2 variables qualitatives ... Test du Chi2 d'indépendance

2 variables quantitatives

modèle $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ Tests Student/ANOVA.

Part V

ANOVA

13 ANOVA à un facteur

Supposons que X soit qualitative, on perd la capacité d'un modèle $\beta_1 X$ ...!

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

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

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.

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 :

X variable réponse

ξ facteur (variable explicative qualitative).

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,

		eff	Moyennes	estimateurs
$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$	n_1	μ_1	$\overline{X_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$	n_2	μ_2	$\overline{X_2} = \frac{1}{n_2} \sum_j X_{2,j}$

ou dans le deuxième cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$	μ_1	$\overline{X_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$	μ_2	$\overline{X_2}$
$\xi = \xi_3$: "traitement expérimental à forte dose"	$X_{3,1}, X_{3,2}, \cdots X_{3,n_3}$	μ_3	$\overline{X_3}$

Modèle : $X_{ij} = \mu_i + \varepsilon_{ij}$ $\Rightarrow$ Moyenne du groupe + erreurs.

$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d

"le facteur ξ n'influence pas" $\Longleftrightarrow$ "les μ_i sont égaux"

X variable réponse continue (quantitative)

ξ facteur (discret qualitatif). $\xi = \xi_1 ; \xi_2 ; \dots ; \xi_a$

De manière générale ;

niveaux de facteur	observations.	effectif	moyenne	Estimateur moyenne
$\xi = \xi_1$	$X_{11}\ X_{12} \dots X_{1n_1}$	n_1	μ_1	$\bar{X}_1 = \frac{1}{n_1} \sum_j X_{1j}$
$\xi = \xi_2$	$X_{2,1}\ X_{22} \dots X_{2n_2}$	n_2	μ_2	$\bar{X}_2 = \frac{1}{n_2} \sum_j X_{2j}$
$\xi = \xi_3$	$X_{3,1}\ X_{32} \dots X_{3n_3}$	n_3	μ_3	
$\vdots$				$\vdots$
$\xi = \xi_a$	$X_{a,1}\ X_{a2} \dots X_{an_a}$	n_a	μ_a	$\bar{X}_a = \frac{1}{n_a} \sum_j X_{aj}$
	global	$n = \sum_i n_i$	μ	$\bar{X} = \frac{1}{n} \sum_{ij} X_{ij}$

$$\bar{X} = \frac{1}{n} \sum_i \underbrace{n_i \bar{X}_i}_{\sum_j X_{ij}}$$

μ_i

On considère le modèle

$$X_i = EX_i + \epsilon_i.$$

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.

Hypothèses $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ i.i.d.

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

Note : $X_{ij} \sim \mathcal{N}(\mu_i, \sigma_\varepsilon^2)$:

$$EX_{ij} = E[\mu_i + \varepsilon_{ij}] = \mu_i + \overbrace{E\varepsilon_{ij}}^{0} = \mu_i$$

$$\mathrm{Var}(X_{ij}) = \mathrm{Var}(\mu_i + \varepsilon_{ij}) = \mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

à estimer (plus rand).

constant car paramètre

Pb : On va tester l'égalité des μ_i, dans ce cas, le facteur n'explique pas X, ou encore, "la totalité des observations correspond à la même population"

13.1 Facteur à deux valeurs - t de Student

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

$$\xi = \xi_1 \quad X_{1,1}, X_{1,2}, \cdots, X_{1,n_1}; \quad n_1 \quad \mu_1 \quad \bar{X}_1 = \frac{1}{n_1}\sum_i X_{1j}$$

$$\xi = \xi_2 \quad X_{2,1}, X_{2,2}, \cdots, X_{2,n_2}, \quad n_2 \quad \mu_2 \quad \bar{X}_2 = \frac{1}{n_2}\sum_i X_{2j}$$

dans chaque groupe, 1 aléa est lié.

connu

Sous H_0, $\bar{X}_1, \bar{X}_2$ aussi.

Modèle :

$$X_{ij} = \underset{\bar{X}_i}{\underset{\uparrow}{\mu_i}} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim \mathcal{N}(0,\sigma) \text{ i.i.d}$$

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma) \text{ indpdtes.}$$

H_0 = "$\mu_1 = \mu_2$".

Estimons l'écart-type de l'erreur $\sigma_\varepsilon = \sigma$, plutôt la variance...

$$S^2 = \frac{1}{\text{nbre d'écarts indpts}} \sum_i (\varepsilon_{ij} - \underset{0}{\bar{\varepsilon}})^2 = \frac{\sum_i \varepsilon_{ij}^2}{n-?} = \frac{\sum_i (X_{ij} - \bar{X}_i)^2}{n-?} = \frac{\sum_j (X_{1j} - \bar{X}_1)^2 + \sum_j (X_{2j} - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

un aléa lié par groupe car on connaît la moyenne sous H_0.

$$s^2 = \frac{\sum_{j=1}^{n_1}(X_{1,j} - \bar{X}_1)^2 + \sum_{j=1}^{n_2}(X_{2,j} - \bar{X}_2)^2}{\underbrace{n_1 + n_2 - 2}_{n-2}}.$$

On note que $(n_1 + n_2 - 2)s^2/\sigma^2$ suit une loi $\chi^2(n_1 + n_2 - 2)$.

Notre but est de tester l'hypothèse

$$\mathcal{H}_0 = \text{“}\mu_1 = \mu_2\text{”},$$

équivalente à “Le facteur n'a pas d'effet sur la variable X”, ou encore “les deux échantillons sont issus de la même population”. Nous allons donc étudier l'estimateur de la différence des moyennes $\bar{X}_1 - \bar{X}_2$, de moyenne nulle par hypothèse nulle, et de variance

Sous $\mathcal{H}_0$, $\mu_1 = \mu_2 = \mu$. $\bar{X}_1 - \bar{X}_2$ est centré sous $\mathcal{H}_0$.

En effet, $\bar{X}_1 = \frac{1}{n_1} \sum_i X_{1j}$ combinaison linéaire de normales indépendantes

donc $\bar{X}_1 \sim \mathcal{N}(m, \sigma^2_{\bar{X}_1})$

$m = E\,\bar{X}_1 = \frac{1}{n_1} \sum_i E(X_{1j}) = \frac{1}{n_1} \sum_i \mu_1 = \frac{n_1 \mu_1}{n_1} = \mu_1 = \mu$ sous $\mathcal{H}_0$.

$\bar{X}_1 \sim \mathcal{N}(\mu, \sigma^2_{\bar{X}_1})$ sous H_0, $\bar{X}_2 \sim \mathcal{N}(\mu, \sigma^2_{\bar{X}_2})$ —

$E(\bar{X}_1 - \bar{X}_2) = 0$.

Car les X_{ij} sont indépendants.

$$Var(\bar{X}_1 - \bar{X}_2) = Var\left(\frac{1}{n_1}\sum_{j=1}^{n_1} X_{1,j} - \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2,j}\right) = \frac{1}{n_1^2} Var\left(\sum_{j=1}^{n_1} X_{1,j}\right) + \frac{1}{n_2^2} Var\left(\sum_{j=1}^{n_2} X_{2,j}\right),$$

$$= \frac{1}{n_1^2} \sum_i \underbrace{Var(X_{1j})}_{\sigma^2} + \frac{1}{n_2^2} \sum_i \underbrace{Var(X_{2j})}_{\sigma^2}$$

par indépendance des échantillons. Il en suit

$$Var(\bar{X}_1 - \bar{X}_2) = \frac{n_1}{n_1^2}\sigma^2 + \frac{n_2}{n_2^2}\sigma^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Sous H_0 :

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1).$$

En estimant σ : par $\sqrt{\dfrac{\sum_i (x_{1j}-\overline{x_1})^2 + \sum_i (x_{2j}-\overline{x_2})^2}{n-2}} = S$

Donc $\dfrac{\overline{x_1}-\overline{x_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n-2}$

Cette quantité sera estimée par

$$s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right).$$

Le test est basé sur la variable

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

On rejettera donc $\mathcal{H}_0$ au seuil α si $|T| \geq q_{1-\alpha/2}^{t(n_1+n_2-2)}$.

Notons que $T \sim \mathcal{F}_{1,n-2}$ . . .

Le numérateur de la statistique T est une mesure de l'écart entre les moyennes échantillonnales, alors qu'au dénominateur figure l'écart type s qui est une mesure de la dispersion à l'intérieur des échantillons. Nous rejetons $\mathcal{H}_0$ lorsque $|T|$ prend une valeur trop grande, c'est-à-dire lorsque l'écart entre les échantillons est trop grand comparé à la dispersion à l'intérieur des échantillons. Nous utiliserons le même principe maintenant dans le cas de plus de deux échantillons.

Gros problème : avec a modalités, comment généraliser

$$\mathcal{H}_0 = \text{"}\mu_1 = \mu_2\text{"} = \text{"}\mu_1 - \mu_2 = 0\text{"}.$$

estimé par $\overline{X}_1 - \overline{X}_2$

reviendrait à $(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2$

$$\Rightarrow \quad \mathcal{H}_0 = \text{"}\mu_1 = \mu_2 = \cdots = \mu_a = \mu\text{"} = \text{"}\sum_i (\mu_i - \mu)^2 = 0\text{"}.$$

$$= \text{"}\sum_i n_i (\mu_i - \mu)^2 = 0\text{"}$$

13.2 Facteur à a modalités

13.2.1 Le modèle

Supposons donc qu'on prélève a échantillons indépendants :

$$
\begin{array}{lcccc}
\xi = \xi_1 & X_{1,1}, X_{1,2}, \quad \cdots \quad , X_{1,n_1}; & \mu_1 & n_1 & \overline{X_1} \\
\xi = \xi_2 & X_{2,1}, X_{2,2}, \quad \cdots \quad , X_{2,n_2}; & \mu_2 & n_2 & \overline{X_2} \\
 & \vdots & \vdots & \vdots & \vdots \\
\xi = \xi_a & X_{a,1}, X_{a,2}, \quad \cdots \quad , X_{a,n_a}; & \mu_a & n_a & \overline{X_a} \\
 & & \mu_{\cdot} & n = \sum_1^a n_i & \overline{X}
\end{array}
$$

<u>Modèle</u> : $X_{ij} = \mu_i + \varepsilon_{ij}$ $\quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ i.i.d.

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma_\varepsilon^2)$$

Estimons σ_ε^2 :

$$S_2 = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}{\underbrace{\sum_i (n_i - 1)}} = \frac{\sum_{ij} (X_{ij} - \bar{X}_{i\cdot})^2}{n-a} = \frac{SCE}{n-a}$$

nombre d'écarts indépendants

↳ en supposant les moyennes connues, un écart par groupe est lié aux autres.

$$\Rightarrow \frac{SCE}{\sigma^2} \sim \chi^2(n-a)$$

13.2.2 Le test de Fisher

L'hypothèse à tester est

$$\mathcal{H}_0 : \text{“}\mu_1 = \mu_2 = \cdots = \mu_a = \mu\text{”} = \text{“}\sum_i n_i (\mu_i - \mu)^2 = 0\text{”}$$

La variable de test "découle" de

$$\underbrace{SCM} = \sum_i n_i (\overline{X_i} - \bar{X})^2 \ldots$$

Somme des carrés due au modèle.

Estimons la variance de $SCM = \sum_{i=1}^{a} n_i(\bar{X}_i - \bar{X})^2$

$\bar{X}_i = \frac{1}{n_i}\sum_j X_{ij}$ $\quad\quad$ $\bar{X} = \frac{1}{n}\sum_i n_i \bar{X}_i$

$\bar{\varepsilon}_i = \sum_j \frac{1}{n_i}(X_{ij} - \bar{X}_i) = \sum_j \frac{1}{n_i} X_{ij} - \frac{n_i \bar{X}}{n_i} = \bar{X}_i - \bar{X}.$

Sous H_0, $\bar{X}_i = \bar{X}$

$\bar{\varepsilon}_i = \frac{1}{n_i}\sum_j \varepsilon_{ij}$

$\bar{\varepsilon}_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{\sqrt{n_i}}\right)$

On observe que

$$\bar{X}_i - \bar{X} = \sum_{j=1}^{n_i} \frac{1}{n_i}(X_{ij} - \bar{X}) = \sum_{j=1}^{n_i} \frac{1}{n_i}\varepsilon_{ij} = \bar{\varepsilon}_i,$$

dont la variance est σ^2/n_i. Par conséquent,

$$\sum_{i=1}^{a} \frac{n_i(\bar{X}_i - \bar{X})^2}{\sigma^2} = \frac{SCM}{\sigma^2} \sim \chi^2(a-1).$$

$$\frac{SCM}{\sigma^2} = \sum_{i=1}^{a}\left(\frac{(X_{ij} - \bar{X})}{\sigma/\sqrt{n_i}}\right)^2 \sim \chi^2(a-1)$$

$\mathcal{N}(0,1)$ indépendantes

$\mu = \frac{1}{n}\sum_i n_i \bar{X}_i$

connu

$\exists$ $a-1$ aléas indpts dans les X_i

On pose

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

Puisque

$$\sum_{i=1}^{a} (n_i - 1) = n - a,$$

on estime σ^2 par $SCE/(n-a)$ et

$$\frac{SCE}{\sigma^2} \sim \chi^2(n-a).$$

$$F = \frac{\left(\frac{SCM}{\sigma^2}\right) / (a-1)}{\left(\frac{SCE}{\sigma^2}\right) / (n-a)} \sim \mathcal{F}_{a-1,n-a}$$

La variable du test est donc

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

⚠ Ce ne sont pas les mêmes degrés de liberté que dans le cadre de la régression linéaire

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.

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

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

13.2.3 Équation de la variance

Posons de plus

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

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

$$SCT = SCM + SCE.$$

Que l'on montrera rigoureusement en TD.

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

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

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

bien lire les D.D.L pour reconnaître la différence entre ANOVA sur facteur et ANOVA sur régression linéaire.

Les résultats d'une analyse de variance sont habituellement présentés sous la forme d'un tableau comme le suivant :

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

+ p value.

Carrés moyens : $\frac{SCM}{a-1}$ $\frac{SCE}{n-a}$ $\frac{SCT}{n-1}$ estimations corrigées des variances

$$F = \frac{SCM/a-1}{SCE/n-a} = \frac{CMM}{CME}$$

facteur	Réponse	Moyennes	Estimations	effectifs
$\xi = \xi_1$	$X_{11}, \dots, X_{1n_1}$	μ_1	$\overline{X}_1$	n_1
$\xi = \xi_2$	$X_{21}, \dots, X_{2n_2}$	μ_2	$\overline{X}_2$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\xi = \xi_a$	$X_{a1}, \dots, X_{an_a}$	μ_a	$\overline{X}_a$	n_a
		μ	$\overline{X} = \frac{\sum_i n_i \overline{X}_i}{n}$	$n = \sum_i n_i$

Modèle : $X_{ij} = \mu_i + \varepsilon_{ij}$ $\qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma)$ i.i.d.

$$SCT = \sum_{ij} (X_{ij} - \overline{X})^2$$

$$SCM = \sum_i n_i (\overline{X_i} - \overline{X})^2$$

$$SCE = \sum_{ij} (X_{ij} - \overline{X_i})^2$$

Équation de la variance

$$SCT = SCM + SCE$$

σ^2 s'estime par $\dfrac{SCE}{n-a}$

$X_{ij} \sim \mathcal{N}(\mu_i, \sigma)$

⚠ pas les mêmes DDL que dans la régression linéaire

ANOVA : H_0 = " le facteur n'a pas d'effet " ⇔ " $\sum_i n_i(\mu_i - \mu)^2 = 0$ "

$$F = \frac{SCM/(a-1)}{SCE/(n-a)}$$

À comparer à $\mathcal{F}_{a-1,\, n-1}$

Pb : Supposons qu'on rejette H_0 ⇒ On peut conclure que le facteur a un effet sur X. Y a-t-il des modalités, des groupements de modalités qui auraient plus d'effet que d'autres ?

$\boxed{\text{Contrast.}}$

13.2.4 En cas de rejet de l'hypothèse d'égalités des moyennes

La table d'analyse de variance nous permet de tester l'hypothèse que les moyennes des populations sont toutes égales. Dans la plupart des cas, le rejet de l'hypothèse soulève de nouvelles questions : si les moyennes ne sont pas toutes égales, où sont les différences ? Nous étudions ici le cas où l'expérimentateur a formulé certaines questions (formulé certaines hypothèses) à priori.

Supposons, par exemple, qu'un expérimentateur veuille comparer trois traitements pour la culture des betteraves :

(i) Un engrais minéral appliqué en avril avant l'ensemencement ;

(ii) Le même engrais appliqué en décembre avant le labourage ;

(iii) Pas de minéraux.

On suppose qu'on rejètte significativement l'égalité des moyennes par l'ANOVA. Le type d'engrais et son application influence le rendement.

Pb : Est-ce la nature de l'engrais qui a un effet ? les 2 premiers sont le même engrais.

la moyenne des 2 premiers groupes est égale à la moyenne du troisième !

Les données portent sur la récolte obtenue dans chacune de ces trois conditions. En supposant que l'hypothèse $\mu_1 = \mu_2 = \mu_3$ sera rejetée, l'expérimentateur sait qu'il voudra ensuite tester l'hypothèse :

$$\boxed{\frac{\mu_1 + \mu_2}{2} = \mu_3,} \iff \frac{\mu_1}{2} + \frac{\mu_2}{2} - \mu_3 = \underbrace{\sum_i \lambda_i \mu_i}_{\varphi} \qquad \begin{cases} \lambda_1 = \frac{1}{2} \\ \lambda_2 = \frac{1}{2} \\ \lambda_3 = -1 \end{cases}.$$

c'est l'hypothèse qu'en moyenne, les minéraux n'ont pas d'effet. Plus généralement, supposons qu'on veuille tester une hypothèse de la forme

$$\mathcal{H}_0 = \text{``}\varphi = \sum_{i=1}^{a} \lambda_i \mu_i = 0\text{''},$$

où λ_i sont des constantes donnés. la fonction linéaire φ sera estimée par

$$\hat{\varphi} = \sum_{i=1}^{n} \lambda_i \bar{X}_i.$$

$\hat{\varphi} = \sum_i \lambda_i \bar{X}_i$ estime $\varphi = \sum_i \lambda_i \mu_i$.

les $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ i.i.d. les $\bar{X}_i = \frac{1}{n_i} \sum_j X_{ij}$ sont des normales en tant que combinaisons linéaires de normales indépendantes.

$$E\bar{X}_i = \frac{1}{n_i} \sum_j EX_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mu_i = \frac{n_i}{n_i} \mu_i$$

et $\operatorname{Var} \bar{X}_i = \operatorname{Var}\left[\frac{1}{n_i} \sum_j X_{ij}\right) = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \operatorname{Var} X_{ij} = \frac{n_i}{n_i^2} \sigma^2 = \frac{\sigma^2}{n_i}$

X_{ij} indépendants

Notons que les $\bar{X}_i$ sont indépendantes et suivent $\mathcal{N}\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right)$

EV pour $\hat{\varphi}$?

$$E\hat{\varphi} = E\left[\sum_i \lambda_i \bar{X}_i\right] = \sum_i \lambda_i E\bar{X}_i = \sum_i \lambda_i \mu_i = \varphi = 0 \text{ sous } H_0$$

$$\operatorname{Var} \hat{\varphi} = \operatorname{Var}\Big(\underbrace{\sum_i \lambda_i \bar{X}_i}_{\text{indépendantes}}\Big) = \sum_i \lambda_i^2 \operatorname{Var}(\bar{X}_i) = \sum_i \frac{\lambda_i^2}{n_i} \sigma^2 = \sigma^2 \sum_i \frac{\lambda_i^2}{n_i}$$

Donc, sous H_0, $\dfrac{\hat{\varphi}}{\sigma\sqrt{\sum_i \frac{\lambda_i^2}{n_i}}} \sim \mathcal{N}(0,1)$

et σ est estimé par $\sqrt{\frac{SCE}{n-a}}$.

$$(n-a)\frac{SCE/n-a}{\sigma^2} \sim \chi^2(n-a)$$

Donc
$$\frac{\dfrac{\hat{\varphi}}{\cancel{\sigma}\sqrt{\sum_i \frac{\lambda_i^2}{n_i}}}}{\sqrt{\cancel{\dfrac{(n-a)}{n-a}}\ \dfrac{SCE/n-a}{\cancel{\sigma^2}}}} \sim t_{n-a}$$

$$\| $$

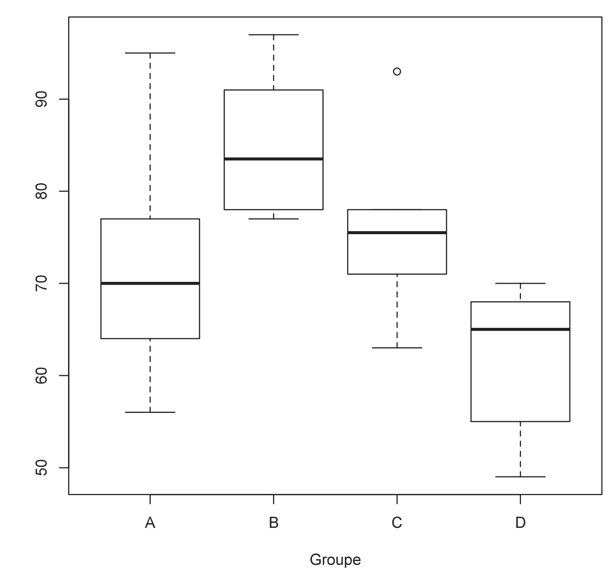
$$\frac{\hat{\varphi}}{\sqrt{\frac{SCE}{n-a}}\sqrt{\sum_i \frac{\lambda_i^2}{n_i}}} \sim t_{n-a}$$

À comparer au quantile théorique de t_{n-a}

13.2.5 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.

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



Données et ANOVA. Poids de matière grasse absorbée par fournée (diminuée de 100 g)

j	**Matière grasse** (indice i) **1**	**2**	**3**	**4**	**Tous**
1	64	78	75	55	
2	72	91	93	66	
3	68	97	78	49	
4	77	82	71	64	
5	56	85	63	70	
6	95	77	76	68	
$\sum_j X_{ij}$	432	510	456	372	1 770
$\bar{X}_i$	72	85	76	62	295
$\sum_j X_{ij}^2$	31 994	43 652	35 144	23 402	134 192
$n_i \bar{X}_i^2$	31 104	43 350	34 656	23 064	132 174
$\sum_j X_{ij}^2 - n_i \bar{X}_i^2$	890	302	488	338	2018
d.l.	5	5	5	5	20

$$\text{Pondéré } s^2 = 2\,018/20 = 100{,}9$$

$$s_{\bar{D}} = \sqrt{2s^2/n} = \sqrt{2 \times 100{,}9/6} = 5{,}80$$

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

SCE
= 2018

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				

DDL de SCE : n − n
24 − 4

H_0 = "égalité de la quantité de graisse absorbée"

= "égalité des moyennes"

= "le type de graisse n'a pas d'effet sur la quantité absorbée"

p < 5%. Rejet significatif de H_0. Le type de graisse influence la quantité absorbée, significativement.

Les deux premières matières grasses sont d'origine animale ; les deux dernières sont d'origine végétale. Nous aimerions tester l'hypothèse que les deux types de matière grasse sont absorbées en moyenne de la même façon.

$\mathcal{H}_0$ = " le type de matière grasse n'influence pas la quantité absorbée ".

Le contraste s'écrit de la manière suivante :

$$\mathcal{H}_0 = \text{"} \underbrace{\frac{\mu_1+\mu_2}{2}}_{\text{moyenne des 2 premières M.G.}} = \underbrace{\frac{\mu_3+\mu_4}{2}}_{\text{moyenne des 2 autres}} \text{"} \iff \mu_1+\mu_2-\mu_3-\mu_4=0.$$

Intuitivement, pas de somme de carré, une combinaison linéaire des moyennes $\Rightarrow$ Student !

Contraste : $\mu_1 + \mu_2 - \mu_3 - \mu_4 = \varphi = \sum_i \lambda_i \mu_i$

avec $\lambda_1 = \lambda_2 = 1$

$\lambda_3 = \lambda_4 = -1$

$\mathcal{H}_0$ = " $\varphi = 0$ ".

Statistique de test : $t = \dfrac{\hat{\varphi}}{\sqrt{\dfrac{SCE}{n-a}}\sqrt{\sum_{i=1}^{a} \dfrac{\lambda_i^2}{n_i}}}$

$\lambda_i^2 = 1$

$n_i = 6$.

$a = 4$

Moyenne
72
85
76
62

$\hat{\varphi} = \bar{X}_1 + \bar{X}_2 - \bar{X}_3 - \bar{X}_4 = 72 + 85 - 76 - 62 = 19$

$$\sqrt{\sum_{i=1}^{a} \frac{\lambda_i^2}{n_i}} = \sqrt{\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}} = \sqrt{\frac{2}{3}} = 0{,}82$$

$$\sqrt{\frac{SCE}{n-a}} = \sqrt{\frac{2018}{24-4}} = \sqrt{\frac{2018}{20}} = \sqrt{100{,}9} = 10$$

Donc $|t| = \dfrac{19}{10 \times 0{,}82} = 2{,}31$

À comparer au quantile de t_{20}

$q = 2{,}086$.

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975
1	0,325	0,727	1,376	3,078	6,314	12,71
2	0,289	0,617	1,061	1,886	2,920	4,303
3	0,277	0,584	0,978	1,638	2,353	3,182
4	0,271	0,569	0,941	1,533	2,132	2,776
5	0,267	0,559	0,920	1,476	2,015	2,571
6	0,265	0,553	0,906	1,440	1,943	2,447
7	0,263	0,549	0,896	1,415	1,895	2,365
8	0,262	0,546	0,889	1,397	1,860	2,306
9	0,261	0,543	0,883	1,383	1,833	2,262
10	0,260	0,542	0,879	1,372	1,812	2,228
11	0,260	0,540	0,876	1,363	1,796	2,201
12	0,259	0,539	0,873	1,356	1,782	2,179
13	0,259	0,538	0,870	1,350	1,771	2,160
14	0,258	0,537	0,868	1,345	1,761	2,145
15	0,258	0,536	0,866	1,341	1,753	2,131
16	0,258	0,535	0,865	1,337	1,746	2,120
17	0,257	0,534	0,863	1,333	1,740	2,110
18	0,257	0,534	0,862	1,330	1,734	2,101
19	0,257	0,533	0,861	1,328	1,729	2,093
20	0,257	0,533	0,860	1,325	1,725	$\boxed{2{,}086}$

$|t| > 9$. Au seuil 5%, on rejette H_0. Les graisses animales sont plus absorbées que les graisses végétales.

$\hat{\varphi} = 14 \Rightarrow \varphi = \underbrace{\mu_1 + \mu_2}_{\text{animales}} - \underbrace{(\mu_3 + \mu_4)}_{\text{végétales}} > 0$.