STA401

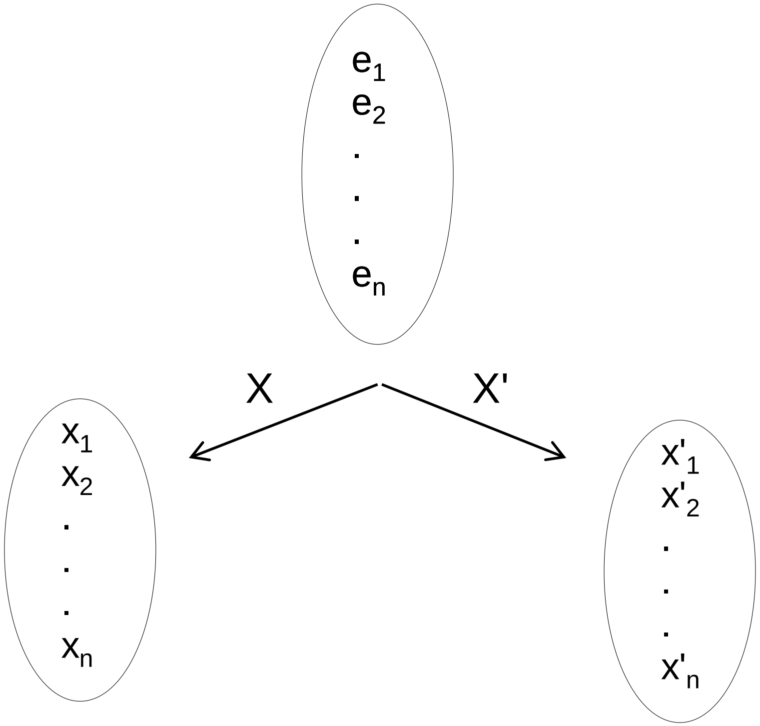
Statistique et Calcul des Probabilités

Responsable *: Carole Durand-Desprez*

CHAPITRE 5 : Autres tests statistiques. Comparaison d'échantillons

I. Test de comparaison de deux séries appariées, (un test paramétrique particulier) :

Modèle



Un seul échantillon d'individus (n)

Une seule et même variable
mesurée par 2 “méthodes“

$\downarrow$

Données appariées

Condition : X - X' suit une loi Normale

Test :

$$\begin{cases} H_0 : \text{Les 2 séries sont identiques} \\ H_1 : \text{Les 2 séries sont différentes} \end{cases} \Leftrightarrow \begin{cases} H_0 : \text{Les 2 méthodes sont identiques} \\ H_1 : \text{Les 2 méthodes sont différentes} \end{cases}$$

Nouvelle variable des différences : D = X – X'.
On remplace les 2 séries de données de valeurs x_i et x_i' par $d_i = x_i - x_i'$

<u>**Condition**</u> : D suit $N(\mu_d \,;\, \sigma_d^2)$

Test ⟺ $\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \quad (ou > 0) \end{cases}$ ⟶ *Test paramétrique de moyenne, σ_d^2 inconnue*

La statistique $T = \dfrac{\overline{D} - 0}{S_d/\sqrt{n-1}} \quad suit \quad T_{n-1}$

Test bilatéral : $Rejet\ de\ H_0 \Leftrightarrow T < -t_{1-\alpha/2}^{n-1} \quad ou \quad T > t_{1-\alpha/2}^{n-1}$

Test unilatéral : $Rejet\ de\ H_0 \Leftrightarrow T > t_{1-\alpha}^{n-1}$

Si on accepte H_0, on conclut que les 2 'méthodes' sont identiques avec une proba $(1-\alpha)$

Si on accepte H_1, on conclut que les 2 'méthodes' sont différentes avec un risque α

Exemple : Amphi : n=122 étudiants. Variable : note au partiel.
X note du correcteur 1; X' note du correcteur 2.
Le correcteur 1 note-t'il plus largement au seuil de 5% ?
On suppose que X-X' suit loi *Normale*.

x : 12 13 11 10 5 4 … ⟶ d : 0 -1 2 0 1 -1 ...
x' : 12 14 9 10 4 5 … ⟶ $\bar{d}$ = 1,5 ; s_d = 5 ; n = 122

$$\begin{cases} H_0 : \textit{Les 2 correcteurs notent pareil} \\ H_1 : \textit{Le correcteur 1 note plus large} \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d > 0 \end{cases}$$

D suit *N*(μ_d ; σ_d^2) donc $T = \dfrac{\overline{D} - 0}{S_d/\sqrt{n-1}}$ *suit* T_{n-1}

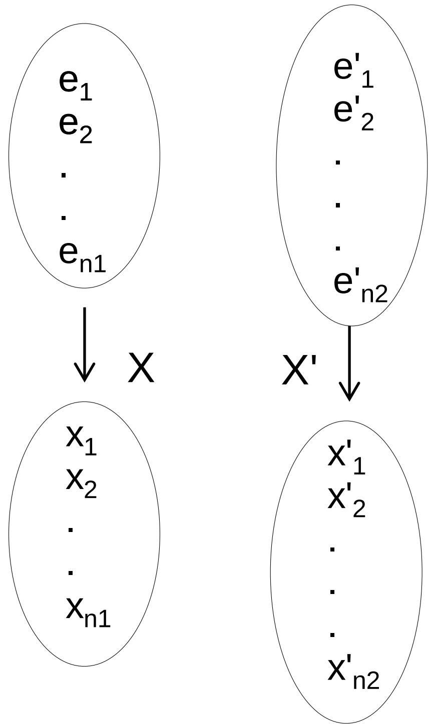
$$T_{calc} = \frac{1{,}5}{5/\sqrt{121}} = 3{,}3 \qquad t_{1-\alpha}^{n-1} \approx u_{1-\alpha} = 1{,}645 \quad (n grand)$$

Test unilatéral : $\textit{Rejet de } H_0 \Leftrightarrow T > t_{1-\alpha}^{n-1}$

Ici, T_{calc}>1,645, donc on accepte H_1. On conclut que le correcteur 1 note plus largement que le correcteur 2 avec un risque de 5% de se tromper.

II. Tests de comparaison de 2 échantillons indépendants :

Modèle



Deux échantillons d'individus différents

Une même variable mesurée

↓

Échantillons indépendants

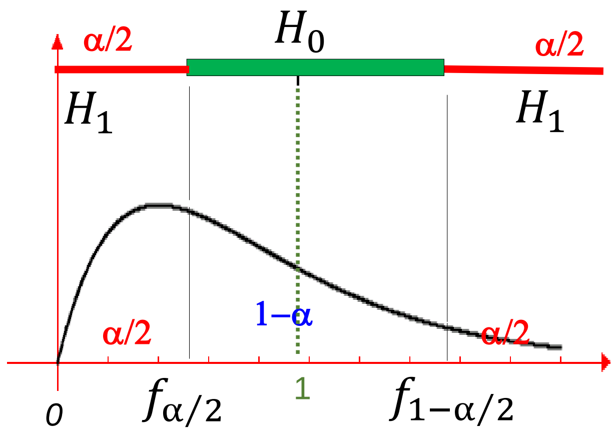
Conditions : X suit $N(\mu_1\,;\,\sigma_1^2)$ *et* X' suit $N(\mu_2\,;\,\sigma_2^2)$ *et elles sont indépendantes.*

Test :
$$\begin{cases} H_0 : \text{Les 2 séries sont identiques} \\ H_1 : \text{Les 2 séries sont différentes} \end{cases} \Leftrightarrow \begin{cases} H_0 : \text{Les 2 échantillons sont identiques} \\ H_1 : \text{Les 2 échantillons sont différents} \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \quad \text{et} \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

a) ***Test de comparaison de 2 variances : test de Fisher***

$$\begin{cases} H_0 \ : \sigma_1^2 = \sigma_2^2 \\ H_1 \ : \sigma_1^2 \neq \sigma_2^2 \end{cases} \Leftrightarrow \begin{cases} H_0 \ : \sigma_1^2/\sigma_2^2 = 1 \\ H_1 \ : \sigma_1^2/\sigma_2^2 \neq 1 \end{cases}$$



La statistique :

$$T = \frac{S'^2_1}{S'^2_2} = \frac{\frac{n_1}{n_1-1}S_1^2}{\frac{n_2}{n_2-1}S_2^2} \quad suit \quad F_{(n_1-1;n_2-1)}$$

Règle de décision :

$$Rejet\ de\ H_0 \quad \Leftrightarrow \quad T < f_{\alpha/2}^{(n_1-1;n_2-1)} \quad ou \quad T > f_{1-\alpha/2}^{(n_1-1;n_2-1)}$$

Remarque : Les tables de Fisher ne donnent que des valeurs de *f* supérieures à 1.

On remarque que $f_{\alpha/2}^{(n_1-1;n_2-1)} = 1/f_{1-\alpha/2}^{(n_2-1;n_1-1)}$

En pratique : On se ramène à T > 1, (si ça n'est pas le cas, échanger le rôle de X et X', ainsi que n_1 et n_2). Alors :

Règle de décision :

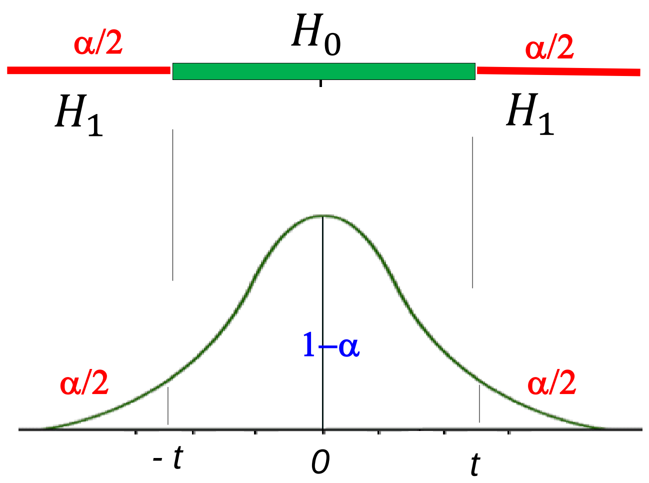
$$Rejet\ de\ H_0 \quad \Leftrightarrow \quad T > f_{1-\alpha/2}^{(n_1-1;n_2-1)}$$

b) ***Test des moyennes : test de Student***

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \quad (ou > 0) \end{cases}$$

$$Si \quad \sigma_1^2 = \sigma_2^2 \ (inconnues)$$

La statistique :



$$T = \sqrt{\frac{n_1 + n_2 - 2}{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\bar{X} - \bar{X}'}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \quad suit \quad T_{(n_1+n_2-2)}$$

Règle de décision :

Test bilatéral : $Rejet\ de\ H_0 \Leftrightarrow T < -t_{1-\alpha/2}^{(n_1+n_2-2)} \quad ou \quad T > t_{1-\alpha/2}^{(n_1+n_2-2)}$

Test unilatéral supérieur : $Rejet\ de\ H_0 \Leftrightarrow T > t_{1-\alpha}^{(n_1+n_2-2)}$

Preuve (construction de la statistique) :

$$\begin{cases} H_0 \ : \ \mu_1 - \mu_2 = 0 \\ H_1 \ : \ \mu_1 - \mu_2 \neq 0 \ \ (ou > 0) \end{cases} \qquad avec \quad \sigma_1^2 = \sigma_2^2$$

X suit $N(\mu_1;\sigma_1^2)$ et X' suit $N(\mu_2;\sigma_2^2)$ par hypothèse et elles sont indépendantes donc :
$\bar{X} + \bar{X}'$ suit $N(\mu_1+\mu_2 ; \sigma_1^2/n_1 + \sigma_2^2/n_2)$.
De plus, la variance globale vérifie : $\sigma^2 = \sigma_1^2 = \sigma_2^2$ *donc :*
$\bar{X} + \bar{X}'$ suit $N(\mu_1+\mu_2 ; \sigma^2 (1/n_1 + 1/n_2))$.

On déduit que la statistique T du test est : $$T = \frac{\bar{X} - \bar{X}'}{\sqrt{\widehat{\sigma^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad suit \quad T_{(n_1+n_2-2)}$$

L'estimateur global de la variance est la moyenne pondérée des estimateurs des variances :

$$\widehat{\sigma^2} = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$$

On conclut alors : $$T = \sqrt{\frac{n_1 + n_2 - 2}{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\bar{X} - \bar{X}'}{\sqrt{n_1 S_1^2 + n_2 S_2^2}} \quad suit \quad T_{(n_1+n_2-2)}$$

c) ***Test des moyennes :*** $Cas\ asymptotique\ (\,n\ \ grand)$

La statistique :

$$T = \frac{\bar{X} - \bar{X}'}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad suit \quad N(0;1)$$

Règle de décision :

Test bilatéral : $Rejet\ de\ H_0 \quad \Leftrightarrow \quad T < -u_{1-\alpha/2} \quad ou \quad T > u_{1-\alpha/2}$

Test unilatéral supérieur : $Rejet\ de\ H_0 \quad \Leftrightarrow \quad T > u_{1-\alpha}$

Synthèse

Test des variances
- → H_1 acceptée alors les 2 échantillons sont différents. STOP
- → H_0 acceptée (variances =), faire le test des moyennes.

Test des moyennes
- → H_1 acceptée alors les 2 échantillons sont différents.
- → H_0 acceptée alors les 2 échantillons sont identiques.

$\boldsymbol{Attention\ au\ risque\ \ \alpha}$

Exemple : Amphi 1 : n=122 étudiants. Amphi 2 : n=101 étudiants.
Variable : note au partie. (X notes amphi 1; X' notes amphi 2)
Les 2 amphis ont-ils eu des notes semblables (en moyenne et en variance) au seuil de 5% ? On suppose que X et X' suivent des lois *Normale*.

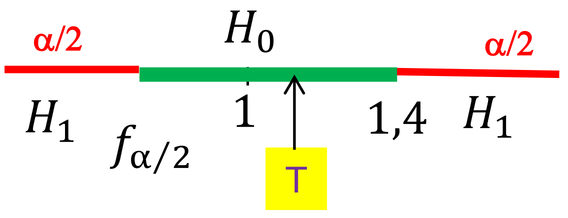
- Les 2 échantillons sont indépendants (individus différents). Lois gaussiennes.

x : 12 11 15 10 9 4 …
x' : 16 13 2 15 9 5 … ⟶ $n_1 = 122$; $\bar{x} = 12$; $s_1'^2 = 5{,}2$
$n_2 = 101$; $\bar{x'} = 13$; $s_2'^2 = 4{,}5$

- Test des variances :

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \qquad T_{calc} = \frac{5{,}2}{4{,}5} = 1{,}155 \qquad f_{1-\alpha/2}^{(n_1-1;n_2-1)} = f_{0{,}975}^{(121;100)} = 1{,}4$$

$$Rejet\ de\ H_0 \quad \Leftrightarrow \quad T < f_{\alpha/2}^{(n_1-1;n_2-1)} \quad ou \quad T > f_{1-\alpha/2}^{(n_1-1;n_2-1)}$$



Ici, $T_{calc} < 1{,}4$ (et T>1) donc on accepte H_0.

- Test des moyennes :

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \qquad n\ grand \qquad T_{calc} = \frac{-1}{0{,}296} = -3{,}378 \qquad u_{1-\alpha/2} = 1{,}96$$

Ici, $T_{calc} < -1{,}96$, donc on accepte H_1. On conclut que les notes moyennes ne sont pas identiques dans les deux amphis, avec un risque de 5% de se tromper.
[Amphi 2 meilleur que amphi 1 car différence des moyennes négative]

III. Test de comparaison de deux proportions :

On note p_1 et p_2 les proportions d'un caractère sur deux populations de taille n_1 et n_2.
On veut déterminer s'il s'agit des deux mêmes popolations.
Condition : Indépendance.

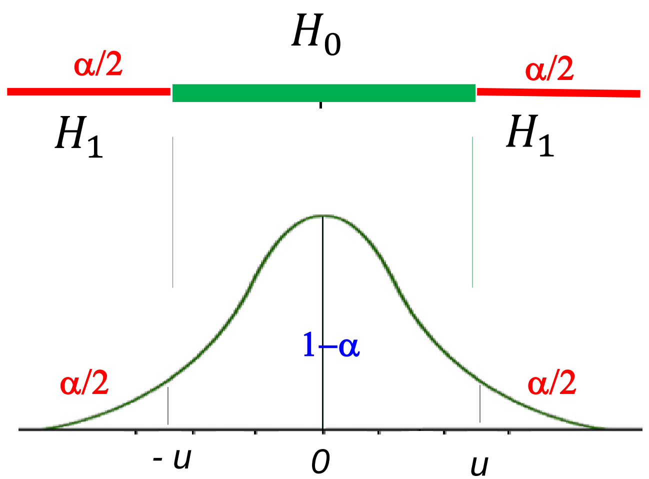
$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases} \Leftrightarrow \begin{cases} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \quad (ou > 0) \end{cases}$$

La statistique : (si n_1 et n_2 sont assez grands)

$$On\ note \quad F = \frac{n_1 F_1 + n_2 F_2}{n_1 + n_2}$$

(moyenne pondérée des fréquences)

$$T = \frac{F_1 - F_2}{\sqrt{F(1-F)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad suit \quad N(0;1)$$



Règle de décision :

Test bilatéral : $Rejet\ de\ H_0 \Leftrightarrow T < -u_{1-\alpha/2} \quad ou \quad T > u_{1-\alpha/2}$

Test unilatéral supérieur : $Rejet\ de\ H_0 \Leftrightarrow T > u_{1-\alpha}$

Exemple : On étudie un certains évènement 'A' dans deux pays différents.
Dans un échantillon de n_1 = 11034 individus, l'évènement A s'est réalisé pour 239.
Dans un échantillon de n_2 = 11037 individus, il s'est réalisé 139 fois.
Peut-on dire que l'évènement A est réduit dans le second pays (au seuil de 5%) ?

Ici $\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{cases}$ avec f_1 = 0,0217 et f_2 = 0,0126; n_1 et n_2 très grands

Attention aux conclusions trop rapides !!!
$f_1 - f_2$ = 0,0091 très petit, il “semblerait“ qu'il n'y ait pas réduction ……..
f_1 / f_2 = 1,72 , il “semblerait“ qu'il y ait 1,7 fois plus d'évènements A dans le pays 1 …….

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = 0{,}017149. \quad \textit{La valeur calculée de la statistique } T \textit{ est} : 5{,}2$$

Test unilatéral supérieur : $\textit{Rejet de } H_0 \Leftrightarrow T > u_{1-\alpha}$ avec $u_{0,95} = 1{,}645$

Ici, T> 1,645, donc on accepte H_1. On conclut que l'apparition de A est réduit dans le second pays , avec un risque de 5% de se tromper.

IV. Tests du khi-deux :

POUR INFORMATION

a) *Test du khi-deux d'ajustement (adéquation) :*

X une variable qualitative (modalités), n la taille de l'échantillon des données.
(c_1 , …, c_r) la répartition en r classes de toutes les valeurs possibles de X
n_i l'effectif observé dans l'échantillon de la classe c_i
p_i la probabilité théorique de la classe c_i pour la loi P

X	c_1 .... c_r	Total
Données	n_1 .... n_r	n
Loi P	np_1 .... np_r	n

$$\begin{cases} H_0 : & X \text{ suit la loi } P \\ H_1 : & X \text{ ne suit pas la loi } P \end{cases}$$

La statistique :

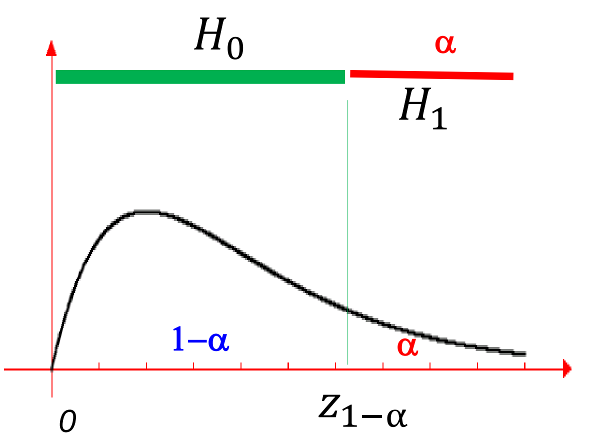
$$T = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \quad suit \quad \aleph_{r-1}$$

S'il y a k paramètres à estimer pour avoir la loi P, alors T suit $\aleph_{r-1-k}$

Règle de décision :

$$Rejet\ de\ H_0 \quad \Leftrightarrow \quad T > z_{1-\alpha}^{r-1}$$

Condition : effectifs $np_i > 5$ et $n > 50$



$$[Remarque : si\ \forall i \quad n_i \approx np_i \ \Rightarrow\ n_i - np_i \approx 0 \ \Rightarrow\ T \approx 0]$$

b) ***Test du khi-deux d'indépendance (contingence) :***

POUR INFORMATION

X et Y deux variables qualitatives de modalités respectives x_i et y_j
n la taille de l'échantillon des données.
n_{ij} l'effectif observé dans l'échantillon croisant les modalités x_i et y_j

X / Y	y_1 .... y_q	Marge ligne
$x_{1.}$ . . $x_{p.}$	n_{11} .... $n_{1q.}$ . . . . . . n_{p1} .... $n_{pq.}$	$n_{1.}$ . . $n_{p.}$
Marge colonne	$n_{.1}$ .... $n_{.q}$	$n.$

$$n_{i.} = \sum_j n_{ij} \qquad n_{.j} = \sum_i n_{ij}$$

Profils–lignes (colonnes) : $\dfrac{n_{ij}}{n_{i.}}$ et $\dfrac{n_{ij}}{n_{.j}}$

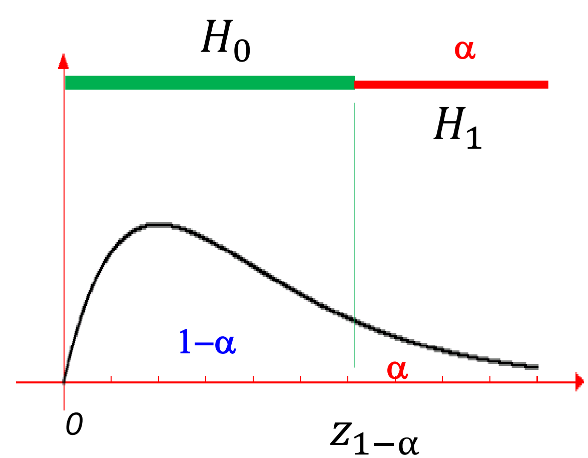
$$\begin{cases} H_0 \;:\; X \text{ et } Y \text{ sont indépendantes} \\ H_1 \;:\; X \text{ et } Y \text{ non indépendantes} \end{cases}$$

La statistique :

$$T = \sum_i \sum_j \frac{\left(n_{ij} - n_{i.}n_{.j}/n\right)^2}{n_{i.}n_{.j}/n} \quad suit \quad \aleph_{(p-1)(q-1)}$$

Règle de décision :

$$Rejet\ de\ H_0 \iff T > z_{1-\alpha}^{r-1}$$



Conditions : effectifs > 5 et n > 50

$$\left[Rem : \; si \;\; \forall i,j \quad \frac{n_{ij}}{n_{i.}} \approx \frac{n_{.j}}{n} \;\Rightarrow\; n_{ij} - n_{i.}n_{.j}/n \approx 0 \;\Rightarrow\; T \approx 0 \right]$$

POUR INFORMATION

Preuve (éléments de la construction de la statistique) :

Si X et Y sont indépendantes *(H_0 vraie), tous les profils-lignes (et colonnes) sont identiques. La connaissance de X ne change pas les distributions conditionnelles de Y.*

Donc, pour tout j : $$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \cdots = \frac{n_{pj}}{n_{p.}}$$

Donc, pour tout i et j : $$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n}$$ *(la moyenne des profils-lignes)*

X et Y sont indépendantes (H_0 vraie) ssi : $n_{ij} = \dfrac{n_{i.}\, n_{.j}}{n}$ *pour tout i et j.*

Ce qui justifie l'emploi de la métrique du Khi-deux pour mesurer l'écart à l'indépendance. D'où la statistique et la loi de ce test (effectifs théoriques sont supérieurs à 5).

$$T = \sum_i \sum_j \frac{\left(n_{ij} - n_{i.}\, n_{.j}/n\right)^2}{n_{i.}\, n_{.j}/n} \text{ suit } \aleph^2_{(p-1)(q-1)}$$

De plus, en développant le carré, on obtient aussi :

$$T = \mathrm{n}\left(\sum_i \sum_j \frac{\left(n_{ij}\right)^2}{n_{i.} n_{.j}} - 1\right)$$

c) ***Test du khi-deux de comparaison d'échantillons (hom***

POUR INFORMATION

X une variable qualitative, de modalités x_i. On a e_1, … e_p : des échantillons d'individus.

Pas d'hypothèse d'indépendance, et de normalité dans ce test.

n la taille de tous les échantillons réunis. n_{ij} l'effectif de x_j dans l'échantillon e_i.

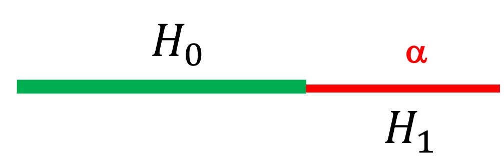
$n_{i.}$ la taille de l'échantillon e_i

X	x_1 .... x_q	Marge ligne
e_1 . . e_p	n_{11} .... n_{1q} . . . . . . n_{p1} .... n_{pq}	$n_{1.}$ . . $n_{p.}$
Marge colonne	$n_{.1}$ .... $n_{.q}$	n

$$\begin{cases} H_0 : \text{Tous les échantillons sont homogènes} \\ H_1 : \text{Tous non homogènes.} \end{cases}$$

La statistique : $$T = \sum_i \sum_j \frac{\left(n_{ij} - n_{i.}n_{.j}/n\right)^2}{n_{i.}n_{.j}/n} = \mathrm{n}\left(\sum_i \sum_j \frac{\left(n_{ij}\right)^2}{n_{i.}n_{.j}} - 1\right) \quad suit \quad \aleph_{(p-1)(q-1)}$$

Règle de décision : $$\text{Rejet de } H_0 \Leftrightarrow T > z_{1-\alpha}^{(p-1)(q-1)}$$



POUR INFORMATION

Exemple 1 :

On souhaite tester l'hypothèse selon laquelle un dé n'est pas truqué.
On lance 600 fois ce dé.

Numéro dé	1	2	3	4	5	6	Total
Observ.	88	109	107	94	105	97	600
P	1/6	1/6	1/6	1/6	1/6	1/16	
Effect. P	100	100	100	100	100	100	600

$$\begin{cases} H_0 : & \textit{X suit la loi Uniforme} \\ H_1 : & \textit{X ne suit pas la loi Uniforme} \end{cases}$$

$$T_{calc} = \frac{(88-100)^2}{100} + \ldots + \frac{(97-100)^2}{100} = 3{,}44$$

Lecture sur $\aleph_5^2$ $\longrightarrow$ p_{val} = 1- P(T<3,44) et 0,1< p<0,5 donc 0,5<p_{val} <0,9

Pour tout risque $\alpha <$ 50%, on accepte H_0 et on conclut que le dé n'est pas truqué.

$$\left[Remarque : \ si \ \alpha = 0{,}05 \ alors \ z_{0,95}^{5} = 11{,}07 \ donc \ T \leq 11{,}07 \ H_0 \ acceptée \right]$$

POUR INFORMATION

Exemple 2 :

Même expérience de Mendel (à l'origine du test). On s'intéresse à deux variables qualitatives sur des pois : Couleur (jaune et vert) et Forme (rond, ridé). On a alors un tableau croisant 2 variables qualitatives (Tableau de contingence, d'effectifs)

X / Y	Rond	Ridé	marges
Jaune	315	101	416
Vert	108	32	140
marges	423	133	556

→

X / Y	Rond	Ridé	marge
Jaune	316,49	99,51	416
Vert	106,51	33,49	140
marge	423	133	556

$$\begin{cases} H_0 \ : \ X \text{ et } Y \text{ indépendantes} \\ H_1 \ : \ X \text{ et } Y \text{ non indépendantes} \end{cases}$$

$$Tableau\ \ théorique\ \ (si\ \ indépendance):$$
$$\frac{n_{1.}n_{.1}}{n} = \frac{423 * 416}{556} = 316{,}49 \quad ect\ldots$$

$$T_{calc} = \frac{(315 - 316{,}49)^2}{316{,}49} + \ldots + \frac{(32 - 33{,}49)^2}{33{,}49} = 0{,}11$$

Lecture sur $\aleph_1^2$ ⟶ p_{val} = 1 - P(T < 0,11) = 1 - 0,016 = 0,984

Pour tout risque α inférieur à 98,4%, on accepte l'hypothèse d'indépendance.

$$\left[Remarque : pour\ \ \alpha = 0{,}05\ \ alors\ \ z_{0,95}^{1} = 3{,}841\ \ donc\ \ T \leq 3{,}841\ \ \rightarrow\ \ H_0\ \ acceptée\right]$$