Cours de Statistique Inférentielle

Jean Christophe meunier

Module 3
Tests de comparaison de moyennes

2ème Bac, Commerce Extérieur
Année académique 2015-2016

EPFC

A. Comparaison entre 2 moyennes
t-tests

t-test pour échantillons indépendants

L'objectif: $\mu_1 - \mu_2$

L'estimateur: $\bar{X}_1 - \bar{X}_2$

Intervalle de confiance à 95%

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

où s_p est la variance commune

On peut montrer que:

$$s_p^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

Le dénominateur est en fait la ddl

ddl = $(n_1 - 1) + (n_2 - 1)$

t-test pour échantillons indépendants

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n_1 + n_2 - 2}$$

Exercice 1

Un échantillon aléatoire de 4 notes est prélevé dans un amphithéâtre : 64, 66, 89 et 77.

Dans un autre amphithéâtre, on tire de façon indépendante un échantillon aléatoire de 3 notes: 56, 71 et 53.

Calculer l'intervalle de confiance pour la différence entre les deux moyennes μ_1 et μ_2.

Construisons le tableau suivant:

Amphi 1			Amphi 2		
X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
64	-10	100	56	-4	16
66	-8	64	71	11	121
89	15	225	53	-7	49
77	3	9			
$\bar{X}_1 = \frac{296}{4} = 74$	0	398	$\bar{X}_2 = \frac{180}{3} = 60$	0	186

Exercice 1 : solution

On peut alors calculer la variance commune s_p^2 :

$$s_p^2 = \frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{398 + 186}{3 + 2} = \frac{584}{5} = 117$$

Au dénominateur figure la valeur de la ddl = 5.

En cherchant la valeur de t dans la table fournie en annexe pour le degré de confiance de 5%, on trouve pour ddl = 5, $t_{0,025} = 2,57$.

En substituant dans:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

on trouve:

$$\mu_1 - \mu_2 = (74 - 60) \pm 2,57\sqrt{117}\sqrt{\frac{1}{4} + \frac{1}{3}} = 14 \pm 21$$

$$= -7 \text{ à } +35$$

t-test pour échantillons appariés

On note X_1 le résultat d'étudiants au 1er semestre

On reprend *les mêmes étudiants* et on note X_2 leurs résultats au 2nd semestre

On souhaite savoir comment leur note a évolué et on calcule les différences $D = X_1 - X_2$

On traite ensuite les 4 différences comme un nouvel échantillon unique et on l'analyse comme n'importe quel échantillon simple:

Etudiant	Notes observées		Différence		
	X_1	X_2	$D = (X_1 - X_2)$	$D - \bar{D}$	$(D - \bar{D})^2$
Jean	64	57	7	-4	16
Marie	66	57	9	-2	4
Meryam	89	73	16	5	25
Sophie	77	65	12	1	1
			$\bar{D} = \frac{44}{4} = 11$	0	$s_D^2 = \frac{46}{3} = 15,3$

Echantillons appariés

On peut construire l'intervalle de confiance pour la distribution D.

L'intervalle de confiance à 95% pour la différence moyenne de la population Δ s'écrit:

$$\Delta = \bar{D} \pm t_{0,025} \frac{s_D}{\sqrt{n}}$$

Dans l'exemple traité, nous avons:

$\Delta = 11 \pm 3{,}18\ (\sqrt{15{,}3} / \sqrt{4}) = 11 \pm 6 = 5$ à 17

Comme la moyenne des différences Δ est égale à la différence des moyennes ($\mu_1 - \mu_2$),

$$\Delta = \mu_1 - \mu_2 \qquad \text{et que: } \bar{D} = \bar{X}_1 - \bar{X}_2$$

on peut écrire:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} \frac{s_D}{\sqrt{n}}$$

L'avantage de l'appariement réside dans le fait que *la marge d'échantillonnage est plus faible*. En effet, nous avions trouvé ± 21 à comparer à ± 6

L'approche du couple apparié est donc meilleure.

Echantillons appariés

t-test pour échantillons appariés

$t_{\bar{d}} = \frac{\bar{d}}{s_{\bar{d}}}$ où $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$ et $\bar{d} = \bar{x}_1 - \bar{x}_2$ (différence moyenne = moyenne des différences!)

Attention: conceptuellement, le *d* utilisé ici est une **différence individuelle entre les mesures de chaque paire d'observations**, et non une différence de moyennes. C'est uniquement grâce à un raccouri arithmétique permis et commode qu'on peut se permettre le calcul de la différence moyenne à l'aide de la formule $\bar{d} = \bar{x}_1 - \bar{x}_2$.

Exercice 2

Supposons que le traitement de départ de MBA de l'ULB est normalement distribué avec une moyenne de 62 000 € et un écart type de 14 500 €. Celui des MBA de l'UCL est normalement distribué avec une moyenne de 60 000€ et un écart type de 18 300€. Si on prend un échantillon aléatoire de 50 ULBistes et un second échantillon aléatoire de 60 UCListes quelle est la probabilité que le traitement moyen de départ d'un ULBiste soit supérieur à celui d'un UCListe ?

Exercice 2 : solution

Nous souhaitons connaître Pr($\bar{X}_1 - \bar{X}_2 > 0$). Nous savons que la différence des moyennes d'échantillon est normalement distribuée avec une moyenne $\mu_1 - \mu_2$ et un écart type

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{14500^2}{50} + \frac{18300^2}{60}} = 3128$$

On peut normaliser et centrer la variable et obtenir:

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{0 - 2000}{3128} = -0,64$$ et donc:

$$\Pr(\bar{X}_1 - \bar{X}_2 > 0) = \Pr(Z > -0,64) = 1 - \Pr(Z < -0,64) = 1 - 0,2611 = 0,7389$$

Exercice 3

Pour mesurer l'impact d'un régime amaigrissant, un club a choisi au hasard un échantillon de 5 individus avant le régime et de 5 autres après. Les poids se présentent ainsi:

Avant: JH 86, KL 99, MM 78, TR 94, MT 86

Après: LW 91, VG 88, EP 74, JC, 81, MW 90

a) Déterminer un intervalle de confiance à 95% pour:
   i. le poids moyen avant régime
   ii. le poids moyen après régime
   iii. la perte de poids moyen durant le régime

b) On a décidé qu'une meilleure manière de tirer des échantillons consistait à peser les mêmes individus avant et après. Dans le désordre, on obtient:

Après: KL 97, MT 80, TR 88, MM 74, JH 78

Sur base de cet échantillon, déterminer un intervalle de confiance à 95% pour la perte moyenne de poids durant le régime

Exercice 3 : solution

a)

Avant			Après		
X_1	$X_1 - \bar{X}_1$	$(X_1 - \bar{X}_1)^2$	X_2	$X_2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
86	-2,6	6,76	91	6,2	38,44
99	10,4	108,16	88	3,2	10,24
78	-10,6	112,36	74	-10,8	116,64
94	5,4	29,16	81	-3,8	14,44
86	-2,6	6,76	90	5,2	27,04
$\bar{X}_1 = \frac{443}{5} = 88,6$	0	263,8 $s_1=\sqrt{52,76}=7.26$	$\bar{X}_2 = \frac{424}{5} = 84,8$	0	206,8 $s_2=\sqrt{41,36}=6.43$

On ne connaît ni μ, ni σ. On doit donc utiliser la formule avec le t de Student, à savoir:

Intervalle de confiance à 95% $\mu = \bar{X} \pm t_{0,025} \frac{s}{\sqrt{n}}$

$\mu_1 = \bar{X}_1 \pm t_{0,025} \frac{s_1}{\sqrt{n_1}}$ $\mu_2 = \bar{X}_2 \pm t_{0,025} \frac{s_2}{\sqrt{n_2}}$

Exercice 3 : solution

Comme $n_1 = n_2 = 5$, dll = 4 avant et après régime et $t_{0,025} = 2,78$ et,

μ_1 a 95% de chances de se trouver dans l'intervalle

$$\bar{X}_1 = 88,6 \quad (78,5 \;;\; 98,7)$$

μ_2 a 95% de chances de se trouver dans l'intervalle

$$\bar{X}_2 = 84,8 \quad (75,8 \;;\; 93,8)$$

La différence entre la moyenne avant et après régime a 95% de chances de se trouver dans l'intervalle suivant:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

où: $$s_p^2 = \frac{\sum (X_1 - \bar{X}_1)^2 + \sum (X_2 - \bar{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

et donc, $\mu_1 - \mu_2$ a 95% de chances de se trouver dans

$$\bar{X}_1 - \bar{X}_2 = 3,8 \quad (0,37 \;;\; 7,23)$$

Exercice 3 : solution

b)

Individu	Poids		Différence		
	$X_{1\,(avant)}$	$X_{2\,(après)}$	$D = (X_1 - X_2)$	$D - \bar{D}$	$(D - \bar{D})^2$
JH	86	78	8	2,8	7,84
KL	99	97	2	-3,2	10,24
MM	78	74	4	-1,2	1,44
TR	94	88	6	0,8	0,64
MT	86	80	6	0,8	0,64
			$\bar{D} = \frac{26}{5} = 5,2$	0	$s_D^2 = 20,8 / 5 = 4,16$

L'intervalle de confiance à 95% pour les échantillons appariés est donné par:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} \frac{s_D}{\sqrt{n}} = 88,6 - 84,8 \pm 2,78 \sqrt{\frac{4,16}{4}} = 3,8 \pm 2,8$$

B. Analyse de Variance (ANOVA) Comparaison de plusieurs moyennes

Introduction

- ANOVA pour '**AN**alysis **O**f **VA**riance'
  - Contrairement à son nom, analyse les moyennes et non les variances
- Extension du test de comparaison de deux moyennes (échantillons indépendants)
- Permet de comparer plus de deux groupes

H_0 et H_A

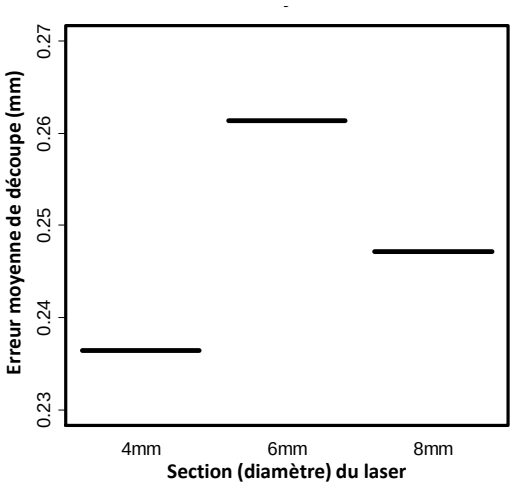
- ANOVA teste l'H_0 selon laquelle :
  - H_0: $\overline{X}_1 = \overline{X}_2 = \ldots = \overline{X}_K$
  - Les moyennes des groupes/échantillons sont toutes égales
- H_A suggère :
  - H_A: $\overline{X}_i \neq \overline{X}_j$ pour quelques *i, j*
  - 'Les moyennes des groupes/échantillons *ne* sont *pas* toutes égales'
  - Autrement dit, 'il existe *au moins une différence significatives entre deux groupes/échantillons* parmi les k groupes/échantillons'

Exemple : machine de découpage laser

- Un constructeur automobile veut acquérir une machine de découpage laser
- La machine existe en 3 versions selon section (diamètre) du laser : 4 mm, 6 mm, 8mm
- Le choix se portera sur la machine la plus précise

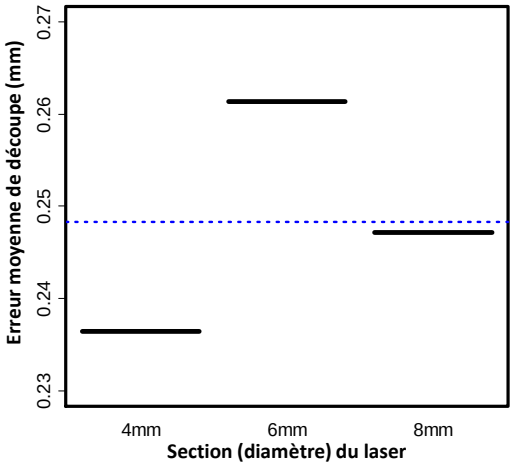
Exemple : machine de découpage laser

- La précision de chacune des trois machines est testée sur 15 véhicules
- Le schéma représente l'erreur moyenne de découpe pour chacune des machines
- ANOVA
  - H_0 : $\overline{X}_1 = \overline{X}_2 = \overline{X}_3$
  - H_A: $\overline{X}_i \neq \overline{X}_j$ pour quelques *i, j*



Exemple : machine de découpage laser

- La moyenne globale des trois échantillons (notée $\overline{\overline{X}}$) = 0.248 mm.
- Si H_0 se vérifie, on s'attend à ce que la distance des moyennes des échantillons soit proches de la moyenne globale ($\overline{\overline{X}}$).



Variabilité inter-échantillon

- Effets liés aux groupes/échantillons
  - (machines 1 vs 2 vs 3)

- Somme des carrés (SC_{inter})
  - Quantifie les différences entre chaque groupe et la moyenne globale

$$SC_{inter} = \sum_{i=1}^{a} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2$$

$$SC_{inter} = 15\left(\overline{X}_{4mm} - \overline{\overline{X}}\right)^2 + 15\left(\overline{X}_{6mm} - \overline{\overline{X}}\right)^2 + 15\left(\overline{X}_{8mm} - \overline{\overline{X}}\right)^2$$

→0,005

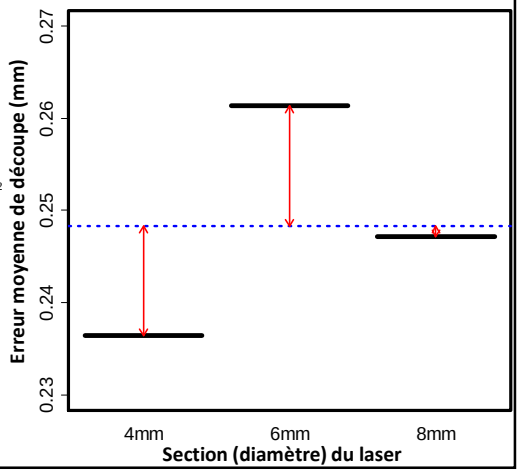
- Carré moyen (CM_{inter})
  - SC pondéré par le nombre de groupe – 1

$$CM_{inter} = \frac{SC_{inter}}{a - 1}$$

→0,005

→3 -1 = 2

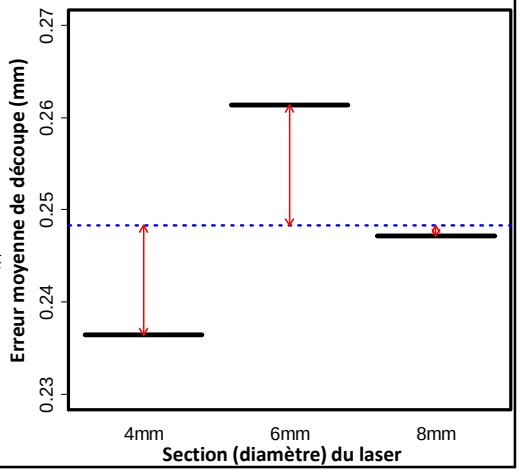
a = nombre de groupes
n = nombre d'observations dans échantillons



Variabilité inter-échantillon

- Remarque :
  si tailles des échantillons/groupes égales
  $CM_{inter} = n * s^2_{\overline{X}}$
  - Ou n = taille des échantillons
  - Ou $s^2_{\overline{X}}$ variance des moyennes des échantillons

$$CM_{inter} = \frac{\sum_{i=1}^{a} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2}{a - 1} = n \frac{\sum_{i=1}^{a} \left(\overline{X}_i - \overline{\overline{X}}\right)^2}{a - 1} = n s^2_{\overline{X}}$$

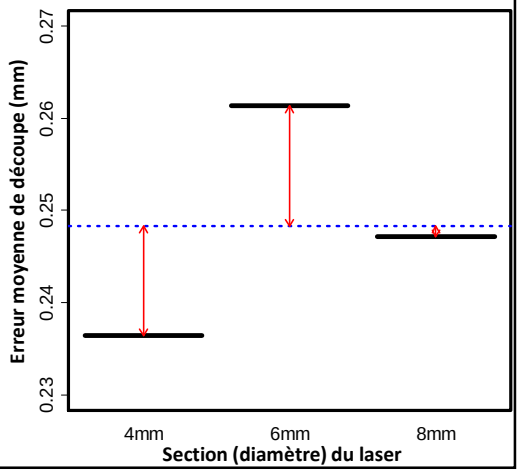


Variabilité inter-échantillons

- Comment interpréter ces effets liés aux groupes ?

- *différence significative entre les groupes* ou liées au *fluctuations du hasard* ?
- Mesurer l'effet lié au hasard
- Confronter variation inter-échantillon à l'effet du hasard

Effet du hasard = Variation intra-échantillon



Variabilité intra-échantillon

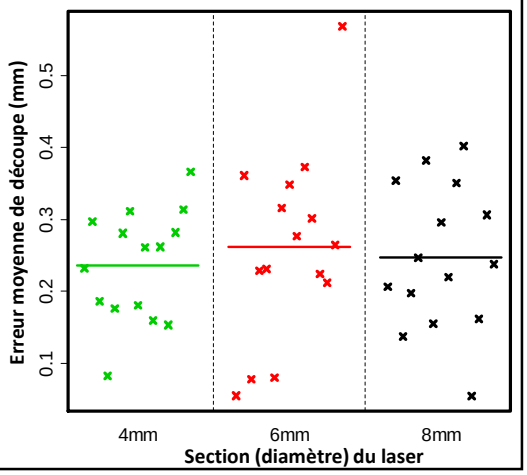
- Effet du hasard
  - Comprend les fluctuations possibles au sein des groupes

- Somme des carrés (SC_{intra})
  - Quantifie les différences au sein des groupes entre les observations individuelles et la moyenne du groupe

$$SC_{intra} = \sum_{i=1}^{a}\sum_{t=1}^{n} \left(X_{it} - \overline{X_i}\right)^2$$

- Carré moyen (CM_{intra})
  - SC pondéré par le nombre totale d'observations – le nombre de groupe

$$CM_{intra} = \frac{SC_{intra}}{N - a}$$



Variabilité intra-échantillon

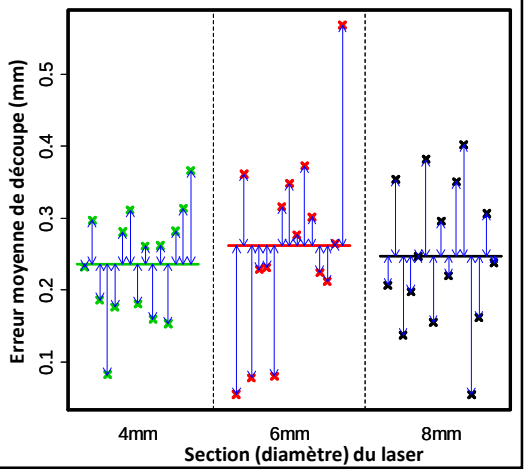
- Effet du hasard
  - Comprend les fluctuations possibles au sein des groupes

- Somme des carrés (SC_{intra})
  - Quantifie les différences au sein des groupes entre les observations individuelles et la moyenne du groupe

$$SC_{intra} = \sum_{i=1}^{a}\sum_{t=1}^{n} \left(X_{it} - \overline{X_i}\right)^2 \rightarrow 0{,}467$$

- Carré moyen (CM_{intra})
  - SC pondéré par le nombre totale d'observations – le nombre de groupe

$$CM_{intra} = \frac{SC_{intra}}{N - a} \rightarrow 0{,}467 \rightarrow 45 - 3 = 42$$



Variabilité intra-échantillon

- Calcul alternatif de CM_{intra}
  - Calculer les variances de chaque groupe et en faire la moyenne

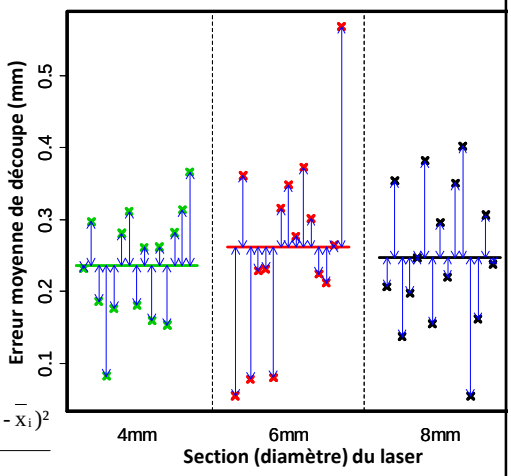
$$CM_{intra} = \frac{\sum_{i=1}^{a} s^2_{\overline{x}_i}}{a}$$

Sachant que,

$$s^2_{\overline{x}_i} = \frac{\sum_{i=1}^{n} (x_{it} - \overline{x}_i)^2}{n - 1}$$

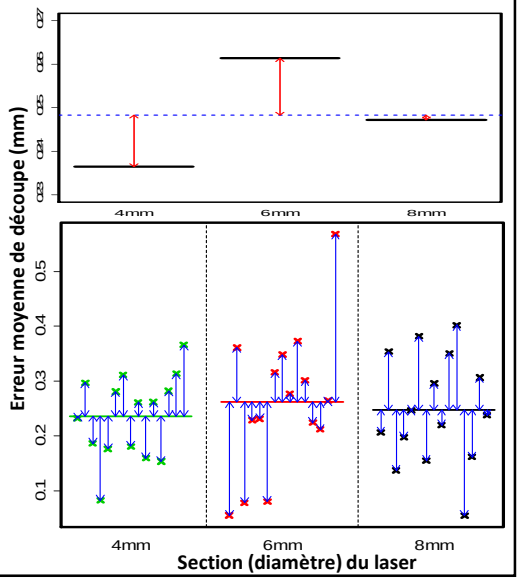
Dès lors,

$$CM_{intra} = \frac{\sum_{i=1}^{a} s^2_{\overline{x}_i}}{a} = \frac{\sum_{i=1}^{a}\sum_{t=1}^{n} (x_{it} - \overline{x}_i)^2}{\sum_{i=1}^{a} (n_i - 1)} = \frac{\sum_{i=1}^{a}\sum_{t=1}^{n} (x_{it} - \overline{x}_i)^2}{N - a}$$



Variabilité intra-échantillon

- CM_{intra} = ***mesure de la variabilité après que l'effet des groupes ait été pris en compte***
  - Varabilité // moyennes de chaque groupe
  - et non variabilité // moyenne globale
- CM_{intra} aussi appelé,
  - CM_{erreur} ou CM_E
- Plus l'effet CM_{intra} est important, moins l'effet groupe a de chance d'être significatif



Test *F*

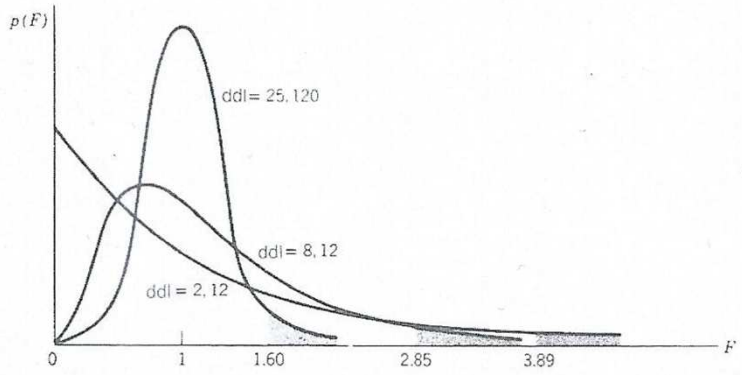
- Permet de savoir si H_0 doit être acceptée ou rejetée au profit de H_A

$$F = \frac{CM_{inter}}{CM_{intra}} \qquad F = \frac{\text{Fluctuations expliquées par l'effet des groupes}}{\text{Fluctuations inexpliquées par l'effet des groupes}}$$

  - Les fluctuations entre les groupes (inter) sont-elles suffisamment grandes par rapport aux fluctuations au sein des groupes (intra → hasard) pour affirmer que les moyennes ne sont pas toutes égales (H_A)
- F suit une loi de probabilité
  - Voir tableau VI pour les probabilités critiques
  - S'interprète à l'aide de deux degré de libérté (ddl)
    - ddl inter (numérateur dans l'équation) = a-1
    - ddl intra (dénominateur dans l'équation) = N-a (si groupe de taille égale = a(n-1))

Loi de distribution F

Il y a une distribution différente pour chaque combinaison de dll. En voici quelques unes:



La zone grisée fournit la valeur critique de F au seuil de 5%. On voit que cette valeur critique se déplace vers la gauche et tend vers 1, lorsque dll augmente.

Tableau ANOVA

- Résumé des différentes opérations

Source de variation	Variation (somme des carrés, SC)	ddl	Variance (carré moyen, CM)	Rapport F
FACTEUR A: différence entre les moyennes des machines $\bar{X}_i$	$SC_A = n\sum_{i=1}^{a}(\bar{X}_i - \bar{\bar{X}})^2$	a – 1	$CM_A = SC_A/(a-1) = ns^2_{\bar{X}}$	$F = \frac{CM_A}{CM_E} = \frac{ns^2_{\bar{X}}}{s^2_p}$
RESIDUELLE (ERREUR) différence entre les observations X_{it} et les moyennes $\bar{X}_i$	$SC_E = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{X}_i)^2$	N-a	$CM_E = SC_E /$ N-a $= s^2_p$	
TOTAL	$SC = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{\bar{X}})^2$	N-1		

Variation ***inter-groupe*** (facteur A, dans ce cas-ci les trois machines)

Variation ***intra-groupe*** (erreur ou variation résiduelle, car non expliquée par le facteur A (machines))

Tableau ANOVA

- Résumé des différentes opérations

Source de variation	Variation (somme des carrés, SC)	ddl	Variance (carré moyen, CM)	Rapport F
FACTEUR A: différence entre les moyennes des machines $\bar{X}_i$	$SC_A = n\sum_{i=1}^{a}(\bar{X}_i - \bar{\bar{X}})^2$	a – 1	$CM_A = SC_A/(a-1) = ns^2_{\bar{X}}$	$F = \frac{CM_A}{CM_E} = \frac{ns^2_{\bar{X}}}{s^2_p}$
RESIDUELLE (ERREUR) différence entre las observations X_{it} et les moyennes $\bar{X}_i$	$SC_E = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{X}_i)^2$	N-a	$CM_E = SC_E /$ N-a $= s^2_p$	
TOTAL	$SC = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{\bar{X}})^2$	N-1		

$SC_{total} = SC_A + SC_E$
$ddl_{total} = ddl_A + ddl_E$

Tableau ANOVA

- Résumé des différentes opérations

Source de variation	Variation (somme des carrés, SC)	ddl	Variance (carré moyen, CM)	Rapport F
FACTEUR A: différence entre les moyennes des machines $\bar{X}_i$	$SC_A = n\sum_{i=1}^{a}(\bar{X}_i - \bar{\bar{X}})^2$	a – 1	$CM_A = SC_A/(a-1) = ns^2_{\bar{X}}$	$F = \frac{CM_A}{CM_E} = \frac{ns^2_{\bar{X}}}{s^2_p}$
RESIDUELLE (ERREUR) différence entre las observations X_{it} et les moyennes $\bar{X}_i$	$SC_E = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{X}_i)^2$	N-a	$CM_E = SC_E /$ N-a $= s^2_p$	
TOTAL	$SC = \sum_{i=1}^{a}\sum_{t=1}^{n}(X_{it} - \bar{\bar{X}})^2$	N-1		

En divisant chaque SC par son ddl on obtient le Carré moyen (CM)

Tableau ANOVA

Dans l'exemple des machines de découpage laser

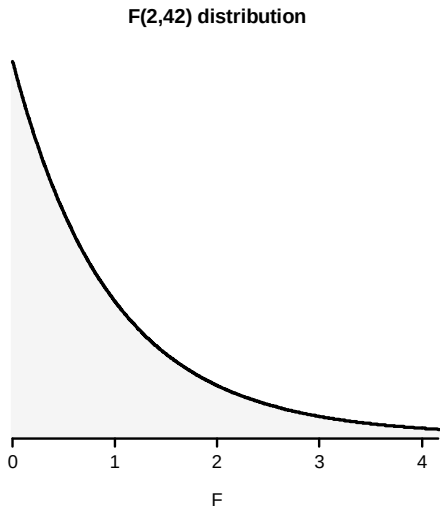
Source de variation	SC	ddl	CM	F	Sig.
Facteur 'Machines' Inter-échantillon	.005	2	.002	.211	.811
Erreur ou variance résiduelle Intra-échantillon	.467	42	.011		
Total	.472	44			

CM_{inter} — CM_{intra} ou CM_{Erreur} — Test F — p valeur

Ex. Machine de découpage laser

p-valeur du test F

Pour avoir la p-valeur, on compare le test F à une distribution F de degrés de liberté 2 (inter) et 42 (intra) : F(2, 42).



Ex. Machine de découpage laser

p-valeur du test F

Pour avoir la p-valeur, on compare le test F à une distribution F de degrés de liberté 2 (inter) et 42 (intra) : F(2, 42).

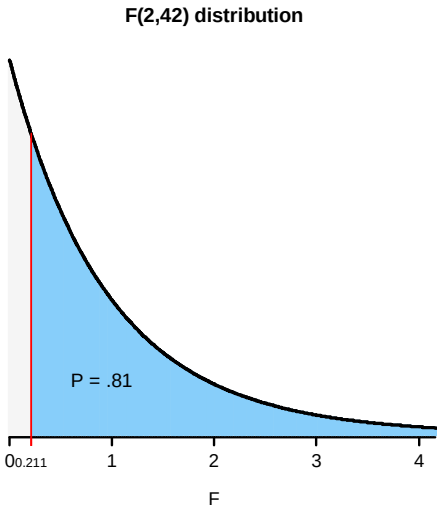
Dans tableau VI, probabilité critique ($p<0,05 \rightarrow F_{0,05}$) pour F(2,42) = 3,23

- au-dela de cette valeur moins de 5 % de chance que les moyennes soient égales (H_0 rejetée)

Dans notre exemple,

$$F = \frac{.0047/2}{.0467/42} = .211$$

- Cette valeur est < 3,23 (dès lors, H_0 acceptée)



H_0 et H_A pour L'ANOVA

- Si H_0 acceptée : $\overline{X}_1 = \overline{X}_2 = \ldots = \overline{X}_K$
  - Les moyennes des groupes/échantillons sont toutes égales
- Si H_0 rejetée, H_A suggère : $X_i \neq \overline{X}_j$ pour quelques *i, j*
  - 'Les moyennes des groupes/échantillons *ne* sont *pas* toutes égales'
  - Autrement dit, 'il existe *au moins une différence significatives entre deux groupes/échantillons* parmi les k groupes/échantillons'

Mais comment savoir quelle paire(s) de moyennes montre une différence significative → **Test Post-hoc** : comparer les moyenne deux à deux

Test Post hoc

- Comparer les moyennes deux à deux
  - Démarche similaire à un t-test pour deux échantillons indépendants
  - Il existe plusieurs types de test post hoc. Le plus simple appelé LSD (Least significant difference)
    - Directement dérivé du t-test pour échantillon indépendant

Intervalle de confiance à 95%

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

On remplace la variance commune du t-test (s_p) par $CM_{intra}/CM_{erreur}/CM_{résiduelle}$

Pour effectuer le t-test, le nombre de degrés de liberté est celui $CM_{intra}/CM_{erreur}/CM_{résiduelle}$

Illustration test post hoc (1)

On souhaite comparer 3 machines opérées par des opératrices. La production est sujette à des fluctuations et on souhaite comprendre les causes. On prélève sur chaque machine un échantillon aléatoire de la production obtenue au cours de 5 périodes différentes d'une heure chacune.

	Machine 1	Machine 2	Machine 3
	47	55	54
	53	54	50
	49	58	51
	50	61	51
	46	52	49
$\bar{X}_i$	*49*	*56*	*51*

Tableau 10.1

Illustration test post hoc (2)

Le tableau ANOVA de l'exemple précédent (données du tableau 10.1) se présente de la manière suivante:

Source	SC	ddl	CM	Rapport F	p_c
MACHINES	130	2	65 (= 130/2)	65/7,83 = 8,3	< 0,01
RESIDUELLE	94	12	7,83 (= 94/12)		
TOTAL	224	14			

Tableau 10.5 – Calculs ANOVA

Le tableau ANOVA permet un contrôle commode des calculs.

La somme des carrés SC (ou *variations*) des 2 premières SC du tableau ci-dessus est égale à la SC_{totale} . On a bien: $\mathbf{SC_{totale} = SC_A + SC_E}$

Les dll s'additionnent également: $\mathbf{dll_{totale} = dll_A + dll_E}$

En divisant chaque SC par son dll, on obtient le carré moyen **CM** (ou variance).

La variance entre machines s'explique par le fait que les colonnes du tableau 10.1 peuvent provenir de populations apparentées (machines à productions différentes).

La variance résiduelle au sein des colonnes est inexpliquée car aléatoire (variation contingente).

Illustration test post hoc (3)

Le tableau ANOVA de l'exemple précédent (données du tableau 10.1) se présente de la manière suivante:

Source	SC	ddl	CM	Rapport F	p_c
MACHINES	130	2	65 (= 130/2)	65/7,83 = 8,3	< 0,01
RESIDUELLE	94	12	7,83 (= 94/12)		
TOTAL	224	14			

Tableau 10.5 – Calculs ANOVA

Intervalle de confiance à 95%

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

On remplace la variance commune du t-test (s_p) par la racine carrée de **CM_{intra}/CM_{erreur}/$CM_{résiduelle}$**

Illustration test post hoc (4)

Le tableau ANOVA de l'exemple précédent (données du tableau 10.1) se présente de la manière suivante:

Source	SC	ddl	CM	Rapport F	p_c
MACHINES	130	2	65 (= 130/2)	65/7,83 = 8,3	< 0,01
RESIDUELLE	94	12	7,83 (= 94/12)		
TOTAL	224	14			

Tableau 10.5 – Calculs ANOVA

Pour effectuer le t-test, le nombre de degrés de liberté (**ddl**) est celui $CM_{intra}/CM_{erreur}/CM_{résiduelle}$

Intervalle de confiance à 95%

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

On remplace la variance commune du t-test (s_p) par la racine carrée de **$CM_{intra}/CM_{erreur}/CM_{résiduelle}$**

Illustration test post hoc (5)

Le tableau ANOVA de l'exemple précédent (données du tableau 10.1) se présente de la manière suivante:

Source	SC	ddl	CM	Rapport F	p_c
MACHINES	130	2	65 (= 130/2)	65/7,83 = 8,3	< 0,01
RESIDUELLE	94	12	7,83 (= 94/12)		
TOTAL	224	14			

formule à partir du tableau ANOVA,

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

où:

- μ_1 et μ_2 sont les 2 moyennes de population à comparer;
- $\bar{X}_1$ et $\bar{X}_2$ sont les 2 moyennes d'échantillons correspondantes (des 2 premières colonnes, per exemple);
- n_1 et n_2 sont les nombres d'observations utilisées pour calculer $\bar{X}_1$ et $\bar{X}_2$;
- s^2 est la variance résiduelle dans le tableau ANOVA (CM_E)

Illustration test post hoc (6)

Construire un intervalle de confiance à 95% pour la différence entre la 2ième et la 3ième machine du tableau 10.1 et dont on donné les calculs ANOVA dans le tableau 10.5

Solution

On remplace les quantités appropriées dans l'équation:

$$\mu_1 - \mu_2 = (\bar{X}_1 - \bar{X}_2) \pm t_{0,025} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

On utilise le $s^2 = 7,83$ du tableau ANOVA 10.5 et son dll correspondant (dll = 12) pour la valeur $t_{0,025}$ et on obtient:

$$\mu_1 - \mu_2 = (56 - 51) \pm 2,18\sqrt{7,83}\sqrt{\frac{1}{5} + \frac{1}{5}}$$

$$= 5,0 \pm 3,9 \simeq 5 \pm 4$$

On peut donc conclure, qu'au seuil de confiance de 95%, la 2ième machine est de 1 à 9 unités plus performante que la 3ième machine.