Premières notions de statistique: échantillon et distribution empirique

Franck Picard

Licence 3 Biosciences, 2022-2023

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

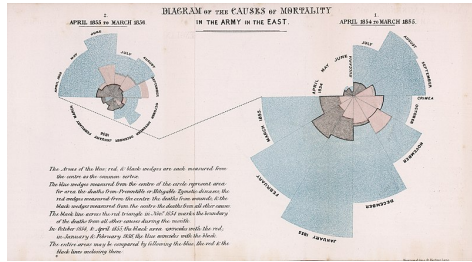
franck.picard@ens-lyon.fr

Pourquoi a-t-on besoin de statistique dans la vie ?

- Etude de phénomènes pouvant être décrits de manières quantitatives ou qualitatives
- Lorsque la totalité de l'information n'est pas disponible, mais uniquement sous la forme d'un échantillon
- La collecte, l'interprétation et la représentation des données constituent le coeur de la discipline statistique
- C'est la science de l'incertitude, de l'étude de la variabilité et des erreurs
- Le cadre mathématique des statistique est la théorie des probabilités

Une perspective historique

- Le terme Statistique émerge au 18e siècle
- Comment obtenir des données chiffrées sur les caractéristiques des états et leur fonctionnement ?
- Comment étudier le cours des céréales et des valeurs commerciales ?
- Etude de la démographie
- Peut on prédire des phénomènes à partir de données chiffrées ?



"Diagram of the causes of mortality in the army in the East" (1858) by F. Nightingale

→ Comment guider la prise de décision sur des critères quantitatifs prenant en compte des erreurs ?

Outline

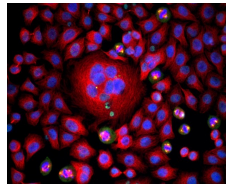
Populations et caractères

- Les études statistiques s'appuient sur des **observations** mesurées sur des **populations** composées d'**individus** sur lesquelles on observe des **caractères**
- La notion d'**individus** devient statistique
- On mesure des **caractères**:
  - Qualitatifs (ni ordonnés ni ajoutés)
  - Ordinaux (ordonnés mais pas ajoutés)
  - Quantitatifs (numériques, discrèts ou continus)

Animaux
- Génotype
- Taille
- Poids
- couleur des yeux

Galaxies
- luminosité
- vitesse
- densité

chromosomes
- longueur
- # de copies

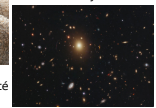


cellules
- taille du noyau
- # de chromosomes
- taille

Quand fait-on de la statistique ?

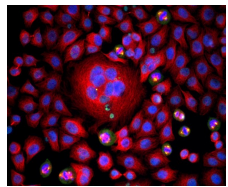
- Quand il est **impossible** ou **inutile** d'observer un caractère sur l'ensemble de la population.
- La **stratégie** statistique consiste à observer les caractères sur une **sous-population** en espérant tirer des observations des conclusions générales à la **population de référence**.
- La première étape consiste donc à identifier cette population d'intérêt (cible)

Animaux
- Génotype
- Taille
- Poids
- couleur des yeux



Galaxies
- luminosité
- vitesse
- densité

chromosomes
- longueur
- # de copies



cellules
- taille du noyau
- # de chromosomes
- taille

La notion d'échantillon

- L'échantillonage consiste à **choisir des individus** de la population générale suivant certaines **contraintes**
- Le résultat de la mesure d'un caractère sur n individus est un n-uplet $(x_1, \ldots, x_n)$, que l'on appelle échantillon de taille n.
- Exemple: $n = 3$ individus, x_i l'âge du $i$ème individu. Après mesure, on dispose d'un 3-échantillon $(x_1, x_2, x_3) = (15, 18, 19)$.
- Plus n est grand, plus on collecte de l'information, mieux on décrira la population d'origine

Première étape de prise en main des données

- Lorsque l'on récolte des données pour les analyser une première étape est de formaliser les données disponibles
- On récolte 500 cocons de Bombyx mori. On en pèse 10 au hasard

Poids	0.64	0.65	0.73	0.60	0.65	0.77	0.82	0.64	0.66	0.72

- Dans cet exemple les données sont quantitatives et univariées, $n = 10$, et on note x_i le poids du cocon i, tel que l'échantillon observé est

$$(x_1 \ldots, x_n) = (0.64, 0.65, \ldots, 0.72).$$

- C'est un vecteur de $\mathbb{R}^n$ (données réelles de taille n)

Données qualitatives

- On génotype 5 individus à deux loci différents

Locus 1	AA	AA	AA	AA	Aa
Locus 2	Bb	bb	Bb	bb	bb

- Dans cet exemple les données sont qualitative et multivariées (deux variables)
- $n = 5$, et on note $x_i = (x_i^1, x_i^2)$ le genotype de l'individu i aux loci 1 et 2.
- L'échantillon observé est

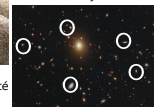
$$(x_1 \dots, x_n) = \begin{bmatrix} AA & AA & AA & AA & Aa \\ Bb & bb & Bb & bb & bb \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 & 2 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

- On peut aussi modéliser les observations par le nombre d'allèles dominants (choix)

Un échantillon est aléatoire

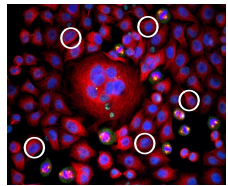
- L'échantillonage étant une procédure aléatoire, un échantillon est par essence aléatoire
- Les résultats issus d'un échantillon sont également des **résultats aléatoires**
- Exemple: on tire $(x_1, x_2, x_3) = (15, 18, 19)$, si on retire un autre 3-échantillon $(x'_1, x'_2, x'_3) = (17, 19, 16)$

Animaux
- Génotype
- Taille
- Poids
- couleur des yeux



Galaxies
- luminosité
- vitesse
- densité

chromosomes
- longueur
- # de copies



cellules
- taille du noyau
- # de chromosomes
- taille

Statistiques descriptives et résumé des données

- Comment définir des indicateurs permettant de synthétiser l'information contenue dans l'échantillon ?
- UNE statistique est une fonction d'un échantillon permettant d'accéder à un certain type d'information.
  - La moyenne renseigne sur la position (barycentre au sens physique)
  - La variance renseigne sur la dispersion autour du barycentre
  - Les quantiles renseignent sur la répartition des individus suivant les valeurs observées.

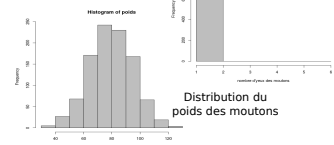
Les échantillons étant aléatoires, les statistiques calculées à partir de ces échantillons seront aussi aléatoires

Outline

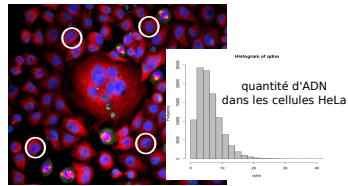
Distribution empirique associée à un échantillon

- C'est la distribution de probabilité sur l'ensemble des modalités
- Chaque observation a le même poids $1/n$
- Dans le cas de variables discrètes on étudie la fréquence des modalités
- Dans le cas de variables continues, on étudie la densité des valeurs observées
- Dans tous les cas on pourra étudier la fonction de répartition empirique

Distribution du nombre d'yeux des moutons



Distribution du poids des moutons



quantité d'ADN dans les cellules HeLa

Données qualitatives et diagramme en bâtons

- Dans le cas d'une variable discrète

$$(x_1 \dots, x_n) = \begin{bmatrix} AA & AA & AA & AA & Aa \\ Bb & bb & Bb & bb & bb \end{bmatrix}$$

*Locus*1	*Freq**	*Locus*2	*Freq*
AA	4	*BB*	0
Aa	1	*Bb*	2
aa	0	*bb*	3
Total	5	*Total*	5

- Si on note $c_1, \dots c_K$ les K modalités ($\{AA, Aa, aa\}$), on construit la loi de probabilité empirique:

$$\forall k \in \{1, \dots, K\}, \quad \widehat{P}(c_k) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i = c_k\}}$$

*En anglais Frequency: comptage

Croisement de données qualitatives

- On peut étudier l'observation croisée de phénotypes dans une **table de contingence**

Locus 1/2	*BB*	*Bb*	*bb*
AA	0	2	2
Aa	0	0	1
aa	0	0	0

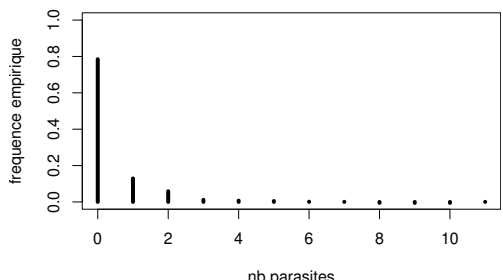
- Si on note $c_1^1, \dots c_{K_1}^1$ les K_1 modalités du locus 1 ($\{AA, Aa, aa\}$), et $c_1^2, \dots c_{K_2}^2$ les K_2 modalités du locus 2 ($\{BB, Bb, bb\}$) on construit la loi de probabilité empirique du couple $\forall (k_1, k_2) \in \{1, \dots, K_1\} \times \{1, \dots, K_2\}$,

$$\widehat{P}(c_{k_1}^1, c_{k_2}^2) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i^1 = c_{k_1}^1, x_i^2 = c_{k_2}^2\}}$$

Construction du diagramme en batons

- On a récolté des châtaignes et compté le nombre de parasites dans chacun des fruits
- $n = 1329$ chataignes, et x_i est le nombre de parasites dans la chataigne i
- Cette variable est observée avec $K = 12$ catégories, $c_k \in \{0, \ldots, 12\}$
- x_i est une variable quantitative discrète

nb parasites	0	1	2	3	4	5	6	7	8	9	10	11
nb de chataignes	1043	172	78	15	10	7	2	1	0	0	0	1



Distribution empirique de lois continues

- Lorsque les observations sont continues, représenter un diagramme en baton n'est pas informatif car chaque modalité est de fréquence 1
- On regroupe les données au sein d'intervalles plus ou moins grands (binning en anglais)
- On compte le nombre d'observations qui tombent dans ces intervalles
- Choisir le nombre et la taille des bins change l'aspect visuel de l'histogramme
- Plus le nombre de bins sera grand, plus les détails seront visibles (lissage)

Definition d'un histogramme

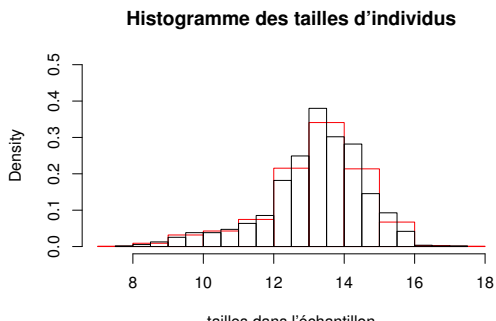
- On observe n variables continues $(x_1, \ldots, x_n)$
- On choisit K bins et des intervalles $[a_0, a_1]$, $]a_1, a_2]$,$\ldots$, $]a_{K-1}, a_K]$ pour regrouper les données en paquets.
- $a_0 = \min_i x_i$, $a_K = \max_i x_i$
- On calcule pour chaque bin la fréquence correspondante

$$\widehat{P}(]a_{k-1}, a_k]) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{a_k - a_{k-1}} 1_{\{x_i \in]a_{k-1}, a_k]\}}$$

- L'histogramme représente ces fréquences rapportées à la longueur des intervalles

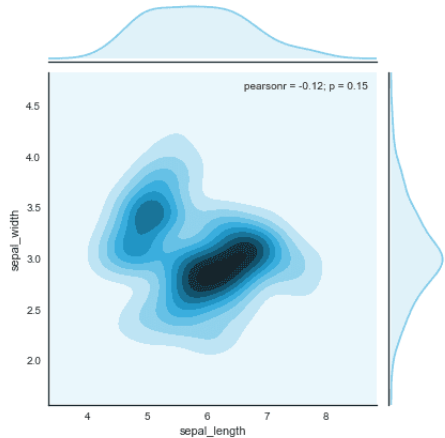
Attention aux représentations visuelles !

- Les histogrammes sont avant tout un outil visuel de représentation de la variabilité des données
- Ils dépendent de paramètres, notamment le nombre et la taille des bins
- Si on a un échantillon de taille n, c'est souvent $\sqrt{n}$ pas qui donne "la meilleure" représentation.



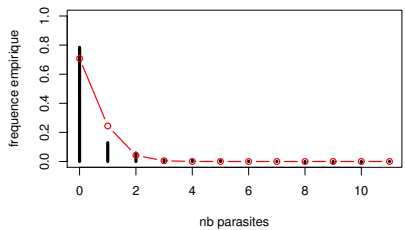
Graphiques en 2 dimensions

- Lorsque les données sont multivariées (2D)
- On peut représenter l'histogramme joint (density plot)
- Les courbes de niveau sont les courbes d'iso densité empirique
- On peut s'intéresser aux relations entre variables (régression)



Comparaison visuelle de distributions

- Les histogrammes peuvent être utilisés pour comparer visuellement des distributions
- On peut comparer des distributions empiriques entre elles
- On peut aussi comparer une distribution empirique avec une distribution théorique
- Exemple du nombre de balanins sur les chataignes et la loi de Poisson.



Loi empirique du nombre de balanins par chataigne et loi de Poisson théorique de paramètre $\lambda = 0.34$.

Point de vue cumulatif

- On peut étudier la répartition des observations d'un point de vue cumulatif
- On compte le nombre d'ascenseurs dans 30 immeubles

nb d'ascenseurs	2	4	6	8	10
nb d'immeubles	10	6	10	2	2

- On peut s'intéresser aux fréquences cumulées

nb ascenseurs	freq. cumulée
2	0.3333333
4	0.5333333
6	0.8666667
8	0.9333333
10	1.0000000

La fonction de répartition empirique

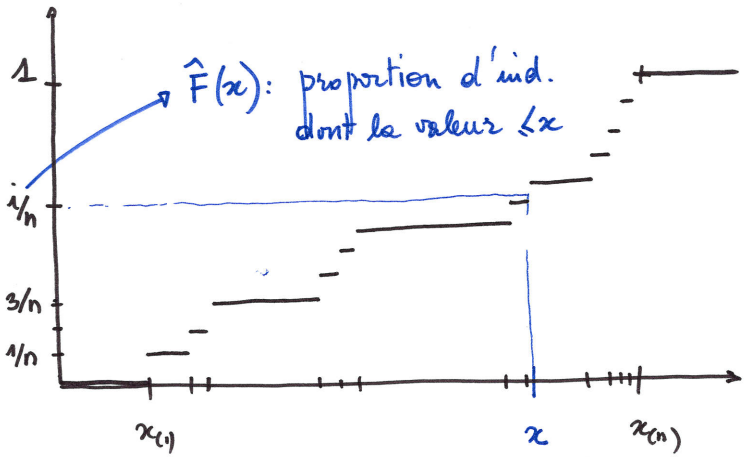
- On appelle statistiques d'ordre de l'échantillon $(x_1, \ldots, x_n)$, les valeurs $x_{(1)}, \ldots, x_{(n)}$ égales aux x_i rangées par ordre croissant.

$$x_{(1)} = \min_{i=1,\ldots,n} \{x_i\} \leq x_{(2)} \leq \cdots \leq x_{(n)} = \max_{i=1,\ldots,n} \{x_i\} .$$

- La fonction de répartition empirique est la proportion d'éléments de l'échantillon qui sont inférieurs ou égaux à x.
- Elle est notée $\widehat{F}(x)$. C'est une fonction de l'ensemble des valeurs prises par $\mathbf{x}$ dans $[0, 1]$, qui vaut :

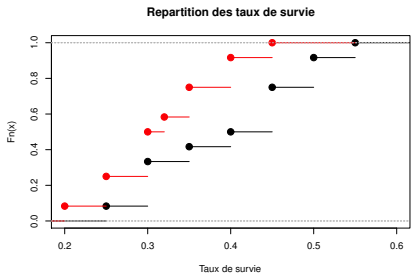
$$\begin{aligned} \widehat{F}(x) &= 0 && \text{pour } x \leq x_{(1)} \\ \widehat{F}(x) &= i/n && \text{pour } x_{(i)} \leq x \leq x_{(i+1)} \\ \widehat{F}(x) &= 1 && \text{pour } x \geq x_{(n)} \end{aligned}$$

La fonction de répartition empirique



Comparaison distributions par la fonction de répartition

- Il peut être plus clair (visuellement) d'utiliser la fonction de répartition pour comparer deux échantillons
- On étudie le taux de survie d'un insecte pendant son développement embryonnaire.
- On considère deux sites d'études (noir et rouge sur le graphique)
- 80% des larves du site S1 on un taux de survie < 0.5 contre 100% du site S2.



Noir : $\widehat{F}$ pour la survie sur le site S1, Rouge pour S2.

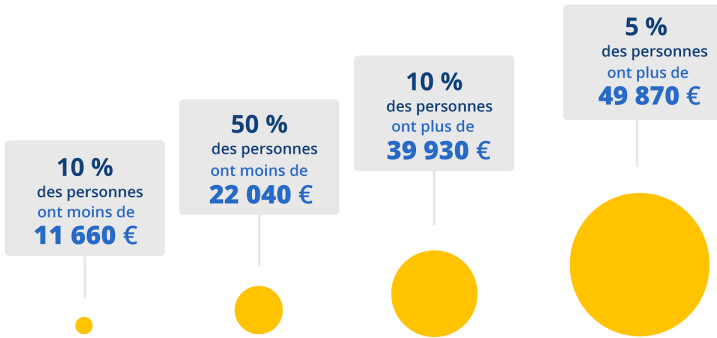
Et si on retournait le point de vue cumulé ?

- On peut s'intéresser aux fréquences cumulées

nb ascenseurs	freq. cumulée
2	0.3333333
4	0.5333333
6	0.8666667
8	0.9333333
10	1.0000000

- Plus de 25% des immeubles ont au moins 2 ascenseurs
- $\sim$50% des immeubles ont au moins 4 ascenseurs
- Plus de 75% des immeubles ont au moins 6 ascenseurs

Les quantiles dans la vie quotidienne



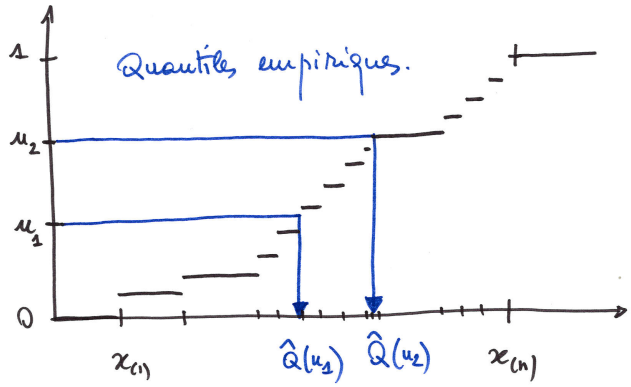
Données INSEE (2019): En 2019, les 10% d'individus les plus modestes (1^{er} décile) ont un niveau de vie inférieur à 11 660 euros par an.

De la fonction de répartition empirique aux quantiles

- La fonction de répartition $\widehat{F}(x)$ renseigne sur la proportion d"observations qui sont inférieures ou égales à l'élément x. C'est une valeur entre 0 et 1.
- On peut se poser la question inverse: quelle serait la valeur de x pour laquelle on aurait $u\%$ des individus sous cette valeur ? C'est une valeur entre $x_{(1)}$ et $x_{(n)}$.
- Les quantiles permettent d'extraire des valeurs caractéristiques des distributions (extrêmes, médianes).

Les quantiles permettent de placer les autres valeurs par rapport à celles ci en terme de centralité ou d'exceptionalité.

Quantiles Empiriques



Quantiles empiriques

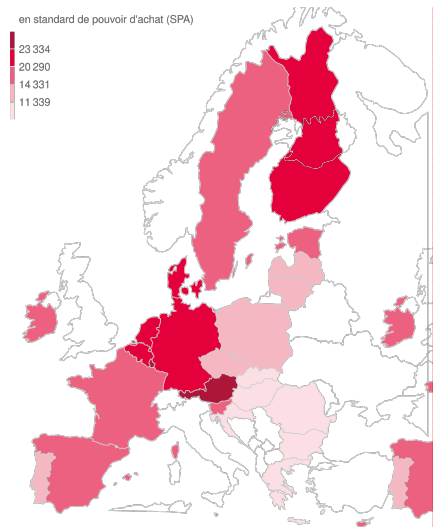
- La fonction quantile empirique de l' échantillon est la fonction $\widehat{Q}$ qui, pour tout $i = 1, \ldots, n$, vaut $x_{(i)}$ sur l'intervalle $]\frac{i-1}{n}, \frac{i}{n}]$.

$$\forall u \in \left]\frac{i-1}{n}, \frac{i}{n}\right] \quad , \quad \widehat{Q}(u) = x_{(i)} \; .$$

- Pour certaines valeurs de u, on donne un nom particulier aux quantiles $\widehat{Q}(u)$ (médiane, quartiles, déciles)
- La médiane est une valeur centrale de l'échantillon : il y a autant de valeurs qui lui sont inférieures que supérieures.

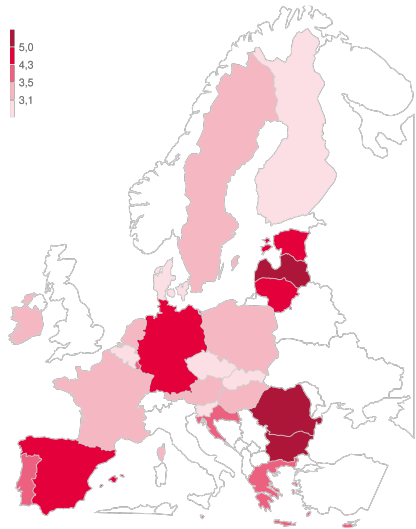
Les quantiles dans la vie quotidienne (2)

- Le Standard de pouvoir d'achat est une unité monétaire artificielle permettant de comparer en volume des indicateurs économiques entre pays
- En 2019, en France, le niveau de vie médian en standard de pouvoir d'achat (SPA) est de 19 151.
- Données INSEE (2019)



Les quantiles dans la vie quotidienne (3)

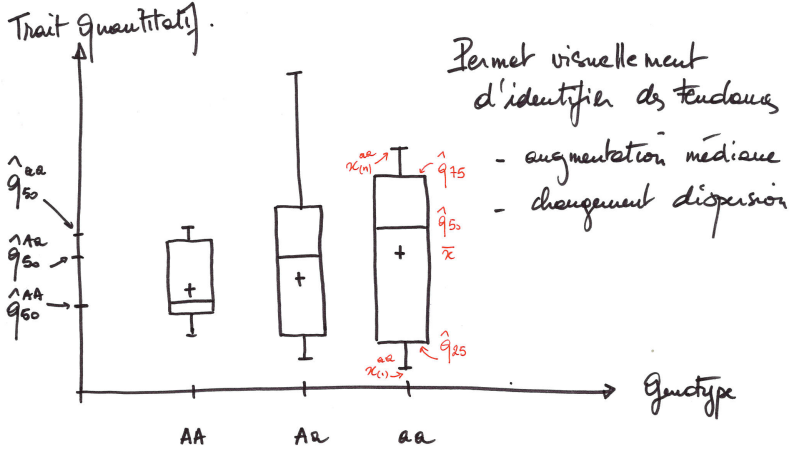
- On peut s'intéresser au rapport des déciles $\widehat{Q}(10\%)$ et $\widehat{Q}(90\%)$
- En 2019, en France, les 10% d'individus les plus aisés ont un niveau de vie 3,4 fois plus élevé que les 10% les plus modestes.
- Bulgarie: 5.8, Allemagne: 4.6, Espagne: 4.8, Slovaquie: 2.7 Données INSEE (2019)



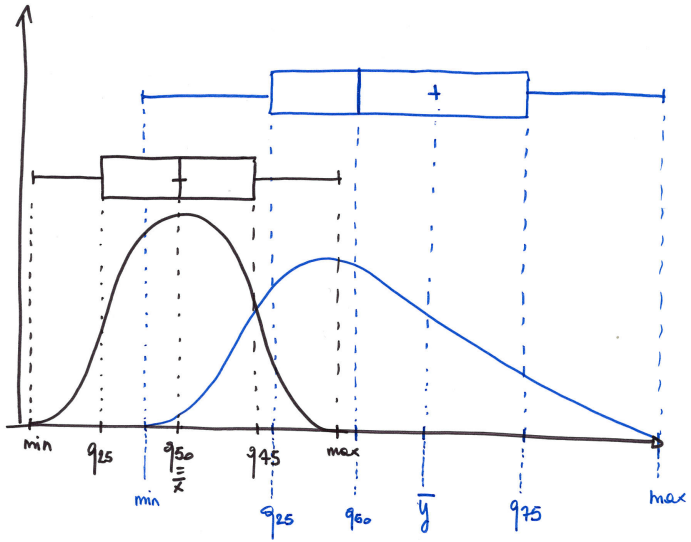
Quantiles et boxplots

- La représentation en boxplot repose sur l'hypothèse que seules quelques valeurs des quantiles peuvent être utilisées pour synthétiser l'information contenue dans une distribution
- L'intervalle de base des quantile est l'inter-quartile range
- Comparer des distributions s'avère être facilité par cette représentation qui permet de visualiser directement l'étendue de la distribution empirique.

Boxplots


Trait quantitatif.
Permet visuellement d'identifier des tendances
- augmentation médiane
- changement dispersion
AA
Aa
aa
Genotype
3 Facteurs Qualitatifs
3 modalités d'un même facteur (Genotype)

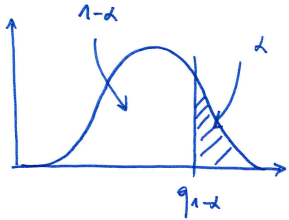
Boxplots et histogrammes



Quantiles et exceptionnalité

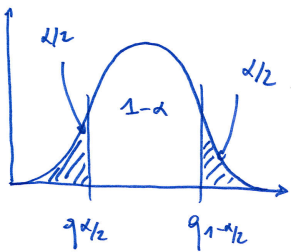
- La notion de quantile est centrale pour comprendre de nombreuses démarches statistiques car les quantiles renseignent sur la **répartition des d'individus** à droite et à gauche du quantile (ex: salaire médian)
- Exemple : si j'ai une nouvelle observation x_{n+1}, comment la 'placer' par rapport aux autres $(x_1, \ldots, x_n)$?
  - si $x_{n+1} > \widehat{q}(0.99999)$ alors x_{n+1} est "très" exceptionnelle **par rapport à la distribution empirique**
  - si $x_{n+1} > q(0.99999)$ alors x_{n+1} est "très" exceptionnelle **par rapport à ce qu'aurait prédit un modèle** (cf. tests)

Calculs utiles sur les quantiles (pour la suite)



$$\mathbb{P}(Z \leq q_{1-\alpha}) = F(q_{1-\alpha}) = 1-\alpha$$

$$\mathbb{P}(Z \geq q_{1-\alpha}) = 1 - F(q_{1-\alpha}) = \alpha.$$

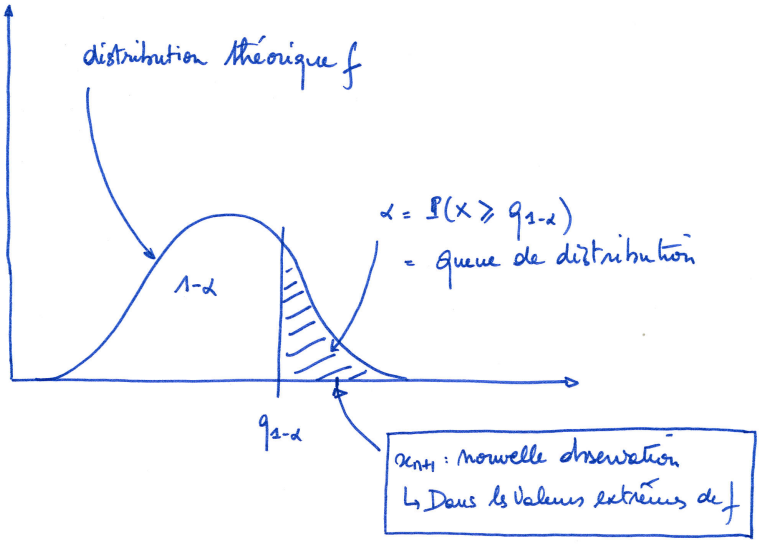


$$\mathbb{P}\left(Z \in [q_{\alpha/2}, q_{1-\alpha/2}]\right) = 1 - \left[\mathbb{P}(Z \leq q_{\alpha/2}) + \mathbb{P}(Z \geq q_{1-\alpha/2})\right]$$

$$= 1 - \left[\alpha/2 + 1 - 1 + \alpha/2\right]$$

$$= 1-\alpha$$

A garder en mémoire



Outline

La moyenne empirique et centre de gravité

- Si l'échantillon est noté $(x_1, \ldots, x_n)$, sa moyenne empirique est :

$$\overline{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- La moyenne empirique de deux échantillons réunis de tailles respectives n_x et $n_{x'}$, de moyennes respectives $\overline{x}$ et $\overline{x'}$ sera le nouveau barycentre:

$$\overline{xx'} = \frac{n_x\overline{x} + n_{x'}\overline{x'}}{n_x + n_{x'}}$$

- Elle est **sensible aux valeurs extrêmes**
- Le **centrage** des données consiste à retrancher la moyenne empirique à toutes les valeurs de l'échantillon qui devient centré $(x_1 - \overline{x}, \ldots, x_n - \overline{x})$

Exemple de détection de valeurs aberrantes

- Dans deux types de forêts on a mesuré les hauteurs de $n = 12$ arbres

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

- Contrairement à la moyenne, la médiane est insensible aux valeurs aberrantes.
- On note x_i^1 la taille de l'arbre i mesuré dans la forêt de type 1.
- On compare les statistiques descriptives des deux types de forêt:

	Moyenne	Ecart-type	Médiane
Type 1	25.66	1.24	26.2
Type 2	25.31	2.60	24.5
Type 2(sans x_{12}^2)	24.85	1.52	24.5

La variance empirique et la dispersion

- C'est un indicateur qui permet de quantifier la **dispersion** d'un échantillon autour de sa moyenne.
- La variance empirique de l'échantillon est notée s^2,:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2 = \overline{x^2} - \overline{x}^2$$

- L'écart-type (s) a l'avantage de s'exprimer, comme la moyenne, dans la même unité que les données.
- On utilise parfois des indicateurs construits à partir de $\overline{x}$ et s^2:
  - le rapport signal sur bruit $\overline{x}^2/s^2$ qui normalise l'intensité du signal par rapport à sa dispersion
  - le coefficient de variation $s/\overline{x}$ qui quantifie le degré de variabilité rapporté à la localisation de la distribution

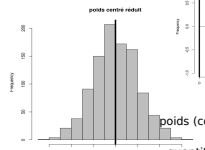
La réduction des données

- Après centrage des données, on peut également les réduire :

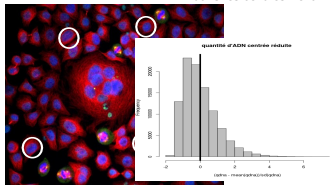
$$\widetilde{\mathbf{x}} = \left(\frac{x_1 - \overline{x}}{s}, \ldots, \frac{x_n - \overline{x}}{s} \right)$$

- On obtient ainsi un nouvel échantillon dont la moyenne est nulle et la variance égale à 1, ces nouvelles données n'ont plus d'unité.
- On peut donc comparer deux échantillons réduits

Distribution du nombre d'yeux (centré réduit) des mout



Distribution du poids (centré réduit) des moutons



quantité d'ADN centrée réduite dans les cellules HeLa

References

- *Introduction to Statistical thinking* Benjamin Yakir, 2010
- *Histoire de la statistique*, Philippe Tassi, Jean-Jacques Droesbecke, Que Sais-Je ? PUF 1997
- *Statistics in Action with R* Marc Lavielle Inria Saclay & Ecole Polytechnique (CMAP)
- *Statistique inférentielle : idées, démarches, exemples*, Jean-Jacques Daudin, Stéphane Robin, Colette Vuillet, PUR, 1999
- Statistiques avec R, Pierre Cornillon et al., PUR 2010
- *Practical Statistics for Life Sciences* Lieven Clement, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium.