adéquation statistique

discrépance

la distribution multinomiale (fréq. obs.)

niveau de signification

distribution associée à la discrépance : loi du χ^2

grand nombre d'événements

carré d'un variable normale et χ^2

calcul exact

Eléments de statistiques pour les data sciences

Cours 4 : La distribution du χ^2 de Pearson

Dr. Ph. Müllhaupt

IGM - EPFL

—

adéquation statistique

discrépance

la distribution multinomiale (fréq. obs.)

niveau de signification

distribution associée à la discrépance : loi du χ^2

grand nombre d'évène-ments

carré d'un variable normale et χ^2

Plan

1. adéquation statistique
2. discrépance
3. la distribution multinomiale (fréq. obs.)
4. niveau de signification
5. distribution associée à la discrépance : loi du χ^2
6. grand nombre d'évènements
7. carré d'un variable normale et χ^2
8. calcul exact du niveau de signification dans le cas binomial

qualité d'ajustement, adéquation statistique

goodness-of-fit

On partitionne le résultat de l'expérience en plusieurs classes, disons m classes distinctes. L'expérience consiste ainsi en un étiquetage de n évèvements [1] selon m classes. Chaque évènement correspond à une des m classes, et chaque classe contient un des n évènements.

classe d'évènements	A_1	A_2	...	A_m	total
nombre d'évènements observés	f_1	f_2	...	f_m	n
nombre d'évènements théoriques	e_1	e_2	...	e_m	n

Il s'agit de déterminer l'adéquation statistique entre la probabilité théorique de l'apparition de chacune des classes avec la fréquence avec laquelle les évènements ont été observés.

1. au sens classique du terme et non au sens probabiliste

discrépance D

La discrépance est une mesure entre la différence entre le nombre théorique et le nombre observé des évènements dans chacune des classes. La différence au carré (au lieu de la valeur absolue, i.e. médiane) est considérée. On ne prend pas la racine carrée et on divise également la somme des carrés de chaque écart, par le nombre théorique des évènements afin d'avoir une mesure à l'unité.

$$\frac{(\text{ nombre d'évènements observés } - \text{ nombre d'évènements théoriques})^2}{\text{nombre d'évènements théoriques}}$$

ce qui se traduit mathématiquement par

définition

$$D = \sum_{k=1}^{m} \frac{(X_k - e_k)^2}{e_k} \qquad d = \sum_{k=1}^{m} \frac{(f_k - e_k)^2}{e_k}$$

exemple : 1 lancé de 10 dès

10 dés de 6 faces chacunes, numérotées de 1 à 6, sont lancés. On compte les évènements associés aux faces qui sont apparues. Il y a donc $m = 6$ classes et $n = 10$ évènements. Un résultat d'un lancé est donné dans le tableau :

face (classe)	1	2	3	4	5	6	total
f_k	2	2	3	1	0	2	10
e_k	$1.\bar{6}$	$1.\bar{6}$	$1.\bar{6}$	$1.\bar{6}$	$1.\bar{6}$	$1.\bar{6}$	10
$\frac{(f_k-e_k)^2}{e_k}$	$0.0\bar{6}$	$0.0\bar{6}$	$1.0\bar{6}$	$0.2\bar{6}$	$1.\bar{6}$	$0.0\bar{6}$	$\boxed{3.2}$

$$\begin{aligned} d &= 0.6 \times \left[3 \times (2 - 1.\bar{6})^2 + (3 - 1.\bar{6})^2 + (1 - 1.\bar{6})^2 + (0 - 1.\bar{6})^2\right] \\ &= 3.2 \end{aligned}$$

exemple : 12 lancés de 10 dès

face (classe)	1	2	3	4	5	6	total
	2	2	3	1	0	2	10
	0	1	3	4	1	1	10
	1	1	3	0	1	4	10
	2	3	0	2	2	1	10
	1	3	0	3	2	1	10
	2	1	1	3	1	2	10
	1	1	2	1	2	3	10
	3	2	0	1	2	2	10
	4	2	2	2	0	0	10
	0	1	2	2	3	2	10
	1	1	4	1	1	2	10
	1	0	2	2	3	2	10
	18	18	22	22	18	22	120

exemple : 12 lancés de 10 dès

face (classe)	1	2	3	4	5	6	total
f_k	18	18	22	22	18	22	120
e_k	20	20	20	20	20	20	120
$\frac{(f_k-e_k)^2}{e_k}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	1.2

$$d = \sum_{k=1}^{6} \frac{4}{20} = 1.2$$

la distribution multinomiale

rappel, la distribution multinomiale, $n = n_1 + n_2 + \ldots n_m$

$$\begin{aligned} \sum_{k=1}^{m} n_k &= n \\ P(n_1, n_2, \ldots, n_m) &= \left(\begin{array}{c} n \\ n_1 n_n \cdots n_m \end{array} \right) \prod_{k=1}^{m} p_k^{n_k} \\ &= \frac{n!}{n_1!\, n_2! \cdots n_m!} \prod_{k=1}^{m} p_k^{n_k} \end{aligned}$$

la distribution multinomiale

cas particulier : binomiale

$$
\begin{aligned}
m &= 2 \\
p_1 &= p \\
p_2 &= (1-p) \\
n_2 &= n - n_1
\end{aligned}
$$

$$P(n_1, n_2) = \frac{n!}{n_1!\,(n-n_1)!} p^k (1-p)^{n-k}$$

Hypothèse statistique, condition expérimentale

On suppose connu les probabilités de l'apparition de chacune des classes k parmis les m classes. Autrement dit, on suppose connues les probabilités :

$$p_k \qquad k = 1 \cdots m$$

Proposition

La probabilité d'apparition des évènements (résultats) observés f_1, f_2, ..., f_m est donnée par la distribution multinomiale :

$$\begin{aligned} n &\triangleq \sum_{k=1}^{m} f_k \\ P(f_1, f_2, \ldots, f_m) &= \binom{n}{f_1 f_2 \cdots f_m} p_1^{f_1} p_2^{f_2} \cdots p_m^{f_m} \end{aligned} \tag{1}$$

2 classes

le cas $m = 2$, $n = f_1 + f_2$

$$
\begin{aligned}
n &= n_1 + n_2 \\
P(f_1, f_2) &= \binom{n}{f_1 \; f_2} p_1^{f_1} p_2^{f_2} \\
&= \frac{n!}{f_1!\,(n-f_1)!} p_1^{f_1} (1-p_1)^{n-f_1}
\end{aligned}
$$

le niveau de signification

Le niveau de signification est défini dans notre contexte comme la probabilité que D dépasse la valeur d

Niveau de signification NS également noté α

$$\boxed{\text{NS} = \alpha \triangleq P(D \geq d)}$$

le niveau de signification

- Lorsque NS est grand (e.g. 20 %), une discrépance aussi grande que celle observée d se manifesterait fréquemment par suite de l'effet de variations dues au hasard si le modèle probabiliste est correct. Il n'y a donc dans les données observées rien d'inconsistent avec le modèle.
- Par contre, si le niveau de signification NS est petit (e.g. 1 %) alors une telle discrépance observée
  d apparaîtra rarement si le modèle probabiliste est correct. Il y donc beaucoup d'évidence contre le modèle utilisé.

distribution associée à la discrépance : χ^2

dans le cas $e_j >> 1$

hypothèses et paramètres

- On suppose que le modèle de probabilité est correct, i.e. p_1, p_2, ..., p_n est représentatif des probabilités de chacune des classes de la loi multinomiale.
- Le nombre de classes théoriques (i.e. le nombre attendu des fréquences des classes) e_j sont tous grands, i.e. $e_j >> 1$, $\forall j$.
- Le paramètre r est le nombre de paramètres estimés à partir du résultat observé.
- le nombre de degrés de liberté est défini par

$$\nu = m - r - 1$$

- r est le nombre de paramètres estimé à partir des données observées.
- m est le nombre de classes utilisées pour calculer d.

distribution de la discrépance

$P(D \geq d)$

Sous les hypothèse mentionnées nous avons

théorème

$$\boxed{\alpha = \mathsf{NS} = P(D \geq d) \approx P\{\chi^2_{(m-r-1)} \geq d\}}$$

Rappel : définition de la distribution du $\chi^2_{(\nu)}$

$$\boxed{f(x) \triangleq \frac{1}{2^{\frac{\nu}{2}}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}$$

Exemple

1 lancé de 10 dés

face (classe)	1	2	3	4	5	6	total
f_k	2	2	3	1	0	2	10
e_k	$1.6\bar{6}$	$1.6\bar{6}$	$1.6\bar{6}$	$1.6\bar{6}$	$1.6\bar{6}$	$1.6\bar{6}$	10
$\frac{(f_k-e_k)^2}{e_k}$	$0.0\bar{6}$	$0.0\bar{6}$	$1.0\bar{6}$	$0.2\bar{6}$	$1.\bar{6}$	$0.0\bar{6}$	$\boxed{3.2}$

$$\text{NS} = \alpha = P(D \geq 3.2) \approx P\{\chi^2_{(5)} \geq 3.2\}$$

Exemple

utilisation de la table du χ^2

	probabilité qu'une valeur supérieure soit atteinte								
ν	0.995	0.990	0.950	0.975	0.900	0.750	0.500	0.250	0.100
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24

$$2.67 < \chi^2_{(5)} < 4.35 \qquad 0.750 \geq \mathsf{NS} = \alpha \geq 0.500$$

Exemple

12 lancés de 10 dés

face (classe)	1	2	3	4	5	6	total
f_k	18	18	22	22	18	22	120
e_k	20	20	20	20	20	20	120
$\frac{(f_k - e_k)^2}{e_k}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	$\frac{4}{20}$	1.2

$$\text{NS} = \alpha \approx P\{\chi^2_{(5)} \geq 1.2\}$$

Exemple

utilisation de la table du χ^2

	probabilité qu'une valeur supérieure soit atteinte								
ν	0.995	0.990	0.950	0.975	0.900	0.750	0.500	0.250	0.100
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24

$$1.15 < \chi^2_{(5)} < 1.61 \qquad 0.975 \geq \mathrm{NS} = \alpha \geq 0.900$$

Exemple

discussion

Dans les deux cas, le niveau de signification est grand. Il n'y a pas d'évidence contre le modèle probabiliste d'un dé avec $p = \frac{1}{6}$ pour chacune des faces.

- dans les deux cas $r = 0$, car on a pas utiliser les données observées pour calculer un paramètre.
- le nombre de degrés de libertés est $\nu = m - r - 1 = 6 - 0 - 1 = 5$. La raison d'avoir 5 au lieu de 6 est expliqué par le fait que le nombre observé dans la 6ème classe (f_6) est connu dés que l'on connait le nombre de chacune des autres faces (f_1, f_2, f_3, f_4, et f_5). Il y a la contrainte $\sum_{i=1}^{6} f_i = n$.

- La première expérience (1 seul jet) est un peu "douteuse" concernant l'hypothèse pour appliquer la loi du χ^2 pour approximer la distribution de la discrépance dans le cas multinomial, car les $e_j = 1.666$ sont tous petits.
- La deuxième expérience est meilleure concernant cette hypothèse car $e_j = 20$.

La deuxième expérience donne peu d'évidence contre $\frac{1}{6}$ de chance par face, car $0.900 \leq \text{NS} \leq 0.975$ est grand.

carré d'une loi normale et χ^2

théorème

Si $X \sim \mathcal{N}(0,1)$ et $Z = X^2$, alors

$$Z \sim \chi^2_{(1)}$$

démonstration

Comme X^2 n'est pas monotone, on ne peut pas appliquer la transformation $Z = X^2$ avec

$$h(x) = f(x) \left| \frac{\partial g}{\partial x} \right|$$

On va calculer à partir des fonctions de répartition et ensuite dériver.

carré d'une loi normale et χ^2

démonstration

$$Z = X^2 \qquad \text{avec } X \sim \mathcal{N}(0,1)$$

$$\begin{aligned} F(z) &= P(Z \leq z) = P(X^2 \leq z) \\ &= P(|X| \leq \sqrt{z}) \\ &= P(-\sqrt{z} \leq X \leq \sqrt{z}) \\ &= G(\sqrt{z}) - G(-\sqrt{z}) \end{aligned}$$

avec G la fonction de répartition de $\mathcal{N}(0,1)$

carré d'une loi normale et χ^2

démonstration

$$
\begin{aligned}
f(z) &= \frac{d}{dz}F(z) = \frac{d}{dz}G(\sqrt{z}) - \frac{d}{dz}G(-\sqrt{z}) \\
\frac{d}{dx}G(x) &= g(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \\
f(z) &= g(\sqrt{z})\frac{d\sqrt{z}}{dz} - g(-\sqrt{z})\frac{d(-\sqrt{z})}{dz} \\
&= \frac{1}{\sqrt{2\pi}}z^{-\frac{1}{2}}e^{-\frac{1}{2}z} \\
&= \frac{1}{2^{\frac{1}{2}}\Gamma(1/2)}z^{\frac{1}{2}-1}\,e^{-z/2}
\end{aligned}
$$

c'est la densité de probabilité de $\chi^2_{(1)}$ car $\Gamma(1/2) = \sqrt{\pi}$

calcul exact et approximé du niveau de signification

le cas binomial

Pour le cas multinomial et en particulier pour le cas binomial il est possible de calculer

$$\mathrm{NS} = P(D \geq d)$$

- de manière exacte, sans utiliser la distribution du χ^2, en utilisant le distribution multinomiale directement. Nous allons illustrer ceci dans le cas binomial.
- de manière approximatif, sans utiliser la distribution du χ^2 en utilisant, dans le cas binomial, le théorème de Laplace-De Moivre qui approxime la distribution binomiale par une distribution normale.

Nous allons illustrer la marche à suivre.

le cas binomial

$m = 2$

Lorsque $m = 2$ il y a seulement deux classes, que l'on peut appeler "réussite" et "échec". Soit X le nombre de réussites et $n - X$ le nombre d'échecs.

Les valeurs attendues (nombre d'évènements théoriques) sont np réussites et $n(1 - p)$ échecs, où p désigne la probabilité de réussite qui dépend du modèle sous-jacent.

le cas binomial

$m = 2$

classe	réussite	échec	total
fréquence observée f_k	X	$n - X$	n
fréquence théorique e_k	np	$n(1-p)$	n

la discrépance D peut se mettre sous la forme

$$\begin{aligned} D &= \frac{(X-np)^2}{np} + \frac{((n-X)-n(1-p))^2}{n(1-p)} \\ &= \boxed{\frac{(X-np)^2}{np(1-p)}} \end{aligned}$$

cas binomial

distribution de la discrépance et approximation

D est exactement distribuée selon la loi binomiale

$$P(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \qquad \text{pour } x = 0, 1, \ldots, n$$

Le théorème de Laplace - De Moivre donne lorsque np et $n(1-p)$ sont grands une approximation par une loi normale de moyenne $\mu = np$ et de variance $\sigma^2 = np(1-p)$, i.e. $\mathcal{N}(np, np(1-p))$ et si on effectue un changement de variables

$$\frac{X - np}{\sqrt{np(1-p)}} \approx \mathcal{N}(0, 1)$$

distribution de la discrépance

cas binomial

Rappel : le carré d'une variable $\mathcal{N}(0,1)$ est distribué selon $\chi^2_{(1)}$. La discrépance tombe, dans le cas binomial, dans ce cas car :

$$D = \frac{(X - np)^2}{np(1-p)} \approx \chi^2_{(1)}$$

à condition que les fréquences observées sont grandes. Nous avons donc le niveau de signification

$$\text{NS} = P(D \geq d) \sim P\{\chi^2_{(1)} \geq d\}$$

et on tombe sur un cas particulier de ce que l'on a déjà vu avec $m = 2$, er $r = 0$ (mais cette fois-ci avec la démonstration si on admet Laplace - De Moivre).

exemple

calcul exact du NS dans le cas binomial

On a choisit un jury composé de 82 personnes choisies au hasard dans une population homogène entre hommes et femmes.

classe	hommes	femmes	total
fréquence observée	58	24	82
fréquence théorique	41	41	82

$$D = \frac{(X-41)^2}{20.5}; \qquad d = \frac{(58-41)^2}{20.5} = \frac{17^2}{20.5} = 14.10$$

$$\text{NS} = P(D \geq d) = P\left\{\frac{(X-41.5)^2}{20.5} \geq \frac{17^2}{20.5}\right\} = P(|X-41| \geq 17)$$

exemple

calcul exact du NS dans le cas binomial

comme $58 - 17 = 41$ et $24 + 17 = 41$ on a

$$\begin{aligned} \text{NS} &= P(X \geq 58) + P(X \leq 24) \\ &= \sum_{x \geq 58} P(x) + \sum_{x \leq 24} P(x) \end{aligned}$$

avec

$$P(x) = \left(\begin{array}{c} 82 \\ x \end{array} \right) \left(\frac{1}{2} \right)^{82}$$

On commence par

$$P(24) = P(58) = \left(\begin{array}{c} 82 \\ 58 \end{array} \right) 0.5^{82} = 6.74 \cdot 10^{-5}$$

exemple

calcul exact du NS dans le cas binomial

relations $P(59) = \frac{24}{59} P(58)$ et symétrie $P(X \geq 58) = P(X \leq 24)$

$$
\begin{aligned}
P(23) = P(59) &= \frac{24}{59} P(58) = 2.7419 \cdot 10^{-5} \\
P(22) = P(60) &= \frac{23}{60} P(59) = 1.051 \cdot 10^{-5} \\
P(21) = P(61) &= \frac{22}{61} P(60) = 3.791 \cdot 10^{-6} \\
P(20) = P(62) &= \frac{21}{62} P(61) = 1.2839 \cdot 10^{-6} \\
P(19) = P(63) &= \frac{20}{63} P(62) = 4.076 \cdot 10^{-7} \\
P(18) = P(64) &= \frac{19}{64} P(63) = 1.210 \cdot 10^{-7} \\
P(17) = P(65) &= \frac{18}{65} P(64) = 3.351 \cdot 10^{-8}
\end{aligned}
$$

exemple

calcul exact du NS dans le cas binomial

en effectuant le somme et en doublant pour la symétrie

$$\begin{aligned} NS &= P(x \geq 58) + P(x \leq 24) = 2 \cdot P(x \geq 58) = 2 \cdot 1.11 \cdot 10^{-}4 \\ &\approx 2.22 \cdot 10^{-4} = 0.000222 \end{aligned}$$

La chance d'obtenir un tel déséquilibre dans la répartition hommes-femmes est très petite, et il est presque certain que la répartition des membres du jury n'a pas été faite de manière équilibrée, de manière non biaisée.

exemple

calcul approximé de NS par la loi normale dans le cas binomial

L'approximation consiste dans le calcul

$$\text{NS} \approx P\{\chi^2_{(1)} \geq 14.10\} = P\{|Z| \geq \sqrt{14.10}\} = P\{|Z| \geq 3.755\}$$

On est en dehors de la table du χ^2 pour $\nu = 1$.

calcul approximé de NS par la loi normale dans le cas binomial

En ce qui concerne l'utilisation de la table de la loi normale, en ligne 3.7 (deuxième page de la table de la loi normale) et colonnes 5 et 6 on a $P(0 < Z \leq 3.755) = 0.4999$ ce qui donne

$$\begin{aligned} \text{NS} &\approx P(|Z| \geq 3.755) \\ &= 1 - 2 \cdot P(0 < Z \leq 3.755) \\ &= 1 - 2 \cdot 0.4999 = 0.0002 \end{aligned}$$

ce qui est très proche de ce que l'on a calculé de manière exacte $\text{NS} = 0.000222$. L'approximation de la binomiale par la loi normale très bonne dans cet exemple.