Tests du χ^2

Université de Bourgogne, Licence 2 Gestion

20 novembre 2024

Test du χ^2 d'indépendance

- Le test du χ^2 d'indépendance permet de déterminer s'il existe une relation entre deux variables X et Y **catégorielles**.
- Les hypothèses sont :
  H_0 : X et Y sont indépendantes.
  H_1 : X et Y sont liées.

Exemple

Dans une entreprise, vous avez collecté des données sur la satisfaction de 200 clients et leur tranche d'âge.

	Très satisfait	Satisfait	Non satisfait
moins de 30 ans	25	40	10
entre 30 et 45 ans	20	30	15
plus de 45 ans	15	25	6

Peut-on dire au niveau 5% que la satisfaction est liée à l'âge ?

Modélisation

- **X=Satisfaction** a comme catégories {Très satisfait, Satisfait, Non satisfait}
  Y=Age a comme catégories { moins de 30, entre 30 et 40, plus de 45 }
- H_0 : X et Y sont indépendantes
  H_1 : X et Y sont liées
- Ajoutons d'abord les effectifs marginaux :

	Très satisfait	Satisfait	Non satisfait	
moins de 30 ans	25	40	10	$n_{1*} = 75$
entre 30 et 45 ans	20	30	15	$n_{2*} = 65$
plus de 45 ans	15	25	5	$n_{2*} = 45$
	$n_{*1} = 60$	$n_{*2} = 95$	$n_{*3} = 30$	$n = 185$

- Sous H_0, les effectifs attendus dans les cellules sont proches de :

$$n_{ij}^{th} = \frac{n_{i*} \times n_{*j}}{n}$$

appelés **effectifs théoriques sous** H_0.

Effectifs théoriques (en rouge)

	Très satisfait	Satisfait	Non satisfait	
moins de 30 ans	$25/\frac{75\times 60}{185} = 24.32$	$40/38.51$	$10/12.16$	$n_{1*} = 75$
entre 30 et 45 ans	$20/21.08$	$30/33.38$	$15/10.54$	$n_{2*} = 65$
plus de 45 ans	$15/14.59$	$25/23.11$	$5/7.29$	$n_{2*} = 45$
	$n_{*1} = 60$	$n_{*2} = 95$	$n_{*3} = 30$	$n = 185$

- ☞ "idée du test" : la statistique à utiliser mesure une distance entre les effectifs observés et les effectifs théoriques des cellules. Si la distance est trop grande, c'est que les variables ne sont pas indépendantes.
- ☞ Notons N_{ij} l'effectif aléatoire de la cellule i,j. Cette distance est mesurée par la statistique :

$$Y = \sum_{i,j} \frac{(N_{i,j} - n_{i,j}^{th})^2}{n_{i,j}^{th}}$$

- ☞ Résultat : *pour un tableau à p lignes et q colonnes la statistique Y suit une loi du* χ^2 *à* $(p-1) \times (q-1)$ *ddl.*

- ☞ ici $(3-1)\times(3-1)=4$ ddl.
- ☞ **Zone de rejet de** H_0 on rejette H_0 si Y est "trop" grand :

$$ZR_{0.05}=\{Y>9.488\}$$

- ☞ $Y^e=\frac{(25-24.32)^2}{24.32}+\frac{(40-38.51)^2}{38.51}+\cdots+\frac{(5-7.29)^2}{7.29}=3.63$
- ☞ Y^e n'est pas dans la zone de rejet de H_0. On ne peut pas dire qu'il y a un lien entre satisfaction et âge.
- ☞ **Attention** pour pouvoir faire un test du χ^2, il faut que tous les effectifs théoriques soit ≥ 5. Si ce n'est pas le cas, il faut regrouper des catégories (cf ex **TD** 17).

Exercice 17

Le responsable ressource humaine s'intéresse à l'indicateur jours de maladies + accidents du travail du bilan social de trois filiales de sa société :

	filiale 1	filiale 2	filiale 3
jours maladies+AT	123	165	149
jours travaillés	2559	2545	2528

Faire un test d'indépendance du χ^2 pour déterminer si les filiales sont toutes équivalentes concernant cet indicateur.

- L'échantillon est constitué de l'ensemble des journées de travail du bilan social de ces filiales.
- X = numéro de la filiale a comme catégories $\{1, 2, 3\}$
- Y = type du jour a comme catégories {maladie, travaillé}
- **Hypothèses**
  H_0 : X et Y sont indépendantes H_1 : elles ne le sont pas
- **Statistique** Calcul des effectifs théoriques $n_{i,j}^{th}$ sous H_0 (en rouge)

	filiale 1	filiale 2	filiale 3	
jours maladies+AT	$123/\frac{437\times 2682}{8069} = 145.2$	$165/146.7$	$149/145.0$	437
jours travaillés	$2559/2536.7$	$2545/2563.2$	$2528/2532.0$	7632
	2682	2710	2677	8069

- Notons $N_{i,j}$ l' effectif aléatoire de la cellule i,j. Sous H_0 $Y = \sum_{i,j} \frac{(N_{i,j} - n_{i,j}^{th})^2}{n_{i,j}^{th}}$ suit une loi du χ^2 à $(2-1)\times(3-1) = 2ddl$
- $ZR_{0.05} = \{Y \geq 5.991\}$
- $Y^e = \frac{(123-145.2)^2}{145.2} + ... + \frac{(2528-2532.0)^2}{2532.0} = 6.11 > 5.991$
- On accepte H_1 : les filiales ne sont pas toutes équivalentes concernant cet indicateur.

Test du χ^2 d'équirépartition

- Le test du χ^2 d'équirépartition permet de tester si une variable catégorielle est **équirépartie** (ou uniforme), c'est-à-dire si chacune de ses catégories a la même probabilité d'être prise.
- Les hypothèses sont :
  H_0 : X est équirépartie
  H_1 : X n'est pas équirépartie

Exemple

☞ On veut tester si un dé n'est pas truqué au niveau 0,05. Pour cela, on lance le dé 60 fois et on obtient les résultats suivants :

face	1	2	3	4	5	6
N_i	15	7	4	11	6	17

Modélisation

- L'échantillon est

$$1, ..., 1 \text{ (15fois)}, 2, ..., 2 \text{ (7fois)}, ..., 6, ..., 6 \text{ (17fois)}$$

- La variable catégorielle est X="face du dé". Elle prend les valeurs (catégories) $i = 1, 2, 3, 4, 5, 6$.
- On note $p_i = p(X = i)$.
- Si le dé n'est pas truqué : $p_1 = \cdots = p_6 = \frac{1}{6}$, c'est-à-dire que X suit une loi équirépartie sur $\{1, \cdots, 6\}$.
- **Hypothèses**
  H_0 : X suit une loi équirépartie
  H_1 : X ne suit pas une loi équirépartie.

- ☞ "idée du test" : on calcule les effectifs théoriques de la loi sous H_0 et la statistique utilisée mesure une distance entre les effectifs observés et les effectifs théorique.
- ☞ On prend ici un échantillon de taille $n = 60$.
- ☞ Les effectifs théoriques sous H_0 sont tous égaux à $n_i^{th} = n \times p_i = \frac{60}{6} = 10$.
- ☞ Notons N_i la statistique désignant l'effectif de la modalité i dans un échantillon de 60 lancers.
- ☞ Sous H_0, les N_i devraient être "assez proches" des $n_i^{th} = 10$.

- ☞ Résultat : *Sous H_0 la statistique*

$$Y = \sum_{i=1}^{r} \frac{(N_i - n_i^{th})^2}{n_i^{th}}$$

suit une loi du χ^2 à $r-1$ ddl

- ☞ (r est le nombre de catégories de X, ici 6.)
- ☞ **Zone de rejet** On accepte H_1 si les N_i sont loin des n_i^{th}, donc pour les grandes valeurs de Y :

$$ZR_\alpha = \{Y \geq x_\alpha\}$$

Dé truqué ?

☞

$face$	1	2	3	4	5	6
N_i	15	7	4	11	6	17
n_i^{th}	10	10	10	10	10	10

☞ $ZR_{0.05} = \{Y \geq 11.07\}$

☞ $Y^e = \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \cdots + \frac{(17-10)^2}{10}$

☞ $Y^e = 13.6$

☞ On peut donc conclure avec un risque d'erreur de 5% que le dé est truqué.

Exercice 19

Le nombre de livres empruntés dans une bibliothèque pendant une semaine est : Lundi : 120 , Mardi : 100 , Mercredi : 115, Jeudi : 120 , Vendredi : 140. Tester si le nombre de livres empruntés est équiréparti sur les jours de la semaine.

Correction

- **Modélisation** on a un échantillon de

$$n = 120 + 100 + 115 + 120 + 140 = 595 \text{ livres}$$

- La variable catégorielle est $X =$ "jour où le livre est emprunté"
- X a $r = 5$ catégories : lundi, mardi, mercredi, jeudi, vendredi
- **Hypothèses** H_0 : X est équirépartie sur les 5 jours
  H_1 : X n'est pas équirépartie

Correction

- **Statistique** Sous H_0, les effectifs théoriques sont de
$$n_i^{th} = \frac{595}{5} = 119$$
- Notons N_i les effectifs aléatoires des différents jours.
- Sous H_0, $Y = \sum_{i=1}^{5} \frac{(N_i - 119)^2}{119}$ suit une loi $\chi^2(4)$ $(r - 1 = 4)$.
- **Zone de rejet** $ZR_{0.05} = \{Y \geq 9.488\}$
- **Valeur expérimentale**
$$Y^e = \frac{(120 - 119)^2}{119} + \frac{(100 - 119)^2}{119} + \cdots + \frac{(140 - 119)^2}{119} = 6.94$$
- Y^e n'est pas dans la zone de rejet de H_0
- donc on peut considérer que les emprunts sont équirépartis sur les jours de la semaine.