Programmation Statistique
TP8 - Test d'hypothèse et comparaison de distributions

Le test de Student

Avec le test de Student (t-test) nous pouvons vérifier plusieurs propriétés d'une distribution ou comparer deux distributions différentes. Par exemple:

- Comparaison de moyenne d'une loi normale à une valeur si la variance est inconnue.
- Comparaison de deux moyennes issues de deux lois normales si leurs variances sont égales et inconnues, ou si leurs variances sont différentes et inconnues (Test t de Welch).
- Test sur des échantillons appariés

L'utilisation typique est pour répondre à des questions du type:

- Est-ce que les salaires des hommes sont significativement plus élevés que ceux des femmes?
- Est-ce que ce médicament fonctionne mieux qu'un placebo?

La fonction pour calculer le test de Student en R est **t.test(...)**:

```
# Comparaison d'une moyenne observée à une moyenne théorique mu
t.test(x, mu=0)
# Test de student non-apparié
# Comparaison des moyennes de deux groupes (x et y)
t.test(x, y)
# Test de student apparié (la taille de x et y doit être la même car il
s'agit de deux observations sur la même population)
t.test(x, y, paired=TRUE)
```

Un paramètre important est le *niveau de confiance* **conf.level**: par défaut conf.level=0.95 ce qui veut dire qu'on est "95% sûrs" que le résultat est correct. On peut être plus strictes en mettant le paramètre à 0.995 ("99.5% sûrs"), ou réduire le niveau de confiance. Un niveau de confiance 0.5 signifie que le résultat n'est pas différent de prendre une décision aléatoire.

Le résultat d'un test de student contient habituellement trois valeurs:

- **t** est la statistique de student, **df** est le degré de liberté, **p-value** est le degré de significativité du test

Si p-value est inférieur à (1-conf.level), alors on peut dire que la différence entre les distributions est statistiquement significative à *conf.level*.

> Exemple: On compare les tailles des hommes et des femmes dans une population; si conf.level=0.95, on obtient un p-value=0.002, alors p.value < 0.05 (1-0.95), donc on peut affirmer au 95% que les tailles des hommes et des femmes sont différentes.

Exercice: On utilisera un nouveau jeu de données, qui contient des notes pour un ensemble d'étudiants. Vous pouvez le récupérer comme ça:

notes <- read.csv("https://lipn.univ-paris13.fr/~buscaldi/mathsv.tsv", sep='\t', dec = ',')

1. dessiner l'histogramme des notes (notes$note) pour les femmes (notes$sexe==”F”) et pour les hommes (notes$sexe==”M”) (2 histogrammes séparés).
2. dessiner la densité des notes avec la fonction density() vue en TP5, en bleu pour les hommes, en rouge pour les femmes (un graphique avec les deux, vous pouvez utiliser plot et lines)
3. comparez avec le test de Student les deux distributions, en utilisant d'abord un niveau de confidence 0.95
    a. Utilisez les données brutes
    b. Utilisez les données interpolées avec density() (si d <- density(...), vous pouvez les récupérer avec d$y)

   Est-ce qu'on peut affirmer que les notes des femmes sont significativement meilleures que celles des hommes? Justifier.
4. Si on modifie le niveau de confiance à 0.995 est-ce que le résultat est différent?

Exercice: retour sur le jeu de données des iris: nous allons vérifier nos résultats et nos analyses qualitatives du TP précédent

5. Utilisez les données des iris. Calculez, avec le test t, si, pour les 3 types de iris, la taille (Length) des pétales est assez importante pour distinguer les 3 types entre eux (3 comparaisons à faire)
6. Même chose avec la taille des sépales