CORRIGÉ

Exercice 3 : *Richesse des villes et espérance de vie à la naissance*

TABLE 1 – ESPÉRANCE DE VIE MOYENNE DES PERSONNES VULNÉRABLES AU SEIN DES DIX VILLES LES PLUS RICHES ET DES DIX VILLES LES PLUS PAUVRES AUX ETATS-UNIS (RICHESSE URBAINE ÉVALUÉE SUR LA BASE DU REVENU MOYEN PAR HABITANT).

Ville	*Richesse*	*Esp.vie*	*Ville*	*Richesse*	*Esp.vie*
New York City	R	79,5	Gary, Ind.	P	74,2
Santa Barbara	R	79,4	Indianapolis	P	74,6
Santa Rosa	R	79	Detroit	P	74,8
Los Angeles	R	79	Louisville, Ky.	P	74,9
San Francisco	R	78,8	Tulsa, Okla.	P	74,9
San Diego	R	78,8	Toledo, Ohio	P	74,9
Portland, Me.	R	78,2	Oklahoma City	P	75
Boston	R	78,1	Dayton, Ohio	P	75,1
Miami	R	78,3	Knoxville, Ten.	P	75,1
Newmark	R	78,2	Little Rock, Ark.	P	75,1

Note : "R" désigne une ville riche (relativement aux autres), et réciproquement "P" désigne une ville pauvre pa rapport aux autres.
Source : Irwin, N. et Bui, Q. (2016).

1. **Calculer la moyenne et la variance marginale de l'espérance de vie à la naissance.**
Tout au long de cet exercice, la variable X sera le niveau de richesse de la ville (variable qualitative), et la variable Y sera l'espérance de vie à la naissance (variable quantitative continue).

   La moyenne marginales correspond à la moyenne arithmétique de l'ensemble des valeurs de la variable quantitative continue, indépendamment des valeurs prises par la variable qualitative.

   Ainsi, dans le cas de données ponctuelles (cas général) :

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{q} y_j$$

   On peut alors écrire :

$$\bar{y} = \frac{1}{20}(79,5 + 79,4 + 79 + ... + 75,1) = \mathbf{76,8}$$

   Même chose pour déterminer la variance (et l'écart-type) :

$$V(y) = \frac{1}{n}\sum_{j=1}^{q} (y_j - \bar{y})^2$$

$$V(y) = \frac{[(79,5-76,8)^2 + (79,4-76,8)^2 + ... + (75,1-76,8)^2]}{20} = \mathbf{4,10}$$

$$\sigma_y = \sqrt{4,10} = \mathbf{2,03}$$

2. **Déterminer la moyenne et la variance de l'espérance de vie conditionnellement au niveau de richesse de la ville.**
   Il s'agit donc de déterminer les moyennes et les variances conditionnelles. Autrement dit, la moyenne de y sous condition que pour la variable x : $i = R$ ou $i = P$.

Concernant les villes les plus riches aux Etats-Unis :

$$\overline{y_R} = \frac{1}{n_R}\sum_{j=1}^{q} y_{jR} = \mathbf{78{,}73}$$

Et :

$$V_R(y) = \frac{1}{n_R}\sum_{j=1}^{q} (y_{jR} - \overline{y_R})^2 = \mathbf{0{,}26}$$

$$\sigma_{yR} = \sqrt{0,26} = \mathbf{0{,}51}$$

Concernant les villes les plus pauvres aux Etats-Unis :

$$\overline{y_P} = \frac{1}{n_P}\sum_{j=1}^{q} y_{jP} = \mathbf{74{,}86}$$

Et :

$$V_P(y) = \frac{1}{n_P}\sum_{j=1}^{q} (y_{jP} - \overline{y_P})^2 = \mathbf{0{,}08}$$

$$\sigma_{yP} = \sqrt{0,08} = \mathbf{0{,}28}$$

3. **Calculer et interpréter le rapport de corrélation.**

   Le rapport de corrélation est une mesure de la force de la liaison existante entre une variable qualitative (catégorielle) et une variable quantitative continue (numérique). Ce rapport η^2 correspond à la part que représente la variance entre les groupes (variance "INTER") dans la variance totale, elle s'interprète en termes de variance expliquée par la variable qualitative. Dit autrement, cette variance expliquée est une mesure du lien entre le facteur (variable qualitative) X et la variable quantitative continue Y, pour apprécier comment Y dépend du fait d'appartenir à une sous-population ou à une autre.

   Ainsi : $\eta^2 = \frac{V_{inter}(y)}{V_{total}(y)}$

   `Un petit rappel de première année`

   La variance globale avec des données ponctuelles :

$$V_G(y) = \frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m} n_k(\overline{y_k} - \overline{y_G})^2 + \frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m} n_k V_k(y)$$

   En définitive, pour η^2 :

$$\eta^2 = \frac{\frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m} n_k(\overline{y_k} - \overline{y_G})^2}{\frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m} n_k V_k(y) + \frac{1}{\sum_{k=1}^{m} n_k}\sum_{k=1}^{m} n_k(\overline{y_k} - \overline{y_G})^2}$$

$$\eta^2 = \frac{3,744}{3,896} = \mathbf{0{,}96}$$

4. **Commenter les résultats obtenus.**

   Dans le cas présenté, on obtient une valeur du rapport de corrélation $\eta^2 = 0,96$, ce qui correspond à une liaison extrêmement élevée entre le prospérité des habitants des villes américaines et l'espérance de vie à la naissance des personnes vulnérables. En d'autres termes, la variabilité observée de l'espérance de vie des habitants vulnérables des villes de l'échantillon s'explique quasiment intégralement par la prospérité de ces centres urbains. La richesse d'une ville est donc bien un facteur déterminant, voir fondamental, de la qualité de vie et de la santé des habitants vulnérables dans notre échantillon.

Références :

Irwin, N. et Bui, Q. (2016) The Rich Live Longer Everywhere. For the Poor, Geography Matters. *The New York Times*, publié le 04/11/2016.
Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A. et Cutler, D. (2016) The Association Between Income and Life Expectancy in the United States, 2001-2014. *Journal of American Medical Association* 315(16) :1750-1766.
Case, A. et Deaton, A. (2015) Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *PNAS, Proceedings of the National Academy of Sciences* 112(49) :15078-15083.

Corrigé [en ligne le 05/10/2016] [003A]