Es L1)

$X, Y \sim Exp(1)$, $\frac{X}{X+Y}$, $X+Y$

$$\begin{cases} U = \frac{X}{X+Y} \\ V = X+Y \end{cases} \rightarrow \begin{cases} X = UV \\ Y = V - UV \end{cases}$$

$$f_{U,V}(u,v) = f_{X,Y}(uv, v(1-u)) \, |det(J)| \overset{*}{=}$$

$$J = \begin{pmatrix} v & u \\ -v & 1-u \end{pmatrix} \rightarrow |det(J)| = v(1-u) + uv = v$$

$$\overset{*}{=} e^{-uv} e^{-(1-u)v} |v|$$

$$= |v| e^{-v} \cdot 1_{(0,1)}(u)$$

u can't miss from this function

it's a product between a gamma and an uniform distribution

$\Downarrow$

u, v are independent

$X_n \sim Bin(n, \vartheta)$ → Sum of Bernoulli (0,1) random variables

a) $X_n/n \xrightarrow{P} \vartheta$ → Average of the Bernoulli r.v.

→ linearity of expectation

$$E(X_n/n) = \frac{1}{n} \underline{E(X_n)} = \frac{1}{n} \underline{n\vartheta} = \vartheta$$

$$Var(X_n/n) = \frac{1}{n^2} Var(X_n) = \frac{1}{n^2} n\vartheta(1-\vartheta) = \frac{\vartheta(1-\vartheta)}{n}$$

↳ for $n \to \infty$, $Var(X_n/n) \to 0$

let $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n/n - \vartheta| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\vartheta(1-\vartheta)}{n\varepsilon^2} \to 0$$

↳ Proved that $\frac{X_n}{n} \to \vartheta$

b) $1 - \frac{X_n}{n} \xrightarrow{P} 1 - \vartheta$

$$E\left(1 - \frac{X_n}{n}\right) = 1 - E\left(\frac{X_n}{n}\right) = 1 - \vartheta$$

$$Var\left(1 - \frac{X_n}{n}\right) = -Var\left(\frac{X_n}{n}\right) = -\frac{\vartheta(1-\vartheta)}{n} \to 0$$

↳ Proved that converges to 0

c) $(X_n/n)(1 - X_n/n) \xrightarrow{P} \vartheta(1-\vartheta)$

↳ True due to algebraic laws of sequences of r.v.

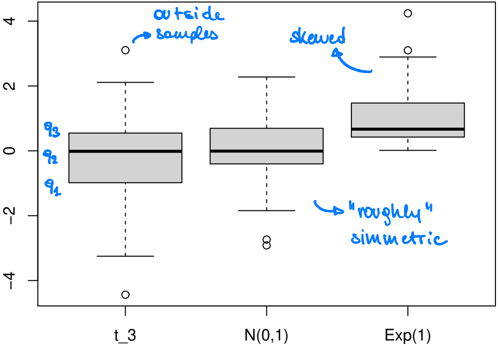
Boxplot (Box-and-whiskers plot)

It's a 5-summary statistics description of a (typically observed) sample.

It provides informations about the location, spread and the shape of the distribution of the sample.

In the vertical orientation:

- the middle line represents q_2, and vertical edges of the box represent q_1 and q_3, resp.
- the upper whisker is the largest $x_i \leq q_3 + 1.5 \cdot \text{iqr}$
- the lower whisker is the smallest $x_i \geq q_1 - 1.5 \cdot \text{iqr}$
- observations outside the whiskers are typically marked by a "*"



box is wider → higher variability

When we see whiskers that are almost equal or we see that the median is roughly in the middle

⇓

The distribution is (roughly) simmetric

Quantile-Quantile plot

The QQ plot is useful for checking if an observed sample is compatible with a population with continuous F.

It works by comparing a list of observed sample quantiles with the corresponding quantiles of a distribution F.

It consists in plotting the pairs $(x_{(i)}, F^{-1}(i/(n+1)))$ and looking for a linear pattern.

Observed quantile

Quantile function (True)

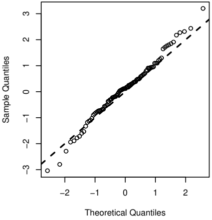
The QQ plot with F the normal distribution is the most widely used. In practice F involves unknown parameters which have to be estimated beforehand.

Good compatibility

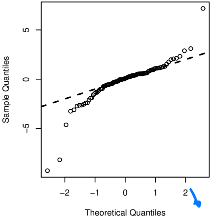
Tails of our funct. are too rapid

I've chosen a simmetric function while it should be skewed

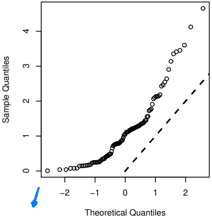
(a) observed vs N(0,1): ok

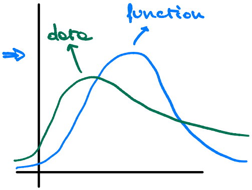


(b) observed vs N(0,1): tails!



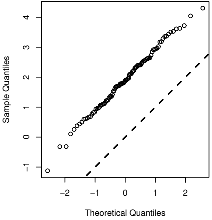
(c) observed vs N(0,1): symmetry!



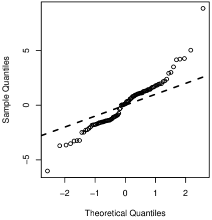


Must find new distribution

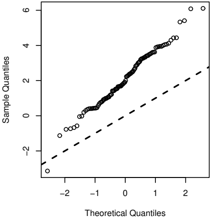
(d) observed vs N(0,1): location!



(e) observed vs N(0,1): scale!



(f) observed vs N(0,1): location and scale!



Un-matched sample space

My function has too low variability

Multivariate data

In realistic applications we may collect observation for several variables, thus a typical dataset looks like

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

where n is the number of observations and p is the number of variables.

The jth column represents the overall sample for the jth variable, and the ith row is the ith sample point for all variables.

For example, the columns could be pollutants s.t. $PM_{2.5}$, PM_{10}, CO_2, etc. and the rows may be values measured hourly.

Summaries for multivariate data

A common query is if the p variables are related to each other.

A first approach could be to plot pairs of variables and inspect the graph for possible associations.

For pairs of variables the sample covariance and the sample Pearson's correlation are widely used measures of association.
In particular, for a pair of variables x, y, the sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}),$$

and the sample Pearson correlation is

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}. \rightarrow \in [-1,1]$$

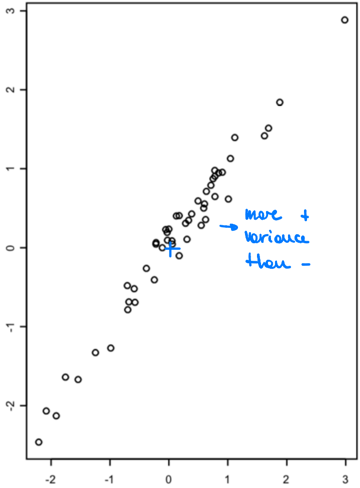
strong positive association

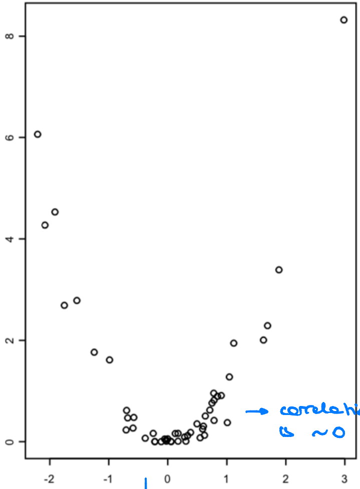
strong negative association

s_{xy} targets it population version σ_{XY}, whereas r_{xy} targets its population version ρ_{XY}.

Caution: lack of correlation $\not\Rightarrow$ lack of association!

→ + covariance





→ correlation is ~0

⇒ can't tell that there's no association

it's a more general term

are associated but not (linearly) correlated

Statistical models

Depends on a parameter that I don't know

Let $X_1, \ldots, X_n$ be random sample with $X \sim F_\theta$. If, in addition, X_i are also independent we call it iid random sample.

identically independently distributed

The joint pdf of the sample is $f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$.

By a statistical model we mean the set

$$\{f(x_1, \ldots, x_n; \theta) : \theta \in \Theta, x_i \in \mathcal{X}\},$$

where Θ is the parameter space, i.e. the set of all possible values for θ.

Typically, $\Theta \subseteq \mathbb{R}^d$ for some integer $d > 0$ and X_i could be a rv or a rve of any dimension.