Eléments de statistiques pour les data sciences

Cours 1 : Introduction, survol historique, organisation du cours

Dr. Ph. Müllhaupt

IGM - EPFL

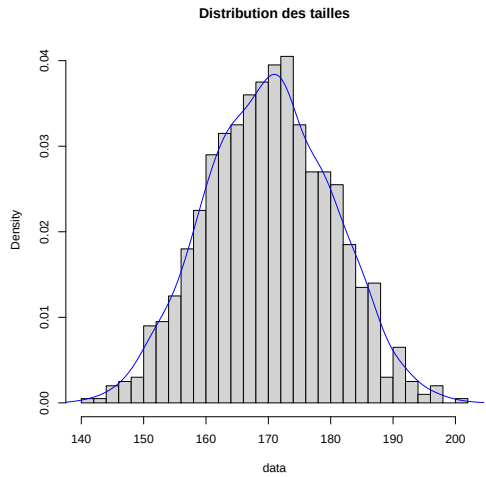
Plan

Qu'est-ce que les statistiques ?

- Les statistiques sont la science de la collecte, de l'analyse et de l'interprétation des données.
- Deux branches principales :
  1. Statistiques descriptives
  2. Statistiques inférentielles

Exemple : Visualisation des données

- Données : Distribution des tailles dans une population
- Méthode : Histogramme et courbe de densité



Conclusion

- Les statistiques sont un outil puissant pour décrire l'incertitude et appréhender les phéonomènes incertains.
- Sous de bonnes hypothèses, permet de prendre des décisions raisonnées (mais pas toujours raisonnables).
- La visualisation des données est souvent utile pour leur interprétation.

Organisation

pour le semestre, en 14 leçons et exercices

1. Introduction
2. Probabilité discrète
3. Probabilité continue
4. Loi du χ^2 de Pearson
5. Statistique t de Student
6. Estimateurs I : la vraisemblance
7. Estimateurs II : biais et minimum de variance
8. Régression linéaire — loi des grands nombres
9. Tests I : Neynman-Pearson
10. Tests II : intervalles de confiance
11. Estimation bayésienne — a priori
12. Tests non paramétriques
13. Test non paramétrique de Wilcoxon
14. Introduction aux méthodes Monte-Carlo

Continuation

préparation pour...

- classification, apprentissage supervisé et non supervisé
- processus stochastique, chaînes de Markov, processus de Levy
- mécanique satistique, ergodicité,
  - (i) entropie topologique (de Kolmogorov-Sinai–systèmes dynamiques déterministes)
  - (ii) entropie statistique (de Boltzmann–thermodynamique, de Shannon–information)
  - (iii) entropie de Rényi
- statistique avancée

Ces sujets sont plus accessibles au niveau Master/PhD que Bacchelor, une fois les bases de probabilité et statistiques solidement acquises, ce qui est un objectif du cours.

Pierre de Fermat

(1601-1665)

Pierre de Fermat (1601–1665) était un mathématicien français, célèbre pour ses contributions majeures aux mathématiques, notamment dans les domaines de l'arithmétique, de la géométrie et de la théorie des probabilités. Fermat est considéré comme l'un des fondateurs de la théorie des probabilités, notamment grâce à sa collaboration avec Blaise Pascal dans les années 1650, qui a posé les bases de la théorie moderne des jeux de hasard.

Pierre de Fermat

contributions

Avec Pascal, Fermat a résolu des problèmes pratiques concernant la répartition des gains dans les jeux de hasard, donnant naissance à la théorie des probabilités. Leur célèbre correspondance, qui portait sur la question du partage équitable d'une mise dans un jeu de dés interrompu, a permis d'établir des concepts essentiels comme la probabilité, l'espérance mathématique et la valeur des mises en fonction des probabilités des événements. Ce travail est souvent considéré comme l'origine de la théorie des probabilités.

Fermat a également introduit la notion de probabilité conditionnelle, bien qu'il n'ait pas formalisé le concept comme on le fait aujourd'hui. Son approche des probabilités reposait sur des principes intuitifs liés à l'équité et à l'incertitude, influençant la manière dont les gens ont commencé à aborder le calcul des chances dans des situations incertaines.

Blaise Pascal

(1623-1662)

Blaise Pascal (1623–1662) était un mathématicien, physicien et philosophe français, l'un des précurseurs des probabilités modernes. Bien que sa carrière ait été marquée par de nombreuses découvertes dans les sciences physiques et la géométrie, il est particulièrement célèbre pour ses contributions à la théorie des probabilités et à la résolution des jeux de hasard, un domaine dans lequel il a jeté les bases des concepts probalistes modernes.

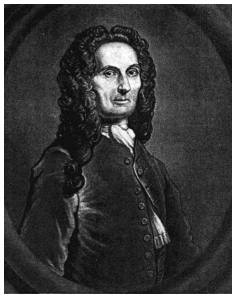
Jacob Bernouilli

(1655-1705)

Jacob Bernoulli (1655-1705) était un mathématicien suisse pionnier du calcul des probabilités, connu pour la loi des grands nombres et la loi de Bernoulli. Son ouvrage Ars Conjectandi a jeté les bases de la théorie moderne des probabilités, et la variable aléatoire de Bernoulli, qui modélise un essai à deux issues (succès ou échec), porte son nom.

Abraham de Moivre

(1667-1754)



Abraham de Moivre (1667–1754) était un mathématicien franco-britannique, pionnier des probabilités et de la théorie des jeux. Né à Paris, il s'installe en Angleterre où il devient membre de la Royal Society. Il est surtout connu pour sa contribution fondamentale à la théorie des probabilités, notamment à travers sa formule de Moivre, qui établit un lien entre les puissances binomiales et les distributions normales. Il a écrit The Doctrine of Chances, un ouvrage marquant dans le domaine des probabilités.

Thomas Bayes

(1701-1761)

Thomas Bayes (1701–1761) était un mathématicien et statisticien anglais, dont le théorème de Bayes joue un rôle fondamental dans le domaine des probabilités et de la statistique. Bayes a introduit une approche probabiliste de l'inférence inductive, permettant de réviser les croyances ou hypothèses à la lumière de nouvelles données.

Thomas Bayes

contributions et influences

Le théorème de Bayes fournit un cadre formel pour réévaluer les probabilités d'une hypothèse à partir d'observations ou de données nouvelles, ce qui en fait un outil essentiel dans de nombreux domaines, notamment les statistiques, l'intelligence artificielle, l'apprentissage automatique (machine learning), et la prise de décision sous incertitude. Bien que Bayes n'ait pas publié ses travaux de manière aussi étendue de son vivant, son théorème a été redécouvert et largement développé au XIXe siècle, et il reste aujourd'hui un pilier de la statistique et de la logique inductive.

Pierre-Simon Laplace

(1749-1827)

Pierre-Simon Laplace (1749–1827) était un mathématicien et astronome français, l'une des figures les plus importantes de la science au XVIIIe siècle. Ses travaux couvrent un large éventail de domaines, notamment les probabilités, la mécanique céleste et la statistique. En probabilités, il a développé la théorie des erreurs, formulé la loi de Laplace et approfondi les travaux sur la distribution normale, contribuant ainsi à la fondation des statistiques modernes. Son ouvrage majeur, Théorie analytique des probabilités, a eu une influence profonde sur la manière dont les probabilités sont comprises et appliquées. En astronomie, ses travaux sur les mouvements des planètes et des satellites ont été essentiels pour la compréhension de la mécanique céleste.

Carl Friedrich Gauss

(1777-1855)

Carl Friedrich Gauss (1777–1855) était un mathématicien et astronome allemand, considéré comme l'un des plus grands génies de l'histoire des sciences. Bien que ses contributions couvrent un large éventail de domaines, notamment l'astronomie, la géométrie, l'algèbre et l'analyse, Gauss est également une figure centrale dans le développement de la théorie des probabilités.

Carl Friedrich Gauss

contributions

Gauss est particulièrement reconnu pour avoir introduit la distribution normale, parfois appelée la "courbe de Gauss", qui est devenue la distribution de probabilité la plus importante et la plus largement utilisée en statistique. Il a démontré que de nombreuses variables aléatoires, sous certaines conditions, suivent cette distribution, un concept fondamental dans la théorie des probabilités et les statistiques modernes. La distribution normale est utilisée pour modéliser une grande variété de phénomènes naturels et sociaux, de la hauteur des individus à l'erreur de mesure en passant par les fluctuations financières.

En outre, Gauss a développé la méthode des moindres carrés, une technique statistique qui permet d'estimer les paramètres d'un modèle de régression en minimisant la somme des carrés des erreurs. Cette méthode est devenue un outil de base en statistique et en économétrie.

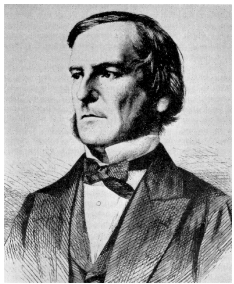
Augustus De Morgan

(1806-1871)

Augustus De Morgan (1806–1871) était un mathématicien et logicien britannique, connu pour ses contributions à la logique formelle et aux probabilités. Il est célèbre pour avoir formulé les lois de De Morgan, qui sont des règles fondamentales de la logique booléenne, et pour avoir joué un rôle clé dans la formalisation des systèmes logiques. En probabilités, il a contribué à la diffusion de la théorie des probabilités en Angleterre et a travaillé sur des concepts comme les événements indépendants et les probabilités conditionnelles. De Morgan a également été l'un des premiers à appliquer les idées de la logique à des problèmes mathématiques pratiques, influençant ainsi la logique moderne et la philosophie des mathématiques.

George Boole

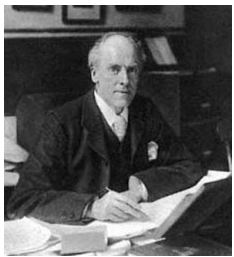
(1815-1864)



George Boole (1815–1864) était un mathématicien et logicien britannique, considéré comme le fondateur de l'algèbre logique, une branche fondamentale des mathématiques modernes. Son ouvrage majeur, The Laws of Thought (1854), a introduit des concepts qui ont transformé la logique en un système algébrique. Boole a formulé des règles algébriques pour les opérations logiques, créant ainsi la base de la logique booléenne, utilisée dans la conception des circuits électroniques et des ordinateurs modernes. En plus de ses contributions en logique, Boole a également travaillé sur la théorie des probabilités, où il a appliqué ses méthodes algébriques pour traiter les problèmes de probabilité, influençant le développement ultérieur de la statistique et de l'informatique.

Karl Pearson

(1857-1936)



Karl Pearson (1857–1936) était un statisticien britannique, l'une des figures les plus influentes dans le développement de la statistique moderne. Il est surtout connu pour avoir introduit des concepts fondamentaux tels que le coefficient de corrélation de Pearson, qui mesure la force et la direction de la relation linéaire entre deux variables, et la chi-carré, un test statistique largement utilisé pour l'analyse de la relation entre des variables catégorielles.

Karl Pearson

contributions

Dans le domaine de la probabilité et des statistiques, Pearson a également contribué à l'élargissement de la théorie des distributions. Il a travaillé sur la formulation de la distribution de Pearson, une famille de distributions de probabilité qui inclut certaines distributions importantes telles que la normale et la gamma. Pearson a par ailleurs introduit le concept de moments statistiques pour décrire les caractéristiques de distribution, comme la moyenne, la variance, et la courbure, contribuant ainsi au développement des méthodes statistiques appliquées.

En plus de ses travaux théoriques, Pearson a été un pionnier de l'application de la statistique dans les sciences sociales et naturelles, notamment en génétique et en biologie, où il a promu l'utilisation des méthodes statistiques pour analyser les phénomènes observés et établir des relations quantitatives fiables.

Ronald A. Fisher

(1890-1962)

Ronald A. Fisher (1890–1962) était un statisticien, généticien et biologiste britannique, considéré comme l'un des fondateurs de la statistique moderne. Il est particulièrement connu pour ses travaux sur la conception expérimentale, l'analyse de variance (ANOVA) et la régression linéaire, qui sont devenus des outils essentiels en statistique. Fisher a également formulé le test de signification statistique qui porte son nom, le test de Fisher, et a introduit la notion de vraisemblance maximale, un concept clé dans les estimations statistiques. En génétique, ses recherches ont jeté les bases de la génétique quantitative. Son ouvrage majeur, The Genetical Theory of Natural Selection (1930), a combiné la théorie de l'évolution avec des concepts statistiques, influençant profondément la biologie évolutive.

Richard Edcumbe Snedecor

(1881-1972)

Richard Edgcumbe Snedecor (1881–1972) était un statisticien américain de renom, largement reconnu pour ses contributions à la théorie statistique et à son application dans divers domaines scientifiques. Il est surtout connu pour ses travaux sur la distribution F, une distribution de probabilité qui joue un rôle fondamental dans l'analyse de la variance (ANOVA) et dans les tests statistiques comparant les variances entre plusieurs groupes. Cette distribution est aujourd'hui un outil essentiel dans la statistique appliquée, notamment en biologie, économie et sciences sociales.

Richard Edcumbe Snedecor

contributions

Snedecor est également l'auteur de l'ouvrage majeur Statistical Methods (1937), qui a eu une influence considérable sur le développement des méthodes statistiques, en particulier dans les applications agricoles et en recherche scientifique. Cet ouvrage est devenu une référence dans le domaine des statistiques et a contribué à l'adoption des méthodes statistiques dans les sciences appliquées.
De plus, Snedecor a joué un rôle crucial dans la diffusion de l'analyse de la variance (ANOVA), une méthode qui permet d'étudier les différences entre plusieurs groupes à partir d'échantillons. Il a aussi élargi les applications des statistiques à la conception d'expériences, en particulier dans l'agriculture, ce qui a permis de transformer la manière dont les données étaient collectées et analysées pour prendre des décisions informées dans divers secteurs.

William Sealy Gosset

(1876-1937)

William Sealy Gosset (1876–1937), mieux connu sous le pseudonyme de "Student", était un statisticien britannique, célèbre pour ses contributions à la théorie statistique, notamment dans le développement de la distribution t de Student. Travaillant pour la brasserie Guinness, Gosset a conçu cette distribution pour traiter les problèmes statistiques liés à de petits échantillons, où la distribution normale ne pouvait pas être utilisée de manière fiable. Gosset a aussi contribué à la mise au point de méthodes pour l'analyse de la variance et à la conception d'expériences, influençant ainsi de manière significative les pratiques statistiques modernes, notamment en recherche scientifique et industrielle.

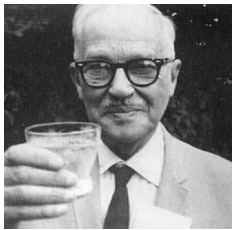
Prasanta Chandra Mahalanobis

(1893-1972)

Prasanta Chandra Mahalanobis (1893-1972) était un statisticien indien connu pour la distance de Mahalanobis, une mesure clé en analyse multivariée, et pour ses contributions à l'échantillonnage et à la planification statistique. La distance de Mahalanobis est largement utilisée en apprentissage automatique (machine learning) et data science pour la détection d'anomalies, la classification et l'analyse de données multivariées.

Jerzy Neyman

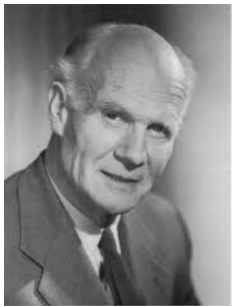
(1894-1981)



Jerzy Neyman (1894–1981) était un statisticien polonais-américain, l'une des figures clés du développement de la statistique au XXe siècle. Il est particulièrement connu pour ses contributions à la théorie des tests d'hypothèses et à la méthode de l'intervalle de confiance, en collaboration avec Egon Pearson. Neyman a formulé le principe du test d'hypothèse de manière rigoureuse, en introduisant les concepts de erreur de type I et erreur de type II, qui sont devenus essentiels pour la pratique statistique. Il a également développé la méthode des plans d'expérience (design expérimentaux), une approche permettant de structurer les expériences pour obtenir des résultats fiables et précis.

Egon Pearson

(1895-1980)



Egon Pearson (1895–1980) était un statisticien britannique, connu pour ses contributions majeures à la théorie des probabilités et des statistiques, en particulier dans le domaine des tests d'hypothèses. Il est surtout célèbre pour son travail avec Jerzy Neyman sur la formulation du cadre des tests statistiques, notamment la distinction entre les erreurs de type I et de type II. Ensemble, ils ont développé la méthode de test d'hypothèses et l'idée d'intervalle de confiance. Pearson a également joué un rôle clé dans le développement du coefficient de corrélation de Pearson, une mesure de la force et de la direction de la relation linéaire entre deux variables.

Rudolf Carnap

(1891-1970)

Rudolf Carnap (1891–1970) était un philosophe et logicien allemand, associé au mouvement du positivisme logique et à l'école de Vienne. Il a grandement contribué à la philosophie des sciences et à la logique formelle, en particulier à travers ses travaux sur la logique inductive et la probabilité. Carnap a cherché à formaliser le raisonnement inductif, un processus fondamental dans les sciences empiriques, qui repose sur l'inférence de généralisations à partir de données limitées.

Rudolf Carnap

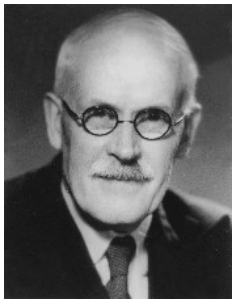
contributions

Dans son ouvrage majeur The Logical Foundations of Probability (1950), Carnap a exploré les bases logiques de la probabilité, en introduisant une conception probabiliste de la logique inductive. Il a proposé un cadre où la probabilité était vue non seulement comme une mesure des croyances rationnelles, mais aussi comme un outil pour formaliser l'induction scientifique, permettant ainsi de traiter l'incertitude dans les généralisations basées sur les observations. Carnap a intégré des modèles probabilistes dans un système logique, reliant ainsi les méthodes formelles de la logique avec les probabilités pour rendre les inférences inductives plus rigoureuses et systématiques.

Son travail sur la probabilité a jeté les bases de la logique probabiliste moderne et influencé le développement de théories contemporaines sur l'incertitude, la logique inductive et la modélisation probabiliste dans les sciences.

Harold Jeffreys

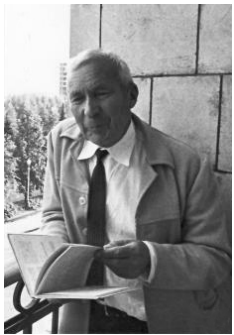
(1891-1989)



Harold Jeffreys (1891–1989) était un mathématicien, statisticien et géophysicien britannique, reconnu pour ses travaux pionniers sur l'inférence bayésienne et ses contributions aux fondements de la théorie des probabilités. Jeffreys a été un ardent défenseur de l'approche bayésienne des probabilités, insistant sur l'interprétation subjective de la probabilité comme un degré de croyance qui se met à jour avec de nouvelles preuves. Il est particulièrement célèbre pour avoir développé le prior de Jeffreys, une distribution a priori objective utilisée dans l'analyse bayésienne, qui reste un outil fondamental pour l'estimation des paramètres lorsqu'il y a peu d'informations a priori.

Andrey Kolmogorov

(1903-1987)



Andrey Kolmogorov (1903-1987) était un mathématicien russe pionnier des probabilités, ayant formulé les bases axiomatiques de la théorie des probabilités moderne en 1933. Son travail a permis de formaliser les probabilités en tant que branche rigoureuse des mathématiques, influençant profondément la statistique, la théorie du hasard et de nombreux domaines appliqués.

Bruno de Finetti

(1906-1985)

Bruno de Finetti (1906–1985) était un mathématicien et statisticien italien, l'un des principaux théoriciens des probabilités au XXe siècle. Il est particulièrement connu pour ses travaux sur la probabilité subjective et la théorie des probabilités bayésiennes, qui remettent en question l'approche fréquentiste de la probabilité. De Finetti a proposé que les probabilités ne soient pas des propriétés objectives des événements, mais des degrés de croyance ou d'incertitude subjectifs, qui peuvent être modifiés à mesure que de nouvelles informations deviennent disponibles.

Bruno de Finetti

contributions

Dans son ouvrage majeur, Theory of Probability (1974), il a formalisé cette idée en développant la notion de "probabilité personnelle", selon laquelle une probabilité peut être interprétée comme une croyance personnelle dans un événement donné, et peut être mise à jour en fonction de l'observation de nouvelles données. De Finetti a également joué un rôle clé dans le développement des modèles bayésiens, en insistant sur l'importance de l'inférence inductive basée sur des connaissances préalables et des observations.
Sa vision de la probabilité a profondément influencé les méthodes statistiques modernes, en particulier l'inférence bayésienne, et il est souvent considéré comme l'un des plus grands défenseurs de l'approche subjective de la probabilité.

Alfréd Rényi

(1921-1970)

Alfréd Rényi (1921-1970) était un mathématicien hongrois connu pour ses contributions en théorie des probabilités, en statistiques et en mathématiques discrètes. Il a joué un rôle clé dans le développement des axiomes de la probabilité, proposant une approche alternative à celle de Kolmogorov. Il s'est également intéressé aux probabilités subjectives, bien que son travail soit resté plus axiomatique que bayésien. En statistique, il a introduit des concepts comme l'entropie de Rényi, qui généralise l'entropie de Shannon et trouve des applications en théorie de l'information et en apprentissage automatique.

Eugène Dynkin

(1924-2014)

Eugène Dynkin (1924-2014) était un mathématicien d'origine russe, reconnu pour ses contributions majeures en probabilités et en algèbre. Il a surtout marqué les domaines des processus stochastiques, en particulier les processus de Markov, et des groupes de Lie. Dynkin a formulé ce qui est aujourd'hui connu sous le nom de diagrammes de Dynkin, une classification des groupes de Lie semi-simples, et il a joué un rôle clé dans l'étude des systèmes dynamiques et de la théorie des probabilités, en introduisant des concepts comme les espaces de Dynkin.

Remarques

entre le survol historique et les leçons

Deux grandes écoles

- Fréquentistes
- Bayésiens

- La plupart du temps, les deux approches convergent sur des conclusions similaires/identiques.
- Parfois, il y a un désacord entre les deux écoles, notamment sur la formulation et l'interprétation du problème : Exemple du problème de Behrens-Fisher.
- Querelle fameuse

$$\text{Fisher} \Leftrightarrow \text{Neyman}$$

Statistique

pour la science des données

- Big Data
- Small Data

- Approches statistiques différentes et/ou complémentaires selon le nombre de phénomène observé (e.g. degré de libertés).
- Concepts comme la vraisemblance est un outil clé et un trait d'union entre le type de problème (small data vs. big data).

Importance de maîtriser les bases !

Examen

et tests

- examen écrit en fin de semestre
- pas d'examen intermédiaire ou examen blanc
- examen comme les exercices
- utilisation de la calculatrice et des tables