Loi du χ^2, test d'ajustement et d'indépendance

LI323

Hugues Richard
(notes de cours: Pierre-Henri Wuillemin)

Université Pierre et Marie Curie (UPMC)
Laboratoire génomique des microorganismes (LGM)

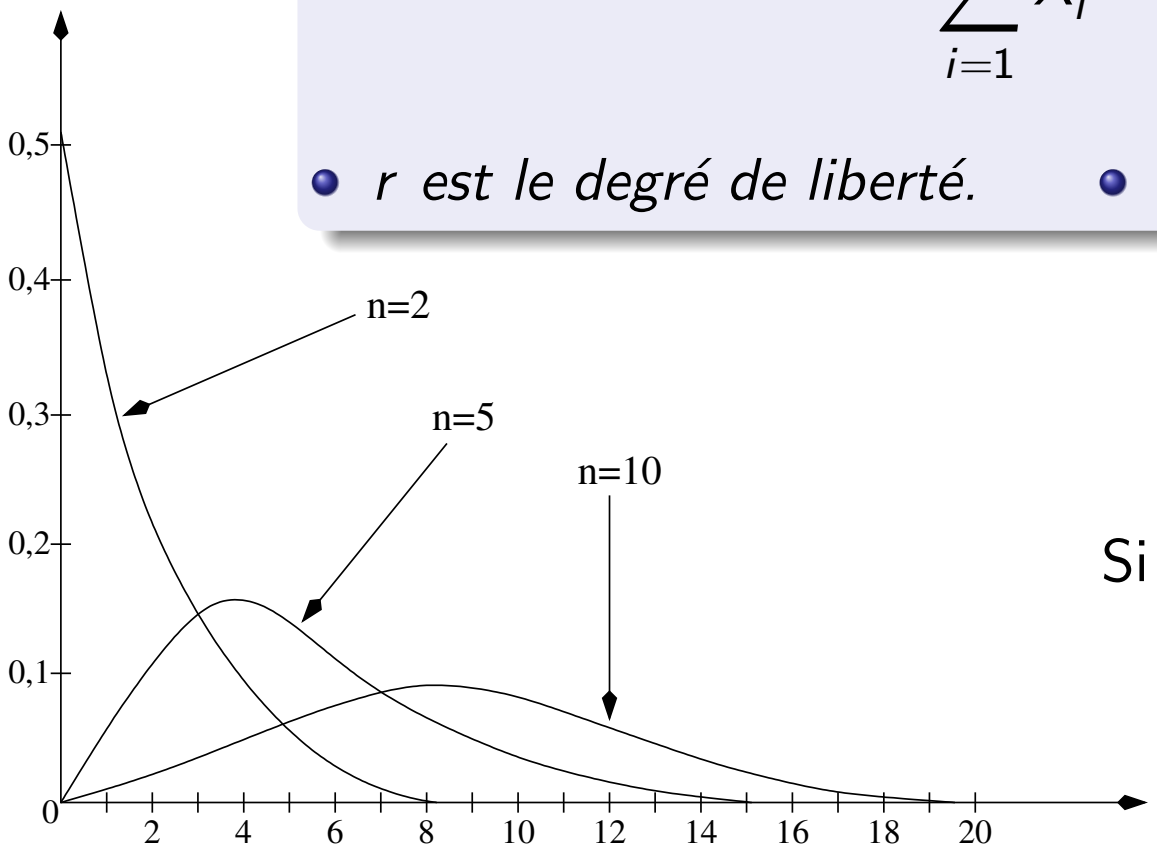
Loi du χ^2

On se souvient que si $(X_i)_{i\in\{1,\cdots,r\}}$ (i.i.d) $\sim \mathcal{N}(0;1)$ alors $\sum_{i=1}^{r} X_i \sim \mathcal{N}(0;n)$

Définition (Loi du χ^2)

$$\sum_{i=1}^{r} X_i^2 \sim \chi^2_{(r)}$$

- *r est le degré de liberté.*
- *Moyenne$= r$ et variance$= 2r$*

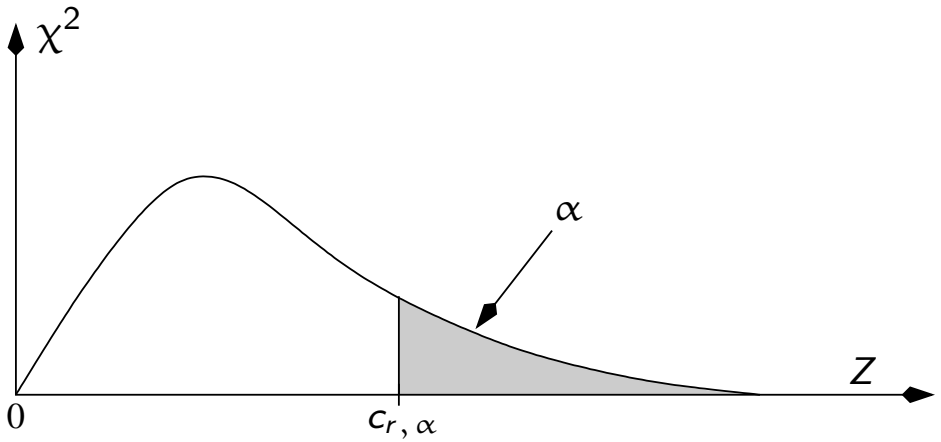


Si $r > 100$ alors

$$\chi^2_{(r)} \approx \mathcal{N}(r;2r).$$

Table du χ^2

Soit $Z \sim \chi^2_{(r)}$ alors on note $c_{r,\alpha}$ telle que $P(Z \geq c_{r,\alpha}) = \alpha$.



$n \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	0,0000393	0,000157	0,000982	0,00393	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9
5	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7
6	0,676	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8	18,5
7	0,989	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5	20,3
8	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22,0
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6
10	2,16	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2	25,2
11	2,60	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8
12	3,07	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2	28,3
13	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8
14	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3
15	4,60	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6	32,8

Intervalle de confiance pour σ^2

Soit $(X_i)_{i\in\{1,\cdots,n\}}$ (i.i.d) $\sim \mathcal{N}(\mu, \sigma^2)$,

On sait que $\overline{X} \sim \mathcal{N}(\mu; \sigma^2/n)$

Définition (Loi de la variance corrigée (Cochran))

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sim \sigma^2 \cdot \chi^2_{(n-1)}$$

D'où, comme $S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$

Soit X une variable aléatoire suivant une loi $\mathcal{N}(\mu; \sigma^2)$. Soit s^2 la variance corrigée observée sur un échantillon de taille n. Alors un intervalle de confiance pour σ^2 de niveau de confiance $1-\alpha$ est donné par :

$$\left[\frac{(n-1)s^2}{c_{(n-1),\frac{\alpha}{2}}}; \frac{(n-1)s^2}{c_{(n-1),1-\frac{\alpha}{2}}}\right].$$

Intervalle de confiance pour σ^2

Calcul

Un distributeur de boissons est réglé de telle sorte que la quantité X de liquide qu'il verse dans un gobelet est distribuée selon une loi normale. Afin d'affiner les réglages, un technicien prélève un échantillon de 10 boissons. Il obtient sur cet échantillon une moyenne de 20 cl avec un écart-type de 1,65 cl. Il veut estimer avec un risque d'erreur de 5% la variance de la quantité de boisson versée par le distributeur.

L'intervalle de confiance est :

$$\left[\frac{9 \times 1,65^2}{c_{9;0,025}}; \frac{9 \times 1,65^2}{c_{9;0,975}}\right].$$

$n \setminus \alpha$	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
9	1,73	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7	23,6

D'après la table de la loi du χ^2, on obtient l'intervalle :

$$\left[\frac{24,5025}{19}; \frac{24,5025}{2,7}\right] = [1,29; 9,07].$$

Test d'hypothèse

Forme de la zone critique pour $H_0 : \sigma^2 = \sigma_0^2$

Contre-hypothèse	forme de la région critique
$H_1 : \sigma^2 < \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;1-\alpha}$
$H_1 : \sigma^2 > \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} > c_{n-1;\alpha}$
$H_1 : \sigma^2 \neq \sigma_0^2$	$\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;1-\alpha/2}$ ou $\frac{(n-1)s^2}{\sigma_0^2} < c_{n-1;\alpha/2}$

Approximation de χ^2 par une loi normale

Lorsque n est supérieur à 30, on a :

$$c_{n;\alpha} \approx \frac{1}{2}\left[z_\alpha + \sqrt{2n-1}\right]^2,$$

où z_α est le quantile d'ordre $1-\alpha$ d'une variable aléatoire Z suivant une

Test d'hypothèse

Le montant de l'assurance pour une automobile utilisée à des fins non commerciales est fonction de la distance moyenne parcourue annuellement par les automobilistes. Des études datant de quelques années ont montré que celle-ci était de 18000 km avec un écart-type de 5000 km. On suppose que la variable *distance parcourue par les automobilistes* suit une loi normale. Le statisticien se demande si l'écart-type doit être revu à la baisse. Pour cela, il a prélevé un échantillon de 400 individus et il a obtenu une variance corrigée égale à $(4700 \text{ km})^2$.
Les hypothèses qu'il formule sont donc les suivantes :

$$H_0 : \sigma^2 = 5000^2 \qquad \text{versus} \qquad H_1 : \sigma^2 < 5000^2$$

Région critique :

$$\frac{(400-1) \times 4700^2}{5000^2} < c_{400-1;1-\alpha}.$$

Puisque $n > 30$, on utilise l'approximation de c :

$$\frac{399 \times 4700^2}{5000^2} < \frac{1}{2}\left[z_{1-\alpha} + \sqrt{2 \times 400 - 1}\right]^2.$$

Si notre statisticien veut un niveau de test égal à 95%, il obtient :

$$\frac{399 \times 4700^2}{5000^2} = 362,5564 \qquad \text{et} \qquad \frac{1}{2}\left[z_{1-\alpha} + \sqrt{2 \times 400 - 1}\right]^2 = \frac{1}{2}\left[1,645 + \sqrt{799}\right]^2 = 447,351.$$

Le statisticien peut en conclure (avec un taux d'erreur de 5%) que la variance a effectivement

Tests d'ajustement

Définition

Test d'ajustement

- *test d'ajustement = test ayant pour issue l'acceptation ou le rejet de l'hypothèse que l'échantillon observé est tiré d'une certaine loi.*
- *contre-hypothèse : ne précise pas de quelle autre loi il aurait pu être tiré.*

- population $\Longrightarrow$ répartie en k classes.
- hypothèse : répartition dans les classes connues.
  $\Longrightarrow p_l$ = proba qu'un individu appartienne à la classe l.
- on tire au hasard des individus dans la population.
- N_l = variable aléatoire «nombre d'individus tirés de classe l»
- soit $D^2_{(n)} = \sum_{l=1}^{k} \frac{(N_l - n.p_l)^2}{n.p_l}$.
  $D^2_{(n)}$ = écart entre théorie et observation
- $D^2_{(n)}$ tend en loi, lorsque $n \to \infty$, vers une loi du χ^2_{k-1}.

Tests d'ajustement (2)

- population répartie en k classes
- échantillon de taille $n \Longrightarrow$ répartition $= (n_1, \ldots, n_k)$
- échantillon tiré selon loi multinomiale $(p_1, \ldots, p_k)$
  $\Longrightarrow (n_1, \ldots, n_k) \approx (n.p_1, \ldots, n.p_k)$
- $D^2_{(n)} = \sum_{l=1}^{k} \frac{(N_l - n.p_l)^2}{n.p_l} \sim \chi^2_{k-1}$
- d^2 valeur prise par $D^2_{(n)}$
  $\Longrightarrow$ si échantillon tiré selon $(p_1, \ldots, p_k)$ alors d^2 petit
- lecture dans une table de d^2_α tel que $P(\chi^2_{k-1} > d^2_\alpha) = \alpha$
- si $d^2 < d^2_\alpha$ alors règle de décision : $(p_1, \ldots, p_k)$ est bien la loi dont est tirée l'échantillon

Exemple de test d'ajustement

Dans un supermarché, on maintient 8 caisses de plus de 10 articles en opération durant les nocturnes du jeudi. Normalement, la clientèle devrait se répartir uniformément entre les caisses. Afin de le vérifier, on a recensé le nombre de clients passés à chacune des caisses un jeudi soir. Les résultats suivants ont été observés :

Numéro de la caisse	1	2	3	4	5	6	7	8	Total
Nombre de clients	72	70	71	52	45	59	67	48	484

Les hypothèses que l'on veut confronter sont les suivantes :

H_0 : la clientèle se répartit uniformément. H_1 : la clientèle ne se répartit pas uniformément.

Si H_0 est vraie, les clients se répartissent uniformément entre toutes les caisses. On devrait donc avoir dans chacune des caisses un effectif de $1/8$, soit $\nu_i = 484/8 = 60,5$. Par conséquent,

$$A = \frac{(72-60,5)^2}{60,5} + \frac{(70-60,5)^2}{60,5} + \frac{(71-60,5)^2}{60,5} + \frac{(52-60,5)^2}{60,5} + \frac{(45-60,5)^2}{60,5} + \frac{(59-60,5)^2}{60,5} + \frac{(67-60,5)^2}{60,5} + \frac{(48-60,5)^2}{60,5} = 15,56$$

Notons que chacun des ν_i est supérieur à 5 et que la taille de l'échantillon (484) est supérieur à 30. Par conséquent, on peut affirmer que A suit approximativement une loi du χ^2_d. Pour calculer les ν_i, nous n'avons pas eu besoin d'estimer de paramètres. Donc le nombre de degrés de liberté de la loi est égal à $I - 1 = 8 - 1 = 7$.

Si l'on fixe le niveau du test à $\alpha = 0,05$, la région critique du test correspond à rejeter H_0 si A prend une valeur supérieure à $c_{7;0,05} = 14,1$. Or $A = 15,56$. On peut donc rejeter H_0, c'est-à-dire conclure que la clientèle n'est pas répartie uniformément entre les 8 caisses.

Tests d'indépendance (1/3)

- 2 caractères X et Y
- classes de X : $A_1, A_2, \ldots, A_I$
- classes de Y : $B_1, B_2, \ldots, B_J$
- échantillon de taille n
- tableau de contingence :

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}

Tests d'indépendance (2/3)

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	*total*
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

$$\frac{n_{ij}}{n} = P(X \in A_i, Y \in B_j)$$

$$P(X \in A_i) = \frac{n_{i.}}{n} = \frac{\sum_{j=1}^{J} n_{ij}}{n} \qquad \text{et} \qquad P(Y \in B_j) = \frac{n_{.j}}{n} = \frac{\sum_{i=1}^{I} n_{ij}}{n}$$

$$X \text{ et } Y \text{ indépendants} \Longrightarrow P(X \in A_i, Y \in B_j) = P(X \in A_i) \times P(Y \in B_j)$$

Tests d'indépendance (3/3)

$X \backslash Y$	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	*total*
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

X et Y indépendants $\Longrightarrow \dfrac{n_{ij}}{n} = \dfrac{n_{i.}}{n} \times \dfrac{n_{.j}}{n}$

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

$$\chi^2_{(I-1)\times(J-1)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

Test d'indépendance – exemple

Un échantillon de 200 contribuables est prélevé afin de vérifier si le revenu brut annuel d'un individu est un caractère dépendant du niveau de scolarité de l'individu. Les résultats de l'échantillon sont résumés dans le tableau suivant :

classe de revenu (en F) \ niveau de scolarité (en années)	[0,7[	[7,12[	[12,14[	14 et plus	total
[0;75000[	17	14	9	5	45
[75000;12000[	12	37	11	5	65
[12000;20000[	7	20	20	8	55
200000 et plus	4	9	10	12	35
total	40	80	50	30	n=200

Celui-ci nous donne les valeurs des effectifs observés n_{ij} ainsi que les sommes en ligne $n_{i.}$ et en colonne $n_{.j}$. Les effectifs espérés théoriquement si H_0 est vraie s'obtiennent en multipliant entre eux les $n_{i.}$ et les $n_{.j}$, et en divisant le tout par n. Par exemple,

$$\nu_{12} = \frac{n_{1.} \times n_{.2}}{n} = \frac{45 \times 80}{200} = 18,$$

$$\nu_{34} = \frac{n_{3.} \times n_{.4}}{n} = \frac{55 \times 30}{200} = 8,25.$$

Test d'indépendance – exemple(2)

Après calcul de tous les ν_{ij}, on obtient :

classe de revenu (en F) \ niveau de scolarité (en années)	[0,7[	[7,12[	[12,14[	14 et plus	total
[0;75000[	9,00	18,00	11,25	6,75	45
[75000;12000[	13,00	26,00	16,25	9,75	65
[12000;20000[	11,00	22,00	13,75	8,25	55
200000 et plus	7,00	14,00	8,75	5,25	35
total	40	80	50	30	n=200

La taille de l'échantillon est grande (200). De plus, on peut remarquer dans ce tableau que tous les ν_{ij} sont supérieurs ou égaux à 5. Par conséquent, sous H_0, la statistique d'ajustement suivante :

$$A = \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{n_{ij}^2}{\nu_{ij}} \right) - n$$

A suit une loi χ^2_d, où $d = (I-1) \times (J-1) = 3 \times 3 = 9$ (puisqu'il y a $I = 4$ classes de valeurs pour le premier caractère, et $J = 4$ classes pour le deuxième caractère). En utilisant un niveau $\alpha = 0,05$, on rejette H_0 si la valeur de A est supérieure à $c_{9;0,05} = 16,9$. Si l'on calcule A, on trouve la valeur : 34,06. Par conséquent, on doit rejeter H_0 : on conclut qu'il y a une dépendance entre le revenu et le niveau de scolarité dans la population étudiée (quelque part, c'est tout de même rassurant !).