Sorbonne Université Année 2018/2019

1ère année ISUP Deuxième semestre

Feuille de TD 4 : Tests

Exercice 1 : Dans les années 70, les athlètes féminines de RDA étaient réputées pour leur forte corpulence et soupçonnées par le comité éthique olympique de dopages via la prise de substances hormonales virilisantes (dites androgènes). Des mesures ont été effectuées sur la quantité d'androgènes par litre de sang chez 9 athlètes, et on obtient les résultats suivants :

3.22 3.07 3.17 2.91 3.40 3.58 3.23 3.11 3.62

On veut tester l'hypothèse nulle "les athlètes de RDA ne sont pas dopées", sachant que chez une femme "lambda", le quantité moyenne d'androgènes est de 3.1

1. Quel test faut-il effectuer ?
2. Quels sont les hypothèses à vérifier ?
3. Commenter l'histogramme suivant :

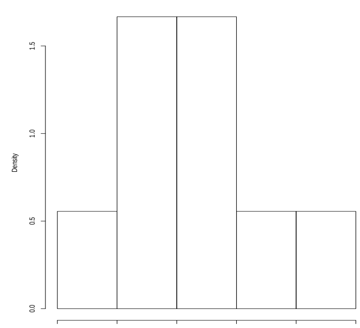

Fig. 1: Histogramme des mesures d'androgènes par litre de sang

4. Faire le test au risque de 5%. Pour s'aider, on pourra utiliser la sortie R suivante :

```
data:  c(3.22, 3.07, 3.17, 2.91, 3.4, 3.58, 3.23, 3.11, 3.62)
t = 1.9968, df = 8, p-value = 0.04046
alternative hypothesis: true mean is greater than 3.1
95 percent confidence interval:
 3.110771      Inf
```

```
sample estimates:
mean of x
 3.256667
```

5. Pourquoi peut-on remettre en question le protocole expérimental et donc la conclusion du test ?

Exercice 2 : Soient $X_1, X_2, \dots, X_p$ des variables aléatoires indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et soient $Y_1, Y_2, \dots, Y_q$ des variables aléatoires indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$. On suppose que les deux échantillons sont indépendants et de même variance, c'est à dire que $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

On s'intéresse à l'estimation de la différence $\mu_1 - \mu_2$.

1. Proposer un estimateur de $\mu_1 - \mu_2$. On le notera D.
2. Etablir la loi de cet estimateur.
3. Proposer un estimateur de σ^2. On le notera S^2 et démontrer que $\frac{(p+q-2)S^2}{\sigma^2}$ suit une loi du Khi-deux à $(p+q-2)$ ddl et S^2 indépendant de $\overline{X}_p$ et $\overline{Y}_q$.
4. Etablir alors que
$$\frac{D - (\mu_1 - \mu_2)}{S\sqrt{\dfrac{1}{p} + \dfrac{1}{q}}} \sim T_{p+q-2}$$
5. Construire un intervalle de confiance pour la différence $(\mu_1 - \mu_2)$ au niveau de confiance $(1-\alpha)$ avec $\alpha \in]0, 1[$.
6. Soient $x_1, \dots, x_{15}$ des réalisations de $X_1, \dots, X_{15}$ et $y_1, \dots, y_9$ des réalisations de $Y_1, \dots, Y_9$. Tester au risque de 1% l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$" sachant que :
$$\sum_{i=1}^{15} x_i = 16.2 \qquad \sum_{i=1}^{15} x_i^2 = 28.7$$
$$\sum_{i=1}^{9} y_i = 8.9 \qquad \sum_{i=1}^{9} y_i^2 = 31.5.$$
On pourra également s'aider des quantiles suivants :
$$\begin{array}{lll} \mathbb{P}[T_{22} \leq 2.074] = 0.975 & \mathbb{P}[T_{23} \leq 2.069] = 0.975 & \mathbb{P}[T_{24} \leq 2.064] = 0.975 \\ \mathbb{P}[T_{22} \leq 2.508] = 0.99 & \mathbb{P}[T_{23} \leq 2.500] = 0.99 & \mathbb{P}[T_{24} \leq 2.492] = 0.99 \\ \mathbb{P}[T_{22} \leq 2.819] = 0.995 & \mathbb{P}[T_{23} \leq 2.807] = 0.995 & \mathbb{P}[T_{24} \leq 2.797] = 0.995 \end{array}$$

Exercice 3 : Lors d'une petite expérimentation sur des souris atteintes d'une maladie mortelle, on a tiré au sort parmi 16 souris, 7 qui reçoivent un nouveau traitement alors que les 9 autres sont des contrôles qui reçoivent un placebo. Leurs durées de survie sont mesurées en jours et donnent les résultats suivants :

Survie (en jours)	
Groupe 1 (Placebo)	Groupe 2 (Traitement)
$n_1 = 9$ mesures 52, 10, 40, 104, 50, 27, 146, 31, 46	$n_2 = 7$ mesures 94, 38, 23, 197, 99, 16, 141
$\sum_{j=1}^{n_1} x_{j,1} = 506$	$\sum_{j=1}^{n_2} x_{j,2} = 608$
$\sum_{j=1}^{n_1} x_{j,1}^2 = 42842$	$\sum_{j=1}^{n_2} x_{j,2}^2 = 79556$

On supposera que les données du groupe 1 sont des réalisations indépendantes d'une variable aléatoire X_1 de loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ et que les données du groupe 2 sont des réalisations indépendantes d'une variable aléatoire X_2 de loi normale $\mathcal{N}(\mu_2, \sigma_2^2)$.

1. Calculer la moyenne des durées de survie des souris des groupe 1 et 2 On les notera m_1 et m_2 respectivement. Commenter les résultats obtenus. En particulier, que peut-on dire sur l'effet du traitement sur la durée de survie ?
2. Construire des intervalles de confiance au niveau de confiance 95% pour les moyennes réelles μ_1 et μ_2 et calculer leurs réalisations. On pourra s'aider des résultats suivants :

   $$\mathbb{P}[T_6 \leq 1.943] = 0.95 \quad \mathbb{P}[T_7 \leq 1.895] = 0.95 \quad \mathbb{P}[T_8 \leq 1.859] = 0.95 \quad \mathbb{P}[T_9 \leq 1.833] = 0.95$$
   $$\mathbb{P}[T_6 \leq 2.447] = 0.975 \quad \mathbb{P}[T_7 \leq 2.365] = 0.975 \quad \mathbb{P}[T_8 \leq 2.306] = 0.975 \quad \mathbb{P}[T_9 \leq 2.262] = 0.975$$

   Commenter les résultats obtenus. En particulier, que peut-on dire sur l'effet du traitement sur la durée de survie ?
3. On suppose que $\sigma_1^2 = \sigma_2^2$. Tester au risque de 1% l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$". On pourra s'aider de la sortie R suivante :

```
Two Sample t-test


data:  c1 and c2
t = -1.163, df = 9.909, p-value = 0.2721
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -99.12612  31.18961
sample estimates:
mean of x mean of y
 52.88889  86.85714
```

4. Quelles conclusions peut-on tirer de cette expérience ? Vous pourrez utiliser le graphique des boîtes à moustaches des durées de survie des 2 groupes pour argumenter.

.

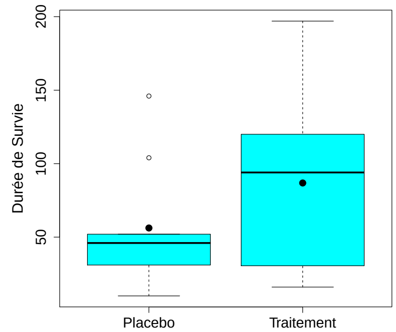

Fig. 2: Boîtes à moustaches des durées de survie par groupe.

Exercice 4 : On veut tester la précision d'une balance en effectuant une série de 15 mesures du poids d'un kilo de riz. On obtient les mesures suivantes :

Poids (en g)
$n = 15$ mesures 996.17, 994.45, 998.78, 997.2, 1007.01, 998.45, 1003.93, 995.23, 997.01, 999.36, 997.64, 993.81, 1004.33, 991.38, 1000.97
$\sum_{j=1}^{n} x_j = 14975.72$
$\sum_{j=1}^{n} x_j^2 = 14951732$

On supposera que les données sont des réalisations indépendantes d'une variable aléatoire X de loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Commenter l'histogramme suivant :

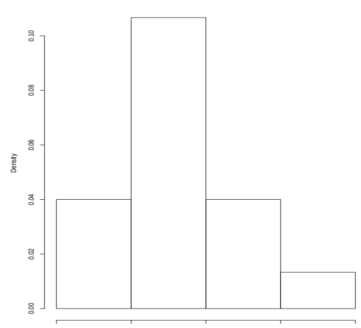

Fig. 3: Histogramme des mesures

2. Calculer le poids mesuré moyen que l'on notera m.

3. Donner une estimation de σ^2. On la notera s^2.
4. Construire un intervalle de confiance à 90% pour la moyenne et calculer sa réalisation. On pourra s'aider des quantiles suivants :

$$\mathbb{P}\left[T_{14} \leq 1.761\right] = 0.95 \quad \mathbb{P}\left[T_{14} \leq 2.977\right] = 0.995 \quad \mathbb{P}\left[T_{15} \leq 1.763\right] = 0.95 \quad \mathbb{P}\left[T_{15} \leq 2.947\right] = 0.995$$

5. Construire un intervalle de confiance à 95% pour la variance et calculer sa réalisation. On pourra s'aider des quantiles suivants :

$$\mathbb{P}\left[K_{14} \leq 26.1\right] = 0.975 \quad \mathbb{P}\left[K_{14} \leq 5.63\right] = 0.025 \quad \mathbb{P}\left[K_{15} \leq 27.49\right] = 0.975 \quad \mathbb{P}\left[K_{15} \leq 6.26\right] = 0.025$$

   où K_p suit une chi deux à p degrés de liberté.
6. Tester au risque de 1% la précision de la balance. On pourra s'aider de la sortie R suivante :

```
One Sample t-test

data:  c
t = -1.4763, df = 14, p-value = 0.162
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
  996.0297 1000.7330
sample estimates:
mean of x
 998.3813
```

.

Exercice 5 : On s'intéresse au salaire journalier des employés d'une entreprise. On obtient les salaires suivants :

Salaires (en euro)
$n = 16$ mesures 41, 40, 50, 45, 41, 41, 40,43, 45, 52, 40, 48, 50, 40, 47, 46
$\sum_{j=1}^{n} x_j = 709$
$\sum_{j=1}^{n} x_j^2 = 31675$

On supposera que les données sont des réalisations indépendantes d'une variable aléatoire X de loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. Calculer le salaire moyen mesuré que l'on notera m.
2. Donner une estimation de σ^2. On la notera s^2.

3. L'entreprise prétend payer en moyenne ses salariés plus de 47 euros par jour. Au risque de 5%, pouvez vous confirmer cette affirmation ? Au risque de 1% ? On pourra s'aider de la sortie R suivante.

```
One Sample t-test


data:  c
t = -2.5949, df = 15, p-value = 0.01015
alternative hypothesis: true mean is less than 47
95 percent confidence interval:
     -Inf 46.12812
sample estimates:
mean of x
  44.3125
```

4. L'entreprise prétend également avoir très peu de différences de salaires au sein de l'entreprise, i.e avoir une variance des salaires $\sigma_0^2 = 5$. On pourra s'aider des résultats suivants :

$$\mathbb{P}\left[K_{15} \leq 6.3\right] \leq 0.025 \quad \mathbb{P}\left[K_{15} \leq 27.5\right] \leq 0.975 \quad \mathbb{P}\left[K_{16} \leq 6.9\right] \leq 0.025 \quad \mathbb{P}\left[K_{16} \leq 28.8\right] \leq 0.975$$

où K_p suit une chi deux à p degrés de liberté.

Exercice 6 : On souhaite comparer les longueurs des mâchoires inférieures de 10 chacals mâles et 10 chacals femelles. On a les mesures suivantes :

Longueur (en mm)	
Groupe 1 (Mâles)	Groupe 2 (Femelles)
$n_1 = 10$ mesures 120, 107, 110, 116, 114, 111, 113, 117, 114, 112	$n_2 = 10$ mesures 110, 111, 107, 108, 110, 105, 107, 106, 111, 111
$\sum_{j=1}^{n_1} x_{j,1} = 1134$	$\sum_{j=1}^{n_2} x_{j,2} = 1086$
$\sum_{j=1}^{n_1} x_{j,1}^2 = 128720$	$\sum_{j=1}^{n_2} x_{j,2}^2 = 117986$

On supposera que les données du groupe 1 sont des réalisations indépendantes d'une variable aléatoire X_1 de loi normale $\mathcal{N}(\mu_1, \sigma^2)$ et que les données du groupe 2 sont des réalisations indépendantes d'une variable aléatoire X_2 de loi normale $\mathcal{N}(\mu_2, \sigma^2)$.

1. Commenter la figure suivante.

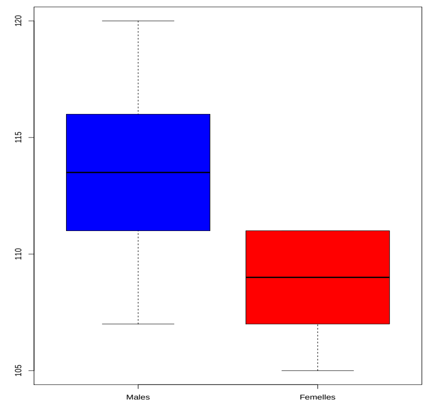

Fig. 4: Boxplots des longueurs des mâchoires chez les chacals mâles (à gauche) et femelles (à droite)

2. Calculer la moyenne des longueurs des mâchoires des groupe 1 et 2. On les notera respectivement m_1 et m_2.
3. Donner une estimation de σ^2. On la notera s^2.
4. Tester au risque de 5% le fait que le sexe des individus n'a pas d'incidence sur la longueur moyenne de leur mâchoire. On pourra s'aider de la sortie R suivante :

```
Welch Two Sample t-test

data:  males and femelles
t = 3.4843, df = 14.894, p-value = 0.00336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.861895 7.738105
sample estimates:
mean of x mean of y
    113.4     108.6
```

5. On suppose maintenant que les variables aléatoires X_1 et X_2 sont de variance σ_1^2 et σ_2^2. Tester au risque de 5% l'égalité de ces variances.

```
F test to compare two variances

data:  males and femelles
F = 2.681, num df = 9, denom df = 9, p-value = 0.1579
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.665931 10.793829
```

```
sample estimates:
ratio of variances
         2.681034
```

Exercice 7 : On considère un groupe de 28 individus souffrant d'un même handicap. Les individus ont été répartis en deux groupe, suivant deux apprentissages différents. Le premier consiste en de l'imitation (les sujets doivent imiter les gestes faits), le second consiste en la guidance (les sujets sont aidés physiquement pour effectuer les gestes). Le tableau ci-dessous donne les scores obtenus par les différents individus.

Scores	
Groupe 1 (Imitation)	Groupe 2 (Guidance)
$n_1 = 15$ mesures 19, 16, 24, 13, 9, 14, 17, 10, 19, 22, 23, 5, 7, 13, 11	$n_2 = 13$ mesures 15, 18, 23, 10, 8, 11, 12, 14, 21, 15, 18, 6, 7
$\sum_{j=1}^{n_1} x_{j,1} = 222$	$\sum_{j=1}^{n_2} x_{j,2} = 178$
$\sum_{j=1}^{n_1} x_{j,1}^2 = 3766$	$\sum_{j=1}^{n_2} x_{j,2}^2 = 2778$

On supposera que les données du groupe 1 sont des réalisations indépendantes d'une variable aléatoire X_1 de loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ et que les données du groupe 2 sont des réalisations indépendantes d'une variable aléatoire X_2 de loi normale $\mathcal{N}(\mu_2, \sigma_2^2)$.

1. Le test de Shapiro effectué sur les données des groupes 1 et 2 donnent des p-values respectivement égales à 0.8433 et 0.8401. Que pouvez vous en déduire ?
2. Commenter le graphique suivant :

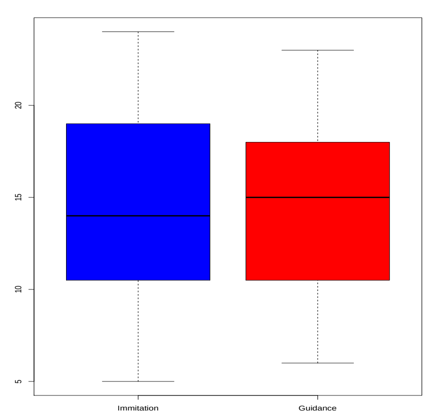

Fig. 5: Boxplots des scores pour la l'imitation (à gauche) et la guidance (à droite)

3. Tester au risque de 5% le fait que les variabilités des scores dans chacun des groupes ne sont pas différentes. On pourra s'aider de la sortie R suivante :

```
F test to compare two variances

data:  imitation and guidance
F = 1.2084, num df = 14, denom df = 12, p-value = 0.7503
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3768805 3.6856808
sample estimates:
ratio of variances
          1.208359
```

4. Tester au risque de 5% le fait que la méthode choisie n'impacte pas le score moyen. On pourra s'aider de la sortie R suivante :

```
Welch Two Sample t-test

data:  imitation and guidance
t = 0.5238, df = 25.924, p-value = 0.6049
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.239792  5.455177
sample estimates:
mean of x mean of y
 14.80000  13.69231
```

Exercice 8 : On étudie l'influence du magnésium sur la croissance d'une moisissure. On procède à deux expériences distinctes :
`Expérience 1`: On cultive la moisissure dans 20 boîtes, on injecte une dose de $5mg$ dans 10 boîtes et une dose de $10mg$ dans les 10 autres. La croissance moyenne dans les 10 premières boîtes est de $1.03\mu m$ avec une variance de $0.05\mu m^2$, la croissance moyenne dans les 10 dernières boîtes est de $1.12\mu m$ avec une variance de $0.1\mu m^2$.
`Expérience 2`: On cultive les moisissures dans 10 boîtes, puis on sépare chaque boîte en deux. Dans une partie, on injecte $5mg$ de magnésium, dans l'autre $10mg$ de magnésium. La différence moyenne est égale à $0.08\mu m$ et la variance de la différence est égale à $0.02\mu m^2$.

1. Expliquer la différence entre les deux expériences.
2. Pour chaque expérience, après avoir rappelé le cadre théorique, mettre en oeuvre un test au risque de 5% pour l'égalité des croissances moyennes. On pourra s'aider des résultats

suivants :

$$\mathbb{P}\left[T_9 \leq 2.262\right] = 0.975 \quad \mathbb{P}\left[T_9 \leq 3.250\right] = 0.995 \quad \mathbb{P}\left[T_{18} \leq 2.101\right] = 0.975 \quad \mathbb{P}\left[T_{18} \leq 2.878\right] = 0.995$$

3. Les deux tests mènent-ils à la même conclusion ?

Exercice 9 : Une entreprise a mis au point un nouveau traitement contre le phylloxera, puceron qui ravage les vignes. Il est testé sur une parcelle de 600 plants sur lesquels on observe les résultats suivants :

Effet	Eradication	Amélioration	Sans effet
Nombre de plants	280	210	110

Les résultats promis par l'entreprise sont de 60% d'éradication, 30% d'amélioration et 10% sans effet.

1. Tester au risque de 1% la véracité des dires de l'entreprise. On pourra s'aider des résultats suivants :

$$\mathbb{P}\left[\chi_2^2 \leq 9.2\right] = 0.99 \quad \mathbb{P}\left[\chi_2^2 \leq 10.6\right] = 0.995 \quad \mathbb{P}\left[\chi_3^2 \leq 11.3\right] = 0.99 \quad \mathbb{P}\left[\chi_3^2 \leq 12.8\right] = 0.995$$

2. On traite une deuxième parcelle avec le traitement habituel. Les résultats observé sur 400 plants sont les suivants :

Effet	Eradication	Amélioration	Sans effet
Nombre de plants	220	90	90

   (a) Proposer un test qui permet de rejeter ou non l'hypothèse "le nouveau traitement est différent de l'ancien".

   (b) Tester au risque de 5% si les traitements sont différents. On pourra s'aider des résultats suivants :

$$\mathbb{P}\left[\chi_2^2 \leq 6.0\right] = 0.95 \quad \mathbb{P}\left[\chi_2^2 \leq 7.4\right] = 0.975 \quad \mathbb{P}\left[\chi_3^2 \leq 7.8\right] = 0.95 \quad \mathbb{P}\left[\chi_3^2 \leq 9.3\right] = 0.975$$

Exercice 10 : On souhaite savoir si le rhésus dépend du groupe sanguin. Pour cela, on dispose du tableau de données suivant :

	O	A	B	AB	Total
Rhésus +	370	381	62	28	
Rhésus -	70	72	12	5	
Total					

Au risque de 5%, tester si le rhésus est indépendant du groupe sanguin. On pourra s'aider des résultats suivants :

$$\mathbb{P}\left[\chi^2_3 \leq 7.8\right] = 0.95 \quad \mathbb{P}\left[\chi^2_8 \leq 15.5\right] = 0.95$$

Exercice 11 : Le couvert végétal du domaine vital d'un orignal (élan d'amérique) se compose de peuplements feuillus (25% de la superficie du domaine vital), de peuplements mixtes (38% de la superficie), de peuplements résineux (25.8%) et d'un marécage (10.4%). Dans ce domaine, l'orignal a été localisé à 511 reprises au cours de l'année. Sur les 511 localisations, 118 se trouvaient dans le feuillus, 201 dans les peuplements mixtes, 110 dans les résineux, et 83 dans les marécages.

1. On veut montrer que l'orignal fréquente préférentiellement certains milieux. Proposer un test statistique pour vérifier cette hypothèse.
2. Réaliser le test et proposer une interprétation biologique du résultat. On pourra s'aider des résultats suivants :

$$\mathbb{P}\left[\chi^2_3 \leq 7.8\right] = 0.95 \quad \mathbb{P}\left[\chi^2_4 \leq 9.5\right] = 0.95$$

Exercice 12 : Dans une étude sur un répulsif de moustique, on compte le nombre de piqûres de chaque personne à partir d'un échantillon de 150 personnes. On obtient

Nb de piqûres	0	1	2	3	4	5	6	>6
Nb d'individus	32	54	34	21	6	2	1	0

Tester au risque de 5% que le nombre de piqûres pour une personne est une variable aléatoire suivant une loi de Poisson de paramètre 1. Pour s'aider, soit $X \sim \mathcal{P}(1)$, on a

k	0	1	2	3	4	5	6	>6
$\mathbb{P}[X = k]$	0.37	0.37	0.18	0.061	0.015	0.0031	0.00051	8.10^-5

On dispose également des résultats suivants :

$$\mathbb{P}\left[\chi^2_7 \leq 14.0\right] = 0.95 \quad \mathbb{P}\left[\chi^2_8 \leq 15.5\right] = 0.95$$