T. D. n VI. Suite tests d'hypothèse

Exercice n°1.

Une nouvelle technique de dosage de sels nutritifs vient d'être mise au point. Sept dosages, effectués à l'aide de cette nouvelle technique, à partir d'échantillons d'eau de mer de la même station, donnent les résultats suivants :

$$1.17,\ 1.16,\ 1.16,\ 1.19,\ 1.19,\ 1.21,\ 1.18\, mg/l.$$

1. La technique utilisée jusque là était caractérisée par un écart-type de $0.05\, mg/l$. Peut-on dire que la nouvelle technique est plus précise que l'ancienne?
2. Déterminer un intervalle de confiance de la moyenne μ.

Exercice n°2.

On considère deux variables aléatoires réelles R et S de densités de probabilité données par :

$$f_R\,(r) = ae^{-ar},\ r > 0,\ 0\,\text{sinon},$$

et

$$f_S\,(s) = be^{-bs},\ s > 0,\ 0\,\text{sinon}.$$

On veut construire des test concernants les paramètres $a > 0$ et $b > 0$ de ces deux distributions en construisant des variables aléatoires de densité connue.

1. Montrer que la variable aléatoire $T = 2aR$ suit une loi du χ^2_2. En déduire espérance et variance de R.
2. Soit un échantillon $\{R_1, \ldots, R_n\}$ *i.i.d.* de R. Soit la variable $Z_n = 2a\,(R_1 + \ldots + R_n)$. Donner sa loi.
3. Soit un autre échantillon *i.i.d.* $\{S_1, \ldots, S_p\}$ de S. Soit la variable $U\,(n,p) = \frac{pa(R_1+\ldots+R_n)}{nb(S_1+\ldots+S_p)}$. Donner la loi de $U\,(n,p)$.
4. **Premier test**. Soit $n = 12$. On a observé :

   $$0.1,\ 1.197,\ 0.152,\ 0.182,\ 0.418,\ 0.192,\ 0.029,\ 0.885,\ 0.161,\ 0.633,\ 0.278,\ 0.008.$$

   Construire un test d'hypothèse $H_0 :\ a = 3$ contre $H_1 :\ a > 3$. Donner le résultat du test.
5. **Second test**. Soit $p = 10$. On a observé :

   $$0.361,\ 0.085,\ 0.293,\ 0.08,\ 0.077,\ 0.02,\ 0.036,\ 0.095,\ 0.197,\ 0.206.$$

   Construire un test d'hypothèse $H_0 :\ a = b$ contre $H_1 :\ a \neq b$. Donner le résultat du test.
6. **Estimation**. En cas de rejet de l'hypothèse $a = b$, déterminer un intervalle de confiance de b et donner une estimation de b.

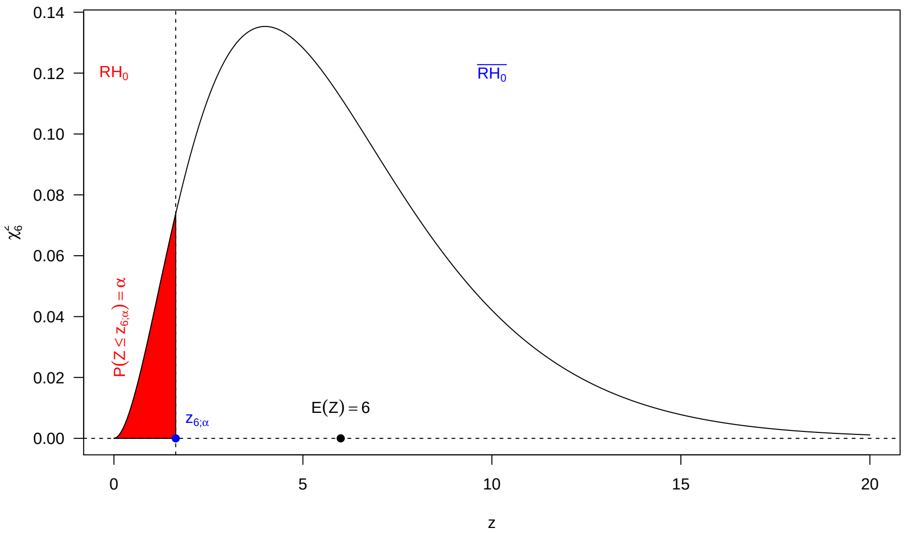

Figure 1: Distribution du χ^2_6 et zone de rejet pour $\alpha = 0.05$

Corrections

Correction Exercice n°1.

Q1- Dans cet exercice, il n'y a aucune indication sur la connaissance de paramètres populationnels. Nous travaillerons avec la réalisation d'un échantillon de $n = 7$ variables aléatoires $\{X_1, , \cdots, X_n\}$ que nous supposerons *i.i.d.* et de même loi mère d'une variable $X \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$. Nous allons donc commencer par estimer l'espérance et la variance de la population grâce aux estimateurs empiriques

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Les valeurs réalisées de ces estimateurs nous donne $\overline{x}_{obs} = 1.18$ et $6 \times s^2_6 = 0.002$. Posons $\sigma_0^2 = 0.05^2$. On va maintenant construire un test pour confronter l'hypothèse nulle $H_0 : \sigma^2 = \sigma_0^2$ contre l'alternative $H_1 : \sigma^2 < \sigma_0^2$ (l'énoncé nous dit "plus précis"). Sous H_0, la variable

$$Z = \frac{(n-1)\, S^2_{n-1}}{\sigma_0^2} = \frac{6S^2_6}{\sigma_0^2} \rightsquigarrow \chi^2_6$$

suit une loi du Chi2 à 6 degrés de liberté. Sous H_1, le protocole étant plus précis, les valeurs de l'échantillon devraient être peu dispersées : Z aura tendance à prendre des valeurs plus petites que sous H_0. Nous avons donc affaire à un test unilatéral avec zone de rejet à gauche (Fig. 1).

Fixons $\alpha = 0.05$, le risque de rejeter à tort l'hypothèse nulle, risque que l'on fixe petit. La valeur $z_{n-1;\alpha} = z_{6;0.05} = 1.635$ représente le quantile d'ordre α de la loi du χ^2_6. La zone rouge de la figure 1 représente la zone de rejet de H_0

$$RH_0 = [0; 1.635[$$

L'échantillon nous donne $s^2_6 = \frac{0.002}{6}$ soit une valeur

$$z_{obs} = \frac{0.002}{0.0025} = 0.8$$

sous H_0. On observe que $z_{obs} < 1.635$ et donc $z_{obs} \in RH_0$. On en déduit donc, qu'avec une probabilité de 0.95, le nouveau protocole expérimental est plus précis que celui d'avant. On peut également calculer la valeur $\alpha_{obs} = P\left(Z \leq z_{obs}\right) = 0.008$, (appelée *p-value*) , très faible dans ce cas.

Q2- L'intervalle de confiance de l'espérance est la fois dépendant de l'estimation de la moyenne de la population mais également de celui de l'écart-type en utilisant un échantillon de faible effectif. On cumule donc deux sources de variabilité liées aux estimations des deux paramètres populationnels. On va s'intéresser à la variable

$$T = \frac{\overline{X} - \mu}{\sqrt{\frac{S_{n-1}^2}{n}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{S_n^2}{n-1}}} = \frac{\overline{X} - \mu}{\sqrt{\frac{S_6^2}{7}}} \rightsquigarrow \mathcal{T}_6$$

qui suit une loi de Student à 6 degrés de liberté.

On connait l'$IC_{1-\alpha}$ lorsque la moyenne et la variance populationnelle sont inconnus

$$IC_{1-\alpha} = \left[\overline{X} - \sqrt{\frac{S_{n-1}^2}{n}} t_{n-1;1-\alpha/2}; \, \overline{X} - \sqrt{\frac{S_{n-1}^2}{n}} t_{n-1;\alpha/2}\right].$$

Avec $\alpha = 0.05$, la valeur observée de $\overline{X}$ et les quantiles de la loi de Student à $n - 1 = 6$ degrés de liberté, on obtient

$$IC_{0.95} = \left[1.18 - \sqrt{\frac{0.002}{6 \times 7}} \times 2.447; \, 1.18 + \sqrt{\frac{0.002}{6 \times 7}} \times 2.447\right] = [1.163; \, 1.197].$$

Avec une probabilité de 0.95, sachant l'écart-type de la population inconnu et estimé avec un échantillon de petite taille, la moyenne de la population se situe dans l'intervalle [1.163 1.197].

Correction Exercice n°2.

Q1-Soit X une variable aléatoire suivant une loi du χ_k^2 à k degrés de liberté. Sa densité est donnée par

$$f_X(x) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right), \, x \geq 0,$$

où

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} \exp(-t) \, dt$$

s'appelle la fonction Gamma. Soit $T = 2aR$ une variable aléatoire. Pour trouver sa densité, on va d'abord chercher l'expression de sa fonction de répartition (FR)

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= P(2aR \leq t) \\ &= P(R \leq \frac{t}{2a}) \\ &= F_R\left(\frac{t}{2a}\right). \end{aligned}$$

La FR de T est donc reliée à celle de R. La densité de T est donnée par la dérivée de sa FR

$$f_T(t) = \frac{d\,F_T}{d\,t}(t).$$

Or, R est une variable aléatoire qui suit une loi exponentielle de paramètre a (on notera $R \rightsquigarrow \mathcal{E}(a)$), on connait (ou on calcule aisément) sa FR que l'on peut évaluer en $\frac{t}{2a}$ avec

$$F_R\left(\frac{t}{2a}\right) = 1 - \exp\left(-a \times \frac{t}{2a}\right), \, t \geq 0.$$

La dérivée de cette fonction nous donne la densité de T

$$f_T(t) = \frac{1}{2} \exp\left(-\frac{t}{2}\right), \, t \geq 0.$$

On reconnait ici une loi exponentielle de paramètre $\frac{1}{2}$ donc $T \rightsquigarrow \mathcal{E}\left(\frac{1}{2}\right)$. Or, une variable X qui suit une loi du χ_2^2 à pour densité

$$f_X(x) = \frac{1}{2\Gamma(1)} \exp\left(-\frac{x}{2}\right),\ x \geq 0.$$

On peut montrer facilement que $\Gamma(1) = 1$ avec

$$\Gamma(1) = \int_0^{+\infty} e^{-t} dt = \left[-e^{-t}\right]_0^{+\infty} = 1,$$

et on en déduit que la densité d'une loi du χ_2^2 est la même que celle d'une loi $\mathcal{E}\left(\frac{1}{2}\right)$ donc

$$T = 2aR \rightsquigarrow \chi_2^2 = \mathcal{E}\left(\frac{1}{2}\right).$$

L'espérance et la variance d'une loi χ_n^2 sont données par

$$\begin{aligned} E(T) &= n \\ V(T) &= 2n. \end{aligned}$$

Dans notre cas, $n = 2$, donc

$$\begin{aligned} E(T) &= 2 \\ V(T) &= 4. \end{aligned}$$

On peut alors calculer l'espérance et la variance de R grâce au fait que

$$\begin{aligned} E(T) &= E(2aR) = 2 \\ V(T) &= V(2aR) = 4. \end{aligned}$$

Les propriétés de linéarité de l'espérance, les propriétés de la variance nous amène à

$$\begin{aligned} 2aE(R) &= 2 \\ 4a^2V(R) &= 4, \end{aligned}$$

on en déduit ainsi que

$$\begin{aligned} E(R) &= \frac{1}{a}, \\ V(R) &= \frac{1}{a^2}, \end{aligned}$$

résultats bien connus d'une variable qui suit une loi exponentielle de paramètre a.

Q2- Nous considérons maintenant la variable $Z_n = 2a(R_1 + \ldots + R_n)$. On sait que les variables R_i sont indépendantes. On peut réécrire $Z_n = T_1 + \cdots + T_n$, somme de n variables aléatoires indépendantes de même loi que la variable T. Or on sait qu'une variable aléatoire, somme de deux lois du χ^2 indépendantes, suit également une loi du χ^2 telle que

$$\chi_n^2 + \chi_p^2 \rightsquigarrow \chi_{n+p}^2.$$

La variable Z est une somme de n variables du χ_2^2. On en déduit donc que

$$Z_n \rightsquigarrow \chi_{2n}^2.$$

Q3- On considère maintenant la variable

$$U(n,p) = \frac{pa(R_1 + \ldots + R_n)}{nb(S_1 + \ldots + S_p)}.$$

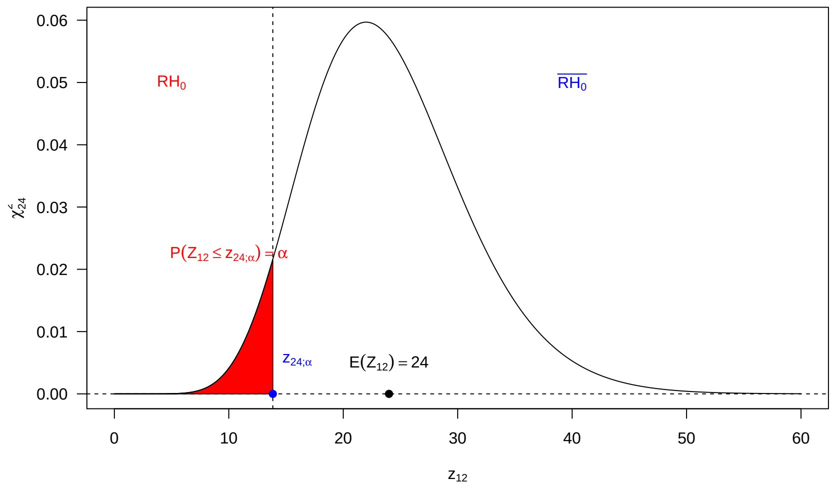

Figure 2: Distribution d'un χ^2_{24} et test unilatéral.

On déduit des premières questions que

$$Y = 2bS \rightsquigarrow \chi^2_2$$

et que la variable $W_p = 2b\left(S_1 + \ldots + S_p\right)$ est telle que

$$W_p \rightsquigarrow \chi^2_{2p}.$$

La variable $U\left(n,p\right)$ correspond donc à un rapport de deux variables aléatoires qui suivent une loi du χ^2 :

$$U\left(n,p\right) = \frac{2p \times 2a\left(R_1 + \ldots + R_n\right)}{2n \times 2b\left(S_1 + \ldots + S_p\right)} = \frac{2pZ_n}{2nW_p}.$$

La densité d'une telle variable aléatoire est connue sous le nom de loi de Fisher-Snedecor de paramètres $(2n, 2p)$ et on note

$$U\left(n,p\right) \rightsquigarrow \mathcal{F}\left(2n, 2p\right).$$

Q4- On a observé $n = 12$ valeurs de l'échantillon $\{R_1, \ldots, R_n\}$. On souhaite savoir si cet échantillon provient d'une loi exponentielle de paramètre $a_0 = 3$. Pour cela, on construit un test statistique pour confronter l'hypothèse nulle $H_0 : a = a_0$ à l'alternative $H_1 : a > a_0$. On s'intéresse naturellement à la statistique de test $Z_{12} = 2a_0\left(R_1 + \ldots + R_{12}\right)$. Sous H_0, $Z_{12} \rightsquigarrow \chi^2_{24}$. Sous H_1, l'échantillon ne provient pas d'une loi exponentielle de paramètre $a = a_0$ mais de paramètre, disons, $a = a_1$. La variable Z_{12} ne suit plus la même distribution mais on peut écrire

$$Z_{12} = \frac{a_0}{a_1} \times 2a_1\left(R_1 + \ldots + R_{12}\right) = \frac{a_0}{a_1} Z'_{12}$$

où $Z'_{12} \rightsquigarrow \chi^2_{24}$ d'après les résultats précédents. On voit que si a_1 augmente, Z_{12} prendra probablement des valeurs inférieures à celles obtenues sous H_0. Le test est donc unilatéral avec zone de rejet à gauche (Fig. 2).

Fixons $\alpha = 0.05$, le risque de rejeter à tort l'hypothèse nulle, risque que l'on fixe petit. La zone de rejet de H_0 est alors égale à :

$$RH_0 = [0; 13.85[$$

où la valeur $z_{2n;\alpha} = z_{24;0.05} = 13.85$ représente le quantile d'ordre α de la loi du χ^2_{24}. Cette valeur correspond à une probabilité de 0.95 de trouver une valeur de Z dans la zone de non-rejet de H_0 si l'échantillon provient d'une population sous $H_0 : a = a_0$

$$P\left(Z_{12} \geq z_{24;\alpha}\right) = P\left(Z_{12} \in \overline{RH}_0 / H_0\right) = 1 - \alpha.$$

L'échantillon nous donne $r_1 + \ldots + r_{12} = 4.235$ soit une valeur $z_{obs} = 2 \times 3 \times 4.235 = 25.41$, sous H_0. On observe que $z_{obs} \in \overline{RH}_0$. On peut également calculer la valeur $\alpha_{obs} = P\left(Z_{12} \leq z_{obs}\right) = 0.616$, ce qui est élevé. On en

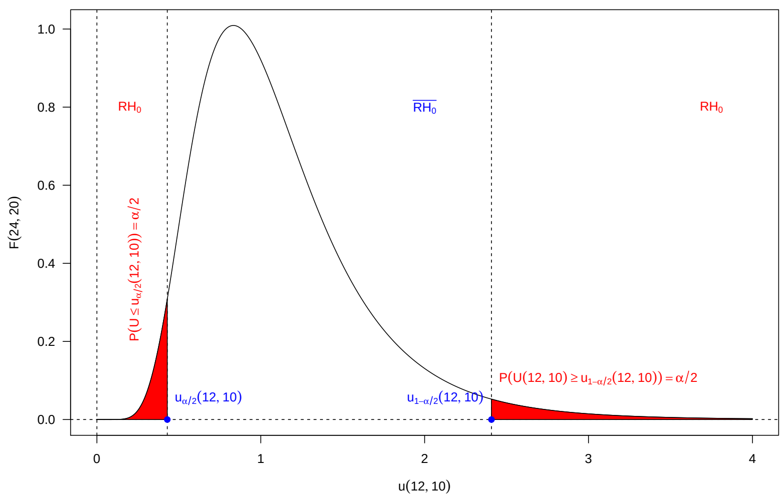

Figure 3: Distribution de Fisher $\mathcal{F}\,(24, 10)$ et test bilatéral.

déduit donc, qu'avec une probabilité de 0.95, l'échantillon provient de $n = 12$ réalisations indépendantes d'une loi $\mathcal{E}\,(a = 3)$.

Q5- On dispose maintenant d'un échantillon $\{R_1, R_2, ..., R_{12}\}$ de taille $n = 12$ et d'un échantillon $\{S_1, S_2, \cdots, S_{10}\}$ de taille $p = 10$, indépendant du précédent. On veut tester $H_0 : a = b$ contre $H_1 : a \neq b$. Pour cela, on choisit la variable

$$U(12, 10) = \frac{10a(R_1 + R_2 + \cdots + R_{12})}{12b(S_1 + S_2 + \cdots + S_{10})}.$$

que l'on sait suivre une loi de Fisher $\mathcal{F}\,(24, 20)$.

Sous $H_0 : 1 = \frac{a}{b}$, la variable

$$U_0(12, 10) = \frac{10(R_1 + R_2 + \cdots + R_{12})}{12(S_1 + S_2 + \cdots + S_{10})} \rightsquigarrow \mathcal{F}\,(24, 20)\,.$$

Sous $H_1 : a \neq b$, on peut écrire

$$U_0\,(12, 10) = \frac{b10a(R_1 + R_2 + \cdots + R_{12})}{a12b(S_1 + S_2 + \cdots + S_{10})} = \frac{b}{a}U\,(12, 10)\,.$$

Cette expression nous permet de voir que si $a > b$, les valeur de $U_0\,(12, 10)$ vont avoir tendance à baisser par rapport à H_0 : la zone de rejet sera à gauche. Dans le cas $a < b$, la zone de rejet de H_0 sera à droite. Nous avons affaire à un test bilatéral (Fig. 3).

Fixons $\alpha = 0.05$, le risque de rejeter à tort l'hypothèse nulle, risque que l'on fixe petit. La zone de non-rejet de H_0 est alors égale à :

$$\overline{RH}_0 = [0.43;\, 2.407]$$

où la valeur $u_{\alpha/2}\,(12, 10) = f_{0.025}\,(24, 20) = 0.43$ représente le quantile d'ordre $\alpha/2$ de la loi de Fisher $\mathcal{F}\,(24, 20)$et $u_{1-\alpha/2}\,(12, 10) = f_{0.975}\,(24, 20) = 2.407$ représente le quantile d'ordre $1 - \alpha/2$. L'échantillon nous donne

$$u_{obs} = \frac{10 \times 4.235}{12 \times 1.45} \simeq 2.434,$$

on voit immédiatement que $u_{obs} \in RH_0$. On en déduit donc, qu'avec une probabilité de 0.95, que les deux échantillons ne proviennent pas de la même loi exponentielle. Cependant, la valeur seuil

$$\alpha_{obs} = P\,(U\,(12, 10) \geq u_{obs}) = 0.0236$$

est très proche de 0.025 car u_{obs} est proche du seuil $f_{0.975}\,(24, 20) = 2.407$. Si le test est significatif, une petite modification de l'échantillon mènerait probablement à des résultats différents.

Q6- Pour construire un intervalle de confiance de b, nous allons considérer la variable aléatoire

$$W_{20} = 2b(S_1 + S_2 + \cdots + S_{10}) = 20b\overline{S}$$

qui suit une loi du χ^2_{20}, où $\overline{S}$ est la moyenne empirique des S_i. On peut alors calculer la probabilité suivante

$$P\left(w_{20;\alpha/2} \leq W_{20} \leq w_{20;1-\alpha/2}\right) = 1 - \alpha$$
$$P\left(\frac{w_{20;\alpha/2}}{20\overline{S}} \leq b \leq \frac{w_{20;1-\alpha/2}}{20\overline{S}}\right) = 1 - \alpha.$$

On a observé $\overline{s}_{obs} = 0.145$ et pour un coefficient de sécurité de 0.95, on a $w_{20;0,025} = 9.59$ et $w_{20;0,975} = 34.17$. D'où, l'intervalle de confiance

$$3.307 \leq b \leq 11.78.$$

On sait que $E\,(S) = \frac{1}{b}$. Un bon estimateur de l'espérance de S est donné par $\overline{S}$. C'est un estimateur sans biais $\left(E\left(\overline{S}\right) = E\,(S)\right)$ et consistant : la limite lorsque $p \to +\infty$ de $V\left(\overline{S}\right) = \frac{1}{p}V\,(S)$ est nulle. On peut donc en déduire une estimation b_{obs} de b avec

$$b_{obs} = \frac{1}{\overline{s}_{obs}} = \frac{1}{0.145} = 6.89.$$