PRE1: APPLIED STATISTICS

Marie-Anne Poursat

Laboratoire de Mathématiques d'Orsay (LMO)
Université Paris-Saclay

Sept 8, 2022

PRE1 : APPLIED STATISTICS

- Teacher : marie-anne.poursat@universite-paris-saclay.fr
  Laboratoire de Mathématiques d'Orsay and CELESTE Team INRIA Saclay
- Organization of the course : 40% lecture, 40 % training exercises, 20% lab, tutorial sessions will propose online notebooks posted on Google colab
  9h–12h, PUIO, D201
- Grading scheme : 100%CC
  - 1 time-limited test (written exercises) on week 4 (sept 29) : 25% of the final grade
  - 1 labwork (due on oct 8) : 15% of the final grade
  - Final test (oct 21) : 60% of the final grade

Program

1. prerequisites : random variables, probability distributions, descriptive statistics
2. Modeling data and fitting distributions, estimators
3. Parameter estimation, Maximum Likelihood methods
4. Laws of estimators : limit theorems, approximate laws, variance estimation, confidence intervals
5. Bootstrap : estimating standard errors and computing confidence intervals
6. Hypothesis testing
7. More on tests ; Final test

Statistical analysis for the data scientist

Objectives of data science

1. Collecting data
2. Processing data
3. Exploring and visualizing data
4. Analysing the data and applying learning to the data
5. Deciding

Steps 3 to 5 : using statistical thinking

Common terms :

- Statistical population and samples
- Random variables, probability
- Discrete and continuous data, probability distributions
- Modeling, fitting, statistical inference
- Classification and regression, machine learning, assessment

Ref : P. Cortez et A. Morais *A Data Mining Approach to Predict Forest Fires using Meteorological Data, Proceedings of the 13th EPIA (2007) pp. 512-523.*

```
'data.frame': 517 obs. of  11 variables:
xyarea month day   FFMC     DC    ISI  temp           RH  wind  rain lburned
 A86      8   sun -1.638  0.474 -1.562  1.53 -0.5692182 -0.74 -0.07   0.000
 A43      8   sun -1.638  0.474 -1.562  1.53 -0.7530703 -0.74 -0.07   2.007
 A24      8   sun -1.638  0.474 -1.562  0.52  1.6370062  0.99 -0.07   4.013
 A74      8   sun -1.638  0.474 -1.562  0.40  1.5757222  1.50 -0.07   2.498
 A14      8   sat  0.680  0.269  0.500  1.16 -0.1402302 -0.01 -0.07   0.000
 A63     11   tue -2.019 -1.779 -1.737 -1.22 -0.8143543  0.27 -0.07   0.000
 ...
```

- identify the variables
- dimension of the data set ?
- format of the values of the variables ? range ?

Data summary :

```
   xyarea     month     day          FFMC                  DC
 A86 : 52    8 :184   sun:95   Min.    :-13.033000   Min.    :-2.1770000   Min.
 A65 : 49    9 :172   mon:74   1st Qu.: -0.081000   1st Qu.:-0.4440000   1st
 A74 : 45    3 : 54   tue:64   Median :  0.173000   Median : 0.4690000   Medi
 A34 : 43    7 : 32   wed:54   Mean   : -0.000039   Mean   : 0.0000387   Mean
 A44 : 36    2 : 20   thu:61   3rd Qu.:  0.409000   3rd Qu.: 0.6690000   3rd
 A24 : 27    6 : 17   fri:85   Max.   :  1.006000   Max.   : 1.2600000   Max.
Other:265   Other:38 sat:84
 ....
```

- quantitative variables : `numeric` (integer, float)
- categorical variables : `dtype='object'` (`xyarea`, `month`, `day`)

Random variables

one of the fundamental ideas of probability theory

A random variable X is essentially a random number.

- a *discrete* random variable : only a finite or at most a countably infinite number of values.
  The probabilities of the outcomes of X are given by the frequency function or *probability mass function* (PMF) : $pmf(x_i) = \mathrm{P}(X = x_i)$, $\sum_i pmf(x_i) = 1$, $i = 1, 2, \ldots$
  Examples : Bernoulli, Binomial, Poisson.
- *Continuous* random variables : a continuum of values, in an interval of $\mathbb{R}$.
  The role of the frequency function is taken by a *probability density function* (PDF) $pdf(x)$ with properties :
  $pdf(x) \geq 0, \quad \int pdf(x)dx = 1.$

$$\mathrm{P}(a < X < b) = \int_a^b pdf(x)dx$$

  Examples : Normal (Gaussian), Exponential, Gamma.

CDFs are useful for comparing distributions

$$cdf(x) = \mathrm{P}(X \leq x), \qquad -\infty < x < \infty$$

- If X is discrete, cdf is an non-decreasing step function,
  $0 \leq cdf(x) \leq 1$.
  The cdf jumps wherever $pmf(x_i) > 0$ and the jump at x_i is $pmf(x_i)$.
- If X is continuous, cdf is a non-decreasing continuous function,
  $0 \leq cdf(x) \leq 1$.
  $cdf(x) = \int_{-\infty}^{x} pdf(x)dx, \quad pdf(x) = cdf'(x),$
  $\mathrm{P}(a < X < b) = cdf(b) - cdf(a)$

Support of the distribution : $S = \{x, pdf(x) > 0\}$.

Quantiles

The *p*th quantile of the distribution of X is the value x_p such that

$$cdf(x_p) = p$$

or

$$x_p = cdf^{-1}(p)$$

if the inverse of cdf is well defined. Otherwise, $cdf^{-1}(p) = inf\{x,\ cdf(x) \geq p\}$.

Exercice : if X is a continuous variable with CDF F and U is a uniform variable on $[0, 1]$, then

- X and $F^{-1}(U)$ have the same distribution F,
- $F(X)$ is a uniform variable on $[0, 1]$.

Expectation

Definition of the expected value or expectation or theoretical mean of X :

- If X is discrete, $\mathrm{E}(X) = \sum_{S} x_i pmf(x_i)$.
- If X is continuous, $\mathrm{E}(X) = \int_{S} x \; pdf(x) \; dx$.

Linearity property : if $X_1, \ldots, X_n$ are n random variables,

$$\mathrm{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i).$$

Variance and standard deviation

Definition : $\mathrm{Var}(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^2\right)$

The standard deviation is the squared root of the variance :

$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$

Notations : $\mu = \mathrm{E}(X)$, $\sigma = \mathrm{sd}(X)$

- if X is discrete, $\mathrm{Var}(X) = \sum_S (x_i - \mu)^2 pmf(x_i)$,
- if X is continuous, $\mathrm{Var}(X) = \int_S (x - \mu)^2 pdf(x)$.

Properties :

- $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$
- $\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X)$
- $\dfrac{X - \mathrm{E}(X)}{\mathrm{sd}(X)}$ is a centered variable with variance 1
- Chebyshev's inequality : $\mathrm{P}(|X - \mu| > t) \leq \dfrac{\sigma^2}{t^2}$

The Exponential Distribution

The family of exponential densities depends on a single parameter $\mu > 0$:

$$pdf(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) 1_{x \geq 0}$$

- $cdf(x) = [1 - e^{-\frac{x}{\mu}}] 1_{x \geq 0}$
- $\mathrm{E}(X) = \int_0^{\infty} \frac{x}{\mu} \exp\left(-\frac{x}{\mu}\right) dx = \mu, \ \mathrm{Var}(X) = \mu^2$
- from $cdf(x) = 1/2$ we have median$=\mu \log 2$

The exponential distribution is often used to model lifetimes or waiting times.

Correlation

- Joint distribution of X and Y : $cdf_{(X,Y)}(x,y) = \mathrm{P}(X \leq x, Y \leq y)$
- X and Y are independent if $cdf_{(X,Y)}(x,y) = cdf_X(x)\, cdf_Y(y)$ for all (x,y).
- *Definition* of the covariance of X and Y :

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X-\mu_X)(Y-\mu_Y)\right)$$

- correlation coefficient is defined by :

$$\mathrm{corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$$

- X and Y independent $\Longrightarrow \mathrm{corr}(X,Y) = 0$ but the reverse is not true !

Descriptive statistics

The first part of the course deals with analytical probability distributions. They are theoretical functions used to model the generative distribution of the data.
The second part of the course is about empirical distributions that are based on empirical observations (finite samples).

Empirical methods are useful for

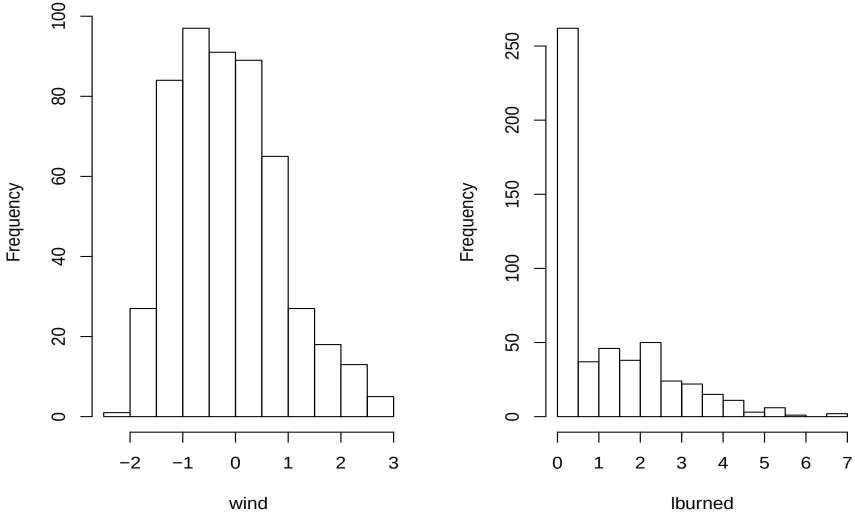
- summarizing data
- revealing the structure of data
- generating graphical representations
- choosing a model

$x_1, x_2, \ldots, x_n$, a serie of numbers = independent realisations of X

- Empirical mean : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median : the 50% sample quantile $x_{(\frac{n+1}{2})}$ if n is odd, $(x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})})/2$ if n is even, where $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

```
  Fires.wind           Fires.lburned
Min.    :-2.0200000   Min.   :0.000
1st Qu.:-0.7400000    1st Qu.:0.000
Median :-0.0100000    Median :0.419
Mean   :-0.0003675    Mean   :1.111
3rd Qu.: 0.4900000    3rd Qu.:2.024
Max.   : 3.0000000    Max.   :6.996
```

Exemple Fires



- range : $x_{(n)} - x_{(1)}$
- empirical variance : $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- 1st quartile Q1 = 25% sample quantile, 3rd quartile Q3, interquartile range=Q3-Q1

```
Fires.day     Fires.DC
 sun:95     Min.    :-2.1770000
 mon:74     1st Qu.:-0.4440000
 tue:64     Median : 0.4690000
 wed:54     Mean   : 0.0000387
 thu:61     3rd Qu.: 0.6690000
 fri:85     Max.   : 1.2600000
 sat:84
```

Discrete sample : counting table

Frequency plots

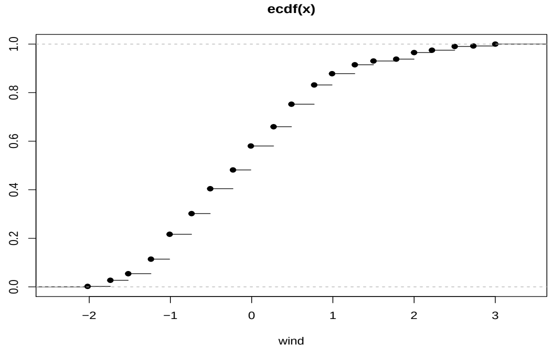
1. qualitative or discrete variables : bar plots, pie charts
2. quantitative variables : box plots, histograms (normalize such that the area is 1)

Probability plots They are useful to assess the fit of data to a theoretical distribution

- The empirical cumulative distribution function (ecdf)
- Quantile-quantile plots or QQplots

Empirical cumulative distribution function : $ecdf(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq x}$

$\rightarrow$ data analogue of the CDF of a random variable.
$\rightarrow$ CDF of the empirical probability with support $\{x_1, \ldots, x_n\}$ and mass $1/n$ at each point.

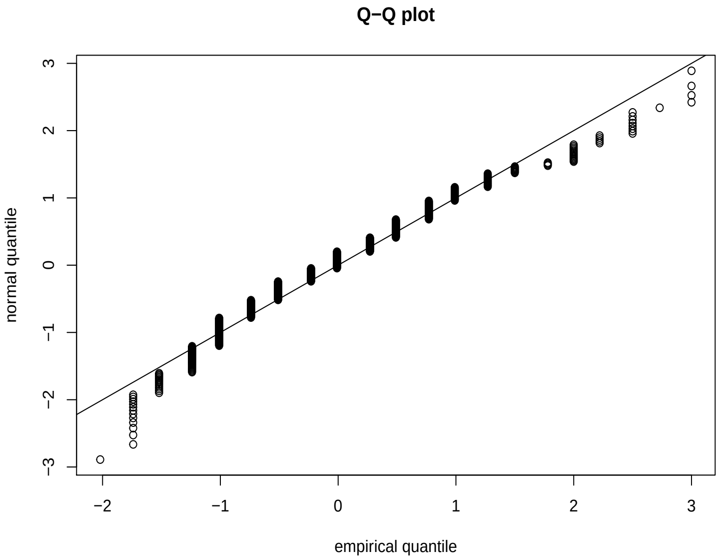


- Consider $x_1, \ldots, x_n$ sample from a uniform(0,1) law
  Ordered sample values : $x_{(1)} < x_{(2)} \leq \ldots < x_{(n)}$
  We have $\mathrm{E}(X_{(j)}) = \dfrac{j}{n+1}$
  $\rightarrow$ plot the ordered sample against the expected values $1/(n+1), \ldots, n/(n+1)$
  $\rightarrow$ if the underlying law is uniform, the plot should look linear.
- If $x_1, \ldots, x_n$ sample from F, plot $F(x_{(j)})$ vs $\frac{j}{n+1}$ or equivalently

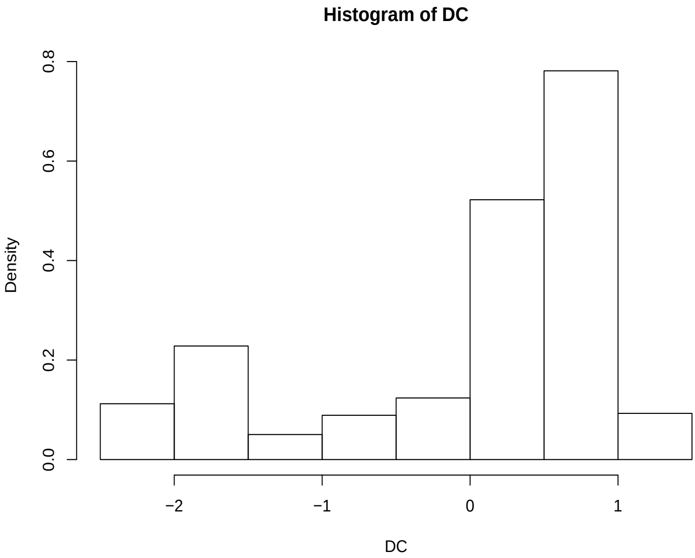
$$x_{(j)} \quad \text{vs} \quad F^{-1}\left(\frac{j}{n+1}\right)$$

- Q-Q plot : empirical quantiles versus the quantiles of F

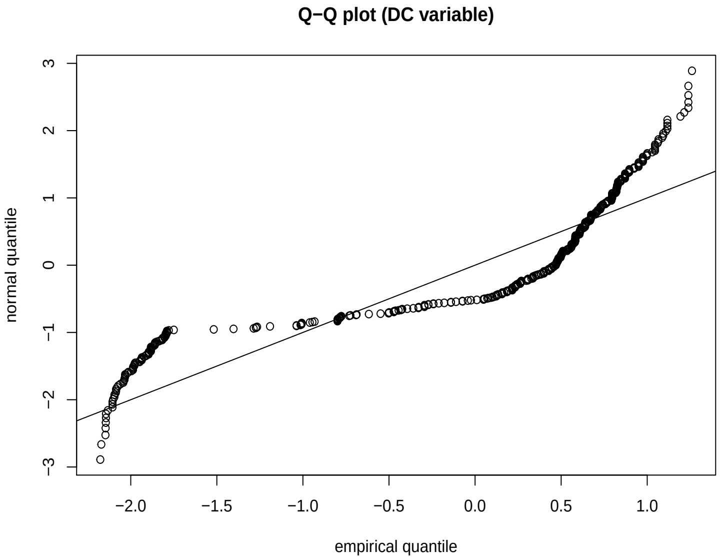
Variable Wind



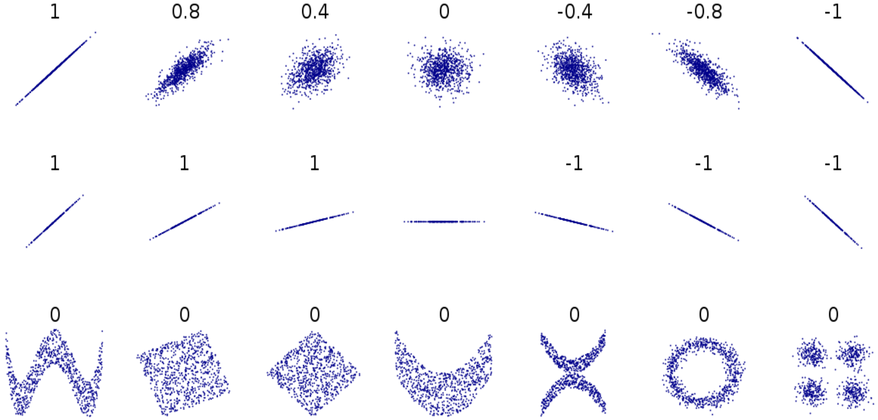
Variable DC (drought code)

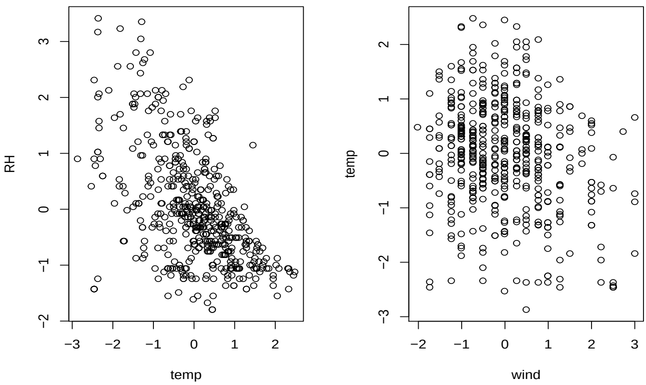


Variable DC (drought code)

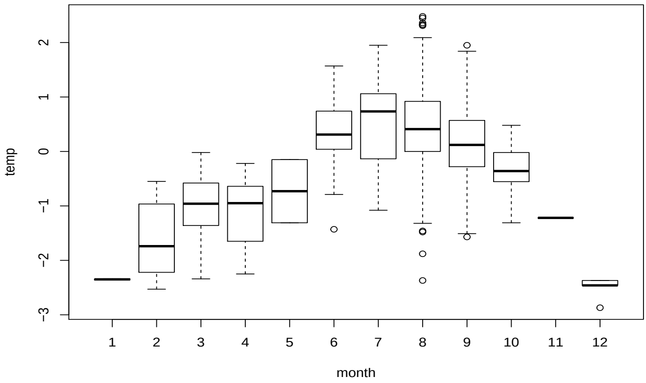


2 quantitative samples : correlation, scatter plots




RH
temp
temp
wind
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

One quantitative sample and one categorical sample :



References

- *Think Stats* Downey
  Analytical distributions : chap. 5, 6
  Empirical distributions and descriptive statistics in Python : chap. 2, 3, 4, 7
- *Mathematical statistics and data analysis* Rice
  Probability distributions : chap. 2, 3, 4
  Descriptive statistics : chap. 10