Acquisition et analyse de données (BING-F4002): travaux pratiques

Marius Gilbert & Marc Dufrêne

Année académique 2015-2016

Séance IV. Rappel sur les tests d'hypothèse et le test de Student

- Chargez le jeu de données iris
- L'intervalle de confiance autour de la moyenne se calcule par x +/- t*s/sqrt(n), avec t* étant le percentile de 95% d'une distribution de Student à n-1 degrés de liberté. En utilisant les fonction mean(), sd() et qt() qui vous donne la valeur du quantile de la distribution t de Student pour un valeur p et un nombre de degrés de liberté donné, calculez i) la moyenne de la longueur de sépales (que vous appellerez SepMn), l'écart-type de cette moyenne (SepSd) et les bornes supérieures (SepMnUpperCI) et inférieures (SepMnLowerCI) de l'intervalle de confiance de cette moyenne. N'oubliez pas que vous pouvez atteindre l'aide de toute fonction à l'aide de la commande ? « nom de fonction ».
- Selon vous, cette moyenne diffère-t-elle significativement de 0, pourquoi ?
- Réalisez une boite de dispersion de la longueur de sépales en fonction de l'espèce. Les longueurs de sépale sont-elles semblables entre les espèces ?
- Construisez un jeux de donnée par espèce, à l'aide de la fonction subset(). Appellez ces jeux de données Ver, Set, et Vir pour les espèces I. versicolor, I. setosa et I. virginica.
- Calculez la moyenne, l'écart-type, et les bornes supérieures et inférieures de l'intervalle de confiance pour la longueur de sépales des individus de l'espèce I. versicolor, et de l'espèce I. setosa. Pour ce faire, vous pouvez utiliser les jeux de donnée créés au point 5) et réutiliser les fonctions que vous avez écrites au point 1).
- Selon vous, les moyennes de longueur de sépale de ces deux espèces peuvent-elles être égales ? Justifiez votre réponse sur base des valeurs estimées au point 6)
- La fonction t.test() permet de faire un test t de comparaison de moyenne entre deux séries d'observations. A l'aide des jeux de données créés au point 5) faites les tests suivants pour comparer les longueurs de sépales:

a. Comparaison de moyenne de I. versicolor et de I. setosa
b. Comparaison de moyenne de I. versicolor et de I. virginica

- Tapez les commandes suivantes :

```
myseq = seq(from = -15, to = 15,by = 0.1)
plot(myseq, dt(myseq, 86.538), type = "l", ylim = c(0,0.5),
     xlim = c(-15,15), ylab = "Density", xlab = "Quantile")
```

- Selon vous, qu'est-ce qui est représenté ici ? Positionnez les valeurs de vos statistiques t calculées en 8 a,b,c,d sur ce graphique. Ces valeurs sont-elles probables sous H0 ? Sur base de ce graphique, donnez 2-3 exemples de valeur de la statistique t qui n'aboutiraient pas à rejeter H0.
- A l'aide de la fonction 'power.t.test()' faite un calcul de puissance du test t, pour une différence de moyenne de 1, et un écart-type correpondant à la moyenne des écart-types des 3 espèces. La puissance du test vous semble-t-elle bonne ? Quelle serait cette puissance s'il n'y avait que 5 mesures par espèce ?