Module 5
Inférence Statistique

Sommaire

Dernière mise à jour le 25 octobre 2022

Introduction

L'inférence statistique est le ...

> *Raisonnement par lequel on tire, à partir des observations faites sur un échantillon prélevé dans une population préalablement définie (population de référence), des conclusions concernant certaines caractéristiques quantitatives ou qualitatives de la population en question.(L'Office québécois de la langue française)*

En d'autres termes, le principe de l'inférence statistique est d'estimer les paramètres d'une population en se basant sur les données observées d'un échantillon. Afin de bien comprendre cette notion très importante en science des données, il sera judicieux de commencer par certaines définitions.

- **Population :** Une population (ou « population mère ») est définie comme l'ensemble de toutes les données collectées pour une étude particulière.
- **Échantillon :** Un échantillon est défini comme un sous-ensemble de la population (un sous-ensemble de données).
- **Individu :** Un individu est défini comme un élément d'une certaine population.

Formellement, l'objectif de l'inférence statistique est d'estimer le paramètre θ qui permet de décrire la population en question. Le paramètre θ correspond très souvent à la moyenne qui est estimé par la moyenne de l'échantillon considéré (telle qu'illustré dans la figure 5.1). Cependant, cet estimé est affecté par une certaine erreur. D'où l'intérêt d'estimer cette erreur en calculant ses bornes supérieure et inférieure qui correspondent à l'intervalle de confiance.

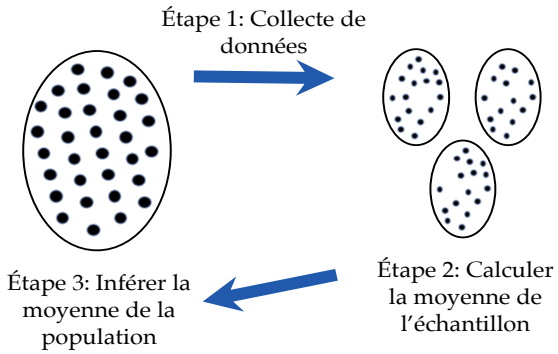

Figure 5.1: Principe de l'inférence statistique.

Soit Z une densité normale symétrique telle qu'illustrée à la Figure 5.2 (a). En relation avec ce qui est dit précédemment, Z peut correspondre à la densité de la moyenne à estimer. L'intervalle de confiance de la variable décrite par Z est situé entre $-z_{\alpha/2}$ et $z_{\alpha/2}$. La valeur de $z_{\alpha/2}$ représente le point pour lequel la probabilité d'observer une valeur

Z plus grande que $z_{\alpha/2}$ est égale à p. Ainsi, la zone grise dans cette figure comprend $(1-\alpha)$ échantillons de toutes les données, c'est-à-dire $(1-\alpha)$ échantillons ont une probabilité *d'être observés* plus grande que $z_{\alpha/2}$. Dans le cas où $\alpha/2 = 0.025$ (c'est-à-dire $\alpha = 0.5$ ou encore $1-\alpha = 0.95$), la valeur de $z_{\alpha/2}$ =1.96 (Figure 5.2 (b)) . La valeur de $z_{\alpha/2}$ peut être évaluée en utilisant la commande `qt` de R (comme illustré dans l'exemple 3.1)

5.1 Intervalle de confiance

Un intervalle de confiance permet d'évaluer la précision de l'estimation d'un paramètre statistique sur un échantillon.
Supposons que l'on désire calculer l'intervalle de confiance pour la moyenne d'une population dont l'écart type est connu. Ce calcul peut se faire en suivant les étapes suivantes :

1. Calculer la moyenne μ et l'écart-type σ de l'échantillon.

$$\mu = \frac{\sum_{i=1}^{n} x}{n} \quad \text{et} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x-\mu)^2}{n}} \tag{5.1}$$

2. Choisir le niveau de confiance associé à l'intervalle de confiance :

$$100(1-\alpha)\%$$

   Les niveaux de confiance les plus couramment utilisés sont 90%, 95 % et 99% ce qui correspond à une valeur de α égale respectivement à 0.1, 0.05 et 0.001.

3. Calculer la marge d'erreur en se basant sur la formule suivante :

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

4. Établir l'intervalle de confiance selon :

$$I_C = \left[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \ , \ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{5.2}$$

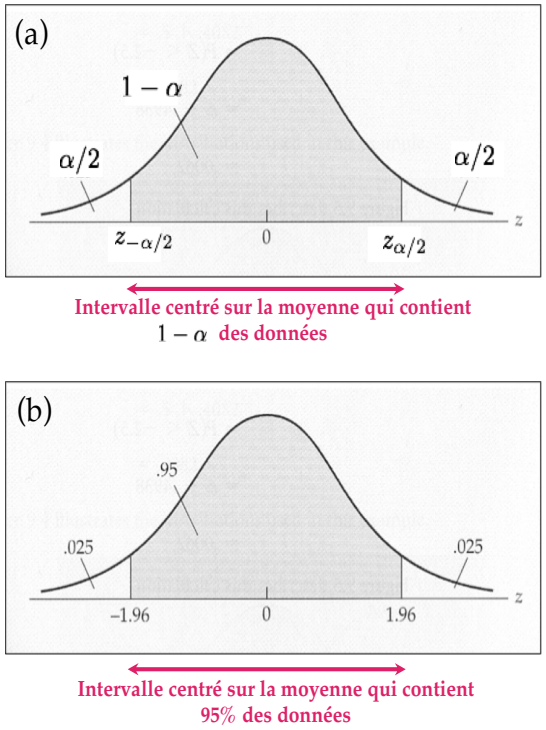

Figure 5.2: Distribution normale.

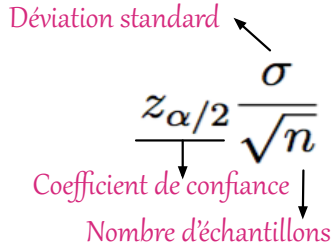

Figure 5.3: Marge d'erreur.

Note : Les bornes de l'intervalle de confiance telles que décrites précédemment sont applicables lorsque la distribution de la population est normale. Si la distribution n'est pas normale et que nous disposons d'un grand échantillon de donnée, les valeurs des bornes sont déterminés approximativement grâce à la relation 5.2 et ce en se basant sur le théorème de limite centrale.

▶ **Exemple 5.1** Revenons à l'ensemble des données tiré de l'étude évaluant la qualité de relation et la quantité service reçue par le patient lors de son séjour à l'hôpital (Exemple 4.1). La base de données comprend les informations de 534 patients ayant séjourné dans des hôpitaux de la région parisienne.

Les lignes de codes suivantes permettent d'interroger la base de données et d'obtenir les limites de l'intervalle de confiance à 95%. La commande `summary(MyData)` permet d'afficher toutes les variables considérées parmi lesquelles on retrouve la variable `age` qui varie entre 18 et 97 ans.

```
library(prettyR)
MyData <- read.csv(file="satisfaction_hopital.csv", header=TRUE, sep=",")
summary(MyData)
describe(MyData$age)
ageM = MyData$age[!is.na(MyData$age)]       # Supression des valeurs manquantes
z <- abs(qt(0.025,length(ageM)))     # Coefficient de confiance 0.05/2
IcInf<- mean(ageM) - z * sd(ageM)/sqrt(length(ageM))        #Borne inf
IcSup<- mean(ageM) + z * sd(ageM)/sqrt(length(ageM))        #Borne sup
```

Pour un niveau de confiance de 95%, nous devons identifier un nombre positif z pour lequel l'aire sous la courbe d'une distribution normale entre $-z$ et z est de 0.95. La fonction `qt` permet de trouver un nombre positif z de sorte que l'aire sous la courbe normale standard entre $-z$ et z est de 0.95%. Nous avons $100(1-\alpha)\% = 95\%$, donc $\alpha/2 = 0,025$.

Il est à noter que `qt (0.025, 534)`, donne une valeur négative qui correspond à $z-\alpha/2$. Cependant, étant donnée que distribution normale est symétrique, nous considérons la

valeur absolue (Se référer à la figure 5.2 (b) pour la visualisation de $z_{-\alpha/2}$ et $z_{\alpha/2}$). Les bornes de l'intervalle de confiance se situent de bord et d'autre de la moyenne.

```
    service          sexe            age          profession
 Min.   :1.000   Min.   :0.0000   Min.   :18.00   Min.   :1.000
 1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:45.00   1st Qu.:3.000
 Median :5.000   Median :0.0000   Median :60.00   Median :4.000
 Mean   :4.549   Mean   :0.4981   Mean   :58.21   Mean   :4.431
 3rd Qu.:7.000   3rd Qu.:1.0000   3rd Qu.:72.00   3rd Qu.:5.500
 Max.   :8.000   Max.   :1.0000   Max.   :97.00   Max.   :8.000
                                  NA's   :6       NA's   :107
```

La variable `age` contient 6 valeurs manquantes (`NA's :6`) qui peuvent être supprimées par la commande `!is.na` (Ligne 5) et stockées dans la variable `ageM`. Le nombre d'échantillons constituant `ageM` est de 527 patients. Les bornes inférieure et supérieure de l'intervalle de confiance sont, ensuite, déterminées en utilisant la relation 5.2 (Lignes 7 et 8).

```
> IcInf
[1] 56.69303
> IcSup
[1] 59.73122
```

Il existe une manière plus directe pour déterminer les bornes de l'intervalle de confiance en utilisant le package `PrettyR`. Les deux manières aboutissent exactement au même résultat.

```
1 library(prettyR)
2 t.test(ageM)$conf.int[1]                    #Borne inf
3 t.test(ageM)$conf.int[2]                    #Borne sup
```


5.2 Tests d'hypothèses

Rappelons que formellement, l'objectif de l'inférence statistique est d'estimer le paramètre θ qui permet de décrire la population en question. Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Leur principe consiste à confirmer ou à contredire une hypothèse concernant la valeur d'un paramètre de la population en se basant sur le test de signification. En d'autres termes, il s'agit de déterminer si l'échantillon de taille n dont nous disposons appartient à une population de moyenne μ_0 au seuil de signification α.

Les tests d'hypothèses se basent sur les étapes suivantes :

1. Formuler les hypothèses. Chaque test de signification commence par une hypothèse nulle H_0 qui représente une supposition qui soit proposée ou bien intuitivement ou bien parce qu'elle peut être utilisée comme point de départ du test d'hypothèse. L'hypothèse complémentaire à l'hypothèse nulle constitue l'hypothèse alternative (notée H_1 ou bien H_a). Le test d'hypothèse peut correspondre à une égalité (Figure 5.4 (a)) ou une inégalité (Figure 5.4 (b) et (c)). Dans le cas de l'égalité, nous avons les relations suivantes :

$$\begin{cases} \mathrm{H}_0 : m = \mu_0 \\ \mathrm{H}_1 : m \neq \mu_0 \end{cases} \tag{5.3}$$

   qui se lit : "L'hypothèse nulle correspond à une distribution dont la moyenne m est égale à μ_0".

2. Choisir la distribution de probabilité du paramètre à estimer. Dans le cas d'un grand nombre d'échantillon ($n > 30$), la distribution la plus utilisée est la densité normale. Dans le cas contraire, d'autres types de distribution sont utilisées. Dans ces cas, les analyses sont plus complexes.

3. Choisir le niveau de confiance associé à l'intervalle de confiance :

$$100(1-\alpha)\%$$

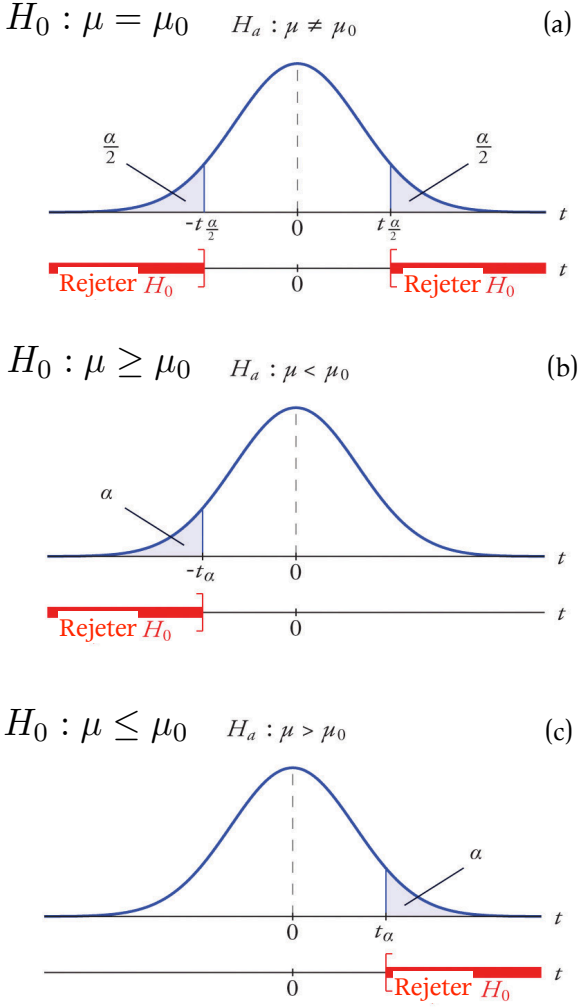

Figure 5.4: Tests d'hypothèse

▶ **Exemple 5.2** Continuons avec l'ensemble de données de l'exemple précédent. Nous cherchons à savoir si l'échantillon de taille $n = 528$ dont nous disposons appartient à une population de moyenne μ_0 au seuil de signification $p = 0,05$.

Il est noter que si le niveau de signification ou valeur de p est ≥ 0.05 , nous acceptons l'hypothèse nulle et nous pouvons conclure qu'il n'y a pas de différence significative entre les deux groupes (ou les mesures). Si la signification ou valeur de p est $< $ à 0.05 (c'est-à-dire que p est dans la zone grise de la figure 5.4, nous rejetons l'hypothèse nulle et nous pouvons conclure qu'il y a une différence significative entre les deux groupes.

Nous testerons deux valeurs de moyenne $\mu_0 = 50$ ans et $\mu_0 = 57$ ans. Les lignes de codes suivantes permettent de tester chacun de ces cas.

Cas ou $\mu_0 = 50$

```
library(prettyR)
MyData <- read.csv(file="satisfaction_hopital.csv", header=TRUE, sep="↩
    ,")
# Application du test de Student
t.test(MyData$age, mu=50)
```

```
data:  MyData$age
t = 10.596, df = 527, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 56.68956 59.73468
sample estimates:
mean of x
 58.21212
```

Le *Output* de l'exemple précédent permet de tirer les conclusions suivantes : La valeur p du test est $< 2.2e-16$, ce qui est inférieur au niveau de signification $\alpha = 0,05$. Nous pouvons, donc, conclure que la moyenne de l'âge des patients est significativement différente de 50 ans avec une valeur de $p < 2.2e-16$. Les bornes supérieure et inférieure de l'intervalle de confiance à 95% sont aussi indiquées. Ces valeurs confirment celles déterminées

précédemment selon la relation 5.2.

Cas ou $\mu_0 = 57$

```
library(prettyR)
MyData <- read.csv(file="satisfaction_hopital.csv", header=TRUE, sep=",")
# Application du test de Student
t.test(MyData$age, mu=57)
```

Contrairement au cas précédent, cet *Output* permet de dire que valeur p du test est supérieur au niveau de signification $\alpha = 0,05$. Nous pouvons conclure que la moyenne de l'age des patients n'est pas significativement différente de 57 ans avec une valeur de $p = 0.1184$.

Dans le cas d'un test d'hypothèse traitant une inégalité, nous devons spécifier dans la commande `t.test`, la paramètre `alternative = "less"` ou `alternative = "greater"` pour les cas de figures 5.4 (b) et (c)).

```
One Sample t-test
data:  MyData$age
t = 1.5639, df = 527, p-value = 0.1184
alternative hypothesis: true mean is not equal to 57
95 percent confidence interval:
 56.68956 59.73468
sample estimates:
mean of x
 58.21212
```

