

Tests d'hypothèses

Tewfik Mahdjoub

Université Abou Bekr Belkaïd, Tlemcen

Novembre 2024

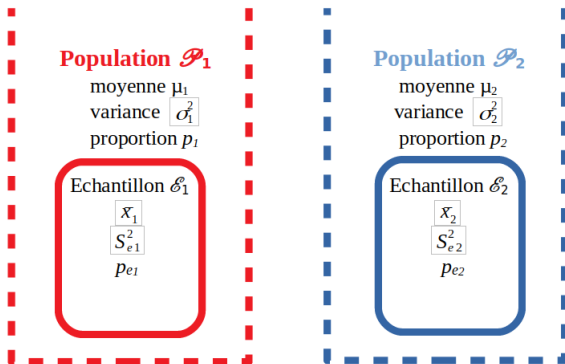


Plan du cours

- 1 Test d'homogénéité de deux variances
- 2 Test d'homogénéité de deux moyennes
 - Variances des populations égales et inconnues
 - Variance des populations différentes et inconnues
 - Variance des populations connues
 - Cas des données appariées
- 3 Inférence sur deux proportions
 - Cas des échantillons grands
 - Cas des échantillons petits
- 4 Test d'adéquation
- 5 Test d'indépendance



Rappel :



Les échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ sont indépendants si les paramètres statistiques de $\mathcal{P}_1$ ne sont pas liés à ceux de $\mathcal{P}_2$.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

On supposera, pour simplifier les calculs, que $s_{e1}^2 \geq s_{e2}^2$



Conditions d'application du test :

- la variable statistique est quantitative
- les 2 échantillons sont aléatoires simples
- les 2 échantillons sont indépendants
- les populations sont normalement distribuées

Remarque : condition (4) $\Rightarrow$ **test paramétrique**.

La statistique du test est :

$$S = \frac{s_{e1}^2}{s_{e2}^2} \quad (1)$$

S suit une loi de Fisher $df(., \nu_1, \nu_2)$ de degré de liberté

$$\nu_1 = n_1 - 1, \nu_2 = n_2 - 1$$

n_1, n_2 tailles respectives des échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$



Exercice : le dosage des transaminases informe sur le fonctionnement hépatique. Lors de la mise au point d'un traitement qui peut perturber la fonction hépatique, deux variantes du protocole thérapeutique sont appliquées à deux groupes indépendants de malades. On suit la sensibilité au traitement par dosage des transaminases dont les valeurs sont reportées dans le tableau suivant :

G1	22	18	25	14	16	22	17	19	30	17	23
G2	31	35	32	30	28	26	27	19	25	20	18
	31	27	29	24							

La variable "dosage des transaminases" est supposée suivre une loi normale.

1. Tester, au seuil $\alpha = 5\%$, l'hypothèse selon laquelle la variance de la population du groupe 1 est égale à celle de la population du groupe 2.
2. Tester, au seuil $\alpha = 5\%$, l'hypothèse selon laquelle la moyenne de la population du groupe 1 est égale à celle de la population du groupe 2.



```
> transG1 = c(22, 18, 25, 14, 16, 22, 17, 19, 30, 17, 23)
> transG2 = c(31, 35, 32, 30, 28, 26, 27, 19, 25, 20, 18, 31, 27, 29, 24)
> var.test(transG1, transG2, alternative = "two.sided", conf.level = 0.95)
```

F test to compare two variances	nature du test
data : transG1 and transG2	fichier de données
F = 0.88392 , num df = 10, denom df = 14, p-value = 0.8629	F :valeur de la statistique, num df :ddl du num, denom df :ddl du denom p-value :probabilité de rejet
alternative hypothesis : true ratio of variances is not equal to 1	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
0.2808897 3.1382817	
sample estimates :	estimations sur l'échantillon
ratio of variances 0.883921	moyenne des échantillons

$F = 0.88392 < 1 \Rightarrow$ refaire le test en plaçant transG2 avant transG1



```
> var.test(transG2, transG1, alternative = "two.sided", conf.level = 0.95)
```

F test to compare two variances	nature du test
data : transG2 and transG1	fichier de données
F = 1.1313, num df = 14, denom df = 10, p-value = 0.8629	F :valeur de la statistique, num df :ddl du num, denom df :ddl du denom p-value :probabilité de rejet
alternative hypothesis : true ratio of variances is not equal to 1	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
0.3186457 3.5601159	
sample estimates :	estimations sur l'échantillon
ratio of variances 1.131323	moyenne des échantillons

p-value = 0.8629 $> \alpha = 0.05 \Rightarrow$ **acceptation de H_0**

Les variances des 2 populations sont considérées égales.



$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Deux cas sont considérés : variances égales et inconnues et variances différentes



Conditions d'application du test :

- la variable statistique est quantitative
- les 2 échantillons sont aléatoires simples
- les 2 échantillons sont indépendants
- les 2 populations ont le même écart-type $\sigma_1 = \sigma_2$
- les populations sont normalement distribuées ou les tailles des échantillons sont supérieures à 30

La statistique du test est :

$$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}} \quad (2)$$

où :

$$s_c^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad (3)$$

S suit une loi de Student $t(., \nu)$ avec $\nu = (n_1 - 1) + (n_2 - 1)$



```
> t.test(transG1, transG2, var.equal = T, alternative = "two.sided",
conf.level = 0.95)
```

Two Sample t-test	nature du test
data : transG1 and transG2	fichier de données
t = -3.4084, df = 24, p-value = 0.002309	t : valeur de la statistique, df :ddl, p-value : probabilité de rejet
alternative hypothesis : true difference in means is not equal to 0	hypothèse alternative
95 percent confidence interval : -10.479763 -2.574782	intervalle de confiance à 95%
sample estimates : mean of x mean of y	estimations sur l'échantillon moyenne des échantillons
20.27273 26.80000	

p-value = 0.002309 < α = 0.05 $\Rightarrow$ rejet de H_0

Les 2 variantes du traitement sont différentes.



L'égalité des variances inconnues est testée par `var.test()`. Dans le cas où elles sont différentes, on calcule alors la statistique suivante, sachant que les conditions d'application :

- la variable statistique est quantitative
- les 2 échantillons sont aléatoires simples
- les 2 échantillons sont indépendants
- les populations sont normalement distribuées ou les tailles des échantillons sont supérieures à 30

sont satisfaites :

$$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (4)$$

S suit une loi de Student à $ddl = \min(n_1 - 1, n_2 - 1)$.



Sur les données transG1 et transG2 on a vérifié que les variances sont égales. Le but de cet exemple est juste pour voir comment fonctionne `var.equal = F`.

```
> t.test(transG1, transG2, var.equal = F, alternative = "two.sided",
conf.level = 0.95)
```

Welch Two Sample t-test	nature du test
data : transG1 and transG2	fichier de données
t = -3.4422, df = 22.442, p-value = 0.002277	t : valeur de la statistique, df :ddl, p-value : probabilité de rejet
alternative hypothesis : true difference in means is not equal to 0	hypothèse alternative
95 percent confidence interval : -10.455367 -2.599178	intervalle de confiance à 95%
sample estimates :	estimations sur l'échantillon
mean of x mean of y 20.27273 26.80000	moyenne des échantillons



Dans le test de Welch, le ddl se calcul par :

$$ddl = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}} \quad (5)$$

C'est pour cela : $df = 22.442$



Conditions d'application du test :

- la variable statistique est quantitative
- les 2 échantillons sont aléatoires simples
- les 2 échantillons sont indépendants
- les populations sont normalement distribuées ou les tailles des échantillons sont supérieures à 30

sont satisfaites :

La statistique S du test est :

$$S = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (6)$$

S suit une loi normale $\text{dnorm}(\cdot, 0, 1)$.

Exercice : écrire, sous  , le script correspondant à ce test.



données appariées : 2 séries de mesures sur le même échantillon.

Exemples :

1. efficacité d'un traitement : mesure avant puis après,
2. efficacité d'un régime pauvre en matière grasse : poids avant et poids après,
3. double correction des même copies.

Conditions d'application du test :

- la variable statistique est quantitative
- les données sont appariées
- l'échantillon est aléatoire
- les populations sont normalement distribuées ou le nombre des paires est grand



La statistique S du test est :

$$S = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \quad (7)$$

S suit une loi de Student $t(., n-1)$.

avec :

- $\bar{d}$: moyenne des différences
- s_d : écart-type des différences
- n : nombre de paires



Exercice :

La double correction d'un examen a donné les résultats suivants :

N° étudiant	1	2	3	4	5	6	7	8	9	10
Correcteur 1	10	14	8	15	7	5	17	14	11	8
Correcteur 2	10	12	9	15	6	5	15	10	12	8

Peut-on conclure, au seuil de confiance de 95%, à une différence de notation entre les 2 correcteurs ?



```
> correcteur1 = c(10, 14, 8, 15, 7, 5, 17, 14, 11, 8)
> correcteur2 = c(10, 12, 9, 15, 6, 5, 15, 10, 12, 8)
> t.test(correcteur1, correcteur2, paired = T, alternative = "two.sided",
conf.level = 0.95)
```

Paired t-test	nature du test
data : correcteur1 and correcteur2	fichier de données
t = 1.4126, df = 9, p-value = 0.1914	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true difference in means is not equal to 0	hypothèse alternative
95 percent confidence interval : :	intervalle de confiance à 95%
-0.4209795 1.8209795	
sample estimates :	estimations sur l'échantillon
mean of the differences 0.7	moyenne de la différences

p-value = 0.1914 $>$ α = 0.05 $\Rightarrow$ acceptation de H_0

Il n'y a pas de différence entre les 2 correcteurs.



Avec le test de Welch :

> t.test(correcteur1, correcteur2, alternative = "two.sided", conf.level = 0.95) il vient :

Welch Two Sample t-test	nature du test
data : correcteur1 and correcteur2	fichier de données
t = 0.4247, df = 17.591, p-value = 0.6762	t :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : true difference in means is not equal to 0	hypothèse alternative
95 percent confidence interval : -2.768591 4.168591	intervalle de confiance à 95%
sample estimates : mean of x mean of y	estimations sur l'échantillon moyenne des échantillons
10.9 10.2	

p-value (appariées) = 0.1914 < p-value (Welch) = 0.6762



$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Deux cas sont considérés : grands échantillons et petits échantillons



Conditions :

- les proportions sont issues de 2 échantillons aléatoires simples
- $n_1 p_1 > 5$, $n_1(1 - p_1) > 5$, $n_2 p_2 > 5$ et $n_2(1 - p_2) > 5$

avec n_1 , n_2 tailles de $\mathcal{E}_1$ et $\mathcal{E}_2$ respectivement.

La statistique du test est :

$$S = \frac{p_{e1} - p_{e2}}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad (8)$$

avec :

$$\bar{p} = \frac{n_1 p_{e1} + n_2 p_{e2}}{n_1 + n_2} \quad (9)$$

S suit une loi normale $d_{\text{norm}}(0,1)$

Nécessité d'écrire un script, sous , pour ce test.



Comparer 2 proportions : cas particulier du test de comparaison de plusieurs proportions basé sur le calcul du χ^2 (test valable même pour les échantillons grands).

On veut tester l'impact du temps de préparation sur la réussite d'un examen :

•	Groupe 1 (10h00)	Groupe 2 (20h00)
Nombre d'étudiants	180	150
Nombre d'étudiants ayant réussi l'examen	126	129



```
> prop.test(c(126,129), c(180,150), alternative = "two.sided", conf.level = 0.95)
```

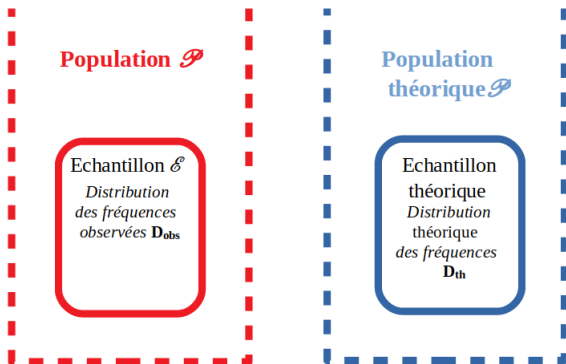
2-sample test for equality of proportions with continuity correction	nature du test
data : c(126, 129) out of c(180, 150)	fichier de données
X-squared = 11.033, df = 1, p-value = 0.0008951	X-squared :valeur de la statistique, df :ddl, p-value :probabilité de rejet
alternative hypothesis : two.sided	hypothèse alternative
95 percent confidence interval :	intervalle de confiance à 95%
-0.25308877 -0.06691123	
sample estimates :	estimations sur l'échantillon
prop 1 prop 2	moyenne des échantillons
0.70 0.86	

p-value = 0.0008951 < $\alpha = 0.05 \Rightarrow$ rejet de H_0

Le temps de préparation impacte sur la réussite de l'examen.



Rappel :



$$H_0 : D_{Obs} = D_{The}$$

$$H_1 : Diff^2 > 0$$



Il s'agit de tester l'hypothèse que les fréquences observées pour les différentes catégories sont en adéquation avec une distribution donnée. Un test d'adéquation est toujours unilatéral à droite.

Conditions d'application :

- données sélectionnées aléatoirement
- pour chaque catégorie, la fréquence attendue est au moins 5

La statistique du test est :

$$S = \chi^2 = \sum_{i=1}^k \frac{(n_i^o - n_i^t)^2}{n_i^t} \quad (10)$$

avec :

- n_i^o : fréquences observées
- n_i^t : fréquences théoriques
- k : nombre de classes



S suit une loi du Khi-deux $\text{chisq}(., \nu_1)$ avec $\nu_1 = ddl = k - 1$

Exemple : un chercheur a développé un modèle théorique pour prédire la couleur des yeux. Après avoir examiné un échantillon aléatoire de parents, il prédit la couleur des yeux du premier enfant. Le tableau suivant donne la couleur des yeux des enfants. Selon sa théorie, il prédit que 87% des enfants devraient avoir les yeux marrons, 8% devraient avoir les yeux bleus et 5% devraient avoir les yeux verts. Tester l'affirmation que les fréquences observées correspondent aux fréquences théoriques, sachant que $\alpha = 5\%$.

•	Yeux marrons	Yeux bleus	Yeux verts
Fréquences	132	17	0



```
> chisq.test(c(132,17,0), p=c(0.87,0.08,0.05))
```

Chi-squared test for given probabilities

data : c(132, 17, 0)

X-squared = 9.6583, df = 2, p-value = 0.007993

$$\mathbf{p\text{-}value = 0.007993} < \alpha = \mathbf{0.05} \Rightarrow \text{rejet de } H_0$$

Les distributions sont différentes.



```
> analyse=chisq.test(c(132,17,0),p=c(0.87,0.08,0.05))
> summary(analyse)
```

•	Length	Class	Mode
statistic	1	-none-	numeric
parameter	1	-none-	numeric
p.value	1	-none-	numeric
method	1	-none-	character
data.name	1	-none-	character
observed	3	-none-	numeric
expected	3	-none -	numeric
residuals	3	-none-	numeric
stdres	3	-none-	numeric



On appelle tableau de contingence, un tableau dans lequel les fréquences correspondent à 2 variables statistiques qualitatives ou quantitatives, la 1ère en ligne et la 2ème en colonne.

Exemple :

Une enquête sur la qualité du régime alimentaire chez un échantillon composé de 22 étudiants, a permis de dresser le tableau suivant :



N°	Genre	Poids(kg)	Age (années)	Alimentation	Codage des variables
1	2	108.64	20	2	Genre :
2	2	51.94	21	2	1 : étudiante
3	2	72.85	18	2	2 : étudiant
4	2	52.07	19	2	Alimentation :
5	1	68.41	18	1	1 : riche en gras
6	2	106.88	21	2	2 : moyenne en gras
7	1	52.19	23	2	3 : pauvre en gras
8	2	93.08	22	1	
9	1	74.24	22	2	
10	2	70.88	19	3	
11	2	55.60	22	2	
12	1	69.22	21	2	
13	2	78.81	20	2	
14	2	96.43	22	3	
15	2	76.67	18	2	
16	1	85.98	20	1	
17	2	91.75	23	2	
18	1	77.39	19	1	
19	2	75.08	18	3	
20	2	100.39	21	2	
21	1	65.96	19	2	
22	1	66.47	22	1	



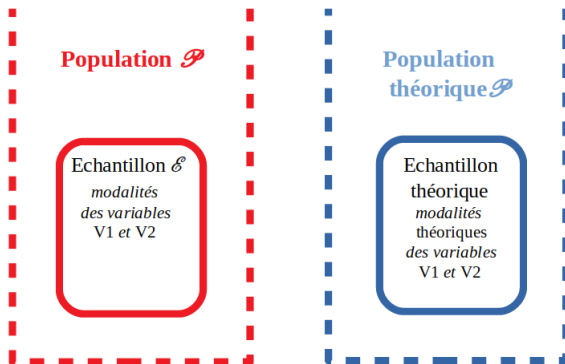
On construit les tableaux de contingence :

	Etudiantes	Etudiants
[51.94, 66.11[		
[66.11, 80.29[		
[80.29, 94.46[		
[94.46, 108.64]		

	18	19	20	21	22	23
[51.94, 66.11[						
[66.11, 80.29[						
[80.29, 94.46[						
[94.46, 108.64]						

	18	19	20	21	22	23
Etudiantes						
Etudiants						





H_0 : il n'y a pas d'association entre les 2 variables

H_1 : les 2 variables sont dépendantes



Conditions d'application du test :

- les données d'échantillon sont aléatoires
- pour chaque case du tableau, la fréquence attendue $n_{ij}^t \geq 5$

La statistique du test est :

$$S = \chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij}^o - n_{ij}^t)^2}{n_{ij}^t} \quad (11)$$

avec :

- n_{ij}^o : fréquences observées
- $n_{i.}$: somme des fréquences de la ligne i
- $n_{.j}$: somme des fréquences de la colonne j
- $n_{ij}^t = \frac{n_{i.} n_{.j}}{n}$: fréquences théoriques
- k et l : nombre de classes des 2 variables

S suit une loi du Khi-deux $\text{chisq}(., \nu_1)$ avec $\nu_1 = ddf = (k - 1)(l - 1)$

Un test d'indépendance est toujours unilatéral droit



```
> contingency=table(Age,Alimentation)
```

```
> contingency
```

```
Alimentation
```

```
Age 1 2 3
```

```
18 1 2 1
```

```
19 1 2 1
```

```
20 1 2 0
```

```
21 0 4 0
```

```
22 2 2 1
```

```
23 0 2 0
```

```
> chisq.test(contingence)
```

```
Pearson's Chi-squared test
```

```
data : contingence
```

```
X-squared = 6.2438, df = 10, p-value = 0.7944
```

```
Warning message :
```

```
In chisq.test(contingence) : Chi-squared approximation may be incorrect
```



```
> chisq.test(contingence, simulate.p.value = T)  
Pearson's Chi-squared test with simulated p-value (based on 2000  
replicates)  
data : contingence  
X-squared = 6.2438, df = NA, p-value = 0.8971
```

Au risque $\alpha = 5\%$, on peut dire que le type d'alimentation et l'âge sont indépendants.

