Comparaisons de proportions

Yohann.Foucher@univ-nantes.fr

Equipe d'Accueil 4275 "Biostatistique, recherche clinique et mesures subjectives en santé", Université de Nantes

Internat d'Odontologie, 29 Février 2012

UNIVERSITÉ DE NANTES

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

itun institut transplantation urologie néphrologie INSERM - UMR 643

Plan

Plan

Grands échantillons

L'ARS se pose la question de la fermeture d'un service de chirurgie à la vue d'un taux de mortalité élevé : 57 patients opérés pour une pose de valve cardiaque n'ont pas survécu à l'intervention parmi les 300 opérations comptabilisées en 2009. Le taux de mortalité pour ce type d'opérations est 6.4% au niveau national. D'un point de vue statistique, doit-on fermer ce centre ?

- X : Variable aléatoire représentant le nombre de décès parmi 300 opérations.
- $\pi = 0.064$: pourcentage de décès au niveau national.
- Choix des hypothèses :
  - H_0 : $P = \pi$, l'échantillon étudié est issu de la population nationale.
  - H_1 : $P > \pi$, la mortalité est supérieure à la moyenne nationale.

Grands échantillons

- Statistique de test sous H_0.

$$X \sim \mathcal{B}(n, \pi)$$

Comme $n > 30$, $n\pi = 300 * 0.064 = 19 > 5$ et $n(1-\pi) = 300 * (1 - 0.064) = 280 > 5$, la loi binomiale précédente peut être approximée par une loi normale :

$$X \sim \mathcal{N}(n\pi, \sqrt{n\pi(1-\pi)})$$

donc

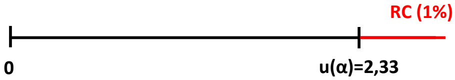
$$P = X/n \sim \mathcal{N}\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Finalement, en centrant et en réduisant P, la statistique de test est :

$$U = \frac{P - \pi}{\sqrt{\pi(1-\pi)/n}} \sim \mathcal{N}(0,1)$$

Grands échantillons

- Risque de première espèce maximum : $\alpha = 1\%$
- Région critique unilatérale positive (RC).



- Application numérique :

$$t = (0.19 - 0.064)/(\sqrt{0.064 * (1 - 0.064)/300}) = 8.91 \in RC$$

- Conclusion : L'hypothèse nulle est rejetée. Le taux de mortalité est significativement* supérieur à la moyenne nationale ($p_c < 1\%$).
  * Evite de véhiculer le sentiment de certitude
- Exercice : Faire la même chose en bilatéral.

Plan

Comparaison de deux proportions

- On considère deux populations $\mathcal{P}_A$ et $\mathcal{P}_B$ desquelles sont extraits deux échantillons de tailles N_A et N_B. On suppose les proportions π_A et π_B d'une caractéristique dans chacune des deux populations. A partir des proportions observés p_A et p_B, on cherche à savoir si les proportions des deux populations peuvent être considérées comme égales, ou bien si elles semblent être différentes.

Comparaison de deux proportions

- Définition des populations et des v.a. :
  - X_A : v.a. continue binaire dans la population $\mathcal{P}_A$ de moyenne π_A.
    $\rightarrow$ On observe un échantillon de taille N_A $\{x_{A,1}, ..., x_{A,N_A}\}$.
  - X_B : v.a. continue binaire dans la population $\mathcal{P}_B$ de moyenne π_B.
    $\rightarrow$ On observe un échantillon de taille N_B $\{x_{B,1}, ..., x_{B,N_B}\}$.
- Choix des hypothèses :
  - H_0 : $\pi_A = \pi_B$ $(= \pi)$
  - H_1 : $\pi_A \neq \pi_B$

Comparaison de deux proportions

- Soit p_i les proportions estimées de π_i telles que :

$$p_i = \frac{\sum_j^{N_i} x_{i,j}}{N_i}$$

- Soit π la proportion commune sous H_0 estimée par :

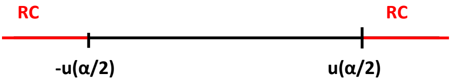
$$p = \frac{N_A p_A + N_B p_B}{N_A + N_B} = \frac{\sum_k \sum_j^{N_k} x_{k,j}}{\sum_k N_k}$$

- Définition de la statistique de test. Sous H_0, si :
  1. N_A et $N_B > 30$,
  2. $\min(N_A\pi, N_B\pi, N_A(1-\pi), N_B(1-\pi)) > 5$,
  3. et les échantillons sont indépendants,

$$U = \frac{P_A - P_B}{\sqrt{\frac{P_A(1-P_A)}{N_A} + \frac{P_B(1-P_B)}{N_B}}} \sim \mathcal{N}(0,1)$$

Comparaison de deux proportions

- Définition de la région critique (α, test bilatéral)



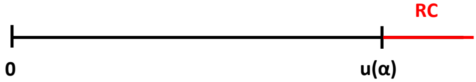
- Application numérique :

$$U = \frac{p_A - p_B}{\sqrt{\frac{p_A(1-p_A)}{N_A} + \frac{p_B(1-p_B)}{N_B}}}$$

- Si $u \in RC \rightarrow p_c < \alpha$.
  - Rejet de H_0 car moins de α % de chance qu'elle soit vraie.
  - Il semble que l'écart entre les proportions des deux populations soit différent.
- Si $u \notin RC \rightarrow p_c > \alpha$.
  - Non rejet de H_0 car plus de α % de chance qu'elle soit vraie.
  - On ne peut pas montrer une différence significative entre les proportions des deux populations.

Cas unilatéral

- Identique au cas bilatéral, mais....
- $H_1 : \pi_A > \pi_B$ (l'hypothèse peut aussi être posée en infériorité)
- Loi normale, α, test unilatéral



- Si $u \in RC \rightarrow p < \alpha$.
  - Il semble que l'échantillon A soit issu d'une population où la proportion π_A est supérieure à la proportion π_B.
- Si $u \notin RC \rightarrow p > \alpha$.
  - On ne peut pas montrer que la proportion de la population A soit supérieure à celle de B.

Plan

Comparaison de 2 proportions (petits échantillons)

- Une des conditions suivantes n'est pas respectée :
  - N_A et $N_B > 30$,
  - $\min(N_A\pi,\ N_B\pi,\ N_A(1-\pi),\ N_B(1-\pi)) > 5$,
- Utilisation du test exact de Fisher.

Test exact de Fisher

- Comparaison de 2 traitements A et B contre infection nosocomiale
  $\Rightarrow$ proportions de patients infectés
- Hypothèses :
  $H_0 : \pi_A = \pi_B = \pi$
  $H_1 : \pi_A \neq \pi_B$
- Etude pilote : N=16 patients
  $\rightarrow$ 7 patients dans le groupe A
  $\rightarrow$ 9 patients dans le groupe B

Test exact de Fisher

On observe :

	Infectés	Non infectés	Total
Groupe A	6	1	7
Groupe B	2	7	9
Total	8	8	16

Idée

Calculer sous H_0 la probabilité (exacte) d'obtenir un écart entre les groupes $\geq$ à celui observé en conservant les mêmes totaux (lignes et colonnes)

Test exact de Fisher

On observe :

	Infectés	Non infectés	Total
Groupe A	6	1	7
Groupe B	2	7	9
Total	8	8	16

Une seule autre possibilité avec les mêmes totaux

	Infectés	Non infectés	Total
Groupe A	7	0	7
Groupe B	1	8	9
Total	8	8	16

Test exact de Fisher

Il faut savoir que la probabilité d'observer la configuration suivante :

	Modalité 1	Modalité 2	Total
Groupe 1	a	b	a+b
Groupe 2	c	d	c+d
Total	a+c	b+d	N

est égale à

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}$$

Test exact de Fisher

On observe :

	Infectés	Non infectés	Total
Groupe A	6	1	7
Groupe B	2	7	9
Total	8	8	16

Probabilité d'obtenir cette configuration : p_1

$$p_1 = \frac{7!9!8!8!}{6!1!2!7!16!}$$

Exemple

Seconde configuration avec même totaux mais des écarts entre les groupes supérieurs à ce qu'on observe :

	Infectés	Non infectés	Total
Groupe A	7	0	7
Groupe B	1	8	9
Total	8	8	16

Probabilité d'obtenir cette configuration : p_2

$$p_2 = \frac{7!9!8!8!}{7!0!1!8!16!}$$

Exemple

$\Rightarrow$ Probabilité d'avoir l'une ou l'autre des configurations possibles :

$$p = p_1 + p_2 = \frac{7!9!8!8!}{6!1!2!7!16!} + \frac{7!9!8!8!}{7!0!1!8!16!} \approx 0.0203$$

- Hypothèse bilatérale $\Rightarrow$ L'écart doit être envisagé dans les 2 sens

$$p_c = 2 * p = 0.0406 < 0.05$$

$\Rightarrow$ Rejet de H_0 au seuil $\alpha = 5\%$

Plan

Comparaison de plus de deux proportions

- Toutes les tests précédents ne sont pas adaptés si plus de 2 proportions à comparer.
- Exemple : Montrer que le pourcentage de rémission est différent selon trois stades de cancer (A, B et C).
- Présentation 3 groupes (stades) $\times$ 2 modalités (rémission oui/non) :

	Modalité 1	Modalité 2	Total
Groupe A	o_{11}	o_{12}	$n_{1.}$
Groupe B	o_{21}	o_{22}	$n_{2.}$
Groupe C	o_{31}	o_{32}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Test du Chi-deux (Chi-square)

	Modalité 1	Modalité 2	Total
Groupe A	o_{11}	o_{12}	$n_{1.}$
Groupe B	o_{21}	o_{22}	$n_{2.}$
Groupe C	o_{31}	o_{32}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

- Choix des hypothèses :
  - H_0 : Il n'y a pas de lien entre les deux variables ($\pi_A = \pi_B = \pi_C$).
  - H_1 : Il y a un lien entre les deux variables (au moins deux proportions sont différentes).
- Sous H_0, calcul des effectifs théoriques, c'est à dire les effectifs qu'on aurait observés en cas d'égalité des proportions.

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$

Test du Chi-deux (Chi-square)

	Modalité 1	Modalité 2	Total
Groupe A	o_{11} e_{11}	o_{12} e_{12}	$n_{1.}$
Groupe B	o_{21} e_{21}	o_{22} e_{22}	$n_{2.}$
Groupe C	o_{31} e_{31}	o_{32} e_{32}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

- Statistique de test
- Sous H_0,

$$\chi^2 = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{2ddl}$$

- Le nombre de degrés de liberté de la loi est défini par le nombre de colonnes - 1 multiplié par le nombre de lignes - 1 : (3-1) $\times$ (2-1) $=$ 2.

Plan

Généralisations

	Modalité 1	Modalité 2	Total
Groupe 1	a	b	a+b
Groupe 2	c	d	c+d
Total	a+c	b+d	N

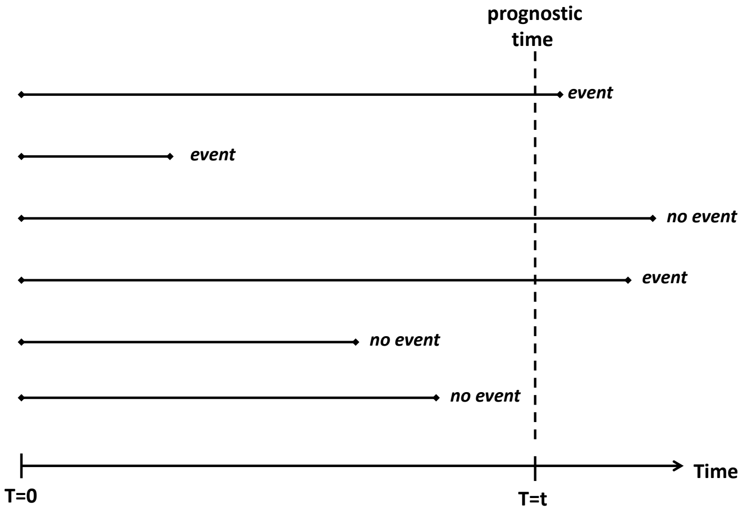
- Odds Ratio :

$$OR = \frac{a \times d}{b \times c}$$

$$IC_{95\%} = OR^{(1 \pm 1.96/\sqrt{\chi^2})}$$

- Pour prendre en compte des facteurs de confusions : régression logistique.
- Attention aux données censurées : nécessité d'utiliser des modèles de survie (Kaplan-Meier, Cox, etc.)

Problème des données censurées



Problème des données censurées