TP8 : Introduction aux tests paramétriques

Objectifs : *Appliquer une règle de décision et savoir l'ajuster de sorte que le risque de première espèce α (celui de refuser à tort l'hypothèse nulle testée) ait une valeur spécifiée à l'avance. On dira que l'on met en oeuvre un test de niveau (ou seuil) α. Comprendre la p-valeur d'un test. Exercices sur données simulées et sur données réelles* `apnee.csv`.

1 Données simulées

Exercice 1 :

On utilisera dans cet exercice des données simulées sous une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Définir $\mu = 2$, $\sigma = 1$, $N = 1000$ et $n = 50$. Générer N tirages d'échantillons de taille n pour une variable de $\mathcal{N}(\mu, \sigma^2)$ et les affecter à `dataG`.
2. On souhaite effectuer le test statistique suivant pour une valeur μ_0 donnée :
$$\mathcal{H}_0 : \mu = \mu_0 \qquad\qquad \mathcal{H}_1 : \mu > \mu_0$$
3. On propose plusieurs tests possibles pour décider entre ces deux hypothèses à l'aide d'un échantillon de X.
   - Test 1 : refus de $\mathcal{H}_0$ lorsque $\bar{x}_n > \mu_0$
   - Test 2 : refus de $\mathcal{H}_0$ lorsque $\bar{x}_n > \mu_0 + 2/\sqrt{n}$
   - Test 3 : refus de $\mathcal{H}_0$ lorsque $\bar{x}_n > \mu_0 + 3/\sqrt{n}$

   Evaluer pour chacun des trois tests proposés la fréquence avec laquelle sera conclue l'hypothèse $\mathcal{H}_1$ sur les N échantillons tirés, lorsque l'on teste la valeur $\mu_0 = 2$. Que représentent ces fréquences ainsi calculées ?
4. On suppose $\sigma = 1$. Proposer un test pour lequel le risque d'accepter $\mathcal{H}_1$ à tort soit de valeur $\alpha = 5\%$ et vérifier que cette nouvelle règle de décision se trompe approximativement $0.05N = 50$ fois parmi les $N = 1000$ échantillons tirés.
5. Exprimer la région critique du test en utilisant la statistique de test centrée et normalisée.
6. Si on suppose à présent σ inconnu que devient le test précédent ?

2 Données réelles

Les données du fichier `apnee.csv` décrivent l'observation de 35 sujets souffrant d'apnée du sommeil (apnee=1) et de 65 sujets ne souffrant pas de ce problème (apnee=0). Pour chaque sujet ont été observés l'âge, le poids, la taille, le tabagisme (1 si fumeur), le sexe (0 pour les hommes) et le nombre de verres d'alcool consommés par jour. Ces données ont été collectées afin d'évaluer les effets potentiels du sexe, de la consommation d'alcool et de tabac ou du poids sur l'apnée du sommeil.

Exercice 2 : test pour μ

1. Charger les données `apnee.csv` avec la fonction `read.table()` et affecter le data.frame à `data`.
2. Extraire du data.frame `data` l'échantillon des mesures de la variable `taille` chez les hommes, avec la commande `data[data$sexe==0,"taille"]` et l'affecter à `tailleH`.
3. On supposera que la taille d'un homme suit une loi normale $\mathcal{N}(\mu, \sigma^2)$. Calculer l'estimation sans biais de la moyenne μ et de la variance σ^2.
4. Proposer un test sur μ d'égalité avec la valeur $\mu_0 = 178$ contre l'alternative $\mu_0 > 178$ au niveau α pour une collection de valeur de α : 0.01,0.02,...,0.1. Il y aura une décision pour chaque α proposé. Entre quelles valeurs de α observe-t-on un changement de décision ?
5. Calculer la valeur de ce niveau critique α^* appelé p-valeur du test.
6. Que retourne la commande `t.test(tailleH,mu=178,alternative="greater")` ?