9.1 Inference for Two-Way Tables

CHI-SQUARE STATISTIC

The **chi-square statistic** is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts. The formula for the statistic is

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

where "observed" represents an observed cell count, "expected" represents the expected count for the same cell, and the sum is over all $r \times c$ cells in the table.

chi-square distribution χ^2

If the expected counts and the observed counts are very different, a large value of X^2 will result. Large values of X^2 provide evidence against the null hypothesis. To obtain a P-value for the test, we need the sampling distribution of X^2 under the assumption that H_0 (no association between the row and column variables) is true. The distribution is called the **chi-square distribution**, which we denote by χ^2 (χ is the lowercase Greek letter chi).

Le test du khi-deux

Université de Montréal

Puisque z suit approximativement une distribution normale standard, alors la statistique (1), égale à z^2, suit approximativement une distribution normale standard au carré, soit une *distribution khi-deux avec un degré de liberté*.

Définition : Soit $Z_1, Z_2, \ldots, Z_p$ des variables aléatoires i.i.d. normale standard et soit $X = \sum_{i=1}^{p} Z_i^2$. Alors X suit une distribution khi-deux à p degrés de liberté, dénotée par χ_p^2.

Note : Comme une distribution khi-deux à p degrés de liberté est la somme de p variables aléatoires indépendantes et identiquement distribuées, alors une distribution khi-deux ressemble de plus en plus à une distribution normale lorsque p devient grand par le *théorème central limite*.

On veut confronter les hypothèses

$$H_0 : p_1 = p_2 \qquad vs \qquad H_a : p_1 \neq p_2.$$

On calcule la statistique

$$X^2 = z^2 = \frac{(X_1 - n_1\hat{p}_c)^2}{n_1\hat{p}_c} + \frac{((n_1 - X_1) - n_1(1-\hat{p}_c))^2}{n_1(1-\hat{p}_c)} + \frac{(X_2 - n_2\hat{p}_c)^2}{n_2\hat{p}_c} + \frac{((n_2 - X_2) - n_2(1-\hat{p}_c))^2}{n_2(1-\hat{p}_c)},$$

où $\hat{p}_c = (X_1 + X_2)/(n_1 + n_2)$ et on rejette pour de grandes valeurs de X^2.

Note : Puisque $X^2 = z^2$, on perd le *signe* et donc la *direction* de la différence entre les proportions des deux populations. C'est pourquoi l'alternative du test khi-deux est toujours bilatérale.

Il y a **deux façons** d'obtenir des données sous forme d'un tableau de contingence $r \times c$. Premièrement, on peut avoir c *populations différentes* (p.e., différentes catégories d'âges) à partir desquelles nous obtenons des échantillons indépendants et pour lesquels nous allons classifier chaque individu selon *un facteur* qui contient r *catégories* (p.e., différents types de problèmes visuels). Dans ce cas, le nombre d'individus dans chaque échantillon n_i est fixé à l'avance, mais le nombre total d'individus dans chaque catégorie est aléatoire. L'hypothèse d'intérêt est qu'il n'y a *pas de différences dans les proportions entre les différentes populations.*

Dans le deuxième cas, n est la taille fixe de l'échantillon pris à partir *d'une seule et unique population.* Sauf que cette fois-ci, on classifie chaque individu selon *deux facteurs contenant r et c catégories*, respectivement (p.e., le statut socio-économique, faible, moyen, élevé et le type de fumeur, présentement, dans le passé, jamais). L'hypothèse d'intérêt est que les *deux facteurs sont indépendants.*

En d'autres mots, par exemple, la probabilité qu'un individu soit d'un statut socio-économique faible **et** soit présentement fumeur est le **produit** de la probabilité qu'un individu soit d'un statut socio-économique faible et celle qu'un individu soit présentement fumeur. Ainsi il n'y aurait *pas d'association* entre le statut socio-économique et le type de fumeur.

Le livre ne fait pas la distinction entre ces deux **modèles** et parle dans les deux cas de **l'hypothèse d'absence d'association.**

Premier modèle : Homogénéité des proportions

Université de Montréal

On choisit des EAS indépendants de taille $n_1, n_2, \ldots, n_c$ de c populations différentes. Chaque individu est classifié selon l'une de r catégories (p.e., type de problèmes visuels). Soit $X_{i(j)}$ le nombre d'individus classifiés dans la $i^{\text{ème}}$ catégorie pour le $j^{\text{ème}}$ échantillon. Définissons $n = \sum_{i=1}^{c} n_i$. Ainsi le tableau est de la forme

	Colonnes (échantillons)				
Rangées (catégories)	1	2	$\cdots$	c	
1	$X_{1(1)}$	$X_{1(2)}$	$\cdots$	$X_{1(c)}$	
2	$X_{2(1)}$	$X_{2(2)}$	$\cdots$	$X_{2(c)}$	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	
r	$X_{r(1)}$	$X_{r(2)}$	$\cdots$	$X_{r(c)}$	
Total	n_1	n_2	$\cdots$	n_c	n

On veut tester l'hypothèse qu'il n'y a pas de différences entre les différentes populations.

Soit $p_{i(j)}$, la probabilité d'être classifié dans la $i^{\text{ème}}$ catégorie pour la $j^{\text{ème}}$ population. L'hypothèse nulle s'exprime ainsi

$$\begin{aligned} H_0 \quad : \quad & p_{1(1)} = p_{1(2)} = \cdots = p_{1(c)} = p_1 \\ & p_{2(1)} = p_{2(2)} = \cdots = p_{2(c)} = p_2 \\ & \vdots \\ & p_{r(1)} = p_{r(2)} = \cdots = p_{r(c)} = p_r \\ vs\ H_a \quad : \quad & H_0^C \end{aligned}$$

Sous H_0, il y a r paramètres (inconnus) : $p_1, p_2, \ldots, p_r$. En fait, il n'y en a que $r-1$ puisque $\sum_{i=1}^{r} p_i = 1$. Comment doit-on les estimer ?

Prenons p_1. On pourrait traiter le problème comme s'il y avait deux catégories seulement : la première et les autres mises ensemble. À ce moment-là, sous l'hypothèse nulle, nous aurions $X_{1(1)}$ succès parmi n_1 essais avec probabilité de succès p_1 dans la première population, $X_{1(2)}$ succès parmi n_2 essais avec probabilité de succès p_1 dans la seconde population, ainsi de suite jusqu'à $X_{1(c)}$ succès parmi n_c essais tous avec probabilité de succès p_1 dans la population c.

Ainsi, sous l'hypothèse nulle, $\sum_{j=1}^{c} X_{1(j)}$ est distribué selon une loi $B(n, p_1)$ et l'on obtient

$$\hat{p}_1 = \frac{\sum_{j=1}^{c} X_{1(j)}}{n}.$$

De façon générale, sous H_0, $\sum_{j=1}^{c} X_{i(j)}$ est distribué selon une loi $B(n,p_i)$ et

$$\hat{p}_i = \frac{\sum_{j=1}^{c} X_{i(j)}}{n}.$$

Si l'hypothèse nulle est vraie, le *nombre espéré* de personnes dans la *cellule* i,j est $n_j p_i$ qui est *estimé* par $n_j \hat{p}_i$.

On a donc des nombres de personnes observés $X_{i(j)}$ et des nombres espérés $n_j \hat{p}_i$. Si H_0 est vraie, ces nombres devraient être proches.

On utilise la statistique du *khi-deux de Pearson* pour mesurer la distance entre les tableaux observé et espéré :

$$\begin{aligned} X^2 &= \sum \frac{(\text{Observé - Espéré})^2}{\text{Espéré}} \\ &= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{i(j)} - n_j \hat{p}_i)^2}{n_j \hat{p}_i}. \end{aligned}$$

Si l'hypothèse nulle est vraie, la statistique X^2 suit approximativement une distribution χ^2 avec $(r-1)(c-1)$ degrés de liberté. La valeur-p approximative du test est

$$P(\chi^2 \geq x^2)$$

où χ^2 est une variable aléatoire avec une distribution khi-deux à $(r-1)(c-1)$ degrés de liberté et x^2 est la valeur observée dans notre échantillon de la statistique X^2.

Note : Si H_0 n'est pas vraie, les valeurs observées et espérées seront différentes et X^2 aura tendance à être grand. C'est pourquoi la valeur-p calcule la probabilité pour des grandes valeurs de X^2 seulement et non pour les petites valeurs aussi.

Second modèle : Indépendance de deux facteurs

Université de Montréal

Cette fois, nous avons un EAS de taille n et l'on classifie les personnes selon deux facteurs, par exemple le sexe et le but poursuivi (avoir de bonnes notes, être populaire ou réussir dans les sports). Donc une seule population (les étudiants) plutôt que c (les étudiants de sexe masculin et féminin).

Soit X_{ij} le nombre d'individus dans l'échantillon qui sont dans la catégorie i du premier facteur et la catégorie j du second facteur.

Le tableau devient

	Facteur 2				
Facteur 1	1	2	$\cdots$	c	Total
1	X_{11}	X_{12}	$\cdots$	X_{1c}	r_1
2	X_{21}	X_{22}	$\cdots$	X_{2c}	r_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	r_r
Total	c_1	c_2	$\cdots$	c_c	n

où les r_i et les c_j sont les totaux (aléatoires) des rangées et des colonnes, respectivement. Donc $n = \sum_{i=1}^{r} r_i = \sum_{j=1}^{c} c_j$.

Soit p_{ij}, la probabilité d'être classifié dans la catégorie i du facteur 1 et la catégorie j du facteur 2 et soit $p_{i\cdot} = \sum_{j=1}^{c} p_{ij}$ et $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$, alors les $p_{i\cdot}$ et les $p_{\cdot j}$ représentent les *probabilités marginales* pour les facteurs 1 et 2, respectivement, par exemple les probabilités d'être un gars ou une fille d'une part et les probabilités de vouloir avoir de bonnes notes, être populaire ou réussir dans les sports, d'autre part.

L'hypothèse d'intérêt est *l'indépendance* des deux facteurs. Sous cette hypothèse,

$$\begin{aligned} p_{ij} &= P(\text{cat } i \text{ pour facteur 1 et cat } j \text{ pour facteur 2}) \\ &= P(\text{cat } i \text{ pour facteur 1})P(\text{cat } j \text{ pour facteur 2}) \\ &= p_{i\cdot}p_{\cdot j}. \end{aligned}$$

Donc

$$H_0 \quad : \quad p_{ij} = p_{i\cdot}p_{\cdot j} \qquad i = 1,\ldots,r, \quad j = 1,\ldots,c$$
$$vs\ H_a \quad : \quad H_0^C.$$

Il faut estimer les paramètres $p_{i\cdot}$ et $p_{\cdot j}$ sous l'hypothèse nulle. On le fait de la façon suivante

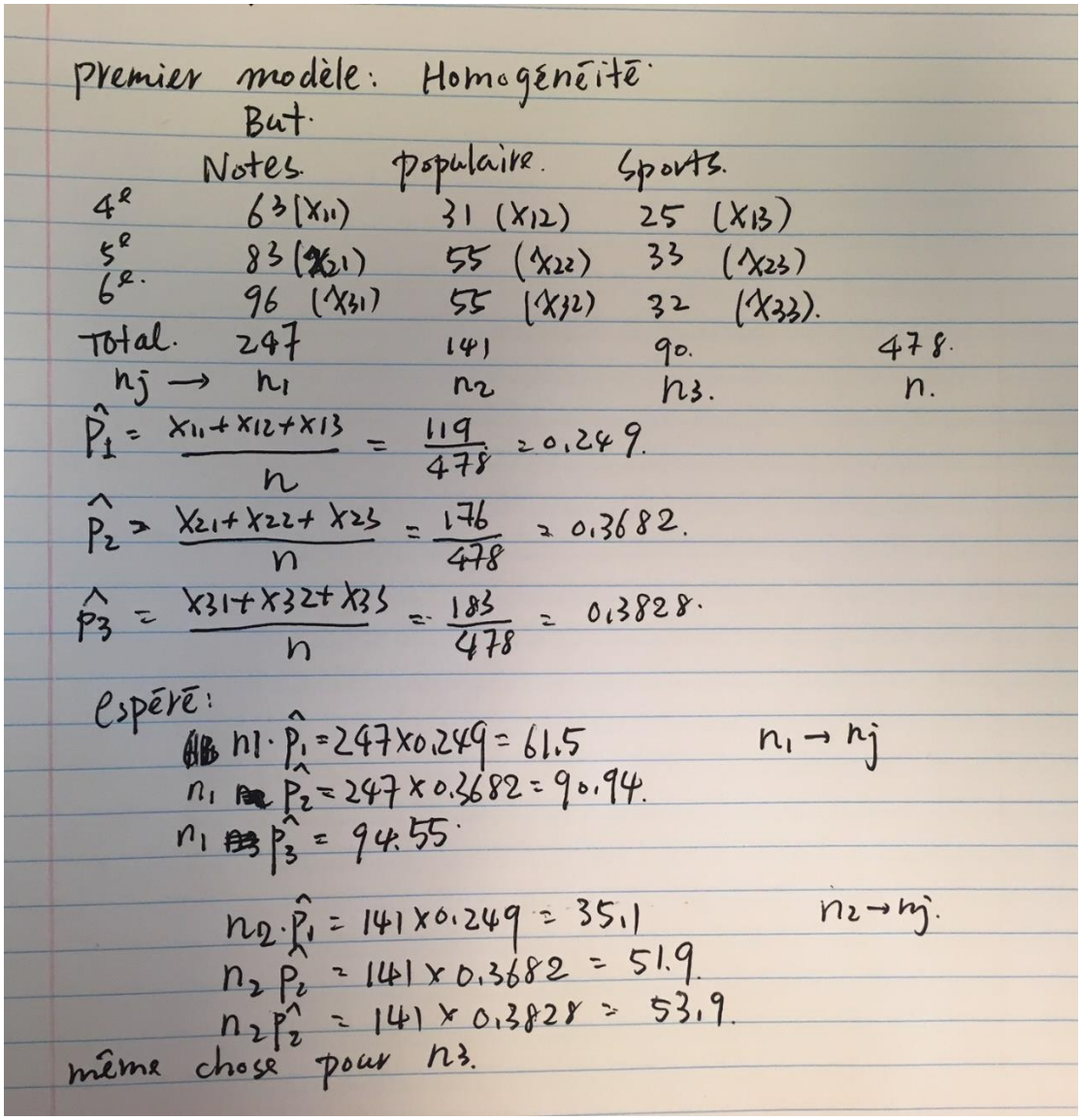
$$\hat{p}_{i\cdot} = \frac{r_i}{n}, \qquad \hat{p}_{\cdot j} = \frac{c_j}{n}.$$

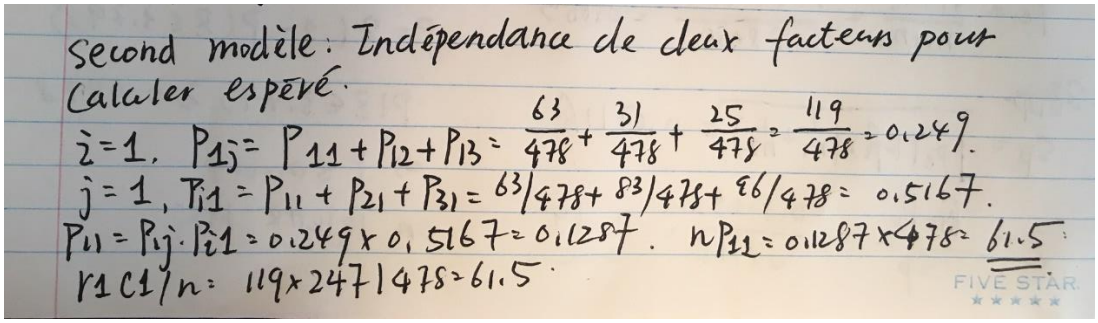
Si H_0 est vraie, le nombre espéré de personnes dans la cellule i,j est $np_{ij} = np_{i\cdot}p_{\cdot j}$ qui est estimé par $n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = r_i c_j/n$.

Note : Le nombre espéré pour ce modèle est le même que pour le modèle précédent (à vérifier). Le test X^2 est donc le même que pour le modèle précédent. Le nombre espéré selon les deux modèles est donc le produit des marginales respectives divisé par le nombre total d'observations.

Exemple:

	Observé But			
Année	Notes	Populaire	Sports	Total
4^e	63	31	25	119
5^e	88	55	33	176
6^e	96	55	32	183
Total	247	141	90	478





	Espéré But			
Année	Notes	Populaire	Sports	Total
4^e	61,5	35,1	22,4	119
5^e	90,9	51,9	33,1	176
6^e	94,6	54,0	34,5	183
Total	247	141	90	478

Note : Le nombre espéré pour ce modèle est le même que pour le modèle précédent (à vérifier). Le test X^2 est donc le même que pour le modèle précédent. Le nombre espéré selon les deux modèles est donc le produit des marginales respectives divisé par le nombre total d'observations.

Note : dl = # paramètres indépendants –# paramètres estimés. Dans ce cas $dl = (rc - 1) - (r - 1) - (c - 1)$ puisque $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$ et $\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{c} p_{\cdot j} = 1$. Donc $dl = rc - 1 - r + 1 - c + 1 = (r - 1)(c - 1)$ comme c'était le cas dans le modèle précédent.

Test d'adéquation à une distribution qualitative

Université de Montréal

Loto Québec utilise des bouliers mécaniques pour ses tirages. Est-ce que chaque numéro entre 0 et 9 a la même probabilité (1/10) d'être tiré ?

Est-ce que le fait d'être né au début de l'année (janvier, février,...) plutôt qu'à la fin de l'année (..., novembre, décembre) a un impact sur le fait de devenir un joueur de hockey professionnel ? En d'autres mots, si la période de l'année n'a aucun impact sur le fait de devenir un hockeyeur professionnel alors la probabilité d'être né en janvier devrait être de 31/365, en février de 28/365, etc. Est-ce le cas ?

On obtient un échantillon de n observations indépendantes et identiquement distribuées d'un tel événement aléatoire. Chaque observation tombe dans l'une de k catégories. En tout, il y a X_i observations dans la $i^{\text{ième}}$ catégorie (le livre parle de n_i). L'hypothèse nulle qu'on souhaite tester est que la probabilité d'être dans la $i^{\text{ième}}$ catégorie est p_i, pour $i = 1, \ldots, k$, où p_i est une constante connue.

Test d'adéquation à une distribution qualitative

Université de Montréal

Les hypothèses à confronter sont :

$$H_0 \ : \ Prob(\text{Catégorie } i) = p_i \qquad vs \qquad H_a \ : \ H_0^C.$$

Le nombre espéré d'observations dans la $i^{\text{ième}}$ catégorie sous H_0 est np_i.

La statistique de test est

$$\begin{aligned} X^2 &= \sum \frac{(\text{Observé - Espéré})^2}{\text{Espéré}} \\ &= \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}. \end{aligned}$$

Sous H_0, la statistique X^2 suit (asymptotiquement) une distribution χ^2 à $k-1$ degrés de liberté.

Note : Pour que l'approximation par la distribution χ^2 soit bonne, il faut que le nombre espéré soit d'au moins 5 dans chaque catégorie.