

Faculté des sciences exactes et sciences de la nature et de la vie

Département : Sciences exactes et sciences de la nature et de la vie de

Module : Biostatistiques

Chapitre III : Tests d'hypothèses

1. Introduction :

Les tests statistiques sont des tests ayant comme objectif l'aide à la décision et la validation d'hypothèses: ils sont utilisés chaque fois que l'on veut comparer un paramètre à une valeur de référence ou plusieurs paramètres entre eux. Les paramètres peuvent être quantitatifs (moyenne, médiane, coefficient de corrélation....) ou qualitatif (pourcentage). Ces paramètres sont issus d'échantillons statistiques qui vont servir à émettre des conclusions sur la population, réelle ou imaginaire, d'où ils sont issus.

2. Choix de l'hypothèse à tester:

2.1. L'hypothèse nulle notée H_0 : c'est l'hypothèse que l'on désire contrôler, elle consiste à dire qu'il **n'existe pas de différence** entre les paramètres comparés ou que la différence observée n'est pas significative et est due aux fluctuations d'échantillonnage.

2.2. L'hypothèse alternative notée H_1 : c'est la négation de l'hypothèse nulle, elle est équivalente à dire « H_0 est fausse ». La décision de rejeter H_0 signifie que H_1 est réalisée ou H_1 est vraie.

3. Les étapes des tests d'hypothèse :

Pour réaliser un test d'hypothèse, il y a un enchaînement d'actions à effectuer : cela commence par la formulation de l'hypothèse et sa traduction en événements probabilistes liés à H_0 . On doit ensuite considérer la statistique d'écart (la loi théorique de la différence) et choisir un seuil (α) de décision. On doit ensuite calculer la valeur de la statistique d'écart pour nos valeurs puis comparer à la valeur théorique de la statistique d'écart pour le seuil choisi et en déduire si on refuse H_0 ou non.

Enfin, le calcul (ou la lecture) de la "p-value" associé au dépassement de la valeur de la statistique d'écart permet de conclure de façon fine sur le fait que la différence est significative ou non.

4. Choix d'un test statistique :

Ce choix dépend de la nature des données, du type d'hypothèse que l'on désire contrôler, des affirmations que l'on peut admettre concernant la nature des populations étudiées (normalité, égalité des variances).

5. Tests de conformité :

Les tests de conformité sont destinés à vérifier si un échantillon peut être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre comme la moyenne, la variance ou la fréquence observée.

5.1. Comparaison d'une répartition observée à une répartition théorique

Test du chi-deux (X^2) :

Est un test utilisé le plus souvent pour comparer des pourcentages (ou des distributions) observés dans deux échantillons (ou plus), et parfois amené à comparer un pourcentage à une valeur référence.

- **Le test du chi-deux (X^2) d'ajustement :**

- On compare une distribution observée à une distribution théorique ou encore un pourcentage observé à un pourcentage théorique.
- On cherche à savoir si la différence observée peut être attribuée aux fluctuations d'échantillonnage ou si elle correspond à une différence réelle.

Considérons une série statistique constituée par N observations. Ces observations sont réparties en k classes qui correspondent soit à des modalités diverses d'un caractère qualitatif, soit à des valeurs d'un caractère quantitatif.

Soit O_i l'effectif observé de la classe i ($O_1 + O_2 + \dots + O_k = N$).

L'effectif théorique de la classe i (c'est-à-dire l'effectif que la théorie permet de prévoir pour la classe i) est Np noté C_i .

- **Condition d'application du test:**

Le test du X^2 ne peut être utilisé que si tous les effectifs calculés sont suffisamment grands ($C_i \geq 5 ; \forall i = 1, \dots, k$).

- **Principe du test :** le test du X^2 se fait selon les étapes suivantes :

1/ On pose l'hypothèse nulle :

H_0 : il y a conformité (ou concordance) entre la répartition théorique et répartition expérimentale.

2/ On calcule :

$$X^2 = \frac{(O_1 - C_1)^2}{C_1} + \frac{(O_2 - C_2)^2}{C_2} + \dots + \frac{(O_k - C_k)^2}{C_k} = \sum_{i=1}^k \frac{(O_i - C_i)^2}{C_i}$$

3/ Conclusion : Etant donné un seuil de signification α ($\alpha = 5\%$ ou 1%), correspondant à un seuil de sécurité ($\alpha' = 95\%$ ou 99%).

On utilise la table X^2 pour déterminer la valeur du X^2 ayant la probabilité α d'être dépassée, avec un nombre de degrés de liberté (ddl) correspondant.

On applique ensuite la règle de décision suivante :

a) Si $X^2 \geq X_\alpha^2$ (c'est-à-dire si X^2 calculé est supérieur ou égal à la valeur dans la table) : l'hypothèse H_0 est **rejetée**

b) Si $X^2 < X_\alpha^2$: l'hypothèse H_0 est retenue, c'est-à-dire la distribution observée est conforme à la distribution théorique.

- **Le degré de liberté (ddl) :**

Toute relation supplémentaire imposée aux effectifs théoriques conduit à réduire d'une unité : **le degré de liberté (ddl)**. On peut donc dire que le degré de liberté est égale au nombre de classes diminué du nombre de relations existant entre la distribution observée et la distribution théorique.

- **La table de la loi de X^2 :** la table de la loi de X^2 donne la valeur de X_α^2 où :

- La première colonne de la table indique le nombre de degrés de liberté,
- La première ligne indique la valeur de α ,
- L'intersection d'une ligne et d'une colonne donne la valeur de X_α^2 .

Ex : Pour un nombre de ddl égal à 4, et pour une valeur de α égale à 5 %, $X_\alpha^2 = 9,49$.

- **Exemple :**

On a croisé deux races de plantes différentes par deux caractères A et B. la première génération est homogène, la seconde génération fait apparaître 4 types de plantes, dont le phénotype est noté : AB, Ab, aB, ab.

Si les caractères se transmettent selon les lois de Mendel, les proportions théoriques des 4 phénotypes sont : 9/16, 3/16, 3/16 et 1/16.

Dans une expérience, un échantillon de 160 plantes a donné :

| | | | |
|-----|----|----|----|
| AB | Ab | aB | ab |
| 100 | 18 | 24 | 18 |

- Cette répartition est-elle conforme aux lois de Mendel au seuil de signification de 5% ?

5.2. Comparaison de deux pourcentages (test de l'écart réduit):

Soit une population dans laquelle une certaine proportion des individus possède un caractère A.

On extrait au hasard dans cette population un échantillon de taille n, le problème consiste à chercher si la divergence constatée entre le pourcentage de l'échantillon et le pourcentage de la population peut être expliquée par les fluctuations d'échantillonnage ou bien le résultat expérimental est-il en contradiction avec la valeur théorique p.

- **Principe du test :**

a) **Première étape :** détermination de l'hypothèse nulle H0 et l'hypothèse alternative H1.

H0 : L'échantillon est représentatif d'une population de référence dans laquelle le caractère étudié a une fréquence théorique p₀.

b) **Deuxième étape :** Vérifier les conditions de validité du test (np₀ et nq₀) ≥ 5).

c) **Troisième étape :** Confronter les valeurs observées aux valeurs attendues sous H0 en calculant l'écart réduit :
$$\varepsilon = \frac{|P - P_0|}{\sqrt{\frac{P_0 Q_0}{n}}}$$

La lecture de la table de la loi de Laplace Gauss., permet de trouver la probabilité que la valeur ε soit supérieure ou égale à la valeur absolue de l'écart réduit observé.

Si cette probabilité est petite, on peut dire que la différence |P0- P| est un événement peu probable sous l'hypothèse nulle.

On ne rejette pas H0 si le degré de signification (probabilité lue dans la table, caractéristique d'un échantillon donné) est supérieur au seuil de probabilité choisi (en général 5%) et on rejette H0 dans le cas contraire.

- **Exemple :**

Une race de souris présente des tumeurs spontanément avec un taux parfaitement connu, soit P = 20%. Dans une expérience portant sur 100 souris, soumises à un certain traitement, on observe 34 cancers, soit p₀ = 34%. On demande si la différence entre p et p₀ est significative.

5. Comparaison d'une moyenne observée à une valeur théorique :

On note m : la moyenne observée dans l'échantillon et μ la moyenne vraie dans la population dont est issu l'échantillon, où l'on observe m ; comparer une moyenne m à la valeur théorique μ revient à savoir si la différence entre la moyenne observée m et la moyenne théorique μ est attribuable aux fluctuations d'échantillonnage ou correspond à une différence réelle entre μ et m .

On a alors deux possibilités (ou hypothèses) :

Hypothèse nulle :

$H_0 : m = \mu$: L'échantillon est représentatif d'une population ; la différence $|m - \mu|$ est petite, on peut considérer qu'elle est le fruit des fluctuations d'échantillonnage

Hypothèse alternative :

$H_1 : m \neq \mu$: L'échantillon n'est pas représentatif d'une population, ce qui veut dire que l'on a : soit un échantillon d'une population différente, soit un échantillon d'un sous-groupe de la population de moyenne ($m \neq \mu$).

Si n est grand $n \geq 30$:

Le test consiste à calculer la quantité ε observée sur l'échantillon et à la comparer à la valeur seuil de **la loi normale centrée réduite** au risque α avec $\varepsilon = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}}$

Interprétation du test : l'interprétation du test est la suivante : On rejette H_0 si $|\varepsilon| \geq \varepsilon_\alpha$

Exemple :

On a prélevé un échantillon de 100 paquets de tabac dans la production d'une machine à paqueter. La mesure du poids de ces paquets a donné une moyenne de 36g.

Vérifier si la moyenne observée est compatible avec l'hypothèse que la machine fabrique en moyenne des paquets de 40g avec un écart-type de 18g au risque de 5%.

Si n est petit $n < 30$:

Le test consiste à calculer la quantité t observée sur l'échantillon et à la comparer à la valeur seuil au risque α : $t_{n-1, \alpha}$ de **la loi de Student à (n-1) ddl**, avec $t = \frac{m - \mu}{\frac{\sigma}{\sqrt{n}}}$

L'interprétation du test est la suivante :

$H_1 : m < \mu$, on rejette H_0 si $t \leq t_{n-1; \alpha}$

6. Tests d'homogénéité :

Les tests d'homogénéité (tests d'égalité) représentent les tests les plus couramment utilisés, ils sont destinés à comparer deux populations à l'aide d'un nombre équivalent d'échantillons.

• Principe des tests :

Le principe de la comparaison consiste à estimer qu'il n'y a pas de différence significative entre les deux populations dont sont issus ces deux échantillons (c'est ce qu'on appelle hypothèse nulle).

D'après le résultat du test :

- Si l'hypothèse nulle doit être rejetée, cela signifie que les deux populations sont différentes,
- Si au contraire, l'hypothèse nulle doit être retenue : les deux populations sont effectivement semblables pour le caractère étudié.

Pour pouvoir conclure que deux populations sont identiques entre elles, il faut pouvoir comparer les paramètres qui les caractérisent (exemple la moyenne, la variance, pourcentage).

6.1. Comparaison de deux moyennes sur des échantillons indépendants :

Soit deux échantillons d'effectif n_1 et n_2 , tirés au sort dans chacune des populations que l'on veut comparer, soit m_1 , m_2 , s^1 et s^2 les moyennes et variances observées dans les deux échantillons.

Le problème consiste à comparer les moyennes de ces deux échantillons ; doit-on attribuer au hasard la différence $d = \bar{X} - \bar{Y}$ des moyennes des deux échantillons, ou correspond à une différence réelle entre les valeurs vraies dans les populations d'où sont issus les échantillons.

1^{er} Cas : Grands échantillons : $n_1, n_2 \geq 30$:

Le test de comparaison se fait selon les étapes suivantes :

1) On pose l'hypothèse nulle H_0 et l'hypothèse alternative H_1 .

H_0 : les deux échantillons proviennent d'une même population c'est-à-dire : $m_1 = m_2$.

H_1 : les deux échantillons ne proviennent pas d'une même population c'est-à-dire : $m_1 \neq m_2$

2) On calcule l'écart réduit des moyennes des deux échantillons :

$$\varepsilon = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

3) Conclusion : au seuil de sécurité de $\alpha = 95\%$

Si : $\varepsilon > 1.96$ (pratiquement 2), alors on rejette l'hypothèse nulle H_0 et on conclue que les deux populations sont différentes.

$\varepsilon \leq 1.96$ (pratiquement 2), alors on retient l'hypothèse nulle H_0 et on conclue qu'il n'y a pas de différence significative entre les deux moyennes, autrement dit on accepte l'hypothèse que les deux populations soient comparables.

Remarque : au seuil de sécurité de $\alpha = 99$, on compare ε à 2.58 (pratiquement 2.6)

Exemple : un chercheur a fait l'étude sur deux échantillons de souris qu'il a capturées en deux endroits différents. Il a obtenu les résultats suivants :

Echantillon 1 : $n_1 = 50$, $\bar{X} = 51g$,

Echantillon 2 : $n_2 = 50$, $\bar{X} = 45g$.

$$\sigma_1^2 = 256 \text{ g}^2, \quad \sigma_2^2 = 144 \text{ g}^2$$

Ces souris peuvent-elles appartenir à la même population au seuil de confiance au seuil de confiance de 95%.

2^{ème} Cas: petits échantillons : n_1 et $n_2 < 30$:

Dans ce cas le test précédent n'est plus applicable. Il faut utiliser le test de Student.

On peut calculer S^2 qui est l'estimation de la variance commune aux deux échantillons : cette estimation est donnée par la formule suivante qui pondère S_1^2 et S_2^2 par leur nombre de degré de liberté.

$$S^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2 - 2}$$

Test de Student :

Le test de Student consiste alors à calculer la quantité : $t = \frac{|\bar{X} - \bar{Y}|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

On compare la valeur de t ainsi calculée avec la valeur de la table de student en fonction du nombre de degré de liberté, dans ce cas $ddl = n_1 + n_2 - 2$ et le seuil de sécurité choisi :

- si t (valeur calculée) $\leq t_{\alpha}^*$ (valeur lue sur la table), tout ce qu'on peut dire est que la différence entre les deux échantillons n'est pas significative.

- si t (valeur calculée) $> t_{\alpha}^*$: on peut déduire que la différence entre les deux échantillons est significative au seuil de sécurité choisi, et on peut affirmer que les deux échantillons n'appartiennent pas à la même population.

6.2. Comparaison de la variance de deux échantillons indépendants (test de Fisher) :

• Principe du test :

Soit deux échantillons indépendants d'effectifs n_1 et n_2 extraits de deux populations.

Notons : s_1^2 et s_2^2 les variances d'une variable X observées dans les deux échantillons.

On souhaite comparer les variances observées.

1^{ère} étape : On fait l'hypothèse que les deux échantillons proviennent de 2 populations dont les variances sont égales : $s_1^2 = s_2^2$

2^{ème} étape : statistique du test :

La statistique associée au test de comparaison de deux variances correspond au rapport des deux variances estimées.

$$F = \frac{s_1^2}{s_2^2} \text{ (si } s_1^2 > s_2^2) \text{ ou } F = \frac{s_2^2}{s_1^2} \text{ (si } s_2^2 > s_1^2)$$

3^{ème} étape :

La valeur de la statistique F calculée (F_{obs}) est comparée avec la valeur F_{seuil} lue dans la table de la loi de Fisher-Snedecor pour un risque d'erreur α fixé et (n_1-1, n_2-1) degrés de liberté.

- Si $F_{\text{obs}} < F_{\text{seuil}}$: l'hypothèse H_0 est acceptée : les deux échantillons sont extraits de deux populations ayant même variances.
- Si $F_{\text{obs}} \geq F_{\text{seuil}}$: l'hypothèse H_0 est rejetée : les deux échantillons sont extraits de deux populations ayant des variances statistiquement différentes s_1^2 et s_2^2

6.3. Comparaison de deux pourcentages observés :

Le cas le plus fréquent est celui où l'on dispose de deux échantillons d'effectifs n_1 et n_2 dans lesquels les nombres d'individus possédant le caractère A sont respectivement k_1 et k_2 d'où

le pourcentage : $P_1 = \frac{k_1}{n_1}$ et $P_2 = \frac{k_2}{n_2}$.

La question est de savoir si la différence observée entre P_1 et P_2 est suffisamment petite pour que l'on puisse admettre que ces deux échantillons sont extraits d'une même population et que cette différence n'est due qu'à des fluctuations d'échantillonnage.

Le test de comparaison de deux pourcentages se fait selon les étapes suivantes :

Première étape : on spécifie l'hypothèse nulle H_0 et l'hypothèse alternative H_1 .

H_0 : Les 2 échantillons sont tirés au hasard d'une même population : $P_1 = P_2$

H_1 : Les 2 échantillons ne sont pas tirés au hasard d'une même population (chacun a été tiré au hasard d'une population différente ou ils ont été tirés d'une même population mais pas au hasard) : $P_1 \neq P_2$

P_1 et P_2 sont les pourcentages théoriques des populations d'où sont extraits les échantillons.

Deuxième étape : l'estimation de P et la vérification des conditions de validité:

Sous l'hypothèse H_0 , la meilleure estimation de la fréquence théorique P est :

$$P = \frac{k_1 + k_2}{n_1 + n_2}$$

Troisième étape : On calcule l'écart réduit :

$$\varepsilon = \frac{|P_1 - P_2|}{\sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

La lecture de la table de la loi de Laplace Gauss, permet de trouver la probabilité que :

Si H_0 est vraie, ε soit supérieur ou égal à la valeur absolue de l'écart réduit observé,

Si cette probabilité est inférieure à 5% on rejette H_0 au seuil de 5% et on accepte H_1 : on dit que la différence ($P_1 - P_2$) est significative au seuil de 5%.

Exemple:

Dans un groupe de 200 malades, on a constitué par tirage au sort une série soumise à un nouveau traitement A et une série soumise au traitement classique B. On a :

Traitement A; $n_A = 102$; 20 échecs soit $P_A = 19,6\%$

Traitement B; $n_B = 98$; 29 échecs soit $P_B = 29,6\%$

Question posée : les traitements A et B ont-ils le même taux d'échecs ?