- Licence Sciences et technologie
- Mention science de la vie – L2

SCIENCES SORBONNE UNIVERSITÉ

INTRODUCTION A LA MODELISATION EN BIOLOGIE (BM1) - 2V382

Modelisation statistique (Cours 4)

Martin LARSEN

www.immulab.fr

Principaux tests univariés sous R

Mention science de la vie – L2

Choix de test dépend de la caractéristique de l'étude

Level of Measurement	Sample characteristics					Correlation
	1 Sample	2 Sample		K Samples (K>2)		
		Independent	Dependent	Independent	Dependent	
Categorical or Nominal	χ^2 conformity or homogeneity	χ^2 independence	McNemar χ^2	χ^2 independence	Cochran's Q (CMH)	
Rank or ordinal	χ^2 or Mann Whitney U*	Mann Whitney U	Wilcoxon Matched Pairs Signed Ranks	Kruskal Wallis H	Friedman's ANOVA	Spearman's rho
Parametric (interval & Ratio)	Z test or t test	T test between groups	T test within groups	1 way ANOVA between groups	1 way ANOVA (within or repeated measure)	Pearson's r
		Factorial (2 way) ANOVA				

* H_0 : μ = valeure theorique, Correspond à un Wilcoxon signed ranks, test, lorsque le data est centré sur la valeur théorique

Test de proportions et χ^2

Mention science de la vie – L2

Comparer deux proportions (Z) ou un distribution (χ^2 de homogénéité)

Prenons le cas de deux grossistes qui obtiennent après un contrôle qualité le tableau suivant :

Obs	A	B	Total
Défaut	12	15	27
Total	96	55	151
P	0.13	0.28	

On cherche à savoir si les deux grossistes se fournissent chez le même fabricant ? On va comparer le nombre de défauts dans les deux échantillons

- defauts <- c(12, 15)
- total <- c(96, 55)
- Comparer deux proportions a la main

X VAD « # defaut »
F=X/n suit N(φ,φ(1-φ)/n)

Conditions:
$n_A\varphi > 5$ ET $n_A(1-\varphi) > 5$
$n_B\varphi > 5$ ET $n_B(1-\varphi) > 5$

Hypothèse:
H_0: $\varphi_A = \varphi_B$
H_1: $\varphi_A \neq \varphi_B$

Sous H_0

$Z_c = \frac{p_B - pA}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_B} + \frac{\hat{p}(1-\hat{p})}{n_A}}}$; Z~N(0,1)

$\hat{p} = \frac{12+15}{96+55} = 0.179$ =>

$Z_c = \frac{0.28-0.13}{\sqrt{\frac{\hat{p}(1-\hat{p})}{96} + \frac{\hat{p}(1-\hat{p})}{55}}} = 2.28$

p-value = 0.02263 => rejet H_0

Comparer deux proportions (Z) ou un distribution (χ^2 de homogénéité)

Prenons le cas de deux grossistes qui obtiennent après un contrôle qualité le tableau suivant :

Obs	A	B	Total
Bonne	84	40	124
Défaut	12	15	27
Total	96	55	151
P	0.13	0.28	

Sous H0 →

Theo	A	B	Total
Bonne	78.8	45.2	124
Défaut	17.2	9.8	27
Total	96	55	151

On cherche à savoir si les deux grossistes se fournissent chez le même fabricant ? On va comparer le nombre de défauts dans les deux échantillons

- defauts <- c(12, 15)
- total <- c(96, 55)
- prop.test(defauts, total, correct = FALSE)

- pieces <- matrix(c(12,15,84,40),nrow=2, byrow=F)
- chisq.test(pieces, correct = FALSE)

Prop.test et chisq.test effectuent une test chi-2 –> équivalence parfait avec test de Z (limite: 2 échantillons).

Hypothèse:
H0: Distribution de défaut et bonne homogène entre grossistes A et B
H1: Distribution de défaut et bonne non-homogène entre grossistes A et B

Conditions:
Deux echantillon independant sous H_0
80% des valeurs attendues >5 (Cochran)

Variable de décision:

$$\chi^2 = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

ddl = $(n_i-1)(n_j-1)$

p-value = $P(\chi^2 > \chi^2_{obs})$

Conclusion statistique/biologique

Comparer deux proportions (Z) ou un distribution (χ^2 de homogénéité)

Prenons le cas de deux grossistes qui obtiennent après un contrôle qualité le tableau suivant :

Obs	A	B	Total
Bonne	84	40	124
Défaut	12	15	27
Total	96	55	151
P	0.13	0.28	

Theo	A	B	Total
Bonne	78.8	45.2	124
Défaut	17.2	9.8	27
Total	96	55	151

On cherche à savoir si les deux grossistes se fournissent chez le même fabricant ? On va comparer le nombre de défauts dans les deux échantillons

- defauts <- c(12, 15)
- total <- c(96, 55)
- prop.test(defauts, total, correct = FALSE)

- pieces <- matrix(c(12,15,84,40),nrow=2, byrow=F)
- chisq.test(pieces, correct = FALSE)

Prop.test et chisq.test effectuent une test chi-2 –> équivalence parfait avec test de Z (limite: 2 échantillons).
Z_{obs}=2.28 => Z^2_{obs}=5.197 = χ^2_{obs} qui suit χ^2 avec (2-1(2-1)=1 ddl

Hypothèse:
H0: Distribution de défaut et bonne homogène entre grossistes A et B
H1: Distribution de défaut et bonne non-homogène entre grossistes A et B

Conditions:
Deux echantillon independant sous H_0
80% des valeurs attendues >5 (Cochran)

Variable de décision:

$$\chi^2 = \sum_{i=1}^{n_i}\sum_{j=1}^{n_j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

ddl = (n_i-1)(n_j-1)

p-value = P(χ^2> χ^2_{obs})

Conclusion statistique/biologique

Test du χ^2 de conformité

- Monsieur GRAIN semencier de Monsieur LEON (cultivateur) affirme que la variété de semences achetées par Mr LEON donne

Jaune: 50% de pieds à épis
Jaune/rouge : 30% de pieds à épis (ou orange plus simple en R)
Rouge: 20% de pieds à épis

- Lors de la récolte Mr LEON **observe** :

Jaune: 48 épis
Jaune/rouge: 22 épis
Rouge: 29 épi

Test du χ^2 de conformité

Hypothèse:
H0: La distribution observée lors de la récolte est conforme a celle annoncée par Mr GRAIN
H1: La distribution observée lors de la récolte n'est pas conforme a celle annoncée par Mr GRAIN

- Script :

```
> ##### CHI CARRE #####
> #### Monsieur Leon et son Mais ###
> ###################################
> # créons un vecteur avec les effectifs observés
> obs<-c(48,22,29)
># Transformons le en table pour avoir des entêtes de colonnes
> obs<-as.table(obs)
> obs
> # affichons les noms de colonnes
> names(obs)
```

Variable de décision:

$$\chi^2 = \sum_{j=1}^{n_i} \frac{(O_i - Ei)^2}{E_i}$$

ddl = (n_i-1)

p-value = P($\chi^2 > \chi^2_{obs}$)

Conclusion statistique/biologique

Test du χ^2 de conformité

- Script :

```
> # changeons les entêtes de colonnes avec les couleurs de grains
> names(obs)<-c("Jaune","Jaune.Rouge","Rouge")
> obs
> addmargins(obs) # tableau de contingence
># calcul des prop observées #
> pi.obs<-obs/margin.table(obs)
# Proportions théoriques
> pi.theo<-c(0.5,0.3,0.2)
```

> chisq.test(obs,p=pi.theo)

Test du χ^2 d'Independence

- En plus de la couleur, on travaille avec le caractère d'enracinement.

	Faible	Fort	Moyen	Tresfort
Jaune	13	6	17	12
Orange	2	7	3	10
Rouge	3	13	8	5

```
> ## On rentre une matrice avec 3 lignes
> mat<-matrix(c(13,2,3,6,7,13,17,3,8,12,10,5),nrow=3)
> ## noms des lignes
> rownames(mat)<-c("Jaune","Orange","Rouge")
> ## noms des colonnes
> colnames(mat)<-c("Faible","Fort","Moyen","Tresfort")
```

Test du χ^2 d'Independence

- Script :

```
> ## Tableau de contingence
> addmargins(mat)
       Faible Fort Moyen Tresfort Sum
Jaune      13    6    17       12  48
Orange      2    7     3       10  22
Rouge       3   13     8        5  29
Sum        18   26    28       27  99
> ## graphe du tableau de contingence
> mosaicplot(mat)
> ## test du chi 2 d'Independence
> chisq.test(mat)
        Pearson's Chi-squared test
data: mat
X-squared = 17.9626, df = 6, p-value = 0.006326
```

Hypothèse:
H0: Couleur et enracinement sont indépendant
H1: Couleur et enracinement ne sont pas indépendant

Conditions:
Un échantillon sur lequel 2 variables catégoriques sont évaluées
Independence sous H_0 (données appariées analysées avec McNemar)
80% des valeurs attendues >5 (Cochran)

Variable de décision:

$$\chi^2 = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \frac{(O_{ij} - Ei_j)^2}{E_{ij}}$$

ddl = $(n_i\text{-}1)(n_j\text{-}1)$

p-value = $P(\chi^2 > \chi^2_{obs})$

Conclusion statistique/biologique

Test du χ^2 appariées avec McNemar

Deux tests de diagnostic du VIH (E et I) évalués chez 100 personnes infectées par le VIH.

	I+	I-	Total
E+	60 (a)	12 (b)	72
E-	5 (c)	23 (d)	28
Total	65	35	100

Applique test de McNemar:

```
> df <- cbind(c(60,5),c(12,23))
> colnames(df) <- c("I+", "I-")
> row.names(df) <- c("E+", "E-")
> mcnemar.test(df, correct=F)
```

$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi^2$ avec 1 ddl

$\chi^2_{obs} = \frac{(12-5)^2}{12+5} = 2.88 < \chi^2_{\alpha}$ =3.84; Ne rejet pas H_0 =>

Conclusion: Les deux tests sont également sensible (note; toutes individus inclus sont malade (VIH+)).

Sous H_0 : Se(E) = P(E+|M) = π_{E+} = Se(I) = P(I+|M) = π_{I+}

Hypothèse:
H0: $\pi_{E+} = \pi_{I+}$ (homogénéité marginal)
H1: $\pi_{E+} \neq \pi_{I+}$

Conditions:
Dépendance sous H_0 (données appariées)
2x2 problème (généraliser avec Cochran–Mantel–Haenszel statistique stratifie)

Note:
Le test de McNemar est identique à comparer une proportion discordante « P(E+ ∩ I-|Discordante) » à une proportion théorique (0,5) en évaluant uniquement les combinaisons discordantes (E+ ∩ I-) et (E- ∩ I+)

```
> prop.test(x=5,n=(5+12),p=0.5,
                                correct=FALSE)
```

Test du χ^2 appariées avec McNemar

Douleur stratifier par couleur de cheveux (indépendance)

Non-appariées	Blond(e)	Brun(e)	Total
Pain	21 (a)	68 (b)	89
No Pain	38 (c)	142 (d)	180
Total	59	210	269

```
> df <- cbind(c(21,38),c(68,142))
> colnames(df) <- c("Blond(e)", "Brun(e)")
> row.names(df) <- c("Pain", "No Pain")
```

Applique test de χ^2 d'indépendance:

```
> chisq.test(df, correct=F) # χ²obs = 0.21 => p=0.64
```

Conclusion: Douleur et couleur de cheveux sont indépendant.

Hypothèse:
H0: Douleur et couleur de cheveux sont indépendant.
H1: Douleur et couleur de cheveux ne sont pas indépendant.

Conditions:
Un échantillon sur lequel 2 variables catégoriques sont évaluées.
Indépendance sous H_0
80% des valeurs attendues>5 (Cochran)

Test du χ^2 appariées avec McNemar

Douleur stratifier par couleur de cheveux (indépendance)

Non-appariées	Blond(e)	Brun(e)	Total
Pain	21 (a)	68 (b)	89
No Pain	38 (c)	142 (d)	180
Total	59	210	269

Douleur avant et après une traitement antidouleur (appariée)

Appariées	Pain After	No Pain After	Total
Pain Before	21 (a)	68 (b)	89
No Pain Before	38 (c)	142 (d)	180
Total	59	210	269

```
> df <- cbind(c(21,38),c(68,142))
> colnames(df) <- c("A+", "A-")
> row.names(df) <- c("B+", "B-")
```

Applique test de χ^2 d'indépendance:

```
> chisq.test(df, correct=F) # χ²obs = 0.21 => p=0.64
```

Conclusion: Douleur et couleur de cheveux sont indépendant.

Applique test de McNemar:

```
> mcnemar.test(df, correct=F)
```

$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi^2$ avec 1 ddl

$\chi^2_{obs} = \frac{(68-38)^2}{68+38} = 8.49 > \chi^2_{\alpha}$ =3.84; Rejet H_0 =>

Conclusion: Douleur avant et après traitement ne sont pas identique/liée -> traitement réduit le douleur. (note; les mêmes individués sont suivie avant et après traitement).

Hypothèse:
H0: $\pi_{B+} = \pi_{A+}$ (homogénéité marginal)
H1: $\pi_{B+} \neq \pi_{A+}$

Conditions:
Dépendance sous H_0 (données appariées)
2x2 problème (généraliser avec Cochran–Mantel–Haenszel statistique stratifie)

Note: Le test de McNemar est identique à comparer une proportion discordante « P(B+ ∩ A-|Discordante) » à une proportion théorique (0.5) en évaluant uniquement les combinaisons discordantes (E+ ∩ I-) et (E- ∩ I+)

Test du χ^2 - Sommaire

	Nombre d'echantillons	Variable qualitative (V_i)	Indépendance sous H_0	Facteurs pour Variable Qualitative	H_0
Conformité	1	1	NA	2+	Distribution de la variable qualitative conforme à la distribution théorique
Homogénéité	2+	1	NA	2+	Variable qualitative distribuée de manière homogène à travers des échantillons
Indépendance	1	2	OUI (Non-appariée)	2+	Variable qualitative sont indépendant
McNemar	1	2	NON (Appariée)	2	$P(V_1)=P(V_2)$ (probabilité marginale égale)

Fisher's exact test fournira une valeur p pour les problèmes basés sur un tableau de contingence 2x2. Surpasse la sensitivité de χ^2 lorsque N est petit. N'est pas une analyse de proportion mais une analyse combinatorique.