Introduction to the Monte Carlo method

MUSE Methods for technical and economic energy analysis

Dr Jonathan Chambers

Introduction

- This aim of this lesson is to introduce the Monte Carlo Simulation method in the context of Energy research
- You should:
  - Gain an understanding of what the Monte Carlo method is
  - Understand the importance of uncertainty and how Monte Carlo methods can be used to address it
  - Understand basic notions of probability distributions in the context of Monte Carlo methods
  - Gain an overview of the Energy modelling example to be used in the exercise

The Monte Carlo (MC) method

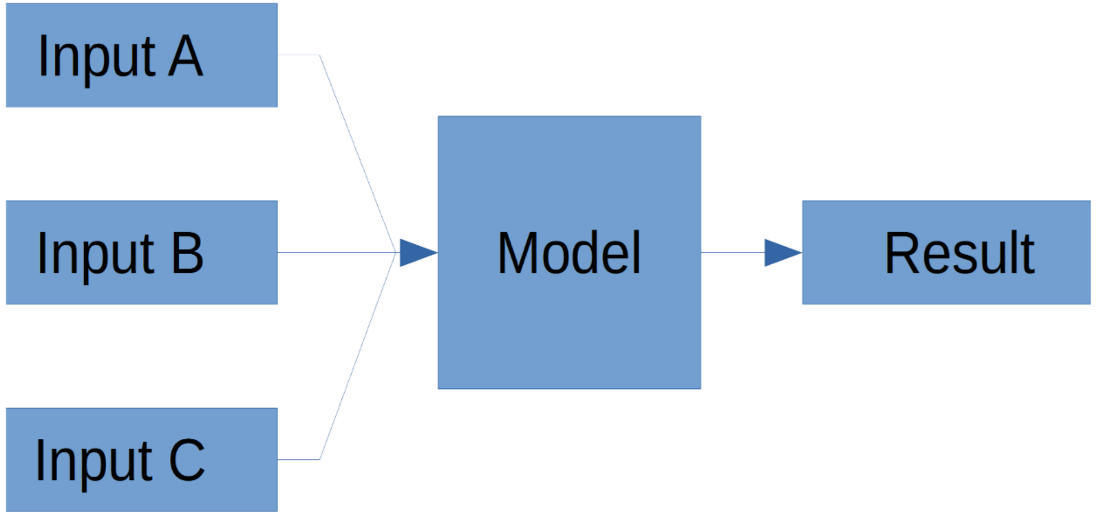
(this is not the Monte Carlo method)

The Monte Carlo (MC) method

- MC simulation is a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- Use randomness to solve problems that are **theoretically deterministic**, **but difficult to solve in practise.**
- ***Steps:***
  1. Build a *deterministic* mathematical model (i.e. not random)
  2. Run the model many times using a *range* of inputs taken from a *random distribution* and *aggregate* the results

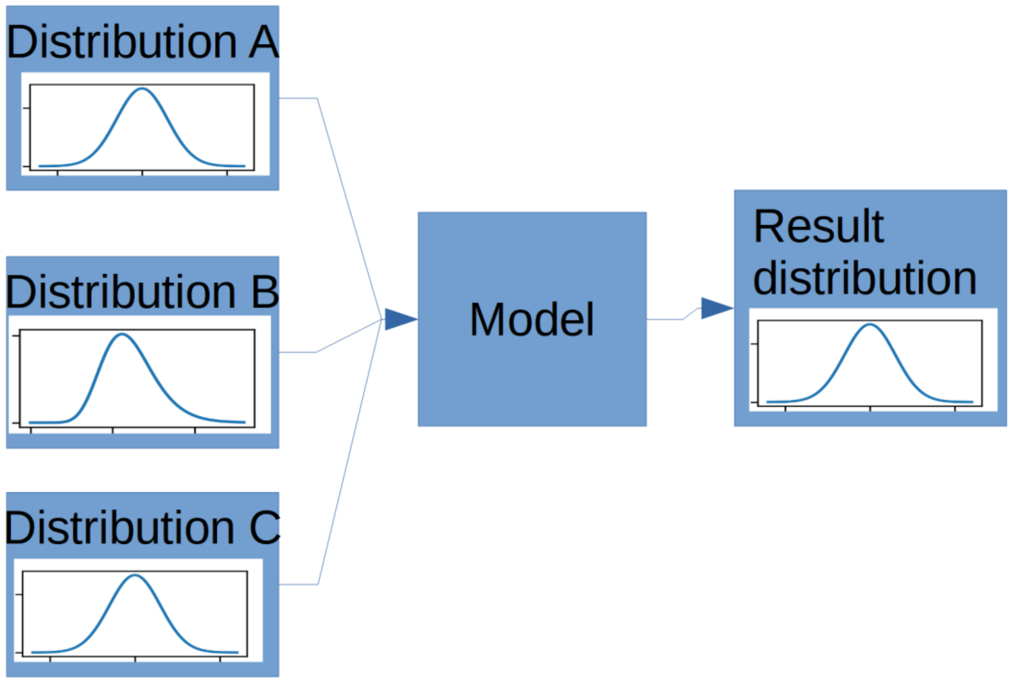
Deterministic models

- A deterministic model takes a set of input data and produces a result.
- Running the same model with the same data will always produce the same result.



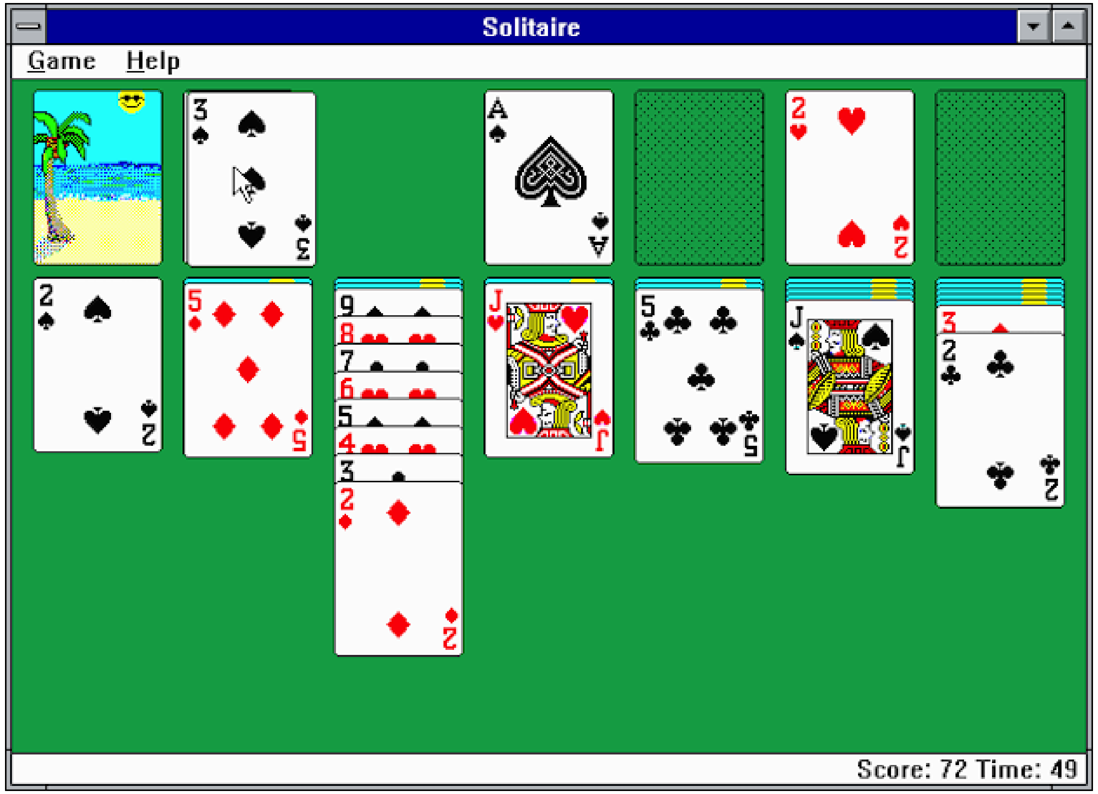
Monte Carlo method

- Build **distributions** of input parameters and insert them into the model.
- Calculate results many times, each time using a different set of random values from the probability functions.
  - Might need 1000s of iterations
- Produces distributions of possible result values.



An example: Solitaire

- *Q: What is the likely configuration of cards at the outcome of a game?*
- There is no random element to the game, only the initial configuration of cards matters.
- Very difficult to write a mathematical description to predict the output!
- BUT: easy to play some games and record the output.
  - Even easier if you can write a computer program to do this for you!



Example and Exercise

- We will discuss the Monte Carlo method in the context of an example:

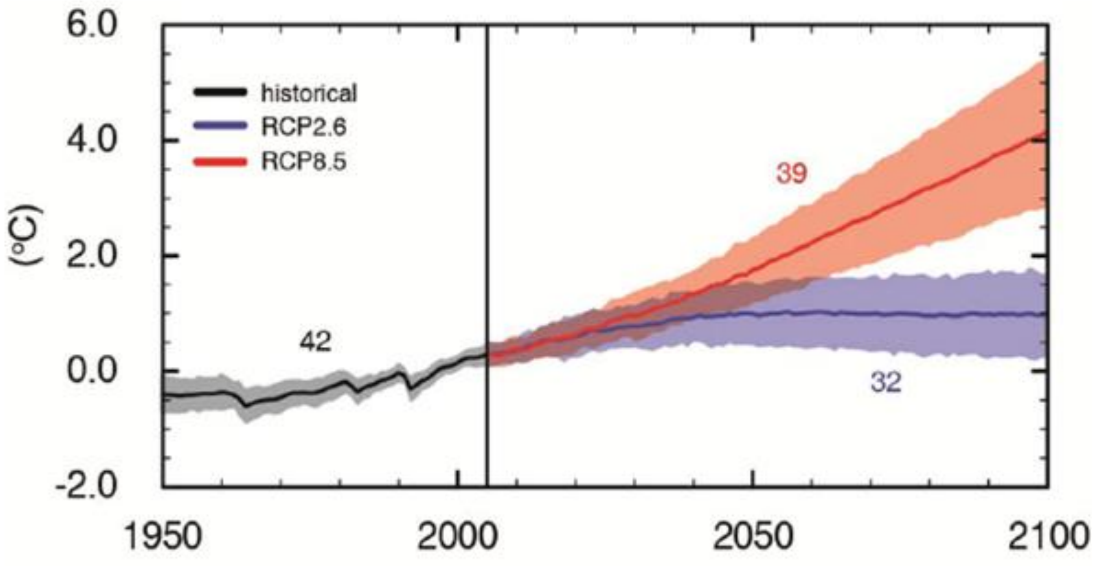
Modelling the energy demand for cooling in Switzerland, today and in the future

- Thursday: introduce the concepts
- Friday: work through the example

Modelling cooling demand and climate change

- Cooling is currently mostly used in office buildings in Switzerland.
- With climate change, expect to see increasing demand.
- **How can we estimate the demand for today and the future for all the office buildings?**

-> We build a model!



Projected global surface temperature change under different emissions scenarios. Zero is set at the average of 1986-2005 levels (Figure SPM.7(a). IPCC Working Group I Assessment Report, Summary for Policy Makers, 2013)

Model part 1:
Yearly cooling energy demand for one building

$$\mathbf{E_{cooling} = B.C. * P_{th,m2} * NH * A / COP}$$

Outputs:

- $E_{cooling}$: yearly final energy demand for cooling.

Inputs:

- B.C. : Building Cooled = 1 if building has cooling system, = 0 if it does not.
- $P_{th,m2}$ **:** Thermal cooling power need per m^2, the amount of heat energy per second (i.e. Watts) to be removed per m^2 of building area.
- NH: the number of hours of operation of the cooling system per year.
- A: cooled office building floor area.
- COP: Coefficient Of Performance of the cooling system.

Model part 2: Yearly cooling energy demand for all buildings

- Find the values of all the parameters in part 1
- Calculate the demand for each building
- Sum the total energy demand over the **number of office buildings** $\mathbf{N_{buildings}}$

$$E_{Total} = \sum_{i=0}^{N_{Buildings}} E_{Cooling,i}$$

Summary of model parameters

- In the exercise, we will find inputs for each of these parameters.

Variable	Symbol	Unit
Total number of office buildings	$N_{buildings}$	-
Is the building cooled?	B.C.	-
Total office (non-residential) area	A	m^2
Cooling thermal power per unit area	$P_{th,m2}$	kW/m^2
Cooling system Coefficient of performance	COP	-
Number of operating hours per year	NH	hours

Scenario modelling 1:
Effect of Climate Change

- We expect climate change to affect:
  - The number of buildings using cooling
  - The number of hours per year that the cooling will be used
- We measure the effect of climate change through the change in **Cooling Degree Days (CDD)**
  - CDD is defined as the sum of degrees that the outside temperature is above a base temperature
  - The cooling energy demand for a building are assumed to be proportional to the CDD

Scenario modelling 2:
Effect of Improved Technology

- We can also imagine that we get better efficienct with better technology and policy
- Improved COP as a result
- We would like to model a 50% improvement in COP

Why do we use the Monte Carlo method in this case?

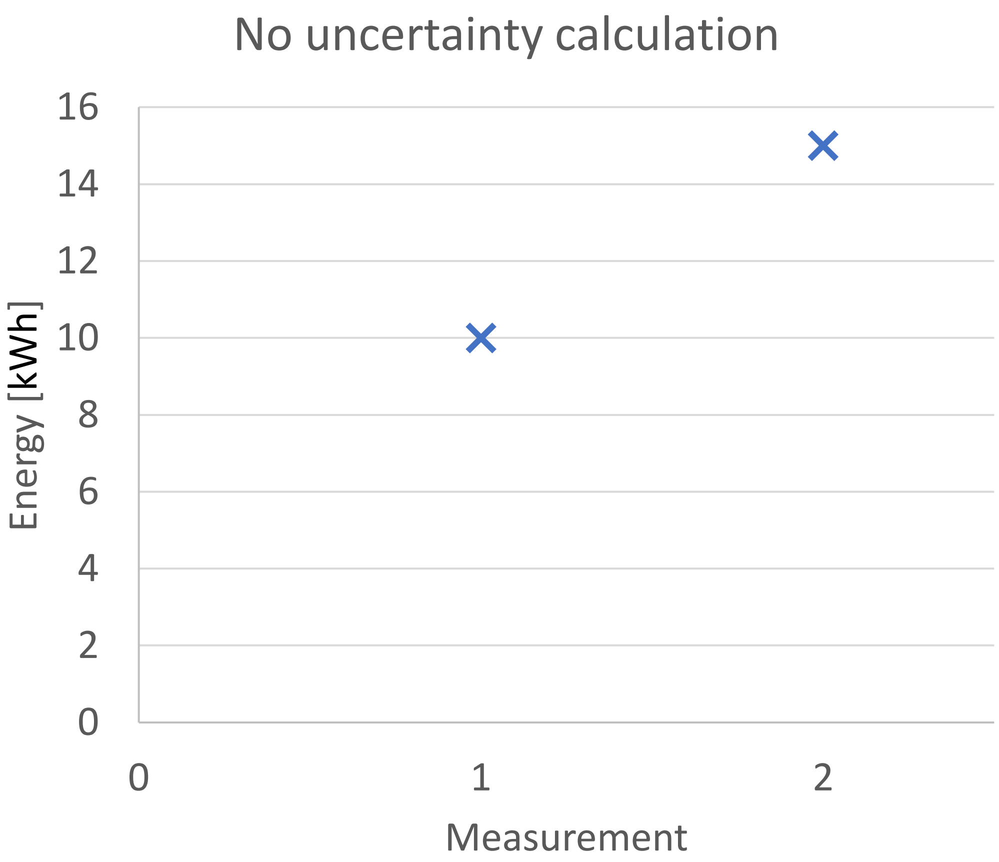
- We need model parameter values for each building
- We could try to find an average value for all buildings
  - But we know this isn't true – each building is different!
- We could try to find a variable value for each building
  - Better, but if we pick at random for each building, we only get one result
- Monte Carlo method:
  - We pick from a *distribution* of values for each paramter
  - We calculate the *total* energy demand for all the buildings for a given set of parameters
  - We do this *many times* to get the *distribution* of results

Why use do we use the Monte Carlo method in this case?

- MC allows us to calculate the **uncertainty** in the model output
  - Measure uncertainty using the **95% confidence interval (CI)**
  - Gives a result X ± CI.

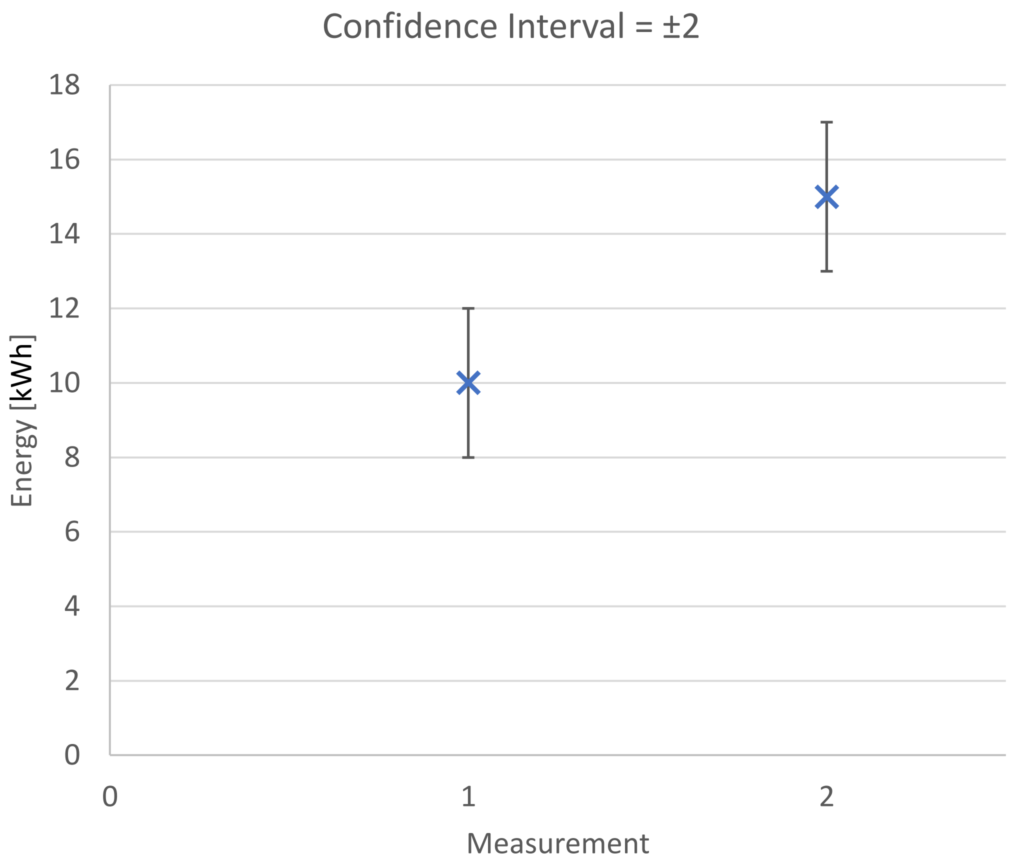
- MC allows us to change model input parameters and determine whether there is a **statistically significant** change in the result

The importance of uncertainty



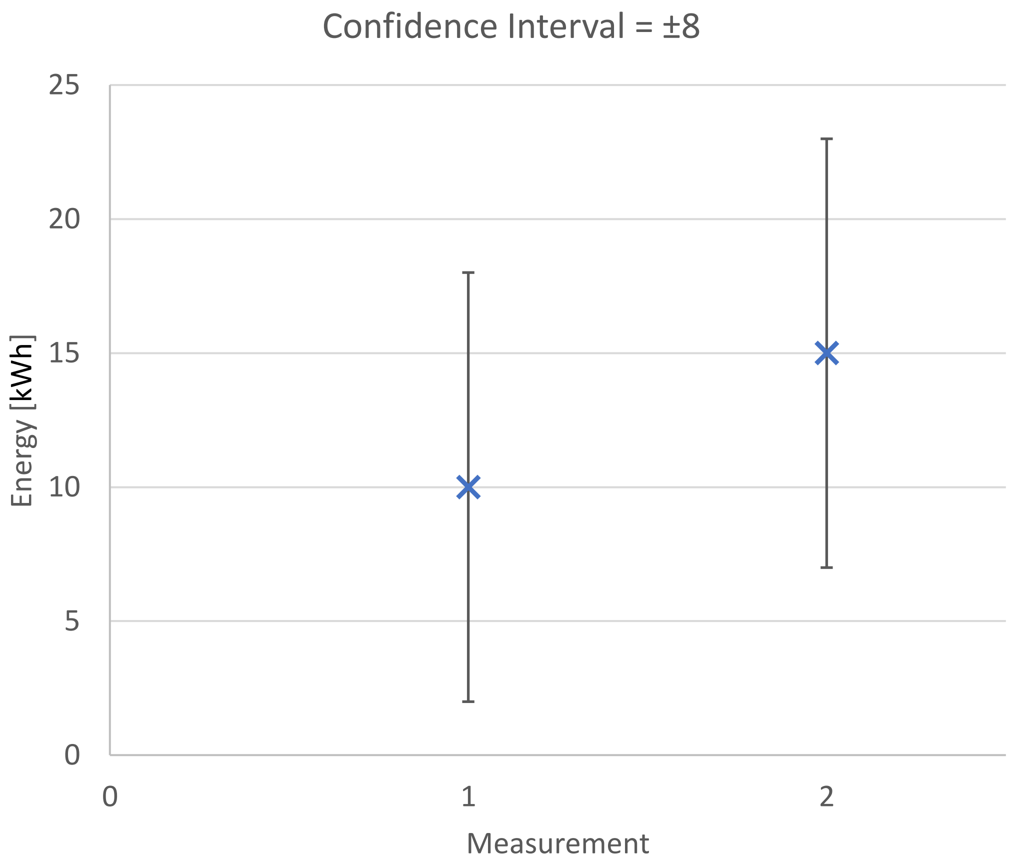
- **Is there a significant difference between these two measurements?**
- The absolute difference is 5.
- *Without an estimate of uncertainty, this doesn't have much meaning!*

The importance of uncertainty



- Confidence Interval = ±2
- In this simple case, we can add the confidence intervals to get the CI of the result.
- Difference = 5±4
- There is a difference, but the uncertainty is quite large!

The importance of uncertainty



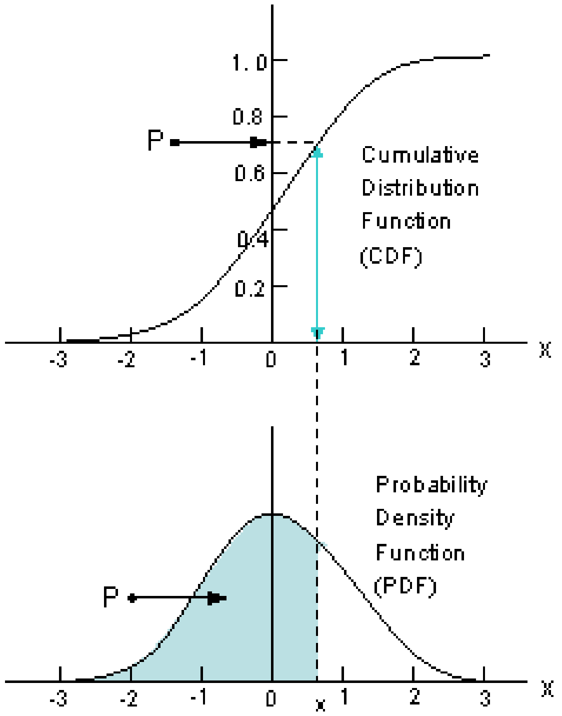
- Confidence interval = ±8
- Already from the graph, it is clear the uncertainties overlap a lot!
- Difference = 5±16
- *We cannot say that the two values are definitely different!*

What is a confidence interval?
Probability Distributions and the Normal distribution

- Probability distributions tell us how likely a given result is
- Example: rolling a dice.
  - Each number has an equal probability, so the distribution is *uniform*
- We describe probability distributions using:
  - **Cumulative Distribution Function**
  - **Probability Density Function**
- The **Normal distribution** appears in many places and has important properties

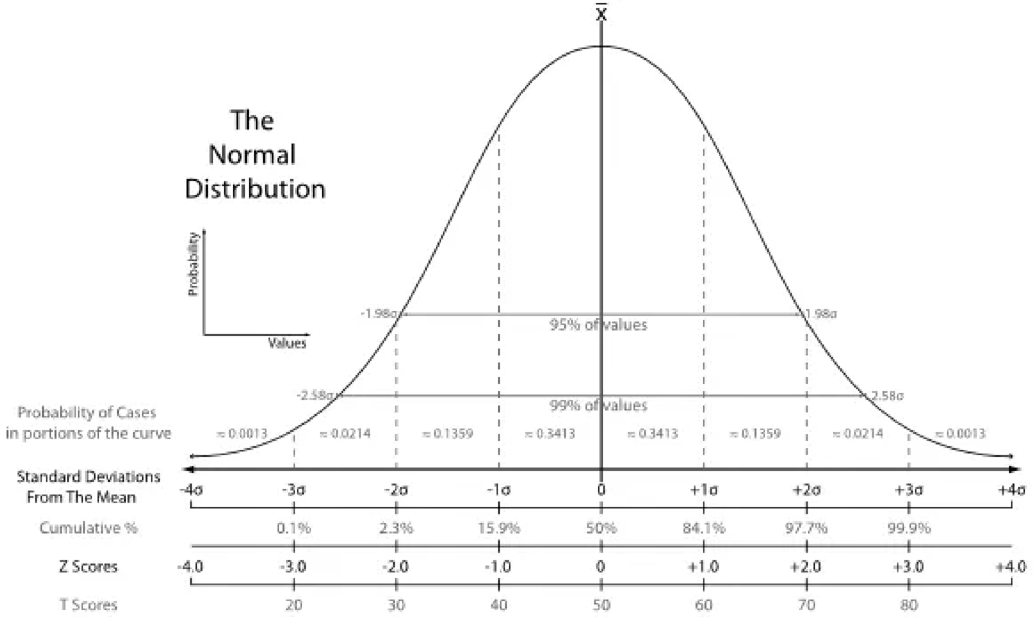
Cumulative Distribution Functions (CDF) & Probability Density Functions (PDFs)

- CDF maps from a uniform distribution between 0 and 1 to the probability of x.
  - Easy to generate a random value between 0 and 1 and use it to get the desired probability value.
- PDF define the range of possibilities and the relative probability of those possibilities.
  - The value of the PDF at *x* gives the probability of *x* occurring.
  - Often see in illustrations: the ‘bell curve’



The normal distribution

- Many analyses assume normal distributions.
- The **mean** is the middle of the distribution.
- The **standard deviation** "σ" is the measure of how 'wide' the distribution is.
- The most commonly encountered distribution because of the **Central Limit Theorem.**



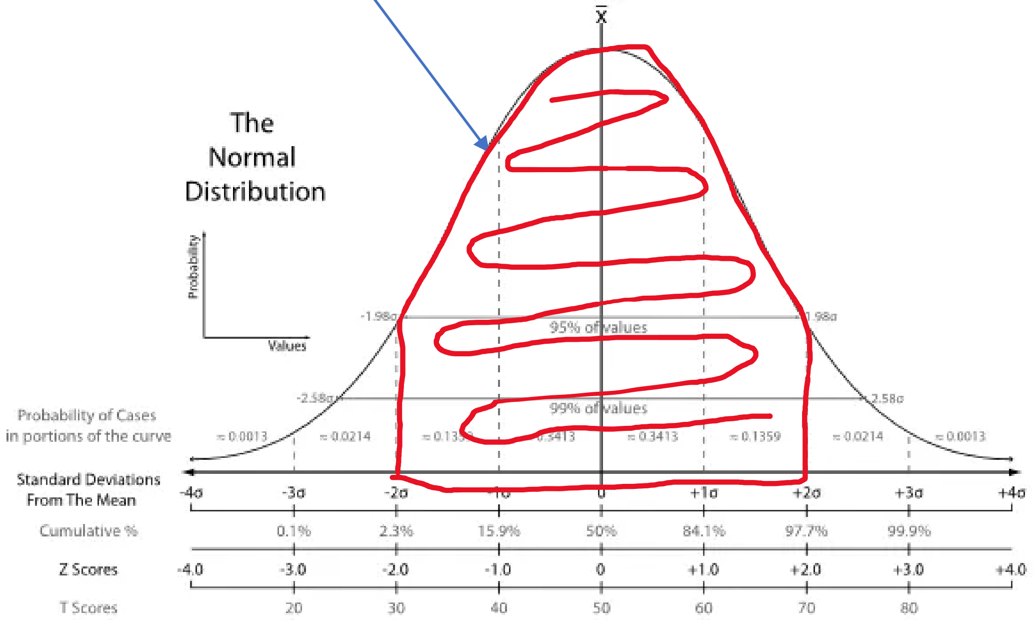
95% confidence interval

- The range within which 95% of the values are expected to be found
- For normally distributed measurements:

$$\mathbf{CI = 1.96 * \sigma}$$

σ is the standard deviation

95% of values



The central limit theorem

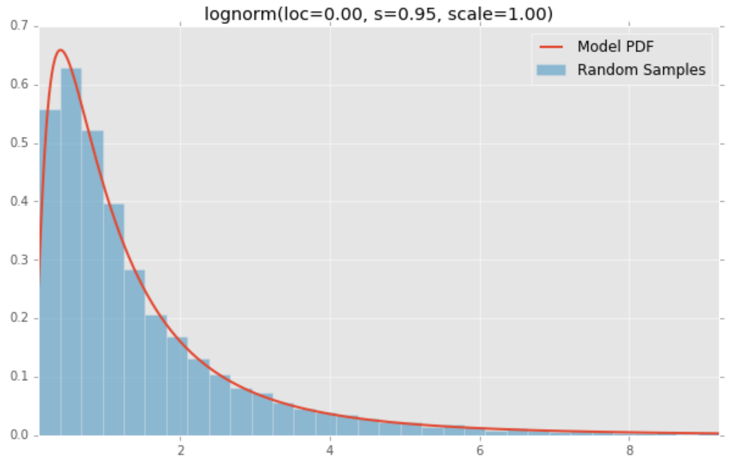
The means of the sum of randomly distributed values tends towards a normal distribution as the sample size gets larger

- Take repeated *samples* from a *population* of values and sum each sample.
    - e.g. measure the energy consumption of N randomly selected houses and sum it up. Repeat
- The *distribution of the sums* will be normally distributed, even though the *distribution of the population values are not.*
    - E.g. the distribution of energy consumption of homes is NOT a normal distribution. But the distribution of the SUM of energy consumption from sets of randomly selected homes IS a normal distribution
- NOTE: this does NOT mean that if you make enough samples of every kind of distribution you will always get a normal distribution!
- This allows use to make useful assumptions when doing statistical analysis!

Important probability distributions for our model

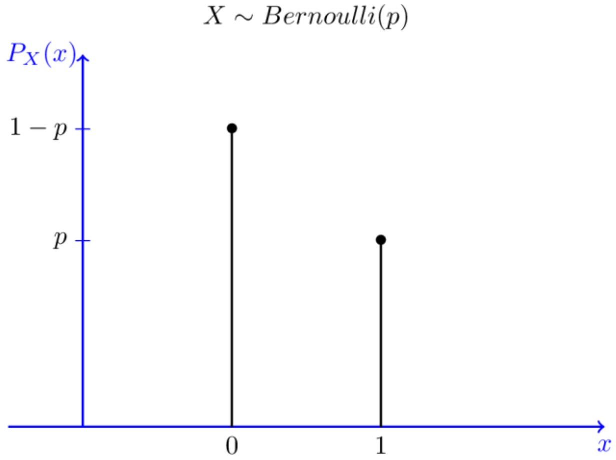
While we expect the **output** of the model to be a normal distribution, the **inputs** can be very different

Log-normal distribution is a continuous probability distribution of a random variable whose **logarithm** is normally distributed.



Values that are always > 0 may be log-normal distributions, e.g. energy consumption

Bernoulli distribution is the **discrete** probability distribution of a random variable which has the value 1 with probability **p** and the value 0 with probability **1-p**.



Values that are true or false follow a Bernoulli distribution, e.g. whether a building has cooling

Distribution parameters

- **Parametric** distributions are controlled by one or more **parameters -** the variables of the equations defined the distribution.
  - E.g. for a normal distribution, the parameters are the mean (μ) and standard deviation (σ):

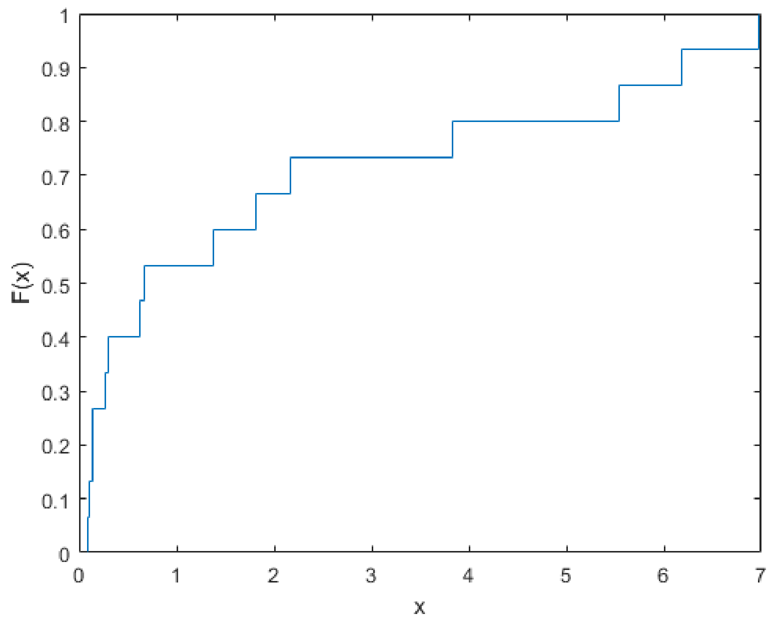
$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

You don't need to remember this equation!

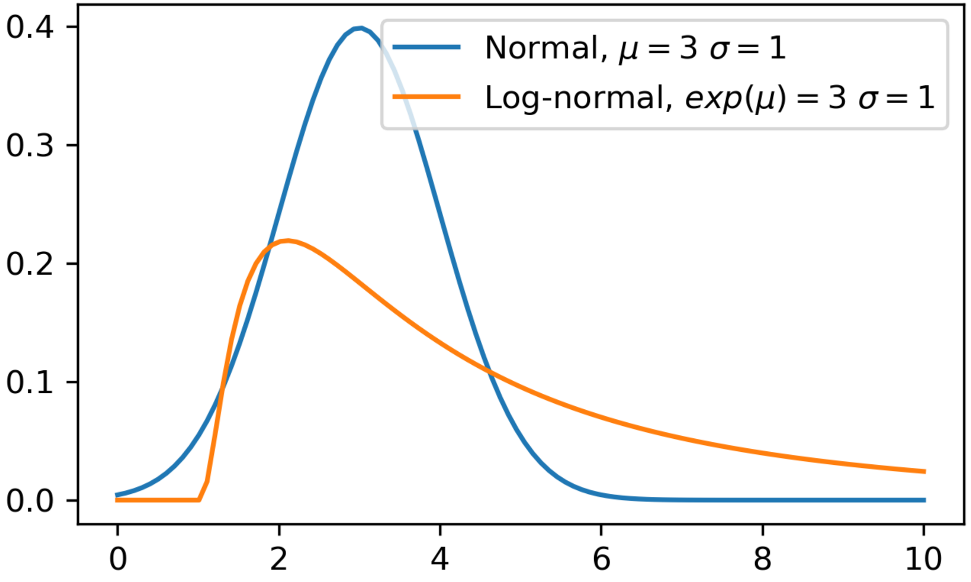
- For our MC model, we need to:
  - Figure out which distribution corresponds to our inputs
  - Calculate the parameters of this distribution
- Often we can do this by calculating the parameters from existing data

Empirical distributions from data

- Can also use collected data directly to define a distribution
  - Represented by an “Empirical CDF”.
- Advantages:
  - Exactly models the input data
  - Don’t need to choose the right distribution to fit the data
- Disadvantages
  - Reproduces errors and biases in the input data
  - Common problem in machine learning: race bias, gender bias from building models from raw data – reproduces all of societies existing problems!



Comparing distributions

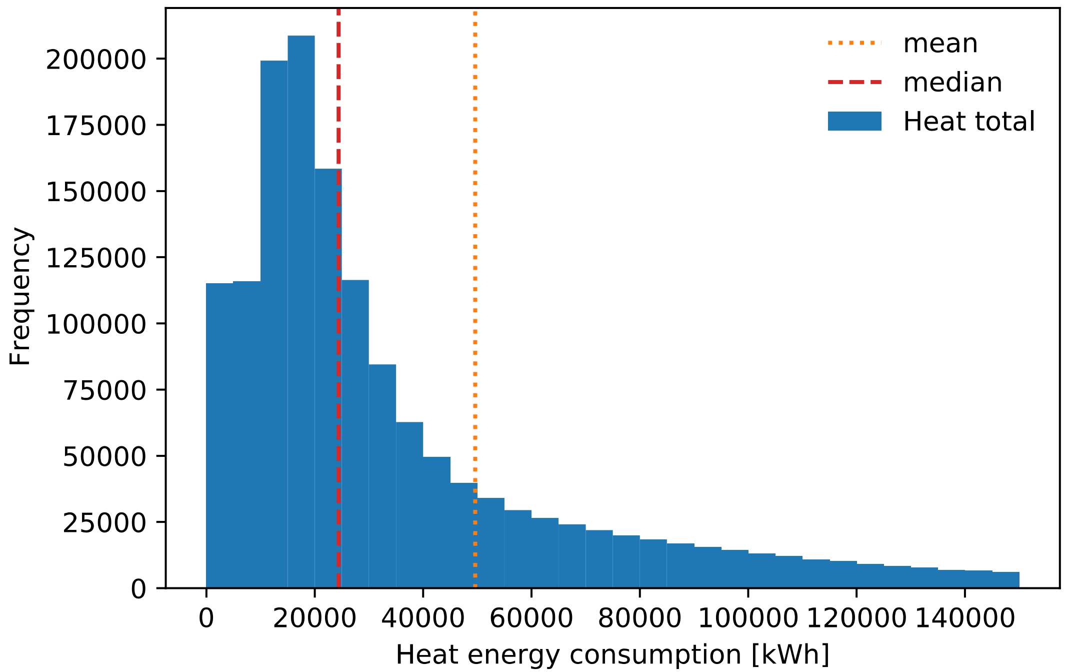


- Want to determine whether data follows a given distribution
- Statistical tools can test this for us
  - They compare “skewness” and “kurtosis” which are both 0 for the Normal distribution.

Aside: importance of distribution shapes

- The shape of distributions is of more than mathematical interest!
- E.g: the difference for a skew distribution between
  - mean (sum of values divided by number of values)
  - median (middle value)
- Which one gives a better impression of what a house 'typically' uses

Distribution of heat energy consumption per year for Swiss homes



Using the distributions

- Distributions produce a range of *values* with a given *probability.*
- Need to *sample* from these distributions.
- Get one value each time.
- Need to perform *multiple iterations* of the sampling.
  - i.e. repeating the same process many times

Putting it together

- Step 1: define a deterministic model based on a set of inputs.
  - This could be a few simple equations
- Step 2: find probability distributions for the model inputs.
  - Can be derived from theory or from data
- Step 3: run the model many times using inputs sampled from the distributions.
- Step 4: collect the results of all the model runs to produce a distribution for the model outputs.

Advantages of the Monte Carlo method

- Relatively simple to develop an MC model from a standard (deterministic) one.
- Results show not only what could happen, but how likely each outcome is.
- Allows to estimate uncertainties
- Integration of different data sources
  - ‘harmonize’ diverse inputs by converting them to probability distributions
  - Also accounts for the uncertainties introduced by combining different datasets.

Disadvantages of the Monte Carlo method

- Computationally expensive – have to run the model many times
- Developing input distributions is not alway straightforward

Summary

- Monte Carlo modelling is a method for probabilistic modelling.
- MC enables you to calculate the distribution of outputs for complex models.
- MC combines deterministic models with inputs sampled from probability distributions.
- MC requires running the model many times, each time sampling from the input probability distributions, to generate a distribution of results.

Some more probability distributions

- Probability distributions and their theory
  https://www.itl.nist.gov/div898/handbook/eda/section3/eda366.htm
- Visualising common distributions
  https://stackoverflow.com/questions/37559470/what-do-all-the-distributions-available-in-scipy-stats-look-like
- Python (SciPy) module for working with distributions
  https://docs.scipy.org/doc/scipy/reference/stats.html

Going further: Markov Chain Monte Carlo

- An often used method is “Markov Chain Monte Carlo” (MCMC).
- Markov Chains are a method for generating the random numbers for the Monte Carlo simulation.
- They are generally important when:
  - The input parameter space is more complex
  - There are (potentially complicated) inter-relations between the input parameters (i.e. the value of one input depends on the other input)
  - When modelling a process which has a ‘state’ such as behaviour of building occupants (e.g. a person can be sleeping, cooking, etc…)

Going further: Unobserved variables

- We considered only observed random variables: “known unknowns”.
- Also possible to take into account unobserved random variables: “unknown unknowns” – advanced MC modelling using prior distributions (assumed distributions for unobserved inputs).
  - Overlap with ‘Bayesian’ statistics, machine learning
  - Learn more: https://docs.pymc.io/learn.html
- Can be useful for inferring the characteristics of complex systems where you can’t observe everything.