RFIDEC — cours 2 : Échantillons, estimations ponctuelles

Christophe Gonzales

LIP6 – Université Paris 6, France

Plan du cours n˚2

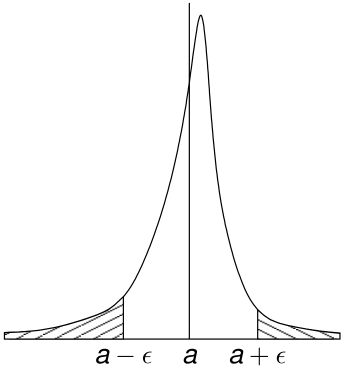
1. Lois des grands nombres
2. Théorème central-limite
3. Estimation ponctuelle à partir d'échantillons
4. Biais dans les estimations

Convergence en probabilité

Définition

- $(X_n)_{n \in \mathbb{N}}$: suite de variables
- a : constante
- (X_n) *converge en probabilité* vers a si, pour tout $\epsilon > 0$ la probabilité que l'écart absolu entre X_n et a dépasse ϵ tend vers 0 quand $n \to \infty$:

$$\lim_{n \to \infty} P(| X_n - a | \geq \epsilon) = 0$$



Aire hachurée tend vers 0 quand $n \to \infty$

Loi faible des grands nombres

Loi faible

- $(X_n)_{n\in\mathbb{N}}$ est une suite de variables aléatoires :
  - de même loi
  - d'espérance m
  - possédant une variance σ^2
  - deux à deux indépendantes
- alors la suite des variables $\overline{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$ converge en probabilité vers m

$\overline{X}_n$ est appelée *moyenne empirique*

$$E(\overline{X}_n) = m$$

$$V(\overline{X}_n) = \frac{\sigma^2}{n}$$

conséquence : échantillons de grandes tailles $\Longrightarrow$ bonne chance d'estimer m

Définition

- $(X_n)_{n\in\mathbb{N}}$: suite de variables
- a : constante
- (X_n) *converge presque sûrement* vers a s'il y a une proba 1 que la suite des réalisations des X_n tende vers a :

$$P\left(\lim_{n\to\infty} X_n = a\right) = 1$$

⚠ Définition plus exigeante que la convergence en probabilité

Loi forte des grands nombres

Loi forte

- $(X_n)_{n \in \mathbb{N}}$: suite de variables aléatoires
  - de même loi
  - d'espérance m
  - possédant une variance σ^2
  - mutuellement indépendantes
- alors la suite des variables $\overline{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$ converge presque sûrement vers m

Interprétation : échantillon de grande taille $\Longrightarrow$ bonne estimation de m

Convergence en loi

Définition

- $(X_n)_{n \in \mathbb{N}}$: suite de variables
- F_n : fonction de répartition de X_n
- X : variable de fonction de répartition F
- La suite X_n *converge en loi* vers X lorsque $F_n(x)$ tend vers $F(x)$ en tout point de continuité de F

Notation : $X_n \xrightarrow{loi} X$

Théorème central-limite

Théorème central-limite

- $(X_n)_{n \in \mathbb{N}}$: suite de variables
  - de même loi
  - d'espérance μ
  - de variance σ^2
  - mutuellement indépendantes
- alors la suite des moyennes empiriques centrées réduites $\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}$ tend en loi vers la loi normale centrée réduite :

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

Lancés de dés à 6 faces

$$\Longrightarrow \begin{cases} X_i & = \text{résultat du jet du } i\text{ème dé} \\ \overline{X}_n & = \text{somme des résultats des dés} \end{cases}$$

distribution de $\overline{X}_n$ pour 1 jet de dé

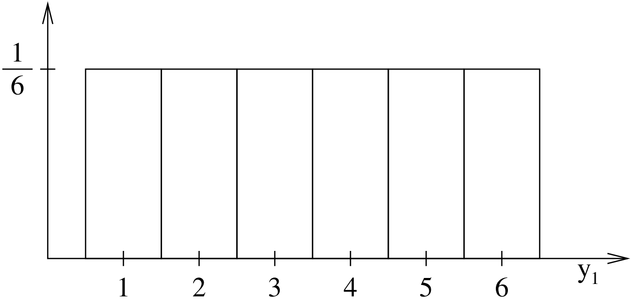


Illustration du théorème central-limite (2/4)

distribution de $\overline{X}_n$ pour 2 jets de dés

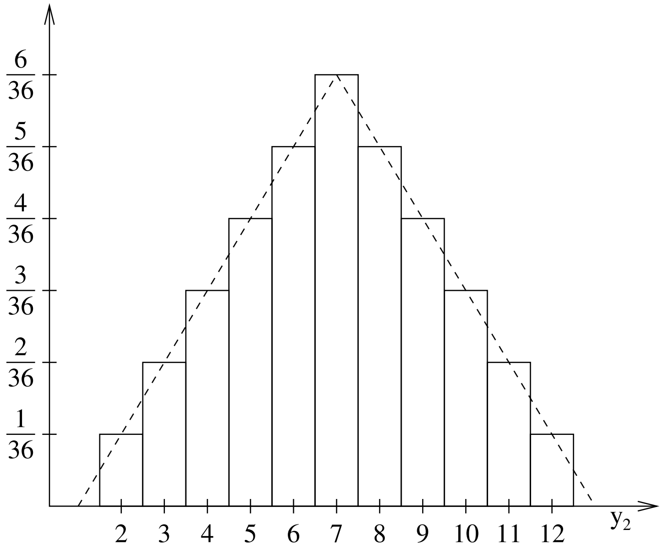
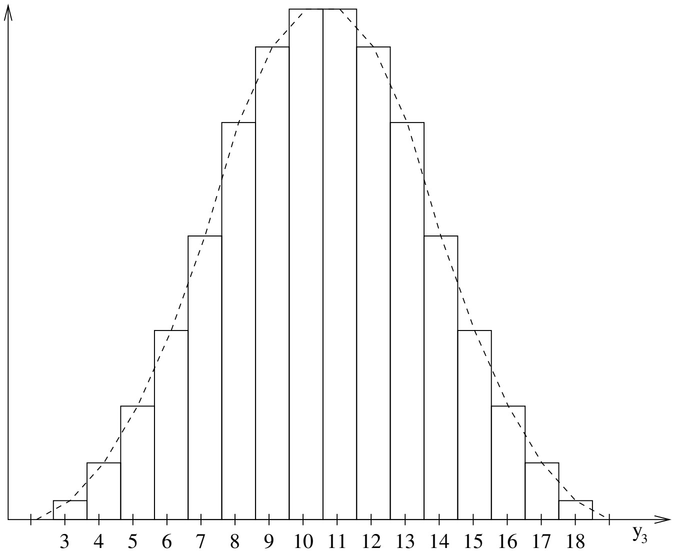
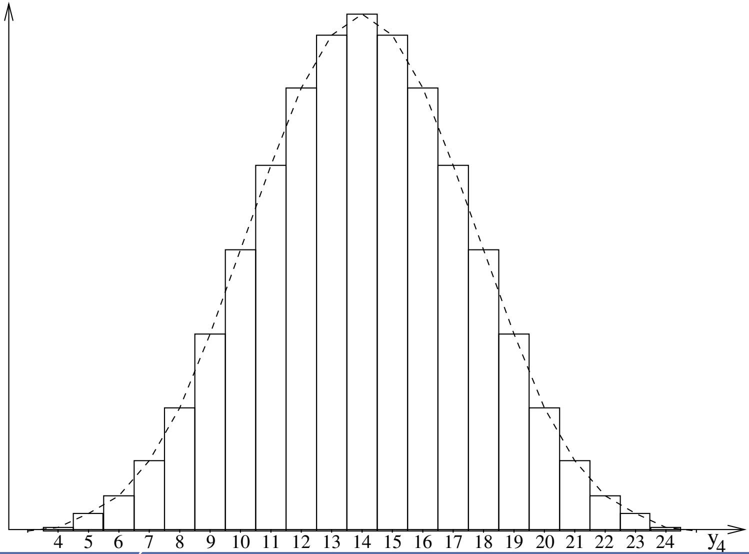


Illustration du théorème central-limite (3/4)

distribution de $\overline{X}_n$ pour 3 jets de dés



distribution de $\overline{X}_n$ pour 4 jets de dés



Statistique inférentielle

Hypothèses :

- les données = les observations
- observations $(x_1, \ldots, x_n)$ = réalisation d'une variable aléatoire multidimensionnelle $X = (X_1, \ldots, X_n)$
- observations = *échantillon empirique*
- l'échantillon est tiré suivant une loi X_0
- chaque variable X_i suit la même loi que X_0

But :

- déduire des informations sur X_0

Exemples :

- sondages électoraux, études de marché
- tests de fiabilité/qualité
- bases de données $\Longrightarrow$ diagnostic

Estimation ponctuelle

Idée force

Dans moult études statistiques :

- population de grande taille

  $\Longrightarrow$ impossible de la connaître précisément

- possibilité de prélever des échantillons

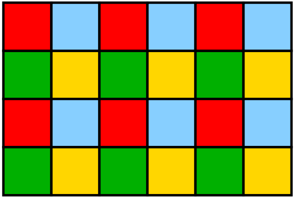
$\Longrightarrow$ étudier l'échantillon et en déduire les caractéristiques de la population

caractéristiques : moyenne μ, variance σ^2, proportion de succès p

Échantillonnage (1/2)

caractéristique de la population $\Longleftarrow$ échantillon

$\Longrightarrow$ problème : comment prélever l'échantillon ?



Définition d'un échantillon

Échantillon d'une population $=$ sous-ensemble représentatif de la population
$\Longrightarrow$ composition de l'échantillon due au hasard

Échantillonnage (2/2)

Prélèvement d'un échantillon

2 manières de prélever un échantillon aléatoirement :

1. Échantillonnage avec remise :
   - choisir un individu au hasard
   - noter la valeur de la variable d'intérêt pour celui-ci
   - remettre l'individu dans la population
   - réitérer le processus
2. Échantillonnage sans remise :
   - même procédé mais sans remettre dans la population les individus sélectionnés

⚠ échantillonnage avec remise $\Longrightarrow$ un individu peut apparaître plusieurs fois dans l'échantillon

⚠ population de grande taille $\Longrightarrow$ avec remise $\approx$ sans remise

Estimation de la moyenne d'une population (1/4)

⚠ résultats suivant valables uniquement pour des échantillons *avec remise*

$\Longrightarrow$ on va travailler avec des échantillons i.i.d :

Échantillon i.i.d

- échantillon de n individus
- X_i = variable aléatoire « valeur du $i$ème individu tiré »
- les X_i sont mutuellement indépendants
- les X_i ont tous la même distribution

$\Longrightarrow$ les X_i sont indépendants et identiquement distribués (i.i.d)

échantillons i.i.d $\Longrightarrow$ bonnes propriétés mathématiques

- X variable aléatoire sur l'ensemble de la population
- espérance : μ, variance : σ^2

- échantillon de n individus $\Longrightarrow$ observation de n valeurs de X
- X_i : variable aléatoire correspondant au $i$ème individu
- échantillon i.i.d $\Longrightarrow$ espérance de X_i : μ, variance de X_i : σ^2

- $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ = moyenne de l'échantillon

Problèmes : que valent $E(\overline{X})$ et $V(\overline{X})$?

Estimation de la moyenne d'une population (3/4)

- $$E(\overline{X}) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n)$$
  $$= \frac{1}{n}\left(E(X_1) + E(X_2) + \cdots + E(X_n)\right)$$
  $$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$$

variables X_i mutuellement indépendantes $\Longrightarrow$

- $$V(\overline{X}) = V\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}V(X_1 + X_2 + \cdots + X_n)$$
  $$= \frac{1}{n^2}\left(V(X_1) + V(X_2) + \cdots + V(X_n)\right)$$
  $$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}$$

Estimation de la moyenne d'une population (4/4)

Théorème 1

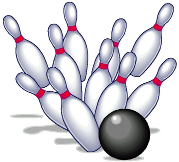
- X : variable aléatoire
- espérance de X : μ, variance de X : σ^2
- échantillon de taille n avec remise sur X
- $\overline{X}$: variable aléatoire « moyenne de l'échantillon »
- Alors : $E(\overline{X}) = \mu$ et $V(\overline{X}) = \frac{\sigma^2}{n}$

Corollaire

- X : variable aléatoire suivant une loi normale $\mathcal{N}(\mu; \sigma^2)$
- échantillon de taille n avec remise sur X
- $\overline{X}$: variable aléatoire « moyenne de l'échantillon »
- Alors : $\overline{X} \sim \mathcal{N}(\mu; \frac{\sigma^2}{n})$

Exemple : calcul du handicap au bowling

Compétition en 2 phases :



- 1ère phase : 6 parties
- calcul du handicap :
  Handicap H = (215 $-$ moyenne des 6 parties)$\times$60%
- 2ème phase : 6 parties (score $+$ le handicap H)

- Calcul du handicap $\Longrightarrow$ estimation de votre niveau
- 6 parties de la 1ère phase = échantillon de taille 6
- X_i $=$ variable aléatoire « score de la $i$ème partie »

idée : scores de l'échantillon $\Longrightarrow$ score moyen de la population

Retour sur l'estimation de la moyenne

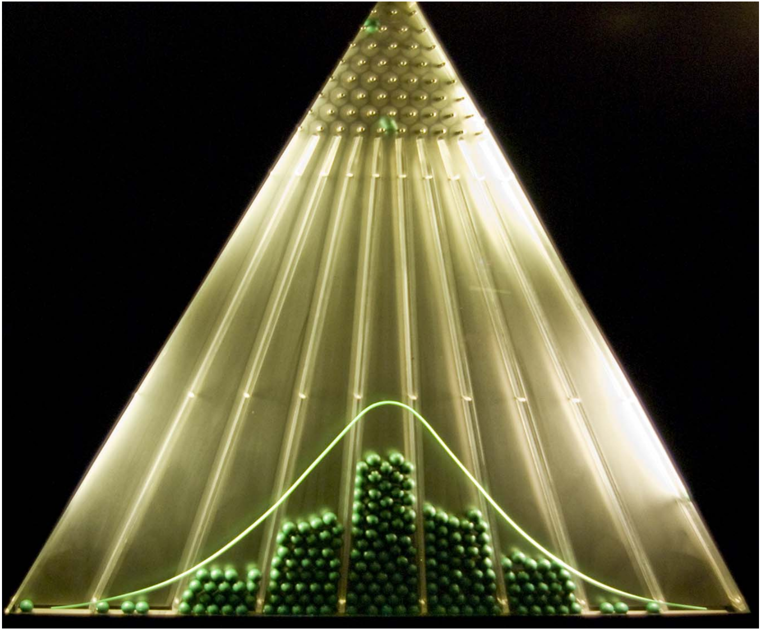
Problème : que faire si X ne suit pas une loi normale ?

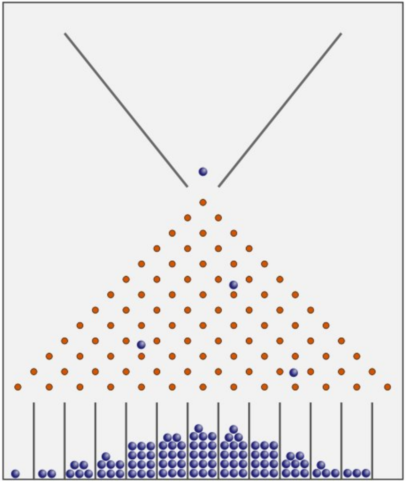
$\Longrightarrow$ sauvé grâce au théorème central-limite :

Théorème 2

- X : variable aléatoire
- espérance de X : μ, variance de X : σ^2
- échantillon de taille n avec remise sur X
- n suffisamment grand ($n \geq 30$ si la distribution de X n'est pas trop dissymétrique, $n \geq 50$ sinon)
- $\overline{X}$: variable aléatoire « moyenne de l'échantillon »
- Alors : $\dfrac{\overline{X} - E(\overline{X})}{\sqrt{V(\overline{X})}} = \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0; 1)$

Exemple 1 : la planche de Galton (1/2)





- chaque niveau $\Longrightarrow$ expérience de Bernoulli
- $\Longrightarrow X \sim$ loi binomiale

  $\Longrightarrow X \not\sim$ loi normale
- théorème 2 $\Longrightarrow$

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{V(\overline{X})}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0; 1)$$

- Grenelle de l'environnement
  $\Longrightarrow$ réduction des déchets
  $\Longrightarrow$ analyse des déchets

- impossible à réaliser sur toute la population
  $\Longrightarrow$ échantillon de taille 100 :

450	320	320	390	410	415	380	390
440	350	400	380	430	400	375	$\cdots$

- moyenne de l'échantillon $=$ 390 kg/an/habitant
- $\overline{X}$: variable aléatoire « moyenne de l'échantillon »
- $\sigma = 20$ supposé connu
- $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \Longrightarrow$ écart-type de $\overline{X} = 20/10 = 2$

Estimation de proportions de succès (1/3)

Problème : 

v.s. 

- lancement de la Wii $\Longrightarrow$ étude de marché
- échantillon de taille $n \Longrightarrow$ succès = Wii, échec = PS3
- $p =$ proportion de succès dans toute la population
- P_i : variable aléatoire « succès du $i$ème individu »

Question : peut-on déduire p en observant les p_i ?

Estimation de proportions de succès (2/3)

Question : peut-on déduire p en observant les p_i ?

- $p =$ proportion de succès dans la population (p pas trop petit)
- échantillon i.i.d de taille n assez grand
- $P_i =$ variable aléatoire « succès du $i$ème individu »
- $P_i \sim$ loi binomiale $\mathcal{B}(1, p)$
- $\overline{P} =$ moyenne de l'échantillon

Théorème central-limite $\Longrightarrow \dfrac{\overline{P} - E(\overline{P})}{\sqrt{V(\overline{P})}} \sim \mathcal{N}(0; 1)$

Or $\overline{P} \sim \frac{1}{n}\mathcal{B}(n; p) \Longrightarrow E(\overline{P}) = \dfrac{np}{n}$ et $V(\overline{P}) = \dfrac{p(1-p)}{n}$

Théorème 3

- échantillon i.i.d de taille n assez grand
- $p =$ proportion de succès dans la population (p pas trop petit)
- $\overline{P} =$ moyenne de l'échantillon
- Alors : $\dfrac{\overline{P} - E(\overline{P})}{\sqrt{V(\overline{P})}} = \dfrac{\overline{P} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \sim \mathcal{N}(0; 1)$

$\Longrightarrow$ on peut estimer p en observant la valeur de $\overline{P}$

Expert : 80%  v.s. 20% 

- échantillon de 100 personnes $\Longrightarrow$ 70 Wii et 30 PS3
- Doit-on croire l'expert ?

si $p = 80\%$ alors Théorème 3 $\Longrightarrow \dfrac{\overline{P} - 0,8}{\sqrt{\frac{0,8 \times 0,2}{100}}} = 25(\overline{P} - 0,8) \sim \mathcal{N}(0;1)$

$$\begin{aligned}\text{Prob}(\overline{P} \leq 0.7) &= \text{Prob}(25(\overline{P} - 0.8) \leq 25 \times (0.7 - 0.8)) \\ &= \text{Prob}(25(\overline{P} - 0.8) \leq -2,5) \approx 0,62\%\end{aligned}$$

Application au réchauffement climatique

- Avril 2007 : étude tms-sofres / CNRS :
  opinion des gens sur le réchauffement climatique
- 1000 personnes de 15 ans et $+$ interrogées
  $\Longrightarrow$ échantillon i.i.d
- 790 pensent qu'il y a un changement climatique
- 210 ne le pensent pas
- $\overline{P}$: proportion de succès moyenne de l'échantillon
- p : proportion de personnes pensant qu'il y a dérèglement climatique dans la population française

$$\frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0;1)$$

$\Longrightarrow$ estimation de $p = 79\%$, écart-type de $\overline{P} \leq \sqrt{\frac{0.25}{1000}} \approx 0,015$

Caractéristiques des estimateurs

Transparents précédents $\Longrightarrow$ si échantillon de grande taille
alors $\overline{X} \approx \mu \Longrightarrow$ estimer μ par $\overline{X}$

Estimateur non biaisé

- Estimateur T d'un paramètre $\theta \Longrightarrow$ valeur estimée $\hat{\theta}$
- T non biaisé si $E(T) = \theta$

⚠ estimateur biaisé $\Longrightarrow$ on surévalue ou sous-évalue θ

Estimateur convergent

- Estimateur T d'un paramètre θ
- T convergent si $E[(T - \theta)^2] \to 0$ lorsque la taille de l'échantillon $\nearrow$

moyenne $\overline{X}$ et proportion de succès $\overline{P}$:
estimateurs non biaisés et convergents

- population $=$ 4 nombres $\{1, 2, 3, 4\}$
- $\mu = \frac{5}{2}$ et $\sigma^2 = \frac{5}{4}$
- échantillon de taille 2 avec remise :

échant.	espérance	variance	échant.	espérance	variance
1 1	$E(\overline{X}) = 1$	$V(\overline{X}) = 0$	1 2	$E(\overline{X}) = \frac{3}{2}$	$V(\overline{X}) = \frac{1}{4}$
1 3	$E(\overline{X}) = 2$	$V(\overline{X}) = 1$	1 4	$E(\overline{X}) = \frac{5}{2}$	$V(\overline{X}) = \frac{9}{4}$
2 1	$E(\overline{X}) = \frac{3}{2}$	$V(\overline{X}) = \frac{1}{4}$	2 2	$E(\overline{X}) = 2$	$V(\overline{X}) = 0$
2 3	$E(\overline{X}) = \frac{5}{2}$	$V(\overline{X}) = \frac{1}{4}$	2 4	$E(\overline{X}) = 3$	$V(\overline{X}) = 1$
3 1	$E(\overline{X}) = 2$	$V(\overline{X}) = 1$	3 2	$E(\overline{X}) = \frac{5}{2}$	$V(\overline{X}) = \frac{1}{4}$
3 3	$E(\overline{X}) = 3$	$V(\overline{X}) = 0$	3 4	$E(\overline{X}) = \frac{7}{2}$	$V(\overline{X}) = \frac{1}{4}$
4 1	$E(\overline{X}) = \frac{5}{2}$	$V(\overline{X}) = \frac{9}{4}$	4 2	$E(\overline{X}) = 3$	$V(\overline{X}) = 1$
4 3	$E(\overline{X}) = \frac{7}{2}$	$V(\overline{X}) = \frac{1}{4}$	4 4	$E(\overline{X}) = 4$	$V(\overline{X}) = 0$

- espérance $\begin{cases} \text{des moyennes des échantillons} = \mu \\ \text{des variances des échantillons} = \frac{5}{8} \neq \frac{5}{4} \Longrightarrow \text{biais !} \end{cases}$

Biais : estimation ponctuelle d'une variance (2/5)

- X : variable aléatoire sur l'ensemble de la population
- σ^2 : variance de X
- échantillon i.i.d.de taille n
- X_i : variable aléatoire correspondant au $i$ème individu
- x_i : valeur observée de X_i
- $\overline{X}$: variable aléatoire « moyenne sur l'échantillon »
- $\overline{x}$: valeur observée de $\overline{X}$
- $s_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 = \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2$
- $s_n^2 =$ variance de l'échantillon

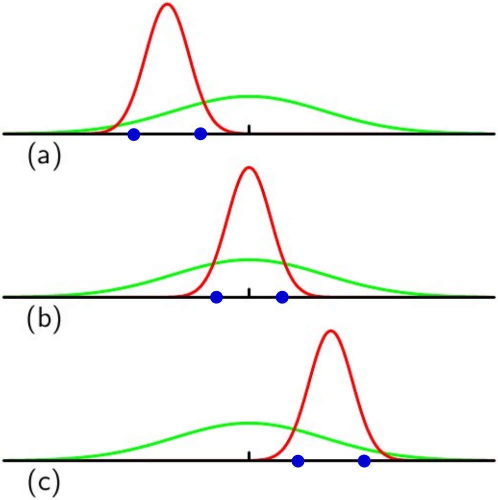
Problème : pourquoi s_n^2 n'est-il pas un bon estimateur de σ^2 ?

- $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$
- s_n^2 est une réalisation de S_n^2
- $E(S_n^2) = E\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E(X_i^2)\right) - E\left(\overline{X}^2\right)$
- Or échantillon i.i.d $\Longrightarrow \forall i,\ E(X_i^2) = E(X^2)$

$$\Longrightarrow E(S_n^2) = E(X^2) - E(\overline{X}^2)$$

- Or $V(X) = E(X^2) - E(X)^2$ et $V(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 = E(\overline{X}^2) - E(X)^2$
  $\Longrightarrow E(S_n^2) = V(X) + E(X)^2 - V(\overline{X}) - E(X)^2 = V(X) - V(\overline{X})$
- Or échantillon i.i.d $\Longrightarrow V(X) = \sigma^2$ et $V(\overline{X}) = \frac{\sigma^2}{n}$

$$\Longrightarrow E(S_n^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$



- courbe verte : la population
- courbes rouges : échantillons
- pts bleus : valeurs observées
- variance S_n^2 sous-estimée : mesurée par rapport à la moyenne de l'échantillon au lieu de la moyenne de la population

Variance corrigée

- X : variable aléatoire sur la population
- σ^2 : variance de X sur cette population
- échantillon de taille n avec remise
- X_i : variable aléatoire correspondant au $i$ème individu
- $\overline{X}$: moyenne des X_i
- alors : $E\left(\frac{n}{n-1}\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{n}\right) = \sigma^2$
- *Variance corrigée :* $\frac{n}{n-1}$ fois la variance de l'échantillon S_n^2