Idup / Master 1

Analyse statistique / JFL

Support de cours

La décomposition de la variance

Le tableau 1 présente les salaires moyens selon le sexe (Hommes et femmes) et la catégorie professionnelle (Cadres et Non Cadres) des salariés des entreprises du secteur privé et public

Tableau 1. Salaires moyens (en équivalent temps plein) selon le sexe et la catégorie sociale en 2014

Catégorie sociale	Hommes		Femmes		Ensemble	
	Salaire moyen	Part de la pop. Totale (%)	Salaire moyen	Part de la pop. Totale (%)	Salaire moyen	Part de la pop. Totale (%)
Cadres	4 407	11,8	3 524	6,2	**4 104**	**18,0**
Non cadres	1 908	47,0	1 685	35,0	**1 813**	**82,0**
Ensemble	**2 410**	**58,8**	**1 961**	**41,2**	**2 225**	**100,0**

Source : Insee, DADS, fichier semi-définitif.

L'objectif est de déterminer laquelle de ces deux caractéristiques (le sexe ou la catégorie professionnelle) contribue le plus à la variance (ou dispersion des valeurs).

On commence par déterminer la variance totale :

$$Var(S) = f_{C,H} \times \left(\bar{s}_{C,H} - \bar{\bar{s}}\right)^2 + f_{C,F} \times \left(\bar{s}_{C,F} - \bar{\bar{s}}\right)^2 + f_{NC,H} \times \left(\bar{s}_{NC,H} - \bar{\bar{s}}\right)^2 + f_{NC,F} \times \left(\bar{s}_{NC,F} - \bar{\bar{s}}\right)^2$$

$$Var(S) = \frac{11{,}8}{100{,}0} \times (4\,407 - 2\,225)^2 + \frac{6{,}2}{100{,}0} \times (3\,524 - 2\,225)^2 + \frac{47{,}0}{100{,}0} \times (1\,908 - 2\,225)^2 + \frac{35{,}0}{100{,}0} \times (1\,685 - 2\,225)^2 = 816\,278\ euros^2$$

$$\sigma(S) = \sqrt{Var(S)} = \sqrt{816\,278} = 903\ euros$$

On va maintenant décomposer cette variance en deux : la variance inter-groupe et la variance intra-groupe. Cette décomposition peut s'opérer de deux façons différentes, selon le choix de la variable qui va distinguer les deux groupes.

1) <u>On distingue les deux groupes selon la catégorie sociale</u>

On commence par calculer la variance inter-groupe, c'est-à-dire la contribution des écarts de salaire entre catégorie sociale à la variance totale des salaires. On détermine donc la moyenne des carrés des écarts entre les salaires moyens des différentes catégories sociales et la moyenne des salaires :

$$Var\ intergroupe = f_C \times (\bar{s}_C - \bar{\bar{s}})^2 + f_{NC} \times (\bar{s}_{NC} - \bar{\bar{s}})^2$$
$$= \frac{18{,}0}{100} \times (4\,104 - 2\,225)^2 + \frac{82{,}0}{100} \times (1\,813 - 2\,225)^2 = 774\,705\ euros^2$$

Puis on calcule la variance intra-groupe, c'est-à-dire la contribution des écarts de salaire selon le sexe à la variance totale. Ce calcul se fait en deux temps :

- On détermine la variance des salaires selon le sexe au sein de chacun des groupes (ici au sein de chaque catégorie sociale) :

> Variance des salaires selon le sexe parmi les Cadres :

$$Var(S/C) = f_{H/C} \times \left(\bar{s}_{C,H} - \bar{s}_C\right)^2 + f_{F/C} \times \left(\bar{s}_{C,F} - \bar{s}_C\right)^2$$
$$= \frac{11{,}8}{18{,}0} \times (4\,407 - 4\,104)^2 + \frac{6{,}2}{18{,}0} \times (3\,524 - 4\,104)^2 = 175\,749\ euros^2$$

> Variance des salaires selon le sexe parmi les Non Cadres :

$$Var(S/NC) = f_{H/NC} \times \left(\bar{s}_{NC,H} - \bar{s}_{NC}\right)^2 + f_{F/NC} \times \left(\bar{s}_{NC,F} - \bar{s}_{NC}\right)^2$$
$$= \frac{47{,}0}{82{,}0} \times (1\,908 - 1\,813)^2 + \frac{35{,}0}{82{,}0} \times (1\,685 - 1\,813)^2 = 12\,119\ euros^2$$

- Puis on fait la moyenne de ces variances selon le sexe pondérée par le poids de chacune des catégories sociales dans la population totale.

$$Var\ intragroupe = f_C \times Var(S/C) + f_{NC} \times Var(S/NC) = \frac{18{,}0}{100} \times 175\,749 + \frac{82{,}0}{100} \times 12\,119$$
$$= 41\,573\ euros^2$$

On vérifie que :

$$Var\ intergroupe + Var\ intragroupe = Var\ totale$$

$$774\,705 + 41\,573 = 816\,278\ euros^2$$

La variance totale est donc la somme de la variance des moyennes (des salaires selon la catégorie sociale = variance intergroupe) et de la moyenne des variances (ici des salaires selon le sexe = variance intragroupe). Dans le cas présent, on peut donc dire que 95 % de la variance des salaires est expliquée par les écarts selon la catégorie sociale.

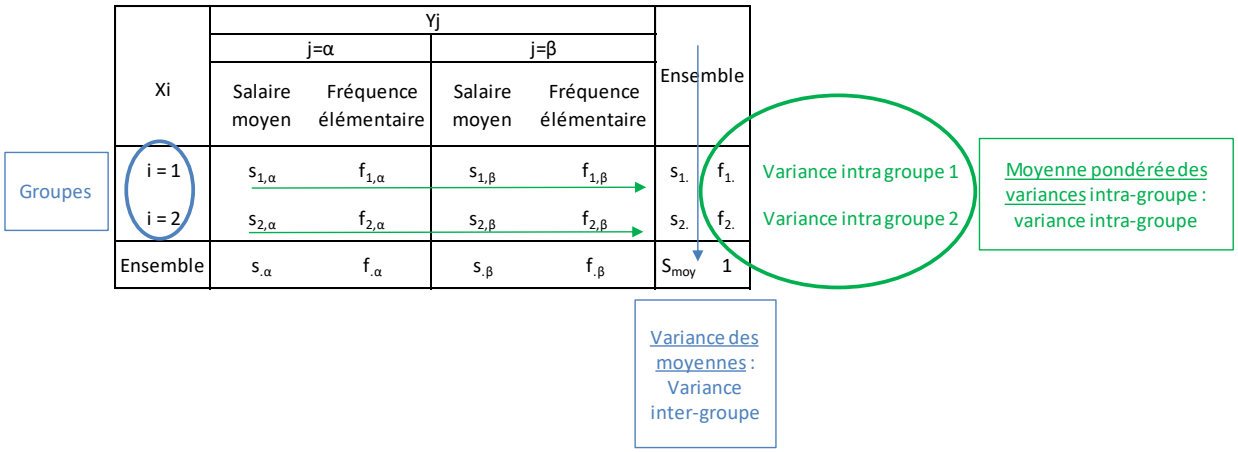
<u>Formalisation (généralisation) du calcul</u>

Nous allons démontrer de manière théorique ce que nous avons constaté de manière empirique, à savoir que :

$$Var\ totale = Var\ intergroupe + Var\ intragroupe$$

$$f_{C,H} \times \left(\bar{s}_{C,H} - \bar{\bar{s}}\right)^2 + f_{C,F} \times \left(\bar{s}_{C,F} - \bar{\bar{s}}\right)^2 + f_{NC,H} \times \left(\bar{s}_{NC,H} - \bar{\bar{s}}\right)^2 + f_{NC,F} \times \left(\bar{s}_{NC,F} - \bar{\bar{s}}\right)^2$$
$$= (f_C \times (\bar{s}_C - \bar{\bar{s}})^2 + f_{NC} \times (\bar{s}_{NC} - \bar{\bar{s}})^2)$$
$$+ \left(f_C \times \left(f_{H/C} \times \left(\bar{s}_{C,H} - \bar{s}_C\right)^2 + f_{F/C} \times \left(\bar{s}_{C,F} - \bar{s}_C\right)^2\right)\right.$$
$$\left. + f_{NC} \times \left(f_{H/NC} \times \left(\bar{s}_{NC,H} - \bar{s}_{NC}\right)^2 + f_{F/NC} \times \left(\bar{s}_{NC,F} - \bar{s}_{NC}\right)^2\right)\right)$$

Le tableau 1 peut être généralisé de la façon suivante :



La relation entre la variance totale et les variances inter-groupe et intra-groupe peut être formalisée de la manière suivante :

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - \bar{\bar{s}}\right)^2 = \sum_{i=1}^{2} f_{i.} \times (s_{i.} - \bar{\bar{s}})^2 + \sum_{i=1}^{2} f_{i.} \sum_{j=\alpha}^{\beta} f_{j/i} \times \left(s_{j/i} - s_{i.}\right)^2$$

Pour arriver à cette relation, il faut commencer par faire apparaître dans la formulation de la variance totale les salaires moyens des groupes définis par la variable X (ici la catégorie sociale) :

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - \bar{\bar{s}}\right)^2 = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - s_{i.} + s_{i.} - \bar{\bar{s}}\right)^2 = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(\left(s_{ij} - s_{i.}\right) + \left(s_{i.} - \bar{\bar{s}}\right)\right)^2$$

On développe ensuite cette relation de type $(a+b)^2 = a^2 + 2ab + b^2$, avec :

$a = s_{ij} - s_{i.}$ et $b = s_{i.} - \bar{\bar{s}}$

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(\left(s_{ij} - s_{i.}\right) + \left(s_{i.} - \bar{\bar{s}}\right)\right)^2$$

$$= \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(\left(s_{ij} - s_{i.}\right)^2 + 2 \times \left(s_{ij} - s_{i.}\right)\left(s_{i.} - \bar{\bar{s}}\right) + \left(s_{i.} - \bar{\bar{s}}\right)^2\right)$$

On décompose ensuite cette relation (qui correspond à la variance totale) en une somme de trois termes :

$$Var\ totale = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - s_{i.}\right)^2 + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times 2 \times \left(s_{ij} - s_{i.}\right)\left(s_{i.} - \bar{\bar{s}}\right) + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{i.} - \bar{\bar{s}}\right)^2$$

Il faut ensuite faire apparaître les fréquences marginales et conditionnelles nécessaires à la formalisation de variances. On formalise donc les fréquences élémentaires en un produit d'une fréquence marginale et d'une fréquence conditionnelle. Ici, la variable qui distingue les groupes est

la variable X (la catégorie sociale) de modalités i (Cadres et Non Cadres). Il faut donc faire apparaître les fréquences marginales pour cette variable ($f_{i.}$) et les fréquences conditionnelles au sein de chacun de ces groupes ($f_{j/i}$).

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{i.}} \times \frac{n_{i.}}{N} = f_{j/i} \times f_{i.}$$

Chacun des trois termes de la somme mise en évidence précédemment s'écrit donc :

$$Var\ totale = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times \left(s_{ij} - s_{i.}\right)^2 + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times 2 \times \left(s_{ij} - s_{i.}\right)(s_{i.} - \bar{\bar{s}}) + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times (s_{i.} - \bar{\bar{s}})^2$$

Identification du sens de chaque terme :

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times \left(s_{ij} - s_{i.}\right)^2 = \sum_{i=1}^{2} f_{i.} \sum_{j=\alpha}^{\beta} f_{j/i} \times \left(s_{ij} - s_{i.}\right)^2 = Variance\ intragroupe$$

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times 2 \times \left(s_{ij} - s_{i.}\right)(s_{i.} - \bar{\bar{s}}) = 2 \times \sum_{i=1}^{2} f_{i.} \times (s_{i.} - \bar{\bar{s}}) \sum_{j=\alpha}^{\beta} f_{j/i} \times \left(s_{ij} - s_{i.}\right) = 0$$

Il s'agit de la moyenne pondérée des écarts à la moyenne des salaires des groupes (catégories sociales). Par définition de la moyenne, la moyenne des écarts à la moyenne est égale à 0.

Il s'agit de la moyenne pondérée des écarts à la moyenne des salaires des sous-catégories au sein des groupes (salaires des sexes au sein de chaque catégorie sociale). Là encore, par définition de la moyenne, la moyenne des écarts à la moyenne est égale à 0.

$$\sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{j/i} \times f_{i.} \times (s_{i.} - \bar{\bar{s}})^2 = \sum_{j=\alpha}^{\beta} f_{j/i} \sum_{i=1}^{2} f_{i.} \times (s_{i.} - \bar{\bar{s}})^2 = \sum_{i=1}^{2} f_{i.} \times (s_{i.} - \bar{\bar{s}})^2 = variance\ intergroupe$$

Dans le cas présent, les valeurs à pondérer sont seulement les carrés des écarts entre les moyennes marginales de la variable X (qui prend les modalités i) et la moyenne générale. À aucun moment les modalités j de la variable Y sont utilisées. De plus, pour chaque modalité de la variable i, la somme des fréquences conditionnelles fj/i est égale à 1. Ce terme vaut donc 1.

2) <u>On distingue les deux groupes selon le sexe</u>

Cette fois-ci la variance inter-groupe correspond à la contribution des écarts de salaire entre sexes à la variance totale des salaires. On détermine donc la moyenne des carrés des écarts entre les salaires moyens des hommes et des femmes et la moyenne des salaires :

$$Var\ intergroupe = f_H \times (\bar{s}_H - \bar{\bar{s}})^2 + f_F \times (\bar{s}_F - \bar{\bar{s}})^2$$
$$= \frac{58{,}8}{100} \times (2\,410 - 2\,225)^2 + \frac{41{,}2}{100} \times (1\,961 - 2\,225)^2 = 48\,850\ euros^2$$

Puis on calcule la variance intra-groupe, c'est-à-dire dans ce cas la contribution des écarts de salaire selon la catégorie sociale à la variance totale. Ce calcul se fait en deux temps :

- On détermine la variance des salaires selon la catégorie sociale au sein de chacun des groupes (ici au sein de chaque sexe) :

> Variance des salaires selon la catégorie sociale parmi les Hommes :

$$Var(S/H) = f_{C/H} \times \left(\bar{s}_{C,H} - \bar{s}_H\right)^2 + f_{NC/H} \times \left(\bar{s}_{NC,H} - \bar{s}_H\right)^2$$
$$= \frac{11{,}8}{58{,}8} \times (4\,407 - 2\,410)^2 + \frac{47{,}0}{58{,}8} \times (1\,908 - 2\,410)^2 = 1\,002\,941\ euros^2$$

> Variance des salaires selon la catégorie sociale parmi les Femmes :

$$Var(S/F) = f_{C/F} \times \left(\bar{s}_{C,F} - \bar{s}_F\right)^2 + f_{NC/F} \times \left(\bar{s}_{NC,F} - \bar{s}_F\right)^2$$
$$= \frac{6{,}2}{41{,}2} \times (3\,524 - 1\,961)^2 + \frac{35{,}0}{41{,}2} \times (1\,685 - 1\,961)^2 = 430\,981\ euros^2$$

- Puis on fait la moyenne de ces variances selon la catégorie sociale pondérée par le poids de chacun des sexes dans la population totale.

$$Var\ intragroupe = f_H \times Var(S/H) + f_F \times Var(S/F) = \frac{58{,}8}{100} \times 1\,002\,941 + \frac{41{,}2}{100} \times 430\,981$$
$$= 767\,428\ euros^2$$

On vérifie que :

$$Var\ intergroupe + Var\ intragroupe = Var\ totale$$

$$48\,850 + 767\,428 = 816\,278\ euros^2$$

La variance totale est donc bien, là encore, la somme de la variance des moyennes (des salaires selon le sexe = variance intergroupe) et de la moyenne des variances (ici des salaires selon la catégorie sociale = variance intragroupe).

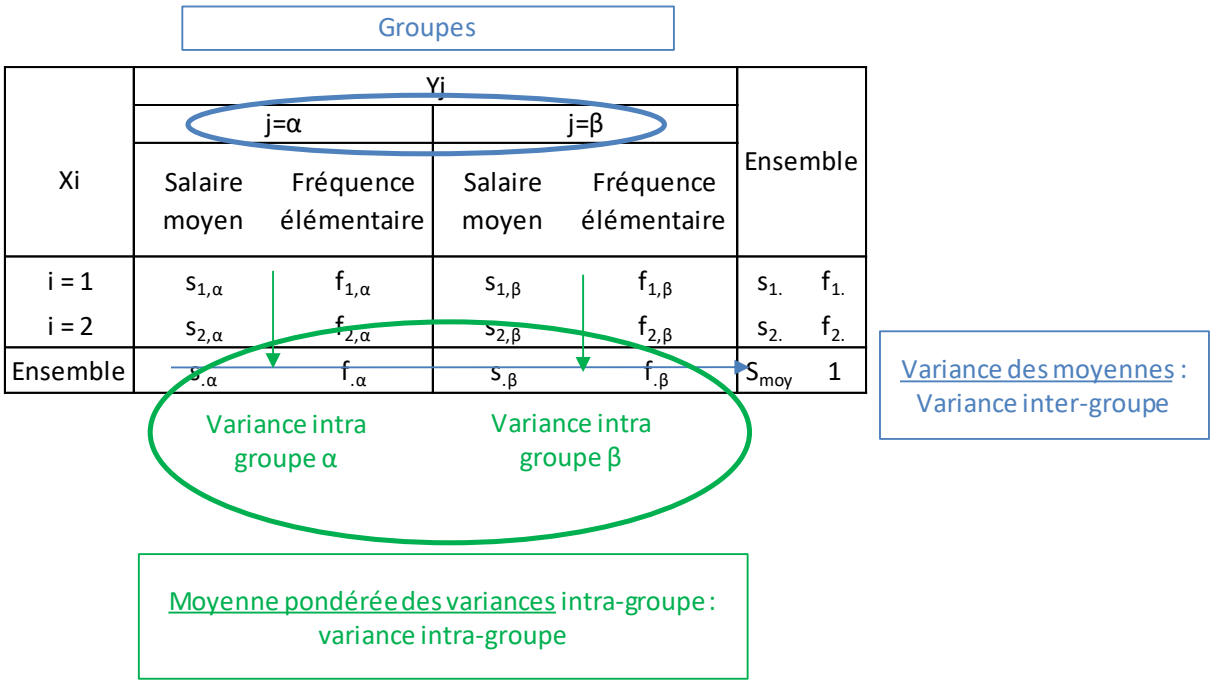
Dans le cas présent, on peut donc dire que 94 % de la variance des salaires est expliquée par les écarts selon la catégorie sociale. Le constat est donc le même que celui fait précédemment bien que le résultat soit très légèrement différent. Si les deux options de calcul conduisent au même constat, les résultats ne sont en revanche pas exactement les mêmes.

Généralisation de la démarche

La démarche est la même que celle détaillée quand les groupes sont distingués selon la catégorie sociale.

Ici, il faut donc faire apparaître les moyennes marginales de la variable Y ainsi que les fréquences marginales de cette variable et les fréquences conditionnelles au sein de chacun des deux groupes définis par cette variable.

Schématiquement, le calcul est donc le suivant :



La formalisation de la démonstration s'écrit :

$$Var\ Totale = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - \bar{\bar{s}}\right)^2 = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - s_{.j} + s_{.j} - \bar{\bar{s}}\right)^2$$

$$= \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(\left(s_{ij} - s_{.j}\right) + \left(s_{.j} - \bar{\bar{s}}\right)\right)^2$$

$$Var\ Totale = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(\left(s_{ij} - s_{.j}\right)^2 + 2 \times \left(s_{ij} - s_{.j}\right)\left(s_{.j} - \bar{\bar{s}}\right) + \left(s_{.j} - \bar{\bar{s}}\right)^2\right)$$

$$= \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{ij} - s_{.j}\right)^2 + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times 2 \times \left(s_{ij} - s_{.j}\right)\left(s_{.j} - \bar{\bar{s}}\right)$$

$$+ \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{ij} \times \left(s_{.j} - \bar{\bar{s}}\right)^2$$

Comme :

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{ij}}{n_{.j}} \times \frac{n_{.j}}{N} = f_{i/j} \times f_{.j}$$

On peut écrire :

$$Var\ totale = \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{i/j} \times f_{.j} \times \left(s_{ij} - s_{.j}\right)^2 + \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{i/j} \times f_{.j} \times 2 \times \left(s_{ij} - s_{.j}\right)\left(s_{.j} - \bar{\bar{s}}\right)$$
$$+ \sum_{i=1}^{2}\sum_{j=\alpha}^{\beta} f_{i/j} \times f_{.j} \times \left(s_{.j} - \bar{\bar{s}}\right)^2$$

$$Var\ totale = \sum_{j=\alpha}^{\beta} f_{.j} \sum_{i=1}^{2} f_{\frac{i}{j}} \times \left(s_{ij} - s_{.j}\right)^2 + 0 + \sum_{i=1}^{2} f_{\frac{i}{j}} \sum_{j=\alpha}^{\beta} f_{.j} \times \left(s_{.j} - \bar{\bar{s}}\right)^2$$
$$= \sum_{j=\alpha}^{\beta} f_{.j} \sum_{i=1}^{2} f_{\frac{i}{j}} \times \left(s_{ij} - s_{.j}\right)^2 + \sum_{j=\alpha}^{\beta} f_{.j} \times \left(s_{.j} - \bar{\bar{s}}\right)^2$$
$$= Variance\ intragroupe + variance\ intergroupe$$