

Données qualitatives

David Causeur

Laboratoire de Mathématiques Appliquées

Agrocampus Rennes

IRMAR CNRS UMR 6625

<http://www.agrocampus-rennes.fr/math/causeur/>

Plan de la présentation

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements

Galetas ... análisis sensorial



Galetas ... análisis sensorial



- Couleur préférée ?
- Forme préférée ?



Galetas ... análisis sensorial

Juges	Galle	Color	Grupo	Sexo
1	I	H	niño	muj.
2	D	H	niño	muj.
3	G	J	niño	muj.
4	A	J	niño	muj.
5	B	J	niño	muj.
⋮	⋮	⋮	⋮	⋮
1192	H	H	adulto	humb.

- Couleur préférée ?
- Forme préférée ?



Problématiques

Lien entre variables qualitatives

- Y-a-t'il un lien entre la forme et la couleur préférée ?



Problématiques

Lien entre variables qualitatives

- Y-a-t'il un lien entre la forme et la couleur préférée ?
- Y-a-t'il un lien entre la préférence et l'âge ?



Problématiques

Lien entre variables qualitatives

- Y-a-t'il un lien entre la forme et la couleur préférée ?
- Y-a-t'il un lien entre la préférence et l'âge ?
- Si oui, quelle est l'intensité de ces liens ?



Problématiques

Lien entre variables qualitatives

- Y-a-t'il un lien entre la forme et la couleur préférée ?
- Y-a-t'il un lien entre la préférence et l'âge ?
- Si oui, quelle est l'intensité de ces liens ?
- Si oui, comment se traduit ce lien en pratique ?
- ...



Les pois de Mendel

Croisement de 2 hétérozygotes

	JAUNE, ROND	JAUNE, anguleux	vert, ROND	vert, anguleux
JAUNE, ROND	JAUNE, ROND	JAUNE, ROND	JAUNE, ROND	JAUNE, ROND
JAUNE, anguleux	JAUNE, ROND	JAUNE, anguleux	JAUNE, ROND	JAUNE, anguleux
vert, ROND	JAUNE, ROND	JAUNE, ROND	vert, ROND	vert, ROND
vert, anguleux	JAUNE, ROND	JAUNE, anguleux	vert, ROND	vert, anguleux

Répartition théorique des phénotypes

JAUNE, ROND	JAUNE, anguleux	vert, ROND	vert, anguleux
$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$



Validation par l'expérience

Expérience : 556 croisements de parents hétérozygotes

JAUNE, ROND	JAUNE, anguleux	vert, ROND	vert, anguleux
315	108	101	32

L'expérience confirme-t'elle la théorie biologique ?



Plan de la présentation

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Plan du module

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - **Modèle multinomial**
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Modèle pour variables qualitatives

X = phénotype d'un plant hybride

$X \in \{ (\text{JAUNE}, \text{ROND}), (\text{JAUNE}, \text{anguleux}), (\text{vert}, \text{ROND}), (\text{vert}, \text{anguleux}), \}$

D'après la théorie biologique,

$$\mathbb{P}(X = JR) = 9/16, \mathbb{P}(X = Ja) = 3/16,$$

$$\mathbb{P}(X = vR) = 3/16, \mathbb{P}(X = va) = 1/16.$$



Modèle multinomial

n épreuves indépendantes à K issues possibles

Épreuve 1	Épreuve 2	...	Épreuve n
E_1	E_1	...	E_1
ou	ou	...	ou
E_2	E_2	...	E_2
ou	ou	...	ou
\vdots	\vdots		\vdots
E_K	E_K	...	E_K

N_k : nombre de réalisations de E_k

$$(N_1, N_2, \dots, N_K) \sim \mathcal{M}(n; \pi_1, \pi_2, \dots, \pi_K)$$



Expérience versus théorie

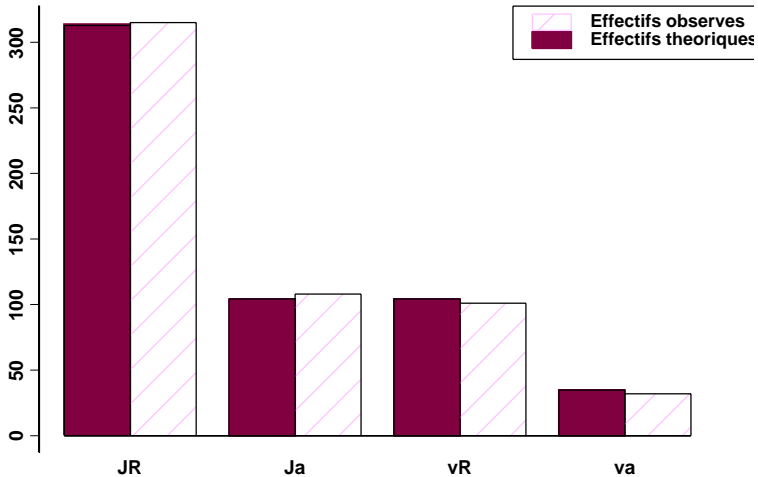
J,R	J,a	v,R	v,a
$n_1 = 315$	$n_2 = 108$	$n_3 = 101$	$n_4 = 32$
56.65 %	19.42 %	18.17 %	5.76 %

$$\left\{ \begin{array}{l} H_0 : (N_1, N_2, N_3, N_4) \sim \mathcal{M}(556; 9/16, 3/16, 3/16, 1/16) \\ H_1 : (N_1, N_2, N_3, N_4) \not\sim \mathcal{M}(556; 9/16, 3/16, 3/16, 1/16) \end{array} \right.$$

	Phénotype			
	JR	Ja	vR	va
Proba.	9/16	3/16	3/16	1/16
Eff. théo.	$(9/16) \times 556$ = 312.75	$(3/16) \times 556$ = 104.25	$(3/16) \times 556$ = 104.25	$(1/16) \times 556$ = 35.75
Eff. obs.	315	108	101	32



Expérience versus théorie





Mesure de l'adéquation entre observations et théorie

	Phénotypes			
	JR	Ja	vR	va
Eff. théo.	$556 \times (9/16)$	$556 \times (3/16)$	$556 \times (3/16)$	$556 \times (1/16)$
Eff. obs.	315	108	101	32

De manière générale ...

	Issues possibles				
	E_1	E_2	...	E_{K-1}	E_K
Effectifs théoriques	$n\pi_1$	$n\pi_2$...	$n\pi_{K-1}$	$n\pi_K$
Effectifs observés	n_1	n_2	...	n_{K-1}	n_K

$$D^2 = \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_2 - n\pi_2)^2}{n\pi_2} + \dots + \frac{(n_K - n\pi_K)^2}{n\pi_K}$$





Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$D^2 = \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned} D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\ &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \end{aligned}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned}
 D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \underbrace{(1 - \pi_1 + \pi_1)}_1
 \end{aligned}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned}
 D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \underbrace{(1 - \pi_1 + \pi_1)}_1 \\
 &= \left(\frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} \right)^2
 \end{aligned}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned}
 D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \underbrace{(1 - \pi_1 + \pi_1)}_1 \\
 &= \left(\frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} \right)^2
 \end{aligned}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned}
 D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \underbrace{(1 - \pi_1 + \pi_1)}_1 \\
 &= \left(\frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} \right)^2 \sim \chi_1^2
 \end{aligned}$$



Cas particulier du modèle binomial ($K = 2$)

Si H_0 est vraie

$$\begin{aligned}
 D^2 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{[n - n_1 - n(1 - \pi_1)]^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1} + \frac{(n_1 - n\pi_1)^2}{n(1 - \pi_1)} \\
 &= \frac{(n_1 - n\pi_1)^2}{n\pi_1(1 - \pi_1)} \underbrace{(1 - \pi_1 + \pi_1)}_1 \\
 &= \left(\frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1 - \pi_1)}} \right)^2 \sim \chi_1^2
 \end{aligned}$$

De manière générale, si H_0 est vraie, alors $D^2 \sim \chi_{K-1}^2$

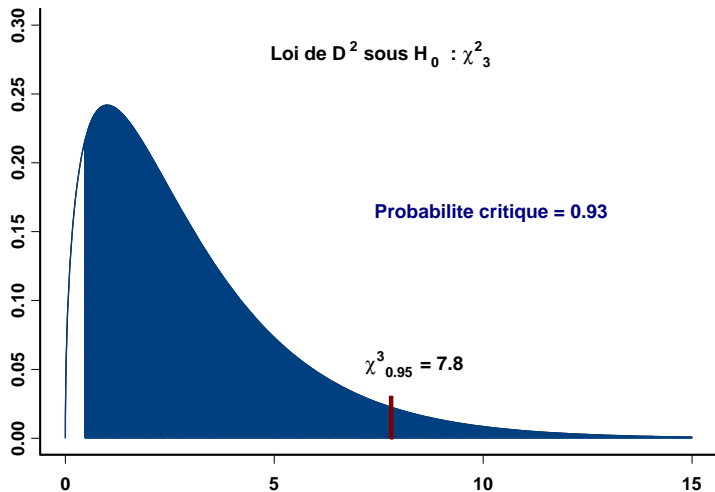


Revenons à notre exemple ...

	Phénotypes				D^2
	JR	Ja	vR	va	
Eff. théo.	$556 \times (9/16)$	$556 \times (3/16)$	$556 \times (3/16)$	$556 \times (1/16)$	
Eff. obs.	315	108	101	32	
$\frac{(n_j - n\pi_j)^2}{n\pi_j}$	0.02	0.13	0.10	0.22	0.47



Revenons à notre exemple ...





Plan du module

- 1 Introduction
- 2 **Test d'ajustement à une loi théorique**
 - Modèle multinomial
 - **Modèle pour variable de comptage**
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements

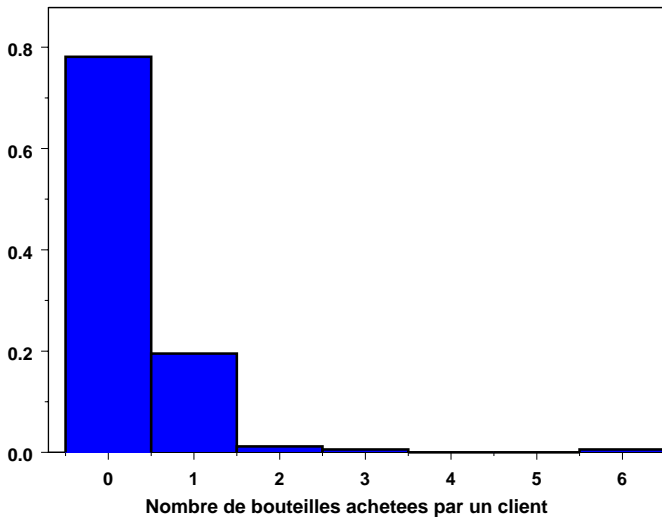


Comportement de clients

Client	Luminosité	Musique	Durée (minutes)	...	Nombre de bouteilles examinées	Nombre de bouteilles extraites	Nombre de bouteilles achetées
1	Douce	Class.	2.0	...	0	0	0
2	Douce	Class.	1.0	...	0	0	0
3	Douce	Class.	0.5	...	0	0	0
4	Douce	Class.	2.3	...	16	0	0
5	Douce	Class.	2.2	...	1	1	0
6	Douce	Class.	1.7	...	1	0	0
7	Douce	Class.	42.7	...	11	3	1
8	Douce	Class.	3.0	...	0	0	0
9	Douce	Class.	0.7	...	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



Comportement de clients





Modèle pour variable de comptage

X : nombre de bouteilles

$$\begin{cases} H_0 : X \sim \mathcal{P}(\lambda) \Leftrightarrow \text{Les clients ne s'influencent pas} \\ H_1 : X \not\sim \mathcal{P}(\lambda) \end{cases}$$

Sous H_0 :

$$\begin{aligned} \pi_k = \mathbb{P}(X = k) &= \frac{\lambda^k}{k!} \exp(-\lambda), \\ \mathbb{E}(X) &= \lambda \end{aligned}$$

Test du caractère Poissonnien

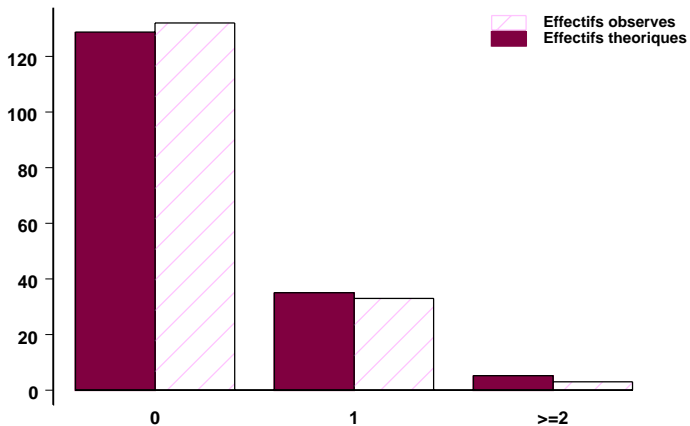
Estimation de λ :

$$\hat{\lambda} = \bar{X} \text{ (ici, } n = 169, \bar{X} = 0.272)$$

Nbre de bout.	Effectifs	Effectifs	$\frac{(n_j - n\pi_j)^2}{n\pi_j}$
	théoriques	observés	
0	$169 \times \frac{0.272^0}{0!} \exp(-0.272) = 128.7$	132	0.08
1	$169 \times \frac{0.272^1}{1!} \exp(-0.272) = 35.0$	33	0.12
2	$169 \times \frac{0.272^2}{2!} \exp(-0.272) = 4.8$	2	
3 et plus	$169 \times \frac{0.272^3}{3!} \exp(-0.272)$ $+ 169 \times \frac{0.272^4}{4!} \exp(-0.272)$ $+ \dots = 0.5$	2	0.29
			$D^2 = 0.49$



Test du caractère Poissonnien





Test du caractère Poissonnien

$$\text{Sous } H_0, D^2 \sim \chi^2_{\underbrace{K-1-1}_{1 \text{ paramètre estimé}}}$$

Ici, probabilité critique : 0.486

Analyse des contributions

	Nombre de bouteilles		
	0	1	≥ 2
Effectifs théoriques	128.7	35.0	5.3
Effectifs observés	132	33	4
$\frac{(n_i - n\pi_i)^2}{n\pi_i}$ D^2	16.85 %	24.00 %	59.16 %

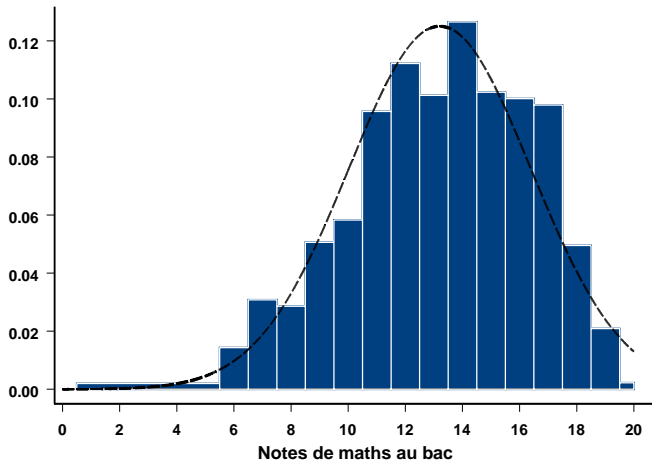


Plan du module

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - **Modèle pour variable continue**
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Homogénéité d'une population





Homogénéité d'une population

$$\begin{cases} H_0 : X \sim \mathcal{N}(\mu; \sigma) \Leftrightarrow \text{Population homogène} \\ H_1 : X \not\sim \mathcal{N}(\mu; \sigma) \end{cases}$$

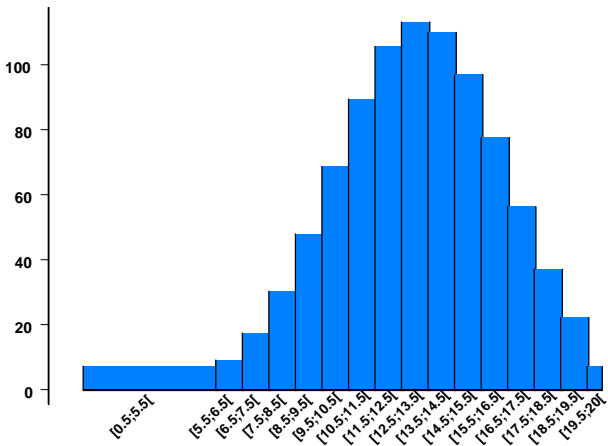
$$\hat{\mu} = \bar{X} = 13.2, \hat{\sigma} = S' = 3.2$$

Soit $U \sim \mathcal{N}(13.2; 3.2)$

Classe	Effectifs	Effectifs	$\frac{(n_i - n\pi_i)^2}{n\pi_i}$
	théoriques	observés	
[0.5; 5.5[$909 \times \mathbb{P}(0.5 \leq U < 5.5) = 7.1$	9	0.52
[5.5; 6.5[$909 \times \mathbb{P}(5.5 \leq U < 6.5) = 9.0$	13	1.78
[6.5; 7.5[$909 \times \mathbb{P}(6.5 \leq U < 7.5) = 17.3$	28	6.62
[7.5; 8.5[$909 \times \mathbb{P}(7.5 \leq U < 8.5) = 30.2$	26	0.58
[8.5; 9.5[$909 \times \mathbb{P}(8.5 \leq U < 9.5) = 47.8$	46	0.07
⋮	⋮	⋮	⋮
			$D^2 = 46.46$

Homogénéité d'une population

Effectifs théoriques





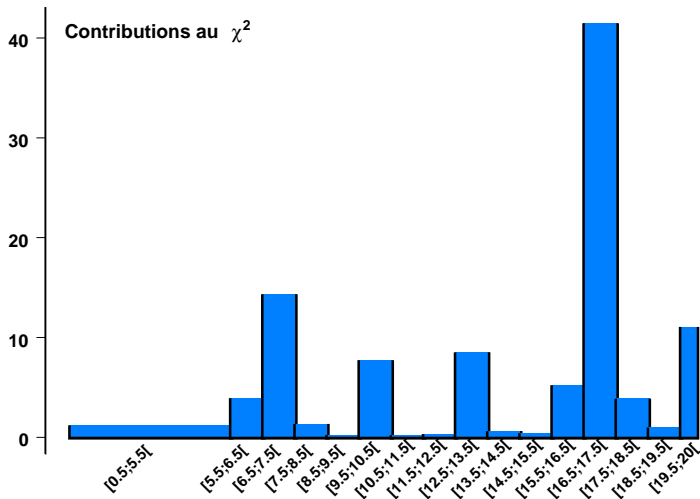
Test de normalité

Sous H_0 , $D^2 \sim \chi^2_{\underbrace{K-1-2}_{2 \text{ paramètres estimés}}}$

Ici, $K = 16$, probabilité critique : 1.2×10^{-5}



Test de normalité





Plan de la présentation

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Lien entre deux variables qualitatives

La réponse à une question dépend-t-elle du sexe ? de la CSP ?

Etude du lien entre sexe et couleur des cheveux

- $\left\{ \begin{array}{l} H_0 : \text{sexe et couleur des cheveux indépendants} \\ H_1 : \text{sexe et couleur des cheveux non indépendants} \end{array} \right.$

		Couleur des cheveux					
		Blond	Roux	Châtain	Brun	Noir de Jais	
Sexe	Garçon	592	119	849	504	36	2100
	Fille	544	97	677	451	14	1783
		1136	216	1526	955	50	3883

Plan du module

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Tableau de contingence

1ère variable	2ème variable						
	B_1	B_2	...	B_j	...	B_J	
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2J}	$n_{2\bullet}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}	$n_{i\bullet}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
A_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}	$n_{I\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet j}$...	$n_{\bullet J}$	$n = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$



Lecture de la dépendance sur le tableau de contingence

Profils lignes

		Couleur des cheveux					
Sexe		Blond	Roux	Châtain	Brun	Noir de Jais	
Garçon		0.28	0.06	0.40	0.24	0.02	1
Fille		0.31	0.05	0.38	0.25	0.01	1

Profils colonnes

		Couleur des cheveux				
Sexe		Blond	Roux	Châtain	Brun	Noir de Jais
Garçon		0.52	0.55	0.56	0.53	0.72
Fille		0.48	0.45	0.44	0.47	0.28
		1	1	1	1	1

Plan du module

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Indépendance

Si A_i et B_j sont indépendants

$$\underbrace{\mathbb{P}(A_i \cap B_j)}_{\pi_{ij}} = \underbrace{\mathbb{P}(A_i)}_{\pi_i} \underbrace{\mathbb{P}(B_j)}_{\pi_j}$$

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet} n_{\bullet j}}{n n}$$

Effectifs théoriques

$$t_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$



Lien entre sexe et couleur des cheveux : test d'indépendance

Effectifs théoriques

Sexe	Couleur des cheveux					
	Blond	Roux	Châtain	Brun	Noir de Jais	
Garçon	614.37	116.82	825.29	516.48	27.04	2100
Fille	521.63	99.18	700.71	438.52	22.96	1783
	1136	216	1526	955	50	3883

Statistique de test

$$D^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - t_{ij})^2}{t_{ij}} \underset{\text{H}_0 \text{ vraie et } n \text{ grand}}{\sim} \chi_{(I-1)(J-1)}^2$$

Exemple : $D^2 = 10.47$, probabilité critique = 0.033





Analyse de la dépendance

Contributions à la dépendance

		Couleur des cheveux					
Sexe	Blond	Roux	Châtain	Brun	Noir de Jais		
Garçon	7.78	0.39	6.51	2.88	28.36	46	
Fille	9.17	0.46	7.66	3.39	33.40	54	
	16.9	0.9	14.2	6.3	61.8	100	

Plan de la présentation

- 1 Introduction
- 2 Test d'ajustement à une loi théorique
 - Modèle multinomial
 - Modèle pour variable de comptage
 - Modèle pour variable continue
- 3 Test d'indépendance
 - Tableau de contingence
 - Construction du test d'indépendance
- 4 Prolongements



Dépouillement d'enquête

Etude consommateur *sensibilité du cuir chevelu* d'un groupe cosmétique

	Ethnie	Sexe	Sensible	< 2 shamp./sem	Séchage brosse	..
1	Indiens	M	Très sensible	Non	Non	
2	Indiens	M	Sensible	Non	Non	
3	Indiens	F	Sensible	Non	Non	
4	Indiens	F	Non sensible	Non	Non	
⋮	⋮	⋮	⋮	⋮	⋮	
4560	Cauc. Am.	F	Sensible	Non	Non	



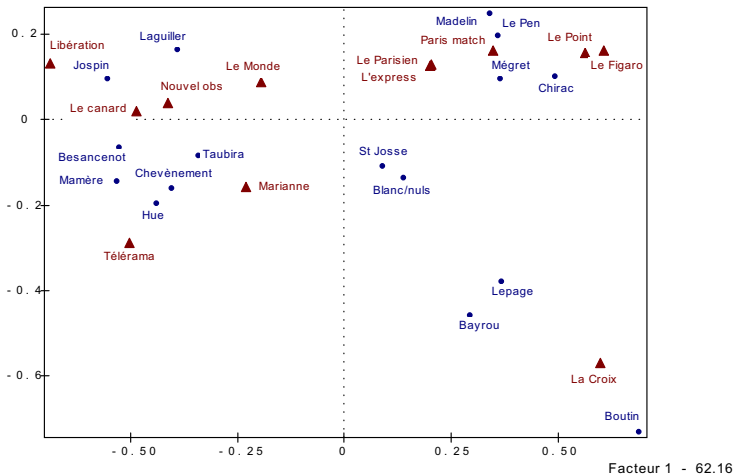
Analyse des correspondances

Vote	Magazine					...
	La Croix	Figaro	Libé	Le Monde	Le Parisien	
Laguiller	0	2	6	5	4	...
Besancenot	2	2	8	6	3	...
Hue	2	0	3	2	4	...
Jospin	3	7	41	26	12	...
⋮	⋮	⋮	⋮	⋮	⋮	



Analyse des correspondances

Facteur 2 - 14.24 %





Bilan

Analyses univariées		
	Exploratoire	Inférentiel
Quanti.	\bar{x}, S'	Intervalle de confiance Tests de conformité
Quali.	$\hat{\pi}_j$	Intervalle de confiance Tests du χ^2 d'ajustement
Analyses bivariées		
	Exploratoire	Inférentiel
Quanti \times Quali	η^2	Comparaison de populations Analyse de variance
Quanti \times Quanti	r_{xy} ou R^2	Régression linéaire simple
Quali \times Quali	D^2	Test du χ^2 d'indépendance