Premières notions de statistique
statistiques descriptives, estimation, modèles et intervalles de confiance

Franck Picard

UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

Outline

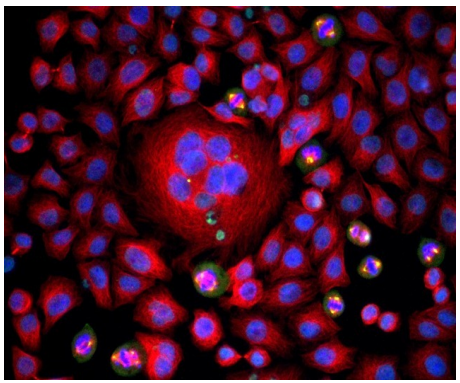
Populations et caractères

- 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

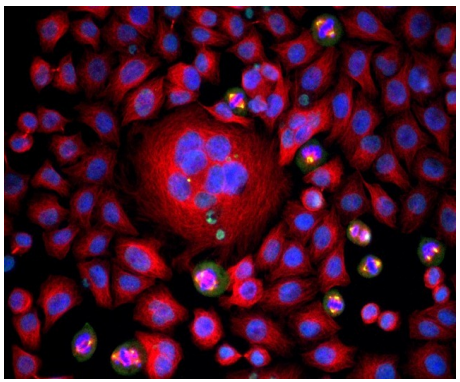
Quand fait-on de la statistique ?

- 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

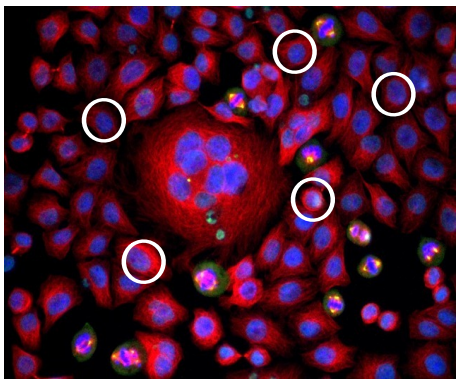
La notion d'échantillon

- 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 $(x_1, x_2, x_3) = (15, 18, 19)$.

Animaux
- Génotype
- Taille
- Poids
- couleur des yeux

Galaxies
- luminosité
- vitesse
- densité

chromosomes
- longueur
- # de copies



cellules
- taille du noyau
- # de chromosomes
- taille

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

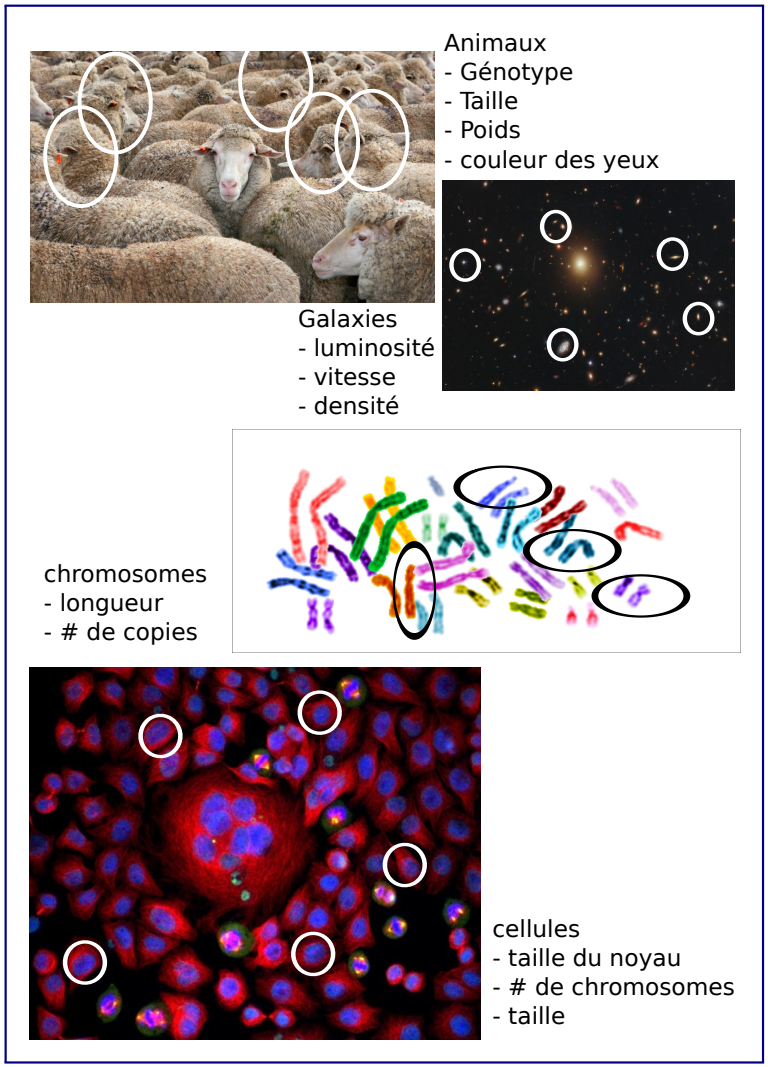
Poids	0.64	0.65	0.73	0.60	0.65	0.77	0.82	0.64	0.66	0.72

- La première étape de formalisation consiste à poser : $n = 10$, et x_i le poids du cocon i, tel que l'échantillon observé est

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

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 $(x'_1, x'_2, x'_3) = (17, 19, 16)$



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.

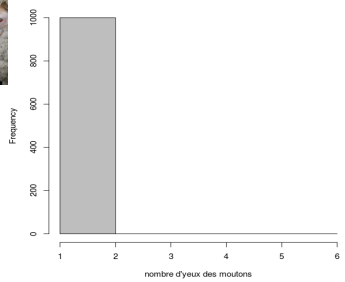
Les échantillons étant aléatoires, les statistiques calculées à partir de ces échantillons seront aussi aléatoires

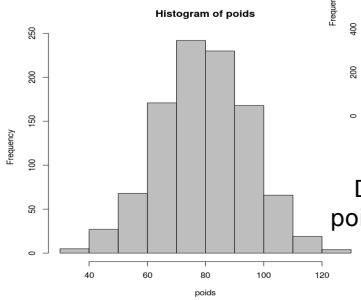
Outline

Distribution empirique associée à un échantillon

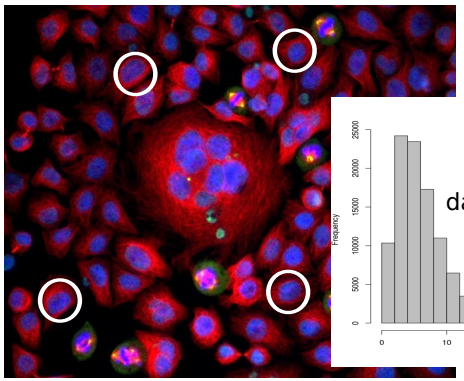
- C'est la distribution de probabilité sur l'ensemble des modalités qui affecte chaque observation du poids $1/n$.
- Si on note $c_1, \ldots, c_k$ les valeurs distinctes prises par les x_i et pour $h = 1, \ldots, k$, $n_h = \sum_{i=1}^{n} 1_{\{x_i = c_h\}}$ l'effectif de la valeur c_h.
- C'est la loi de probabilité $\widehat{P}$ sur l'ensemble $\{c_1, \ldots, c_k\}$, telle que : $\widehat{P}(c_h) = \dfrac{n_h}{n}$.

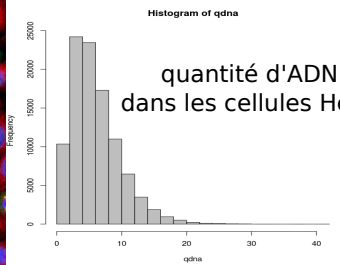
Distribution du nombre d'yeux des moutons





Distribution du poids des moutons





quantité d'ADN dans les cellules HeLa

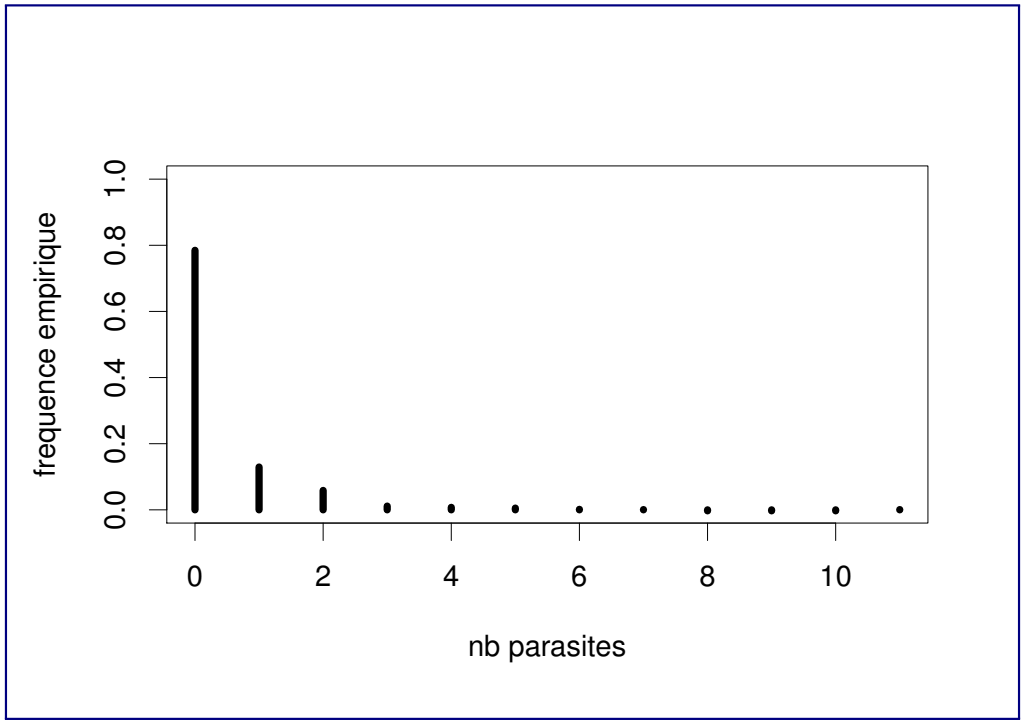
Distribution empirique de lois discrètes

- Lorsque nombre de valeurs différentes k est faible devant la taille de l'échantillon n
- Le **diagramme en bâtons** consiste à représenter les valeurs différentes $c_1, \ldots, c_k$ en abscisse, avec au-dessus de chacune une barre verticale de longueur égale à sa fréquence expérimentale $\widehat{P}(c_h)$.
- Pour des échantillons qualitatifs, on utilise aussi des représentations en **camembert** (pie-chart), ou en **barre**.
- Elles consistent à diviser un disque ou un rectangle proportionnellement aux différentes fréquences.
- Pour explorer différentes représentations graphiques `http://pbil.univ-lyon1.fr/R/pdf/lang04.pdf`.

Construction du diagramme en batons

- Le balanin de la châtaigne *Curculio elephas* est un insecte parasite de la châtaigne. On a récolté des châtaignes et compté le nombre de parasites dans chacun des fruits

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 et histogrammes

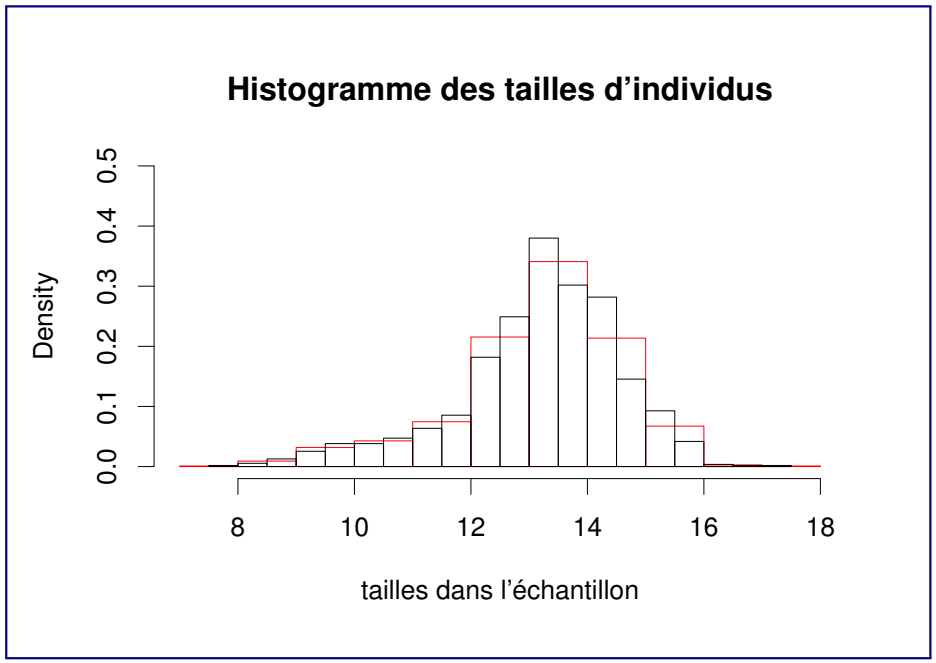
- On choisit k classes et des intervalles $[a_0, a_1]$, $]a_1, a_2]$,..., $]a_{k-1}, a_k]$ pour regrouper les données en paquets.
- On calcule pour chaque $]a_{h-1}, a_h]$ la fréquence correspondante

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

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

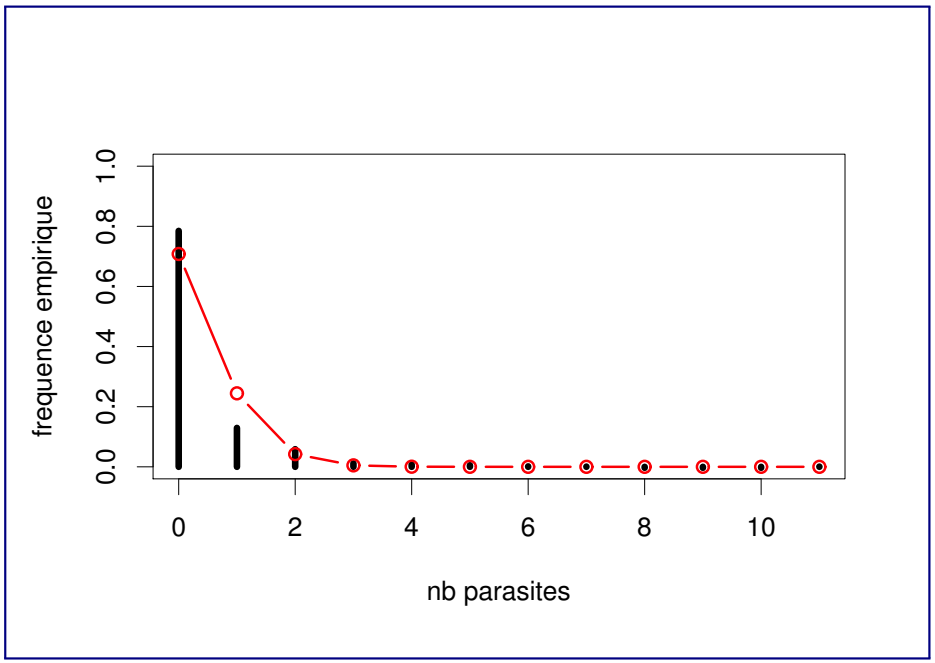
Attention aux représentations visuelles

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



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 $\lambda = 0.34$.

La fonction de répartition empirique

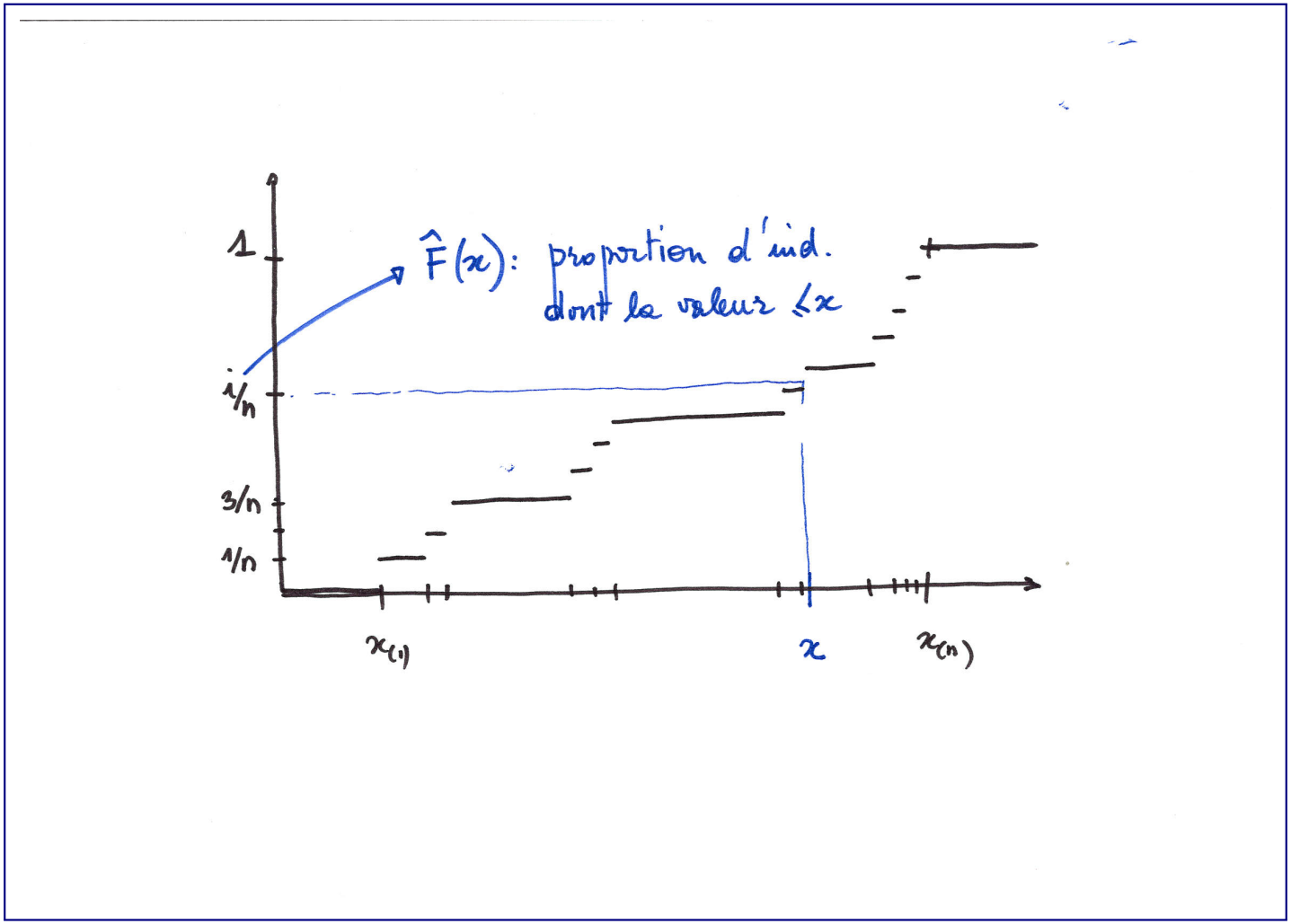
- On appelle statistiques d'ordre de l'échantillon $(x_1, \ldots, x_n)$, les valeurs $x_{(1)}, \ldots, x_{(n)}$ égales aux x_i rangées par ordre croissant.

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

- 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 **x** dans $[0, 1]$, qui vaut :

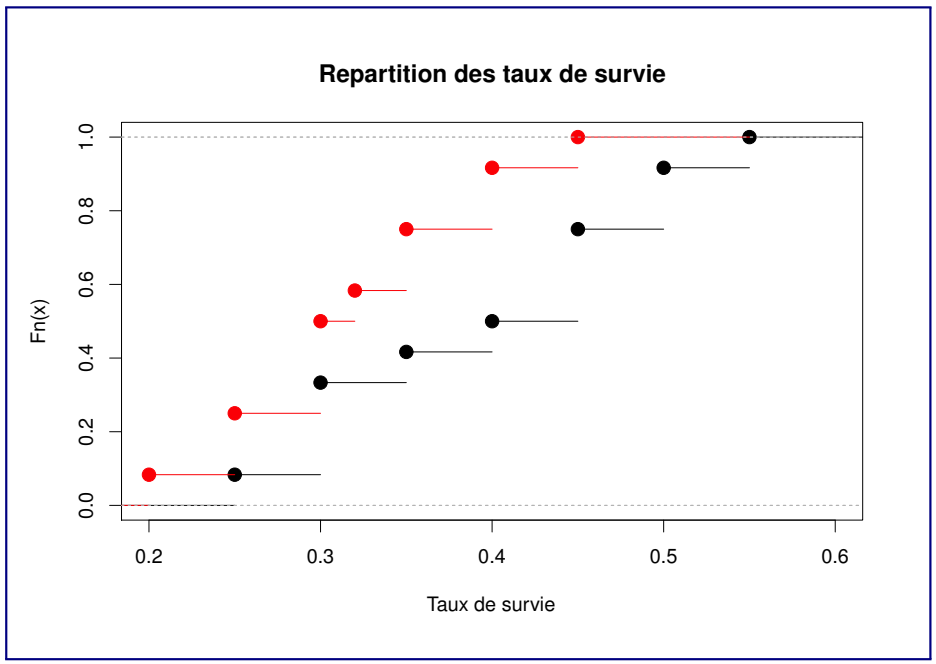
$$\begin{aligned} \widehat{F}(x) &= 0 && \text{pour } x \leq x_{(1)} \\ \widehat{F}(x) &= i/n && \text{pour } x_{(i)} \leq x \leq x_{(i+1)} \\ \widehat{F}(x) &= 1 && \text{pour } x \geq x_{(n)} \end{aligned}$$

La fonction de répartition empirique



Comparaison visuelle de 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.

Outline

La moyenne empirique et centre de gravité

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

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

- Elle respecte une propriété d'**associativité**: 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:

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

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

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

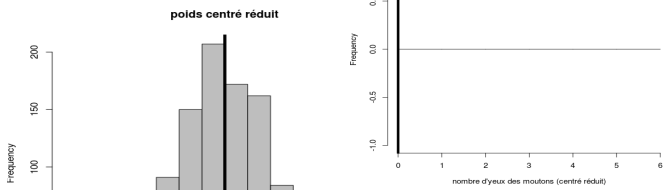
La réduction des données

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

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

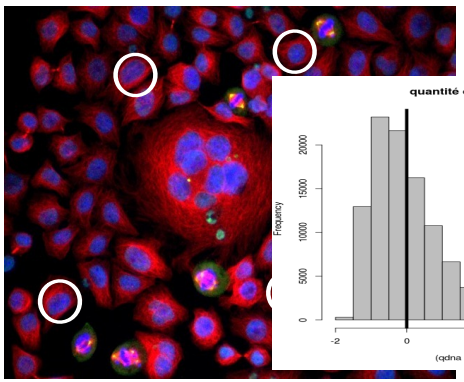
- On obtient ainsi un nouvel échantillon dont la moyenne est nulle et la variance égale à 1, ces nouvelles données n'ont plus d'unité.
- On peut donc comparer deux échantillons réduits

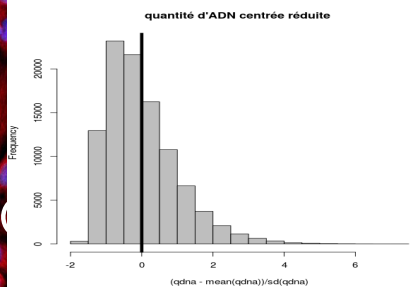
Distribution du nombre d'yeux (centré réduit) des mout



Distribution du poids (centré réduit) des moutons

quantité d'ADN centrée réduite dans les cellules HeLa





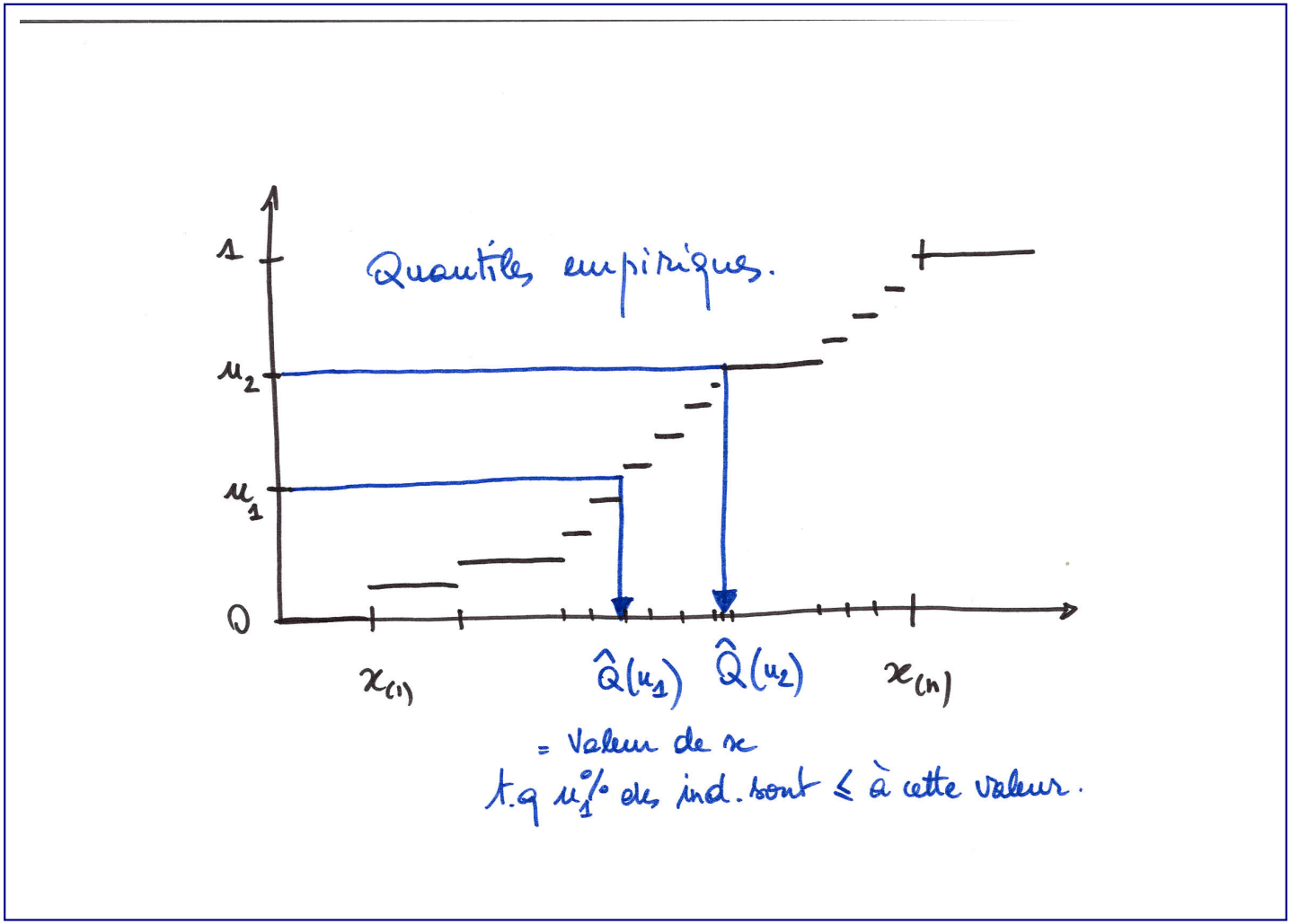
Outline

De la fonction de répartition empirique aux quantiles

- La fonction de répartition $\widehat{F}(x)$ renseigne sur la proportion de points de l'échantillon qui sont inférieurs ou égaux 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.
- Si l'échantillon est dissymétrique, avec une distribution très étalée vers la droite, la médiane pourra être nettement plus petite que la moyenne.
- Contrairement à la moyenne, la médiane est insensible aux valeurs aberrantes.

Exemple de détection de valeurs aberrantes

- Dans deux types de forêts on a mesuré les hauteurs de 12 arbres choisis au hasard et indépendemment. On souhaite savoir si la taille des arbres dépend du type de forêt.

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	32

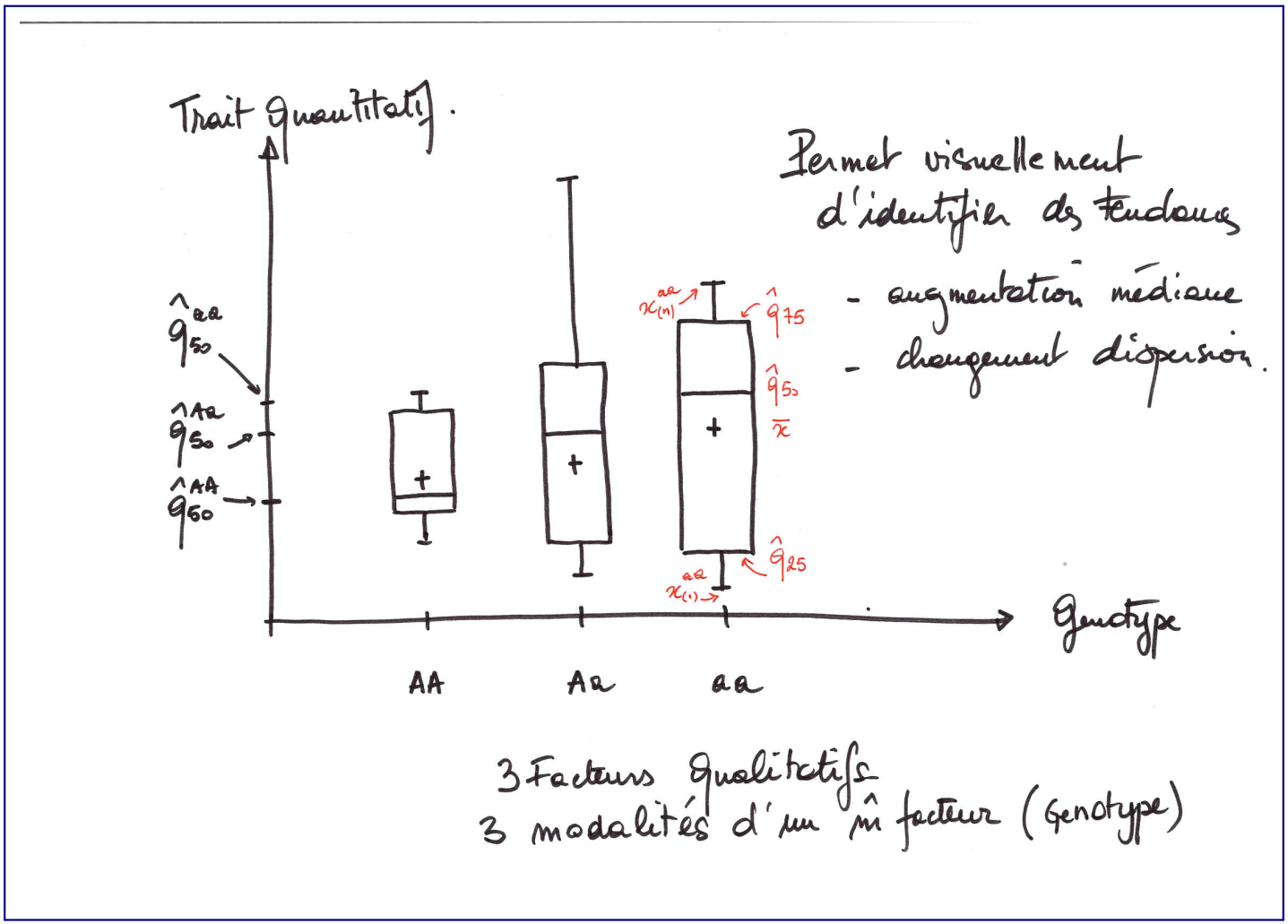
- On note x_i^1 la taille de l'arbre i mesuré dans la forêt de type 1.
- On compare les statistiques descriptives des deux types de forêt:

	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

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.

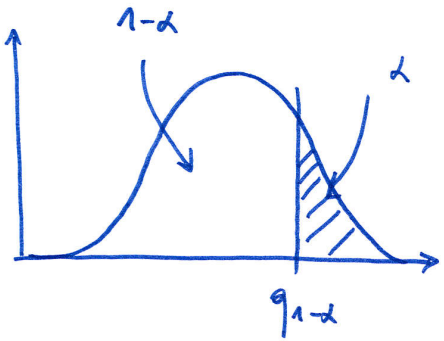
Boxplots


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

Quantiles et exceptionnalité

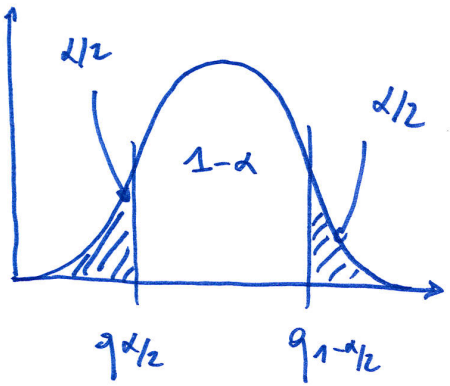
- La notion de quantile est centrale pour comprendre de nombreuses démarches statistiques car les quantiles renseignent sur la **répartition des d'individus** à droite et à gauche du quantile (ex: salaire médian)
- Exemple : si j'ai une nouvelle observation x_{n+1}, comment la 'placer' par rapport aux 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)



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

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

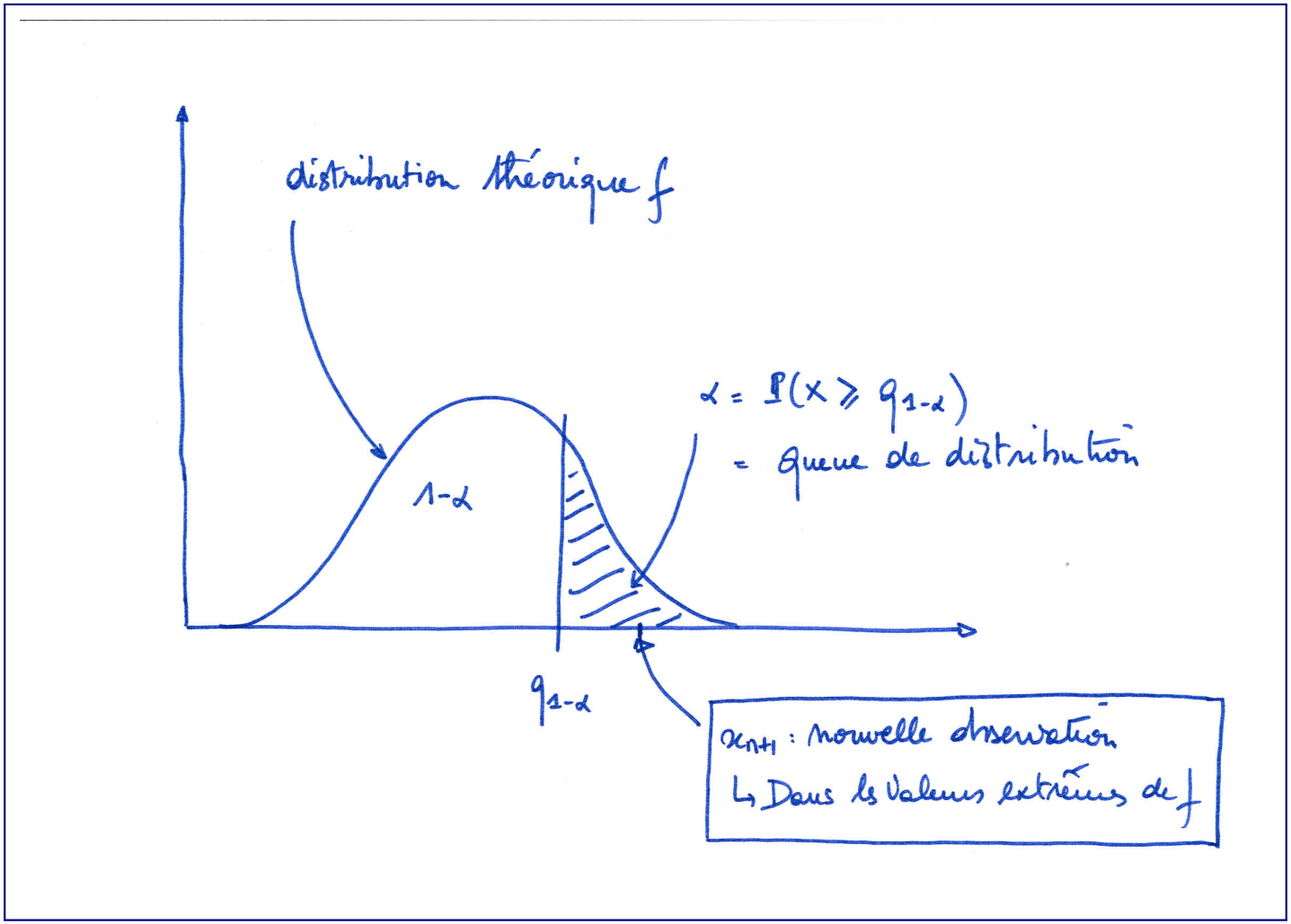


$$\mathbb{P}\left(Z \in [q_{\alpha/2}, q_{1-\alpha/2}]\right) = 1 - \left[\mathbb{P}(Z \leq q_{\alpha/2}) + \mathbb{P}(Z \geq q_{1-\alpha/2})\right]$$

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

$$= 1-\alpha$$

A garder en mémoire



Lien avec les tests (exceptionalité) et avec les intervalles de confiance

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)

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**: $(x_1, \ldots, x_n)$, n réalisations de $(X_1, \ldots, X_n)$
- La loi jointe des observations est:

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

▸ Rappels Variables Aléatoires

Echantillon Indépendant et Identiquement distribué (i.i.d.)

- On considère un n échantillon aléatoire $\mathbf{X}$ (vecteur de taille n).
- La notion d'échantillon iid 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:

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

Moments d'un Echantillon i.i.d.

- les X_i ont tous la même loi donc la même espérance:

$$\forall i \in \{1, \ldots, n\}, \mathbb{E}(X_i) = \mu$$

- L'espérance de $\mathbf{X}$ est donc un vecteur de taille n:

$$\mathbb{E}(\mathbf{X}) = [\mathbb{E}(X_1), ...\mathbb{E}(X_n)] = \mu \times [1, \ldots, 1]$$

- Variance: les X_i ont la même variance:

$$\forall i \in \{1, \ldots, n\}, \mathbb{V}(X_i) = \sigma^2$$

- Covariance: les X_i sont indédendants, donc

$$\forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, n\} \text{cov}(X_i, X_j) = 0.$$

- Matrice de Variance/Covariance $\mathbb{V}(\mathbf{X})$ $(n \times n)$: $\mathbb{V}(X_1), ...\mathbb{V}(X_n)$ est sur la diagonale et $\text{cov}(X_i, X_j)$ sur les termes extra-diagonaux
- Exemple de données pour lesquelles l'hypothèse iid n'est pas valide: données temporelles, spatiales, familiales, etc

Retour sur les tailles 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.

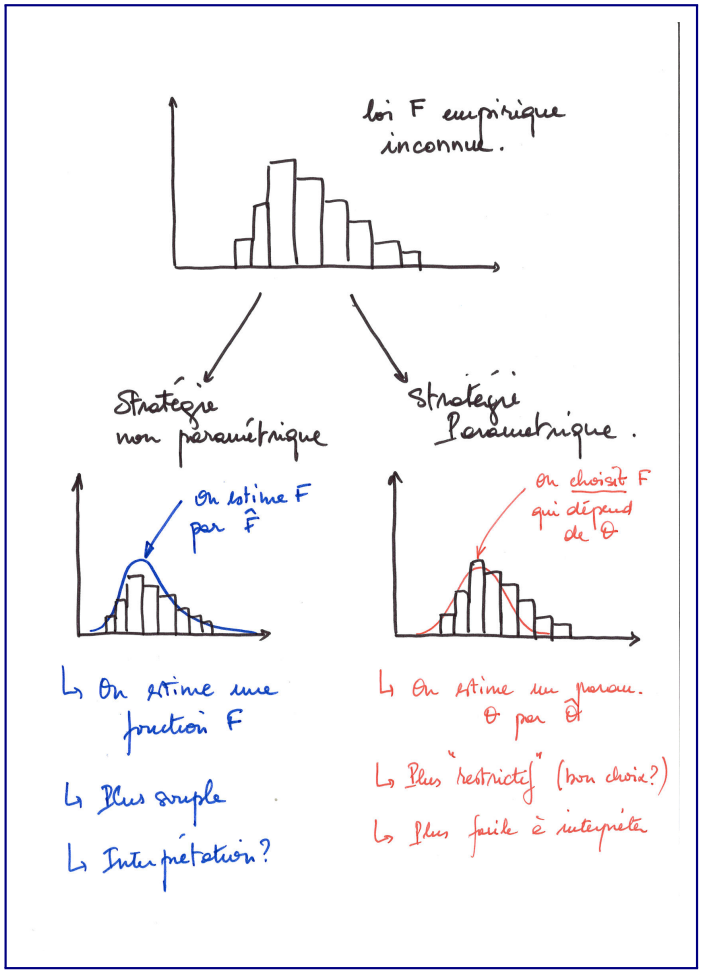
Type 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	27
Type 2	22.5	22.9	23.74	24.0	24.4	24.5	25.3	26	26.4	26.7	32

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.
- Dans la suite on notera F_θ le modèle de distribution paramétré par θ (notation générique).

Le cadre des modèles paramétrique

- Deux stratégies possibles:
  - pas d'hypothèse sur F
  - On suppose que F appartient à une famille de lois connue
- En général ces lois dépendent d'un nombre fini de paramètres qui deviennent les seules inconnues du problème
- Supposons que l'on sache identifier les paramètres d'intérêt



Retour sur les tailles 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, telle que:

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

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

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

Exemple de modèle pour données de comptage

- En période d'épizootie de péripneumonie contiagieuse bovine (PPCCB), on s'intéresse au nombre de troupeaux inféctés dans les départements francais.
- On effectue un relevé et on obtient:

Département	Nb de troupeaux infectés
01	120
02	200
...	

- On définit x_i le nombre observé de troupeaux 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 ?

Exemple de modèle pour des observations qualitatives-1

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

pop1	"AA"	"Aa"	"Aa"	"AA"	"Aa"	"aA"
pop2	"Aa"	"aa"	"AA"	"Aa"	"aa"	"aa"

- On note x_i^1 la variable qui vaut 1 si l'individu i dans la population 1 est hétérozygote, 0 sinon.
- On suppose que x_i^1 est la réalisation de X_i^1. Les individus étant non apparentés, on suppose que les X_i^1 sont indépendants de loi

$$\forall i \in \{1, \ldots, n_1\},\ X_i^1 \sim \mathcal{B}(p_1)$$

Exemple de modèle pour des observations qualitatives-2

- A partir des mêmes données on peut poser un autre modèle
- Supposons que l'on s'intéresse à la fréquence des génotypes
- On note x_i^1 la variable telle que:

$$x_i^1 = \begin{cases} 0 & \text{si} \quad x_i^1 = \texttt{"AA"} \\ 1 & \text{si} \quad x_i^1 = \texttt{"Aa"} \\ 2 & \text{si} \quad x_i^1 = \texttt{"aa"} \end{cases}$$

- On suppose que x_i^1 est la réalisation de X_i^1. Les individus étant non apparentés, on suppose que les X_i^1 sont indépendants de loi

$$\forall i \in \{1, \ldots, n_1\}, \; X_i^1 \sim \mathcal{M}(1, p_{\texttt{AA}}^1, p_{\texttt{Aa}}^1, p_{\texttt{aa}}^1),$$

- La loi multinomiale généralise la loi Binomiale à plus de deux modalités, avec $p_{\texttt{AA}} + p_{\texttt{Aa}} + p_{\texttt{aa}} = 1$. ▸ Modèle multinomial

Outline

La somme des carrés et la notion de distance-dispersion

- L'interprétation géométrique des indicateurs statistique est un élément central de nombreuses méthodologies
- On peut considérer que $(x_1 ..., x_n)$ est un vecteur de taille n noté $\mathbf{x}$
- On peut s'interroger sur la distance des observations (éléments de $\mathbf{x}$) à une valeur m

$$d^2(\mathbf{x}, m) = \|\mathbf{x} - m\|^2 = \sum_{i=1}^{n} (x_i - m)^2$$

- La somme des carrés quantifie la distance de chaque point de l'échantillon à une valeur de localisation
- $\overline{x}$ est la valeur de m qui minimise cette distance (centre de gravité)

$$d^2(\mathbf{x}, \overline{x}) = \min_m \|\mathbf{x} - m\|^2$$

Distance au modèle

- Une fois le modèle choisi, il faut confronter le modèle aux données
- La première étape consiste souvent à estimer le paramètre d'espérance (position)
- On considère un modèle de distribution tel que $X_i \sim F$ d'espérance

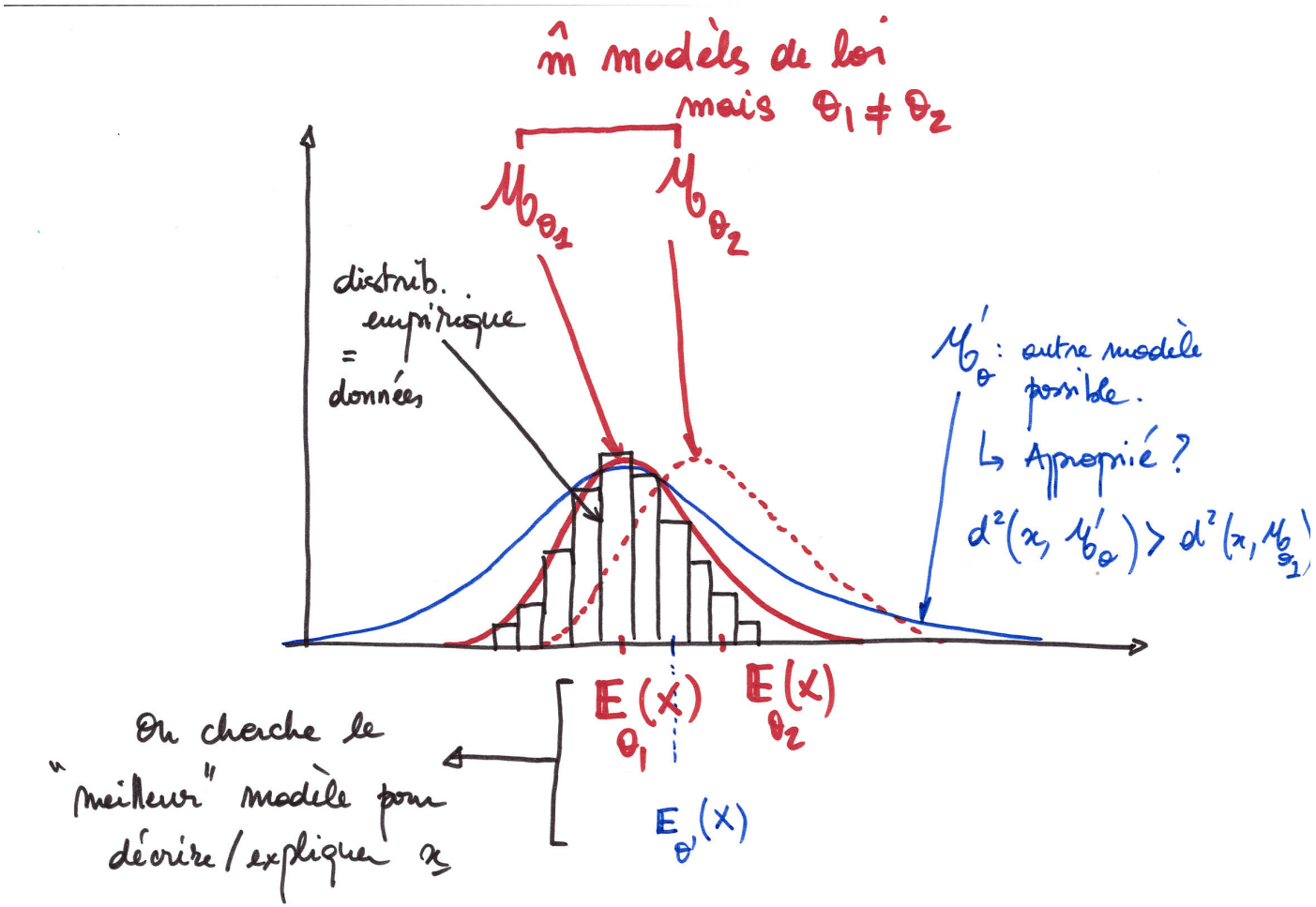
$$\mathbb{E}(\mathbf{X}) = \mu$$

- On cherche donc la quantité m qui minimise la somme des carrés:

$$d^2(\mathbf{X}, m) = \|\mathbf{X} - m\|^2 = \sum_{i=1}^{n} (X_i - m)^2$$

- L'idée centrale de la démarche est de considérer que les données sont figées et que c'est le modèle qui bouge

Notion de distance au modèle



Exemples d'estimation par moindre-carrés

- On a posé un modèle gaussien sur la taille des arbres tel que

$$X_i^1 \sim \mathcal{N}(\mu_1, \sigma_1^2).$$

- On souhaite estimer μ_1 par moindre carrés:

Type 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.8	26.9	27

$$\begin{aligned} d^2(\mathbf{x}, \mu_1) &= (x_1^1 - \mu_1)^2 + (x_2^1 - \mu_1)^2 \ldots + (x_{12}^1 - \mu_1)^2 \\ &= (23.4 - \mu_1)^2 + (24.4 - \mu_1)^2 \ldots + (27 - \mu_1)^2 \end{aligned}$$

- C'est une fonction de μ_1. Pour trouver le $\widehat{\mu}_1$ qui minimise ce critère, on dérive et on annule la dérivée. On trouve bien la moyenne empirique:

$$\widehat{\mu}_1(\mathbf{x}) = \bar{\mathbf{x}} = 25.66$$

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

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

Paramètre / Estimateur / Estimation

- μ est le paramètre de la loi F qui a été choisie pour décrire les observations
- $\widehat{\mu}(\mathbf{x}) = \overline{x}$ est une estimation: sa valeur dépend de l'échantillon $\mathbf{x}$
- $\widehat{\mu}(\mathbf{X}) = \overline{X}$ est un estimateur. C'est une statistique qui est une fonction de $\mathbf{X}$
- $\overline{X}$ est une variable aléatoire dont la loi est définie par le modèle F
- On peut donc caractériser un estimateur par son espérance $\mathbb{E}(\widehat{\mu}(\mathbf{X}))$ et sa variance $\mathbb{V}(\widehat{\mu}(\mathbf{X}))$

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 $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- On estime μ par $\overline{X}$:

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

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

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

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

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

Interprétation

- $\frac{\sigma^2}{n}$ est la variance de l'estimateur $\widehat{\mu}(\mathbf{X})$ de l'espérance μ
- Plus n augmente, plus on est précis dans la caractérisation du paramètre μ
- La loi de $\widehat{\mu}(\mathbf{X})$ est donnée par le modèle (somme de Gaussiennes iid)

$$\widehat{\mu}(\mathbf{X}) = \overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

- On peut aussi considérer sa version centrée réduite:

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

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 :

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

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

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

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

Le Théorème Limite Central

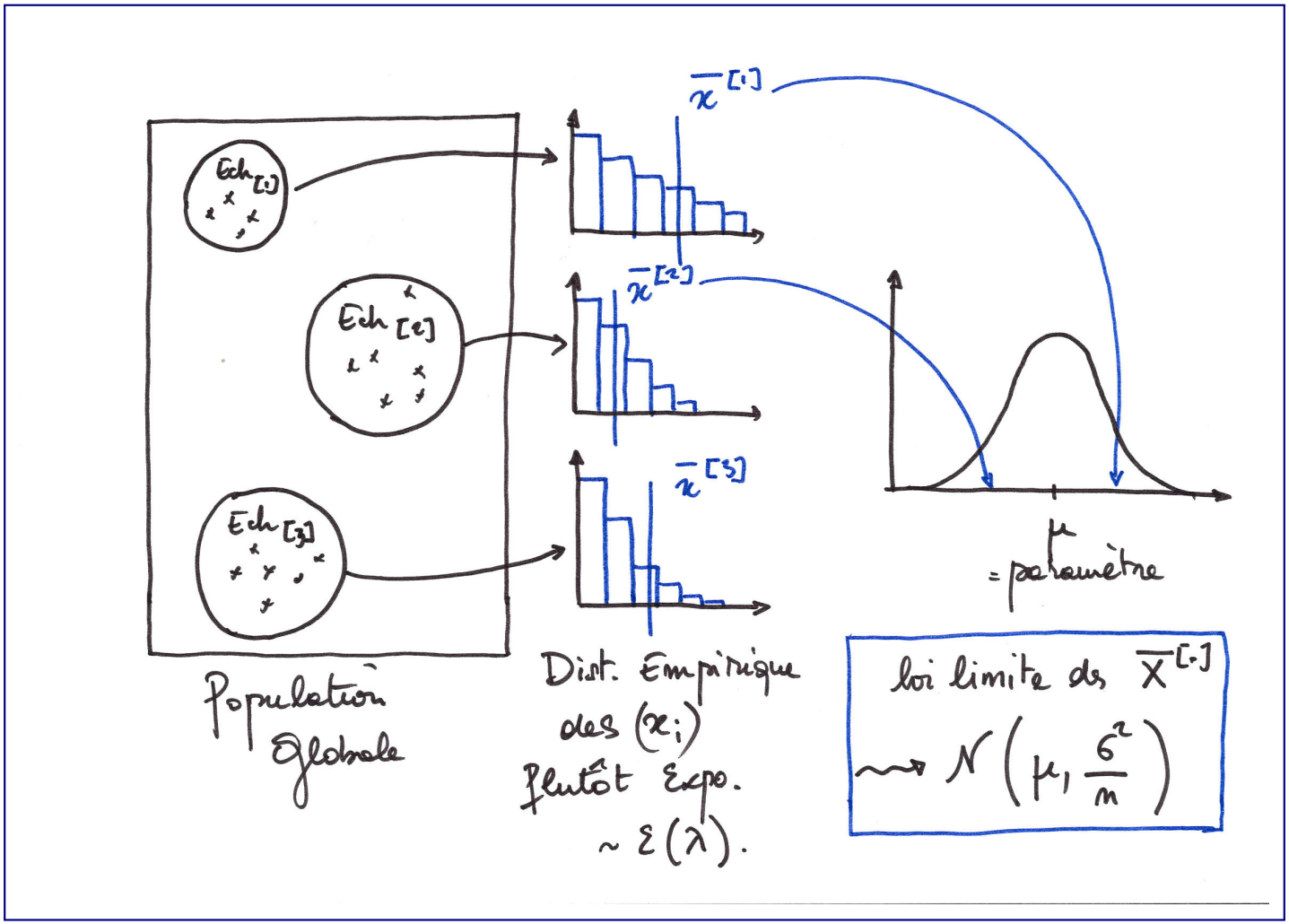
- Si les X_i sont des variables aléatoires iid, si leur espérance μ, et leur variance σ^2 existent alors

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow[\mathcal{L}]{} \mathcal{N}(0,1)$$

- Interprétation: même si les observations ne sont pas modélisées par une loi Gaussienne, l'estimateur de l'espérance fondé sur $\overline{X}$ 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



Exemple dans le cas Bernoulli-1

- On observe $(x_1, \ldots, x_n)$ et on suppose que cet échantillon constitue une réalisation de n variables aléatoires indépendantes et identiquement distribuées (iid) $(X_1, \ldots, X_n)$, avec $X_i \sim \mathcal{B}(p)$
- La loi de Y le nombre de succès observés sur n expériences:

$$Y = \sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p)$$

- On choisit $\widehat{p}(\mathbf{X}) = \overline{X}$ comme estimateur de la probabilité de succès p
- On peut caractériser son espérance et sa variance:

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

Exemple dans le cas Bernoulli-2

- Le TCL est un théorème **asymptotique**, quand le nombre d'observations est suffisant ($np > 5$).

Grâce au TCL on peut supposer:

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

Estimateur de la variance dans le cas Gaussien

- On observe $\mathbf{x}$ et on suppose qu'il constitue une réalisation d'un n-échantillon iid $\mathbf{X}$, avec $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- On estime σ^2 par $S_n^2(\mathbf{X}) = \overline{X^2} - \overline{X}^2$. C'est un ▸ estimateur de type moment
- C'est un estimateur biasé: $\mathbb{E}(S_n^2(\mathbf{X})) = \frac{n-1}{n}\sigma^2$
- On peut considérer un autre estimateur (sans biais)

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

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

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

Pour la culture (et la suite)

- Il existe différentes techniques d'estimation, la technique des moindre-carrés n'est qu'un exemple
- On peut citer la technique du maximum de vraisemblance qui est une des plus utilisées
- Les estimateurs de type moment qui consistent à identifier moments théoriques et empiriques
- L'estimation robuste qui permet de limiter l'influence des outliers

▸ maximum de vraisemblance ▸ estimateur de type moment

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 que le vrai paramètre ne soit pas compris dans l'intervalle considéré
- On cherche un ensemble de valeurs que l'on peut raisonnablement attribuer au paramètre.

Construction de l'intervalle

- Considérons l'estimation du paramètre d'espérance μ d'une loi gaussienne à partir d'un échantillon $\mathbf{X}$.
- On souhaite quantifier la probabilité que le paramètre μ tombe dans un intervalle au vu des données:

$$\mathbb{P}\{\mu \in [A(\mathbf{X}), B(\mathbf{X})]\}$$

- On souhaiterait ensuite controler cette probabilité, en fixant un risque α que le "vrai" paramètre n'appartienne pas à l'intervalle:

$$\mathbb{P}\{\mu \in [A(\mathbf{X}), B(\mathbf{X})]\} = 1 - \alpha$$

- L'intervalle considéré est un **intervalle aléatoire** dont les bornes varient d'un échantillon à un autre.

Vers les intervalles de confiance

- Pour déterminer les bornes de l'intervalle, on va considérer la loi de l'estimateur du paramètre:

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

- Etant donné la loi de l'estimateur, on connait ses quantiles. On note souvent u_α le quantile d'ordre α de la loi gaussienne centrée réduite.
- En supposant σ connue, on a donc:

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

- On peut donc construire un intervalle de confiance de μ à partir de l'intervalle de dispersion de son estimateur

Intervalle de confiance pour le paramètre d'espérance à variance connue

- A partir de l'intervalle de dispersion de $\widehat{\mu}(\mathbf{X})$:

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

- 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:

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

Intervalle de confiance pour le paramètre d'espérance à variance inconnue

- Si la variance est inconnue, on doit l'estimer avec $S^2(\mathbf{X})$.
- Dans ce cas les fluctuations de l'estimateur autour de son espérance n'ont plus la même loi:

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

- Les bornes de l'intervalle sont donc modifiées et utilisent les quantiles de la ▸ loi de Student :

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

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 $\alpha = 5\%$

$$IC_{1-\alpha}(\mu_1) = [24.82, 26.49] \qquad IC_{1-\alpha}(\mu_2) = [23.62, 25.66]$$

- Est ce que :

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

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:

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

- Cette approximation donne les bornes de l'intervalle de confiance d'une proportion tel que: $\mathbb{P}\{p \in [A(\mathbf{X}), B(\mathbf{X})]\} = 1 - \alpha$
- 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

$$\frac{\widehat{p}(\mathbf{X})(1-\widehat{p}(\mathbf{X}))}{n}$$

Outline

Variables aléatoires discrètes

- Les valeurs prises ($X(\Omega)$) par les variables sont discrètes (comptages, modalités,...)
- Leur loi est complètement déterminée par $\mathbb{P}(X = k)$ pour tous les k dans $X(\Omega)$, avec

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

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

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

- L'espérance et la variance des lois discrètes sont définies par

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

- Exemples: loi de Bernoulli, Binomiale, de Poisson, Geométrique

Variables aléatoires continues

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

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

- L'espérance et la variance des lois continues sont définies par

$$\mathbb{E}(X) = \int_{X(\Omega)} xf(x)dx, \qquad \mathbb{V}(X) = \int_{X(\Omega)} (x - \mathbb{E}(X))^2 f(x)dx$$

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

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...)

◂ Variables Aléatoires

La loi du chi2 et loi des sommes de carrés

- C'est la loi du carré de variables gaussiennes: si $X_i \sim \mathcal{N}(0,1)$ alors $X_i^2 \sim \chi^2(1)$
- Si on considère $(X_1, ...X_k)$ iid de même loi $\mathcal{N}(0, \sigma^2)$ alors

$$S^2 = \sum_i X_i^2 \sim \sigma^2 \chi^2(k), \quad \mathbb{E}(S^2) = k, \quad \mathbb{V}(S^2) = 2k$$

- Si on considère un vecteur aléatoire gaussien iid $\mathbf{X}$, la loi du χ^2 est la loi de la norme $\|\mathbf{X}\|^2$
- k est appelé le degré de liberté (ddl, dF) : c'est le nombre de composantes **indépendantes** de S^2

◂ Estimation de la variance

Moments d'une loi

- On considère un modèle $\mathcal{M}_\theta$ de loi F_θ avec $\theta \in \mathbb{R}^d$ par exemple
- Les moments **théoriques** caratérisent totalement la loi F_θ, et sont définis par:

$$\mathbb{E}_\theta(X^\ell) = \int x^\ell f_\theta(x) dx, \qquad \mathbb{E}_\theta(X^\ell) = \sum x^\ell \mathbb{P}_\theta(X = x)$$

- Si on considère que θ paramétrise la loi de X alors on peut trouver une relation entre θ et $(\mathbb{E}_\theta(X^\ell))_\ell$
- Exemple dans le cas gaussien $\theta = (\mu, \sigma^2)$:

$$\mathbb{E}_\theta(X^1) = \mu, \qquad \mathbb{E}_\theta(X^2) = \sigma^2 + \left(\mathbb{E}_\theta(X^1)\right)^2$$

- En inversant le système on peut relier les moments aux paramètres

Moments empiriques / Moments théoriques

- Les moments **théoriques** caratérisent totalement une loi, et sont définis par:

$$\mathbb{E}_\theta(X^k) = \int x^k f_\theta(x) dx, \qquad \mathbb{E}_\theta(X^k) = \sum x^k \mathbb{P}_\theta(X = x)$$

Exemples : l'espérance μ et la variance σ^2

- Les moments **empiriques** sont des fonctions des observations:

$$m_k(x_1, \ldots, x_n) = \frac{1}{n} \sum_i x_i^k = \overline{x^k}$$

Exemples : la moyenne $\overline{x}$ et la variance empirique s^2

De l'intérêt du modèle

- Comment relier les moments empiriques et théoriques?

on suppose que la loi empirique des observations $(x_1, \ldots, x_n)$ peut être approchée par la loi théorique de $(X_1, \ldots, X_n)$ sous le modèle $\mathcal{M}_\theta$

- Quelles seraient les propriétés des moments empiriques sous $\mathcal{M}_\theta$?

$$M_k(X_1, \ldots, X_n) = \frac{1}{n} \sum_i X_i^k$$

- Sous $\mathcal{M}_\theta$ les moments empiriques sont des variables aléatoires !
- Exemple : $M_1(X_1, \ldots, X_n) = \overline{X}$, $S^2 = \overline{X^2} - M_1^2(X_1, \ldots, X_n)$

La loi de $M_k(X_1, \ldots, X_n)$ est déterminée par le modèle $\mathcal{M}_\theta$

◂ Estimation de la variance ◂ Pour la suite

Le modèle multinomial

- On note $N = (N_1, \ldots, N_k)$ un vecteur de comptages aléatoires t.q. $\sum_k N_k = n$.
- On dit que $N \sim \mathcal{M}(n, p_1, \ldots, p_k)$ si, pour $\sum_{j=1}^k p_j = 1$, et $\sum_{j=1}^k n_j = n$

$$\mathbb{P}\{N_1 = n_1, \ldots, N_k = n_k\} = \frac{n! p_1^{n_1} \ldots, p_k^{n_k}}{n_1! \ldots n_k!},$$

- Cette loi modélise n tirages consécutifs avec remise dans une urne à k catégories n_1 boules de type 1, ... n_k boules de type k, avec des boules de type j en proportion p_j dans l'urne.
- Du fait de la contrainte $\sum_{j=1}^k n_j = n$, les comptages de différentes catégories ne sont pas indépendants !

$$\text{cov}\left(N_j, N_{j'}\right) = -p_j p_{j'}$$

◂ Exemple de modèles pour les lois discrètes

Notion de vraisemblance

- On appelle vraisemblance du modèle $\mathcal{M}_\theta$ au vu de l'observation $(x_1, \ldots, x_n)$ la fonction de densité ayant servi à définir le modèle,

$$\mathcal{L}_x(\mathcal{M}_\theta) = \mathbb{P}_{\mathcal{M}_\theta}(x_1, \ldots, x_n)$$

- C'est une fonction de l'échantillon **x** et de θ. La démarche statistique consiste à considérer que c'est une fonction de θ, les données **x** étant considérées comme "fixées"
- Dans le cas d'un échantillon i.i.d. la vraisemblance devient:

$$\mathcal{L}_x(\theta) = \mathbb{P}_\theta(x_1, \ldots, x_n) \overset{iid}{=} \prod_{i=1}^{n} \mathbb{P}_\theta(x_i; \theta)$$

Exemple dans le cas Binomial

- On considère un échantillon de 20 patients dans lequel 5 malades sont observés
- p le pourcentage de malades dans la population
- Dans un modèle Binomial $\mathbb{P}(X = k) = C_n^k p^k (1-p)^{n-k}$
- La vraisemblance de l'échantillon :

$$\mathcal{L}(\mathbf{x}; p) = C_{20}^5 p^5 (1-p)^{20-5}$$

- $\mathcal{L}(\mathbf{x}; p = 0.1) = 0.03$, $\mathcal{L}(\mathbf{x}; p = 0.25) = 0.2$
- La probabilité jointe de l'échantillon est plus élevée quand la valeur du paramètre est 0.25.

Probabilités ou Statistiques ?

- La démarche est théorique, n'a besoin d'aucune "réalité"
- La démarche est déductive : on fait des hypothèses et on démontre des théorèmes mathématiques
- Fixe θ et étudie $\mathbb{P}_\theta(\mathbf{x})$

- La démarche est liée aux observations $\mathbf{x}$
- La démarche est inductive : elle part des $\mathbf{x}$ et remonte à la loi qui les a engendrées
- Dispose de $\mathbf{x}$ (fixe!), et étudie θ avec $\mathcal{L}_x(\theta)$

On parle de **probabilité d'une observation**

On parle de **vraisemblance d'un modèle**

L'estimateur du maximum de vraisemblance

- Si on considère le modèle $\mathcal{M}_\theta$, plusieurs valeurs de θ sont possibles
- Lorsqu'on dispose d'observations, on peut alors chercher le "meilleur modèle", celui dont la vraisemblance est la meilleure:

$$\widehat{\theta}(\mathbf{x}) = \arg\max_{\theta \in \Theta} \{\mathcal{L}_x(\theta)\}$$

- La maximisation de la vraisemblance nécessite la résolution de l'équation:

$$\frac{\partial \log \mathcal{L}_x(\theta)}{\partial \theta} = 0$$

- L'estimateur du maximum de vraisemblance $\hat{\theta}(\mathbf{x})$ est la solution de cette équation

Exemple dans le cas Gaussien

- On observe $(x_1, \ldots, x_n)$ et on suppose que cet échantillon constitue une réalisations de n variables aléatoires indépendantes et identiquement distribuées (iid) $(X_1, \ldots, X_n)$, avec $X_i \sim \mathcal{N}(\mu, \sigma^2)$
- La vraisemblance du modèle est définie par

$$\mathbb{P}_{\mathcal{M}_\theta}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \frac{1}{\sigma^n \sqrt{2\pi}} \exp\left(-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} \right)$$

- L'estimateur du maximum de vraisemblance de l'espérance $\widehat{\mu}^{\mathsf{MV}}$ est obtenu en résolvant l'équation

$$\frac{\partial \log \mathbb{P}_{\mathcal{M}_\theta}(x_1, \ldots, x_n)}{\partial \mu} = 0$$

- Dans le cas Gaussien $\widehat{\mu}^{\mathsf{MV}} = \overline{X}$!

◂ Pour la suite

Loi de Student

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

$$T \sim \mathcal{T}(n)$$

- Les moments de cette loi sont:

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

◂ Intervalle de confiance