

TESTS D'AJUSTEMENT: TEST DE NORMALITÉ DE SHAPIRO-WILK TEST D'INDÉPENDANCE

Benchikh Tawfik

Faculté de Médecine, UDL, SBA
1^{ère} année Médecine

20 Mars 2024

PLAN DE COURS

1 AJUSTEMENT ANALYTIQUE ET PRINCIPAUX TESTS

2 TEST D'INDÉPENDANCE:

- Pour mettre en oeuvre un test d'ajustement, il faut:
 - prélever un échantillon suffisamment important de la population étudiée,
 - classer les observations par ordre croissant dans le cas d'une variable aléatoire discrète, les répartir en classes (par ordre croissant) pour une variable aléatoire continue, d'égale amplitude ou d'égale probabilité,
 - définir une variable de décision D permettant de mesurer les écarts entre la distribution théorique F et la distribution empirique F^* de l'échantillon.

- Pour vérifier la concordance des deux distributions, on doit:
 - définir les hypothèses $H0$ et $H1$,
 - ◇ $H0$: les observations suivent une distribution théorique spécifiée
 $F = F_0$,
 - ◇ $H1$: les observations ne suivent pas la distribution théorique spécifiée
 $F \neq F_0$.
 - accepter un risque de première espèce α de refuser l'hypothèse $H0$ alors qu'elle est vraie,

LES DÉMARCHES DE TEST (2)

- Calculer la valeur d de la variable de décision D (à partir des valeurs données par l'échantillon),
- énoncer une règle de décision:
 - ◇ on rejette l'hypothèse H_0 si la valeur calculée d est supérieure à une valeur d_0 n'ayant qu'une probabilité α d'être dépassée par la variable D ,
 - ◇ sinon, on garde l'hypothèse H_0 et on considère que la distribution théorique spécifiée peut décrire le phénomène étudié, c'est-à-dire $F = F_0$.

- Le test du chi-deux utilise des propriétés de la loi multinomiale. Deux cas sont à distinguer:
 - ❶ la fonction de répartition F est entièrement spécifiée, en particulier, les paramètres sont connus,
 - ❷ on connaît seulement la forme de la loi de distribution, les paramètres de la fonction de répartition F sont estimés à partir d'un échantillon.

- Soit X la variable aléatoire parente, de fonction de répartition F .
- On considère une partition du domaine de définition en r intervalles $I_1, \dots, I_r$ d'égale étendue ou non.
- Pour chaque intervalle I_i , on considère l'ensemble E_i tel que:

$$E_i = \{\omega : X(\omega) \in I_i\} \quad p_i = \Pr(E_i)$$

- $n * p_i$ est égal à la fréquence absolue (effectif) théorique espérée dans la classe I_i que l'on compare à la fréquence observée (effectif) N_i dans cette même classe I_i .

TEST DU CHI-DEUX: VARIABLE DE DÉCISION

- C'est la statistique:

$$D^2 = \sum_{i=1}^r \frac{(N_i - n * p_i)^2}{n * p_i}$$

- Si l'hypothèse H_0 est vraie (concordance acceptable entre la distribution théorique et les valeurs observées), cette quantité ne peut pas être trop grande.
- En fait, Pearson a montré que la statistique D^2 suit une loi du chi-deux, à n degrés de liberté quelle que soit la loi considérée, quand le nombre n d'observations tend vers l'infini.
- Le nombre n de degrés de liberté est égal à:
 - $(r - 1)$ si la distribution théorique est entièrement déterminée, aucun paramètre n'ayant été estimé,
 - $(r - 1 - k)$ si k paramètres ont été estimés à partir des observations, pour définir complètement la distribution.

- On rejette l'hypothèse H_0 si la valeur de la statistique D^2 obtenue à partir de l'échantillon est supérieure à une valeur n'ayant qu'une probabilité α d'être dépassée par la variable χ^2 considérée.
- Sinon, on garde l'hypothèse H_0 et on considère que la distribution théorique spécifiée est acceptable, c'est-à-dire $F = F_0$.

- La distribution limite de la statistique D^2 est indépendante de la loi F , ce test peut donc être utilisé dans de nombreuses situations.
- Les effectifs de chaque classe doivent être supérieurs à cinq. Si cette condition n'est pas vérifiée, on regroupe les classes d'effectifs trop faibles.

TEST DU CHI-DEUX: EXEMPLE

- Sur 300 étudiants en Médecine (Fmed-SBA), la moyenne de la taille est de 1.75m avec un écart type estimé de 0.1m. Ces deux paramètres sont **estimés** à partir des données de cet échantillon.
- Vous avez observé 8 étudiants avec une taille inférieure à 1.55m ; 40 avec une taille entre 1.55 et 1.65; 102 avec une taille entre 1.65 et 1.75 et 150 avec une taille supérieure à 1.75m
- La distribution de la taille des étudiants en Médecine peut elle être considérée comme suivant une loi normale ?

- Moyenne estimée = 1.75 , Écart type estimé = 0.1;

Obs	<1.55 8	[1.55-1.65[40	[1.65-1.75[102	≥ 1.75 150	Total 300
Prob si Loi norm	0.0228	0.1359	0.3413	0.5	
Effec Théo	6.84	40.77	102.39	150	300

- $\chi^2_{cal} = 0.04$,
- DDL = 4 - 1 - 2 = 1; $\chi^2_{th} = 3,84$
- Donc $\chi^2_{cal} = 0.04 < \chi^2_{th} = 3,84$ et $p > 0,05$.
- La distribution observée ne diffère pas de manière statistiquement significative d'une loi normale de paramètre (1.75; 0.1).

PRINCIPE

- On mesure deux variables aléatoires qualitatives dans une population notées X et Y .
 - On veut savoir si ces variables sont indépendantes, c'est-à-dire, si la connaissance d'une des v.a. peut influencer la loi de probabilité de l'autre.
- Le test d'indépendance du Khi2 permet de déterminer si deux variables qualitatives son indépendantes ou non, ou autrement dit, si les réponses de l'une conditionnent les réponses de l'autre.
- Pour cela, nous testons les deux hypothèses suivantes:
 - Hypothèse nulle : H_0 : "les deux caractères sont indépendants" (X et Y sont indépendantes)
 - Hypothèse alternative : H_1 : "les deux caractères ne sont pas indépendants" (X et Y sont liées)
- **Remarque:** Le test ne permet pas de connaître le sens de la dépendance.

DÉMARCHES

Le test se fait selon les étapes suivantes:

- Le test s'applique sur un tableau de contingence, expression qui désigne le tableau de croisement des deux variables qualitatives.
 - Un tableau de contingence est un tableau dans lequel les fréquences correspondent à deux variables: une variable est utilisée en ligne et l'autre en colonne.

$Y \setminus X$	modalité 1	...	modalité j	...	modalité k	Total
échantillon 1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
échantillon i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
échantillon h	n_{h1}	...	n_{hj}	...	n_{hk}	$n_{h.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.k}$	$n_{..}$

DÉMARCHES

- On calcule à partir du tableau de contingence pour chaque case ij , l'effectif calculé ou théorique en utilisant la formule suivante:

$$c_{ij} = \frac{n_{i.} * n_{.j}}{n_{..}}$$

- Le χ^2_{cal} relatif à l'ensemble des données est obtenu par la formule suivantes:

$$\chi^2_{cal} = \sum_{i=1}^h \sum_{j=1}^k \frac{(n_{ij} - c_{ij})^2}{c_{ij}} .$$

DÉMARCHES

Le test se fait selon les étapes suivantes:

- **Conditions d'application:** $c_{ij} \geq 5$, pour tout i, j .
- Le nombre d'échantillons étant h et le nombre de classes étant k , le nombre de degré de liberté est donné par la formule : $ddl = (h - 1)(k - 1)$.
- Etant donné un seuil de confiance $1 - \alpha$, on utilise le tableau de χ^2 pour déterminer la valeur du $\chi_{th}^2 = \chi_{1-\alpha}^2$ avec un $ddl = (h - 1)(k - 1)$.
- Conditions d'application: $c_{ij} \geq 5$. pour tout i, j .
- **Décision:** on applique la règle de décision suivante:
 - Si $\chi_{cal}^2 \geq \chi_{th}^2$, on rejette H_0 et on accepte H_1 avec le risque α .
 - Si $\chi_{cal}^2 < \chi_{th}^2$, l'hypothèse H_0 est retenue et H_1 rejetée avec taux de sécurité $1 - \alpha$.

- Un hôpital enregistre le nombre de malades reçus dans un pavillon durant deux (2) périodes: Période Normale et Période avec Grippe. On note le nombre de morts et de survivants parmi ces mêmes malades, les résultats sont les suivants.
 - Période normale: nombre de morts 49, nombre de survivant 91.
 - Période avec Grippe: nombre de morts 49, nombre de survivant 91.
- Y-t-il une relation entre le nombre de malades (morts et vivants) et le type de période (normale, avec grippe) ?

TEST DU CHI-DEUX: SOLUTION

- On pose d'abord les hypothèses :
 - H0 : "le nombre de malades (morts et vivants) est indépendant du type de période (normale, avec grippe)"
 - H1 : "le nombre de malades n'est pas indépendant du type de période"
- Construisons le tableau de contingence avec le calcul des effectifs théoriques par la formule :

$$c_{ij} = \frac{n_{i.} * n_{.j}}{n_{..}}$$

Période	Nombre	Morts	survivant	Total
Période normale		49	91	$n_{1.} = 140$
		50.7	89.297	
Période avec grippe		18	27	$n_{2.} = 45$
		16.297	28.7	
Total		$n_{.1} = 67$	$n_{.2} = 118$	$n_{..} = 185$

TEST DU CHI-DEUX: SOLUTION

- Sachant que tous les $c_{ij} \geq 5$, pour tout i, j : alors on procède au calcul de avec la formule :

$$\begin{aligned}\chi_{cal}^2 &= \sum_{i,j} \frac{(n_{ij} - c_{ij})^2}{c_{ij}} \\ &= \frac{(49 - 50.7)^2}{50.7} + \frac{(91 - 89.297)^2}{89.297} + \frac{(18 - 16.297)^2}{16.297} + \frac{(27 - 28.7)^2}{28.7} \\ &= 0.3681.\end{aligned}$$

- On déduit ensuite le χ_{th}^2 :

- $h = 2$ et $k = 2 \Rightarrow ddl = (2 - 1) * (2 - 1) = 1$.
- $\alpha = 5\%$
- D'où $\chi_{th}^2 = 3.84$.

- Conclusion : on a donc comme résultat :

$\chi_{cal}^2 = 0,3681 < \chi_{th}^2 = 3,84$. L'hypothèse nulle H_0 est retenue au seuil de confiance de 95%. Donc le nombre de malades (morts et vivants) est indépendant du type de période (normale, avec grippe) c'est-à-dire qu'il n'existe aucun lien entre eux, alors la période grippale est identique à la période normale en terme de taux de mortalités ou de survivants.