TP Analyse bivariée
Croisement Qualitatif-Quantitatif
Corrigé

Exercice 1 — *Décomposition de la variance*

Dans une population Ω de taille n, on observe deux variables :

- une qualitative, $x=\{x_k\}_{k=1,\dots,n}$, à p modalités notées, $m_1,\dots,m_p$
- une quantitative continue $y=\{y_k\}_{k=1,\dots,n}$ de moyenne $\bar{y}$ et de variance s_y^2 .

On suppose que les modalités de la série x définissent des sous-populations

$$\Omega = \Omega_1 \cup \dots \cup \Omega_p \quad \text{où } \Omega_i \cap \Omega_j = \varnothing ,$$

de tailles respectives $n_1,\dots,n_p$.

On peut alors considérer les restrictions de la caractéristiques y sur chacune des sous-populations et calculer les indicateurs numériques usuels pour chaque modalités de x,

- moyennes : $\overline{y_i}$, i=1,…,p
- variances : s_i^2, i=1,…,p

Montrer que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{p} n_i \overline{y_i}$$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{p} n_i (\overline{y_i} - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{p} n_i s_i^2 = s_E^2 + s_R^2$$

A quoi correspondent les termes s_E^2 et s_R^2 ?

On définit un indice de liaison entre les deux caractéristiques x et y par le rapport de corrélation

$$s_{y/x} = \sqrt{\frac{s_E^2}{s_y^2}}$$

Donner un encadrement de $S_{y/x}$. A quoi correspondent les cas $S_{y/x}=0$ et $S_{y/x}=1$?

Exercice 2 — *Données : SalairesData.xls*

Le fichier présente les salariés d'une entreprise ayant 3 sites (A, B et C). On y indique leur sexe, leur salaire annuel, leur catégorie (CS : cadre supérieur, CM : cadre moyen, OE : ouvrier employé), leur âge et leur site.

1) Construire un tableau répertoriant le salaire des individus par catégorie, puis par site.

2) Sur un même graphique, représenter les boites de Tuckey pour chaque catégorie. Commenter

3) Faire le même graphique mais par site. Sur quel site vaut-il mieux travailler à votre avis ?

4) Pouvez-vous justifier votre réponse à l'aide d'un indicateur numérique ?

Exercice 3 *Données : EnsSuperieurData.xls*

Le fichier EnsSuperieur.xls comptabilise le nombre d'étudiants par sexe dans l'enseignement supérieur de premier et deuxième cycles. Il s'agit chiffres relevés par Eurostat en 2008.

Illustrer et commenter ces chiffres en travaillant dans un premier temps sur le nombre d'étudiants et ensuite sur le taux d'étudiants pour 1000 habitants.

Exercice 1

Notons $y_{k,l}$,ki=1,…,n_l, la k^{ème} observation de la caractéristique Y appartenant à la sous-population Ω_l .

- Décomposition de la moyenne

$$\overline{y_l} = \frac{1}{n_l}\sum_{k=1}^{n_l} y_{k,l} \Rightarrow n_l\overline{y_l} = \sum_{k=1}^{n_l} y_{k,l} \Rightarrow \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l} = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l} = \frac{1}{n}\sum_{k=1}^{n} y_k = \overline{y}$$

- Décomposition de la variance (Calculs très longs ne pas laisser chercher)

$$s_E^2 = \frac{1}{n}\sum_{l=1}^{p} n_l(\overline{y_l} - \overline{y})^2 = \frac{1}{n}\sum_{l=1}^{p} n_l(\overline{y_l}^2 - 2\overline{y_l}\overline{y} + \overline{y}^2) = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \frac{2}{n}\overline{y}\underbrace{\sum_{l=1}^{p} n_l\overline{y_l}}_{=\overline{y}} + \frac{1}{n}\overline{y}^2\underbrace{\sum_{l=1}^{p} n_l}_{=n}$$

$$= \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - 2\overline{y}^2 + \overline{y}^2 = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \overline{y}^2$$

Nous avons

$$s_l^2 = \frac{1}{n_l}\sum_{k=1}^{n_l}(y_{k,l} - \overline{y_l})^2 ,$$

d'où

$$s_R^2 = \frac{1}{n}\sum_{l=1}^{p} n_l s_l^2 = \frac{1}{n}\sum_{l=1}^{p} n_l \frac{1}{n_l}\sum_{k=1}^{n_l}(y_{k,l} - \overline{y_l})^2 = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l}(y_{k,l}^2 - 2y_{k,l}\overline{y_l} + \overline{y_l}^2)$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{2}{n}\sum_{l=1}^{p}\overline{y_l}\underbrace{\sum_{k=1}^{n_l} y_{k,l}}_{=n_l\overline{y_l}} + \frac{1}{n}\sum_{l=1}^{p}\overline{y_l}^2\underbrace{\sum_{k=1}^{n_l} 1}_{=n_l}$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{2}{n}\sum_{l=1}^{p}\overline{y_l}n_l\overline{y_l} + \frac{1}{n}\sum_{l=1}^{p}\overline{y_l}^2 n_l = \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2$$

Donc

$$s_E^2 + s_R^2 = \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2 - \overline{y}^2 + \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \frac{1}{n}\sum_{l=1}^{p} n_l\overline{y_l}^2$$

$$= \frac{1}{n}\sum_{l=1}^{p}\sum_{k=1}^{n_l} y_{k,l}^2 - \overline{y}^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - \overline{y}^2 = s_y^2$$

Le premier terme correspond à la *variance expliquée* par la partition de la série observée X et le deuxième terme est un reste appelé *variance résiduelle*.

- Rapport de corrélation

$$s_y^2 = s_E^2 + s_R^2 \Rightarrow 0 \le s_E^2 \le s_y^2 \Rightarrow 0 \le \frac{s_E^2}{s_y^2} \le 1 \Rightarrow 0 \le s_{y/x} \le 1$$

Cela signifie que s_E^2 représente le pourcentage de variabilité de y expliquée par x.

- Si $s_{y/x}=0$ alors la variance expliquée est nulle, il n'y a donc aucun lien entre y et x
- Si $s_{y/x}=1$ alors la variance expliquée est égale à la variance de y donc y est entièrement expliquée par x.

Attenttion : $s_{y/x} \neq s_{x/y}$. Cela signifie que si $s_{y/x}$ est proche de 1 alors x permet d'expliquer y mais pas réciproquement.

Salaires_Correction.xls [Mode de compatibilité] - Microsoft Excel

Regroupement du salaire par catégorie

A	B	C
140	117	118
120	114	117
115	56	115
95	55	113
60	43	58
50	42	50
49	32	45
49	30	35
38	27	27
28	27	26
27	26	25
26	26	25
26	26	25
25	24	24
25	24	24
25	24	24
25	24	24
24	23	24
24	23	23
24	22	23
24	21	22
24	21	22
24	21	22
24	21	22
23	21	21
23	21	18

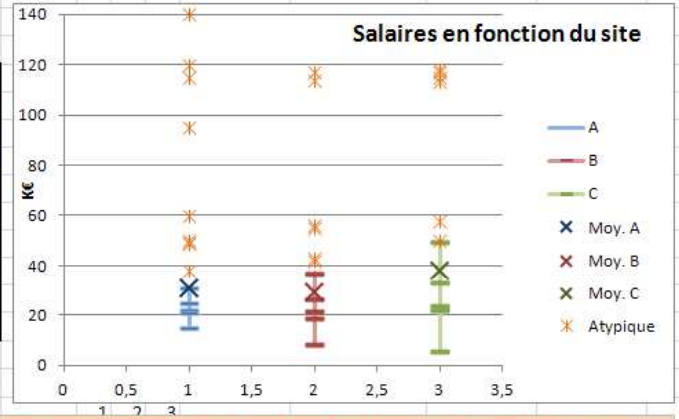
Récapitulatif

	A	B	C	Total
Effectif	60	40	30	130
Moyenne	30,80	29,33	38,13	32,04
Variance	618,03	472,07	1010,38	675,21
Médiane	22	21,5	24	
Q1	21	19	22	
Q3	25	26,25	33	
m	15	8,13	5,5	
M	31	37,13	49,5	

Décomposition de la variance

variance intra	663,66
variance inter	11,55
somme inter+intra	675,21
Rapport	0,02

!!! Attention !!!
la variance est sans biais, il faut * par (n-1)/n pour retrouver la variance biaisée de la décomposition de la variance



Les indicateurs de tendance (moyenne et médiane) sont du même ordre de grandeur sur les trois sites.
On note que le salaire médian est à une vingtaine de K€ alors que le salaire moyen est à une trentaine de K€. La distribution des salaires est donc asymétrique. Les petits salaires sont regroupés proches de la médiane tandis que le salaire moyen est attiré par les très hauts salaires. Cette observations est confirmée avec les nombreux hauts salaires atypiques .

Le rapport de dispersion confirme qu'il y a peu de différences entre les sites puisque uniquement 2% de la variabilité des salaires est expliquée par le site.

Salaires_Correction.xls [Mode de compatibilité] - Microsoft Excel

Regroupement du salaire par catégorie

CS	CM	OE
140	60	26
120	58	26
118	56	26
117	55	26
117	50	26
115	50	26
115	49	25
114	49	25
113	45	25
95	43	25
	42	25
	38	25
	35	25
	32	24
	30	24
	28	24
	27	24
	27	24
	27	24
	27	24
		24
		24
		24
		24
		24

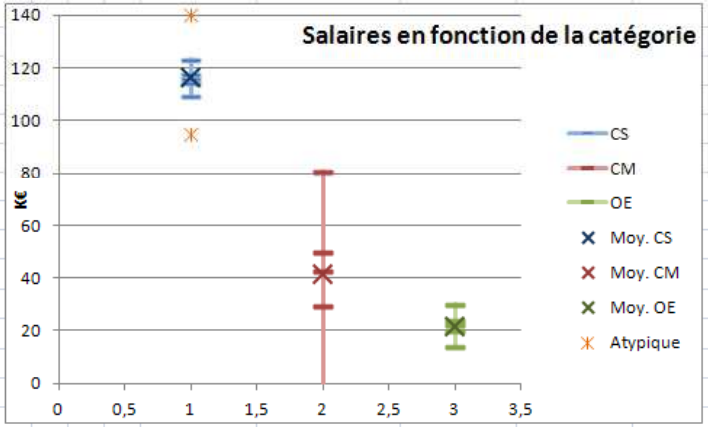
Récapitulatif

	CS	CM	OE	Total
Effectif	10	20	100	130
Moyenne	116,40	41,40	21,73	32,04
Variance	105,24	128,94	5,98	675,21
Médiane	116	42,5	22	
Q1	114,25	29,5	20	
Q3	117,75	50	24	
m	109	-1,25	14	
M	123	80,75	30	

Décomposition de la variance

variance intra	32,53
variance inter	642,68
somme inter+intra	675,21
Rapport	0,95

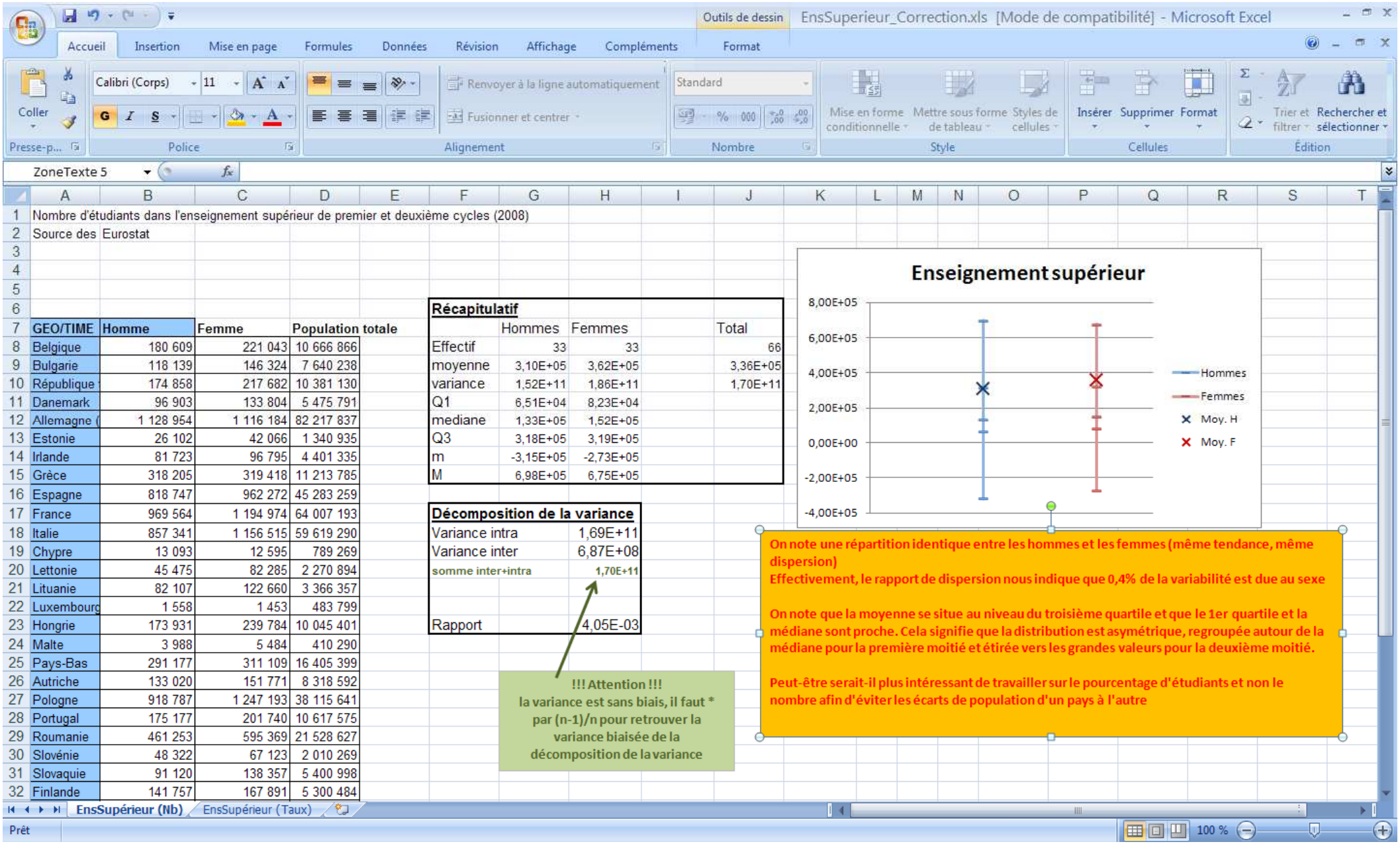
!!! Attention !!!
la variance est sans biais, il faut * par (n-1)/n pour retrouver la variance biaisée de la décomposition de la variance



Les indicateurs de tendance (moyenne et médiane) augmente naturellement avec la catégorie. Les deux indicateurs sont à peu près égaux, ce qui implique une distribution symétrique des salaires de part et d'autre de la tendance centrale.

On note très peu de dispersion dans les catégories extrêmes. Ce qui signifie qu'il n'y a pas beaucoup de différences de salaires dans ces catégories. On note toutefois deux éléments atypiques dans la catégorie CS. A contrario la dispersion est très grande dans la catégorie intermédiaire qui implique certainement des profils très différents.

Le rapport de dispersion confirme que 95% de la variabilité des salaires est expliquée par la catégorie.



Nombre d'étudiants dans l'enseignement supérieur de premier et deuxième cycles (2008)
Source des Eurostat

GEO/TIME	Homme	Femme	Population totale
Belgique	180 609	221 043	10 666 866
Bulgarie	118 139	146 324	7 640 238
République	174 858	217 682	10 381 130
Danemark	96 903	133 804	5 475 791
Allemagne (	1 128 954	1 116 184	82 217 837
Estonie	26 102	42 066	1 340 935
Irlande	81 723	96 795	4 401 335
Grèce	318 205	319 418	11 213 785
Espagne	818 747	962 272	45 283 259
France	969 564	1 194 974	64 007 193
Italie	857 341	1 156 515	59 619 290
Chypre	13 093	12 595	789 269
Lettonie	45 475	82 285	2 270 894
Lituanie	82 107	122 660	3 366 357
Luxembourg	1 558	1 453	483 799
Hongrie	173 931	239 784	10 045 401
Malte	3 988	5 484	410 290
Pays-Bas	291 177	311 109	16 405 399
Autriche	133 020	151 771	8 318 592
Pologne	918 787	1 247 193	38 115 641
Portugal	175 177	201 740	10 617 575
Roumanie	461 253	595 369	21 528 627
Slovénie	48 322	67 123	2 010 269
Slovaquie	91 120	138 357	5 400 998
Finlande	141 757	167 891	5 300 484

Récapitulatif

	Hommes	Femmes	Total
Effectif	33	33	66
moyenne	3,10E+05	3,62E+05	3,36E+05
variance	1,52E+11	1,86E+11	1,70E+11
Q1	6,51E+04	8,23E+04	
mediane	1,33E+05	1,52E+05	
Q3	3,18E+05	3,19E+05	
m	-3,15E+05	-2,73E+05	
M	6,98E+05	6,75E+05	

Décomposition de la variance

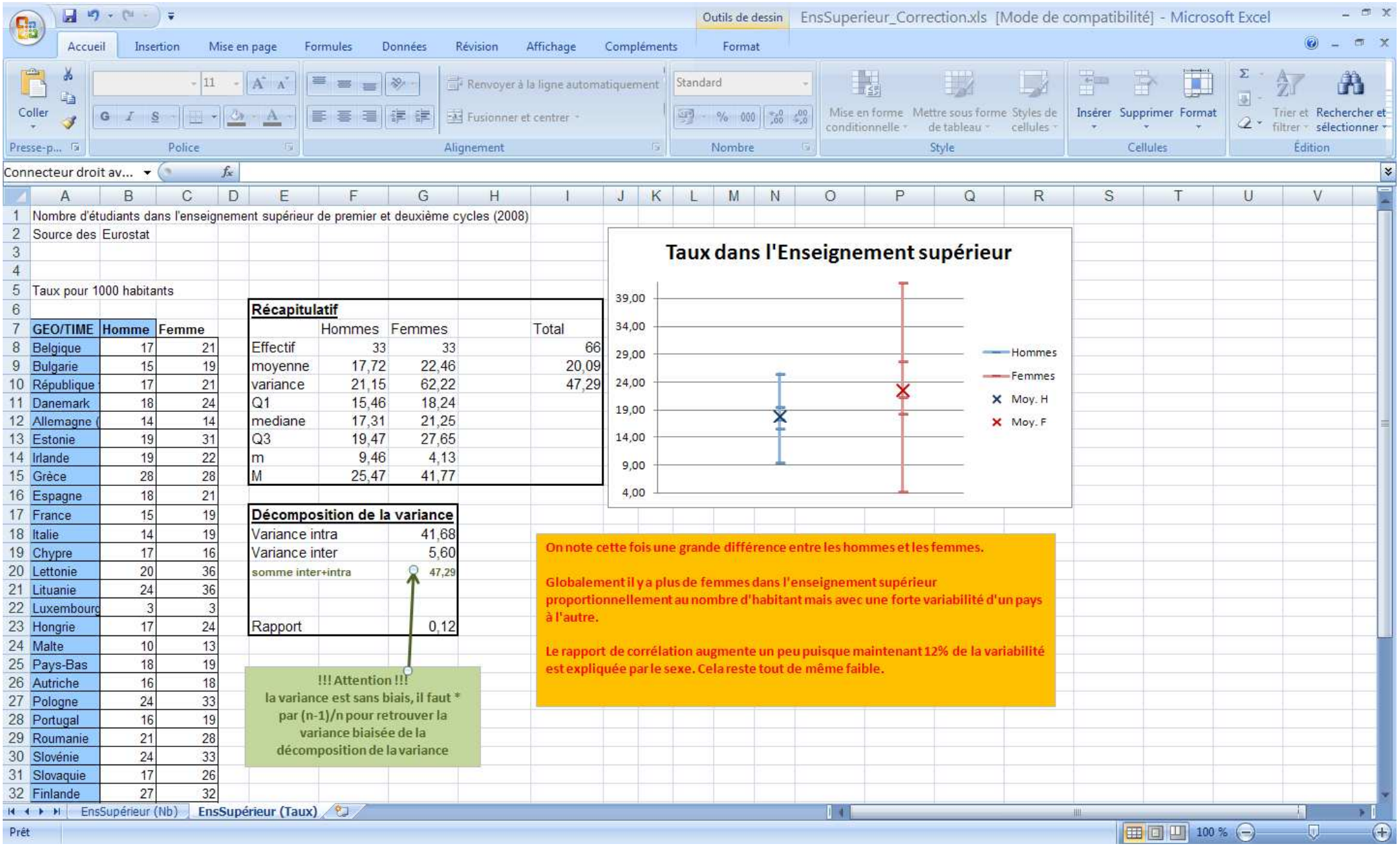
Variance intra	1,69E+11
Variance inter	6,87E+08
somme inter+intra	1,70E+11
Rapport	4,05E-03

!!! Attention !!!
la variance est sans biais, il faut * par (n-1)/n pour retrouver la variance biaisée de la décomposition de la variance

On note une répartition identique entre les hommes et les femmes (même tendance, même dispersion)
Effectivement, le rapport de dispersion nous indique que 0,4% de la variabilité est due au sexe

On note que la moyenne se situe au niveau du troisième quartile et que le 1er quartile et la médiane sont proche. Cela signifie que la distribution est asymétrique, regroupée autour de la médiane pour la première moitié et étirée vers les grandes valeurs pour la deuxième moitié.

Peut-être serait-il plus intéressant de travailler sur le pourcentage d'étudiants et non le nombre afin d'éviter les écarts de population d'un pays à l'autre


EnsSuperieur_Correction.xls [Mode de compatibilité] - Microsoft Excel
Outils de dessin
Accueil
Insertion
Mise en page
Formules
Données
Révision
Affichage
Compléments
Format
Connecteur droit av...
Nombre d'étudiants dans l'enseignement supérieur de premier et deuxième cycles (2008)
Source des Eurostat
Taux pour 1000 habitants
GEO/TIME | Homme | Femme
Belgique 17 21
Bulgarie 15 19
République 17 21
Danemark 18 24
Allemagne (14 14
Estonie 19 31
Irlande 19 22
Grèce 28 28
Espagne 18 21
France 15 19
Italie 14 19
Chypre 17 16
Lettonie 20 36
Lituanie 24 36
Luxembourg 3 3
Hongrie 17 24
Malte 10 13
Pays-Bas 18 19
Autriche 16 18
Pologne 24 33
Portugal 16 19
Roumanie 21 28
Slovénie 24 33
Slovaquie 17 26
Finlande 27 32
Récapitulatif
Hommes Femmes Total
Effectif 33 33 66
moyenne 17,72 22,46 20,09
variance 21,15 62,22 47,29
Q1 15,46 18,24
mediane 17,31 21,25
Q3 19,47 27,65
m 9,46 4,13
M 25,47 41,77
Décomposition de la variance
Variance intra 41,68
Variance inter 5,60
somme inter+intra 47,29
Rapport 0,12
!!! Attention !!!
la variance est sans biais, il faut * par (n-1)/n pour retrouver la variance biaisée de la décomposition de la variance
Taux dans l'Enseignement supérieur
39,00
34,00
29,00
24,00
19,00
14,00
9,00
4,00
Hommes
Femmes
Moy. H
Moy. F
On note cette fois une grande différence entre les hommes et les femmes.
Globalement il y a plus de femmes dans l'enseignement supérieur proportionnellement au nombre d'habitant mais avec une forte variabilité d'un pays à l'autre.
Le rapport de corrélation augmente un peu puisque maintenant 12% de la variabilité est expliquée par le sexe. Cela reste tout de même faible.
EnsSupérieur (Nb)
EnsSupérieur (Taux)
Prêt
100 %