Méthode Bootstrap et applications en R

Préface :

- en stat, "bootstrapping" renvoie à faire de l'inférence à partir d'une distribution d'échantillonnage d'une statistique en rééchantillonnant l'échantillon lui-même avec remise, comme si nous avions affaire à une population finie.
- A condition que la distrib° rééchantillonnée imite bien l'originale, l'inférence est robuste. La précision s'améliore quand la taille de l'échantillon grandit en vertu du théorème central limite, lorsque celui-ci s'applique.
- L'idée du rééchantillonnage est apparu dans les travaux de R.A. Fisher (1935) et d'E.J.G. Pitman (1937, 1938).
- Concernant le bootstrap et le tirage avec remise, en utilisant des approx. Monte Carlo, ce sont les travaux d'Efron (Brad) qui (papier Annals of Stats, 1977).

Dans le papier d'Efron, le lien du bootstrap avec d'autres techniques de rééchantillonnage est présenté, c'est le début de l'intérêt de la communauté scientifique/statistiques. (ex: jackknife)

- L'avantage du bootstrap combiné aux approx. Monte Carlo est de pouvoir faire de l'inférence statistique sans hypothèses fortes (ex: hyp. paramétrique ⇒ de lois de proba...).

I INTRODUCTION

I- Histoire

- Le bootstrap fait partie des méthodes de rééchantillonnage (stats non-paramétriques).
- Quelques techniques sont + anciennes que le bootstrap : les méthodes de permutation de Fisher (1935) et Pitman (1937, 1938) ; le jackknife avec Quenouille (1949).

(John Tukey)

- Le bootstrapping est devenu applicable d'un point de vue pratique depuis l'avènement de l'informatique et le couplage avec les méthodes Monte Carlo.
- A la base, le but du bootstrap était de construire une approximation du jackknife.
- Le Jackknife est utile à la base pour les petits échantillons, car très consommateur en ressources de calcul.
- Bootstrap ordinaire implique un tirage avec remise n fois d'un échantillon de taille n
  $\Rightarrow n^n$ possible distinct ordered bootstrap samples (bien que certains soient équivalents sous l'hypothèse d'échangeabilité liées aux permutations).
  $\Rightarrow$ l'énumération des $\neq$ échantillons bootstrap est infaisable dès que n devient grand.
- L'idée de prendre des échantillons Monte Carlo de taille n avec remise à partir de l'échantillon d'origine existait déjà (Julian Simon, 1969), et permettait de s'affranchir de l'hypothèse paramétrique d'un modèle de probabilité par la génération de l'échantillon d'origine.
- Efron (1979) a connecté le bootstrapping avec le Jackknife, la méthode Delta, la validation croisée et les tests de permutation. Il a été le 1er à démontrer que le bootstrap était compétitif avec ces autres méthodes pour estimer l'écart-type d'un estimateur.
- Efron a également vite compris que le bootstrap pouvait être utile aussi pour le calcul d'intervalle de confiance, de tests d'hypothèses, ... (voir Efron and Gong (1983), ...)
- Le bootstrap s'est énormément développé dans les années 80 et 90.
- Auteurs connus pour le bootstrap: Efron, Tibshirani, Chernick, Therneau, Michael Martin, Peter Hall, David Freedman.
- Malgré des résultats théoriques prometteurs dans les années 80, quelques exemples montrent que le bootstrap n'est parfois pas convistent... L'estimateur ne CV pas vers le vrai paramètre pd la taille de l'échantillon devient infinie).

- Généralement, le bootstrap est consistent lorsque le TCL s'applique.
  Une condition suffisante est la condition de Lyapanov: existence du $2+\delta^e$ moment de la distribution de la population.
- Il existe des versions du bootstrap pour remédier aux pb d'inconsistance dans certains cas bien connus.

II - Definition et lien avec la méthode Delta et d'autres méthodes de rééchantillonnage

- Definition informelle : le but du bootstrapping est d'estimer un paramètre (via Monte Carlo) basé sur des données (ex: moyenne, médiane, écart-type). On s'intéresse aussi aux propriétés de la distribution de l'estimation de ce paramètre, pour construire des intervalles de confiance par exemple. Mais sans faire d'hyp. restrictives sur la (forme de la) distribution d'origine des données observées.

- Pour le cas i.i.d., la brique de base est la distribution empirique : c'est la distribution discrète qui donne un poids égal à chaque observation (poids = $\frac{1}{n}$ si taille échantillon = n).
  Notée F_n en général. (cf schéma)

- La plupart des paramètres que nous étudierons (ex: la moyenne d'un échantillon) sont des fonctions de la distribution inconnue F de la population, des fonctions qui prennent F en entrée et retournent un réel correspondant. Dans notre cas, on s'intéresse aux fonctionnelles des FdR. Par exemple,

  → si $E[X] = \mu$ où $X \sim F$, alors ? $\mu = \int x \, dF(x)$ $\quad \hat{\mu}_n = \int x \, dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$

  → si $Var(X) = \sigma^2$ alors ? $\sigma^2 = \int (x-\mu)^2 \, dF(x)$

  ⇒ On note que les estimations données par ces paramètres en pratique (par ex la moyenne) sont issues des mêmes fonctionnelles appliquées à F_n - (la FdR empirique donc).

- L'idée du bootstrap est d'utiliser ce que l'on sait à partir des données, sans introduire d'hypothèses non-justifiées à propos de la distribution de (la population / l'échantillon).

- Principe du bootstrap : quand on note F la distribution de la population, et $T(F)$ la fonctionnelle qui définit le paramètre d'intérêt (ex : $T(F) = \mu = \int x\,dF(x)$).
En estimant ce paramètre depuis un échantillon de taille n, il faut considérer que F_n joue le rôle de F et que F_n^* (la distribut° bootstrap) joue le rôle de F_n dans le processus de rééchantillonnage.

- L'échantillon de base est composé de n observat° iid de distribut° F. Notez que l'estimation du paramètre d'intérêt vaut $T(F_n)$. Dans un bootstrapping, on laisse F_n jouer le rôle de F et on prend n observations iid provenant de F_n. Puisque F_n est la FdR empirique, il s'agit simplement de tirer aléatoirement avec remise les observations de l'échantillon d'origine. Car $P(X_i = x_i) = \frac{1}{n}$ à chaque tirage doit être satisfait (d'où la remise).

- Ex : $n = 5$, $X_1 = 7$ $X_3 = 3$ $X_5 = 6$ $X_2 = 5$ $X_4 = 9$ $\Rightarrow \bar{\mu}_n = 6$ dans cet échantillon

Tirer aléatoirement avec remise dans cet échantillon génère un "échantillon bootstrap".

Échantillon bootstrap : on le note $X_1^*, X_2^*, X_3^*, X_4^*, X_5^*$; cet échantillon permet de définir la distribution bootstrap de rééchantillonnage avec remise, notée F_n^*.
$\Rightarrow$ estimation bootstrap $= T(F_n^*)$.

Ex avec l'échantillon d'origine précédent : $X_1^* = 5$ $X_3^* = 7$ $X_5^* = 5$ $X_2^* = 9$ $X_4^* = 7$ $\Rightarrow \bar{\mu}_n^* = 6{,}6$

Donc l'estimateur bootstrap $= 6{,}6 \neq 6$ (échantillon origine)

- Remarquez que certaines valeurs/observ. de l'échantillon d'origine sont répétées ou omises dans l'échantillon bootstrap.

- Si on répète la génération d'un nouvel échantillon, bootstrap, on aura (probablement quasi-certainement) une estimation de la moyenne encore différente.

- Si on répète l'opération plein de fois on aura plein de valeurs ≠ pour $\bar{x}_n^*$.
  ⇒ on pourra faire un histogramme des valeurs de la moyenne
  ⇒ on l'appellera l'approximation Monte Carlo de la distribution bootstrap.
  ⇒ La moyenne de toutes ces valeurs (donc la moyenne des moyennes ici) sera proche de 6 puisque la moyenne théorique de la distribut° bootstrap est la moyenne de l'échantillon d'origine. Cependant, l'histogramme issu du rééchantillonnage permet de rendre compte de la variabilité de ces estimations, de leur asymétrie, écart-type, etc...

- En théorie, l'estimation bootstrap exacte du paramètre peut être calculée en moyennant sur tous les échantillons bootstraps possibles. (dans le cas précédent, l'estimation exacte vaudrait 6).
Mais en pratique, puisqu'il y a n^n échantillons bootstraps distincts possibles, ça devient vite infaisable (ex: n=10, ⇒ 10 000 000 000!) ⇒ en pratique, on utilise une approx. Monte-Carlo!

- Si on génère aléatoirement M (M=10 000 par ex.) échantillons bootstrap, la distribut° des estimations bootstrap approchera la distribution bootstrap des l'estimation.
Plus M est grand, plus l'histogramme approche la (vraie) distribution bootstrap ~~de l'estimation~~. (rarement calculable à cause des limites de calcul notamment lorsque on ne connaît pas l'expression analytique)

- Fonctionnement de l'approx. Monte Carlo :
  ① Générer un échantillon par tirage avec remise depuis la distribution empirique des données, on a ainsi créé un échantillon bootstrap.
  ② Calculer $T(F_n^*)$, l'estimation bootstrap de $T(F)$: on remplace l'échantillon d'origine par l'échantillon bootstrap et l'estimation bootstrap de $T(F)$ à la place de l'estimation de $T(F)$ (on prend $\hat{T}(F_n)$) vis les données d'origine!
  ③ Répéter ① et ② M fois avec M grand.

⚠: IMPORTANT: l'application de Monte Carlo approx au bootstrap implique 2 sources d'erreur:
  1- l'approx. Monte Carlo de la distribution bootstrap, qui peut être considérablement réduite si M est grand.
  2- l'approximation de la distribution bootstrap F_n^* par la distrib. originelle F de la population.

* Fin du 1er cours de 2h

- Si $T(F_n^*) \xrightarrow{n\to+\infty} T(F)$ alors le bootstrapping fonctionne.
  ⇒ ça marche dans bcp de cas, mais pas toujours.
- Rappel : Th. de Glivenko-Cantelli : $F_n \longrightarrow F$ uniformément ($\mathbb{P}(\lim_{n\to\infty} \|F_n - F\|_\infty = 0) = 1$; $\sup_x |F_n(x) - F(x)| \xrightarrow{n\to\infty} 0$?)
- Souvent nous savons que l'estimateur empirique est convergent. Par ex.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to+\infty} E[X] \quad \text{d'après la loi des grands nombres.}$$

Donc on a (1) $T(F_n) \xrightarrow{n\to+\infty} T(F)$.

Mais cela dépend aussi de la fonctionnelle T et de sa régularité (dérivée, ...)

⇒ On a donc aussi besoin de (2) $T(F_n^*) \xrightarrow{n\to+\infty} T(F_n)$, plus précisément

(2) $T(F_n^*) - T(F_n) \xrightarrow{n\to+\infty} 0$

⇒ Rq : Le bootstrap reste basé sur une théorie asymptotique : des études par simulations pour la taille de l'échantillon est petite sont nécessaires pour évaluer la pertinence de son usage.

① - Le Jackknife

- Introduit par Quenouille (1949)
- Le but était d'améliorer l'estimation en corrigeant son biais. [estimateur → biais ou variance observé]
- Tukey (1958) montre ensuite que le jackknife peut aussi être utilisé pour estimer l'écart-type d'une estimation.
- Le nom "jackknife" (couteau suisse) vient du fait que la technique peut être utilisée pour 4 objectifs.
- Le bootstrapping utilisent des échantillons bootstrap pour estimer une variabilité de l'estimation, alors que le jackknife utilisent les "pseudo-observations".

Question : "l'estimateur bootstrap" est-il $T(F_n^*)$? NON
ou la moyenne des $T(F_n^*)$? OUI (il y a application de Monte Carlo)

- Tout d'abord, notons $\hat{\mu}$ une estimation basée sur un échantillon d'observations iid de taille n, $(X_1, \dots, X_n)$ où $X_i \sim F$.

Comme précédemment, on note F_n la FdR empirique et

$$\begin{cases} \mu = T(F) \text{ une fonctionnelle} \\ \hat{\mu} = T(F_n) \\ \hat{\mu}_{(i)} = T(F_{n(i)}) \text{, où } F_{n(i)} \text{ est la FdR empirique pour l'échantillon sans l'observat. i.} \end{cases}$$

ex : $\hat{\mu}$ est la variance de la population (de l'échantillon), l'estimation jackknife de variance des σ^2 de $\hat{\mu}$ vaut ($\sigma^2 = Var_{JACKNIFE}(\hat{\mu})$)

→ variance de la variance : $\sigma^2_{JACK} = \frac{n}{n-1} \sum_{i=1}^{n} (\hat{\mu}_{(i)} - \mu^*)^2$ avec $\mu^* = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_{(i)}$ (variance estimée sans l'observat. i ; moyenne des variances)

L'estimation jackknife de l'écart-type pour $\hat{\mu}$ (la variance) vaut $\sqrt{\sigma^2_{JACK}}$.

- Tukey définit les pseudo-observations, ou pseudo-valeurs : (en toute généralité)

$$\tilde{\mu}_i = \hat{\mu} + (n-1)(\hat{\mu} - \hat{\mu}_{(i)})$$

($\hat{\mu}$: variance empirique) → pour revenir au même que précédemment = aboutir à une estim. type Monte Carlo

⇒ l'estimation jackknife du paramètre μ vaut $\mu_{JACK} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mu}_i$. (même principe que Monte Carlo)

L'estimation est la moyenne des pseudo-valeurs.

- On écrit par ex l'estimation de la variance σ^2 de l'estimation $\hat{\mu}$ avec les pseudo-valeurs :

$$\sigma^2_{JACK} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\tilde{\mu}_i - \mu_{JACK})^2$$

variance de la variance : moyenne des variances de variance (μ_{JACK} : moyenne des variances ; $\frac{1}{n}$: moyenne)

⇒ dans cette forme, on voit que la variance est l'estimat. usuelle pour la variance d'une moyenne. Dans ce cas, c'est la moyenne des pseudo-valeurs.

- ~~Le lien entre~~ L'estimation jackknife ~~et l'estimé~~ de l'écart-type est une estimat. bootstrap où l'estimation du paramètre d'intérêt est remplacé par une approx. linéaire de lui-même. ⇒ il y a donc un lien étroit entre les 2...

② La méthode Delta : (estimateur fonction d'un autre estimateur)

On est souvent intéressé par les moments d'un estimateur. En particulier, c'est souvent la variance de l'estimateur qui nous intéresse.

Pour illustration, soit $\varphi = f(\alpha)$ avec ($\alpha, \varphi \in \mathbb{R}$, f une fonction dérivable par rapport à α.

$\Rightarrow$ ∃ une expansion de Taylor en α_0 : au 1er ordre cela donne, pour $\alpha \in v(\alpha_0)$,

$$\varphi = f(\alpha) = f(\alpha_0) + (\alpha - \alpha_0) f'(\alpha_0) + o(\alpha - \alpha_0)$$

$$\Rightarrow \quad \varphi = f(\alpha) \simeq f(\alpha_0) + (\alpha - \alpha_0) f'(\alpha_0)$$

$$\text{ou} \quad f(\alpha) - f(\alpha_0) \simeq (\alpha - \alpha_0) f'(\alpha_0)$$

$$\Rightarrow \quad \left(f(\alpha) - f(\alpha_0)\right)^2 \simeq (\alpha - \alpha_0)^2 \left(f'(\alpha_0)\right)^2$$

Maintenant, disons que φ est une v.a., on prend l'espérance dans l'égalité précédente :

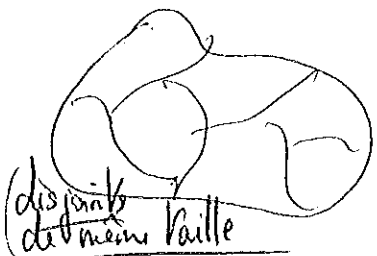
$$E\left[\left(f(\alpha) - \underbrace{f(\alpha_0)}_{\text{cte}}\right)^2\right] \simeq E\left[(\alpha - \alpha_0)^2\right] \underbrace{\left(f'(\alpha_0)\right)^2}_{\text{cte}}$$

Cette dernière égalité est l'approximation par la méthode Delta de la variance de $\varphi = f(\alpha)$ puisque le terme de gauche est à peu près la variance de φ et celui de droite la variance d'α multipliée par la constante $\left(f'(\alpha_0)\right)^2$ si l'on choisit α_0 comme la moyenne de α.

③ La validation croisée

La validation croisée est une procédure générale en modélisation statistique.
Elle est ~~par exemple~~ utilisée pour de la sélection de modèle parmi un ensemble de modèles, par exemple pour :

- ordre d'un processus AR
- select de variable en régression logistique / régression linéaire multiple
- nombre de composantes d'un modèle mélange fini
- choix d'un modèle de classificat paramétrique
- élagage d'arbre de classification, ...

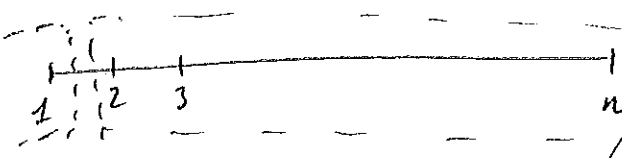


(disjoints deux à deux)

L'idée est de partager aléatoirement les données en 2 sous-ensembles : 1 pour calibrer (estimer) le modèle, l'autre pour le valider. On moyenne ensuite les erreurs du modèle données sur toutes les combinaisons pour choisir le meilleur modèle.

(4) Le sous-échantillonnage

- Développé par Hartigan (1969) ⇒ théorie des intervalles de confiance par sous-échantillonnage aléatoire.
- Récemment, cette théorie a été reliée au bootstrap, pour des données indépendantes ou dépendantes. Voir Politis et al. (1999); Politis and Romano (1994)....
- Définition: soient $S_1, S_2, \dots, S_{B-1}$; $(B-1)$ des (2^n-1) sous ensembles non-vides des entiers $1, 2, \dots, n$



2^n car pour chaque élément dénombrement

× 3

1,2 - 3	1-2-3	2
1,3 - 2	1,2,3	3
2,3 - 1	1	

$2^3 - 1 = 7$ possibilités

Ces $(B-1)$ sous-ensembles sont sélectionnés aléatoirement sans remise, et les estimateurs sont ensuite faites.

~~Déf: Valeur des applications~~

Preuve: ($E = \{1, 2, \dots, n\}$ si $\text{Card}(E) = n$ (ce qui est notre cas dans notre exemple) alors $\text{Card}(\mathcal{P}(E)) = 2^n$ où $\mathcal{P}(E)$ est l'ensemble des parties de E)

Par récurrence: • $n = 0$ ⇒ par d'élément ⇒ l'ensemble-vide a un seul sous-ensemble (lui-même).
⇒ $2^0 = 1$ donc ok

• supposons la prop. vraie pour $n = m$ ⇒ 2^m parties
pour $n = m+1$: l'ensemble E est donc non-vide. Soit a un élément de E.

Ou preuve: on peut aussi compter les parties à $0, 1, 2, \dots, n$ éléments d'un ensemble E à n éléments:

$$\text{Card}(\mathcal{P}(E)) = \underbrace{C_n^0}_{\text{ensemble-vide}} + C_n^1 + C_n^2 + \dots + C_n^n = (1+1)^n = 2^n$$

Si l'on enlève l'ensemble-vide, on a donc $2^n - 1$ parties.

Puis on poursuit le même raisonnement que précédemment à partir des échantillons créés?

III - Variété des applications

- On est tenté d'appliquer le bootstrap un peu partout... mais on a vu que cela ne marche pas toujours... alors comment le sait-on ? ↪ soit prouvé théoriquement (Thés de CV) ↪ soit prouvé par simus.

Par exemple en régression, il y a 2 approches au bootstrapping :

1) "bootstrapping residuals" : on fit le modèle puis on bootstrappe les résidus.

2) "bootstrapping vectors or cases" : on bootstrap sur $(y_i, x_{1i}, \dots, x_{ki})$ pour $i = 1, \dots, n$ puis on fit chaque fois un modèle ⇒ mieux si on a des doutes sur la forme paramétrique choisie pour la modélisation.

- Quelque soit l'allure des données, le bootstrap peut fonctionner : V cas d'heavy-tailed, highly skewed, ... puisqu'il n'y a pas d'hyp. param. (en fait)
- Simplicité de mise en oeuvre : Monte Carlo... et ordinateurs puissants aujourd'hui.

Mais ⚠ : il n'est pas toujours facile de voir quand le bootstrap ne fonctionne pas, et de savoir pourquoi...

IV - Bootstrap et R

- Beaucoup de librairies qui traite du sujet : (> help.search("bootstrap") ?? bootstrap
- Package ("bootstrap" "boot"
- Pour le rééchantillonnage, la fonction de base est "sample(.)".

V - Histoire et biblio

cf p 25

VI - Exercices

cf p. 26

Exemple: considérons l'estimateur du max. de vraisemblance de la variance σ^2 d'une v.a. gaussienne : $X_i \sim \mathcal{N}(\mu, \sigma^2)$

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - X_b)^2 \quad \text{avec} \quad X_b = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Cet estimateur a un biais : $b = -\frac{\sigma^2}{n}$ connu ici et dont l'expression analytique existe !

→ La (vraie) estimation bootstrap du biais vaut b. (comme concentration = 6 tout à l'heure).
Cet inconnue (& expression est inconnue !!!)

→ On définit $B^* = E(\theta^* - \hat{\theta})$ avec $\begin{cases} \hat{\theta} = S_n^2 \text{ estimateur du max de vraisemblance} \\ \theta^* = \frac{1}{n} \sum_{i=1}^{n} (X_i^* - X_b)^2 \text{ : estimateur bootstrap} \end{cases}$

L'approx. Monte Carlo de B^* vaut $B_{MC} = \frac{1}{N} \sum_{j=1}^{N} B_j^*$ avec B_j^* une estimation du biais pour le j^e échantillon bootstrap.

Ainsi $B_j^* = \theta_j^* - \hat{\theta}$ et $\theta_j^* = \frac{1}{n} \sum_{i=1}^{n} (X_{ij}^* - X_b)^2$

X_{ij}^* est l'i^e observation du j^e échantillon bootstrap.

X_b : moyenne de l'échantillon d'origine (⚠ : pas celle de l'échantillon bootstrap !)

Illustration 1 : machine

- Le but de détecter un biais est de corriger un estimateur en le débiaisant.

⚠ : la réduction du biais se fait en général au détriment de l'↑ de la variance de l'estimateur.
Il faut débiaiser l'estimateur si le carré du biais est > ↑ de la variance ⇒ MSE de $\hat{\theta}$ ↓
Sinon ce n'est pas forcément pertinent.
(car MSE = biais² + variance)

② Exemples en analyse linéaire discriminante (classification error rate)

③ Exemple simple en analyse linéaire discriminante et estimation du taux d'erreur bootstrap

On imagine 2 classes dans la population (de 10 individus) :
- → 5 individus ∈ classe 1, où la classe 1 $\sim \mathcal{N}\left((0,0)^T, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$
- → 5 " " " 2, " " $\sim \mathcal{N}\left((1,1)^T, \begin{pmatrix} 1 & ? \\ 0 & 1 \end{pmatrix}\right)$

On génère les individus pour chaque classe (on connaît donc ici les classes !) :
- → classe 1 : $(2,052; 0,339)^T$; $(1,083; -1,32)^T$; $(0,083; -1,524)^T$; $(1,278; -0,453)^T$; $(-1,226; -0,606)^T$
- → classe 2 : $(1,307; 2,268)^T$; $(-0,548; 1,741)^T$; $(2,498; 0,813)^T$; $(0,832; 1,409)^T$; $(1,498; 2,063)^T$

⇒ On va générer 6 échantillons bootstrap de taille 10 (taille de la population), et

Exercice 1 : on retrouve dans le beurre de cacahuètes des résidus d'Aflatoxine.
Dans les 12 lots de beurre de cacahuètes à disposition, on mesure la concentration d'Aflatoxine résiduelle : 4,34 5,06 4,53 5,07 4,33 5,16 4,38 4,43 4,33 4,72 4,32 4,96

a) Combien d'échantillons bootstrap possibles peut-on créer ?
R : 12^{12} puisqu'on en tire 12 et qu'il y a 12 possibilités à chaque tirage.

b) En utilisant R et fonction sample(.), générer 5 échantillons d'entiers entre 1 et 12. de taille 12.

c) Pour chacun des échantillons, créés, calculer la moyenne de la concentration en aflatoxine.

d) Calculer la moyenne de ces moyennes, et comparer la à la moyenne de l'échantillon d'origine.

e) Trouver le minimum et le maximum des 5 moyennes provenant des échantillons bootstrap. C'est un intervalle de confiance grossier sur la moyenne. Si vous aviez créé 1000 échantillons bootstrap, un intervalle de confiance à 95% de la moyenne aurait été défini par les 25e et 975e moyennes lorsqu'elles sont classées par ordre croissant.

---

2. ESTIMATION

On parlera ici d'estimation ponctuelle, et de comment le bootstrap peut être utilisé pour estimer le biais d'une estimation, afin de l'améliorer après ajustement.
Historiquement, le bootstrap a été introduit pour estimer l'écart-type d'une estimation, puis pour l'ajustement d'un biais (contrairement au jackknife)

I. Estimation du biais

① Ajustement bootstrap

$E[\hat{\theta} - \theta] = 0 \Rightarrow \hat{\theta}$ estimateur sans biais de θ.
(v.a. ; 1ère)

Soit $E(X)$ l'espérance de X. Soit $\hat{\theta}$ une estimation du paramètre θ.
Posons $X = \hat{\theta} - \theta$
Le biais d'un estimateur $\hat{\theta}$ de θ est donné par $b = E(X) = E(\hat{\theta} - \theta)$.

Rq : la distribut bootstrap d'un estimateur peut donner une idée de la loi de l'estimateur si le bootstrap s'applique. (ex : EMV qui est gaussien se retrouve).

II - Estimer un paramètre de localisation

Cette partie parle de l'estimation ponctuelle d'indicateurs de localisation (moyenne, médian). Dans le cas général de distributions "classiques" (moments finis jusqu'à l'ordre 4), la moyenne est le meilleur estimateur de localisation, et le plus adapté pour les distribut symétriques. Mais cet indicateur de localisation n'est pas toujours défini (ex : loi de Cauchy), et n'est pas le meilleur indicateur si les distribut présentent de fortes asymétries.

(1) Estimer une moyenne :

- Pour les populations avec un moment d'ordre 1 fini, on peut estimer la moyenne.
- Pour des distrib. symétriques, la moyenne est la meilleure mesure pour la tendance centrale. Comment le bootstrap se comporte-t-il pour déterminer cette estimation ?
- Dans le cas gaussien par ex., la moyenne empirique est l'EMV de μ si $(X \sim \mathcal{N}(\mu, \sigma^2))$. La moyenne empirique est aussi l'ESBVM (Estimateur Sans Biais et de Variance Minimale) de μ. Comment le bootstrap estimator peut-il faire mieux ? → en fait il ne peut PAS !

Dans ce cas précis, on peut déterminer l'estimation bootstrap sans approx. Monte Carlo. En effet car la moyenne empirique est aussi la moyenne de toutes les estimations bootstrap de moyennes empiriques ⇒ l'estimation bootstrap vaut la moyenne empirique !

② Estimer une médiane

Dans le cas de distributions fortement asymétriques, la médiane ou le mode sont de meilleurs indicateurs de tendance centrale.
Par exemple pour la distribution de Cauchy, la médiane ~~empirique~~ théorique ("population median") est bien définie et la médiane empirique ("sample median") est un estimateur consistant de la médiane théo.
Dans ce cas, aussi, l'estimateur bootstrap de la médiane est aussi consistant, ~~mais~~ et n'a pas d'avantages supplémentaires.
Cependant, si l'on veut estimer l'écart-type des indicateurs (moyenne, médiane), le bootstrap permet de le quantifier.

III - Estimer la dispersion

de la distribution de la population ? ou de la distrib. de l'estimateur ?

Quand le moment d'ordre 2 existe, l'écart-type peut être estimé.
Pour des estimateurs non-biaisés, cet écart-type est appelé "standard error of the estimate".
L'écart-type est souvent considéré comme le meilleur indicateur de dispersion, en tout cas pour des distributions classiques. De plus, cet indicateur intervient dans des résultats fondamentaux (ex: inégalité de Chebyshev) qui nous aide à comprendre la dispersion des données.
Malheureusement, cet indicateur n'existe parfois pas (ex: Cauchy). On se base alors sur un autre indicateur: l'"interquartile range".

① Estimer l'écart-type d'un estimateur:

- Certains auteurs ont donné un calcul exact de l'estimateur bootstrap de l'écart-type d'une médiane par exemple. Il suffit ensuite de le comparer avec l'approximation Monte Carlo du bootstrap sur un échantillon réel.

Application: pour simplifier, considérons un nombre impair d'observations X_i: $n = 2m+1$
Pour estimer la médiane, ce sera donc $X_{(m+1)}$, où $X_{(i)}$ désigne la i^e observation classée par ordre croissant.

On peut donner la variance exacte de la distribution bootstrap de la médiane en utilisant des propriétés célèbres de statistiques d'ordre.

Soit $X^*_{(1)}, X^*_{(2)}, \dots, X^*_{(n)}$ la statistique d'ordre d'un échantillon bootstrap issu de $X_{(1)}, \dots, X_{(n)}$.
Soit $X_{(i)}$ la i^e plus petite observation de l'échantillon d'origine, et

$$N_i^* = \#\{j : X_j^* = x_{(i)}, \ i = 1, 2, \dots, n\}.$$

$\Rightarrow$ N_i^* est le nombre d'observations dans l'échantillon bootstrap égales à l'observation i ordonnée dans l'échantillon d'origine.

On peut montrer que $\sum_{i=1}^{k} N_i^* \sim \text{Binomiale}(n, p = \frac{k}{n})$.

Maintenant, soit P^* la probabilité sous rééchantillonnage bootstrap. On a :

$$P^*\left(\underbrace{X^*_{(m+1)}}_{\text{médiane échantillon bootstrap}} > x_k\right) = P^*\left(\underbrace{\sum_{i=1}^{k} N_i^*}_{\sim \text{Binomiale}} \leq n\right) = \sum_{j=0}^{n} C_n^j \left(\frac{k}{n}\right)^j \left(\frac{n-k}{n}\right)^{n-j}$$

En utilisant des relations entre sommes binomiales et fonctions beta incomplètes, on peut déterminer une formule explicite pour l'estimateur bootstrap de l'écart-type de la médiane.
$\Rightarrow$ On n'a pas besoin d'approx. Monte Carlo.

- Dans d'autres problèmes d'estimation de paramètres, on n'a pas de formule explicite. L'approx. Monte Carlo est alors nécessaire.

(Soit $\hat{\theta}$ l'estimateur (basé sur les observat.) de θ.
Soit $\hat{\theta}_i^*$ l'estimateur bootstrap de θ pour l'échantillon bootstrap i.
Soit θ^* la moyenne des $\hat{\theta}_i^*$.

Alors, comme décrit dans Efron (1982), l'estimateur bootstrap de l'écart-type de l'estimateur $\hat{\theta}$ est donné par

$$SD_b = \sqrt{\frac{1}{B-1} \sum_{k=1}^{B} (\hat{\theta}_k^* - \theta^*)^2}$$

B : nb échantillons bootstrap.
issu de la distribut. bootstrap de l'estimateur.
écart-type de l'estimateur bootstrap de l'estimation

② Estimer "l'interquartile range" :

Un moyen naturel d'estimer cette quantité est de prendre le 25e et le 75e percentile de la distribution bootstrap pour obtenir l'estimation. Il faut donc au préalable classer les estimations bootstrap. Application : machine.

IV Régression linéaire