

---

## TP10 : Comparaisons d'échantillons : tests sur les paramètres.

---

**Objectifs :** *Savoir faire des tests de comparaison des paramètres de deux échantillons. Exercices sur données réelles du fichier **her.csv***

Les exercices de ce TP portent sur les données contenues dans les fichiers **her.csv** ou **her.txt**. Ce jeu de données, HER (Health Exam Results), provient de l' US department of Health and Human Services, National Center for Health Statistics, et est extrait de la troisième enquête nationale sur la santé et l'alimentation aux USA. Les variables collectées sont les suivantes :

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

Pour charger les données on utilise `read.table()` et on nommera **data** le tableau des données :

```
data<-read.table("her.csv",header = TRUE, sep = " ")# avec le format csv
data<-read.table("her.txt",header = TRUE, sep = "\t")# avec le format txt
```

### Exercice 1:

1. Créer les vecteurs **sys** et **dia** contenant respectivement les échantillons de pression systolique et pression diastolique. Ces échantillons sont-ils appariés ? Si oui créer l'échantillon observé des différences entre les deux pressions et le nommer **diff**.
2. Peut-on supposer que la différence des deux pressions suit une loi normale ?
3. Donner estimation et intervalle de confiance de niveau 95% pour la moyenne de la différence entre les deux pressions.
4. On se demande si la pression systolique moyenne est la même que la pression diastolique moyenne. Faire un test de comparaison des deux échantillons permettant de répondre. On calculera la statistique de test et la p-valeur.
5. Retrouver le résultat précédent en utilisant `t.test()` soit sur l'échantillon des différences **diff**, soit sur les 2 échantillons des pressions **sys** et **dia** en précisant à l'aide de l'option **paired** si ils sont appariés ou non.
6. Montrer par un test statistique que la pression systolique moyenne est supérieure d'au moins 40 à la pression diastolique moyenne. La conclusion du test est-elle très significative ?

### Exercice 2:

On souhaite à présent comparer la pression systolique chez les hommes et chez les femmes afin d'apprécier si le sexe a un effet sur cette pression.

1. Créer les vecteurs `sysF` et `sysH` contenant respectivement les échantillons de pression systolique chez les femmes et chez les hommes. Ces échantillons sont-ils appariés ? On souhaite comparer les moyennes des pressions systoliques selon le sexe. Proposer un graphique permettant de répondre à cette question.
2. Afin de valider cette conjecture suggérée par le précédent graphique on se propose d'effectuer un test de comparaison de moyennes. Quelles hypothèses doit-on poser sur les variables pression systolique chez les hommes et pression systolique chez les femmes pour effectuer ce test ? Quelles sont les hypothèses  $\mathcal{H}_0$  et  $\mathcal{H}_1$  du test ?
3. Vérifier que le modèle posé sur ces deux variables est raisonnable ? Si cela ne semble pas le cas que conviendrait-il de faire ici ?
4. Extraire des deux échantillons les valeurs extrêmes signalées par les `boxplot` les affecter aux échantillons réduits `sysFR` et `sysHR`. La normalité est-elle mieux satisfaite sur les échantillons où ont été écartés les individus très extrêmes ?
5. Avant de comparer les moyennes, faire un test d'égalité des variances sur les échantillons complets (avec les trois individus extrêmes). Calculer la statistique de test et la p-valeur du test bilatéral. Qu'en conclure ?
6. Retrouver le précédent résultat avec la fonction `var.test`. Refaire le même test sur les échantillons réduits. Interpréter.
7. Pour finir proposer un test sur les moyennes permettant de montrer que la pression systolique moyenne des hommes est significativement plus élevée que celle des femmes.
8. La conclusion du test sur les moyennes est-elle très différente selon que l'on utilise les échantillons complets où ceux dans lesquels on a écartés les trois individus extrêmes ?

### Exercice 3

On s'intéresse à présent à la probabilité d'avoir une valeur de cholesterol supérieure à 350 selon que l'on est un homme,  $p_H$  ou une femme  $p_F$ .

1. Faire un graphique présentant les boxplots des échantillons de cholestérol sur les deux sous populations afin de les comparer.
2. Proposer un test permettant de juger de l'effet du sexe sur la probabilité d'avoir une valeur de cholesterol supérieure à 350.
3. Mettre en oeuvre le test en utilisant la procédure `prop.test`.