Université Badji Mokhtar - Annaba
FACULTE DES SCIENCES

عنابـــــــة -جامعـــــــــة باجـــــي مختـــــــار
كليـــــــــة العلـــــــــوم

Cours (4) Biostatistique
Comparaison de distributions de variables qualitatives

Dr.SAADOUN AHMED
saadounahmed1@yahoo.fr

I-Comparaison de distributions de variables qualitatives

On s’intéresse a la variable ≪ couleur des cheveux ≫ mesurée sur une échelle qualitative avec les $J = 5$ valeurs possibles suivantes : ≪ blond ≫, ≪ roux ≫, ≪ châtain ≫, ≪ brun ≫ et ≪ noir ≫. On veut comparer cette variable qualitative entre $I = 4$ populations d’écoliers qui ont respectivement les yeux bleus, les yeux verts, les yeux bruns et les yeux noirs, le but étant de montrer qu’il existe un lien entre la couleur des cheveux et la couleur des yeux, autrement dit que la distribution de la couleurs des cheveux diffère d’une population a l’autre. On essaiera donc de rejeter l’hypothèse nulle suivante :

$H0$: la distribution de la variable ≪ couleur des cheveux ≫ est la même dans les 4 populations aux ≪ couleurs des yeux ≫ différentes.

Pour ce faire, on utilise les données récoltées auprès de 4 échantillons représentatifs de ces populations d'écoliers, qui sont résumées dans la table de contingence de dimension 4×5 suivante (contenant les *fréquences observées*) :

couleur yeux	couleur cheveux					
	blond	roux	châtain	brun	noir	total
bleu	326	38	241	110	3	718
	(45 %)	(5 %)	(34 %)	(15 %)	(0 %)	(100 %)
vert	688	116	584	188	4	1580
	(44 %)	(7 %)	(37 %)	(12 %)	(0 %)	(100 %)
brun	343	84	909	412	26	1774
	(19 %)	(5 %)	(51 %)	(23 %)	(1 %)	(100 %)
noir	98	48	403	681	85	1315
	(7 %)	(4 %)	(31 %)	(52 %)	(6 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

On a vu que la distribution d'une variable qualitative avec J valeurs possibles peut être caractérisée par J proportions (en fait $J - 1$ proportions sont suffisantes).
Dans cette table, on a donc calcule ces $J = 5$ proportions pour chacun des $I = 4$ échantillons.

La dernière ligne de la table nous donne en outre la ≪ distribution moyenne ≫ de la variable ≪ couleur des cheveux ≫ calculée sur le total des observations (on a ici 27 % de cheveux blonds, 5 % de cheveux roux, 40 % de cheveux châtains, 26 % de cheveux bruns, 2 % de cheveux noirs). A première vue, ces données semblent effectivement contredire l'hypothèse nulle. La distribution de la couleur des cheveux est par exemple très différente entre les écoliers aux yeux bleus (majorité de cheveux blonds), les écoliers aux yeux bruns (majorité de cheveux châtains) et les écoliers aux yeux noirs (majorité de cheveux bruns). Rappelons toutefois qu'a cause du hasard de l'échantillonnage, on observera inévitablement certaines différences entre ces quatre distributions, et ceci même si l'hypothèse nulle était vraie. Par contre, de trop grosses différences seront suspectes.

Toute la question est donc de déterminer si les différences observées entre ces quatre distributions sont suspectes ou non, autrement dit si elles peuvent être mises sur le compte du hasard de l'échantillonnage (auquel cas on ne rejettera pas $H0$) ou non (auquel cas on rejettera $H0$). Pour répondre a cette question, on utilise une statistique de test. Rappelons qu'une statistique de test est une mesure de la distance entre les données et l'hypothèse nulle dont on connait (approximativement) la distribution sous $H0$. Afin de définir une telle distance, on notera tout d'abord que si nos données respectaient a la lettre $H0$, on aurait la même distribution de la couleur des cheveux dans chacun de ces 4 échantillons, ces 4 distributions étant alors égales a la distribution moyenne (calculée dans la dernière ligne de la table ci-dessus).
On aurait alors la table de contingence suivante (contenant les *fréquences attendues* sous l'hypothèse nulle) :

couleur yeux	couleur cheveux					
	blond	roux	châtain	brun	noir	total
bleu	193.9	38.1	284.8	185.4	15.7	718
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
vert	426.7	83.9)	626.8	408.0	34.6	1580
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
brun	479.1	94.2	703.7	458.1	38.9	1774
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
noir	355.2	69.8	521.7	339.6	28.8	1315
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

Les fréquences attendues pour la première ligne de cette table s'obtiennent par exemple de la manière suivante : 718 *x* 1455/5387 = 193.9 cheveux blonds, 718 *x* 286/5387 = 38.1 cheveux roux, 718 *x* 2137/5387 = 284.8 cheveux châtains, 718 *x* 1391/5387 = 185.4 cheveux bruns et 718 *x* 118/5387 = 15.7 cheveux noirs.

Les calculs sont analogues pour les autres lignes de cette table. La statistique de test sera une mesure de la distance globale entre les fréquences observées et les fréquences attendues. Elle se calcule en sommant les 4×5 termes suivants :

$$
\begin{aligned}
t_{stat} &= \frac{(326-193.9)^2}{193.9} + \frac{(38-38.1)^2}{38.1} + \frac{(241-284.8)^2}{284.8} + \frac{(110-185.4)^2}{185.4} + \frac{(3-15.7)^2}{15.7} \\
&\quad + \frac{(688-426.7)^2}{426.7} + \frac{(116-83.9)^2}{83.9} + \frac{(584-626.8)^2}{626.8} + \frac{(188-408-0)^2}{408.0} + \frac{(4-34.6)^2}{34.6} \\
&\quad + \frac{(343-479.1)^2}{479.1} + \frac{(84-94.2)^2}{94.2} + \frac{(909-703.7)^2}{703.7} + \frac{(412-458.1)^2}{458.1} + \frac{(26-38.9)^2}{38.9} \\
&\quad + \frac{(98-355.2)^2}{355.2} + \frac{(48-69.8)^2}{69.8} + \frac{(403-521.7)^2}{521.7} + \frac{(681-339.6)^2}{339.6} + \frac{(85-28.8)^2}{28.8} \\
&= 1240.0
\end{aligned}
$$

Au-dela de notre exemple particulier, si on dénote par O_i les fréquences observées ($O_i$désignant le nombre de fois que l'on observe la valeur possible j dans le groupe i, pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$), les fréquences attendues E_i sont définies comme suit :

$$E_{ij} = \frac{\sum_{i=1}^{J} O_{ij} \times \sum_{i=1}^{I} O_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} O_{ij}}$$

et la statistique de test est définie comme suit :

$$T_{stat} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

On rappellera ici la distinction de notation entre le concept de la variable aléatoire T_{stat} et sa réalisation t_{stat} dans notre échantillon. Pour un test du *khi-deux* applique a une table de contingence, les quatre étapes d'un test statistique se présentent alors de la manière suivante :

• la première étape consiste a définir une statistique de test T_{stat} comme on vient de le faire ci-dessus.

• la deuxième étape consiste a établir mathématiquement la distribution de cette statistique de test sous l'hypothèse nulle ; il se trouve que la distribution de T_{stat} sous $H0$ est ici (approximativement) une *distribution du khi-deux* avec $(I-1)(J-1)$ *dl*, l'approximation étant bonne si la majorité (par exemple 80 %) des fréquences attendues sont supérieures a 5, auquel cas le test du khi-deux est dit valide (dans notre exemple, il s'agit donc de $(4-1)(5-1) = 12$ *dl*)

• la troisième étape consiste a calculer la réalisation t_{stat} de la variable aléatoire T_{stat} dans notre échantillon (dans notre exemple $t_{stat} = 1240.0$)

• la quatrième étape consiste a comparer la valeur observée (dans notre exemple $t_{stat} = 1240.0$) avec la distribution théorique de T_{stat} sous $H0$ (dans notre exemple, une distribution du khi-deux avec 12 *dl*), l'idée étant de rejeter hypothèse nulle si la statistique de test observée t_{stat} est incompatible (trop grande) par rapport a la distribution théorique.

En ce qui concerne cette quatrième étape, on adopte la règle de rejet suivante :

$$\text{On rejette } H_0 \text{ au seuil } \alpha \text{ si } t_{stat} \geq \chi^2_{1-\alpha,(I-1)(J-1)}.$$

Le graphique du haut de la figure 7.1 nous montre la région de rejet au seuil de 5 % lorsque *dl* = 12. Dans notre exemple, on rejette *H*0 au seuil de 5 % car t_{stat} = 1240.0 est beaucoup plus grand que $\chi^2_{0.95,12}$= 21.03. On a donc réussi a prouver statistiquement notre hypothèse scientifique, a savoir que la distribution de la couleur des cheveux n'est pas la même dans ces quatre populations (et donc qu'il y a un lien entre couleur des cheveux et couleur des yeux).

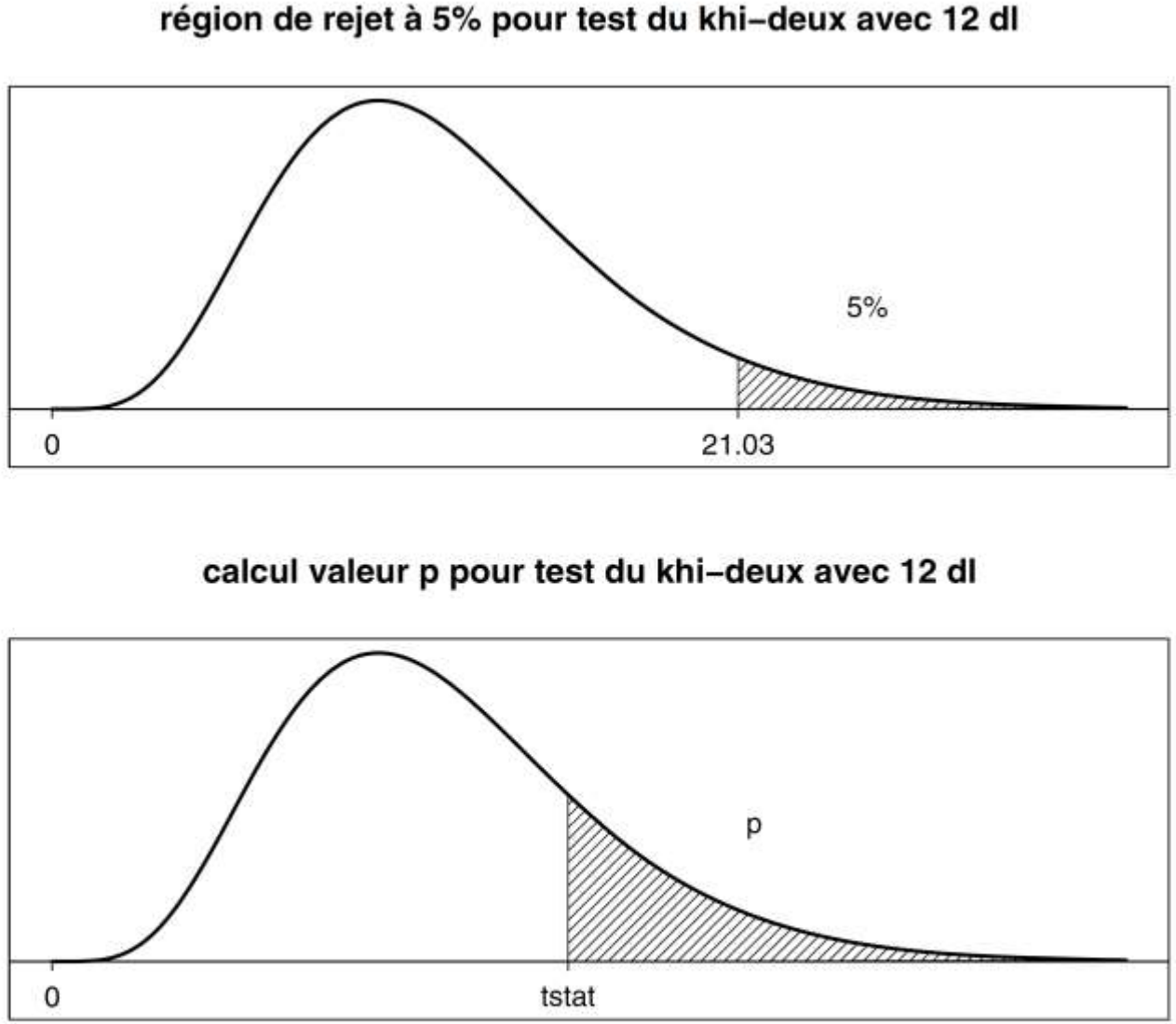

Figure 7.1 – Région de rejet à 5 % et calcul de la valeur p pour test du χ^2 avec 12 dl.

II-Comparaison d’une distribution qualitative avec distribution de référence

II-Comparaison d’une distribution qualitative avec distribution de référence

Le test du khi-deux vu dans la section précédente s’utilise notamment lorsqu’il s’agit de comparer la distribution d’une variable qualitative entre deux groupes. On va voir dans cette section un test du khi-deux qui nous permet

de comparer la distribution d’une variable qualitative avec une distribution de référence, donnée par exemple dans la littérature.

Nous reprenons pour cela l’exemple introduit au chapitre 2 a propos des groupes sanguins récoltes auprès de $n = 158$ individus d’une population du nord de l’Europe. Nous avions les fréquences observées suivantes :

groupe O	groupe A	groupe B	groupe AB	total
82	62	10	4	158
52 %	39 %	6 %	3 %	100 %

L'hypothèse scientifique d'une telle étude sera par exemple que la distribution des groupes sanguins dans cette population du nord de l'Europe diffère de celle de la France, ou l'on a (selon la littérature) 43 % de groupe 0, 45 % de groupe A, 9 % de groupe B et 3 % de groupe AB. On essaiera alors de rejeter l'hypothèse nulle suivante :

*H*0 : il y a 43 % de groupe O, 45 % de groupe A, 9 % de groupe B et 3 % de groupe AB.

Pour ce faire, on utilise une idée similaire a celle vue dans la section précédente.

Si nos données respectaient a la lettre l'hypothèse nulle, les fréquences attendues sous l'hypothèse nulle seraient les suivantes :

couleurs des yeux différentes revient a dire que la distribution de la couleur des yeux est la même dans des populations aux couleurs des cheveux différentes (et *vice versa*). Ces deux hypothèses nulles sont également équivalentes a l'hypothèse nulle suivante :

*H*0 : les variables ≪ couleur des cheveux ≫ et ≪ couleur des yeux ≫ sont indépendantes.

Cette hypothèse nulle nous dit que le fait de connaitre la couleur des yeux d'un écolier ne nous informe en rien sur sa couleur des cheveux, de même que le fait de connaitre la couleur des cheveux d'un écolier ne nous informe en rien sur sa couleur des yeux. Ainsi, le test du khi-deux que l'on voit dans cette section est parfois appelé ≪ test d'indépendance pour variables qualitatives ≫. Notons par ailleurs que ces trois hypothèses nulles nous suggèrent trois façons différentes d'échantillonner les données.

La première nous suggère d'échantillonner parmi 4 populations d'écoliers aux couleurs des yeux différentes et de mesurer la couleur des cheveux ; la deuxième d'échantillonner parmi 5 populations d'écoliers aux couleurs des cheveux différentes et de mesurer la couleur des yeux ; la troisième d'échantillonner parmi une population d'écoliers et de mesurer a la fois la couleur des cheveux et la couleur des yeux (ce qui a été fait dans cet exemple).
L'équivalence de ces trois hypothèses nulles implique que l'on peut appliquer ce même test du khi-deux, quelle que soit la façon dont on a échantillonne les données.

groupe O	groupe A	groupe B	groupe AB	total
67.9	71.1	14.2	4.7	158
43 %	45 %	9 %	3 %	100 %

Pour juger si les différences entre fréquences observées et fréquences attendues peuvent être mises sur le compte du hasard de l'échantillonnage, on calcule la statistique de test suivante :

$$t_{stat} = \frac{(82 - 67.9)^2}{67.9} + \frac{(62 - 71.1)^2}{71.1} + \frac{(10 - 14.2)^2}{14.2} + \frac{(4 - 4.7)^2}{4.7} = 5.44.$$

Plus généralement, si on a une variable qualitative avec *I* valeurs possibles et que l'on dénote par O_i les fréquences observées (le nombre de fois que l'on observe la valeur possible *i* pour $i = 1, \ldots, I$), et par E_i les fréquences attendues, la statistique de test est définie comme suit :

$$T_{stat} = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i}$$

On peut ici établir mathématiquement que si l'hypothèse nulle était vraie, cette statistique de test aurait (approximativement) une distribution du *khi-deux* avec $I-1$ *dl*. Cette approximation sera bonne si les fréquences attendues sont supérieures a 5, auquel cas le test du khi-deux est dit valide (notons que cette condition n'est pas tout a fait satisfaite dans notre exemple, ou l'on a une fréquence attendue de 4.7). On adopte alors la règle de rejet suivante :

$$\text{On rejette } H_0 \text{ au seuil } \alpha \text{ si } t_{stat} \geq \chi^2_{1-\alpha,I-1}.$$

La valeur *p* se calcule de manière analogue a ce que l'on a vu dans la section précédente, c'est-a-dire que l'on aura $t_{stat} = \chi^2_{1-p,I-1}$. Dans notre exemple, t_{stat} = 5.44 est plus petit que $\chi^2_{0.95,3}$ = 7.81, de sorte que la distance entre fréquences observées et fréquences attendues n'est pas assez grande pour rejeter l'hypothèse nulle au seuil de 5 % (un logiciel statistique nous donnera *p* = 0.14 ; autrement dit, la valeur t_{stat} = 5.44 correspond au quantile 86 % d'une distribution du khi-deux avec 3 *dl*) 4. Il se pourrait donc que la distribution des groupes sanguins dans cette population du nord de l'Europe soit identique a celle que l'on a en France (du moins, on n'a pas réussi a prouver le contraire).

Merci pour votre attention