STATISTIQUE COMPUTATIONNELLE – 1

NOTIONS DE PROBABILITÉS – LOIS

1.1

Introduction

Il y a trois grands domaines statistiques, chaque domaine faisant appel à des techniques diverses. On a ainsi

- la production des données. On utilise les techniques de sondages ou de plans d'expérience.
- l'exploration des données. On met en oeuvre des méthodes de la statistique descriptive et, en particulier, les méthodes géométriques d'analyse des données. Ces méthodes ont pour but de résumer ou de visualiser des grands ensembles de données, sans faire référence à un modèle probabiliste.
- la modélisation statistique. Elle a pour but de formaliser un phénomène par un modèle probabiliste et de confronter ce modèle aux données.

Ce cours sera axé principalement sur le dernier domaine dont le but est de confronter les modèles à la réalité. Pour réaliser cette confrontation nous aurons parfois besoin de l'analyse de données. La dernière partie de ce cours y sera consacrée.

1.2

Les ingrédients de base

Ils sont au nombre de trois

- la population sous examen, dont la taille est considérée comme étant infinie ;

– l'échantillon de taille fini, issu de la population, et
– le paramètre par rapport auquel nous allons étudier la population.

Exemple : Supposons qu'on joue pile ou face avec un nombre infini de pièces et soient $\xi_1, \xi_2, \ldots$ les résultats, avec $x_i = 0$ si pile et $x_i = 1$ si face. L'ensemble de résultats forme la *population* $\Xi = \{\xi_1, \xi_2, \ldots\}$. On observe que la valeur de chaque résultat est indépendante de la valeur des autre résultats. Cette indépendance des éléments de la population est une propriété essentielle en statistique. Si on veut calculer la moyenne du nombre d'apparitions de face (ce qui correspond à la probabilité d'apparition de face, calculée selon l'approche fréquentielle) on doit poser $\mu = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \xi_i$ qui est ce qu'on appelle la *moyenne théorique*. Concrètement ce calcul ne peut se faire que de façon approchée, en fixant n à une valeur N_0, c-à-d. en considérant un sous-ensemble $X = \{x_1, x_2, \ldots\}$ de la population Ξ, avec $x_i = \xi_{\sigma(i)}$ où $\sigma(i)$ est un indice issu de N avec $\sigma(i) \neq \sigma(j)$ pour $i \neq j$. L'ensemble X s'appelle *échantillon* et il y a des règles précises pour sa création. En utilisant l'échantillon, on peut avoir une estimation (approximation) de la valeur de la moyenne théorique μ : $\widehat{\mu} = \bar{x} = m_X = \frac{1}{N_0} \sum_{i=1}^{N_0} x_i$ qui est la moyenne empirique. Les trois notations $\widehat{\mu} = \bar{x} = m_X$ sont en statistique équivalentes. Mais en traitement du signal, la première n'est pas utilisée, la deuxième indique une moyenne temporelle et la troisième une moyenne spatiale qui, seule, correspond à la formule que nous utilisons ici. Supposons maintenant qu'on a plusieurs populations de pièces en provenance des différents procédés de fabrication et on cherche à calculer la moyenne du nombre d'apparitions de face. Cette moyenne μ est le paramètre que nous voulons étudier et qu l'on notera par θ. On aura donc $\theta = \mu$ et dans le cas de l'estimation $\widehat{\theta} = m_X$.

Dans l'exemple ci-dessus, nous avons que la moyenne peut être différente selon la population considérée. Elle dépend donc de la probabilité P de l'apparition de face dans chaque population. Nous pouvons donc poser $\theta = \theta(P)$. Si la fonction θ est injective, on dit que nous avons un *modèle paramétrique*.

1.3 Problèmes statistiques

On suppose que nous travaillons avec un modèle paramétrique avec $\theta \in \Theta \subset \mathbb{R}^q$. En statistique, on s'occupe essentielement des problèmes suivants :

Estimation ponctuelle On cherche à calculer une valeur pour une fonction $g(\theta)$ où θ est la vraie valeur du paramètre θ et g est une fonction connue. On dit que nous avons un problème d'estimation ponctuelle, puisque résultat est un point de l'espace $g(\Theta)$. (N.B. Si on prend g comme la fonction identité, l'estimation ponctuelle revient au calcul de la vraie valeur de θ.)

Intervalle de confiance On cherche à préciser un ensemble connexe (au sens topologique du terme) de $g(\Theta)$ dans lequel se trouve $g(\theta)$.

Test d'hypothèse On partage l'espace des paramètres Θ en deux sous-ensembles Θ_0 et Θ_1, que l'on appelle hypothèses. Le problème à résoudre avec un test d'hypothèse est d'indiquer si la vraie valeur du paramètre θ se trouve dans Θ_0 ou Θ_1.

1.4

Rappels des probabilités

Définition fréquentielle de la probabilité.

On considère une expérience $\mathcal{E}$ dont le résultat est aléatoire serait un des événements élémentaires $\Omega = \{\omega_1, \ldots \omega_N\}$. Pour calculer la probabilité d'apparition d'un événement on suppose qu'on a les résultats d'un nombre n d'expériences[1]. Soit n_k le nombre de fois où l'événement élémentaire ω_k se réalise, avec $n_k \in \{ 0,\ 1,\ \ldots\ ,\ n \}$ et $k = 1,\ 2,\ \ldots,\ N$. Nous appelons *fréquence d'apparition* de l'événement élémentaire ω_k pour les n expériences, le rapport :

$$f_n(\omega_k) = \frac{n_k}{n}\ ; \qquad \forall\ \omega_k \in \Omega; \quad n \geq 1$$

On fait l'hypothèse que la limite $\lim_{n\to\infty} f_n(\omega_k)$ existe pour tout $k = 1,\ 2,\ \ldots,\ N$. Dans ce cas nous pouvons définir la probabilité d'un événement élémentaire ω_k comme la fréquence d'apparition de cet événement, quand n devient grand, i.e.

$$P(\omega_k) = \lim_{n\to\infty} f_n(\omega_k)\ ; \qquad \forall\ \omega_k \in \Omega$$

Définie de cette façon, la probabilité est en fait une application $P : \Omega \to [0,1]$ qui possède un certain nombre de propriétés et, en particulier il y en a deux qu'on oublie ou on utilise mal :

En règle générale on s'intéresse à des événements composés qui forment donc des sous-ensembles de Ω. Soient donc $A, B \subset \Omega$. On a

Propriété de la sous-additivité

$$\forall A, B \subset \Omega\ :\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Propriété de l'additivité des événements disjoints

$$\forall A, B \subset \Omega \text{ avec } A \cap B = \emptyset :\ P(A \cup B) = P(A) + P(B)$$

Conditionnement – Indépendance

Probabilité conditionnelle de la réalisation de A si B est réalisé

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

Nous avons la propriété suivante :

$$\forall\ A,\ B \subset \Omega : P(A \cap B) = P(B)\,P(A/B) = P(A)\,P(B/A)$$

et le théorème de Bayes

$$P(B_i/A) = \frac{P(B_i)\ P(A/B_i)}{P(A)} = \frac{P(B_i)\ P(A/B_i)}{\sum_{j=1}^{\infty} P(B_j)\ P(A/B_j)}$$

---

1. par exemple si l'expérience est le lancement d'un dé, on suppose que nous avons un nombre n de dès que nous lançons en même temps et non pas qu'on a lancé le dé n fois

où $B_1, \ldots$ est une partition de Ω.

Notons que les événements A et B sont indépendants si et seulement si

$$P(A/B) = P(A) \ et \ P(B/A) = P(B)$$

Variables aléatoires

Si à chaque résultat possible d'une expérience aléatoire on associe une valeur numérique, on construit une variable aléatoire. Cette variable est donc une fonction X et le résultat numérique d'une expérience sera noté par x. Formellement, considérons un sous-ensemble (en fait une tribu) $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ de l'ensemble puissance de Ω. La variable aléatoire est une application dans un ensemble discret $E : X : \mathcal{A} \rightarrow E \subseteq \mathbb{R}$ avec la propriété $\forall x \in E : X^{-1}(x) \in \mathcal{A}$.

On parle de la probabilité pour qu'un v.a. X ait comme résultat x que l'on note $P(X = x)$ ou encore $P(x)$. Il s'agit en réalité de la probabilité $P\left(X^{-1}(x)\right)$ c-à-d. de la probabilité associée à l'événement dont la valeur numérique via X est x.

Propriété fondamentale : Si X, Y deux v.a., alors $X + Y$ et $X \cdot Y$ les sont aussi.

Espérance mathématique d'une v.a. $X : E(X) = \lim_{n \rightarrow \infty} \sum_{k=1}^{n} x_k P(X = x_k)$ à condition que la limite existe.

L'espérance mathématique est un opérateur linéaire, à savoir

$$\begin{aligned} E(\lambda X + \mu) &= \lambda E(X) + \mu \\ E(X + Y) &= E(X) + E(Y) \end{aligned}$$

La variance d'une v.a.X mesure la dispersion des valeurs de la v.a. X par rapport à son espérance mathématique :

$$V(X) = E(X - E(X))^2 = E\left(X^2\right) - E(X)^2$$

Selon cette définition la variance d'une constante est nulle : $V(a) = 0$ et aussi $V(-X) = V(X)$

La variance est invariante par rapport à la translation : $V(\lambda X + \mu) = \lambda^2 V(X)$

La racine carrée de la variance est l'écart-type $s(X) = \sqrt{V(X)}$.

Soit la v.a. X. On peut définir une nouvelle v.a. $Z_X = \frac{X - E(X)}{\sqrt{V(X)}}$ qui est de moyenne nulle et de variance unité. Cette v.a., qui joue un grand rôle en Statistique, s'appelle *v.a. centrée réduite* .

Si L'ensemble E est un continuum de valeurs, alors nous avons une v.a. continue. Dans ce cas on définit la fonction de répartition de la v.a. $X : F_x : \mathbb{R} \rightarrow \mathbb{R}$ par la relation $F_X(b) = P(X \leq b)$. La dérivée de cette fonction est la fonction de densité $f_X(x)$ de la variable aléatoire X. Nous avons

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx = F_X(b) - F_X(a)$$

L'espérance mathématique de la v.a. continue est $E(X) = \int_{\mathbb{R}} x f_X(x)\, dx$ et sa variance $V(X) = \int_{\mathbb{R}} (x - E(X))^2 f_X(x)\, dx$

Un autre grandeur statistique est la médiane m d'une v.a. X qui est une valeur telle que $P(X \leq m) = 0.5$ et $P(X \geq m) = 0.5$ ou, encore, $F_X(m) = 0.5$.

1.5

Lois des probabilités

Loi binomiale

Elle décrit le comportement de plusieurs expériences identiques qui ont deux résultats possibles.

Si X v.a.suit la loi binomiale, on note : $X \sim B(n,p)$. On a

Probabilité $P_X(X=x) = C_n^x p^x (1-p)^{n-x}$, $.x = 0,1,\ldots$

Espérance mathématique $E(X) = np$

Variance $V(X) = np(1-p)$

Loi de Poisson

On l'utilise lorsque on s'intéresse au nombre de fois qu'un événement se réalise pendant un intervalle de temps ou au nombre d'articles de type spécifique qui se trouvent soit dans une région d'un espace soit dans une portion d'un volume

Si X v.a.suit la loi binomiale, on note : $X \sim P(\theta)$ ou $X \sim P(\lambda, n)$. On a

Probabilité $P_X(X=x) = \frac{e^{-\theta}\,\theta^x}{x!}$; $x = 0,1,\ldots$ avec $\theta = \lambda n$ ($\theta > 0$ est le nombre moyen de réalisations de l'événement A pendant les n unités relatives à l'experience).

Espérance mathématique $E(X) = \theta$

Variance $V(X) = \theta$

Loi uniforme discrète

Une expérience aléatoire qui se produit une seule fois et qu'elle a plusieurs résultats possibles équiprobables, suit la loi uniforme discrète

Si X v.a.suit la loi binomiale, on note : $X \sim U(q)$. On a

Probabilité $P_X(X=k) = \frac{1}{q}$; $k = 1,2,\ldots,q$

Espérance mathématique $E(X) = \frac{q+1}{2}$

Variance $V(X) = \frac{q^2-1}{12}$

Loi uniforme continue

Une expérience aléatoire qui peut prendre n'importe quelle valeur dans l'intervalle $[a,b]$ de façon équiprobable, suit la loi uniforme continue.

Si X v.a.suit la loi binomiale, on note : $X \sim U([a,b])$. On a

Densité de probabilité $f_X(x) = \frac{1}{b-a}$; $x \in [a,b]$

Espérance mathématique $E(X) = \mu$

Variance $V(X) = \frac{(b-a)^2}{12}$

Loi normale (de Laplace-Gauss

Il s'agit de la limite des lois binomiale et de Poisson lorsque le nombre d'expériences est grand.

Si X v.a.suit la loi binomiale, on note : $X \sim \mathcal{N}(\mu, \sigma^2)$. On a

Densité de probabilité $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2\,\sigma^2}\ (x-\mu)^2}$; $x \in \mathbb{R}$

Espérance mathématique $E(X) = \frac{a+b}{2}$

Variance $V(X) = \sigma^2$

Loi de Weibull

Elle représente très bien les phénomènes de vieillissement des appareils et elle peut donc être utilisée pour des études de fiabilité.

Si on a une v.a. Y qui suit la loi exponentielle $E\left(\frac{1}{(\theta)}\right)$ de paramètre $(\theta)^{-1}$, avec $\theta > 0$. Alors la v.a. $X = Y^b, b > 0$ suit la loi de Weibull.

Les différents paramètres utilisés par la loi sont :

- b paramètre de forme ou, encore, de pente (sans unités). Il influe sur la forme de la courbe de la distribution.
- x_0 paramètre de position ou de localisation. Il exprime la plus petite valeur que peut prendre la v.a. X et de ce fait il caractérise l'origine de la distribution. Le plus souvent il est pris égal à zéro.
- θ paramètre d'échelle (exprimé selon les mêmes unités que la v.a. X). Dans le cas des études de fiabilité, il est caractéristique de la durée de vie des appareils.

Si X v.a.suit la loi de Weibull, on note : $X \sim W(b, \theta, x_0)$.

Densité de probabilité $f_X(x) = \frac{b}{\theta}\left(\frac{x-x_0}{\theta}\right)^{b-1} \cdot \exp\left[-\left(\frac{x-x_0}{\theta}\right)^b\right]$; $x, \theta, x_0, b\,(x-x_0) \in \mathbb{R}_+$

Espérance mathématique $E(X) = x_0 + \theta\Gamma\left(1+\frac{1}{b}\right)$

Variance $V(X) = \theta^2\left[\Gamma\left(1+\frac{2}{b}\right) - \Gamma^2\left(1+\frac{1}{b}\right)\right]$

Remarques : La fonction de fiabilité est donnée par :

$$R(x) = 1 - F(x) = \exp\left\{-\left(\frac{x-x_0}{\theta}\right)^b\right\}$$

Le taux de défaillance de la loi de Weibull est donné par :

$$\lambda(x) = \frac{b}{\theta}\left(\frac{x-x_0}{\theta}\right)^{b-1}$$

L'étude de ce taux nous montre que :

- Si $\beta < 1$, on a $\lambda(t)$ qui décroît en fonction de t.
- Si $b = 1$, on a $\lambda(t)$ qui est constant égal à $\frac{1}{\theta}$.
- Si $b > 1$, on a $\lambda(t)$ qui croissant en fonction de t.
- $\lambda(t)$ est linéairement croissant si $b = 2$.

1.6

Exercices

EXERCICE 1.1 *(Qualité de la moyenne et de la médiane) Chaque ligne de la table suivante fournit le temps en heures passé par trois groupes de dix personnes pour la lecture des journaux pendant mois.*

3	4	4	5	5	5	5	5	5	6
6	6	6	7	7	7	7	7	8	8
9	10	10	10	10	10	10	12	55	60

La moyenne est de 10.4 heures et la médiane est de 7 heures.

Commentaires sur la représentativité de la moyenne et de la médiane et comparaison entre les deux grandeurs.

EXERCICE 1.2 *(Propriétés de la moyenne) Soient $x_1,\ldots,x_n$ et $y_1,\ldots,y_n$ deux ensembles de n mesures chacun. On note par*

- *$\bar{x},\bar{y}$ leur moyenne,*
- *$d_i = x_i - \bar{x}; i = 1,\ldots,n$ la déviation de la i-ième mesure de la moyenne de l'ensemble des mesures*

(1) Montrer que $\sum_{i=1}^{n} d_i = 0$

(2) On pose $z_i = x_i + y_i; i = 1,\ldots,n$. Montrer que $\bar{z} = \bar{x} + \bar{y}$

EXERCICE 1.3 *Des ampoules électriques sont fabriquées à deux unités de production différentes. La durées de vie moyenne des ampoules est de $\bar{x}_1$ =1495 et $\bar{x}_2$ =1875 heures respectivement et l'écart-type $s_1 = 280$ et $s_2 = 310$.*

Évaluer leur dispersion absolue et aussi relative. Commentaires.

EXERCICE 1.4 *On suppose que le nombre de filles et de garçons est le même. Trouver la distribution de la probabilité des filles et des garçons dans les familles avec trois enfants.*

EXERCICE 1.5 *Dans une loterie on peut gagner un lot de 5000 euros avec une probabilité de 0.001 et un lot de 2000 euros avec une probabilité de 0.003.*

Calculer le prix équitable d'un billet.

EXERCICE 1.6 *Un inconvénient de la retransmission par sous-réseau est la perte des informations due au fait que plusieurs hôtes essaient d'accéder au canal en même temps. Supposons qu'une période de temps est divisée en certain nombre d'intervalles. Pour chaque intervalle la probabilité qu'un hôte demande à utiliser le canal est p. Si deux ou plusieurs hôtes veulent utiliser le canal en même temps, nous avons une collision et les messages sont perdus. S'il y a n hôtes, calculer la probabilité pour que au moins deux hôtes veulent utiliser le canal en même temps.*

EXERCICE 1.7 *Un programme informatique calcule 1000 nombres aléatoires avec précision de deux chiffres décimaux et ensuite il fait fait leur somme. Calculer la probabilité que la valeur absolue de l'erreur totale d'arrondi est ≥ 1.*

EXERCICE 1.8 *Le diamètre des vis fabriquées par une machine suit la loi normale avec moyenne 10mm et écart-type 1mm. Une autre machine fabrique des écrous, dont le diamètre intérieur suit la loi normale avec moyenne 11mm et écart-type 0.5mm. Si une vis et un écrou sont choisis au hasard, quelle est la probabilité que la vis puisse entrer dans l'écrou ?*

EXERCICE 1.9 *Nous avons à notre disposition un sac avec des fruits.*

(1) On suppose que la probabilité d'avoir un fruit avarié est de 0.1. Le sac est refusé si, lors des tirages avec remise, on trouve au moins 5 fruits avariés. On commence à extraire un par un les fruits et à les examiner. Quelle est la probabilité qu'on refuse le sac en ayant examiné 25 fruits ?

(2) Supposons qu'une personne contrôle des sacs avec de fruits et qu'elle refuse 5 sacs en moyenne par jour. Quelle est la probabilité de refuser demain moins de 5 sacs ? Exactement 5 sacs ?

(3) Cette même personne déclare n'avoir à refuser que 3 sacs par semaine en moyenne. Quelle est la probabilité d'avoir à refuser 5 sacs ou plus dans un espace de vingt jours ?

EXERCICE 1.10 *On étudie la durée de vie d'une ampoule électrique. On note par X la v.a. qui à chaque ampoule prélevé au hasard associe sa durée de vie exprimée en mois. On suppose que X suit la loi de Weibull avec comme paramètres*

$$b = 2.4 \ \textit{et} \ \theta = 50$$

Déterminer le temps au bout duquel une ampoule doit être changée, en tenant compte que sa probabilité de survie doit rester supérieure à 09%.

Exercices à faire chez soi

EXERCICE 1.11 *Une compagnie d'assurance a calculé que le risque de décès d'une personne à cause d'une maladie rare dans une année est de* 0.00001. *La compagnie assure* 100000 *personnes qui risquent attraper cette maladie. Quelle est la probabilité d'avoir à payer, l'année prochaine, la prime de décès à quatre personnes ?*

EXERCICE 1.12 *La quantité d'électricité consommée par une ampoule de 100w pendant une période de 10 minutes est distribuée selon la loi normale avec comme moyenne 50 unités et comme écart-type 4 unités. Si une ampoule consomme 59, ou plus, unités pendant une période de 10 minutes, elle est considérée comme défectueuse. Quelle est la probabilité qu'une ampoule choisie pour être testée, soit défectueuse ?*

EXERCICE 1.13 *Dans un lot de pièces manufacturées, il y a une proportion D des pièces défectueuses. Pour effectuer un contrôle de qualité, nous prélevons n pièces dont r sont défectueuses.*

(1) Donner la probabilité $P(r = 0)$ en fonction de D et de n.

(2) Si $D = 10$ %, quelle est la valeur de n pour que $P(r = 0) < 5$ % ?

(3) Si $D = 5$ %, combien de pièces faut-il prélever pour que l'on ait au moins 90% de chances d'avoir au moins une pièce défectueuse ?

(4) Si $N = 50$, pour quelles valeurs de D a-t-on $P(r = 0) < 1$ % ?

(5) Si $n = 10$, quelles sont les valeurs de D qui permettent la présence d'au moins une pièce défectueuse dans au moins 95% des contrôles ?

EXERCICE 1.14 *Un central téléphonique traite, entre 14h00 et 16h00, trois appels par minute. Quelle est la probabilité de recevoir 6 appels pendant une minute ? Quelle est la probabilité de recevoir 2 appels au plus pendant deux minutes ?*