M1/Analyse statistique

3. LES CARACTÉRISTIQUES DE TENDANCE CENTRALE
Partie 3. Les taux & la correction des effets de structure

Généralités : les taux en statistique

Le taux en statistique est une proportion. Il s'agit d'un rapport entre deux populations :

- au **dénominateur**, on place la population mère (ou population de référence);

- au **numérateur** figure un sous-ensemble de cette population, dont on cherche à mesurer le poids au sein de cette dernière.

Exemple : le taux de chômage est la proportion de chômeurs (actifs sans emploi et à la recherche d'un emploi) au sein de la population des actifs (actifs occupés + chômeurs). Ces grandeurs sont saisies au même moment.

$$Tx_{Ch.} = \frac{Chômeurs}{Actifs\ occupés + Chômeurs} = \frac{Chômeurs}{Ensemble\ des\ actifs}$$

Généralités : les taux est une moyenne

Le taux en statistique est une caractéristique de tendance centrale. Il s'agit d'une MOYENNE ARITHMÉTIQUE.

Exemple : Reprenons le taux de chômage.

On peut décomposer les chômeurs selon leur âge : les 15-29 ans, les 30-49 ans, les 50-64 ans et les 65 ans et +.

$$(1)\ Tx_{Ch.} = \frac{Ch_{15-29} + Ch_{30-49} + Ch_{50-64} + Ch_{65+}}{Actifs}$$

$$Tx_{Ch.} = \frac{Ch_{15-29}}{Actifs} + \frac{Ch_{30-49}}{Actifs} + \frac{Ch_{50-64}}{Actifs} + \frac{Ch_{65+}}{Actifs}$$

$$(2)\ Tx_{Ch.} = \frac{Tx_{15-29} * Act_{15-29}}{Actifs} + \frac{Tx_{30-49} * Act_{30-49}}{Actifs} + \frac{Tx_{50-64} * Act_{50-64}}{Actifs} + \frac{Tx_{65+} * Act_{65+}}{Actifs}$$

$$(3)\ Tx_{Ch.} = Tx_{15-29} * f_{Act_{15-29}} + Tx_{30-49} * f_{Act_{30-49}} + Tx_{50-64} * f_{Act_{50-64}} + Tx_{65+} * f_{Act_{65+}}$$

$$(4)\ Tx_{Ch.} = \sum_{i=15}^{65+} \left(Tx_i * f_{Act_i} \right)$$

novembre 20 Analyse des données démographiques / JFL 3

(1) On peut décomposer la population des chômeurs en fonction de leur âge. L'ensemble des chômeurs est la somme des chômeurs âgés de 15-29 ans, 30-49 ans, etc.

(2) On a établi que le taux était le rapport entre une sous-population et la population plus globale dont elle est issue: ici les chômeurs, qui est une catégorie d'actifs, divisés par l'ensemble des actifs.

On peut définir le même indicateur au sein de chaque groupe d'âge.

Ainsi, le taux de chômage des personnes âgées de 15-29 ans est le rapport, au sein de la population des 15-29 ans, entre les chômeurs et les actifs. Il s'agit donc du rapport entre les chômeurs âgés de 15-29 ans et les actifs âgés de 15-29 ans.

$$Tx_i = \frac{Ch_i}{Act_i}$$

On peut donc également écrire que le nombre de chômeurs est le produit entre le taux de chômage et le nombre d'actifs :

$$Tx_i * Act_i = Ch_i$$

On se sert de cette expression pour proposer une nouvelle formule du taux de chômage.

(3) Le rapport entre les actifs âgés de 15-29 ans par exemple et l'ensemble des actifs correspond à la proportion d'actifs âgés de 15-29 ans au sein de la population active. Il s'agit donc de la proportion de 15-29 ans parmi la population active. On peut noter cette proportion *f* (pour fréquence relative).

$$f_i = \frac{Act_i}{\sum_{i=15}^{65+} Act_i} = \frac{Act_i}{Actifs}$$

(4) Le taux de chômage est donc la moyenne des taux de chômage par âge pondérés par le poids de chaque catégorie d'âge au sein de la population active.

Généralités : les taux en démographie

Le taux en démographie est un rapport entre deux types de grandeur différentes :

- au **numérateur** figure des événements saisis sur une période de temps qui, par convention, est toujours l'année. Il peut s'agir de décès, de naissances, etc.

- au **dénominateur**, on place la population moyenne au sein de laquelle ces événements ont été observés.

Exemple : le taux de mortalité en France en 2009 est le rapport entre le nombre de décès dénombrés au cours de l'année 2009 et la population moyenne de l'année 2009. Cet indicateur est appelé en démographie *taux brut de mortalité* (TBM).

$$TBM_{2009} = \frac{Décès\ en\ 2009}{Pop.\ moyenne\ 2009}$$

Plus généralement :

$$Taux_N = \frac{Evénements_N}{Pm_N}$$

Le taux en démographie est une proportion. Il s'agit par exemple de la proportion de décès par unité de population (nombre de décès pour 1 000, 10 000 personnes, etc.)

Mais on peut également assimiler cette proportion à une probabilité. On peut d'ailleurs estimer les probabilités de décès à partir des taux de mortalité, moyennant quelques hypothèses.

novembre 20 Analyse des données démographiques / JFL 4

Le taux « démographique » est une caractéristique de tendance centrale qui a une dimension ANNUELLE. On mesure donc la prévalence d'un phénomène sur une durée réelle ou fictive d'une année.

Au numérateur, le nombre d'événements doit donc être annualisé.

C'est déjà le cas quand la période étudiée est l'année (le calcul du taux brut de mortalité au cours d'une année donnée).

Quand la période d'observation n'est pas l'année, il faut ramener le nombre d'observations sur une période d'une année.

Exemple : on cherche à déterminer le taux de mortalité pour la période 2000-2009, soit une période de 10 ans. On cherche donc la valeur MOYENNE prise par cet indicateur au cours de chacune des années de la décennie. Si l'on rapporte le nombre total de décès observés au cours des années 2000-2009, on va cumuler au numérateur 10 années d'observation, tandis qu'au dénominateur la population moyenne correspondra à celle observée au milieu de la période (cf. ci-dessous), c'est-à-dire à un instant donné. Le résultat d'un tel calcul sera 10 fois plus important que le taux moyen observé CHAQUE année.

Il faut donc annualiser le numérateur : on divise le nombre total de décès sur la période par l'amplitude de cette période (ici 10 ans). On calcule donc le nombre annuel moyen de décès au cours de cette période (on fait l'hypothèse sous-jacente que les décès se répartissent de manière linéaire tout au long de la période = autant de décès d'une année à l'autre).

La population moyenne correspond à l'effectif moyen au sein de laquelle ces événements ont été observés. Dans le cas d'une variation linéaire de la population il s'agit de la moyenne entre les populations recensées au début et à la fin de la période d'observation (ici les 1/1 2000 et 1/1/2010)

Exemple : On veut déterminer le taux de mortalité pour le mois de janvier 2009. Cette fois-ci la période de référence est inférieure à l'année. Pour annualiser le nombre de décès en janvier, il faut donc projeter le nombre de décès que l'on aurait sur toute l'année 2009 si l'intensité était celle observée au cours du mois de janvier. On va donc multiplier le nombre de décès dénombrés en janvier par 12 (12 mois). On peut affiner le calcul en tenant compte de la variabilité de la durée des mois de l'année. Dans ce cas, on calcule le nombre moyen de décès observés chaque jour au cours du mois de janvier ($D_{janvier}/31$), puis on multiplie cette moyenne quotidienne par le nombre de jours pour une année (365).

Pour calculer le taux, il reste à diviser ces décès annualisés par la population moyenne de la période d'observation : il s'agit ici de la moyenne des effectifs des populations des 1er janvier et février 2009.

Les taux est une moyenne

Le taux en démographie est également une caractéristique de tendance centrale. Il s'agit d'une MOYENNE ARITHMÉTIQUE.

Exemple : Reprenons le taux mortalité.

On peut décomposer les décès selon leur âge : à 0 an, 1 an, 2 ans, … 103 ans, 104 ans, etc.

$$(1)\ TBM = \frac{D_0 + D_1 + D_2 + \ldots + D_{103} + D_{104}}{Pm}$$

$$TBM = \frac{D_0}{Pm} + \frac{D_1}{Pm} + \frac{D_2}{Pm} + \ldots + \frac{D_{103}}{Pm} + \frac{D_{104}}{Pm}$$

$$(2)\ TBM = \frac{Tx_0 * Pm_0}{Pm} + \frac{Tx_1 * Pm_1}{Pm} + \frac{Tx_2 * Pm_2}{Pm} + \ldots + \frac{Tx_{103} * Pm_{103}}{Pm} + \frac{Tx_{104} * Pm_{104}}{Pm}$$

$$(3)\ TBM = (Tx_0 * f_0) + (Tx_1 * f_1) + (Tx_2 * f_2) + \ldots + (Tx_{103} * f_{103}) + (Tx_{104} * f_{104})$$

$$(4)\ TBM = \sum_{i=15}^{104} (Tx_i * f_i)$$

novembre 20 Analyse des données démographiques / JFL 5

On peut transformer le rapport entre décès et population moyenne en une moyenne de taux pondérés par le poids de chacun des âges au sein de la population totale. Il suffit, par rapport à ce qui a été fait pour le taux de chômage, de remplacer « chômeurs » par « décès », et « actifs » par « population totale ».

(1) On peut décomposer les décès en fonction de leur âge. L'ensemble des décès est la somme des décès à 0 an, 1 an, 2 ans, etc.

(2) On a établi que le taux était le rapport entre des événements et la population plus globale qui « produit » ces événements : ici les décès divisés par la population au sein de laquelle sont observés les décès.

On peut définir le même indicateur pour chaque âge.

Ainsi, le taux de mortalité à 103 ans est le rapport entre les décès de personnes âgées de 103 ans et le nombre moyen de personnes âgées de 103 ans.

$$Tx_i = \frac{D_i}{Pm_i}$$

On peut donc également écrire, par exemple, que le nombre de décès à 103 ans est le produit entre le taux de mortalité à 103 ans et le nombre de personnes âgées de 103 ans :

$$Tx_i * Pm_i = D_i$$

On se sert de cette expression pour proposer une nouvelle formule du taux brut de mortalité.

(3) Le rapport entre les personnes âgées de 103 ans par exemple et l'ensemble des personnes correspond au poids des 103 ans au sein de la population. On peut noter cette proportion *f* (pour fréquence relative).

$$f_i = \frac{Pm_i}{\sum_{i=0}^{104} Pm_i} = \frac{Pm_i}{Pm}$$

(4) Le taux brut de mortalité est donc la moyenne des taux de mortalité par âge pondérés par le poids de âge au sein de la population.

Attention aux comparaisons

La valeur d'un taux moyen est donc le produit d'une interaction entre les taux par âge (par exemple) et la structure par âge.

Quand on fait des comparaisons, on compare donc le produit de cette interaction.

Par exemple, si l'on compare le TBM des populations A et B, on compare deux séries de données différentes : les Tx_i et les f_i.

$$TBM^A = \sum_{i=15}^{104} \left(Tx_i^A * f_i^A \right)$$

$$TBM^B = \sum_{i=15}^{104} \left(Tx_i^B * f_i^B \right)$$

- SI les structures sont différentes (ici les structures par âge),
- ET s'il y a un lien entre le phénomène étudié (ici les décès) et la variable utilisée pour décomposer la population (ici l'âge),

ALORS la comparaison des moyennes va être biaisée. Elle est affectée par un **effet de structure**.

novembre 20 Analyse des données démographiques / JFL 6

Par exemple, les taux de mortalité par âge de la population française sont tous inférieurs (et très nettement) à ceux du Brésil.

Pourtant, le taux brut de mortalité du Brésil est inférieur à celui de la France.

La mortalité est d'autant plus importante que l'âge augmente. Age et mortalité sont donc corrélés. Or, la population brésilienne présente une structure par âge très différente de celle de la France : les jeunes y sont surreprésentés.

De ce fait, la contribution des taux de mortalité aux jeunes âges est très importante dans le calcul de la moyenne des taux (le TBM) au Brésil, ce qui n'est pas le cas en France, où la population est plus équilibrée.

La structure par âge, jeune au Brésil, affecte donc en quelque sorte des coefficients élevés aux taux de mortalité aux jeunes âges (les moins importants), tandis qu'elle minimise le poids des taux aux âges élevés (les plus importants). La moyenne des taux est donc tirée artificiellement vers les valeurs les plus faibles. Ce n'est pas le cas en France, d'où une comparaison « favorable » au Brésil.

(cf. document Excel sur la comparaison de la mortalité en France et au Brésil)

Correction des effets de structure par la *méthode de la population type*

La méthode la population-type consiste à pondérer, pour chacune des populations faisant l'objet d'une comparaison, les taux dont on cherche à calculer la moyenne par une même série de coefficients.

En d'autres termes, on adopte une structure commune : une structure-type.

On calcule alors un taux comparatif. Dans le cas d'une comparaison des niveaux moyens de mortalité par âge, on calcule des taux comparatifs de mortalité (TCM):

$$TCM^{A} = \sum_{i=15}^{104}\left(Tx_i^{A} * f_i^{Type}\right)$$

$$TCM^{B} = \sum_{i=15}^{104}\left(Tx_i^{B} * f_i^{Type}\right)$$

On compare toujours les produits de l'interaction entre la mortalité et la structure par âge. Mais ici on a neutralisé l'effet de l'âge en adoptant une structure type. TCM^A et TCM^B diffèrent seulement par la série de taux de mortalité par âge.

novembre 20 Analyse des données démographiques / JFL 7

Quand on compare deux populations, on peut adopter la structure de l'une des deux populations comme structure de référence. Par exemple, si l'on compare les niveaux moyens de mortalité de la France et du Brésil, on peut calculer le taux brut de mortalité que l'on aurait au Brésil si sa structure par âge était la même que celle de la population française.

$$TCM^{Brésil} = \sum_{i=0}^{104} Tx_i^{Brésil} * f_i^{France}$$

Ce taux fictif est la moyenne des taux REELS pondérés par la structure de la France. On le compare au TBM de la France : $TCM^{Brésil}$ vs. TBM^{France}

$$TCM^{Brésil}\ vs.\ TBM^{France}$$

$$\sum_{i=0}^{104}\left(Tx_i^{Brésil} * f_i^{France}\right) vs \sum_{i=0}^{104}\left(Tx_i^{France} * f_i^{France}\right)$$

En adoptant une même structure par âge, on neutralise l'effet de structure. Les indicateurs produits permettent donc de comparer les niveaux moyens de mortalité compte tenu des différences de structure par âge.

Correction des effets de structure par la *méthode des taux-types*

On peut corriger les effets de structure en utilisant une autre méthode : la méthode des taux-types.

On cherche par exemple à comparer le niveau de mortalité d'une commune à celle de la France, afin de répondre à la question suivante : la mortalité de la commune est-elle différente de celle observée en moyenne en France ?

A l'échelle de la commune, on connaît pour chaque année (t), le nombre de décès annuel (D), la population totale (P) et la structure par âge (f_i). On peut calculer le TBM = D/Pm

On va tester cet indicateur contre la valeur que prendrait le TBM en France si la structure par âge était celle de la commune étudiée. Cet indicateur est appelé le taux-type de mortalité (TTM).

La comparaison se fait sous la forme d'un rapport : l'indicateur comparatif de mortalité (ICM)

$$TBM_t = \frac{D_t^{commune}}{Pm_t} = \sum_{i=0}^{\omega}(t_{i;t}^{commune} \times f_{i;t}^{commune})$$

$$TTM_t = \sum_{i=0}^{\omega}(t_{i;t}^{référence} \times f_{i;t}^{commune})$$

$$ICM = \frac{TBM_t}{TTM_t} = \frac{\dfrac{D_t^{commune}}{Pm_t}}{\sum_{i=0}^{\omega}(t_{i;t}^{référence} \times f_{i;t}^{commune})} \; ou \; \frac{\sum_{i=0}^{\omega}(t_{i;t}^{commune} \times f_{i;t}^{commune})}{\sum_{i=0}^{\omega}(t_{i;t}^{référence} \times f_{i;t}^{commune})}$$

novembre 20 Analyse des données démographiques / JFL 8

L'utilisation de la *méthode de la population-type* suppose que l'on dispose de la série des taux pour chacune des populations que l'on souhaite comparer.

A des échelles démographiques fines, ce n'est pas toujours possible. Par exemple, on ne dispose pas de la série des taux de mortalité par groupe d'âges quinquennaux à l'échelle d'une commune.

Le recours à la *méthode des taux-types* suppose de connaître :

- un indicateur moyen du phénomène à mesurer (par exemple le taux brut de mortalité);

- la série des taux pour la population de référence (taux-types);

- la structure de la population de référence et de la population que l'on « teste ».

On va calculer un taux (moyen) type qui va servir de référence : par exemple le taux-type de mortalité, qui est la moyenne des taux de référence (les taux-types) pondérés par la structure de la population que l'on « teste ».

La comparaison entre la valeur REELLE mesurée pour la population testée (par exemple le taux brut de mortalité) et la valeur FICTIVE calculée pour la population de référence (le taux type de mortalité dans ce cas) se fait sous la forme d'un rapport. Comme tout indicateur comparatif, la valeur servant de référence est placée au dénominateur, la valeur que l'on teste est pour sa part placée au numérateur.

Cet indicateur prend différents noms : l'indicateur comparatif de mortalité (ICM), l'indice standardisé de mortalité (ISM) ou encore en anglais le standardized mortality rate (SMR).

Si cet indicateur est supérieur à 1, alors l'événement étudié est plus fréquent dans la population étudiée; s'il est inférieur à 1, alors l'événement étudié est moins fréquent que dans la population de référence; s'il est égal à 1, il n'y a pas de différence.

Lecture de l'ICM

Prenons l'exemple de la comparaison du taux brut de mortalité d'une commune française.

Par exemple :

Si l'ICM est égal à 1,5, on dit que la mortalité dans la commune étudiée est 1,5 fois supérieure à celle de la France.

Si l'ICM est égal à 0,7, on peut dire que la mortalité dans la commune étudiée est 0,7 fois supérieure à celle de la France. Mais cette expression n'est pas facile à comprendre. Pour améliorer la compréhension, on peut calculer l'inverse de 0,7 (1/0,7). On peut alors dire que la mortalité dans la commune est 1,4 fois MOINS élevée qu'en France (ou que la mortalité est 1,4 fois supérieure en France).

Le choix de la méthode de standardisation

Le choix de la méthode de standardisation dépend des données dont on dispose.

Pour mettre en œuvre *la méthode de la population type*, il faut nécessairement disposer de la série des taux (t_i) dont on cherche à comparer les moyennes pour chacune des populations entrant dans le champ de la comparaison.

Pour mettre en œuvre *la méthode des taux-types*, il faut disposer de la série des taux (t_i) pour la population de référence, et connaître la structure (f_i) (celle dont on cherche précisément à neutraliser les effets) pour la population de référence et celle que l'on étudie. Enfin, il est nécessaire de connaître la moyenne des taux (par exemple le TBM dans le cadre de l'étude de la mortalité) pour cette population étudiée compte tenu de sa structure.