Brief reminder on statistics

by Eric Marsden <eric.marsden@risk-engineering.org>

1 Summary statistics

If you have a sample of *n* values x_i, the *mean* (sometimes called the average), μ, is the sum of the values divided by the number of values; in other words moyenne

$$\mu = \mathbb{E}(x) \overset{\text{def}}{=} \frac{1}{n} \sum_i x_i$$

The mean represents the central tendency of a set of values (its expected value); it lets you make informed decisions on how you expect *x* to behave on average over the long run.

Properties of expected value as a mathematical operator

If *c* is a constant and *X* and *Y* are random variables, then

- $\mathbb{E}(c) = c$
- $\mathbb{E}(cX) = c\mathbb{E}(X)$
- $\mathbb{E}(c + X) = c + \mathbb{E}(X)$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

The *variance* measures its spread, or dispersion. The variance of a set of values is

$$\sigma^2 = Var(x) \overset{\text{def}}{=} \frac{1}{n} \sum_i (x_i - \mu)^2$$

The term $x_i - \mu$ is called the "deviation from the mean", so variance is the mean squared deviation, which is why it is denoted σ^2. The square root of variance, , is called the *standard deviation*. écart type

Properties of variance as a mathematical operator

If *c* is a constant and *X* and *Y* are random variables, then

- $Var(c) = 0$
- $Var(c + X) = Var(X)$
- $Var(cX) = c^2\, Var(X)$
- $Var(X + Y) = Var(X) + Var(Y)$, **assuming that X and Y are independent**
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$ if X and Y are dependent

2 Distributions

Summary statistics are concise, but you can get a more complete view of the data by looking at its *distribution*, for example with a *histogram* (*cf.* figure 1).

A **probability mass function** maps the possible values of *x* against their respective probabilities of occurrence, $p(x)$. Note that

- $p(x)$ $[0, 1]$
- the area under a probability function is always 1

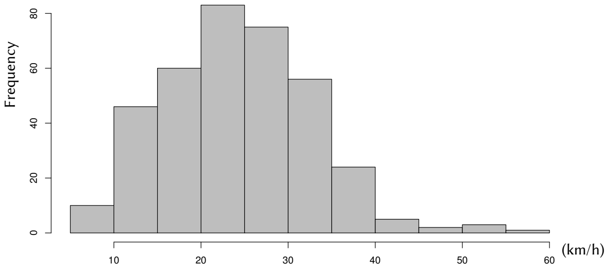


Figure 1 – *Example histogram of maximum daily wind speed in Toulouse, for 2010*

A *percentile* is the value of a variable below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found. The 25th percentile is also known as the first quartile (Q_1), the 50th percentile as the *median* or second quartile (Q_2), and the 75th percentile as the third quartile (Q_3). The *interquartile range* (difference between 25th and 75th percentiles) is a way of measuring the spread of values in a set.

médiane

The *cumulative distribution function* (*cf.* figure 2) maps values to their percentile rank in a distribution. To evaluate $CDF(x)$ for a particular value of x, we compute the fraction of the values in the sample that are less than or equal to x.

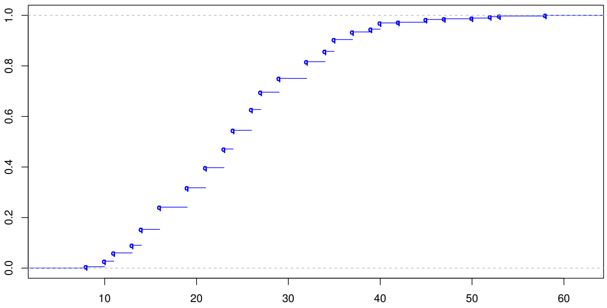


Figure 2 – *Cumulative distribution function for the maximum daily wind speed in Toulouse*

The *reliability function* of a distribution, used in reliability engineering, is simply one minus the CDF.

When doing risk analysis, we are often more interested in *extreme values* than in the sample mean (*highest level of a river, strongest wind, weakest metal pipe, biggest load on a bridge, etc.*). This requires careful modeling of the *tails* of the distribution (the 1% and 99% percentiles, for example).

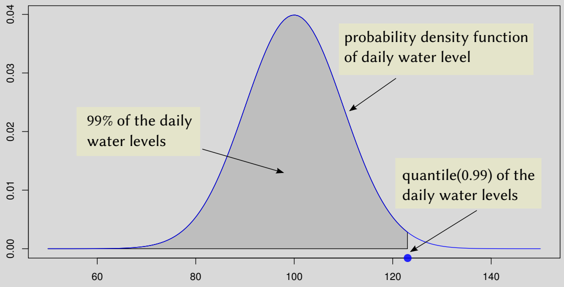
The *quantile function* is the inverse of the cumulative distribution function: for a given probability, it tells you the value which the random variable will be lower than, with that probability.

A one-hundred-year flood

Example

A one-hundred-year flood is the level of flood water expected to be equaled or exceeded every 100 years on average. The 100-year flood is more accurately referred to as the 1% annual exceedance probability flood, since it is a flood that has a 1% chance of being equaled or exceeded in any single year. It is simply `quantile(0.99)` (the 99% centile) of the water level.

crue centenale



2.1 The uniform distribution

The *discrete uniform distribution* (*cf.* figure 3) concerns a finite number of equally spaced values which are equally likely to be observed: each of the n values has a probability of $1/n$. A simple example of the discrete uniform distribution is throwing a fair dice. The possible values are 1, 2, 3, 4, 5, 6; and each time the die is thrown, the probability of a given score is 1/6[1].

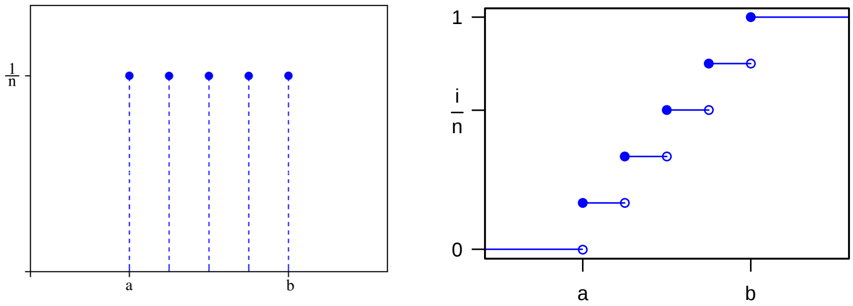


Figure 3 – *The probability mass function (left) and cumulative distribution function (right) of the discrete uniform distribution*

The *continuous uniform distribution* (*cf.* figure 4) is a family of probability distributions such that for each member of the family, all intervals of the same length on the distribution's support are equally probable. The support is defined by the two parameters, a and b, which are its minimum and maximum values. The distribution is often abbreviated $U(a, b)$.

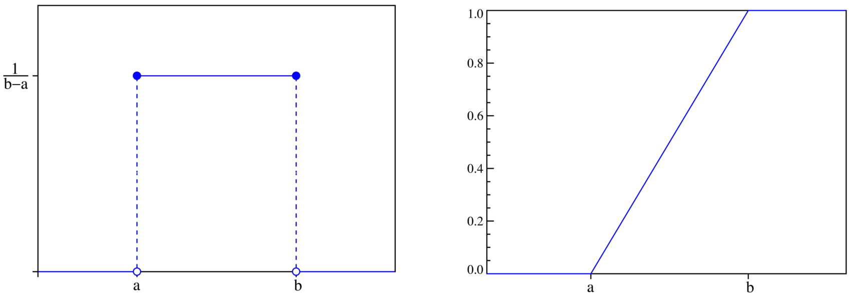


Figure 4 – *The probability mass function (left) and cumulative distribution function (right) of the continuous uniform distribution*

2.2 The normal distribution

The *normal* (or Gaussian) distribution (*cf.* figure 5) is a continuous probability distribution that is often used as a first approximation to describe real-valued random variables that tend to cluster around a single mean value[2]. The "bell" shape of the normal distribution makes it a convenient choice for modelling a large variety of random variables encountered in practice (*measurement errors, distribution of heights or weights in a population, velocities of the molecules in the ideal gas*). The normal distribution arises as the outcome of the **central limit theorem**, which states that under mild conditions the sum of a large number of random variables is distributed approximately normally.

[1] Note that if *two* dice are thrown and their values added, the uniform distribution no longer fits, since the values from 2 to 12 do not have equal probabilities.

[2] See `wikipedia.org/wiki/Normal_distribution` for more information.

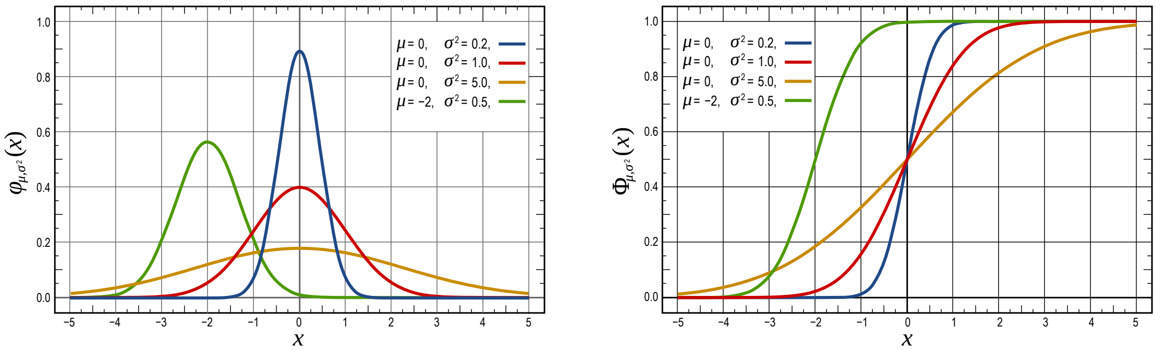


Figure 5 – *The probability mass function (left) and cumulative distribution function (right) of the normal distribution*

probability density function	$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
cumulative distribution function	$\frac{1}{2}\left[1 + erf(\frac{x-\mu}{\sqrt{2\sigma^2}})\right]$
mean (= median)	μ
variance	σ^2

Table 1 – Characteristics of the normal distribution

2.3 The exponential distribution

The *exponential distribution* (*cf.* figure 6) is a family of continuous probability distributions, commonly used in reliability engineering. The exponential distribution is used to model situations where events occur continuously and independently at a constant average rate (for example, failures of an electronic component).

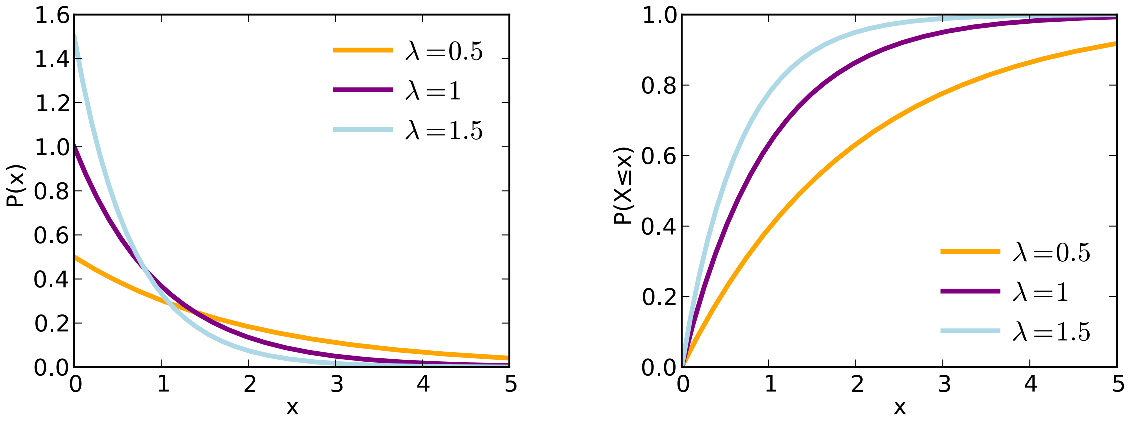


Figure 6 – *The probability mass function (left) and cumulative distribution function (right) of the exponential distribution*

Here is called the *rate parameter*, and represents a constant failure rate[3], in failures per unit of measurement (*failures per hour, per cycle, etc.*). is the inverse of the MTTF (*mean time to failure*).

[3] Note that this means that components are assumed to have no wear-out or infancy problems!

probability density function	$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
cumulative distribution function	$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$
mean (= MTTF)	$\mathbb{E}[X] = \frac{1}{\lambda}$
variance	$\mathrm{Var}[X] = \frac{1}{\lambda^2}$
median	$\mathrm{m}[X] = \frac{ln\, 2}{\lambda} < \mathrm{E}[X]$

Table 2 – Characteristics of the exponential distribution

The exponential distribution is *memoryless*: if a random variable T is exponentially distributed, then

$$Pr(T > s + t \mid T > s) = Pr(T > t) \quad \forall s, t \geq 0$$

3 Analyzing data

A *Q-Q plot* (or quantile-quantile plot) is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles are chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q-Q plot will approximately lie on a line, but not necessarily on the line $y = x$.

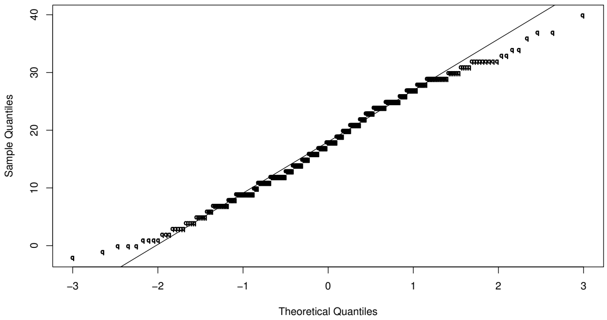


Figure 7 – *Normal Q-Q plot for daily maximum temperature in Toulouse (2010)*

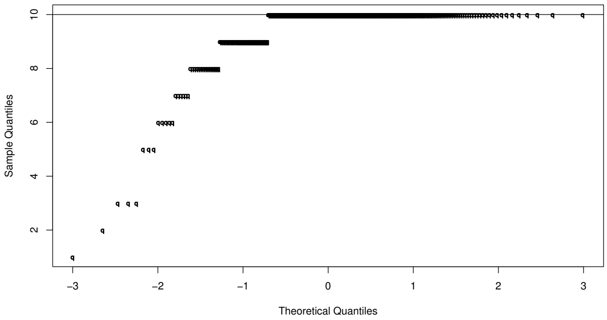
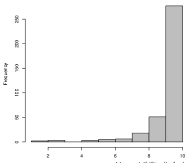


Figure 8 – *Normal Q-Q plot for mean daily visibility in Toulouse (2010, vertical axis measured in kilometers)*

Figures 7 and 8 show normal quantile-quantile plots respectively for daily maximum temperature and mean daily visibility in Toulouse in 2010. The plots show that maximum temperature



are approximately normally distributed, but that visibility does not follow a normal distribution (the sample is strongly skewed towards high values of visibility, as the histogram on the right shows).