Pb général des stats : lien/effet entre deux variables.

— 2 variables qualitatives : test du Chi-2

— 2 variables quantitatives : Régression linéaire avec test de Fisher (Anova). ← nombreuses modalités

Part IV

ANOVA

Nous avons utilisé l'ANOVA dans l'étude du modèle linéaire

$$Y = \beta_1 \xi + \beta_0 + \epsilon,$$

(ξ : X quantitative ; $\beta_1\xi + \beta_0$: affine ; ϵ : erreur.)

avec certaines hypothèses sur ξ, Y, ϵ de normalité et de non corrélation des termes d'erreur. Nous avions utilisé les test de Student et de Fisher afin de vérifier la non nullité de β_1, ce qui entrainerait l'absence d'effet de ξ sur Y, par l'étude des moyennes ou des variances.

ξ qualitative / X quantitative

facteur — réponse.

⇒ regrouper les valeurs de X dans des groupes (échantillons) dont la propriété commune est la valeur de ξ.

La variable explicative ξ était alors quantitative. Il n'est cependant pas rare de rencontrer une variable explicative qualitative. Le passage par une régression linéaire n'a plus de sens dès que la multiplication $\beta_1\xi$ n'en a plus. Prenons par exemple la variable ξ à deux modalités :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",

ou plus :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",
- ξ_3 : "traitement expérimental à forte dose".

)) nombre de modalités de ξ = niveaux de facteur

2

3

!

à niveaux de facteur

On regroupe les observations de X sur les niveaux de facteur.

La variable ξ s'appelle le facteur. On pourra chercher à expliquer une variable réponse X, par exemple le taux d'une hormone. Pour chaque valeur ξ_i, on obtient un échantillon indépendant X_i. Dans le premier cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$

,

taille n_1, n_2, n — Moyennes μ_1, μ_2 — $n = n_1 + n_2$

ou dans le deuxième cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$
$\xi = \xi_3$: "traitement expérimental à forte dose"	$X_{3,1}, X_{3,2}, \cdots X_{3,n_3}$

.

n_1, n_2, n_3, n — μ_1, μ_2, μ_3 — $n = n_1 + n_2 + n_3$

On considère le modèle

moyenne sous-population μ_ℓ — effet aléatoire.

$$X_{ij} = EX_i + \epsilon_{ij}$$

La question est de savoir si les $\mu_i = EX_i$ sont identiques (ξ n'as pas d'effet sur X) ou différents selon les valeurs ξ_i. Dans ce cas, ξ influence X.

écart-type constant

Hypothèse du modèle : les $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ i.i.d.

Modèle $X_{ij} = \mu_i + \varepsilon_{ij}$.

$\xi = \xi_i$ → $j$ème observation dans la population i

Pb: Si ξ a un effet sur X alors les μ_i sont différentes.

Si ξ n'a pas d'effet sur X, alors $\mu_1 = \mu_2 = \dots = \mu_a = \mu$

μ est la moyenne de la population totale.

Estimation ponctuelle des moyennes:

Sur $\xi = \xi_i$: $\overline{X_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ avec l'effectif n_i, les observations X_{ij}.

→ estime μ_i

Sur la population totale: $\overline{X} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{ij}$ → estime μ.

↑ sur tous les échantillons, toutes les observations.

Hypothèse nulle à tester : $\mathcal{H}_0$ = " $\mu_1 = \mu_2 = \cdots = \mu_a = \mu$ ".

= " ξ n'explique pas X ".

10.1 Facteur à deux valeurs - t de Student

On considère deux échantillons indépendants de tailles n_1 et n_2, respectivement :

		moyenne
$\xi = \xi_1$	$X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}$;	$\bar{X}_1$
$\xi = \xi_2$	$X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}$,	$\bar{X}_2$

$\mathcal{H}_0$ = " $\bar{X}_1 - \bar{X}_2 = 0$ "

↳ stat du test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}}$$ variable de Student.

Comment généraliser $\bar{X}_1 - \bar{X}_2 = 0$ à plus de modalités de facteur ?

10.2 Facteur à a modalités

10.2.1 Le modèle

Supposons donc qu'on prélève a échantillons indépendants :

		effectifs	moyennes	Estimation des moyennes
$\xi = \xi_1$	$X_{1,1}, X_{1,2}, \cdots, X_{1,n_1};$	n_1	μ_1	$\overline{X_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}$
$\xi = \xi_2$	$X_{2,1}, X_{2,2}, \cdots, X_{2,n_2};$	n_2	μ_2	$\vdots$
$\vdots$	$\vdots$			
$\xi = \xi_a$	$X_{a,1}, X_{a,2}, \cdots, X_{a,n_a};$	n_a	μ_a	$\overline{X_a} = \frac{1}{n_a}\sum_{j=1}^{n_a} X_{aj}$
	total	n	N	$\overline{X} = \frac{1}{n}\sum_i \sum_j X_{ij}$

où $X_{i,j}$ représente la $j$ème observation du $i$ème échantillon, ($i = 1, \cdots, a$ et $j = 1, \cdots, n_i$). Les échantillons indépendants sont issus des populations normales de moyenne $\mu_1, \cdots, \mu_a$ et de variance commune σ^2. On pose donc le modèle

estimé par $\overline{X_i}$

$$X_{i,j} = \mu_i + \varepsilon_{ij},$$

où les ε_{ij} sont des lois normales $\mathcal{N}(0, \sigma)$ indépendantes.

10.2.2 Le test de Fisher

L'hypothèse à tester est

$$\mathcal{H}_0 : \text{“}\mu_1 = \mu_2 = \cdots = \mu_a = \mu\text{”}.$$

Chacune des espérances μ_i des échantillons sont estimées par les moyennes

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j},$$

et par la moyenne totale

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^{a} n_i \bar{X}_i$$

pour μ. Une bonne façon de tester l'égalité de toutes les moyennes est de les comparer à la moyenne commune $\bar{X}$:

$$\bar{X}_1 = \bar{X}_2 = \dots = \bar{X}_a = \bar{X} \iff \sum_{i=1}^{a} (\bar{X}_i - \bar{X})^2 = 0 \iff \sum_{i=1}^{a} n_i (\bar{X}_i - \bar{X})^2 = 0$$

multiplier par les effectifs

ou mieux, en faisant apparaître le rapport de force de chaque moyenne grâce à l'effectif de chaque échantillon

On étudie donc $$SCM = \sum_{i=1}^{a} n_i (\bar{X}_i - \bar{X})^2,$$

où SCM signifie somme des carrés due au modèle.

"On s'attend à ce que SCM soit très proche de zéro si H_0 est vraie"

Note: SCM a $(a-1)$ ddl indépendants:

puisque $\bar{X} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{a} n_i \bar{X}_i$. On peut déduire un $\bar{X}_i$ de $\bar{X}$ et des autres $\bar{X}_i$. $(a-1)$ ddl indépendants

$n_i \bar{X}_i$ Si $\bar{X}$ considéré comme connu et les $\bar{X}_i$ comme aléatoire.

On réduit la variable pour obtenir une loi connue.

$$\overline{X_i} - \mu \underset{\text{sous } \mathcal{H}_0}{=} \overline{X_i} - \mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} - \mu_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} - \frac{n_i \mu_i}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i}(X_{ij} - \mu_i)$$

($\sum_{i=1}^{n_i} \mu_i = n_i \mu_i$, $X_{ij} - \mu_i = \varepsilon_{ij}$)

$$= \frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_{ij} = \overline{\varepsilon_i}$$

$\overline{X_i} - \mu$ s'estime $\overline{X_i} - \overline{X}$.

$$\mathrm{Var}(X_i - \mu) = \mathrm{Var}\,\overline{\varepsilon_i} = \mathrm{Var}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} \varepsilon_{ij}\right) = \frac{1}{n_i^2}\mathrm{Var}\left(\sum_{j=1}^{n_i} \varepsilon_{ij}\right)$$

$\mathrm{Var}(aX) = a^2 \mathrm{Var}\,X$

indépendance des ε_{ij}

$$= \frac{1}{n_i^2}\sum_{j=1}^{n_i} \mathrm{Var}(\varepsilon_{ij}) \qquad (\mathrm{Var}(\varepsilon_{ij}) = \sigma^2)$$

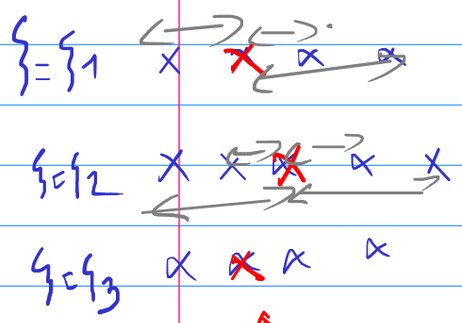
$$= \frac{1}{n_i^2}\sum_{j=1}^{n_i} \sigma^2 = \frac{n_i \sigma^2}{n_i^2} = \frac{\sigma^2}{n_i}$$

$\sqrt{n_i}(\overline{X_i} - \overline{X})$ a pour variance $(\sqrt{n_i})^2 \frac{\sigma^2}{n_i} = \sigma^2$.

Estimation de σ^2

σ^2 ← variance de l'erreur.

1 écart de moins que l'effectif total.



$n_1 - 1$

$n_2 - 1$

$n_3 - 1$ ...

Dans la population entière :

$(n_1 - 1) + (n_2 - 1) + \dots + (n_a - 1) = n - a$

← en estimant la variance, on suppose qu'il y a un individu moyen.

les erreurs sont centrées : $\mathrm{Var}(\xi) = \frac{1}{n-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\xi_{ij} - \overline{\xi})^2$ ($\overline{\xi}$: 0 par hypothèse)

$= \frac{1}{n-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} \xi_{ij}^2$ (ξ_{ij} : $X_{ij} - \overline{X_i}$)

$= \frac{1}{n-a} \underbrace{\sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X_i}\right)^2}$

SCE = somme des carrés des erreurs.

Donc σ^2 est estimé par $SCE / n-a$ // $\to n-a$ écarts indépendants

Rappel : $F = \dfrac{\chi^2(n_1)/n_1}{\chi^2(n_2)/n_2} \sim \mathcal{F}_{n_1,n_2}$ (Fisher)

$\chi^2(n) = \sum_{i=1}^{n} X_i^2$ où les X_i sont $\mathcal{N}(0,1)$ i.i.d.

SCM est une somme de carrés de variance σ^2. $\frac{SCM}{\sigma^2} \sim \chi^2(a-1)$.

$SCE/n-a$ est une somme de carrés qui estime σ^2. $\frac{SCE}{\sigma^2} \sim \chi^2(n-a)$

La variable du test est donc

$$F = \frac{SCM/(a-1)}{SCE/n-a} \sim \mathcal{F}_{a-1,n-a}.$$

(χ^2 divisés par leurs ddl.)

Nous rejetons $\mathcal{H}_0$ au seuil α si

$$F = \frac{CMM}{CME} = \frac{SCM/(a-1)}{SCE/n-a} \geq q_{1-\alpha}^{\mathcal{F}_{a-1,n-a}},$$

où q est le quantile d'ordre $1-\alpha$ de la dite loi.

Remarquons que nous rejetons $\mathcal{H}_0$ seulement si F est trop grand et non si F est trop petit car un F grand signifie que les $\bar{X}_i$ sont trop dispersés, et donc que les μ_i ne semblent pas être tous égaux.

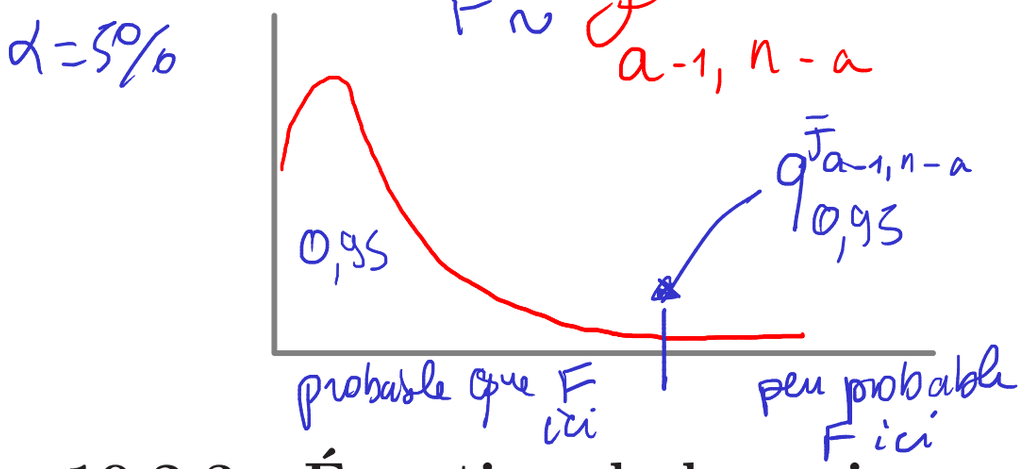
Variable (statistique) du test :

$$F = \frac{SCM/(a-1)}{SCE/(n-a)} \sim \mathcal{F}_{a-1,n-a} \text{ sous } \mathcal{H}_0.$$

$\Leftrightarrow$ $SCM \simeq 0$.

Sous H_0 :

$F \sim \mathcal{F}_{a-1,\, n-a}$



Conclusion :

Au seuil 5%, si $F > q_{0,95}$, on rejette H_0 et donc, on retient que ξ a un effet sur X.

10.2.3 Équation de la variance

Posons de plus

$$SCT = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2,$$

somme des carrés des différences entre les observations et la moyenne globale.

pour la dispersion totale. On peut aisément établir l'équation de la variance suivante.

$$SCT = SCM + SCE.$$

(Analogue à celle des régressions linéaire)

Cette décomposition met en évidence le fait que la dispersion totale des données (SCT) est formée d'une partie (SCM) expliquée par le fait que les populations sont différentes, et d'une autre partie (SCE) qu'on attribue au hasard. Autrement dit, SCE représente les différences individuelles alors que SCM représente les différences entre les groupes. On rejette l'hypothèse que les populations d'origine des groupes sont de même moyenne si les différences entre les groupes sont trop grandes par rapport aux différences individuelles. Cette analyse est appelée analyse de variance. Les calculs se font plus aisément à l'aide des formules suivantes :

$$SCM = \sum_{i=1}^{a} n_i \bar{X}_i^2 - n\bar{X}^2,$$

$$SCT = \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2,$$

$$SCE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{a} n_i \bar{X}_i^2.$$

Source	**Somme des carrés**	**d.l.**	**Moyenne des carrés**	F
Modèle	$\mathrm{SCM} = \sum_{i=1}^{a} n_i \bar{X}_i^2 - n\bar{X}^2$	$a-1$	$\frac{\mathrm{SCM}}{a-1}$	$F = \frac{\mathrm{CMM}}{\mathrm{CME}}$
Erreur	$\mathrm{SCE} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2$	$n-a$	$\frac{\mathrm{SCE}}{n-a}$	
Total	$\mathrm{SCT} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}^2$	$n-1$	$\frac{\mathrm{SCT}}{n-1}$	

10.2.4 Exemple

Nous allons reprendre ici un exemple du livre de Snedecor and Cochran (1989). Pendant leur cuisson les beignets absorbent de la matière grasse en quantité variable. On peut se demander si la quantité absorbée dépend de la matière grasse utilisée ? Pour chacune des quatre matières grasses, on a constitué six fournées de 24 beignets chacune. La mesure est la quantité, en grammes, de matière grasse absorbée, par fournée. On a simplifié les calculs en leur soustrayant 100 g. Les données de ce genre constituent une classification à une seule entrée, ou à une seule voie ou classification simple; on dit aussi à un seul facteur, chaque matière grasse représentant une classe, ou niveau du facteur.

Résumé des données.

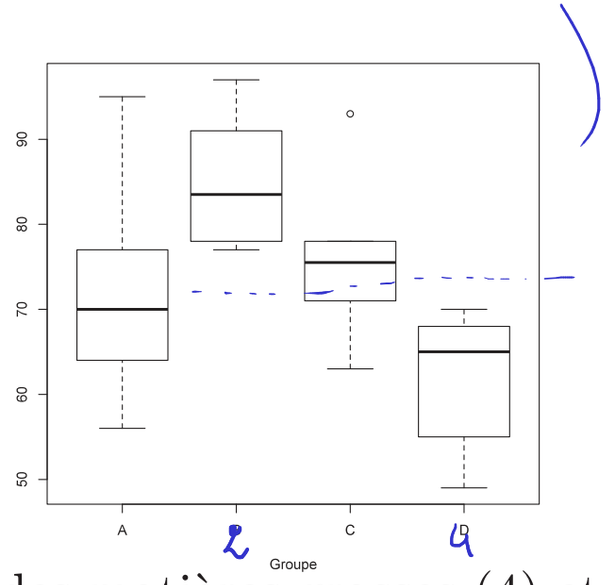
RAPPORT DÉTAILLÉ

Groupes	n_i	*Somme*	*Moyenne*	*Variance*
1	6	432	72	178
2	6	510	85	60,4
3	6	456	76	97,6
4	6	372	62	67,6

4 différents types de graisse.

Avant de commencer l'analyse, notons que les quatre absorptions totales de M.G. diffèrent de façon visible : de 372 pour la 4e à 510 pour la 2e.

Boxplot. Quantité de matière grasse absorbée, par fournée, en grammes.



II y a en effet une séparation assez net entre les résultats individuels des matières grasses (4) et (2), 70 est la plus haute valeur donne par la M.G. (4) tandis que 77 est la plus basse pour la M.G. (2). Pour les autres paires d'échantillons, on observe un certain chevauchement des résultats.

On rappelle que le test de Fisher de l'ANOVA a pour hypothèse nulle

$$\mathcal{H}_0 = \text{“les moyennes sur chaque groupe sont égales”}.$$

P-value
0,007 < 5%.
On rejette Ho.

ANALYSE DE VARIANCE

Source	*S. C.*	*d. l.*	*C.M.*	*F*	*Prob.*	$F_{3,20;0,05}$
Inter groupes	1636,5	3	545,5	5,41	0,0069	3,10
Intra groupes	2018,0	20	100,9			
Total	3654,5	23				

Plus simplement, vu la p-value, on rejette significativement l'hypothèse nulle. De même, on en conclut que le type de matière grasse influe sur la quantité de matière grasse absorbée.

Puisque p-value < 0,05, on en déduit que le type de matière grasse influence significativement la quantité de matière grasse absorbée.