Cours 4:
Statistique inférentielle

Échantillonnage

A- Statistique inférentielle et échantillon

B- Théorie de l'échantillonnage

C- Distributions d'échantillonnage

D- Simulation d'échantillons

A- Statistique inférentielle et échantillon

A- 1 Introduction

Etude Statistique = étude des caractéristiques (variables statistiques) d'un ensemble d'objets (population, composée d'individus) .

- Recensement : les valeurs des variables sont disponibles sur l'ensemble de la population $\Rightarrow$ statistique descriptive (pas besoin de stat inférentielle)

  Ex : Recensement de la population française, notes obtenues par tous les candidats à un examen, salaires de tous les employés d'une entreprise, ...

  Pbme : coûteux, long, impossible (population infinie), mesures destructrices (ex : tests en vieillissement accélérés)

- **Sondage :**
  - On n'étudie qu'une partie de la population : un échantillon. Les méthodes permettant de réaliser un échantillon de bonne qualité (sui ressemble à la population dont il est issu) sont étudiées en théorie de l'échantillonnage.
  - On cherche alors à extrapoler à la population entière les propriétés mises en évidence sur l'échantillon $\Rightarrow$ statistique inférentielle

A-2 Les hypothèses de la statistique inférentielle

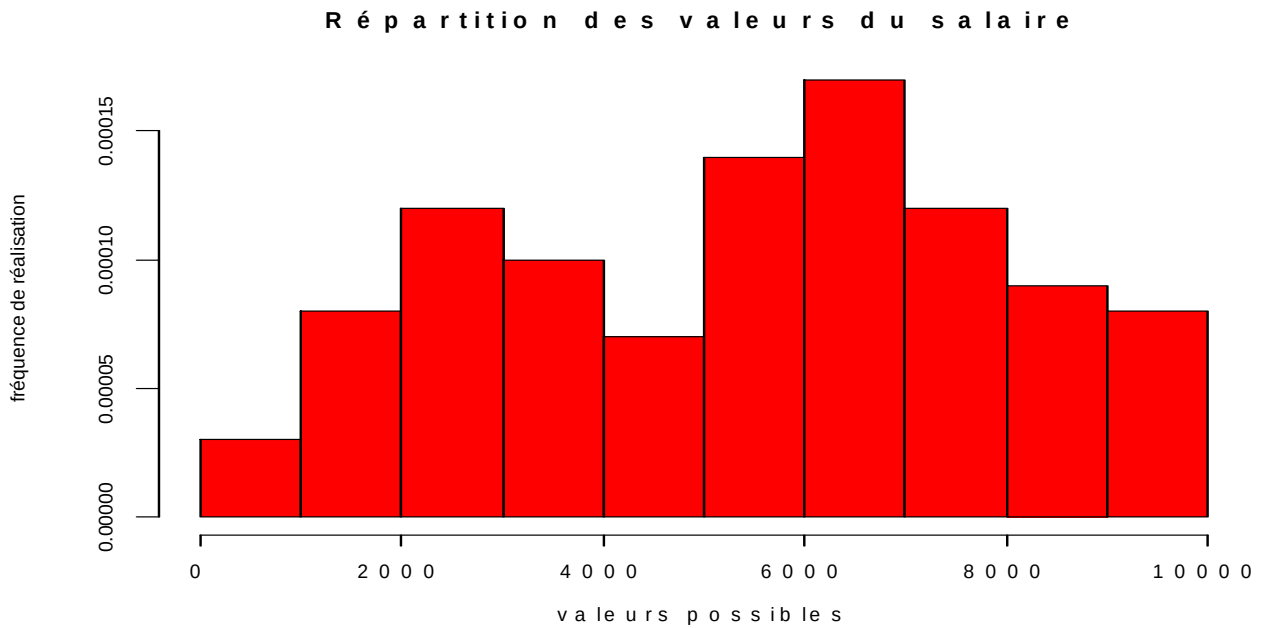
- ✓ **La population est considérée comme infinie (très grande)**

- ✓ **les variables statistiques qui la décrivent peuvent être considérées comme des v.a.**

  La valeur prise par la variable statistique X pour un individu donné de la population ne peut pas être déterminée a priori et dépend d'un grand nombre de paramètres : On peut considérer sa valeur comme fonction du résultat d'une expérience aléatoire.

A-2 Les hypothèses de la statistique inférentielle

Ex : répartition des salaires des salariés dans la population francaise : série (x1,...xn), vue comme n réalisations de la variable aléatoire X=salaire



A-2 Les hypothèses de la statistique inférentielle

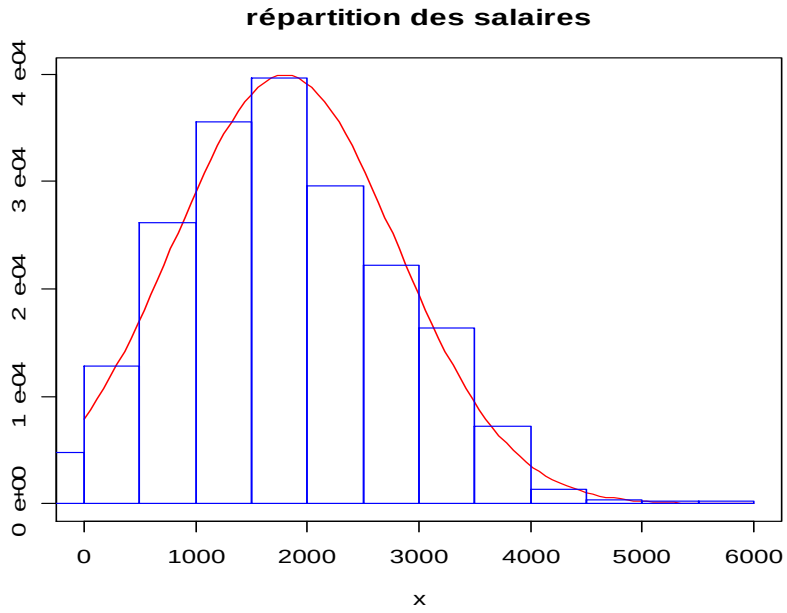
- ✓ **La répartition des valeurs de ces variables sont caractérisées par des lois de probabilités**

  La répartition d'une variable statistique X sur la population est décrite par une loi de probabilité,

- ➢ caractérisée par une densité de probabilité (X continue)ou une séquence de fréquences relatives à chacune de ses valeurs (X discrète)

- ➢ possédant des caractéristiques (E(X), V(X), autres paramètres résumant la distribution.)

Ex: si l'on suppose que les salaires sont soumis à un grand nombre de petites fluctuations d'origines diverses, X suit une loi normale tronquée à zero.



A-2 Les hypothèses de la statistique inférentielle

Les variations simultanées de deux ou plusieurs variables statistiques sont décrites par une loi jointe

- caractérisée par une densité jointe (variables continues) ou une séquence des fréquences jointes (variables discrètes).
  - Ex : les variations simultanées du salaire et de l'age des salariés pourront être décrites par une fonction de densité jointe f(x,z).

- possédant différentes caractéristiques, par exemple un vecteur espérance, une matrice de variance covariance , un coefficient de corrélation linéaire.

A-2 Les hypothèses de la statistique inférentielle

- ✓ **Ces lois de probabilités sont généralement**

  - **Totalement Inconnues :** nous ne connaissons rien de la loi - problème de statistique inférentielle non-paramétrique

  - **Partiellement inconnues** : nous connaissons la famille à laquelle appartient la loi (sa forme) mais pas ses ou un certain nombre de ses paramètres (Ex : X obéit à une loi normale, mais on ne connaît ni son espérance ni sa variance – problème de statistique inférentielle paramétrique.

A-3 Les objectifs de la statistique inférentielle

L'objectif de la statistique inférentielle est d'identifier ces lois, au vu d'un échantillon de valeurs des variables obtenu par sondage dans la population, grâce à différent types de méthodes :

- Méthodes d'estimation : permettent d'approcher les lois ou certaines de leurs caractéristiques (ex : approcher , à partir de l'échantillon l'espérance E(Y) de la variable Y=salaire,….)
- Méthodes de tests d'hypothèses : permettent de confirmer ou d'infirmer des hypothèses faites sur ces lois (ex : décider si, au vu de l'échantillon, l'affirmation « E(Y)=1500 euros » est plausible.)
- Méthodes de modélisation et prévision : permettent d'expliquer et de prévoir la loi d'une variable a partir de s valeurs prises par d'autres (ex: au vu de l'échantillon, les variations de salaires sont expliquées presque exclusivement par l'age X des salariées : $Y=f(X)+\varepsilon$) .

La pertinence de ces méthodes repose en premier lieu sur la qualité du sondage effectué $\Rightarrow$ théorie de l'échantillonnage.

B- Théorie de l’échantillonnage

B-1 Introduction

- ✓ **Théorie de l'échantillonnage** = Etude des liaisons existant entre une population et les échantillons de cette population, prélevés par sondage.

  - ➢ Méthodes d'échantillonnage : ensemble des méthodes permettant de réaliser un sondage (de prélever un échantillon de données) au sein d'une population, de manière à reproduire un échantillon aussi représentatif que possible de cette population.

  - ➢ Evaluation de ces méthodes : le système d'échantillonnage sera jugé d'après la qualité des approximations des paramètres de la population, calculées sur l'échantillon prélevé . Pour cela, on étudiera la loi des caractéristiques classiques d'un échantillon (moyenne arithmétique , variance empirique,…)

B-2 Les méthodes d'échantillonnage

- ✓ **Les méthodes empiriques** : les plus utilisées par les instituts de sondage. Leur précision ne peut pas être calculée et leur réussite dépend de l'expertise des enquêteurs.

  - *Echantillonnage sur la base du jugement* : Echantillon prélevé à partir d'avis d'experts, qui connaissent bien la population et sont capable de dire quelles sont les entités représentatives.

    Pbme: l'avis des experts est subjectif.

  - *Echantillonnage par la méthode des quotas :* Echantillon prélevé librement à condition de respecter une composition donnée à l'avance (sexe, âge, CSP,...).

    Pbme : repose sur la pertinence des catégories retenues.

B-2 Les méthodes d'échantillonnage

- ✓ **Les méthodes aléatoires** : Reposent sur le tirage au hasard d'échantillons et sur le calcul des probabilités.

  - *Echantillonnage aléatoire simple* : On prélève dans la population, des individus au hasard, sans remise : tous les individus ont la même probabilité d'être prélevés, et ils le sont indépendamment les uns des autres.

  - *Echantillonnage aléatoire stratifié* : Suppose que la population soit stratifiée, i.e. constituée de sous-populations homogènes, les strates. (ex : stratification par tranche d'age). Dans chaque strate, on fait un échantillonnage aléatoire simple, de taille proportionnelle à la taille de strate dans la population (échantillon représentatif). Les individus de la population n'ont pas tous la même probabilité d'être tirés. Nécessite une homogénéité des strates. *Augmente la précision des estimations*.

  - *Echantillonnage par grappe* : on tire au hasard des grappes ou familles d'individus, et on examine tous les individus de la grappe (ex: on tire des immeubles puis on interroge tous les habitants). La méthode est d'autant meilleure que les grappes se ressemblent et que les individus d'une même grappe sont différents, contrairement aux strates.

B-2 Les méthodes d'échantillonnage

Dans toute la suite du cours, on se place dans le cadre d'un échantillonnage aléatoire simple, sauf mention contraire.

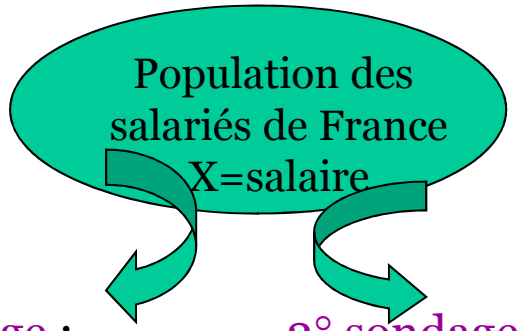
B-3 Notion d'échantillon aléatoire

- ✓ **Quelle que soit la technique d'échantillonnage utilisée, le contenu du jeu de données prélevé varie d'un sondage à l'autre**

On pourrait répéter le sondage un grand nombre de fois, on obtiendrait la plupart du temps une répartition différente des valeurs prélevées.

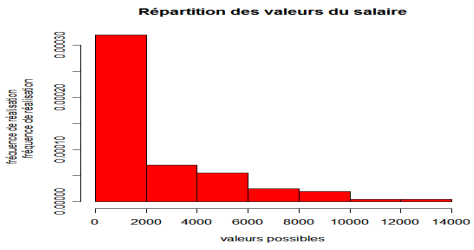
Le résultat d'un sondage est aléatoire

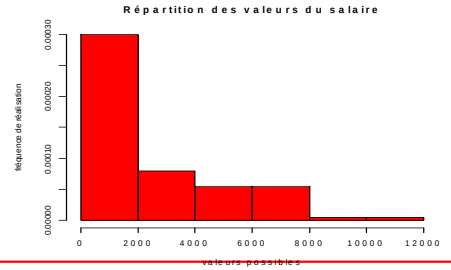
Sondage de 100 salariés



1° sondage : $(x_1, \ldots . x_{100})$

2° sondage: $(x'_1, \ldots . x'_{100})$





$\overline{x} = 2050.7 \quad s_X = 2959.1$

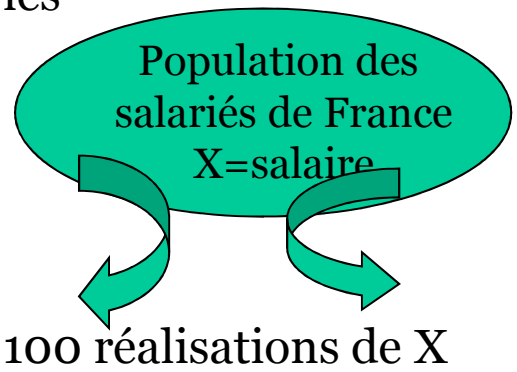
$\overline{x} = 2153.8 \quad s_X = 3002.2$

B-3 Notion d'échantillon aléatoire

- ✓ **Deux façons différentes de modéliser cet aléa**

- ➢ 1° Modélisation :
  L'échantillon prélevé consiste en n réalisations X(ω1) , …, X(ω) de la v.a. X.

- Sondage aléatoire simple de 100 salariés



1° sondage : $(x_1, \ldots . x_{100})$

$= (X(\omega_1), \ldots . X(\omega_{100}))$

2° sondage : $(x'_1, \ldots . x'_{100})$

$= (X(\omega'_1), \ldots . X(\omega'_{100}))$

B-3 Notion d'échantillon aléatoire

- 2° Modélisation : On associe au premier individu tiré une variable aléatoire X_1 de même loi que X

  Elle vaut, $x_1, x'_1, x''_1 ..$ selon le sondage. On fait de même pour les n-1 autres individus.

  L'objet $(X_1,...,X_n)$, où Xi est la valeur de X pour le i° individu tiré, est un vecteur de v.a. i.i.d. de même loi que X. Un tirage correspond à une seule réalisation de celui-ci.:

  $$(x_1,...,x_n)=(X_1(\omega),...,X_n(\omega))$$

  $(X_1,...,X_n)$ **est appelé l'échantillon aléatoire**.

- Sondage aléatoire simple de 100 salariés

1° sondage : 2° sondage :

X1(w) X2(w) X100(w) X1(w') X2(w')X100(w')

x1	x2	...	x100

x'1	x'2	...	x'100

B-4 Etude des statistiques classiques

- ✓ Objectif : étudier la loi des statistiques classiques de l'échantillon aléatoire (les distributions d'échantillonnage), en fonction de la distribution de la variable parente, lorsque la taille de l'échantillon augmente.

- ✓ Définition d'une statistique = variable aléatoire, définie comme une fonction de l'échantillon aléatoire

$$S = f(X_1, \ldots, X_n)$$

Lorsque $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ la réalisation de S vaut $s = f(s_1, \ldots, s_n)$

- ✓ Exemples de statistiques : Moyenne empirique de l'échantillon, variance empirique, covariance empirique, fonction de répartition,.....

B-4 Etude des statistiques classiques

✓ Rq: En statistique inférentielle, les indicateurs usuels de la statistique descriptive deviennent des statistiques de l'échantillon aléatoire

- Moyenne « empirique » : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \rightarrow \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Variance « empirique » : $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \rightarrow S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$
- Moments « empiriques »

C- Distributions d’échantillonnage

C-1 Notations

- ✓ On s'intéresse à la caractéristique X d'une population (X=v.a.). On pose $E(X) = m, V(X) = \sigma^2$

- ✓ On note $(X_1, \ldots, X_n)$ l'échantillon aléatoire associé à un sondage aléatoire simple de n individus de cette population et $(x_1, \ldots, x_n)$ une réalisation de celui ci (1 sondage particulier)

Empirique veut dire « de l'échantillon »

C-1 Moyenne empirique

- ✓ Définition :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- ✓ Loi et moments :

Loi inconnue en général

$$i)\ E(\bar{X}_n) = m, \quad ii)\ V(\bar{X}_n) = \frac{\sigma^2}{n}$$

- ✓ Propriétés asymptotiques :

  - ➢ Loi des grands nombres

$$iii)\ \bar{X}_n \xrightarrow{P} m\ ;\ \bar{X}_n \xrightarrow{p.s.} m$$

  - ➢ Théorème central limite (TCL)

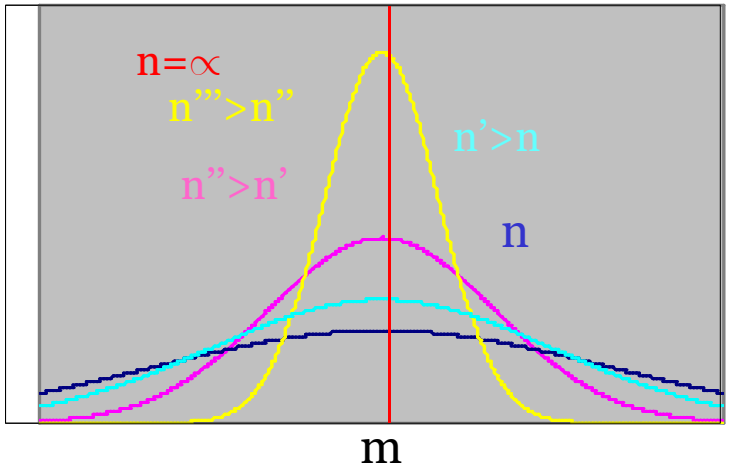
$$iv)\ \sqrt{n}\,\frac{\bar{X}_n - m}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$\sqrt{n}\left(\bar{X}_n - m\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma)$$

Info : $\bar{X}_n$ « approche» m : c'est un estimateur de m. Il est :

- sans biais (i))
- asymptotiquement efficace (ii))
- fortement convergent (iii))
- la loi de l'erreur d'approximation est approximativement gaussienne lorsque n est grand (iv).

distribution de la moyenne



C-1 Moyenne empirique

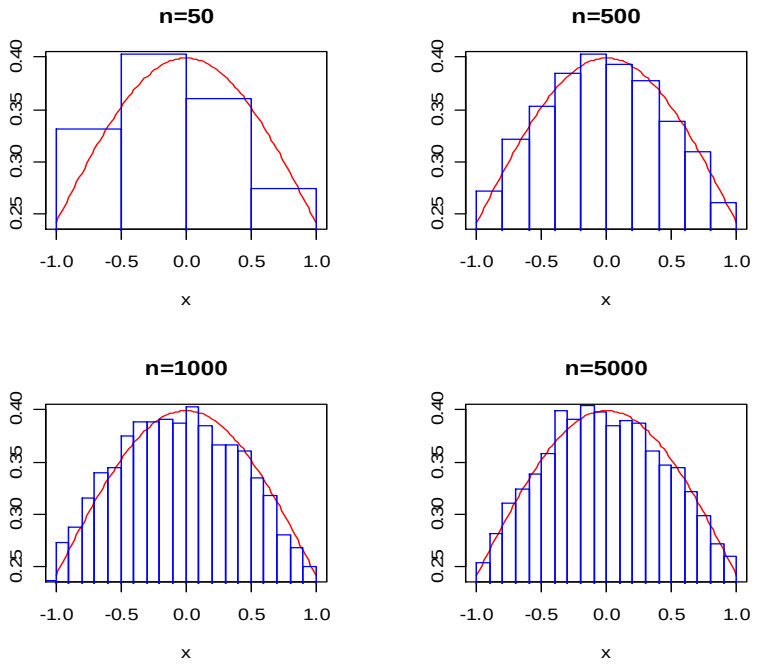
Interprétation statistique :

i) et ii) ≈ si l'on prélève un grand nombre k d'échantillons de taille n, et que l'on calcule leur moyenne, la moyenne des k valeurs moyennes ainsi obtenues vaut à peu près m, et la variance de ces k valeurs est d'autant plus faible que n est grand.

iii) ≈ lorsque la taille de l'échantillon prélevé est très grande, les k moyennes valent presque toutes m.

iv) ≈ si l'on prélève un grand nombre k d'échantillons de grande taille n et que l'on calcule leurs moyennes renormalisées, l'histogramme des k valeurs est proche de la densité de la loi normale centrée réduite.

TCL : histogramme de la série normalisée des moyennes de 10000 échantillons de taille 50, 500, 1000, 5000 de E(1)



C-1 Moyenne empirique

✓ Application : loi d'un pourcentage

On tire dans une urne de Bernouilli composée d'une proportion p de boules rouges n boules avec remise. On note X le nombre aléatoire de boules rouges Q la fréquence empirique :

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

$$E(Q) = p; \quad V(Q) = \frac{p(1-p)}{n}$$

Lorsque n est grand, $Q \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

C-2 Variance empirique

✓ Définitions :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Rq : ce n'est pas une somme de va indépendantes

Autres expressions :

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2$$

$$(*)\ S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - (\bar{X} - m)^2$$

✓ Loi et Moments : La loi est généralement inconnue.

$$E(S_n^2) = \frac{n-1}{n}\sigma^2, \quad V(S_n^2) = \frac{n-1}{n^3}\left((n-1)\mu^4 - (n-3)\sigma^4\right)$$

$$V(S_n^2) \sim \frac{\mu^4 - \sigma^4}{n}$$

Outils de démonstration pour la variance : on utilise (*) et $Cov(X_i, \bar{X}_n) = \frac{\sigma^2}{n}$

C-2 Variance empirique

✓ Lien entre $\bar{X}_n$ et S_n^2 : ils sont asymptotiquement non corrélés :

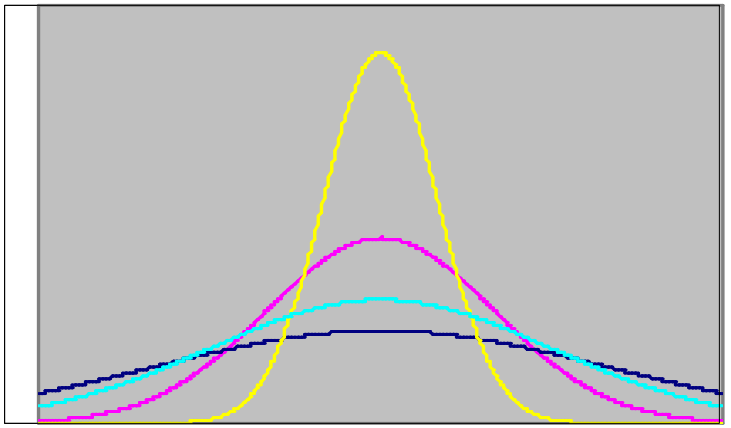
$$Cov(\bar{X}_n, S_n^2) = \frac{\mu^3}{n}\left(1 - \frac{1}{n}\right)$$

✓ Propriétés asymptotiques :

$$S_n^2 \xrightarrow{P} \sigma^2 \; ; \; S_n^2 \xrightarrow{p.s.} \sigma^2$$

Dém : On utilise la condition suffisante de Convergence en probabilité.

$$\sqrt{n}\frac{S_n^2 - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$



Dém : $S_n^2 = T_n - (\bar{X} - m)^2$ avec $\sqrt{n}\frac{T_n - \sigma^2}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{L} N(0,1)$ et $(\bar{X} - m)^2 \xrightarrow{P} 0$

C-3 Moyenne et la variance empirique : Cas gaussien

Si X suit une loi $\mathcal{N}(m, \sigma)$. Alors :

$$Cov(S_n^2, \bar{X}_n) = 0 : S_n^2 \perp \bar{X}_n$$

$$\bar{X}_n \sim \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

$$\frac{nS_n^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\sqrt{n-1}\left(\frac{\bar{X}_n - m}{\sqrt{S_n^2}}\right) \sim \mathcal{T}(n-1)$$

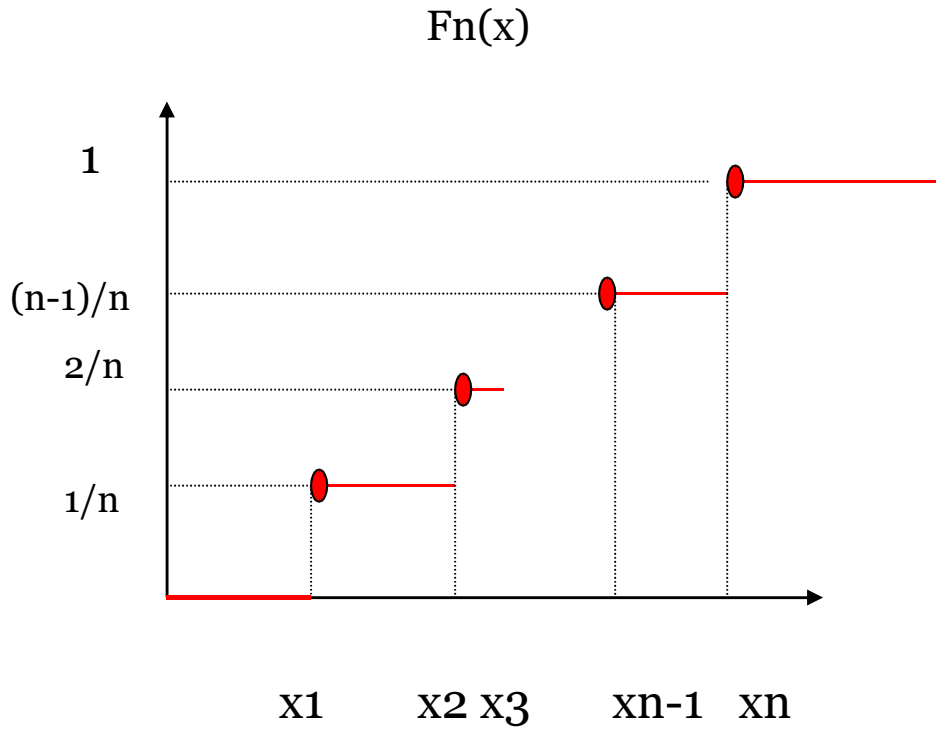
RQ : une combinaison linéaire de v.a. gaussiennes indépendantes est gaussienne.

C-4 Fonction de répartition empirique

✓ Définition :

$$F_n(x) = \frac{\text{nombre de } X_i \le x}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \le x}$$

- Pour chaque valeur $x \in \mathbb{R}$, $F_n(x)$ est une variable aléatoire
- Pour une réalisation $(x_1,\ldots,x_n)$ donnée de l'échantillon aléatoire, c'est une fonction en escalier à valeurs dans [0,1], croissante, continue à droite dans [0,1], de sauts égaux à 1/n.



C-4 Fonction de répartition empirique

✓ Loi et moments à x fixé

$$i)\ nF_n(x) \sim B(n, F(x))$$
$$ii)\ E(F_n(x)) = F(x)$$
$$iii)\ V(F_n(x)) = \frac{F(x)(1-F(x))}{n}$$

✓ Propriétés asymptotiques

➢ Lois des grands nombres :

$$iv)\ F_n(x) \xrightarrow{P} F(x)\ ;\ F_n(x) \xrightarrow{p.s.} F(x)$$

➢ Théorème central limite (TCL)

$$v)\ \sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, F(x)(1-F(x))\right)$$

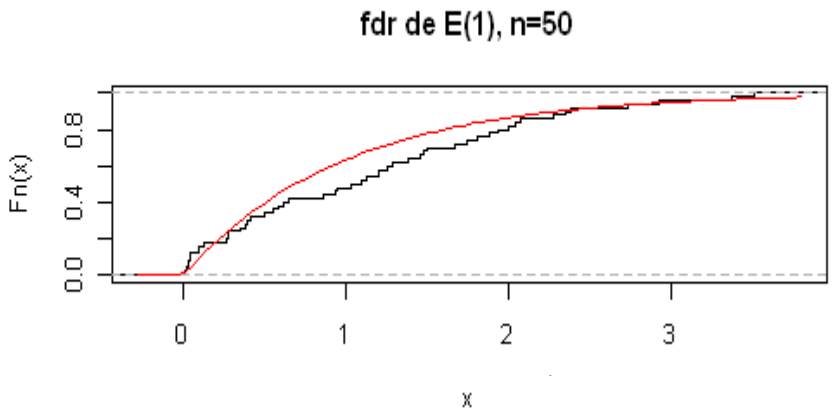
Info : Pour tout $x \in \mathbb{R}$, $F_n(x)$ « approche» $F(x)$: c'est un estimateur de $F(x)$. Il est :

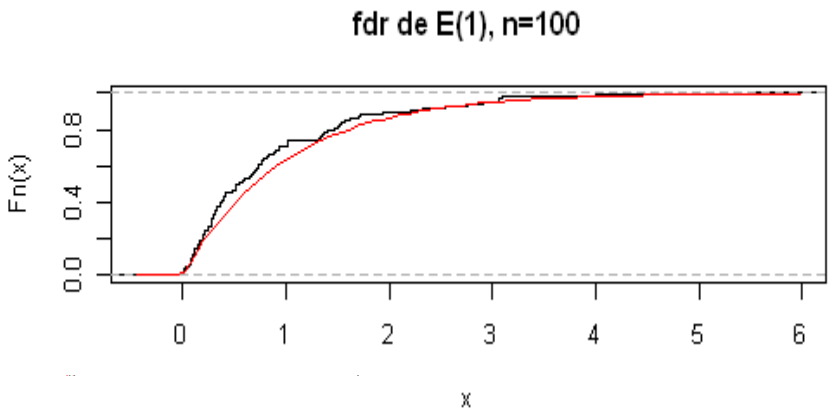
- sans biais (ii))
- asymptotiquement efficace (iii))
- fortement convergent (iv))

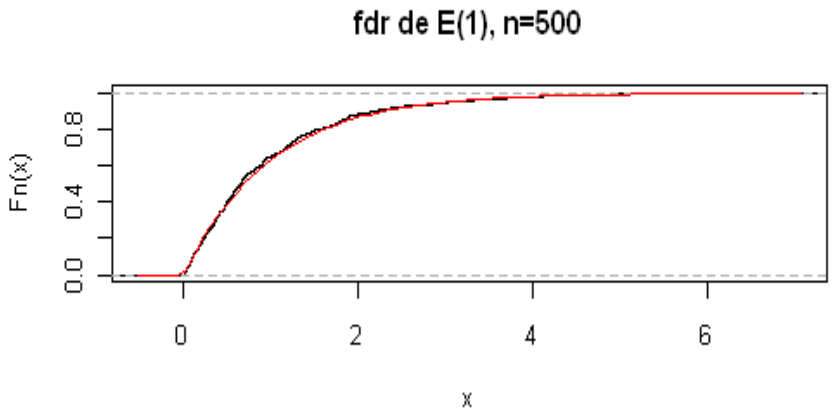
Outils de dém: $Y_i = 1_{X_i \le x} \sim B(F(x))$

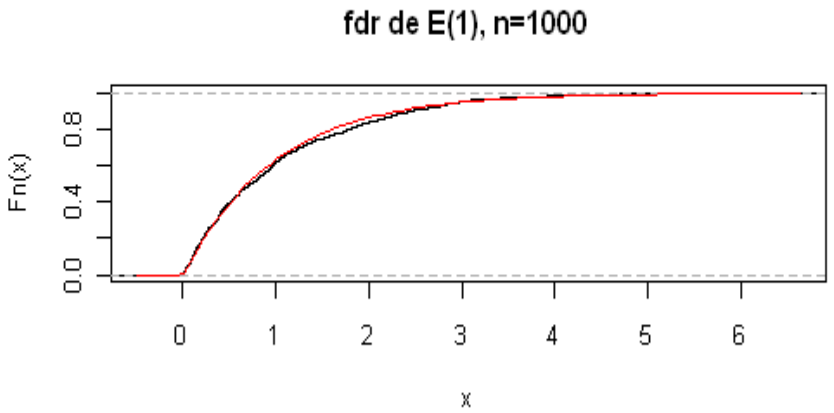
donc $$nF_n(x) = \sum_{i=1}^{n} Y_i \sim \mathcal{B}(n, F(x))$$

C-4 Fonction de répartition empirique









D-1 Simulations d'un échantillon: Cas général

✓ Théorème d'inversion

Soit F une fonction de répartition sur $\mathbb{R}$. On note $F^{-1}(y) = \inf\{x \in \mathbb{R} / F(x) \geq y\}$ l'inverse généralisé de F (vaut l'inverse habituelle lorsque F est continue et strictement croissante). Soit U deloi uniforme sur [0,1]. Alors,

1. $X = F^{-1}(U)$ a pour fonction de répartition F

2. Si F est continue sur $\mathbb{R}$ et X de fdr F, U=F(X) suit une loi uniforme sur [0,1].

D-2 Simulations d'un échantillon : cas continu

✓ Simulation d'une loi continue

Simulation de n réalisations X de loi F:

- on simule n réalisations d'une loi uniforme sur [0,1] (tirage au hasard de n nombres sur cet intervalle) : u1,…,un

- On calcule $\forall i = 1,....,n, \quad x_i = F^{-1}(u_i)$. Ce sont n réalisations de X de loi F.

D-2 Simulation d'un échantillon : cas discret

✓ Simulation d'une loi discrète

Soit $\left(p_i = P(X = x_i)\right)_{1 \le i \le n}$ la loi de probabilité discrète d'une variable aléatoire à valeurs dans $\{x_1, \ldots, x_n\}$. On note $s_k = P(X \le x_k) = \sum_{i=1}^{k} p_i$ et $F(x) = \sum_{k=1}^{n} s_{k-1} 1_{x_{k-1} \le x < x_k}$ la fonction de répartition de cette loi en tout point.
Soient $u_1, \ldots, u_n$ n réalisations d'une variable de loi uniforme sur [0,1].

Alors $\forall i = 1, \ldots, n,\ x_k^* = F^{-1}(u_i) = \sum_{k=1}^{n} x_k 1_{s_{k-1} < u_i \le s_k}$

Sont n réalisations d'une variable aléatoire discrète de loi F.