Formation Logiciel R

Institut Pasteur
de Nouvelle-Calédonie

1

5° séance

STATISTIQUES AVEC LE LOGICIEL R

2

TEST DE
χ^2

Domaines d'application

- **Comparer des distributions**
  - **variables qualitatives nominales, ordinales, binaires,**
  - **quantitatives discrètes**
- **χ^2 de conformité**
  - **comparaison d'une distribution observée à une distribution dans une population**
- **χ^2 d'homogénéité**
  - **Deux ou plusieurs distributions observées**
- **χ^2 d'indépendance**
  - **Liaison entre les distributions de deux variables d'un même échantillon**

Principe du test de χ^2 (1)

- **Comparer les effectifs des classes**
- **Sous H_0 : l'échantillon observé provient de la population (χ^2 conformité)**
- **Sous H_0 : les échantillons étudiés proviennent de la même population (χ^2 d'homogénéité)**

5

Conditions d'application

- Les effectifs théoriques calculés dans chaque case doivent être supérieurs ou égaux à 5.

6

Test χ^2 pour comparer une fréquence observée à une fréquence théorique

χ^2 de conformité ou d'ajustement

7

Variable **qualitative nominale**

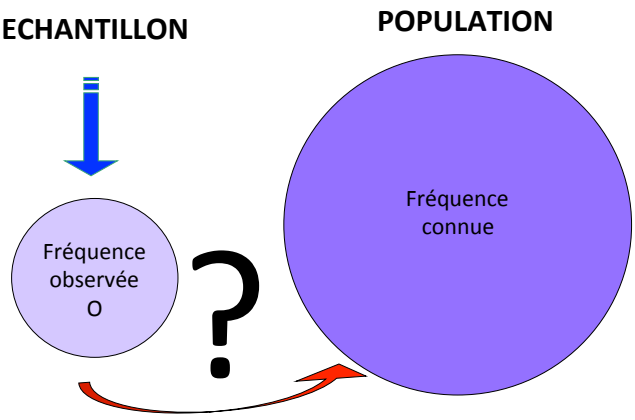
Paramètre étudié : **effectifs observés et attendus**

H_0 **: distribution échantillon = distribution théorique**

H1 bilatérale **: dist. échant ≠ dist.théoriq**

Condition d'application :

Effectifs attendus sup ou égaux à 5



8

Calcul du test de χ^2

	Echantillon	Popul	Théo
A1	O_1	F1	N*F1
A2	O_2	F2	N*F2
….	….	…	
Ai	Oi	Fi	N*Fi
Total	N	100%	

Les effectifs théoriques sont les effectifs attendus:

Ci = N x Fi

$$\chi^2 = \Sigma [(o_{ij} - c_{ij})^2 / c_{ij}]$$

ddl = r - 1

9

Exemple

	Echantillon	%Pop =f	284*F
0-19	73	24,6	69,9
20-39	82	28,1	79,8
40-59	75	26,0	73,8
60-74	36	13,6	38,6
>74	18	7,7	21,9
Total	284	100%	

Les effectifs théoriques sont les effectifs attendus:

Ci = N x Fi

$$\chi^2 = (73-69,9)^2/69,9 + \ldots+(18-21,9)^2/21,9 = 1.09$$

ddl = 5 – 1=4

10

Test du χ^2 de conformité

binom.test(n1,n,pt)

binom.test(23,100,0.4)

 # la comparaison d'une proportion observée à une valeur théorique

n1= nombre de cas observé

n=effectif total

pt= proportion théorique

Situation unilatérale

>binom.test(n1,n, pt, alternative = "greater")

>binom.test(n1,n ,pt, alternative = "less")

11

Test de χ^2 pour comparer des distributions observées entre plusieurs échantillons

χ^2 d' homogénéité

12

Variables **qualitatives nominales ou binaires**

Paramètres étudiés : **effectifs des classes des échantillons**

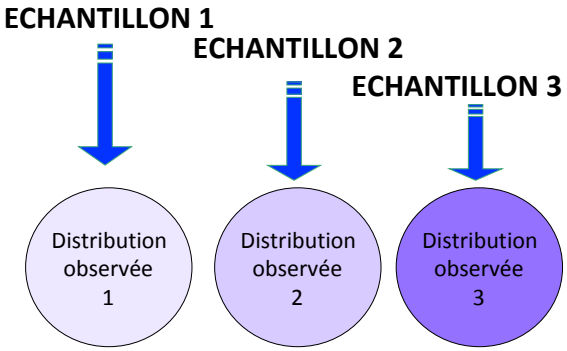
H_0 : $\%_1 = \%_2 = \%_3$

H1 bilatérale : $\%_1 \neq \%_2 \neq \%_3$

Conditions d'application :

Effectifs théoriques sup ou égaux à 5

Echantillons indépendants



$$\chi^2 = \Sigma \left[(o_{ij} - c_{ij})^2 / c_{ij} \right]$$

ddl = (r -1) * (k -1)

13

Tableau de contingence

Echantillons

Variable	E1	E2	…	Ei	Total
A1	O11 C11	O12 C12		O1j C1j	t1
…					
Ai	Oi1 Ci1			Oij Cij	ti
Total	n1	n2	….	nj	N

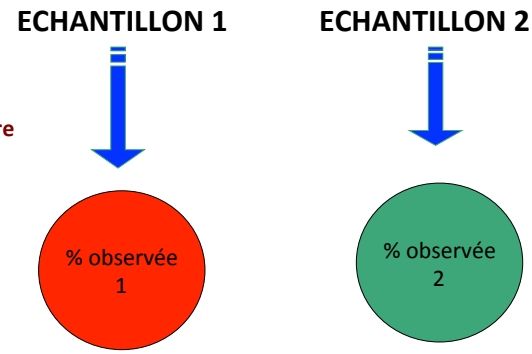
$$\chi^2 = \Sigma \left[(o_{ij} - c_{ij})^2 / c_{ij} \right]$$

ddl = (r -1) * (k -1)

14

Test de χ^2 pour comparer deux pourcentages

15



Variable **qualitative binaire**

Paramètres étudiés : **effectifs des classes des échantillons**

H_0 : $P_1 = P_2$

H1 bilatérale : $P_1 \neq P_2$

H1 unilatérale :

P1 > P2 ou P1 < P2

Condition d'application :

Effectifs théoriques sup ou égaux à 5

Echantillons indépendants

$$\chi^2 = \Sigma \left[(o_{ij} - c_{ij})^2 / c_{ij} \right]$$

$$ddl = (r - 1) * (k - 1)$$

Si effectif théorique <5 : test exact de Fisher ou Yates

16

Tableau de contingence

Echantillons

Variable	E1	E2	Total
présent	a	b	t_1
absent	c	d	t_2
Total	n_1	n_2	N
%	a/n_1	b/n_2	

$$\chi^2 = N *(ad - bc)^2 / n_1*n_2*t_1*t_2$$

ddl = 1

α / ddl	0,0001	0,001	0,01	0,02	0,03	0,04	0,05	0,10	0,20	0,30	0,50	0,90
1	15,13	10,83	6,63	5,41	4,71	4,22	3,84	2,71	1,64	1,07	0,45	0,02
2	18,42	13,82	9,21	7,82	7,01	6,44	5,99	4,61	3,22	2,41	1,39	0,21
3	21,10	16,27	11,34	9,84	8,95	8,31	7,81	6,25	4,64	3,66	2,37	0,58

17

TESTS DE LIAISON

χ^2 d' indépendance

χ^2 de tendance

18

χ^2 d'Indépendance

- **Liaison entre 2 variables qualitatives**
  (ex : couleur des yeux en fonction de la couleur des cheveux)
- **Sous H0 :**
  Distribution de A identique quelque soit la classe de B

Variables	B1	B2	B3	Total
A1	10	10	10	30
A2	20	20	20	60
A3	30	30	30	90
Total	60	60	60	180

19

χ^2 d'Indépendance

- **Liaison entre 2 variables qualitatives**
  (ex : couleur des yeux en fonction de la couleur des cheveux)
- **Sous H1 :**
  Distribution de A varie en fonction de la classe de B
- **Conditions d'application:**
  Effectifs théoriques ≥ 5

Variables	B1	B2	B3	Total
A1	10	10	30	50
A2	20	40	20	80
A3	30	10	10	50
Total	60	60	60	180

$$\chi^2 = \Sigma \left[(o_{ij} - c_{ij})^2 / c_{ij} \right]$$

$$ddl = (r - 1) * (k - 1)$$

Même test mais interprétation différente

20

χ^2 de Tendance

- **Liaison entre variable qualitative binaire et une variable qualitative de type ordinale**
  - (exemple: répartition par tranche d'âge en fonction du sexe)
- **Sous H0:**
  - % identique quelque soit la tranche d'âge

Variables	M	F	Total
0 - 9	5	5	10
10-19	15	15	30
20-29	25	25	50
Total	45	45	90

21

χ^2 de Tendance

- **Liaison entre variable qualitative binaire et une variable qualitative de type ordinale**
  - (exemple: répartition par tranche d'âge en fonction du sexe)
- **Sous H1:**
  - Distribution de l'âge est dépendante du sexe.
  - La proportion de filles diminue avec l'âge
- **Conditions d'application:**
  Effectifs théoriques ≥ 5
  Liaison supposée linéaire

Variables	M	F	Total
0 - 9	5	25	30
10-19	15	15	30
20-29	25	5	30
Total	45	45	90

22

Tableau de contingence

Variable B ordinale

Variable A	x1	x2	…	xi	Total
+	O11 C11	O12 C12		O1j C1j	t1
-	O21 C21			O2j C2j	t2
Total	n1	n2	….	nj	N

$$\chi^2 = N^3 * [\Sigma x_i (o_{ij} - c_{ij})]^2 / t_{1*} t_{2*} [N * \Sigma (nixi^2 - (\Sigma nixi)^2)]$$

ddl = 1

23

Test du χ^2

Table de contingence

```
a<-table(Mydata$SEXE,Mydata$NIVSCOL)
margin.table(a,1)
b<-
round(prop.table(table(Mydata$SEXE,Mydata$NIVSCOL),
1)*100,digit=1)
Conting<-rbind(a[1,],b[1,],a[2,],b[2,])
dimnames(Conting)<-
list(c("Male","%","Female","%"),c("NonSc","Prim","Secon
d","Universit"))
Conting
```

24

Test du χ^2

Conditions de validité?

chisq.test(**a**)$expected

chisq.test(a, correct=F)

Utilisable également lorsque x est sous la forme d'un tableau de contingence.

L'argument correct peut prendre les valeurs "T" ou "F" selon que l'on utilise la correction de Yates ou non.

25

Test du χ^2

Table de contingence

```
a.1<-table(Mydata$SEXE,Mydata$PALU)
b.1<-round(prop.table(a.1,1)*100,digit=1)
Conting.1<-rbind(a.1[1,],b.1[1,],a.1[2,],b.1[2,])
dimnames(Conting.1)<-
list(c("Male","%","Female","%"),c("Palu
Neg","Palu +"))
Conting.1
```

26

Test du χ^2

Conditions de validité?

chisq.test(**a.1**)$expected

chisq.test(a.1, correct=F)

L'argument correct peut prendre les valeurs "T" ou "F" selon que l'on utilise la correction de Yates ou non.

27

Test de χ^2 corrigé de Yates

- $\chi^2 = \Sigma [(|oij – cij| - 0,5)^2 / cij]$

  ➔ **ddl = 1**

- **Si effectif théorique <5 : test exact de Fisher ou Yates**

28

Test de χ^2 corrigé de Yates

>chisq.test(Mydata$SEXE,Mydata$NIVSCOL, correct=T) # réalise le test du Chi-deux de comparaison de x et y.

>fisher.test(table(Mydata$SEXE,Mydata$NIVSCOL))

Utilisable également lorsque x est sous la forme d'un tableau de contingence.

29

Test exact de Fisher

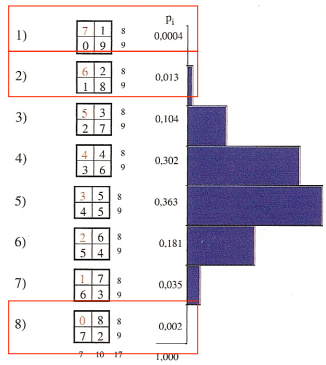
Pour les tableaux de contingence à 4 cases, si effectifs théoriques < 5.

Calculer la probabilité d'avoir observé une configuration donnant un écart au moins aussi grand que l'écart observé entre les 2 pourcentages que l'on compare

pi = n1!n2!t1!t2! / a!b!c!d!N! ➔ p = Σ pi

30

	Echantillons		
Variable	E1	E2	Total
présent	6	2	8
absent	1	8	9
Total	7	10	17
p	6/7 85,7%	2/10 20,0%	



p1 – p2 = |85,7 – 20,0| = 65,7

P = 0.0004+0.013+0.002=0.0154

31

Test exact de Fischer

>fisher.test(Mydata$SEXE,Mydata$NIVSCOL) # réalise le test du Chi-deux de comparaison de x et y.

>fisher.test(table(Mydata$SEXE,Mydata$NIVSCOL))

Utilisable également lorsque x est sous la forme d'un tableau de contingence.

32

Test pour séries appariées

>mcnemar.test() # réalise le test du Chi-deux de séries appariées

Utilisable lorsque x est sous la forme d'un tableau de contingence.

33

	Malade	Non Malade	Total
Malade	794	150	944
Non Malade	86	570	656
Total	880	720	1600

34

ANOVA:
Analyse de Variance

Unité d'épidémiologie
IPD

35

Test F
de
Fischer-Snedecor

Comparer deux variances

36

Domaines du test

➔ Vérification des conditions d'application d'égalité des variances dans un test T de Student

➔ Comparaison de moyennes (Anova)

37

Principe du test

Comparaison des variances par leur rapport
H0: les deux variances sont égales
➔ leur rapport = 1
H1: les deux variances sont différentes (bilatérale)

Calcul du rapport:

$$F_0 = \frac{s_1^2}{s_2^2}$$

38

Condition d'application du test

Distributions supposées normales dans les deux populations d'où proviennent les deux échantillons.

39

ANOVA

Deux variables: 1 Catégorie, 1 Quantitative

La Question: La moyenne de la variable quantitative dépend-elle du groupe de la variable catégorielle?

NB : Si la variable catégorielle n'a que deux valeurs
➔ test T ou Z (comparaison de deux moyennes)

ANOVA est utilisé pour 3 groupes ou plus

40

La difference entre les groupes dépend de:

- la difference des moyennes
- de l'intervalle de confiance
- de la taille de l'échantillon

ANOVA determine P-value à partir du test F de Fischer Snedecor

41

ANOVA

H_0: les moyennes des différents groupes sont égales.

H_a: Toutes les moyennes ne sont pas égales

- Ce qui ne veut pas dire comment elles varients ou laquelle est différente.

42

Conditions d'application

- Les distributions sont supposées normales dans les différents groupes
- Les variances de chaque groupe sont égales

43

Contrôle de la Normalité

- par hypothèse sur la population
- histogrammes pour chaque groupe
- courbe normale pour chaque groupe

Le plus souvent on fait l'hypothèse que les mesures dans la population tendent vers une distribution normale

44

Contrôle des variances

Variable	traitement	N	Moyenne	Mediane	EcT
jours	A	8	7.250	7.000	1.669
	B	8	8.875	9.000	1.458
	P	9	10.111	10.000	1.764

On compare le plus petit et le plus grand EcT:

- + grand: 1.764
- + petit : 1.458
- 1.458 x 2 = 2.916 > 1.764

Test de Bartlett – Test de Levene

45

Principe de l'Anova

ANOVA s'intéresse à deux sources de variation dans les données et compare leur taille.

- variation ENTRE les groupes
  écart moyen entre chaque moyenne et la moyenne générale

$$(\overline{x}_i - \overline{x})^2$$

- variation DANS les groupes
  la variation moyenne des individus à l'intérieur des groupes.

$$(x_{ij} - \overline{x}_i)^2$$

46

En pratique

Comparer graphiquement les deux sous populations

```
Mydata$ASTH <-factor(Mydata$ASTH)
 Mydata$VALEURF <-as.numeric(Mydata$VALEURF)
boxplot(VALEURF~ASTH, ylab= "FIEVRE", xlab= "Asthenie",data=Mydata)
```

Présenter la valeur moyenne par sous population

```
by(Mydata$VALEURF, Mydata$ASTH, summary)
```

Réaliser un test de normalité des données dans chacun des groupes

```
Select.0<-Mydata[Mydata$ASTH== 0,]

shapiro.test(Select.0$VALEURF)
 …
Select.3<-Mydata[Mydata$ASTH==3,]

shapiro.test(Select.3$VALEURF")
```

47

En pratique

Réaliser ensuite un test d'égalité des variance

bartlett.test(Mydata$VALEURF~Mydata$ASTH)

Réaliser ensuite l'analyse

- **Regaov<-lm(VALEURF~ASTH,data=Mydata)**
- **anova(Regaov)**

Response: VALEURF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ASTH	1	9.695	9.6947	8.602	0.00376 **
Residuals	195	219.771	1.1270		

48

Test de Bartlett

>bartlett.test(Mydata$VALEURF~Mydata$PALU)

Comparaison des variances pour savoir si les conditions d'utilisation des tests sont vérifiées

49

Test t de Student

>t.test(x, y, alternative="two.sided", paired="F", var.equal="T") : #réalise le test t de Student de comparaison de **x et y.**

- L'argument alternative peut prendre les valeurs "two.sided", "greater" ou "less" selon que le test est réalisé en situation bilatérale, unilatéral de supériorité ou unilatéral d'infériorité ;
- L'argument paired peut prendre les valeurs logiques "F" ou "T" selon que le test est non apparié ou apparié ;
- L'argument var.equal peut prendre les valeurs logiques "F" ou "T" selon que l'hypothèse d'égalité des variances n'est pas vérifiée ou qu'elle est vérifiée.

50

En pratique

Comparer graphiquement les deux sous populations

```
boxplot(VALEURF~PALU, ylab= "FIEVRE", xlab=  "Paludisme", data=Mydata)
```

Quel est la valeur moyenne par sous population

```
by(Mydata$VALEURF, Mydata$PALU, summary)
```

Réaliser un test de normalité des données dans chacun des groupes

```
Select.0<-Mydata[,"PALU"]== 0
shapiro.test(Mydata[Select.0,"VALEURF”])
Select.1<-Mydata[,"PALU"]==1
shapiro.test(Mydata[Select.1,”VALEURF"])
```

51

Réaliser ensuite un test d'égalité des variances

```
var.test(VALEURF~PALU,conf.level=.95,
data=Mydata)
```

Comparaison de 2 moyennes

```
>t.test(Mydata$VALEURF~Mydata$PALU,var.equal=F)
>wilcox.test(Mydata$VALEURF~Mydata$PALU)
```

52

Intervalle de confiance d'une moyenne

Intervalle de confiance d'une moyenne

```
t.test(Mydata$VALEURF,conf.level=0.05)
```

Intervalle de confiance par groupe

```
Pgroup<-split(Mydata$VALEURF,Mydata$PALU)
Pgroup
t.test(Pgroup$ "1",conf.level=0.05)
t.test(Pgroup$ "2",conf.level=0.05)
```

53

Test de Kruskall wallis

```
kruskal.test(Mydata$VALEURF~Mydata$PALU)
```

Test non paramétrique de comparaison des moyennes ➔ comparaison de médiane

54

TEST DU COEFFICIENT DE CORRELATION

- Variables **quantitatives**
- Paramètre étudié : **coefficient de corrélation**
- Séries comparées **: séries appariées**
- H0 **: absence de liaison entre X et Y : ρ = 0**
- H1 bilatérale **: liaison entre X et Y : ρ ≠ 0**
- H1 unilatérale **:**
- **Liaison positive ρ > 0**
- **Liaison négative ρ < 0**

55

TEST DU COEFFICIENT DE CORRELATION

- Conditions d'application :
  - **Les variables X et Y sont aléatoires**
  - **L'association entre X et Y est linéaire**
  - **Les observations de chaque variable sont indépendantes**
  - **Les distributions de Y liées à chaque valeur de X sont normales et de variance constante et vice versa.**

56

Quelques notions importantes

Attention, une corrélation n'est pas forcément une relation de cause à effet !

Les corrélations mesurent la relation linéaire entre 2 variables

Un coefficient de corrélation doit être représentatif de la relation existant entre 2 variables

Il faut donc visualiser cette relation sur un graphique avant tout calcul de corrélation

57

Corrélation

```
>cor.test(x, y, alternative="two.sided", paired=F,
  method=c("pearson")) #réalise le test de corrélation
  entre x et y.
```

- L'argument alternative peut prendre les valeurs "two.sided", "greater" ou "less" selon que le test est réalisé en situation bilatérale, unilatéral de supériorité ou unilatéral d'infériorité ;
- L'argument method peut prendre les valeurs " pearson ", " kendall " ou " spearman " selon la méthode choisie pour estimer les coefficients de corrélation
- Le coefficient r de Pearson nécessite que les variables soient distribuées normalement

58

Pearson's product-moment correlation

data: Mydata$DMAL and Mydata$VALEURF
t = -0.9074, df = 198, p-value = 0.3653
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.20129417 0.07506084
sample estimates:
       cor
-0.06435032

59

Mydata$DMAL <-as.numeric(Mydata$DMAL)
On va comparer la durée de maladie en fonction de la valeur de la fièvre
Représenter le nuage de points

```
plot(DMAL~VALEURF,data=Mydata,pch=15,col="red",cex=.5)
```

Réaliser la régression

```
Reglin<-lm(DMAL~VALEURF,data=Mydata)
anova(Reglin)
summary(Reglin)
abline(Reglin)
cor.test(Mydata$DMAL,Mydata$VALEURF,
alternative="two.sided",method="spearman",conf.level=0.95)
```

60

- Call:
- lm(formula = DMAL ~ VALEURF, data = Mydata)

- Residuals:
- Min 1Q Median 3Q Max
- -2.4755 -1.3729 -0.1675 1.5681 2.8838

- Coefficients:
- Estimate Std. Error t value Pr(>|t|)
- (Intercept) 11.2744 4.3653 2.583 0.0105 *
- VALEURF -0.1027 0.1132 -0.907 0.3653
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Residual standard error: 1.716 on 198 degrees of freedom
- Multiple R-squared: 0.004141, Adjusted R-squared: -0.0008886
- F-statistic: 0.8233 on 1 and 198 DF, p-value: 0.3653

61