Les tests statistiques

Introduction

- *Population et individus*

Une population est un ensemble d'objets sur lesquels une étude se porte. Ces objets sont appelés individus.

- *Caractère/variable*

Toute propriété étudiée chez les individus d'une population est appelée caractère.

Introduction

- *Nature d'un caractère*

Un caractère est dit :

- quantitatif s'il mesure une quantité ou un nombre (le nombre de personnes dans une salle, le salaire, temps de réalisation d'une travail en heures. . .),
- qualitatif/catégoriel s'il mesure une catégorie (la couleur des yeux d'une femme, la marque du téléphone portable d'un étudiant, la présence ou l'absence d'un défaut de fabrication dans l'emballage d'un produit. . .). Les valeurs sont appelées modalités.

- *Échantillon*

Un échantillon est un ensemble d'individus issus d'une population.

Introduction

- *Données*

Les données sont les observations de caractères sur les individus d'un échantillon.

- *Estimation paramétrique*

L'enjeu de l'estimation paramétrique est d'évaluer/estimer avec précision un ou plusieurs paramètres inconnus émanant de caractères à partir des données.

- *Moyenne et écart-type corrigé*

La moyenne et l'écart-type corrigé des données sont les principales mesures statistiques intervenant en estimation paramétrique

En notant X un caractère numérique (il peut être quantitatif, ou qualitatif avec un codage numérique), n le nombre d'individus d'un échantillon et x1,x2,…,xn les données associées, on définit :

La moyenne de x1,x2,…,xn :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

C'est une estimation ponctuelle de la valeur moyenne de X.

L'écart-type corrigé de x1,x2,…,xn :

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

C'est une estimation ponctuelle de la variabilité de X autour de sa moyenne. La valeur obtenue a la même unité que X.

Exemple

Population	Ensemble des pommes d'une ferme
Individu	Pomme
Caractère	Poids d'une pomme (en grammes)
Paramètre inconnu	Poids moyen d'une pomme
Échantillon	7 pommes choisies au hasard ($n = 7$)
Données	x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 162 | 155 | 148 | 171 | 151 | 165 | 154 (par exemple, x_1 est le poids de la première pomme de l'échantillon, soit 162 grammes)
Objectif	Évaluer le poids moyen inconnu d'une pomme à l'aide des données $x_1, \ldots, x_7$
Moyenne	$\overline{x} = \frac{1}{7}\sum_{i=1}^{7} x_i = 158$
Écart-type corrigé	$s = \sqrt{\frac{1}{7-1}\sum_{i=1}^{7}(x_i - \overline{x})^2} = 8.246211$

Bases des tests statistiques

- *Hypothèses*

On oppose deux hypothèses complémentaires : H0 et H1

L'hypothèse H0 formule ce que l'on souhaite rejeter/réfuter,

L'hypothèse H1 formule ce que l'on souhaite montrer.

Par exemple, si on veut montrer l'hypothèse « taille en moyenne non égales », H0 et H1 s'opposent sous la forme :

*H0: "*taille en moyenne égales *" contre H1:* "taille en moyenne non égales "

Bases des tests statistiques

- *Notion de risque*

Le risque (de première espèce) est le pourcentage de chances de rejeter H0, donc d'accepter H1 , alors que H0 est vraie.

On veut que ce risque soit aussi faible que possible.

Il s'écrit sous la forme : $100\alpha\%$, avec $\alpha \in]0; 1[$ (par exemple, 5%, soit $= 0:05$).

Le réel α est alors la probabilité de rejeter H0 alors que H0 est vraie.

Le rejet de H0 est dit "significatif" si elle est rejetée au risque 5%.

- *Test statistique*

Un test statistique est une procédure qui vise à apporter une réponse à la question :Est-ce que les données nous permettent de rejeter H0, donc d'accepter H1, avec un faible risque de se tromper ?

Bases des tests statistiques

- *Types de test statistique sur un paramètre :*

Lorsque le test statistique porte sur un paramètre inconnu θ , on dit que le test est

Bilatéral si: H1 est de la forme H1: $\theta \neq$ …

Unilatéral à gauche (sens de <) si: H1 est de la forme H1: $\theta <$ …

Unilatéral à droite (sens de >) si : H1 est de la forme H1: $\theta >$ …

p-valeur

La p-valeur est le plus petit réel $\alpha \in]0; 1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H0 alors que H0 est vraie.

Bases des tests statistiques

- *Degré de significativité*

La p-valeur nous donne un degré de significativité du rejet de H0.

Le rejet de H0 est dit :

- significatif si p-valeur $\alpha \in$]0,01; 0,05], symbolisé par *,
- très significatif si p-valeur $\alpha \in$]0,001; 0,01], symbolisé par **,
- hautement significatif si p-valeur < 0,001, symbolisé par ***

Il y a non rejet de H0 si p-valeur > 0,05.

Bases des tests statistiques

- *En cas de non-rejet de H0*

S’il y a non-rejet de H0, sauf convention, on ne peut rien conclure du tout (avec le risque considéré).

Éventuellement, on peut dire que H0 est plausible (elle "semble pouvoir être admise").

En revanche, peut-être qu’un risque de départ plus élevé ou la disposition de plus de données peuvent conduire à un rejet de H0

Tests paramétriques

- *Comparaison de deux moyennes observes*

Nous travaillons sur sur échantillon de 32 individus sur lesquels nous évaluerons l'effet du tabac sur la tension artérielle.

1. Quelle est la moyenne globale de la tension artérielle systolique?
2. Quelle est la variance globale de la tension artérielle systolique?
3. Affichez les graphiques (boxplot et histogramme) de tension artérielle systolique.

Tests paramétriques

- *Comparaison de deux moyennes observes*

Nous travaillons sur sur échantillon de 32 individus sur lesquels nous évaluerons l'effet du tabac sur la tension artérielle.

1. m= **140,8** mmHg
2. s^2= **252,9** $mmHg^2$
3. boxplot(data$TAS) hist(data$TAS)

Tests paramétriques

- *Comparaison de deux moyennes observes*

Nous travaillons sur sur échantillon de 32 individus sur lesquels nous évaluerons l'effet du tabac sur la tension artérielle.

1. Combien y'a-t-il de fumeurs et de non fumeurs?
2. Donnez une description de ces deux groupes (moyennes, variances et graphiques)

Tests paramétriques

Hypothèses

H0: $\mu_{NF} = \mu_F$ la TAS est en moyenne identique chez les fumeurs et les non fumeurs

H1: $\mu_{NF} \neq \mu_F$ la TAS moyenne est différente chez les fumeurs et les non fumeurs

Prédiction sous H0:

Sous H0 et si les conditions d'applications sont respectées alors

$$T = \frac{\left|\mu_{NF} - \mu_F\right|}{\sqrt{\sigma^2\left(\frac{1}{n_{NF}} + \frac{1}{n_F}\right)}} \rightarrow T_{n_{NF}+n_F-2}$$

Loi de Student

Tests paramétriques

Conditions d'applications

1. $X \rightarrow \mathrm{N}(\mu;\sigma^2)$ ou $n_{NF} > 30\, ET\ n_F > 30$
2. *Egalité des variances*
3. *Indépendance des individus*

Confrontation test de Student

t.test(TAS[Tabac==0], TAS[Tabac==1], var.equal=TRUE)

1. Hypothèses
2. Prédiction sous H0
3. Confrontation **test de Student**

t.test(TAS[Tabac==0], TAS[Tabac==1], var.equal=TRUE)

Test de Student

1. Hypothèses
2. Prédiction sous H0
3. Confrontation **test de Student**

t.test(TAS[Tabac==0], TAS[Tabac==1], var.equal=TRUE)

Tension artérielle chez les non fumeurs

1. Hypothèses
2. Prédiction sous H0
3. Confrontation **test de Student**

t.test(TAS[Tabac==0], TAS[Tabac==1], var.equal=TRUE)

Tension artérielle chez les fumeurs

1. Hypothèses
2. Prédiction sous H0
3. Confrontation **test de Student**

t.test(TAS[Tabac==0], TAS[Tabac==1], var.equal=TRUE)

Condition d'égalité des variance

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

```
  Two Sample t-test

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 0.0003113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.106070  -9.337067
sample estimates:
mean of x mean of y
130.8667  149.5882
```

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

Test de Student pour 2 échantillons indépendants

```
Two Sample t-test

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 0.0003113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.106070  -9.337067
sample estimates:
mean of x mean of y
130.8667  149.5882
```

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

```
  Two Sample t-test

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 0.0003113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.106070  -9.337067
sample estimates:
mean of x mean of y
130.8667  149.5882
```

données

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

```
  Two Sample t-test

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.106070  -9.337067
sample estimates:
mean of x mean of y
130.8667  149.5882
```

t_0 calculé sous H0

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

Two Sample t-test

data: TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 0.0003113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-28.106070 -9.337067
sample estimates:
mean of x mean of y
130.8667 149.5882

Petit « p »

1. Hypothèses
2. Prédiction sous H0
3. Confrontation

t.test(TAS[Tabac==0],TAS[Tabac==1], var.equal=TRUE)

```
  Two Sample t-test

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
t = -4.0742, df = 30, p-value = 0.0003113
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.106070  -9.337067
sample estimates:
mean of x mean of y
130.8667  149.5882
```

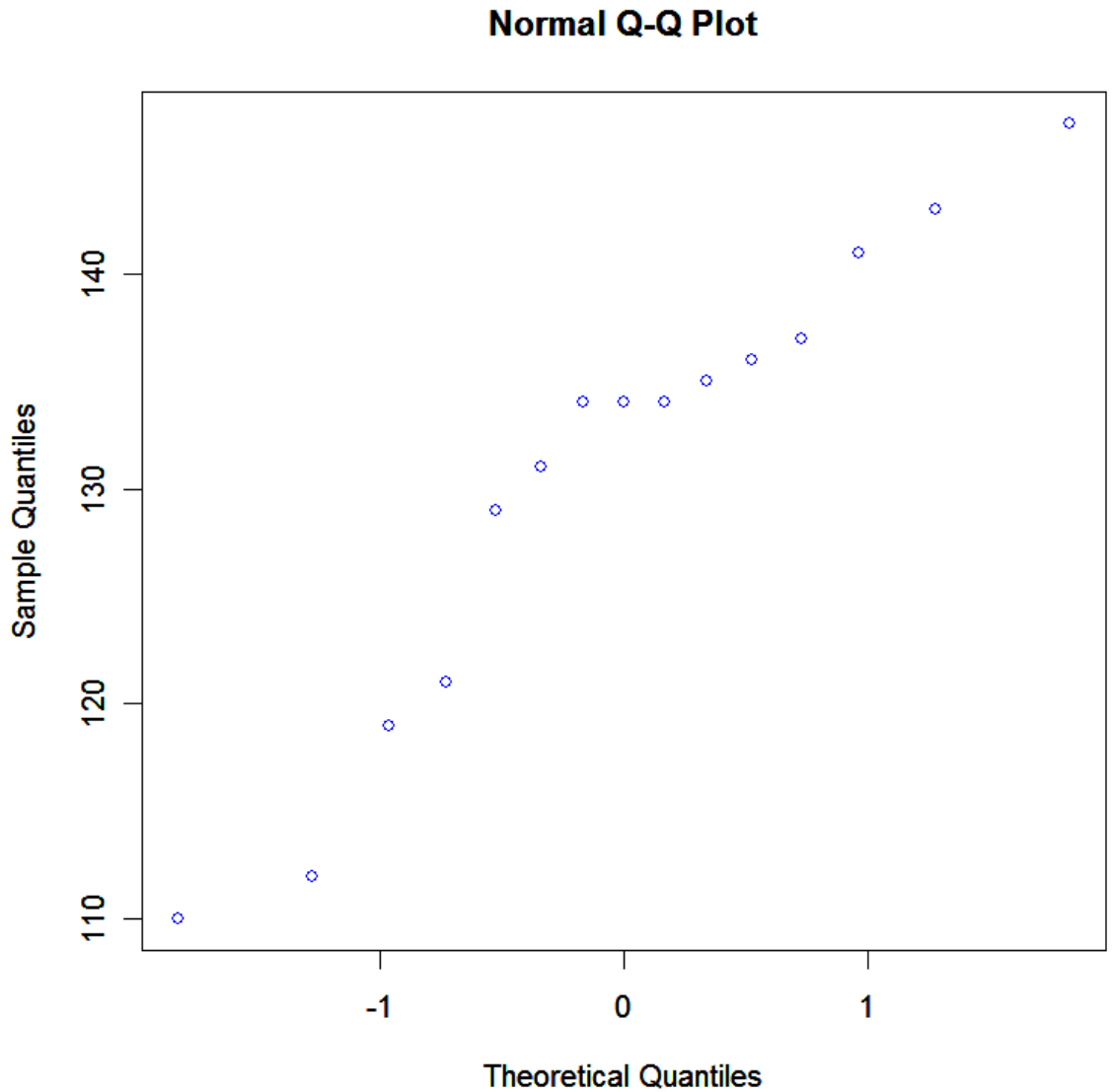
H1

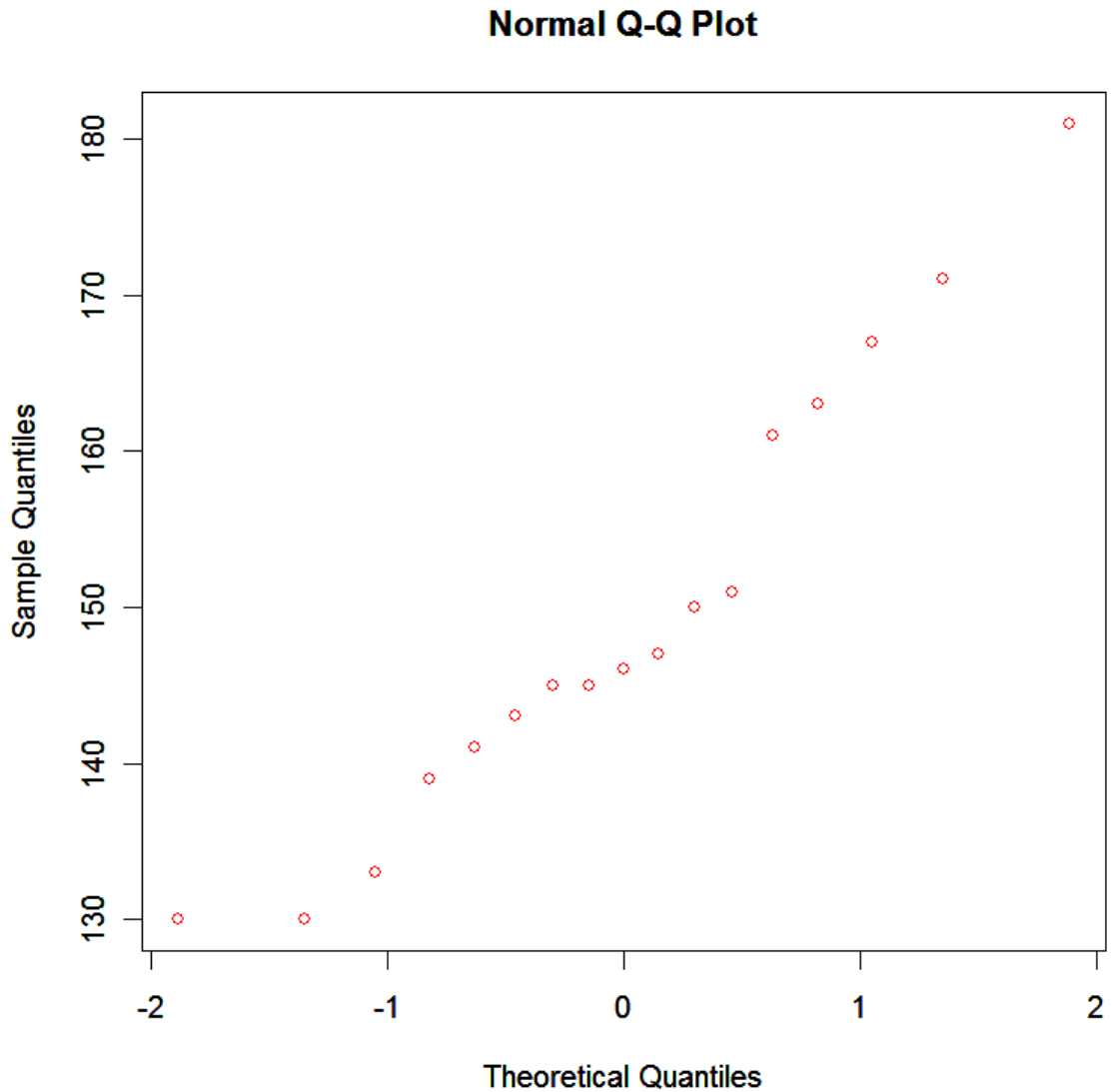
1. Hypothèses
2. Prédiction sous H0
3. Confrontation
4. Interprétation

- $p < 0,05$
- Test significatif
- On rejette H0, au risque α=5%
- Il y a une différence entre les 2 groupes
- Dans le sens **“les fumeurs ont une TAS moyenne plus élevée que les non-fumeurs”**

Vérification des Conditions d’application

- Normalité
  - Histogrammes
  - qq-plot: *qqnorm(TAS [Tabac==0], col="blue")*
  - Test non paramétrique


Normal Q-Q Plot
Sample Quantiles
110
120
130
140
-1
0
1
Theoretical Quantiles


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
130
140
150
160
170
180
-2
-1
0
1
2

Vérification des Conditions d'application

- Egalité des variances
  - Test de comparaison de 2 variances
    - H0: égalité
    - H1: différence

    var.test(TAS[Tabac==0], TAS[Tabac==1])

    ou var.test(TAS~Tabac)

var.test(TAS[Tabac==0], TAS[Tabac==1])

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

var.test(TAS[Tabac==0], TAS[Tabac==1])

Comparaison de 2 variances

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

var.test(TAS[Tabac==0], TAS[Tabac==1])

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

Données

var.test(TAS[Tabac==0], TAS[Tabac==1])

F test to compare two variances

data: TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, d value = 0.2773
alternative hypothesis: true s is not equal to 1
95 percent confidence interval:
0.1976701 1.6278629
sample estimates:
ratio of variances
0.5568401

F_0 calculé sous H0

var.test(TAS[Tabac==0], TAS[Tabac==1])

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

Petit « p »

var.test(TAS[Tabac==0], TAS[Tabac==1])

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

H1

var.test(TAS[Tabac==0], TAS[Tabac==1])

```
 F test to compare two variances

data:  TAS[Tabac == 0] and TAS[Tabac == 1]
F = 0.5568, num df = 14, denom df = 16, p-value = 0.2773
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1976701 1.6278629
sample estimates:
ratio of variances
         0.5568401
```

- p>0,05
- Test Non significatif
- Non rejet de H0 au risque β
- On ne met pas en évidence de différence significative entre les 2 variances

EXERCICE

En utilisant la même méthode y a-t-il une différence entre la moyenne de la TAS des hommes (code 1) et celle des femmes (code 0) ?

Références

Jean Bouyer: *Méthodes statistiques, Médecine-Biologie*, éditions INSERM

Pr Jean Gaudart ,Cours statistique univariée, Faculté de Médecine de Marseille

Christophe Chesneau. Introduction aux tests statistiques avec R