

# TD - 2.1

## Échantillonnage – Estimation

Vincent PAYET

« *Trouvez-moi un homme qui vit seul et dont la cuisine est sale en permanence, et six fois sur dix je vous montrerai un homme exceptionnel.* »  
— Charles Bukowski, Nouveaux contes de la folie ordinaire

**Objectifs** : cette fiche propose une utilisation de la simulation pour illustrer le cours d'échantillonnage et une utilisation des fonctions personnelles pour l'estimation par intervalles de confiance.

## Table des matières

<b>1</b>	<b>Échantillonnage</b>	<b>1</b>
1.1	Les lois de probabilité . . . . .	2
1.2	Distribution d'échantillonnage de la moyenne . . . . .	3
<b>2</b>	<b>Estimation</b>	<b>4</b>
2.1	Construire un Intervalle de Confiance . . . . .	4
2.2	Une fonction pour estimer un IC . . . . .	5
<b>3</b>	<b>Application</b>	<b>5</b>

## 1 Échantillonnage

Votre cours vous a donné des éléments sur la théorie des distributions d'échantillonnage. Par exemple, si une variable ne suit pas une loi normale mais que les échantillons sont d'effectifs suffisants, on considère que la loi de la moyenne tend vers une loi normale. Plus généralement, la question de l'échantillonnage est “que peut-on attendre d'un échantillon ?”.

Ici nous allons expérimenter *in silico* en réalisant des simulations de processus aléatoires. Nous pouvons générer des populations et réaliser des échantillons sur celles-ci. La fonction `sample()` notamment permet de faire un échantillon en partant d'une population :

```
> population <- 1:50
> sample(population,5) # vous pouvez jouer au loto...

[1] 16 32 36 6 34

> # avec ou sans répétition ?
> sample(population,20, replace=TRUE)

[1] 30 5 21 23 47 46 49 41 27 35 4 29 18 47 24 39 29 11 19 4

> # vous pouvez ensuite utiliser cet échantillon. Calculons la moyenne d'un autre tirage :
> mean(sample(population,5))

[1] 22,4
```

**Exercice :** On simule un processus :

```
> sample(c("N","B"),size=100, replace=TRUE, prob=c(0.9,0.1))
```

- Identifiez la variable aléatoire.
- Quelle est la loi de probabilité de cette variable aléatoire ?

## 1.1 Les lois de probabilité

Comment créer des populations dont on connaît la loi de probabilité ? Plusieurs lois sont utilisables dans  $\mathbb{R}$  (voir TDR1.3, [1]), notamment la loi normale (**norm**) et la loi uniforme (**unif**). Pour une loi, on peut obtenir 4 éléments : la fonction de répartition (**p**), les quantiles (**q**), la densité (**d**) ou un échantillon aléatoire, *random* en anglais (**r**). Il faut pour cela écrire le nom de la fonction avec la lettre **p**, **q**, **d** ou **r** et le nom de la loi :

```
> pnorm(1.96) # probabilité P(T<1,96) si T normale N(0;1)

[1] 0,9750021

> qnorm(0.95) # valeur t de la loi normale standard telle que P(T<t)=0,975

[1] 1,644854

> dnorm(0)      # ordonnée de la fonction de densité pour t=0

[1] 0,3989423

> rnorm(5)      # 5 valeurs aléatoires pour une variable aléatoire normale N(0;1)

[1] 0,7043652 -1,0866482 0,6521222 0,3977304 0,3964501
```

**Exercice :** Réalisez un histogramme (avec la fonction **hist()**) de 30 valeurs d'une loi normale centrée-réduite. Qu'en pensez-vous ?

## 1.2 Distribution d'échantillonnage de la moyenne

Soit  $X$  une variable aléatoire uniforme (par défaut entre 0 et 1). Nous allons créer une population de 10 000 individus puis réaliser 100 échantillons de 30 individus. Afin de comparer la loi de la variable initiale à la loi de la moyenne, nous observons les histogrammes des distributions de  $X$  et de  $\bar{X}$ .

```
> pop <- runif(10000)
> MoysCalc <- NULL # on initialise l'objet MoysCalc avant la boucle de calcul
> #
> # calculons la moyenne de 100 échantillons d'effectif n=30 :
> for (i in 1:100) {
+   MoysCalc[i] <- mean(sample(pop,30))
+ }
> sd(pop); mean(pop)

[1] 0,2895879

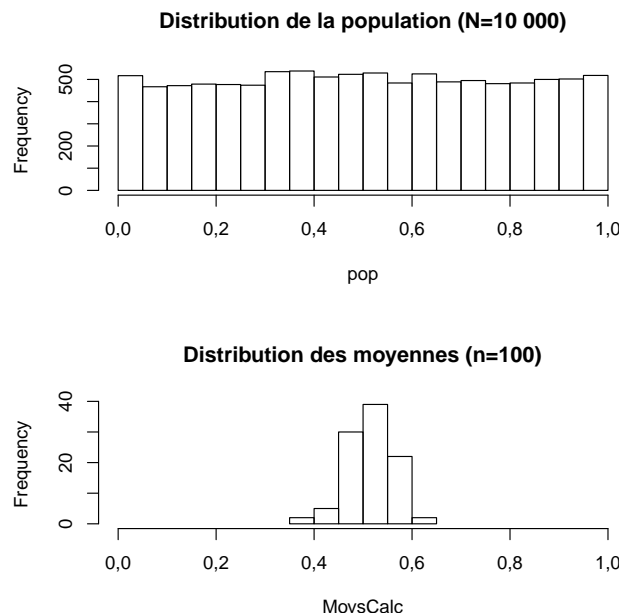
[1] 0,5019808

> sd(MoysCalc); mean(MoysCalc)


[1] 0,05216405

[1] 0,5009741

> par(mfrow=c(2,1)) # permet de positionner deux graphiques verticalement
> hist(pop, main="Distribution de la population (N=10 000)")
> hist(MoysCalc,xlim=c(0,1),main="Distribution des moyennes (n=100)")
```



Comparez les moyennes puis les écart-types (`sd()` calcule l'écart-type estimé) de la population puis de la distribution des moyennes. Quelles sont les formules du cours ?

La boucle `for` est connue et similaire à ce que vous connaissez par ailleurs.  permet des écritures plus condensées avec la fonction `replicate()`, notamment. Essayez cette écriture à la place de toute la boucle `for` précédente :

```
> hist(replicate(100,mean(sample(pop,300) )),xlim=c(0,1))
```

### Exercice :

1. Commentez les figures obtenues pour cette première simulation.
2. Retrouvez les relations entre les paramètres de la population et ceux de l'échantillon.
3. Que se passe-t-il si on fait 10 fois plus d'échantillons ?
4. Que se passe-t-il avec des échantillons d'effectif 300 ?
5. Refaites la même simulation avec une population d'origine suivant une loi normale. Il faut modifier judicieusement les bornes de l'histogramme des moyennes.

## 2 Estimation

La connaissance d'un échantillon permet de calculer des valeurs ponctuelles d'estimation de paramètres de la population. Rappelez les estimateurs ponctuels de la moyenne de la variance et d'une proportion. Quelle valeur renvoie la fonction `var()` ?

Ces estimations doivent si possible être accompagnées par des estimations par intervalle de confiance.

### 2.1 Construire un Intervalle de Confiance

Pour rappel, dans une population de variance connue  $\sigma_x^2$ , l'écart-type  $\sigma_m$  de la variable moyenne vaut  $\frac{\sigma_x}{\sqrt{n}}$ . On l'appelle erreur-standard. L'intervalle de confiance à 95% de la moyenne vaut :

$$IC_{95\%} = \bar{x} \pm t_{0,975} \frac{\sigma_x}{\sqrt{n}}$$

On obtient les valeurs des quantiles de la lois normales,  $t_{0,975}$  et  $t_{0,025}$  ainsi :

```
> qnorm(0.975)
[1] 1,959964
> qnorm(0.025)
[1] -1,959964
```

Dans le cas où la variance n'est pas connue, on l'estime avec  $\widehat{\sigma_x^2}$  et la loi de la moyenne est alors une loi de student à  $n - 1$  ddl. La loi de student se manipule avec `pt`, `qt`, `dt` et `rt`.

**Exercice :** Dans un centre de soin, on a mesuré la taille de 57 jeunes de 18 à 20 ans. On a trouvé  $\bar{x} = 175$  et  $s^2 = 27,5$ . On veut estimer la taille moyenne des jeunes dans la population. Proposez une estimation de la moyenne par intervalle de confiance à 95% puis à 99%. Il s'agit de calculer les bornes des intervalles.

## 2.2 Une fonction pour estimer un IC

Nous allons écrire une fonction qui retourne les bornes de l'intervalle de confiance. Une fonction est introduite avec un nom, le mot clef **function**, des paramètres et un contenu. Voici la création d'une fonction simple et son utilisation :

```
> UnNom <- function(a,b){  
+   S <- a+b  
+   S  
+ }  
> UnNom(3,5)  
  
[1] 8
```

### Exercice :

1. Que fait la fonction UnNom ?
2. Renommez-la judicieusement.
3. Créez une fonction pour calculer un intervalle de confiance à partir des éléments suivants à mettre en paramètre :
  - moyenne
  - variance
  - effectif de l'échantillon
  - niveau de confiance, vous pouvez lui attribuer une valeur par défaut.

## 3 Application

On a relevé les diamètres en cm de 15 arbres sur une parcelle forestière :

```
c(9, 8, 11, 10, 8, 6, 11, 9, 10, 9, 12, 8, 6, 9, 14)
```

1. Donnez une estimation ponctuelle de la moyenne et de la variance des arbres de la parcelle.
2. Donnez une estimation de la moyenne par intervalle de confiance à 95% et 99%.
3. Si on suppose que le diamètre suit une loi normale, quelle est la probabilité de trouver un arbre avec un diamètre supérieur à 15 cm ?

## Références

- [1] V. Payet. *TDR3, Distributions, formes de distribution et lois théoriques*. ISARA, 2012.