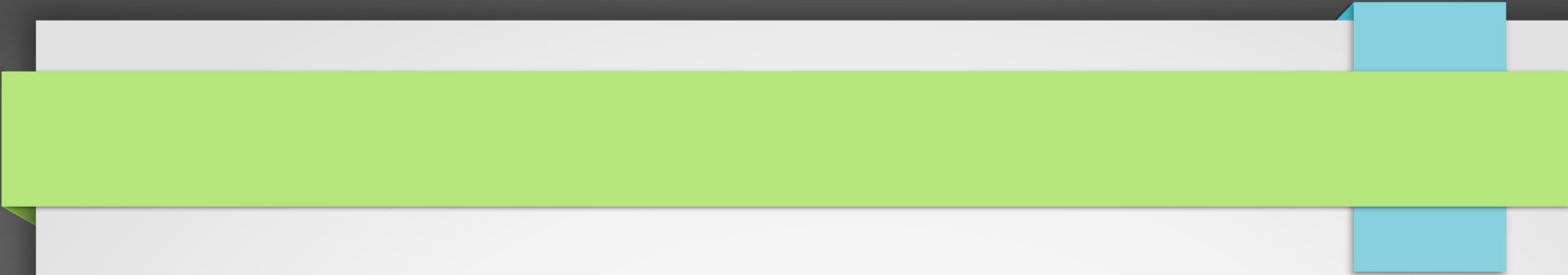




# Quand et comment mettre en œuvre une démarche de recherche quantitative ?

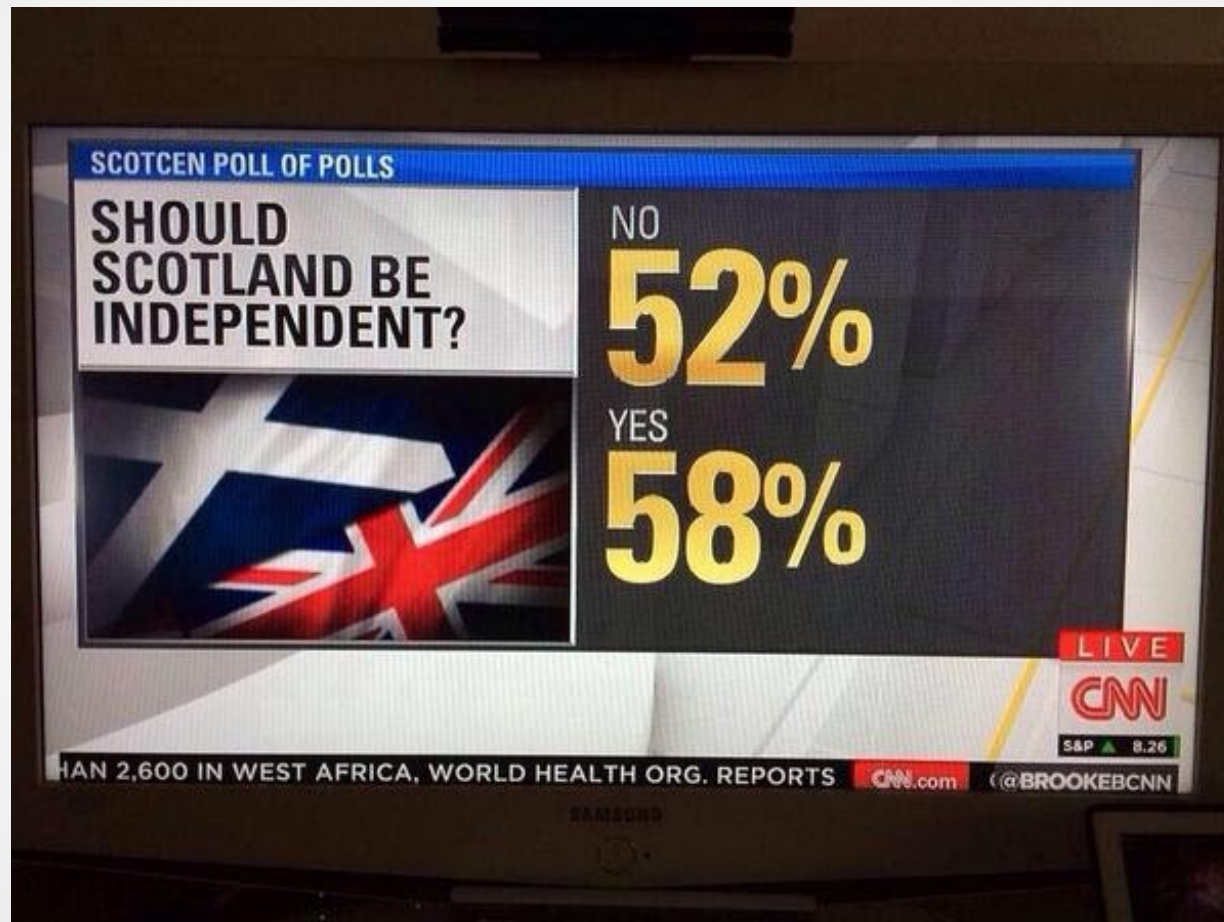
Dr J.C. Bartier  
Anesthésiste-Réanimateur  
**RESURAL**

[Jeanclaude.bartier@gmail.com](mailto:Jeanclaude.bartier@gmail.com)



S'il ne me restait plus qu'une heure à vivre,  
j'aimerais la passer dans un cours de statistiques :  
elle me semblerait tellement plus longue...

Tout le monde utilise les statistiques...



# Introduction

- La Recherche quantitative est définie comme une recherche qui effectue
  - la modélisation mathématique
  - et l'estimation statistique ou l'inférence statistique
  - ou un moyen pour tester des théories objectives
- en examinant les relations entre les **variables**.

# Introduction

- La statistique est l'outil principal de cette recherche
- En général un projet commence par une collection de données
  - Basée sur une théorie ou une hypothèse
- Suivi par l'application de méthodes statistiques
  - Descriptives
  - Inferentielles
- En général il est nécessaire de collecter un volume important de données qui doivent être
  - Validées
  - Vérifiées
  - Enregistrées

# Introduction

- La partie la plus difficile de tout le travail statistique est de commencer.
- Et l'une des choses les plus difficiles pour commencer est de choisir le bon type d'analyse statistique.
- Le choix dépend de
  - la nature de vos données
  - de la question particulière à laquelle on essaie de répondre.
- « La vérité est qu' il n'existe aucun substitut à l'expérience: le moyen de savoir ce qu'il faut faire, c'est d'avoir fait correctement beaucoup de fois avant. »
- *Statistics: An Introduction Using R, Second Edition. Michael J. Crawley. 2015 John Wiley & Sons, Ltd. Published 2015 by John Wiley & Sons, Ltd.*



Comment partager des données avec  
un statisticien

# Dualité

- le statisticien doit être capable de comprendre les données dans l'état où elles arrivent.
  - Il est important de voir les données brutes,
  - de comprendre les étapes du circuit du traitement,
  - être en mesure d'intégrer des sources cachées de variabilité dans l'analyse des données dans l'analyse.
- D'autre côté, pour de nombreux types de données, les étapes de traitement sont bien documentées et normalisées.
- Ainsi, le travail de conversion des données de la forme brute vers une forme directement analysable peut être effectuée avant de consulter un statisticien.
- Cela peut considérablement accélérer le temps de traitement, car le statisticien n'a pas à faire tout le prétraitement des données.



# Ce que vous devez fournir au statisticien

- Les données brutes.
- un ensemble de données ordonné  
[tidy data] (<http://vita.had.co.nz/papers/tidy-data.pdf>)
- Un lexique décrivant
  - chaque variable
  - et la fourchette de valeurs possibles
- La recette que vous avez utilisé pour passer de l'étape 1 aux étapes 2 et 3

# Les données brutes (raw data)

- Il est essentiel d'inclure les données sous forme brute (native)
- Voici quelques exemples de forme brute des données:
  - Fichier binaire fourni par un appareil de mesure
  - Un classeur Excel non formaté
  - Fichier au format Json fourni par twitter
  - Les chiffres saisis à la main tout en regardant à travers un microscope...

# Exemple de fichier binaire

```
0000000 0000 0001 0001 1010 0010 0001 0004 0128
0000010 0000 0016 0000 0028 0000 0010 0000 0020
0000020 0000 0001 0004 0000 0000 0000 0000 0000
0000030 0000 0000 0000 0010 0000 0000 0000 0204
0000040 0004 8384 0084 c7c8 00c8 4748 0048 e8e9
0000050 00e9 6a69 0069 a8a9 00a9 2828 0028 fdfe
0000060 00fc 1819 0019 9898 0098 d9d8 00d8 5857
0000070 0057 7b7a 007a bab9 00b9 3a3c 003c 8888
0000080 8888 8888 8888 8888 288e be88 8888 8888
0000090 3b83 5788 8888 8888 7667 778e 8828 8888
00000a0 d61f 7abd 8818 8888 467c 585f 8814 8188
00000b0 8b06 e8f7 88aa 8388 8b3b 88f3 88bd e988
00000c0 8a18 880c e841 c988 b328 6871 688e 958b
00000d0 a948 5862 5884 7e81 3788 1ab4 5a84 3eec
00000e0 3d86 dcb8 5cbb 8888 8888 8888 8888 8888
00000f0 8888 8888 8888 8888 8888 8888 8888 0000
0000100 0000 0000 0000 0000 0000 0000 0000 0000
*
0000130 0000 0000 0000 0000 0000 0000 0000
000013e
```

A hex dump of the 318 byte Wikipedia favicon, or Wikipedia's W.svg. The first column numerates the line's starting address, while the \* indicates repetition.

# Exemple de fichier JSON

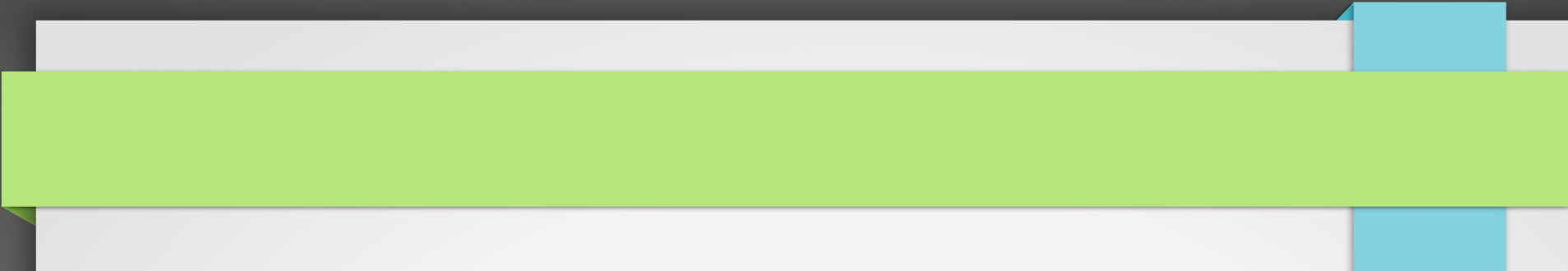
```
{
  "firstName": "John",
  "lastName": "Smith",
  "isAlive": true,
  "age": 25,
  "height_cm": 167.6,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021-3100"
  },
  "phoneNumbers": [
    {
      "type": "home",
      "number": "212 555-1234"
    },
    {
      "type": "office",
      "number": "646 555-4567"
    }
  ],
  "children": [],
  "spouse": null
}
```

# Les données brutes sont au bon format si:

- L'utilisation de ses données avec un logiciel échoue
- Aucune donnée n'a été manipulée
- Aucune donnée n'a été retirée
- Aucune donnée n'a été résumée de quelque manière que ce soit

# The tidy data set

- tidy = rangé, ordonné, utilisable.
- messy, messiness = malpropre, désordre
- Les principes généraux des tidy data ont été énoncé par
  - [Hadley Wickham](<http://had.co.nz/>)
  - dans [this paper (<http://vita.had.co.nz/papers/tidy-data.pdf>)
  - et dans cette [vidéo](<http://vimeo.com/33727555>).

- 
- 1. chaque variable mesurée doit figurer dans une colonne et une seule
  - 2. Chaque observation différente d'une variable doit figurer sur une ligne différente
  - 3. There should be one table for each "kind" of variable
  - 1. If you have multiple tables, they should include a column in the table that allows them to be linked

# Le lexique (code book)

- Les mesures qui sont faites devront être décrites avec plus de détail que ce qui figure dans la feuille de calcul . Le lexique contient ces informations . Au minimum, il devrait contenir :
  - Informations sur les variables ( y compris les unités !)
  - Informations sur les choix qui ont été faits
  - Informations sur le protocole de l'étude



# Le lexique (code book)

- Autre informations utiles
  - Comment a été conçue l'étude ?
  - Comment a été organisé la collecte de données
    - sont-ce les 20 premiers patients rencontrés dans la clinique ?
    - les patients ont-ils été sélectionnés sur certaines caractéristiques comme l'âge ?
    - Les traitements sont-ils attribués au hasard ?

# Le lexique (code book)

- Le format habituel pour ce document est un fichier type Word .
- Il devrait y avoir une section intitulée
  - « **Conception de l'étude** » qui décrit de façon détaillée la façon dont vous avez recueilli les données .
  - « **codage** » qui décrit chaque variable et ses unités.

# Comment coder les variables

- Dans une feuille de calcul, on peut distinguer quelques grandes catégories de données en fonction de leur nature [data type] ([http://en.wikipedia.org/wiki/Statistical\\_data\\_type](http://en.wikipedia.org/wiki/Statistical_data_type)):
  - Continues (quantitatives)
  - Discrète (qualitatives)
    - Nominales (catégorielles)
    - Ordinales
    - Intervalles
    - Ratio
  - Données manquantes
  - Données censurées

# Comment coder les variables

- variables **continues**
  - Ce sont toute variable mesurée sur une échelle quantitative continue comme par exemple le poids mesuré en kg.
- variables **ordinales**
  - ([http://en.wikipedia.org/wiki/Ordinal\\_data](http://en.wikipedia.org/wiki/Ordinal_data))
  - sont des variables réparties en niveaux, en nombre limité (< 100)
  - Ces niveaux sont ordonnés et l'ordre à un sens. (exemple une échelle de réponse (Likert): mauvais, passable, bon).
- variables **catégorielles**
  - [Categorical data]([http://en.wikipedia.org/wiki/Categorical\\_variable](http://en.wikipedia.org/wiki/Categorical_variable))
  - sont des variables où il existe plusieurs catégories, mais qui ne sont pas ordonnées. L'exemple classique est le sexe: homme ou femme.
- données **manquantes**
  - [Missing data]([http://en.wikipedia.org/wiki/Missing\\_data](http://en.wikipedia.org/wiki/Missing_data))
  - sont des données manquantes et irrécupérables.
  - Elles sont codées avec le **symbole NA** (not available).
- données **censurées**
  - [Censored data]([http://en.wikipedia.org/wiki/Censoring\\_\(statistics\)](http://en.wikipedia.org/wiki/Censoring_(statistics)))
  - sont des données **manquantes mais ont sait pourquoi**.
  - exemple classique : un patient perdu de vue.
  - Ils doivent aussi être codés `NA` quand vous n'avez pas les données. Mais vous devriez également ajouter une nouvelle colonne à vos données appelé, "VariableNameCensored" qui devrait avoir des valeurs de `true` si censuré et `false` si pas. Dans le lexique, il faut expliquer pourquoi ces valeurs sont manquantes. Il est absolument essentiel de mentionner à l'analyste, si il y a une raison connue pour que ces certaines données soient manquantes.
- On ne doit pas supprimer les valeurs manquantes

# Taxonomie des variables discrètes

Niveau de mesure	Groupes mutuellement exclusifs	Rangs ordonnés	Valeurs équidistantes	Référence nulle non arbitraire	exemples
Nominal	X				Status marital
Ordinal	X	X			Niveau de stress (1-7)
Intervalle	X	X	X		Echelle de dépression (1-100)
Ratio	X	X	X	X	Poids (kg)

(Stevens 1946)

# Tips

- En général éviter de coder variables catégorielles ou ordinales comme des nombres.
  - Lorsque vous entrez la valeur pour le sexe, il devrait être «homme» ou «femme».
  - Les valeurs ordinales dans le jeu de données doivent être "mauvais", "passable", "bon" et non 1, 2, 3.
- Cela permettra
  - d'éviter les ambiguïtés potentielles
  - et aidera à identifier les erreurs de codage.
- Les logiciels savent très bien manipuler ce type de données

# 3 objectifs

- Lire un article scientifique
- Comprendre/réaliser un protocole
- Ecrire un article

# Anatomie d'un article

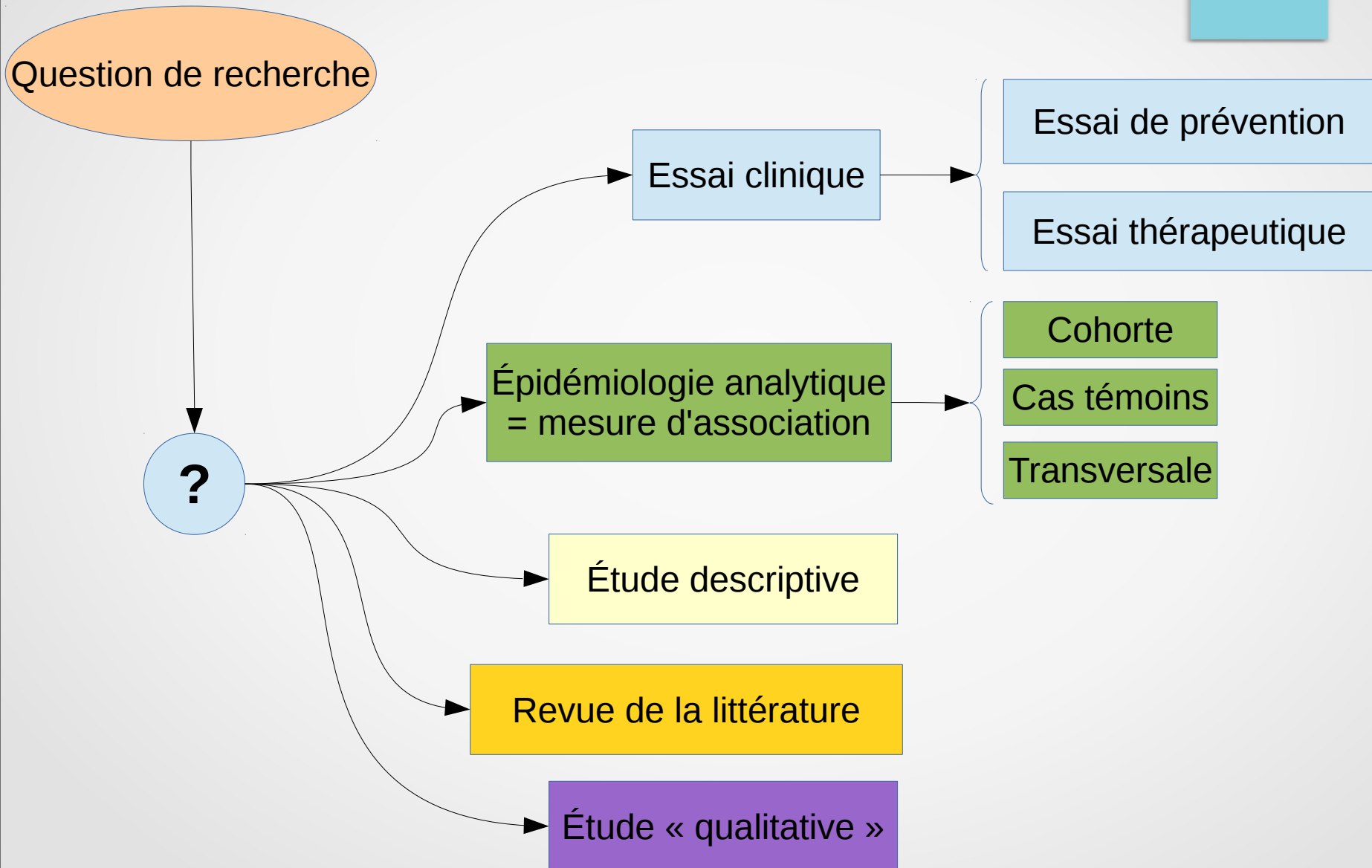
- Introduction
- Matériel et méthode
- Résultat
- Discussion
- Conclusion
- Bibliographie



# Quelle est la question ?

- Toujours poser la « question de recherche »
  - Unique et précise : Résumer un travail par une question simple
  - Pourquoi ? (pourquoi s'intéresser à cette question)
- Exemple
  - Thème : L'obésité a-elle une influence sur la réussite d'une analgésie péridurale ?
  - Question : existe t'il une relation entre un IMC élevé et l'échec d'une analgésie péridurale ?
  - Titre de l'article / mémoire
  - Plus difficile que cela en a l'air...

# Quelle méthodologie ?



# Donnée (Data)

- Ce que l'on observe, mesure
- Matière première du statisticien
- Tout travail scientifique génère et/ou manipule des données (data : Point sur lequel on fonde un raisonnement )
- 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

# Données (Data)

- 2 grands types
  - Quantitatives (continues)
  - Qualitatives (on ne peut pas les additionner) ou discrètes
    - Catégories non recouvrantes
    - Habituellement moins de 100 catégories
    - Ex : sexe (Homme, Femme, inconnu)
  - On peut assez facilement transformer une variable quantitative en qualitative (discrétisation). L'inverse n'est pas vrai :
    - EX : age et tranche d'age (INSEE)
- La distinction est importante dans le choix des méthodes statistiques applicables.

# 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)**,
  - Une colonne correspond à une **variable (ce que je mesure)**.

# Anatomie d'un tableur

- Lignes
  - Première ligne = en-tête (**Header**) = nom colonne
  - 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							

header

## 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



# 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) :  
Ceci\_est\_un\_exemple
  - pas de caractères accentués
  - Encodage : UTF-8
  - Ne pas fusionner les cellules

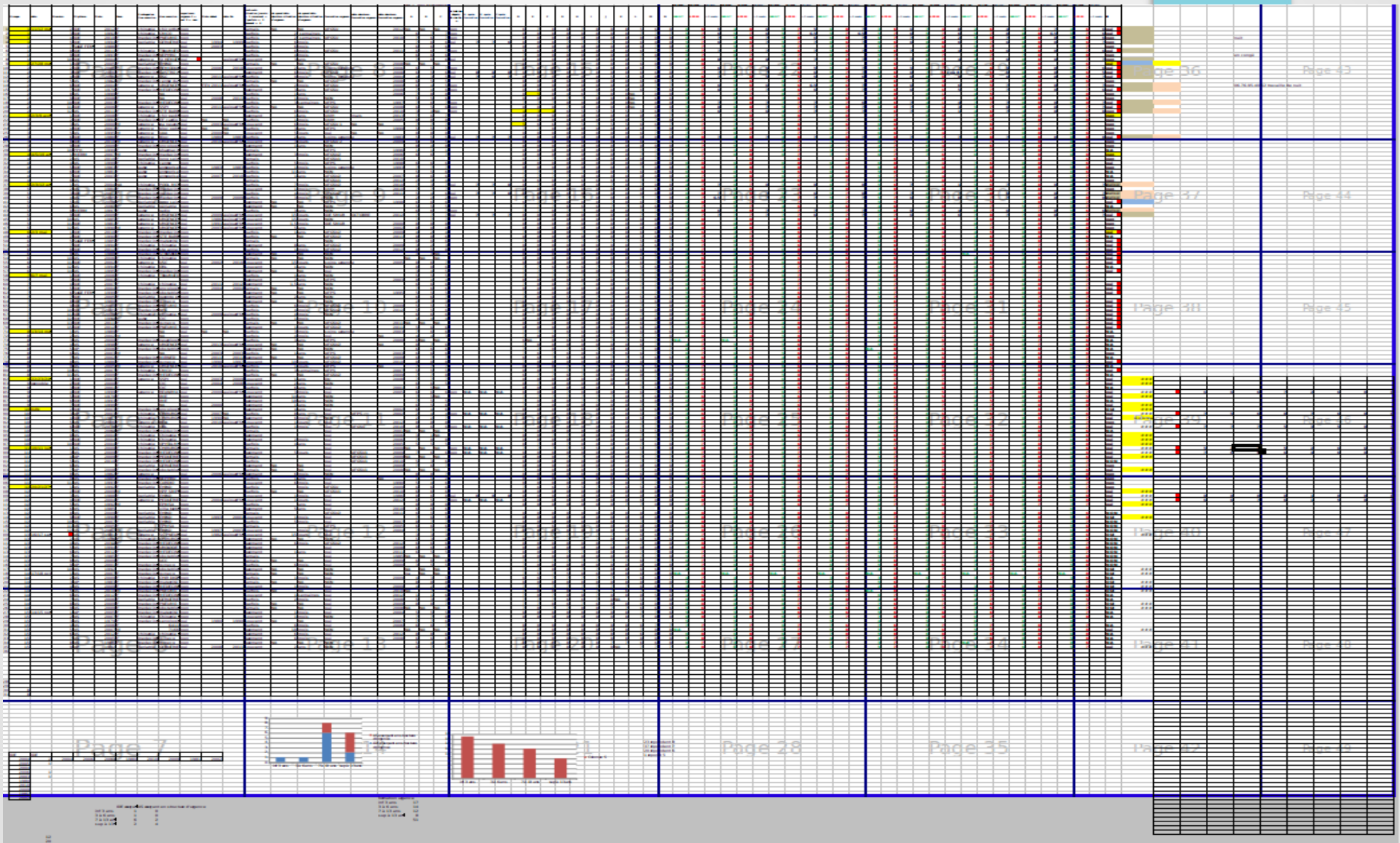
# Mesures

- 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).

# Présentation des données

- 80 % du temps de travail d'un statisticien est consacré au nettoyage des données
- Tableur (Excel, Libre office, etc.)
  - Très bon moyen de saisie des données
  - Très mauvais outil de traitement des données
- Logiciels de traitement (R, SPSS, STATA,...)
  - Mauvais moyens de saisie
  - Bon moyen de traitement
- Bons outils = tableur + logiciel spécialisé
  - Recherche reproductible

# Ce qu'il ne faut pas faire...



Une philosophie : recherche reproductible (reproducible research)

# Le format CSV

- CSV = comma separated virgule (données séparées par une virgule)
- Format universel (fichier/enregistrer sous)
- 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

## Tableau

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
```

# Organiser son travail

- Tableur
  - Données brutes (toujours conserver les originaux)
  - Données nettoyées (toujours travailler sur des copies)
  - Choisir son style
    - Solitaire : Open Office (Excel)
    - Collaboratif : cloud (Google drive) attention à la confidentialité...
- Notebook
  - Livre de bord horodaté
    - Métadonnées (dictionnaire des variables)
    - Où j'en suis
    - To Do (ce qu'il reste à faire)
- Dossier DOC
  - Articles
  - Biblio
- Dossier mémoire/thèse/article



## 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 aux points 2 et 3

Contacter le statisticien au stade du projet, jamais à la fin du travail...

# 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)

# Application

# Naissances dans une maternité anglaise

- Données provenant de 500 naissances uniques (singleton births) dans un hôpital de Londres.
- Métadonnées
- Un tableau (data frame) de 500 observations concernant les 8 variables suivantes:

Variable	Signification
id	Identifiant pour la mère et l'enfant
bweight	Poids de naissance
lowbw	Poids inférieur à 2500 g (0 = non, 1 = oui)
gestwks	Durée de la gestation (semaines)
preterm	Durée de gestation inférieure à 37 semaines (0 = non, 1 = oui)
matage	Age maternel
hyp	Hypertension gravidique (0 = non, 1 = oui)
sex	Sexe de l'enfant (1 = garçon, 2 = fille)

# Principaux paramètres statistiques

- Statistiques descriptives
  - Préalable à tout travail
  - Moyenne, variance, médiane, étendue
  - Généralement 1 variable à la fois (univarié)
- Deux variables (bivarié)
  - Étude du croisement des variables

# tests

	quantitatif	qualitatif
quantitatif	Coef. Corrélation Régression linéaire	
qualitatif	Boxplot ANOVA	Chi 2

- A partir de 3 groupes :
  - Analyse de la variance
    - Ex comparaison de l'age des femmes présentant une hémorragie au 1<sup>er</sup>, 2ème et 3ème trimestre

# Births

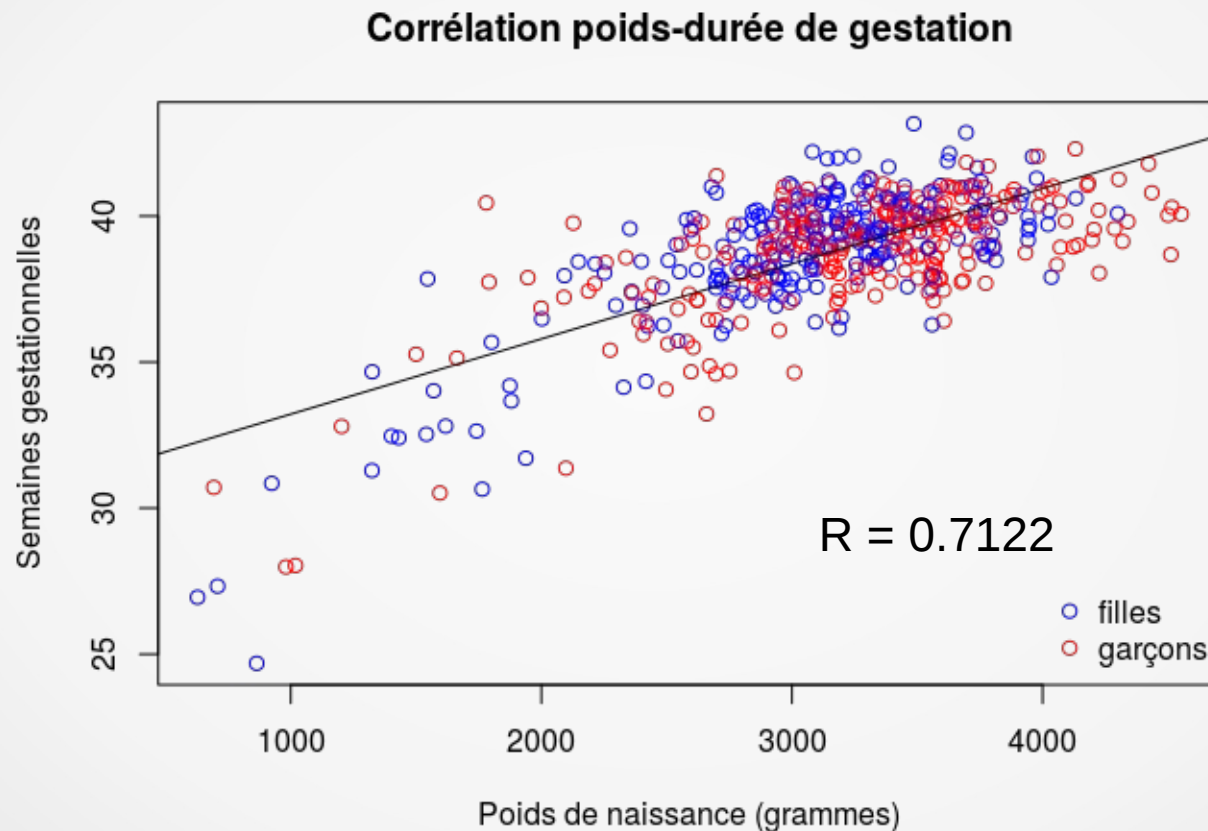
## Data

	id	bweight	lowbw	gestwks	preterm	matage	hyp	sex
1	1	2974	0	38.52	0	34	0	2
2	2	3270	0	NA	NA	30	0	1
3	3	2620	0	38.15	0	35	0	2
4	4	3751	0	39.80	0	31	0	1
5	5	3200	0	38.89	0	33	1	1
6	6	3673	0	40.97	0	33	0	2

## Statistiques univariées :

	id	bweight	lowbw	gestwks	preterm	matage	hyp	sex
Min.	: 1.0	Min. : 628	0:440	Min. :24.69	0 :427	Min. :23.00	0:428	1:264
1st Qu.	:125.8	1st Qu.:2862	1: 60	1st Qu.:37.94	1 : 63	1st Qu.:31.00	1: 72	2:236
Median	:250.5	Median :3188		Median :39.12	NA's: 10	Median :34.00		
Mean	:250.5	Mean :3137		Mean :38.72		Mean :34.03		
3rd Qu.	:375.2	3rd Qu.:3551		3rd Qu.:40.09		3rd Qu.:37.00		
Max.	:500.0	Max. :4553		Max. :43.16		Max. :43.00		
				NA's :10				

# Exemple de corrélation



Ne pas confondre corrélation et relation de cause à effet  
(QI et pointure)



# Comparaison de 2 moyennes : principe des tests d'hypothèse

Q : le poids de naissance des garçons est-il plus élevé que celui des filles ?

On part d'une hypothèse neutre (hypothèse nulle) :

« il n'y a pas de différence de poids à la naissance entre les garçons et les filles »

**Test de Student** t-test

```
data:  births$bweight by births$sex
```

```
t = 3.4916, df = 492.914, p-value = 0.0002617
```

```
Hypothèse alternative: true difference in means is greater than 0
```

```
Intervalle de confiance à 95 %:
```

```
86.17495 307.96706
```

```
Poids moyen Garçons  
3229.902
```

```
Poids moyen fille  
3032.831
```

Conclusion : p est inférieur à 5 % (0,05). On rejette l'hypothèse nulle et on retient l'hypothèse alternative : en moyenne, le poids de naissance des garçons est plus élevé que celui des filles.

# Deux variables quantitatives

Q : un poids faible à la naissance est indépendant de présence d'une HTA chez la mère ?

Poids faible	HTA gravidique		Sum
	non	oui	
non	388	52	440
oui	40	20	60
Sum	428	72	500

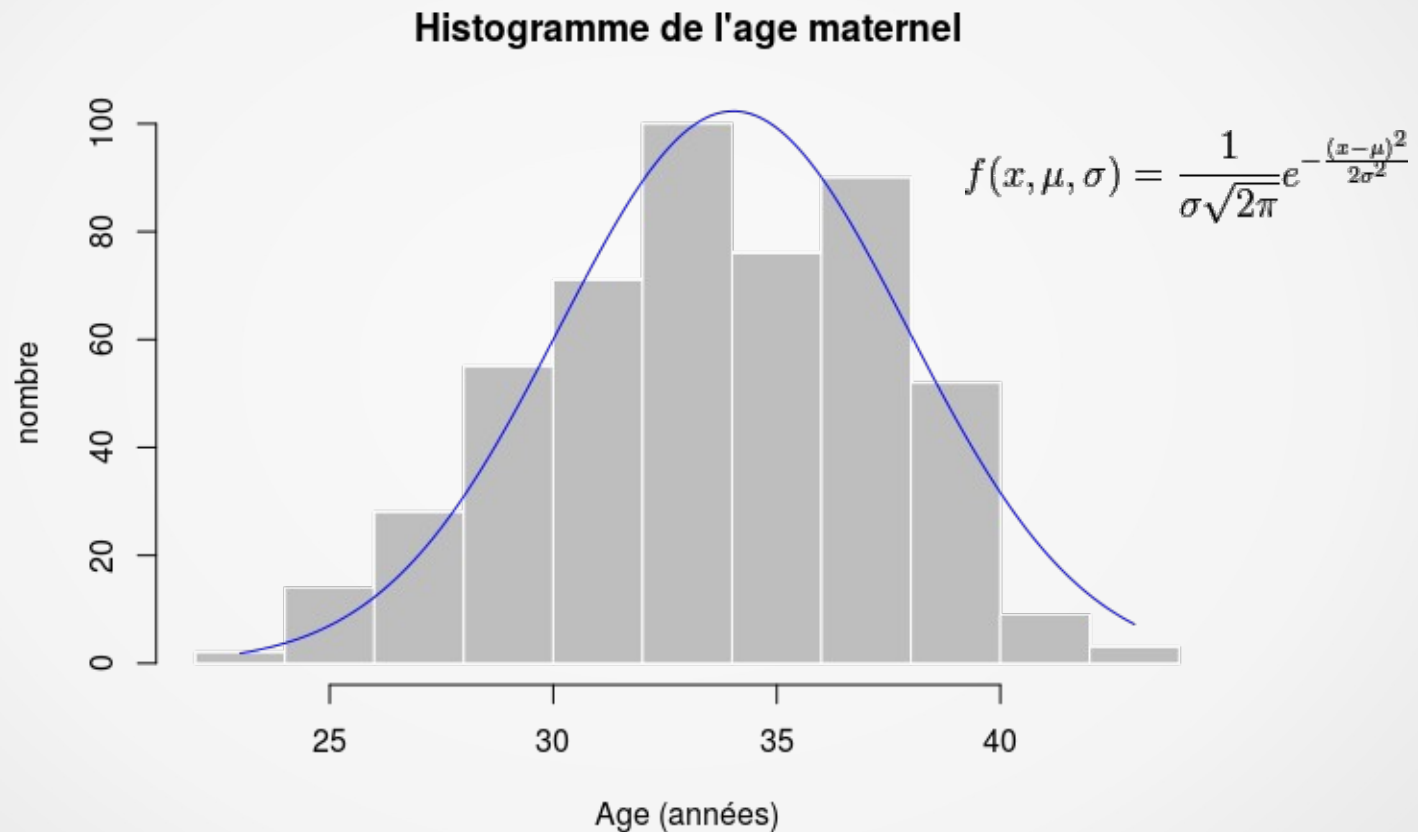
Pearson's **Chi-squared** test with Yates' continuity correction  
data: t  
X-squared = 18.12, df = 1, **p-value = 2.073e-05**

Conclusion : les 2 variables ne sont pas indépendantes. L'existence d'une HTA chez la mère a une influence sur le poids de l'enfant.

# Notion de hasard

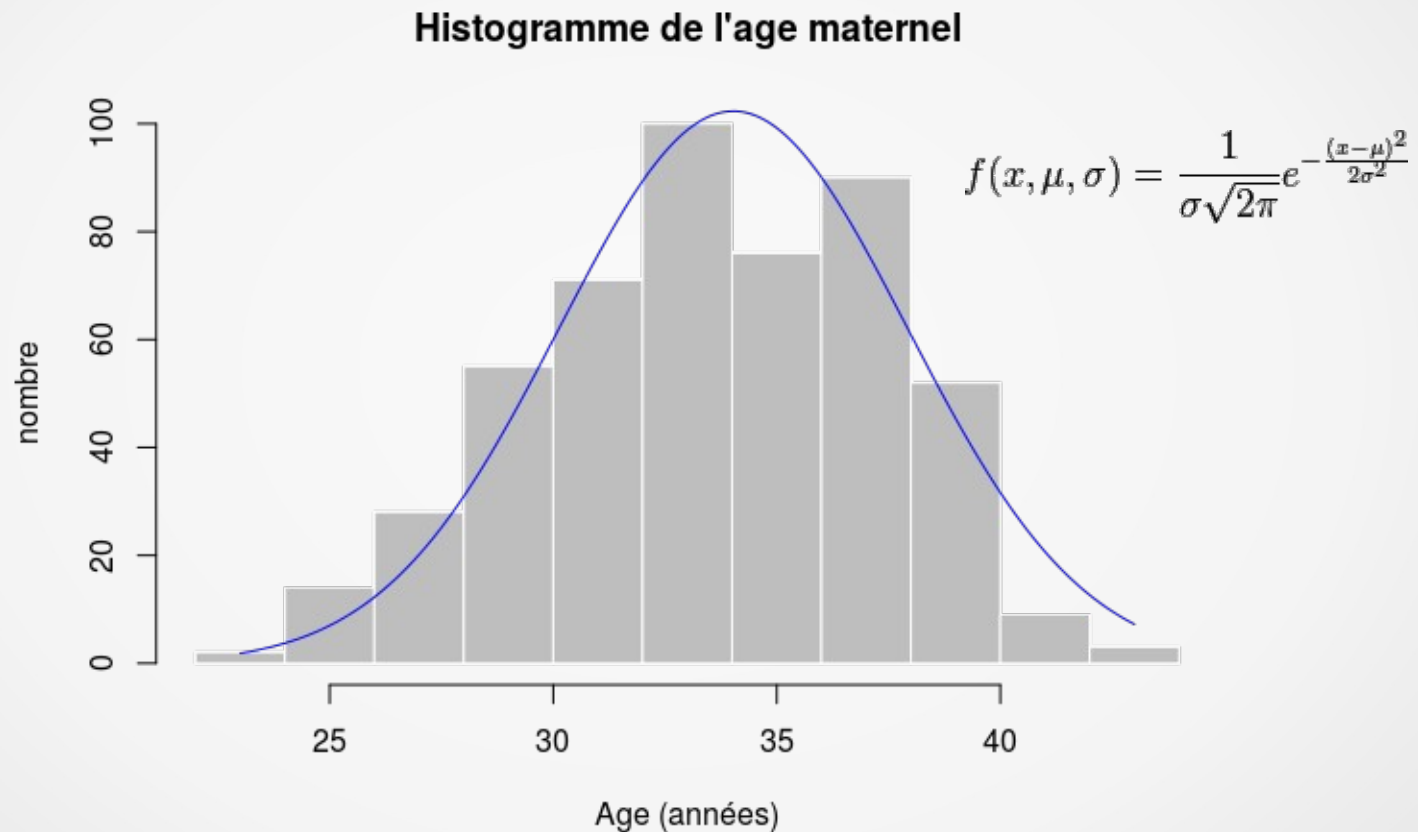
- Jet d'une pièce : la même pièce donne des résultats différents
- Lois statistiques
  - Loi normale (Gauss - Laplace)
  - Loi de Student (Gosset)
  - Loi de Poisson

# La loi normale



Deux paramètres : moyenne et variance permettent de résumer des centaines d'observations.

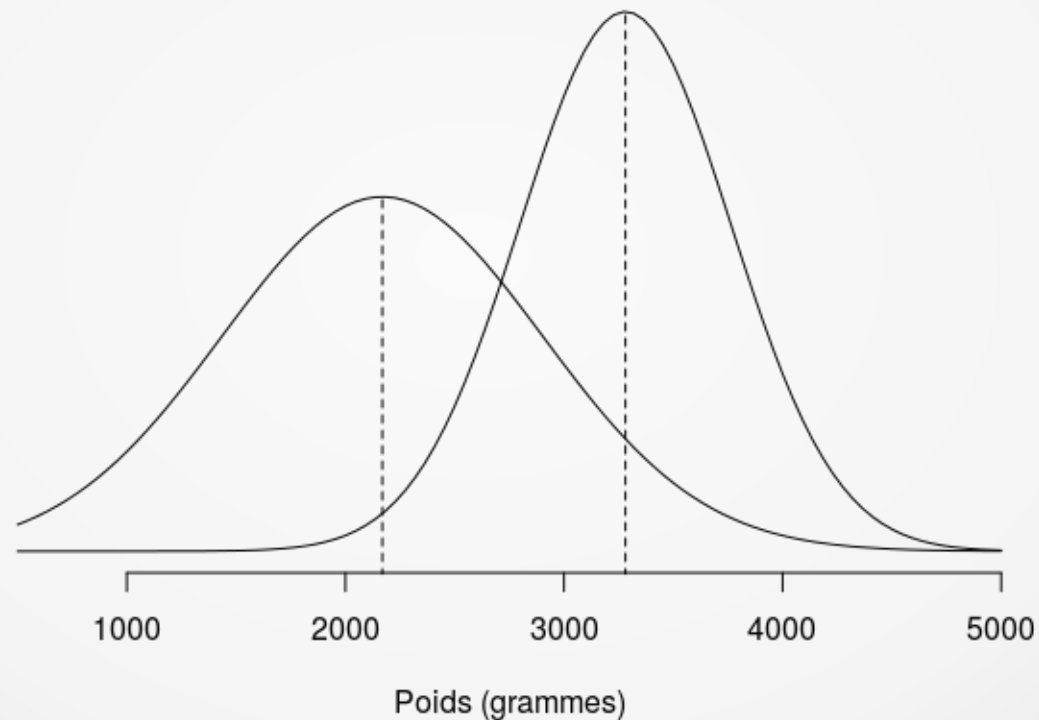
# La loi normale



Deux paramètres : moyenne et variance permettent de résumer des centaines d'observations.

# Sensibilité et spécificité

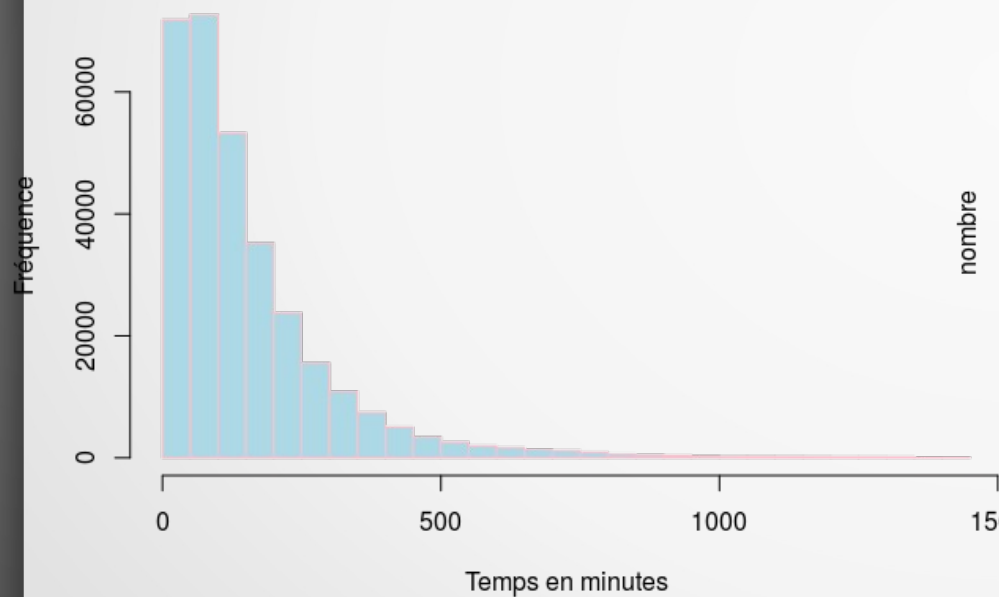
**Poids des enfants prématurés et nés à terme**



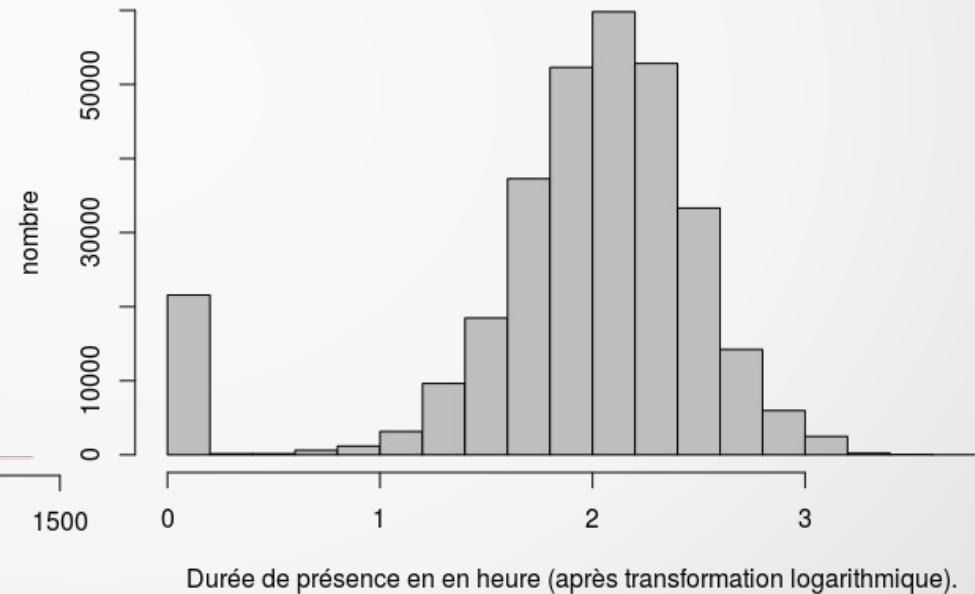
En fonction du poids peut-on dire si un enfant est né à terme ou pas ?

# Normaliser en transformant

Durée de présence



Durée de présence au SU en 2013



# Population et échantillon

- Etudes prospectives rétrospectives, cas témoins, cohorte
- Simple et double aveugle





Merci pour votre attention

SAV : [jeanclaude.bartier@gmail.com](mailto:jeanclaude.bartier@gmail.com)