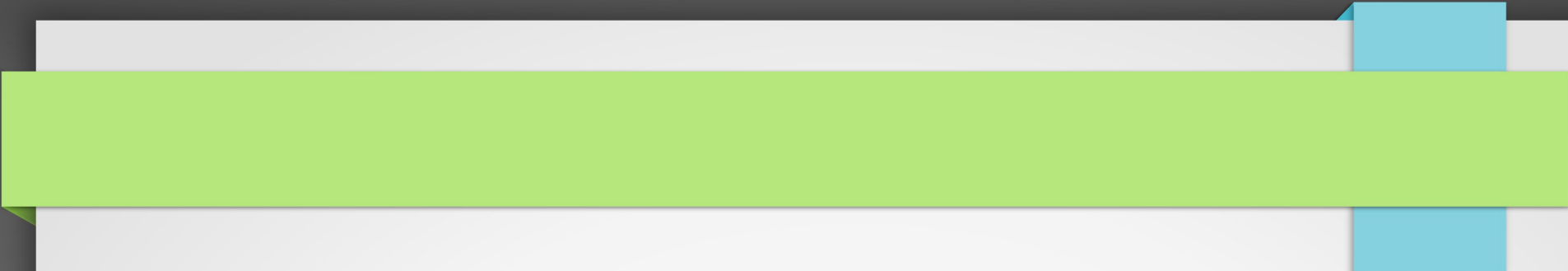


Recherche en pédagogie

Comment entrer les données de recherche dans un tableau, pour qu'elles soient exploitables statistiquement ?

- 
- Tout travail scientifique génère et/ou manipule des données (data : Point sur lequel on fonde un raisonnement <> setting)
 - 80 % du temps consacré à un travail scientifique est consacré à la préparation des données
 - Le traitement statistique des données est le principal frein à une production rapide de résultats
 - Mauvaise présentation des données
 - Difficultés de compréhension entre le producteur de données et le statisticien

80% of the effort of analysis is spent just getting the data ready to analyse

Utiliser le bon outil

- Saisie des données
 - Un tableur (open office)
 - Pour les données brutes
 - Pour les données nettoyées
 - Pour vous (si vous y tenez...)
- Exploitation statistique des données
 - Un logiciel de statistiques
- La communication entre les deux
 - Format CSV

Le format CSV

- CSV = comma separated virgule (données séparées par une virgule)
- Format universel
- Les colonne du tableur sont remplacées par des virgules
- Le nom du fichier se termine par **.csv**
- Il existe des variantes (tab,;)
- [\[http://fr.wikipedia.org/wiki/Comma-separated_values\]](http://fr.wikipedia.org/wiki/Comma-separated_values)

Format csv

Tableur

hopitaux.csv - LibreOffice Calc

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	calendrier	Sel	Col	Odi	Wis	Geb	Hus	3Fr	Mul	Hag	Dia	Alk	Sav	moyenne
2	2013-01-01	111	208	84	32	45	126	59	1	131	88	1	1	73.9166666666667
3	2013-01-02	80	197	69	49	42	125	38	1	112	89	1	1	67
4	2013-01-03	78	160	55	35	42	121	39	1	83	73	1	1	57.4166666666667
5	2013-01-04	65	170	67	24	30	121	42	1	92	93	1	1	58.9166666666667
6	2013-01-05	85	150	70	38	44	102	46	1	100	87	1	1	60.4166666666667
7	2013-01-06	68	167	79	36	43	93	38	1	90	77	1	1	57.8333333333333
8	2013-01-07	84	171	57	31	34	119	37	180	84	79	1	1	73.1666666666667
9	2013-01-08	84	168	69	23	44	113	42	1	82	69	1	1	58.0833333333333
10	2013-01-09	56	185	61	30	43	96	36	1	91	85	1	1	57.1666666666667
11	2013-01-10	75	184	64	37	45	77	44	1	81	66	1	1	56.3333333333333
12	2013-01-11	71	168	56	26	24	128	32	1	67	70	1	1	53.75
13	2013-01-12	64	163	72	31	33	116	34	1	95	75	1	1	57.1666666666667
14	2013-01-13	75	149	67	29	41	76	53	1	97	71	1	1	55.0833333333333
15	2013-01-14	64	173	63	35	39	123	37	168	103	86	1	1	74.4166666666667
16	2013-01-15	89	151	49	33	48	113	36	157	93	81	1	1	71
17	2013-01-16	63	160	57	26	29	117	50	149	98	68	1	1	68.25
18	2013-01-17	71	187	54	22	38	103	41	153	78	66	1	1	67.9166666666667
19	2013-01-18	78	184	71	25	33	105	43	81	93	73	1	1	65.6666666666667
20	2013-01-19	88	159	58	26	37	69	47	178	103	82	1	1	70.75

.CSV

hopitaux.csv x

```
1 calendrier,Sel,Col,Odi,Wis,Geb,Hus,3Fr,Mul,Hag,Dia,Alk,Sav,moyenne
2 2013-01-01,111,208,84,32,45,126,59,1,131,88,1,1,73.9166666666667
3 2013-01-02,80,197,69,49,42,125,38,1,112,89,1,1,67
4 2013-01-03,78,160,55,35,42,121,39,1,83,73,1,1,57.4166666666667
5 2013-01-04,65,170,67,24,30,121,42,1,92,93,1,1,58.9166666666667
6 2013-01-05,85,150,70,38,44,102,46,1,100,87,1,1,60.4166666666667
7 2013-01-06,68,167,79,36,43,93,38,1,90,77,1,1,57.8333333333333
8 2013-01-07,84,171,57,31,34,119,37,180,84,79,1,1,73.1666666666667
9 2013-01-08,84,168,69,23,44,113,42,1,82,69,1,1,58.0833333333333
10 2013-01-09,56,185,61,30,43,96,36,1,91,85,1,1,57.1666666666667
11 2013-01-10,75,184,64,37,45,77,44,1,81,66,1,1,56.3333333333333
12 2013-01-11,71,168,56,26,24,128,32,1,67,70,1,1,53.75
13 2013-01-12,64,163,72,31,33,116,34,1,95,75,1,1,57.1666666666667
14 2013-01-13,75,149,67,29,41,76,53,1,97,71,1,1,55.0833333333333
15 2013-01-14,64,173,63,35,39,123,37,168,103,86,1,1,74.4166666666667
16 2013-01-15,89,151,49,33,48,113,36,157,93,81,1,1,71
17 2013-01-16,63,160,57,26,29,117,50,149,98,68,1,1,68.25
18 2013-01-17,71,187,54,22,38,103,41,153,78,66,1,1,67.9166666666667
19 2013-01-18,78,184,71,25,33,105,43,81,93,73,1,1,65.6666666666667
20 2013-01-19,88,159,58,26,37,69,47,178,103,82,1,1,70.75
```

Les ensembles de données

- ensemble de données, fichier, feuille de calcul, tableur = dataset
- Format rectangulaire
 - en langage mathématique ***matrice***
 - en terme de base données ***table***
 - en langage de tableur ***feuille de calcul***
- Un dataset est habituellement un tableau
 - constitué de **lignes** et de **colonnes**.
 - Une ligne contient une **observation (une personne)**,
 - tandis que chaque colonne correspond à une **variable**.

Anatomie d'un tableur

- Lignes
 - Première ligne = en-tête (**Header**)
 - 1 ligne = 1 unité expérimentale (observation)
- Colonnes
 - 1 colonne = 1 variable (age, profession , score...)
 - 1ere colonne = **identifiant**

Anatomie d'un tableur

	A	B	C	D	E	F	G
1	titre						
2	Objectif(s)						
3	Investigateur						
4	Institution	CESU 67					
5							
6	unité	AS	Date de	DEC	Note	Pre	Post
7	expérimental	IDE	naissance	TRAD	globale	test	test
8		AMB					
9							
10	ID	METIER	AGE	GROUPE	NOTE	PT	OT
11	1	DE	22	DEC	10	6	6
12	2	DE	25	DEC	8	5	6
13	3	DE	23	DEC	7	5	5
14	4	DE	31	DEC	8	4	5
15	5	S	20	DEC	7	3	5
16	6	S	21	TRAD	9	5	8
17	7	S	24	TRAD	10	8	8
18	8	S	25	TRAD	10	7	7
19	9	MB	28	TRAD	5	5	4
20	10	MB	30	TRAD	6	4	4
21							

Anatomie d'un tableur (2)

- Un certain nombre de règles sont à respecter pour que les données soient exploitables.
- La feuille de saisie est divisée en 3 zones horizontales:
 - la première, facultative, sert à stocker les métadonnées [<http://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e>]
 - la seconde stocke le nom des colonnes
 - la troisième contient les données proprement dites
- La troisième zone est divisée verticalement en 2:
 - les colonnes les plus à gauche contiennent les variables de type "facteur" (factorielles : profession, type de pédagogie)
 - les colonnes les plus à droite contiennent les mesures

Anatomie d'un tableur

	A	B	C	D	E	F	G
1	titre						
2	Objectif(s)						
3	Investigateur						
4	Institution	CESU 67					
5							
6	unité	AS	Date de	DEC	Note	Pre	Post
7	expérimental	IDE	naissance	TRAD	globale	test	test
8		AMB					
9							
10	ID	METIER	AGE	GROUPE	NOTE	PT	OT
11	1	DE	22	DEC	10	6	6
12	2	DE	25	DEC	8	5	6
13	3	DE	23	DEC	7	5	5
14	4	DE	31	DEC	8	4	5
15	5	AS	20	DEC	7	3	5
16	6	AS	21	TRAD	9	5	8
17	7	AS	24	TRAD	10	8	8
18	8	AS	25	TRAD	10	7	7
19	9	AMB	28	TRAD	5	5	4
20	10	AMB	30	TRAD	6	4	4
21							

Large datasets can be split over several worksheets in the same workbook, for instance one worksheet containing the experiment details and one or more worksheets containing layout factors, treatment factors and measurement variables.

EXPERIMENT DETAILS Program 4 Project 4 experiment GLMT Principal scientist A. M. Heineman Collaborators E.K. Mengitch, B. Amadallo, A.D. Olang Location Maseno, Western Kenya Design Randomized Complete Block Design Objective To study the effects of mulch on crop performance date Aug 89 End of first cropping season				MEASUREMENT VARIABLES Experimental plots Experimental blocks Type of mulch Mulch intensity air dry cob weight Sample cob weight Total air cob weight (kg) (g) (kg)		
LAYOUT AND TREATMENT FACTORS Plot! Block! Type! Intensity!				a_cob	b_cob	c_cob
1	1	leu	5	1.27	305	17.7
2	1	con	0	1.95	161	19.9
3	1	gli	10	2.6	208	28.3
4	1	leu	10	2.7	220	25.7
5	1	gli	5	2.42	187	21.08
1	2	leu	5	2.3	178	26.63
2	2	gli	5	2.59	202	24.88
3	2	con	0	2.3	186	25.46
4	2	gli	10	2.27	179	12.5
5	2	con	0	2.39	198	16.06
1	3	con	0	2.45	196	14.35
2	3	gli	5	1.37	112	15.26
3	3	gli	10	2.45	185	25.56
4	3	leu	10	2.48	196	26.61
5	3	leu	5	2.34	187	22.01
1	4	gli	5	2.5	203	25.31
2	4	leu	5	2.4	195	18.91
3	4	leu	10	2.14	168	24.73
4	4	con	0	2.27	183	23.78
5	4	gli	10	2.49	195	22.87

SHORT
DESCRIPTIVE
COLUMN
HEADERS

STRATUM
AND
FACTOR
LEVELS

OBSERVATIONS

Métadonnées

- Métadonnées
 - ce sont des données à propos des données.
 - Elles servent à comprendre le contexte dans lequel se fait l'étude:
 - nom de l'auteur, des investigateurs
 - version du questionnaire
 - méthodes de description des variables: 'pour la rubrique sexe préciser H ou F'.
 - Cette zone est facultative ou faire l'objet d'un document séparé

Nom des colonnes

- une seule ligne.
- Sert à stocker le nom opérationnel des colonnes.
- Règles:
 - le nom est succinct (5 caractères)
 - il ne doit pas contenir d'espace ou de caractères qui pourrait être mal interprété (,;^)
 - remplacer les espace par des underscore (caractère 8)
 - pas de caractères accentués
 - Encodage : UTF-8

Mesures

- toujours utiliser le point décimal (jamais la virgule)
- une colonne ne peut contenir que des chiffres ou du texte, jamais des deux.
- Les dates de préférence au format ISO :
 - AAAA-MM-JJ ex. 2014-02-06
 - AAAA-MM-JJ HH:MM:SS ex. 2014-02-06 10:25:36
- Un logiciel ne sait pas distinguer les codes de couleur
- Un logiciel gère très bien les lettres
 - Non : 1, 1, 2, 2, 2, 0, 1, 2
 - Oui : H, H, F, F, F, NA, H, F
- Ne pas mélanger les majuscules et les minuscules
 - Homme <> Homme <> HOMME
- Valeur manquante
 - La déclarer explicitement (ne pas laisser de blanc)
 - Ne pas utiliser le zéro ou une valeur négative
 - Valeur par défaut : NA (not available).

Tidy data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
1	date	Méthode	individu	groupe	SEXE	METIE	AGE	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
2	10/1/2013	D	1	AVANT	H	Labo	53	3	4	3	3	2	3	2	5	5	3	2	1	5	4	2	2	5	3	4	3
3	10/1/2013	D	2	AVANT	F	Labo	20	5	5	2	5	4	4	2	6	6	5	6	1	5	4	4	5	1	4	4	4
4	10/1/2013	D	3	AVANT	F	Labo	34	5	6	4	5	3	4	4	5	5	3	5	2	3	4	4	3	1	1	4	3
5	10/1/2013	D	4	AVANT	F	AS	27	6	6	4	5	1	4	1	6	4	4	6	2	3	5	4	4	1	3	1	4
6	10/1/2013	D	5	AVANT	F	IDE	41	5	6	6	5	2	2	2	6	6	5	6	1	5	5	6	4	1	1	3	4
7	10/1/2013	D	1	APRES	H	Labo	53	3	4	4	4	3	2	2	4	4	3	4	2	3	4	3	4	5	3	3	3
8	10/1/2013	D	2	APRES	F	Labo	20	5	5	5	4	3	5	2	6	5	4	6	2	6	4	4	4	1	3	4	4
9	10/1/2013	D	3	APRES	F	Labo	34	5	6	5	5	2	4	3	5	5	1	5	3	5	5	4	3	1	2	4	3
10	10/1/2013	D	4	APRES	F	AS	27	6	6	5	5	1	4	1	6	5	1	6	1	4	5	5	4	1	1	1	4
11	10/1/2013	D	5	APRES	F	IDE	41	6	6	6	6	1	2	1	6	6	1	6	1	5	6	5	2	1	6	1	4

Google drive

Variables

- Synonymes : champs, colonne, attributs, caractéristiques.
- On distingue :
 - **variable d'entrée** (input variables, intrants) qui s'imposent à l'observateur
 - variables mesurées, prédicteurs, covariables, *variables indépendantes*
 - Elles sont généralement notées X1, X2, X3...
 - **variables de sortie** (output variables, extrants) qui sont « influencées » par les variables d'entrée
 - cible, réponse, variables dépendantes. Ce sont des variables .
 - Elles sont généralement notées Y1, Y2, Y3...
 - On appelle **identifiants** des variables qui ne participent pas à l'analyse mais qui servent uniquement à identifier de manière unique et non ambiguë chaque observation (n° d'ordre).
- En statistique on essaie de
 - mesurer les variables (statistique descriptive),
 - d'établir s'il existe un lien entre des variables d'entrées et de sortie,
 - si ce lien est dû au hasard ou non
 - si on peut établir un modèle (modélisation) qui partant des premières permet de retrouver les secondes

4 éléments sont nécessaires pour le statisticien

- 1. Les données brutes.
- 2. Les données nettoyées (tidy)
- 3. Un lexique décrivant chaque variable et ses valeurs possibles.
- 4. La recette utilisée pour passer du point 1 au points 2 et 3

Le lexique

- Il contient tout ce qui est nécessaire pour comprendre les tableaux de données
 - 1. Informations sur les variables (y compris les unités !)
 - 2. Informations sur les choix/compromis qui ont été faits, les codages, etc.
 - 3. Informations sur le protocole de l'étude
- Document de type word

Types de données

- Les variables stockent différents type de données
- Le type de la donnée influence le traitement statistique
- On divise habituellement les variables en 2 grandes catégories
 - Variable qualitatives
 - Variables quantitatives
-
- Les variables quantitatives ou numériques sont celles qui se mesurent avec des chiffres. Les variables quantitatives peuvent être discrètes (age) ou continues (poids). On peut discréditer des variables continues, l'inverse n'est plus difficile (date en secondes).
 - On peut leur appliquer des opérations simples (addtion)
 - Paramètres : moyenne, écart-type
 - Test paramétriques : student, analyse de la variance

Variables Qualitatives

- données qualitatives (nom, qualité d'un objet) ou **nominales** sont représentés par des chaînes de caractères:
 - sexe(homme, femme),
 - couleur,
 - les sentiments ou opinions (pas d'accord du tout, plutôt pas d'accord, neutre, plutôt d'accord, tout à fait d'accord), etc.
- Par ordre de complexité croissante on distingue :
 - variables **catégorielles**:
 - ne peut prendre qu'une valeur et une seule au sein d'une liste finie de valeurs (variable discrète),
 - par exemple le sexe. L'ordre des variables n'intervient pas. Un cas particulier sont les variables binaires (vrai, faux).
 - variables **ordinales** sont des variables catégorielles ordonnées
 - où les catégories sont mutuellement exclusives et possèdent un lien hiérarchique entre elles, sans que ce lien puisse être quantifié.
 - Les items de type Likert entrent dans cette catégorie.
 - Par exemple une échelle d'anxiété allant de 1 (pas du tout anxieux) à 7 (extrêmement anxieux). Les échelons ne sont pas équidistants: un score de 6 traduit une anxiété plus grande qu'un score à 3 mais une anxiété deux fois plus importante
 - Une échelle de likert (échelle composite constituée d'un ensemble d'items de Likert),
 - le score de glasgow, entre dans cette catégorie (polémique persistante).
 - variables de type **intervalles** sont des variables ordinales dont les échelons sont équidistants: thermomètre.
 - **ratios**: ce sont des intervalles possédant un zéro invariant comme la température en degré kelvin. L'âge peut entrer dans cette catégorie (quoique le point de départ soit controversé: conception, naissance)
- Paramètres de base : médiane, quantiles
- Tests non paramétriques (khi2, mann whitney...)

En résumé

- Données brutes (raw data)
 - Feuille de papier
- Données non préparées (messy data)
 - Données transcrites dans un tableur
 - les entête de colonne contiennent des chiffres au lieu de lettres
 - des variables multiples stockées dans la même colonne
 - les variables sont stockées en lignes et en colonnes
- Données nettoyées (tidy data)
 - 1. chaque variable mesurée doit figurer dans une colonne
 - 2. Chaque observation différente d'une variable doit figurer sur une ligne différente
- Le passage raw data ->tidy data entraine une altération des données. Il faut toujours fournir
 - Les données brutes (pour comprendre)
 - Les données nettoyées (pour gagner du temps)
 - La description du processus de nettoyage (quels compromis)

Messy data

- les entête de colonne contiennent des chiffres au lieu de lettres
- des variables multiples stockées dans la même colonne
- les variables sont stockées en lignes et en colonnes

Messy data

	grossesse	Pas de grossesse
Homme	0	5
Femme	1	4

Combien de variables ?

Tidy Data

grossesse	sexe	fréquence
non	femme	4
non	homme	5
oui	femme	1
oui	homme	0

Revenu et religion aux USA

	\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k	\$75-100k
Agnostique	27	34	60	81	76	137	122
Athée	12	27	37	52	35	70	73
Bouddhiste	17	21	30	34	33	558	62
Catholique	418	617	732	670	638	1116	949
Evangéliste	575	869	1064	982	881	1486	949
Hindou	1	9	7	9	11	34	47
Protestant	228	244	236	238	197	223	131
T.Jehovah	20	27	24	24	21	30	15
Juifs	19	19	25	25	30	95	69
Inconnu	15	14	15	11	10	35	21

Les en-têtes de colonne sont des valeurs pas des variables

Religion	Revenu	fréquence
Agnostique	\$10k	27
Athée	\$10k	12
Boudhiste	\$10k	17
catholique	\$10k	418
Evangeliste	\$10k	575
Hindou	\$10k	1
Protestant	\$10k	228
T. Jehovah	\$10k	20
Inconnu	\$10k	19
Religion	\$10-20k	15
Agnostique	\$10-20k	34
Athée	\$10-20k	27
...