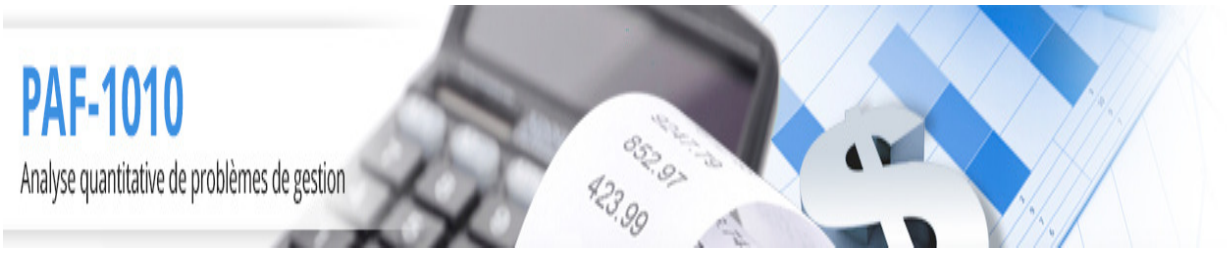
MODULE 12

TESTS
DU
KHI-DEUX



Louis Houde
Département de Mathématiques et d'informatique
Université du Québec à Trois-Rivières

12 Tests du khi-deux

La statistique du khi-deux est particulièrement adaptée pour les observations qualitatives. On développe dans ce module une serie de tests pour ce type de données

Objectifs et compétences

L'objectif de cette partie est de montrer à l'étudiant les méthodes pour l'analyse des données de type qualitatif.

L'étudiant sera en mesure de

- établir les hypothèses statistiques
- choisir le test adapté
- calculer la statistique du test du khi-deux et effectuer le test associé
- interpréter les résultats du test

Tests et statistique

Les différents tests du khi-deux

Le khi-deux est une statistique permettant de comparer les effectifs (fréquences) observés dans un échantillon avec des fréquences théoriques qui découlent des hypothèses statistiques. On s'intéresse dans ce module à quatre situations dans lesquelles la statistique est applicable pour effectuer un test d'hypothèse

Ajustement On suppose que la loi de probabilité de la variable aléatoire qualitative (ou quantitative avec peu de modalités) est connue et on veut vérifier c'est le cas. C'est le cas classique du lancer d'un dé. On suppose que chaque face a une probabilité identique et on veut vérifier si le dé est équilibré.

Homogénéité La variable aléatoire qualitative provient de k populations et on veut vérifier si la loi de probabilité est la même dans chaque population. On a donc k échantillons et on mesure la même caractéristique dans chacune d'elles. C'est le cas lorsqu'on veut savoir si la satisfaction (en quelques catégories) par rapport au service de transport en commun est semblable entre trois villes canadienne.

Indépendance On mesure deux variables aléatoires qualitatives dans une population et on veut savoir si ces variables sont indépendantes c'est-à-dire si la connaissance d'une des v.a. peut influencer la loi de probabilité de l'autre. C'est le cas lorsqu'on veut vérifier si la satisfaction (en quelques catégories) par rapport au service de transport en commun est indépendant de la fréquence d'utilisation (en quelques catégories) de ces transports. Il n'y a qu'une petite nuance entre l'homogénéité et l'indépendance.

Égalité de proportions On est dans le contexte d'un test d'homogénéité mais la variable n'a que deux modalités que l'on peut qualifier de "succès" ou d'"échec" ET il n'y a que deux populations. Le fait de se demander si les deux populations ont la même distribution pour la variable mesurée c'est la même chose que de vérifier si les deux proportions de succès sont identiques. Cela mérite une section particulière puisque c'est le seul test du khi-deux qui peut se décliné en unilatéral ou bilatéral. On utilise ce test lorsqu'on veut savoir si le taux de réussite chez les hommes dans un programme d'administration est le même que le taux de réussite chez les femmes.

Les tests du khi-deux demandent un calcul assez long et malheureusement ils ne sont pas disponibles directement dans le logiciel Excel. Il faut donc apprendre à faire le calcul avec la calculatrice tout en considérant que lors d'un examen on tentera de réduire le plus possible la complexité du calcul requis.

Statistique du test

L'idée des tests du khi deux est de comparer les valeurs observées et les valeurs moyennes qu'on observerait si l'hypothèses nulle est vraie. Considérons le cas d'un test visant à vérifier si un dé est équilibré c'est-à-dire si chacune des faces avait la même probabilité $(1/6)$. Si on lance le dé 500 fois on devrait retrouver en moyenne $500 * 1/6 = \frac{250}{3} = 83.333$ fois la valeur "1" et 83.333 fois la valeur "2", etc. Supposons qu'on observe 90 valeur "1" sur les 500 lancers, 74 fois la valeur "2", 68 fois la valeur "3", 105 fois la valeur "4", 85 fois la valeur "5" et finalement $500 - (90 + 74 + 68 + 105 + 84) = 79$ fois la valeur "6". On cherche à établir si la différence entre les valeurs observées et les valeurs théoriques est importante ou simplement due à une variation aléatoire.

Posons n_i la valeur observée pour le nombre de fois que le "i" est sorti et T_i la valeur moyenne attendue. Si on fait simplement la différence entre les deux on obtient toujours 0 :

$$\sum (n_i - T_i) = \sum n_i - \sum T_i = n - n = 0$$

ce qui n'est pas particulièrement pratique. La statistique du khi deux utilise donc la différence au carré : $\sum (n_i - T_i)^2$. Or cette dernière façon de considérer les différences entre les valeurs qui donne un poids trop grand pour les petites valeurs de n_i : si on a une valeur théorique de 10 pour une modalité et une valeur observée de 5 alors la différence est la même que si on a une modalité avec une valeur théorique de 500 et une valeur

observée de 505. Il y a dans les deux cas une différence de 5 unités mais dans le premier cela correspond à une diminution de 50% et dans le deuxième à $5/500 * 100 = 1\%$. Pour éviter cette disproportion pour une modalité en particulier la statistique du khi deux est donnée par

$$\sum \frac{(n_i - T_i)^2}{T_i}$$

soit la différence relative. Dans tous les cas le principe est le même, seule la formulation des fréquences théoriques diffèrent selon les hypothèses.

Test d'ajustement du khi-deux

Le test d'ajustement du khi-deux permet de vérifier qu'une variable qualitative ou quantitative discrète mesurée dans une population suit une loi de probabilité théorique connue. Considérons un dé à six faces et supposons que l'on veuille vérifier s'il est bien équilibré. On peut effectuer un test pour chaque face séparément ou utiliser la loi de probabilité de la variable aléatoire qui donne le nombre de points sur la face visible du dé. Dans ce cas il suffit de confronter les hypothèses

$$H_0 \quad : \quad \pi_i = \frac{1}{6} \text{ pour chaque } i = 1, 2, \ldots 6$$
$$H_1 \quad : \quad \pi_i \neq \frac{1}{6} \text{ pour au moins un } i$$

On peut tester l'ensemble des faces en une seule opération à l'aide d'un test d'ajustement du khi-deux.

On cherche à déterminer s'il y a une différence dans le nombre de créations d'entreprises dans l'année (les saisons plus spécifiquement). Les hypothèses à confronter sont

$$H_0 \quad : \quad \pi_i = \frac{1}{4} \text{ pour } i = \text{"été", "printemps", "automne", "hiver"}$$
$$H_1 \quad : \quad \pi_i \neq \frac{1}{4} \text{ pour au moins un } i$$

où π est la probabilité de créer une entreprise.

Soit X une v.a. discrète de support S_X et loi de probabilité

$$f(x_i) = \pi_i \text{ pour } x_i \in S_X$$

et considérons les hypothèses statistiques :

$$H_0 : \pi_i = \pi_{i0} \text{ pour chaque } i$$
$$H_1 : \pi_i \neq \pi_{i0} \text{ pour au moins un } i$$

où π_{i0} sont des constantes connues.

Le test d'ajustement du khi-deux de niveau α pour confronter ces hypothèses est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \geq \chi^2_{k-1;\alpha}$$

où

$$n_i = np_i$$
$$T_i = n\pi_{i0}$$

et $\chi^2_{k-1;\alpha}$ est le point critique de niveau α pour une loi khi-deux de paramètre $k-1$.

Conditions d'application : Le test approximatif est valide si

a. $T_i \geq 1$ pour chaque i
b. Il y a un maximun de 20% des valeurs T_i qui sont moins grandes que 5

Remarque 12.1 *Les deux conditions d'application sont connues comme étant la règle de Cochran.*

Exemple 12.1 ★*Dans le but de vérifier si un dé est bien équilibré une machine "lance" le dé 1000 fois et on observe le nombre de points sur la face visible du dé. Les résultats sont donnés dans le tableau suivant :*

Face	*1*	*2*	*3*	*4*	*5*	*6*
Observations	*180*	*167*	*158*	*210*	*135*	*150*

Faire un test au niveau 5% pour vérifier si le dé est équilibré.

Solution : *Considérons la v.a. qui donne le nombre de points sur la face visible du dé, on veut confronter les hypothèses*

$$H_0 : \pi_i = \frac{1}{6} \text{ pour chaque } i = 1, 2, \ldots 6$$
$$H_1 : \pi_i \neq \frac{1}{6} \text{ pour au moins un } i$$

Le test d'ajustement du khi-deux est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \geq \chi^2_{k-1;\alpha}$$

où $k = 6$ et $\alpha = 0.05$. On obtient

x_i	*1*	*2*	*3*	*4*	*5*	*6*
T_i	166.67	166.67	166.67	166.67	166.67	166.67

et ainsi les conditions d'application du test du khi-deux sont respectées.

On observe

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \\ &= \frac{(180 - 166.67)^2}{166.67} + \frac{(167 - 166.67)^2}{166.67} + \cdots \\ &= 20.468\end{aligned}$$

Or $\chi^2_{5;0.05} = 11.07$ donc on rejette H_0 et on doit conclure avec un niveau de 5% que le dé n'est pas équilibré.

Exemple 12.2 ★★*Une étude sur la création d'entreprises vise à vérifier s'il y a une variabilité au cours de l'année. On observe 52 créations d'entreprises en 2007 et la distribution selon les saisons est la suivante :*

Saison	*Été*	*Automne*	*Hiver*	*Printemps*
Créations	*10*	*21*	*8*	*13*

Faire un test au niveau 10% pour vérifier s'il y a une fluctuation dans l'année.

Solution : *On veut confronter les hypothèses*

$$\begin{aligned}H_0 &: \pi_i = \frac{1}{4} \text{ pour } i = \text{"été", "printemps", "automne", "hiver"} \\ H_1 &: \pi_i \neq \frac{1}{4} \text{ pour au moins un } i\end{aligned}$$

où π_i est la probabilité de création de l'entreprise à la saison i. Le test de niveau α est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \geq \chi^2_{k-1;\alpha}$$

k étant le nombre de saisons soit 4. On obtient

$$T_i = 52 * \frac{1}{4} = 13$$

pour chaque saison et ainsi les conditions d'application du test d'ajustement sont re-

spectées. Selon l'échantillon on observe

$$\begin{aligned}\chi^2 &= \frac{(10-13)^2}{13}+\frac{(21-13)^2}{13}+\\ &\quad \frac{(8-13)^2}{13}+\frac{(13-13)^2}{13}\\ &= 7.538\,5\end{aligned}$$

tandis que le point critique est

$$\chi^2_{k-1;\alpha}=\chi^2_{3;0.1}=6.251\,4$$

On rejette alors H_0 au niveau 10% et on peut dire qu'il y a une différence selon les saisons.

Test d'indépendance pour deux variables discrètes

Lorsque deux variables dicrètes ou qualitatives sont mesurées sur les mêmes individus on est en présence d'une population et de deux mesures. Il est alors intéressant de vérifier si ces variables aléatoires sont indépendantes c'est-à-dire si elles ont une influence l'une sur l'autre. La notion même de dépendance doit être définie. Intuitivement, il y a indépendance entre deux v.a. si le fait de connaître le résultat d'une ne donne aucune information sur le résultat de la deuxième. Plus précisément, il y a indépendance entre deux v.a. X et Y si

$$\Pr(X=x \text{ et } Y=y)=\Pr(X=x)\times\Pr(Y=y)$$

ce qui revient à dire que

$$\Pr(X=x \mid Y=y)=\Pr(X=x)$$

et

$$\Pr(Y=y \mid x=x)=\Pr(Y=y)$$

Cette définition rejoint la définition d'indépendance entre deux événements définie dans la section sur les probabilité. Les hypothèses statistiques à confronter pour X et Y deux variables aléatoires qualitatives ou quantitatives discrètes sont

$$\begin{aligned}H_0 &: \Pr(X=x \text{ et } Y=y)=\Pr(X=x)\Pr(Y=y) \text{ pour tout } x,y \qquad (12.1)\\ H_1 &: \Pr(X=x \text{ et } Y=y)\neq\Pr(X=x)\Pr(Y=y) \text{ pour au moins un } x,y\end{aligned}$$

Cette formulation de l'indépendance étant un peu rébarbative on écrit généralement les hypothèses :

$$\begin{aligned}H_0 &: X \text{ et } Y \text{ sont indépendantes}\\ H_1 &: X \text{ et } Y \text{ sont dépendantes}\end{aligned}$$

sous entendu que cela correspond à la formulation ci-haut.

Pour effectuer le test d'indépendance on utilise la statistique du khi-deux. Cette dernière est assez complexe à calculer c'est pourquoi on passe par le tableau de contingence des observations et le tableau des valeurs attendues ou théoriques. Il est alors plus facile de calculer la valeur de la statistique.

Tableau de contingence

Lorsque deux v.a. sont discrètes, il est possible de représenter les résultats d'un échantillon de taille n par un tableau de contingence :

$X \backslash Y$	mod 1	$\cdots$	mod j	$\cdots$
mod 1	n_{11}		n_{1j}	
$\vdots$				
mod i			n_{ij}	
$\vdots$				

où n_{ij} est le nombre de sujets pour lesquels la v.a. X a la modalité i et la v.a. Y a la modalité j. En plus de ces informations il est intéressant de mettre dans le tableau les marginales pour la v.a. X et la v.a. Y, c'est-à-dire les fréquences par variable aléatoire

$X \backslash Y$	mod 1	$\cdots$	mod j	$\cdots$	
mod 1	n_{11}		n_{1j}		$n_{1.}$
$\vdots$					
mod i			n_{ij}		$n_{i.}$
$\vdots$					
	$n_{.1}$		$n_{.j}$		n

où n et la taille d'échantillon, $n_{i.}$ est la fréquence de la modalité i de la v.a. X et $n_{.j}$ est la fréqence de la modalité j de la v.a. Y. On a $n_{1.}/n$ une estimation de la probabilité que la v.a. X prenne la modalité 1, $n_{.1}/n$ une estimation de la probabilité que la v.a. Y prenne la modalité 1 et n_{11}/n une estimation de la probabilité que les v.a. X et Y prennent les modalités i et j respectivement.

Statistique du khi-deux

S'il y a indépendance on devrait avoir

$$\frac{n_{ij}}{n} \simeq \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$$

Posons $T_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ la fréquence attendue pour les modalités i et j s'il y avait indépen-

dence. La statistique pour le test du khi deux est donnée par

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}}$$

où k est le nombre de modalités de X et m est le nombre de modalités de Y. Cette statistique est une mesure de la dépendance entre les v.a. X et Y.

Le test d'hypothèses pour confronter

$$H_0 \quad : \quad \text{les v.a. sont indépendantes}$$

$$H_1 \quad : \quad \text{les v.a. sont dépendantes}$$

est de rejeter H_0 si $\chi^2 \geq \chi^2_{(k-1)(m-1);\alpha}$, c'est-à-dire si la statistique est plus grande que le point critique de niveau α d'une loi khi deux à $(k-1)\,(m-1)$ degrés de liberté.

Conditions d'application :

Ce test approximatif est valide si (règle de Cochran)

- $T_{ij} \geq 1$ pour tout i et j
- Il n'y a pas plus de 20% des valeurs T_{ij} plus petites que 5.

Remarque 12.2 *Le logicels EXCEL possède une fonction permettant de faire le calcul du seuil de signification empirique si on dispose des fréquences observées et des fréquences attendues :*

=TEST.KHIDEUX(PLAGE OBSERVÉE;PLAGE ATTENDUE)

Pour obtenir la valeur de la statistique du khi-deux, il faut faire la formule suivante

=SOMME((PO-PA) ^2/PA)

où PO et PA sont respectivement les plages observée et attendue.

Exemple 12.3 ★★*Pour cibler la clientèle d'un nouveau produit de consommation, une entreprise fait un sondage auprès de 321 personnes. L'intérêt dans le produit est noté par "aucun intérêt", "un intérêt mineur" ou un "intérêt important". La situation familiale (au moins un enfant à charge : oui ou non) est notée également. On cherche à vérifier si l'intérêt dans le produit dépend de la situation familiale. Les résultat sont les suivants*

Enfant	*aucun*	*mineur*	*important*
oui	*10*	*12*	*3*
non	*7*	*38*	*9*

On a donc 79 personnes qui répondent. On veut vérifier s'il y a un lien entre les deux mesures au niveau 5%

Solution *: On cherche à confronter les hypothèses* H_0 *: indépendance entre la v.a. famille et intérêt dans le produit et* H_1 *: dépendance entre la v.a. famille et intérêt dans le produit. Le niveau est fixé à 5% c'est-à-dire que la probabilité de dire qu'il y a dépendance étant donné qu'il y a indépendance entre ces deux variables est de 5%.*

Le test est de rejeter H_0 *si*

$$\chi^2 = \sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \geq \chi^2_{(m-1)(k-1);\alpha} = \chi^2_{2;0.05} = 5.991\,5$$

On obtient le tableau de contingence suivant :

n_{ij}	*aucun*	*mineur*	*important*	
oui	*10*	*12*	*3*	*25*
non	*7*	*38*	*9*	*54*
	17	*50*	*12*	*79*

et le tableau des fréquences théoriques :

T_{ij}	*aucun*	*mineur*	*important*	
oui	*5.400*	*15.823*	*3.800*	*25*
non	*11.620*	*34.177*	*8.203*	*54*
	17	*50*	*12*	*79*

Il y a une cellule sur 6 qui contient une valeur attendue plus petite que 5. Cela correspond à $1/6 * 100 = 16.667\%$ *des valeurs attendues, soit moins de 20%. Le test est donc valide.*

La statistique observée est

$$\begin{aligned}\sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} &= \frac{(10 - 5.400)^2}{5.400} + \ldots \\ &= 7.401\end{aligned}$$

Comme la statistique est plus grande que le point critique on accepte H_0.

On peut aussi utiliser EXCEL pour faire les calculs. On obtiendrait les tableaux suivants :

OBSERVATIONS

	aucun	mineur	important	
famille	10	12	3	25
non	7	38	9	54
	17	50	12	79

VALEURS ATTENDUES

	aucun	mineur	important	
famille	5,37974684	15,8227848	3,79746835	25
non	11,6202532	34,1772152	8,20253165	54
	17	50	12	79

Khi deux	niveau de signification
7,40118126	0,02470893

Calcul de la statistique du khi deux et de son niveau de signification

Exemple 12.4 ★★*Un chercheur veut vérifier si deux universités ont un même barême*

pour l'attribution des cotes. Pour ce faire il choisit un échantillon de 21000 étudiants provenant des deux université et il regarde les cotes attribuées aux étudiants de 2001 :

Cote	*A*	*B*	*C*	*D*	*E*
Université I	*605*	*1400*	*1789*	*300*	*70*
Université II	*2014*	*4178*	*8032*	*2005*	*607*

En fait, on cherche à vérifier si la répartition des cotes est dépendante des universités c'est-à-dire si les variables "université" et "cote" sont des v.a. indépendantes au niveau 5%

***Solution** : Les hypothèses statistiques sont*

$$H_0 \quad : \quad \textit{les v.a. sont indépendantes}$$
$$H_1 \quad : \quad \textit{les v.a. sont dépendantes}$$

et le test du khi-deux est utilisé. Le test est de rejeter H_0 si

$$\chi^2 = \sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \geq \chi^2_{(m-1)(k-1);\alpha} = \chi^2_{4;0.05} = 9.487\,7$$

On obtient le tableau de contingence suivant :

Cote	*A*	*B*	*C*	*D*	*E*	
Université I	*605*	*1400*	*1789*	*300*	*70*	*4164*
Université II	*2014*	*4178*	*8032*	*2005*	*607*	*16836*
	2619	*5578*	*9821*	*2305*	*677*	*21000*

Les fréquences théoriques sont données par :

Cote	*A*	*B*	*C*	*D*	*E*	
Université I	*519.310*	*1106.038*	*1947.364*	*457.049*	*134.239*	*4164*
Université II	*2099.690*	*4471.962*	*7873.636*	*1847.951*	*542.761*	*16836*
	2619	*5578*	*9821*	*2305*	*677*	*21000*

La statistique observée est

$$\sum\sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = \frac{(605 - 519.310)^2}{519.310} + \ldots$$
$$= 236.808$$

On rejette donc H_0 au niveau 5% et on peut dire qu'il y a dépendance.

Si on utilise le logiciel EXCEL, on obtient les résultats suivants :

OBSERVATIONS

Université	A	B	C	D	E	
I	605	1400	1789	300	70	4164
II	2014	4178	8032	2005	607	16836
	2619	5578	9821	2305	677	

ATTENDUES

Université	A	B	C	D	E	
I	519,310286	1106,03771	1947,364	457,048571	134,239429	4164
II	2099,68971	4471,96229	7873,636	1847,95143	542,760571	16836
	2619	5578	9821	2305	677	21000

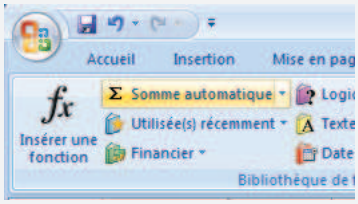
STATISTIQUE DU KHI DEUX	236,808
NIVEAU DE SIGNIFICATION	0,000

On rejette H_0 au niveau 5% et on peut dire que la répartition des cotes dépend de l'université.

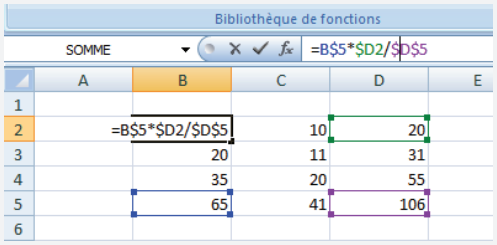
La fonction TEST.KHIDEUX de EXCEL permet d'obtenir le niveau expérimental du test mais il est aussi intéressant d'obtenir la valeur de la statistique. Cette dernière utilise les possibilités de calcul matriciel du logiciel. Voici une procédure détaillée pour obtenir les fréquences théoriques :

Calcul des fréquences attendues par EXCEL

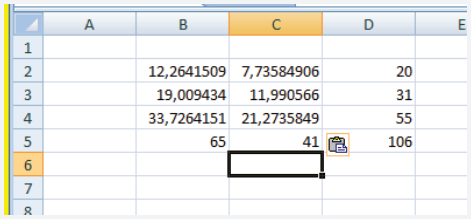
La première étape est de créer les totaux sur les lignes et sur les colonnes avec l'outil somme automatique :



On doit par la suite faire un copier – coller spécial (valeur) pour recréer un tableau des valeurs. On entre alors la formule suivante



et on fait un copier-coller de la formule dans toutes les cellules pour donner le résultat suivant :



On a alors la plage des valeurs attendues nécessaire pour évaluer la valeur de la statistique et pour calculer le seuil de signification empirique.

Test d'homogénéite pour k populations

Considérons une variable sur une échelle nominale qui est mesurée dans k populations. Dans la population j cette variable a une loi de probabilité donnée par $f_j(x_i)$ pour chaque x_i dans le support (commun à toutes les populations). Une question intéressante est de vérifier si ces populations sont régies par la même loi de probabilité. Les hypothèses statistiques sont

$$H_0 \quad : \quad f_j(x_i) = f_l(x_i) \text{ pour tous } j, l \text{ et pour chaque } i$$
$$H_1 \quad : \quad f_j(x_i) \neq f_l(x_i) \text{ pour un certain } (j,l) \text{ et un certain } i$$

c'est-à-dire que les k lois de probabilité sont identiques contre l'hypothèse alternative qu'il y a au moins une loi de probabilité qui est différente des autres.

Ce contexte est assez fréquent comme l'illustre les exemples suivants :

Exemple 12.5 ★*On veut comparer le salaire des femmes et des hommes dans une grande entreprise. On prend un échantillon d'hommes et un échantillon de femmes puis on note le salaire selon "faible", "moyen" et "élevé". Il y a donc 2 populations (hommes et femmes) puis une variable sur une échelle à tout le moins nominale (catégorie de salaire). On veut savoir si les femmes ont un salaire équivalent aux hommes.*

Exemple 12.6 ★*Une université veut vérifier si les cotes qu'elle accorde à ses étudiants sont similaires aux autres universités. On prend un échantillon de 2000 étudiant de l'UdeM, 2000 de McGill, 2000 de l'UQTR et 2000 de l'UQAM. On veut vérifier si la répartition des cotes est semblable entre les universités.*

Exemple 12.7 ★*Un courtier veut vérifier si le rendement des placements (négatif, nul, positif) est lié à la catégorie de l'industrie (mine, services, manufacture, placement). Il collige des échantillons de 40 compagnies dans chacun des secteurs et il établit pour chaque compagnie le type de rendement. On veut savoir si le rendement du placement dépend du type de secteur. Cela revient à comparer les distributions du rendement en fonction du type d'industrie.*

Le test est similaire au test du khi-deux pour deux variables dans une population (test d'indépendance) : la statistique est la même en considérant qu'il y a une variable qui

indique la population mais qu'elle est fixée. On a alors le schéma suivant

$X \backslash$POP	POP1 $\cdots$	POP j $\cdots$	
mod 1	n_{11}	n_{1j}	$n_{1.}$
$\vdots$			
mod i		n_{ij}	$n_{i.}$
$\vdots$			
	$n_{.1}$	$n_{.j}$	n

et le test est de rejeter H_0 si

$$\chi^2 \geq \chi^2_{(k-1)(m-1);\alpha}$$

où

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{(n_{ij} - T_{ij})^2}{T_{ij}}$$

Les valeurs n_{ij} et T_{ij} sont telles que définies dans la section précédante.

Exemple 12.8 ★★*On veut comparer le salaire des femmes et des hommes dans une grande entreprise. On prend un échantillon de 120 hommes et un échantillon de 150 femmes puis on note le salaire selon "faible", "moyen" et "élevé". On observe*

	faible	*moyen*	*élevé*
Homme	*10*	*70*	*40*
Femme	*30*	*60*	*60*

Peut-on dire avec un niveau de 5% que les hommes et les femmes ont un niveau de salaire différent.

Solution : On veut comparer les hypothèses

H_0 : *les distributions sont identiques*

H_1 : *les distribution sont différentes*

avec un niveau de $\alpha = 5\%$.

Le test consiste à rejeter H_0 *si* $\chi^2 \geq \chi^2_{2;\alpha} = 5.99$

On observe $\chi^2 = 11.58$ *et* $\widehat{\alpha}$ *=0,003059747 selon EXCEL. On rejette donc* H_0 *au niveau 5% et on peut dire que les hommes et les femmes sont traités différemment au niveau du salaire.*

Exemple 12.9 ★★*On observe la ville et l'utilisation du transport en commun par un*

sondage dans chaque ville

Transport en commum	*Jamais*	*peu*	*régulièrement*
MTL	*10*	*80*	*120*
TR	*9*	*25*	*8*
Qué	*20*	*40*	*32*

Solution : *On veut comparer les hypothèses*

$$H_0 \quad : \quad \textit{les distributions sont identiques}$$
$$H_1 \quad : \quad \textit{les distribution sont différentes}$$

avec un niveau de $\alpha = 5\%$. *Le test est de rejeter* H_0 *si*

$$\chi^2 \geq \chi^2_{2*2;\alpha} = 9.487\,7$$

Pour calculer la statistique on complète le tableau avec les marginales :

10	*80*	*120*	***210***
9	*25*	*8*	***42***
20	*40*	*32*	***92***
39	**145**	**160**	***344***

Puis on donne le tableau avec les valeurs attendues :

23.808	88.517	97.674	***210***
$4.761\,6$	17.703	19.535	***42***
10.43	38.779	42.791	***92***
39	**145**	**160**	***344***

La statistique du Khi-deux est

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij}-T_{ij})^2}{T_{ij}} \\ &= \frac{(10-23.808)^2}{23.808}+\frac{(80-88.517)^2}{88.517}+\cdots \\ &= 39.064\end{aligned}$$

Comme elle est plus grande que le point critique on rejette H_0 *et on peut conclure que les trois villes sont différentes.*

Remarque 12.3 *Le test sur l'indépendance et sur l'homogénéité sont en fait identique la différence réside dans le fait que dans le test d'indépendance on mesure deux variables tandis que dans le test d'homogénéité on mesure une seule variable mais dans plusieurs populations.*

OBSERVÉES

	Début	Fin	
Favorable	635	327	962
Non favorable	1396	962	2358
	2031	1289	3320

ATTENDUES

	Début	Fin	
Favorable	588,501	373,5	962
Non favorable	1442,5	915,5	2358
	2031	1289	3320

KHI DEUX	13,3238
P-VALUE	0,00026

Test sur deux proportions

Le même test du khi deux peut être utilisé pour vérifier l'égalité de deux proportions (deux échantillons)

$$H_0 \quad : \quad \pi_1 = \pi_2$$
$$H_1 \quad : \quad \pi_1 \neq \pi_2$$

puisque si on pose X la v.a. qui donne la population (1 ou 2) et Y la v.a. qui donne succès ou échec, cela correspond aux hypothèses

$$H_0 \quad : \quad \text{les v.a. sont indépendantes}$$
$$H_1 \quad : \quad \text{les v.a. sont dépendantes}$$

Exemple 12.10 ★★*Lors d'un sondage électoral au début de la compagne, 635 personnes sur 2031 étaient en faveur d'un certain candidat tandis qu'à une semaine des élections, 327 sur 1289 étaient en faveur du même candidat. Peut-on dire à un niveau de 10% que l'opinion a changer ?*

Pour répondre à cette question on veut confronter les hypothèses

$$H_0 \quad : \quad \pi_1 = \pi_2$$
$$H_1 \quad : \quad \pi_1 \neq \pi_2$$

et le test de niveau 10% pour confronter ces hypothèses est de rejeter H_0 *si* $\chi^2 \geq \chi^2_{1;.1} = 2.\,705\,5$. *Or on observe*

	Début	*Fin*
favorable	*635*	*327*
non favorable	1396	962

et la statistique du khi deux telle que calculée par EXCEL est

On rejette H_0 *au niveau 5%. On doit donc conclure que la proportion de personne favorable au candidat a changé au cours de la campagne électorale.*

Remarque 12.4 *Il est possible de faire un test unilatéral en utilisant une statistique du khi deux. Pour faire le test, il faut diviser le niveau de signification par 2 lorsqu'on utilise ce dernier pour prendre la décision ou utiliser le point critique* $\chi^2_{(k-1)(m-1);2\alpha}$

Exemple 12.11 ★★*On pense qu'il y a une proportion plus grande de personnes qui utilisent le transport en commun à Montréal qu'à Toronto. Pour valider cette hypothèse un échantillon de 350 résidents de Toronto est constitué et sur ce nombre il y en a 155 qui utilisent régulièrement le transport en commun. À Montréal, sur un échantillon de 500 personnes il y en a 260 qui utilisent régulièrement le transport en commun.*

Peut-on dire avec un niveau de 10% que les Montréalais utilisent plus le transport en commun qu'à Toronto ?

Solution : *On veut confronter les hypothèses*

$$\begin{aligned} H_0 &: \pi_M = \pi_T \\ H_1 &: \pi_M > \pi_T \end{aligned}$$

où π représente la probabilité d'utiliser régulièrement les transports en commun.

Le test de niveau 10% est de rejeter H_0 si

$$\chi^2 \geq \chi^2_{1;.2} = 1.642\,4$$

Or on observe

Transport	*Montréal*	*Toronto*
Oui	*260*	*155*
non	*240*	*195*

Le tableau des fréquences théoriques est donné par

Transport	*Montréal*	*Toronto*	
Oui	244. 12	170. 88	*415*
non	255. 88	179. 12	*435*
	500	*350*	*850*

et la statistique du khi-deux est

$$\begin{aligned} \chi^2 &= \frac{(260-244.12)^2}{244.12} + \frac{(155-170.88)^2}{170.88} \\ &\quad + \frac{(240-255.88)^2}{255.88} + \frac{(195-179.12)^2}{179.12} \\ &= 4.902\,1 \end{aligned}$$

On rejette H_0 et on peut dire que les montréalais utilisent plus les transports en commun que les torontois au niveau 10%.

Résumé

Les tests du khi-deux permettent de faire des tests d'hypothèses sur des variables qualitatives ou quantitatives avec peu de modalités. Le principe est toujours le même : on compare les fréquences observées et les fréquences théoriques selon les hypothèses.

Pour comparer une distribution à une loi de probabilité les fréquences théoriques sont

données par $T_i = n\pi_{io}$ et le test est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - T_i)^2}{T_i} \geq \chi^2_{k-1;\alpha}$$

Pour comparer la distribution d'une même caractéristique dans plusieurs populations les fréquences théoriques sont données par $T_{ij} = \frac{n_{i.}n_{.j}}{n}$ et le test est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \geq \chi^2_{(k-1)(m-1);\alpha}$$

Pour vérifier si deux variables sont indépendantes les fréquences théoriques sont données par $T_{ij} = \frac{n_{i.}n_{.j}}{n}$ et le test est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \geq \chi^2_{(k-1)(m-1);\alpha}$$

Pour comparer deux proportions les fréquences théoriques sont données par $T_{ij} = \frac{n_{i.}n_{.j}}{n}$ et le test est de rejeter H_0 si

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \geq \chi^2_{1;\alpha}$$

Dans tous les cas il faut vérifier la règle de Cochran : il ne doit pas y avoir de fréquences théoriques plus petites que 1 et au maximun 20% peuvent être plus petites que 5.