Alexia MONSAVOIR
06.23.42.24.30

isaralyon
Une école d'ingénieurs au cœur de la vie

ISARA - 2ème année - STATISTIQUE
Cours : Parties 4 et 5
Mme B. Bottollier-Lemallaz

ECHANTILLONNAGE - ESTIMATION / TEST

PREREQUIS

Avant d'aborder le cours de 2ème année, l'étudiant-e doit être capable de:

- calculer la moyenne, la variance, la SCE et l'écart-type dans un échantillon de taille n,
- d'utiliser des tables statistiques (loi normale centrée réduite, Student, X^2, Fisher),
- définir une hypothèse nulle et de la tester dans le cas d'un test non paramétrique d'indépendance.
- calculer et interpréter les coefficients de régression et de détermination ainsi que la liste les résidus dans le cas d'une régression linéaire simple.

NOTATIONS ET FORMULES DE LA 1ERE PARTIE DU COURS DE 1ERE ANNEE

Dans la population	Dans un échantillon
N nombre d'individus	n nombre d'individus
1) On mesure X: (variable aléatoire quantitative) sur chaque individu	
μ ou E(X) moyenne des X des N individus	m ou $\bar{x}$ moyenne de X des n individus
$\sigma^2(x)$ variance des X dans la population	$s^2(x)$ variance des X dans l'échantillon
σ (x) écart-type dans la population	s (x) écart-type dans l'échantillon
2) On attribut X: (variable aléatoire qualitative) sur chaque individu	
p : proportion d'individu dans la population de taille N ayant la caractéristique X	f = k/n: fréquence d'apparition de la caractéristique X dans l'échantillon de taille n

Les formules à retenir :

Soit une série statistique (xi ; ni) :$\{(x_1 ; n_1), (x_2 ; n_2),\ldots.(x_i ; n_i),\ldots.(x_k ; n_k)\}$ $\quad n = \sum_{i=1}^{k} n_i$

Moyenne $\bar{x}$ ou m	Somme des Carrés des Ecarts :	Variance s^2_x
$\bar{x} = m = \sum_{i=1}^{k} n_i x_i / n$	$SCE_x = \sum_{i=1}^{k} n_i (x_i - \bar{x})^2$ $= (\sum_{i=1}^{k} n_i x_i^2) - n\,\bar{x}^2$ $= SC - n*m^2$	Formule de définition : $s^2_x = SCE_x / n$ Formule du calcul manuel: $s^2_x = (\sum_{i=1}^{k} n_i x_i^2 / n) - \bar{x}^2$ **Ecart-type s_x**

CE QU'IL FAUT SAVOIR DE LA 2EME PARTIE DU COURS DE 1ERE ANNEE

Changements d'origine et d'unité : variable centrée réduite (T)

Si $t_i = \frac{Xi - \mu}{\sigma(x)}$ Alors $\bar{t} = 0$ et $s(t) = 1$

Utilisation des tables statistiques : loi Normale centrée réduite, loi de Student, loi du X^2, loi de Fisher-Snedecor.

NOTATIONS ET FORMULES DE LA 3EME PARTIE DU COURS DE 1ERE ANNEE

I – Régression linéaire et corrélation

Y est la **variable dépendante** à expliquer ou variable de réponse
X est la **variable explicative** ou **variable indépendante** ou encore **régresseur**.

La droite de régression de y en fonction de x (droite de y pour x fixé; Dy/x) a pour équation:

$\hat{y}i = bxi + a$ ou encore $yi = bxi + a + ei$

Les significations d'indicateurs à retenir :
Soit une série statistique double (xi ; yi) :{$(x_1 ; y_1), (x_2 ; y_2),\ldots.(x_i ; y_i),\ldots.(x_n ; y_n)$}

La covariance
Elle mesure le lien entre 2 variables X et Y, covariance nulle = indépendance des variables

Rappel :
Formule de définition:
$cov_{xy} = SPE_{xy} / n$
$SPE_{xy} = \sum_{i=1}^{k} (x_i - \bar{x})(y_i - \bar{y})$
Formule de calcul manuel:
$cov_{xy} = (\sum_{i=1}^{k} x_i y_i / n) - \bar{x}\,\bar{y}$
$cov = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$

Les coefficients de la régression
b : mesure la variation de Y quand X augmente de 1 unité
a : estimation de y pour x = 0

Rappel :
$a = \bar{y} - b\,\bar{x}$
$b = SPE_{xy} / SCE_x$
$= cov_{xy} / s^2_x$
$b = \frac{cov}{s^2}$

Coefficient de détermination R^2
Il mesure la part de la variance totale qui est expliquée par la régression.
Il représente le pourcentage des variations de Y expliquées par les variations de X.

Rappel :
$R^2 = SCE(\hat{y}) / SCE(y)$
$R^2 = SPE_{xy}^2 / (SCE_x * SCE_y)$
$R^2 = cov_{xy}^2 / (s^2_x * s^2_y)$
$0 \leq R^2 \leq +1$
$R^2 = \left(\frac{cov^2}{s^2_x \times s^2_y}\right)^2$

☺ Quelques propriétés bien utiles !!!

$\Sigma ei = 0$ $\quad \Sigma yi = \Sigma \hat{y}i$ $\quad SCE\,y = SCE\,\hat{y} + SCE\,e$

☺ Etude des résidus pour analyser la qualité de la régression
- Normalité des résidus
- Résidus standardisés
- Indépendance des résidus

II – Test du X^2

Il s'agit de vérifier si une série d'effectifs observés et conforme à la distribution d'une série d'effectifs théoriques.
Critère statistique calculé : X^2 calculé
Hypothèse nulle H_0 : il y a conformité entre les 2 séries d'effectifs.
Hypothèse alternative H_1 : il n'y a pas conformité.
Pour un risque de première espèce fixé a priori (α) et un ddl (degré de liberté) donnés on cherche dans la table le critère théorique (X^2 théorique) à ne pas dépasser pour conserver l'hypothèse nulle.
Si X^2 calculé > X^2 théorique on rejette H_0 avec moins de α% de risque de se tromper.
Si X^2 calculé < X^2 théorique on conserve H_0, on ne peut pas mettre en évidence la non conformité.

Partie 4: ECHANTILLONNAGE ESTIMATIONS

OBJECTIFS :

- identifier si un problème fait appel à la théorie de l'échantillonnage ou aux estimations,
- expliquer en ses propres mots et par des graphes la théorie de l'échantillonnage,
- associer les notations aux définitions,
- identifier si un tirage est avec ou sans remise dans un énoncé,
- identifier si la variable aléatoire d'une étude est une moyenne ou une fréquence,
- définir les formules de l'espérance mathématique et de la variance de la variable aléatoire étudiée,
- calculer l'espérance mathématique et l'écart type de la variable aléatoire d'une étude,
- identifier la loi suivie par la variable aléatoire d'une étude,
- calculer les limites de l'intervalle de prédiction dans lequel doit se trouver la variable aléatoire de l'échantillon de taille n à partir des paramètres connus de la population pour un risque donné,
- faire la distinction entre estimateur et estimation au niveau des définitions et des notations,
- définir les propriétés des estimateurs,
- définir les formules des estimateurs ponctuels des moyennes, fréquence, variance et écart-type,
- calculer les limites de l'intervalle d'estimation dans lequel doit se trouver le paramètre de la population à partir de paramètres connus dans l'échantillon de taille n et pour un risque donné,
- calculer les marges d'erreurs relative et absolue d'un intervalle,
- calculer la taille n d'un échantillon à prélever pour établir un intervalle d'estimation en fonction d'un risque d'erreur et d'une précision donnés.
- formuler en toute rigueur les conclusions relatives aux intervalles,

Dans cette partie nous allons pouvoir répondre aux questions suivantes :

- ✓ Que peut-on attendre du paramètre d'un échantillon aléatoire issu d'une population de paramètres connus?
- ✓ Comment un paramètre de la population mère peut-il être estimé à partir des observations d'un échantillon aléatoire?

I - Théorie de l'échantillonnage

Supposons que l'on dispose de la liste de toutes les unités qui constituent une population, sans omission ni répétition. Cette liste constitue une **base de sondage.** On peut attribuer à chaque individu un numéro unique puis prélever par tirage au sort n individus pour constituer un **échantillon aléatoire** où chaque unité de la population a une probabilité connue, non nulle d'être choisie.
On peut construire un échantillon aléatoire comme suit :

Tirage sans remise : (tirage exhaustif) les unités tirées successivement ou ensemble ne sont pas remises dans la population. C'est le cas lorsque le **taux de sondage** $\frac{n}{N}$ ***100 est supérieur à 10%.**

Dans cette situation on utilisera le coefficient d'exhaustivité $\mathbf{K = \frac{N - n}{N - 1}}$.

Tirage avec remise : (tirage indépendant) chaque unité tirée au hasard dans la base de sondage est observée puis remise à la population avant qu'une autre unité soit tirée. C'est le cas lorsque **N n'est pas défini comme un nombre fini** ou **si le taux de sondage est inférieur 10%.** Dans cette situation on utilisera le coefficient d'exhaustivité **K = 1** dans les calculs.

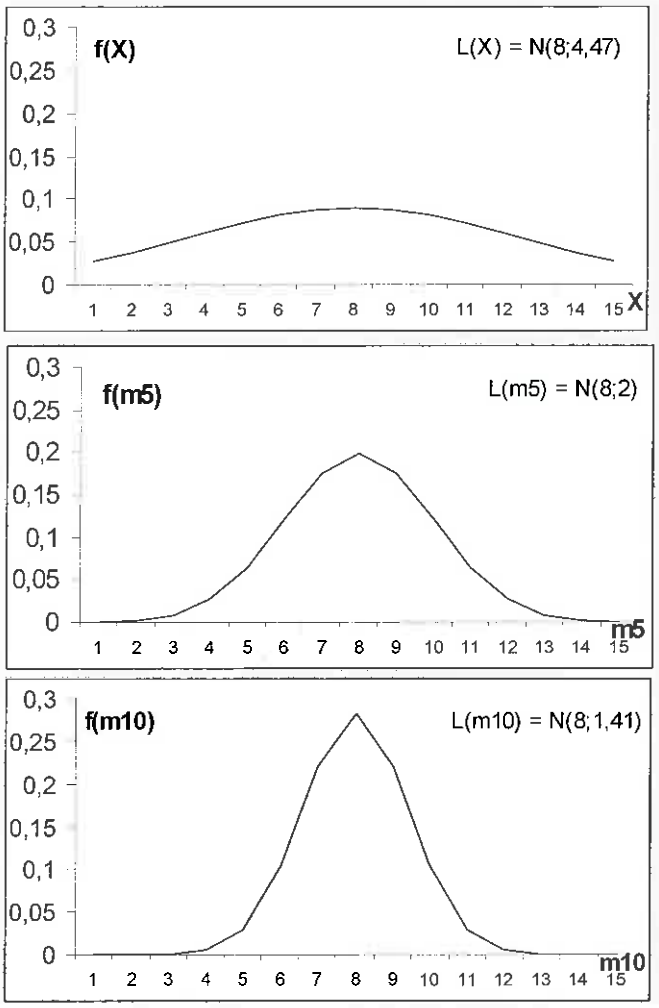
Dans ce cours nous n'envisagerons que **l'échantillonnage au hasard simple**, méthode pour laquelle tous les échantillons possible de même taille ont la même probabilité d'être choisis et tous les individus de la population ont une chance égale de faire partie de l'échantillon.

I – 1 - Distribution d'échantillonnage de la moyenne

Soit une population dans laquelle la variable aléatoire quantitative X étudiée est centrée sur $E_x = \mu$ et dont la variance est égale à σ^2_x
On considère tous les échantillons de même taille n qui sont issus de cette population (q échantillons).
Sur chaque échantillon on mesure la moyenne m_j.
On obtient donc la série des q moyennes : $m_1, m_2, \ldots\ldots m_j, \ldots m_q$.
E_m moyenne des q moyennes m_j des q échantillons de taille n.
σ^2_m variance des q moyennes m_j des q échantillons de taille n.
σ_m écart-type des q moyennes m_j des q échantillons de taille n.

La variable aléatoire est la moyenne arithmétique m_j de la variable étudiée dans les échantillons de taille n.

Règle de l'approximation normale : " dans les échantillons aléatoires de taille n, la moyenne de l'échantillon m varie autour de la moyenne de la population μ avec un écart-type égal à $\sigma_m = \sigma_x / \sqrt{n}$. Donc, quand n augmente, la distribution d'échantillonnage de m est de plus en plus concentrée autour de son objectif μ et devient de plus en plus proche de la distribution normale de Gauss. " (Consulter les courbes qui suivent).



Espérance mathématique : $E_m = \mu = E_x$

écart-type : $\sigma_m = \dfrac{\sigma_x}{\sqrt{n}} \sqrt{K}$

Si $L(X) = N(\mu ; \sigma_x)$ **ou** si $L(X) \neq N$ avec $n \geq 30$

Alors $L(m_n) = N(E_m ; \sigma_m)$

Et on peut écrire pour α donné:

$$P[\mu - t_{(1-\alpha/2)}\, \sigma_m < m < \mu + t_{(1-\alpha/2)}\, \sigma_m] = (1 - \alpha)$$

I – 2 - Distribution d'échantillonnage des fréquences

Soit une population dans laquelle la variable aléatoire qualitative X étudiée est centrée sur $E_x = p$ et dont la variance est égale à $\sigma_x = np(1-p)$. Dans cette situation X = 1 ou 0 selon que l'individu observé possède ou non la caractéristique X.
On considère tous les échantillons de taille n qui sont issus de cette population (q échantillons).
Sur chaque échantillon on mesure la fréquence $f_j = k_j / n$
On a donc la série des fréquences : $f_1, f_2, \ldots f_j \ldots f_q$

La variable aléatoire est la fréquence de la variable observée dans l'échantillon de taille n.

E(f) moyenne des q fréquences f_j des q échantillons de taille n.
σ^2_f variance des q fréquences des q échantillons de taille n.
σ_f écart-type des q fréquences des q échantillons de taille n.

Dans les échantillons aléatoires de taille n, la fréquence f de l'échantillon varie autour de la proportion p de la population avec un écart-type σ_f

Donc, quand n augmente, la distribution d'échantillonnage de f est de plus en plus concentrée autour de p et devient plus proche de la loi Normale.

Espérance mathématique : $E_f = p$

écart-type : $\sigma_f = \sqrt{\frac{p(1-p)}{n}} \cdot \sqrt{K}$ ← coef d'exhaustivité

Si $\left| \sqrt{\frac{p}{1-p}} - \sqrt{\frac{1-p}{p}} \right| / \sqrt{n} \leq 0.34$ et $n > 5$ ④ **ou** Si $n\,p$ **et** $n\,(1-p) \geq 5$

Alors $L(f_n) = N(E_f ; \sigma_f)$
Et on peut écrire pour α donné:

$$P\left[p - t_{(1-\alpha/2)}\,\sigma_f < f < p + t_{(1-\alpha/2)}\,\sigma_f\right] = (1-\alpha)$$

④ $\left| \sqrt{\frac{p}{q}} - \sqrt{\frac{q}{p}} \right| / \sqrt{n} \leqslant 0{,}34$

↑ autre condition

II - Les estimations

II – 1 - Définitions et propriétés → à savoir pour le QCM

On appelle *estimateur,* toute fonction statistique des valeurs observées sur un échantillon utilisée pour estimer un paramètre inconnu de la population. Toute valeur prise par l'estimateur est une *estimation* du paramètre de la population. Un estimateur est donc une variable aléatoire dont les propriétés sont les suivantes :

Un estimateur est **sans biais** si son espérance mathématique est égale au paramètre que l'on cherche à estimer quelque soit n.

Un estimateur est **convergent** (ou correct) si son espérance mathématique tend vers le paramètre que l'on cherche à estimer et si sa variance tend vers 0 lorsque la taille de l'échantillon tend augmente

Un estimateur est **absolument correct** s'il est sans biais et si sa variance tend vers 0 lorsque la taille de l'échantillon augmente

Un estimateur est **efficace** s'il est absolument correct et si sa variance est minimale parmi les estimateurs sans biais possibles.

II – 2 - Estimations ponctuelles :

L'estimation ponctuelle est la réalisation d'un estimateur dans un échantillon donné.

C'est donc la valeur que l'on attribue au paramètre inconnu que l'on cherche à définir à l'aide d'un échantillon, si l'on doit fournir une valeur unique de ce paramètre.

Estimation d'une moyenne : $\hat{\mu} = m$

Estimation d'une proportion : $\hat{p} = f = \frac{k}{n}$

Estimation d'une variance : $\hat{\sigma}^2_x = \frac{ns^2(x)}{n-1} = \frac{SCEx}{n-1}$

Estimation d'un écart-type : $\hat{\sigma}_x$

II – 3 - Estimation par intervalle de confiance

Il s'agit de calculer une fourchette de valeurs estimées d'un paramètre inconnu d'une population mère, généralement centrée sur l'estimation ponctuelle de ce paramètre, et dont l'amplitude est définie par le choix d'un coefficient de confiance.

Si la loi de probabilité de l'estimateur est connue, il est possible de déterminer, à partir de la valeur calculée sur un échantillon (ou estimation) des limites entre lesquelles se trouve presque certainement comprise la vraie valeur de la caractéristique. Quand n est suffisamment grand, l'estimateur bien choisi d'une caractéristique se distribue généralement suivant une loi voisine de la loi Normale. Celle-ci est alors définie par sa moyenne et son écart-type. Soient θ la vraie valeur inconnue et d l'estimateur. Si l'on peut déterminer l'écart type σ (d) on peut affirmer que l'intervalle $[d - t_{1-\alpha/2}\ \sigma_d ; d + t_{1-\alpha/2}\ \sigma_d]$ a une probabilité égale à α de ne pas contenir la vraie valeur θ.

Tout intervalle de confiance à $(1 - \alpha)\%$ peut être reformulé sous une forme unilatérale en plaçant toute le risque d'erreur de $\alpha\%$ d'un seul côté. Ceci signifie qu'on est beaucoup plus exigeant d'un côté alors que l'on reste très vague de l'autre.

II – 3 – 1 - Estimation d'une moyenne ; variance de la population connue.

Si $L(X) = N(\mu ; \sigma_x)$ **ou** si $n > 30$

Alors $L(m_n) = N(E_m ; \sigma_m)$

Et on peut écrire pour α donné:

$P[m - t_{1-\alpha/2}.\sigma_m < \mu < m + t_{1-\alpha/2}.\sigma_m] = 1 - \alpha$	avec $\sigma_m = \frac{\sigma_x}{\sqrt{n}}\sqrt{K}$

II – 3 – 2 - Estimation d'une moyenne; variance de la population inconnue

Si $L(X) = N(\mu ; \hat{\sigma}_x)$ **et** si $n < 30$

Alors $L(m_n) = S(\nu)(E_m ; \sigma_m)$

Et on peut écrire pour α donné:

$P[m - t_{1-\alpha/2}(\nu).\sigma_m < \mu < m + t_{1-\alpha/2}(\nu).\sigma_m] = 1 - \alpha$ ↳ dans la table de student	avec $\nu = ddl = n - 1$ et $\sigma_m = \frac{\hat{\sigma}_x}{\sqrt{n}}\sqrt{K} = \frac{s_x}{\sqrt{n-1}}\sqrt{K}$

N.B. Si le ddl est supérieur à 30 on utilisera une loi normale. (les prises vont jusqu'à 90...)

II – 3 – 3 - Estimation d'une proportion

Si n prélèvements indépendants et si N est de taille finie le prélèvement doit être avec remise.

On peut écrire pour α donné:

$$P[p_1 < p < p_2] = (1 - \alpha)$$

ça n'existe pas sans remise...

$n < 100$	$n \geq 100$ et $0.1 \leq f \leq 0.9$	$n \geq 100$ et $f \leq 0.1$
$p_1 = \frac{k}{k+(n-k+1)*F_{1-\alpha/2}(\nu_1,\nu_2)}$ * avec $\nu_1 = 2(n-k+1)$ et $\nu_2 = 2k$ $p_2 = \frac{(k+1)*F_{1-\alpha/2}(\nu_1,\nu_2)}{n-k+(k+1)*F_{1-\alpha/2}(\nu_1,\nu_2)}$ * avec $\nu_1 = 2(k+1)$ et $\nu_2 = 2(n-k)$ Remarque : Ne pas apprendre cette formule mais savoir l'appliquer * loi de Fisher	⇒ cas d'une loi normale $p_1 = f - t_{1-\alpha/2} * \sigma(f)$ avec $\sigma_f = \sqrt{\frac{f(1-f)}{n}}.\sqrt{K}$ $p_2 = f + t_{1-\alpha/2} * \sigma(f)$ Remarque : pour cette formule on peut se trouver dans la situation sans remise. Elle est à connaître !	$p_1 = \frac{1}{2n} * X^2_{\alpha/2}(\nu)$ * $\nu = 2k$ $p_2 = \frac{1}{2n} * X^2_{1-\alpha/2}(\nu)$ * $\nu = 2k + 2$ Remarque : Ne pas apprendre cette formule mais savoir l'appliquer * loi du χ^2

II – 3 – 4 - Estimation d'une variance

Si on a n prélèvements indépendants et si la distribution de la variable suit une loi Normale dans la population on peut écrire pour α donné:

II-3-4-Estimation d'une variance

$$P\left[\frac{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}{X^2_{1-\alpha/2}(\nu)} < \sigma^2_x < \frac{\sum_{i=1}^{i=n}(x_i-\bar{x})^2}{X^2_{\alpha/2}(\nu)}\right] = 1-\alpha \qquad \nu = n-1$$

II – 3 – 5 - Estimation d'un écart-type

Si on a n prélèvements indépendants (n>15) et si la distribution de la variable suit une loi Normale dans la population. Les limites de l'intervalle de confiance d'un écart-type sont les racines carrées des limites correspondantes de l'intervalle de confiance de la variance.

$$\sqrt{P_{v_1}} < \sigma_x < \sqrt{P_{v_2}}$$

II – 3 – 6 - Précision et taille d'échantillon

La précision d'une mesure ou marge d'erreur peut s'apprécier en termes absolus ou en termes relatifs. Si la précision d'une estimation est fixée on pourra calculer la taille de l'échantillon nécessaire à l'estimation.

La précision absolue d'une estimation est utilisée dans le cas d'une estimation par intervalle de confiance d'un paramètre; elle est égale à la demi-différence entre les limites supérieure et inférieure de l'intervalle

Marge d'erreur absolue = ME abs = $t.\sigma_m$ ou $t.\sigma_f$

La précision relative est égale au rapport entre la précision absolue et la valeur sur laquelle est centré l'intervalle (multiplié par 100 pour un pourcentage)

Marge d'erreur relative = ME rel = $100.t.\sigma_m/m$ ou $100.t.\sigma_f/f$ ou $100.t.\sigma_m/\mu$ ou $100.t.\sigma_f/p$

$$ME_{rel} = ME_{abs} \times \frac{100}{m}$$

Partie 5: LES TESTS D'HYPOTHESE

OBJECTIFS

- identifier si un problème relève de la théorie des tests,
- identifier la variable aléatoire d'un test,
- identifier la loi de la variable aléatoire d'un test,
- calculer l'espérance mathématique et l'écart type de la variable aléatoire d'un test,
- définir les hypothèses nulle et alternative d'un test en fonction de la question posée,
- calculer 2 paramètres parmi les 4 possibles : n, Π, α, β dans le cas de l'acceptation ou du rejet d'un lot,
- interpréter concrètement les risques de première et deuxième espèce,
- choisir le bon critère statistique à calculer en fonction des éléments dont on dispose dans le problème,
- calculer et interpréter le critère statistique,
- déterminer la p-value dans le cas d'un critère statistique calculé qui suit une loi normale,
- identifier le critère statistique théorique (choix de la table, du ddl, de la probabilité)
- formuler une conclusion rigoureuse au test,

I - Généralités et définitions

Un test d'hypothèse consiste à définir une règle de décision concernant la validité d'une hypothèse portant sur la valeur d'une caractéristique (moyenne, variance, proportion...) d'une distribution dans une population dont on observe un échantillon aléatoire.
La procédure générale d'un test comprend les éléments suivants:

1. définir la variable aléatoire étudiée, sa loi, et les paramètres connus,
2. définir l'hypothèse nulle Ho et son alternative H_1 ; un test statistique est par nature négatif,
3. fixer à priori une valeur α pour le risque de refuser Ho alors qu'elle serait vraie ; accepter Ho ne signifie pas que cette hypothèse est vraie, mais seulement que les observations disponibles n'ont pas permis de mettre en évidence qu'elle n'était pas vraie,
4. définir un critère statistique dont on connaît la loi quand Ho est vraie,
5. définir, à l'aide de la table statistique adéquate, une région critique de rejet de H_0 telle que αsoit la probabilité, si Ho est vraie, pour que le critère statistique appartienne à cette région,
6. énoncer la règle de décision correspondant à la valeur numérique prise par le critère statistique, à savoir rejeter Ho si ce critère est dans la région critique de rejet, ou accepter Ho si le critère est dans la région d'acceptation, région complémentaire de la précédente.

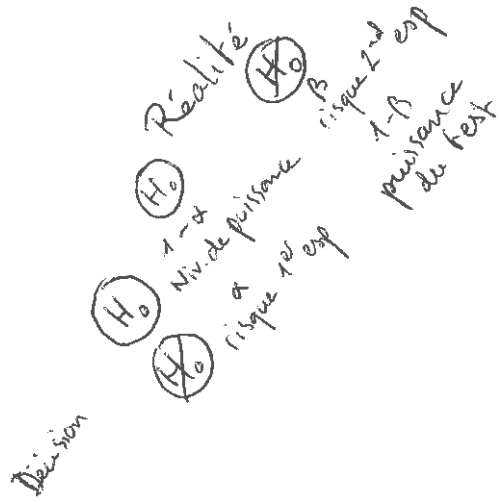
Rejeter Ho alors que celle-ci est vraie se fait avec un risque de 1ère espèce α.
Accepter Ho alors que celle-ci est fausse se fait avec un risque de 2ème espèce β.

Remarque : Si α n'est pas fixé on devra calculer la p-value (risque réel de se tromper si on décide de rejeter H0) en utilisant la table 1, cette situation n'est possible, pour vous, qu'avec une variable aléatoire qui suit une loi normale, dans ce cas les étapes 3 et 6 n'existent pas et on formule la conclusion en fonction de la p-value.

Attention! On accepte H0 parce que l'on n'a pas pu mettre en évidence que H1 soit vraie!

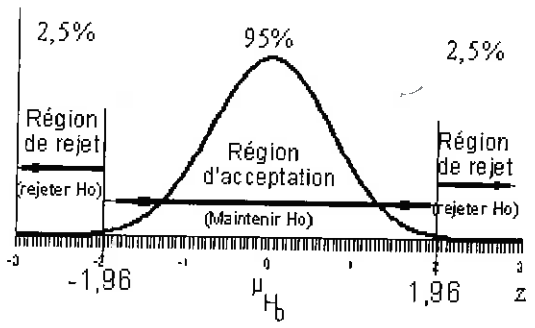
α : risque de se tromper en rejetant H_0
β : risque de se tromper en gardant H_0

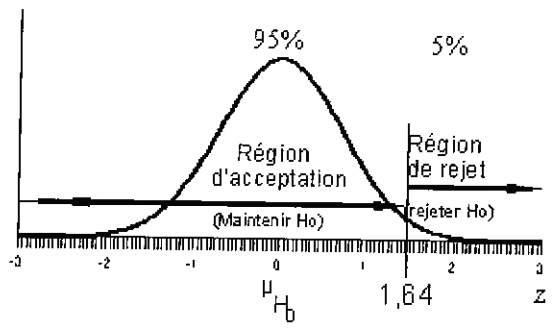
si on a pas α → calcul p-value



TEST BILATÉRAL	TEST UNILATÉRAL
Lorsqu'on ne peut spécifier de direction particulière pour l'hypothèse alternative, on dit que le test est <u>bilatéral</u> $H_0 : \mu = p$	Lorsqu'on peut spécifier une direction particulière pour l'hypothèse alternative, on dit que le test est <u>unilatéral</u> $\mu \geq \leq p$.
Dans ce cas, les hypothèses sont de la forme H_0: θ = valeur présumée vs H_1: $\theta \neq$ valeur présumée (où θ est le paramètre)	Dans ce cas, les hypothèses sont de la forme H_0: θ = valeur présumée vs H_1 : $\theta <$ valeur présumée (unilatéral à gauche) ou H_0: θ = valeur présumée vs H_1: $\theta >$ valeur présumée (unilatéral à droite)
Dans ce cas, il importe peu que le paramètre soit plus grand ou plus petit ce qui compte, c'est qu'il diffère de la valeur supposée en hypothèse, et c'est là la *seule conclusion possible* quand l'hypothèse nulle est rejetée.	Dans ce cas, le rejet de l'hypothèse nulle permet de conclure que la valeur du paramètre est, respectivement, inférieure ou supérieure, à la valeur présumée
Dans ce type de test, il y a deux régions de rejet, situées aux extrémités de la distribution et chacune est d'aire $\alpha/2$	Dans ce type de test, il y a une seule région de rejet, située du côté spécifié par l'hypothèse alternative et d'aire α

Graphiquement, on a par exemple (α = 0,05)





		Réalité	
		H_0 vraie	H_0 fausse
Décision	Non-rejet de H_0	Niveau de confiance $1 - \alpha$	Rejet à tort de H_1 risque de second espèce β
	Rejet de H_0	Rejet à tort de H_0 risque de première espèce α	Puissance du test $1 - \beta$

II - Choix entre deux paramètres

Les paramètres que nous étudierons sont la moyenne et la fréquence. Dans les 2 cas, le problème consistera à décider de quelle population est extrait un échantillon donné, le nombre de populations possibles étant réduit à 2.

II – 1 - Choix entre 2 moyennes

Soit un échantillon de taille n et de moyenne m. La population dont il est issu est-elle centrée sur $\mu = \mu_1$ ou sur $\mu = \mu_2$?

$H_0 : \mu_1 = \mu$ hypothèse rejetée avec un risque de 1ère espèce α

$H_1 : \mu_2 = \mu$ hypothèse rejetée avec un risque de 2ème espèce β

Π = valeur critique au dessus de laquelle on rejette H_0 (si $m > \Pi$ pour $\mu_1 < \mu_2$)		Π = valeur critique au-dessous de laquelle on rejette H_0 (si $m < \Pi$ pour $\mu_1 > \mu_2$)	
$t_{1-\alpha} = \frac{\pi - \mu_1}{\sigma_1(m)}$	$t_\beta = \frac{\pi - \mu_2}{\sigma_2(m)}$	$t_\alpha = \frac{\pi - \mu_1}{\sigma_1(m)}$	$t_{1-\beta} = \frac{\pi - \mu_2}{\sigma_2(m)}$

II – 2 - Choix entre 2 fréquences

Soit un échantillon de taille n sur lequel on observe la fréquence f d'apparition d'un caractère. La population dont il est issu est-elle centrée sur $p = p_1$ ou sur $p = p_2$?

$H_0 : p_1 = p$ hypothèse rejetée avec un risque de 1ère espèce α

$H_1 : p_2 = p$ hypothèse rejetée avec un risque de 2ème espèce β

Π = valeur critique au dessus de laquelle on rejette H_0 (si $f > \Pi$ pour $p_1 < p_2$)		Π = valeur critique au dessous de laquelle on rejette H_0 (si $f < \Pi$ pour $p_1 > p_2$)	
$t_{1-\alpha} = \frac{\pi - p_1}{\sigma_1(f)}$	$t_\beta = \frac{\pi - p_2}{\sigma_2(f)}$	$t_\alpha = \frac{\pi - p_1}{\sigma_1(f)}$	$t_{1-\beta} = \frac{\pi - p_2}{\sigma_2(f)}$

III - Comparaison d'une moyenne à une valeur donnée

Soit un échantillon de taille n et de moyenne m. La population dont il est issu est-elle centrée sur $\mu = a$?

III – 1 - Variance de la population connue

$H_0 : \mu = a$

$H_1 : \mu \neq a$ (bilatéral) ou $\mu < a$ (ou $\mu > a$) (unilatéral)

Variable aléatoire: m_n

Loi de la v.a: $L(m_n) = N(E_m ; \sigma_m)$ si $L(X) = N(E_X ; \sigma_x)$ ou si $n > 30$

Critère statistique calculé : $t_{calculé} = \frac{(m - E_m)}{\sigma_m}$	avec $\sigma^2_m = \frac{\sigma^2_x}{n} * K$
Critère statistique théorique : $t_{1-\alpha/2}$ (bilatéral) ou $t_{1-\alpha}$ (unilatéral) lu dans la table 2 pour α donné a priori	

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t$ théo) on a un risque inférieur à α% de se tromper.

Autre méthode : On peut calculer le risque réel de se tromper (α calculé), ce risque s'appelle la p-value.

p value (bilatéral) = $2*P(T > |t\ calculé|)$ et p value (unilatéral) = $P(T > |t\ calculé|)$

III – 2 - Variance de la population inconnue

$H_0 : \mu = a$

$H_1 : \mu \neq a$ (bilatéral) ou $\mu < a$ (ou $\mu > a$) (unilatéral)

Variable aléatoire: m_n

Loi de la v.a: $L(m_n) = S(\nu)(E_m ; \sigma_m)$ si $L(X) = N(E_X ; \hat{\sigma}_x)$ et si $n < 30$

Critère statistique calculé : $t_{calculé} = \frac{(m - E_m)}{\sigma_m}$	avec $\sigma^2_m = \frac{s^2_x}{n-1} * K = \frac{\hat{\sigma}^2_x}{n} * K$ et $\nu = n-1$
Critère statistique théorique : $t_{1-\alpha/2}(\nu)$ (bilatéral) ou $t_{1-\alpha}(\nu)$ (unilatéral) lu dans la table de Student pour α donné a priori. Si ddl >30 alors t est lu dans la table de la loi Normale.	

A ddl>30

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t$ théo) on a un risque inférieur à α% de se tromper.

IV - Comparaison d'une proportion à une valeur donnée

Soit un échantillon de taille n sur lequel on observe la fréquence f d'apparition d'un caractère. La population dont il est issu est-elle centrée sur p = a ?

$H_0 : p = a$

$H_1 : p \neq a$ (bilatéral) ou $p < a$ (ou $p > a$) (unilatéral)

Variable aléatoire: f_n

Si la loi de la v.a: $L(f_n) = N(E_f ; \sigma_f)$

Critère statistique calculé : $t_{calc} = \frac{(f - E_f)}{\sigma_f}$	avec $\sigma^2_f = (p(1-p)/n) * K$ et $E_f = p$
Critère statistique théorique : $t_{1-\alpha/2}$ (bilatéral) ou $t_{1-\alpha}$ (unilatéral) lu dans la table 2 pour α donné a priori.	

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t$ théo) on a un risque inférieur à α% de se tromper.

Autre méthode : On peut calculer le risque réel de se tromper (α calculé), ce risque s'appelle la p-value.

p value (bilatéral) = $2*P(T > |t\ calculé|)$ et p value (unilatéral) = $P(T > |t\ calculé|)$

V - Comparaison d'une variance à une valeur donnée

$H_0 : \sigma^2 = \sigma^2_0$
$H_1 : \sigma^2 \neq \sigma^2_0$ (bilatéral) ou $\sigma^2 < \sigma^2_0$ ($\sigma^2 > \sigma^2_0$) (unilatéral)

Variable aléatoire: SCE / σ^2_0
Loi de la v.a: loi du X^2 avec $\nu = n - 1$

Critère statistique calculé : $X^2_{calc} = \frac{SCEx(\bar{x})}{\sigma^2_0}$

Conclusion : Pour α donné a priori, on accepte l'hypothèse nulle si :
$X^2_{\alpha/2}(\nu) < X^2_{calc} < X^2_{1-\alpha/2}(\nu)$ (bilatéral)
$X^2_{\alpha}(\nu) < X^2_{calc}$ (ou $X^2_{calc} < X^2_{1-\alpha}(\nu)$) (unilatéral)
Sinon on rejette l'hypothèse nulle avec un risque inférieur à α% de se tromper.

VI - Comparaison de 2 variances

Deux populations de variances σ_1^2 et σ_2^2 inconnues.
Deux échantillons aléatoires, tirés de façon indépendante dans chacune des populations d'où sont prélevées respectivement n_1 et n_2 unités indépendantes.
La distribution de la variable, dans chacune des populations, suit une loi Normale sinon l'effectif de chaque échantillon doit au moins être égal à 30.

$H_0 : \sigma_1^2 / \sigma_2^2 = 1$
$H_1 : \sigma_1^2 / \sigma_2^2 \neq 1$ (bilatéral) ou $\sigma_1^2 / \sigma_2^2 > 1$(unilatéral).

Variable aléatoire : $\hat{\sigma}_1^2 / \hat{\sigma}_2^2$ (on suppose que la numérotation des échantillons conduit à reporter au numérateur la plus forte des variances estimées).
Loi de la v.a : Loi F (ν_1 ; ν_2) avec $\nu_1 = n_1 - 1$ = ddl de la variance du numérateur et $\nu_2 = n_2 - 1$ = ddl de la variance du dénominateur.

Critère statistique calculé : $F_{calc} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$ avec $\hat{\sigma}_2^2 < \hat{\sigma}_1^2$

Critère statistique théorique : $F_{1-\alpha/2}(\nu_1 ; \nu_2)$ (bilatéral) ou $F_{1-\alpha}(\nu_1 ; \nu_2)$ (unilatéral) pour α donné a priori.

Conclusion : Si F calc < F théo on conserve H_0.

VII - Comparaison de 2 séries

VII – 1 - Echantillons indépendants

(Voir synthèse dans le tableau qui suit)

Soient 2 échantillons de taille respectives n_1 et n_2 et de moyennes m_1 et m_2. La question est de savoir si les échantillons 1 et 2 sont bien issus de 2 populations respectivement centrées sur μ_1 et μ_2 avec $\mu_1-\mu_2 = a$.

$H_0 : \mu_1 - \mu_2 = a$

$H_1 : \mu_1 - \mu_2 \neq a$ (bilatéral) ou $\mu_1 - \mu_2 < a$ (ou $\mu_1 - \mu_2 > a$) (unilatéral).

Variable aléatoire: Δm

. ***Cas 1 : variances des populations connues et populations normales si n<30 ou n > 30***

Loi de la v.a : $L(\Delta m) = N(\mu_1-\mu_2 ; \sigma_d(m)) = N(0 ; \sigma_d(m))$ si $a = 0$

. ***Cas 2 : variances des populations inconnues et n >30.***

Loi de la v.a : $L(\Delta m) = N(\mu_1-\mu_2 ; \sigma_d(m)) = N(0 ; \sigma_d(m))$ si $a = 0$

. ***Cas 3 : variances des populations inconnues et populations normales et n < 30***

Loi de la v.a : $L(\Delta m) = S(\nu)(\mu_1-\mu_2 ; \sigma_d(m))$ avec $\nu = n_1 + n_2 - 2 = S(\nu)(0 ; \sigma_d(m))$ si $a = 0$

Dans tous les cas : Critère statistique calculé : $t_{calc} = \dfrac{(m_1 - m_2) - (\mu_1 - \mu_2)}{\sigma_d(m)}$	avec $\sigma_d(m) = \sqrt{\sigma_1^2(m) + \sigma_2^2(m)}$

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t$ théo) on a un risque inférieur l à α% de se tromper.

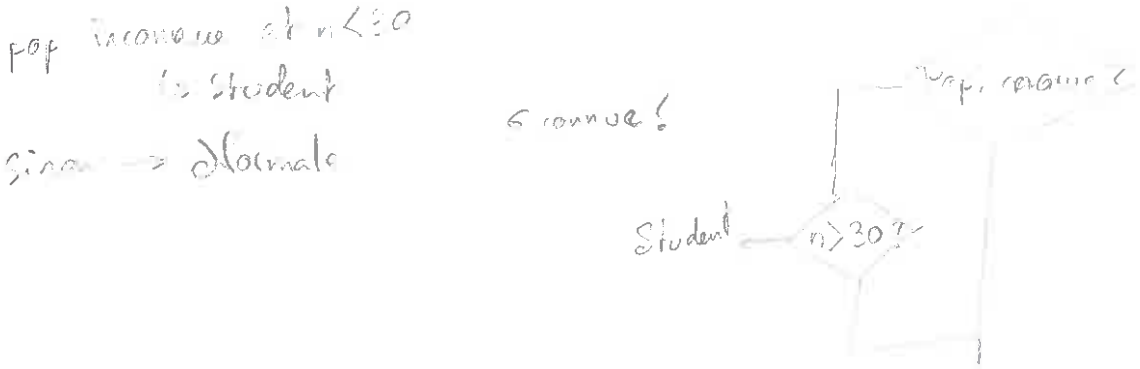


Tableau de synthèse de comparaison de 2 moyennes d'échantillons indépendants pour a = 0

Populations:		μ_1	$\sigma^2_1(x)$	et		μ_2	$\sigma^2_2(x)$	
Echantillons:	n_1	m_1	$s^2_1(x)$	et	n_2	m_2	$s^2_2(x)$	

Hypothèses:	H0: $\mu_1 - \mu_2 = 0$ Hypothèse nulle	H1: $\mu_1 - \mu_2 \neq 0$ (test bilatéral) Hypothèse alternative	ou H'1: $\mu_1 - \mu_2 < 0$ (test unilatéral) ou $\mu_1 - \mu_2 > 0$

Variances des populations connues **et** populations normales si n<30 ou Variances des populations connues **et** n > 30	Variances des populations inconnues **et** n >30	Variances des populations inconnues **et** populations normales **et** n < 30 **!!!Vérifier en premier l'égalité des variances des populations à l'aide du test F !!!** → ce test F doit être bilatéral ds ts les cas
$t\ calculé = \dfrac{(m_1 - m_2)}{\sqrt{\dfrac{\sigma_1^2(x)}{n_1}K_1 + \dfrac{\sigma_2^2(x)}{n_2}K_2}}$	$t\ calculé = \dfrac{(m_1 - m_2)}{\sqrt{\dfrac{\hat{\sigma}_1^2(x)}{n_1}K_1 + \dfrac{\hat{\sigma}_2^2(x)}{n_2}K_2}} = \dfrac{(m_1 - m_2)}{\sqrt{\dfrac{s_1^2(x)}{n_1 - 1}K_1 + \dfrac{s_2^2(x)}{n_2 - 1}K_2}}$	$t\ calculé = \dfrac{(m_1 - m_2)}{\sqrt{\dfrac{\hat{\sigma}^2(x)}{n_1}K_1 + \dfrac{\hat{\sigma}^2(x)}{n_2}K_2}}$ avec $\hat{\sigma}^2(x) = \dfrac{n_1 s_1^2(x) + n_2 s_2^2(x)}{n_1 + n_2 - 2}$
t théorique lu dans la table 2 de la loi Normale pour α donné a priori. t théorique = $t_{(1-\alpha/2)}$ (bilatéral) ou $t_{(1-\alpha)}$ (unilatéral) Autre méthode : on peut calculer le risque réel de se tromper (α calculé), ce risque s'appelle la p-value. p value (bilatéral) = 2*P (T > |t calculé|) et p value (unilatéral) = P (T > |t calculé|)		t théorique lu dans la table de la loi de Student pour α donné a priori t théorique = $t_{(1-\alpha/2)(\nu)}$ (bilatéral) ou $t_{(1-\alpha)(\nu)}$ (unilatéral) et (ν) = ddl = $(n_1 + n_2 - 2)$ **remarque: si ddl > 30 on utilise la loi normale**

Dans tous les cas:
- si \| t calc \| < t théorique on conserve H_0.
- si \| t calc \| > t théorique on rejette H_0 avec moins de α% de risque de le faire à tort.

VII – 2 - Echantillons appariés

Des échantillons appariés sont des échantillons dont les éléments sont liés 2 à 2 par une relation limitant les risques d'incidence de facteurs susceptibles d'affecter la variable aléatoire à laquelle on s'intéresse, autre que celui dont on veut garder le contrôle.

unité	1	2	...	i	...	n
X	x_1	x_2	...	x_i	...	x_n
X'	x'_1	x'_2	...	x'_i	...	x'_n
d	$d_1 = x_1 - x'_1$	$d_2 = x_2 - x'_2$	...	$d_i = x_i - x'_i$	...	$d_n = x_n - x'_n$

On travaille donc sur la série des n différences $d_i = x_i - x'_i$

Dans l'échantillon on a n différences di de moyenne $\bar{d}$ et d'écart type estimé $\hat{\sigma}_d$. La question est de savoir si la moyenne des différences sur cette série correspond à une différence théorique égale à la valeur a.

$H_0 : \delta = a$

$H_1 : \delta \neq a$ (bilatéral) ou $\delta < a$ (ou $\delta > a$) (unilatéral).

Variable aléatoire: $\bar{d}_n$

Loi de la v.a : $L(\bar{d}_n) = S(\nu)\ (\delta; \sigma(\bar{d}))$ avec ν =ddl = n-1

$= S(\nu)\ (0; \sigma(\bar{d}))$ si a = 0.

Critère statistique calculé : $t_{calc} = \dfrac{\bar{d} - \delta}{\sigma(\bar{d})}$ avec $\sigma(\bar{d}) = (\hat{\sigma}_d / \sqrt{n}) * \sqrt{K}$
Critère statistique théorique : $t_{1-\alpha/2}(\nu)$ (bilatéral) ou $t_{1-\alpha}(\nu)$ (unilatéral) dans la table de la loi Student pour α donné a priori. Remarque: si ddl > 30 on utilise la loi normale.

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t\ théo$) on a un risque inférieur à α% de se tromper.

VIII - Comparaison de 2 proportions

Soit 2 populations comportant respectivement des proportions p_1 et p_2 inconnues d'unités possédant un caractère étudié ; 2 échantillons d'effectif n_1 et n_2 supérieurs à 100, aléatoires et indépendants, extraits de chacune des populations. Les prélèvements correspondent à des tirages non exhaustifs. Les nombres d'individus possédant le caractère spécifié dans les échantillons sont notés k_1 et k_2.

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$ (bilatéral) ou $p_1 <$ (ou >) p_2 (unilatéral).

Variable aléatoire: Δf

Loi de la v.a : $L(\Delta f) = N(0 ; \sigma_d(f))$

Critère statistique calculé :

$t_{calc} = \frac{(f_1 - f_2)}{\sigma_d(f)}$

avec $\sigma_d(f) = \sqrt{\sigma_1^2(f) + \sigma_2^2(f)} = \sqrt{\frac{p_0 q_0}{n_1} K_1 + \frac{p_0 q_0}{n_2} K_2}$

et $p_0 = \frac{k_1 + k_2}{n_1 + n_2}$ $\quad q_0 = 1 - p_0$

Critère statistique théorique : $t_{1-\alpha/2}$ (bilatéral) ou $t_{1-\alpha}$ (unilatéral) lu dans la table 2 pour α donné a priori.

Conclusion : Si on rejette l'hypothèse nulle ($|t\ calc| > t\ théo$) on a un risque inférieur à α% de se tromper.

Autre méthode : On peut calculer le risque réel de se tromper (α calculé), ce risque s'appelle la p-value.

p value (bilatéral) = 2*P (T > |t calculé|) et p value (unilatéral) = P (T > |t calculé|)

IX - Comparaison de plusieurs variances : test de Bartlett

Ce test est utilisé lorsque **l'on doit vérifier** l'homogénéité des variances intra groupes de plusieurs séries statistiques.

k populations de variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_j^2, \ldots\ldots \sigma_k^2$ inconnues.
k échantillons aléatoires, tirés de façon indépendante dans chacune des population d'où sont prélevées respectivement $n_1, n_2, \ldots\ldots n_k$ unités indépendantes.
La distribution de la variable, dans chacune des populations, suit une loi Normale et aucune des variances empiriques n'est nulle ni trop petite.

H_0 : les variances sont homogènes.
H_1 : au moins une variance est supérieure aux autres.

Critère statistique calculé :

$$\chi^2_{calculé} = \left(\frac{2.3026}{C}\right) * \left[\nu * \log_{10} \hat{\sigma}^2 - \sum_{j=1}^{j=k} (\nu_j * \log_{10} \hat{\sigma}_j^2)\right]$$

Remarque 1 : Ne pas apprendre cette formule mais savoir l'appliquer. Elle sera donnée le jour de l'examen !

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{j=1}^{j=k} \frac{1}{\nu_j} - \frac{1}{\nu}\right]$$

$\nu_j = n_j - 1$ $\quad \nu = \sum_{j=1}^{j=k} \nu_j$

$\hat{\sigma}_j^2 = \frac{SCE_{(\bar{x}_j)}}{n_j - 1}$ $\quad \hat{\sigma}^2 = \frac{1}{\nu} \sum_{j=1}^{j=k} \nu_j \hat{\sigma}_j^2$

Critère statistique théorique : $X^2_{1-\alpha}$ (k-1)

Conclusion : X^2 calc < X^2 théo on ne peut pas mettre en évidence qu'au moins une variance est supérieure aux autres, on garde l'hypothèse de l'homogénéité des variances.

!!!!!!Remarque 2 : Ce test est très important à savoir faire, il est très utilisé dans la suite du programme…et même en 3ème année !!!!!!!!