chp 3 Test non paramétrique

III/ Comparaisons de proportions

1/ Comparaison d'une proportion à une valeur donnée

Soit un n-éch d'une population présentant un caractère A

On note par X le nombre d'inds ayant le caractère A.

$$X = \sum_{i=1}^{n} \mathbb{1}_A(X_i)$$

si X_i présente A $\mathbb{1}_A(X_i) = 1$ sinon $\mathbb{1}_A(X_i) = 0$.

On note par p la proportion d'individus présentant le caractère A

$\hat{p} = \frac{1}{n} X$ un estimateur sans biais de p.

La question qu'on se pose : La proportion est elle égale à une pr p_0

. Test bilatéral : $H_0 : p = p_0$ vs $H_1 : p \neq p_0$

On rejette H_0 si $u_{ob} = \frac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > u_{1-\alpha/2}$.

on $X \sim B(n, p)$. (somme de bernoulli ind)

$E(X) = np$ et $var(X) = np(1-p)$

pour n assez grand $X \sim N(np, \sqrt{np(1-p)})$ (En général, n et n-...

D'où sous H_0 : $u_{ob} \sim N(0,1)$

$$u_{1-\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2).$$

[Rq : Ds certains cas, (rarement) on utilise la stat de décision suivante

$$u_{obs} = 2\sqrt{n}\left|\arcsin\sqrt{\frac{x}{n}} - \arcsin\sqrt{p_0}\right| > u_{1-\alpha/2} \quad u_{obs} \sim N(0,...$$

arc sin est calculé pour les angles exprimés en radians (pas en degrés

. Test unilatéral $H_0 : p = p_0$ vs $H_1\, p > p_0$.

La règle de décision $u_{obs} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} > u_{1-\alpha}$

[pour le second test $u_{ob} = 2\sqrt{n}\left|\arcsin\sqrt{\frac{x}{n}} - \arcsin\sqrt{p_0}\right| > u_{1-\alpha}$

II/ Comparaison de deux proportions :

Deux échantillons de tailles n_1 et n_2 respec

p_1 La proportion d'inds de l'ech 1 ayant le caractère A

p_2 —— —— 2 —— A

On compare les deux proportions p_1 et p_2.

On peut presenter les resultats ds le tableau de contingence suivant

	Ech 1	Ech 2	Totaux
A	a	b	$a+b$
$\bar{A}$	$n_1 - a$	$n_2 - b$	n_1+n_2-a-b
Tautaux	n_1	n_2	$\boxed{n_1+n_2}$

$H_1 : p_1 \neq p_2$.

Methode asymptotique : les tailles d'ech suffisament grandes.

H_0 : $p_1 = p_1 = p_0$ contre H_1 $p_1 \neq p_2$.

La statistique de decision : $U = \frac{1}{n_1}\sum_{i=1}^{n_1} 1\!\!1_A(x_i) - \frac{1}{n_2}\sum_{i=1}^{n_2} 1\!\!1_A(y_i)$

$RC = [\,|U| > k_\alpha\,]$.

On rejette le fait que $p_1 = p_2$ au risque α pour k_α donnee tq $P(RC) = \alpha$

$U \sim N\left(0, \sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)$ $\quad P\left(U/\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} > k_\alpha/\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = \alpha/2$

Pour un test unilateral H_0 : $p_1 = p_2 = p_0$ contre H_1 : $p_1 > p_2$ (ou $p_1 < p_2$)

au seuil α, on rejette H_0 si $P(U > k_\alpha) = \alpha$.

$$P\left(\frac{U}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > k_\alpha/\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right) = \alpha.$$

Exp

Sur 96 pieces prevenant d'un fournisseur A, il y 12 qui sont deffectueuses et sur 55 pieces provenant d'un autre fournisseur B, il y 15 deffectueuses

Peut on dire, que la proportion des pieces deffectueuses de A est la même que celle de B, au seuil $\alpha = 5\%$.

H_0 : $p_1 = p_2 = p$ contre H_1 : $p_1 \neq p_2$.

$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{96} 1\!\!1_{\{def\}}(x_i) = \frac{12}{96} = \frac{1}{8}$ $\quad \hat{p}_2 = \frac{15}{55} = \frac{3}{11}$

$U = \hat{p}_1 - \hat{p}_2 \sim N\left(0, \sqrt{p(1-p)}\sqrt{\frac{1}{96}+\frac{1}{55}}\right)$

p inconnue on l'estime par $\hat{p} = \frac{12+15}{96+55} = \frac{27}{151} = 0,18$

$P\left(\left|U/\sqrt{p(1-p)\left(\frac{1}{96}+\frac{1}{55}\right)}\right| > k_\alpha\right) = 0,05$ $\quad k_\alpha = F^{-1}_{N(0,1)}\left(1 - \frac{0,05}{2}\right) = 1,96$

$\frac{U}{\sqrt{0,18(1-0,18)\left(\frac{1}{96}+\frac{1}{55}\right)}}_{obs} = \frac{1/8 - 3/11}{\sqrt{}} = 2,28$

... rejette H_0.

270 | 15
140 | 0,1
42

0,05 |
10 | 0,

Test de Cramer Von Mises.

Soit $\{x_1, \dots, x_n\}$ n-éch de X de f.r F

$F^\alpha(x)$ f.r. connue.

H_0 : $F(x) = F^\alpha(x)$ $\forall x$ contre H_1 : $\exists x$ $F(x) \neq F^\alpha(x)$

On suppose que tout les parametres de F^α sont connus.

indicateur d'ecart du test est

$$I = \int_{-\infty}^{+\infty} |F^\alpha(x) - F(x)|^2 \, dF(x)$$

La distribution de cet indicateur a été tabulee. On demontre que.

$$I = \frac{1}{2n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F^\alpha(x_i)\right]^2$$

Où les valeurs de l'ech ordonnées ↗.

Au seuil α, on rejette H_0 si la valeur de I dépasse k_α.

Rq: Le test de Cramer-Von Mises a les mêmes applications que le test de Kolmogorov. La difference entre ces deux tests reside dans le fait que pour le test de ~~C.V.M~~ Kolmogorov seul l'ecart max entre la distribution empirique et la distribution d'ajustement entre en conséderation alors que le test de Cramer V.M, l'indicateur d'ecart prend mieux en compt l'ensemble des données en ce sens que la somme des ecarts intervient. Le test de Kolmogorov, est donc beaucoup plus sensible à l'existence des pts aberrants ds un échantillon que le test de Cramer V.M. On pense generalement que ce dernier est plus puissant, mais cela n'a pas été demontré theoriquement.

2) Test de χ^2

Il permet de comparer la densité de la loi à l'hystogramme construit à partir des obs.

Le pb avec l'hystogramme est le choix tjs arbitraire des classes.

On suppose néanmoins que k classes sont choisies.

Le principe de test de χ^2 est de comparer le pourcentage d'obs observé ds la classe numéro i, que nous noterons par $\hat{p}_i$ au pourcentage d'obs contenues ds cette classe. que nous noterons par p_i

Soient les k classes $C_1 = [a_1\ a_2[\ ; \text{---}\ C_k = [a_k\ a_{k+1}[$

$p_1 = P(C_1) = F_0(a_2) - F_0(a_1)\ ; \text{---}\ ; p_k = P(C_k) = F_0(a_{k+1}) - F_0(a_k)$

Test : $H_0 : F(x) = F_0(x)\ \forall x$ contre $H_1 : \exists x\ F(x) \neq F_0(x)$

La stat de decision $\chi^2_{obs} = \sum_{i=1}^{k} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2$

La distribution de χ^2_{ob} est χ^2_{k-1} $(\sum N_i = n)$

N_i = effectif de la classe C_i.

$N_i \sim N(np_i ; np_i(1-p_i))$ $\left(\frac{N_i - np_i}{\sqrt{np_i(1-p_i)}} \right)^2 \sim N(0,1)$

On admet aussi que si p_i est petit $np_i(1-p_i) \sim np_i$ par suite

$$\sum_{i=1}^{r} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2 \sim \chi^2_{k-1} \quad (n \text{ assez grand})$$

le test de χ^2 est applicable pour $np_i > 10$ ~~generalement~~ sinon on regroupe les classes.

Exp

Soit un dé à 6 faces dont on veut verifier la regularité.

X le nbre apparu apres lancement.

Tester $X \sim \mathcal{U}_{[1, \text{--}, 6]}$ contre $F_X(x) \neq F_{\mathcal{U}}(x)$ pour x

ie $H_0 : p_i = 1/6\ \forall i$ contre $H_1 : \exists i$ tq $p_i \neq 1/6$.

Pour cela on lance le dé 100 fois on obtient

i	1	2	3	4	5	6
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	31

Ces données sont-elles compatibles avec la distr sous H_0 ?

C_i = Face i

$$P_{ef} \sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 24{,}56$$

Au seuil $\alpha = 0{,}05$, on trouve χ^2_5 $k = 11{,}1$

$\chi^2_{ob} > \chi^2_5$ on rejette H_0.

Rq : si sous H_0 la forme de la distribution est connue à r paramètres inconnus. On estime les paramètres. χ^2_{obs} la statistique de décision suit ds ce cas χ^2_{k-1-r}.

III/ Tests d'indépendance de deux caractères.

1) Test de χ^2 et ~~test de Mann~~)

Soit une population sur laquelle on observe de caractères X et Y on ~~s'interesse de~~ propose de tester l'indépendance des deux caractères

On suppose que les inds de la populations peuvent être regroupés en k classes selon X et s classes selon Y.

Rq : si X ou Y sont qualitatifs, les classe sont les modalités

On note par $N_{i\cdot}$ l'effectif de classe C_i sur X f_i fréquence

$N_{\cdot j}$ C_j' sur Y f_j fréquence.

Sous H_0 : $X \perp Y \Leftrightarrow f_{ij} = f_i \times f_j$

La statistique de décision $d^2 = \sum_{i,j} \frac{(f_{ij} - f_i \times f_j')^2}{n \times f_i \times f_j} \sim \chi^2_{(k-1)(s-1)}$

On rejette H_0 au seuil α, si $d > k_\alpha$.

Exp : On interroge 100 personnes de catégories professionnelles X_1 et X_2 et lisant le journal Y_1 ou Y_2. On obtient

A-t-on indépendance entre X et Y au seuil 5%.

Y \ X	X_1	X_2	n_j'
Y_1	27	9	36
Y_2	12	52	64
n_i	39	61	100

... sont gaussiennes.

Indep de X et Y $\Leftrightarrow$ $\rho = 0$

$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ comparer r à 0

Proposition

Si X et Y sont indép^té et si R est le coefficient de correlation empirique on a

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2} \sim T_{n-2}$$

Test $X \perp Y$ $RC = [|R| > k]$.

Rq : on donne .

$$Z = \frac{1}{2} Log\left(\frac{1+r}{1-r}\right) \sim N\left(\frac{1}{2} log \frac{1+\rho}{1-\rho} ; \frac{1}{\sqrt{n-3}}\right)$$

Rq

si ~~X~~ et Y ne sont pas gaussiennes, ce test precedent reste valable mais il test la correlation lineaire non pas l'indépendance.

$$\frac{N_{ij}}{n} \sim N\left(p_{ij}, \sqrt{\frac{np_{ij}}{n^2}}\right) = N\left(p_{ij}, \sqrt{\frac{p_{ij}}{n}}\right)$$

$$\left(\frac{\frac{N_{ij}}{n} - p_{ij}}{\sqrt{\frac{p_{ij}}{n}}}\right)^2 \sim \chi^2_1$$