# CORRIGÉ

## Exercice 1 : *Exercice tiré de l'examen blanc 2015/2016*

TABLE 1 – ÉCHANTILLON D'ADULTES BRÉSILIENS (18-60 ANS) ÉTUDIÉS SUIVANT LEUR INDICE DE MASSE CORPORELLE (IMC) ET LEUR NIVEAU D'ÉTUDE.

| | | *Classe d'IMC* | | |
|---|---|---|---|---|
| | | Bonne santé | Embonpoint | Obésité |
| *Éducation* | Primaire | 27 | 5 | 1 |
| | Secondaire | 24 | 35 | 2 |
| | Supérieur | 6 | 43 | 19 |

*Source :* Inspiré de Kakeshita et Sousa Almeida (2008).

1. **Caractériser les séries statistiques $X$ (niveau d'éducation) et $Y$ (classe d'IMC).**
   Pour caractériser la distribution statistique $X$ (niveau d'éducation), il est nécessaire de présenter :
   — La population : 162 adultes brésiliens (18–60 ans).
   — Par conséquent, l'unité statistique est un adulte brésilien.
   — Le caractère étudié : nous étudions ces adultes selon leur niveau d'étude, selon trois catégories : niveau primaire, secondaire ou supérieur.
   Sachant que nous étudions une caractéristique (un état ou une qualité) présentée par les individus, il s'agit bien d'un caractère non-dénombrable, non mesurable, non-quantifiable, et donc par conséquent qualitatif. Ce caractère comprend trois modalités : primaire, secondaire et supérieur.

   De la même manière, pour caractériser la distribution statistique $Y$ (classe d'IMC), il est nécessaire de présenter :
   — La population : 162 adultes brésiliens (18–60 ans).
   — Par conséquent, l'unité statistique est un adulte brésilien.
   — Le caractère étudié : nous étudions ces adultes selon leur classe d'indice de masse corporelle , selon trois catégories : bonne santé, embonpoint et obésité.
   Sachant que nous étudions une caractéristique (un état ou une qualité) présentée par les individus, il s'agit bien d'un caractère non-dénombrable, non mesurable, non-quantifiable, et donc par conséquent qualitatif. Ce caractère comprend trois modalités : bonne santé, embonpoint et obésité.

2. **Que représentent les marges d'un tableau de contingence ? Indiquez-les pour le tableau 1.**
   D'abord, nous constatons que le tableau de contingence nous permet d'aller plus loin que l'analyse statistique univariée. En effet, il nous permet de savoir comment se distribuent les effectifs de chaque modalité d'un caractère, suivant les modalités de l'autre. Dit autrement, il nous permet de croiser l'information. Ainsi, une population de $n$ individus est décrite selon deux variables statistiques $X, Y$, il s'agit donc d'analyser $n$ selon deux dimensions, on parle alors d'analyse bivariée.

`Détail du tableau vu en TD`

Au sein de ce tableau de contingence, les colonnes et les lignes "total" prennent le nom de "marges". Ainsi, les marges du tableau correspondent aux distributions marginales des deux variables mises en relation.

Par convention, on va remplacer l'indice qui varie par un "**.**" :

Effectifs marginaux de $X$ (en ligne) :

$$\sum_{j=1}^{q} n_{ij} = n_{i.}$$

Effectifs marginaux de $Y$ (en colonne) :

$$\sum_{i=1}^{p} n_{ij} = n_{.j}$$

A l'intersection de la colonne $n_{i.}$ et de la ligne $n_{.j}$, se trouve le "total des totaux", à savoir : $n_{..}$

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = n_{..}$$

Mais aussi :

$$\sum_{j=1}^{q} n_{i.} = n_{..}$$

Ou encore :

$$\sum_{i=1}^{p} n_{.j} = n_{..}$$

TABLE 2 – ÉCHANTILLON D'ADULTES BRÉSILIENS (18-60 ANS) ÉTUDIÉS SUIVANT LEUR INDICE DE MASSE CORPORELLE (IMC) ET LEUR NIVEAU D'ÉTUDE.

| | | *Classe d'IMC* | | | |
|---|---|---|---|---|---|
| | | Bonne santé | Embonpoint | Obésité | **Total** ($n_{i.}$) |
| *Éducation* | Primaire | 27 | 5 | 1 | **33** |
| | Secondaire | 24 | 35 | 2 | **61** |
| | Supérieur | 6 | 43 | 19 | **68** |
| | **Total** ($n_{.j}$) | **57** | **83** | **22** | **162** ($n_{..}$) |

*Source :* Inspiré de Kakeshita et Sousa Almeida (2008).

3. **Donner la signification de :** $n_{11}$**,** $n_{1.}$**,** $n_{.3}$**,** $n_{..}$**.**
   - $n_{11}$ : il s'agit de l'effectif conjoint se situant à l'intersection de la ligne 1 appartenant à la variable $X$ et la colonne 1 appartenant à la variable $Y$. $n_{11} = 27$, soit 27 adultes brésiliens de notre échantillon **présentent à la fois** un niveau d'éducation primaire et un bon état de santé (relativement à leur IMC).
   - $n_{1.}$ : il s'agit de l'effectif marginal de la modalité 1 sur la variable $X$. $n_{1.} = 33$, soit 33 adultes brésiliens de notre échantillon présentent un niveau d'éducation primaire **indépendamment** de leur classe d'IMC.
   - $n_{.3}$ : il s'agit de l'effectif marginal de la modalité 3 sur la variable $Y$. $n_{.3} = 22$, soit 22 adultes brésiliens de notre échantillon sont considérés en situation d'obésité (en fonction de leur IMC) **indépendamment** de leur niveau d'éducation.
   - $n_{..}$ : il s'agit de l'effectif total, autrement dit la population considérée, ici 162 adultes brésiliens tout niveau d'éducation et classe d'IMC confondu.

4. **Comparer et commenter :** $f_{31}$ **et** $f_{33}$**,** $f_{11}$ **et** $f_{13}$**.**
   A partir des effectifs conjoints nous pouvons indiquer les fréquences conjointes (même idée que les fréquences relatives dans une distribution univariée) :

$$f_{ij} = \frac{n_{ij}}{n_{..}}$$

   Ainsi :
   - $f_{31} = \frac{n_{31}}{n_{..}} = \frac{6}{162} = 3{,}7\%$, il y a donc 3,7% des adultes brésiliens de notre échantillon qui sont en bonne santé et qui présentent un niveau d'éducation supérieur.

— $f_{33} = \frac{n_{33}}{n_{..}} = \frac{19}{162} = 11,7\%$, il y a donc 11,7% des adultes brésiliens de notre échantillon qui sont obèses et qui présentent un niveau d'éducation supérieur.

Il y a donc en proportion de l'échantillon plus d'obèses parmi les individus les plus éduqués.

— $f_{11} = \frac{n_{11}}{n_{..}} = \frac{27}{162} = 16,7\%$, il y a donc 16,7% des adultes brésiliens de notre échantillon qui sont en bonne santé et qui présentent un niveau d'éducation primaire.
— $f_{13} = \frac{n_{13}}{n_{..}} = \frac{1}{162} = 0,6\%$, il y a donc 0,6% des adultes brésiliens de notre échantillon qui sont obèses et qui présentent un niveau d'éducation primaire.

Réciproquement à l'interprétation précédente, il y a donc en proportion de l'échantillon moins d'obèses parmi les individus les moins éduqués.

5. **Déterminer et interpréter les profils en ligne et en colonne.**
On notera $f_{i/j}$ ou $f_i^j$ la fréquence de $x = x_i$, sous condition de $y = y_j$. Autrement dit, il s'agit d'exprimer la fréquence de l'effectif d'une modalité d'une variable conditionnellement à la modalité de l'autre variable.

Ainsi :

$$f_{i/j} = \frac{n_{ij}}{n_{.j}}$$

Ainsi, nous pouvons présenter les profils en colonne, de $x$ selon la première modalité de $y$ :

$$f_{1/1} = \frac{n_{11}}{n_{.1}} = \frac{27}{57} = 47,4\%$$

$$f_{2/1} = \frac{n_{21}}{n_{.1}} = \frac{24}{57} = 42,1\%$$

$$f_{3/1} = \frac{n_{31}}{n_{.1}} = \frac{6}{57} = 10,5\%$$

Nous pouvons alors dire qu'au sein des brésiliens en bonne santé dans notre échantillon, la proportion de ceux ayant un niveau d'éducation primaire est de 47,4%, 42,1% pour le niveau secondaire et seulement 10,5% pour le niveau supérieur. Ainsi, au sein des individus en bonne santé, plus le niveau d'étude augmente et plus la proportion d'adultes en bonne santé dans l'échantillon a tendance à diminuer.

Pour la seconde modalité de $y$ :

$$f_{1/2} = \frac{n_{12}}{n_{.2}} = \frac{5}{83} = 6,0\%$$

$$f_{2/2} = \frac{n_{22}}{n_{.2}} = \frac{35}{83} = 42,2\%$$

$$f_{3/2} = \frac{n_{32}}{n_{.2}} = \frac{43}{83} = 51,8\%$$

Nous pouvons alors dire qu'au sein des brésiliens présentant de l'embonpoint dans notre échantillon, la proportion de ceux ayant un niveau d'éducation primaire est de 6%, 42,2% pour le niveau secondaire et 51,8% pour le niveau supérieur. Corroborant le résultat précédent, au sein des individus en sur-poids, plus le niveau d'étude augmente et plus la proportion d'adultes présentant de l'embonpoint dans l'échantillon a tendance à augmenter.

Pour la troisième modalité de $y$ :

$$f_{1/3} = \frac{n_{13}}{n_{.3}} = \frac{1}{22} = 4,5\%$$

$$f_{2/3} = \frac{n_{23}}{n_{.3}} = \frac{2}{22} = 9\%$$

$$f_{3/3} = \frac{n_{33}}{n_{.3}} = \frac{19}{22} = 86,4\%$$

Nous pouvons alors dire qu'au sein des brésiliens en situation d'obésité dans notre échantillon, la proportion de ceux ayant un niveau d'éducation primaire est de 4,5%, 9% pour le niveau secondaire et 86,4% pour le niveau supérieur. Corroborant le résultat précédent, au sein des obèses, plus le niveau d'étude augmente et plus la proportion d'adultes en situation d'obésité dans l'échantillon a tendance à augmenter.

Nous pouvons présenter les profils en ligne, de $y$ selon une modalité de $x$ :

$$f_{j/i} = \frac{n_{ij}}{n_{i.}}$$

Ainsi, nous pouvons présenter les profils en ligne, de $y$ selon la première modalité de $x$ :

$$f_{1/1} = \frac{n_{11}}{n_{1.}} = \frac{27}{33} = 81,8\%$$

$$f_{2/1} = \frac{n_{21}}{n_{1.}} = \frac{5}{33} = 15,2\%$$

$$f_{3/1} = \frac{n_{31}}{n_{1.}} = \frac{1}{33} = 3,0\%$$

Nous pouvons alors dire qu'au sein des brésiliens présentant un niveau primaire dans notre échantillon, la proportion de ceux en bonne santé est de 81,8%, 15,2% pour ceux en situation d'embonpoint et seulement 3% pour les obèses. Ainsi, au sein des individus présentant un niveau primaire, plus le rapport poids/taille augmente et plus la proportion d'adultes ayant un niveau d'éducation primaire a tendance à diminuer dans l'échantillon.

Pour la seconde modalité de $x$ :

$$f_{1/2} = \frac{n_{21}}{n_{2.}} = \frac{24}{61} = 39,3\%$$

$$f_{2/2} = \frac{n_{22}}{n_{2.}} = \frac{35}{61} = 57,3\%$$

$$f_{3/2} = \frac{n_{23}}{n_{2.}} = \frac{2}{61} = 3,3\%$$

Nous pouvons alors dire qu'au sein des brésiliens présentant un niveau secondaire dans notre échantillon, la proportion de ceux en bonne santé est de 39,3%, 57,3% pour ceux en situation d'embonpoint et seulement 3,3% pour les obèses. Ainsi, au sein des individus ayant un niveau secondaire, plus le rapport poids/taille augmente et plus la proportion d'adultes présentant un niveau d'éducation secondaire a tendance à augmenter dans l'échantillon.

Pour la seconde modalité de $x$ :

$$f_{1/3} = \frac{n_{31}}{n_{3.}} = \frac{6}{68} = 8,8\%$$

$$f_{2/3} = \frac{n_{32}}{n_{3.}} = \frac{43}{68} = 63,2\%$$

$$f_{3/3} = \frac{n_{33}}{n_{3.}} = \frac{19}{68} = 31,1\%$$

Nous pouvons alors dire qu'au sein des brésiliens présentant un niveau supérieur dans notre échantillon, la proportion de ceux en bonne santé est de 39,3%, 57,3% pour ceux en situation d'embonpoint et seulement 3,3% pour les obèses. Ainsi, au sein de ceux ayant un niveau supérieur, plus le rapport poids/taille augmente et plus la proportion d'adultes présentant un niveau d'éducation supérieur a tendance à augmenter dans l'échantillon.

6. **Mesurer la dépendance entre ces deux variables.**
   Deux variables sont totalement indépendantes si les variations de l'une n'entrainent aucune variation de l'autre. A l'inverse les variables sont totalement dépendantes lorsque **les fréquences conditionnelles sont égales aux fréquences marginales**.

$$f_{i/j} = f_{i.}$$

Ainsi :

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n_{..}}$$

Donc :

$$n_{ij} = \frac{n_{i.} * n_{.j}}{n_{..}}$$

Dans notre cas :

$$f_{1.} = \frac{n_{1.}}{n_{..}} = \frac{33}{162} = 20,4\%$$

$$f_{2.} = \frac{n_{2.}}{n_{..}} = \frac{61}{162} = 37,6\%$$

$$f_{3.} = \frac{n_{3.}}{n_{..}} = \frac{68}{162} = 42\%$$

On vérifie donc que $f_{1/1} \neq f_{1.}$... Les deux variables ne sont donc pas indépendantes. Toutefois, cette méthode manque un peu de rigueur.

Plus formellement, une statistique existe afin de nous simplifier la tâche. On peut calculer la statistique du $\chi^2$, l'interprétation est simple : plus cette valeur est élevée ($\chi^2 > 0$, nous verrons pourquoi) et plus on va s'éloigner de la situation d'indépendance.

Cette statistique se calcule en deux temps :

— On réalise dans un premier temps le produit des marges : $\hat{n_{ij}} = \frac{n_{i.} * n_{j.}}{n_{..}}$ [on retrouve ici le calcul précédent].
On va ensuite mesurer l'écart entre la situation observée (réelle) et cette situation d'indépendance théorique grâce à la statistique du $\chi^2$ :

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(n_{ij} - \hat{n_{ij}})^2}{\hat{n_{ij}}}$$

Ou :

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} (\frac{n_{ij}}{\hat{n_{ij}}})^2 - n_{..}$$

TABLE 3 – TABLEAU DU PRODUIT DES MARGES.

| | | *Classe d'IMC* | | | |
|---|---|---|---|---|---|
| | | Bonne santé | Embonpoint | Obésité | **Total** ($n_{i.}$) |
| *Éducation* | Primaire | 11,61 | 16,91 | 4,48 | **33** |
| | Secondaire | 21,46 | 31,25 | 8,28 | **61** |
| | Supérieur | 23,93 | 34,84 | 9,23 | **68** |
| | **Total** ($n_{.j}$) | **57** | **83** | **22** | **162** ($n_{..}$) |

*Source :* Inspiré de Kakeshita et Sousa Almeida (2008).

Dans notre cas :

$$\chi^2 = (\frac{(27-11,61)^2}{11,61}) + (\frac{(5-16,91)^2}{16,91}) + (\frac{(1-4,48)^2}{4,48}) + ... + (\frac{(19-9,23)^2}{9,23})$$

$$\chi^2 = 20,40 + 8,39 + 2,70 + 0,30 + 0,45 + 4,76 + 13,43 + 1,91 + 10,34 = \mathbf{62,68}$$

On vérifie alors que $\chi^2 > 0$, et on conclue donc à la dépendance entre les deux variables.

Bonus **Construire et interpréter le test d'indépendance ($\chi^2$) au seuil de 5%.** Table

En dépassant le cadre strict des statistiques descriptives, nous pouvons tester le lien de dépendance entre deux variables à l'aide du test du $\chi^2$. Quatre étapes sont nécessaires :

(a) Formuler les hypothèses du test du $\chi^2$ :
On pose :
— $H_0$ : les variables $X$ et $Y$ sont indépendantes.
— $H_1$ : les variables ne sont pas indépendantes.

(b) Détermination des degrés de liberté (ddl) :
— On détermine $p$ le nombre de lignes et $q$ le nombre de colonnes, dans la mesure où : la distance du $\chi^2$ suit la loi du $\chi^2$ à $(p-1)(q-1)ddl$.

(c) On définit le seuil (la marge d'erreur) : ici 5%.
— Cela dépend de la précision attendue par l'analyste lors du test, par exemple l'influence d'un médicament ne sera pas évaluée avec la même précision que la réussite des étudiants aux examens.
— Dans notre cas, cela signifie que nous avons 5% de chance de rejeter $H_0$ à tord.
— Il ne faut pas sous-estimer les probabilités faibles !

(d) Enfin, on va chercher dans la table du *chi*$^2$ afin de repérer la valeur critique au croisement de la ligne ddl et du seuil d'erreur, on va donc chercher $K_{95\%} = F(0,05;4)$

(e) On compare enfin cette statistique à la valeur du *chi*$^2$ trouvée précédemment.
— Si la valeur du $\chi^2$ est inférieure à la valeur critique (issue de la table), cela conduit à accepter $H_0$ et à conclure à l'indépendance entre les deux variables.
— Inversement si la valeur est supérieure.

Dans notre cas :

$$\chi^2 = 62,68$$

Dans la table, on observe que :

$$K_{95\%} = F(0,05;4) = 0,71$$

On doit donc rejeter l'hypothèse $H_0$ et accepter l'hypothèse alternative $H_1$, afin de conclure à la dépendance entre les deux variables au seuil de 5%. Autrement dit, il existe bien une dépendance entre le niveau d'éducation et le rapport poids/taille au seuil de 5% dans notre échantillon.

*De façon plus systématique encore, il semble bien qu'il existe une certaine dépendance entre le rapport poids/taille et le niveau d'éducation. Dans le cas contraire toutes catégories présenteraient des fréquences équivalentes. Les plus éduqués ont certainement plus de moyens et donc effectuent moins d'activité physique quotidiennement, la relation est sans doute forte avec la possession d'une automobile. De plus, la perception de l'embonpoint ou de l'obésité peut varier en fonction des régions. Dans les régions plus rurales, l'empbonpoint peut être synonyme de bonne santé (souvent le cas lorsque la population de référence était majoritairement sous-nutrie ou mal-nutrie), au contraire des normes sociales dans les régions plus urbanisées. Enfin le droit du travail reste peu procteur au niveau du partage temps de travail et temps dédié à la vie de familiale. Ainsi les plus éduqués, sont peut-être les plus à mêmes à avoir recours aux produits à faible qualité nutritive issus de l'agroalimentaire par manque de temps ou correspondance aux standards internationaux.*

Références :

Kakechita, I.S. et Sousa Almeida, S. (2008) The relationship between body mass index and body image in Brazilian adults. *Psychology and Neuroscience* 1(2) :103-107.
Mintem, G.C., Petrucci Gigante, D. et Lessa Horta, B. (2015) Change in body weight and body image in young adults: a longitudinal study. *BMC Public Health* 15 :222.

Corrigé [en ligne le : 05/10/2016] [001A]