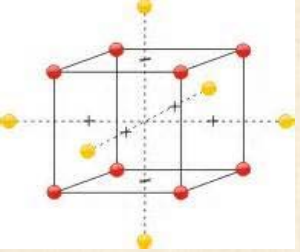


Génistat Conseils

POLYTECHNIQUE MONTRÉAL

Approche globale pour entreprendre une étude statistique

Bernard CLÉMENT, PhD

P L A N (*)

(*) cette présentation est partiellement basée sur des documents provenant de mon site WEB

http://www.groupes.polymtl.ca/mth6301/

Statistique : définition - applications

Statistique

- collecte, analyse, interprétation, présentation, visualisation de données numériques et autres
- augmenter la connaissance avec des données en vue de …

But prendre les meilleures décisions basées sur des **données existantes** (historiques) ou données **à recueillir** dans des conditions d'**incertitude** et de **variabilité**

Applications : TOUS les domaines de l'activité humaine

Anthropologie, Biologie, Criminologie, …Ingénierie….. , Zoologie

science interdisciplinaire par excellence !

" *Statistical thinking* will one day be as necessary for efficient citizenship as the ability to read and write. " H.G. Wells (1866-1946)

" I keep saying the *sexy job in the next ten years will be statisticians*. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s ? "

Hal Varian, PhD, chief economist, Google The McKinsey Quarterly 2009

QUELQUES APPLICATIONS DE LA STATISTIQUE

Industrie et affaires

- détection de fraudeurs d'appels téléphoniques
- réduction des « pourriels »
- évaluation de la fiabilité et la sécurité de produit
- analyse des cotes boursières
- optimisation des procédés de fabrication
- conception (DESIGN) d'un nouveau produit
- contrôle de la qualité des produits
- data mining des banques de données etc …

Science et technologie / ingénierie / biotechnologie

- identification des courriels indésirables
- concentration du radon dans une résidence
- prédictions métrologiques
- variations climatiques etc …

Médecine / sciences de la vie / pharmaceutique

- modélisation du déclanchement d'une attaque d'Anthrax
- étude sur la propagation du SIDA
- identification DNA / génomiques

Science sociales et humaines …. etc …

PROCESSUS et MÉTHODES STATISTIQUES

PROCESSUS → VARIABILITÉ → DONNÉES → AMÉLIORATION

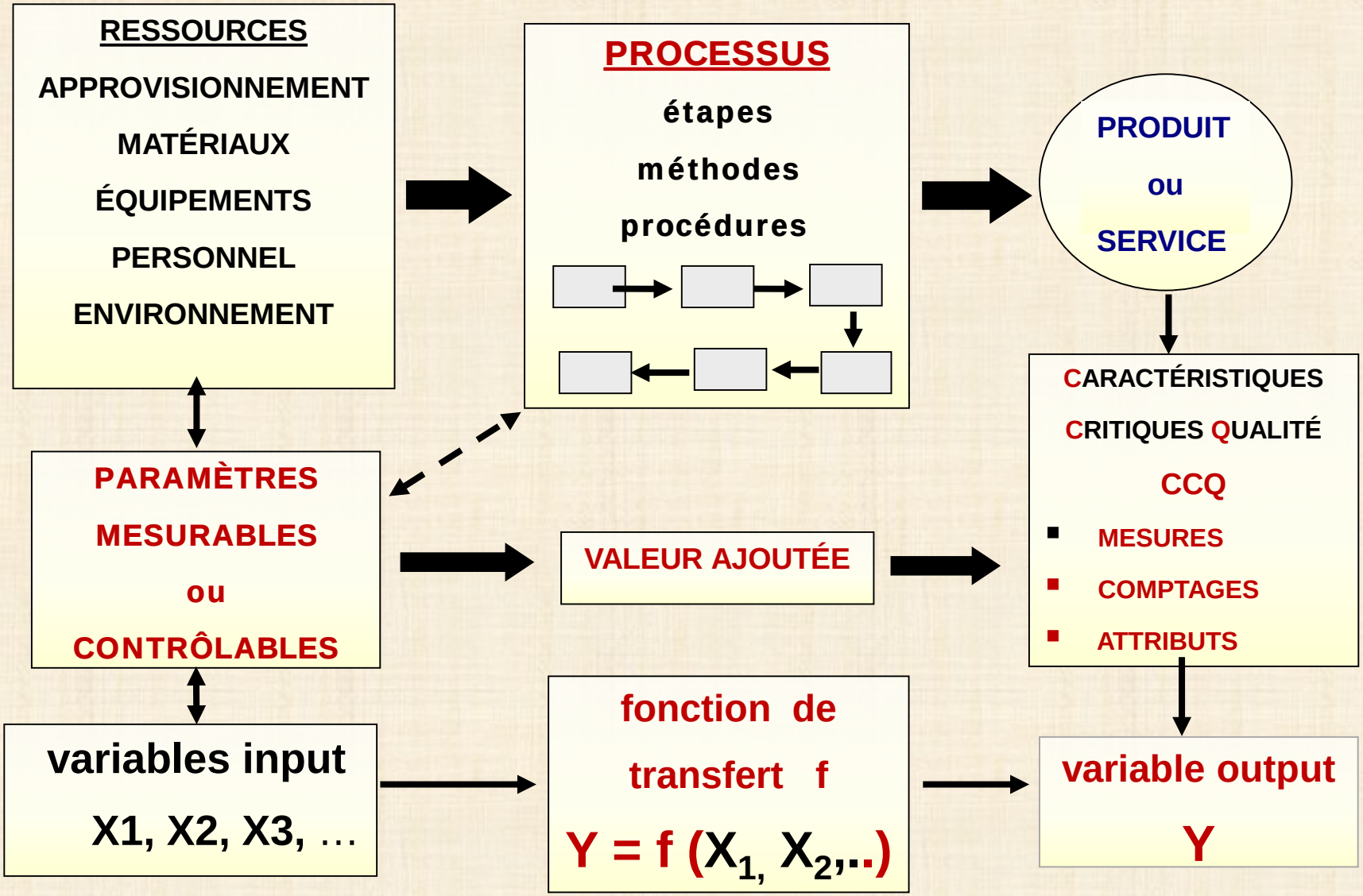
PENSÉE STATISTIQUE ⟹ MÉTHODES STATISTIQUES

FOURNISSEURS → PROCESSUS 1 → PROCESSUS 2 → . . . . → CLIENTS

TRAVAIL = EST UN SYSTÈME DE PROCESSUS INTERDÉPENDANTS

- LA VARIABILITÉ EXISTE DANS TOUS LES PROCESSUS
- LA CLÉ : COMPRENDRE et RÉDUIRE LA VARIABILITÉ
- L'ÉTUDE de la VARIABILITÉ → MÉTHODES STATISTIQUES

PROCESSUS/SYSTÈME



PROCESSUS (= SYSTÈME) : S I P O C

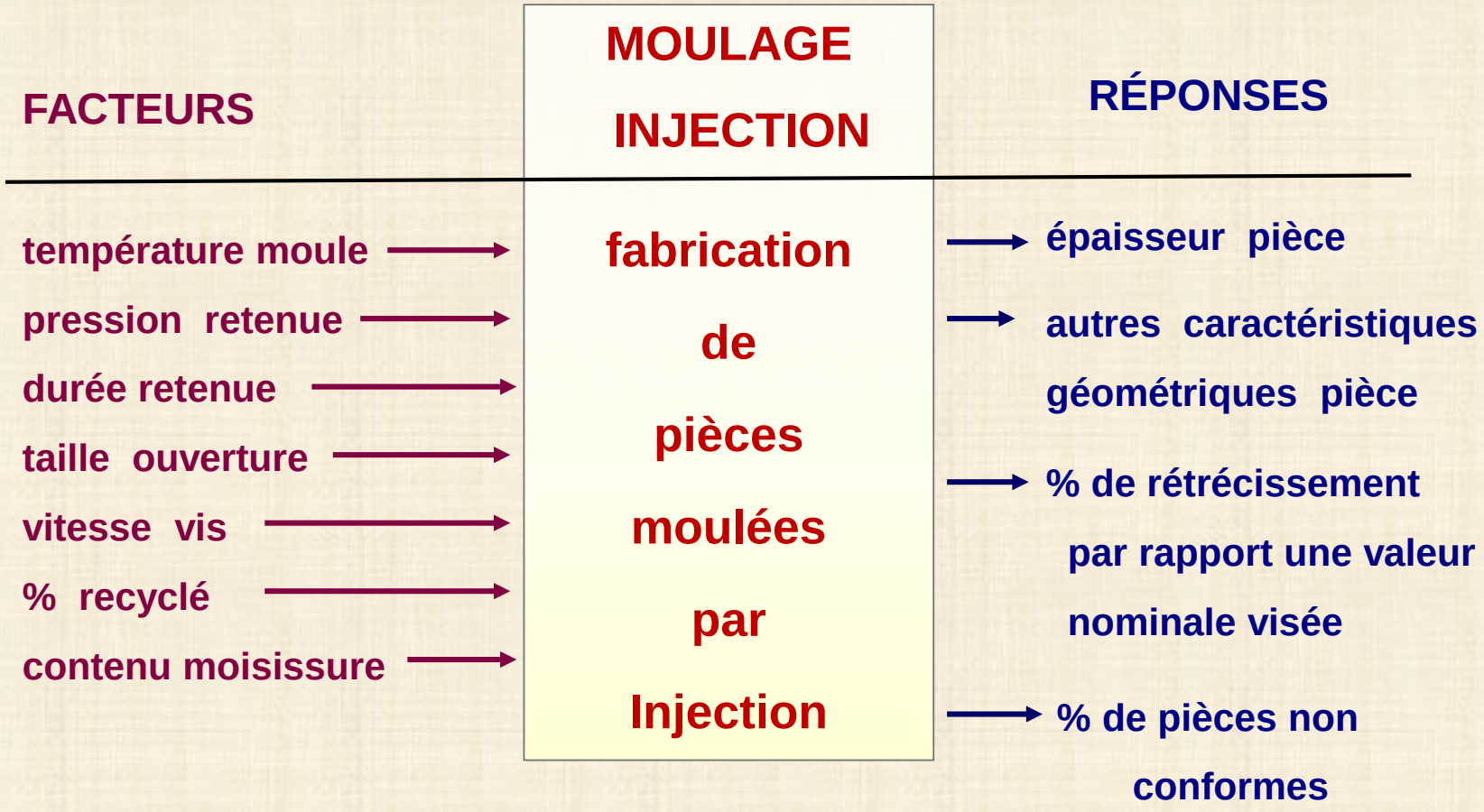
Suppliers	Input : X	Processus	Output : Y	Clients

FACTEURS	GÉNÉRAL	RÉPONSES
Personnel →	mélange	indicateurs en relation avec
Matériaux →	d'inputs	
Équipement →	qui	→ qualité produit
Politiques →	génèrent	
Procédures →	un	→ qualité service
Méthodes →		
Environnement →	output	→ exécution tâche

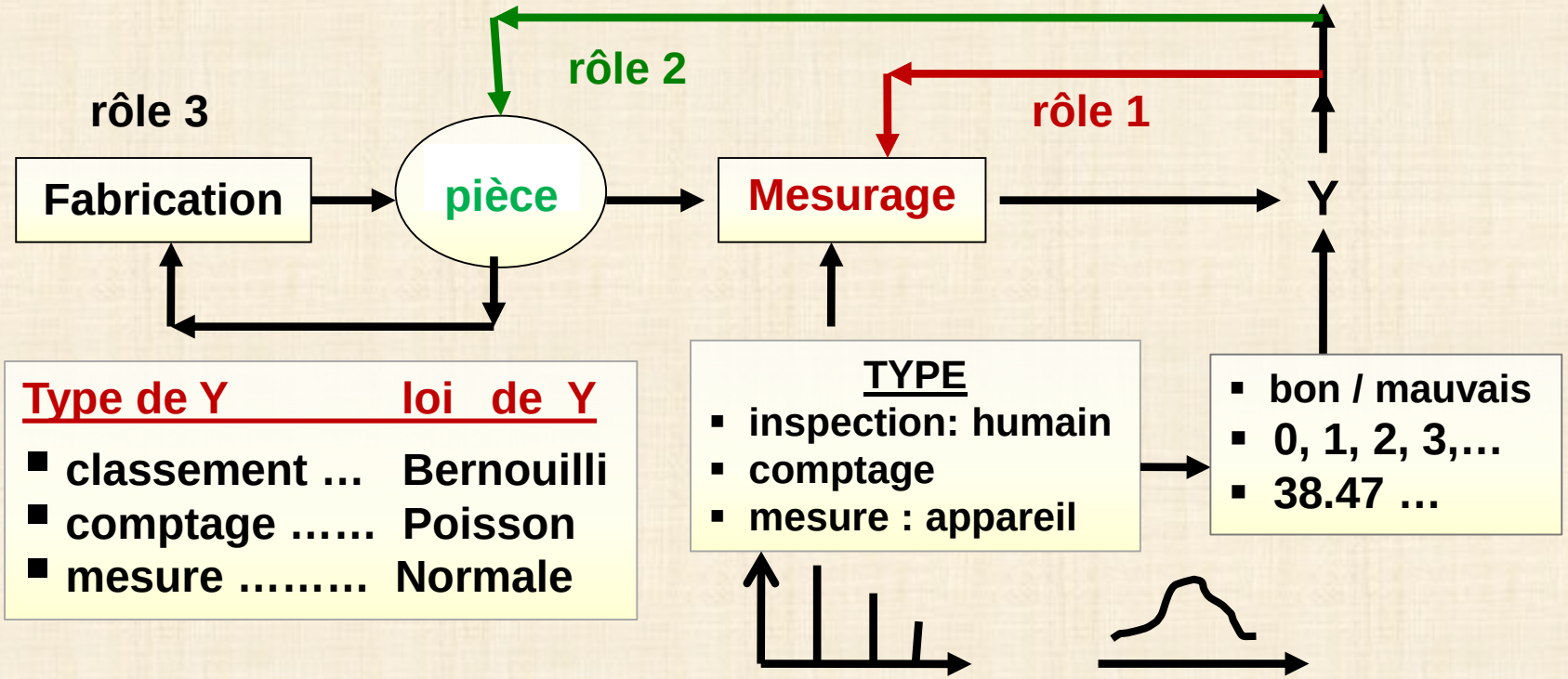
CATÉGORIE de PROCESSUS / SYSTÈME

- DESIGN (CONCEPTION) : produit / procédé
- FABRICATION
- MESURAGE
- TRANSACTIONEL / ADMINISTRATIF

PROCÉDÉ F A B R I C A T I O N : exemple



2 PROCESSUS INSÉPARABLES: Fabrication + Mesurage



Type de Y	loi de Y
classement …	Bernouilli
comptage ……	Poisson
mesure ……	Normale

TYPE
- inspection: humain
- comptage
- mesure : appareil

- bon / mauvais
- 0, 1, 2, 3,…
- 38.47 …

rôle des données

rôle 1 Analyser le processus de mesurage : étude R&R
reproductible? répétable? erreur mesure =?

rôle 2 classer la pièce: conforme? non conforme?

rôle 3 Analyser le processus de fabrication:
étude de capabilité C_{pk} =? stable? capable?

Type d'études statistiques

CARACTÉRISTIQUE	OBSERVATIONNELLE (mode passif)	EXPÉRIMENTALE (mode actif)
provenance des données	- historiques - suivi de processus dans des conditions normales d'opération; -le processus n'est pas manipulé (volontairement perturbé);	on fait varier le processus sous différentes conditions (variables) dans le but d'obtenir des données
quantité de données	-généralement abondantes; -on peut obtenir des observations additionnelles	fixe et limitée
qualité de données	peut présenter des difficultés : changements non documentés, données manquantes etc.	excellente
coût	généralement faible	généralement élevé
but	modélisation et exploration	détecter des changements
hypothèse sous-jacente	homogénéité des données (*)	hétérogénéité causée par les perturbations induites
méthodes d'analyse	- carte données individuelles et étendues mobiles XmR; - carte Xbar&R ou Xbar&S afin de vérifier l'homogénéité des données - méthodes de Data Mining	- analyse de la variance - analyse de régression - autres méthodes - cartes XmR, Xbar&R,…

* **L'homogénéité des données est fondamentale à l'étape de l'analyse.**
Cette question est clarifiée dans l'article suivant : Wheeler, Donald J. (2009) *The four Questions of Data Analysis*
http://www.qualitydigest.com/inside/quality-insider-column/four-questions-data-analysis.html

Type d'études statistiques

mode actif

mode passif

Expérience planifiée
traitements appliqués
aux unités expérimentales
selon un protocole (design)

structure traitements

protocole expérimental

- complètement aléatoire
- blocs randomisés
- autres…

MTH8301 Planification et analyse d'expériences

Sondages, enquêtes, recensements
= études énumératives

plan d'échantillonnage
des unités statistiques
pas de traitements
appliqués aux unités

sciences humaines
sciences sociales

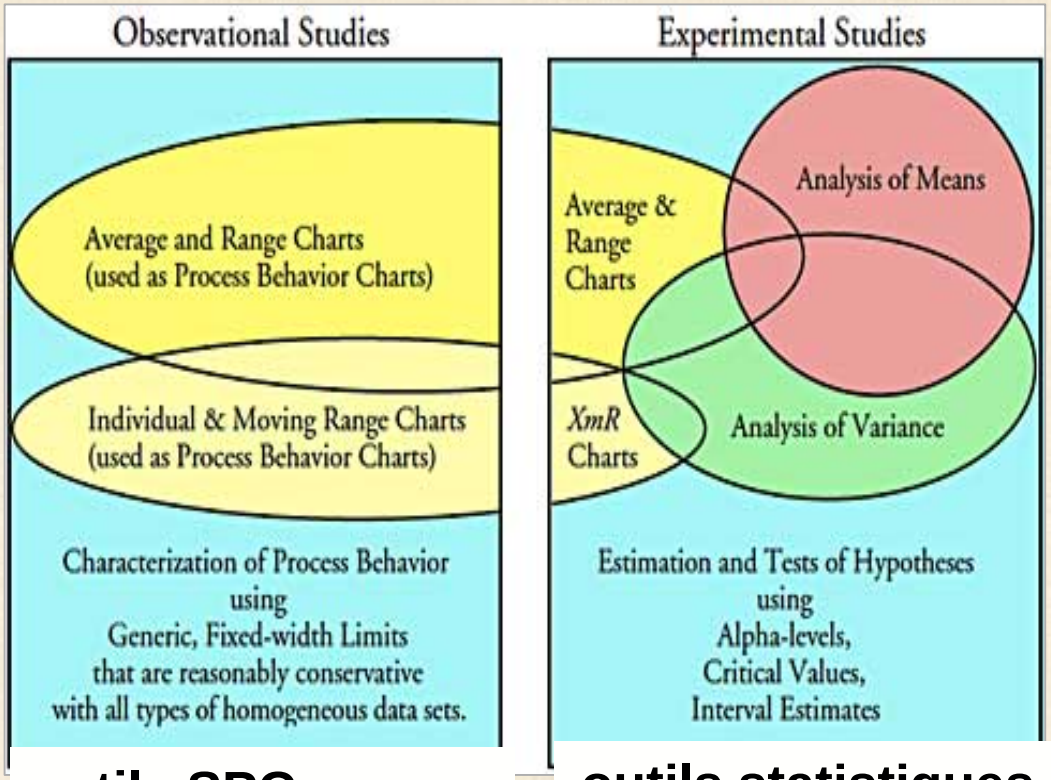
Étude observationnelle
données collectées
au fil du temps
(instants d'observations)
sans une très grande planification statistique

banque de données

mégadonnées (big data)

Type d'études statistiques

observationnelle **expérimentale**



outils SPC
cartes comportement processus

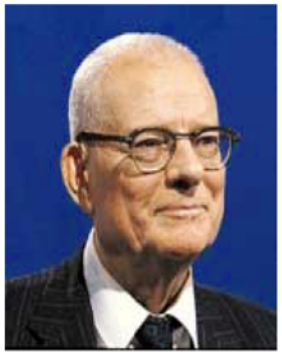
outils statistiques
tests, ANOVA, régression, etc.

énumérative **analytique**

Two Types of Studies

Enumerative Studies:	Analytic Studies:
Census	Selection of suppliers
Inventory	Poll to determine strategy
Exit survey at polls	Experiment to improve performance
Acceptance sampling	Drug testing

W. E. Deming's Two Types of Studies



Chapter 7 from *Some Theory of Sampling*, 1950

The aim of any experiment is to provide a rational basis for action

Enumerative study: an experiment in which action will be taken on the universe.

Analytic study: an experiment in which action will be taken on a cause system to improve performance in the future.

Donald J. Wheeler PhD
Statistics and SPC 03/03/2014
http://www.qualitydigest.com/

études statistiques en industrie: majoritairement ANALYTIQUES

ÉTUDES STATISTIQUES

SPC	INFÉRENCE STATISTIQUE
ÉTUDES OBSERVATIONNELLES	ÉTUDES EXPÉRIMENTALES
Caractérisation du comportement du processus	Analyse des données une seule fois
Utilisation de limites de contrôle conservatrices	Estimation des paramètres de modèles statistiques
Découvertes facteurs inconnus et perturbateurs	Développement de modèles prédictifs structures internes

Beaucoup de confusion sur l'usage des méthodes :
le SPC est complètement mis de côté à l'ère des mégadonnées (BigData)
alors que la majorité des données sont de nature observationnelles.
On analyse toujours avec des méthodes pour les études expérimentales …

SPC

plus de détails

CONSTATS UNIVERSELS

- La qualité du produit dépend du processus.
- Le processus doit être étudié avec le produit.
- Le comportement du processus varie dans le temps

La VARIABILITÉ est TOUJOURS PRÉSENTE

- Sans surveillance, TOUS les processus se désorganisent et se dégradent à cause de l' ENTROPIE

Pour s'en sortir, une solution qui a fait ses preuves :

CARTES de CONTRÔLE des PROCESSUS

remarque : le terme *CONTRÔLE* prête à **beaucoup de confusion**.

Les cartes ne contrôlent pas le processus mais elles donnent une image du COMPORTEMENT du processus par l'intermédiaire de mesures sur le produit. Il serait préférable d'appeler ces cartes :

cartes de comportement du processus

résultats que l'on obtient avec les cartes

- analyser les fluctuations de Y
- quantifier ces fluctuations
- comprendre deux catégories de variabilité
- réduire la variabilité
- statuer si le processus est STABLE (concept à définir)
- évaluer la capacité du processus à l'aide d'indices (à définir) relativement à des limites de spécification (tolérances)

très important

RÉSUMÉ

- cartes de contrôle sont des graphiques du comportement du processus
- BILAN de santé du PROCESSUS

http://www.groupes.polymtl.ca/mth6301/mth8301/Clement/Clement-Introduction_SPC.pdf

Planification d'expériences

L'expérimentation (série de tests) est nécessaire

- caractériser et optimiser les procédés
- évaluer les propriétés des matériaux / designs / systèmes
- déterminer les tolérances des composantes / systèmes
- réduire temps pour le design des produits
- qualifier les procédés de fabrication
- améliorer la fiabilité des produits
- obtenir des produits et des procédés robustes

Toutes les expériences sont planifiées mais

- beaucoup sont mal planifiées
- certaines sont bien planifiées en utilisant la planification statistique des essais

DOE : Design Of Experiment

Planification d'expériences

PHASE	ÉTAPES
P: planifi - cation	1 Définir PROCESSUS / problématique / objectifs
	2 Choisir les variables de RÉPONSE (S) Y à mesurer
	3 Choisir les VARIABLES facteurs X et l'espace de variation
	4 Choisir et comparer des PLANS EXPÉRIMENTAUX
D:exécution	5 PRÉPARER pour l'expérience
	6 CONDUIRE l'expérience
S: analyse	7 ANALYSE statistique des résultats
A: transfert	8 AGIR avec les conclusions de l'analyse

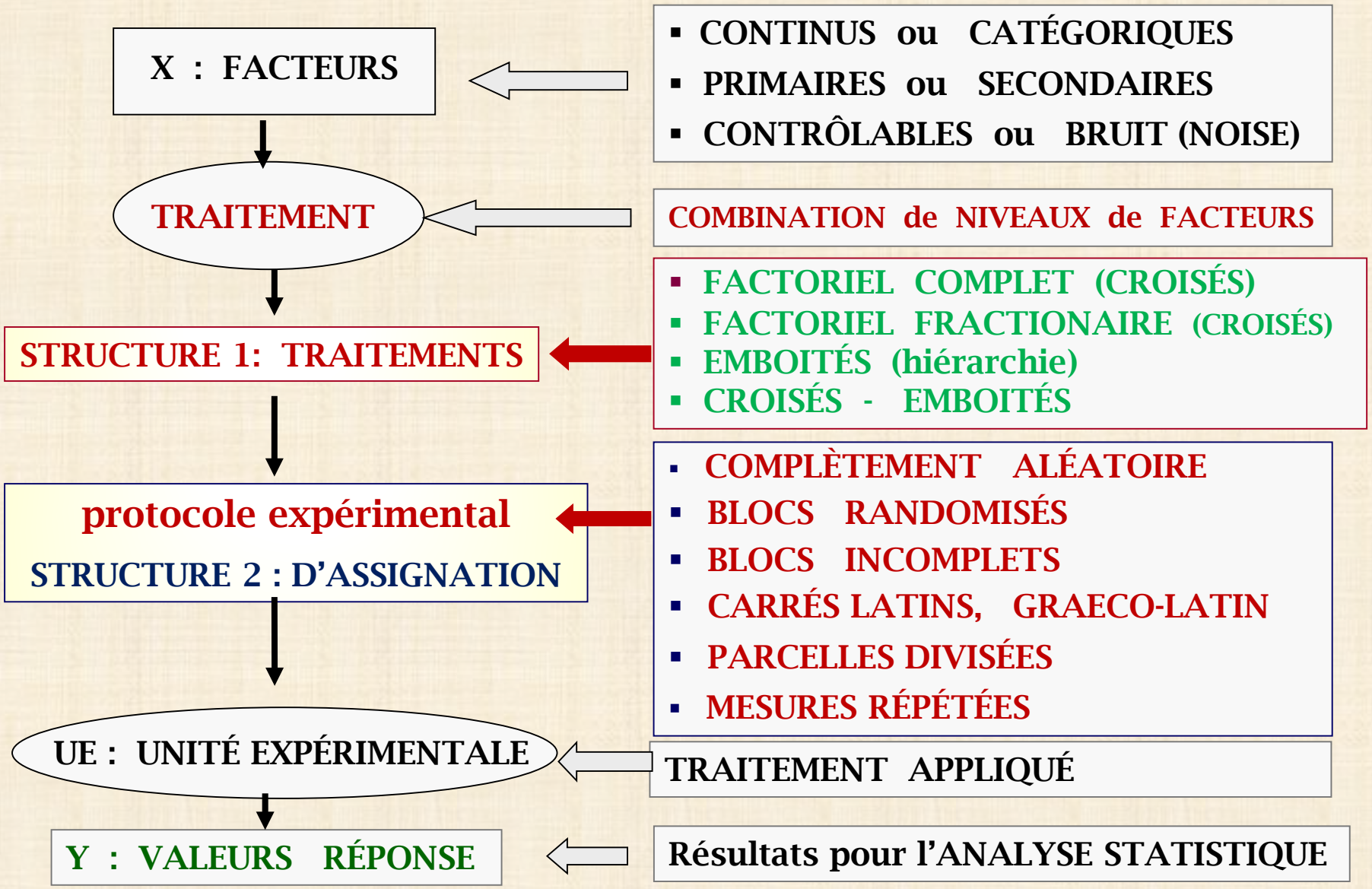
quel plan ?

comment ?

P D S A

Shewhart - Deming

EXPÉRIMENTATION : 2 structures



ÉTUDE EXPÉRIMENTALE : exemple

Influence de 5 variables (A, B, C, D, E) sur une réponse Y_WetAngle
p = 6 variables n = 32 essais (observations)

S. Bisgaard - J. Q. Tech. vol 32 (2000) no 1 p. 39-56
Box, Bisgaard, Quality Engineering, vol 8 (1996) no 4, p. 705-708
Les facteurs A B C D définissent les WholePlot - E est le facteur SplitPlot

	1 StdOrder	2 RunOrder	3 bloc	4 WholePlot	5 A_Pressure	6 B_Power	7 C_GasFlow	8 D_GasType	9 E_PaperType	10 Y_WetAngle
1	5	1	1	1	-1	-1	1	Oxygen	E1	37,6
2	21	2	1	1	-1	-1	1	Oxygen	E2	43,5
3	2	3	1	2	1	-1	-1	Oxygen	E1	41,2
4	18	4	1	2	1	-1	-1	Oxygen	E2	38,2
5	10	5	1	3	1	-1	-1	SiCl4	E1	56,8
6	26	6	1	3	1	-1	-1	SiCl4	E2	56,2
7	14	7	1	4	1	-1	1	SiCl4	E1	47,5
8	30	8	1	4	1	-1	1	SiCl4	E2	43,2
9	11	9	1	5	-1	1	-1	SiCl4	E1	25,6
10	27	10	1	5	-1	1	-1	SiCl4	E2	33,0
11	3	11	1	6	-1	1	-1	Oxygen	E1	55,8
12	19	12	1	6	-1	1	-1	Oxygen	E2	62,9
13	13	13	1	7	-1	-1	1	SiCl4	E1	13,3
14	29	14	1	7	-1	-1	1	SiCl4	E2	23,7
15	6	15	1	8	1	-1	1	Oxygen	E1	47,2
16	22	16	1	8	1	-1	1	Oxygen	E2	44,8
17	16	17	1	9	1	1	1	SiCl4	E1	49,5
18	32	18	1	9	1	1	1	SiCl4	E2	48,2
19	9	19	1	10	-1	-1	-1	SiCl4	E1	5,0
20	25	20	1	10	-1	-1	-1	SiCl4	E2	18,1
21	15	21	1	11	-1	1	1	SiCl4	E1	11,3
22	31	22	1	11	-1	1	1	SiCl4	E2	23,9
23	1	23	1	12	-1	-1	-1	Oxygen	E1	48,6
24	17	24	1	12	-1	-1	-1	Oxygen	E2	57,0
25	8	25	1	13	1	1	1	Oxygen	E1	48,7
26	24	26	1	13	1	1	1	Oxygen	E2	44,4
27	7	27	1	14	-1	1	1	Oxygen	E1	47,2
28	23	28	1	14	-1	1	1	Oxygen	E2	54,6
29	4	29	1	15	1	1	-1	Oxygen	E1	53,5
30	20	30	1	15	1	1	-1	Oxygen	E2	51,3
31	12	31	1	16	1	1	-1	SiCl4	E1	41,8
32	28	32	1	16	1	1	-1	SiCl4	E2	37,8

ÉTUDE OBSERVATIONNELLE : exemple

Influence de 13 caractéristiques physico-chimique sur la qualité du vin (Portugal) p = 13 variables n = 6496 observations

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine.
1599 red wine 4898 white wine
http://www.vinhoverde.pt/en/
Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available.
(e.g. there is no data about grape types, wine brand, wine selling price, etc.).
The inputs include objective tests (e.g. pH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts).
Each expert graded the wine quality between 0 (very bad) and 10 (excellent).
Data mining methods were applied to model the dataset under a regression approach.
Variables
input variables X (based on physicochemical tests)
1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 residual sugar
5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide
8 - density 9 - pH 10 - sulphates 11 - alcohol
Output variable Y (based on sensory data) Y = QUALITY : score between 0 (=very bad) and 10 (=excellent)

	1 ID6497	2 QUALITY	3 color	4 fixed acidity	5 volatile acidity	6 citric acid	7 residual sugar	8 chlorides	9 free sulfur dioxide	10 total sulfur dioxide	11 density	12 pH	13 sulphates	14 alcohol
1	1	5	red	7,4	0,70	0,00	1,90	0,076	11	34	0,9978	3,51	0,56	9,4
2	2	5	red	7,8	0,88	0,00	2,60	0,098	25	67	0,9968	3,20	0,68	9,8
3	3	5	red	7,8	0,76	0,04	2,30	0,092	15	54	0,9970	3,26	0,65	9,8
4	4	6	red	11,2	0,28	0,56	1,90	0,075	17	60	0,9980	3,16	0,58	9,8
5	5	5	red	7,4	0,70	0,00	1,90	0,076	11	34	0,9978	3,51	0,56	9,4
6	6	5	red	7,4	0,66	0,00	1,80	0,075	13	40	0,9978	3,51	0,56	9,4
7	7	5	red	7,9	0,60	0,06	1,60	0,069	15	59	0,9964	3,30	0,46	9,4
8	8	7	red	7,3	0,65	0,00	1,20	0,065	15	21	0,9946	3,39	0,47	10,0
9	9	7	red	7,8	0,58	0,02	2,00	0,073	9	18	0,9968	3,36	0,57	9,5

	1 ID2	2 ID	3 couleur	4 fixed_acidity	5 volatile_acidity	6 citric_acid	7 residual_sugar	8 chlorides	9 free_sulfur_dioxide	10 total_sulfur_dioxide	11 density	12 pH	13 sulphates	14 alcohol
6475	6475	5684	blanc	6,5	0,33	0,24	14,50	0,048	20,0	96	0,99456	3,06	0,30	11,50
6476	6476	5718	blanc	6,5	0,43	0,31	3,60	0,046	19,0	143	0,99022	3,15	0,34	12,00
6477	6477	5767	blanc	6,3	0,17	0,32	1,00	0,040	39,0	118	0,98886	3,31	0,40	13,10
6478	6478	5795	blanc	7,1	0,45	0,24	2,70	0,040	24,0	87	0,98862	2,94	0,38	13,40
6479	6479	5932	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6480	6480	5933	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6481	6481	5934	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6482	6482	5935	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6483	6483	5936	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6484	6484	5937	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6485	6485	5938	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6486	6486	5939	blanc	6,8	0,24	0,29	2,00	0,044	15,0	96	0,99232	3,23	0,64	10,40
6487	6487	5940	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6488	6488	6365	blanc	5,2	0,30	0,34	1,50	0,038	18,0	96	0,98942	3,56	0,48	13,00
6489	6489	6366	blanc	6,4	0,32	0,25	5,00	0,055	28,0	138	0,99171	3,27	0,50	12,40
6490	6490	6386	blanc	4,4	0,32	0,39	4,30	0,030	31,0	127	0,98904	3,46	0,36	12,80
6491	6491	6387	blanc	3,9	0,23	0,40	4,20	0,030	29,0	118	0,98900	3,57	0,36	12,80
6492	6492	6402	blanc	5,8	0,28	0,34	2,20	0,037	24,0	125	0,98986	3,36	0,33	12,80
6493	6493	2374	blanc	9,1	0,27	0,45	10,60	0,035	28,0	124	0,99700	3,20	0,46	10,40
6494	6494	2420	blanc	6,6	0,36	0,29	1,60	0,021	24,0	85	0,98965	3,41	0,61	12,40
6495	6495	2427	blanc	7,4	0,24	0,36	2,00	0,031	27,0	139	0,99055	3,28	0,48	12,50
6496	6496	2476	blanc	6,9	0,36	0,34	4,20	0,018	57,0	119	0,98980	3,28	0,36	12,70

Principes pour concevoir des études expérimentales / analytiques de bonne qualité scientifique

1. Objectifs clairement définis
2. Périmètre expérimental nettement établi
3. Séparation des effets de chaque facteur
4. Élimination de biais
5. Précision suffisante pour atteindre le but
6. Décomposition des sources de variabilité
7. Approche séquentielle
8. Exécution sans embûches

approche processus

X : entrées → PROCESSUS → Y : sorties / réponse

Quelles sont les variables CRITIQUES X affectant les variables de réponse Y ?	IDENTIFICATION
Quelle est la FONCTION de TRANSFERT f entre les variables critiques X et la variable de réponse variable Y ?	MODÉLISATION f X → Y = f (X)
Comment CONTRÔLER la réponse Y à un niveau désiré nominal - maximum - minimum en fixant les variables X à des niveaux spécifiques ?	CONTRÔLE et OPTIMISATION

ENTREPRENDRE UN PROJET D'EXPÉRIMENTATION (1/5)

1. Décrire sommairement (diagramme de flux) du processus (produit ou procédé de fabrication) qui fera l'objet du projet d'expérimentation.
   - produit : conception / re conception / modification
   - procédé fabrication : conception / modification
2. Définir le but principal de l'expérience (penser aux réponses) et les objectifs associés.
3. Identifier la ou les variables de réponse (variables output du processus).
4. Identifier l'ensemble de tous les facteurs pouvant affecter la (les) variables de réponse.
   suggestion : diagramme d'Ishikawa →
5. identifier les facteurs qui seront maintenus constants au cours des essais.
6. Identifier les facteurs (variables primaires) que l'on fera varier au cours des essais.

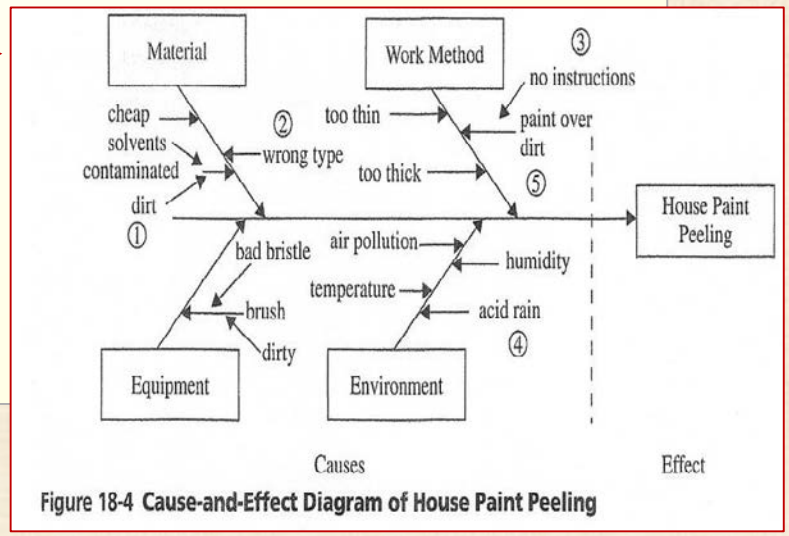


Figure 18-4 Cause-and-Effect Diagram of House Paint Peeling

ENTREPRENDRE UN PROJET D'EXPÉRIMENTATION (2/5)

7. Préciser s'il y a des facteurs (variables) secondaires nuisibles connus qui seront contrôlés. remarque : différence entre 6 et 7 ?
   Facteurs primaires : ceux qui sont à l'origine du projet
   Facteurs secondaires nuisibles : ceux qui varient en cours d'expérimentation et que l'on ne peut pas maintenir constants.
   Si on peut les contrôler (fixer) alors on peut construire un plan en blocs qui permet de neutraliser leurs effets réels ou non sur la réponse.

8. Identifier les variables (facteurs) non contrôlées mais que l'on peut mesurer. Ils sont tenu en compte à titre de covariables lors de l'analyse des données de l'expérience.

9. Préciser la valeur minimale et la valeur maximale (intervalle de variation) de chaque facteur primaire quantitatif que l'on fera varier au cours des essais.
   remarque : explorer le plus grand espace possible mais éviter les régions problématiques

10. Préciser la liste des modalités de chaque facteur primaire qualitatif.
    remarque : 9 et 10 constitue l'espace d'expérimentation qui sera explorer avec les essais

ENTREPRENDRE UN PROJET D'EXPÉRIMENTATION (3/5)

11. Anticiper la relation (augmentation / diminution) de la réponse avec chaque facteur
12. Préciser comment seront mesurés les variables de réponse.

    remarque : une étude du processus de mesurage est-elle nécessaire?
13. Selon l'état des connaissances sur le processus, proposer un ou plusieurs plans :

    plan de tamisage pour séparer les facteurs importants et ceux qui ne le sont pas

    séparation claire et nette des effets principaux et des effets d'interaction

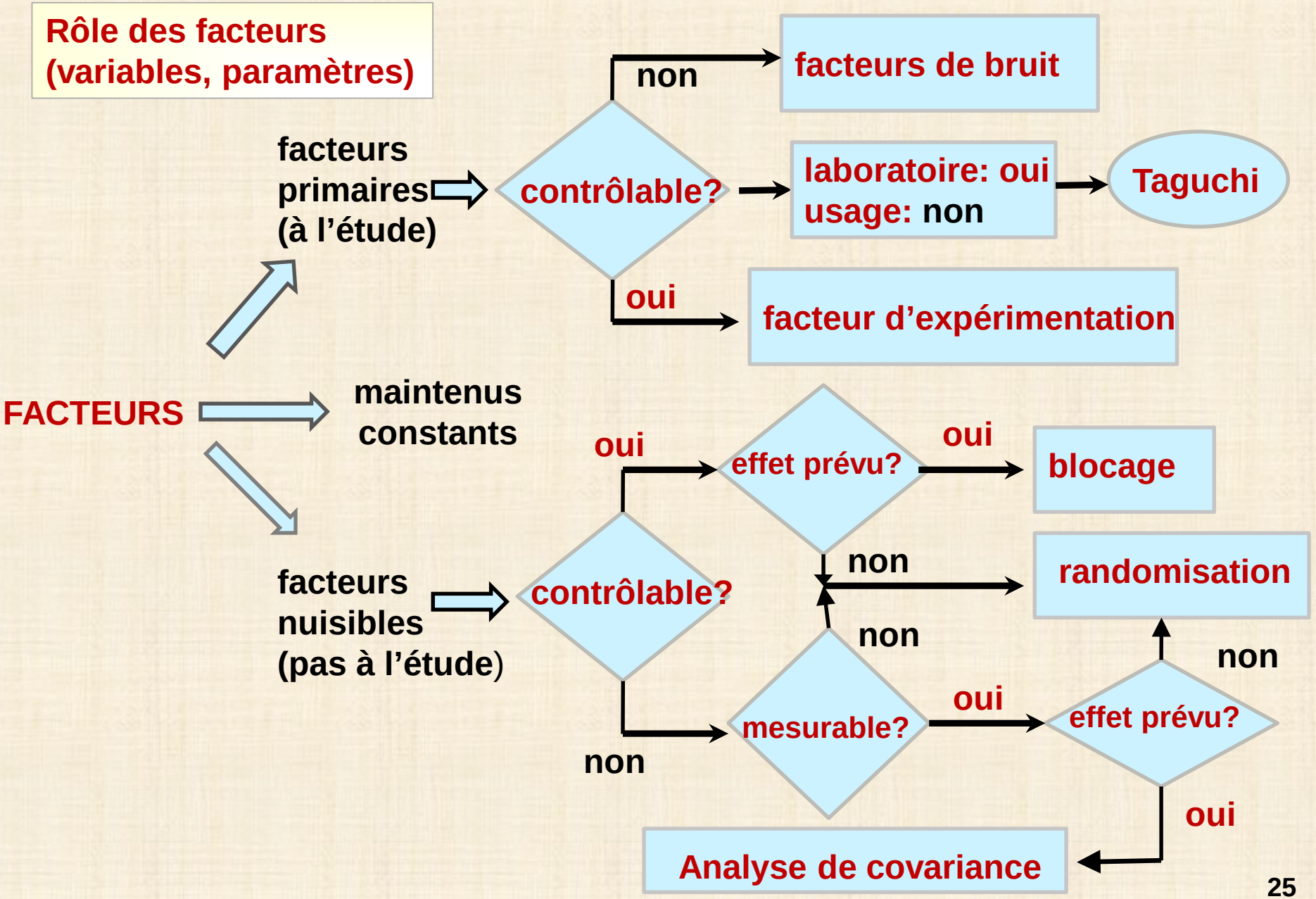
    plan pour l'optimisation des réponses
14. Déterminer le nombre de répétition (n) de chaque essai.
15. Considérer l'ajout d'essais au centre de l'espace expérimental.
16. Existe-t-il des relations mathématiques connues entre la réponse et les facteurs?
17. Y a t –il des RESTRICTIONS À LA RANDOMISATION complète (ordre au hasard) des essais?

    remarque : certains facteurs sont-ils difficiles à changer?
18. Préciser tous les détails (protocole expérimental) pour l'exécution des essais.
19. Prévoir un budget et un échéancier pour le projet.

    Inclure des tests pour valider la ou les solutions résultant du projet.

ENTREPRENDRE UN PROJET D'EXPÉRIMENTATION (4/5)

Rôle des facteurs (variables, paramètres)



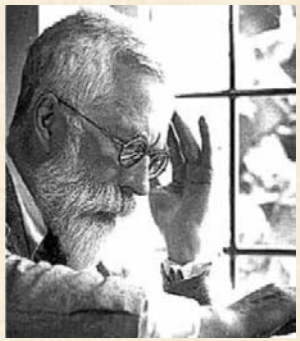
ENTREPRENDRE UN PROJET D'EXPÉRIMENTATION (5/5)

en immobilier : location location location

vie de couple : comunication communication communication

études statistiques : planification planification planification

« To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. » **Sir Ronald Fisher (1890-1962)**



Étapes du processus de modélisation statistique

1. Identification	processus / problème / variables
2. Observation	plan collecte des données
3. Spécification	modèle pour analyse
4. Estimation	paramètres du modèle
5. Décomposition	variabilité
6. Validation	analyse résidus / validation croisée
7. Exploitation	- optimisation
	- résolution problème
	- décision
	- action

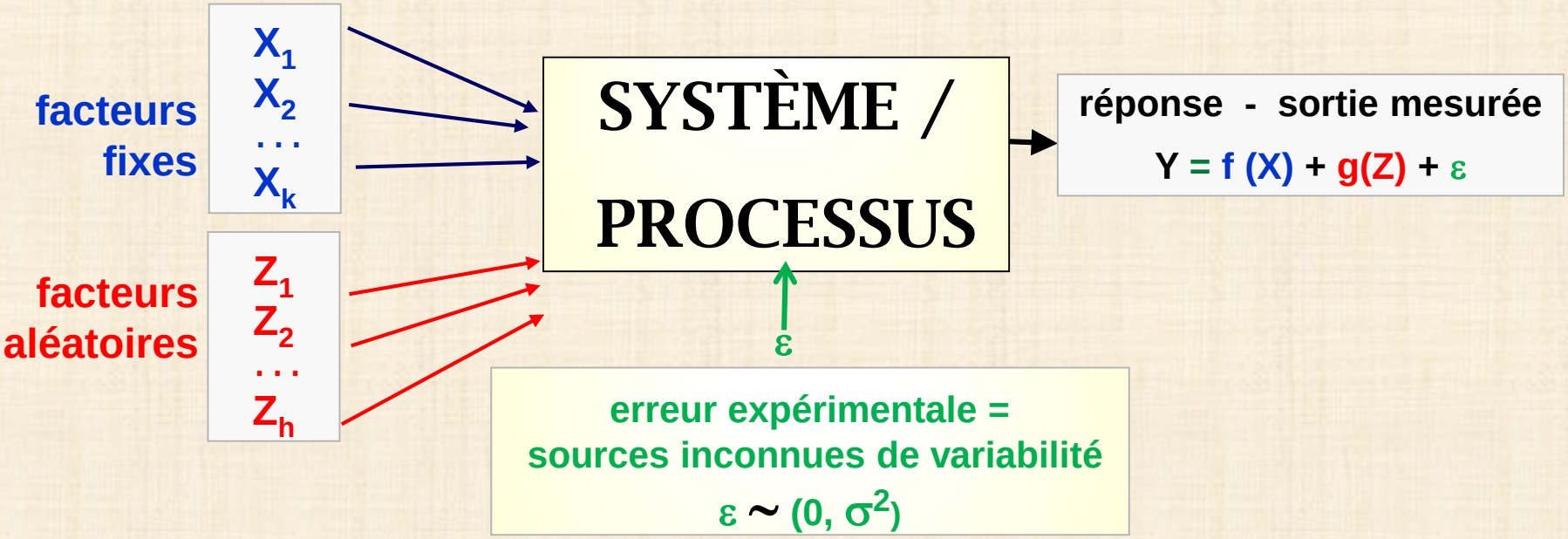
ANALYSE STATISTIQUE : comprendre / prédire / optimiser
SYSTÈME / PROCESSUS

TOUTE analyse statistique repose sur un MODÈLE

- relation entre input X (fixe) , Z (aléatoire) et output Y
- connaissance de la structure des données:
  - plan de collecte des données / nature variables / rôle des variables
  - type d'influence des variables / unités statistiques (expérimentales)

$X_1, X_2, \ldots, Z_1, Z_2, \ldots$: facteurs, variables contrôlées en expérimentation
mode actif : données expérimentales

variables observées / mesurées en observation
mode passif – données observationnelles



ANALYSE STATISTIQUE : comprendre / prédire / optimiser

SYSTÈME / PROCESSUS

VARIABLES

RÔLE

Y : réponse , output, à expliquer

peut être: binaire (0, 1), multinomiale, continue, multidimensionnelle

X, Z : explicatives, régresseurs, input

inter / intra relativement aux unités expérimentales

NATURE

X (fixes) : continues, catégoriques (facteurs)

Z (aléatoires) : continues, catégoriques

INFLUENCE

X : affecte la centralité (moyenne) de Y : effets fixes

Z : affecte la dispersion (variance) de Y : effets aléatoires

MODÈLES effets fixes , effets aléatoires , mixtes (fixes + aléatoires)

$$Y = f(X_1, X_2, \ldots, X_k ; \beta_0, \beta_1, \beta_2, \ldots) + g(Z_1, Z_2, .., Z_h ; \sigma_1^2, \sigma_2^2, \ldots) + \varepsilon(0, \sigma^2)$$

ANALYSE STATISTIQUE : comprendre / prédire / optimiser

VARIABLES

Nature: continue - catégorique

Rôle: explicatives (X = input) - à expliquer (Y = output = réponse)

Liste des X complète? k = nombre OK?

Mesure de Y - processus de mesure / erreur? justesse?

STRUCTURE et le PLAN de collecte des données

expérience planifiée - quel plan statistique?

- combien de données? n?

données observées sans plan expérimental – qualité?

Terme d'erreur expérimentale - distribution normale? importance?

préoccupation obsessive de la normalité !

Forme de f - connue – linéaire / non linéaire (cas plutôt rare)

- inconnue - quelle approximation? – polynomiale?
- techniques de sélection des variables pour modéliser
- qualité du modèle ajusté? critères?

Ajustement du modèle - analyse de sensibilité des X

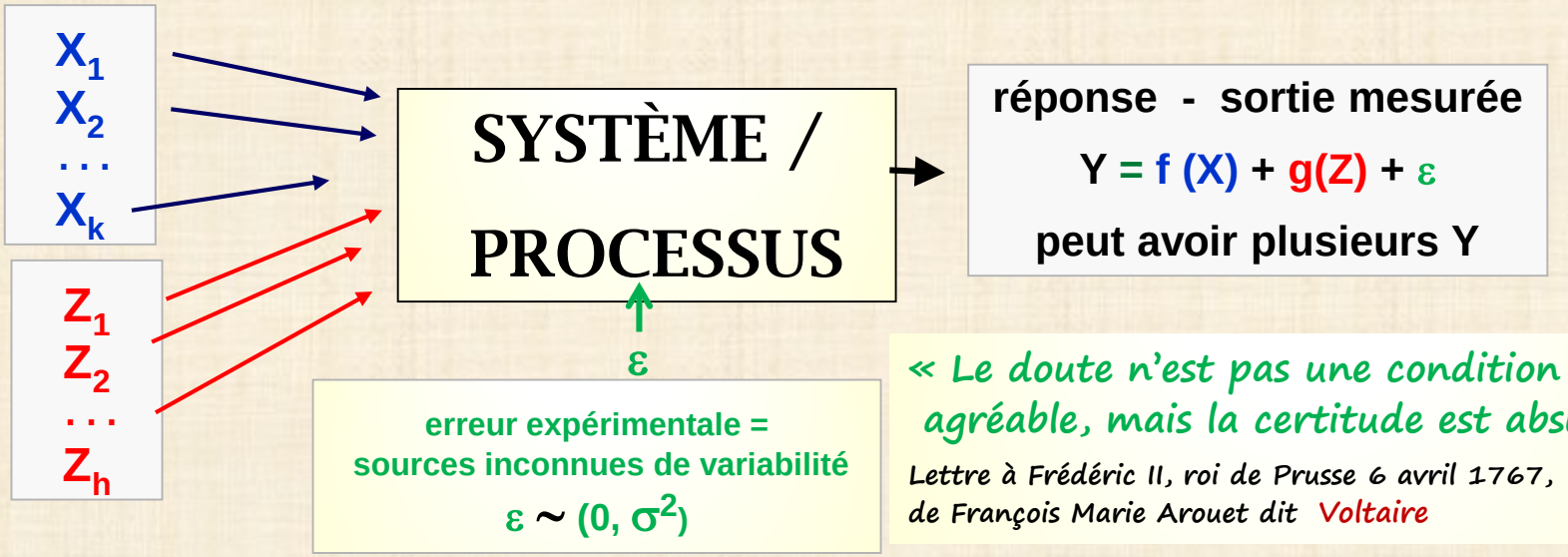
Évaluation de qualité du modèle - analyse des résidus

- validation croisée

ANALYSE STATISTIQUE : étapes

1. Spécification d'un modèle statistique
2. Estimation des paramètres du modèle
3. Décomposition de la variabilité : ANOVA
4. Tests d'hypothèses sur les paramètres
5. Analyse diagnostique des résidus
   - vérification des hypothèses de base
   - identification d'observations influentes
   - transformation Box-Cox de réponse Y
6. Si nécessaire : itération des étapes 1 à 5
7. Optimisation de la réponse (s'il y a lieu)
8. Graphiques de la réponse

ANALYSE STATISTIQUE : comprendre / prédire / optimiser



« Le doute n'est pas une condition agréable, mais la certitude est absurde. »

Lettre à Frédéric II, roi de Prusse 6 avril 1767, de François Marie Arouet dit Voltaire

Aucune restriction concernant la nature des X et Y

X: catégorique, entière, continue, contrôlées, aléatoires

Y: binaire(0, 1), multinomiale, entière, continue

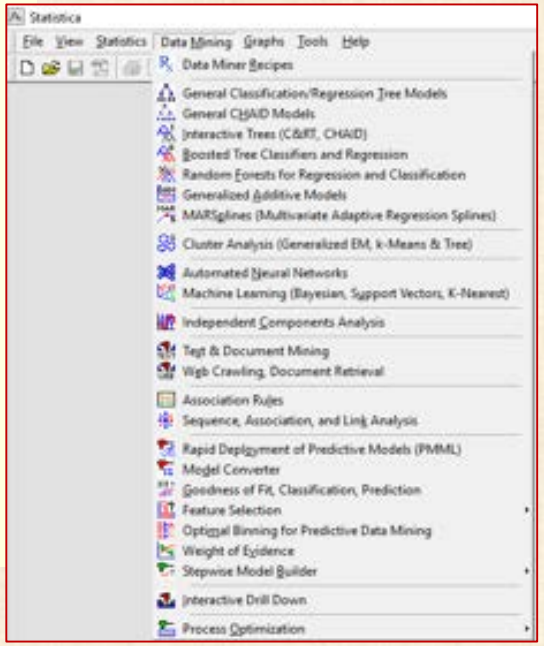
Modèles et algorithmes (nombreux !)

linéaire, linéaire généralisé, arbres, réseaux neurones, PLS, etc. ..

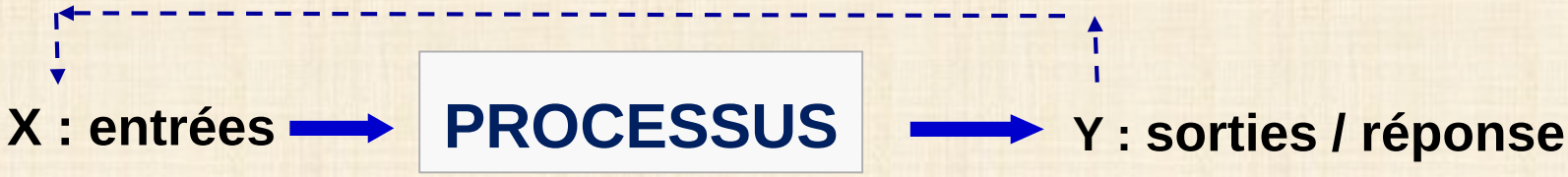
p = nombre de variables n = nombre d'observations

On peut avoir plus de variables que d'obsevations !

data mining



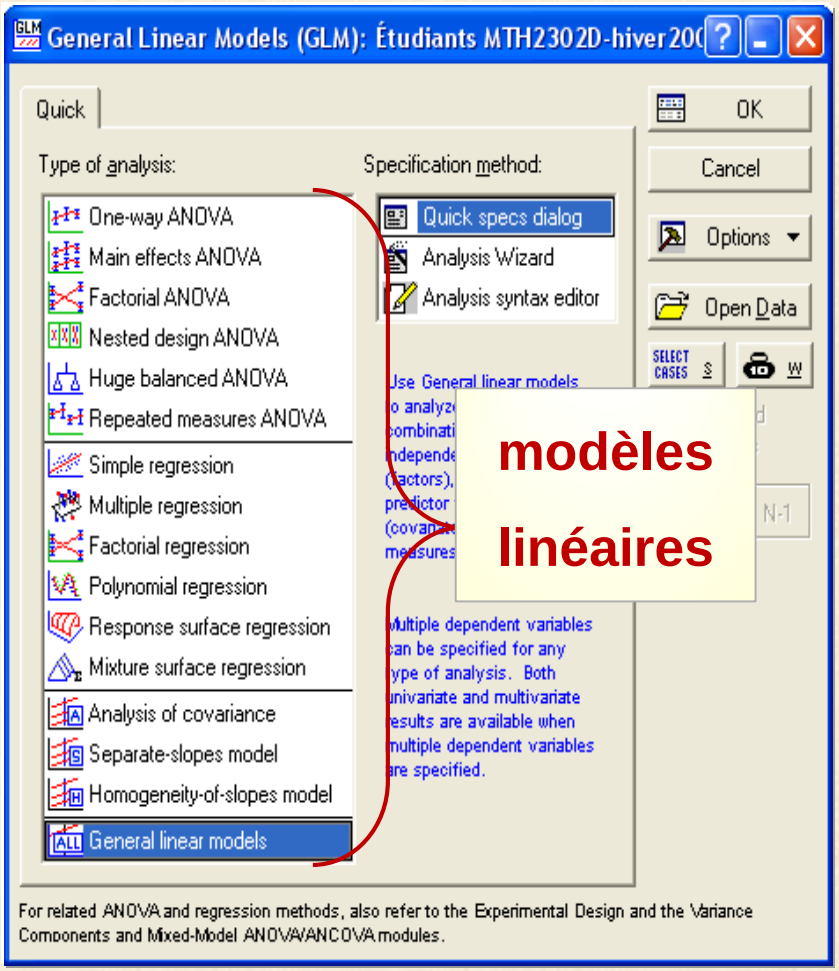
Étude des relations entrées-sorties



COMPARAISON	Modèle de prédiction	Modèle d'analyse de variance
But	développement d'un modèle prédictif de la réponse	identification des effets significatifs sur la réponse
Source des données	historiques / observationnelles	résultat d'un plan d'expérimentation
Nombre d'observations	grand: centaines, milliers…	petit : dizaines
Variables d'entrée	continues / quantitatives	catégoriques / qualitatives
Nombre de valeurs distinctes des variables d'entrée	autant qu'il y a d'observations	nombre restreint généralement moins de 10
Utilisation des variables indicatrices (0-1)	occasionnelle	employées systématiquement pour représenter les modalités
Emphase et difficulté	forme et la qualité du modèle	spécification du modèle reflétant la complexité du plan expérimental
Structure des données	simple	complexe

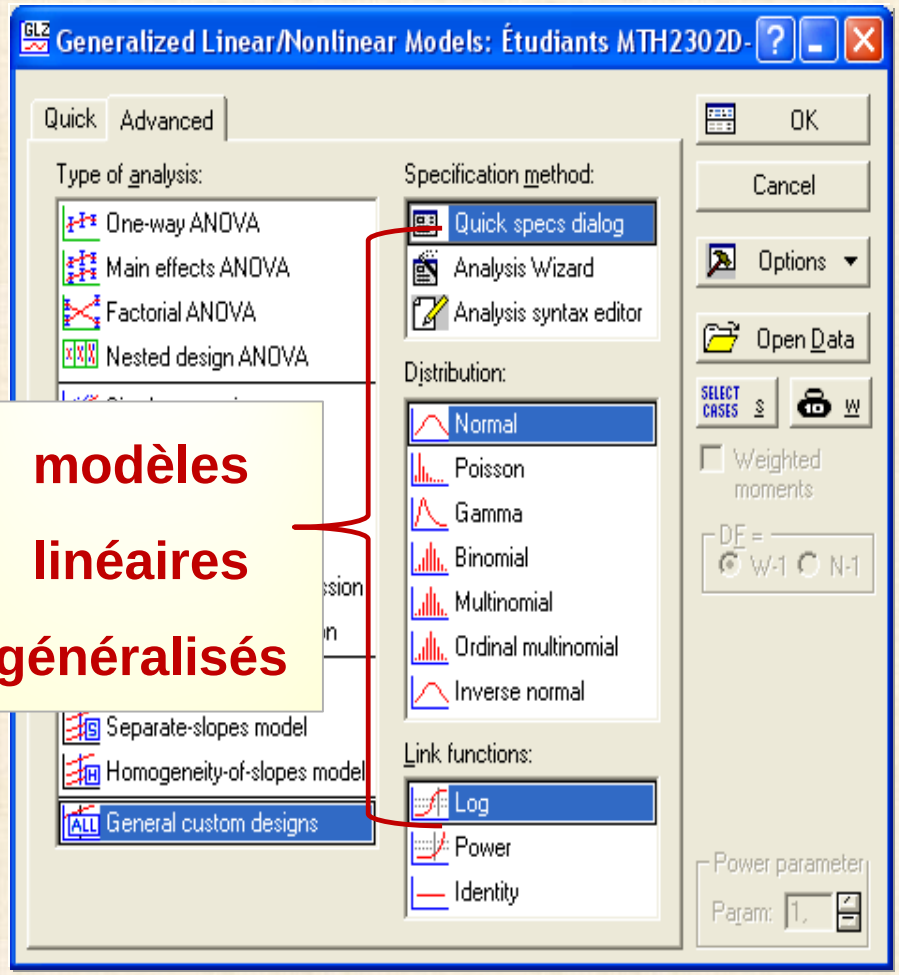
modèles disponibles avec Statistica + modèles avec data mining

GLM : General Linear Model



modèles linéaires

GLZ : Generalized Linear/Nonlinear Model



modèles linéaires généralisés

" I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s ? "

Hal Varian, PhD, chief economist, Google

a sexy statistician

Bernard Clément

22 ans

Bernard Clément, PhD
Statisticien

Génistat Conseils

1205-80 Berlioz
Verdun, QC
Canada H3E 1N9

Cell : 514-677-7896
genistat@sympatico.ca

Bernard Clément PhD

Génistat Conseils
Courriel : genistat@sympatico.ca
Tel : (514) 677-7896

Département de mathématiques
et de génie industriel
École Polytechnique de Montréal
Tél. : (514) 340-4711 poste 4944
Courriel : bernard.clement@polymtl.ca

Bernard Clément, PhD est professeur titulaire au département de mathématiques et de génie industriel de l'École Polytechnique de Montréal affiliée à l'Université de Montréal. Il possède plus de 30 années d'expérience en enseignement des méthodes de statistiques appliquées et en management de la qualité aux ingénieurs et scientifiques.

Il a fondé Génistat Conseils Inc., une firme de consultation spécialisée en design et analyse d'études statistiques. Son produit principal est le transfert d'expertise, de connaissance et de management pour l'amélioration de la qualité des produits et des procédés.

Sa liste de clients comprend IBM, Sidbec-Dosco, Noranda Research Center, Bolting Technology Council, Nortel, Institut de Recherche en Biotechnologie, Compagnie Générale des Eaux (Vivendi), Bell, Postes Canada, DALSA semiconducteurs, Cardianove, Warnex, Camoplast, CIRANO et plusieurs établissements de recherche.

Il est membre élu de l'International Statistical Institute et membre de American Society of Quality. Il a été vice-président du Canada Quality Council, administrateur de l'Association québécoise de la Qualité (Mouvement Québécois Qualité), et il fut président de la Société Statistique de Montréal. Il a été membre du comité ISO du Standard Council du Canada.

Consultation - recherche - formation

- Planification et analyse d'expériences industrielles (DOE)
- Maîtrise statistique des processus (SPC)
- Études statistiques appliquées: design, analyse, data mining
- Management de la qualité et Six Sigma
- Ingénierie robuste de Taguchi : design de produit et de procédé
- Logiciels statistiques : STATISTICA, MINITAB, JMP, SAS, DESIGN-EXPERT

Site Internet http://www.cours.polymtl.ca/mth6301