Analyse de variance ANOVA

ANOVA

📖 **Analyse de variance ANOVA**

Il ne s'agit pas d'un test des variances mais un test sur les moyennes !

- **Objectif**
  - ✓ Comparer les moyennes de k séries statistiques (échantillons) indépendantes qui diffèrent selon les un ou plusieurs facteurs.
  - ✓ La comparaison porte sur une variable quantitative.
  - ✓ Si le test est significatif, les moyennes des k échantillons diffèrent globalement, sans précision sur l'origine de la différence
  - ✓ Il faut ensuite effectuer des comparaisons multiples de moyennes fondées sur le résultat de l'ANOVA.

ANOVA

- **Exemple**

Mettre en évidence l'effet d'un médicament sur la taille d'un fibrome

Médicament	**A**	**B**	**C**
Taille du fibrome	7.5	9.2	13.5
	5.2	10.3	12.1
	4.1	8.7	15.1
	6	14.4	-
	5.4	-	-
Effectif	5	3	4

L'effet du médicament ? ⇒ ANOVA pour étudier l'effet des variables qualitatives sur une variable quantitative !

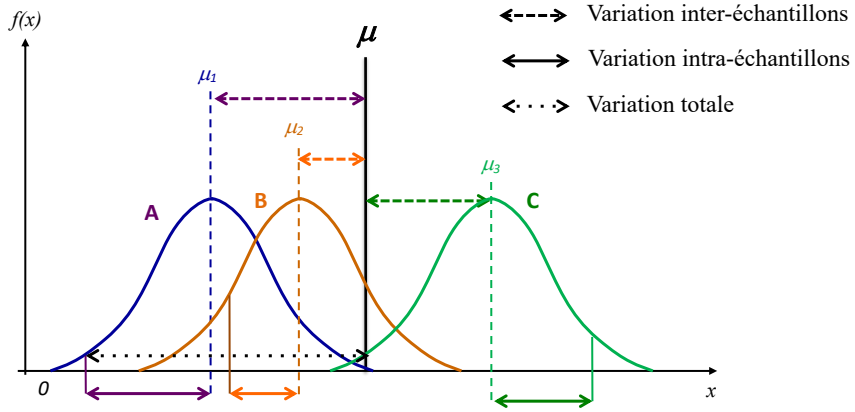
- **Terminologie**
  - *Facteur (variable qualitative)* : prend un nombre fini de valeurs. Ex. type du médicament
  - *Niveau (modalités)* : les valeurs prises par un facteur. Ex : niveaux A, B, C
  - *Test de l'effet d'un facteur* : tester si les moyennes des *modalité* sont égales
  - *La variable étudiée* : Y , a valeurs numériques (Taille du fibrome)

2

2

ANOVA

- **Principe mathématique de l'analyse de variance**



Une variation *inter-échantillons*, due aux écarts entre les moyennes de chaque échantillon et la moyenne générale, *et qui traduit l'effet du facteur* : **Variation factorielle**

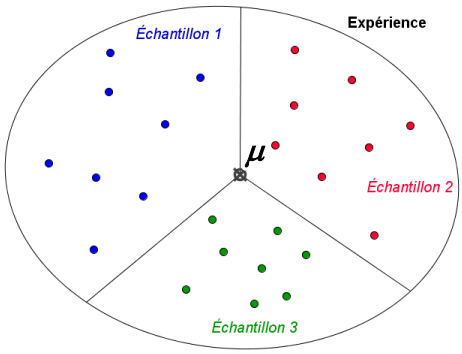
Une variation *intra-échantillon* qui cumule les écarts de chaque valeur individuelle de la variable à leur moyenne d'échantillon. Cette dispersion provient des *fluctuations aléatoires d'échantillonnage* : **Variation résiduelle**

3

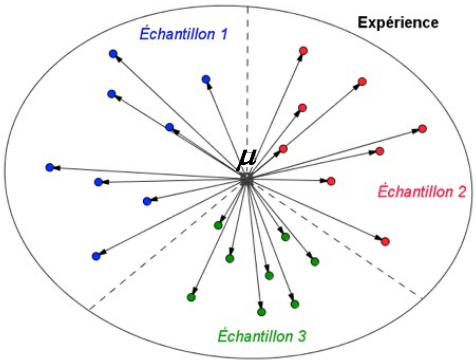
3

ANOVA

- **Schématisation**



Données initiales



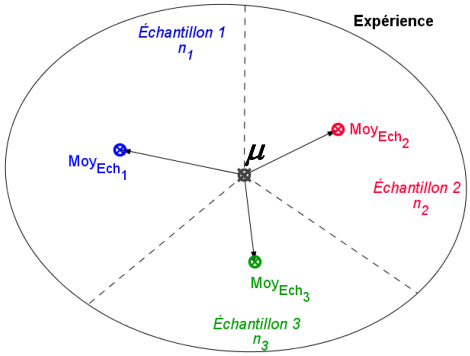
Variabilité totale

- ✓ Soit une Expérience avec k échantillons de n_i individus
- ✓ Le nombre total d'individus est $N = \Sigma\, n_i$
- ✓ μ est la moyenne générale
- ✓ Variabilité totale : les écarts de tous les individus par rapport à la moyenne générale (μ)
- ✓ Calcul de la Somme des Carrés des Écarts à la moyenne totale (SCE_T)
- ✓ Degrés de liberté (DDL) associés : n -1

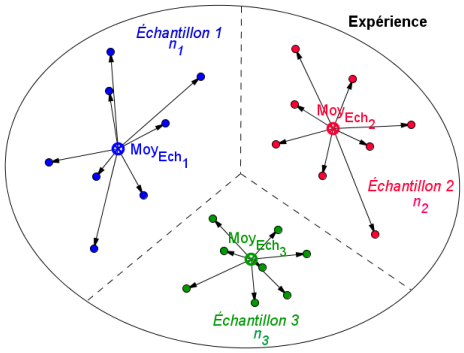
4

4

ANOVA



Variabilité factorielle



Variabilité résiduelle

- ✓ Variabilité factorielle : les écarts des moyennes des échantillons (μ_k) par rapport à la moyenne générale (μ)
- ✓ Calcul de la Somme des Carrés des Écarts à la moyenne factorielle (SCE_F)
- ✓ DDL associés : k-1.
- ✓ Variabilité résiduelle : les écarts des n_i individus par rapport à aux moyennes des échantillons (μ_k) correspondants
- ✓ Calcul de la Somme des Carrés des Écarts à la moyenne résiduelle (SCE_R).
- ✓ DDL associés : n-k.

5

5

ANOVA

- **Hypothèses**

H_0 : Toutes les moyennes sont identiques

H_1 : Au moins une des moyennes est différente des autres

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu \\ H_1 : \exists j, \mu_j \neq \mu \end{cases}$$

- **Le risque de première espèce ou risque α**

C'est le risque de rejeter l'hypothèse H_0 alors que celle-ci est vraie.

Un risque de 5% est classiquement utilisée (plus raisonnable) ; dans certains cas 1%.

Au risque 5 % on estime que la probabilité pour que la différence observée soit due aux fluctuations d'échantillonnage est suffisamment faible pour accepter H_0.

6

6

ANOVA

- ***Les écart à la moyenne peuvent s'écrire de la manière suivante :***

$$x_{ij} - \bar{x} = \left(\bar{x}_j - \bar{x}\right) + \left(x_{ij} - \bar{x}_j\right)$$

Écart à la moyenne globale | **Écart entre Les groupes** | **Écart à l'intérieur des groupes**

- ***En passant au carré et en faisant les Σ, on obtient l'équation d'analyse de la variance :***

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}\right)^2 = \sum_{j=1}^{k}\left(\bar{x}_j - \bar{x}\right)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}_j\right)^2$$

SCE_T (Somme carrés des écarts totaux) Exprime la variabilité totale des observations — **Totale**

SCE_F (Somme carrés des écarts factoriels) Exprime la variabilité expliquées par les facteurs — **Intergroupe**

SCE_R (Somme carrés des écarts résiduels) Exprime la variabilité non expliquées par les facteurs — **Intragroupe**

7

7

ANOVA

- **Degrés de liberté (ddl ou DF en anglais)**
  - ✓ Totale : $ddl_T = n - 1$
  - ✓ Résiduelle : $ddl_R = n - k$
  - ✓ Factorielle : $ddl_F = k - 1$

- **Carrés moyens**

$$Carré\ moyen\ (CM) = \frac{Somme\ carés\ des\ écarts\ (SCE)}{Degréde\ liberté\ (DDL)}$$

$$CM_T = \frac{SCE_T}{ddl_T} \quad ; \quad CM_F = \frac{SCE_F}{ddl_F} \quad ; \quad CM_R = \frac{SCE_R}{ddl_R}$$

- **Statistique du test et prise de décision**

$$F = \frac{CM_F}{CM_R}$$

Sous H_0, la variable aléatoire ***F*** suit une loi de ***Snédécor*** (loi de *Fischer*) à $\nu_1 = k - 1$ et $\nu_2 = n - k$ degrés de liberté

8

8

ANOVA

- **Calculs**

✓ ***Pour chaque échantillon k de taille n_i, on calcule :***

- Moyenne

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$

✓ ***Pour l'ensemble de l'expérience :***

- Moyenne générale

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{i=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{k} n_i \, \bar{x}_i$$

$$avec \quad n = \sum_{i=1}^{k} n_i$$

✓ **Variance totale**

$$SCE_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x} \right)^2$$

✓ **Variance factorielle (variance intergroupe) :**

Dispersion des valeurs d'un échantillon à l'autre (influence du facteur)

$$SCE_F = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{x} \right)^2$$

✓ **Variance résiduelle (variance intragroupe) :**

Dispersion des valeurs à l'intérieur des échantillons (variabilité individuelle)

$$SCE_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_i \right)^2$$

9

9

ANOVA

Le tableau d'analyse de la variance est alors

Source de variation	Somme carré des écarts (SCE)	Degré de liberté (ddl)	Carré moyen (CM)	*F*	*P*-value
Factorielle	SCE_F	$k-1$	CM_F	CM_F/CM_R	*Logiciel*
Résiduelle	SCE_R	$n-k$	CM_R		
Totale	SCE_T	$n-1$			

On lit dans la table de *Fisher-Snédécor* la valeur $f_{1-\alpha}$

Exemple : Pour un $F_{6,4}$: on lit dans la table α = 5% la valeur du croisement de la ligne $\nu_1 = ddl_f = 6$ et de la colonne $\nu_2 = ddl_R = 4$

ν_2 \ ν_1	1	2	3	4	5	6	8	10	15	20	30	∞
1	161	200	216	225	230	234	239	242	246	248	250	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64	3,51	3,44	3,38	3,23

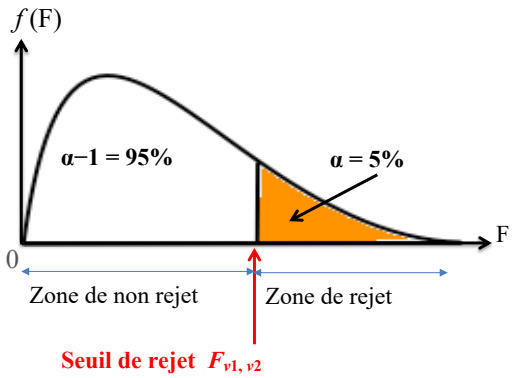
Donc : $F_{6,4} = 6.16$

10

10

ANOVA

Zones de rejet de l'hypothèse nulle pour une distribution de *Snédécor* et un test unilatéral



✓ **Si $F_{calculé} > F_{seuil\ critique}$ ➔ on rejette l'hypothèse H_0 on garde H_1**

✓ **Si $F_{calculé} < F_{seuil\ critique}$ ➔ on garde l'hypothèse H_0**

11

11

ANOVA

- **Exemple** A partir de ce tableau, déterminer si l'effet des trois facteurs est différent avec un risque α de 5%.

F 1	F 2	F 3
5	7	7
4	5	8
3	6	9

Les calcules

- Moyenne pour chaque facteur : $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} \quad \Rightarrow \quad \bar{x}_1 = 4\,; \quad \bar{x}_2 = 6\,; \quad \bar{x}_3 = 8$

- Moyenne générale : $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_j\,\bar{x}_j \quad \Rightarrow \quad \bar{x} = \frac{(3\times 4)+(3\times 6)+(3\times 8)}{9} = 6$

- Variance totale : $SCE_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2$

$$SCE_T = \left[(5-6)^2+(5-6)^2+(5-6)^2\right]+\left[(7-6)^2+(5-6)^2+(6-6)^2\right]+\left[(7-6)^2+(8-6)^2+(9-6)^2\right]$$

$$SCE_T = 30$$

12

12

ANOVA

- Variance factorielle (variance intergroupe) :

$$SCE_F = \sum_{j=1}^{k} n_j\left(\bar{x}_j-\bar{x}\right)^2$$

$$SCE_F = \left[3\times(4-6)^2\right]+\left[3\times(6-6)^2\right]+\left[3\times(8-6)^2\right] = 24$$

- Variance résiduelle (variance intragroupe) :

$$SCE_R = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij}-\bar{x}_i\right)^2$$

$$SCE_R = \left[(5-4)^2+(5-4)^2+(5-4)^2\right]+\left[(7-6)^2+(5-6)^2+(6-6)^2\right]+\left[(7-8)^2+(8-8)^2+(9-8)^2\right]$$

$$SCE_R = 6$$

- Degrés de liberté *ddl* :

Totale	: ddl_T	=	n − 1	= 9−1= **8**
Résiduelle	: ddl_R	=	n − k	= 9−3= **6**
Factorielle	: ddl_F	=	k − 1	= 3−1= **2**

13

13

ANOVA

- Carrés moyens CM :

$$CM_T = \frac{SCE_T}{ddl_T} = \frac{30}{8} = 3.75$$

$$CM_F = \frac{SCE_F}{ddl_F} = \frac{24}{2} = 12$$

$$CM_R = \frac{SCE_R}{ddl_R} = \frac{6}{6} = 1$$

- Calcul de la statistique ***F***:

$$F = \frac{CM_F}{CM_R} = \frac{12}{1} = 12$$

- Le tableau d'analyse de la variance :

Source de variation	Somme carré des écarts (SCE)	Degré de liberté (ddl)	Carré moyen (CM)	*F*	*P*-value
Factorielle	**24**	**2**	**12**	**12**	**0.0...**
Résiduelle	**6**	**6**	**1**		
Totale	**30**	**8**			

14

ANOVA

- Lire dans la table à α=5% :

$F_{\nu 1, \nu 2}$ (ν_1= **2** et ν_2= **6**)

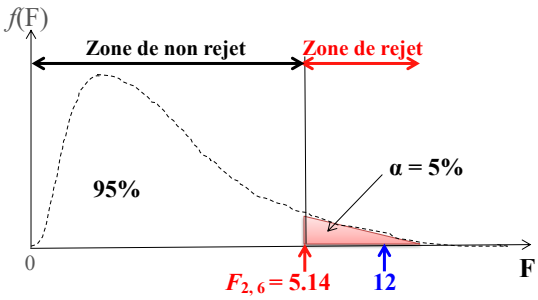
Donc : $F_{2,6} = 5.14$

ν_2 \ ν_1	1	2	3	4
1	161	200	216	225
2	18,5	19,0	19,2	19,2
3	10,1	9,55	9,28	9,12
4	7,71	6,94	6,59	6,39
5	6,61	5,79	5,41	5,19
6	5,99	5,14	4,76	4,53
7	5,59	4,74	4,35	4,12

- Prise de décision :

12 > 5.14 ➔ F $_{calculé}$ > F $_{seuil\ critique}$

Donc on rejet l'hypothèse H_0



➔ **Conclusion :**

L'effet des trois facteurs est différent

15