ITEM 17.4

Test de comparaison de deux probabilités

Comparer deux groupes.

Voir Principe d'un test d'hypothèse, fiche transversale 4 pour les principes généraux des tests statistiques et Démarche pour la rédaction-type d'un exercice, fiche transversale 5 pour la démarche type.

POURQUOI EST-CE IMPORTANT ?

Dans de nombreuses situations, on se demande si la probabilité qu'un événement survienne est modifiée entre deux conditions. Par exemple, si un traitement est efficace, on s'attend à ce que la probabilité de guérison augmente lorsqu'on le prend. La question n'appelant que deux réponses, on réalisera pour y répondre un test statistique (test d'hypothèse).

RAPPELS DE COURS

On a vu (voir Estimer une probabilité par une proportion, item 16.3) que l'estimateur ponctuel, P_G, de la probabilité π_G était la proportion dans l'échantillon G et que, sous certaines conditions, sa loi pouvait être approchée par une loi normale. En particulier, si $n_G\,\pi_G > 5$ et $n_G\,(1-\pi_G) > 5$, avec n_G la taille de l'échantillon G, on a

$$\frac{P_G - \pi_G}{\sqrt{\dfrac{\pi_G(1-\pi_G)}{n_G}}} \sim \mathcal{N}(0;1)$$

TEST DE COMPARAISON DE Π_A À Π_B

HYPOTHÈSES DU TEST

Pour toutes les situations au programme du concours de l'internat, on cherchera à montrer que les deux probabilités diffèrent : l'hypothèse nulle sera donc

$$H_0 \text{ « } \pi_A = \pi_B \text{ »}$$

où π_A est la probabilité de l'événement étudié dans la condition A (groupe A) et π_B celle dans la condition B (groupe B).
L'hypothèse alternative H_1 peut prendre trois formes :

- H_1 « $\pi_A > \pi_B$ », unilatérale (supériorité) ;
- H_1 « $\pi_A < \pi_B$ », unilatérale (infériorité) ;
- H_1 « $\pi_A \neq \pi_B$ », bilatérale (différence).

Une seule de ces 3 hypothèses alternatives doit être choisie. On la détermine en fonction de la question posée.

PROPORTION COMMUNE

Si H_0 est vraie, $\pi_A = \pi_B$. Les deux estimations p_A et p_B sont donc deux estimations de la même valeur. Il est plus intéressant d'en construire une seule, plus précise, en fusionnant les deux échantillons. On l'appelle proportion commune et elle vaut

$$P_C = \frac{n_A P_A + n_B P_B}{n_A + n_B}$$

- P_C correspond au « nombre de cas favorables » rapporté au « nombre de cas possibles » en fusionnant les deux échantillons.
- P_C est aussi la moyenne pondérée des deux estimations séparées, les pondérations étant les tailles des échantillons.
- P_C est une variable aléatoire ; p_C est sa réalisation obtenue dans l'expérience étudiée.

STATISTIQUE DE TEST

La statistique de test utilisée est

$$C = \frac{P_B - P_A}{\sqrt{P_C(1-P_C)\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right)}}$$

Remarque

On peut tout aussi bien utiliser la différence $P_A - P_B$ au numérateur. Cela change seulement le signe. Attention cependant, pour un test unilatéral, à rester cohérent avec la zone de rejet.

Si l'hypothèse H_0 est vraie et *si les deux échantillons sont indépendants*, C est centré (E(C) = 0) et approximativement réduit (V(C) ≈ 1). En revanche, C n'est ni centré ni réduit si H_1 est vraie.
Si l'on peut approcher les lois de P_A, P_B et P_C par une gaussienne, alors C suivra, si H_0 est vraie, une loi normale centrée réduite.

À retenir

On vérifiera dans chaque groupe que l'approximation est justifiée en utilisant la *proportion commune*, donc en vérifiant que $n_A\,p_C > 5$, $n_A\,(1-p_C) > 5$, $n_B\,p_C > 5$ et $n_B\,(1-p_C) > 5$. On rencontre aussi la convention ≥ 5.

RÉGIONS DE DÉCISION DU TEST

On considère qu'un risque α a été choisi. La zone de non-rejet de l'hypothèse nulle va dépendre de l'hypothèse H_1 choisie.
On note u_q le quantile d'ordre q d'une loi normale centrée réduite.

- Pour H_1 « $\pi_A > \pi_B$ » : non-rejet de H_0 si $C \in]-\infty ; u_{1-\alpha}]$.
- Pour H_1 « $\pi_A < \pi_B$ » : non-rejet de H_0 si $C \in [u_\alpha ; +\infty[$.
- Pour H_1 « $\pi_A \neq \pi_B$ » : non-rejet de H_0 si $C \in [u_{\alpha/2} ; u_{1-\alpha/2}]$.

À retenir

Avec les tables de l'internat (qui sont écrites pour le cas bilatéral), on trouve les valeurs seuils ainsi :

- $u_{1-\alpha}$ est la valeur dans la table correspondant à 2α ;
- u_α est l'opposé de la valeur dans la table correspondant à 2α ;
- $u_{1-\alpha/2}$ se lit dans la table directement avec le α donné dans l'énoncé et $u_{\alpha/2} = -u_{1-\alpha/2}$.

EXEMPLES

CONCOURS 2019

EXERCICE 3

Un essai thérapeutique randomisé a été réalisé pour comparer la qualité de vie de deux groupes de 100 patients chacun, souffrant d'insuffisance cardiaque chronique. Le groupe A bénéficiait de la prise en charge clinique habituelle en ambulatoire tandis que le groupe B bénéficiait d'un programme d'éducation thérapeutique.
27 % des patients du groupe A et 15 % des patients du groupe B vivaient seuls. Ces deux proportions sont-elles significativement différentes ?

CORRIGÉ

POSITIONNEMENT DU PROBLÈME

La variable d'intérêt est la variable qualitative binaire « vivre seul » (réponse : oui ou non). On cherche à comparer la probabilité de l'événement « vivre seul » entre les deux groupes, à partir de leurs estimations ponctuelles, les proportions de patients vivant seuls.
On note π_A la probabilité de vivre seul pour un individu du groupe A et π_B celle pour un individu du groupe B ; on note $p_A = 0{,}27$ et $p_B = 0{,}15$ les proportions observées dans ces deux groupes.

HYPOTHÈSES DU TEST

On se demande si ces probabilités diffèrent. Ainsi, on cherche à démontrer l'hypothèse H_1 « $\pi_A \neq \pi_B$ », et l'hypothèse H_0 est donc « $\pi_A = \pi_B$ ».

PROPORTION COMMUNE

On calcule $p_C = \dfrac{100 \times 0{,}27 + 100 \times 0{,}15}{100+100} = \dfrac{0{,}27+0{,}15}{2} = 0{,}21$.

STATISTIQUE DE TEST

La statistique de test est $C = \dfrac{P_B - P_A}{\sqrt{P_C(1-P_C)\left(\dfrac{1}{n_A}+\dfrac{1}{n_B}\right)}}$; on a $n_A = n_B = 100$, donc $n_A\, p_C = n_B\, p_C = 21 > 5$ et $n_A\,(1 - p_C) = n_B\,(1 - p_C) = 79 > 5$: l'approximation gaussienne est valable ; on considère donc que $C \sim \mathcal{N}(0 ; 1)$.

RÉGIONS DE DÉCISION DU TEST

Compte tenu de l'hypothèse H_1, on rejettera H_0 si la valeur observée de la statistique de test, c_{obs}, est « trop différente de 0 » : si elle est supérieure à $u_{1-\alpha/2}$ ou inférieure à $u_{\alpha/2}$, où α est le risque choisi. La zone de non-rejet de H_0 est donc l'intervalle $[u_{\alpha/2} ; u_{1-\alpha/2}]$.
Pour $\alpha = 5$ %, la lecture de la table donne $u_{1-\alpha/2} = -u_{\alpha/2} = 1{,}960$.

RÉALISATION DU TEST

On calcule $c_{obs} = \dfrac{0{,}15-0{,}27}{\sqrt{0{,}21 \times 0{,}79 \times \left(\dfrac{1}{100}+\dfrac{1}{100}\right)}} = -2{,}083$. Cette valeur, inférieure à $u_{0{,}025}$, est dans la zone de rejet de l'hypothèse H_0 : au risque $\alpha = 5$ %, les deux proportions sont bien significativement différentes et les probabilités de vivre seuls diffèrent entre les deux groupes.

CONCOURS MAI 2012

EXERCICE 1

Une étude portant sur les facteurs de risque de la maladie de Parkinson a été réalisée sur un échantillon de 120 malades. Ces malades ont été répartis en 2 groupes selon leur score à un test neuropsychologique DRS (*Dementia Rating Scale*). Le groupe 1 est constitué de malades dont les facultés mentales sont « normales » et le groupe 2 de malades atteints de « démence ». Un groupe « contrôle » (groupe 3) est constitué de 80 volontaires sains de même âge. Les caractéristiques démographiques et cliniques des sujets sont rassemblées dans le tableau ci-dessous.

Caractéristiques	Groupe 1 *Normal*	Groupe 2 *Démence*	Groupe 3 *Contrôle*
Effectif	65	55	80
Pourcentage d'hommes	63,1 %	78,2 %	62,5 %

L'augmentation apparente du pourcentage d'hommes du groupe 2 par rapport au groupe contrôle est-elle significative ?

CORRIGÉ

On note π_X la probabilité qu'un patient du groupe X soit un homme et p_X la proportion d'hommes dans le groupe X, estimation ponctuelle de X.
On veut montrer qu'il y a plus de chances qu'un patient du groupe 2 soit un homme que pour un patient du groupe 3 : on pose donc l'hypothèse H_1 « $\pi_2 > \pi_3$ » et comme hypothèse H_0 « $\pi_2 = \pi_3$ ». La statistique de test est $C = \dfrac{P_3 - P_2}{\sqrt{P_C(1-P_C)\left(\dfrac{1}{n_3}+\dfrac{1}{n_2}\right)}}$; on attend des valeurs négatives de cette statistique de test si H_1 est vraie.

La proportion commune vaut $p_C = \dfrac{55 \times 0{,}782 + 80 \times 0{,}625}{55+80} = 0{,}689$;

on a $n_2\, p_C = 37{,}9 > 5$, $n_2\,(1 - p_C) = 17{,}11 > 5$ et comme $n_3 > n_2$, les deux autres conditions sont aussi vérifiées. Ainsi, C suit (si H_0 est vraie) une loi normale centrée réduite.
Comme on attend des valeurs négatives de la statistique de test sous H_1, on rejettera H_0 si $C < u_\alpha$; pour $\alpha = 5$ %, $u_\alpha = -1{,}65$: la zone de non-rejet de H_0 est donc $[-1{,}65 ; +\infty[$.
On calcule $c_{obs} = -1{,}936$: on rejette donc l'hypothèse H_0 au risque $\alpha = 5$ % ; la proportion d'hommes dans le groupe 2 est bien significativement augmentée.

Remarque

On a ici choisi un test unilatéral car l'énoncé évoque une augmentation de la probabilité d'être un homme. Cependant, il n'est pas certain que l'on attende *a priori* une augmentation (le choix d'une augmentation est induit par l'observation d'une proportion plus élevée) ; dans ce cas, on peut aussi envisager un test bilatéral.
Attention, pour un test bilatéral, le calcul de c_{obs} est le même, mais cette fois la zone de non-rejet est $[-1{,}96 ; +1{,}96]$ et là, le test n'est pas significatif !