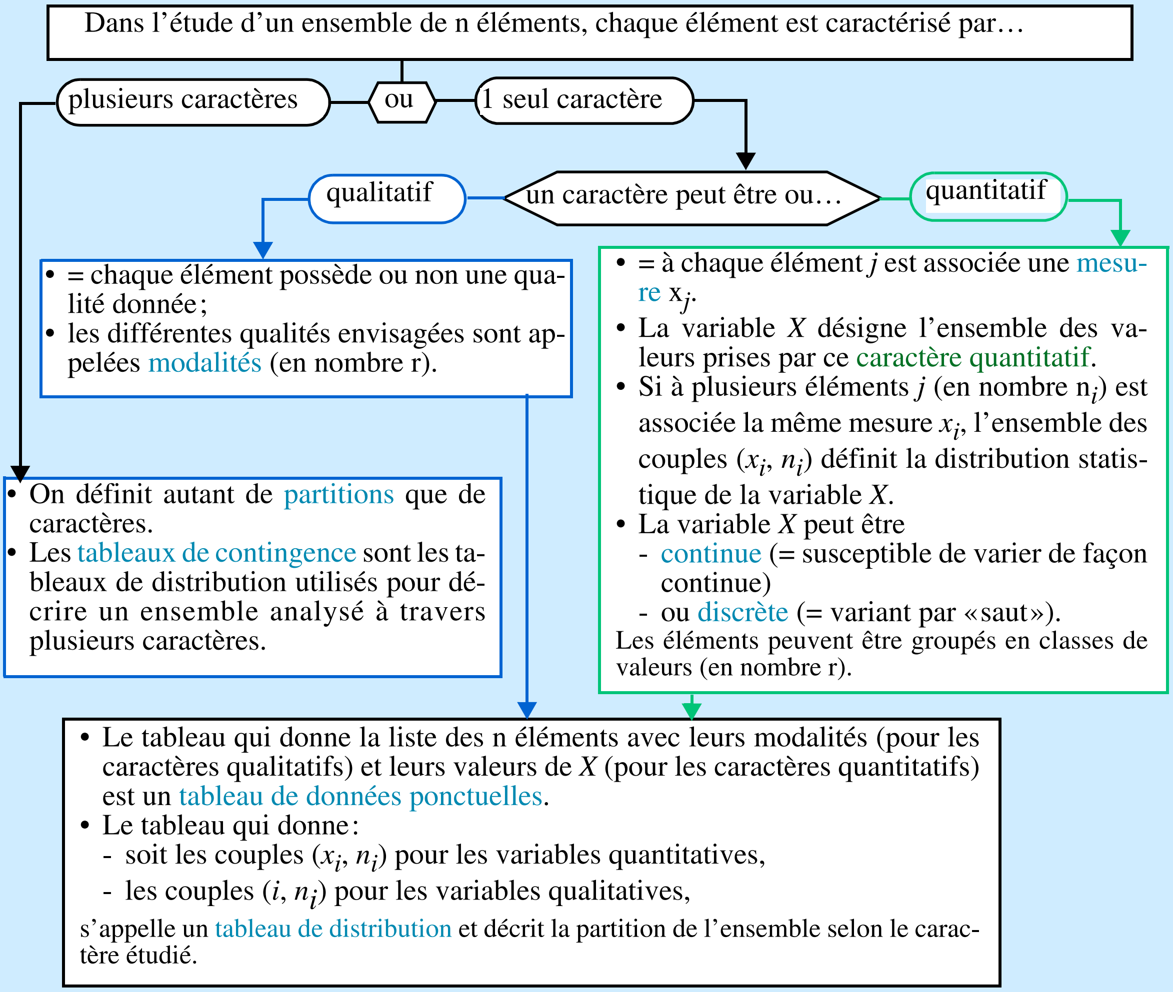
CHAPITRE XI

RÉCAPITULATION DES PRINCIPAUX RÉSUMÉS SYNOPTIQUES

RÉSUMÉ DU CHAPITRE I - STATISTIQUE DESCRIPTIVE



Dans l'étude d'un ensemble de n éléments, chaque élément est caractérisé par…

plusieurs caractères

ou

1 seul caractère

un caractère peut être ou…

qualitatif

quantitatif

- = chaque élément possède ou non une qualité donnée ;
- les différentes qualités envisagées sont appelées modalités (en nombre r).

- = à chaque élément j est associée une mesure x_j.
- La variable X désigne l'ensemble des valeurs prises par ce caractère quantitatif.
- Si à plusieurs éléments j (en nombre n_i) est associée la même mesure x_i, l'ensemble des couples (x_i, n_i) définit la distribution statistique de la variable X.
- La variable X peut être
  - continue (= susceptible de varier de façon continue)
  - ou discrète (= variant par «saut»).

Les éléments peuvent être groupés en classes de valeurs (en nombre r).

- On définit autant de partitions que de caractères.
- Les tableaux de contingence sont les tableaux de distribution utilisés pour décrire un ensemble analysé à travers plusieurs caractères.

- Le tableau qui donne la liste des n éléments avec leurs modalités (pour les caractères qualitatifs) et leurs valeurs de X (pour les caractères quantitatifs) est un tableau de données ponctuelles.
- Le tableau qui donne :
  - soit les couples (x_i, n_i) pour les variables quantitatives,
  - les couples (i, n_i) pour les variables qualitatives,

s'appelle un tableau de distribution et décrit la partition de l'ensemble selon le caractère étudié.

Observations faites à des dates différentes (ou sur des groupes différents)

- indice élémentaire : $\mathbf{I}(X)_{t_1/t_0} = \frac{x_{t_1}}{x_{t_0}} \times 100$
- indice synthétique pour résumer un ensemble d'indices élémentaires

		Indice	
		des prix	des quantités
Indice	de Laspeyres	$\mathbf{L}(p)_{t_1/t_0} = \frac{\sum_{i=1}^{n} p_{i,t_1} q_{i,t_0}}{\sum_{j=1}^{n} p_{j,t_0} q_{j,t_0}}$	$\mathbf{L}(q)_{t_1/t_0} = \frac{\sum_{i=1}^{n} p_{i,t_0} q_{i,t_1}}{\sum_{j=1}^{n} p_{j,t_0} q_{j,t_0}}$
	de Paasche	$\mathbf{P}(p)_{t_1/t_0} = \frac{\sum_{i=1}^{n} p_{i,t_1} q_{i,t_1}}{\sum_{j=1}^{n} p_{j,t_0} q_{j,t_1}}$	$\mathbf{P}(q)_{t_1/t_0} = \frac{\sum_{i=1}^{n} p_{i,t_1} q_{i,t_1}}{\sum_{j=1}^{n} p_{j,t_1} q_{j,t_0}}$

Calculette Statistique
Index
Table des Matières

Récapitulation des principaux résumés synoptiques

Problème : synthétiser les informations relatives à un ensemble d'éléments examinés dans l'optique de

variable...

... qualitative

- Synthèse numérique: fréquences n_i/n
- Synthèse par l'image:
  - Diagramme à secteurs (rect. ou circ.)
  - Diagramme à barres
  - Courbe de Pareto

...quantitative

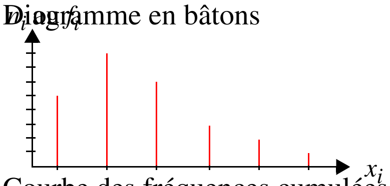
synthèse..

... graphique

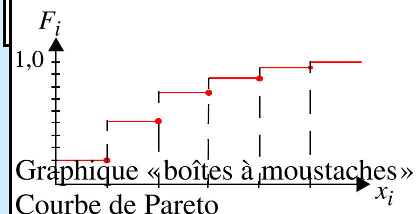
... par des paramètres

Variable discrète

Diagramme en bâtons



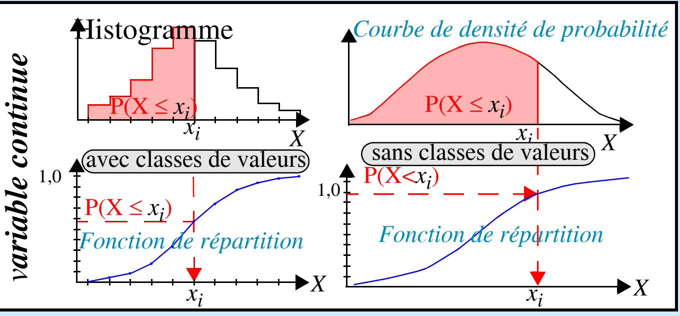
Courbe des fréquences cumulées



Graphique «boîtes à moustaches»

Courbe de Pareto

variable continue



étude d'une seule variable

non

oui

Paramètres de...

..tendance centrale

...dispersion

- Moyenne:
  - $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ (tableau de données ponctuelles)
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \sum_{i=1}^{r} f_i x_i$ (tableau de distribution)
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.} \cdot x_i$ (tableau de contingence)
- Médiane: M_e = valeur de X telle qu'il y ait autant d'observations ayant une valeur inférieure (ou égale) à M_e que d'observations ayant une valeur supérieure à M_e: $F(X \le M_e) = 0{,}5$.
- Quartile: Q_1 tel que $F(X \le Q_1) = 0{,}25$; Q_3 tel que $F(X \le Q_3) = 0{,}25$; $Q_2 = M_e$
- Mode: M_o = valeur de X dont la fréquence observée est maximale.

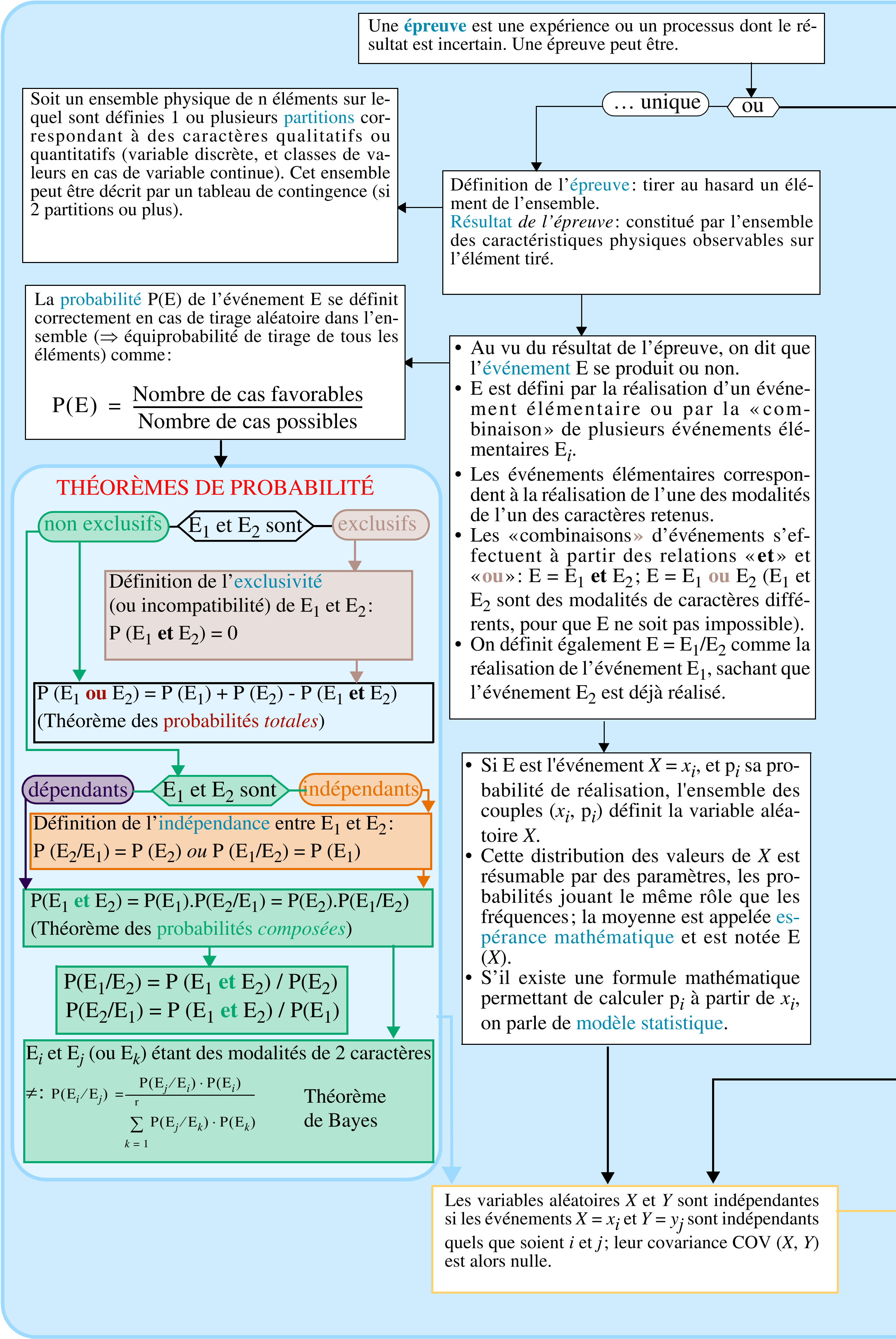
- Variance: V(X) ou σ^2:

		Tableau de	
		données ponctuelles	distribution
relation de	définition	$\frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})^2$	$\frac{1}{n}\sum_{i=1}^{r}n_i(x_i-\bar{x})^2$ $= \sum_{i=1}^{r}f_i(x_i-\bar{x})^2$
	calcul	$\frac{1}{n}\sum_{j=1}^{n}x_j^2-\bar{x}^2$	$\frac{1}{n}\sum_{i=1}^{r}n_ix_i^2-\bar{x}^2$ $= \sum_{i=1}^{r}f_ix_i^2-\bar{x}^2$

- Écart-type $\sigma = \sqrt{\sigma^2}$
- Coefficient de variation: $\sigma / \bar{x}$
- Intervalle interquartile: $Q_3 - Q_1$

- Calcul de la covariance entre X et Y:
  - $\mathrm{COV}(X, Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}$ (tableau de données ponctuelles)
  - $\mathrm{COV}(X, Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \bar{x}\cdot\bar{y}$ (tableau de contingence)
- $Z = \mathbf{a}X + \mathbf{b}Y \Rightarrow \left| \begin{array}{l} \bar{z} = \mathbf{a}\bar{x} + \mathbf{b}\bar{y} \\ \mathrm{V}(Z) = \mathbf{a}^2\mathrm{V}(X) + \mathbf{b}^2\mathrm{V}(Y) + 2\mathbf{ab}\,\mathrm{COV}(X,Y) \end{array} \right.$

RÉSUMÉ DU CHAPITRE II - INTRODUCTION AUX PROBABILITÉS



Une **épreuve** est une expérience ou un processus dont le résultat est incertain. Une épreuve peut être.

… unique — ou

Définition de l'épreuve : tirer au hasard un élément de l'ensemble.
Résultat *de l'épreuve* : constitué par l'ensemble des caractéristiques physiques observables sur l'élément tiré.

Soit un ensemble physique de n éléments sur lequel sont définies 1 ou plusieurs partitions correspondant à des caractères qualitatifs ou quantitatifs (variable discrète, et classes de valeurs en cas de variable continue). Cet ensemble peut être décrit par un tableau de contingence (si 2 partitions ou plus).

- Au vu du résultat de l'épreuve, on dit que l'événement E se produit ou non.
- E est défini par la réalisation d'un événement élémentaire ou par la « combinaison » de plusieurs événements élémentaires E_i.
- Les événements élémentaires correspondent à la réalisation de l'une des modalités de l'un des caractères retenus.
- Les « combinaisons » d'événements s'effectuent à partir des relations « **et** » et « **ou** » : E = E_1 **et** E_2 ; E = E_1 **ou** E_2 (E_1 et E_2 sont des modalités de caractères différents, pour que E ne soit pas impossible).
- On définit également E = E_1/E_2 comme la réalisation de l'événement E_1, sachant que l'événement E_2 est déjà réalisé.

La probabilité P(E) de l'événement E se définit correctement en cas de tirage aléatoire dans l'ensemble (⇒ équiprobabilité de tirage de tous les éléments) comme :

$$P(E) = \frac{\text{Nombre de cas favorables}}{\text{Nombre de cas possibles}}$$

THÉORÈMES DE PROBABILITÉ

non exclusifs — E_1 et E_2 sont — exclusifs

Définition de l'exclusivité (ou incompatibilité) de E_1 et E_2 :
$P(E_1 \textbf{ et } E_2) = 0$

$P(E_1 \textbf{ ou } E_2) = P(E_1) + P(E_2) - P(E_1 \textbf{ et } E_2)$
(Théorème des probabilités *totales*)

dépendants — E_1 et E_2 sont — indépendants

Définition de l'indépendance entre E_1 et E_2 :
$P(E_2/E_1) = P(E_2)$ *ou* $P(E_1/E_2) = P(E_1)$

$P(E_1 \textbf{ et } E_2) = P(E_1).P(E_2/E_1) = P(E_2).P(E_1/E_2)$
(Théorème des probabilités *composées*)

$P(E_1/E_2) = P(E_1 \textbf{ et } E_2) / P(E_2)$
$P(E_2/E_1) = P(E_1 \textbf{ et } E_2) / P(E_1)$

E_i et E_j (ou E_k) étant des modalités de 2 caractères ≠ :

$$P(E_i/E_j) = \frac{P(E_j/E_i) \cdot P(E_i)}{\sum_{k=1}^{r} P(E_j/E_k) \cdot P(E_k)}$$

Théorème de Bayes

- Si E est l'événement $X = x_i$, et p_i sa probabilité de réalisation, l'ensemble des couples (x_i, p_i) définit la variable aléatoire X.
- Cette distribution des valeurs de X est résumable par des paramètres, les probabilités jouant le même rôle que les fréquences ; la moyenne est appelée espérance mathématique et est notée $E(X)$.
- S'il existe une formule mathématique permettant de calculer p_i à partir de x_i, on parle de modèle statistique.

Les variables aléatoires X et Y sont indépendantes si les événements $X = x_i$ et $Y = y_j$ sont indépendants quels que soient i et j ; leur covariance COV (X, Y) est alors nulle.

Index

Table des Matières

… constituée de plusieurs épreuves élémentaires

Ensemble physique n'existant pas; 3 causes:

- non-matérialité des éléments de l'ensemble: matérialisation par liste des résultats possibles (faire très attention à leur équiprobabilité et se ramener au cas de l'ensemble physique),
- référence au passé: P(E) = fréquence observée sur le passé,
- probabilité subjective définie a priori comme la «chance» que l'on donne à la réalisation de E.

- À chaque épreuve élémentaire, on associe un ensemble d'événements élémentaires exclusifs. On définit donc une partition par épreuve élémentaire.
- L'événement est défini à partir d'une «combinaison» d'événements élémentaires (comme dans le cas d'une épreuve unique). Sa probabilité se calcule à partir des théorèmes de probabilités et des résultats à des problèmes de dénombrement.

Problèmes de dénombrement

- nombre P_n de permutations (nombre de *mots* différents de k lettres différentes, à partir d'un alphabet de n = k lettres différentes): $P_n = k!$
- nombre A_n^k d'arrangements (nombre de *mots* différents de k lettres différentes, à partir d'un alphabet de n lettres différentes):

$$A_n^k = \frac{n!}{(n-k)!}$$

- nombre C_n^k de combinaisons (nombre de *groupes* différents de k lettres différentes, à partir d'un alphabet de n lettres différentes):

$$C_n^k = \frac{n!}{(n-k)!\,k!}$$

$$Y = k \cdot X \Rightarrow \left| \begin{array}{l} E(Y) = k \cdot E(X) \\ V(Y) = k^2 \cdot V(X) \end{array} \right.$$

$$Z = \mathbf{a}X + \mathbf{b}Y \Rightarrow \left| \begin{array}{l} E(Z) = \mathbf{a} \cdot E(X) + \mathbf{b} \cdot E(Y) \\ V(Z) = \mathbf{a}^2 V(X) + \mathbf{b}^2 V(Y) \\ \qquad + 2\mathbf{ab}\,COV(X,Y) \end{array} \right.$$

RÉSUMÉ DU CHAPITRE III - LES PRINCIPALES DISTIBUTIONS

LOI BINOMIALE $\mathscr{B}$(n ; p)

FORME DE LA LOI

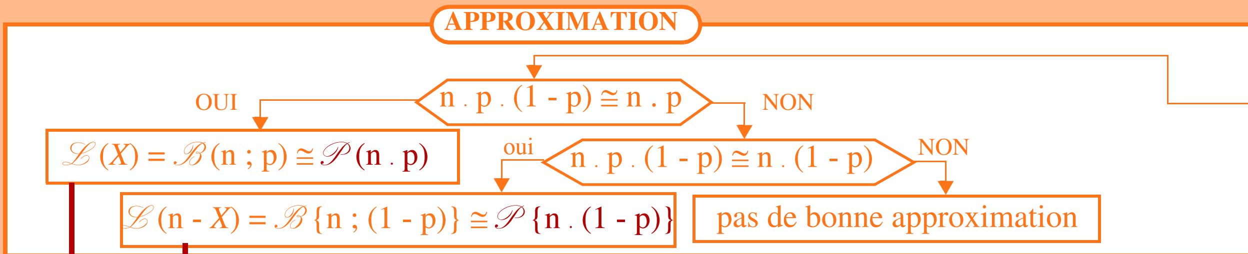
- $P(X = x) = C_n^x \cdot p^x \cdot (1-p)^{n-x} = \frac{n!}{x! \cdot (n-x)!} \cdot p^x \cdot (1-p)^{n-x}$
- $P(f_n = f = \frac{x}{n}) \equiv P(X = x)$: loi Binomiale en proportion

PROPRIÉTÉS DE LA LOI

$E(X) = n \,.\, p$	$V(X) = n \,.\, p \,.\, (1-p)$
$E(f_n) = p$	$V(f_n) = \frac{p \cdot (1-p)}{n}$

$\mathscr{L}(X_1) = \mathscr{B}(n_1 ; p)$
$\mathscr{L}(X_2) = \mathscr{B}(n_2 ; p)$ $\Rightarrow \mathscr{L}(X_1+X_2) = \mathscr{B}(n_1+n_2 ; p)$

APPROXIMATION



$n \,.\, p \,.\, (1 - p) \cong n \,.\, p$

OUI

$\mathscr{L}(X) = \mathscr{B}(n ; p) \cong \mathscr{P}(n \,.\, p)$

NON

$n \,.\, p \,.\, (1 - p) \cong n \,.\, (1 - p)$

oui

$\mathscr{L}(n - X) = \mathscr{B}\{n ; (1 - p)\} \cong \mathscr{P}\{n \,.\, (1 - p)\}$

NON

pas de bonne approximation

LOI de POISSON $\mathscr{P}(\lambda)$

GÉNÉRATION ET SYMBOLES UTILISÉS

Génération $\{$ processus de Poisson
approximation d'une loi Binomiale

λ: paramètre de Poisson

FORME DE LA LOI

$P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$

PROPRIÉTÉS DE LA LOI

$$\left.\begin{array}{l} E(X) = V(X) = \lambda \\ \frac{P(X = x)}{P(X = x - 1)} = \frac{\lambda}{x} \end{array}\right\}$$ utilisées dans la reconnaissance empirique d'une loi de Poisson

$\mathscr{L}(X_1) = \mathscr{P}(\lambda_1)$
$\mathscr{L}(X_2) = \mathscr{P}(\lambda_2)$ $\Rightarrow \mathscr{L}(X_1+X_2) = \mathscr{P}(\lambda_1+\lambda_2)$

APPROXIMATION

$\lambda > 20 \Rightarrow$
$\mathscr{L}(X) = \mathscr{P}(\lambda) \cong \mathscr{N}(\lambda, \sqrt{\lambda})$

LOI DE STUDENT $\mathscr{L}(X) = \mathscr{S}t_{\nu}(\bar{x}, \sigma)$

- Symboles:
  - les mêmes que ceux de la loi Normale, plus ν (= nombre de degrés de liberté),
  - loi continue tabulée utilisée pour l'estimation dans les petits échantillons.
- Approximation: $\nu > 30 \Rightarrow \mathscr{S}t_{\nu}(\bar{x}, \sigma) \cong \mathscr{N}(\bar{x}, \sigma)$.
- Remarques sur les problèmes possibles: 5 données interviennent: α, $\bar{x}$, σ, x_{α}, ν; la connaissance de 4 d'entre elles engendre celle de la 5^e.

Index

Table des Matières

THÉORIQUES.

LOI BINOMIALE $\mathcal{B}$(n ; p)

GÉNÉRATION ET SYMBOLES UTILISÉS

- n = nombre d'épreuves indépendantes de même type
- p = probabilité de réalisation de l'événement étudié, au cours de l'une quelconque des épreuves
- X = variable aléatoire définie comme le nombre observable de réalisations de l'événement étudié au cours des n épreuves ($0 \leq X \leq n$)

  ⇒ variable aléatoire n – X = nombre observable de non-réalisations de l'événement étudié au cours des n épreuves ($0 \leq n - X \leq n$)
- $f_n = X/n$ = variable aléatoire définie comme la fréquence observable d'épreuves conduisant à la réalisation de l'événement ($0 \leq f_n \leq 1$)
- x = nombre de réalisations de l'événement étudié au cours des n épreuves
- f $= x$ / n = fréquence observée de réalisation de l'événement étudié au cours des n épreuves

APPROXIMATION

NON

$$n > 5 \ \text{ET} \ \left| \sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}} \right| \cdot \frac{1}{\sqrt{n}} < 0,3$$

OUI

$\mathcal{L}(X) = \mathcal{B}(n\,;\,p) \cong \mathcal{N}\{n\,p\,;\sqrt{np(1-p)}\}$	$\mathcal{L}(f_n) = \mathcal{B}(n\,;\,p) \cong \mathcal{N}\left\{p\,;\sqrt{\frac{p(1-p)}{n}}\right\}$

LOI NORMALE : $\mathcal{L}(X) = \mathcal{N}(\bar{x}, \sigma)$

GÉNÉRATION ET SYMBOLES UTILISÉS

- X = variable aléatoire continue de moyenne $\bar{x}$ et d'écart-type σ ; $-\infty < X < +\infty$
- effet d'un très grand nombre de causes indépendantes, à effets additifs, chacune d'entre elles ayant un effet négligeable par rapport à l'ensemble des autres.
- $T = \dfrac{X - \bar{x}}{\sigma}$; variable centrée réduite associée à X ; de moyenne $\bar{t} = 0$ et d'écart-type = 1

FORME DE LA LOI

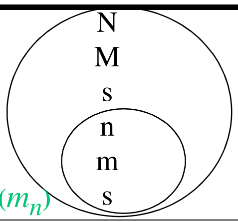
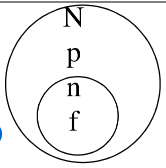
$$P(X < x_\alpha) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_\alpha} e^{-\left(\frac{x-\bar{x}}{\sigma}\right)^2/2} dx = \alpha$$; loi tabulée.

- $\bar{x} = M_e = M_o$; courbe symétrique par rapport à un axe perpendiculaire à $\bar{x}$
- $\left.\begin{array}{l}\mathcal{L}(X_1) = \mathcal{N}(\bar{x}_1;\sigma_1)\\ \mathcal{L}(X_2) = \mathcal{N}(\bar{x}_2;\sigma_2)\end{array}\right| \Rightarrow \mathcal{L}(X) = \mathcal{L}(X_1 \pm X_2) = \mathcal{N}\left([\bar{x}_1 \pm \bar{x}_2];\sqrt{\sigma_1^2 + \sigma_2^2}\right)$; généralisable à la somme d'un nombre quelconque de variables suivant une loi Normale.
- Théorème de la limite centrale :

  $$Y = \sum_{i=1}^{n} X_i \Rightarrow \mathcal{L}(Y) = \mathcal{N}(m;\sigma), \text{ avec } m = \sum_{i=1}^{n} \bar{x}_i \text{ et } \sigma^2 = \sum_{i=1}^{n} \sigma_i^2 \text{ si } n \geq 30.$$
- loi remplacée par la loi de Student lorsque σ est estimé (cf. chapitre IV)
- Remarques sur les problèmes possibles : 4 données interviennent : α, $\bar{x}$, σ, x_α ; la connaissance de 3 d'entre elles engendre celle de la 4^e^.

RÉSUMÉ DU CHAPITRE IV - L'INDUCTION STATISTIQUE.

Un échantillon de n éléments se définit comme l'un des sous-ensembles possibles tiré d'une population-mère de N éléments

		VARIABLES	
		ALÉATOIRES	CERTAINES
CARACTÈRES	QUANTITATIFS	X, m_n, s_n	N M s n m s $\sigma^{*2}=V(m_n)$
	QUALITATIFS	f_n	N p n f $\sigma^{*2}=V(f_n)$

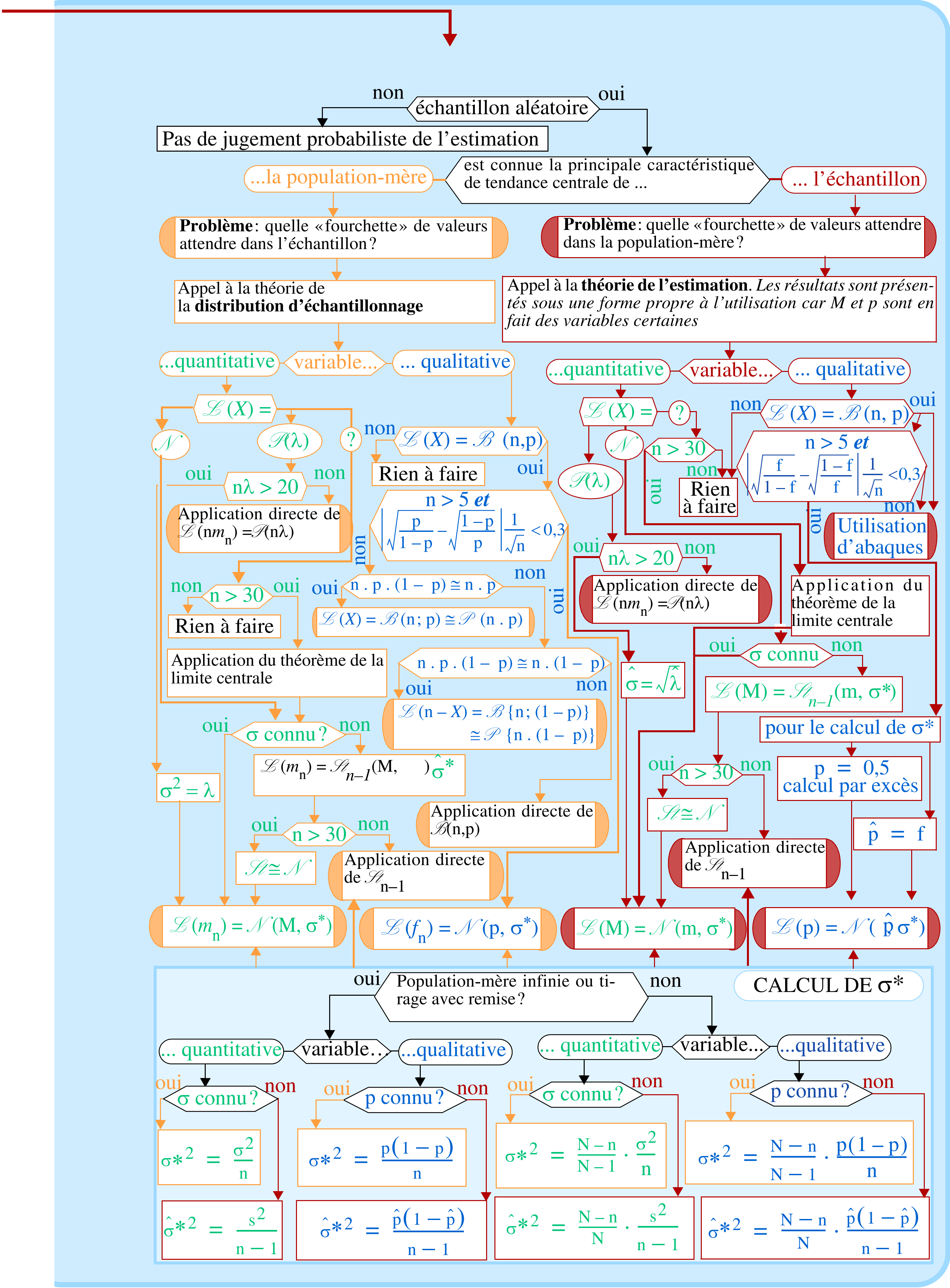
Un tirage est dit exhaustif lorsque l'on prélève les éléments simultanément, ou successivement sans remise de l'élément qui vient d'être tiré dans la population-mère.

- Un échantillon aléatoire est composé d'éléments tirés «au hasard» dans la population-mère, c'est-à-dire ayant tous la même probabilité a priori d'appartenir à l'échantillon; cette égalité de chance suppose un recensement préalable (= base de sondage) de la population-mère.
- Les sondages aléatoires sont à un degré (utilisation de table de nombres au hasard, tirage systématique) ou à plusieurs degrés.
- Les échantillons empiriques (méthodes des quotas…) sont utilisés en l'absence de base de sondage et, en toute rigueur, ne permettent pas de jugement en probabilité.

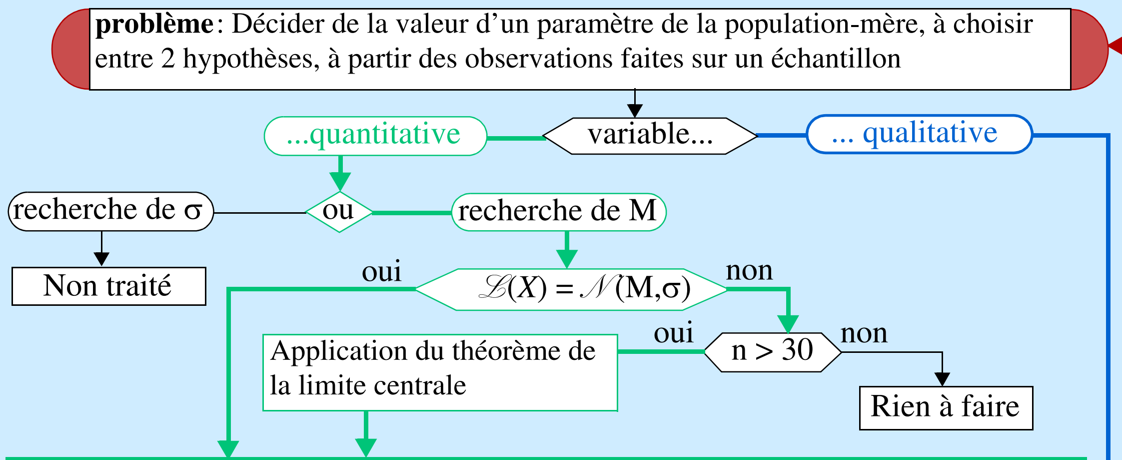
- Il existe C_N^n échantillons différents de taille n, tirés d'une population de taille N. Chacun est caractérisable par des paramètres.
- La distribution d'échantillonnage d'un paramètre donné (ex.: $\bar{x}$) est la distribution statistique de ce paramètre dans tous les échantillons possibles de taille n, ce qui permet de définir la variable aléatoire $\bar{x}_n$ (ou m_n) lorsque le tirage est aléatoire.
- Cette distribution d'échantillonnage du paramètre étudié peut à son tour être caractérisée par des paramètres, comme toute distribution statistique.
- Il ne faut pas confondre, pour une variable quantitative, la variance de la population-mère (σ^2), la variance observée sur un échantillon donné (s^2) et la variance de la distribution d'échantillonnage des moyennes observables dans un échantillon (σ^{*2}).

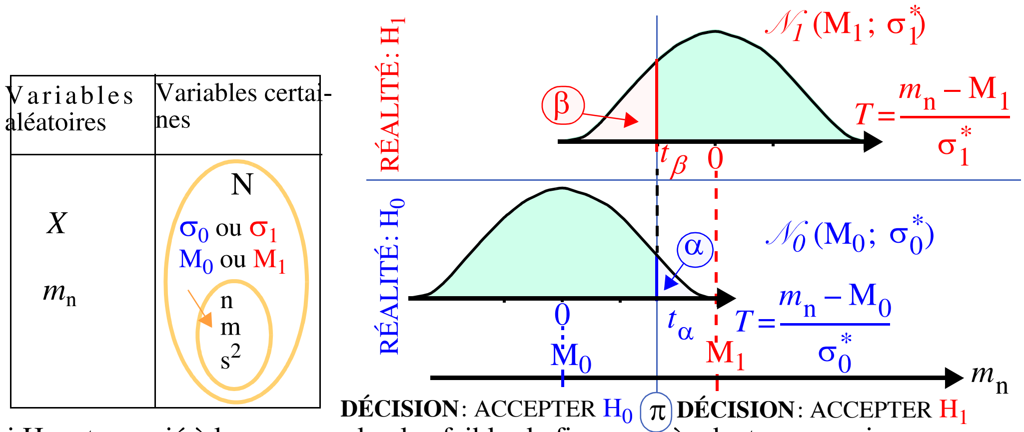
- L'estimateur $\hat{\theta}_n$ est une approximation d'un paramètre inconnu θ de la population-mère, à partir des données d'un échantillon aléatoire de taille n.
- Une estimation ponctuelle est la valeur prise par un estimateur dans un échantillon particulier.
- Principales qualités d'un estimateur:
  - sans biais: $E(\hat{\theta}_n) = \theta$
  - convergent: $V(\hat{\theta}_n) \to 0$ si $n \to N$
  - à variance $V(\hat{\theta}_n)$ minimale dans les échantillons de taille n

  sans biais + convergent = estimateur absolument correct; avec variance minimale = estimateur efficace
- Principales estimations à utiliser: $\hat{p} = f$; $\hat{M} = m$; $\hat{\sigma}^2 = \dfrac{ns^2}{n-1}$

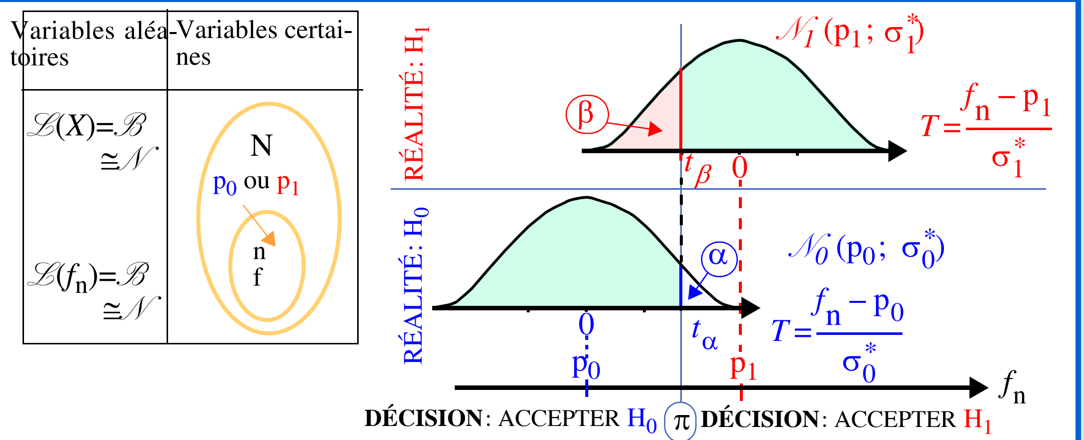

non
échantillon aléatoire
oui
Pas de jugement probabiliste de l'estimation
...la population-mère
est connue la principale caractéristique de tendance centrale de ...
... l'échantillon
Problème: quelle «fourchette» de valeurs attendre dans l'échantillon?
Problème: quelle «fourchette» de valeurs attendre dans la population-mère?
Appel à la théorie de la **distribution d'échantillonnage**
Appel à la **théorie de l'estimation**. *Les résultats sont présentés sous une forme propre à l'utilisation car M et p sont en fait des variables certaines*
...quantitative
variable...
... qualitative
$\mathcal{L}(X) =$
$\mathcal{N}$
$\mathcal{P}(\lambda)$
?
non
$\mathcal{L}(X) = \mathcal{B}$ (n,p)
Rien à faire
oui
non
$n\lambda > 20$
oui
Application directe de $\mathcal{L}(nm_n) = \mathcal{P}(n\lambda)$
n > 5 *et* $\left|\sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}}\right| \frac{1}{\sqrt{n}} < 0{,}3$
non
non
n > 30
oui
oui
$n \cdot p \cdot (1-p) \cong n \cdot p$
non
oui
Rien à faire
$\mathcal{L}(X) = \mathcal{B}(n;p) \cong \mathcal{P}(n \cdot p)$
Application du théorème de la limite centrale
$n \cdot p \cdot (1-p) \cong n \cdot (1-p)$
oui
non
$\mathcal{L}(n-X) = \mathcal{B}\{n;(1-p)\} \cong \mathcal{P}\{n \cdot (1-p)\}$
oui
σ connu?
non
$\sigma^2 = \lambda$
$\mathcal{L}(m_n) = \mathcal{S}_{n-1}(M, \hat{\sigma}^*)$
Application directe de $\mathcal{B}$(n,p)
oui
n > 30
non
$\mathcal{S} \cong \mathcal{N}$
Application directe de $\mathcal{S}_{n-1}$
$\mathcal{L}(m_n) = \mathcal{N}(M, \sigma^*)$
$\mathcal{L}(f_n) = \mathcal{N}(p, \sigma^*)$
...quantitative
variable...
... qualitative
$\mathcal{L}(X) =$
?
non
$\mathcal{L}(X) = \mathcal{B}$(n, p)
oui
$\mathcal{N}$
n > 30
non
Rien à faire
$\mathcal{P}(\lambda)$
oui
n > 5 *et* $\left|\sqrt{\frac{f}{1-f}} - \sqrt{\frac{1-f}{f}}\right| \frac{1}{\sqrt{n}} < 0{,}3$
oui
non
Utilisation d'abaques
oui
$n\lambda > 20$
non
Application directe de $\mathcal{L}(nm_n) = \mathcal{P}(n\lambda)$
Application du théorème de la limite centrale
oui
σ connu
non
$\hat{\sigma} = \sqrt{\hat{\lambda}}$
$\mathcal{L}(M) = \mathcal{S}_{n-1}(m, \sigma^*)$
pour le calcul de σ*
oui
n > 30
non
p = 0,5 calcul par excès
$\mathcal{S} \cong \mathcal{N}$
$\hat{p} = f$
Application directe de $\mathcal{S}_{n-1}$
$\mathcal{L}(M) = \mathcal{N}(m, \sigma^*)$
$\mathcal{L}(p) = \mathcal{N}(\hat{p}\ \sigma^*)$
CALCUL DE σ*
oui
Population-mère infinie ou tirage avec remise?
non
... quantitative
variable...
...qualitative
... quantitative
variable...
...qualitative
oui
σ connu?
non
oui
p connu?
non
oui
σ connu?
non
oui
p connu?
non
$\sigma^{*2} = \frac{\sigma^2}{n}$
$\sigma^{*2} = \frac{p(1-p)}{n}$
$\sigma^{*2} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$
$\sigma^{*2} = \frac{N-n}{N-1} \cdot \frac{p(1-p)}{n}$
$\hat{\sigma}^{*2} = \frac{s^2}{n-1}$
$\hat{\sigma}^{*2} = \frac{\hat{p}(1-\hat{p})}{n-1}$
$\hat{\sigma}^{*2} = \frac{N-n}{N} \cdot \frac{s^2}{n-1}$
$\hat{\sigma}^{*2} = \frac{N-n}{N} \cdot \frac{\hat{p}(1-\hat{p})}{n-1}$

RÉSUMÉ DU CHAPITRE V - INTRODUCTION À LA THÉORIE



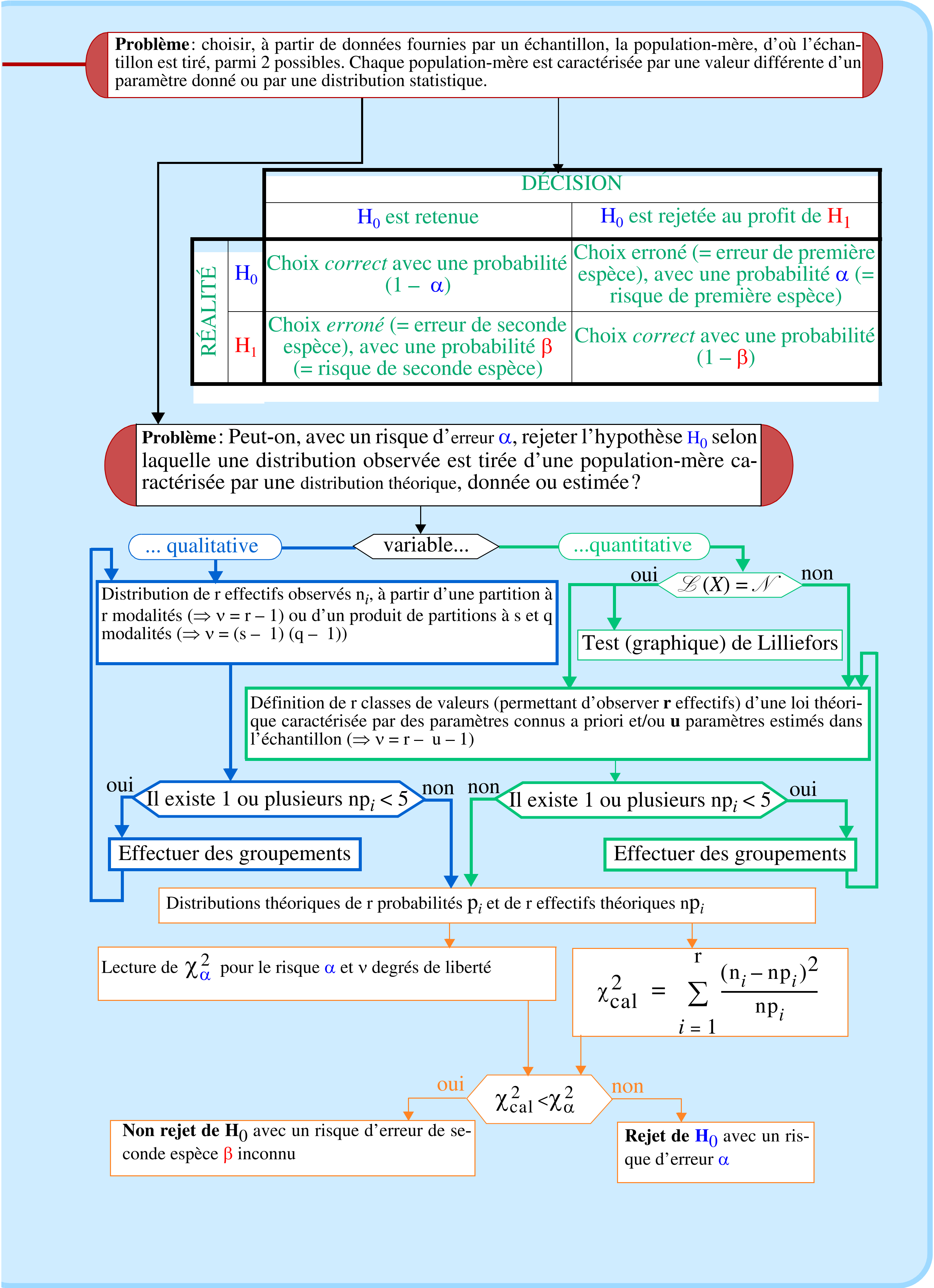


- Ici H_0 est associé à la moyenne la plus faible ; la figure est à adapter au cas inverse
- Double problème de distribution d'échantillonnage, chaque problème étant caractérisé par 5 paramètres, dont 2 communs :
  - le premier, où interviennent : M_0, σ_0, α, n, π,
  - le second, où interviennent : M_1, σ_1, β, n, π.
- si $\sigma = \sigma_0 = \sigma_1 = ?$, alors estimation de σ ($\hat{\sigma}^2 = \frac{n \cdot s^2}{(n-1)}$) et la loi Normale est remplacée par la loi de Student à $\nu = n - 1$ degrés de liberté.



- Ici H_0 est associé à la proportion la plus faible ; la figure est à adapter au cas inverse
- Double problème de distribution d'échantillonnage, chaque problème étant caractérisé par 4 paramètres, dont 2 communs :
  - le premier, où interviennent : p_0, α, n, π,
  - le second, où interviennent : p_1, β, n, π.

DES TESTS.



Problème : choisir, à partir de données fournies par un échantillon, la population-mère, d'où l'échantillon est tiré, parmi 2 possibles. Chaque population-mère est caractérisée par une valeur différente d'un paramètre donné ou par une distribution statistique.

		DÉCISION	
		H_0 est retenue	H_0 est rejetée au profit de H_1
RÉALITÉ	H_0	Choix *correct* avec une probabilité (1 – α)	Choix erroné (= erreur de première espèce), avec une probabilité α (= risque de première espèce)
	H_1	Choix *erroné* (= erreur de seconde espèce), avec une probabilité β (= risque de seconde espèce)	Choix *correct* avec une probabilité (1 – β)

Problème : Peut-on, avec un risque d'erreur α, rejeter l'hypothèse H_0 selon laquelle une distribution observée est tirée d'une population-mère caractérisée par une distribution théorique, donnée ou estimée ?

Distribution de r effectifs observés n_i, à partir d'une partition à r modalités (⇒ ν = r – 1) ou d'un produit de partitions à s et q modalités (⇒ ν = (s – 1) (q – 1))

Test (graphique) de Lilliefors

Définition de r classes de valeurs (permettant d'observer **r** effectifs) d'une loi théorique caractérisée par des paramètres connus a priori et/ou **u** paramètres estimés dans l'échantillon (⇒ ν = r – u – 1)

Effectuer des groupements

Effectuer des groupements

Distributions théoriques de r probabilités p_i et de r effectifs théoriques np_i

Lecture de χ^2_α pour le risque α et ν degrés de liberté

$$\chi^2_{cal} = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

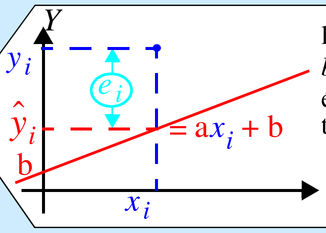
Non rejet de H_0 avec un risque d'erreur de seconde espèce β inconnu

Rejet de H_0 avec un risque d'erreur α

RÉSUMÉ DU CHAPITRE VI- L'AJUSTEMENT LINÉAIRE.

Etude simultanée de 2 caractères quantitatifs ⇒ observation de n couples (x_i, y_i) ou de triplets (x_i, y_j, n_{ij}), avec dépendance causale de X sur Y postulée pour des motifs d'ordre extra-statistique.

non — les n observations sont considérées comme constituant un échantillon — oui



oui — la dépendance causale se traduit par la liaison : $y_i = ax_i + b + e_i$ (régression linéaire de Y en X) où X et Y peuvent éventuellement avoir subi préalablement une transformation (élévation à une puissance, logarithme…) — non

Régression non linéaire : non traité

Problème : recherche d'une valeur pour *a* et *b* (car une infinité de droites peuvent prétendre, chacune d'entre elles, résumer correctement le nuage de points).

Un critère de choix possible peut être de trouver le couple (a, b) qui rend minimum $\sum_{i=1}^{n} e_i^2 \Rightarrow$

Résultats analytiques

méthode des moindres carrés :

$$a = \frac{\mathrm{COV}(X, Y)}{\mathrm{V}(X)} \quad \text{avec}$$

$$b = \bar{y} - a\bar{x}$$

tableau de données ponctuelles :

$$\mathrm{COV}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

tableau de contingence :

$$\mathrm{COV}(X, Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \bar{x}\cdot\bar{y}$$

Problème : analyse de la «qualité» de la liaison linéaire

. variance totale = variance-résiduelle + variance expliquée : $V(Y) \equiv \sigma_y^2 = \sigma_r^2 + \sigma_e^2$

. Variance résiduelle : $\sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - (ax_i + b)]^2 = (1 - r^2)\sigma_y^2$

. Coefficient de détermination : $r^2 = \frac{[\mathrm{COV}(X, Y)]^2}{\mathrm{V}(X)\cdot\mathrm{V}(Y)} = \frac{\sigma_e^2}{\sigma_y^2}$

. Coefficient de corrélation : $r = \frac{\mathrm{COV}(X, Y)}{\sqrt{\mathrm{V}(X)\cdot\mathrm{V}(Y)}} = \frac{\frac{1}{n}\sum x_i \cdot y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\sum y_i^2 - \bar{y}^2\right)}}$

- *Graphique* des résidus standardisés e_i/σ_r en fonction des x_i, pour détecter d'éventuelles erreurs de spécification et/ou des points aberrants.

Index

Table des Matières

Variables aléatoires	Variables certaines
ε_i a_n b_n r_n^2	N $y_i = \alpha x_i + \beta + \varepsilon_i$ ρ^2 n $y_i = ax_i + b + e_i$ r^2

Une première série de 4 hypothèses :

- - variables mesurées sans erreur
- - absence de biais systématique : $E(\varepsilon_i) = 0$
- - absence de liaison entre les écarts :
- $COV(\varepsilon_i, \varepsilon_j) = 0$, pour tout couple i,j
- - dispersion identique autour de y_i :
- $V(\varepsilon_i) = \sigma$, pour tout i ,

permet d'affirmer que :

- $\hat{\alpha} = a$ et $\hat{\beta} = b$;
- $\hat{\sigma}^2 = \dfrac{\sum e_i^2}{n-2} = \dfrac{n}{n-2}(1 - r^2)V(Y)$

estimateurs absolument corrects et efficaces

- $\hat{\rho}^2 = 1 - \dfrac{(1 - r^2)(n - 1)}{n - 2}$: estimateur correct

L'hypothèse selon laquelle $\mathcal{L}(\varepsilon_i) = \mathcal{N}(0, \sigma)$ permet de fournir des estimations par intervalle de confiance. Appel à la **théorie de l'estimation**.*Les résultats sont présentés sous une forme propre à l'utilisation car α et β sont en fait des variables certaines*

- $\mathcal{L}(\alpha) = \mathcal{S}t_{n-2}(a, \hat{\sigma}_{a_n})$ avec $\hat{\sigma}_{a_n} = \sqrt{\dfrac{\hat{\sigma}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\dfrac{(1 - r^2)V(Y)}{(n-2)V(X)}}$
- $\mathcal{L}(\beta) = \mathcal{S}t_{n-2}(b, \hat{\sigma}_{b_n})$ avec $\hat{\sigma}_{b_n} = \sqrt{\dfrac{\hat{\sigma}^2 \sum\limits_{i=1}^{n} x_i^2}{n \sum\limits_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\dfrac{(1 - r^2)V(Y)}{(n-2)V(X)}[V(X) + \bar{x}^2]}$
- $\mathcal{L}(\hat{y}_j) = \mathcal{S}t_{n-2}(a \cdot x_j + b, \hat{\sigma}_j)$ avec

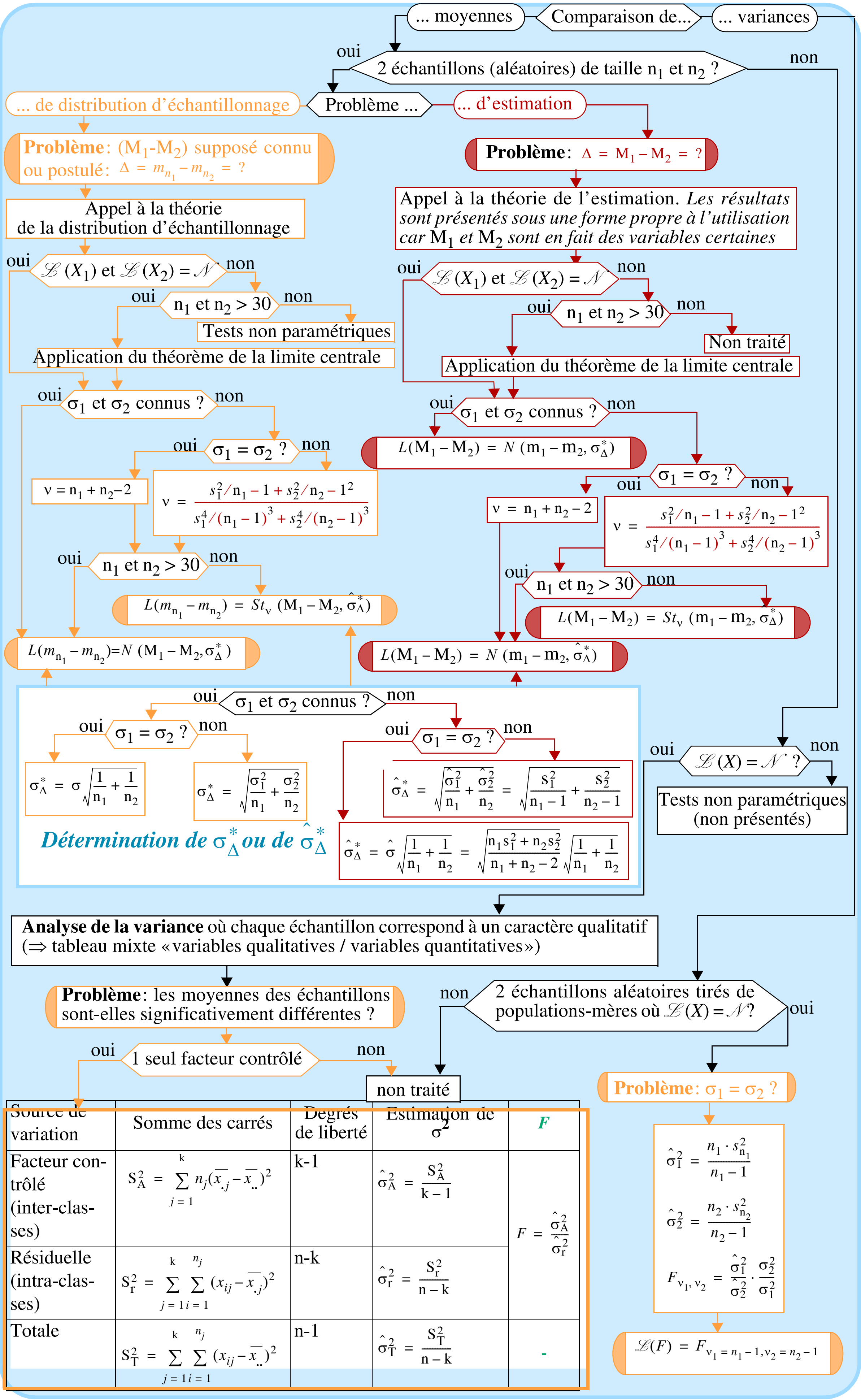
$$\hat{\sigma}_j = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{(1 - r^2)V(Y)}{(n-2)V(X)}[(1 + n)V(X) + (x_j - \bar{x})^2]}$$

- $P\left(\dfrac{-1 + e^{[2(r - t_\alpha.s)]}}{1 + e^{[2(r - t_\alpha.s)]}} < \rho < \dfrac{-1 + e^{[2(r + t_\alpha.s)]}}{1 + e^{[2(r + t_\alpha.s)]}}\right)$ =1-2α, avec $s = \sqrt{\dfrac{1}{n-1} + \dfrac{4 - r^2}{2(n-1)^2}}$ et n>11
- ρ^2 est significativement différent de 0 si la pente α l'est aussi.

Calculette Statistique
Index
Table des Matières

RÉSUMÉ DU CHAPITRE VII - COMPARAISON D'ÉCHANTILLONS

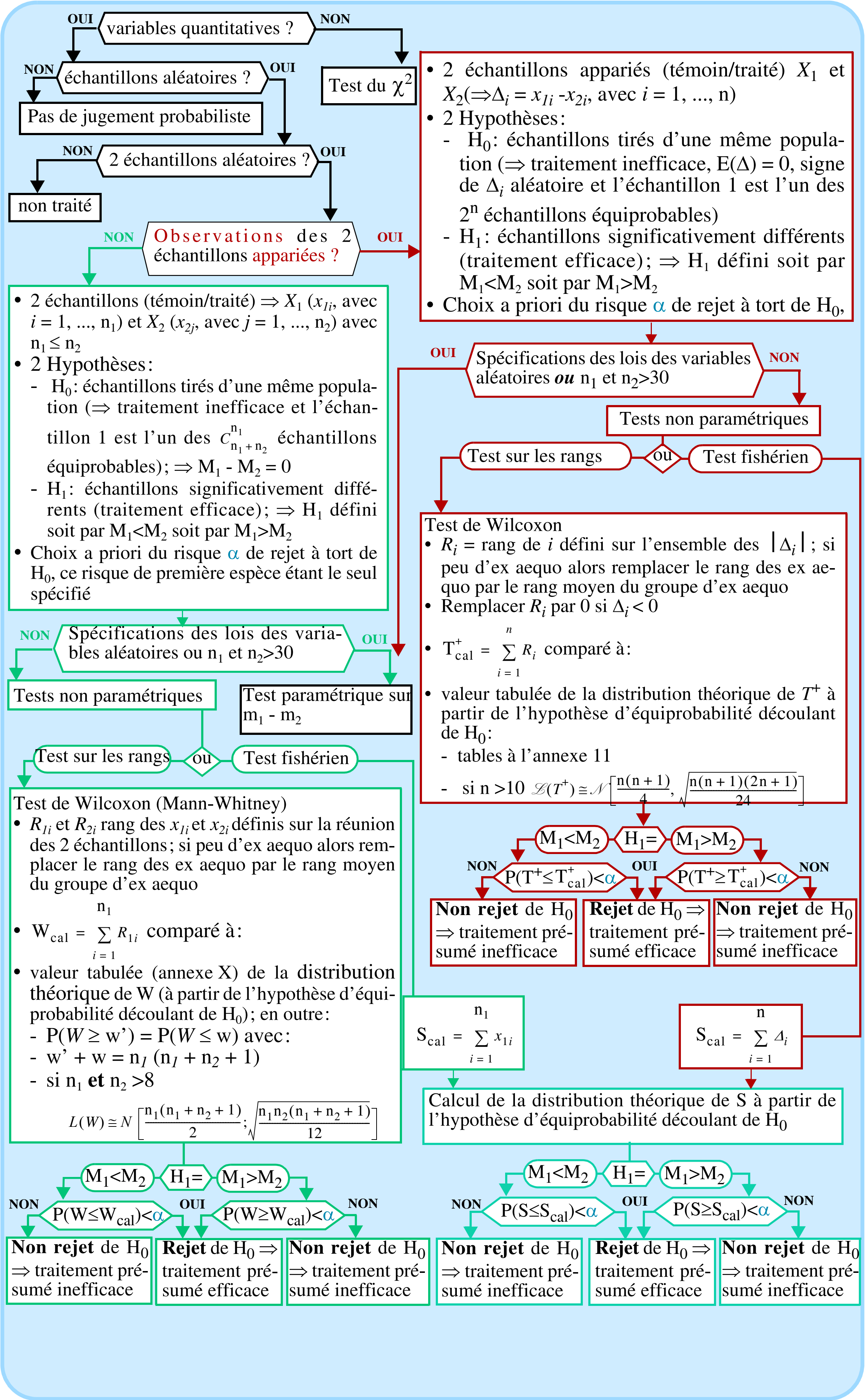
TESTS PARAMÉTRIQUES



Source de variation	Somme des carrés	Degrés de liberté	Estimation de σ^2	***F***
Facteur contrôlé (inter-classes)	$S_A^2 = \sum_{j=1}^{k} n_j(\overline{x}_{.j} - \overline{x}_{..})^2$	k-1	$\hat{\sigma}_A^2 = \frac{S_A^2}{k-1}$	$F = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_r^2}$
Résiduelle (intra-classes)	$S_r^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \overline{x}_{.j})^2$	n-k	$\hat{\sigma}_r^2 = \frac{S_r^2}{n-k}$	
Totale	$S_T^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \overline{x}_{..})^2$	n-1	$\hat{\sigma}_T^2 = \frac{S_T^2}{n-k}$	-

RÉSUMÉ DU CHAPITRE VII - COMPARAISON D'ÉCHANTILLONS

TESTS NON PARAMÉTRIQUES



- 2 échantillons appariés (témoin/traité) X_1 et X_2(⇒$\Delta_i = x_{1i} - x_{2i}$, avec i = 1, ..., n)
- 2 Hypothèses :
  - H_0 : échantillons tirés d'une même population (⇒ traitement inefficace, E(Δ) = 0, signe de Δ_i aléatoire et l'échantillon 1 est l'un des 2^n échantillons équiprobables)
  - H_1 : échantillons significativement différents (traitement efficace) ; ⇒ H_1 défini soit par M_1<M_2 soit par M_1>M_2
- Choix a priori du risque α de rejet à tort de H_0,

- 2 échantillons (témoin/traité) ⇒ X_1 (x_{1i}, avec i = 1, ..., n_1) et X_2 (x_{2j}, avec j = 1, ..., n_2) avec $n_1 \leq n_2$
- 2 Hypothèses :
  - H_0 : échantillons tirés d'une même population (⇒ traitement inefficace et l'échantillon 1 est l'un des $C^{n_1}_{n_1+n_2}$ échantillons équiprobables) ; ⇒ $M_1 - M_2 = 0$
  - H_1 : échantillons significativement différents (traitement efficace) ; ⇒ H_1 défini soit par M_1<M_2 soit par M_1>M_2
- Choix a priori du risque α de rejet à tort de H_0, ce risque de première espèce étant le seul spécifié

Test de Wilcoxon
- R_i = rang de i défini sur l'ensemble des $|\Delta_i|$; si peu d'ex aequo alors remplacer le rang des ex aequo par le rang moyen du groupe d'ex aequo
- Remplacer R_i par 0 si $\Delta_i < 0$
- $T^+_{cal} = \sum_{i=1}^{n} R_i$ comparé à :
- valeur tabulée de la distribution théorique de T^+ à partir de l'hypothèse d'équiprobabilité découlant de H_0 :
  - tables à l'annexe 11
  - si n >10 $\mathscr{L}(T^+) \cong \mathscr{N}\left[\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right]$

Non rejet de H_0 ⇒ traitement présumé inefficace

Rejet de H_0 ⇒ traitement présumé efficace

Non rejet de H_0 ⇒ traitement présumé inefficace

Test de Wilcoxon (Mann-Whitney)
- R_{1i} et R_{2i} rang des x_{1i} et x_{2i} définis sur la réunion des 2 échantillons ; si peu d'ex aequo alors remplacer le rang des ex aequo par le rang moyen du groupe d'ex aequo
- $W_{cal} = \sum_{i=1}^{n_1} R_{1i}$ comparé à :
- valeur tabulée (annexe X) de la distribution théorique de W (à partir de l'hypothèse d'équiprobabilité découlant de H_0) ; en outre :
  - P($W \geq$ w') = P($W \leq$ w) avec :
  - w' + w = n_1 ($n_1 + n_2 + 1$)
  - si n_1 **et** n_2 >8

$$L(W) \cong N\left[\frac{n_1(n_1+n_2+1)}{2} ; \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right]$$

$$S_{cal} = \sum_{i=1}^{n_1} x_{1i}$$

$$S_{cal} = \sum_{i=1}^{n} \Delta_i$$

Calcul de la distribution théorique de S à partir de l'hypothèse d'équiprobabilité découlant de H_0

Non rejet de H_0 ⇒ traitement présumé inefficace

Rejet de H_0 ⇒ traitement présumé efficace

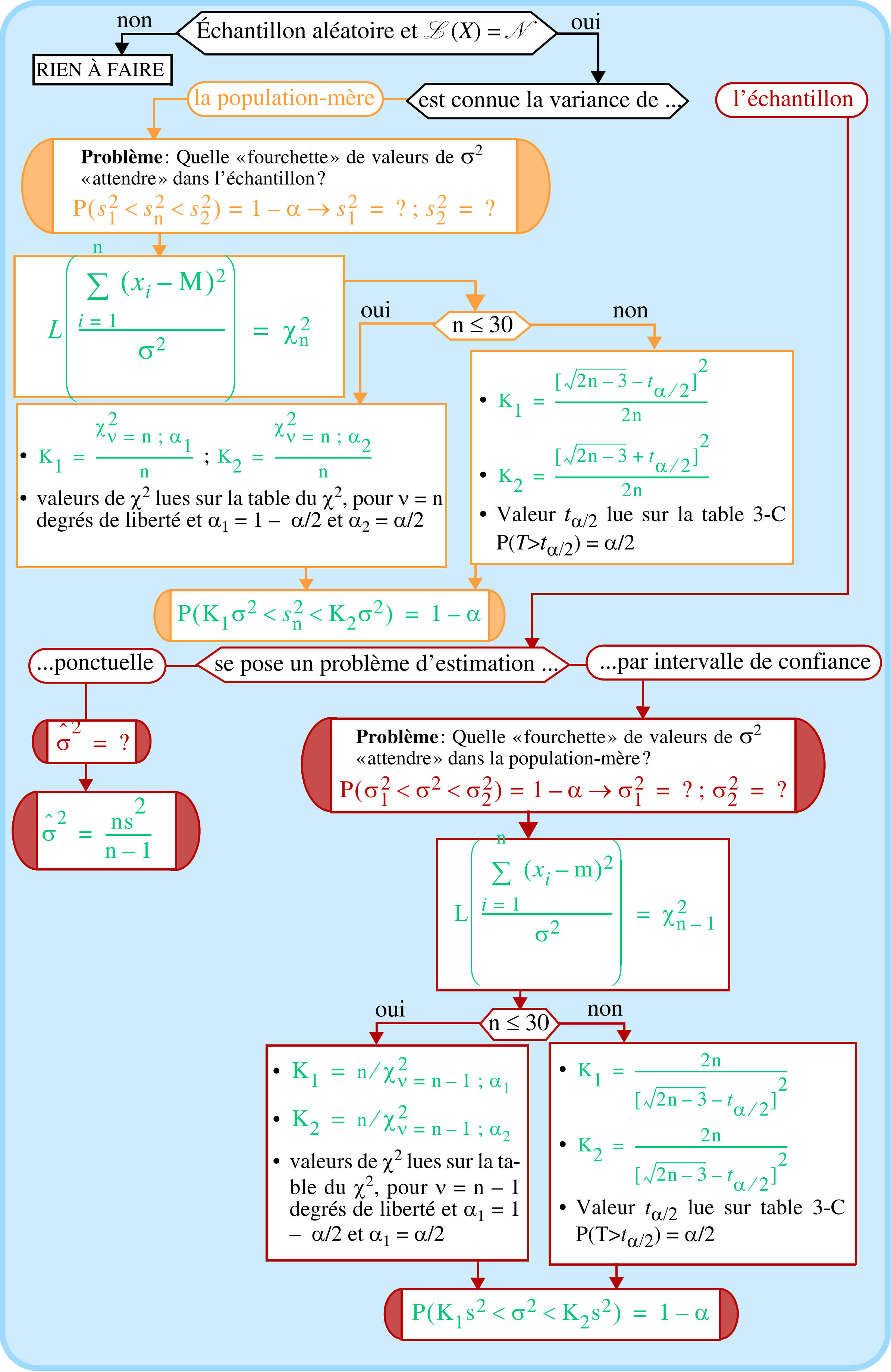
Non rejet de H_0 ⇒ traitement présumé inefficace

Non rejet de H_0 ⇒ traitement présumé inefficace

Rejet de H_0 ⇒ traitement présumé efficace

Non rejet de H_0 ⇒ traitement présumé inefficace

INDUCTION STATISTIQUE - CAS DE LA VARIANCE (CHAPITRE IV)



Échantillon aléatoire et $\mathscr{L}(X) = \mathscr{N}$

non : RIEN À FAIRE

oui : est connue la variance de ...

la population-mère

Problème: Quelle «fourchette» de valeurs de σ^2 «attendre» dans l'échantillon ?

$$P(s_1^2 < s_n^2 < s_2^2) = 1 - \alpha \to s_1^2 = ?\ ;\ s_2^2 = ?$$

$$L\left(\frac{\sum_{i=1}^{n}(x_i - M)^2}{\sigma^2}\right) = \chi^2_n$$

$n \leq 30$

oui :

- $K_1 = \frac{\chi^2_{\nu = n\,;\,\alpha_1}}{n}$; $K_2 = \frac{\chi^2_{\nu = n\,;\,\alpha_2}}{n}$
- valeurs de χ^2 lues sur la table du χ^2, pour $\nu = n$ degrés de liberté et $\alpha_1 = 1 - \alpha/2$ et $\alpha_2 = \alpha/2$

non :

- $K_1 = \frac{[\sqrt{2n-3} - t_{\alpha/2}]^2}{2n}$
- $K_2 = \frac{[\sqrt{2n-3} + t_{\alpha/2}]^2}{2n}$
- Valeur $t_{\alpha/2}$ lue sur la table 3-C $P(T > t_{\alpha/2}) = \alpha/2$

$$P(K_1\sigma^2 < s_n^2 < K_2\sigma^2) = 1 - \alpha$$

l'échantillon

se pose un problème d'estimation ...

...ponctuelle

$\hat{\sigma}^2 = ?$

$$\hat{\sigma}^2 = \frac{ns^2}{n-1}$$

...par intervalle de confiance

Problème: Quelle «fourchette» de valeurs de σ^2 «attendre» dans la population-mère ?

$$P(\sigma_1^2 < \sigma^2 < \sigma_2^2) = 1 - \alpha \to \sigma_1^2 = ?\ ;\ \sigma_2^2 = ?$$

$$L\left(\frac{\sum_{i=1}^{n}(x_i - m)^2}{\sigma^2}\right) = \chi^2_{n-1}$$

$n \leq 30$

oui :

- $K_1 = n/\chi^2_{\nu = n-1\,;\,\alpha_1}$
- $K_2 = n/\chi^2_{\nu = n-1\,;\,\alpha_2}$
- valeurs de χ^2 lues sur la table du χ^2, pour $\nu = n - 1$ degrés de liberté et $\alpha_1 = 1 - \alpha/2$ et $\alpha_1 = \alpha/2$

non :

- $K_1 = \frac{2n}{[\sqrt{2n-3} - t_{\alpha/2}]^2}$
- $K_2 = \frac{2n}{[\sqrt{2n-3} - t_{\alpha/2}]^2}$
- Valeur $t_{\alpha/2}$ lue sur table 3-C $P(T > t_{\alpha/2}) = \alpha/2$

$$P(K_1 s^2 < \sigma^2 < K_2 s^2) = 1 - \alpha$$