

PROBABILITÉS ET STATISTIQUE

Mise en garde

Ce polycopié est à usage personnel. Il ne doit être ni copié, ni modifié, ni diffusé mais peut être annoté.

Avant-propos

Ce cours de probabilités et statistique est destiné aux étudiant.e.s du Département des Technologies de l'Information et de la Communication (TIC) de la Haute École d'Ingénierie et de Gestion du Canton de Vaud, HEIG-VD. Les objectifs du cours sont clairement énoncés dans la fiche de cours.

Les notes de cours contiennent peu d'exemples. L'étudiant.e les complètera par les exemples et exercices présentés aux séances de cours. Les théorèmes de base sont énoncés avec précision et rigueur mais souvent sans preuve. On préfère une approche intuitive des concepts probabilistes et statistiques. Certaines citations figurant dans les notes n'ont pas été traduites en français pour ne pas déformer leur sens original.

À la fin de presque tous les chapitres figurent des exercices ainsi que leurs solutions numériques. L'étudiant.e pourra consolider ses connaissances en plus des séries d'exercices distribuées en classe.

Le logiciel de statistique utilisé au cours est R. Il a été créé vers 1994 par Ross Ihaka et Robert Gentleman de l'Université d'Auckland en Nouvelle-Zélande. R est davantage qu'un simple logiciel de statistique. Il s'agit non seulement d'un outil d'analyse statistique et graphique mais aussi d'un langage reposant sur le langage S créé par AT&T Bell Laboratories. John M. Chambers, l'un des créateurs de S, a reçu en 1998 le *Software System Award* de la prestigieuse ACM ("Association for Computing Machinery")¹.

Les possibilités offertes par R sont vastes et permettent à l'utilisateur d'effectuer des analyses de données très pointues. R est reconnu pour sa flexibilité. En effet, les résultats d'une analyse sont stockés dans un "objet"; il est alors possible de n'afficher que la partie des résultats qui intéresse l'utilisateur. Cette facilité n'est pas offerte par la plupart des logiciels classiques.

R est distribué librement sous les termes de la *GNU General Public Licence* sur le site web

www.r-project.org.

Il est développé et distribué par le groupe du *R Development Core Team* composé de plusieurs statisticiens provenant du milieu académique et du milieu industriel.

De prime abord, R paraît compliqué, peu convivial et destiné à des experts en statistique. Rassurez-vous, ce n'est pas le cas! Après un apprentissage, peut-être pénible et laborieux, R devient un outil statistique très apprécié (ce n'est pas un bluff!) et très efficace. Il est d'ailleurs utilisé dans le milieu académique et dans l'industrie.

Plusieurs dessins figurant dans les notes sont tirés du livre "*The Cartoon Guide to Statistics*" de L. Gonick et W. Smith. Les références exactes (chapitres et pages) n'ont délibérément pas été mentionnées dans les notes de cours. Un grand merci à Enrico Chavez qui lui aussi par ses dessins illustre avec justesse certains concepts délicats des probabilités et de la statistique. Finalement, quelques dessins ont été empruntés à Burki, dessinateur de presse, ainsi qu'à quelques autres.

Je tiens à remercier Marcel Baumgartner (Nestlé) pour sa lecture attentive du manuscrit et pour ses remarques pertinentes ainsi qu'Anthony Davison (EPFL), Pierre-Louis Aubert (HEIG-VD) et Diego Kuonen (Statoo Consulting) pour leur aide à la rédaction des notes de cours et pour leurs conseils.

BONNE ANNÉE SCOLAIRE !

Jacques Zuber
Bureau H06, interne 76459
email jacques.zuber@heig-vd.ch

Références

- Dalgaard, P. (2008). *Introductory Statistics with R*. Second Edition. New York : Springer.
- Daly, F., Hand, D.J., Jones, M.C., Lunn, A.D., and McConway, K.J. (1995). *Elements of Statistics*. Addison-Wesley, Harlow, England.
- Drouilhet, R., Lafaye de Micheaux, P., et Liquet, B. (2011). *Le logiciel R : Maîtriser le langage, Effectuer des analyses statistiques*. Paris : Springer-Verlag.
- Gonick, L. & Smith, W. (1993). *The Cartoon Guide to Statistics*. HarperCollins, New-York.
- Maindonald, J. & Braun, J. (2010). *Data Analysis and Graphics Using R. An Example-Based Approach*, Third Edition. Cambridge : Cambridge University Press.
- Morgensthaler, S. (2014). *Introduction à la statistique (4ème édition)*. Presses Polytechniques et Universitaires Romandes, PPUR, Lausanne.
- Mountassir, M. (2016). *Probabilités et statistique*. Modulo, Montréal.
- Ross, S. M. (2014). *Initiation aux probabilités* (Traduction de la neuvième édition américaine). Presses Polytechniques et Universitaires Romandes, PPUR, Lausanne.
- Wild, C. J. & Seber, G. A. F. (2000). *Chance Encounter, A First Course in Data Analysis and Inference*, Wiley, New York.

Ressources pour R

- Hashtag #rstats :
<https://twitter.com/hashtag/rstats?src=hash>
- Hornik, K. (2018). *The R FAQ* :
<http://cran.r-project.org/faqs.html>
- The R Seek :
<http://www.rseek.org/>
- The Quick-R :
<http://www.statmethods.net/>
- The R Graph Gallery :
<http://r-graph-gallery.com/>

¹. Citation de l'ACM au sujet de S : "forever altered the way people analyze, visualize, and manipulate data... S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers."

Table des matières

1 Qu'est-ce que la statistique ?	
1.1 Introduction	3
1.2 Quelques définitions de base	27
1.3 Et la statistique...	35
2 Analyse exploratoire des données	
2.1 Introduction	3
2.2 Techniques de visualisation pour une variable	10
Variable quantitative	10
1. Diagramme en points (<i>dot plot</i>)	14
2. Diagramme branche-et-feuilles (<i>stem-and-leaf</i>)	17
3. Histogramme (<i>histogram</i>)	23
4. Les statistiques élémentaires (<i>numerical summaries</i>)	33
5. Boîte à moustaches (<i>boxplot</i>)	52
Variable qualitative	68
1. Diagramme en camembert (<i>pie chart</i>)	69
2. Diagramme en barres (<i>bar graph</i>)	71
2.3 Techniques de visualisation pour deux variables	74
1. Nuage de points (<i>scatter plot</i>)	74
Covariance	79
Coefficient de corrélation	80
2. Matrice de nuages de points (<i>scatter plot matrix</i>)	86
2.4 Conseils pour créer de bons graphiques en statistique	91
3 Probabilités élémentaires	
3.1 Introduction	3
3.2 Ensemble fondamental (ou univers) et événements	10
3.3 Mesure de probabilité (ou loi de probabilité)	17
3.4 Indépendance (première idée... qu'on approfondira...)	20
3.5 Ensembles fondamentaux à événements élémentaires équiprobables	21
4 Probabilités conditionnelles et indépendance	
4.1 Probabilités conditionnelles	3
4.2 Théorème de Bayes	12
Application : filtre bayésien anti-spam	16
4.3 Indépendance	27
Application : évaluation de la fiabilité d'un système	33
a) Système en série	36
b) Système en parallèle	38
4.4 Tirages et schéma de Bernoulli	40
a) Tirages avec remise, schéma de Bernoulli	42
b) Tirages sans remise	45
5 Variables aléatoires	
5.1 Généralités sur les variables aléatoires	3
5.2 Variables aléatoires discrètes	15
5.2.1 Histogramme, fonction de répartition	18
Histogramme	18
Fonction de répartition	20
5.2.2 Espérance mathématique	23
5.2.3 Variance, écart-type	25
Variance	25
Écart-type	27
5.3 Variables aléatoires continues	29
5.3.1 Fonction de répartition	42
5.3.2 Interprétation intuitive de la fonction de densité	48
5.3.3 Espérance mathématique	49
5.3.4 Variance, écart-type	50
Variance	50
Écart-type	50
5.4 Propriétés de l'espérance mathématique et de la variance	52
5.4.1 Propriétés de l'espérance mathématique	52
5.4.2 Propriétés de la variance	55
5.4.3 Changements d'origine et d'échelle	56
6 Distributions usuelles	
6.1 Lois discrètes	3
6.1.1 Loi de Bernoulli	5

6.1.2 Loi binomiale	6
6.1.3 Loi géométrique	10
6.1.4 Introduction aux processus de Poisson	12
6.1.5 Loi de Poisson	18
Résumé des caractéristiques des lois discrètes usuelles	22
6.2 Lois continues	23
6.2.1 Loi uniforme	26
6.2.2 Loi exponentielle	31
6.2.3 Loi normale (ou loi de Laplace – Gauss)	37
Résumé des caractéristiques des lois continues usuelles	60
7 Variables aléatoires simultanées	
7.1 Introduction	3
7.2 Cas discret	4
Loi de probabilité simultanée (ou conjointe)	7
Lois marginales	9
Covariance	10
Corrélation	12
Indépendance	14
7.3 Cas continu	15
Fonction de densité conjointe (ou simultanée)	16
Fonctions de densité marginales	17
Covariance	18
Corrélation	18
Indépendance	19
7.4 Somme de variables aléatoires	21
7.4.1 X et Y variables aléatoires indépendantes et continues	22
Convolution	23
7.4.2 X et Y variables aléatoires indépendantes et discrètes	28
7.5 Espérance mathématique d'une somme de variables aléatoires	32
7.6 Propriétés de la covariance, de la corrélation et autres	33
8 Le théorème central limite	
8.1 Introduction	3
8.2 Illustrations du théorème central limite	4
8.3 Le théorème central limite	8
9 Modèles statistiques et estimation de paramètres	
9.1 Modèles statistiques et échantillon	3
9.2 Estimateurs	7
9.3 Propriétés d'un estimateur	10
a) Biais	12
b) Carré moyen de l'erreur	14
c) Efficacité	16
9.4 L'estimation par le maximum de vraisemblance	18
9.5 Estimation par intervalle	29
Estimation par intervalle pour l'espérance μ dans le cas d'une distribution normale	36
Cas 1 : variance σ^2 connue	36
Cas 2 : variance σ^2 inconnue	42
Estimation par intervalle pour la variance σ^2 dans le cas d'une distribution normale	48
Estimation par intervalle pour le paramètre p dans le cas d'une distribution binomiale	53
10 Tests d'hypothèses	
10.1 Qu'entend-on par "inférence statistique" ?	3
10.2 Tests d'hypothèses	6
10.3 Quelques tests d'hypothèses	35
10.3.1 Tests sur les paramètres de lois normales	35
10.3.2 Test d'adéquation d'un modèle théorique	55
10.3.3 Test d'indépendance de deux variables catégoriques	66
10.4 Mises en garde	77
10.5 Rudiments de la théorie de la décision	79
Puissance d'un test	90
10.6 Et ensuite...	95
11 Introduction au data mining orienté vers le business	
11.1 Introduction et motivation	3
11.2 Qu'est-ce que le data mining ?	19
11.3 Nature des données	31
11.4 Le processus d'extraction des connaissances	37
11.5 Quelques techniques de data mining	54
Apprentissage non supervisé	68
Les règles d'associations logiques	69
Le clustering (groupement)	80
Quelques autres méthodes d'apprentissage non supervisé	101

Apprentissage supervisé	102
Classification et arbres de classification	103
Réseaux de neurones	117
Quelques autres méthodes d'apprentissage supervisé	134
11.6 Conclusion	150
11.7 Références et ressources	157

Annexes

- Formulaire de probabilités
- Intervalles de confiance
- Tableau des tests sur les paramètres de lois normales
- Table de la distribution normale centrée réduite
- Quantiles de la distribution de Student
- Quantiles de la distribution khi carré
- Quelques documents complémentaires

Chapitre 1

Qu'est-ce que la statistique ?

1.1 Introduction

- De nos jours, les données jouent un rôle de plus en plus prépondérant dans les entreprises, sociétés et organisations.
- On assiste d'ailleurs à un flux considérable de données issues de saisies automatisées. Pour l'illustrer, WallMart, une grande chaîne de distribution américaine, enregistre chaque jour plus de 20 millions de transactions à partir de ses points de vente.
- Cependant, une étude du Gartner Group montrait que moins de 15 % des données stockées et moins de 5 % des données manipulées étaient analysées.

On ressent donc un besoin croissant en analyse et en étude de données.

Contenu

1.1 Introduction

1.2 Quelques définitions de base

1.3 Et la statistique...



"We must treat data as a company asset."

Nestlé's GLOBE ("Global Business Excellence") programme

"The problem isn't that specialised companies lack the data they need, it's that they don't go and look for it, they don't understand how to handle it."

Hans Rosling

- Le besoin en analyse et en étude de données est lié, par exemple, à des problèmes vitaux pour le positionnement concurrentiel (différenciation par rapport à la concurrence, gain de productivité, amélioration de la performance, avantage concurrentiel).
- Pour exploiter efficacement des données et en tirer des informations pertinentes, on utilise principalement des **méthodes statistiques**. Cependant, la statistique ne se limite pas à l'**analyse de données**. En effet, elle s'occupe aussi
 - ◊ de la **collecte**,
 - ◊ de l'**interprétation**

des données observées ou mesurées. Ces activités de la statistique ne sont pas distinctes; au contraire, elles dépendent fortement les unes des autres.



La collecte des données n'est pas toujours aisée.

- La statistique est utilisée dans plusieurs disciplines :

- la science du traitement du signal;
- l'hydrologie et la climatologie;
- la biologie;
- la physique;
- la recherche en médecine, en pharmacie, en psychologie et en nutrition;
- les sciences économiques et sociales;
- le marketing;
- le sport;
- ...



Exemples d'applications de la statistique :

- déetecter les effets d'une substance (drogue, médicament), les comparer et observer s'ils sont statistiquement significatifs;
- reconnaître des schémas dans des séquences d'ADN;
- analyser le comportement du consommateur dans un système de vente par correspondance;
- prévoir les ventes dans une grande chaîne de distribution et anticiper au mieux les tendances du marché;
- constituer une segmentation (typologie) des clients d'une banque pour cibler des opérations de marketing ou des attributions de crédit;

De: "Amazon.com" <store-news@amazon.com>
 A: jzuber@netplus.ch
 Date: Ven, 9 Mars 2007, 4:21
 Sujet: Amazon.com recommends Probability and Statistics and more

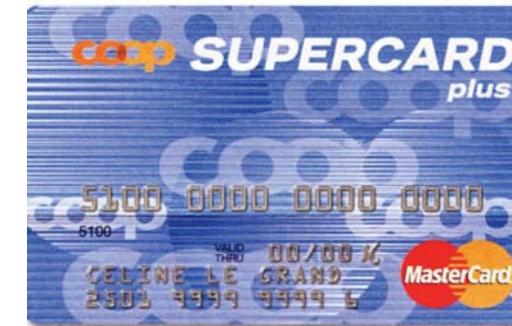
In this message:
 Jacques Zuber, Amazon.com has new recommendations for you based on items
 you purchased or told us you own.

We recommend:
 * Probability and Statistics
 * Concrete
 * How to Lie With Charts: Second Edition
 * Seeing Through Statistics (with CD-ROM and InfoTrac)
 * Ordinal Data Modeling (Statistics for Social Science and Behavioral
 Sciences)
 * Models for Discrete Longitudinal Data (Springer Series in Statistics)
 * The Statistical Evaluation of Medical Tests for Classification and
 Prediction (Oxford Statistical Science Series)
 * Matched Sampling for Causal Effects

...
 We recommend: Probability and Statistics
 by 3rd Edition DeGroot and Schervish
http://www.amazon.com/dp/1428813802/ref=pe_ar_xl

Price: \$9.95

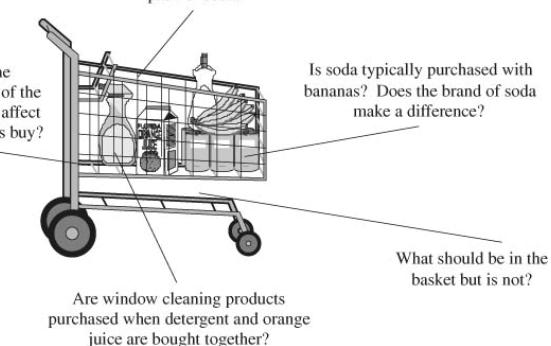
Recommended because you purchased or rated:
 * Probability and Statistics (3rd Edition)



Exemples d'applications de la statistique (*suite*) :

6. rechercher, spécifier puis cibler les niches de marché les plus profitables (banque) ou au contraire les plus risquées (assurance);
7. optimiser des procédés ou des produits à l'aide de plans d'expériences grâce auxquels il est aussi possible d'obtenir un maximum d'informations à un coût minimal;
8. contrôler la qualité, anticiper les défauts d'un produit issu d'un procédé d'usinage et détecter au plus vite l'origine d'une défaillance;
9. découvrir des structures et des relations se trouvant dans de grandes bases de données complexes ("data mining" en anglais, fouille de données ou prospection de données en français);
10. ...

In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.



Les concepts de base de la statistique sont

**la variabilité
et
l'incertitude.**

"There is nothing as certain and unchanging as uncertainty and change."

John F. Kennedy (1917–1963)

▷ La variabilité :

des variations dues à de petits, parfois grands changements, d'ailleurs souvent imprévisibles, se manifestent dans de très nombreuses situations.

Exemples :

- a) les êtres humains sont différents les uns des autres et physiologiquement leur organisme se modifie de jour en jour, voire même de minute en minute. En effet, la pression du sang, la taille et le poids d'un individu varient constamment;
- b) dans son travail, le chercheur est confronté à des erreurs expérimentales et des erreurs de mesure dont il doit en tenir compte lors de l'analyse des données mesurées.



Une certaine variabilité, non ?

▷ L'incertitude :

une grande partie de notre vie est faite d'incertitude que l'on souhaite quantifier. Ne se pose-t-on pas souvent la question : "Que va-t-il se passer ?" ou encore "Que va-t-il m'advenir demain, dans un mois, dans une année, dans dix ans ?".

~ Par la statistique,

en dépit de la variabilité à laquelle sont soumises les données observées ou mesurées, on souhaite en tirer des conclusions tout en contrôlant totalement le niveau d'incertitude. Cette démarche se résume à exploiter les données, à y découvrir des informations quantifiables et pertinentes, des tendances. La statistique devient un outil précieux d'aide à la décision.

"My basic idea is that the world has changed so much, what people need isn't more data but a new mindset."

Hans Rosling

"The world we live in is awash with data, that comes pouring in from everywhere around us. On its own, this data is just noise and confusion. To make sense of data, to find the meaning in it, we need a powerful branch of science: statistics!"

"For Today's Graduate, Just One Word: Statistics."

Steve Lohr, New York Times, August 5, 2009

Hans Rosling

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."

Hal Varian, chief economist at Google

"If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."

Hal Varian, chief economist at Google

1.2 Quelques définitions de base

- **variable** : expression numérique d'un caractère observé; dans un contexte plus général citons **type d'information** comme terme synonyme.

Une variable peut être

- ◊ **quantitative**;
- ◊ **qualitative**.

- a) une **variable quantitative** est utilisée par exemple dans des campagnes de mesures. Elle peut être *discrete* ou *continue*.

- Exemples de **variables quantitatives discrètes** :
 - nombre d'enfants dans une famille;
 - nombre de pièces défectueuses dans un lot produit par un procédé d'usinage;
 - nombre de transactions journalières (autour de 20 millions) enregistrées par une grande chaîne de distribution à partir de ses 2000 points de vente.
- Exemples de **variables quantitatives continues** :
 - poids d'un être humain;
 - taille d'un adulte;
 - température relevée à Yverdon-les-Bains.

b) une variable qualitative peut être *catégorique* ou *ordinale*.

- Exemples de variables qualitatives catégoriques :

- le sexe (masculin ou féminin);
- le pays (Suisse, Suède, Australie, ...).

- Exemples de variables qualitatives ordinaires :

- la qualité du son émis par un baladeur mp3 testé par les Organisations européennes de consommateurs peut être jugée *bonne, moyenne ou mauvaise*;
- le revenu d'un habitant d'un pays donné peut être qualifié *très bas, bas, moyen, élevé ou très élevé*.

1.2 Quelques définitions de base (suite)

- observation** : valeur mesurée ou observée pour une variable donnée;

exemple : nombre de buts inscrits dans ses matches à domicile par Liverpool lors de la saison 1994–1995 jusqu'au 5 février 1995 : 24;

- individu, unité expérimentale, unité d'observation** : source des valeurs d'une ou plusieurs variables;

exemple : pour la variable comptant le nombre de buts inscrits dans ses matches à domicile lors de la saison 1994–1995 jusqu'au 5 février 1995, Liverpool peut être considéré comme une unité d'observation;

- population** : collection finie ou hypothétique d'individus;

exemple : ensemble des 22 équipes de la première division anglaise de football durant la saison 1994–1995;

Remarques :

- si les valeurs prises par une variable quantitative se répètent beaucoup, la variable est considérée comme discrète. En revanche, si les valeurs se répètent peu, la variable est considérée comme continue;
- dans ce cours, on évitera d'effectuer une distinction stricte entre le type de variables. On se limitera à différencier les variables qualitatives des variables quantitatives.

- données** : collection d'observations d'une ou plusieurs variables;

exemple : nombres de buts inscrits à domicile (HG) et à l'extérieur (AG) lors de la saison 1994–1995 du championnat de première division anglaise de football jusqu'au 5 février 1995 :

Équipe	HG	AG	Équipe	HG	AG
Blackburn Rovers	39	19	Chelsea	19	15
Manchester United	27	21	Manchester City	27	8
Newcastle United	28	16	Aston Villa	13	19
Liverpool	24	21	Southampton	18	19
Nottingham Forest	22	18	Crystal Palace	9	12
Tottenham Hotspur	22	22	Queen's Park Rangers	23	15
Leeds United	20	14	Everton	23	4
Sheffield Wednesday	18	18	West Ham United	14	10
Wimbledon	18	13	Coventry City	12	13
Norwich City	19	6	Ipswich Town	19	10
Arsenal	15	15	Leicester City	16	8

Source : Wild C.J. & Seber G.A.F. (2000). *Chance Encounter, A First Course in Data Analysis and Inference*. NY:Wiley

- **échantillon** : sous-ensemble d'une population définie.

L'échantillon est sélectionné afin de représenter la population pour les variables considérées;

exemple : échantillon de 1000 habitants d'un pays qui en comprend 7 millions;

- **échantillonnage** : procédure d'extraction d'un échantillon à partir d'une population;

- **classe** : soit un intervalle de grandeur appropriée regroupant des observations quantitatives, soit une catégorie regroupant des observations qualitatives;

exemple : une étude a été menée dans un atelier mécanique pour vérifier le diamètre de tiges tournées sur un tour automatique. Les tiges dont le diamètre est compris entre 38.5 et 39.5 millimètres forment une classe;

- **paramètre** : caractéristique d'une population;

exemple : la moyenne de buts inscrits à domicile dans la population formée des 22 équipes de football de première division anglaise lors de la saison 1994–1995 jusqu'au 5 février 1995 est approximativement 20.23;

- **statistique** : valeur d'une caractéristique calculée à partir de l'échantillon;

exemple : parmi les 22 équipes appartenant au championnat de première division anglaise de football durant la saison 1994–1995, 11 ont été choisies au hasard pour former un échantillon. Les équipes sélectionnées sont Blackburn Rovers, Manchester United, Nottingham Forest, Leeds United, Arsenal, Manchester City, Queen's Park Rangers, Everton, West Ham United, Ipswich Town et Leicester City. La moyenne de buts inscrits à domicile jusqu'au 5 février 1995 dans cet échantillon de 11 équipes vaut approximativement 22.27.

1.3 Et la statistique...

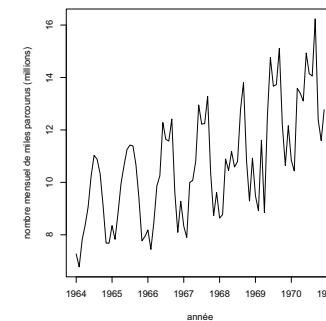
La statistique est souvent décrite comme étant formée de deux pôles :

- **l'analyse exploratoire des données** : elle est composée principalement de méthodes graphiques et elle permet de détecter les structures spécifiques (tendances, formes, allures des distributions, observations atypiques);

- **l'inférence statistique** : elle conduit à des conclusions statistiques à partir de données en utilisant des notions de la théorie des probabilités.

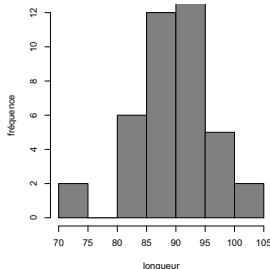
→ Le cours sera divisé en trois parties : l'analyse exploratoire des données, une initiation aux probabilités et une introduction à l'inférence statistique.

▷ Tendances annuelle et saisonnière



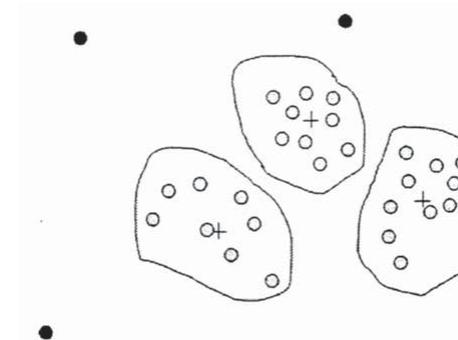
Nombre mensuel de millions de miles parcourus au Royaume-Uni par transport aérien entre 1964 et 1970.

▷ Forme de la distribution



Longueur des 40 coyotes femelles faisant partie d'une étude menée au Canada.

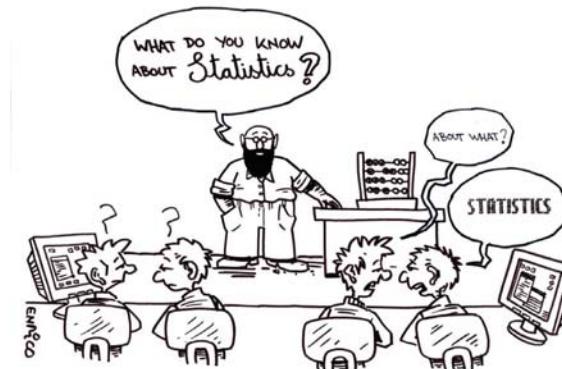
▷ Observations atypiques



Observations atypiques.

Mais avant de débuter le cours, quelques mises au point s'imposent :

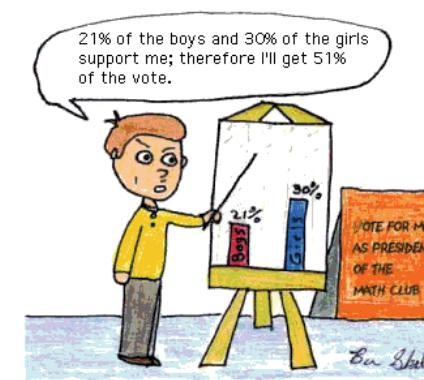
- la statistique est un outil très puissant. On ignore parfois (souvent ?) ce qu'elle représente, ses principes et ses méthodes;
- par la statistique, on cherche à tirer des conclusions sur des données soumises à la variabilité tout en contrôlant totalement le niveau d'incertitude. La statistique est donc une science exacte qui ne doit pas être considérée comme une vilaine boîte noire !
- les résultats statistiques sont parfois difficilement interprétables et doivent être manipulés avec précision, prudence et sans malveillance. Il faut toujours mentionner le contexte;
- une analyse statistique peut être effectuée seulement si on connaît parfaitement les méthodes à utiliser, leurs pièges et leurs limites.



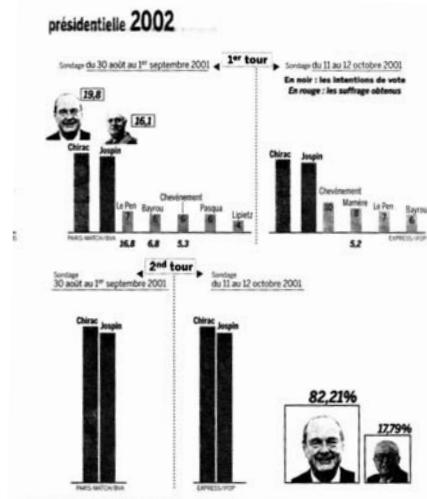
À y remédier rapidement !



No comment !



Quelle interprétation !



C'est pour un sondage...

Les sondages ne disent pas toujours la vérité. Tenez: interrogés dans le cadre d'une étude comparative sur la qualité de l'enseignement dans les Universités, les étudiants de la Faculté Economie et sciences sociales de l'Uni de Berne ont pris l'initiative de se passer le mot, via Internet, afin d'encourager les étudiants fréquentant l'établissement à répondre positivement. Objectif: faire en sorte que la Faculté soit favorablement notée, pour qu'elle bénéficie d'une nouvelle notoriété vis-à-vis de l'extérieur, augmentant du même coup les chances

des étudiants sur le marché de l'emploi. Plutôt que de faire des reproches sur cette façon de procéder, il vaudrait peut-être mieux louer la démarche de ces étudiants qui ont bien compris non seulement les règles du jeu économique – l'opportunité de « bien vendre » un produit –, mais aussi le fonctionnement des lois du marché. Autrement dit, ces derniers se sont en l'occurrence avérés être des économistes en herbe plutôt prometteurs! Si, par la même occasion, les sondages devaient en prendre un petit coup, ce ne serait pas plus mal...

Ce genre d'analyses comparatives est fréquent dans certains milieux académiques, et de nombreux bureaux de consultants en ont fait leur fonds de commerce. Il faut néanmoins noter que les procédures utilisées lors de ces enquêtes ne sont pas toujours des plus transparentes. Si le public se montre parfois sceptique vis-à-vis des sondages en général, l'exemple de l'Université de Berne a le mérite de montrer que nous pouvons tous être tentés de répondre aux questions d'un sondage de manière à faire pencher la balance du côté de nos propres aspirations.



Beat Kappeler,
éditorialiste
au *Temps* et
à la *NZZ*.

Moi aussi, lorsque je suis consulté par téléphone, je me permets parfois quelques libertés.

Il est démontré que la crédibilité d'un sondage nécessite au moins 300 répondants dans le cas d'un oui/non et plus de mille lorsque les questions sont à choix multiple. Une seule règle: ne nous fions jamais à un sondage qui n'indiquerait pas le nombre des personnes interrogées.



"In earlier times, they had no statistics, and so they had to fall back on lies."

Anonymous



"It is easy to lie with statistics, but easier to lie without them."

Fred Mosteller



Prof. Diego Kuonen @DiegoKuonen · 19 août
Data + Science + Statistics = Sexy!

Voir la traduction

Data + Science - Statistics = Failure!

Data - Science - Statistics > IT!

#DataScience #Statistics #ML



Kuonen, D. (19 août 2016).
@DiegoKuonen



Prof. Diego Kuonen @DiegoKuonen · 15 août

'Minds & #Machines: #Forecasting in age of #ArtificialIntelligence'

>dupress.com/articles/art-o...

#BigData #AI HT @mitsmr



Such issues routinely arise in applied work and are a major reason why models can guide—but typically cannot replace—human experts. Figuratively speaking, the equation should be not “algorithms > experts” but instead, “experts + algorithms > experts.”



Kuonen, D. (15 août 2016).
@DiegoKuonen

EXERCICE : QU'EST-CE QUE LA STATISTIQUE ?

Exercice 1

Les nombres de buts inscrits à domicile (HG) et à l'extérieur (AG) lors de la saison 1994–1995 du championnat de première division anglaise de football jusqu'au 5 février 1995 figurent dans le tableau ci-dessous.

Équipe	HG	AG	Équipe	HG	AG
Blackburn Rovers	39	19	Chelsea	19	15
Manchester United	27	21	Manchester City	27	8
Newcastle United	28	16	Aston Villa	13	19
Liverpool	24	21	Southampton	18	19
Nottingham Forest	22	18	Crystal Palace	9	12
Tottenham Hotspur	22	22	Queen's Park Rangers	23	15
Leeds United	20	14	Everton	23	4
Sheffield Wednesday	18	18	West Ham United	14	10
Wimbledon	18	13	Coventry City	12	13
Norwich City	19	6	Ipswich Town	19	10
Arsenal	15	15	Leicester City	16	8

Source : Wild, C.J. & Seber, G.A.F. (2000). *Chance Encounter, A First Course in Data Analysis and Inference*. NY:Wiley

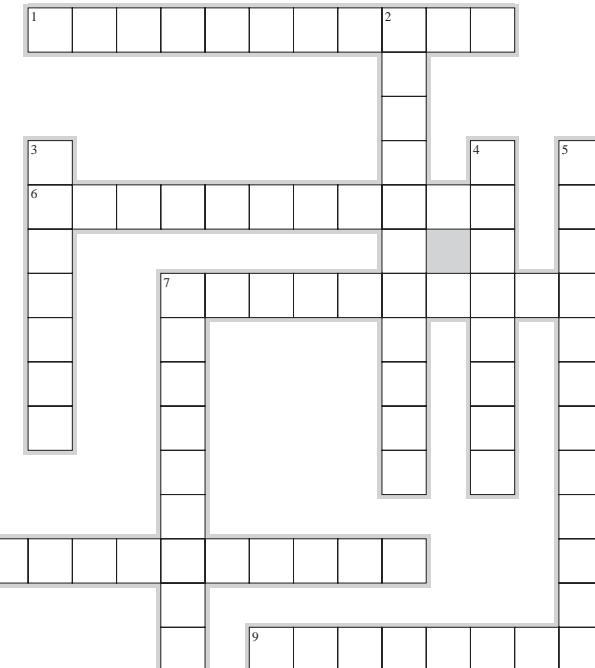
Compléter la grille de la page suivante à l'aide des indications ci-dessous.

Horizontal

1. La moyenne des buts inscrits à domicile par 11 équipes formant un échantillon est une
6. Le nombre de buts inscrits lors de ses matches à domicile par Arsenal est une
7. L'ensemble des 22 équipes constitue une
8. 11 équipes choisies au hasard parmi les 22 équipes forme un
9. Le nombre de buts inscrits à domicile est une

Vertical

2. Le nombre de buts inscrits à l'extérieur n'est pas une variable de nature
3. Le tableau ci-dessus contient les
4. Pour la variable comptant le nombre de buts inscrits dans ses matches à domicile, l'équipe Everton peut être considérée comme un
5. Le nombre de buts inscrits à domicile est une variable de nature
7. La moyenne des buts inscrits à domicile par les 22 équipes est un ...



Created with EclipseCrossword - www.eclipsecrossword.com

Chapitre 2

Analyse exploratoire des données

Contenu

2.1 Introduction

2.2 Techniques de visualisation pour une variable

2.3 Techniques de visualisation pour deux variables

2.4 Conseils pour créer de bons graphiques en statistique

2.1 Introduction

- Une fois les données soigneusement collectées, on souhaite les analyser pour en découvrir ses principales caractéristiques. Cette démarche permettant d'avoir un aperçu global des données s'appelle l'**analyse exploratoire**.
- Les objectifs spécifiques de l'analyse exploratoire sont multiples :
 - ◊ comprendre,
 - ◊ résumer,
 - ◊ décrire,
 - ◊ représenter

les données en vue d'interpréter la réalité.



"The world we live in is awash with data, that comes pouring in from everywhere around us. On its own, this data is just noise and confusion. To make sense of data, to find the meaning in it, we need a powerful branch of science: statistics!"

Hans Rosling

- L'analyse exploratoire suggère également des modèles et des hypothèses qui seront par la suite formalisés puis traités grâce à des outils probabilistes
↗ modélisation et inférence statistique.
- Les techniques d'analyse exploratoire sont **graphiques** (*l'adage ne dit-il pas qu'un graphique peut exprimer davantage que mille mots ?*) complétées de résultats **numériques**.
- Décrire les données à l'aide de graphiques est primordial. Mais...
 - ◊ quels graphiques doit-on utiliser ?
 - ◊ comment peut-on les interpréter ?

- Dans ce chapitre, nous répondrons à ces questions en présentant des techniques de visualisation pour
 - ◊ une seule variable qualitative ou continue;
 - ◊ deux variables continues.
- Dans le cas d'une variable continue, les techniques de visualisation nous permettront de disposer de premières informations sur la **répartition**, la **forme générale** de la distribution des données. Pour deux variables continues, elles reconnaîtront, si elle existe, la **relation** entre les deux variables. En résumé, les techniques de visualisation sont chargées de détecter la **structure** existant dans les données.

- Notons que le choix du graphique dépend
 - ◊ du type de variables (qualitative ou quantitative);
 - ◊ du nombre d'observations;
 - ◊ du message qu'on souhaite transmettre.

Nous reviendrons tout au long du chapitre sur cette dépendance.

- Bon à savoir...

"Construire de bons graphiques est à la fois un art et une science."



2.2 Techniques de visualisation pour une variable

Différents types de graphiques seront présentés dans ce paragraphe. Considérons une seule variable continue mesurée plusieurs fois. On dispose ainsi de n observations

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

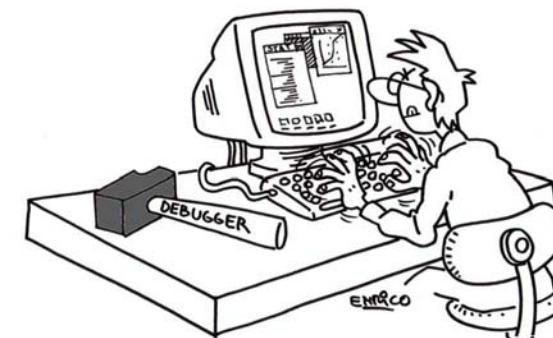
de la variable. Cette série peut être rangée dans l'ordre croissant des valeurs. Comme nous le verrons ailleurs dans le paragraphe, cette classification est très utile pour construire des graphiques ou pour calculer plusieurs caractéristiques de la distribution des données. Les valeurs ordonnées de la série seront notées

$$x_{1|n}, x_{2|n}, \dots, x_{n|n},$$

$x_{1|n}$ est le minimum et $x_{n|n}$ le maximum.

Les techniques de visualisation seront appliquées aux mêmes données. Dans un procédé industriel, l'épaisseur en *micropouces* d'un placage en or sur 90 panneaux de circuits imprimés a été relevée. Les données observées figurent dans le tableau suivant :

24	24	26	28	28	29	29	29	30	31
32	32	32	32	32	33	33	35	35	35
35	35	36	36	36	36	37	38	38	39
39	39	39	40	40	41	42	43	43	44
45	47	47	47	48	48	48	49	49	49
50	50	50	51	51	51	52	52	52	53
54	54	55	56	56	56	57	57	57	57
57	57	58	58	59	59	60	60	60	60
61	61	62	62	63	64	64	65	66	69



Le logiciel de statistique R sera utilisé pour analyser les données.

Source : "Improving a gold plating process using Taguchi Methods", Sixth symposium on Taguchi Methods
November 1988, American Supplier Institute, Dearborn, Michigan

The R-Files

The truth is in the data

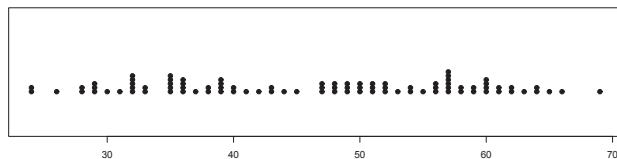


Diagramme en points de l'épaisseur en micropouces du placage en or.

1. Diagramme en points (*dot plot*)

Particulièrement utile pour :

visualiser la distribution, détecter des groupes, des disparités et des observations atypiques, i.e des données "très" éloignées de la majeure partie du reste des observations.

Construction :

les données sont triées par ordre croissant et placées sur un axe. Chaque observation est représentée par un point. Les points sont placés les uns sur les autres autant de fois que la valeur a été observée.

Inconvénient :

le diagramme en points ne peut pas être facilement interprétable pour des jeux de données de taille supérieure à 20, voire même à 30 observations;

↗ diagramme branche-et-feuilles et histogramme.

2. Diagramme branche-et-feuilles (*stem-and-leaf*)

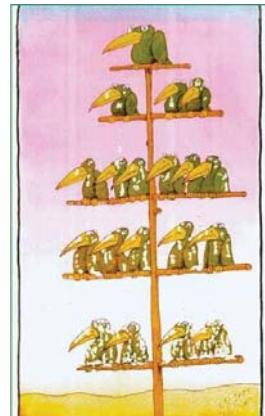
Particulièrement utile pour :

visualiser la distribution, détecter des groupes, des disparités et des observations atypiques.

Construction :

1. regrouper les données dans différentes classes :

on peut par exemple séparer le nombre formant une donnée en deux parties $a | b$ où a représente la *branche* et b la *feuille* (la feuille est formée d'un seul chiffre); exemple : $32 \rightarrow 3 | 2$. Chaque classe est étiquetée par l'une des branches a ;



Construction (*suite*) :

2. écrire une ligne par classe :

chaque ligne se compose de la *branche a*, d'un séparateur | suivi des *feuilles b* des données se trouvant dans la classe;

3. ranger les *feuilles* de chaque classe par ordre croissant.

Avantage :

le diagramme branche-et-feuilles montre les valeurs numériques. Il permet une bonne visualisation de 15 à 150 observations.

2	44
2	688999
3	012222233
3	5555566667889999
4	0012334
4	5777888999
5	000111222344
5	56667777778899
6	00001122344
6	569

Diagramme branche-et-feuilles de l'épaisseur en micropouces du placage en or.

Remarque :

pour alléger le graphique et/ou le raccourcir, il convient souvent d'arrondir les valeurs de la liste des données et/ou de choisir (choix difficile !) une échelle. Pour l'illustrer, considérons les précipitations annuelles en millimètres de Genève entre 1826 et 1843 :

583	890	777	958	875	926	524	756	619
730	688	528	901	884	969	1258	850	939

Les données observées sont arrondies à la dizaine (583 est arrondi à 580), l'échelle est 100 mm (la feuille 5 | 8 du diagramme représente la valeur $5.8 \times 100 \text{ mm} = 580 \text{ mm}$, valeur arrondie de la donnée observée 583 mm) et l'incrément est 1 unité (les valeurs de la branche sont 5, 6, ..., 12).



Diagramme branche-et-feuilles des précipitations annuelles de Genève entre 1826 et 1843.

3. Histogramme (*histogram*)

Particulièrement utile pour :

visualiser la distribution, détecter des groupes, des disparités et des observations atypiques.

Pour construire un histogramme, il est utile de disposer d'une **table de fréquences** qui peut être considérée comme un résumé des valeurs observées.

Construction de la table de fréquences :

1. ranger les données par ordre croissant;
2. regrouper les données dans différentes classes si n est assez grand.
Autrement dit, diviser l'axe en une partition de k intervalles disjoints (en général entre 5 et 15) de même longueur h appelée **l'amplitude**;
3. compter le nombre d'observations figurant dans chaque classe (ce nombre est appelé **l'effectif** ou la **fréquence absolue**);
4. construire la table de fréquences. Celle-ci contient : les limites des classes i.e les extrémités des intervalles, les centres des classes i.e les milieux des intervalles, les fréquences absolues des classes, les fréquences relatives des classes i.e la proportion des observations contenues dans les intervalles.

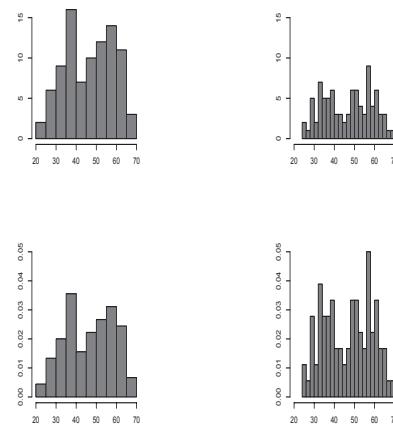
classe	centre	fréquence absolue	fréquence relative
20 – 25 [–]	22.5	2	0.02
25 – 30 [–]	27.5	6	0.07
30 – 35 [–]	32.5	9	0.10
35 – 40 [–]	37.5	16	0.18
40 – 45 [–]	42.5	7	0.08
45 – 50 [–]	47.5	10	0.11
50 – 55 [–]	52.5	12	0.13
55 – 60 [–]	57.5	14	0.16
60 – 65 [–]	62.5	11	0.12
65 – 70 [–]	67.5	3	0.03
total		90	1

Table de fréquences de l'épaisseur en micropouces du placage en or.

Construction de l'histogramme :

la table de fréquences permet une construction rapide de l'histogramme, graphique formé d'une suite de rectangles contigus. Chaque rectangle (appelé aussi **boîte**) couvre un intervalle (une classe). Il est centré au point milieu de l'intervalle et sa hauteur est donnée soit par la fréquence absolue soit pour faire en sorte que l'aire des rectangles réunis soit égale à 1.

En choisissant judicieusement l'amplitude des classes, en d'autres termes, la **largeur** des boîtes, on obtient un bon aperçu de la distribution des données. Néanmoins, le choix d'une largeur adéquate n'est pas immédiat. Le principe de base (très peu précis !) se résume à "peu de données, peu d'intervalles; plus de données, plus d'intervalles". Il faut parfois procéder par tâtonnement pour obtenir une largeur appropriée.

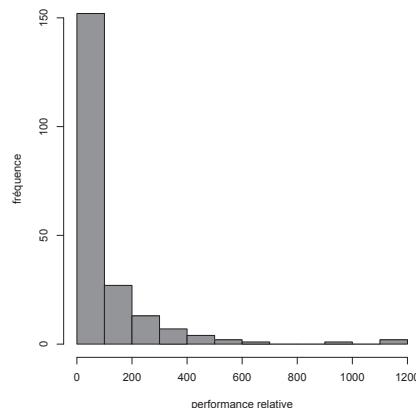


Histogrammes de l'épaisseur en micropouces du placage en or; 10 (gauche) et 23 (droite) classes avec fréquences absolues (haut) et aire des rectangles égale à 1 (bas).

Exemple :

la performance relative au processeur IBM 370/158-3 de 209 processeurs d'ordinateurs a été déterminée.

L'histogramme des performances se trouve à la page suivante. On y remarque que les performances sont davantage condensées vers les petites valeurs que vers les grandes. On dit alors que la distribution des valeurs observées est asymétrique. Comme la queue de distribution est allongée du côté positif, on parle d'asymétrie positive.

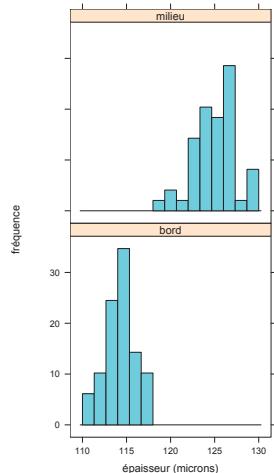


Histogramme des performances relatives de 209 processeurs d'ordinateurs.

Exemple :

dans une production industrielle, 49 rouleaux de film en plastique ont été choisis au hasard. Au bout de chaque rouleau, on a mesuré l'épaisseur en micromètres au bord et au milieu.

On se propose de comparer les épaisseurs selon l'endroit de mesure. L'histogramme figurant en page suivante met en évidence une différence significative. On distingue clairement que les rouleaux sont plus épais au milieu qu'au bord.



Histogrammes des épaisseurs au bord et au milieu des rouleaux.

Avantage :

l'histogramme peut être appliqué tout aussi bien à un petit nombre de données qu'à un grand nombre de données. Il est simple à construire, très compréhensible et facilement interprétable.

Inconvénients :

les principaux défauts de l'histogramme sont la perte d'informations en raison de l'absence des valeurs des observations et le choix délicat de la largeur des boîtes.

Remarque :

le diagramme branche-et-feuilles peut être vu comme un histogramme particulier obtenu par rotation de 90° dans le sens inverse des aiguilles de la montre.

4. Les statistiques élémentaires (*numerical summaries*)

Les statistiques élémentaires sont souvent considérées comme un résumé synthétique de la distribution des données. Elles complètent les représentations graphiques et sont très utiles pour effectuer des comparaisons précises entre plusieurs échantillons liés (par exemple, plusieurs séries de mesures de la même variable réalisées à des conditions différentes : mesures de l'épaisseur de rouleaux d'aluminium laminé, mesures prises au bord et au milieu en une position précise du rouleau).

Les statistiques élémentaires sont formées d'un petit nombre de valeurs numériques appelées **indicateurs**.

Les indicateurs sont classés en trois catégories :

- les indicateurs de **tendance centrale** qui informent sur le "milieu" (la position, le centre) d'une distribution : la **moyenne**, la **médiane** et accessoirement le **mode**;
- les indicateurs de **dispersion** qui renseignent sur la variabilité de la distribution : l'**étendue**, l'**écart-type**, l'**étendue interquartile** et le **coefficient de variation**;
- les indicateurs de **forme** qui mesurent le degré d'asymétrie ou d'aplatissement d'une distribution : l'**asymétrie** et l'**aplatissement**. Ces indicateurs seront étudiés plus loin dans le chapitre.



a) les indicateurs de tendance centrale :

- la **moyenne** (*mean, average*) : la **moyenne** (arithmétique) de l'échantillon est le nombre \bar{x} défini par

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (1)$$

Exemple : l'épaisseur moyenne du placage en or sur les panneaux des circuits imprimés est de 46.2 *micropouces*;

- la **médiane** (*median*) : les valeurs observées étant classées par ordre croissant, la **médiane**, notée $\text{med}(x)$ ou \tilde{x} , est la valeur qui partage l'ensemble des observations en deux parties de même grandeur. Autrement dit, 50 % des données sont plus petites qu'elle et 50 % plus grandes.

"A statistician can have his head in an oven and his feet in ice, and he will say that on the average he feels fine."

Anonymous

Exemples :

- la médiane des 8 nombres suivants

11 13 15 16 19 21 22 25

$$\text{est } \frac{x_{4|8} + x_{5|8}}{2} = 17.5;$$

- la médiane des épaisseurs du placage en or sur les panneaux des circuits imprimés est de 48 *micropouces*.

Remarque :

la moyenne et la médiane de l'épaisseur du placage en or sur les panneaux des circuits imprimés sont relativement proches l'une de l'autre. Quelle différence intrinsèque existe-t-il entre la **moyenne** et la **médiane** ?

Nous allons répondre à cette question par l'exemple numérique qui suit.

Calcul de la médiane :

- si n est *impair*, la médiane vaut

$$x_{(n+1)/2|n}.$$

Il s'agit de la valeur observée centrale.

Exemple : la médiane des 7 nombres suivants

1 4 7 9 10 12 14

est $x_{4|7} = 9$;

- si n est *pair*, la médiane est

$$\frac{x_{(n/2)|n} + x_{(n/2+1)|n}}{2}.$$

La médiane est la moyenne des données observées voisines du centre.

Remarque (*suite*) :

les temps de survie en jours des six premières personnes de cœur transplanté selon un certain programme médical sont

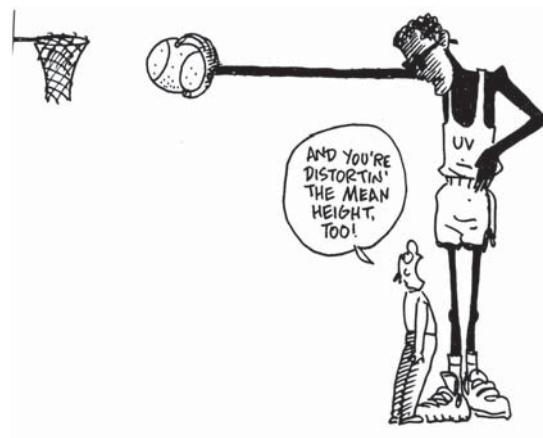
3 15 46 64 126 632.

La moyenne des données est $\bar{x} = 147.7$ jours; la médiane vaut $\text{med}(x) = (46 + 64)/2 = 55$ jours.

Ces résultats nous conduisent à la conclusion :

la moyenne est beaucoup plus sensible aux valeurs extrêmes (atypiques) que la médiane.

C'est d'ailleurs pour cette raison que la médiane est dite **indicateur robuste** de la tendance centrale d'une distribution statistique.



- **le mode (mode)** : le mode, bien qu'il soit classé dans la catégorie des indicateurs de tendance centrale, n'en est pas vraiment un.

Le **mode** (ou **valeur dominante**), noté M_0 , est la valeur de la variable statistique la plus observée. Autrement dit, il s'agit de la valeur qui a été observée le plus grand nombre de fois.

Exemples :

1. les modes des 10 nombres suivants

1 4 7 7 10 12 12 14 25 32

sont 7 et 12. Ces valeurs ont été observées exactement 2 fois.

2. le mode des épaisseurs du placage en or sur les panneaux des circuits imprimés est de 57 micropouces. Cette valeur a été relevée exactement 6 fois.

Remarques :

- il est possible qu'il n'existe aucun mode. On peut aussi rencontrer exactement un mode (distribution unimodale), deux modes (distribution bimodale), même plus. Dans l'exemple 1 de la page précédente, la distribution est bimodale;
- l'utilisation du mode est précieuse dans le contrôle industriel. Une distribution comportant plusieurs modes peut indiquer un mélange de populations différentes ayant leurs caractéristiques propres. Ceci peut s'expliquer entre autres par un mélange de matières premières ou par un déréglage soudain d'une machine. Dans ce type de phénomènes, la moyenne arithmétique ne peut plus être considérée comme mesure de tendance centrale.

b) les indicateurs de dispersion :

- **l'étendue (range)** : l'**étendue**, notée R , est la différence entre la plus grande et la plus petite observation de l'échantillon, i.e

$$x_{n|n} - x_{1|n}.$$

L'étendue est fortement influencée par les valeurs extrêmes. De ce fait, elle constitue un indicateur apprécié dans la maîtrise statistique des procédés industriels;

- **l'écart-type (standard deviation)** : l'**écart-type** s d'une liste de données est la racine carrée de la somme des carrés des écarts à la moyenne divisée par $(n - 1)$:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}. \quad (2)$$

Exemples :

- considérons les 5 nombres suivants

5 2 3 4 8.

La moyenne de ces nombres est $\bar{x} = 22/5 = 4.4$ et l'écart-type vaut

$$s = \sqrt{\frac{(5 - 4.4)^2 + (2 - 4.4)^2 + (3 - 4.4)^2 + (4 - 4.4)^2 + (8 - 4.4)^2}{4}} = 2.3;$$

- l'écart-type des épaisseurs du placage en or sur les panneaux des circuits imprimés est de 11.794 micropouces.

Remarque :

la quantité s^2 est la **variance** de l'échantillon. En pratique, elle est moins utilisée que l'écart-type étant donné qu'elle ne s'exprime pas dans la même unité que les observations.

- **l'étendue interquartile (interquartile range)** : comme nous l'avons vu précédemment, la médiane divise l'échantillon ordonné en deux parties de même grandeur. On peut envisager des partages plus fins de l'échantillon ordonné comme par exemple un partage en quatre parties de même grandeur. Les trois valeurs permettant cette division sont appelées les **quartiles**. On les définit comme suit :

- ◊ le **premier quartile** est la valeur notée $\hat{q}(25\%)$ telle que le 25 % des observations est plus petit qu'elle;
- ◊ le **second quartile** est la médiane;
- ◊ le **troisième quartile** est la valeur notée $\hat{q}(75\%)$ telle que le 75 % des observations est plus petit qu'elle.

Exemple : les quartiles des 9 nombres suivants

2 4 5 6 6 8 10 10 12

sont $\hat{q}(25\%) = \frac{4+5}{2} = 4.5$, médiane = 6 et $\hat{q}(75\%) = \frac{10+10}{2} = 10$.

La liste des cinq valeurs

$\text{minimum} = x_{1|n}$, $\hat{q}(25\%)$, médiane, $\hat{q}(75\%)$, $\text{maximum} = x_{n|n}$,
appelée en anglais *Five-Number summary*, donne un résumé
numérique simple et pratique d'une distribution.

Un troisième indicateur de dispersion (le moins sensible aux données atypiques) est l'**étendue interquartile** définie par

$$\hat{q}(75\%) - \hat{q}(25\%). \quad (3)$$

Cet indicateur est utile et très parlant mais son usage se réduit au stade exploratoire.

Exemple : l'étendue interquartile des nombres suivants

2 4 5 6 6 8 10 10 12

vaut $\hat{q}(75\%) - \hat{q}(25\%) = 10 - 4.5 = 5.5$.

Remarque :

on peut généraliser la notion de quartiles aux $\alpha\%-quantiles$. Le $\alpha\%-quantile$ est la valeur notée $\hat{q}(\alpha\%)$ telle que le $\alpha\%$ des observations est plus petit qu'elle. Des logiciels, S-PLUS ou R par exemple, proposent des méthodes précises pour déterminer les quartiles d'une liste de données. Il est fort probable que les quartiles obtenus par ces méthodes plus générales diffèrent de ceux que nous avons introduits dans ce cours. Néanmoins, les valeurs des quartiles restent relativement proches les unes des autres suivant les méthodes pour des échantillons de tailles moyennes ou grandes.

- **le coefficient de variation (coefficient of variation)** : le coefficient de variation, noté CV , est obtenu en divisant l'écart-type s par la moyenne arithmétique,

$$CV = \frac{s}{\bar{x}} \text{ avec } \bar{x} \neq 0. \quad (4)$$

Il s'agit d'un indicateur de dispersion sans dimension qui exprime la précision relative des observations. En effet, plus le coefficient de variation est faible, plus la série des valeurs observées est homogène, i.e concentrée autour de la moyenne \bar{x} . Dans ce cas, la moyenne devient un bon résumé de l'ensemble des valeurs de la série.

Exemple (suite) :

une indication sur le degré d'homogénéité d'une distribution est donnée par le coefficient de variation :

- ◊ pour la production de lampes au sodium :

$$CV_{sodium} = \frac{5.8}{115} = 0.05;$$

- ◊ pour la production de lampes à vapeur de mercure :

$$CV_{mercure} = \frac{3.2}{40} = 0.08.$$

Comme CV_{sodium} est plus petit que $CV_{mercure}$, la production des lampes au sodium présente un rendement énergétique plus homogène que celui des lampes à vapeur de mercure.

Exemple inspiré de G. Baillargeon (2002), p. 87–88.

Exemple :

une entreprise produit des lampes au sodium utilisées en particulier pour l'éclairage des autoroutes ainsi que des lampes à vapeur de mercure. En se basant sur les résultats du mois de mars figurant dans le tableau ci-dessous, on se demande laquelle des deux productions est de qualité plus homogène.

	lampe au sodium	lampe à vapeur de mercure
nombre d'essais	35	30
puissance [watts]	400	400
rendement énergétique moyen [lumens / watt]	115	40
écart-type [lumens / watt]	5.8	3.2

Source : Baillargeon, G. (2002). *Statistique Appliquée et outils d'amélioration de la qualité*, 2^e édition, SMG, Trois-Rivières ouest, Canada.

5. Boîte à moustaches (boxplot)

Particulièrement utile pour :

représenter une liste de données de manière compacte, comparer la position et la variabilité de différents groupes de données, détecter des observations atypiques.

Construction :

la construction d'une boîte à moustaches se base sur le résumé simple et pratique d'une distribution : $\hat{q}(25\%)$, $\hat{q}(50\%)$ et $\hat{q}(75\%)$.

1. On construit d'abord une boîte qui s'étend du premier quartile au troisième quartile en indiquant la médiane par un segment à travers la boîte. Le 50 % des valeurs observées se trouve dans la boîte.

2. Les **moustaches** sont ensuite dessinées afin d'obtenir des informations sur le comportement des queues de distribution des données et pour détecter d'éventuelles valeurs atypiques. Elles sont formées de deux segments ajoutés de chaque côté de la boîte de la manière suivante :

- calculer les bornes

$$\hat{q}(25\%) - 1.5 \cdot \{\hat{q}(75\%) - \hat{q}(25\%)\}$$

et

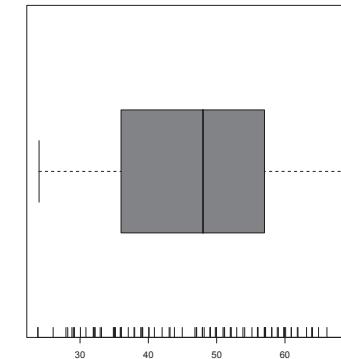
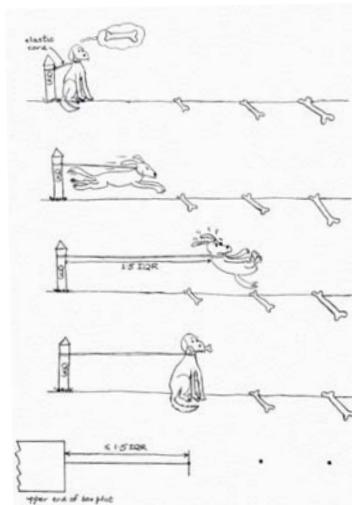
$$\hat{q}(75\%) + 1.5 \cdot \underbrace{\{\hat{q}(75\%) - \hat{q}(25\%)\}}_{\text{étendue interquartile}};$$

- déterminer la plus petite et la plus grande des valeurs observées situées entre les deux bornes. Ces valeurs sont appelées les **valeurs adjacentes**;

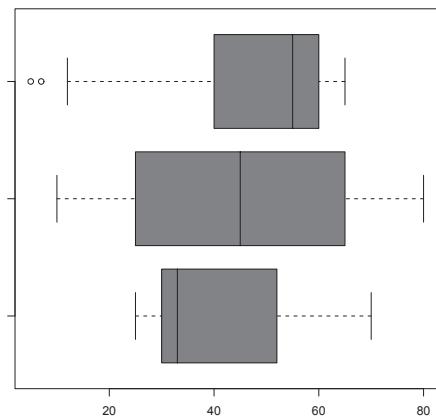
- tracer les segments qui relient les deux valeurs adjacentes à la boîte.

Remarques :

- les observations se situant en dehors des moustaches sont considérées comme valeurs atypiques. Elles nécessitent une attention particulière;
- la boîte à moustaches a été conçue de telle sorte que si les données sont issues d'une distribution normale (elle sera introduite plus tard dans le cours), approximativement le 99.5 % des observations se situe entre les moustaches. C'est pour cette raison qu'apparaît le facteur 1.5 dans le calcul des bords du graphique;
- la boîte à moustaches est très utile pour comparer des échantillons liés ou pour observer le degré d'asymétrie de la distribution.



Boîte à moustaches de l'épaisseur du placage en or sur les panneaux des circuits imprimés .

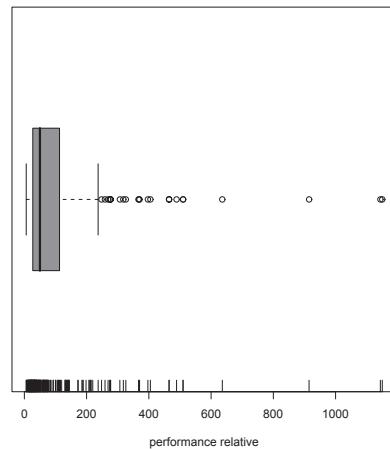


Distributions à différents degrés d'asymétrie (en haut : degré d'asymétrie négatif, au milieu : degré d'asymétrie nul, en bas : degré d'asymétrie positif).

Exemple :

la boîte à moustaches des performances relatives au processeur IBM 370/158-3 de 209 processeurs d'ordinateurs figure à la page suivante.

En raison de la position de la médiane dans la boîte et des longueurs des moustaches, le graphique montre une asymétrie positive. On y remarque également plusieurs valeurs atypiques.

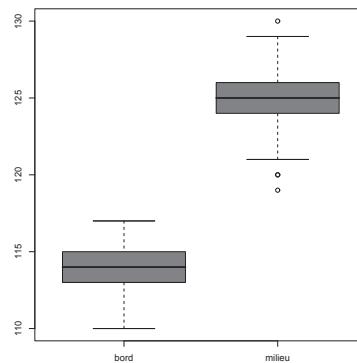


Boîte à moustaches des performances relatives de 209 processeurs d'ordinateurs.

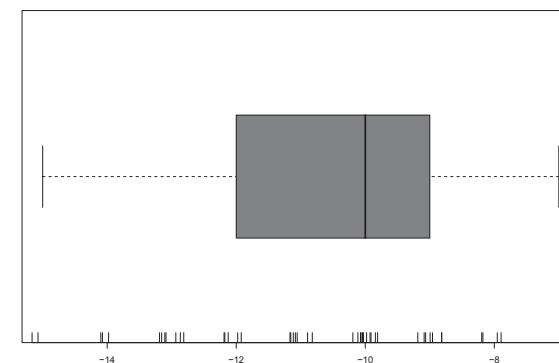
Exemple :

pour comparer l'épaisseur au bord et au milieu de 49 rouleaux fabriqués par une entreprise, on avait utilisé des histogrammes en parallèle. Une méthode plus efficace pour comparer des échantillons liés consiste à construire des boîtes à moustaches en parallèle. Elles se trouvent dans la figure de la page suivante. La différence existant entre les mesures au bord et au milieu se remarque distinctement.

Toutefois, la construction de ces boîtes à moustaches en parallèle n'est pas très logique. En effet, chaque rouleau est la source de deux observations. Par conséquent, on devrait plutôt s'intéresser pour chaque rouleau à la différence existant entre la mesure prise au bord et celle provenant du milieu du rouleau. Ainsi, construire la boîte à moustaches des différences est plus approprié pour comparer les endroits où sont relevées les épaisseurs.



Boîtes à moustaches des épaisseurs au bord et au milieu des rouleaux de film en plastique.



Boîte à moustaches des différences des épaisseurs entre le bord et le milieu des rouleaux.

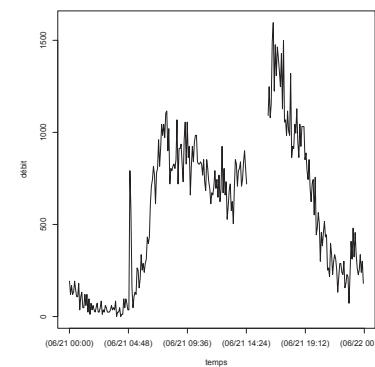
Une **série temporelle (ou chronologique)** est une suite d'observations qui arrivent les unes à la suite des autres, souvent à intervalles réguliers. Citons par exemple,

- ◊ la fluctuation dans le temps du nombre d'employés d'une entreprise;
- ◊ le cours journalier d'une action pendant une période donnée;
- ◊ le nombre hebdomadaire de pièces défectueuses d'un certain type produites dans une entreprise horlogère.

La technique la plus simple pour visualiser des données chronologiques consiste à placer le temps sur un axe, par exemple l'année. On associe ensuite un point à chaque unité de temps observée de telle sorte que la hauteur par rapport à l'axe soit égale à la valeur mesurée correspondante. Finalement, deux points consécutifs sont reliés par un segment.

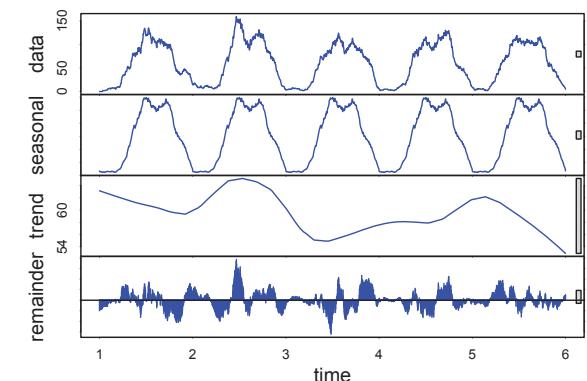
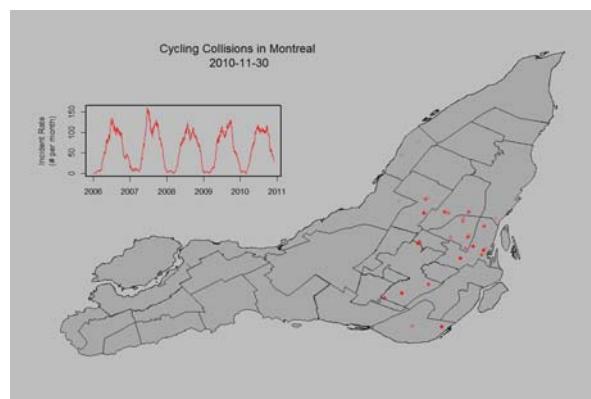
Exemple :

débit du trafic automobile au portail du tunnel de Glion le 21 juin 2004.

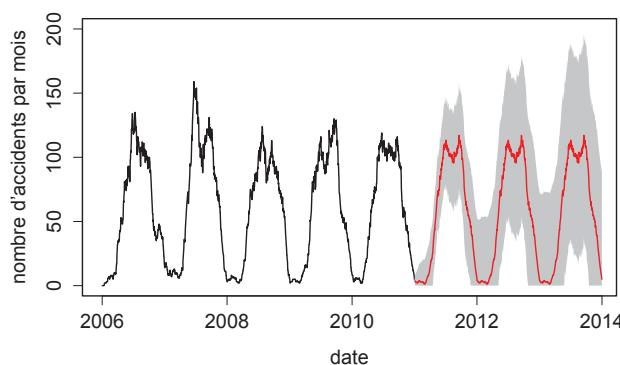


Exemple :

nombre d'accidents impliquant des cyclistes à Montréal entre 2006 et 2010.



Décomposition de la série temporelle.



Prévision du nombre d'accidents entre 2011 et 2013.

Les deux principaux graphiques pour représenter une variable qualitative sont

- ◊ le diagramme en camembert;
- ◊ le diagramme en barres.

Ils permettent de visualiser les proportions observées pour une variable qualitative (catégorique) soit à l'aide de portions d'un camembert soit à l'aide de boîtes placées les unes à côté des autres dont les hauteurs respectives sont données par les fréquences absolues ou relatives.

1. Diagramme en camembert (*pie chart*)

Exemple : le tableau ci-dessous énumère les parts de marché en 2005 des moteurs de recherche.

société	part de marché (%)
Google	46.2
Yahoo!	22.5
msn	12.6
AOL.fr	5.4
autres	13.3

Source : Nielsen / NetRatings.

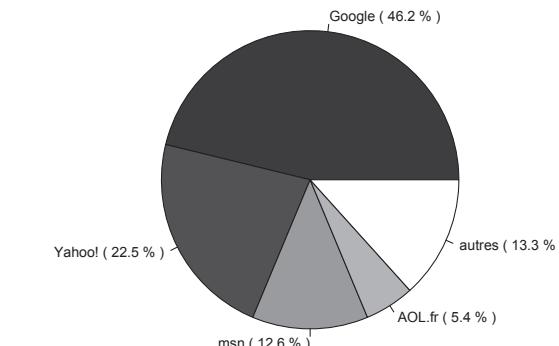


Diagramme en camembert des parts de marché des moteurs de recherche.

→ quelle critique faites-vous à ce graphique ?

2. Diagramme en barres (*bar graph*)

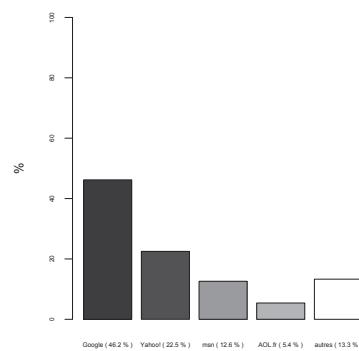


Diagramme en barres des parts de marché des moteurs de recherche.

“Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements.”

W.S. Cleveland and McGill

"Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data."

R Help



2.3 Techniques de visualisation pour deux variables

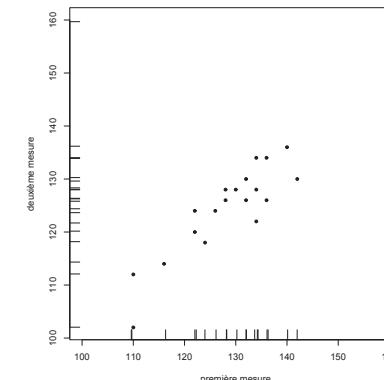
Il arrive souvent qu'on fasse deux observations sur chaque élément d'un échantillon comme par exemple la taille et le poids d'un individu. On souhaite ensuite trouver et comprendre la relation existante entre les deux variables.

Le graphique le plus simple pour visualiser le lien entre deux variables continues est le **nuage de points**.

1. Nuage de points (*scatter plot*)

Construction :

on utilise un système d'axes orthogonaux. La première variable est représentée en abscisse et la deuxième en ordonnée.

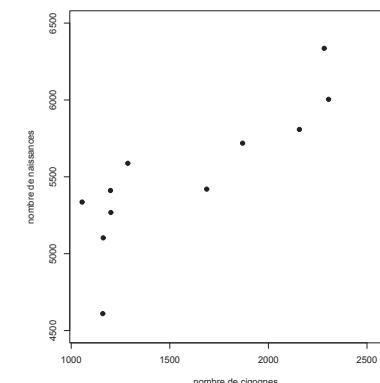


Tension artérielle systolique relevée à deux reprises sur 20 individus : la première 5 minutes après le début d'un entretien d'embauche, la deuxième plus tard.

Exemple :

dans l'album "Le fils d'Astérix" de Goscinny & Uderzo, Obélix avait rêvé que les cigognes étaient passées sur le village pour y déposer les commandes de bébés et que l'une d'elles avait fait l'erreur d'en déposer une devant sa hutte. Pendant plusieurs années, on a relevé à Copenhague le nombre de cigognes qui y sont passées et le nombre d'enfants nés ces années-là. Le graphique de nuage de points du nombre de naissances versus le nombre de cigognes figure à la page suivante.

Une relation linéaire semble exister entre les deux variables. Cependant, être en relation ne signifie pas toujours cause à effet. Une grande prudence dans l'interprétation est recommandée en statistique



Graphique de nuage de points entre le nombre de naissances et le nombre de cigognes passées à Copenhague.

Considérons deux variables X et Y dont on dispose de n observations présentées sous la forme de couples de nombres,

$$\left(\begin{array}{c} x_1 \\ y_1 \end{array} \right), \left(\begin{array}{c} x_2 \\ y_2 \end{array} \right), \dots, \left(\begin{array}{c} x_n \\ y_n \end{array} \right).$$

Pour quantifier le lien existant entre deux variables, on utilise la covariance définie par

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}), \quad (5)$$

avec $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Le degré de dépendance linéaire existant entre deux grandeurs X et Y est mesuré par le coefficient de corrélation noté r et défini par

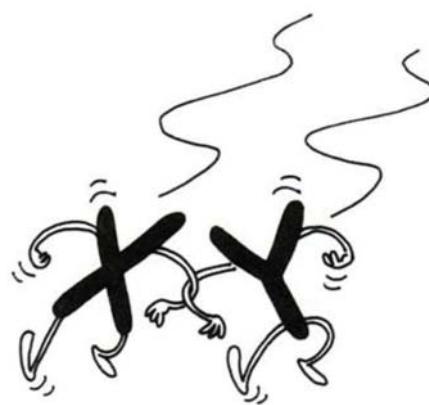
$$\begin{aligned} r &= \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n x_i \cdot y_i - n \bar{x} \bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n \bar{x}^2) \cdot (\sum_{i=1}^n y_i^2 - n \bar{y}^2)}}. \end{aligned} \quad (6)$$

Exemple : considérons les deux séries de nombres :

x_i	8	4	5	-1
y_i	-2	0	2	6

Par calculs directs, on obtient $\bar{x} = 4$, $\bar{y} = 1.5$, $\sum_{i=1}^4 x_i \cdot y_i = -12$, $\sum_{i=1}^4 x_i^2 = 106$ et $\sum_{i=1}^4 y_i^2 = 44$. Ainsi,

$$r = \frac{-12 - 4 \cdot 4 \cdot 1.5}{\sqrt{(106 - 4 \cdot 16) \cdot (44 - 4 \cdot 2.25)}} = -0.939.$$



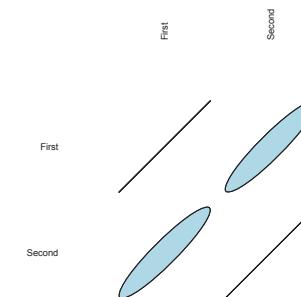
Remarques :

- la covariance dépend des unités respectives des grandeurs observées. En revanche, le coefficient de corrélation s'exprime sans unité. Plus précisément, ce coefficient est un nombre réel compris entre -1 et $+1$. Il s'ensuit que le coefficient de corrélation permet de comparer des variables très différentes comme par exemple le poids et la taille d'un individu;
- la corrélation peut s'écrire comme la covariance de deux grandeurs divisée par la racine carrée du produit des variances respectives;
- le coefficient de corrélation quantifie la partie linéaire de la relation existant entre deux grandeurs observées.

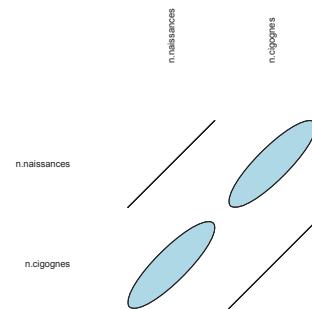
Le logiciel de statistique R permet d'illustrer graphiquement le coefficient de corrélation. Plus précisément, la forme (ellipse, cercle, segment) et la grandeur le représentent.

Dans l'interprétation des graphiques obtenus, il ne faut jamais oublier que le coefficient de corrélation se spécialise en quantifiant uniquement la partie **linéaire** de la relation existant entre deux grandeurs observées. On aura donc des difficultés à identifier un autre type de relation en se basant uniquement sur ce graphique.

Pour éviter de se faire piéger, il convient d'accompagner le coefficient de corrélation par un graphique de nuage de points.



Graphique pour illustrer le coefficient de corrélation entre les mesures des tensions artérielles.



Graphique pour illustrer le coefficient de corrélation entre le nombre de naissances et le nombre de cigognes passées à Copenhague.

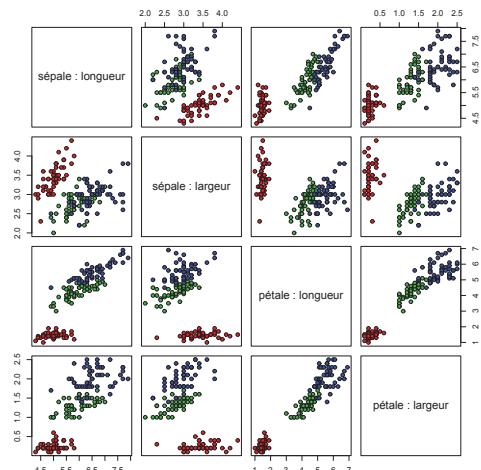
2. Matrice de nuages de points (*scatter plot matrix*)

Particulièrement utile pour :

déetecter et comprendre les relations existant entre plusieurs variables.

Construction :

chaque variable est représentée en fonction d'une autre variable dans un nuage de points. Les différents graphiques obtenus sont ensuite assemblés sous la forme d'une matrice de nuages de points.



Matrice de nuages de points de quatre variables observées sur des iris appartenant à trois espèces : Setosa (rouge), Versicolor (vert) et Virginica (bleu).

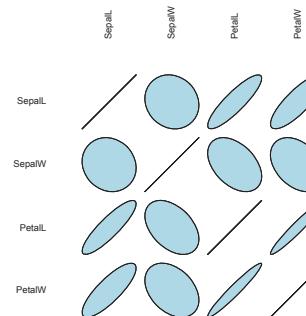


Illustration des coefficients de corrélation entre les variables mesurées sur les iris.

- Pour conclure, notons que les outils de l'analyse exploratoire des données, principalement graphiques, permettent de dépister les sources de problèmes telles que les
 - ◊ observations atypiques, erronées ou manquantes;
 - ◊ modalités trop rares;
 - ◊ distributions particulières (asymétrie, multimodalité, épaisseur des queues de distribution);
 - ◊ incohérences;
 - ◊ relation particulières;
 - ◊ ...

"If data analysis is to be well done, much of it must be a matter of judgement, and 'theory', whether statistical or non-statistical, will have to guide, not command."

John W. Tukey (1915–2000)

2.4 Conseils pour créer de bons graphiques en statistique

"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency."

Edward R. Tufte

"There are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not."

Edward R. Tufte

En statistique, les graphiques doivent

- illustrer clairement, efficacement les données avec précision;
- ne pas attirer l'attention sur la conception du graphique (méthode, technologie, design) mais plutôt pousser à la réflexion;
- éviter de déformer les informations contenues dans les données
→ empêcher aux graphiques de mentir !
- représenter beaucoup de nombres sur un minimum de place;
- rendre les grands jeux de données cohérents;

Mises en garde

- dans un graphique, montrer les variations des données, non pas les variations existant dans les figures du graphique;
- la “mauvaise représentation” d'un graphique peut être mesurée par le facteur suivant :

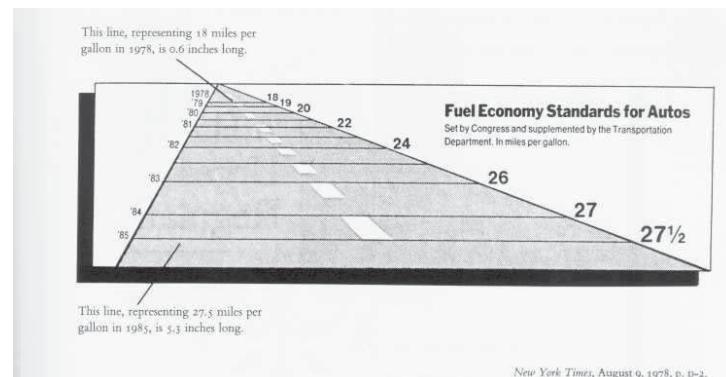
$$\frac{\text{grandeur de l'effet dans le graphique}}{\text{grandeur de l'effet dans les données}}$$

Le cas idéal est naturellement 1.

Question : que peut-on dire du graphique de la page suivante ?

En statistique, les graphiques doivent

- faciliter l'effort de visualisation et de compréhension tout en gardant les données dans leur contexte;
- encourager l'œil à comparer naturellement différentes parties des données;
- présenter les données à différents niveaux allant de l'aperçu global à une petite structure;
- prendre naissance d'une intention claire : description, exploration ou décoration;
- dans la conception des graphiques en statistique, on veut concilier la **simplicité de la forme, du design** avec la **complexité des jeux de données** tout en révélant la vérité contenue dans les données.

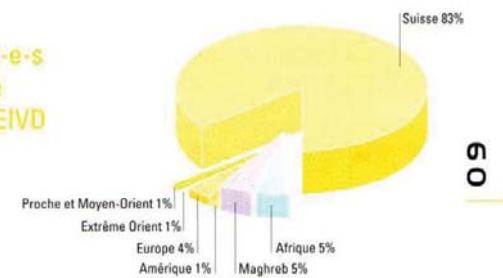


Projection de la diminution de la consommation d'essence des automobiles aux Etats-Unis.

Mises en garde (suite)

- sus aux ambiguïtés dans les graphiques; par exemple, les labels doivent être clairs et précis ↗ clareté !
- pas de fioritures inutiles, pas de bric-à-brac ↗ se concentrer sur l'information quantifiable;
- éviter les surfaces vides superflues;
- le concepteur d'un mauvais graphique porte une certaine responsabilité dans l'interprétation du graphique effectuée, par exemple, par une personne influente;

 Lieu de domicile des étudiant-e-s lors de l'obtention du diplôme leur ayant permis l'accès à l'EIVD



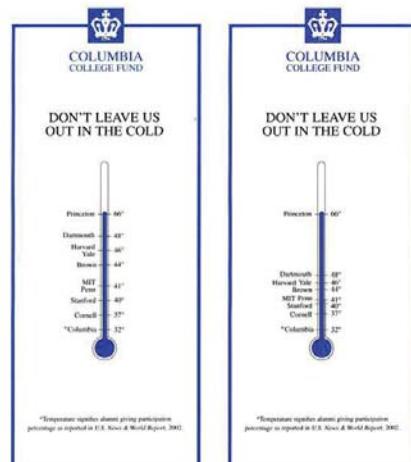
Mises en garde (suite)

- plutôt qu'un mauvais graphique, il est préférable de ne rien avoir du tout !
- ne faites pas dire n'importe quoi à un graphique ou à un résultat statistique ! C'est pas beau, pas beau du tout !
- ...

Exercice : que peut-on dire des graphiques des pages suivantes ?



Risque de compromettre sa santé en consommant de l'alcool.



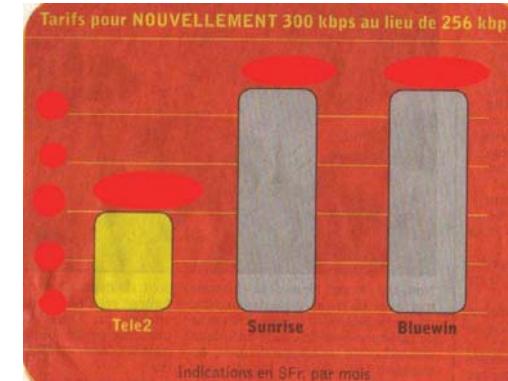
Températures relevées dans plusieurs villes des Etats-Unis.



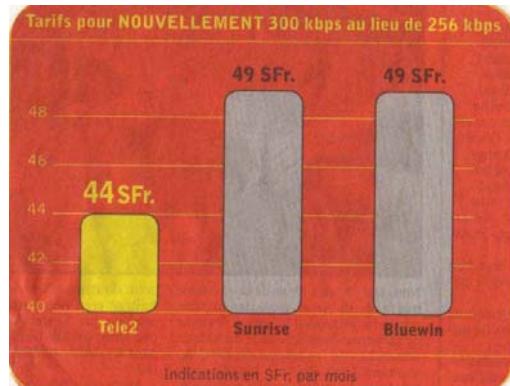
Résultat d'un sondage sur la composition d'une famille suisse.



Résultat d'un sondage sur la satisfaction de sa vie en Suisse.



Graphique sans échelle.

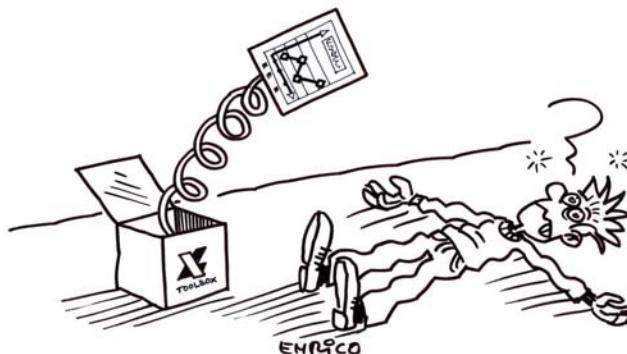


Même graphique avec échelle.

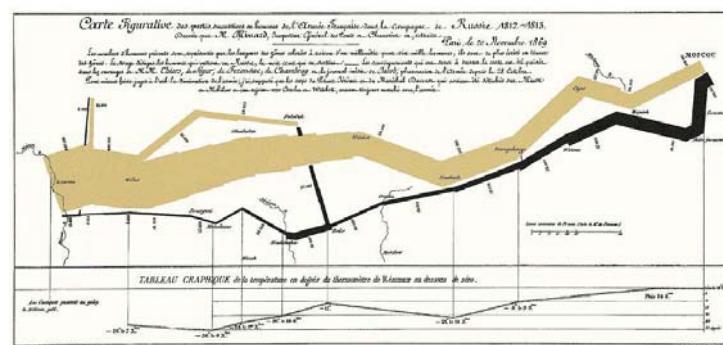
"Graphical excellence is the well-designed presentation of interesting data – a matter of substance, of statistics, and of design."

"Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

Edward R. Tufte



Créer de bons graphiques n'est pas une mince affaire !



Un joyau ! Graphique retracant la campagne de Napoléon en Russie; Minard (1861).

EXERCICES : ANALYSE EXPLORATOIRE DES DONNÉES

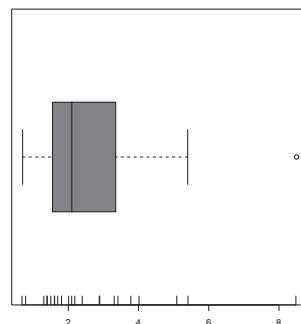
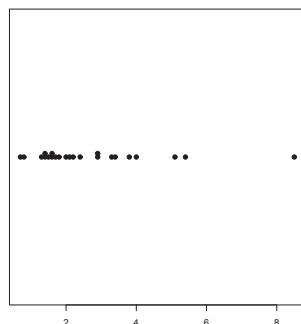
Exercice 1

Le temps nécessaire en secondes à un ordinateur pour compléter un fichier en puisant des données dans une gigantesque base de données a été relevé.

2.0 2.4 1.6 4.0 1.3 1.4 3.4 2.1 5.1 5.4 3.3 2.9
8.5 2.9 1.7 3.8 0.8 0.7 1.4 1.5 1.8 2.2 1.6

- Construire un diagramme en points, un diagramme branche-et-feuilles et une boîte à moustaches des données observées.
- Commenter la distribution des valeurs observées : valeur(s) atypique(s), asymétrie, aplatissement.

Solutions : a)



0	78
1	34456678
2	012499
3	348
4	0
5	14
6	
7	
8	5

- une valeur atypique (8.5), asymétrie positive, difficile de répondre avec assurance à l'aplatissement de la distribution; disons distribution moyennement aplatie.

Exercice 2

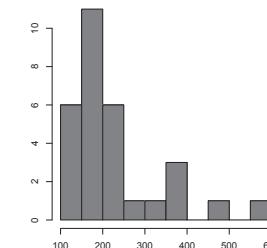
Dans le cadre d'une étude menée à l'université d'Auckland (Nouvelle-Zélande), 20 femmes âgées de moins de 40 ans ont suivi un régime basé sur l'hydrate de carbone. Les quantités en milligrammes consommées quotidiennement figurent dans le tableau qui suit.

199 162 327 145 149 351 453 374 287 151 201 375 223 230 193
229 206 144 152 164 121 190 158 145 129 168 173 189 589 247

- Écrire les données dans un tableau des fréquences en partant de la valeur 100 et en utilisant une amplitude de 50.

- Tracer l'histogramme correspondant au tableau des fréquences. Que peut-on en déduire ?

Solution : b)



En se basant sur l'histogramme, il existe certainement deux valeurs atypiques (453 et 589). On constate une asymétrie positive, pas aussi distincte que sur une boîte à moustaches.

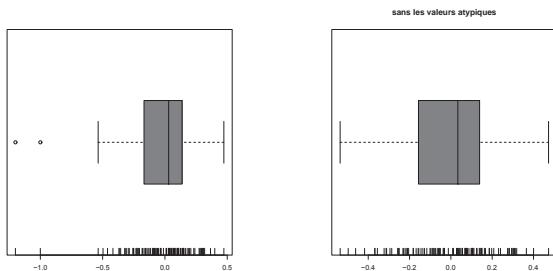
Exercice 3

Une firme produit des profilés en aluminium. Elle souhaite acheter une machine très sophistiquée pour diminuer les déchets lors de la coupe des profilés. Avant de passer commande, le directeur du secteur "Presses" de la firme souhaite vérifier la précision de la machine. Il se propose de couper 102 profilés à une longueur prédefinie puis de mesurer précisément l'erreur de coupe en millimètres (mm) à l'aide d'un instrument laser. L'erreur est négative pour les pièces trop courtes, positive pour les pièces trop longues. Les résultats obtenus figurent dans le fichier **coupe.txt** (données fictives).

- Construire une boîte à moustaches des données.
- Commenter la distribution des données : valeur(s) atypique(s), asymétrie.
- Déterminer les statistiques élémentaires des données en utilisant la fonction **summary** du logiciel de statistique **R**.
- La moyenne des erreurs de coupe est-elle un indicateur approprié de tendance centrale des données ?
- En se basant sur la boîte à moustaches et les statistiques élémentaires, peut-on affirmer que la machine est précise ?

Solutions : a) deux boîtes à moustaches figurent en page 3 : l'une à partir de toutes les valeurs observées, l'autre sans les valeurs atypiques b) la boîte à moustaches de gauche révèle deux valeurs atypiques, les plus petites valeurs observées -1.2 et -1, et une distribution légèrement asymétrique (asymétrie négative), presque nulle en ôtant les deux valeurs atypiques c) les statistiques élémentaires fournies par **R** sont

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.2000	-0.1660	0.0298	-0.0270	0.1350	0.4730



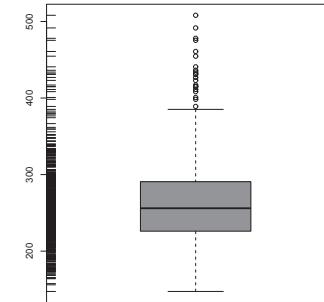
d) la moyenne des erreurs de coupe ne semble pas être un indicateur approprié de la tendance centrale des données. En effet, elle est influencée par les deux valeurs atypiques de telle sorte que la médiane et la moyenne ne sont pas très proches l'une de l'autre (médiane positive, moyenne négative) e) en se basant sur les valeurs observées, la machine paraît être précise si on excepte les valeurs atypiques (il faudrait d'ailleurs le vérifier; il s'agit peut-être d'une erreur de report ou de manipulation de la machine). La médiane et la moyenne sont proches de zéro.

Exercice 4

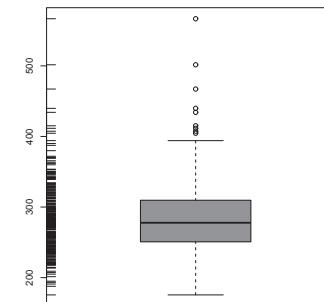
Le temps en minutes mis par 292 femmes choisies au hasard parmi les participantes pour courir le marathon de New-York en 2002 a été relevé. Une partie des valeurs observées ont été classées par ordre croissant et figurent ci-dessous.

```
[1] 175.53 187.73 190.93 193.32 194.95 201.35 204.93 207.15 209.07 213.77
[11] 213.92 214.63 217.48 218.38 218.72 219.43 219.50 220.25 220.58 221.52
...
[141] 276.57 277.08 277.38 277.60 277.63 277.68 277.77 278.12 278.15 278.80
[151] 278.87 279.82 280.07 280.10 280.88 280.90 281.13 281.27 281.78 282.25
...
[271] 359.87 361.85 363.47 365.57 369.35 369.35 370.20 370.60 371.98 379.85
[281] 387.40 390.00 394.08 404.55 407.43 411.57 415.13 434.18 439.83 467.18
[291] 501.55 566.78
```

- Calculer la médiane des temps observés.
- Le premier et le troisième quartiles des temps réalisés par les participantes valent respectivement 250.69 et 309.66. Construire soigneusement la boîte à moustaches des temps observés en ajoutant les valeurs atypiques si on en trouve.
- La boîte à moustaches des temps réalisés par 708 hommes choisis eux aussi au hasard parmi les participants à l'édition 2002 du marathon de New-York se trouve en page 4.
- À l'aide de la boîte à moustaches, commenter la distribution des temps réalisés par les hommes et traiter en particulier les valeurs atypiques et l'asymétrie. Préciser si l'asymétrie est positive, négative ou nulle.
- Calculer l'étendue interquartile des temps mis par les femmes pour courir le marathon. À l'aide de la boîte à moustaches, est-elle plus grande, plus petite ou égale à celle des hommes ? L'étendue interquartile est-elle un indicateur de dispersion robuste ou non ?



Solutions : a) 277.725 b)



- la médiane se trouve légèrement plus bas que le milieu de la boîte et la moustache supérieure est légèrement plus longue que la moustache inférieure. En ne considérant pas les valeurs atypiques, on peut dire que la distribution est symétrique, peut-être légèrement asymétrique, asymétrie positive. Toutefois, les valeurs atypiques supérieures introduisent une asymétrie positive plus marquée. Tout compte fait, en raison des valeurs atypiques supérieures, la distribution est asymétrique, asymétrique positive. On rencontre des valeurs atypiques du côté du maximum. Ce phénomène est fréquent dans les courses à pied d) 58.97. En se basant sur la boîte à moustaches, l'étendue interquartile des temps mis par les hommes n'est pas significativement différente de 60. Ainsi, les étendues interquartile des hommes et des femmes sont assez semblables. L'étendue interquartile est un indicateur robuste de dispersion étant donné qu'elle fait intervenir les quartiles, peu sensibles aux valeurs extrêmes.

Exercice 5

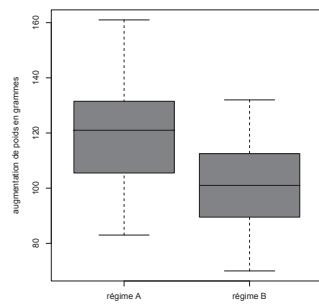
Un nutritionniste s'intéresse à la différence entre deux régimes *A* et *B* qui ont été appliqués à deux groupes de rats. On a relevé les accroissements en grammes du poids des animaux. Les résultats obtenus figurent dans le fichier **regime.txt**. La variable **gain.A** indique l'augmentation de poids des 12 rats soumis au régime *A* et **gain.B** celle des 7 rats soumis à l'autre régime.

- Calculer les statistiques élémentaires des poids selon les régimes en utilisant la fonction **summary** du logiciel de statistique **R**. En comparant les valeurs obtenues, que peut-on en conclure ?
- Pour comparer les poids suivant les régimes, construire le graphique qui vous semble être le plus approprié.
- Peut-on discerner une différence significative entre les régimes à partir du graphique tracé en b) ?
- La boîte à moustaches de l'augmentation des poids des rats soumis au régime *A* nous donne-t-elle une information sur la dispersion des valeurs observées ?

Solutions : a) les statistiques élémentaires fournies par **R** sont

	gain.A	gain.B
Min.	: 83	Min. : 70.0
1st Qu.	: 106	1st Qu. : 89.5
Median	: 121	Median : 101.0
Mean	: 120	Mean : 101.0
3rd Qu.	: 130	3rd Qu. : 112.5
Max.	: 161	Max. : 132.0

En comparant les statistiques élémentaires, l'augmentation du poids est plus grande pour les rats soumis au régime *A* qu'au régime *B* b) les boîtes à moustaches placées en parallèle ont été tracées. Elles permettent une comparaison aisée des régimes *A* et *B*. En effet, on dispose pour chaque régime d'une bonne vue d'ensemble des valeurs observées c) on peut discerner une différence significative entre les régimes à partir des boîtes à moustaches placées en parallèle. La boîte de gauche (régime *A*) est située plus haut que celle de droite (régime *B*) d) toute boîte à moustaches contient une information sur la dispersion des données : l'étendue interquartile



$\hat{q}(75\%) - \hat{q}(25\%)$ donnée par la hauteur de la boîte. Visuellement, on remarque que les deux boîtes ont quasiment la même hauteur. Les dispersions des valeurs observées des deux régimes semblent être relativement proches l'une de l'autre.

Exercice 6

Les hot-dogs américains sont réputés pour contenir une dose excessive de sodium due principalement à un emploi abusif de sel. Trois types de hot-dogs ont attiré l'attention d'un groupe de consommateurs : le hot-dog de bœuf, le mélange bœuf et de porc et le hot-dog de poulet. La quantité de sodium en milligrammes contenue dans les hot-dogs préparés par les principaux fast-food américains a été mesuré. Les résultats obtenus se trouvent dans l'objet **sodium**. Les variables s'appellent respectivement **beef**, **meat** et **poultry**. On se propose d'utiliser le logiciel de statistique **R** pour répondre aux questions posées.

- Calculer les minima, quartiles, moyennes, écarts-type, maxima et étendues interquartile du sodium contenu dans les trois types de hot-dogs. En comparant les valeurs obtenues, que peut-on conclure ?

Remarque : utiliser les commandes **summary** et **sd** pour calculer les valeurs cherchées. Comme les trois types de hot-dogs n'ont pas le même nombre d'observations, l'option **na.rm=T** doit être utilisée dans la fonction **sd** pour ne pas tenir compte des valeurs manquantes symbolisées par **NA** (Not Available).

- Tracer les boîtes à moustaches en parallèle des trois types de hot-dogs.
- Que peut-on discerner dans ce graphique ? Les conclusions qui en découlent corroborent-elles celles établies en a) ?

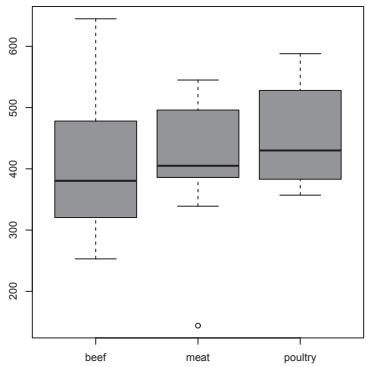
Les données se trouvent dans le fichier **sodium.txt**.

Solutions : a) les statistiques élémentaires fournies par le logiciel **R** sont

	beef	meat	poultry
Min.	: 253.0	Min. : 144.0	Min. : 357.0
1st Qu.	: 321.2	1st Qu. : 386.0	1st Qu. : 383.0
Median	: 380.5	Median : 405.0	Median : 430.0
Mean	: 401.1	Mean : 418.5	Mean : 459.0
Sd	: 102.43	Sd : 93.87	Sd : 84.74
3rd Qu.	: 477.5	3rd Qu. : 496.0	3rd Qu. : 528.0
Max.	: 645.0	Max. : 545.0	Max. : 588.0

Globalement, il semble qu'il existe une différence entre les trois types de hot-dogs, avec, dans l'ordre croissant, le hot-dog de bœuf, le mélange bœuf-porc et le hot-dog de poulet, ceci en se basant sur les quartiles, les moyennes et les médianes. Toutefois, il est difficile de conclure en raison de la variabilité plus grande pour le hot-dog de bœuf que pour les deux autres hot-dogs. De plus, le minimum pour le mélange bœuf-porc (bien plus petit que le 1er quartile) laisse présager une (plusieurs) valeur(s) atypique(s). Pour clarifier et nous permettre de conclure avec assurance, un bon graphique serait approprié b) les boîtes à moustaches placées en parallèle ont été tracées en page 7 c) on y remarque globalement une différence entre les trois types de hot-dogs en n'accordant pas une trop grande importance aux moustaches. En se basant sur les étendues interquartile, la variabilité du mélange bœuf-porc est plus petite que celle des deux autres types de hot-dogs. Le minimum de ce mélange est bien une valeur atypique. Les conclusions établies en c) corroborent celles provenant de a). Cependant, il est beaucoup plus facile de tirer des conclusions en comparant les boîtes à moustaches plutôt qu'en comparant les statistiques élémentaires.

Solutions : a)



Exercice 7

Le nombre d'abonnements de l'arrondissement de Lausanne entre 1970 et 1988 (état en fin décembre de chaque année) figure dans le tableau suivant :

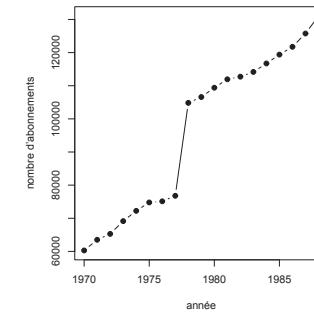
Année	Nombre d'abonnements	Année	Nombre d'abonnements	Année	Nombre d'abonnements
1970	60'286	1977	76'785	1984	116'724
1971	63'506	1978	104'830	1985	119'393
1972	65'286	1979	106'589	1986	121'744
1973	69'136	1980	109'404	1987	125'759
1974	72'239	1981	111'944	1988	131'021
1975	74'775	1982	112'730		
1976	75'118	1983	114'171		

- a) Construire un diagramme approprié pour représenter les données.
- b) Pour analyser la croissance du nombre d'abonnements de téléphone au fil des années, il convient de représenter graphiquement la différence du nombre d'abonnements de deux années consécutives. Pour le faire, on peut utiliser l'instruction suivante du logiciel de statistique R,

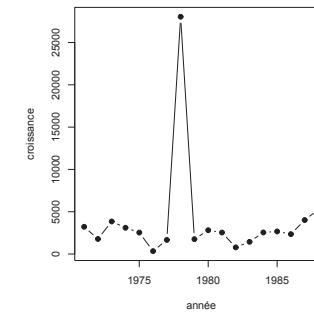
```
> plot.ts(ts(diff(telephone$abonnement), start=c(1971,1), frequency=1),
+ + xlab="année", ylab="croissance", type="b")
```

Commenter le graphique obtenu.

Les valeurs observées se trouvent dans le fichier **telephone.txt**.



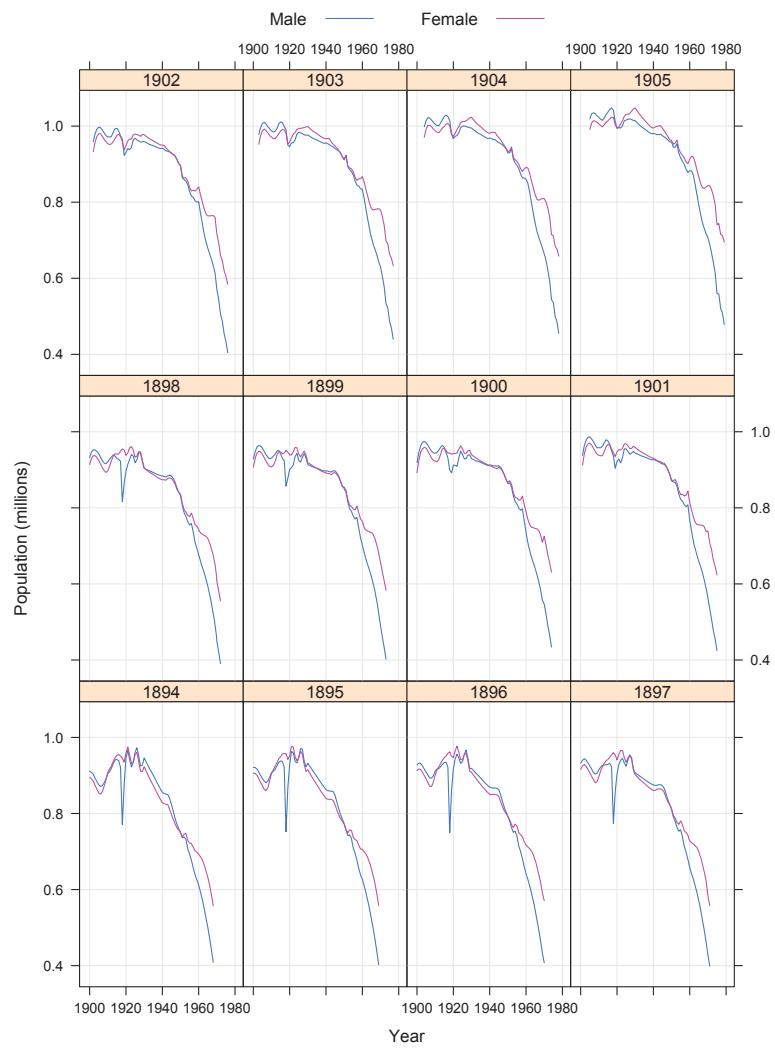
- b) on remarque un "pic". Il est dû à un saut enregistré en 1978 lorsque le nombre de nouveaux abonnés était de 28'045, valeur nettement supérieure à la moyenne des autres années.



Exercice 8

La population des États-Unis en millions d'habitants a été estimée depuis de nombreuses années par l'office du recensement. On s'est intéressé, par exemple, à l'évolution de la population selon l'âge (**Age**) et le sexe (**Sex**) pour des individus nés entre 1894 et 1905. Cette évolution au fil des ans entre 1900 et 1990 a été représentée graphiquement dans la figure se trouvant en page 9; les légendes **Male** et **Female** signifient respectivement homme et femme.

- a) Décrire la tendance générale de l'évolution de la population se dégageant du graphique.
- b) Constate-t-on une différence générale entre les femmes et les hommes ?
- c) Remarque-t-on un phénomène inhabituel sur chacun des graphiques de la figure ? Si oui, préciser lequel.



Solutions : a) plus le temps passe, plus la population décroît; on peut même y remarquer une décroissance exponentielle b) oui, à deux niveaux. En bas âge, on remarque davantage de garçons que de filles, cette tendance s'inverse par la suite c) on remarque les effets de la première guerre mondiale en particulier pour les garçons nés entre 1894 et 1900.

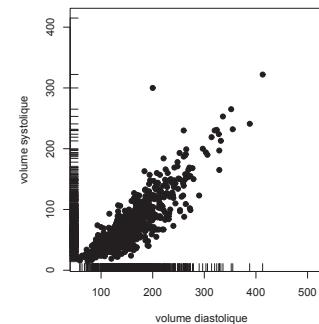
Exercice 9

À la suite d'une attaque cardiaque, la taille du cœur de plusieurs patients a été mesurée dans deux états différents ("end-systolic volume"; variable `sysvol` et "end-diastolic volume"; variable `diavol`).

- Construire un graphique de nuage de points `sysvol` versus `diavol`.
- Deux observations se distinguent des autres. Peut-on les considérer comme des observations atypiques ?

Les données se trouvent dans le fichier `heart.txt`.

Solutions : a)



b) bien que les volumes ne soient habituellement pas aussi élevés, une des deux observations particulières se situe dans la tendance dessinée par l'ensemble des observations. L'autre en revanche est effectivement une observation atypique. Elle nécessite une attention particulière concernant son origine (outil de mesure, report d'observation,...). Mise en garde : ne pas se fier aux projections sur les axes ! Elles ne reflètent aucunement la tendance entre les deux variables et peuvent cacher des observations atypiques comme l'illustre cet exemple.

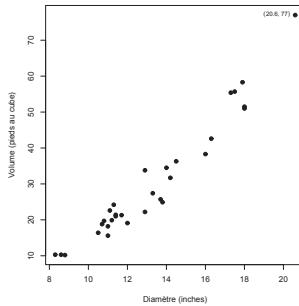
Exercice 10

Dans un parc américain, le diamètre du tronc de 31 arbres de même type ("black cherry tree") a été mesuré pour chacun d'eux à une hauteur de 4.5 inches du sol. Les arbres ont ensuite été abattus (aie ! Idéfix le déteste...). On a ensuite relevé le volume de bois de chaque arbre en pieds au cube. Les valeurs observées se trouvent dans le fichier `trees.txt`. Les variables `Diametre` et `Volume` indiquent respectivement le diamètre du tronc et le volume de bois de l'arbre.

- Construire un graphique de nuage de points `Y = Volume` versus `X = Diametre`.
- Commenter le graphique obtenu : valeur(s) atypique(s), type de relation éventuelle entre les variables.
- Calculer le coefficient de corrélation entre les deux variables. Ce résultat est-il en accord avec le graphique de nuage de points ?

- d) Quel point du graphique semble particulièrement influencer la tendance linéaire des données ?

Solutions : a)



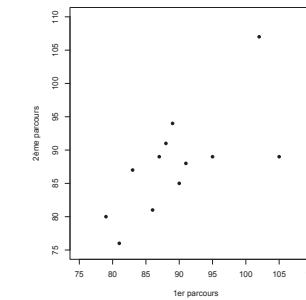
b) l'observation (20.6, 77) semble éloignée des autres observations, les valeurs 20.6 et 77 étant relativement grandes. Cependant, cette observation respecte parfaitement la tendance linéaire entre les variables mise en évidence par le graphique. Ainsi, elle ne sera pas considérée comme observation atypique c) la valeur du coefficient de corrélation entre les deux variables vaut 0.96712. Elle est très proche de 1. Une forte relation linéaire peut exister entre les deux variables. Ce résultat est en parfait accord avec le graphique de nuage de points tracé en a). Toutefois, il faut rester prudent si on se base uniquement sur le coefficient de corrélation; il peut être insidieux dans certains cas d) le point (20.6, 77) semble particulièrement influencer la tendance linéaire des données. Si ce point n'existe pas, on pourrait presque envisager une relation quadratique en raison des points de diamètres et volumes les plus petits.

Exercice 11

Les résultats obtenus par douze joueuses de golf sur un certain parcours ont été relevés. Les joueuses ont effectué deux fois le parcours. Rappelons que la vainqueur est celle qui a réalisé le plus petit score. Les résultats se trouvent dans le fichier **golf.txt**.

- Construire un graphique de nuage de points représentant pour chaque joueuse son résultat au second parcours en fonction de celui réalisé au premier parcours.
- Existe-t-il une relation entre les deux variables, c'est-à-dire entre les résultats des deux parcours ? Si oui, quel type de relation est-ce ?
- Estimer la valeur du coefficient de corrélation entre les deux variables. Corrobore-t-elle votre constatation faite en b).
- En observant le graphique, les observations (102, 107) et (105, 89) se démarquent des autres. L'une d'elles peut être considérée comme observation atypique. Laquelle ?
- En ne considérant pas l'observation atypique, la corrélation entre les deux variables va-t-elle augmenter ?

Solutions : a)



b) le nuage de points indique une relation linéaire entre les résultats des deux parcours c) le coefficient de corrélation entre les résultats du deuxième et du premier parcours est positif, modérément proche de 1. Il devrait se situer entre 0.6 et 0.8. La valeur du coefficient de corrélation fournie par le logiciel de statistique **R** est 0.69. Elle corrobore la constatation faite en b) d) l'observation (105, 89) peut être considérée comme atypique. En effet, elle se démarque de la tendance linéaire dessinée par les points représentant les autres observations e) en ne considérant pas l'observation atypique, la corrélation entre les deux variables va augmenter et s'approcher davantage de 1. Cette augmentation s'explique par le fait qu'en ôtant le point atypique, le nuage de points dessine une plus forte tendance linéaire.

Exercice 12

L'humoriste Pierre Desproges avait affirmé que "La moitié des Français est plus cons que la moyenne"¹. Mais à votre avis, lorsque l'on parle de moyenne, rencontre-t-on forcément autant de valeurs au-dessus qu'au-dessous de la moyenne ?

Solution : non, on n'aura pas forcément autant de valeurs au-dessus qu'au-dessous de la moyenne \bar{x} . En revanche, on aura autant de valeurs (à savoir la moitié) au-dessus qu'au-dessous de la médiane. La confusion entre médiane et moyenne est assez fréquente.

Considérons la suite de valeurs

95, 100, 105, 190.

Dans cette série, la moyenne \bar{x} vaut 122.5. On a une valeur au-dessus et trois valeurs au-dessous de la moyenne et, par conséquent, pas autant de valeurs au-dessus qu'au-dessous.

¹Cette citation n'engage aucunement votre professeur de statistique !

Chapitre 3

Probabilités élémentaires

3.1 Introduction

- La théorie des probabilités permet de décrire et modéliser des **phénomènes aléatoires**, i.e des phénomènes qui, lorsqu'ils sont observés sous des conditions déterminées, ne mènent pas toujours à la même issue.
- Les probabilités ont pris naissance dans l'étude des jeux de chance.
- Les bases de la théorie des probabilités ont été établies notamment par Blaise Pascal (1623–1662), Pierre de Fermat (1601–1668), Pierre-Simon de Laplace (1749–1827) et les Bernoulli répartis sur trois générations entre le XVII^e et le XVIII^e siècle.
- La modélisation probabiliste occupe une place importante autant dans le monde scientifique (météorologie, fiabilité, simulation, ...), économique et politique que dans la vie de tous les jours (lotto, casino, tiercé, ...).

Contenu

- 3.1 Introduction
- 3.2 Ensemble fondamental (ou univers) et événements
- 3.3 Mesure de probabilité (ou loi de probabilité)
- 3.4 Indépendance
- 3.5 Ensembles fondamentaux à événements élémentaires équiprobables



Exemples :

- jet d'une pièce de monnaie ou d'un dé;
- risques d'accidents;
- prédictions de résultats sportifs;
- nombre de fusibles défectueux dans un lot de vingt pièces;
- nombre de personnes entrant dans un établissement donné (banque, poste, ...);
- temps d'attente entre l'arrivée de deux messages électroniques;
- période de garantie de certains fours à micro-ondes;

Exemples (*suite*) :

- durée de vie d'un type de composants électroniques;
- fiabilité d'un dispositif électronique;
- temps requis pour effectuer le montage d'une pièce;
- largeur en centimètres d'une fente entaillée dans une pièce métallique; la largeur standard étant de 2 centimètres;
- trafic aérien (nombre d'arrivées et de départs) à l'aéroport de Cointrin durant une période de pointe;
- croissance d'une population naturelle (démographie);
- ...



- Dans le chapitre précédent nous avons résumé, synthétisé et visualisé les valeurs observées d'une variable quantitative à l'aide d'un histogramme. À plusieurs reprises, nous avons précisé que l'allure générale de l'histogramme pouvait nous conduire à des distributions particulières dites **distributions de probabilités**. Parmi ces distributions figure la distribution normale (gaussienne) vraisemblablement la plus connue lorsqu'on parle de modélisation probabiliste.
- Avant d'introduire les distributions usuelles de probabilités, fixons les objectifs de cette initiation aux probabilités et effectuons une brève introduction aux probabilités élémentaires dans ce chapitre et aux probabilités conditionnelles dans le chapitre suivant.

Objectifs

- présenter mathématiquement les principaux éléments de la théorie des probabilités;
- construire un modèle mathématique pour de nombreux phénomènes réels;
- présenter différentes applications à travers une foule d'exemples.

Quelques définitions de base (*suite*)...

- **Événement élémentaire** : événement formé d'une seule issue possible.
Notation : $A = \{\omega\}$.
- **Événement impossible** : événement qui n'aura jamais lieu. En d'autres termes, l'événement ne contient aucune réalisation.
Notation : \emptyset (ensemble vide).
- **Événement certain** : événement qui se réalisera à coup sûr. En d'autres termes, *quelque chose se passera !*
Notation : Ω .

3.2 Ensemble fondamental (ou univers) et événements

Quelques définitions de base...

- **Expérience (ou épreuve) aléatoire (ou stochastique)** : expérience dont il est impossible de prévoir le résultat (ou l'issue).
- **Ensemble fondamental (ou univers)** : ensemble de toutes les issues possibles de l'expérience aléatoire. L'ensemble fondamental est aussi appelé l'**espace fondamental** ou l'**univers**.
Notation : Ω .
- **Événement** : ensemble d'issues possibles d'une expérience aléatoire. En d'autres termes, un événement est un sous-ensemble de Ω .
Notation : A, B, C, \dots

Opérations sur les événements...

À partir de deux événements A et B , on peut définir de nouveaux événements à l'aide d'opérations.

- i) **Première opération** : **union**

Le nouvel événement noté $A \text{ ou } B$ en termes d'événements et $A \cup B$ en termes ensemblistes sera réalisé si **au moins un** des deux événements surviendra.

- ii) **Deuxième opération** : **intersection**

Le nouvel événement noté $A \text{ et } B$ en termes d'événements et $A \cap B$ en termes ensemblistes sera réalisé si **les deux** événements surviendront.

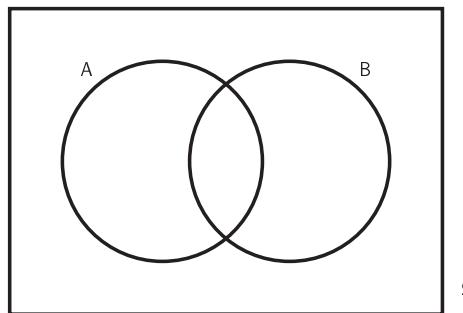


Illustration de l'union de deux événements A et B à l'aide d'un diagramme de Venn.

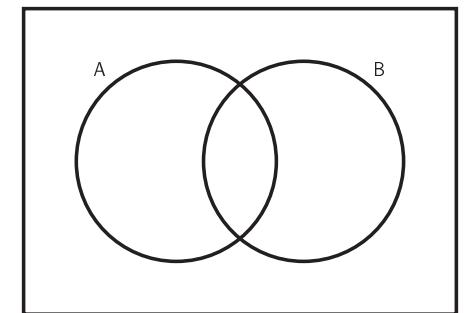


Illustration de l'intersection de deux événements A et B à l'aide d'un diagramme de Venn.

Opérations sur les événements (*suite*)...

iii) Troisième opération : complémentation

L'événement complémentaire à A noté *non A* en termes d'événements et \bar{A} en termes ensemblistes sera réalisé si A ne surviendra pas.

Incompatibilité de deux événements

Deux événements A et B sont **incompatibles** ou **exclusifs** s'ils ne pourront pas se réaliser en même temps, i.e si $A \cap B = \emptyset$. Autrement dit si l'événement A et B est impossible.

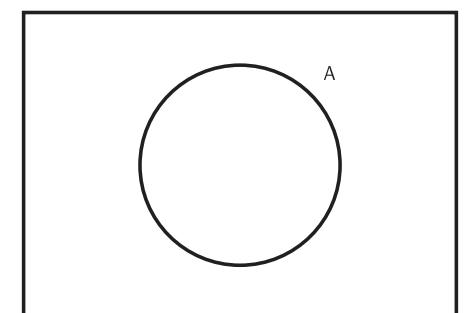


Illustration de l'événement complémentaire de l'événement A à l'aide d'un diagramme de Venn.

3.3 Mesure de probabilité (ou loi de probabilité)

Définir une mesure de probabilité sur Ω signifie associer à chaque événement A de Ω un nombre $P(A)$ qui satisfait les trois axiomes suivants :

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. Si A et B sont deux événements incompatibles,

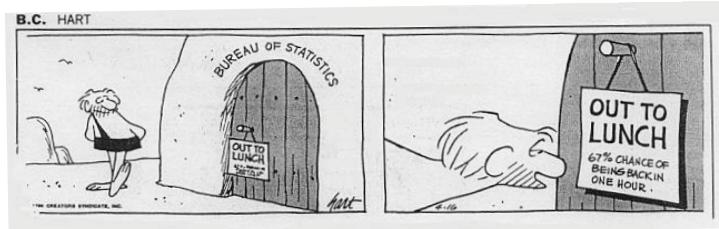
$$P(A \cup B) = P(A) + P(B).$$

Quelques résultats immédiats...

- i) $P(\emptyset) = 0.$
- ii) $P(\overline{A}) = 1 - P(A).$
- iii) Si $A \subset B$, $P(A) \leq P(B).$
- iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B).$

Remarque :

l'axiome 3 peut être généralisé à toute séquence d'événements mutuellement incompatibles $A_1, A_2, A_3 \dots$ (i.e d'événements pour lesquels $A_i \cap A_j = \emptyset$ si $i \neq j$).



3.4 Indépendance (première idée... qu'on approfondira...)

Définition 1 L'événement A est *indépendant* de l'événement B si le fait de savoir que B s'est déroulé n'influence pas la probabilité de A .

En langage probabiliste, l'indépendance entre les événements A et B se traduit par

$$A \text{ et } B \text{ sont indépendants} \Leftrightarrow P(A \cap B) = P(A) \cdot P(B).$$

Remarque :

l'indépendance est supposée dans de nombreuses applications comme par exemple dans la répétition d'un lancer de dé, dans le tirage de boules dans une urne avec remise, dans le fonctionnement d'un dispositif d'alarmes dans une banque. La notion d'indépendance est fondamentale dans le calcul des probabilités.

3.5 Ensembles fondamentaux à événements élémentaires équiprobables

Pour de nombreuses expériences, il est naturel d'admettre que chaque élément (chaque événement élémentaire) de l'ensemble fondamental a la même probabilité de se réaliser.

Plus précisément, considérons une expérience dont l'ensemble fondamental Ω est fini, disons $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$. Il est alors souvent naturel de supposer que

$$P(\{\omega_1\}) = P(\{\omega_2\}) = \dots = P(\{\omega_n\}),$$

ce qui implique par les axiomes des probabilités que

$$P(\{\omega_i\}) = \frac{1}{n}, \quad i = 1, 2, \dots, n. \quad (1)$$

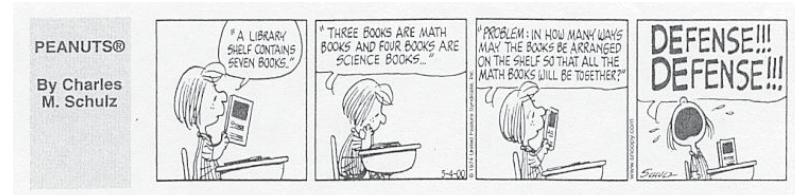
Du résultat (1) et des axiomes 2 et 3 des probabilités, il résulte sous hypothèse d'équiprobabilité que pour tout événement A d'un ensemble fondamental fini Ω ,

$$\begin{aligned} P(A) &= \frac{\text{nombre d'événements élémentaires de } A}{\text{nombre total d'événements élémentaires}} \\ &= \frac{\text{nombre de cas favorables à } A}{\text{nombre total de cas possibles}} \end{aligned} \quad (2)$$

Comme la probabilité $P(A)$ modélise la notion intuitive de "chance de réalisation de A ", la relation (2) se justifie.

Remarques :

- calculer $P(A)$ revient à déterminer le nombre d'éléments de A et de Ω . Ainsi, on aura très souvent recours à l'**analyse combinatoire**. Malheureusement, le dénombrement n'est pas toujours facile !
- pour que l'hypothèse d'équiprobabilité se justifie dans une expérience, il faut se veiller à définir un ensemble fondamental adéquat;



L'analyse combinatoire est parfois bien compliquée...

Exemple :

deux pièces de monnaie équilibrées sont jetées de suite. Considérons les deux ensembles fondamentaux

a) $\Omega_1 = \{\text{PILE \& PILE, PILE \& FACE, FACE \& FACE}\}$

b) $\Omega_2 = \{(\text{PILE, PILE}), (\text{PILE, FACE}), (\text{FACE, PILE}), (\text{FACE, FACE})\}.$

L'un des deux ensembles ne permet pas d'utiliser l'hypothèse d'équiprobabilité. Lequel ?

Remarques (suite) :

- comme mentionné précédemment, il est naturel de supposer équiprobables toutes les issues d'une expérience :
 - ▷ cette hypothèse se justifie par exemple pour le jet d'une pièce de monnaie équilibrée (deux issues possibles PILE ou FACE) ou pour le lancer d'un dé non pipé (six résultats possibles 1, 2, ..., 6);
 - ▷ il s'agit parfois d'une hypothèse simplificatrice mais tout à fait acceptable en première approximation. Par exemple, une personne a autant de chance d'avoir son anniversaire le vingt-quatre décembre que n'importe quel autre jour de l'année;

Remarques (suite) :

- ▷ dans certaines situations, l'hypothèse d'équiprobabilité ne peut pas être admise. Si $\Omega = \{\text{GAGNÉ, PERDU}\}$ décrit le résultat obtenu, il est raisonnable de supposer les deux issues équiprobables si vous affrontez Forrest Gump à "pile ou face" avec une pièce de monnaie équilibrée. Il n'en est pas de même si l'épreuve est un sprint de 100 m ou une partie de ping-pong.

"On réalise en fin de compte que la théorie des probabilités n'est tout simplement que le bon sens réduit à du calcul. Elle nous fait apprécier avec exactitude ce que l'esprit bien fait sent déjà par une sorte d'instinct, souvent sans être capable d'en rendre compte..."

P.-S. de Laplace (1749–1827)

"J'ai découvert les probabilités et ai très vite décidé que c'était ce que je voulais faire !"

Nicole El Karoui, *Le Monde*, "La bosse des maths",
édition du mardi 16 mai 2006, p.24



Qui se souvient que le 53 a rendu fous les Italiens ?

EXERCICES : PROBABILITÉS ÉLÉMENTAIRES

Exercice 1

Cent disques plastiques ont été testés pour analyser leur résistance aux égratignures et aux chocs. Les résistances sont qualifiées de *haute* ou *basse*. Les résultats obtenus figurent dans le tableau ci-dessous.

résistance aux égratignures	résistance aux chocs	
	haute	basse
haute	70	9
basse	16	5

Un disque a été choisi au hasard.

- a) Calculer les probabilités d'avoir choisi un disque

- de haute résistance aux égratignures et aux chocs;
- de haute résistance aux égratignures ou de haute résistance aux chocs.

- b) Quelle(s) hypothèse(s) faites-vous pour déterminer ces probabilités ?

- c) Les événements “un disque est de haute résistance aux égratignures” et “un disque est de haute résistance aux chocs” sont-ils incompatibles ?

Solutions : a) 0.7, 0.95 b) équiprobabilité c) ils ne sont pas incompatibles.

Exercice 2

Un réservoir cylindrique est utilisé comme réserve d'eau pour une certaine ville. Son approvisionnement n'est pas complètement prévisible : chaque jour, la quantité d'eau arrivant dans le réservoir peut augmenter le niveau de ce dernier de 6 m, 7 m ou 8 m; chaque possibilité ayant la même chance de survenir. La demande en eau est aussi variable et peut, avec des chances égales, diminuer le réservoir de 5 m, 6 m ou 7 m d'eau.

- a) Déterminer les combinaisons possibles d'arrivées et de pertes d'eau dans le réservoir un jour donné.
b) En supposant qu'au début de la journée le niveau dans le réservoir est de 7 m, quelle est la probabilité qu'il reste au moins 9 m d'eau dans le réservoir à la fin de la journée ?

Solutions : a)

arrivée d'eau en m	demande d'eau en m	différence en m
6	5	1
6	6	0
6	7	-1
7	5	2
7	6	1
7	7	0
8	5	3
8	6	2
8	7	1

- b) 1/3.

Exercice 3

Considérons deux événements A et B . Si $P(A) = 0.4$ et $P(B) = 0.7$, calculer la probabilité $P(A \cap B)$ dans les cas suivants :

- a) A et B indépendants;
- b) A et B incompatibles;
- c) $A \subset B$;
- d) $P(A \cup B) = 1$;
- e) $P(\overline{A} \cup \overline{B}) = 0.7$.

Solutions : a) 0.28 b) l'hypothèse d'incompatibilité n'est pas valable ou les probabilités sont fausses c) 0.4 d) 0.1 e) 0.3.

Exercice 4

Un étudiant de la HEIG-VD suit un cours de programmation orientée objet et un cours de mathématiques. La probabilité qu'il réussisse l'examen de programmation orientée objet est 0.5, celle de l'examen de mathématiques vaut 0.7 et celle qu'il réussisse les deux examens est 0.3. Calculer les probabilités que l'étudiant

- a) réussisse au moins un des deux examens;
- b) échoue aux deux examens;
- c) échoue en mathématiques mais réussisse en programmation orientée objet.

Solutions : a) 0.9 b) 0.1 c) 0.2.

Exercice 5

En raison de la hausse des températures, deux capteurs ont été installés pour observer le recul d'un certain glacier. Les capteurs sont très sensibles à plusieurs facteurs environnementaux et peuvent ainsi tomber en panne. Une étude a montré que la probabilité que le capteur A tombe en panne un jour donné vaut 0.02, celle que le capteur B flanche un jour donné est 0.045 et celle qu'ils soient simultanément en panne vaut 0.005.

- a) Calculer la probabilité qu'au minimum un des deux capteurs tombe en panne un jour donné.
- b) Déterminer la probabilité que l'un des deux capteurs soit en panne un jour donné mais pas l'autre.

Solutions : a) 0.06 b) 0.055.

Exercice 6

Dans une étude, on s'intéresse à la capacité que possède un certain “hacker” pour trouver en un temps donné les mots de passe permettant d'accéder à trois centres de calculs. La probabilité

que le “hacker” trouve le mot de passe des centres de calculs valent respectivement 0.22, 0.3 et 0.28. La probabilité qu'il trouve le mot de passe des deux premiers centres est 0.11, celle pour le premier et le troisième vaut 0.14 et celle pour déterminer le mot de passe du deuxième et du troisième centre est 0.1. Finalement, le “hacker” identifie les trois mots de passe avec probabilité 0.06.

- a) Calculer la probabilité que le “hacker” ne trouve aucun mot de passe.
- b) Déterminer la probabilité qu'il identifie au minimum deux des trois mots de passe.

Solutions : a) 0.51 b) 0.23.

Exercice 7

Trois amis déposent leur chapeau au vestiaire en entrant dans un restaurant et choisissent au hasard en sortant un des trois chapeaux.

Calculer les probabilités suivantes :

- a) aucun des trois amis ne prend son propre chapeau;
- b) exactement deux des trois amis prennent leur propre chapeau;
- c) les trois amis choisissent leur propre chapeau.

Solutions : a) 1/3 b) 0 c) 1/6.

Exercice 8

Une classe de la HEIG-VD est formée de 20 étudiants. En ne considérant pas le 29 février, déterminer la probabilité qu'au moins deux étudiants de cette classe aient leurs anniversaires le même jour. Quelle hypothèse faites-vous pour calculer cette probabilité ?

Solution : 0.42 en supposant équiprobabilité dans l'apparition des jours d'anniversaire.

Exercice 9

Gaston, en visite à Paris, traverse le Pont Mirabeau avec un cornet contenant cinq olives mûres et une olive gâtée. Il les prend une par une au hasard. Si l'olive est mûre, il la mange; si elle est gâtée, il jette le cornet à la Seine.

- a) Calculer la probabilité qu'il mange exactement trois olives.
- b) Parmi les olives mûres se trouvent deux vertes et trois noires. Déterminer la probabilité qu'il mange d'abord toutes les olives vertes puis toutes les olives noires.

Solutions : a) 1/6 b) 1/60.

Exercice 10

Un jardinier a mélangé trois oignons de tulipes rouges avec trois oignons de tulipes jaunes. Il les plante régulièrement en cercle en les prenant au hasard l'un après l'autre. Les oignons sont distinguables. Calculer la probabilité que les fleurs jaunes forment un triangle rectangle.

Solution : 0.6.

Exercice 11

Parmi les vingt brochets qui font la terreur des petits poissons d'un lac, cinq sont capturés, marqués puis relâchés. Plus tard, quatre brochets sont attrapés. Déterminer la probabilité que deux d'entre eux soient marqués. Quelle hypothèse faites-vous pour calculer cette probabilité ?

Solution : 0.21.

Exercice 12

Les 5ème et 6ème étages réunis des Éditions Dupuis sont formés de 12 employés; 6 travaillent au 6ème et les autres au 5ème. Pour affronter le bureau Smith dans un match de volleyball à 6 joueurs par équipe, Gaston Lagaffe décide, par souci d'équité, de choisir au hasard les 6 joueurs formant l'équipe des Éditions Dupuis. Calculer la probabilité que chaque étage compose la moitié de l'équipe.

Solution : 0.43.

Exercice 13

Quatre cartes sont tirées au hasard d'un jeu de jass. Calculer la probabilité d'obtenir

- a) 4 as;
- b) au moins 2 as.

Solutions : a) $1.698 \cdot 10^{-5}$ b) $5.271 \cdot 10^{-2}$.

Exercice 14

Considérons deux dés parfaitement équilibrés. Le Chevalier de Méré, noble de la cour de Louis XIV qui adorait les jeux de chance, pensait qu'il était plus probable d'obtenir un 6 en jetant 4 fois un dé plutôt que d'obtenir une paire de 6 en lançant 24 fois deux dés.

- a) Calculer la probabilité d'obtenir au moins un 6 en jetant 4 fois un des deux dés.
- b) Déterminer la probabilité d'obtenir au moins une paire de 6 en lançant 24 fois les deux dés.
- c) Le Chevalier de Méré avait-il raison ?

Solutions : a) 0.52 b) 0.49 c) le Chevalier de Méré avait tout à fait raison.

Exercice 15

Le lancer de deux dés non pipés est répété n fois.

- a) Calculer la probabilité qu'un double six apparaisse au moins une fois.
- b) Déterminer la valeur à attribuer à n pour que la probabilité calculée en a) soit supérieure à 1/2.

Solutions : a) $1 - \left(\frac{35}{36}\right)^n$ b) $n \geq 25$.

Chapitre 4

Probabilités conditionnelles et indépendance

4.1 Probabilités conditionnelles

Une probabilité n'est bien définie qu'en fonction des informations disponibles sur le résultat de l'expérience aléatoire. Il en résulte en particulier que la valeur d'une probabilité peut changer si une partie de l'information concernant le résultat de l'expérience est acquise au cours de son déroulement. La notion de connaissances partielles de l'information occupe une place importante dans le calcul des probabilités.

Supposons qu'on s'intéresse à la probabilité de l'événement A . Si **on sait** qu'un événement B est réalisé, il se peut que la probabilité de A soit modifiée par cette information.

On appelle **probabilité conditionnelle** de A étant donné B (ou en connaissant B ou encore en sachant que B est réalisé) cette nouvelle probabilité notée $P(A | B)$.

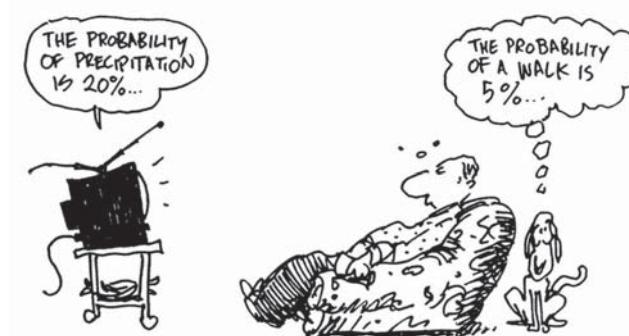
Contenu

4.1 Probabilités conditionnelles

4.2 Théorème de Bayes

4.3 Indépendance

4.4 Tirages et schéma de Bernoulli



Question :

que vaut la probabilité $P(A | B)$?

Réponse :

faisons comme si Ω était remplacé par B (changement de référentiel !). Alors,

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (1)$$

Il va de soi que la probabilité $P(B)$ est supposée être différente de zéro.

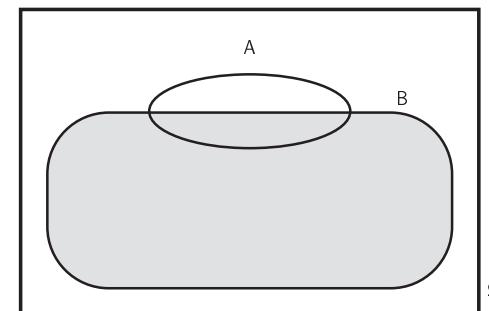


Illustration du calcul de la probabilité conditionnelle $P(A | B)$.

Remarques :

- il se peut que la probabilité conditionnelle $P(A | B)$ soit strictement plus petite, strictement plus grande ou égale à $P(A)$;
- on notera que $P(B | B) = 1$;
- un cas particulier important :

si A est inclus dans B , alors $A \cap B = A$ et donc

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

Deux résultats...

i) Théorème de multiplication :

$$\begin{aligned} P(A \cap B) &= P(A | B) \cdot P(B) \\ &= P(B | A) \cdot P(A). \end{aligned}$$

ii) Théorème des probabilités totales :

- D'abord une définition...

Définition 1 Soit un ensemble fondamental Ω . On appelle **partition** de Ω , ou encore **système complet d'événements**, tout ensemble H_1, H_2, \dots, H_k de sous-ensembles 2 à 2 disjoints de Ω dont la réunion est Ω .

- Ensuite le théorème...

Soient Ω un ensemble fondamental, H_1, H_2, \dots, H_k une partition de Ω et A un événement. Alors,

$$\begin{aligned} P(A) &= P(A|H_1) \cdot P(H_1) + P(A|H_2) \cdot P(H_2) \\ &\quad + \dots + P(A|H_k) \cdot P(H_k). \end{aligned}$$

Application du théorème des probabilités totales

Soient A et B deux événements quelconques. Comme B et \bar{B} forment une partition de Ω , on aura selon le théorème des probabilités totales,

$$\begin{aligned} P(A) &= P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B}) \\ &= P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot (1 - P(B)). \end{aligned} \quad (2)$$

Question :

comment peut-on interpréter le résultat (2) ?

Une possibilité :

la probabilité de l'événement A est une moyenne pondérée de la probabilité conditionnelle de A lorsque B s'est réalisé et de la probabilité conditionnelle du même A lorsque B n'a pas eu lieu. Les poids sont donnés par les probabilités des événements conditionnans.

La formule (2) est extrêmement utile puisqu'elle nous permet dans bien des cas de déterminer la probabilité d'un événement en commençant par le conditionner selon l'apparition ou non d'un autre événement. Dans de nombreuses situations, il est difficile de calculer **directement** la probabilité d'un événement mais il est en revanche possible de la calculer connaissant ses probabilités conditionnelles si certains événements sont réalisés.

4.2 Théorème de Bayes

Le théorème de Bayes qui fait appel aux théorèmes de multiplication et de probabilités totales est très important. Par exemple, il donna naissance à une autre approche de la statistique. Nous présenterons d'abord la version simple du théorème puis sa version composée.

◊ Formule de Bayes [version simple].

Supposons que A et B soient deux événements d'un ensemble fondamental Ω , avec $P(B) \neq 0$. Alors,

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

◊ Formule de Bayes [version composée].

Soient une partition H_1, H_2, \dots, H_k et un événement B d'un ensemble fondamental Ω , avec $P(B) \neq 0$. Pour tout indice j , $1 \leq j \leq k$, on aura,

$$\begin{aligned} P(H_j | B) &= \frac{P(H_j \cap B)}{P(B)} \\ &= \frac{P(B | H_j) \cdot P(H_j)}{P(B | H_1) \cdot P(H_1) + \dots + P(B | H_k) \cdot P(H_k)}. \end{aligned}$$

Remarque :

la formule de Bayes est davantage qu'une simple "formule mathématique". L'interprétation qui suit l'illustre.

Interprétons les événements H_j comme étant les états dans lesquels peut se trouver un système et B le résultat d'une expérience réalisée dans le but de déterminer dans lequel des états H_j le système se trouve. Alors,

- $P(H_j)$: probabilité **a priori**; opinion de l'observateur **avant** l'expérience;
- $P(B | H_j)$: probabilité d'observer B **si** le système se trouve dans l'état H_j ;
- $P(H_j | B)$: probabilité **a posteriori**; opinion de l'observateur après l'expérience sachant que B s'est réalisé.

Morale :

la formule de Bayes apparaît comme une **règle méthodologique**. En effet,

elle précise comment on doit modifier son opinion en tenant compte d'une expérience.

Application : filtre bayesian anti-spam

Dans le milieu informatique, le courrier électronique publicitaire (spam) est un véritable fléau. Le nombre de messages spam est quotidiennement en hausse. On prévoit même que le taux de messages électroniques publicitaires avoisinera le 70 % des messages d'ici 2007.

Actuellement pour combattre le problème grandissant du courriel publicitaire, l'outil le plus efficace est sans doute le filtre bayesian. Grâce à ce filtre, dont le fonctionnement repose sur le théorème de Bayes, on parvient maintenant à un taux de détection de spam de plus de 98 %.

Mais comment fonctionne le filtre bayesian pour détecter aussi bien le courriel publicitaire ?

de: postmaster@heig-vd.ch <postmaster@heig-vd.ch> à:
sujet: Rapport de quarantaine Malicleaner
pour: ZUBER Jacques <Jacques.Zuber@heig-vd.ch>

Rapport de quarantaine pour les 7 derniers jours pour l'adresse jacques.zuber@heig-vd.ch — Du 08.08.10 au 15.08.10
Etat global de la quarantaine, paramétrage, statistiques ou aide: https://mail0.heig-vd.ch

Rappel des actions que vous pouvez effectuer directement depuis ce rapport

[Libérer le message](#) [Aperçu du message](#) [Demander l'ajustement du filtre](#)

En quarantaine : 460 messages

Action	Date	Expéditeur	Objet	Score
	8-8-2010 00:21:26	natalia_co_@handmaker...	100 Cialis Pills x 20mg 2899 buy cialis...	██████
	8-8-2010 00:24:21	flowlabh_@salisade.c...	15mg x 60 Codeine \$264.00 (+4 FreeViagr...	██████
	8-8-2010 02:07:13	prosperity_@abonora.co...	Best for your nimrod	██████
	8-8-2010 02:14:07	novome_@lobrasl...	65% off. Affordable Watches	██████
	8-8-2010 02:55:55	lavada.yee...@gatesalber...	iggy 10	██████
	8-8-2010 03:18:28	acassoneqf_@unitymedia...	No More Side Effects	██████
	8-8-2010 04:40:36	timika.ie_@ge.cokec...	Explode, intenseOrgaans, Increase Vol...	██████
	8-8-2010 04:45:30	blindlydr_@twa.com	More drive for in-out stocking	██████
	8-8-2010 05:30:12	michaela.p_@aspenonet...	Get 5% discount with every ViagraPills ...	██████
	8-8-2010 09:59:38	s.onike_@gbm.com	_Buy And Download Cheap OEMSoftwares! ...	██████
	8-8-2010 07:10:26	samela.su_@strykerco...	CheapGenericViagra+cialisLevitra Order ...	██████
	8-8-2010 07:24:47	ihattie_to...@levy.net	Cialis (generic)... We accept VISA, Mas...	██████

	10-8-2010 09:41:28	jacques.zu...@heig-vd.ch...	Invite to jacques.zuber: come and save ...	██████
	10-8-2010 09:46:17	r-help-bou...@r-project...	Re: [R] How to extract the conf.level o...	██████
	10-8-2010 10:09:23	iyxofafod...@comcastbus...	Invite to jacques.zuber: come and save ...	██████
	10-8-2010 10:09:08	renemcgur...@malaysia.n...	Les traditions et la modernit...	██████
	10-8-2010 10:17:30	r-help-bou...@r-project...	Re: [R] Identifying integers (as oppose...	██████
	10-8-2010 10:30:49	aaburny98...@asianet.co...	Invite to jacques.zuber: come and save ...	██████
	10-8-2010 10:50:26	dbruboet@politico.c...	Wollen Sie ein VIP werden? - Jetzt hier...	██████
	10-8-2010 11:37:44	sonajchina...@madriver.c...	Codeine 30mg x 90 \$396.00 (+4 FreeViagr...	██████
	10-8-2010 12:16:44	blanchel...@pelicanhot...	_Buy And Download Cheap OEMSoftwares! ...	██████
	10-8-2010 13:14:23	rathskele...@acc-inter...	Your Order with Amazon.com	██████
	10-8-2010 13:37:30	tupsrywzat...@kuleuvn.b...	Ne pas partir à Las Vegas avec ces offr...	██████
	10-8-2010 13:48:10	ipunpo142...@ptnet.pl	Upgrade your man-tool	██████
	10-8-2010 13:55:19	lyduhy5982...@windstream...	Be the winner in bed now	██████

heig-vd

Aide à la recherche et de Gestion

Cours de Probabilités et Statistique

JZR

Chapitre 4 : Probabilités conditionnelles et indépendance

19

heig-vd

Aide à la recherche et de Gestion

Cours de Probabilités et Statistique

JZR

Chapitre 4 : Probabilités conditionnelles et indépendance

20

de: Romain Francois <romain.francois@dbmail.com>
sujet: Re: [R] How to extract the conf.level out of t.test() data
pour: Etienne Stockhausen <einh0r2002@web.de>
copié à: r-help@r-project.org <r-help@r-project.org>

Le 09/08/10 20:39, Etienne Stockhausen a écrit :

```
Good afternoon everybody,  
  
I'm writing a little function to visualise hypothesis testing. Therefore  
I need to extract the confidence level of a t-test. Here a little example:  
  
x <- str(t.test(1:10))  
gives  
  
List of 9  
 $ statistic : Named num 5.74  
 ..- attr(*, "names")= chr "t"  
 $ p.value : num 0.000278  
 $ conf.int : atomic [1:2] 3.33 7.67  
 ..- attr(*, "conf.level")= num 0.95  
 $ estimate : Named num 5.5  
 ..- attr(*, "names")= chr "mean of x"  
 $ null.value : Named num 0  
 ..- attr(*, "names")= chr "mean"  
 $ alternative: chr "two.sided"  
 $ method : chr "One Sample t-test"  
 $ data.name : chr "1:10"  
 - attr(*, "class")= chr "htest"  
Now T...>
```

de: Romain Francois <romain.francois@dbmail.com>
sujet: Re: [R] Identifying integers (as opposed to real #) in matrix
pour: David Katz <dkatz@ucdavis.edu>
copié à: r-help@r-project.org <r-help@r-project.org>

Le 10/08/10 10:05, David Katz a écrit :

```
Is there a way to identify (for subsequent replacement) which rows in a  
matrix are comprised entirely of *integers*? I have a large set of  
"nx3" matrices  
where each row either consists of a set of 3 integers or a set of 3 real  
numbers. A given matrix might looks something like this:
```

	[,1]	[,2]	[,3]
[1,]	121.0000	-98.0000	276.0000
[2,]	10.1234	25.4573	-188.9204
[3,]	121.0000	-98.0000	276.0000
[4,]	-214.4982	-99.1043	-312.0495
....			
[n,]	99.0000	1.0000	-222.0000

Ultimately, I'm going to replace the values in the integer-only rows with
"NAs." But first I need r to recognize the integer-only rows. I assume
whatever function I write will be keyed off of the ".0000s", but have no
clue how to write that function. Any ideas?

David Katz

heig-vd

Aide à la recherche et de Gestion

Cours de Probabilités et Statistique

JZR

Chapitre 4 : Probabilités conditionnelles et indépendance

heig-vd

Aide à la recherche et de Gestion

Cours de Probabilités et Statistique

Avant que le courrier puisse être filtré, l'utilisateur a besoin de générer une base de données de mots et signes particuliers comme hypothèque, viagra, sexe, \$, miracle, adresses et domaines IP, ... rassemblés à partir d'exemples de courrier spam (courrier non désiré) et ham (courrier légitime).

Par simplification, notons ces mots ou signes $1, 2, \dots, n$. Pour chacun de ces mots ou signes, on détermine les probabilités suivantes :

p_i : probabilité qu'un mot choisi au hasard dans un message électronique est le mot i en sachant que le message est un spam;

q_i : probabilité qu'un mot choisi au hasard dans un message électronique est le mot i en sachant que le message n'est pas un spam.

Écrivons les probabilités p_i et q_i à l'aide des événements :

M_i : "le mot choisi au hasard dans le message électronique est le mot i ";

S : "le message électronique est un spam".

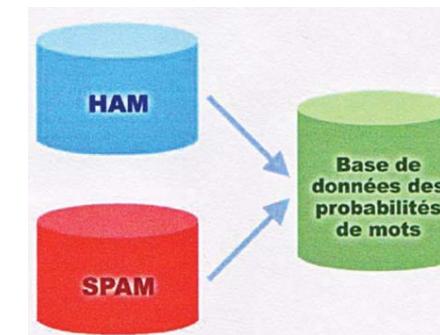
Ainsi,

$$p_i = P(M_i|S) \quad \text{et} \quad q_i = P(M_i|\bar{S}).$$

Probabilités p_i et q_i calculées pour les mots et signes choisis, le filtre est prêt à l'emploi.

À l'arrivée d'un nouveau message électronique, celui-ci est analysé et décomposé en mots et les mots les plus importants sont isolés, en particulier les plus significatifs des messages spam ou non spam. Le filtre calcule alors la probabilité que ce message soit un spam ou non. Si la probabilité dépasse un certain seuil (par exemple 0.9), le message est considéré comme spam.

Les probabilités p_i et q_i des mots i ($i = 1, \dots, n$) sont calculées directement à partir d'exemples de courrier spam et ham propres à la compagnie, à l'établissement scolaire ou à l'organisation.



Pour illustrer le fonctionnement du filtre bayesien, supposons que la proportion de messages spam d'une certaine compagnie vaut 0.9 et que pour le mot "hypothèque" noté 1, $p_1 = 0.05$ et $q_1 = 0.001$.

Pour ce mot, on a alors

$$p_1 = P(M_1|S) = 0.05 \quad \text{et} \quad q_1 = P(M_1|\bar{S}) = 0.001.$$

Un nouveau message électronique vient d'arriver et le mot "hypothèque" y apparaît exactement une fois. En appliquant le théorème de Bayes, la probabilité que le message électronique soit un spam est

$$P(S|M_1) = \frac{P(M_1|S) \cdot P(S)}{P(M_1|S) \cdot P(S) + P(M_1|\bar{S}) \cdot P(\bar{S})} = 0.998.$$

Remarques :

- comme le choix des mots et signes et l'analyse de courrier ham et spam sont propres à l'utilisateur, le filtre bayesien correspondra parfaitement aux besoins de celui-ci;
- la liste des mots clés se diffèrent selon les activités de l'établissement. Par exemple, le mot "hypothèque" est considéré comme suspect pour un établissement scolaire mais pas du tout pour une banque. Ainsi, le filtre bayesien apprend les habitudes de messagerie des établissements;
- la période d'apprentissage qui comprend l'analyse du courrier ham et spam et le calcul des probabilités dure environ quinze jours;
- on peut néanmoins avoir de fausses alertes en particulier au début de la mise en activité du filtre.

4.3 Indépendance

Définition 2 L'événement A est **indépendant** de l'événement B si le fait de savoir que B s'est déroulé n'influence pas la probabilité de A .

En langage probabiliste, cette définition se traduit par

$$P(A | B) = P(A).$$

Or, par définition d'une probabilité conditionnelle,

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

Ainsi, on obtient

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

Remarques (suite) :

Le filtre bayesian

- est une approche intelligente étant donné qu'elle reconnaît non seulement les mots clés qui identifient le spam mais aussi ceux qui dénotent un message valide. Elle va au-delà d'une simple vérification de la présence de mots clés et d'un classement direct d'un message dans la catégorie spam;
- s'adapte et évolue continuellement en apprenant à partir de nouveaux spam et de nouveaux e-mails sortants. Par exemple lorsque les "spammeurs" ont utilisé "f-r-e-e" au lieu de "free", ils sont parvenus à passer à travers le filtre jusqu'à ce que le mot soit intégré dans la base de données;
- fonctionne pour n'importe quelle langue et n'importe quel langage.

Par conséquent, A est indépendant de B si

$$P(A \cap B) = P(A) \cdot P(B).$$

Comme dans le raisonnement, les rôles de A et B sont interchangeables, il en résulte que lorsque A est indépendant de B , B l'est aussi de A . Ainsi, on débouche sur la définition,

Définition 3 Deux événements A et B sont **indépendants** si

$$P(A \cap B) = P(A) \cdot P(B). \quad (3)$$

Deux événements sont **dépendants** s'ils ne sont pas indépendants.

En d'autres termes, deux événements A et B sont **indépendants** si la réalisation de l'un ne modifie pas la probabilité de réalisation de l'autre.



Remarques :

a) l'indépendance de deux événements en probabilités peut être évidente. Par exemple, une pièce de monnaie équilibrée est jetée deux fois de suite. On définit les événements :

- ◊ A : "face apparaît au premier jet";
- ◊ B : "pile apparaît au second jet".

Cependant, l'indépendance n'est pas toujours ressentie de manière intuitive aussi évidente. Par exemple, considérons les événements :

- ◊ C : "le même côté de la pièce sort deux fois de suite";
- ◊ D : "le nombre de face est strictement inférieur à 2".

Sont-ils indépendants ? La réponse n'est pas évidente...

Remarques (suite) :

- b) si les événements A et B sont incompatibles, ils ne sont pas indépendants;
- c) si A et B sont indépendants, A et \overline{B} le sont aussi. Autrement dit, lorsque A est indépendant de B , la probabilité que A survienne n'est influencée ni par l'information que B est réalisé ni par celle que B ne l'est pas;
- d) à noter...

- ◊ A et B incompatibles $\Rightarrow P(A \cup B) = P(A) + P(B);$
- ◊ A et B indépendants $\Rightarrow P(A \cap B) = P(A) \cdot P(B).$

Indépendance (totale) de n événements

Il est évidemment possible de généraliser la notion d'indépendance (totale) à plus de deux événements.

Définition 4 Un ensemble de n événements A_1, A_2, \dots, A_n est dit (*totalelement*) indépendant si pour tout sous-ensemble A_1, A_2, \dots, A_r , $r \leq n$, on a

$$P(A_1 \cap A_2 \cap \dots \cap A_r) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_r).$$

La vérification de l'indépendance de n événements est bien longue... Imaginez le nombre de tests à effectuer pour un ensemble de 5, 10 ou 20 événements.

Application : évaluation de la fiabilité d'un système

- La fiabilité joue un rôle prépondérant dans le milieu industriel pour, par exemple, éviter du retard dans la production ou minimiser les déchets. Elle s'occupe en particulier de l'aptitude de dispositifs techniques (machines, composants électroniques) à accomplir des tâches.
- Pour simplifier, supposons que chaque dispositif se trouve soit en état de fonctionner soit dans l'incapacité de fonctionner i.e hors service.
- Les pannes ou défaillances sont imprévisibles et sont donc soumises au hasard. Il est donc tout à fait naturel de recourir aux probabilités pour traiter des problèmes industriels de fiabilité.

- La fiabilité d'un dispositif peut être définie comme étant la probabilité que le dispositif fonctionne correctement pendant un intervalle de temps donné, c'est-à-dire aucune défaillance se produit pendant ce laps de temps.
- Dans le milieu industriel, les dispositifs (ou composants) sont placés et agencés de différentes manières et forment ainsi des systèmes. Ces dispositifs agissent de manière interactive.
- L'évaluation de la fiabilité d'un système dépend donc de l'état dans lequel se trouve chaque composant et de la disposition de ces derniers (par exemple en série ou en parallèle).

a) système en série



Système en série formé de deux composants A_1 et A_2 .

- Pour simplifier, supposons que les composants fonctionnent indépendamment les uns des autres et considérons seulement les systèmes en série et en parallèle.

- Définissons les événements

F : "le système fonctionne" ;

A_i : "le composant i fonctionne"

avec $i = 1, \dots, n$.

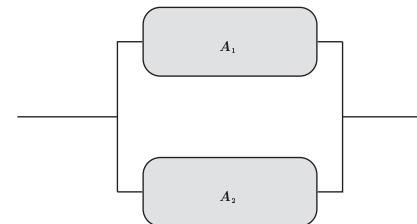
- On cherche à évaluer la fiabilité du système. Plus précisément, on se propose de calculer la probabilité $P(F)$ en fonction des probabilités $P(A_i)$, $i = 1, \dots, n$.

Le système en série fonctionne uniquement si tous les composants fonctionnent. Ainsi,

$$P(F) = P(A_1 \cap A_2 \cap \dots \cap A_n).$$

Par indépendance,

$$P(F) = \prod_{i=1}^n P(A_i).$$



Système en parallèle formé de deux composants A_1 et A_2 .

Le système en parallèle fonctionne si au moins un de ses composants fonctionne. Une défaillance se produit uniquement lorsque tous les composants sont défaillants. Il est donc préférable de raisonner par événement complémentaire. Ainsi,

$$P(\overline{F}) = P(\overline{A_1} \cap \overline{A_2} \cap \dots \cap \overline{A_n}).$$

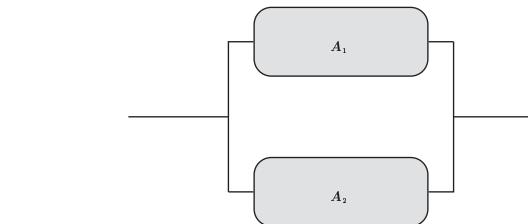
Par indépendance,

$$P(\overline{F}) = \prod_{i=1}^n P(\overline{A_i}) = \prod_{i=1}^n (1 - P(A_i)).$$

Ainsi,

$$P(F) = 1 - \prod_{i=1}^n (1 - P(A_i)).$$

b) système en parallèle



4.4 Tirages et schéma de Bernoulli

Dans de nombreuses applications, l'expérience globale se décompose en une suite d'expériences partielles répétées dans des conditions similaires. Si par exemple, l'expérience de base consiste à répéter le jet d'une pièce de monnaie, on peut considérer chaque jet comme l'une des expériences partielles.

Les exemples les plus représentatifs de ce type d'expériences sont les tirages de boules de couleurs différentes dans une urne.

Souvent les problèmes de probabilité qui en découlent peuvent être résolus à l'aide de diagrammes en arbre.

Sans restriction de la généralité, nous allons nous restreindre à une urne composée de boules de deux couleurs différentes.

Le tirage d'une boule de l'une des deux couleurs peut être interprété comme la réalisation d'un événement; le tirage d'une boule de l'autre couleur correspond à la réalisation de l'événement complémentaire.

Deux types de tirages sont à considérer :

- { les épreuves totalement indépendantes : **tirages avec remise**;
- les expériences non distinctes : **tirages sans remise**.

a) tirages avec remise, schéma de Bernoulli

Les épreuves partielles sont

- identiques (elles ont toutes le même (sous-)ensemble fondamental et sont toutes affectées de la même fonction de probabilité);
- totalement indépendantes.

↗ répétition d'épreuves identiques totalement indépendantes.

Exemple :

une urne contient des boules noires et des boules blanches. On tire deux boules **avec remise**. Si p est la probabilité de sortir une boule noire de l'urne, déterminer les issues de l'expérience et leurs probabilités.

a) tirages avec remise, schéma de Bernoulli (suite)

Méthode : utiliser un schéma en arbre.

Les probabilités conditionnelles se trouvent sur les branches de l'arbre. Elles se multiplient pour donner les probabilités des événements élémentaires représentés par les feuilles.

Justification : $P(E \cap F) = P(E | F) \cdot P(F)$.

première épreuve deuxième épreuve résultat

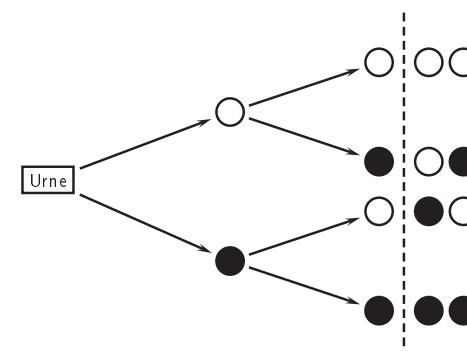


Schéma en arbre du tirage avec remise de deux boules.

b) tirages sans remise

Les expériences partielles ne sont pas identiques.

~~~ répétition d'expériences non distinctes.

Exemple :

une urne contient  $n$  boules,  $k$  de couleur noire et  $n - k$  de couleur blanche. On tire deux boules **sans remise**. Déterminer les issues de l'expérience et calculer leurs probabilités.

Méthode : utiliser un schéma en arbre.

Justification :  $P(E \cap F) = P(E | F) \cdot P(F)$ .

première expérience    deuxième expérience    résultat

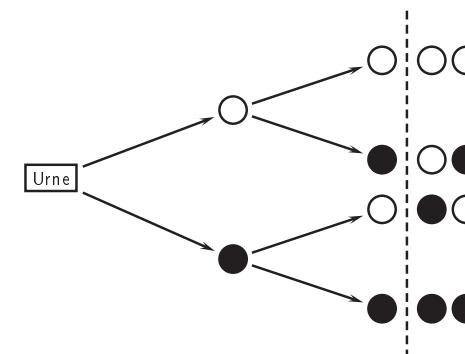


Schéma en arbre du tirage sans remise de deux boules.

Remarque :

dans de nombreux problèmes de probabilité, le dénombrement direct du nombre de cas favorables à un événement n'est pas toujours immédiat. Souvent la division de l'expérience globale en expériences partielles est très pratique et évite des calculs fastidieux.

*“Il est remarquable que la théorie des probabilités, qui a pris son origine dans l'étude des jeux de chance, soit devenue l'objet le plus important de la connaissance humaine. Les questions les plus importantes de la vie ne sont en réalité, pour l'essentiel, que des problèmes de probabilité.”*

P.-S. de Laplace (1749–1827)

## EXERCICES : PROBABILITÉS CONDITIONNELLES ET INDÉPENDANCE

### Exercice 1

Considérons deux événements  $A$  et  $B$  pouvant se réaliser dans une expérience aléatoire.

- Si  $P(\bar{A}) = 0.7$ ,  $P(B) = 0.5$  et  $P(A \cup B) = 0.6$ , calculer la probabilité  $P(A | B)$ .
- Si  $P(A) = 0.3$ ,  $P(\bar{B}) = 0.4$  et  $P(B | A) = 0.4$ , déterminer la probabilité  $P(A \cup B)$ .

**Solutions :** a) 0.4 b) 0.78.

### Exercice 2

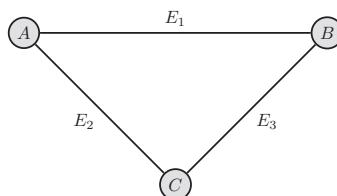
Gaston a acheté des pêches au supermarché : deux fois plus de jaunes que de blanches. Quand il prend une pêche au hasard, il y a une chance sur deux pour qu'elle soit avariée, une chance sur quatre qu'elle soit à la fois blanche et avariée. Il choisit une pêche au hasard.

- En sachant que la pêche tirée par Gaston est blanche, déterminer la probabilité que la pêche soit avariée.
- La pêche choisie par Gaston est avariée. Calculer la probabilité qu'elle soit blanche.

**Solutions :** a) 3/4 b) 1/2.

### Exercice 3

Un voyageur de commerce doit se rendre de la ville  $A$  au petit village de montagne  $B$ . Comme l'indique la figure ci-dessous, deux possibilités s'offrent à lui : une route qui relie directement  $A$  à  $B$  et un détour par le village  $C$ . Suivant les conditions météorologiques, les routes ne sont pas toujours praticables pendant les longs mois d'hiver.



Désignons par  $E_1$ ,  $E_2$  et  $E_3$  les événements

- $E_1$  : "la route entre  $A$  et  $B$  est ouverte";
- $E_2$  : "la route entre  $A$  et  $C$  est ouverte";
- $E_3$  : "la route entre  $C$  et  $B$  est ouverte".

Pour un certain jour d'hiver, considérons les probabilités :

$$P(E_1) = 2/5 \quad P(E_2) = 3/4 \quad P(E_3) = 2/3$$

$$P(E_3 | E_2) = 4/5 \quad P(E_1 | E_2 \cap E_3) = 1/2.$$

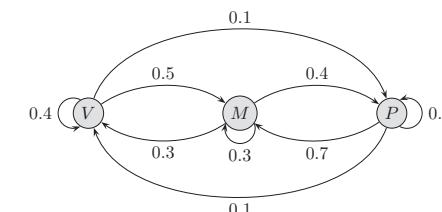
- En supposant que le voyageur de commerce doit passer par le village  $C$ , déterminer la probabilité qu'il puisse se rendre de  $A$  à  $B$ .

- Calculer la probabilité qu'il puisse se rendre de  $A$  à  $B$  par l'un ou l'autre des deux itinéraires possibles.

**Solutions :** a) 0.6 b) 0.7.

### Exercice 4

L'approvisionnement en eau d'une ville australienne est assurée par un réservoir. Pour simplifier, on admet que le réservoir peut se trouver seulement dans trois états : vide ( $V$ ), à moitié plein ( $M$ ) et plein ( $P$ ). La quantité d'eau récoltée dans le réservoir ainsi que la demande en eau de la ville ne sont pas totalement prévisibles et varient pendant la saison. Les probabilités de transition du réservoir d'un état à un autre figurent dans le graphe suivant :



Pour illustrer la lecture du graphe, supposons que le réservoir soit vide au début de la saison. La probabilité qu'on le trouve à nouveau vide en fin de saison est 0.4 et celle qu'il soit à moitié plein vaut 0.5.

On suppose que le réservoir soit plein au début d'une saison donnée.

- Déterminer la probabilité que le réservoir contienne encore de l'eau à la fin de la saison.
- On admet que les probabilités de transition demeurent identiques la saison suivante. Sous cette hypothèse, calculer la probabilité que le réservoir soit plein à la fin de la deuxième saison.

**Solutions :** a) 0.9 b) 0.33.

### Exercice 5

Dans une école d'ingénierie trois parkings notés  $A$ ,  $B$  et  $C$  sont disponibles pour garer son automobile. Pour trouver une place de parc, un professeur procède toujours de la même manière : il prospecte dans l'ordre le parking  $A$ , le parking  $B$  et enfin le  $C$  et parque sa voiture sitôt une place libre trouvée. Les parkings  $A$  et  $B$  sont gratuits alors que le parking  $C$  est payant. On suppose qu'aucune autre possibilité de parquer soit possible. Une étude statistique a montré qu'un matin donné les probabilités de trouver une place disponible dans les parkings  $A$ ,  $B$  et  $C$  valent respectivement 0.2, 0.1 et 0.5. De plus, si aucune place de parc est disponible dans le parking  $A$ , la probabilité que le professeur trouve à se parquer dans le parking  $B$  vaut 0.04 et si les parkings  $B$  et  $C$  sont tous deux complets, la probabilité de tomber sur une place disponible dans le parking  $C$  est 0.4.

- Calculer la probabilité que le professeur ne trouve aucune place de parc non payante un matin donné.

- b) Déterminer la probabilité qu'un matin donné le professeur trouve une place libre dans l'un des trois parkings pour y laisser son automobile.

**Solutions :** a) 0.768 b) 0.5392.

#### Exercice 6

Quand on téléphone à Labévue entre 20h00 et 21h00, on a neuf chances sur dix de tomber sur son répondeur. Il l'utilise lorsqu'il est à la maison deux fois sur trois pour ne pas être dérangé par un importun et le branche systématiquement lorsqu'il est absent. Vous appellerez Labévue par téléphone un jour donné entre 20h00 et 21h00. Calculer la probabilité qu'il sera là.

*Indication :* considérer les événements  $A$  : "on tombe sur le répondeur de Labévue" et  $B$  : "Labévue est à la maison".

**Solution :** 3/10.

#### Exercice 7

Une compagnie d'assurances répartit les conducteurs en deux classes : personnes à risque modéré et à haut risque. Ses statistiques montrent que les pourcentages de ses assurés impliqués dans un accident sur une période d'une année sont respectivement suivant les classes de 20% et de 40%. On suppose que 30% de la population appartient à la classe à haut risque.

- Calculer la probabilité qu'un nouvel assuré soit victime d'un accident pendant l'année qui suit la signature de son contrat.
- Un nouveau signataire a un accident pendant l'année qui suit la signature du contrat. Déterminer la probabilité qu'il fasse partie de la classe à haut risque.

**Solutions :** a) 0.26 b) 0.46.

#### Exercice 8

Une petite fabrique possède trois machines  $A$ ,  $B$  et  $C$  qui assurent respectivement 45%, 35% et 20% de la production totale. Par machine, les déchets sont respectivement de 10%, 10% et 15% du total des pièces produites.

- Calculer la proportion défectueuse de la production totale.
- Calculer la probabilité qu'une pièce provienne de la machine  $B$  sachant qu'elle est défectueuse.

**Solutions :** a) 0.11 b) 0.32.

#### Exercice 9

Une petite société de métallurgie est divisée en deux secteurs : l'administration et l'exploitation. Un jour donné, les trois serveurs informatiques de la société désignés par  $A$ ,  $B$  et  $C$  ont traité respectivement 50, 75 et 100 requêtes. Par serveur, les pourcentages des requêtes provenant de l'administration valent respectivement 50%, 60% et 70%. Ce jour-là, une requête de l'administration est choisie au hasard. Déterminer la probabilité qu'elle ait été traitée par le serveur  $C$ . Quelle hypothèse faites-vous pour calculer cette probabilité ?

**Solution :** 0.5. Hypothèse d'équiprobabilité; chaque requête a même probabilité d'être choisie.

#### Exercice 10

Deux séquences de signaux binaires 100010 et 101011 ont été émises respectivement par les émetteurs  $X$  et  $Y$ . La transmission des signaux est soumise à aucune perturbation. Les statistiques du récepteur  $Z$  montrent que le 40% des séquences reçues proviennent de l'émetteur  $X$  et le reste de l'émetteur  $Y$ . Une des deux séquences vient d'être enregistrée en  $Z$ . L'un des six signaux formant cette séquence est choisi au hasard. En sachant qu'il s'agit d'un 1, déterminer la probabilité que la séquence reçue soit celle émise par  $X$ .

*Indication :* pour calculer la probabilité demandée, définissons les événements suivants :

$A$  : "la séquence reçue est émise par l'émetteur  $X$ " et  $C$  : "le signal reçu est un 1".

**Solution :** 1/4.

#### Exercice 11

Laboratoire d'analyses médicales

Paris, le 12 mai 2004

Monsieur,

Vous êtes récemment venu dans notre hôpital pour un test de dépistage d'une maladie rare, qui touche en France une personne sur dix mille. Nous sommes au regret de vous annoncer que ce test, efficace à 99 %, s'est révélé positif.

Vous recevez cette lettre de votre hôpital. Quelle est, selon vous, la probabilité que vous soyiez réellement malade ?

**Solution :** face à une lettre aussi alarmante, la plupart des gens pensent qu'ils ont 99% de risques d'être malade. Pourtant, cette probabilité est ici inférieure à ... 1% !

tiré de "Science & Vie", septembre 2004.

#### Exercice 12

On dispose de deux pièces de monnaie :

- la première pièce est équilibrée; la probabilité d'obtenir un "pile" est 1/2;
- la deuxième pièce est truquée; la probabilité d'obtenir un "pile" est 3/4.

On prend une pièce au hasard, on la jette trois fois de suite et on obtient trois fois "pile". Quelle est la probabilité qu'on ait lancé la pièce truquée ?

**Solution :** 27/35.

#### Exercice 13

On considère deux urnes : la première, désignée par  $A$ , contient 5 boules blanches et 7 noires; la seconde, notée  $B$ , est formée de 3 boules blanches et 12 noires. Une pièce de monnaie équilibrée est jetée. Si pile sort, on tire une boule de l'urne  $A$ . On prend une boule dans l'urne  $B$  si la pièce montre face.

- a) Calculer la probabilité de tirer une boule blanche.  
 b) En sachant qu'une boule blanche a été tirée, déterminer la probabilité que le jet de la pièce avant le tirage de la boule ait donné face.

**Solutions :** a) 37/120   b) 12/37.

#### Exercice 14

La pollution de l'air dans une certaine ville est causée principalement par l'industrie et les automobiles. Selon l'office de l'environnement de la ville, il est possible dans les cinq prochaines années de maintenir les excès de ces deux sources de pollution dans des normes acceptables avec probabilités respectives 0.75 et 0.6. On suppose que si la pollution causée par l'une des deux sources est admissible mais pas celle occasionnée par l'autre, la probabilité que le niveau global de pollution dans cette ville reste acceptable vaut 0.8.

Considérons les événements :

- $I$  : "la pollution causée par l'industrie est acceptable";
- $A$  : "la pollution générée par les automobiles est admissible".

On admet que les impacts de l'industrie et des automobiles sur la pollution de la ville sont indépendants.

- a) Calculer les probabilités  $P(A \cap I)$ ,  $P(A \cap \bar{I})$ ,  $P(\bar{A} \cap I)$  et  $P(\bar{A} \cap \bar{I})$ .  
 b) Déterminer la probabilité que le niveau global de pollution dans cette ville soit acceptable dans cinq ans.  
 c) En sachant que dans cinq ans le niveau global de pollution dans cette ville ne soit pas acceptable, calculer la probabilité que cet excès soit uniquement dû aux automobiles.

**Solutions :** a) 0.45, 0.15, 0.30, 0.10   b) 0.81   c) 0.32.

#### Exercice 15

Depuis quelques semaines, M. de Mesmaeker utilise la même stratégie pour gagner à la roulette. Il ne mise que sur la couleur *rouge* et seulement si les dix numéros sortis précédemment étaient de couleur *noire*. En se basant sur la rareté des séquences de 11 numéros noirs, il est persuadé que ses chances de gagner sont très grandes. Que pensez-vous de cette stratégie ?

**Solution :** la stratégie est mauvaise en raison de l'indépendance.

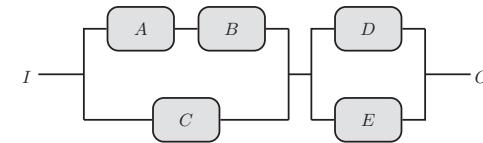
#### Exercice 16

Un système électronique est formé de trois composants  $A$ ,  $B$  et  $C$ . Il fonctionne si  $A$  fonctionne et si  $B$  ou  $C$  fonctionnent. En supposant que les composants sont indépendants les uns des autres et qu'ils fonctionnent avec des probabilités respectives de 0.9, 0.95 et 0.95, déterminer la probabilité que le système tombe en panne.

**Solution :** 0.1.

#### Exercice 17

Un système électronique est formé des composants  $A$ ,  $B$ ,  $C$ ,  $D$  et  $E$  placés selon le dispositif ci-dessous.



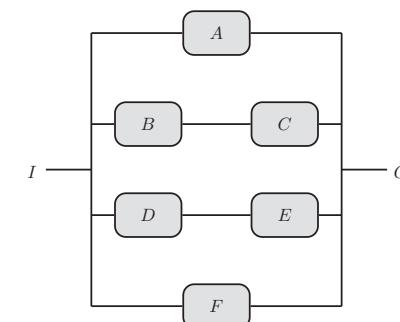
Le système est opérationnel si les composants  $A$  et  $B$  fonctionnent simultanément ou si le composant  $C$  fonctionne et si au moins un des composants  $D$  ou  $E$  fonctionne; autrement dit, si un chemin est possible entre les points  $I$  et  $O$ . On suppose que les composants fonctionnent indépendamment les uns des autres. Les probabilités que les composants  $A$ ,  $B$  et  $C$  fonctionnent sont égales et valent toutes 0.9, pour les composants  $D$  et  $E$ , elles sont les deux égales à 0.95.

Calculer la probabilité que le système soit opérationnel.

**Solution :** 0.9785.

#### Exercice 18

Un système électronique est formé des composants  $A$ ,  $B$ ,  $C$ ,  $D$ ,  $E$  et  $F$  placés selon le dispositif ci-dessous.



Le système est opérationnel si le composant  $A$  fonctionne ou si les composants  $B$  et  $C$  fonctionnent simultanément ou si les composants  $D$  et  $E$  fonctionnent simultanément ou si le composant  $F$  fonctionne; autrement dit, si un chemin est possible entre les points  $I$  et  $O$ . On suppose que les composants fonctionnent indépendamment les uns des autres. Les probabilités que les composants  $B$ ,  $C$ ,  $D$  et  $E$  fonctionnent sont égales et valent toutes 0.9, pour les composants  $A$  et  $F$ , elles valent 0.95.

Calculer la probabilité que le système ne fonctionne pas.

**Solution :**  $9.025 \cdot 10^{-5}$ .



## Chapitre 5

### Variables aléatoires

### Contenu

5.1 Généralités sur les variables aléatoires

5.2 Variables aléatoires discrètes

5.3 Variables aléatoires continues

5.4 Propriétés de l'espérance mathématique et de la variance

### 5.1 Généralités sur les variables aléatoires

- Nous introduirons dans ce chapitre les variables aléatoires et les distributions de probabilités qui jouent un rôle important non seulement dans le milieu industriel mais aussi dans la vie de tous les jours.
- D'une certaine manière, nous avons déjà abordé la notion de distribution dans l'analyse exploratoire des données mais sous un angle pratique.
- Pour ainsi dire, le calcul des probabilités peut être vu comme étant l'aspect théorique des notions pratiques introduites dans l'analyse exploratoire des données.

- Plus précisément, la **loi forte des grands nombres** permet d'établir un lien entre les notions probabilistes (aspect théorique) et les notions statistiques (aspect pratique). Cette loi nous indique que la fréquence relative d'un événement tend vers sa probabilité lorsque l'expérience est répétée indéfiniment.
- La loi forte des grands nombres est sans doute le résultat le plus célèbre de la théorie des probabilités.
- Les notions probabilistes sont en fait associées à une population hypothétique alors que les notions statistiques sont de leur côté associées à un nombre restreint d'observations (échantillon).

Il arrive fréquemment que lors d'une expérience aléatoire, on s'intéresse davantage à une **fonction** du résultat plutôt qu'au résultat en lui-même. Illustrons cette idée à l'aide des exemples suivants :

1. jet de deux dés

on s'intéresse à la somme obtenue, 7 par exemple, plutôt qu'au fait de savoir si c'est le couple (1, 6) qui est apparu ou (2, 5), (3, 4), (4, 3), (5, 2) ou (6, 1);

2. jet d'une pièce de monnaie

on s'intéresse au nombre de fois où pile est apparu plutôt qu'à la séquence détaillée des piles ou faces.

Ce concept de fonctions réelles définies sur l'ensemble fondamental nous conduit à la notion de variables aléatoires.

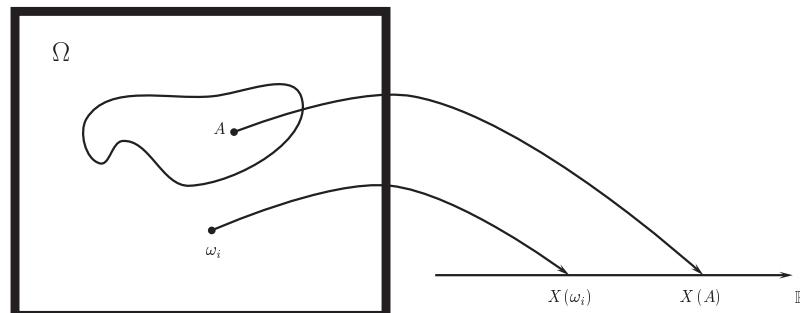


Illustration d'une variable aléatoire réelle.

**Définition 1** Soit un espace probabilisé d'ensemble fondamental  $\Omega$  et de mesure de probabilité  $P$ . Une **variable aléatoire** réelle définie sur  $\Omega$  est une application de  $\Omega$  dans  $\mathbb{R}$  (ou dans une partie de  $\mathbb{R}$ ) :

$$\begin{aligned} X &: \Omega \rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega). \end{aligned}$$

- $\Omega$  : ensemble fondamental d'une expérience, ensemble des issues;
- $\mathcal{H}$  : ensemble des valeurs prises par la variable aléatoire  $X$ .

$\mathcal{H}$  peut être :

- ◊ discret  $\rightsquigarrow$  variable aléatoire discrète;
- ◊ continu  $\rightsquigarrow$  variable aléatoire continue.

Exemples :

1. nombre de pièces non conformes dans un échantillon de taille 10 prélevé au hasard dans la production  $\mathcal{H} = \{0, 1, \dots, 10\}$ ;
2. nombre d'appels téléphoniques parvenant à un standard durant un intervalle de temps fixé  $\mathcal{H} = \mathbb{N}$ ;
3. proportion de réponses "oui" à une question posée dans un sondage  $\mathcal{H} = \{x \in \mathbb{Q}, 0 \leq x \leq 1\}$ ;
4. nombre de piles pour  $n$  parties de pile ou face  $\mathcal{H} = \{0, 1, \dots, n\}$ ;  
cas particulier :  $n = 3$  en utilisant une pièce équilibrée  
– ensemble fondamental  $\Omega$  :

$$\Omega = \{(P, P, P), (P, P, F), (P, F, P), (P, F, F), (F, P, P), (F, P, F), (F, F, P), (F, F, F)\};$$

#### 4. nombre de piles pour $n$ parties de pile ou face (*suite*)

- définition de la variable aléatoire  $X$  :

$$X : \Omega \longrightarrow \mathbb{R}$$

$\omega \longmapsto X(\omega) = \text{nombre de piles de l'événement } \omega;$

- ensemble des valeurs prises par  $X$  :

$$\mathcal{H} = \{0, 1, 2, 3\};$$

- loi de probabilité de  $X$  (probabilités des valeurs prises par  $X$ ) :

- ▷  $P(X = 0) = P(\{(F, F, F)\}) = 1/8;$
- ▷  $P(X = 1) = P(\{(P, F, F), (F, P, F), (F, F, P)\}) = 3/8;$
- ▷  $P(X = 2) = P(\{(P, P, F), (P, F, P), (F, P, P)\}) = 3/8;$
- ▷  $P(X = 3) = P(\{(P, P, P)\}) = 1/8;$

#### 4. nombre de piles pour $n$ parties de pile ou face (*suite*)

la loi de probabilité de  $X$  est souvent donnée sous la forme d'un tableau :

|            |     |     |     |     |                     |
|------------|-----|-----|-----|-----|---------------------|
| $X = x$    | 0   | 1   | 2   | 3   |                     |
| $P(X = x)$ | 1/8 | 3/8 | 3/8 | 1/8 | $\sum P(X = x) = 1$ |
|            |     |     |     |     |                     |

#### 5. variable aléatoire de Bernoulli

$$\mathcal{H} = \{0, 1\};$$

on réalise une expérience dont le résultat sera interprété soit comme un succès soit comme un échec. On définit alors une variable aléatoire  $X$  en lui attribuant la valeur 1 lors d'un succès et 0 lors d'un échec.

#### 5. variable aléatoire de Bernoulli (*suite*)

- définition de la variable aléatoire  $X$  :

$$X : \omega \longmapsto X(\omega) : \text{succès ou échec};$$

- ensemble des valeurs prises par  $X$  :

$$\mathcal{H} = \{0, 1\};$$

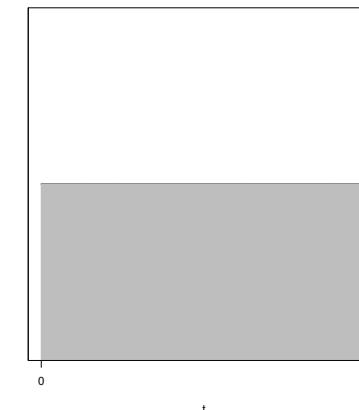
- loi de probabilité de  $X$  :

|            |         |     |                     |
|------------|---------|-----|---------------------|
| $X = x$    | 0       | 1   |                     |
| $P(X = x)$ | 1 - $p$ | $p$ |                     |
|            |         |     | $\sum P(X = x) = 1$ |

où  $p$  est la probabilité d'un succès.

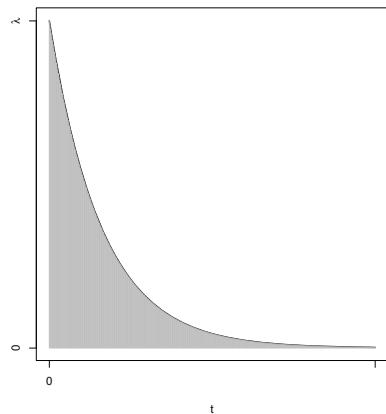
#### 6. temps d'attente d'un bus

$$\mathcal{H} = [0, T];$$



7. durée de vie en heures d'un composant électronique

$$\mathcal{H} = [0, \infty[;$$



## 5.2 Variables aléatoires discrètes

**Définition 2** On appelle **variable aléatoire discrète** une variable aléatoire qui ne prend qu'un nombre fini ou dénombrable de valeurs. Pour une variable aléatoire discrète  $X$ , on définit sa **loi de probabilité** par  $p(x) = P(X = x)$ .

Si  $X$  peut prendre les valeurs  $x_1, x_2, \dots$ , alors

- i)  $p(x_i) = P(X = x_i) \geq 0$ ,  $i = 1, 2, \dots$ ;
- ii)  $p(x) = 0$  pour toutes les autres valeurs de  $x$ ;
- iii)  $\sum p(x_i) = 1$ .

Remarques :

- a) comme le montrent ces exemples, l'ensemble  $\mathcal{H}$  des valeurs prises par une variable aléatoire peut être **fini** (exemples 1, 3, 4 et 5), **infini dénombrable** (exemple 2) ou **infini non dénombrable** (exemples 6 et 7);
- b) les réalisations d'une variable aléatoire discrète sont isolées; l'ensemble  $\mathcal{H}$  est fini ou infini dénombrable;
- c) l'ensemble des valeurs possibles pour une variable aléatoire continue est infini non dénombrable. Une variable aléatoire continue est définie sur un intervalle réel.

Exemples :

1. nombre de piles pour  $n$  parties de pile ou face;
2. variable aléatoire de Bernoulli;
3. variable aléatoire  $X$  uniforme discrète

- $\mathcal{H} = \{1, 2, \dots, n\}$ ;
- loi de probabilité de  $X$  :

$$p(x) = \begin{cases} P(X = x) = 1/n & \text{si } x = 1, \dots, n, \\ 0 & \text{sinon.} \end{cases}$$

Application :

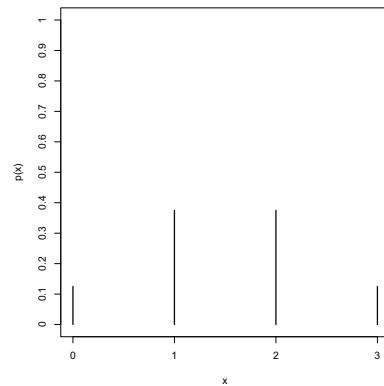
jet d'un dé équilibré ( $n = 6$ ).

Exemples (*suite*) :

4. nombre d'articles vendus par jour dans un magasin;
5. nombre de messages électroniques reçus par un étudiant de la HEIG-VD dans un intervalle de temps de 20 minutes;
6. ...

Remarque :

une loi de probabilité est donnée soit par la liste des probabilités (comme par exemple pour la variable aléatoire de Bernoulli) soit par une formule générale (comme par exemple pour la variable uniforme discrète).



Histogramme de la loi de probabilité de la variable aléatoire discrète introduite dans l'exemple 4 en page 8.

### 5.2.1 Histogramme, fonction de répartition

Pour visualiser la loi de probabilité d'une variable aléatoire discrète  $X$ , on construit fréquemment un **histogramme**. Les valeurs  $x_i$  prises par  $X$  sont placées en abscisse; à chacune d'elles est associé un "bâton", une "aiguille" de hauteur  $P(X = x_i)$ .

Inconvénient :

l'**histogramme** est une représentation graphique trop ponctuelle (elle est trop concentrée sur les valeurs prises par la variable aléatoire). On préfère associer à la variable une fonction définie sur tout  $x$  réel

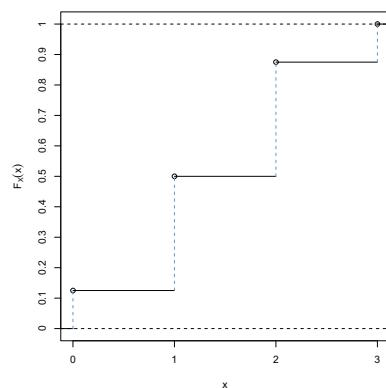
~→ **fonction de répartition**.

**Définition 3** La **fonction de répartition**  $F_X$  d'une variable aléatoire discrète  $X$  définie sur  $\Omega$  est donnée par

$$F_X(x) = P(X \leq x). \quad (1)$$

Propriétés

- i)  $F_X$  prend ses valeurs dans  $[0, 1]$ ;
- ii)  $F_X$  est monotone non décroissante;
- iii)  $F_X(x) = \sum_{x_i \leq x} P(X = x_i)$ .



Graph de la fonction de répartition de la variable aléatoire discrète introduite dans l'exemple 4 en page 8.

Remarques :

- a) pour une variable aléatoire discrète  $X$ ,  $F_X$  est une fonction en "escalier" (elle est continue à droite en  $x = x_i$ );
- b) pour chaque valeur  $x_i$  prise par  $X$ ,  $F_X$  fait un bond de hauteur  $P(X = x_i)$ ;
- c)  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  et  $\lim_{x \rightarrow +\infty} F_X(x) = 1$ ;
- d) la fonction de répartition indique de quelle manière évoluent les probabilités lorsqu'on parcourt de gauche à droite l'axe des réalisations de la variable aléatoire.

## 5.2.2 Espérance mathématique

**Définition 4** Soient  $X$  une variable aléatoire discrète sur  $\Omega$ ,  $\mathcal{H} = \{x_i\}_{i=1}^n$  l'ensemble des valeurs prises par  $X$  et  $P(X = x_i) = p_i$  la loi de probabilité de  $X$ . L'**espérance mathématique** de  $X$ , notée  $E(X)$  ou  $\mu_X$ , est le nombre réel défini par

$$E(X) = \mu_X = \sum_{i=1}^n x_i \cdot P(X = x_i) = \sum_{i=1}^n x_i \cdot p_i. \quad (2)$$

L'espérance mathématique peut être considérée comme étant la notion correspondante dans les probabilités de la moyenne arithmétique introduite en analyse exploratoire des données.

En effet, la moyenne arithmétique est un indicateur de tendance centrale de la distribution des valeurs observées. L'espérance mathématique est un paramètre de tendance centrale pour la distribution de probabilités.

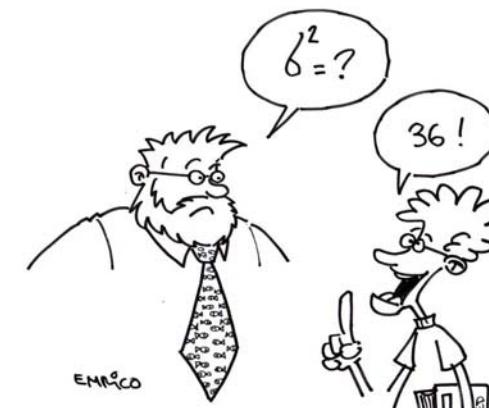
Les termes de **moyenne** et **espérance mathématique** ont souvent été utilisés comme synonymes. Pour éviter une quelconque confusion, il est préférable de réservé l'appellation moyenne à la variable aléatoire moyenne de plusieurs variables aléatoires et à sa réalisation. L'espérance mathématique est en fait un nombre réel qui n'a rien d'aléatoire.

### 5.2.3 Variance, écart-type

L'espérance mathématique est un paramètre qui caractérise la **tendance centrale** ("le milieu") d'une distribution (ou loi) de probabilités. Le paramètre qui caractérise la **dispersion** de la distribution est la **variance**, i.e l'espérance mathématique du carré de l'écart  $X - E(X)$ .

**Définition 5** Soient  $X$  une variable aléatoire discrète sur  $\Omega$ ,  $\mathcal{H} = \{x_i\}_{i=1}^n$  l'ensemble des valeurs prises par  $X$  et  $P(X = x_i) = p_i$  la loi de probabilité de  $X$ . La **variance** de  $X$ , notée  $\text{Var}(X)$ ,  $\sigma^2(X)$  ou  $\sigma_x^2$ , est le nombre réel positif défini par

$$\text{Var}(X) = \sigma^2(X) = \sigma_x^2 = \sum_{i=1}^n (x_i - E(X))^2 \cdot p_i. \quad (3)$$



#### Remarques :

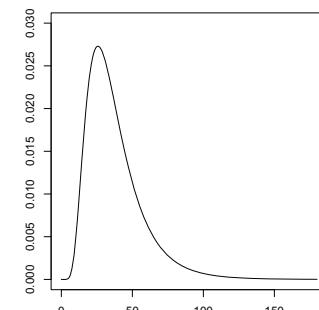
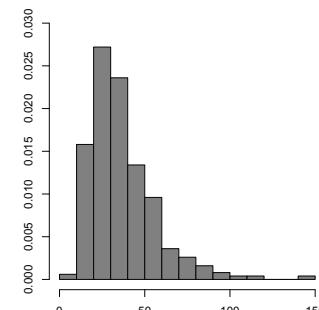
- a) l'unité de la variance n'est pas identique à celle de la variable aléatoire !  
→ **écart-type**  $\sigma(X)$  ou  $\sigma_x$  défini par
- math display="block">\sigma(X) = \sqrt{\text{Var}(X)}. \quad (4)
- L'écart-type est donc un paramètre de dispersion dans les mêmes unités que la variable;
- b) comme l'espérance mathématique, la variance et l'écart-type sont des nombres réels qui n'ont rien d'aléatoire;
- c) la variance et l'écart-type sont toujours des nombres réels **positifs** ! Un premier diagnostic sur l'exactitude de vos calculs !

#### Remarques (suite) :

- d) nous avons rencontré les notions correspondantes de la variance et de l'écart-type en analyse exploratoire des données. Elles jouent un rôle tout à fait identique mais portent malheureusement les mêmes noms.  
Soyons vigilants à ne pas les confondre !
- e) on peut montrer que la variance d'une variable aléatoire discrète  $X$  s'écrit sous sa forme simplifiée
- math display="block">\text{Var}(X) = E(X^2) - E(X)^2 \quad (5)
- où  $E(X^2) = \sum_{i=1}^n x_i^2 \cdot p_i$ ;
- f) les définitions de l'espérance mathématique et de la variance pour une variable aléatoire discrète  $X$  peuvent être généralisées à un nombre infini de réalisations  $x_1, x_2, \dots$

### 5.3 Variables aléatoires continues

- En analyse exploratoire des données, nous avons visualisé la distribution des valeurs observées d'une variable statistique continue à l'aide d'un histogramme qui pouvait présenter une forme plus ou moins symétrique comme l'indique la figure de la page suivante.
- Si on augmente indéfiniment le nombre d'observations en diminuant la largeur des intervalles de classe jusqu'à ce qu'elle soit indéfiniment petite, les boîtes formant l'histogramme vont se multiplier, devenir de plus en plus étroites et, à la limite, vont se fondre en une surface délimitée par le graphe d'une fonction continue, appelée **fonction de densité**, et l'axe des abscisses.
- Ainsi, on tombe sur des variables aléatoires définies sur des intervalles finis ou infinis.



Histogramme d'une variable statistique continue (gauche)

et modèle probabiliste (droite).

#### Exemples :

1. temps d'attente d'un bus;
2. choix d'un point sur un segment;
3. taille du corps humain;
4. longueur d'une pièce d'aluminium;
5. durée de vie d'un transistor;
6. durée d'une conversation téléphonique;
7. usure de pièces;
8. ...

**Définition 6** On appelle **variable aléatoire continue** une variable aléatoire qui peut prendre toutes les valeurs d'un intervalle (un intervalle borné, une demi-droite ou  $\mathbb{R}$  tout entier).

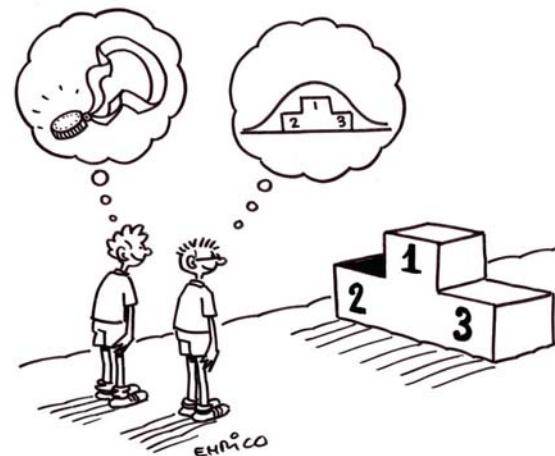
La mesure de probabilité est alors "étalée" sur l'intervalle.

↪ conséquence :

**tous les points ont une probabilité nulle !!!**

Explication intuitive :

il y a trop de points dans un intervalle pour que chacun d'eux puisse avoir une probabilité non nulle, aussi petite soit-elle.



- Comme les probabilités ponctuelles sont nulles, la notion de base pour une variable aléatoire continue est la probabilité d'intervalle

$$P(a < X \leq b).$$

- L'utilisation d'une probabilité d'intervalle se justifie en pratique. En effet, on est par exemple davantage intéressé par la probabilité que le diamètre d'une vis soit compris dans l'intervalle de tolérance imposé par le client plutôt que de connaître la probabilité que le diamètre mesure exactement la valeur standard souhaitée par le client, valeur qui ne sera d'ailleurs pas atteinte à coup sûr.
- Il nous reste à définir la notion de variable aléatoire continue de manière rigoureuse.

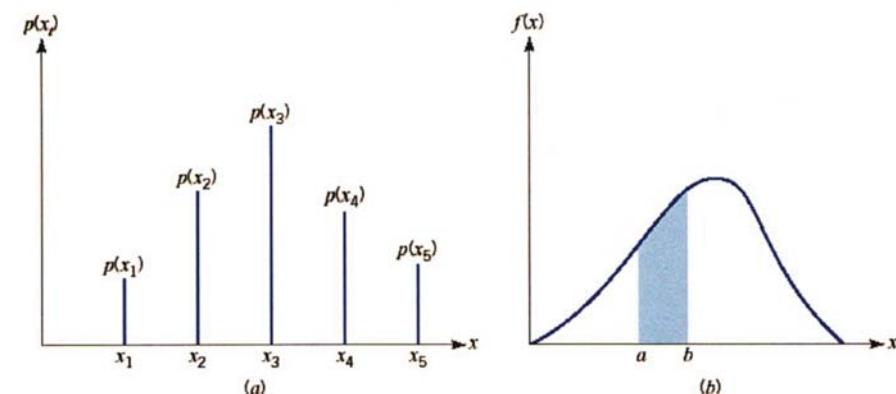
**Définition 7**  $X$  est une **variable aléatoire continue** s'il existe une fonction  $f_X$  non négative telle que

$$P(X \in A) = \int_A f_X(u) du, \quad (6)$$

où  $A$  est un ensemble de nombres réels (souvent un intervalle).

La fonction  $f_X$  est appelée la **fonction de densité de probabilité** de  $X$ . Si aucune ambiguïté n'est à craindre,  $f_X$  est souvent notée  $f$ .

La fonction de densité nous indique de quelle manière la masse de probabilité est distribuée sur un intervalle fini ou infini. Pour cette raison, on peut la considérer comme étant la notion correspondante dans le cas continu de la loi de probabilité.



Histogramme d'une loi de probabilité discrète (gauche) et fonction de densité (droite).

### Propriétés de la fonction de densité $f_X$

i)  $P(X \in ]-\infty, \infty[) = \int_{\mathcal{H}} f_X(u) du = 1;$

ii)  $P(a < X \leq b) = \int_a^b f_X(u) du;$

iii)  $P(X = a) = 0 !!!$

iv)  $f_X(u) \geq 0, \forall u \in \mathbb{R}.$

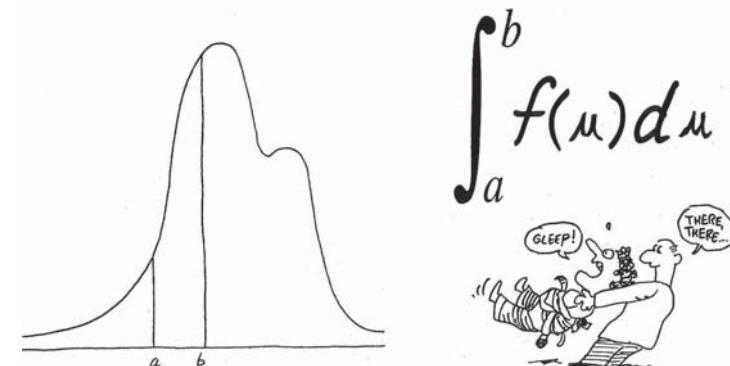
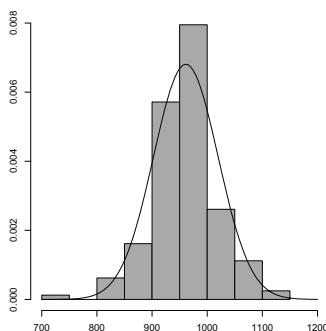


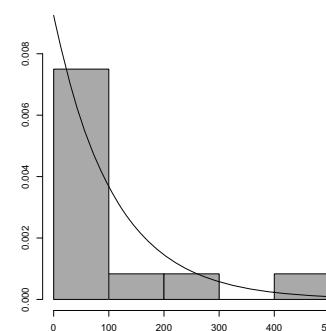
Illustration de la propriété ii) d'une fonction de densité.

▷ Exemple 1 : rendement d'une machine dans un procédé d'usinage.



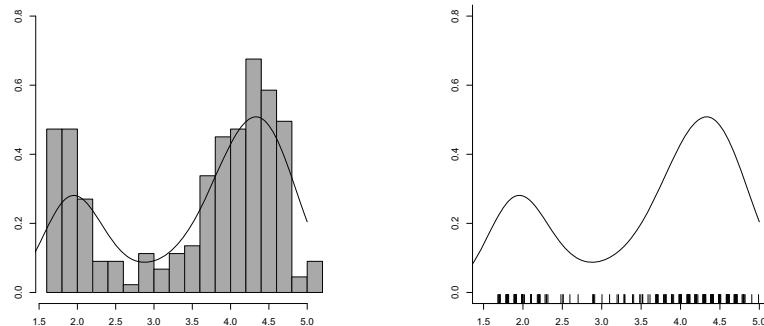
Histogramme (gauche) et diagramme en aiguilles (droite).

▷ Exemple 2 : durée de service entre deux pannes consécutives.



Histogramme (gauche) et diagramme en aiguilles (droite).

- ▷ **Exemple 3 :** durée d'une éruption volcanique.



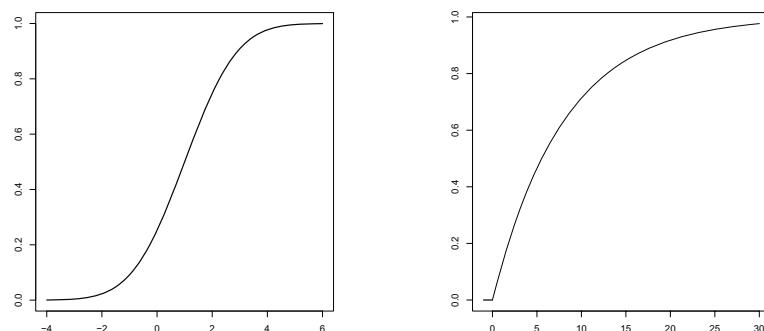
Histogramme (gauche) et diagramme en aiguilles (droite).

### 5.3.1 Fonction de répartition

La fonction de répartition d'une variable aléatoire continue  $X$  est analogue à celle d'une variable aléatoire discrète.

**Définition 8** La *fonction de répartition*  $F_X$  d'une variable aléatoire continue  $X$  est définie par

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du. \quad (7)$$



Exemples de fonctions de répartition de variables aléatoires continues.

### Propriétés

- i)  $F_X$  prend ses valeurs dans  $[0, 1]$ ;
- ii)  $F_X$  est monotone non décroissante;
- iii)  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  et  $\lim_{x \rightarrow +\infty} F_X(x) = 1$ ;
- iv)  $F_X$  est une **fonction continue** dans l'intervalle  $[0, 1]$ ;

## Propriétés (suite)

v)  $F'_x(x) = f_x(x);$

vi) relation entre la probabilité d'intervalle et la fonction de répartition :

$$\begin{aligned} P(a < X \leq b) &= \int_a^b f_x(u) du \\ &= \underbrace{\int_{-\infty}^b f_x(u) du}_{F_X(b)} - \underbrace{\int_{-\infty}^a f_x(u) du}_{F_X(a)} \\ &= F_X(b) - F_X(a). \end{aligned}$$

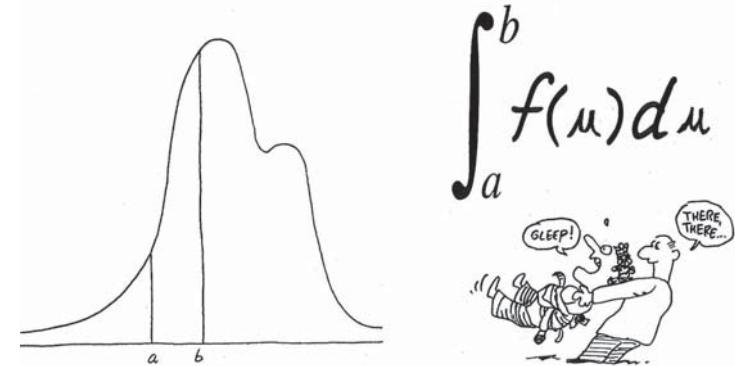


Illustration de la propriété vi) d'une fonction de répartition.

## Remarques :

- a) la fonction de répartition d'une variable aléatoire continue est une probabilité d'intervalle particulière (celle de la demi-droite  $]-\infty, x]$ );
- b) la propriété vi) reste valable pour une variable aléatoire discrète;
- c) connaître la fonction de répartition et connaître la densité est mathématiquement équivalent. Dans la plupart des cas, la loi de probabilité d'une variable aléatoire continue est définie par la densité qui, en général, est plus simple à exprimer.

### 5.3.2 Interprétation intuitive de la fonction de densité

Supposons que la fonction de densité  $f_x$  d'une variable aléatoire continue  $X$  soit continue en  $u = a$ . Soit  $\varepsilon > 0$  un nombre réel relativement petit. Par la propriété ii) en page 37, on a

$$P\left(a - \frac{\varepsilon}{2} < X \leq a + \frac{\varepsilon}{2}\right) = \int_{a-\varepsilon/2}^{a+\varepsilon/2} f_x(u) du \approx \varepsilon \cdot f_x(a).$$

Ainsi, la probabilité que  $X$  prenne une valeur dans un intervalle de longueur  $\varepsilon$  centré en  $a$  est approximativement  $\varepsilon \cdot f_x(a)$ . On en conclut que  $f_x(a)$  est une sorte de **mesure de la probabilité** que  $X$  soit proche de  $a$ .

### 5.3.3 Espérance mathématique

L'espérance mathématique d'une variable aléatoire continue  $X$  est analogue à celle d'une variable aléatoire discrète. La densité de probabilité de  $X$  remplace les probabilités ponctuelles.

**Définition 9** Soit  $X$  une variable aléatoire continue de densité  $f_X$ .

L'**espérance mathématique** de  $X$  est le nombre réel noté  $E(X)$  ou  $\mu_X$  défini par

$$E(X) = \int_{-\infty}^{\infty} u \cdot f_X(u) du. \quad (8)$$

Remarques :

- a) comme dans le cas discret, il existe une formule plus pratique pour calculer la variance d'une variable aléatoire continue

$$\text{Var}(X) = E(X^2) - E(X)^2 \quad (10)$$

$$\text{où } E(X^2) = \int_{-\infty}^{\infty} u^2 \cdot f_X(u) du;$$

- b) indépendamment de son type (discret ou continu), une définition équivalente de la variance d'une variable aléatoire  $X$  est

$$\text{Var}(X) = E[(X - \mu_X)^2], \quad (11)$$

où  $\mu_X$  représente l'espérance mathématique de  $X$ .

En considérant cette définition générale, on remarque que la variance est effectivement une mesure de la **dispersion** de la distribution de  $X$  autour de l'espérance  $\mu_X$ .

### 5.3.4 Variance, écart-type

Les définitions de la variance et de l'écart-type d'une variable aléatoire continue sont similaires à celles d'une variable aléatoire discrète.

**Définition 10** Soit  $X$  une variable aléatoire continue de densité  $f_X$ .

La **variance** de  $X$  est le nombre réel positif noté  $\text{Var}(X)$ ,  $\sigma^2(X)$  ou  $\sigma_X^2$  défini par

$$\text{Var}(X) = \int_{-\infty}^{\infty} (u - E(X))^2 \cdot f_X(u) du. \quad (9)$$

L'**écart-type** est la racine carrée de la variance.

### 5.4 Propriétés de l'espérance mathématique et de la variance

Les propriétés qui suivent sont valables, sans spécification particulière, quelque soit le type de la variable aléatoire.

#### 5.4.1 Propriétés de l'espérance mathématique

Soient  $X$  une variable aléatoire,  $a$  et  $b$  deux constantes réelles.

- i)  $E(b) = b;$
- ii)  $E(X + b) = E(X) + b;$
- iii)  $E(a \cdot X) = a \cdot E(X);$

### 5.4.1 Propriétés de l'espérance mathématique (*suite*)

iv)  $E(a \cdot X + b) = a \cdot E(X) + b$

(linéarité de l'espérance mathématique);

v)  $E(\varphi(X))$  où  $\varphi$  est une fonction réelle :

a) *X variable aléatoire discrète* :

si  $\mathcal{H} = \{x_i\}_{i=1}^n$  est l'ensemble des valeurs prises par  $X$   
et  $p_i = P(X = x_i)$  la loi de probabilité de  $X$ ,

$$E(\varphi(X)) = \sum_{i=1}^n \varphi(x_i) \cdot p_i;$$

### 5.4.1 Propriétés de l'espérance mathématique (*suite*)

v)  $E(\varphi(X))$  où  $\varphi$  est une fonction réelle :

b) *X variable aléatoire continue* :

si  $f_X$  est la densité de  $X$ ,

$$E(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(u) \cdot f_X(u) du;$$

### 5.4.2 Propriétés de la variance

Soient  $X$  une variable aléatoire,  $a$  et  $b$  deux constantes réelles.

i)  $\text{Var}(b) = 0$ ;

ii)  $\text{Var}(X + b) = \text{Var}(X)$ ;

iii)  $\text{Var}(a \cdot X) = a^2 \cdot \text{Var}(X)$ ;

iv)  $\text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X)$ ;

### 5.4.3 Changements d'origine et d'échelle

Il arrive fréquemment qu'une variable aléatoire  $X$  soit transformée par une application affine

$$aX + b,$$

où  $a$  et  $b$  sont des constantes réelles.

Question :

Comment se modifient l'espérance mathématique et la variance sous la transformation affine ?

changement d'origine

$$X \mapsto X + b$$

changement d'échelle

$$X \mapsto a X$$

changement d'origine et d'échelle

$$X \mapsto a X + b$$

$$E(X + b) = E(X) + b$$

$$E(a X) = a E(X)$$

$$E(aX + b) = a E(X) + b$$

$$\text{Var}(X + b) = \text{Var}(X)$$

$$\text{Var}(a X) = a^2 \text{Var}(X)$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

*Changements d'origine et d'échelle de l'espérance mathématique et de la variance.*

Remarques :

- a) la variance ne se modifie pas lors d'un changement d'origine;
- b) une variable d'espérance nulle est appelée **variable aléatoire centrée**;
- c) une variable d'espérance nulle et de variance 1 est dite **variable aléatoire centrée réduite**;
- d) il est possible d'associer à toute variable aléatoire  $X$  une variable  $Z$ ,

$$Z = g(X) = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - E(X)}{\sigma(X)},$$

qui, par construction, est centrée réduite. La transformation  $g$  est très pratique; elle nous permet notamment d'utiliser la plupart des tables.



## EXERCICES : VARIABLES ALÉATOIRES

### Exercice 1

Une pièce de monnaie est lancée jusqu'à l'apparition d'un premier face ou d'une séquence de quatre piles consécutifs. La pièce est truquée; pile apparaît avec probabilité 3/4.

- Déterminer l'ensemble fondamental  $\Omega$ .
- Désignons par  $X$  le nombre de piles. Déterminer sa loi de probabilité.

**Solutions :** a)  $\Omega = \{(F), (P, F), (P, P, F), (P, P, P, F), (P, P, P, P)\}$  b)

| $X = x$    | 0   | 1    | 2    | 3      | 4      |
|------------|-----|------|------|--------|--------|
| $P(X = x)$ | 1/4 | 3/16 | 9/64 | 27/256 | 81/256 |

### Exercice 2

Une pièce de monnaie équilibrée est jetée quatre fois de suite par un croupier.

- Déterminer l'ensemble fondamental  $\Omega$  de l'expérience aléatoire.
- Désignons par  $X$  la différence entre le nombre de piles et le nombre de faces. Déterminer sa loi de probabilité.

**Solution :** b)

| $X = x$    | -4   | -2  | 0   | 2   | 4    |
|------------|------|-----|-----|-----|------|
| $P(X = x)$ | 1/16 | 1/4 | 3/8 | 1/4 | 1/16 |

### Exercice 3

Dans un jeu, une pièce de monnaie est lancée. Si elle tombe sur face, le joueur gagne 2 Frs; si elle tombe sur pile, le joueur ne reçoit aucun gain. Tout joueur doit verser une mise de 1 Frs à chaque jet de la pièce. Mister Prob adopte la stratégie suivante : s'il gagne au premier jet, il arrête de jouer. S'il perd au premier jet, il joue encore deux fois et se retire ensuite du jeu. Désignons par  $X$  le gain ou la perte de Mister Prob lorsqu'il arrête de jouer. On suppose que la pièce est truquée; face apparaît avec probabilité 1/3.

- Déterminer les valeurs que peut prendre la variable aléatoire  $X$ .
- Donner la loi de probabilité de  $X$ .
- Calculer la probabilité  $P(X > 0)$ .

**Solutions :** a)  $\mathcal{H} = \{-3, -1, 1\}$  b)

| $X = x$    | -3   | -1   | 1     |
|------------|------|------|-------|
| $P(X = x)$ | 8/27 | 8/27 | 11/27 |

c) 11/27.

### Exercice 4

Un bit transmis par un canal digital est reçu sans erreur avec probabilité 0.9. On suppose que les transmissions d'un bit sont indépendantes. Désignons par  $X$  le nombre de bits reçus avec erreur parmi les quatre prochains bits transmis par le canal.

- Déterminer l'ensemble fondamental  $\Omega$  de l'expérience aléatoire à l'aide des événements :

$O$  : "un bit est reçu sans erreur" et  $E$  : "un bit est reçu avec erreur".

- Construire la loi de probabilité de  $X$ .

**Solutions :** a)

$$\Omega = \{(OOOO), (OOOE), (OOEO), (OOEE), (OEOO), (OEOE), (OEEE), \\ (EOOO), (EOOE), (EOEO), (EOEE), (EEOO), (EEOE), (EEE0), (EEEE)\}$$

b)

| $X = x$    | 0      | 1      | 2      | 3      | 4      |
|------------|--------|--------|--------|--------|--------|
| $P(X = x)$ | 0.6561 | 0.2916 | 0.0486 | 0.0036 | 0.0001 |

### Exercice 5

La loi de probabilité d'une variable aléatoire  $X$  est

$$P(X = x) = \frac{2x + 1}{25}, \quad x = 0, 1, 2, 3, 4.$$

Calculer la probabilité  $P(2 \leq X < 4)$ .

**Solution :** 12/25.

### Exercice 6

Considérons la variable aléatoire  $X$  dont la loi de probabilité est donnée par

| $X = x$    | 1       | 2      | 3      | 4      | 5      |
|------------|---------|--------|--------|--------|--------|
| $P(X = x)$ | 137/300 | 77/300 | 47/300 | 27/300 | 12/300 |

- Construire soigneusement l'histogramme et la fonction de répartition de  $X$ .
- Déterminer l'espérance mathématique et la variance de  $X$ .
- Calculer la probabilité  $P(X \leq 1)$ .

**Solutions :** b)  $E(X) = 2$ ,  $\text{Var}(X) = 4/3$  c) 137/300.

### Exercice 7

Un ingénieur soumet ses trois batchs  $A$ ,  $B$  et  $C$  aux ordinateurs du service informatique. Les probabilités respectives qu'ils soient exécutés par l'ordinateur le plus performant sont

$$P(A) = 0.5, \quad P(B) = 0.8 \text{ et } P(C) = 0.2.$$

On suppose que l'attribution ou non des batchs à cet ordinateur est totalement indépendante.

Désignons par  $X$  le nombre de batchs exécutés par l'ordinateur le plus performant.

- Déterminer la loi de probabilité de  $X$ .
- Construire soigneusement l'histogramme et la fonction de répartition de  $X$ .
- Déterminer l'espérance mathématique et la variance de  $X$ .

d) Calculer la probabilité  $P(X \leq 2)$ .

**Solutions :** a)

| $X = x$    | 0    | 1     | 2     | 3    |
|------------|------|-------|-------|------|
| $P(X = x)$ | 2/25 | 21/50 | 21/50 | 2/25 |

c)  $E(X) = 3/2$ ,  $\text{Var}(X) = 57/100$  d) 0.92.

### Exercice 8

La fonction de répartition d'une variable aléatoire  $X$  est

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0, \\ \frac{x}{4} & \text{si } 0 \leq x < 1, \\ \frac{1}{2} + \frac{x-1}{4} & \text{si } 1 \leq x < 2, \\ \frac{11}{12} & \text{si } 2 \leq x < 3, \\ 1 & \text{si } 3 \leq x. \end{cases}$$

Calculer les probabilités

- a)  $P(X = 1)$ ;
- b)  $P\left(\frac{1}{2} < X < \frac{3}{2}\right)$ .

**Solutions :** a) 1/4 b) 1/2.

### Exercice 9

Pour aplanir le terrain en vue de la construction d'une bretelle d'autoroute, le chef de chantier projette d'utiliser trois bulldozers. Désignons par  $X$  le nombre de bulldozers encore en état de fonctionnement à la fin de l'aplanissement du terrain. On suppose que la loi de probabilité de  $X$  soit

| $X = x$    | 0     | 1     | 2     | 3     |
|------------|-------|-------|-------|-------|
| $P(X = x)$ | 0.008 | 0.096 | 0.384 | 0.512 |

Calculer le coefficient de variation  $\delta_X$  de  $X$  défini par  $\frac{\sigma_X}{\mu_X}$  où  $\sigma_X$  et  $\mu_X$  représentent respectivement l'écart-type et l'espérance mathématique de  $X$ .

**Solution :** 0.2887.

### Exercice 10

Deux options  $A$  et  $B$  sont proposées à un client pour faire fructifier ses investissements. Dans le tableau ci-dessous figurent les profits potentiels en dollars et les probabilités associées.

| option A |             | option B |             |
|----------|-------------|----------|-------------|
| profit   | probabilité | profit   | probabilité |
| -1500    | 0.2         | -2500    | 0.2         |
| -100     | 0.1         | -500     | 0.1         |
| 500      | 0.4         | 1500     | 0.3         |
| 1500     | 0.2         | 2500     | 0.3         |
| 3500     | 0.1         | 3500     | 0.1         |

a) En se basant sur l'espérance mathématique, déterminer l'option la plus rentable.

b) L'écart-type est une mesure de la variabilité des profits et par conséquent du risque de l'investissement. Quelle option est la plus risquée ?

**Solutions :** a) 540 \$ et 1000 \$; l'option  $B$  est plus rentable b) 1390.83 \$ et 2012.46 \$; l'option  $B$  est plus risquée.

### Exercice 11

Imaginons qu'on souhaite transmettre les réalisations  $x_1, x_2, \dots, x_n$  d'une variable aléatoire discrète  $X$  d'un point d'observation  $A$  à un point de réception  $B$  à l'aide d'un canal de communication ne pouvant transférer que des 0 ou des 1. Ainsi, les valeurs prises par  $X$  devront être codées en chaînes formées uniquement de 0 et de 1 avant d'être transmises. Pour éviter toute ambiguïté, on exige qu'un code ne puisse pas être une extension d'un autre.

Comme exemple, supposons que les réalisations de  $X$  sont  $x_1, x_2, x_3$  et  $x_4$ . Comme code possible, on peut envisager

$$x_1 \leftrightarrow 00, \quad x_2 \leftrightarrow 01, \quad x_3 \leftrightarrow 10, \quad x_4 \leftrightarrow 11. \quad (1)$$

Ainsi, si  $X$  prend la valeur  $x_1$ , le message envoyé en  $B$  sera 00, il vaudra 01 si  $X = x_2$  et ainsi de suite. Un autre code possible est

$$x_1 \leftrightarrow 0, \quad x_2 \leftrightarrow 10, \quad x_3 \leftrightarrow 110, \quad x_4 \leftrightarrow 111. \quad (2)$$

En revanche, le codage

$$x_1 \leftrightarrow 0, \quad x_2 \leftrightarrow 1, \quad x_3 \leftrightarrow 00, \quad x_4 \leftrightarrow 01$$

n'est pas admis étant donné que  $x_3$  et  $x_4$  sont des extensions de  $x_1$ .

Un objectif du codage consiste tout naturellement à minimiser le nombre espéré de bits nécessaires pour transmettre l'information. Ainsi, un code est dit *plus efficace* qu'un autre si son nombre espéré de bits est plus petit que celui nécessaire à l'autre code.

a) Supposons que la loi de probabilité de la variable aléatoire  $X$  est

| $X = x$    | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|------------|-------|-------|-------|-------|
| $P(X = x)$ | 1/2   | 1/4   | 1/8   | 1/8   |

et considérons les codes donnés ci-dessus par (1) et (2).

Pour cette distribution de  $X$ , quel est le code le plus efficace ?

b) Considérons en toute généralité une variable aléatoire discrète  $X$  prenant ses valeurs dans l'ensemble  $\{x_1, x_2, \dots, x_n\}$  avec probabilités correspondantes  $p_1, p_2, \dots, p_n$ . La grandeur

$$H(X) = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$$

est appelée, en théorie de l'information, *entropie* de la variable aléatoire  $X$ . Par convention, si l'une des probabilités  $p_i$  est nulle, on pose  $0 \cdot \log_2(0) = 0$ . L'entropie représente en quelque sorte la quantité d'incertitude relative à la variable aléatoire  $X$ . Autrement dit, on considère  $H(X)$  comme l'*information* liée à l'observation de  $X$ .

Calculer l'entropie de  $X$  selon la distribution donnée en a).

c) Selon le théorème du codage sans bruit, tout codage nécessite un nombre espéré de bits au minimum égal à l'entropie de  $X$ .

L'un des deux codes présentés ci-dessus est-il le code le plus efficace qu'il puisse exister pour  $X$  ? Est-ce surprenant ?

**Solutions :** a) le code donné par (2) nécessite un nombre espéré de 2 bits alors que celui nécessaire au code donné par (1) pour transmettre l'information est de 1.75 bit. Ainsi, le code donné par (2) est plus efficace que celui donné par (1) b) selon la distribution donnée,  $H(X) = 1.75$  c) le code donné par (2) est le code le plus efficace qu'il puisse exister. Le résultat n'est pas très surprenant en raison de la construction du code.

### Exercice 12

Un jeu consiste à jeter une pièce de monnaie équilibrée jusqu'à l'apparition du premier pile. Un joueur gagne  $2^n$  Frs si pile apparaît au  $n$ -ième jet.

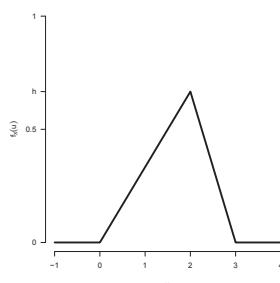
Posons  $X$  le gain du joueur.

- a) Déterminer l'ensemble  $\mathcal{H}$  des valeurs prises par  $X$ .
- b) Donner les probabilités associées sous la forme d'une formule générale en fonction de  $n$ .
- c) Montrer que le gain espéré est infini, i.e.  $E(X) = \infty$ .
- d) En admettant que vous possédez une immense fortune, seriez-vous disposé à verser 10 millions de francs pour pouvoir prendre part une fois au jeu ?

**Solutions :** a)  $\mathcal{H} = \{2^n\}_{n=1}^{\infty}$  b)  $P(X = 2^n) = \left(\frac{1}{2}\right)^n$ ,  $n = 1, 2, \dots$  d) sûrement pas.

### Exercice 13

La longueur  $X$  d'une pièce fabriquée en série est distribuée selon une fonction de densité  $f_X$  dont le graphe est représenté dans la figure ci-dessous. Déterminer la constante  $h$ .



**Solution :** 2/3.

### Exercice 14

La durée en mois d'un certain travail de maçonnerie peut être représentée par une variable aléatoire  $X$  de fonction de répartition  $F_X$ ,

$$F_X(x) = \begin{cases} 0 & \text{si } x < 1, \\ x^2 - 2x + 1 & \text{si } 1 \leq x \leq 2, \\ 1 & \text{si } x > 2. \end{cases}$$

- a) Déterminer la fonction de densité  $f_X$  de  $X$ .

- b) Calculer la probabilité  $P(X > 1.5)$  en utilisant uniquement la fonction de répartition  $F_X$ .

**Solutions :** a)  $f_X(u) = 2u - 2$  si  $1 \leq u \leq 2$  et  $f_X(u) = 0$  sinon b) 3/4.

### Exercice 15

Selon la demande du client, le diamètre du trou percé dans une pièce métallique est fixé à 12.5 millimètres. Des tests ont mis en évidence des perturbations dans le procédé d'usinage qui font apparaître des trous de diamètres supérieurs à la valeur cible. Des expériences ont montré que le diamètre du trou de la pièce peut être représenté par une variable aléatoire  $X$  de fonction de densité

$$f_X(u) = \begin{cases} 20 \cdot e^{-20(u-12.5)} & \text{si } u \geq 12.5, \\ 0 & \text{sinon.} \end{cases}$$

Le client accepte une pièce pour autant que le diamètre du trou soit compris entre 12.5 et 12.6 millimètres. Calculer la proportion de pièces acceptées par le client.

**Solution :**  $1 - e^{-2}$ .

### Exercice 16

Considérons la variable aléatoire  $X$  de fonction de densité

$$f_X(u) = \begin{cases} e^{-(u-4)} & \text{si } u > 4, \\ 0 & \text{sinon.} \end{cases}$$

- a) Déterminer la fonction de répartition  $F_X$  de  $X$  et construire son graphe.
- b) Calculer la probabilité  $P(2 \leq X < 5)$ .

**Solutions :** a)  $F_X(x) = 1 - e^{-(x-4)}$  si  $x > 4$  et 0 sinon b)  $1 - e^{-1}$ .

### Exercice 17

La hauteur en mètres des vagues dans la Mer du Nord peut être représentée par une variable aléatoire  $X$  dont la fonction de densité est

$$f_X(u) = \begin{cases} \frac{2u}{\beta} \cdot e^{-\frac{u^2}{\beta}} & \text{si } u > 0, \\ 0 & \text{sinon.} \end{cases}$$

À partir d'observations, le paramètre  $\beta$  a été estimé à 8. Déterminer la fonction de répartition  $F_X$  de  $X$ .

**Solution :**  $F_X(x) = 1 - e^{-x^2/8}$  si  $x > 0$  et 0 sinon.

### Exercice 18

La durée de vie d'une savonnette mesurée en jours peut être décrite par une variable aléatoire  $X$  dont la fonction de densité est donnée par

$$f_X(u) = \begin{cases} \lambda^2 u e^{-\lambda u} & \text{si } u > 0, \\ 0 & \text{sinon} \end{cases}$$

avec  $\lambda = 1/10$ .

- a) Déterminer la fonction de répartition  $F_X$  de  $X$ .  
b) Calculer la probabilité qu'une savonnette dure plus de 30 jours en sachant qu'elle a déjà vécu plus de 10 jours.

**Solutions :** a)  $F_X(x) = 1 - e^{-\lambda x} \cdot (\lambda x + 1)$  si  $x > 0$  et 0 sinon    b)  $2e^2$ .

#### Exercice 19

Le temps de réaction en secondes d'un muscle à un certain stimulus peut être représenté par une variable aléatoire  $X$  de fonction de densité

$$f_X(u) = \begin{cases} \frac{3}{2} \cdot \frac{1}{u^2} & \text{si } 1 \leq u < 3, \\ 0 & \text{sinon.} \end{cases}$$

Calculer le temps de réaction espéré.

**Solution :**  $3/2 \cdot \ln(3)$ .

#### Exercice 20

La résistance latérale de la charpente d'une petite construction peut être représentée par une variable aléatoire  $X$  dont la fonction de densité est

$$f_X(u) = \begin{cases} c \cdot (2u - u^2) & \text{si } 0 < u < 2, \\ 0 & \text{sinon.} \end{cases}$$

- a) Déterminer la constante  $c$ .  
b) Donner la fonction de répartition de  $X$  et construire soigneusement sa représentation graphique.  
c) Calculer la résistance latérale espérée de la charpente.  
d) Déterminer la valeur  $\tilde{x}$ , appelée médiane de la distribution de  $X$ , définie par  $P(X \leq \tilde{x}) = 0.5$ .

**Solutions :** a)  $c = 3/4$     b)

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ \frac{3}{4}x^2 - \frac{1}{4}x^3 & \text{si } 0 < x < 2, \\ 1 & \text{si } x \geq 2. \end{cases}$$

c)  $E(X) = 1$     d)  $\tilde{x} = 1$ .

#### Exercice 21

L'épaisseur en micromètres ( $\mu m$ ) d'une couche de peinture enduite sur certaines vis peut être représentée par une variable aléatoire  $X$  dont la fonction de densité est donnée par

$$f_X(u) = \begin{cases} c \cdot u^{-2} & \text{si } 100 < u < 120, \\ 0 & \text{sinon.} \end{cases}$$

- a) Montrer que la constante  $c$  vaut 600.  
b) Déterminer l'épaisseur espérée de la couche de peinture.

- c) Calculer  $E(X^2)$  puis en déduire la variance de  $X$ .

**Solutions :** b)  $600 \cdot \ln(1.2) \approx 109.39 \mu m$     c)  $E(X^2) = 12'000$ ;  $\text{Var}(X) \approx 33.19 \mu m^2$ .

#### Exercice 22

Considérons une variable aléatoire  $X$  dont la fonction de densité est donnée par

$$f_X(u) = \begin{cases} 1 & \text{si } 0 < u < 1, \\ 0 & \text{sinon.} \end{cases}$$

Désignons par  $Y$  la variable aléatoire  $Y = e^X$ . Déterminer l'espérance de  $Y$ .

**Solution :**  $e - 1$ .

## Chapitre 6

### Distributions usuelles



## 6.1 Lois discrètes

Dans ce chapitre, nous introduirons les principales classes de variables aléatoires, autrement dit les distributions usuelles de probabilités.

Elles nous permettent de modéliser différentes situations pratiques comme par exemple le nombre de pièces défectueuses dans un échantillon prélevé aléatoirement dans la production ou la durée d'une conversation téléphonique.

Rappelons que ces distributions ou modèles probabilistes ne sont qu'une représentation, avec qualités et défauts, de la réalité.

Nous présenterons d'abord les distributions discrètes usuelles puis les distributions continues standards.

## Contenu

6.1.1 Loi de Bernoulli

6.1.2 Loi binomiale

6.1.3 Loi géométrique

6.1.4 Introduction aux processus de Poisson

6.1.5 Loi de Poisson

### 6.1.1 Loi de Bernoulli (J. Bernoulli, 1654–1705, mathématicien suisse)

**Utilisation :** on réalise une expérience dont l'issue est interprétée soit comme un succès soit comme un échec. On définit alors une variable aléatoire  $X$  en lui attribuant la valeur 1 lors d'un succès et 0 lors d'un échec.

- un paramètre :  $p$  ( $0 \leq p \leq 1$ )

- $\mathcal{H} = \{0, 1\}$

- loi de probabilité :

$$P(X = x) = \begin{cases} p & \text{si } x = 1, \\ 1 - p = q & \text{si } x = 0. \end{cases}$$

- $E(X) = p$ ,  $\text{Var}(X) = pq$



Pour résoudre le problème du Chevalier de Méré, il faut utiliser la distribution, hum, hum, hum... binomiale !

### 6.1.2 Loi binomiale

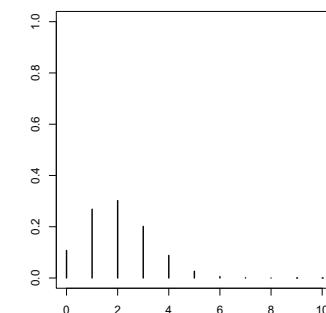
**Utilisation :**  $n$  épreuves indépendantes de Bernoulli se réalisent, chacune d'elles ayant  $p$  comme probabilité de succès. On définit une variable aléatoire  $X$  qui compte le nombre de succès sur l'ensemble des  $n$  épreuves.

- deux paramètres :  $n$  ( $n \in \mathbb{N}^*$ ) et  $p$  ( $0 \leq p \leq 1$ )

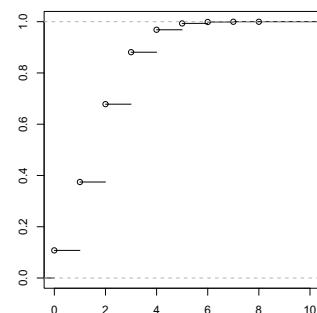
- $\mathcal{H} = \{0, 1, \dots, n\}$

- loi de probabilité :

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} p^x q^{n-x}, \quad x = 0, \dots, n.$$



Loi de probabilité (gauche) et fonction de répartition (droite) d'une variable aléatoire issue d'une distribution binomiale de paramètres  $n = 10$  et  $p = 0.2$ .



Comme  $X$  est une somme de  $n$  variables de Bernoulli identiques et indépendantes, on a immédiatement

- $E(X) = np$
- $\text{Var}(X) = np(1 - p) = npq$

**Notation :**  $\mathcal{B}(n, p)$ .

En posant  $n = 1$ , on obtient la loi de Bernoulli.

### 6.1.3 Loi géométrique

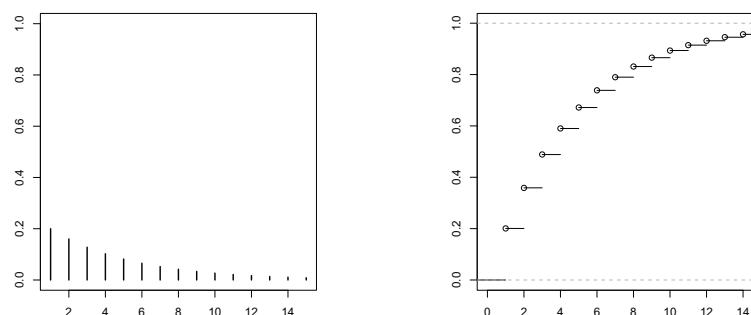
**Utilisation :** on répète des épreuves indépendantes de Bernoulli ayant chacune la probabilité  $p$  d'être un succès. Désignons par  $X$  le nombre d'épreuves nécessaires pour obtenir le premier succès.

- un paramètre :  $p$  ( $0 < p \leq 1$ )
- $\mathcal{H} = \{1, 2, \dots\}$
- loi de probabilité :

$$P(X = x) = p \underbrace{(1-p)}_q^{x-1} = p q^{x-1}, \quad x = 1, 2, \dots$$

- $E(X) = \frac{1}{p}$ ,  $\text{Var}(X) = \frac{q}{p^2}$

**Notation :**  $\mathcal{G}(p)$ .



Loi de probabilité (gauche) et fonction de répartition (droite) d'une variable aléatoire issue d'une distribution géométrique de paramètre  $p = 0.2$ .

### 6.1.4 Introduction aux processus de Poisson

(S.-D. Poisson, 1781–1840, mathématicien français)

Les situations dans lesquelles un événement particulier se reproduit à intervalles réguliers au cours du temps s'observent fréquemment dans la vie de tous les jours. On assiste à un flux d'événements qui se réalisent les uns à la suite des autres.

Exemples d'événements particuliers :

- tremblement de terre;
- entrée d'une personne dans un établissement donné (banque, poste, station essence, ...);
- panne;
- ...

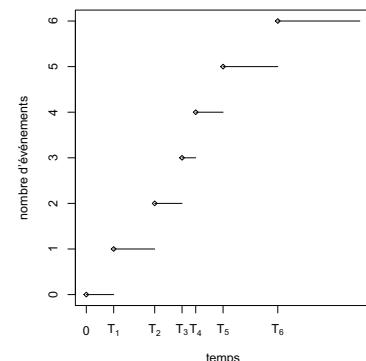


Schéma d'un processus de Poisson.

### Question :

Comment peut-on modéliser de tels phénomènes ?

### Convention et définition :

- par convention, la suite des événements débute au temps  $t = 0$ ;
- $N(t)$  : variable aléatoire discrète qui compte le nombre total d'événements survenus entre 0 et  $t$ .

### Exemples :

- nombre de séismes survenant pendant une période donnée;
- nombre d'électrons libérés par une cathode durant un laps de temps donné;
- nombre de décès parmi les assurés d'une compagnie d'assurance-vie sur une période donnée;
- nombre d'accidents à un endroit fixe pendant une année;
- nombre de requêtes traitées par un serveur durant un intervalle de temps donné;
- ...

### Hypothèses :

1. un seul événement arrive à la fois;
2. le nombre d'événements se produisant pendant une période  $T$  ne dépend que de la durée de cette période;
3. les événements sont indépendants.

### Objectif :

Calculer la probabilité  $P(N(t) = x)$ ,  $x = 0, 1, \dots$

Sous les hypothèses 1-3, on peut montrer que

$$P(N(t) = x) = e^{-\lambda t} \cdot \frac{(\lambda t)^x}{x!}, \quad x = 0, 1, \dots \quad (1)$$

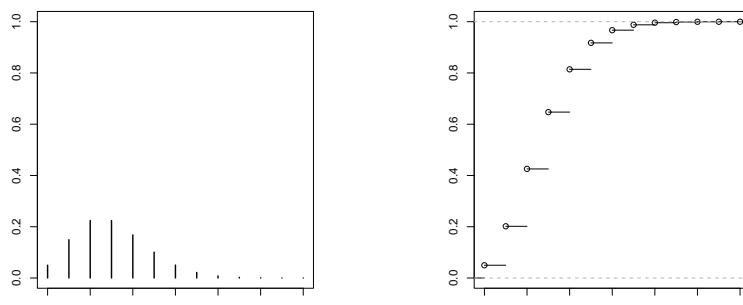
où  $\lambda$  est un nombre réel strictement positif.

**Définition 1** Le nombre d'occurrences d'événements remplissant les hypothèses 1-3 est une variable aléatoire dite de **Poisson** de paramètre  $\lambda t$ . Les événements se réalisent alors selon un **processus de Poisson** de paramètre  $\lambda$ .

Comme

$$P(N(t + \Delta t) = N(t) + 1) \sim \lambda \cdot \Delta t \text{ lorsque } \Delta t \rightarrow 0,$$

le paramètre  $\lambda > 0$  est appelé le **taux d'accroissement**. En principe inconnu, ce paramètre est estimé à la suite d'expériences.



Loi de probabilité (gauche) et fonction de répartition (droite) d'une variable aléatoire issue d'une distribution de Poisson de paramètre  $\lambda = 3$ .

### 6.1.5 Loi de Poisson

**Utilisation :** les variables aléatoires de Poisson sont principalement utilisées pour compter le nombre d'occurrences d'événements dans diverses expériences. Soit  $X$  une variable aléatoire de Poisson.

- un paramètre :  $\lambda$  ( $\lambda > 0$ )
- $\mathcal{H} = \{0, 1, \dots\}$
- loi de probabilité :

$$P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots$$

- $E(X) = \lambda$ ,  $\text{Var}(X) = \lambda$

Notation :  $\mathcal{P}(\lambda)$ .

Remarque :

$e^{-\lambda}$  est une constante de normalisation car  $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$ .

Application de la loi de Poisson : approximation de la loi binomiale

Si  $n$  est grand et  $p$  petit (pas facile d'être plus précis !!!), une variable aléatoire de Poisson de paramètre  $\lambda = np$  peut être utilisée pour approcher une variable aléatoire binomiale de paramètres  $n$  et  $p$ .

Parmi les phénomènes pouvant être modélisés par une loi de Poisson, citons :

- le nombre de fautes d'impression par page dans un livre;
- le nombre de pièces défectueuses dans une livraison importante, la production étant de bonne qualité;
- le nombre d'individus dépassant l'âge de 100 ans dans une communauté;
- le nombre de faux numéros téléphoniques composés en un jour;
- ...

| loi de $X$  | loi de probabilité de $X$                 | $\mathcal{H}$        | paramètre(s)                          | $E(X)$    | $Var(X)$    |
|-------------|-------------------------------------------|----------------------|---------------------------------------|-----------|-------------|
| binomiale   | $\binom{n}{x} p^x (1-p)^{n-x}$            | $x = 0, 1, \dots, n$ | $n \in \mathbb{N}^*, 0 \leq p \leq 1$ | $np$      | $np(1-p)$   |
| géométrique | $p(1-p)^{x-1}$                            | $x = 1, 2, \dots$    | $0 < p \leq 1$                        | $1/p$     | $(1-p)/p^2$ |
| Poisson     | $e^{-\lambda} \cdot \frac{\lambda^x}{x!}$ | $x = 0, 1, \dots$    | $\lambda > 0$                         | $\lambda$ | $\lambda$   |

Résumé des caractéristiques des lois discrètes usuelles.



## 6.2 Lois continues

Dans ce paragraphe, nous présenterons les principales distributions continues utilisées dans la pratique. Parmi elles figure la distribution normale (ou distribution gaussienne) qui joue un rôle fondamental non seulement dans le contrôle industriel mais aussi dans toute la statistique et dans la théorie des probabilités. Cette distribution est incontournable et porte bien son nom. Le premier mathématicien à l'avoir utilisée est Abraham de Moivre, pionnier du consulting en statistique, en 1733 dans son approximation d'une distribution binomiale par une distribution normale.

## Contenu

6.2.1 Loi uniforme

6.2.2 Loi exponentielle

6.2.3 Loi normale (ou loi de Laplace – Gauss)

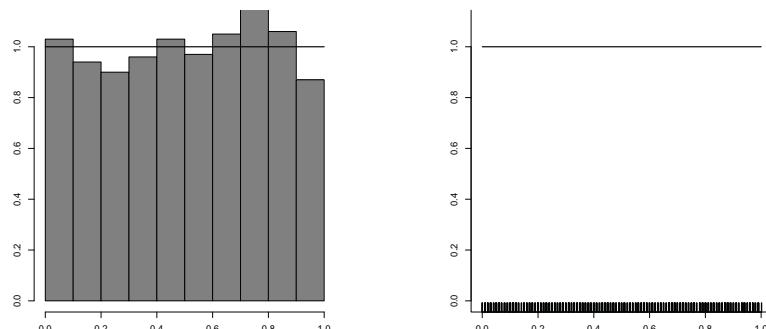
### 6.2.1 Loi uniforme

**Utilisation :** la loi uniforme est utilisée pour représenter des phénomènes aléatoires continus uniformément répartis sur un intervalle. On définit alors une variable  $X$  dite uniformément distribuée sur un intervalle  $[a, b]$  avec  $a, b \in \mathbb{R}$ ,  $a < b$ .

Exemples :

- heure indiquant la fin d'un batch informatique entre son heure de début d'exécution et une durée maximale de 8 heures;
- choix d'un point sur un segment;
- distance de l'endroit d'une panne à une ville donnée;
- ...

▷ Exemple : générateur de nombres (pseudo-)aléatoires dans  $[0, 1]$ .



Histogramme (gauche) et diagramme en aiguilles (droite).



Copyright © 2001 United Feature Syndicate, Inc.

- deux paramètres :  $a, b \in \mathbb{R}$  avec  $a < b$  (définition de l'intervalle)

- $\mathcal{H} = [a, b]$

- fonction de densité :

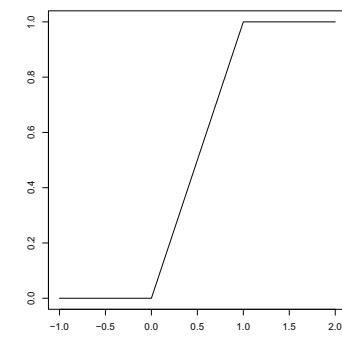
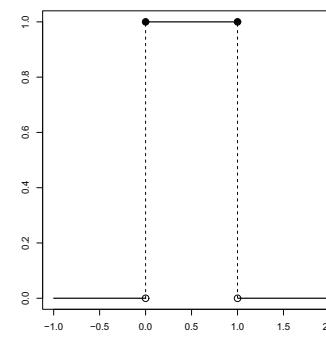
$$f_X(u) = \frac{1}{b-a} \quad \text{si } a \leq u \leq b, \quad f_X(u) = 0 \quad \text{sinon.}$$

- fonction de répartition :

$$F_X(x) = \begin{cases} 0 & \text{si } x < a, \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b, \\ 1 & \text{si } x > b. \end{cases}$$

- $E(X) = \frac{a+b}{2}, \quad \text{Var}(X) = \frac{(b-a)^2}{12}$

Notation :  $\mathcal{U}(a, b)$ .



Densité (gauche) et fonction de répartition (droite) d'une variable aléatoire uniformément distribuée sur  $[0, 1]$ .

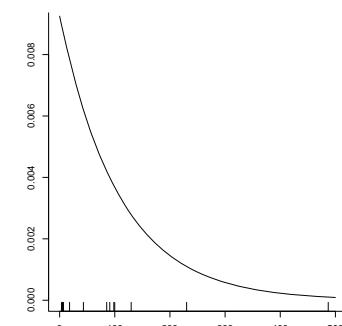
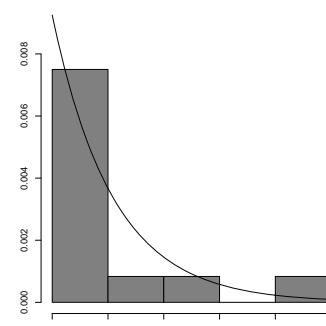
## 6.2.2 Loi exponentielle

**Utilisation :** la loi exponentielle est souvent utilisée pour décrire le temps d'attente d'un événement spécifié. Désignons par  $X$  une variable aléatoire issue d'une distribution exponentielle.

Exemples :

- temps d'attente d'un phénomène poissonnien de taux  $\lambda$  : temps d'attente du premier événement ou temps entre deux événements consécutifs;
- durée de vie d'un transistor;
- durée d'une conversation téléphonique;
- ...

▷ Exemple : heures de service entre deux pannes consécutives.



Histogramme (gauche) et diagramme en aiguilles (droite).

- un paramètre :  $\lambda$  ( $\lambda > 0$ )

- $\mathcal{H} = [0, \infty[$

- fonction de densité :

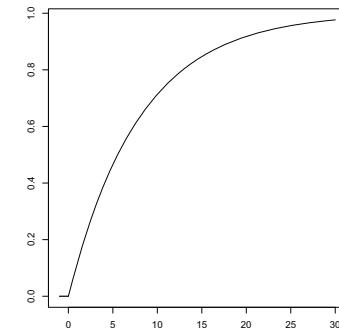
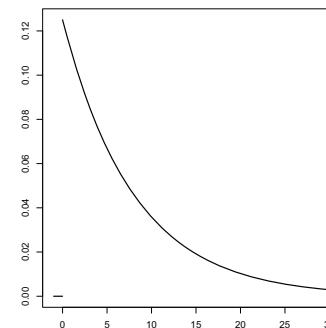
$$f_X(u) = \begin{cases} \lambda e^{-\lambda u} & \text{si } u \geq 0, \\ 0 & \text{si } u < 0. \end{cases}$$

- fonction de répartition :

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0, \\ 1 - e^{-\lambda x} & \text{si } x \geq 0. \end{cases}$$

- $E(X) = \frac{1}{\lambda}$ ,  $\text{Var}(X) = \frac{1}{\lambda^2}$

Notation :  $\mathcal{E}(\lambda)$ .



Densité (gauche) et fonction de répartition (droite)  
d'une variable aléatoire exponentielle de paramètre  $\lambda = 1/8$ .

### Propriété d'absence de mémoire

**Définition 2** Une variable aléatoire est dite **sans mémoire** si pour tous  $s$  et  $t$  positifs,

$$P(X > s + t | X > t) = P(X > s). \quad (2)$$

### Interprétation

imaginons que  $X$  représente la durée de vie d'un certain appareil. La propriété d'absence de mémoire signifie que si l'appareil fonctionne encore après  $t$  heures de service, la distribution de sa durée de vie à partir de ce moment-là est la **même** que la distribution de la durée de vie de l'appareil neuf.

En d'autres termes, l'appareil fonctionne sans mémoire du temps d'usage déjà écoulé.

Selon la définition des probabilités conditionnelles, la relation (2) est équivalente à

$$\frac{P(X > s + t, X > t)}{P(X > t)} = P(X > s),$$

ou encore

$$P(X > s + t) = P(X > s) \cdot P(X > t). \quad (3)$$

Puisque  $e^{-\lambda(s+t)} = e^{-\lambda s} \cdot e^{-\lambda t}$ , la relation (3) est vérifiée par toute variable aléatoire issue d'une distribution exponentielle.

La classe des variables aléatoires exponentielles est donc **sans mémoire**.

### 6.2.3 Loi normale (ou loi de Laplace – Gauss)

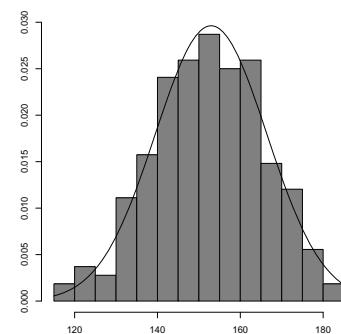
(P.-S. de Laplace, 1749–1827, astronome, mathématicien et physicien français et C.-F. Gauss, 1777–1855, mathématicien, physicien et astronome allemand)

**Utilisation :** la distribution normale doit sa notoriété au théorème central limite qui sera abordé plus tard dans le cours. Ce théorème est l'un des plus importants de la théorie des probabilités; il sert de base pour expliquer qu'en pratique de très nombreux phénomènes aléatoires suivent approximativement une distribution normale.

Exemples :

- taille d'un individu choisi au hasard;
- erreur de mesure d'une quantité physique;
- ...

▷ Exemple : chaleur mensuelle consommée.



Histogramme (gauche) et diagramme en aiguilles (droite).

Conclusion :

Pour représenter la chaleur mensuelle consommée, il convient d'utiliser une variable aléatoire  $X$  dont le graphe de la fonction de densité a la forme d'une "cloche" symétrique par rapport à un axe vertical. La surface sous la courbe possède une région dans laquelle la "masse probabiliste" est plus dense.

**Définition 3** On dit que la variable aléatoire continue  $X$  est issue d'une **distribution normale (ou gaussienne)** si elle possède pour fonction de densité

$$f_X(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}}, \quad -\infty < u < \infty.$$

- deux paramètres :  $\mu \in \mathbb{R}$  et  $\sigma^2 \in \mathbb{R}^+$
- $\mathcal{H} = \mathbb{R}$
- fonction de densité :

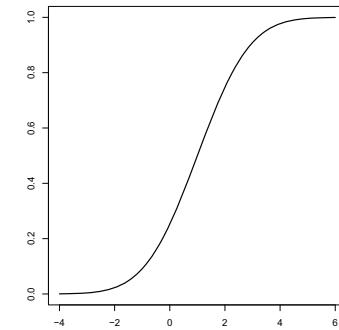
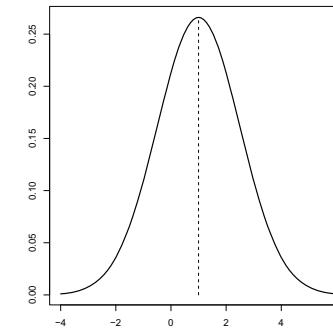
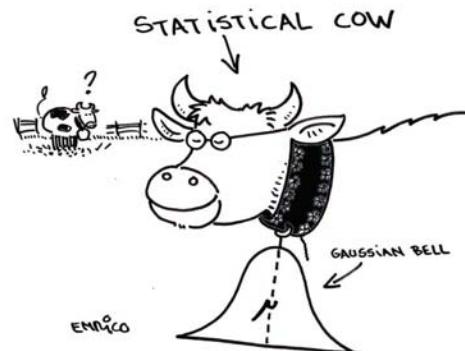
$$f_X(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}}, \quad -\infty < u < \infty$$

- fonction de répartition :

$$F_X(x) = \int_{-\infty}^x f_X(u) du, \quad -\infty < x < \infty$$

- $E(X) = \mu, \quad \text{Var}(X) = \sigma^2$

Notation :  $\mathcal{N}(\mu, \sigma^2)$ .

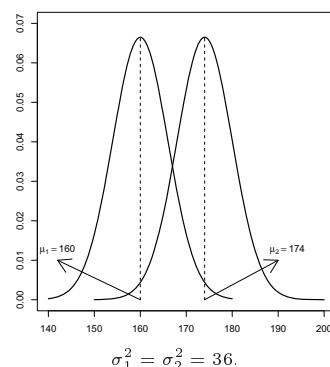


Densité (gauche) et fonction de répartition (droite) d'une variable aléatoire

issue d'une distribution normale telle que  $\mu = 1$ ,  $\sigma^2 = 2.25$ .

### Quelques remarques...

- a) Que se passe-t-il si on modifie l'espérance  $\mu$  de  $X$  ?



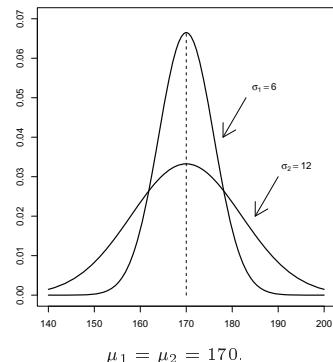
◇ Constatation :

Modifier la valeur de l'espérance  $\mu$  de  $X$  revient à déplacer la courbe de la fonction de densité le long de l'axe des abscisses.

La grandeur  $\mu$  peut être considérée comme un **paramètre de position** de la distribution de  $X$ .

### Quelques remarques (suite)...

b) Que se passe-t-il si on augmente l'écart-type  $\sigma$  de  $X$  ?



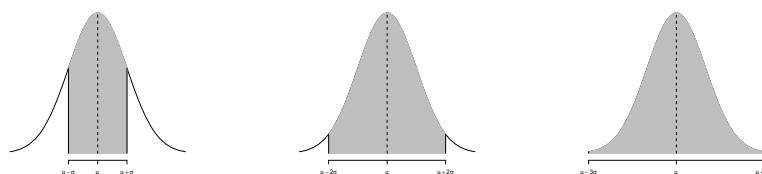
◊ Constatation :

Augmenter la valeur de l'écart-type  $\sigma$  de  $X$  revient à étirer et aplatisir la courbe de la fonction de densité.

La grandeur  $\sigma$  peut être considérée comme un **paramètre d'échelle** de la distribution de  $X$ .

### Quelques remarques (suite)...

c) Existe-t-il une relation entre la probabilité et l'écart-type ?



Les aires des surfaces hachurées sous les graphes des fonctions de densité valent respectivement 0.683, 0.954 et 0.997.

◊ Constatation :

Si  $X$  est une variable aléatoire de distribution normale, sa réalisation a une probabilité de

- 68 % de se trouver dans l'intervalle  $[\mu - \sigma, \mu + \sigma]$ ;
- 95 % de se trouver dans l'intervalle  $[\mu - 2\sigma, \mu + 2\sigma]$ ;
- 99.7 % de se trouver dans l'intervalle  $[\mu - 3\sigma, \mu + 3\sigma]$ .

Et alors...

supposons que  $X$  soit une variable aléatoire de distribution normale. Pour toutes les valeurs prises par l'espérance  $\mu$  et la variance  $\sigma^2$ , les probabilités que la réalisation de  $X$  appartienne aux intervalles centrés en  $\mu$  et de rayons  $\sigma$ ,  $2\sigma$  et  $3\sigma$  sont toutes constantes, à savoir 68 %, 95 % et 99.7 %.

Ces probabilités constantes nous conduisent à introduire une distribution normale particulière à partir de laquelle une probabilité d'intervalle d'une variable aléatoire normale quelconque pourra être calculée.

### Distribution normale de référence :

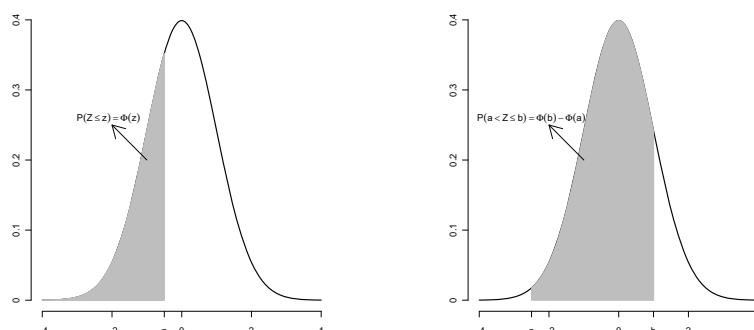
la loi normale centrée réduite  $\mathcal{N}(0, 1)$  de fonction de densité

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}, \quad -\infty < u < \infty,$$

et de fonction de répartition

$$\Phi(z) = \int_{-\infty}^z \varphi(u) du, \quad -\infty < z < \infty,$$

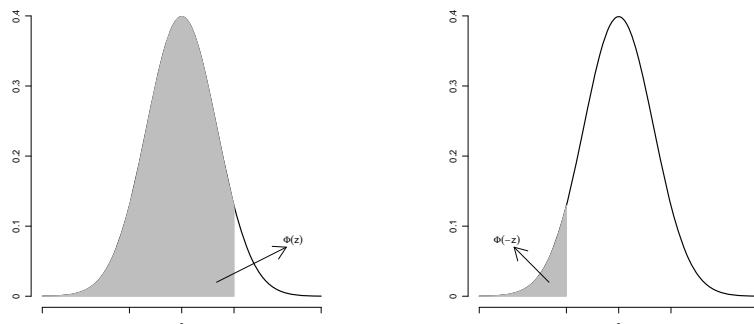
est la distribution de référence à partir de laquelle toute probabilité d'intervalle d'une variable aléatoire normale quelconque pourra être calculée.



Probabilités et fonction de répartition  $\Phi$ .

### Propriétés

- i)  $\varphi$  est une fonction symétrique;
- ii)  $\Phi(-z) = 1 - \Phi(z)$  (par symétrie);



Fonction de répartition  $\Phi$  appliquée à  $z$  et à  $-z$ .

Remarque :

le calcul de la fonction  $\Phi(z)$  pour un argument  $z$  donné est compliqué.

Toutefois...

des tables indiquent les valeurs  $\Phi(z)$  pour des arguments  $z$  non-négatifs. Pour les arguments négatifs, il convient d'utiliser la propriété ii). Les calculettes, certains tableurs et les logiciels de statistique disposent d'une fonction permettant le calcul de  $\Phi(z)$ .

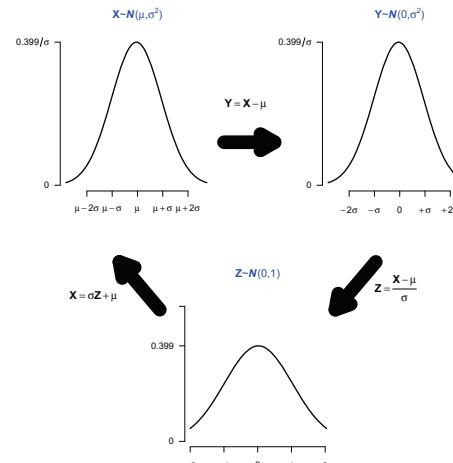
### Propriétés (suite)

iii) soit  $X$  une variable aléatoire normale d'espérance  $\mu$  et de variance  $\sigma^2$ . La variable aléatoire

$$Z = g(X) = \frac{X - \mu}{\sigma}$$

est de distribution normale centrée réduite. Ainsi,

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(\underbrace{\sigma Z + \mu}_{X} \leq x) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right); \end{aligned}$$



Transformation d'une variable aléatoire normale de paramètres  $\mu$  et  $\sigma^2$

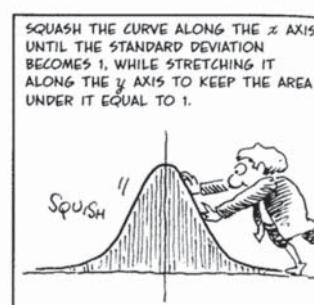
en une variable aléatoire normale centrée réduite.

Propriétés (*suite*) : additivité de la loi normale

- iv) soient  $X_1$  et  $X_2$  deux variables aléatoires indépendantes de distribution normale de paramètres respectifs  $(\mu_1, \sigma_1^2)$  et  $(\mu_2, \sigma_2^2)$ . Leur somme  $X_1 + X_2$  est une variable aléatoire de distribution normale de paramètres  $(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ .

Ainsi, la loi normale est dite **additive** par rapport à ses paramètres.

Cette propriété peut être généralisée à  $n$  variables aléatoires indépendantes de distribution normale.



Approximation normale d'une répartition binomiale établie en 1733 par A. de Moivre pour  $p = 1/2$  puis généralisée en 1812 par P.-S. de Laplace à toute valeur  $0 \leq p \leq 1$ .

### Application de la loi normale : approximation de la loi binomiale

Soit  $X$  une variable aléatoire de distribution binomiale de paramètres  $n$  et  $p$ . Si  $np(1-p)$  est grand (à nouveau pas facile d'être précis !!!), la probabilité

$$P\left(a < \frac{X - np}{\sqrt{np(1-p)}} \leq b\right)$$

peut être approchée par  $\Phi(b) - \Phi(a)$ , avec  $a < b$ .

En règle générale, l'approximation est satisfaisante dès que

$$np(1-p) > 10.$$

| loi de $X$    | densité $f_X(u)$                                                       | $\mathcal{H}$      | paramètre(s)                                    | $E(X)$      | $\text{Var}(X)$ |
|---------------|------------------------------------------------------------------------|--------------------|-------------------------------------------------|-------------|-----------------|
| uniforme      | $\frac{1}{b-a}$                                                        | $u \in [a, b]$     | $a, b \in \mathbb{R}, a < b$                    | $(a+b)/2$   | $(b-a)^2/12$    |
| exponentielle | $\lambda e^{-\lambda u}$                                               | $u \geq 0$         | $\lambda > 0$                                   | $1/\lambda$ | $1/\lambda^2$   |
| normale       | $\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}}$ | $u \in \mathbb{R}$ | $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ | $\mu$       | $\sigma^2$      |

Résumé des caractéristiques des lois continues usuelles.



## EXERCICES : DISTRIBUTIONS USUELLES

### Exercice 1

Un épicer reçoit un lot de pommes d'une centrale de distribution et constate que 25% des pommes livrées sont avariées. Il charge un employé de préparer des emballages de cinq pommes chacun. Celui-ci, fort négligent, ne se donne pas la peine de trier les fruits et de jeter les pommes avariées.

Désignons par  $X$  le nombre de pommes avariées contenues dans un emballage.

- Déterminer la loi de probabilité de  $X$  et ses paramètres.
- Tout client qui trouve dans l'emballage qu'il achète deux fruits avariés ou plus revient au magasin se plaindre. Calculer la probabilité qu'un client donné se plaigne auprès de l'épicier.

**Solutions :** a)  $X \sim \mathcal{B}(5, 1/4)$  b) 0.367.

### Exercice 2

Des études ont montré qu'en moyenne 10% des appareils électroniques d'une certaine marque tombent en panne durant la période de garantie. Un marchand a vendu 10 appareils de ce type. Désignons par  $X$  le nombre d'appareils qu'il devra remplacer durant la période de garantie.

- Donner la loi de probabilité de  $X$  et ses paramètres.
- Déterminer la probabilité que le marchand doive remplacer plus de deux appareils.
- Calculer l'espérance mathématique de  $X$ .

**Solutions :** a)  $X \sim \mathcal{B}(10, 0.1)$  b) 0.07 c) 1.

### Exercice 3

Dix-huit échantillons d'air ont été soigneusement prélevés dans de petits ballons et ont ensuite été analysés. Selon des expériences antérieures, chaque ballon a une probabilité de 10% de contenir une molécule très rare. On suppose que la présence de la molécule est indépendante de l'échantillon. Calculer la probabilité qu'au moins quatre ballons contiennent la molécule.

**Solution :** 0.098.

### Exercice 4

Pour vérifier le bon fonctionnement d'un procédé d'usinage, des échantillons de 20 pièces sont prélevés dans la production. Les prélèvements s'effectuent chaque heure au hasard. Des analyses ont montré qu'un pour-cent des pièces examinées sont défectueuses. Désignons par  $X$  le nombre de pièces défectueuses contenues dans l'échantillon,  $\mu_X$  et  $\sigma_X$  son espérance et son écart-type. Les responsables de la qualité ont décidé d'intervenir pour rectifier le procédé si  $X$  dépasse  $\mu_X + 3 \cdot \sigma_X$ . Cette méthode est utilisée dans le contrôle d'un procédé d'usinage.

- Déterminer la probabilité qu'il faille rectifier le procédé.
- Si la proportion de pièces défectueuses passe à 4%, calculer la probabilité que plus d'une pièce de l'échantillon soient défectueuses.

c) Quelle(s) hypothèse(s) faites-vous pour calculer ces probabilités ?

**Solutions :** a) 0.0169 b) 0.1897 c) indépendance dans la qualité des pièces; toutes les pièces ont la même probabilité d'être défectueuses.

### Exercice 5

Chaque matin, du lundi au vendredi, un professeur se rend en automobile à la HEIG-VD. Sur son trajet se trouve un carrefour, souvent surchargé, où la probabilité que le feu de signalisation soit vert vaut 0.2. On admet que l'état du feu de signalisation (rouge ou vert) représente chaque matin une expérience aléatoire indépendante.

- Calculer la probabilité que jeudi soit le premier matin de la semaine où le professeur tombe sur le feu vert.
- Déterminer la probabilité que le pauvre professeur, fou de rage, ne tombe pas sur le feu vert dix fois de suite en ne considérant que les jours ouvrables de la semaine du lundi au vendredi.

**Solutions :** a) 0.102 b) 0.107.

### Exercice 6

Avant le début d'un tournoi de basketball, Ray avait réussi approximativement le 70% de ses essais à 3 points pendant sa carrière. On suppose que les essais à 3 points de Ray sont indépendants et que chacun d'eux a une probabilité 0.7 de se concrétiser.

- Déterminer la probabilité que Ray inscrive au plus 2 paniers à 3 points en 6 essais.
- Calculer la probabilité que Ray inscrive son premier panier à 3 points à son 5ème essai.
- Pendant le tournoi, Ray a seulement inscrit 2 paniers à 3 points sur 6 essais. Son coach pense qu'il ne s'agit que d'une "poisse" passagère. En se basant sur la probabilité calculée en a) et sur son habileté habituelle, pensez-vous que la prestation de Ray dans le tournoi s'explique uniquement par cette "poisse" passagère ?

**Solutions :** a) 0.0705 b) 0.0057 c) en théorie, Ray a une probabilité de 0.0705 d'inscrire au plus deux paniers à 3 points sur 6 essais. En supposant indépendance entre les matches et même probabilité de réussir au plus deux paniers à 3 points sur 6 essais, sa performance au tournoi devrait se reproduire en moyenne tous les  $1/0.0705 \approx 14$  matches (espérance d'une distribution géométrique de paramètre  $p = 0.0705$ ). Or, Ray a certes participé à un tournoi mais a certainement disputé moins de 14 matches. On peut ainsi douter que sa performance soit due uniquement au hasard.

### Exercice 7

Le nombre de messages qui arrivent à un canal de communication dans un intervalle de  $t$  secondes peut être représenté par un processus de Poisson de paramètre 0.3. Calculer la probabilité qu'exactement trois messages arrivent en 10 secondes.

**Solution :** 0.224.

### Exercice 8

Le nombre de jobs reçus par une certaine imprimante peut être décrit par un processus de Poisson. Les jobs arrivent au rythme moyen d'un job toutes les 6 minutes.

- a) Calculer la probabilité qu'au moins 4 jobs arrivent entre 10h00 et 10h30.
- b) Déterminer la probabilité qu'aucun job arrive entre 12h00 et 12h15.
- c) Calculer la probabilité que le temps éoulé entre l'arrivée du 20ème et du 21ème job dépasse un quart d'heure.

**Solutions :** a) 0.735 b) 0.082 c) 0.082.

### Exercice 9

Depuis l'installation de nombreux radars fixes, les contrôles de vitesse sur les autoroutes suisses s'intensifient. En moyenne, un véhicule dépasse toutes les deux minutes la vitesse autorisée sur la ceinture lausannoise un matin donné. On suppose que les excès de vitesse peuvent être modélisés par un processus de Poisson.

- a) Déterminer la probabilité qu'un seul excès de vitesse arrive en 6 minutes.
- b) Calculer la probabilité qu'au maximum deux excès de vitesse sont enregistrés en une minute.

**Solutions :** a)  $3e^{-3}$  b)  $13/8e^{-1/2}$ .

### Exercice 10

Wayne Gretzky est l'un des meilleurs compteurs qu'a connu la Ligue Nord-Américaine de Hockey-sur-Glace (National Hockey League). Le nombre de points par match réalisés par un joueur de hockey est la somme des buts inscrits et des assists. Pour sa première saison en NHL (1979–1980), le nombre de points réalisés par Gretzky peut être représenté par une variable aléatoire  $X$  issue d'une distribution de Poisson de paramètre  $\lambda = 1.734$ .

- a) Déterminer la probabilité que Gretzky inscrive exactement un point.
- b) La saison 1979–1980 comprenait 79 parties. Estimer le nombre de matches dans lesquels Gretzky a réalisé exactement un point.

**Solutions :** a) 0.306 b) 24.19.

### Exercice 11

On suppose que le nombre d'erreurs par page dans un certain livre de probabilités suit une loi de Poisson de paramètre  $\lambda = 1/2$ . Calculer la probabilité de découvrir au moins une erreur dans une page de ce livre.

**Solution :**  $1 - e^{-1/2}$ .

### Exercice 12

Le nombre de gagnants à la loterie de Kaamelott peut être représenté par une variable aléatoire issue d'une distribution de Poisson de paramètre 1.23.

- a) Déterminer la probabilité que le jackpot de la loterie soit décroché par une seule personne.
- b) Le jackpot de la loterie vient d'être décroché mais on ignore le nombre de gagnants. En disposant de cette information, calculer la probabilité qu'il n'y ait qu'un seul gagnant.

**Solutions :** a) 0.36 b) 0.51.

### Exercice 13

Un réseau de neurones, outil statistique, est utilisé pour reconnaître les caractères, par exemple les lettres de l'alphabet ou les chiffres arabes, dans un très long texte manuscrit. On approche le nombre de caractères interprétés incorrectement par le réseau de neurones par une variable aléatoire issue d'une distribution de Poisson d'espérance 5.

- a) Calculer la probabilité qu'au plus 2 caractères d'un très long texte donné soient transcrits incorrectement par le réseau de neurones.
- b) En sachant que le réseau de neurones a déjà interprété incorrectement au moins un caractère d'un très long texte donné, déterminer la probabilité qu'exactement 2 caractères soient transcrits incorrectement.

**Solutions :** a) 0.125 b) 0.085.

### Exercice 14

Donald Knuth<sup>2</sup>, dans son "Art of Computer Programming", s'engage à verser deux dollars au premier lecteur qui lui signale une erreur dans un de ses livres. Pour simplifier, on suppose que le tome 2 contient une erreur et qu'un lecteur assidu a une chance sur 100'000 de la trouver et de la signaler. Si ce livre est lu par 500'000 lecteurs assidus, calculer la probabilité que Donald Knuth doive payer deux dollars.

**Solution :** désignons par  $X$  la variable aléatoire qui indique le nombre de lecteurs parmi les 500'000 assidus qui peuvent découvrir l'erreur. En supposant indépendance et équiprobabilité,  $X \sim \mathcal{B}(n = 500'000, p = 1/100'000)$ . Comme  $n$  est très grand,  $p$  très faible et  $np = 5$  est inférieur à 10, on peut utiliser l'approximation de la distribution binomiale par la distribution de Poisson. Ainsi,  $P(X > 0) = 1 - P(X = 0) \approx 1 - e^{-5} \approx 0.993$ .

### Exercice 15

Le courant électrique en milliampère passant dans un certain fil de cuivre peut être décrit par une variable aléatoire  $X$  de fonction de densité

$$f_X(u) = \begin{cases} 0.05 & \text{si } 0 \leq u \leq 20, \\ 0 & \text{sinon.} \end{cases}$$

- a) Déterminer la loi de probabilité de  $X$  et ses paramètres.
- b) Calculer le courant espéré passant dans le fil.

**Solutions :** a)  $X \sim \mathcal{U}(0, 20)$  b) 10.

### Exercice 16

La durée de vie en jours (temps éoulé avant un crash) d'un serveur Apache sous Linux peut être modélisée par une variable aléatoire issue d'une distribution exponentielle. La durée de vie espérée du serveur est de 200 jours.

<sup>2</sup>Mathématicien et informaticien américain, créateur du système typographique TeX. Cette création devait l'occuper pendant son semestre sabbatique de 1978. Il y consacra finalement environ 10 ans.

- a) Déterminer la probabilité que le serveur fonctionne pendant au moins 100 jours.  
b) En sachant que le serveur fonctionne toujours après 100 jours, déterminer la probabilité qu'il tienne encore au moins 50 jours supplémentaires.

**Solutions :** a)  $e^{-1/2}$  b)  $e^{-1/4}$ .

### Exercice 17

On suppose que le temps d'attente entre l'arrivée de deux messages électroniques peut être modélisé par une variable aléatoire issue d'une distribution exponentielle. Gaston a constaté que le temps d'attente moyen entre l'arrivée de deux messages est de deux heures.

- a) Déterminer la probabilité que Gaston doive attendre plus de 20 minutes pour recevoir son premier message électronique de la journée.  
b) Calculer la probabilité que Gaston ne reçoive aucun message électronique durant une période de deux heures.

**Solutions :** a)  $e^{-1/6}$  b)  $e^{-1}$ .

### Exercice 18

Le temps d'attente d'un taxi libre devant les Éditions Bidule peut être représenté par une variable aléatoire issue d'une distribution exponentielle. En moyenne, un taxi libre passe devant les Éditions Bidule toutes les trois minutes.

- a) Calculer la probabilité que Gaston attende plus de 10 minutes avant qu'un taxi libre s'arrête devant les Éditions Bidule.  
b) Déterminer la probabilité que Gaston trouve un taxi pendant les 30 premières secondes.

**Solutions :** a)  $e^{-10/3}$  b)  $1 - e^{-1/6}$ .

### Exercice 19

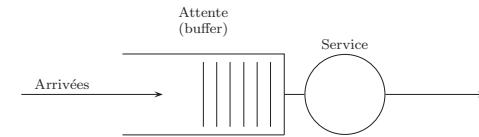
Considérons une variable aléatoire  $X$  issue d'une distribution exponentielle d'espérance  $1/\lambda$ .

- a) Calculer la probabilité  $P(X > 1/\lambda)$ .  
b) Déterminer la probabilité  $P(X > 2/\lambda)$ .  
c) Trouver une formule en fonction de  $k$  et de  $\lambda$  pour calculer la probabilité  $P(X > k/\lambda)$ ,  $k = 0, 1, 2, \dots$

**Solutions :** a)  $e^{-1}$  b)  $e^{-2}$  c)  $e^{-k}$ ,  $k = 0, 1, 2, \dots$

### Exercice 20

Un serveur de calcul reçoit des requêtes à fréquence régulière. Elles sont traitées indépendamment les unes après les autres dans leur ordre d'arrivée. On suppose que le nombre de requêtes qui arrivent au serveur peut être modélisé par un processus de Poisson. Les requêtes sont envoyées vers le serveur à un rythme moyen de 7 requêtes par seconde. Le temps de service d'une requête peut être représenté par une variable aléatoire issue d'une distribution exponentielle. En moyenne, le serveur est capable de traiter une requête en 0.1 seconde. On admet que le temps de service est indépendant du processus d'arrivée.



- a) Calculer la probabilité qu'en l'espace de deux secondes, au moins cinq requêtes arrivent au serveur.  
b) Déterminer la probabilité que la durée de traitement d'une requête par le serveur prenne au moins deux secondes.

**Solutions :** a)  $0.9982 \approx 1$  b)  $e^{-20} \approx 0$ .

### Exercice 21

On suppose que le temps en heures pour réparer une automobile peut être décrit par une variable aléatoire issue d'une distribution exponentielle de paramètre  $1/2$ .

- a) Calculer la probabilité que le temps de réparation d'une voiture excède 2 heures.  
b) La réparation d'une automobile a déjà dépassé 9 heures. Déterminer la probabilité qu'elle prenne au moins 10 heures.

**Solutions :** a)  $e^{-1}$  b)  $e^{-1/2}$ .

### Exercice 22

Considérons une variable aléatoire  $X$  issue d'une distribution normale de paramètres  $\mu = 3$  et  $\sigma^2 = 9$ . Calculer les probabilités suivantes :

- a)  $P(2 < X \leq 5)$ ;  
b)  $P(X > 0)$ .

**Solutions :** a) 0.3779 b) 0.8413.

### Exercice 23

Désignons par  $X$  une variable aléatoire issue d'une distribution normale d'espérance  $\mu$  et de variance  $\sigma^2$ . Calculer la probabilité  $P(\mu - \sigma < X \leq \mu + 2\sigma)$ .

**Solution :** 0.81859.

### Exercice 24

Des études cliniques ont montré que le temps de réaction d'un conducteur automobile à un stimulus visuel peut être décrit par une variable aléatoire issue d'une distribution normale d'espérance 0.4 seconde et d'écart-type 0.05 seconde.

- a) Calculer la probabilité que le temps de réaction soit compris entre 0.4 et 0.5 seconde.

b) Déterminer le temps de réaction  $x$  tel que  $P(X > x) = 0.8$ .

**Solutions :** a) 0.4772 b) 0.358.

### Exercice 25

Considérons une variable aléatoire  $X$  issue d'une distribution normale d'espérance 60 et d'écart-type 15.

- Calculer la probabilité que la variable aléatoire  $X$  soit supérieure ou égale à 30.
- Déterminer le nombre réel  $x_{0.1}$  tel que  $P(X \leq x_{0.1}) = 0.1$ . Ce nombre est appelée le 10%-quantile de la distribution de la variable aléatoire  $X$ .
- Déterminer l'écart-type  $\sigma$  de  $X$  tel que  $P(52.16 < X < 67.84) = 0.95$ . L'espérance de  $X$  reste 60.

**Solutions :** a)  $\Phi(2) = 0.97725$  b) 40.8 c)  $\sigma = 4$ .

### Exercice 26

Le petit Nicolas glisse dans son cartable son sac de billes. Le diamètre d'une bille en millimètres ( $mm$ ) peut être représenté par une variable aléatoire issue d'une distribution normale d'espérance 12 et d'écart-type 1. Le sac contient un trou qui peut laisser échapper dans le cartable les billes qui ont un diamètre inférieur à 10 mm.

- Calculer la probabilité qu'une bille puisse sortir du sac.
- Déterminer le nombre réel  $x_{0.1}$  tel que  $P(X \leq x_{0.1}) = 0.1$ .
- Le sac contient 20 billes. Désignons par  $Y$  le nombre de billes qui peuvent sortir du sac et tomber dans le cartable. Déterminer la distribution de  $Y$  ainsi que son (ses) paramètre(s).
- Calculer la probabilité que parmi les 20 billes que contient le sac deux billes au minimum s'en échappent.

**Solutions :** a)  $1 - \Phi(2) = 0.02275$  b) 10.72 c)  $Y \sim \mathcal{B}(20, 0.02275)$  d) 0.075.

### Exercice 27

Un ingénieur en informatique vient de décrocher un projet. Pour mener à bien son entreprise, il se propose de décomposer le projet en deux tâches notées  $A$  et  $B$  de telle sorte qu'elles débutent simultanément le 1er juin, soient exécutées en parallèle et se terminent toutes deux au plus tard le 30 juin, date imposée par le client, mais pas forcément en même temps. L'ingénieur suppose que les temps pour réaliser les deux tâches peuvent être modélisés par les variables aléatoires  $X_A$  et  $X_B$  issues de distributions normales dont les temps espérés et les écarts-type sont respectivement 25 jours et 5 jours pour la tâche  $A$ , 26 jours et 4 jours pour la tâche  $B$ . On suppose que les tâches  $A$  et  $B$  sont statistiquement indépendantes et qu'elles débutent selon le planning le 1er juin. Calculer la probabilité que le projet se termine en temps voulu, à savoir au plus tard le 30 juin.

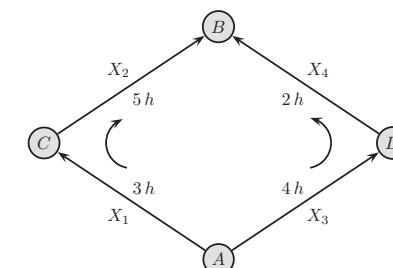
**Solution :** 0.292.

### Exercice 28

Dans sa tournée, un voyageur de commerce doit se rendre de la ville  $A$  à la ville  $B$ . Il dispose de deux itinéraires : le premier en passant par la ville  $C$  et le second par la ville  $D$ . Aucune liaison directe entre  $A$  et  $B$  existe. Les temps en heures ( $h$ ) que passe sur la route le voyageur de commerce pour se déplacer de  $A$  à  $B$  via les villes  $C$  et  $D$  peuvent être représentés par des variables aléatoires indépendantes  $X_1, X_2, X_3$  et  $X_4$  définies par

- $X_1$  : "durée du trajet entre les villes  $A$  et  $C$ ";
- $X_2$  : "durée du trajet entre les villes  $C$  et  $B$ ";
- $X_3$  : "durée du trajet entre les villes  $A$  et  $D$ ";
- $X_4$  : "durée du trajet entre les villes  $D$  et  $B$ ".

On suppose que ces variables aléatoires sont toutes issues d'une distribution normale. Les temps espérés pour se déplacer d'une ville à l'autre se trouvent dans la figure ci-dessous et le coefficient de variation de chacune de ces variables aléatoires vaut 0.2.



- Le coefficient de variation  $\delta_X$  d'une variable aléatoire  $X$  est donné par  $\frac{\sigma_X}{\mu_X}$ . Calculer les écarts-type des variables  $X_1, X_2, X_3$  et  $X_4$ .
- Déterminer  $x$  tel que la probabilité que la durée du trajet  $X_3$  entre les villes  $A$  et  $D$  soit supérieure à  $x$  soit égale à 0.95.
- Calculer la probabilité que le trajet entre la ville  $A$  et  $B$  via  $C$  dure moins de 9 heures.
- Déterminer la probabilité que la durée du trajet entre  $A$  et  $B$  via  $C$  soit plus courte que celle via  $D$  en considérant la variable aléatoire  $T = T_1 - T_2$  où  $T_1$  représente la durée du trajet via  $C$  et  $T_2$  celle du trajet via  $D$ .

**Solutions :** a)  $\sigma_{X_1} = 0.6$ ,  $\sigma_{X_2} = 1$ ,  $\sigma_{X_3} = 0.8$ ,  $\sigma_{X_4} = 0.4$  b) 2.69 c) 0.805 d) 0.087.

# Chapitre 7

## Variables aléatoires simultanées

### 7.1 Introduction

Jusqu'à présent nous avons avant tout traité séparément des distributions de variables aléatoires. Toutefois, il arrive fréquemment qu'on s'intéresse à des événements relatifs à deux variables **simultanément**, voire même à plus de deux variables. Ainsi, on définit une **distribution simultanée (ou conjointe)**.

Dans ce chapitre, les concepts associés aux variables aléatoires simultanées seront brièvement introduits par l'intermédiaire d'exemples. Deux cas sont à considérer :

- ◊ **cas discret**;
- ◊ **cas continu**.

### Contenu

- 7.1 Introduction
- 7.2 Cas discret
- 7.3 Cas continu
- 7.4 Somme de variables aléatoires
- 7.5 Espérance mathématique d'une somme de variables aléatoires
- 7.6 Propriétés de la covariance, de la corrélation et autres

### 7.2 Cas discret

#### Exemple 1 :

on jette une fois deux dés non biaisés simultanément. Désignons par  $X$  et  $Y$  les variables aléatoires :

- $X$  : nombre de faces paires;  
 $Y$  : nombre de faces inférieures ou égales à 3.

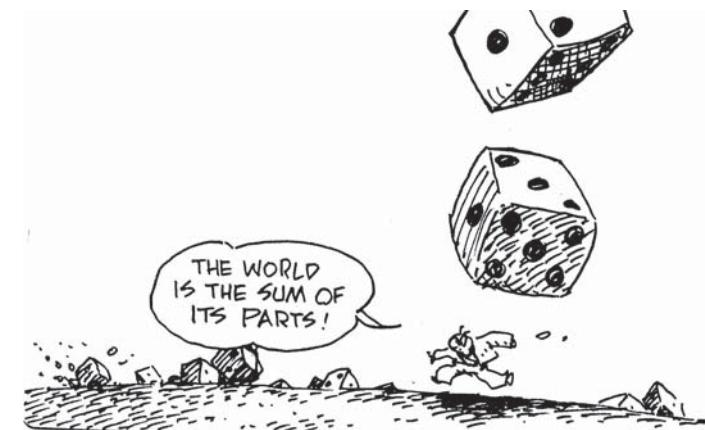
Ainsi, deux variables aléatoires sont définies sur le **même** ensemble fondamental  $\Omega$ , i.e sur les 36 résultats possibles obtenus en jetant les deux dés.

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| (6,1) | (6,2) | (6,3) | (6,4) | (6,5) | (6,6) |
| (5,1) | (5,2) | (5,3) | (5,4) | (5,5) | (5,6) |
| (4,1) | (4,2) | (4,3) | (4,4) | (4,5) | (4,6) |
| (3,1) | (3,2) | (3,3) | (3,4) | (3,5) | (3,6) |
| (2,1) | (2,2) | (2,3) | (2,4) | (2,5) | (2,6) |
| (1,1) | (1,2) | (1,3) | (1,4) | (1,5) | (1,6) |

**[1er dé]**

**[2ème dé]**

Résultats possibles du jet simultané de deux dés et application des variables  $X$  et  $Y$ .



No comment !

### Exemple 1 (suite) :

- Les ensembles des valeurs prises par  $X$  et  $Y$  sont
- $$\mathcal{H}_X = \{0, 1, 2\} \quad \text{et} \quad \mathcal{H}_Y = \{0, 1, 2\}.$$

- Objectif :

calculer les probabilités

$$P(X = x_i, Y = y_j) \quad \text{avec} \quad x_i = i \quad \text{et} \quad y_j = j, \quad i, j = 0, 1, 2.$$

Autrement dit, déterminer la **loi de probabilité simultanée (ou conjointe)** de  $X$  et  $Y$ .

### Exemple 1 (suite) :

- La loi de probabilité simultanée de  $X$  et  $Y$  peut être donnée sous la forme d'un tableau :

|         | $X = 0$ | $X = 1$ | $X = 2$ |
|---------|---------|---------|---------|
| $Y = 0$ | 1/36    | 4/36    | 4/36    |
| $Y = 1$ | 4/36    | 10/36   | 4/36    |
| $Y = 2$ | 4/36    | 4/36    | 1/36    |

Exemple de calcul :

$$P(X = 1, Y = 0) = P(\{(6, 5), (5, 4), (5, 6), (4, 5)\}) = 4/36.$$

## Exemple 1 (suite) :

- Sur les marges du tableau, on inscrit les lois de probabilité de  $X$  seul et de  $Y$  seul. Ces lois sont appelées les **lois marginales** de  $X$  et de  $Y$ .

|            | $X = 0$ | $X = 1$ | $X = 2$ | $P(Y = y)$ |
|------------|---------|---------|---------|------------|
| $Y = 0$    | 1/36    | 4/36    | 4/36    | 9/36       |
| $Y = 1$    | 4/36    | 10/36   | 4/36    | 18/36      |
| $Y = 2$    | 4/36    | 4/36    | 1/36    | 9/36       |
| $P(X = x)$ | 9/36    | 18/36   | 9/36    | 1          |

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j) \rightarrow X \sim \mathcal{B}(2, 1/2).$$

$$P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) \rightarrow Y \sim \mathcal{B}(2, 1/2).$$

## Exemple 1 (suite) :

- Calcul de la covariance de  $X$  et  $Y$  :
  - loi de probabilité de  $Z = X \cdot Y$  :

| $Z = z$    | 0     | 1     | 2    | 4    |
|------------|-------|-------|------|------|
| $P(Z = z)$ | 17/36 | 10/36 | 8/36 | 1/36 |

◊ espérance mathématique de  $Z$  :  $E(Z) = 5/6$ ;

◊ comme  $X \sim \mathcal{B}(2, 1/2)$  et  $Y \sim \mathcal{B}(2, 1/2)$ , on a

$$E(X) = 1 \text{ et } E(Y) = 1.$$

Ainsi,

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y) = 5/6 - 1 = -1/6.$$

## Exemple 1 (suite) :

- Une première question se pose :

existe-t-il un lien entre  $X$  et  $Y$  ?

→ la **covariance** est une mesure du lien, de la “liaison”, existant entre deux variables aléatoires  $X$  et  $Y$ .

**Définition 1** La **covariance** des variables aléatoires  $X$  et  $Y$  est le nombre réel défini par

$$\text{Cov}(X, Y) = E([X - E(X)] \cdot [Y - E(Y)]). \quad (1)$$

**Note** : on peut écrire  $\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y)$ .

## Exemple 1 (suite) :

- Il est aussi possible d'utiliser la **corrélation** entre les variables aléatoires  $X$  et  $Y$ . Il s'agit d'une mesure du degré de **linéarité** entre les deux variables.

**Définition 2** La **corrélation** entre les variables aléatoires  $X$  et  $Y$  est définie par

$$\rho(X, Y) = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}. \quad (2)$$

**Note** : la corrélation est un nombre *sans dimension*; elle ne dépend pas des échelles choisies pour  $X$  et  $Y$ .

### Exemple 1 (suite) :

- Calcul de la corrélation de  $X$  et  $Y$  :

comme  $X \sim \mathcal{B}(2, 1/2)$  et  $Y \sim \mathcal{B}(2, 1/2)$ , on a  $\text{Var}(X) = 1/2$  et  $\text{Var}(Y) = 1/2$ . Ainsi,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = \frac{-1/6}{1/2} = -1/3.$$

- Une seconde question se pose :

les variables  $X$  et  $Y$  sont-elles indépendantes ?

### Exemple 1 (suite) :

- D'abord une définition...

**Définition 3** Les variables aléatoires discrètes  $X$  et  $Y$  sont **indépendantes** si

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j), \forall (i, j); \quad (3)$$

...puis une réponse :

dans notre exemple, les variables  $X$  et  $Y$  **ne** sont **pas** **indépendantes** car

$$\underbrace{P(X = 0, Y = 0)}_{1/36} \neq \underbrace{P(X = 0)}_{9/36} \cdot \underbrace{P(Y = 0)}_{9/36}.$$

Ce résultat corrobore-t-il notre intuition ?

### 7.3 Cas continu (brève approche)

#### Exemple 2 :

on choisit arbitrairement un point dans le carré unité  $[0, 1] \times [0, 1]$ . Désignons par  $X$  sa première composante et par  $Y$  sa seconde composante.

Ainsi, on s'intéresse à un couple de variables aléatoires continues  $(X, Y)$  défini dans le sous-ensemble  $[0, 1] \times [0, 1]$  du plan.

Les ensembles des valeurs prises par  $X$  et  $Y$  sont

$$\mathcal{H}_x = [0, 1] \quad \text{et} \quad \mathcal{H}_y = [0, 1].$$

### Exemple 2 (suite) :

- Objectif :

trouver une fonction  $f_{X, Y}(u, v) \geq 0$  telle que

$$\begin{aligned} P(X \leq x, Y \leq y) &:= F_{X, Y}(x, y) \\ &= \int_{-\infty}^x \int_{-\infty}^y f_{X, Y}(u, v) \, du \, dv. \end{aligned}$$

Autrement dit, déterminer la **fonction de densité conjointe (ou simultanée)** de  $X$  et  $Y$ . La fonction  $F_{X, Y}$  est appelée la **fonction de répartition** de  $(X, Y)$ .

### Exemple 2 (suite) :

- La fonction de densité conjointe qui convient au problème est

$$f_{X,Y}(u, v) = \begin{cases} 1 & \text{si } 0 \leq u \leq 1 \text{ et } 0 \leq v \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

- Par analogie au cas discret, les **fonctions de densité marginales** de  $X$  et de  $Y$  sont définies respectivement par

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v) dv \text{ et } f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y) du.$$

Dans notre exemple, elles valent

$$f_X(x) = \int_0^1 1 dv = 1 \text{ et } f_Y(y) = \int_0^1 1 du = 1,$$

avec  $0 \leq x \leq 1$  et  $0 \leq y \leq 1$ .

### Exemple 2 (suite) :

- Les variables  $X$  et  $Y$  sont-elles indépendantes ?

la définition de l'indépendance de  $X$  et  $Y$  dans le cas continu est la suivante :

**Définition 4** Les variables aléatoires continues  $X$  et  $Y$  sont **indépendantes** si

$$f_{X,Y}(u, v) = f_X(u) \cdot f_Y(v). \quad (4)$$

Dans notre exemple, les variables  $X$  et  $Y$  sont **indépendantes**.

Ce résultat corrobore-t-il notre intuition ?

### Exemple 2 (suite) :

- Calcul de la **covariance** de  $X$  et  $Y$  :

$$\diamond E(X \cdot Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u v f_{X,Y}(u, v) du dv = 1/4;$$

comme  $X \sim \mathcal{U}(0, 1)$  et  $Y \sim \mathcal{U}(0, 1)$ , on a

$$E(X) = 1/2 \text{ et } E(Y) = 1/2.$$

Ainsi,

$$\text{Cov}(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y) = 1/4 - 1/4 = 0.$$

- La **corrélation** des deux variables est

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} = 0.$$

### Remarques :

- la loi de probabilité de deux variables aléatoires discrètes doit vérifier la condition

$$\sum_i \sum_j P(X = x_i, Y = y_j) = 1.$$

De même, la fonction de densité conjointe de deux variables aléatoires continues  $X$  et  $Y$  doit remplir la condition

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv = 1;$$

- la loi de probabilité simultanée et la fonction de densité conjointe de deux variables aléatoires déterminent les lois marginales et les fonctions de densité marginales. La proposition réciproque n'est pas **toujours** vraie.

## 7.4 Somme de variables aléatoires

- On se propose de déterminer la distribution de la somme  $X + Y$  de deux variables aléatoires indépendantes  $X$  et  $Y$  (discrètes ou continues) en se basant sur leurs lois de probabilité **marginales** ou sur leurs fonctions de densité **marginales**.
- Si les variables  $X$  et  $Y$  sont dépendantes, il est souvent impossible d'obtenir la distribution de la somme de  $X$  et  $Y$  à l'aide de leurs lois de probabilité marginales ou de leurs fonctions de densité marginales. Pour cette raison, nous nous contenterons d'une petite remarque concernant le cas de dépendance.

a) Fonction de répartition de  $Z$  :

$$\begin{aligned}
 F_Z(z) &= F_{X+Y}(z) = P(X + Y \leq z) \\
 &= \int \int_{x+y \leq z} f_X(x) f_Y(y) dx dy \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{z-y} f_X(x) f_Y(y) dx dy \\
 &= \int_{-\infty}^{+\infty} \underbrace{\left( \int_{-\infty}^{z-y} f_X(x) dx \right)}_{F_X(z-y)} f_Y(y) dy \\
 &= \int_{-\infty}^{+\infty} F_X(z-y) f_Y(y) dy
 \end{aligned} \tag{5}$$

~ $\rightsquigarrow$  convolution des fonctions  $f_X$  et  $f_Y$ .

### 7.4.1 $X$ et $Y$ variables aléatoires indépendantes et continues

- Objectif :

déterminer la distribution de la somme  $Z$  de deux variables aléatoires continues et indépendantes  $X$  et  $Y$  de fonctions de densité respectives  $f_X$  et  $f_Y$ .

- On se propose de calculer :
  - la fonction de répartition  $F_Z$  de  $Z$ ;
  - la fonction de densité  $f_Z$  de  $Z$ .

b) Fonction de densité de  $Z$  :

$$\begin{aligned}
 f_Z(z) &= \frac{d}{dz} \int_{-\infty}^{+\infty} F_X(z-y) f_Y(y) dy \\
 &= \int_{-\infty}^{+\infty} \frac{d}{dz} F_X(z-y) f_Y(y) dy \\
 &= \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy
 \end{aligned} \tag{6}$$

~ $\rightsquigarrow$  convolution des fonctions  $f_X$  et  $f_Y$ .

**Exemple 3 :**

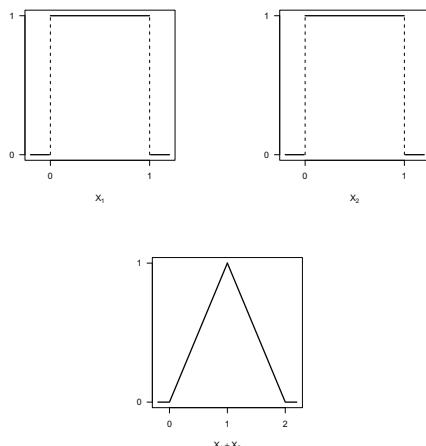
considérons deux variables aléatoires uniformes et indépendantes  $X \sim \mathcal{U}(0, 1)$  et  $Y \sim \mathcal{U}(0, 1)$ . On se propose de déterminer la fonction de densité de la variable aléatoire  $Z = X + Y$ .

- On a :

$$f_X(u) = \begin{cases} 1 & \text{si } 0 \leq u < 1, \\ 0 & \text{sinon} \end{cases} \quad \text{et} \quad f_Y(v) = \begin{cases} 1 & \text{si } 0 \leq v < 1, \\ 0 & \text{sinon.} \end{cases}$$

- En appliquant la formule (6), on obtient :

$$f_Z(z) = \int_0^1 f_X(z - y) \cdot 1 \, dy = \int_0^1 f_X(z - y) \, dy.$$



Fonctions de densité de deux variables aléatoires indépendantes et uniformes sur  $[0, 1]$  ainsi que la fonction de densité de leur somme.

**Exemple 3 (suite) :**

- Deux cas sont à considérer :

▷  $0 \leq z \leq 1$  :

$$f_Z(z) = \int_0^z dy = z;$$

▷  $1 \leq z \leq 2$  :

$$f_Z(z) = \int_{z-1}^1 dy = 2 - z.$$

- Ainsi,

$$f_Z(z) = \begin{cases} z & \text{si } 0 \leq z \leq 1, \\ 2 - z & \text{si } 1 \leq z \leq 2, \\ 0 & \text{sinon.} \end{cases}$$

**7.4.2  $X$  et  $Y$  variables aléatoires indépendantes et discrètes**

- Plutôt que de se lancer directement dans la recherche d'une formule générale donnant la distribution d'une somme de variables aléatoires discrètes et indépendantes, traitons l'exemple simple de la somme de deux variables aléatoires de Bernoulli.
- De Moivre et Gosset lancent indépendamment une pièce de monnaie identique et non biaisée. La variable aléatoire  $X$  prend la valeur 1 si la pièce lancée par de Moivre tombe sur pile (succès) et 0 si elle tombe sur face (échec). On attribue la valeur 1 à la variable aléatoire  $Y$  si la pièce jetée par Gosset tombe sur pile et 0 si elle tombe sur l'autre côté de la pièce.

↝ On cherche la distribution de  $Z = X + Y$  par convolution.

- Sans faire le moindre calcul, on sait que la variable  $Z = X + Y$  est issue d'une distribution binomiale de paramètres  $n = 2$  et  $p = 1/2$ . En effet,  $Z$  compte le nombre de succès de probabilité  $1/2$  lors du jet indépendant de deux pièces de monnaie identiques et non biaisées.
- On peut établir le même résultat de manière analytique. Remarquons d'abord que les réalisations de  $Z$  sont 0, 1 et 2. Comme l'événement  $\{Z = z\}$  est l'union des événements incompatibles  $\{X = z - y, Y = y\}$  pour  $y = 0, 1$ , on a

$$\begin{aligned} P(Z = z) &= \sum_{y=0}^z P(X = z - y, Y = y) \\ &= \sum_{y=0}^z P(X = z - y) \cdot P(Y = y). \end{aligned}$$

Remarques :

- la détermination par convolution de la loi de probabilité de la somme des résultats obtenus en lançant indépendamment deux pièces de monnaie est horriblement laborieuse et pénible en comparaison avec la résolution à l'aide d'un diagramme en arbre !
- si  $X \sim \mathcal{B}(n, p)$  et  $Y \sim \mathcal{B}(m, p)$  sont des variables aléatoires indépendantes, il est possible de montrer, par convolution, que  $X + Y$  est une variable aléatoire binomiale de paramètres  $n + m$  et  $p$ ;
- les formules pour obtenir la distribution de la somme  $Z$  de deux variables aléatoires dépendantes  $X$  et  $Y$  sont, respectivement,
  - cas continu :  $f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(z - y, y) dy$ ;
  - cas discret :  $P(Z = z) = \sum_{y=0}^z P(X = z - y, Y = y)$  en supposant que  $X$  et  $Y$  ne prennent que des valeurs positives.

- Plus précisément,

$$\begin{aligned} - P(Z = 0) &= P(X = 0, Y = 0) = 1/2 \cdot 1/2 = 1/4; \\ - P(Z = 1) &= \underbrace{P(X = 1, Y = 0)}_{1/4} + \underbrace{P(X = 0, Y = 1)}_{1/4} = 1/2; \\ - P(Z = 2) &= P(X = 1, Y = 1) = 1/2 \cdot 1/2 = 1/4. \end{aligned}$$

Autrement dit,

|            |     |     |     |
|------------|-----|-----|-----|
| $Z = z$    | 0   | 1   | 2   |
| $P(Z = z)$ | 1/4 | 1/2 | 1/4 |

## 7.5 Espérance mathématique d'une somme de variables aléatoires

Soient  $m$  variables aléatoires  $X_1, X_2, \dots, X_m$  définies sur le même ensemble fondamental  $\Omega$  et d'espérances finies. Alors,

$$E(X_1 + X_2 + \dots + X_m) = E(X_1) + E(X_2) + \dots + E(X_m).$$

Ecriture équivalente :

$$E\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m E(X_i).$$

## 7.6 Propriétés de la covariance, de la corrélation et autres

Soient  $a, b, c$  et  $d$  des constantes réelles et  $X, Y$  et  $Z$  trois variables aléatoires.

**Propriétés de la covariance :**

- i)  $\text{Cov}(X, Y) = \text{Cov}(Y, X);$
- ii)  $\text{Cov}(X, a) = 0;$
- iii)  $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y);$
- iv)  $\text{Cov}(aX + bY, cZ + d) = ac \text{Cov}(X, Z) + bc \text{Cov}(Y, Z);$
- v)  $\text{Cov}(X, X) = \text{Var}(X)$  d'où  
 $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2 \text{Cov}(X, Y).$

Quelques remarques concernant l'indépendance de deux variables :

- i) si  $X$  et  $Y$  sont deux variables aléatoires indépendantes alors

- $E(X \cdot Y) = E(X) \cdot E(Y);$
- $\text{Cov}(X, Y) = 0;$
- $\text{Corr}(X, Y) = 0;$
- $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y);$   
 cette propriété se généralise à  $m$  variables aléatoires indépendantes;

- ii)  $\text{Cov}(X, Y) = 0 \neq X$  et  $Y$  indépendantes.

**Propriétés de la corrélation :**

- i)  $-1 \leq \text{Corr}(X, Y) \leq 1;$
- ii)  $\text{Corr}(X, Y) = \text{Corr}(Y, X);$
- iii)  $\text{Corr}(X, X) = 1$  et  $\text{Corr}(X, -X) = -1;$
- iv)  $\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$  si  $a$  et  $c$  sont non nuls et de même signe.

**Remarque :**

une corrélation nulle n'implique pas qu'il n'existe pas de lien entre deux variables aléatoires. Elle indique seulement qu'il n'y a pas de relation linéaire entre ces deux variables.

*"Uncertainty is the only certainty there is, and knowing how to live with insecurity is the only security."*

John Allen Paulos



## EXERCICES : VARIABLES ALÉATOIRES SIMULTANÉES

### Exercice 1

Considérons les variables aléatoires  $X$  et  $Y$  dont la loi de probabilité simultanée est donnée dans le tableau ci-dessous.

| $Y \backslash X$ | -1  | 1   |
|------------------|-----|-----|
| -1               | 1/9 | 2/9 |
| 0                | 2/9 | 1/9 |
| 1                | 1/9 | 2/9 |

- a) Déterminer la covariance de  $X$  et  $Y$ .
- b) Montrer que les variables aléatoires  $X$  et  $Y$  ne sont pas indépendantes.

**Solutions :** a) 0   b)  $P(X = 1, Y = -1) \neq P(X = 1) \cdot P(Y = -1)$ .

### Exercice 2

Considérons les variables aléatoires  $X$  et  $Y$  dont la loi de probabilité simultanée est :

| $Y \backslash X$ | 1   | 2   |
|------------------|-----|-----|
| 1                | 1/8 | 1/8 |
| 2                | 1/4 | 1/2 |

- a) Déterminer les lois de probabilité marginales de  $X$  et  $Y$ .
- b) Les variables aléatoires  $X$  et  $Y$  sont-elles indépendantes ?
- c) Calculer la loi de probabilité de la variable  $W = X/Y$ .
- d) Déterminer la probabilité  $P(X > Y)$ .
- e) Calculer la probabilité conditionnelle  $P(X = 1 | Y = 1)$ .

**Solutions :** a)

| $Y \backslash X$ | 1   | 2   | $P(Y = y)$ |
|------------------|-----|-----|------------|
| 1                | 1/8 | 1/8 | 1/4        |
| 2                | 1/4 | 1/2 | 3/4        |
| $P(X = x)$       | 3/8 | 5/8 | 1          |

b) comme  $P(X = 1, Y = 1) \neq P(X = 1) \cdot P(Y = 1)$ , les variables aléatoires  $X$  et  $Y$  ne sont pas indépendantes   c)

| $W = w \backslash P(W = w)$ | 1/2 | 1   | 2   |
|-----------------------------|-----|-----|-----|
| 1/2                         | 1/4 | 5/8 | 1/8 |

d) 1/8   e) 1/2.

### Exercice 3

La loi de probabilité simultanée des variables aléatoires  $X$  et  $Y$  figure dans le tableau ci-dessous.

| $Y \backslash X$ | 0   | 1   | 2   | 3   |
|------------------|-----|-----|-----|-----|
| 0                | 0.2 | 0   | 0   | 0   |
| 1                | 0   | 0.1 | 0.1 | 0   |
| 2                | 0   | 0.1 | 0.1 | 0   |
| 3                | 0   | 0   | 0   | 0.4 |

- a) Représenter la loi de probabilité simultanée de  $X$  et  $Y$  dans un graphique de nuage de points.
- b) Déterminer la corrélation de  $X$  et  $Y$ .
- c) Interpréter le graphique tracé en a) et le résultat obtenu en b).

**Solution :** b)  $\text{Cov}(X, Y) = 1.26$ ,  $\text{Var}(X) = 1.36$ ,  $\text{Var}(Y) = 1.36$ ;  $\text{Corr}(X, Y) = 0.926$ .

### Exercice 4

Considérons les variables aléatoires  $X$  et  $Y$  dont la loi de probabilité simultanée est donnée dans le tableau ci-dessous.

| $Y \backslash X$ | 1   | 2   | 3   | 4   |
|------------------|-----|-----|-----|-----|
| 0                | 0   | 0   | 0   | 1/8 |
| 1                | 1/8 | 1/8 | 1/8 | 0   |
| 2                | 1/4 | 1/8 | 0   | 0   |
| 3                | 1/8 | 0   | 0   | 0   |

- a) Déterminer les lois de probabilité marginales de  $X$  et de  $Y$ .
- b) Calculer  $E(X)$ ,  $E(Y)$ ,  $E(X^2)$ ,  $E(Y^2)$ ,  $\text{Var}(X)$ ,  $\text{Var}(Y)$ ,  $E(X \cdot Y)$ ,  $\text{Cov}(X, Y)$  et  $\text{Corr}(X, Y)$ .

**Solution :** b)  $E(X) = 3/2$ ,  $E(Y) = 15/8$ ,  $E(X^2) = 3$ ,  $E(Y^2) = 37/8$ ,  $\text{Var}(X) = 3/4$ ,  $\text{Var}(Y) = 71/64$ ,  $E(X \cdot Y) = 17/8$ ,  $\text{Cov}(X, Y) = -11/16$ ,  $\text{Corr}(X, Y) = -11/\sqrt{213}$ .

### Exercice 5

Deux canaux de communication notés respectivement  $A$  et  $B$  ont été établis pour transmettre des bits à travers un certain réseau. Désignons par  $X$  la variable aléatoire qui prend la valeur 1 si un bit transmis par  $A$  est reçu sans erreur et 0 s'il est reçu avec erreur. De manière analogue,  $Y$  est la variable aléatoire de réalisation 1 si le bit transmis par  $B$  est reçu sans erreur et 0 s'il est reçu avec erreur. La loi de probabilité simultanée de  $X$  et  $Y$  figure dans le tableau ci-dessous.

| $Y \backslash X$ | 0      | 1      |
|------------------|--------|--------|
| 1                | 0.0648 | 0.7452 |
| 0                | 0.0152 | 0.1748 |

Les variables aléatoires  $X$  et  $Y$  sont-elles indépendantes ?

**Solution :**  $F_X(x) = (1 - e^{-\lambda x})^2$  si  $x > 0$  et 0 sinon.

**Solution :** les variables aléatoires  $X$  et  $Y$  sont indépendantes.

#### Exercice 6

Dans la transmission d'informations digitales, les bits sont soumis à trois types de distorsion : haute, modérée et faible avec probabilités respectives 0.01, 0.04 et 0.95. On suppose que trois bits viennent d'être transmis et que la distorsion de chaque bit est indépendante de celle des deux autres. Calculer la probabilité que parmi les trois bits, exactement deux d'entre eux sont transmis avec haute distorsion et un avec faible distorsion.

**Solution :** 0.000285.

#### Exercice 7

Considérons trois variables aléatoires discrètes  $X$ ,  $Y$  et  $Z$  dont la loi de probabilité simultanée figure dans le tableau ci-dessous.

| $X = x$ | $Y = y$ | $Z = z$ | $P(X = x, Y = y, Z = z)$ |
|---------|---------|---------|--------------------------|
| 1       | 2       | 3       | 0.2                      |
| 2       | 1       | 1       | 0.1                      |
| 2       | 2       | 1       | 0.45                     |
| 2       | 3       | 2       | 0.25                     |

- a) Déterminer la loi de probabilité de la variable aléatoire  $W = XY + YZ + ZX$ .
- b) Calculer l'espérance de  $W$ .

**Solutions :** a)

$$\begin{array}{c|cccc} W = w & 5 & 8 & 11 & 16 \\ \hline P(W = w) & 0.1 & 0.45 & 0.2 & 0.25 \end{array}$$

b)  $E(W) = 10.3$ .

#### Exercice 8

Soient  $X_1$ ,  $X_2$  et  $X_3$  trois variables aléatoires d'espérances respectives 0 et de variances respectives 1. Si les variables sont deux à deux non corrélées, autrement dit si  $\text{Corr}(X_i, X_j) = 0$ ,  $\forall i, j = 1, 2, 3$  avec  $i \neq j$ , calculer  $\text{Corr}(X_1 + X_2, X_2 + X_3)$  en utilisant les propriétés de la corrélation.

**Solution :** 1/2.

#### Exercice 9

On suppose que le percolateur de la cafétéria de la HEIG-VD est composé de deux groupes indépendants. La durée de fonctionnement de chacun de ces groupes peut être représenté par une variable aléatoire  $X_i$ ,  $i = 1, 2$ , issue d'une distribution exponentielle de paramètre  $\lambda$ . Le service est assuré seulement si les deux groupes sont en état de marche. Désignons par  $X$  la durée de vie du percolateur. Déterminer la fonction de répartition de  $X$ , à savoir  $F_X(x) = P(X_1 \leq x \text{ et } X_2 \leq x)$  en fonction de  $\lambda$ .

## Chapitre 8

### Le théorème central limite

## Contenu

#### 8.1 Introduction

#### 8.2 Illustrations du théorème central limite

#### 8.3 Le théorème central limite

#### 8.1 Introduction

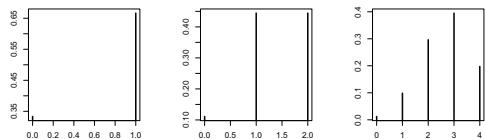
Les théorèmes limites constituent les résultats théoriques les plus importants de la théorie des probabilités. Parmi eux figure le **théorème central limite** (TCL) dont l'utilisation est très fréquente en probabilités et en statistique. Dans le contexte d'épreuves répétées, il montre que la distribution de la somme d'un grand nombre de variables aléatoires est approximativement normale et explique de ce fait le rôle important de la distribution normale.

#### 8.2 Illustrations du théorème central limite

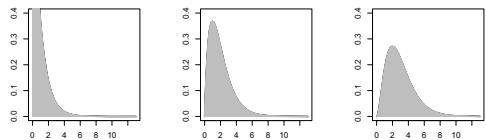
Les graphiques qui suivent représentent l'histogramme ou la densité de la somme de variables aléatoires

- ◊ de Bernoulli;
- ◊ de Poisson;
- ◊ exponentielles.

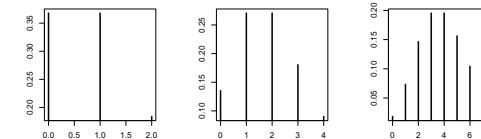
Ils illustrent la convergence de la distribution de ces sommes vers une distribution normale.



Histogrammes de la somme de 1, 2, 4, 8, 16 et 32 variables aléatoires indépendantes de Bernoulli avec  $p = 2/3$ .



Densités de la somme de 1, 2, 3, 4, 5 et 6 variables aléatoires indépendantes exponentielles avec  $\lambda = 1$ .



Histogrammes de la somme de 1, 2, 4, 8, 16 et 32 variables aléatoires indépendantes de Poisson avec  $\lambda = 1$ .

## 8.3 Le théorème central limite

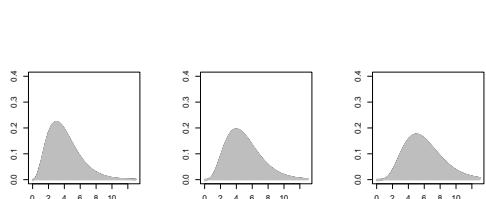
### ◊ Quelques définitions et résultats préalables

Soit  $X_1, X_2, \dots, X_n$  une suite de  $n$  variables aléatoires **indépendantes et identiquement distribuées** (elles sont toutes issues d'une même distribution (pas forcément connue !) et ont toutes la même espérance  $E(X_i) = \mu$  et la même variance  $\text{Var}(X_i) = \sigma^2$ ).

**Définition 1** On définit respectivement par

$$S_n = \sum_{i=1}^n X_i \quad \text{et} \quad M_n = \left( \sum_{i=1}^n X_i \right) / n$$

les variables aléatoires **somme** et **moyenne** des  $n$  variables  $X_i, i = 1, \dots, n$ .



Densités de la somme de 1, 2, 3, 4, 5 et 6 variables aléatoires indépendantes exponentielles avec  $\lambda = 1$ .

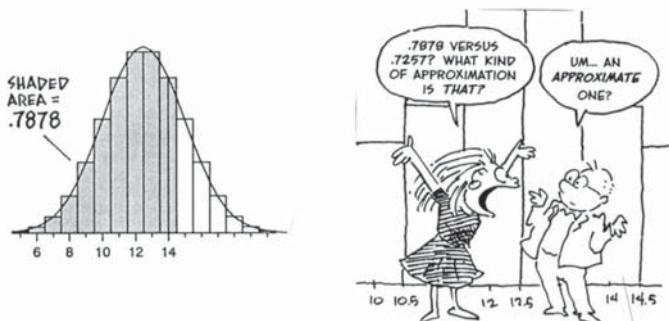
◊ Propriétés de la somme  $S_n$  et de la moyenne  $M_n$

- i)  $E(S_n) = n\mu$       et       $E(M_n) = \mu$ ;  
 ii)  $\text{Var}(S_n) = n\sigma^2$       et       $\text{Var}(M_n) = \sigma^2/n$ .

~ Conséquence

$S_n$  et  $M_n$  ont une variable aléatoire centrée réduite identique, à savoir,

$$Z_n = \frac{S_n - n\mu}{\sqrt{n} \cdot \sigma} = \frac{S_n/n - \mu}{\sqrt{n}/n \cdot \sigma} = \frac{M_n - \mu}{\sigma/\sqrt{n}}.$$



Peut-on obtenir une meilleure approximation pour une distribution discrète ?

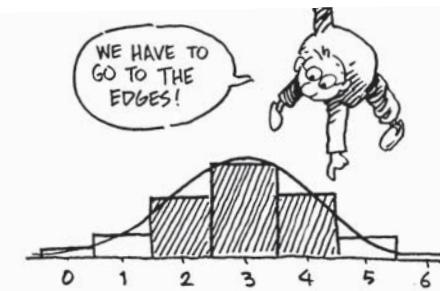
## Le théorème central limite

Soit une suite infinie  $X_1, X_2, \dots, X_n, \dots$  de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), de même espérance mathématique  $E(X_i) = \mu$  et de même variance  $\text{Var}(X_i) = \sigma^2$ .

Lorsque  $n \rightarrow \infty$ , la probabilité d'intervalle

$$P(a < Z_n \leq b) \text{ converge vers } \Phi(b) - \Phi(a),$$

où  $Z_n$  sont les variables aléatoires centrées réduites correspondant aux sommes  $S_n$  ou aux moyennes  $M_n$  des  $n$  premières variables de la suite et  $\Phi$  est la fonction de répartition d'une variable aléatoire normale centrée réduite.



$$\Pr(a \leq X \leq b) \approx \Pr\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

Oui ! En utilisant le coefficient correcteur.



## EXERCICES : LE THÉORÈME CENTRAL LIMITÉ

### Exercice 1

Des moules ont été livrées au poissonnier du quartier. On admet que le poids en grammes ( $g$ ) d'une moule peut être représenté par une variable aléatoire d'espérance  $100\text{ g}$  et d'écart-type  $15\text{ g}$ . Les moules sont livrées par caisse de 50 unités. Soucieux de satisfaire sa clientèle, le poissonnier accepte une caisse pour autant que la moyenne des poids des moules qui s'y trouvent ne soit pas inférieure à  $97\text{ g}$ . Calculer la probabilité qu'une caisse soit refusée par le poissonnier.

**Solution :** 0.0793.

### Exercice 2

Un camion doit livrer des colis dont le poids peut être représenté par une variable aléatoire d'espérance  $\mu = 50\text{ kg}$  et d'écart-type  $\sigma = 5\text{ kg}$ . Calculer la probabilité approchée que le poids total de 40 de ces colis ne dépasse pas 2.05 tonnes.

**Solution :** 0.94295.

### Exercice 3

Considérons une variable aléatoire  $X$  issue d'une distribution binomiale de paramètres  $n = 20$  et  $p = 0.8$ . Déterminer la probabilité  $P(X \geq 14)$

- a) par approximation normale avec puis sans coefficient correcteur,
- b) exactement.

Comparer les résultats obtenus.

**Solutions :** a) avec coefficient correcteur : 0.91337 puis sans coefficient correcteur : 0.85609   b) 0.91331.

### Exercice 4

Une compagnie aérienne a constaté que 4% de ses clients ayant réservé leurs places se désistent au dernier moment et ne se présentent pas au guichet. Sa politique consiste alors à pratiquer de l'"overbooking". Sans prendre trop de risques, elle vend 75 places pour un vol assuré par un avion à capacité de 73 places. Calculer la probabilité qu'un passager se présentant au guichet obtienne une place.

**Solution :** 0.8106.

### Exercice 5

Le nombre d'inscriptions à un cours de programmation en C++ dans une certaine École d'Ingénieurs peut être décrit par une variable aléatoire issue d'une distribution de Poisson d'espérance 100. Le professeur se chargeant du cours a décidé que si le nombre d'inscriptions est supérieur ou égal à 120, il divisera les étudiants en deux groupes et donnera deux fois le cours. Calculer la probabilité que le professeur doive partager les étudiants en deux groupes.

**Solution :** 0.0256.

### Exercice 6

Le nombre de spams reçus par Lagaffe en une journée peut être représenté par une variable aléatoire issue d'une distribution de Poisson d'espérance 20. Déterminer la probabilité approchée que le nombre de spams reçus par Lagaffe un jour donné soit supérieur à 25.

**Solution :** 0.1093.

### Exercice 7

Le volume du trafic aérien d'un certain aéroport (décollages et atterrissages) à une heure de pointe d'un jour donné peut être décrit par une variable aléatoire d'espérance 200 avions et d'écart-type 60 avions. Si l'aéroport peut supporter un trafic de 350 avions par heure, déterminer la probabilité que la capacité maximale de l'aéroport soit dépassée à l'heure de pointe de ce jour-là.

**Solution :** 0.0062.



## Chapitre 9

### Modèles statistiques et estimation de paramètres

#### 9.1 Modèles statistiques et échantillon

- Deux des tâches des plus importantes de la statistique sont la **modélisation** et l'**estimation**. On cherche à partir d'observations (souvent à partir de leurs représentations graphiques) à identifier la loi de probabilité théorique d'une variable statistique. Autrement dit, on cherche à reconstituer le mieux possible le modèle probabiliste d'une situation aléatoire.
- Cette recherche présente souvent peu de difficultés. Les distributions que nous avons introduites dans ce cours s'appliquent en effet à des situations typiques et bien définies.
- Les données sont traitées comme des réalisations de variables aléatoires  $X_i$ , "copies" indépendantes de la variable aléatoire modèle  $X$ .

## Contenu

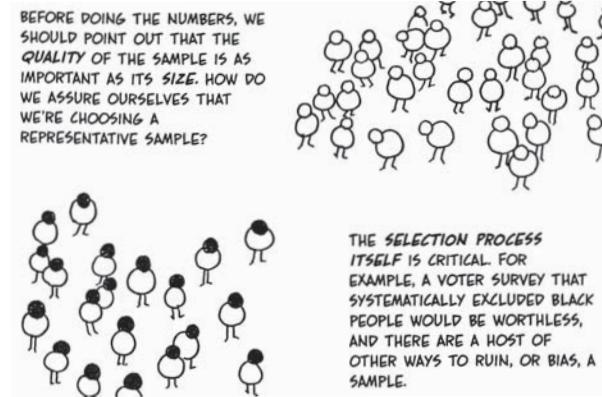
- 9.1 Modèles statistiques et échantillon
- 9.2 Estimateurs
- 9.3 Propriétés d'un estimateur
- 9.4 L'estimation par le maximum de vraisemblance
- 9.5 Estimation par intervalle

#### Exemple 1 :

une pièce de monnaie est jetée six fois de suite. La séquence obtenue est  $(P, P, F, P, F, P)$  où  $P$  indique "Pile" et  $F$  "Face". Cette séquence peut aussi s'écrire sous la forme  $(1, 1, 0, 1, 0, 1)$ . On peut modéliser ces valeurs observées, notées  $x_i$  ( $1 \leq i \leq 6$ ), par une distribution de Bernoulli. Par conséquent, les valeurs  $x_i$  peuvent être considérées comme des réalisations de variables aléatoires indépendantes  $X_i$  ( $1 \leq i \leq 6$ ) issues d'une distribution de Bernoulli dont le paramètre  $p = P(\text{"Pile"})$  est constant mais **inconnu**:

☞ un modèle statistique est une famille de distributions théoriques.

Modèle choisi, il reste à **estimer** le ou les paramètre(s) attaché(s) à la loi de probabilité théorique à partir de données observées sur un échantillon de taille  $n$  représentatif de la population.



**CHOOSE THE SAMPLE AT **random**.**

## 9.2 Estimateurs

**Définition 1** Toute fonction  $T$  des variables aléatoires  $X_1, \dots, X_n$ , i.e de l'échantillon, est appelée **statistique**.

Si cette fonction est utilisée pour estimer certains paramètres de la distribution des variables aléatoires,  $T(X_1, \dots, X_n)$  est un **estimateur**.

Notations :

- $T = T(X_1, \dots, X_n)$  : estimateur du paramètre inconnu
- $t = T(x_1, \dots, x_n)$  : réalisation de  $T$  au moyen des données observées  $x_i$ ;
- estimation du paramètre inconnu.

Exemples :

1.  $T = \bar{X}$ ,  $t = \bar{x}$ ;
2.  $T = \sum_{i=1}^n (X_i - \bar{X})^2 = S_{XX}$ ,  $t = \sum_{i=1}^n (x_i - \bar{x})^2 = s_{xx}$ ;
3. dans l'exemple 1, le paramètre  $p$ , inconnu, indique la probabilité d'obtenir un pile. Un estimateur naturel de  $p$  est

$$T = T(X_1, \dots, X_n) = \frac{\text{nombre de piles}}{\text{nombre total de jets}} = \frac{\text{nombre de "1"}}{\text{nombre total de jets}}.$$

La séquence observée étant  $(1, 1, 0, 1, 0, 1)$ , la réalisation de  $T$  vaut

$$t = T(1, 1, 0, 1, 0, 1) = \frac{4}{6} = \frac{2}{3}.$$

$T$  comme fonction des  $n$  variables aléatoires  $X_1, \dots, X_n$  de l'échantillon est elle-même une variable aléatoire. Sa loi, appelée **distribution d'échantillonnage**, dépend de la loi des  $X_i$ .

Il n'est malheureusement pas toujours possible de déduire la distribution de  $T$  à partir de celle des  $X_i$ . Il faudra souvent se contenter de l'espérance  $E(T)$  et de la variance  $\text{Var}(T)$ . Ces deux caractéristiques de la distribution de  $T$  permettent quand même de disposer d'une bonne information partielle sur la loi de  $T$ . Il est d'ailleurs possible dans certains cas (par exemple si  $T = \bar{X}$ ) d'obtenir une loi approximative de  $T$  par l'intermédiaire du théorème central limite (TCL).

#### Notation :

supposons que le paramètre inconnu soit  $\theta$ . Pour indiquer son rattachement à  $\theta$ , un estimateur de  $\theta$  est souvent noté  $\hat{\theta}$ .



À la recherche d'un "bon" estimateur !

### 9.3 Propriétés d'un estimateur

On souhaite disposer d'un estimateur tel que sa valeur calculée sur un échantillon soit la plus proche possible de la vraie valeur du paramètre de la distribution théorique (distribution fréquemment appelée distribution sous-jacente).

Soit  $T = \hat{\theta}$  un estimateur de la vraie valeur du paramètre  $\theta$ . La qualité de  $\hat{\theta}$  dépend naturellement de la différence

$$\hat{\theta} - \theta.$$

Plusieurs aspects de la qualité de l'estimateur  $\hat{\theta}$  seront traités dans ce paragraphe, à savoir le **biais**, le **carré moyen de l'erreur** et l'**efficacité**.

#### a) Biais :

**Définition 2** Le **biais** de l'estimateur  $\hat{\theta}$  de  $\theta$  est défini par

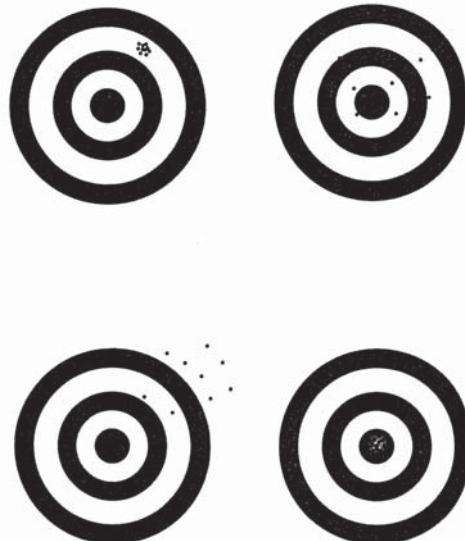
$$b_{\hat{\theta}}(\theta) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta. \quad (1)$$

- ◊ si  $b_{\hat{\theta}}(\theta) < 0$ ,  $\hat{\theta}$  sous-estime  $\theta$ ;
- ◊ si  $b_{\hat{\theta}}(\theta) > 0$ ,  $\hat{\theta}$  sur-estime  $\theta$ ;
- ◊ si  $b_{\hat{\theta}}(\theta) = 0$ ,  $\hat{\theta}$  est un estimateur sans biais de  $\theta$ .

Une première condition pour que  $\hat{\theta}$  soit considéré comme un “bon” estimateur de  $\theta$  est qu’il n’engendre pas un écart systématique entre  $\theta$  et  $\hat{\theta}$ . Autrement dit,  $\hat{\theta}$  doit être un estimateur sans biais de  $\theta$ . Une version plus faible de cette condition est que son biais soit approximativement zéro.

**Remarque :**

une procédure inadéquate d’échantillonnage, une méthodologie analytique peu fiable, une calibration instrumentale erronée sont de fréquentes sources de biais.



### b) Carré moyen de l’erreur :

**Définition 3** Le *carré moyen de l’erreur* de l’estimateur  $\hat{\theta}$  de  $\theta$  est défini par

$$\begin{aligned} \text{CME}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\ &= \text{Var}(\hat{\theta}) + b_{\hat{\theta}}^2(\theta). \end{aligned} \quad (2)$$

Si on dispose de deux estimateurs  $\hat{\theta}_1$  et  $\hat{\theta}_2$  d’un même paramètre  $\theta$ , on choisira naturellement celui dont le carré moyen de l’erreur est le plus petit.

### c) Efficacité :

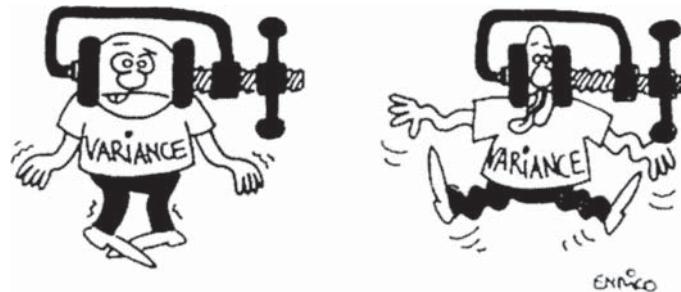
Supposons que  $\hat{\theta}_1$  et  $\hat{\theta}_2$  soient deux estimateurs sans biais du même paramètre  $\theta$ . Ainsi,

$$\text{CME}(\hat{\theta}_1) = \text{Var}(\hat{\theta}_1) \quad \text{et} \quad \text{CME}(\hat{\theta}_2) = \text{Var}(\hat{\theta}_2).$$

**Définition 4** L’estimateur  $\hat{\theta}_1$  est dit plus *efficace* que l’estimateur  $\hat{\theta}_2$  si

$$\begin{aligned} \text{Var}(\hat{\theta}_1) &\leq \text{Var}(\hat{\theta}_2), \\ \text{pour } \hat{\theta}_1 \text{ et } \hat{\theta}_2 \text{ deux estimateurs sans biais de } \theta. \end{aligned}$$

Parmi tous les estimateurs non biaisés de  $\theta$ , on choisira évidemment le plus efficace.



Parmi tous les estimateurs non biaisés, on choisira celui de plus petite variance.

*"When it is not in our power to follow what is true, we ought to follow what is most probable."*

René Descartes (1596–1650)

## 9.4 L'estimation par le maximum de vraisemblance

- Il existe plusieurs méthodes pour estimer le ou les paramètre(s) inconnu(s) attaché(s) à un modèle statistique. Nous présenterons dans ce paragraphe la méthode du **maximum de vraisemblance**. Il s'agit d'une approche générale dans le contexte d'estimation de paramètres inconnus à l'aide de données.
- L'estimateur du maximum de vraisemblance  $\hat{\theta}$  d'un paramètre  $\theta$  est celui qui donne à l'échantillon observé la plus grande vraisemblance (probabilité) d'être obtenu par le modèle choisi.
- Rappelons qu'un échantillon est formé de  $n$  variables aléatoires  $X_i$  indépendantes et identiquement distribuées (i.i.d.). L'estimateur du maximum de vraisemblance est construit à partir des valeurs  $x_i$  prises par les variables  $X_i$ .

Considérons un échantillon  $X_1, \dots, X_n$  provenant d'une distribution dont le paramètre  $\theta$  est inconnu.

**Définition 5** La **fonction de vraisemblance** est définie par

$$V(\theta) = \begin{cases} \prod_{i=1}^n f(x_i; \theta) & \text{cas continu;} \\ \prod_{i=1}^n P(X = x_i) & \text{cas discret,} \end{cases} \quad (3)$$

où  $f(\cdot; \theta)$  est la fonction de densité des variables  $X_i$  dans le cas continu.

Les  $x_i$  indiquent les valeurs prises par les  $X_i$  dans l'expérience; ils sont donc fixes. En revanche, le paramètre inconnu  $\theta$  est **variable**. La fonction de vraisemblance le prend comme argument.

### Exemple 2 :

désignons par  $X$  une variable aléatoire issue d'une distribution exponentielle de paramètre  $\lambda$  inconnu. On souhaite estimer  $\lambda$  par la méthode du maximum de vraisemblance en se basant sur un échantillon de taille  $n$ . La fonction de densité des variables  $X_i$ , "copies" de la variable aléatoire modèle  $X$ , est donnée par

$$f(x_i; \lambda) = \lambda \cdot \exp(-\lambda x_i).$$

La fonction de vraisemblance  $V(\lambda)$  devient

$$\begin{aligned} V(\lambda) &= \prod_{i=1}^n f(x_i; \lambda) \\ &= \lambda \cdot \exp(-\lambda x_1) \cdots \lambda \cdot \exp(-\lambda x_n) \\ &= \lambda^n \exp(-\lambda(x_1 + \cdots + x_n)) \\ &= \lambda^n \exp\left(-\lambda \cdot \sum_{i=1}^n x_i\right). \end{aligned}$$

### Remarques :

- a) chercher le maximum du logarithme de la fonction de vraisemblance,  $\ln(V(\theta))$ , est souvent plus facile que déterminer celui de  $V(\theta)$ . Comme les deux fonctions atteignent leur maximum au même point, les deux possibilités sont tout à fait équivalentes;
- b) la méthode du maximum de vraisemblance peut facilement être généralisée à l'estimation de plusieurs paramètres. Pour estimer un seul paramètre  $\theta$ , il convient de résoudre l'équation  $\frac{d}{d\theta} \ln(V(\theta)) = 0$  puis de vérifier que la solution obtenue soit effectivement un maximum. Pour plusieurs paramètres  $\theta_1, \dots, \theta_r$ , le procédé est analogue : il s'agit de résoudre un système d'équations puis de contrôler que la solution donne bien un maximum.

### Remarque :

la fonction de vraisemblance est formée d'un produit de facteurs (fonctions de densité évaluées aux valeurs  $x_i$  ou probabilités  $P(X = x_i)$ ) en raison de l'hypothèse d'indépendance dans l'échantillon.

**Définition 6** L'estimateur du maximum de vraisemblance  $\hat{\theta}$  est le point qui maximise la vraisemblance  $V(\theta)$  :

$$V(\hat{\theta}) \geq V(\theta) \text{ pour tout } \theta.$$

Ainsi, l'estimateur du maximum de vraisemblance permet d'obtenir l'échantillon observé à l'aide du modèle sous-jacent avec la plus grande vraisemblance (probabilité).

### Exemple 2 (suite) :

le logarithme de la fonction de vraisemblance s'écrit sous la forme

$$\ln(V(\lambda)) = n \ln(\lambda) - \lambda \cdot \sum_{i=1}^n x_i.$$

La dérivée de cette fonction par rapport à  $\lambda$  vaut

$$\frac{d}{d\lambda} \ln(V(\lambda)) = \frac{n}{\lambda} - \sum_{i=1}^n x_i.$$

Les valeurs potentielles  $\hat{\lambda}$  qui annulent la dérivée vérifient la relation

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^n x_i. \quad (4)$$

### Exemple 2 (suite... et fin) :

l'unique solution de l'équation (4) est

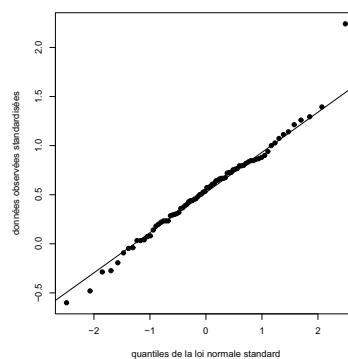
$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}.$$

Comme la dérivée seconde de la fonction  $\ln(V(\lambda))$  est négative pour toutes les valeurs de  $\lambda > 0$ ,  $\hat{\lambda}$  correspond bien à un maximum.

Modèle choisi, estimation calculée, le graphique **quantiles versus quantiles** permet de juger la qualité de l'ajustement du modèle. L'idée consiste à comparer dans un graphique de nuage de points les quantiles des données observées aux quantiles de la distribution théorique. Si la loi supposée est la vraie loi théorique, les points du graphe doivent se situer approximativement sur une droite. Si ce n'est pas le cas, le modèle n'est pas adéquat.

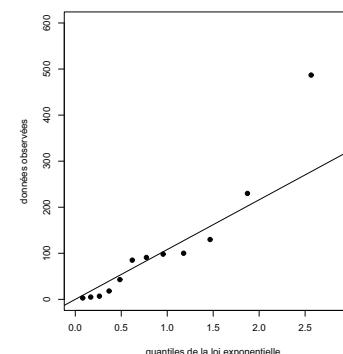
### Remarque :

il n'est pas toujours évident de juger le bon ajustement d'un modèle, particulièrement si le nombre d'observations est relativement petit.



*Graphique quantiles versus quantiles de la différence entre la moyenne du contrôle continu et l'examen pour les étudiants El, 1ère année, EIVD, décembre 2003.*

### Exemple 4 : heures de service entre deux pannes consécutives.



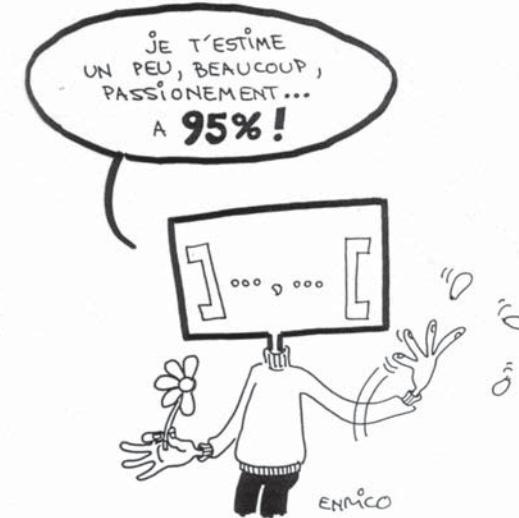
*Graphique quantiles versus quantiles.*

## 9.5 Estimation par intervalle

Aussi précise soit elle, l'estimation ne peut jamais être exacte. Ainsi, au lieu d'estimer le paramètre  $\theta$  par un nombre  $\hat{\theta}$ , on construit un intervalle aléatoire appelé **intervalle de confiance** qui couvre  $\theta$  avec une grande probabilité.

Objectif :

estimer  $\theta$  à l'aide d'un intervalle de confiance  $[I, S]$ .



Principe :

considérons un estimateur  $T = T(X_1, \dots, X_n)$  d'un paramètre  $\theta$  et désignons par  $f_T(\theta)$  sa fonction de densité.

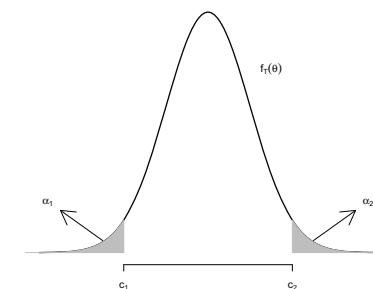
La construction de l'intervalle de confiance s'effectue en deux étapes :

1. si on se fixe un **risque d'erreur**  $\alpha$  ou, ce qui est équivalent, un **seuil de confiance**  $1 - \alpha$ , il est possible de déterminer les valeurs  $c_1$  et  $c_2$  de telle sorte que

$$\alpha_1 = P(T \leq c_1) \text{ et } \alpha_2 = P(T \geq c_2) \text{ avec } \alpha = \alpha_1 + \alpha_2.$$

Ainsi,

$$P(c_1 \leq T \leq c_2) = 1 - \alpha;$$



Première étape dans la construction d'un intervalle de confiance.

Principe (*suite*) :

2. comme la loi de l'estimateur  $T$  dépend de  $\theta$ , on peut, dans certains cas, transformer l'expression  $c_1 \leq T \leq c_2$  en une expression équivalente  $I \leq \theta \leq S$ . Ainsi,

$$P(I \leq \theta \leq S) = 1 - \alpha.$$

**Définition 7** L'intervalle  $[I, S]$  est dit **intervalle de confiance** à un seuil de confiance  $1 - \alpha$  pour le paramètre  $\theta$ . Les statistiques  $I$  et  $S$  sont appelées les **bornes de confiance inférieure et supérieure**.

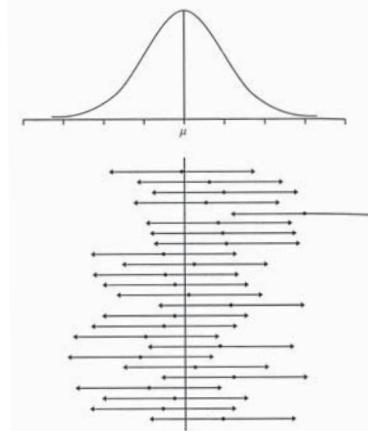
Le risque d'erreur  $\alpha$  est choisi très petit. En pratique, il vaut 5 % ou 1 %.

Remarques :

- a) interprétation de l'intervalle de confiance  $[I, S]$  :

l'intervalle de confiance est tel que la probabilité que la variable aléatoire  $I$  soit inférieure à  $\theta$  et que la variable aléatoire  $S$  soit supérieure à  $\theta$  vaut  $1 - \alpha$ . Le paramètre  $\theta$  n'étant pas une variable aléatoire, il n'est pas correct d'interpréter le niveau de confiance comme une probabilité et d'affirmer que  $\theta$  se situe dans l'intervalle avec probabilité  $1 - \alpha$ . Néanmoins, il se peut qu'on se permette un abus de langage en prétendant qu'un intervalle de confiance numériquement déterminé a une probabilité de  $1 - \alpha$  de contenir le paramètre inconnu  $\theta$ ;

- b) l'application du principe n'est pas toujours aisée. En effet, isoler  $\theta$  dans l'expression  $c_1 \leq T \leq c_2$  peut être compliqué, voire même impossible.



Vingt-cinq intervalles de confiance à 95 % pour l'espérance  $\mu$

dans le cas d'une distribution normale.

Exemple 5 :

estimation par intervalle pour l'espérance  $\mu$  dans le cas d'une distribution normale.

Sur la base d'un échantillon  $X_1, \dots, X_n$  d'une loi normale d'espérance  $\mu$  et de variance  $\sigma^2$ , connu puis inconnu, on se propose de déterminer un intervalle de confiance pour  $\mu$  à risque d'erreur  $\alpha$ . En d'autres termes, on cherche un intervalle  $[I, S]$  tel que

$$P(I \leq \mu \leq S) = 1 - \alpha.$$

**Cas 1 :** variance  $\sigma^2$  connue.

- L'estimateur standard de  $\mu$  à partir duquel sera construit l'intervalle de confiance est  $\hat{\mu} = \bar{X}$ .

- La variable aléatoire

$$T^* = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

suit une loi normale d'espérance 0 et de variance 1.

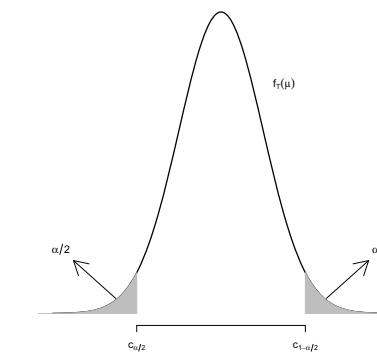
- En utilisant la table de la fonction de répartition d'une variable aléatoire normale centrée réduite, on trouve les valeurs  $c_1 = c_{\alpha/2}$  et  $c_2 = c_{1-\alpha/2}$  telles que

$$P(c_{\alpha/2} \leq T^* \leq c_{1-\alpha/2}) = 1 - \alpha.$$

Plus précisément,

$$\alpha/2 = \Phi(c_{\alpha/2}) \text{ et } 1 - \alpha/2 = \Phi(c_{1-\alpha/2}).$$

Les nombres  $c_{\alpha/2}$  et  $c_{1-\alpha/2}$  sont appelés les **quantiles théoriques** de la distribution normale centrée réduite.



Première étape dans la construction d'un intervalle de confiance pour l'espérance  $\mu$  avec variance connue.

- Comme la fonction de répartition  $\Phi$  est symétrique, il s'ensuit que  $c_{\alpha/2} = -c_{1-\alpha/2}$ . Par conséquent,

$$P(-c_{1-\alpha/2} \leq T^* \leq c_{1-\alpha/2}) = 1 - \alpha,$$

ou encore, selon la définition de  $T^*$ ,

$$P\left(-c_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq c_{1-\alpha/2}\right) = 1 - \alpha.$$

- En isolant  $\mu$ , on obtient un intervalle de confiance pour  $\mu$  à un seuil de confiance  $1 - \alpha$ ,

$$P\left(\underbrace{\bar{X} - c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}}_{I} \leq \mu \leq \underbrace{\bar{X} + c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}}_{S}\right) = 1 - \alpha.$$

En résumé, l'intervalle de confiance pour l'espérance  $\mu$  d'une variable aléatoire  $X$  issue d'une distribution normale de variance  $\sigma^2$  connue se détermine de la manière suivante :

- choisir un seuil de confiance  $1 - \alpha$ ;
- déterminer le quantile théorique  $c_{1-\alpha/2}$  dans la table de la fonction de répartition d'une variable aléatoire normale centrée réduite (si  $1 - \alpha = 95\%$ ,  $c_{1-\alpha/2} = 1.96$ ; si  $1 - \alpha = 99\%$ ,  $c_{1-\alpha/2} = 2.575$ );
- calculer la moyenne arithmétique  $\bar{x}$  des valeurs observées  $x_1, \dots, x_n$ ;
- au seuil de confiance  $1 - \alpha$ , l'intervalle de confiance pour  $\mu$  est

$$\left[ \bar{x} - c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Remarque :

la longueur de l'intervalle de confiance vaut  $L = 2 c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ . On constate alors que

- a) si le seuil de confiance  $1 - \alpha$  augmente,  $L$  augmente aussi;
- b)  $L$  est proportionnel à  $\frac{1}{\sqrt{n}}$ ; par conséquent, si  $n$  est multiplié par 4,  $L$  est divisé par 2;
- c)  $L$  est proportionnel à  $\sigma$ .

### Cas 2 : variance $\sigma^2$ inconnue.

Le paramètre  $\sigma^2$  étant inconnu, il n'est plus possible d'utiliser la statistique

$$T^* = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

◊ Question : que peut-on faire ?

◊ Réponse : utiliser un "bon" estimateur de la variance  $\sigma^2$  !

- Le choix de l'estimateur de  $\sigma^2$  se porte naturellement sur

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2,$$

estimateur sans biais de  $\sigma^2$ .

- Comme  $S$  est une variable aléatoire, la statistique qui en découle

$$T^* = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

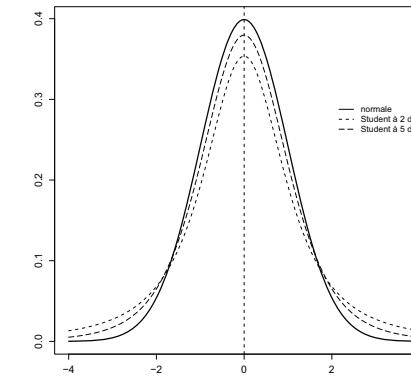
n'est plus issue d'une distribution normale. Cependant, on peut montrer que  $T^*$  suit une loi de Student à  $n - 1$  degrés de liberté (ddl =  $n - 1$ ).



William Gosset (1876–1937), brasseur et mathématicien britannique plus connu sous le pseudonyme de Student.

### Remarques :

- le nombre de degrés de liberté (ddl) est égal au nombre de valeurs observées moins le nombre de paramètres déjà estimés (dans notre cas, un paramètre  $\mu$  estimé par  $\bar{X}$ );
- si  $n$  est grand ( $n \geq 30$ ), la loi de Student est très proche de la loi normale;
- les procédures pour déterminer l'intervalle de confiance pour l'espérance, variance connue et variance inconnue, se distinguent par le quantile théorique. Dans le premier cas, il est tiré de la table de la loi normale centrée réduite; dans le second cas, il provient de la table de la loi de Student.



Fonctions de densité de variables normale et de Student.

En résumé, l'intervalle de confiance pour l'espérance d'une variable aléatoire normale  $X$  de variance  $\sigma^2$  inconnue se détermine de la manière suivante :

- choisir un seuil de confiance  $1 - \alpha$ ;
- déterminer le quantile théorique  $t_{n-1, 1-\alpha/2}$  dans la table des quantiles de la loi de Student à  $n - 1$  degrés de liberté;
- calculer la moyenne arithmétique  $\bar{x}$  et la variance  $s^2$  des valeurs observées  $x_1, \dots, x_n$ ;
- au seuil de confiance  $1 - \alpha$ , l'intervalle de confiance pour  $\mu$  est

$$\left[ \bar{x} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right].$$

### Exemple 6 :

estimation par intervalle pour la variance  $\sigma^2$  dans le cas d'une distribution normale.

À partir d'un échantillon  $X_1, \dots, X_n$  d'une loi normale d'espérance  $\mu$  et de variance  $\sigma^2$ , on cherche un intervalle de confiance pour le paramètre  $\sigma^2$  à risque d'erreur  $\alpha$ .

- L'estimateur standard de  $\sigma^2$  est

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2.$$

Cet estimateur est sans biais.

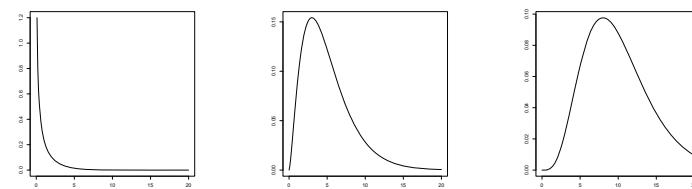
- On peut montrer que la variable aléatoire

$$T^* = (n - 1) \cdot \frac{S^2}{\sigma^2}$$

suit une loi particulière, appelée loi du  $\chi^2$  (khi carré ou khi deux) à  $n - 1$  degrés de liberté.

**Remarque :**

soient  $n$  variables aléatoires  $X_i$ ,  $i = 1, \dots, n$ , indépendantes issues d'une distribution normale centrée réduite. La variable  $Y = \sum_{i=1}^n X_i^2$  suit une loi du khi carré à  $n$  degrés de liberté. C'est d'ailleurs de cette manière qu'est fréquemment introduite la distribution du khi carré dans des cours de probabilités.

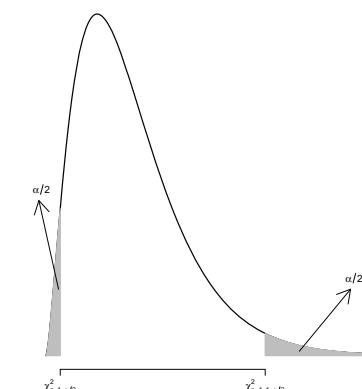


Densité de variables aléatoires issues de distribution de khi carré à 1, 5 et 10 degrés de liberté.

La procédure pour obtenir l'intervalle de confiance recherché est la suivante :

- choisir un seuil de confiance  $1 - \alpha$ ;
- dans la table contenant les quantiles théoriques de la distribution  $\chi^2$  à  $n - 1$  degrés de liberté, déterminer les valeurs  $\chi_{n-1, \alpha/2}^2$  et  $\chi_{n-1, 1-\alpha/2}^2$  telles que
$$F_{\chi_{n-1}^2}(\chi_{n-1, \alpha/2}^2) = \alpha/2 \text{ et } F_{\chi_{n-1}^2}(\chi_{n-1, 1-\alpha/2}^2) = 1 - \alpha/2;$$
- calculer  $(n - 1) \cdot s^2$  à partir des données observées  $x_1, \dots, x_n$ ;
- au seuil de confiance  $1 - \alpha$ , l'intervalle de confiance pour  $\sigma^2$  est

$$\left[ \frac{(n - 1) \cdot s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n - 1) \cdot s^2}{\chi_{n-1, \alpha/2}^2} \right].$$



Deuxième étape dans la construction d'un intervalle de confiance pour  $\sigma^2$ .

### Exemple 7 :

estimation par intervalle pour le paramètre  $p$  dans le cas d'une distribution binomiale.

Considérons un échantillon  $X_1, \dots, X_n$  issu d'une distribution de Bernoulli de paramètre  $p$ ,  $0 \leq p \leq 1$ . On souhaite construire un intervalle de confiance pour  $p$  à risque d'erreur  $\alpha$ .

- L'estimateur standard de  $p$  est  $\hat{P} = \bar{X}$  tel que  $E(\hat{P}) = p$  et  $\text{Var}(\hat{P}) = p(1-p)/n$ .
- Si la taille de l'échantillon est suffisamment grande, on peut admettre l'approximation normale pour l'estimateur  $\bar{X}$ .

On remarque que les statistiques  $I$  et  $S$  contiennent  $p$  !!!

~~ Que faire ?

Une idée : remplacer  $p$  par  $\hat{P} = \bar{X}$  dans les statistiques  $I$  et  $S$ .

L'intervalle de confiance

$$\left[ \bar{X} - c_{1-\alpha/2} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + c_{1-\alpha/2} \cdot \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right]$$

est, approximativement, de seuil de confiance  $1 - \alpha$  pour le paramètre  $p$ .

- On a donc

$$P\left(-c_{1-\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq c_{1-\alpha/2}\right) \simeq 1 - \alpha,$$

où  $c_{1-\alpha/2}$  est le  $(1 - \alpha/2)$ -quantile théorique de la distribution normale centrée réduite.

- Ainsi, on obtient un intervalle de confiance pour  $p$  à un seuil de confiance  $1 - \alpha$ ,

$$P\left(\underbrace{\bar{X} - c_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}}_{I} \leq p \leq \underbrace{\bar{X} + c_{1-\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}}_{S}\right) \simeq 1 - \alpha.$$

La méthode pour obtenir l'intervalle de confiance pour  $p$  dans le cas d'une distribution binomiale est :

1. choisir un seuil de confiance  $1 - \alpha$ ;
2. dans la table des quantiles théoriques de la distribution normale centrée réduite, déterminer la valeur  $c_{1-\alpha/2}$  telle que

$$\Phi(c_{1-\alpha/2}) = 1 - \alpha/2;$$

3. calculer la moyenne arithmétique  $\bar{x}$  des valeurs observées  $x_1, \dots, x_n$ ;
4. approximativement au seuil de confiance  $1 - \alpha$ , l'intervalle de confiance pour  $p$  est

$$\left[ \bar{x} - c_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + c_{1-\alpha/2} \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right].$$

## EXERCICES : MODÈLES STATISTIQUES ET ESTIMATION DE PARAMÈTRES

### Exercice 1

*Classe Éco* était une émission hebdomadaire de la Télévision Suisse Romande consacrée “aux conséquences économiques d’un événement ou d’une décision et leur incidence directe ou indirecte sur la vie quotidienne des gens” (définition tirée du site de l’émission). Dans son édition du lundi 22 mars 2004, *Classe Éco* avait soumis aux téléspectateurs la proposition suivante de l’USAM (Union suisse des arts et métiers) pour doper l’économie suisse : “Lorsque leur efficacité professionnelle diminue à cause de l’âge, les employés doivent être prêts à renoncer à une partie de leur salaire.” Dès le début de l’émission, les téléspectateurs avaient la possibilité par SMS non gratuits de réagir positivement ou négativement à la proposition de l’USAM. Les résultats du “mini sondage SMS” sont les suivants :

4488 SMS reçus    90 % contre    10 % pour

Expliquer pourquoi ce sondage est certainement biaisé.

### Exercice 2

- Considérons un échantillon  $X_1, X_2, \dots, X_n$  d’une distribution de Bernoulli  $\mathcal{B}(1, p)$ . Montrer que l’estimateur  $\bar{X}$  de  $p$  est un estimateur sans biais.
- Buffon, un naturaliste français du dix-huitième siècle, lança 4040 fois une pièce de monnaie et obtint pile à 2048 reprises. Il se demanda alors si la pièce de monnaie était bien équilibrée (sans biais). Calculer l’estimation de la proportion  $p$  de piles.
- L’écart entre l’estimation de  $p$  et la valeur que prendrait  $p$  si la pièce de Buffon était équilibrée peut être considéré comme un critère naturel pour juger si la pièce était biaisée ou non. Le calcul de cette différence est-il suffisant pour conclure ?

Solutions : b)  $\frac{2048}{4040} \simeq 0.5069$     c) non.

### Exercice 3

Supposons que  $\hat{\theta}_1$  et  $\hat{\theta}_2$  soient deux estimateurs sans biais d’un paramètre inconnu  $\theta$ . Si  $\text{Var}(\hat{\theta}_1) = 10$  et  $\text{Var}(\hat{\theta}_2) = 4$ , quel estimateur est le meilleur ? En quel sens, est-il le meilleur ?

**Solution :**  $\hat{\theta}_2$  est un meilleur estimateur de  $\theta$  que  $\hat{\theta}_1$  au sens de l’efficacité.

### Exercice 4

On suppose qu’un certain phénomène aléatoire peut être modélisé par une variable aléatoire continue  $X$  dont la fonction de densité est donnée par

$$f_X(u) = \begin{cases} \frac{1}{2} \cdot (1 + \theta u) & \text{si } -1 \leq u \leq 1, \\ 0 & \text{sinon.} \end{cases}$$

- Calculer l’espérance mathématique de  $X$  en fonction du paramètre inconnu  $\theta$ .
- Sur la base d’un échantillon, déterminer l’estimateur  $\hat{\theta}$  de  $\theta$  par la méthode des moments.
- Montrer que  $\hat{\theta}$  est un estimateur sans biais de  $\theta$ .

Solutions : a)  $E(X) = \frac{\theta}{3}$     b)  $\hat{\theta} = 3\bar{X}$     c)  $E(\hat{\theta}) = E(3\bar{X}) = 3E(\bar{X}) = 3 \cdot \frac{\theta}{3} = \theta$ .

### Exercice 5

On suppose que le temps d’attente devant le guichet d’un office postal peut être modélisé par une variable aléatoire  $X$  issue d’une distribution exponentielle de paramètre inconnu  $\lambda$ .

- Calculer l’espérance mathématique de  $X$  en fonction de  $\lambda$ .
- En se basant sur l’espérance de  $X$ , proposer un estimateur naturel  $\hat{\lambda}$  du paramètre  $\lambda$ .
- À l’aide des statistiques élémentaires calculées par le logiciel de statistique R,

| Min.  | 1stQu. | Median | Mean  | Sd     | 3rdQu. | Max.   |
|-------|--------|--------|-------|--------|--------|--------|
| 0.010 | 1.275  | 2.495  | 3.491 | 2.9891 | 4.985  | 14.210 |

calculer l’estimation de  $\lambda$  en utilisant son estimateur  $\hat{\lambda}$  déterminé en b). Les données originales sont fictives.

- Un outil graphique très pratique pour juger la qualité de l’ajustement du modèle aux données est le graphique quantiles versus quantiles. Il s’agit en fait d’un graphique de nuage de points dans lequel les quantiles observés se trouvent sur l’un des deux axes et les quantiles théoriques correspondants issus de la distribution exponentielle figurent sur l’autre axe. Si l’adéquation entre les quantiles observés et ceux issus de la distribution présumée est bonne, les points seront relativement bien alignés.

En se basant sur le graphique de droite, que peut-on dire de la qualité du modèle ajusté ?

Solutions : a)  $E(X) = 1/\lambda$     b)  $\hat{\lambda} = 1/\bar{X}$     c)  $\hat{\lambda} = 1/\bar{x} = 0.29$     d) l’ajustement du modèle est de bonne qualité. Ainsi, le modèle exponentiel est raisonnable.

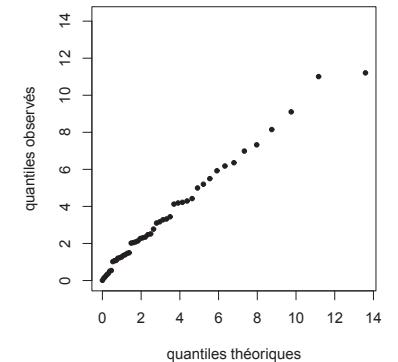
### Exercice 6

Le nombre d’appels téléphoniques reçus par une petite PME entre 8h00 et 9h00 a été relevé pendant six semaines. Les fréquences observées figurent dans le tableau ci-dessous :

| nombre d’appels observé | 5 | 6 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|-------------------------|---|---|---|---|----|----|----|----|----|----|
| fréquence               | 2 | 2 | 3 | 4 | 6  | 4  | 4  | 2  | 1  | 2  |

On suppose que le nombre d’appels téléphoniques reçus par la PME peut être modélisé par une variable aléatoire issue d’une distribution de Poisson  $\mathcal{P}(\lambda)$ .

- Proposer un estimateur naturel  $\hat{\lambda}$  du paramètre  $\lambda$ .
- En utilisant a), calculer l’estimation de  $\lambda$  au moyen des données figurant dans le tableau.



- c) Calculer la probabilité théorique  $P(X = 9)$  à l'aide de la distribution de Poisson puis en déduire la fréquence espérée associée.
- d) Comparer la fréquence espérée calculée en c) et la fréquence observée du tableau.

**Solutions :** a)  $\hat{\lambda} = \bar{X}$  b)  $\frac{304}{30} = 10.13$  c) 0.123 et 3.7.

### Exercice 7

Les bouteilles distribuées par la firme Gaulois & Cie sont supposées contenir 300 millilitres (*ml*) d'une potion magique permettant d'acquérir une force hors du commun. Comme le procédé de remplissage est artisanal, une variation de la quantité de potion contenue dans les bouteilles est inévitable. La quantité de liquide des bouteilles peut être représentée par une variable aléatoire issue d'une distribution normale avec un écart-type de 3 *ml*. On a pu prélever qu'un petit échantillon dont les résultats sont les suivants :

299.4 297.7 301.0 298.9 300.2 297.0.

Déterminer un intervalle de confiance pour la quantité de potion espérée  $\mu$  contenue dans les bouteilles à un risque d'erreur de 10% puis de 1%. En se basant sur l'échantillon prélevé, le procédé de remplissage des bouteilles fonctionne-t-il correctement ?

**Solution :** [297.02, 301.05] pour un seuil de confiance à 90% et [295.88, 302.19] pour un seuil de confiance à 99%. Les deux intervalles de confiance couvre la valeur cible 300. Ainsi, en se basant sur l'échantillon prélevé, le processus artisanal de remplissage des bouteilles fonctionne correctement aux risques d'erreur de 10% et 1%.

### Exercice 8

Un nouveau lot de minerai est testé pour déterminer si le contenu moyen de nickel est cohérent avec la valeur habituelle de 3.25% observée dans les lots extraits précédemment. Les résultats obtenus sont les suivants :

3.27 3.23 3.31 3.34 3.26 3.24 3.25 3.37 3.29 3.33

- a) On suppose que les valeurs observées sont issues d'une distribution normale. À l'aide des statistiques élémentaires calculées par le logiciel de statistique **R**,

| Min.   | 1stQu. | Median | Mean   | Sd      | 3rdQu. | Max.   |
|--------|--------|--------|--------|---------|--------|--------|
| 3.2300 | 3.2525 | 3.2800 | 3.2890 | 0.04701 | 3.3250 | 3.3700 |

calculer un intervalle de confiance bilatéral à 95% pour le contenu moyen de nickel du lot.

- b) Existe-t-il une différence significative entre le nouveau lot et les anciens lots au niveau du contenu moyen de nickel ?
- c) La quantité soustraite et ajoutée à la moyenne arithmétique pour calculer un intervalle de confiance d'une valeur moyenne s'appelle la marge d'erreur. On admet que l'écart-type de la population est connu et vaut 0.05. Quelle taille minimale devrait avoir le lot pour obtenir une marge d'erreur d'au plus 0.015 dans un intervalle de confiance à 95% pour le contenu moyen de nickel du lot ?

**Solutions :** a) [3.255, 3.323] b) comme l'intervalle de confiance ne couvre tout juste pas la valeur habituelle (espérée) de 3.25%, il existe une différence significative entre le nouveau lot et les anciens au niveau du contenu moyen de nickel et ce à un risque d'erreur de 5%. Néanmoins, restons prudents sur la portée pratique de cette conclusion c)  $n = 43$ .

### Exercice 9

On suppose que l'estimation  $\hat{p}$  d'une proportion  $p$  calculée à partir des valeurs observées d'un échantillon de taille 100 vaut  $\hat{p} = 0.21$ . Calculer les intervalles de confiance pour  $p$  aux seuils de confiance à 90%, 95% et 99% ainsi que les longueurs de ces intervalles.

**Solution :** [0.143, 0.277] pour un seuil de confiance à 90%, [0.130, 0.290] pour un seuil de confiance à 95%, [0.105, 0.315] pour un seuil de confiance à 99%. Longueur  $L$  des intervalles : 90%,  $L = 0.134$ ; 95%,  $L = 0.160$ ; 99%,  $L = 0.210$ .

### Exercice 10

Plusieurs PME de Suisse Romande se demandaient si elles allaient installer le système d'exploitation Windows Vista l'année suivante ou conserver Windows XP. Elles craignaient notamment l'instabilité du nouveau système malgré la publicité qui vantait les "100 raisons de vivre une expérience totalement inédite avec Windows Vista". Un sondage fictif (prudence d'usage face aux sondages !) avait été réalisé. Il révélait que parmi les 301 PME interrogées, 208 ne pensaient pas installer Windows Vista l'année suivante.

- a) Déterminer un intervalle de confiance à 99% pour la proportion de PME de Suisse Romande qui ne songeaient pas installer Windows Vista l'année suivante.
- b) On admet que la proportion de PME qui ne songeaient pas installer Windows Vista l'année suivante vaut 0.55. Calculer la taille minimale que doit avoir l'échantillon pour obtenir une marge d'erreur d'au plus 2% à un seuil de confiance de 99%.

**Solutions :** a) [0.622, 0.760] b) 4106.

### Exercice 11

Les étudiants des deux classes de 2<sup>ème</sup> année de la filière informatique, orientation logiciel, de l'EIVD avaient subit le même travail écrit de probabilités et statistique le 3 juillet 2003. Les résultats obtenus figurent dans le fichier **te.txt**. Enregistrez-les dans l'object **te** de **R**.

- a) Construire les boîtes à moustaches en parallèle des valeurs observées suivant les classes. Faites en sorte que le cadre du graphique soit carré.
- b) Pour chaque classe, calculer les intervalles de confiance à 95% pour les espérances des notes obtenues par les étudiants.
- c) Les commandes suivantes de **R** permettent d'ajouter les intervalles de confiance déterminés en b) au graphique des boîtes à moustaches :

```
> attach(te)
> m.te<-tapply(note,classe,mean)
> sd.te<-tapply(note,classe,sd)
> 1.te<-tapply(note,classe,length)
> xi<-c(1.45,1.55)
> points(xi,m.te,col="blue",pch=16)
```

```

> arrows(xi,m.te-(qt(0.975,1.te-1)*sd.te/sqrt(1.te)),xi,
+ m.te+(qt(0.975,1.te-1)*sd.te/sqrt(1.te)),code=0,col="blue",lwd=2)
> detach("te")

```

- d) En se basant sur les boîtes à moustaches et les intervalles de confiance, existe-t-il une différence significative entre les résultats des étudiants des deux classes ?
- e) L'hypothèse de normalité pour la construction des intervalles de confiance se justifie-t-elle en se basant sur les boîtes à moustaches ?

**Solutions :** b) pour la classe A, l'intervalle de confiance à 95 % est [4.29, 4.78]; au même niveau de confiance, l'intervalle de confiance pour la classe B vaut [4.34, 4.82] d) non e) oui.

### Exercice 12

Les directions de deux écoles canadiennes, Glooscap et Coldbrook, ont décidé de comparer les programmes d'éducation physique. Pour chaque école, un échantillon représentatif des élèves de sixième a été déterminé. La course à pied faisant partie du programme, les temps en secondes qu'ont mis les étudiants formant les deux échantillons pour parcourir une certaine distance ont été relevés :

|           |       |       |       |       |       |       |       |       |       |
|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Glooscap  | 14.37 | 14.58 | 13.02 | 12.73 | 15.14 | 13.16 | 12.04 | 12.81 | 13.03 |
|           | 13.13 | 13.85 | 11.89 |       |       |       |       |       |       |
| Coldbrook | 12.01 | 13.17 | 11.13 | 13.95 | 13.28 | 12.30 | 13.65 | 11.42 | 12.01 |
|           | 11.08 | 11.73 | 12.13 | 11.11 |       |       |       |       |       |

Les données figurent dans le fichier **runtime.txt**. Enregistrez-les dans l'objet **runtime** de **R**.

- a) Construire les boîtes à moustaches en parallèle des valeurs observées selon les écoles. Faites en sorte que le cadre du graphique soit carré.
- b) Pour chaque école, calculer les intervalles de confiance à 95 % pour les temps moyens réalisés par les élèves.
- c) Les commandes suivantes de **R** permettent d'ajouter les intervalles de confiance déterminés en b) au graphique des boîtes à moustaches :

```

> m.runtime<-tapply(runtime$runtime,runtime$school,mean)
> sd.runtime<-tapply(runtime$runtime,runtime$school,SD)
> l.runtime<-tapply(runtime$runtime,runtime$school,length)
> xi<-c(1.45,1.55)
> points(xi,m.runtime,col="blue",pch=16)
> arrows(xi,m.runtime-(qt(0.975,1.runtime-1)*sd.runtime/sqrt(l.runtime)),
+ xi,m.runtime+(qt(0.975,1.runtime-1)*sd.runtime/sqrt(l.runtime)),code=0,
+ col="blue",lwd=2)

```

- d) En se basant sur les boîtes à moustaches et les intervalles de confiance, existe-t-il une différence significative entre les temps moyens des élèves des deux écoles ?
- e) L'hypothèse de normalité pour la construction des intervalles de confiance se justifie-t-elle en se basant sur les boîtes à moustaches ?

**Solutions :** b) pour Glooscap, l'intervalle de confiance à 95 % est [12.683, 13.942]; pour Coldbrook, l'intervalle de confiance à 95 % vaut [11.630, 12.827] d) oui e) oui.

### Exercice 13

La quantité de sucre contenu dans un sirop de pêche est supposée être issue d'une distribution normale. L'écart-type  $s$  calculé sur un échantillon de taille 10 vaut  $4.8 \text{ mg}$ .

- a) Construire un intervalle de confiance à 90 % pour la variance  $\sigma^2$ .
- b) En se basant sur l'intervalle de confiance déterminé en a), peut-on considérer une variance  $\sigma^2$  égale à 18 comme étant plausible ?
- c) Si le risque d'erreur passe à 5 %, la longueur de l'intervalle de confiance va-t-elle diminuer ?
- d) Supposons que l'estimation de l'écart-type  $s$  a été calculée sur un échantillon de taille 20. Déterminer l'intervalle de confiance à 90 % pour  $\sigma^2$ . Est-il plus long ou plus court que celui calculé en a) ?

**Solutions :** a) l'intervalle de confiance pour la variance  $\sigma^2$  au seuil de confiance de 90 % est [12.27, 62.27] b) comme l'intervalle de confiance pour  $\sigma^2$  couvre la valeur hypothétique  $\sigma_0^2 = 18$ , on peut considérer une variance de 18 comme étant plausible et ce à un risque d'erreur de 10 % c) l'intervalle de confiance pour la variance  $\sigma^2$  au seuil de confiance de 95 % est [10.91, 76.80]. Si le risque d'erreur diminue, la longueur de l'intervalle de confiance augmente. On s'attendait tout à fait à ce type de conclusion d) si l'estimation  $s$  de l'écart-type avait été calculée sur un échantillon de taille 20, l'intervalle de confiance pour  $\sigma^2$  au seuil de confiance de 90 % serait [14.54, 43.34]. La longueur de l'intervalle vaut 28.8. Il est plus court que celui calculé en a).



# Chapitre 10

## Tests d'hypothèses

### Contenu

- 10.1 Qu'entend-on par "inférence statistique" ?
- 10.2 Tests d'hypothèses
- 10.3 Quelques tests d'hypothèses
- 10.4 Mises en garde
- 10.5 Rudiments de la théorie de la décision
- 10.6 Et ensuite...

### 10.1 Qu'entend-on par "inférence statistique" ?

- L'objectif de l'**inférence statistique** consiste à établir des conclusions précises sur une population au moyen des valeurs observées d'un échantillon. Très souvent, on cherche à tirer des conclusions sur les paramètres inconnus de la population.
- Certaines conclusions peuvent provenir de l'analyse exploratoire des données. Toutefois, les graphiques ne conduisent pas systématiquement vers des conclusions évidentes. Pour cette raison, le statisticien a recours à des méthodes plus formelles et plus convaincantes qu'un simple graphique.
- Ainsi, de nombreuses applications de la statistique sont de type **inférentiel**.

- Parmi les types d'inférence statistique les plus utilisés se trouvent
  - 1. les **intervalles de confiance**, notion introduite au chapitre précédent. Rappelons que les intervalles de confiance sont appropriés pour **estimer** un paramètre de la population;
  - 2. les **tests d'hypothèses**, concept statistique qui sera traité dans ce chapitre. Les tests d'hypothèses permettent, à partir de valeurs observées, de **prendre une décision** sur le refus ou non d'une hypothèse statistique portant sur la population.



*En route pour les tests d'hypothèses ! Mais en douceur...*

## 10.2 Tests d'hypothèses

La démarche pour tester une hypothèse statistique sera présentée à l'aide d'un exemple réel. Les nouveaux concepts statistiques seront introduits au fur et à mesure que la résolution du problème exposé dans l'exemple progresse.

### Exemple 1 :

les responsables du service des eaux d'une ville souhaitent mener une étude pour analyser la concentration de nitrate contenu dans l'eau courante. Ils se demandent si divers facteurs ont pu provoquer une augmentation significative de concentration par rapport à la concentration espérée de  $0.492 \mu\text{g/ml}$

→ on va procéder à une inférence sur la "vraie" espérance de la population

### Remarques :

→ deux hypothèses sont à formuler :

- ▷ l'**hypothèse nulle**  $H_0 : \mu = \mu_0 = 0.492 \mu\text{g/ml}$   
(la concentration est restée constante)
- ▷ l'**hypothèse alternative**  $H_1 : \mu > \mu_0 = 0.492 \mu\text{g/ml}$   
(la concentration a augmenté)

→ l'hypothèse nulle  $H_0$  sera testée contre l'hypothèse alternative  $H_1$ .

- les hypothèses  $H_0$  et  $H_1$  sont choisies en fonction des faits que l'expérimentateur souhaite démontrer. Dans l'exemple 1, l'hypothèse à tester est  $H_0 : \mu = 0.492 \mu\text{g/ml}$ . En général, cette hypothèse doit indiquer une position que l'on désire réfuter sur la base de l'expérience réalisée. Cette stratégie s'explique par le fait qu'il est impossible d'être certain qu'une hypothèse soit correcte; en revanche, on peut évaluer l'évidence que montrent les valeurs observées contre l'hypothèse nulle.

En résumé, l'objectif d'un **test statistique dit de signification** consiste à quantifier l'évidence contre l'hypothèse nulle, cette hypothèse indiquant que les données observées ne présentent rien d'inhabituel;

- les hypothèses seront toujours définies avant de voir les données;

### Remarques (suite) :

- l'hypothèse nulle est communément une égalité comme par exemple  
 $H_0 : \mu = \mu_0$ ;
- il existe deux types d'hypothèses alternatives :
  - a) celles qui ne considèrent qu'une direction de comparaison comme par exemple  $H_1 : \mu < \mu_0$  ou  $H_1 : \mu > \mu_0$   
 ↵ **test unilatéral**;
  - b) celles qui envisagent les deux directions possibles d'une comparaison comme par exemple  $H_1 : \mu \neq \mu_0$   
 ↵ **test bilatéral**;
- le choix du type d'alternatives doit s'effectuer avant de visualiser les données; en cas de doute, il est conseillé d'utiliser une alternative bilatérale.

Comment teste-t-on l'hypothèse nulle  $H_0$  contre l'hypothèse alternative  $H_1$  ?

- Une fois les hypothèses nulle et alternative soigneusement définies, il nous reste à trouver une procédure pour tester l'hypothèse nulle, plus précisément, pour évaluer l'évidence contre l'hypothèse nulle que contiennent les valeurs observées.
- L'outil de base pour mesurer cette évidence est une

**statistique de test**.

Le choix de la statistique de test dépend en principe du paramètre que l'on souhaite tester et du type de données.

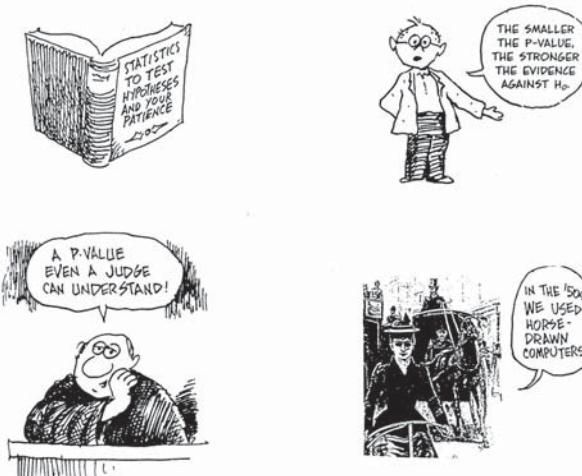
- Si on observe un écart significatif entre la valeur de la statistique et la valeur hypothétique dans la direction de l'hypothèse alternative, on dispose de suffisamment d'évidence contre l'hypothèse nulle.
- Pour évaluer l'évidence contre l'hypothèse nulle contenue dans l'échantillon, deux outils sont utilisés : les **probabilités** (évidemment !!!) et la **distribution d'échantillonnage de la statistique de test**.
- La statistique de test est une fonction de l'échantillon et de certaines valeurs connues. Ainsi, sa distribution d'échantillonnage est entièrement déterminée quand l'hypothèse nulle est vraie.

- Supposons que la réalisation de la statistique de test soit éloignée de la valeur théorique, valeur déterminée à l'aide de la distribution d'échantillonnage de la statistique de test sous l'hypothèse  $H_0$ . Deux issues sont alors possibles :
  - ↝ un échantillon étrange a été observé par chance;
  - ↝ l'hypothèse nulle est fausse; ainsi la distribution d'échantillonnage de la statistique de test n'est pas celle à laquelle on s'y attendait.
- Il est impossible de savoir avec certitude laquelle de ces deux issues est vraie en réalité. Cependant, on peut les évaluer en **probabilités**
- ↝ utilisation de la **p-valeur**.

- La  $p$ -valeur est la probabilité sous l'hypothèse nulle d'observer la statistique de test comme étant au moins aussi extrême (en général, aussi grande ou aussi petite) que la valeur observée à partir des données.

- Une petite  $p$ -valeur

- ▷ indique que la réalisation de la statistique de test se situe dans les queues de la distribution dont elle serait issue si l'hypothèse nulle était vraie
- ▷ fait improbable sous cette hypothèse;
- ▷ entraîne alors une contradiction entre les données et l'hypothèse nulle;
- ▷ implique ainsi une évidence contre l'hypothèse nulle.



*Utilisation de la  $p$ -valeur dans la théorie des tests d'hypothèses.*

- En pratique, une bonne règle pour qualifier l'évidence contre l'hypothèse nulle est résumée dans le tableau ci-dessous :

| $p$ -valeur                     | évidence contre $H_0$ |
|---------------------------------|-----------------------|
| $p$ -valeur $\geq 0.10$         | aucune                |
| $0.10 > p$ -valeur $\geq 0.05$  | faible                |
| $0.05 > p$ -valeur $\geq 0.01$  | modérée               |
| $0.01 > p$ -valeur $\geq 0.001$ | forte                 |
| $p$ -valeur $< 0.001$           | très forte            |

- Si la  $p$ -valeur d'un test est inférieure à une valeur  $\alpha$  spécifiée, le test est dit **statistiquement significatif à un niveau de signification  $\alpha$** . L'hypothèse nulle  $H_0$  est alors rejetée à un niveau de signification  $\alpha$ .
- $\alpha$ , choisi par l'expérimentateur **avant** la collecte des données, est appelé le **niveau ou seuil de signification** du test. Il s'agit en fait du niveau d'évidence prédéfini qui pousserait l'expérimentateur à rejeter l'hypothèse nulle. En se fixant des objectifs exploratoires, un niveau de 5 % ou même de 10 % peut être adéquat alors que pour des buts confirmatoires, il est préférable d'utiliser un niveau plus bas, par exemple 1 % ou moins.
- Dans vos rapports, ne vous contentez pas d'indiquer le rejet ou non de l'hypothèse nulle; précisez toujours le niveau de signification  $\alpha$  !

- En résumé,

procédure pour tester une hypothèse statistique

1. formuler soigneusement l'hypothèse nulle  $H_0$  et l'hypothèse alternative  $H_1$ ;
2. fixer le seuil de signification  $\alpha$ ;
3. choisir une statistique de test appropriée  $T = T(X_1, \dots, X_n)$  et déterminer sa distribution sous l'hypothèse  $H_0$ ;
4. recueillir les valeurs observées de l'échantillon  $x_1, \dots, x_n$ ;
5. calculer la valeur observée de la statistique de test  $t = T(x_1, \dots, x_n)$ ;
6. évaluer la  $p$ -valeur;
7. déterminer la masse d'évidence contre l'hypothèse nulle  $H_0$  et conclure en fonction de la  $p$ -valeur.

### Exemple 1 (suite) :

- dix échantillons d'eau ont été soigneusement prélevés et analysés chimiquement en laboratoire. La concentration en nitrate mesurée en  $\mu\text{g/ml}$  a été relevée :

0.513 0.524 0.529 0.481 0.492 0.499 0.518 0.490 0.494 0.501

Source : Wild C.J. & Seber G.A.F. (2000). *Chance Encounter, A First Course in Data Analysis and Inference*. NY:Wiley, p.410.

- avant d'appliquer la statistique de test aux données, il faut vérifier leur normalité hypothétique. Le graphique quantiles versus quantiles des valeurs observées montre que les points sont relativement proches de la droite. Par conséquent, bien que le nombre de données soit petit (prudence recommandée dans les conclusions), le graphique ne contre-indique pas le fait que les données puissent provenir d'une distribution normale, constat identique sur le diagramme en points.

### Exemple 1 (suite) :

- l'estimateur standard du paramètre  $\mu$  est  $\bar{X}$ . En supposant que les valeurs observées proviennent d'une distribution normale (hypothèse à vérifier siège les données recueillies), on a  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$

→ la statistique de test pour résoudre le problème est donc la version standardisée de  $\bar{X}$ ,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}},$$

où  $\sigma$ , paramètre inconnu, est remplacé par son estimateur standard  $S$ ;

→ sous l'hypothèse  $H_0$ , nous savons (voir le chapitre traitant les intervalles de confiance) que la statistique de test  $T$  est de distribution de Student  $t$  à  $n - 1$  degrés de liberté.

### Exemple 1 (suite) :

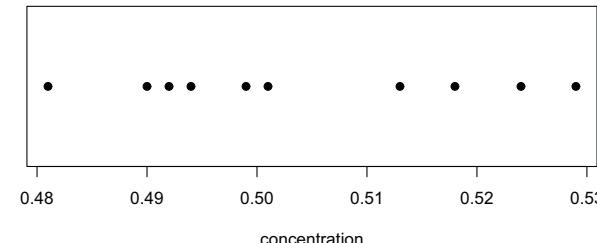
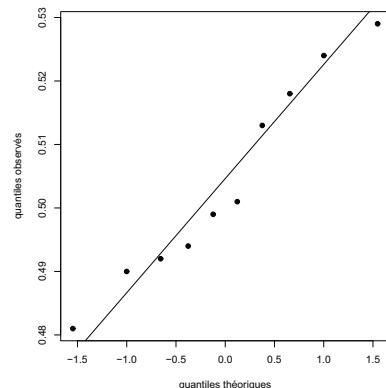


Diagramme en points des valeurs observées.

## Exemple 1 (suite) :



Graphique quantiles versus quantiles des valeurs observées.



## Exemple 1 (suite) :

- en utilisant les données, la valeur observée  $t$  de la statistique de test  $T$  vaut, sous l'hypothèse nulle  $H_0 : \mu = \mu_0 = 0.492 \text{ } \mu\text{g/ml}$ ,

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.5041 - 0.492}{0.016/\sqrt{10}} = 2.391;$$

- dans notre cas, la  $p$ -valeur est la probabilité sous l'hypothèse nulle d'observer la statistique de test  $T$  comme étant au moins aussi grande que celle observée en utilisant les données. Ainsi, elle vaut

$$p\text{-valeur} = P(T \geq t = 2.391 \mid H_0) = 0.02025;$$

comme la  $p$ -valeur est plus petite que le niveau de signification  $\alpha$ , fixé préalablement à 0.05 avant la collecte des échantillons, nous disposons suffisamment d'évidence contre l'hypothèse nulle  $H_0 : \mu = 0.492$  qui est donc rejetée au niveau de signification de 5 %.

## Exemple 1 (suite) :

- l'intervalle de confiance unilatéral à 95 % pour la statistique de test  $T$  est  $(-\infty, t_{n-1, 1-\alpha}) = (-\infty, 1.8331)$
- cet intervalle de confiance est appelée la **région de non rejet** de l'hypothèse nulle. La **région de rejet** est  $(1.8331, \infty)$ . Les valeurs qui marquent la limite entre la région de rejet et de non rejet sont les **valeurs critiques**;
- comme la valeur observée  $t = 2.391$  de la statistique de test n'est pas comprise dans l'intervalle de confiance, l'hypothèse nulle est rejetée au niveau de signification de 5 %;
- le  $t$ -intervalle de confiance unilatéral à 95 % pour  $\mu$  vaut  $(\bar{x} - t_{n-1, 1-\alpha} s/\sqrt{n}, \infty) = (0.49482, \infty)$ . Comme la valeur  $\mu_0 = 0.492$  n'est pas couverte par l'intervalle, il s'ensuit que l'hypothèse nulle  $H_0 : \mu = \mu_0 = 0.492$  est rejetée au niveau de signification de 5 %.

### Exemple 1 (suite) :

▷ instructions utilisées dans le logiciel de statistique R

```
>attach(nitrate)
>t.test(concentration,mu=0.492,alternative="greater",conf.level=0.95)

▷ sortie

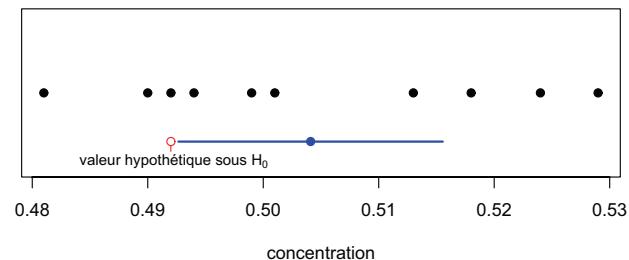
t = 2.391, df = 9, p-value = 0.02025
alternative hypothesis: true mean is greater than 0.492
95 percent confidence interval:
0.49482      Inf
sample estimates:
mean of x
0.5041
```

### Remarques (suite) :

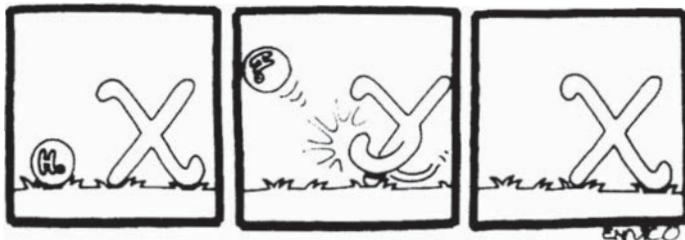
- dualité entre les intervalles de confiance et les tests d'hypothèses :  
un test de signification bilatéral rejette l'hypothèse nulle  $H_0 : \mu = \mu_0$  à un niveau de signification  $\alpha$  précisément lorsque la valeur  $\mu_0$  tombe hors de l'intervalle de confiance pour  $\mu$  au seuil de confiance  $1 - \alpha$ ;
- au cas où l'hypothèse nulle  $H_0 : \mu = \mu_0$  ne peut pas être rejetée à un niveau de signification  $\alpha$ , toutes les valeurs de l'intervalle de confiance pour  $\mu$  au seuil de confiance  $1 - \alpha$  sont plausibles. Alors, toutes les valeurs de l'intervalle de confiance pourraient être la vraie valeur de  $\mu_0$  ! Ainsi, on comprend mieux la raison pour laquelle on cherche à refuser l'hypothèse nulle. De plus, ne dites jamais qu'on accepte l'hypothèse  $H_0$ . Cette affirmation va clairement à l'encontre de la méthodologie des tests de signification.

### Remarque : autre interprétation de la statistique de test

- pour tester une hypothèse nulle  $H_0$ , on se propose d'observer l'"écart" entre les données récoltées et l'hypothèse nulle au lieu de déterminer l'évidence contre l'hypothèse nulle. Cet écart est mesuré par la statistique de test. En général, si la valeur de cette statistique calculée à partir des données est grande (ou petite suivant les cas), l'écart est significatif, tout comme l'hypothèse nulle. Ainsi, on rejette  $H_0$ . La chance d'avoir obtenu un tel écart, en supposant que l'hypothèse nulle soit vraie, est évaluée par la  $p$ -valeur;
- les deux démarches, l'une reposant sur l'"écart" entre les données recueillies et l'hypothèse nulle, l'autre sur l'évidence contre cette même hypothèse sont tout à fait équivalentes.



*Dualité entre les intervalles de confiance et les tests de signification bilatéraux.*



On cherche à refuser l'hypothèse nulle !

### Remarques (suite) :

- supposons qu'une valeur atypique a été observée. On se demande alors quelle sera son influence sur un test d'hypothèse. Nous allons répondre à cette question en reprenant l'exemple 1 et en remplaçant la plus grande valeur observée 0.529 par 0.581. On se propose de tester  $H_0 : \mu = 0.492$  contre  $H_1 : \mu \neq 0.492$ . La nouvelle donnée est incontestablement une valeur atypique.

- ▷ L'écart entre la moyenne du nouvel échantillon (0.5093) et la "vraie" valeur  $\mu_0 = 0.492$  a clairement augmenté. On pourrait alors penser que l'hypothèse  $H_0$  pourrait être plus fortement réfutée. Ce n'est justement pas le cas !
- ▷ En passant de 0.0160 à 0.0285, l'écart-type de l'échantillon a augmenté. Cette modification a généré une *p*-valeur plus élevée ainsi qu'un élargissement de l'intervalle de confiance pour  $\mu$ .

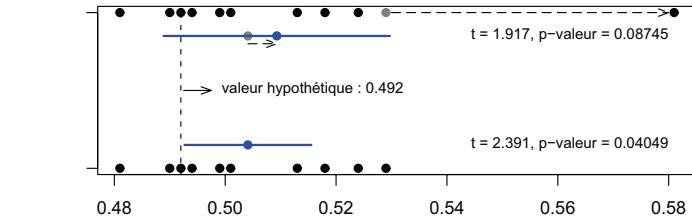


Diagramme en points des données originales (en bas) et des données modifiées (en haut) pour illustrer l'effet d'une valeur atypique sur le test et sur l'intervalle de confiance.

- ▷ Sortie avec R en utilisant les données originales

```
t = 2.391, df = 9, p-value = 0.04049
alternative hypothesis: true mean is not equal to 0.492
95 percent confidence interval:
0.49265      0.51555
mean of x
0.5041
```

☞ l'hypothèse nulle est rejetée au niveau de signification de 5 %.

▷ Sortie avec R en utilisant les données modifiées

```
t = 1.9172, df = 9, p-value = 0.08745
alternative hypothesis: true mean is not equal to 0.492
95 percent confidence interval:
0.48889      0.52971
sample estimates:
mean of x
0.5093
```

- ~> l'hypothèse nulle n'est pas rejetée au niveau de signification de 5%;
- ~> l'influence d'une valeur atypique sur un test peut être très importante !

## 10.3 Quelques tests d'hypothèses

Dans ce paragraphe, nous nous proposons de présenter des tests s'appliquant aux paramètres d'une ou plusieurs populations ainsi que les deux tests du khi carré (le test d'adéquation d'un modèle théorique à des données observées et le test d'indépendance de deux variables catégoriques).

### 10.3.1 Tests sur les paramètres de lois normales

- Considérons  $n$  variables aléatoires indépendantes  $X_1, X_2, \dots, X_n$  issues d'une distribution normale d'espérance  $\mu$  et de variance  $\sigma^2$ , i.e  

$$X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2).$$

### Conseil :

construire un graphique adéquat avant d'utiliser des méthodes formelles d'analyse (intervalles de confiance et tests d'hypothèses). En effet, les graphiques

- forment l'outil le plus simple pour déceler des informations contenues dans les données;
- révèlent souvent des particularités surprenantes et très intéressantes;
- permettent d'éviter l'utilisation de méthodes inappropriées ainsi que des conclusions infondées;
- jouent un rôle central pour vérifier les hypothèses nécessaires dans l'application de méthodes formelles, par exemple, pour vérifier la normalité de l'échantillon (nous y reviendrons en détail dans le paragraphe suivant).

- Inférence pour l'espérance  $\mu$  de la population :

| $H_0$         | Statistique de test                                                | $H_1$                                              | Région de rejet                                                                                                         |
|---------------|--------------------------------------------------------------------|----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| $\mu = \mu_0$ | $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$<br>$\sigma$ connu | $\mu < \mu_0$<br>$\mu > \mu_0$<br>$\mu \neq \mu_0$ | $Z \leq c_{\alpha}$<br>$Z \geq c_{1-\alpha}$<br>$Z \leq c_{\alpha/2}$ et<br>$Z \geq c_{1-\alpha/2}$                     |
| $\mu = \mu_0$ | $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}},$<br>$\sigma$ inconnu    | $\mu < \mu_0$<br>$\mu > \mu_0$<br>$\mu \neq \mu_0$ | $T \leq t_{n-1, \alpha}$<br>$T \geq t_{n-1, 1-\alpha}$<br>$T \leq t_{n-1, \alpha/2}$ et<br>$T \geq t_{n-1, 1-\alpha/2}$ |

- ~> hypothèse :  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2).$

- Comparaison de deux espérances  $\mu_1$  et  $\mu_2$  :

| $H_0$                 | Statistique de test                                                                                                                                                                  | $H_1$                                                                      | Région de rejet                                                                                                                                 |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| $\mu_1 - \mu_2 = d_0$ | $Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}},$<br>$\sigma_1$ et $\sigma_2$ connus                                                           | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$ | $Z \leq c_\alpha$<br>$Z \geq c_{1-\alpha}$<br>$Z \leq c_\alpha/2$ et<br>$Z \geq c_{1-\alpha}/2$                                                 |
| $\mu_1 - \mu_2 = d_0$ | $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_p \sqrt{(1/n_1) + (1/n_2)}},$<br>$\sigma_1 = \sigma_2$ et inconnus,<br>$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$ | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$ | $T \leq t_{n_1+n_2-2, \alpha}$<br>$T \geq t_{n_1+n_2-2, 1-\alpha}$<br>$T \leq t_{n_1+n_2-2, \alpha/2}$ et<br>$T \geq t_{n_1+n_2-2, 1-\alpha/2}$ |

~ hypothèses :  $X_{1;1}, \dots, X_{1;n_1} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ ,  
 $X_{2;1}, \dots, X_{2;n_2} \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$  et indépendance.

- Comparaison de deux espérances  $\mu_1$  et  $\mu_2$  (suite) :

| $H_0$                 | Statistique de test                                                                                                                                                                                               | $H_1$                                                                      | Région de rejet                                                                                                         |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| $\mu_1 - \mu_2 = d_0$ | $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}},$<br>$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{(S_1^2/n_1)^2/n_1 - 1 + (S_2^2/n_2)^2/n_2 - 1},$<br>$\sigma_1 \neq \sigma_2$ et inconnus | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$ | $T \leq t_{\nu, \alpha}$<br>$T \geq t_{\nu, 1-\alpha}$<br>$T \leq t_{\nu, \alpha/2}$ et<br>$T \geq t_{\nu, 1-\alpha/2}$ |
| $\mu_1 - \mu_2 = d_0$ | $T = \frac{\bar{D} - d_0}{S_d / \sqrt{n}},$<br>observations appariées                                                                                                                                             | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$ | $T \leq t_{n-1, \alpha}$<br>$T \geq t_{n-1, 1-\alpha}$<br>$T \leq t_{n-1, \alpha/2}$ et<br>$T \geq t_{n-1, 1-\alpha/2}$ |

~ hypothèses : normalité et indépendance pour le premier cas, dépendance pour le second.

- Inférence pour la variance  $\sigma^2$  de la population :

| $H_0$                     | Statistique de test                   | $H_1$                                                                                  | Région de rejet                                                                                                                                          |
|---------------------------|---------------------------------------|----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\sigma^2 = \sigma_0^2$   | $X^2 = \frac{(n-1) S^2}{\sigma_0^2},$ | $\sigma^2 < \sigma_0^2$<br>$\sigma^2 > \sigma_0^2$<br>$\sigma^2 \neq \sigma_0^2$       | $X^2 \leq \chi_{n-1, \alpha}^2$<br>$X^2 \geq \chi_{n-1, 1-\alpha}^2$<br>$X^2 \leq \chi_{n-1, \alpha/2}^2$ et<br>$X^2 \geq \chi_{n-1, 1-\alpha/2}^2$      |
| $\sigma_1^2 = \sigma_2^2$ | $F = \frac{S_1^2}{S_2^2},$            | $\sigma_1^2 < \sigma_2^2$<br>$\sigma_1^2 > \sigma_2^2$<br>$\sigma_1^2 \neq \sigma_2^2$ | $F \leq F_{n_1-1, n_2-1, \alpha}$<br>$F \geq F_{n_1-1, n_2-1, 1-\alpha}$<br>$F \leq F_{n_1-1, n_2-1, \alpha/2}$<br>$F \geq F_{n_1-1, n_2-1, 1-\alpha/2}$ |

~ hypothèses : normalité pour les deux cas et indépendance pour le second cas.

### Remarques :

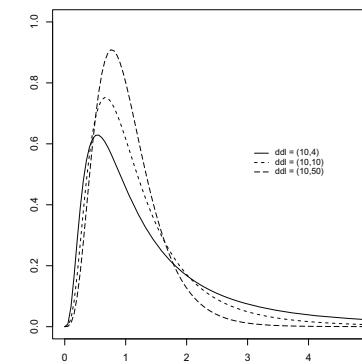
- la distribution  $F$  utilisée pour tester l'hypothèse  $\sigma_1^2 = \sigma_2^2$  mérite tout naturellement quelques précisions étant donné qu'elle n'a pas encore été traitée dans le cours :
  - la distribution  $F$  constitue l'une des contributions à la statistique du biologiste et mathématicien britannique R. A. Fisher (1890–1962) qui l'a introduite pour comparer les espérances de plusieurs populations;
  - supposons que  $\sigma_1^2$  et  $\sigma_2^2$  sont les variances (inconnues) de deux populations et considérons  $S_1^2$  et  $S_2^2$  leurs estimateurs standards respectifs calculés sur deux échantillons indépendants de tailles respectives  $n_1$  et  $n_2$ . On peut montrer, si  $\sigma_1^2 = \sigma_2^2$  et sous hypothèse de normalité, que la variable aléatoire  $F = S_1^2/S_2^2$  est issue d'une distribution  $F$  de Fisher à deux paramètres  $n_1 - 1$  et  $n_2 - 1$  notée  $F_{n_1-1, n_2-1}$ .

### Remarques (suite) :

- remarques sur la distribution  $F$  de Fisher (suite) :

- ▷ les paramètres  $\nu_1$  et  $\nu_2$  d'une variable aléatoire provenant d'une distribution  $F$  sont aussi appelés **degrés de liberté**;
- ▷ la distribution  $F$  peut être vue comme une extension de la distribution de Student. En effet, la distribution d'une variable aléatoire  $t^2$ , où  $t$  est une variable aléatoire de Student à  $\nu$  degrés de liberté, est  $F_{1,\nu}$ ;
- ▷ relation intéressante pour la lecture de la table des quantiles :

$$F_{\nu_1, \nu_2, \alpha} = \frac{1}{F_{\nu_2, \nu_1, 1-\alpha}}.$$



*Illustration de la fonction de densité de variables aléatoires issues de distribution  $F$  à différents degrés de liberté.*

### Remarques (suite) :

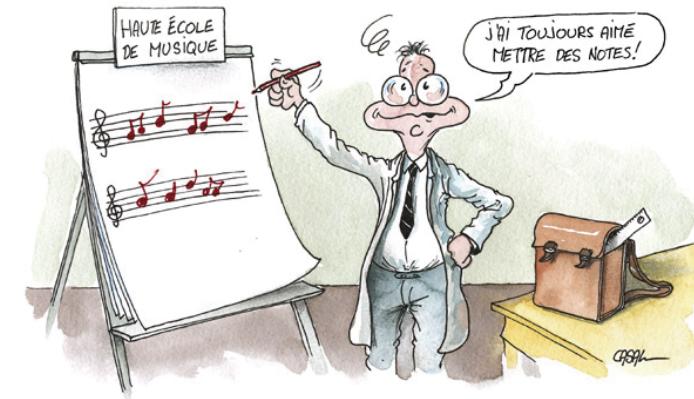
- les tests d'hypothèses décrits dans ce paragraphe sont exacts uniquement si la distribution des variables est normale;
- cependant, ces procédures restent valables au cas où l'hypothèse de normalité est légèrement violée. Les résultats ne seront pas sérieusement affectés si la distribution est proche de la distribution normale. D'ailleurs, si aucune valeur atypique évidente n'a été décelée ou si la distribution ne présente pas une forte asymétrie, les procédures demeurent relativement bonnes.

### Remarques (suite) :

- il va sans dire qu'avant d'entreprendre l'un des tests exposés dans ce paragraphe, l'hypothèse de normalité doit être vérifiée. Deux méthodes sont à disposition : un graphique quantiles versus quantiles et un test formel de normalité, citons par exemple le test de Shapiro-Wilk (commande `shapiro.test` dans la librairie `stats` du logiciel R);
- pour comparer les espérances de deux populations, certains tests présentés ci-dessus exigent l'égalité des variances des échantillons. Pour le vérifier, on peut utiliser le test décrit en page 39. En pratique, si les longueurs des boîtes à moustaches placées en parallèle sont relativement proches l'une de l'autre, on se permet souvent d'admettre l'égalité des variances.

### Exemple 2 :

- les étudiants des deux classes de 2ème année (classe A et classe B) de la filière informatique, orientation logiciel, de l'EIVD-Yverdon ont effectué le même travail écrit de probabilités et statistique le 3 juillet 2003. Les notes obtenues sont comprises entre 1 (note minimale) et 6 (note maximale)
- le professeur de statistique se demande s'il existe une différence significative entre les deux classes.

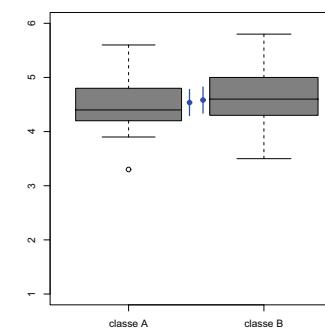


Dessin de Casal, dessinateur de presse.

### Exemple 2 (suite) :

- les boîtes à moustaches en parallèle et les intervalles de confiance à 95 % pour les espérances des notes obtenues par les étudiants n'indiquent pas une différence évidente entre les médianes et les espérances des deux classes. En observant la longueur des boîtes, la variance du deuxième échantillon semble légèrement plus grande que celle du premier
- test  $t$  de comparaison des espérances de deux populations.

### Exemple 2 (suite) :

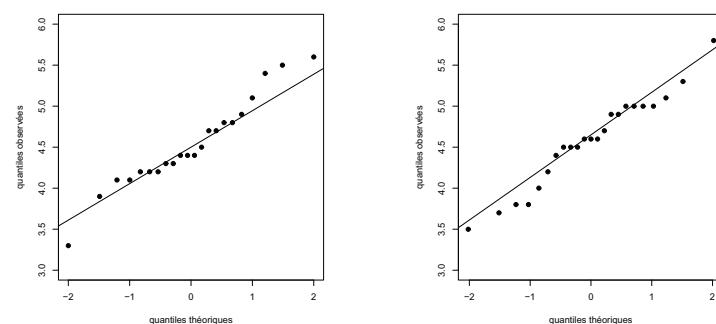


Boîtes à moustaches en parallèle des résultats  
obtenus par les deux classes.

### Exemple 2 (suite) :

- avant de procéder au test formel, vérifions ses conditions d'application :
  - bien que les tailles des échantillons soient relativement petites et que la boîte à moustaches révèle une légère asymétrie pour la classe A, les graphiques quantiles versus quantiles montrent que le modèle normal est raisonnable pour les deux échantillons;
  - les variances des notes selon les classes valent respectivement  $s_A^2 = 0.30$  et  $s_B^2 = 0.31$ . Comme ces valeurs sont assez proches, on admet la plausible égalité des variances de la population. Notons qu'on aurait pu recourir au test formel décrit en page 39 pour vérifier cette condition d'application
- pour tester l'hypothèse nulle  $H_0 : \mu_A - \mu_B = d_0 = 0$  contre l'hypothèse alternative  $H_1 : \mu_A - \mu_B \neq 0$ , on utilise un  $t$  test avec  $\sigma_A = \sigma_B$ ,  $\sigma_A$  et  $\sigma_B$  inconnus.

### Exemple 2 (suite) :



Graphiques quantiles versus quantiles; classe A (gauche), classe B (droite).

### Exemple 2 (suite) :

- la statistique de test  $T$  est donnée par
 
$$T = \frac{\bar{X}_A - \bar{X}_B}{S_p \cdot \sqrt{(1/n_A) + (1/n_B)}},$$
 où  $S_p^2$  est le "pooled" estimateur de la variance commune et inconnue  $\sigma^2$ ;
- le professeur de statistique avait décidé d'utiliser un niveau de signification de 5%;
- par les données, on obtient,  $n_A = 22$ ,  $\bar{x}_A = 4.54$ ,  $n_B = 23$ ,  $\bar{x}_B = 4.58$  et  $s_p^2 = 0.31$
- la valeur observée  $t$  de la statistique de test  $T$  vaut  $t = -0.28027$ .

### Exemple 2 (suite) :

- la valeur  $t$  doit être comparée au quantile de la distribution de Student à  $n_A + n_B - 2 = 22 + 23 - 2 = 43$  degrés de liberté. Comme  $t_{43, 0.025} = -2.0167 (= -t_{43, 0.975}) < t < t_{43, 0.975} = 2.0167$ , la valeur observée  $t$  de la statistique de test ne tombe pas dans la région de rejet;
- le logiciel R a fourni les résultats suivants :

```
t = -0.2803, df = 43, p-value = 0.7806
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.37900    0.28651
sample estimates:
mean of x mean of y
4.5364        4.5826
```

### Exemple 2 (suite) :

- la  $p$ -valeur du test vaut 0.7806. Comme elle est nettement supérieure au niveau de signification de 0.05, on conclut qu'on ne dispose d'aucune évidence pour rejeter l'hypothèse nulle selon laquelle il n'existe pas de différence entre les espérances des notes des deux classes.

En résumé, l'hypothèse nulle ne peut pas être rejetée à un niveau de signification de 5 %.

*“Enseignez le test de Student à quelqu'un et il sera heureux la journée durant; enseignez le principe de la régression à quelqu'un et il sera heureux pour la semaine à venir; enseignez la statistique à quelqu'un et il aura durant toute sa vie des problèmes.”*

Un statisticien inconnu

### 10.3.2 Test d'adéquation d'un modèle théorique

**Objectif** : tester l'hypothèse selon laquelle des valeurs observées proviennent d'une distribution théorique spécifiée.

**Méthode :**

- partager l'axe réel en  $k$  intervalles  $I_1, I_2, \dots, I_k$  appelés **classes**. Pour chaque classe  $I_j$ , notons  $O_j$  les **fréquences observées** dans l'intervalle  $I_j$  i.e le nombre d'observations qui se trouvent dans  $I_j$ ;
- en utilisant la distribution de probabilités théorique, calculer les probabilités  $p_j = P(X \in I_j)$ ,  $j = 1, \dots, k$ ,  $X$  étant une variable aléatoire issue de cette distribution. On en déduit les **fréquences théoriques**  $e_j = n \cdot p_j$ ,  $j = 1, \dots, k$ , où  $n$  représente le nombre total de valeurs observées;

- calculer la réalisation de la statistique de test

$$T = \sum_{j=1}^k \frac{(O_j - e_j)^2}{e_j},$$

qui est une mesure de l'écart entre les deux distributions de fréquences.

- ▷ Sous l'hypothèse  $H_0$ , la distribution de cette statistique peut être approchée par une loi khi carré  $\chi^2$  à  $k - 1$  degrés de liberté si le nombre espéré  $e_j$  dans chacune des classes est suffisamment grand ( $\geq 5$ ). Si ce n'est pas le cas, il faut diviser différemment l'axe réel.
- ▷ Au cas où  $r$  paramètres inhérents à la distribution théorique ont dû être estimés,  $r$  degrés de liberté supplémentaires doivent être soustraits  $\rightsquigarrow k - r - 1$  degrés de liberté au total.

Remarques :

- la statistique de test a été introduite par le mathématicien britannique K. Pearson (1857–1936) qui a aussi établi la théorie générale de la corrélation;
- cette statistique peut aussi s'écrire

$$T = \sum_{j=1}^k \frac{O_j^2}{e_j} - n.$$

Cette formule est souvent plus commode à utiliser.



Wayne Gretzky, le plus célèbre 99 de la NHL.

Exemple 3 :

- Wayne Gretzky est probablement le meilleur compteur qu'a connu jusqu'à nos jours la ligue nord-américaine de hockey-sur-glace (NHL, National Hockey League);
- désignons par  $X$  le nombre de points par match (buts + assists) inscrits par Gretzky lors de sa première saison, 1979–1980, en NHL, saison durant laquelle il a disputé 79 parties. Les fréquences observées figurent dans le tableau deux pages plus loin.

Exemple 3 (suite) :

| nombre de points par match<br>inscrits par Gretzky | fréquences observées |
|----------------------------------------------------|----------------------|
| $X = x_j$                                          | $o_j$                |
| 0                                                  | 17                   |
| 1                                                  | 21                   |
| 2                                                  | 22                   |
| 3                                                  | 10                   |
| 4+                                                 | 9                    |

Fréquences observées du nombre de points par match  
inscrits par Wayne Gretzky lors la saison 1979–1980.

### Exemple 3 (suite) :

- on se demande si les valeurs observées auraient été générées par une distribution de Poisson de paramètre  $\lambda$

~ l'estimation de  $\lambda$  obtenue par la méthode du maximum de vraisemblance est

$$\hat{\lambda} = \bar{x} = 1.734;$$

~ sous l'hypothèse d'une distribution de Poisson, les probabilités théoriques valent

$$p_j = P(X = x_j) = e^{-1.734} \cdot \frac{1.734^{x_j}}{x_j!}, \quad x_j = 0, 1, 2, \dots$$

On en déduit les fréquences théoriques  $e_j = n \cdot p_j$  qui figurent dans le tableau de la page suivante.

### Exemple 3 (suite) :

| nombre de points par match<br>inscrits par Gretzky | probabilités<br>théoriques | fréquences théoriques |
|----------------------------------------------------|----------------------------|-----------------------|
| $X = x_j$                                          | $p_j$                      | $e_j = n \cdot p_j$   |
| 0                                                  | 0.177                      | 13.95                 |
| 1                                                  | 0.306                      | 24.17                 |
| 2                                                  | 0.265                      | 20.97                 |
| 3                                                  | 0.153                      | 12.12                 |
| 4+                                                 | 0.098                      | 7.77                  |
|                                                    |                            | 1                     |
|                                                    |                            | 79                    |

Fréquences théoriques des points par match  
inscrits par Wayne Gretzky lors la saison 1979–1980.

### Exemple 3 (suite) :

- la valeur observée de la statistique de test de type khi carré vaut

$$t = 1.7042;$$

- comme la  $p$ -valeur du test,

$$P(T > t = 1.7042 \mid H_0) = 0.636,$$

est nettement plus grande que le niveau de signification  $\alpha = 0.05$ , on ne peut pas rejeter l'hypothèse  $H_0$ .

Le modèle de Poisson est donc plausible à ce niveau de signification.

*"A statistician is a person whose lifetime ambition is to be wrong 5 percent of the time."*

Anonymous

Remarques :

- comme on a été contraint d'estimer le paramètre de la distribution de Poisson, le nombre de degrés de liberté de la distribution  $\chi^2$  sous l'hypothèse nulle est  $5 - 1 - 1 = 3$ ;
- dans ces tests d'adéquation, aucune hypothèse alternative précise n'est formulée. Les seules décisions possibles sont : rejet ou non rejet de l'hypothèse  $H_0$  où  $H_0$  indique qu'il n'y a aucune différence entre les distributions des fréquences théoriques et observées.

- Un tableau de contingence se présente de la manière suivante :

|       |               | B             |          |               |          |               |                       |              |
|-------|---------------|---------------|----------|---------------|----------|---------------|-----------------------|--------------|
|       |               |               |          |               |          |               |                       |              |
| A     | 1             | $n_{11}$      | $n_{12}$ | ...           | $n_{1j}$ | ...           | $n_{1c}$              | $n_{1\cdot}$ |
|       | 2             | $n_{21}$      | $n_{22}$ | ...           | $n_{2j}$ | ...           | $n_{2c}$              | $n_{2\cdot}$ |
| :     | :             | :             | :        | :             | :        | :             | :                     | :            |
| i     | $n_{i1}$      | $n_{i2}$      | ...      | $n_{ij}$      | ...      | $n_{ic}$      | $n_{i\cdot}$          |              |
| :     | :             | :             | :        | :             | :        | :             | :                     |              |
| r     | $n_{r1}$      | $n_{r2}$      | ...      | $n_{rj}$      | ...      | $n_{rc}$      | $n_{r\cdot}$          |              |
| total | $n_{\cdot 1}$ | $n_{\cdot 2}$ | ...      | $n_{\cdot j}$ | ...      | $n_{\cdot c}$ | $n_{\cdot \cdot} = n$ |              |

▷  $n_{ij}$  : nombre d'individus (ou objets) faisant partie de la classe  $i$  de la variable  $A$  et de la classe  $j$  de la variable  $B$ ;

▷  $n_{i\cdot} = \sum_{j=1}^c n_{ij}$  et  $n_{\cdot j} = \sum_{i=1}^r n_{ij}$ .

### 10.3.3 Test d'indépendance de deux variables catégorielles

- Deux variables catégorielles  $A$  et  $B$ , appelées aussi **facteurs**, sont observées sur des individus ou sur des objects.
- Comme exemples de variables catégorielles, citons le sexe (masculin ou féminin), le pays (Suisse, Brésil, Australie, ...).
- La variable  $A$  forme  $r$  groupes (ou **classes**); la variable  $B$  définit  $c$  groupes (ou **classes**).
- Les fréquences observées sont recueillies dans un tableau, appelé **tableau de contingence**.

### Exemple 4 :

dans le tableau de contingence ci-dessous figurent les résultats d'une enquête réalisée en 1984 aux États-Unis :

| salaire en \$   | satisfaction de son emploi |               |                       |                | total |  |
|-----------------|----------------------------|---------------|-----------------------|----------------|-------|--|
|                 | très insatisfait           | peu satisfait | moyennement satisfait | très satisfait |       |  |
|                 |                            |               |                       |                |       |  |
| < 6'000         | 20                         | 24            | 80                    | 82             | 206   |  |
| 6'000 – 15'000  | 22                         | 38            | 104                   | 125            | 289   |  |
| 15'000 – 25'000 | 13                         | 28            | 81                    | 113            | 235   |  |
| > 25'000        | 7                          | 18            | 54                    | 92             | 171   |  |
| total           | 62                         | 108           | 319                   | 412            | 901   |  |

Source : General Social Survey of the U.S. National Data Programm.

#### Exemple 4 (suite) :

les participants à l'enquête ont été classés selon deux variables catégoriques :

▷ le salaire avec  $r = 4$  classes :

$< \$6'000$ ,  $\$6'000 - \$15'000$ ,  $\$15'000 - \$25'000$  et  $> \$25'000$ ;

▷ la satisfaction de son emploi avec  $c = 4$  classes :

très insatisfait, peu satisfait, moyennement satisfait, très satisfait.

- Dans les tableaux de contingence, on se demande s'il existe une relation (un lien) entre les deux variables catégoriques. En terme d'hypothèse nulle, on se propose de tester si les deux variables sont indépendantes

↝  $H_0$  : il n'existe pas de relation entre les deux variables (entre la variable "ligne" et la variable "colonne")

↝  $H_1$  : une relation existe entre les deux variables.

#### Remarque :

l'hypothèse  $H_1$  ne spécifie pas le type de relation pouvant exister entre les deux variables. Elle peut comprendre différentes sortes de liens. Ainsi, on ne peut pas écrire  $H_1$  comme une égalité ou une inégalité.

- Idée pour tester l'hypothèse  $H_0$  contre l'hypothèse  $H_1$  :

comparer les fréquences observées  $N_{ij}$  avec les fréquences espérées (théoriques)  $e_{ij}$  déterminées sous l'hypothèse d'indépendance des deux variables catégoriques.

- Calcul des fréquences théoriques  $e_{ij}$  :

sous l'hypothèse d'indépendance des deux variables, la probabilité  $p_{ij}$  qu'un individu tombe dans la cellule  $ij$  est égale à

$$p_{ij} \stackrel{\text{indépendance}}{=} \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n^2}.$$

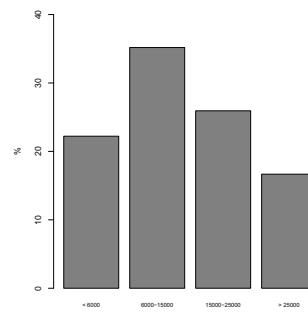
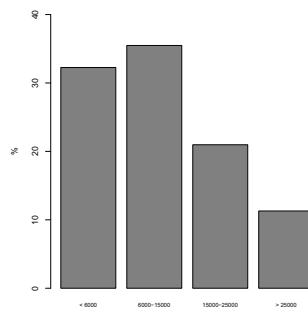
$$\rightsquigarrow e_{ij} = n \cdot p_{ij} = n \cdot \frac{n_{i\cdot} \cdot n_{\cdot j}}{n^2} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}.$$

- Pour tester l'hypothèse d'indépendance, on fait appel à la statistique de test de Pearson

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(N_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(N_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}}.$$

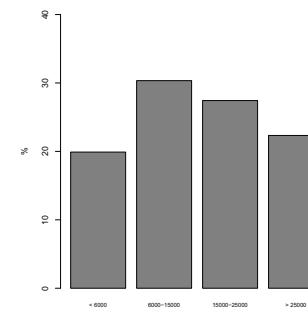
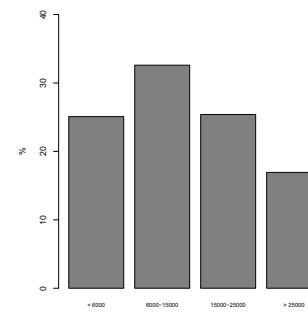
- Sous l'hypothèse  $H_0$ , la distribution de la statistique de test  $T$  peut être approchée, si  $n$  est suffisamment grand, par la distribution khi carré à  $(r - 1) \cdot (c - 1)$  degrés de liberté.

#### Exemple 4 (suite) :



Diagrammes en barres pour les niveaux très insatisfait (gauche)  
et peu satisfait (droite).

#### Exemple 4 (suite) :



Diagrammes en barres pour les niveaux moyenement satisfait (gauche)  
et très satisfait (droite).

#### Exemple 4 (suite) :

▷ instructions utilisées dans le logiciel de statistique R

```
>salsa<-table(salaire,satisfaction)
>summary(salsa)
```

▷ sortie

Number of cases in table: 901

Number of factors: 2

Test for independence of all factors:

Chisq = 12, df = 9, p-value = 0.21

~~> l'hypothèse nulle n'est pas rejetée au niveau de signification de 5%;

~~> on ne dispose pas de suffisamment d'évidence contre l'hypothèse d'indépendance des deux variables catégoriques.

#### Remarques :

- leçon à tirer de l'exemple 4 : pour être plus précis, les variables salaire et satisfaction de son emploi sont des variables qualitatives **ordinaires**. En effet, elles définissent un ordre dans les classes de salaire et dans les degrés de satisfaction. Le test khi carré ne tient pas compte de l'ordre. Des tests plus appropriés (plus puissants) prennent en considération l'ordre dans les variables et, par conséquent, détectent mieux les relations existantes;
- le tableau de contingence est aussi appelé une **table à deux voies**.

## 10.4 Mises en garde

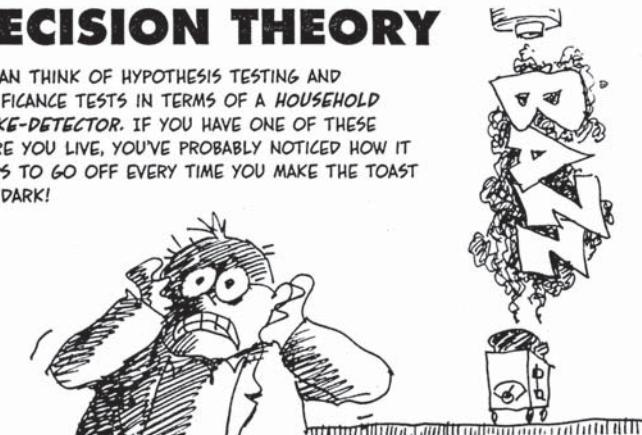
- effectuer un test de signification est souvent très simple en utilisant un bon logiciel (comme R, S-PLUS, STATISTICA ou Minitab). Cependant, comprendre ce que l'on fait est bien plus difficile tout comme interpréter correctement les résultats obtenus;
- toujours vérifier les conditions d'application d'un test de signification ainsi que la qualité des valeurs observées;
- choisir préalablement un niveau de signification  $\alpha$  prend un sens que pour prendre une décision;
- il n'existe pas de frontière précise entre "significatif" et "non significatif". En effet, une différence entre des  $p$ -valeurs de 0.049 et 0.051 ne se justifie pas en pratique. Ainsi, il convient de ne pas considérer 5 % comme une règle universelle définissant ce qui est significatif et ce qui ne l'est pas;
- un test non significatif n'implique pas que l'hypothèse nulle est en réalité vraie;
- ne jamais citer une  $p$ -valeur pour l'existence d'un effet (par exemple une différence entre les espérances de deux populations) en omettant d'indiquer l'intervalle de confiance. En effet, l'intervalle de confiance estime la "grandeur" de la  $p$ -valeur et nous évite de prendre une décision trop hâtive en se basant uniquement sur la  $p$ -valeur;
- la signification statistique n'est pas la même que la signification liée à la pratique et à l'expérience.

## 10.5 Rudiments de la théorie de la décision

- Les principaux concepts de la théorie de la décision en statistique seront présentés en s'inspirant d'un exemple tiré du livre "The Cartoon Guide to Statistics", L. Gonick & W. Smith (1993), HarperCollins.
- Le détecteur d'incendie d'une cuisine est souvent très sensible et réagit probablement dès que, par mégarde, vous avez laissé trop longtemps un toast dans le grille-pain.

## DECISION THEORY

WE CAN THINK OF HYPOTHESIS TESTING AND SIGNIFICANCE TESTS IN TERMS OF A HOUSEHOLD SMOKE-DETECTOR. IF YOU HAVE ONE OF THESE WHERE YOU LIVE, YOU'VE PROBABLY NOTICED HOW IT TENDS TO GO OFF EVERY TIME YOU MAKE THE TOAST TOO DARK!



~ deux types d'erreur :

▷ **erreur de première espèce (ou erreur de type I)** :

l'alarme du détecteur retentit sans qu'un feu se soit déclaré;

▷ **erreur de deuxième espèce (ou erreur de type II)** :

le feu s'est déclaré mais l'alarme ne l'a pas signalé.

- Tout cuisinier sait comment éviter une erreur de première espèce : enlever la batterie du détecteur ! Elle ne sonnera plus, feu ou pas feu !

~ diminution de l'erreur de première espèce mais en contre-partie augmentation de l'erreur de deuxième espèce;

~ inversément, la réduction de l'erreur de deuxième espèce, par exemple en rendant l'alarme extrêmement sensible, peut augmenter l'erreur de première espèce, i.e le nombre de fausses alarmes.



- La table de décision suivante résume la situation :

|              | pas de feu                | feu                       |
|--------------|---------------------------|---------------------------|
| pas d'alarme | pas d'erreur              | erreur de deuxième espèce |
| alarme       | erreur de première espèce | pas d'erreur              |

- En définissant l'hypothèse nulle  $H_0$  par  $H_0$  : "pas de feu" et l'hypothèse alternative  $H_1$  par  $H_1$  : "feu", la table de décision devient

|          |                | réalité                   |                           |
|----------|----------------|---------------------------|---------------------------|
|          |                | $H_0$ vraie               | $H_1$ vraie               |
| décision | $H_0$ vraie    | décision correcte         | erreur de deuxième espèce |
|          | rejet de $H_0$ | erreur de première espèce | décision correcte         |

- Dans les tests de signification exposés dans ce chapitre, nous avons introduit l'**erreur de première espèce**, i.e la probabilité de refuser à tort l'hypothèse nulle  $H_0$  si elle est vraie. En termes probabilistes, l'erreur de première espèce s'écrit sous la forme

$$\alpha = P(\text{rejeter } H_0 \mid H_0).$$

Question : quelle est son interprétation ?

Réponse :  $1 - \alpha$  mesure la confiance qu'on peut avoir en entendant retentir la sonnerie de l'alarme que l'alarme soit fondée (vraie)

~ une confiance élevée signifie qu'on assiste rarement à de fausses alertes.

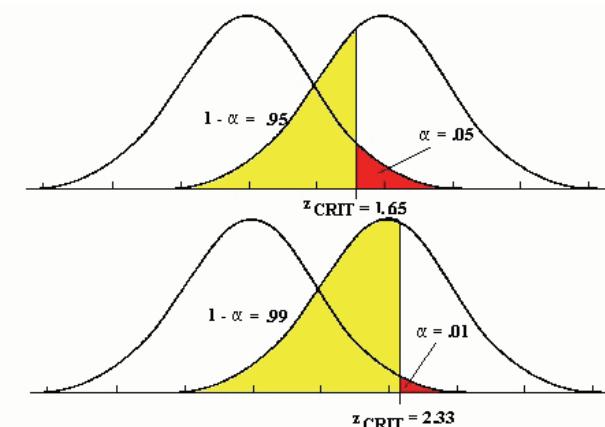
- On peut aussi raisonner avec l'**erreur de deuxième espèce**, plus précisément se demander avec quelle chance il est possible de commettre une erreur de ce type. Cette interrogation se traduit dans notre exemple par la question

"quelle est la sensibilité du détecteur d'incendie lorsque l'hypothèse alternative est vraie ?".

En probabilités, l'erreur de deuxième espèce s'écrit sous la forme

$$\beta = P(\text{ne pas rejeter } H_0 \mid H_1).$$

Elle doit être choisie avant le début de l'expérience.



**Erreur de 1ère espèce:**  
On répond «NON» alors que la vérité est «OUI»

**Erreur de 2ème espèce:**  
On répond «OUI» alors que la vérité est «NON»



Trois types d'erreur !!!

- Une caractéristique essentielle d'un test est sa **puissance** définie par

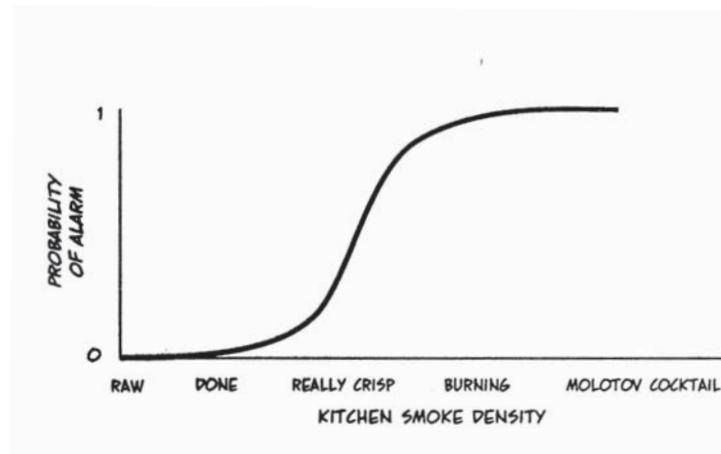
$$1 - \beta = P(\text{rejeter } H_0 \mid H_1).$$

La puissance mesure la capacité que possède le test pour détecter une hypothèse alternative

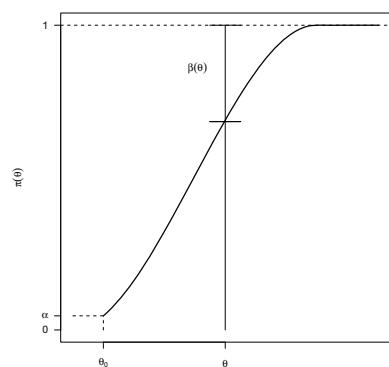
→ on souhaite que la probabilité  $1 - \beta$  grimpe le plus rapidement possible à 1.

- Par exemple, la puissance indique la chance de découvrir une différence réelle existant entre les espérances de deux populations.
- En principe, les tests ne développent qu'une faible puissance lorsque la taille de l'échantillon est petite.
- Pour visualiser la puissance d'un test, il est utile de construire un graphique. Pour l'illustrer, supposons qu'on désire tester l'hypothèse nulle  $H_0 : \theta = \theta_0$  contre l'hypothèse alternative  $H_1 : \theta > \theta_0$ . Pour chaque valeur  $\theta$  de  $H_1$ , on calcule

$$\pi(\theta) = 1 - \beta(\theta) = P(\text{rejeter } H_0 \mid \theta > \theta_0).$$

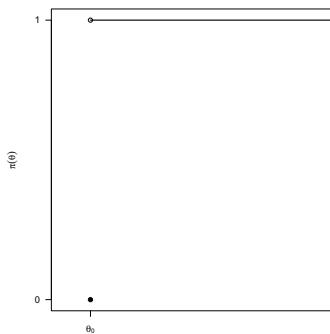


Graphique représentant la puissance du test dans le cas du détecteur d'incendies.



Graphique représentant la puissance d'un test en fonction de  $\theta$ .

~ situation idéale :



Le test peut distinguer exactement les hypothèses  $H_0$  et  $H_1$  :

- ▷  $\pi(\theta_0) = \alpha = 0$ ;
- ▷  $\pi(\theta) = 1 - \beta = 1$  si  $\theta \neq \theta_0$ .

## 10.6 Et ensuite...

### • Inférence pour des proportions

Lorsqu'on est confronté à des variables qualitatives, on s'intéresse fréquemment à la proportion de la population qui tombe dans une certaine catégorie.

Exemple : une pièce de monnaie a été lancée 4040 fois. Elle est tombée 2048 fois sur "pile". On se demande alors si la pièce est bien équilibrée. Désignons par  $p$  la probabilité que la pièce tombe sur "pile". On se propose donc de tester l'hypothèse nulle  $H_0 : p = 0.5$  contre l'hypothèse alternative  $H_1 : p \neq 0.5$ .

### • Inférence pour plusieurs espérances

~ analyse de variance (analysis of variance : ANOVA);

~ idée : comparer la variabilité interne des échantillons avec la variabilité entre les échantillons;

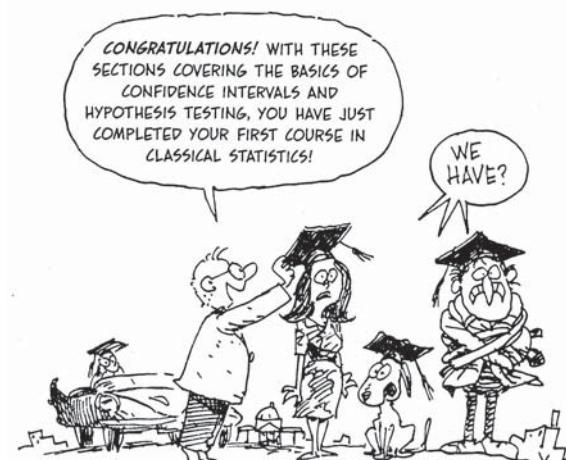
~ si la variabilité interne des échantillons est beaucoup plus petite que celle entre les échantillons, on dispose d'une certaine évidence contre l'égalité des espérances des populations.

Exemple : comparaison de plusieurs méthodes pour apprendre à lire.

- Tests non paramétriques

Les tests décrits au paragraphe 10.3.1 sont des **tests paramétriques**. Leur statistique de test dépend du paramètre que l'on souhaite tester (par exemple  $\mu$  et  $\sigma^2$ ) et du type de données (les tests supposent la normalité de la distribution des données). Un test **non paramétrique** est utile lorsque une ou plusieurs hypothèse(s) de la méthode paramétrique (comme par exemple la normalité des données) ne sont pas satisfaites. Ces tests se basent souvent sur les rangs. Plusieurs tests paramétriques ont leurs équivalents non paramétriques; citons, par exemple, le test des rangs de Mann-Whitney (alternative non paramétrique du test de Student lorsque la normalité des données n'est pas remplie), le test de Wilcoxon (alternative non paramétrique du test de Student pour des données appariées).

- ...



*"The difference between 'significant' and 'not significant' is not itself statistically significant."*

Andrew Gelman and Hal Stern

*"Understanding the models, particularly their limitations and sensitivity to assumptions, is the new task we face. Many of the banking and financial institution problems and failures of the past decade can be directly tied to model failure or overly optimistic judgements in the setting of assumptions or the parameterization of a model."*

Tad Montross, chairman and CEO of GenRe in "Model Mania"



## EXERCICES : TESTS D'HYPOTHÈSES

### Exercice 1

Un travail écrit est formé de 20 questions à 4 réponses possibles chacune. Chaque question a une seule réponse juste. Si un étudiant répond correctement à au moins 9 des 20 questions, on considère qu'il possède suffisamment de connaissances sur le sujet examiné et, par conséquent, ne fait pas que deviner la réponse.

- Écrire les hypothèses nulle et alternative auxquelles nous conduit l'énoncé.
- Proposer une statistique de test pour tester l'hypothèse nulle  $H_0$ .
- Déterminer la distribution d'échantillonnage de la statistique de test sous l'hypothèse nulle  $H_0$ .

**Solutions :** a)  $H_0 : p = p_0 = 1/4$  et  $H_1 : p > p_0 = 1/4$  b) la statistique de test pour tester  $H_0$  est naturellement  $T = \sum_{i=1}^{20} X_i$  où  $X_i$  est une variable aléatoire telle que

$$X_i = \begin{cases} 1 & \text{si l'étudiant répond correctement à la } i\text{-ème question,} \\ 0 & \text{sinon.} \end{cases}$$

c) sous l'hypothèse nulle  $H_0$ , on a  $T = \sum_{i=1}^{20} X_i \sim \mathcal{B}(20, 1/4)$ ; on suppose indépendance entre les questions posées.

### Exercice 2

Dans les années 1970, les athlètes féminines de l'Allemagne de l'Est étaient réputées pour leur forte corpulence. Le comité d'éthique olympique de l'époque, mettant en doute cette étonnante "virilité", avait fait appel aux services du Docteur Volker Fischbach. Celui-ci sélectionna 9 athlètes féminines présentant des caractéristiques morphologiques identiques puis effectua des analyses mesurant la quantité de substances hormonales virilisantes (dites androgènes) par litre de sang. Les résultats obtenus sont les suivants :

3.22 3.07 3.17 2.91 3.40 3.58 3.23 3.11 3.62

- Le Dr. Fischbach a testé l'hypothèse que les athlètes est-allemandes n'étaient pas dopées. En sachant que chez une femme la quantité moyenne d'androgène est de 3.1, écrire l'hypothèse nulle  $H_0$  en langage mathématique.
- Les athlètes est-allemandes sont soupçonnées avoir une quantité d'androgène dépassant la moyenne habituelle. Traduire cet énoncé en une hypothèse alternative  $H_1$ .
- Le Dr. Fischbach a testé l'hypothèse nulle  $H_0$  contre l'hypothèse alternative  $H_1$  au seuil de signification de 5 %. A-t-il rejeté l'hypothèse  $H_0$  à ce seuil de signification ?
- Pour effectuer le test, on a supposé que les valeurs observées sont les réalisations de variables aléatoires issues d'une distribution normale. Cette hypothèse se justifie-t-elle en se basant uniquement sur un diagramme en points ?

**Solutions :** a)  $H_0 : \mu = \mu_0 = 3.1$  où  $\mu$  représente la quantité moyenne (espérée) d'androgène par litre de sang chez une femme b)  $H_1 : \mu > \mu_0 = 3.1$  c) comme  $t_{obs} (2.00) > t_{theo} (t_{8,0.95} = 1.86)$ , le Dr. Fischbach a rejeté l'hypothèse  $H_0$  à un seuil de signification de 5 % d) oui.

### Exercice 3

Des chaussures de deux types  $A$  et  $B$  ont été attribuées aléatoirement au pied droit et au pied gauche de dix enfants. On souhaite comparer l'usure des semelles des chaussures. Les résultats obtenus figurent dans le tableau suivant :

| enfant | chaussure $A$ | chaussure $B$ |
|--------|---------------|---------------|
| 1      | 13.2 (G)      | 14.0 (D)      |
| 2      | 8.2 (G)       | 8.8 (D)       |
| 3      | 10.9 (D)      | 11.2 (G)      |
| 4      | 14.3 (G)      | 14.2 (D)      |
| 5      | 10.7 (D)      | 11.8 (G)      |
| 6      | 6.6 (G)       | 6.4 (D)       |
| 7      | 9.5 (G)       | 9.8 (D)       |
| 8      | 10.8 (G)      | 11.3 (D)      |
| 9      | 8.8 (D)       | 9.3 (G)       |
| 10     | 13.3 (G)      | 13.6 (D)      |

G : chaussure gauche

D : chaussure droite

Source : Box, G. E. P., Hunter, W. G. and Hunter J. S. (2005). *Statistics for Experimenters*. Second Edition. NY:Wiley.

Les données se trouvent dans le fichier **shoe.txt**. Enregistrez-les dans l'objet **shoe** de **R**.

Pour comparer les deux types de chaussures, on vous propose de :

- calculer les différences entre les chaussures de type  $B$  et de type  $A$  pour chaque enfant;
- construire le diagramme en points apparié en utilisant les commandes suivantes de **R** :

```
> oldpar<-par()
> attach(shoe)
> difference<-MatB-MatA
> par(mar=c(14,4.1,14,4.1))
> stripchart(data.frame(cbind(MatA,MatB)),xlim=c(6,15),group.names=c("A",
+ "B"),pch=1)
> segments(MatA,1,MatB,2)
> par(oldpar)
> detach("shoe")
```

Que peut-on en conclure en comparant les semelles des deux types de chaussures ?

- malgré la petite taille de l'échantillon, tracer la boîte à moustaches des différences. Que peut-on en conclure en se basant sur la médiane ?
- en supposant que les différences proviennent d'une distribution normale, effectuer un  $t$ -test apparié bilatéral au seuil de signification de 5 %. Quelle conclusion peut-on en tirer ?

**Solutions :** b) les valeurs de  $B$  sont en grande majorité plus grandes que les valeurs correspondantes de  $A$ . Ainsi, les semelles des chaussures de type  $B$  semblent s'user plus rapidement que celle de type  $A$  d) on teste l'hypothèse nulle  $H_0 : \mu_A = \mu_B$  contre l'hypothèse alternative  $H_1 : \mu_A \neq \mu_B$  avec données appariées. Comme la  $p$ -valeur (0.008539) est plus petite que le niveau de signification ( $\alpha = 0.05$ ), on possède suffisamment d'évidence pour rejeter l'hypothèse nulle  $H_0$  au profit de l'hypothèse alternative  $H_1$ . De plus, l'intervalle de confiance ne couvre pas 0. Ainsi, il existe une différence entre les deux types de chaussures à un niveau de signification de 5 %.

#### Exercice 4

Pour des familles formées d'au moins deux garçons, une conjecture bien étrange prétend qu'une différence existe entre les longueurs espérées des têtes des deux premiers fils à l'âge adulte.

- a) Les longueurs des têtes du premier fils et du second fils ont été mesurées dans 25 familles. Calculées à l'aide du logiciel R, les statistiques élémentaires des différences entre les longueurs des têtes sont respectivement

| Min.   | 1stQu. | Median | Mean | Sd    | 3rdQu. | Max.  |
|--------|--------|--------|------|-------|--------|-------|
| -11.00 | -4.00  | 1.00   | 1.88 | 7.535 | 8.00   | 16.00 |

En supposant que les résultats obtenus sont les réalisations de variables aléatoires issues d'une distribution normale, calculer un intervalle de confiance bilatéral à 95 % pour les différences espérées entre les longueurs des têtes des deux fils.

- b) Pour tester la conjecture, considérons l'hypothèse nulle  $H_0 : \mu_d = 0$  et l'hypothèse alternative  $H_1 : \mu_d \neq 0$  où  $\mu_d$  représente l'espérance des différences entre les longueurs des têtes du premier fils et du second fils. En utilisant les mesures prises sur les 25 familles, on a effectué un  $t$ -test apparié au niveau de signification de 5 %. Les résultats partiels sont

...

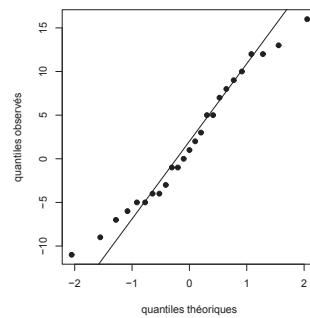
$t = 1.2475$ ,  $df = 24$ ,  $p\text{-value} = 0.2242$

alternative hypothesis: true difference in means is not equal to 0

...

Que peut-on conclure du test ? Justifiez clairement votre réponse de deux manières différentes.

- c) Dans les parties a) et b), nous avons supposé que les différences sont les réalisations de variables aléatoires issues d'une distribution normale. Cette hypothèse est-elle raisonnable en se basant sur le graphique quantiles versus quantiles se trouvant dans la figure ci-dessous ?



**Solutions :** a)  $[-1.23, 4.99]$  b) comme la  $p$ -valeur (0.2242) n'est pas inférieure au niveau de signification (0.05), on ne peut rejeter l'hypothèse  $H_0$  à ce niveau de signification-là. Comme  $t_{obs} = 1.2475$  est compris entre  $t_{24,0.025} = -t_{24,0.975} = -2.064$  et  $t_{24,0.975} = 2.064$ , on ne peut rejeter l'hypothèse  $H_0$  à un niveau de signification de 5 % c) comme les points sont en grande majorité au voisinage de la droite, l'hypothèse de normalité est raisonnable.

#### Exercice 5

Pour tester la qualité nutritive d'une nouvelle variété de maïs, 20 poussins mâles âgés d'un jour ont mangé des rations contenant le nouveau maïs. Un groupe de contrôle formé de 20 autres poussins de même âge a reçu des portions identiques sauf qu'elles contenaient du maïs normal. Les augmentations en grammes à l'issu d'une période de 21 jours figurent dans le tableau ci-dessous.

| sans nouvelle variété |     |     |     | avec nouvelle variété |     |     |     |
|-----------------------|-----|-----|-----|-----------------------|-----|-----|-----|
| 380                   | 321 | 366 | 356 | 361                   | 447 | 401 | 375 |
| 283                   | 349 | 402 | 462 | 434                   | 403 | 393 | 426 |
| 356                   | 410 | 329 | 399 | 406                   | 318 | 467 | 407 |
| 350                   | 384 | 316 | 272 | 427                   | 420 | 477 | 392 |
| 345                   | 455 | 360 | 431 | 430                   | 339 | 410 | 326 |

- a) On se demande si le poids des poussins augmente en moyenne plus rapidement en utilisant la nouvelle variété de maïs au lieu de l'ancienne. Traduire cet énoncé par une hypothèse nulle  $H_0$  et une hypothèse alternative  $H_1$ .
- b) Les statistiques élémentaires selon l'utilisation ou non de la nouvelle variété de maïs sont respectivement

| sans             | avec             |
|------------------|------------------|
| Min. : 272.00    | Min. : 318.00    |
| 1st Qu. : 341.00 | 1st Qu. : 387.75 |
| Median : 358.00  | Median : 406.50  |
| Mean : 366.30    | Mean : 402.95    |
| Sd : 50.81       | Sd : 42.73       |
| 3rd Qu. : 399.75 | 3rd Qu. : 427.75 |
| Max. : 462.00    | Max. : 477.00    |

À la suite d'un test  $F$ , l'égalité des variances des deux populations est plausible. En supposant l'égalité des variances, calculer la valeur  $t_{obs}$  de la statistique  $T$  du  $t$ -test utilisé pour comparer les deux variétés de maïs.

- c) Déterminer la valeur théorique  $t_{theo}$  pour un niveau de signification de 5 %. Peut-on rejeter l'hypothèse  $H_0$  à ce niveau-là ?
- d) Quelle hypothèse sur les observations avons-nous faite pour appliquer le  $t$ -test ?

**Solutions :** a) on teste l'hypothèse nulle  $H_0 : \mu_{sans} = \mu_{avec}$  contre l'hypothèse alternative  $H_1 : \mu_{sans} < \mu_{avec}$  où  $\mu_{sans}$  représente l'augmentation espérée du poids des poussins sans utilisation de la nouvelle variété de maïs et  $\mu_{avec}$  l'augmentation espérée du poids des poussins avec utilisation de la nouvelle variété b)  $t_{obs} = -2.469$  c) comme  $t_{obs} (-2.469) < t_{theo} (t_{38,0.05} = -1.686)$ , on rejette l'hypothèse  $H_0$  à un seuil de signification de 5 % d) on a supposé que les observations provenaient d'une distribution normale et l'indépendance entre les groupes de poussins.

### Exercice 6

Avant d'entreprendre l'écriture d'un programme, un ingénieur se propose de tester deux langages de programmation très différents. Il demande à douze programmeurs expérimentés et maîtrisant parfaitement les deux langages de programmation d'écrire une fonction standard dans les deux langages. Les temps nécessaires en minutes pour coder la fonction figurent dans le tableau ci-dessous.

| programmeur | langage 1 | langage 2 |
|-------------|-----------|-----------|
| 1           | 17        | 18        |
| 2           | 16        | 14        |
| 3           | 21        | 19        |
| 4           | 14        | 11        |
| 5           | 18        | 23        |
| 6           | 24        | 21        |
| 7           | 16        | 10        |
| 8           | 14        | 13        |
| 9           | 21        | 19        |
| 10          | 23        | 24        |
| 11          | 13        | 15        |
| 12          | 18        | 20        |

- a) Pour étudier le temps de codage des deux langages, on a testé

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2,$$

où  $\mu_1$  et  $\mu_2$  représentent respectivement les espérances des temps de programmation dans les deux langages.

À l'aide du diagramme en points appariés construit dans la Figure 1, peut-on observer une différence entre les temps de codage dans les deux langages ?

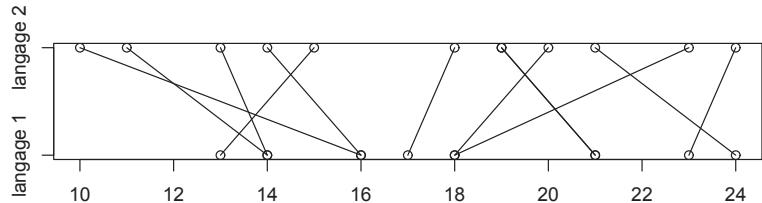


Figure 1: Temps de codage de la fonction dans les langages 1 et 2.

- b) En supposant que les données proviennent d'une distribution normale, déterminer la région de rejet de l'hypothèse  $H_0$  à un niveau de signification de 5%.
- c) Les résultats partiels du  $t$ -test apparié obtenus à l'aide du logiciel de statistique R figurent à la page suivante.

```
...
t = 0.779, df = 11
alternative hypothesis: true difference in means is not equal to 0
...
```

Peut-on rejeter l'hypothèse  $H_0$  à un niveau de signification de 5% ?

- d) Parmi les probabilités 0.45, 0.03 et 0.0045 laquelle est la  $p$ -valeur du test ?

**Solutions :** a) comme les segments ne sont en majorité pas verticaux (égalité des temps de programmation), le graphique laisse apparaître une différence entre les temps de codage b) comme  $\alpha = 0.05$ , les valeurs théoriques  $t_{11,0.975}$  et  $t_{11,0.025} = -t_{11,0.975}$  valent respectivement +2.201 et -2.201. Ainsi, la région de rejet de l'hypothèse nulle  $H_0$  est  $(-\infty, -2.201) \cup (+2.201, +\infty)$  c) comme la région de rejet ne couvre pas la valeur  $t = t_{obs} = 0.779$ , on ne rejette pas l'hypothèse  $H_0$  à un niveau de signification de 5%. On remarque que le graphique tracé en a) laisse apparaître une différence entre les temps de codage. Par le test d'hypothèses effectué en c), on conclut que cette différence n'est pas significative à un niveau de signification de 5% d) comme l'hypothèse  $H_0$  n'est pas rejetée à un niveau de signification de 5%, la  $p$ -valeur du test vaut 0.45.

### Exercice 7

La sortie partielle d'un test d'hypothèses obtenue à l'aide du logiciel de statistique R est

```
Paired t-test ...
t = 3.6927, df = 13, p-value = 0.001354
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
1.8215 Inf
...
```

- a) Reconstituer l'hypothèse nulle et l'hypothèse alternative.
- b) Interpréter les résultats de trois manières différentes.
- c) Quel est le niveau de signification du test ?
- d) Quelle hypothèse faites-vous sur la distribution des valeurs observées ?
- e) Si l'hypothèse nulle est rejetée à un niveau de signification de 1%, la sera-t-elle aussi à un niveau de 5% ?

**Solutions :** a) on a testé l'hypothèse  $H_0 : \mu_1 - \mu_2 = 0$  où  $\mu_1$  et  $\mu_2$  sont les espérances inconnues des deux populations contre l'hypothèse alternative  $H_1 : \mu_1 - \mu_2 > 0$  à un niveau de signification de 5%. Les données sont appariées b) comme la  $p$ -valeur (0.001354) est inférieure au niveau de signification  $\alpha = 0.05$ , on rejette l'hypothèse  $H_0$  à un niveau de signification de 5%. Comme  $\alpha = 0.05$ , la valeur théorique  $t_{13,0.95}$  vaut 1.771. Ainsi, la région de rejet de l'hypothèse  $H_0$  est  $(1.771, +\infty)$ . Comme la région de rejet couvre la valeur  $t = t_{obs} = 3.6927$ , on rejette l'hypothèse  $H_0$  à un niveau de signification de 5%. Comme l'intervalle de confiance pour la différence  $\mu_1 - \mu_2$  (1.8215,  $+\infty$ ) ne couvre pas la valeur hypothétique  $d_0 = 0$ , l'hypothèse  $H_0$  est réfutée au niveau de signification de 5% c) le niveau de signification du test est 5% d) on suppose la normalité des observations e) oui, si la  $p$ -valeur du test est inférieure à 0.01, elle sera aussi de 0.05.

### Exercice 8

Quatre dés sont jetés simultanément. Désignons par  $X$  la variable aléatoire qui compte le nombre de dés indiquant un six.

- On suppose que les dés sont équilibrés. Sous cette hypothèse, déterminer les réalisations  $x_j$  et la distribution théorique de  $X$ .
- Les dés sont lancés 10'000 fois de suite. Les fréquences observées  $o_j$  figurent dans le tableau ci-dessous.

| $x_j$ | 0    | 1    | 2    | 3   | 4  |
|-------|------|------|------|-----|----|
| $o_j$ | 5023 | 3687 | 1132 | 148 | 10 |

Effectuer un test d'équation du modèle théorique défini en a) aux valeurs observées à un niveau de signification de 1%. Que peut-on conclure ?

**Solutions :** a)  $\mathcal{H} = \{0, 1, 2, 3, 4\}$  et  $X \sim \mathcal{B}(4, 1/6)$  b) comme la  $p$ -valeur (0.002) est inférieure au niveau de signification de 1% (0.01), on rejette l'hypothèse  $H_0$  : la variable aléatoire  $X$  est issue d'une distribution  $\mathcal{B}(4, 1/6)$ .

### Exercice 9

Une firme pharmaceutique a développé un certain vaccin contre une maladie. 200 personnes ont participé à un test dont l'objectif consiste à savoir si le vaccin protège ou non contre la maladie. Les résultats obtenus figurent dans le tableau ci-dessous.

| A            | B        |              |
|--------------|----------|--------------|
|              | atteints | non atteints |
| vaccinés     | 10       | 30           |
| non vaccinés | 65       | 95           |

Effectuer un test d'indépendance à un niveau de signification de 5% en utilisant les commandes suivantes de R :

```
> A<-rep(c("vaccinés","non vaccinés"),c(40,160))
> B<-rep(c("atteints","non atteints","atteints","non atteints"),c(10,30,65,95))
> maladie<-table(A,B)
> summary(maladie)
```

**Solution :** comme la  $p$ -valeur (0.068) est supérieure au niveau de signification de 5% (0.05), on ne peut réfuter l'hypothèse d'indépendance à un niveau de signification de 5%.

### Exercice 10

Il est bien connu que la consommation d'alcool et de nicotine pendant la période de grossesse peut nuire gravement à la santé de l'enfant. Dans une étude menée sur 452 futures mères, on souhaite comprendre la nature de la relation existante entre alcool et nicotine afin d'évaluer les effets sur l'enfant. Les participantes à l'étude ont été classées en trois catégories pour la consommation en nicotine et en quatre catégories pour l'alcool. Les résultats obtenus figurent dans le tableau de la page 8.

| alcool (ounces/jour) | nicotine (milligrammes/jour) |      |            |
|----------------------|------------------------------|------|------------|
|                      | 0                            | 1-15 | 16 ou plus |
| 0                    | 105                          | 7    | 11         |
| 0.01-0.10            | 58                           | 5    | 13         |
| 0.11-0.99            | 84                           | 37   | 42         |
| 1.00 ou plus         | 57                           | 16   | 17         |

Source : Ann P. Streissguth et al., "Intrauterine alcohol and nicotine exposure : attention and reaction time in 4-year-old children", *Developmental Psychology*, 20 (1984).

- Les résultats partiels d'un test d'indépendance entre l'alcool et la nicotine au niveau de signification de 1% obtenus à l'aide du logiciel de statistique R sont

```
...
Number of factors: 2
Test for independence of all factors:
Chisq = 42, df = 6, p-value = 1.6e-07
...
```

Interpréter les résultats fournis par R. Quelles conclusions peut-on tirer du test ?

- Si l'hypothèse nulle  $H_0$  était refusée, serait-on en mesure de préciser quelle relation existe entre l'alcool et la nicotine ?
- Vrai ou Faux. L'erreur de deuxième espèce  $\beta$  du test vaut  $1 - \alpha$  où  $\alpha$  est l'erreur de première espèce du test.

**Solutions :** a) comme la  $p$ -valeur est plus petite que le niveau de signification fixé à 1%, on rejette l'indépendance des variables à ce niveau-là. Autrement dit, une relation existe entre l'alcool et la nicotine à un risque d'erreur de 1% b) si l'hypothèse  $H_0$  est refusée, on ne peut préciser la forme de la relation entre les deux variables. On sait juste qu'elle existe c) faux ! Il s'agit de deux probabilités conditionnelles mais l'événement conditionnant est différent, une fois  $H_0$ , une fois  $H_1$ .

### Exercice 11

Il est bien connu que les garçons semblent s'adonner davantage aux jeux vidéo que les filles. Pour remettre en question ce cliché, 89 étudiant(e)s en statistique d'une école américaine ont été choisis au hasard en 1994 et questionnés sur leur comportement face aux jeux vidéo. Les résultats obtenus figurent dans le tableau ci-dessous.

| Aime jouer | Sexe     |         |
|------------|----------|---------|
|            | mASCULIN | fÉMININ |
| Oui        | 43       | 26      |
| Non        | 8        | 12      |

Source : Nolan, D. & Speed T. (2000). *Stat Labs: Mathematical Statistics Through Applications*. NY:Springer.

- Les résultats partiels d'un test d'indépendance entre le sexe et le comportement face aux jeux vidéo obtenus à l'aide du logiciel de statistique R figurent à la page suivante.

```

...
Number of factors: 2
Test for independence of all factors:
Chisq = 3.16, df = 1
...

```

Esquisser le graphe de la fonction de densité de la statistique de test sous l'hypothèse d'indépendance.

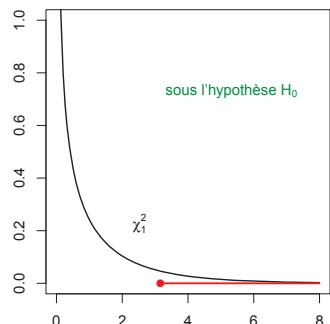
- b) Déterminer la valeur théorique  $t_{theo}$  pour un niveau de signification de 5% puis en déduire la région de rejet de l'hypothèse d'indépendance à ce niveau-là.
- c) Ajouter la région de rejet au graphique construit en a).
- d) Peut-on rejeter l'hypothèse d'indépendance entre le sexe et le comportement face aux jeux vidéo à un niveau de signification de 5% ?
- e) Parmi les probabilités 0.28, 0.076 et 0.0076 laquelle est la  $p$ -valeur du test ?

**Solutions :**

a) sous l'hypothèse  $H_0$ , la statistique de test  $T$  pour tester l'indépendance entre le sexe et le comportement face aux jeux vidéo est une variable aléatoire issue d'une distribution  $\chi^2_1$ . Le graphe de la fonction de densité de  $T$  sous l'hypothèse  $H_0$  se trouve dans la figure de droite

b) comme le niveau de signification  $\alpha$  est fixé à 0.05, la valeur théorique  $t_{theo} = \chi^2_{1, 0.95}$  est 3.84. La région de rejet de l'hypothèse d'indépendance est donc  $[3.84, +\infty)$

c) la superposition de la région de rejet au graphique construit en a) se trouve en rouge dans la figure de droite.



d) comme la statistique observée  $t = t_{obs} = 3.16$  n'est pas couverte par la région de rejet, nous ne disposons pas de suffisamment d'évidence pour réfuter l'indépendance entre le sexe et le comportement face aux jeux vidéo. Ainsi, en se basant sur ces données, les garçons et les filles ont un comportement similaire face aux jeux vidéo

e) comme nous ne pouvons pas rejeter l'hypothèse d'indépendance à un niveau de signification de 5%, les  $p$ -valeurs possibles sont 0.28 et 0.076. Cependant, les valeurs  $t_{obs}$  et  $t_{theo}$  étant proches l'une de l'autre, il est probable que la  $p$ -valeur est 0.076.



## Chapitre 11

### Introduction au data mining orienté vers le business

#### 11.1 Introduction et motivation

La quantité d'information stockée dans les systèmes informatiques des entreprises, sociétés et organisations croît de manière extrêmement rapide :

- le volume des données exploitables croît exponentiellement chaque année;
- l'unité de stockage des grandes bases de données est le tera-octet (1024 Go), soit environ la taille d'une bibliothèque de plus de 2 millions de livres;
- WallMart, la grande chaîne de distribution américaine, enregistre chaque jour plus de 20 millions de transactions à partir de ses points de vente.

## Contenu

- 11.1 Introduction et motivation
- 11.2 Qu'est-ce que le data mining ?
- 11.3 Nature des données
- 11.4 Le processus d'extraction des connaissances
- 11.5 Quelques techniques de data mining
- 11.6 Conclusion
- 11.7 Références et ressources

Plusieurs compagnies collectent des centaines de gigabytes sur leurs clients sans les analyser. Une étude du Gartner Group montrait que moins de 15 % des données stockées et moins de 5 % des données manipulées étaient effectivement analysées.

Or, les besoins en analyse et en étude de données (analyse marketing, analyse du risque, détection de fraudes, ...) croissent de 15 % à 45 % chaque année et sont liés à des problèmes vitaux pour le positionnement concurrentiel (différenciation par rapport à la concurrence, avantage concurrentiel);

→ besoin croissant d'outils d'aide à la décision permettant par exemple d'extrapoler à partir des données existantes (prévision).



Le développement des moyens informatiques de stockage (bases de données) et de calcul permet maintenant le traitement et l'analyse d'ensemble de données très volumineux. Le perfectionnement des interfaces et la popularisation de nouvelles méthodes algorithmiques et d'outils graphiques conduisent à l'apparition et à la commercialisation de logiciels intégrant un sous-ensemble de méthodes statistiques et algorithmiques sous la terminologie de **data mining** en anglais, **fouille de données** ou **prospection de données** en français.

En d'autres termes, on ressent un besoin d'outils de data mining dont l'objectif consiste à extraire des informations, des connaissances pertinentes cachées dans de grandes bases de données souvent complexes. Les techniques du data mining ne constituent qu'une composante du processus plus général d'extraction des connaissances (*KDD : Knowledge Discovery in Databases*).



*En route pour la mine à connaissances !*



## Applications du data mining

### 1. Grande distribution et vente par correspondance :

- analyser le comportement du consommateur;
- déterminer les différents profils du consommateur (fidélité);
- identifier les exigences des consommateurs;
- détecter des associations pertinentes entre les produits et les clients;
- mettre en œuvre une stratégie de marketing ciblé;
- prévision des ventes;
- localisation de nouvelles succursales;
- promotion et développement de nouveaux articles.



De: "Amazon.com" <store-news@amazon.com>

A: jzuber@netplus.ch

Date: Ven, 9 Mars 2007, 4:21

Sujet: Amazon.com recommends Probability and Statistics and more

Jacques Zuber, Amazon.com has new recommendations for you based on items you purchased or told us you own.

In this message:

- \* Probability and Statistics
- \* Concrete
- \* How To Lie With Charts: Second Edition
- \* Seeing Through Statistics (with CD-ROM and InfoTrac )
- \* Ordinal Data Modeling (Statistics for Social Science and Behavioral Sciences)
- \* Models for Discrete Longitudinal Data (Springer Series in Statistics)
- \* The Statistical Evaluation of Medical Tests for Classification and Prediction (Oxford Statistical Science Series)
- \* Matched Sampling for Causal Effects

...

We recommend: Probability and Statistics

by 3rd Edition DeGroot and Schervish  
[http://www.amazon.com/dp/1428813802/ref=pe\\_ar\\_x1](http://www.amazon.com/dp/1428813802/ref=pe_ar_x1)

Price: \$9.95

Recommended because you purchased or rated:  
 \* Probability and Statistics (3rd Edition)

...

### 2. Banques :

- analyse de risque (prêt bancaire : planification des ressources, prévision, détermination des tendances du marché, suivi des fichiers clients pour attribuer les prêts);
- détection de fraudes (cartes de crédit : identification de transactions suspectes ou incohérentes).

### 3. Assurances :

- analyse des sinistres (recherche de critères explicatifs du risque);
- détection de fraudes (détection d'individus mettant en scène des accidents factices);
- déterminer les différents profils du fraudeur.

#### 4. Opérateurs des télécommunications :

- modélisation prédictive des clients partants;
- identification des facteurs poussant un utilisateur à changer de fournisseur;
- planification des ressources du réseau;
- détection de fraudes (téléphones portables : recherche de schémas s'écartant d'une norme standard).

#### 5. Astronomie :

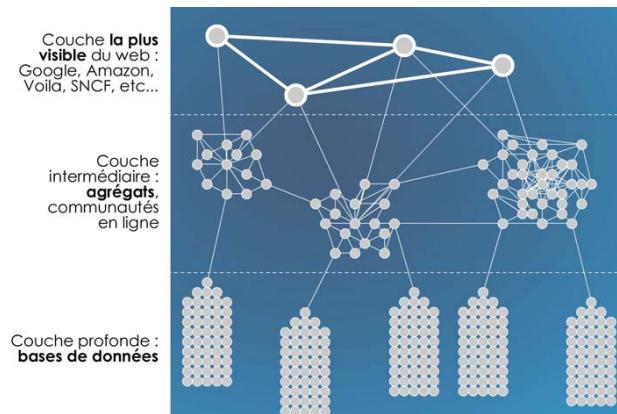
- classification et découverte de nouveaux corps.

#### 6. Internet :

- analyse du comportement d'utilisateurs d'internet;
- analyse de l'efficacité d'une stratégie de marketing par internet.

#### 7. Industrie :

- contrôle de la qualité et anticipation des défauts;
- gestion des stocks, des ventes d'un groupe afin de prévoir et anticiper au mieux les tendances du marché.



*Idée de la construction d'une base de données liée au webmining.*

L'exemple fictif classique...

#### 1. Problématique :

tous les achats des clients d'une chaîne de supermarchés sont enregistrés électroniquement et forme ainsi une base de données de grande envergure.

À l'aide des données, on souhaite répondre aux questions suivantes :

- former des groupes de clients achetant les mêmes produits;
- découvrir des relations entre les produits vendus.

## 2. Objectifs de l'analyse :

- disposer la marchandise de manière optimale dans les magasins;
- faciliter la recherche des produits dans les supermarchés.

## 3. Résultats de l'analyse à l'aide du data mining :

- une forte relation existe entre la vente des langes pour bébés et la vente de bière !
- l'achat commun de bière et de langes se réalise surtout par des hommes;
- les achats de bière et de langes s'effectuent en grande partie le soir et pendant le week-end.

## 11.2 Qu'est-ce que le data mining ?

- Le processus d'extraction des connaissances (*KDD : Knowledge Discovery in Databases*) consiste à extraire des connaissances pertinentes dans de grandes bases de données complexes en vue d'analyses puis de prises de décision.
- Connaissances  $\neq$  Données

▷ **données** : ce qui peut être saisi et stocké;

**exemples** : données sur les clients d'une société pratiquant la vente par correspondance, données démographiques, données géographiques, données concernant le contenu d'un entrepôt de marchandises, ...

## 4. Interprétation des résultats :

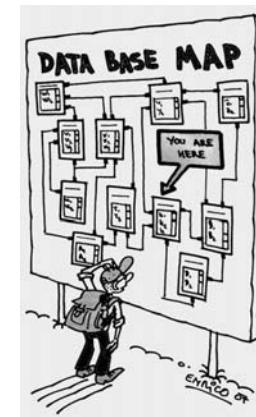
les hommes ont vraisemblablement été priés par leur épouse d'acheter des langes sur le chemin de retour après leur journée de travail ou pendant le week-end. Par la même occasion, la plupart d'entre eux n'ont pu succomber à la tentation d'acheter des bières.

## 5. Et alors...

on décide de placer sur les étalages des supermarchés les amuse-gueule (chips et autres) et les bières à côté des langes.

## 6. Conséquence :

rapidement après la réorganisation des magasins, la vente des amuse-gueule a augmenté de 27 %.



- Connaissances  $\neq$  Données (*suite*)

- ▷ les données se transforment en **informations** lorsqu'elles deviennent pertinentes dans le processus d'aide à la décision;

exemples :  $X$  vit dans la ville  $Z$ ,  $S$  est âgé de  $Y$  ans,  $X$  et  $S$  ont déménagé,  $W$  possède un compte dans la banque  $B$ , ...

- ▷ les informations deviennent des **connaissances** quand elles contribuent au processus d'aide à la décision;

exemples : une quantité  $Q$  du produit  $A$  est utilisée dans la ville  $Z$ , les clients appartenant à la classe  $L$  emploient  $N\%$  de  $C$  durant la période  $T$ , ...

*"We are drowning in information but starved for knowledge."*

John Naisbitt

- Des données aux connaissances

Données  $\rightarrow$  Informations  $\rightarrow$  Connaissances  $\rightarrow$  Décisions  $\rightarrow$  Actions.

Exemples de décisions : vente promotionnelle du produit  $A$  dans la ville  $Z$ , mise en œuvre d'une campagne publicitaire destinée aux clients d'un certain profil, ...

- L'utilité des connaissances extraites se traduit par leur intégrabilité dans un processus d'aide à la décision (d'élaboration de stratégies de vente ou de marketing) dont le **data mining** n'en est qu'une composante (analyse des données).
- Le data mining est un mélange de plusieurs disciplines, en particulier la statistique, l'informatique, l'algorithmique, l'intelligence artificielle, l'apprentissage automatique et la visualisation de données.

- En raison de la combinaison de plusieurs disciplines, le data mining peut être défini de différentes manières. En voici quelques définitions...

◊ “Processus de sélection, d'exploration, de modification et de modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues”.

SAS Institute Inc.

◊ “Le data mining est le processus d'exploration et d'analyse de grands volumes de données à l'aide de techniques automatiques ou semi-automatiques afin de découvrir des schémas et des règles pertinents”.

Michael J. A. Berry et Gordon S. Linoff

- Quelques définitions du data mining (*suite*)

◊ “À l'aide d'une variété de techniques, le data mining identifie des pépites d'informations ou connaissances dans des masses de données. Le data mining extrait des informations en vue de les utiliser comme base à la décision, à la prévision et à l'estimation. Les données sont volumineuses, sans grande valeur et peu utiles dans leur état brut. En revanche, les informations cachées dans les données sont précieuses”.

Clementine User's Guide, un logiciel pour le data mining

◊ “Le data mining consiste à trouver des structures (schémas, modèles statistiques, relations) dans des bases de données”.

Usama Fayyad, Surajit Chaudhuri et Paul Bradley

- Quelques définitions du data mining (*suite*)

◊ “Le data mining est un processus qui utilise une variété de méthodes d'analyse de données afin de découvrir des schémas et des relations dans les données, utiles pour effectuer des prévisions fiables”.

Herbert Edelstein, Président de Two Crows Corporation

◊ “Comment trouver un diamant dans un tas de charbon sans se salir les mains”.

Accroche publicitaire pour le data mining

- Le data mining n'est ni un produit à acheter ni une boîte noire. Il s'agit plutôt d'une discipline qui doit être exercée.

- Du point de vue statistique, le data mining est principalement formé de méthodes de l'analyse de données multivariées (multidimensionnelles).

- Data mining ↔ statistique classique

a) **statistique classique** : les données sont collectées avec en arrière plan une problématique soigneusement formulée;

*exemples* : augmenter la taille d'un échantillon lors d'un sondage ou d'une étude d'opinion, planifier des expériences afin d'optimiser des procédés;

b) **data mining** : les données ne sont pas collectées pour répondre à une problématique. On souhaite découvrir les structures et relations dans des bases de données;

*exemples* : banque de données d'une société pratiquant de la vente par correspondance, trafic des transactions bancaires.

*“What distinguishes data mining from statistical analysis is not so much the amount of data we analyse or the methods we use but that we integrate what we know about the database, the means of analysis and the business knowledge.”*

Paolo Giudici

- Le data mining nous permet d'apprendre une méthodologie très générale :
  - ▷ choisir et préparer les données;
  - ▷ utiliser de manière adéquate les techniques du data mining;
  - ▷ interpréter correctement les résultats avec aisance et assurance.

*"The purpose of collecting data is to provide a basis for action."*

W. Edwards Deming

### 11.3 Nature des données

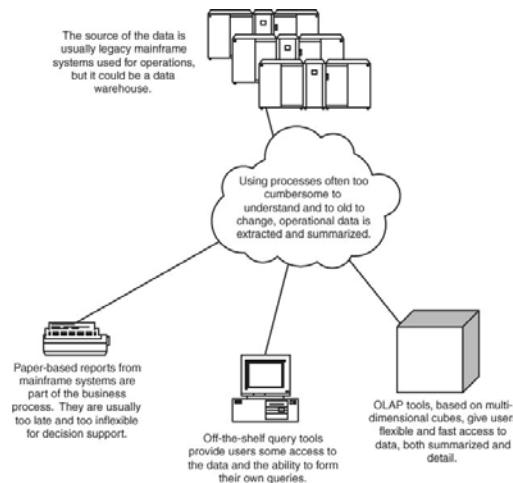
- Les données produites dans les entreprises et organisations sont principalement regroupées dans des bases de données, ensemble de données souvent très volumineux.
- Bases de données ↗ warehouse

le contexte informationnel du data mining est celui des **data warehouse**, entrepôt de données en français. Il s'agit d'un ensemble de bases de données relatives à une problématique :

- ▷ gestion des stocks, des ventes d'un groupe afin de prévoir et anticiper au mieux les tendances du marché;

- Bases de données ↗ warehouse (*suite*)

- ▷ suivi des fichiers clients d'une banque associés à des données socio-économiques, à l'annuaire, en vue de constituer une segmentation (typologie) pour optimiser des opérations de marketing ou les attributions de crédits;
- ▷ suivi des paramètres de production en contrôle de qualité pour détecter au plus vite l'origine d'une défaillance;
- ▷ ...
- L'organisation des données dans une warehouse autorise un accès rapide. Le traitement analytique en ligne OLAP (on-line analytical processing) permet à l'utilisateur d'obtenir des résumés succincts, principalement des statistiques élémentaires comme par exemple le nombre de transactions par produit ou par succursale. OLAP forme une partie du data mining.



La fouille de données peut s'effectuer à partir de différentes sources.

- Pourquoi le “boom” du data mining ?

- ▷ des volumes considérables de données souvent issues de saisies automatisées (on-line transaction processing, OLTP) sont produits dans des entreprises et organisations (codes sur les marchandises, distributeurs automatiques d’argent, cartes de crédit, achats par internet). Apparemment toute sorte de données sont collectées comme par exemple lors d’une commande par téléphone ou par internet;
- ▷ les ordinateurs, les réseaux informatiques sont de plus en plus performants et fiables;
- ▷ des algorithmes et des logiciels très efficaces ont été développés;
- ▷ une concurrence acharnée sévit sur le marché :

“Si vous ne faites pas mieux, votre concurrent le fera !”.

Anonyme

- Des bases de données organisées en warehouse. Très bien, mais...
  - ▷ environnement informatique hétérogène faisant intervenir des sites distants à travers le réseau de l’entreprise (intranet) ou même des accès extérieurs (internet);
  - ▷ taille des bases de données gigantesque
    - ~ méthodes d’analyse particulières;
  - ▷ bases de données pas toujours statiques, plutôt dynamiques;
  - ▷ bases de données sous différents formats
    - ~ création d’une warehouse avec format commun et définitions cohérentes;
  - ▷ données contaminées ~ qualité des données ? ~ nettoyage
    - ~ nécessité de préparer les données (une étape du KDD) !
  - ▷ pas de stationnarité (“Population Drift”);
  - ▷ échantillon représentatif ? (“Selection Bias”);



## 11.4 Le processus d'extraction des connaissances

Les techniques du data mining ne constituent qu'une composante du processus plus général d'extraction des connaissances (KDD) dont la décomposition selon le CRISP-DM (CRoss Industry Standard Process) est la suivante :

1. poser le problème;
2. rechercher les données et sélectionner les données pertinentes;
3. préparer les données;
4. modéliser;
5. évaluer;
6. diffuser des connaissances acquises pour la prise de décision.

*"No study is better than the quality of its data."*

Anonymous

### 1. Poser le problème :

formulation et structuration du problème, intégration de connaissances préalables pertinentes, spécification des objectifs à atteindre, compréhension du projet.

### 2. Rechercher les données et sélectionner les données pertinentes :

identification des informations exploitable, collecte et sélection de données, manipulation des données pour vérifier leur qualité et détecter certains problèmes, création d'un échantillon représentatif, définition d'hypothèses concernant les informations cachées.

### 3. Préparer les données (nettoyage, insertion, réduction, sélection des variables, transformation) :

les données proviennent fréquemment de sources externes. Elles peuvent donc se présenter sous différents formats, être incomplètes et souvent "sales". L'objectif de cette phase consiste à préparer les données pour la phase suivante. Les tâches à accomplir dans cette étape sont :

- ▷ la détection de valeurs atypiques (valeur atypique : valeur suspecte due à un appareil défectueux, à un malentendu, à une défaillance lors du transfert des données, aux limites de la technologie, à une incohérence, à une erreur humaine, ...);



| cid | Date      | Amount |
|-----|-----------|--------|
| 1   | 3/1/2000  | \$150  |
| 1   | 3/5/2000  | \$50   |
| 1   | 3/29/2000 | \$200  |
| 1   | 4/2/2000  | \$300  |
| 1   | 4/6/2000  | \$200  |
| 2   | 3/2/2000  | \$100  |
| 2   | 3/5/2000  | \$250  |
| 2   | 3/10/2000 | \$100  |
| 2   | 3/11/2000 | \$50   |
| 2   | 4/7/2000  | \$200  |
| 3   | 4/2/2000  | \$300  |
| 4   | 3/17/2000 | \$250  |
| 4   | 4/25/2000 | \$100  |



| cid | Month | Year | Amount | Visits |
|-----|-------|------|--------|--------|
| 1   | 3     | 2000 | \$400  | 3      |
| 1   | 4     | 2000 | \$500  | 2      |
| 2   | 3     | 2000 | \$500  | 4      |
| 2   | 4     | 2000 | \$200  | 1      |
| 3   | 4     | 2000 | \$300  | 1      |
| 4   | 3     | 2000 | \$250  | 1      |
| 4   | 4     | 2000 | \$100  | 1      |

Exemple de préparation de données.

### 3. Préparer les données (suite) :

- ▷ le traitement des données manquantes (valeur manquante due à une valeur d'une variable non disponible, à une incohérence, à une défaillance d'un appareil, à une erreur humaine, à un oubli d'une modification, à la définition d'une nouvelle variable, ...);
- ▷ l'intégration des données (nom de variables, unité, incohérence due à des redondances);
- ▷ la réduction, compression (data warehouse trop grande, trop complexe);
- ▷ la normalisation;
- ▷ l'agrégation et discrétilisation;
- ▷ la transformation (souvent beaucoup de transformations !);
- ▷ ...

~ "Quality decisions come from quality data !"

### 3. Préparer les données (suite) :

les algorithmes du data mining nécessitent une disposition des données dans un tableau : les lignes représentent les individus (clients) et les colonnes indiquent les variables (attributs, caractères).

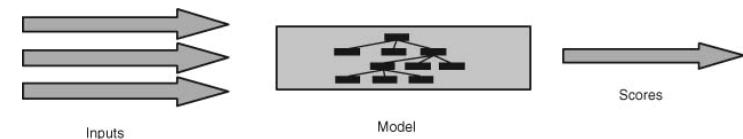
| NCarte      | Montant | Periode | Nachats |
|-------------|---------|---------|---------|
| C0106437190 | 129     | 2       | 1       |
| C0105854709 | 179     | 2       | 2       |
| C0106265943 | 49      | 2       | 1       |
| C0102754347 | 8.99    | 2       | 1       |
| C0101781222 | 49      | 2       | 1       |
| C0103227828 | 74      | 2       | 1       |
| C0100847164 | 59.99   | 2       | 1       |
| C0104537625 | 94      | 2       | 1       |
| C0103507264 | 329     | 2       | 1       |

#### 4. Modéliser :

extraction des connaissances au sens strict, détermination de l'objectif et de la stratégie d'analyse : exploration, description ou prédition puis choix de la méthode, des algorithmes, mise en œuvre des outils informatiques appropriés pour aboutir à une modélisation.

#### 5. Évaluer :

évaluer minutieusement le modèle et vérifier sur la base de critères à préciser (qualité d'ajustement, de prévision, simplicité, représentations graphiques) qu'il réponde favorablement aux objectifs fixés.



*La modélisation constitue une étape de l'extraction de connaissances.*

#### 6. Diffuser des connaissances acquises pour la prise de décision :

diffusion de l'information vers les utilisateurs, présentation des connaissances obtenues, intégration directe de l'information dans un processus de décision.

Souvent la décision finale est prise par le responsable fonctionnel, non pas par l'analyste des données.

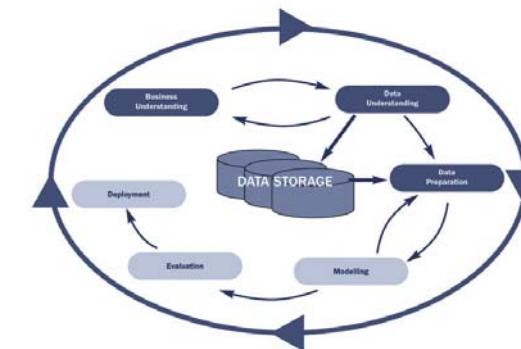
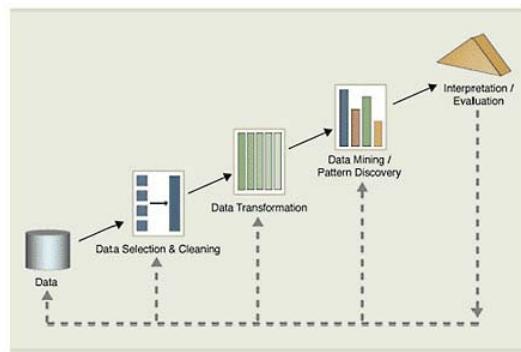
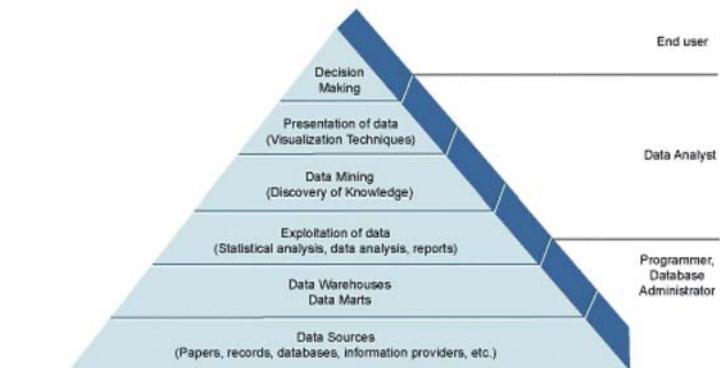


Figure 5.2 A generic data mining methodology

*Résumé du processus d'extraction de connaissances.*



Résumé du processus d'extraction de connaissances avec data mining.



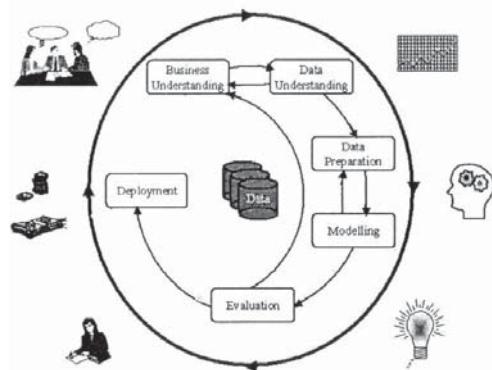
Séparation des tâches dans la recherche de connaissances.

### Remarques :

- le processus d'extraction des connaissances n'est pas rigide. Des allers-retours entre les phases sont requises. Il va sans dire que le processus s'arrête uniquement lorsqu'une solution adéquate est trouvée !
- les trois premières étapes de l'extraction des connaissances sont importantes !!! Elles représentent le 60% de l'effort dans la recherche d'informations;
- la règle du "Garbage in, garbage out !"

### Remarques (suite) :

- formuler correctement le problème, s'entourer de personnes compétentes, entretenir une excellente communication entre les partenaires : trois principes indispensables pour assurer la réussite du projet;
- dans la collecte des données, des problèmes peuvent surgir : où peut-on obtenir les données ? Disposons-nous de toutes les données ? Est-ce que certaines d'entre elles se situent encore dans les ordinateurs personnels des partenaires au projet ?
- une expertise humaine de la part d'un informaticien, d'un statisticien (méconnaissance des limites et pièges des méthodes statistiques) et d'un utilisateur confirmé est indispensable pour mener à bien le projet.



*Phases de la recherche de connaissances selon le CRISP-DM.  
Des allers-retours entre les phases sont requises.*

## 11.5 Quelques techniques de data mining

- Le principal domaine de la statistique utilisé dans le data mining est l'**analyse de données multivariées**. Plusieurs mesures ou observations ( $\leadsto$  variables, attributs) sont relevées sur un ou plusieurs échantillons d'individus.
- Les principales activités du data mining sont :
  - ▷ la classification,
  - ▷ l'estimation,
  - ▷ la prédiction,
  - ▷ les règles d'associations logiques,
  - ▷ le clustering (groupement),
  - ▷ la description et visualisation.

- Les trois premières activités du data mining regroupent des méthodes d'analyse de données multivariées appartenant à

**l'apprentissage supervisé**

et les trois dernières à

**l'apprentissage non supervisé**.



- **apprentissage supervisé :**

(prédition d'appartenance à une classe, découverte directe de connaissances)

**Contexte :** toutes les observations d'un jeu de données sont classées dans des groupes connus. Une structure est donnée.

**Objectif :** en utilisant les données, constituer un modèle qui permettra de décrire une variable particulière (une entrée ou une réponse) en fonction des attributs des données restantes.

- **apprentissage supervisé (suite) :**

**Exemples :**

1. Contrôle de qualité : production défective / non défective

- ▷ déterminer l'origine de pièces défectueuses. En d'autres termes, déterminer les caractéristiques d'une partie défectueuse de la production;
- ▷ détecter des pièces défectueuses en se basant sur certains paramètres de production.

2. Procédé de production :

- ▷ déterminer les variables qui conduisent à un meilleur rendement.

- **apprentissage non supervisé :**

(découverte de classes, découverte indirecte de connaissances)

**Contexte :** des groupes auxquels peuvent appartenir les observations ne sont pas connus.

**Objectif :** recherche de structures, de relations entre variables, de schémas ou similarités entre les observations sans l'utilisation de valeur(s) cible(s) ou groupes prédéfinis.

**Exemples :**

1. Marketing : grouper en classes les clients d'un groupe de supermarchés selon les achats réalisés.
2. Pièces de musique : grouper des pièces de musique selon certains critères.

1. Classification :

**Objectif :** examiner les caractéristiques d'un nouvel objet ou d'un nouvel individu pour déterminer son affectation à l'une des classes prédéfinies. Les variables réponse sont qualitatives (discrètes).

**Exemple :** grouper les candidats à l'attribution de crédits bancaires selon le risque jugé bas, moyen ou haut;

~> apprentissage supervisé.

## 2. Estimation :

**Objectif** : valeurs d'entrée connues, estimer une valeur de réponse inconnue à l'aide de variables explicatives. Les variables réponse sont quantitatives (continues).

**Exemples** : estimer la durée de vie d'un client, estimer la probabilité de réponse à une souscription;

~> apprentissage supervisé.

## 3. Prédiction :

**Objectif** : démarche identique à la classification et à l'estimation sauf qu'on se situe dans un état prédictif.

**Exemple** : prédire quels clients quitteront un opérateur des télécommunications dans les six prochains mois;

~> apprentissage supervisé.

## 5. Clustering (groupement) :

### 4. Règles d'associations logiques ou groupement d'affinités :

**Objectif** : déceler les objets qui vont ensemble.

**Exemple** : analyser le chariot des clients d'un supermarché (panier de la ménagère) ~> "market basket analysis";

~> apprentissage non supervisé.

**Objectif** : création d'une partition de l'ensemble des données. En marketing, on parle alors de segmentation lorsqu'il s'agit de la clientèle d'une banque ou d'une entreprise de ventes par correspondance. Au contraire de la classification, le clustering n'est pas lié à des classes prédefinies.

**Exemple** : segmentation des clients mécontents des services offerts par une banque;

~> apprentissage non supervisé.

## 6. Description et visualisation :

**Objectif** : explorer les données pour découvrir, examiner et comprendre des dépendances entre plusieurs variables (boîte à moustaches, histogramme, matrice de nuages de points, graphique dynamique), identifier les principales structures existant dans les données, détecter les valeurs atypiques ("outliers"), illustrer des résultats (diagramme en arbre et dendrogramme);

~> apprentissage non supervisé.

### Remarque :

les techniques de visualisation interviennent dans plusieurs étapes du processus d'extraction des connaissances : recherche et sélection des données pertinentes, préparation des données, modélisation, évaluation et diffusion des connaissances acquises.

Dans le reste du paragraphe, quelques méthodes statistiques de data mining seront présentées, à savoir les **règles d'associations logiques** et le **clustering** dans la catégorie **apprentissage non supervisé** ainsi que la **classification**, les **arbres de classification** et les **réseaux de neurones** pour l'**apprentissage supervisé**.

De très brèves indications sur plusieurs autres méthodes seront citées;

~> en route pour les méthodes statistiques du data mining !!!

apprentissage non supervisé

## Les règles d'associations logiques :

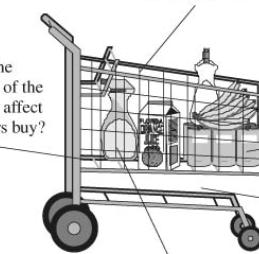
dans cet exposé introductif, on s'attardera sur l'application classique de l'apprentissage symbolique, à savoir l'**analyse du panier de la ménagère** ("Market basket analysis").

Dans une telle analyse, on est en quête de connaissances sur les clients pour savoir ce qu'ils achètent et en déduire des informations telles que

- qui sont-ils ?
- pourquoi achètent-ils certains produits ?

→ quels produits ont tendance à être achetés ensemble et lesquels sont susceptibles d'être mis en promotion ?

In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.



How do the demographics of the neighborhood affect what customers buy?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

Are window cleaning products purchased when detergent and orange juice are bought together?

What should be in the basket but is not?

*Analyse du panier de la ménagère.*

Les informations obtenues à la suite d'une analyse du panier de la ménagère sont utiles pour, par exemple,

- ouvrir de nouvelles succursales;
- savoir quels produits mettent en promotion;
- prévoir la disposition des marchandises dans le supermarché.

## Exemples :

- les libellés d'une facture d'achats par cartes de crédit, par exemple location d'une automobile et d'une chambre d'hôtel, fournissent des informations sur le prochain produit qu'un client sera susceptible d'acheter;
- les services optionnels acquis par les clients d'un opérateur de télécommunications (appels déviés, ISDN, ...) aident à organiser un regroupement des services en "bouquets" afin d'optimiser les revenus;
- les services offerts par une banque à ses clients (comptes en banque, investissements, crédits, ...) permettent d'identifier les clients susceptibles d'être intéressés par d'autres services.

La "market basket analysis" constitue un point de départ dans l'analyse lorsqu'on ignore les schémas à rechercher.

**Exemple :** échantillon d'une base de données de transactions de clients.

| TRS | NCL | Date   | Objet | Qte |
|-----|-----|--------|-------|-----|
| 111 | 201 | 5/1/99 | Stylo | 2   |
| 111 | 201 | 5/1/99 | Encre | 1   |
| 111 | 201 | 5/1/99 | Lait  | 3   |
| 111 | 201 | 5/1/99 | Jus   | 6   |
| 112 | 105 | 6/3/99 | Stylo | 1   |
| 112 | 105 | 6/3/99 | Encre | 1   |
| 112 | 105 | 6/3/99 | Lait  | 1   |
| 113 | 106 | 6/5/99 | Stylo | 1   |
| 113 | 106 | 6/5/99 | Lait  | 1   |
| 114 | 201 | 7/1/99 | Stylo | 2   |
| 114 | 201 | 7/1/99 | Encre | 2   |
| 114 | 201 | 7/1/99 | Jus   | 4   |

Par exemple, la transaction TRS 111 contient les articles {Stylo, Encre, Lait, Jus}.

Dans notre contexte, on s'intéresse aux règles d'associations logiques.

**Exemple :** 60 % des consommateurs qui achètent les produits  $X$  et  $Y$  achètent aussi  $Z$ .

On parcourt l'ensemble de toutes les règles d'associations possibles en utilisant deux mesures intéressantes : la **confiance** et le **support**.

- La **confiance** de la règle " $X \Rightarrow Y$ " est définie par :

$$X \Rightarrow Y \text{ a une confiance } c \text{ si } P(Y|X) = c.$$

- Le **support** de la règle " $X \Rightarrow Y$ " est défini par :

$$X \Rightarrow Y \text{ a un support } s \text{ si } P(Y \text{ et } X) = s.$$

Par conséquent,

$$X \Rightarrow Y \text{ a une confiance } c \text{ si } P(Y|X) = \frac{\text{support de } X \text{ et } Y}{\text{support de } X} = c.$$

- Exemple :** échantillon d'une base de données de transactions de clients (*suite*).

- Le support pour la règle " $\{\text{Stylo}\} \Rightarrow \{\text{Lait}\}$ " est

$$P(\{\text{Stylo}\} \text{ et } \{\text{Lait}\}) = \frac{3}{4} = 75\%.$$

- La confiance pour la même règle vaut

$$P(\{\text{Lait}\} | \{\text{Stylo}\}) = \frac{P(\{\text{Stylo}\} \text{ et } \{\text{Lait}\})}{P(\{\text{Stylo}\})} = \frac{0.75}{4/4} = 75\%.$$

Pour traiter les règles d'associations logiques, il existe des algorithmes (par exemple "**Apriori Algorithm**", "**Generalized Rule Induction**") qui permettent de détecter les plus intéressantes.

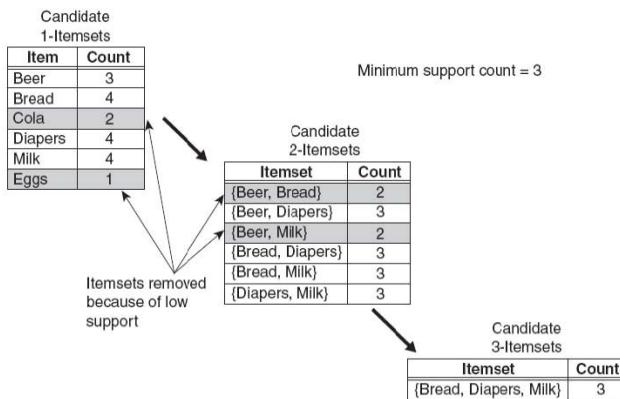


Illustration du fonctionnement de l'algorithme A priori.

(tiré de Tan, Steinbach et Kumar, 2006)

*"There are far more papers published on algorithms to discover association rules than there are papers published on applications of association rules."*

David Hand, Heikki Mannila and Padhraic Smyth

### Les règles d'associations logiques : conclusion

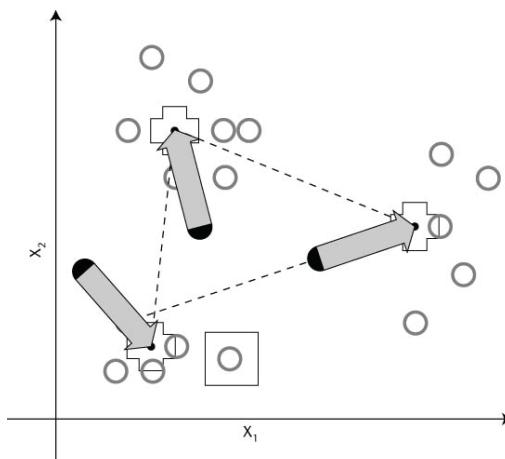
- les règles d'associations logiques permettent de savoir de quelle manière les produits / services sont reliés entre eux et comment on peut les regrouper. Elles sont faciles à comprendre et fournissent de précieuses informations;
- les algorithmes de règles d'associations logiques découvrent les associations qui pourraient également être trouvées par l'intermédiaire de techniques de visualisation.

### Le clustering (groupement) :

en clustering, on cherche à former des groupes d'observations à partir de similarités ou distances entre observations. Ce processus automatique vise à regrouper les observations dans des classes homogènes, les plus distinctes possibles.

Pour l'échantillon d'apprentissage, les groupes (catégories, clusters) ne sont pas connus *a priori* tout comme, en principe, leur nombre.

Les deux principaux types de clustering sont les **méthodes hiérarchiques** et **non hiérarchiques**.



On vise à créer des groupes homogènes.

## 1. Les méthodes hiérarchiques :

les méthodes hiérarchiques consistent

- soit à agglomérer petit à petit des groupes d'observations
  - ↗ méthode hiérarchique ascendante;
- soit à partitionner (diviser) de plus en plus les groupes d'observations
  - ↗ méthode hiérarchique descendante.

### a) clustering hiérarchique ascendant :

le clustering hiérarchique ascendant est un algorithme itératif.  
L'initialisation consiste à déterminer un tableau de distances (dissimilarités) entre les  $n$  observations à grouper.

Quelques distances possibles :

- ▷ distance euclidienne :  $d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2};$
- ▷ distance Manhattan :  $d(x, y) = \sum_{i=1}^n |x_i - y_i|.$

### a) clustering hiérarchique ascendant (suite) :

l'algorithme démarre de la partition triviale des  $n$  singletons puis cherche, à chaque étape, à constituer des classes par aggrégation (fusion) de deux éléments de la partition de l'étape précédente pour finalement s'arrêter à l'agrégation en une seule classe. Pour pouvoir l'exécuter, une distance entre les groupes est nécessaire. En notant  $C_1$  et  $C_2$  deux clusters, les choix de distances possibles sont :

- ▷ saut minimum ("single linkage") :

$$d(C_1, C_2) = \min\{d(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \in C_1, \mathbf{x}_j \in C_2\};$$

- ▷ saut maximum ou diamètre ("complete linkage") :

$$d(C_1, C_2) = \sup\{d(\mathbf{x}_i, \mathbf{x}_j) \mid \mathbf{x}_i \in C_1, \mathbf{x}_j \in C_2\};$$

### a) clustering hiérarchique ascendant (suite) :

- ▷ saut moyen (“average linkage”):

$$d(C_1, C_2) = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d(\mathbf{x}_i, \mathbf{x}_j),$$

où  $\mathbf{x}_i \in C_1$  et  $\mathbf{x}_j \in C_2$ ;  $n_1$  et  $n_2$  indiquent respectivement le cardinal des clusters  $C_1$  et  $C_2$ ;

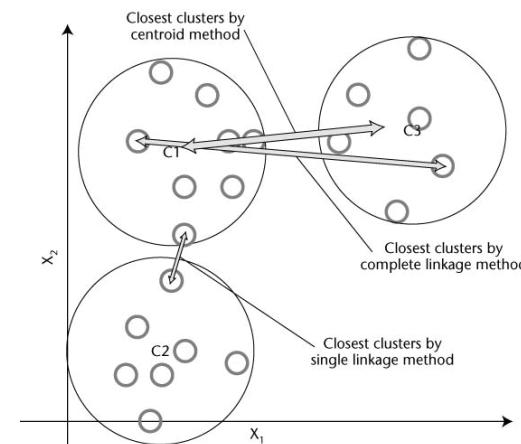
- ▷ saut de Ward :

$$d(C_1, C_2) = \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \cdot \| \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 \|^2,$$

où  $\tilde{\mathbf{x}}_1$  et  $\tilde{\mathbf{x}}_2$  sont les barycentres respectifs des clusters  $C_1$  et  $C_2$ .

Après fusion des clusters, un nouveau barycentre

$(n_1\tilde{\mathbf{x}}_1 + n_2\tilde{\mathbf{x}}_2)/(n_1 + n_2)$  est calculé.



### b) clustering hiérarchique descendant :

le clustering hiérarchique descendant est le procédé inverse de l’ascendant :

on démarre de la partition triviale de l’ensemble des observations pour s’arrêter à l’autre partition triviale, les singltons formés par les observations.

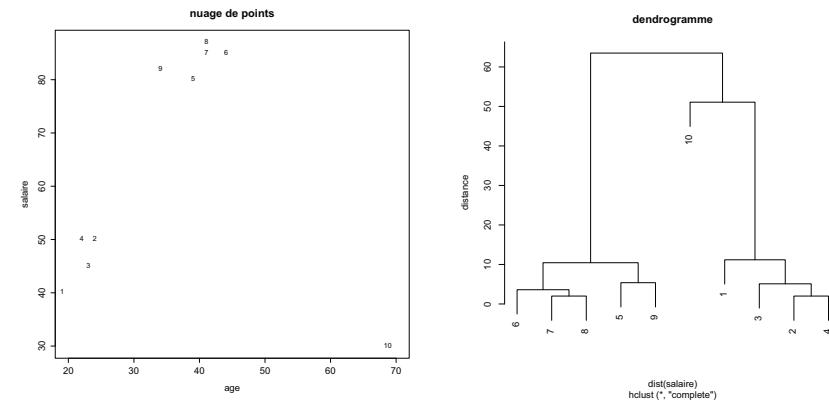
### Remarques :

- il va sans dire qu'à chaque itération des algorithmes de clustering hiérarchique, une mise à jour du tableau des distances entre clusters doit être effectuée;
- le résultat d'une analyse hiérarchique est présenté presque toujours par un graphique, le *dendrogramme*, représentation sous forme d'arbre des agrégations successives jusqu'à la réunion en une seule classe. La hauteur d'une branche est proportionnelle à la perte de variance inter-classe au niveau d'aggrégation considéré;
- le nombre de clusters à retenir est le point délicat !!! Il est souvent déterminé en maximisant le rapport de la variance inter- et intra-classe;
- en raison de métriques différentes, les méthodes hiérarchiques conduisent à des clusters aux caractéristiques différentes.

**Exemple :** jeu de données contenant deux variables continues Age et Salaire.

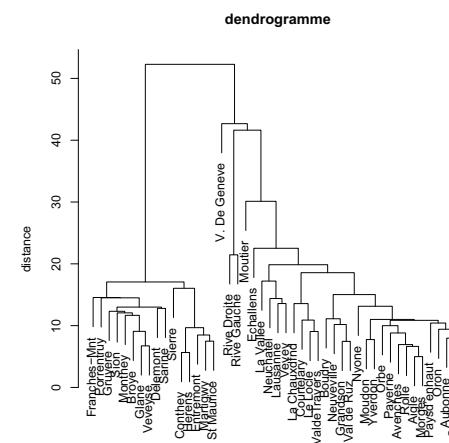
|    | Age | Salaire |
|----|-----|---------|
| 1  | 20  | 40      |
| 2  | 25  | 50      |
| 3  | 24  | 45      |
| 4  | 23  | 50      |
| 5  | 40  | 80      |
| 6  | 45  | 85      |
| 7  | 42  | 85      |
| 8  | 42  | 87      |
| 9  | 35  | 82      |
| 10 | 70  | 30      |

**Question :** combien de groupes se cachent dans ces données ?



**Exemple :** données concernant la situation socio-économique de 47 régions de Suisse Romande vers 1888.

|            | Fertilité | Agriculture | Contrôle | Education | Catholique |
|------------|-----------|-------------|----------|-----------|------------|
| Courtelary | 80.2      | 17.0        | 15       | 12        | 9.96       |
| Delemont   | 83.1      | 45.1        | 6        | 9         | 84.84      |
| ...        |           |             |          |           |            |
| Lausanne   | 55.7      | 19.4        | 26       | 28        | 12.11      |
| LaVallee   | 54.3      | 15.2        | 31       | 20        | 2.15       |
| ...        |           |             |          |           |            |
| RiveDroite | 44.7      | 46.6        | 16       | 29        | 50.43      |
| RiveGauche | 42.8      | 27.7        | 22       | 29        | 58.33      |



↗ on distingue deux groupes importants; Genève se détache des deux classes.

*"A hierarchical method suffers from the defect that it can never repair what was done in previous steps."*

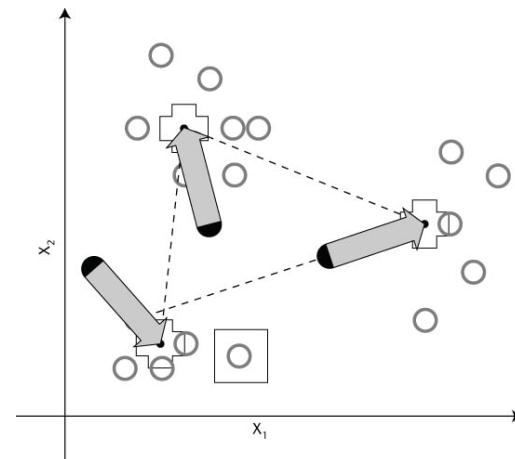
Leonard Kaufman and Peter J. Rousseeuw

### Méthode du *k*-means (méthode des nuées dynamiques) :

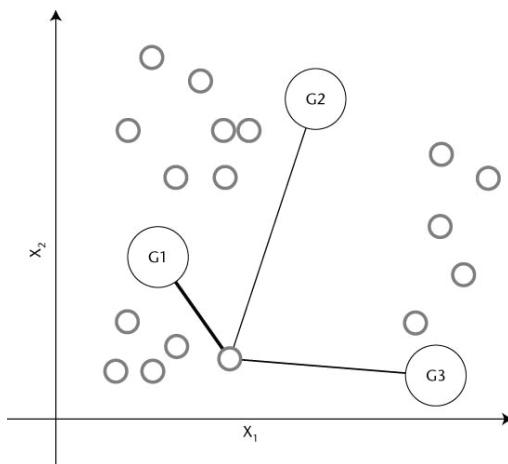
la méthode du *k*-means démarre avec *k* groupes (par exemple le nombre *k* est tiré au hasard), par conséquent, par *k* centres (ou noyaux) de clusters. À chaque itération, tout individu est remis en question dans son affectation et réassigné de façon à optimiser un certain critère, comme, par exemple, réallouer au groupe dont le centre est le plus proche au sens de la distance euclidienne. Les centres de chaque classe sont ensuite recalculés; ils deviennent les nouveaux noyaux. La procédure s'arrête lorsque les classes se sont stabilisées.

### 2. Les méthodes non hiérarchiques :

l'objectif de ces méthodes consiste à construire une partition de l'échantillon, le nombre d'éléments de la partition noté *k* étant fixé à l'avance. Différents types d'algorithmes ont été créés autour du principe de réallocation dynamique des individus à des clusters tant qu'un certain critère n'a pas été atteint. Comme exemple de critères, citons la minimisation de la somme des distances euclidiennes au carré au centre des groupes. La méthode la plus répandue est le *k*-means, méthode à centres mobiles ou encore méthode des nuées dynamiques.



*Le choix des centres initiaux est crucial !*



Toute observation est remise en question dans son affectation à chaque itération.

## Clustering : conclusion (suite)

- comparaison clustering hiérarchique et non hiérarchique :

▷ méthodes hiérarchiques :

- avantage** : mise en œuvre facile sur ordinateur (algorithme très simple à écrire);
- désavantages** : choix difficile d'un représentant caractéristique du cluster; temps d'exécution relativement long;
- la méthode fournit systématiquement un résultat identique pour le même jeu de données.

## Clustering : conclusion

- le clustering est une technique utile et facile pour explorer les données. Des observations atypiques sont rapidement détectées;
- les classes résument les données;
- les méthodes de clustering sont très peu efficaces si les classes existent déjà;
- aucune technique de clustering est bonne, mauvaise ou meilleure qu'une autre. La méthode appropriée dépend de la nature des données à analyser;

## Clustering : conclusion (suite)

- comparaison clustering hiérarchique et non hiérarchique :

▷ méthodes de partitionnement :

- avantages** : le centre d'un cluster est une représentation naturelle de la classe, un bon résumé du contenu du groupe. Les clusters formés satisfont un certain critère d'optimalité ou du moins approximativement;
- désavantages** : choix initial du nombre de classes  $k$ ; temps d'exécution relativement long;
- le résultat final dépend d'une partition initiale des données.

Quelques autres méthodes d'apprentissage non supervisé :

- **analyse en composantes principales** : réduction de la dimension et exploration de la structure des données;
- **analyse des correspondances** : analyse de tableaux de contingence en utilisant une mesure de correspondance entre les lignes ou colonnes;
- **projection poursuite** : recherche de combinaisons linéaires de variables pour faire en sorte que des clusters par rapport à ces variables apparaissent;
- **analyse factorielle** : réduction de la dimension et détection de structures existant dans les variables;
- **échelonnement multidimensionnel** : représenter des points en dimension réduite de façon à respecter au mieux des distances entre ces points.

## apprentissage supervisé

Classification et arbres de classification :

l'objectif de la classification consiste principalement à obtenir une règle optimale pour allouer un nouvel individu à l'un des groupes définis préalablement. La règle d'affectation est déterminée à partir des données qui constituent un ensemble d'apprentissage  $\mathcal{L}$ .

Les groupes sont soit connus préalablement (pièce défectueuse / non défectueuse) soit obtenus après avoir trié les observations.

La classification se base sur une partition de l'ensemble des individus en  $k$  classes :  $A_1, \dots, A_k$ . La règle d'allocation d'une observation est donnée par un opérateur de classification  $C(\cdot, \mathcal{L})$ . Ainsi, pour une observation  $x$ ,  $C(x, \mathcal{L}) = c$  si  $x$  appartient à  $A_c$ .

### Arbres de classification et de régression :

(CART : Classification And Regression Tree)

construction d'arbres binaires modélisant une classification ou une régression. Ces arbres correspondent à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (classification) ou quantitative (régression). Nous nous limiterons dans cet exposé introductif au cas des arbres de classification.

Les données sont formées de  $p$  variables quantitatives ou qualitatives et d'une variable réponse  $Y$  qualitative sur un ensemble de  $n$  individus.

Construction d'un arbre binaire de classification :

### 1. Principe : déterminer une séquence de nœuds

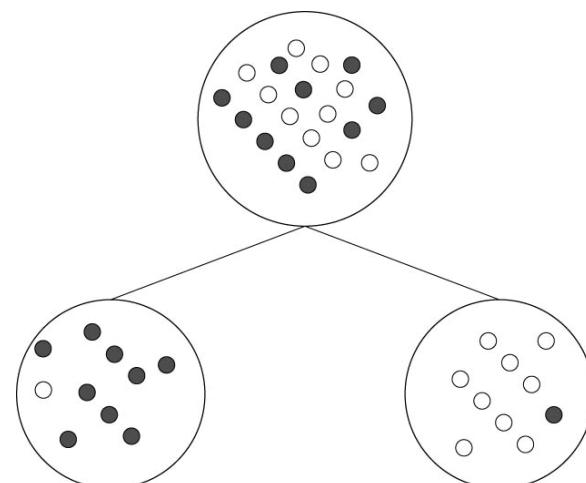
- nœud : choix conjoint d'une variable parmi les variables explicatives et d'une division;
- division : une valeur seuil pour une variable quantitative ou un partage en deux groupes selon les modalités d'une variable qualitative;  
exemples : valeur seuil 30 pour la variable âge et division en deux groupes  $< 30$  et  $\geq 30$ ; division en deux groupes selon les modalités *masculin*, *féminin* de la variable qualitative sexe;  
☞ partition de l'ensemble des individus en deux classes.

Cette procédure est itérée sur chacun des sous-ensembles ainsi créés.

Construction d'un arbre binaire de classification (*suite*) :

### 2. Critère de division :

un critère de division permet de sélectionner le "meilleur" partage parmi tous ceux admissibles (aucun des deux nœuds descendants n'est vide) en considérant toutes les variables. Il repose sur la définition d'une fonction d'"hétérogénéité" (de désordre d'un nœud), le but étant de partager les individus en deux groupes les plus "homogènes" au sens de la variable à expliquer. La meilleure division retenue pour un nœud donné est celle (variable, seuil) qui rend maximale la réduction de l'hétérogénéité sur l'ensemble des nœuds de l'arbre. Les fonctions d'hétérogénéité les plus utilisées sont celles reposant sur la notion d'entropie ou sur le critère de concentration de Gini ou encore sur une statistique de test de type  $\chi^2$ .



*Division d'un arbre de classification.*

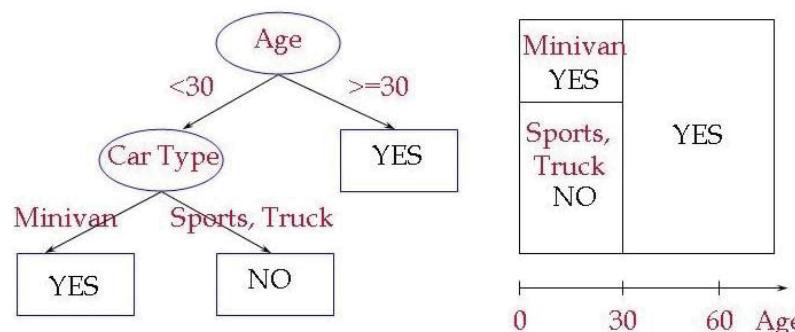
Construction d'un arbre binaire de classification (*suite*) :

### 3. Règle d'arrêt :

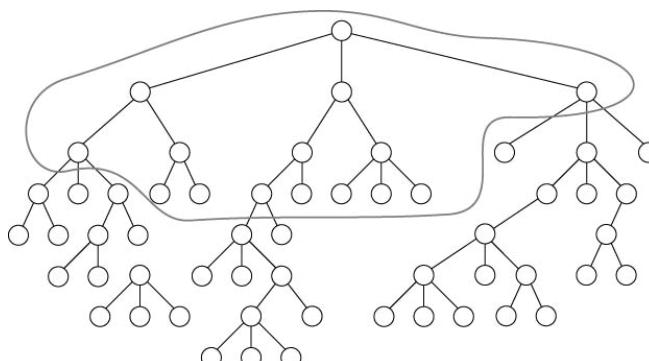
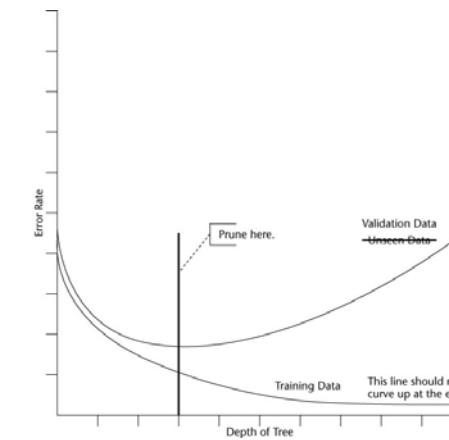
la croissance de l'arbre s'arrête à un nœud donné (nœud terminal ou feuille) lorsqu'il est homogène (plus de partition admissible ou, pour éviter un découpage inutilement fin, si le nombre d'observations qu'il contient est inférieur à une valeur limite à choisir en général entre 1 et 10).

### 4. Affectation :

chaque feuille ou nœud terminal est affecté à une modalité  $Y$ .

*Exemple d'arbre de classification.***Remarques :**

- pour évaluer la qualité de la classification, un critère raisonnable est, par exemple, le pourcentage d'observations mal classées;
- les variables explicatives qui apparaissent les premières dans l'arbre sont les plus importantes;
- toutes les variables ne vont pas forcément être utilisées pour construire l'arbre de classification;
- il est possible que des arbres extrêmement raffinés, et donc des modèles de prévision très instables, surviennent dans certaines situations complexes. On est alors confronté à des problèmes de **sur-ajustement** ("overfitting") qui peuvent être notamment traités par une procédure d'**élagage** ("pruning").

*Élagage d'un arbre de classification.**Principe pour élaguer un arbre de classification.*

### Remarques (suite) :

- e) en pratique, il s'avère que le choix du critère de division importe moins que celui du niveau d'élagage;
- f) un critère élégant pour élaguer un arbre de classification est la règle du "un écart-type". Un exercice en fin de chapitre y est consacré;
- g) dans un arbre de régression, la variable à expliquer, modéliser ou prédire est quantitative. Les principes de construction d'un arbre de régression et de classification sont les mêmes. Cependant, l'objectif du critère de division pour la régression consiste à trouver à chaque étape la variable induisant une partition en deux classes telle que la variabilité interne aux classes soit minimale et du coup celle entre les classes la plus grande. Ainsi, on remarque que le critère de division repose bien sur une notion de fonction d'hétérogénéité.

### Arbres de classification : conclusion

- les méthodes basées sur des arbres binaires sont faciles à utiliser. De plus, elles permettent de traiter une grande variété de problèmes, elles requièrent moins d'hypothèses que des méthodes classiques et semblent particulièrement adaptées au cas où les variables explicatives sont nombreuses;
- les résultats obtenus sous forme d'arbres de classification sont simples à interpréter et constituent un atout efficace pour l'aide à la décision;
- les arbres de classification contiennent des informations sur les relations existant entre variables explicatives et variables à expliquer;

### Classification : conclusion (suite)

- pour de petits jeux de données, les méthodes basées sur les arbres binaires ne parviennent pas à mettre en évidence la structure des données. En revanche, elles sont capables de découvrir les structures se trouvant dans de grandes bases de données complexes. Ces structures peuvent d'ailleurs être difficilement décelables en utilisant des modèles de régression classiques;
- les valeurs atypiques influencent peu les résultats obtenus par la méthode des arbres de classification.

### Classification : conclusion

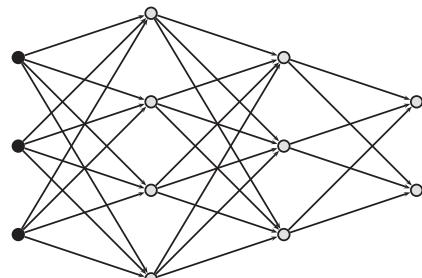
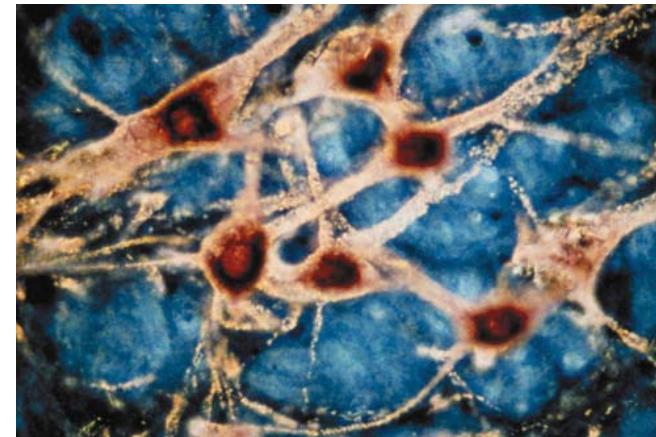
- la classification est un problème classique en statistique. Il s'agit vraisemblablement de l'une des techniques de data mining des plus répandues;
- combiner habilement la classification et les techniques des bases de données est un sujet de recherche très prometteur;
- un nombre croissant de variables ne devrait pas affecter les performances de la plupart des méthodes de classification;
- aucune technique de classification est bonne, mauvaise ou meilleure qu'une autre. La méthode appropriée dépend de la nature des données à analyser.

## Réseaux de neurones :

les réseaux de neurones formels reposent sur le principe du fonctionnement du cerveau humain. Très souples et adaptatifs, ils permettent après un processus d'apprentissage une prédiction de nouvelles observations.

Les principaux réseaux de neurones se distinguent par le nombre de neurones, l'organisation du graphe orienté (en couches, complets, ...), i.e l'architecture, et par le mode de calcul.

Dans ce paragraphe, nous nous limiterons au perceptron multicouches (PMC). Il est composé de couches successives comme l'indique la figure de la page suivante. Les neurones d'une couche ne sont pas connectés entre eux.

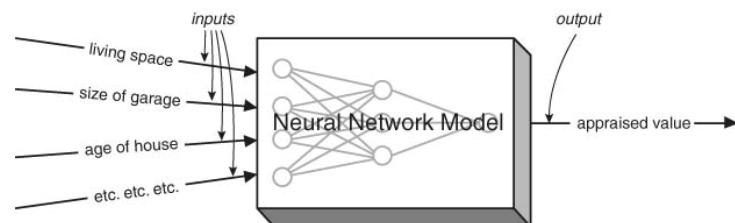


*Exemple de perceptron multicouches : trois couches avec deux couches cachées.*

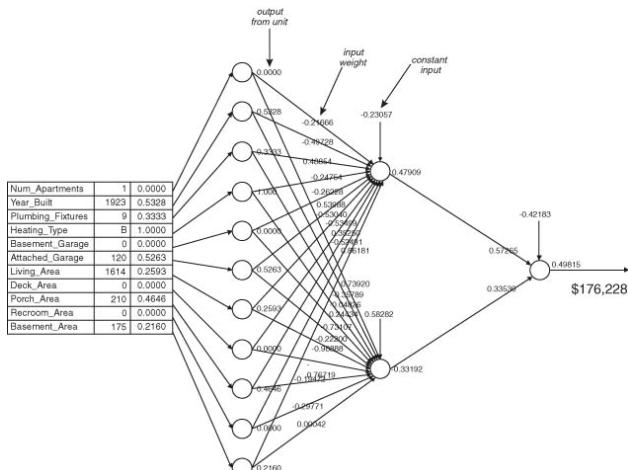
*Les unités d'entrée sont indiquées en noir.*

## Mode de calcul

- Dans un perceptron multicouches, une couche d'entrée lit les signaux entrant ("inputs"), un neurone par entrée  $x_j$  ( $j = 1, \dots, p$ ). Une couche en sortie fournit la réponse du système  $y$  ("output"). La réponse peut être quantitative ou qualitative.
- Un perceptron multicouches réalise une suite de transformations vectorielles : une couche reçoit un vecteur d'entrée et le transforme en un vecteur de sortie. Les dimensions d'entrée et de sortie peuvent être différentes.
- Les transformations vectorielles se réalisent sur toutes les couches, l'une après l'autre.



Modélisation à l'aide d'un réseau de neurones.



Fonctionnement d'un perceptron multicouches.

### Mode de calcul (suite)

- Plus précisément, le  $k$ -ième neurone de la première couche reçoit les signaux d'entrée ("inputs")  $x_j$  ( $j = 1, \dots, p$ ) et les transforme en une sortie  $z_k$  selon une fonction d'activation  $f$  de telle sorte que

$$z_k = \phi(x_1, \dots, x_p) = f\left(\alpha_{k0} + \sum_{j=1}^p \alpha_{kj} x_j\right),$$

où les  $p + 1$  paramètres  $\alpha_{kj}$  sont inconnus. Les valeurs des paramètres seront estimées dans la phase d'apprentissage.

- Les paramètres  $\alpha_{kj}$  ( $j = 1, \dots, p$ ) sont appelés les poids synaptiques (poids de connexion) et  $\alpha_{k0}$  est le biais.
- Par commodité, les poids synaptiques sont inscrits au-dessus des arcs.

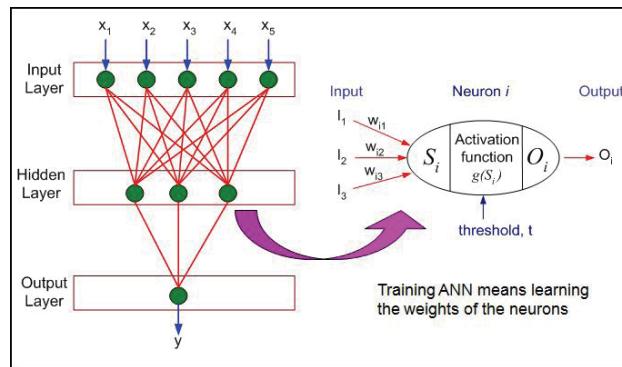
### Mode de calcul (suite)

- Plusieurs choix sont possibles pour la fonction d'activation  $f$  :

$$\begin{aligned} & \triangleright f(u) = u; \\ & \triangleright f(u) = 1/(1 + e^{-u}). \end{aligned}$$

Ces deux fonctions sont très pratiques pour faciliter la phase d'apprentissage du réseau de neurones. D'autres fonctions existent mais s'adaptent moins bien aux algorithmes utilisés dans le processus d'apprentissage.

- La fonction d'activation est aussi appelée fonction de transfert.



Fonctionnement d'un perceptron à une couche cachée.

(tiré de Tan, Steinbach et Kumar, 2006)

### Mode de calcul (suite)

- La deuxième couche du perceptron reçoit en entrée les  $K$  valeurs de la première couche et selon le même principe que pour la première couche les transforme en un vecteur de sortie.
- Le procédé se répète aux couches restantes du perceptron selon le même principe.
- Tous les neurones d'une même couche d'un perceptron auront la même fonction d'activation. Les couches peuvent avoir des fonctions d'activation différentes. Au moins dans une des couches, la fonction de transfert ne sera pas linéaire ( $f(u) = u$ ).

### Phase d'apprentissage

- La procédure d'apprentissage du perceptron s'effectue à partir de  $n$  observations  $(x_{i1}, \dots, x_{ip}; y_i)$ ,  $i = 1, \dots, n$ .
- Par commodité, tous les poids synaptiques et tous les biais du réseau sont regroupés dans le vecteur  $\alpha$ .
- Si la variable réponse  $y$  du système est quantitative, la phase d'apprentissage consiste à déterminer le vecteur des poids  $\hat{\alpha}$  qui minimise la fonction

$$Q(\alpha) = \sum_{i=1}^n \left( y_i - \phi(x_{i1}, \dots, x_{ip}; \alpha) \right)^2,$$

où  $\phi(x_{i1}, \dots, x_{ip}; \alpha)$  représente le modèle induit par le réseau de neurones selon ses fonctions d'activation.

### Phase d'apprentissage (suite)

- En minimisant la fonction  $Q$ , on cherche à approcher au mieux les valeurs observées  $y_i$  à l'aide du réseau de neurones au sens des moindres carrés.
- L'algorithme de **rétropropagation du gradient** permet de modifier itérativement les poids de chaque neurone pour faire décroître la fonction  $Q$ .

### Critères d'arrêt

- borne sur le temps de calcul (nombre d'itérations de l'algorithme);
- valeur à atteindre (on s'arrête quand la fonction  $Q$  passe en dessous d'un seuil);
- vitesse de progression (on s'arrête quand la fonction  $Q$  ne décroît plus suffisamment ou quand le vecteur  $\alpha$  se stabilise).

### Remarque :

les modèles issus de réseaux de neurones sont “sur-paramétrés”. Ainsi, on est confronté à des problèmes de **sur-ajustement (“overfitting”)**.

La première méthode pour y remédier consiste à construire le plus petit réseau capable d'ajuster l'échantillon d'apprentissage. De plus, la phase d'apprentissage sur un tel réseau est plus rapide.

Pour pallier les problèmes de “sur-paramétrisation”, les stratégies plus récentes se basent sur la **régulation**. On cherche un bon compromis entre l'ajustement et la variation. Une possibilité pour y parvenir consiste à introduire une fonction de pénalisation dans le critère à optimiser :  $Q(\alpha) + \lambda \cdot J(\alpha)$ . Plus le paramètre  $\lambda > 0$  est grand, plus les poids tendront vers 0. Des méthodes permettent efficacement d'estimer  $\lambda$ .

### Réseaux de neurones : conclusion (suite)

- les perceptrons multicouches possèdent d'indéniables qualités en particulier lorsque le nombre de variables explicatives (“inputs”) rend les modèles statistiques traditionnels inutilisables;
- les critiques principales énoncées à l'encontre du perceptron multicouches concernent
  - ▷ les difficultés liées à l'apprentissage  
(temps de calcul, taille de l'échantillon, stagnation à un optimum local défavorable, expérience de l'utilisateur);
  - ▷ son statut de boîte noire;
  - ▷ la difficulté d'interprétation des résultats même pour un expert.

### Réseaux de neurones : conclusion

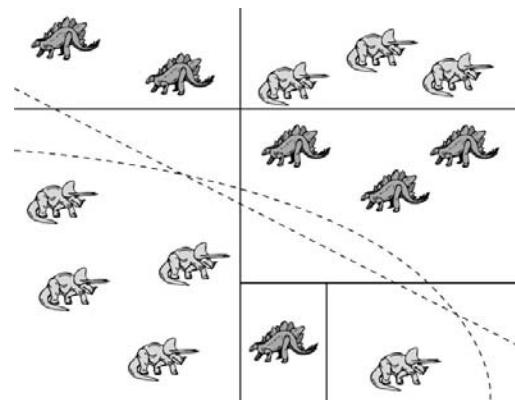
- les réseaux de neurones sont très répandus et très utilisés. Les modèles prédictifs qui en découlent sont efficaces en régression et classification;
- les réseaux de neurones sont très souples et très adaptatifs. Ils peuvent modéliser quasiment n'importe quoi !
- les champs d'application des perceptrons multicouches sont multiples : discrimination, prévision d'une série temporelle, reconnaissance de forme, . . .;

*“Neural networks are statistics for amateurs. A properly designed network, when learning and responding, performs good statistical inference, based on what it saw when it learned and what it sees when it responds. Most neural networks conceal the statistics from the user.”*

James A. Anderson, Andras Pellionisz and Edward Rosenfeld

*"There has been a great deal of hype surrounding neural networks, making them seem magical and mysterious. . . they are just nonlinear statistical models."*

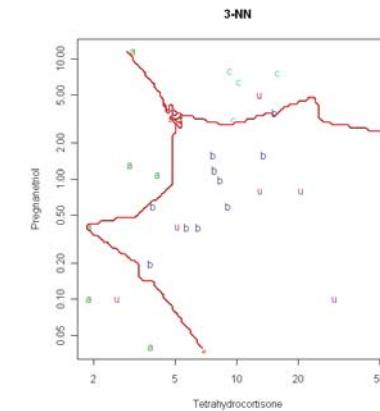
Travor Hastie, Robert Tibshirani and Jerome Friedman



Classification par arbre et discriminations linéaire et quadratique.

Quelques autres méthodes d'apprentissage supervisé :

- **analyse discriminante** : recherche des variables permettant de séparer le mieux possible deux ou plusieurs groupes. L'analyse discriminante (linéaire, quadratique, logistique) permet, lorsque cette séparation est bien marquée, d'assigner un nouvel individu à l'un des groupes;
- **classification par les voisins les plus proches** : la classification par les voisins les plus proches s'effectue en utilisant une distance entre les observations comme par exemple la distance euclidienne ou un moins la corrélation entre deux variables. La classification d'une nouvelle observation  $x$  s'effectue en deux étapes : déterminer les  $k$  voisins les plus proches de l'ensemble d'apprentissage puis prédire la classe d'affectation de  $x$  par le système de vote majoritaire;



Classification par les voisins les plus proches.

- classification par agrégation** : construction de prédicteurs d'agrégation à partir de versions perturbées de l'ensemble d'apprentissage. Citons par exemple le “bagging” (agrégation par rééchantillonnage i.e par réPLICATION d'échantillons –“bootstrap”), les forêts aléatoires (amélioration du “bagging” par l'ajout d'une randomisation) et le “boosting” (rééchantillonnage adaptatif et autre système de vote);
- analyse de variance multivariée** : test pour différences significatives entre moyennes en comparant / analysant les variances;
- régression multiple** : analyse des relations entre plusieurs variables explicatives indépendantes (“input”) et une variable réponse (“output”). Les relations peuvent être linéaires ou non linéaires. La régression est une méthode d'apprentissage supervisé très utilisée. Elle fera l'objet d'un travail pratique pour le cas linéaire (modèles statistiques linéaires);

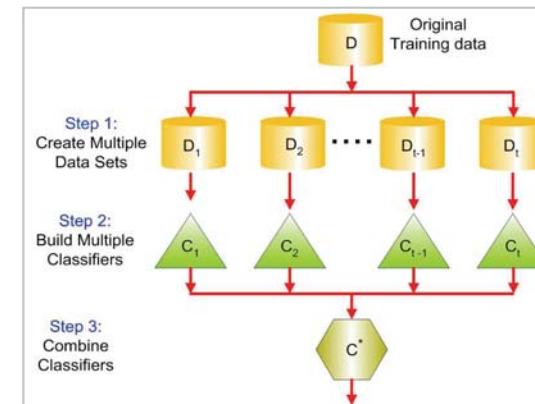


Illustration du fonctionnement de la classification par agrégation.

(tiré de Tan, Steinbach et Kumar, 2006)

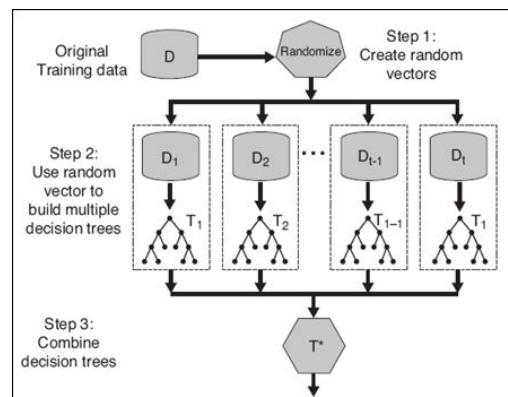
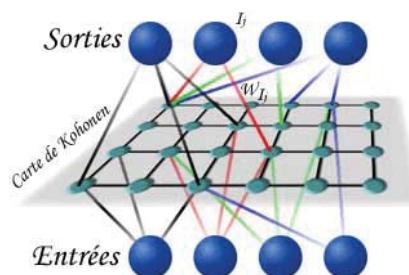


Illustration de la construction de forêts aléatoires.

(tiré de Tan, Steinbach et Kumar, 2006)

- cartes auto-organisatrices de Kohonen** : comme les réseaux de neurones, elles s'inspirent du fonctionnement du cerveau. Elles se basent en fait sur la correspondance topologique entre les zones du cerveau et les capteurs sensoriels : par exemple, deux zones proches dans le cortex visuel correspondent à deux zones proches sur la rétine. Inspiré par ce principe de correspondance, Teuvo Kohonen proposa un modèle de carte topologique auto-adaptative.

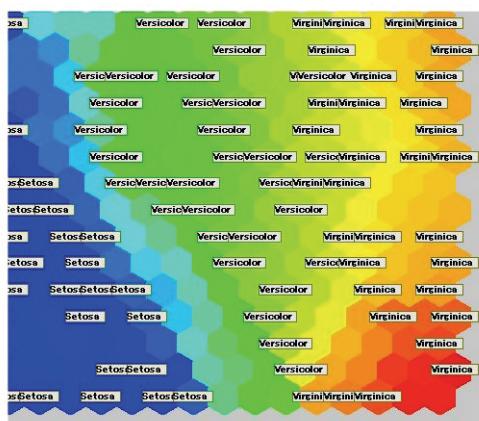
Une carte de contrôle de Kohonen est formée d'un réseau de neurones, en général à deux dimensions. Chaque neurone de la carte est relié à chaque neurone d'une couche d'entrée.



Architecture d'un modèle de Kohonen.

- **cartes auto-organisatrices de Kohonen (suite)** : grossièrement, la couche d'entrée va stimuler et activer les neurones de la carte. Pour représenter leurs réactions aux stimuli, une carte en relief est construite. Pour un ensemble de stimuli, certains neurones réagiront plus que d'autres. Ainsi, les endroits les plus intéressants sur la carte en relief seront les "pics", autrement dit les endroits qui correspondent aux neurones les plus sensibles aux stimuli.

Les cartes de Kohonen sont utilisées dans la recherche de groupes dans un ensemble de données (processus d'apprentissage non supervisé) ou dans la classification (démarche d'apprentissage supervisé). Dans la classification, chaque neurone de la carte est affecté à l'une des classes après une phase d'apprentissage. Il est possible qu'un neurone soit attribué à aucune classe. Un nouvel individu sera alloué au groupe correspondant au neurone le plus stimulé par son passage.



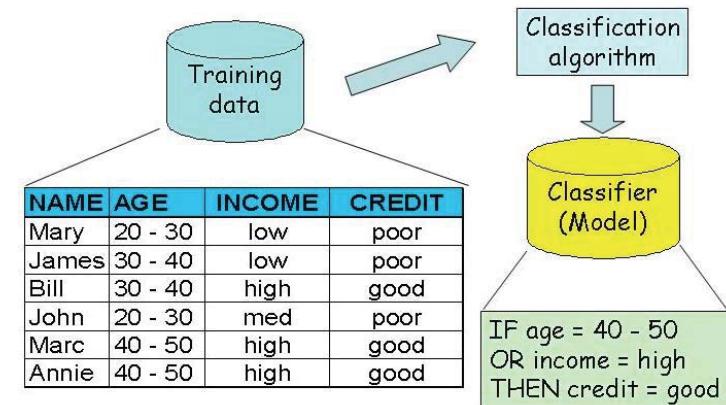
Classification des 150 iris réparties en 3 groupes  
à l'aide d'une carte auto-organisatrice de Kohonen.

*"All models are wrong, but some are useful."*

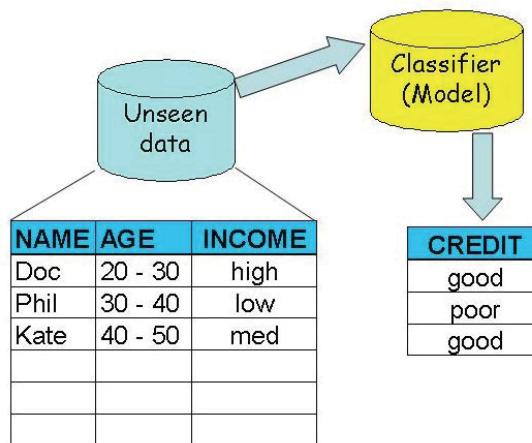
George E. P. Box

### Remarques :

- a) le choix de la technique du data mining dépend de la nature de la tâche à accomplir et de la nature des données;
- b) les techniques utilisées dans le data mining sont complémentaires : elles sont le plus souvent employées conjointement pour améliorer la qualité des résultats produits;
- c) les méthodes utilisées par le data mining permettent l'exploration, la description et la prévision, les trois facettes essentielles de l'aide à la décision;
- d) n'oubliez jamais la phase de validation des modèles prédictifs constitués !  
Trois approches existent : échantillon test, validation croisée et "bootstrap" (rééchantillonnage, réplication d'échantillons).



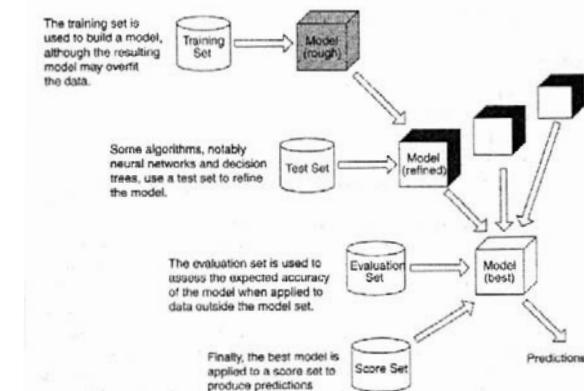
Classification par échantillon test.



Classification par échantillon test (suite).



Le "bootstrap"; une origine bien étrange...



Démarche de validation de modèles prédictifs.

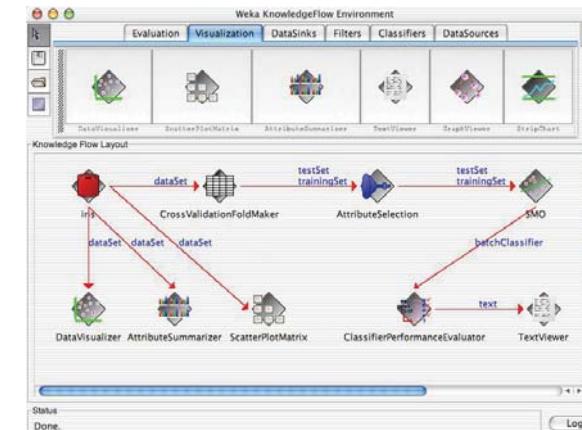
- le data mining n'est pas qu'une technologie : une part importante des performances des outils est conditionnée par une bonne formulation du problème. La formation des utilisateurs et l'intégration des responsables fonctionnels sont primordiales pour toute démarche du data mining;
- la réussite d'un projet de data mining s'appuie, d'une part sur une bonne compréhension des méthodes statistiques et algorithmiques utilisées, d'autre part sur une interprétation des résultats effectuée avec aisance et assurance;
- dans beaucoup de situations mal contrôlées, les méthodes statistiques doivent être utilisées avec prudence ! Ne jamais considérer le data mining comme une boîte noire !

## 11.6 Conclusion

- le data mining, prospection ou fouille de données, est actuellement connu; on ignore souvent ce qu'il représente;
- le data mining, composante de la recherche des connaissances, comprend un sous-ensemble de méthodes statistiques et algorithmiques → analyse de données multivariées;
- une utilisation intelligente des techniques de data mining représente une réelle opportunité;
- la technologie data mining n'est pas autonome. Les techniques du data mining se fondent progressivement dans les outils existants;

- il n'y a pas de choix *a priori* d'une méthode statistique. Il est recommandé d'appliquer plusieurs techniques en parallèle. Uniquement l'expérience et une validation soignée permettent de se déterminer. C'est la raison pour laquelle des logiciels généralistes comme SAS, module Entreprise Miner, Clementine de SPSS, STATISTICA ne font pas de choix et offrent ces méthodes en parallèle pour mieux s'adapter aux données, aux habitudes des utilisateurs et à la mode;
- la protection de la vie privée va devenir une préoccupation importante dans le domaine du data mining;

- un marché important, un secteur de recherche très prometteur i.e non seulement le développement de nouvelles méthodes statistiques et algorithmiques (par exemple les algorithmes adaptatifs permettant une mise à jour et une intégration des connaissances acquises en temps réel) mais aussi d'interfaces (par exemple celle entre le logiciel statistique R et les bases de données relationnelles) et d'outils graphiques sophistiqués (graphiques dynamiques);
- le WEB mining et le Text mining (statistique textuelle) sont actuellement des domaines d'application à la mode du data mining. Le WEB mining se propose d'analyser le comportement d'internautes sur un site (marchand ou non). Le Text mining s'applique à des données textuelles et cherche des connaissances et relations à partir de documents disponibles;
- les algorithmes bio-inspirés et les Supports Vector Machines (SVM, fréquemment utilisés dans la discrimination et dans la régression) sont aussi d'actualité comme techniques de data mining.



Interface de WEKA, logiciel libre de data mining.

*"The problem isn't that specialised companies lack the data they need, it's that they don't go and look for it, they don't understand how to handle it."*

Hans Rosling

*"Understanding the models, particularly their limitations and sensitivity to assumptions, is the new task we face. Many of the banking and financial institution problems and failures of the past decade can be directly tied to model failure or overly optimistic judgements in the setting of assumptions or the parameterization of a model."*

Tad Montross, chairman and CEO of GenRe in "Model Mania"

## 11.7 Références et ressources

### Livres et articles

- Berry, M. J. A. & Linoff, G. S. (2000). *Mastering Data Mining*. New York: Wiley.
- Berry, M. J. A. & Linoff, G. S. (2004). *Data Mining Techniques for Marketing, Sales and CRM* (2nd, extended edition). New York: Wiley.
- Berthold, M. & Hand, D. J., Eds. (2003). *Intelligent Data Analysis* (2nd, extended edition). Berlin: Springer.
- Besse, Ph. & Baccini, A. (2005). Data mining I. Exploration Statistique. *Publications du Laboratoire de Statistique et Probabilités*, Université Paul Sabatier, Toulouse.

- Besse, Ph. (2005). Data mining II. Modélisation Statistique & Apprentissage. *Publications du Laboratoire de Statistique et Probabilités*, Université Paul Sabatier, Toulouse.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.
- Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Kuonen, D. (2005). Methodological Training in Statistical Data Mining Related to Drug Development. Course in Statistics. Statoox Consulting, Lausanne.
- Fayyad, U. M., Pitagetsky-Shapiro, G., Smyth, P. & Uthurusamy, R., Eds. (1996). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press.

- Giudici, P. (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. London: Wiley.
- Han, J. & Kamber, M. (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.
- Hand, D. J., Mannila, H. & Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
- Hastie, T., Tibshirani, R. & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Rajman, M. (1998). Data Mining : la ruée vers l'or gris, Le Data Mining : qu'est-ce que c'est ? *Proceedings des rencontres du CAST*, EPFL.
- Tan, P. N., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining*. New-York: Addison Wesley

### Ressources

- 'Boosting Research' site:  
[www.boosting.org](http://www.boosting.org)
- Breiman's homepage:  
[www.stat.berkeley.edu/users/breiman/](http://www.stat.berkeley.edu/users/breiman/)
- 'CRISP-DM Process Guide and User Manual':  
[www.crisp-dm.org](http://www.crisp-dm.org)
- Friedman's homepage:  
[www-stat.stanford.edu/~jhf](http://www-stat.stanford.edu/~jhf)
- Hastie's homepage:  
[www-stat.stanford.edu/~hastie](http://www-stat.stanford.edu/~hastie)

- KDnuggets site (portail du data mining avec toute l'actualité du domaine):  
[www.kdnuggets.com/](http://www.kdnuggets.com/)
- Ricco Rakotomalala's homepage (liens, supports de cours en ligne):  
[chirouble.univ-lyon2.fr/~ricco/cours/](http://chirouble.univ-lyon2.fr/~ricco/cours/)
- Statoo Consulting's data mining related links:  
[www.statoo.com/en/resources/anthill/Datamining/](http://www.statoo.com/en/resources/anthill/Datamining/)
- Web-datamining site (portail français du data mining avec toute l'actualité du domaine):  
[www.web-datamining.net/](http://www.web-datamining.net/)
- Weka 3: Data Mining Software in Java  
[www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/)

*"The world we live in is awash with data, that comes pouring in from everywhere around us. On its own, this data is just noise and confusion. To make sense of data, to find the meaning in it, we need a powerful branch of science: statistics!"*

Hans Rosling

*"If you are looking for a career where your services will be in high demand, you should find something where you provide a scarce, complementary service to something that is getting ubiquitous and cheap. So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."*

Hal Varian, chief economist at Google





## EXERCICES : INTRODUCTION AU DATA MINING ORIENTÉ VERS LE BUSINESS

### Exercice 1

Dans une analyse du panier de la ménagère, on est en quête de connaissances sur les clients pour, par exemple, déterminer les produits qui ont tendance à être achetés ensemble. Considérons un échantillon d'une base de données comprenant les achats réalisés dans un supermarché :

| TRS | NCL | Date   | Objet | Qte |
|-----|-----|--------|-------|-----|
| 111 | 201 | 5/1/99 | Stylo | 2   |
| 111 | 201 | 5/1/99 | Encre | 1   |
| 111 | 201 | 5/1/99 | Lait  | 3   |
| 111 | 201 | 5/1/99 | Jus   | 6   |
| 112 | 105 | 6/3/99 | Stylo | 1   |
| 112 | 105 | 6/3/99 | Encre | 1   |
| 112 | 105 | 6/3/99 | Lait  | 1   |
| 113 | 106 | 6/5/99 | Stylo | 1   |
| 113 | 106 | 6/5/99 | Lait  | 1   |
| 114 | 201 | 7/1/99 | Stylo | 2   |
| 114 | 201 | 7/1/99 | Encre | 2   |
| 114 | 201 | 7/1/99 | Jus   | 4   |

Par exemple, la transaction TRS 111 contient les articles {Stylo, Encre, Lait, Jus}.

Dans l'échantillon des transactions, calculer le support et la confiance des règles

- a) “{Stylo}  $\Rightarrow$  {Lait}”;
- b) “{Encre}  $\Rightarrow$  {Stylo}”.

Solutions : a) 3/4, 3/4 b) 3/4, 1.

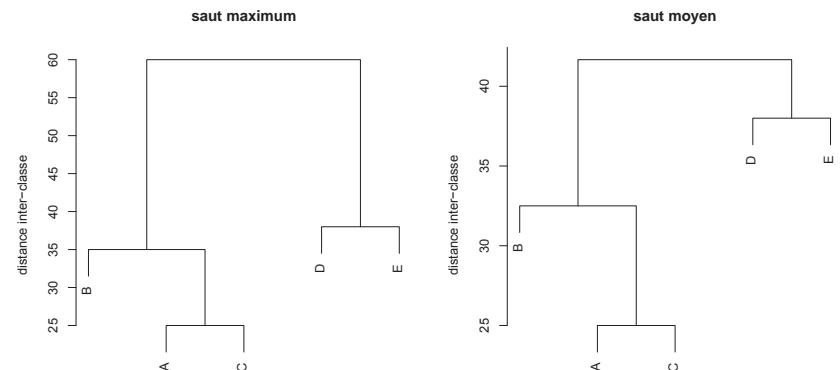
### Exercice 2

Dans le tableau ci-dessous figurent les distances séparant 5 villes notées A, B, C, D et E :

|   | A  | B  | C  | D  | E  |
|---|----|----|----|----|----|
| A | 0  | 35 | 25 | 40 | 55 |
| B | 35 | 0  | 30 | 60 | 40 |
| C | 25 | 30 | 0  | 25 | 30 |
| D | 40 | 60 | 25 | 0  | 38 |
| E | 55 | 40 | 30 | 38 | 0  |

- a) Construire le dendrogramme du clustering hiérarchique ascendant (par agglomération, fusion) en utilisant le saut maximum (“complete linkage”) comme distance inter-classes.
- b) Construire le dendrogramme du clustering hiérarchique ascendant en utilisant le saut moyen (“average linkage”) comme distance inter-classes.
- c) Les deux distances utilisées conduisent-elles aux mêmes dendrogrammes ?

Solutions : a) et b)

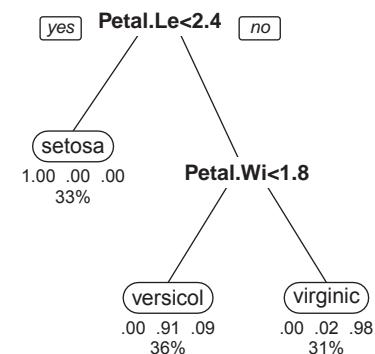


c) en principe, les dendrogrammes basés sur le saut maximum et sur le saut moyen ne sont pas les mêmes. Cependant, ils peuvent ne pas se différencier sur les groupes mais sur les distances comme dans cet exercice.

### Exercice 3

Dans cet exercice, nous introduirons les arbres de classification implémentés par T. M. Therneau et E. J. Atkinson en 1997. Les méthodes seront appliquées sur 150 iris réparties en 3 groupes. Les données ont été présentées au chapitre consacré à l'analyse exploratoire des données.

L'arbre de classification des iris obtenu à l'aide du logiciel de statistique R se trouve ci-dessous.



L'arbre est formé de 3 feuilles et possède 2 divisions. En se basant sur les résultats ci-dessous obtenus à l'aide de **R** doit-on élaguer l'arbre de classification et le rendre optimal à l'aide de la règle du "un écart-type" en partant d'un paramètre de complexité CP fixé à 0.01 ?

|   | CP   | nsplit | relerror | xerror | xstd   |
|---|------|--------|----------|--------|--------|
| 1 | 0.50 | 0      | 1.00     | 1.18   | 0.0502 |
| 2 | 0.44 | 1      | 0.50     | 0.68   | 0.0610 |
| 3 | 0.01 | 2      | 0.06     | 0.09   | 0.0291 |

**Solution :** en appliquant la règle du "un écart-type", nous choisirons l'arbre complet. En effet, la valeur limite vaut  $0.09 + 0.0291 = 0.1191$ . La plus grande valeur de l'erreur de validation croisée **xerror** inférieure à la valeur limite 0.1191 est 0.09, valeur qui correspond à un nombre de divisions **nsplit** égal à 2. Ainsi, aucun élagage n'est nécessaire.

#### Exercice 4

Des débris de verre ont été prélevés et classés dans 6 groupes selon leur origine (fenêtres de maison, pare-brise, bouteilles par exemple). On a mesuré en laboratoire les quantités de différents composants (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe) dans ces débris.

- a) On se propose de construire un arbre de classification avec un paramètre de complexité fixé à 0.001 en prenant l'origine des débris de verre **WinF**, **WinNF**, **Veh**, **Con**, **Tabl** et **Head** comme variable réponse et les composants de ces débris comme variables de classification.
1. L'arbre de classification construit à l'aide de la librairie **rpart** de **R** se trouve dans la Figure 2. Quelle est la variable la plus importante ?

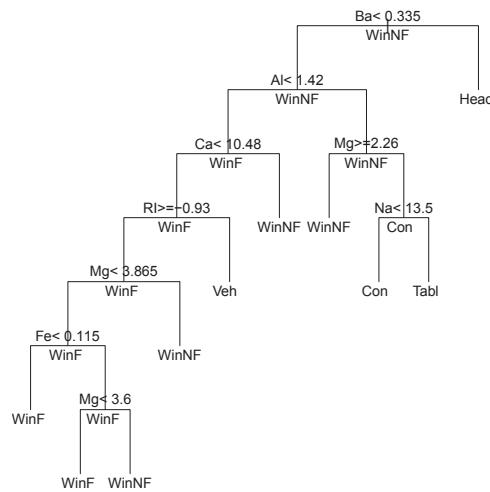


Figure 2: Arbre de classification de débris de verre.

2. Comme on peut le remarquer dans la Figure 2, l'arbre de classification est trop grand et doit être élagué. Pour le faire on fait appel à la règle du "un écart-type". Le graphique permettant d'élaguer l'arbre à l'aide de cette méthode se trouve dans la Figure 3. Quel est le nombre de divisions et la taille de l'arbre élagué par application de la règle du "un écart-type" ?

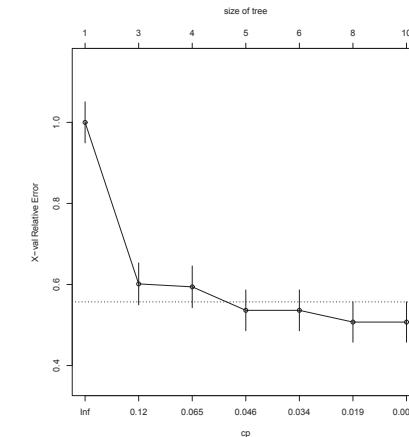


Figure 3: Graphique permettant d'élaguer l'arbre de classification.

3. Vrai ou Faux. Justifiez brièvement votre réponse.
- La taille de l'arbre de classification croît lorsque le paramètre de complexité décroît.
  - On élague un arbre de classification pour obtenir un bon compromis entre l'ajustement et l'efficacité de prédiction.
- b) On se propose de comparer la performance de classification de l'arbre de décision à celle des Séparateurs à Vaste Marge (en anglais Support Vector Machines, SVM) pour notre jeu de données. Pour comparer les prédictions, on utilise la validation croisée, méthode de rééchantillonnage.
1. Dans le tableau ci-dessous se trouvent les prédictions des classes espérées des 214 débris données par des arbres de classification ainsi que pour chaque débris sa classe observée.

| observed | predicted |       |     |     |      |      |
|----------|-----------|-------|-----|-----|------|------|
|          | WinF      | WinNF | Veh | Con | Tabl | Head |
| WinF     | 57        | 12    | 0   | 0   | 0    | 1    |
| WinNF    | 17        | 54    | 0   | 4   | 0    | 1    |
| Veh      | 12        | 4     | 1   | 0   | 0    | 0    |
| Con      | 0         | 3     | 0   | 9   | 0    | 1    |
| Tabl     | 2         | 4     | 0   | 3   | 0    | 0    |
| Head     | 2         | 1     | 0   | 0   | 0    | 26   |

Calculer le pourcentage des débris mal classés par les arbres de décision.

2. Dans le tableau ci-dessous figurent les prédictions des classes espérées des 214 débris données par des séparateurs à vaste marge ainsi que pour chaque débris sa classe observée.

| observed | predicted |       |     |     |      |      |
|----------|-----------|-------|-----|-----|------|------|
|          | WinF      | WinNF | Veh | Con | Tabl | Head |
| WinF     | 43        | 22    | 5   | 0   | 0    | 0    |
| WinNF    | 15        | 56    | 4   | 0   | 0    | 1    |
| Veh      | 8         | 6     | 3   | 0   | 0    | 0    |
| Con      | 0         | 9     | 0   | 4   | 0    | 0    |
| Tabl     | 1         | 5     | 0   | 0   | 3    | 0    |
| Head     | 0         | 9     | 0   | 0   | 0    | 20   |

Calculer le pourcentage des débris mal classés par les séparateurs à vaste marge.

3. Comparer les résultats obtenus en 1. et en 2. par les deux opérateurs de classification. Quel est le meilleur opérateur pour notre jeu de données ?

**Solutions :** a) 1. Ba 2. nombre de divisions : 4, taille de l'arbre : 5 3. i) vrai ii) vrai b) 1. 31.31% 2. 39.72% 3. le meilleur opérateur de classification pour notre jeu de données est l'arbre de classification.

### Exercice 5

Le prix d'achat et l'impôt annuel ont été relevés pour vingt-quatre maisons. On se demande s'il existe une relation linéaire entre ces deux variables. L'impôt constitue la variable explicative indépendante et le prix d'achat est la variable réponse, variable dépendante.

- a) Obtenus à l'aide du logiciel de statistique R, les résultats partiels d'une régression linéaire simple ajustée aux observations sont

```
Call:
lm(formula = prix.vente ~ impot, data = maisons
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.32      2.57    5.18 3.4e-05 ***
impot        3.32      0.39    8.52 2.1e-08 ***
...
Residual standard error: 2.96 on 22 degrees of freedom
Multiple R-Squared: 0.767, Adjusted R-squared: 0.757
F-statistic: 72.6 on 1 and 22 DF, p-value: 2.05e-08
```

Que représente le coefficient de détermination  $R^2$  ?

- b) Que teste-t-on au moyen des graphiques de la Figure 5 ? Effectuer un premier diagnostic du modèle à l'aide de ces graphiques.  
c) En utilisant les résultats obtenus par R en a), tester la nullité de la pente de la droite de régression à un niveau de signification de 1%.  
d) Déterminer la valeur de vente espérée pour un impôt de 7.50.

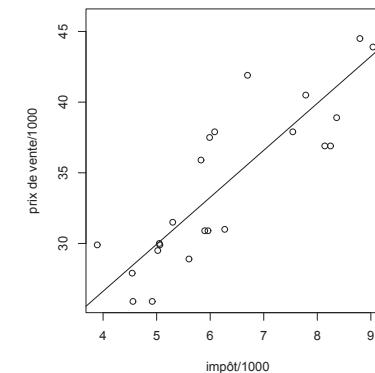


Figure 4: Nuage de points prix de vente versus impôt.

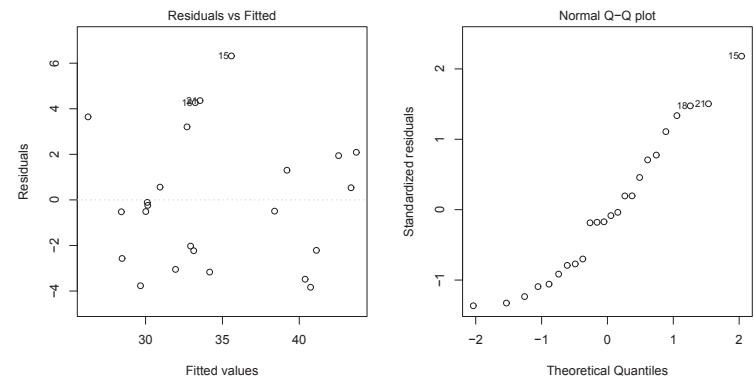


Figure 5: Graphiques de diagnostic du modèle de régression linéaire.

**Solutions :** a) le coefficient de détermination  $R^2$  représente en quelque sorte une mesure de la qualité de l'ajustement du modèle linéaire aux données. Il mesure en fait de quelle manière la régression explique la variabilité totale contenue dans la variable réponse b) par les deux graphiques, on teste l'hypothèse de normalité des erreurs et l'homoscédasticité de ces dernières. Plus simplement, on teste l'hypothèse  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ,  $\sigma^2$  constant. Le graphique de droite (quan-

tiles versus quantiles) montre que l'hypothèse de normalité des erreurs est raisonnable. Les points se trouvent au voisinage d'une droite à l'exception des points n° 15, 18 et 21. Comme presque tous les points du graphique de gauche (à l'exception des points n° 15, 18 et 21) se trouvent dans une bande centrée en 0 (espérance nulle) et comprise entre -4 et +4, l'hypothèse d'homoscédasticité est raisonnable c) on teste l'hypothèse  $H_0 : \beta_1 = 0$  où  $\beta_1$  est la pente (inconnue) de la droite de régression linéaire. Comme la  $p$ -valeur est très inférieure au niveau de signification (0.01), on peut rejeter l'hypothèse  $H_0$  à un niveau de signification de 1% d) la valeur de vente espérée pour un impôt de 7.5 vaut  $13.32 + 3.32 \cdot 7.5 = 38.22$ .

#### Exercice 6

Le nombre de miles parcourus par gallon (unité américaine qui équivaut à 3.8 litres) et le poids en pounds de soixante automobiles ont été relevés. On se demande s'il existe une relation linéaire entre ces deux variables. Le poids (Weight) constitue la variable explicative indépendante et le nombre de miles parcourus par gallon (Mileage) est la variable réponse, variable dépendante.

- a) Obtenus à l'aide du logiciel de statistique **R**, les résultats partiels d'une régression linéaire simple ajustée aux observations sont

```
Call:
lm(formula = Mileage ~ Weight, data = car.test.frame)
...
Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) 48.349347  1.979414   24.4 < 2e-16 ***
Weight      -0.008193  0.000673   -12.2 < 2e-16 ***
...
Residual standard error: 2.56 on 58 degrees of freedom
Multiple R-Squared:  0.719, Adjusted R-squared:  0.714
F-statistic: 148 on 1 and 58 DF, p-value: < 2e-16
...
```

À l'aide du coefficient de détermination  $R^2$  et du graphique de nuage de points de la Figure 6, qualifier la qualité de l'ajustement du modèle linéaire aux données observées.

- b) En utilisant les résultats obtenus par **R** en a), tester la nullité de la pente de la droite de régression à un niveau de signification de 1%.
- c) À l'aide du modèle de régression, déterminer le nombre espéré de miles par gallon pour une automobile de poids égal à 2750 pounds.
- d) Un arbre de régression optimal avec le nombre de miles par gallon comme variable réponse et le poids comme variable explicative se trouve dans la Figure 7. À l'aide de cet arbre de régression, déterminer le nombre espéré de miles par gallon pour une automobile de poids égal à 2750 pounds.

**Solutions :** a) en se basant sur le graphique du nuage de points et sur le coefficient de détermination (0.719) ainsi que sur le coefficient de détermination ajusté (0.714), l'ajustement du modèle aux observations est relativement bon b) on teste l'hypothèse  $H_0 : \beta_1 = 0$  où  $\beta_1$  est la pente (inconnue) de la droite de régression linéaire contre l'hypothèse alternative  $H_1 : \beta_1 \neq 0$ . Comme la  $p$ -valeur du test est bien inférieure au niveau de signification (0.01), on rejette la nullité de la pente de la droite de régression à un niveau de signification de 1% c) le nombre espéré de miles par gallon pour une automobile de 2750 pounds est  $48.349347 - 0.008193 \cdot 2750 \approx 25.819$  d) selon l'arbre de régression, le nombre espéré de miles par gallon pour une auto de 2750 pounds est 24.

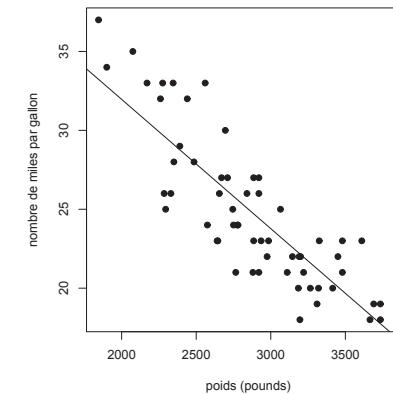


Figure 6: Nuage de points du nombre de miles parcourus par gallon versus poids.

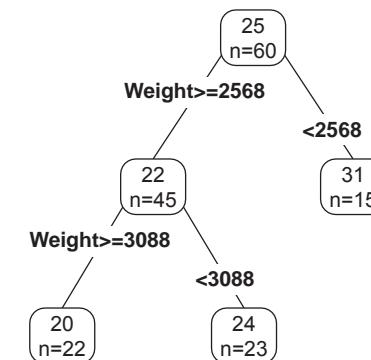


Figure 7: Arbre de régression optimal.

### Exercice 7

Pour chacune des affirmations ci-dessous, décider si elle est vraie ou fausse.

1. Une fois la distance entre groupes choisie, le clustering hiérarchique construira toujours le même dendrogramme.

Vrai  Faux

2. Dans l'analyse des règles d'associations logiques, les règles à supports relativement bas (par exemple dépassant tout juste 20 %) et à confiances élevées (par exemple au-delà de 80 %) sont intéressantes pour l'analyste.

Vrai  Faux

3. Le paramètre de complexité CP indique l'erreur de prédiction pour un arbre de classification élagué à un certain niveau.

Vrai  Faux

4. Le paramètre de complexité décroît lorsque la taille de l'arbre de classification ou de régression croît.

Vrai  Faux

5. La partition finale du  $k$ -means (méthode des nuées dynamiques) ne dépend pas de la partition initiale des données.

Vrai  Faux

6. Le temps d'apprentissage qu'exige un réseau de neurones dépend du nombre d'observations atypiques.

Vrai  Faux

**Solutions :** 1. Vrai 2. Vrai 3. Faux 4. Vrai 5. Faux 6. Faux.

### Exercice 8

Écrire en pseudo-code l'algorithme du clustering hiérarchique ascendant.



## Formulaire de probabilités

### 1. Probabilités élémentaires

$P(A)$

- $\Omega$  : ensemble fondamental;
- $A$  : événement,  $\Omega$  : événement certain,  $\emptyset$  : événement impossible,  $\overline{A}$  : événement complémentaire;
- axiomes de probabilités :
  1.  $0 \leq P(A) \leq 1$ ;
  2.  $P(\Omega) = 1$ ;
  3. Si  $A$  et  $B$  sont deux événements incompatibles,  $P(A \cup B) = P(A) + P(B)$ ;
- $P(\emptyset) = 0$ ;
- $P(\overline{A}) = 1 - P(A)$ ;
- Si  $A \subset B$ ,  $P(A) \leq P(B)$ ;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ ;
- $P(A) = \frac{\#\text{(cas favorables à } A\text{)}}{\#\text{(cas possibles)}}$ .

### 2. Probabilités conditionnelles et indépendance

$P(A | B)$

- définition :  $P(A | B) = \frac{P(A \cap B)}{P(B)}$ ;
- théorème de multiplication :  $P(A \cap B) = P(B | A) \cdot P(A) = P(A | B) \cdot P(B)$ ;
- théorème des probabilités totales :  $P(A) = P(A | H_1) \cdot P(H_1) + \dots + P(A | H_k) \cdot P(H_k)$  où  $H_1, \dots, H_k$  forment une partition de  $\Omega$ ;
- formule de Bayes (version simple) :  $P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$ ;
- formule de Bayes (version composée) :  $P(H_j | B) = \frac{P(B | H_j) \cdot P(H_j)}{P(B | H_1) \cdot P(H_1) + \dots + P(B | H_k) \cdot P(H_k)}$  si  $H_1, \dots, H_k$  forment une partition de  $\Omega$ ;
- indépendance de  $A$  et  $B$  :  $P(A | B) = P(A)$  et  $P(B | A) = P(B)$ ;  $P(A \cap B) = P(A) \cdot P(B)$ .

### 3. Variables aléatoires

$X$

- définition :

$$\begin{aligned} X &: \Omega \rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega), \end{aligned}$$

- $\Omega$  : ensemble fondamental d'une expérience, ensemble des issues;
- $\omega$  : événement élémentaire de l'expérience;
- $\mathcal{H}$  : ensemble des valeurs pouvant être prises par la variable aléatoire  $X$ ,  $\mathcal{H} \subseteq \mathbb{R}$ ;

- $\mathcal{H}$  peut être :

- ◊ discret  $\rightsquigarrow$  variable aléatoire discrète;
- ◊ continu  $\rightsquigarrow$  variable aléatoire continue;

## Annexes

---

Formulaire de probabilités

Intervalles de confiance

Tableau des tests sur les paramètres de lois normales

Table de la distribution normale centrée réduite

Quantiles de la distribution de Student

Quantiles de la distribution khi carré

Quelques documents complémentaires

- loi de probabilité (pour une variable aléatoire discrète) :

$$p(x) = P(X = x)$$

si  $X$  peut prendre les valeurs  $x_1, x_2, \dots$ ,

–  $p(x_i) \geq 0, \quad i = 1, 2, \dots;$

$$-\sum p(x_i) = 1;$$

- fonction de densité (pour une variable aléatoire continue) :

$$f_X, f$$

–  $f_X(u) \geq 0, \quad \forall u \in \mathbb{R};$

$$-\int_{-\infty}^{\infty} f_X(u) du = 1;$$

$$-P(a < X \leq b) = \int_a^b f_X(u) du;$$

- fonction de répartition :

$$F_X, F$$

–  $F_X(x) = P(X \leq x);$

–  $F'_X(x) = f_X(x)$  si  $X$  est une variable aléatoire continue;

$$-P(a < X \leq b) = F_X(b) - F_X(a);$$

- espérance mathématique :

$$\mathbb{E}(X), \mu(X), \mu_X, \mu$$

– définition :

$$\mathbb{E}(X) = \begin{cases} \sum x_i \cdot p_i & \text{si } X \text{ est discrète;} \\ \int_{-\infty}^{\infty} u \cdot f_X(u) du & \text{si } X \text{ est continue,} \end{cases}$$

où  $\mathcal{H} = \{x_1, x_2, \dots\}$  et  $P(X = x_i) = p_i$  avec  $\sum p_i = 1$ ;

– propriétés :

$$1. \mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}(X) + b;$$

$$2. \mathbb{E}(\varphi(X)) = \begin{cases} \sum \varphi(x_i) \cdot p_i & \text{si } X \text{ est discrète;} \\ \int_{-\infty}^{\infty} \varphi(u) \cdot f_X(u) du & \text{si } X \text{ est continue,} \end{cases}$$

où  $\varphi$  est une fonction réelle;

- variance :

$$\text{Var}(X), \sigma^2(X), \sigma_X^2, \sigma^2$$

– définition :

$$\text{Var}(X) = \begin{cases} \sum (x_i - \mathbb{E}(X))^2 \cdot p_i & \text{si } X \text{ est discrète;} \\ \int_{-\infty}^{\infty} (u - \mathbb{E}(X))^2 \cdot f_X(u) du & \text{si } X \text{ est continue,} \end{cases}$$

où  $\mathcal{H} = \{x_1, x_2, \dots\}$  et  $P(X = x_i) = p_i$  avec  $\sum p_i = 1$ ;

– propriétés :

$$1. \text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X);$$

$$2. \text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X);$$

$$\bullet \text{ écart-type : } \sigma(X) = \sqrt{\text{Var}(X)};$$

$$\sigma(X), \sigma_X, \sigma$$

- si  $X$  est une variable aléatoire d'espérance  $\mathbb{E}(X) = \mu$  et de variance  $\text{Var}(X) = \sigma^2$ , la variable aléatoire définie par

$$Z = \frac{X - \mathbb{E}(X)}{\sqrt{\text{Var}(X)}} = \frac{X - \mu}{\sigma}$$

est une variable aléatoire centrée, i.e  $\mathbb{E}(Z) = 0$ , et réduite, i.e  $\text{Var}(Z) = 1$ .

#### 4. Distributions usuelles

- binomiale

$$-\text{loi de probabilité : } p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n;$$

– paramètres :  $n \in \mathbb{N}^*$ ,  $0 \leq p \leq 1$ ;

– espérance mathématique :  $np$ ;

– variance :  $np(1-p)$ ;

$$\mathcal{B}(n, p)$$

- géométrique

$$-\text{loi de probabilité : } p(x) = p(1-p)^{x-1}, \quad x = 1, 2, \dots$$

– paramètre :  $0 < p \leq 1$ ;

– espérance mathématique :  $\frac{1}{p}$ ;

– variance :  $\frac{1-p}{p^2}$ ;

$$\mathcal{G}(p)$$

- Poisson

$$-\text{loi de probabilité : } p(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots$$

– paramètre :  $\lambda > 0$ ;

– espérance mathématique :  $\lambda$ ;

– variance :  $\lambda$ ;

$$\mathcal{P}(\lambda)$$

- uniforme

$$-\text{fonction de densité : } f_X(u) = \begin{cases} \frac{1}{b-a} & \text{si } u \in [a, b]; \\ 0 & \text{sinon;} \end{cases}$$

$$\mathcal{U}(a, b)$$

- paramètres :  $a, b \in \mathbb{R}$ ,  $a < b$ ;
- fonction de répartition :

$$F_X(x) = \begin{cases} 0 & \text{si } x < a; \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b; \\ 1 & \text{si } x > b; \end{cases}$$

- espérance mathématique :  $\frac{a+b}{2}$ ;
- variance :  $\frac{(b-a)^2}{12}$ ;

- exponentielle

- fonction de densité :  $f_X(u) = \begin{cases} \lambda e^{-\lambda u} & \text{si } u \geq 0; \\ 0 & \text{si } u < 0; \end{cases}$

- paramètre :  $\lambda > 0$ ;
- fonction de répartition :

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0; \\ 1 - e^{-\lambda x} & \text{si } x \geq 0; \end{cases}$$

- espérance mathématique :  $\frac{1}{\lambda}$ ;
- variance :  $\frac{1}{\lambda^2}$ ;

- normale (Laplace – Gauss)

- fonction de densité :  $f_X(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(u-\mu)^2}{2\sigma^2}}$ ,  $u \in \mathbb{R}$ ;
- paramètres :  $\mu \in \mathbb{R}$ ,  $\sigma^2 \in \mathbb{R}^+$ ;
- fonction de répartition :

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R},$$

où  $\Phi$  est la fonction de répartition d'une variable aléatoire normale centrée réduite;

- espérance mathématique :  $\mu$ ;
- variance :  $\sigma^2$ .

## 5. Variables aléatoires simultanées

$$\boxed{\mathcal{E}(\lambda)}$$

- cas discret :
- loi de probabilité simultanée (ou conjointe) de  $X$  et  $Y$  :  $P(X = x_i, Y = y_j)$ ;
- loi de probabilité marginale de  $X$  :  $P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$ ;

loi de probabilité marginale de  $Y$  :  $P(Y = y_j) = \sum_i P(X = x_i, Y = y_j)$ ;

- cas continu :

fonction de densité conjointe (ou simultanée) de  $X$  et  $Y$  :  $f_{X,Y}(u, v)$ ;

$$\boxed{f_{X,Y}}$$

fonction de densité marginale de  $X$  :  $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v) dv$ ;

fonction de densité marginale de  $Y$  :  $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u, y) du$ ;

- fonction de répartition de  $X$  et de  $Y$  :

$$\boxed{F_{X,Y}}$$

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

- covariance de  $X$  et de  $Y$  :

$$\boxed{\text{Cov}(X, Y)}$$

$$\text{Cov}(X, Y) = E\left([X - E(X)] \cdot [Y - E(Y)]\right) = E(X \cdot Y) - E(X) \cdot E(Y);$$

- propriétés de la covariance :

1.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ ;
2.  $\text{Cov}(X, a) = 0$ ;
3.  $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$ ;
4.  $\text{Cov}(aX + bY, cZ + d) = ac \text{Cov}(X, Z) + bc \text{Cov}(Y, Z)$ ;
5.  $\text{Cov}(X, X) = \text{Var}(X) \rightsquigarrow \text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) \pm 2 \text{Cov}(X, Y)$ ;

- corrélation entre  $X$  et  $Y$  :

$$\boxed{\rho(X, Y), \text{Corr}(X, Y)}$$

$$\rho(X, Y) = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}};$$

- propriété de la corrélation :

1.  $-1 \leq \text{Corr}(X, Y) \leq 1$ ;
2.  $\text{Corr}(X, Y) = \text{Corr}(Y, X)$ ;
3.  $\text{Corr}(X, X) = 1$  et  $\text{Corr}(X, -X) = -1$ ;
4.  $\text{Corr}(aX + b, cY + d) = \text{Corr}(X, Y)$  si  $a$  et  $c$  sont non nuls et de même signe;

- indépendance :

- cas discret :  $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$ ,  $\forall (i, j)$ ;

- cas continu :  $f_{X,Y}(u, v) = f_X(u) \cdot f_Y(v)$ ;

- si  $X$  et  $Y$  sont deux variables aléatoires indépendantes alors

$$\triangleright E(X \cdot Y) = E(X) \cdot E(Y);$$

$$\triangleright \text{Cov}(X, Y) = 0;$$

$$\triangleright \text{Corr}(X, Y) = 0;$$

$$\triangleright \text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y);$$

- $\text{Cov}(X, Y) = 0 \not\Rightarrow X$  et  $Y$  variables aléatoires indépendantes;

- $E\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m E(X_i)$  où  $X_1, X_2, \dots, X_m$  sont  $m$  variables aléatoires;

• convolution :

$X, Y$  deux variables aléatoires indépendantes continues de fonctions de densité respectives  $f_X$  et  $f_Y$ . Posons  $Z = X + Y$ .

$$- F_Z(z) = P(X + Y \leq z) = \int_{-\infty}^{+\infty} F_X(z - y) f_Y(y) dy$$

(convolution des fonctions  $F_X$  et  $f_Y$ );

$$- f_Z(z) = \frac{d}{dz} \int_{-\infty}^{+\infty} F_X(z - y) f_Y(y) dy = \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y) dy$$

(convolution des fonctions  $f_X$  et  $f_Y$ ).

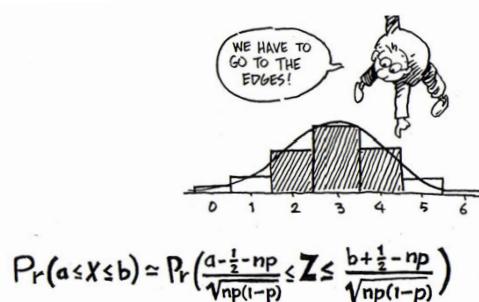
## 6. Le théorème central limite

Soit une suite infinie  $X_1, X_2, \dots, X_n, \dots$  de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), de même espérance mathématique  $E(X_i) = \mu$  et de même variance  $\text{Var}(X_i) = \sigma^2$ .

Lorsque  $n \rightarrow \infty$ , la probabilité d'intervalle

$P(a < Z_n \leq b)$  converge vers  $\Phi(b) - \Phi(a)$ ,

où  $Z_n$  sont les variables aléatoires centrées réduites correspondant aux sommes  $S_n$  ou aux moyennes  $M_n$  des  $n$  premières variables de la suite et  $\Phi$  est la fonction de répartition d'une variable aléatoire normale centrée réduite.



## Intervalles de confiance

- Intervalle de confiance pour l'espérance  $\mu$  d'une variable aléatoire  $X$  issue d'une distribution normale de variance  $\sigma^2$  connue au seuil de confiance  $1 - \alpha$  :

$$\left[ \bar{x} - c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + c_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

- Intervalle de confiance pour l'espérance  $\mu$  d'une variable aléatoire  $X$  issue d'une distribution normale de variance  $\sigma^2$  inconnue au seuil de confiance  $1 - \alpha$  :

$$\left[ \bar{x} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right].$$

- Intervalle de confiance pour la variance  $\sigma^2$  d'une variable aléatoire  $X$  issue d'une distribution normale au seuil de confiance  $1 - \alpha$  :

$$\left[ \frac{(n-1) \cdot s^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1) \cdot s^2}{\chi_{n-1, \alpha/2}^2} \right].$$

- Intervalle de confiance approché pour le paramètre  $p$  dans le cas d'une distribution binomiale au seuil de confiance  $1 - \alpha$  :

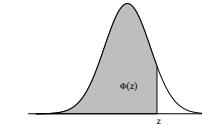
$$\left[ \hat{p} - c_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].$$



Tableau des tests sur les paramètres de lois normales

| $H_0$                     | Statistique de test                                                                                                                                                                                                           | $H_1$                                                                                  | Région de rejet                                                                                                                                             |
|---------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\mu = \mu_0$             | $Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}},$<br>$\sigma$ connu                                                                                                                                                            | $\mu < \mu_0$<br>$\mu > \mu_0$<br>$\mu \neq \mu_0$                                     | $Z \leq c_\alpha$<br>$Z \geq c_{1-\alpha}$<br>$Z \leq c_{\alpha/2}$ et<br>$Z \geq c_{1-\alpha/2}$                                                           |
| $\mu = \mu_0$             | $T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}},$<br>$\sigma$ inconnu                                                                                                                                                               | $\mu < \mu_0$<br>$\mu > \mu_0$<br>$\mu \neq \mu_0$                                     | $T \leq t_{n-1, \alpha}$<br>$T \geq t_{n-1, 1-\alpha}$<br>$T \leq t_{n-1, \alpha/2}$ et<br>$T \geq t_{n-1, 1-\alpha/2}$                                     |
| $\mu_1 - \mu_2 = d_0$     | $Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}},$<br>$\sigma_1$ et $\sigma_2$ connus                                                                                                    | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$             | $Z \leq c_\alpha$<br>$Z \geq c_{1-\alpha}$<br>$Z \leq c_{\alpha/2}$ et<br>$Z \geq c_{1-\alpha/2}$                                                           |
| $\mu_1 - \mu_2 = d_0$     | $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_p \sqrt{(1/n_1) + (1/n_2)}},$<br>$\sigma_1 = \sigma_2$ et inconnus,<br>$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$                                          | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$             | $T \leq t_{n_1+n_2-2, \alpha}$<br>$T \geq t_{n_1+n_2-2, 1-\alpha}$<br>$T \leq t_{n_1+n_2-2, \alpha/2}$ et<br>$T \geq t_{n_1+n_2-2, 1-\alpha/2}$             |
| $\mu_1 - \mu_2 = d_0$     | $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}},$<br>$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}},$<br>$\sigma_1 \neq \sigma_2$ et inconnus | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$             | $T \leq t_{\nu, \alpha}$<br>$T \geq t_{\nu, 1-\alpha}$<br>$T \leq t_{\nu, \alpha/2}$ et<br>$T \geq t_{\nu, 1-\alpha/2}$                                     |
| $\mu_1 - \mu_2 = d_0$     | $T = \frac{\bar{D} - d_0}{S_d / \sqrt{n}},$<br>observations appariées                                                                                                                                                         | $\mu_1 - \mu_2 < d_0$<br>$\mu_1 - \mu_2 > d_0$<br>$\mu_1 - \mu_2 \neq d_0$             | $T \leq t_{n-1, \alpha}$<br>$T \geq t_{n-1, 1-\alpha}$<br>$T \leq t_{n-1, \alpha/2}$ et<br>$T \geq t_{n-1, 1-\alpha/2}$                                     |
| $\sigma^2 = \sigma_0^2$   | $X^2 = \frac{(n-1) S^2}{\sigma_0^2},$                                                                                                                                                                                         | $\sigma^2 < \sigma_0^2$<br>$\sigma^2 > \sigma_0^2$<br>$\sigma^2 \neq \sigma_0^2$       | $X^2 \leq \chi^2_{n-1, \alpha}$<br>$X^2 \geq \chi^2_{n-1, 1-\alpha}$<br>$X^2 \leq \chi^2_{n-1, \alpha/2}$ et<br>$X^2 \geq \chi^2_{n-1, 1-\alpha/2}$         |
| $\sigma_1^2 = \sigma_2^2$ | $F = \frac{S_1^2}{S_2^2},$                                                                                                                                                                                                    | $\sigma_1^2 < \sigma_2^2$<br>$\sigma_1^2 > \sigma_2^2$<br>$\sigma_1^2 \neq \sigma_2^2$ | $F \leq F_{n_1-1, n_2-1, \alpha}$<br>$F \geq F_{n_1-1, n_2-1, 1-\alpha}$<br>$F \leq F_{n_1-1, n_2-1, \alpha/2}$ et<br>$F \geq F_{n_1-1, n_2-1, 1-\alpha/2}$ |

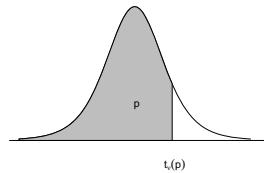
Distribution normale centrée réduite : valeur de  $\Phi(z)$  en fonction de  $z$



Pour  $z < 0$ , utiliser la propriété :  $P(Z \leq z) = \Phi(z) = 1 - \Phi(-z)$ ,  $z \in \mathbb{R}$ .

| $z$ | 0      | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | .50000 | .50399 | .50798 | .51197 | .51595 | .51994 | .52392 | .52790 | .53188 | .53586 |
| 0.1 | .53983 | .54380 | .54776 | .55172 | .55567 | .55962 | .56356 | .56750 | .57142 | .57535 |
| 0.2 | .57926 | .58317 | .58706 | .59095 | .59483 | .59871 | .60257 | .60642 | .61026 | .61409 |
| 0.3 | .61791 | .62172 | .62552 | .62930 | .63307 | .63683 | .64058 | .64431 | .64803 | .65173 |
| 0.4 | .65542 | .65910 | .66276 | .66640 | .67003 | .67364 | .67724 | .68082 | .68439 | .68793 |
| 0.5 | .69146 | .69497 | .69847 | .70194 | .70540 | .70884 | .71226 | .71566 | .71904 | .72240 |
| 0.6 | .72575 | .72907 | .73237 | .73565 | .73891 | .74215 | .74537 | .74857 | .75175 | .75490 |
| 0.7 | .75804 | .76115 | .76424 | .76730 | .77035 | .77337 | .77637 | .77935 | .78230 | .78524 |
| 0.8 | .78814 | .79103 | .79389 | .79673 | .79955 | .80234 | .80511 | .80785 | .81057 | .81327 |
| 0.9 | .81594 | .81859 | .82121 | .82381 | .82639 | .82894 | .83147 | .83398 | .83646 | .83891 |
| 1.0 | .84134 | .84375 | .84614 | .84850 | .85083 | .85314 | .85543 | .85769 | .85993 | .86214 |
| 1.1 | .86433 | .86650 | .86864 | .87076 | .87286 | .87493 | .87698 | .87900 | .88100 | .88298 |
| 1.2 | .88493 | .88686 | .88877 | .89065 | .89251 | .89435 | .89617 | .89796 | .89973 | .90147 |
| 1.3 | .90320 | .90490 | .90658 | .90824 | .90988 | .91149 | .91309 | .91466 | .91621 | .91774 |
| 1.4 | .91924 | .92073 | .92220 | .92364 | .92507 | .92647 | .92786 | .92922 | .93056 | .93189 |
| 1.5 | .93319 | .93448 | .93574 | .93699 | .93822 | .93943 | .94062 | .94179 | .94295 | .94408 |
| 1.6 | .94520 | .94630 | .94738 | .94845 | .94950 | .95053 | .95154 | .95254 | .95352 | .95449 |
| 1.7 | .95543 | .95637 | .95728 | .95818 | .95907 | .95994 | .96080 | .96164 | .96246 | .96327 |
| 1.8 | .96407 | .96485 | .96562 | .96638 | .96712 | .96784 | .96856 | .96926 | .96995 | .97062 |
| 1.9 | .97128 | .97193 | .97257 | .97320 | .97381 | .97441 | .97500 | .97558 | .97615 | .97670 |
| 2.0 | .97725 | .97778 | .97831 | .97882 | .97932 | .97982 | .98030 | .98077 | .98124 | .98169 |
| 2.1 | .98214 | .98257 | .98300 | .98341 | .98382 | .98422 | .98461 | .98500 | .98537 | .98574 |
| 2.2 | .98610 | .98645 | .98679 | .98713 | .98745 | .98778 | .98809 | .98840 | .98870 | .98899 |
| 2.3 | .98928 | .98956 | .98983 | .99010 | .99036 | .99061 | .99086 | .99111 | .99134 | .99158 |
| 2.4 | .99180 | .99202 | .99224 | .99245 | .99266 | .99286 | .99305 | .99324 | .99343 | .99361 |
| 2.5 | .99379 | .99396 | .99413 | .99430 | .99446 | .99461 | .99477 | .99492 | .99506 | .99520 |
| 2.6 | .99534 | .99547 | .99560 | .99573 | .99585 | .99598 | .99609 | .99621 | .99632 | .99643 |
| 2.7 | .99653 | .99664 | .99674 | .99683 | .99693 | .99702 | .99711 | .99720 | .99728 | .99736 |
| 2.8 | .99744 | .99752 | .99760 | .99767 | .99774 | .99781 | .99788 | .99795 | .99801 | .99807 |
| 2.9 | .99813 | .99819 | .99825 | .99831 | .99836 | .99841 | .99846 | .99851 | .99856 | .99861 |
| 3.0 | .99865 | .99869 | .99874 | .99878 | .99882 | .99886 | .99889 | .99893 | .99896 | .99900 |
| 3.1 | .99903 | .99906 | .99910 | .99913 | .99916 | .99918 | .99921 | .99924 | .99926 | .99929 |
| 3.2 | .99931 | .99934 | .99936 | .99938 | .99940 | .99942 | .99944 | .99946 | .99948 | .99950 |
| 3.3 | .99952 | .99953 | .99955 | .99957 | .99958 | .99960 | .99961 | .99962 | .99964 | .99965 |
| 3.4 | .99966 | .99968 | .99969 | .99970 | .99972 | .99973 | .99974 | .99975 | .99976 | .99976 |
| 3.5 | .99977 | .99978 | .99978 | .99979 | .99980 | .99981 | .99982 | .99983 | .99983 | .99983 |
| 3.6 | .99984 | .99985 | .99985 | .99986 | .99986 | .99987 | .99987 | .99988 | .99988 | .99989 |
| 3.7 | .99989 | .99990 | .99990 | .99990 | .99991 | .99991 | .99992 | .99992 | .99992 | .99992 |
| 3.8 | .99993 | .99993 | .99993 | .99994 | .99994 | .99994 | .99994 | .99995 | .99995 | .99995 |
| 3.9 | .99995 | .99995 | .99996 | .99996 | .99996 | .99996 | .99996 | .99996 | .99996 | .99997 |

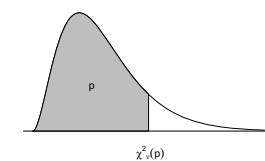
## Distribution $t_\nu$ de Student



Quantiles  $t_\nu(p)$  pour la distribution  $t$  de Student à  $\nu$  degrés de liberté.

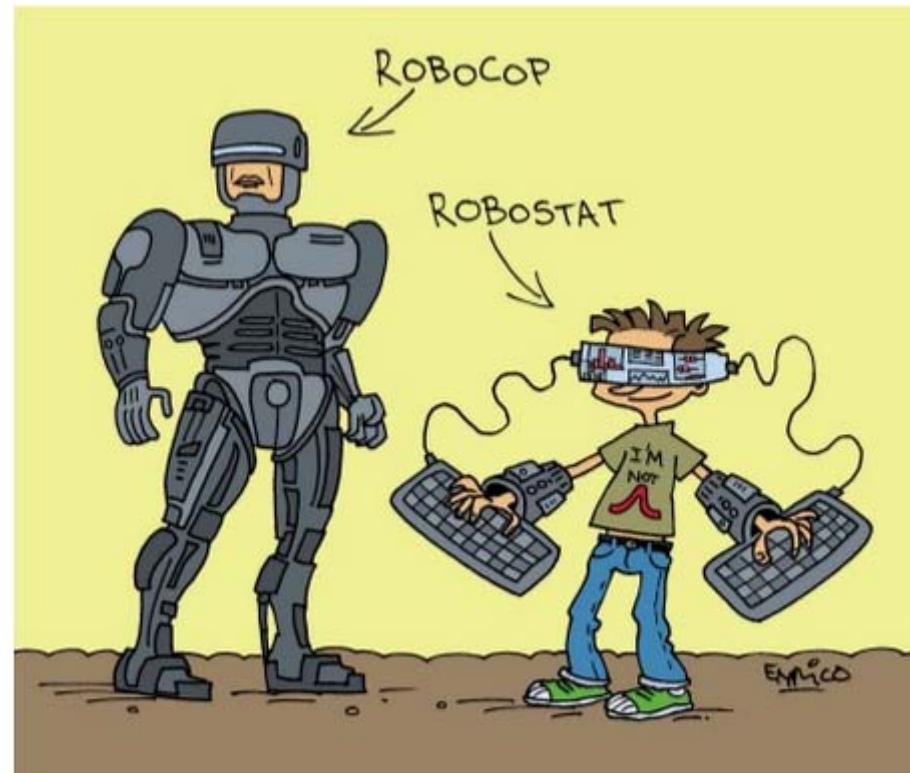
| $\nu$    | .75   | 9     | 95    | .975  | .99   | .995  |
|----------|-------|-------|-------|-------|-------|-------|
| 1        | 1.000 | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 |
| 2        | .8165 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3        | .7649 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4        | .7407 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5        | .7267 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6        | .7176 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7        | .7111 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8        | .7064 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9        | .7027 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10       | .6998 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |
| 11       | .6974 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 |
| 12       | .6955 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 |
| 13       | .6938 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 |
| 14       | .6924 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 |
| 15       | .6912 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16       | .6901 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17       | .6892 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |
| 18       | .6884 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 |
| 19       | .6876 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 |
| 20       | .6870 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 |
| 21       | .6864 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 |
| 22       | .6858 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 |
| 23       | .6853 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 |
| 24       | .6848 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 |
| 25       | .6844 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 |
| 26       | .6840 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 |
| 27       | .6837 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 |
| 28       | .6834 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 |
| 29       | .6830 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 |
| 30       | .6828 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 |
| 40       | .6807 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 |
| 50       | .6794 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 |
| 100      | .6770 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 |
| $\infty$ | .6745 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 |

## Distribution $\chi^2_\nu$

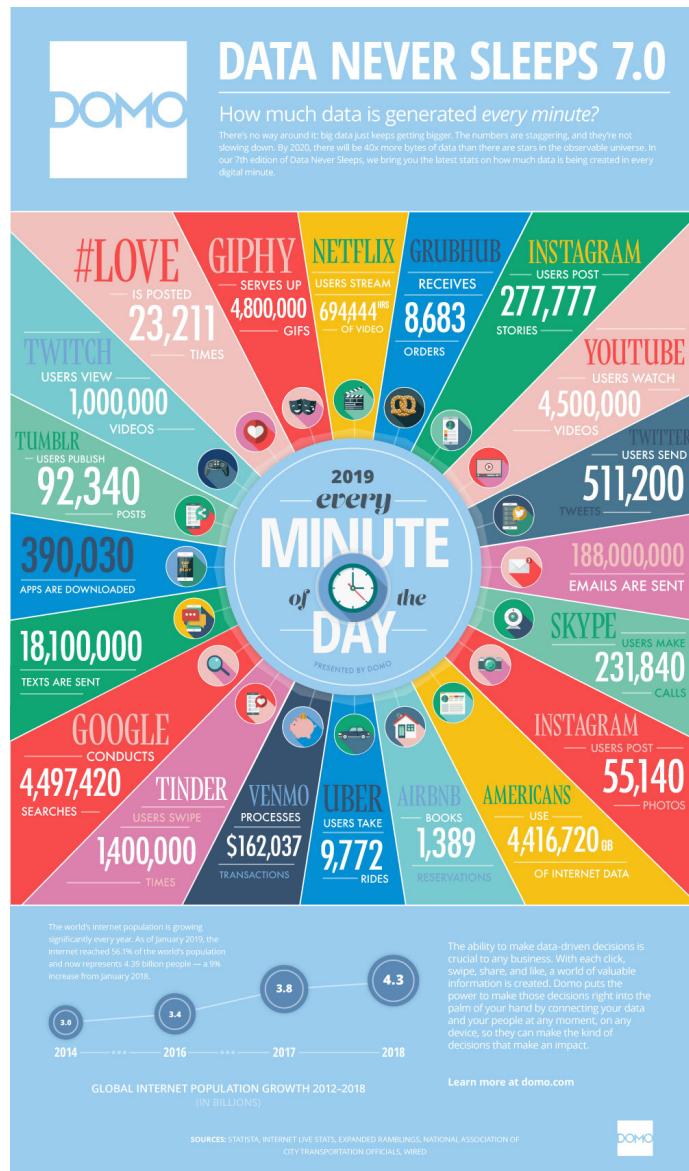


Quantiles  $\chi^2_\nu(p)$  de la distribution khi carré à  $\nu$  degrés de liberté.

| $\nu$ | .005  | .01   | .025  | .05   | .10   | .25  | .50  | .75  | .90  | .95  | .975 | .99  | .995 | .999 |
|-------|-------|-------|-------|-------|-------|------|------|------|------|------|------|------|------|------|
| 1     | 0     | 0.002 | 0.10  | 0.039 | 0.158 | 1.02 | 4.55 | 1.32 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 | 10.8 |
| 2     | .0100 | .0201 | .0506 | .103  | .211  | .575 | 1.39 | 2.77 | 4.61 | 5.99 | 7.38 | 9.21 | 10.6 | 13.8 |
| 3     | .0717 | .115  | .216  | .352  | .584  | 1.21 | 2.37 | 4.11 | 6.25 | 7.81 | 9.35 | 11.3 | 12.8 | 16.3 |
| 4     | .207  | .297  | .484  | .711  | 1.06  | 1.92 | 3.36 | 5.39 | 7.78 | 9.49 | 11.1 | 13.3 | 14.9 | 18.5 |
| 5     | .412  | .554  | .831  | 1.15  | 1.61  | 2.67 | 4.35 | 6.63 | 9.24 | 11.1 | 12.8 | 15.1 | 16.7 | 20.5 |
| 6     | .676  | .872  | 1.24  | 1.64  | 2.20  | 3.45 | 5.35 | 7.84 | 10.6 | 12.6 | 14.4 | 16.8 | 18.5 | 22.5 |
| 7     | .989  | 1.24  | 1.69  | 2.17  | 2.83  | 4.25 | 6.35 | 9.04 | 12.0 | 14.1 | 16.0 | 18.5 | 20.3 | 24.3 |
| 8     | 1.34  | 1.65  | 2.18  | 2.73  | 3.49  | 5.07 | 7.34 | 10.2 | 13.4 | 15.5 | 17.5 | 20.1 | 22.0 | 26.1 |
| 9     | 1.73  | 2.09  | 2.70  | 3.33  | 4.17  | 5.90 | 8.34 | 11.4 | 14.7 | 16.9 | 19.0 | 21.7 | 23.6 | 27.9 |
| 10    | 2.16  | 2.56  | 3.25  | 3.94  | 4.87  | 6.74 | 9.34 | 12.5 | 16.0 | 18.3 | 20.5 | 23.2 | 25.2 | 29.6 |
| 11    | 2.60  | 3.05  | 3.82  | 4.57  | 5.58  | 10.3 | 13.7 | 17.3 | 19.7 | 21.9 | 24.7 | 26.8 | 31.3 |      |
| 12    | 3.07  | 3.57  | 4.40  | 5.23  | 6.30  | 8.44 | 11.3 | 14.8 | 18.5 | 21.0 | 23.3 | 26.2 | 28.3 | 32.9 |
| 13    | 3.57  | 4.11  | 5.01  | 5.89  | 7.04  | 9.30 | 12.3 | 16.0 | 19.8 | 22.4 | 24.7 | 27.7 | 29.8 | 34.5 |
| 14    | 4.07  | 4.66  | 5.63  | 6.57  | 7.79  | 10.2 | 13.3 | 17.1 | 21.1 | 23.7 | 26.1 | 29.1 | 31.3 | 36.1 |
| 15    | 4.60  | 5.23  | 6.26  | 7.26  | 8.55  | 11.0 | 14.3 | 18.2 | 22.3 | 25.0 | 27.5 | 30.6 | 32.8 | 37.7 |
| 16    | 5.14  | 5.81  | 6.91  | 7.96  | 9.31  | 11.9 | 15.3 | 19.4 | 23.5 | 26.3 | 28.8 | 32.0 | 34.3 | 39.3 |
| 17    | 5.70  | 6.41  | 7.56  | 8.67  | 10.1  | 12.8 | 16.3 | 20.5 | 24.8 | 27.6 | 30.2 | 33.4 | 35.7 | 40.8 |
| 18    | 6.26  | 7.01  | 8.23  | 9.39  | 10.9  | 13.7 | 17.3 | 21.6 | 26.0 | 28.9 | 31.5 | 34.8 | 37.2 | 42.3 |
| 19    | 6.84  | 7.63  | 8.91  | 10.1  | 11.7  | 14.6 | 18.3 | 22.7 | 27.2 | 30.1 | 32.9 | 36.2 | 38.6 | 43.8 |
| 20    | 7.43  | 8.26  | 9.59  | 10.9  | 12.4  | 15.5 | 19.3 | 23.8 | 28.4 | 31.4 | 34.2 | 37.6 | 40.0 | 45.3 |
| 21    | 8.03  | 8.90  | 10.3  | 11.6  | 13.2  | 16.3 | 20.3 | 24.9 | 29.6 | 32.7 | 35.5 | 38.9 | 41.4 | 46.8 |
| 22    | 8.64  | 9.54  | 11.0  | 12.3  | 14.0  | 17.2 | 21.3 | 26.0 | 30.8 | 33.9 | 36.8 | 40.3 | 42.8 | 48.3 |
| 23    | 9.26  | 10.2  | 11.7  | 13.1  | 14.8  | 18.1 | 22.3 | 27.1 | 32.0 | 35.2 | 38.1 | 41.6 | 44.2 | 49.7 |
| 24    | 9.89  | 10.9  | 12.4  | 13.8  | 15.7  | 19.0 | 23.3 | 28.2 | 33.2 | 36.4 | 39.4 | 43.0 | 45.6 | 51.2 |
| 25    | 10.5  | 11.5  | 13.1  | 14.6  | 16.5  | 19.9 | 24.3 | 29.3 | 34.4 | 37.7 | 40.6 | 44.3 | 46.9 | 52.6 |
| 26    | 11.2  | 12.2  | 13.8  | 15.4  | 17.3  | 20.8 | 25.3 | 30.4 | 35.6 | 38.9 | 41.9 | 45.6 | 48.3 | 54.1 |
| 27    | 11.8  | 12.9  | 14.6  | 16.2  | 18.1  | 21.7 | 26.3 | 31.5 | 36.7 | 40.1 | 43.2 | 47.0 | 49.6 | 55.5 |
| 28    | 12.5  | 13.6  | 15.3  | 16.9  | 18.9  | 22.7 | 27.3 | 32.6 | 37.9 | 41.3 | 44.5 | 48.3 | 51.0 | 56.9 |
| 29    | 13.1  | 14.3  | 16.0  | 17.7  | 19.8  | 23.6 | 28.3 | 33.7 | 39.1 | 42.6 | 45.7 | 49.6 | 52.3 | 58.3 |
| 30    | 13.8  | 15.0  | 16.8  | 18.5  | 20.6  | 24.5 | 29.3 | 34.8 | 40.3 | 43.8 | 47.0 | 50.9 | 53.7 | 59.7 |
| 40    | 20.7  | 22.2  | 24.4  | 26.5  | 29.1  | 33.7 | 39.3 | 45.6 | 51.8 | 55.8 | 59.3 | 63.7 | 66.8 | 73.4 |
| 50    | 28.0  | 29.7  | 32.4  | 34.8  | 37.7  | 42.9 | 49.3 | 56.3 | 63.2 | 67.5 | 71.4 | 76.2 | 79.5 | 86.7 |
| 60    | 35.5  | 37.5  | 40.5  | 43.2  | 46.5  | 52.3 | 59.3 | 67.0 | 74.4 | 79.1 | 83.3 | 88.4 | 92.0 | 99.6 |
| 70    | 43.3  | 45.4  | 48.8  | 51.7  | 55.3  | 61.7 | 69.3 | 77.6 | 85.5 | 90.5 | 95.0 | 100  | 104. | 112. |







Volume de données générées sur internet par minute (juillet 2019)

James, J. (2019) *What 'Data Never Sleeps 7.0' Says—and Doesn't Say*. Blog Post, July 2019.  
[www.domo.com/blog/what-data-never-sleeps-7-0-says-and-doesnt-say](http://www.domo.com/blog/what-data-never-sleeps-7-0-says-and-doesnt-say)



## Digital Vision for Supercomputing & Big Data



**93%**  
of UK adults own a smartphone<sup>5</sup>



**75%**  
of mobile data traffic will be video content in 2020<sup>6</sup>



**50bn**  
connected objects around the world by 2020<sup>7</sup>



### Global live internet stats<sup>8</sup>

44,005GB of internet traffic per second  
59,945 Google searches per second  
68,990 YouTube videos viewed per second  
2,583,435 emails sent per second (67% of which is SPAM)



**3.6 billion**

internet users worldwide<sup>9</sup>



**Stored data**  
is growing at 4x the speed of the world's economy<sup>10</sup>



**1.7 megabytes**  
of new information will be created every second for every human being on the planet by the year 2020<sup>11</sup>



**100.2 zettabytes**  
projected global Big Data traffic by 2020<sup>12</sup>



**1 trillion-fold**  
The increase in computing performance 1956-2015<sup>13</sup>



**20 million PCs**

The equivalent performance of the Bull Sequana supercomputer by 2020<sup>14</sup>



**97%**

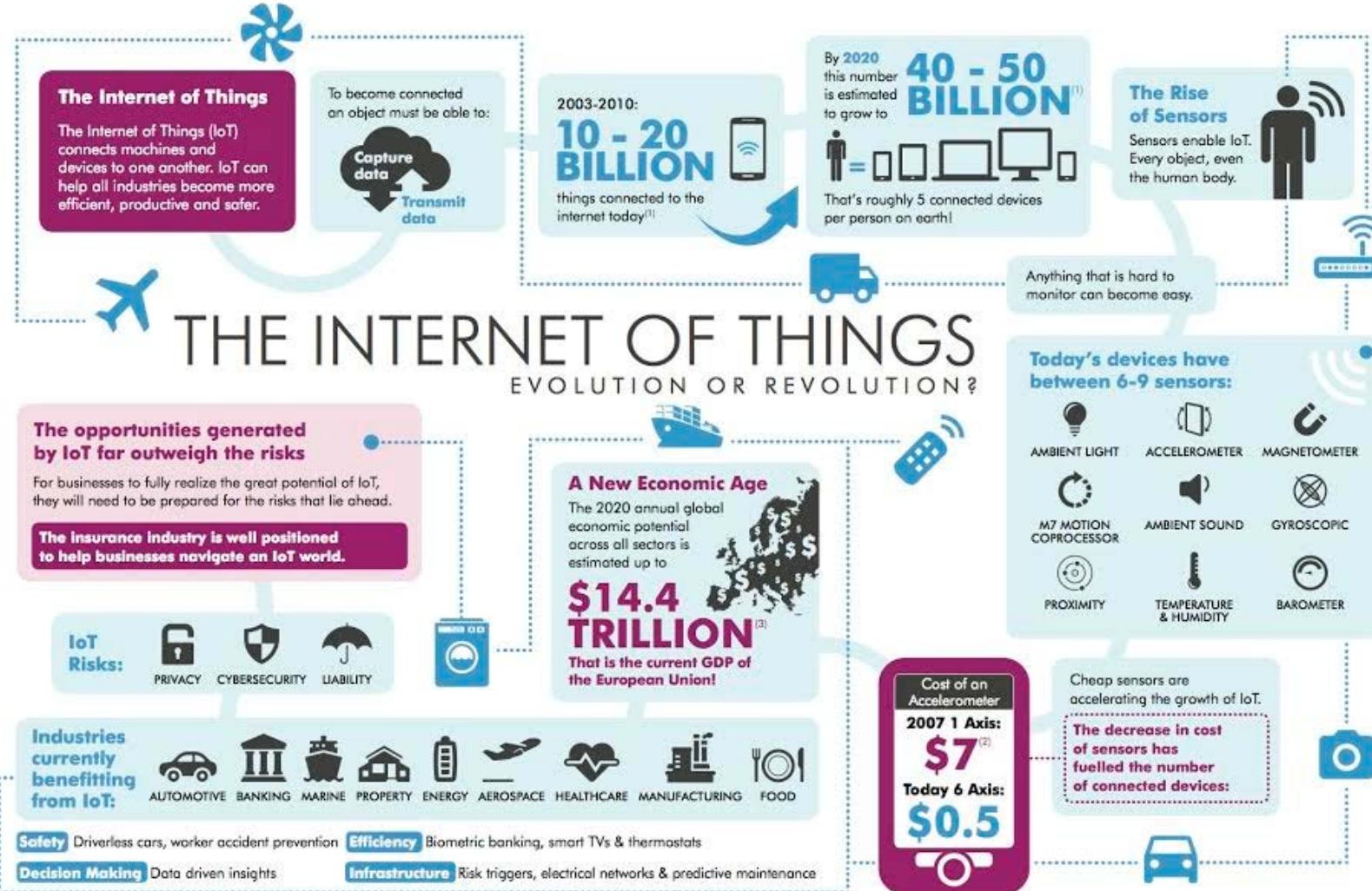
of global companies that adopted HPC could no longer compete or survive without it<sup>15</sup>

**Atos**

Trusted Partner for your Digital Journey

[atos.net/en-gb/united-kingdom/digital-vision-programme/digital-vision-supercomputing-big-data/digital-vision-supercomputing-big-data-infographic-uk](http://atos.net/en-gb/united-kingdom/digital-vision-programme/digital-vision-supercomputing-big-data/digital-vision-supercomputing-big-data-infographic-uk)





Visit [www.aig.com/iot](http://www.aig.com/iot)

Source: (1) Dubrovic, Shawn, "Digital Destiny." (2) CISCO: The Internet of Things How the Next Evolution of the Internet Is Changing Everything, 2011. (3) RAND: Europe's policy options for a dynamic and trustworthy development of the Internet of Things. American International Group, Inc. (AIG) is a leading global insurance organization serving customers in more than 100 countries and jurisdictions. AIG companies serve commercial, institutional, and individual customers through one of the most extensive worldwide property-casualty networks of any insurer. In addition, AIG companies are leading providers of life insurance and reinsurance services in the United States. AIG common stock is listed on the New York Stock Exchange and the Tokyo Stock Exchange. Additional information about AIG can be found at [www.aig.com](http://www.aig.com) | YouTube: [www.youtube.com/aig](http://www.youtube.com/aig) | Twitter: @AIGInsurance | LinkedIn: [www.linkedin.com/company/aig](http://www.linkedin.com/company/aig) AIG is the marketing name for the worldwide property-casualty, life and reinsurance, and general insurance operations of American International Group, Inc. For additional information, please visit our website at [www.aig.com](http://www.aig.com). All products and services are written or provided by subsidiaries or affiliates of American International Group, Inc. Products or services may not be available in all countries, and coverage is subject to actual policy language. Non-insurance products and services may be provided by independent third parties. Certain property-casualty coverages may be provided by a surplus lines insurer. Surplus lines insurers do not generally participate in state guaranty funds, and insureds are therefore not protected by such funds. © American International Group, Inc. All rights reserved.



**40 ZETTABYTES**

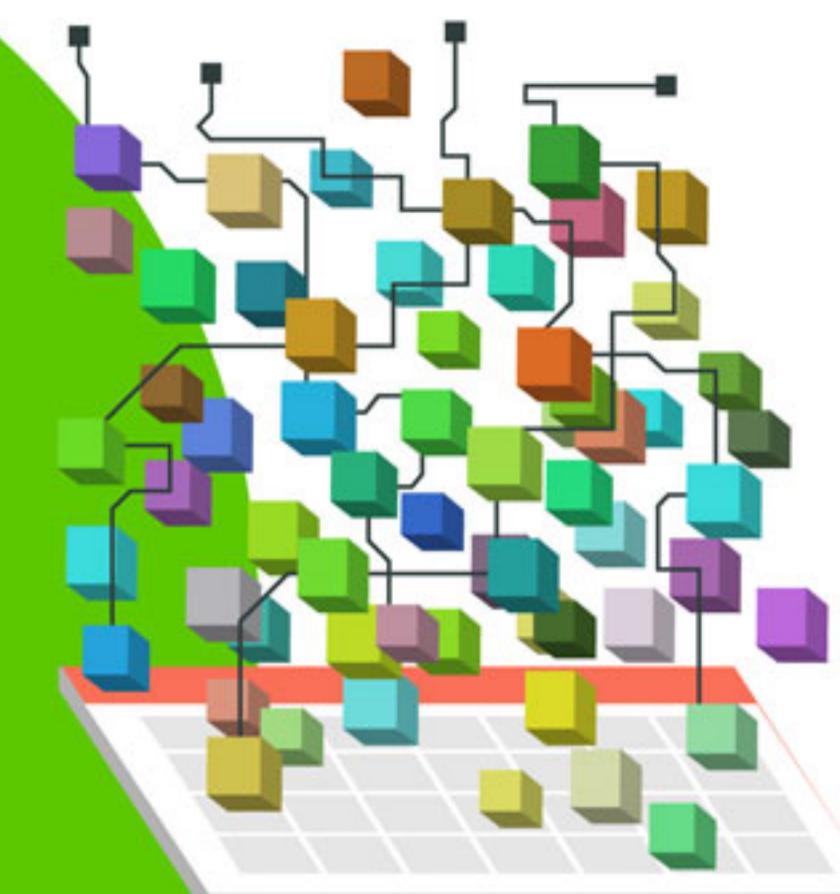
[ 43 TRILLION GIGABYTES ]

of data will be created by 2020, an increase of 300 times from 2005

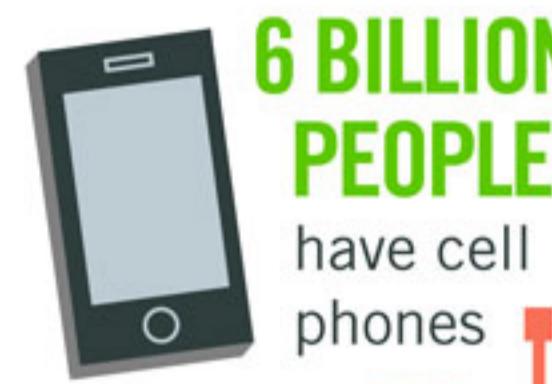


It's estimated that  
**2.5 QUINTILLION BYTES**

[ 2.3 TRILLION GIGABYTES ]  
of data are created each day

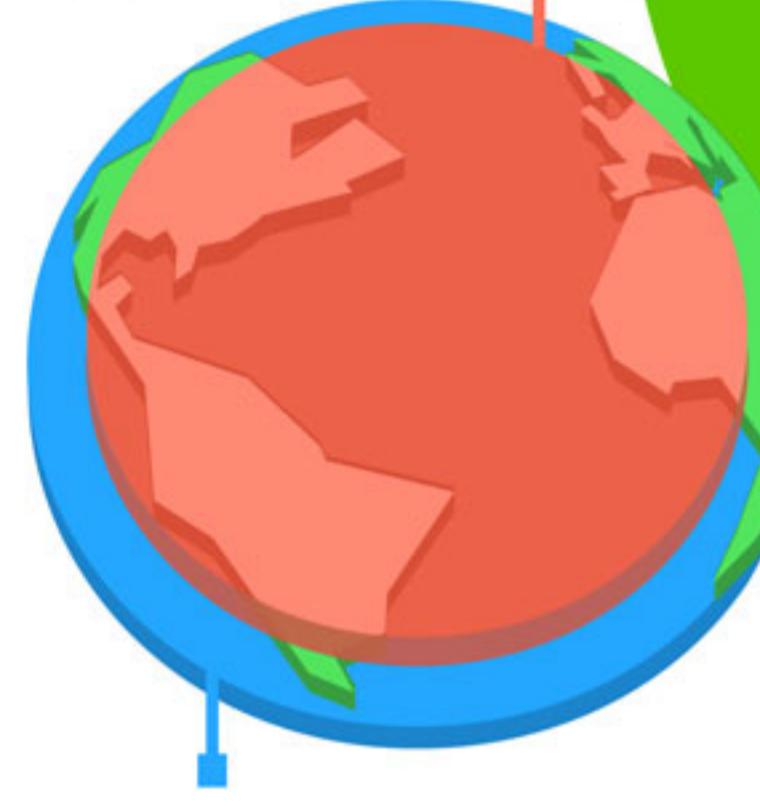


## Volume SCALE OF DATA



**6 BILLION PEOPLE**

have cell phones

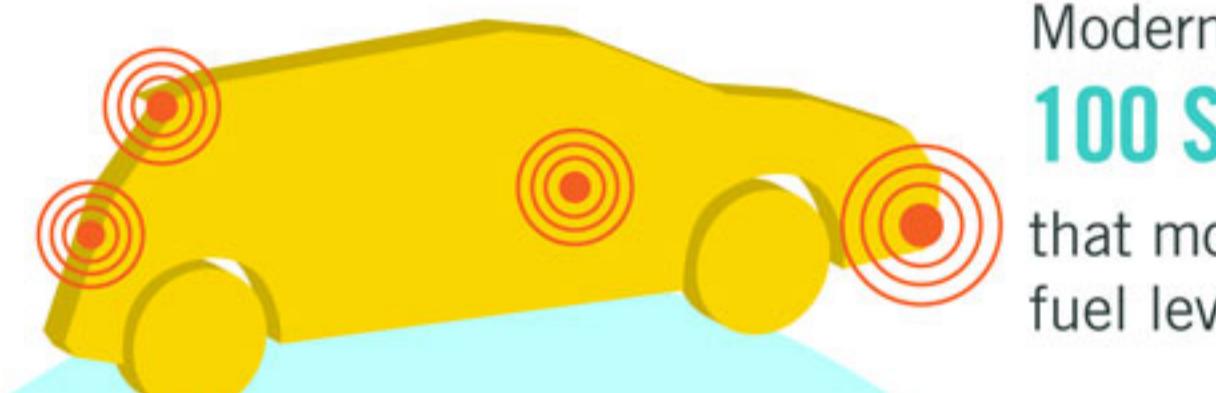


WORLD POPULATION: 7 BILLION

The New York Stock Exchange captures  
data at a rate of

**1 TB OF TRADE INFORMATION**

during each trading session



Modern cars have close to  
**100 SENSORS**

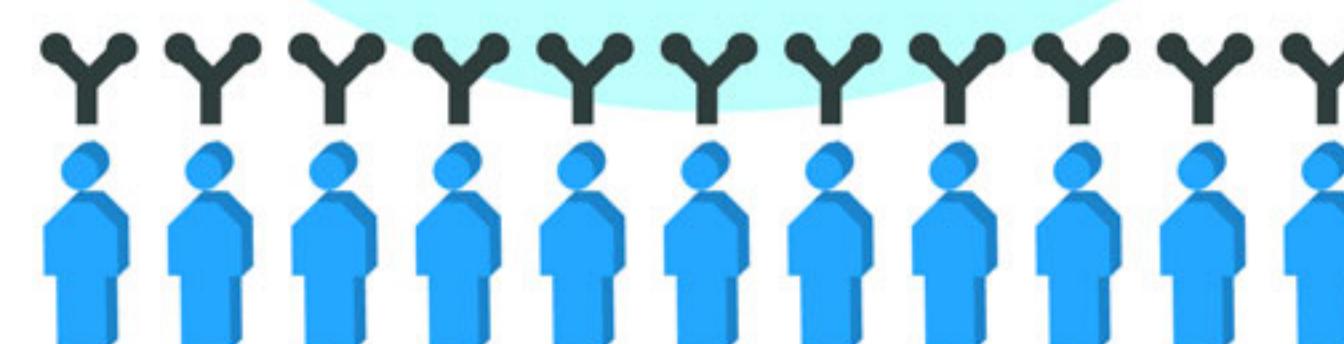
that monitor items such as  
fuel level and tire pressure

## Velocity ANALYSIS OF STREAMING DATA

By 2016, it is projected  
there will be

**18.9 BILLION  
NETWORK CONNECTIONS**

– almost 2.5 connections  
per person on earth



# The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume**, **Velocity**, **Variety** and **Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015  
**4.4 MILLION IT JOBS**  
will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

**150 EXABYTES**

[ 161 BILLION GIGABYTES ]



**30 BILLION PIECES OF CONTENT**

are shared on Facebook every month

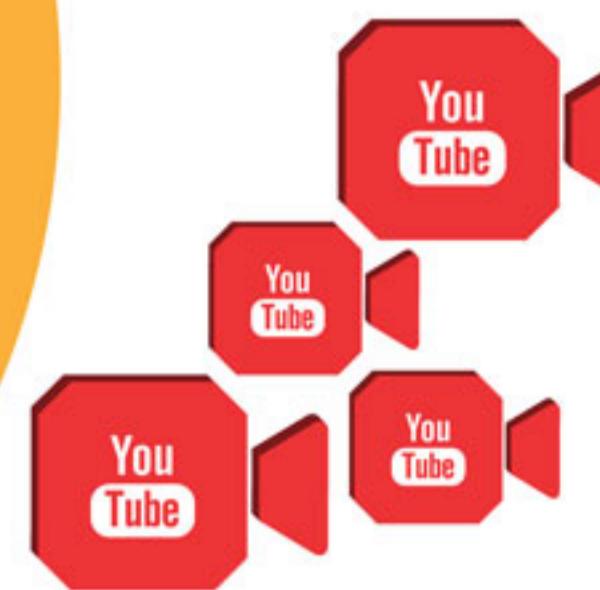


## Variety DIFFERENT FORMS OF DATA



By 2014, it's anticipated there will be  
**420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO**  
are watched on YouTube each month



**400 MILLION TWEETS**  
are sent per day by about 200 million monthly active users

**1 IN 3 BUSINESS LEADERS**

don't trust the information they use to make decisions



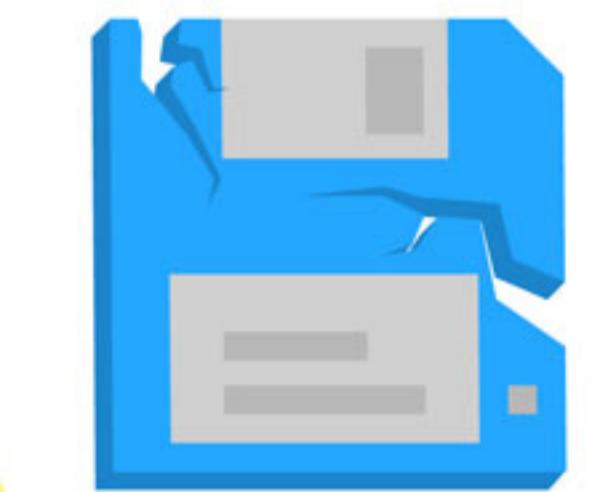
**27% OF RESPONDENTS**

in one survey were unsure of how much of their data was inaccurate

## Veracity UNCERTAINTY OF DATA

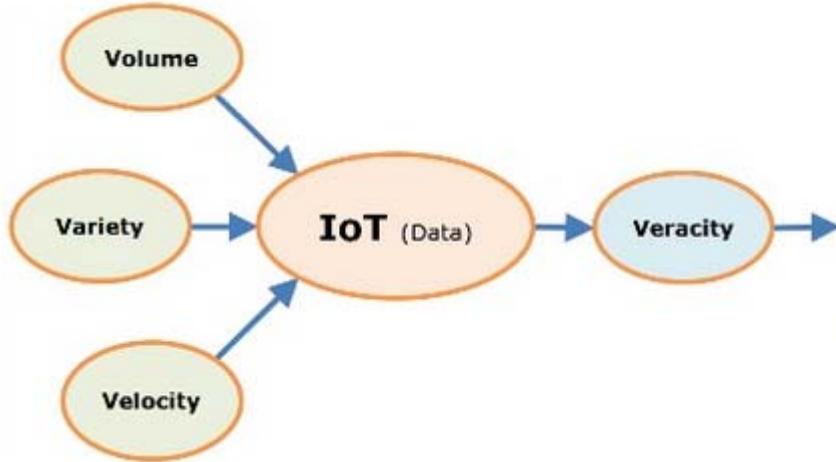
Poor data quality costs the US economy around

**\$3.1 TRILLION A YEAR**



IBM

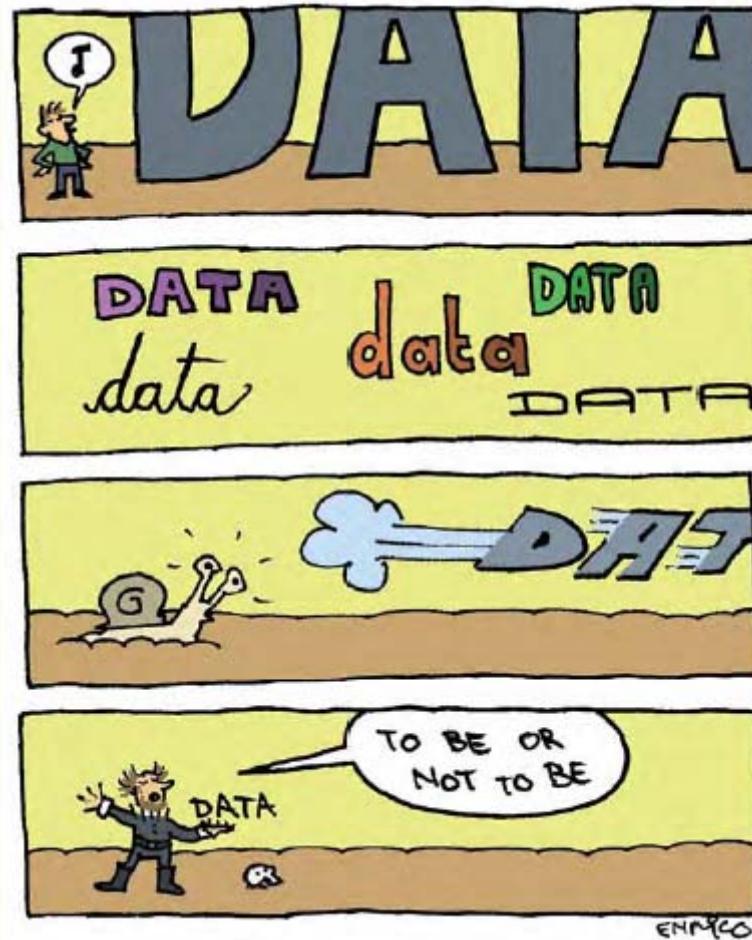




### Les cinq V des mégadonnées

Kuonen, D. (2017). *Glocalised Smart Statistics and Analytics of Things. Core Challenges and Key Issues for Smart (Official) Statistics at the Edge*.  
STS021: From Big Data to Smart Statistics, ISI2017, Marrakech, MA







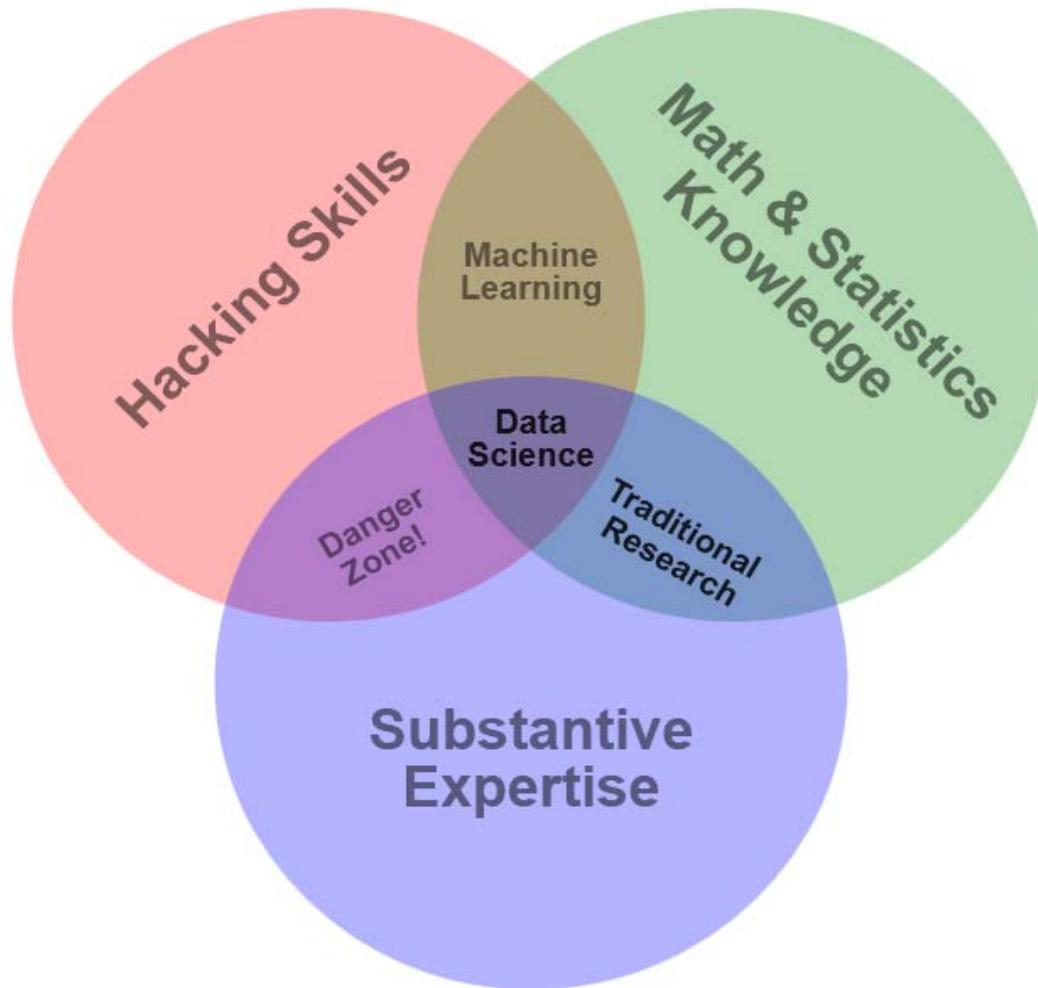


Diagramme de Venn des mégadonnées : à la croisée de différentes disciplines

Conway, D. (2010). [drewconway.com/zia/2013/3/26/the-data-science-venn-diagram](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)



# MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

## MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

## DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

## PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS



## COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



CONNAISSANCE DES ARTS PLUS

ECONOMIE ET POLITIQUE

BILANS GRATUITS

ÉDIAIS

ACTUALITÉS

**LesEchos.fr**



CONNEXION | INSCRIPTION | f | t  
Recherchez sur Les Echos

Recevez nos newsletters | Le journal du jour

## SECTEURS

# Le commerce en ligne français s'arrache les « data miners »

**POUR ACCÉDER AU MEILLEUR DE L'INFORMATION**

**ABONNEZ-VOUS MAINTENANT**

15/07 | 10:22 | mis à jour le 16/07 à 07:38 | 4 commentaires

Tweeter 500      J'aime 366      Share 133      40

**Les spécialistes du maniement des statistiques, souvent issus des écoles d'ingénieurs, sont recrutés à prix d'or, jusqu'à 500.000 euros annuels.**



Le chômage, les « data miners » ou « fouilleurs de données » ne connaissent pas les e-commerçants s'arrachent ces as des maths, qui lont parler vos données pour deviner votre marque de baskets préférée. « Je suis un pur mathéux », explique David Bessis, un chercheur de 40 ans, normandien.

Spécialités: « l'algèbre, la géométrie et la topologie ».

Après une dizaine d'années de recherche fondamentale à l'université de Yale et au CNRS, M. Bessis s'est lancé en affaires. Il a colonisé TinyClues, une start-up qui « cherche de petits indices dans de grandes masses de données », pour de grands noms du commerce en ligne. Parmi des milliers de paramètres, l'ordinateur trouve des corrélations entre l'hébergeur de votre e-mail (@yahoo.fr, @gmail.com) ou vos pseudos, et vos achats. Votre prénom en dit beaucoup sur votre âge et le milieu dont vous êtes issu, et la moindre seconde pendant laquelle votre souris s'attarde sur une page est décortiquée pour choisir, par exemple, quelle publicité vous envoyer. « C'est très subtil culturellement (...) avec des bases de données regroupant plus de 10% de la population française, on dépasse les préjugés », explique M. Bessis.

Avec l'essor du commerce en ligne, ce métier devient stratégique. Les sites « qui déplient les meilleures pratiques en matière de connaissance du client affichent des taux de croissance annuels moyens significativement supérieurs », note Eric Hazan, directeur associé de McKinsey. Mais les « data-miners » ne courent pas les rues. CDiscount recrute des statisticiens. Priceminister aussi, pour sa base de données cumulant 15 millions de clients, près d'un Français sur quatre.

Trouver la perle rare, « ce n'est pas évident », dit à l'AFP Pierre Kosciusko-Morizet, son patron et fondateur. « Il faut vraiment aimer les chiffres et les maths ». Le profil idéal, c'est « un ingénieur intéressé par le business. C'est exceptionnel », ajoute M. Kosciusko-Morizet, qui estime la France pas si mal placée grâce à ses formations en maths fondamentales. « C'est un marché très tendu, il y a clairement un déficit d'offre », confirme Stéphane Trepoz, patron de Sarenza.

## On s'arrache les meilleurs

Son site emploie 5 « data miners » sur 40 informaticiens, pour « draguer légèrement plutôt que lourdement » les clients, c'est-à-dire leur envoyer des publicités bien ciblées. « Tout le monde s'arrache les meilleurs », témoigne Romain Niccoli, cofondateur et patron de la R&D chez Criteo. Ce spécialiste français de la pub en ligne cherche à pourvoir 100 postes dans la recherche et développement.

Les salaires sur le marché montent « très facilement », de 40.000 à 100.000 euros bruts annuels, selon les patrons de Sarenza et de Criteo. Ils culminent à 250.000 voire 500.000 euros pour un « data scientist », un expert recruté par l'une des dix entreprises les plus en pointe dans le monde, selon M. Bessis.

Mais « ingénieurs ou docteurs ne viennent pas naturellement, ça nous a pris un certain temps pour leur expliquer » l'intérêt de ce travail, explique Romain Niccoli, lui-même diplômé des Mines. « Il y a encore quelques années quand on démarchait les candidats c'était même difficile d'obtenir un rendez-vous », ajoute-t-il. « C'est compliqué pour une entreprise mainstream driffrer ces gens car ils aiment une ambiance de laboratoire de recherche, avec des collègues très brillants », ajoute M. Bessis.

Plutôt que le-commerce, certains vont vers l'assurance ou la finance, « qui offre des problèmes théoriques assez complexes », ou encore la santé publique, aussi consommatrice de données, explique Stéphane Tufféry, professeur à l'ENSAI et à Rennes-1, dont « 100% des élèves trouvent un travail rapidement ».

SOURCE AFP

Café Digital : Arnaud Barey  
(Voyagermoinscher.com)



## La baisse du titre Facebook...

**JUSTIFIÉE**      **INJUSTIFIÉE**      **VALIDEZ**

9271 vote(s)

Par Ludovic Desaulez

**Tablette à tout prix**

Tout comme Nestlé avec son Nespresso, qui a réussi

à propulser le café dans l'univers du luxe, Apple a installé la tablette en classe...

**Arnaud Barey :**

**VoyagerMoinsCher.com sur la vague de l'ultra dernière minute**

Par Les Echos

Le cofondateur du Voyagermoinscher.com décrypte les tendances phares de cet été 2012 dans le tourisme en ligne, notamment l'émergence de...

Tous les billets

**LE FIL**

13:22 | Juncker estime que la Slovénie ne demandera pas de plan de sauvetage (AFP)

12:57 | Libor : la Banque des règlements internationaux s'en mêle

12:54 | Le déficit commercial du Portugal diminue de moitié en juillet (AFP)

12:50 | Timide hausse des Bourses européennes à la mi-séance (Reuters)

12:45 | La Grèce et la troïka s'efforcent d'aplanir leurs divergences (Reuters)

12:38 | Evasion fiscale: Combadiés "ignore" le travail du Sérial (satellite centriste) (AFP)

**POUR ACCÉDER AU MEILLEUR DE L'INFORMATION**

**ABONNEZ-VOUS MAINTENANT**

12:32 | Impôt sur le revenu : 16 millions de foyers



# Fouille dans les données

Le data mining est le terme utilisé pour désigner l'identification de tendances récurrentes à partir de stocks de données volumineux. Ces modèles livrent souvent des informations précieuses sur différents types de données commerciales, notamment en ce qui concerne le comportement des clients. Cette fouille marketing – qui date de l'époque des pharaons – a longtemps été mise aux oubliettes pour ressortir de nos jours, suite à l'engorgement d'informations numériques.

Le data mining est un procédé récent, certes. Mais la génération de données à partir d'un grand nombre d'informations date de fort longtemps. Wikipedia.org relève qu'il faut remonter au cinquième siècle avant Jésus-Christ jusqu'au pharaon Amasis pour sapercevoir que celui-ci organise le premier recensement de sa population. A cette époque, pas d'informatique, mais il lui a bien fallu trier un grand nombre d'informations pour s'organiser. Ce n'était certes pas du data mining, mais plutôt du data collecting. Ce n'est qu'au XVII<sup>eme</sup> siècle qu'on commence à vouloir analyser les données pour en rechercher des caractéristiques communes. En 1763, le mathématicien britannique Thomas Bayes démontre la possibilité de déterminer, non seulement des probabilités à partir des observations issues d'une expérience, mais aussi les paramètres relatifs à ces probabilités. Déjà pas mal. Dans les années 1950, les gros calculateurs font leur apparition. Encore onéreux, ceux-ci opèrent grâce à des techniques de calcul par lots. Simultanément, des méthodes et des techniques voient le jour telles que la segmentation, la classification (méthodes des nuées dynamiques), première version des réseaux de neurones (Perceptron), 1960 voit arriver dans les entreprises les arbres de décision. En 1969, paraît l'ouvrage Myron Tribus qui généralise les méthodes du calcul automatique (le début du langage Basic). L'expression « data mining » avait d'ailleurs à cette époque de conquête lunaire, une connotation péjorative, exprimant le mépris des statisticiens pour les démarches de recherche de corrélation sans hypothèses de départ.

## Minage numérique

« Son origine remonterait à la première conférence annuelle « Knowledge Discovery in Databases » (KDD) de 1989 ([www.kdnuggets.com](http://www.kdnuggets.com)), relève Sandro Saitta, président de la « Swiss Association for Analytics ». Il faut donc attendre jusqu'au début de l'an 2000 pour que les grands groupes informatiques s'intéressent au domaine, à l'instar d'Amazon, puis de Google et eBay. Or, voilà, les ordinateurs, puis maintenant les

tablettes et les smartphones, nous ont habitué à dépendre d'eux. A tel point que leur stockage et leur exploitation sont devenus quasiment impossibles avec des outils de gestion de données classique. « Cette problématique nous concerne tous puisque nous produisons une quantité astronomique de données digitales chaque jour. Ces données viennent de partout : de capteurs, de sites internet, de médias sociaux, etc. Ce sont des textes, des images, des vidéos, des signaux GPS, pour ne citer que quelques exemples », relève Sandro Saitta.

## Basé sur la statistique et l'apprentissage automatique

Le data mining, textuellement minage de données – mais souvent traduit en français par fouille de données – est donc un domaine qui consiste à collecter, agrégier et analyser une grande quantité de données dans le but de les expliquer ou de faire des prédictions. Son objectif ? Extraire de la connaissance actionnable à partir de données existantes. Le data mining se base principalement sur des méthodes venant du monde de la statistique et de l'apprentissage automatique (machine learning). En d'autres termes, le data mining cherche à identifier des tendances parmi les données dans le but de prendre une décision.

En résumé, le data mining se situe à l'intersection de trois domaines : les statistiques, le machine learning et les bases de données. « En statistiques, la procédure habituelle est de faire des hypothèses sur les données et ensuite de tester ces hypothèses. Quand les ensembles de données deviennent très grands, cette tâche devient difficile, voire impossible. Avec le data mining, on applique des algorithmes pour répondre à un problème business. Le but n'étant généralement pas de faire des hypothèses, mais de faire des prédictions », explique encore le président de la « Swiss Association for Analytics ».

## Détection de fraudes

Aujourd'hui, l'utilisation industrielle ou opérationnelle du data mining dans le monde professionnel permet de résoudre des problèmes très divers, allant de la

gestion de la relation client à la maintenance préventive, en passant par la détection de fraudes ou encore l'optimisation de sites internet. L'exploration de données fait suite à l'informatique décisionnelle. Celle-ci permet de constater un fait, tel que le chiffre d'affaires, de l'expliquer, de le prévoir. Une chose est sûre. Le data mining est voué à un grand essor. L'Université de Genève (UNIGE) offre des cours qui traitent notamment de l'apprentissage supervisé (classification, régression), des techniques d'évaluation et d'expérimentation, de l'agrégation de modèles (bagging, boosting, stacking), arbres et règles de décision, etc. Le groupe français Altran (présent à Lausanne, Clarens et Genève notamment) offre également des services de data mining qui vont, par exemple, de la détermination des probabilités et des potentiels de vente croisée, des probabilités d'annulation et de résiliation, de détection contre la fraude ou même de modèles de risques et modélisation de l'historique des événements (calcul de valeurs extrêmes).

## 41 % d'entreprises désireuses d'améliorer l'utilisation de leurs données

En 2012, selon la plate-forme d'intégration américaine Talend ([www.talend.com](http://www.talend.com)), 41 % des entreprises mondiales ont affirmé avoir une stratégie pour améliorer la gestion et l'utilisation de leurs données. « Faire parler » les informations brutes et identifier des groupes d'individus aux comportements similaires, voilà les objectifs recherchés par les algorithmes utilisés par le data mining. « Cette approche permet de pousser plus loin l'analyse des données nécessaires à la bonne connaissance des contacts et d'améliorer les prises de décisions des marketeurs. Les choix de ces derniers ne se basent plus seulement sur des intuitions mais sur des faits concrets ! », estime la société française Dolist.net ([www.dolist.net](http://www.dolist.net)), spécialiste de l'email marketing. Sage cons-tatation. (rké)

Info : [www.swiss-analytics.ch](http://www.swiss-analytics.ch)

# «Toute entreprise sérieuse se doit d'utiliser le data mining»

Explorer une astronomique quantité de données, puis les analyser dans le but de les expliquer ou d'en faire des prédictions... Le data mining est devenu une méthode de marketing très convoitée par les grandes firmes. A part Google, Amazon ou eBay, la Suisse se met aussi à traiter ces infos, à l'instar de Swisscom, Migros ou Nestlé. Analyse.



Alain Blumenthal

## Monsieur Saitta, en quoi le domaine du marketing peut-il tirer profit de cette extraction des connaissances à partir de grands volumes de données ?

Le domaine du marketing et de l'analyse client (Customer Relationship Management, CRM) est certainement un des plus concerné par le data mining. Etant donné que les entreprises développent la collecte de données clients, ces données peuvent être utilisées à des fins de marketing. Les applications les plus fréquentes sont la vente croisée (cross-selling) et la vente incitative (up-selling) : quel est le produit susceptible d'être le plus sollicité par chaque client ? Pour cela, le data mining permet d'analyser les données clients et de comprendre les tendances qui définissent des profils utilisateurs. Dans le passé, j'ai travaillé sur un projet de ciblage comportemental pour une société de télécommunication suisse. Le but était d'utiliser le data mining pour prédire la probabilité à afficher pour chaque visiteur du site web. Le résultat est aisément mesurable, puisqu'on peut comparer le taux de clics des visiteurs cibles par le data mining avec un groupe de contrôle choisi aléatoirement. Un autre exemple est la prévention des débauchements (churn management). La question à laquelle le data mining répond s'avère celle-ci : quels sont les risques de voir parvenir un client vers la concurrence ? Avoir ce genre d'information est très précieux pour une entreprise, car cela lui permet d'agir en fonction, par exemple en faisant une offre promotionnelle afin de fidéliser la clientèle à risque.

## Comment les entreprises et les organismes – grands magasins, grandes firmes, états – s'y prennent-ils pour analyser leurs données ?

Il y a plusieurs manières pour une entreprise possédant des données de faire du data mining. Elle peut le faire en interne, en embauchant un data miner (aussi appelé «data scientist» plus récemment). Celui-ci pourra alors mettre en place des projets de data mining au sein de l'entreprise. Il se créera alors une collaboration entre le data miner et plusieurs acteurs au sein de l'en-

Sandro Saitta : «Etant donné qu'il faut des données, beaucoup de données, les grandes entreprises sont les premières clientes du data mining».

treprise. Le client interne, celui pour qui le data miner réalise l'analyse, fera le lien avec la compréhension business du problème à résoudre. Le service informatique fournira l'accès aux bases de données. Enfin, les experts-domaines contribueront à la compréhension des données et à leurs particularités. On voit donc que le facteur humain est crucial pour le bon fonctionnement d'un projet de data mining. Du point de vue des outils, le data miner utilise principalement deux types de software : les systèmes de management de bases de données (Oracle, MySQL) et les solutions d'analyses statistiques (SAS, R, MATLAB). Le processus est similaire avec un consultant externe.

## Quels sont les enjeux économiques du data mining ?

Pour pouvoir rester compétitive, une entreprise doit vendre plus, réduire ses coûts de logistiques au minimum ou encore optimiser sa production. Pour autant que l'on possède des données de bonne qualité, ces défis peuvent être relevés avec des techniques de data mining. D'un point de vue économique, l'idée derrière le data mining est celle d'améliorer la prise de décision avec des données (data-driven decision making). Etant donné qu'il faut des données, beaucoup de données, les grandes entreprises sont les premières, les grandes entreprises sont les pre-

mières clientes du data mining. On pense en premier à des entreprises qui ont fait du data mining leur business, comme Google, Amazon, ou Ebay. Cela dit, en Suisse aussi, les entreprises font du data mining : Swisscom, Migros ou Nestlé font du data mining. Pour moi, toute entreprise sérieuse qui possède de grands volumes de données se doit d'utiliser le data mining.

## « Le problème n'est pas le data mining en soi, mais la collecte de données qui peuvent être considérées comme sensibles »

mères clientes du data mining. On pense en premier à des entreprises qui ont fait du data mining leur business, comme Google, Amazon, ou Ebay. Cela dit, en Suisse aussi, les entreprises font du data mining : Swisscom, Migros ou Nestlé font du data mining. Pour moi, toute entreprise sérieuse qui possède de grands volumes de données se doit d'utiliser le data mining.

### En quoi l'armée pourrait-elle se servir du data mining pour espionnage ou pour lutter contre une cyberguerre ?

Les gouvernements font du data mining pour plusieurs raisons. Les applications les plus fréquentes sont l'optimisation de processus existants, la détection de fraudes ou encore la lutte contre l'évasion fiscale. On peut donc imaginer que la plupart des utilisations du data mining faites par les gouvernements sont légitimes ou du moins bénéfiques pour les honnêtes citoyens. Dans le cas de l'armée ou des services de renseignements, les choses deviennent plus discutables. Quelles données un gouvernement a-t-il le droit de collecter et d'analyser dans

le but d'améliorer la sécurité de la population ? Vos données bancaires, les sites web que vous fréquentez, vos données médicales... Où est la limite ? On peut imaginer l'intérêt des gouvernements pour l'analyse de toutes les données à disposition aux fins de prédire quel individu a le plus de chance d'être une menace pour le pays. Le pays en question peut évidemment avoir des désavantages plus critiques, comme la veille technologique ou la concurrence économique.

**Y a-t-il du data mining derrière PRISM et si oui, quel genre d'application ?**

Sans connaître les secrets des projets comme PRISM, on ne risque pas de se tromper en disant qu'il s'agit sûrement de la connaissance à partir d'une grande quantité de données. Il y a donc clairement du data mining derrière ce genre de projets. Quand on lit certains articles et autres blogs sur internet, on constate que de nombreuses personnes possèdent une mauvaise image du data mining. Il y a eu tellement de scandales sur l'analyse de données, notamment aux USA, que le data mining est souvent perçu à tort comme de l'espionnage effectué par le gouvernement. Cela dit, il faut préciser que le vrai problème n'est pas le data mining en soi, mais la collecte de données pouvant être considérées comme sensibles. En ce

qui concerne les applications, je pense par exemple à l'analyse de réseaux sociaux. Cela consiste à surveiller des sites comme Twitter, Facebook ou LinkedIn pour mettre en évidence des individus suspects.

**Comment voyez-vous l'avenir du data mining tenant compte du fait que les volumes vont sans cesse prendre de l'ampleur ?**

De nos jours on parle souvent de Big Data pour faire référence au volume, à la variété et la vitesse à laquelle les données sont générées. Cela dit, il s'agit d'un problème d'architecture et de format de base de données ainsi que de parallelisation. Le défi du data mining est d'obtenir de la connaissance actionnable, quel que soit le volume de données. C'est plutôt du côté des applications possibles qu'il faut se tourner. Hier, le data mining était utilisé dans le monde académique ainsi que dans le marketing. Demain, le data mining sera partout : vie personnelle, publicité, santé, fabrication, gouvernement ou encore ressources humaines. Le livre

« Analytics », donne un large panel d'applications possibles.

### Un big-bang du data mining serait-il possible ?

Il n'y aucune raison d'imaginer un big-bang du data mining puisque que celui-ci s'intègre de manière transparente dans les applications. Aujourd'hui, qui pense au data mining lorsque sa carte de crédit est bloquée, qu'en lui offre une assurance complémentaire ou qu'en lui propose un nouveau contrat de téléphone mobile ? Pourtant, le data mining est derrière ces offres et ces recommandations ! Le data mining fait déjà partie intégrante de nos vies. Ce que l'avoir nous réserve, ce n'est pas un big-bang, mais une multitude d'applications dans des domaines que l'on n'imagine pas encore aujourd'hui. 

Interview: Roland Keller  
Rédacteur responsable  
SWISS ENGINEERING RTS

### Regard sur

**Sandro Salta**  
Né le 23 avril 1981, Sandro Salta occupe la fonction de chercheur en data mining dans une entreprise de sécurité depuis 2011. Il obtient son master en informatique à l'EPFL (2004) suite à un projet de data mining sur la prédition de policien, en collaboration avec MeteoSwiss. La possibilité d'utiliser des algorithmes pour faire apprendre une tâche

à un ordinateur le motive à continuer par un doctorat en data mining à l'EPFL. Durant sa thèse, il applique le data mining pour aider les ingénieurs civils à interpréter les données qu'ils collectent. Il propose alors d'utiliser des approches de clustering [groupement] et feature selection [sélection de paramètres] pour aider les ingénieurs à prendre une décision lors de diagnostics de structures. En 2006, il lance le blog « Data Mining Research » ([www.dataminingblog.com](http://www.dataminingblog.com)) et discute de recherche, applications, livres, logiciels et jobs dans le domaine du data mining. Il décroche son doctorat en 2008 et part appliquer le data mining dans le monde de la finance. En 2009, il rejoint FinScore, une société de consulting et travaille sur un projet de ciblage comportemental pour la publicité en ligne. Ce projet est réalisé pour le compte d'une société de télécommunication suisse. En 2011, il rejoint une entreprise de sécurité en tant que « Data Mining Research Engineer » où il applique le data mining dans des projets variés. En 2012, il crée, sur son temps libre et avec l'aide d'autres passionnés, la « Swiss Association for Analytics », dont il devient le président.

# Une association pour mieux comprendre les données

Dans le but de faire comprendre et de promouvoir le data mining mais également l'analyse de données, le machine learning, la statistique et le web analytique, l'Association suisse pour l'analytique « Swiss Association for Analytics » a pignon sur rue sur internet grâce au réseau LinkedIn. Mais l'organisation est aussi active sur le terrain, créant des rencontres de réseautage.



Pour qu'une entreprise puisse bénéficier du data mining, il lui faut des données, des outils et du savoir-faire.

Depuis début 2012, une association suisse pour l'analyse des données « Swiss Association for Analytics » est en activité à Lausanne. Son objectif principal vise à sensibiliser les entreprises suisses aux possibilités du data mining. « Que ce soit dans les domaines bancaire, financier, pharmaceutique, télé-communication ou e-commerce importe peu. Pour qu'une entreprise puisse bénéficier du data mining, il lui faut des données, des outils et le savoir-faire. Notre association peut aider les entreprises à choisir les données sur lesquelles travailler, à découvrir les outils existant sur le marché ou encore à obtenir le savoir-faire (où trouver les personnes possédant ce savoir-faire) », explique son fondateur et président Sandro Saitta.

Comment vous est venue l'idée de créer ce

groupement ? « Fin 2011, j'ai été invité en Belgique pour donner une présentation organisée par BAQMaR, une association d'analytique belge. L'événement était local, mais spécifique et ciblé sur l'industrie. Dans l'avion, sur le chemin du retour, je me suis dit qu'une telle initiative était une excellente idée et devrait voir le jour en Suisse aussi », précise-t-il. Du coup, l'affaire était lancée : « Un groupement de ce genre n'existeait tout simplement pas en Suisse. Seuls des événements marketing (organisés par des vendeurs de logiciels) ainsi que des événements académiques avaient lieu. »

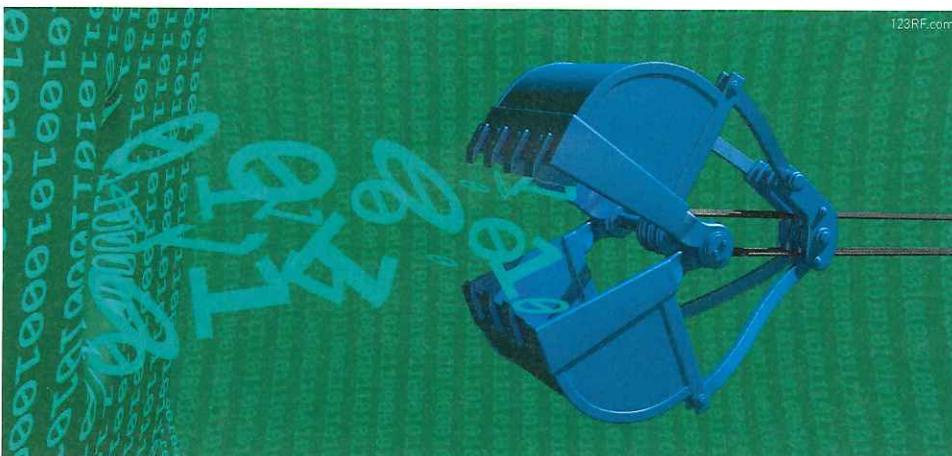
**Information et promotion**  
L'association est donc une organisation à but non lucratif qui a pour objectif de promouvoir le domaine analytique en Suisse. Par analytique, sont inclus l'analyse de données, le data mining, le machine learning, la statistique et le web analytique, entre autres. Le comité d'organisation est actuellement composé de cinq personnes qui garantissent les prestations suivantes :

- Faire passer le message concernant l'analytique
- Montrer la valeur ajoutée aux entreprises suisses
- Fournir une plate-forme de réseautage pour les praticiens suisses
- Promouvoir le contact avec d'autres organisations ayant des objectifs similaires.
- « Pour atteindre ces objectifs, nous avons plusieurs moyens. Par exemple, nous avons créé un groupe LinkedIn qui compta

plus de 350 membres (l'adhésion est pour le moment gratuite). L'orientation de l'association est clairement axée sur l'industrie. On y discute applications, cas d'études, défis, articles, offres d'emploi, etc.», précise le président.

#### Rencontre le 9 septembre, à Lausanne

L'association organise aussi des événements sur l'analytique. En mars de cette année, a eu lieu à Lausanne le premier événement de ce genre en Suisse. Plusieurs experts de divers horizons sont venus présenter des cas concrets d'analytique. Le CEO de la spin-off de l'EPFL Predigo a présenté le principe des recommandations de produits sur Internet. Un spécialiste de Nestlé a discuté du forecasting (prévision) de la demande. Enfin, un professeur de HEC-VD est venu introduire certains concepts clés en data mining. «Nous avons compté plus de 45 personnes, ce qui est clairement un succès. C'est pourquoi nous organisons un deuxième événement le 9 septembre, à 18h, à Lausanne», ajoute Sandro Saitta. Le thème choisi est Visualizing Analytics (analytique visuel). Ces séminaires permettent aussi aux experts dans le domaine de se rencontrer et déchanter leurs points de vue ou



leurs problèmes d'analyses: «Toutes les présentations sont en anglais pour accueillir un public plus large. Avis aux intéressés: nous sommes toujours à la recherche d'otateurs et de sponsors».

#### Opportunité de réseautage

Grâce à cette association, les entreprises suisses peuvent découvrir ce qu'est véritablement le data mining, en tous cas en prendre les tenants et aboutissants. Autrement dit de voir ce qui est réalisé dans ce domaine et surtout comment le data mining est appliqué. L'organisation estime que «la prise de décision en entreprise est plus efficace si elle est basée sur des données (data-driven decision making). Selon Sandro Saitta, «l'époque de l'expert qui se base uniquement sur son instinct est révolue». La «Swiss Association for Analytics» est aussi une excellente opportunité de réseautage, que ce soit pour sonder le marché, pour lancer une collaboration ou même... pour trouver un emploi. De bonne augure! (rke) ©

Info:  
[www.swiss-analytics.ch](http://www.swiss-analytics.ch)  
[www.predigo.com](http://www.predigo.com)

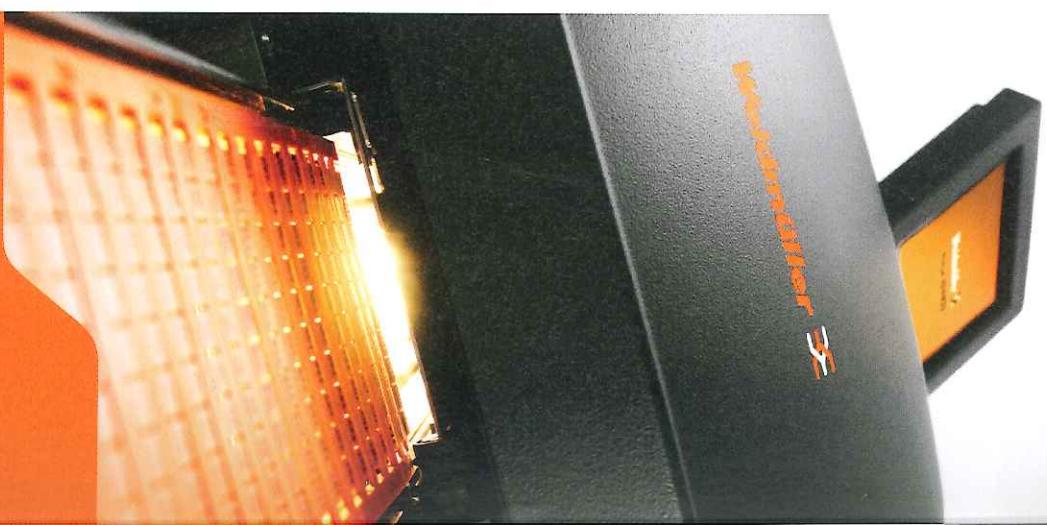
#### entre autres

##### SRT-15: Les gros échanges de données

Lancé en octobre 2010, le projet SRT-15 s'est intéressé au traitement et à l'acheminement de données des grands réseaux d'entreprises. L'équipe de l'Institut d'informatique de l'Université de Neuchâtel (UNINE) emmenée par le professeur Pascal Felber et Etienne Rivière, maître-assistant, s'est vu décerner récemment le prix du meilleur article scientifique lors d'une conférence internationale au Texas, pour une communication décrivant une de ses principales contributions au projet. Arrivé à son terme en mars 2013, le projet SRT-15 a en outre obtenu la mention «excellent» de la part du comité international d'experts chargé de juger l'ensemble de ses deux ans et demi d'activité.

Ce projet visait à construire une plateforme informatique adaptée à ces nouveaux enjeux, conjuguant facilité de programmation, performance et sécurité.

La contribution principale de l'UNINE est une nouvelle approche pour transmettre et rediriger des contenus, entre un grand nombre d'entités informatiques, en fonction des intérêts exprimés par les destinataires pour les contenus générés par les autres composants. Cet acheminement fondé sur le contenu se fait toutefois en protégeant les données et intérêts des applications et de leurs utilisateurs.



**Envie de combiner marquage et progrès. Notre PrintJet ADVANCED vous garantit un marquage systématisé**  
**Let's connect.**

Vous êtes fondées sur les bases les plus élevées de l'automatisation progressive et vous voulez changer vos besoins de marquage pour l'avenir? Que ce soit les IP 20 ou IP 67 - Allez-y toute simplement: Notre PrintJet Advanced est la meilleure solution du marché que vous pouvez trouver. Une solution automobile, 6000 repères en 45 minutes, imprime ainsi les marques en métal, layouts pré-installés, touchpanel intuitif, 25 langues et de nombreuses autres fonctionnalités innovantes qui sauront vous inspirer.

Let's connect. [www.printjet-advanced.com](http://www.printjet-advanced.com)

Grâce à «Swiss Association for Analytics», le data mining devient plus aisément à comprendre.

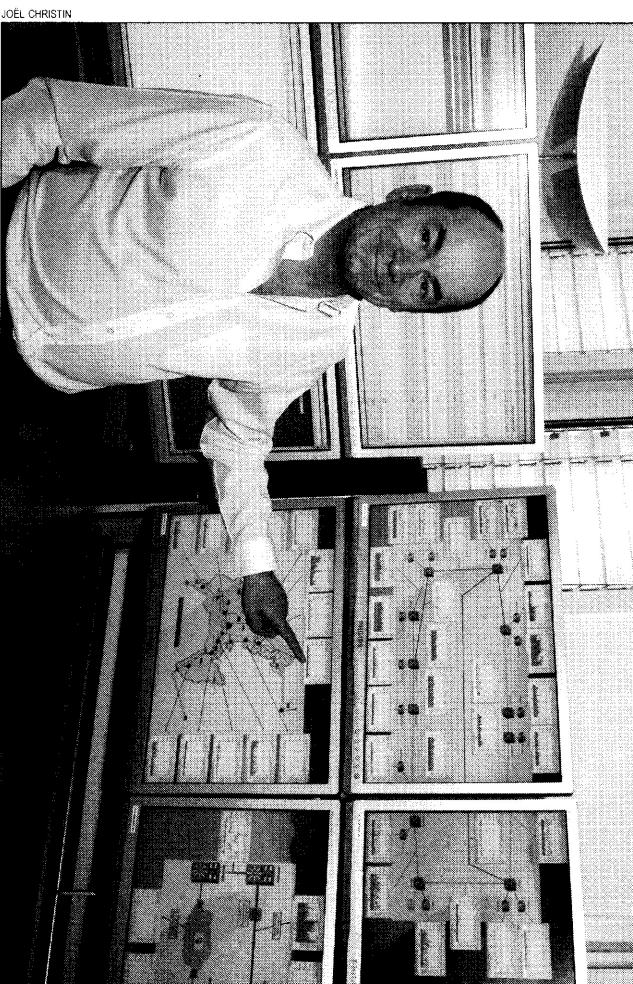
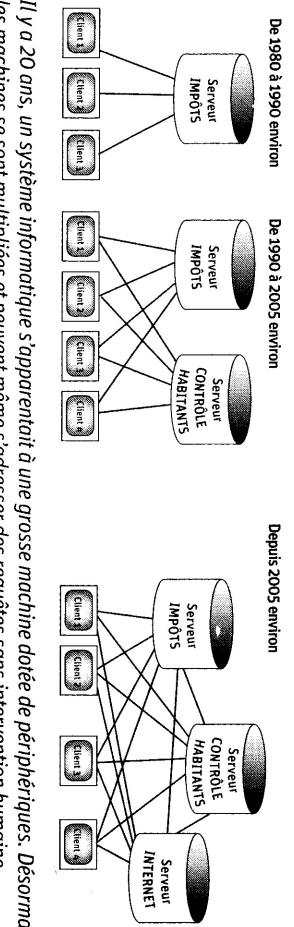


# Les probabilités pour faire face à la complexité

*L'évolution des techniques et des habitudes, ainsi que la multiplicité des métiers exercés à l'Etat rendent la sécurité informatique si complexe qu'elle doit être abordée sous l'angle probabiliste.*

«La complexité des systèmes est exponentielle. Une vision complète devient impossible. Les anciens critères de sécurité ne sont plus applicables.» Responsable de la récente unité Sécurité des systèmes d'information (SSI) à la DSJ, André Bourget compte jusqu'à partir d'aujourd'hui pour évoluer vers une sécurité informatique de l'Etat basée sur les probabilités, seule possibilité pour appréhender des interconnexions innombrables, encore centuplées par la multiplicité des métiers exercés à l'Etat.

Auparavant, il suffisait de contrôler les accès à la machine et les données qui transistaient jusqu'à l'unique serveur auquel elle était reliée. Mais les serveurs informatiques se sont multipliés et sont désormais reliés entre eux. On peut par exemple imaginer un système de radar pour les automobiles entièrement automatisé. Une machine contrôle la vitesse, prend une photo et envoie ces données à une autre machine susceptible d'interroger des registres de numéros d'immatriculation sur des serveurs de différents cantons ou pays. Le tout est transmis à une troisième machine pour la facturation. Le nombre potentiellement infini de connexions et de requêtes engendrées par ces processus rend impossible les vérifications exhaustives. Dès lors, comme pour les contrôles radar, la sécurité informatique planifiera une partie de ses contrôles dans le domaine opérationnel selon des statistiques de risques. – JC



## Rattachement à la DSJ

### Le rôle de l'utilisateur

Jusqu'à fin 2009, la sécurité informatique était l'affaire de l'OSIC (Office de la sécurité informatique), rattaché au DSE. La nouvelle organisation de l'informatique a conduit à dissoudre l'OSIC, remplacé par deux structures: la SSI (Sécurité des systèmes d'information) et l'ASSTI (Audit de sécurité du système d'information et télécom).

Jusque dans les années 2000, la sécurité informatique est restée un domaine bien balisé. Depuis, l'évolution technique et le développement de la mobilité ont changé la donne. Avec l'apparition des ordiphones et autres machines multimédias transportables partout, le comportement réfléchi et conscient des utilisateurs devient central. Chacun sait qu'il faut se méfier des courriels de provenance inconnue. Ou qu'un clic de souris trop rapide sur un site internet peut ouvrir la porte à des virus ou à des chevaux de Troie susceptibles de s'approprier des mots de passe. Mais il faut aussi avoir toujours à l'esprit que la plupart des utilisateurs de l'administration sont en contact avec des données sensibles. La simple perte d'un téléphone mobile – quand ce n'est pas un ordinateur portable – contenant un identifiant ou d'autres données peut être lourde de conséquences.



## Business Computing

### Data Analysts Captivated by R's Power



[More Articles in Technology »](#)



**Get DealBook by E-Mail**

**Subscribe to Technology RSS Feeds**

|                                    |                               |
|------------------------------------|-------------------------------|
| <a href="#">Technology News</a>    | <a href="#">Bits Blog</a>     |
| <a href="#">Internet</a>           | <a href="#">Personal Tech</a> |
| <a href="#">Start-Ups</a>          | <a href="#">Pogue's Posts</a> |
| <a href="#">Business Computing</a> |                               |

**MOST POPULAR - TECHNOLOGY**

[E-MAILED](#)

[BLOGGED](#)

[VIEWED](#)

- [1. Game Maker Without a Rule Book](#)
- [2. Unboxed: Tech's New Wave, Driven by Data](#)
- [3. Google Struggles to Unseat Amazon as the Web's Most Popular Mall](#)
- [4. State of the Art: Smartphone? Presto! 2-Way Radio](#)
- [5. William Moggridge, Designer and Laptop Pioneer, Dies at 69](#)
- [6. Bits: Big Data in Your Blood](#)
- [7. App Smart: For a Baby, Mobile Screens to Shake and Rattle](#)
- [8. Amazon Updates Its Kindle Line of E-Readers](#)
- [9. Amazon to Allow Kindle Fire Users to Pay to Block Ads](#)
- [10. Gadgetwise: Another Way to Take Your iPad Underwater](#)

[Go to Complete List »](#)

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By [ASHLEE VANCE](#)

Published: January 6, 2009

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as [Google](#), [Pfizer](#), [Merck](#), [Bank of America](#), the InterContinental Hotels Group and Shell use it.

But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use.

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google, which uses the software widely. "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

It is also free. R is an open-source program, and its popularity reflects a shift in the type of software used inside corporations. Open-source software is free for anyone to use and modify. [IBM](#), [Hewlett-Packard](#) and [Dell](#) make billions of dollars a year selling servers that run the open-source Linux operating system, which competes with Windows from Microsoft. Most Web sites are displayed using an open-source application called [Apache](#), and companies increasingly rely on the open-source MySQL database to store their critical information. Many people view the end results of all this technology via the Firefox Web browser, also open-source software.

R is similar to other programming languages, like C, Java and Perl, in that it helps people perform a wide variety of computing tasks by giving them access to various commands.

For statisticians, however, R is particularly useful because it contains a number of built-in mechanisms for organizing data, running calculations on the information and creating graphical representations of data sets.

| WORLD                             | U.S.               | N.Y. / REGION                     | BUSINESS                 | TECHNOLOGY                | SCIENCE                            | HEALTH                    | SPORTS                    | OPINION                   | ARTS                          | STYLE                         | TRAVEL                        | JOBs                 | REAL ESTATE                 | AUTOS                 |
|-----------------------------------|--------------------|-----------------------------------|--------------------------|---------------------------|------------------------------------|---------------------------|---------------------------|---------------------------|-------------------------------|-------------------------------|-------------------------------|----------------------|-----------------------------|-----------------------|
| <a href="#">Search Technology</a> | <a href="#">Go</a> | <a href="#">Inside Technology</a> | <a href="#">Internet</a> | <a href="#">Start-Ups</a> | <a href="#">Business Computing</a> | <a href="#">Companies</a> | <a href="#">Companies</a> | <a href="#">Bits Blog</a> | <a href="#">Personal Tech</a> | <a href="#">Personal Tech</a> | <a href="#">Personal Tech</a> | <a href="#">Jobs</a> | <a href="#">Real Estate</a> | <a href="#">Autos</a> |

[Advertiser](#) | [NYTimes.com](#) | [Feedback](#) | [Help](#) | [Privacy](#) | [Terms of Service](#) | [About Us](#) | [Work With Us](#) | [Contact Us](#)

Some people familiar with R describe it as a supercharged version of Microsoft's Excel spreadsheet software that can help illuminate data trends more clearly than is possible by entering information into rows and columns.

What makes R so useful — and helps explain its quick acceptance — is that statisticians, engineers and scientists can improve the software's code or write variations for specific tasks. Packages written for R add advanced algorithms, colored and textured graphs and mining techniques to dig deeper into databases.

Close to 1,600 different packages reside on just one of the many Web sites devoted to R, and the number of packages has grown exponentially. One package, called *BiodiversityR*, offers a graphical interface aimed at making calculations of environmental trends easier.

Another package, called *Emu*, analyzes speech patterns, while *GenABEL* is used to study the human genome.

The financial services community has demonstrated a particular affinity for R; dozens of packages exist for derivatives analysis alone.

"The great beauty of R is that you can modify it to do all sorts of things," said Hal Varian, chief economist at Google. "And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants."

R first appeared in 1996, when the statistics professors Ross Ihaka and Robert Gentleman of the University of Auckland in New Zealand released the code as a free software package.

According to them, the notion of devising something like R sprang up during a hallway conversation. They both wanted technology better suited for their statistics students, who needed to analyze data and produce graphical models of the information. Most comparable software had been designed by computer scientists and proved hard to use.

Lacking deep computer science training, the professors considered their coding efforts more of an academic game than anything else. Nonetheless, starting in about 1991, they worked on R full time. "We were pretty much inseparable for five or six years," Mr. Gentleman said. "One person would do the typing and one person would do the thinking."

Some statisticians who took an early look at the software considered it rough around the edges. But despite its shortcomings, R immediately gained a following with people who saw the possibilities in customizing the free software.

John M. Chambers, a former Bell Labs researcher who is now a consulting professor of statistics at [Stanford University](#), was an early champion. At Bell Labs, Mr. Chambers had helped develop S, another statistics software project, which was meant to give researchers of all stripes an accessible data analysis tool. It was, however, not an open-source project.

The software failed to generate broad interest and ultimately the rights to S ended up in the hands of Tibco Software. Now R is surpassing what Mr. Chambers had imagined possible with S.

"The diversity and excitement around what all of these people are doing is great," Mr. Chambers said.

While it is difficult to calculate exactly how many people use R, those most familiar with the software estimate that close to 250,000 people work with it regularly. The popularity of R at universities could threaten SAS Institute, the privately held business software company that specializes in data analysis software. SAS, with more than \$2 billion in annual revenue, has been the preferred tool of scholars and corporate managers.

"R has really become the second language for people coming out of grad school now, and there's an amazing amount of code being written for it," said Max Kuhn, associate director of nonclinical statistics at Pfizer. "You can look on the SAS message boards and see there is a proportional downturn in traffic."

SAS says it has noticed R's rising popularity at universities, despite educational discounts on its own software, but it dismisses the technology as being of interest to a limited set of

people working on very hard tasks.

"I think it addresses a niche market for high-end data analysts that want free, readily available code," said Anne H. Milley, director of technology product marketing at SAS. She adds, "We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet."

But while SAS plays down R's corporate appeal, companies like Google and Pfizer say they use the software for just about anything they can. Google, for example, taps R for help understanding trends in ad pricing and for illuminating patterns in the search data it collects. Pfizer has created customized packages for R to let its scientists manipulate their own data during nonclinical drug studies rather than send the information off to a statistician.

The co-creators of R express satisfaction that such companies profit from the fruits of their labor and that of hundreds of volunteers.

Mr. Ihaka continues to teach statistics at the University of Auckland and wants to create more advanced software. Mr. Gentleman is applying R-based software, called Bioconductor, in work he is doing on computational biology at the Fred Hutchinson Cancer Research Center in Seattle.

"R is a real demonstration of the power of collaboration, and I don't think you could construct something like this any other way," Mr. Ihaka said. "We could have chosen to be commercial, and we would have sold five copies of the software."

A version of this article appeared in print on January 7, 2009, on page B6 of the New York edition.

Follow the IHT for world news, cartoons and opinion.

[More Articles in Technology »](#)

SIGN IN TO E-MAIL

PRINT

REPRINTS

## Related Articles

**FROM THE NEW YORK TIMES**  
Data Analysis Are Mesmerized by the Power of Program R

(January 7, 2009)

## INSIDE NYTIMES.COM

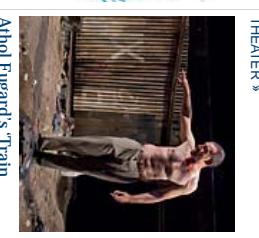
BUSINESS »



OPINION »



THEATER »



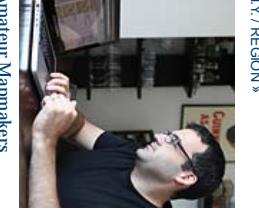
OPINION »

### Op-Ed: What Keeps the Chinese Up at Night

In China, a spiritual hunger has taken hold even as physical hunger has receded.



N.Y. REGION »



Construction and Real Estate Hinder China

Antibiotics May Carry Serious Side Effects

Gary Gutting: What Work Is Really For

Athol Fugard's Train Driver

Home

World

U.S.

N.Y./Region

Business

Technology

Science

Health

Sports

Opinion

Arts

Style

Travel

Jobs

Real Estate

Automobiles

Back to Top

Copyright 2009 The New York Times Company

Privacy Policy

Terms of Service

Search

Corrections

RSS

First Look

Help

Contact Us

Work for Us

Site Map





# Sciences et découvertes

**Adieu le diagramme camembert!** Rendre lisibles des données en un coup d'œil est devenu une véritable science. Le designer anglais David McCandless

Anne-Muriel Brouet

Trop, c'est trop. L'océan d'informations dans lequel nous plonge Internet fait que plus rien n'émerge. Tourse noire dans les marais de Facebook, les sables mouvants d'Outlook et les inondations de Google. On se retrouve incapable de hiérarchiser et même de comprendre. Heureusement David McCandless vient à la rescoufse. Gourou d'une nouvelle science baptisée la visualisation des données, le journaliste et graphiste anglais propose de substituer le texte par l'image. Et il y réussit plutôt bien.

Attention ! Il ne s'agit pas de dessiner des camemberts pour remplacer un tableau de chiffres. «Je hais les diagrammes en fromage, j'en ai trop vu, ils me donnent la nausée», grince ainsi David McCandless.

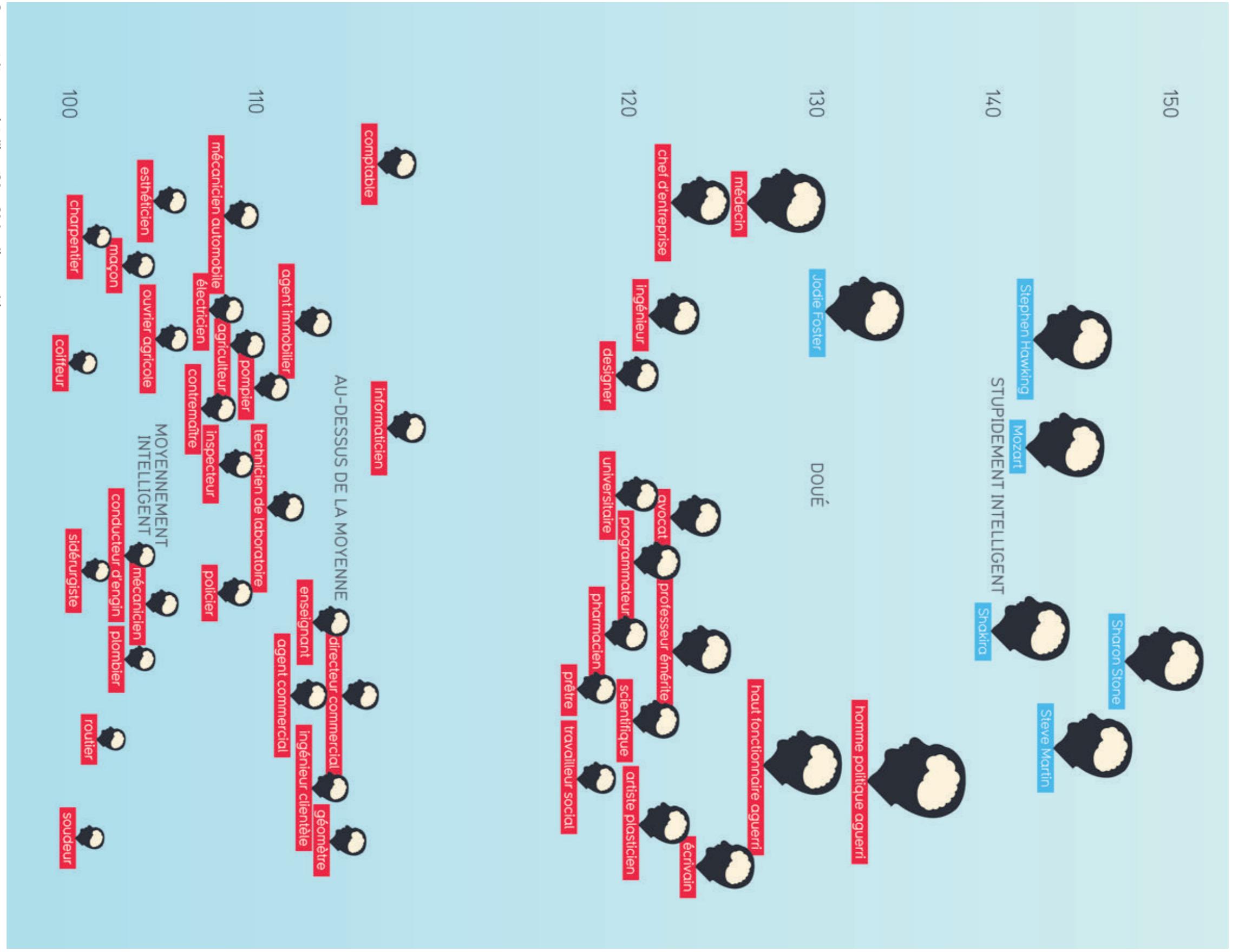
Le fait est qu'aujourd'hui «les informations sérieuses sont relayées en mots, graphiques et diagrammes. Les images ne constituent qu'un habillement cosmétique. C'est en train de changer», assure David McCandless, qui travaille notamment pour le quotidien britannique *The Guardian* et le magazine geek américain *Wired*. Comment ? En donnant «un visage aux informations et aux idées». Souvent avec humour.

La palette du graphiste comporte des éléments simplissimes, couleurs, lignes, formes (souvent circulaires), pictogrammes. S'ajoutent une bonne dose d'intelligence et une volonté féroce de sortir des pièges des médias traditionnels, surchargés et embourbés dans la banalité, le manque d'idée et... le trop-plein d'informations.

Ainsi, en cinq couleurs, il élabore une «matrice morale» permettant de déceler d'un coup d'œil la religion la plus tolérante (New Age) ou la pratique la plus réprobée (l'adultère). En jouant sur la taille des caractères, il met immédiatement en évidence la compagnie aérienne qui a connu le plus d'accidents mortels au cours des 50 dernières années (Aeroflot sans surprise). Avec une spirale de sphères décroissantes, il nous révèle les dangers de mort qui nous menacent. Il est ainsi clair que nous avons 100% de risque de mourir d'une manière ou d'une autre et que l'attaque de requin arrive en avant dernière position, avant la chute de météorite.

Le choix des sujets est dicté «par ma curiosité autant que par mon ignorance». Elle correspond plutôt bien à la nôtre.

**Datavision** David McCandless, octobre 2011, Editions Robert Laffont.

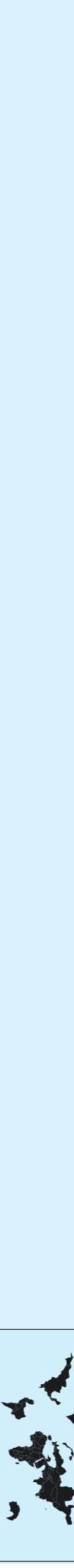


# Sciences et découvertes

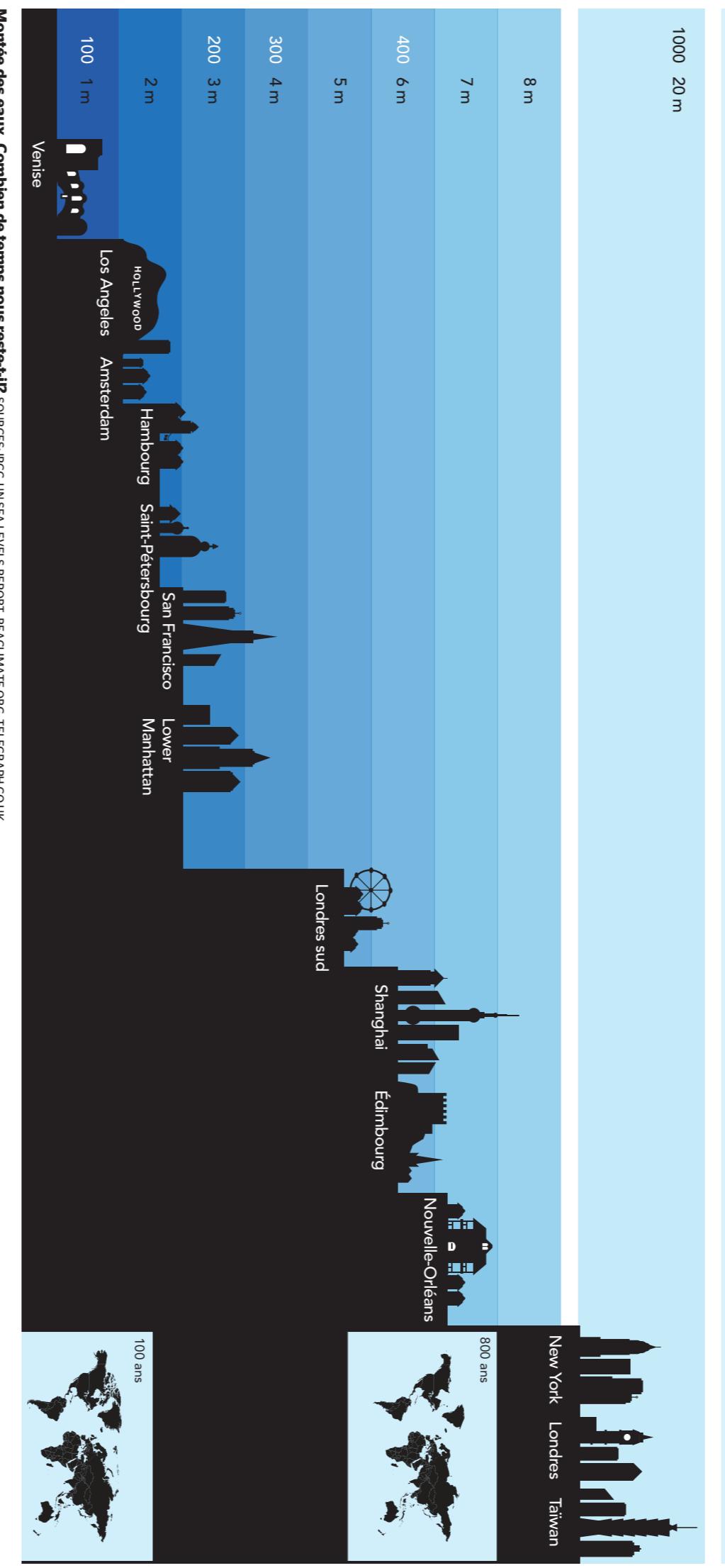
**Place à l'info belle et attrayante**  
Mc Candless s'est fait un nom dans ce domaine, devant le gourou de la visualisation

année élévation 

8000 80 m

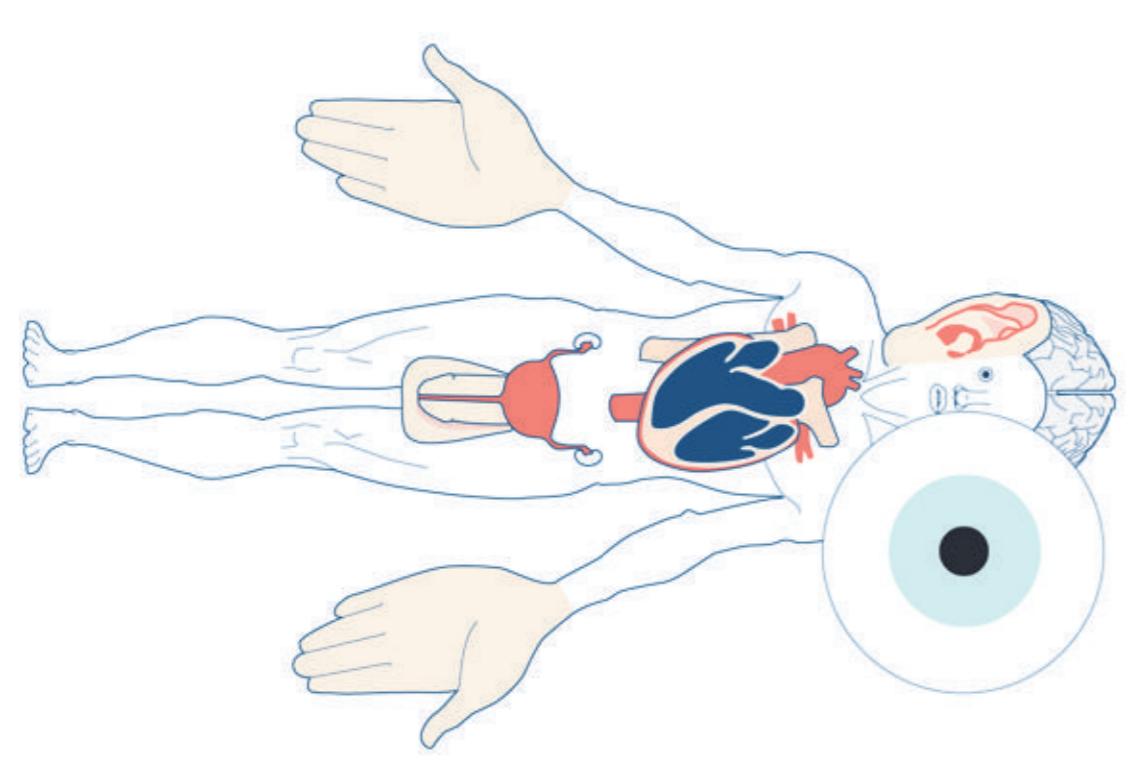


Montée des eaux. Combien de temps nous reste-t-il? SOURCES: IPCC, UN SEA LEVELS REPORT, REACLIMATE.ORG, TELEGRAPH.CO.UK



Les calories sortantes, brûlées en moyenne en 30 minutes de... SOURCES: CROISEMENT DE DONNÉES DE PLUSIEURS SITES DE RÉGIMES

Compte qualité



Le corps humain selon les résultats de recherche Google.

WIKIPEDIA



