écrivons la fonction de répartition de V :

$$\begin{aligned}\mathbb{P}(\mathrm{U} \leqslant t) &= \mathbb{P}(\mathrm{X} \leqslant t \text{ et } \mathrm{Y} \leq t) \\ &= \mathbb{P}(\mathrm{X} \leqslant t)\mathbb{P}(\mathrm{Y} \leq t) \quad \text{(X, Y indépendantes)} \\ &= \mathrm{F}(t)\mathrm{G}(t).\end{aligned}$$

D'où la densité de V : $f\mathrm{G} + \mathrm{F}g$.
3. On a dans ce cas

$$f(t) = g(t) = \begin{cases} 1 & \text{si } t \in [0,1] \\ 0 & \text{sinon.}\end{cases}$$

et

$$\mathrm{F}(t) = \mathrm{G}(t) = \begin{cases} t & \text{si } t \in [0,1] \\ 0 & \text{sinon.}\end{cases}$$

On en déduit la densité de U :

$$h_{\mathrm{U}}(t) = \begin{cases} 2(1-t) & \text{si } t \in [0,1] \\ 0 & \text{sinon.}\end{cases}$$

et celle de V :

$$h_{\mathrm{V}}(t) = \begin{cases} 2t & \text{si } t \in [0,1] \\ 0 & \text{sinon.}\end{cases}$$

Corrigé de l'exercice 6 1. La densité de X et Y est $f(t) = \begin{cases} 1 & \text{si } t \in [0,1] \\ 0 & \text{sinon.} \end{cases}$

Z ne peut prendre que des valeurs entre 0 et 2, sa densité est donc nulle en dehors de [0,2]. Pour $t \in [0,2]$, la densité de Z est donnée par le produit de convolution

$$\begin{aligned} h(t) &= \int_{-\infty}^{+\infty} f(x)f(t-x)\,\mathrm{d}x \\ &= \int_0^1 f(x)f(t-x)\,\mathrm{d}x \qquad (f(x) = 0 \text{ si } x \notin [0,1]) \end{aligned}$$

D'autre part $f(t-x)$ n'est non nulle que si $x \in [t-1,t]$; et donc

$$h(t) = \begin{cases} \displaystyle\int_0^t 1\,\mathrm{d}x = t & \text{si } t \in [0,1] \\ \displaystyle\int_{t-1}^1 1\,\mathrm{d}x = 2-t & \text{si } t \in [1,2]. \end{cases}$$

2. La densité de X et Y est $f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t \geqslant 0 \\ 0 & \text{sinon.} \end{cases}$

Z ne peut prendre que des valeurs positives, sa densité est nulle pour $t < 0$. Si $t \geqslant 0$, sa densité est

$$\begin{aligned} h(t) &= \int_{-\infty}^{+\infty} f(x)f(t-x)\,\mathrm{d}x \\ &= \int_0^{+\infty} f(x)f(t-x)\,\mathrm{d}x \qquad (f(x) = 0 \text{ si } x < 0) \\ &= \int_0^t \left(\lambda e^{-\lambda x}\right)\left(\lambda e^{-\lambda(t-x)}\right)\mathrm{d}x \quad (f(t-x) = 0 \text{ si } x > t) \\ &= \int_0^t \lambda^2 e^{-\lambda t}\,\mathrm{d}x \\ &= \lambda^2 t e^{-\lambda t}. \end{aligned}$$

On reconnaît la loi $\Gamma(2,\lambda)$.

A.5 Exercices du chapitre 5

Corrigé de l'exercice 1 On a $\mathbb{P}(X \geqslant t) = 1 - \mathbb{P}(X \leqslant t) = e^{-\lambda t}$.

u étant positif, on a $(X \geqslant t+u \text{ et } X \geqslant t) = (X \geqslant t+u)$. Donc :

$$\begin{aligned} \mathbb{P}(X \geqslant t+u | X \geqslant t) &= \frac{\mathbb{P}(X \geqslant t+u)}{\mathbb{P}(X \geqslant t)} \\ &= \frac{e^{-\lambda(t+u)}}{e^{-\lambda t}} \\ &= e^{-\lambda u}. \end{aligned}$$

On a donc $\mathbb{P}(X \geqslant t+u | X \geqslant t) = \mathbb{P}(X \geqslant u)$: dans un processus de Poisson, le fait d'avoir déjà attendu un temps t n'augmente pas la probabilité d'occurence d'un événement avant qu'un temps u se soit écoulé (on dit que le processus « n'a pas de mémoire »).

Corrigé de l'exercice 2 Soient f et g les densités de X et Y, et F, G leur fonction de répartition. On a vu dans un exercice de la fiche précédente que la densité de Z est $h = f \cdot (1-G) + g \cdot (1-F)$.

On a

$$\begin{array}{ll} f(x) = \lambda_1 e^{-\lambda_1 x} & g(x) = \lambda_2 e^{-\lambda_2 x} \\ F(x) = 1 - e^{-\lambda_1 x} & G(x) = 1 - e^{-\lambda_2 x} \end{array}$$

et donc $h(x) = \lambda_1 e^{-(\lambda_1+\lambda_2)x} + \lambda_2 e^{-(\lambda_1+\lambda_2)x}$. En posant $\lambda = \lambda_1 + \lambda_2$ on a $h(x) = \lambda e^{-\lambda x}$.

On voit que Z suit une loi $\exp(\lambda)$, ce qui correspond au résultat annoncé dans le paragraphe sur la superposition des processus de Poisson.

Corrigé de l'exercice 3 Soient f et F la densité et la fonction de répartition de X. La fonction de répartition de Y est

$$\begin{aligned}\mathbb{P}(\mathrm{Y} \leqslant x) &= \mathbb{P}(a\mathrm{X} \leqslant x)\\ &= \mathbb{P}\left(\mathrm{X} \leqslant \frac{x}{a}\right)\\ &= \mathrm{F}\left(\frac{x}{a}\right).\end{aligned}$$

On dérive pour trouver la densité de Y, qui est donc $\frac{1}{a} f\left(\frac{x}{a}\right) = \frac{\lambda}{a} e^{-\left(\frac{\lambda}{a}\right)x}$. On reconnaît la densité de la loi $\exp\left(\frac{\lambda}{a}\right)$.

Corrigé de l'exercice 4 On applique la formule des probabilités conditionnelles :

$$\begin{aligned}\mathbb{P}(\mathrm{X}=k|\mathrm{Z}=\ell) &= \frac{\mathbb{P}\big((\mathrm{X}=k) \text{ et } (\mathrm{Z}=\ell)\big)}{\mathbb{P}(\mathrm{Z}=\ell)}\\ &= \frac{\mathbb{P}\big((\mathrm{X}=k) \text{ et } (\mathrm{Y}=\ell-k)\big)}{\mathbb{P}(\mathrm{Z}=\ell)}\\ &= \frac{\mathbb{P}(\mathrm{X}=k)\,\mathbb{P}(\mathrm{Y}=\ell-k)}{\mathbb{P}(\mathrm{Z}=\ell)} \quad \text{(X et Y indépendantes)}\\ &= \frac{\frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{\ell-k}}{(\ell-k)!}e^{-\lambda_2}}{\frac{(\lambda_1+\lambda_2)^\ell}{\ell!}e^{-(\lambda_1+\lambda_2)}}\\ &= \frac{\ell!}{k!(\ell-k)!}\frac{\lambda_1^k\lambda_2^{\ell-k}}{(\lambda_1+\lambda_2)^\ell}\\ &= \binom{\ell}{k}\frac{\lambda_1^k}{(\lambda_1+\lambda_2)^k}\frac{\lambda_2^{\ell-k}}{(\lambda_1+\lambda_2)^{\ell-k}}\\ &= \binom{\ell}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{\ell-k}\\ &= \binom{\ell}{k}p^k(1-p)^{\ell-k}.\end{aligned}$$

On reconnaît la loi binomiale $\mathscr{B}in(\ell,p)$. Dans le processus de Poisson $\mathscr{S}$ obtenu par superposition des processus $\mathscr{S}_1$ et §$_2$ de taux λ_1 et λ_2, qui est donc un processus de Poisson de taux $\lambda = \lambda_1 + \lambda_2$, la proportion d'événements issus du processus $\mathscr{S}_1$ est $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$.

Autrement dit, si on considère ℓ événements du processus $\mathscr{S}$, la probabilité pour chacun d'entre eux d'être issu du processus $\mathscr{S}_1$ est p. Les événements étant indépendants, le nombre d'événements de $\mathscr{S}_1$ soit bien suivre une binomiale $\mathscr{B}in(\ell,p)$.

Dans le cas du rétinoblastome dans le modèle de Knudson, l'apparition des tumeurs dans chacun des deux yeux est un processus de Poisson de même taux λ, et le nombre total de tumeurs suit un processus de Poisson de taux 2λ. Pour un total fixé de trois tumeurs, le nombre X de tumeurs à un œil donné (par exemple le gauche) suit une loi $\mathscr{B}in\left(3,\frac{1}{2}\right)$, et donc

$$\mathbb{P}(\mathrm{X}=0 \text{ ou } \mathrm{X}=3) = \mathbb{P}(\mathrm{X}=0)+\mathbb{P}(\mathrm{X}=3) = \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}.$$

Corrigé de l'exercice 5

1. Soient X_g et X_d les nombres de tumeurs aux yeux gauche et droit. Dans le modèle de Knudson l'apparition des tumeurs suit un processus de Poisson, indépendament à chaque œil : on a $\mathrm{X}_g, \mathrm{X}_d \sim$

$\mathscr{P}(\lambda)$. Le nombre total Xde tumeurs est $X = X_g + X_d$, qui suit une loi de Poisson de paramètre $\lambda_K = \lambda + \lambda$, d'où le résultat.

2. Pour fixer les idées, considérons les tumeurs à l'oeil gauche (le choix de l'œil étant indifférent). Remarquons la relation $\mathbb{P}(X = k) = \frac{\lambda}{k}\mathbb{P}(X = k-1)$ qui permet de calculer les valeurs de $\mathbb{P}(X = k)$ de proche en proche.

Pour $\lambda_K = 2$ (resp. 3) on a $\lambda = 1$ (resp. 1,5). On trouve les valeurs suivantes.

	$\lambda = 1$	$\lambda = 1{,}5$
$\mathbb{P}(X = 0)$	36,8%	22,3%
$\mathbb{P}(X = 1)$	36,8%	33,5%
$\mathbb{P}(X = 2)$	18,4%	25,1%
$\mathbb{P}(X = 3)$	6,1%	12,6%
$\mathbb{P}(X = 4)$	1,5%	4,7%
$\mathbb{P}(X \geqslant 5)$	0,4%	1,8%

On calcule ensuite les proportions des classes $(X = 1)$ à $(X \geqslant 5)$ parmi les $X > 0$, ce qui correspond à la formule $\mathbb{P}(X = k|X > 0) = \dfrac{\mathbb{P}(X = k)}{\mathbb{P}(X > 0)} = \dfrac{\mathbb{P}(X = k)}{1 - \mathbb{P}(X = 0)}$.

	$\lambda = 1$	$\lambda = 1{,}5$
$\mathbb{P}(X = 1\|X > 0)$	58,2%	43,1%
$\mathbb{P}(X = 2\|X > 0)$	29,1%	32,3%
$\mathbb{P}(X = 3\|X > 0)$	9,7%	16,2%
$\mathbb{P}(X = 4\|X > 0)$	2,4%	6,0%
$\mathbb{P}(X \geqslant 5\|X > 0)$	0,6%	2,3%

3. La comparaison laisse penser que la « vraie » valeur de λ se situe entre 1 et 1,5, et donc que la vraie valeur de λ_K est entre 2 et 3.

4. On calcule $\overline{x} = \frac{117}{66} = 1{,}77$. On a donc $\lambda \approx 1{,}77 - \exp(-0{,}9 \times 0{,}77) = 1{,}27$.

On réalise un χ^2 de conformité :

	$(X = 1)$	$(X = 2)$	$(X = 3)$	$(X = 4)$	$(X \geqslant 5)$
Observés	35	17	9	4	1
Attendus	32.7	20.8	8.8	2.8	0.7

Pour être dans les conditions de Cochran, il faut fusionner les dernières catégories :

	$(X = 1)$	$(X = 2)$	$(X \geqslant 3)$
Observés	35	17	14
Attendus	32.7	20.8	12.3

On calcule un χ^2 de 1.09, à un degré de liberté : on ne rejette pas le modèle.

Corrigé de l'exercice 6 1. Si le Chevalier gagne la manche suivante (ce qui se produit avec probabilité $\frac{1}{2}$) il a gagné la partie. S'il la perd, les deux joueurs sont à deux manches partout ; à la manche suivante, chacun des deux a une chance sur deux de gagner la partie. Le Chevalier a donc une probabilité $\frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}$ de gagner. Son espérance de gain est donc de $\frac{3}{4}64 = 48$ pistoles.

2. Soit $r = N - k$. C'est le nombre de manches que le premier joueur doit encore gagner pour remporter la partie. Supposons que le jeu se poursuive (potentiellement à l'infini), même après que la partie soit finie : c'est un processus de Bernoulli.

La loi de X : le rang du r-ième succès du premier joueur est une loi binomiale négative de paramètres r et $p = \frac{1}{2}$, c'est-à-dire

$$P(X = i) = \binom{i-1}{r-1}\left(\frac{1}{2}\right)^i \quad (i \geqslant r).$$

Notez qu'au moment du r-ième succès du premier joueur, le second joueur a gagné $X - r$ manches. Le premier joueur a gagné la partie s'il a gagné r nouvelles manches avant que le second joueur en ait gagné $N - \ell$, c'est-à-dire si $X - r < N - \ell$, ou encore $X < r + N - \ell$.

La probabilité que le premier joueur gagne est donc

$$\sum_{i=r}^{r+\mathrm{N}-\ell-1}\binom{i-1}{r-1}\left(\frac{1}{2}\right)^i=\sum_{i=\mathrm{N}-k}^{2\mathrm{N}-k-\ell-1}\binom{i-1}{\mathrm{N}-k-1}\left(\frac{1}{2}\right)^i.$$

Le calcul pratique de cette quantité demande une grande patience ou un petit ordinateur.

Corrigé de l'exercice 7 1. X suit une loi de Poisson $\mathscr{P}(\lambda=1)$.

2. On sait que $\mathbb{P}(\mathrm{X}=k)=\frac{\lambda}{k!}e^{-\lambda}$, ce qui permet de calculer $\mathbb{P}(\mathrm{X}=0)=0{,}367$, $\mathbb{P}(\mathrm{X}=1)=0{,}367$ et $\mathbb{P}(\mathrm{X}=2)=0{,}184$.

On a

$$\begin{aligned}\mathbb{P}(\mathrm{X}\geq 3) &= 1-\sum_{i=0,1,2}\mathbb{P}(\mathrm{X}=i)\\ &= 0{,}082.\end{aligned}$$

3. $\mathrm{Y}\sim\mathscr{P}\left(\lambda=\frac{1}{7}\right)$. On calcule $\mathbb{P}(\mathrm{Y}=0)=0{,}867$ et $\mathbb{P}(\mathrm{Y}=1)=0{,}124$.

4. Si on choisit comme unité de temps la semaine, on a $\mathrm{T}\sim\exp(\lambda=1)$. On pourrait choisir la journée, auquel cas la loi serait $\exp\left(\lambda=\frac{1}{7}\right)$.

5. La fonction de répartition de la loi exponentielle est $\mathrm{F}(t)=1-e^{-\lambda t}$.

On a $\mathbb{P}(\mathrm{T}>1)=e^{-1}=0{,}367$ et $\mathbb{P}(\mathrm{T}>2)=e^{-2}=0{,}135$.

Corrigé de l'exercice 8

1. La moyenne vaut $\frac{1}{64}2455=38{,}4$. On lit facilement la médiane en 32$^{\mathrm{e}}$ position, $q_{0,5}=21$. De même les premiers et troisièmes quartiles valent $q_{0,25}=9$ et $q_{0,75}=40$.

2. L'espérance de la loi exponentielle est $1/\lambda$ donc on prend $\widehat{\lambda}=1/38{,}4=0{,}026$.

3. La fonction de répartition de la loi exponentielle est $\mathrm{F}(t)=1-e^{-\lambda t}$, donc la fonction quantile est $q_\alpha=-\frac{1}{\lambda}\log(1-\alpha)$. On estime donc $q_{0,5}=\frac{1}{\lambda}\log 2=26{,}6$, $q_{0,25}=11$, et $q_{0,75}=53$.

4. C'est $\mathbb{P}(\mathrm{T}<12)=1-e^{-\lambda\cdot 12}=0{,}27$, et $\mathbb{P}(\mathrm{T}<60)=0{,}79$.

5. Chaque d_i a variance $\frac{1}{\lambda^2}$; la variance de $\overline{d}$ vaut $\frac{1}{64^2}\times 64\frac{1}{\lambda^2}=\frac{1}{64\lambda^2}$. De plus d'après le théorème central limite $\overline{d}$ est approximativement normale. On estime son écart-type à $\frac{1}{8\lambda}=\frac{1}{8\times 0{,}026}=4{,}8$. On en déduit l'intervalle de confiance pour $\frac{1}{\lambda}$

$$[38{,}4-1{,}96\times 4{,}8;34{,}8+1{,}96\times 4{,}8]=[29{,}0;47{,}8].$$

On a ensuite l'intervalle de confiance suivant pour λ : $[0{,}021;0{,}034]$, et pour $q_{0,5}=\frac{1}{\lambda}\log 2=[20{,}1;33{,}13]$.

A.6 Exercices du chapitre 6

Corrigé de l'exercice 1

1. On sait que si $\mathrm{X}\sim\mathscr{N}(\mu,\sigma^2)$ alors $\mathrm{Y}=a\mathrm{X}+b\sim\mathscr{N}(a\mu+b,a^2\sigma^2)$. On a donc $\mathrm{Y}_1\sim\mathscr{N}(0{,}36)$, $\mathrm{Y}_2\sim\mathscr{N}\left(0,\frac{1}{36}\right)$, $\mathrm{Y}_3\sim\mathscr{N}(0{,}1)$. On a $\mathrm{Y}_4=\mathrm{Y}_3^2$ donc $\mathrm{Y}_4\sim\chi^2(1)$.

2. On a par exemple $\mathbb{P}(\mathrm{Y}_1>0|\mathrm{X}>5)=1$, donc X et Y_1 ne sont pas indépendantes.

3. On a $\mathrm{X}+\mathrm{Y}_1=2\mathrm{X}-5\sim\mathscr{N}(5{,}144)$.

Corrigé de l'exercice 2 1. On a $\mathrm{cov}(\mathrm{X},a\mathrm{X}+b\mathrm{Y})=a\mathrm{cov}(\mathrm{X},\mathrm{X})+b\mathrm{cov}(\mathrm{X},\mathrm{Y})=a$ car $\mathrm{cov}(\mathrm{X},\mathrm{X})=\mathrm{var}(\mathrm{X})=1$ et $\mathrm{cov}(\mathrm{X},\mathrm{Y})=0$, les variables X et Y étant indépendantes.

2. On a $\mathrm{var}(\mathrm{Z})=a^2\,\mathrm{var}(\mathrm{X})+2ab\mathrm{cov}(\mathrm{X},\mathrm{Y})+b^2\,\mathrm{var}(\mathrm{Y})=a^2+b^2$.

3. Il faut $a^2+b^2=1$, soit $0{,}64+b^2=1$, d'où $b=\pm 0{,}6$. On a alors $\mathrm{corr}(\mathrm{X},\mathrm{Z})=\mathrm{cov}(\mathrm{X},\mathrm{Z})/\sigma_{\mathrm{X}}\sigma_{\mathrm{Z}}=0{,}8$.

Corrigé de l'exercice 3 1. On a $\text{cov}(X,Y) = r\sigma_X\sigma_Y = 0{,}7 \times \sqrt{2 \times 2{,}3} = 1{,}5$.

2. On a $E(Z) = E(Y) - E(X) = 1 - 1{,}8 = -0{,}8$, et

$$\text{var}(Z) = \text{var}(X) - 2\text{cov}(X,Y) + \text{var}(Y) = 2 - 2 \times 1{,}50 + 2{,}3 = 1{,}3.$$

3. Comme (X,Y) est suit une loi normale bivariée, la loi de $Z = Y - X$ est gaussienne : $Z \sim \mathcal{N}(\mu_Z = -0{,}8, \sigma_Z^2 = 1{,}3)$, et $Z' = \frac{Z-\mu_Z}{\sigma_Z} \sim \mathcal{N}(0,1)$, avec $\sigma_Z = \sqrt{1{,}3} = 1{,}14$. Alors

$$\begin{aligned}\mathbb{P}(Y \geqslant X) &= \mathbb{P}(Z \geqslant 0)\\ &= \mathbb{P}\left(\frac{Z-\mu_Z}{\sigma_Z} \geqslant \frac{0-\mu_Z}{\sigma_Z}\right)\\ &= \mathbb{P}(Z' \geqslant 0{,}7)\\ &= 0{,}24.\end{aligned}$$

Corrigé de l'exercice 4 (pas de corrigé)

Corrigé de l'exercice 5

1. On a $E(Z) = E(Y) + aE(X) = 3 + 2a$, et $\text{var}(Z) = \text{var}(Y) + 2a\text{cov}(X,Y) + a^2\,\text{var}(X) = 2 + 2a + a^2$.

2. On a $\text{cov}(Z,X) = \text{cov}(Y,X) + a\,\text{var}(X) = 1 + a$, $\text{cov}(Z,Y) = \text{var}(Y) + a\text{cov}(X,Y) = 2 + a$, et $\text{cov}(Z, Y-X) = \text{cov}(Z,Y) - \text{cov}(Z,X) = 1$.

3. Pour $a = -1$, c'est-à-dire $Z = Y - X$, on a $\text{cov}(X,Z) = 0$. Comme (X,Y) est gaussien, on en déduit que X et Z sont toutes deux gaussiennes et indépendantes.

Corrigé de l'exercice 6

1. C'est $\mu \pm 1{,}96\sigma = [3{,}216; 4{,}784]$.

2. Pour $n = 4$, l'écart-type de $\overline{X}_n$ est $0{,}4/\sqrt{4} = 0{,}2$ donc l'intervalle de fluctuation est $4 \pm 1{,}96 \times 0{,}2 = [3{,}608; 4{,}392]$.

3. L'écart-type de $\overline{X}_n$ est $0{,}4/\sqrt{(}n)$; la largeur de l'intervalle de fluctuation est donc $2 \times 1{,}96 \times 0{,}4/\sqrt{n} = 1{,}568/\sqrt{n}$.

On veut que cette largeur soit $< 0{,}4$ soit $\sqrt{n} > 1{,}568/0{,}4$; on trouve $n > 15{,}4$ soit, n étant entier, $n \geqslant 16$.

Corrigé de l'exercice 7

1. On se ramène à une loi $\mathcal{N}(0,1)$ en posant $Z = \frac{1}{0{,}8}(B_1 - 4)$. On a $\mathbb{P}(B_1 < 5) = \mathbb{P}(Z < 1{,}25) = 0{,}89$.

2. On procède de la même façon : cette fois $B_1 \sim \mathcal{N}(\mu = 6{,}5, \sigma^2 = 0{,}64)$ donc $Z = \frac{1}{0{,}8}(B - 6{,}5)$ suit une loi $\mathcal{N}(0,1)$; on veut la probabilité que $Z > \frac{1}{0{,}8}(5 - 6{,}5) = -1{,}875$. Avec la table de loi normale on trouve $P(Z > -1{,}875) = 0{,}97$.

3. On a $\mathbb{P}(Z > -2{,}33) = 0{,}99$. Pour avoir une sensibilité de 99% il faut prendre $s = 6{,}5 - 2{,}33 \times 0{,}8 = 4{,}64$. On a alors une spécifité $P(B_1 < 4{,}64)$ avec $B_1 \sim \mathcal{N}(\mu = 4, \sigma^2 = 0{,}64)$, donc égale à $P(Z < 0{,}8) = 0{,}79$.

4. On a $\text{cov}(B_1, B_2) = \text{cor}(B_1, B_2)\sqrt{\text{var}(B_1)\,\text{var}(B_2)} = -0{,}2$.

5. B suit une loi normale d'espérance $4 + 1{,}6 \times 2 = 7{,}2$ et de variance $\text{var}(B_1) + 2 \times 1{,}6 \times \text{cov}(B_1, B_2) + 1{,}6^2\,\text{var}(B_2) = 0{,}4096$.

6. On a les mêmes variances et corrélation qu'auparavant, donc on a toujours $\text{cov}(B_1, B_2) = -0{,}2$ et $\text{var}(B) = 0{,}4096$; on a maintenant $E(B) = 6{,}5 + 1{,}6 \times 2{,}45 = 10{,}42$.

7. Dans la population saine, $\mathbb{P}(B > 9) = P\left(Z > \frac{9-7.2}{\sqrt{0{,}4096}}\right) = \mathbb{P}(Z > 2{,}81) = 0{,}0025$ (ce qui correspond à une spécificité de 99,75% et dans la population infectée $\mathbb{P}(B > 9) = P\left(Z > \frac{9-10.42}{\sqrt{0{,}4096}}\right) = \mathbb{P}(Z > -2.22) \simeq 0{,}99$ (sensibilité 99%).

Corrigé de l'exercice 8

1. On se ramène à une loi normale centrée réduite en posant $Z' = \frac{Z-175}{7} \sim \mathcal{N}(0,1)$, si $Z \sim \mathcal{N}(175, 7^2)$

est la taille d'un homme pris au hasard :

$$\begin{aligned}\mathbb{P}(170 < Z < 180|\text{homme}) &= \mathbb{P}\left(\frac{170-175}{7} \leqslant Z' \leqslant \frac{180-175}{7}\right)\\ &\simeq \mathbb{P}\left(-0{,}71 < Z' < 0{,}71\right)\\ &\simeq 1 - 2\mathbb{P}(Z' > 0{,}71)\\ &\simeq 1 - 2 \times 0{,}24\\ &\simeq 0{,}52\end{aligned}$$

De même pour une femme :

$$\begin{aligned}\mathbb{P}(170 < Z < 180|\text{femme}) &= \mathbb{P}\left(\frac{170-165}{7} \leqslant Z' \leqslant \frac{180-165}{7}\right)\\ &\simeq \mathbb{P}\left(0{,}71 < Z' < 2{,}14\right)\\ &\simeq 0{,}984 - 0{,}761\\ &\simeq 0{,}22\end{aligned}$$

2. Par la formule des probabilités totales, c'est

$$\mathbb{P}(170 < Z < 180) = 0{,}5 \times 0{,}52 + 0{,}5 \times 0{,}23 = 0{,}375$$

3. Symmétrie de la loi normale aidant, on imagine volontiers que la taille médiane est à mi-chemin entre l'espérance de celle des hommes et et de celle des femmes : donc, 170.
On peut calculer :

$$\begin{aligned}\mathbb{P}(Z < 170) &= 0{,}5 \times \mathbb{P}(Z < 170|\text{homme}) + 0{,}5 \times \mathbb{P}(Z < 170|\text{femme})\\ &= 0{,}5 \times \mathbb{P}(Z' < -0{,}71) + 0{,}5 \times \mathbb{P}(Z' < 0{,}71)\\ &= 0{,}5\end{aligned}$$

4. On a donc $\mathrm{E}(Z) = 0{,}5 \times 175 + 0{,}5 \times 165 = 170$, et

$$\begin{aligned}\mathrm{E}\left(Z^2\right) &= 0{,}5 \times \mathrm{E}\left(Z^2\middle|\text{homme}\right) + 0{,}5 \times \mathrm{E}\left(Z^2\middle|\text{femme}\right)\\ &= 0{,}5 \times \left(7^2 + 175^2\right) + 0{,}5 \times \left(7^2 + 165^2\right)\\ &= 28674\end{aligned}$$

d'où une variance $\mathrm{var}(Z) = 28674 - 170^2 = 74$, correspondant à un écart-type de 8,6 cm.

5. La taille étant mesurée au centimètre près, il s'agit en fait d'une taille comprise entre 179,5 et 180,5 cm.
On calcule

$$\begin{aligned}\mathbb{P}(179{,}5 < Z < 180{,}5|\text{homme}) &= \mathbb{P}\left(\frac{179{,}5-175}{7} \leqslant Z' \leqslant \frac{180{,}5-175}{7}\right)\\ &\simeq \mathbb{P}\left(0{,}64 < Z' < 0{,}79\right)\\ &\simeq 0{,}7852 - 0{,}7389\\ &\simeq 0{,}046\end{aligned}$$

et

$$\begin{aligned}\mathbb{P}(179{,}5 < Z < 180{,}5|\text{femme}) &= \mathbb{P}\left(\frac{179{,}5-165}{7} \leqslant Z' \leqslant \frac{180{,}5-165}{7}\right)\\ &\simeq \mathbb{P}\left(2{,}07 < Z' < 2{,}21\right)\\ &\simeq 0{,}98645 - 0{,}98077\\ &\simeq 0{,}0057\end{aligned}$$

Et donc $\mathbb{P}(Z = 180) = 0{,}5 \times 0{,}046 + 0{,}5 \times 0{,}0057 \simeq 0{,}026$. On conclut par la formule de Bayes :

$$\begin{aligned}\mathbb{P}(\text{homme}|Z = 180) &= \frac{\mathbb{P}(Z = 180|\text{homme}) \times \mathbb{P}(\text{homme})}{\mathbb{P}(Z = 180)} \\ &\simeq \frac{0{,}046 \times 0{,}5}{0{,}5 \times 0{,}046 + 0{,}5 \times 0{,}0057} \\ &\simeq 0{,}89\end{aligned}$$

On peut aussi le faire avec la formule des odds ratio :

$$\begin{aligned}\frac{\mathbb{P}(\text{homme}|Z = 180)}{\mathbb{P}(\text{femme}|Z = 180)} &= \frac{\mathbb{P}(\text{homme})}{\mathbb{P}(\text{femme})} \times \frac{\mathbb{P}(Z = 180|\text{homme})}{\mathbb{P}(Z = 180|\text{femme})} \\ &\simeq \frac{0{,}046}{0{,}0057} \\ &\simeq 8\end{aligned}$$

et donc $\mathbb{P}(\text{homme}|Z = 180) = 8/9 = 0{,}89$.

Remarque : Au lieu de faire le calcul avec $\mathbb{P}(179{,}5 < Z < 180{,}5)$, on pouvait directement utiliser la densité de la loi normale évaluée en $z = 180$ (on obtient grosso modo le même résultat). Pourquoi ?

Corrigé de l'exercice 9 1. Il suffit de prendre $X_1, \ldots, X_n$ de loi $\mathscr{P}(\frac{1}{n}\lambda)$. Il est facile de vérifier, en prenant simplement $\lambda = 1$, que l'approximation par une loi normale est de mauvaise qualité.

2. Quand α devient petit, α^2 et α^3 deviennent beaucoup plus petit que α. Il suffit par exemple de prendre α assez petit pour que $6\alpha^2 < 0{,}5\,\alpha$ et $8\alpha^3 < 0{,}5\,\alpha$.

3. On pose $\alpha = \frac{1}{n}\lambda$, avec n assez grand pour que les conditions de la question 2 soient réalisées ; et on écrit $X = X_1 + \cdots + X_n$ avec les X_i indépendantes de loi $\mathscr{P}(\alpha)$.

Le théorème de Berry-Esséen implique que l'erreur commise par l'approximation normale de X est inférieure à

$$0{,}48\frac{\rho}{\alpha^{3/2}\sqrt{n}} < 0{,}48\frac{2\alpha}{\alpha^{3/2}\sqrt{n}} < \frac{1}{\sqrt{\alpha n}} = \frac{1}{\sqrt{\lambda}}.$$

On en déduit le résultat voulu.

A.7 Exercices du chapitre 7

Corrigé de l'exercice 1 On commence par la fonction de répartition de P. Notons que P prend ses valeurs dans [0,1] donc sa densité en nulle en dehors de cet intervalle. Pour $t \in [0,1]$

$$\begin{aligned}\mathbb{P}(P \leqslant t) = \mathbb{P}(1 - F(X) \leqslant t) = \mathbb{P}(F(X) \geqslant 1 - t) = \mathbb{P}(X \geqslant F^{-1}(1 - t)) \\ = 1 - \mathbb{P}(X < F^{-1}(1 - t)) = 1 - F(F^{-1}(1 - t)) = 1 - (1 - t) = t,\end{aligned}$$

et on en déduit que la densité de P est

$$g(t) = \begin{cases} 1 & \text{si } t \in [0,1] \\ 0 & \text{sinon.} \end{cases}$$

Corrigé de l'exercice 2

On a facilement $E(X) = \mu = 1$ et $\text{var}(X) = \sigma^2 = \frac{1}{3}\alpha^2$.

La densité de X est

$$f(x) = \begin{cases} \frac{1}{2\alpha} & \text{si } 1 - \alpha \leqslant x \leqslant 1 + \alpha \\ 0 & \text{sinon} \end{cases}$$

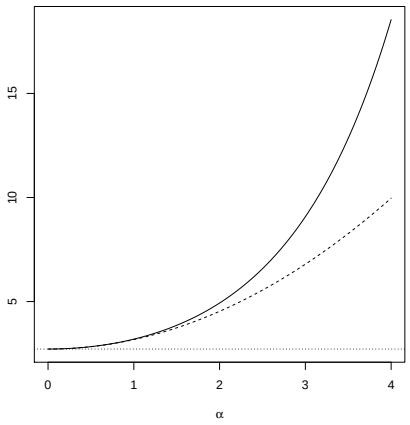
On a donc

$$\begin{aligned}\mathrm{E}\left(e^{\mathrm{X}}\right) &= \int_{1-\alpha}^{1+\alpha} e^x \times \frac{1}{2\alpha} dx \\ &= \frac{1}{2\alpha}\left[e^x\right]_{1-\alpha}^{1+\alpha} \\ &= \frac{1}{2\alpha}\left(e^{1+\alpha} - e^{1-\alpha}\right)\end{aligned}$$

L'approximation d'ordre 1 est $\phi(\mu) = e = 2{,}71$.
Pour calcule l'approximation d'ordre 2 avec $\phi(x) = e^x$ on a $\phi''(x) = e^x$ et donc $\phi''(\mu) = e$, d'où

$$\mathrm{E}\left(e^{\mathrm{X}}\right) \simeq \phi(\mu) + \frac{1}{2}\phi''(\mu)\sigma^2 = e + \frac{1}{3}e^2\alpha^2.$$

Traçons un graphe pour α entre 0 et 5 :



Le trait plein représente la valeur exacte de $\mathrm{E}\left(e^{\mathrm{X}}\right)$, le pointillé l'approximation par la méthode Delta d'ordre 2 – et le pointillé fin l'approximation d'ordre 1, qui ne dépend pas de α. On constate que plus α est petit (c'est-à-dire plus var(X) est petit), meilleure est l'approximation, et que l'utilisation de l'approximation d'ordre 2 améliore grandement la qualité de l'approximation.

Corrigé de l'exercice 3

Si X est la concentration mesurée, on s'intéresse à $\mathrm{Y} = \log \mathrm{X}$.
La dérivée de $\phi(x) = \log(x)$ est $\phi'(x) = \frac{1}{x}$, la dérivée seconde est $\phi''(x) = -\frac{1}{x^2}$. Notons $\mu_{\mathrm{X}} = \mathrm{E(X)}$ et $\sigma^2_{\mathrm{X}} = \mathrm{var(X)}$. Par la methode du Delta, on a

$$\mathrm{E(Y)} \simeq \log(\mu_{\mathrm{X}}) - \frac{\sigma^2_{\mathrm{X}}}{2\mu^2_{\mathrm{X}}},$$

En utilisant les écart-types estimés s_{X} pour les concentrations, on en tire une table d'approximations pour les moyennes des logarithmes des concentrations.

t	0	1	2	3	4
moyenne des log	3,00	2,66	2,31	1,90	1,37

Table 25. Moyennes des logarithmes des concentrations

On remarque que le terme $\frac{\sigma^2_{\mathrm{X}}}{2\mu^2_{\mathrm{X}}}$ est négligeable pour les premières valeurs, mais devient sensible pour les dernières.

Corrigé de l'exercice 4

1. Il suffit de poser $\varepsilon = \frac{Z-\mu}{\sigma}$. La loi de ε est bien la loi normale centrée réduite.

2. On a

$$\phi(Z) = \phi(\mu + \sigma\varepsilon) = \phi(\mu) + \sigma\varepsilon\phi'(\mu + c),$$

avec c entre 0 et $\sigma\varepsilon$; et donc

$$\frac{\phi(Z) - \phi(\mu)}{\phi'(\mu)\sigma} = \varepsilon \times \left(\frac{\phi'(\mu + c)}{\phi'(\mu)}\right).$$

3. On écrit $Z_n = \mu + \sigma_n\varepsilon_n$. On a, d'après la question précédente,

$$Y_n = \frac{\phi(Z_n) - \phi(\mu)}{\phi'(\mu)\sigma_n} = \varepsilon_n \times \left(\frac{\phi'(\mu + c_n)}{\phi'(\mu)}\right)$$

avec c_n entre 0 et $\sigma_n\varepsilon_n$. Quand n est grand, c_n s'approche de 0 et $\phi'(\mu + c_{Z_n})$ s'approche de $\phi'(\mu)$. On en conclut que la loi de Y_n s'approche de la loi $\mathcal{N}(0,1)$.

Cette preuve manque de rigueur mathématique mais les grandes idées sont là.

Corrigé de l'exercice 5

1. On obtient les résultats suivants.

Patient	σ	μ	cv
1	15,80	738,1	0,021
2	10,93	256,7	0,043

2. On a, pour $Y = \log(X)$

$$\sigma_Y^2 = \mathrm{var}(Y) \simeq \left(\frac{1}{E(X)}\right)^2 \times \mathrm{var}(X)$$

et donc, en prenant la racine carrée, l'écart-type de Y est approximativement $\frac{\sigma_X}{\mu_X} = \mathrm{cv}(X)$.

3. On obtient

Patient	σ_Y
1	0,021
2	0,042

Sur ces deux séries de mesures, l'approximation est très satisfaisante.

4. On commence par donner un intervalle de confiance sur σ_Y^2 avec la procédure standard pour la variance d'une loi normale, puis on en déduit un intervalle de confiance sur $\sigma_Y \simeq \mathrm{cv}(X)$.

L'intervalle de confiance pour la variance est

$$\sigma^2 \in \left[\frac{n-1}{x^{n-1}_{0,975}} \times S^2 ; \frac{n-1}{x^{n-1}_{0,025}} \times S^2\right],$$

d'où l'IC pour l'écart-type

$$\sigma \in \left[\sqrt{\frac{n-1}{x^{n-1}_{0,975}}} \times S ; \sqrt{\frac{n-1}{x^{n-1}_{0,025}}} \times S\right].$$

Ici, $n = 3$, $x^2_{0,975} = 7{,}38$ et $x^2_{0,025} = 0{,}056$, d'où les IC suivants :

Patient	σ_Y	IC
1	0,021	[0,011 ; 0,134]
2	0,042	[0,022 ; 0,265]

Remarque sur le choix de l'estimateur de la variance. On pouvait pour les questions 1 et 3, préférer utiliser un estimateur de la variance en $\frac{1}{n}$ et non $\frac{1}{n-1}$. L'estimateur de la variance en $\frac{1}{n-1}$ correspond à l'idée que les trois mesures sont représentatives d'une série potentiellement infinie de mesures, dont on veut estimer le coefficient de variation ; on peut utiliser l'estimateur en $\frac{1}{n}$ simplement pour quantifier la variation entre les trois mesures réalisées.

On obtient avec ce choix les résultats suivants (l'approximation reste de bonne qualité).

Patient	σ	μ	cv	σ_Y
1	12,90	738,1	0,017	0,017
2	8,92	256,7	0,035	0,034

Pour la question 4, on n'a plus le choix : fournir un intervalle de confiance n'a de sens que si on considère qu'on estime le coefficient de variation d'une série d'expériences.

Corrigé de l'exercice 6

1. X suit une loi binomiale $\mathscr{B}in(n,p)$.

2. On a $\mathrm{E}\left(\widehat{p}\right) = \frac{1}{n}\mathrm{E}(\mathrm{X}) = p$ (pas de biais), et $\mathrm{var}\left(\widehat{p}\right) = \frac{1}{n^2}\mathrm{var}(\mathrm{X}) = \frac{p(1-p)}{n}$.

3. On a $\phi(p) = \log p - \log(1-p)$ d'où $\phi'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$. La variance de $\widehat{\phi} = \phi\left(\widehat{p}\right)$ est donc approximativement

$$\phi'(p)^2 \mathrm{var}\left(\widehat{p}\right) = \left(\frac{1}{p(1-p)}\right)^2 \times \frac{p(1-p)}{n} = \frac{1}{np(1-p)}.$$

4. On a

$$\frac{1}{np(1-p)} = \frac{1}{np} + \frac{1}{n-np}$$

et $\mathrm{E}(\mathrm{X}) = np$. On en déduit que $\frac{1}{\mathrm{X}} + \frac{1}{n-\mathrm{X}}$ est un estimateur « raisonnable » de la variance de $\widehat{\phi}$.

Corrigé de l'exercice 7 1. Appliquer le théorème de Bayes.

2. On a $\mathrm{E}(\mathrm{A}) = n\mathbb{P}(\mathrm{exp}|\mathrm{att})$, $\mathrm{E}(\mathrm{B}) = n\mathbb{P}(\mathrm{non\ exp}|\mathrm{att})$, $\mathrm{E}(\mathrm{C}) = m\mathbb{P}(\mathrm{exp}|\mathrm{non\ att})$ et $\mathrm{E}(\mathrm{D}) = m\mathbb{P}(\mathrm{non\ exp}|\mathrm{non\ att})$. On peut donc proposer d'estimer l'OR par

$$\widehat{\mathrm{OR}} = \frac{\mathrm{A/B}}{\mathrm{C/D}} = \frac{\mathrm{AD}}{\mathrm{CB}}.$$

On pourrait utiliser une variante de la méthode du Delta d'ordre 1 pour montrer que si n et m sont grands, l'espérance de cette approximation est proche de l'OR.

3. On a

$$\log\widehat{\mathrm{OR}} = \log\left(\frac{\mathrm{A}}{\mathrm{B}}\right) + \log\left(\frac{\mathrm{C}}{\mathrm{D}}\right).$$

D'autre part $\frac{\mathrm{A}}{\mathrm{B}}$ et $\frac{\mathrm{C}}{\mathrm{D}}$ sont indépendants, d'où le résultat demandé. D'après l'exercice précédent

$$\mathrm{var}\left(\log\left(\frac{\mathrm{A}}{\mathrm{B}}\right)\right) = \frac{1}{\mathrm{A}} + \frac{1}{\mathrm{B}} \text{ et } \mathrm{var}\left(\log\left(\frac{\mathrm{C}}{\mathrm{D}}\right)\right) = \frac{1}{\mathrm{C}} + \frac{1}{\mathrm{D}},$$

et on conclut.

4. On calcule $\widehat{\mathrm{OR}} = 1{,}83$, et donc $\log\widehat{\mathrm{OR}} = 0{,}61$, avec une variance approximative 0,082. On a l'intervalle de confiance pour $\log\mathrm{OR}$

$$0{,}61 \pm 1{,}96 \times \sqrt{0{,}082} = [0{,}048\,;\,1{,}17].$$

On passe à l'exponentielle pour un intervalle de confiance sur l'OR : $[1{,}05, 3{,}22]$.

5. On travaille sur $\log\mathrm{OR}$, toujours avec l'approximation normale : sous H_0, OR est nul, et la variance de $\log\widehat{\mathrm{OR}}$ est approximativement $4 \times \frac{1}{50} = 0{,}08$. On calcule donc un z-score

$$z = \frac{0{,}61}{\sqrt{0{,}08}} = 2{,}16.$$

La p-valeur est $\mathbb{P}(|Z| > 2{,}16) = 2 \times \mathbb{P}(Z > 2{,}16) = 2 \times (1 - 0{,}9846) = 0{,}031$.

6. Le χ^2 est égal à 4,51 avec 1 ddl. Pour calculer la p-valeur on utilise le fait qu'un $\chi^2(1)$ est le carré d'une loi normale standard : la p-valeur est donc

$$\mathbb{P}\left(|Z| > \sqrt{4.51}\right) = 2\mathbb{P}(Z > 2{,}12) = 2 \times (1 - 0{,}983) = 0{,}034.$$

Sur cet exemple, les deux tests paraissent très proches.

A.8 Exercices du chapitre 8

Corrigé de l'exercice 1

1. D'après le paragraphe 8.5.1, la loi de $\arcsin\sqrt{\widehat{p}}$ est approximativement normale

$$\mathcal{N}\left(\mu = \arcsin\sqrt{p}, \sigma^2 = \frac{1}{4n}\right)$$

où $n = 900$ est la taille de l'échantillon, donc sa variance est approximativement $1/3600$ et son écart-type $\sigma = 1/60$.

2. Le quantile de niveau 0,975 de la loi normale est 1,96 et on a donc avec probabilité 0,95

$$\mu - 1{,}96\sigma \leqslant \arcsin\sqrt{\widehat{p}} \leqslant \mu + 1{,}96\sigma$$

soit, en remplaçant μ par $\arcsin\sqrt{p}$, le résultat demandé.

3. On en déduit par la méthode du pivot qu'avec probabilité égale à 0,95

$$\arcsin\sqrt{\widehat{p}} - 1{,}96\sigma \leqslant \arcsin\sqrt{p} \leqslant \arcsin\sqrt{\widehat{p}} + 1{,}96\sigma$$

soit, avec $\widehat{p} = 0{,}01$ et $\sigma = 1/60$,

$$0{,}0675 \leqslant \arcsin\sqrt{p} \leqslant 0{,}01328.$$

On en déduit, en prenant d'abord le sinus puis le carré des bornes, l'intervalle de confiance suivant pour p :

$$0{,}0045 \leqslant p \leqslant 0{,}018.$$

Corrigé de l'exercice 2

1. L'espérance de $\overline{X}$ est $E\left(\overline{X}\right) = E(X_1) = \lambda$, et sa variance est $\text{var}\left(\overline{X}\right) = \frac{1}{n}\text{var}(X_1) = \frac{\lambda}{n}$.

2. Une somme de variables indépendantes suivant des lois de Poisson $\mathcal{P}(\lambda_1), \ldots, \mathcal{P}(\lambda_n)$ suit une loi $\mathcal{P}(\lambda_1 + \cdots + \lambda_n)$, donc ici $S \sim \mathcal{P}(n\lambda)$.

3. On pose $U = 2\sqrt{S}$. La loi de U est approximativement une normale $\mathcal{N}\left(2\sqrt{n\lambda}, 1\right)$, d'où

$$\mathbb{P}\left(-1{,}96 \leqslant U - 2\sqrt{n\lambda} \leqslant 1{,}96\right) = 0{,}95,$$

d'où on tire facilement le résultat demandé :

$$\mathbb{P}\left(2\sqrt{S} - 1{,}96 \leqslant 2\sqrt{n\lambda} \leqslant 2\sqrt{S} + 1{,}96\right) \simeq 0{,}95.$$

4. On a donc, en divisant par $2\sqrt{n}$ les inégalités précédentes,

$$\mathbb{P}\left(\sqrt{\frac{S}{n}} - \frac{1{,}96}{2\sqrt{n}} \leqslant \sqrt{\lambda} \leqslant \sqrt{\frac{S}{n}} + \frac{1{,}96}{2\sqrt{n}}\right) \simeq 0{,}95,$$

puis, en remplaçant $\frac{S}{n}$ par $\overline{X}$ et mettant au carré,

$$\mathbb{P}\left(\left(\sqrt{\overline{X}} - \frac{1{,}96}{2\sqrt{n}}\right)^2 \leqslant \lambda \leqslant \left(\sqrt{\overline{X}} + \frac{1{,}96}{2\sqrt{n}}\right)^2\right) \simeq 0{,}95.$$

(Remarque : cet intervalle de confiance est un intervalle approché ; en particulier, il ne sera valable que si $\sqrt{\overline{X}} - \frac{1,96}{2\sqrt{n}} > 0$.)

5. On a donc $\overline{x} = \frac{s}{n} = 30,5$, et l'intervalle de confiance se calcule par

$$\left[\left(\sqrt{30,5} - \frac{1,96}{2\sqrt{10}}\right)^2 ; \left(\sqrt{30,5} + \frac{1,96}{2\sqrt{10}}\right)^2\right] ;$$

on obtient [27,17;34,02].

A.9 Exercices du chapitre 9

Corrigé de l'exercice 1

1. C'est la formule des probabilités totales :

$$\begin{aligned}\mathbb{P}(\text{vote A}) &= \mathbb{P}(\text{vote A}|\text{ouvrier})\mathbb{P}(\text{ouvrier}) + \mathbb{P}(\text{vote A}|\text{paysan})\mathbb{P}(\text{paysan}) + \mathbb{P}(\text{vote A}|\text{artisan})\mathbb{P}(\text{artisan}) \\ &= 0,4 \times 0,2 + 0,4 \times 0,7 + 0,2 \times 0,6 \\ &= 0,48,\end{aligned}$$

soit $p_A = 48\%$ des voix.

2. (a) La loi de X est une loi binomiale $\mathscr{B}in(n = 1000, p = p_A = 0,48)$ (notons que comme le sondage est a priori sans remise, c'est-à-dire qu'on n'interroge pas deux fois le même individu, on utilise ici implicitement l'hypothèse que la Syldavie a une grande population). Son espérance est $np = 480$ et sa variance $np(1-p) = 249,6$.

On peut l'approcher par une loi normale $\mathscr{N}(\mu = 480, \sigma^2 = 249,6)$.

(b) L'estimation de p_A se fait par $\widehat{p_A} = \frac{1}{n}X$ (avec $n = 1000$). L'espérance de $\widehat{p_A}$ est égale à $p_A = 0,48$, et sa variance est $\frac{1}{n}p_A(1-p_A) = 0,0002496$. L'intervalle de pari est donc

$$0,48 \pm 1,96 \times \sqrt{0,0002496} = [44,9\% \ ; \ 51,1\%].$$

3. Notons X_1, (respectivement X_2 et X_3) le nombre d'ouvriers (respectivement de paysans et d'artisans) qui déclarent voter pour le candidat A. On a les lois suivantes : $X_1 \sim \mathscr{B}in(n = 400, p = 0,2)$, $X_2 \sim \mathscr{B}in(n = 400, p = 0,7)$, $X_3 \sim \mathscr{B}in(n = 200, p = 0,6)$. Leurs espérances respectives sont $\mu_1 = 80$, $\mu_2 = 280$ et $\mu_3 = 120$; leurs variances respectives sont $\sigma_1^2 = 64$, $\sigma_2^2 = 84$ et $\sigma_3^2 = 48$.

L'approximation normale reste valable pour X_1, X_2 et X_3. Le nombre totale de déclarations de votes pour le candidat est $X = X_1 + X_2 + X_3$, qui suit donc approximativement une loi normale $\mathscr{N}(\mu = \mu_1 + \mu_2 + \mu_3 = 480, \sigma_2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 = 196)$.

L'estimation de p_A obtenue par $\widehat{p_A} = \frac{1}{n}X$ est donc approximativement normale, d'espérance 0,48 et variance 0,000196, d'où un intervalle de pari légèrement plus étroit que le précédent :

$$0,48 \pm 1,96 \times \sqrt{0,000196} = [45,2\% \ ; \ 50,7\%].$$

Corrigé de l'exercice 2 1. On calcule $\widehat{\mu} = \overline{x}$ l'estimation ponctuelle de μ :

$$\widehat{\mu} = \overline{x} = \frac{1}{n}\sum x_i = \frac{1}{20}22,4 = 1,12.$$

Pour l'estimation $\widehat{\sigma^2} = s^2$ de σ^2, on calcule d'abord l'estimation non corrigée $\tilde{s}^2$:

$$\tilde{s}^2 = \frac{1}{n}\sum x_i^2 - (\overline{x})^2 = 2,79,$$

puis

$$s^2 = \frac{n}{n-1}\tilde{s}^2 = 2,94.$$

Pour calculer l'intervalle de confiance de niveau 90% pour μ, on lit dans la table le quantile $t^{19}_{0,95} = 1{,}729$; on a donc l'intervalle

$$1{,}12 \pm 1{,}729\sqrt{\frac{2{,}94}{20}} = [0{,}46; 1{,}78]\,.$$

Le quantile $t^{19}_{0,975}$ est égal à 2,093, et l'intervalle de confiance de niveau 95% est

$$1{,}12 \pm 2{,}093\sqrt{\frac{2{,}94}{20}} = [0{,}32; 1{,}92]\,.$$

Le calcul de l'intervalle de confiance de niveau 90% sur σ^2 nécessite les quantiles de la loi $\chi^2(19)$: $x^{19}_{0,05} = 10{,}12$ et $x^{19}_{0,95} = 30{,}14$. On en tire l'intervalle :

$$\left[19 \times \frac{2{,}94}{30{,}14}; 19 \times \frac{2{,}94}{10{,}12}\right] = [1{,}85; 5{,}52]\,.$$

Pour le niveau 95%, on a $x^{19}_{0,025} = 8{,}91$ et $x^{19}_{0,975} = 32{,}85$, d'où on tire de même l'intervalle de confiance [1,70; 6,27].

2. Si on suppose que $\sigma^2 = 4$ est connu a priori, on construit les intervalles de confiance à partir des quantiles de la loi normale : $z_{0,95} = 1{,}64$ et $z_{1,975} = 1{,}96$: on obtient

$$1{,}12 \pm 1{,}64\sqrt{\frac{4}{20}} = [0{,}39; 1{,}85]\,,$$

et

$$1{,}12 \pm 1{,}96\sqrt{\frac{4}{20}} = [0{,}24; 2{,}00]\,.$$

Corrigé de l'exercice 3 Pour le premier échantillon, on a déjà calculé à l'exercice précédent $\widehat{\mu}_1 = 1{,}12$ et $s^2_1 = 2{,}94$. On calcule pour le second $\widehat{\mu}_2 = 0{,}94$, et $s^2_2 = 3{,}68$.

Pour la variance, on a les quantiles $f^{29,19}_{0,95} = 2{,}077$ et $f^{29,19}_{0,05} = \frac{1}{f^{19,29}_{0,95}} \simeq \frac{1}{1{,}945} = 0{,}51$, d'où l'intervalle de confiance au niveau 90%

$$\left[\frac{2{,}94}{3{,}68} \times 0{,}51; \frac{2{,}94}{3{,}68} \times 2{,}077\right] = [0{,}41; 1{,}66].$$

De même, de $f^{29,19}_{0,975} = 2{,}402$ et $f^{29,19}_{0,025} = \frac{1}{f^{19,29}_{0,975}} \simeq \frac{1}{2{,}21} = 0{,}45$, on tire l'intervalle de confiance au niveau 95% [0,36; 1,92].

La différence $\mu_1 - \mu_2$ est donc estimée par $\widehat{\mu}_1 - \widehat{\mu}_2 = 0{,}18$. On va supposer que $\sigma^2_1 = \sigma^2_2$ (ce qui semble légitime vu l'intervalle de confiance). L'estimation de cette variance commune est :

$$\frac{19 \times 2{,}94 + 29 \times 3{,}68}{19 + 29} = 3{,}39,$$

et l'intervalle de confiance au niveau $1 - \alpha$ pour $\mu_1 - \mu_2$ est

$$0{,}18 \pm t^{48}_{1-\frac{\alpha}{2}}\sqrt{3{,}39\left(\frac{1}{20} + \frac{1}{30}\right)}.$$

On a $t^{48}_{0,95} = 1{,}677$, d'où l'intervalle de confiance au niveau 90% : $[-0{,}71; 1{,}07]$; et $t^{48}_{0,975} = 2{,}011$, d'où l'intervalle de confiance au niveau 95% : $[-0{,}89; 1{,}25]$.

Corrigé de l'exercice 4 1. On calcule tout d'abord $\widehat{\mu} = \frac{1}{40}23 = 0{,}575$. La variance empirique est $s^2 = 47{,}92$.

Comme on ne suppose que la loi est normale, on construit un intervalle de confiance basé sur le théorème de la limite centrale :

$$0{,}575 \pm 1{,}96\sqrt{\frac{47{,}92}{40}} = [-1.57; 2{,}72].$$

2. Si la loi est normale, on utilise le quantile $t^{39}_{0,975} = 2{,}023$:

$$0{,}575 \pm 2{,}023\sqrt{\frac{47{,}92}{40}} = [-1.64; 2{,}79].$$

Paradoxalement, supposer que la loi est normale augmente la taille de l'intervalle de confiance. La raison en est que, dans ce dernier cas, le calcul effectué grâce aux quantiles de la loi de Student est exact ; dans le premier cas, on ne fait qu'un calcul approché, basé sur le théorème de la limite centrale. L'intervalle de confiance obtenu est plus douteux.

Corrigé de l'exercice 5 On estime $\widehat{p} = \frac{4}{159} = 2{,}5\%$. On n'est pas dans les conditions d'utilisation de l'intervalle de confiance de Wald, on va donc utiliser l'intervalle de confiance sur $\Phi(0{,}025) = 0{,}159$. La borne inférieure est

$$\Phi\left(\frac{3{,}5}{159}\right) - 1{,}96\frac{1}{2\sqrt{159}} = 0{,}071,$$

la borne supérieure est

$$\Phi\left(\frac{4{,}5}{159}\right) + 1{,}96\frac{1}{2\sqrt{159}} = 0{,}247.$$

On en déduit l'intervalle de confiance sur p : $\left[\sin(0{,}071)^2; \sin(0{,}247)^2\right] = [0{,}5\%; 5{,}9\%]$.

Corrigé de l'exercice 6 1. On a bien sûr $p = \frac{1}{n}$.

2. On répète N expériences de Bernoulli indépendantes, et on compte les succès (boule noire) qui se produisent avec probabilité p, la loi de X est donc une loi binomiale $\mathscr{B}in(\mathrm{N}, p)$.

3. On calcule l'estimation de p : $\widehat{p} = \frac{x}{\mathrm{N}} = \frac{740}{5000} = 0{,}148$. Un intervalle de confiance à 95% se calcule par la formule $\widehat{p} \pm 1{,}96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{\mathrm{N}}}$, on obtient : $[0{,}1382; 0{,}1578]$.

4. On en déduit un intervalle de confiance à 95% pour n, entre $\frac{1}{0{,}1578} = 6{,}34$ et $\frac{1}{0{,}1382} = 7{,}24$. Comme n est entier, cet intervalle de confiance se réduit à un seul élément, 7.

Corrigé de l'exercice 7 1. $\overline{\mathrm{X}} \sim \mathscr{N}\left(\mu, \frac{1}{n}\sigma^2\right)$.

2. $\overline{\mathrm{X}}$ et X_{n+1} sont indépendantes, donc $(\mathrm{X}_{n+1} - \overline{\mathrm{X}}) \sim \mathscr{N}\left(0, \left(1 + \frac{1}{n}\right)\sigma^2\right)$.

3. L'écart-type de $(\mathrm{X}_{n+1} - \overline{\mathrm{X}})$ est $\sqrt{\frac{n+1}{n}}\sigma$, donc Y est égale à une variable gaussienne centrée $(\mathrm{X}_{n+1} - \overline{\mathrm{X}})$, divisée par son écart-type : c'est bien une variable gaussienne centrée réduite.

4. On a $\mathrm{Y} \sim \chi^2(n-1)$ (cf cours).

5. Notons que dans le modèle gaussien, $\overline{\mathrm{X}}$ et S^2 sont indépendantes ; on en déduit que Y et Z sont indépendantes. La définition de T est celle d'une variable suivant une loi de Student à $n-1$ degrés de liberté.

6. On a $\mathbb{P}(t^9_{0,025} \leqslant \mathrm{T} \leqslant t^9_{0,975}) = 0{,}95$. On lit dans la table $t^9_{0,975} = 2{,}26$; d'autre part, par symétrie de la fonction de distribution de la loi de Student, on a $t^9_{0,025} = -t^9_{0,975}$.

7. On a $\overline{x} = \frac{1}{10}207 = 20{,}7$ et

$$s^2 = \frac{10}{9}\left(\frac{1}{10}4297 - (20{,}7)^2\right) = 1{,}34,$$

d'où $s = \sqrt{1{,}34} = 1{,}16$ (estimation de σ).

8. On calcule

$$\overline{x} \pm t^{n-1}_{0,975}\frac{s}{\sqrt{n}},$$

avec $n = 10$ d'où $t^9_{0,975} = 2{,}26$, d'où l'intervalle de confiance pour μ : $[19{,}87; 21{,}53]$.

9. On calcule

$$\left[\frac{(n-1)}{x^{n-1}_{0,975}}s^2; \frac{(n-1)}{x^{n-1}_{0,035}}s^2\right]$$

où, pour $n = 10$, les quantiles de la loi de $\chi^2(9)$ sont $x^9_{0,975} = 19{,}023$ et $x^9_{0,025} = 2{,}700$, d'où l'intervalle de confiance pour σ^2 : $[0{,}63; 4{,}46]$.

10. On reprend le résultat démontré en partie I : dans le cas où $n = 10$, on a

$$\mathbb{P}(-2{,}26 \leqslant \mathrm{T} \leqslant 2{,}26) = 0{,}95).$$

La définition de T est

$$\begin{aligned} \mathrm{T} &= \frac{\mathrm{Z}}{\sqrt{\frac{\mathrm{Y}}{n-1}}} \\ &= \frac{\sigma \mathrm{Z}}{\mathrm{S}} \\ &= \sqrt{\frac{n}{n+1}}\left(\frac{\mathrm{X}_{n+1} - \overline{\mathrm{X}}}{\mathrm{S}}\right). \end{aligned}$$

et dans le cas où $n = 10$, on a donc avec probabilité 95%

$$\begin{aligned} -2{,}26 \leqslant \sqrt{\frac{10}{11}}\left(\frac{\mathrm{X}_{11} - \overline{\mathrm{X}}}{\mathrm{S}}\right) &\leqslant 2{,}26 \\ -2{,}26 \times \sqrt{\frac{11}{10}} \times \mathrm{S} \leqslant \mathrm{X}_{11} - \overline{\mathrm{X}} &\leqslant 2{,}26 \times \sqrt{\frac{11}{10}} \times \mathrm{S} \\ \overline{\mathrm{X}} - 2{,}26 \times \sqrt{1{,}1} \times \mathrm{S} \leqslant \mathrm{X}_{11} &\leqslant \overline{\mathrm{X}} + 2{,}26 \times \sqrt{1{,}1} \times \mathrm{S} \end{aligned}$$

En remplaçant $\overline{\mathrm{X}}$ et S par les valeurs calculées à la question 7, on en tire $\mathbb{P}(17{,}95 \leqslant \mathrm{X}_{11} \leqslant 23{,}45) = 0{,}95$.

A.10 Exercices du chapitre 10

Corrigé de l'exercice 1 1. C'est exactement la même question que pour le tirage de deux dés :

$$\begin{aligned} \mathbb{P}(|\mathrm{X}| > 1{,}96 \text{ ou } |\mathrm{Y}| > 1{,}96) &= 1 - \mathbb{P}(|\mathrm{X}| \leqslant 1{,}96 \text{ et } |\mathrm{Y}| \leqslant 1{,}96) \\ &= 1 - \mathbb{P}(|\mathrm{X}| \leqslant 1{,}96)\mathbb{P}(|\mathrm{Y}| \leqslant 1{,}96) \quad \text{car X et Y sont indépendantes} \\ &= 1 - (0{,}95)^2 \\ &= 0{,}0975. \end{aligned}$$

La zone d'acceptation est le carré défini par $|\mathrm{X}| \leqslant 1{,}96$ et $|\mathrm{Y}| \leqslant 1{,}96$.

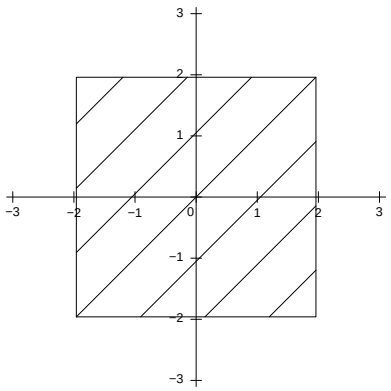


Figure 77. Zone d'acceptation du test de la question 1

2. Soit $\alpha_1 = \mathbb{P}(|\mathrm{X}| > c)$. D'après ce qui précède on a $\alpha = 1 - (1 - \alpha_1)^2$. On veut $\alpha = 0{,}05$ donc $(1 - \alpha_1)^2 = 0{,}95$ puis $(1 - \alpha_1) = 0{,}96468$ et pour finir $\alpha_1 = 0{,}02532$.

On n'a que des tables unilatérales : on écrit donc $\alpha_1 = 2\mathrm{P}(\mathrm{X} > c)$; on doit trouver c tel que $\mathrm{P}(\mathrm{X} > c) = 0{,}01266$; on lit dans la table $c \simeq 2{,}24$.

3. Par définition de la loi de χ^2, $Z \sim \chi^2(2)$.

4. On rejettera l'hypothèse d'un état non pathologique quand $Z > 5{,}99$, le quantile de niveau 0,95 de la loi $\chi^2(2)$.

La zone d'acceptation de ce test est un disque de rayon $\sqrt{5{,}99} \simeq 2{,}45$.

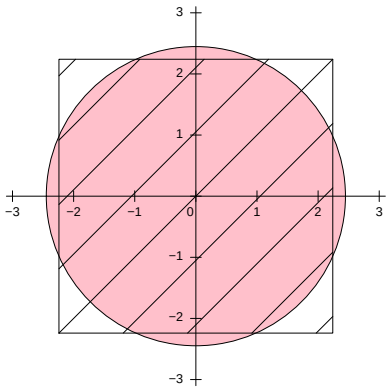


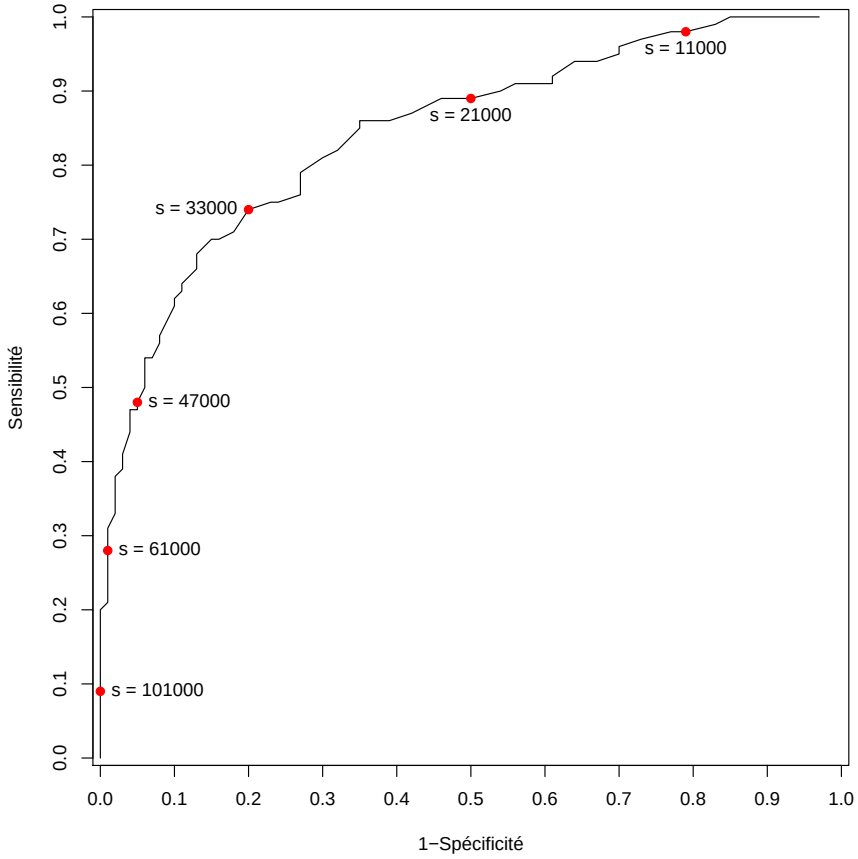
Figure 78. Zone d'acceptation des tests des questions 2 et 4

Corrigé de l'exercice 2

1. On demande que 95% des témoins soient en-dessous du seuil, c'est-à-dire $F(s) = 0{,}95$. On prend donc $s = 47000$, et la sensibilité est la proportion de cas au-dessus du seuil : $1 - G(s) = 0{,}48$. (La réponse $s = 47940$ avec une sensibilité 0,47 était admise également)

2. Pour $s = 21000$, on lit la spécifité $F(s) = 0{,}50$ et la sensibilité $1 - G(s) = 0{,}89$. Pour $s = 46000$, on a $F(s) = 0{,}94$ et $1 - G(s) = 0{,}50$.

3. On peut utiliser les points précédents, auxquels on ajoutera quelques points avant de tracer la courbe à main levée (ici on a tracé la courbe obtenue avec toutes les valeurs seuils de la table).



Corrigé de l'exercice 3

1. Il est naturel de supposer que les accidents de la route, et en particulier les accidents mortels, surviennent selon un processus de Poisson. La loi du nombre total de tués pendant une certaine période doit donc suivre (approximativement) une loi de Poisson. Le paramètre de cette loi dépend de la longueur de la période considérée, donc il doit dépendre du mois. Il dépendra également d'autres paramètres qui peuvent varier d'une année à l'autre (conditions météos en particulier, ou politique de prévention et sécurité routière – *cf.* question suivante) : en cela cette modélisation n'est qu'une approximation. S'y ajoute le fait que les décès en sont pas indépendants (accidents impliquants plusieurs véhicules, véhicules transportants plusieurs personnes...).

2. Tous les X_i et Y_i sont indépendants. Si on suppose que rien ne change entre 2010 et 2011, on a pour chaque i (le numéro du mois), $X_i, Y_i \sim \mathscr{P}(\lambda_i)$ approximativement (le paramètre dépend du mois mais pas de l'année). On a donc $U_i, V_i \sim \mathscr{N}(2\sqrt{\lambda_i}, 1)$ et $\Delta_i = (V_i - U_i) \sim \mathscr{N}(0,2)$ (la situation rappelle le cas classique des séries appariées, en plus simple car les variances sont connues).

La somme $\Delta_1 + \cdots + \Delta_4$ suit donc approximativement une loi $\mathscr{N}(0,8)$, il suffit de la diviser par $\sqrt{8}$ pour obtenir une loi normale centrée réduite. On fait un test au niveau 5% en comparant la valeur obtenue au quantile 0,95 de la loi normale (test unilatéral car sous H_1 les Δ_i sont positifs).

On calcule $\frac{1}{\sqrt{8}}(\delta_1 + \cdots + \delta_4) = 2{,}92$: le test est significatif. On peut également calculer le degré de signification, $p = 0{,}00175$.

Corrigé de l'exercice 4 1. Par définition de la loi de χ^2, on a $Z \sim \sigma^2 \chi^2(35)$.

2. Sous H_0, on a $Z \sim \chi^2(35)$. On rejettera H_0 quand $Z \geq x^{35}_{0,95}$, le quantile de niveau 0,95 de la loi $\chi^2(35)$, qui (d'après la table des quantiles) est égal à 49,80.

3. On estime σ^2 par $\frac{1}{35} z = 0{,}86$. D'après la question précédente, on ne rejette pas H_0 pour $z \leqslant 49{,}80$.

4. On a donc $Z' = \frac{Z}{1,45} \sim \chi^2(35)$. Le risque β est la probabilité de ne pas rejeter H_0

$$\beta = \mathbb{P}(Z \leqslant 49,80) = \mathbb{P}\left(\frac{Z}{1,45} \leqslant \frac{49,80}{1,45}\right) = \mathbb{P}(Z' \leqslant 34,34) \simeq 0,5$$

d'après la table des quantiles.

Corrigé de l'exercice 5 1. La probabilité de n'avoir aucune tache est nulle pour les araignées A, on en conclut qu'il s'agit certainement d'une araignée B. Je refuse donc le traitement (ceci implique d'avoir une confiance absolue dans la table ci-dessus...).

2. La probabilité de d'avoir plus de neuf taches est nulle pour les araignées B, alors que c'est le cas de 27% des araignées A. J'accepte le traitement.

3. (a) Le risque α est la probabilité de rejeter H_0 quand H_0 est vraie, c'est-à-dire $\mathbb{P}(N > s|B)$:

$$\begin{aligned}\alpha &= \mathbb{P}(N > 5|B)\\ &= \mathbb{P}(N = 6|B) + \mathbb{P}(N = 7|B) + \mathbb{P}(N = 8|B) + \mathbb{P}(N = 9|B) + \mathbb{P}(N \leqslant 10|B)\\ &= 0,05 + 0,02 + 0,01 + 0,00 + 0,00\\ &= 0,08\end{aligned}$$

La puissance $(1-\beta)$ est la probabilité de rejeter H_0 quand H_0 est fausse, c'est-à-dire $\mathbb{P}(N > s|A)$:

$$\begin{aligned}1-\beta &= \mathbb{P}(N > 5|A)\\ &= 0,15 + 0,15 + 0,13 + 0,10 + 0,17\\ &= 0,70\end{aligned}$$

(b) Pour $s = 3$, le même calcul donne $\alpha = 0,35$ et $1-\beta = 0,92$.
Pour $s = 5$, on obtient $\alpha = 0,01$ et $1-\beta = 0,40$.

A.11 Exercices du chapitre 11

Corrigé de l'exercice 1 1. On calcule $\overline{x} = \frac{13990}{50} = 279,8$. L'estimation non corrigée de la variance est $\tilde{s}^2 = \frac{3917264}{50} - (279,8)^2 = 57,24$, l'estimation corrigée est $s^2 = \frac{50}{49} \times 57,24 = 58,41$ jours. L'écart-type estimé est $\sqrt{58,41} = 7,64$ jours.

2. Considérons d'abord la variance : on teste $H_0 : \sigma^2 = \sigma_0^2$ contre $H_1 : \sigma^2 \neq \sigma_0^2$, avec $\sigma_0^2 = 6,75^2 = 45,56$. Sous H_0, la loi de $T = \frac{(n-1)}{\sigma_0^2} S^2$ est une loi $\chi^2(n-1)$.

On doit donc rejeter H_0 quand $T < x_{0,025}^{n-1}$ ou $T > x_{0,975}^{n-1}$, les x_α^{n-1} étant les quantiles de la loi $\chi(n-1)$. Pour $n-1 = 49$ ils valent respectivement 31,55 et 70,22. La valeur observée de T est $t = 49 \times \frac{58,41}{45,56} = 62,82$: on ne rejette pas H_0.

On ne dispose cependant que rarement d'une table de χ^2 à 49 degrés de libertés, et l'ordinateur n'est pas supposé exister. On peut alors approcher la loi de χ^2 par une loi normale de même moyenne et de même variance (rappelons qu'une variable suit une loi $\chi^2(d)$ si elle est la somme de d variables normales au carré ; on applique le théorème de la limite centrale).

La loi de T est donc approximativement $\mathcal{N}(n-1, 2(n-1))$; on centre et réduit T :

$$T_1 = \frac{T-(n-1)}{\sqrt{2(n-1)}} = \sqrt{\frac{n-1}{2}}\left(\frac{S^2}{\sigma_0^2} - 1\right);$$

la loi de T_1 est approximativement $\mathcal{N}(0,1)$.

Calculons la valeur observée de T_1 : $t_1 = \sqrt{\frac{49}{2}}\left(\frac{58,41}{45,56} - 1\right) = 1,40$. On ne rejette pas H_0.

Considérons maintenant la moyenne : on teste $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$, où $\mu_0 = 280$ jours. On va utiliser la valeur a priori de l'écart-type plutôt que la valeur observée ; l'autre solution est défendable également et mène à un test t.

Sous H_0, l'estimation de μ suit une loi normale de moyenne $\mu_0 = 280$ et d'écart-type $\frac{6,75}{\sqrt{50}}$. On calcule l'écart réduit, qui suit une loi normale $\mathcal{N}(0,1)$:

$$z = \frac{279,8 - 280}{\frac{6,75}{\sqrt{50}}} = -0,21.$$

Là encore, on conserve H_0.

Corrigé de l'exercice 2 1. On calcule $s_y^2 = 53,53$. La déviation à la variance attendue (45,56) étant moins importante que celle observée chez Martial ($s_x^2 = 58,41$) on sait qu'on ne va pas rejetter l'hypothèse de conformité à la variance a priori.

Pour comparer les deux variances observées, on fait un test F pour tester $H_0 : \sigma_x^2 = \sigma_y^2$ contre $H_1 : \sigma_x^2 \neq \sigma_y^2$.

Sous H_0, le quotient $\frac{S_x^2}{S_y^2}$ suit une loi F(49,49).

On calcule $\frac{s_x^2}{s_y^2} = \frac{58,41}{53,53} = 1,09$.

On rejette H_0 (au risque $\alpha = 5\%$) si $\frac{s_x^2}{s_y^2}$ n'est pas dans $[F_{0.025}^{49,49}; F_{0.975}^{49,49}]$. On a $F_{0.025}^{49,49} = 1/F_{0.975}^{49,49} = 0.57$. On ne rejette pas H_0.

2. On peut hésiter à faire un test t, ce à quoi on est incité par la question qui précède. Cependant on a une valeur a priori $\sigma_0 = 6,75$ pour la variance, et on a tout lieu de la croire fondée ; on va donc l'utiliser.

On fait un test unilatéral (l'hypothèse que les Charolaises ont une durée de gestation plus élevée étant posée a priori) : $H_0 : \mu_x = \mu_y$ contre $H_1 : \mu_x < \mu_y$, où μ_x (respectivement μ_y) est la durée moyenne de la gestation des vaches de M. Bottafoin (respectivement J. Carnicero).

On calcule donc $\overline{y} = 289,24$, et l'écart réduit :

$z = \frac{\overline{y} - \overline{x}}{\sigma_0 \sqrt{\frac{2}{n}}} = \frac{289,24 - 279,8}{6,75\sqrt{\frac{2}{50}}} = 6,99$. On a $z > 1,64$, on rejette H_0.

Corrigé de l'exercice 3 Le test F ne rejette pas l'égalité des variances si $F_{0,025}^{d_1,d_2} \leqslant \frac{s_x^2}{s_y^2} \leqslant F_{0,975}^{d_1,d_2}$, et donc si

$$s_y^2 \leqslant \frac{s_x^2}{F_{0,025}^{d_1,d_2}} = F_{0,975}^{d_2,d_1} s_x^2 \text{ et } s_y^2 \geqslant \frac{s_x^2}{F_{0,975}^{d_1,d_2}}.$$

Pour $d_1 = 100, d_2 = 5$, on a $F_{0,975}^{5,100} = 2,70$ et $F_{0,975}^{100,5} = 6,08$;

Pour $d_1 = 100, d_2 = 100$, on a $F_{0,975}^{100,100} = 1,48$;

Pour $d_1 = 5, d_2 = 5$, on a $F_{0,975}^{5,5} = 7,15$.

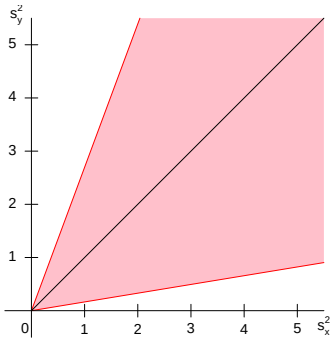


Figure 79. $d_1 = 100, d_2 = 5$

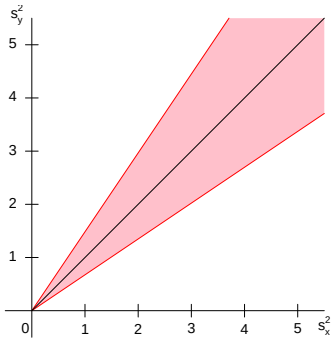


Figure 80. $d_1 = 100, d_2 = 100$

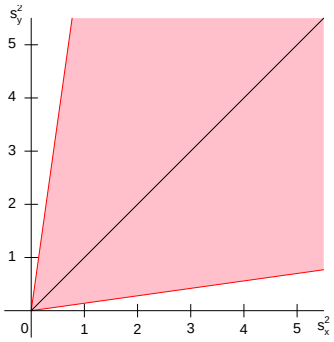


Figure 81. $d_1 = 5, d_2 = 5$

La zone d'acceptation est coloriée, la zone de rejet est tout le reste du plan.
On remarque notamment que le graphe est symétrique par rapport à la droite $y = x$ quand $d_1 = d_2$.

Corrigé de l'exercice 4 On calcule $\widehat{p} = \frac{25}{40} = 0{,}625$ et $\Phi(\widehat{p}) = 0{,}9117$

1. En estimant $\frac{p(1-p)}{n}$ par $\frac{\widehat{p}(1-\widehat{p})}{n}$, on obtient l'intervalle de confiance de niveau 95% sur $\widehat{p}$:

$$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0{,}625 \pm 0{,}15 = [0{,}475; 0{,}775].$$

On vérifie qu'on est bien dans les conditions d'utilisation de l'approximation normale sur tout l'intervalle (np et $n(1-p) > 5$).
D'autre part on a un intervalle de confiance sur $\Phi(\widehat{p})$:

$$\Phi(\widehat{p}) \pm 1.96\frac{1}{2\sqrt{n}} = 0{,}9117 \pm 0{,}1550 = [0{,}7567; 1{,}0667].$$

Pour retrouver un intervalle de confiance sur $\widehat{p}$ il suffit d'appliquer la fonction réciproque de Φ, $\Phi^{-1}(x) = \sin(x)^2$. On obtient l'intervalle $[\sin(0{,}7567)^2; \sin(1{,}0667)^2] = [0{,}471; 0{,}767]$.

2. On teste $H_0 : p = \frac{1}{2}$ contre $H_1 : p \neq \frac{1}{2}$ au risque $\alpha = 5\%$. On peut déjà prévoir que les tests vont être négatifs (les intervalles de confiance contiennent largement 0,5).

Première méthode : on utilise $\widehat{p}$. On pose $Z = \frac{\widehat{p}-0{,}5}{\sqrt{\frac{0{,}5\cdot 0{,}5}{n}}}$. Z suit une loi normale centrée réduite ; ici on calcule $z = 1{,}5811$. On ne rejette pas H_0.

Seconde méthode : on utilise $\Phi(\widehat{p})$. On pose $Z = \frac{\Phi(\widehat{p})-\Phi(0{,}5)}{\frac{1}{2\sqrt{n}}} \sim \mathcal{N}(0,1)$. On calcule $z = 1{,}598$. On ne rejette pas H_0.

Corrigé de l'exercice 5 1. Sous H_0, la loi de $\Phi(\widehat{p})$ est approximativement $\mathcal{N}\left(\Phi(p_0), \frac{1}{4n}\right)$. En posant

$$Z = \frac{\Phi(\widehat{p}) - \Phi(p_0)}{\sqrt{\frac{1}{4n}}},$$

on a $Z \sim \mathcal{N}(0,1)$ (sous S_0), et $\mathbb{P}(|Z| > 1{,}96) = 0{,}05$.

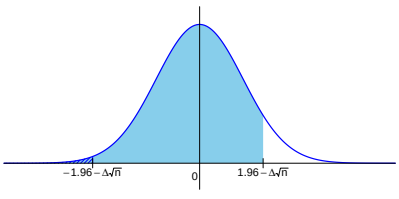
D'autre part on a $Z = 2\sqrt{n}\left(\Phi(\widehat{p}) - \Phi(p_0)\right)$. d'où le résultat.

2. La loi de $\Phi(\widehat{p})$ sous H_1 est $\mathcal{N}\left(\Phi(p_1), \frac{1}{4n}\right)$. En posant $Z_1 = 2\sqrt{n}\left(\Phi(\widehat{p}) - \Phi(p_1)\right)$, on a $Z_1 \sim \mathcal{N}(0,1)$ sous H_1.

On a

$$\begin{aligned}
\beta &= \mathbb{P}\left(-1{,}96 \leqslant 2\sqrt{n}\left(\Phi(\widehat{p}) - \Phi(p_0)\right) \leqslant 1{,}96\right) \\
&= \mathbb{P}\left(-1{,}96 \leqslant Z_1 + 2\sqrt{n}\left(\Phi(p_1) - \Phi(p_0)\right) \leqslant 1{,}96\right) \\
&= \mathbb{P}\left(-1{,}96 - \Delta\sqrt{n} \leqslant Z_1 \leqslant 1{,}96 - \Delta\sqrt{n}\right) \\
&= \mathbb{P}\left(Z_1 \leqslant 1{,}96 - \Delta\sqrt{n}\right) - \mathbb{P}\left(Z \leqslant -1{,}96 - \Delta\sqrt{n}\right).
\end{aligned}$$

3. La probabilité $\mathbb{P}\left(Z \leqslant 1{,}96 - \Delta\sqrt{n}\right)$ est en coloriée en bleu sur la figure ; la probabilité $\mathbb{P}\left(Z \leqslant -1{,}96 - \Delta\sqrt{n}\right)$ est hachurée.



Dès que $\Delta\sqrt{n}$ est assez grand, l'aire hachurée devient négligeable. On a $P(Z \leqslant 1{,}96 - \Delta\sqrt{n}) \approx \beta$ et donc $z_\beta \approx 1{,}96 - \Delta\sqrt{n}$.

4. On a donc $\Delta = 2\left(\Phi(0{,}515) - \Phi(0{,}5)\right) = 0{,}03$. Pour avoir $1 - \beta = 0{,}8$, c'est-à-dire $\beta = 0{,}2$ et $z_\beta = -0.84$, il faut $1{,}96 - \Delta\sqrt{n} \approx -0{,}84$, d'où $\sqrt{n} \approx \frac{1{,}96+0{,}84}{0{,}03} = 91{,}33$, d'où $n \approx 8711$.

A.12 Exercices du chapitre 12

Corrigé de l'exercice 1

1. Le sex ratio étant équilibré, c'est $\frac{1}{2}(176+168) = 172$.

2. On peut utiliser le formalisme de l'anova : la variance calculée sur l'échantillon est CMT = 73,50, d'où SCT = 99 × 73,50 = 7276,50. D'autre part on peut calculer SCF, par exemple par

$$\text{SCF} = \frac{1}{100} 50 \times 50 \times (176-168)^2 = 1600$$

ou encore

$$\text{SCF} = 50 \times 176^2 + 50 \times 168^2 - \frac{1}{100} 17200^2 = 1600.$$

On en déduit SCR = SCT − SCF = 5676,50, et la variance commune aux hommes et aux femmes est estimée par CMR = $\frac{1}{98}$SCR = 57,92.

Une autre solution est d'écrire directement que la variance commune est

$$\frac{\sum_{i=1}^{50}(x_i^{\text{H}}-176)^2 + \sum_{i=1}^{50}(x_i^{\text{F}}-168)^2}{98}.$$

D'une part on a

$$\sum_{i=1}^{50}(x_i^{\text{H}}-176)^2 + \sum_{i=1}^{50}(x_i^{\text{F}}-168)^2 = \sum_{i=1}^{100} x_i^2 - 2\times 176 \times \sum_{i=1}^{50} x_i^{\text{H}} - 2 \times 168 \times \sum_{i=1}^{50} x_i^{\text{F}} + 50 \times 176^2 + 50 \times 168^2$$

avec $\sum_{i=1}^{50} x_i^{\text{H}} = 50 \times 176 = 8800$, $\sum_{i=1}^{50} x_i^{\text{F}} = 8400$, et d'autre part on a

$$\frac{1}{99}\left(\sum_{i=1}^{100} x_i^2 - \frac{1}{100}\left(\sum_{i=1}^{100} x_i\right)^2\right) = 73{,}50$$

avec $\sum_i x_i = 17200$. On en déduit

$$\sum_{i=1}^{100} x_i^2 = 99 \times 73{,}50 + \frac{1}{100} 17200^2 = 2965676{,}50$$

et on peut finalement calculer la variance commune

$$\frac{1}{98}\left(2965676{,}50 - 2\times 176 \times 8800 - 2 \times 168 \times 8400 + 50 \times 176^2 + 50 \times 168^2\right) = 57{,}92.$$

Corrigé de l'exercice 2 On calcule tout d'abord la somme des carrés dans chacun des groupes :
$\text{SC}_1 = x_{1+}^2 - \frac{1}{n_1}(x_{1+})^2 = 267{,}3067 - \frac{1}{12}55{,}17^2 = 13{,}66$, $\text{SC}_2 = 40{,}81$ et $\text{SC}_3 = 80{,}98$.
On en déduit $\text{SCR} = \text{SC}_1 + \text{SC}_2 + \text{SC}_3 = 135{,}45$.
On a $n = n_1 + n_2 + n_3 = 100$, $x_{++} = x_{1+} + x_{2+} + x_{3+} = 500{,}52$, $x_{++}^2 = x_{1+}^2 + x_{2+}^2 + x_{3+}^2 = 2644{,}114$. On calcule $\text{SCT} = x_{++}^2 - \frac{1}{n}(x_{++})^2 = 138{,}91$.
On peut ensuite calculer SCF = SCT − SCR = 3,46.
La table d'anova est

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	3,46	2	1,73	F = 1,23
résidus	135,45	97	1,40	
Total	138,91	99	1,40	

La valeur de F est à comparer au quantile de niveau 0,95 de F(2,97), qui vaut environ 3,09 : on ne rejette pas l'hypothèse d'égalité des groupes, pour un test de risque α = 5%.

Corrigé de l'exercice 3 1. L'hypothèse est que pour la dose c, le TCP moyen est de $\mu + c \cdot r$ où r est une constante inconnue. Dans notre cas on a donc

$$\begin{cases} \alpha_1 &= 10 \cdot r \\ \alpha_2 &= 15 \cdot r \\ \alpha_3 &= 20 \cdot r \end{cases}$$

ce qui équivaut à $\alpha_2 - \alpha_1 = \alpha_3 - \alpha_2$ (l'évolution du TCP est la même quand on passe de 10 à 15 mg et quand on passe de 15 à 20 mg), d'où le résultat demandé.

2. La moyenne $X_{i\bullet}$ est une estimation de $\mu + \alpha_i$, donc $X_{1\bullet} - 2X_{2\bullet} + X_{3\bullet}$ est une estimation de $\mu + \alpha_2 - 2(\mu + \alpha_2) + \mu + \alpha_3 = \alpha_1 - 2\alpha_2 + \alpha_3$.

La loi de chaque $X_{i\bullet}$ est la loi normale $\mathcal{N}\left(\mu_i, \frac{1}{n_i}\sigma^2\right)$, donc la loi de $\widehat{C}$ est $\mathcal{N}\left(C, \left(\frac{1}{n_1} + \frac{4}{n_2} + \frac{1}{n_3}\right)\sigma^2\right)$.

3. On calcule $\widehat{C} = \frac{700}{12} - 2\frac{745}{10} + \frac{829}{10} = -7{,}77$. La variance σ^2 est estimée par CMR = 462,9 (calcul fait dans le cours), d'où un écart-type estimé pour $\widehat{C}$:

$$\sqrt{\left(\frac{1}{n_1} + \frac{4}{n_2} + \frac{1}{n_3}\right)\text{CMR}} = 16{,}43.$$

L'intervalle de confiance est

$$[\widehat{C} \pm t^{29}_{0,975}\sqrt{\left(\frac{1}{n_1} + \frac{4}{n_2} + \frac{1}{n_3}\right)\text{CMR}}] = [-7{,}77 \pm 2{,}05 \times 16{,}43] = [-41{,}45; 25{,}91].$$

L'intervalle de confiance contient 0, on ne rejette pas l'hypothèse d'un effet-dose linéaire.

Corrigé de l'exercice 4

1. On réalise donc une anova à un facteur : l'origine géographique.

Remarquez qu'on pourrait débattre longuement de savoir s'il s'agit d'un modèle à effet fixe (on s'intéresse à l'efficacité du traitement dans les quatres régions précises où on a sélectionné les patients) ou d'un modèle à effets aléatoires (on s'intéresse à la variabilité de l'effet du traitement à l'échelle mondiale).

On peut compléter le tableau de données en calculant les sommes de carrés dans chaque groupe :

	n_i	x_{i+}	x^2_{i+}	SC_i
Groupe 1	10	568	37 186	4 923,6
Groupe 2	10	621	42 965	4 400,9
Groupe 3	10	879	85 781	8 516,9
Groupe 4	10	732	55 390	1 807,6
Total	40	2 800	221 322	25 322

On a donc SCR = $\sum_i SC_i$ = 19649, SCT = 25322 et SCF = SCT − SCR = 5673.

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	5 673	3	1 891	F = 3,46
résidus	19 649	36	545,8	
Total	25 322	39	649,3	

La valeur de F est à comparer avec le quantile 0,95 de F(3,36) = 2,86 : on rejette l'hypothèse d'égalité.

2. En fusionnant les groupes 1 et 2 d'une part, 3 et 4 d'autre part, on obtient :

	n_i	x_{i+}	x^2_{i+}	SC_i
Groupe 1+2	20	1 189	80 151	9 464,95
Groupe 3+4	20	1 611	141 171	11 404,95

La somme des carrés associée au modèle à deux paramètres $\mu_1 = \mu_2$ et $\mu_3 = \mu_4$ est $SC_{1,2} + SC_{3,4} = 20869{,}9$.

La table d'anova « sans tests », pour les trois modèles en concurrence, est la suivante :

Source	Somme des carrés	degrés de liberté	Carrés moyens
(c) (4 par.)	19 659	36	545,8
(b) (2 par.)	20 869,9	38	549,2
(a) (1 par.)	25 322	39	649,3

On a déjà comparé (c) et (a) à la question 1 : le résultat était que le modèle à 4 paramètres (c) explique significativement mieux les observations que le modèle (a).

Comparons (b) et (a).

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(a)-(b)	4 452,1	1	4 452,1	8,1
(b) (2 par.)	20 869,9	38	549,2	
(a) (1 par.)	25 322	39	649,3	

La statistique F = 8,1 est à comparer au quantile d'ordre 0,95 de F(1,38), qui vaut environ 4 : on rejette l'hypothèse nulle (le modèle (a) suffit à expliquer les observations) au profit du modèle (b).

Comparons maintenant (b) et (c).

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
(b)-(c)	1 210,9	2	605,5	1,33
(c) (4 par.)	19 659	36	545,8	
(b) (2 par.)	20 869,9	38	549,2	

La statistique F = 1,33 est à comparer au quantile d'ordre 0,95 de F(2,36), qui vaut environ 3,2 : le modèle (c) n'explique pas significativement mieux les observations que le modèle (b).

Des trois modèles, on retient finalement le modèle (b).

Corrigé de l'exercice 5 1. On calcule $SC_1 = 7173{,}56 - \frac{1}{10}(253{,}40)^2 = 752{,}40$, et de même $SC_2 = 965{,}02$, $SC_3 = 1111{,}34$, $SC_4 = 670{,}00$, d'où SCR = 3498,76. On a $n = 65$, $x_{++} = 1861{,}20$ et $x_{++}^2 = 57628{,}06$ d'où SCT = 4334,75. On peut compléter la table de Fisher :

Sources	SC	ddl	CM	F
Facteur	835,99	3	278,66	4,86
Résidus	3498,76	61	57,36	
Total	4334,75	64	67,73	

Pour un test au risque $\alpha = 0{,}05$, il faut comparer la valeur de F = 4,86 au quantile $F_{0,95}^{3,61} = 2{,}7$: on rejette H_0.

2. On estime la variance résiduelle σ^2 par CMR = 57,36. La loi de CMR est $\frac{\sigma^2}{61}\chi^2(61)$, d'où l'intervalle de confiance à 90% pour σ^2 :

$$\left[\frac{61\text{CMR}}{x_{0,95}^{61}}; \frac{61\text{CMR}}{x_{0,05}^{61}}\right]$$

On lit dans les tables les quantiles $x_{0,05}^{61} \simeq 44$ et $x_{0,95}^{61} \simeq 80$, d'où l'intervalle de confiance [43,7;79,5].

3. La moyenne est estimée par $\frac{502{,}50}{20} = 25{,}125$. On calcule un intervalle de confiance en se basant sur la variance estimée par CMR :

$$25{,}125 \pm t_{0,95}^{61}\sqrt{\frac{\text{CMR}}{20}}.$$

On lit dans la table de la loi t le quantile $t^{61}_{0,95} \simeq 1,67$, d'où l'intervalle de confiance [22,3;28,0].

4. La moyenne du groupe 3 est $\frac{592,60}{20} = 29,63$. Pour la comparaison, on utilise un contraste, qui mène à un écart réduit

$$\frac{29,63 - 25,125}{\sqrt{\text{CMR}\left(\frac{1}{20} + \frac{1}{20}\right)}} = 1,88.$$

Pour un test bilatéral au risque $\alpha = 0,05$, on compare au quantile $t^{61}_{0,975} \simeq 2,00$: la différence n'est pas significative.

Corrigé de l'exercice 6 1. Taux d'anticorps moyen $\overline{x}_\text{A} = \frac{1}{10}152,8 = 15,28$.

Pour calculer l'intervalle de confiance, on calcule d'abord la somme de carrés associée (qui servira à nouveau à la question 2) :

$$\begin{aligned}\text{SC}_\text{A} &= \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2 \\ &= 2392,66 - \frac{1}{10}15,28^2 = 57,876\end{aligned}$$

On en déduit une estimation de la variance du taux d'anticorps chez les patients de l'hôpital A :

$$\widehat{\sigma}^2 = \frac{1}{n-1}\text{SC}_\text{A} = \frac{1}{9}57,879 = 6,431$$

et donc un intervalle de confiance basé sur le quantile de niveau 97,5 % de la loi t à 9 ddl, $t^9_{0,975} = 2,262$:

$$15,28 \pm 2,262\sqrt{6,431/10} = [13,47\,;\,17,09]$$

2. Les moyennes sont $\overline{x}_\text{B} = 14,67$ et $\overline{x}_\text{C} = 14,43$; les sommes de carrés

$$\begin{aligned}\text{SC}_\text{B} &= 53.821 \\ \text{SC}_\text{C} &= 40.681\end{aligned}$$

et les intervalles de confiance s'obtiennent de même :

$$14,67 \pm 2,262\sqrt{5,98/10} = [12,92\,;\,16,42] \qquad \text{(hôpital B)}$$
$$14,43 \pm 2,262\sqrt{4,52/10} = [12,91\,;\,15,95] \qquad \text{(hôpital C)}$$

3. Si on suppose que la variance est la même dans les trois hôpitaux, on estime la variance commune grâce à la somme des carrés résiduelle :

$$\begin{aligned}\text{SCR} &= \text{SC}_\text{A} + \text{SC}_\text{B} + \text{SC}_\text{C} \\ &= 152,378\end{aligned}$$

et $\widehat{\sigma} = \frac{1}{27}\text{SCR} = 5,644$.

Les intervalles de confiance sont maintenant basés sur un quantile de loi t à 27 ddl, $t^{27}_{0,975} = 2,052$:

$$15,28 \pm 2,052\sqrt{5,644/10} = [13,74\,;\,16,82] \qquad \text{(hôpital A)}$$
$$14,67 \pm 2,052\sqrt{5,644/10} = [13,13\,;\,16,21] \qquad \text{(hôpital B)}$$
$$14,43 \pm 2,052\sqrt{5,644/10} = [12,89\,;\,15,97] \qquad \text{(hôpital C)}$$

4. On réalise une anova. La somme des carrés totaux est

$$\begin{aligned}\text{SCT} &= (2392,66 + 2205,91 + 2122,03) - \frac{1}{30}(152,8 + 146,7 + 144,3)^2 \\ &= 156,2187\end{aligned}$$

On peut établir la table de Fisher

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
facteur	3,841	2	1,920	F = 0,34
résidus	152,378	27	5,644	
Total	156,219	29		

La différence n'est pas significative.

A.13 Exercices du chapitre 13

Corrigé de l'exercice 1 Le plan 1 est équilibré : les effectifs dans les lignes sont en proportion 1 : 1 : 1, et dans les colonnes 2 : 3 : 4.

Le plan 2 ne l'est pas : les effectifs de la première colonne sont en proportion 2 : 4 : 3, ceux de la seconde en proportion 1 : 2 : 4.

Le plan 3 est équilibré : les effectifs dans les lignes sont en proportion 1 : 2 : 3, et dans les colonnes 1 : 5.

Corrigé de l'exercice 2 On commence par le calcul des $SC_{ij} = x_{ij+}^2 - \frac{1}{n_{ij}}(x_{ij+})^2$.

	Traitement A	Traitement B	Traitement C
Hommes	$SC_{11} = 17{,}58$	$SC_{12} = 9{,}63$	$SC_{13} = 20{,}79$
Femmes	$SC_{21} = 11{,}79$	$SC_{22} = 21{,}23$	$SC_{23} = 25{,}96$

Par exemple : $SC_{11} = x_{11+}^2 - \frac{1}{n_{11}}(x_{11+})^2 = 733{,}14 - \frac{1}{4}53{,}5^2 = 17{,}58$.
On a $SCR = \sum_{ij} SC_{ij} = 106{,}98$
On calcule ensuite

$$\begin{aligned} SCF_A &= \frac{1}{n_{1+}}(x_{1++})^2 + \frac{1}{n_{2+}}(x_{2++})^2 - \frac{1}{n}(x_{+++})^2 \\ &= \frac{1}{14}189{,}3^2 + \frac{1}{21}324{,}2^2 - \frac{1}{35}513{,}5^2 \\ &= 30{,}86 \end{aligned}$$

et de même $SCF_B = 6{,}38$.

On a $SCF_{AB} = SCT - SCF_A - SCF_B - SCR = 13{,}04$.

On peut remplir la table d'anova :

Source	Somme des carrés	degrés de liberté	Carrés moyens	F
Sexe	30,86	1	30,86	F = 8,38
Traitement	6,38	2	3,19	F = 0,86
Interaction	13,04	2	6,52	F = 1,77
Résidus	106,98	29	3,68	
Total	157,26	34		

Après comparaison aux valeurs critiques, on constate que seul l'effet du sexe est significatif.

Corrigé de l'exercice 3 1. On complète le tableau de sommes de carrés par $SC_{11} = 0{,}02$ et $SC_{12} = 2{,}587$. On en déduit $SCR = 36{,}52$. On calcule d'autre part $SCT = 71{,}29$, $SCF_{Sexe} = 4{,}73$, $SCF_{Traitement} = 18{,}32$, et enfin $SCF_{Interaction} = 11{,}72$. D'où la table de Fisher :

Sources	SC	ddl	CM	F
Sexe	4,73	1	4,73	3,89
Traitement	18,32	2	9,16	7,52
Interaction	11,72	2	5,86	4,81
Residus	36,52	30	1,22	
Total	71,29	35	2,04	

On compare les statistiques F aux quantiles $F^{1,30}_{0,95} = 4{,}171$ et $F^{2,30}_{0,95} = 3{,}316$: l'effet du sexe n'est pas significatif, l'effet du traitement est mis en évidence ainsi que la présence d'une interaction entre sexe et traitement.

2. La valeur moyenne de l'efficacité dans ce groupe est $\frac{30{,}9}{6} = 5{,}15$. La meilleure estimation de la variance est donnée par le carré moyen résiduel, CMR = 1,22 avec 30 degrés de liberté ; on utilise donc le quantile de loi t : $t^{30}_{0,975} = 2{,}0423$, pour calculer l'intervalle de confiance [4,23;6,07].

Corrigé de l'exercice 4 1. Le plan d'expérience est bien équilibré : les colonnes sont toutes au ratio 2 :3.

2. On complète le tableau des sommes de carrés, par exemple : $SC_{12} = x^2_{12+} - \frac{1}{n_{12}}(x_{12+})^2 = 211{,}64 - \frac{1}{6}(35{,}4)^2 = 2{,}78$.

	Traitement A	Traitement B	Traitement C
Hommes	$SC_{11} = 2{,}528$	$SC_{12} = 2{,}78$	$SC_{13} = 2{,}828$
Femmes	$SC_{21} = 4{,}228$	$SC_{22} = 8{,}676$	$SC_{23} = 6{,}16$

On calcule $SCR = \sum_{ij} SC_{ij} = 27{,}2$.

On calcule aussi $SCT = x^2_{+++} - \frac{1}{n}(x_{+++})^2 = 1780{,}67 - \frac{1}{40}(259{,}3)^2 = 99{,}76$.

Les sommes de carrés factoriels sont

$$SCF_{\text{Trait,}} = \sum_j \frac{1}{n_{+j}}(X_{+j+})^2 - \frac{1}{n}(x_{+++})^2 = \frac{1}{10}(44{,}2)^2 + \frac{1}{15}(99{,}5)^2 + \frac{1}{15}(115{,}6)^2 - \frac{1}{40}(259{,}3)^2 = 65{,}36$$

et

$$SCF_{\text{Sexe}} = \sum_j \frac{1}{n_{i+}}(X_{i++})^2 - \frac{1}{n}(x_{+++})^2 = \frac{1}{16}(102{,}4)^2 + \frac{1}{24}(156{,}9)^2 - \frac{1}{40}(259{,}3)^2 = 0{,}18.$$

On réalise la table de Fisher.

Sources	SC	ddl	CM	F
Sexe	0,18	1	0,18	0,23
Traitement	65,36	2	32,68	40,85
Interaction	7,02	2	3,51	4,39
Residus	27,20	34	0,80	
Total	99,76	39	2,56	

Les quantiles auxquels comparer ces valeurs sont $F^{1,34}_{0,95} = 4{,}1$ et $F^{2,34}_{0,95} = 3{,}3$.

On a donc un effet significatif du traitement, pas d'effet significatif du sexe mais une interaction entre le sexe et le traitement.

3. Chez les hommes : $\widehat{\mu}_{12} = \frac{1}{6}35{,}4 = 5{,}90$ et $\widehat{\mu}_{13} = \frac{1}{6}47{,}5 = 7{,}92$, soit une différence estimée à 7,92 − 5,90 = 2,02, avec un écart-type $\sqrt{\sigma^2\left(\frac{1}{n_{12}} + \frac{1}{n_{13}}\right)}$; la variance résiduelle σ^2 étant estimée par le carré

moyen résiduel CMR = 0,80 avec 34 degrés de liberté. Pour finir on a un intervalle de confiance

$$2{,}02 \pm t^{34}_{0{,}975}\sqrt{\frac{0{,}80}{3}}$$
$$2{,}02 \pm 2{,}0322 \times 0{,}516$$

$$[0{,}97; 3{,}07]$$

De même chez les femmes la différence est estimée à $\frac{1}{9}68{,}1 - \frac{1}{9}64{,}1 = 0{,}44$, l'écart-type est estimé par $\sqrt{0{,}80 \times \frac{2}{9}} = 0{,}422$ et on obtient l'intervalle de confiance

$$[-0{,}41; 1{,}31]$$