

T. D. n VII. Révision tests d'hypothèse

Exercice n°1.

Une technique de dosage de sels nutritifs permet de fabriquer des échantillons calibrés d'eau de mer avec un écart-type de 8 mg/l . Un nouveau procédé de fabrication sera adopté s'il assure une réduction substantielle de la variabilité. Dix mesures sont réalisées sur des échantillons fabriqués avec la nouvelle méthode :

725, 722, 727, 718, 723, , 731, 719, 724, 726, 725 mg/l .

1. Peut-on adopter la nouvelle technique?
2. Déterminer un intervalle de confiance de la variance.

Exercice n°2.

On a doser des métaux lourds sur des échantillons de poissons. On souhaite ici comparer deux échantillons de taille variable et provenant de deux sites A et B . Les résultats ($\mu\text{g/g}$) sont les suivants :

	n	\bar{x}_{obs}	s_{obs}^2
A	11	3.92	0.3443
B	9	4.18	0.4760

1. Peut-on affirmer qu'il y a une différence entre les variances des deux sites?.
2. Peut-on conclure à une différence de contamination entre les sites?
3. Donner une estimation par intervalle de confiance de la moyenne pour chaque site.

Corrections

Les valeurs numériques des quantiles sont déterminées à l'aide du logiciel *R* en utilisant les fonctions :

- `qt(p,df)`: renvoie le quantile d'ordre p d'une loi de Student avec df degrés de liberté
- `qf(p,df1,df2)`: renvoie le quantile d'ordre p d'une loi de Fisher avec $(df1, df2)$ degrés de liberté
- `qchisq(p,df)`: renvoie le quantile d'ordre p d'une loi du χ^2 avec df degrés de liberté
- `qnorm(p,mu,sigma)`: renvoie le quantile d'ordre p d'une loi \mathcal{N} de moyenne mu et d'écart-type $sigma$

Correction Exercice n°1.

Q1- Soit X la variable représentant les mesures (mg/l). On supposera que $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, les paramètres de la gaussienne étant inconnus. On suppose l'échantillon $\{X_1, \dots, X_{10}\}$ *i.i.d* et de même loi mère que X . On utilise les estimateurs classiques : pour estimer la moyenne populationnelle μ , on utilise la moyenne empirique $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ et pour estimer la variance σ^2 , l'estimateur sans biais $S_{n-1}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$, avec $n = 10$.

On sait que la variable

$$Z = \frac{(n-1)S_{n-1}^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2.$$

On veut tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ contre l'alternative $H_1 : \sigma^2 < \sigma_0^2$ avec $\sigma_0^2 = 8^2 mg/l$. Sous H_0 , la variable

$$Z = \frac{9S_9^2}{\sigma_0^2} \rightsquigarrow \chi_9^2.$$

Sous H_1 , la variable S_9^2 et donc Z prendront des valeurs plus petites puisque les valeurs des X_i seront moins dispersées (expérience plus précise). On est amené à construire un test unilatéral avec zone de rejet à gauche. Fixons le niveau du test à $\alpha = 0.05$. La borne de rejet du test est donnée par le quantile d'ordre α du χ_9^2 , c'est à dire $z_{9;0.05} = 3.325$. La zone de rejet de H_0 est donc donnée par

$$RH_0 = [0; 3.325[.$$

On a observé les valeurs $\bar{x}_{obs} = 724 mg/l$, $s_{obs}^2 = \frac{130}{9} mg^2/l^2$ et on en déduit $z_{obs} = \frac{130}{8^2} = 2.03$. On constate que $z_{obs} \in RH_0$: l'hypothèse H_0 est rejetée. Avec une probabilité de 0.95, la nouvelle méthode est donc meilleure que l'ancienne.

Q2- On sait que $Z \rightsquigarrow \chi_9^2$. On peut calculer la probabilité que Z soit comprise entre deux quantiles d'ordre fixé :

$$P(z_{9;\alpha/2} \leq Z \leq z_{9;1-\alpha/2}) = 1 - \alpha.$$

Si on fixe le risque $\alpha = 0.05$, on obtient $z_{9;0.025} = 2.70$ et $z_{9;0.975} = 19.02$. L' $IC_{0.95}$ est donc le suivant

$$\begin{aligned} 2.70 &\leq \frac{9S_9^2}{\sigma^2} \leq 19.02 \\ \frac{9S_9^2}{19.02} &\leq \sigma^2 \leq \frac{9S_9^2}{2.70}. \end{aligned}$$

Les bornes de cet intervalle sont aléatoires et dépendent de la valeur prise par S_9^2 . On a observé $s_{obs}^2 = \frac{130}{9}$ et donc

$$6.83 = \frac{130}{19.02} \leq \sigma^2 \leq \frac{130}{2.70} = 48.14.$$

Il y a donc 95% de chance d'avoir un écart-type compris entre ces deux valeurs.

Correction Exercice n°2.

Q1- On admet que les dosages sur chaque site sont les réalisations d'une gaussienne $\mathcal{N}(\mu_A, \sigma_A^2)$ pour le site A et $\mathcal{N}(\mu_B, \sigma_B^2)$ pour le site B . On va tester l'hypothèse $H_0 : \sigma_A^2 = \sigma_B^2$ contre l'alternative $H_1 : \sigma_A^2 \neq \sigma_B^2$. Il

s'agit donc ici d'un test bilatéral. Pour cela, on considère les estimateurs sans biais $S_A^2 = \frac{1}{n_A-1} \sum_i (X_i - \bar{X}_A)^2$ et $S_B^2 = \frac{1}{n_B-1} \sum_j (X_j - \bar{X}_B)^2$. La statistique de test sera une variable qui suit une loi de Fisher-Snedecor de paramètres $(n_A - 1, n_B - 1)$ soit

$$Z = \frac{\sigma_B^2 S_A^2}{\sigma_A^2 S_B^2} \rightsquigarrow \mathcal{F}(n_B - 1, n_A - 1).$$

Sous H_0 , $Z = \frac{S_A^2}{S_B^2}$ suit également une loi $\mathcal{F}(n_B - 1, n_A - 1)$ car $\sigma_A^2 = \sigma_B^2$. Sous H_1 , ce rapport prendra des valeurs plus grandes ou plus petites que sous H_0 . le test est bilatéral : la zone de rejet de H_0 se situe donc à droite et à gauche. Pour un niveau de test α , elle est de la forme

$$RH_0 = [0; f_{\alpha/2}(n_B - 1, n_A - 1) \cup] f_{1-\alpha/2}(n_B - 1, n_A - 1); +\infty[$$

où les valeurs seuils $f_{\alpha/2}(n_B - 1, n_A - 1)$ et $f_{1-\alpha/2}(n_B - 1, n_A - 1)$ sont les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de la distribution $\mathcal{F}(n_B - 1, n_A - 1)$. Dans notre cas, $\alpha = 0.05$, $f_{0.025}(8, 10) = 0.233$ et $f_{0.975}(8, 10) = 3.855$. La zone de rejet de H_0 devient

$$RH_0 = [0; 0.233 \cup] 3.855; +\infty[.$$

On a observé

$$z_{obs} = \frac{s_A^2}{s_B^2} = \frac{0.3443}{0.4760} = 1.38,$$

en supposant que les variances observées sont calculées dans leur version sans biais. On constate que $z_{obs} \in \overline{RH_0}$, l'hypothèse nulle n'est pas rejetée. L'hypothèse d'égalité des variances est acceptable. Pour information, la valeur seuil observée est telle que

$$P(Z \leq z_{obs}) = 0.38.$$

Q2- On vient de montrer que l'hypothèse $\sigma_A^2 = \sigma_B^2$ était acceptable. Dans ces conditions, un estimateur sans biais de la variance commune entre les deux séries peut s'exprimer sous la forme

$$S^2 = \frac{(n_A - 1) S_A^2 + (n_B - 1) S_B^2}{n_A + n_B - 2}.$$

Dans ce cas, la variable

$$Z = \frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sqrt{S^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} \rightsquigarrow \mathcal{T}_{n+p-2}.$$

On veut ici tester l'hypothèse $H_0 : \mu_A = \mu_B$ contre l'alternative $H_0 : \mu_A \neq \mu_B$. C'est un test bilatéral. Sous H_0 , la variable Z suivra donc une loi de Student \mathcal{T}_{n+p-2} . Sous H_1 , celle-ci aura tendance à prendre des valeurs soit plus petites soit plus grande que sous H_0 . La zone de non-rejet de H_0 sera donc de la forme

$$\overline{RH_0} = [t_{n_A+n_B-2; \alpha/2}; t_{n_A+n_B-2; 1-\alpha/2}]$$

où les seuils sont les quantiles d'ordre indiqué pour un niveau de test α fixé. Dans notre cas, $\alpha = 0.05$, et $t_{18; 0.025} = -2.1$, $t_{18; 0.975} = 2.1$, la loi étant symétrique, la zone de non-rejet devient

$$\overline{RH_0} = [-2.1; 2.1].$$

On a mesuré

$$s_{obs}^2 = \frac{10 \times 0.3443 + 8 \times 0.4760}{18} = 0.403,$$

on en déduit que sous H_0

$$z_{obs} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_{obs}^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}} = \frac{3.92 - 4.18}{\sqrt{0.403 \times \left(\frac{1}{11} + \frac{1}{9} \right)}} = -0.911.$$

Cette valeur observée appartient à $\overline{RH_0}$. L'hypothèse nulle n'est pas rejetée. On peut donc admettre que les deux sites ont des niveaux de pollution équivalents en moyenne. Ce test de comparaison de deux espérances à

partir de deux moyennes empiriques est appelé *test d'homogénéité de Student*. La condition restrictive d'égalité des variances nécessite, la plupart du temps, une vérification préliminaire à l'aide d'un test de Fisher-Snedecor.

Q3- On peut compléter le test précédent en donnant un intervalle de confiance pour des moyenne μ_A et μ_B . On sait que dans le cas de population gaussienne dont les paramètres sont estimés à partir des échantillons, les $IC_{1-\alpha}$ sont donnés par

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

Dans le cas de la population A , pour $\alpha = 0.05$,

$$\begin{aligned} IC_A &= \left[3.92 - \sqrt{\frac{0.3443}{11}} t_{10; 0.975}; 3.92 + \sqrt{\frac{0.3443}{11}} t_{10; 0.025} \right] \\ &= \left[3.92 - \sqrt{\frac{0.3443}{11}} \times 2.23; 3.92 + \sqrt{\frac{0.3443}{11}} \times 2.23 \right] \\ &= [3.525; 4.314]. \end{aligned}$$

Il y a donc 95 % de chance d'avoir une moyenne populationnelle μ_A telle que

$$3.525 \leq \mu_A \leq 4.314.$$

Dans le cas de la population B , pour $\alpha = 0.05$,

$$\begin{aligned} IC_B &= \left[4.18 - \sqrt{\frac{0.4760}{9}} t_{8; 0.975}; 4.18 + \sqrt{\frac{0.4760}{9}} t_{8; 0.025} \right] \\ &= \left[4.18 - \sqrt{\frac{0.4760}{9}} \times 2.3; 4.18 + \sqrt{\frac{0.4760}{9}} \times 2.3 \right] \\ &= [3.65; 4.71]. \end{aligned}$$

Il y a donc 95 % de chance d'avoir une moyenne populationnelle μ_B telle que

$$3.65 \leq \mu_B \leq 4.71.$$

On voit qu'il y a un fort recouvrement des intervalles sur les sites A et B . Ces résultats confortent l'issue du test précédent.