

Statistique



Compléments au polycopié

http://perso.ensem.inpl-nancy.fr/Gerard.Vinsard/enseignement_2006/doc/prep_Statistiques.pdf

G. Vinsard

Gerard.Vinsard@univ-lorraine.fr

4 avril 2013

Indépendance de X et $N S^2$: le vecteur de V.A.

► $X_1, \dots, X_N$ sont N variables aléatoires, indépendantes et identiquement distribuées selon la loi de Gauss-Laplace de moyenne μ et d'écart-type σ . Le vecteur formé par les X_n

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}$$

a une densité de probabilité

$$\rho_X(x) = \left(\frac{1}{\sqrt{2\pi} \sigma} \right) \exp^{-\frac{|x-\mu \mathbb{I}|^2}{2 \sigma^2}} \text{ avec } \mathbb{I} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ et } \underbrace{\left| \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \right|}_{= x} = \sum_{n=1}^N x_n^2$$

c'est à dire que

$$\mathbb{P} \left(X \in \begin{pmatrix} [x_1, x_1 + \delta x_1] \\ \vdots \\ [x_N, x_N + \delta x_N] \end{pmatrix} \right) = \rho_X(x) \delta x_1 \dots \delta x_N$$

Indépendance de X et $N S^2$: transformation du vecteur

► Pour R , une matrice de rotation dans $\mathbb{R}^N$ quelconque mais telle que

$${}^t R = R^{-1}$$

le vecteur X peut être mis en correspondance avec un autre vecteur Y comme

$$X = R (Y - \mu \mathbb{I}) + \mu \mathbb{I} \iff Y = {}^t R (X - \mu \mathbb{I}) + \mu \mathbb{I}$$

La densité de probabilité du vecteur de variables aléatoires Y est

$$\rho_Y(y) = \rho_X(R (y - \mu \mathbb{I}) + \mu \mathbb{I}) \underbrace{|\det R|}_{=1}$$

soit donc

$$\rho_Y(y) = \left(\frac{1}{\sqrt{2\pi} \sigma} \right) \exp^{-\frac{|y - \mu \mathbb{I}|^2}{2 \sigma^2}}$$

qui est le produit des densités de probabilité des Y_n : ces variables aléatoires sont donc indépendantes.

Indépendance de X et $N S^2$: calculs algébriques

► Le choix d'une matrice R telle que sa première colonne soit

$$\frac{1}{\sqrt{N}} \mathbb{I}$$

correspond à

$$Y_1 = \sqrt{N} \underbrace{\frac{1}{N} \sum_{n=1}^N X_n}_{= \bar{X}}$$

et donc

$$\begin{aligned} \sum_{n=1}^N (X_n - \bar{X})^2 &= |X - \bar{X} \mathbb{I}|^2 = \left| R(Y - \mu \mathbb{I}) + \mu \mathbb{I} - Y_1 \frac{\mathbb{I}}{\sqrt{N}} \right|^2 \\ &= \left| R(Y - \mu \mathbb{I}) - R\left((Y_1 - \sqrt{N} \mu) \frac{\mathbb{I}}{\sqrt{N}}\right) \right|^2 \\ &= \sum_{n=2}^N (Y_n - \mu)^2 \end{aligned}$$

Indépendance de X et $N S^2 : \chi^2$ et S.F.

► Et donc

$$N S^2 = \sum_{n=1}^N \frac{(X_n - \bar{X})^2}{\sigma^2} = \sum_{n=2}^N \frac{(Y_n - \mu)^2}{\sigma^2}$$

La somme des N carrés de variables aléatoires de Gauss-Laplace non indépendantes (et de toute façon pas réduites) du premier membre est égale à la somme des carrés de $N - 1$ variables aléatoires de Gauss-Laplace centrées et réduites du second membre ; c'est à dire à **une variable aléatoire du χ^2 à $N - 1$ ddl.**

► D'autre part

$$\bar{X} = \sqrt{N} Y_1$$

est indépendante de S^2 qui est une fonction des $Y_2, \dots, Y_N$, et donc

$$\frac{\bar{X}}{\sqrt{N S^2}}$$

suit une loi de Student-Fisher à $N - 1$ ddl.

Indépendance de $\bar{X}$ et $N S^2$: remarque contrefactuelle

- Le fait que les X_n soient des variables aléatoires de Gauss-Laplace est un élément important.
- Si elles étaient par exemple des variables aléatoires de Bernoulli, pour le cas $N = 2$

$$\bar{X} = \frac{1}{2}(X_1 + X_2) \text{ et } N S^2 = \frac{1}{2} (X_1 - X_2)^2$$

Si X_1 et X_2 sont indépendantes alors le tableau des probabilités est

X_2	X_1		
		0	1
0	$(1 - \bar{w})^2$	$\bar{w} (1 - \bar{w})$	
1	$\bar{w} (1 - \bar{w})$	$\bar{w}^2$	

celui de $2 \bar{X}$ et $2 N S^2$ est alors

$2 N S^2$	$2 \bar{X}$	0	1	2
		$(1 - \bar{w})^2$	0	$\bar{w}^2$
0		0	$2 \bar{w} (1 - \bar{w})$	0
1				

qui montre que $\bar{X}$ et $N S^2$ ne sont pas indépendantes.

Test d'indépendance de deux variables qualitatives

► Il s'agit de répondre à des questions comme :

On a examiné les cheveux et les sourcils de 46742 individus qu'on a classé comme suit ¹

Sourcils	Cheveux		
		Clairs	Foncés
Clairs		30472	3238
Foncés		3364	9468

Les couleurs de cheveux et sourcils sont-elles indépendantes ?

1. D'après un exercice de C. Chambon, les données sont sans doute réelles.

Les cheveux et les sourcils

► Si elles le sont alors en introduisant les notations N_{pq} , $N_{.q}$, $N_{p.}$ comme

S. \ C.	Clairs	Foncés	totaux
Clairs	$N_{11} = 30472$	$N_{12} = 3238$	$N_{1.} = 30472 + 3238 = 33710$
Foncés	$N_{21} = 3364$	$N_{22} = 9468$	$N_{2.} = 3364 + 9468 = 12832$
Totaux	$N_{.1} = 33836$ $= 3364 + 30472$	$N_{.2} = 12706$ $= 9468 + 3238$	$N = N_{.1} + N_{.2} = N_{1.} + N_{2.}$ $= 46542$

on obtient le tableau des 'effectifs théoriques'

Sourcils \ Cheveux	Clairs	Foncés
Clairs	$\frac{N_{1.} N_{.1}}{N} = 24507$	$\frac{N_{1.} N_{.2}}{N} = 9202$
Foncés	$\frac{N_{2.} N_{.1}}{N} = 9328$	$\frac{N_{2.} N_{.2}}{N} = 3503$

qui est à comparer au tableau des effectifs réels.

Le test *ad hoc* (1/2)

- Si $\overline{w}_c$, $\overline{w}_s$ les probabilités d'avoir les cheveux clairs et d'avoir les sourcils clairs étaient connues et que ces deux facteurs soient indépendants alors le tableau des effectifs théoriques serait

		Cheveux	
Sourcils		Clairs	Foncés
Clairs		$N \overline{w}_s \overline{w}_c$	$N \overline{w}_s (1 - \overline{w}_c)$
Foncés		$N (1 - \overline{w}_s) \overline{w}_c$	$N (1 - \overline{w}_s) (1 - \overline{w}_c)$

et

$$\frac{(N_{11} - N \overline{w}_s \overline{w}_c)^2}{N \overline{w}_s \overline{w}_c} + \frac{(N_{12} - N \overline{w}_s (1 - \overline{w}_c))^2}{N \overline{w}_s (1 - \overline{w}_c)} + \frac{(N_{21} - N \overline{w}_c (1 - \overline{w}_s))^2}{N \overline{w}_c (1 - \overline{w}_s)} + \frac{(N_{22} - N (1 - \overline{w}_c) (1 - \overline{w}_s))^2}{N (1 - \overline{w}_c) (1 - \overline{w}_s)}$$

serait la réalisation d'une loi du χ^2 à 4 - 1 ddl (le nombre N est connu).

- Mais $\overline{w}_c$ et $\overline{w}_s$ ne sont pas connues, elles doivent donc être estimées par

$$\overline{w}_c^* = N_{.1}/N ; \quad \overline{w}_s^* = N_{1.}/N$$

ce qui enlève encore deux degrés de liberté.

Le test *ad hoc* (2/2)

► Il vient alors que si les deux caractères sont indépendants la quantité

$$\sum_{p=1}^2 \sum_{q=1}^2 \frac{\left(N_{pq} - \frac{N_{p.} N_{.p}}{N} \right)^2}{\frac{N_{p.} N_{.p}}{N}}$$

est une réalisation d'une loi du χ^2 à $4 - 2 - 1 = 1$ ddl.

D'une façon générale, si le tableau comportait P lignes et Q colonnes (il y aurait plus de détails sur les couleurs des cheveux et sourcils), et que les deux caractères soient indépendant alors

$$\sum_{p=1}^P \sum_{q=1}^Q \frac{\left(N_{pq} - \frac{N_{p.} N_{.p}}{N} \right)^2}{\frac{N_{p.} N_{.p}}{N}}$$

serait une réalisation d'une loi du χ^2 à $(P - 1)(Q - 1)$ ddl.

Le résultat

$$\sum_{p=1}^2 \sum_{q=1}^2 \frac{\left(N_{pq} - \frac{N_{p.} N_{.p}}{N} \right)^2}{\frac{N_{p.} N_{.p}}{N}} = 19288$$

la valeur seuil à 0.5% du χ^2 à 1 ddl est 7.88. La conclusion est que les couleurs de cheveux et sourcils ne sont pas indépendantes

(et de loin : R donne 0 pour le risque à droite de 19288).

Approche bayésienne des sondages

- Une population est composée d'individus qui votent pour 'A' en proportion $\bar{w}$ et pour 'B' en proportion $1 - \bar{w}$;
- Si les individus d'un échantillon de taille N sont interrogés sur leur choix, la variable aléatoire X = 'nombre d'individus déclarant voter pour A' suit une loi binomiale

$$C_N^n \bar{w}^n (1 - \bar{w})^{N-n}$$

- Cette probabilité peut être réinterprétée en termes de probabilité conditionnelle. D'abord on introduit une variable aléatoire $\bar{\Omega}$ 'valeurs possibles de $\bar{w}$ ' et ensuite

$$\mathbb{P}(X = n / \bar{\Omega} = u) = C_N^n u^n (1 - u)^{N-n}$$

C'est la probabilité que X vaille n sachant que $\bar{\Omega}$ vaut u .

Approche bayésienne des sondages

- La grandeur intéressante est

$$\mathcal{P} = \mathbb{P}(\bar{\Omega} > \frac{1}{2} / X = n)$$

c'est la probabilité qu'il y ait plus de 50% d'électeurs de 'A' dans la population sachant que l'échantillon de taille N en contient n .

- Si un nombre Q est donnée, elle se réécrit

$$\mathcal{P} = \mathbb{P} \left(\bigcup_{q=0}^{Q-1} \left(\frac{1}{2} + \frac{q}{2Q} < \bar{\Omega} < \frac{1}{2} + \frac{q+1}{2Q} \right) / X = n \right)$$

soit

$$\mathcal{P} = \sum_{q=0}^{Q-1} \mathbb{P} \left(\left(\frac{1}{2} + \frac{q}{2Q} < \bar{\Omega} < \frac{1}{2} + \frac{q+1}{2Q} \right) / X = n \right)$$

et donc en utilisant la formule d'inversion des probabilités conditionnelles

$$\mathcal{P} = \sum_{q=0}^{Q-1} \frac{\mathbb{P} \left(X = n / \left(\frac{1}{2} + \frac{q}{2Q} < \bar{\Omega} < \frac{1}{2} + \frac{q+1}{2Q} \right) \right) \mathbb{P} \left(\frac{1}{2} + \frac{q}{2Q} < \bar{\Omega} < \frac{1}{2} + \frac{q+1}{2Q} \right)}{\mathbb{P}(X = n)}$$

Approche bayésienne des sondages

► Lorsque $Q \rightarrow \infty$, si $u = \frac{1}{2} + \frac{q}{2Q}$ et $du = \frac{1}{2Q}$

$$\mathbb{P}\left(\frac{1}{2} + \frac{q}{2Q} < \overline{\Omega} < \frac{1}{2} + \frac{q+1}{2Q}\right) \rightarrow \rho(u) du$$

où ρ est la densité de probabilité qu'il faudra connaître *a priori*.

$$\mathbb{P}\left(X = n / \left(\frac{1}{2} + \frac{q}{2Q} < \overline{\Omega} < \frac{1}{2} + \frac{q+1}{2Q}\right)\right) \rightarrow \mathbb{P}(X = n / \overline{\Omega} = u)$$

$$\sum_{q=0}^Q \rightarrow \int_{\frac{1}{2}}^1$$

► La probabilité $\mathbb{P}(\overline{\Omega} > \frac{1}{2} / X = n)$ s'écrit alors

$$\mathcal{P} = \left(\int_{\frac{1}{2}}^1 \mathbb{P}(X = n / \overline{\Omega} = u) \rho(u) du \right) / \mathbb{P}(X = n)$$

soit en remplaçant $\mathbb{P}(X = n / \overline{\Omega} = u)$ par son expression

$$\mathcal{P} = \left(\int_{\frac{1}{2}}^1 C_N^n u^n (1-u)^{N-n} \rho(u) du \right) / \mathbb{P}(X = n)$$

Approche bayésienne des sondages

► $\mathbb{P}(X = n)$ est la probabilité que N vaille n sans préjuger de la valeur de $\overline{\Omega}$ soit donc

$$\begin{aligned}\mathbb{P}(X = n) &= \int_0^1 \mathbb{P}(X = n / \overline{\Omega} = u) \rho(u) du \\ &= \int_0^1 C_N^n u^n (1 - u)^{N-n} \rho(u) du\end{aligned}$$

► Et donc au final (en simplifiant par C_N^n) la probabilité pour que $\overline{\Omega}$ dépasse 1/2 dans la population est

$$\mathcal{P} = \frac{\int_{\frac{1}{2}}^1 u^n (1 - u)^{N-n} \rho(u) du}{\int_0^1 u^n (1 - u)^{N-n} \rho(u) du}$$

c'est une expression calculable lorsque ρ est donnée **mais toute la difficulté est là.**

Approche bayésienne des sondages

- Si on suppose que les valeurs de $\bar{\Omega}$ sont équiprobables sur $[0, 1]$, $\rho(u) = 1$ et

$$\mathcal{P} = \frac{\int_{\frac{1}{2}}^1 u^n (1-u)^{N-n} du}{\int_0^1 u^n (1-u)^{N-n} du}$$

- La fonction β_{ir} incomplète régularisée est définie comme

$$\beta_{ir}(x, a, b) = \frac{\int_0^x u^{a-1} (1-u)^{b-1} du}{\int_0^1 u^{a-1} (1-u)^{b-1} du}$$

et son algorithme de calcul effectif est implanté un peu partout, il vient

$$\mathcal{P} = 1 - \beta_{ir}(1/2, n+1, N-n+1)$$

Approche bayésienne des sondages

► Numériquement

$$N = 800, n = 428 \longrightarrow \mathcal{P} = 0.976$$

► L'approche traditionnelle est que pour l'hypothèse $\mu < 1/2$ c'est seulement dans 2.5% des cas que n peut être supérieur à

$$\frac{1}{2} N + \frac{1.96 \sqrt{N}}{2} \left(= 427.72 \text{ pour } N = 800 \right)$$

► L'approche Bayésienne conduit aux mêmes conclusions que l'approche traditionnelle lorsqu'on suppose l'équiprobabilité de $\overline{\Omega}$.

Si la probabilité que $\overline{\Omega}$ soit plus grande que $1/2$ est de 0.976 c'est qu'il y a moins de $1 - 0.976 = 0.024$ chances d'obtenir X au delà de 428.

Approche bayésienne des sondages

- Mais l'approche Bayésienne permet d'introduire des connaissances *a priori* de la situation en jouant sur la densité ρ .

Dans le cas des sondages par exemple il est presque certain que la densité de $\bar{\Omega}$ n'est pas uniforme sur $[0, 1]$, elle est plutôt plus grande au voisinage de $1/2$ qu'aux extrêmes.

- C'est donc une approche en germe plus fertile.

Néanmoins elle est également plus susceptible de conduire à des trafics divers puisqu'elle repose sur des estimations *a priori*.

- Pour plus d'informations sur la statistique bayésienne voir <http://christophe.genolini.free.fr/recherche/aTelecharger/Bayes.pdf>