8. Tests paramétriques

On va traiter ici deux situations distinctes. Soit on dispose des mesures de la valeur d'un caractère sur les individus d'un échantillon issu d'une population, et l'on souhaite décider si les paramètres associés (moyenne ou variance) ont une valeur conforme à celle prédite par une loi théorique connue (*tests de conformité*) ; soit on dispose des mesures du caractère sur deux échantillons issus de deux populations distinctes, et on souhaite comparer les estimations des paramètres entre elles, pour savoir si les deux populations considérées sont homogènes (*tests d'homogénéité*).

La stratégie est la même pour tous les tests statistiques : on considère une *hypothèse nulle* (H_0) : « l'estimation obtenue pour le paramètre considéré est conforme à la valeur prédite par la loi théorique », ou « les estimations du paramètre obtenues dans les différents échantillons sont homogènes ». Puis on construit une variable aléatoire dont on connaît la loi **si** (H_0) **est vraie**. On calcule sa valeur à partir des données recueillies. En fonction de la probabilité que la variable aléatoire prenne effectivement cette valeur (lue dans la table correspondant à sa loi), soit :

— on rejette (H_0), avec risque d'erreur dit *de première espèce* mesuré par α, si cette probabilité est faible (inférieure à α) ;

— on accepte (H_0), avec risque d'erreur dit *de deuxième espèce* (non mesuré), sinon.

On souhaite en général éviter le risque de première espèce, c'est-à-dire qu'on minimise le risque de rejeter (H_0) à tort ; en contrepartie, on est conduit à n'accepter (H_0) qu'avec prudence : « au risque d'erreur α, l'hypothèse (H_0) n'est pas incompatible avec les résultats observés ».

Avant de rentrer dans le détail des tests, notons que des techniques analogues permettent de tester la conformité et l'homogénéité de la fréquence d'apparition d'une valeur d'un caractère, ainsi que la conformité et l'homogénéité de la distribution des valeurs d'un caractère, et même l'indépendance de deux caractères (*tests du* χ^2).

1 Conformité

On suppose connue la loi de la variable aléatoire X correspondant au caractère étudié sur la population. En particulier, on note μ son espérance et σ^2 sa variance. On notera $\overline{X}$ (resp. S_e^2) la variable aléatoire qui prend pour valeurs la moyenne (resp. la variance) sur les échantillons de taille n. Les valeurs du caractère sur l'échantillon de taille n qui a été étudié permettent de calculer les estimations ponctuelles $\overline{x}$ et s_x.

Intéressons nous d'abord à la **moyenne** ; notre hypothèse nulle (H_0) est : « $\overline{x}$ est conforme à μ », autrement dit la moyenne observée est compatible, modulo les fluctuations d'échantillonnage, à la loi de X.

Si $n \geq 30$, sous l'hypothèse (H_0), la variable aléatoire $U = \frac{\overline{X}-\mu}{s_x/\sqrt{n}}$ suit sensiblement la loi normale centrée réduite $\mathcal{N}$. Le risque de première espèce α étant fixé, on lit u_α dans la table de l'écart réduit tel que $P(|\mathcal{N}| \geq u_\alpha) = \alpha$. On note $u = \frac{\overline{x}-\mu}{s_x/\sqrt{n}}$ la valeur de U sur notre échantillon.

- si $u \notin]-u_\alpha; u_\alpha[$, on rejette ($H_0$), avec risque d'erreur α ;
- si $u \in]-u_\alpha; u_\alpha[$, on ne peut pas rejeter (H_0).

Lorsque $n \leq 30$, sous l'hypothèse (H_0) et en supposant que X **suit une loi normale**, la variable aléatoire $T = \frac{\overline{X}-\mu}{s_x/\sqrt{n}}$ suit une loi de Student à $n-1$ degrés de liberté. On fait comme ci-dessus en remplaçant u_α par t_α lu dans la table de Student tel que $P(|T| \geq t_\alpha) = \alpha$.

Venons en maintenant à la **variance**, en supposant que X **suit une loi normale**. L'hypothèse nulle (H_0) est : « s_x^2 est conforme à σ^2 », autrement dit la variance débiaisée est compatible, modulo les fluctuations d'échantillonnage, à la loi de X. Sous l'hypothèse (H_0), la variable aléatoire $Y^2 = (n-1)\frac{S_e^2}{\sigma^2}$ suit la loi du χ^2 à $n-1$ degrés de liberté. On note $y^2 = (n-1)\frac{s_x^2}{\sigma^2}$ la valeur prise par Y^2 sur notre échantillon.

Pour $n \leq 30$, on lit a et b dans la table de la loi du χ^2 à $n-1$ degrés de liberté tels que $P(Y^2 \geq b) = \frac{\alpha}{2}$ et $P(Y^2 \geq a) = 1 - \frac{\alpha}{2}$.

- si $y^2 \notin]a;b[$, on rejette (H_0), avec risque d'erreur α ;
- si $y^2 \in]a;b[$, on ne peut pas rejeter (H_0).

Pour $n \geq 30$, on utilise la variable aléatoire $U = \sqrt{2Y^2} - \sqrt{2n-3}$, qui suit sensiblement la loi normale centrée réduite. On note u sa valeur sur l'échantillon et on conclut comme pour la conformité de la moyenne.

2 Comparaison

On étudie une variable aléatoire X dans deux populations P_1 et P_2, dans lesquelles on a étudié deux échantillons E_1 et E_2 de tailles respectives n_1 et n_2.

Pour $i \in \{1,2\}$, on note μ_i et σ_i^2 la moyenne et la variance de X dans P_i, $\overline{x}_i$ et s_i^2 les estimations ponctuelles de la moyenne et de la variance obtenues dans E_i ; enfin, on note $\overline{X}_i$ et S_i^2 les variables aléatoires qui prennent pour valeurs respectives la moyenne et la variance de X sur les échantillons de taille n_i de P_i.

En ce qui concerne la **moyenne**, l'hypothèse nulle (H_0) est « $\mu_1 = \mu_2$ », c'est-à-dire « la différence éventuelle entre $\overline{x}_1$ et $\overline{x}_2$ n'est pas significative ».

Si $n_1 \geq 30$ et $n_2 \geq 30$, sous l'hypothèse (H_0), la variable aléatoire $U = (\overline{X}_1 - \overline{X}_2)\Big/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ suit sensiblement la loi normale centrée réduite. On calcule la valeur u de U pour notre couple d'échantillons (c'est-à-dire en remplaçant $\overline{X}_i$ par $\overline{x}_i$) ; on lit u_α tel que $P(|\mathcal{N}| \geq u_\alpha) = \alpha$ dans la table de l'écart réduit et on conclut comme ci-dessus.

Si $n_1 \leq 30$ **ou** $n_2 \leq 30$, on suppose que X **suit une loi normale dans** P_1 **et dans** P_2 **et que** $\sigma_1 = \sigma_2 = \sigma$. Alors, sous l'hypothèse (H_0), la variable aléatoire $T = (\overline{X}_1 - \overline{X}_2)\Big/\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ suit sensiblement la loi de Student à $n_1 + n_2 - 2$ degrés de liberté. On va voir ci-dessous comment tester la condition $\sigma_1 = \sigma_2$; lorsqu'elle est remplie, on prendra comme estimation de σ :

$$\hat{\sigma} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \ .$$

Il ne reste qu'à calculer la valeur t de T pour notre couple d'échantillons et à conclure comme plus haut.

Pour ce qui est de la **variance**, si X **suit une loi normale dans** P_1 **et dans** P_2, l'hypothèse nulle (H_0) est « $\sigma_1^2 = \sigma_2^2$ », c'est-à-dire « la différence éventuelle entre s_1^2 et s_2^2 n'est pas significative » (et s'explique par les fluctuations d'échantillonnage).

Sous (H_0), la variable aléatoire $F = \frac{S_1^2}{S_2^2}$ suit la *loi de Snédécor* à $(n_1 - 1, n_2 - 1)$ degrés de liberté. On calcule la valeur f de F sur notre couple d'échantillons, en prenant garde que f soit

supérieur ou égal à 1 (sinon, on permute les deux échantillons). On lit dans la table de la loi de Snédécor à $(n_1 - 1, n_2 - 1)$ degrés de liberté le nombre f_α tel que $P(F \geq f_\alpha) = \frac{\alpha}{2}$, alors :

- si $f \geq f_\alpha$, on rejette (H_0), avec risque d'erreur α ;
- si $f < f_\alpha$, on ne peut pas rejeter (H_0).

Exercice 1 (théorique)
On revient sur les test de conformité de la moyenne pour de grands échantillons et de la variance pour de petits échantillons.

(a) Montrer que $P(|\mathcal{N}| \geq u_\alpha) = \alpha$ équivaut à $P(-u_\alpha \leq \mathcal{N} \leq u_\alpha) = 1 - \alpha$.

(b) Montrer que la condition $u \in]-u_\alpha; u_\alpha[$ équivaut à ce que l'espérance théorique μ appartienne à l'intervalle de confiance de la moyenne au risque d'erreur α.

(c) Montrer que $P(Y^2 \notin]a, b[) = \alpha$ si a et b sont choisis tels que $P(Y^2 \geq b) = \frac{\alpha}{2}$ et $P(Y^2 \geq a) = 1 - \frac{\alpha}{2}$. Est-ce le seul choix possible pour a et b ?

Exercice 2
Les spécifications d'un certain médicament indiquent que chaque comprimé doit contenir 2,5g de substance active. 100 comprimés sont choisis au hasard dans la production et analysés. Ils contiennent en moyenne 2,6g de substance active, avec un écart-type estimé $s = 0{,}4$g.

Peut-on dire que le médicament respecte les spécifications (au risque d'erreur $\alpha = 5\%$) ?

Exercice 3
A la suite d'un traitement sur une variété de rongeurs, on prélève un échantillon de 5 animaux ; ils pèsent respectivement 83g, 81g, 84g, 80g et 85g. A la même époque, un grand nombre de mesures a permis d'établir que les rongeurs non traités avaient un poids moyen de 87,6g.

Le poids moyen des rongeurs traités diffère-t-il significativement de cette norme au seuil de 5% ? On suppose que le poids des rongeurs suit une loi normale.

Exercice 4
On désire comparer le travail d'une nouvelle doseuse pour boîtes de haricots verts à la norme habituelle de l'usine pour laquelle l'écart-type est $\sigma = 4$g.

(a) On prélève parmi les boîtes remplies par cette nouvelle machine un échantillon de taille 10 sur lequel on obtient un écart-type estimé $s = 4{,}84$g. Peut-on considérer que ce résultat est conforme à la norme souhaitée ? Faire le test avec $\alpha = 0{,}05$, en supposant que la variable aléatoire donnant le poids de chaque boîte suit une loi normale.

(b) Même question en supposant que les valeurs numériques ont été obtenues à partir d'un échantillon de taille 50.

Exercice 5
On a prélevé deux échantillons de pommes pour les peser. Le premier échantillon, constitué de 100 pommes cueillies au début de la récolte, a pour moyenne 120g et pour écart-type estimé 20g ; Le second, constitué de 150 pommes cueillies à la fin de la récolte, a pour moyenne 150g et pour écart-type estimé 10g.

La différence entre les poids moyens à ces deux époques de la récolte est-elle significative ?

Exercice 6

Pour déterminer le poids moyen d'épis de blé appartenant à deux variétés, on a procédé à 10 pesées pour chacune. Les moyennes obtenues sont $\overline{x}_1 = 107{,}7$ cg et $\overline{x}_2 = 168{,}5$ cg. On admet que le poids des graines est distribué dans chaque variété suivant une loi de Gauss et que les variances des deux distributions peuvent être considérées comme égales. Les estimations obtenues pour celles-ci sont $s_1^2 = 432{,}9$ et $s_2^2 = 182{,}7$.

Les deux moyennes sont-elles significativement différentes au risque $\alpha = 5\%$?

Exercice 7

On revient sur l'exercice précédent. Tester l'hypothèse selon laquelle les variances des deux distributions peuvent être considérées comme égales.

Exercice 8

Dans un article de la revue « Biometrica », le biologiste Latter donne la longueur L en mm des œufs de coucoux trouvés dans les nids de deux espèces d'oiseaux :

— dans des nids de petite taille (Roitelet) :

19,8 ; 22,1 ; 21,5 ; 20,9 ; 22,0 ; 21,0 ; 22,3 ; 21,0 ; 20,3 ; 20,9 ; 22,0 ; 22,0 ; 20,8 ; 21,2 ; 21,0 .

— dans des nids de taille plus grande (Fauvette) :

22,0 ; 23,9 ; 20,9 ; 23,8 ; 25,0 ; 24,0 ; 23,8 ; 21,7 ; 22,8 ; 23,1 ; 23,5 ; 23,0 ; 23,0 ; 23,1 .

(a) Donner une estimation ponctuelle de la moyenne et de la variance de L pour chacune des deux populations : dans les nids de Roitelet et dans les nids de Fauvette.

(b) En supposant que L suit une loi de Gauss dans chacune des deux populations, montrer que les variances ne sont pas significativement différentes.

(c) Tester l'hypothèse selon laquelle le Coucou adapte la taille de ses œufs à la taille du nid dans lequel il pond.