Comparaison de fréquences

I – Test du X^2

1 – X^2 d'ajustement = d'adéquation = de conformité

Objectifs :

- comparer des fréquences observées à des fréquences théoriques

H_0 : effectifs observés conformes à distribution théorique
H_1 : effectifs observés non conformes à distribution théorique

Sous H_0 :
Ci (*calculés*) et Oi (*observés*)

Si Ci ≥5 : $X_C^2 = \sum_{i=1}^{n} \frac{(Oi-Ci)^2}{Ci}$

α
X^2_{table}
ν = (n-1) ddl

Comparaison :

- Si $X_C^2 > X_t^2$: au risque α, on rejette H_0 : non-conformité
- Si $X_C^2 < X_t^2$: au risque α, on accepte H_0 : on ne montre pas de différence significative entre la distribution observée et la distribution théorique

Exercice 1 :
Mendel → croise des fleurs blanches et rouges :
n = 100
22 rouges
46 roses
32 blanches

Rappel : loi de Mendel ¼ ½ ¼

Sous H_0, Ci :

	Rouges	Roses	Blanches
Oi	22	46	32
Ci	25	50	25

Ci ≥ 5 donc :
$X_c^2 = \frac{(22-25)^2}{25} + \frac{(46-50)^2}{50} + \frac{(32-25)^2}{25} = \frac{9}{25} + \frac{16}{50} + \frac{49}{25} = \frac{132}{50} = 2{,}64$
α = 5%
ν = (3-1) = 2ddl
$X_t^2 = 5{,}99$

$X_C^2 < X_t^2$: au risque 5%, on accepte H_0 : on ne montre pas que la distribution observée ne suit pas la loi de Mendel

Exercice 2 :
Même problème mais n = 400

	Rouges	Roses	Blanches
Oi	88	184	128
Ci	100	200	100

$X_c{}^2 = \frac{(88-100)^2}{100} + \frac{(184-200)^2}{200} + \frac{(128-100)^2}{100} = 10{,}56$

$X_c{}^2 > X_t{}^2$: au risque 5%, on rejette H_0 : la distribution observée ne suit pas la loi de Mendel

2 – X² d'homogénéité = d'indépendance

Objectifs :
- à utiliser quand plusieurs variables qualitatives
- homogénéité de 2 traitements ? (=indépendance ?)
- effectifs observés comparés entre eux (pas de comparaison de pourcentages)

Exemple :
2 échantillons A, B
G = guéris

	G	$\overline{G}$
A	n_1	n_2
B	n_3	n_4

	G	Amélioration	Stabilité	Aggravation
A				
B				
C				

Tableau : lignes x colonnes

$Ci = \frac{Tl \times Tc}{Tg}$ avec Tl : total ligne ; Tc : total colonne ; Tg : total grille

	G		$\overline{G}$		Total
	Oi	Ci	Oi	Ci	
A	45	38,33	55	61,67	100
B	70	76,67	130	123,33	200
Total	115 115		185 185		300

Sous H_0, Ci = effectif théorique guéri avec le traitement A ou B

$X_c{}^2 = \sum_{i=1}^{n} \frac{(Oi\text{-}Ci)^2}{Ci} = \frac{(45-38{,}33)^2}{28{,}33} + \frac{(55-61{,}67)^2}{61{,}67} + \frac{(70-76{,}67)^2}{76{,}67} + \frac{(130-123{,}33)^2}{123{,}33} = 2{,}82$

α = 5%
ν = (l-1)(c-1) = (ligne-1)(colonne-1) = (2-1)(2-1) = 1 ddl (pour les colonnes, Oi et Ci ne comptent que pour 1 colonne et non 2)

$X_t{}^2 = 3{,}84$
$X_c{}^2 < X_t{}^2$: au risque 5%, on accepte H_0 donc on ne montre pas de différence significative

Exercice :
3 groupes A, B, C atteints de leucémie aiguë
3 traitements différents 1, 2 et 3
Observation des symptômes

délai \ traitement	1		2		3		
	Oi	Ci	Oi	Ci	Oi	Ci	
< 1 mois	35	38,5	41	37,7	36	25,8	112
1 à 1,9 mois	16	20,3	17	19,9	26	18,8	59
2 à 3,9 mois	18	17,9	17	17,5	17	16,6	52
4 à 7,9 mois	20	14,8	17	14,5	6	13,7	43
≥ 8 mois	9	6,5	4	6,4	6	6,1	19
	98		96		91		285

H_0 : pas de différence de délai selon le traitement
H_1 : il existe une différence de délai

Ci : efficacité théorique si délai :
$C_{11} = \frac{112 \times 98}{285} = 38,5$
$C_{12} = \frac{112 \times 96}{285} = 37,7$

Ci doit être plus grand que 5 (pas d'importance par contre pour les Oi) donc on regroupe les données pour que tous les Ci ≥ 5.

Calcul du X_C^2 :
$X_C^2 = \frac{\sum_{i=1}(Oi-Ci)^2}{Ci} = 13,17$

$\alpha = 5\%$
$\nu = (l-1)(c-1) = 4 \times 2 = 8$ ddl
$X_t^2 = 15,51$

$X_C^2 < X_t^2$: au risque 5%, on accepte H_0, on ne montre pas de différence significative

3 – X^2 corrigé de Yates (pour tableaux 2l x 2c)

Si Ci < 5 :
- regrouper les effectifs observés, si cela ne change rien au problème posé
- corrigé de Yates : X_C^2 surestimé

$$X^2_{corrigé} = \frac{\sum \left[|Oi-Ci| - \frac{1}{2}\right]^2}{Ci}$$

Exemple :
n = 20 souris traitées, 8 tumeurs (40%)
en théorie, 20% de tumeurs dans la population $\mathcal{P}$

	T	$\bar{T}$	
Oi	8	12	20
Ci	4	16	20

H_0 : le traitement ne modifie pas le taux de tumeurs
H_1 : le traitement modifie le taux de tumeurs

$X_C^2 = \frac{(|18-4|-0{,}5)^2}{4} + \frac{([12-16|-0{,}5)^2}{16} = 3{,}83$

α = 5%
ν = 1 ddl
$X_t^2 = 3{,}84$

$X_C^2 < X_t^2$: au risque 5%, on accepte H_0, on ne montre pas de différence significative

II – Test de l'écart-réduit

1 – Comparer fréquence observée et fréquence théorique

p_0 n
p population
fréquence connue

H_0 : pas de différence significative
H_1 : il existe une différence significative

Sous H_0 :

Si np et nq ≥ 5 (p + q = 1), $U_C = \frac{p_0 - p}{\sqrt{\frac{pq}{n}}}$ suit une loi normale $\mathcal{N}(0 ; 1)$

α, U_t

Comparaison :
- Si $|U_C| > U_t$: au risque α, on rejette H_0, on accepte H_1
- Si $|U_C| < U_t$: au risque α, on accepte H_0 : on ne montre pas de différence significative

2 – Comparer 2 fréquences observées

p_A n_A
p_B n_B

H_0 : pas de différence significative
H_1 : il existe une différence significative

Sous H_0 :

Si $n_A p$, $n_A q$, $n_B p$, $n_B q \geq 5$, alors $p = \frac{n_A p_A + n_B p_B}{n_A + n_B}$ (et $q = 1 - p$)

et $U_C = \frac{p_A - p_B}{\sqrt{\frac{pq}{n_A} + \frac{pq}{n_B}}}$ suit une loi normale $\mathcal{N}(0 ; 1)$

alfa, U_t

Comparer $|U_C|$ et U_t

3 – Exercices

Exercice 1 :

Population 50% ♂, 50% ♀

n = 20 4 ♂, 16 ♀ touchés par la maladie

La maladie est-elle en rapport avec le sexe ?

$p_0 = \frac{16}{20} = 0{,}8$

$p = 0{,}5$

$\left.\begin{array}{l} np = 10 \\ nq = 10 \end{array}\right\}$ test de l'écart-réduit

H_0 : la maladie touche autant de filles que de garçons

H_1 : la maladie touche différemment les filles et les garçons

Sous H_0 :

$$U_C = \frac{p_0 - p}{\sqrt{\frac{pq}{n}}} = \frac{-0{,}8+0{,}5}{\sqrt{\frac{0{,}5 \times 0{,}5}{20}}} = 2{,}68$$

$\alpha = 5\% \rightarrow U_t = 1{,}96$

$|U_C| > U_t$: au risque 5%, on rejette H_0, la maladie touche plus les ♀ que les ♂

Test du X^2 :

$X_C^2 = 2{,}68^2 = U_C^2$

	♀	♂	
Oi	16	4	20
Ci	10	10	20

$$X_C^2 = \frac{(16-10)^2}{10} + \frac{(4-10)^2}{10} = 7{,}2$$

$\alpha = 5\% \rightarrow X_t^2 = 3{,}84$

$X_C^2 > X_t^2$: au risque 5%, on rejette H_0

Exercice 2 :

n = 927 enfants malades
408 $\overline{T}$ → 104 complications → 304 sans complications
519 T → 166 complications → 353 sans complications
Est-ce que le traitement est efficace ?

$n_A = 408$ $\quad$ $p_A = \frac{104}{408} = 0{,}255$

$n_B = 519$ $\quad$ $p_B = \frac{166}{519} = 0{,}320$

H_0 : la fréquence de complication est la même avec ou sans traitement
H_1 : la fréquence de complication est différente avec ou sans traitement

Sous H_0 :

$$U_C = \frac{p_A - p_B}{\sqrt{\frac{pq}{n_A} + \frac{pq}{n_B}}} = -2{,}16$$

α = 5% → $U_t = 1{,}96$
$|U_C| > U_t$: au risque 5%, on rejette H_0, il existe une différence de fréquence de complication avec ou sans traitement

Autre méthode : test X^2 (homogénéité)

	C		$\overline{C}$		
	Oi	Ci	Oi	Ci	
T	166	157,16	353	357,83	519
$\overline{T}$	104	118,83	304	289,16	408
	270		657		927

$X_C^2 = 4{,}67$

α = 5% → $X_t^2 = 3{,}84$
$X_C^2 > X_t^2$: au risque 5%, on rejette H_0, il existe une différence de fréquence de complication avec ou sans traitement

III – Test de Mac-Nemar

Variables qualitatives, résultats sur même sujet avant et après traitement donc pas indépendants.

avant / après	positif	négatif
positif	n_1	n_2
négatif	n_3	n_4

N = nombre total de patients = $n_1 + n_2 + n_3 + n_4$

2 fréquences non comparables directement, puisque même personne avant et après traitement.

H_0 : pas de différence significative entre les fréquences
H_1 : il existe une différence significative entre les fréquences

fréquences (+,-) (-,+) théoriques de 0,5

+avt, -après -avt, +après

$\frac{n_2}{n_2+n_3}$ ou $\frac{n_3}{n_2+n_3}$

$$U_C = \frac{\frac{n_2}{n_2+n_3} - 0,5}{\sqrt{\frac{0,5 \times (1-0,5)}{n_2+n_3}}} = \frac{n_2 - n_3}{\sqrt{n_2+n_3}}$$

$\alpha = 5\% \rightarrow U_t = 1,96$

Comparaison :

- Si $|U_C| > U_t$: au risque α, on rejette H_0 donc le traitement est efficace
- Si $|U_C| < U_t$: au risque α, on accepte H_0

Test avec X² :

Oi	n_2	n_3
Ci	$\frac{n_2+n_3}{2}$	$\frac{n_2+n_3}{2}$

$$X_C^2 = \frac{\left(n_2 - \frac{n_2+n_3}{2}\right)^2}{\frac{n_2+n_3}{2}} + \frac{\left(n_3 - \frac{n_2+n_3}{2}\right)^2}{\frac{n_2+n_3}{2}} = \frac{(n_2 - n_3)^2}{(n_2 + n_3)}$$

$X_t^2 \rightarrow \alpha$, $\nu = 1$ ddl

$X_C^2 = U_C^2$

Exemple :

Patients asthmatiques

Test d'un aérosol

Etude de la gêne respiratoire

Effectifs dépendants : pas X²

Pas de gêne avant, pas de gêne après = 30

Pas de gêne avant, gêne après = 12

Gêne avant, pas de gêne après = 40

Gêne avant, gêne après = 18

=

après \ avant	gêne	pas gêne
gêne	18	12
pas gêne	40	30

$U_C = \frac{12-40}{\sqrt{12+40}} = -3,88$

$U_t = 1,96$

ou

$X_C^2 = \frac{(12-40)^2}{12+40} = 15,08$

$X_t^2 = 3,84$

$|U_c| > U_t$

$X_c^2 > X_t^2$

au risque 5%, on rejette H_0, il existe une différence avec ou sans traitement => efficacité