Fundaments of Machine learning for and with engineering applications

Enrico Riccardi[1]

Department of Energy Resources, University of Stavanger (UiS).[1]

Jan 22, 2025

Statistics

Definition

Statistics is the science of acquiring and utilizing data

- It comprises tools for data collection, summarization, and interpretation.
- The aim is identifying the underlying structure, trends, and relationships inherent in the data.
- Is it all statistics then? Yes.
- **Numbers to data, data to information**

Data properties

Before we talk about machine learning, we need to refresh some terminology.

Population

The universe of all possible outcomes and events.

Sample

A finite subset extracted from the population.

Exhaustivity

The samples covered the population spectra.

Representativity

The population is properly described by the samples.

Big data

We speak of big data when dataset are very large: i.e. many instances and features Models have thus a large set of parameters (and often no one has a clue anymore of what is going on).

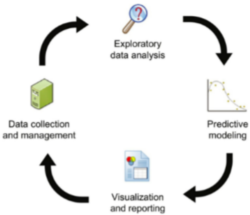
- Volume of data
- Variety different types of data sources with different length and scale.
- Frequency of data generation

Sampling

Samples shall have no bias (to be randomly selected). If not, the bias has to be corrected for.

Cycle of data

1. Data is collected
2. Checked upon
3. Some modelling
4. Analysis and visualization



Data quality

1. Data has to be acquired and integrated
2. Data are passed to a quality analysis and control
3. Data cleaning, consistency check. Most of time goes here

Preliminary Modeling

Main tasks:

1. Hunt for redundancy
2. Reduce dimensionality
3. AnOmAlIes removal

- Descriptive modeling (unsupervised learning)
- Predictive modeling (supervised learning)
- The model can be used to guide data acquisition (risky!)

Visualization and reporting

- The data has to be condensed into a visualization to provide input for decisions.
- Depending on the goal, very very different visualizations are possible.
- Use a model to indicate what is undersampled or oversampled.

Summarizing and visualizing data as a starting point for more analysis later on.

- Computing summary statistics (e.g. means and variance)
- Determining conditional probabilities of cause+effect relationships
- Calculating correlation and rank correlation coefficient between two variables
- Visualizing univariate, bivariate and multivariate data
- ...

Exploratory data analysis

Summarizing and visualizing data as a starting point for more analysis later on.

- ...
- Estimating probability coverage levels for different distributions
- Analyzing behaviour of normal distributions
- Calculating confidence interval and sampling distribution for the mean
- Testing for significance of difference in means
- Comparing two different distributions for statistical equivalence
- Developing a nonparametric regression model from given data
- Reducing data dimensionality
- Grouping data

Random Variables

- A random variable is a real valued function that assigns a value to each outcome in the sample space
- A random variable (RV) can be either discrete or continuous
  - Discrete RV
  - Continuous RV

- The probability mass function (PMF),P, of a discrete RV, X, denotes the probability that the RV is equal to a specified value, a.

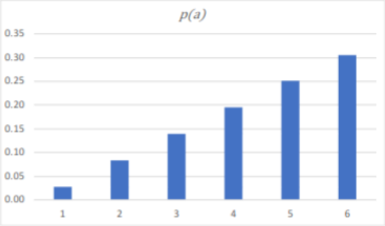
$$\mathrm{p(a) = p(X = a)}$$

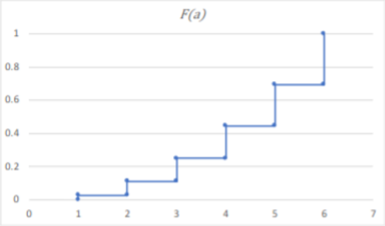
- The cumulative distribution function (CDF), F, denotes the sum

$$F(a) = P(X \leq a) = \sum_0^a f(x)dx$$

Random Variables

a	1	2	3	4	5	6
$p(a)$	1/36	3/36	5/36	7/36	9/36	11/36
$F(a)$	1/36	4/36	9/36	16/36	25/36	1





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
- etc

Wind turbine example 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)
  - Sariance (Var) or coefficient of variation
- Measures of shape
  - Skewness, modality

Histograms

- Task 1: make a histogram from a 2d random distribution
- Task 2: make a 2d heat map from a 2d random distribution

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

Exercise

Plot a histogram from different random number distributions and bin sizes.

Cumulative Histogram

- Cumulative frequency
- Each data point can be plotted individually
- It helps to read quantiles and compare distributions
- Practice with your jupyter notebook

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

* 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

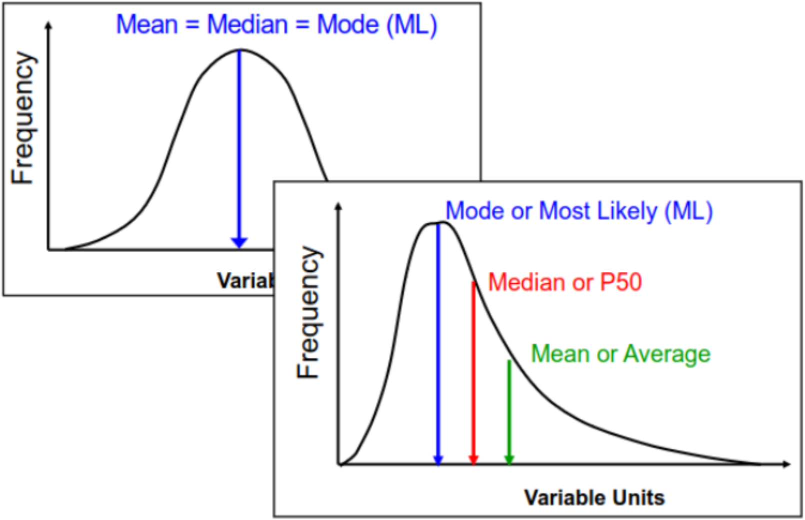
Mode

- 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.

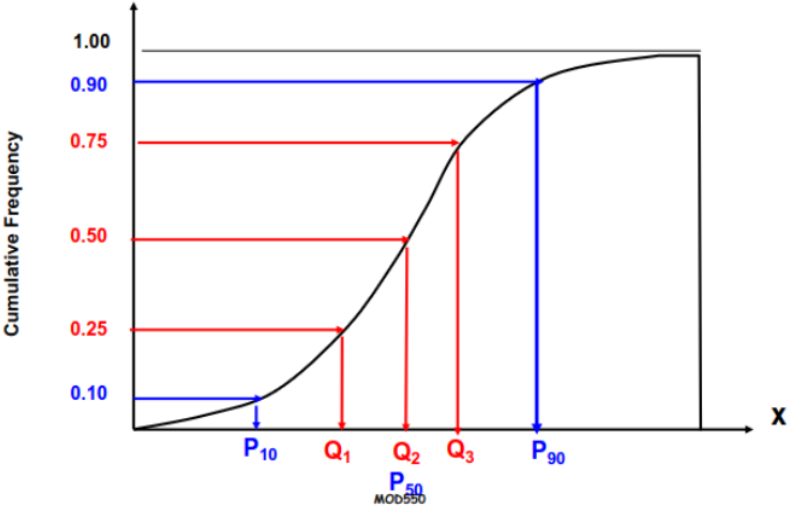
Percentiles

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 = maximum - minimum

Inter-quantile Range

IQR = Q3 - Q1

Mean Deviation from the Mean

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

Mean Absolute Deviation

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

Variance

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

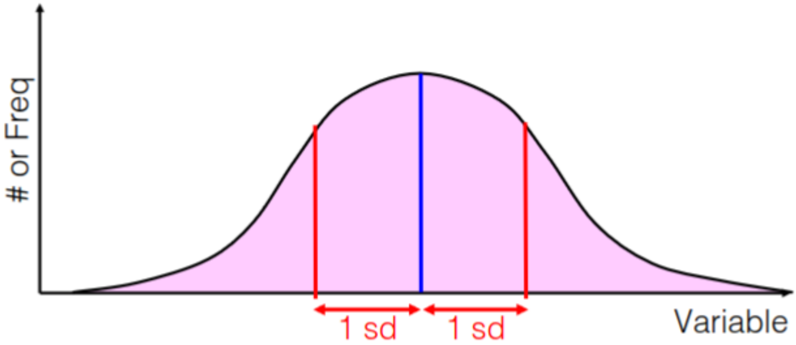
Variance

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

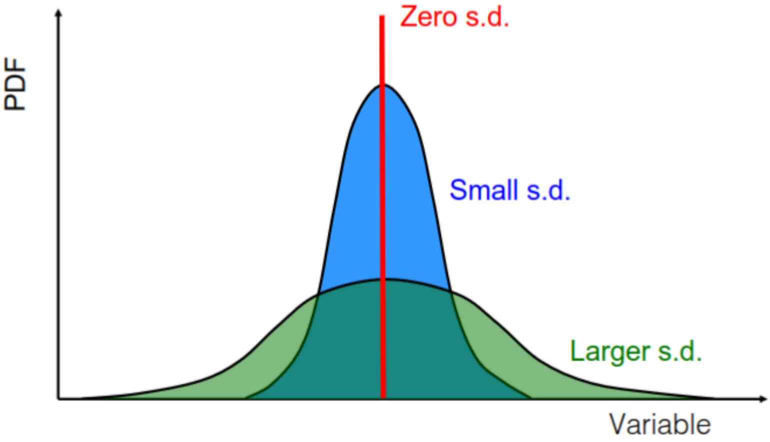
Standard Deviation (SD)

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

Standard Deviation



Standard Deviation



Measures of dispersion

Standard Deviation (SD)

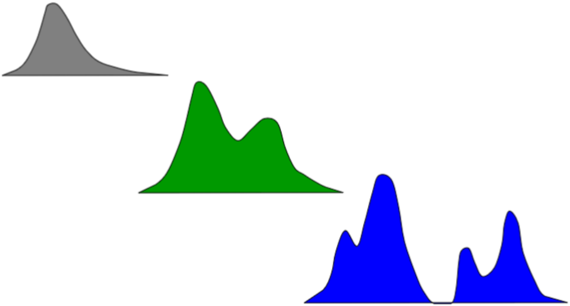
$$SE_x = \frac{s_x}{\sqrt{n}}$$

Coefficient of Variability

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

Modality

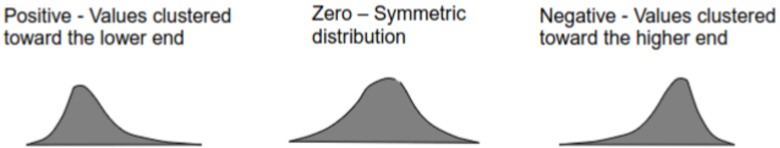
- Unimodal
- Bimodal
- Polymodal



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

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 uncertainity

PDF: $f(x) =$

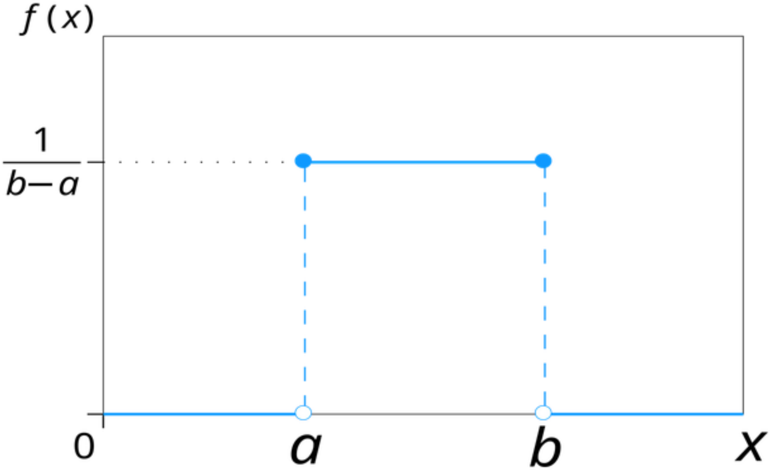
$/frac1b - a, a \leq x \leq b$

CDF: $F(x) =$

$/fracx - ab - a$

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: $f(x) =$

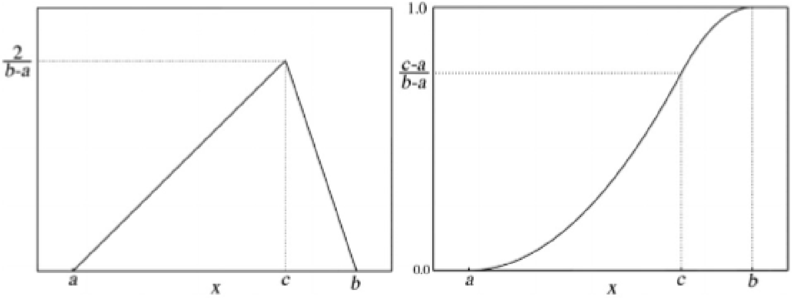
$$\frac{2(x-a)}{(b-a)(c-a)};\ \text{if } a \leq x \leq c$$
$$\frac{2(b-x)}{(b-a)(c-a)};\ \text{if } c \leq x \leq b$$

CDF: $F(x) =$

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

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(s=

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 Symmeytic 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