Chapitre 6

Les tests d'hypothèse

2 - Les tests du χ^2 (chi 2)

(a) Tester la liaison entre deux variables qualitatives

Tester la liaison entre deux variables

Exemple du chapitre précédent:

Fumer augmente-t-il le risque de développer une certaine maladie ?

p_1 = proportion de malades chez les non fumeurs
p_2 = proportion de malades chez les fumeurs

Test d'hypothèse:

H0: $p_1=p_2$
H1: $p_1 \neq p_2$

Echantillonnage
n_1=197 non fumeurs, 12 malades
n_2=178 fumeurs, 23 malades

D'un point de vue probabiliste: $p_1=P(M|\text{ non } F)$, $p_2=P(M|F)$
avec M = « être malade », F = « fumer »

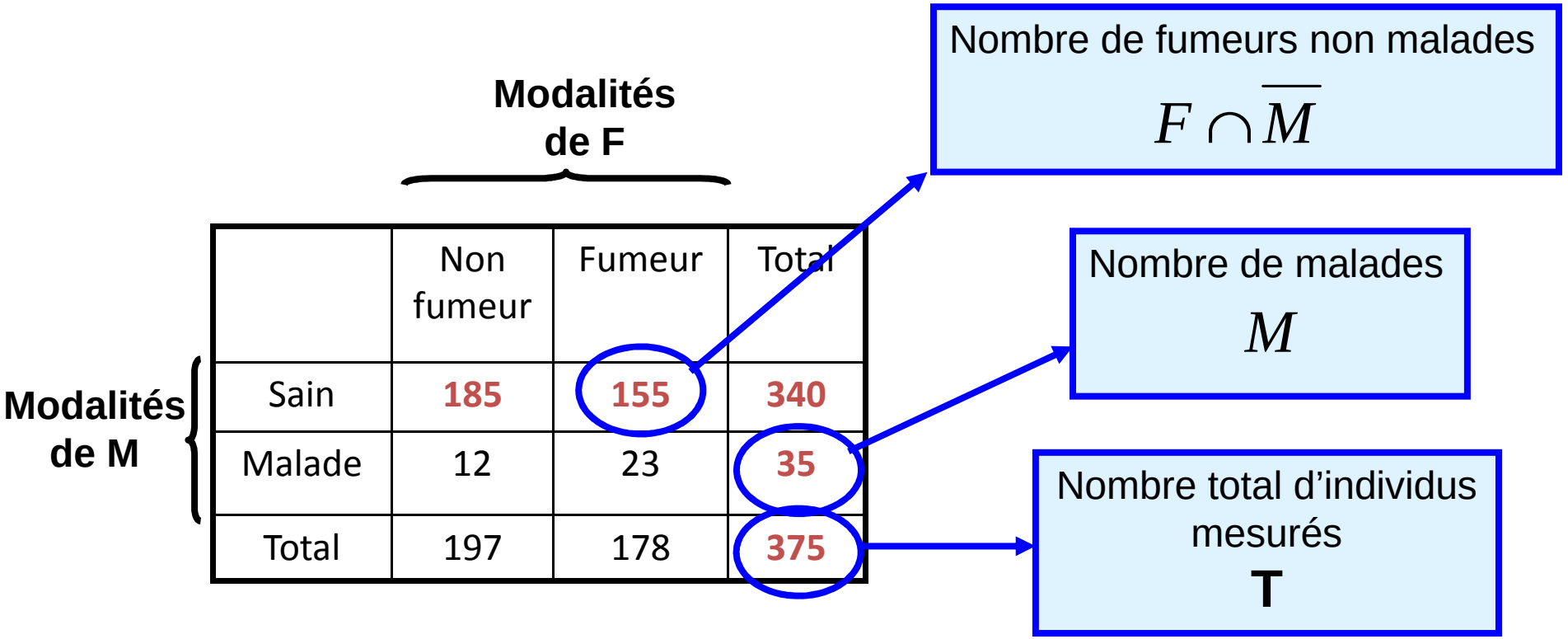
Mathématiquement: $p_1=p_2 \Leftrightarrow$M et F indépendantes

Tester la liaison entre deux variables

Autre écriture du test d'hypothèse (équivalente)

H0: M et F indépendantes
H1: M et F liées

On peut construire une table de contingence:



Tester la liaison entre deux variables

En absence de liaison entre M et F, on aurait $P(M \cap F) = P(M) \times P(F)$

Le nombre de fumeurs malades serait

$$effectif\,(M \cap F) = P(M \cap F) \times T = P(M) \times P(F) \times T = \frac{M}{T} \times \frac{F}{T} \times T = \frac{M \times F}{T}$$

La table de contingence serait (on parle de **table théorique**)

	Non fumeur	Fumeur	Total
Sain	$\frac{\overline{M} \times \overline{F}}{T}$	$\frac{\overline{M} \times F}{T}$	$\overline{M}$
Malade	$\frac{M \times \overline{F}}{T}$	$\frac{M \times F}{T}$	M
Total	$\overline{F}$	F	T

Application

Numérique

	Non fumeur	Fumeur	Total
Sain	178.6	161.4	340
Malade	18.4	16.6	35
Total	197	178	375

Tester la liaison entre deux variables

On peut comparer les tables observées et théoriques

Table observée

	Non fumeur	Fumeur	Total
Sain	185	155	340
Malade	12	23	35
Total	197	178	375

Table théorique

	Non fumeur	Fumeur	Total
Sain	178.6	161.4	340
Malade	18.4	16.6	35
Total	197	178	375

Ces deux tables sont différentes:

$\Rightarrow$ c'est normal (effets d'échantillonnage)

Par contre, si H0 est vraie (M et F indépendantes) les deux tables ne devraient pas être trop « distantes »

Le chi2 (χ^2) mesure la distance entre les deux tables

Tester la liaison entre deux variables

Mathématiquement on a:

$$\chi^2_{obs} = \sum_{cases} \frac{(effectif\ observé - effectif\ théorique)^2}{effectif\ théorique}$$

Réduction de l'écart entre observation et théorie (nous ramène à des lois connues)

Distance entre observations et effectifs théoriques

Si H0 est vraie, alors χ^2_{obs} est la réalisation d'un $\chi^2(1\ ddl)$

Il suffit de comparer la valeur trouvée au seuil d'un chi2 à 1 ddl

Tester la liaison entre deux variables

Application numérique:

Table observée

	Non fumeur	Fumeur	Total
Sain	185	155	340
Malade	12	23	35
Total	197	178	375

Table théorique

	Non fumeur	Fumeur	Total
Sain	178.6	161.4	340
Malade	18.4	16.6	35
Total	197	178	375

$$\chi^2_{obs} = \frac{(185-178.6)^2}{178.6} + \frac{(155-161.4)^2}{161.4} + \frac{(12-18.4)^2}{18.4} + \frac{(23-16.6)^2}{16.6} = 4.71$$

On lit dans la table $\chi^2_{seuil}(\alpha = 0.05,\ 1\ ddl) = 3.84$

$\chi^2_{obs} > 3.84$ il y a donc une liaison significative entre M et F (ici effet de F sur M)

Tester la liaison entre deux variables

Remarque:

Deux méthodes pour traiter une même question (effet de F sur M ?)

1) Comparaison de fréquences

|z|=2.17 comparé à ε_α=1.96

2) Chi 2

$\chi^2_{obs} = 4.71$ comparé à $\chi^2_{seuil}(\alpha = 0.05,\ 1\ ddl) = 3.84$

Les deux tests sont **parfaitement équivalents** car

$$\chi^2_{obs} = z^2$$
$$\chi^2_{seuil}(\alpha,\ 1\ ddl) = \varepsilon^2_\alpha$$

(Bien le cas ici: 2.17^2=4.71 et 1.96^2=3.84)

Et donc $|z| < \varepsilon_\alpha \Leftrightarrow \chi^2_{obs} < \chi^2_{seuil}(\alpha,\ 1\ ddl)$

Tester la liaison entre deux variables

Test de la liaison entre deux variables ayant plus de 2 modalités:

Soient X et Y deux variables ayant n_1 et n_2 modalités (X1,…, Xn_1) pour X et (Y1,…, Yn_2) pour Y

On peut construire la table de contingence observée

	X1	…	Xi	…	Xn_1	Total
Y1	$O_{1,1}$	…	$O_{1,i}$	…	$O_{1,n1}$	N_1
…	…	…	…	…	…	…
Yj	$O_{j,1}$	…	$O_{j,i}$	…	$O_{j,n1}$	N_j
…	…	…	…	…	…	…
Yn_2	$O_{n2,1}$	…	$O_{n2,i}$	…	$O_{n2,n1}$	N_{n2}
Total	M_1	…	M_i	…	M_{n1}	T

Tester la liaison entre deux variables

On calcul les effectifs théoriques: $H_{i,j} = \frac{M_i \times N_j}{T}$

On teste:

H0: X et Y indépendantes
H1: X et Y liées

On calcule le chi2 observé de la même manière $\chi^2_{obs} = \sum_{i,j} \frac{(O_{j,i} - H_{j,i})^2}{H_{j,i}}$

On montre (maths) que si H0 est vraie alors le chi2 observé est la réalisation d'un chi 2 à $(n_1\text{-}1)\times(n_2\text{-}1)$ ddl

$\Rightarrow$ On compare χ^2_{obs} à $\chi^2_{seuil}\left(\alpha,\ (n_1-1)\times(n_2-1)\ ddl\right)$

Condition d'application: les effectifs théoriques (les $H_{j,i}$) doivent être tous ≥ 5
Sinon: on regroupe des modalités (de X ou Y, au choix)

Tester la liaison entre deux variables

Exemple:

Effet du génotype (AA, Aa ou aa) sur la vitesse d'évolution d'un cancer chez la souris

X = génotype
Y = stade du cancer 1 an après exposition au cancérigène

	AA	Aa	aa	Total
Saines	12	5	4	21
Stade 1	17	15	13	45
Stade 2	22	31	43	96
Stade 3	16	45	39	100
Total	67	96	99	262

Table théorique:

	AA	Aa	aa	Total
Saines	$\frac{21\times 67}{262}=5.4$	$\frac{21\times 96}{262}=7.7$	$\frac{21\times 99}{262}=7.9$	21
Stade 1	$\frac{45\times 67}{262}=11.5$	$\frac{45\times 96}{262}=16.5$	$\frac{45\times 99}{262}=17.0$	45
Stade 2	$\frac{96\times 67}{262}=24.5$	$\frac{96\times 96}{262}=35.2$	$\frac{96\times 99}{262}=36.3$	96
Stade 3	$\frac{100\times 67}{262}=25.6$	$\frac{100\times 96}{262}=36.6$	$\frac{100\times 99}{262}=37.8$	100
Total	67	96	99	262

Condition d'application: les effectifs théoriques sont bien tous ≥ 5

Tester la liaison entre deux variables

Chi2 observé:

$$\chi^2_{obs} = \frac{(12-5.4)^2}{5.4} + \frac{(5-7.7)^2}{7.7} + \ldots + \frac{(45-36.6)^2}{36.6} + \frac{(39-37.8)^2}{37.8} = 22.3$$

Valeur seuil:

Nombre de ddl = (3-1) × (4-1) = 6
On prend α=0.05

$$\Rightarrow \quad \chi^2_{seuil}(\alpha = 0.05,\ 6\ \ ddl) = 12.59$$

$\chi^2_{obs} > \chi^2_{seuil}(\alpha = 0.05,\ 6\ \ ddl) \Rightarrow$ H0 rejetée: effet du génotype (AA, Aa, aa) sur la vitesse d'évolution du cancer chez la souris (AA évoluent moins vite)

(b) Tester l’ajustement de données à une loi de probabilité

Le chi2 d'ajustement

Exemple:

X = Nombre d'étudiants gauchers dans des groupes de TT de 4 personnes

X	0	1	2	3	4	Total
effectif	17	24	15	8	2	66

H0: X~B(n,p) (répartition aléatoire des gauchers dans les groupes de TT)
H1: X ne suit pas une B(n,p) (honteuse mise à l'écart des gauchers)

On connaît n (=4) mais pas le paramètre p de la binomiale

$\Rightarrow$ Estimation

$$\hat{p} = proportion\ de\ gauchers = \frac{0\times17+1\times24+2\times15+3\times8+4\times2}{(17+24+15+8+2)\times4} = 0.33$$

Le chi2 d'ajustement

Exemple:

X = Nombre d'étudiants gauchers dans des groupes de TT de 4 personnes

X	0	1	2	3	4	Total
Effectifs observés	17	24	15	8	2	66
Effectifs théoriques	13.6	26.4	19.1	6.2	0.7	66

$$T \times C_n^k \hat{p}^k (1-\hat{p})^{n-k}$$

(T=66, n=4, k= 0, 1, 2, 3 ou 4)

Regroupement (l'effectif théorique de X = 4 est inférieur à 5)

X	0	1	2	3-4	Total
Effectifs observés	17	24	15	10	66
Effectifs théoriques	13.6	26.4	19.1	6.9	66

Le chi2 d'ajustement

Calcul du chi2 observé:

$$\chi^2_{obs} = \frac{(17-13.6)^2}{13.6} + \frac{(24-26.4)^2}{26.4} + \frac{(15-19.1)^2}{19.1} + \frac{(10-6.9)^2}{6.9} = 3.32$$

Nombre de ddl: formule générale:

nb de ddl = nb de modalités (après regroupement) – 1 – nb de paramètres estimés
= 4 – 1 – 1
= 2

$$\chi^2_{seuil}(\alpha = 0.05,\ 2\ \ ddl) = 5.99 \Rightarrow \chi^2_{obs} < \chi^2_{seuil}(\alpha = 0.05,\ 2\ \ ddl)$$

Conclusion: on ne rejette pas H0: distribution binomiale des gauchers dans les groupes de TT **possible**

Ce que l'on a vu ici avec une loi binomiale marche avec n'importe quelle loi de probabilité (normale, poisson,…)

Bilan regroupements et ddl

CHI2 d'ajustement

- Loi binomiale $\beta(n, p)$
  - Regroupement si effectifs théoriques < 5
  - Paramètres estimés: parfois p (1 ou 2)
- Loi de poisson $P(\lambda)$
  - Regroupement en fin de distribution et en début si effectifs théoriques < 5
  - Paramètres estimés: parfois λ (1 ou 2)
- Loi normale $N(\mu,\delta)$
  - Regroupement en début et fin de distribution
  - Paramètres estimés: parfois μ et δ (1 ou 3)

CHI2 d'homogénéité et d'indépendance

Degrés de liberté : (nbre de modalités de X – 1) * (nbre de modalités de Y – 1)
Regroupements de **classes** si effectifs théoriques <5

Conditions d'application des tests et intervalles de confiance

Dans tous les cas (sans exception!!):
- L'échantillon doit être représentatif de la population
- Les mesures doivent être indépendantes

en plus dans le cas où n<30 pour l'**étude d'une moyenne** (IC ou test):
- La ou les variable(s) mesurée(s) doi(ven)t être distribuée(s) suivant une loi normale
- si on compare deux moyennes avec n_1 et n_2<30, on suppose l'égalité des variances

En plus pour l'**étude d'une fréquence** (IC ou test): en plus
- Il faut que le ou les échantillons soi(en)t de taille n≥30
- Les np et nq doivent être ≥5, plus exactement

Intervalle de confiance	Test d'égalité à une fréquence théorique	Test d'égalité de deux fréquences observées
nf, n(1-f) ≥5	np, nq ≥5	n_1f, n_2f, $n_1(1-f)$, $n_2(1-f)$≥5 (f= fréquence commune observée)

En plus pour un **test du chi2**
- n ≥ 50
- tous les effectifs théoriques doivent être ≥ 5