Fundaments of Machine learning for and with engineering applications

Enrico Riccardi[1]

Department of Mathematics and Physics, University of Stavanger (UiS).[1]

Sep 10, 2025

University of Stavanger

Sampling

- What are the effective sampling strategies? (Wind turbine example)
- Solar Panels to determine the efficiency of the source (Usage patterns, energy production forecast)
- Drilling (penetration rate)
- Corrosion extension
- Concrete Rigidity
- Experimental design!

Wind turbine example

Turbine	Height	X	Y	Wind Speed	Air Density	Temperature	Power Output	Rotor Diameter	Hub Height	Air Pressure	Turbulence Intensity
WT-1	80	752.1	3945	7.5	1.225	15	1500	82	80	1013	0.1
WT-1	80	752.2	3945	8	1.223	15	1600	82	80	1012	0.12
WT-1	80	752.3	3945	7.8	1.224	16	1550	82	80	1013	0.11
WT-2	90	753.5	3946	6.5	1.226	14	1400	85	90	1012	0.15
WT-2	90	753.6	3946	7	1.225	14	1500	85	90	1011	0.13
WT-2	90	753.7	3946	7.2	1.227	14	1520	85	90	1012	0.14

Sampling approaches

Experimental design

Grid, parallel, series.

Sampling without replacement

SPR (single point representation).

Sampling with replacement

The number of the members of the population does not change.

Univariate statistics

- Easy to displaying data:
  - histogram
  - frequency plots
  - cumulative
- Measures of Location
  - Mean, median, mode
  - Quartiles, Percentiles, Quantiles
- Measure of Dispersion (Spread)
  - Standard deviation (sd)
  - Variance (Var) or coefficient of variation
- Measures of shape
  - Skewness, modality

Frequency plots and Histograms

Given a set of data

1. Look for min and max values
2. Divide the range of values into a number of sensible class intervals (bins)
3. Count
4. Make a frequency table (or percentage)
5. Plot (see jupyter notebook)

Does this histogram represent uncertainty?

No. It shows variability, but it can be used to quantify uncertainty.

Class widths

- Class widths (bin sizes) are usually CONSTANT
  - the height of each bar is proportional to the number of values in it
- If class width are VARIABLE
  - the AREA of each bar is proportional to the number of values in it
- For small samples, the shape of the histogram can be very sensitive to the number and definition of the class intervals

Cumulative Histogram

- Cumulative frequency
- Each data point can be plotted individually
- It helps to read quantiles and compare distributions

Measure of Location: Central Tendency, MEAN

$$m_x = < x > = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Each point weighted equally by $\frac{1}{n}$ (assumption)

- Every element is the data set contributes to the values of the mean
- An average provides a common measure for comparing one set of data to another
- The mean is influenced by the extreme values in the data set
- The mean may not be an actual element if the dataset
- The sum of all deviation from the mean is zero, and the sum of squared deviation is minimized when those deviations are measured from the mean

Means

- Arithmetic
  - Mean of raw data

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$

- Geometric
  - n^{th} root of product

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

- Geometric

$$* \text{ Mean of logarithms } exp\left(\frac{1}{n} \sum_{i=1}^{n} ln(x_i)\right)$$

- Harmonic
  - Mean of inverses

$$\left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Median

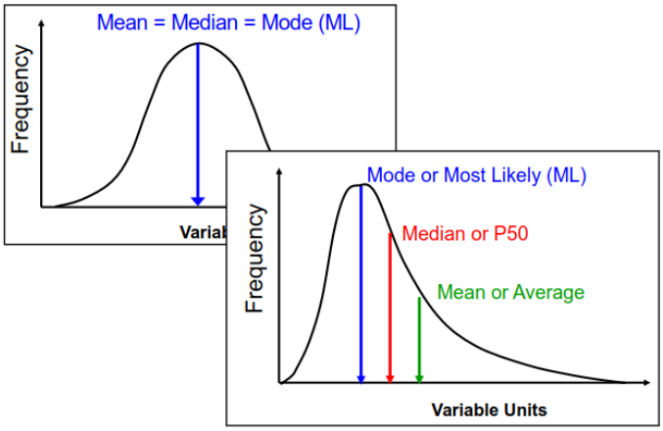
```
if n is odd:
    median = x[(n+1)/2]
else:
    median = x[n/2] + x[(n/2)+1]
```

- On a cumulative density plot, the value of the x-axis that corresponds to 50 % of the y-axis
- Not influenced by extreme values
- May not be contained in the dataset (if n is even)
- For a perfectly symmetrical dataset, means = median

Mode

- The mode is the most frequently occurring data element or the most likely or most probable value (for a pmf)
- A data set may have more than one mode and it thus called multimodal
- A mode is always a data element in the set
- For a perfectly symmetrical dataset, means = median = mode

Distribution Descriptors



Quantiles

Quartiles

The data split into quarters.

Deciles

The data are split into tenths. The fifth decile is also the median.

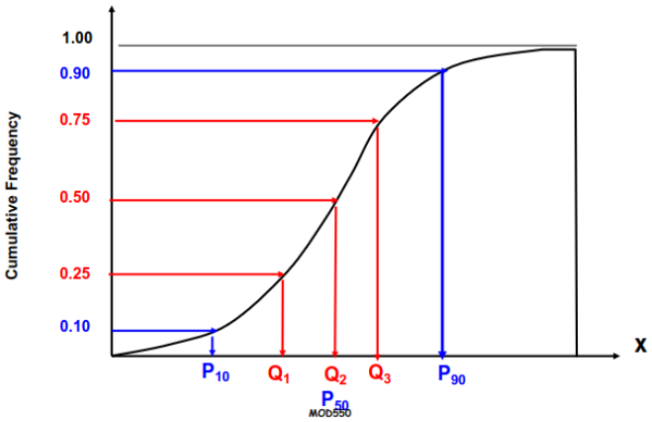
Percentiles

The data are split into hundredths. P10, P25, P50, P75 and P90 are the most commonly used.

Quantiles

A generalization of splitting data into any fraction

Distribution Descriptors



Dispersion (Spread)

Range

$$R = \text{maximum - minimum}$$

Inter-quantile Range

$$IQR = Q3 - Q1$$

Mean Deviation from the Mean

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

Mean Absolute Deviation

$$MAD = \sum_{i=1}^{n} |x_i - \bar{x}|/n$$

Variance

The variance is the average of squared differences between the sample data points and their mean

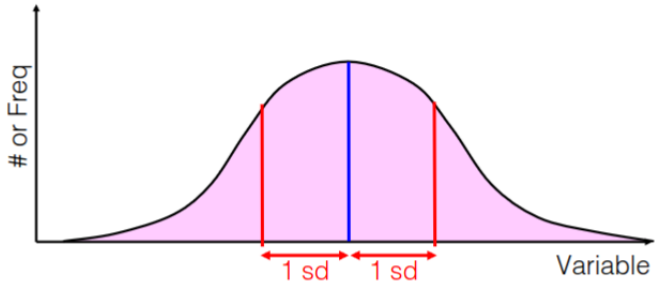
Variance

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

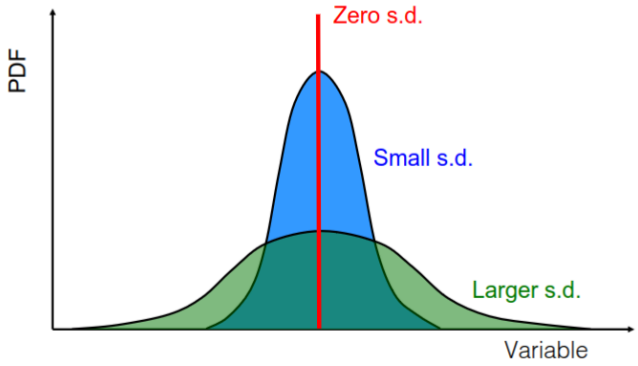
Standard Deviation (SD)

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

Standard Deviation



Standard Deviation



Measures of dispersion

Standard Deviation (SD)

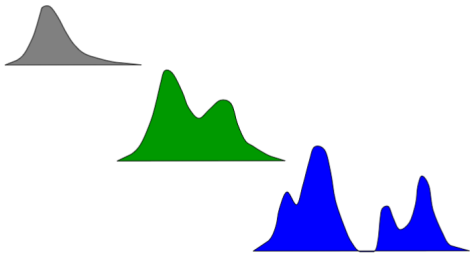
$$SE_x = \frac{s_x}{\sqrt{n}}$$

Coefficient of Variability

$$CV = \frac{s_x}{\bar{x}}$$

Modality

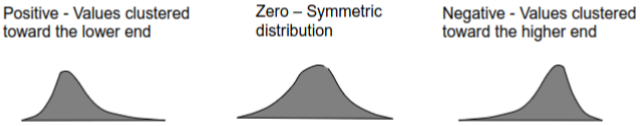
- Unimodal
- Bimodal
- Polymodal



Skewness

It measures the symmetry in a distribution

$$Sk = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{s^3}$$



A bit out of fashion with ML

Distribution Models

Distribution

means of expressing uncertainty or variability

Models

- Uniform: useful when only upper and lower bounds are known
- Triangular: useful when estimates of min, max, mode [P10, P50, P90] are available
- Normal: symmetric model of random errors or unbiased uncertainties with mean of standard deviation specified
  - Very common for observed data
  - Additive processes tend to be normal as a result of the Central Limit Theorem
- Log normal comes from multiplicative uncertainties with mean and standard deviation specified
- Many more!

Uniform Distribution

- The uniform distribution is useful as a rough model for representing low states of knowledge when only the upper and lower bounds are known.
- All possible values within the specified maximum and minimum values are equally likely (b=max, a=min):
- It can express maximum uncertainty

PDF: $f(x) =$

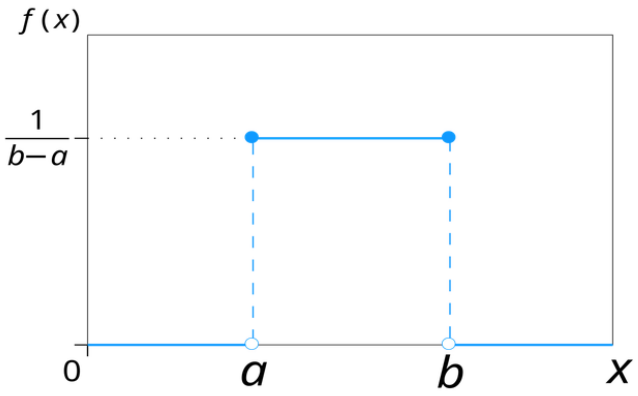
$$\frac{1}{b-a}, a \leq x \leq b$$

CDF: $F(x) =$

$$\frac{x-a}{b-a}$$

Notation: $X \sim U(a, b)$

Uniform Distribution



Triangular distribution

- The triangular distribution can be used for modeling situations, where non extremal (central) values are more likely than the upper and lower bounds.
- Take min, mode and max as inputs. Typically on the basis of subjective judgement:

PDF: $f(x) =$

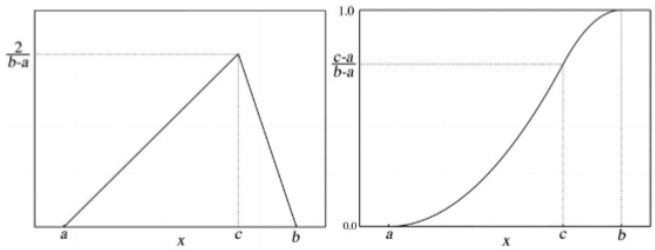
$\frac{2(x-a)}{(b-a)(c-a)}$; if $a \leq x \leq c$
$\frac{2(b-x)}{(b-a)(c-a)}$; if $c \leq x \leq b$

CDF: $F(x) =$

$\frac{(x-a)^2}{(b-a)(c-a)}$; if $a \leq x \leq c$
$1 - \frac{(b-x)^2}{(b-a)(c-a)}$; if $c \leq x \leq b$

Triangular Distribution

Notation: $X \sim T(a, b, c)$



It can be symmetric or asymmetric

Normal Distribution

- The normal distribution ('bell curve' or Gaussian) for modeling unbiased uncertainties and random errors of the additive kind of symmetrical distributions of many material processes and phenomena.
- A commonly cited rational for assuming normal distribution is the central limit theorem, which states that the sum of independent observations asymptotically approaches a normal distribution regardless of the shape of the underlying distributions

PDF:

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$; $-\infty \leq x \leq \infty$

CDF: $F(x) =$

has no closed form solution but is often presented using the complementary error function solution

Normal Distribution

Notation: $X \sim G(\mu, \sigma)$

It is a Symmetric distribution around the mean

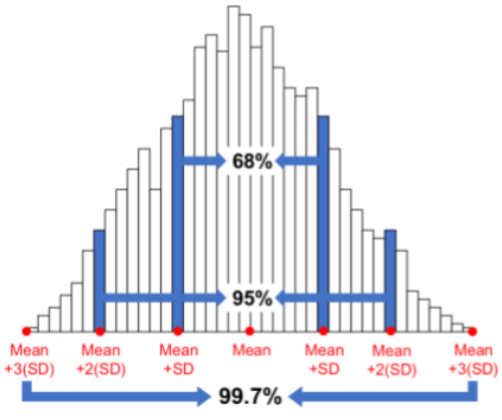
μ is the mean, σ is the standard deviation

$\mu \pm \sigma$: 68.3%*probability*

$\mu \pm 2\sigma$: 95.4%*probability*

$\mu \pm 3\sigma$: 99.7%*probability*

Normal Distribution



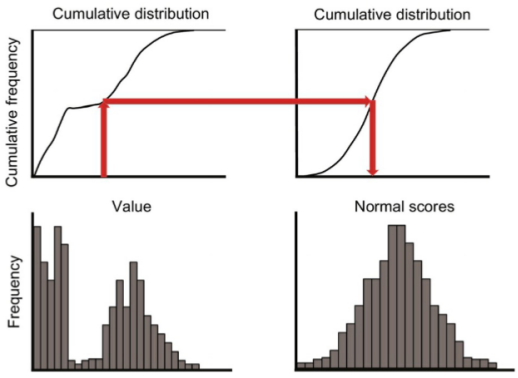
Data transformations

- Often, it is useful to transform a sample distribution into the space of an equivalent normal distribution, where many statistical operations can be easily performed and visualized
- The approach involves a rank-preserving one-to-one transformation.
- Transforming the data so that their distribution matches a prescribed (target) distribution.
- Sometimes we must transform the data...

Normal Score Transformation

1. From data to cumulative distribution.
2. From cumulative distribution and map back.

O Quantile-to-quantile normal score transformation



Match Quantiles

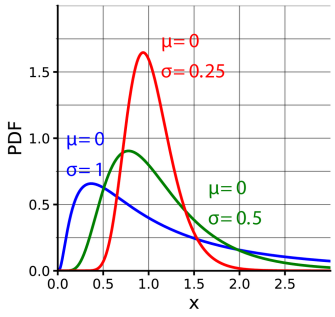
Log - Normal distribution

For a log-normal distribution, we define the standard normal variate as

$\alpha = means\ of\ ln(x)$

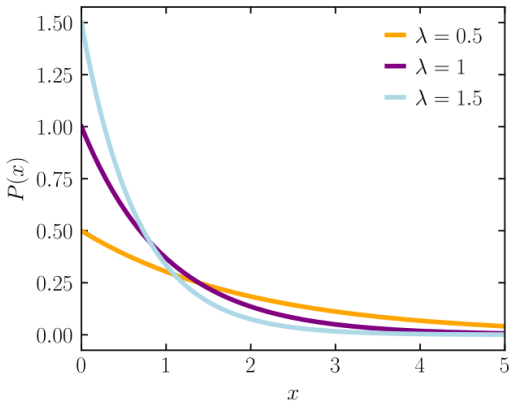
$\beta = SD\ of\ ln(x)$

Notation: $ln(X) \sim G(\mu, \sigma)$



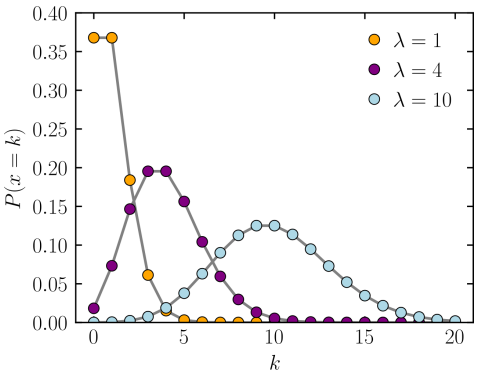
Boltzmann distribution

Another extremely famous and used distributions (computational chemistry):



Poisson distribution

Another extremely famous and used distributions (criminal justice):



the beauty of it is that it can be derived exactly.

Correlated variables

Let's make a few example of correlated variables:

1. Basketball playing skill and hight
2. Age and hair loss
3. Oil and gas production vs oil price
4. Wind and wind turbine efficiency
5. Wind and energy production
6. Rain and energy production
7. Solar irradiance and energy production

How to approach

There are a set of questions that one shall pose when relating two variables.

1. Statistical dependence

Two variables have their distribution and, even if very similar, are unrelated

1. Causal dependence

Two variables depend on each other.

Discussion point

How does this relate to soft and hard modeling?

Visualization

- Scatterplots (MatPLotLib)

It is one of the simplest ways to graphically display their relationship (it can be 3D).

- Heathmaps (seabon)

Just a 2d histogram but with a better name.

- Correlation matrix plots (pandas)

Matrix of plots

Correlation

The covariance or joint variance between two random variables is an extension of the concept of variance.

$$Cov[X, Y] = \sigma_{xy} = E[(X - \bar{X})(Y - \bar{Y})] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})$$

$$= \frac{N}{N-1}(E[XY] - E[X]E[Y])$$

- Generalization of variance
- Consider the covariance of a variable with itself

Note:

Variance: always positive
Covariance: positive or negative

Correlation Analysis

- The correlation between two random variables is a measure of the strength of their linear relationship
- Parametric Correlation

Measures a linear (Pearson) dependence between two variables (x and y) is known as parametric correlation test because it depends

- Non Parametric Correlation

Spearman and Kendall: they are rank-based correlation coefficients done for categorical data.

Pearson

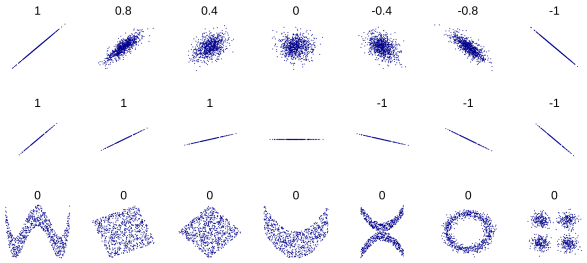
The Pearson's correlation coefficient ρ between to

$$\rho_{X,Y} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

- Assumes normal distribution (σ is a standard deviation)
- $\rho_{X,Y}$ ranges between -1 to + 1 (Strong negative, weak and Strong Positive)

Paerson

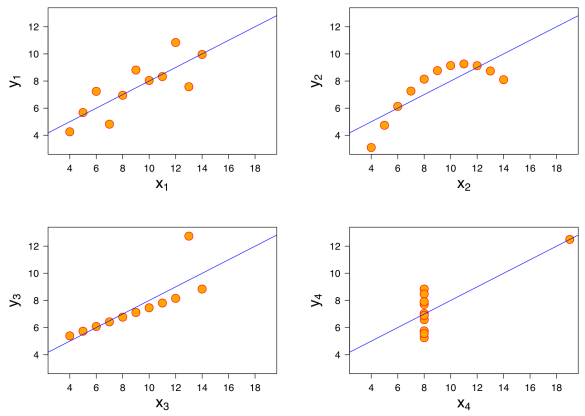
Why should we bother? ... check the assumption here!



Regression Coefficient - limitation

Anscorbe's Quartet: Four different pairs of variables

4 distributions with the same means (7.5), standard deviation (4.12), correlation (0.81) and regression line (y=3 + 0.5x)

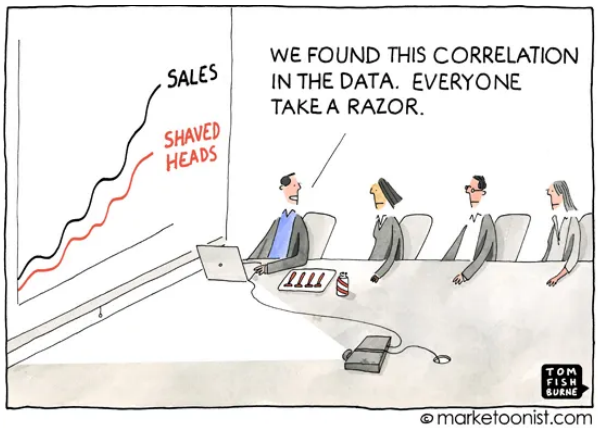


Spearman Rank Correlation

The Spearman correlation evaluates a monotonic relationship between two variables - Continuous or Ordinal and it is based on the *ranked* values for each variable rather than the raw data-

- Rank correlation compares the ranks (ordering)
- Calculated the same way as the Pearson correlation coefficient but using ranks instead of values

Correlation does NOT indicate Causation



Uncorrelated and Independent Random Variables

The two random variables X and Y are said to be

Uncorrelate

if : $Cov(X, Y) = 0$

Independent

if : $f_{XY}(x, y) = f_X(x) f_Y(y)$

- Correlation without Causation

Two variables might be correlated (and hence dependent) due to a coincidence, a lurking variable, or confounding factor.

- Common cause

Two variables might be dependent because they are both influenced by a third variable

Causal Analysis!