Statistique inductive

Introduction

Introduction

On étudie le caractère d'une population. Etant donné une variable aléatoire X, on examine un échantillon $X_1 \dots X_n$. Lors d'une épreuve, cette échantillon donne les valeurs réelles $x_1 \dots x_n$, encore appelé *série statistique*.

	Observée (variable aléatoire)	*De la série statistique (réel)*
Moyenne	$\overline{X} = \frac{1}{n}\sum_i X_i$	$\bar{x} = \frac{1}{n}\sum_i x_i$
Variance	$S_X{}^2 = \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2 = \overline{X^2} - \overline{X}^2$	$V_x = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^2 = \overline{x^2} - \bar{x}^2$

Pour les séries statistiques doubles[1], on donne le formulaire suivant.

	Observée (variable aléatoire)	*De la série statistique (réel)*
Covariance	$S_{XY}{}^2 = \frac{1}{n}\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$ $= \overline{XY} - \overline{X}\,\overline{Y}$	$C_{xy} = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$ $= \overline{xy} - \bar{x}.\bar{y}$
Corrélation	$R_{XY} = \frac{S_{XY}{}^2}{S_X S_Y}$	$r_{xy} = \frac{C_{xy}}{\sqrt{V_x V_y}}$

On remarquera que $S_{XX}{}^2 = S_X{}^2$.

Introduction à la statistique des données

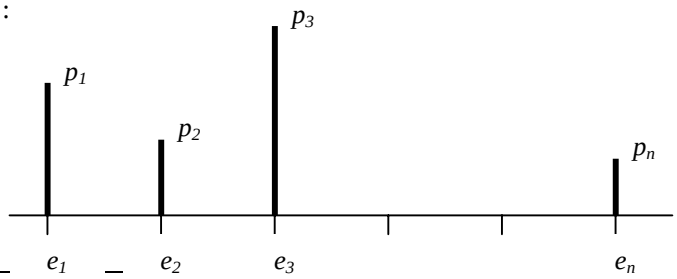
Considérons l'échantillon $X : \begin{cases} X_1 \cdots X_m \\ x_1 \cdots x_m \end{cases}$. Soit $e_1 \cdots e_n$ les valeurs prises par les x_i.

- *Cas de n petit (loi discrète)*

On suppose que le nombre de valeurs prises par la série statistique est petit, c'est-à-dire n est petit. On note *la fréquence de la valeur* e_i, $p_i = \frac{nombre\,de\,x\,égal\,à\,e_i}{m}$, et $p_i' = P\left(X = e_i\right)$.

Théorème : «A moins de ne pas avoir de chance, les p_i sont proches des p_i'. »

Diagramme bâton :



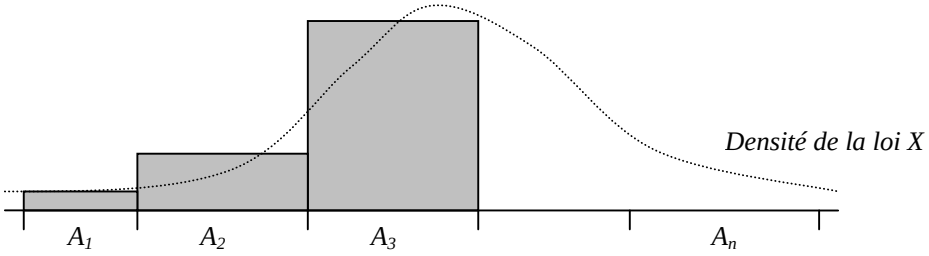
[1] Comprenant toutes deux n échantillons

- *Cas de n grand (loi diffuse)*

Dans le cas d'une loi diffuse, on partage l'intervalle des valeurs prises en *classe* A_i. On exige que chaque classe comporte au moins 3 éléments.

On redéfinit *la fréquence de la classe* A_i comme $p_i = \frac{nombre\ de\ x \in A_i}{m}$, et $p_i' = P(X \in A_i)$.

Histogramme :



Les p_i représentent les aires et non pas les hauteurs !

De la même façon, on a que les p_i sont proches des p_i' ; ce qui se traduit sur le graphique, pour une classe donnée, par des aires grises proches de l'intégrale de la densité.

Grâce au dessin par histogramme, on peut approcher la densité de la loi *X*.

- *Vocabulaire*

- On appelle *mode*, la valeur ou la classe ayant la fréquence la plus importante.
- La *médiane* est telle qu'elle partage également en deux le nombre des x_i.
- On nomme *étendue* la différence entre la plus grande valeur prise et la plus petite.

Régression linéaire

Cf. TD…

Estimation ponctuelle

On souhaite estimer une grandeur *a* inconnue. Le plus souvent, *a* est le paramètre d'une loi de probabilité. Considérons un échantillon $X : \begin{cases} X_1 \cdots X_m \\ x_1 \cdots x_m \end{cases}$. On cherche *un estimateur* $A(X_1 \cdots X_m)$, tel qu'on espère que $A(x_1 \cdots x_m)$ soit proche de *a*.

- *Définitions*

- Un estimateur est *sans biais* si $E(A) = a$.
- Un estimateur est *convergent* si $A(X_1 \cdots X_m)$ converge en probabilité vers a lorsque $m \to \infty$.
- Un estimateur sans biais et convergent est dit *correct*. $\overline{X}$ est un estimateur correct de l'espérance de *X*.
- Un estimateur est dit *de variance minimum*, si pour tout autre estimateur *B* on a : $E\left((A-a)^2\right) \leq E\left((B-a)^2\right)$. Un estimateur de variance minimum est sans biais.
- Un estimateur est *exhaustif* si la loi de $(X_1 \cdots X_m \mid A)$ ne dépend pas de *a*. ? ? ?
- Un estimateur est *robuste* si il est insensible aux données aberrantes.

- *Estimateur au maximum de vraisemblance*

On définit *la fonction de vraisemblance* $v:(x_1 \cdots x_m) \mapsto P(X_1 = x_1 \cdots X_m = x_m)$ dans le cas discret, et $v:(x_1 \cdots x_m) \mapsto f(x_1 \cdots x_m)$ dans le cas diffus avec densité *f*. La fonction de vraisemblance caractérise la vraisemblance des données. On va donc chercher $a = \varphi(x_1 \cdots x_m)$qui rend *v* maximum, c'est-à-dire tel que $\frac{\partial v}{\partial a} = 0$[2]. On prend $A = \varphi(X_1 \cdots X_m)$, *l'estimateur au maximum de vraisemblance.*

- *Inégalité de Frechet-Cramer-Rav*

Soit *X* une loi de densité $f(x_1 \cdots x_m, a)$ avec *a* le paramètre à estimer. On appelle *la borne de Cramer-Rav*, le réel $B(x,a) = \frac{1}{m.E\left(\left(\frac{\partial \log(f(x,a))}{\partial a}\right)^2\right)}$.

Soit *A* un estimateur sans biais, on démontre :

- $E(A) = a$;
- $Var(A) = E\left((A-a)^2\right)$;
- $Var(A) \geq B(x,a)$.

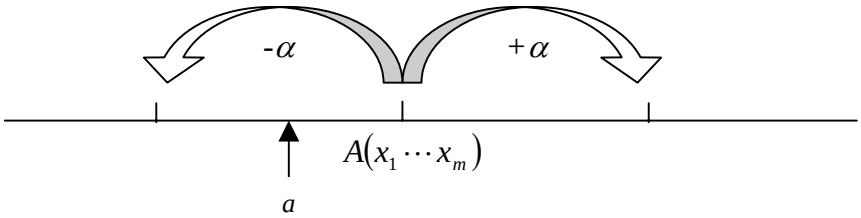
- *Efficacité d'un estimateur*

On définit alors *le coefficient d'efficacité* d'un estimateur : $0 \leq \frac{B(x,a)}{Var(A)} \leq 1$. Le coefficient d'efficacité maximum est 1. On cherche toujours un estimateur qui soit le plus précis possible, tel donc que sa variance soit la plus petite possible. On dit que l'estimateur a une *efficacité maximum* lorsque $Var(A) = B(x,a)$; la borne est atteinte. Par conséquent, on ne peut pas trouver d'estimateur qui soit plus précis !

Estimation par intervalle de confiance

Considérons un échantillon $X : \begin{cases} X_1 \cdots X_m \\ x_1 \cdots x_m \end{cases}$. On cherche *un estimateur* $A(X_1 \cdots X_m)$ d'une grandeur *a*.

Soit $\alpha > 0$, on cherche *un intervalle de confiance* autour de l'estimateur $A(x_1 \cdots x_m)$ qui soit de la forme $[A(x_1 \cdots x_m) - \alpha, A(x_1 \cdots x_m) + \alpha]$.



On se donne généralement *un seuil* ε égal à *0,05* ou *0,01*, ce qui garantit *un niveau de confiance* de 95% ou 99% dans l'intervalle.

L'estimation par intervalle de confiance est telle que $P(|A(x_1 \cdots x_m) - a| < \alpha) \geq 1 - \varepsilon$ (*niveau de confiance*) ou encore $P(|A(x_1 \cdots x_m) - a| > \alpha) < \varepsilon$ (*seuil*).

Plus le seuil est petit, plus le niveau de confiance exigé est élevé, et plus l'intervalle de confiance va être large.

[2] on vérifiera que l'on a bien à faire à un maximum.

Quelques lois de probabilités utiles en statistique

Loi du *chi-deux*

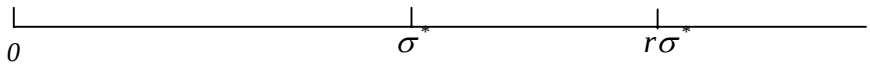
Soient $X_1 \dots X_n$ indépendants, qui suivent $N(0,1)$. $U_n = \sum_{i=1}^{n} {X_i}^2$ suit la loi du $\aleph^2$, à *n* degrés de liberté.

Estimateur de variance d'une loi normale, et intervalle de confiance

Soient *X* qui suit $N(m,\sigma)$, et un échantillon $X_1 \dots X_n$.

- *Cas où m est connue*

On a ${V_n}^* = {\sigma_n}^{*2} = \frac{1}{n}\sum_n (X_i - m)^2$ l'estimateur de la variance, sans biais car $E({V_n}^*) = Var(X)$. Alors $\frac{n{V_n}^*}{\sigma^2} = n\left(\frac{{\sigma_n}^*}{\sigma}\right)^2$ suit la loi du $\aleph^2$, à *n* degrés de liberté.



Déterminons l'intervalle de confiance pour σ ,au seuil ε. Seule la partie droite de l'intervalle de confiance est significative (pas de valeur absolue). On cherche *r* tel que $P(\sigma > r\sigma^*) \le \varepsilon$ ou $P(\sigma < r\sigma^*) \ge 1-\varepsilon$, ce qui donne $P\left(\frac{n}{r^2} < U_n\right) \ge 1-\varepsilon$. La table du $\aleph^2$ à *n* degrés de liberté nous donne $\frac{n}{r^2}$. On déduit *r*.

- *Cas où m est inconnue*

On remplace *m* par son estimateur $\overline{X}$. Par conséquent, on a ${V_{n-1}}^* = {\sigma_{n-1}}^{*2} = \frac{1}{n-1}\sum_n \left(X_i - \overline{X}\right)^2$ l'estimateur de la variance, sans biais. Notons bien que si l'on prenait $\frac{1}{n}$ plutôt que $\frac{1}{n-1}$ dans l'expression de cet estimateur, on introduirait un biais. Ce point est sans importance dès que *n* devient grand. Alors $(n-1)\left(\frac{{\sigma_{n-1}}^*}{\sigma}\right)^2$ suit la loi du $\aleph^2$, à *n - 1* degrés de liberté.

Déterminons l'intervalle de confiance pour σ ,au seuil ε. On cherche *r* tel que $P(\sigma > r{\sigma_{n-1}}^*) \le \varepsilon$, ce qui donne $P\left(\frac{n-1}{r^2} < U_{n-1}\right) \ge 1-\varepsilon$. La table du $\aleph^2$ à *n - 1* degrés de liberté nous donne $\frac{n-1}{r^2}$. On déduit *r*.

Loi de Student

- *Définition*

Soient $X_0 \dots X_n$ indépendants, un échantillon de $N(0,1)$. Alors $St = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}}} = \frac{X_0}{\sqrt{\overline{X^2}}}$ suit la loi de *Stundent* à n degrés de liberté.

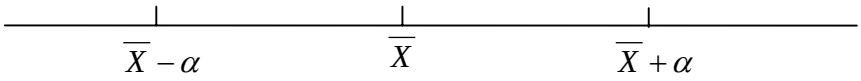
Les tables de *Student* donne $P(|St| > x)$. Ces tables ne sont établies que pour $n \leq 30$, après on considère que *St* suit à peu près $N(0,1)$. En effet, la loi des grands nombres montre que *St* converge en probabilité vers la loi normale réduite.

- *Application de cette loi*

Soient $X_1 \dots X_n$ indépendants, un échantillon de $N(m,\sigma)$. Alors $\left(\frac{\overline{X} - m}{\sigma^*_{n-1}}\right)\sqrt{n}$ suit *Student* à *n - 1* degrés de liberté.

Intervalle de confiance pour l'espérance m d'une loi normale

Déterminons l'intervalle de confiance pour *m* ,au seuil *ε*.



- *Cas où σ est connu*

On cherche α tel que $P(|\overline{X} - m| > \alpha) \leq \varepsilon$ ou $P(|\overline{X} - m| < \alpha) \geq 1 - \varepsilon$. $\overline{X}$ suit $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. On se ramène à la loi normale réduite : $P\left(\frac{|\overline{X} - m|}{\sigma}\sqrt{n} < \frac{\alpha}{\sigma}\sqrt{n}\right) \geq 1 - \varepsilon$ et on déduit $\frac{\alpha}{\sigma}\sqrt{n}$.

- *Cas où σ est inconnu*

Comme σ est inconnue, on ne peut pas se ramener à la loi normale réduite. On va donc utiliser l'estimateur de l'écart-type σ^*_{n-1} et la loi de Student : c'est-à-dire, on cherche α tel que $P\left(\frac{|\overline{X} - m|}{\sigma^*_{n-1}}\sqrt{n} > \frac{\alpha}{\sigma^*_{n-1}}\sqrt{n}\right) \leq \varepsilon$, avec $\left(\frac{\overline{X} - m}{\sigma^*_{n-1}}\right)\sqrt{n}$ qui suit *Student* à *n - 1* degrés de liberté.

- *Propriété importante, cas où σ est inconnu, grand échantillon*

Dans le cas d'un grand échantillon $(n > 30)$, on a $\left(\frac{\overline{X} - m}{\sigma^*_{n-1}}\right)\sqrt{n}$ qui suit la loi normale réduite. C'est encore vrai si $X_1 \dots X_n$ est un échantillon quelconque (pas forcément une loi normale) ! Ce théorème est une conséquence de la loi des grands nombres, et du théorème central limite.

Loi de Snedecor

Considérons $\begin{cases} X_1 \dots X_m \\ Y_1 \dots Y_n \end{cases}$ deux échantillons indépendants, qui suivent la loi normale réduite. Par définition, on a $S_{m,n} = \dfrac{\frac{1}{m}\sum_m X_i^2}{\frac{1}{n}\sum_n Y_i^2} = \dfrac{\overline{X^2}}{\overline{Y^2}}$ qui suit la loi de *Snedecor* à $\begin{pmatrix} m \\ n \end{pmatrix}$ degrés de liberté.

Tests d'hypothèses

Généralités

On cherche à tester une hypothèse.

1. **Le bon choix de l'hypothèse :** Considérons le problème classique du critère de qualité : par exemple, le vendeur vient de recevoir un lot de 1000 pièces et il souhaite qu'il y ait moins de 50 pièces défectueuses dans le lot. *L'hypothèse H à tester est celle dont on a confiance dans le rejet,* c'est-à-dire *"Il y a plus de 50 pièces défectueuses dans le lot"*. Il ne faut pas confondre avec l'hypothèse contraire $\overline{H}$ qui représente ce en quoi on a confiance dans l'acceptation. Soit θ un paramètre de la loi de *X* sur lequel doit portee le test. Par exemple, X suit $N(m,\sigma)$ et $\theta = m$. On distingue deux sortes d'hypothèses : *les hypothèses composées* *("$m \geq m_0$ ")* et *les hypothèses simples ("$m = m_0$ ").*

2. **La condition de rejet de l'hypothèse** : On définit alors une *condition ou région de rejet* de *l'hypothèse H*. Par exemple $\{\overline{X} \leq a\}$ pour "$m \geq m_0$ ", et $\{|\overline{X} - m| \geq a\}$ pour "$m = m_0$ ".

3. **Condition de seuil :** *On cherche à déterminer a tel que* $P(rejet\ H\ /\ H\ vraie) \leq \varepsilon$ *(*risque de 1ère espèce) avec ε un seuil donné. On exprime que la probabilité que l'on se trouve dans la région de rejet de *H*, sachant que *H* est vraie, est inférieure au seuil ε. En effet, *le seuil représente le niveau de confiance que l'on souhaite avoir dans le rejet,* qui est en général de *0.05, 0.01 ou 0.1*. Plus ε est petit, et plus le niveau de confiance exigé pour l'acceptation de $\overline{H}$ est élevé.

4. **Principe de la puissance maximum** : On obtient en résultat une inégalité sur *a*, le plus souvent $a \leq a_0\ ou\ a \geq a_0$. On applique *le principe de la puissance maximum*, qui cherche à grandir au maximum la région de rejet. Ce qui impose $a = a_0$.

5. **Interprétation du résultat** : Il faut envisager deux cas. Ainsi, si on a effectivement $\overline{X} \leq a_0$ alors on peut rejeter l'hypothèse *H* avec un niveau de confiance de 90%, 95% ou 99% , ce qui revient à accepter $\overline{H}$. En revanche, si $\overline{X} \leq a_0$ n'est pas vérifié, alors il n'y a pas de rejet.

- *Test de Bayes*

On affecte au risque de 1ère espèce $P_1 = P(rejet\ H\ /\ H\ vraie)$ un coût C_1 et au risque de 2nde espèce $P_2 = P(accepter\ H\ /\ H\ \ faux)$ un coût C_2. On définit le coût moyen d'erreur par $C = C_1 P_1 + C_2 P_2$. On cherche le minimum de *C*.

Comparaison de deux moyennes

Considérons $\begin{cases} X : X_1 \dots X_m \\ Y : Y_1 \dots Y_n \end{cases}$ deux échantillons indépendants.

- *Grands échantillons (taille ≥ 30)*

On ne suppose rien sur *X* et *Y*. On pose $m_X = E(X)$ et $m_Y = E(Y)$. $T = \dfrac{(\overline{Y} - \overline{X}) - (m_Y - m_X)}{\sqrt{\dfrac{1}{m}S_X{}^2 + \dfrac{1}{n}S_Y{}^2}}$ suit à peu près la loi normale réduite.

On veut tester l'hypothèse *H :"* $m_X = m_Y$ *"*. On prend la condition de rejet $\left\{\left|\overline{Y} - \overline{X}\right| \geq a\right\}$. Seulement d'un point de vue strictement calculatoire, il serait plus habile de prendre $\left\{\dfrac{\left|\overline{Y} - \overline{X}\right|}{\sqrt{\dfrac{1}{m}S_X{}^2 + \dfrac{1}{n}S_Y{}^2}} \geq a\right\}$. On cherche à déterminer *a* pour un seuil ε donné. On écrit $P(\text{rejet } H \text{ / } H \text{ vraie}) \leq \varepsilon$, c'est-à-dire $P(|T| \geq a) \leq \varepsilon$ avec *T* qui suit $N(0,1)$. On en déduit $a \geq a_0$. Le principe de la puissance maximale impose $a = a_0$.

- *Petits échantillons*

On suppose en plus que X et Y suivent des lois normales de même écart-type. On définit $U = \sqrt{\dfrac{mn(m+n-2)}{m+n}} \dfrac{\left[(\overline{Y} - \overline{X}) - (m_Y - m_X)\right]}{\sqrt{mS_X{}^2 + nS_Y{}^2}}$. *U* suit la loi de *Student* à $m+n-2$ degrés de liberté.

On veut tester l'hypothèse *H :"* $m_X = m_Y$ *"*. Comme précédemment, on cherche à se ramener au théorème ; donc, on prend la condition de rejet $\left\{\sqrt{\dfrac{mn(m+n-2)}{m+n}} \dfrac{\left|\overline{Y} - \overline{X}\right|}{\sqrt{mS_X{}^2 + nS_Y{}^2}} \geq a\right\}$. On cherche à déterminer *a* pour un seuil ε donné. On écrit $P(\text{rejet } H \text{ / } H \text{ vraie}) \leq \varepsilon$, c'est-à-dire $P(|U| \geq a) \leq \varepsilon$ avec *U* qui suitla loi de *Student* à $m+n-2$ degrés de liberté. En application du principe de la puissance maximale, on déduit $a = a_0$.

<u>Test des longueurs (comparaison de deux lois)</u>

Considérons $\begin{cases} X : X_1 \dots X_m \\ Y : Y_1 \dots Y_n \end{cases}$ deux échantillons indépendants. On veut tester l'hypothèse *H :" X et Y suivent la même loi"*. On classe par ordre croissant (ou décroissant) les *x* et les *y* en les regroupant : $xx, y, xxx, yyyy, x$. On appelle *L* le nombre des longueurs. Ici, $L = 5$. Si *L* est petit, cela signifie que les *X* et les *Y* ne se mélange pas bien ; par conséquent, il ne s'agira pas de la même loi. On traduit l'hypothèse *H* dans cette même idée, en disant que tous les ordres possibles sont équiprobables. La condition de rejet est $\{L \leq l\}$. On cherche à déterminer *l* tel que $P(L \leq l \text{ / } H \text{ vraie}) \leq \varepsilon$.

- *Petits échantillons*

Supposons $m \leq n$. On s'intéresse au *X*, les résultats sont symétriques pour *Y*. Soit *s* le nombre de longueurs relatif au *X*. On a :

- $P(L = 2s) = \dfrac{2C_{m-1}^{s-1}C_{n-1}^{s-1}}{C_{m+n}^{m}}$ pour $1 \leq s \leq m$
- $P(L = 2s+1) = \dfrac{C_{m-1}^{s}C_{n-1}^{s-1} + C_{m-1}^{s-1}C_{n-1}^{s}}{C_{m+n}^{m}}$ pour $1 \leq s < m$

- $P(L = 2m+1) = \dfrac{C_{m-1}^{n}}{C_{m+n}^{m}}$

- *Grands échantillons ($m, n \geq 7$ ou 8)*

L suit $N(m_L, \sigma_L)$ avec $m_L = 1 + \dfrac{2mn}{m+n}$ et $\sigma_L = \dfrac{2mn(m^2+n^2-m-n)}{(m+n)^2}$.

Test du rang (comparaison de deux lois)

Considérons $\begin{cases} X : X_1 \dots X_m \\ Y : Y_1 \dots Y_n \end{cases}$ deux échantillons indépendants. On veut tester l'hypothèse *H :" X et Y suivent la même loi"*. On classe par ordre croissant (ou décroissant) les *x* et les *y* : *x x y x x x y y y y x* . Soit *T* la somme des rangs de *X*. Ici, $T = 1+2+4+5+6+11 = 29$.

- *Définition du test*

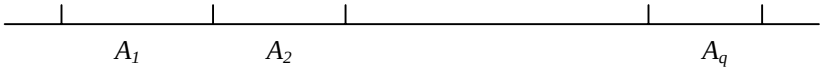
Y est stochastiquement supérieure à *X*, noté $Y \geq X$ si et seulement si $\forall z, P(Y \leq z) \leq P(X \leq z)$. On teste l'hypothèse *H : "X et Y suivent la même loi"* contre *K : "Y est stochastiquement supérieur à X"*. La condition de rejet de *H* est $\{T \leq t\}$. On cherche à déterminer *t* tel que $P(T \leq t \ / \ H \ vraie) \leq \varepsilon$.

- *Théorème*

Si *H* est vraie, si *n* et *m* sont "grands", alors *T* suit à peu près $N(m_T, \sigma_T)$ avec $m_T = \dfrac{m}{2}(m+n+1)$ et $\sigma_T{}^2 = \dfrac{mn}{12}(m+n+1)$.

Test du $\aleph^2$

Considérons un échantillon $X : X_1 \dots X_n$. A partir de l'histogramme, on imagine la loi *P* de *X*. On teste l'hypothèse *H : "X suit la loi P"*. Ce test va permettre de vérifier si un échantillon est conforme à une loi. On effectue un partage de la droite réelle en *q* classes. Chaque classe doit comporter un nombre suffisant de x_i.



On rappelle :

- la fréquence de la classe des A_j : $p_j = \dfrac{Card(A_j)}{n}$
- la fréquence théorique de la classe des A_j : $p'_j = P(X \in A_j)$ calculé avec la loi *P*

- *Règle heuristique pour le partage en classe*
  1. Le nombre de classes doit être supérieure à 4 fois le nombre de paramètre de la loi. Par exemple pour la loi normale, il faudra au moins 8 classes.
  2. Le nombre d'éléments attendus pour la classe A_j doit être telle que $np_j \geq 5$, sauf éventuellement pour deux classes où il doit être > 1.

- *Théorème*

Si *H* est vraie, $Y_n = n\sum_{j=1}^{q} \dfrac{1}{p'_j}(p_j - p'_j)^2$ converge en loi vers la loi du $\aleph^2$ à *q – 1* degrés de liberté.

- *Condition de rejet*

La condition de rejet de *H* est $\{Y_n \leq \alpha \text{ ou } Y_n \geq \beta\}$. Si $Y_n \leq \alpha$, la fréquence des p_j est trop différente des fréquences théoriques p'_j pour que *X* suivent la loi *P*. Si $Y_n \geq \beta$, l'échantillon est trop bon, et il a certainement été truqué !

Analyse de la variance

Test de comparaison de *S* moyennes (S ≥ 3)

$X_1, X_2 \dots X_S$ suivent des lois normales avec même écart-type : $N(m_i, \sigma)$ pour $1 \leq i \leq s$. On souhaite tester l'hypothèse *H : "* $m_1 = m_2 = \cdots = m_S$ *"*. Pour chaque X_i, on dispose d'un échantillon $X_{i,j}$ avec $j = 1 \cdots n_i$. Soit *n* la somme des n_i. On suppose tous les $X_{i,j}$ indépendants.
On définit :

- $\overline{X_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}$, l'estimateur de X_i ;
- $\overline{X} = \frac{1}{n} \sum_{i,j} X_{i,j} = \frac{1}{n} \sum_i n_i \overline{X_i}$, la moyenne estimée globale ;
- $V_i^* = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{i,j} - \overline{X_i}\right)^2$, l'estimateur de variance de X_i ;
- $\overline{V^*} = \frac{1}{n-S} \sum_i (n_i - 1) V_i^* = \frac{1}{n-S} \sum_{i,j} \left(X_{i,j} - \overline{X_i}\right)^2$
- $\sigma^{2^*} = \frac{1}{S-1} \sum_i n_i \left(\overline{X_i} - \overline{X}\right)^2$, l'estimateur de σ^2 (si *H* vraie).

- *Théorème*

Si *H* est vraie, $W = \frac{\sigma^{2^*}}{\overline{V^*}}$ suit *Snedecor* à $\begin{pmatrix} S-1 \\ n-S \end{pmatrix}$ degrés de liberté. La condition de rejet de *H* est $\{W \geq w\}$.
Donc on cherche *w* tel que $P(W \geq w \,/\, H \text{ vraie}) \leq \varepsilon$.

Test de l'interaction et de l'influence de deux facteurs *A* et *B* sur un caractère *X*

- *Préliminaires*

Soit *X* un caractère sur lequel peuvent agir deux facteurs *A* et *B*. *A* peut prendre des états $A_1 \dots A_p$, *B* peut prendre des états $B_1 \dots B_q$. Lorsque $A = A_i$ et $B = B_j$, le caractère obtenu est $X_{i,j}$. α_i représente l'influence de la cause $A = A_i$, β_j représente l'influence de la cause $B = B_j$, $\gamma_{i,j}$ représente l'influence de l'interaction entre les deux causes. On suppose que $X_{i,j}$ suit $N(m_{i,j}, \sigma^2)$, avec σ^2 fixé.

Considérons l'équation $\mu + \alpha_i + \beta_j + \gamma_{i,j} = m_{i,j}$ pour $i = 1 \cdots p$ et $j = 1 \cdots q$. La somme des α_i est nulle, de même que la somme des β_j. La somme des $\gamma_{i,j}$ est nulle par rapport à *i* et par rapport à *j*. On a $\mu = \overline{m}$, $\alpha_i = \overline{m_{i,\bullet}} - \overline{m}$, $\beta_j = \overline{m_{\bullet,j}} - \overline{m}$.

On définit les hypothèses suivantes :

- *H_I : "Il n'y a pas d'interaction."*, c'est-à-dire $\forall i, j, \gamma_{i,j} = 0$.

- H_A *: "Le facteur A n'a pas d'influence."*, c'est-à-dire $\forall i, \alpha_i = 0$.
- H_B *: "Le facteur B n'a pas d'influence."*, c'est-à-dire $\forall j, \beta_j = 0$.

- *Cas où il n'y a pas d'interaction*

On se place dans le cas où il n'y a pas d'interaction entre les facteurs *A* et *B* : $\forall i, j, \gamma_{i,j} = 0$. *Cf. Notations…*

$\tilde{\sigma}^2 = \frac{pq}{\sigma^2} \overline{S_{\bullet,\bullet}}^2$ est un estimateur sans biais de σ^2, et il suit la loi du $\aleph^2$ à $(p-1)(q-1)$ degrés de liberté. Par conséquent, on peut effectuer une estimation par intervalle de confiance de σ^2.

Test de H_A : Si H_A est vraie, alors $W_A = (q-1)\frac{S_A^2}{\overline{S_{\bullet,\bullet}}^2}$ suit la loi de *Snedecor* à $\begin{Bmatrix} p-1 \\ (p-1)(q-1) \end{Bmatrix}$ degrés de liberté. La condition de rejet est $\{W_A \geq w_A\}$. On cherche w_A tel que $P(W_A \geq w_A \ / \ H \ vraie) \leq 0.05$.

- *Cas plus général (avec échantillons)*

(…)