Chapitre 4

Tests statistiques

Soit une hypothèse H_0 concernant une population. Sur la base des résultats d'échantillons extraits de cette population on est amené à accepter ou rejeter l'hypothèse H_0. Les règles de décision sont appelées tests statistiques. H_0 désigne l'hypothèse dite hypothèse nulle et par H_1 on note l'hypothèse dite hypothèse alternative.

On a H_0 vraie et H_1 fausse ou bien H_0 fausse et H_1 vraie.

Tests d'homogénéité

A partir d'un échantillon de taille n_1 extrait d'une population P_1 et d'un échantillon de taille n_2 extrait d'une population P_2, le test permet de décider :

$$\begin{cases} H_0 : \theta_0 = \theta_1 \\ H_1 : \theta_0 \neq \theta_1 \end{cases}$$

où θ_0 et θ_1 sont les deux valeurs d'un même paramètre des deux populations P_1 et P_2.

4.1 Test de Student (comparaison de deux moyennes)

Soient X et Y deux variables aléatoires indépendants de moyennes m_1 et m_2 et d'écart-type σ_1 et σ_2. On dépose de deux échantillons indépendants $\{X_1; X_2; ...; X_{n_1}\}$ tel que X_i suit la même loi $N(m_1, \sigma_1)$ et $\{Y_1; Y_2; ...; Y_{n_2}\}$ tel que Y_i suit la même loi $N(m_2, \sigma_2)$. On cherche à décider si les moyennes m_1 et m_2 sont significativement différentes ou non, on utilise le test de Student :

a- Si $n_1 \geq 30$, $n_2 \geq 30$ et σ_1 , σ_2 sont connus.

On teste au seuil de signification α

$$\begin{cases} H_0 : m_1 = m_2 \\ H_1 : m_1 \neq m_2 \end{cases}$$

-On accepte H_0 (c.à.d il n'ya pas différance significative entre les moyennes de deux échantillons) si

$$z \in]-u; u[$$

où $z = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ et la valeur u est lue dans la table normale centrée réduite $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

-On rejette H_0 si $z \notin]-u; u[$ (Il ya une différance significative).

Remarque 1 Si σ_1 et σ_2 sont inconnues, on les remplace par les estimateurs

$\hat{S}_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$ et $\hat{S}_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$ respectivement, c.à.d

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}}$$

Exemple 1 Une machine remplit des paquets de café, on prélève un échantillon de paquets de taille $n_1 = 120$ de poids moyen 48.53 g et d'écart type 2.8 g, le lendemain on prélève un échantillon de taille $n_2 = 270$ de moyen 50.08 g et l'écart type 3.1 g.

Au seuil de signification 5% (risque d'erreur), qu'il existe une différance significative entre les poids moyens des paquets ?

Echantillon 1	Echantillon 2
$n_1 = 120$	$n_2 = 270$
$\bar{x} = 48.53$	$\bar{y} = 50.08$
$\sigma_1 = 2.8$	$\sigma_2 = 3.1$

Il s'agit du test $H_0 : m_1 = m_2$

$$\begin{aligned} z &= \frac{48.53 - 50.08}{\sqrt{\frac{(2.8)^2}{120} + \frac{(3.1)^2}{270}}} \\ &= -4.88 \end{aligned}$$

$$\begin{aligned}\Phi(u) &= 1-\frac{0.05}{2}\\ &= 0.975\end{aligned}$$

Dans la table $N(0,1)$, on trouve $u=1.96$, $z\notin[-1.96;1.96]$ donc on rejette H_0, il ya une différance significative entre les poids moyens des paquets.

b- si $n_1<30$, $n_2<30$ et σ_1 , σ_2 égaux et inconnus ($\sigma_1=\sigma_2=\sigma$)

-On accepte H_0 (c.à.d il n'ya pas différance significative entre les moyennes de deux échantillons) si

$$z\in\left]-t_{n_1+n_2-2,\frac{\alpha}{2}};t_{n_1+n_2-2,\frac{\alpha}{2}}\right[$$

où

$$z=\frac{\overline{x}-\overline{y}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

avec

$$S=\sqrt{\frac{(n_1-1)\,\hat{S}_1^2+(n_2-1)\,\hat{S}_2^2}{n_1+n_2-2}}$$

et la valeur $t_{n_1+n_2-2,\frac{\alpha}{2}}$ *est lue dans* la table de Student à $k=n_1+n_2-2$ dégrées de liberté (ddl) et $\gamma=\frac{\alpha}{2}$.

-On rejette H_0 si $z\notin\left]-t_{n_1+n_2-2,\frac{\alpha}{2}};t_{n_1+n_2-2,\frac{\alpha}{2}}\right[$ (Il ya une différance significative).

Exemple 2 Le poids d'un médicament conditionné en boites est réparti suivant une loi normale $N(m,\sigma)$. Deux échantillons de tailles respectives $n_1=12$ et $n_2=18$ ont pour moyennes $\overline{x}=22.235$ g et $\overline{y}=21.988$ g et écart type (estimateur) $\hat{S}_1=0.18$ g et $\hat{S}_2=0.23$ g

Qu'il existe une différance significative entre les poids moyens des deux échantillons pour un seuil de signification de 5% ?

Echantillon 1	Echantillon 2
$n_1=12$	$n_2=18$
$\overline{x}=22.235$	$\overline{y}=21.988$
$\hat{S}_1=0.18$	$\hat{S}_2=0.23$

Il s'agit du test $H_0: m_1=m_2$

$$S=\sqrt{\frac{(12-1)\,(0.18)^2+(18-1)\,(0.23)^2}{12+18-2}}=0.211\,77$$

donc

$$z = \frac{(22.235 - 21.988)}{0.211\,77 \times \sqrt{\frac{1}{12} + \frac{1}{18}}} = 3.\,129$$

Dans la table de loi de Student, on trouve

$$t_{n_1+n_2-2,\frac{\alpha}{2}} = t_{28,0.025} = 2.048,$$

$z \notin [-2.048; 2.048]$ donc on rejette H_0, il ya une différance significative entre les moyennes des deux échantillons.

4.2 Comparaison de deux proportions

Soient deux population P_1 et P_2, on extrait un échantillon de population P_1 de taille n_1 et on extrait un échantillon de taille n_2 dans la population P_2.

On compare deux proportions inconnues p_1 et p_2. On souhaite tester si ce sont les mêmes. L'hypothèse nulle à tester est H_0 : «$p_1 = p_2$»contre H_1 : «$p_1 \neq p_2$».

On dispose de deux séries d'observations, de taille n_1 pour p_1 qu'on estime par f_1 et de taille n_2 pour p_2 qu'on estime par f_2.

-On accepte H_0 (c.à.d on admet alors l'égalité des proportions) si

$$z \in]-u; u[$$

où

$$z = \frac{f_1 - f_2}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

avec

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

et la valeur u est lue dans la table normale centrée réduite $N(0,1)$ telle que $\Phi(u) = 1 - \frac{\alpha}{2}$.

-On rejette H_0 si $z \notin]-u; u[$ (Il ya une différance significative entre les proportions des deux échantillons).

Exemple 3 On expérimente un vaccin contre une maladie M sur des animaux. Un échantillon aléatoire de taille $n_1 = 80$ animaux vaccinés montre que 42 d'entre eaux ont contracté la maladie. Un échantillon aléatoire de taille

$n_2 = 113$ animaux non vaccinés montre que 76 d'entre eaux ont contacté la maladie.

Peut-on dire au seuil de signification de 5% que le vaccin est inefficace ?

On décide : $H_0 : p_1 = p_2$

$n_1 = 80, n_2 = 113, f_1 = \frac{42}{80}$ et $f_2 = \frac{76}{113}$, donc on a :

$$f = \frac{80\left(\frac{42}{80}\right) + 113\left(\frac{76}{113}\right)}{80 + 113} = 0.611$$

alors

$$z = \frac{\frac{42}{80} - \frac{76}{113}}{\sqrt{0.611\,(1 - 0.611\,)\left(\frac{1}{80} + \frac{1}{113}\right)}} = -2.071\,6$$

Dans la table $N(0, 1)$, on trouve $u = 1.96$, $z \notin [-1.96; 1.96]$ donc on rejette H_0, au seuil de signification de 5% la différance entre les proportions est significative.

4.3 Test de Fisher (comparaison de deux variances)

Soient X et Y deux variables aléatoires indépendants de moyennes m_1 et m_2 et d'écart-type σ_1 et σ_2. On dépose de deux échantillons indépendants $\{X_1; X_2; ...; X_{n_1}\}$ tel que X_i suit la même loi $N(m_1, \sigma_1)$ et $\{Y_1; Y_2; ...; Y_{n_2}\}$ tel que Y_i suit la même loi $N(m_1, \sigma_1)$. On cherche à décider si les variances σ_1^2 et σ_2^2 sont significativement différentes ou non, on utilise le test de Fisher :

On pose l'hypothèse $H_0 : \sigma_1 = \sigma_2$ (les deux populations ont la même variance) et

$$F = \begin{cases} \frac{\hat{S}_1^2}{\hat{S}_2^2} & \text{si } \hat{S}_1^2 > \hat{S}_2^2 \\ \frac{\hat{S}_2^2}{\hat{S}_1^2} & \text{si } \hat{S}_1^2 < \hat{S}_2^2 \end{cases}$$

où $\hat{S}_1^2$ est un estimateur de σ_1^2 et $\hat{S}_2^2$ est un estimateur de σ_2^2 c.à.d

$$\hat{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2;\ \hat{S}_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(y_j - \bar{y})^2$$

Si $F < F^{\alpha}_{n_1-1,n_2-1}$ on accepte H_0 (on admet alors l'égalité des variances)

Si $F > F^{\alpha}_{n_1-1,n_2-1}$ on rejette H_0 (il ya différance significative entre les variances des deux échantillons), avec la valeur $F^{\alpha}_{n_1-1,n_2-1}$ est lue dans la table de Fisher au risque d'erreur α et à $n_1 - 1$ et $n_2 - 1$ dégrées de liberté (ddl).

Exemple 4 Reprenons les données des 2 échantillons

Ech 1	7	18	9	9	18	27	12	10	32	6	37
Ech 2	12	15	14	16	22	17	25	9	18	/	/

Qu'il existe une différance significative entre les variances des deux échantillons pour un seuil de signification de 5%.

On pose l'hypothèse $H_0 : \sigma_1 = \sigma_2$

Ech 1	Ech 2
$n_1 = 11$	$n_2 = 9$
$\bar{x} = 16.82$	$\bar{y} = 16.44$
$S_1^2 = 114.96$	$S_2^2 = 23.78$

donc on a

$$F = \frac{S_1^2}{S_2^2} = \frac{114.96}{23.78} = 4.834$$

Dans la table de Fisher, on trouve :

$$F^{0.05}_{10,8} = 3.347$$

$F > F^{0,05}_{10,8}$ donc on rejette H_0 , il ya une différance significative entre les variances de deux échantillons.

4.4 Les Tests du Khi-deux

On peut distinguer trois types de test du Khi-deux χ^2 :

· Le test du χ^2 d'adéquation (H_0 : « le caractère X suit-il une loi particulière ? »),

· Le test du χ^2 d'homogénéité (H_0 : « le caractère X suit-il la meme loi dans deux populations données ? ») ,

· Le test du χ^2 d'indépendance (H_0 : « les caractères X et Y sont-ils indépendants ? »).

Ces trois tests ont un principe commun qui est le suivant : on répartit les observations dans k classes dont les effctifs sont notés $n_1 = N_1(w), ..., n_k =$

$N_k(w)$. L'hypothèse H_0 permet de calculer les effectifs théoriques, notés $n_{1,th}, ..., n_{k,th}$. On rejette H_0 si les effectifs observés sont trop differents des effectifs théoriques.

On accepte H_0 si

$$h \notin]\chi_{k-1-m,\alpha}; +\infty[$$

où

$$h = \sum_{i=1}^{k} \frac{(n_i - n_{i,th})^2}{n_{i,th}}$$

où la valeurs $\chi_{k-1-m,\alpha}$ est leu dans la table du Khi-deux avec $(k-1-m)$ dégrées de liberté (ddl)$(\gamma = \alpha)$ avec k est le nombre de classes et m est le nombre de paramètres estimés nécessaires au calcul des effectifs théoriques.

On rejette H_0 si

$$h \in]\chi_{k-1-m,\alpha}; +\infty[$$

Exemple 5 Un croisement entre roses rouges et blanches a donné en seconde génération des roses rouges, roses et blanches. Sur un échantillon de taille 600, on a troué les résultats suivants :

couleur	effectifs
rouges	141
roses	315
blanches	144

Peut-on affirmer que les résultats sont conformes aux lois de *Mendel* ?

Il s'agit donc de tester

$H_0 : p_{rouges} = 0.25, p_{roses} = 0.5, p_{blanches} = 0.25$ au risque disons $\alpha = 0,05$.

On dresse alors le tableau suivant :

couleur	effectifs observés n_i	effectifs théoriques $n_{i,th}$
rouges	141	$0.25 \times 600 = 150$
roses	315	$0.5 \times 600 = 300$
blanches	144	$0.25 \times 600 = 150$

Ici on a $k = 3$ classes et $m = 0$ (aucun paramètre a estimer pour pouvoir calculer les effectifs théoriques) donc $k - 1 - m = 2$; on calcule ensuite $]\chi^2_{k-1-m,\alpha}; +\infty[$ à l'aide de la table du Khi-deux et on obtient $\chi^2_{2,0.05} = 5.99$. Enffin, on calcule

$$\begin{aligned} h &= \sum_{i=1}^{k} \frac{(n_i - n_{i,th})^2}{n_{i,th}} \\ &= \frac{(141-150)^2}{150} + \frac{(315-300)^2}{300} + \frac{(144-150)^2}{150} \\ &= 1.53 \end{aligned}$$

donc $h \notin]5.99; +\infty[$.

On ne rejette pas H_0 au risque d'erreur $\alpha = 0,05$ (On accepte H_0), on ne peut pas dire que les observations contredisent la loi de *Mendel*.

4.5 Test de Kruskal-wallis (Test sur échantillons indépendants)

Le test de Kruskal-Wallis est un test à utiliser lorsque vous êtes en présence de k échantillons indépendant, afin de déterminer si les échantillons proviennent d'une même population ou si au mois un échantillon provient d'une population différente des autres. Il permet de tester si k échantillons ($k > 2$) proviennent de la même population, ou de population ayant des caractéristiques identiques, au sens d'un paramètre de position.

Principe du test de Kruskal-wallis

Si on désigne par M_i le paramètre de position l'échantillon i, les hypothèses nulle H_0 et alternative H_1 du test de Kruskal-wallis sont les suivantes :

-H_0 : $M_1 = M_2 = ... = M_k$

-H_1 : il existe au moins un couple (i, j) tel que $M_i \neq M_j$

1/Classer les données sous forme de tableau

Noter l'effectif de chaque série

Exemple pratique :

On veut comparer 3 milieux de culture différents A, B et C, pour cela on compte le nombre de colonies bactériennes dans chaque milieu sur plusieurs jours.

Milieu	$J1$	$J2$	$J3$	$J4$	$J5$	$J6$
A	7	4	3	2	4	–
B	5	4	4	1	3	5
C	6	7	6	5	7	6