L'analyse de la variance
ANOVA

L'analyse de la variance est un test statistique qui permet de déterminer si un ou plusieurs facteurs influent sur la réponse observée sur un échantillon. Nous verrons les deux cas suivants :

- ***ANOVA à un facteur***
- ***ANOVA à 2 facteurs (avec répétitions des mesures)***

FACTEUR = variable qualitative

REPONSE = variable quantitative

Exemples

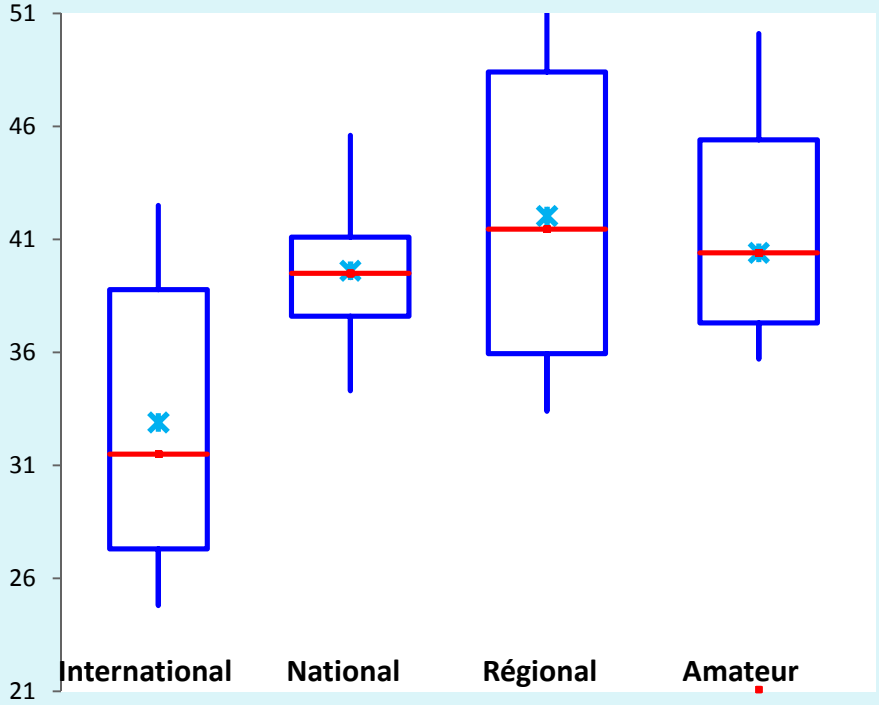
- Y-a-t'il un lien entre le niveau d'un élève au bac et la CSP de ses parents?
- La puissance et la provenance d'une voiture ont-elles un impact sur sa consommation?

ANOVA à un facteur

Dans le cas où il n'y a qu'un seul facteur cela revient à faire un croisement quantitatif-qualitatif.

Exemple : Y-a-t'il un lien entre le niveau d'anxiété d'un joueur de tennis et le niveau de la compétition?

International	National	Régional	Amateur
24,8	45,6	33,4	21,1
26,7	41,1	34,6	35,7
27,5	34,3	36,4	37,3
30,6	37,6	39,1	39,4
…	…	…	…
38,2		47,9	44,5
40,5		49,9	45,4
42,5		51,2	49,8
			50,1



ANOVA à un facteur

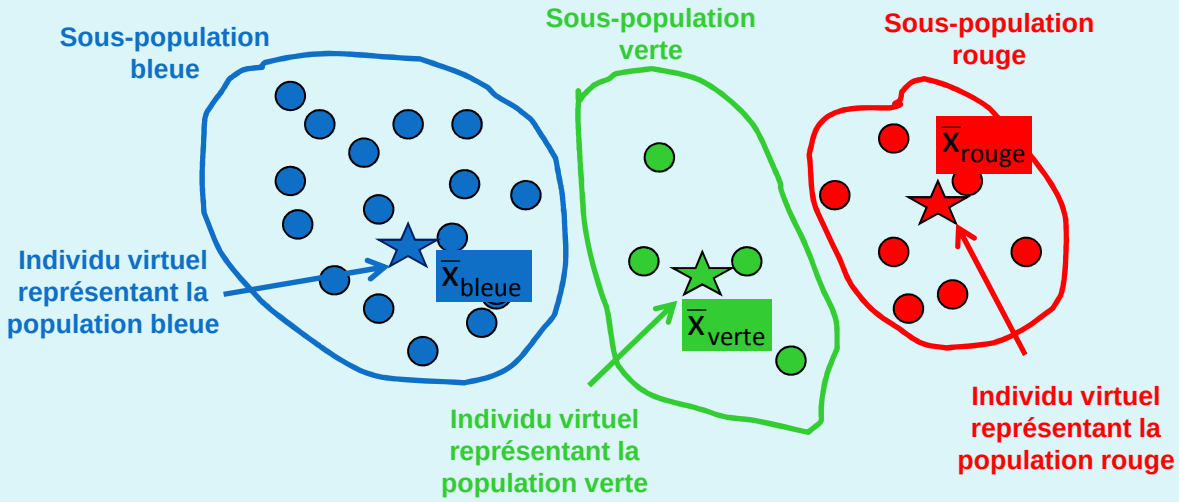
Notons A le facteur supposé influent et A_i ses modalités
(compétition : 4 niveaux)

	Modalités			
	A_1	A_2		A_p
	x_1^1	x_2^1		x_p^1
			…	
	$x_1^{n_1}$	$x_2^{n_2}$		$x_p^{n_p}$
Moyennes	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_p$
Effectifs	n_1	n_2		n_p

On partitionne la population en sous-populations :

➢ **une sous-population pour chaque modalité du facteur**

- Pour chaque sous-population déterminée par les modalités du facteur, on crée un individu virtuel dont la valeur est égale à la moyenne des valeurs des individus de la sous-population.
- On crée donc une nouvelle population formée de ces individus virtuels. Chaque individu aura un poids de n_k, l'effectif de chaque sous-population. On parle de la variation expliquée par le facteur *(variation inter)*



Variance expliquée

$$var_A = \frac{1}{n}\sum_{k=1}^{p} n_k (\bar{x}_k - \bar{x})^2$$

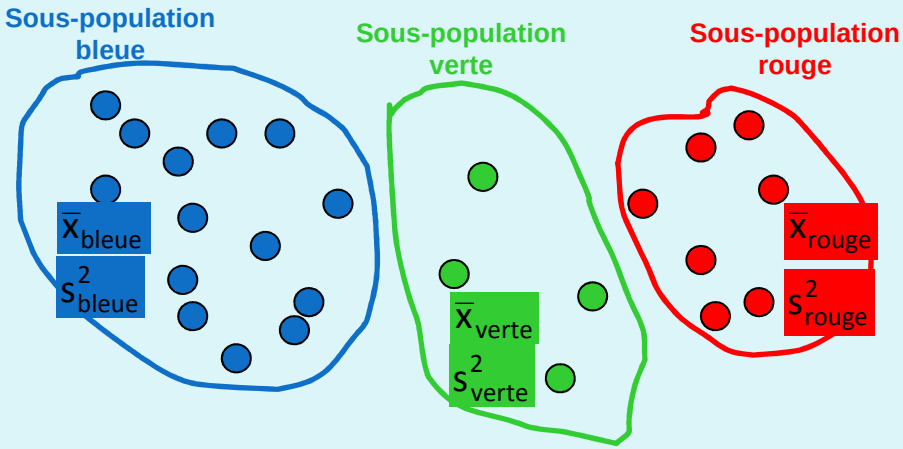
ANOVA à un facteur

- On étudie ensuite la réponse observée sur chaque sous-population en calculant la moyenne et la variance pour chaque modalité.
- Les variances calculées ne représentent pas d'écart entre les modalités (puisque intrinsèques à chaque modalité). On parle alors de variance résiduelle dans le sens où cette variance n'est pas expliquée par le facteur.

	Modalités			
	A_1	A_2		A_p
	x_1^1	x_2^1		x_p^1
			...	
	$x_1^{n_1}$	$x_2^{n_2}$		$x_p^{n_p}$
Moyennes	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_p$
Variances	s_1^2	s_2^2		s_p^2

Variance résiduelle

$$var_R = \frac{1}{n}\sum_{k=1}^{p} n_k \, s_k^2$$



Décomposition de la variance

$$var_{totale} = var_A + var_R$$

Rapport de corrélation

$$\frac{var_A}{var_{totale}}$$

= pourcentage de variabilité de la réponse observée expliquée par le facteur

ANOVA à un facteur

Modalités				
	A_1	A_2		A_p
	x_1^1	x_2^1		x_p^1
			…	
	$x_1^{n_1}$	$x_2^{n_2}$		$x_p^{n_p}$
Moyennes	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_p$
Variances	s_1^2	s_2^2		s_p^2
	X_1	X_2	…	X_p

On suppose que les observations pour la modalités A_i sont les réalisations d'une variable aléatoire X_i, telle que

$$X_i \sim N(\mu_i, \sigma^2)$$

C'est-à-dire que chaque observation s'écrit sous la forme

$$x_i^k = \underbrace{\mu_i}_{\text{Effet moyen sur } A_i} + \underbrace{\varepsilon_i^k}_{\text{Erreur}}$$

L'analyse de la variance consiste à effecteur le test d'hypothèses suivant

$$\begin{cases} H_0 : \mu_1=\mu_2=\ldots=\mu_p=\mu \\ H_1 : \exists i,j \quad \mu_i \neq \mu_j \end{cases}$$

Effets moyens égaux quelle que soit la modalité
⇔
Facteur non influent

Trois hypothèses :

- **A influe linéairement sur la moyenne: $\mu_i=\mu+\alpha_i$ où α_i est l'effet de la modalité A_i**
- **A n'influe pas sur la variance (variance constante) : $\varepsilon_i \sim N(0,\sigma^2)$**
- **Les échantillons (pour chaque modalité) sont gaussiens**

ANOVA à un facteur

Décomposition de la variance *vs* sommes des carrés des écarts

$$S^2 = S_A^2 + S_R^2$$

Variance totale = **variance facteur** + **variance résiduelle**

$$S^2 = \sum_i \sum_j (X_i^j - \overline{X})^2 \qquad S_A^2 = \sum_{i=1}^{p} n_i (\overline{X}_i - \overline{X})^2 \qquad S_R^2 = \sum_i \sum_j (X_i^j - \overline{X}_i)^2$$

Si H_0 est vraie alors

$$F = \frac{S_A^2/(p-1)}{S_R^2/(n-p)}$$

suit une loi Fisher-Snedecor $F_\alpha(p\text{-}1, n\text{-}p)$

⇨ **Si le rapport est supérieur à la valeur critique de la loi de Fisher-Snedecor, on conclut à une influence significative de A (H_1) avec α chance de se tromper.**

On ne connaît pas le risque (β) associé à la décision inverse mais il est d'autant plus grand que S_R est grand devant S_A

ANOVA à un facteur

Tableau type d'analyse de la variance

Source de variation	Somme des carrées	Degré de liberté	Carré moyen
Expliquée	S_A^2	p-1	$C_A^2 = \frac{S_A^2}{p-1}$
Résiduelle	S_R^2	n-p	$C_R^2 = \frac{S_R^2}{n-p}$
Totale	S^2	n-1	

$$\Rightarrow F = \frac{C_A^2}{C_R^2}$$

Exemple : Compétition

Source de variation	Somme des carrées	Degré de liberté	Carré moyen
Expliquée	384,18	p-1=3	128,06
Résiduelle	1377,96	n-p=26	53
Totale	1762,14	n-1=29	

$$\Rightarrow F = \frac{128,06}{53} = 2,42$$

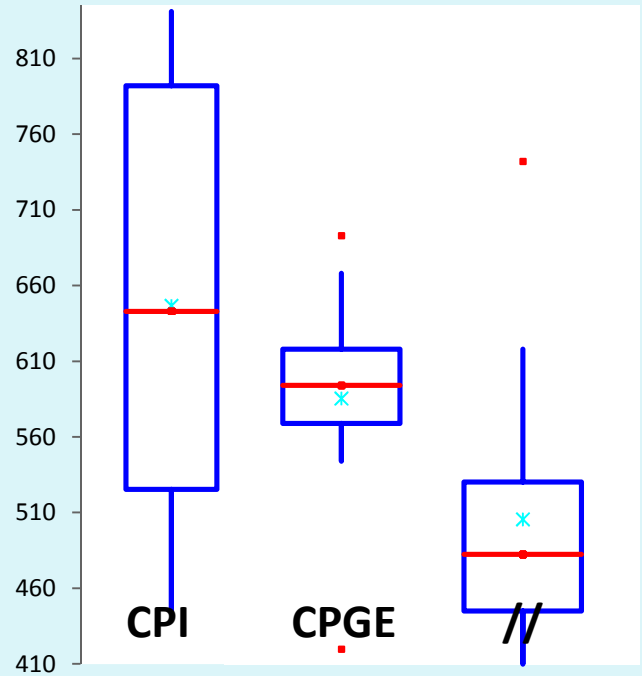
La valeur de F est inférieure au seuil 1% d'une variable F(3;26) qui est 4,64. On conclut que le niveau de compétition n'influe pas sur l'anxiété.

ANOVA à un facteur

Autre exemple : Y-a-t'il un lien entre le score au TOEIC et la provenance de l'étudiant?

CPI	CPGE	//
470	594	495
792	569	594
544	668	445
792	420	410
643	544	509
594	693	445
643	569	420
643	618	460
445	594	470
841		618
495		742
841		495
519		445
792		445
		594
		504

Source de variation	Somme des carrées	Degré de liberté	Carré moyen
Expliquée	149847,38	p-1=2	74923,69
Résiduelle	431545,17	n-p=36	11987,37
Totale	581392,56	n-1=38	



$$F = \frac{74923,69}{11987,37} = 6,25$$

La valeur de F est supérieure au seuil 1% d'une variable F(2;36) qui est 5,25. On conclut que les notes au TOEIC dépendent de la provenance de l'étudiant avec 1% de chance de se tromper.

ANOVA à un facteur

Analyse des contrastes

Admettre l'hypothèse H_1, ne signifie pas que toutes les moyennes $\mu_1,\ldots,\mu_p$ sont différentes. Pour avoir plus de précision, il faut faire des tests deux à deux :

$$\forall i,j \in \{1,\ldots,p\} \qquad \begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$$

- Comparaison de moyennes
- Méthode de Scheffé

si $$\left|\bar{x}_i - \bar{x}_j\right| > \sqrt{(p-1)F_\alpha(p-1,n-p)C_R^2}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

alors H_0 est rejetée

!!!!!!!!!!!!!

$\mu_i=\mu_j$ et $\mu_j=\mu_k$ mais $\mu_i\neq\mu_k$

Exemple : Notes au TOEIC suivant provenance de l'étudiant

	Moyenne
CPI	**647**
CPGE	**585**
//	**506**

	Diff moy	*Val. Scheffé*	*Décision*
CPI-CPGE	**61,27**	**151,55**	**pas de différence significative**
CPGE-//	**79,76**	**147,79**	**pas de différence significative**
CPI - //	**141,03**	**129,81**	**différence significative**

ANOVA à deux facteurs

Facteur A à p niveaux / Facteur B à q niveaux / r répétitions

Facteur B		Facteur A				
		A_1	A_2	A_i	A_p	Moy.
	B_1					
	B_2					
	B_j			$\bar{x}_{ij\bullet}$; x_{ij1}, x_{ij2}, …, x_{ijr}		$\bar{x}_{\bullet j\bullet}$
	B_q					
	Moy.			$\bar{x}_{i\bullet\bullet}$		$\bar{x}$

L'analyse de la variance permet de tester les trois hypothèses H_0 suivantes :

H_0 : Y-a-t-il interaction
($\gamma_{ij} = 0$ pour tous i, j) ?

H_0 : Y-a-t-il un effet ligne ou facteur A
($\alpha_i = 0$ pour tous i) ?

H_0 : Y-a-t-il un effet colonne ou facteur B
($\beta_j = 0$ pour tous j) ?

Toujours deux hypothèses : $X_{ij} \sim N(\mu_{ij}, \sigma^2)$

Effet moyen

Effet de B_j

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Effet de A_i

Effet interaction

ANOVA à deux facteurs

Décomposition de la variance

$$S^2 = S_A^2 + S_B^2 + S_{AB}^2 + S_R^2$$

var. A
+ var. B
+ var. interaction
+ var. résiduelle

= Var. totale

Source de variation	Somme des carrées	Degré de liberté	Carré moyen	F
A	S_A^2	p-1	$C_A^2 = \frac{S_A^2}{p-1}$	$F_A = \frac{C_A^2}{C_R^2}$
B	S_B^2	q-1	$C_B^2 = \frac{S_B^2}{q-1}$	$F_B = \frac{C_B^2}{C_R^2}$
AB	S_{AB}^2	(p-1)(q-1)	$C_{AB}^2 = \frac{S_{AB}^2}{(p-1)(q-1)}$	$F_{AB} = \frac{C_{AB}^2}{C_R^2}$
R	S_R^2	pq(r-1)	$C_R^2 = \frac{S_R^2}{pq(r-1)}$	
Totale	S^2	pqr-1		

- **Si $F_{AB} > F_\alpha((p-1)(q-1);pq(r-1))$ alors on considère que l'interaction AB a une influence avec un risque α de se tromper**
- **Si $F_A > F_\alpha(p-1;pq(r-1))$ alors on considère que le facteur A a une influence avec un risque α de se tromper**
- **Si $F_B > F_\alpha(q-1;pq(r-1))$ alors on considère que le facteur B a une influence avec un risque α de se tromper**

Rq. On ne connaît pas le risque (β) associé à la décision inverse mais il est d'autant plus grand que S_R est grand devant S_A