

Correction de la feuille de TD 5 : 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 ?

Un test d'inégalité de moyenne $H_0 : \mu \leq 3.1$.

2. Quels sont les hypothèses à vérifier ? Il vaut vérifier que ce sont les réalisations de variables aléatoires X_i indépendantes et identiquement distribuées et $X_1 \sim \mathcal{N}(\mu, \sigma^2)$.

3. Commenter l'histogramme suivant :

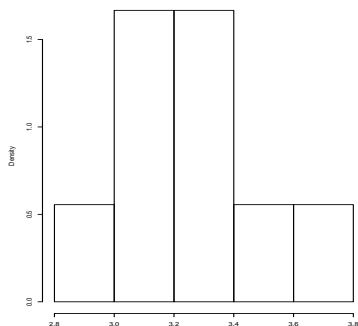


Fig. 1: Histogramme des mesures d'androgènes par litre de sang

Au vu du protocole, on peut penser que les réalisations sont bien indépendantes. Au vu de l'histogramme, on peut penser que ce sont biens des réalisations de variables aléatoires gaussiennes.

4. Faire le teste 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

```

On teste au risque de 5% :

$$H_0 : " \mu \leq 3.1 " \quad \text{contre} \quad H_1 : " \mu > 3.1 "$$

Sous H_0 , il existe $\mu' \leq 3.1$ tel que

$$Z(\mu') = \sqrt{n} \frac{\bar{X}_n - \mu'}{S_n} \sim T_{n-1},$$

et on obtient donc la zone de rejet

$$ZR = \{ Z(3.1) > t_{n-1, 1-\alpha} \}$$

où $t_{n-1, 1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi de Student à $n - 1$ degrés de liberté. Ici, la p -value est inférieure à 0.05 et on rejette donc H_0 .

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

Il est stupide de comparer la moyenne d'androgènes d'athlètes avec celles de femmes "lambda". Ce test ne répond donc pas à la question : "est-ce que les athlètes allemandes sont dopées".

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 . Comme \bar{X}_p et \bar{Y}_q sont des estimateurs naturels de μ_1, μ_2 , on va considérer l'estimateur

$$D = \bar{X}_p - \bar{Y}_q.$$

2. Etablir la loi de cet estimateur. Comme D est une combinaison linéaire de variables aléatoires suivant des lois normales et indépendantes, D suit une loi normale et on a par

linéarité de l'espérance

$$\mathbb{E}[D] = \mathbb{E}[\bar{X}_p] - \mathbb{E}[\bar{Y}_q] = \mu_1 - \mu_2.$$

De la même façon, comme \bar{X}_p et \bar{Y}_q sont indépendantes,

$$\mathbb{V}[D] = \mathbb{V}[\bar{X}_p] + \mathbb{V}[\bar{Y}_q] = \frac{\sigma^2}{p} + \frac{\sigma^2}{q}.$$

On a donc

$$D \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{p} + \frac{1}{q}\right)\right).$$

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. On a deux estimateurs naturels de σ^2 qui sont

$$S_X^2 = \frac{1}{p-1} \sum_{i=1}^p (X_i - \bar{X}_p)^2 \quad \text{et} \quad S_Y^2 = \frac{1}{q-1} \sum_{i=1}^q (Y_i - \bar{Y}_q)^2.$$

De plus on a

$$\frac{(p-1)S_X^2}{\sigma^2} \sim \chi_{p-1}^2 \quad \text{et} \quad \frac{(q-1)S_Y^2}{\sigma^2} \sim \chi_{q-1}^2.$$

De plus, par indépendances des échantillons, S_X^2 et S_Y^2 sont indépendants. Par le théorème de Cochran, on obtient alors

$$\frac{(p-1)S_X^2}{\sigma^2} + \frac{(q-1)S_Y^2}{\sigma^2} \sim \chi_{p+q-2}^2.$$

On obtient donc l'estimateur

$$S^2 = \frac{1}{p+q-2} ((p-1)S_X^2 + (q-1)S_Y^2) = \frac{1}{p+q-2} \left(\sum_{i=1}^p (X_i - \bar{X}_p)^2 + \sum_{i=1}^q (Y_i - \bar{Y}_q)^2 \right),$$

et par construction,

$$\frac{(p+q-2)S^2}{\sigma^2} \sim \chi_{p+q-2}^2.$$

4. Etablir alors que

$$\frac{D - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

On peut réécrire la variable précédente comme

$$\frac{D - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{D - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{p} + \frac{1}{q}}} \times \frac{1}{\sqrt{\frac{(p+q-2)S^2}{\sigma^2(p+q-2)}}}.$$

Comme

$$\frac{D - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{p} + \frac{1}{q}}} \sim \mathcal{N}(0, 1) \quad \text{et} \quad \frac{(p + q - 2)S^2}{\sigma^2} \sim \chi_{p+q-2}^2.$$

et comme S_X^2 est indépendant de \bar{X}_p et S_Y^2 est indépendants de \bar{Y}_q , on a S^2 indépendants de \bar{X}_p et \bar{Y}_q , et en particulier, S^2 est indépendants de $\bar{X}_p - \bar{Y}_q = D$, on obtient

$$\frac{D - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2}$$

5. Construire un intervalle de confiance pour la différence $(\mu_1 - \mu_2)$ au niveau de confiance $(1 - \alpha)$ avec $\alpha \in]0, 1[$.

Soit $t_{p+q-2, 1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $p + q - 2$ degrés de liberté. On a

$$\begin{aligned} 1 - \alpha &= \mathbb{P} \left[-t_{p+q-2, 1-\alpha/2} \leq \frac{D - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{p} + \frac{1}{q}}} \leq t_{p+q-2, 1-\alpha/2} \right] \\ &= \mathbb{P} \left[-t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}} \leq D - (\mu_1 - \mu_2) \leq t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}} \right] \\ &= \mathbb{P} \left[D - t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}} \leq \mu_1 - \mu_2 \leq D + t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}} \right] \end{aligned}$$

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha} = \left[D - t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}}; D + t_{p+q-2, 1-\alpha/2} S \sqrt{\frac{1}{p} + \frac{1}{q}} \right].$$

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 $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$ sachant que :

$$\begin{aligned} \sum_{i=1}^{15} x_i &= 16.2 & \sum_{i=1}^{15} x_i^2 &= 28.7 \\ \sum_{i=1}^9 y_i &= 8.9 & \sum_{i=1}^9 y_i^2 &= 31.5. \end{aligned}$$

Avant de partir comme des bourrins, on a

$$\bar{x}_{15} = \frac{16.2}{15} = 1.08 \quad \text{et} \quad \bar{y}_9 = \frac{8.9}{9} = 0.99$$

et

$$s^2 = \frac{1}{22} \left(\sum_{j=1}^{15} x_j^2 - 15\bar{x}_{15}^2 + \sum_{i=1}^9 y_i^2 - 9\bar{y}_9^2 \right) = 1.54$$

Attention : la rédaction suivante est celle qui sera attendue pour tous les tests :

On teste au risque de 1% l'hypothèse $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$.

On a la statistique de test

$$Z = \frac{D}{S \sqrt{\frac{1}{p} + \frac{1}{q}}} \sim T_{p+q-2} \quad \text{sous } H_0.$$

On a la zone de rejet

$$ZR = \{|Z| \geq t_{22,0.995}\}$$

où $t_{22,0.995}$ est le quantile d'ordre 0.995 de la loi de Student à 22 degrés de liberté. Ici, $t_{22,0.995} = 2.819$.

On a $|z| = 0.174 < 2.819$. On ne rejette donc pas H_0 .

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	$n_2 = 7$ mesures
52, 10, 40, 104, 50, 27, 146, 31, 46	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?

On a

$$m_1 = \frac{\sum_{i=1}^9 x_{i,1}}{9} = 56.22 \quad \text{et} \quad m_2 = \frac{\sum_{i=1}^7 x_{i,2}}{7} = 86.86$$

Le traitement semble avoir une influence sur la survie, mais il n'est pas possible pour le moment de répondre sérieusement à cette question.

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 :

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

Commenter les résultats obtenus. En particulier, que peut-on dire sur l'effet du traitement sur la durée de survie ?

On a

$$\begin{aligned} s_1^2 &= \frac{1}{8} \left(\sum_{i=1}^9 x_{i,1}^2 - 9m_1^2 \right) = 1799 \\ s_2^2 &= \frac{1}{6} \left(\sum_{i=1}^7 x_{i,2}^2 - 7m_2^2 \right) = 4457 \end{aligned}$$

Comme

$$\sqrt{9} \frac{\bar{X}_{9,1} - \mu_1}{S_1} \sim T_8$$

et comme $t_{8,0.975} = 2.31$, on obtient l'intervalle de confiance

$$IC_{0.95}(\mu_1) = \left[m_1 \pm t_{8,0.975} \frac{s_1}{\sqrt{9}} \right] = [23.56; 88.88]$$

De la même façon, on a

$$\sqrt{7} \frac{\bar{X}_{7,2} - \mu_2}{S_2} \sim T_6$$

et comme $t_{6,0.975} = 2.45$, on obtient l'IC

$$IC_{0.95}(\mu_2) = [25.03; 148.68]$$

Les deux intervalles s'intersectent, ce qui peut laisser supposer qu'il n'y a pas d'influence du traitement.

3. On suppose que $\sigma_1^2 = \sigma_2^2$. Tester au risque de 1% l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$. 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

```

A noter que comme $\sigma_1^2 = \sigma_2^2 = \sigma^2$, on a l'estimation

$$s^2 = \frac{1}{p+q-2} ((p-1)s_1^2 + (q-1)s_2^2) = 2934.$$

On teste au risque de 1% $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$. On a la statistique de test

$$Z = \sqrt{\frac{9 \times 7}{9+7}} \frac{\bar{X}_{9,1} - \bar{X}_{7,2}}{S} \sim T_{14} \text{ sous } H_0$$

On a la zone de rejet

$$ZR = \{|Z| > t_{14,0.995}\}.$$

où $t_{14,0.995}$ est le quantile d'ordre 0.995 de la loi de Student à 14 degrés de liberté. Ici, la p value vaut $0.2721 > 0.01$ et on ne rejette donc pas H_0 .

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.

Au vu des boxplots, ce résultats peut sembler surprenant. En réalité, le test repose sur le fait que les variances soient égales, ce qui semble faux au vu des boxplots.

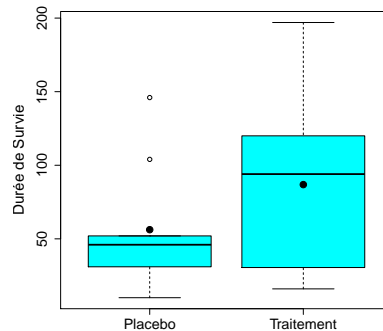


Fig. 2: Boîtes à moustaches des durées de survie par groupe.

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)
$n_1 = 15$ 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
$\sum_{j=1}^n x_j = 14975.72$
$\sum_{j=1}^{n_1} x_j^2 = 14951732$

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

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

$$m = 998.3813$$

2. Donner une estimation de σ^2 . On la notera s^2 .

$$s^2 = \frac{1}{n-1} \left(\sum_{j=1}^n x_j^2 - nm^2 \right) = 18.03347$$

3. Construire un intervalle de confiance à 90% pour la moyenne et calculer sa réalisation. On pourra s'aider des quantiles suivants :

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

On a

$$\sqrt{n} \frac{\bar{X}_n - \mu}{S} \sim T_{n-1}$$

et donc

$$IC = \left[m \pm t \frac{s}{\sqrt{n}} \right] = [996.4505; 1000.312]$$

avec $t = 1,761$

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

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

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

On a

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2,$$

et donc

$$IC = \left[\frac{(n-1)s^2}{k_{1-\alpha/2}}; \frac{(n-1)s^2}{k_{\alpha/2}} \right] = [9.673125; 44.84344]$$

avec $k_{\alpha/2} = 5.63$ et $k_{1-\alpha/2} = 26.1$.

5. 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
```

On test au risque de 1% l'hypothèse nulle $H_0 : \mu = 1$ contre $H_1 : \mu \neq 1$. On a la statistique de test

$$\sqrt{n} \frac{\bar{X}_n - 1000}{S} \sim T_{n-1} \quad \text{sous } H_0$$

On a

$$ZR = \{|z| > t_{14,0.995}\}$$

où $t_{14,0.995}$ est le quantile d'ordre 0.995 de la loi de Student à 14 degrés de liberté.

On a une p -value égale à $0.162 > 0.01$ et on ne rejette donc pas H_0 .

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 .

$$m = 44.3125$$

2. Donner une estimation de σ^2 . On la notera s^2 .

$$s^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - m)^2 = \frac{1}{n-1} \left(\sum_{j=1}^n x_j^2 - nm^2 \right) = 17.1625$$

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
```

On test au risque de 1% (ou 5%) l'hypothèse nulle $H_0 : \mu \geq 47$ contre $H_1 : \mu < 47$. Sous H_0 , il existe $\mu' \geq 47$ tel que

$$Z(\mu') = \sqrt{n} \frac{\bar{X}_n - \mu}{S} \sim T_{n-1} \quad \text{sous } H_0$$

On a la zone de rejet $ZR = \{Z(47) < -t_{47,0.975}\}$ (au risque de 5%) et $ZR = \{Z(47) < -t_{47,0.995}\}$ (au risque de 1%).

Ici la p -value vaut 0.01015 et on ne rejette donc pas H_0 au risque de 1% mais on la rejette au risque de 5%.

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

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

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

On teste au risque de 5%

$$H_0 : \sigma^2 = 5 \quad \text{contre} \quad H_1 : \sigma^2 \neq 5$$

On a la statistique de test

$$Z = 15 \frac{S^2}{5} \sim \chi_{15}^2$$

sous H_0 . On a la zone de rejet

$$ZR = \{k_{0.025} > Z\} \cup \{k_{0.975} < Z\}$$

où $k_{0.025}$ et $k_{0.975}$ sont les quantiles d'ordre 0.025 et 0.975 de la loi du chi deux à 15 degrés de libertés. Ici, $z = 51.4875$ et $k_{0.975} = 27.48839$, et on rejette donc H_0 .

Exercice 6 : On s'intéresse à la longueur de la mâchoire inférieures de 10 chacals mâles et 10 chacals femelles.

Longueur (en mm)	
Groupe 1 (Mâles)	Groupe 2 (Femelles)
$n_1 = 10$ mesures	$n_2 = 10$ mesures
120, 107, 110, 116, 114, 111, 113, 117, 114, 112	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. Calculer la moyenne des longueurs des mâchoires des groupe 1 et 2. On les notera m_1 et m_2 respectivement.

$$m_1 = 113.4, \quad m_2 = 108.6$$

2. Donner une estimation de σ^2 . On la notera s^2 .

$$\begin{aligned} s^2 &= \frac{1}{n_1 + n_2 - 2} \left(\sum_{j=1}^{n_1} (x_{j,1} - m_1)^2 + \sum_{j=1}^{n_2} (x_{j,2} - m_2)^2 \right) \\ &= \frac{1}{n_1 + n_2 - 2} \left(\sum_{j=1}^{n_1} x_{j,1}^2 - n_1 m_1^2 + \sum_{j=1}^{n_2} x_{j,2}^2 - n_2 m_2^2 \right) \\ &= 9.49 \end{aligned}$$

3. 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
```

On test au risque de 5%

$$H_0 : " \mu_1 = \mu_2 " \quad \text{contre} \quad H_1 : " \mu_1 \neq \mu_2 " .$$

On a la statistique de test

$$Z = \frac{\sqrt{n_1 n_2} (\bar{X}_1 - \bar{X}_2)}{\sqrt{n_1 + n_2} S} \sim T_{p+q-2} \quad \text{sous } H_0$$

On a donc la zone de rejet

$$ZR = \{ |Z| > t_{18,0.975} \}$$

où $t_{18,0.975}$ est le quantile d'ordre 0.975 de la loi de Student à 18 degrés de liberté. La p -value est égale à 0.00336 et on rejette donc H_0 .

4. 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. On pourra s'aider de la sortie R suivante :

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
```

On test au risque de 5%

$$H_0 : " \sigma_1^2 = \sigma_2^2 " \quad \text{contre} \quad H_1 : " \sigma_1^2 \neq \sigma_2^2 "$$

On a la statistique de test

$$Z = \frac{S_{X_1}^2}{S_{X_2}^2} \sim F_{9,9} \quad \text{sous } H_0$$

On a donc la zone de rejet

$$ZR = \{ Z < f_{9,9,0.025} \} \cup \{ Z > f_{9,9,0.975} \}$$

où $f_{9,9,0.025}$ et $f_{9,9,0.975}$ sont les quantiles d'ordre 0.025 et 0.975 de la loi de Fisher de paramètres 9,9. Ici la p -value est égale à 0.1579 et on ne rejette donc pas H_0 .

Exercice 7 : On considère un groupe de 28 individus souffrant d'un même handicap. Les individus ont été répartis en deux groupes, 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	$n_2 = 13$ mesures
19, 16, 24, 13, 9, 14, 17, 10, 19, 22, 23, 5, 7, 13, 11	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 ?

Au risque de 5% (et plus), on ne rejette pas l'hypothèse selon laquelle les variables suivent des lois normales.

2. 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
```

On teste au risque de 5% l'hypothèse nulle $H_0 : \sigma_1^2 = \sigma_2^2$ contre l'hypothèse alternative $H_1 : \sigma_1^2 \neq \sigma_2^2$.

On a la statistique de test

$$\frac{S_1^2}{S_2^2} \sim F_{14,12} \quad \text{sous } H_0$$

La zone de rejet est définie par $\left\{ \frac{1}{f_{14,12,0.975}} > z \right\} \cup \left\{ z > f_{14,12,0.975} \right\}$ où $f_{14,12,0.975}$ est le quantile d'ordre 0.975 de la loi de Fisher de paramètres 14,12. Ici la p -value est égale à 0.7503 et on ne rejette donc pas H_0 .

3. 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
```

On teste l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ contre l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$. On a la statistique de test

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{X}_1 - \bar{X}_2}{S} \sim T_{n_1 + n_2 - 2} \quad \text{sous } H_0$$

La zone de rejet est définie par

$$\{|Z| > t_{26,0.975}\}$$

où $t_{26,0.975}$ est le quantile d'ordre 0.975 de la loi de Student à 26 degrés de liberté. Ici la p -value est égale à 0.6049 et on ne rejette donc pas H_0 .

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.

Dans la première expérience, les deux échantillons sont indépendants alors que ce n'est pas le cas de la deuxième. Dans le premier cas, on va donc mettre en oeuvre un test de Student "classique" tandis que dans le deuxième, on va mettre en oeuvre le test de Student dans le cas apparié.

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 :

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

Expérience 1: On suppose que $x_{1,1}, \dots, x_{10,1}$ sont des réalisations de variables aléatoires indépendantes $X_{i,1}$ suivant la même loi que $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $x_{1,2}, \dots, x_{10,2}$ sont des réalisations de variables aléatoires indépendantes $X_{i,2}$ suivant la même loi que $X \sim \mathcal{N}(\mu_2, \sigma_2^2)$. On suppose également que les $X_{i,1}$ et $X_{j,2}$ sont indépendants et $\sigma_1^2 = \sigma_2^2$. On teste au risque de 5% :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2.$$

On a la statistique de test

$$Z = \frac{\sqrt{10 \times 10} \bar{X}_{10,1} - \bar{X}_{10,2}}{10 + 10} \sim T_{18} \quad \text{sous } H_0.$$

On a la zone de rejet

$$ZR = \{|Z| > t_{18,0.975}\}.$$

Ici, $t_{18,0.975} = 2.1$ et

$$s^2 = \frac{1}{18}(9 \times 0.05 + 9 \times 0.1) = 0.075 \quad \text{et} \quad z_{obs} = 5 \frac{1.03 - 1.12}{\sqrt{0.075}} = -1.64$$

et donc $|z_{obs}| \leq t_{18,0.975}$ et on ne rejette donc pas H_0 .

On suppose que les $x_{i,1} - x_{i,2}$ sont des réalisations de couples de variables aléatoires $(X_{i,1}, X_{i,2})$ identiquement distribués (avec $\mu_1 = \mathbb{E}[X_{1,1}]$ et $\mathbb{E}[X_{1,2}]$) et que $Y_i = X_{i,1} - X_{i,2} \sim \mathcal{N}(\mu, \sigma^2)$, avec $\mu = \mu_1 - \mu_2$. On teste au risque de 5% :

$$H_0 : \mu_1 = \mu_2 \quad \text{contre} \quad H_1 : \mu_1 \neq \mu_2$$

On a la statistique de test

$$Z = \sqrt{20} \frac{\bar{X}_{10,1} - \bar{X}_{10,2}}{S}$$

et on obtient la zone de rejet

$$ZR = \{|Z| > t_{19,0.975}\}$$

Ici $t_{19,0.975} = 2.093$ et

$$z_{obs} = \sqrt{19} \frac{0.08}{\sqrt{0.02}} = 10.75$$

et on rejette donc H_0 .

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 :

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

On test au risque de 1%

$$H_0 : "La loi est P_0" \quad \text{contre} \quad H_1 : "la loi n'est pas P_0"$$

On a la statistique de test :

$$Z = \sum_{k=1}^K \frac{(E_k - N_k)^2}{N_k} \sim \chi_2^2 \quad \text{sous } H_0$$

et la zone de rejet

$$ZR = \{Z > k_{0.99}\}$$

où $k_{0.99}$ est le quantile d'ordre 0.99 de la loi du chi deux à 2 degrés de liberté. Ici $k_{0.99} = 9.2$ et $z_{obs} = 64.44$ et on rejette donc H_0 .

2. On traite une deuxième parcelle avec le traitement habituel. Les résultats observés 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".

On peut partir du principe que le deuxième échantillon nous donne "sa loi" et faire un chi deux d'adéquation.

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

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

On a le tableau (attention mettre d'abord le deuxième effectif en proportion...)

Effet	Eradication	Amélioration	Sans effet	Total
Nombre de plants	280	210	110	600
Effectifs théoriques	330	135	135	600
"Distance"	7.6	41.7	4.6	53.9

On teste au risque de 5% si le nouveau traitement est différent de l'ancien, i.e si ils ont la même loi. On a la statistique de test :

$$Z = \sum_{k=1}^K \frac{(E_k - N_k)^2}{N_k} \sim \chi_2^2 \quad \text{sous } H_0$$

et la zone de rejet

$$ZR = \{Z > k_{0.95}\}$$

où $k_{0.95}$ est le quantile d'ordre 0.95 de la loi du chi deux à 2 degrés de liberté. Ici

$k_{0.95} = 6$ et $z_{obs} = 53.87$ et on rejette donc H_0 .

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

	O	A	B	AB	Total
Rhésus +	370	381	62	28	$n_{1.} = 841$
Rhésus -	70	72	12	5	$n_{2.} = 159$
Total	$n_{.1} = 440$	$n_{.2} = 453$	$n_{.3} = 74$	$n_{.4} = 33$	1000

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

$$\mathbb{P} [\chi_3^2 \leq 7.8] = 0.95 \quad \mathbb{P} [\chi_8^2 \leq 15.5] = 0.95$$

On commence déjà par calculer les

$$E_{i,j} = \frac{n_{i.} n_{.j}}{n}$$

	1	2	3	4
1	370.04	380.97	62.23	27.75
2	69.96	72.03	11.77	5.25

On obtient le tableau suivant :

On teste au risque de 5% l'hypothèse nulle H_0 : "le rhésus est indépendant du groupe sanguin" contre l'hypothèse alternative H_1 : "le rhésus et le groupe sanguin ne sont pas indépendants".

On a la statistique de test

$$Z = \sum_{i=1}^2 \sum_{j=1}^4 \frac{(N_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi_{1*3}^2 \quad \text{sous } H_0$$

On a donc la zone de rejet $ZR = \{Z \geq k_{0.95}\}$. Ici $k_{0.95} = 7.815$ et $z = 0.02$. Donc on ne rejette pas H_0 .

Exercice 11 : Le couvert végétal du domaine vital d'un orignal (élan d'amérique) se compose de peuplement feuillus (25% de la superficie du domaine vital), de peuplements mixtes (38% de la superficie), de peuplements résineux (25.8%) et d'un marécage (10.2%). Dans ce domaine, l'original 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'original fréquente préférentiellement certains milieux. Proposer un test statistique pour vérifier cette hypothèse.

On veut tester si la variable domaine vital suit la même loi que la variable peuplement, loi que l'on notera P_0 .

2. Réaliser le test et proposer une interprétation biologique du résultat. On pourra s'aider des résultats suivants :

$$\mathbb{P} [\chi_3^2 \leq 7.8] = 0.95 \quad \mathbb{P} [\chi_4^2 \leq 9.5] = 0.95$$

On a

$$e = (118, 201, 110, 83)$$

$$n_k = (127.750, 194.180, 131.838, 52.122)$$

On teste au risque de 5% l'hypothèse nulle H_0 : "la variable domaine vitale suit la loi P_0 " contre H_1 : "la variable domaine végétal ne suit pas la loi P_0 ".

On a la statistique de test :

$$Q^2 = \sum_{k=1}^4 \frac{(E_k - N_k)^2}{N_k} \sim \chi_3^2 \quad \text{sous } H_0$$

On a la zone de rejet $ZR = \{q^2 > k_{0.95}\}$. Ici $k_{0.95} = 14$ et $q^2 = 22.9$. Donc on rejette H_0 .

On peut en conclure que l'original doit avoir des préférences de couvert végétal.

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 :

$$\mathbb{P}[\chi_7^2 \leq 14.0] = 0.95 \quad \mathbb{P}[\chi_8^2 \leq 15.5] = 0.95$$

On a le tableau

Nb de piqûres	0	1	2	3	4	5	6	>6	Total
Nb d'individus	32	54	34	21	6	2	1	0	150
Effectif théorique	55.5	55.5	27	9.15	2.25	0.465	0.0765	0.043	150
Distance	9.95	0.04	1.81	15.34	6.25	5.067	11.15	0.004	49.62

On teste au risque de 5% l'hypothèse nulle H_0 : "le nombre de piqûres suit une loi de poisson" contre H_1 : "le nombre de piqûres ne suit pas une loi de poisson".

On a la statistique de test :

$$Q^2 = \sum_{k=0}^7 \frac{(E_k - N_k)^2}{N_k} \sim \chi_7^2 \quad \text{sous } H_0$$

On a la zone de rejet $ZR = \{q^2 > k_{0.95}\}$ où $k_{0.95}$ est le quantile d'ordre 0.95 de la loi du chi deux à 7 degrés de libertés. Ici $k_{0.95} = 14$ et on rejette donc H_0 .