ECUE21.1 Science des données (DATA)

Chloé-Agathe Azencott (CBIO)

Printemps 2020 – Mines ParisTech

Compétences

C1	Maîtriser des méthodes statistiques usuelles permettant de traiter convenablement des cas simples d'analyse de données
C2	Maîtriser des méthodes usuelles d'exploration des données
C3	Connaître les limites d'applications des méthodes vues en cours
C4	Pouvoir se référer à un cas d'application avec des données réelles en lien avec une discipline autre que celle de l'analyse des données
C5	Savoir évaluer la complexité numérique de quelques algorithmes
C6	Connaître des méthodes d'apprentissage statistique (machine learning) supervisé et des méthodes d'apprentissage statistique non supervisé
C7	Savoir valider et sélectionner un modèle statistique

Chapitre 1 Introduction

Notions : statistique descriptive, statistique inférentielle, apprentissage statistique, population

Objectifs pédagogiques :

— Donner une définition de la science des données

— Donner une définition de la statistique

— Donner une définition de l'apprentissage statistique, ou apprentissage automatique, ou encore *machine learning*

1.1 Qu'est-ce que la science des données ?

La science des données, ou *data science*, est un domaine dont la définition dépend des personnes qui la donnent. On s'accorde néanmoins généralement sur l'idée qu'il s'agit d'une science interdisciplinaire, qui s'appuie sur les mathématiques (et notamment les probabilités, la statistique et l'optimisation) et l'informatique (et notamment l'algorithmique, les bases de données, l'architecture distribuée, et l'analyse numérique) mais aussi sur des connaissances spécifiques au domaine d'application, autrement dit à la nature des données étudiées (finance, commerce, physique, biologie, sociologie, etc.).

C'est un domaine multiforme qui fait beaucoup parler de lui, et on se réfère souvent à un article du *Harvard Business Review* intitulé « *Data Scientist : the Sexiest Job of the 21st Century* » [1].

La science des données permet par exemple de mieux comprendre les besoins de la clientèle d'une entreprise ; de dimensionner des serveurs ; d'améliorer la distribution de l'électricité ; d'analyser des données génomiques pour suggérer de nouvelles hypothèses biologiques ; d'optimiser la livraison de colis ; de détecter des fraudes ; de recommander des livres, films, ou autres produits adaptés à nos goûts ; ou de personnaliser des publicités.

Dans les années à venir, la science des données, et en particulier le *machine learning*, nous permettra vraisemblablement d'améliorer la sécurité routière (y compris grâce aux véhicules autonomes), la réponse d'urgence aux catastrophes naturelles, le développement de nouveaux médicaments, ou l'efficacité énergétique de nos bâtiments et industries.

1.2 Objectifs de ce cours

Ce cours se concentre sur les aspects mathématiques (statistique et modélisation) et informatiques (utilisation pratique) de la science des données. Il fait appel à des notions que vous avez découvertes en Probabilités, en Optimisation, et en Outils Numériques pour les Mathématiques.

Le premier but de ce cours est de démystifier la science des données, le Big Data, l'intelligence artificielle telle qu'on en parle de nos jours et de vous donner les clés nécessaires à recevoir les informations sur le sujet d'un œil critique.

Le deuxième but de ce cours est de poser les bases mathématiques et algorithmiques de l'exploitation de données. Les domaines de la statistique inférentielle et de l'apprentissage automatique sont vastes, et vous aurez, si vous le souhaitez, amplement l'occasion de les explorer en deuxième et troisième année.

1. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Socle minimal À l'issue de ce cours, vous devriez avoir acquis *a minima* les compétences suivantes :

1. Interpréter une p-valeur ;
2. Éviter les principaux écueils de visualisation de données ;
3. Expliciter le principe de la minimisation du risque empirique, et l'illustrer sur un ou deux exemples d'algorithme d'apprentissage automatique ;
4. Expliquer comment un algorithme d'apprentissage automatique peut reproduire un biais ;
5. Reconnaître une situation pouvant prêter au surapprentissage ;
6. Sélectionner et valider un modèle d'apprentissage automatique ;
7. Décrire un réseau de neurones artificiel comme un modèle paramétrique ;
8. Lister quelques succès et limites de l'apprentissage profond ;
9. Donner des exemples d'algorithmes d'apprentissage automatique permettant d'apprendre des modèles non linéaires ;
10. Expliquer quelques uns des enjeux éthiques liés à l'intelligence artificielle.

Les sections marquées d'une étoile (★) dans le poly sont celles qui ne font pas partie de ce socle minimal. Ne pensez pas cependant qu'elles soient plus difficiles !

Acquisition de données Aucune approche statistique ne pourra créer un bon modèle à partir de données qui ne sont pas pertinentes – c'est le concept *garbage in, garbage out* qui stipule qu'un algorithme d'apprentissage auquel on fournit des données de mauvaise qualité ne pourra rien en faire d'autre que des prédictions de mauvaise qualité.

Bien que ce cours soit consacré aux outils mathématiques et, dans une moindre mesure, informatiques de la science des données, il ne faut pas négliger qu'une part importante du de *machine learner* ou de *data scientist* est un travail d'ingénierie consistant à préparer les données afin d'éliminer les données aberrantes, gérer les données manquantes, choisir une représentation pertinente, etc.

Big data De plus en plus, les quantités de données disponibles imposent de transformer les algorithmes utilisés et de faire appel à des architectures de calcul et de base de données distribuées. Nous n'abordons pas non plus ce point dans ce cours. Le cours optionnel Large Scale Machine Learning, proposé en semaine bloquée au printemps, vous permettra de découvrir ce domaine.

1.3 Qu'est-ce que la statistique ?

Le terme de « statistique » est dérivé du latin « *status* » (signifiant « état »). Historiquement, **les statistiques** concernent l'étude méthodique, par des procédés numériques (inventaires, recensements, etc.) des faits sociaux qui définissent un état.

Par contraste, **la statistique** est un ensemble de méthodes des mathématiques appliquées permettant de décrire et d'analyser des phénomènes dont la nature rend une étude exhaustive de tous leurs facteurs impossible. Ces méthodes permettent d'étudier des données, ou observations, consistant en la mesure d'une ou plusieurs caractéristiques d'un ensemble de personnes ou objets équivalents.

1.3.1 Vocabulaire

L'ensemble de personnes ou d'objets équivalents étudiés est appelé **la population.** Il peut s'agir d'une population au sens « courant » du terme (par exemple, l'ensemble de la population française, ou l'ensemble des individus d'une espèce animale sur un territoire) mais aussi plus largement d'un ensemble plus générique d'objets que l'on cherche à étudier (par exemple, l'ensemble des pièces produites par une chaîne de montage, un ensemble de particules en physique, etc.)

Chacun des éléments de la population est appelé **individu.**

Les caractérisiques que l'on mesure pour chacun de ces individus sont appelées les **variables** ; les individus pour lesquels ces caractéristiques ont été mesurées sont appelées des **observations**. Un ensemble de n observations $(x_1,x_2,\ldots,x_n)$ d'une variable est appelée **série statistique.**

Par exemple, si j'étudie les données climatiques pour la station météo de Paris-Montsouris en 2019 (cf. table 1.1), il s'agit d'une population de 365 individus. Cette population peut contenir 8 variables : températures minimale, maximale et moyenne ; vitesse maximale du vent ; ensoleillement ; précipitations totales ; pressions atmosphériques minimale et maximale.

Lorsque la population à étudier est trop grande pour qu'il soit possible d'observer chacun de ses individus, on étudie alors une partie seulement de la population. Cette partie est appelée **échantillon**. On parle alors de **sondage**, par opposition à un **recensement**, qui consiste à étudier tous les individus d'une population. Nous reviendrons sur la notion d'échantillon dans le chapitre 3.

Par exemple, la population des élèves de première année des Mines est composée de 125 individus. Si je recueille l'âge, le département de naissance et le nombre de frères et sœurs de 20 de ces élèves, j'aurai mesuré 3 variables sur un échantillon de 20 observations.

On distinguera plusieurs types de variables :

— les **variables quantitatives** : des caractéristiques numériques qui s'expriment naturellement à l'aide de nombres réels. Ces variables peuvent être **discrètes** si le nombre de valeurs qu'elles peuvent prendre est fini ou dénombrable (ex : âge, nombre de frères et sœurs) ou **continues** (ex : températures, taille, pression atmosphérique)

— les **variables qualitatives** : des caractéristiques qui, bien qu'elles puissent être encodées numériquement (ex : département de naissance), relèvent plutôt de catégories et sur lesquelles les opérations arithmétiques de base (somme, moyenne) n'ont aucun sens. On parle de variables **nominales** s'il n'y a pas d'ordre total sur l'ensemble de ces catégories (ex : département de naissance) ou **ordinales** s'il y en a (ex : entièrement d'accord, assez d'accord, pas vraiment d'accord, pas du tout d'accord).

Remarque : seuiller des variables quantitatives permet de les transformer en variables qualitatives ordinales. Par exemple, une variable d'âge peut être transformée en catégories (< 18, 18 – 20, 20 – 35, etc.)

Le tableau 1.2 montre un exemple d'un échantillon de 20 individus d'une population de données de remboursements d'un acte biologique bien précis : le dosage de l'antigène tumoral 125. Ces données sont issues de la base de dépenses de biologie médicale en France mise à disposition par l'Assurance Maladie [2]. La population complète, de 604 individus, est disponible dans le fichier `data/OPEN_BIO_2018_7325.csv`.

Chaque individu de cette population (i.e. ligne du tableau) correspond à un ensemble de dosages et est décrit par 5 variables : la tranche d'âge des patients et patientes ; leur région ; le nombre de dosages ; et enfin les montants remboursés et remboursables. Dans ce tableau, l'âge est une variable qualitative ordinale ; la région une variable qualitative ; et les nombres et montants des variables quantitatives. On pourra choisir de traiter le nombre de remboursements comme une variable discrète ou continue.

1.3.2 Statistique descriptive

La **statistique descriptive,** aussi appelée **statistique exploratoire,** consiste à caractériser une population par la détermination d'un certain nombre de grandeurs qui la décrivent. Son objectif est de

2. http://open-data-assurance-maladie.ameli.fr/biologie/index.php

synthétiser l'information contenue dans un ensemble d'observations et de mettre en évidence des propriétés de cet ensemble. Elle permet aussi de suggérer des hypothèses relatives à la population dont sont issues les observations. Il s'agit principalement de calculer des indicateurs (par exemple des moyennes) et de visualiser les données par des graphiques. La visualisation peut être enrichie par des techniques d'apprentissage non-supervisé (cf section 1.4.2) qui permettent de réduire le nombre de variables ou de regrouper ensemble les individus semblables. La statistique descriptive est traitée au chapitre 2 2.

1.3.3 Statistique inférentielle

Aussi appelée **statistique décisionnaire,** ou encore **inférence statistique,** la **statistique inférentielle** consiste à tirer des conclusions sur une population à partir de l'étude d'un échantillon de celle-ci. Les données observées sont considérées comme un échantillon d'une population. Il s'agit alors d'étendre des propriétés constatées sur l'échantillon à la population. L'inférence statistique repose beaucoup sur les probabilités : on considèrera les observations comme les réalisations de variables aléatoires, ce qui permettra d'approcher les caractéristiques probabilistes de ces variables aléatoires à l'aide d'indicateurs calculés sur l'échantillon. Les chapitres 3 et 4 traîtent de statistique inférentielle.

1.4 Qu'est-ce que l'apprentissage statistique ?

Qu'est-ce qu'apprendre, comment apprend-on, et que cela signifie-t-il pour une machine ? La question de l'*apprentissage* fascine les spécialistes de l'informatique et des mathématiques tout autant que neurologues, pédagogues, philosophes ou artistes.

Dans le cas d'un programme informatique, on parle d'**apprentissage statistique,** ou **apprentissage automatique**, ou encore *machine learning*, quand ce programme a la capacité d'apprendre sans que cette modification ne soit explicitement programmée. Cette définition est celle donnée par Arthur Samuel en 1959. On peut ainsi opposer un programme *classique*, qui utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme *d'apprentissage automatique*, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

Supposons par exemple qu'une entreprise veuille connaître le montant total dépensé par un client ou une cliente à partir de ses factures. Il suffit d'appliquer un algorithme classique, à savoir une simple addition : un algorithme d'apprentissage n'est pas nécessaire.

Par contraste, supposons maintenant que l'on veuille utiliser ces factures pour déterminer quels produits le client est le plus susceptible d'acheter dans un mois. Bien que cela soit vraisemblablement lié, nous n'avons manifestement pas toutes les informations nécessaires pour ce faire. Cependant, si nous disposons de l'historique d'achat d'un grand nombre d'individus, il devient possible d'utiliser un algorithme d'apprentissage automatique pour qu'il en tire un modèle prédictif nous permettant d'apporter une réponse à notre question.

Ce point de vue informatique sur l'apprentissage automatique justifie que l'on considère qu'il s'agisse d'un domaine différent de celui de la statistique. Cependant, nous aurons l'occasion de voir que la frontière entre inférence statistique et apprentissage est souvent mince. Il s'agit ici, fondamentalement, de **modéliser** un phénomène à partir de données considérées comme autant d'observations de celui-ci.

Attention

Bien que l'usage soit souvent d'appeler les deux du même nom, il faut distinguer l'**algorithme d'apprentissage** automatique du **modèle appris** : le premier utilise les données pour produire le second, qui peut ensuite être appliqué comme un programme classique.

On distingue plusieurs types de problèmes en apprentissage automatique. Nous nous intéresserons dans ce cours à l'apprentissage supervisé et à l'apprentissage non-supervisé, en ignorant entre autres l'apprentissage par renforcement principalement utilisé en robotique.

1.4.1 Apprentissage supervisé

L'**apprentissage supervisé,** ou **apprentissage prédictif,** est peut-être le type de problèmes de machine learning le plus facile à appréhender : son but est d'apprendre à faire des *prédictions*, à partir d'une liste d'exemples **étiquetés**, c'est-à-dire accompagnés de la valeur à prédire. Les étiquettes servent de « prof » et supervisent l'apprentissage de l'algorithme. Un exemple classique est celui de l'annotation d'images : il s'agit par exemple de déterminer si une image représente ou non un chat.

Étant données n observations $\{x_i\}_{i=1,\dots,n}$ décrites dans un espace $\mathcal{X}$, et leurs étiquettes $\{y_i\}_{i=1,\dots,n}$ décrites dans un espace $\mathcal{Y}$, on suppose que les étiquettes peuvent être obtenues à partir des observations grâce à une fonction $\phi : \mathcal{X} \to \mathcal{Y}$ fixe et inconnue : $y_i = \phi(x_i) + \epsilon_i$, où ϵ_i est un bruit aléatoire. Il s'agit alors d'utiliser les données pour déterminer une fonction $f : \mathcal{X} \to \mathcal{Y}$ telle que, pour tout couple $(x,\phi(x)) \in \mathcal{X} \times \mathcal{Y}$, $f(x) \approx \phi(x)$. On suppose généralement pour cela que les couples (x_i,y_i) sont les réalisations d'un vecteur aléatoire (X,Y) vérifiant $Y = \phi(X) + \epsilon$ et l'on cherche à déterminer ϕ.

Nous aborderons en détail l'apprentissage supervisé dans les chapitres 7 à 9.

1.4.2 Apprentissage non supervisé

Dans le cadre de l'**apprentissage non supervisé**, les données ne sont pas étiquetées. Il s'agit alors de modéliser les observations pour mieux les comprendre. Ces techniques sont ainsi complémentaires de celles de la statistique exploratoire.

Parmi les exemples d'apprentissage non supervisé, on compte notamment

— la **réduction de dimension**, que nous aborderons au chapitre 5), qui permet de créer un petit nombre de variables qui résument les mesures prises sur les observations. Il s'agit de trouver une représentation des données dans un espace de dimension plus faible que celle de l'espace dans lequel elles sont représentées à l'origine. Cela permet de réduire les temps de calcul et l'espace mémoire nécessaire au stockage les données, mais aussi souvent d'améliorer les performances d'un algorithme d'apprentissage supervisé entraîné par la suite sur ces données.

— le **partitionnement**, ou **clustering**, qui permet de réduire la taille d'un échantillon en regroupant les individus présentant des caractéristiques homogènes. Nous ne traiterons pas de ce sujet dans ce cours. Il sera notamment abordé dans le cours optionnel d'apprentissage automatique proposé en semaine bloquée à l'automne.

1.5 Et l'intelligence artificielle, dans tout ça ?

Le machine learning est une branche de l'intelligence artificielle. En effet, un système incapable d'apprendre peut difficilement être considéré comme intelligent. L'intelligence artificielle, définie comme l'ensemble des techniques mises en œuvre afin de construire des machines capables de faire preuve d'un comportement que l'on peut qualifier d'intelligent, fait aussi appel aux sciences cognitives, à la neurobiologie, à la logique, à l'électronique, à l'ingénierie et bien plus encore.

Le terme d'« intelligence artificielle » stimulant plus l'imagination, il est de plus en plus souvent employé en lieu et place de celui d'apprentissage automatique.

1.6 Bonnes pratiques

L'essor récent de l'intelligence artificielle, à travers notamment les développements en apprentissage profond que nous aborderons brièvement au chapitre 9, suscite de vifs débats philosophiques, éthiques et moraux dans notre société. Sans entrer en profondeur dans ces débats, ce qui relèverait d'un cours d'éthique ou de philosophie et non plus d'un cours de mathématiques appliquées, nous en aborderons quelques points clés dans le chapitre 6, dédiée aux bonnes pratiques en science des données. En science des données, il serait malhonnête de prétendre pouvoir détacher les mathématiques et l'informatique du contexte de leur utilisation.

1.7 Sources

Le contenu de ce poly s'appuie en partie sur des documents mis à disposition en ligne par Stéphane Canu, Laure Reboul, et Joseph Salmon, que je remercie vivement, ainsi que les ouvrages *Probabilités, analyse des données et Statistique* (Technip) de Gilbert Saporta et *Introduction au Machine Learning* (Dunod InfoSup) de Chloé-Agathe Azencott.

Pour aller plus loin

- L'article *50 Years of Data Science* de David Donaho aborde les différences entre statistique, science des données, et apprentissage automatique et donne une vision d'ensemble de ces domaines.

T min °C	T max °C	T moy °C	Vent km/h	Ensoleillement min	Précipitations mm	P min hPa	P max hPa
7.6	9.6	8.4	22.2	0	0	1034	1036.6
5.6	7.2	6.3	24.1	0	0	1037.3	1041.3
4.1	6.6	5.4	16.7	0	0	1040.2	1041.8
3.1	6	4.7	20.4	0	0	1039.5	1041.7
4.2	5.9	5	20.4	0	0	1037.5	1039.6
4.3	6.8	5.6	16.7	0	0	1036.5	1038
6.8	8.6	7.4	20.4	0	0.6	1030.5	1037.2
7.4	9.7	8.5	24.1	120	0	1025.9	1029.7
4	7.7	5.2	29.6	42	0.8	1024.1	1026.4
2.1	5.5	4	18.5	30	0	1026.6	1029.5
4.2	8.3	6.2	14.8	0	1.2	1028.1	1030.5
6.7	9.1	8	22.2	0	0.8	1021.7	1030.6
8.8	11.9	10.5	31.5	30	0.8	1014	1021.1
8.5	10.9	8.8	29.6	0	0	1014	1024.8
6.9	8.6	7.6	16.7	0	0	1020.3	1025.3
1.9	7.8	5.2	27.8	276	3	1007.9	1019.6
4	8.5	5.4	27.8	0	0	1007.5	1019.7
0.9	6.1	2.4	18.5	342	0	1017.2	1021
-1.7	2.8	0.3	14.8	78	4	1009.7	1016.8
1.9	3	2.5	20.4	0	0.8	1010.1	1021.8
-2.2	3.6	0.1	13	480	0	1019.2	1024.9
-2.4	1.7	-0.1	20.4	0	7.4	998.4	1018.3
0.6	2.1	1.3	24.1	6	1	995.6	1007.4
-0.4	2.5	1.2	20.4	0	1	1008.4	1015.5
1.7	5.5	3.9	13	0	1.2	1015.2	1018.4
5.5	9.6	8.4	29.6	24	1.6	998.1	1016.9
4.4	8.2	6.4	37	6	4.4	989.8	1001.4
2.7	6.9	4.4	25.9	216	0	1001.7	1011.7
-0.8	5.2	1.7	24.1	18	19.6	989.5	1011.7
0.5	5.2	2.3	33.3	252	0.4	990	998.9
-1	2.5	1.3	25.9	24	1.2	983.5	998

TABLEAU 1.1 – Exemple de 8 variables pour 31 observations (celles du mois de janvier 2019) de la population des données climatiques pour la station météo de Paris-Montsouris. Ces données sont disponibles dans le fichier `data/meteo_data.csv`.

Âge ans	Région	Nombre d'actes	Montants remboursés (€)	Montants remboursables (€)
> 60	76	26	377,96	402,80
> 60	75	1 401	14 054,37	21 332,15
> 60	44	5 299	65 928,93	80 447,00
> 60	32	1 706	25 137,65	26 032,65
> 60	32	2 596	37 877,02	39 336,15
> 60	27	14	159,85	211,35
> 60	24	3 565	50 770,46	54 076,15
> 60	11	396	5 226,55	6 060,05
> 60	5	260	4 496,91	4 676,40
> 60	93	162	2 303,56	2 466,10
> 60	76	578	8 499,53	8 793,10
40-59	76	13	172,26	199,80
40-59	44	102	1 204,93	1 557,20
40-59	11	48	555,39	733,05
40-59	84	14	190,21	217,85
40-59	32	126	1 350,06	1 912,15
20–39	32	749	7 941,69	11 362,40
20–39	32	24	289,35	365,25
20–39	5	918	9 704,10	16 550,40
20–39	11	106	1 073,32	1 618,35

TABLEAU 1.2 – Population de remboursements du dosage de l'antigène 125 dans le sang en 2018, composée de 20 individus décrits par 5 variables et extraite du fichier `data/OPEN_BIO_2018_7325.csv`. Région : 5 = Régions et Départements d'outre-mer. 11 = Ile-de-France. 24 = Centre-Val de Loire. 27 = Bourgogne-Franche-Comté. 32 = Hauts-de-France. 44 = Grand-Est. 75 = Nouvelle-Aquitaine. 76 = Occitanie. 84 = Auvergne-Rhône-Alpes. 93 = Provence-Alpes-Côte d'Azur et Corse.

Première partie

Notions de statistique

Chapitre 2 Statistique descriptive

Notions : individu, population, fréquences, indicateurs de tendance centrale, indicateurs de liaison, table de contingence.

Objectif pédagogique : Caractériser une variable statistique, ou la relation entre deux variables statistiques, à travers des représentations graphiques et le calcul d'indicateurs numériques.

Le rôle de la statistique descriptive est de caractériser une population par la détermination d'un certain nombre de grandeurs qui la décrivent. Ce chapitre présente quelques unes des visualisations et des indicateurs numériques les plus fréquemments utilisés pour décrire une unique variable statistique, ou la relation entre deux variables statistiques.

Il s'agit ici uniquement de *décrire* les données. La statistique descriptive ne nous permet pas de faire de l'*inférerence*, c'est-à-dire de tirer des conclusions sur ces données. Elle nous permet par contre de faire des hypothèses, comme par exemple :

— Telle variable semble suivre une distribution uniforme sur un intervalle ;

— Telle variable semble dépendre de telle autre ;

— Telle variable semble prendre une valeur plus élevée dans un segment de la population que dans un autre.

Exercice : En découvrant les exemples de ce chapitre, demandez-vous quelles hypothèses les valeurs d'indicateurs et les visualisations graphiques vous suggèrent. Quand vous rencontrez des indicateurs numériques ou des visualisations de données dans d'autres matières ou projets, ou dans les media d'information, demandez-vous dans quelle section de ce chapitre elles entrent ; quelle est la taille de la population et/ou de l'échantillon ; quelle est la nature des variables mesurées ; quelles hypothèses elles vous permettent de formuler.

2.1 Statistique descriptive unidimensionelle

Il s'agit ici de mettre en évidence les principales caractéristiques d'une unique variable statistique x observée sur n individus, à travers la série statistique $(x_1, x_2, \ldots, x_n)$.

2.1.1 Fréquences

L'étude d'une série statistique passe par la construction d'une **table des fréquences,** soit des valeurs elles-mêmes dans le cas d'une variable qualitative ou discrète, soit d'intervalles de ces valeurs. La construction de ces intervalles permet de transformer la série statistique en **série classée.**

Exemple

Prenons l'exemple des données de remboursement dans `data/OPEN_BIO_2018_7325.csv` dont sont extraits les 20 individus du tableau 1.2. La **table des fréquences** des âges est donnée dans le tableau ci-dessous :

Tranche d'âge (ans)	0 – 19	20 – 39	40 – 59	>60
Fréquence	7%	18%	31%	43%

Pour une variable qualitative, la table de fréquences peut aussi être visualisée grâce à un **diagramme en bâtons,** comme illustré sur la figure 2.1.

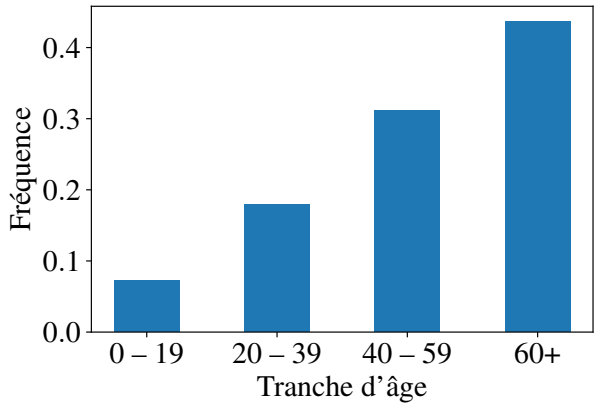


FIGURE 2.1 – Diagramme en bâtons de la fréquence des tranches d'âges dans les données de remboursement.

Dans le cas d'une variable continue, la constitution des classes de valeurs d'une série statistique est une étape importante. La **règle de Sturges** propose de découper les valeurs observées en $\lfloor 1+\log_2(n) \rfloor$ intervalles de même taille $\frac{\max(x_i)-\min(x_i)}{k}$. Cependant, cette règle suppose que la variable analysée suive une distribution gaussienne ; elle n'est pas appropriée, par exemple, si les valeurs s'étalent sur plusieurs échelles de grandeur, auquel cas une transformation logarithmique s'imposera.

Exemple

Prenons par exemple, 31 observations de la température minimale (en °C) pour la station météo de Paris-Montsouris, telles que relevées dans la première colonne de la table 1.1.

Nous disposons de $n = 31$ observations, qu'il s'agit, en appliquant la règle de Sturges, de grouper en 5 intervalles d'amplitude 2,24°C. La table des fréquences des températures minimales est donnée dans le tableau ci-dessous :

T min (°C)	< -0,16	-0,16 – 2,08	2,08 – 4,32	4,32 – 6,56	> 6,56
Fréquence	0,19	0,19	0,29	0,10	0,23

Pour une variable continue, la table des fréquences peut être traduite en **histogramme,** comme illustré sur la figure 2.2.

Utiliser des fréquences plutôt que des comptes permet de comparer des populations de taille différente. De plus, la distribution des fréquences d'une série statistique de la variable x, représentée visuellement par un histogramme, peut être considérée comme une approximation de la distribution de la probabilité de cette variable dans la population.

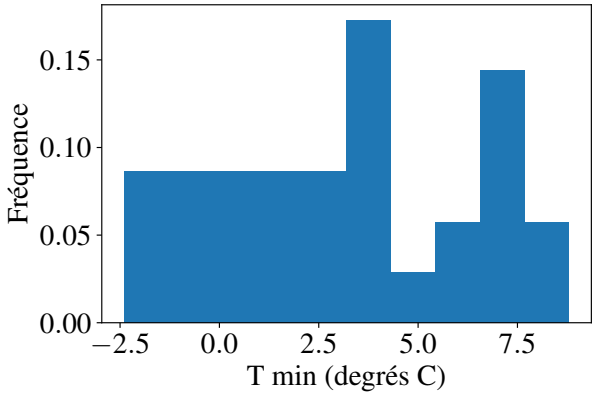


FIGURE 2.2 – Histogramme des températures minimales dans le tableau 1.1.

Fréquences cumulées On peut aussi choisir de représenter plutôt les **fréquences cumulées.**

Exemple

Pour notre série de températures minimales, la table des fréquences cumulées est donnée dans le tableau ci-dessous :

T min (°C)	< -0,16	< 2,08	< 4,32	< 6,56	< 8,80
Fréquence	0,19	0,38	0,67	0,77	1,0
T min (°C)	> -2,40	> -0,16	> 2,08	> 4,32	> 6,56
Fréquence	1,0	0,81	0,62	0,33	0,23

Une table des fréquences cumulées croissantes et décroissantes peut directement être traduite en **courbes des fréquences cumulées,** comme illustré sur la figure 2.3.

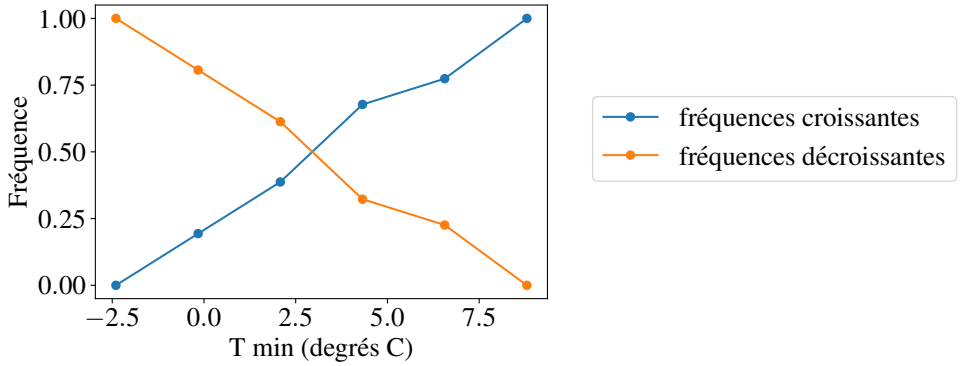


FIGURE 2.3 – Courbes des fréquences cumulées pour les températures minimales du tableau 1.1.

2.1.2 Indicateurs numériques

Enfin, des **indicateurs numériques** permettent de compléter cette description. On distinguera les **indicateurs de tendance centrale** qui indiquent l'ordre de grandeur des valeurs de la série statistique et où ces valeurs se rassemblent, des **indicateurs de dispersion** qui indiquent l'étalement de ces valeurs.

Indicateurs de tendance centrale Les indicateurs de tendance centrale comportent :

— la **moyenne arithmétique**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

La moyenne arithmétique peut être très sensible à la présence de valeurs aberrantes.

— la **médiane,** qui correspond à une fréquence cumulée de 50%,

— le **mode,** qui est la valeur la plus fréquente dans la série statistique. Le mode n'a réellement de sens que pour une variable discrète ; dans le cas d'une variable continue, on parlera plutôt, lorsque la série est classée, de **classe modale** qui est la classe la plus fréquente.

Exemple

Pour notre série de températures minimales,

— la moyenne arithmétique vaut $3{,}2°\mathrm{C}$;

— la médiane vaut $4°\mathrm{C}$;

— la classe modale est $2{,}1$ – $4{,}4°\mathrm{C}$.

Indicateurs de dispersion Les indicateurs de dispersion comportent :

— la **variance de la série statistique**

$$s^2(x_1,x_2,\ldots,x_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

— la **variance d'échantillonage**

$$s^{*2}(x_1,x_2,\ldots,x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

La variance d'échantillonage est d'autant plus proche de la variance que le nombre d'observations est grand. Nous verrons dans la section 3.4.1 qu'il s'agit d'un estimateur **non-biaisé** de la variance de la population.

— l'**écart-type** qui est la racine carrée de la variance,

— le **coefficient de variation**

$$\mathrm{CV}(x_1,x_2,\ldots,x_n) = \frac{s^2(x_1,x_2,\ldots,x_n)}{\bar{x}}$$

Le coefficient de variation permet d'apprécier la variabilité d'une variable en fonction de sa valeur moyenne, et n'a de sens que pour une variable donnée sur une échelle dotée d'un zéro absolu, c'est-à-dire dans laquelle une valeur de $2z$ peut effectivement être considérée comme deux fois plus qu'une valeur de z (ce n'est pas le cas pour une température en degrés Celsius : $10°\mathrm{C}$ n'est pas « deux fois plus chaud » que $5\ °\mathrm{C}$). De plus, il est numériquement instable quand $\bar{x}$ est proche de 0.

Exemple

La variance de notre série de températures minimales vaut $10{,}02°C^2$, tandis que la variance d'échantillonage vaut $10{,}36°C^2$. Les écarts-types correspondants valent tous les deux $3{,}2°C$. Le coefficient de variation n'a pas de sens en degrés Celsius.

Remarques

— L'écart-type d'une variable, qui s'exprime dans la même unité que la variable, est beaucoup plus facile à interpréter que la variance. On donne plus facilement un sens à $3{,}2°C$ qu'à $10{,}02°C^2$.

— L'écart-type est utilisé pour définir une erreur de mesure. Imaginons que l'on prenne 10 fois la même mesure, obtenant ainsi une population de 10 mesures, de moyenne arithmétique m et d'écart-type σ ; on rapporte alors une valeur de $m \pm \sigma$. Cette remarque est une brève incursion dans le domaine de la *métrologie*.

Enfin, les **quantiles** permettent aussi de déterminer la dispersion d'une variable. Les q-quantiles divisent les valeurs prises par la variable en q intervalles de mêmes fréquences. Le p-ème q-quantile de $(x_1,x_2,\ldots,x_n)$ est défini comme la valeur Q_p^q telle que

$$\text{Freq}(x \leq Q_p) = \frac{p}{q}.$$

Lorsque $q = 4$, on parle de **quartiles.** Lorsque $q = 10$, on parle de **déciles.**

Exemple

Les trois quartiles de notre série de températures minimales sont $0{,}8°C$, $4{,}0°C$ et $5{,}6°C$: 25% des valeurs observées sont inférieures à $0{,}8°C$, 50% sont inférieures à 4,0 $°C$ et 75% sont inférieures à $5{,}6°C$. Le deuxième quartile correspond bien à la médiane.

Une **boîte à moustaches** (ou *boxplot*) permet de résumer visuellement ces indicateurs, comme illustré sur la figure 2.4. La boîte à moustaches est composée d'un rectangle, d'une largeur arbitraire et délimité en bas par la valeur du premier quartile et en haut par la valeur du troisième quartile ; d'une barre horizontale au niveau de la médiane ; et de deux segments joignant chacun les extrémités du rectangle aux valeurs les plus extrêmes. Représenter les valeurs prises par la variable par un nuage de points superposé à ce rectangle permet d'en faciliter l'interprétation.

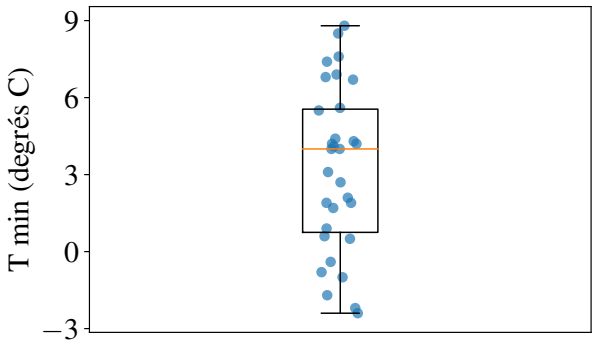


FIGURE 2.4 – Boîte à moustaches des températures minimales du tableau 1.1.

2.2 Statistique descriptive bidimensionelle

Il s'agit ici de mettre en évidence une éventuelle **liaison,** c'est-à-dire une variabilité simultanée, entre deux variables statistiques x et y, observées sur n individus, à travers les séries statistiques $(x_1,x_2,\ldots,x_n)$ et $(y_1,y_2,\ldots,y_n)$.

Cette liaison peut être causale ou non. Mettre en évidence une causalité est délicat, et dépasse le cadre de ce cours.

Comprendre la liaison entre deux variables nous permet de comprendre

— Si une variable peut dépendre d'une autre : la température minimale dépend-elle de l'ensoleillement ?

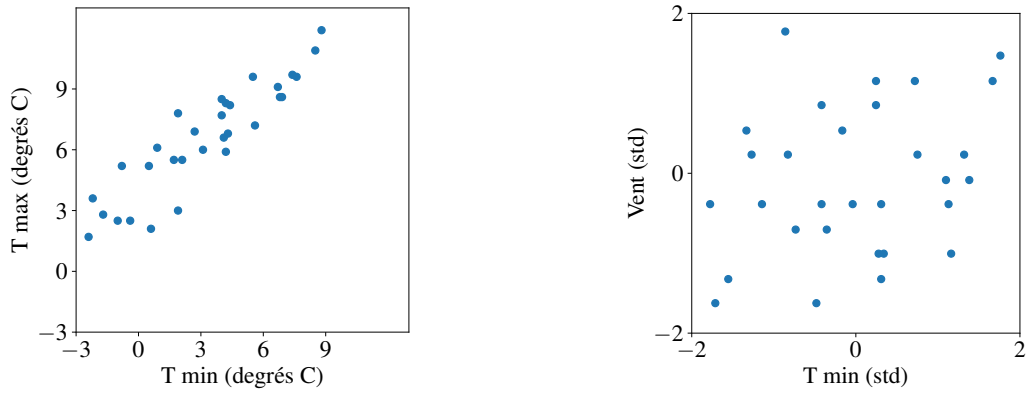
— Si une variable peut permettre de prédire une autre : la température minimale permet-elle de prédire la température maximale ?

— Si une variable peut être remplacée par une autre : ai-je besoin de prendre en compte et la température minimale et la température maximale, ou la température moyenne suffit-elle ?

2.2.1 Liaison entre deux variables quantitatives

Nuage de points Pour visualiser la liaison entre deux variables quantitatives, on utilise généralement un **nuage de points**. Si x et y sont homogènes, c'est-à-dire exprimées dans la même unité, on utilisera la même échelle sur les deux axes, comme sur la figure 2.5a. Sinon, on préfèrera généralement centrer et réduire les variables au préalable, comme sur la figure 2.5b :

$$x_i \leftarrow \frac{x_i - \bar{x}}{\sigma_x} \qquad \text{avec } \bar{x} = \sum_{i=1}^{n} x_i \text{ et } \sigma_x = \sqrt{\sum_{i-1}^{n} (x_i - \bar{x})^2}.$$



(A) Températures maximales vs minimales. (B) Vent vs températures minimales.

FIGURE 2.5 – Nuages de points pour des paires de variables du tableau 1.1.

Indicateurs de liaison entre deux variables quantitatives Pour quantifier la liaison entre deux variables quantitatives, on utilise principalement

— **la covariance**

$$s(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

— **le coefficient de corrélation de Pearson,** qui est égal à la covariance entre les variables centrées réduites, et compris entre -1 et 1 :

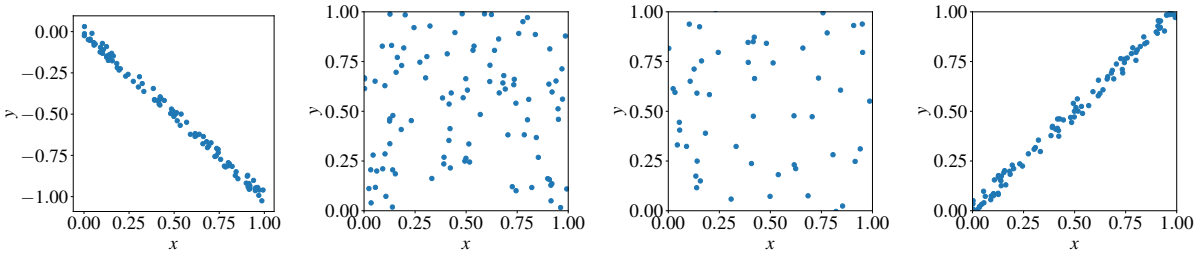
$$r(x,y) = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x, \sigma_y.}$$

À noter que la covariance et le coefficient de corrélation de Pearson mesurent des liaisons *linéaires* entre deux variables. Une corrélation de Pearson proche de 1 ou de -1 indique une relation linéaire ;

une corrélation de Pearson proche de 0 indique une absence de corrélation. D'autres mesures, comme l'information mutuelle (hors cadre de ce cours), permettent de mesurer des liaisons *non-linéaires*.

Exemple

Pour les données du tableau 1.1, la covariance entre la température minimale et la température maximale vaut $7{,}69°C^2$; leur corrélation de Pearson vaut 0,91. La corrélation de Pearson entre vent et température minimale vaut 0,28. La figure 2.6 illustre le rapport entre corrélation de Pearson et nuage de points.



(A) r(x, y) = -1,00. (B) r(x, y) = 0,03. (C) r(x, y) = 0,53. (D) r(x, y) = 1,00.

FIGURE 2.6 – Nuages de points entre deux variables simulées et leur corrélation de Pearson.

Indicateurs de liaison entre une variable qualitative et une variable quantitative Pour étudier la liaison entre une variable qualitative x, ayant K modes (ou valeurs différentes) dans la série statistique $(x_1,x_2,\ldots,x_n)$, et une variable quantitative y, on considère que la variable x permet de définir p sous-populations. Il s'agit alors d'évaluer s'il existe des différences, pour la variable y, entre ces sous-populations.

Visuellement, on utilisera une série de boîtes à moustaches, comme illustré sur la figure 2.7.

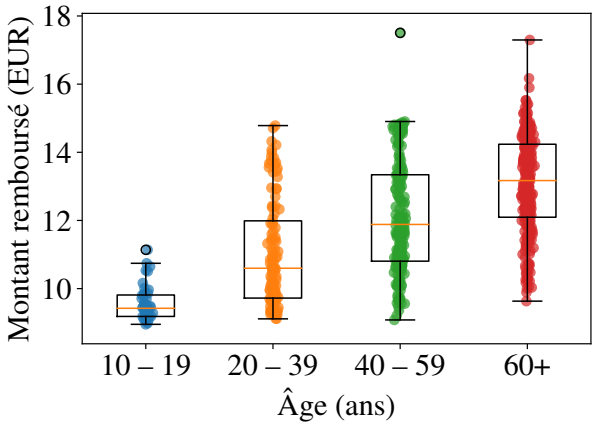


FIGURE 2.7 – Montants remboursés par acte, par tranche d'âge, pour les données de remboursement.

La **variance expliquée** par x de y la moyenne des carrés des écarts entre la moyenne de y dans chaque sous-population et la moyenne de y dans toute la population, pondérée par la taille des sous-populations :

$$\sigma_E^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2,$$

où $\bar{y}_k$ est la moyenne de y dans la sous-population k et $\bar{y}$ la moyenne de y dans la population totale.

La **variance résiduelle** est la moyenne des variances des sous-populations, pondérées par leur taille :

$$\sigma_R^2 = \frac{1}{n}\sum_{k=1}^{K} n_k \sigma_k^2,$$

où n_k est le nombre d'individus dans la sous-population k et σ_k^2 est la variance de y dans cette sous-population.

On peut montrer que $\sigma_y^2 = \sigma_R^2 + \sigma_E^2$.

Le **rapport de corrélation** est la part de variation de y expliquée par x. Compris entre 0 et 1, il est d'autant plus élevé que la liaison entre les deux variables est forte :

$$e^2 = \frac{\sigma_E^2}{\sigma_y^2}.$$

Exemple

Pour les montants remboursés par acte de nos données de remboursement, la variance de ces montants est de $3{,}30^2$, tandis que la variance expliquée par l'âge est de $1{,}09^2$, ce qui donne un rapport de corrélation de $0{,}33$.

Indicateurs de liaison entre deux variables qualitatives Pour étudier la liaison entre une variable qualitative x, ayant K modes (ou valeurs différentes) dans la série statistique $(x_1, x_2, \ldots, x_n)$, et une variable qualitative y, ayant L modes dans la série statistique $(y_1, y_2, \ldots, y_n)$, on utilise généralement une **table de contingence** A de taille $K \times L$. Il s'agit de compter, pour chaque mode de x et chaque mode de y, combien d'individus présentent ces deux modes : A_{ij} est le nombre d'invididus pour lesquels $x = i$ et $y = j$.

Si l'on appelle $N = \sum_{k=1}^{K}\sum_{l=1}^{L} A_{kl}$ le nombre total d'individus, $N_{i.} = \sum_{l=1}^{L} A_{il}$ le nombre d'individus dans la ligne i et $N_{.j} = \sum_{k=1}^{K} A_{kj}$ le nombre d'individus dans la colonne j, alors l'absence de liaison entre x et y se traduit par

$$\frac{N_{ij}}{N} = \frac{N_{i.}}{N}\frac{N_{.j}}{N} \text{ pour tout } 1 \leq i \leq K, 1 \leq j \leq L.$$

L'écart entre les valeurs prises de part et d'autre de cette égalité se mesure grâce à la **distance du chi2**, définie par

$$d_{\chi^2} = \sum_{i=1}^{K}\sum_{j=1}^{L} \frac{\left(A_{ij} - \frac{N_{i.}N_{.j}}{N}\right)^2}{\frac{N_{i.}N_{.j}}{N}}$$

Exemple

La table de contingence pour les variables « âge » et « région » des données de remboursement est donnée dans le tableau 2.1. La distance du chi2 pour cette table de contingence est de 11,21, ce qui suggère une dépendance entre les variables « âge » et « région » dans les données.

Nous verrons dans la PC1 comment transformer cette distance en un test statistique de dépendance.

La table de contingence peut être visualisée grâce à deux **diagrammes en barres empilées** : on peut choisir de visualiser, pour chaque mode de x, la proportion relative des modes de y, ou inversement, pour chaque mode de y, la proportion relative des modes de x. Ces deux choix sont illustrés sur les figures 2.8 et 2.9 respectivement.

	Région												
Âge	5	11	24	27	28	32	44	52	53	75	76	84	93
0–19	3	5	3	3	3	3	4	3	2	3	3	4	4
20–39	8	18	4	7	4	9	11	6	4	7	8	10	11
40–59	11	26	11	13	13	16	15	10	8	12	17	15	19
> 60	15	31	18	16	21	21	22	12	12	19	24	23	26

TABLEAU 2.1 – Table de contingence pour l'âge et la région des données de remboursement.

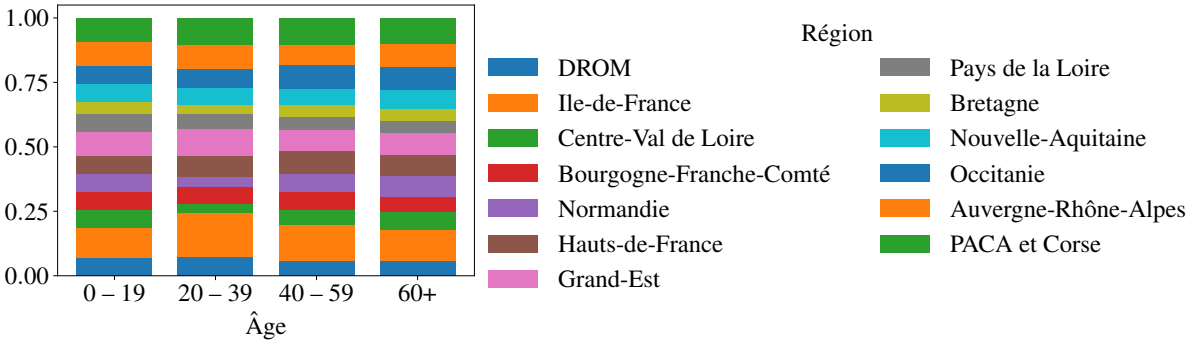


FIGURE 2.8 – Diagramme en barres représentant, pour chaque tranche d'âges, la proportion relative d'individus de chaque région dans la table de contingence du tableau 2.1.

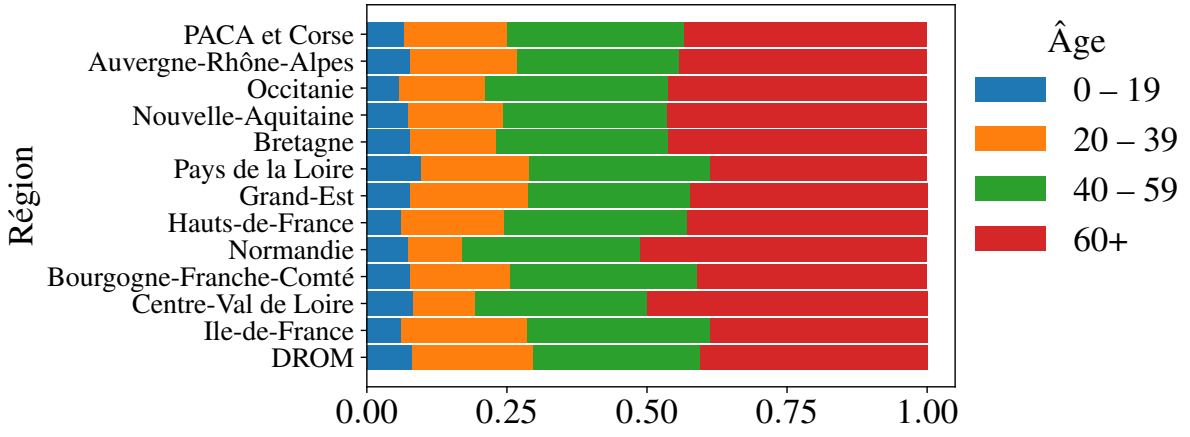


FIGURE 2.9 – Diagramme en barres représentant, pour chaque région, la proportion relative d'individus de chaque tranche d'âge dans la table de contingence du tableau 2.1.

Chapitre 3 Estimation

Notions : échantillon aléatoire, estimateur, estimation, biais d'un estimateur, convergence d'un estimateur, estimation par maximisation de la vraisemblance, estimation de Bayes.

Objectifs pédagogiques :

- Choisir un estimateur, en particulier en déterminant des propriétés telles que son biais ou sa précision.
- Proposer un estimateur, en particulier par maximisation de la vraisemblance.

3.1 Inférence statistique

Alors que la statistique descriptive se contente de *décrire* une population ou un échantillon de celle-ci, l'inférence statistique cherche à tirer des conclusions sur une population à partir de l'étude d'un échantillon de celle-ci.

3.2 Échantillonnage

Lorsque la population à étudier est trop grande pour qu'il soit possible d'observer chacun de ses individus, on étudie alors une partie seulement de la population. Cette partie est appelée **échantillon**. On parle alors de **sondage**, par opposition à un **recensement**, qui consiste à étudier tous les individus d'une population.

Hypothèses de l'échantillonnage Pour tirer parti d'un échantillon, nous allons avoir besoin des hypothèses suivantes :

- La taille de la population est infinie ;
- Les variables mesurées sur la population peuvent être considérées comme des variables aléatoires, dont les mesures sont des réalisations. Les lois de probabilité suivies par ces variables peuvent appartenir à une famille connue (e.g. loi gaussienne, loi de Poisson, etc.) ou être totalement inconnues. Dans le premier cas, on parlera de **statistique inférentielle paramétrique** ; dans le deuxième, de **statistique inférentielle non-paramétrique**.

Objectifs de la statistique inférentielle La statistique inférentielle a alors pour but d'**identifier les lois de probabilité des variables aléatoires** en décrivant les variables. Cela peut prendre les formes suivantes :

- L'estimation, qui permet d'approcher les paramètres des lois (paramètre p d'une loi de Bernoulli, indice et paramètre d'échelle d'une loi Gamma) ou certaines de leurs caractéristiques (espérance, variance, moments d'ordre supérieur, quartiles, etc.). C'est le sujet de ce chapitre.
- Les tests d'hypothèse, qui permettent d'infirmer ou de confirmer des hypothèses faites sur ces lois, leurs paramètres ou leurs caractéristiques. Il s'agit par exemple de décider s'il est plausible que l'espérance d'une variable soit supérieure à une certaine valeur ; ou qu'une variable suive une loi normale. C'est le sujet du prochain chapitre.

3.2.1 Échantillonnage aléatoire

Dans la suite de ce chapitre, nous allons considérer que l'échantillon obtenu par sondage est obtenu par **échantillonnage aléatoire simple** : on prélève des individus dans la population au hasard, sans

remise. Chaque individu de la population a la même probabilité $1/N$ d'être prélevé, où N est la taille de la population (on rappelle que $N \to \infty$) et les individus sont prélevés indépendamment les uns des autres.

Autres techniques d'échantillonnage D'autres techniques d'échantillonnage sont possibles, comme l'échantillonnage aléatoire *stratifié*, dans lequel la population est partitionnée en strates selon une caractéristique (par exemple, par tranche d'âge), et l'échantillon est obtenu en procédant à un échantillonnage aléatoire simple dans chacune des strates, permettant d'obtenir pour chaque strate un échantillon de taille proportionnelle à la taille de strate dans la population. Ainsi, les individus n'ont pas tous la même probabilité d'être tirés : celle-ci dépend de la taille de la strate à laquelle ils appartiennent.

Représentativité Avant de tirer des conclusions d'un échantillon aléatoire, il est important de s'assurer que celui-ci est représentatif de la population étudiée. Ainsi, les premières études cliniques démontrant l'efficacité de l'aspirine pour réduire le risque d'infarctus du myocarde chez les patients à risque portaient sur des échantillons composées principalement d'hommes ; ce n'est que bien plus tard que la communauté médicale a réalisé que ce n'est pas le cas chez les femmes.

Deux échantillons $(x_1,x_2,\dots,x_n)$ et $(x'_1,x'_2,\dots,x'_n)$ de tailles identiques n de la même population seront donc différents. On modélise cette variabilité en considérant que chacun des individus x_i ou x'_i est la réalisation d'une même variable aléatoire X_i, où $(X_1,X_2,\dots,X_n)$ est un vecteur aléatoire, dont les composantes sont indépendantes et identiquement distribuées.

- $(X_1,X_2,\dots,X_n)$ est appelé **échantillon aléatoire** ;
- $(x_1,x_2,\dots,x_n)$ et $(x'_1,x'_2,\dots,x'_n)$ sont deux échantillons, c'est-à-dire deux *réalisations* de cet échantillon aléatoire.

Un indicateur statistique de l'échantillon est alors la réalisation d'une variable aléatoire fonction de l'échantillon aléatoire.

Exemple

La moyenne d'un échantillon, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, est la réalisation d'une variable aléatoire M_n définie par

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

qui est une fonction de l'échantillon aléatoire $(X_1,X_2,\dots,X_n)$.

3.3 Estimation ponctuelle

Soit $(\Omega,\mathcal{A},\mathbb{P})$ un espace probabilisé, E un espace mesurable, et X une variable aléatoire à valeurs dans E. En pratique, dans la suite de ce chapitre, nous considèrerons des variables aléatoires réelles ($E = \mathbb{R}$ ou une partie de $\mathbb{R}$ telle que $\mathbb{R}_+$ ou $\mathbb{N}$), mais les idées qui y sont présentées peuvent être étendues à $\mathbb{R}^d$ ou à des espaces plus sophistiqués.

Soit $(X_1,X_2,\dots,X_n)$ un échantillon aléatoire. Les X_i sont indépendants et identiquement distribuées, de même loi $\mathbb{P}_X$ que X. Soit $(x_1,x_2,\dots,x_n)$ un échantillon, autrement dit une réalisation de cet échantillon aléatoire.

Soit $\theta \in \mathbb{R}$ une quantité déterministe (i.e. il ne s'agit pas d'une variable aléatoire), qui dépend uniquement de $\mathbb{P}_X$. Le but de l'estimation ponctuelle est d'approcher au mieux la valeur de θ.

Si l'on fait l'hypothèse que X suit une loi exponentielle (statistique inférentielle paramétrique), on peut chercher à estimer le paramètre θ de cette loi. On peut aussi chercher à estimer l'espérance de $\mathbb{P}_X$, un de ses moments, un quantile, etc.

3.3.1 Définition d'un estimateur

On appelle **estimateur** de θ une statistique de l'échantillon aléatoire $(X_1,X_2,\dots,X_n)$, c'est à dire une variable aléatoire fonction de $(X_1,X_2,\dots,X_n)$: un estimateur Θ_n de θ peut être défini par

$$\Theta_n = g(X_1,X_2,\dots,X_n), \qquad g : E \to \mathbb{R}.$$

Étant donné un échantillon $(x_1,x_2,\dots,x_n)$ de X, on appelle **estimation** de θ la valeur

$$\hat{\theta}_n = g(x_1,x_2,\dots,x_n) \in \mathbb{R},$$

qui est donc une réalisation de Θ_n.

Résumé Étant donné une variable aléatoire réelle X à valeurs dans E, un entier $n \in \mathbb{N}^*$, et une valeur θ à estimer qui ne dépend que de la loi de X,

- un échantillon aléatoire $(X_1,X_2,\dots,X_n)$ est un vecteur aléatoire, dont les composantes sont iid de même loi que X ;
- un échantillon $(x_1,x_2,\dots,x_n) \in \mathbb{R}^n$ est une réalisation de ce vecteur aléatoire ;
- un estimateur de θ est une variable aléatoire Θ_n fonction de $(X_1,X_2,\dots,X_n)$: $\Theta_n = g(X_1,X_2,\dots,X_n)$, avec $g : E \to \mathbb{R}$;
- une estimation de θ est une réalisation $\hat{\theta}_n$ de Θ_n : $\hat{\theta}_n = g(x_1,x_2,\dots,x_n) \in \mathbb{R}$.

3.3.2 Exemple : estimation de la moyenne par la moyenne empirique

Considérons maintenant que X est de carré intégrable ($X \in \mathcal{L}^2$), d'espérance m et de variance σ^2.

La **moyenne empirique** de X est une variable aléatoire M_n, définie par

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3.1}$$

M_n est un estimateur de m : étant donné un échantillon $(x_1,x_2,\dots,x_n)$, la valeur $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ est une estimation de m.

À ce stade, rien ne nous permet de dire que M_n est un *bon* estimateur de m ; en effet, nous pourrions aussi définir $\frac{2}{n}\sum_{i=1}^{n} X_i$ comme estimateur de la moyenne.

Question : Quelles sont selon vous les *propriétés* de M_n qui nous font préférer poser M_n comme nous l'avons fait ?

Réponse

- $\mathbb{E}(M_n) = m$.
  Nous verrons que l'on dit que M_n est un estimateur *non-biaisé* de m (cf. section 3.4.1) ;
- $\mathbb{V}(M_n) = \frac{\sigma^2}{n}$ (voir calcul section 3.7.1) : plus l'échantillon est grand, plus la variance de l'estimateur est faible, autrement dit plus sa réalisation $\hat{m}_n$ sera proche de son espérance m.
  On parle ici de la *précision* de M_n (cf. section 3.4.3) ;
- Par la loi faible des grands nombres, $M_n \xrightarrow{\mathbb{P}} m$.
  Nous verrons que l'on dit que M_n est un estimateur *convergent* de m (cf. section 3.4.4) ;
- Par la loi forte des grands nombres, $M_n \xrightarrow{\text{p.s.}} m$.
  Nous verrons que l'on dit que M_n est un estimateur *fortement convergent* de m (cf. section 3.4.4).

3.4 Propriétés d'un estimateur

Nous considérons toujours dans cette section un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de taille $n \in \mathbb{N}^*$ d'une variable aléatoire réelle X de loi $\mathbb{P}_X$, et un estimateur Θ_n de θ.

Notre but ici est maintenant de caractériser Θ_n.

3.4.1 Biais d'un estimateur

Le **biais** d'un estimateur Θ_n de la quantité θ est défini par

$$\mathrm{B}(\Theta_n) = \mathbb{E}(\Theta_n) - \theta. \tag{3.2}$$

Θ_n est dit **non-biaisé** si $\mathrm{B}(\Theta_n) = 0$, autrement dit si son espérance vaut θ.

La figure 3.1 illustre les distributions de 3 estimateurs d'une même quantité θ. On suppose ici que ce sont des gaussiennes. Les estimateurs Θ et Θ'' sont non-biaisés. Θ' est biaisé : son espérance vaut $\theta + \epsilon$.

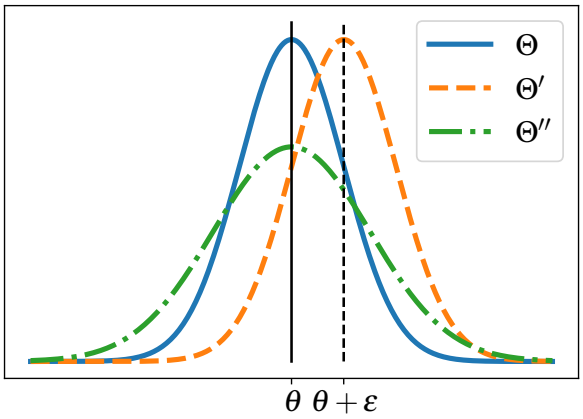


FIGURE 3.1 – Distribution de 3 estimateurs de θ.

3.4.2 Exemple : Estimation non-biaisée de la variance

Considérons X est de carré intégrable ($X \in \mathcal{L}^2$), d'espérance m et de variance σ^2.

La **variance empirique** de X est une variable aléatoire S_n, définie par

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2, \tag{3.3}$$

où M_n est la moyenne empirique telle que définie précédemment.

S_n est un estimateur de σ^2.

Cependant, son biais vaut $\frac{n-1}{n}\sigma^2$ (voir calcul section 3.7.2).

On propose donc la **variance empirique corrigée,** définie par

$$S_n^* = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2, \tag{3.4}$$

et qui est non-biaisé.

Néanmoins, le biais de la variance empirique tend vers 0 lorsque n tend vers $+\infty$. On parle alors d'un estimateur **asymptotiquement non-biaisé.**

3.4.3 Précision d'un estimateur

Reprenons la figure 3.1. Les deux estimateurs Θ et Θ'' sont non-biaisés. Cependant, Θ'' a une plus grande variance ; une de ses réalisation a une probabilité plus grande que pour Θ d'être éloignée de θ.

C'est cette notion que l'on utilise pour mesurer la précision d'un estimateur. Dans le cas d'un estimateur non-biaisée, sa précision est définie comme sa variance.

Dans le cas général d'un estimateur biaisé, il faut aussi prendre en compte le biais. Un estimateur biaisé mais avec une faible variance pourra donner de meilleures estimations (c'est-à-dire plus proches de la vraie valeur) qu'un estimateur moins biaisé mais avec une plus grande variance.

On utilise pour quantifier la précision d'un estimateur ponctuel générique son **erreur quadratique moyenne,** définie comme

$$\mathrm{EQM}(\Theta_n) = \mathbb{E}((\Theta_n - \theta)^2) = \mathbb{V}(\Theta_n) + \mathrm{B}(\Theta_n)^2. \tag{3.5}$$

Compromis biais-variance Il est tout à fait possible qu'un estimateur biaisé ait une meilleure précision qu'un estimateur non-biaisé, si ce dernier a une plus grande variance !

3.4.4 Convergence d'un estimateur ★

On souhaite aussi d'un estimateur qu'il permette de s'approcher d'autant mieux de la quantité qu'il estime que le nombre d'échantillons est grand. On parle ici de la convergence d'une série de variables aléatoires réelles, $(\Theta_n)_{n\in\mathbb{N}^*}$, vers une valeur réelle, θ ; il s'agit donc en fait de considérer la convergence vers une variable aléatoire Θ qui vaut θ presque partout.

On dit que l'estimateur Θ_n de θ **est convergent** si

$$(\Theta_n)_{n\in\mathbb{N}^*} \xrightarrow{\mathbb{P}} \theta. \tag{3.6}$$

Si de plus $(\Theta_n)_{n\in\mathbb{N}^*} \xrightarrow{\text{p.s.}} \theta$, on dit alors que Θ_n est un estimateur **fortement convergent** de θ.

Propriété Un estimateur sans biais et de variance asymptomatiquement nulle est convergent.

Preuve La preuve en a été faite dans l'exercice « Convergence vers une constante » de Probabilité III. Pour rappel, posons Θ_n un estimateur non biaisé et de variance asymptotiquement nulle de $\theta \in \mathbb{R}$. Par définition, $\mathbb{E}(\Theta_n) = \theta$ et $\mathbb{V}(\Theta_n) \xrightarrow[n\to+\infty]{} 0$, Θ_n est donc d'espérance et de variance bornées et ainsi dans $\mathcal{L}^2$. Enfin,

$$\mathbb{E}((\Theta_n - \theta)^2) = \mathbb{V}(\Theta_n) + (\mathbb{E}(\Theta_n) - \theta)^2,$$

et donc $\mathbb{E}((\Theta_n - \theta)^2) \xrightarrow[n\to+\infty]{} 0$, ce qui signifie que $\Theta_n \xrightarrow{\mathcal{L}^2} \theta$ et donc $\Theta_n \xrightarrow{\mathbb{P}} \theta$. □

Remarque On utilise en anglais le terme de "*consistent*", ce qui conduit les francophones à parfois parler d'estimateur consistant plutôt que convergent.

3.4.5 Exercice (estimation de la moyenne)

Nous cherchons à déterminer le poids moyen des bébés à la naissance en France. Pour cela, nous disposons d'un échantillon $(x_1,x_2,\ldots,x_n)$ de n mesures obtenues dans plusieurs maternités à travers le pays.

Nous supposons que cet échantillon est une réalisation d'un échantillon $(X_1,X_2,\ldots,X_n)$ de variables aléatoire réelles indépendantes et identiquement distribuées, d'espérance m et de variance σ^2.

On propose deux estimateurs de m :

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ et } Z_n = \frac{1}{2}(X_n + X_{n-1}).$$

Montrer que M_n et Z_n sont sans biais. Lequel choisir pour approcher m ?

(Solution : section 3.7.3.)

3.5 Estimation par maximum de vraisemblance

Nous considérdons toujours dans cette section un échantillon aléatoire $(X_1,X_2,\ldots,X_n)$ de taille $n \in \mathbb{N}^*$ d'une variable aléatoire réelle X, et une quantité $\theta \in \mathcal{S} \subseteq \mathbb{R}$ à estimer. Nous notons $\mathbb{P}_X$ la loi de X.

Nous venons de voir comment caractériser un estimateur Θ_n afin de choisir le meilleur estimateur parmi plusieurs. Mais comment proposer un estimateur de θ ?

Supposons que $(x_1,x_2,\ldots,x_n)$ est une réalisation de $(X_1,X_2,\ldots,X_n)$. La technique que nous allons voir consiste à maximiser la vraisemblance de l'échantillon, autrement dit la probabilité d'observer cet échantillon étant donnée la valeur estimée de θ.

Exemple

Nous nous intéressons à la réussite d'élèves au baccalauréat en Île-de-France, et disposons d'observations issues de plusieurs lycées de la région.

Nous modélisons l'observation « réussite » ou « échec » comme la réalisation d'une variable aléatoire X, de domaine $E = \{0,1\}$ (0 correspondant à « échec » et 1 à « réussite »), et suivant une loi de probabilité $\mathbb{P}_X$. Un choix classique pour cette loi de probabilité est d'utiliser une loi de Bernoulli de paramètre p :

$$\mathbb{P}_X(X = x) = p^x(1-p)^{1-x}.$$

Nos observations constituent un échantillon $(x_1,x_2,\ldots,x_n)$, qui est une réalisation de l'échantillon aléatoire $(X_1,X_2,\ldots,X_n)$ de composantes indépendantes et identiquement distribuées de même loi que X.

Nous cherchons à estimer p à partir de cet échantillon.

Supposons que notre échantillon contient $n = 500$ élèves, dont $b = 450$ ont eu le bac.

La valeur $p = 50\%$ est peu vraisemblable ; la valeur $p = 90\%$ l'est beaucoup plus. C'est cette notion que nous allons formaliser par la suite.

On appelle **vraisemblance** de l'échantillon $(x_1,x_2,\ldots,x_n)$ la fonction de θ définie comme :

$$L(x_1,x_2,\ldots,x_n;\theta) = \mathbb{P}_X(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|\theta) = \prod_{i=1}^{n} \mathbb{P}_X(X_i = x_i|\theta),$$

cette dernière égalité étant dûe à l'indépendance des éléments de l'échantillon aléatoire.

On appelle alors **estimation par maximum de vraisemblance** (*maximum likelihood estimate* ou *MLE* en anglais) de θ une valeur $\widehat{\theta}_{\text{MLE}}$ qui maximise la vraisemblance, autrement dit la probabilité d'observer l'échantillon $(x_1,x_2,\ldots,x_n)$ étant donné θ :

$$\widehat{\theta}_{\text{MLE}} \in \arg\max_{\widehat{\theta}\in\mathcal{S}} \prod_{i=1}^{n} \mathbb{P}_X(X_i = x_i|\widehat{\theta}). \tag{3.7}$$

Un estimateur par maximum de vraisemblance de θ est une variable aléatoire réelle $\widehat{\Theta}_{\text{MLE}}$ dont la valeur quand $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ est donnée par $\widehat{\theta}_{\text{MLE}}$.

Pour simplifier les calculs, on choisira souvent de maximiser non pas directement la vraisemblance mais son logarithme :

$$\widehat{\theta}_{\text{MLE}} \in \arg\max_{\widehat{\theta} \in \mathcal{S}} \sum_{i=1}^{n} \log \mathbb{P}_X(X_i = x_i | \widehat{\theta}). \tag{3.8}$$

Exemple

Reprenons notre exemple de réussite au baccalauréat.

L'estimation par maximum de vraisemblance de p est

$$\begin{aligned}\widehat{p}_{\text{MLE}} &= \arg\max_{p \in [0,1]} \sum_{i=1}^{n} \log \mathbb{P}_X(X_i = x_i | p) = \arg\max_{p \in [0,1]} \sum_{i=1}^{n} \log \left(p^{x_i} (1-p)^{1-x_i} \right) \\ &= \arg\max_{p \in [0,1]} \sum_{i=1}^{n} x_i \log p + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p).\end{aligned}$$

La fonction $L : p \mapsto \sum_{i=1}^{n} x_i \log p + (n - \sum_{i=1}^{n} x_i) \log(1-p)$ est concave, nous pouvons donc la maximiser en annulant sa dérivée :

$$\frac{\partial L}{\partial p} = \sum_{i=1}^{n} x_i \frac{1}{p} - \left(n - \sum_{i=1}^{n} x_i \right) \frac{1}{1-p},$$

ce qui nous donne

$$(1 - \widehat{p}_{\text{MLE}}) \sum_{i=1}^{n} x_i - \widehat{p}_{\text{MLE}} \left(n - \sum_{i=1}^{n} x_i \right) = 0$$

et donc

$$\widehat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{b}{n}. \tag{3.9}$$

L'estimateur par maximum de vraisemblance de p est ainsi tout simplement la moyenne empirique de l'échantillon. Dans notre exemple, $p = 450/500 = 90\%$.

Propriété L'estimateur par maximum de vraisemblance est convergent.

3.6 Estimation de Bayes ★

Supposons que plutôt que de ne pas connaître du tout la valeur du paramètre θ, nous ayons une bonne idée des valeurs qu'il peut prendre. Cette information peut être très utile, surtout quand le nombre d'observations est faible.

Pour l'utiliser, nous allons utiliser une variable aléatoire Θ, dont la loi $\mathbb{P}_\Theta$ est la **distribution a priori**, c'est-à-dire définie avant d'avoir observé un échantillon, des valeurs de θ.

Il va maintenant s'agir d'utiliser la loi de Bayes pour exprimer la **distribution a posteriori** de Θ :

$$\mathbb{P}(\Theta | X_1, X_2, \ldots, X_n) = \frac{\mathbb{P}(X_1, X_2, \ldots, X_n | \Theta) \mathbb{P}(\Theta)}{\mathbb{P}(X_1, X_2, \ldots, X_n)}. \tag{3.10}$$

La distribution a posteriori de Θ, $\mathbb{P}(\Theta | X_1, X_2, \ldots, X_n)$, s'exprime en fonction de sa distribution a priori $\mathbb{P}(\Theta)$, de la vraisemblance $\mathbb{P}(X_1, X_2, \ldots, X_n | \Theta)$, et de la probabilité marginale de l'échantillon $(X_1, X_2, \ldots, X_n)$, $\mathbb{P}(X_1, X_2, \ldots, X_n)$.

En d'autres termes, les observations permettent d'ajuster la distribution a priori de Θ en sa distribution a posteriori. Cette idée est au cœur de **l'inférence bayésienne.**

Là où l'estimation par maximum de vraisemblance cherche à maximiser la vraisemblance, l'estimation de Bayes cherche à minimiser l'erreur postérieure. La formulation est générique ; on peut imaginer utiliser plusieurs définitions de cette erreur.

Une des définitions les plus fréquentes, qui est celle que nous utiliserons par la suite, consite à considérer comme fonction d'erreur l'erreur quadratique moyenne (définie section 3.4.3).

L'**estimation de Bayes pour l'erreur quadratique moyenne** de θ est alors définie par

$$\widehat{\theta}_{\text{Bayes}} \in \arg\min_{\widehat{\theta}\in\mathcal{S}} \mathbb{E}((\Theta - \widehat{\theta})^2), \tag{3.11}$$

cette espérance étant prise sur les distributions jointes de Θ et de $(X_1,X_2,\ldots,X_n)$.

Propriété L'estimation de Bayes pour l'erreur quadratique moyenne est l'espérance de la distribution postérieure de Θ :

$$\widehat{\theta}_{\text{Bayes}} = \mathbb{E}(\Theta|X_1 = x_1,X_2 = x_2,\ldots,X_n = x_n). \tag{3.12}$$

Preuve En effet, posons $\widehat{\theta} \in \mathcal{S}$. Alors

$$\mathbb{E}((\Theta - \widehat{\theta})^2) = (\mathbb{E}(\Theta) - \widehat{\theta})^2 + \mathbb{E}(\Theta^2) - \mathbb{E}(\Theta)^2.$$

Comme ni $\mathbb{E}(\Theta^2)$ ni $\mathbb{E}(\Theta)^2$ ne dépendent de $\widehat{\theta}$, $\widehat{\theta}_{\text{Bayes}}$ est obtenu en minimisant $(\mathbb{E}(\Theta) - \widehat{\theta})^2$ et donc $\widehat{\theta}_{\text{Bayes}} = \mathbb{E}(\Theta) - \mathbb{E}(\Theta|(X_1,X_2,\ldots,X_n))$. □

Exemple

Reprenons notre exemple de taux de réussite au baccalauréat. Nous supposons maintenant que p est une réalisation d'une variable aléatoire Θ qui suit une loi bêta de paramètres (α,β) (cf. section 3.7.4).

Pour calculer l'estimateur de Bayes de p, il nous faut connaître la loi

$$\mathbb{P}(\Theta|X_1 = x_1,X_2 = x_2,\ldots,X_n = x_n).$$

La loi de Bayes, combinée à l'hypothèse d'indépendance et de distribution identique des X_i, nous permet d'écrire

$$\begin{aligned}\mathbb{P}(\Theta = p|X_1,X_2,\ldots,X_n) &= \frac{\mathbb{P}(X_1,X_2,\ldots,X_n|p)\mathbb{P}(p)}{\mathbb{P}(X_1,X_2,\ldots,X_n)} \\ &= \frac{1}{\mathbb{P}(X_1,X_2,\ldots,X_n)B(\alpha,\beta)} \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} p^{\alpha-1}(1-p)^{\beta-1} \\ &= \frac{1}{\mathbb{P}(X_1,X_2,\ldots,X_n)B(\alpha,\beta)} p^{b+\alpha-1}(1-p)^{n-b+\beta-1}.\end{aligned}$$

On reconnaît ici la densité d'une nouvelle loi bêta. Ainsi $\Theta|X_1,X_2,\ldots,X_n$ suit une loi bêta de paramètres $(b+\alpha)$ et $(n-b+\beta)$.

L'estimation de Bayes de p est ainsi

$$\widehat{p}_{\text{Bayes}} = \mathbb{E}(\Theta|X_1 = x_1,X_2 = x_2,\ldots,X_n = x_n) = \frac{(b+\alpha)}{(b+\alpha)+(n-b+\beta)} = \frac{b+\alpha}{n+\alpha+\beta}.$$

Cette première égalité est obtenue d'après la formule donnant l'espérance d'une loi bêta (cf section 3.7.4).

Remarque importante On peut réécrire cette estimation sous la forme

$$\widehat{p}_{\text{Bayes}} = \frac{\alpha+\beta}{n+\alpha+\beta}\mathbb{E}[\Theta] + \frac{n}{n+\alpha+\beta}\widehat{p}_{\text{MLE}}.$$

Ainsi, l'estimation de Bayes du paramètre p est une combinaison linéaire de l'espérance de sa distribution a priori et de son estimation par maximum de vraisemblance.

De plus, le coefficient multiplicatif de l'espérance a priori décroît en fonction de la taille n de l'échantillon, tandis que le coefficient multiplicatif de l'estimation par maximum de vraisemblance croît en fonction de n. Ainsi, plus l'échantillon est grand, plus l'estimateur de Bayes fait confiance aux données, et s'éloigne de l'espérance a priori du paramètre, dont on est plus proche avec un petit échantillon.

La figure 3.2 illustre cet exemple.

Remarque Le choix d'une loi bêta ne s'est pas fait au hasard. On retrouve ici les lois conjuguées présentées en exercice de Probabilités IV. En inférence bayésienne, on dit qu'une loi a priori et une loi a posteriori sont conjuguées lorsqu'elles appartiennent à la même famille. En particulier, la loi bêta est conjuguée à elle-même pour une vraisemblance de Bernoulli.

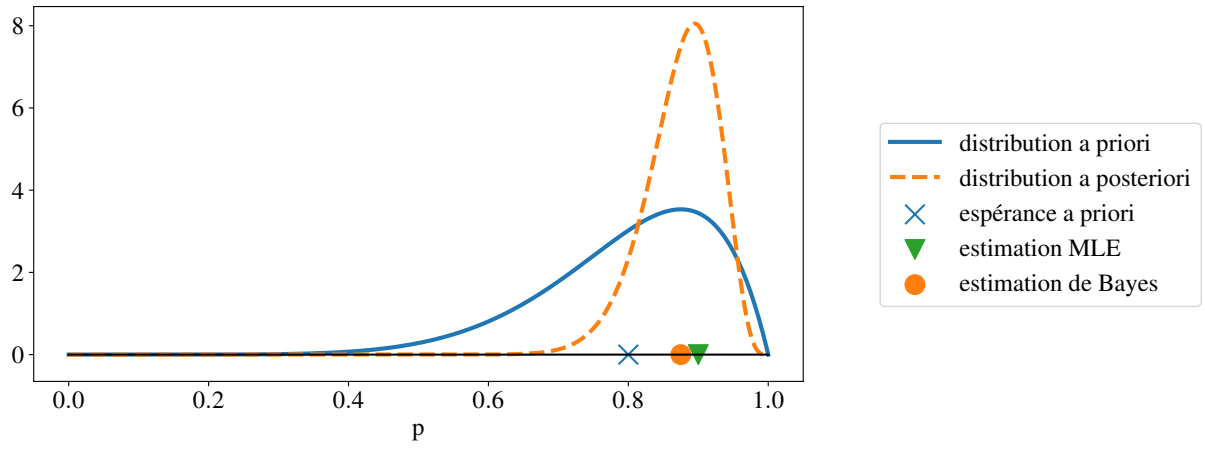


FIGURE 3.2 – Loi a priori et a posteriori pour le paramètre p dans l'exemple du taux de réussite au baccalauréat. Sans voir de données, $p = 0{,}80$, c'est-à-dire l'espérance de sa loi a priori (croix bleue). En utilisant uniquement l'échantillon, $p = 0{,}90$, c'est-à-dire son estimation par maximum de vraisemblance (triangle vert). L'estimation de Bayes (rond orange) est intermédiaire.

3.7 Compléments

3.7.1 Variance de la moyenne empirique

Soit X une variable aléatoire réelle de carré intégrable, d'espérance m et de variance σ^2. Soient $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées, de même loi que X.

Par définition de la variance, $\sigma^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ donc $\mathbb{E}(X^2) = \sigma^2 + m^2$.

Posons $M_n = \frac{1}{n}\sum_{i=1}^n X_i$.

$$\mathbb{V}(M_n) = \mathbb{E}(M_n^2) - \mathbb{E}(M_n)^2 = \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right) - m^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n X_i \sum_{j=1}^n X_j\right) - m^2$$
$$= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n \left(X_i^2 + \sum_{j\neq i}^n X_i X_j\right)\right) - m^2 = \frac{1}{n}\left(\mathbb{E}(X^2) + \sum_{j\neq i}^n \mathbb{E}(X)^2\right) - m^2,$$

Et donc

$$\mathbb{V}(M_n) = \frac{1}{n}\left(\sigma^2 + m^2 + (n-1)m^2\right) - m^2 = \frac{\sigma^2}{n}.$$

3.7.2 Biais de la variance empirique

Soit X une variable aléatoire réelle de carré intégrable, d'espérance m et de variance σ^2. Soient $X_1, X_2, \ldots, X_n$ indépendantes et identiquement distribuées, de même loi que X.

Posons $M_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - M_n)^2$. Alors

$$\mathbb{E}(S_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}((X_i - M_n)^2) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}(X_i^2) + \mathbb{E}(M_n^2) - 2\mathbb{E}(X_i M_n)\right)$$
$$= \mathbb{E}(X^2) + \mathbb{E}(M_n^2) - \frac{2}{n}\sum_{i=1}^n \mathbb{E}(X_i M_n).$$

Nous avons montré lors du calcul de la variance de la moyenne empirique que $\mathbb{E}(X_i^2) = \sigma^2 + m^2$ et que $\mathbb{E}(M_n^2) = m^2 + \frac{\sigma^2}{n}$.

De plus, par linéarité de l'espérance,

$$\mathbb{E}(M_n^2) = \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right) M_n\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i M_n),$$

et donc

$$\mathbb{E}(M_n^2) - \frac{2}{n}\sum_{i=1}^n \mathbb{E}(X_i M_n) = -\mathbb{E}(M_n^2).$$

On obtient ainsi

$$\mathbb{E}(S_n) = (\sigma^2 + m^2) - (m^2 + \frac{\sigma^2}{n}) = \frac{n-1}{n}\sigma^2.$$

La variance empirique est donc biaisée et son biais vaut

$$\mathrm{B}(S_n) = \mathbb{E}(S_n) - \sigma^2 = -\frac{1}{n}\sigma^2.$$

3.7.3 Solution de l'exercice 3.4.5

La démonstration pour la moyenne empirique M_n a été faite plus haut.

En ce qui concerne Z_n,

$$\mathbb{E}(Z_n) = \frac{1}{2}(\mathbb{E}(X_n) + \mathbb{E}(X_{n-1})) = m.$$

Nous avons assez naturellement envie d'utiliser M_n, qui utilise toutes les observations, plutôt que Z_n, qui n'en utilise que deux.

Pour nous en convaincre, nous pouvons comparer les variances de M_n et Z_n. La variance de la moyenne empirique est $\mathbb{V}(M_n) = \frac{\sigma^2}{n}$ (voir plus haut). La variance de Z_n, elle, vaut

$$\mathbb{V}(Z_n) = \frac{1}{4}\left(\mathbb{V}(X_n) + \mathbb{V}(X_{n-1})\right) = \frac{\sigma^2}{2},$$

la première égalité étant obtenue par indépendance de X_n et X_{n-1}.

Z_n est ainsi un estimateur bien moins précis que M_n dès que $n > 2$.

3.7.4 Loi Beta

La densité de probabilité de la *loi bêta* de paramètres $\alpha,\beta > 0$, définie sur $0 \leq u \leq 1$, est donnée par :

$$f_{\alpha,\beta}(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha,\beta)} \tag{3.13}$$

où $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ et Γ est la fonction gamma. L'espérance de cette loi est $\frac{\alpha}{\alpha+\beta}$.

Pour aller plus loin

- Un exercice sur la fonction de répartition empirique vous a été proposé dans le poly de Probabilité III.
- On peut construire un estimateur par la *méthode des moments*, qui consiste à faire coincider les moments théoriques de $\mathbb{P}_X$ (qui dépendent donc de θ) avec les moments empiriques de l'échantillon. La loi des grands nombres justifie en effet d'approcher la moyenne par la moyenne empirique. Cette méthode est généralement moins précise que le maximum de vraisemblance.
- Plus la variance d'un estimateur est faible, plus cet estimateur peut-être considéré comme précis. La *borne de Cramér-Rao* est une borne inférieure de cette variance pour un estimateur sans biais, en se basant sur l'information de Fisher. On dit qu'un estimateur est *efficace* s'il est non-biaisé et que sa variance tend vers sa borne de Cramér-Rao.

Chapitre 4 Tests d'hypothèse

Notions : hypothèse nulle, hypothèse alternative, statistique de test, p-valeur, tests multiples.

Objectifs pédagogiques :

— Reconnaître une situation dans laquelle un test statistique est approprié.

— Poser les hypothèses de test nulle et alternative correspondant à un énoncé.

— Interpréter une statistique de test ou une p-valeur.

4.1 Principe d'un test statistique

Le but d'un test statistique est de déterminer la fiabilité d'une observation faite sur une échantillon.

Exemple

Si je lance une pièce 5 fois et obtiens 5 fois pile, puis-je en déduire que la pièce est déséquilibrée ? Ou ce résultat est-il dû au hasard de l'échantillonnage ? Qu'en est-il si j'obtiens le même résultat après 50 lancers ?

Un test statistique permet de déterminer si l'échantillon observé permet d'invalider une hypothèse qu'il était raisonnable de formuler avant d'observer les données.

Exemple

Reprenons l'exemple du lancer de pièce. Sous l'hypothèse que la pièce est équilibrée, la probabilité π d'obtenir « pile » pour un lancer est $0{,}5$ et celle d'obtenir pile pour 5 lancers est $0{,}5^5 = 3\%$. Cette probabilité est faible, mais non négligeable : on a 3% de chance d'obtenir un résultat aussi extrême que celui observé sur un échantillon.

Pour 50 lancers, cette probabilité tombe à $0{,}5^{50} = 9{,}10^{-16}$. Cette probabilité est extrêmement faible, et l'échantillon ne soutient pas l'hypothèse selon laquelle la pièce est équilibrée : nous pouvons la rejeter.

4.2 Formalisme

Soit $(X_1,X_2,\ldots,X_n)$ un échantillon aléatoire de taille $n \in \mathbb{N}^*$ d'une variable aléatoire réelle X de loi $\mathbb{P}_X$. Rappelons que les composantes X_i de ce vecteur aléatoire sont indépendantes et identiquement distribuées, de même loi que X. Les notions présentées dans ce chapitre s'appliquent aussi à des variables aléatoires de nature plus complexe (par exemple, des valeurs aléatoires multi-dimensionnelles) mais nous nous limitons aux variables aléatoires réelles par souci de simplicité. Nous supposons aussi disposer d'un échantillon $(x_1,x_2,\ldots,x_n)$ qui est une réalisation de $(X_1,X_2,\ldots,X_n)$.

Un test statistique repose sur les éléments suivants :

— Une **hypothèse nulle,** notée $\mathcal{H}_0$. L'hypothèse nulle est celle que l'on chercher à rejeter.

— Une **hypothèse alternative,** notée $\mathcal{H}_1$ ou $\mathcal{H}_a$. C'est en général la négation de $\mathcal{H}_0$.

— Une **statistique de test,** T, qui sert à mesurer à quel point un échantillon « dévie » de l'hypothèse nulle.

— Un **niveau de signification,** $0 < \alpha < 1$, qui est la probabilité de rejeter l'hypothèse nulle alors qu'elle est correcte.

Le but de cette section est de développer ces notions.

4.2.1 Hypothèses de test

Conduire un test d'hypothèse nécessite de formuler deux hypothèses :

— Une **hypothèse nulle,** notée $\mathcal{H}_0$. Cette hypothèse doit être précise et permettre de faire des calculs. Le but du test est de déterminer s'il est raisonnable de rejeter cette hypothèse.

— Une **hypothèse alternative,** notée $\mathcal{H}_1$ ou $\mathcal{H}_a$. Cette hypothèse est une forme de négation de $\mathcal{H}_0$, et c'est l'hypothèse que l'on adoptera si l'hypothèse nulle est rejetée.

L'hypothèse nulle est souvent une hypothèse formulée sur la valeur un paramètre $\theta \in \mathcal{S} \subseteq \mathbb{R}$ caractérisant la loi $\mathbb{P}_X$ de l'échantillon aléatoire. Il s'agit alors de tester

$$\mathcal{H}_0 : \theta = \theta_0, \tag{4.1}$$

où $\theta_0 \in \mathcal{S}$ est une valeur déterministe fixée à l'avance.

L'hypothèse nulle peut cependant être de nature plus complexe, par exemple :

— « Deux variables statistique X et Y sont indépendantes » (c'est le cas du test d'indépendance du χ^2 que nous verrons dans la PC1).

— « Deux échantillons $(x_1,x_2,\ldots,x_n)$ et $(y_1,y_2,\ldots,y_n)$ sont des réalisations de la même distribution » (c'est le cas du test de Wilcoxon-Mann-Whitney, qui dépasse le cadre de ce programme) ;

Présomption d'innocence De même que le principe de la présomption d'innocence veut que l'on recueille suffisamment de preuves pour rejeter l'innocence, en théorie des tests statistiques il y a présomption de $\mathcal{H}_0$. Il s'agit donc de savoir si l'échantillon observé (les preuves) est suffisant pour rejeter $\mathcal{H}_0$, ce dont on conclura $\mathcal{H}_1$. Par contre, si l'on ne rejette pas $\mathcal{H}_0$, cela peut venir soit de ce que $\mathcal{H}_0$ est vraie, soit de ce que nous n'avons pas suffisamment de données pour rejeter $\mathcal{H}_1$. Ainsi, $\mathcal{H}_0$ doit être une hypothèse raisonnable, mais que l'on aimerait avoir des raisons de réfuter.

Dans le cadre d'une expérience scientifique, l'hypothèse $\mathcal{H}_0$ correspond ainsi à l'état actuel des connaissances. Le but d'un test statistique est de déterminer si les données qui semblent contredire cette hypothèse sont effectivement suffisamment improbables sous $\mathcal{H}_0$ pour justifier de la réfuter. Dans le cadre d'un essai clinique, par exemple, l'hypothèse $\mathcal{H}_0$ se doit d'être défavorable au nouveau médicament (« le nouveau médicament est inefficace » ou « le nouveau médicament n'est pas plus efficace que les traitements connus »). Le but du test statistique est de déterminer si les données récoltées jusqu'à présent sont suffisantes pour réfuter cette hypothèse.

Exemple

Dans le cas de notre lancer de pièce,

— X est une variable aléatoire discrète qui suit une loi de Bernoulli de paramètre π. $\mathbb{P}_X(1) = \pi$ et $\mathbb{P}_X(0) = 1 - \pi$;

— l'échantillon aléatoire est un vecteur $(X_1,X_2,\ldots,X_n)$ de n composantes, iid de même loi que X ;

— une série de lancers est une réalisation $(x_1,x_2,\ldots,x_n)$ de ce vecteur aléatoire. Dans le cas de 5 lancers tous tombant sur « pile », cet échantillon est $(1,1,1,1,1)$ et $n = 5$.

— l'hypothèse nulle est $\mathcal{H}_0 : \pi = 0{,}5$.

Dans le cas où l'on cherche à tester la valeur d'un paramètre θ d'une population, l'hypothèse alternative peut prendre deux formes :

— $\theta \neq \theta_0$, ou en d'autres termes,

$$\mathcal{H}_1 : \theta < \theta_0 \text{ ou } \theta > \theta_0. \tag{4.2}$$

On parle alors de test **bilatéral** (*two-sided test* en anglais).

— Si seulement l'une des deux parties de cette hypothèse alternative nous intéresse, ou est possible, on parle de test **unilatéral** (*one-sided test* en anglais). Il s'agit alors de tester soit

$$\mathcal{H}_1 : \theta < \theta_0, \tag{4.3}$$

soit

$$\mathcal{H}_1 : \theta > \theta_0. \tag{4.4}$$

De même que l'on élabore $\mathcal{H}_0$ de sorte à ce qu'elle soit la plus plausible avant d'avoir observé les données, on élabore $\mathcal{H}_1$ en fonction de ce que l'on espère découvrir.

Reprenons l'exemple d'un essai clinique sur un nouveau médicament. Si l'hypothèse $\mathcal{H}_0$ est « le nouveau médicament n'a pas d'effet », on peut poser l'hypothèse alternative $\mathcal{H}_1$: « le nouveau traitement a un effet positif sur l'état des patients ». On espère ici non seulement rejeter l'hypothèse nulle, mais aussi suggérer une efficacité du traitement. Cette hypothèse est plus précise que l'hypothèse alternative selon laquelle « le nouveau traitement a un effet sur l'état des patients », cet effet pouvant être négatif.

Exemple

Dans le cas de notre lancer de pièce, l'hypothèse alternative dans le cadre d'un test bilatéral est

$$\mathcal{H}_1 : \pi \neq 0{,}5.$$

Si nous rejetons $\mathcal{H}_0$, notre conclusion sera que la pièce n'est pas équilibrée.

Dans le cadre d'un test bilatéral, par exemple

$$\mathcal{H}_1 : \pi > 0{,}5,$$

si nous rejetons $\mathcal{H}_0$, notre conclusion sera que la pièce n'est pas équilibrée, et qu'elle favorise « pile ».

Il ne s'agit donc pas du même test.

4.2.2 Statistique de test et p-valeur

Une **statistique de test** T est une statistique de l'échantillon aléatoire. Il s'agit donc d'une variable aléatoire réelle, fonction de $(X_1,X_2,\ldots,X_n)$: $T = g(X_1,X_2,\ldots,X_n)$ Cette statistique de test sert à mesurer à quel point un échantillon « dévie » de l'hypothèse nulle.

Une statistique de test est ainsi choisie de sorte à avoir une loi différente sous $\mathcal{H}_0$ et sous $\mathcal{H}_1$, et de sorte à ce que sa loi sous $\mathcal{H}_0$ soit connue : c'est ce qui permettra de déterminer un critère de rejet de $\mathcal{H}_0$ garantissant le niveau de signification choisi.

La plupart des test statistiques reposent sur des statistiques de test dont le développement a été long et minutieux. Le choix entre plusieurs statistiques candidates pour un même problème est un choix difficile, qui repose entre autres sur la validité des hypothèses sur la distribution de l'échantillon aléatoire ou sur sa taille qui permettent de déterminer sa loi sous $\mathcal{H}_0$.

Remarque Pour des tests portant sur un paramètre ($\mathcal{H}_0 : \theta = \theta_0$), la statistique de test est souvent basée sur la différence entre un estimateur de ce paramètre et sa valeur sous $\mathcal{H}_0$.

Exemple

Reprenons l'exemple du lancer de pièce.

Dans la section 4.1, nous avons choisi comme statistique de test T le nombre de pile obtenus dans l'échantillon :

$$T = \sum_{i=1}^{n} X_i.$$

Sous $\mathcal{H}_0$, autrement dit si $\pi = 0{,}5$, la loi de T est déterminée par

$$\mathbb{P}(T = k) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = k\right) \text{ pour } k = 0{,}1, \ldots, n.$$

On reconnait ici une loi binomiale de paramètres n et π.

4.2.3 Niveau de signification

Nous avons maintenant posé $\mathcal{H}_0$, $\mathcal{H}_1$, et une statistique de test T dont nous connaissons la loi $\mathbb{P}_{T0}$ sous $\mathcal{H}_0$. Il nous faut maintenant déterminer le **domaine de rejet** du test, autrement dit l'ensemble $\mathcal{I} \subseteq \mathbb{R}$ de ses valeurs qui conduisent à rejeter $\mathcal{H}_0$.

Pour ce faire, nous avons besoin de fixer le **niveau de signification** (*significance level*), $0 < \alpha < 1$, qui est la probabilité de rejeter l'hypothèse nulle alors qu'elle est correcte. Ce seuil est fixé à l'avance, généralement parmi $\alpha = 1\%$, $\alpha = 5\%$ ou $\alpha = 10\%$, et détermine à quel point le test est strict.

Ainsi, il s'agit de déterminer $\mathcal{I} \subseteq \mathbb{R}$ de sorte à ce que $\mathbb{P}_{T0}(T \in \mathcal{I}) = \alpha$.

Exemple

Dans l'exemple du lancer de pièce, nous avons choisi le nombre de pile comme statistique de test T. Sous $\mathcal{H}_0 : \pi = 0{,}5$, T suit une loi binomiale de paramètres n (le nombre de lancers) et π.

Posons $\alpha = 5\%$.

Considérons le test unilatéral $\mathcal{H}_1 : \pi > 0{,}5$. Si nous rejetons $\mathcal{H}_0$, nous en conclurons que la pièce est biaisée en faveur du côté pile. Cela signifie que nous souhaitons rejeter $\mathcal{H}_0$ quand le nombre de pile dans l'échantillon est grand. Il est ici naturel de considérer un domaine de rejet de la forme $\mathcal{I} =]t_0{,}n]$. En d'autres termes, nous allons rejeter $\mathcal{H}_0$ si la réalisation t de T sur notre échantillon est plus grande qu'un seuil t_0, fixé tel que $\mathbb{P}_{T0}(T > t_0) = \alpha$.

En d'autres termes, si F_{T0} est la fonction de répartition de T sous $\mathcal{H}_0$, t_0 est fixé de sorte à ce que $F_{T0}(t_0) = \alpha$. Dans notre exemple avec $n = 5$ et $\alpha = 0{,}05$, cela fixe $t_0 = 4$.

Le test consiste donc à rejeter l'hypothèse nulle si tous les 5 lancers aboutissent à pile.

Considérons maintenant le test unilatéral $\mathcal{H}_1 : \pi < 0{,}5$. Rejeter $\mathcal{H}_0$ conduit à conclure que la pièce est biaisée en faveur du côté face. Nous considérons maintenant un domaine de rejet de la forme $\mathcal{I} = [0{,}t_0[$, et t_0 est déterminé par $\mathbb{P}_{T0}(T < t_0) = \alpha$. Avec $n = 5$ et $\alpha = 0{,}05$, cela fixe $t_0 = 1$. Le test consiste donc à rejeter l'hypothèse nulle si aucun des 5 lancers n'aboutit à pile.

Enfin, considérons le test bilatéral $\mathcal{H}_1 : \pi \neq 0{,}5$. Rejeter $\mathcal{H}_0$ conduit à conclure que la pièce est biaisée, en faveur de l'un ou de l'autre de ses côtés. Nous considérons alors un domaine de rejet de la forme $\mathcal{I} = [0{,}t_l[\cup]t_r{,}n]$. Il nous faut donc choisir t_l et t_r de sorte à ce que $\mathbb{P}_{T0}(T < t_l) + \mathbb{P}_{T0}(T > t_r) = \alpha$. Il est assez naturel de fixer alors $\mathbb{P}_{T0}(T < t_l) =$

$\mathbb{P}_{T0}(T > t_r) = \frac{\alpha}{2}$. Avec $n = 5$ et $\alpha = 0{,}05$, on obtient $t_l = 0$ et $t_r = 5$ et il n'est donc jamais possible de rejeter l'hypothèse nulle.

Le test que nous venons de définir s'appelle le **test binomial.**

Remarque importante On observe ici que, parmi les trois hypothèses alternatives envisagées, seul le test statistique unilatéral $\mathcal{H}_1 : \pi > 0{,}5$ nous permet de rejeter l'hypothèse nulle. C'est une observation générale : un test unilatéral est plus puissant qu'un test bilatéral ; cependant il n'est utile que si on sait de quel côté le définir.

4.2.4 Valeur critique

Dans le cas d'un test sur la valeur d'un paramètre θ, c'est-à-dire avec pour hypothèse nulle

$$\mathcal{H}_0 : \theta = \theta_0,$$

le domaine de rejet sera de la forme

- $\mathcal{I} =]t_r, +\infty[$ pour le test unilatéral à droite, pour lequel $\mathcal{H}_1 : \theta > \theta_0$;
- $\mathcal{I} =]-\infty, t_l[$ pour le test unilatéral à gauche, pour lequel $\mathcal{H}_1 : \theta > \theta_0$;
- $\mathcal{I} =]-\infty, t_l[\cup]t_r, +\infty[$ pour le test bilatéral, pour lequel $\mathcal{H}_1 : \theta \neq \theta_0$.

On cherchera souvent à utiliser une statistique de test symmétrique, de sorte à pouvoir utiliser $t_r = -t_l$. Dans ce cas $t_0 = t_t$ est appelée **valeur critique** du test et est telle que

- $\mathbb{P}_{T0}(T > t_0) = \alpha$ pour le test unilatéral à droite ;
- $\mathbb{P}_{T0}(T < -t_0) = \alpha$ pour le test unilatéral à gauche ;
- $\mathbb{P}_{T0}(|T| > t_0) = \alpha$ pour le test bilatéral.

4.2.5 p-valeur

La **p-valeur** (*p-value* en anglais) d'un test statistique est définie dans le cas où le test statistique peut être réalisé en comparant une statistique de test T [1] à une valeur critique t_0.

Dans ce contexte, étant donné un échantillon $(x_1, x_2, \ldots, x_n)$ et la réalisation t de T sur cet échantillon, on appelle **p-valeur** la probabilité $\mathbb{P}_{T0}(T > t)$ pour un test unilatéral à droite (respectivement, $\mathbb{P}_{T0}(T < -t)$ pour un test unilatéral à gauche, et $\mathbb{P}_{T0}(|T| > t)$ pour un test bilatéral).

L'hypothèse nulle est rejetée si la p-valeur est plus petite que le niveau de signification. On dit alors que la p-valeur est **significative.**

En d'autres termes, la p-valeur peut être interprétée comme la probabilité d'obtenir, sous l'hypothèse nulle, un résultat au moins aussi extrême que celui observé.

On rapporte ainsi généralement comme résultat d'un test non pas la statistique de test réalisée sur l'échantillon observé, mais la p-valeur correspondante.

On lira ainsi dans des publications scientifiques des assertions suivies de « $(p < 0{,}05)$ », signifiant que l'assertion en question est l'hypothèse alternative d'un test dont l'hypothèse nulle a été rejetée avec une p-valeur inférieure à 5%.

---

1. ou sa valeur absolue $|T|$, sans perte de généralité, puisque l'on peut alors utiliser la statistique $U = |T|$.

Exemple

Le test que nous avons défini dans l'exemple de la pièce de monnaie s'appelle le test binomial. Il est implémenté dans `scipy.stats` :

```
t = 5 # nb pile
n = 5 # taille échantillons
pi = 0.5
import scipy.stats as st
st.binom_test(t, n, pi, alternative='greater') # unilatéral à droite
```

Attention

On fera attention à ne pas sur-interpréter la p-valeur. En particulier, la p-valeur *n'est pas* la probabilité que l'hypothèse nulle soit vraie : $\mathbb{P}(t|\mathcal{H}_0) \neq \mathbb{P}(\mathcal{H}_0|t)$.

4.2.6 Erreurs de première et deuxième espèce

Deux types d'erreurs sont possibles quand on fait un test d'hypothèse :

- Rejeter l'hypothèse nulle alors qu'elle est correcte : on parle d'une **erreur de première espèce**, ou **erreur de Type I** (*Type I error* en anglais).
- Accepter l'hypothèse nulle alors qu'elle est en fait fausse : on parle d'une **erreur de deuxième espèce**, ou **erreur de Type II** (*Type II error* en anglais).

Moyen mnémotechnique Ces deux types d'erreurs sont numérotés dans le même ordre que dans l'histoire du garçon qui criait au loup : d'abord les villageois pensaient qu'il y avait un loup alors qu'il n'y en avait pas (erreur de première espèce), mais à la fin les villageois pensaient qu'il n'y avait pas de loup alors qu'il y en avait un (erreur de deuxième espèce). Ici, l'hypothèse nulle est l'hypothèse correspondant à l'état « par défaut » du village, à savoir sans loup [2].

Le niveau de signification α est ainsi la probabilité de commettre une erreur de première espèce.

La probabilité de commettre une erreur de deuxième espèce est généralement noté β. La probabilité de rejeter $\mathcal{H}_0$ à raison, $1 - \beta$, est appelée la **puissance** du test (*power* en anglais).

4.3 Comparaison d'une moyenne observée à une moyenne théorique

Dans cette section, nous allons dérouler un autre exemple de test statistique.

Nous souhaitons tester l'hypothèse selon laquelle les pigeons du Jardin du Luxembourg ont un poids moyen de 300g. Nous disposons de mesures pour 40 pigeons, capturés et pesés par des élèves de l'École, dont la moyenne est de 312g et l'écart-type 31g.

Définissons une variable aléatoire réelle X de carré intégrable. X modélise le poids d'un pigeon. Posons μ l'espérance de X et σ^2 sa variance.

2. Les fans de *Battlestar Galactica* pourront construire leur propre moyen mnémotechnique à partir de Starbuck qui refuse de faire valider les élèves pilotes après avoir fait valider Zak.

4.3.1 Hypothèses de test

Question : Comment modéliser ce problème ? Que poser pour $\mathcal{H}_0$ et $\mathcal{H}_1$?

Réponse

Nous posons $n = 40$; les poids des 40 pigeons, $(x_1,x_2,\ldots,x_n)$, sont la réalisation de l'échantillon aléatoire $(X_1,X_2,\ldots,X_n)$ composé de variables aléatoires indépendantes et identiquement distribuées de même loi que X.

Nous posons l'hypothèse nulle à tester

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

avec $\mu_0 = 300\text{g}$.

Nous n'avons aucun a priori sur le poids des pigeons du Jardin du Luxembourg, et formulons donc l'hypothèse alternative bilatérale

$$\mathcal{H}_1 : \mu \neq \mu_0.$$

4.3.2 Statistique de test

Pour tester $\mathcal{H}_0$, nous souhaitons déterminer la probabilité d'observer une moyenne empirique $\widehat{m}$ de 312g si l'espérance de X est de 300g. En posant M_n la moyenne empirique de l'échantillon, nous souhaitons déterminer $\mathbb{P}(M_n = \widehat{m}|\mu = \mu_0)$.

Le théorème central limite nous indique que

$$\frac{\sqrt{n}(M_n - \mu)}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Avec $n = 40$, nous pouvons supposer que cette limite est suffisamment proche d'être atteinte pour poser

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

Nous ne connaissons pas la variance σ^2 de X ; cependant nous pouvons l'estimer grâce à l'écart-type empirique $\widehat{\sigma} = 31\text{g}$, et utiliser

$$\frac{\sqrt{n}(M_n - \mu)}{\widehat{\sigma}} \sim \mathcal{N}(0,1). \tag{4.5}$$

Nous ne remplaçons pas μ par son estimation $\widehat{m}$: ce n'aurait aucun sens, car nous cherchons justement à tester sa valeur.

Question : Comment utiliser l'équation (4.5) pour définir une statistique de test ?

Réponse

Si $\mathcal{H}_0$ est vraie, alors $\mu = \mu_0$ et la variable aléatoire réelle

$$Z = \frac{\sqrt{n}(M_n - \mu_0)}{\widehat{\sigma}} \tag{4.6}$$

est une gaussienne standard : $Z \sim \mathcal{N}(0,1)$.

Z est donc une variable aléatoire réelle dont nous connaissons la distribution sous l'hypothèse nulle. Cela en fait une bonne candidate à être statistique de test.

Sous $\mathcal{H}_0$, on s'attend à ce que la réalisation de Z sur l'échantillon observé soit proche de 0. Nous réalisons un test bilatéral et n'avons aucun a priori sur le signe de Z. Ainsi, nous rejetterons $\mathcal{H}_0$ si la réalisation de $|Z|$ est « trop grande » pour être plausible.

Le test statistique permettant de tester si la moyenne d'un échantillon vaut une valeur prédéterminée consiste donc à rejeter $\mathcal{H}_0$ si $|Z| > z_0$.

Question : Étant donné un niveau de signification α, quelle est la valeur critique z_0 ?

Réponse

Nous souhaitons que la probabilité de rejeter $\mathcal{H}_0$ alors qu'elle est vraie soit égale à α. En d'autres termes, nous cherchons z_0 tel que

$$\mathbb{P}(|Z| > z_0) = \alpha, \text{ sachant } Z \sim \mathcal{N}(0,1).$$

La densité de Z étant symétrique, on cherche donc z_0 telle que

$$\mathbb{P}(Z < -z_0) = \frac{\alpha}{2}.$$

Ceci est illustré sur la figure 4.1.

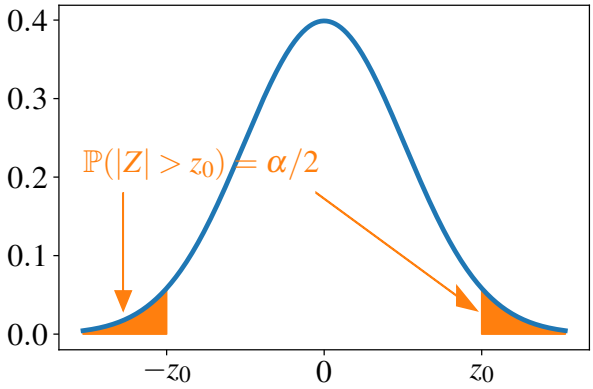


FIGURE 4.1 – Densité d'une gaussienne centrée réduite. L'aire colorée vaut α et correspond au domaine de rejet du test.

Question : Peut-on rejeter l'hypothèse selon laquelle les pigeons du jardin du Luxembourg ont un poids moyen de 300g ?

Réponse

Cela dépend du niveau de signification que l'on choisit.

Calculons tout d'abord la réalisation de la statistique de test Z sur notre échantillon : $z = 2{,}45$.

Posons $\alpha = 0{,}05$. Alors $z_0 \approx 1{,}96$. On a bien $z > z_0$ et on rejette l'hypothèse nulle. On dit que l'écart entre M_n et μ_0 est **statistiquement significatif.**

Posons maintenant $\alpha = 0{,}01$. Le domaine de rejet est plus restreint ; $z_0 \approx 2{,}58$. On ne peut pas rejeter l'hypothèse nulle. L'écart entre M_n et μ_0 n'est pas statistiquement significatif.

Cet exemple est illustré sur la figure 4.2.

4.3.3 p-valeur

Question : Quelle est la p-valeur correspondant à la valeur de test $z = 2{,}45$?

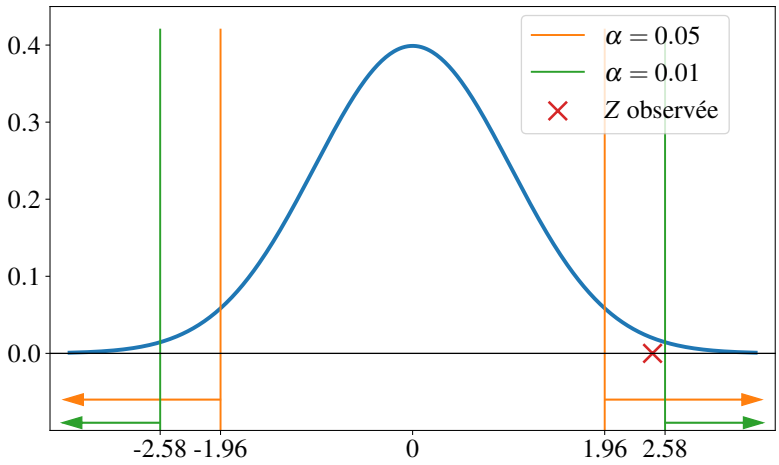


FIGURE 4.2 – Densité d'une gaussienne centrée réduite. La valeur $z = 2{,}45$ est dans le domaine de rejet pour $\alpha = 0{,}05$ mais pas pour $\alpha = 0{,}01$.

Réponse

La p-valeur est

$$\mathbb{P}(|Z|) \geq |z| = \mathbb{P}(Z \leq -|z|) + \mathbb{P}(Z \geq |z|) = 2\mathbb{P}(Z \leq -|z|) = 2\Phi(-|z|) = 0{,}018.$$

où Φ est la fonction de répartition d'une gaussienne standard.

Cette p-valeur est bien inférieure au seuil de signification $\alpha = 0{,}05$, mais supérieure au seuil de signification $\alpha = 0{,}01$.

4.3.4 Test unilatéral à droite

Supposons maintenant que nous nous demandons si les pigeons du Jardin du Luxembourg, qui nous semblent particulièrement bien nourris de restes des sandwicheries environnantes, ne seraient pas plus lourds que la moyenne de 300g. Il s'agit maintenant de faire un test unilatéral à droite, pour lequel

$$\mathcal{H}_1 : \mu > \mu_0.$$

Question : Comment cela transforme-t-il notre test d'hypothèse ?

Réponse

Le test statistique consiste maintenant à rejeter $\mathcal{H}_0$ si $Z > z_r$ (sans valeur absolue). En particulier, toutes les valeurs négatives nous font accepter $\mathcal{H}_0$, contrairement au cas bilatéral.

La valeur critique z_r est telle que

$$\mathbb{P}(Z > z_r) = \alpha, \text{ sachant } Z \sim \mathcal{N}(0{,}1).$$

La densité de Z étant symmétrique, on cherche donc z_0 telle que

$$\Phi(-z_r) = \alpha.$$

Pour $\alpha = 0{,}05$, la valeur critique est $z_r = 1{,}64$. Pour $\alpha = 0{,}01$, la valeur critique est $z_r = 2{,}33$. L'hypothèse nulle est rejetée dans les deux cas.

Le test unilatéral est plus puissant pour les valeurs du bon côté.

Cet exemple est illustré sur la figure 4.3.

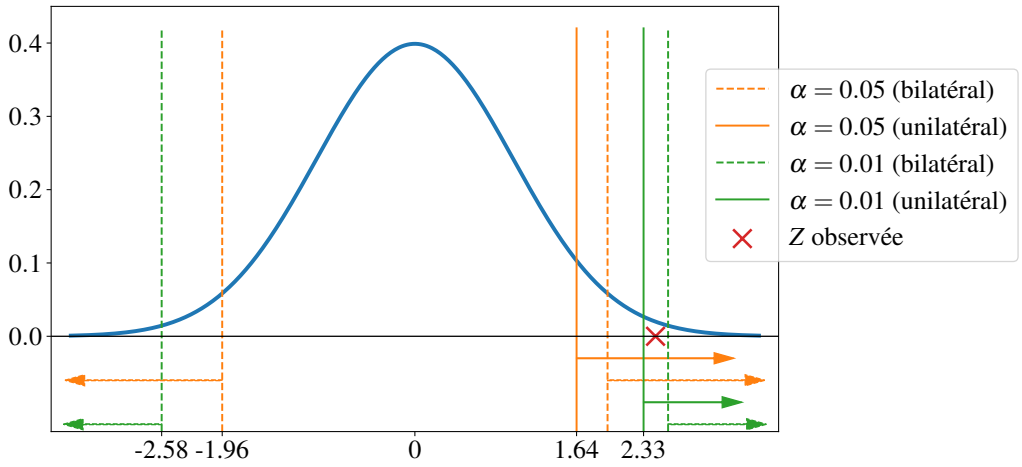


FIGURE 4.3 – Densité d'une gaussienne centrée réduite. La valeur $z = 2{,}45$ est dans le domaine de rejet pour $\alpha = 0{,}05$ et pour $\alpha = 0{,}01$ dans le cas du test unilatéral.

4.3.5 Intervalle de confiance ★

Reprenons le test bilatéral.

Étant donné α, nous avons déterminé z_0 de sorte à ce que

$$\mathbb{P}(|Z| > z_0) = \alpha, \text{ sachant } Z \sim \mathcal{N}(0{,}1).$$

En d'autres termes,

$$\mathbb{P}(-z_0 \leq Z \leq z_0) = 1 - \alpha.$$

(On pourra se référer à la figure 4.1.)

D'après la définition de Z (équation (4.6)), cela est équivalent à

$$\mathbb{P}\left(M_n - \frac{\widehat{\sigma}}{\sqrt{n}} z_0 \leq \mu \leq M_n + \frac{\widehat{\sigma}}{\sqrt{n}} z_0\right) = 1 - \alpha.$$

Ainsi l'intervalle

$$\left[M_n - \frac{\widehat{\sigma}}{\sqrt{n}} z_0,\ M_n + \frac{\widehat{\sigma}}{\sqrt{n}} z_0\right]$$

est un **intervalle de confiance** à $(1 - \alpha)$ pour la taille moyenne μ (voir Probabilités V).

Dans notre exemple, l'intervalle de confiance à 95% pour la valeur moyenne du poids d'un pigeon est $[302{,}4\text{g} ; 321{,}6\text{g}]$. $\mu_0 = 300\text{g}$ n'est pas dans l'intervalle de confiance ; on adopte l'hypothèse alternative selon laquelle $\mu \neq \mu_0$.

L'intervalle de confiance à 99% est $[299{,}4\text{g} ; 324{,}6\text{g}]$. Cet intervalle contient μ_0. On ne peut pas rejeter l'hypothèse nulle.

Exercice : Calculer l'intervalle de confiance à 95% et à 99% pour le test d'hypothèse unilatéral à droite. (Solution : cf section 4.5.1.)

4.3.6 Tests de comparaison de moyenne ★

Le test que nous avons étudié dans cette section, qui permet de comparer la moyenne d'un échantillon suffisamment large pour être dans la limite du théorème centrale limite ($n \geq 30$) à sa moyenne

théorique, s'appelle un **test Z**, ou *Z-test* en anglais, par référence à la notation Z couramment utilisée pour une variable normalement distribuée de moyenne 0 et variance 1.

Dans le cas d'un échantillon de faible taille, le théorème central limite ne s'applique pas. Si l'on suppose X normalement distribuée, on peut alors appliquer un test de Student, ou test t (*t-test* en anglais), ainsi appelé car la statistique de test suit une loi de Student.

Des variantes de ces tests Z et t peuvent aussi être utilisés pour comparer les moyennes de deux échantillons, appariés ou non. On dit que deux échantillons aléatoires $(X_1,X_2,\ldots,X_n)$ et $(Y_1,Y_2,\ldots,Y_n)$ sont appariés quand les variables X_i et Y_i décrivent le même individu i. Il peut par exemple s'agir de mesures répétées sur les mêmes individus, soit prises par deux appareils différents, soit prises avant et après un traitement.

4.4 Tests d'hypothèses multiples

Question Imaginons l'expérience de pile ou face suivante : je lance 15 fois une pièce équilibrée, et demande à une personne en face de moi de prédire avant chaque lancer si je vais obtenir pile ou face. Supposons que cette personne me donne la bonne réponse 12 fois. A-t-elle un don de voyance ?

Réponse

Pour répondre à cette question, posons X une variable de Bernouilli de paramètre p modélisant, pour un lancer de pièce, le succès de la personne : X vaut 0 si la personne n'a pas donné la bonne prédiction et 1 sinon. L'hypothèse nulle est

$$\mathcal{H}_o : p = 0{,}5 \text{ (la personne n'a pas de don de voyance).}$$

Nous pouvons ici poser une hypothèse alternative unilatérale à droite :

$$\mathcal{H}_1 : p > 0{,}5 \text{ (la personne a un don de voyance).}$$

Il s'agit du test binomial que nous avons défini dans la section 4.2, mais ici la variable modélise non pas le résultat du lancer de pièce mais la correction de la réponse donnée par mon cobaye.

La statistique de test T est le nombre de succès dans l'échantillon. Sous $\mathcal{H}_0$, $T \sim \mathcal{B}(n,p)$. La p-valeur correspondant à 12 succès est donc $\mathbb{P}_{\mathcal{H}_0}(T \geq 12) = 1 - \mathbb{P}(T \leq 11) = 0{,}018$. Cette p-valeur est significative pour $\alpha = 5\%$.

Question Supposons maintenant que je fasse ce test avec toute la classe. Vous êtes derrière votre ordinateur et ne communiquez pas entre vous. Trois élèves passent mon test de pyschisme, autrement dit tombent juste au moins 12 fois sur 15. Dois-je appeler la presse ?

Réponse

Supposons une promo de $m = 125$ élèves. Nous posons maintenant Y une variable de Bernouilli de paramètre π modélisant le succès d'une personne sur 15 lancers. Nous faisons ici un nouveau test statistique sur Y,

$$\mathcal{H}'_0 : \pi = 0{,}018 \qquad \text{et} \qquad \mathcal{H}'_1 : \pi > 0{,}018.$$

Il s'agit toujours d'un test binomial. La statistique de test U est le nombre d'élèves passant le test. Sous $\mathcal{H}'_0$, $U \sim \mathcal{B}(m,\pi)$. La p-valeur est ici $\mathbb{P}_{\mathcal{H}'_0}(T \geq 3) = 1 - \mathbb{P}(T \leq 2) = 0{,}39$. Cette p-valeur n'est pas significative !

Cet exemple illustre le principe suivant : plus on fait de tests, et plus on a de chances de voir apparaître une p-valeur significative.

Il est nécessaire de corriger cet effet : on parle de **correction** ou **ajustement de tests d'hypothèse multiples.** La plus simple et plus utilisées de ces corrections, proposée par la biostatisticienne Olive

Jean Dunn, est connue sous le nom de **correction de Bonferroni** : il s'agit simplement de diviser le niveau de signification par le nombre de tests

$$\alpha \leftarrow \frac{\alpha}{m}.$$

Cette correction se justifie de la façon suivante : notons p_1,p_2,$\ldots$,p_m les p-valeurs obtenue pour m tests, testant chacun $\mathcal{H}_0$ vs. $\mathcal{H}_1$, et supposons que $\mathcal{H}_0$ est vraie pour les m_0 premiers tests . Alors

$$\mathbb{P}(\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right) \leq \sum_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right) = \frac{m_0\alpha}{m} \leq \alpha.$$

4.5 Compléments

4.5.1 Solution de l'exercice section 4.3.5 ★

Nous avons déterminé la valeur critique z_r de sorte à ce que

$$\mathbb{P}(Z \leq z_r) = 1 - \alpha.$$

Comme $Z = \frac{\sqrt{n}(M_n - \mu_0)}{\widehat{\sigma}}$, cela est équivalent à

$$\mathbb{P}\left(\mu \geq M_n - \frac{\widehat{\sigma}}{\sqrt{n}}z_r\right) = 1 - \alpha.$$

Ainsi l'intervalle

$$\left[M_n - \frac{\widehat{\sigma}}{\sqrt{n}}z_r, +\infty\right]$$

est un intervalle de confiance unilatéral à $(1-\alpha)$ pour la taille moyenne μ.

Dans notre exemple, l'intervalle de confiance unilatéral à droite à 95% pour la valeur moyenne du poids d'un pigeon est $[303{,}9\text{g}, +\infty]$. Cet intervalle contient μ_0 et on ne peut pas rejeter l'hypothèse nulle. À 99%, cet intervalle est $[300{,}6\text{g}, +\infty[$. Ces résultats sont cohérents.

Pour aller plus loin

- On dit d'un test statistique qu'il est sans biais si sa puissance est supérieure au niveau de signification : $1 - \beta > \alpha$.
- On dit d'un test statistique qu'il converge si la suite des erreurs de deuxième espèce converge vers 0 : $1 - \beta \xrightarrow[n\to+\infty]{} 0$.
- Pour plus de détails sur la sur-interprétation des p-valeurs en sciences et le *p-hacking* (consistant à ne conserver, parmi de nombreux tests conduits sur les données, ceux qui donnent des p-valeurs significatives), on pourra se reporter aux références suivantes.

[1] John PA Ioannidis. Why most published research findings are false. *PLoS medicine,* 2(8) :e124, 2005.

[2] Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn and Michael D Jennions. The extent and consequences of p-hacking in science. *PLoS biology*, 13(3) :e1002106, 2015.

[3] Ronald L Wasserstein, Nicole A Lazar, et al. The ASA's statement on p-values : context, process, and purpose. *The American Statistician*, 70(2) :129–133, 2016.

[4] Susan Holmes. Statistical proof ? The problem of irreproducibility. *Bulletin (New Series) of the American Mathematical Society*, 55(1), 2018.