

Feuille de TP Python 05

Statistiques : Lois du χ^2 , de STUDENT, Tests.

L'objet de ce TP est essentiellement de mettre en place une méthodologie pour effectuer deux tests statistiques usuels concernant la conformité d'une moyenne (en cas de faible effectif) et l'adéquation de résultats à une loi théorique donnée à l'avance.

1 Introduction

On suppose que X est une variable aléatoire réelle et la suite $(X_i)_{i \geq 1}$ est un échantillon¹ de X

Les statistiques les plus importantes que l'on calcule sur un échantillon $(X_i)_{i \geq 1}$ sont :

1. La moyenne empirique d'ordre N :

$$\bar{X}_N = \frac{1}{N} \sum_{k=1}^N X_k;$$

2. L'erreur quadratique d'ordre N et l'estimateur sans biais de la variance :

$$S_N^2(m) = \frac{1}{N} \sum_{k=1}^N (X_k - m)^2 \text{ et } \Sigma_N^2 = \frac{1}{N-1} \sum_{k=1}^N (X_k - \bar{X}_N)^2.$$

On rappelle que le test de conformité de l'espérance de X à une moyenne m donnée à l'avance est basée sur le calcul de la statistique

$$T_N = \frac{\bar{X}_N - m}{\Sigma_N}$$

et le fait que pour N grand, T_N est distribuée² « approximativement » suivant une loi normale $\mathcal{N}(0, 1)$.

La question que l'on aborde aujourd'hui dans un premier temps porte sur ce qui se passe lorsque N « n'est pas grand³ » mais que l'on teste des hypothèses plus fortes que la simple valeur de l'espérance. Ici, on va faire une hypothèse de normalité, *i.e.* le fait que $X \sim \mathcal{N}(m, \sigma^2)$. Deux lois importantes pour ces questions seront définies.

Dans un second temps, on introduira et travaillera sur un test d'adéquation de données à une loi donnée à l'avance. Les outils (lois particulières) développés dans le premier temps seront réutilisés.

2 Deux lois à densité importantes en statistiques

2.1 La loi du χ^2 à d degrés de liberté.

Définition 1. Soit $d \in \mathbb{N}^*$. La « loi du χ^2 à d degrés de liberté », notée $\chi^2(d)$ est la loi de la norme euclidienne au carré d'un vecteur aléatoire, à valeurs dans $\mathbb{R}^d$ dont les composantes forment un d -échantillon de loi $\mathcal{N}(0, 1)$.

Autrement dit, si $(Y_1, \dots, Y_d)$ est une famille de d v.a. indépendantes de loi $\mathcal{N}(0, 1)$, alors $U = \sum_{k=1}^d Y_k^2$ suit une loi $\chi^2(d)$.

En reprenant les notations de l'introduction, si de plus $X \sim \mathcal{N}(m, \sigma^2)$,

1. (facile, juste définition) $\left(\frac{\sqrt{N}}{\sigma}\right)^2 \cdot S_N^2$ suit une loi $\chi^2(N)$; S_N et $\bar{X}_N$ ne sont pas indépendantes; $\frac{\sqrt{N}}{\sigma}(\bar{X}_N - m) \sim \mathcal{N}(0, 1)$
2. (difficile, nécessite un changement de coordonnées et des notions sur l'indépendance de composantes de vecteurs gaussiens, hors programme) $U_N = \left(\frac{\sqrt{N-1}}{\sigma}\right)^2 \cdot \Sigma_N^2$ suit une loi $\chi^2(N-1)$; U_N et $\bar{X}_N$ sont indépendantes;

1. Rappel : c'est une suite de v.a. mutuellement indépendantes, toutes distribuées suivant la même loi que X
2. Noter que dans le cours sur le test de conformité à la moyenne, la quantité $T_N = \frac{\bar{X}_N - m}{\Sigma_N(m)}$ est utilisée, à ce point ça n'y change rien, les différences étant absorbées par le « approximativement »
3. oui, on peut faire de la statistique inférentielle sur de petits échantillons !!

Théorème 2. Une variable aléatoire U de loi $\chi^2(d)$ est une variable aléatoire à densité dont une densité est donnée⁴ par la fonction δ définie par

$$\forall u \in \mathbb{R}, \delta(u) = \begin{cases} 0 & \text{si } u < 0 \\ C_d \cdot u^{\frac{d}{2}-1} e^{-\frac{1}{2}u} & \text{si } u \geq 0 \end{cases}.$$

Son espérance et sa variance valent⁵ respectivement :

$$\mathbb{E}(U) = d, \mathbb{V}(U) = 2d.$$

2.2 La loi t de STUDENT à d degrés de liberté.

Définition 3. Soit $d \in \mathbb{N}^*$. La « loi de STUDENT⁶ à d degrés de liberté », notée parfois $t(d)$ est la loi du quotient

$$T = \frac{Z}{\sqrt{\frac{U}{d}}}$$

où $Z \sim \mathcal{N}(0, 1)$, $U \sim \chi^2(d)$ sont indépendantes.

Autrement dit, si $(Z, Y_1, \dots, Y_d)$ est une famille de d v.a. indépendantes de loi $\mathcal{N}(0, 1)$, alors $T = \frac{Z}{\sqrt{\frac{1}{d} \sum_{k=1}^d Y_k^2}}$ suit une loi de STUDENT $t(d)$.

En reprenant les notations de l'introduction, si de plus $X \sim \mathcal{N}(m, \sigma^2)$,

- (problème, loi inconnue/non référencée) S_N et $\bar{X}_N$ ne sont pas indépendantes ; la loi de $\frac{\bar{X}_N}{S_N}$ est inconnue.
- (difficile, mais loi connue) $\left(\frac{\sqrt{N-1}}{\sigma}\right)^2 \cdot \Sigma_N^2$ (de loi $\chi^2(N-1)$) et $\bar{X}_N$ (de loi $\mathcal{N}\left(m, \frac{\sigma^2}{N}\right)$) sont indépendantes ;

$$T_N = \sqrt{N} \frac{\bar{X}_N - m}{\Sigma_N}$$

suit une loi $t(N-1)$ de STUDENT à $N-1$ degrés de liberté. On peut noter, est c'est là toute l'astuce, que cette loi ne dépend ni de m , ni de σ^2 .

Théorème 4. Une variable aléatoire T de loi $t(d)$ est une variable aléatoire à densité dont une densité est donnée⁷ par la fonction δ définie par

$$\forall t \in \mathbb{R}, \delta(t) = \frac{c_d}{\left(1 + \frac{t^2}{d}\right)^{\frac{d+1}{2}}}.$$

Son espérance et sa variance valent⁸ respectivement :

$$\mathbb{E}(T) = \begin{cases} \text{non def.} & \text{si } d \leq 1 \\ 0 & \text{si } d > 1 \end{cases}, \mathbb{V}(T) = \begin{cases} \text{non def.} & \text{si } d \leq 2 \\ \frac{d}{d-2} & \text{si } d > 2 \end{cases}.$$

2.3 Travail à effectuer

- 10 Ecrire une fonction `Chi2(d)` retournant une valeur simulée de v.a. $\chi^2(d)$. On pourra avoir recours, pour simuler une v.a. $\mathcal{N}(0, 1)$, à une fonction du module `scipy.stats` ou à la technique exposée dans le TP 4.
2. Vérifier (reprendre le TP 4) l'adéquation⁹ de votre simulation aux résultats du théorème 2. On pourra tenter de reproduire la figure Fig. 1
3. Ecrire une fonction `Student(d)` retournant une valeur simulée de v.a. de STUDENT à d degrés de liberté.
- 15 4. Vérifier (reprendre le TP 4) l'adéquation¹⁰ de votre simulation aux résultats du théorème 4.

4. C_d est une constante de normalisation fixée par le fait que $\int_{\mathbb{R}} \delta(u) du = 1$.

5. Ces résultats se montrent totalement dans le cadre de notre programme (la loi du χ^2 est une « loi γ ») et font l'objet d'exercices récurrents : savez vous les redémontrer ?

6. W. GOSSET a utilisé ce pseudonyme alors qu'il était, étudiant, stagiaire chez GUINNESS

7. c_d est une constante de normalisation fixée par le fait que $\int_{\mathbb{R}} \delta(u) du = 1$.

8. Ces résultats se montrent probablement dans le cadre de notre programme. Bon, ce n'est pas si facile que ça non plus...

9. La loi et ses différentes caractéristiques sont définies dans le module `scipy.stats`, dans l'objet `chi2` (simulation `chi2.rvs(d)`, densité `chi2.pdf(x, d)`, fonction de répartition `chi2.cdf(x, d)`, fonction des quantiles, `chi2.ppf(q, d)`...)

10. Dans le module `scipy.stats`, on peut utiliser l'objet `t` qui définit la loi via ses différentes caractéristiques. On peut (pour libérer le nom de variable `t`) par exemple importer cet objet via `from scipy.stats import t as student` et ensuite utiliser `student.rvs`, `student.pdf`, `student.cdf`, `student.ppf`...

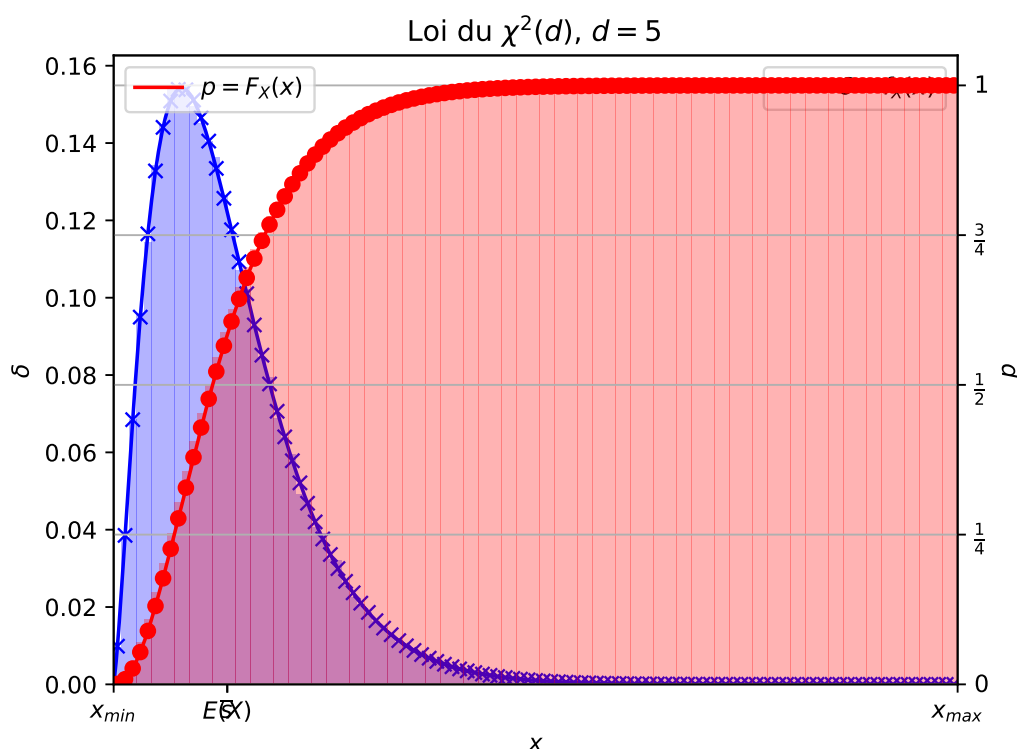


FIGURE 1 – La loi du χ^2 pour $d = 5$ degrés de liberté.

3 Test de conformité à la moyenne sous hypothèse Gaussienne

On dispose d'une série statistique $x = (x_n)_{1 \leq n \leq N}$ réelle que l'on suppose être une réalisation d'un N -échantillon $(X_1, \dots, X_N)$ d'une certaine variable aléatoire réelle X .

Le nombre réel m étant donné, on souhaite tester la conformité de la série x à l'hypothèse :

$$H_0 : \exists \sigma > 0, X \sim \mathcal{N}(m, \sigma^2)$$

D'après ce que nous avons vu précédemment, sous cette hypothèse, la quantité statistique (calculée en fonction de x et m)

$$t_m = \sqrt{N} \frac{\bar{x} - m}{\tilde{\sigma}} \quad \text{où } \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n, \quad \tilde{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2,$$

est une réalisation de la v.a. $T = T_N$ définie en introduction.

Sous l'hypothèse H_0 , la loi de T_N est connue, c'est une loi à densité δ continue et on peut évaluer la p -valeur de la quantité t_m , à savoir

$$p = \mathbb{P}(T \in J_{t_m}) \quad \text{où } J_{t_m} = \{\tau \in \mathbb{R}, \delta(\tau) \leq \delta(t_m)\}$$

- 5 La p -valeur de la quantité t_m , réalisation de la v.a. T est la probabilité d'observer pour T une valeur moins susceptible d'être observée que t , ce qui justifie la formule précédente.

Comme le graphe de δ est connu, l'ensemble J_{t_m} est une réunion de deux intervalles, symétrique par rapport à 0 et la valeur de p est calculable par une intégrale (cf. Fig 2).

3.1 Travail à effectuer

10 Dans un script séparé des scripts précédents (intervalle-confiance-student.py)

1. (a) En utilisant la fonction `t.cdf` du module `scipy.stats`, qui calcule la fonction de répartition d'une v.a. de STUDENT, écrire une fonction `pValeurStudent(t, d=5)` qui retourne la p -valeur décrite précédemment de la quantité t . Faire afficher (cas $d = 5$) le graphe de cette p -valeur en fonction de t .
- (b) En utilisant la fonction `t.ppf` du module `scipy.stats` qui calcule la fonction des quantiles d'une v.a. de STUDENT, écrire une fonction `a_Student(p, d=5)` qui retourne la valeur $\alpha \in \mathbb{R}_+$ telle que `pValeurStudent(alpha) = p`.
- 15 (c) Ecrire, à la main, sur votre feuille, pour d variant de 1 à 15, la liste des valeurs α_d telles que `pValeurStudent(alpha_d) = 0.05`. (Faites évidemment le calcul à la machine...)

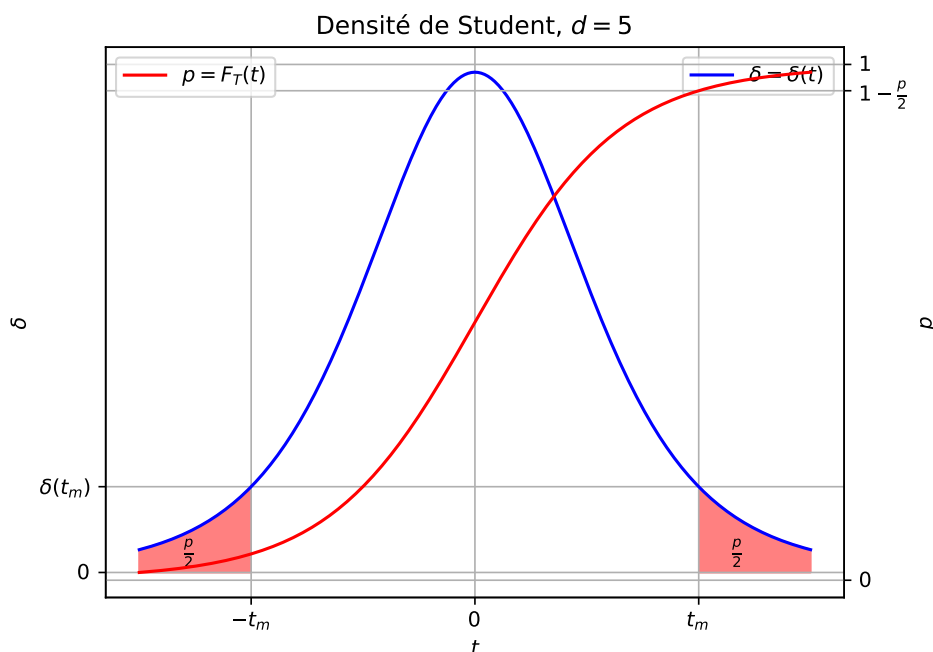


FIGURE 2 – Dans le cas de STUDENT, par symétrie et décroissance de la densité sur $\mathbb{R}_+$, J_{t_m} est constitué de l'union de $[t_m, +\infty[$ et de son opposé, c'est la « base » de la zone rosée.

2. Ecrire une fonction `EchantillonGaussien(NS = 5, m0 = 2)` qui retourne une réalisation d'un NS-échantillon $\mathcal{N}(m, \sigma^2)$ où $m = m0$ et σ^2 est inconnu (tiré au sort d'une façon quelconque).
3. Etant donné une série statistique numérique $x = (x_1, \dots, x_N)$ réalisation d'un échantillon d'une v.a. X et une valeur m , on peut calculer $t_m = \sqrt{N} \frac{\bar{x} - m}{\tilde{\sigma}_N}$ la réalisation (calculable à partir de cette série x) de la variable aléatoire T définie dans l'introduction. (On pose, comme en introduction $\bar{x}$ la moyenne de la série statistique x et $\tilde{\sigma}_N = \frac{1}{N-1} \sum_{n=1}^N (x_n - \bar{x})^2$ sa variance corrigée au sens de STUDENT.)
Ecrire une fonction `Graphe_pValeur(x, Im)` qui, étant donné un échantillon Gaussien (par exemple issu de la fonction de la question précédente) x , affiche le graphe de la fonction π définie par

$$\forall m \in \text{Im}, \pi(m) = p\text{-valeur de } t_m \text{ pour l'échantillon } x$$

NB : Pour raison graphique, l'intervalle I_m peut être pris centré en $\bar{x}$, de demi-largeur $2\tilde{\sigma}_N$

4. (Théorie). Déterminer une valeur de α (en fonction de N le nombre d'observations dans la série x) pour que en posant

$$I_x = \left[\bar{x} - \alpha \cdot \frac{\tilde{\sigma}_N}{\sqrt{N}}, \bar{x} + \alpha \cdot \frac{\tilde{\sigma}_N}{\sqrt{N}} \right],$$

la fonction π définie précédemment vérifie :

$$\forall m \in I_x, \pi(m) \leq 0.05 \text{ et } \forall m \notin I_x, \pi(m) > 0.05$$

Faites le lien avec la première question de cette série et corrigez la fonction de la question précédente pour que le graphe affiche (dans l'axe des abscisses), l'intervalle I_m ainsi que la valeur $m0$ valant par défaut 2.

5. Imaginez que vous menez NE = 1000 fois la suite¹¹ d'opérations suivante :

- (a) Tirer au sort un échantillon $x = \text{EchantillonGaussien}()$;
- (b) Tracer le graphe de π pour cet échantillon x ;

Dans combien (approximativement) de cas allez vous observer le fait que $m0$ est dans l'intervalle I_x ?

Pouvez vous écrire une fonction `CheckTheorie(m0 = 2, NS = 5, NE = 1000)` qui vérifie ce fait ?

L'intervalle I_x est appelé « Intervalle de confiance de STUDENT au niveau 5% » pour la moyenne de X . Il est composé des valeurs de m pour lesquelles la p -valeur de t_m est inférieure à 5%.

11. La valeur $m0$ vaut 2 par défaut

4 Variables discrètes : Test de conformité à une distribution donnée

On suppose maintenant que X est une v.a. prenant ses valeurs dans l'ensemble fini $\mathcal{X} = \{\xi_1, \dots, \xi_M\}$ (Les éléments de cet ensemble ne sont pas forcément des nombres, ce sont les « modalités » de la v.a. X et peuvent être des éléments qualitatifs (sexe, couleur des yeux)).

On dispose d'une observation $x = (x_n)_{1 \leq n \leq N}$ d'un N -échantillon $(X_n)_{1 \leq n \leq N}$ de X et donc pour chaque $m \in \{1, \dots, M\}$, de la fréquence d'apparition de ξ_m dans l'observation x , à savoir

$$f_m = \frac{\#\{n \in \{1, \dots, N\}, x_n = \xi_m\}}{N} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{\{x_n = \xi_m\}}$$

On peut poser :

$$\forall m \in \{1, \dots, M\}, F_m = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{\{X_n = \xi_m\}}.$$

5 Chaque v.a. $N.F_m$ est une v.a. $\sim \mathcal{B}(N, \mathbb{P}(X = \xi_m))$ et $f = (f_m)_{1 \leq m \leq M}$ est une réalisation du vecteur aléatoire $F = (F_m)_{1 \leq m \leq M}$.

Pour un vecteur de probabilité $p = (p_m)_{1 \leq m \leq M}$ (avec $\forall m, p_m > 0$), on se pose la question de la conformité de cet ensemble de fréquences avec l'hypothèse :

$$H_0 : \forall m \in \{1, \dots, M\}, \mathbb{P}(X = \xi_m) = p_m$$

Il s'agit donc de tester la conformité de l'observation des fréquences avec une hypothèse portant sur la distribution de X .

C'est l'objectif du test du χ^2 de PEARSON, basé sur le théorème limite suivant, dans l'esprit du théorème de la limite centrale :

Théorème 5. Si $p = (\mathbb{P}(X = \xi_m))_{1 \leq m \leq M}$ alors, en posant

$$\Delta = N \cdot \sum_{m=1}^M \frac{(F_m - p_m)^2}{p_m},$$

pourvu que N soit assez grand, Δ est approximativement une v.a. de loi $\chi^2(M-1)$.

Cela donne le test suivant de conformité de l'observation au fait que de la distribution de X soit donnée par le vecteur p .

10 Pourvu que N soit suffisamment¹² grand :

1. On calcule sur les données $\delta = N \cdot \sum_{m=1}^M \frac{(f_m - p_m)^2}{p_m}$ qui est une réalisation de Δ . On remarque que si H_0 est supposée vraie alors Δ est distribuée suivant une loi $\chi^2(M-1)$.
2. On calcule (table ou fonction Python adéquate) $\pi(\delta)$, la p -valeur de δ pour la distribution $\chi^2(M-1)$ et
 - (a) Si $\pi(\delta) \leq 0.05$, on rejette H_0 au niveau de confiance 5% ;
 - 15 (b) si $\pi(\delta) \geq 0.05$, on accepte H_0 au niveau de confiance 5% ;

4.1 Travail demandé

Dans un script nommé `experimentation-chi2.py` :

1. (a) Ecrire une fonction `pValeurChi2(delta, d=5)` qui retourne la p -valeur décrite précédemment de la quantité `delta`. Faire afficher (cas $d = 5$) le graphe de cette p -valeur en fonction de `delta`. On pourra utiliser la fonction `chi2.cdf` du module `scipy.stats`. On utilisera l'approximation, valable pour $\delta \gg d - 2$, `pValeurChi2(delta, d=5) \simeq 1 - chi2.cdf(delta)`.
 - (b) En utilisant la fonction `chi2.ppf` du module `scipy.stats` (fonction des quantiles d'une v.a. $\chi^2(d)$), écrire une fonction `a_Chi2(p, d=5)` qui retourne la valeur de $\alpha \in \mathbb{R}^+$ telle que `pValeurChi2(alpha) = p`.
 - (c) Ecrire, à la main, sur votre feuille, pour d variant de 1 à 15, la liste des valeurs α_d telles que `pValeurChi2(alpha_d) = 0.05`. (Faites évidemment le calcul à la machine...)
2. Ecrire une fonction `EchantillonXFinie(p = [1/6]*6, NS = 100)` qui retourne une réalisation d'un NS-échantillon d'une variable X à valeurs dans $\{0, \dots, \text{len}(p) - 1\}$ dont la loi est donnée par la liste `p`.
3. Etant donné une série statistique numérique $x = (x_1, \dots, x_N)$ réalisation d'un échantillon d'une v.a. X à valeurs dans $\{0, \dots, M - 1\}$ et un vecteur de probabilité p , écrire une fonction `Delta(x, p)` calculant δ_p la réalisation (calculable à partir de cette série x et du vecteur p) de la variable aléatoire Δ définie dans le texte.
4. Imaginez que vous menez $NE = 1000$ fois la suite d'opérations suivante :
 - (a) Tirer au sort un échantillon `x = EchantillonXFinie()` ;
 - (b) Calculer `delta = Delta(x, p)` et `pv = pValeurChi2(delta)` ;
 Dans combien (approximativement) de cas allez vous observer le fait que `delta` est inférieur à `a_Chi2(0.05, d = 5)` ?
 - 35 Pouvez vous écrire une fonction `CheckTheorieChi2(p = [1/6]*6, NS = 100, NE = 1000)` qui vérifie ce fait ?
5. (Théorie) Vérifier que pour le cas $M = 1$, le test proposé est équivalent au test de conformité de moyenne pour une variable $X \sim \mathcal{B}(p)$.

12. Ceci dépend essentiellement de la valeur du vecteur p