Semaine 13
Exemples de tests d'adéquation et test d'ajustement

Test d'adéquation sur une variable discrète

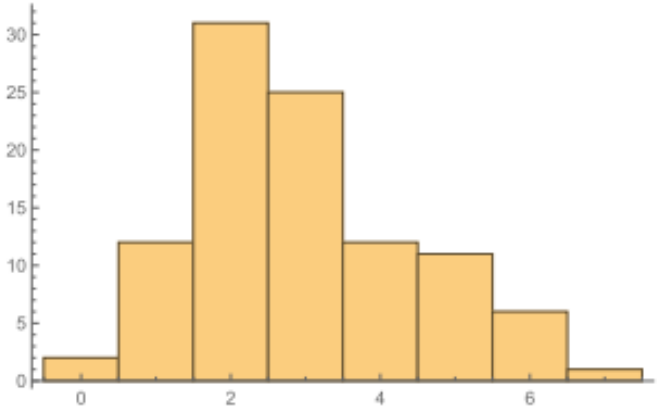
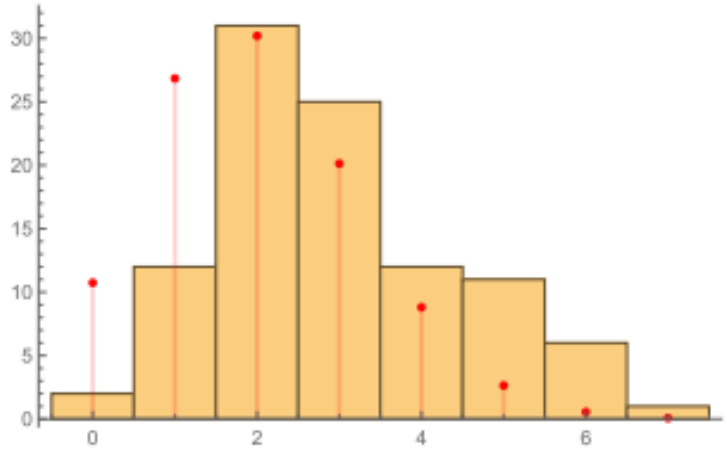


Diagramme à bades des données



En rouge : distribution binomiale$(10, 1/5)$

Étant donné des observations, on peut faire un test d'hypothèse

H_0 : Distribution suit une loi P $\quad$ H_1 : ne suit pas P

Test d'adéquation sur une variable discrète

$$T_{\text{stat}} = \sum_{j=1}^{k} \frac{(O_i - A_i)^2}{A_i}$$

où $k =$ nb de modalités

$O_i =$ valeur observée pour la i-ième modalité

$A_i =$ valeur attendue pour la i-ième modalité

Attention ! k n'est pas nombre d'observations. Le nombre d'observation est

$$n = O_1 + \cdots + O_k = A_1 + \cdots + A_k.$$

$T_{\text{stat}} \approx \chi^2_{k-1}$

Exemple

On veut comparer les données à une distribution uniforme.

Modalités	1	2	3	4	5	6	Total
Fréquence	24	30	27	26	20	21	148

H_0 : distribution est uniforme

Modalités	1	2	3	4	5	6	Total
Probabilité	1/6	1/6	1/6	1/6	1/6	1/6	1
Fréquence	$24,\overline{6}$	$24,\overline{6}$	$24,\overline{6}$	$24,\overline{6}$	$24,\overline{6}$	$24,\overline{6}$	148

$$t_{\text{stat}} = \frac{(24-24,\overline{6})^2}{24,\overline{6}} + \frac{(30-24,\overline{6})^2}{24,\overline{6}} + \frac{(27-24,\overline{6})^2}{24,\overline{6}} + \frac{(26-24,\overline{6})^2}{24,\overline{6}} + \frac{(20-24,\overline{6})^2}{24,\overline{6}} + \frac{(21-24,\overline{6})^2}{24,\overline{6}} = 2,89$$

$$dl = 6 - 1 = 5$$

$$T_{\text{stat}} \approx \chi^2_5$$

$$\text{valeur-}p = \mathbb{P}(T_{\text{stat}} \geq t_{\text{stat}} | H_0) \approx 71\%$$

On ne rejette pas H_0.

Nombre infini de modalités

Supposons que la distribution de la population a une infinité de modalité et que l'on veuille tester contre une loi géométrique de paramètre q.
On peut approximer cela en garder les $k-1$ premières modalités (celles plus probables) et en regroupant les modalités k à ∞ en une seule classe.
Si X suit la distribution de la population, on peut formuler l'hypothèse nulle comme suit

$$H_0 : \begin{cases} \mathbb{P}(X=1) = p_1 \\ \quad \vdots \\ \mathbb{P}(X=k-1) = p_{k-1} \\ \mathbb{P}(X \geq k) = p \end{cases}$$

où $p_i = (1-q)^{i-1}q$ et $p = \sum_{n=k}^{\infty}(1-q)^{n-1}q$.

Exemple

Test d'adéquation, infinité de modalités

On veut comparer les données à une distribution géométrique(0,8). D'abord, remarquons que si $X \sim$ géométrique(0,8), alors $\mathbb{P}(X \geq 4) = 0{,}008$, donc il semble raisonnable de combiner les observations des modalités 4 et plus.

Modalités	1	2	3	4+	Total
Fréquence	48	18	2	2	70

H_0 : distribution est uniforme

Modalités	1	2	3	4+	Total
Probabilité	0,8	0,16	0,032	0,008	1
Fréquence	56	11,2	2,24	0,56	70

Exemple

Test d'adéquation, infinité de modalités (suite)

$$t_{\text{stat}} = \frac{(48-56)^2}{56} + \frac{(18-11{,}2)^2}{11{,}2} + \frac{(2-2{,}24)^2}{2{,}24} + \frac{(2-0{,}56)^2}{0{,}56} = 3{,}70$$

Le degré de liberté est $4 - 1 = 3$.

On trouve valeur-$p = \mathbb{P}(T_{\text{stat}} \geq t_{\text{stat}} | H_0) = 0{,}296$. On ne rejette pas H_0 avec un seuil de 5%.

Test d'indépendance

Test d'indépendance

Soit X et Y deux variables catégorielles indépendantes, où X possède n modalités et Y, k modalités.
H_0 : X et Y sont indépendantes.
Pour la i-ième modalité de X et la j-ième de Y, il y a

$$O_{ij} = \text{valeur observée}, \qquad A_{ij} = \text{valeur attendue}.$$

La statistique de test

$$T_{\text{stat}} = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

suit (approximativement) une loi du χ^2 avec $dl = (n-1)(k-1)$ degrés de liberté sous H_0.

Couleur de yeux et couleur de cheveux indépendants

On se demande si $X =$ couleur des yeux et $Y =$ couleur des cheveux sont indépendantes.

Pour remplir la table des données observées, il est peut-être plus commode de penser à l'indépendance comme suit :

$$\mathbb{P}(\text{couleur yeux}|\text{couleur cheveux}) = \mathbb{P}(\text{couleur yeux}).$$

Couleur de yeux et couleur de cheveux indépendants

Tableau des observations

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 %)

(Les pourcentages sont indiqués «en ligne». On aurait pu aussi le faire «en colonne».)

Couleur de yeux et couleur de cheveux indépendants

Remplir le tableau des observations attendues

Prenons seulement la colonne blond pour le moment.

Observées

couleur yeux	blond
bleu	326 (45 %)
vert	688 (44 %)
brun	343 (19 %)
noir	98 (7 %)
total	1455 (27 %)

$\longrightarrow$

couleur yeux	blond
bleu	(27 %)
vert	(27 %)
brun	(27 %)
noir	(27 %)
total	1455 (27 %)

Attendues

Si X =«couleur cheveux» et Y =«couleur yeux» sont indépendantes, alors il devrait y avoir le même pourcentage de blond parmi toutes les couleurs d'yeux, c'est-à-dire

$\mathbb{P}(blond|blue) = \mathbb{P}(blond) = 27\%, \quad \mathbb{P}(blond|vert) = \mathbb{P}(blond) = 27\%$

Couleur de yeux et couleur de cheveux indépendants

Remplir le tableau des observations attendues

couleur yeux	couleur cheveux blond	roux	châtain	brun	noir	total
bleu						718 (100 %)
vert						1580 (100 %)
brun						1774 (100 %)
noir						1315 (100 %)
total	1455 (27 %)	286 (5 %)	2137 (40 %)	1391 (26 %)	118 (2 %)	5387 (100 %)

$$\text{blond/bleu} = 718 \times 27\% = \frac{718 \times 1455}{5387}$$

Couleur de yeux et couleur de cheveux indépendants

Tableau des observations attendues

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

Couleur de yeux et couleur de cheveux indépendants

Statistique de test et conclusion

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

1. T_{stat} suit une loi du χ^2 avec $(4-1)\times(5-1) = 12$ degrés de liberté
2. on trouve $p = \mathbb{P}(T_{\text{stat}} > 1240) \approx 0$, donc on rejette H_0
3. Conclusion : la couleur des cheveux et la couleur des yeux ne sont pas indépendantes.

Exemple 2

(Emprunté des notes d'Élise Davignon)

On a recueilli les données suivantes auprès d'un échantillon de $n = 906$ élèves du collège de sorcellerie Poudlard. On s'intéresse à l'indépedance des variables X =«revenu familial» et Y=«maison à Poudlard» avec un seuil $\alpha = 1\%$.

Revenu familial \ Maison de Poudlard f	Gryffondor	Serpentard	Poufsouffle	Serdaigle	Total
Élevé	41	153	160	112	466
Faible	20	107	192	121	440
Total	61	260	352	233	906

Table des données observées

Exemple 2

Suite 1

Table des données attendues

Revenu familial / Maison de Poudlard *e*	Gryffondor	Serpentard	Poufsouffle	Serdaigle	Total
Élevé					466
Faible					440
Total	61	260	352	233	906

Exemple 2

Suite 2

Table des données observées

Revenu familial \ Maison de Poudlard f	Gryffondor	Serpentard	Poufsouffle	Serdaigle	Total
Élevé	41	153	160	112	466
Faible	20	107	192	121	440
Total	61	260	352	233	906

Table des données attendues

Revenu familial \ Maison de Poudlard e	Gryffondor	Serpentard	Poufsouffle	Serdaigle	Total
Élevé	31,38	133,73	181,05	119,84	466
Faible	29,62	126,27	170,95	113,16	440
Total	61	260	352	233	906

Exemple 2

Suite 3

On trouve $t_{\text{stat}} = 17{,}89$.
Le degré de liberté est $(4-1) \times (2-1) = 3$.
valeur-$p \approx 0{,}04\%$
Conclusion : on rejette l'hypothèse que les variables sont indépendantes.

Test d'indépendance et corrélation

Attention

La valeur-p ne mesure pas l'intensité du lien entre X et Y !
Autrement dit, même si p est très très petit, cela ne veut pas dire que les variables sont fortement liées ou faiblement liées.
C'est la **corrélation** qui mesure l'intensité de l'association (linéaire) entre les variables.