HOMOGÉNÉITE (C5, F3, H4, I2, I4)
(15 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion courante d'**homogénéité** (comme celle, corrélative, d'**hétérogénéité**) possède un contenu « technique » précis :

(a) en **calcul des probabilités**, cette notion s'oppose à celle de **variabilité** d'une **loi de probabilité** ;

(b) en **Statistique**, elle résulte de la comparaison entre des **populations** constituées d'**unités statistiques** données (cf **problème à plusieurs échantillons**). On observe alors, sur ces unités, diverses **variables** qui les décrivent. Les comparaisons portent sur ces variables : **variables numériques**, **variables qualitatives**, **variables morphologiques**.

On peut, le cas échéant, distinguer entre :

(a) l'**homogénéité interne** : celle relative à un même **ensemble** (population ou **échantillon**) ;

(b) l'**homogénéité externe** : celle relative à des ensembles distincts.

Cependant, l'approche statistique peut se développer de façon comparable dans les deux cas.

(i) **Homogénéité interne** (cf eg **analyse de la variance**, **décomposition de la variance**). De manière générale :

(a) tout ensemble constitué d'éléments (unités) « identiques » peut être qualifié d'**ensemble strictement homogène** ou d'**ensemble totalement homogène** ;

(b) à l'opposé, tout ensemble constitué d'éléments (unités) deux à deux « distincts » peut être qualifié d'**ensemble strictement hétérogène** ou d'**ensemble totalement hétérogène** (cf **hétérogénéité**) ;

(c) entre ces deux notions extrêmes, on peut considérer divers **degrés d'homogénéité** ou divers **degrés d'hétérogénéité**. Ainsi, tout ensemble d'éléments dont au moins deux sont distincts peut être qualifié d'**ensemble faiblement hétérogène** et tout ensemble dont au moins deux éléments sont identiques d'**ensemble faiblement homogène** (cf aussi **complexité**).

Dans ce qui précède, on appelle **éléments identiques**, ou **unités identiques** :

(a) soit le même élément répliqué au moins deux fois. Dans ce cas, ce n'est pas l'élément lui-même qui compte, mais une **caractère statistique** (ou plusieurs) observable(s) sur cet élément ;

(b) soit des éléments distincts deux à deux mais assimilables entre eux car leurs caractères possèdent les mêmes valeurs.

A contrario, des éléments distincts diffèrent d'au moins une valeur d'un des caractères.

Ainsi, l'**observation** d'un **caractère statistique** η à valeurs dans un ensemble $\mathcal{Y}$ (numérique ou non), sur un **échantillon** A constitué d'**unités statistiques** $a \in A$, conduit à noter $y = \eta(a)$ la valeur observée sur l'unité a, $\forall\, a \in A$. Par suite, deux unités $\alpha \in A$ et $\beta \in A$ peuvent être considérées comme identiques ssi soit $\beta = \alpha$ (assimilation), soit $\eta(\beta) = \eta(\alpha)$ (égalité des « valeurs »).

(ii) **Homogénéité externe** (cf eg **problème à plusieurs échantillons**). Soit $X_1 = (X_{11}, ..., X_{1N}), ..., X_k = (X_{k1}, ..., X_{kN})$ des **échantillons aléatoires** indépendants (ie des **échantillons iid**) de même taille, resp générés par des **va** $\xi_1, ..., \xi_k$ (non nécessairement définies sur le même ensemble fondamental Ω), mais toutes à valeurs dans le même **espace d'observation** $\mathcal{X}$:

(1) $\quad \xi_i : \Omega_i \mapsto \mathcal{X}, \qquad \forall\, i \in N_k^*.$

On dit que deux échantillons X_i et X_j (avec $j \neq i$) proviennent de **populations homogènes entre elles au sens strict**, ou de **populations strictement homogènes**, Ω_i et Ω_j ssi ξ_i et ξ_j ont même **loi de probabilité** :

(2) $\quad \mathcal{L}(\xi_j) = \mathcal{L}(\xi_j)$, ou encore $P^{\xi(j)} = P^{\xi(i)}$.

en notant resp $\xi(i)$ et $\xi(j)$ pour désigner ξ_i et ξ_j .

On dit aussi que X_i et X_j sont des **échantillons homogènes entre eux** (au sens strict).

A contrario, les populations (donc les distributions qu'elles décrivent) sont plus ou moins hétérogènes entre elles selon qu'une **caractéristique légale** prend des valeurs plus ou moins différentes entre elles : ceci vaut notamment pour une caractéristique de **dispersion** ou pour un **paramètre d'échelle**.

(iii) En pratique, on s'intéresse seulement à certaines caractéristiques légales : eg **espérance**, **dispersion**. Lorsque ces caractéristiques sont égales entre deux populations données, on dit parfois que ce sont des **populations homogènes (entre elles) au sens large**. On peut éventuellement préciser :

(a) si elles sont **homogènes au premier ordre**, ou dans L^1, ie si les espérances $E\,\xi_i$ et $E\,\xi_j$, supposées exister, sont égales ;

(b) si elles sont **homogènes au second ordre**, ou dans L^2, ie si, outre l'égalité des espérances précédentes, les variances $V\,\xi_i$ et $V\,\xi_j$ sont égales (avec $j \neq i$), etc.

La même terminologie s'applique aux **échantillons** extraits des populations.

Ainsi, le **problème de W.U. BEHRENS - R.A. FISHER** consiste à tester l'égalité des espérances (**moyennes**) de deux populations, les variances étant supposées inégales.

(iv) Le vocabulaire précédent est comparable à celui relatif à la **stationnarité** d'un **processus stochastique**, dans la mesure où l'on se réfère aux moments des **lois** ou de leurs caractéristiques.

(v) Il existe de nombreux **tests d'homogénéité** : tests d'égalité des lois ou de leurs **fonctions de répartition** (homogénéité stricte), ou seulement tests d'égalité de leurs espérances, de leurs variances, de certains quantiles, etc.

(vi) Par définition, une population qui n'est pas reconnue comme homogène à une autre est donc appelée **population hétérogène** (hétérogénéité « externe »). Cette notion est ainsi à distinguer de l'**hétérogénéité interne** d'une population, qui peut être dûe eg au fait qu'elle possède une variance « importante » (**loi à queue épaisse**), ou qu'elle constitue un « mélange » de populations différentes (cf **mélange de lois**). Dans ce sens, l'homogénéité interne stricte se réfère à une **loi de DIRAC** (toutes les valeurs d'une variables sont identiques).

De même qu'il existe plusieurs formes de **dépendance** entre **événements** ou entre **va**, il existe ainsi plusieurs formes d'hétérogénéité entre populations (resp entre échantillons issus de ces dernières).

L'expression courante d' « *échantillons provenant (ou issus) de populations différentes* » peut s'interpréter de deux façons :

(a) dans la première, les populations de base Ω_i , munies de leurs **tribus** $\mathcal{T}_i$ et de **probabilités** P_i (avec $i \in N^*_k$), sont considérées comme distinctes. La **variable parente**, définie comme en (1), qui génère l'échantillon X_i , admet donc pour lp $P^{\xi(i)} = \xi_i\,(P_i)$;

(b) dans la seconde, les populations de base Ω_i sont identiques, de même que les tribus $\mathcal{T}_i$ et les probabilités P_i (soit $\Omega_i = \Omega_0$, $\mathcal{T}_i = \mathcal{T}_0$ et $P_i = P_0$, $\forall\ i \in N^*_k$). La va parente :

(3) $\quad \xi_i : \Omega_0 \mapsto \mathcal{X}$

qui génère X_i admet donc, $\forall\ i \in N_k^*$, pour lp la loi $P^{\xi(i)} = \xi_i\,(P_0)$.

Dans les deux cas, un **test d'homogénéité** (eg stricte) porte sur l'**hypothèse** :

(4) $H_0 : \gamma_i = \gamma_0 , \quad \forall\, i \in N_k^*,$

où $\gamma_i = c\,(P^{\xi(i)})$ est la **caractéristique** étudiée, associée à la loi $P^{\xi(i)}$, et où $c : \mathscr{P} \mapsto \Gamma$ est une **application caractéristique** définie sur l'ensemble $\mathscr{P}$ des lois $P^{\xi(i)}$ considérées (en notant, pour simplifier, $\xi(i)$ au lieu de ξ_i).