LES ANALYSES DE VARIANCE (ANOVA)

COMPARAISON DE PLUSIEURS MOYENNES OBSERVÉES

Benchikh Tawfik

Faculté de Médecine, UDL, SBA

1ère année Médecine

20 Mars 2024

PLAN DU COURS

Objectif de la ANOVA

- La problématique de l'**ANOVA** consiste à utiliser les moyennes observées sur les échantillons pour conclure à des différences significatives sur les moyennes dans les sous-populations.
- Il s'agit d'un test statistique permettant de comparer les moyennes de plusieurs variables aléatoires **indépendantes gaussiennes de même variance**.

OBJECTIF DE L'ANOVA

- L'objectif de **ANOVA:** étudier l'effet des variables qualitatives sur une variable quantitative.
- On applique l'ANOVA (**des modèles factoriels**) quand on dispose:
  - d'une variable quantitative à expliquer,
  - d'une ou de plusieurs variables qualitatives explicatives, appelées **facteurs**.

Exemple

- 21 candidats, 3 examinateurs (resp. 6,8 et 7 étudiants)

Examinateur	A	B	C
Notes	$10, 11, 11$ $12, 13, 15$	$8, 11, 11, 13$ $14, 15, 15, 16$	$10, 13, 14, 14$ $15, 16, 16$
Effectis	6	8	7
Moyenne	12	13	14

- Quelles est l'"effet d'examinateur" sur les notes des étudiants ?

TERMINOLOGIE

1. **facteur** (variable qualitative): prend un nombre fini de modalités (une classe). Il est totalement contrôlé(fixées par l'expérimentateur).
   - Exemple: facteur "examinateur".
2. **niveau**: les différentes valeurs prises par un facteur (les modalités).
   - Exemple: niveaux A, B, C.
3. **test de l'effet d'un facteur**: tester si les moyennes des populations sont égales.
4. La variable étudiée: Y, à valeurs numériques. Nous l'appelons la réponse (response). Dans l'exemple: $Y =$(**Note).**

Notations et les Données

1. Pour les observations nous utilisons deux indices:
   - le premier indice indique le numéro du groupe dans la population (exemple: "Examinateur"),
   - le second indice indique le numéro de l'observation dans l'échantillon.
2. Pour le premier indice, nous utilisons i (ou encore i',i'',i_1,i_2).
3. Pour le second indice, nous utilisons j (ou encore j',j'',j_1,j_2).

NOTATIONS ET LES DONNÉES

1. Ainsi les observations sont en général notées par:

$$y_{ij}\ ;\ i = 1, \ldots, k \quad \text{et } j = 1, \ldots, n_i$$

   - où i est l'indice du groupe (où de l'échantillon) définie par le facteur explicatif (niveau), et $I = \{i = 1, \ldots, k\}$ (le nombre d'échantillons),
   - n_i le nombre d'expériences dans le groupe i (taille des échantillons).

2. **Définition:** Lorsque les échantillons sont de même taille, nous disons que l'expérience est équilibrée.

3. Si les tailles des échantillons sont différentes, alors elles sont notées comme précedemment par: n_i,où $i = 1, \ldots, k$.

Notations et les Données: résumé

1. Un seul facteur F
2. k niveaux
3. k échantillons de tailles respectives $n_1, \ldots, n_k$.
4. Effectif total $n = \sum_{i=1}^{k} n_i$.
5. A chaque expérience, on mesure la valeur de la variable Y.

DONNÉES

1. Données sous forme d'un tableau:

Niveau (Population)	Nombre d'observation (Effectif)	Valeurs de Y
1	n_1	$y_{11}, y_{12}, \ldots, y_{1n_1}$
2	n_2	$y_{21}, y_{22}, \ldots, y_{2n_2}$
$\vdots$	$\vdots$	$\ldots$
k	n_k	$y_{k1}, y_{k2}, \ldots, y_{kn_k}$

Les Données

1. Moyennes empiriques (moyenne dans chaque classe):
   - Pour niveaux i : $\overline{Y_{i.}} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$.
2. Moyenne globale
   - $Y_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$.
   - $\overline{Y_{..}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$.

Conditions de test

1. les k échantillons sont indépendants.
2. Les y_{ij} sont des réalisations de la v.a. $Y_{ij} \rightsquigarrow \mathcal{N}(m_i, \sigma^2)$.
3. Y_{ij} et Y_{ts} indépendantes pour $i \neq t$ et $j \neq s$.
4. L'écart-type (théorique) est le même pour tous les niveaux.
5. La moyenne (théorique) peut varier avec le niveau.
6. On veut savoir si les moyennes m_i sont toutes égales ou non.

ESTIMATION DES PARAMÈTRES

- Sous l'hypothèse de normalité et d'indépendance des échantillons,
  1. $\overline{Y_{i.}}$ est un estimateur sans biais de m_i et

$$\hat{m}_i = \overline{Y_{i.}} \rightsquigarrow \mathcal{N}(m_i, \frac{\sigma^2}{n_i}) .$$

  2. L'estimateur de σ^2 est:

$$S_n'^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i.}})^2 .$$

TEST DE COMPARAISON DES MOYENNES

- L'ANOVA consiste à construire le test d'hypothèse suivant:

$$\begin{cases} H_0 : & \text{toutes les moyennes sont identiques} \\ H_1 : & \text{au moins une des moyennes est différente des autres .} \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0 : & m_1 = m_2 = \ldots = m_k = m \\ H_1 : & \exists\, i\,, j \in \{1, \ldots, k\} \quad \text{tels que } m_i \neq m_j\,. \end{cases}$$

TEST DE COMPARAISON DES MOYENNES

- La variabilité totale est: $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{..}})^2$.
- On peut écrire: $Y_{ij}-\overline{Y_{..}}=(Y_{ij}-\overline{Y_{i.}})+(\overline{Y_{i.}}-\overline{Y_{..}})$.
- et on obtient:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{..}})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y_{i.}})^2+\sum_{i=1}^{k}n_i(\overline{Y_{i.}}-\overline{Y_{..}})^2$$

- Variabilité totale = variabilité résiduelle + variabilité due au modèle: $SST=SSR+SSL$.
- SSL est la somme des carrés inter-groupes et SSR est la somme des carrés intra-groupes.

TEST DE COMPARAISON DES MOYENNES: CALCULS

- On obtient:

$$SSL = SCR = \sum_{i=1}^{k} n_i(\overline{Y_{i.}})^2 - n(\overline{Y_{..}})^2 \,,$$

$$SSR = SCF = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij})^2 - \sum_{i=1}^{k} n_i(\overline{Y_{i.}})^2 \,,$$

TEST DE COMPARAISON DES MOYENNES

- Pour tester l'hypothèse H_0 on utilise la statistique:

$$F = \frac{SSL/(k-1)}{SSR/(n-k)} \sim F(k-1, n-k) \quad (\text{sous } H_0)$$

est une réalisation d'une variable aléatoire F qui suit une loi de Fisher à $(k-1)$ degrés de liberté au numérateur et $(n-k)$ degrés de liberté au dénominateur.

- Pour un risque α fixé,la zone d'acceptation est: $[0, f_{(k-1,n-k,1-\alpha)}]$.

TABLEAU D'ANALYSE DE VARIANCE: LOGICIEL

- Le tableau de variation donne un résumé des calculs effectués pour l'analyse de variance.

Source de variation	Dégres de liberté	Somme des carrés	Carrés moyens	F	$p-value$
Expliqué (facteur)	$k-1$	$SLL(SCF)$	$SLL/(k-1) = CMF$	$F = CMF/CMR$	
Résidus	$n-k$	$SSR(SCR)$	$SSR/(n-k) = CMR$		
Total	$n-1$	$SST(SCT)$			

TEST DE COMPARAISON DES MOYENNES: EXEMPLE

- $Y_{ij} = m_i + \varepsilon_{ij}$ avec $i = 1;2;3$; $j = 1;\ldots;n_i$; $n_1 = 6$, $n_2 = 8$, $n_3 = 7$.
- H_0 : pas d'effet examinateur sur la notation.
- $H_0 : m_1 = m_2 = m_3 = m$ contre $H_1 : \exists i \neq j$ tel que $m_i \neq m_j$.
- On obtient $SSL = 12.95$ et $SSR = 98$.
- Donc $f_{cal} = \frac{SSL/(3-1)}{SSR/(21-3)} = 1.19$.
- La zone d'acceptation est $[0, f_{(k-1,n-k,1-\alpha)}] = [0, 3.55]$: $f_{th} = 3.55$.
- Donc H_0 est accepter: les examinateurs ont le même système de notation.

TEST DE COMPARAISON DES MOYENNES: REMARQUE

- Le rejet de l'hypothèse d'égalité des moyennes ne signifie pas que tous les m_i sont différents entre eux.
- On cherche souvent à tester l'égalité entre deux moyennes:

$$H_0 : \ m_h = m_j \qquad \text{contre} \quad H_1 : \ m_h \neq m_j \, .$$

- On utilise la statistique de test:

$$T = \frac{|\overline{Y_{h.}} - \overline{Y_{j.}}|}{\sqrt{\frac{SSR}{n-k}}\sqrt{\frac{1}{n_h} + \frac{1}{n_j}}}$$

(t_{n-k}: loi de Student à $n-k$ degrés de liberté.)

- La zone d'acceptation $[-t_{n-k;1-\alpha/2}, t_{n-k;1-\alpha/2}]$.

ANOVA: Exercice

- On veut étudier l'effet de deux médicaments sur le taux de lymphocytes d'animaux de laboratoires.
- On construit un plan factoriel dans lequel il y a trois groupes d'animaux d'effctifs 10 animaux par groupe.
- On garde un des groupes comme témoin et l'on administre les médicaments A et B aux deux autres groupes.

Groupe témoin	$272\ ,193\ ,432\ 259; 386; 349; 320, 247; 260; 478$
Groupe traité par A	$468\ ,383\ ,375\ ,398, 534; 451; 474; 278, 255; 528$
Groupe traité par B	$368\ ,290\ ,325\ ,298, 314; 350; 378; 321, 275; 401$

Exercice

- Les données correspondent au modèle d'ANOVA: une variable de groupe, une variable continue dont on veut comparer les moyennes.
- Indications numériques : $\sum_j x_{1,j} = 3196$, $\sum_j x_{2,j} = 4094$, $\sum_j x_{3,j} = 3320$ (somme de chaque ligne).

EXERCICE

- La taille globale des 3 échantillon est:

  (A) 10 (B) 20 $(\mathbf{C})$ **30** (D) 40 (E) 60.

- La moyenne globale est :

  (A)256.67 $(\mathbf{B})$**353.34** (C)415.33 (D)435.96 (E)563.75.

- la variabilité expliqué SSL est:

  $(\mathbf{A})$**47361.9** (B)51426.85 (C)54211.17 (D)62516.54 (E)65785.76.

EXERCICE

- Quelle sont les degrés de liberté:

  $(A)\ (2, 30)$ $(\mathbf{B})\ (\mathbf{2, 27})$ $(C)\ (3, 27)$ $(D)\ (3, 30)$ $(E)\ (29, 2)$.

- Quelle est la valeurs de la statistique calculée F sachant que

  $SSR = 176130$:$(\mathbf{A})\ \mathbf{3.63}$ $(B)\ 2.42$ $(C)\ 4.003$ $(D)\ 2.689$ $(E)\ 6.84$

- la valeurs de la statistique théorique F:(niveau de confiance $= 95\%$)

  $(A)\ 2.96$ $(B)\ 2.922$ $(\mathbf{C})\ \mathbf{3.354}$ $(D)\ 4.61$ $(E)\ 12.59$

- A quoi correspond le risque alpha ? (A.) à la probabilité de conclure à une différence significative. (B.) à la probabilité de conclure à tort à une absence de différence significative. (C.) à la probabilité de ne pas conclure H1 alors que H1 est vraie. (D.) à la probabilité d'accepter H0 alors que H0 est vraie. **(E.) à la probabilité de rejeter H0 alors que H0 est vraie.**