# Méthodes Monte-Carlo pour l'inférence statistique

Master parcours SSD - UE Statistique Computationnelle

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

# Introduction

Inférence statistique :

- Induire les caractéristiques d'une population à partir d'un échantillon (issu de cette population).
- Deux grandes questions : fournir des estimations de ces caractéristiques et faire des tests d'hypothèses.

# Introduction

Inférence statistique :

- Induire les caractéristiques d'une population à partir d'un échantillon (issu de cette population).
- Deux grandes questions : fournir des estimations de ces caractéristiques et faire des tests d'hypothèses.

Méthodes Monte-Carlo pour l'inférence :

- Tirer des échantillons à partir d'un modèle probabiliste de la population.
- Evaluer empiriquement l'incertitude de l'estimation.

# Introduction



Inférence statistique :

- Induire les caractéristiques d'une population à partir d'un échantillon (issu de cette population).
- Deux grandes questions : fournir des estimations de ces caractéristiques et faire des tests d'hypothèses.

Méthodes Monte-Carlo pour l'inférence :

- Tirer des échantillons à partir d'un modèle probabiliste de la population.
- Evaluer empiriquement l'incertitude de l'estimation.

$\Rightarrow$ Applications :

- étudier la distribution d'échantillonnage d'un estimateur
- estimer les propriétés d'un test statistique

# Estimateur / Estimation

Soit $(X_1, ..., X_n)$ un $n$-échantillon distribué selon la loi de $X$.

- Un estimateur $\hat{\theta}$ d'un paramètre $\theta$ est une fonction de l'échantillon :

$$\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$$

# Estimateur / Estimation

Soit $(X_1, ..., X_n)$ un $n$-échantillon distribué selon la loi de $X$.

- Un estimateur $\hat{\theta}$ d'un paramètre $\theta$ est une fonction de l'échantillon :
$$\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$$
- C'est lui même une variable aléatoire qui possède sa propre distribution.
- On parle de distribution d'échantillonnage (sampling distribution).

# Estimateur / Estimation



Soit $(X_1, ..., X_n)$ un $n$-échantillon distribué selon la loi de $X$.

- Un estimateur $\hat{\theta}$ d'un paramètre $\theta$ est une fonction de l'échantillon :
$$\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$$
- C'est lui même une variable aléatoire qui possède sa propre distribution.
- On parle de distribution d'échantillonnage (sampling distribution).
- Une estimation est la valeur de l'estimateur pour une réalisation $(x_1, ..., x_n)$ de l'échantillon.

# Méthodes MC & inférence - caractérisation d'un estimateur et intervalles de confiance

# Méthodes MC pour l'estimation

- L'approche MC (couverte ici) vise à caractériser les propriétés d'un estimateur d'une grandeur / paramètre que l'on connaît (et donc qu'on peut contrôler / fixer).

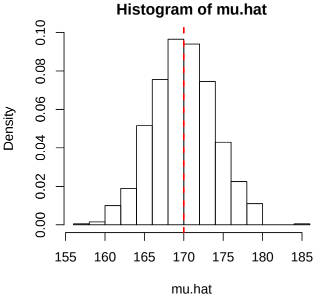
- Typiquement : le paramètre d'une loi de probabilité
  - (on parle parfois de bootstrap paramétrique)

# Méthodes MC pour l'estimation

- L'approche MC (couverte ici) vise à caractériser les propriétés d'un estimateur d'une grandeur / paramètre que l'on connaît (et donc qu'on peut contrôler / fixer).
- Typiquement : le paramètre d'une loi de probabilité
  - (on parle parfois de bootstrap paramétrique)
- Elle consiste à :
  1. tirer $m$ $n$-échantillons $\left(X_1^{(j)}, \dots, X_n^{(j)}\right)_{j=1,\dots,m}$, en fixant le paramètre $\theta$ à estimer.
  2. calculer les $m$ estimations $\hat{\theta}^{(j)}$, $j = 1, \dots, m$.
  3. étudier la distribution d'échantillonnage de $\hat{\theta}$ à partir de ces $m$ réalisations.

# Illustration

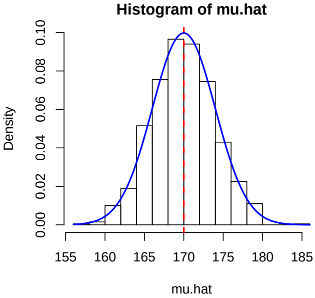
- On fait l'hypothèse que la taille des étudiants est distribuée normalement selon $\mathcal{N}(\mu = 170, \sigma = 20)$.
- Illustration de la distribution d'échantillonnage de l'estimateur "moyenne empirique" de $\mu$ : variabilité attendue sur 1000 échantillons de $n = 25$ élèves.




```
> m = 1000; n  = 25
> mu  = 170; sigma = 20
> mu.hat  = replicate(m,
expr = {x=rnorm(n,mu,sigma); mean(x)}))
> hist(mu.hat, prob = TRUE)
> abline(v=mu, lty=2, lwd=2, col=2)
```

# Illustration

- On fait l'hypothèse que la taille des étudiants est distribuée normalement selon $\mathcal{N}(\mu = 170, \sigma = 20)$.
- Illustration de la distribution d'échantillonnage de l'estimateur "moyenne empirique" de $\mu$ : variabilité attendue sur 1000 échantillons de $n = 25$ élèves.




```
> m = 1000; n  = 25
> mu  = 170; sigma = 20
> mu.hat  = replicate(m,
expr = {x=rnorm(n,mu,sigma); mean(x)})
> hist(mu.hat, prob = TRUE)
> abline(v=mu, lty=2, lwd=2, col=2)

> curve(dnorm(x, mu, sigma/sqrt(n)),
add = TRUE, col = "blue", lwd = 2)
```

# Motivations : à quoi ça sert ?

- Les paramètres des lois usuelles sont bien connus.
  - on dispose d'estimateurs performants (e.g., non biaisés et de variance minimale).
  - on connait leur distribution d'échantillonnage : on peut leur associer des intervalles de confiance.

# Motivations : à quoi ça sert ?

- Les paramètres des lois usuelles sont bien connus.
  - on dispose d'estimateurs performants (e.g., non biaisés et de variance minimale).
  - on connait leur distribution d'échantillonnage : on peut leur associer des intervalles de confiance.
- Leurs propriétés sont souvent basées sur des hypothèses (e.g., de normalité) et/ou des résulats asymptotiques.

# Motivations : à quoi ça sert ?

- Les paramètres des lois usuelles sont bien connus.
  - on dispose d'estimateurs performants (e.g., non biaisés et de variance minimale).
  - on connait leur distribution d'échantillonnage : on peut leur associer des intervalles de confiance.
- Leurs propriétés sont souvent basées sur des hypothèses (e.g., de normalité) et/ou des résulats asymptotiques.
- Dans les applications réelles, ces hypothèses ne sont pas toujours vérifiées
  - pas toujours tout à fait normal, peu d'observations.

# Motivations : à quoi ça sert ?



- Les paramètres des lois usuelles sont bien connus.
  - on dispose d'estimateurs performants (e.g., non biaisés et de variance minimale).
  - on connait leur distribution d'échantillonnage : on peut leur associer des intervalles de confiance.
- Leurs propriétés sont souvent basées sur des hypothèses (e.g., de normalité) et/ou des résulats asymptotiques.
- Dans les applications réelles, ces hypothèses ne sont pas toujours vérifiées
  - pas toujours tout à fait normal, peu d'observations.

$\Rightarrow$ l'approche MC permet (entre autres) de quantifier l'impact d'hypothèses non vérifiées sur les propriétés de l'estimateur.

# Caractérisation d'un estimateur



Pour caractériser un estimateur $\hat{\theta}$, on peut s'intéresser :

- à son biais : $\text{Biais}(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$
- à son erreur quadratique moyenne : $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$
- à son erreur type $\text{se}(\hat{\theta})$, définie comme l'écart type de sa distribution d'échantillonnage.

# Caractérisation d'un estimateur



Pour caractériser un estimateur $\hat{\theta}$, on peut s'intéresser :

- à son biais : $\text{Biais}(\hat{\theta}) = E\left[\hat{\theta}\right] - \theta$
- à son erreur quadratique moyenne : $\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$
- à son erreur type $\text{se}(\hat{\theta})$, définie comme l'écart type de sa distribution d'échantillonnage.

Ces critères permettent notamment :

- de caractériser la précision d'un estimateur en fonction de la taille $n$ de l'échantillon
- de comparer la performance de différents estimateurs

# Illustration : estimer la moyenne d'une loi normale

On souhaite estimer la moyenne d'une loi normale à partir d'un échantillon de taille $n = 20$.

# Illustration : estimer la moyenne d'une loi normale



On souhaite estimer la moyenne d'une loi normale à partir d'un échantillon de taille $n = 20$.

Estimateur naturel : moyenne empirique : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$

- On sait qu'il est sans biais (loi des grands nombres)
- On connait son erreur-type : $se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$.
  - $var(X_1 + ... + X_n) = n\sigma^2$, donc $var(\bar{X}_n) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$

# Illustration : estimer la moyenne d'une loi normale

On souhaite estimer la moyenne d'une loi normale à partir d'un échantillon de taille $n = 20$.

Estimateur naturel : moyenne empirique : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- On sait qu'il est sans biais (loi des grands nombres)
- On connait son erreur-type : $se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$.
  - $var(X_1 + ... + X_n) = n\sigma^2$, donc $var(\bar{X}_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$
- En pratique, on ne connait pas $\sigma^2$ et on utilise la variance empirique comme estimateur de la variance :

$$\hat{se}(\bar{x_n}) = \frac{1}{\sqrt{n}}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}.$$

# Illustration : estimer la moyenne d'une loi normale

On souhaite estimer la moyenne d'une loi normale à partir d'un échantillon de taille $n = 20$.

Estimateur naturel : moyenne empirique : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

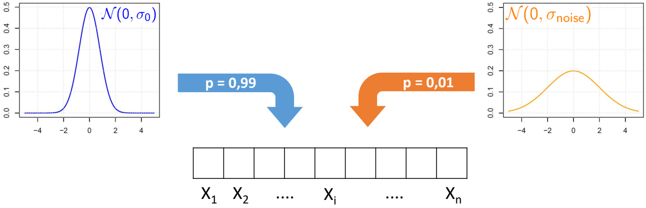
- On sait qu'il est sans biais (loi des grands nombres)
- On connait son erreur-type : $se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$.
  - $var(X_1 + ... + X_n) = n\sigma^2$, donc $var(\bar{X}_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$
- En pratique, on ne connait pas $\sigma^2$ et on utilise la variance empirique comme estimateur de la variance :

$$\hat{se}(\bar{x}_n) = \frac{1}{\sqrt{n}}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}.$$

$\Rightarrow$ pourquoi aller chercher plus loin ?

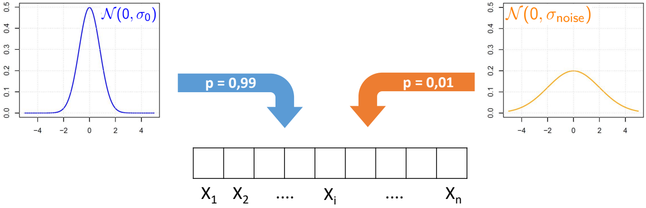
# Illustration : estimer la moyenne d'une loi normale

Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :



# Illustration : estimer la moyenne d'une loi normale

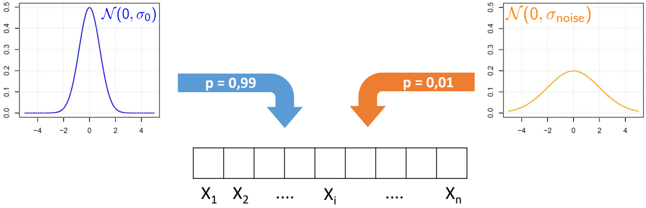
Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :




D'autres estimateurs pourraient être plus robustes :

# Illustration : estimer la moyenne d'une loi normale

Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :

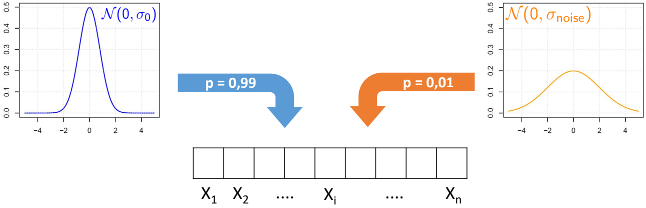



D'autres estimateurs pourraient être plus robustes :

- la médiane,

# Illustration : estimer la moyenne d'une loi normale

Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :

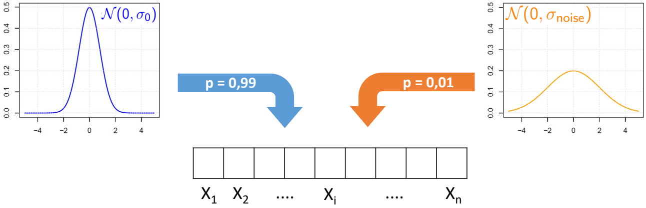



D'autres estimateurs pourraient être plus robustes :

- la médiane,
- la moyenne empirique "trimmée" (trimmed) où on élimine la plus grande et la plus petite observation,

# Illustration : estimer la moyenne d'une loi normale

Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :

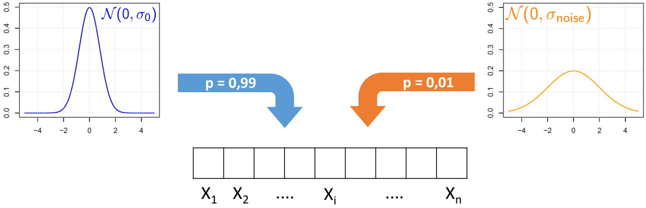



D'autres estimateurs pourraient être plus robustes :

- la médiane,
- la moyenne empirique "trimmée" (trimmed) où on élimine la plus grande et la plus petite observation,
- la moyenne empirique "trimée" d'ordre $k$ où on supprime les $k$ plus petites et $k$ plus grandes observations.

# Illustration : estimer la moyenne d'une loi normale

Et si nos observations n'étaient pas tout à fait normales mais "contaminées" par 1% d'"outliers" :




D'autres estimateurs pourraient être plus robustes :

- la médiane,
- la moyenne empirique "trimmée" (trimmed) où on élimine la plus grande et la plus petite observation,
- la moyenne empirique "trimée" d'ordre $k$ où on supprime les $k$ plus petites et $k$ plus grandes observations.

$\Rightarrow$ Problème : on ne connaît pas leurs propriétés.

# Illustration : estimer la moyenne d'une loi normale

Stratégie MC : simuler des échantillons et caractériser empiriquement ces estimateurs :

# Illustration : estimer la moyenne d'une loi normale

Stratégie MC : simuler des échantillons et caractériser empiriquement ces estimateurs :

1. tirer $m$ $n$-échantillons $\left(X_1^{(j)}, ..., X_n^{(j)}\right)_{j=1,...,m}$, en fixant le paramètre $\theta$ à estimer.

# Illustration : estimer la moyenne d'une loi normale

Stratégie MC : simuler des échantillons et caractériser empiriquement ces estimateurs :

1. tirer $m$ $n$-échantillons $\left(X_1^{(j)}, ..., X_n^{(j)}\right)_{j=1,...,m}$, en fixant le paramètre $\theta$ à estimer.
2. calculer les $m$ estimations $\hat{\theta}^{(j)}$, $j = 1, ..., m$.

# Illustration : estimer la moyenne d'une loi normale

Stratégie MC : simuler des échantillons et caractériser empiriquement ces estimateurs :

1. tirer $m$ $n$-échantillons $\left(X_1^{(j)}, ..., X_n^{(j)}\right)_{j=1,...,m}$, en fixant le paramètre $\theta$ à estimer.
2. calculer les $m$ estimations $\hat{\theta}^{(j)}$, $j = 1, ..., m$.
3. étudier la distribution d'échantillonnage des $\hat{\theta}^{(j)}$ :

$$\text{Biais}(\hat{\theta}) : \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}^{(j)} - \theta = \bar{\hat{\theta}} - \theta$$

$$\text{MSE}(\hat{\theta}) : \frac{1}{m}\sum_{j=1}^{m} (\hat{\theta}^{(j)} - \theta)^2$$

$$\text{Erreur type - se}(\hat{\theta}) : \left(\frac{1}{m-1}\sum_{j=1}^{m} \left(\hat{\theta}^{(j)} - \bar{\hat{\theta}}\right)^2\right)^{1/2}$$

(NB : ça n'a en général pas d'importance d'utiliser la version biaisée ou non de l'écart type car on simule en général de nombreux échantillons)

# Illustration : estimer la moyenne d'une loi normale - procédure R

```
> n = 20; m = 1000; k = 5
> e = matrix(0, m, 4)
> for(i in 1:m){
x = sort(rnorm(n))
e[i,1] = mean(x)
e[i,2] = mean(x[2:(n-1)])
e[i,3] = mean(x[(k+1):(n-k)])
e[i,4] = median(x)
}
> mse = apply(e, 2, function(x){mean(x^2)})
> se = apply(e, 2, function(x){sqrt(sum((x - mean(x))^2)/m)})

> mse
[1] 0.05165281 0.05317642 0.06306673 0.07978597

> se
[1] 0.2272608 0.2305817 0.2511308 0.2824580
```

$\Rightarrow$ TP : étude de la robustesse de ces estimateurs.

# Estimation d'un niveau de confiance - motivation

- En pratique, une estimation s'accompagne souvent d'un niveau de confiance, formalisé comme un intervalle de confiance.

# Estimation d'un niveau de confiance - motivation

- En pratique, une estimation s'accompagne souvent d'un niveau de confiance, formalisé comme un intervalle de confiance.
- Ces intervalles sont souvent obtenus sous des hypothèses de normalité de la population.
  - qui peuvent être justifiée si (on pense que) la loi est effectivement normale.
  - qui sont sinon liées à des approximations asymptotiques (e.g., théorème central limite).

# Estimation d'un niveau de confiance - motivation

- En pratique, une estimation s'accompagne souvent d'un niveau de confiance, formalisé comme un intervalle de confiance.
- Ces intervalles sont souvent obtenus sous des hypothèses de normalité de la population.
  - qui peuvent être justifiée si (on pense que) la loi est effectivement normale.
  - qui sont sinon liées à des approximations asymptotiques (e.g., théorème central limite).
- On peut appliquer le même type d'approche pour estimer le vrai niveau de confiance d'une procédure d'estimation quand on s'éloigne des hypothèses de normalité.

# Estimation d'un niveau de confiance - principe



Soit $X$ la variable aléatoire étudiée et $\theta$ le paramètre à estimer (à partir d'un échantillon de taille $n$).

On va s'appuyer sur une procédure de Monte Carlo suivante :

- Pour chaque répétition $j = 1, ..., m$ :
  - générer le $j$ème $n$-échantillon $\left(X_1^{(j)}, ..., X_n^{(j)}\right)$.
  - calculer l'intervalle de confiance $C_j$ correspondant.
  - vérifier si $\theta \in \mathcal{C}_j$.

Le niveau de confiance empirique est égal à la proportion d'intervalles de confiance contenant $\theta$.

# Estimation d'un niveau de confiance - illustration

- On cherche à estimer la variance $\sigma^2$ d'une variable aléatoire $X$.

# Estimation d'un niveau de confiance - illustration

- On cherche à estimer la variance $\sigma^2$ d'une variable aléatoire $X$.
- Si elle est normalement distribuée, et qu'on dispose d'un $n$-échantillon $(X_1, ..., X_n)$, alors

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

où $S_n^2$ est la variance empirique.

# Estimation d'un niveau de confiance - illustration

- On cherche à estimer la variance $\sigma^2$ d'une variable aléatoire $X$.
- Si elle est normalement distribuée, et qu'on dispose d'un $n$-échantillon $\left(X_1, ..., X_n\right)$, alors

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

  où $S_n^2$ est la variance empirique.
- Un intervalle de confiance à $100(1-\alpha)\%$ pour $\sigma^2$ est donné par :

$$\left[(n-1)S_n^2/\chi^2_{1-\alpha/2} \;\; ; \;\; (n-1)S_n^2/\chi^2_{\alpha/2}\right],$$

  où $\chi^2_\alpha$ est le quantile d'ordre $\alpha$ de la distribution $\chi^2(n-1)$.

# Estimation d'un niveau de confiance - illustration

- On peut vérifier la définition de cet intervalle de confiance en simulant une loi normale :

```
> m = 1000; n = 20; sigma = 2; alpha = 0.05
> I1 = numeric(m); I2 = numeric(m)
> for(i in 1:1000){
x = rnorm(n, mean = 0, sd = sigma)
I1[i] = (n-1)*var(x)/qchisq(1-alpha/2, df = n-1)
I2[i] = (n-1)*var(x)/qchisq(alpha/2, df = n-1)}
> print( mean(sigma^2 > I1 & sigma^2 < I2) )
```

# Estimation d'un niveau de confiance - illustration

- On peut vérifier la définition de cet intervalle de confiance en simulant une loi normale :

```
> m = 1000; n = 20; sigma = 2; alpha = 0.05
> I1 = numeric(m); I2 = numeric(m)
> for(i in 1:1000){
x = rnorm(n, mean = 0, sd = sigma)
I1[i] = (n-1)*var(x)/qchisq(1-alpha/2, df = n-1)
I2[i] = (n-1)*var(x)/qchisq(alpha/2, df = n-1)}
> print( mean(sigma^2 > I1 & sigma^2 < I2) )
```

- Cette procédure nous donne – comme attendu – approximativement 95%.
- On sait néanmoins que cette définition d'intervalle de confiance est assez sensible aux écarts à la normalité...

$\Rightarrow$ TP : évaluer la robustesse de cette procédure.

# Méthodes MC et estimation - résumé

En s'appuyant sur des techniques de simulation, l'approche MC permet de caractériser empiriquement la précision d'un estimateur en fonction de la taille de l'échantillon.

# Méthodes MC et estimation - résumé

En s'appuyant sur des techniques de simulation, l'approche MC permet de caractériser empiriquement la précision d'un estimateur en fonction de la taille de l'échantillon.

C'est une approche souvent plus simple à mettre en oeuvre que des développements mathématiques visant à affiner les approximations asymptotiques.

# Méthodes MC et estimation - résumé



En s'appuyant sur des techniques de simulation, l'approche MC permet de caractériser empiriquement la précision d'un estimateur en fonction de la taille de l'échantillon.

C'est une approche souvent plus simple à mettre en oeuvre que des développements mathématiques visant à affiner les approximations asymptotiques.

Elle permet notamment de comparer la performance de plusieurs estimateurs et d'évaluer empiriquement les niveaux de confiance associés quand on s'éloigne de leurs hypothèses de validité.

- taille d'échantillon et/ou loi de la variable aléatoire.

# Méthodes MC & inférence - tests statistiques

# Méthodes MC et tests statistiques - introduction



Test d'hypothèse : évaluer la validité d'une hypothèse statistique en fonction d'un échantillon.

- valeur théorique vs estimation et fluctuation d'échantillonnage.

# Méthodes MC et tests statistiques - introduction

Test d'hypothèse : évaluer la validité d'une hypothèse statistique en fonction d'un échantillon.

- valeur théorique vs estimation et fluctuation d'échantillonnage.

Faire un choix entre deux hypothèses statistiques :

- l'hypothèse nulle notée $H_0$
- une hypothèse alternative notée $H_1$

# Méthodes MC et tests statistiques - introduction

Test d'hypothèse : évaluer la validité d'une hypothèse statistique en fonction d'un échantillon.

- valeur théorique vs estimation et fluctuation d'échantillonnage.

Faire un choix entre deux hypothèses statistiques :

- l'hypothèse nulle notée $H_0$
- une hypothèse alternative notée $H_1$

Démarche générale :

1. Définir une statistique de test et sa distribution sous $H_0$.
2. Choisir un seuil de significativité, et en déduire la zone de rejet de $H_0$.
3. Evaluer la statistique de test sur un échantillon et prendre la décision : rejeter ou accepter $H_0$.

# Méthodes MC et tests statistiques - introduction

Exemple : on veut tester l'hypothèse $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ pour une v.a. $X$ de loi $\mathcal{N}(\mu, \sigma)$, de variance $\sigma^2$ connue, à partir d'un échantillon de taille $n$.

# Méthodes MC et tests statistiques - introduction



Exemple : on veut tester l'hypothèse $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ pour une v.a. $X$ de loi $\mathcal{N}(\mu, \sigma)$, de variance $\sigma^2$ connue, à partir d'un échantillon de taille $n$.

Procédure :

1. On va se baser sur la moyenne empirique pour estimer $\mu$.

# Méthodes MC et tests statistiques - introduction



Exemple : on veut tester l'hypothèse $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ pour une v.a. $X$ de loi $\mathcal{N}(\mu, \sigma)$, de variance $\sigma^2$ connue, à partir d'un échantillon de taille $n$.

Procédure :

1. On va se baser sur la moyenne empirique pour estimer $\mu$.
2. Sous $H_0$, on sait que $\bar{X}_n \to \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$.

# Méthodes MC et tests statistiques - introduction



Exemple : on veut tester l'hypothèse $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ pour une v.a. $X$ de loi $\mathcal{N}(\mu, \sigma)$, de variance $\sigma^2$ connue, à partir d'un échantillon de taille $n$.

Procédure :

1. On va se baser sur la moyenne empirique pour estimer $\mu$.
2. Sous $H_0$, on sait que $\bar{X}_n \to \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$.
3. On déduit notre région de rejet au seuil de significativité $\alpha$ : $T = \mu_0 + t_{1-\alpha} \times \sigma/\sqrt{n}$.

# Méthodes MC et tests statistiques - introduction

Exemple : on veut tester l'hypothèse $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$ pour une v.a. $X$ de loi $\mathcal{N}(\mu, \sigma)$, de variance $\sigma^2$ connue, à partir d'un échantillon de taille $n$.

Procédure :

1. On va se baser sur la moyenne empirique pour estimer $\mu$.
2. Sous $H_0$, on sait que $\bar{X}_n \to \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$.
3. On déduit notre région de rejet au seuil de significativité $\alpha$ : $T = \mu_0 + t_{1-\alpha} \times \sigma/\sqrt{n}$.
4. Si la réalisation $\bar{x}_n$ est supérieure à $T$, on rejette $H_0$.

(voir schéma...)

# Méthodes MC et tests statistiques - introduction

Deux types d'erreurs :

# Méthodes MC et tests statistiques - introduction

Deux types d'erreurs :

- rejeter $H_0$ à tort = le risque de première espèce.
  - on le note $\alpha$.
  - il est **défini a priori** : c'est le seuil de significativité choisi.

# Méthodes MC et tests statistiques - introduction

Deux types d'erreurs :

- rejeter $H_0$ à tort = le risque de première espèce.
  - on le note $\alpha$.
  - il est **défini a priori** : c'est le seuil de significativité choisi.
- accepter $H_0$ à tort = le risque de seconde espèce
  - on le note $\beta$.
  - il est propre à une **hypothèse alternative spécifique**.

# Méthodes MC et tests statistiques - introduction



Deux types d'erreurs :

- rejeter $H_0$ à tort = le risque de première espèce.
  - on le note $\alpha$.
  - il est **défini a priori** : c'est le seuil de significativité choisi.
- accepter $H_0$ à tort = le risque de seconde espèce
  - on le note $\beta$.
  - il est propre à une **hypothèse alternative spécifique**.

| | | Décision | |
|---|---|---|---|
| | | $H_0$ vraie | $H_0$ fausse |
| Réalité | $H_0$ vraie | $1-\alpha$ | $\alpha$ |
| | $H_0$ fausse | $\beta$ | $1-\beta$ |

(voir schéma...)

# Méthodes MC et tests statistiques - introduction

Deux notions importantes :

# Méthodes MC et tests statistiques - introduction

Deux notions importantes :

- la puissance du test = la probabilité de rejeter $H_0$ à raison ($\sim$ la probabilité de détecter l'hypothèse alternative).
  - elle vaut par définition $1 - \beta$.

# Méthodes MC et tests statistiques - introduction



Deux notions importantes :

- la puissance du test = la probabilité de rejeter $H_0$ à raison ($\sim$ la probabilité de détecter l'hypothèse alternative).
  - elle vaut par définition $1 - \beta$.

- la p-valeur du test = la probabilité d'observer sous $H_0$ une valeur plus élevée de la statistique de test que celle observée sur l'échantillon.
  - le plus faible $\alpha$ auquel on aurait pu rejeter l'hypothèse nulle compte tenu de notre observation.

(voir schéma...)

# Méthodes MC et tests statistiques

L'approche MC peut être déclinée pour étudier les performances d'un test statistique en terme :

- de risque de première espèce : le risque (empirique) de rejeter à tort l'hypothèse nulle est-il conforme à celui attendu ?
- de puissance : estimer empiriquement la probabilité de rejeter l'hypothèse nulle **pour une hypothèse alternative donnée**.

# Méthodes MC et tests statistiques

L'approche MC peut être déclinée pour étudier les performances d'un test statistique en terme :

- de risque de première espèce : le risque (empirique) de rejeter à tort l'hypothèse nulle est-il conforme à celui attendu ?
- de puissance : estimer empiriquement la probabilité de rejeter l'hypothèse nulle **pour une hypothèse alternative donnée**.

Cette approche peut notamment être utile pour évaluer la performance d'un test quand le nombre d'observations est limité, ou pour comparer la puissance de différents tests.

- dimensionnement du "sample size" de l'étude
- (voir schéma...)

# Méthodes MC et tests statistiques

Procédure pour mesurer empiriquement le risque de 1ère espèce d'un test :

# Méthodes MC et tests statistiques

Procédure pour mesurer empiriquement le risque de 1ère espèce d'un test :

- Pour $j = 1, .., m$
  - générer le $j$-ème $n$-échantillon $\left(X_1^{(j)}, ..., X_n^{(j)}\right)$ selon l'hypothèse nulle
  - calculer la statistique de test $T_j$
  - vérifier si l'hypothèse nulle est rejetée ou non (au seuil de significativité voulu)

# Méthodes MC et tests statistiques

Procédure pour mesurer empiriquement le risque de 1ère espèce d'un test :

- Pour $j = 1, .., m$
  - générer le $j$-ème $n$-échantillon $\left(X_1^{(j)}, ..., X_n^{(j)}\right)$ selon l'hypothèse nulle
  - calculer la statistique de test $T_j$
  - vérifier si l'hypothèse nulle est rejetée ou non (au seuil de significativité voulu)
- Le risque de 1ère espèce empirique est égal à la proportion de tests rejetés.
  - NB : on les rejette à tort.

# Méthodes MC et tests statistiques

Procédure pour mesurer empiriquement la puissance d'un test :

# Méthodes MC et tests statistiques



Procédure pour mesurer empiriquement la puissance d'un test :

- Pour $j = 1, .., m$
  - générer le $j$-ème $n$-échantillon $\left(X_1^{(j)}, ..., X_n^{(j)}\right)$ selon l'hypothèse alternative à évaluer
  - calculer la statistique de test $T_j$
  - vérifier si l'hypothèse nulle est rejetée ou non (au seuil de significativité voulu)

# Méthodes MC et tests statistiques

Procédure pour mesurer empiriquement la puissance d'un test :

- Pour $j = 1, .., m$
  - générer le $j$-ème $n$-échantillon $\left(X_1^{(j)}, ..., X_n^{(j)}\right)$ selon l'hypothèse alternative à évaluer
  - calculer la statistique de test $T_j$
  - vérifier si l'hypothèse nulle est rejetée ou non (au seuil de significativité voulu)
- La puissance empirique est égale à la proportion de tests rejetés.
  - NB : on les rejette à raison.

# Remarques et conclusion

# Conclusions

Monte-Carlo pour l'inférence :

- simuler des échantillons à partir d'un modèle probabiliste
- évaluer empiriquement l'incertitude de l'estimation

# Conclusions

Monte-Carlo pour l'inférence :

- simuler des échantillons à partir d'un modèle probabiliste
- évaluer empiriquement l'incertitude de l'estimation

4 "recettes" génériques :

1. caractérisation d'un estimateur
2. etimation d'un niveau de confiance
3. estimation du risque de 1ère espèce d'un test
4. estimation de la puissance d'un test

# Conclusions

Monte-Carlo pour l'inférence :

- simuler des échantillons à partir d'un modèle probabiliste
- évaluer empiriquement l'incertitude de l'estimation

4 "recettes" génériques :

1. caractérisation d'un estimateur
2. etimation d'un niveau de confiance
3. estimation du risque de 1ère espèce d'un test
4. estimation de la puissance d'un test

$\Rightarrow$ Simple à mettre en oeuvre.

# Conclusions

Monte-Carlo pour l'inférence :

- simuler des échantillons à partir d'un modèle probabiliste
- évaluer empiriquement l'incertitude de l'estimation

4 "recettes" génériques :

1. caractérisation d'un estimateur
2. etimation d'un niveau de confiance
3. estimation du risque de 1ère espèce d'un test
4. estimation de la puissance d'un test

$\Rightarrow$ Simple à mettre en oeuvre.

$\Rightarrow$ Utile pour dimensionner un problème et/ou quantifier l'impact de l'écart aux hypothèses.

# Remarques

- Méthode Monte Carlo : *toute méthode d'inférence statistique ou d'analyse numérique s'appuyant sur des techniques de simulation [de variables aléatoires]* (Rizzo, 2007, §6.1).
- Une autre application importante non couverte = simulation à large échelle d'un système pour étudier sa sensibilité aux fluctuations de ses entrées.
  - "sensitivity analysis" et/ou "uncertainty analysis"
- L'approche générale décrite dans la section "inférence" est parfois appellée bootstrap paramétrique. Le prochain cours s'intéressera au bootstrap "classique".
  - ré-échantillonnage à partir de l'échantillon.

# Mise en oeuvre R : la fonction replicate



Avec les approches MC, on fait beaucoup de boucles...

La fonction replicate permet de les faire pour vous :

```
> m = 1000;
> n  = 25
> mu  = 170; sigma = 20
> mu.hat = replicate(m,
           expr = {x = rnorm(n,mu,sigma); mean(x)})
```

Utilisation :

- 1er argument : nombre de réplications à faire
- 2ème argument : calcul à faire
- en sortie : un vecteur contenant les *m* valeurs obtenues

# Références

Maria L. Rizzo. *Statistical Computing with R*. CRC Press, 2007.