

Exercice 3 : La distribution des salaires dans le secteur privé et semi-public en 2004

CORRIGÉ

- 1) Calculez la variance totale des salaires à partir des données du tableau 1.
- 2) Puis décomposez la variance des salaires. Quelle variable pèse le plus dans la décomposition de la variance : le sexe ou la PCS ?

Tableau 1 : Salaires annuels nets moyens (en euros) des postes à temps complet et répartition des effectifs en 2004 (en %) selon le sexe et la profession et la catégorie sociale (PCS) dans le secteur privé et semi-public

PCS	Hommes			Femmes			Ensemble	
	% con- dit.	% élé- ment.	salaire	% con- dit.	% élé- ment.	salaire	%	salaire
Cadres*	18,6		46 514	12,5			16,4	43 569
Profs. Intermédiaires	22,5		23 693	28,1			24,5	22 464
Employés	11,1		16 259	44,4			23,1	15 441
Ouvriers	47,8		16 730	15,0			36,0	16 419
Ensemble	100,0		23 784	100,0			100,0	22 129

* Y compris chefs d'entreprise salariés.

Champ : salariés à temps complet du secteur privé et semi-public.

Source : DADS, Insee (fichier 2004 semi-définitif).

Corrigé

Il faut commencer par compléter le tableau 1. Voici une manière de le faire.

On commence par estimer les fréquences élémentaires à partir des données dont on dispose. Par exemple pour la catégorie des cadres :

On sait que : $f_C = f_{C,H} + f_{C,F}$

On connaît f_C qui est égal à 16,4 %.

On cherche précisément les fréquences élémentaires (parts des cadres masculins et des cadres féminins dans la population totale). On peut les formuler autrement :

$$f_{C,H} = \frac{n_{C,H}}{N} = \frac{n_{C,H}}{n_H} \times \frac{n_H}{N} = f_{C/H} \times f_H$$

D'une manière plus générale, on peut écrire : $f_{i,j} = f_{i/j} \times f_j$

Pour les femmes cadres, la relation est donc :

$$f_{C,F} = f_{C/F} \times f_F$$

La part des cadres dans la population totale s'écrit donc :

$$f_C = f_{C,H} + f_{C,F} = f_{C/H} \times f_H + f_{C/F} \times f_F$$

On connaît trois des cinq termes de cette relation : f_C , $f_{C/H} = 18,6\%$, $f_{C/F} = 12,5\%$

Il est facile d'en déduire les proportions d'hommes et de femmes. En effet, comme : $f_H + f_F = 1$,

$$f_C = f_{C/H} \times f_H + f_{C/F} \times (1 - f_H)$$

$$f_C = f_{C/H} \times f_H + f_{C/F} - f_{C/F} \times f_H$$

$$f_C - f_{C/F} = f_H \times (f_{C/H} - f_{C/F})$$

$$f_H = \frac{f_C - f_{C/F}}{f_{C/H} - f_{C/F}}$$

$$f_F = 1 - f_H$$

Application numérique :

$$f_H = \frac{16,4 - 12,5}{18,6 - 12,5} = 0,64$$

$$f_F = 0,36$$

On est maintenant en mesure de calculer les fréquences relatives élémentaires :

$$f_{C,H} = 18,6 \times 0,64 = 11,9 \%$$

$$f_{C,F} = 12,5 \times 0,36 = 4,5 \%$$

Avec ces fréquences élémentaires, il est également possible de déterminer le salaire moyen des femmes cadres :

$$SM_C = SM_{C,H} \times f_{H/C} + SM_{C,F} \times f_{F/C}$$

$$SM_C = SM_{C,H} \times \frac{f_{C,H}}{f_C} + SM_{C,F} \times \frac{f_{C,F}}{f_C}$$

$$SM_C \times f_C = SM_{C,H} \times f_{C,H} + SM_{C,F} \times f_{C,F}$$

$$SM_{C,F} = \frac{SM_C \times f_C - SM_{C,H} \times f_{C,H}}{f_{C,F}} = \frac{43\,569 \times 16,4 - 46\,514 \times 11,9}{4,5} = 35\,780 \text{ €}$$

On peut faire cette série de calculs pour toutes les PCS et ainsi compléter le tableau 1. On peut calculer le salaire moyen des femmes de deux façons différentes :

$$SM_F = \sum_{i=1}^4 (f_{i/F} \times SM_{i,F}) = 0,125 \times 35\,780 + \dots + 0,15 \times 14\,659 = 19\,187 \text{ €}$$

$$SM = f_H \times SM_H + f_F \times SM_F$$

$$SM_F = \frac{SM - f_H \times SM_H}{f_F} = \frac{22\,129 - 0,64 \times 23\,784}{0,36} = 19\,187 \text{ €}$$

Tableau récapitulatif :

PCS	Hommes			Femmes			Ensemble	
	%	% élémentaire	salaire	%	% élémentaire	salaire	%	salaire
Cadres*	18,6	11,9	46 514	12,5	4,5	35 780 €	16,4	43 580
Profs. Intermédiaires	22,5	14,4	23 693	28,1	10,1	20 715 €	24,5	22 479
Employés	11,1	7,1	16 259	44,4	16,0	15 078 €	22,9	15 568
Ouvriers	47,8	30,6	16 730	15,0	5,4	14 659 €	36,2	16 325
Ensemble	100,0	64,0	23 784	100,0	36,0	19 187 €	100,0	22 129

Commentaires :

Quelle que soit la PCS, les femmes ont dans le secteur privé et semi-public des salaires inférieurs à ceux des hommes. La différence est particulièrement importante pour les cadres : au sein de cette catégorie, les hommes ont un salaire 1,3 fois plus important que celui des femmes, soit une différence de 30 % ! Ce rapport est égal à 1,14 pour les professions intermédiaires et les ouvriers, et il est de 1,08 pour les employés.

1) La variance totale du salaire ($Var(S)$) correspond ici à la moyenne des carrés des écarts des différents salaires (s_{ij}) par rapport au salaire moyen (\bar{s}). Dans le cas présent, il y a, au carrefour des variables sexe (deux modalités) et PCS (4 modalités), 8 sous-populations. Chacune est définie par le sexe (y_j) et la PCS (x_i).

$$Var(S) = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (s_{ij} - \bar{s})^2$$

$$Var(S) = \left[f_{Cadres, Hommes} \times (s_{Cadres, Hommes} - \bar{s})^2 \right] + \dots + \left[f_{Ouvriers, Femmes} \times (s_{Ouvriers, Femmes} - \bar{s})^2 \right]$$

$$Var(S) = \left[0,119 \times (46\,514 - 22\,129)^2 \right] + \dots + \left[0,054 \times (14\,659 - 22\,129)^2 \right]$$

$$Var(S) = 102\,049\,757 \text{ €}^2$$

$$\sigma_s = \sqrt{Var(S)} = \sqrt{102\,049\,757} = 10\,102 \text{ €}$$

$$CV = \frac{\sigma_s}{\bar{s}} = \frac{10\,102}{22\,129} = 46 \%$$

La variance est de 102 049 757 €² et l'écart-type est de 10 102 €. Le coefficient de variation, qui est le rapport entre l'écart-type et la moyenne est de 46 %, ce qui est assez élevé.

2) Il est possible de décomposer la variance et de mesurer la contribution respective du sexe et de la PCS sur les inégalités de salaires.

On peut commencer par calculer la variance des salaires selon le sexe. Il s'agit d'une mesure de la dispersion des salaires expliquée par cette caractéristique. On détermine cette variance à partir des salaires moyens selon le sexe. Il s'agit d'une variance de moyennes (VM).

$$\begin{aligned} \text{Var}(\bar{S}_{sexe}) &= \sum_{j=1}^2 f \cdot_j \times (\bar{s}_j - \bar{s})^2 \\ \text{Var}(\bar{S}_{sexe}) &= [0,64 \times (23\,784 - 22\,129)^2] + [0,36 \times (19\,187 - 22\,129)^2] \\ \text{Var}(\bar{S}_{sexe}) &= 4\,869\,501 \text{ €}^2 \\ \sigma_{\bar{S}_{sexe}} &= \sqrt{4\,869\,501} = 2\,207 \text{ €} \\ CV &= \frac{\sigma_{\bar{S}_{sexe}}}{\bar{S}} = \frac{2\,207}{22\,129} = 10\% \end{aligned}$$

Le coefficient de variation des salaires moyens selon le sexe est assez faible (10 %). Le sexe assure donc une faible part de la dispersion des salaires selon la PCS et le sexe. La PCS doit donc vraisemblablement expliquer la plus grande part de la dispersion des salaires.

On peut précisément calculer en complément de cette variance expliquée par le sexe, celle expliquée par la PCS. On ne peut toutefois pas calculer la variance expliquée par la PCS comme on vient de le faire pour la variable « sexe ». En effet, c'est deux caractéristiques ne sont pas indépendantes : les hommes sont par exemple proportionnellement plus nombreux à être cadres, tandis que les femmes sont de leur côté sur-représentées parmi les employés. Il y a un effet d'interaction entre les deux variables. De ce fait, la valeur de la variance liée à l'une des deux variables tient compte inévitablement (bien que de manière très réduite) de la variance liée à l'autre variable.

La valeur calculée précédemment de la variance expliquée par le sexe tient déjà compte de cette interaction. Il faut donc éviter de tenir compte une seconde fois de cette dernière. Pour cela, on va calculer la variance expliquée par la PCS au sein de chaque catégorie de sexe. Il s'agit donc à chaque fois de la variance expliquée par la PCS nette de l'influence du sexe (puisqu'on se situe au sein d'une même sous-population). On va ensuite calculer la moyenne de ces deux variances selon la PCS pondérée par le poids de chacun de ces deux groupes. De cette manière, on tient compte de l'inégale répartition des salariés selon la PCS au sein d'une même catégorie de personnes (les hommes, les femmes) pour le calcul de chaque variance, puis, au moment du calcul de la variance moyenne, du poids respectif de chacun de ces deux groupes. La prise en compte de ces deux variables est successive et non simultanée, raison pour laquelle on a neutralisé l'interaction entre les deux variables.

La variance calculée est donc ici une moyenne de deux variances ; on parle d'une moyenne des variances (MV).

Variance expliquée par la PCS au sein de chaque sexe :

$$\begin{aligned} \text{Var}(S_{PCS / Hommes}) &= \sum_{i=1}^4 f_{i / Hommes} \times (s_{i, Hommes} - \bar{s}_{Hommes})^2 \\ \text{Var}(S_{PCS / Femmes}) &= \sum_{i=1}^4 f_{i / Femmes} \times (s_{i, Femmes} - \bar{s}_{Femmes})^2 \end{aligned}$$

Généralisation :

$$\text{Var}(S_{i/j}) = \sum_{i=1}^4 f_{i/j} \times (s_{i,j} - \bar{s}_j)^2$$

La moyenne de ces deux variances s'écrit :

$$\overline{\text{Var}}(S_{i/j}) = \sum_{j=1}^2 f_{.j} \sum_{i=1}^4 f_{i/j} \times (s_{i,j} - \bar{s}_j)^2$$

Application numérique :

$$\text{Var}(S_{PCS / Hommes}) = [0,186 \times (46\,514 - 23\,784)^2] + \dots + [0,478 \times (16\,730 - 23\,784)^2]$$

$$\text{Var}(S_{PCS / Hommes}) = 126\,169\,509 \text{ €}^2$$

$$\sigma_{S_{PCS / Hommes}} = \sqrt{126\,169\,509} = 11\,233 \text{ €}$$

$$CV = \frac{\sigma_{S_{PCS / Hommes}}}{\bar{S}_{Hommes}} = \frac{11\,233}{23\,784} = 47 \%$$

$$\text{Var}(S_{PCS / Femmes}) = [0,125 \times (35\,780 - 19\,187)^2] + \dots + [0,150 \times (14\,659 - 19\,187)^2]$$

$$\text{Var}(S_{PCS / Femmes}) = 45\,643\,805 \text{ €}^2$$

$$\sigma_{S_{PCS / Femmes}} = \sqrt{45\,643\,805} = 6\,756 \text{ €}$$

$$CV = \frac{\sigma_{S_{PCS / Femmes}}}{\bar{S}_{Femmes}} = \frac{6\,756}{19\,187} = 35 \%$$

$$\overline{\text{Var}}(S_{PCS / Sexe}) = 0,64 \times 126\,169\,509 + 0,36 \times 45\,643\,805 = 97\,180\,255 \text{ €}^2$$

$$\sigma_{PCS / Sexe} = \sqrt{97\,180\,255} = 9\,858 \text{ €}$$

$$CV = \frac{\sigma_{PCS / Sexe}}{\bar{S}} = \frac{9\,858}{22\,129} = 45 \%$$

La dispersion des salaires selon la PCS est plus importante parmi les hommes (CV = 47 %) que parmi les femmes (CV = 35 %). Au total, la dispersion des salaires selon la PCS est plus importante que celle calculée à partir des salaires moyens selon le sexe.

De manière empirique, on arrive au constat selon lequel la somme de la variance des salaires moyens selon le sexe (variance des moyennes correspondant ici à la variance expliquée par le sexe) et de la moyenne des variances expliquées par la PCS pour chaque sexe (moyenne des variances correspondant ici à la mesure de la variance expliquée par la PCS) est égale à la variance totale :

$$VM + MV = VT$$

$$\text{Var}(\bar{S}_j) + \overline{\text{Var}}(S_{i/j}) = \text{Var}(S)$$

$$\text{Var}(S_{Sexe}) + \overline{\text{Var}}(S_{PCS / Sexe}) = \text{Var}(S)$$

$$4\,869\,501 + 97\,180\,255 = 102\,049\,756$$

La variance expliquée par le sexe est de 5 % environ (4 869 501 / 102 049 756), tandis que la PCS explique 95 % de la variance (97 180 255 / 102 049 756).

Démonstration-Formalisation statistique :

On peut démontrer cette relation. On fait intervenir dans la relation les moyennes marginales en colonnes, correspondant ici au salaire moyen pour chacun des sexes, toutes PCS confondues (\bar{s}_j) :

$$Var(S) = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (s_{ij} - \bar{s})^2$$

$$Var(S) = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (s_{ij} - \bar{s}_j + \bar{s}_j - \bar{s})^2$$

$$Var(S) = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} \left[(s_{ij} - \bar{s}_j) + (\bar{s}_j - \bar{s}) \right]^2$$

$$Var(S) = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} \left[(s_{ij} - \bar{s}_j)^2 + \left[2 \times (s_{ij} - \bar{s}_j) \times (\bar{s}_j - \bar{s}) \right] + (\bar{s}_j - \bar{s})^2 \right]$$

On peut faire apparaître trois termes :

$$a) \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (s_{ij} - \bar{s}_j)^2$$

$$b) \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} \left[2 \times (s_{ij} - \bar{s}_j) \times (\bar{s}_j - \bar{s}) \right]$$

$$c) \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (\bar{s}_j - \bar{s})^2$$

On peut réécrire chacun des termes en transformant les fréquences élémentaires par le produit d'une fréquence conditionnelle et d'une fréquence marginale. En effet :

$$\frac{n_{ij}}{n_{..}} = \frac{n_{ij}}{n_{.j}} \times \frac{n_{.j}}{n_{..}}$$

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

Sens de a) :

$$a = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} (s_{ij} - \bar{s}_j)^2$$

$$a = \sum_{i=1}^4 \sum_{j=1}^2 (f_{i/j} \times f_{.j}) (s_{ij} - \bar{s}_j)^2$$

$$a = \sum_{j=1}^2 f_{.j} \sum_{i=1}^4 f_{i/j} (s_{ij} - \bar{s}_j)^2$$

$$a = \sum_{j=1}^2 f_{.j} \times Var(S_{i/j})$$

$$a = \overline{Var}(S_{i/j})$$

Le terme (a) correspond donc à la moyenne des variances pour chaque sous-population définie par une modalité j de la variable Y. Dans le cas présent, il correspond à la moyenne des variances au sein

de chaque groupe défini uniquement par le sexe. Il s'agit donc de la moyenne des variances expliquées par la PCS au sein de chaque sexe.

Sens de b) :

$$b = \sum_{i=1}^4 \sum_{j=1}^2 (f_{i/j} \times f_{\cdot j}) \left[2 \times (s_{ij} - \bar{s}_j) \times (\bar{s}_j - \bar{s}) \right]$$

$$b = 2 \times \left[\sum_{i=1}^4 f_{i/j} (s_{ij} - \bar{s}_j) \right] \times \left[\sum_{j=1}^2 f_{\cdot j} (\bar{s}_j - \bar{s}) \right]$$

Le terme (b) correspond au produit de deux moyennes des écarts à la moyenne.

Le premier correspond à la moyenne pondérée des écarts des salaires des différentes catégories i de X (ici la PCS) parmi les personnes caractérisées par la modalité j de Y (ici le sexe) par rapport au salaire moyen de l'ensemble des personnes ayant en commun la modalité j de Y (ici le même sexe).

La seconde est la moyenne pondérée des écarts des salaires des personnes caractérisées par la modalité j de Y (ici le sexe) par rapport au salaire moyen de l'ensemble des personnes, quelle que soit la modalité j de Y (ici hommes et femmes réunis).

Par définition, la moyenne des écarts à la moyenne est égale à 0. Le terme (b) est donc nul.

Sens de c) :

$$c = \sum_{i=1}^4 \sum_{j=1}^2 f_{ij} \left[(\bar{s}_j - \bar{s})^2 \right]$$

$$c = \sum_{i=1}^4 \sum_{j=1}^2 (f_{i/j} \times f_{\cdot j}) (\bar{s}_j - \bar{s})^2$$

$$c = \sum_{i=1}^4 f_{i/j} \sum_{j=1}^2 f_{\cdot j} (\bar{s}_j - \bar{s})^2$$

$$c = \sum_{i=1}^4 f_{i/j} \text{Var}(\bar{S}_j) = \text{Var}(\bar{S}_j) \sum_{i=1}^4 f_{i/j} = \text{Var}(\bar{S}_j) \times 1 = \text{Var}(\bar{S}_j)$$

Le terme (c) est bien égal à la variance des salaires moyens selon la modalité i de Y, soit dans le cas présent le sexe. Il s'agit plus généralement de la variance des moyennes.

On a donc bien :

$$\text{Var}(S) = \text{Var}(\bar{S}_j) + \overline{\text{Var}(S_{i/j})}$$

$$\text{Var}(S) = \sum_{j=1}^2 f_{\cdot j} (\bar{s}_j - \bar{s})^2 + \sum_{j=1}^2 f_{\cdot j} \sum_{i=1}^4 f_{i/j} (s_{ij} - \bar{s}_j)^2$$

Attention : les variances s'additionnent mais pas les écarts-types.

Complément : on pourrait très bien décomposer la variance en calculant d'abord la variance des salaires moyens selon la PCS, puis la moyenne des variances selon le sexe au sein de chaque PCS. Dans le cas présent, la variance expliquée par la PCS tient compte de l'interaction entre les variables sexe et PCS, tandis que la variance sexe est nette de cette interaction (qui a donc déjà été prise en compte). On va donc trouver des résultats très légèrement différents par rapport à ceux qui viennent d'être calculés, mais les parts relatives de la variance totale expliquée par le sexe et la PCS sont les

mêmes. En outre, la somme de la variance des moyennes et de la moyenne des variances est toujours égale à la variance totale.

Application numérique :

a) Variance des moyennes (dispersion des salaires moyens selon la PCS)

$$\text{Var}(\bar{S}_{PCS}) = \sum_{i=1}^4 f_i \cdot (\bar{s}_i - \bar{s})^2$$

$$\text{Var}(\bar{S}_{PCS}) = [0,164 \times (43\,569 - 22\,129)^2] + \dots + [0,360 \times (16\,419 - 22\,129)^2]$$

$$\text{Var}(\bar{S}_{PCS}) = 97\,494\,897 \text{ €}^2$$

$$\sigma_{\bar{S}_{PCS}} = \sqrt{97\,494\,897} = 9\,874 \text{ €}$$

$$CV = \frac{\sigma_{\bar{S}_{PCS}}}{\bar{s}} = \frac{9\,874}{22\,129} = 45 \%$$

b) Moyenne des variances (dispersion moyenne des salaires selon le sexe au sein de chaque PCS)

$$\text{Var}(S_{\text{Sexe} / \text{Cadres}}) = \left[\frac{0,119}{0,164} \times (46\,514 - 43\,569)^2 \right] + \left[\frac{0,045}{0,164} \times (35\,780 - 43\,569)^2 \right] = 22\,937\,226 \text{ €}^2$$

$$\text{Var}(S_{\text{Sexe} / \text{Pr.int.}}) = \left[\frac{0,144}{0,245} \times (23\,693 - 22\,464)^2 \right] + \left[\frac{0,101}{0,245} \times (20\,715 - 22\,464)^2 \right] = 2\,148\,941 \text{ €}^2$$

$$\text{Var}(S_{\text{Sexe} / \text{Employés}}) = \left[\frac{0,071}{0,231} \times (16\,259 - 15\,441)^2 \right] + \left[\frac{0,160}{0,231} \times (15\,078 - 15\,441)^2 \right] = 297\,219 \text{ €}^2$$

$$\text{Var}(S_{\text{Sexe} / \text{Employés}}) = \left[\frac{0,306}{0,360} \times (16\,730 - 16\,419)^2 \right] + \left[\frac{0,054}{0,360} \times (14\,659 - 16\,419)^2 \right] = 546\,734 \text{ €}^2$$

$$\overline{\text{Var}}(S_{\text{Sexe} / PCS}) = (0,164 \times 22\,937\,226) + (0,245 \times 2\,148\,941) + (0,231 \times 297\,219) + (0,360 \times 546\,734)$$

$$\overline{\text{Var}}(S_{\text{Sexe} / PCS}) = 4\,554\,859 \text{ €}^2$$

$$\sigma_{\text{Sexe} / PCS} = \sqrt{4\,554\,859} = 2\,134 \text{ €}$$

$$CV = \frac{\sigma_{\text{Sexe} / PCS}}{\bar{s}} = \frac{2\,134}{22\,129} = 10 \%$$

c) Variance totale

$$\text{Var}(S) = \text{Var}(\bar{S}_{PCS}) + \overline{\text{Var}}(S_{\text{Sexe} / PCS})$$

$$\text{Var}(S) = 97\,494\,897 + 4\,554\,859 = 102\,049\,756$$

Sans surprise compte tenu de ce qui a été dit précédemment, la variance expliquée par la PCS rend compte de 95 % de la variance totale.