Analyse de la Variance à un facteur

- permet de généraliser la démarche du test de Student à plus de 2 groupes, et surtout à plusieurs types de facteurs.

  ↗ qualitatifs : ANOVA
  → Quantitatifs : Régression
  ↘ les deux : ANCOVA

  } modèle linéaire

- plus généralement permet l'analyse et la mise au point d'exp. basées sur du matériel exp. variable.

→ Retour sur les médicaments :

. patients : . prennent soit med A, B, C ou 1, 2, 3
  . n_1, n_2, n_3

. patient 1 … n_1 : y_{11} … y_{1n_1}
  patient 1 … n_2 : y_{21} … y_{2n_2}
  patient 1 … n_3 : y_{31} … y_{3n_3}

y_{ij} → j : indice de répétition $j = 1 \dots n_i$

y_{ij} → i : indice du traitement contrôlé ⇓ effet fixe.

y_{ij} = charge virale pour le j^e pat. qui prend le médoc de type i.

- Facteur : médicament
- niveau : 1, 2, 3

Modèle : $Y_{ij} \sim N(\mu_i, \sigma^2)$.

ANOVA. $Y_{ij} = \mu_i + \varepsilon_{ij}$

espérance

erreur aléatoire
- iid
- $N(0, \sigma^2)$, σ^2 cste.

- Si on refait la même démarche que Student :
  $\mathcal{H}_0 = \{ \mu_i = \mu_{i'} \}$

  ↳ Répond à la question : les types i et i' ont-ils le même effet ?

  ↳ Or on peut poser 2 types de questions :
  - effet O/N
  - lequel ? a le + fort effet.

La Reparamétrisation : $\mu_i = \mu + \alpha_i$.

effet général

effet du médicament de type i.

- on rajoute un paramètre
  - ↗ On gagne en interprétation
  - → On perdre qqchose : μ et α_i n'ont pas vraiment de sens en eux même, ce n'est que $\mu + \alpha_i$ qui en a un.

- Paramètres du modèle : de 2 types :

$$\begin{cases} \theta = (\mu, \alpha_1, \alpha_2, \alpha_3) = \text{paramètres de l'espérance} \\ \sigma^2 = \text{variance résiduelle.} \end{cases}$$

- Notation matricielle $Y = X\theta + E$
- Notation Y_{it}, $Y_{i\circ}$

- La Table d'Analyse de la Variance.

⚠ Important de distinguer un facteur de ses niveaux,

- facteur = source de variabilité
- niveau = modalités du facteur.

- Idée de l'Analyse de la Variance : étudier les différentes sources de variabilité.

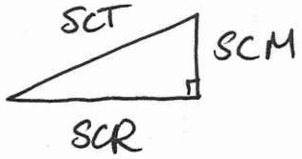
- $SCT = \sum_{ij} (Y_{ij} - Y_{\circ\circ})^2$: Variance totale
  : comme si il ∄ de groupe

- $SCM = \sum_{ij} (Y_{i\circ} - Y_{\circ\circ})^2$: si on resume l'info par la moyenne du groupe.
  - Variance expliquée par le modèle.

  Between Var. — moy. pour niveau i — moy. générale

- $SCR = \sum_{ij} (Y_{ij} - Y_{i\circ})^2$: somme des carrés résiduelle.
  - Within Variance.

Représentation graphique.



Source	df	SS	MS	
Modèle	$I-1$	SCM	$SCM/I-1$	$\sigma^2 \chi^2(I-1)$
Resid	$n-I$	SCR	$SCR/n-I$	$\sigma^2 \chi^2(n-I)$
Total	$n-1$	SCT	$SCT/n-1$	$\sigma^2 \chi^2(n-1)$

- 1ère Question : la variabilité expliquée par le modèle est-elle plus grande que la résiduelle ?

$$\frac{SCM/(I-1)}{SCR/(n-I)} \sim F_{I-1,n-I}.$$

↳ Si ce test n'est pas significatif, alors ce n'est pas la peine de regarder la suite.

⇒ Ce correspond à tester l'hypothèse nulle $H_0 = \{\mu_1 = \dots = \mu_I\}$.

- 2e Question : la part de variabilité expliquée par le modèle.

$$R^2 = \frac{SCM}{SCT}$$ = pouvoir explicatif / prédictif.

→ Interprétation.

. Estimation des paramètres.

$$S(\mu, \alpha) = \sum_{ij} \left(Y_{ij} - (\mu + \alpha_i)\right)^2$$

$$\begin{cases} \dfrac{\partial S}{\partial \mu} = 0 \\ \dfrac{\partial S}{\partial \alpha} = 0 \end{cases} \qquad \begin{cases} Y_{++} - n\hat{\mu} - \sum_i n_i \alpha_i = 0 \\ Y_{i+} - n_i \hat{\mu} - n_i \hat{\alpha}_i = 0 . \end{cases}$$

→ Il faut mettre une contrainte pour inverser le système.

→ $\alpha_I = 0$: . Contrainte "utile" d'un point de vue numérique.

$$\hat{\mu} = Y_{I\bullet} \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{I\bullet} .$$

. les estimateurs s'interprètent comme les écarts à un groupe de ref qui est choisi arbitrairement ç le dernier.

→ $\sum \alpha_i = 0$: . elle suppose que les effets se compensent globalement !

. si n_i = cste, dispositif équilibré.

$$\hat{\mu} = Y_{\bullet\bullet} \qquad \hat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}$$

$\hat{\mu}$ ↓ moyenne générale

$\hat{\alpha}_i$ ↓ écart à la moyenne générale

- si déséquilibre.

$$\hat{\mu} = \frac{1}{I} \sum Y_{i\cdot} \qquad \hat{\alpha}_i = Y_{i\cdot} - \hat{\mu}$$

difficile à interpréter.

→ $\sum n_i \alpha_i = 0$ $\quad \hat{\mu} = Y_{\cdot\cdot} \quad , \quad \hat{\alpha}_i = Y_{i\cdot} - Y_{\cdot\cdot}$.

↳ redonne les estimateurs naturels.

↳ les valeurs des estimateurs des paramètres dépendent des contraintes

⇒ mais certaines comb. lin. sont invariantes = CL Est.

↳ De manière générale, $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$ = prédict°, invariants

↳ rassurant !..

. autre exemple : $\hat{\alpha}_i - \hat{\alpha}_{i'} = Y_{i\cdot} - Y_{i'\cdot}$

↳ la comparaison des groups est ⊥ de la contrainte.

⇒ faire des tests sur les paramètres (α_i) dependra des contraintes, alors que faire de tests sur $\alpha_i - \alpha_{i'}$ non.

paramètre	H_0 $\mu+\alpha_i$	H_0 μ_i.
μ	$\mu=0$	$\mu_3=0$
α_1	$\alpha_1=0$	$\mu_1=\mu_3$
α_2	$\alpha_2=0$	$\mu_2=\mu_3$
α_3.	$\alpha_3=0$	$0=0$.

avec $(\alpha_3=0)$.

⇒ le sens de l'hypothèse dépend de la contrainte dans le modèle $\mu+\alpha_i$ alors qu'il n'en dépend pas dans μ_i.

⇒ ⚠ Faire un test sur les paramètres du modèle peut être dangereux.

↳ Mais en général, ce qui nous intéresse c'est la comparaison de groupes.

Comparaison de groupes :

→ Student : $T = \dfrac{Y_{i\cdot} - Y_{i'\cdot}}{\sigma\sqrt{\frac{1}{n_i}+\frac{1}{n_{i'}}}} \sim \mathcal{T}(n_i+n_{i'}-2)$

la puissance augmente avec $n_i+n_{i'}-2$.

→ dans l'Anova, on augmente la puissance en utilisant une estimation de la variance fondée sur l'ensemble des groupes.

↳ uniqt si $\sigma^2 = \text{cste}$ $T \sim \mathcal{T}(n-I)$.

. les Comparaisons Multiples .

. Si on compare I Groupes $\rightarrow$ $I(I-1)/2$ Tests

Bonf : $\alpha^* \leq \alpha \times I(I-1)/2$.

. Analyse des résidus .