

TP6 : Test sur un échantillon

Objectifs : savoir utiliser les tests d'une moyenne et d'une probabilité sur un échantillon.

On rappelle la règle de décision pour n'importe quel test (moyenne, proportion, deux échantillons, etc).

Pour un test de niveau α ,

- si la p-valeur $< \alpha$, on rejette H_0
- si la p-valeur $> \alpha$, on ne rejette pas H_0

Exercice 1

La base de données HER (Health Exam Results) est issue d'un service américain de santé. Les variables collectées sont les suivantes

iden : identifiant de l'individu	dias : pression diastolique (mmHg)
sex : 0 pour les hommes, 1 pour les femmes	chol : cholestérol (mg)
age : en années	BMI : indice de masse corporelle (kg/m ²)
ht : taille (cm)	leg : longueur de la partie haute de la jambe (cm)
wt : poids (kg)	elbow : largeur du coude (cm)
waist : tour de taille (cm)	wrist : largeur du poignet (cm)
pulse : pouls (battements par minute)	arm : circonférence du bras (cm)
sys : pression systolique (mmHg)	treat : traitement (1 si traité, 0 sinon)

1. Charger le jeu de données `her.csv`. On s'intéresse à la variable BMI, appeler B cette variable. Quel est le type de la variable BMI ? Suggérer une loi pour modéliser la variable BMI.
2. Calculer les statistiques descriptives de la variable B. Appeler n le nombre de données de la variable B. Tracer sa distribution empirique.
3. Tester si la moyenne de BMI est différente de 25 pour un niveau de 5% :
`t.test(B, mu=25)`
Quelle décision prenez vous ? Quelle est la conclusion du test ?
4. Tester si la moyenne de BMI est plus grande que 25 pour un niveau de 5% :
`t.test(B, mu=25, alternative = "greater")`
Quelle décision prenez vous ? Quelle est la conclusion du test ?
5. Diriez vous que en moyenne
 - (a) les hommes ont un BMI au dessus de 23 ?
 - (b) les femmes ont un BMI en dessous de 22 ?
 - (c) que les patients traités ont un BMI au dessus de 25 ?
 - (d) que les patients non traités ont un BMI en dessous de 26 ?

Exercice 2

Pour une certaine maladie, un traitement a 70% de réussite. Un laboratoire propose un nouveau traitement et prétend qu'il est plus efficace que le traitement de référence. Pensez vous que ce nouveau traitement doit être autorisé si

1. Parmi 100 patients recevant le nouveau traitement, 74 ont été guéris. Vous commencerez par rentrer les valeurs de n , x et p_0 puis vous utiliserez la fonction `prop.test` pour tester si le taux de réussite du nouveau traitement est différent du traitement de référence :

```
prop.test(x,n, p0)
```

Utiliser l'option `correct=FALSE` pour appliquer l'approximation gaussienne du test. Est ce que le résultat change ?

2. Vous testerez si le nouveau traitement est plus efficace que le traitement de référence à l'aide d'un test unilatéral :

```
prop.test(x,n, p0, alternative="greater")
```

3. Parmi 200 patients recevant le nouveau traitement, 148 ont été guéris.
4. Parmi 1000 patients recevant le nouveau traitement, 740 ont été guéris.

Exercice 3

Dans la population américaine la proportion de personnes qui ont une valeur de cholestérol supérieure à 260 est de $p_0 = 55\%$. On veut savoir si la proportion p de personnes ayant une valeur de cholestérol supérieure à 260 dans la base `her.txt` est égale à cette proportion de référence p_0 .

1. Charger la base de données `her.txt`.
2. Donner les hypothèses du test que l'on va appliquer.
3. Extraire la variable `chol`.
4. Créer une variable binaire qui vaut 1 quand le cholestérol est inférieur à 260 :

```
cholSup260<-(chol>260)
```
5. Rentrer les valeurs de n , x et p_0 :

```
n <- length(chol)
x <- sum(cholSup260)
p0 <- 0.55
```
6. Calculer la p-valeur du test en utilisant la fonction de R `prop.test`. Quelle est la décision du test ?
7. Proposer un test unilatéral adapté et en calculer la p-valeur avec `prop.test`.

Exercice 4

1. On travaille sur le jeu de données `proteine.csv`. On étudie la variable `hydro`. Extraire les données de cette variable et les stocker dans `H`.
2. Calculer les statistiques descriptives de `H`. Tracer sa distribution empirique. Qu'observez vous ?
3. Tester si l'hydropathie moyenne est plus petite que $\mu_0 = -1$. Donner la p-valeur et commenter.
4. Appliquer ce test uniquement aux protéines issues de Viridiplantae.
5. Même question pour les Metazoa. Commenter.