Statistique inférentielle

M1 du Master MMAS

James Ledoux

Dépt de mathématiques, Univ. Poitiers

12 juillet 2009

Statistique

Ensemble de méthodes permettant d'analyser des ensembles d'observations (ou de données)

- Méthodes relevant des mathématiques :
  Calcul des probabilités, Algèbre linéaire, Calcul scientifique, ...
- Utilisation intensive de l'outil informatique :
  SGBD, Calcul scientifique, Calculs distribués, ...

« Deux » classes de méthodes en statistique :

1. Statistique descriptive ou **exploratoire** : module au 1er semestre
2. Statistique **inférentielle** : contenu du cours

Statistique exploratoire

Objectifs :

présenter
résumer
structurer
| les données

Statistique exploratoire

- mise en évidence des propriétés de la population étudiée
- suggérer des hypothèses

Problème : données entachées d'incertitude et présentant des variations pour plusieurs raisons :

- le déroulement des phénomènes observés n'est pas prévisible à l'avance
- toute mesure est entachée d'erreur
- seuls quelques individus sont observées et on doit extrapoler les conclusions de l'étude à toute la population

Statistique inférentielle

Objectif de la statistique inférentielle : quantifier/maîtriser le hasard ou l'incertitude

- Analyser les variations dans les observations (robustesse des conclusions) : étendre à une population globale des phénomènes observés sur un « échantillon »
- Proposer un modèle probabiliste du phénomène aléatoire sur la base de la phase exploratoire
  cadre cohérent pour manipuler ces modèles : calcul des probabilités
- Valider ou infirmer les hypothèses émises
- Prévision et prise de décision

Ingénieur : prendre des décisions au des informations disponibles et contrôler les risques d'erreur

Une bibliographie succinte

- D. Dacunha-Castelle et M. Duflo. *Probabilités et Statistiques* Vol. 1. Masson, 1982.
- E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, 2004.
- E.L. Lehmann et G. Casella. *Theory of Point Estimation*. Springer, 2003.
- E. L. Lehmann et J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2006.
- A. Monfort. *Cours de statistique mathématique*. Economica, 1997.
- G. Saporta. *Probabilités, Analyse de Données et Statistique*. Technip, 2006.

Démarche statistique

Statistique et théorie des probabilités : « deux » disciplines complémentaires

1. Théorie des probabilités : branche des mathématiques pures
2. Probabilités appliquées : modèles probabilistes du déroulement de phénomènes aléatoires concrets

   prévisions préalablement à toute expérience

Durée de vie X d'un système : $X \sim \mathrm{Exp}(\lambda)$

$$\mathbb{P}\{X > t\} = 1 - \exp(-\lambda t) \quad \mathbb{E}[X] = 1/\lambda$$

n ampoules identiques sont mises en fonctionnement au même instant. Elles fonctionnent de manière indépendantes. Alors le nombre d'ampoules en pannes à l'instant t est de loi : $\mathrm{Bin}(n, \mathbb{P}\{X \leq t\}) = \mathrm{Bin}(n, 1 - \exp(-\lambda t))$.
En moyenne, on aura $n(1 - \exp(-\lambda t))$ ampoules en panne à l'instant t.

Problème : modèle exponentiel ? Si oui alors $\lambda = ?$

... statistique

3 Statistique

- Mettre en place une expérimentation pour recueillir les données et les analyser (Plans d'expériences)

Ampoule

On a mesuré la durée de vie en h de 10 ampoules identiques :

$$x_1 = 91.6,\ x_2 = 35.7,\ x_3 = 251.3,\ x_4 = 24.3,\ x_5 = 5.4,$$
$$x_6 = 67.3,\ x_7 = 170.9,\ x_8 = 9.5,\ x_9 = 118.4,\ x_{10} = 57.1$$

Modèle statistique : la durée de vie d'une ampoule n'étant pas prévisible à l'avance, on considère x_i comme la réalisation d'une variable aléatoire X_i :

- Puisque c'est le même type d'ampoules, les v.a. $X_1, \ldots, X_n$ sont de même loi
- Puisque les ampoules sont censées avoir fonctionner de manière indépendantes, alors les v.a. $X_1, \ldots, X_n$ sont supposées indépendantes

↪ variabilité sur les données

1. Au vu des observation, est-il raisonnable de supposer que la durée de vie d'une ampoule est une v.a. de loi exponentielle ? Si non, quel autre loi choisir ?
   Pb : choix de modèles ou tests d'adéquation
2. Si le modèle exponentiel est correct, comment proposer une valeur ou un ensemble de valeurs pour λ ?
   Pb : estimation paramétrique
3. Dans ce cas, peut-on garantir que λ est inférieur à un certain seuil λ_0 ? cela garantira que $\mathbb{E}[X] = 1/\lambda \geq 1/\lambda_0$, autrement dit que les ampoules sont suffisamment fiable en moyenne
   Pb : test d'hypothèse paramétrique
4. Sur un parc de 100 ampoules, à combien estime-t-on le nombre de panne en moins de 50 heures ?
   Pb : prévision

Dans ce cours

1. **Estimation :** proposer une approximation des grandeurs inconnues au vu des observations qui soit la plus proche possible de la réalité
2. **Tests d'hypothèses** : se prononcer sur la validité d'une hypothèse lié au problème, loi, comparaison à un seuil, l'objectif de fiabilité est-il satisfait ? Attribuer un niveau de confiance dans ses réponses

Privilégier l'application à la théorie

Domaines : informatique, médecine, contrôle de qualité, sciences de la vie, de l'environnement, ...

Statistiques paramétriques vs non-paramétriques

x observation d'une v.a. X

La statistique inférentielle associe à cette observation un modèle statistique $(E_X, \mathscr{E}, \mathcal{P})$

- E_X est l'espace des observations, i.e. ensemble des valeurs possibles de x : on supposera que cet ensemble est discret ou $\mathbb{R}^n$ pour un certain n
- $\mathscr{E}$ est la tribu des observables associée à E_X
- $\mathcal{P}$ est l'ensemble des lois de probabilité possible pour X, définie sur $\mathscr{E}$ qu'on identifiera à l'ensemble des fonctions de répartitions pour X.

Abus de notation

La v.a. est définie comme une application mesurable définie sur un certain espace probabilisé. Ainsi $\mathbb{P}_\theta$ dans le modèle statistique $(E_X, \mathscr{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ devrait être notée $\mathbb{P}_\theta^X$ (ou $X(\mathbb{P}_\theta)$) pour mettre en évidence qu'il s'agit de la loi image par X d'une certaine mesure de probabilité sous-jacente.

1. **Statistique paramétrique :** $\mathcal{P}$ est en bijection avec un ensemble de paramètres appartenant à un espace de dimension finie.

   Alors on peut considérer que $\mathcal{P}$ est un ensemble de lois indicées par un paramètre θ

$$\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta \subset \mathbb{R}^p\}$$

- $\theta = \lambda \in \Theta =]0, +\infty[$ et $F_\theta(t) = \big(1 - \exp(-\lambda t)\big)1_{\mathbb{R}_+}$
- $\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+$ et

$$F_\theta(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(u-m)^2/2\sigma^2)\, du$$

2. **Statistiques non paramétriques :** $\mathcal{P}$ ne peut pas être mis sous la forme précédente.

   Par exemple $\mathcal{P}$ est
   - l'ensemble des lois absolument continues sur $\mathbb{R}$
   - l'ensemble des lois de support $[0, 1]$
   - ...

   - Objectifs restent les mêmes que le cas paramétrique : extraire de l'information pertinentes de données
   - En général, l'avantage du cadre non-paramétrique est de s'affranchir de choisir $\mathcal{P}$ dans une famille usuelle de lois paramétrées.
   - Par contre, si les observations sont tirées d'un modèle précis, les procédures non-paramétriques sont moins performantes que celles dédiées au modèle sous-jacent

Remarque : il existe un cas intermédiaire lorsqu'un paramètre est présent mais ne caractérise pas la loi sous-jacente. On parle de statistiques semi-paramétriques.

Modèle statistique dominé et vraisemblance

Dans tout le cours, le modèle $(E_X, \mathscr{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est supposé **dominé** :

Définition 1 (Modèle dominé)

On dit que le modèle $(E_X, \mathscr{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est dominé s'il existe une mesure (σ-finie) μ telle que, pour tout $\theta \in \Theta$, la probabilité $\mathbb{P}_\theta$ est absolument continue par rapport à μ ($\mathbb{P}_\theta << \mu$).
*Une application $L : (\theta, x) \mapsto \mathbb{R}_+$ telle que, pour tout $\theta \in \Theta$, $x \mapsto L(\theta, x)$ est une densité de $\mathbb{P}_\theta$ relativement à μ ($L(\theta, \cdot) = d\mathbb{P}_\theta/d\mu$), est appelée une **vraisemblance du modèle**.*

Cas fondamentaux

Nos deux exemples fondamentaux pour μ seront

- mesure de Lebesgue sur $\mathbb{R}^p$: $L(\theta, \cdot)$ est une densité de probabilité usuelle
- mesure de comptage sur un ensemble dénombrable : $L(\theta, x)$ est la probabilité pour que x soit observé

Abus de notation

$\mathbb{E}_\theta[g(X)]$ désignera l'intégrale de $x \mapsto g(x)$ relativement à la mesure de proba. $\mathbb{P}_\theta \equiv \mathbb{P}_\theta^X$:

$$\mathbb{E}_\theta[g(X)] = \int_{E_X} g(x) d\mathbb{P}_\theta(x) = \int_{E_X} g(x) L(\theta, x) d\mu(x)$$

Remarque 1

Dans un modèle dominé, pour tout $\theta \in \Theta$, on a $L(\theta, \cdot) > 0$ $\mathbb{P}_\theta$-ps. En général, la vraisemblance n'est pas strictement positive pour μ-presque tout x. Elle le sera ssi $\mu \sim \mathbb{P}_\theta$ pour tout $\theta \in \Theta$ et alors $d\mu/d\mathbb{P}_\theta = 1/L(\theta, \cdot)$.

Introduisons la définition suivante :

Définition 2 (Modèle homogène)

On dit que le modèle $(E_X, \mathscr{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ est homogène s'il existe une mesure dominante μ telle que la vraisemblance $L(\theta, x)$ associée est strictement positive pour μ-presque tout x (ou de manière équivalente, $\forall(\theta, \theta') \in \Theta^2$, $\mathbb{P}_\theta << \mathbb{P}_{\theta'}$).

Pour simplifier la présentation tout en conservant les moyens d'analyser les modèles standards :

Hypothèse de base pour tout le cours

le modèle statistique $(E_X, \mathscr{E}, (\mathbb{P}_\theta)_{\theta\in\Theta})$ sera dominé relativement à μ.

Des exemples classiques

1. Des lois (discrètes) absolument continues par rapport à la mesure de comptage : $\mathbb{P}_\theta$ est
   - la loi de Bernoulli $\mathrm{Ber}(\theta)$, la loi binomial $\mathrm{Bin}(n, \theta)$, $\mathrm{Geo}(\theta)$ avec $\theta \in \Theta =]0, 1[$
   - la loi de Poisson $\mathrm{Pois}(\theta)$ avec $\theta \in \Theta =]0, +\infty[$,
   - ...
2. Des lois absolument continues par rapport à la mesure de Lebesgue : $\mathbb{P}_\theta$ est
   - la loi exponentielle $\mathrm{Exp}(\theta)$ où $\theta \in \Theta =]0, +\infty[$,
   - la loi gamma $\mathrm{Ga}(\theta)$ avec $\theta = (\alpha, \lambda) \in \Theta =]0, +\infty[^2$,
   - la loi gaussienne unidimensionnelle $\mathcal{N}(\theta)$ avec $\theta = (m, \sigma^2) \in \Theta = \mathbb{R}\times]0, +\infty[$,
   - la loi gaussienne multidimensionnelle $\mathcal{N}(\theta)$ où $\theta = (m, V) \in \Theta = \mathbb{R}^p \times \mathcal{L}^{++}$ où $\mathcal{L}^{++}$ est l'ensemble des matrices **définies** positives
   - ...

Loi gamma

1. Loi gamma $\mathrm{Ga}(\theta)$ avec $\theta = (\alpha, \lambda) \in \Theta =]0, +\infty[^2$:

$$f(\theta, x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha \exp(-\lambda x) x^{\alpha-1} \mathbf{1}_{]0,+\infty[}(x)$$

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} \exp(-x) dx \qquad \Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

2. $\mathbb{E}_\theta[X] = \alpha/\lambda$ et $V_\theta(X) = \alpha/\lambda^2$
3. Si $X \sim \mathrm{Ga}(\alpha, \lambda)$ alors $aX \sim \mathrm{Ga}(\alpha, \lambda/a)$ pour $a > 0$.
4. Propriétés de convolution :

$$\mathrm{Ga}(\alpha, \lambda) \star \mathrm{Ga}(\beta, \lambda) = \mathrm{Ga}(\alpha + \beta, \lambda)$$

5. Pour une vraisemblance exponentielle $\mathrm{Exp}(\lambda)$, la famille $\mathcal{G}$ des lois gamma est conjuguée, i.e. si on choisit une loi a priori pour λ dans $\mathcal{G}$, la loi a posteriori de λ se trouve également dans $\mathcal{G}$
6. $\mathrm{Ga}(1,\lambda) = \mathrm{Exp}(\lambda)$ (file d'attente, durée de vie, ...)
7. $\mathrm{Ga}(n,\lambda) = \mathrm{Erl}_n$ est la loi d'Erlang, i.e. la convolution $\star_n \mathrm{Exp}(\lambda)$ (file d'attente, fiabilité, ...)
8. Relation avec la loi du chi-deux : $\chi^2_n = \mathrm{Ga}(n/2, 1/2)$ et

$$\mathbb{E}_\theta[X] = n, \quad V_\theta(X) = 2n,$$
$$\chi^2_{n_1} \star \chi^2_{n_2} = \chi^2_{n_1+n_2}$$

Applications : modèle général de durée de vie, l'archétype de la famille de lois a priori en statistique bayesienne

Vecteur et lois gaussiens (cf cours/TD de proba)

1. La densité : $\forall x \in \mathbb{R}^p$

$$f(x) = \frac{1}{(2\pi)^{p/2}|\det(V)|^{1/2}} \exp\left((x-m)^\top V^{-1}(x-m)\right)$$

2. Les composantes sont nécessairement de loi gaussienne
3. Les composantes sont indépendantes ssi elles sont non-corrélées ssi la matrice de covariance V est diagonale
4. Les n composantes sont i.i.d. ssi la matrice de variance $V = \sigma^2 I_n$ où $\sigma^2 > 0$
5. Le carré de la norme $\|X\|_2^2$ d'un vecteur gaussien de paramètres $m = \mathbf{0}, V = I_n$ suit une loi du χ^2_n
(et $\|X\|_2^2/\sigma^2 \sim \chi^2_n$ pour un vecteur gaussien $(m,V) = (\mathbf{0}, \sigma^2 I_n)$)

Théorème de Cochran

Si $X = (X_1, \ldots, X_n)$ est un vecteur gaussien de paramètre $(\mathbf{0}, \sigma^2 I_n)$ alors

1. Le transformé de X par une transformation orthogonale est un vecteur gaussien de paramètre $(\mathbf{0}, \sigma^2 I_n)$
2. Soit $E_1 \oplus \cdots \oplus E_p$ une décomposition de $\mathbb{R}^n$ en p sous-espaces orthogonaux de dimension respective $r_1, \ldots, r_p$.
Alors les vecteurs aléatoires $P_{E_1}X, \ldots, P_{E_p}X$, projection orthogonale de X sur les différents sous-espaces, sont indépendants et les v.a. $\|P_{E_1}X\|_2^2/\sigma^2, \ldots, \|P_{E_p}X\|_2^2/\sigma^2$ sont de loi $\chi^2_{r_1}, \ldots, \chi^2_{r_p}$.

Loi de Student (W. Gosset)

Loi de Student

La loi de Student à $n \geq 1$ degrés de liberté, notée st_n, est la loi du ratio

$$\frac{Z_1}{\sqrt{Z_2/n}} \qquad \textit{Informel } st_n = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_n/E[\chi^2_n]}}$$

où

- Z_1 et Z_2 sont indépendantes
- $Z_1 \sim \mathcal{N}(0,1)$ et $Z_2 \sim \chi^2_n$ à n degrés de liberté

$$\forall x \in \mathbb{R}, \quad f_{st_n}(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

- $\mathbb{E}_\theta[X] = 0$ et $V_\theta(X) = n/(n-2)$ si $n > 2$

Applications : l'étude de la moyenne d'un échantillon indépendant de v.a. d'une même loi gaussienne de variance inconnue, tests statistiques, modélisation des v.a. avec excès de kurtosis,...

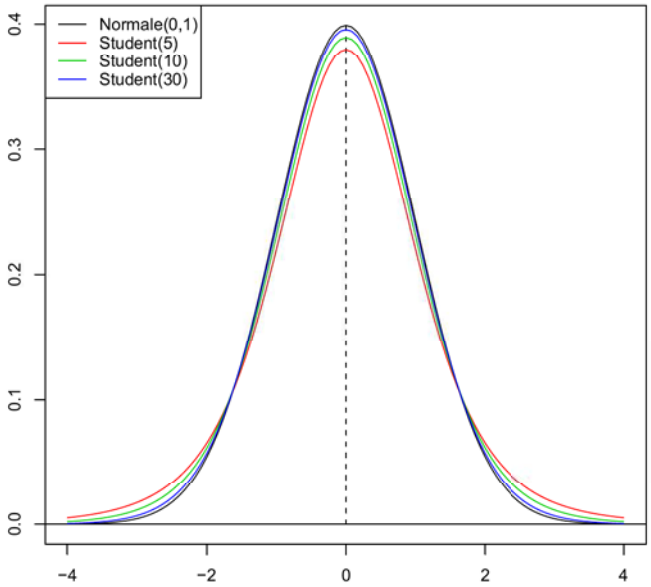

FIGURE 1: Densités de type Student et la loi normale centrée-réduite

Loi de Fisher-Snedecor

Loi de Fisher-Snedecor

Soit $X_1 \sim \chi^2_{n_1}$ et $X_2 \sim \chi^2_{n_2}$ indépendantes alors la loi du rapport

$$\frac{X_1/n_1}{X_2/n_2} \qquad \textit{Informel } f(n_1,n_2) = \frac{\chi^2_{n_1}/\mathbb{E}[\chi^2_{n_1}]}{\chi^2_{n_2}/\mathbb{E}[\chi^2_{n_2}]}$$

est appelée loi de Fisher notée $f(n_1,n_2)$

1. $\mathbb{E}_\theta[X] = n_2/(n_2-2)$ si $n_2 \geq 3$ et $V_\theta(X) = \left(n_2/(n_2-2)\right)^2 \frac{2(n_1+n_2-2)}{n_1(n_2-4)}$ si $n_2 \geq 5$
2. Si T suit une loi Student st_n, alors T^2 suit une loi $f(1,n)$
3. Si T suit une loi $f(n_1,n_2)$ alors $1/T$ suit une loi $f(n_2,n_1)$

Applications : analyse de la variance, tests statistiques (exemple : en régression)

$$f_{n_1,n_2}(x) = \mathbf{1}_{]0,+\infty[}(x) \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} {n_1}^{n_1/2} {n_2}^{n_2/2} \frac{x^{n_1/2-1}}{(n_1x+n_2)^{(n_1+n_2)/2}}$$

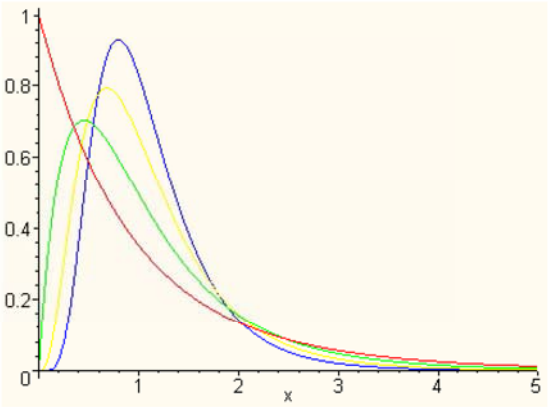

FIGURE 2: Densité de la loi $f(n_1,20)$ avec $n_1 = 2,4,8,16$.

Famille exponentielle

Définition 3 (Famille exponentielle)

On dit que la famille $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ est une famille exponentielle de dimension p relativement à la mesure dominante μ s'il existe des applications mesurables

- $\alpha : \Theta \mapsto \mathbb{R}^p$, $\beta : \Theta \mapsto \mathbb{R}_+$,
- $t : E_X \mapsto \mathbb{R}^p$ *et* $\xi : E_X \mapsto \mathbb{R}_+$

telles qu'une vraisemblance du modèle statistique s'écrive

$$(1) \qquad \forall (\theta,x) \in \Theta \times E_X, \quad L(\theta,x) = \beta(\theta)\xi(x)\exp\left(\langle t(x), \alpha(\theta)\rangle\right)$$

Une telle écriture impose que

$$c_\mu(\alpha(\theta)) := \int_{E_X} \xi(x)\exp\left(\langle t(x),\alpha(\theta)\rangle\right) d\mu(x) < +\infty$$

Notons que $\beta(\theta) = 1/c_\mu(\alpha(\theta))$.

- *t est est la statistique naturelle*
- *$\alpha(\theta)$ le paramètre naturel et*
- *l'espace naturel des paramètres est l'ensemble $N_\mu := \{a \in \mathbb{R}^p,\ c_\mu(a) < +\infty\}$*