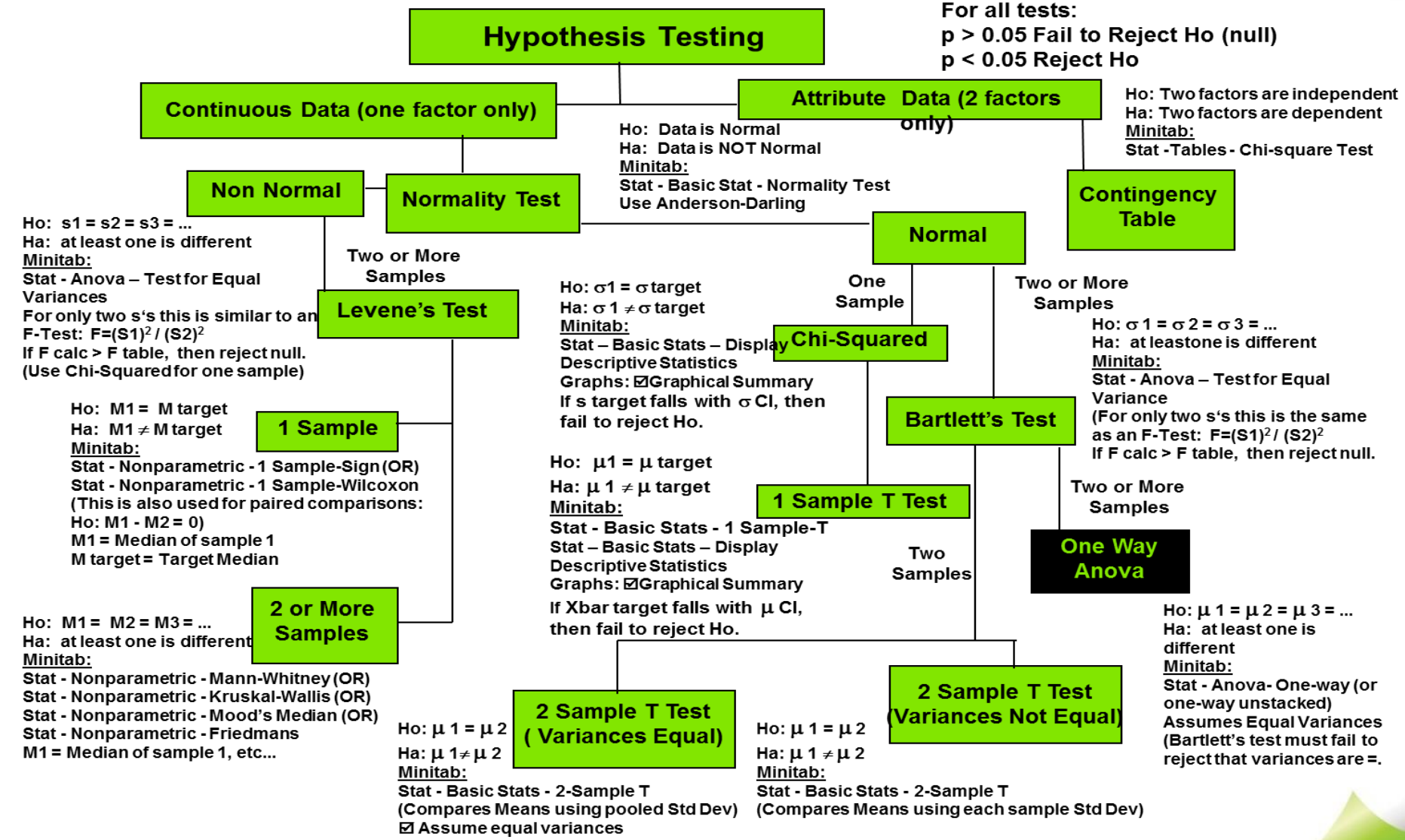
Analyse des variances (ANOVA)

Agenda

- **Caractéristiques de l'Analyse de la Variance**
- **Méthode Générale**
  - Le modèle
  - Les suppositions du modèle
  - Test du supposition
  - La somme des carrées
  - Calcul du test statistique
  - Exercice
  - Le carrée d' Epsilon
- **Exemple**

Hypothesis Testing Roadmap



Analyse de la Variance

- **L'Analyse de Variance est un outil utilisé pour détecter s'il existe une différence statistique (μ ou σ) entre plusieurs facteurs et si cette différence est attribuée au hasard ou à une cause spécifique (les paramètres viennent de même population ou non.)**

- **Cet outil utilise les statistiques pour déterminer si la variation dans un facteur est supérieure ou inférieure à la variation entre les facteurs. Si la variation entre les facteurs est supérieure à la variation entre les niveaux de facteurs, alors on dit que le facteur est significatif**

- **Les outputs sont généralement mesurées sous forme d'intervalle/Echelle (Rendement, température, voltes, % impuretés, etc…)**

- **Les inputs ou facteurs sont des données catégoriques.**

- **On veut répondre à la question:**

"Existe t il une différence significative entre les facteurs _____&____&___…?

Méthode ANOVA

Step 1: Statuer le problème pratique (Graph data)

Step 2: Statuer l'hypothèse nulle et alternative

Step 3: Choisir le test statistique approprié – ANOVA

- Les moyennes sont indépendantes et normalement distribués
- Les variances sont égaux pour tous les facteurs

Step 4: Statuer le niveau alpha (5%)

Step 5: Calculer la taille des échantillons

Step 6: Développer le plan d'échantillonnage

Step 7: Construire le tableau ANOVA

Step 8: Interpréter p-value (statistique F) pour l'effet des facteurs

P-value <.05, REJETER H_o
Autrement l'hypothèse nulle ne peut pas être rejetée

Calculer le carrée epsilon

Step 9: Faire la supposition pour l'erreur (analyse des résidus)

Erreurs sont indépendantes et distribuées normalement

Step 10: Traduire la conclusion statistique en langage processus

Etape 02: Définir l'hypothèse

Développer les énoncés d'hypothèses

Ho: $\mu_{Site1} = \mu_{Site}2 = \mu_{Site3} = \mu_{Site4}$

Ha: au moins une moyenne d'un site diffère

- **Modèle Mathématique:**

$$Y_{ti} = \mu + \tau_i + \varepsilon_{ti}$$

Y_{ti} = La réponse du traitement

μ = La moyenne totale

τ_i = Traitement

ε_{ti} = Erreure aléatoire

Note: H_o assume pas d'effet du traitement

Hypothèse Mathématique:

$$H_o : \tau's = 0$$

$$H_a : \tau_k \neq 0$$

Hypothèse Conventionnel:

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu \ldots$$

$$H_a : \text{Au minimum un } \mu_k \text{ est différent}$$

Etape 03:

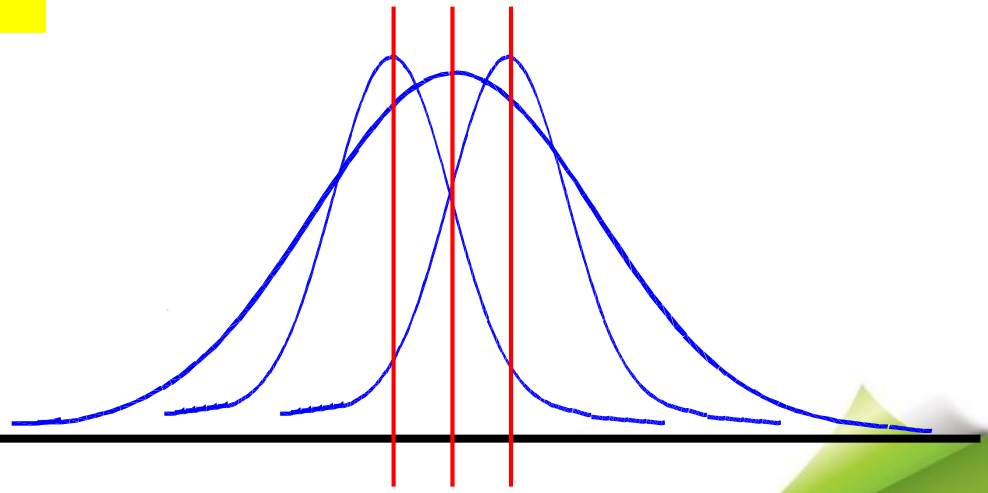
- **Les moyennes sont indépendantes et normalement distribués**
  - Des prélèvement aléatoire durant l'expérience
  - Assurer une taille adéquate de l'échantillon
  - Vérifier le test de normalité

- **Les variances des populations sont égales à tous les niveaux des facteurs (test d'égalité des variances)**

$$H_0 : \sigma_{pop1} = \sigma_{pop2} = \sigma_{pop3} = \sigma_{pop4} = \ldots$$

$$H_a : \text{au moins un est différent}$$

Note: La différence entre facteurs peut être impactée par la variance d'un facteur si les variance ne sont pas égales. La supposition de l'égalité des variances est généralement vrai surtout si on a un test équilibré (même nombre d'observation)



Step 7: Tableau ANOVA

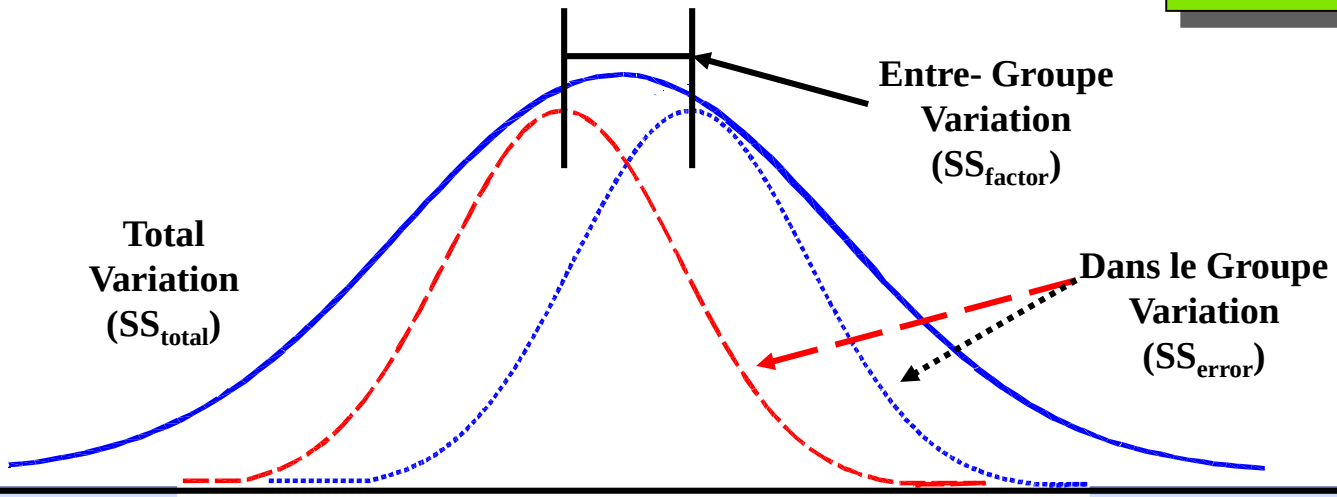
SOURCE	SS	df	MS	Test Statistic
Factor	SS_{factor}	g-1	$MS_{factor}= SS_{factor}/(g-1)$	$F= MS_{factor}/ MS_{error}$
Error	SS_{error}	g(n-1)	$MS_{error}= SS_{error}/[g(n-1)]$	
Total	SS_{total}	ng-1		

$$\sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{\bar{x}})^2 = \sum_{j=1}^{g} n*(\bar{x}_j-\bar{\bar{x}})^2 + \sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

SS_{total} **$SS_{facteurs}$** **SS_{erreur}**

g = nombre des sous groupes

n = nombre des échantillons par facteur



Steps 8-10

Step 8: Interpréter le p-value (ou le F-statistic) pour l'effet des facteurs

- P-value < .05, rejeter H_o
- Autrement assumer que l'hypothèse nulle est vrai.
- Calculer le carrée d'epsilon des facteurs et l'erreur

$$\varepsilon^2_{factor} = \frac{SS_{Between}}{SS_{Total}}$$

$$\varepsilon^2_{error} = \frac{SS_{Error}}{SS_{Total}}$$

Step 9: Faire la supposition pour l'erreur (analyse des résidus)

- Erreurs sont indépendantes et distribuées normalement
- Effectuer l'histogramme des résidus, le test de normalité , le charte graphique (erreur par rapport à la moyenne)

Step 10: Traduire la conclusion statistique en langage processus

ANOVA – Etape 1

Step 1: Statuer le problème pratique (Graph data)

Une société financière possède quatre sites différents qui traitent les affaires de crédit. Le tableau ci-dessous contient les données de productivité sur le nombre moyen de cas traités par heure pour un échantillon d'employés sur chacun des quatre sites.

Site 1	Site 2	Site 3	Site 4
14.9	15.7	17.3	15.2
15.7	16.6	17.2	14.8
15.2	16.5	17.4	14.3
15.8	16	17.2	14.9
15.1	15.7	17	15.4
16.3	16.4	17.6	14.9
14.4	16.7	17.4	14.6
15.9	16.8	17.3	15.1
	16.3	16.5	15
	16.5	16.7	14.7

Que pouvez-vous dire des données ?
Y-a-t-il une différence entre les sites ?

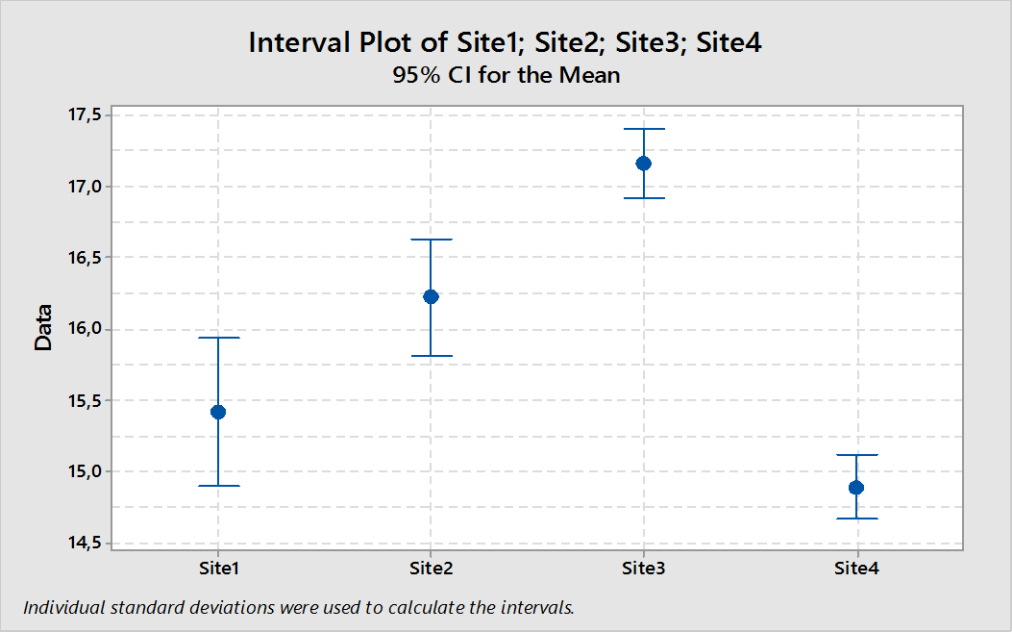
ANOVA – Etape 2

Step 2: Statuer l'hypothèse nulle et alternative

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_{...}$$

$$H_a : \text{Au minimum un } \mu_k \text{ est différent}$$

Graphique des principaux effets pour le nbre de cas/heure



Il semble que la productivité de cas par heure soit la plus élevée au site 3.

Comment pouvons-nous en être sûrs ?

Les graphiques des effets principaux ne sont pas des tests statistiques !

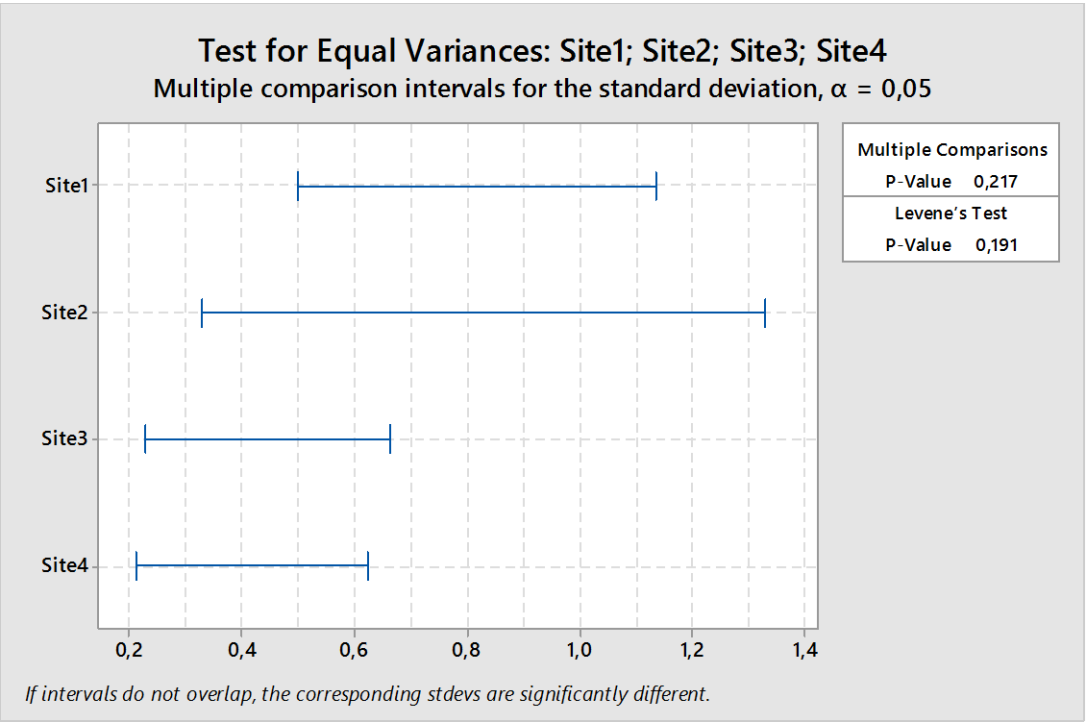
ANOVA – Etape 3

Step 3: Choisir le test statistique approprié – ANOVA

Test d'homogèniété de la variance

Ho: $\sigma 1 = \sigma 2 = \sigma 3 = \sigma 4$

Ha: au moins une diffère



Méthode Bonferroni (utiliser α/2p (p nbr facteurs)

	σ_1	σ_2	σ_3	σ_4
	0,620	0,571	0,337	0,314
CI_H	1,591	1,263	0,746	0,695
CI_L	0,370	0,358	0,211	0,197

La valeur-P doit être > 0.05 pour qu'on ne rejette pas Ho.

```
Sample    N     StDev           CI
 Site1    8  0,619764  (0,323699; 1,72527)
 Site2   10  0,571159  (0,235072; 1,84977)
 Site3   10  0,337310  (0,147666; 1,02703)
 Site4   10  0,314289  (0,166249; 0,79196)

Individual confidence level = 98,75%

                          Test
Method                Statistic  P-Value
Multiple comparisons          —    0,217
Levene                     1,67    0,191
```

P>α on échoue à rejeter H0, donc les variance sont égaux

ANOVA – Etape 7

$$\sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{\bar{x}})^2=\sum_{j=1}^{g}n*(\bar{x}_j-\bar{\bar{x}})^2+\sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

Site1	Site2	Site3	Site4
14,9	15,7	17,3	15,2
15,7	16,6	17,2	14,8
15,2	16,5	17,4	14,3
15,8	15	17,2	14,9
15,1	15,7	17	15,4
16,3	16,4	17,6	14,9
14,4	16,7	17,4	14,6
15,9	16,8	17,3	15,1
	16,3	16,5	15
	16,5	16,7	14,7

Y_{1bar}	Y_{2bar}	Y_{3bar}	Y_{4bar}
15,41	16,22	17,16	14,89
		Y_{barbar}	15,92

Calcul somme des carrées dû aux facteur

Site1	Site2	Site3	Site4
2,066	0,896	15,361	10,622
$\sum N*(Y_{ibar}-Y_{barbar})^2$			28,94

Somme des carrées erreur

Site1	Site2	Site3	Site4
0,263	0,270	0,020	0,096
0,083	0,144	0,002	0,008
0,045	0,078	0,058	0,348
0,150	1,488	0,002	0,000
0,098	0,270	0,026	0,260
0,788	0,032	0,194	0,000
1,025	0,230	0,058	0,084
0,238	0,336	0,020	0,044
	0,006	0,436	0,012
	0,078	0,212	0,036
$\sum(Yij - Ybar)^2$			7,54

Somme des carrées total

Site1	Site2	Site3	Site4
1,042	0,049	1,903	0,519
0,049	0,462	1,637	1,256
0,519	0,336	2,189	2,626
0,015	0,848	1,637	1,042
0,673	0,049	1,165	0,271
0,144	0,230	2,820	1,042
2,312	0,607	2,189	1,744
0,000	0,773	1,903	0,673
	0,144	0,336	0,848
	0,336	0,607	1,490
$\sum\sum(Yij - Ybarbar)^2$			36,482

ANOVA – Etape 7

Analyse de Variance à sens unique

Source	Df	SS	MS	F	P
Facteurs	3	28,94	9,65	43,52	0,00
Erreur	34	7,54	0,22		
Total	37	36,48			

Sources de variabilité

Quantité d'information Degrés de liberté

Mesure quantitative de la variabilité expliquée par chaque source

Estimation des variances

La mesure statistique utilisée pour déterminer si un facteur est significatif

Erreur de Type I (valeur-P)

ANOVA – Etape 7

Calculer les Moyennes des carrés.

$$\text{MS Niveau facteur} = \frac{\text{SS niveau facteur}}{n - 1}$$

$$\text{MS Erreur} = \frac{\text{SS Erreur}}{n - 1}$$

Lorsque les sommes des valeurs au carré sont divisées par le nombre approprié de degrés de liberté, la moyenne des valeurs au carré donne une bonne estimation de la variabilité.

$$F = \frac{\text{MS Niveau facteur}}{\text{MS Erreur}}$$

- **Est faible, l'erreur joue peut-être un GRAND rôle comme facteur. On ne peut pas prouver que le facteur est fortement responsable des différences de réussites. *Ne pas rejeter Ho.***
- **Est grande, le facteur joue un rôle significatif dans les différences de réussites. *On peut rejeter Ho*.**

Le taux F sert à déterminer la valeur P!

ANOVA – Etape 8

Evaluer la valeur-P (Loi distribution F).

Dans notre exemple, on a une valeur-P = 0.000.

Par conséquent, Ho peut être rejetée et nous pouvons conclure que la moyenne des cas traités par heure est différente dans au moins un site.

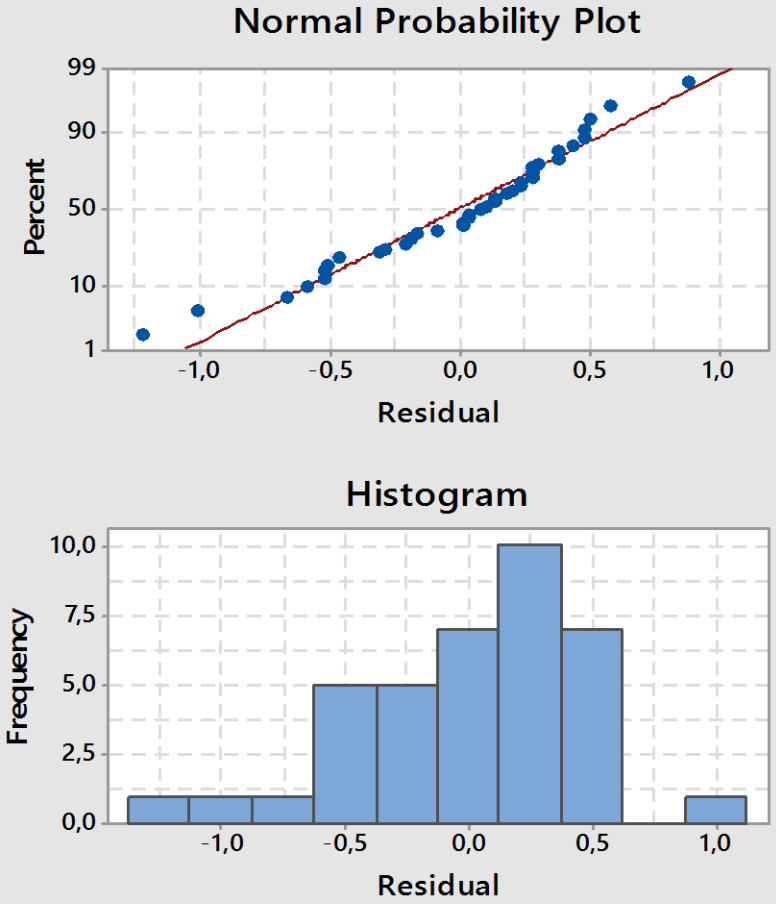
Autrement dit, la variation entre les sites =0,993 (écart type globale) est supérieure à la variation dans chaque site.

$$\varepsilon^2_{factor} = \frac{SS_{Between}}{SS_{Total}} = \frac{28,94}{36,48} = 79,34\% \qquad \varepsilon^2_{error} = \frac{SS_{Error}}{SS_{Total}} = \frac{7,54}{36,48} = 20,66\%$$

79,34% de la variance est expliqué par les sites, donc au minimum un site est différents des autres

ANOVA – Etape 9

Faire la supposition pour l'erreur (analyse des résidus)



Erreurs sont indépendantes et distribuées normalement

Plan en blocs aléatoires

- Un plan en blocs randomisés est un type de plan couramment utilisé pour réduire l'effet de la variabilité lorsqu'elle est associée à des unités discrètes (par exemple, emplacement, opérateur, usine, lot, date). Le cas le plus fréquent consiste à randomiser une réplique de chaque combinaison de traitements dans chaque bloc. En général, les blocs ne présentent pas d'intérêt intrinsèque et sont considérés comme des facteurs aléatoires. La supposition habituelle est que l'interaction bloc par traitement est nulle et devient le terme d'erreur pour tester les effets des traitements. Si vous appelez la variable de bloc "Bloc", les termes du modèle seraient Bloc, A, B et A*B. Vous spécifieriez également cette variable Bloc en tant que facteur aléatoire

Math Model:

$$Y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti}$$

Hypotheses:

$$H_o : \beta_i's = 0 \qquad H_o : \tau_t's = 0$$

$$H_a : \beta_i's \neq 0 \qquad H_a : \tau_t's \neq 0$$

Example – Step 1

- **Problème pratique: un ingénieur de production veut tester l'effet sur la productivité de plusieurs types d'ingrédients. Le test est effectué selon un standard et en mesurant le temps qui va prendre chaque types. L'expérience sera faite en utilisant 04 différents opérateurs.**

- **L'ingénieur sait que les opérateurs vont être une source de variabilité qui va impacté la différence entre les types d'ingrédients. Le facteur opérateur va être considéré comme un bloc aléatoire**

Operator	TypeA	TypeB	TypeC	TypeD
1	15.5	14.4	16.2	15.0
2	18.7	17.3	18.1	17.7
3	16.2	16.0	16.8	15.5
4	14.1	14.5	15.1	13.7

Example – Step 3

- **Hypothèse nulle & hypothèse alternative:**

$$Y_{ti} = \mu + \beta_i + \tau_t + \varepsilon_{ti}$$

Operator Effect:

$$H_o : \beta_i's = 0$$
$$H_a : \beta_i's \neq 0$$

Block

Type Effect:

$$H_o : \tau_t's = 0$$
$$H_a : \tau_t's \neq 0$$

Treatment

ANOVA – Etape 7-8

Analyse Anova (facteur Opérateur non inclus)

$$\sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{\bar{x}})^2 = \sum_{j=1}^{g} n*(\bar{x}_j-\bar{\bar{x}})^2 + \sum_{j=1}^{g}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

Opérateur	Type1	Type2	Type3	Type4
1	15,5	14,4	16,2	15
2	18,7	17,3	18,1	17,7
3	16,2	16	16,8	15,5
4	14,1	14,5	15,1	13,7
	Y_{1bar}	Y_{2bar}	Y_{3bar}	Y_{4bar}
	16,13	15,55	16,55	15,475
	Y_{barbar}			15,93

Somme des carrées erreur

Type1	Type2	Type3	Type4
0,391	1,323	0,122	0,226
6,631	3,063	2,403	4,951
0,006	0,202	0,063	0,001
4,101	1,103	2,102	3,151
$\sum(Yij - Ybar)^2$			29,84

Somme des carrées total

Type1	Type2	Type3	Type4
0,181	2,326	0,076	0,856
7,701	1,891	4,731	3,151
0,076	0,006	0,766	0,181
3,331	2,031	0,681	4,951
$\sum\sum(Yij - Ybarbar)^2$			32,930

Calcul somme des carrées dû aux facteurs

Type1	Type2	Type3	Type4
0,160	0,562	1,562	0,810
$\sum N*(Y_{ibar} - Y_{barbar})^2$			3,09

Exemple – Step 7-8

- **Construction du table ANOVA (opérateur non inclus):**

Source	Df	SS	MS	F	P
Facteurs	3	3,09	1,03	0,41	**0,75**
Erreur	12	29,84	2,49		
Total	15	32,93			

$$\varepsilon^2_{factor} = \frac{SS_{Between}}{SS_{Total}}$$

Efacteur = 9,40%

$$\varepsilon^2_{error} = \frac{SS_{Error}}{SS_{Total}}$$

Eerreur= 90,60%

La valeur-p-value est largement supérieur à 5% ce qui nous amène à accepter l'hypothèse nulle et conclure qu'il n'y a pas de différence entre les types des ingrédients. Cette conclusion est elle réellement vraie?

Exemple – Step 7-8

Construction du table ANOVA (opérateur inclus):

Source	Df	SS	MS	F	P
Type	3	3,09	1,03	5,59	0,0192
Opérateur	3	28,18	9,39	50,92	0,00
Erreur	9	1,66	0,18		
Total	15	32,93			

La valeur-p-value est très faible comparée à 5% ce qui nous amène à rejeter l'hypothèse nulle et conclure qu'il y a une différence significative entre les types des ingrédients

$$\varepsilon^2_{opérateur} = \frac{SS_{Between}}{SS_{Total}}$$

Eopérateur = 85,56%

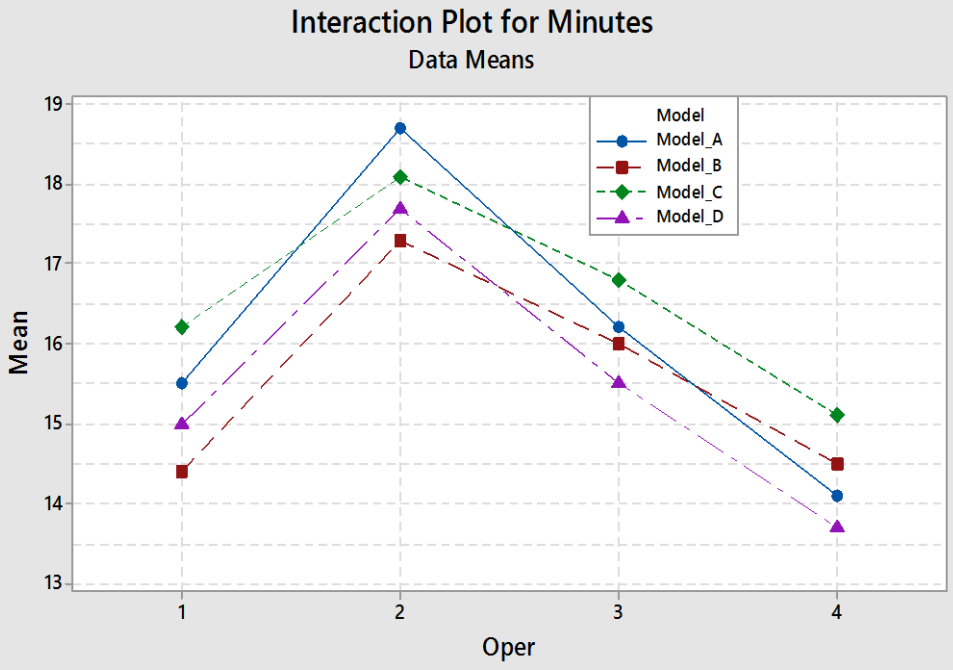
$$\varepsilon^2_{Type} = \frac{SS_{Between}}{SS_{Total}}$$

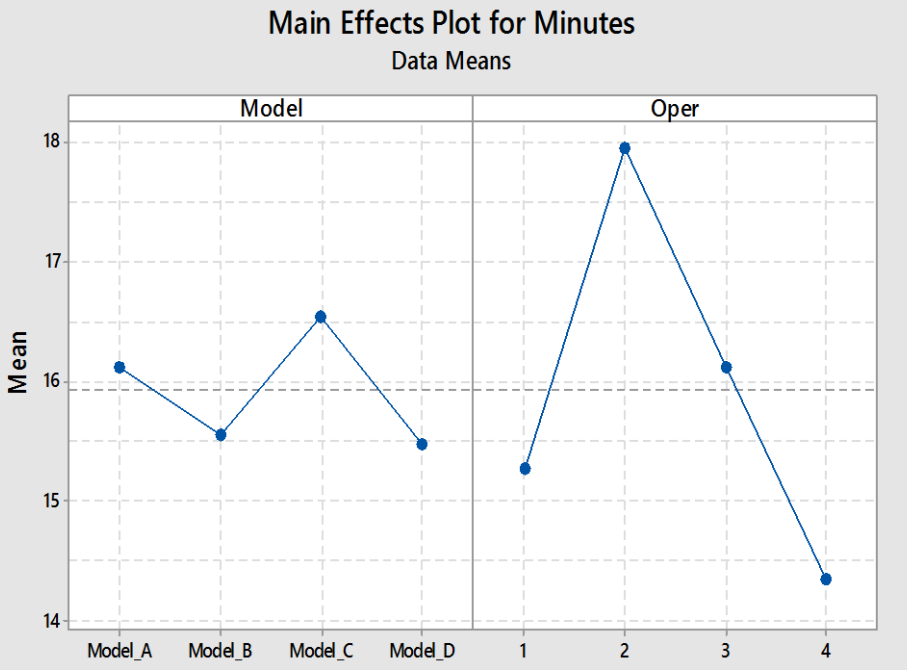
Efacteur = 9,40%

$$\varepsilon^2_{error} = \frac{SS_{Error}}{SS_{Total}}$$

Eerreur= 5,04%

Exemple – Step 7-8





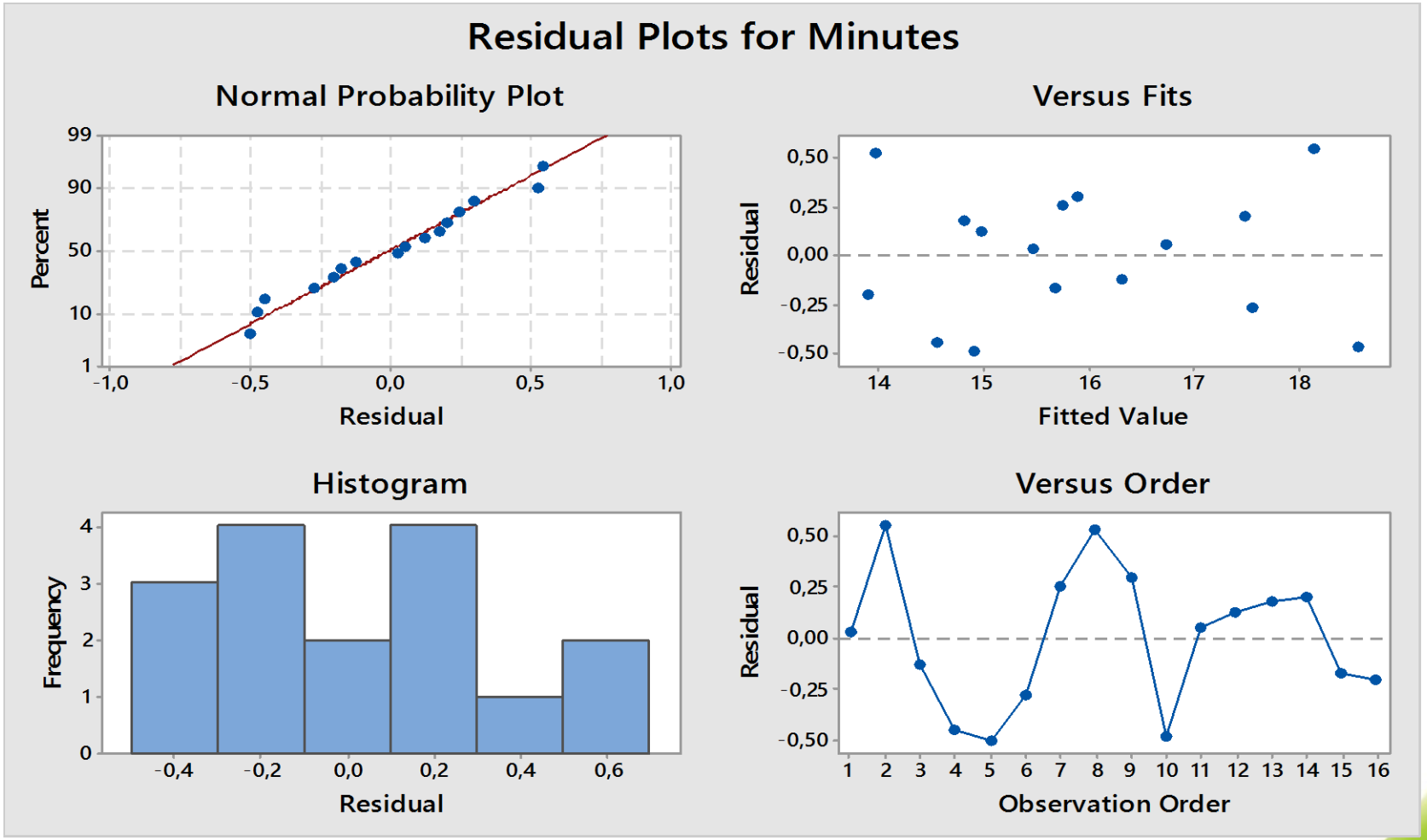
Interaction: Le croisement des courbes signifie l'existence d'une interaction entre les opérateurs et les ingrédients

L'effet principale montre une grande variabilité dans le facteur opérateur comparé aux ingrédients

Exemple – Step 9

- **Analyse des graphes du résidu**

Erreurs sont indépendantes et distribuées normalement



Questions?