Test du Chi2 (χ^2)

Dr Oumar Bassoum
FMPO-ISED/UCAD

Objectifs éducationnels

- Différencier un test de conformité d'un test d'indépendance
- Dessiner un tableau de contingence
- **Lire la table du X^2 et calculer le nombre de degré de liberté**
- Réaliser un test de conformité
- Réaliser un test d'indépendance

Plan du cours

I. Généralités sur le Chi2

II. Test de conformité entre une distribution théorique et une distribution expérimentale

III. Indépendance de variables aléatoires. Tableau de contingence

IV. Test de Fischer

V. Les limites du Chi2

Généralités sur le test du Chi2 (χ^2)

- C'est un test utilisé pour:
  - comparer une distribution théorique avec un distribution expérimentale ;
  - tester l'indépendance de deux variables aléatoires.

Test de conformité entre une distribution théorique et une distribution expérimentale

2. **Mesure de l'ampleur de l'écart entre les 2 distributions :**

- Soit une distribution constituée de n observations classées en k classes : dans une classe i nous aurons :
  - O_i : l'effectif observé dans la classe ki
  - E_i : l'effectif espéré dans la classe ki ;
- Pour mesurer l'écart entre la distribution théorique et la distribution observée, on calcule pour chaque classe $\frac{(Oi-Ei)^2}{Ei}$ c'est-à-dire l'Écart Quadratique relatif.

Test de conformité entre une distribution théorique et une distribution expérimentale

2. **Mesure de l'ampleur de l'écart entre les 2 distributions :**

- Pour l'ensemble des cases, on aura :
- $\chi^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$

  Le χ^2 sera faible si les écarts entre les valeurs observées et les valeurs théoriques sont petits.

  Le χ^2 sera grand dans le cas contraire.

- PEARSON a montré que $\chi^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei}$ suit une loi du χ^2 sous l'hypothèse H0.

Test de conformité entre une distribution théorique et une distribution expérimentale

3. Nombre de degrés de liberté (ddl):

- Le nombre de degrés de liberté associé au calcul du χ^2 prend la valeur $\gamma = k - 1$, si la distribution théorique est entièrement spécifiée.

Test de conformité entre une distribution théorique et une distribution expérimentale

4. <u>Hypothèses statistiques</u> :

Soit :

H_0 : les observations suivent une distribution théorique spécifiée ;

H_1 : les observations ne suivent pas la distribution théorique spécifiée.

Test de conformité entre une distribution théorique et une distribution expérimentale

6. **Exemple 1** :

On jette 120 fois un dé. On a observé les fréquences suivantes :

1	2	3	4	5	6
14	16	28	30	18	14

Peut-on en conclure au seuil de $\alpha = 0.05$ que le dé est bien équilibré ?

Test de conformité entre une distribution théorique et une distribution expérimentale

6. Exemple 1 : Solution

- Si le dé est bien équilibré, on doit théoriquement obtenir :
  - 120 x 1/6 = 20 fois 1 ; 20 fois le 2 ; …. ; d'où le tableau de contingence.

Face	1	2	3	4	5	6
Répartition théorique	20	20	20	20	20	20
Répartition observée	14	16	28	30	18	14

Test de conformité entre une distribution théorique et une distribution expérimentale

6. Exemple 1 : Solution

$$\chi 2 = \frac{(14-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(14-20)^2}{20}$$

χ2 = 12.8

Test de conformité entre une distribution théorique et une distribution expérimentale

6. **Exemple 1 : Solution**

- **Hypothèses**:
  - H0 : le dé est équilibré
  - H1 : le dé n'est pas équilibré.

- **Le nombre de degrés de liberté** :
  - On n'utilise pas la distribution expérimentale pour définir la distribution théorique. D'où : v = 6-1 = 5
    nddl = 5 = v

Test de conformité entre une distribution théorique et une distribution expérimentale

6. **Exemple 1 : Solution**

- **Règle de décision** :
  - Lecture de $\chi^2_{0.05;\,5}$ à α = 0.05 et sur la ligne ddl = 5, on lit $\chi^2_{seuil} = 11.070$

- **Conclusion** :
  - Puisque $\chi^2 = 12.8$ et $\chi^2_{seuil} = 11.070$, on rejette l'hypothèse nulle selon laquelle, le dé est équilibré.

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:**

- **Soient 2 variables aléatoires A et B, chacune ayant 2 modalités :**

- **Exemple :**
  - **A --------→ sexe (masculin ou féminin) et**
  - **B --------→ réponse à un médicament (bonne ou mauvaise).**

- **On peut construire le tableau suivant :**

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (suite)

		REPONSES		TOTAL
		Bonne	Mauvaise	
SEXE	Masculin	17	13	30
	Féminin	20	10	30
	TOTAL	37	23	60

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (suite)
   - On peut ainsi calculer les effectifs théoriques de chaque case .

- ***Règle*** : *on multiplie le total de la ligne par le total de la colonne et on divise par le nombre total d'individus*.
- Ex : cas (1,2) = 30 x 23/60 = 11.5

	REPONSES		**TOTAL**
	Bonne	Mauvaise	
Masculin	17 (18.5)	13 (11.5)	30
Féminin	20 (18.5)	10 (11.5)	30
TOTAL	37	23	60

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (suite)

- ***Nombre de degré de liberté*** :
  - Si on se donne les effectifs totaux ; si on place 1 valeur dans le tableau à 4 cases, toutes les autres valeurs sont fixées, d'où :
    - υ = 1 ddl
    - υ = (nombre de lignes – 1) x (nombre de colonnes – 1) = (2-1) x (2-1) = 1

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (suite)

- ***Calcul du χ2:***
  - $\chi 2 =$ $(17-18.5)^2/18.5 + (13 - 11.5)^2/11.5 + (20 - 18.5)^2/18.5 + (10 - 11.5)^2/11.5$
  - ***$\chi 2 = 0.63$***

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (suite)

- ***Hypothèses*** *:*
  - *H_0: les 2 caractères (Sexe, Réponse) sont indépendants ;*
  - *H_1 : les 2 caractères ne sont pas indépendants.*

- ***Seuil de signification*** *: $\alpha = 0.05$*
  - *Lecture du $\chi^2_{0.05\,;\,1}$ dans la table :*
  - *$\chi^2_{seuil} = 3.84$*

Indépendance de variables aléatoires ; Tableau de contingence

1. **Cas des tableaux 2 x2:** (fin)

- ***Décision*** *:*
  - *$\chi 2 = 0.63$ $\chi 2 < \chi 2$ seuil*
  - *On retient donc l'hypothèse H_0 (les 2 caractères sont indépendants). Il n'existe pas d'association entre la réponse au médicament et le sexe du malade.*

Indépendance de variables aléatoires ; Tableau de contingence

2. **Cas des petits échantillons:**

Les effectifs théoriques calculés doivent être supérieurs à 5 pour que le χ2 soit interprétable.

Le nombre total de sujets doivent être supérieur à 30.

Si ces conditions ne sont pas réalisées, on peut utiliser la correction de Yates à condition que les effectifs ne soient pas trop éloignés de 5.

On calcule le χ2 en retranchant ½.

χ2 Yates = $$\sum_{i=1}^{k} \frac{\left(Oi - Ei - 0.5\right)^2}{Ei}$$

Indépendance de variables aléatoires ; Tableau de contingence

2. Cas des petits échantillons: (suite)

- **Exemple :**

Sur un échantillon de 40 souris soumises à un carcinogène on a observé en fonction du sexe, le nombre de souris avec ou sans cancer.

	Cancer	Pas cancer	
Mâle	3	17	20
Femelle	6	14	20
	9	31	40

Question: Y-a-t-il une répartition différente des différents cancers?

Indépendance de variables aléatoires ; Tableau de contingence

2. **Cas des petits échantillons:** (suite)

- **Exemple :**

Sur un échantillon de 40 souris soumises à un carcinogène on a observé en fonction du sexe, le nombre de souris avec ou sans cancer.

	Cancer	Pas cancer	
Mâle	3 (4,5)	17 (15,5)	20
Femelle	6 (4,5)	14 (15,5)	20
	9	31	40

Question: Y-a-t-il une répartition différente des différents cancers?

Indépendance de variables aléatoires ; Tableau de contingence

2. **Cas des petits échantillons:** (suite)

- **Exemple :** Solution
  - Calcul des effectifs théoriques:
    - 9 x 20/ 40 = **4.5**
    - 31 x 20 / 40 = 15.5
  - Calcul du χ^2:
    - La cellule (1,1) a un effectif de **4.5 < 5**. Il faut utiliser la correction de Yates pour corriger le $\chi 2$
    - $$\chi^2_{Yates} = \frac{(|3-4.5|-0.5)^2}{4.5} + \frac{(|17-15.5|-0.5)^2}{15.5} + \frac{(6-4.5-0.5)^2}{4.5} + \frac{(14-15.5-0.5)^2}{15.5}$$
    - χ^2_{Yates} = 1/4.5 + 1/15.5 + 1/4.5 + 1/15.5 = 0.573
    - $X^2_{0.05;1}$ = 3.84
  - Il n'y a pas de différence statistiquement significative entre la répartition des cancers chez les mâles et chez les femelles.

2. Cas des petits échantillons: (fin)

Remarques :

Cette correction encore appelée correction de continuité est approximative.

On peut faire un calcul exact : c'est la méthode de FISHER (1930). Le principe est basé sur le calcul de la probabilité d'avoir une répartition particulière.

Ce calcul doit être fait si l'effectif est inférieur à 20 ou si un effectif d'une case est inférieur à 5.

Indépendance de variables aléatoires ; Tableau de contingence

3. Cas des tableaux m lignes x n colonnes:

(suite)

On calcul de la même manière les effectifs théoriques.

- $E(1,1) = S_1 \times T_1 / N$
- $E(2,2) = S_2 \times T_1 / N$

Etc.

Indépendance de variables aléatoires ; Tableau de contingence

3. Cas des tableaux m lignes x n colonnes:

On calcul aussi de la même manière le χ^2

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Le nombre de degrés de liberté est égale à :

$\upsilon = (m-1) \times (n-1)$

La valeur seuil est donnée par la table du χ^2 à α = 5%

$\chi^2_{0.05\,;\upsilon}$

On rejette l'hypothèse nulle H0 si $\chi^2 > \chi^2_{0.05\,;\upsilon}$

Indépendance de variables aléatoires ; Tableau de contingence

3. Cas des tableaux m lignes x n colonnes:

Exemple :

On a sélectionné 3864 mariages au hasard en 1980. On a recherché le devenir de ces mariages 5 ans après en fonction de la résidence des jeunes ménages. On obtient la table de contingence suivante :

	Rural	Petite ville	Grande ville	
Encore mariés	287 (275.6)	1124 (1096.2)	2081 (2120.2)	3492
séparés	18 (29.4)	89 (116.8)	265 (255.8)	372
	305	1213	2346	3864

Y-a-t-il une indépendance entre le lieu de résidence et l'état marital ?

Indépendance de variables aléatoires ; Tableau de contingence

3. **Cas des tableaux m lignes x n colonnes:**

Exemple : Solution

Hypothèses :

H0 : indépendance entre les états maritaux et la résidence.

H1 : dépendance entre les deux facteurs

Seuil α = 0.05

Calcul des effectifs théoriques:

- (1,1) = (305 x 3492) / 3864 = 275.6
- (2,2) = (1213 x 372) / 3864 = 116.8

Etc.

Indépendance de variables aléatoires ; Tableau de contingence

3. Cas des tableaux m lignes x n colonnes:

Exemple : Solution

- **Calcul du $\chi 2$**
  - $\chi 2 = (287-275.6)^2/ 275.6 + (1124 - 1096.2)^2/1096.2 + (2081-2120.2)^2/2120.2 + (18 - 29.4)^2/29.4 + (89 -116.8)^2/116.8 + (265 - 255.8)^2/255.8 = 19.68$
  - $\chi 2 = 19.68$
- **Nombre de degrés de liberté**
  - $\upsilon = (3-1) \times (2-1) = 2$
- **Recherche du $\chi^2_{0.05\,;2}$ sur la table**
  - $\chi^2_{0.05\,;2} = 5.991$

Indépendance de variables aléatoires ; Tableau de contingence

3. Cas des tableaux m lignes x n colonnes:

Exemple : Solution

- **Conclusion**:
  - Il existe une dépendance entre la résidence et le statut marital, cinq ans après le mariage.
  - Niveau de significativité sur la table $\chi^2_{0.01\,;2} = 13.815$
  - $\chi^2_{calculé} > \chi^2_{0.01\,;2}$
  - L'hypothèse H0 est rejetée avec un niveau de significativité p > 0.001.

Test de Fischer

- Ce test donne des résultats exacts pour des tables 2 x 2 ;

- Mais il n’est utile que lorsque l’on s’attend à obtenir des valeurs espérées petites, c’est-à-dire lorsque le test du $\chi 2$ n’est pas applicable.

Test de Fischer

- **Exemple**:

  Un échantillon de 40 souris est soumis à un agent carcinogène. On dénombre en fonction du sexe, le nombre de souris avec ou sans cancer.

	Cancer	**Pas cancer**	
Mâle	2 (3)	18 (17)	20
Femelle	4 (3)	16 (17)	20
	6	34	40

<u>Question</u>: Peut-on dire qu’il existe une répartition différente du nombre de cancers dans les 2 groupes ?

Test de Fischer

- **Exemple**: Solution
  - **Calcul des effectifs théoriques** :
    - Case (1,1) = (6 x 20)/ 40 = 3
    - Case (1,2) = (34 x 20)/ 40 = 17
  - On ne doit pas utiliser le test du **$\chi 2$** même corrigé par la correction de Yates car les effectifs théoriques sont trop faibles [3 pour la case (1,1) et 3 pour la case (2,1)].

Test de Fischer

- **Exemple**: Solution

Théorie :

- Supposons que P1 soit la probabilité de développer un cancer chez la souris mâle et P2 la probabilité de développer un cancer chez la souris femelle.

- On souhaite tester l'hypothèse nulle.
  - H0 : $P1 = P2 = P$ vis-à-vis de l'hypothèse alternative
  - H1 : $P1 \neq P2$

Test de Fischer

- **Exemple**: Solution
- Théorie (suite)
  - Si les effectifs marginaux sont fixés, on peut calculer la probabilité d'observer une répartition particulière.

	Cancers	Pas cancers	
Mâle	a	b	a +b
Femelle	c	d	c+d
	a+c	b+d	n

Test de Fischer

- **Exemple**: Solution
- Théorie (suite)
  - La probabilité exacte d'observer une répartition particulière sera :
    - Pr (a,b,c,d) = $$\frac{((a+b)!(c+d)!(a+c)!(b+d)!)}{n!a!b!c!d!}$$

Test de Fischer

- **Exemple**:
- Calcul de la probabilité d'observer la répartition (2,18,4,16)
- Pr (2,18,4,16) = $\frac{20!20!6!34!}{40!2!18!4!16!} = \frac{19x20x17x18x19x20x5x6}{2x35x36x37x38x39x40}$

$$= \frac{19x17x15}{7x37x2x39} = 0{,}2398$$

Comparaison de deux pourcentages observés

Cas de séries appariées:

On souhaite comparer 2 traitements A et B sur 100 malades, chacun recevant successivement les 2 traitements dans un ordre aléatoire.

Le résultat est noté (+) ou (-)

On obtient 100 couples de réponses :

Comparaison de deux pourcentages observés: Cas de séries appariées

Résultats avec A	B	Nombre de malades
-	-	35
-	+	5
+	-	15
+	+	45
TOTAL		10

Comparaison de deux pourcentages observés: Cas de séries appariées

	0

Comparaison de deux pourcentages observés: Cas de séries appariées

- **On peut également faire un χ^2:**

	A	B	
+	60 (55)	50 (55)	110
-	40 (45)	50 (45)	90
	100	100	200

Comparaison de deux pourcentages observés: Cas de séries appariées

- Les 2 séries ne sont pas indépendantes, la question posée est :
  - quel est le meilleur médicament ?

- Les réponses concordantes (- - et + +) n'apporte rien. On utilisera les réponses discordantes pour conclure.
  - On a observé 20 discordances soit (- +) 5 fois et (+ -) 15 fois.

- L'hypothèse nulle devrait donner 10 discordances dans un sens et 10 dans l'autre sens

Comparaison de deux pourcentages observés: Cas de séries appariées

- **χ^2 Mc Némar**

B

A

	+	-	
+	**45**	**15**	**60**
-	**5**	**35**	**40**
	50	**50**	**100**

Comparaison de deux pourcentages observés: Cas de séries appariées

- Khi2 de Mc Nemar:

Khi2 = = 5

- on rejette l'hypothèse nulle

Limites du χ^2

- Le **χ2** ne donne aucune information:
  - sur le sens de la liaison,
  - ni sur la force de la liaison.

MERCI DE VOTRE ATTENTION