COURS DE STATISTIQUES (24h)

- Introduction
- Statistiques descriptives (4 h)
- Rappels de Probabilités (4 h)
- Echantillonnage (4 h)
- Estimation ponctuelle (6 h)
- Introduction aux tests (6 h)

Qu'est-ce que la statistique?

- Les statistiques (descriptives) sont nées de l'activité de recueil des données répondant aux besoins d'organisation et de gouvernement des grands empires (armée, impôts, organisation des richesses). Ex: premiers recensements connus vers 3000 ans avant notre ère en Sumérie.

- Les statistiques sont aujourd'hui utilisées dans tous les secteurs d'activité :

  - Industrie : fiabilité, contrôle qualité, ….

  - Economie et finance: sondages, enquête d'opinion, assurance, marketing

  - Santé, environnement,…

  - Partout où l'on dispose de données

- ont connu un grand essor avec l'arrivage des ordinateurs performants

Qu'est-ce que la statistique?

- Vient du latin *status* = « état ». Le terme statisticum apparaît à la fin du XVII° siècle.

- Statistique = ensemble de méthodes permettant de décrire et d'analyser des observations (ou données). Ces observations consistent généralement en la mesure d'une ou plusieurs caractéristiques communes sur un ensemble de personnes ou d'objets équivalents.

- Remarque : une statistique = grandeur calculée à partir des observations recueillies (ex : moyenne d'âge des élèves d'une même classe, balance commerciale de la France, etc..)

Quelques définitions de base

- L'ensemble de personnes ou d'objets équivalents étudié s'appelle la population.
- Chaque objet d'une population s'appelle un individus ou unité statistique.
- Les caractéristique que l'on mesure s'appellent des variables.

  Les mesures s'appellent des observations.
- La série d'observations recueillies s'appelle série statistique. Elle est généralement retranscrite dans un tableau de données.

Rq : La statistique traite des propriétés des population plus que des individus particuliers de ces populations.

Quelques définitions de base

Exemple 1 : On s'intéresse aux débits annuels du Nil entre 1871 et 1970.

Variable étudiée=débit annuel ;

population= 100 années de 1871 à 1970.

Un individu= 1900 par exemple.

Série statistique (unidimensionnelle):

```
[1] 1120 1160  963 1210 1160 1160  813 1230 1370 1140  995
   935 1110  994 1020  960 1180  799  958 1140 1100 1210 1150
 [24] 1250 1260 1220 1030 1100  774  840  874  694  940  833
   701  916  692 1020 1050  969  831  726  456  824  702 1120
 [47] 1100  832  764  821  768  845  864  862  698  845  744
   796 1040  759  781  865  845  944  984  897  822 1010  771
 [70]  676  649  846  812  742  801 1040  860  874  848  890
   744  749  838 1050  918  986  797  923  975  815 1020  906
 [93]  901 1170  912  746  919  718  714  740
```

Quelques définitions de base

Exemple 2 : On s'intéresse à la fécondité en relation avec certains indicateurs socio-économiques dans 47 provinces francophones suisses vers 1888.

La série statistique (multidimensionnelles) est donnée dans le tableau de données suivant :

population

1 individus

1 variable

1 observation

	Fertility	Agriculture	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	12	9.96	22.2
Delemont	83.1	45.1	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	93.40	20.2
Moutier	85.8	36.5	7	33.77	20.3
Neuveville	76.9	43.5	15	5.16	20.6
Porrentruy	76.1	35.3	7	90.57	26.6
.....					

Fertility=indice de fécondité

Agriculture= % de males agriculteurs

Education= % d'individus ayant étudié après le primaire

Catholic=% de catholiques

Infant.Mortality=% mortalité infantile

Quelques définitions de base

- Recensement= Etude de tous les individus d'une population. Difficile en pratique lorsque les populations sont grandes pour des questions de coût et de temps.

$$\neq$$

- Sondage= recueil d'une partie de la population. La partie des individus étudiés s'appelle l'échantillon. Le recueil d'un échantillon à partir de la population initiale se fait par des techniques statistiques, appelées méthodes d'échantillonnage.

Quelques définitions de base

- Il existe différent types de variables

  - ✓ Variables quantitatives : caractéristiques numériques (taille, age,…). S'expriment par des nombres réels sur lesquels les opérations arithmétiques de base (somme, moyenne,…) ont un sens. Peuvent être discrètes (nombre fini ou dénombrable de valeurs : age,…) ou continues (toutes les valeurs réelles sont susceptibles d'être prises : taille,…).

  - ✓ Variables qualitatives : caractéristiques non numériques dans le sens où les opérations de base n'ont pas de sens. Peuvent être nominales (sexe,..) ou ordinales lorsque l'ensemble des catégories est muni d'un ordre total (très résistant, assez résistant, peu résistant,..). Les différents niveaux d'une variable qualitative s'appellent des modalités (ou catégories).

Quelques définitions de base

INFO

Une variable quantitative peut être mise sous forme qualitative ordinale en constituant des classes d'appartenance.

Exemple : On considère la population des salariés de France, le salaire mensuel S est une variable quantitative. On peut construire la variable SS qualitative ordinale à quatre modalités (S<6000 : modalité 1 ; 6000<S<10000 : modalité 2 ; 10000<S<20000 : modalité 3 ; S>20000 : modalité 4).

La création des amplitudes des classes est un problème délicat, qui nécessite un arbitrage entre information et simplification.

Les différentes problématiques de la statistique

- La statistique descriptive (ou exploratoire)

  - ✓ Objectifs :
    - résumer, synthétiser l'information contenue dans une série statistique, mettre en évidence ses propriétés.
    - suggérer des hypothèses relatives à la population dont est issu l'échantillon.

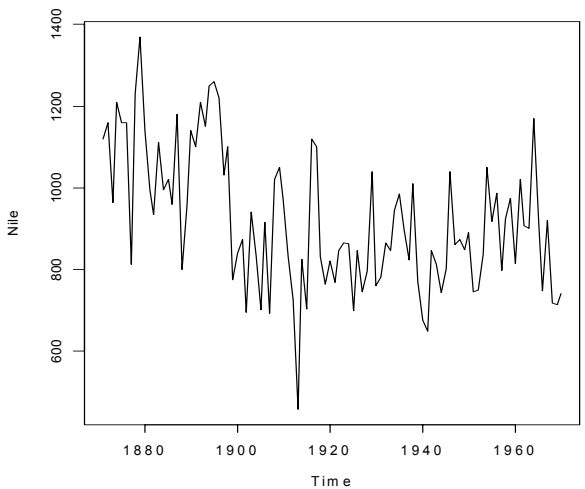
  - ✓ Outils utilisés :
    - Tableaux (table des fréquences,..)
    - Graphiques (box-plots, histogrammes,..)
    - indicateurs (moyenne, corrélation,..).

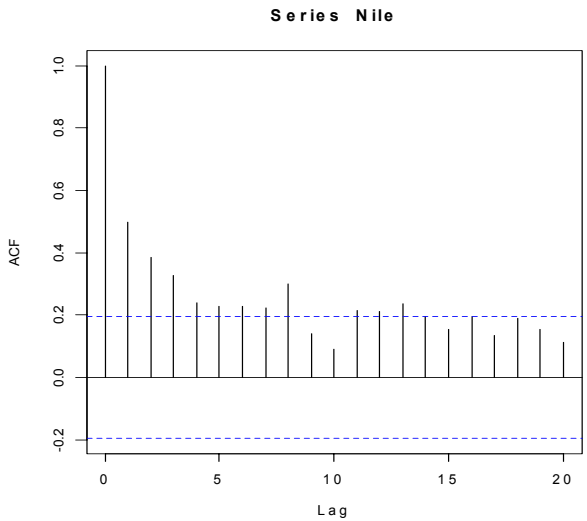
  - ✓ Méthodes :
    - Statistique descriptive classiques (uni et bidimensionnelles) Méthodes d'ADD.

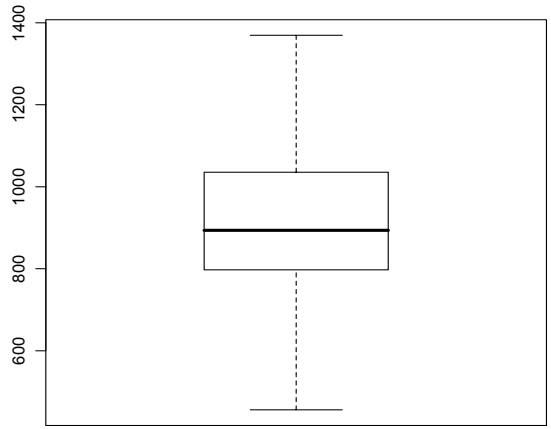
Les différentes problématiques de la statistique

Exemple 1 :

Graphiques :







Indicateurs :

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 456.0   798.5   893.5   919.4  1033.0  1370.0
```

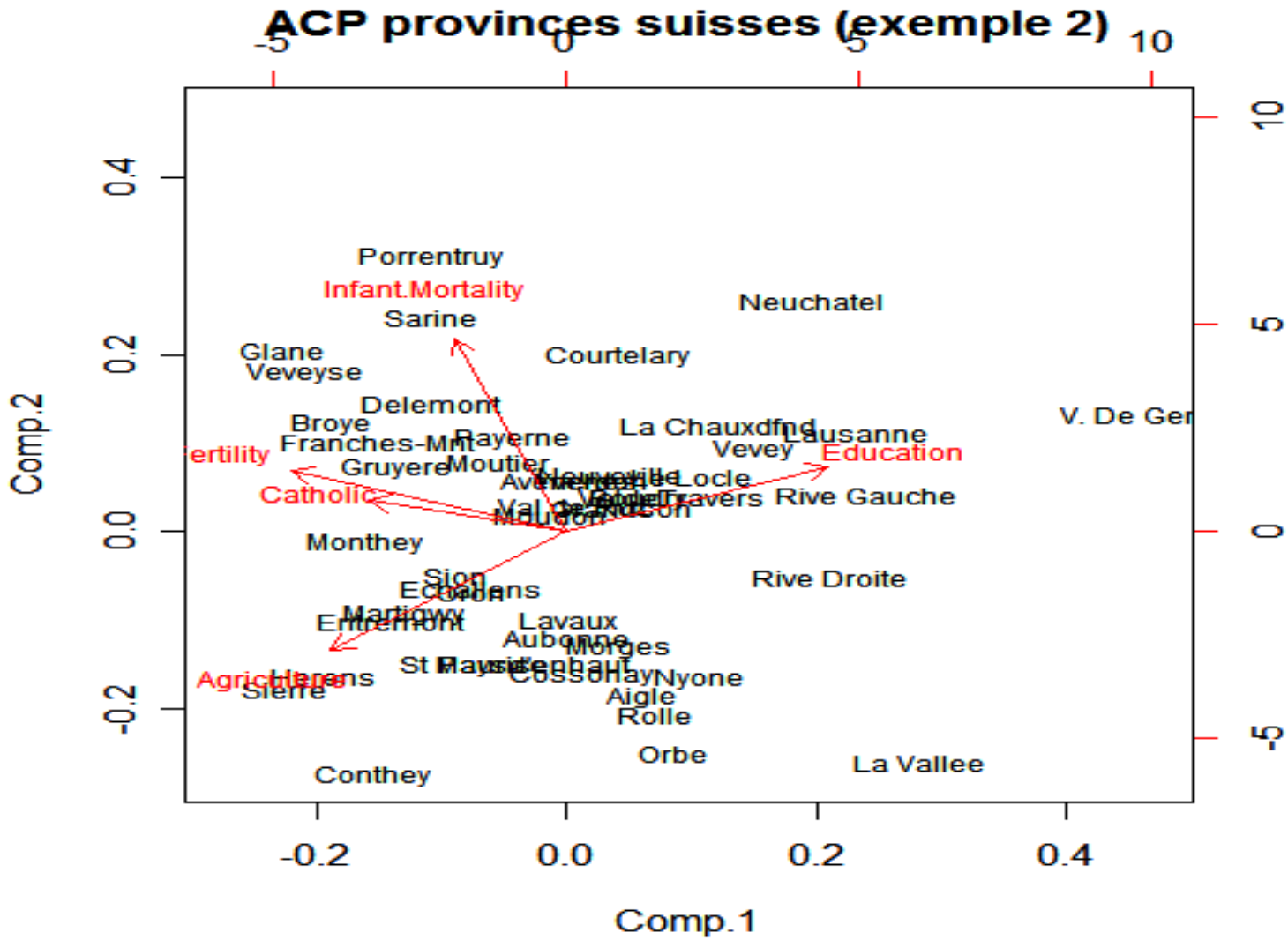
Les différentes problématiques de la statistique

INFO

La statistique descriptive s'est enrichie ces dernières années de nombreuses techniques de visualisation de données multidimensionnelles, connues sous le nom d'analyse des données, puis de data mining. Parmi ces méthodes on trouve :

- ✓ les méthodes de classification (partitionnement, CAH), visant à réduire la taille de l'échantillon en classant les individus dans des groupes de caractéristiques homogènes.

- ✓ les méthodes d'analyse factorielle (ACP, AFCM,…) qui cherchent à réduire le nombre de caractéristiques d'une population en les résumant par un petit nombre de composantes synthétiques.


ACP provinces suisses (exemple 2)
Comp.1
Comp.2
-0.2
0.0
0.2
0.4
-5
0
5
10
Porrentruy
Infant.Mortality
Sarine
Neuchatel
Glane
Veveyse
Courtelary
Delemont
Broye
Franches-Mnt
Rayerne
Moutier
Gruyere
Catholic
Vevey
Education
Locle
Rive Gauche
Monthey
Rive Droite
Lavaux
Morges
Nyone
Aigle
Rolle
Orbe
La Vallee
Conthey
Sierre

Les différentes problématiques de la statistique

- La statistique inférentielle (ou décisionnelle)

Inférence. Opération par laquelle on passe d'une vérité à une autre vérité, jugée telle en fonction de son lien avec la première. (*Petit Larousse*)

  - ✓ Spécificité :
    - La série de données est considéré comme un échantillon d'une population
    - suppose un modèle probabiliste sur la population.
    - Nécessite des méthodes d'échantillonnage.

  - ✓ Objectifs :
    - étendre (inférer) les propriétés constatées sur l'échantillon à la population.
    - Valider ou infirmer des hypothèses sur la population énoncées a priori ou formulées après une phase exploratoire.

  - ✓ Méthodes :
    - Estimation : approcher des paramètres de la population à partir de l'échantillon.
    - Tests : valider ou d'infirmer des hypothèses émises sur ces paramètres.
    - Modélisation et de prévision : recherche d'une relation entre une variable et plusieurs autres, valable pour l'ensemble de la population.

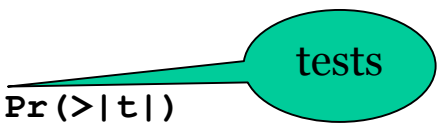
Les différentes problématiques de la statistique

Ex 2 : Modélisation par RLM : $F \approx 62.1 - 0.15A - 0.98E + 0.12C + 1.08I$

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       62.10131    9.60489   6.466 8.49e-08 ***
Agriculture       -0.15462    0.06819  -2.267  0.02857 *
Education         -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic           0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality   1.07844    0.38187   2.824  0.00722 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-Squared: 0.6993,     Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```



Rôle de la théorie des probabilités dans les problèmes de statistique

Probabilités = théorie permettant de modéliser des phénomènes aléatoires

Statistiques = repose sur l'observation de données issues d'un phénomène concret.

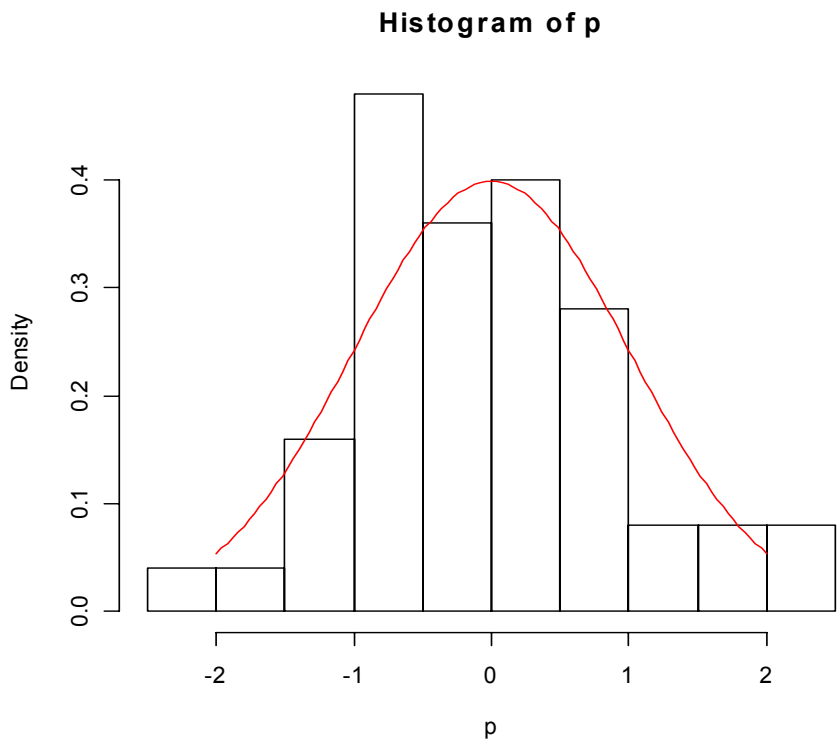
- Le rôle des probabilités est nul en statistique descriptive, prépondérant en statistique inférentielle.

- Les caractéristiques d'une grande population peuvent être considérées comme des variables aléatoires (on recode celles sont qualitatives). Les observations recueillies dans une série statistique peuvent être considérées comme des réalisations de ces variables.

- Lorsque l'échantillonnage est bien fait, on pourra approcher les caractéristiques théoriques (probabilistes) de la population (loi de probabilités etc…) à l'aide de statistiques calculées à partir d'un échantillon.

Rôle de la théorie des probabilités dans les problèmes de statistique

- Série de 50 observations issue d'une population gaussienne



- Série de 1000 observations issue d'une population gaussienne