

Exercices de biostatistique

Rappel: pour visualiser la formule associée aux résultats obtenus, il vous suffit d'aller cliquer sur la case concernée(uniquement dans excel et non avec "Adobe Acrobat") !!

Tests d'hypothèse sur les pourcentages

Exercice 1

Sur 1270 vaches laitières d'une région étudiée, 67 % d'entre-elles possèdent un certain gène.

a) A-t-on une majorité de bêtes portant ce gène? b) Est-elle significative?

a) La proportion $\hat{p}$ est de 0,67 et donc q est égal à $(1-0,67)=0,33$ et la taille de l'échantillon est 1270. $N \cdot p$ étant >5 , on peut utiliser une distribution normale et dire qu'il s'agit d'une majorité, il reste maintenant à vérifier si elle est significative ou non!

b) Hypothèse nulle: pas de majorité, 50 % le possédant et 50 % ne le possédant pas!

$$\begin{array}{ll} \mu = & 50\% \\ \hat{p} = & 67\% \end{array}$$

$$\sigma = \sqrt{\frac{pq}{n}} \quad 0,01403034$$

$$z = 12,1166002$$

Cette valeur de z étant très élevée, on peut en déduire que l'hypothèse nulle est rejetée et donc que la majorité observée est bel et bien significative!

Exercice 2

Dans une population, une affection particulière apparaît dans 30 % des cas.

Un échantillon de taille $n=64$ indique une proportion d'atteints de $\hat{p} = 0,46875$.

Le pourcentage est-il significativement différent de 0,3?

Quel test? Uni-ou bilatéral? Donner 2 méthodes d'analyse!

Hypothèse nulle: pas de différence entre les proportions rencontrées!

a) Test de z unilatéral droit:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$2,94594152$$

Cette valeur de z étant largement supérieure à celle dans la table au seuil 95 %, qui est de 1,645(unilatéral droit) et donc, nous rejetons l'hypothèse nulle, les proportions sont différentes!

b) Test de χ^2 , toujours bilatéral:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$8,67857143$$

$$\text{avec } \chi^2 = Z^2$$

Ei	Oi
19,2	30
44,8	34

La conclusion est la même qu'avant, les proportions sont différentes!

Exercice 3

Dans quelle situation, 0,6875 sera une majorité significative au seuil 5%?

Il s'agit d'un test unilatéral droit, donc $Z(0,95)=1,645$!

Suivant la formule déjà rencontrée:

$$Z = \frac{0.6875 - 0.5}{\sqrt{\frac{0.5^2}{n}}} = 0.375\sqrt{n}$$

Soit $n = (1,645/0,375)^2 =$

19,2428444

Si la taille de l'échantillon est de 20.

Exercice 4

Deux vaccins ont donné les pourcentages de survivants suivants: 15/34 et 12/60.
Ces pourcentages sont-ils significativement différents?

Hypothèse nulle: les pourcentages ne sont pas significativement différents, taux de survie identiques!

La comparaison de deux pourcentages observés requière l'utilisation de la formule suivante:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{avec: } \bar{p} = \frac{(y_1 + y_2)}{(n_1 + n_2)} \quad \text{et} \quad \bar{q} = 1 - \bar{p}$$

$\bar{p} =$ 0,28723404 et $\bar{q} =$ 0,71276596

Z= 2,48310259

Ho est rejetée au seuil 5 % car pour ce seuil la valeur de Z=1,96(test bilatéral).

Exercice 5

Quatre taureaux ont les pourcentages d'anormaux suivants chez leurs descendants.

T1	15%	(n=100)
T2	11%	(n=200)
T3	20%	(n=240)
T4	10%	(n=120)

Ces pourcentages sont-ils significativement différents?

La comparaison de plus de 2 pourcentages requière l'utilisation d'une table de contingence RX2.

Taureau	n	% anormaux	Anormaux	Normaux
1	100	15	15	85
2	200	11	22	178
3	240	20	48	192
4	120	10	12	108
Somme			97	660

H 0: $p_1=p_2=p_3=p_4=p=$ 0,1469697 et donc $q=$ 0,8530303

Attendus: Anormaux Normaux
 14,6969697 85,3030303 $\chi^2=$ 9,6828844
 29,3939394 170,606061 Cette valeur étant supérieure à celle trouvée dans la
 35,2727273 204,727273 table de χ^2 0,95 avec 3 ddl et donc, H 0 est rejetée.
 17,6363636 102,363636 ou encore $P(\chi^2 <) =$ 0,0214633

Exercice 6

Un généticien pense qu'un allèle (variant génétique) est présent dans une population avec une fréquence de 30 %. Des mesures effectuées sur une autre population ont permis de calculer que, dans cette autre population, la fréquence est de 50 %. Combien de sujets le généticien doit-il évaluer pour pouvoir rejeter, dans sa population, l'hypothèse nulle $H_0: p=0,5$ avec une puissance $> 80 \%$.

Les 2 populations suivent une distribution normale avec:

$$N_1(0.3, \sqrt{\frac{0.3 * 0.7}{n}})$$

$$N_2(0.5, \frac{0.5}{\sqrt{n}})$$

La première chose à faire est de trouver la valeur de p:

$$-1.645 = \frac{p - 0.5}{\frac{0.5}{\sqrt{n}}}$$

et donc :

$$p = 0.5 - \frac{0.8225}{\sqrt{n}}$$

La valeur de z correspondant à une puissance $>80 \%$ est obtenue comme suit: 0,84162139

$$Z = 0.84162139 = \frac{p - 0.3}{\sqrt{\frac{0.3 * 0.7}{n}}} = \frac{0.2 * \sqrt{n} - 0.8225}{\sqrt{0.21}}$$

Et donc $n = 36,447564$
Il faudra donc évaluer au minimum 37 sujets!!!

Exercice 7

Suite aux résultats suivants obtenus, vérifiez l'intérêt d'un régime alimentaire strict, chez le chien, dans la lutte contre l'infarctus du myocarde? Cet intérêt est-il significatif?

	Vivants	Morts	Totaux
Régime strict	28	22	50
Pas de régime	12	38	50
Totaux	40	60	100

Utilisez deux méthodes!

H_0 : Pas d'intérêt du régime étudié dans la lutte contre l'infarctus. Pas de différence entre traités et non traités.

Calcul des "attendus":

	Vivants	Morts	Totaux
Régime strict	20	30	50
Pas de régime	20	30	50
Totaux	40	60	100

Calcul du χ^2 :

10,6666667

$P(\chi^2 <)$:

0,00109084

H_0 est rejetée car moins de 0,2% d'obtenir un tel χ^2 , avec 1 degré de liberté. Et donc, il y a bien un intérêt de l'instauration d'un tel régime, pour lutter contre l'infarctus du myocarde.

Calcul de Z:

3,26598632

0,00109096

Ce qui mène bien à : $Z^2 = \chi^2$.

Exercice 8

Dans une étable de 30 vaches, 12 sont atteintes d'une affection parasitaire.
L'année précédente, il y avait 28 vaches dans l'étable et seulement 6 étaient atteintes.
Y a-t-il aggravation du phénomène cette année ?

Comparaison de deux proportions: 12/30 et 6/28

p1 0,4
p2 0,21428571
p 0,31034483
q 0,68965517

Approche Z

$z = (p1-p2)/\text{racine}(pq(1/n1+1/n2)) = 1,52768109$ $P = 0,12659178$
 $z^2 = 2,33380952$

Approche Chi²

12	18	30
6	22	28
18	40	58

9,31034483	20,6896552	Attendus
8,68965517	19,3103448	

Chi² = 2,33380952 P = 0,12659178

Exercice 9

Une maladie génétique à caractère dominant ségrège dans 30 familles (parents plus enfants).
Dans chaque famille, un seul des parents est atteint. La maladie est supposée rare,
ce qui implique qu'il est peu probable qu'un individu atteint soit homozygote. Parmi les 50 enfants
auscultés, 20 sont diagnostiqués atteints par l'affection. Ces chiffres supportent-ils l'hypothèse
de caractère dominant ?
Déterminez les limites en dehors desquelles l'hypothèse de dominance devra être rejeté au seuil 5%.

Maladie dominante - Un parent hétérozygote => P (enfant atteint) = 0.5

Sur 50 enfants, on attend (sous H0) 25 atteints. On en a 20.
Quelle est la probabilité d'avoir 20 atteints (ou -) alors qu'on en attendait 25 ?

Approximation gaussienne:

$\mu = 25$
 $\sigma = \text{racine}(npq) = 3,53553391$
 $zs -1,41421356$
 $P(z < zs) = 0,07864965$

Loi binomiale:

$0,10131938 = B(0) + B(1) + \dots B(20)$

Limites (5% bilatéral)

LI = $25 - 1,96 * 3,536 = 18,0703535$
LS = $25 + 1,96 * 3,536 = 31,9296465$

Exercice 10

Trois préparations d'un même vaccin contre la maladie de Marek dans la volaille ont assuré la protection suivante:

Préparation	Vaccinés	Atteints
I	54	16
II	55	25
III	38	4

Ces trois préparations ont-elles des pouvoirs protecteurs différents ?

On pourrait comparer les vaccins 2 à 2 en utilisant la procédure décrite dans la question précédente, ce qui permettrait de répondre à la question de savoir si deux préparations particulières ont des pouvoirs protecteurs différents. Il y a trois objections à cette approche:

- 1) La question ici est plus globale On désire savoir s'il y a des différences entre préparations, sans spécifier quelle préparation diffère de telle autre.
- 2) L'approche peut devenir lourde si le nombre de préparations augmente (en fait, le nombre de tests à effectuer est $n(n-1)/2$, où n est le nombre de préparations).
- 3) En effectuant plusieurs tests, on augmente le risque de faux positifs (en moyenne, si on fait 20 tests, on trouvera 1 faux positif !).

En conclusion, une autre approche est souhaitable. Une approche globale est de voir la table comme une table de contingence, ce qui donne:

Observé

Prépa	Atteints	Sains	Total
I	16	38	54
II	25	30	55
III	4	34	38
Total	45	102	147

Attendu

Prépa	Atteints	Sains	Total
I	16,5306122	37,4693878	54
II	16,8367347	38,1632653	55
III	11,6326531	26,3673469	38
Total	45	102	147

La valeur de κ^2 se calcule comme d'habitude:

$$\kappa^2 = (16 - 16,53)^2/16,53 + \dots + (34 - 26,37)^2/26,37 = \mathbf{12,9461898}$$

Dans la table avec deux degrés de liberté, la probabilité d'une telle valeur (ou d'une valeur plus grande) est de : **0,00154444**

Il est donc très peu probable (moins d'une chance sur 600) d'avoir de telles différences entre échantillons si les préparations sont équivalentes. La conclusion logique est que les préparations diffèrent.

Exercice 11

Deux préparations A et B sont administrées à 50 patients. 16 réagissent mieux à la première préparation, 22 à la seconde (les autres ne réagissent pas mieux à l'une qu'à l'autre). Y a-t-il en fait plus de patients qui préfèrent B à A ? (Utilisez deux méthodes)

Données paarées :seules les réponses discordantes nous intéressent

H0: pas de différence => $(16+22)/2 = 19$ patients préfèrent chaque prépa.
cad $p = 0.5$

Test (Mac Nemar):

$$\chi^2 = (16-19)^2/19 + (22-19)^2/19 = 18/19 = 0,94736842 \Rightarrow \text{NS}$$

Test binomial

$$\mu = 19$$

$$X = 22$$

$$\sigma = \sqrt{npq} = 3,082207$$

$$z = 0,97332853$$

$$0,94736842$$

Intervalle de confiance

$$P\{ (16/38) - 1,96 \cdot \sqrt{(16/38) \cdot (22/38) / 38} \leq p \leq (16/38) + 1,96 \cdot \sqrt{(16/38) \cdot (22/38) / 38} \} = 0,95$$

$$pi = 0,26609552$$

$$ps = 0,57803541$$

=> $p=0,5$ appartient a l'intervalle de confiance => acceptation de H0

Exercice 12

Les sujets atteints d'une certaine affection du sang présentent une anomalie au niveau d'une paire chromosomique particulière. Néanmoins, on n'observe pas l'anomalie chaque fois qu'on examine cette paire chromosomique. On voudrait estimer la proportion des cas où cette anomalie est présente; sur 120 sujets atteints, on examine 5 mitoses, et l'on obtient le tableau suivant:

Mitoses avec anomalies	0	1	2	3	4	5
Fréquence	6	31	42	29	10	2

Dans quel pourcentage cette anomalie est-elle présente ?

La proportion est évidemment estimée par le nombre de mitoses anormales sur le nombre total de mitoses observées.

On a observé $(120 \cdot 5)$ mitoses, soit 600 mitoses.

Le nombre de mitoses anormales est:

$6 \cdot 0$	soit	0
$31 \cdot 1$		31
$42 \cdot 2$		84
$29 \cdot 3$		87
$10 \cdot 4$		40
$2 \cdot 5$		10

ce qui donne un total de **252** mitoses anormales.

La proportion est donc de $(252/600)$, soit **0,42**

Une autre possibilité serait de calculer la moyenne (pondérée) du nombre de mitoses anormales

# mit. ano.	0	1	2	3	4	5
Fréq.	6	31	42	29	10	2
Fréq. Relative	0,05	0,25833333	0,35	0,24166667	0,08333333	0,01666667

La moyenne pondérée est alors: $(0.05 \cdot 0) + (0.258333 \cdot 1) + \dots + (0.01666 \cdot 5) = \mathbf{2,1}$

Ce nombre est le nombre attendu de mitoses anormales dans 5 mitoses. La probabilité qu'une mitose particulière soit anormale est donc: $(2,1 / 5) =$ **0,42**

L'intervalle de confiance peut être estimé par: $p \pm Z(0.025) \cdot \sqrt{p \cdot q / n}$

On obtient alors: $p \pm 1.96 \cdot \sqrt{0.42 \cdot 0.58 / 600} = 0,42 \pm 0,04$ au seuil 95%.

La vraie valeur de p est donc comprise (avec 95 % de chances) entre **0,39** et **0,45**.

Exercice 13

Dans une région, une maladie a une prévalence de 10 %. Pour tester si une région voisine est dans les mêmes conditions concernant cette maladie, 200 individus sont testés, et 30 d'entre eux sont déclarés malades (on supposera que le test de détection est parfait, pour la simplicité). Utilisez une méthode exacte et trois méthodes approximatives pour répondre à la question: la région testée est-elle plus atteinte que la région de référence ?

En premier lieu, il faut constater que l'échantillon prélevé dans la "région voisine" est plus atteint que la région témoin, car $(30 / 200) =$ **0,15** qui est supérieur à 0,10.

Deux possibilités pour expliquer cette observation: soit la fréquence dans la région voisine est effectivement 10%, et la valeur mesurée sur l'échantillon ne s'écarte de cette valeur que par fluctuation d'échantillonnage, soit effectivement la prévalence dans cette région est plus élevée. Pour distinguer entre les deux possibilités, on peut se poser la question suivante: si la fréquence est effectivement de 10% dans la région voisine (H_0), quelle est la probabilité d'observer 30 atteints (ou plus) dans un échantillon de taille 200 ? Si cette probabilité est anormalement petite (par exemple $< 5\%$), on admettra alors que cette hypothèse est peut-être vraisemblable, et qu'il est plus raisonnable de considérer que la prévalence est plus grande dans la région voisine.

L'approche binomiale fournit (avec un ordinateur !) une méthode exacte de calcul de cette probabilité:

n	200
p	0,1
r	30
Proba	$= 1 - \text{LOI.BINOMIALE}(r; n; p; \text{VRAI})$ soit, 0,00950831

Il y a donc moins d'une chance sur 100 d'obtenir une pareille fluctuation, ce qui nous conduit à rejeter l'hypothèse nulle: il y a une prévalence plus grande de la maladie dans la région voisine.

Si aucun ordinateur n'est disponible, le calcul évoqué plus haut est très fastidieux, et il est préférable de se tourner vers des méthodes approximatives. L'approximation classique de la loi binomiale est d'utiliser une distribution normale; formellement, on calcule:

$$z = (X - \mu) / \sigma = (r - n \cdot p) / \sqrt{n \cdot p \cdot q}$$

et on teste si la valeur de z est anormale. On obtient ici:

$$z = (30 - 200 \cdot 0.1) / \sqrt{200 \cdot 0.1 \cdot 0.9}$$

soit, $z =$ **2,3570226**

La probabilité d'une telle valeur (ou d'une valeur) plus grande est donnée dans les tables.

$$P(z > 2,357) =$$
 0,00921105

La valeur est très proche du résultat fourni par la méthode exacte.

Une approche équivalente est de raisonner sur les pourcentages. L'approximation gaussienne donne alors:

$$z = (X - \mu) / \sigma = (p_{\text{est}} - p) / \sqrt{(p * q)/n}$$

Dans cette équation, p_{est} est la valeur estimée (sur l'échantillon) de p , et vaut: $(30 / 200) = 0.15$. Il est clair que cette formulation est complètement équivalente à la précédente (on a simplement divisé le numérateur et le dénominateur par n).

Une troisième méthode approximative serait de raisonner en termes de χ^2 . On a alors:

	Observés	Attendus (si H_0 est vraie)
Malades	30	20
Sains	170	180

On peut calculer que $\chi^2 = (30 - 20)^2 / 20 + (170 - 180)^2 / 180 =$ **5,55555556**

A nouveau, on peut trouver la probabilité d'être supérieur à cette valeur dans une table de χ^2 avec 1 degré de liberté. On obtient:

$$P(\chi^2 > 5,5555) =$$
 0,01842213

Cette probabilité est donc la probabilité de s'écarter de 10 ou plus des valeurs attendues. Elle est par conséquent bilatérale (on s'écarte de + ou - 10), et la probabilité unilatérale s'obtient en divisant cette probabilité par 2, soit: **0,00921106**

Comme on le sait, un χ^2 avec 1 degré de liberté est équivalent à un z^2 , ce qui apparaît clairement ici (remarquez que $z^2 = 2.357^2 = 5.5555 = \chi^2$)

Exercice 14

Deux traitements sont administrés à deux groupes d'animaux pour améliorer leurs performances zootechniques, le but de l'expérience étant de comparer l'efficacité des deux approches. L'amélioration se traduisant par une prise de poids, une augmentation de poids de 10 kgs dans la période examinée sera considérée (arbitrairement...) comme significative. À l'issue du test, 16 animaux parmi les 38 du premier groupe, et 22 parmi les 40 du second groupe ont grossi de plus de 10 kgs. Y a-t-il une différence d'efficacité entre les deux traitements ? Et si non, quelle devait être le nombre d'individus "améliorés" dans le second groupe pour pouvoir considérer que le second régime est plus efficace que le premier

Une différence d'efficacité des traitements se traduirait dans ce cas par une différence significative des proportions d'"améliorés" entre les deux échantillons. Remarquez toutefois que "efficacité" signifie dans ce contexte "capacité à grossir de 10 kgs ou plus pendant la période

On a donc une efficacité de 16 sur 38 pour le premier traitement, soit **0,42105263** et pour le second traitement, de 22 sur 40, soit **0,55**

Sous l'hypothèse nulle (pas de différence entre les traitements), les deux échantillons pourraient être considérés comme issus de la même population, avec une efficacité de 38 sur 78.

En résumé,	p_1	0,42105263	Efficacité du premier traitement
	p_2	0,55	Efficacité du second traitement
	p	0,48717949	Efficacité globale (si H_0 est vraie)

On peut comparer les deux traitements en utilisant la statistique z:

$$z = (p_1 - p_2) / \sqrt{p * q * (1/n_1 + 1/n_2)}$$

qui donne ici: $z = (0,421 - 0,55) / \sqrt{0,487 * 0,513 * (1/38 + 1/40)} = -1,13883232$

La probabilité d'obtenir une valeur aussi petite de z ou plus petite mesure la probabilité que le premier échantillon s'écarte aussi fort (en négatif) de la valeur estimée pour la population (p).
Le test est bilatéral (on regarde si p1 diffère de p2), et la probabilité qui nous intéresse est donc:

$$P(z < -1,13883232) + P(z > 1,13883232) = 2 * P(z > 1,13883232) = 0,2547732$$

Une telle différence n'est donc pas rare (sous l'hypothèse nulle), et on ne peut pas rejeter H0.
Combien aurait-il fallu d'individus améliorés dans le second groupe pour rejeter H0 (au seuil 5%)?

La valeur z correspondante est z = **-1,96**

Si on appelle X le nombre d'individus améliorés dans le groupe 2, on peut écrire que:

	Améliorés	Total		
p1	16	38	=>	0,42105263
p2	X	40	=>	X/40
p	16+X	78	=>	(16+X)/78
q	78 - (16 + X)	78	=>	(62-X)/78

$$z = (16/38 - X/40) / \sqrt{(16+X)/78 * (62-X)/78 * (1/38 + 1/40)} = -1.96$$

en augmentant X, on arrive finalement à:

X	25	z	-1,80299454
X	26	z	-2,02735058

Il faut donc 26 individus améliorés ou plus dans le second groupe pour pouvoir conclure à l'efficacité du traitement 2 par rapport au traitement 1 (avec une erreur de type I de 5%).