Lecture nº 3.

Test χ^2.

"How close to the truth to you want to come, sir?"

Tests non-parametriques.

Nous disposons d'un échantillon issu d'une population inconnue .
Les données sont donc des réalisations de variables aléatoires, de même loi .
Cette loi n'est pas supposée appartenir à une famille paramétrique particulière.

Tests non paramétriques sont applicables à tout type de populations (ne nécessitent pas l'estimation des paramètres μ, σ).

En pratique, les tests non-paramétriques sont utilisés quand:

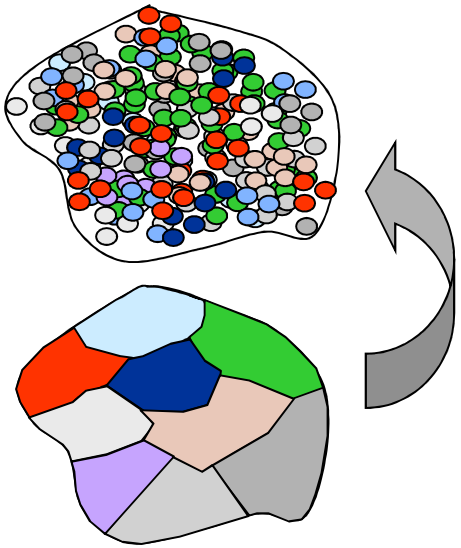
- les données sont non paramétriques (rangs, appréciations, conventions etc.);
- les populations ne sont pas normalement distribuées;
- les variances dans les populations comparées ne sont pas égales;
- les échantillons sont de petite taille.

Test χ^2 de conformité.

(variables discrètes ou continues regroupés en classes)

Dans une population d'effectif N,
on définie k événements E_1, E_2, ..., E_k
formant un système complet
d'événements.

Selon le modèle théorique
de la population, $P(E_i) = p_i$.



PROBLEME :

le modèle théorique
est-il conforme
à la réalité
observée ?

Dans un échantillon théorique de taille **n**, les effectifs (calculés) de chaque événement sont $C_i = np_i$

Dans un échantillon réel de taille **n**,
les effectifs observés des événements E_i
sont O_i.

Hypothèse H_0 :

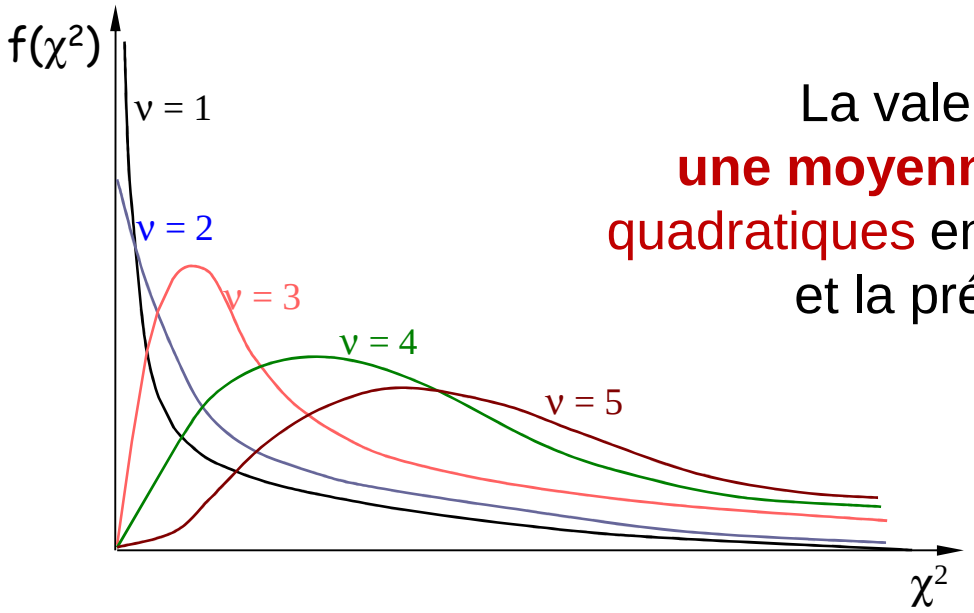
La distribution observée
est conforme
à la distribution théorique choisie

HYPOTHESE H_0 :
La distribution observée est conforme à la distribution théorique choisie.

THEOREME: Sous l'hypothèse H_0, la variable χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

suit la loi de χ^2 à $\nu = k - 1$ degrés de liberté



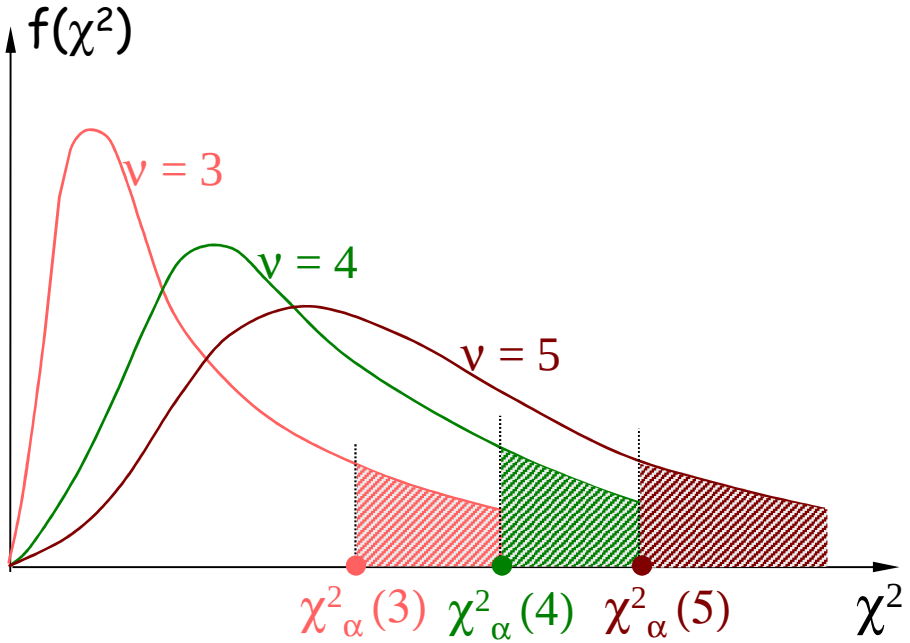
La valeur du χ^2 est donc
une moyenne pondérée d'écarts
quadratiques entre les effectifs observés
et la prévision théorique.

si $\nu \to \infty$, la loi χ^2 tend vers la loi normale (de Gauss).

DECISION:

On fixe α (risque de première espèce).

Les tables de χ^2 permettent de lire $\chi^2_\alpha(\nu)$ telle, que $P(\chi^2 \geq \chi^2_\alpha) = \alpha$

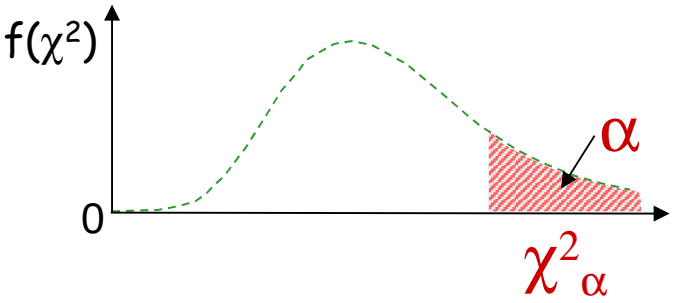


$\Rightarrow$ si $\chi^2 < \chi^2_\alpha \rightarrow H_0$ ne peut pas être rejetée ;

$\Rightarrow$ si $\chi^2 > \chi^2_\alpha \rightarrow H_0$ est écartée au risque α.

Tables de la loi χ^2.

ν/α	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.001
1	0.0002	0.001	0.004	0.016	2.71	3.84	5.02	6.63	10.83
2	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	13.83
3	0.12	0.22	0.35	0.58	6.25	7.81	9.35	11.34	16.27
4	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	18.47
5	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	20.52
6	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	22.46
7	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.47	24.32
8	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	26.13
9	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	27.88
10	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	29.59
11	3.05	3.82	4.57	5.58	17.27	19.67	21.92	24.72	31.26
12	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	32.91
13	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	34.53
14	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	36.12
15	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	37.70
16	5.81	6.91	7.96	9.31	23.54	26.30	28.84	32.00	39.25
17	6.41	7.56	8.67	10.08	24.77	27.59	30.19	33.41	40.79
18	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.80	42.31
19	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	43.82
20	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	45.32
21	8.90	10.28	11.59	13.24	29.61	32.67	35.48	38.93	46.80
22	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	48.27
23	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	49.73
24	10.86	12.40	13.85	15.66	33.20	36.41	39.37	42.98	51.18
25	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	52.62
26	12.20	13.84	15.38	17.29	35.56	38.88	41.92	45.64	54.05
27	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	55.48
28	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	56.89
29	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	58.30
30	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	59.70



Lorsque ν >30, la variable Y

$$Y = \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

suit '*à peu près*' la loi N(0,1)

EXEMPLE 1 : On veut savoir si l'apparition d'une maladie est liée au groupe sanguin. Sur **200** malades observés, on a dénombré: 104 du groupe O, 76 du A, 18 du B, 2 du AB. Dans la population entière, la répartition entre les groupes est : O - 47%, A - 43%, B - 7%, AB - **3%.** Conclusion ?

Nous devons comparer la loi théorique (répartition entre groupes dans la population saine) et la loi observée (cette répartition dans la population malade)

HYPOTHESE H_0: La répartition des groupes sanguins est la même dans les deux populations.

Calcul de χ^2 :

groupe	O	A	B	AB	Total
pi	0.47	0.43	0.07	0.03	1
$C_i = np_i$	94	86	14	6	200
O_i	104	76	18	2	200

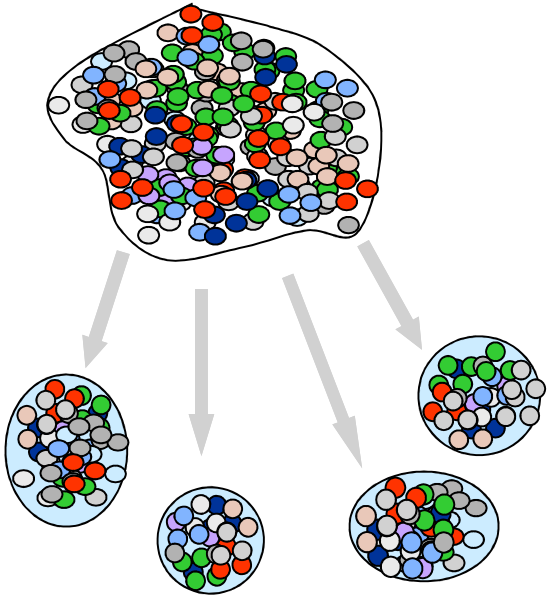
$$\chi^2 = \frac{(104-94)^2}{94} + \frac{(76-86)^2}{86} + \frac{(18-14)^2}{14} + \frac{(2-6)^2}{6} \approx 6.04$$

$\nu = 4 - 1 = 3$
$\alpha = 0.05$
$\chi^2_\alpha(3,0.05) = 7.81$

$\chi^2 < \chi^2_\alpha$

Conclusion : au risque α, l'hypothèse H_0 ne peut pas être rejetée ; la maladie n'est pas liée au groupe sanguin.

Test χ^2 d'homogeneité.



Dans une population P, un caractère A peut prendre k valeurs: A_1, A_2, ..., A_k.

On dispose de m échantillons E_1, E_2, ..., E_m provenant de la population P.

L'effectif de la valeur A_i observé dans l'échantillon E_j est O_{ij}.

PROBLEME : la différence entre les échantillons est-elle significative ?

HYPOTHESE H_0 :
Les différences entre les échantillons sont dues aux fluctuations d'échantillonnage.
(autrement dit, les échantillons sont issus de la même population P).

Calcul des effectifs théoriques :

Sous l'hypothèse H_0
les échantillons peuvent être réunis en un seul, de taille N

$$N = \sum_{i=1}^{k} \sum_{j=1}^{m} O_{ij}$$

La probabilité de l'évènement A_i est donc (théoriquement) :

$$p_i = \frac{\sum_{j=1}^{m} O_{ij}}{N} = \frac{S_i}{N}$$

effectif théorique

$$C_{ij} = n_j p_i = \frac{n_j S_i}{N}$$

THEOREME : Sous l'hypothèse H_0, la variable χ^2

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

suit la loi de χ^2 à $\nu = (k - 1)(m - 1)$ degrés de liberté.

EXEMPLE 2 : Le tableau ci-dessous regroupe les résultats de l'évolution d'une maladie M à l'issue de l'emploi des traitements A, B et C.
Peut-on dire que les traitements A, B et C sont différents ?

	guérison	amélioration	état stationnaire	Totaux
A	280 / 300	210 / 200	110 / 100	600
B	220 / 200	90 / 133	90 / 67	400
C	250 / 250	200 / 167	50 / 83	500
Totaux	750	500	250	1500

$\nu = (3-1)(3-1) = 4$

$\alpha = 0.05$

$\chi^2_\alpha(4, 0.05) = 9.49$

$\chi^2 > \chi^2_\alpha$

HYPOTHESE H_0 : les traitements A, B et C sont identiques.

Sous l'hypothèse H_0,

P(guérison) = 0.5

P(amélioration) = 1/3

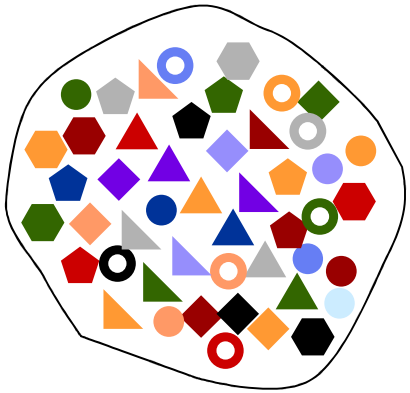
P(état stationnaire) = 1/6

$$\chi^2 = \frac{(280-300)^2}{300} + \frac{(210-200)^2}{200} + \frac{(110-100)^2}{100} + \frac{(220-200)^2}{200} + \frac{(90-133)^2}{133} + \frac{(90-67)^2}{67} + \frac{(250-250)^2}{250} + \frac{(200-167)^2}{167} + \frac{(50-83)^2}{83} \approx 46.27$$

Conclusion: au risque α, l'hypothèse H_0 est rejetée ;les traitements sont différents.

Test χ^2 d'indépendance (corrélation) des caractères.

(variante du test d'homogeneité)



Dans une population, chaque individu possède deux caractères:

X (X_1, X_2, ..., X_k) (ex: forme 

et **Y (Y_1, Y_2, ..., Y_m).** (ex: couleur 

On connaît les nombres O_{ij} d'individus présentant en même temps les caractères X_i et Y_j.

PROBLEME : les caractères **X** et **Y** sont-ils liés entre eux ?

HYPOTHESE H_0 :

Les caractères **X** et **Y** sont indépendants.
(autrement dit, il n'existe aucune corrélation entre eux).

Calcul des effectifs théoriques :

L'effectif total de toutes les valeurs est :

$$N = \sum_{i=1}^{k} \sum_{j=1}^{m} O_{ij}$$

L 'effectif de X_i :

$$S_i = \sum_{j=1}^{m} O_{ij}$$

L 'effectif de Y_j :

$$T_j = \sum_{i=1}^{k} O_{ij}$$

Sous l'hypothèse initiale, si X et Y sont indépendants

$$\frac{C_{ij}}{N} = \frac{S_i}{N} \cdot \frac{T_j}{N}$$

$$C_{ij} = \frac{S_i T_j}{N}$$

THEOREME : Sous l'hypothèse H_0, la variable χ^2

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - C_{ij})^2}{C_{ij}}$$

suit la loi de χ^2 à $\nu = (k - 1)(m - 1)$ degrés de liberté.

EXEMPLE 3: L'étude porte sur des malades dans un hôpital psychiatrique:
on observe s'ils présentent (ou non) une tendance suicidaire (caractère **X**).
Leurs maladies ont été classées en *psychoses* et *névroses* (caractère **Y**).
Y a-t-il un lien entre les tendances suicidaires et le type de maladie ?

	tendance	sans tendance	totaux
psychoses	20 / 40	180 / 160	200
nevroses	60 / 40	140 / 160	200
totaux	80	320	400

$\nu = (2-1)(2-1) = 1$

$\alpha = 0.05$

$\chi^2_\alpha(1, 0.05) = 3.84$

HYPOTHESE H_0 : Il n'y a pas de lien entre la tendance suicidaire et le type de la maladie.

Sous l'hypothèse H_0,

P(tendance suicidaire) = 0.2

P(sans tendance) = 0.8

$$\chi^2 = \frac{(20-40)^2}{40} + \frac{(180-160)^2}{160} + \frac{(60-40)^2}{40} + \frac{(140-160)^2}{160} \approx 25$$

$\chi^2 > \chi^2_\alpha$

Conclusion : au risque α, l'hypothèse H_0 est rejetée ;
il y a un lien entre la tendance suicidaire et le type de maladie.