Tests statistiques

a) Définition de l'hypothèse
b) Risque d'erreur et règle de décision
c) Test de conformité ou d'ajustement
   - Test du Chi 2 (χ^2)
   - Comparaison d'un pourcentage observé à un pourcentage théorique
   - Comparaison d'une moyenne observée à une moyenne théorique
d) Test d'homogénéité
   - Comparaison des moyennes avec le test d'homogénéité
   - Comparaison des pourcentages avec le test d'homogénéité
e) Test d'homogénéité de plusieurs échantillons
f) Test de comparaison des variances
g) Test d'indépendance

Tests Statistiques

Il s'agit d'une étape de la statistique très importante qui sert à éclaircir les décisions qui peuvent être prise dans différents domaines, et ceux avec le plus de précision possible. Pour répondre à des questions de décision il faut utiliser une procédure permettant l'acceptation ou le rejet d'hypothèse posée. Cette procédure s'appelle **test d'hypothèse.**

La confirmation ou l'infirmation d'une hypothèse est toujours fait avec une certaine probabilité que l'on voudra aussi forte que possible.

Définition de l'hypothèse : En pratique, on a 2 types d'hypothèse exclusives H_0 et H_1:

H_0 : hypothèse nulle qui sera rejetée uniquement et qui n'amène pas de changement et d'action à entreprendre dans le cas contraire (si on l'accepte). C'est l'hypothèse á vérifier.

Exemple: H_0: La Personne X est innocente d'un crime.

H_1 : qui s'appelle aussi hypothèse alternative ou contre-hypothèse qui sera acceptée lorsque H_0 est rejetée. C'est une hypothèse qui amène un changement et qui implique une action à entreprendre. **Exemple:** H_1 : la personne X est coupable.

Risque d'erreur et règle de décision :

Avant d'arriver à accepter ou rejeter une hypothèse, il faut étudier la règle de processus menant à une telle décision. Pour établir cette règle de décision il faut tenir compte de la distribution d'échantillonnage de l'estimateur (loi normale) du paramètre à étudier et des risques d'erreur que cette distribution entraîne.

Logiquement on a 4 situations selon H_0 soit vraie ou fausse et selon qu'on a l'accepte ou on la rejette.

Réalité / Décision	**H_0 est vraie**	**H_0 est fausse**
H_0 acceptée	Bonne décision	Erreur de 2ème espèce
H_0 rejetée	Erreur de 1ère espèce	Bonne décision

Dans 2 de ces situations on prend une bonne décision, on doit donc chercher à faire en sorte que les probabilités que ces 2 situations se produisent soit grande autrement dit minimiser la probabilité de commettre des erreurs.

On dit qu'on commet une erreur de 1ère espèce si on rejette H_0 et que H_0 est vraie, on note par α la probabilité de commettre une erreur de 1ère espèce.

On dit qu'on commet une erreur de 2ème espèce si on accepte H_0 et que H_0 est fausse, on note par β la probabilité de commettre une erreur de 2ème espèce.

Erreur de 1ère espèce = α = P (H_0 rejetée / vraie).

Erreur de 2ème espèce = β = P(H_0 acceptée / fausse).

Cette probabilité α s'appelle **le niveau de signification du test** (ou seuil de signification).

Généralement α est fixée á l'avance suivant la nature du problème (généralement on prend α = 5% ou 1%)

I- Test de conformité ou d'ajustement :

I-1. Comparaison d'une répartition observée à une répartition théorique (Test du χ^2) :

On veut savoir si une répartition expérimentale est bien conforme à une répartition théorique par le biais du test du χ^2.

Si on suppose que la répartition de la population suit une loi théorique donnée, on va observer un écart entre l'effectif observé d'une classe et l'effectif théorique de cette même classe. Dans ce cas on est amené à utiliser la somme des écarts quadratique entre l'effectif observé et théorique qui n'est autre que le **χ^2 observé**.

$$\chi^2 = \frac{\sum (O_i - C_i)^2}{C_i}$$ O_i : effectif observé ; C_i : effectif théorique.

Le test de χ^2 se fait selon les étapes suivantes :

- On pose l'hypothèse nulle H_0

  H_0 : il y a conformité entre la répartition théorique et observé,
- Il faut fixer α à l'avance,
- On calcule le χ^2 observé,
- Au seuil α et á un degré de liberté ddl correspondant, on lit sur la table du **χ^2** , le χ^2_α théorique.
- La conclusion sera ainsi :

  a- $\chi^2_{observé} \geq \chi^2_{\alpha\ théorique}$ $\Rightarrow$ H_0 est rejetée.

  b- $\chi^2_{observé} < \chi^2_{\alpha\ théorique}$ $\Rightarrow$ H_0 est acceptée.

<u>Remarque :</u>

Pour appliquer le test χ^2, l'effectif théorique par classe doit au moins égal 5; $\mathbf{C_i \geq 5.}$

Exemple: On a croisé 2 variétés de plantes différentes ayant comme caractère A et B. La 1ère génération est homogène. La 2ème génération fait apparaître 4 phénotypes : AB, Ab, aB, ab

Si les caractères se transmettent selon les lois de Mendel les proportions théoriques de 4 phénotypes sont: 9/16, 3 /16, 3/16, 1/16. L'expérience sur un échantillon de 160 plantes a donnée:

AB : 100 , Ab : 18 , aB : 24 , ab : 18.

Cette répartition est elle conforme aux lois de Mendel à un seuil de signification de 5% ?

Solution : €

H_0 : La répartition observée est conforme aux lois de Mendel avec $\alpha = 0,05$.

Phénotype	AB	Ab	aB	Ab	Total
Proportion théorique	9/16	3/16	3/16	1/16	1
Effectif théorique C_i	9/16 *160 = 90	3/16 * 160 = 30	3/16 * 160 = 30	1/16 * 160 = 10	160
Effectif observé O_i	100	18	24	18	160

$$\chi^2_{observé} = \frac{\sum (O_i - C_i)^2}{C_i} = \frac{(100-90)^2}{90} + \frac{(18-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(18-10)^2}{10}$$

$\chi^2_{observé} = 12,51$

$ddl = K - 1 = 4 - 1 = 3$

$\alpha = 0,05$

$\chi^2_{0,05;3} = 7,815$ (théorique, lu sur la table de χ^2).

$\chi^2_{observé} > \chi^2_{théorique} \Rightarrow H_0$ est rejetée au seuil de significat ion $\alpha = 5\%$

ou bien H_0 est rejetée au seuil de securité de 95%.

I-2. Comparaison d'un pourcentage observé à pourcentage théorique

La comparaison entre un pourcentage (ou proportion) observé **p** sur un échantillon expérimental et le un pourcentage théorique $\mathbf{p_0}$ de la population de l'échantillon est basée sur l'écart réduit **ε**. A savoir $\varepsilon = \frac{|p - p_0|}{\sqrt{\frac{p_0 q_0}{n}}}$ au seuil de signification 5 %.

Si $\varepsilon < 1{,}96$ (≈ 2) la différence n'est pas significative.

Si $\varepsilon \geq 1{,}96$ la différence est significative au seuil de 5%.

Au seuil de 1 % :

Si $\varepsilon < 2{,}576$ ($\approx 2{,}6$) la différence n'est pas significative.

Si $\varepsilon \geq 2{,}576$ la différence est significative.

Exemple: Une race de souris présente des tumeurs spontanées avec un taux parfaitement connu soit $\mathbf{p_0}$ = 20%. Dans une expérience portant sur 100 souris, on observe 34 atteintes, soit **p** = 34%. On demande si la différence entre p_0 et p est significative.

Solution: Ho: pas de différence significative entre p et p_o

$$\varepsilon = \frac{|0{,}34 - 020|}{\sqrt{\frac{0{,}2 \times 0{,}8}{100}}} = 3{,}50$$

$\varepsilon = 3{,}5 > 1{,}96 \Rightarrow$ on rejette H_o , donc la différence est significative entre p et p_o seuil de 5%

Appliquons le même exemple en employant le χ^2

Solution

	Tumeur	Pas de tumeur	total
Effectif théorique $C_i = np$	20%	80%	100%
Effectif observée O_i	34%	66%	100%
% théorique « P »	20%	80%	100%

$$\chi^2 = \frac{\sum (O_i - Ci)^2}{C_i} = \frac{(34-20)^2}{20} + \frac{(66-80)^2}{80} = 12{,}25$$

$$\chi^2_{0{,}05} = 3{,}841$$

Remarque : On remarque que le $\chi_{observé} = \varepsilon^2$

$$12{,}25 = (3{,}50)^2$$

$\chi^{lu}_{0{,}05} = t^2$

$3{,}841 = (1{,}96)^2$

En effet la méthode de comparaison par l'écart réduit et le test du χ^2 sont absolument superposables.

I-3. Comparaison d'une moyenne observée à une moyenne théorique :

Soit à comparer un échantillon expérimental de moyenne $\overline{X}$ à une population dont la moyenne m et l'écart type σ sont connus.

Prenons le cas des grands échantillons où $n \geq 30$, la moyenne $\overline{X}$ suit donc une loi normale

$$\overline{X} \rightsquigarrow N\left(m, \frac{\sigma_{population}}{\sqrt{n}}\right)$$

La transformation t qui correspond à la valeur critique de Student suit une loi normale CR

$$t \rightsquigarrow N(0,1)$$

$$t = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \text{ (variable de student)}$$

$H_0 : m = m_0$ $(\alpha = 5\%)$

Si $|t| < 1{,}96$ $\Rightarrow$ la différence n'est pas significative

$\Rightarrow H_0$ est acceptée.

Si $|t| \geq 1{,}96$ $\Rightarrow$ la différence est significative

$\Rightarrow H_0$ est rejetée.

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 m = 36g.

On demande si la moyenne observée est compatible avec l'hypothèse que la machine fabrique en moyenne des paquets de m_0 = 40g avec un écart type de 18g (α = 5%).

Solution:

$\overline{X} = 36, \quad m_0 = 40$

$$t = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{36 - 40}{\frac{18}{\sqrt{100}}} = 2{,}22$$

$|t| > 1{,}96$ $\Rightarrow$ la différence est significative.

La moyenne observée est différente de la moyenne théorique au seuil α = 5%.

II. Test d'homogénéité :

Supposons qu'on a 2 échantillons pris dans 2 endroits différents. Peut on considérer que ces 2 échantillons proviennent de la même population ou 2 populations différents ?

Le principe de la comparaison consiste à poser H_0.

H_0 : il n'y a pas de différence significative entre les 2 échantillons.

On procède au test au seuil de signification α (ou au seuil de sécurité 1 - α).

Si H_0 est rejetée cela signifie que les 2 populations sont différentes.

Si H_0 est acceptée : il y'a 2 explications possibles :

- Soit les 2 échantillons sont effectivement semblables.
- Soit les 2 échantillons sont réellement différentes, mais la taille des échantillons est insuffisante pour pouvoir mettre la différence en évidence.

Pour pouvoir conclure que 2 populations sont identiques entre elles, il faut comparer les paramètres qui les caractérisent tel que : la moyenne, la variance, % ……

II-1. Comparaison de deux moyennes avec le test d'homogénéité :

Soit 2 échantillons : X_{i1}, avec i = 1,2,3,…..,n_1 de moyenne $\overline{X}_1 = \frac{\sum_{i=1}^{n_1} X_{i1}}{n_1}$

et X_{i2}, avec i = 1,2,3,…..,n_2 de moyenne $\overline{X}_2 = \frac{\sum_{i=1}^{n_2} X_{i2}}{n_2}$

Avant de comparer les moyennes $\overline{X}_1$ et $\overline{X}_2$, on étudie d'abord l'intersection des intervalles de confiance des moyennes m_1, m_2.

- Si n > 30 (grand échantillon)

  L'intervalle de confiance IC :

  IC : $\overline{X}_1 \pm t_\alpha \frac{\sigma_{ech1}}{\sqrt{n_1 - 1}}$ et $\overline{X}_2 \pm t_\alpha \frac{\sigma_{ech2}}{\sqrt{n_2 - 1}}$

  à 5 % t = 1,96

  à 1 % t = 2,58

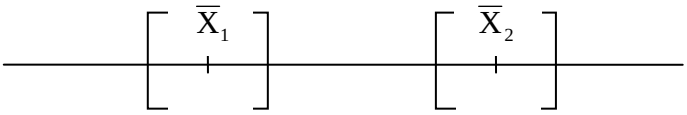
- Si n ≤ 30 (petit échantillon)

  IC : $\overline{X}_1 \pm t_\alpha \frac{\sigma_{ech1}}{\sqrt{n_1 - 1}}$ et $\overline{X}_2 \pm t_\alpha \frac{\sigma_{ech2}}{\sqrt{n_2 - 1}}$

  t : utiliser la table de Student, α, á ddl = n_i -1

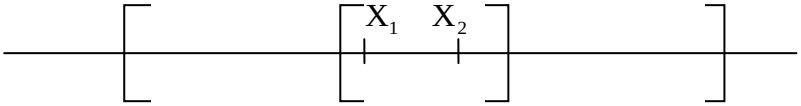
  Trois cas peuvent se présenter:

  a- Intervalles de confiances disjoints : IC (m_1) ∩ IC (m_2) = 0, schématiquement on peut la représenter par la figure suivante :



On conclue qu'il y a une différence signification entre les moyennes des 2 populations.

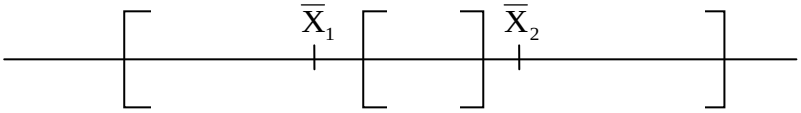
b- Intervalles de confiances non disjoints : $IC(m_1) \cap IC(m_2) \neq 0$



Ou encore $\overline{X}_1 \in IC(m_2)$
$\overline{X}_2 \in IC(m_1)$

Dans ce cas on conclue que la différence entre les 2 moyennes des 2 populations n'est pas significative au seuil de sécurité considéré.

c- Intervalles de confiance non disjoints : $IC(m_1) \cap IC(m_2) \neq 0$



mais $\overline{X}_1 \notin IC(m_2)$
$\overline{X}_2 \notin IC(m_1)$

Dans ce cas pour pouvoir conclure si la différence des 2 moyennes est significative ou pas, on procède au test de comparaison des moyennes en utilisant le test de l'écart réduit **ε**)

- Si n_1 et $n_2 > 30$, on procède ainsi:

1- On pose l'hypothèse $H_0 : m_1 = m_2$, en supposant que les 2 échantillons proviennent de la même population

2- On calcule $\varepsilon = \dfrac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\dfrac{\sigma^2_{ech1}}{n_1} + \dfrac{\sigma^2_{ech2}}{n_2}}}$

3- En Conclusion et au seuil de 5 % si $\varepsilon \geq 1{,}96 \Rightarrow$ on rejette H_0,
et si $\varepsilon < 1{,}96 \Rightarrow$ on accepte H_0.

Exemple: On a fait l'étude sur 2 échantillons de souris qu'il a capturés en 2 endroits différents, on a obtenu les résultats suivants :

Echantillon 1 : $n_1 = 50$, $\overline{X}_1 = 51g$, $\sigma_1^2 = 256g^2$.

Echantillon 2 : $n_2 = 50$, $\overline{X}_1 = 45g$, $\sigma_1^2 = 144g^2$.

Ces souris appartiennent t-elles à la même population ?

Solution:

$H_0 : m_1 = m_2$

- IC (m_1) ?

$$\overline{X}_1 - 1{,}96\frac{\sigma_1}{\sqrt{n_1 - 1}} \le m_1 \le \overline{X}_1 + 1{,}96\frac{\sigma_1}{\sqrt{n_1 - 1}}$$

$$51 - 1{,}96\frac{16}{\sqrt{49}} \le m_1 \le 51 + 1{,}96\frac{16}{\sqrt{49}}$$

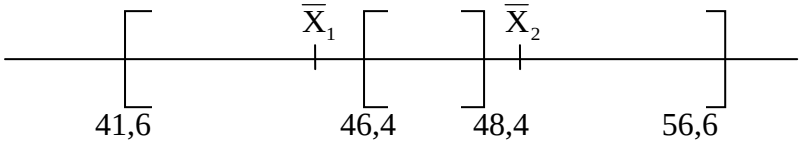
$$m_1 \in [46{,}4;56{,}6]$$

- IC (m_2) ?

$$\overline{X}_2 - 1{,}96\frac{\sigma_2}{\sqrt{n_2 - 1}} \le m_2 \le \overline{X}_2 + 1{,}96\frac{\sigma_2}{\sqrt{n_2 - 1}}$$

$$45 - 1{,}96\frac{12}{\sqrt{49}} \le m_2 \le 45 + 1{,}96\frac{12}{\sqrt{49}}$$

$$m_1 \in [41{,}6;48{,}4]$$



$\overline{X}_1 \notin IC(m_2)$
$\overline{X}_2 \notin IC(m_1)$ les 2 Intervalles de confiance sont non disjoints, on calcule donc **ε**

$$\varepsilon = \frac{|51 - 45|}{\sqrt{\frac{256}{50} + \frac{144}{50}}} = 2{,}48$$

Au seuil de 5 % , $\varepsilon > 1{,}96 \Rightarrow H_0$ est rejetée.

Les 2 populations de souris sont différentes.

- Si n_1 et $n_2 < 30$

$H_0 : m_1 = m_2$ (les 2 échantillons appartiennent à la même population),

Il est montré qu'une bonne estimation de σ (écart – type de la population) est fournie par S^2. Le S^2 remplace donc σ^2.

$$S^2 = \frac{\sum_1^{n_1}\left(X_{1i} - \overline{X}_1\right)^2 + \sum_1^{n_2}\left(X_{2i} - \overline{X}_2\right)^2}{(n_1 - 1)(n_2 - 1)} = \frac{n_1\sigma^2_{ech1} + n_2\sigma^2_{ech2}}{n_1 + n_2 - 2}$$

Au lieu de l'expression de l'écart réduit ε on utilise le **test de Student**, avec

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

En Conclusion, on compare le t observé ainsi calculé avec le $\mathbf{t_\alpha}$ théorique lu à partir de la table de Student, donné en fonction de α et un ddl égal á **ddl = $n_1 + n_2 - 2$** :

- Si $t \geq t_\alpha \Rightarrow$ la différence est significative : H_0 est rejetée c-a-d les 2 échantillons n'appartiennent pas à la même population.
- Si $t < t_\alpha \Rightarrow$ la différence n'est pas significative : H_0 est acceptée, les 2 échantillons proviennent de la même population.

Exemple: Dans une étude d'anesthésie, on compare les effets de 2 somnifères, on à noté les durées de sommeil en minute qui ont suivi les injections d'une dose bien définie.

Somnifère 1: 170, 175, 187, 180, 190, 165, 175, 174, 173, 181.

Somnifère 2: 155, 160, 164, 150, 160, 159, 154, 156, 160, 167, 153, 158.

Solution :

$\overline{X}_1 = 177$, $n_1 = 10.$

$\overline{X}_2 = 158$, $n_2 = 12$ $\qquad S^2 = 38{,}4$

$$t = \frac{|177 - 158|}{\sqrt{38{,}4}\sqrt{\frac{1}{10} + \frac{1}{12}}} = 7{,}2$$

$\alpha = 5\%.$

$ddl = n_1 + n_2 - 2 = 10 + 12 - 2$

$t \approx 2{,}09 \qquad (t \geq t_\alpha)$

La différence est significative : H_0 est rejetée, et le somnifère 1 est plus significatif de point de vue longue durée de sommeil que le somnifère 2.

II-2. Comparaison de 2 pourcentages avec le test d'homogénéité :

Soient 2 échantillons X_1 et X_2 dont les quels le n^{bre} d'individus possédant un certains caractère A sont respectivement K_1 et K_2 d'où le pourcentage (%), $p_1 = \frac{K1}{n_1}$ et $p_2 = \frac{K_2}{n_1}$.

Peut on considérer que ces 2 échantillons sont extraits d'une même population ?

On pose $H_0 : p_1 = p_2$

Comme dans le cas de test de comparaison des moyennes on étudie d'abord l'intersection des intervalles de confiances des 2 échantillons.

$$IC : p_1 \pm t_\alpha \sqrt{\frac{p_1(1-p_1)}{n_1}} \quad , \quad p_2 \pm t_\alpha \sqrt{\frac{p_2(1-p_2)}{n_2}}$$

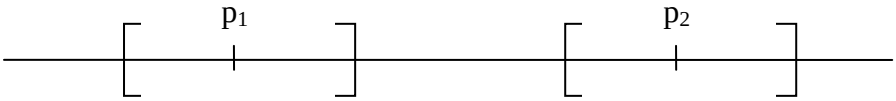
$\alpha = 5\%$ $t_\alpha = 1{,}96$

$\alpha = 1\%$ $t_\alpha = 2{,}6$

Comme précédemment, Trois cas peuvent se présenter:

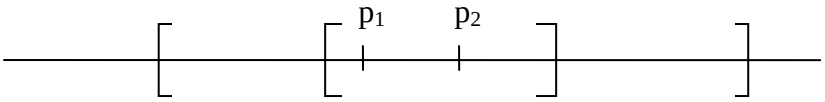
A noter $\hat{p} \rightarrow$ population , $p \rightarrow$ échantillon

a. Intervalles de confiance (IC) disjoints: $IC(\hat{p}_1) \cap IC(\hat{p}_2) = 0$



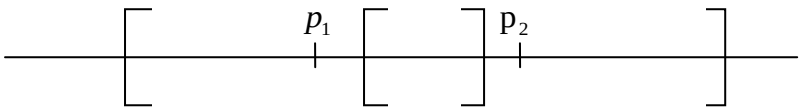
On conclue qu'il y a une différence significative entre les 2 pourcentages

b. Intervalles de confiance non disjoints:



$p_1 \in IC(\hat{p}_2)$
$p_2 \in IC(\hat{p}_1)$ On conclue que la différence entre les 2 % n'est pas significative
$IC(\hat{p}_1) \cap IC(\hat{p}_2) \neq 0$

c. Intervalles de confiance non disjoints:



$p_1 \notin IC(\hat{p}_2)$
$p_2 \notin IC(\hat{p}_1)$
$IC(\hat{p}_1) \cap IC(\hat{p}_2) \neq 0$

Dans ce cas on doit faire le test en utilisant l'écart réduit ε et on a 2 cas :

- n_1 et $n_2 > 30$, p_1 et p_2 pas trop voisin de 0 et 1.

$$p_1 = \frac{K_1}{n_1} \qquad K_1 = n_1 p_1$$

$$p_2 = \frac{K_2}{n_2} \qquad K_2 = n_2 p_2$$

Au total : $K_1 + K_2 = n_1p_1 + n_2p_2$ d'individus qui portent le caractère A dans les 2 échantillons.

On estime le % moyen du caractère A entre les 2 échantillons

$$p = \frac{n_1 p_1 + p_2 n_2}{n_1 + n_2}$$

1. $H_0 : p_1 = p_2$

2. $$\varepsilon = \frac{|p_1 - p_2|}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. Conclusion : pour $\alpha = 5\%$,

- Si $\varepsilon \geq 1{,}96 \Rightarrow$ la différence est significative, H_0 est rejetée les 2 échantillons n'appartiennent pas à la même population.
- Si $\varepsilon < 1{,}96 \Rightarrow$ la différence est significative, H_0 est acceptée.

Exemple: Pour déceler la présence d'une maladie M chez un individu, on a utilisé 2 tests différents sur 2 séries d'observations (2 échantillons):

1ère test : sur 300 personnes présentant effectivement la maladie M, le test 1 a décelé la présence de la maladie chez 243 individus.

2ème test : sur 200 autres malades, le test 2 a décelé la présence de la maladie chez 152 individus.

Peut on admettre que les 2 tests ont un pouvoir de détection sensiblement égal ?

Solution:

Echantillon 1: $n_1 = 300$, $K_1 = 243$.

$$p_1 = \frac{K_1}{n_1} = \frac{243}{300} = 0{,}81$$

Echantillon 2: $n_2 = 200$, $K_2 = 152$.

$$p_2 = \frac{K_2}{n_2} = \frac{152}{200} = 0{,}76$$

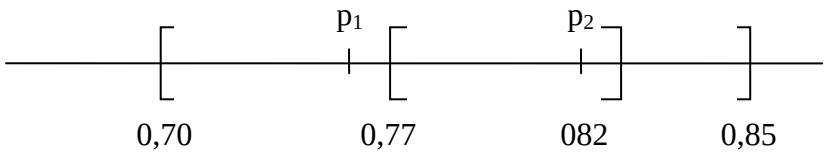
$$IC(\hat{p}_1) = ?$$

$$0{,}81 - 1{,}96\sqrt{\frac{0{,}81 \times 0{,}19}{300}} \le \hat{p}_1 \le 0{,}81 + 1{,}96\sqrt{\frac{0{,}81 \times 0{,}19}{300}}$$

$$IC(\hat{p}_2) = ?$$

$$0{,}76 - 1{,}96\sqrt{\frac{0{,}76 \times 0{,}24}{200}} \le \hat{p}_2 \le 0{,}76 + 1{,}96\sqrt{\frac{0{,}76 \times 0{,}24}{200}}$$

$$IC(\hat{p}_2) = [0{,}70;0{,}82]$$



On doit faire le test

$$p = \frac{K_1 + K_2}{n_1 + n_2} = \frac{152 + 243}{500} = 0{,}79$$

$$\varepsilon = \frac{|0{,}81 - 0{,}76|}{\sqrt{0{,}79 \times 0{,}21\left(\frac{1}{300} + \frac{1}{200}\right)}} = 1{,}35$$

Conclusion :

Pour $\alpha = 5\%$ on a $\varepsilon < 1{,}96 \Rightarrow$ H_0 est acceptée, les 2 tests ont un pouvoir de détection sensiblement égal.

III. Test d'homogénéité de plusieurs échantillons :

Soient plusieurs échantillons d'effectifs n_1, n_2, n_3, …… n_m, et soient K_1, K_2, K_3, …..K_m les effectifs d'individus portant un caractère A.

p_1, p_2, p_3, …..p_m sont les % d'individus portant un caractère A, et q_1, q_2, q_3, …..q_m les % qui ne possèdent pas ce caractère.

$$p_1 = \frac{K_1}{n_1}, p_1 = \frac{K_1}{n_1}, ..., p_m = \frac{K_m}{n_m}$$

$$q_1 = \frac{n_1 - K_1}{n_1}, q_2 = \frac{n_2 - K_2}{n_2}, ..., q_m = \frac{n_m - K_m}{n_m}$$

Tableau des effectifs expérimentaux :

	Présence caractère A	Absence du caractère A	Total
Echantillon 1	K_1	$n_1 - K_1$	n_1
Echantillon 2	K_1	$n_2 - K_2$	n_2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
Echantillon m	K_1	$n_m - K_m$	n_m

Le problème se pose comme suit:

Peut-on considérer que ces échantillons sont extraits d'une même population ?

On pose donc, H_0: « les échantillons proviennent de la même population. »

On estime le % du caractère A dans la population P_0 :

$$p_0 = \frac{K_1 + K_2 + ... + K_m}{n_1 + n_2 + ... + n_m}$$

On calcule les effectifs théoriques pour chaque échantillon C_i:

$C_i = n_i p_0$

$ech_1 \Rightarrow C_1 = n_1 p_0$

⋮

$ech_m \Rightarrow C_m = n_m p_0$

Ainsi on établit le tableau des effectifs théoriques

Tableau des effectifs théoriques :

	Présence caractère A	Absence du caractère A	Total
Echantillon 1	$C_1 = n_1 p_0$	$n_1 - C_1$	n_1
Echantillon 2	$C_2 = n_2 p_0$	$n_2 - C_2$	n_2
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
Echantillon m	$C_m = n_m p_0$	$n_m - C_m$	n_m

On calcule donc le χ^2 observé, comme si on faisait un test de conformité d'une répartition expérimentale à une répartition théorique..

$$\chi^2 = \frac{\sum (K_i - C_i)^2}{C_i}$$

Effectif expérimental: $K_1, K_2, \ldots\ldots\ldots\ldots, K_m$.

Effectif théorique: $C_1, C_2, \ldots\ldots\ldots\ldots, C_m$.

Au seuil de signification $\alpha = 5\%$, et pour un ddl = m – 1 (avec m : n^{bre} d'échantillons):

Si $\chi^2 < \chi^2_\alpha \Rightarrow H_0$ est acceptée.

Si $\chi^2 \geq \chi^2_\alpha \Rightarrow H_0$ est rejetée.

Exemple: Une maladie est traitée dans 4 hôpitaux différents, en appliquant dans chaque hôpital un traitement différent, on a enregistré les observations suivantes:

	Cas de guérison	Cas de non guérison	N^{bre} total des malades traités	% de guérison
Hôpital 1	123	28	151	81,4
Hôpital 2	95	19	114	83,3
Hôpital 3	152	63	215	70,6
Hôpital 4	132	53	185	71,3
Total	502	163	665	75,6

Peut-on considérer que l'efficacité des 4 traitements est la même au seuil $\alpha = 5\%$?

Solution : H_0 : l'efficacité des 4 traitements est la même

$p_0 = \frac{502}{665} = 0{,}756$, et les effectifs théoriques sont résumé dans le tableau suivant :

	Cas de guérison	Cas de non guérison	N^{bre} total des malades traités	% de guérison
	$C_i = n_i p_0$	$n_i - C_i$		
Hôpital 1	144	37	151	75,6
Hôpital 2	86	28	114	75,6
Hôpital 3	162	53	215	75,6
Hôpital 4	140	43	185	75,6

Total	502	163	665	75,6

$$\chi^2 = \frac{(123-144)^2}{144} + \frac{(95-86)^2}{86} + \frac{(152-162)^2}{162} + \frac{(132-140)^2}{140} + \frac{(28-37)^2}{37} + \frac{(19-28)^2}{28} + \frac{(63-53)^2}{53} + \frac{(53-48)^2}{48}$$

$$\chi^2 = 11{,}11$$

$$m = 4-1 = 3$$

$$\chi^2_{0,05;3} \approx 7{,}82$$

Puisque $\chi^2_\alpha < \chi^2$ donc H_0 est rejetée, il y a une différence significative entre les 4 traitements.

IV- Test de comparaison des variances :

Soit á comparer deux variances en posant l'hypothèse nulle H_0

$H_0 : \sigma_1^2 = \sigma_2^2$ on utilise :

1- test de Fischer Snedecor :

$$F_{obs} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \qquad \sigma_1^2 = \frac{\sum (x_i - \overline{X})^2}{2} = \frac{SCE}{n} (\text{échantillon})$$

$$F_{obs} = \frac{\frac{SCE_1}{n_1 - 1}}{\frac{SCE_2}{n_2}}$$

F_α lu à partir de la table de Fischer.

Si $F_{obs} > F_\alpha \Rightarrow H_0$ est rejetée.

2- Test de Bartlett : Il s'agit de tester l'égalité de plusieurs variances

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_p^2$

$$\chi^2_{obs} = \frac{(n-p)\log \hat{\sigma}^2 - \sum_{i=1}^{p} \left[(n-1)\log \sigma_i^2\right]}{1 + \frac{1}{3(p-1)}\left[\sum_{i=1}^{p} \frac{1}{n_i - 1} - \frac{1}{n-p}\right]}$$

$\hat{\sigma}^2 = \frac{SCE}{n-p}$ l'ensemble des échantillons

$\hat{\sigma}_i^2 = \frac{SCE_i}{n_i - 1}$

Il faut noter que l'égalité des variances est une des conditions de l'analyse statistique paramétrique telle que l'analyse de la variance (voir chapitre qui suivent)