Tests statistiques et autres méthodes

Nadine Mandran
Grenoble

Juin 2010

La démarche quantitative
Les données

- Les variables **indépendantes** ou **facteurs** ou **explicatives**
  - Celles qui sont fixées avant l'expérimentation
    - Deux classes d'âges (lycées et collégiens)
    - Le niveau d'accès des étudiants à un tuteur virtuel
- Les variables **dépendantes** ou **à expliquer**
  - Celles qui seront mesurées
    - Temps de réalisation des exercices
- Les **indicateurs**
  - Celles qui sont construites après
    - Ratio temps de réalisation de l'exercice i / temps total de réalisation

La démarche quantitative
Les variables

- Les variables qualitatives ou nominales
  - Genre (Homme, Femme)
  - Diplôme (Bep, BAC, Bac+2, …)
- Les variables quantitatives continues
  - Age
  - Temps d'une action
  - Nombre de réussites ou d'erreurs
  - Taille
  - Les mesures répétées danger …
- Les variables quantitatives ordinales
  - Notation
  - Niveaux d'utilisation d'un logiciel
  - Echelle de satisfaction
  - Age en classes
- Les variables séquentielles
- Les séries chronologiques

La démarche quantitative
L'échantillonnage

- **Les quotas**
  - Caler les résultats sur les données du recensement de la population ou sur des informations globales sur la population étudiée.
  - Le plus pratiqué car moins couteux
- **La méthode aléatoire**
  - Tirer au hasard dans une base de sondage.
  - La probabilité d'inclusion d'un individu est connu sans biais.
  - Couteux car l'individu tiré au hasard ne doit pas être abandonné
- **L'exhaustivité (le recensement)**
  - Avoir la totalité des individus ou des actions faites par les individus
- **Le plan expérimental**
  - Le nombre de sujets est fixé a priori en fonction des facteurs expérimentaux

Démarche quantitative
Le plan expérimental

- A partir des hypothèses à tester
- Identifier les facteurs
- Identifier les variables
- Dénombrer le nombre d'individus

- Exemple de plan : Utilisation de Copex Chimie (C.d'Ham et I.Girault, Metah)
  - 120 étudiants en 1ère année de physique, répartis en 5 groupes
  - Facteurs : les niveaux d'accès au tuteur et à la description
    - full accès - accès limité - description limitée – accès et description limités – no tutor - pas de logiciel
  - Variables : l'accès aux leçons et les erreurs commises (50 variables), le temps de la session
  - Indicateurs : Les 50 variables sur le temps de la session
  - Les « pas de logiciel » : groupe contrôle

La démarche quantitative
Combien de sujets ?

Individu statistique ou unité statistique

– « La population statistique est l'ensemble sur lequel des méthodes et techniques de présentation, de description et d'inférence statistique sont appliquées. Il ne s'agit donc pas forcément d'une population au sens biologique du terme ». *(J.Vaillant 2005)*

– « Population et individus : La population est l'ensemble des individus (ou unités statistiques) auxquels on décide de s'intéresser . Sa taille, habituellement désignée par N, est grande, ou même infinie. Le choix de la population étudiée dépend du problème qui est à l'origine de la démarche statistique, et de la façon dont on décide de le traiter. » *(Vocabulaire de la statistique descriptive)*

– Exemples

– l'action faite par un utilisateur :

  • 3 utilisateurs et 600 actions : 1800 individus statistiques

– Le diagnostic établi par un enseignant pour un élève

  • 10 diagnostics * 4 enseignants * 20 élèves : 800 individus statistiques

Un exemple : deux classes

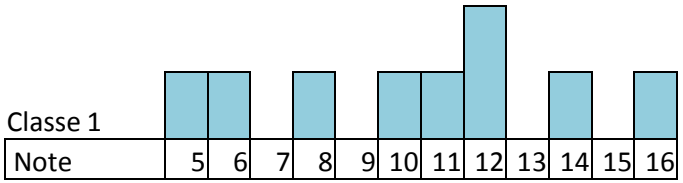
classe 1	classe 2
8	10
12	11
16	10
5	10
6	11
12	11
14	11
11	10
10	10

Moyenne :
Classe 1 : 10,4
Classe 2 : 10,4

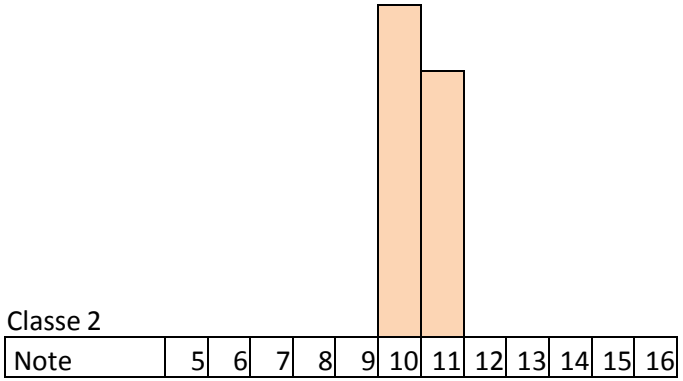
Médiane
Classe 1 : 11
Classe 2 : 10

Etendue
Classe 1 : 11
Classe 2 : 1

Ecart type
Classe 1 : 3,61
Classe 2 : 0,53



Coefficient de Variation %
Classe 1 : 34,7
Classe 2 : 5,1



La démarche quantitative
le test statistique

- Est-ce que 11 = 12 ?

- Cela dépend :
  - De l'échantillon, de sa taille
  - De la loi de distribution
  - De la nature de la mesure (quantitatif ou qualitative)
  - %, moyenne, coefficients, …..

La démarche quantitative
le test statistique : principes généraux

- Objectif

  Valider ou invalider une hypothèse en prenant un certain **risque**
- Démarche générale simplifiée d'un test en stat
- Poser H0 hypothèse nulle et la contre hypothèse H1
  - H0 : la taille des hommes est égale à celle des femmes
  - H1 : la taille des hommes est différente de celle des femmes

- Calculer un « indicateur statistique » qui dépend de la nature des variables, des objectifs et qui est lié à une loi de distribution
  - Loi de Student
  - Loi du Chi2, Pearson
  - Loi de Fisher
  - …
- Calculer des degrés de libertés
- Calculer la probabilité de se tromper en rejetant H0, risque α, le seuil de significativité. (risque de 1ère espèce : rejeter H0 quand elle est vraie)
- (risque de 2ème espèce : accepter H0 quand elle est fausse)

La démarche quantitative
le test statistique

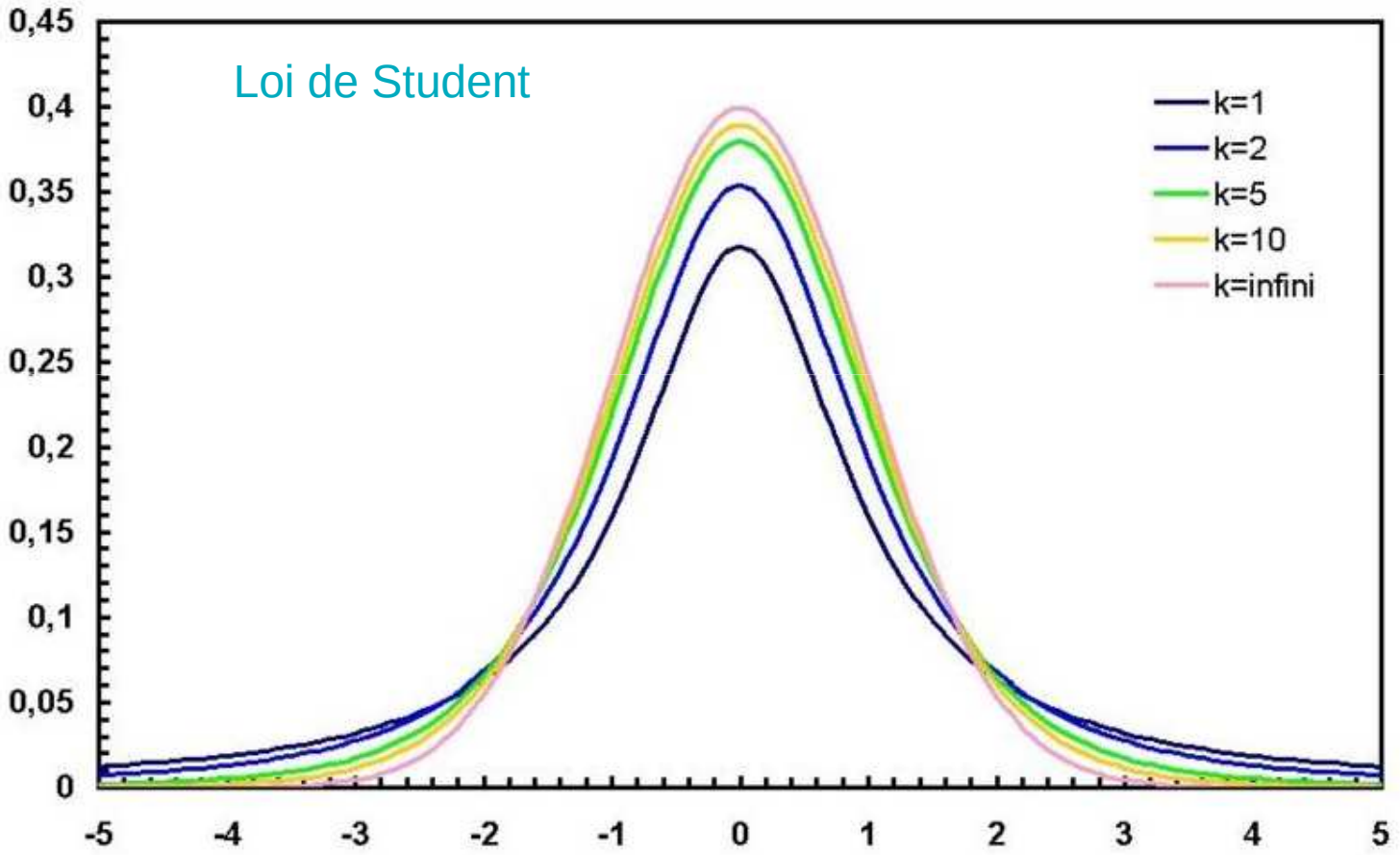
- Poser H0 hypothèse nulle et la contre hypothèse H1
  - H0 : la taille des hommes est égale à celle des femmes
  - H1 : la taille des hommes est différente de celle des femmes
- Calculer un « indicateur statistique » qui est lié à une loi de distribution
  - Facteur explicatif : qualitatif Homme/femme
  - Variables à expliquer quantitative : taille

	Homme	Femme
Moyenne	1,75	1,67
Ecart type	0,13	0,11
N	25	25

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1 X_2} \cdot \sqrt{\frac{2}{n}}}$$

- Faire le calcul du T = 28,2
- Degré de libertés (ddl) degree of freedom (df) : (n1+n2)-2 =48
- Les p-values :
- Probabilité d'obtenir une valeur de t supérieur = 0,01
- Probabilité d'obtenir une valeur de z supérieur = 0,0001
- Le test t impose la normalité de la distribution : (Kolmogorov-Smirnov, Shapiro et wilk)

La démarche quantitative
le test statistique



Exemple sur Copex Chimie

	Faible		Fort	
	Moyenne	Ecart type	Moyenne	Ecart type
Durée des sessions	116,8	17,3	120,6	19,9
N	54			

H0 : La moyenne des durées de sessions avec un niveau de rétroaction faible est égale à celle obtenue avec un niveau de rétroaction fort

H1 : H0 : La moyenne des durées de sessions avec un niveau de rétroaction faible est différente à celle obtenue avec un niveau de rétroaction fort

Ecart entre les moyennes	-3,82
Variance de Faible	298,55
Variance de Fort	394,41
Variance de faible /N	5,53
Variance de fort /N	7,30
t observé	**-1,07**
ddl	**106**
T théorique 5%	1,96
T théorique 10%	1,64
T théorique à 20%	1,28

Un exemple en live

- Une expérience pour tester des menus sur téléphone portable
  - Facteurs expérimentaux :
    - Centre ou ailleurs
    - Trait ou non
  - Variables temps des marques
    - Durée de la marque 1
    - Durée totale des marques

Hypothèse et hypothèse alternative

- H0 : le temps de la marque 1 quand le participant démarre du centre est égale au temps de marque 1 quand le participant peut partir d'ailleurs

- H1 : le temps de la marque 1 quand le participant démarre du centre est différent au temps de marque 1 quand le participant peut partir d'ailleurs

- **Réalisation des tests sur Tanagra**

La démarche quantitative
Variables quantitatives

- Analyse de variance
  - Comparaisons de moyennes en prenant en considération les variances dans les groupes et entre les groupes.
  - Une comparaison des moyennes selon les facteurs expérimentaux et de leurs interactions
  - Décomposition de la variance
    - Variance totale
    - Variances liées aux facteurs (moyenne du facteur-moyenne globale)
    - Variance résiduelle
    - Rapport des variances des facteurs/la variance résiduelle
  - Test de Fischer : rapport de la variance expliquée par les facteurs et la variance résiduelle.
  - Le test de Fischer à deux valeurs pour le degré de liberté (nombre de niveaux des facteurs -1, nombre d'individus – nombre de facteurs)

Analyse de variance « intuitive »

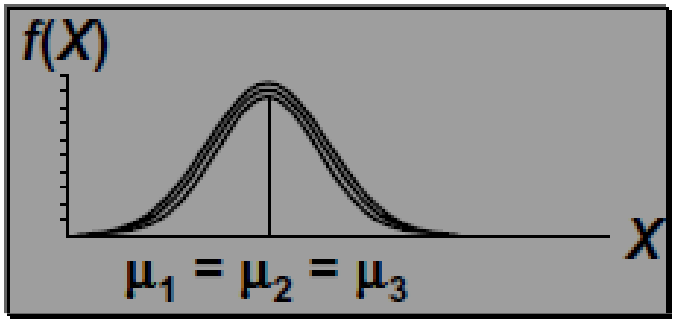
H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

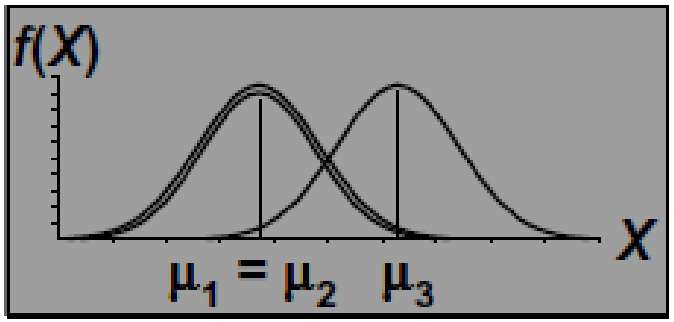
- Toutes les moyennes des poulations sont égales

H_1: Les moyennes des populations ne sont pas toutes égales

- Au moins une des moyennes est différente
- Ne signifie pas:

  $\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$





La démarche quantitative
paramétriques ou non paramétriques

- Deux hypothèses à vérifier
  - Normalité de la distribution (tests de Kolmogorov Smirnov, aplatissement et asymétrie, Shapiro)
  - Homogénéité des variances (tests de Bartlett)
- Si ces conditions ne sont pas vérifiées :
    - Tests non paramétriques, dit tests de rangs
      - Test Kruskall et Wallis, Wilcoxon, Mann et Whitney, Friedman, .
    - Transformation des données
    - Logarithmique, Arc sinus racine
- On peut également se contenter de la normalité des résidus

- Comparaison entre les modalités Tests de Tukey ou de Wilcoxon
  - Différence franchement significative HSD
  - Comparer les niveaux de facteurs entre eux.

La démarche quantitative
exemple d'analyse de variance

- Expérience sur Diagelec : Outils de diagnostic

- Deux facteurs expérimentaux :
  - Des outils d'évaluation D2,D3,D4,D5,D6,D7
  - Des types à évaluer : connaissance, compétences, erreur
- Une variable « degré de croyance qui varie de 0 à 4 »

Analysis Variable : E4 E4						
Type	N Obs	Nb	Moyenne	Écart-type	Minimum	Maximum
c	4992	4992	3.5833333	0.8747823	0	4.0000000
e	4444	4444	3.4718722	0.9023295	0	4.0000000
s	3650	3650	3.4383562	0.9215055	0	4.0000000

Analysis Variable : E4 E4						
DIAG	N Obs	Nb	Moyenne	Écart-type	Minimum	Maximum
D2	1105	1105	2.8733032	1.0564941	0	4.0000000
D3	1102	1102	3.4627949	0.9405462	0	4.0000000
D4	2918	2918	3.5281014	0.9125634	0	4.0000000
D5	2324	2324	3.7538726	0.6891369	0	4.0000000
D6	1907	1907	3.2863136	0.9813777	0	4.0000000
D7	3730	3730	3.6434316	0.7788027	0	4.0000000

La démarche quantitative
exemple d'analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	17	1058.80070	62.28239	85.42	<.0001
Error	13068	9528.36642	0.72914		
Corrected Total	13085	10587.16713			

Signification du modèle

Source	DF	Type I SS	Carré moyen	Valeur F	Pr > F
DIAG	5	751.0824023	150.2164805	206.02	<.0001
Type	2	27.3789430	13.6894715	18.77	<.0001
DIAG*Type	10	280.3393561	28.0339356	38.45	<.0001

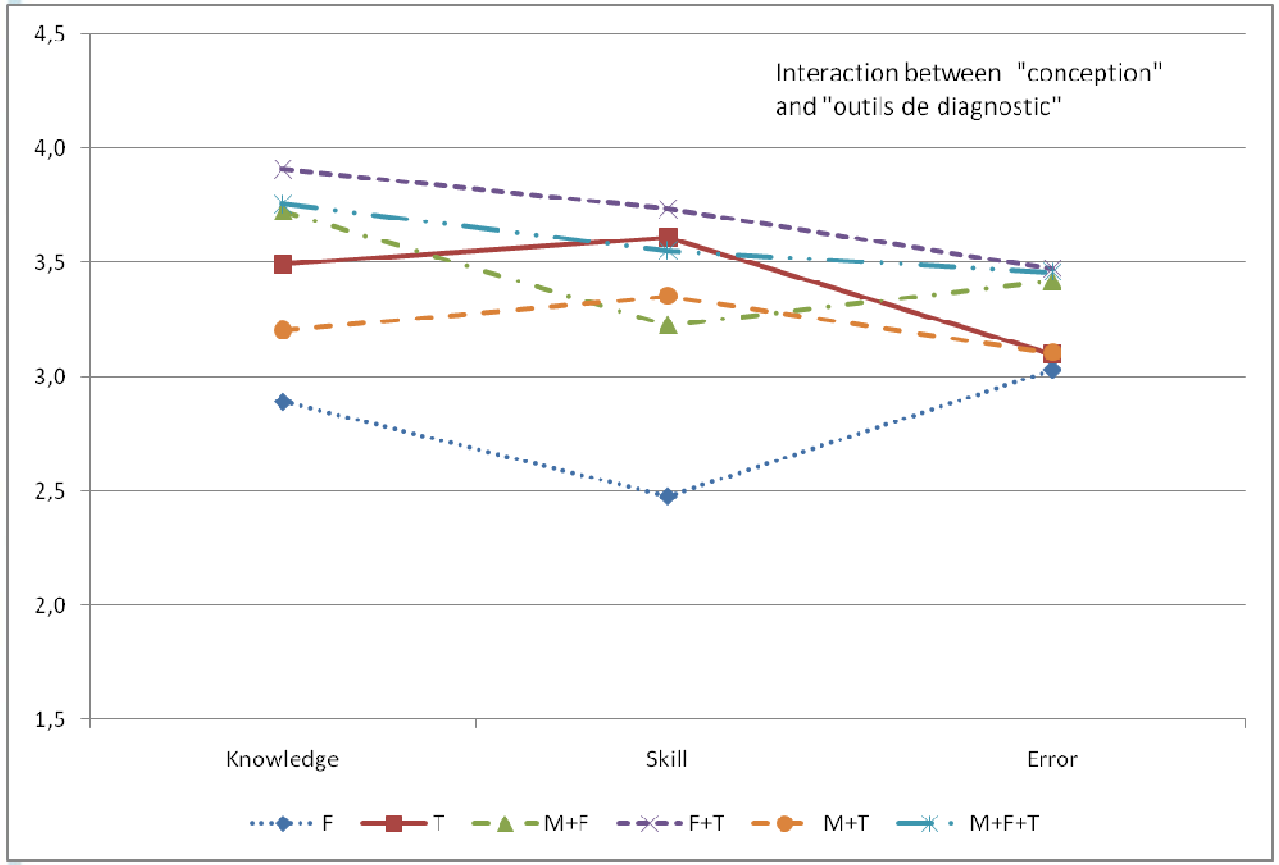
Signification des facteurs

Où se trouve les différences ? Tests de Tukey

La démarche quantitative
exemple d'analyse de variance

Comparaisons significatives au niveau 0.05 indiquées par ***.				
DIAG Comparaison	Différence entre les moyennes	Simultané 95Limites de confiance %		
D5 - D7	0.11044	0.04612	0.17476	***
D5 - D4	0.22577	0.15811	0.29344	***
D5 - D3	0.29108	0.20206	0.38009	***
D5 - D6	0.46756	0.39236	0.54276	***
D5 - D2	0.88057	0.79164	0.96950	***
D7 - D5	-0.11044	-0.17476	-0.04612	***
D7 - D4	0.11533	0.05518	0.17548	***
D7 - D3	0.18064	0.09719	0.26408	***
D7 - D6	0.35712	0.28861	0.42563	***
D7 - D2	0.77013	0.68677	0.85349	***
D4 - D5	-0.22577	-0.29344	-0.15811	***
D4 - D7	-0.11533	-0.17548	-0.05518	***
D4 - D3	0.06531	-0.02074	0.15136	
D4 - D6	0.24179	0.17012	0.31345	***
D4 - D2	0.65480	0.56883	0.74076	***
D3 - D5	-0.29108	-0.38009	-0.20206	***
D3 - D7	-0.18064	-0.26408	-0.09719	***
D3 - D4	-0.06531	-0.15136	0.02074	
D3 - D6	0.17648	0.08439	0.26857	***
D3 - D2	0.58949	0.48588	0.69310	***
D6 - D5	-0.46756	-0.54276	-0.39236	***
D6 - D7	-0.35712	-0.42563	-0.28861	***
D6 - D4	-0.24179	-0.31345	-0.17012	***
D6 - D3	-0.17648	-0.26857	-0.08439	***
D6 - D2	0.41301	0.32100	0.50502	***
D2 - D5	-0.88057	-0.96950	-0.79164	***
D2 - D7	-0.77013	-0.85349	-0.68677	***
D2 - D4	-0.65480	-0.74076	-0.56883	***
D2 - D3	-0.58949	-0.69310	-0.48588	***
D2 - D6	-0.41301	-0.50502	-0.32100	***

La démarche quantitative
exemple d'analyse de variance



La démarche quantitative

Test de la normalité

- Les tests statistiques paramétriques sont valables quand la loi de distribution suit une loi normale.
- Les manières de tester l'adéquation à la loi normale
  - Les histogrammes
  - Les indicateurs
  - Coefficient d'assymétrie et d'applatissement
  - Test Kolmogorov Smirnov
  - Test Shapiro
  - ….. (exemple sur Tanagra – lecture des tests)
- Les solutions quand la distribution n'est pas normale
- 1 - transformer les variables :
  - Fonctions utilisées : Log népérien, Racine, Arcsinus
  - Vérifier la normalité de ces variables transformées
  - Utiliser des tests paramétriques
- 2 – utiliser des tests non paramétriques ou de rangs
- 3 – Faire de l'anova et tester la loi de distribution des résidus

La démarche quantitative

Les tests non paramétriques ou tests de rangs

- Les calculs statistiques ne sont plus faits sur les valeurs des variables
- Les indicateurs classiques comme la moyenne et écart type ne sont pas utilisés
- On utilise le rang des valeurs
- Les valeurs de la variable à étudier sont triées par ordre croissant et ensuite une nouvelle variable est créé qui est le rang de la valeur
- Ces rangs servent alors de variables
- Travaille sur la moyenne des rangs

La démarche quantitative

Les tests non paramétriques ou tests de rangs

- Définition des variables et facteurs
- Exercice avec les données
- H0 : la durée des marques est égale quelque soit le point de départ du tracé et de la présence du trait ou non
- H1 : la durée des marques est différente en fonction du point de départ du tracé et de la présence du trait ou non

- Sur Tanagra
- Test de la normalité
- Comparaison de moyennes avec deux facteurs expérimentaux avec le test de Kruskall et Wallis

La démarche quantitative

Analyse de l'indépendance

- Deux variables qualitatives
- Nombre de modalités restreintes
- Indépendance et pas de corrélation

La démarche quantitative

Etude de l'indépendance

- Quatre tuteurs qui réalisent des actions sur une interface qui comportent trois niveaux d'utilisation
- Au préalable les tuteurs sont interrogés sur leur pratique de suivi d'une classe de TP sans outil informatique
- Les actions sont enregistrées => un fichier de traces

La démarche quantitative
Etude de l'indépendance

Equipe Metah

Figure 1. FORMID-Suivi, extrait du niveau 1.

Source : « Expérimentation d'un environnement flexible pour la supervision de travaux pratiques basés sur des simulations » Viviane Guéraud, Jean-Michel Adam, Anne Lejeune, Nadine Mandran, Nicolas Vézian, Michel Dubois, EIAH 2009.

La démarche quantitative

Etude de l'indépendance

Equipe Metah

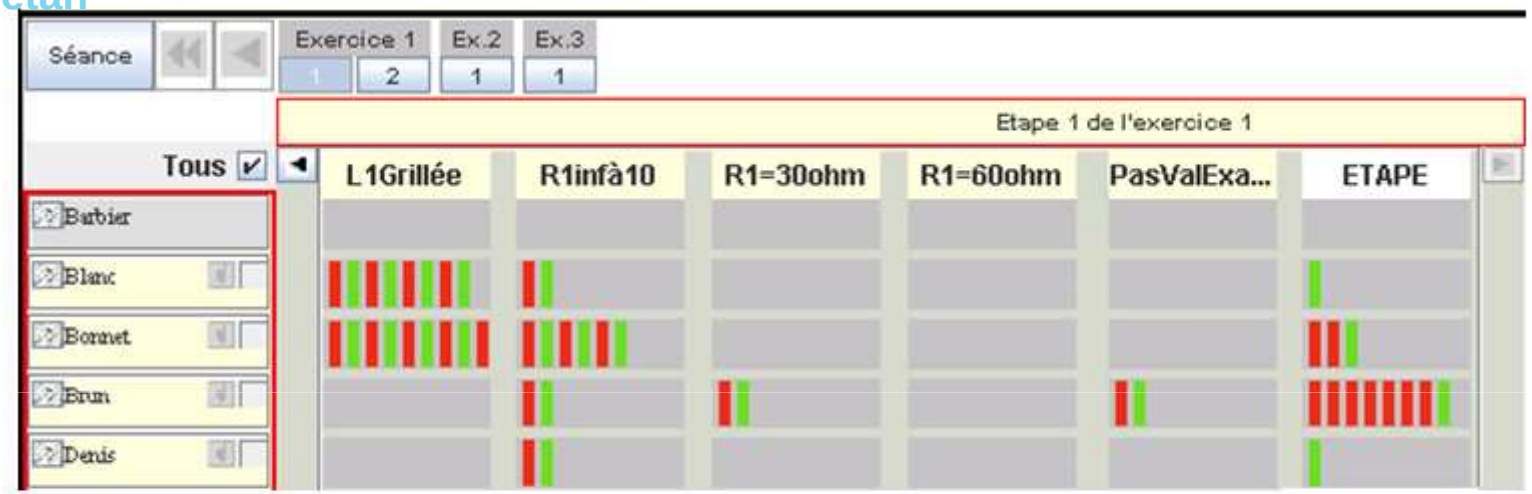


Figure 2. FORMID-Suivi, extrait du niveau 2 (apprenants dans l'étape 1 de l'exercice 1).

Figure 3. FORMID-Suivi, extrait du niveau 3 (Duval dans l'étape 1 de l'exercice 2)

La démarche quantitative

Etude de l'indépendance

- Objectifs :
  - Décrire l'utilisation effective du logiciel
  - Comparer les utilisations entre les tuteurs et les niveaux du logiciel
  - Des variables qualitatives : tuteur et niveau
  - Unité statistique : Action

La démarche quantitative
Etude de l'indépendance

- Le tableau croisé : Fréquence

Fréquence	Niveau 1	Niveau 2	Niveau 3	Total
Gerard	24	452	121	597
Kevin	74	246	3	323
Nathalie	206	18	268	492
Pierre	437	50	404	891
Total	741	766	796	2303

La démarche quantitative
Etude de l'indépendance

- Le tableau croisé : % total

% total	Niveau 1	Niveau 2	Niveau 3	Total
Gerard	1,0	19,6	5,3	25,9
Kevin	3,2	10,7	0,1	14,0
Nathalie	8,9	0,8	11,6	21,4
Pierre	19,0	2,2	17,5	38,7
Total	32,2	33,3	34,6	100,0

La démarche quantitative
Etude de l'indépendance

- Le tableau croisé : % ligne

% ligne	Niveau 1	Niveau 2	Niveau 3	Total
Gerard	4,0	75,7	20,3	100,0
Kevin	22,9	76,2	0,9	100,0
Nathalie	41,9	3,7	54,5	100,0
Pierre	49,0	5,6	45,3	100,0
Total	32,2	33,3	34,6	100,0

La démarche quantitative
Etude de l'indépendance

- Le tableau croisé : % colonne

% colonne	Niveau 1	Niveau 2	Niveau 3	Total
Gerard	3,2	59,0	15,2	25,9
Kevin	10,0	32,1	0,4	14,0
Nathalie	27,8	2,3	33,7	21,4
Pierre	59,0	6,5	50,8	38,7
Total	100,0	100,0	100,0	100,0

La démarche quantitative

Etude de l'indépendance

% action à un niveau donné pour un tuteur

% ligne	Niveau 1	Niveau 2	Niveau 3
Gerard	4,0	75,7	20,3
Kevin	22,9	76,2	0,9
Nathalie	41,9	3,7	54,5
Pierre	49,0	5,6	45,3
Total	32,2	33,3	34,6

% action totale pour un tuteur

Total
25,9
14,0
21,4
38,7
100,0

Gérard a fait 597 actions sur 2303 soit 25.9% des actions totales
Parmi les actions qu'il a faites 75% sont faites au niveau 2 (452/597)

=> Sur représentation en niveau 2 versus sous représentation en niveau 1

La démarche quantitative
Etude de l'indépendance

- Le test d'indépendance
  - H0 : indépendance entre tuteurs et niveaux
    - Clics au hasard sur l'interface
    - Une même utilisation
  - H1 : Liaison entre les tuteurs et les niveaux
    - Une utilisation différente selon les tuteurs

  - La grandeur statistique : Chi2 de Pearson

Statistique	DF	Valeur	Proba.
Khi-2	6	1312.8695	<.0001
Test du rapport de vraisemblance	6	1532.8885	<.0001
Khi-2 de Mantel-Haenszel	1	8.6900	0.0032
Coefficient Phi		0.7550	
Coefficient de contingence		0.6026	
V de Cramer		0.5339	

La démarche quantitative

Etude de l'indépendance

Fréquence Pourcentage Pourct. en ligne Pourct. en col.	Table de tuteur par Niveau_			
	Niveau_(Niveau)			
tuteur(tuteur)	Niveau 1	Niveau 2	Niveau 3	Total
G	24 1.04 4.02 3.24	452 19.63 75.71 59.01	121 5.25 20.27 15.20	597 25.92
K	74 3.21 22.91 9.99	246 10.68 76.16 32.11	3 0.13 0.93 0.38	323 14.03
N	206 8.94 41.87 27.80	18 0.78 3.66 2.35	268 11.64 54.47 33.67	492 21.36
P	437 18.98 49.05 58.97	50 2.17 5.61 6.53	404 17.54 45.34 50.75	891 38.69
Total	741 32.18	766 33.26	796 34.56	2303 100.00

La démarche quantitative

Etude de l'indépendance

- Exercice avec les données de J.Francone
- Test de l'indépendance entre le type de marque 1 faite et l'erreur sur cette marque
- Test du chi 2 et v de Cramer
- Tableau % ligne ou colonne
- Contribution du Chi2

La démarche quantitative
les méthodes

- Les objectifs
  - Décrire
  - Tester l'adéquation
  - Tester l'indépendance
  - Comparer
  - Etudier la liaison
  - Modéliser
  - Résumer un ensemble de données
  - Classifier des données
- Les lois de distribution
  - Normale
  - Poisson
  - Binomiale
  - …

La démarche quantitative
les méthodes

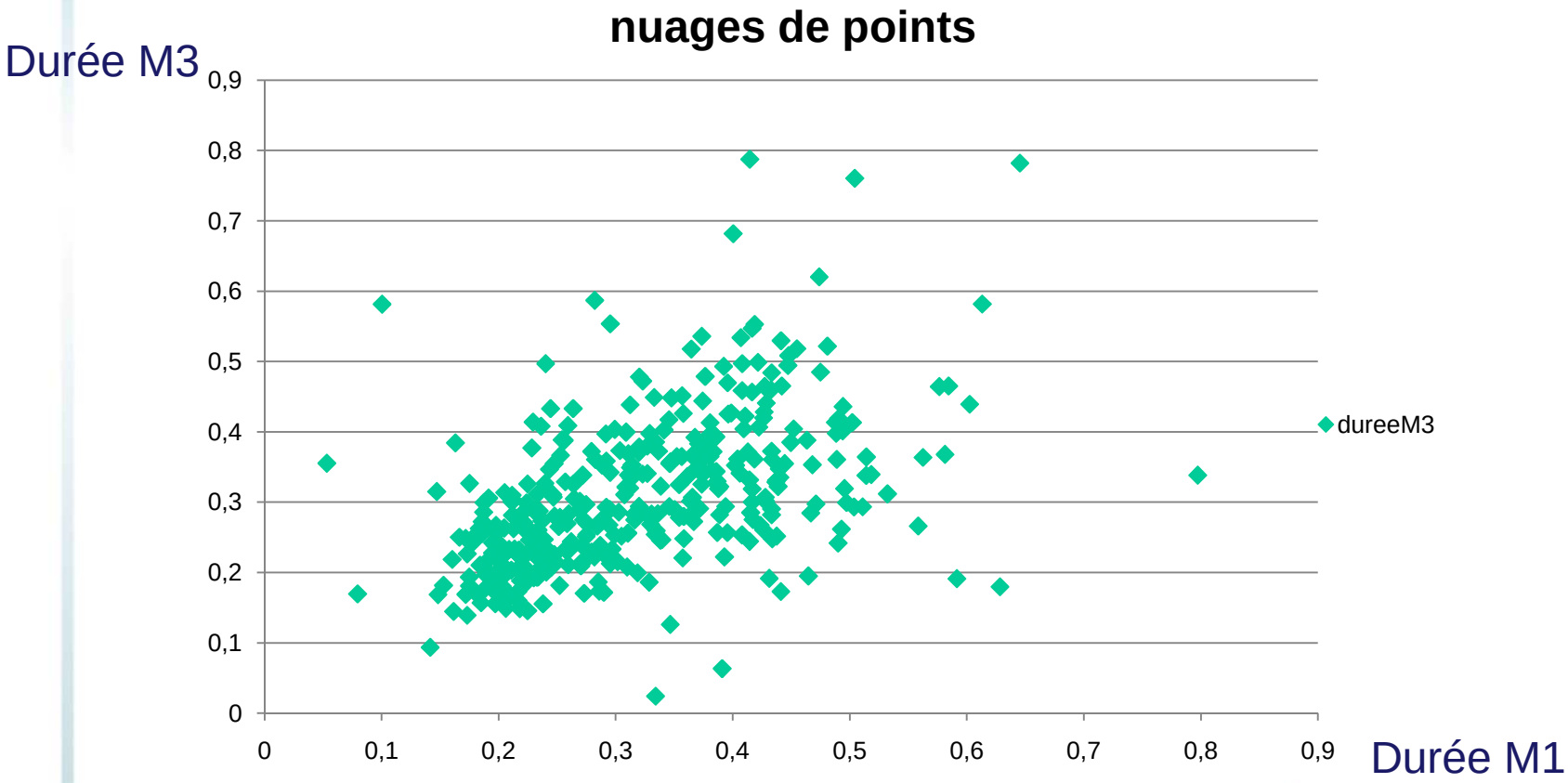
	Qualitatives	Quantitatives	Ordinales	Textuelles	Séquentielles	Chronologiques
Décrire graphiquement	Histogramme	Histogrammes/Nuages/courbes	Histogrammes/Nuages/courbes		Distance de levenstein	Courbes
Décrire quantitativement	Pourcentages	Moyennes, ….	Moyennes, ….	Dénombrement du lexique/ analyse lexicale	Coefficient de confiance et de cohérence	
Tester l'adéquation		Coefficient d'aplatissement/ coefficient d'asymétrie, Test de Kolmogorov Smirnov	*Coefficient d'aplatissement/ coefficient d'asymétrie, Test de Kolmogorov Smirnov*			
Tester l'indépendance	Test du Chi2/ V de Cramer					
Comparer	Test de proportions	Test de student/ Analyse de variances/ tests non paramétriques			Comparaison de chaîne/Distance de levenstein/ *tests classiques*	techniques des séries temporelles
Etudier la liaison		Coefficient de corrélation, Rho de Spearman, test	tau de Kendall			
Modéliser	Régression logistique	Régression linéaire, non linéaire				Modélisation des séries chronologiques
Résumer un ensemble de données	Analyse factorielles des correspondances simples ou multiples	Analyse en composantes principales/ analyse discriminante	Analyse en composantes principales/ analyse discriminante	Analyse automatisée		
Classifier des données	Classification automatique	Classification automatique	Classification automatique	Analyse automatisée		

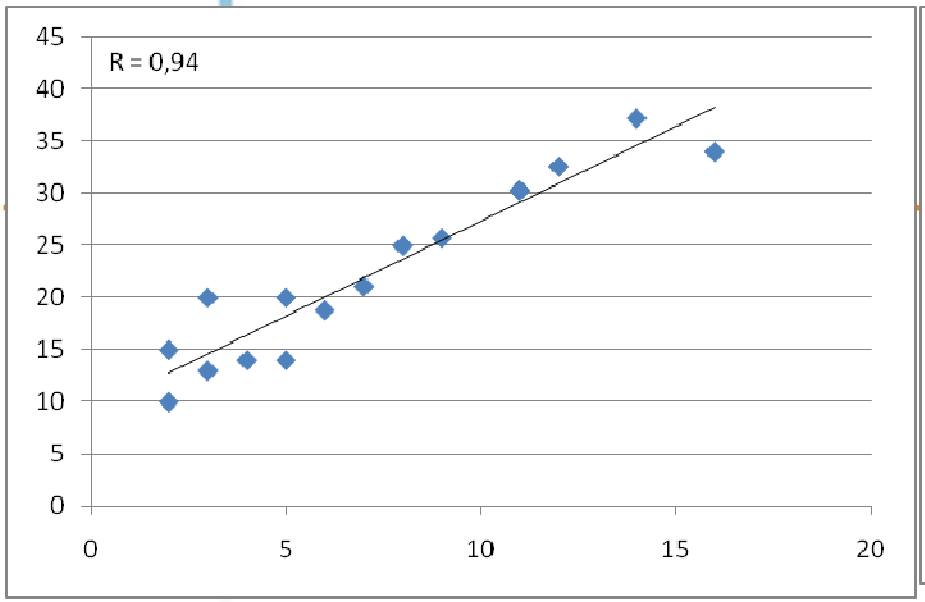
La démarche quantitative
les méthodes

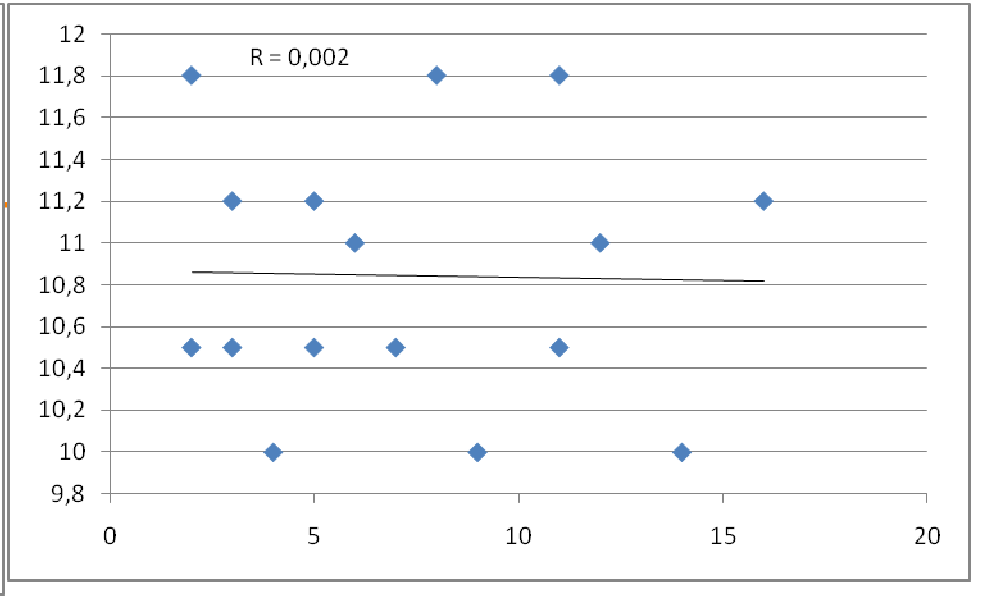
Variables à expliquer	Facteurs explicatifs	
	Qualitatif	**Quantitatif**
Qualitatif		*Avec une mise en classes des variables*
	Test du chi2 table de contingence	**Test du chi2 table de contingence**
	Test Z	**Test Z**
	Analyse factorielle des correspondances (ACM)	Analyse factorielle des correspondances (ACM)
	Régression logistique	Régression logistique
	Segmentation	segmentation
	Classification (CAH, CAD)	Classification (CAH, CAD)
Quantitatif	**tests T, Z**	**Analyse de corrélation (R2)**
	Analyse de variance (ANOVA)	**Régression linéaire ou non linéaire**
	Analyse en composantes principales (ACP)	Analyse en composantes principales (ACP)
	Analyse discriminante (AD)	Classification (CAH, CAD)
	Classification (CAH, CAD)	
	Tests non paramétriques	
Textuelles	Analyse lexicale Analyse du discours Computer assisted qualitative data analyse software (CAQDAS) Logiciels : Sphinx, Alceste, Tropes, Prospero, Spad T, DTM	

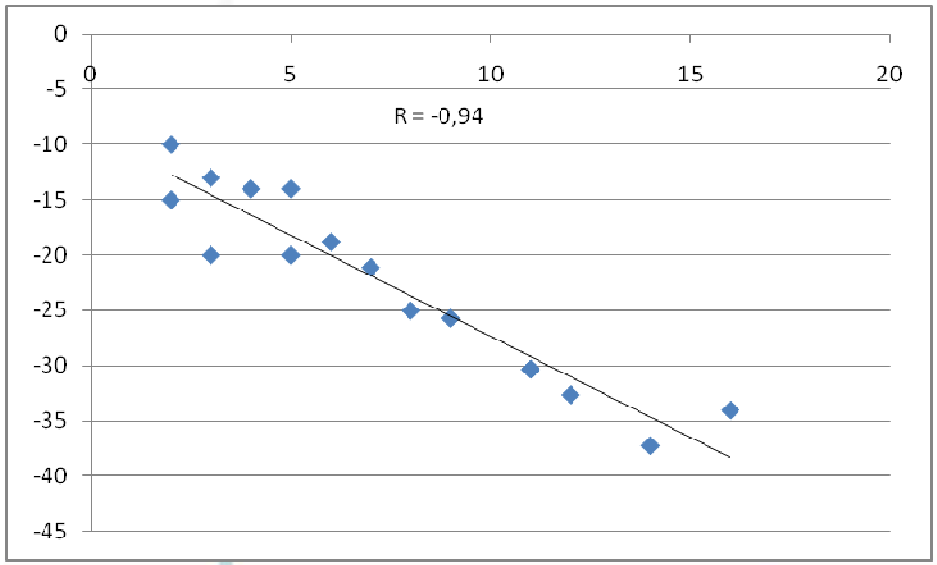
La démarche quantitative

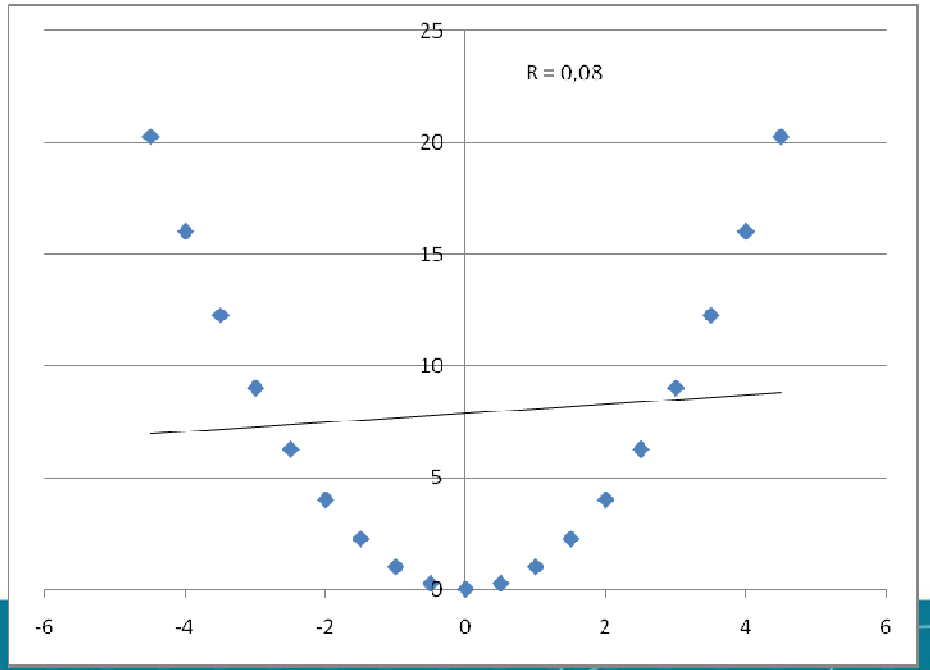
L'étude de la liaison entre deux variables quantitatives




R = 0,94
45
40
35
30
25
20
15
10
5
0
0
5
10
15
20


R = 0,002
12
11,8
11,6
11,4
11,2
11
10,8
10,6
10,4
10,2
10
9,8
0
5
10
15
20


R = -0,94
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
0
5
10
15
20


R = 0,08
25
20
15
10
5
0
-6
-4
-2
0
2
4
6

La démarche quantitative

L'étude de la liaison entre deux variables quantitatives

- Objectifs :
  - Etudier la liaison entre deux variables numériques
  - Coefficient de corrélation linéaire
  - Un test de signification associé
  - Un coefficient compris entre -1 et 1
  - H0 : le coefficient est égale à 0
  - H1 : le coefficient est différent de 0

Y	X	r	r^2	t	Pr(>|t|)
dureeM3	dureeM1	0,4773	0,2278	10,6158	0,0000

La démarche quantitative
La régression

- Objectifs :
  - Modéliser la liaison entre deux ou plusieurs variables numériques : y=ax+b
  - Estimation d'une valeur prédite et d'une valeur résiduelle
  - Deux tests statistiques :
  - Global sur le modèle
    - H0 : Il n'existe pas de modèle
    - H1 : Il existe un modèle
  - Sur les coefficients
    - H0 : les coefficients sont égaux à 0
    - H1 : les coefficients sont différents de 0
  - Analyser les résidus : graphes de nuages de points

La démarche quantitative

La régression

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	1,0176	1	1,0176	112,6955	0,0000
Residual	3,4493	382	0,0090		
Total	4,4669	383			

Coefficients

Attribute	Coef.	std	t(382)	p-value
Intercept	0,168090	0,014211	11,828148	0,000000
dureeM1	0,439533	0,041404	10,615812	0,000000

La démarche quantitative

La régression

Att. name	Full statistics

Statistics

Average	0,0000
Median	-0,0135
Std dev. [Coef of variation]	0,0949 [-99999,0000]
MAD [MAD/STDDEV]	0,0703 [0,7409]
Min * Max [Full range]	-0,29 * 0,44 [0,73]
1st * 3rd quartile [Range]	-0,06 * 0,05 [0,11]
Skewness (std-dev)	0,8333 (0,1245)
Kurtosis (std-dev)	2,7512 (0,2484)

Err_Pred_lmreg_1

Histogram

Values	Count	Percent	Histogram
x_<_-0,2181	5	1,30%	
-0,2181_=<_x_<_-0,1453	7	1,82%	
-0,1453_=<_x_<_-0,0725	59	15,36%	
-0,0725_=<_x_<_0,0003	138	35,94%	
0,0003_=<_x_<_0,0731	108	28,13%	
0,0731_=<_x_<_0,1458	44	11,46%	
0,1458_=<_x_<_0,2186	14	3,65%	
0,2186_=<_x_<_0,2914	3	0,78%	
0,2914_=<_x_<_0,3642	3	0,78%	
x>=_0,3642	3	0,78%	

La démarche quantitative

Logiciels statistiques

- SAS
- R
- Tanagra
- Spad
- Sphinx
- Modalisa
- SPSS
- DTM
- Chic
- ….

- ….

La démarche quantitative
Que retenir ….

- Les stats n'ont rien d'obligatoire
- Fixer un plan expérimental a priori
  - Variables indépendantes
  - Variables dépendantes
  - L'unité statistique
- Produire les données en respectant le plan expérimental
- Créer les indicateurs pour l'analyse
- Valider les données et tester l'adéquation aux lois (normale)
  - Indicateurs de base
  - Graphiques
- Poser les questions auxquelles les stats doivent répondre
- Choisir la méthode appropriée

Conclusion

- Pas de solution clé en main
  - Explorer les méthodes des SHS et des autres disciplines
  - Imaginer des combinaisons de méthodes originales
  - Inventer des jeux

- Des résultats de qualité des étapes essentielles
  - Identifier les différentes dimensions à mesurer pour répondre aux problèmes
  - Connaitre le contexte du produit ou du concept
  - Construire et rédiger protocole expérimental
  - Cibler les sujets et les recruter
  - Produire le matériel expérimental

  - Etablir un planning de réalisation de l'expérience
  - Faire les expériences, produire les données
  - Valider les données
  - Analyser les données
  - Rédiger les résultats

Deux manières de se tromper

- **Risque de 1ère espèce : risque α**
- Rejeter l'hypothèse nulle (H0) alors que celle-ci est vraie. Ce risque est parfaitement connu, c'est le seuil de probabilité que l'on se fixe pour rejeter l'hypothèse.

- **Risque de 2ème espèce : risque β**
- Accepter l'hypothèse H0 alors que celle-ci est fausse, c'est à dire que c'est H1 qui est vraie). Le fait d'accepter H0 est en effet la conséquence de ne pas avoir pu mettre en évidence une différence significative. En réalité cette différence peut exister, mais cette différence est trop faible pour être démontrée à l'aide du test utilisé. Le risque β peut être quantifié à partir de la loi de distribution des deux échantillons.

Questions ?

Coefficient de Variation

- Insee : Le coefficient de variation (CV) est le rapport de l'écart-type à la moyenne. Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande. Il est généralement exprimé en pourcentage. Sans unité, il permet la comparaison de distributions de valeurs dont les échelles de mesure ne sont pas comparables.