

11. Tests d'hypothèses (partie 2/2)

MTH2302D

S. Le Digabel, École Polytechnique de Montréal

A2017

(v2)

Plan

1. Introduction
2. Tests d'hypothèses avec 2 échantillons
3. Tests sur la normalité
4. Test d'ajustement du Khi-deux de Pearson
5. Test d'indépendance entre deux variables
6. Analyse de la variance et test de l'égalité de plusieurs moyennes

1. Introduction

2. Tests d'hypothèses avec 2 échantillons

3. Tests sur la normalité

4. Test d'ajustement du Khi-deux de Pearson

5. Test d'indépendance entre deux variables

6. Analyse de la variance et test de l'égalité de plusieurs moyennes

Introduction

Nous avons vu les tests paramétriques sur un échantillon. Nous allons finir les tests paramétriques avec le cas “2 échantillons”, puis nous allons voir certains tests non paramétriques :

- ▶ Tests sur la normalité (Shapiro-Wilk).
- ▶ Test d'ajustement du Khi-deux.
- ▶ Test d'indépendance entre deux variables (test du Khi-deux).
- ▶ Analyse de la variance (ANOVA).

1. Introduction

2. Tests d'hypothèses avec 2 échantillons

3. Tests sur la normalité

4. Test d'ajustement du Khi-deux de Pearson

5. Test d'indépendance entre deux variables

6. Analyse de la variance et test de l'égalité de plusieurs moyennes

Tests d'hypothèses avec 2 échantillons

Le [formulaire](#) sur le site du cours résume les tests d'hypothèses pour différentes situations, et en particulier pour deux échantillons.

Exemple 1

Une analyse statistique descriptive des données sur des mesures du diamètre (en cm) des pièces produites par deux chaînes de production A et B a donné les résultats suivants :

Chaîne	A	B
nb. observations	51	51
\bar{x}	1.375147	1.374982
s	0.000167	0.000172

Au seuil critique 1% peut-on affirmer que le diamètre des pièces de la chaîne A est supérieur en moyenne à celui des pièces de la chaîne B ?

1. Introduction
2. Tests d'hypothèses avec 2 échantillons
- 3. Tests sur la normalité**
4. Test d'ajustement du Khi-deux de Pearson
5. Test d'indépendance entre deux variables
6. Analyse de la variance et test de l'égalité de plusieurs moyennes

Tests sur la normalité

But : vérifier si les données d'un échantillon proviennent d'une population normale.

Méthode graphique

Soit $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ les données de l'échantillon rangées dans l'ordre croissant.

On compare ces données aux centiles d'une loi normale en traçant dans le plan les point $(X_{(j)}, z_j)$, où

$$z_j = F_Z^{-1} \left(\frac{j - 1/2}{n} \right) \text{ avec } j = 1, 2, \dots, n \text{ et } Z \sim N(\bar{X}, S^2).$$

Si les données proviennent d'une loi normale alors $X_{(j)} \simeq z_j$ et les points obtenus s'alignent approximativement sur une droite.

Le graphe ainsi obtenu est un *graphe de probabilité normal*.

Tests sur la normalité (suite)

Tests de normalité :

Il s'agit de tester

$$H_0 : X \sim \text{Normale.}$$

$$H_1 : X \text{ ne suit pas une loi normale.}$$

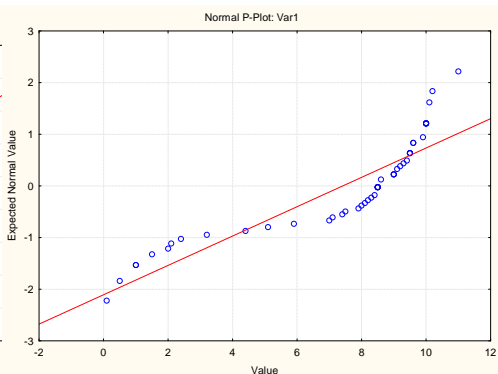
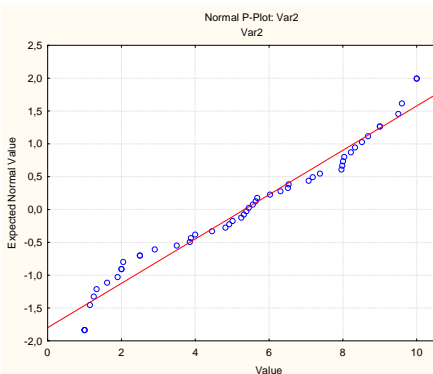
Il existe plusieurs types de tests qui sont généralement réalisés à l'aide d'un logiciel. Les logiciels calculent la statistique du test et la P -value correspondante. On distingue, entre autres : le test d'**Anderson-Darling**, le test d'**Agostino-Pearson**, le test de **Geary**, le test du **Khi-deux**, le test de **Kolmogorov-Smirnov** (statistique D) et le test de **Shapiro-Wilk** (statistique W). Nous nous intéressons en particulier au test de **Shapiro-Wilk**.

Tests sur la normalité : test de Shapiro-Wilk

Ce test calcule une statistique W représentant le carré d'un coefficient de corrélation entre les $X_{(j)}$ observés et les centiles théoriques $z_{(i)}$ d'une loi normale $N(0, 1)$.

- ▶ On peut montrer que $c \leq W \leq 1$, où $c \simeq 0.70$.
- ▶ En pratique, on rejette $H_0 : X \sim \text{Normale}$ lorsque la valeur de W est petite (proche de 0.70).

Exemple 2



Shapiro-Wilk : $W = 0.94294$

Shapiro-Wilk : $W = 0.79552$

Test de Shapiro-Wilk (suite)

Remarques

1. Le logiciel STATISTICA calcule w et surtout $P\text{-value} = P(W < w)$ qui est nécessaire pour la décision.
2. Lorsque l'hypothèse de normalité est acceptée (une grande $P\text{-value}$) il est important de confirmer l'hypothèse à l'aide des différents graphiques (quantile-quantile. etc.). Car, comme dans tout test statistique, l'acceptation de H_0 n'est pas une preuve que l'hypothèse soit vraie.

1. Introduction
2. Tests d'hypothèses avec 2 échantillons
3. Tests sur la normalité
- 4. Test d'ajustement du Khi-deux de Pearson**
5. Test d'indépendance entre deux variables
6. Analyse de la variance et test de l'égalité de plusieurs moyennes

Test d'ajustement du Khi-deux

- ▶ On cherche à vérifier si les données, x_1, \dots, x_n dont on dispose proviennent d'une population distribuée selon une loi particulière $F(x, \theta)$.
- ▶ À partir d'un échantillon aléatoire X_1, \dots, X_n de taille n d'une variable X , on va tester les hypothèses :

$$H_0 : X \sim F(x, \theta)$$

$$H_1 : X \neq F(x, \theta).$$

Test d'ajustement du Khi-deux : méthode

- ▶ On procède à un regroupement des observations selon k valeurs (ou intervalles). On obtient ainsi un tableau dont la forme générale est :

Valeurs (x_i)	V_1	V_2	...	V_i	...	Total
Effectifs observés (O_i)	O_1	O_2	...	O_i	...	n
Effectifs attendus (E_i)	E_1	E_2	...	E_i	...	n

Les O_i sont les effectifs observés, tandis que les E_i sont les effectifs attendus lorsque H_0 est vraie.

- ▶ Si on constate des E_i petits, regrouper des classes.
- ▶ On calcule les effectifs attendus $E_i = n \times p_i^{(0)}$ où

$$p_i^{(0)} = P(X \in V_i | H_0 \text{ est vraie}), i = 1, 2, \dots, k \text{ et } \sum_{i=1}^k p_i^{(0)} = 1.$$

Test d'ajustement du Khi-deux : méthode (suite)

- ▶ On calcule la statistique du test

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

- ▶ La statistique χ_0^2 représente une sorte de “distance” globale entre les effectifs observés et les effectifs attendus. Plus elle est grande moins l'hypothèse H_0 est plausible.
- ▶ Lorsque H_0 est vraie, χ_0^2 est distribuée selon une loi khi-deux à $\nu = k - p - 1$ degrés de liberté, où :
 - ▶ k est le nombre de classes retenues.
 - ▶ p est le nombre de paramètres estimés.
- ▶ Pour un niveau critique α donné, le test consiste à **rejeter H_0** si $\chi_0^2 > \chi_{\alpha;\nu}^2$.

Exemple 3

On dispose des données suivantes sur une variable X :

Valeurs (x_i)	1	2	3	Total
Effectifs observés (O_i)	28	18	12	58

Tester l'hypothèse selon laquelle les données proviennent d'une population distribuée selon une loi géométrique, i.e. :

$$H_0 : X \sim G(p).$$

Utiliser $\alpha = 0.05$.

Exemple 4

On dispose des données suivantes sur une variable X :

Intervalle	$[0, 0,5[$	$[0,5 1,0[$	$[1,0 1,5[$	$[1,5 2,0[$	$[2,0 2,5[$	$[2,5 3,0[$	$[3,0, \infty[$
Nombre observé	2	23	17	4	2	0	2

Tester l'hypothèse selon laquelle les données proviennent d'une population distribuée selon une loi normale, i.e.

$$H_0 : X \sim N(\mu, \sigma^2).$$

Utiliser $\alpha = 0.05$. La moyenne et l'écart type de l'échantillon sont $\bar{X} = 1.168$ et $S = 0.591$.

1. Introduction
2. Tests d'hypothèses avec 2 échantillons
3. Tests sur la normalité
4. Test d'ajustement du Khi-deux de Pearson
- 5. Test d'indépendance entre deux variables**
6. Analyse de la variance et test de l'égalité de plusieurs moyennes

Test d'indépendance entre deux variables

Il arrive en pratique que l'on étudie plusieurs variables simultanément. Dans le cas particulier de deux variables, on peut être amené à vérifier s'il existe un lien entre les deux. La méthode du khi-deux permet d'effectuer ce test.

Exemples

- ▶ On aimerait vérifier si, dans une population donnée, les hommes et les femmes ont la même opinion au sujet du tabagisme. On dit alors qu'on effectue le test de l'indépendance entre le sexe (X) et l'opinion (Y).
- ▶ On veut vérifier si le type de pneu (X) est dépendant du kilométrage parcouru avant usure (Y).

Test d'indépendance : méthode

Il s'agit dans ces cas d'un test non paramétrique des hypothèses :

H_0 : X et Y sont indépendantes

H_1 : X et Y sont dépendantes.

Afin d'effectuer un tel test, on prélève un échantillon de taille n de la population que l'on classe conjointement selon les r modalités de X et les c modalités de Y . On obtient alors un **tableau de contingence**.

Test d'indépendance : méthode (suite)

Le tableau de contingence a la forme suivante :

X \ Y	y_1	y_2	\cdots	y_j	\cdots	y_c	Total
x_1	O_{11}	O_{12}	\cdots	O_{1j}	\cdots	O_{1c}	$\sum_{j=1}^c O_{1j}$
x_2	O_{21}	O_{22}	\cdots	O_{2j}	\cdots	O_{2c}	$\sum_{j=1}^c O_{2j}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	O_{i1}	O_{i2}	\cdots	O_{ij}	\cdots	O_{ic}	$\sum_{j=1}^c O_{ij}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_r	O_{r1}	O_{r2}	\cdots	O_{rj}	\cdots	O_{rc}	$\sum_{j=1}^c O_{rj}$
Total	$\sum_{i=1}^r O_{i1}$	$\sum_{i=1}^r O_{i2}$	\cdots	$\sum_{i=1}^r O_{ij}$	\cdots	$\sum_{i=1}^r O_{ic}$	n

Test d'indépendance : méthode (suite)

Tout comme le cas du test d'ajustement, le principe du test du Khi-deux consiste à comparer les effectifs observés O_{ij} aux effectifs attendus E_{ij} si H_0 est vraie. Si les deux variables sont indépendantes, les effectifs attendus E_{ij} (avec $i = 1, \dots, r$ et $j = 1, \dots, c$) sont calculés à partir du tableau de contingence :

$$E_{ij} = \frac{1}{n} \left(\sum_{k=1}^c O_{ik} \right) \times \left(\sum_{l=1}^r O_{lj} \right).$$

La statistique du test est

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Test d'indépendance : méthode (suite)

- ▶ Lorsque H_0 est vraie, la statistique χ_0^2 suit une loi Khi-deux à $\nu = (r - 1) \times (c - 1)$ degrés de liberté.
- ▶ Pour un niveau critique α donné, le test consiste à rejeter H_0 si $\chi_0^2 > \chi_{\alpha;\nu}^2$.

Exemple 5

Une flotte d'autobus est équipée de quatre types de pneus (A, B, C, D). On mesure le kilométrage parcouru avant usure des pneus. On construit trois classes de kilométrage (en milliers) pour lesquelles on a obtenu les résultats suivants :

Observé	A	B	C	D	Total
< 20	26	23	15	32	96
[20; 30]	118	93	116	121	448
> 30	56	84	69	47	256
Total	200	200	200	200	800

Tester si les deux variables sont indépendantes au seuil critique $\alpha = 0.05$.

1. Introduction
2. Tests d'hypothèses avec 2 échantillons
3. Tests sur la normalité
4. Test d'ajustement du Khi-deux de Pearson
5. Test d'indépendance entre deux variables
- 6. Analyse de la variance et test de l'égalité de plusieurs moyennes**

Contexte

- ▶ On cherche à planifier une expérience afin d'analyser la relation entre une variable particulière d'intérêt et un certain facteur ou traitement.
- ▶ On considère le cas où l'expérience vise à comparer les moyennes de la variable d'intérêt pour plusieurs niveaux du même facteur.
- ▶ **Analyse de la variance** : méthode qui permet de comparer les moyennes de plusieurs niveaux d'un facteur.
- ▶ L'hypothèse utilisée sera H_0 : "tous les traitements ont la même moyenne".

Exemple 6

Un fabricant de sacs en papier veut améliorer la résistance à la traction de son papier. On pense que celle-ci est liée à la concentration en bois dur de la pâte (entre 5 et 20%).

- ▶ Variable d'intérêt : résistance à la traction du papier.
- ▶ Facteur : concentration en bois dur, avec quatre niveaux (5%, 10%, 15% et 20%).
- ▶ Expérience : fabriquer six éprouvettes par niveau de concentration et tester dans un ordre aléatoire la résistance à la traction de chacune des éprouvettes.

Exemple 6 (suite)

Concentration en bois dur (%)	Observations						Sommes	Moyennes
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

Cas général

- ▶ a traitements indexés par $i \in I = \{1, 2, \dots, a\}$.
- ▶ n observations par traitement, indexées par $j \in J = \{1, 2, \dots, n\}$.
- ▶ Le nombre total d'observations est $N = a \times n$.
- ▶ Y_{ij} : j ème observation obtenue durant le traitement i .

Traitement	Observations			
1	Y_{11}	Y_{12}	\dots	Y_{1n}
2	Y_{21}	Y_{22}	\dots	Y_{2n}
\dots	\dots	\dots	\dots	\dots
a	Y_{a1}	Y_{a2}	\dots	Y_{an}

Cas général (suite)

- ▶ Pour chaque traitement, on considère la v.a.
 $Y_i \sim N(\mu + \tau_i, \sigma^2)$, $i \in I$.
- ▶ On pose $Y_{ij} = \mu + \tau_i + E_{ij}$ avec
 - ▶ μ : moyenne globale de tous les traitements.
 - ▶ τ_i : *effet* du i ème traitement, avec $\sum_{i \in I} \tau_i = 0$. Peut être vu comme la différence entre la moyenne du traitement i et la moyenne globale μ .
 - ▶ E_{ij} : erreur aléatoire avec $E_{ij} \sim N(0, \sigma^2)$.
- ▶ On définit :

	sommés	moyennes
par traitement	$Y_{i.} = \sum_{j=1}^n Y_{ij}$	$\bar{Y}_{i.} = \frac{Y_{i.}}{n}$
globale	$Y_{..} = \sum_{i=1}^a \sum_{j=1}^n Y_{ij} = \sum_{i=1}^a Y_{i.}$	$\bar{Y}_{..} = \frac{Y_{..}}{N}$

Cas général (suite)

Traitement	Observations				Sommes	Moyennes
1	Y_{11}	Y_{12}	\dots	Y_{1n}	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$
2	Y_{21}	Y_{22}	\dots	Y_{2n}	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots
a	Y_{a1}	Y_{a2}	\dots	Y_{an}	$Y_{a\cdot}$	$\bar{Y}_{a\cdot}$
					$Y_{\cdot\cdot}$	$\bar{Y}_{\cdot\cdot}$

Variabilité des données

La *variabilité totale des données* est mesurée par SS_T la somme totale des carrés. Elle peut se décomposer en *variabilité entre les traitements* (SS_I) et *variabilité ou erreur à l'intérieur des traitements* (SS_E) :

$$\begin{aligned}
 SS_T &= \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{..})^2 = \left(\sum_{i=1}^a \sum_{j=1}^n Y_{ij}^2 \right) - \frac{Y_{..}^2}{N} \\
 &= n \sum_{i=1}^a (\bar{Y}_{i.} - \bar{Y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_{i.})^2 \\
 &= \left(\sum_{i=1}^a \frac{Y_{i.}^2}{n} \right) - \frac{Y_{..}^2}{N} + \left(\sum_{i=1}^a \sum_{j=1}^n Y_{ij}^2 \right) - \left(\sum_{i=1}^a \frac{Y_{i.}^2}{n} \right) \\
 &= SS_I + SS_E .
 \end{aligned}$$

(SS =Sum of Squares).

Exemple 6 (suite)

Concentration en bois dur (%)	Observations						Sommes	Moyennes
	1	2	3	4	5	6		
5	7	8	15	11	9	10	$Y_{1.} = 60$	$\bar{Y}_{1.} = 10.00$
10	12	17	13	18	19	15	$Y_{2.} = 94$	$\bar{Y}_{2.} = 15.67$
15	14	18	19	17	16	18	$Y_{3.} = 102$	$\bar{Y}_{3.} = 17.00$
20	19	25	22	23	18	20	$Y_{4.} = 127$	$\bar{Y}_{4.} = 21.17$
							$Y_{..} = 383$	$\bar{Y}_{..} = 15.96$

- ▶ $a = 4$, $n = 6$ et $N = 24$.
- ▶ $SS_T = 6625 - 383^2/24 = 512.96$.
- ▶ $SS_I = 6494.83 - 383^2/24 = 382.79$.
- ▶ $SS_E = SS_T - SS_I = 130.17$.

Hypothèse nulle et lois suivies

- ▶ On veut vérifier si les a moyennes de traitements sont égales, c'est-à-dire si les effets des traitements sont nuls. On pose les hypothèses :
$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0.$$
$$H_1 : \tau_i \neq 0 \text{ pour au moins un traitement } i.$$
- ▶ $\frac{SS_T}{\sigma^2} \sim \chi_{N-1}^2$, $\frac{SS_I}{\sigma^2} \underset{H_0}{\sim} \chi_{a-1}^2$, $\frac{SS_E}{\sigma^2} \sim \chi_{N-a}^2$.
- ▶ Les degrés de liberté correspondent au nombre de valeurs aléatoires qui ne peuvent être déterminées ou fixées par une équation. C'est aussi le nombre d'observations moins le nombre de relations entre ces observations.

Hypothèse nulle et lois suivies (suite)

- ▶ Statistique du test :

$$F_0 = \frac{SS_I / (a - 1)}{SS_E / (N - a)} = \frac{MS_I}{MS_E} \underset{H_0}{\sim} F_{a-1, N-a}$$

(*MS = Mean of Squares*).

- ▶ On peut montrer que $E(MS_E) = \sigma^2$ et $E(MS_I) = \sigma^2 + \frac{n}{a-1} \sum_{i=1}^a \tau_i = \sigma^2$.
- ▶ Si H_0 est vraie, MS_E et MS_I sont des estimateurs sans biais de σ^2 . Sinon, on aura, en moyenne, une grande valeur pour F_0 .
- ▶ L'hypothèse nulle est donc rejetée si $F_0 > F_{\alpha; a-1, N-a}$.

Tableau d'analyse de la variance

Source de variation	Somme des carrés	Nombre de ddl	Moyenne des carrés	F_0
Entre les traitements	SS_I	$a - 1$	MS_I	$\frac{MS_I}{MS_E}$
Erreur dans les traitements	SS_E	$N - a$	MS_E	
Total	SS_T	$N - 1$		

Tableau d'analyse de la variance pour l'exemple 6

Source de variation	Somme des carrés	Nombre de ddl	Moyenne des carrés	F_0
Conc. bois dur	382.79	3	127.60	19.6
Erreur	130.17	20	6.51	
Total	512.96	23		

Pour $\alpha = 1\%$, on a le quantile de Fisher $F_{0.01;3,20} = 4,94$. On rejette donc H_0 et on conclut que la concentration en bois dur influe de façon significative sur la résistance du papier.