Chapitre III. Construction de tests en présence de 2 distributions

1) Test sur la corrélation

1.1. Deux variables quantitatives : le coefficient de corrélation linéaire

Le coefficient de corrélation linéaire cherche à mesurer la liaison linéaire qui peut exister entre deux variables X et Y observées sur les mêmes individus.

$$\rho_{XY} = E\left(\frac{X - E(X)}{\sigma_X} \times \frac{Y - E(Y)}{\sigma_Y}\right)$$

qui sera estimé par le **coefficient de corrélation linéaire empirique :**

$$\hat{\rho}(X,Y) = \hat{\rho}_{XY} = \frac{1}{n}\sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\hat{\sigma}_X} \frac{(Y_i - \overline{Y})}{\hat{\sigma}_Y}$$

et on a $-1 \leq \hat{\rho}_{X,Y} \leq 1$.

Plus ce coefficient se rapproche de 1, plus les variables sont corrélées positivement, c'est-à-dire qu'elles varient dans le même sens. Plus il se rapproche de -1, plus elles varient en sens opposé.

S'il se rapproche de 0, leurs variations ne sont pas liées linéairement.

Exemple 1 : Deux séries de notes observées sur 12 individus

X :	14	13	17	15	14	15	16	12	14	12	13	13
Y :	13	11	16	15	12	13	15	10	14	12	13	12

Ex1 : $\hat{\rho} = 0.874$.

Exemple 2: “fichier notes”

	Math	Phys	Chim	Angl	Fran	Hist
1	15	17	18	9	8	10
2	6	7	5	10	7	5
3	7	4	4	13	15	19
4	18	19	19	18	14	16
5	8	12	10	10	11	9
6	15	14	19	12	6	8
7	6	10	5	19	13	16
8	14	16	12	17	11	15
9	8	7	8	9	10	10
10	7	9	7	7	5	5
11	9	10	11	12	14	13
12	14	18	15	6	7	5
13	5	7	9	18	16	16
14	6	11	10	9	5	8
15	14	16	18	16	11	10
16	16	12	12	14	11	14
17	14	16	15	8	7	10
18	9	8	13	12	9	10
19	6	4	7	15	14	17
20	8	4	3	8	9	8
21	12	15	13	15	12	13
22	7	4	3	17	15	13
23	5	8	7	9	9	9
24	16	14	17	7	9	6
25	7	11	12	11	10	9

Variable	Moyenne	Ecart-type	Minimum	Maximum
Mathém	10.08	4.17	5.00	18.00
Physique	10.92	4.69	4.00	19.00
Chimie	10.88	5.07	3.00	19.00
Anglais	12.04	3.94	6.00	19.00
Français	10.32	3.22	5.00	16.00
Histoire	10.96	4.03	5.00	19.00

	Math.	Phys.	Chim.	Angl.	Fran.	Hist.
Mathém	1.00	0.82	0.83	-0.00	-0.15	-0.05
Physique	0.82	1.00	0.87	-0.04	-0.29	-0.18
Chimie	0.83	0.87	1.00	-0.05	-0.25	-0.17
Anglais	-0.00	-0.04	-0.05	1.00	0.76	0.80
Français	-0.15	-0.29	-0.25	0.76	1.00	0.85
Histoire	-0.05	-0.18	-0.17	0.80	0.85	1.00

Ce tableau s'appelle le tableau ou matrice des corrélations.

!! Attention

Un coefficient de corrélation ne traduit pas nécessairement une relation de cause à effet : “Une bonne note en math n’implique pas une bonne note en chimie.”

Autre exemple : la corrélation entre le revenu et le débit de carte bancaire est fortement positif. Il existe ici une relation évidente : plus le revenu est élevé plus le débit de carte bancaire va augmenter et pas le contraire!!!

!! La relation n’est pas contenue dans les données.

1.2. Quand doit-on considérer le coefficient de corrélation comme significativement non nul?

C'est encore un problème de test statistique avec H_0 : "$\rho = 0$".

En se plaçant sous l'hypothèse "absence de lien linéaire" entre 2 variables quantitatives et sous les hypothèses du théorème central limite :

$$\frac{\hat{\rho}}{\sqrt{\frac{1}{n-1}}} \sim \mathcal{N}(0,1)$$

On décidera de la dépendance entre 2 variables quantitatives lorsque

$\sqrt{n-1}\hat{\rho}$ est grand en valeur absolue.

1.3. Rejet de l'hypothèse H_0 "$\rho = 0$" : "absence de lien linéaire" (contre H_1 "$\rho \neq 0$") si

$$\boxed{|\hat{\rho}| > \frac{l_{\alpha/2}}{\sqrt{n-1}}}$$

Exemple : pour $\alpha = 5\%$, on rejettera l'hypothèse "absence de lien linéaire" si $|\hat{\rho}| > \frac{1.96}{\sqrt{n-1}}$.

C'est à dire, rejet :

si $\lvert\hat{\rho}\rvert > 0.653$	$(n = 10)$	si $\lvert\hat{\rho}\rvert > 0.197$	$(n = 100)$
si $\lvert\hat{\rho}\rvert > 0.450$	$(n = 20)$	si $\lvert\hat{\rho}\rvert > 0.0877$	$(n = 500)$
si $\lvert\hat{\rho}\rvert > 0.364$	$(n = 30)$	si $\lvert\hat{\rho}\rvert > 0.0620$	$(n = 1000)$
si $\lvert\hat{\rho}\rvert > 0.314$	$(n = 40)$	si $\lvert\hat{\rho}\rvert > 0.0196$	$(n = 10\,000)$
si $\lvert\hat{\rho}\rvert > 0.280$	$(n = 50)$	si $\lvert\hat{\rho}\rvert > 0.00196$	$(n = 1\,000\,000)$

2) Un test d'adéquation : le test du χ^2

2.1. Exemples

Exemple 1 : "Lancer de dés"

Une expérience consiste à lancer deux dés, et à relever la somme des chiffres lus. On fait l'expérience $n = 1000$ fois, et on obtient :

S	2	3	4	5	6	7	8	9	10	11	12
n_k	32	56	81	115	142	160	143	105	89	53	24

Exemple 2 : "Les familles de 8 enfants"

On a observé, en étudiant 53680 familles de 8 enfants, les résultats suivants (k désigne le nombre de garçons et n_k le nombre de familles ayant k garçons) :

k	0	1	2	3	4	5	6	7	8
n_k	215	1485	5331	10649	14959	11929	6678	2092	342

Exemple 3 : “10 000 premières décimales du nombre π”

La répartition de ces décimales est donnée dans le tableau suivant :

décimale	0	1	2	3	4	5	6	7	8	9
effectifs	968	1026	1021	974	1012	1046	1021	970	948	1014

Se répartissent-elles de manière uniforme?

Exemple 4 : “Sondage sur le niveau d’acceptation d’un nouveau système”

Appréciation Âge des sondés	Très difficile	Assez difficile	Peu/pas difficile
de 18 à 29 ans	81	138	132
de 30 à 40 ans	126	131	94
41 ans et plus	203	78	69

2.2. Distance entre 2 tableaux :

Le tableau des observations $[n_k]$ et le tableau sous hypothèse d'un modèle avec les effectifs théoriques $[\tilde{n}_k]$

$$\chi^2 = \sum_k \frac{(n_k - \tilde{n}_k)^2}{\tilde{n}_k}$$

Cette distance sera donc d'autant plus grande que le tableau des observations sera loin du tableau sous l'hypothèse du modèle "théorique".

Pourquoi diviser par $\tilde{n}_k$?

Théo	1000	100	10
Obs.	1010	110	20
Ecart	négligeable	faible	important
$\chi^2_{cellule}$	1/10	1	10

2.3. Quand doit-on considérer cette distance du χ^2 comme grande?

C'est un problème décisionnel $\Rightarrow$ problème de test statistique.

Pour chaque situation on a une hypothèse H_0 "adéquation à un modèle" et une alternative H_1 = non H_0, et on aura :

rejet de H_0 si $\sum_k \frac{(n_k - \tilde{n}_k)^2}{\tilde{n}_k} \geq l_{\nu,\alpha}$

On décidera de rejeter l'adéquation au modèle lorsque la distance du χ^2 sera supérieure à une valeur limite, qui dépend d'un degré de liberté et de l'erreur de 1^{re} espèce choisie.

Pour répondre à cette question, on s'aidera d'une table appelée "Table du χ^2" qui donne cette valeur limite pour chaque degré de liberté et différentes erreurs de 1^{re} espèce.

3) Analyse de la variance ou Anova

3.1. Introduction

Objectif : étudier l'effet d'une ou plusieurs variables qualitatives sur une variable quantitative.

Le cas d'une variable qualitatitive : on observe sur des individus à la fois une variable quantitative et une variable qualitative. On cherche alors à savoir si les différentes modalités de la variable qualitative influencent la variable quantitative.

Exemple : On considère 6 échantillons de patients correspondant à des localisations différentes. Pour chaque patient, on observe une donnée clinique :

1	2	3	4	5	6
1602	1472	1548	1435	1493	1585
1615	1477	1555	1438	1498	1592
1624	1485	1559	1448	1509	1598
1631	1493	1563	1449	1516	1604
	1496	1575	1454	1521	1609
	1504		1458	1523	1612
	1510		1467		
			1475		

Informations disponibles pour chaque patient :

- Y : donnée clinique
- X : code de la localisation

  Question : peut-on considérer que la localisation a une influence sur la donnée clinique des patients?

  ... ou encore : la variable X a-t-elle une influence sur la variable Y?

  ... ou encore : la modélisation de l'espérance μ de Y doit-elle être différente selon les modalités de X ou non?

Vocabulaire

On appelle facteur (ou facteur explicatif, cause contrôlée) la variable qualitative X qui sert à expliquer Y.

On parle de niveaux d'un facteur (ou traitements) pour désigner les différentes modalités de cette variable.

Lorsqu'on étudie l'effet de plusieurs facteurs sur Y, on peut regarder leurs effets cumulés mais aussi l'effet de leur interaction, i.e. le croisement de deux modalités a une influence particulière sur Y.

On appelle unité expérimentale le sujet que l'on soumet à un traitement et sur lequel on mesure Y.

De façon générale ici, on suppose que l'on ne soumet pas une unité expérimentale à plusieurs traitements, autrement dit on suppose qu'il n'y a pas de répétition de la mesure de Y. Il y a donc autant d'observations que d'unités expérimentales.

ANOVA : pourquoi?

- Analyse de la variance :
  - Variations inter-groupes : écart entre les moyennes des groupes, dispersion des moyennes autour de la moyenne globale.
  - Variations intra-groupes : écart entre les données à l'intérieur des groupes, dispersion des données autour de leur moyenne de groupe.

- $ANOVA$ ↗ technique de comparaison de moyennes

- $ANOVA$ ↘ décider de l'égalité des moyennes

3.2. ANOVA à un facteur contrôlé

Approche intuitive sur un exemple

Un étudiant a mesuré le temps de parcours pour se rendre à la fac selon trois types de trajet.

T_1	17.5	20.0	18.0	17.0	16.5
T_2	15.1	16.0	13.0	12.0	14.5
T_3	10.0	13.0	10.0	11.0	12.0

Structure générale des données :

On a le tableau des données suivant :

Facteur	niveau 1	...	niveau k	...	niveau K
Observations	y_{11}	...	y_{k1}	...	y_{K1}
indépendantes	y_{12}	...	y_{k2}	...	y_{K2}
de la variable	$\vdots$		$\vdots$		$\vdots$
quantitative	y_{1n_1}	...	y_{kn_k}	...	y_{Kn_K}

- y_{ki} : i^e observation du niveau k
- n_k : nombre d'observations du niveau k

- $n = \sum_{k=1}^{K} n_k$: nombre total d'observations.
- $\overline{y}_{k\bullet} = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ki}$: moyenne des observations pour le niveau k.
- $\overline{y}_{\bullet\bullet} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} y_{ki}$: moyenne globale.

!!! Attention

La moyenne des observations $\overline{y}_{\bullet\bullet}$ n'est pas égale à la moyenne des moyennes par niveau du facteur $\overline{y}_{k\bullet}$.

Quelques hypothèses naturelles

Les Y_{ki} sont les résultats aléatoires de l'expérience étudiée et on suppose que leur "valeur espérée" est un paramètre noté μ_k qui ne dépend que du niveau du facteur contrôlé.

- μ_k est appelé l'effet fixe du niveau k du facteur contrôlé. C'est un paramètre inconnu : l'espérance de l'observation Y_{ki}.
- Les Y_{ki} sont aussi supposés être indépendants (donc associés à des sujets distincts).

Dans cette modélisation, il y a donc K paramètres inconnus liés à l'espérance : un paramètre pour chacun des niveaux du facteur.

Equation d'analyse de la variance :

$\sum_{ki}(y_{ki} - \overline{y}_{\bullet\bullet})^2$	=	$\sum_{ki}(y_{ki} - \overline{y}_{k\bullet})^2$	+	$\sum_{k} n_k(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2$
dispersion totale	=	dispersion INTRA	+	dispersion INTER
SC_T	=	SC_R	+	SC_F
↑		↑		↑
Somme des Carrés	=	Somme des Carrés	+	Somme des Carrés
Totale		Résiduelle		due au facteur contrôlé

Degrés de liberté associés aux SC

ddl(SC_T)	=	ddl(SC_R)	+	ddl(SC_F)
n-1	=	n-K	+	K-1

Données :

1	2	3	4	5	6	
1602	1472	1548	1435	1493	1585	
1615	1477	1555	1438	1498	1592	
1624	1485	1559	1448	1509	1598	
1631	1493	1563	1449	1516	1604	
	1496	1575	1454	1521	1609	
	1504		1458	1523	1612	
	1510		1467			
			1475			
4	7	5	8	6	6	n_k
1618	1491	1560	1453	1510	1600	$\overline{y}_{k\bullet}$
470	1152	404	1296	760	534	$\sum_i (y_{ki} - \overline{y}_{k\bullet})^2$

Analyse de la variance sur l'exemple des patients :

k	1	2	3	4	5	6
$\overline{y}_{k\bullet}$	1618	1491	1560	1453	1510	1600

$\overline{y}_{\bullet\bullet} = \frac{1}{n}\sum_{ki} y_{ki} = \frac{1}{n}\sum_{k} n_k \overline{y}_{k\bullet} = 1527.58$. Ici $n = 36$.

Tableau d'analyse de la variance

Source de dispersion	Somme des carrés	ddl
INTER	125144.8 $\sum_k n_k(\overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet})^2$	5 $K-1$
INTRA	4616 $\sum_{ki}(y_{ki} - \overline{y}_{k\bullet})^2$	30 $n-K$
TOTALE	129760.8 $\sum_{ki}(y_{ki} - \overline{y}_{\bullet\bullet})^2$	35 $n-1$

Test de l'égalité des espérances :

- l'hypothèse H_0 : absence d'effet du facteur contrôlé $\Rightarrow$ égalité des espérances $\mu_1 = \mu_2 = \ldots = \mu_K$.
- l'hypothèse alternative H_1 : effet significatif du facteur contrôlé $\Rightarrow$ différence des espérances μ_k, $k = 1, \ldots, K$.

$$\frac{INTER}{INTRA} * (n-K) = \frac{\sum_{ki}(y_{ki} - \overline{y}_{\bullet\bullet})^2 - \sum_{ki}(y_{ki} - \overline{y}_{k\bullet})^2}{\left(\sum_{ki}(y_{ki} - \overline{y}_{k\bullet})^2\right)/(n-K)}$$

suit sous H_0 une loi de χ^2 à $(K-1)$ degrés de liberté dès que le nombre d'observations est grand.

Re-écriture du paramètre μ_k et interprétation

$$\mu_k = \mu + \alpha_k, \qquad k = 1, \ldots, K \tag{1}$$

avec $K + 1$ paramètres pour l'espérance dont seulement K sont libres et identifiables $\Rightarrow$ il y a sur-paramétrisation.

Différentes contraintes peuvent alors être envisagées :

- $\sum_k \alpha_k = 0$
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \frac{1}{K} \sum_{k=1}^{K} \mu_k = \overline{\mu} \qquad \text{et} \qquad \alpha_k = \mu_k - \overline{\mu}$$

  Le paramètre μ représente alors l'effet moyen général.

  Le paramètre α_k représente alors l'effet différentiel du niveau k à la "moyenne".

- $\alpha_1 = 0$, par défaut dans de nombreux logiciels de statistiques.
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \mu_1 \qquad \text{et} \qquad \alpha_k = \mu_k - \mu_1$$

  Le paramètre μ représente alors l' effet du niveau 1 du facteur.

  Le paramètre α_k représente alors l'effet différentiel du niveau k à l'effet du niveau 1.

  Ici le traitement 1 sert de référence mais on peut prendre l'un quelconque des K traitements comme référence.

Estimation des paramètres :

Par moindres carrés : $\min_{\mu_k} \sum_k^K \sum_{i=1}^{n_k} (y_{ki} - \mu_k)^2$

- $\hat{\mu}_k = \overline{y}_{k\bullet} = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ki}$
- contraste "sum" :

$$\hat{\mu} = \hat{\overline{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mu}_k = \frac{1}{K} \sum_{k=1}^{K} \overline{y}_{k\bullet} \quad \text{et} \quad \hat{\alpha}_k = \hat{\mu}_k - \hat{\mu} = \overline{y}_{k\bullet} - \frac{1}{K} \sum_{k=1}^{K} \overline{y}_{k\bullet}$$

Remarque : si pour $\forall k \in \{1, \ldots, K\}$, $n_k = I$ (c'est-à-dire que les nombres d'observations par niveau sont les mêmes) alors $\hat{\mu} = \overline{y}_{\bullet\bullet}$ et $\hat{\alpha}_k = \overline{y}_{k\bullet} - \overline{y}_{\bullet\bullet}$.

- contraste "treatment" :

$$\hat{\mu} = \hat{\mu}_1 = \overline{y}_{1\bullet} \quad \text{et} \quad \hat{\alpha}_k = \hat{\mu}_k - \hat{\mu}_1 = \overline{y}_{k\bullet} - \overline{y}_{1\bullet}$$