Tests d'adéquation et d'indépendance du Chi-deux

Frédérique Leblanc

Problèmes :

1. Tester l'adéquation entre la loi d'une variable qualitative ou quantitative discrète X et une loi cible donnée
2. Tester l'indépendance entre deux variables X et Y qualitatives ou quantitatives discrètes

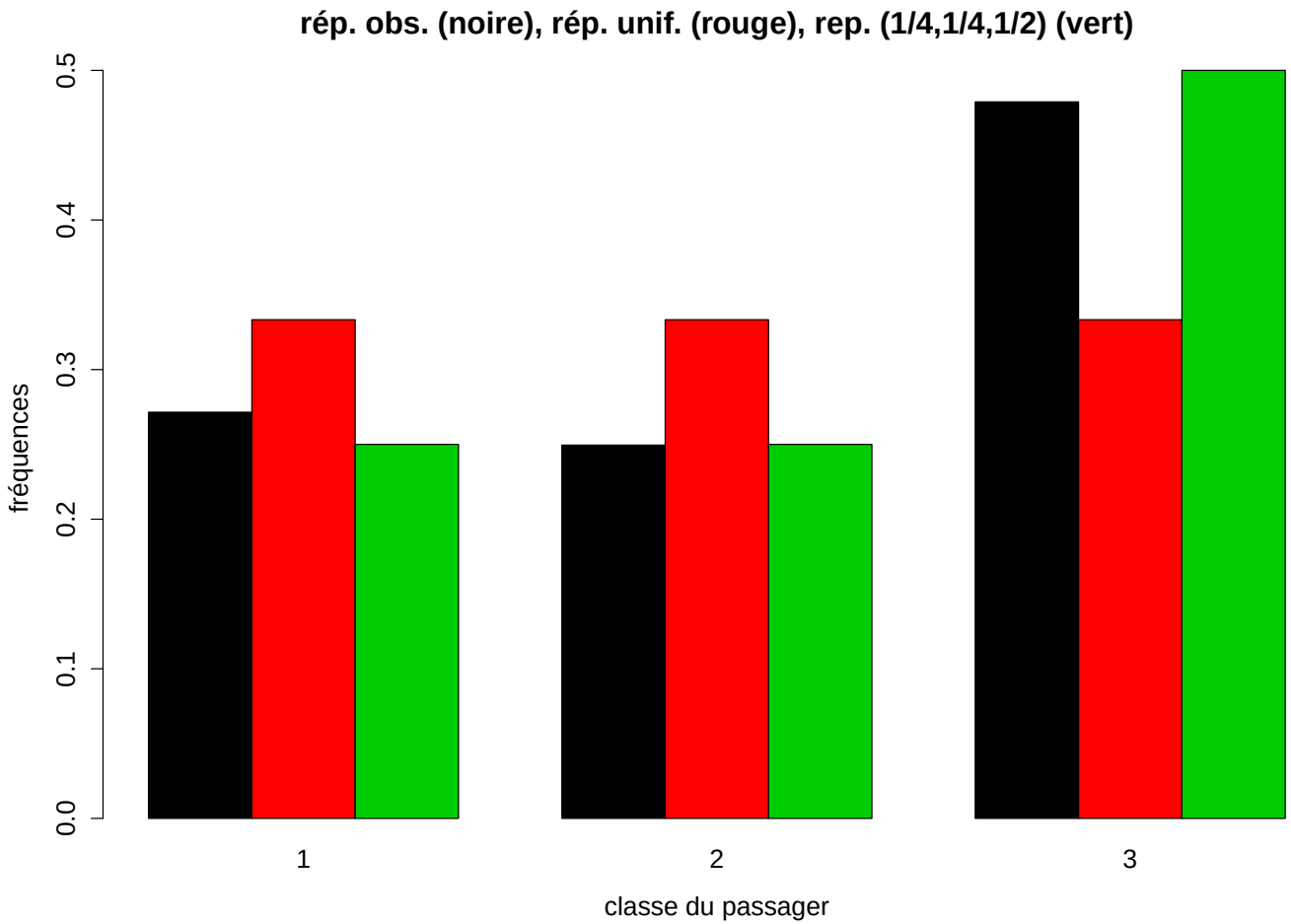
Données :

1. Un échantillon de taille n de X : $x_1, ..., x_n$
2. Deux échantillons de X et de Y appariés et de même taille n : $(x_1, y_1), ..., (x_n, y_n)$

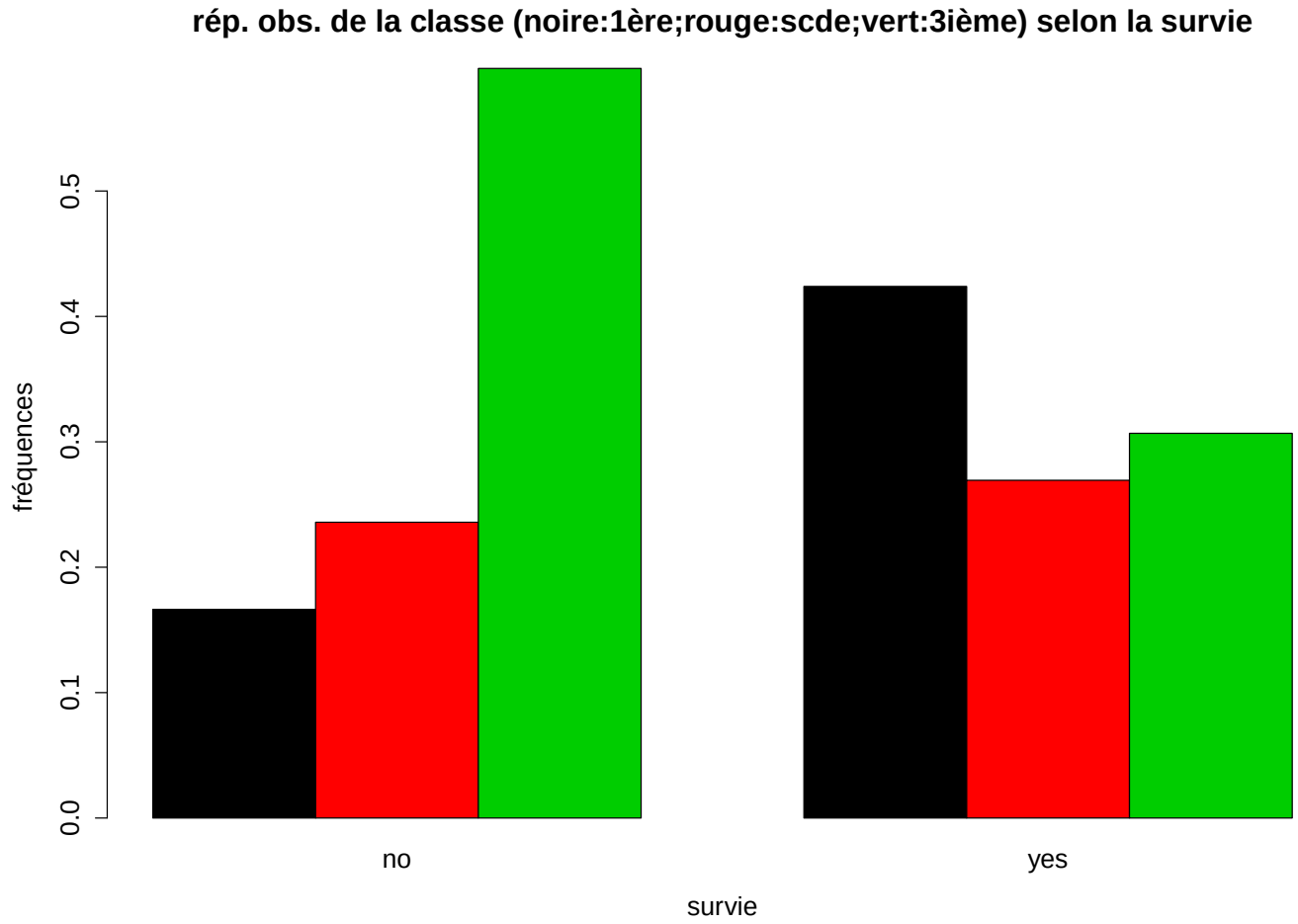
dans le fichier `titanic.csv`

1. X la classe (à trois modalités $\{m_1, m_2, m_3\} = \{1, 2, 3\}$) suit-elle la loi cible définie par (p_1^*, p_2^*, p_3^*) ?
   On pourrait tester par exemple une répartition uniforme avec $(p_1^*, p_2^*, p_3^*) = (1/3, 1/3, 1/3)$
2. X la classe (à valeurs dans $\{m_1, m_2, m_3\} = \{1, 2, 3\}$) et Y la survie (à valeurs dans $\{\tilde{m}_1, \tilde{m}_2\} = \{0, 1\}$) sont elles indépendantes ?

Exemple 1 : Répartition observée de `pclass`, et deux repartitions théoriques (rouge et vert)



Exemple 2 : Répartitions observées de `pclass` selon `survived`



Modèle : $X_1, ..., X_n$ échantillon de la variable X à valeurs dans $\{m_1, m_2, ..., m_q\}$

Notations :

- $(p_1, ..., p_q) = (P(X = m_1), ..., P(X = m_q))$
  loi de proba de X **inconnue**
- N_k : nombre de variables parmi $X_1, ..., X_n$ qui prennent la modalité m_k.
- Effectifs observés : n_k réalisation de N_k pour l'échantillon observé $x_1, ..., x_n$.
- Probabilités théoriques : $(p_1^*, ..., p_q^*)$ **connues**
- Effectifs attendus (ou théoriques) : $n_k^* = np_k^*$ effectif attendu de la modalité m_k si la loi de X est $(p_1^*, p_2^*, ..., p_q^*)$.

Les hypothèses du test d'adéquation

$$\mathcal{H}_0 : \quad X \quad \text{suit} \quad \text{la} \quad \text{loi} \quad (p_1^*, p_2^*, ..., p_q^*)$$

$$\mathcal{H}_1 : \quad X \quad \text{ne} \quad \text{suit} \quad \text{pas} \quad \text{la} \quad \text{loi} \quad (p_1^*, p_2^*, ..., p_q^*)$$

ou de façon équivalente

$$\mathcal{H}_0 : (p_1, p_2, ..., p_q) = (p_1^*, p_2^*, ..., p_q^*)$$

$$\mathcal{H}_1 : (p_1, p_2, ..., p_q) \neq (p_1^*, p_2^*, ..., p_q^*)$$

ou encore

$$\mathcal{H}_0 : \delta^2 = 0 \quad \mathcal{H}_1 : \delta^2 > 0$$

avec $\delta^2 = \sum_{k=1}^{q} (np_k - np_k^*)^2/(np_k^*)$

- **Conditions d'application de ce test :** $np_k^* \geq 5$
- **Estimateur de** p_k **:** N_k
- **Statistique de test : Estimateur de** δ^2 **:** D^2 défini par

$$D^2 = \sum_{k=1}^{q} \frac{(N_k - np_k^*)^2}{np_k^*}.$$

- **Loi de la statistique de test si** $np_k^* \geq 5$**:** Sous $\mathcal{H}_0$ la loi de D^2 est un $\mathcal{X}^2_{q-1}$
- **Rejet de** $\mathcal{H}_0$ **au seuil** α : $\{D^2 > z_{q-1,1-\alpha}\}$
- **p-valeur** : $P(D^2 > D^2_{calc}) = 1 - F(D^2_{calc})$ où F est la Fdr d'une loi du $\mathcal{X}^2_{q-1}$

En pratique :

Soit les données sont présentées avec le tableau des effectifs, soit on dispose des données brutes et on le construit à l'aide de la fonction `table()` de R. On peut décrire les calculs à l'aide du tableau :

modalités	m_1	...	m_k	...	m_q	total
eff. obs.	n_1	...	n_k	...	n_q	n
prob. théo.	p_1^*	...	p_k^*	...	p_q^*	1
eff. att.	np_1^*	...	np_k^*	...	np_q^*	n
contrib d^2	$\frac{(n_1-np_1^*)^2}{np_1^*}$	...	$\frac{(n_k-np_k^*)^2}{np_k^*}$	...	$\frac{(n_q-np_q^*)^2}{np_q^*}$	D^2_{calc}

puis on calcule la p-valeur : $\alpha^* = P(D^2 > D^2_{calc})$

Exemple 1 : Adéquation à la loi uniforme dans les données `titanic.csv` de la variable `class` à valeurs dans $\{1, 2, 3\}$?

$$\mathcal{H}_0 : (p_1, p_2, p_3) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \qquad \mathcal{H}_1 : (p_1, p_2, p_3) \neq \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

mod.	1	2	3	total
eff. obs.	284	261	501	1046
pr. théo.	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	1
eff. att.	$\frac{1046}{3} = 348,67$	$\frac{1046}{3} = 348,67$	$\frac{1046}{3} = 348,67$	1046
cont. d^2	$\frac{(284-348,67)^2}{348,67}$	$\frac{(261-348,67)^2}{348,67}$	$\frac{(501-348,67)^2}{348,67}$	100, 59

p-valeur : $P_{\mathcal{H}_0}(D^2 > 100, 59) \approx 0$
La répartition de la classe n'est donc clairement pas uniforme

Exercice : Adéquation à la loi $\{1/4, 1/4, 1/2\}$ dans les données `titanic.csv` de la variable `class` à valeurs dans $\{1, 2, 3\}$?

$$\mathcal{H}_0 : (p_1, p_2, p_3) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right) \qquad \mathcal{H}_1 : (p_1, p_2, p_3) \neq \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$$

On obtient dans ce cas une p-valeur de $23,9\%$

Conclusion littérale : A moins de prendre un risque de se tromper supérieur à $23,9\%$ on ne peut pas refuser $\mathcal{H}_0$ (la répartition de la classe est donc bien $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right)$).

Retrouver ce résultat avec la fonction $TEST - \mathcal{X}^2 GOF$ de la calculatrice

Remarques :

- On peut étendre ce test au variables quantitatives continues en remplaçant m_k par la classe C_k lorsque l'ensemble des valeurs de X est partitionné en q classes.
  Dans ce cas la loi de D^2 est toujours un $\mathcal{X}^2$ mais à $q-1-r$ degrés de liberté avec r=nombre de paramètres estimés dans la loi cible proposée décrite par $(p_1^*, ..., p_q^*)$.
  On peut ainsi tester l'adéquation d'une variable continue X a une loi normale ayant r paramètres inconnus.
- Si l'un des effectifs théoriques < 5 on regroupera deux (ou plusieurs) modalités consécutives ou deux ou plusieurs classes dans le cas continu.

Modèle : $(X_1, Y_1), ..., (X_n, Y_n)$, échantillon aléatoire d'un couple de variables aléatoires X et Y à valeurs dans $\mathcal{X} = \{m_1, ..., m_q\}$ et $\mathcal{Y} = \{\tilde{m}_1, ..., \tilde{m}_p\}$.

Notations :

- N_{ij} la var. aléa. : nb. de (X_k, Y_k) de $(X_1, Y_1)...., (X_n, Y_n)$ qui prennent les modali tés $(m_i, \tilde{m}_j)$.
- $N_{k.}$: nb de X_i qui prennent la modalité m_k

$$N_{k.} = \sum_{j=1}^{p} N_{kj}$$

- $N_{.k}$: nb de Y_j qui prennent la modalité $\tilde{m}_k$

$$N_{.k} = \sum_{i=1}^{q} N_{ik}$$

Données : Soit on a les données brutes soit directement les données en effectifs présentées dans un tableau appelé *Tableau de contingence*

Y / X	$\tilde{m}_1$	...	$\tilde{m}_j$	...	$\tilde{m}_p$	Total
m_1	n_{11}	...	n_{1j}	...	n_{1p}	$n_{1.}$
$\vdots$						
m_i	n_{i1}	...	n_{ij}	...	n_{ip}	$n_{i.}$
$\vdots$						
m_q	n_{q1}	...	n_{qj}	...	n_{qp}	$n_{q.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.p}$	$n_{..} = n$

Indépendance entre X et Y est par def.:

$$P(X = m_i, Y = \tilde{m}_j) = P(X = m_i)P(X = \tilde{m}_j), \quad \text{pour tout} \quad (i, j)$$

On veut tester

$\mathcal{H}_0 : X$ et Y indépendantes $\mathcal{H}_1 : X$ et Y non indépendantes

On peut de façon équivalente définir $\mathcal{H}_0 : \delta^2 = 0$ avec

$$\delta^2 = n \cdot \sum_{i,j} \frac{(P(X = m_i, Y = \tilde{m}_j) - P(X = m_i)P(X = \tilde{m}_j))^2}{P(X = m_i)P(X = \tilde{m}_j)}.$$

On peut donc ramener le problème à un test sur le paramètre δ^2 soit :

$$\mathcal{H}_0 : \delta^2 = 0 \quad \text{contre} \quad \mathcal{H}_1 : \delta^2 > 0.$$

- **Conditions d'application de ce test :** $n_{i.}n_{.j}/n \geq 5$ pour tout (i,j)
- **Estimateur de** $P(X = m_i)$: $N_{i.}/n$
- **Estimateur de** $P(Y = \tilde{m}_j)$: $N_{.j}/n$
- **Statistique de test :** D^2 **Estimateur de** δ^2 défini par

$$D^2 = \sum_{i,j} \frac{\left(N_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}}.$$

- **Loi de la statistique de test sous** $\mathcal{H}_0$ **si** $n_{i.}n_{.j}/n \geq 5$ **:** D^2 suit une loi du $\mathcal{X}^2_{(p-1)(q-1)}$
- **Rejet de** $\mathcal{H}_0$ **au seuil** α : $\{D^2 > z_{(p-1)(q-1),1-\alpha}\}$
- **p-valeur** : $P(D^2 > D^2_{calc}) = 1 - F(D^2_{calc})$ où F est la Fdr d'une loi du $\mathcal{X}^2_{(p-1)(q-1)}$

En pratique :

Soit les données sont présentées avec le tableau de contingence, soit on dispose des données brutes et on le construit à l'aide de la fonction `table()` de R. On peut décrire les calculs à l'aide du tableau de contingence qui donne **les effectifs observés** $n_{i,j}$ et du tableau des **effectifs attendus** contenant les $n_{.i}n_{j.}/n$ On calcule ensuite les pq termes de contribution au calcul D^2 en construisant le tableau à double entrées des **contributions** et qui contient en ligne i et colonne j:

$$\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

Exemple 2 : Indépendance entre la classe occupée par le passager `pclass` (variable X à valeurs dans $\{1, 2, 3\}$) et l'indicatrice de sa survie `survived` (variable Y à valeurs dans $\{"no", "yes"\}$) ?

$\mathcal{H}_0 : X$ et Y indépendantes $\mathcal{H}_1 : X$ et Y non indépendantes

tableau de contingence : obtenu par la commande R
`table(titanic$pclass,titanic$survived)`

effectifs observés

X \ Y	*no*	*yes*	Total
1	103	181	284
2	146	115	261
3	370	131	501
Total	619	427	1046

effectifs attendus

X \ Y	*no*	*yes*	Total
1	168, 07	115,93	284
2	154, 45	106,55	261
3	296, 48	204,52	501
Total	619	427	1046

Statistique de test et p-valeur:

$$D^2_{calc} = 107,5; \qquad P_{\mathcal{H}_0}(D^2 > D^2_{calc}) = 4,5 \cdot 10^{-24}$$

Conclusion : On conclut que Survie et Classe sont dépendantes avec un très faible risque de se tromper ($> 4,5 \cdot 10^{-24}$)