# MOD300 Anvendt Python programmering og modellering

Enrico Riccardi[1]

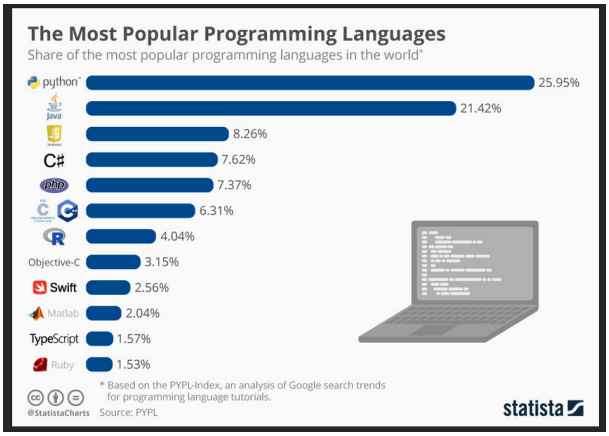
Department of Mathematics and Physics, University of Stavanger (UiS).[1]

Oct 2, 2025

University of Stavanger



## Popularity




## Coding standards

| Lifetime | Use |
|---|---|
| 1-shot | 🚫 |
| Week+ | Git + Github/GitLab |
| 3 month+ | + Testing |
| 6 month+ | + Documentation, automated testing |

| Dev/Users | Use |
|---|---|
| 1 | Push to main |
| 2+ | + Branches, merging |
| 2+ (+students) | + Code review |
| 2+ (+external) | + Release branch |

## Version control

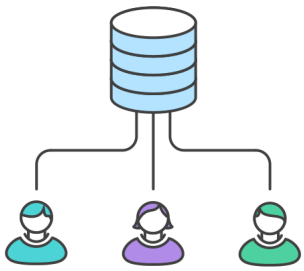

## Git

Git is a distributed version control system that tracks changes in any set of computer files, usually used for coordinating work among programmers who are collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows (thousands of parallel branches running on different computers). [Wiki]

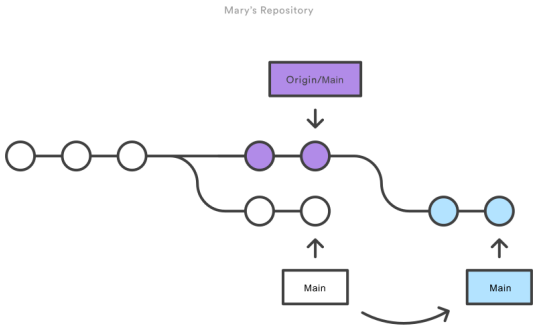
**Let's try to be more accessible.**

Git is a computer program/tool to save and download files on a hosting server (e.g. GitHub and GitLab).

## Centralized workflow

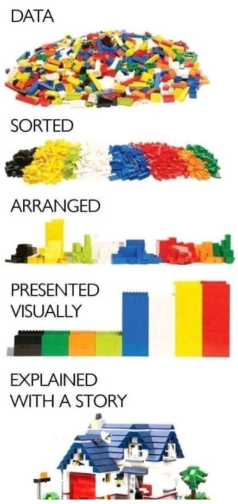


## How does it work -in short-



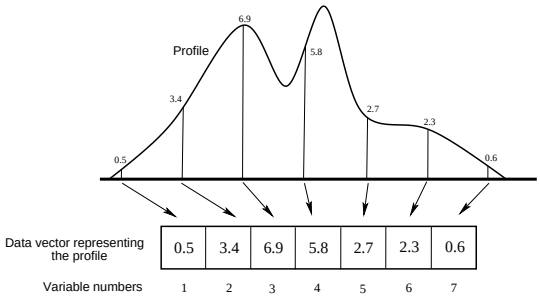

## It can help your CV! maybe...


## Data



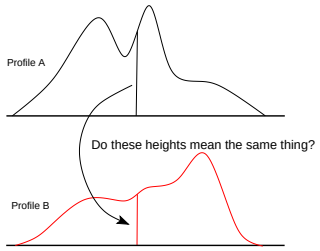

## Sampling point representation (SPR)

- An intuitive way to represent curves and spectra is the **sampling point representation**.
- We sample at regular intervals where each sample point is represented by a variable




## Sampling point representation (SPR)

- SPR is useful until point *i* in a curve has the same meaning of the point *i* in another curve.




- Which parts of the profiles or shapes are comparable, i.e. have the same meaning?

## Statistics (recaps)

**Definition**

Statistics is the science of acquiring and utilizing data

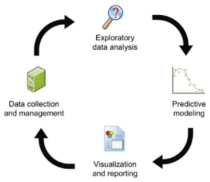
- It comprises tools for data collection, summarization, and interpretation.
- The aim is identifying the underlying structure, trends, and relationships inherent in the data.
- Is it all statistics then? Yes.
- **Numbers to data, data to information**

## 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




## Preliminary Modeling

**Main tasks:**

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

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

## 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. p(a) = p(X = a)

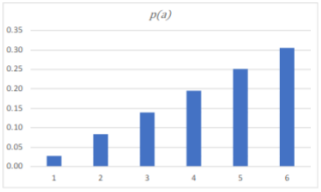
The cumulative distribution function (CDF), F, denotes the sum $F(a) = P(X \leq a) = \sum_0^a f(x)dx$

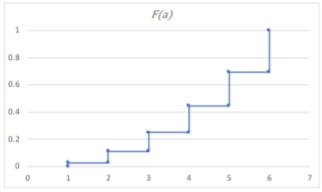
## Wind turbine example

| 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 |

## 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 |




## 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?**

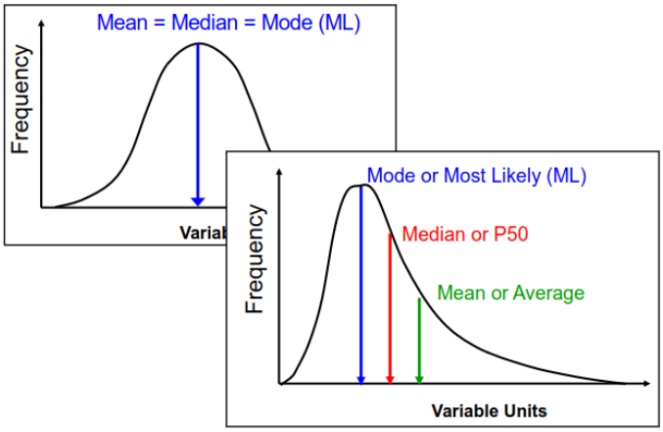
No. It shows variability, but it can be used to quantify uncertainty.

## Median

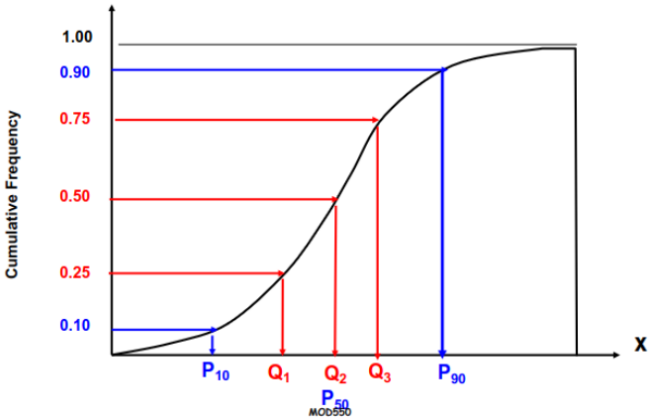
```
if n is odd:
    median = x[(n+1)/2]
else:
    median = x[n/2] + x[(n/2)+1]
```

- 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

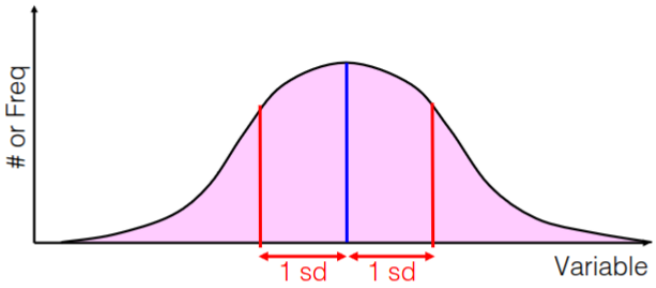
## Distribution Descriptors



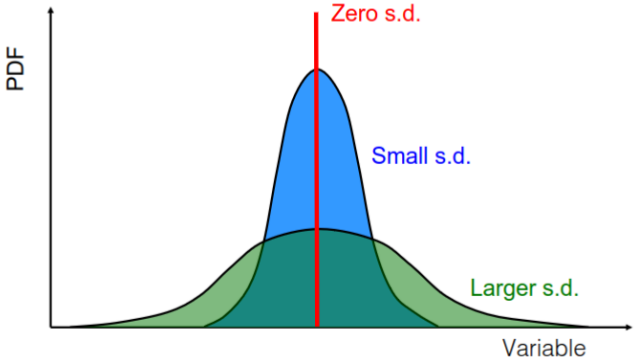

## Distribution Descriptors




## Standard Deviation




## Standard Deviation




## Measures of dispersion

**Standard Deviation (SD)**

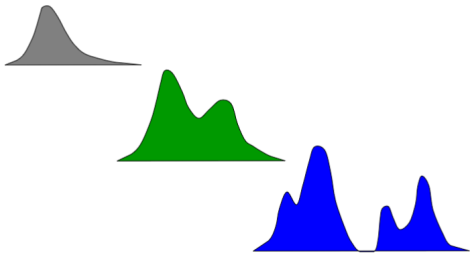
$$SE_x = \frac{s_x}{\sqrt{n}}$$

**Coefficient of Variability**

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

## Modality

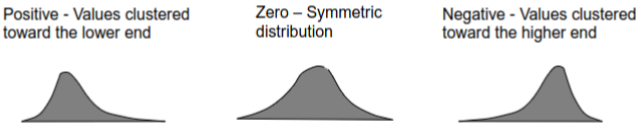
- Unimodal
- Bimodal
- Polymodal



## 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

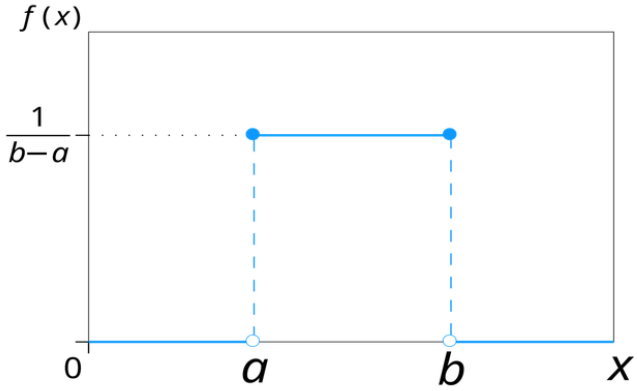
**Distribution**

Means of expressing uncertainty or variability

Models

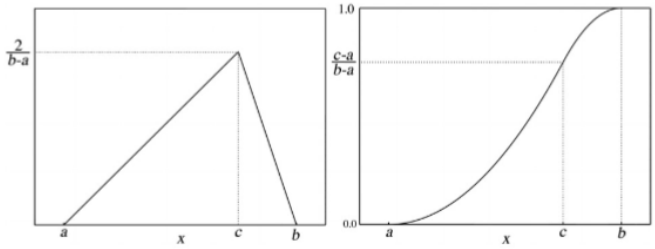
- Uniform: useful when only upper and lower bounds are known
- Triangular: useful when estimates of min, max, mode [P10, P50, P90] are available
- Normal: symmetric model of random errors or unbiased uncertainties with mean of standard deviation specified
  - Very common for observed data
  - Additive 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




## Triangular Distribution

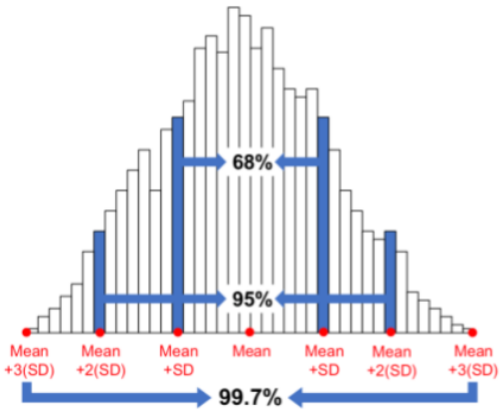
Notation: $X \sim T(a, b, c)$




It can be symmetric or asymmetric

## Normal Distribution

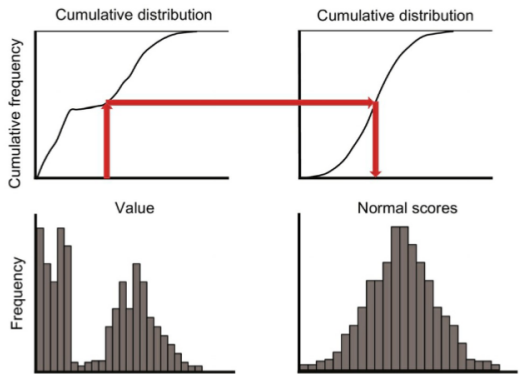
Notation: $X \sim G(\mu, \sigma)$



## 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

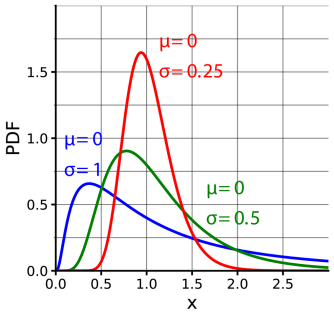
## Log - Normal distribution

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

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

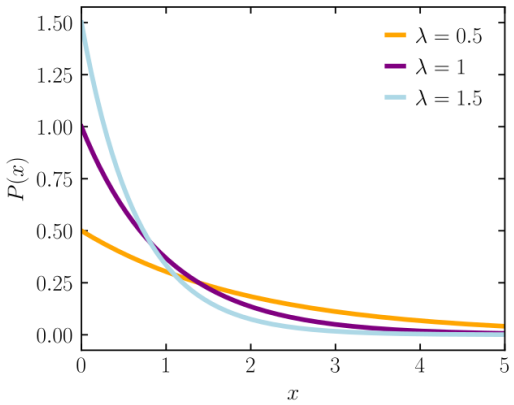
$\beta = SD\ of\ ln(x)$

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



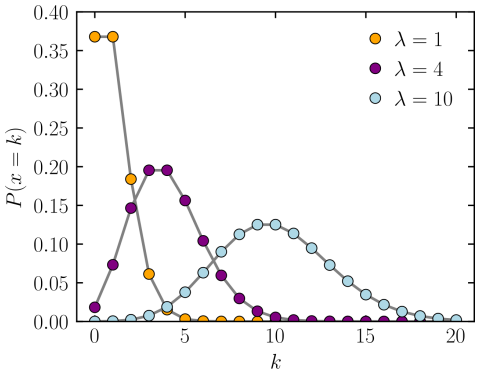

## Boltzmann distribution

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.

## Radial distribution function