Probability and Statistics for Data Science

S. Gadat - Toulouse School of Economics
Lecture 2 - Functions of Random Variables - Statistics

2023

Outline

Outline

Transfer theorem

You are expected to be able to solve the following problems.

Transfer theorem - 1D example

Exercise : For any random variable X of density f w.r.t. the Lebesgue measure, identify the distribution of $Y = e^X$.
When $X \sim \mathcal{N}(\mu, \sigma^2)$, Y is a log-normal random variable.

Transfer theorem - 1D example

Exercise : For any random variable X of density f w.r.t. the Lebesgue measure, identify the distribution of $Y = X^{-1}$.
When $X \sim \mathcal{N}(0,1)$, identify the density of Y. When X is a centered Cauchy random variable, show that Y is also a Cauchy random variable.

Transfer theorem - 1D example

Exercise : When $X \sim \mathcal{N}(0,1)$, identify the density of $Y = X^2$ (chi-square distribution).

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $(U,V) = \left(\frac{X+Y}{\sqrt{2}}, \frac{X-Y}{\sqrt{2}}\right)$.

Transfer theorem - 2D example

Exercise : Consider X and Y two random variables, identify the density of $U = XY$ and the density of $V = \frac{X}{Y}$. Simplify a bit the results when X and Y are independent.

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $U = \frac{X}{Y}$.

Transfer theorem - 2D example

Exercise : When X and Y are *independent* $\mathcal{N}(0,1)$ random variable, identify the density of $(R,\theta) = (\sqrt{X^2+Y^2}, \tan^{-1}(Y/X))$.

Outline

Goal

We aim to introduce some key tools that allow to make some easier computations. Among them :

- The generating function G_X, defined for any integer valued random variable X :

$$G_X : s \longmapsto \sum_{k=0}^{\infty} \mathbb{P}(X = k)s^k = \mathbb{E}[s^X]$$

- The moment generating function M_X :

$$M_X : s \longmapsto \mathbb{E}[e^{sX}]$$

  M_X also refers to the Laplace transform.

- The characteristic function φ_X :

$$\varphi_X : s \longmapsto \mathbb{E}[e^{\mathrm{i}sX}],$$

  which is also the Fourier transform of the density of f.

Generating function

In what follows, X will denote an integer valued random variable, distributed according to to a statistical model, parametrized by $\mathbb{P}_\theta, \theta \in \Theta$.

- The generating function G_X, defined for any integer valued random variable X :
$$G_X : s \longmapsto \sum_{k=0}^{\infty} \mathbb{P}(X = k) s^k$$
- G_X is formally an infinite series, and I will not annoy you about theoretical convergence aspects. We will only keep in mind that G_X is defined $\forall s \in [0, 1]$.
- We observe that :
$$G_X(s) = \mathbb{E}[s^X]$$
- We furthermore have interesting relationships :
$$G'_X(1) = \mathbb{E}[X] \qquad G^{(n)}(0) = n!\mathbb{P}(X =_n)$$

Generating function

Some computations

- $X \sim \mathcal{B}(p)$, Bernoulli distribution of parameter p. Compute G_X :

$$G_X(p) = 1 + p(s-1)$$

- $X \sim \mathcal{P}(\lambda)$, Poisson distribution of parameter λ. Compute G_X :

$$G_X(s) = e^{\lambda(s-1)}$$

- Each time, there is a one to one map between the parameter and G_X.

Generating function

General result

Theorem

- *The map $\mathcal{L}(X) \longmapsto G_X$ is injective, i.e. G_X completely caracterizes the distribution of X.*
- *If X, Y are two independent r.v., then*

$$G_{X+Y}(s) = G_X(s)G_Y(s).$$

- *Assume that $X_1, \ldots, X_n$ are i.i.d. and define $S_n = X_1 + \ldots + X_n$, then*

$$G_{S_n}(s) = G_X(s)^n.$$

Generating function

Application. From the previous results, prove that :

- If $X \sim \mathcal{P}(\lambda)$ and $Y \sim \mathcal{P}(\mu)$ are independent, then $X + Y \sim \mathcal{P}(\lambda + \mu)$.
- If $X \sim \mathcal{B}(n, p)$ and $Y \sim \mathcal{B}(m, p)$ are independent, then $X + Y \sim \mathcal{B}(n + m, p)$.
- Consider $X_1, \ldots, X_n, \ldots$ an infinite sequence of i.i.d. r.v. and N an independent integer valued r.v. We define $S_N = X_1 + \ldots + X_N$. Then

$$G_{S_N}(s) = G_N(G_X(s)).$$

Deduce that

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X].$$

Moment Generating function

Even close at the first sight, the MGF of a random variable X (also referred to as the Laplace transform), is slightly different :

Definition (MGF)

We define Λ_X as :

$$\forall u > 0 \qquad \Lambda_X(u) = \mathbb{E}[e^{uX}]$$

Several remarks :

- Λ_X is not defined for any $u > 0$!
- Compute the following MGF :
  - Bernoulli $\mathcal{B}(p)$
  - Poisson $\mathcal{P}(\lambda)$
  - Exponential $\mathcal{E}(\lambda)$

Moment Generating function

Important properties :

Theorem

- *For any pair of independent r.v.* (X, Y) *:*

$$\Lambda_{X+Y}(u) = \Lambda_X(u)\Lambda_Y(u).$$

- *There is a one to one association between* $\mathcal{L}(X)$ *and* Λ_X *: the MGF fully characterizes the distribution of* X*.*
- *When the moments of* X *exist, we have*

$$\forall k \geq 0 \qquad \mathbb{E}[X^k] = \Lambda^{(k)}(0).$$

The last property justifies the "Moment Generating function" name.

Characteristic function

Final important transform : the Fourier transform / characteristic function

Definition (Fourier transform)

We define φ_X as :

$$\forall \xi \in \mathbb{R} \qquad \varphi_X(\xi) = \mathbb{E}[e^{i\xi X}]$$

- Defined for any $\xi \in \mathbb{R}^d$
- Ultimate transform, powerful !
- Necessitates to handle complex functions :-(
- May be generalized to vectors :

$$\forall \xi \in \mathbb{R}^d \qquad \varphi_X(\xi) = \mathbb{E}[e^{i\langle \xi, X\rangle}]$$

Characteristic function

Important properties :

Theorem

- *For any pair of independent r.v.* (X, Y) *:*

$$\varphi_{X+Y}(\xi) = \varphi_X(\xi)\varphi_Y(\xi).$$

- *There is a one to one association between* $\mathcal{L}(X)$ *and* φ_X *: the characteristic function fully characterizes the distribution of* X*.*
- *For any* a, b *and* X *a r.v. :*

$$\varphi_{aX+b}(\xi) = e^{ib\xi}\varphi_X(a\xi)$$

Characteristic function

Some more or less easy computations

- Bernoulli, Binomial, $q = 1 - p$:

$$\varphi_X(\xi) = (q + pe^{i\xi})^n$$

- Poisson distribution

$$\varphi_X(\xi) = e^{\lambda(e^{i\xi}-1)}$$

- Geometric distribution

$$\varphi_X(\xi) = \frac{pe^{i\xi}}{1 - qe^{i\xi}}$$

Characteristic function

Some more or less easy computations

- Exponential

$$\varphi_X(\xi) = \frac{\lambda}{\lambda - i\xi}$$

- Laplace

$$\varphi_X(\xi) = \frac{1}{1+\xi^2}$$

- Gaussian

$$\varphi_X(\xi) = e^{-\xi^2/2}$$

- Cauchy

$$\varphi_X(\xi) = e^{-|\xi|}$$

Characteristic function

Fundamental result

Theorem (Levy theorem)

A sequence of r.v. (X_n) *verifies* $X_n \longrightarrow X$ *in distribution if and only if*

$$\forall \xi \in \mathbb{R} \qquad \varphi_{X_n}(\xi) \longrightarrow \varphi_X(\xi)$$

Basic tool to establish the central limit theorem.
We essentially use the characteristic function to identify distributions of random variable, in particular when manipulating Gaussian vectors.

Outline

Sampling experiment

A sampling experiment is the repetition of n **identical** and **independent** primary experiments.

If $(\Omega, \mathcal{A}, \mathcal{P})$ is the model adopted for one primary experiment, then the model for the sampling experiment of size n is denoted by $(\Omega, \mathcal{A}, \mathcal{P})^{\bigotimes n}$ and given by

- the population set is the cartesian product Ω^n
- the measurable events σ-algebra is **generated** by the set of cartesian products $B_1 \times \cdots \times B_n$ where B_j's are measurable events of the σ-algebra $\mathcal{A}$. Here : "**generated**" means that the events are obtained by complement and by countable unions of such cartesian products.
- if X is the r.v. of interest for the primary experiment, then let $X_1, \ldots, X_n$ be the n independent and identically distributed r.v. (resulting from the sampling experiment) with the same law as the underlying X. Therefore, the set of probability laws of the sample $(X_1, \ldots, X_n)$ defines the probability family of the model.

Density in the sampling experiment

If $P \equiv P_X$ is the probability law of X in the primary experiment, we denote by $P^{\bigotimes n}$ the joint probability law of $(X_1, \ldots, X_n)$.

We consider two cases depending on whether P is discrete or continuous (note : mixtures do exist) :

1. Discrete case : the law $P^{\bigotimes n}$ of $(X_1, \ldots, X_n)$ is described by its p.m.f.

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \Pi_{i=1}^{n} \mathbb{P}(X = x_i)$$

   where $\mathbb{P}(X = x) \equiv P_X(x)$ is the p.m.f. of the underlying distribution P;

2. Continuous case : the law $P^{\bigotimes n}$ is described by its p.d.f.

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \Pi_{i=1}^{n} f_X(x_i)$$

   where $f_X(x)$ is the p.d.f. of the underlying distribution P.

Sampling experiment for an exponential model

Theorem

If $(\Omega, \mathcal{A}, \mathcal{P})$ is an exponential model in the primary experiment, then the resulting model $(\Omega, \mathcal{A}, \mathcal{P})^{\bigotimes n}$ in the sampling experiment of size n is also an exponential model.

If $T_1, \ldots, T_r$ are the "sufficient statistics" of the primary model, then the following statistics

$$\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_r(X_i)$$

are "sufficient statistics" for the sampling model.

The empirical distribution

In the sampling experiment, if $(x_1, \ldots, x_n)$ is a realization of the random sample $(X_1, \ldots, X_n)$, we define a discrete law P_n called the **empirical law** associated to the sample in the following way :

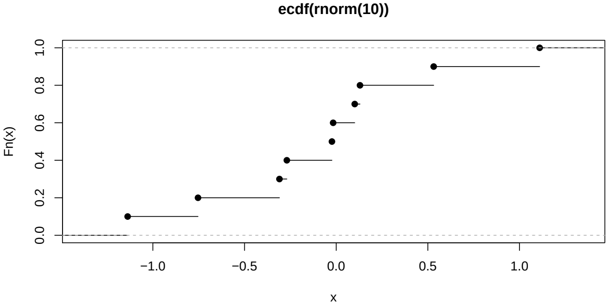
- P_n is the **discrete uniform** law on the sample values $\{x_1, \ldots, x_n\}$ which puts a mass equal to $\frac{1}{n}$ on each data point x_i.
- Its corresponding cumulative distribution function F_n, called the **empirical distribution** function, is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$

- The empirical c.d.f. $F_n(x)$ is an approximation of the population c.d.f. $F_X(x) = \mathbb{P}(X \leq x)$: we will prove that $F_n(x)$ converges to $F_X(x)$ when the sample size n increases to infinity.

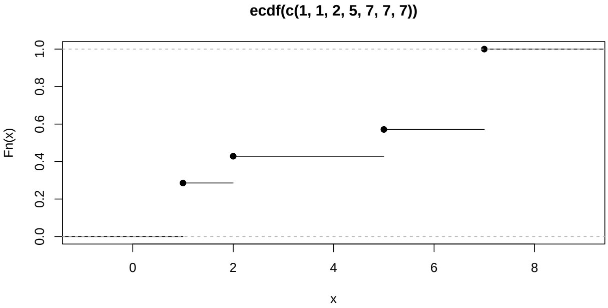
Graph of the empirical distribution function : case of no ties

```
F10 <- ecdf(rnorm(10))
plot(F10)
```



Graph of the empirical distribution function : case of ties

```
F2=ecdf(c(1,1,2,5,7,7,7))
plot(F2)
```



What is a statistic ?

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a model.

A **statistic** $T : x = \mathrm{obs} \in \Omega \mapsto T(x)$ is a **measurable map** from Ω to a measurable space $\mathcal{Y}$.

Example : Consider a sampling experiment of size n based on the second example (slide 9), where each X_i represents the number of claims for year i.

In this example, the quantity $T(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ is a statistic with $\mathcal{Y}$ equal to the set of real numbers $\mathbb{R}$ equipped with the Borel sigma-algebra.

Here : $T(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$ is a random variable, also called a statistic by misuse of language. To make short, one usually write rather $T = \frac{1}{n}\sum_{i=1}^n X_i$. In practical terms, a statistic is a measurable function $T(X_1, \ldots, X_n)$ of the observed random variables.

Population versus Sample characteristics

The characteristics of a probability law are :
its density, its distribution function, its mean, its variance, more generally its moments, its quantiles, etc.

In the sampling experiment,

- the **population characteristics** (or **theoretical characteristics**) are the characteristics of the underlying population or theoretical law P_X (unknown) ;
- the empirical characteristics (or sample characteristics) are the characteristics of the corresponding empirical law P_n.

We will also prove that the empirical characteristics converge to the population ones when the sample size n increases to infinity.

Population and Empirical mean

In the sampling experiment,

- the population mean (or theoretical mean) is the mean of P_X which is equal to $\mathbb{E}(X)$;
- the empirical mean is the mean of P_n which is equal to

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

One can define a random version, also called empirical mean, by

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

As such, $\bar{x}$ is a realization of the random variable $\bar{X}$.

Population and Empirical moments

The population (theoretical) moment of order k is :

1. $\mathbb{E}(X^k)$ (uncentered)
2. $\mathbb{E}[(X - \mathbb{E}(X))^k]$ (centered)

The empirical moments of order k is :

1. $\frac{1}{n}\sum_{i=1}^n x_i^k$ (uncentered)
2. $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^k$ (centered)

Empirical variance is :

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example : an empirical cdf

Population (or theoretical) law :
Consider the uniform discrete distribution on the set $\{1, 2, 3\}$.

A sample of size 5 yields : $1, 3, 2, 2, 1$.

- Population (or theoretical) probability mass function (denoted by $\mathbb{P}_X$) :

$$\mathbb{P}_X(1) = \mathbb{P}_X(2) = \mathbb{P}_X(3) = \frac{1}{3}$$

- Empirical probability mass function (denoted by $\mathbb{P}_5$)

$$\mathbb{P}_5(1) = \frac{2}{5}, \quad \mathbb{P}_5(2) = \frac{2}{5}, \quad \mathbb{P}_5(3) = \frac{1}{5}.$$

Same example : moments

Population (or theoretical) mean

$$\mathbb{E}(X) = \frac{1+2+3}{3} = 2$$

Empirical mean (or sample mean)

$$\bar{x} = \frac{1+3+2+2+1}{5} = \frac{9}{5}$$

Population quantiles : examples

Recall that the quantile of order $\alpha \in]0,1]$ of the population distribution P_X is

$$q_\alpha(X) = F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}.$$

Examples :

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, and if Φ is the cdf of $\mathcal{N}(0,1)$ then

$$q_\alpha = \mu + \sigma\,\Phi^{-1}(\alpha).$$

- If X follows the logistic distribution with pdf

$$f_X(x) = \frac{\exp(-\frac{x-a}{b})}{b(1+\exp(-\frac{x-a}{b}))^2},$$

then

$$q_\alpha = a + b\ln\left(\frac{\alpha}{1-\alpha}\right).$$

Empirical quantiles

Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics
(sample values $X_1, \ldots, X_n$ sorted in increasing order).

The previous rules applied to the empirical distribution function yield two cases in the scenario of no ties :

1. First case : $n\alpha$ is an integer

$$\hat{q}_\alpha = X_{(n\alpha)} = X_{([n\alpha])}$$

2. Second case : $n\alpha$ is not an integer

$$\hat{q}_\alpha = X_{([n\alpha]+1)}$$

where $[n\alpha]$ is the integer part of $n\alpha$.

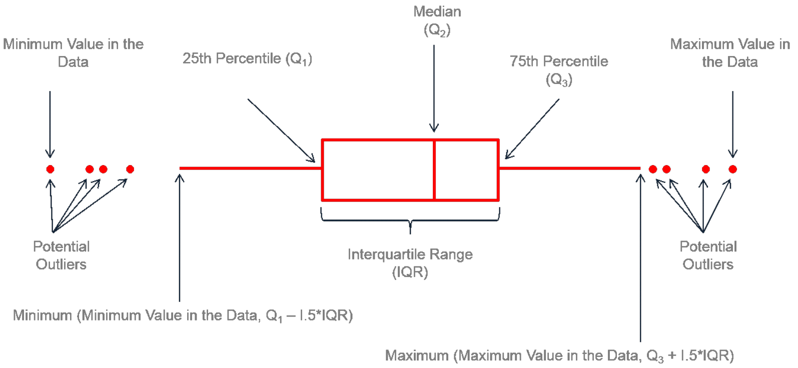
Alternatively in this second case :

$$\hat{q}_\alpha = \frac{1}{2}(X_{([n\alpha])} + X_{([n\alpha]+1)})$$

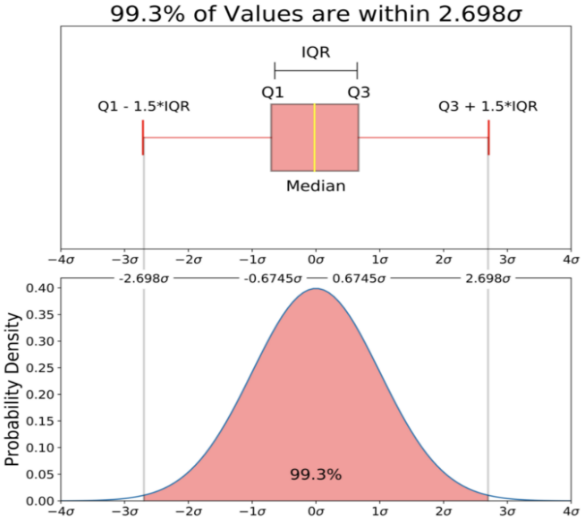
which coincides with classical empirical **median** when $\alpha = 1/2$.

Empirical quantiles applications : boxplots

Box Plot Anatomy :



Box Plot (cont'd)



Empirical quantiles applications : Probability plots

A **probability plot** is a **graphical tool** for comparing **two data sets** :

- either two sets of empirical observations,
- or one empirical set against a theoretical set,
- or (more rarely) two theoretical sets against each other.

It commonly means one of these three plots :

- P-P plot, "Probability-Probability" or "Percent-Percent" plot
  [plot of $F_X(z)$ against $F_Y(z)$] ;
- Q-Q plot, "Quantile-Quantile" plot
  [plot of $F_X^{-1}(\alpha)$ against $F_Y^{-1}(\alpha)$] ;
- (special case :) Normal probability plot,
  a Q-Q plot against the standard normal distribution.

Normal probability plot : preliminary lemmas

Lemma 1

- If the R.V. X has a cdf F_X and if U has a uniform distribution on $[0,1]$, then the random variable $F_X^{-1}(U)$ has the same distribution as X.
- If the R.V. X has a cdf F_X which is invertible, then the variable $F_X(X)$ has a uniform distribution on $[0,1]$.

This lemma is used by statistical softwares to generate samples from a given law starting with uniform samples.

Lemma 2 If $U_1,\dots,U_n$ is a sample from the uniform distribution on the interval $[0,1]$, then

$$\mathbb{E}(U_{(r)}) = \frac{r}{n+1}.$$

Normal probability plot : theory

Given a sample $(X_1, \ldots, X_n)$, the **normal probability plot** is the plot of the points

$$\left(X_{(r)}, \Phi^{-1}\left(\frac{r}{n+1}\right)\right) \quad \text{for} \quad r = 1, \ldots, n.$$

It is a Q-Q plot for F_n^{-1} against Φ^{-1}.

The principle is based on the following theorem :

Theorem. If $X_1, \ldots, X_n$ are i.i.d. with cdf F, then

$$\mathbb{E}\left[F\left(X_{(r)}\right)\right] = \frac{r}{n+1}.$$

Application : R function 'qqnorm'

QQ plot for location – scale family

Similarly, one can see whether the distribution of X belongs to a given **location–scale family** of distributions :

$$F_X(x) = F_0\left(\frac{x-\mu}{\sigma}\right),$$

where μ is the **location parameter** and σ is the **scale parameter**.

Same theorem $\rightarrowtail$ plot of $\left(X_{(r)}, F_0^{-1}\left(\frac{r}{n+1}\right)\right)$ is approximately aligned.

Examples : R function 'qqt' of package 'limma', 'qqPlot' from package 'qualityTools' for Beta, Cauchy, χ^2, Poisson, etc.

Normal probability plot : practice

Application : forget about the expectation in the theorem, roughly the points $\left(X_{(r)}, \Phi^{-1}\left(\frac{r}{n+1}\right)\right)$ should be aligned if the sample comes from a gaussian $\mathcal{N}(\mu, \sigma^2)$.

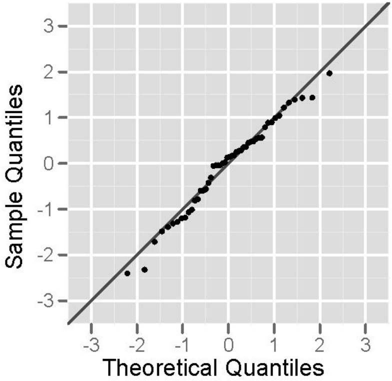
In practice, the quantile order $\frac{r}{n+1}$ is replaced by more sophisticated forms.

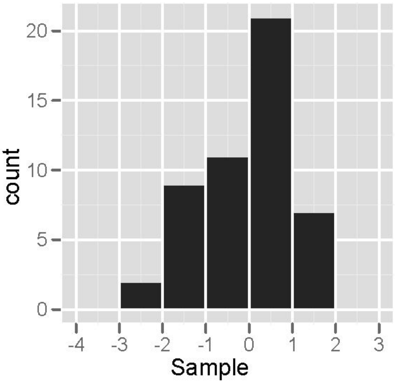
As a reference, a straight line can be fit to the points :
The further the points deviate from this line, the greater the indication of departure from normality.

Appreciation depends upon the sample size.

Normal probability plot : examples

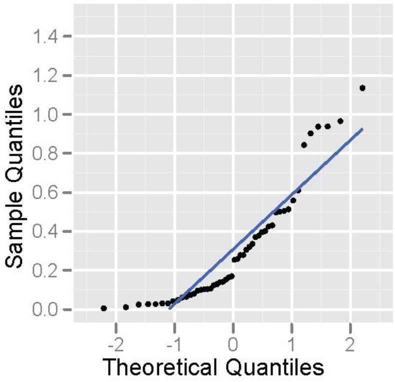
Sample of size 50 from a gaussian, from Wikipedia Commons

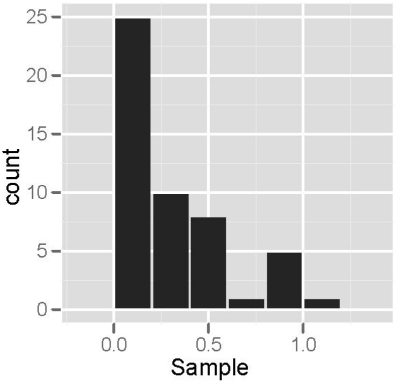




Normal probability plot : examples

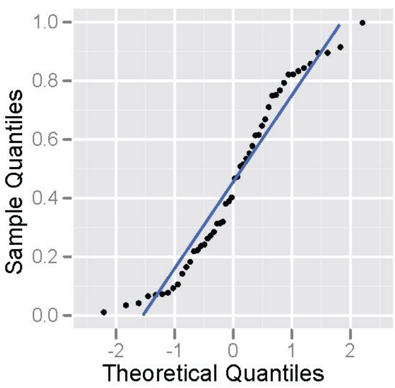
Sample of size 50 from a right-skewed distribution, from Wikipedia Commons

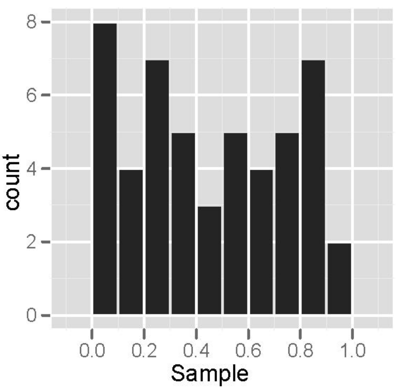




Normal probability plot : examples

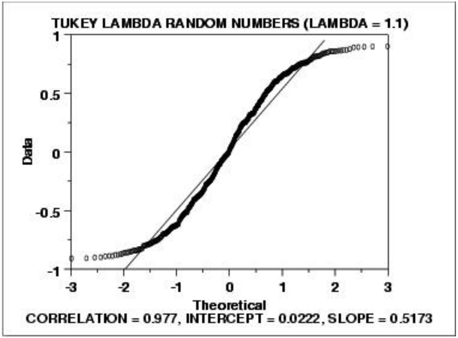
Sample of size 50 from a uniform, S-shape, from Wikipedia Commons





Normal probability plot : examples

Fat tail distribution



Quantiles : application to actuarial risk appraisal

In actuarial science, an aggregate loss is a random variable.
The **Value at Risk** (VaR) is a quantile of the distribution of aggregated losses (over a given time period) at a high probability level p ;
It is used in the determination of capital necessary to withstand such adverse outcomes (severe losses) :

$$Var_p(X) = F_X^{-1}(p) \quad \text{for } p \text{ close to 1 (high quantile).}$$

The **Tail-Value at risk** (or expected shortfall) is another more informative measure. Given a probability level p, the $TVar_p(X)$ is equal to the expected loss given that the loss exceeds the pth quantile of X (*i.e.* $Var_p(X)$) :

$$TVar_p(X) = \mathbb{E}(X \mid X > F_X^{-1}(p))$$

It can be shown that it is an average of all VaR values above the security level p and thus contains more information about the distribution of X in the tails than the VaR.

Quantiles : application to financial risk appraisal

Example : if $Var_p(X) = 100{,}000$ euros for $p = 0.99$ and the time period is one year, it means that there is a probability of

$$1 - p = 0.01$$

that the company will experience a loss of more than 100,000 euros over the next year.

If moreover $TVar_p(X) = 150{,}000$ euros for $p = 0.99$ and the time period is one year, it means that the expected loss will be 150,000 euros knowing that the company experiences a loss exceeding 100,000 euros next year.