Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

Cours 11 – Une variable numérique : dispersion et variance

Ce cours est consacré à la variance et à l'écart-type ; on commence par faire un rappel sur la variance comme indice de dispersion, mesure de l'éparpillement des observations ; puis on étudie la manière dont elle se décompose à l'aide des variances conditionnelles, dispersions propres aux sous-populations. Dans ce cours nous supposerons encore que la variable numérique est Y, les sous-populations étant induites par X.

1 déf **Intervalle inter-quartile** : c'est l'intervalle des valeurs situées entre les premier et troisième quartiles, $q_{25\%}$ et $q_{75\%}$; par définition, la moitié des observations « centrales » se trouvent dans cet intervalle, un quart étant à sa gauche et un autre quart à sa droite. Son amplitude s'appelle `l'écart inter-quartile`. C'est une mesure de la dispersion qui s'exprime dans la même unité que Y : plus il est petit, plus les valeurs « centrales » sont ramassées.

Variance et écart-type

2 déf **La variance d'une série E de n observations y_i de Y** est la dispersion quadratique moyenne de E autour de la moyenne :

$$(1)\quad var(y) = \frac{\text{disp}_q(\overline{y})}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

Puisque la moyenne est la valeur la plus proche de E pour la distance quadratique, on pourrait définir la variance comme la plus petite distance quadratique moyenne d'une valeur à E.

3 **L'écart-type d'une série E de n observations y_i de Y** est la racine carrée de la variance ; on le note $\sigma(y)$ (on prononce « sigma »), ce qui permet de noter la variance $\sigma^2(y)$ (ou plus simplement σ^2) ; l'écart-type quantifie la dispersion des observations dans la même unité que Y.

4 **Formules de la variance.** De manière synthétique la définition (1) s'énonce « variance = moyenne des carrés des écarts à la moyenne » ; on démontre qu'elle est équivalente à une seconde formule : « variance = moyenne des carrés moins carré de la moyenne » :

$$(2)\quad \sigma^2(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$$

Cette seconde formule, souvent plus pratique, doit être utilisée avec précaution car elle est sensible aux erreurs d'arrondis.

5 **Calcul de la variance.** Si on dispose des observations y_i, on utilise l'une ou l'autre des formules (1) ou (2).

Exemple des données cliniques du groupe 1 :

Groupe 1	65	72	85	93	104	110
Groupe 2	45	51	58	67	79	
Groupe 3	93	98	116	121	123	

La moyenne $\overline{y_{G1}}$ est égale à $529/6$; en arrondissant au centième, les formules (1) et (2) donnent :

$\frac{(65-88{,}17)^2+...+(110-88{,}17)^2}{6} = 259{,}81$ et $\frac{65^2+...+110^2}{6} - 88{,}17^2 = 259{,}22$.

Le tableau suivant donne les résultats des deux formules pour différents arrondis de la moyenne ; on notera qu'il faut 3 décimales pour que les deux formules donnent sensiblement la même valeur :

							Moyenne	Variance
y_i	65	72	85	93	104	110	88,1667	
y_i^2	4225	5184	7225	8649	10816	12100	8033,1667	259,7997
Moyenne							Variance (1)	Variance (2)
88,2	538,24	262,44	10,24	23,04	249,64	475,24	259,807	253,927
88,17	536,85	261,47	10,05	23,33	250,59	476,55	259,807	259,218
88,167	536,71	261,372	10,03	23,358	250,684	476,68	259,806	259,747
88,1667	536,696	261,3622	10,028	23,3608	250,6934	476,693	259,806	259,8

6 Distribution discrète. A partir de la distribution d'une variable discrète, la variance se calcule par une formule équivalente qui consiste à regrouper dans un même terme les n_j valeurs égales à m'_j, comme pour la moyenne :

(1d) $\sigma^2(y) = \frac{1}{n}\sum_{j=1}^{p} n_j * (m'_j - \overline{y})^2$ et (2d) $\sigma^2(y) = \frac{1}{n}\sum_{j=1}^{p} n_j * m'^2_j - \overline{y}^2$

On notera que dans ces formules, la variable muette j parcourt les p modalités de Y et non les n observations.

Exemple du nombre d'enfants de 0 à 6 ans dans les couples (exemple 1 du Cours 9) :

Nbre enfants [0-6]	1	2	3	4	5	Total	Moy / Var(1d)
Couples	3269	2174	963	120	10	6536	
$n_j * m'_j$	3269	4348	2889	480	50	11036	1,688
$n_j * (m'_j - \overline{y}_{couples})^2$	1547,36	211,63	1657,65	641,44	109,69	4167,77	0,638
m'^2_j	1	4	9	16	25	Moyenne	Variance (2d)
$n_j * m'^2_j$	3269	8696	8667	1920	250	3,489	0,639

7 Distribution continue. A partir de la distribution d'une variable continue, la variance ne peut qu'être approximée, puisqu'on ne connaît pas les valeurs ; en supposant qu'elles sont uniformément réparties dans chaque modalité $[b_j; b_{j+1}[$, on place virtuellement les n_j valeurs au centre $c_j = \frac{b_j+b_{j+1}}{2}$ qui se confond avec leur moyenne, et on obtient les approximations suivantes :

(1c) $\sigma^2 \approx \frac{1}{n}\sum_{j=1}^{p} n_j * (c_j - \overline{y})^2$ et (2c) $\sigma^2 \approx \frac{1}{n}\sum_{j=1}^{p} n_j * c_j^2 - \overline{y}^2$

Exemple du revenu dans le sud (exemple 2 du Cours 9) :

Modalités	]0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 30[	Total	Moyenne	Variance
Centres c_j	2,5	7,5	12,5	22,5			
c_j^2	6,25	56,25	156,25	506,25			
Eff. Y_{sud}	28	42	30	24	124	10,484	
$n_j * (c_j - \overline{y}_{sud})^2$	1784,84	373,98	121,93	3465,22	5745,97	46,34	46,34
$n_j * c_j^2$	175	2362,5	4687,5	12150	19375	156,25	46,34

8 ⚠ Erreur de l'approximation. La moyenne et la variance sont évaluées en faisant la même hypothèse, une répartition uniforme des observations dans la modalité-intervalle ; mais cette hypothèse conduit à des erreurs plus importantes pour le calcul de la variance, comme nous allons le voir sur l'exemple suivant.

Considérons la distribution continue :

Y	[0 ; 3[	[3 ; 6[
Effectif	2	2

La moyenne est approximée par la formule $\frac{1}{4}(2 * \frac{0+3}{2} + 2 * \frac{3+6}{2}) = 3$ et la variance par la formule $\frac{1}{4}(2 * (\frac{0+3}{2} - 3)^2 + 2 * (\frac{3+6}{2} - 3)^2) = 2{,}25$.

Si l'hypothèse d'uniforme répartition est fausse, en supposant par exemple que les observations effectives sont 1 1,5 5 et 5,5, de moyenne 3,25 et de variance 4,06, les approximations de la moyenne et de la variance donnent des valeurs incorrectes, ce qui n'est pas étonnant.

Si par contre l'hypothèse d'uniforme répartition est juste, en supposant par exemple que les observations effectives sont 1 2 4 et 5, de moyenne 3 et de variance 2,5, l'approximation de la moyenne donne comme prévu une valeur correcte, alors que celle de la variance donne encore une valeur incorrecte. Cela est dû au fait que l'approximation néglige la dispersion réelle des observations à l'intérieur de chaque modalité, en les regroupant toutes au centre.

La conclusion est que, dans le cas d'une variable continue, la variance doit autant que possible se calculer sur les observations elles-mêmes.

9 Propriétés.

1. $\sigma(y+a) = \sigma(y)$ et $\sigma^2(y+a) = \sigma^2(y)$: on ne modifie pas la dispersion d'une série d'observations en décalant à gauche ou à droite d'une même valeur a toutes les observations.
2. $\sigma(a*y) = a*\sigma(y)$ et $\sigma^2(a*y) = a^2*\sigma^2(y)$: si on multiplie chaque observation par une même valeur a, l'écart-type est également multiplié par a, alors que la variance est multipliée par son carré a^2.
3. Réduction : la variable $\frac{Y}{\sigma(y)} = \frac{1}{\sigma(y)} * Y$, obtenue en divisant toutes les observations y_i par $\sigma(y)$, est de variance égale à 1, d'après la propriété précédente ; on dit qu'elle est `réduite`.
4. Centrage et réduction : La variable $\frac{Y-\overline{y}}{\sigma(y)}$ est de moyenne nulle et de variance égale à 1 ; on dit qu'elle est `centrée et réduite`.
5. Intervalle de dispersion : l'intervalle de dispersion centré autour de la moyenne $ID(\overline{y}) = [\overline{y}-2*\sigma(y), \overline{y}+2*\sigma(y)]$ contient approximativement 95% des observations si la distribution est symétrique et unimodale.

Variance et sous-populations

10 La variabilité de Y qui se manifeste dans la dispersion des observations a deux origines possibles : la *variabilité intrinsèque* de Y due au fait que sa mesure sur un individu comporte une part contingente, et *l'hétérogénéité des sous-populations* pour ce caractère.

La première origine s'exprime par la dispersion de Y dans chaque sous-population, mesurée par les variances conditionnelles, la seconde par les différences entre les moyennes conditionnelles.

11 déf **Variances conditionnelles.** Les variances conditionnelles de la variable numérique Y sont les variances des distributions conditionnelles de Y. On note σ_i l'écart-type de Y dans la sous-population induite par la ième modalité m_i de X, et σ_i^2 la variance.

Si l'on dispose des données brutes y_j, la variance conditionnelle σ_i^2 est alors la dispersion quadratique des $n_{i\bullet}$ observations de la sous-population autour de la moyenne $\overline{y_i}$:
$\sigma_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} (y_j - \overline{y_i})^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{n_{i\bullet}} y_j^2 - (\overline{y_i})^2$.

Elle mesure la dispersion *interne* à la sous-population, sans tenir compte des autres sous-populations, et par conséquent l'effet de la variabilité intrinsèque de Y concernant cette sous-population.

12 déf **Décomposition de la variance.** La double origine de la variabilité de Y se formalise dans la formule de décomposition de la variance ; en effet on peut démontrer que la variance totale des observations de l'échantillon est la somme de deux quantités :

la variance intra : la moyenne des variances conditionnelles pondérée par la taille des sous-populations ; elle quantifie la part de la variabilité intrinsèque de Y dans la variance totale ;

la variance inter : la variance des moyennes conditionnelles également pondérée par la taille des sous-populations ; elle quantifie l'hétérogénéité des sous-populations.

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} \sigma_i^2 + \frac{1}{n} \sum_{i=1}^{k} n_{i\bullet} (\overline{y_i} - \overline{y})^2$$

En termes synthétiques la décomposition de la variance s'énonce « variance totale = variance intra + variance inter », ou encore « variance totale = moyenne des variances + variance des moyennes ».

13 Calcul. On calcule d'abord pour chaque sous-population la somme des valeurs et la somme des carrés des valeurs.

Avec les données cliniques, cela donne :

							Somme
Groupe 1	65	72	85	93	104	110	529
y_i^2	4225	5184	7225	8649	10816	12100	48199
Groupe 2	45	51	58	67	79		300
y_i^2	2025	2601	3364	4489	6241		18720
Groupe 3	93	98	116	121	123		551
y_i^2	8649	9604	13456	14641	15129		61479

On rassemble ensuite ces résultats dans un tableau (les valeurs en italique), et on le complète avec une précision suffisante pour ne pas introduire d'erreur importante dans les calculs de carrés (3 décimales sont généralement suffisantes) :

1. On calcule $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; dans l'exemple $\overline{y} = \frac{1}{16}1380 = 86{,}25$.
2. On calcule les moyennes $\overline{y_i}$ et les variances σ_i^2 conditionnelles ($88{,}167 = \frac{529}{6}$, $259{,}747 = \frac{48199}{6} - 88{,}167^2$).
3. On calcule la variance inter par la formule $\frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\overline{y_i}^2 - \overline{y}^2$ (var inter $= 7835{,}04 - 86{,}25^2 = 395{,}98$).
4. On calcule enfin la variance intra par la formule $\frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}\sigma_i^2$ (var intra $= 3037{,}28/16 = 189{,}83$).

Arrondi 3

	Effectif	$\sum y_i$	$\sum y_i^2$	$\overline{y_i}$	σ_i^2	$n_{i\bullet}\overline{y_i}^2$	$n_{i\bullet}\sigma_i^2$	
Groupe 1	*6*	*529*	*48199*	88,167	259,747	46640,519	1558,482	
Groupe 2	*5*	*300*	*18720*	60	144	18000	720	
Groupe 3	*5*	*551*	*61479*	110,2	151,76	60720,2	758,8	
Total	**16**	1380	128398			125360,719	3037,282	
Moyenne		**86,25**	8024,875			7835,045	189,83	
		Variance	**585,813**		**Inter**	**395,983**	**189,83**	**Intra**

On peut vérifier la formule de décomposition de la variance en comparant la variance totale qui se calcule par la formule $\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2$ ($\sigma^2 = 8024{,}88 - 86{,}25^2 = 585{,}82$), avec la somme des variances inter et intra (395,98+189,83=585,81).

Programme de travail

Savoir définir :

- la variance et l'écart-type ;
- la formule de décomposition de la variance ;
- la variance intra ;
- la variance inter.

Savoir expliquer :

- la signification de la variance et de l'écart-type ;
- la signification de la variance intra ;
- la signification de la variance inter.

Savoir énoncer :

- les propriétés de la variance et de l'écart-type (§9).

Savoir faire :

- centrer et/ou réduire une variable ;
- calculer les variances globales et conditionnelles, dans tous les cas de figure ;
- calculer la variance intra ;
- calculer la variance inter ;
- vérifier la formule de décomposition de la variance.