Eléments de statistique
Leçon 2 - Notion d'échantillon

Louis Wehenkel

Département d'Electricité, Electronique et Informatique - Université de Liège
B24/II.93 - L.Wehenkel@ulg.ac.be

MATH0487-1 : 3BacIng, 3BacInf - 25/9/2018

Find slides: http://montefiore.ulg.ac.be/∼lwh/Stats/

Objectifs de la statistique inférentielle

- La statistique descriptive fournit des outils permettant résumer les informations contenues dans un ensemble de données recueillies à partir d'une 'expérience'.
- L'objectif de la statistique inférentielle est de tirer à partir de telles données des hypothèses plausibles sur la nature générale de l'expérience
  - Par exemple, à partir d'une BD médicale, on souhaite déterminer les symptômes les plus discriminants d'une maladie.
  - A partir de la mesure des durées de vie d'un lot de pièces, on souhaite déterminer les plages de variation des durées de vie des autres pièces qui seront fabriquées dans le futur, ou bien détecter des dérives dans la chaîne de fabrication nécessitant une maintenance.
- De manière générique, l'inférence statistique consiste à généraliser des informations obtenues lors d'une série d'expériences à un ensemble de situations analogues mais différentes.

Structure de donnée de base en statistique inférentielle

- On dispose d'un tableau de données $\mathbf{D} \in \mathbb{R}^{n \times p}$, dont les n lignes sont appelées des 'individus', et les p colonnes sont des variables utilisées pour mesurer les caractéristiques des individus.
  - NB: en pratique, on a souvent que n et/ou p sont grands (p.ex. plusieurs centaines, ou milliers).
  - Les variables peuvent être numériques ou bien catégorielles
    - Une variable catégorielle m-aire peut prendre m valeurs mutuellement exclusives (p.ex. "sain" vs "malade");
    - on les code cependant souvent sous forme numérique (p.ex. 0, 1), même si elles ne représentent pas une grandeur quantitative.
  - Notations/terminologie: on appelle $\mathbf{D}$ l'échantillon.
    - élément (i, j) de $\mathbf{D}$: $\mathbf{D}_{i,j}$,
    - i-ème ligne de $\mathbf{D}$: $\mathbf{D}_{i,.}$,
    - j-ème colonne de $\mathbf{D}$: $\mathbf{D}_{.,j}$,

Exemple de tableau **D**

- Les résultats obtenus au cours de Probas en juin 2012.
- Individus: les students (ici $n = 182$)
- Variables: les cotes obtenues aux différents travaux et questions de l'examen (ici $p = 10$ variables numériques (entières))
- Exemples de questions d'inférence statistique
  - Après avoir corrigé les 20 premières copies, que peut on dire de la valeur probable des moyennes sur l'ensemble du tableau ?
  - Après avoir corrigé les 182 copies, que peut on dire de la difficulté intrinsèque des différentes questions, c'est-à-dire, au delà de l'échantillon d'étudiants sondés en juin 2012 ?
  - Si on pouvait répéter l'expérience, avec 100 autres étudiants dans les mêmes conditions, quel serait le taux de réussite ?

Objectifs pédagogiques

- Le but de cette partie du cours est de formaliser mathématiquement la notion d'échantillon **D**.
- On se servira du modèle probabiliste $(\Omega, \mathcal{E}, P)$ pour définir la notion d'échantillon i.i.d. (ainsi que des variantes de ce modèle canonique).
- On mettra en oeuvre le calcul de probabilités pour caractériser les propriétés des statistiques calculées à partir d'un échantillon (aussi appelées distributions d'échantillonnage).
- On discutera les techniques dites de 'contrôle (ou design) d'expérience' visant à optimiser certaines de ces propriétés, quand le nombre d'individus n est limité (par le temps, ou le coût d'acquisition de données).

Modèle probabiliste de l'échantillon 'i.i.d'

Pour simplifier, supposons pour le moment que $p = 1$ (échantillon uni-dimensionnel), et que la (seule) variable mesurée est à valeurs numériques.

Algorithme i.i.d. $[(\Omega, \mathcal{E}, P), \mathcal{X}, n]$:

- On suppose fixé un espace de probabilité $(\Omega, \mathcal{E}, P)$ et une v.a. $\mathcal{X}$ définie sur cet espace. On fixe a priori la taille n de l'échantillon.
  (NB: Ω peut être fini, infini-dénombrable, ou non-dénombrable; $\mathcal{X}$ peut-être discrète, continue, ou mixte.)
- Les n lignes de $\mathbf{D}_n$ sont obtenues au moyen d'une expérience aléatoire "répétée", comme suit
  - L'obtention de la première ligne du tableau, consiste à observer la valeur de $\mathcal{X}$ pour un élément ω de Ω pris au hasard selon la loi P.
  - L'obtention de la seconde ligne du tableau, consiste à répéter l'expérience une seconde fois, indépendamment du résultat de la première
  - ...
  - l'obtention de la n-ième ligne du tableau, consiste à répéter l'expérience une n-ième fois, indépendamment du résultat des $n-1$ premières.

Modèle probabiliste de l'échantillon 'i.i.d'

- Les n valeurs observées dans $\mathbf{D}_n$ sont par conséquent des réalisations de n variables aléatoires, indépendantes et identiquement distribuées selon la même loi que la variable $\mathcal{X}$.
- Si $F_{\mathcal{X}}(\cdot)$ désigne la fonction de répartition (uni-dimensionnelle) de $\mathcal{X}$, la fonction de répartition conjointe du vecteur de n-valeurs de l'échantillon aléatoire $\boldsymbol{\mathcal{D}}_n$ est donc

$$F_{\boldsymbol{\mathcal{D}}_n}(x_1, \ldots, x_n) = F_{\mathcal{X}}(x_1) F_{\mathcal{X}}(x_2) \cdots F_{\mathcal{X}}(x_n) \qquad (1)$$

- Si $\mathcal{X}$ est continue de densité $f_{\mathcal{X}}(\cdot)$, le vecteur $\boldsymbol{\mathcal{D}}$ l'est aussi, et dans ce cas sa densité s'écrit

$$f_{\boldsymbol{\mathcal{D}}_n}(x_1, \ldots, x_n) = f_{\mathcal{X}}(x_1) f_{\mathcal{X}}(x_2) \cdots f_{\mathcal{X}}(x_n) \qquad (2)$$

- Remarque: l'hypothèse i.i.d. implique que $F_{\mathcal{D}_n}$ et $f_{\mathcal{D}_n}$ restent invariantes si on permute les lignes du tableau.

Distributions d'échantillonnage de statistiques uni-variées

- Le vecteur des valeurs observées dans le tableau $\mathbf{D}_n$ peut donc être vu comme une réalisation de l'expérience aléatoire modélisée par l'espace $(\Omega_{\mathcal{D}_n}, \mathcal{E}_{\mathcal{D}_n}, P_{\mathcal{D}_n})$, espace produit de $(\Omega_{\mathcal{X}}, \mathcal{E}_{\mathcal{X}}, P_{\mathcal{X}})$ n-fois avec lui-même. Voir sections 2.1.1 et 2.1.2 des notes du cours de probabilités.
- On appelle statistique univariée, toute variable aléatoire à valeurs réelles définie génériquement (i.e. pour toute valeur de $n \in \mathbb{N}_0$) sur $(\Omega_{\mathcal{D}_n}, \mathcal{E}_{\mathcal{D}_n}, P_{\mathcal{D}_n})$.
- Une statistique univariée définit par conséquent une suite de variables aléatoires à valeurs réelles.
- Nous allons introduire les principales statistiques univariées, et étudier comment leurs lois de probabilité évoluent en fonction de la valeur de n et de la loi de la variable parente $\mathcal{X}$.

Distribution d'échantillonnage de la moyenne d'échantillon

Il s'agit de la variable aléatoire définie par $m_x(\mathbf{D}_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$, où les x_i sont les n valeurs de la v.a. $\mathcal{X}$ observées dans $\mathbf{D}_n$.

On a les résultats suivants : (sous l'hypothèse $V\{\mathcal{X}\} < \infty$; explications au cours oral)

- $E\{m_x\} = \mu_{\mathcal{X}}$
- $V\{m_x\} = \frac{\sigma^2_{\mathcal{X}}}{n}$
- Loi forte des grands nombres: $m_x(\mathbf{D}_n) \xrightarrow{p.s.} \mu_{\mathcal{X}}$
- Théorème central limite: $\frac{m_x(\mathbf{D}_n)-\mu_{\mathcal{X}}}{\sigma_{\mathcal{X}}/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0;1)$

HOMEWORK: Montrer que pour toute valeur de n, et $\forall \epsilon > 0$, l'inégalité de Bienaymé-Tchebyshev permet donc de construire un intervalle autour de $\mu_{\mathcal{X}}$ qui contiendra avec une probabilité $1-\epsilon$ la valeur de $m_x(\mathbf{D}_n)$. La largeur de cet intervalle décroit en $1/\sqrt{n}$ lorsque n croit. Calculer son expression en fonction de $\mu_{\mathcal{X}}, \sigma_{\mathcal{X}}, n$, et ϵ.

Cas où $\mathcal{X}$ est une variable de Bernoulli

Rappel: $\mathcal{X}$ est une variable de Bernoulli si sa valeur vaut 1 avec la probabilité p_1 et 0 avec la probabilité $p_0 = 1 - p_1$.

- Dans ce cas $m_x(\mathbf{D}_n)$ vaut la proportion $f_1(\mathbf{D}_n) = n_1/n$ de fois qu'on observe la valeur $\mathcal{X} = 1$ dans $\mathbf{D}_n$.
- On a $E\{f_1\} = p_1$ et $V\{f_1\} = \frac{p_1(1-p_1)}{n}$.
- $f_1(\mathbf{D}_n) \xrightarrow{p.s.} p_1$
- Si n est suffisamment grand: $f_1 \sim \mathcal{N}(p_1; \sqrt{\frac{p_1(1-p_1)}{n}})$

  En pratique si $\min(np_1, np_0) > 5$.
- Interprétation sur un exemple, voir slide suivant

Exemple : lancer de deux dés

- L'expérience parente est un lancer simultané de deux dés
  (A ne pas confondre avec un double lancer d'un même dé.)
- La variable $\mathcal{X}$ est la somme des deux faces supérieures.
  On a $\mathcal{X}(\Omega) \triangleq \Omega_{\mathcal{X}} = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
- On ne sait rien quant au fait que les dés sont pipés ou non.
- A la place, on dispose d'une table comprenant $n = 20$ valeurs de $\mathcal{X}$, obtenues en effectuant 20 lancers simultanés des deux dés et en reportant chaque fois la somme des valeurs obtenues.
- On aimerait bien exploiter ces données pour prédire la probabilité d'obtenir une valeur $\mathcal{X} < 8$ au lancer suivant.

Distribution d'échantillonnage de la fonction de répartition

Pour rappel : $\hat{F}_x(y) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i < y)$.

- Pour une valeur de y fixée, $\hat{F}_x(y)$ définit une statistique univariée, qui indique la proportion des valeurs de l'échantillon qui sont strictement inférieures à y.
- La variable $\mathcal{Z} \stackrel{\triangle}{=} 1(\mathcal{X} < y)$ est une variable de Bernoulli, avec $p_1 = F_{\mathcal{X}}(y)$.
- On a $\hat{F}_x(y) = \frac{1}{n}\sum_{i=1}^{n} z_i$. NB: Les n copies de $\mathcal{Z}$ construites à partir des n copies de $\mathcal{X}$ sont indépendantes, puisque les $\mathcal{X}$ le sont par hypothèse.
- Par conséquent: $\forall y : \hat{F}_x(y)(\mathbf{D}_n) \xrightarrow{p.s.} F_{\mathcal{X}}(y)$. (Convergence ponctuelle)
- De plus on a le Théorème de Glivenko-Cantelli :

$$D_{n,\mathcal{X}}^{KS} = \sup_y \left| \hat{F}_x(y)(\mathbf{D}_n) - F_{\mathcal{X}}(y) \right| \xrightarrow{p.s.} 0$$

(Convergence uniforme)

Distance de Kolmogorov-Smirnov

(NB: c'est à illustrer par des graphiques)

- La distance de Kolmogorov-Smirnov entre deux fonctions de répartition $F_{\mathcal{X}}(\cdot)$ et $F_{\mathcal{Y}}(\cdot)$ est définie par

$$D^{KS}(F_{\mathcal{X}}, F_{\mathcal{Y}}) \stackrel{\triangle}{=} \sup_{z \in \mathbb{R}} |F_{\mathcal{X}}(z) - F_{\mathcal{Y}}(z)| \leq 1.$$

- La statistique de Kolmogorov-Smirnov d'une v.a. obtenue à partir d'un échantillon $\mathbf{D}_n$ de taille n est la distance de Kolmogorov-Smirnov entre la fonction de répartition empirique déduite de l'échantillon et la fonction de répartition de la variable parente $\mathcal{X}$. C'est une v.a. dénotée par $D^{KS}_{n,\mathcal{X}}$.

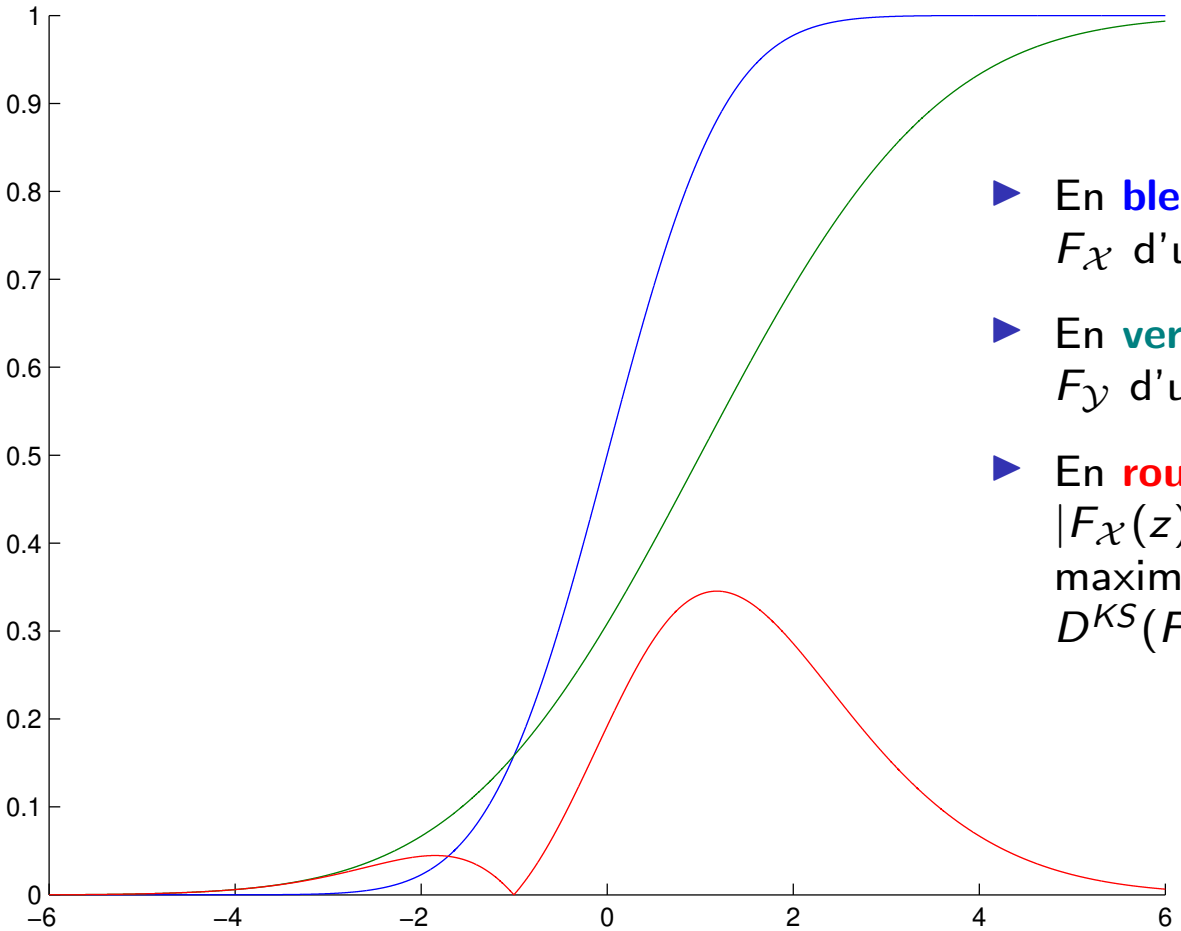
- On peut se convaincre aisément que si $\mathcal{Y}$ est une fonction strictement croissante de $\mathcal{X}$, alors $D^{KS}_{n,\mathcal{Y}} = D^{KS}_{n,\mathcal{X}}$. C'est aussi le cas si $\mathcal{Y}$ est une fonction strictement déccroissante de $\mathcal{X}$.

  Homework : dessiner une fonction de répartition quelconque d'une v.a. $\mathcal{X}$, puis en déduire le graphique de la fonction de répartition de $\mathcal{Y} \stackrel{\triangle}{=} -\mathcal{X}$.

- On peut en déduire que la distribution d'échantillonnage de $D^{KS}_{n,\mathcal{X}}$ ne dépend en fait pas de l'allure de $F_{\mathcal{X}}$, pour autant que $F_{\mathcal{X}}$ soit continue.

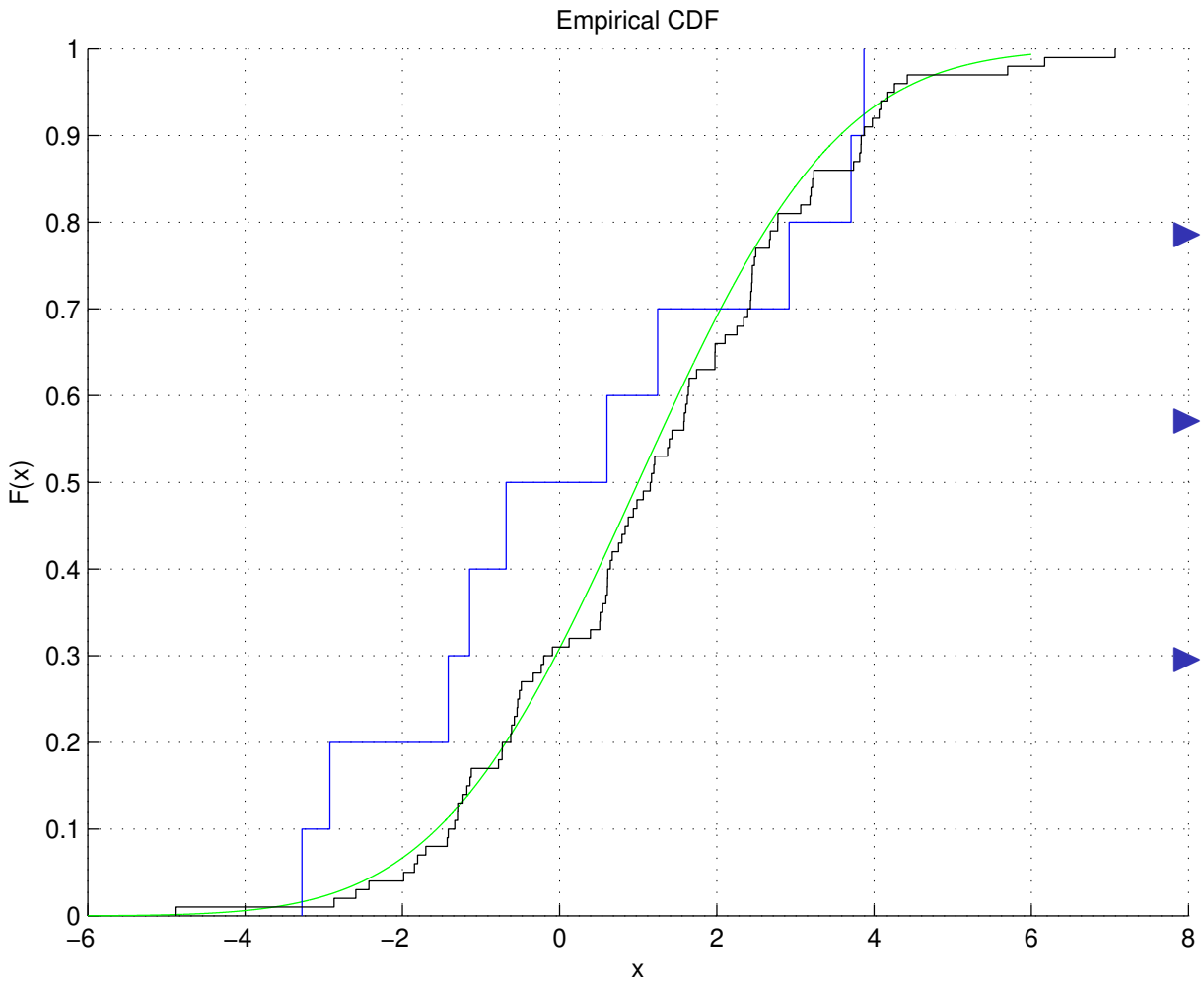
  (ATTENTION: en cas de discontinuités c'est plus complexe !!!!)

Distance de KS entre deux fonctions de répartition



- En **bleu** la fonction de répartition $F_{\mathcal{X}}$ d'une v.a. $\mathcal{N}(0,1)$
- En **vert** la fonction de répartition $F_{\mathcal{Y}}$ d'une v.a. $\mathcal{N}(1,2)$
- En **rouge** la fonction $|F_{\mathcal{X}}(z) - F_{\mathcal{Y}}(z)|$ dont la valeur maximale vaut $D^{KS}(F_{\mathcal{X}}, F_{\mathcal{Y}}) \approx 0.35$.

Convergence de la fonction de répartition empirique



- En **vert** la fonction de répartition $F_{\mathcal{Y}}$ d'une v.a. $\mathcal{N}(1, 2)$
- En **bleu** une fonction de répartition empirique observée pour la même variable avec $n = 10$
- En **noir** une fonction de répartition empirique observée pour la même variable avec $n = 100$

Etude de la statistique s^2 (variance d'échantillon)

Il s'agit de la variable aléatoire définie par $s_x^2(\mathbf{D}_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2$, où les x_i sont les n valeurs de la v.a. $\mathcal{X}$ observées dans $\mathbf{D}_n$.

- On a (évidemment) $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - (m_x)^2$, et par conséquent, lorsque $n \to \infty$, on a

$$E\{s_x^2\} \xrightarrow{p.s.} E\{\mathcal{X}^2\} - (E\{\mathcal{X}\})^2 = \sigma_{\mathcal{X}}^2,$$

  puisque $\frac{1}{n}\sum_{i=1}^{n}(x_i)^2 \xrightarrow{p.s.} E\{\mathcal{X}^2\}$ et que $m_x \xrightarrow{p.s.} E\{\mathcal{X}\}$.

- Cependant, pour une valeur finie de n, on n'a pas $E\{s_x^2\} = \sigma_{\mathcal{X}}^2$! En effet, on a $\forall n \in \mathbb{N}_0$:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathcal{X}})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 + (m_x - \mu_{\mathcal{X}})^2$$

  et donc $E\{s_x^2(\mathbf{D}_n)\} = \sigma_{\mathcal{X}}^2(1 - 1/n)$.

Etude de la statistique s^2 (variance d'échantillon)

Raisonnement détaillé :

- La relation

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathcal{X}})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 + (m_x - \mu_{\mathcal{X}})^2,$$

est un conséquence directe du fait que $\forall a \in \mathbb{R}$ on a

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 + (m_x - a)^2.$$

(Développer $\frac{1}{n}\sum_{i=1}^{n}((x_i - m_x) - (a - m_x))^2$ en tenant compte du fait que $\sum_{i=1}^{n}(x_i - m_x) = 0$.)

- Donc

$$E\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\mathcal{X}})^2\} = E\{\frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2\} + E\{(m_x - \mu_{\mathcal{X}})^2\},$$

c'est-à-dire que

$$\frac{1}{n}n\sigma_{\mathcal{X}}^2 = E\{s_x^2\} + \frac{1}{n}\sigma_{\mathcal{X}}^2.$$

ou encore

$$E\{s_x^2\} = \frac{n-1}{n}\sigma_{\mathcal{X}}^2.$$

Etude de la statistique s_x^2 (variance d'échantillon)

Quelques résultats suplémentaires:

- La statistique s_x^2 sous-estime en moyenne la valeur de $\sigma_{\mathcal{X}}^2$.
- On utilise pour cette raison souvent à la place de s_x^2 la version corrigée (sans biais) dénotée s_{n-1}^2 définie par

$$s_{n-1}^2 \triangleq \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)^2.$$

- s_{n-1}^2 est dit un "estimateur sans biais" de $\sigma_{\mathcal{X}}^2$ car $E\{s_{n-1}^2\} = \sigma_{\mathcal{X}}^2$.

  Certains auteurs utilisent la notation s_n^2 à la place de ce que nous avons désigné par s_x^2; d'autres auteurs réservent la notation s_x^2 à ce que nous avons appelé s_{n-1}^2. Il faut donc un peu se méfier.
- Puisque $E\{\sqrt{\mathcal{X}}\} \leq \sqrt{E\{\mathcal{X}\}}$ (inégalité de Jenssen), la statistique s_{n-1} sous-estime néanmoins la valeur de l'écart-type $\sigma_{\mathcal{X}}$.

Cas où $\mathcal{X}$ est gaussienne

Lorsque la variable parente $\mathcal{X}$ suit une loi gaussienne, on peut montrer plusieurs résultats supplémentaires :

- m_x suit une loi gaussienne $\mathcal{N}(\mu_{\mathcal{X}}; \frac{\sigma_{\mathcal{X}}}{\sqrt{n}})$ quelque soit n.
- $n\frac{s_x^2}{\sigma_{\mathcal{X}}^2}$ suit une loi χ^2_{n-1} ... à expliquer...
- Les statistiques m_x et s_x^2 sont indépendantes (et c'est la seule situation où cela est vrai).
- La variance de s_x^2 vaut $\frac{2(n-1)}{n^2}\sigma_{\mathcal{X}}^4$.

Lois en χ^2: par définition, la loi en χ^2_p (Khi-deux à p degrés de liberté) est la loi d'un v.a. obtenue en calculant la somme des carrés de p variables normales indépendantes, centrées et réduites. Notez que l'on perd un degré de liberté à cause du fait que dans s_x^2 on estime la moyenne à partir de l'échantillon...

Autres paramètres de position et de dispersion

- Les statistiques de position autres que la moyenne (médiane, quartiles, quantiles, valeurs extrêmes) ont des distributions qui peuvent aussi être calculées de façon directe à partir de la connaissance de $F_{\mathcal{X}}$ qui caractérise la variable parente.
- Par exemple, si nous désignons par x_{min} la valeur la plus faible observée dans l'échantillon, on peut montrer que

$$F_{x_{\text{min}}}(y) = 1 - (1 - F_{\mathcal{X}}(y))^n$$

et que

- si x_{max} désigne la valeur la plus élevée observée dans l'échantillon, on a

$$F_{x_{\text{max}}}(y) = (F_{\mathcal{X}}(y))^n.$$

- Voir ouvrages de référence pour l'étude des distributions d'échantillonnage des ces grandeurs, qui relèvent d'applications plus ou moins directes du calcul de probabilités.

Distributions d'échantillonnage de statistiques multi-variées

- Lorsqu'on dispose d'un tableau à p colonnes, le modèle théorique i.i.d. s'entend directement : il faut remplacer dans ce qui précède la v.a. réelle $\mathcal{X}$ par une v.a. vectorielle à valeurs dans $\mathbb{R}^p$, et la fonction de répartition $F_{\mathcal{X}}$ par la fonction de répartition conjointe $F_{\boldsymbol{\mathcal{X}}}$.
- Les résultats 'uni-variés', que nous venons de développer, s'appliquent évidemment à chacune des colonnes prise séparément.
- On peut également, sous différentes hypothèses dériver les distributions d'échantillonnage de statistiques multi-variées (i.e. des grandeurs calculées à partir du tableau $\mathbf{D}_n$ et qui font intervenir les valeurs de plusieurs de ses colonnes).
- Par exemple, sous l'hypothèse où $\boldsymbol{\mathcal{X}} \sim \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$, on peut montrer que le vecteur des moyennes $\boldsymbol{m}_x \sim \mathcal{N}(\boldsymbol{\mu}; \frac{1}{n}\boldsymbol{\Sigma})$. On peut aussi caractériser les lois décrivant l'écart entre le vecteur $\boldsymbol{m}_x$ et le vecteur $\boldsymbol{\mu}_{\boldsymbol{\mathcal{X}}}$.

Autres modèles d'échantillonnage

Je pense qu'à ce stade il est important de clôturer cette leçon en prenant du recul par rapport au modèle i.i.d. en montrant qu'il n'est pas toujours raisonnable (cas des petites populations) ou bien judicieux (en termes d'efficacité).

Les slides suivants, se veulent intuitifs et non formels, car nous reviendront sur la plupart des idées qui y sont introduites.

Tirage sans remise dans une population finie

- Lorsque la population mère est finie (souvent vrai en pratique), on effectue généralement un tirage parmi les individus en évitant de considérer deux fois le même individu.
- Ce modèle est différent du modèle théorique 'i.i.d', mais lorsque $n \ll N$ il peut s'en approcher suffisamment pour négliger le fait que dans un tirage sans remise, les différentes observations ne sont en réalité pas indépendantes.
- Le modèle le plus simple est celui où on tire sans remise dans une urne de taille finie N: à chaque tirage, on choisit au hasard parmi les individus qui restent encore dans l'urne.
- Cela revient à choisir au hasard un des C_N^n sous-ensembles de taille n de la population mère de taille N.
- On montre que la moyenne d'échantillon reste un estimateur sans biais de la moyenne de la population. On montre aussi que la variance de cet estimateur est multipliée par le facteur $(N-n)/(N-1) \in [0;1]$ par rapport au cas du tirage avec remise.

  Réfléchir au sens de cette formule lorsque $n = 1$, et lorsque $n = N$.

Stratification

- La stratification consiste en des tirages séparés effectués dans des sous-populations de la population mère.
- Lorsque les sous-populations sont plus homogènes du point de vue des grandeurs étudiées, la stratification permet d'obtenir des estimations plus précises à nombre égal d'observations.
- Exemple illustratif:
  - Supposons que mon but soit d'estimer la moyenne des cotes de juin 2012 le plus précisément possible, en corrigeant seulement 20 copies (tirées sans remise, parmi les 182 copies). Supposons aussi que je dispose d'une information complémentaire pour chaque étudiant, à savoir son pourcentage global $\mathcal{B}$ obtenu en premier bac?
  - Je propose de tirer 10 étudiants au hasard parmi ceux qui ont eu une moyenne en Bac1 inférieure à la médiane de ces 182 valeurs de $\mathcal{B}$, et 10 étudiants au hasard parmi les autres étudiants, de corriger leurs 20 copies, et d'estimer la moyenne de la classe par la moyenne des 20 cotes obtenues.
  - Je prétends que cette valeur est un estimateur sans biais de la moyenne de la classe, et qu'elle est bien plus précise que la valeur que j'aurai obtenue, si je n'avais pas stratifié mon échantillon.

Précautions à prendre au moment de l'inférence

- Analyser un tableau de données pour estimer des grandeurs de population, sans savoir comment les données ont été collectées consiste à jouer à l'apprenti sorcier.
- Participer activement à la collecte de données est certainement une excellente façon de savoir comment les données ont été collectées. Sinon, il faut interroger ceux qui ont effectué la collecte, et essayer d'en savoir le plus.
- Les connaissances en statistique sont utiles non seulement pour analyser des données, mais aussi pour choisir des plan d'expérience de façon à optimiser la représentativité de l'échantillon en fonction des objectifs poursuivis.

Sondages

- De façon générale, les techniques de sondage on pour but de tirer des individus dans une population concrète, afin d'estimer avec la meilleure précision des paramètres de population d'intérêt.
- Exemple de méthodes relatives au sondage:
  - Tirage à probabilités inégales
    On tire les échantillons selon une loi différente que celle qui définit l'espérance qu'on cherche à estimer, et on corrige l'estimateur. Si la loi de tirage est bien corrélée (positivement) avec la v.a. $\mathcal{X}$ dont on souhaite estimer l'espérance, la variance de l'estimateur est fortement réduite.
  - Stratification
    Cf. ce qui précède. Il est possible d'optimiser la façon de stratifier, si on a une connaissance a priori des variances intra-strates.
  - Sondages par grappes
    On divise la population en 'grappes', on tire des grappes au hasard, et on observe tous les éléments de chaque grappe tirée; il faut corriger l'estimateur, en fonction du rapport entre la taille des grappes et leur probabilité d'être choisie.
  - Redressement a posteriori
    On essaye d'exploiter la dépendance entre la variable cible et une autre variable, afin de mieux exploiter les informations de l'échantillon obtenu par tirage aléatoire simple (avec ou sans remise).

Plans d'expérience

- En pratique (voir plus loin dans ce cours), l'inférence statistique a souvent pour but de d'étudier la relation entre une v.a. cible $\mathcal{Y}$ (appelée réponse), et une ou plusieurs variables explicatives $\mathcal{X}_1, \mathcal{X}_2, \ldots$.
- Par exemple, on souhaite expliquer $\mathcal{Y}$ par un modèle linéaire $f(\mathcal{X}) = \lambda_0 + \sum_{i=1}^{p} \lambda_i \mathcal{X}_i$
- Les plans d'expérience visent dans ce cas à choisir les valeurs des $\mathcal{X}_1, \mathcal{X}_2, \ldots$ pour lesquelles on fera une expérience pour déterminer $\mathcal{Y}$, dans le but d'estimer le plus précisément les paramètres λ_i.
- On reviendra sur ces techniques dans la suite du cours.

Méthode de Monte Carlo

- La méthode de Monte Carlo consiste à appliquer les idées développées dans cette leçon au calcul d'une intégrale multiple d'une fonction de plusieurs variables (souvent un nombre tellement élevé qu'on ne peut même pas espérer utiliser des techniques d'intégration formelle ou numérique usuelles).
- L'idée de base est de faire le parallèle entre une intégrale multiple, et le calcul d'une espérance mathématique, puis de remplacer le calcul de cette espérance mathématique par une moyenne calculée à partir d'un échantillon de valeurs des variables sur lesquelles porte la fonction à intégrer.
- **HOMEWORK :** relire les sections 3.12 (en particulier 3.12.2) et 4.5 (en particulier 4.5.2) du syllabus du cours de probabilités.

Quiz

(Réponses données au cours suivant)

Répondre par 'Vrai' ou 'Faux' aux questions suivantes:

1. Un tirage avec remise implique que les lignes suivantes de $\mathbf{D}_n$ sont indépendantes des lignes précédentes ?
2. Considérons les deux types d'échantillonnage suivants : tirage au sort avec remise et tirage au sort sans remise. La différence entre les deux types d'échantillonnage est d'autant plus grande que la taille de la population est élevée ?
3. La valeur de la médiane est très sensible aux valeurs aberrantes (*outliers*) ?
4. Quand la taille n de l'échantillon augmente, la variance de la moyenne empirique de l'échantillon ($V\{m_x\}$) diminue ?
5. La statistique de Kolmogorov-Smirnov d'une v.a. obtenue à partir d?un échantillon $\mathbf{D}_n$ de taille n permet de comparer la fonction de répartition empirique obtenue à partir de l'échantillon et la fonction de répartition de la variable parente ?
6. L'estimateur s^2_{n-1} est souvent utilisé à la place de s^2_x car c'est un estimateur sans biais de :
   - 6.1 L'écart-type de la population ?
   - 6.2 La variance de la population ?
7. La moyenne empirique m_x d'un échantillon $\mathbf{D}_n$ de taille n suit toujours une loi gaussienne quel que soit n ?


Brett Ryder