Auto-évaluation – Analyse de fréquences et test du khi carré (χ^2)

Solutions

1. a) Lequel des éléments s'avère un problème pour le test du khi-carré (χ^2) ?

Réponse : **b) Fréquences théoriques < 5.** La restriction concerne les fréquences théoriques et non observées. L'hétérogénéité de la variance est une supposition de modèles qui dépendent de la distribution normale, tels que le test-t de Student.

b) Un tableau de contingence résume les fréquences pour chaque combinaison des niveaux de deux variables catégoriques différentes. *Réponse :* **Vrai**

c) Nommez un test équivalent au χ^2 qui s'applique dans les mêmes conditions que ce dernier.

Réponse : Le test-G est un test équivalent au χ^2.

2. Un concessionnaire automobile désire mieux cibler sa clientèle dans une ville. Il demande une étude d'observation afin de déterminer si les résidants de quatre arrondissements (A, B, C, D) d'une même ville utilisent plutôt des voitures produites par des manufacturiers domestiques (GM, Chrysler, Ford) ou produites par des compagnies étrangères (p.ex., Honda, Toyota, Subaru, Mitsubishi, Mazda, Nissan, Suzuki, Kia, Hyundai, BMW, VolksWagen, Mercedes, Mini, Volvo, SAAB, SMART). On échantillonne aléatoirement 300 voitures dans les quartiers de chacun des quatre arrondissements. On note la marque de chaque voiture. Le tableau 1 présente les données. a) Créez un objet approprié pour stocker les données de

Tableau 1 – Tableau de données des fréquences pour chaque type de véhicule (importé ou domestique) dans les quatre arrondissements (A, B, C, D).

	A	B	C	D
Domestique	125	223	62	180
Importé	175	77	238	120

ce tableau de contingence.

Réponse : On peut stocker les données dans une matrice :

```
> ##matrice
> autos <- matrix(data = c(125, 223, 62, 180,
                           175, 77, 238, 120), nrow = 2, ncol = 4, byrow = TRUE)
> ##on peut ajouter les étiquettes
> colnames(autos) <- c("A", "B", "C", "D")
> rownames(autos) <- c("Domestique", "Importe")
> ##visualisons le tout
> autos
             A   B   C   D
Domestique 125 223  62 180
Importe    175  77 238 120
```

b) Énoncez les hypothèses statistiques pour l'analyse de ces données, ainsi que le seuil de signification.

Réponse : On peut tester les hypothèses statistiques suivantes, soit formulées en terme d'indépendance ou encore en terme de proportions. Les deux formulations sont équivalentes.

Formulation en termes d'indépendance :

H_0 (indépendance) : L'origine des véhicules est indépendante de l'arrondissement de la ville.

H_a (non-indépendance) : L'origine des véhicules n'est pas indépendante de l'arrondissement de la ville.

$\alpha = 0.05$

OU

Formulation en termes de proportions :

H_0 : La proportion des véhicules d'origine domestique ne diffère pas selon l'arrondissement de la ville.

H_a : La proportion des véhicules d'origine domestique diffère selon l'arrondissement de la ville.

$\alpha = 0.05$

c) Choisissez l'analyse appropriée et justifiez votre choix.

Réponse : On peut analyser les données du tableau de contingence avec un test du χ^2 puisque les données ont été obtenues à l'aide d'un dispositif complètement aléatoire et parce que nous avons 5 fois plus d'observations qu'il y a de cellules dans le tableau (300 fois plus d'observations). Par conséquent, aucune fréquence théorique n'est inférieure à 5.

d) À l'aide de `R`, effectuez l'analyse que vous avez choisie précédemment et interprétez les résultats.

Réponse : Si on a choisi le χ^2, on procédera comme suit :

```
> chisq.test(x = autos)

        Pearson's Chi-squared test


data:  autos
```

```
X-squared = 194.36, df = 3, p-value < 2.2e-16
```

On observe un χ^2 de 194.36 avec trois degrés de liberté et $P(\chi^2_{0.05,3} > 194.36) < 0.0001$. On rejette H_0 et on conclut que l'origine des véhicules n'est pas indépendante de l'arrondissement. L'interprétation alternative, en termes de proportion, indique que la proportion de véhicules d'origine domestique varie avec le quartier. En d'autres mots, la compagnie peut viser certains quartiers afin d'optimiser sa campagne de publicité.