

# Chapitre 6 Le Logiciel R et Statistique Inférentielle

## Introduction

Dans ce chapitre on donne les fonctions R les plus couramment utilisées afin d'obtenir les estimations, les intervalles de confiance pour les paramètres classiques : moyenne, proportion, variance les tests d'hypothèses les plus classiques.

Nous allons fonder tous les exemples de ce chapitre sur le fichier de données *intimamedia.csv*

### Présentation du fichier:

La paroi artérielle est constituée de trois couches : l'intima, la media et l'adventice. L'épaisseur de l'intima-media est un marqueur reconnu d'athérosclérose. Elle a été mesurée par échographie sur un échantillon de 110 sujets. Des informations sur les principaux facteurs de risque ont aussi été recueillies.

Description	Unité ou Codage	Variable
Sexe	1=Homme; 2= Femme	SEXE
Age le jour de la visite	Année	AGE
Taille	Cm	taille
Poids	Kg	poids
Statut tabagique	0=Ne fume pas 1=A urité de fumer 2=Fume	tabac
Estimation de consommation pour les fumeurs et ex-fumeurs	Nombre de poquets/année	paqan
Activité physique	0=Non; 1=Oui	SPORT
Mesure de l'intima-média	Mm	mesura
Consommation d'alcool	0=Ne boit pas 1=Boit occasionnellement 2=Boit régulièrement	alcool

## 6.1 ESTIMATION PONCTUELLE

Le tableau ci-dessous présente les fonctions R qui donnent les estimateurs des paramètres les plus connus.

Paramètre	Notation	Estimateur	Estimation	Fonction R
moyenne	$\mu$	$\bar{X}$	$\bar{x}$	mean()
variance	$\sigma^2$	$\hat{\sigma}^2$	$\hat{\sigma}^2$	var()
médiane	$m_n$	$\tilde{M}_n$	$\tilde{m}_n$	median()
corrélation	$\rho$	$R$	$r$	cor()
proportion	$p$	$\hat{p}$	$\hat{p}$	mean()

## 6.1 INTERVAL DE CONFIANCE

### 6.1.1 Intervalles de confiance pour une moyenne

#### • Cas des grands échantillons ( $n > 30$ ) ou des petits échantillons avec hypothèse de normalité

L'intervalle de confiance est obtenu grâce à la fonction `t.test()`.

Exemple d'application:

On s'intéresse à l'estimation par intervalle de confiance de la moyenne du poids des personnes de l'étude.

```
> t.test(poids,conf.level=0.9)$conf.int
```

```
[1] 65.16024 67.80436
```

Nous obtenons l'intervalle de confiance [65.16,67.80] de niveau de confiance 0.9.

#### • Cas des petits échantillons

Dans le cas où aucune hypothèse n'est faite sur les données, nous conseillons d'utiliser une approche par *bootstrap*. Il est possible d'utiliser les fonctions `boot()` et `boot.ci()` disponibles dans le *package* `boot`.

Exemple d'application: Nous disposons d'un échantillon, représentatif de la population féminine, de dix femmes ayant les taux de cholestérol suivants (en g/l) :

```
> taux <- c(3,1.8,2.5,2.1,2.7,1.9,1.5,1.7,2,1.6)
```

Sans hypothèse gaussienne des données, nous proposons un intervalle de confiance de niveau 95 % du taux moyen de cholestérol des femmes

```

> require(boot)
> moyenne <- function(x, indices) mean(x[indices])
> taux.boot <- boot(taux, moyenne, R = 999, stype = "i", sim = "ordinary")
> boot.ci(taux.boot, conf = 0.95, type = c("norm", "basic", "perc", "bca"))

```

```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates
CALL :
boot.ci(boot.out = taux.boot, conf = 0.95, type =
c("norm", "basic", "perc", "bca"))
Intervals :
Level      Normal          Basic
95%   ( 1.787,  2.366 )   ( 1.770,  2.340 )
Level      Percentile      BCa
95%   ( 1.82,  2.39 )   ( 1.83,  2.41 )
Calculations and Intervals on Original Scale

```

### 6.1.1 Intervalles de confiance pour une proportion $p$

- Cas des grands échantillons ( $np \sim 5$  et  $n(1 - p) \sim 5$ )

Il est possible d'utiliser la fonction `binom.approx()` du *package* `epitools`.

*Exemple d'application* : On s'intéresse à l'estimation par intervalle de confiance de la proportion d'hommes (codes 2 ci-dessous) chez les personnes étudiés

```

> require(epitools)
> table(sexe) # Repartition de la variable sexe.
> binom.approx(141,226)[c("lower","upper")] # Calcul de l'ic

```

La fonction `prop.test()` fournit également un intervalle de confiance de la proportion

- Cas des petits échantillons : calcul exact

Il est possible d'utiliser la fonction `binom.test()`.

```

> binom.test(141,226)$conf # Calcul de l'ic avec n=226.

```

La fonction `binom.exact()` du *package* `epitools` renvoie le même intervalle de confiance.

### Intervalles de confiance pour une variance

- Cas des échantillons avec une hypothèse de normalité

```

> require(sigma2tools)
> sigma2.test(poids,conf.level=0.9)$conf

```

Cas des échantillons sans hypothèse de normalité

Il est possible d'utiliser les fonctions `boot()` et `boot.ci()` disponibles dans le *package* `boot`.

~ *Exemple d'application* : Reprenons les données du taux de cholestérol

des femmes et calculons un intervalle de confiance de la variance du

taux de cholestérol sans hypothèse de normalité des données.

```

> taux <- c(3,1.8,2.5,2.1,2.7,1.9,1.5,1.7,2,1.6)
> require(boot) # Charger le package boot.
> variance <- function(x,indices) var(x[indices])
> taux.boot <- boot(taux, variance, R = 999, stype = "i", sim = "ordinary")
> boot.ci(taux.boot, conf = 0.95, type = c("norm", "basic", "perc", "bca"))

```

```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates
CALL :
boot.ci(boot.out = taux.boot, conf = 0.95, type =
c("norm", "basic", "perc", "bca"))
Intervals :
Level      Normal          Basic
95%   ( 0.1060,  0.4412 )   ( 0.1026,  0.4448 )
Level      Percentile      BCa
95%   ( 0.0521,  0.3943 )   ( 0.1201,  0.4670 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable

```

Remarque : pour de grands échantillons sans hypothèse de normalité, on peut utiliser une approche asymptotique en utilisant :

```

> require(asympTest)
> asymp.test(poids,par="var")$conf

```

Ici vous trouvez un tableau récapitulatif pour les différents intervalles de confiance

Type	Condition de validité	Fonction R
proportion	$np \geq 5$ et $n(1 - p) \geq 5$ aucune	<code>prop.test(x)\$conf</code> <code>binom.test(x)\$conf</code>
moyenne	$n > 30$ ou normalité	<code>t.test(x)\$conf</code>
variance	normalité	<code>sigma2.test(x)\$conf</code>
médiane	aucune	<code>wilcox.test(x)\$conf</code>
corrélation	binormale	<code>cor.test(x)\$conf</code>

D'une manière similaire on donne un tableau récapitulatif pour les différents tests statistiques

Nature	Données	Conditions de validité	Fonction R
Tests paramétriques :			
moyenne	1 échantillon	$n > 30$ ou normalité	<code>t.test(x,...)</code>
	2 échantillons	normalité et variances égales	<code>t.test(x,y,...)</code>
	2 échantillons 2 éch. appariés	normalité $n > 30$ ou normalité	<code>t.test(x,y,var.equal=F)</code> <code>t.test(x,y,paired=T)</code>
variance	1 échantillon	normalité	<code>sigma2.test(x,...)</code>
	2 échantillons	normalité	<code>var.test(x,y,...)</code>
	2 échantillons	grand échantillon	<code>asyp.test(x,y,...)</code>
corrélation	1 échantillon	normalité, $\mathcal{H}_0 : \rho = \rho_0$	<code>cor.test(x,y..)</code>
	2 échantillons	normalité	<code>cor.test.2.sample(x,y,...)</code>
proportion	1 échantillon	$np \geq 5$ et $n(1 - p) \geq 5$	<code>prop.test(x,...)</code>
	1 échantillon		<code>binom.test(x,...)</code>
	2 échantillons	grand échantillon	<code>prop.test(x,y,...)</code>
Tests d'indépendance :			
$\chi^2$ d'indépendance	tableau de contingence tableau $2 \times 2$	effectifs théoriques $\geq 5$	<code>chisq.test(,correct=F)</code>
$\chi^2$ de Yates		effectifs théoriques $\geq 2.5$	<code>chisq.test()</code>
Fisher exact			<code>fisher.test()</code>
Tests d'adéquation :			
Shapiro-Wilk	1 échantillon	effectifs théoriques $\geq 5$	<code>shapiro.test(x,...)</code>
$\chi^2$ d'ajustement	1 échantillon		<code>chisq.test()</code>
Kolmogorov-Smirnov	1 échantillon		<code>ks.test(x,.)</code>
	2 échantillons	<code>ks.test(x,y)</code>	
Tests de position :			
médiane	1 échantillon	$\min(n_1, n_2) \geq 10$ $\min(n_1, n_2) \leq 10$	<code>binom.test(x,)</code>
test du signe	2 échantillons		<code>fisher.test(x,y,)</code>
	2 éch. appariés		<code>binom.test(x,y,paired=T)</code>
Mann-Whitney	2 échantillons		<code>wilcox.test(x,y,exact=F)</code>
Mann-Whitney	2 échantillons		<code>wilcox.test(x,y)</code>
Wilcoxon	2 éch. appariés		<code>wilcox.test(x,y,paired=T)</code>

Plus d'explications seront fournies dans la session de travaux pratiques sur le logiciel R ! (cf. Tp03)