MESURES DE VARIATION

À LA FIN DE CE CHAPITRE VOUS DEVEZ ÊTRE CAPABLE DE :

- COMPRENDRE POURQUOI ON MESURE LA VARIABILITÉ
- COMPRENDRE LES CONCEPTS DE VARIATION DE L'ÉCART-TYPE
- COMPRENDRE LE COEFFICIENT DE VARIATION
- COMPRENDRE QUAND APPLIQUER LES DIFFÉRENTES MESURES DE VARIABILITÉ

CHAPITRE 6

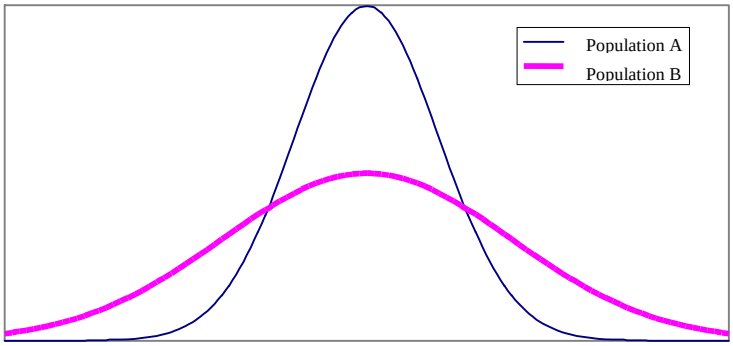
MESURES DE VARIATION

Le concept de variation

quelquefois les moyennes ne suffisent pas

Une mesure de la valeur moyenne peut apporter beaucoup d'informations utiles sur un jeu d'observations, mais dans la plupart des cas ce n'est pas suffisant pour nous dire quoi que ce soit sur la variable. Regardons par exemple la figure 6.1 ci-dessous :

Figure 6.1 Comparaison entre deux distributions



les distributions sont différentes mais donnent quand même les mêmes moyennes

Alors que les deux distributions ont les mêmes valeurs moyennes, si on mesure les moyennes, les médianes ou les dominantes on ne peut pas dire que les distributions soient les mêmes. Pour les décrire et les comparer on a besoin d'informations supplémentaires; on a besoin de moyens alternatifs pour décrire les distributions. Après la valeur moyenne, la propriété la plus importante que l'on doit mesurer est la variabilité de la distribution. À la figure 6.1 on peut voir que la distribution B est bien plus variable (ou étendue) que la distribution A. Dans ce chapitre on examinera différentes manières de mesurer la variabilité.

le niveau réel de variabilité

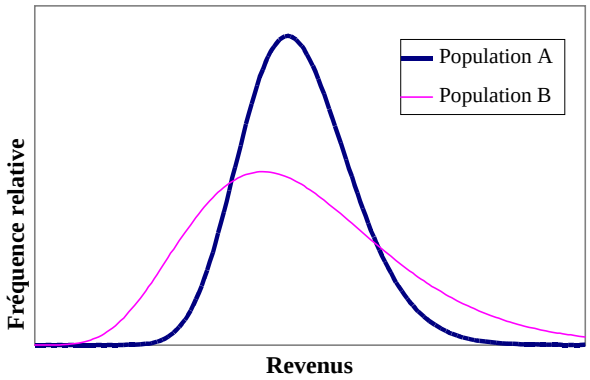
On veut mesurer la variabilité pour deux raisons majeures. D'abord on peut être intéressé par le niveau réel de variabilité et le comparer à une autre distribution. Si on examine des distributions de revenus par exemple, alors le gouvernement peut avoir un intérêt non seulement dans le niveau moyen des revenus, mais aussi dans la variabilité du niveau de revenus entre les personnes, et aussi entre des régions différentes d'un pays. Beaucoup de politiques sont élaborées pour aider à la redistribution des revenus des plus riches vers les plus pauvres (réduisant ainsi la variabilité des apports de revenus), donc on a besoin de mesurer la variabilité pour voir si elle change avec le temps.

la variabilité due à la variation de l'échantillon

La deuxième raison qui nous pousserait à mesurer la variabilité est l'utilisation de l'échantillionage. On doit alors prendre en compte la variabilité. On veut être capable de distinguer des différences qui

pourraient se produire simplement par hasard (c'est-à-dire, dans la sélection des échantillons) et ceux qui indiquent un changement réel. Par exemple :

Figure 6.2 Comparaison des niveaux de revenus de deux populations



la variabilité n'est pas nécessairement reflétée dans les moyennes

La population A représente la distribution des revenus annuels par ménage dans une région et la population B représente la même distribution dans une autre région. Elles ont toutes les deux la même moyenne de 1.800 dollars par an, mais on ne peut pas dire que les deux distributions soient les mêmes. La distribution des revenus de la population B est bien plus étendue que celle de la population A. Elle a aussi, par conséquent, un plus grand degré de variabilité.

deux mesures différentes de la variabilité

Il est clair qu'on doit non seulement comparer les mesures de position quand on examine des populations, mais aussi les mesures de variabilité. Dans ce chapitre on étudiera deux mesures de variabilité :

a. la mesure de la distance **entre deux valeurs représentatives** de la population, et

b. la mesure de la distance **entre chaque unité** de la population **et une valeur centrale**.

l'étendue et la déviation standard pour des données non groupées

On prendra comme exemple l'étendue et la déviation standard (ou variance) de données non groupées. On couvrira des techniques plus compliquées (comme de trouver la déviation standard quand les données sont groupées dans une distribution de fréquences) dans la formation avancée.

L'étendue

plus grand – plus petit

La façon la plus simple de mesurer la variation ou la répartition d'un jeu d'observations, est de calculer **l'étendue**. On définit l'étendue d'une série d'observations par la différence entre la plus petite et la plus grande des valeurs dans la série. Elle est simple à comprendre et facile à calculer, et a de ce fait des attraits évidents. On l'utilise souvent, mais elle est utile seulement quand les valeurs de la variable sont réparties assez régulièrement dans l'étendue. Elle a quelques inconvénients flagrants qui font qu'on l'utilise peu en pratique. En voici quelques uns des plus importants :

inconvénients

a. l'étendue étant la différence entre la plus grande des valeurs et la plus petite, elle est très affectée par les très grandes ou très petites observations. L'inclusion d'une seule donnée bizarre (rare ou inhabituelle) affectera grandement l'étendue.

b. L'étendue dépend du nombre d'observations. Si on augmente le nombre d'observations on ne peut qu'augmenter l'étendue; on ne peut pas la réduire. Ce qui veut dire qu'il est difficile de comparer l'étendue de deux distributions dont le nombre d'observations est différent.

c. Alors que l'étendue est très facile à calculer, elle ignore toutes les données qui se trouvent entre la valeur la plus haute et la valeur la plus basse. Si, par exemple, on examine les trois séries de données suivantes :

Série 1	3, 5, 7, 9, 11, 13, 15, 17, 17, 17, 17
Série 2	3, 5, 5, 5, 17, 17, 17, 17, 17, 17, 17
Série 3	3, 6, 7, 8, 10, 11, 14, 14, 15, 16, 17

On voit que les étendues de ces trois séries sont les mêmes (17 - 3 = 14), mais que les degrés de variation ne sont en aucun cas les mêmes;

d. Il est difficile de calculer l'étendue de données groupées en distribution de fréquences. La seule chose que l'on peut faire, c'est de prendre la différence entre la limite inférieure de la classe la plus basse et la limite supérieure de la classe la plus haute. Ça dépent évidemment des définitions des classes, et c'est impossible à calculer avec les classes ouvertes. On y arrive quand même avec un peu de jugeotte si on connait bien le sujet observé. Pour des raisons pratiques on ferme généralement les classes ouvertes avec une valeur fixe.

exemple

Prenons un autre exemple, les valeurs des importations dans quelques Îles du Pacifique en 1995 :

Tableau 6.1 Total des importations par pays, 1995 (en milliers de AUD)

Pays	Valeur des importations (milliers de AUD)
Îles Cook	65.363
Îles Fidji	1.172.052
Kiribati	47.547
Îles Marshall	100.073
Papouasie-Nlle-Guinée	1.741.935
Samoa	126.689
Îles Salomon	224.254
Tonga	98.047
Tuvalu	12.535
Vanuatu	124.521

Source : SPESS 14, 1998, Communauté du Pacifique, Nouméa

méthode

L'étendue des valeurs importées est la différence entre la plus grande et la plus petite des valeurs, dans ce cas :

Étendue = $ (1.741.935.000 – 12.535.000) = $1.729 million

en général on ne calcule pas l'étendue pour les données groupées

En général on ne calcule pas l'étendue d'une distribution de fréquences groupées à cause des inconvénients mentionnés plus haut. On peut néanmoins obtenir une approximationment en prenant la différence entre la limite supérieure de la dernière classe et la limite inférieure de la dernière classe. On doit noter qu'il peut être parfois très difficile, ou même n'avoir aucun sens de la calculer, si l'une de ces classes, ou les deux, sont ouvertes. Prenons encore une fois l'exemple du revenu monétaire annuel de deux régions d'un pays :

Tableau 6.2 Comparaison d'étendues de revenus monétaires pour deux régions

Revenu monétaire annuel des ménages			
Région A		Région B	
Revenus ($)	Fréquence (Nombre de ménages)	Revenus ($)	Fréquence (Nombre de ménages)
Moins que 500 (200*)	137	Moins que 1.000 (500*)	86
500–999	278	1.000–1.999	137
1.000–1.499	406	2.000–2.999	64
1.500–1.999	331	3.000–3.999	47
2.000–4.999	188	4.000–6.999	130
5.000–9.999	259	7.000–9.999	62
10.000–19.999	138	10.000 & plus (20.000*)	88
20.000 & plus (30.000*)	14		
Total	1.751	Total	614

Source : Tableau 5.1 (données fictives) * = Limites supposées

intervalles de classe ouverts

Bien sûr, on ne peut pas calculer l'étendue des revenus ici à cause des intervalles de classes ouverts à chaque bout. Mais, si on doit vraiment calculer l'étendue des revenus pour les deux populations on devra faire des suppositions. Ces suppositions pourront être mal ou bien-fondées, mais si on doit prendre une décision on devra mettre des valeurs dans les classes ouvertes. Dans l'exemple ci-dessus les valeurs supposées sont :

Valeurs supposées ($)	*Région*	
	'A'	'B'
Limite inférieure (première classe)	200	500
Limite supérieure (dernière classe)	30.000	20.000

étendue

L'étendue des revenus peut être maintenant calculée comme suit :

Région 'A'	*Région 'B'*
Étendue des revenus = $(30.000 - 200) = $29.800	Étendue des revenus = $(20.000 - 500) = $19.500

indiquer clairement les suppositions

À l'exemple ci-dessus, bien qu'on ait dérivé l'étendue des revenus à 29.800 dollars pour la région A et 19.500 dollars pour la région B, l'étendue ne veut rien dire si on se rend compte ensuite que les valeurs supposées sont incorrectes. Les statisticiens et les planificateurs sont souvent confrontés à des problèmes de ce genre dans leur travail de tous les jours, et ils sont souvent obligés de prendre des décisions comme choisir l'étendue des revenus pour les régions A et B. Le plus important c'est que les suppositions faites pour générer un résultat soient clairement indiquées.

utiliser d'autres points que le plus grand et le plus petit

On peut contourner la plupart des problèmes que pose l'étendue comme mesure de variation en utilisant d'autres points dans la distribution au lieu des deux extrèmes. On peut aussi mesurer ce qu'on appelle le semi-interquartile, ou la déviation quartile (c'est-à-dire qu'on mesure la différence moyenne entre les quartiles supérieurs et inférieurs). Pour de plus amples informations, voir le chapitre 5 "Plus d'infos sur les mesures de position". La déviation quartile ne fait pas partie de ces notes, mais elle est expliquée en détails dans le cours avancé.

utiliser les percentiles

On peut aussi utiliser la différence entre, disons le **10ème** et le **90ème percentile** (c'est-à-dire les valeurs qui correspondent à 10% et 90% des valeurs observées). Ces deux valeurs sont très utiles comme mesures de variation. Elles ne sont pas affectées par l'une ou l'autre des valeurs extrèmes, ou les valeurs aberrantes, elles dépendent moins du nombre d'observations, et elle tendent à se différencier suivant différents jeux d'observations. Dans le cas de distributions de fréquences non groupées, on peut presque toujours les calculer. Dans le cas de distributions de fréquences groupées, on a un problème quand un des percentiles ou un des quartiles se trouve dans une classe ouverte.

Écart-type

écart-type comme mesure de répartition

Bien que l'étendue soit une mesure simple de variation ou de répartition, elle a beaucoup d'inconvénients. On a besoin d'une mesure qui va éviter les inconvénients en fournissant quand même une bonne mesure de variation. Une de ces mesures est **l'écart moyen**, où on mesure la distance des observations à partir de la moyenne. L'écart moyen inclue cependant des valeurs absolues, et il est difficile de traiter celles-ci en mathématique. **L'écart-type** est basé sur les mêmes principes que l'écart moyen, et on élimine les signes de l'écart par la moyenne en les mettant au carré.

méthode

Comment fonctionne l'écart-type ? Comme la moyenne, l'écart-type prend en compte toutes les valeurs observées. S'il n'y a pas du tout de dispersion dans une distribution, toutes les valeurs observées seront les mêmes. La moyenne sera aussi la même que cette valeur répetée. Si tout le monde avait la même taille de 1,80 mètres, la moyenne serait 1,80 mètres. Aucune valeur observée ne dévierait ou ne diffèrerait de la moyenne. Mais avec la dispersion, les valeurs observées dévient de la moyenne, certaines de beaucoup, d'autres de peu. Le fait de citer l'écart-type d'une distribution est un moyen d'indiquer un certain chiffre "moyen" par lequel les valeurs dévient de la moyenne. Plus grande est la dispersion, plus grands sont les écarts et plus grand est l'écart-type.

principe de l'écart-type

On calcule l'écart-type en additionnant les carrés des écarts des valeurs individuelles par rapport à la moyenne de la distribution, en divisant cette somme par le nombre d'unités dans la distribution, et en extrayant alors la racine carrée de ce chiffre.

Expliquons maintenant en détail la procédure pour calculer l'écart-type.

On prend une population qui consiste en N valeurs $x_1, x_2, x_3 \ldots x_N$ avec une moyenne μ (prononcer mu ou mou), **l'écart-type d'une population** se définit avec la formule :

Formule

$$\textbf{Ecart-type } (\sigma) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Définition

Pour décrire la formule on prendra les étapes de calcul une par une.

. *D'abord on calcule* μ

On calcule μ de la même manière que $\overline{X}$ au chapitre précédent (on additionne tous les chiffres et on divise par le nombre de chiffres). On l'appelle μ quand il s'agit d'une population, plutôt que $\overline{X}$ quand il s'agit d'un échantillon.

. *on soustrait* μ *de chaque valeur* x : $(x_i - \mu)$

. *on met chacune de ces valeurs au carré* : $(x_i - \mu)^2$

. *on additionne ces valeurs :* $\sum (x - \mu)^2$

. *on divise par le nombre d'unités dans la population (N):* $\dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

. *on prend le carré du total :* $\sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

On indique l'écart-type par σ (la lettre grecque sigma minuscule).

variance

Le carré de l'écart-type est appelé la variance, et est désigné par σ^2 . Quand on met au carré le résultat d'une formule qui a une racine carrée, le signe de la racine carrée est annulé et disparait. On a alors :

formule

$$\textbf{Variance } (\sigma^2) = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

variance de l'échantillon

Si on a un échantillon et qu'on veut calculer la **variance de l'échantillon** (ou l'**écart-type de l'échantillon**) pour faire une estimation de la valeur de la population, la formule change légèrement. Dans ce cas, s^2 indique la variance de l'échantillon, $\bar{x}$ la moyenne de l'échantillon, et n la taille de l'échantillon. La formule de la variance de l'échantillon devient alors :

$$\textbf{Variance de l'échantillon } (s^2) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}$$

et **l'écart-type de l'échantillon** devient :

$$\textbf{Ecart-type de l'échantillon } (s) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

échantillon = (n - 1)

Ces formules sont en réalité les mêmes que celles des populations, excepté qu'on a utilisé la formule *(n - 1)* au lieu de *N*. Le plus important c'est que quand on calcule la variance ou l'écart-type d'un échantillon on divise par *(n - 1)*. Quand on calcule la variance ou l'écart-type d'une population on divise par *N*.

écart-type important = répartition importante

Il faut noter que plus les valeurs des unités individuelles diffèrent de la moyenne, plus grands vont être les carrés de ces différences. Par conséquent plus grande va être la somme des carrés et plus grand va être l'écart-type S . Donc, plus la répartition est importante, plus l'écart-type est grand.

exemple

On va examiner maintenant le calcul de l'écart-type en utilisant les données suivantes :

Tableau 6.3 PIB à prix courants, Fidji, 1988–1995 (en millions de FID)

Colonne 1	*Colonne 2*	*Colonne 3*	*Colonne 4*
Année	PIB (en millions de FID)	Écart par rapport à la moyenne ($\bar{x}$ = 1 920,5)	Écart au carré $(x_i - \bar{x})^2$
1988	1.433,3	-487,2	237.363,84
1989	1.588,0	-362,5	131.406,25
1990	1.728,0	-192,5	37.056,25
1991	1.834,8	-85,7	7.344,49
1992	2.019,0	+98,5	9.702,25
1993	2.170,7	+250,2	62.600,04
1994	2.268,9	+348,4	121.382,56
1995	2.351,3	+430,8	185.588,64
Total	15.364,0	0	792.444,32

Source : Current Economic Statistics, July 1996, Bureau of Statistics, Suva, Fiji.

trouver d'abord la moyenne arithmétique

À la colonne 1 on a les huit dernières années et à la colonne 2 les chiffres du PIB pour ces années.

Pour calculer l'écart-type on calcule d'abord $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

= (1.433,3+1.558,0+1.728,0+1.834,8+2.019,0+2.170,7+2.268,9+2.351,3)/8

= 1.920,5 millions de FID

À la colonne 3 on soustrait la moyenne des valeurs du PIB aux valeurs du PIB de chaque année.

À la colonne 4 on met au carré les écarts et on fait la somme de ces carrés, ce qui donne un total de $792.444,32*10^{12}$.

les données d'une population

Si les données proviennent d'une population, alors pour dériver l'écart-type on divise la somme des carrés des écarts par le nombre d'observations ou d'industries (N = 8) et on en extrait la racine carrée. On a donc :

Écart-type d'une population (σ) = $\sqrt{\frac{792.444,32}{8}} = \sqrt{99.055,54} = 314,73$ millions FID

données d'un échantillon

Cependant, si les données proviennent d'un échantillon de la population, alors pour dériver l'écart-type on divise la somme des carrés des déviations par le nombre d'observations ou d'industries moins un (n-1 = 7), et on prend la racine carrée. Dans ce cas on a :

Écart-type de l'échantillon (S) = $\sqrt{\frac{792,444.32}{7}}$ = $\sqrt{113,206.33}$ = 336,46 millions FID

Dans l'exemple, on pense que les données proviennent sûrement d'un échantillon, alors on divise par 7.

délicat avec un grand nombre de chiffres

Bien que ce soit assez facile à calculer pour un petit nombre de chiffres, c'est une procédure très lourde pour des jeux de chiffres importants. Premièrement on doit déterminer la moyenne du jeu de données, ensuite on calcule les écarts de la moyenne pour chaque observation, on les met au carré et on les additionne. Même avec une calculette ces opérations prennent beaucoup de temps. Il vaut mieux utiliser un ordinateur pour faire les calculs.

réorganiser la formule

On peut faciliter grandement les calculs en réorganisant la formule pour la variance. Alors, pour un échantillon, on a :

formule de l'échantillon

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \left(\left[\sum_{i=1}^{n} x_i\right]^2 / n\right)}{n-1}$$

formule de la population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - \left(\left[\sum_{i=1}^{N} x_i\right]^2 / N\right)}{N}$$

étapes de la variance de l'échantillon

Étudions cette formule de près :

Pour la variance de l'échantillon on met d'abord au carré chaque valeur individuelle x : x_i^2

On calcule alors la somme de ces carrés : $\sum_{i=1}^{n} x_i^2$ Appelons-la total A.

On calcule aussi le total des valeurs individuelles x : $\sum_{i=1}^{n} x_i$

On met ce total au carré : $\left[\sum_{i=1}^{n} x_i\right]^2$

on divise par n (le nombre de chiffres dans l'échantillon) : $\left(\left[\sum_{i=1}^{n} x_i\right]^2 \Big/ n\right)$ Appelons le résultat total B

On prend alors A - B et on divise par (n - 1) : $\dfrac{\sum_{i=1}^{n} x_i^2 - \left(\left[\sum_{i=1}^{n} x_i\right]^2 \Big/ n\right)}{n-1}$

pas si compliqué qu'il ne paraît

Bien que la deuxième formule paraisse plus compliquée, elle est en fait bien plus facile que d'utiliser la calculette. Regardons les valeurs de l'échantillon ci-dessous, qui sont les mêmes observations que celles du tableau 6.3

exemple

1.433,3 1.558,0 1.728,0 1.834,8 2.019,0 2.170,7 2.268,9 2.351,3

total et moyenne des observations

a. Pour calculer la variance de l'échantillon on doit obtenir en premier le total et la moyenne des observations. On a :

Σx_i = 15.364,0 n = 8 $\bar{x}$ = 1.920,5

Les écarts de la moyenne sont :

-487,2 -362,5 -192,5 -85,7 +98,5 +250,2 +348,4 +430,8

La somme des carrés de l'écart est 792.444,32. Alors, l'écart est :

s^2 = 792,444.32 / 7 = 113,206.33

deuxième méthode

b. Pour calculer l'écart en utilisant la deuxième méthode, on a besoin de :

Σx_i = 15.364,0 et Σx_i^2 = 30.299.437,12

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\left[\sum_{i=1}^{n} x_i\right]^2 \Big/ n\right)}{n-1}$$

$= [30.299.437,12 - \{(15.364,0)^2 / 8\}] / 7$

$= (30.299.006,32 – 29.506.562,00) / 7$

s^2 = *113.206,33*

la deuxième méthode est plus facile et plus rapide

On voit alors que si on utilise la fonction mémoire d'une calculette, le deuxième calcul peut être fait sans avoir à noter les résultats intermédiaires. On remarque aussi qu'en utilisant l'une au l'autre des méthodes l'écart dérivé reste le même (113.206,33), mais que la deuxième méthode est plus facile et plus rapide.

Propriétés de l'écart-type

rappelez-vous

Quand on utilise l'écart-type il est important de se souvenir de quelques points :

- ☆ on utilise l'écart-type seulement pour mesurer la répartition par rapport à la moyenne;
- ☆ l'écart-type n'est jamais un chiffre négatif;
- ☆ l'écart-type est affecté par les valeurs extrêmes (appelées valeurs aberrantes). Une seule valeur aberrante peut grandement augmenter l'écart-type, et altérer le dessin de la répartition
- ☆ plus la répartition est importante, plus l'écart-type est grand.

Coefficient de variation

la moyenne ajoute un sens à l'écart-type

L'écart-type en soi n'est pas très significatif, à moins qu'on lui associe la moyenne arithmétique. Par exemple, un écart-type de 100 dollars quand la moyenne des revenus est 10.000 dollars montre une variation relative bien plus grande qu'un écart-type de 100 dollars pour une valeur PIB moyenne de 10 millions de dollars. De même lorsqu'on compare la variabilité de deux populations qui ont des unités de mesure différentes (comme les niveaux de revenus en Papouasie-Nouvelle-Guinée (en Kinas) et ceux du Vanuatu (en Vatus)).

intéressé à l'écart par rapport à la moyenne

Donc, la variabilité dans un jeu de données peut être mesurée par rapport à une mesure centrale comme la moyenne arithmétique. Cette mesure est donnée par le **coefficient de variation**, qui est le rapport de l'écart-type à la moyenne arithmétique, exprimé en général en pourcentage, et donné par la formule :

formule

$$\textbf{\textit{Coefficient de variation (C.V.)}} = (\sigma / \bar{x}) \times 100$$

(L'expression $\times$ 100 transforme le chiffre en pourcentage)

on peut comparer les données

La comparaison de la variabilité de deux séries de chiffres implique donc de comparer leurs coefficients de variation respectifs. On peut comparer les coefficients de variation quand :

- les moyennes des distributions comparées sont très éloignées, ou
- les données ont des unités différentes.

pourcentage

Les unités sont converties à un dénominateur commun (un pourcentage).

exemple

Si on regarde les données de PIB du tableau 6.3 on peut calculer le coefficient de variation :

$$\begin{aligned} \text{C.V.(PIB)} &= (\sigma / \bar{x}) * 100 \\ &= 314{,}73 / 1.920{,}5 * 100 \\ &= 16{,}39\% \end{aligned}$$

exemple

On va prendre des données fictives pour illustrer le coefficient de variation.

La moyenne des revenus des propriétaires de maisons en Australie est 40.000 dollars, avec un écart-type de 4.000 dollars. À Kiribati, la moyenne est 12.000 dollars avec un écart-type de 1.200 dollars (on remarque que les moyennes sont très éloignées et les écarts-types différents). Comparez et interprétez la dispersion relative dans les deux groupes de revenus.

solution

La première chose qui vient à l'esprit, c'est que la dispersion est plus grande dans les revenus australiens parce que l'écart-type est plus grand. <ais si on convertit les deux mesures à des termes relatifs en utilisant le coefficient de variation, on trouve que la dispersion relative est la même.

Australie

CV (Australie) = $(\sigma / \bar{X}) * 100$

= \$4.000/\$40.000 * 100

= 10%

Kiribati

CV (Kiribati) = $(\sigma / \bar{X}) * 100$

= \$1.200/\$12.000 * 100

= 10%

CV identique

En fait, les revenus d'Australie et de Kiribati ont le même taux de variation.

exemple

On peut aussi comparer deux types de données différents – les revenus et les âges des propriétaires de maisons. On peut comparer l'étendue des revenus de propriétaires en Australie avec par exemple l'étendue de l'âge des propriétaires. La moyenne d'âge des propriétaires est de 40 ans avec un écart-type de 10 ans.

âge

CV (âge) = $(\sigma / \bar{X}) * 100$

= (10/40) * 100

= 25%

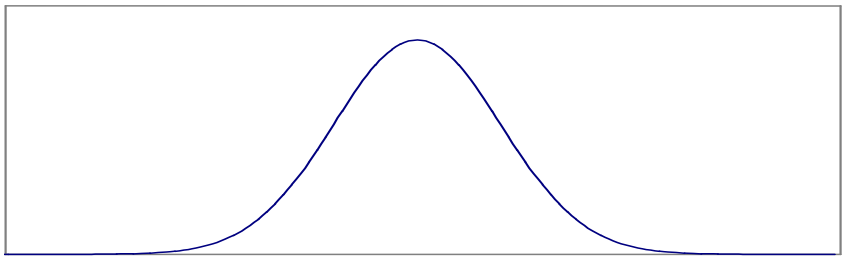
CV (revenus) = 10%

On peut voir qu'il y a une dispersion relative plus grande dans les âges des propriétaires que dans leurs revenus.

Distribution normale

utilisée de manière intensive

La **distribution normale** est utilisée de manière intensive en théorie statistique.



propriétés

La distribution normale a plusieurs propriétés-clés :

- elle est symétrique;
- elle a une forme de cloche;
- la moyenne de la distribution se trouve au sommet de la courbe; et
- la zone qui se trouve en-dessous de la courbe est toujours égale à 1.

toujours avoir les quatre

Les distributions normales peuvent avoir des moyennes et des écarts-types différents, mais elles ont toujours ces quatre propriétés-clés.

des exemples tous les jours

Bien des phénomènes de la vie de tous les jours peuvent être décrits par la courbe normale, par exemple la taille des gens. Un petit nombre de personnes dans la population est de petite taille, un petit nombre est très grand, et la majorité de la population a une taille moyenne. Beaucoup d'autres phénomènes ont également une distribution normale, par exemple des résultats de tests ou des poids.

On pourrait parler sans fin des distributions normales, mais pour l'instant c'est tout ce que vous avez besoin de savoir.

Intervalle de confiance pour un écart-type

analyse de données d'une distribution normale

Lorsqu'on fait l'analyse de données de distributions normales, on utilise l'écart-type et la moyenne pour calculer où les données se placent dans certains intervalles de confiance. Le plus important sur les intervalles de confiance, c'est que pour chaque jeu de données de distribution normale :

intervalle de confiance

- ☆ environ 68% des données se trouvent dans l'intervalle $\bar{x} - s < \bar{x} < \bar{x} + s$

  (c'est-à-dire que 68% des données se situent dans l'étendue entre la moyenne moins l'écart-type et la moyenne plus l'écart-type)

- ☆ environ 95% des données se situent dans l'intervalle $\bar{x} - 2s < \bar{x} < \bar{x} + 2s$
- ☆ environ 99% des données se situent dans l'intervalle $\bar{x} - 3s < \bar{x} < \bar{x} + 3s$

quand $\bar{x}$ = la moyenne

s = l'écart-type

68% d'intervalle de confiance

Si on regarde les données de PIB du tableau 6.3 on peut calculer l'intervalle de confiance de 68% à :

68% intervalle de confiance : $(\bar{x} - s \ , \ \bar{x} + s)$

(1.920,5 - 336,46 , 1.920,5 + 336,46)

(1.584,04 , 2.256,96)

C'est-à-dire que 68% des données se situent dans l'étendue comprise entre 1.584,04 millions et 2.256,96 millions de dollars fidjiens.

95% d'intervalle de confiance

On peut calculer l'intervalle de confiance de 95% à :

95% intervalle de confiance : $(\bar{x} - 2s \ , \ \bar{x} + 2s)$

(1.920,5 - 2(336,46) , 1.920,5 + 2(336,46))

(1.920,5 - 672,92 , 1.920,5 + 672,92)

(1.247,58 , 2.593,42)

C'est-à-dire que 95% des données se situent dans l'étendue comprise entre 1.247,58 millions et 2.593,42 millions de dollars fidjiens.

Résumé des mesures de variabilité

L'ÉTENDUE

- ☺ on la calcule facilement, saufpour les distributions de fréquences, et on la comprend facilement;
- ☹ elle est basée sur les deux observations extrèmes, ce qui la rend très instable;
- ☹ elle est difficile à manipuler en mathématique;
- ☹ elle ne procure aucune information sur le fonctionnement général de la distribution;
- ➡ elle devrait seulement être utilisée comme un guide approximatif du niveau de variabilité.

VARIANCE/ÉCART-TYPE

- ☺ c'est une mesure de variabilité qui utilise des informations provenant de toutes les observations;
- ☺ avec quelques manipulations, les calculs sont assez simples;
- ☺ elle a un rôle central en mathématique et en théorie statistique, et est largement répandue;
- ☹ elle peut être affectée par les valeurs extrèmes;
- ➡ c'est la mesure de variabilité la plus utilisée.

COEFFICIENT DE VARIATION

- ☺ il est indépendant des unités observées. Par conséquent, il est utile pour comparer des distributions où les unités sont différentes;
- ☹ un des inconvénients du coefficient de variation est qu'il est instable quand la moyenne arithmétique est proche de zéro.

Une dernière charactéristique de la distribution

comprendre la structure sous-jacente

Résumer un jeu de données aide à comprendre la structure sous-jacente et le modèle de distribution des valeurs de la variable. On essaye de résumer les données en les réduisant à quelques mesures qui vont donner une indication des valeurs centrales, de la variation des valeurs, de la concentration de fréquences et de la forme de la distribution. La distribution de fréquences décrit la population étudiée, et les mesures de position et de variation aident à caractérisér la distribution par des mesures simples.

distributions asymétriques

Une autre façon de caractériser une distribution est d'étudier son **asymétrie** (c'est-à-dire, si elle n'est pas symétrique, est-ce que les observations sont concentrées dans les valeurs hautes ou basses). On trouve des exemples de distributions asymétriques dans les revenus, ce qui nous intéresse c'est la taille des propriétés foncières, et la taille des ménages. Pour ces distributions, on cherche le type d'asymétrie, s'il y a plus de valeurs élevées que de basses, ou l'inverse.

queue 'droite'

On dit d'une distribution qu'elle a une **asymétrie positive** si les valeurs élevées sont concentrées sur la gauche de la distribution et les valeurs basses sur la droite (c'est-à-dire que si la distribution a une queue 'à droite' et qu'elle a plus de valeurs basses que de valeurs élevées).

queue 'gauche'

On dit d'une distribution qu'elle a une **asymétrie négative** losque les valeurs hautes sont concentrées sur la droite de la distribution et les valeurs basses sur la gauche (c'est-à-dire que la distribution a une queue 'à gauche' et qu'elle a plus de valeurs élevées que de valeurs basses).

trois caractéristiques principales

On peut dire qu'une distribution a trois grandes caractéristiques intéressantes quand on étudie une population :

1 ses valeurs centrales;

2 ses variations à partir des valeurs centrales;

3 si la distribution est symétrique à partir des valeurs centrales, et dans le cas contraire si elle penche vers la droite ou vers la gauche.