

Statistiques et informatique appliquées aux sciences sociales

Tableau croisé : définition et représentation graphique des écarts à l'indépendance

Mardi 3/10/2023 15h30-17h Salle D2 Censier

Licence de sciences sociales 3^e année
Université Paris 1 Panthéon Sorbonne

Plan de la séance

- Présentation de résultats rapides : % ligne et écarts à la moyenne, lecture rapide
- Retour aux effectifs et représentations graphiques
- Commentaire
- Choix d'autres variables
- Évaluer la force des écarts
- Quelle information est significative ?

Tableau des pourcentages en ligne

Effectifs genre	Priorités entre carrière et vie familiale			
	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	14,5%	11,1%	74,4%	100,0%
1:femme	12,3%	10,5%	77,2%	100,0%
Ensemble	13,4%	10,8%	75,8%	100,0%

écart à la $\frac{\text{pourcentage en ligne} - \text{pourcentage moyen}}{\text{pourcentage moyen}} = \frac{14,5\% - 13,4\%}{13,4\%} = +1,1 \text{ pts}$

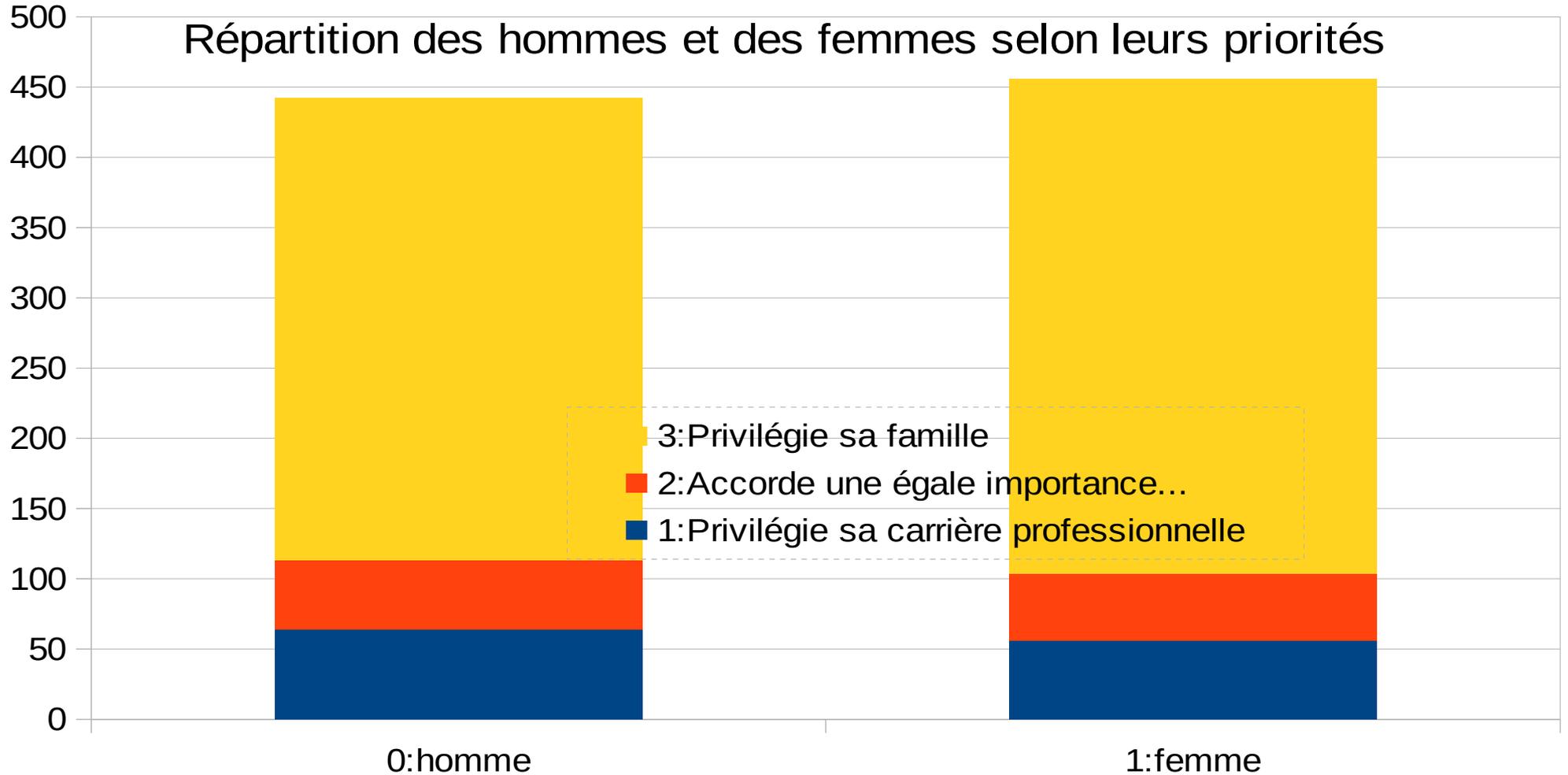
Écarts à la moyenne

Écarts à la moyenne	Priorités entre carrière et vie familiale			
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	+1,1 pts	+0,3 pts	-1,4 pts	0,0
1:femme	-1,1 pts	-0,3 pts	+1,4 pts	0,0

Information simplifiée (signes des écarts)

Profil	Priorités entre carrière et vie familiale		
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille
0:homme	+	+	-
1:femme	-	-	+

Représentation graphique des effectifs



Étude du lien entre deux variables

On cherche à étudier le lien entre les deux variables (en ligne et en colonne) en revenant aux effectifs et en construisant des indicateurs synthétiques et des représentations graphiques

- La quasi-totalité des indicateurs traitent de manière *symétrique* la variable en ligne et celle en colonne
- On part des effectifs constatés dans le tableau et on mesure d'abord des écarts en effectifs par rapport à une situation théorique, celle de l'indépendance entre les variables du tableau
- L'indépendance correspondrait à un tableau ayant les mêmes marges (totaux en ligne et en colonne) mais où les % en lignes seraient identiques aux % moyens (idem pour les % colonnes)

Calcul d'effectifs à l'indépendance

Effectifs	Priorités entre carrière et vie familiale			
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	59,1	?	?	442
1:femme	?	?	?	456
Ensemble	120	97	681	898

Définition d'un effectif à l'indépendance :

Dans la situation d'indépendance entre les variables en ligne et en colonne, l'effectif des hommes « carriériste » doit être tel que leur proportion parmi les hommes soit la même que celle des « carriéristes » parmi les deux sexes.

La condition s'écrit donc

$$\frac{\text{effectif à l'indépendance}}{442} = \frac{120}{898} \text{ soit } \frac{\text{effectif à l'indépendance}}{\text{total en ligne}} = \frac{\text{total en ligne} \times \text{proportion moyenne}}{\text{total en ligne}} = \frac{\text{total en ligne} \times \frac{\text{total en colonne}}{\text{total général}}}{\text{total général}} = \frac{\text{produits des marges}}{\text{total général}}$$

Tableau des effectifs à l'indépendance

Effectifs théoriques	Priorités entre carrière et vie familiale			
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	59,1	47,7	335,2	442
1:femme	60,9	49,3	345,8	456
Ensemble	120	97	681	898

$$59,1 = 13,4 \% * 442$$

On vérifie que la somme des éléments d'une ligne redonne bien le total :

$$13,4 \%*442 + 10,8 \%*442 + 75,8 \%*442 = (13,4\%+10,8\%+75,8\%)*442 = 100 \%*442 = 442$$

$$\text{Écrit autrement : } 59,1+49,3+335,2=442$$

La somme des éléments d'une colonne redonne aussi le total de la colonne en question :

$$13,4 \%*442+13,4 \%*456=13,4 \%*(442+456)=13,4 \%*898=120/898*898=120$$

$$\text{Écrit autrement : } 59,1+60,9=120$$

En fait, } $\frac{\text{produits des marges}}{\text{total général}}$ { est symétrique, ce qui signifie que les même propriétés se vérifient en colonnes et que toutes les marges sont identiques.

Écart à l'indépendance

écarts à l'indépendance	Priorités entre carrière et vie familiale			
	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	4,9	1,3	-6,2	0
1:femme	-4,9	-1,3	6,2	0
Ensemble	0	0	0	0

L'écart à l'indépendance mesure la différence entre l'effectif observé, par exemple 64 hommes « carriéristes », et l'effectif dans la situation que l'on aurait dans la situation (théorique) d'indépendance, c'est à dire si la même proportion d'hommes était « carriériste » que parmi les deux sexes

Soit $64 - 59,1 = +4,9$

Par construction, les écarts à l'indépendance se compensent sur une ligne ou une colonne.
Somme des écarts positifs = $6,2 + 4,9 + 1,3 = 12,4$

Principe de représentation des écarts

Les *surfaces* doivent être proportionnelles aux écarts à l'indépendance

- On part d'une échelle proportionnelle à l'effectif avec un carré de 10cm sur 10 cm représentant l'effectif total : 1 cm² représente donc 1 % de l'effectif (ici : environ 9 personnes)
- On veut aussi visualiser les écarts positifs et négatifs de part et d'autre d'une ligne horizontale, avec si possible des rectangles dont la *hauteur* est proportionnelle à l'écart à *la moyenne*
- On ajuste la largeur des rectangles au % colonne de chaque sexe
- On obtient ainsi trois profils (1 par modalité en ligne)
- La représentation trouve sa justification dans la formule (diapositive suivante)

Formules (représentation graphique)

Proportionnel à la surface des rectangles

La surface des rectangles est *algébrique* :
- si l'écart est négatif, le rectangle est positionné au dessous d'une ligne horizontale qui sert de référence
- si l'écart est positif, il est positionné au dessus.

$$\begin{aligned} \frac{\text{écart à l'indépendance}}{\text{effectif total}} &= \frac{\text{effectif case}}{\text{effectif total}} - \frac{\text{marge ligne} \cdot \text{marge colonne}}{[\text{effectif total}]^2} \\ &= \left[\frac{\text{effectif case}}{\text{marge ligne}} - \frac{\text{marge colonne}}{\text{effectif total}} \right] \cdot \frac{\text{marge ligne}}{\text{effectif total}} \\ &= \left[\text{écart à la moyenne} \right] \cdot \left[\text{pourcentage colonne} \right] \end{aligned}$$

Proportionnel à la *hauteur* des rectangles

Proportionnel à la *largeur* des rectangles

Comme pour les surfaces, cette hauteur est *algébrique* (+ vers le haut, - vers le bas).

Données pour la représentation graphique des écarts

Écarts à la moyenne		Priorités entre carrière et vie familiale		
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	% colonne
0:homme	+1,1 pts	+0,3 pts	-1,4 pts	49,2 %
1:femme	-1,1 pts	-0,3 pts	+1,4 pts	50,8 %

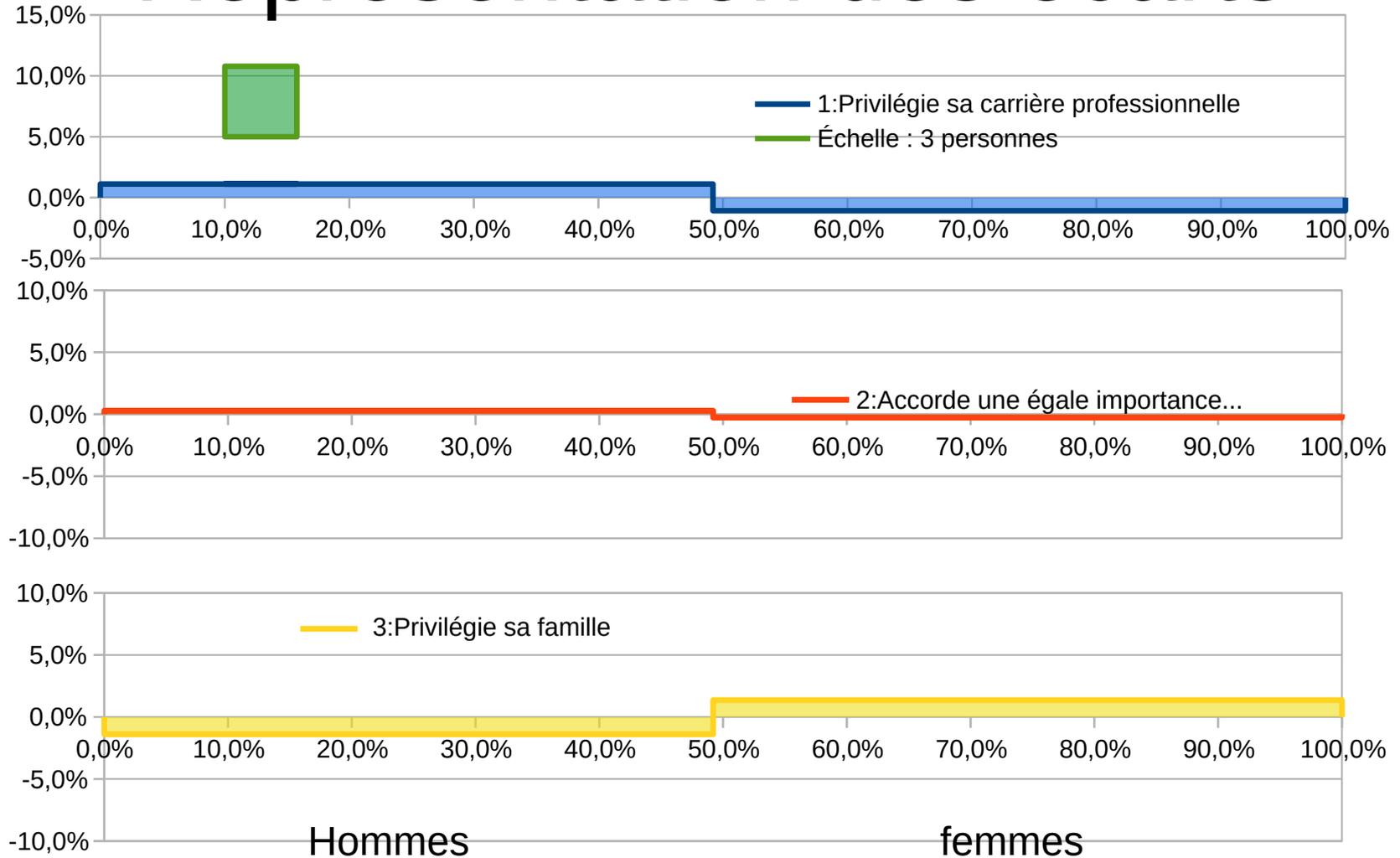
Vérification sur la 1ère case du tableau :

$$\frac{+1,1}{100} \times \frac{49,2}{100} \times 898 \approx +4,9$$

Échelle : 100 % de l'effectif soit 898 personnes seront représentées par un carré de 10x10 cm². Si on prend un carré de 1x1 cm² on obtient un centième de cet effectif soit un peu moins de 9 personnes. Attention : un rectangle de 10 cm sur 1 cm représente bien lui 10 % de l'effectif (89). Sur la diapositive suivante nous avons calculé une échelle représentant 3 personnes soit $3/898 \approx 0,334$ % de l'effectif soit 0,334 cm².

Pour obtenir un carré de cette surface on cherche la longueur d'un côté $\sqrt{0,334} \approx 0,578$ celle-ci nous est donnée par la racine carrée de 0,334, on calcule soit 0,6 cm de côté environ

Représentation des écarts



Lecture rapide de quelques tableaux avec d'autres variables de la même enquête

- Quelques rappels : distinction entre ce que dit le tableau et ce qu'on en dit
- Principe d'organisation : du général au particulier
- Principe de parcimonie : sélectionner et synthétiser l'information
- => omettre certaines informations déjà connues ou non essentielles
- => insister sur ce qui est important ou nouveau, en particulier la variable « d'intérêt »
- Traduire en français, simplifier les fractions et les arrondis (petits calculs mentaux parfois nécessaires)
- Transformer « ce que dit le tableau » en *résultats*
- Interprétation ou « ce qu'on dit du tableau » : Hypothèses ou éléments de connaissance permettant d'expliquer d'où viennent les résultats
- Travail par approximations successives : on explique d'abord les marges même si c'est évident avant de détailler la façon dont chaque profil ligne s'écarte de la moyenne
- On présente ainsi les résultats évidents ou attendus avant de présenter des résultats plus fins, surprenants voire fragiles
- Garder en tête ce qu'on sait sur l'enquête (population-cible, effectifs disponibles, etc.)

Différences entre pays

% ligne	Priorités entre carrière et vie familiale				
Pays	1:Privilégie sa carrière professionnel	2:Accorde une égale importance...	3:Privilégie sa famille	9:Ne sait pas ou ne souhaite pas répondre	Ensemble
1:Allemagne	11,4%	26,8%	61,4%	0,4%	100,0%
2:France	13,6%	10,7%	75,7%	0,1%	100,0%
3:Espagne	9,9%	16,9%	72,7%	0,5%	100,0%
4:Suisse	8,9%	24,3%	66,2%	0,6%	100,0%
5:Pologne	9,3%	23,3%	62,1%	5,3%	100,0%
6:Belgique	10,0%	21,0%	65,7%	3,4%	100,0%
Total Résultat	10,5%	20,5%	67,3%	1,7%	100,0%

NB : les effectifs utilisés ramènent chaque pays à 900 individus – ce qui a une incidence uniquement sur la dernière ligne (qui ne tient pas compte du poids des « grands » pays).

Genre et priorités en Allemagne (vs France)

% ligne	Priorités entre carrière et vie familiale				
Champ : Allemagne	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	9:Ne sait pas ou ne souhaite pas répondre	Ensemble
Genre					
0:homme	13,8%	27,2%	58,3%	0,7%	100,0%
1:femme	8,9%	26,5%	64,5%	0,1%	100,0%
Ensemble	11,4%	26,8%	61,4%	0,4%	100,0%

% ligne	Priorités entre carrière et vie familiale				
Champ : France (rappel)	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	9:Ne sait pas ou ne souhaite pas répondre	Ensemble
Genre					
0:homme	14,5%	11,0%	74,3%	0,2%	100,0%
1:femme	12,3%	10,5%	77,3%		100,0%
Ensemble	13,4%	10,7%	75,8%	0,1%	100,0%

Mobilité et priorités

% ligne	Priorités entre carrière et vie familiale				
Mobilité (résumée)	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	9:Ne sait pas ou ne souhaite pas répondre	Ensemble
0:Non mobile	11,6%	10,1%	78,2%	0,1%	100,0%
1:Mobile	23,9%	14,6%	61,5%		100,0%
Ensemble	13,4%	10,7%	75,8%	0,1%	100,0%

Âge et priorités

% ligne	Priorités entre carrière et vie familiale				
Âge décennal	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	9:Ne sait pas ou ne souhaite pas répondre	Ensemble
1:25-34	17,2%	9,4%	73,5%		100,0%
2:35-44	11,1%	11,3%	77,4%	0,2%	100,0%
3:45-54	12,2%	11,4%	76,4%		100,0%
Ensemble	13,4%	10,7%	75,8%	0,1%	100,0%

Effets combinés du genre et de la mobilité

% ligne		Priorités entre carrière et vie familiale			
France Genre	Mobilité (résumée)	1:Privilégie sa car rière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemb le
0:homme	<i>Non mobile</i>	12,8%	9,8%	77,4%	100,0%
	<i>Mobile</i>	21,9%	16,3%	61,8%	100,0%
	Ensemble	14,5%	11,0%	74,4%	100,0%
1:femme	<i>Non mobile</i>	10,6%	10,3%	79,1%	100,0%
	<i>Mobile</i>	27,5%	11,5%	60,9%	100,0%
	Ensemble	12,3%	10,5%	77,3%	100,0%
Ensemble	<i>Non mobile</i>	11,6%	10,1%	78,3%	100,0%
	<i>Mobile</i>	23,9%	14,6%	61,5%	100,0%
	Ensemble	13,4%	10,7%	75,9%	100,0%

Écart maximal et % de l'écart maximal

Effectifs / écarts maximaux	Priorités entre carrière et vie familiale			
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	120,0	97,0	225,0	442
1:femme	0,0	0,0	456,0	456
Ensemble	120	97	681	898
Écarts / PEM	Priorités entre carrière et vie familiale			
genre	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
0:homme	60,9	49,3	-110,2	0
1:femme	-60,9	-49,3	110,2	0
Ensemble	0	0	0	0

$$\text{PEM} = 12,4 / 220,4 = 5,6 \%$$

Des écarts pondérés au Khi²

On cherche à présent à construire un indicateur mesurant l'information apportée par une case du tableau toujours par rapport à la situation théorique d'indépendance

- On s'attend à ce que les écarts à l'indépendance ne puissent pas être beaucoup plus grands que les effectifs attendus (théoriques) en valeur absolue.
- On construit donc un poids égal à l'écart à l'indépendance divisé par l'effectif théorique.
- On multiplie ce poids par l'écart à l'indépendance pour obtenir un écart pondéré, qui est de signe positif
- La somme de ces écarts pondérés est appelé le Khi² du tableau
- On connaît la loi que suivrait ce Khi² si les variables étaient indépendantes, compte tenu du nombre de lignes et de colonnes
- (à condition qu'aucun effectif théorique ne soit inférieur à 5 individus)

Tableau permettant le calcul du Khi²

écarts pondérés	Priorités entre carrière et vie familiale			
	1:Privilégie sa carrière professionnelle	2:Accorde une égale importance...	3:Privilégie sa famille	Ensemble
genre				
0:homme	0,41	0,03	0,11	0,56
1:femme	0,40	0,03	0,11	0,54
Ensemble	0,81	0,07	0,23	1,10

Khi²=1,1

Degré de liberté = 2

prob(Khi²>1,1)=0,576244463474145

Suite (la prochaine fois)

- χ^2 et écart maximal
- But de l'écart maximal : différencier les cas où les écarts à l'indépendance sont importants de ceux où ils sont peu marqués ou quasi-nuls, avoir une idée des situations « mal spécifiées »
- But du test du χ^2 : avoir un critère pour décider si il est raisonnable d'inférer les résultats obtenus sur un échantillon aléatoire à la population entière