Calculs sur les pourcentages

0_ Loi normale

La loi normale N (m ; â), c'est la loi définie par les fonctions de densité :

f(x) = 1/â (2à)^{½} * e$^{-(x-m)²/2â²}$ où m et â ì þ *$^{+}$

Si X suit une variable aléatoire qui suit la loi normale N (m ; â) ou X Χ N (m ; â) alors : m = E(X) et â ² = V(X).

Lorsque l'on lit la table de Ë (t) de N (0 ; 1), on a :

P(X < t) = F(t) = Ë (t) et Ë (a) = 1 – Ë (-a)

Lorsque l'on lit le quantile õ$_{Ñ}$:
Pour Ñ = 0,05, õ$_{0,05}$ = 1,96 et pour Ñ = 0,01, õ$_{0,01}$ = 2,58
Calcul : P(X < õ$_{Ñ}$) = 1 – Ñ/2

Soit P, une population constituée uniquement de deux catégories A et B. Un échantillon E de P est supposé obtenu par tirage au sort simple, aléatoire et non exhaustif (= avec remise).

I_ Intervalle de confiance

A. Intervalle de confiance et risque

p_0 est le pourcentage (= la fréquence) des individus de la catégorie A dans E.

p est le pourcentage de la population initiale P.

Définition : L'intervalle de confiance au risque Ñ pour p est un intervalle où p a la probabilité (1 – Ñ) de se trouver.

Remarques :

_ Ñ est la probabilité ou le risque de se tromper en affirmant que p est dans l'intervalle de confiance au risque Ñ pour p.

_ (1 – Ñ) est dit le coefficient de sécurité ou est aussi dit le seuil de signification.

à On pose I$_{Ñ}$ = p_0 G õ$_{Ñ}$ * †($p_0 q_0$ / n)

où q_0 = 1 – p_0 et õ$_{Ñ}$ est la valeur donnée par la table de la loi normale : X Χ N (0;1) et P(|X| > õ$_{Ñ}$) = Ñ.

Définition : E est dit un grand échantillon si : np_0, nq_0, nÑ$_i$, $n($ 1 - Ñ$_i$) > 5 ; i = 12.

On supposera par la suite que E est un grand échantillon si n > 30.

Remarques : Plus le risque est faible, plus l'intervalle est large.

à Si d est la précision et ne doit pas dépasser d_0 :

d = | p – p_0 | = Õ$_{Ñ}$ * †(p_0q_0 / n)

è d^2 = Õ$_{Ñ}^2$ * p_0q_0 / n < d_0^2 Ainsi n > Õ$_{Ñ}^2$ * p_0q_0 / d_0^2

B. Test de Conformité

On veut comparer une valeur théorique p à une valeur p_0 observée sur un échantillon de taille n.
On teste l'hypothèse nulle : H_0 = "*p est la proportion des individus de la population de la catégorie A.*"

à On pose : Õ$_0$ = | p_0 – p | / †(pq / n) avec n > 30

Test :

q Si Õ$_{Ñ}$ T Õ$_0$, on refuse H_0 au seuil de signification Ñ
q Si Õ$_{Ñ}$ > Õ$_0$, on ne peut refuser H_0 au risque Ñ

C. Test d'homogénéité

On considère ici deux populations d'individus de catégories A et B et on extrait un échantillon de chacune. On observe une proportion p_0 et p'_0 d'individus de la catégorie A sur les deux échantillons. On veut comparer p_0 et p'_0. Pour cela on formule l'hypothèse H_0 = "*Les deux populations ont la même proportion p d'individus de la catégorie A.*"

à On pose Õ$_0$ = | p_0 – p'_0 | / â$_d$ et â$_d$ = †(p_0q_0 / n_0 + $p'_0q'_0$ / n'_0)

On effectue ensuite le même test que précédemment.

II Test du Khi-2

Soient æ$_1$ … æ$_Ý$ avec Ý des variables aléatoires indépendantes telles que :
æ$_i$ X N (0 ; 1) et X$_Ý$ = Œ æ²$_i$.
Définition : X$_Ý$ suit une loi dite du æ² à Ý degrés de liberté.
On peut montrer que la densité du X$_Ý$ est donnée par :

f(x) = 1/2$^{Ý/2}$Æ(Ý/2) * x$^{2/Ý}$ e$^{-x/2}$ où Æ(x) = €$_0^{'}$ t^{x-1}e^{-t} dt

Propriété : _ Si x$_i$ … x$_n$ sont n variables aléatoires avec x$_i$ X N (0 ; 1) et si æ$_i$ … æ$_n$ sont liés par k relations, alors : Œ x$_i^2$ X æ²$_{n-k}$.

_ Si X X $æ^2_n$, Y X $æ^2_m$ et X et Y sont indépendantes, alors (X + Y) X $æ^2_{n+m}$.

A. Valeur de Khi-2

Si Ñ g [0 ; 1 [, il existe un unique nombre positif noté $æ^2_{Ñ,Ý}$ tel que :

$$P(æ^2_{Ñ,Ý} \text{ T } æ^2_{Ý}) = Ñ$$

à $æ^2_{Ñ,Ý}$ sera donné par la table pour un Ñ donné et un coefficient de liberté Ý défini :
Ý = (nb de lignes – 1) * (nb de colonnes – 1)

Contexte : On considère une distribution observée sur un échantillon E de taille n_T réparti en k classes. O, dispose :

_ d'une série d'effectifs observés : O_1, O_2 …
_ d'une série d'effectifs théoriques : $C_1 = n_T * p_1$ … $C_k = n_T * p_k$

Objectif : Comparer la distribution donnée par E et la distribution théorique (à un seuil de risque Ñ).

Validité : On suppose que E est non exhaustif et aléatoire, le nombre de classes est peu élevé (k < 20) et les effectifs $C_i > 5$.

B. Test de conformité

On pose : $æ^2_0 = Œ (O_i – C_i)^2 / C_i$

H_0 : "*Dans la population observée, la distribution des fréquences observées est celle décrite par le modèle théorique.*"

Propriété : Si $æ^2_0$ U $æ^2_{Ñ;k-1}$, on refuse l'hypothèse au seuil de risque Ñ, sinon on ne peut refuser l'hypothèse.

Remarque : Dans le premier cas, la probabilité de se tromper en refusant l'hypothèse n'étant que du risque Ñ, on prend ce risque. Dans le deuxième cas, le test ne fournit aucune exigence contre l'hypothèse.

à Si k = 2 et p_0 (respectivement p) le pourcentage observé (respectivement théorique) de l'une des deux classes, le test sur les pourcentage avec $Õ_0$ est équivalent au test æ².

à Lorsque le degré de liberté est 1, on effectue la correction de Yates en remplaçant :
$æ_0 = Œ (| O_i – C_i | - ½)^2 / C_i$

C. Test d'échantillonnage

Le modèle théorique est connu et on veut tester l'hypothèse nulle suivante : H_0 "*L'échantillon* E *extrait de la population est représentatif*."

à On la test comme dans le B

Remarque : On note x_{moy} et â $_e$ la moyenne et l'écart-type donnés par E si l'espérance mathématique m et l'écart-type s de la distribution théorique ne sont pas connus. On les estime par : m = x_{moy} et s = †(n / n – 1) * â $_e$.

Lorsque Ý > 30 : (2 XÝ)^{½} X N ((2Ý - 1)^{½}; 1)

D. Test d'homogénéité

On observe un caractère sur une population Ð . On constitue m échantillons E_i … E_m de tailles n_i.

On considère un partage en k classes pour chaque i i { 1 ; … ; m }. On note O_{ij}, l'effectif des observations de E_i dans la classe j.

On fait ensuite un tableau de contingence :

Échantillon	Classe j	…	Classe k	Total
E_i	O_{ij}			n_i
…				
E_m			O_{mk}	n_m
Total	t_j		t_k	n_T

è æ$^2_0$ = Ì (O_{ij} – C_{ij})² / C_{ij}

On teste H_0 = "*Les échantillons E_i … E_m sont extraits d'une même population.*"

à Pour tester H_0, on commence par estimer les probabilités de chaque classe en utilisant H_0 elle-même.

On pose E = Š E_i et | E | = n_T.

- q La probabilité pour la classe j est $p_j = t_j / n_T$
- q L'effectif théorique pour la classe j est $C_{ij} = n_i t_j / n_T$
- q æ^2_i = Œ (O_{ij} – C_{ij})² / C_{ij}
- q æ$^2_0$ = Œ æ^2_i

Propriétés : Si æ$^2_0$ U æ$^2_{Ñ;Ý}$, on refuse H_0 au seuil Ñ.
Si æ$^2_0$ U æ$^2_{Ñ;Ý}$, on ne peut refuser H_0 au seuil Ñ.

III Test de Student

Soit X X N (0 ; 1) et Y X æ^2_Y, deux variables aléatoires indépendantes.

A. Définition

La variable aléatoire $t_Ý$ = X / †(Y / Ý), elle suit une loi de probabilité dite loi de Student de degré de liberté Ý.

Cette loi est définie par la fonction de densité suivante :

$$f(x) = A[1 + x^2 / Ý\ ]^{-(Ý+1)/2}$$, A est une constante positive

B. Propriétés

E(tÝ) = 0 et â (tÝ) = †[Ý / (Ý-2)] ainsi que Ý U 3.

Remarque : si Ý tend vers l'infini, tÝ X N (0 ; 1).

C. Table

Pour Ñ g] 0 ; 1], il existe un nombre positif unique $t_{Ñ;Ý}$ tel que :

$$P(|t_Ý| \text{ U } t_{Ñ;Ý}) = Ñ.$$

à Connaissant Ñ et Ý, la table donne $t_{Ñ;Ý}$.

Ex : $t_{0,05;6} = 2,447$ et $t_{0,05;8} = 2,306$

VI _ Calcul sur les moyennes

On prélève un échantillon non exhaustif de taille n. On note x_i … x_n les observations relatives du caractère étudié.

à On note : $x_{moy} = (Œ\ x_i) / n$ et $â_e^2 = (Œ\ x_i^2) / n – x_{moy}^2$.

A. Intervalle de confiance

Définition : L'intervalle de confiance au risque Ñ pour la moyenne m est un intervalle où m a la probabilité 1 – Ñ de se trouver.

à On estime : $m = x_{moy}$ et $s = â_e * †[n / (n – 1)]$.

è On suppose que X X N (0 ; 1) et n > 10.

Propriété : On a l'intervalle au risque Ñ pou m :

$$I_Ñ = x_{moy} \text{ G } t_{Ñ;n-1} * s / †n$$

Remarques :_ Lorsque *n > 30*, on peut remplacer $t_{Ñ;n-1}$ par $Õ_Ñ$.
_ Le risque de se tromper en affirmant que m g $I_Ñ$ est Ñ.
_ $I_Ñ$ est aussi dit l'indice de confiance au seuil de sécurité 1 - Ñ.

B. Comparaison d'une moyenne théorique à une moyenne observée

On s'intéresse à un caractère dans une population et on suppose à priori que la moyenne de la variable aléatoire X correspondant à ce caractère est m et on veut tester cette supposition, c'est-à-dire l'hypothèse H_0 : "*m est la moyenne du caractère dans la population*".

On extrait un échantillon non exhaustif de taille n. On suppose que n > 10 ou que X suit un loi normale.

à On note x_{moy} la moyenne observée sur l'échantillon. On pose :

$$t_0 = | x_{moy} – m | / (s / †n)$$

Test : _ Si t_0 U $t_{Ñ;n-1}$, on refuse H_0 au seuil de risque Ñ
_ Si $t_0 < t_{Ñ;n-1}$, on ne peut refuser H_0 au seuil de risque Ñ

C. Composition de deux moyennes

l Test paramétrique :

On suppose que l'on a deux échantillons prélevés sur deux populations distribuées, chacune suivant une loi normale de moyenne m_i et de variance $â_i^2$.
On suppose que $â_1 = â_2 = s$.
à On veut tester l'hypothèse H_0 : "*m1 et m2 sont égales*".

Remarque : Si H_0 est rejetée, on conclut que les moyennes sont différentes au seuil prédéterminé sinon, on conclut qu'aucune différence significative n'est montrée entre les deux moyennes.

On estime $m_i = x_{imoy}$ où x_{imoy} est la moyenne de l'échantillon 1 puis 2 et
$s^2 = s_1^2(n_1 – 1) + s_2^2(n_2 – 1) / (n_1 + n_2 – 2)$.

Et on pose :
$$t_0 = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{s \times \sqrt{1/n_1 - 1/n_2}}$$

Propriété : soit Ñ g] 0 ; 1 [:
_ t_0 U $t_{Ñ;n1+n2-2}$ à Il existe une différence entre les deux moyennes au seuil Ñ.
_ $t_0 < t_{Ñ;n1+n2-2}$ à Aucune différence n'a pu être observée.

Remarques :

à Si $n_1 = n_2 = n$, le calcul est simplifié :
$$t_0 = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}$$

à Si n_1 et $n_2 > 30$, alors t_0 peut être remplacé par :
$$\varepsilon_0 = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2}{n_1} - \frac{s_2^2}{n_2}}}$$

l Test de Wilcoxon :

Il s'agit de tester l'homogénéité de deux échantillons E et E' de tailles n et n' extraits de deux populations P et P' de façon indépendante et non exhaustive. P présente une moyenne m et P' présente une moyenne m'.

Le test de Wilcoxon ne suppose pas que les populations sont distribuées suivant une loi binomiale. Il est dit non paramétrique.

On suppose par la suite que n T n'. On veut tester l'hypothèse H_0 : "*Il n'y a pas de différence entre les moyennes m et m'*".

<u>Le rang</u> :

On range les observations sur E N E' par ordre croissant de 1 à n+n'. On attribue à chaque valeur x_i (observée) le nombre ái défini par :

á$_i$(x_i) = la moyenne des rangs de valeurs identiques à x_i.

<u>Table de Wilcoxon</u> :

Pour Ñ g] 0 ; 1 [et n et n', deux entiers avec n T n', la table de Wilcoxon fournit un intervalle] è$_1$; è$_2$ [.

<u>Test</u> : _ Si Œ h] è$_1$; è$_2$ [, il existe une différence significative entre m et m'.
_ Si Œ h] è$_1$; è$_2$ [, aucune différence significative n'est observée.