# TD1 STATISTIQUES 2 / HPC - BIG DATA 2023

## RAPPELS - ESTIMATEURS

# Exercice 1 :

Soit la variable aléatoire X, mesure de R entachée de l'erreur e :
$X = R + e$ , avec $e \sim N(0,\sigma^2)$ , $\sigma^2$ étant inconnue.
On a $E[X]=R$, $V[X]=V[e]=\sigma^2$, et donc $X \sim N(R,\sigma^2)$.
On cherche un estimateur S' non biaisé de la surface $S = \pi R^2$.
On dispose de n mesures indépendantes $x_i$ de R (n pouvant être petit), considérées comme des réalisations de n variables aléatoires $X_i$ indépendantes distribuées selon la loi de X.

Pour construire l'estimateur S' de S, deux stratégies 'naturelles' pouvaient être adoptées :

- moyenner les n surfaces $S_i = \pi X_i^2$ → $S_1' = \frac{1}{n}\sum_{i=1}^{n} S_i$
- exploiter la moyenne empirique de l'échantillon $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ → $S_2' = \pi \bar{X}^2$

Stratégie 1 :

$$E[S_1'] = \frac{\pi}{n}\sum_{i=1}^{n} E[X_i^2] = \frac{\pi}{n}\sum_{i=1}^{n}\left(E[X_i]^2 + V[X_i]\right) = \pi\left(R^2 + \sigma^2\right) \quad \rightarrow \quad BIAIS[S_1'] = \pi\sigma^2$$

Stratégie 2 :

$$E[S_2'] = \pi E[\bar{X}^2] = \pi\left(E[\bar{X}]^2 + V[\bar{X}]\right) = \pi\left(R^2 + \frac{\sigma^2}{n}\right) \quad \rightarrow \quad BIAIS[S_2'] = \frac{\pi\sigma^2}{n}$$

Les 2 estimateurs sont biaisés mais la stratégie 2 est plus intéressante : $S_2'$ est asymptotiquement non biaisé (car permettant aux erreurs de mesure de se compenser), le débiaisage n'est alors pas nécessaire si vous disposez d'un échantillon suffisant.

Pour débiaiser ces estimateurs il suffit de leur retrancher une variable aléatoire dont l'espérance égale le biais. Il faut donc exploiter un estimateur non biaisé de $\sigma^2$.

Vous avez vu en 1ere année que l'estimateur 'naturel' de la variance est biaisé (biais induit par l'exploitation de la moyenne empirique), avec biais = $-\sigma^2/n$, on utilise donc ici l'estimateur débiaisé :

$$\sigma^{2*} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Les 2 estimateurs suivants constituent donc des estimateurs non biaisés de S et répondent à la question :

$$S_1' - \pi\sigma^{2*} = \frac{\pi}{n}\sum_{i=1}^{n} X_i^2 - \frac{\pi}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \qquad S_2' - \frac{\pi\sigma^{2*}}{n} = \pi\bar{X}^2 - \frac{\pi}{n(n-1)}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

# Exercice 2 :

1) a)

Soit F la fonction de répartition de X :
F(x) = P(X<x) = 0 si x<0
= x/a sur [0,a]
= 1 si x>a

T=MAX($X_i$) i=1,...,n ; soit G la fonction de répartition de T :
G(t) = P(T<t) = P(MAX($X_i$)<t) = P($X_1$<t,…,$X_n$<t), puis par indépendance des $X_i$ on obtient :

$$G(t) = \prod_{i=1}^{n} P(X_i < t) = F(t)^n = \left(\frac{t}{a}\right)^n$$ sur [0,a], G(t)=0 si t<0 et G(t)=1 si t>a.

La fonction de densité g(t) est alors obtenue par dérivation :

$$g(t) = \frac{nt^{n-1}}{a^n}$$ sur [0,a], et g(t)=0 pour t<0 ou t>a.

b)

$$E[T] = \int_0^a t g(t) dt = \frac{na}{n+1}$$

On en déduit un estimateur non biaisé de a, en exploitant les propriétés de la variable T :

$$\hat{a}_1 = \frac{n+1}{n} T$$

2) a)

Un problème se pose lors de l'application de la méthode de maximisation de vraisemblance lorsque le support de la loi exploitée dépend des paramètres à estimer, ce qui est le cas avec une loi uniforme.
On dispose de n mesures indépendantes $x_i$ , considérées comme des réalisations de n variables aléatoires $X_i$ indépendantes distribuées selon la loi de X de fonction de densité f définie par :
f(x)=1/a sur [0,a], et f(x)=0 pour x<0 ou x>a.

La fonction de vraisemblance L, fonction de l'unique variable a, est définie par :

$$L(a) = \prod_{i=1}^{n} f(x_i; a) = \frac{1}{a^n} si, MAX(x_1,...,x_n) \le a, et, MIN(x_1,...,x_n) \ge 0$$ , sinon L(a)=0.

Cette fonction n'est pas dérivable en son maximum, atteint pour a=MAX($x_1$,…,$x_n$).

L'estimateur obtenu par maximisation de la vraisemblance est donc $\hat{a}_2$=MAX($X_1$,…,$X_n$)=T, estimateur biaisé puisque BIAIS[$\hat{a}_2$] = E[$\hat{a}_2$]-a = E[T] - a = - a/(n+1).

b)

Le MSE combine biais et variance de l'estimateur, en effet en développant son expression :
MSE[â] = E[(â-a)$^2$] = V[â] + (BIAIS[â])$^2$.
E[$T^2$] = n$a^2$/(n+2) → V[T] = E[$T^2$]-E[T]$^2$ = n$a^2$/[(n+2)(n+1)$^2$]
Et donc on obtient : MSE[$\hat{a}_2$]=MSE[T]=2$a^2$/[(n+1)(n+2)]
Le calcul concernant l'estimateur non biaisé $\hat{a}_1$ obtenu en 1) b) donnerait :
MSE[$\hat{a}_1$]=V[$\hat{a}_1$]=[(n+1)/n]$^2$.V[T] = $a^2$/[n(n+2)]=[(n+1)/(2n)].MSE[$\hat{a}_2$] < MSE[$\hat{a}_2$] pour n>1