FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE DU MODULE

Dans ce module, nous fournissons en termes très généraux quelques notions mathématiques et statistiques fondamentales nécessaires à l'analyse des données et à l'élaboration de modèles ayant des applications pratiques.

Les participants se familiariseront avec les **concepts clés** afin de faciliter leur futur apprentissage.

Cette introduction n'est pas destinée à remplacer une formation formelle et est au mieux **incomplète;** veuillez consulter les ouvrages de référence pour plus de détails.

APERÇU

1. Modélisation
2. Distributions
3. Théorème de la limite centrale
4. Estimation
5. Théorème de Bayes
6. Algèbre matricielle
7. Valeurs propres et vecteurs propres
8. Régression
9. Optimisation

MODÉLISATION

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Comprendre la différence entre la modélisation à partir des principes de base et de la modélisation statistique.

Avoir une connaissance pratique du processus de modélisation.

Mieux connaître les pièges et défis liés à la modélisation.

[http://dailycatdrawings.tumblr.com; nº309]

Monde réel

Modèle

Théorie

Repérage des détails pertinents pour la **description** et la **traduction** d'objets du monde réel en variables de modèle

LES MODÈLES EN GÉNÉRAL

Principes de base de la modélisation

- Examiner un système
- Écrire un ensemble de règles et d'équations qui décrivent l'essence du système
- Ignorer les détails qui compliquent les choses et qui sont « moins » importants

Modélisation statistique

- Habituellement, un ensemble d'équations comprenant des paramètres
- Les paramètres sont appris (le modèle est « entraîné ») à l'aide de multiples observations de données
- Échantillon de données c. population

HEURISTIQUE DE LA MODÉLISATION

Dans un sens, la modélisation est un processus **simple** (et **basé sur des formules**?), guidé par l'**intuition** et par l'**expérience** à chaque étape.

Voici les étapes de base de l'élaboration d'un modèle statistique :

- **Définition des objectifs**
  - Que tentons-nous d'accomplir?
  - Dans quelles situations le modèle sera-t-il utilisé et quel est le résultat que nous essayons de prévoir?
- **Collecte des données**
  - Quelles sont les données accessibles?
  - Combien d'enregistrements de données aurons-nous?
  - En général, les modélisateurs veulent autant de données que possible

HEURISTIQUE DE LA MODÉLISATION

Étapes de base de l'élaboration d'un modèle statistique (suite) :

- **Choix de la structure du modèle**

  - Doit-on exécuter une régression linéaire, une régression logistique ou un modèle non linéaire? De quel genre?

  - Le choix de la structure du modèle exige de l'expérience et une connaissance approfondie des forces et des faiblesses de chaque technique

- **Préparation des données**

  - Rassembler les données sous une forme appropriée pour le modèle

  - Encoder les données en entrées, en utilisant autant que possible des connaissances spécialisées

  - Séparer les données dans les ensembles d'apprentissage, d'essais et de validation souhaités

HEURISTIQUE DE LA MODÉLISATION

Étapes de base de l'élaboration d'un modèle statistique (suite) :

- **Sélection et suppression d'attributs**
  - Les variables sont examinées pour déterminer leur importance dans le modèle et sont sélectionnées ou éliminées
  - La liste des variables admissibles appropriées est classée par ordre d'importance
- **Élaboration des modèles admissibles**
  - Commencer par des modèles linéaires de base et essayer de les améliorer à l'aide de modèles non linéaires plus complexes
  - Garder à l'esprit l'environnement dans lequel le modèle sera mis en œuvre
- **Finalisation du modèle**
  - Sélectionner parmi les modèles admissibles le modèle le plus approprié pour la mise en œuvre
- **Mise en œuvre et surveillance**
  - Intégrer le modèle dans le processus système requis; mettre en œuvre des étapes de surveillance pour examiner le rendement du modèle

LES PIÈGES DE LA MODÉLISATION

Les pièges courants entourant le processus de modélisation :

- **Définition des objectifs**
  - Manque de clarté dans la définition du problème
  - Mauvaise compréhension de la façon dont le modèle sera utilisé et de l'environnement dans lequel il sera utilisé
- **Collecte des données**
  - Utilisation de données trop anciennes ou autrement non pertinentes pour l'avenir
  - Ne pas tenir compte d'autres sources ou ensembles de données clés qui pourraient être disponibles
- **Choix de la structure du modèle**
  - Utilisation d'une méthode de modélisation qui n'est pas adaptée à la nature des données (tailles, dimensions, bruit...)

LES PIÈGES DE LA MODÉLISATION

Les pièges courants entourant le processus de modélisation (suite) :

- **Préparation des données**
  - Ne pas nettoyer les données ou ne pas tenir compte des valeurs aberrantes
  - Ne pas correctement mettre les données à l'échelle
  - Ne pas suffisamment réfléchir à l'élaboration de variables spécialisées
  - Ne pas disposer de données provenant d'importantes catégories d'enregistrement de données
- **Sélection et suppression d'attributs**
  - Conserver trop de variables, ce qui complexifie la modélisation, l'interprétation, la mise en œuvre ou l'entretien des modèles
  - Trop se fier à la simple élimination des variables corrélées

LES PIÈGES DE LA MODÉLISATION

Les pièges courants entourant le processus de modélisation (suite) :

- **Élaboration des modèles possibles**
  - Surajustement
  - Ne pas entraîner ou tester correctement les modèles possibles examinés
  - Ne pas faire de régression linéaire simple à utiliser comme base de référence
- **Finalisation du modèle**
  - Ne pas reconstruire le modèle final de façon optimale en utilisant toutes les données utiles
  - Choisir de manière incorrecte le modèle final sans tenir compte de certaines contraintes de mise en œuvre
- **Mise en œuvre et surveillance**
  - Erreurs dans le processus de mise en œuvre : flux d'entrée de données, codages de variables, erreurs d'algorithme
  - Ne pas surveiller le rendement du modèle

DISTRIBUTIONS

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Quelles questions pouvez-vous poser pour vous aider à choisir une distribution de modèle pour un attribut des données?

Quelles sont les fonctions de densité courantes?

Quelle est la moyenne et quelle est la variance de certaines fonctions de densité courantes?

Quand devons-nous utiliser des distributions à plusieurs variables?

DONNÉES ET DISTRIBUTIONS

Si un attribut de données peut être caractérisé par une distribution, poser ces **quatre questions fondamentales** :

1. La variable ne peut-elle prendre que des valeurs **discrètes**?
   - Le fait que la déclaration d'un contribuable soit vérifiée ou non est une variable *discrète*, mais le montant corrigé découlant de la vérification est une variable *continue*
2. La distribution des données est-elle **symétrique**?
   - Si ce n'est pas le cas, dans quelle **direction** se situe l'asymétrie?
   - Les valeurs aberrantes de **droite** et de **gauche** sont-elles également probables?

DONNÉES ET DISTRIBUTIONS

3. La variable a-t-elle des limites **supérieures** et **inférieures** théoriques?
   - Certains éléments comme l'âge ou la taille ne peuvent être inférieurs à zéro
   - Certains éléments, comme les marges d'exploitation, ne peuvent pas dépasser une certaine valeur (100 % dans ce cas)
4. Quelle est la probabilité d'avoir des **valeurs extrêmes** dans la distribution?
   - Pour certaines données, les valeurs extrêmes sont peu fréquentes alors que pour d'autres, elles sont plus fréquentes

Comment faut-il adapter ces questions lorsqu'il s'agit de **distributions à plusieurs variables**?

DISTRIBUTIONS FONDAMENTALES

Les distributions empiriques sont souvent lissées par des **distributions paramétriques**, définies *par* une **fonction de densité** et par un ensemble de paramètres qui doivent être tirés des données.

Les distributions de base de l'analyse des données sont les suivantes :

- La **distribution uniforme** $U(a, b)$ sur l'intervalle $[a, b]$ ou $U(x_1, \dots, x_n)$ sur l'ensemble discret $\{x_1, \dots, x_n\}$; vraisemblablement la distribution la plus simple
- La distribution normale $N(\mu, \sigma^2)$ sur la droite réelle $\mathbb{R}$; probablement la distribution la plus fréquemment utilisée (mais pas toujours correctement)
- Une grande variété de distributions **spéciales** qui sont utilisées dans des applications allant de la modélisation des consommateurs et de la finance à la recherche opérationnelle (**de Poisson**, **exponentielle**, **log-normale**, **binomiale**, etc.)

ESPÉRANCE MATHÉMATIQUE ET MOMENTS

Compte tenu d'une fonction de densité f et d'une fonction $g(X)$,
l'**espérance mathématique** $\mathrm{E}_f(g(X))$ **de** g **en fonction de** f est la **moyenne pondérée**

$$\mathrm{E}_f\big(g(X)\big) = \int_\Omega g(X)f(X)\,dX, \text{ où } \Omega = \mathrm{dom}(f).$$

Les **moments** d'une distribution sont définis par

$$m_i = \mathrm{E}\big(X^i\big), \text{ pour } i = 0, \ldots,$$

Notez que par définition, $m_0 = 1$. La **moyenne** et la **variance** de la distribution sont respectivement données par $m_1 = \mathrm{E}(X)$ et par $m_2 - m_1^2 = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$.

Distribution	Fonction de densité $f(x)$	Moyenne	Variance	Notes
uniforme $U(a,b)$	$\frac{1}{b-a}$ pour $a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	La plupart des langages fournissent des générateurs de valeurs aléatoires pour $U(a,b)$; sert à générer des v.a. avec d'autres distributions
de Gauss $N(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ pour $x \in \mathbb{R}$	μ	σ^2	Si $X \sim N(\mu,\sigma^2)$, alors $\frac{X-\mu}{\sigma} \sim N(0,1)$ (et *vice-versa*); utilisée très souvent
de Poisson $P(\lambda), \lambda \geq 0$	$\frac{\lambda^x}{x!} e^{-\lambda}$ pour $x = 0,1,2,\ldots$	λ	λ	Estime le nombre d'événements qui se produisent dans un intervalle de temps continu (nombre d'appels reçus dans des intervalles d'une heure)
binomiale $\mathcal{B}(N,p), N \in \mathbb{N}$, $p \in [0,1]$	$\binom{N}{x} p^x (1-p)^{N-x}$ pour $x = 0,1,\ldots,N$	Np	$Np(1-p)$	Décrit la probabilité exacte de x succès dans N essais indépendants si la probabilité de succès d'un seul essai est p (nombre de faces dans N tirages à pile ou face)
log-normale $\Lambda(\mu,\sigma^2)$	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2}$ pour $x > 0$	$e^{(\mu+\sigma^2/2)}$	$e^{(2\mu+\sigma^2)}\left[e^{\sigma^2}-1\right]$	Si $\ln X \sim N(\mu,\sigma^2)$, alors $X \sim \Lambda(\mu,\sigma^2)$ (et *vice-versa*); désaxée vers la droite

DISTRIBUTIONS À PLUSIEURS VARIABLES

Les distributions univariées sont des outils de modélisation utiles, surtout lorsque les variables considérées sont **indépendantes**.

Dans la pratique, ce n'est généralement pas le cas. Une **distribution à plusieurs variables** $P(X_1, \dots, X_n)$ donne la probabilité que chaque valeur $X_1, \dots, X_n$ se situe dans une aire de distribution donnée. La **distribution normale à plusieurs variables** $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ comprend une fonction de densité

$$f(x_1, \dots, x_n) := f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

où $\boldsymbol{\mu}$ est le vecteur moyen et $\boldsymbol{\Sigma}$ la matrice de covariance.

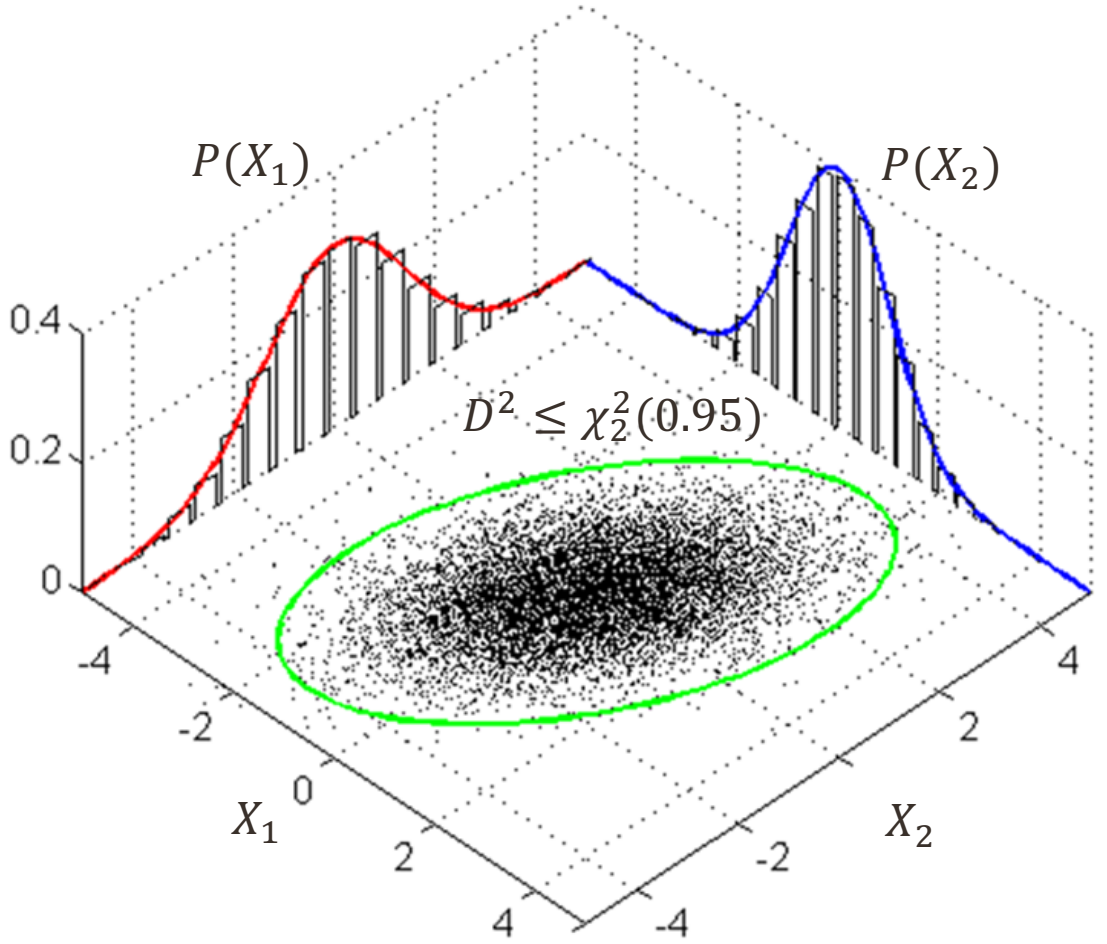
[image tirée de Wikipédia]

DISTRIBUTIONS À PLUSIEURS VARIABLES

Si $\boldsymbol{\Sigma}$ est définie positive, la distribution normale à plusieurs variables est **non décomposable**.

$D = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$ est la distance de **Mahalanobis**.

Pour générer un échantillon $\boldsymbol{x}$ à partir de $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, prenons $\boldsymbol{z} \sim N(\mathbf{0}, \mathbf{I})$ et définissons $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{z}$, où $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}$ est la décomposition de Cholesky.



EXERCICES

Écrivez en langage R ou en Python un script qui vous permet d'extraire des échantillons « aléatoires » des différentes distributions discutées dans cette section.

THÉORÈME DE LA LIMITE CENTRALE

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce que le théorème de la limite centrale?

Dans quelles conditions le théorème de la limite centrale est-il utile?

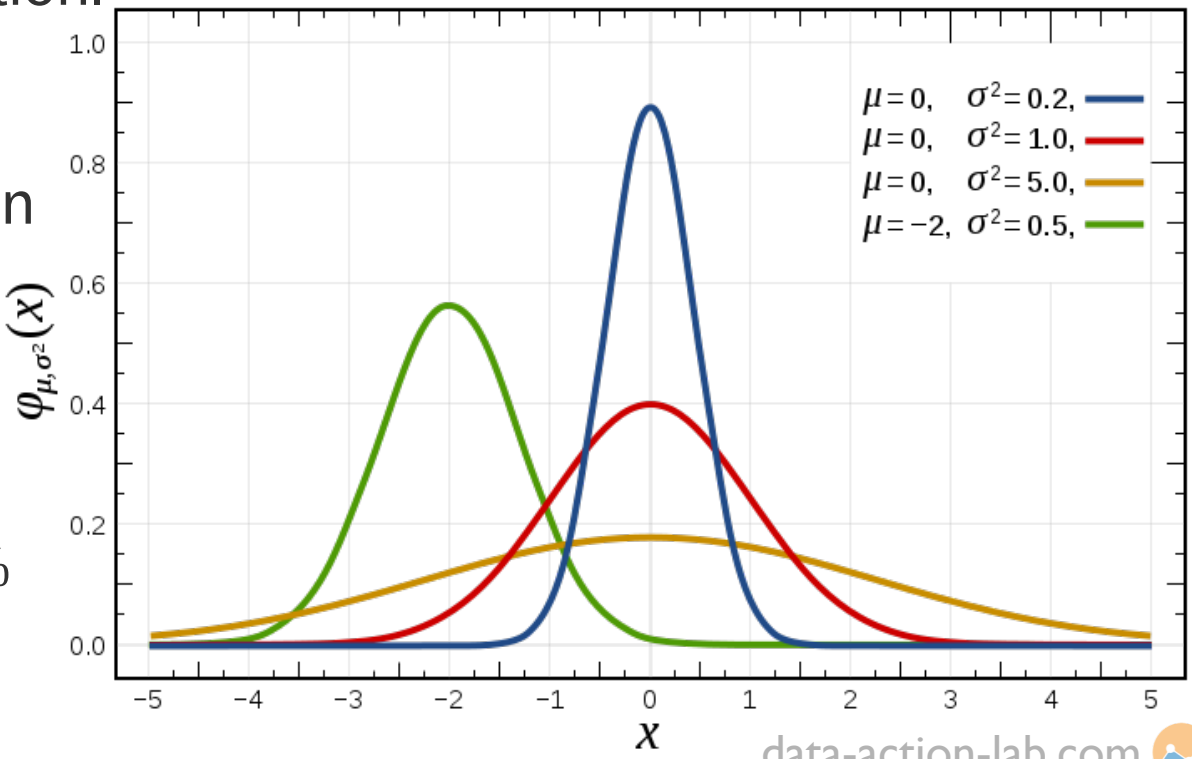
[image tirée de Wikipédia]

DISTRIBUTION NORMALE

L'équation $N(\mu, \sigma^2)$ est **entièrement caractérisée** par la moyenne μ et par l'écart-type σ, ce qui réduit les besoins d'estimation.

La probabilité qu'une valeur soit extraite peut être obtenue si nous savons combien de multiples de σ la séparent de μ

- à l'intérieur de σ par rapport à μ : $\approx 68\%$
- à l'intérieur de 2σ par rapport à μ : $\approx 95\%$
- à l'intérieur de 3σ par rapport à μ : $\approx 99.7\%$



DISTRIBUTION NORMALE

La distribution normale est la mieux adaptée aux données répondant aux exigences minimales suivantes :

- Forte tendance des données à prendre une valeur centrale
- Les écarts positifs et négatifs par rapport à cette valeur centrale sont également probables
- La fréquence des écarts diminue rapidement à mesure que l'on s'éloigne de la valeur centrale

La symétrie des écarts conduit à une **asymétrie** égale à zéro; une faible probabilité d'écarts importants par rapport à la valeur centrale n'entraîne aucun **aplatissement**.

Son omniprésence dans les affaires humaines est liée au **théorème de la limite centrale**.

THÉORÈME DE LA LIMITE CENTRALE

Soit $x_1, x_2, \ldots, x_n$ un **échantillon aléatoire** de toute (?) distribution avec la moyenne μ et la variance σ^2. Si les observations de l'échantillon sont **indépendantes** les unes des autres, alors la distribution de la moyenne

$$w = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

est **à peu près normale** (lorsque $n \to \infty$) avec une moyenne et une variance étant définies par

$$\mu_w = \frac{1}{n} E(x_1 + \cdots + x_n) = \mu, \ \sigma_w^2 = \frac{1}{n^2} E(x_1 + \cdots + x_n - n\mu)^2 = \frac{1}{n}\sigma^2.$$

Le théorème de la limite centrale joue un rôle important au regard de la prévalence de la distribution normale dans les affaires humaines.

TAILLE DE L'ÉCHANTILLON

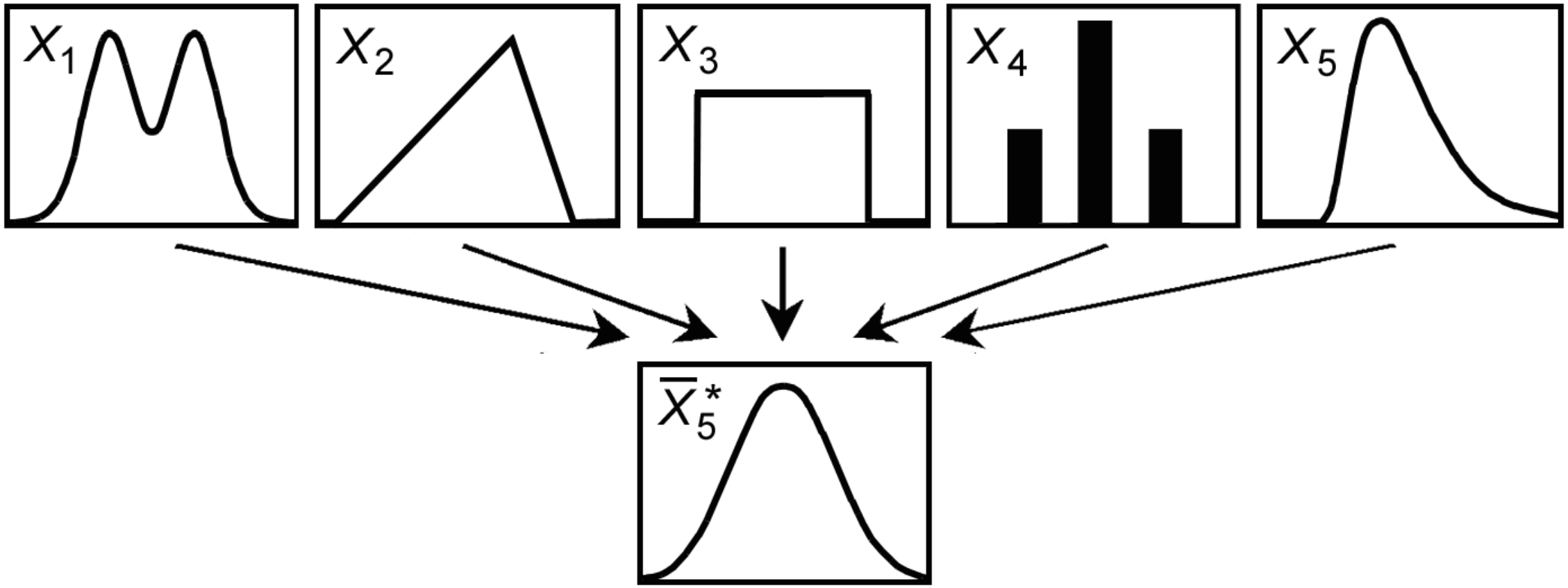
Si la population sous-jacente est **normale**, la distribution de la moyenne de l'échantillon est également **normale**, quelle que soit la taille de l'échantillon n.

Si la population sous-jacente est **approximativement symétrique**, la distribution de la moyenne de l'échantillon est **approximativement normale** pour de petits échantillons n.

Si l'échantillon de la population est **asymétrique** (ou **disparate**), la taille de l'échantillon doit généralement atteindre au moins 30 valeurs avant que la distribution de la moyenne de l'échantillon devienne **approximativement normale**.

[https://www.value-at-risk.net/central-limit-theorem]

THÉORÈME DE LA LIMITE CENTRALE EN ACTION



EXERCICES

Un grand monte-charge peut transporter un maximum de 9 800 livres. Supposons qu'un chargement contenant 49 boîtes doive être transporté. Par expérience, le poids des boîtes suit une distribution avec une moyenne $\mu = 205$ livres et un écart-type $\sigma = 15$ livres.

En programmant en R ou en Python, estimez la probabilité que les 49 boîtes puissent être chargées et transportées en toute sécurité dans le monte-charge.

ESTIMATION

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce que l'estimation, au sens statistique du terme?

À quoi sert l'estimation?

Qu'est-ce que le biais, au sens statistique du terme?

ESTIMATION

L'un des objectifs des statistiques est d'essayer de **comprendre une grande population** sur la base des informations disponibles dans un petit échantillon.

En particulier, nous nous intéressons aux **paramètres** de population, qui sont estimés à l'aide de statistiques d'échantillonnage appropriées.

Par exemple, nous pouvons utiliser la **moyenne de l'échantillon** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ comme estimation de la **moyenne réelle de la population** μ.

ESTIMATION

L'**estimateur** est une variable aléatoire; l'**estimation** est un nombre.

Comme autre exemple, l'**écart-type de l'échantillon** S est un estimateur de l'**écart-type de la population** réelle σ et de la valeur calculée de S

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

est une estimation de σ.

Un estimateur W de ω est sans biais si $\mathrm{E}(W) = \omega$.

CONCEPTS MATHÉMATIQUES DE BASE

Soit les **variables aléatoires** $X_1, \dots, X_n, b_1, \dots, b_n \in \mathbb{R}$ et E, V, Cov les opérateurs relatifs à l'**espérance mathématique**, à la **variance** et à la **covariance**, respectivement, à savoir :

- $\mathrm{E}(X_i) = \mu_i$
- $\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$
- $\mathrm{V}(X_i) = \mathrm{Cov}(X_i, X_i) = \mathrm{E}(X_i^2) - \mathrm{E}^2(X_i) = \mathrm{E}(X_i^2) - \mu_i^2 = \sigma_i^2$ et

$$\mathrm{E}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i \mathrm{E}(X_i) = \sum_{i=1}^{n} b_i \mu_i$$

$$\mathrm{V}\left(\sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} b_i^2 \mathrm{V}(X_i) + \sum_{i \neq j} b_i b_j \,\mathrm{Cov}(X_i, X_j)$$

CONCEPTS MATHÉMATIQUES DE BASE

Le **biais** d'une estimation est la moyenne de l'erreur dans l'estimation si l'étude est répétée indépendamment plusieurs fois dans les mêmes conditions.

La **variation** d'une estimation est la mesure dans laquelle l'estimation varierait par rapport à sa valeur moyenne dans le scénario idéal décrit ci-dessus.

La **variance de la population** d'une estimation est une mesure de l'erreur qui incorpore les deux éléments :

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{V}(\hat{\beta}) + \mathrm{Biais}^2(\hat{\beta}),$$

où $\hat{\beta}$ est un estimateur de β.

CONCEPTS MATHÉMATIQUES DE BASE

Si l'estimation $\hat{\beta}$ est sans biais, $\mathrm{E}(\hat{\beta} - \beta) = 0$ alors un **intervalle de confiance** approximatif à 95 % (IC à 95 %) pour β est donné approximativement par

$$\hat{\beta} \pm 2\sqrt{\widehat{\mathrm{V}}(\hat{\beta})},$$

Où $\widehat{\mathrm{V}}(\hat{\beta})$ est une estimation **spécifique au plan d'échantillonnage** de $\mathrm{V}(\hat{\beta})$.

Mais qu'est-ce qu'un IC à 95 % exactement?

EXERCICE

Le temps total de fabrication d'un composant particulier est connu comme suivant une distribution normale pour laquelle la moyenne μ et la variance σ^2 ne sont pas connues. Dans une expérience, 10 composants sont fabriqués; le temps dans l'échantillon est donné comme suit :

	1	2	3	4	5	6	7	8	9	10
Temps	63,8	60,5	65,3	65,7	61,9	68,2	68,1	64,8	65,8	65,4

Quelles sont les meilleures estimations pour μ et σ^2? Fournir un IC à 95 % pour μ.

THÉORÈME DE BAYES

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce qu'une probabilité conditionnelle et quand est-elle utile?

Quelles sont les règles mathématiques qui régissent la probabilité?

Qu'est-ce que le théorème de Bayes et quand est-il utile?

PROBABILITÉS CONDITIONNELLES

Nous nous intéressons souvent à la probabilité qu'un événement se produise **en fonction de l'occurrence d'un autre événement**.

Voici quelques exemples :

- La probabilité qu'un train arrive à l'heure étant donné qu'il est parti à l'heure
- La probabilité qu'un PC tombe en panne compte tenu du système d'exploitation installé
- La probabilité qu'un bit transmis sur un canal soit vu à la réception comme étant un 1 étant donné que le bit transmis était un 1
- La probabilité qu'un site Web soit visité compte tenu du nombre d'hyperliens qui y mènent

Les questions de ce type sont traitées à l'aide de la probabilité conditionnelle.

PROBABILITÉS CONDITIONNELLES

La **probabilité conditionnelle** est la probabilité qu'un événement se produise en fonction de l'occurrence d'un autre événement.

La probabilité conditionnelle de A étant donné B, $P(A|B)$ est définie par

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

La probabilité que deux événements A et B se produisent est obtenue en appliquant la règle de multiplication

$$P(A \cap B) = P(B)\,P(A|B) = P(A)\,P(B|A)$$

PROBABILITÉS CONDITIONNELLES

Exemple (un classique) : une famille a deux enfants (non jumeaux). Quelle est la probabilité que le plus jeune enfant soit une fille étant donné qu'au moins un des enfants est une fille? Supposons que les garçons et les filles ont autant de chances de naître.

Solution : Soient A et B les événements que le plus jeune enfant est une fille et qu'au moins un enfant est une fille, respectivement :

$$A = \{GG, BG\}, \qquad B = \{GG, BG, GB\}$$

Alors $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{3}$ (et non pas ½, comme on pourrait le supposer naïvement).

RÈGLES DE PROBABILITÉ

Soit I l'information de base pertinente; X, Y, Y_k sont les propositions et $-X$ est la proposition que X est fausse.

La **plausibilité** de X étant donné I est indiquée par $P(X|I)$ allant de 0 (faux) à 1 (vrai).

Règle de la somme : $P(X|I) + P(-X|I) = 1$

Règle du produit : $P(X, Y|I) = P(X|Y, I) \times P(Y|I)$

Théorème de Bayes : $P(X|Y, I) \times P(Y|I) = P(Y|X, I) \times P(X|I)$

Règle de marginalisation : $P(X|I) = \sum P(X, Y_k|I)$, où $\{Y_k\}$ sont des valeurs exhaustives disjointes

THÉORÈME DE BAYES

La règle de la somme et la règle du produit sont les **règles de base en probabilité**.

Le théorème de Bayes et la **règle de marginalisation** sont de simples corollaires de ces règles de base.

Le théorème de Bayes est parfois écrit sous une forme légèrement différente

$$P(X|Y,I) = \frac{P(Y|X,I) \times P(X|I)}{P(Y|I)}$$

THÉORÈME DE BAYES

Mise en place : Supposons qu'une expérience a été menée pour déterminer le degré de validité d'une hypothèse particulière et que des données expérimentales ont été recueillies.

Question d'analyse des données centrales : Étant donné tout ce que l'on savait *avant* l'expérience, les données recueillies appuient-elles (ou invalident-elles) l'hypothèse?

Tout au long de l'expérience, X indique que l'hypothèse en question est vraie, Y indique que l'expérience a produit les données réelles observées, et I indique (comme toujours) l'information de base pertinente.

THÉORÈME DE BAYES

Question d'analyse des données centrales (reprise) :

Quelle est la valeur de $P(\text{l'hypothèse est vraie} \mid \text{données observées}, I)$?

Problème : Cette valeur est presque toujours impossible à calculer directement.

Solution : À l'aide du théorème de Bayes,

$$P(\text{hypothèse} \mid \text{données}, I) = \frac{P(\text{données} \mid \text{hypothèse}, I) \times P(\text{hypothèse} \mid I)}{P(\text{données} \mid I)},$$

il se peut que les termes à droite soient plus faciles à calculer.

THÉORÈME DE BAYES

En termes familiers, la probabilité

- $P(\text{hypothèse} \mid I)$ que l'hypothèse soit vraie avant l'expérience est la **probabilité a priori**
- $P(\text{hypothèse} \mid \text{données}, I)$ que l'hypothèse soit vraie une fois que les données expérimentales sont prises en compte est la **probabilité a posteriori**
- $P(\text{données} \mid \text{hypothèse}, I)$ que les données expérimentales soient observées en supposant que l'hypothèse est vraie est la **vraisemblance**
- $P(\text{données} \mid I)$ que les données expérimentales soient observées indépendamment de toute hypothèse est la **donnée probante**

Une hypothèse donnée comprend un modèle (potentiellement implicite) qui peut être utilisé pour calculer ou établir approximativement la **vraisemblance**.

THÉORÈME DE BAYES

La détermination de la **probabilité a priori** est une source de controverse considérable

- Des estimations prudentes (probabilité a priori non instructive) conduisent souvent à des résultats raisonnables
- En l'absence d'information, choisir la probabilité a priori d'entropie maximale

La **preuve** est plus difficile à calculer sur des bases théoriques. Pour évaluer la probabilité de l'observation des données, il faut un accès à un modèle dans le cadre de I. Soit

- ce modèle était bon, donc il n'est pas nécessaire de poser une nouvelle hypothèse
- ce modèle était mauvais, donc nous ne pouvons pas faire confiance à nos calculs

THÉORÈME DE BAYES

Heureusement, les données probantes sont rarement requises sur les problèmes d'estimation des paramètres (bien qu'elles soient essentielles pour le choix du modèle) :

- avant l'expérience, il existe de nombreuses hypothèses concurrentes
- les données antérieures et les probabilités seront différentes, mais pas les données probantes
- les données probantes ne sont pas nécessaires pour différencier les différentes hypothèses

Le théorème de Bayes est souvent présenté comme suit :

$$P(\text{hypothèse} \mid \text{données}, I) \propto P(\text{données} \mid \text{hypothèse}, I) \times P(\text{hypothèse} \mid I)$$

ou simplement comme postérieur $\propto$ probabilité×antérieur, c'est-à-dire que les **croyances devraient être mises à jour en présence de nouveaux renseignements**.

EXERCICE

Supposons qu'un test de diagnostic d'une maladie particulière ait un taux de réussite très élevé.
Si un patient

- est atteint de la maladie, le test donne correctement un résultat « positif » avec une probabilité de 0,99;
- n'est pas atteint de la maladie, le test donne correctement un résultat « négatif » avec une probabilité de 0,95.

Supposons en outre que seulement 0,1 % de la population est atteinte de la maladie. Quelle est la probabilité qu'un patient dont le test est positif ne soit pas atteint de la maladie?

ALGÈBRE MATRICIELLE

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

Neo : Qu'est-ce que c'est que la Matrice?
Trinity : La réponse est là, quelque part. Elle te cherche aussi, et elle te trouvera, si tu veux qu'elle te trouve.

(*Matrix*, les sœurs Wachowski)

OBJECTIFS D'APPRENTISSAGE

Quel est le principal objet mathématique utilisé dans l'algèbre linéaire?

Pourquoi les matrices sont-elles pertinentes dans la science et l'analyse des données?

Quelles sont certaines des opérations matricielles?

ALGÈBRE LINÉAIRE

Une **matrice** est un outil mathématique important qui permet d'organiser facilement l'information, de simplifier la notation et de faciliter l'application d'algorithmes aux données.

La plupart des outils statistiques nécessitent des données **rectangulaires** :

- chaque colonne contient une **variable** (caractéristique, champ, attribut)

  - indicateur, cible, question dans une enquête, etc.

- chaque ligne contient une **observation** (cas, unité, article)

  - pays, répondant à l'enquête, sujet d'une expérience, etc.

- chaque cellule contient une **valeur** (mesure) pour une variable et une observation particulières

  - PIB par habitant pour le Canada, réponse à une question précise, âge, etc.

OPÉRATIONS MATRICIELLES

Une matrice est une grille rectangulaire d'**éléments** disposés en **lignes** et en **colonnes**.

Les matrices sont souvent utilisées en algèbre pour résoudre des valeurs inconnues dans les équations linéaires, ainsi qu'en géométrie.

Addition de matrice : les matrices peuvent être additionnées (« par **élément** ») tant que leurs **dimensions** sont les mêmes (c'est-à-dire que les deux matrices ont le même nombre de lignes et de colonnes), comme suit :

$$\begin{bmatrix} 3 & -2 \\ 4 & 1 \end{bmatrix} + \begin{bmatrix} 4 & 6 \\ 8 & 3 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 12 & 4 \end{bmatrix}$$

OPÉRATIONS MATRICIELLES

Multiplier une matrice par un scalaire : une matrice de n'importe quelle dimension peut être multipliée par un scalaire en multipliant chaque élément par le scalaire.

$$-1\times\begin{bmatrix}2 & 1\\ 3 & -5\\ 4 & 6\end{bmatrix}=\begin{bmatrix}-2 & -1\\ -3 & 5\\ -4 & -6\end{bmatrix}$$

Multiplier les matrices : deux matrices Aet B peuvent être multipliées si leurs dimensions sont **compatibles** (c.-à-d.,$\dim(A)=n\times p$ et $\dim(B)=p\times k$). Le produit $C=AB$ est tel que $\dim(C)=n\times k$.

OPÉRATIONS MATRICIELLES

L'élément de la i^{e} ligne et de la j^{e} colonne du produit $C = AB$ est donné par

$$c_{i,j} = a_{i,1}b_{1,j} + \cdots + a_{i,p}b_{p,j}$$

Pour les matrices 2×2, cela se réduit à

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \times \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}$$

Par exemple,

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 3 & 0 & 5 \end{bmatrix}, B = \begin{bmatrix} -2 \\ 3 \\ 0 \end{bmatrix} \Rightarrow AB = \begin{bmatrix} 4\times(-2) + 2\times3 + 1\times0 \\ 3\times(-2) + 0\times3 + 5\times0 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \end{bmatrix}$$

OPÉRATIONS MATRICIELLES

Transposer une matrice : l'échange des lignes et des colonnes d'une matrice s'appelle la **transposition** de la matrice – cela est indiqué par un « T » :

$$\begin{bmatrix} 6 & 0 & -2 \\ 2 & 1 & 3 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} 6 & 2 \\ 0 & 1 \\ -2 & 3 \end{bmatrix}$$

Lorsqu'elle est appliquée à une base de données, la transposition a pour effet d'intervertir les rôles des cas et des observations.

Pour les matrices carrées d'ordre n (c.-à-d. $\dim = n \times n$), il existe deux matrices spéciales : la matrice **nulle** 0_n (constituée uniquement de zéros), et la **matrice identité** I_n (les entrées diagonales sont des 1, toutes les autres sont des 0).

OPÉRATIONS MATRICIELLES

Dans le cas des matrices carrées, deux quantités finissent souvent par jouer un rôle fondamental : la **trace** et le **déterminant**.

La **trace** est la somme des éléments de la diagonale principale :

$$\mathrm{tr}\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = a_{11} + a_{22} + \cdots + a_{nn}$$

OPÉRATIONS MATRICIELLES

Le **déterminant** peut être calculé de façon récursive. Indiquons A par $n \times n$.

1. Pour $n = 1$, $\det[\, a \,] = a$;

2. Pour $n = 2$, $\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$

3. Pour une valeur générale n, posons $D_{i,j}(A)$ comme étant le déterminant de la matrice $(n-1)\times(n-1)$ obtenu en supprimant la i^{e} colonne et la j^{e} colonne de A. Le **développement de Laplace** de $\det A$ le long de la première colonne est

$$(-1)^{1+1}a_{11}\, D_{1,1}(A) + (-1)^{2+1}a_{21}\, D_{2,1}(A) + \cdots + (-1)^{j+1}a_{j1}\, D_{j,1}(A) + \cdots + (-1)^{n+1}a_{n1}\, D_{n,1}(A).$$

OPÉRATIONS MATRICIELLES

Le déterminant peut être développé le long de n'importe quelle ligne/colonne sans changer sa valeur :

$$\det\begin{bmatrix}1 & 0 & -2\\ 4 & 2 & 6\\ 10 & 8 & 0\end{bmatrix} = 1\times\det\begin{bmatrix}-2 & 6\\ 8 & 0\end{bmatrix} - 0\times\det\begin{bmatrix}4 & 6\\ 10 & 0\end{bmatrix} + (-2)\times\det\begin{bmatrix}4 & -2\\ 10 & 8\end{bmatrix} = -152$$

ou

$$\det\begin{bmatrix}1 & 0 & -2\\ 4 & -2 & 6\\ 10 & 8 & 0\end{bmatrix} = -0\times\det\begin{bmatrix}4 & 6\\ 10 & 0\end{bmatrix} + (-2)\times\det\begin{bmatrix}1 & -2\\ 10 & 0\end{bmatrix} - 8\times\det\begin{bmatrix}1 & -2\\ 4 & 6\end{bmatrix} = -152$$

et

$$\mathrm{tr}\begin{bmatrix}1 & 0 & -2\\ 4 & -2 & 6\\ 10 & 8 & 0\end{bmatrix} = 1 + (-2) + 0$$

OPÉRATIONS MATRICIELLES

Le déterminant est lié à l'**inverse** d'une matrice.

En arithmétique numérique, chaque nombre $a \neq 0$ a un inverse b indiqué par a^{-1} ou $1/a$ de sorte que $ba = ab = 1$. De même, une matrice carrée A peut avoir un inverse $B = A^{-1}$ où $AB = BA = I_n$.

Divers :

- Les matrices non carrées ne possèdent pas d'inverses.
- Les matrices carrées n'ont pas toutes un inverse (seulement celles avec $\det(A) \neq 0$).
- Une matrice qui a un inverse est dite **non singulière**.

OPÉRATIONS MATRICIELLES

Si $ad - bc \neq 0$, alors la matrice $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ a un inverse (unique) :

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Pour $n > 2$, d'autres méthodes de calcul existent, comme l'**élimination de Gauss :** si une séquence d'opérations de ligne $(yR_j + xR_i \rightarrow R_j, R_j \leftrightarrow R_i)$ appliquée à une matrice carrée A la réduit à une matrice identité I du même ordre, alors la même séquence d'opérations appliquée à I la réduit à A^{-1}.

OPÉRATIONS MATRICIELLES

Si nous ne pouvons pas réduire A à I, alors, A^{-1}n'existe pas. Ceci deviendra manifeste avec l'apparition d'une ligne de zéros. Il n'y a pas de méthode unique pour passer de A à I et c'est l'expérience qui permet de choisir la méthode optimale.

Il est plus efficace d'effectuer les deux réductions simultanément;

$$[A|I] = \left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 1 & 4 & 3 & 0 & 1 & 0 \\ 2 & 7 & 7 & 0 & 0 & 1 \end{array}\right] \overset{\substack{R_2-R_1\to R_2 \\ R_3-2R_1\to R_3}}{\Longrightarrow} \left[\begin{array}{ccc|ccc} 1 & 3 & 3 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 1 & -2 & 0 & 1 \end{array}\right]$$

$$\overset{\substack{R_1-3R_2\to R_1 \\ R_3-R_2\to R_3}}{\Longrightarrow} \left[\begin{array}{ccc|ccc} 1 & 0 & 3 & 4 & -3 & 0 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right] \overset{R_1-3R_3\to R_1}{\Longrightarrow} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 7 & 0 & -3 \\ 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right] = [I|A^{-1}]$$

EXERCICES

Dans R, établissez des matrices carrées 3×3 A, B, C et calculez les éléments suivants :

- $A + B, BC, CB, A^{\mathrm{T}}, CA^{\mathrm{T}}$
- $\mathrm{tr}(A), \mathrm{tr}(3A), \mathrm{tr}(C), \mathrm{tr}(-C), \mathrm{tr}(3A - C)$
- $\det(A), \det(A^{\mathrm{T}}), \det(B), \det(C), \det(BC)$
- A^{-1}, B^{-1}, C^{-1}, si les déterminants respectifs sont $\neq 0$
- $\det(A^{-1}), \det(B^{-1}), \det(C^{-1})$, si les matrices respectives sont inversibles

Pouvez-vous déduire des règles à partir de ces calculs?

VALEURS PROPRES ET VECTEURS PROPRES

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce qu'une valeur propre?

Qu'est-ce qu'un vecteur propre?

Qu'est-ce qu'un cas d'utilisation de ces concepts mathématiques?

VECTEURS PROPRES ET VALEURS PROPRES

Un **vecteur propre** d'une matrice A est un vecteur $\boldsymbol{v} \neq \mathbf{0}$ de sorte que, pour certains scalaires λ, $A\boldsymbol{v} = \lambda\boldsymbol{v}$.

La valeur λs'appelle une **valeur propre** de A associée à $\boldsymbol{v}$.

Les valeurs propres d'une matrice $n \times n$ A répondent à $\det(A - \lambda I_n) = 0$. Le côté gauche est un polynôme dans λ; il est appelé **polynôme caractéristique** de A, représenté par $p_A(\lambda)$.

Pour trouver les valeurs propres de A, nous trouvons les racines de $p_A(\lambda)$.

EXEMPLE

Posons $A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$. Alors, $p_A(\lambda) = \det(A - \lambda I) = (\lambda - 3)(\lambda + 2)$. Ainsi, $\lambda_1 = 3$ et $\lambda_2 = -2$ sont les valeurs propres de A.

Pour trouver les vecteurs propres correspondant à une valeur propre λ, nous résolvons le système d'équations linéaires donné par $(A - \lambda I)\boldsymbol{v} = 0$.

Calculons les vecteurs propres correspondant à $\lambda_1 = 3$, en résolvant

$$(A - 3I)\boldsymbol{v} = \begin{bmatrix} 2-3 & -4 \\ -1 & -1-3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

EXEMPLE

Ceci donne les équations suivantes :

$$-v_1 - 4v_2 = 0 \ , \qquad -v_1 - 4v_2 = 0$$

Si nous laissons $v_2 = t$, alors, $v_1 = -4t$; ainsi, tous les vecteurs propres correspondant à $\lambda_1 = 3$ sont des multiples de $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$.

Un calcul similaire montre que tous les vecteurs propres correspondant à $\lambda_2 = -2$ sont des multiples de $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

DÉCOMPOSITION EN ÉLÉMENTS PROPRES

Si une matrice A $n \times n$ possède des vecteurs propres n linéairement indépendants, alors A peut être **décomposé** de la manière suivante :

$$A = B\Lambda B^{-1},$$

où Λ est une matrice diagonale dont les entrées diagonales sont les valeurs propres de A et les colonnes de B sont les vecteurs propres correspondants de A.

EXEMPLE

Nous avons vu que les valeurs propres de $A = \begin{bmatrix} 2 & -4 \\ -1 & -1 \end{bmatrix}$ sont $\lambda_1 = 3$ et$\lambda_2 = -2$, et que les vecteurs propres correspondants sont $\begin{bmatrix} -4 \\ 1 \end{bmatrix}$ et $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Ainsi, $\Lambda = \begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}$, $B = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}$, et

$$A = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}\frac{1}{-4\times1 - 1\times1}\begin{bmatrix} 1 & -1 \\ -1 & -4 \end{bmatrix}$$

$$= \begin{bmatrix} -4 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 & 0 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} -1/5 & 1/5 \\ 1/5 & 4/5 \end{bmatrix}$$

EXERCICES

Calculez la décomposition en éléments propres des matrices A, B, C que vous avez établies dans le module précédent.

RÉGRESSION

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce que la modélisation par régression?

Pouvez-vous nommer certains types de modélisation par régression?

Dans quels cas la modélisation par régression est-elle utile?

MODÉLISATION PAR RÉGRESSION

Les méthodes de modélisation de données les plus courantes sont les régressions, tant **linéaires** que **logistiques.**

- ~90 % des applications de données réelles finissent par utiliser une régression simple comme modèle final, généralement après une préparation très minutieuse des données, un codage et la création de variables.

Plusieurs raisons expliquent leur utilisation fréquente :

- Généralement faciles à comprendre et à former
- La fonction objectif de l'erreur quadratique moyenne (EQM) a une solution linéaire de forme fermée
- Le système d'équations peut généralement être résolu par inversion matricielle ou manipulation linéaire

MODÉLISATION PAR RÉGRESSION

Structure de données d'une tâche de modélisation générale est représenté par

X_1	X_2	$\cdots$	X_p	Y
x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

Nous tenons compte des variables p indépendantes X_i pour essayer de prédire la variable dépendante Y.

Afin de simplifier l'analyse qui suit, nous présentons la notation matricielle $\boldsymbol{X}[n\times p], \boldsymbol{Y}[n\times 1], \boldsymbol{\beta}[p\times 1]$, où n est le nombre d'observations et p est le nombre de variables indépendantes.

RÉGRESSION LINÉAIRE

L'hypothèse de base de la régression linéaire est que la variable dépendante y peut être **approximée** par une combinaison linéaire des variables indépendantes comme suit :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

où $\beta \in \mathbb{R}^p$ doit être déterminé en fonction de l'ensemble d'apprentissage, et pour lequel

$$\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0, \qquad \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\boldsymbol{X}) = \sigma^2 I.$$

En règle générale, les erreurs sont également supposées être normalement distribuées, c'est-à-dire :

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(0, \sigma^2 I).$$

RÉGRESSION LINÉAIRE

Si $\hat{\beta}_i$ est l'estimation du coefficient β_i réel, le modèle de **régression linéaire** associé aux données est le suivant

$$\widehat{\boldsymbol{Y}}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

Sous forme matricielle, le problème de régression nécessite une solution $\widehat{\boldsymbol{\beta}}$ à **l'équation normale** $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$.

Lorsque la matrice symétrique positive définie $\boldsymbol{X}^T\boldsymbol{X}$ est inversable, le coefficient rajusté est simplement $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$. Notez que $\boldsymbol{X}^T\boldsymbol{X}$ est une matrice $p \times p$, ce qui rend l'inversion relativement « plus facile » à calculer, lorsque n a une valeur élevé.

RÉGRESSION LINÉAIRE GÉNÉRALISÉE

Les modèles linéaires généralisés (MLG) accroissent la portée des modèles statistiques linéaires en acceptant les variables de réponse ayant une distribution conditionnelle **non normale**.

Sauf pour la **structure d'erreur**, un MLG est essentiellement le même que pour un modèle linéaire :

$$Y_i \sim \text{une certaine distribution avec moyenne } \mu_i, \text{ où } g(\mu_i) = x_i^T \beta$$

Ainsi, un MLG compte trois parties :

- une composante **systématique** $x_i^T \beta$
- une composante **aléatoire** - distribution spécifiée pour Y_i
- une fonction de **liaison** g

RÉGRESSION LINÉAIRE GÉNÉRALISÉE

Nous pourrions préciser **n'importe quelle** distribution pour la variable dépendante Y...

- mais les mathématiques du MLG ne fonctionnent bien que pour la **famille exponentielle** de distributions (la plupart des distributions statistiques courantes figurent dans cette famille : normales, binomiales, de Poisson, gamma, etc.).

La régression linéaire est un exemple de MLG :

- composante systématique : $x_i^T \beta$
- composante aléatoire :$Y_i \sim N(\mu_i, \sigma^2)$
- lien : $g(\mu) = \mu$ le lien d'identité

EXEMPLE

Aux premiers stades d'une épidémie, le taux d'apparition de nouveaux cas augmente de façon exponentielle avec le temps.

Si μ_i est le nombre prévu de nouveaux cas par jour t_i, un modèle prenant la forme

$$\mu_i = \gamma \exp(\delta t_i)$$

pourrait convenir. Si nous prenons la valeur log des deux côtés, nous obtenons

$$\log(\mu_i) = \log(\gamma) + \delta t_i = \beta_0 + \beta_1 t_i = (1, t_i)^T (\beta_0, \beta_1).$$

lien — composante systématique

En outre, puisque nous mesurons le nombre de nouveaux cas (un dénombrement), la distribution de **Poisson** pourrait être un choix raisonnable.

composante aléatoire

LES AVANTAGES DU MLG

Nul besoin de transformer Y pour obtenir une distribution normale

Le choix du lien est **distinct** du choix de la composante aléatoire

- souplesse de modélisation accrue

Si le lien produit des **effets additifs**, la variance constante n'est pas requise

Les modèles sont rajustés par estimation de la LM

- propriétés optimales des estimateurs

Les **outils d'inférence** et les **vérifications de modèle** s'appliquent à d'autres MLG.

- Règle de Wald, test du rapport de vraisemblances, somme des carrés des écarts, résidus, intervalles de confiance, etc.

Voir `PROC GENMOD` dans SAS, ou `glm()` dans R

EXERCICE

Une pièce d'automobile est fabriquée par une entreprise une fois par mois, en lots dont la taille varie en fonction de la demande. Les données ci-dessous représentent les observations sur la taille du lot (y) et le nombre d'heures de travail des employés (x) pour dix cycles de production récents.

Ajustez un modèle de régression simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ pour $i \neq j$, et $\mathrm{V}(\varepsilon_i) = \sigma^2$ si les observations sont les suivantes :

$$\boldsymbol{Y} = [73,50,128,170,87,108,135,69,148,132]^{\mathrm{T}},$$

$$\boldsymbol{x} = [30,20,60,80,40,50,60,30,70,60]^{\mathrm{T}}.$$

OPTIMISATION

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

OBJECTIFS D'APPRENTISSAGE

Qu'est-ce que l'optimisation?

Dans quels cas l'optimisation est-elle utile?

Qu'est-ce qu'une fonction coût?

Pourquoi les minima et les maxima sont-ils pertinents pour l'optimisation?

Quelles techniques peuvent être utilisées pour réaliser l'optimisation?

OPTIMISATION

Supposons que nous devions **optimiser** une fonction **coût** f: $\mathbb{R}^n \rightarrow \mathbb{R}$ (économique) (la fonction de vraisemblance maximale de régression linéaire, par exemple).

La recherche d'un maximum pour f équivaut à la recherche d'un minimum pour -f.

L'objectif consiste à trouver les valeurs des paramètres $\boldsymbol{x}$ qui minimisent cette fonction :

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

La fonction coût pourrait être soumise à un certain nombre de contraintes

$$c_i(\boldsymbol{x}) = 0, i = 1, \ldots, m;\ c_j(\boldsymbol{x}) \geq 0, j = 1, \ldots, k;\ \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^n.$$

OPTIMISATION

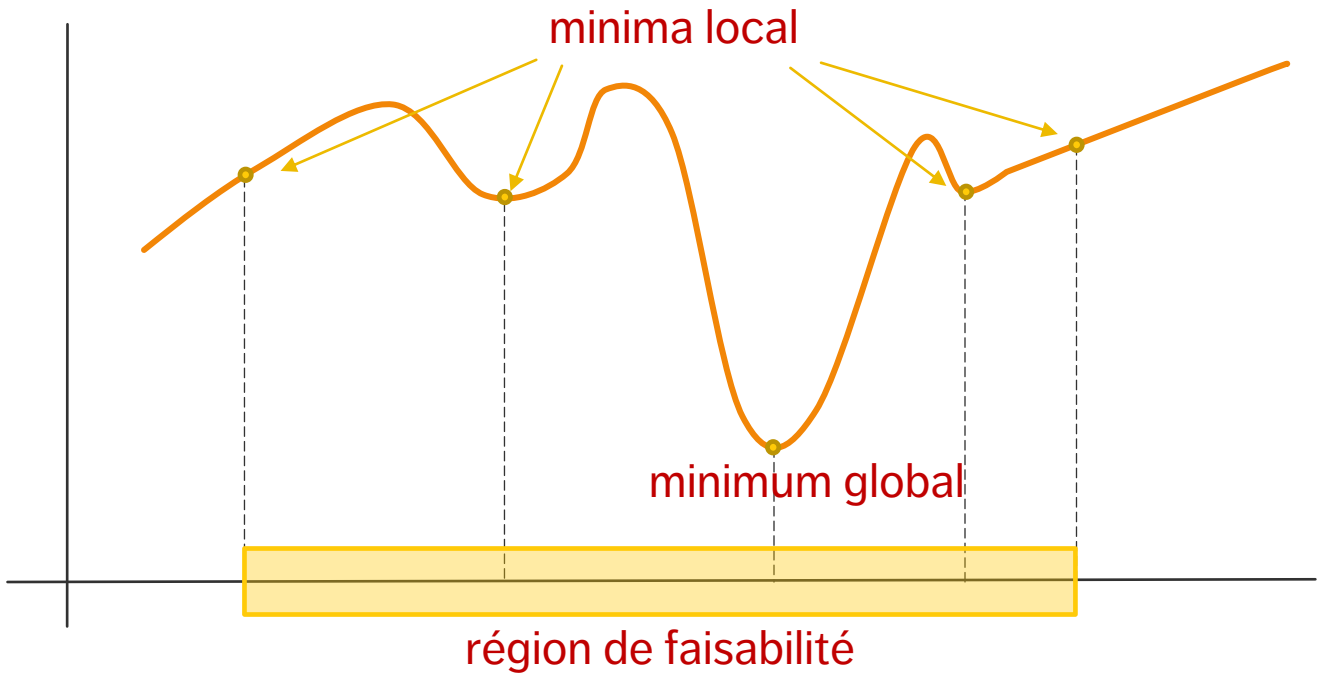
Le problème d'optimisation peut être considéré comme un **problème de décision** qui consiste à trouver le « meilleur » vecteur $\mathbf{x}$ parmi tous les vecteurs possibles dans $\Omega \subseteq \mathbb{R}^n$.

Ce vecteur est appelé le **minimiseur** de f parmi Ω. Il peut y avoir plusieurs minimiseurs, ou aucun.

Si $\Omega = \mathbb{R}^n$, alors nous désignons le problème comme un problème d'optimisation **sans contrainte**.

En général, il ne s'agit pas d'un problème banal (consulter la littérature).

TYPE DE MINIMA



Dans de nombreux cas, l'optimisation est une entreprise **numérique**. Le minima trouvé dépend du **point de départ** de l'algorithme.

MÉTHODE DU RECTANGLE D'OR

La **recherche du rectangle d'or** est une technique permettant de trouver l'extremum (minimum ou maximum) d'une fonction strictement unimodale en rétrécissant successivement la plage de valeurs dans laquelle l'extremum se trouve.

La technique tire son nom du fait que l'algorithme maintient les valeurs de fonction pour les triples de points dont les distances forment un **nombre d'or**.

MÉTHODE DU RECTANGLE D'OR

Posons que $[a, b]$ est l'intervalle de la fourchette courante (c.-à-d. que l'optimiseur se trouve dans $[a, b]$), et que $f(a), f(b)$ a déjà été calculé. Désignons $\varphi = (1 + \sqrt{5})/2$.

1. Soit $c = b - \frac{(b-a)}{\varphi}$, $d = a + \frac{(b-a)}{\varphi}$;
2. Si $f(c), f(d)$ ne sont pas disponibles, calculez-les;
3. Si $f(c) < f(d)$ (pour trouver un minimum – pour trouver un maximum, inversez l'ordre) alors, déplacez les données : $(b, f(b)) \leftarrow (d, f(d))$ and $(d, f(d)) \leftarrow (c, f(c))$ et mettez à jour $c = b - (a - b)/\varphi$ et $f(c)$;
4. Sinon, déplacez les données $(a, f(a)) \leftarrow (c, f(c))$ and $(c, f(c)) \leftarrow (d, f(d))$ et mettez à jour $d = a + (b - a)/\varphi$ et $f(d)$;
5. L'intervalle $[c, d]$ encadre l'optimiseur. Continuez jusqu'à ce que la tolérance soit atteinte.

MÉTHODE DE NEWTON

En calcul, nous apprenons qu'une fonction $f: \Omega \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ qui se comporte suffisamment bien atteint son maximum/minimum soit à un **point critique** (c.-à-d. où $\nabla f = \mathbf{0}$) soit sur la **frontière du domaine** $\partial\Omega$.

Ainsi, pour définir les optimiseurs éventuels, nous devons être capables de résoudre des systèmes généraux de la forme $g(\boldsymbol{x}) = \mathbf{0}$.

La **méthode de Newton** est une méthode puissante permettant de trouver les racines des fonctions.

MÉTHODE DE NEWTON

Pour $n = 1$, l'algorithme figure ci-dessous (cela est assez semblable dans le cas général).

Soit $x = c$ le zéro (inconnu) d'une fonction différentiable f dans un intervalle ouvert contenant c.

1. faire une première approximation x_1 « proche » de c
2. déterminer une nouvelle approximation à l'aide de la formule $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$.
3. Si $|x_2 - x_1|$ est inférieur à la précision souhaitée (qui doit être précisée), x_2 sert d'approximation finale. Sinon, revenir à l'étape 2 et calculer une nouvelle approximation.

EXERCICES

Utilisez la méthode du rectangle d'or et la méthode de Newton pour trouver une racine de

$$f(x) = e^{-x}\sin(x) \text{ et } g(x) = x\ln(x).$$

RÉFÉRENCES

FONDEMENTS STATISTIQUES ET MATHÉMATIQUES

RÉFÉRENCES

Wu, J., Coggeshall, S. *Foundations of Predictive Analytics,* CRC Press, 2012.

Bruce, P., Bruce, A. *Practical Statistics for Data Scientists, 50 Essential Concepts*, O'Reilly, 2017.

Jaynes, E.T. *Probability Theory: the Logic of Science*, Cambridge, 2003.

https://ece.uwaterloo.ca/~dwharder/NumericalAnalysis/11Optimization/newton/

https://www.math.ucdavis.edu/~thomases/W11_16C1_lec_3_11_11.pdf

https://web.as.uky.edu/statistics/users/pbreheny/760/S13/notes/1-24.pdf

https://socialsciences.mcmaster.ca/jfox/Courses/SPIDA/GLMs-notes.pdf

https://onlinecourses.science.psu.edu/stat504/node/216

http://stattrek.com/regression/regression-example.aspx?Tutorial=AP

http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_matrix_simple_regr.pdf

RÉFÉRENCES

http://www.personal.soton.ac.uk/jav/soton/HELM/workbooks/workbook_30/30_3_lu_decomposition.pdf

https://www.math.hmc.edu/calculus/tutorials/eigenstuff/

https://people.duke.edu/~ccc14/sta-663/LinearAlgebraMatrixDecompWithSolutions.html

https://www.georgebrown.ca/uploadedFiles/TLC/_documents/Basic%20Matrix%20Operations.pdf

https://bgsu.instructure.com/courses/901773/pages/p5-learning-using-bayes-rule?module_item_id=6367315

http://www4.stat.ncsu.edu/~bmasmith/ST371S11/Conditional-Probability-and-Independence.pdf

https://www2.isye.gatech.edu/~brani/isyebayes/bank/handout1.pdf

Rubrique « Joint probability distribution » de Wikipédia

Rubrique « Multivariate normal probability distribution » de Wikipédia