Premières notions de statistique: modèles aléatoires et estimation

Franck Picard

Licence 3 Biosciences, 2023-2024

Laboratoire Biologie et Modélisation de la Cellule, CNRS ENS-Lyon

franck.picard@ens-lyon.fr

Outline

Pourquoi des modèles ?

- Le résultat d'une expérience n'est pas strictement prévisible, il est donc aléatoire.
- On modélise les données comme la **réalisation de variables aléatoires**.
- Les outils probabilistes/statistiques permettent d'extraire la part de l'information qui est reproductible
- La démarche statistique suppose qu'un échantillon **ne renseigne que partiellement** sur l'ensemble de la population d'intérêt

Les modèles probabilistes permettent de prendre des décisions sur des bases probabilistes par la formalisation un risque (erreur)

Objectifs du cours

- Reconnaître les situations usuelles : données discrètes / continues
- Associer les modèles usuels: Binomial/Multinomial, Poisson, uniforme, gaussien
- Connaitre les moyennes et les variances de chaque modèle
- Associer la forme des densités / fonctions de répartition aux modèles

Variables aléatoires et vecteurs aléatoires

- Une variable aléatoire X est une **fonction** d'une expérience aléatoire.
- Une réalisation x d'une variable aléatoire X est la valeur de la variable après avoir effectué l'expérience.
- **on considère que les données sont des réalisations de variables aléatoires**:
- Si on note $\mathbf{x} = (x_1, \ldots, x_n)$ le vecteur des observations,
- On suppose que $\mathbf{x}$ est la réalisation d'un **vecteur aléatoire** $\mathbf{X} = (X_1, \ldots, X_n)$
- La loi jointe des observations est:

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(X_1 = x_1, ..., X_n = x_n)$$

Echantillon Indépendant et Identiquement distribué

- On considère un n échantillon aléatoire $\mathbf{X}$
- La notion d'échantillon i.i.d. est centrale et repose sur les hypothèses:
  - Les X_i sont indépendants entre eux
  - Les X_i sont tous de même distribution
- Dans ce cadre, le n échantillon aléatoire correspond à n répétitions indépendantes de la **même distribution**.
- La loi jointe d'un n-échantillon est donc:

$$\begin{aligned}\mathbb{P}(\mathbf{X}=\mathbf{x}) &= \mathbb{P}(X_1=x_1,...,X_n=x_n)\\ &= \mathbb{P}(X_1=x_1)\times...\times\mathbb{P}(X_n=x_n)\\ &= \prod_i \mathbb{P}(X_i=x_i)\end{aligned}$$

Outline

Généralités sur les variables aléatoires discrètes

- Les valeurs prises $X(\Omega)$ par les variables sont discrètes
- On distingue souvent les cas binaires, catégoriels, quantitatifs discrets (comptages)
- La loi des variables discrètes est complètement déterminée par $\mathbb{P}(X = k)$ pour tous les k dans $X(\Omega)$,
- La propriété fondamentale:

$$\sum_{k \in X(\Omega)} \mathbb{P}(X = k) = 1$$

- On note F la fonction de répartition de X définie par

$$F_X(k) = \mathbb{P}\{X \leq k\}$$

- Exemple de lois: Bernoulli, Binomiale, Multinomiale, Poisson, Géométrique

Espérance des variables aléatoires discrètes

- L'espérance et d'une loi discrète est définie par

$$\mathbb{E}(X) = \sum_{k \in X(\Omega)} k\mathbb{P}(X = k)$$

- L'espérance d'une fonction de X est :

$$\mathbb{E}\Big(f(X)\Big) = \sum_{k \in X(\Omega)} f(k)\mathbb{P}(X = k)$$

- Propriété de linéarité: si (a, b) sont deux constantes réelles et X une variable aléatoire

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

- Une variable aléatoire centrée est d'espérance nulle:

$$X - \mathbb{E}(X)$$

Variance des variables aléatoires discrètes

- La variance quantifie l'écart quadratique moyen d'une variable à son espérance t.q.

$$\mathbb{V}(X) = \mathbb{E}\Big(X - \mathbb{E}(X)\Big)^2 = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2 = \sum_{k \in X(\Omega)} \Big(k - \mathbb{E}(X)\Big)^2 \mathbb{P}(X = k)$$

- L'écart-type est de même unité que la variable:

$$\sqrt{\mathbb{V}(X)}$$

- Propriété d'échelle: si (a, b) sont deux constantes réelles et X une variable aléatoire

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

- Une variable aléatoire réduite est de variance 1:

$$\frac{X}{\sqrt{\mathbb{V}(X)}}$$

- Réduire des variables s'avère important si on veut les comparer à la même échelle

Somme de variables aléatoires

- L'espérance d'une somme de variable aléatoire est la somme des espérances

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

- La variance d'une somme de variable aléatoire

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2 \times \mathbb{E}\Big(\left[X - \mathbb{E}(X)\right] \left[Y - \mathbb{E}(Y)\right] \Big)$$

- La covariance entre deux variables aléatoires

$$\text{cov}(X,Y) = \mathbb{E}\Big(\left[X + \mathbb{E}(X)\right] \left[Y - \mathbb{E}(Y)\right] \Big)$$

- Si deux variables sont indépendantes, alors leur covariance est nulle (pas de réciproque)

Moment de la somme d'un échantillon i.i.d.

- Si on considère un n-échantillon i.i.d. de même loi F_θ et d'espérance μ et de variance σ^2
- On s'intéresse souvent à la variable cumulée:

$$Y_n = \sum_{i=1}^{n} X_i$$

- Si $\mathbf{X}$ est identiquement distribué, alors (par linéarité)

$$\mathbb{E}(Y_n) = \mathbb{E}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n \times \mathbb{E}(X_i) = n\mu$$

- Si $\mathbf{X}$ est indépendant et identiquement distribué, alors (par indépendance)

$$\mathbb{V}(Y_n) = \mathbb{V}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \mathbb{V}(X_i) = n \times \mathbb{V}(X_i) = n\sigma^2$$

Le modèle de Bernoulli

- Intervient dans des situations où l'issue de l'expérience est binaire
- Les valeurs prises $X(\Omega)$ par les variables sont $\{0, 1\}$
- La loi de la variable binaire est telle que:

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p,$$

- Lorsque l'expérience est répétée n fois de manière indépendante, on note X_i la $i$ème répétition:

$$\forall i \in \{1, \ldots, n\}, \quad X_i \sim \mathcal{B}(p)$$

- C'est le même paramètre qui régit la loi de l'échantillon.
- p est le paramètre de fréquence des issues positives dans la population générale

Exemple de modèle pour des observations binaires

- On s'intéresse à la fréquence des hétérozygotes dans une population. Le gène d'intérêt comporte deux allèles `"A"` et `"a"`.
- On génotype plusieurs individus non apparentés dans la population:

Genotype	`"AA"`	`"Aa"`	`"Aa"`	`"AA"`	`"Aa"`	`"aA"`
x_i	0	1	1	0	1	1

- On note x_i la variable qui vaut 1 si l'individu i dans la population est hétérozygote, 0 sinon.

Exemple de modèle pour des observations binaires - 2

- On suppose que x_i est la réalisation d'une variable aléatoire X_i.
- Les individus étant non apparentés, on suppose que les X_i sont indépendants
- On suppose que les X_i sont identiquement distribués de loi Binomiale de paramètre p:

$$\forall i \in \{1, \ldots, n\},\ X_i \sim \mathcal{B}(p) = \mathcal{B}(1, p)$$

- La loi des X_i est de la forme

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p,$$

- Le paramètre p caractérise la fréquence théorique des hétérozygotes dans la population générale

Modèle Binomial

- On peut s'intéresser au résultat cumulé de toutes les expériences:

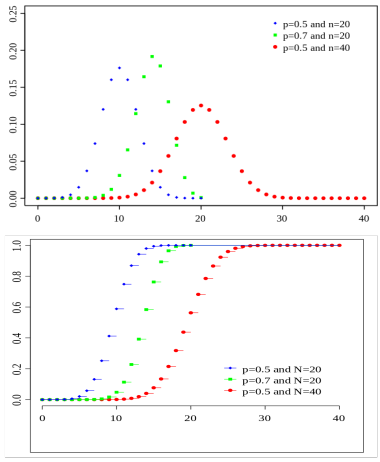
$$Y_n = \sum_{i=1}^{n} X_i$$

- C'est un nombre entier dans $\{0, \ldots, n\}$.
- Les expériences étant indépendantes:

$$Y_n \sim \mathcal{B}(n, p), \quad \mathbb{P}(Y_n = k) = C_n^k p^k (1-p)^{n-k}$$

- Les 2 premiers moments de la loi sont:

$$\mathbb{E}(Y_n) = np, \quad \mathbb{V}(Y_n) = np(1-p)$$



Loi Binomiale

Exemple de modèle pour des observations qualitatives

- Supposons que l'on s'intéresse à la fréquence des génotypes

Genotype	`"AA"`	`"Aa"`	`"Aa"`	`"AA"`	`"Aa"`	`"aA"`

- On note x_i la variable qualitative à $K = 3$ modalités, qui sont `"AA"`,`"Aa"` ou `"aa"`.
- On suppose que x_i est la réalisation de X_i.
- Les individus étant non apparentés, on suppose que les X_i sont indépendants de loi

$$\forall i \in \{1, \dots, n\},\ X_i \sim \mathcal{M}(1, p_{\texttt{AA}}, p_{\texttt{Aa}}, p_{\texttt{aa}}),$$

- La loi multinomiale généralise la loi Binomiale à plus de deux modalités, avec

$$p_{\texttt{AA}} + p_{\texttt{Aa}} + p_{\texttt{aa}} = 1.$$

Le modèle multinomial (1)

- On observe une variable $X_i = (X_i^1, \ldots, X_i^K)$ t.q. $X_i^k = 1$ si la catégorie k est observée:

Modalités	x_1	x_2	x_3	x_4	x_5	x_6	y_6
`"AA"`	1	0	0	1	0	0	2
`"Aa"`	0	1	1	0	1	1	4
`"aa"`	0	0	0	0	0	0	0

- Chaque modalité a une probabilité p_k d'occurrence

$$\forall k \in \{1, \ldots, K\}, \quad \mathbb{P}(X_i^k = 1) = p_k, \quad X_i \sim \mathcal{M}(1, p_1, \ldots, p_K)$$

Le modèle multinomial (2)

- On considère un n échantillon $(X_1, \ldots, X_n)$, et on cumule l'information t.q.

$$Y_n = (Y_n^1, \ldots Y_n^K), \quad Y_n^k = \sum_{i=1}^{n} X_i^k, \quad Y_n \sim \mathcal{M}(n, p_1, \ldots, p_K).$$

$$P(Y_n^1 = y_n^1, \ldots, Y_n^K = y_n^K) = \frac{n!}{y_n^1! \ldots y_n^K!} p_1^{y_n^1} \ldots p_k^{y_n^K}$$

- Les moments d'une loi multinomiale

$$\mathbb{E}(Y_n^k) = np_k, \quad \mathbb{V}(Y_n^k) = np_k(1 - p_k)$$

- Les comptages ne sont pas indépendants car $\sum_{k=1}^{K} Y_n^k = n$ donc

$$\text{cov}(Y_n^k, Y_n^{k'}) = -p_k p_{k'}$$

Notion de loi limite (1)

- A partir du modèle de Bernoulli, on peut accéder à d'autres lois fondamentales
- Si on considère une variable aléatoire qui est un comptage d'événements telle que son espérance est:

$$\mathbb{E}(Y_n) = np$$

- Le comptage attendu est composé d'un nombre d'expériences et d'une probabilité de survenue
- Que se passe-t-il si les comptages attendus sont rares ?

$$p \to 0$$

- Que se passe-t-il si les comptages attendus sont fréquents ?

$$np \to +\infty$$

- Les régimes limites posent des questions calculatoires, on va donc chercher des approximations

Notion de loi limite (2)

- n représente la quantité d'information disponible.
- Si $n \to +\infty$ on accède à la population d'intérêt (idéale):

"Que se passerait-il si on disposait d'un échantillon de taille infinie ?"

"si toute l'information était disponible ?"

- L'asymptotique $np \to +\infty$ correspond à $n \to +\infty$ et p fixé
- L'asymptotique en $np \to 0$ correspond au régime ou le paramètre est de très faible intensité

$$p \to 0$$

- On peut étudier ces régimes par simulation ou par le calcul mathématique

Loi limite en fréquence faible

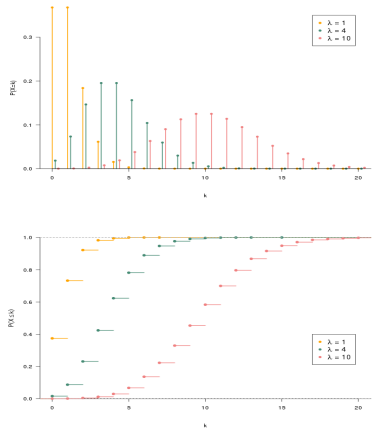
- On considère une loi de comptage avec

$$\mathbb{E}(Y_n) = \lambda = np$$

- On considère que les événements sont rares: p faible
- On considère qu'on observe une infinité d'observations

$$\mathbb{P}(Y_n = k) = C_n^k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

- On s'interroge sur la convergence en loi de la variable Y_n



Loi de Poisson

Loi de Poisson et loi des comptages

- on considère p faible et $n \to +\infty$

$$\mathbb{P}(Y_n = k) = C_n^k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \to \frac{e^{-\lambda}\lambda^k}{k!}$$

- Cette loi limite est la loi de Poisson d'intensité λ, à valeur dans $\mathbb{N}$:

$$Y \sim \mathcal{P}(\lambda), \quad \mathbb{P}(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- Les moments de la loi de Poisson:

$$\mathbb{E}(Y) = \lambda, \quad \mathbb{V}(Y) = \lambda$$

- La variabilité des lois de comptages augmente avec l'intensité du comptage

Exemple de modèle pour données de comptage

- En période d'épidémie, on s'intéresse au nombre d'individus inféctés dans les départements francais.
- On effectue un relevé et on obtient:

Département	Nb d'individus infectés
01	120
02	200
...	

- On définit x_i le nombre observé d'individus infectés dans le département i.
- On suppose que x_i est la réalisation d'une variable aléatoire X_i. On suppose que les X_i sont indépendants (est-ce réaliste ?)
- On supposer que les X_i sont iid, de même loi $X_i \sim \mathcal{P}(\lambda)$. Interprétation de λ ?

Outline

Variables aléatoires continues

- Les valeurs prises $X(\Omega)$ par les variables appartiennent à des ensembles continus
- La probabilité d'un point est nulle ! On raisonne par intervalle :

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(a \leq X \leq b)$$

- La loi d'une variable continue est définie par sa densité $f \geq 0$ ou par sa fonction de répartition F:

$$\int_{X(\Omega)} f(x)dx = 1, \ F(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f(u)du$$

- Exemples: loi normale, exponentielle, gamma, Cauchy (pas d'espérance)

Moments d'une loi continue

- L'espérance d'une variable aléatoire continue est un paramètre défini tel que:

$$\mathbb{E}(X) = \int xf(x)dx = \mu$$

- La variance d'une variable aléatoire continue est l'espérance de la distance entre la variable et son espérance:

$$\mathbb{V}(X) = \mathbb{E}(X - \mu)^2$$

$$\mathbb{V}(X) = \int (x - \mu)^2 f(x)dx$$

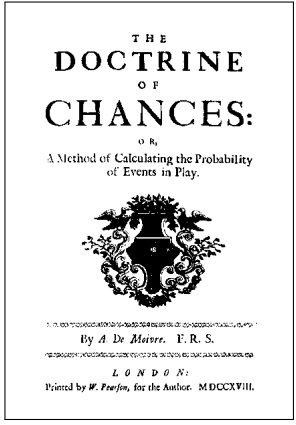
- Une variable centrée réduite:

$$Z = \frac{X - \mu}{\sigma}$$

Convergence vers la loi normale

- Quelle est la loi limite du modèle Binomial lorsque $n \to +\infty$?
- Historiquement, le résultat a été exploré par Bernoulli, de Moivre, Gauss et Laplace (entre autres)
- La loi Binomiale $\mathcal{B}(n, p)$ converge vers une loi normale quand $n \to +\infty$

$$\mathcal{B}(n, p) \to \mathcal{N}\Big(np, np(1-p)\Big)$$

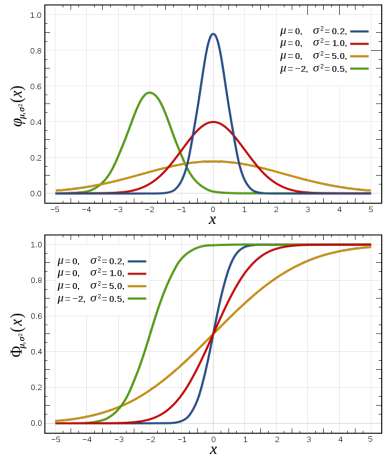

THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability
of Events in Play.
By A. De Moivre. F. R. S.
LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.

Loi Normale

- C'est la loi de référence pour les variables continues
- Elle est caractérisée par deux paramètres: espérance μ et variance σ^2
- Sa densité est de la forme, $\forall x \in \mathbb{R}$

$$\phi(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \times \frac{(x-\mu)^2}{\sigma^2}\right\}$$

- Sa fonction de répartition est souvent notée $\Phi(x, \mu, \sigma)$



Loi Normale

Modélisation de la taille d'arbres

- On note x_i^1 la taille de l'arbre i mesuré dans la forêt de type 1, (resp. x_i^2).
- On suppose que x_i^1 est la réalisation d'une variable aléatoire X_i^1.
- On peut supposer que les tailles des arbres sont indépendantes les unes des autres: on suppose que les X_i^1 sont indépendants.
- On peut également supposer que la distribution empirique des x^i peut être approchée par une même loi.

Forêt 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	27
Forêt 2	22.5	22.9	23.74	24.0	24.4	24.5	25.3	26	26.4	26.7	32

Modélisation de la taille des arbres

- On note x_i^1 la taille de l'arbre i mesuré dans la forêt de type 1, (resp. x_i^2).
- On suppose que x_i^1 est la réalisation d'une variable aléatoire X_i^1.
- On peut supposer que les tailles des arbres sont indépendantes les unes des autres: on suppose que les X_i^1 sont indépendants.
- On peut également supposer que la distribution empirique des x^i peut être approchée par une même loi, telle que:

$$\forall i \in \{1, \dots, n_1\},\ X_i^1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\forall i \in \{1, \dots, n_2\},\ X_i^2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

- Etant donné qu'on a deux types de forêt, comment interpréter le modèle tel que $\mu_1 = \mu_2$ et celui tel que $\sigma_1 = \sigma_2$?

Les tables de la loi (normale et autres !)

- C'est un outil de base qui donne les quantiles et la fonction de répartition (Φ) de la loi normale **centrée réduite**
- Si on considère une variable $X \sim \mathcal{N}(\mu, \sigma^2)$, on se ramènera toujours à la loi centrée réduite

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

- Si on souhaite calculer : $\mathbb{P}\{X \leq x\}$

$$\mathbb{P}\left\{\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right\} = \mathbb{P}\left\{Z \leq \frac{x - \mu}{\sigma}\right\} = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

- Il existe des tables pour beaucoup de lois usuelles (Fisher, χ^2...)

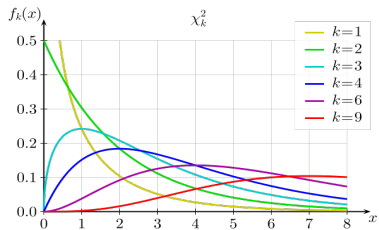
Loi du χ^2

- C'est une loi continue pour des valeurs positives de densité

$$f(x;n) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{x/2}$$

- Elle dépend d'un paramètre qui s'appelle le degré de liberté (n)
- Elle est très utilisée car elle modélise la loi des carrés de variables gaussiennes
- La fonction Γ généralise les factoriels pour des nombres réels

$$\forall x > 0,\ \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$



Loi du χ^2

La loi du chi2 et loi des sommes de carrés

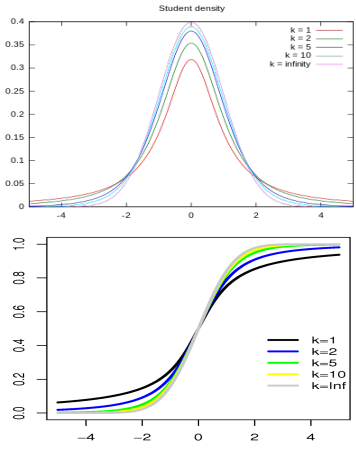
- Si $X_i \sim \mathcal{N}(0,1)$ alors
$$X_i^2 \sim \chi^2(1)$$
- Si on considère $(X_1, ...X_n)$ iid de même loi $\mathcal{N}(0,1)$ alors
$$S^2 = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n), \quad \mathbb{E}(S^2) = n, \quad \mathbb{V}(S^2) = 2n$$
- n est appelé le degré de liberté (ddl, dF) : c'est le nombre de composantes indépendantes de S^2
- Loi importante pour la loi de l'estimateur de la variance

Loi de Student (1)

- C'est une loi symétrique pour les variables continues à valeurs réelles

$$f(x;n) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

- Elle a la propriété d'avoir plus de masse en queue de distribution
- Quand $n \to +\infty$ elle peut être approchée par une loi normale



Loi de Student

Loi de Student (2)

- Si U est une variable aléatoire telle que $U \sim \mathcal{N}(0,1)$, et V est une variable aléatoire telle que $V \sim \chi^2(n)$.
- On suppose que U et V sont indépendantes
- La variable aléatoire T est distribuée suivant une loi appelée loi de Student à n degrés de liberté:

$$T = \frac{U}{\sqrt{V/n}} \sim \mathcal{T}(n)$$

- Les moments de cette loi sont:

$$\mathbb{E}(T) = 0, \qquad \mathbb{V}(T) = n/(n-2)$$

Modèle statistique et modèles de distribution

- C'est souvent l'étape la plus difficile !!! Comment choisir le modèle ?
- La première étape c'est toujours d'étudier les statistiques descriptives et les distributions empiriques pour déterminer quel type de distribution conviendrait
- Comment choisir la distribution ?
  - la nature du caractère étudié (qualitatif, ordinaux, quantitatifs)
  - les connaissances que l'on a du phénomène (valeurs supports)
  - la taille de l'échantillon
- Le modèle comporte souvent des hypothèses (notamment l'indépendance) qu'il est impératif de vérifier
- Les paramètres d'un modèle doivent toujours être interprétés.

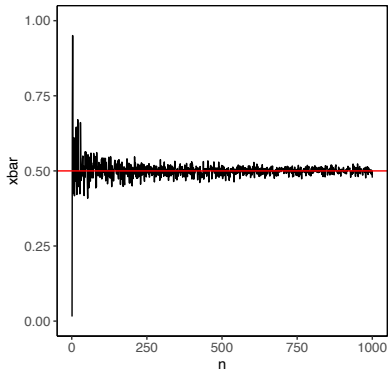
Outline

Loi des grands nombres

- C'est un des théorèmes limites fondamentaux en statistique et probabilité (18e siècle)
- Il s'intéresse à la convergence de la moyenne empirique quand le nombre d'observations est infini

$$\overline{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Le théorème établit que la moyenne empirique converge vers l'espérance
- C'est le théorème qui pose les fondements de l'estimation de l'espérance



$X_i \sim \mathcal{U}([0,1]), \quad \mu = 1/2$

Notion de concentration

- Soit $X_1, \ldots, X_n$ une suite de variables aléatoires iid d'espérance μ, alors la moyenne empirique $\overline{\mathbf{X}}_n$ converge en probabilité vers cette espérance t.q.

$$\forall \epsilon > 0, \quad \mathbb{P}\Big(\mid \overline{\mathbf{X}}_n - \mu \mid > \epsilon \Big) \underset{\infty}{\to} 0$$

- La moyenne empirique converge vers l'espérance théorique (loi faible des grands nombres)
- Justifie l'estimation du paramètre d'espérance par la moyenne empirique quelque soit la loi. Exemples : loi uniforme, loi de Bernoulli
- C'est un premier type d'estimateur, estimateur de type moment qui consiste à identifier les moments empiriques et les moments théoriques
- On parlera d'estimateur convergent (consistent en anglais)

Théorème de la limite centrale

- Si la LGN dit vers quelle quantité $\overline{\mathbf{X}}_n$ converge, elle ne dit pas à quel régime
- Le TCL donne la distribution asymptotique de la moyenne empirique
- Cette loi asymptotique ne dépend pas de la loi des observations X_i !
- Rappels: si

$$\mathbb{E}(X_i) = \mu, \quad \mathbb{V}(X_i) = \sigma^2$$

alors

$$\mathbb{E}(\overline{\mathbf{X}}_n) = \mu, \quad \mathbb{V}(\overline{\mathbf{X}}_n) = \sigma^2/n$$

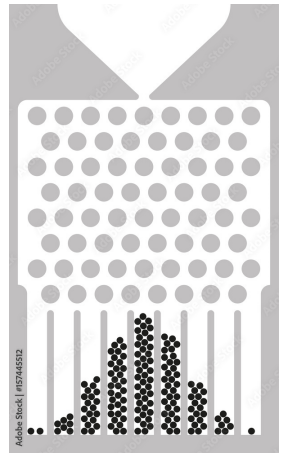


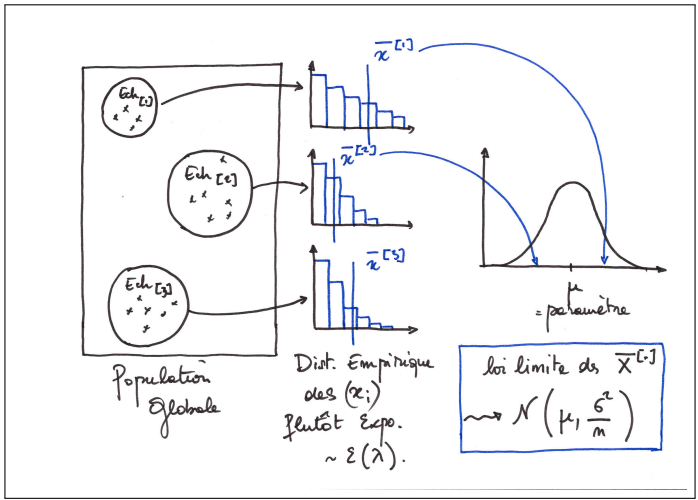
Planche de Galton

Le Théorème Limite Centrale : énoncé

- Si les X_i sont des variables aléatoires iid, si leur espérance μ, et leur variance σ^2 existent alors
$$\frac{\overline{\mathbf{X}}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1)$$
- On suppose que σ^2 est connue.
- Interprétation: même si les observations ne sont pas modélisées par une loi Gaussienne, l'estimateur de l'espérance fondé sur $\overline{\mathbf{X}}_n$ pourra être considéré comme gaussien si le nombre d'observations est suffisant.

Utiliser les moments empiriques + le TLC permet de construire un estimateur et d'avoir sa loi asymptotique !

Le Théorème Limite Central



Outline

Notion d'inférence statistique

- Définition (Académie) 15e siècle. Emprunté du latin *inferre*, pris au sens de "produire un raisonnement, une conclusion".
- C'est l'essence même de la statistique: produire un raisonnement sur une population à partir d'un échantillon
- L'inférence comporte en général deux étapes:

1. Estimation des paramètres : comment utiliser au mieux l'information contenue dans l'échantillon pour renseigner sur les paramètres d'un modèle
2. Tests : une fois les paramètres estimés, comment prendre une décision en quantifiant et contrôlant un risque ?

Estimation ponctuelle

- Il existe plusieurs techniques d'estimation: moments empiriques, moindre-carrés, maximum de vraisemblance
- La démarche est toujours la même:

1 on observe un échantillon $(x_1, \ldots, x_n)$ qui provient d'une population d'intérêt

2 on modélise ces observations par des variables aléatoires iid $(X_1, \ldots, X_n)$ qui suivent une certaine loi F_θ qui dépend d'un paramètre θ:

$$X_i \sim \mathcal{B}(p), \quad \theta = p$$

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad \theta = (\mu, \sigma)$$

3 à partir d'une statistique des observations, on construit un estimateur de θ, noté $\widehat{\theta}(\mathbf{X})$

Le modèle binomial

- Retour sur la fréquence des hétérozygotes dans une population

```
"AA"  "Aa"  "Aa"  "AA"  "Aa"  "aA"
```

- On note x_i la variable qui vaut 1 si l'individu i dans la population 1 est hétérozygote, 0 sinon. On suppose que x_i est la réalisation d'une variable aléatoire X_i.
- Les individus étant non apparentés, on suppose que les X_i sont indépendants
- On suppose que les X_i sont identiquement distribués de loi Binomiale de paramètre p:

$$\forall i \in \{1, \ldots, n\},\ X_i \sim \mathcal{B}(p)$$

- Le paramètre p caractérise la fréquence théorique des hétérozygotes dans la population générale

Estimation d'une proportion

- Un estimateur naturel de la proportion d'hétérozygotes est la proportion empirique
- On note $\widehat{p}(\mathbf{X})$ l'estimateur de cette proportion.
- C'est une variable aléatoire qui dépend du modèle mis sur $\mathbf{X}$.
- Un estimateur de type moment:

$$\widehat{p}(\mathbf{X}) = \overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- L'espérance et la variance de l'estimateur:

$$\mathbb{E}\Big(\widehat{p}(\mathbf{X})\Big) = p, \quad \mathbb{V}\Big(\widehat{p}(\mathbf{X})\Big) = \frac{p(1-p)}{n}$$

- La réalisation de l'estimateur $\widehat{p}(\mathbf{X})$ dépend des observations $\mathbf{x}$ et est appelée estimation:

$$\widehat{p}(\mathbf{x}) = 4/6$$

Un estimateur est une variable aléaoire

Ech	1	2	3	4	5	6	$\widehat{p}$
1	"AA"	"Aa"	"Aa"	"AA"	"Aa"	"aA"	4/6
2	"AA"	"AA"	"AA"	"AA"	"AA"	"AA"	0/6
3	"Aa"	"Aa"	"Aa"	"AA"	"Aa"	"aA"	5/6
⋮							

- Si on répétait la mesure sur d'autres échantillons de la même population, alors on obtiendrait des estimations différentes.
- Pour un échantillon donné, nous n'avons qu'une réalisation de l'estimateur

Estimation d'une proportion: asymptotique

- Grâce à la loi des grands nombres, on sait que la proportion est bien estimée par la fréquence empirique

$$\widehat{p}(\mathbf{X}) \to p$$

- Grâce au théorème central limite on connait la loi asymptotique de l'estimateur de la proportion

$$\mathbb{E}\Big(\widehat{p}(\mathbf{X})\Big) = p, \quad \mathbb{V}\Big(\widehat{p}(\mathbf{X})\Big) = \frac{p(1-p)}{n}$$

$$\frac{\widehat{p}(\mathbf{X}) - p}{\sqrt{p(1-p)/n}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1)$$

Le modèle gaussien

- On note x_i la taille de l'arbre i mesuré dans une forêt. On suppose que x_i est la réalisation d'une variable aléatoire X_i.

  23.4 24.4 24.6 24.9 25 26.2 26.3 26.8 26.8 26.9 27

- On suppose que les X_i sont iid t.q.

$$\forall i \in \{1, \dots, n\},\ X_i \sim \mathcal{N}(\mu, \sigma^2)$$

- On note $\widehat{\mu}(\mathbf{X})$ l'estimateur de l'espérance. Un estimateur de type moment:

$$\widehat{\mu}(\mathbf{X}) = \overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- L'espérance et la variance de l'estimateur:

$$\mathbb{E}\Big(\widehat{\mu}(\mathbf{X})\Big) = \mu, \quad \mathbb{V}\Big(\widehat{\mu}(\mathbf{X})\Big) = \frac{\sigma^2}{n}$$

- L'estimation de l'espérance est $\widehat{\mu}(\mathbf{x}) = 25.66$

Un estimateur est une variable aléaoire

Ech	1	2	3	4	5	6	7	8	9	10	$\widehat{\mu}$
1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	25.53
2	25.51	24.14	28.16	26.42	22.40	24.71	23.04	26.26	24.78	26.03	25.14
3	27.02	24.90	24.57	25.68	26.87	29.23	24.91	24.87	24.41	27.44	25.99
4	25.80	24.76	25.82	26.97	26.71	25.74	25.18	25.90	24.00	26.59	25.75
⋮											

- Si on répétait la mesure sur 10 autres arbres du même type de forêt, alors on obtiendrait des estimations différentes.
- Pour un échantillon donné, nous n'avons qu'une réalisation de l'estimateur

Estimateur de l'espérance dans le cas Gaussien

- On observe $\mathbf{x}$ et on suppose qu'il constitue une réalisation d'un n-echantillon iid $\mathbf{X}$, avec

$$X_i \sim \mathcal{N}(\mu, \sigma^2)$$

- Un estimateur de l'espérance est donné par:

$$\widehat{\mu}(\mathbf{X}) = \overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- L'espérance de l'estimateur de l'espérance:

$$\mathbb{E}\Big(\widehat{\mu}(\mathbf{X})\Big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}(n \times \mu) = \mu$$

- La variance de l'estimateur de l'espérance (avec l'hypothèse iid)

$$\mathbb{V}\Big(\widehat{\mu}(\mathbf{X})\Big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) = \frac{1}{n^2}(n \times \sigma^2) = \frac{\sigma^2}{n}$$

Estimateur de la variance dans le cas Gaussien

- On peut estimer σ^2 par la variance empirique

$$\widehat{\sigma}^2(\mathbf{X}) = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}}_n)^2$$

- C'est un estimateur biasé:

$$\mathbb{E}\Big(\widehat{\sigma}^2(\mathbf{X})\Big) = \frac{n-1}{n}\sigma^2$$

- On peut considérer un autre estimateur (sans biais)

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i}(X_i - \overline{\mathbf{X}}_n)^2$$

- La loi de $S_{n-1}^2(\mathbf{X})$ est une loi du chi2 (somme de Gaussiennes iid au carré)

$$\frac{(n-1)}{\sigma^2}S_{n-1}^2(\mathbf{X}) \sim \chi^2(n-1)$$

Paramètre / Estimateur / Estimation

- Les observations x_i sont modélisées par des variables aléatoires X_i de loi F_θ
- θ est le paramètre de la loi qui décrit la population théorique
- $\widehat{\theta}(\mathbf{X})$ est un estimateur de θ. C'est une statistique qui est une fonction de $\mathbf{X}$
- $\widehat{\theta}(\mathbf{x})$ est une estimation de θ: sa valeur dépend de l'échantillon $\mathbf{x}$
- $\widehat{\theta}(\mathbf{X})$ est une variable aléatoire dont la loi est définie par le modèle F_θ
- On peut donc caractériser un estimateur par son espérance $\mathbb{E}\big(\widehat{\theta}(\mathbf{X})\big)$ et sa variance $\mathbb{V}\big(\widehat{\theta}(\mathbf{X})\big)$
- Les estimateurs de l'espérance et de la variance reposent en général sur les statistiques:

$$\sum_{i=1}^{n} X_i, \quad \sum_{i=1}^{n} X_i^2$$

- La loi asymptotique des estimateurs est donnée par le TCL

Pourquoi s'intéresse-t-on à la loi de l'estimateur ?

- Si on ne disposait pas d'un modèle, on pourrait quand même accéder à $\overline{x}$
- Intuitivement on se demande toujours si cette estimation est précise ou non.
- Exemple: si le taux d'incidence de la grippe dépasse un pourcentage, on lance une campagne de vaccination
- L'idée sous jacente est de s'interroger sur le nombre d'individus nécessaires pour obtenir une estimation précise

Qu'est ce qu'un bon estimateur?

- On considère un paramètre θ qui caractérise une loi de probabilité F_θ
- On observe $\mathbf{x}$ une réalisation de $\mathbf{X}$ et on suppose que $X_i \sim F_\theta$
- On construit un estimateur $\widehat{\theta}(\mathbf{X})$ de θ
- On définit le **bias** et la **variance** de l'estimateur :

$$\mathbb{B}\Big(\widehat{\theta}(\mathbf{X})\Big) = \mathbb{E}\Big(\widehat{\theta}(\mathbf{X})\Big) - \theta, \qquad \mathbb{V}\Big(\widehat{\theta}(\mathbf{X})\Big) = \mathbb{E}\left(\Big[\widehat{\theta}(\mathbf{X}) - \mathbb{E}\Big(\widehat{\theta}(\mathbf{X})\Big)\Big]^2\right)$$

- On peut également définir les carrés moyens comme un **Risque** pour un estimateur:

$$\mathbb{R}\Big(\widehat{\theta}(\mathbf{X})\Big) = \mathbb{E}\Big(\widehat{\theta}(\mathbf{X}) - \theta\Big)^2 = \mathbb{B}^2\Big(\widehat{\theta}(\mathbf{X})\Big) + \mathbb{V}\Big(\widehat{\theta}(\mathbf{X})\Big)$$

- Un bon estimateur sera caractérisé par un faible risque (biais faible ou nul et variance minimale).

Outline

Limite de l'estimation ponctuelle

- Les données récoltées permettent d'obtenir *une* estimation d'un paramètre (par exemple une estimation de l'espérance de la taille d'individus).
- On sait que la valeur observée de l'estimateur change en fonction des données.
- Cette estimation s'appelle *estimation ponctuelle*. Quelles conclusions peut-on en tirer ?
- On préfère raisonner en terme d'intervalle avec un risque l'intervalle ne recouvre pas le "vrai" paramètre
- On cherche un ensemble de valeurs que l'on peut raisonnablement attribuer au paramètre.

Construction de l'intervalle (1)

- Considérons l'estimation du paramètre d'espérance μ d'une loi gaussienne à partir d'un échantillon $\mathbf{X}$.
- On souhaite trouver un intervalle IC($\mathbf{X}$) qui renseigne sur la confiance que les données apportent à la procédure d'estimation
- Cet intervalle dépend des données, et donc du modèle. Il est aléatoire:

$$\text{IC}(\mathbf{X}) = [A(\mathbf{X}), B(\mathbf{X})]$$

- Le paramètre μ quant à lui est fixé !
- Petite coquetterie de notation: On souhaite trouver l'intervalle qui recouvre μ et pas μ qui varie dans un intervalle car μ est fixé

$$\mu \supset \text{IC}(\mathbf{X}) \quad \text{plutôt que} \quad \mu \in \text{IC}(\mathbf{X}).$$

Construction de l'intervalle (2)

- Cet intervalle est aléatoire, on peut donc quantifier la probabilité que l'intervalle encadre le paramètre μ au vu des données:

$$\mathbb{P}\Big\{\mu \supset \mathsf{IC}(\mathbf{X})\Big\}$$

- On introduit le risque α que l'intervalle ne recouvre pas le paramètre
- On introduit le niveau de confiance $1-\alpha$
- On souhaiterait ensuite controler la probabilité de recouvrement à un certain niveau, t.q.

$$\mathbb{P}\Big\{\mu \supset \mathsf{IC}_{1-\alpha}(\mathbf{X})\Big\} = 1-\alpha$$

- $\mathsf{IC}_{1-\alpha}(\mathbf{X})$ s'appelle l'intervalle de confiance de niveau $1-\alpha$ ou de risque α

Vers l'intervalle de confiance d'une espérance

- Pour déterminer les bornes de l'intervalle, on considère la loi de l'estimateur (σ connue):

$$\frac{\widehat{\mu}(\mathbf{X}) - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

- u_α le quantile d'ordre α de la loi gaussienne centrée réduite.
- On utilise les quantiles pour construire l'intervalle de dispersion de l'estimateur:

$$\mathbb{P}\left\{\frac{\widehat{\mu}(\mathbf{X}) - \mu}{\sigma/\sqrt{n}} \supset [u_{\alpha/2}; u_{1-\alpha/2}]\right\} = 1 - \alpha$$

- L'intervalle de confiance de μ est construit à partir de l'intervalle de dispersion de son estimateur $\widehat{\mu}(\mathbf{X})$.

Intervalle de confiance d'une espérance (σ connue)

- On peut donc construire un intervalle de confiance de μ:

$$\widehat{\mu}(\mathbf{X}) - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \widehat{\mu}(\mathbf{X}) - u_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- Pour les **distributions symétriques**: $u_{\alpha/2} = -u_{1-\alpha/2}$.
- Les bornes de l'intervalle sont donc:

$$\begin{aligned} A(\mathbf{X}) &= \widehat{\mu}(\mathbf{X}) - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \\ B(\mathbf{X}) &= \widehat{\mu}(\mathbf{X}) + u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Quelle confiance pour de l'intervalle de confiance ?

- La taille de l'intervalle de confiance reflète la confiance qu'on peut accorder à un estimateur

$$\mathrm{IC}_{1-\alpha}(\mathbf{X};\mu) = \left[\widehat{\mu}(\mathbf{X}) - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \widehat{\mu}(\mathbf{X}) + u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right]$$

- Elle dépend de la dispersion attendue des observations σ
- Elle dépend de la masse des queues de distribution prévue par le modèle $u_{1-\alpha/2} = 1.96$
- Elle dépend du nombre d'observations au régime $1/\sqrt{n}$: si on veut augmenter la précision d'un facteur 10, on doit augmenter n d'un facteur 100!
- Si on doit estimer en plus le paramètre de variance, il faut prendre en compte cette étape

Intervalle de confiance d'une espérance (σ inconnue)

- Si la variance est inconnue, on doit l'estimer avec

$$\widehat{\sigma}^2(\mathbf{X}) = S^2_{n-1}(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{\mathbf{X}}_n\right)^2$$

- La loi de cet estimateur est une loi $\sigma^2\chi^2(n-1)$
- Les fluctuations de l'estimateur autour de son espérance n'ont plus la même loi:

$$\frac{\widehat{\mu}(\mathbf{X}) - \mu}{\widehat{\sigma}(\mathbf{X})/\sqrt{n}} \sim \mathcal{T}(n-1)$$

- Les bornes de l'intervalle utilisent donc les quantiles de la loi de Student :

$$\begin{aligned} A(\mathbf{X}) &= \widehat{\mu}(\mathbf{X}) - t_{1-\alpha/2,n-1} \times \frac{\widehat{\sigma}(\mathbf{X})}{\sqrt{n}} \\ B(\mathbf{X}) &= \widehat{\mu}(\mathbf{X}) + t_{1-\alpha/2,n-1} \times \frac{\widehat{\sigma}(\mathbf{X})}{\sqrt{n}} \end{aligned}$$

Application aux tailles d'arbres

Type 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	27
Type 2	22.5	22.9	23.74	24.0	24.4	24.5	25.3	26	26.4	26.7	.

	Moyenne	Ecart-type	nb obs	$t_{1-\alpha/2,n-1}$	$A(\mathbf{X})$	$B(\mathbf{X})$
Type 1	25.66	1.24	11	2.23	24.82	26.49
Type 2	24.64	1.43	10	2.26	23.62	25.66

- On obtient deux intervalles de confiance au niveau $\alpha = 5\%$

$$\mathrm{IC}_{1-\alpha}(\mathbf{x}; \mu_1) = [24.82, 26.49] \qquad \mathrm{IC}_{1-\alpha}(\mathbf{x}; \mu_2) = [23.62, 25.66]$$

- Est ce que :

$$\mathbb{P}\{\mu_1 \supset [24.82, 26.46]\} = 1 - \alpha \,?$$

Intervalle de confiance pour une proportion

- Dans le modèle de Bernoulli, pour avoir accès aux quantiles, on utilise en général l'approximation gaussienne grâce au TCL:

$$\frac{\widehat{p}(\mathbf{X}) - p}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0,1)$$

- Cette approximation donne les bornes de l'intervalle de confiance d'une proportion tel que:

$$\mathbb{P}\Big\{ p \supset [A(\mathbf{X}), B(\mathbf{X})] \Big\} = 1 - \alpha$$

- Avec l'approximation gaussienne, on a:

$$\begin{aligned} A(\mathbf{X}) &= \widehat{p}(\mathbf{X}) - u_{1-\alpha/2} \sqrt{\frac{\widehat{p}(\mathbf{X})(1-\widehat{p}(\mathbf{X}))}{n}} \\ B(\mathbf{X}) &= \widehat{p}(\mathbf{X}) + u_{1-\alpha/2} \sqrt{\frac{\widehat{p}(\mathbf{X})(1-\widehat{p}(\mathbf{X}))}{n}} \end{aligned}$$

- On a estimé la variance de l'estimateur par la variance empirique