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 $(\mu g/g)$ sont les suivants :

	n	$\overline{x}_{obs}$	s^2_{obs}
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 $X \rightsquigarrow \mathcal{N}\left(\mu, \sigma^2\right)$, les paramètres de la gaussienne étant inconnus. On suppose l'échantillon $\{X_1, \cdots, X_{10}\}$ $i.i.d$ et de même loi mère que X. On utilise les estimateurs classiques : pour estimer la moyenne populationnelle μ, on utilise la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et pour estimer la variance σ^2, l'estimateur sans biais $S^2_{n-1} = \frac{1}{n-1}\sum_i \left(X_i - \overline{X}\right)^2$, avec $n = 10$.

On sait que la variable

$$Z = \frac{(n-1)\, S^2_{n-1}}{\sigma^2} \rightsquigarrow \chi^2_{n-1}.$$

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

$$Z = \frac{9S^2_9}{\sigma_0^2} \rightsquigarrow \chi^2_9.$$

Sous H_1, la variable S^2_9 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^2_9, 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 $\overline{x}_{obs} = 724\, mg/l$, $s^2_{obs} = \frac{130}{9}\, mg^2/l^2$ 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 $Z \rightsquigarrow \chi^2_9$. On peut calculer la probabilité que Z soit comprise entre deux quantiles d'ordre fixé :

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

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

$$2.70 \leq \frac{9S^2_9}{\sigma^2} \leq 19.02$$
$$\frac{9S^2_9}{19.02} \leq \sigma^2 \leq \frac{9S^2_9}{2.70}.$$

Les bornes de cet intervalle sont aléatoires et dépendent de la valeur prise par S^2_9. On a observé $s^2_{obs} = \frac{130}{9}$ 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 $\mathcal{N}\left(\mu_A, \sigma^2_A\right)$ pour le site A et $\mathcal{N}\left(\mu_B, \sigma^2_B\right)$ pour le site B. On va tester l'hypothèse $H_0 : \sigma^2_A = \sigma^2_B$ contre l'alternative $H_1 : \sigma^2_A \neq \sigma^2_B$. Il

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

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

Sous H_0, $Z = \frac{S_A^2}{S_B^2}$ suit également une loi $\mathcal{F}\left(n_B - 1, n_A - 1\right)$ 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

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

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

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

On a observé

$$z_{obs} = \frac{s_A^2}{s_B^2} = \frac{0.3443}{0.4760} = 0.723$$

en supposant que les variances observées soient 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 (p-value) est telle que

$$P\left(Z \leq z_{obs}\right) = 0.33.$$

Cette dernière valeur est celle de la fonction de répartion de la loi de Fisher en $z = z_{obs}$. Elle est calculée avec le logiciel R et la commande `pf(0.723,df1 = 8,df2 = 10)`.

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

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

Dans ce cas, la variable

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

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 soient plus petites soient plus grande que sous H_0. La zone de non-rejet de H_0 sera donc de la forme

$$\overline{RH}_0 = \left[t_{n_A+n_B-2;\alpha/2};\ t_{n_A+n_B-2;1-\alpha/2}\right]$$

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

$$\overline{RH}_0 = \left[-2.1;\ 2.1\right].$$

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

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

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

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

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.