04.Modelling Workload

Three types of random variables can be used:

- Discrete
- Continuous
- Mixed – continuous distributions that also have discrete outcomes.
  Discrete random variables associate a set of probabilities to a discrete number of possible outcomes. Usually the value goes from 0 to 1 and the more is near 1 the more the event has the probability of happening.
  If the outcomes are numbers, the Cumulative Distribution Function can be defined: the probability F(x) of getting an outcome less or equal to x.
  In Probability Distributions are used in performance evaluation to model whenever we have choice. They are used to guide random choices.
  Continuous random variables generate outcomes on a continuous space: $\Omega = [0, \infty)$.
  The probability of having an exact value is always zero: p(X = x) = 0.
  However, we can compute the probability that an outcome is included in a range: $p(x_A \leq X \leq x_B) \geq 0$.
  They are characterized by the probability density function f(x) and the cumulative distribution function F(x):

  $$p(x_A \leq X \leq x_B) = \int_{x_A}^{x_B} f(x)dx = F(x_B) - F(x_A)$$

  The PDF should be such that its integral over the distribution support (generally, from 0 to ∞), is equal to one.

  $$F(x) = \int_{-\infty}^{x} f(y)dy, \int_{-\infty}^{\infty} f(y)dy = 1, f(x) = \frac{dF(x)}{dx}$$

  In performance evaluation they are usually adopted to characterize the inter-arrival times and the service times of the components of a model. For this reason they are generally defined only for positive values of the random variable.

  $$F(x) = \int_{0}^{x} f(y)dy, \int_{0}^{x} f(y)dy = 1$$

  For a probability distribution, the Expected value, according to a function g(x), is defined as:

  $$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

  When g(x)=x^n, the corresponding expected value is called the n^{th} moment of the distribution.

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx$$

> Let us call μ = E[X] the first moment of the distribution. The n^{th} central moment of the distribution is defined as: $E[(X-\mu)^n]$
> Let us also call $\sigma = \sqrt{E[(X-\mu)^2]}$. The n^{th} standardized moment of the distribution is defined as:
>
> $$E[(\frac{X-\mu}{\sigma})^2]$$
>
> Central moments are meaningful for n > 1, and are insensitive to the position of the distribution.
> Standardized moments are meaningful for n > 2, and are insensitive to both the scale and the position of the distribution.

The square root of the variance is known as the standard deviation: its feature is that it uses the same units as the mean. This means that we can compare the standard deviation to the mean and since they can be compared we can define the coefficient of variation as their ratio:

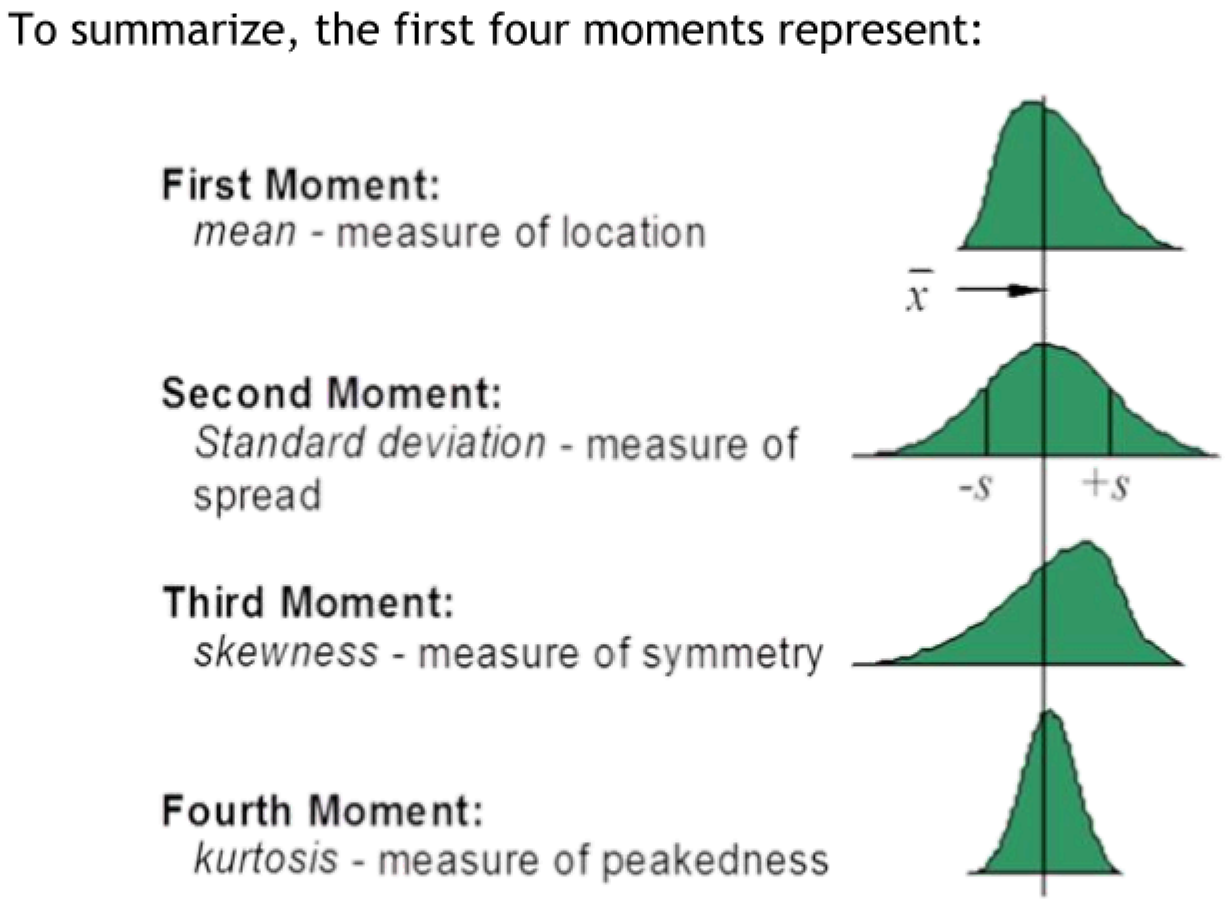
$$\sigma^2 = Var[X], \sigma = \sqrt{Var[X]}, c_v = \frac{\sigma}{\mu}$$

The third standardized moment is called the skewness: Skewness represents whether the distribution is symmetric, or if it has more probability mass to the left or to the right of its mean:

$$\gamma = E[(\frac{X-\mu}{\sigma})^3] = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

The fourth standardized moment (minus 3) determines a characteristic of the distribution called the excess Kurtosis. The -3 is introduced to have the Kurtosis of the Standard Normal Distribution equal to 0.

$$mod$$



To summarize, the first four moments represent:

First Moment:
mean - measure of location

Second Moment:
Standard deviation - measure of spread

Third Moment:
skewness - measure of symmetry

Fourth Moment:
kurtosis - measure of peakedness

Moments are important in performance evaluation for several reasons:

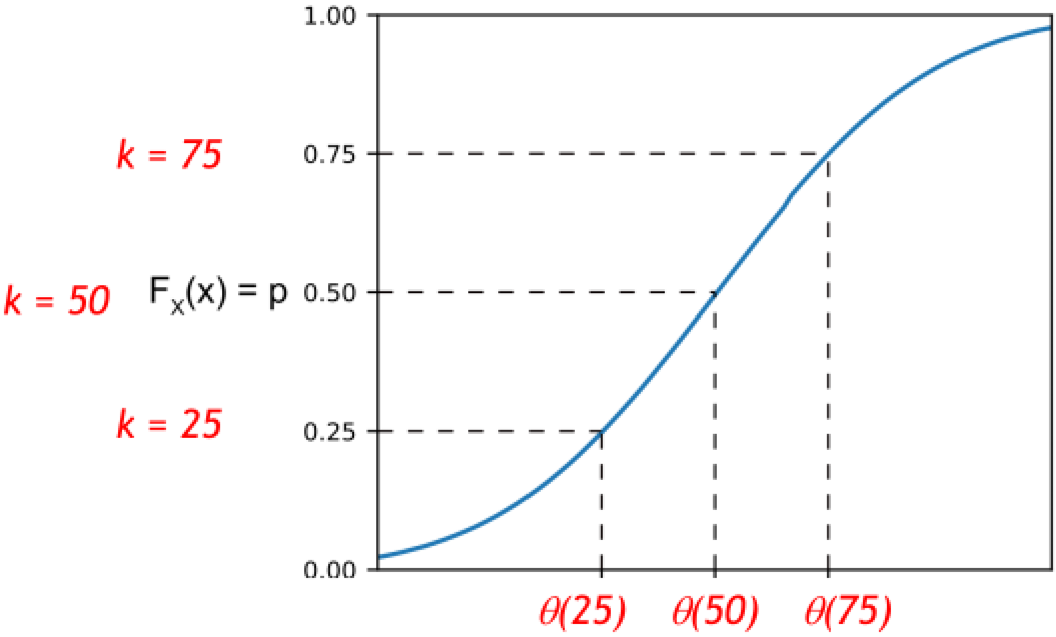
- There are many results that are sensitive only to the first or one or two moments of the distributions.
- Beside being computed analytically on distributions, they can be easily derived from measures, log files and data sets.
- Moments of the collected data can guide the modeler to choose and parametrize the most appropriate distributions.
  For distributions on discrete support, moments, mean, variance and other properties can be computed as finite sums.

$$\mu = E[X] = \sum_i x_i p(x_i),\, E[X^k] = \sum_i x_i^k p(x_i)$$

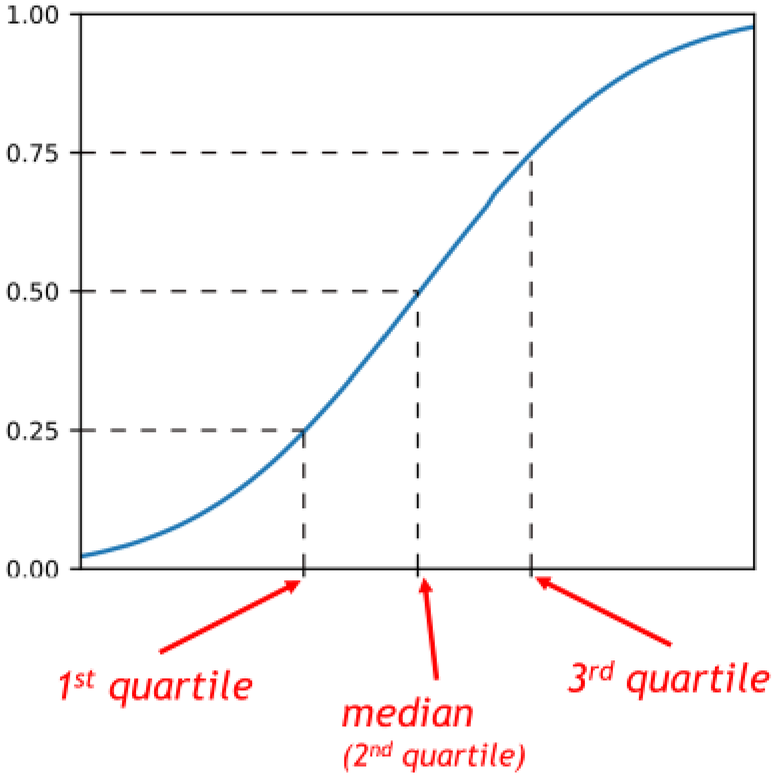
$$Var[X] = E[X^2] - \mu^2 = \sum_i x_i^2 p(x_i) - \mu^2$$

Percentile

The k^{th} percentile of a distribution is the value of the random variable, for which the CDF equals k/100: $\theta(K) = F^{-1}(k/100)$

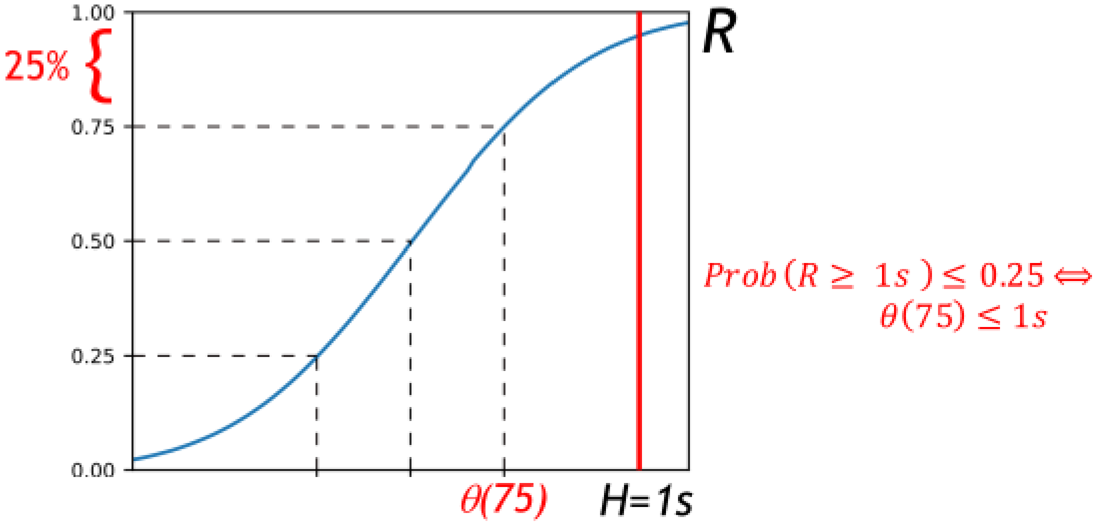


This is useful because it tells the percentile of values that are less than a threshold. The 25^{th}, 50^{th} and 75^{th} percentiles have special names: they are respectively called the 1^{st} quartile, the median (or 2^{nd} quartile), and the 3^{rd} quartile.



In performance evaluation percentiles are used to assess properties: for example, they can be used to ensure that the response time R of a station is greater than a given threshold H only for a limited percentage k of jobs.

$$Prob(R \geq H) \leq \frac{k}{100} \iff \theta(100-k) \leq H$$



Relations between Traces and Moments

If we have a set of sample we have a trace of random value and if we can believe that all these trace belong to a probability distribution we want to find this probability distribution. Estimating the PDF of a distribution from a set of samples is not an easy task. Approximating the CDF is instead much simpler. We can imagine that the entries in a file are all time instant in the same probability distribution. So we can sort all the elements to smaller to bigger. So we can check if they have the same probability. From this we can compute the CDF putting the sorted value on the x-axis and increasing the step of one overhand on the y-axis. The more elements we have the more the CDFs will be. The CDF can be defined as $F_X(x) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \leq x)$
Assuming that each sample is equally probable, moments can be approximated with discrete sums.

$$\mu = E[X]\frac{1}{N}\sum_{i=1}^{N} x_i, \, E[X^k] = \frac{1}{N}\sum_{i=1}^{N} x_i^k$$

Correlation

We can calculate it using the crosscovariance