Data Analysis

Ivica Puljak
University of Split, FESB, Split, Croatia
Ivica.Puljak@cern.ch

Starting the new era

In the future, calendar of particle physics will be

Before Higgs (BH) After Higgs (AH)

July 4, 2012

04.07.2012: Higgs within reach

Proton-proton collision in the CMS experiment producing four high-energy muons (red lines). The event shows characteristics expected from the decay of a Higgs boson but it is also consistent with background Standard Model physics processes (Image: CMS)

At a seminar on 4 July, the ATLAS and CMS experiments at CERN presented their latest results in the search for the long-sought Higgs boson. Both experiments see strong indications for the presence of a new particle, which could be the Higgs boson, in the mass region around 126 gigaelectronvolts (GeV).

01.08.2012: ATLAS and CMS submit Higgs-search papers

Protons collide in the CMS detector at 8 TeV, forming 2 beams which decay into electrons (green lines) and muons (blue). Such an event is compatible with the decay of a Standard Model Higgs boson (Image: CMS)

The ATLAS and CMS collaborations today submitted papers to the Journal of Physics Letters detailing the latest on their searches for the Higgs boson. The teams report even stronger evidence for the presence of a new Higgs-like particle than announced on 4 July.
The Standard Model

![Diagram of the Standard Model]

---

Higgs mass: theoretical constraints

- Problem: Higgs mass is free parameter
  \[ M_H^2 = 2\lambda v^2 \quad \ldots \quad v = 246 \text{ GeV} \]
- Theoretical constraints
  - Unitarity (no probabilities > 1)
    \[ M_H < 700 - 800 \text{ GeV} \]
  - Triviality (Higgs self coupling remains finite)
    \[ M_H^2 < \frac{4\pi^2}{3 \ln(\Lambda/v)} \]
  - Stability (of vacuum)
    \[ M_H^2 > \frac{4m_W^4}{\pi^2} \ln(\Lambda/v) \]

Collisions in LHC

- Proton - Proton
  ~1300 bunches/beam
  Protons/bunch
  10\text{ TeV (4x10^{12} eV)}
  Beam energy
  40 MHz
  Bunch collision frequency
  10 - 10^9 Hz
  Proton collision frequency

**Event selection:**
1 u 10 000 000 000 000
Higgs boson at LHC

Higgs boson (M_H ~125 GeV) produced every ~10 seconds @ L=5x10^{33} cm^{-2} s^{-1}

Higgs boson: decay channels

Decay channel | Mass region
--- | ---
H → γγ | 110-150
H → bb | 110-135
H → ττ | 110-140
H → WW → 2l 2ν | 110-600
H → ZZ → 4l | 110-600
H → ZZ → 2l 2τ | 180-600
H → ZZ → 2l 2j | 226-600
H → ZZ → 2l 2ν | 250-600

The most sensitive channels for low mass Higgs:

H → γγ
H → ZZ → 4l

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

The CMS Collaboration

Abstract

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at √s = 7 and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: gg, ZZ, WW, t+t, and b+b. An excess of events is observed above the expected background, with statistical significance of 5 standard deviations, at a mass near 125 GeV, signaling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The mass is well determined in the two decay modes with the best mass resolution: γγ and ZZ, a fit to these signals gives a mass of 125.3 ± 0.4 (stat.) ± 0.5 (syst.) GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

This paper is dedicated to the memory of our colleagues who worked on CMS but have since passed away.

In recognition of their many contributions to the achievement of this observation.

Submitted to Physics Letters B

1. Puljak: Data Analysis

CSC2019 19 – 20 September 2019, Cluj-Napoca
I. Puljak: Data Analysis

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb⁻¹ collected at √s = 7 TeV in 2011 and 5.8 fb⁻¹ at √s = 8 TeV in 2012. Individual searches in the channels H → ZZ → 4ℓ and H → WW → ℓ⁺ℓ⁻γγ in the 7 TeV data are combined with previously published results of searches for H → ZZ, W⁺W⁻, H → γγ, and H → τ⁺τ⁻ in the 8 TeV data. Clear evidence for the production of a neutral boson with a measured mass of 125.1 ± 0.8 (syst) GeV is presented. The observation, which has a significance of 5.9 standard deviations, corresponds to a background fluctuation probability of 1.7 × 10⁻⁵, is compatible with the production and decay of the Standard Model Higgs boson.

H → ZZ → ℓ⁺ℓ⁻ℓ⁺ℓ⁻ events distribution

Expectations vs measurements
I. Puljak: Data Analysis

H \rightarrow ZZ \rightarrow l^+l^-l^+l^- events distribution

Figure 3: The diphoton invariant mass distribution with each event weighted by the $s/(s+b)$ value of its category. The three separate data/mc likelihood fits are shown. The colored bands represent the 68% and 95% confidence level (CL) upper limits for the 125 GeV band. The inset shows the central part of the unweighted invariant mass distribution.

Figure 4: The distributions of the invariant mass of diphoton events after all selections for the combined 7 TeV and 8 TeV data sample. The tail of the sample is shown in (a) and a weighted version of the same sample in (c). The weighted histogram is superimposed. The residuals of the data and weighted data with respect to the expected background component are shown in (b) and (d).

Figure 8: The observed local significance for the combination of events in the (a) Higgs boson search [139] and the indirect constraints from the global fit to precision electroweak measurements [18] are shown. The bands indicate the 68% and 95% CL bands.

H \rightarrow b\bar{b}: example of Multivariate analysis (MVA)

For the multivariate analysis, a boosted decision tree (BDT) [115, 116] is trained to give a high output value (score) for signature events and a low value for events with good diphoton invariant mass resolution, based on the following observables: (i) the photon quality determined from electromagnetic shower shape and isolation variables; (ii) the expected mass resolution; (iii) the per-event estimate of the probability of locating the diphoton vertex within 10 mm of its true location along the beam direction; and (iv) kinematic characteristics of the photons and the diphoton system. The kinematic variables are constructed so as to contain no information about the invariant mass of the diphoton system. The diphoton events not satisfying the dip select-

Figure 11: Distribution of BDT scores in the high-$m_{ll}$ subchannel of the ZZ (Higgs) search. The data with best-fit nuisance parameter values have been applied. The signal expected from a Higgs boson ($m_h = 125$ GeV), including WZ/ZZ events where the charged lepton is not reconstructed, is also added to the background and also overlaid for comparison with the diboson background.

Figure 12: The 95% CL limits on the signal strength $\sigma/s_b$ for a Higgs boson decaying to two b quarks. The observed (solid) and expected (dashed) limits at 95% CL are shown. The local significance for the combination of events in the (a) Higgs boson search [139] and the indirect constraints from the global fit to precision electroweak measurements [18] are shown. The bands indicate the 68% and 95% CL bands.

Example of limits
p-value and hypothesis testing

![Image of a graph showing p-values as a function of the Higgs boson mass. The graph includes observed and expected values at different mass points.](image)

### Measuring properties

Asymptotically distributed as a χ² distribution with two degrees of freedom, the resulting 68% and 95% CL contours for the $H\to\gamma\gamma$ and $H\to WW^{(*)}\to 4\ell$ channels are shown in Figure 11. The observed (solid) $\sigma_{\mu}$ is a function of the SM Higgs boson mass $m_H$. The dashed line shows the expected local $\sigma_{\mu}$ under the hypothesis of a SM Higgs boson signal at that mass with its ασ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6 σ.

![Image of a graph showing the observed $\sigma_{\mu}$ versus the SM Higgs boson mass $m_H$. The graph includes observed and expected values at different mass points.](image)

### Conclusions of papers - ATLAS

#### M. Conclusion

Searches for the Standard Model Higgs boson have been performed in the $H\to ZZ^{(*)}\to 4\ell$, $H\to\gamma\gamma$ and $H\to WW^{(*)}\to 4\ell$ channels with the ATLAS experiment at the LHC using 5.8–5.9 fb$^{-1}$ of pp collisions recorded during April to June 2012 at a centre-of-mass energy of 8 TeV. These results are combined with earlier results [17], which are based on an integrated luminosity of 4.6–4.8 fb$^{-1}$ recorded in 2011 at a centre-of-mass energy of 7 TeV, except for the $H\to ZZ^{(*)}\to 4\ell$ and $H\to\gamma\gamma$ channels, which have been updated with the improved analysis presented here. The Standard Model Higgs boson is excluded at 95% CL in the mass range 111–359 GeV, except for the narrow region 122–135 GeV. In this region, an excess of events with significance 1.8 σ, corresponding to $σ_{\mu}=1.7\pm0.3$ fb$^{-1}$, is observed. The excess is driven by the two channels with the highest mass resolution, $H\to ZZ^{(*)}\to 4\ell$ and $H\to\gamma\gamma$, and is equally sensitive but low-resolution $H\to WW^{(*)}\to 4\ell$ channel. Taking into account the centre mass range of the search, 110–600 GeV, the global significance of the excess is 3.1 σ, which corresponds to $σ_{\mu}=1.7\pm0.3$ fb$^{-1}$. The results provide conclusive evidence for the discovery of a new particle with mass 126.0 ± 0.4 (stat) ± 0.8 (syst) GeV. The signal strength parameter $\mu$ has the value $1\pm0.3$ at the 95% CL, which is consistent with the SM Higgs boson hypothesis $\mu=1$. The selected events are consistent with pair production of vector bosons whose net electric charge is zero, identifying the new particle as a neutral boson. The observation in the diphoton channel, through the spin-1 hypothesis ([10], [11]), has been confirmed. Although these results are compatible with the hypothesis that the new particle is the Standard Model Higgs boson, more data are needed to access its nature in detail.
Conclusions of papers - CMS

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at \( \sqrt{s} = 7 \) and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb\(^{-1}\) at 7 TeV and 5.3 fb\(^{-1}\) at 8 TeV. The search is performed in five decay modes: \( \gamma\gamma \), ZZ, WW, \( \tau^+\tau^- \), and b\(\bar{b}\). An excess of events is observed above the expected background, with a local significance of 5.0\(\sigma\), at a mass near 125 GeV, signalling the production of a new particle. The expected local significance for a standard model Higgs boson of that mass is 3.8\(\sigma\). The global p-value in the search range of 115–130 (110–145) GeV corresponds to 4.6\(\sigma\) (4.5\(\sigma\)). The excess is most significant in the two decay modes with the best mass resolution, \( \gamma\gamma \) and ZZ, and a fit to these signals gives a mass of 125.3 ± 0.4 (stat.) ± 0.5 (syst.) GeV. The decay to two photons indicates that the new particle is a boson with spin different from one. The results presented here are consistent, within uncertainties, with expectations for a standard model Higgs boson. The collection of further data will enable a more rigorous test of this conclusion and an investigation of whether the properties of the new particle imply physics beyond the standard model.

We’ll come back to this at the end of lectures
Outline of Lecture Series

1. Introduction, Monte Carlo methods and distributions
2. Estimators and confidence intervals
3. Confidence intervals
4. Hypothesis testing

In this lecture

- **Introduction to data analysis**
  - Confirmatory and exploratory data analysis
  - Quantitative vs graphical techniques
  - Experimental vs observational studies
  - Exploring the data

Data analysis, statistics and probability

- **Data analysis** is the process of transforming raw data into usable information

  ![Diagram of data analysis process](RAW data → Data analysis → Usable information)

- Data analysis uses **statistics** for presentation and interpretation (explanation) of data
  - **Descriptive statistics**
    - Describes the main features of a collection of data in quantitative terms
  - **Inductive statistics**
    - Makes **inference** about a random process from its observed behavior during a finite period of time

- A mathematical foundation for statistics is the **probability theory**

Data Analysis

Lecture 1: Introduction to data analysis and Monte Carlo methods
Confirmatory and exploratory data analysis

**Confirmatory** data analysis = Statistical **hypothesis testing**
- A method of making statistical decisions using experimental data
- Two main methods
  - **Frequentist** hypothesis testing
    - Hypothesis is either true or not
  - **Bayesian** inference
    - Introduces a “degree of belief”

**Exploratory** data analysis
- Uses data to suggest hypothesis to test
- Complements confirmatory data analysis
- Main objectives:
  - Suggest hypothesis about the causes of observed phenomena
  - Assess assumptions on which statistical inference will be based
  - Select appropriate statistical tools and techniques
  - Eventually suggest further data collection

Quantitative vs graphical techniques

**Quantitative techniques** yield numeric or tabular output
- Hypothesis testing
- Analysis of variance
- Point estimation
- Interval estimation

**Graphical techniques**
- Used for gaining insight into data sets in terms of testing assumptions, model selection, estimator selection ...
- Provide a convincing mean of presenting results
- Includes: graphs, histograms, scatter plots, probability plots, residual plots, box plots, block plots, biplots
- Four main objectives:
  - Exploring the **content** of a data set
  - Finding **structure** in data
  - Checking **assumptions** in statistical models
  - Communicate the results of an analysis

Experimental vs observational studies

**Experimental studies**
- Measure the system → Manipulate the system → Measure again and compare
- Example: Study of whether and how much a free coffee would improve working performance of scientists in Building 40 at CERN

**Observational studies**
- No experimental manipulation
- Data are gathered and analysed
- Example:
  - Study of correlation between number of beers drunk in a pub on Wednesday evening on performance on the exam the day after
  - Be careful who pays! → see later
  - One could discuss whether to manipulate or not the system 😊

Experiments – basic steps

**Planning**
- Select subject to study
- Select an information source

**Design and Building**
- Design an experiment
- Build and test a model (f.g. MC simulation)
- Once happy with the model build the experiment

**Collecting data**
- Employ descriptive statistics to summarize data
- Suppress details
- Early exploratory analysis

**Analysing data**
- Statistical inference
- Reach a consensus what observations tell about an underlaying reality

**Presenting Documenting**
- Publish article and disseminate results
- Enjoy in the fruits of the hard work!
LHC experiments – basic steps

Planning
• Started ~ 30 years ago (Aachen 1989)
• Core teams from previous experiments UA1&2

Design Building
• ‘Best’ experimental design chosen (CMS, ATLAS, ALICE and LHCb)
• Detailed MC simulations performed before started to build

Collecting data
• Trigger and DAQ carefully planed and built
• MC simulation used for optimization

Analysing data
• Statistical inference → a part of work done at this school too (learning methods&tools)
• For the consensus → let’s see

Presenting Documenting
• Many articles published
• And first discoveries announced and published!

What we (will) measure at LHC?

Something we already know
• At the very beginning of the LHC operation
• For example: production of W and Z bosons

Something that (probably) exists but wasn’t measured yet
• Simply because we are exploring new energy domain
• Standard Model processes
• But surprises are always possible

Hopefully something new but reasonably expected
• Although “reasonably” is not very well defined ☺
• For example we all expected to find the Higgs boson → and we did find it!
• Heavy neutrinos?

Maybe something new but less likely
• New heavy bosons (Z’, W’)
• Micro black holes
• Extra dimensions

Something completely unexpected
• Well, it’s hard to look for unexpected ☹

Some of the physicists’ jargon

Cross section ($\sigma$)
• A measure of ‘frequency’ of the physical process
• Units: barns ($10^{-28}$ cm$^2$)
  • Typical values: femtobarns (fb), picobarns (pb)

Luminosity (L)
• Or instantenous luminosity
• A measure of collisions ‘frequency’
  • Typical (at Tevatron/Early LHC): $L = 10^{32}$ cm$^{-2}$s$^{-1}$

Integrated luminosity ($\mathcal{L} = \int L dt$)
• A measure of number of accumulated collisions after a certain time period
• Units: (cross section)$^{-1}$ ...
  • E.g. 1 fb$^{-1}$ = 1000 pb$^{-1}$
  • Typical (Tevatron/Early LHC): few fb$^{-1}$

Number of events ($N$)
• Number of (expected) events ($N$) after a certain time of running
  \[ N = \sigma \cdot \mathcal{L} \]
Measuring physical objects

Data analysis - general picture

Events collected after some time of LHC running

N ~ 100/s x 10^7 s/year
N ~ 10^9 events per year

Objects = reconstructed objects i.e. electrons, photons, jets, muons...

Analysis steps in typical LHC analysis

1. Event reconstruction
2. Event selection
3. Background estimation
4. Systematic uncertainties
5. Yields and kinematics distributions
6. Kinematic discriminant
7. Statistical analysis and results

For example, let's suppose the TRUE state of nature is:
Higgs boson exists with the mass of m_H(true) = 134.26 GeV

Object 1 == electron
p_x
p_y
p_z
E
**Signal vs background(s)**

- **Signal**: an event coming from the physical process under study  
  - Example: $H \rightarrow ZZ \rightarrow e^+e^-e^+e^-$ (henceforth both $e^+$ and $e^-$ are 'electron')

- **Background**: any other event  
  - 'Dangerous' background is any other process giving at least 4 electrons in the final state  
    - But be careful: electrons seen by detector are reconstructed objects and in some cases when some other objects (e.g., jets) are misreconstructed as electrons  
  - 'Trivial' backgrounds are all other backgrounds and are easily rejected by a simple requirement of having at least 4 electrons in the final state

**Signal: $pp \rightarrow H \rightarrow ZZ \rightarrow 4e$**  
**'Dangerous' background: $pp \rightarrow ZZ \rightarrow 4e$**

---

**Exploring the data**

- Once data are collected $\rightarrow$ exploratory data analysis  
  - Heavily use of graphical techniques  

- Example: **data reduction $=$ Preselection**  
  - Goal: **getting rid of all unuseful events**  
  - Unusefulness is not uniquely defined:  
    - We have a certain interest to keep some background events for better control and its measurement from data  
  - Some numbers:  
    - $\sim 10^9$ events collected per year (after trigger)  
    - $\sim 1$ MB event size on a tape (rought estimate)  
    - $\Rightarrow \sim 1$ PB of data collected per year $\rightarrow$ non manageable at once  
  - Interested physical processes are rare  
    - F.g. just a handful ($\sim 10$) $H \rightarrow ZZ \rightarrow 4e$ events per year  
  - So be careful when choosing criteria for data reduction not to lose too many signal events
Example: $H \rightarrow ZZ \rightarrow 4e$ in CMS

**Very basic cuts:** High Level Trigger + $\geq 3$ electrons, any charge and $p_T^{1,2,3} > 10, 10, 5$ GeV/c

**Preselection cuts:**
- $\geq 2$ ee pairs of identified, opposite charge and same flavor leptons with $p_T > 5$ GeV/c; $|\eta| < 2.5$
- At least two $m_{ee} > 12$ GeV/c$^2$
- At least one $m_{ee} > 100$ GeV/c$^2$
- Loose track based isolation

**After these steps**
- Some background gone
- Some heavily reduced
- Some still resisting

**Full selection** needed for the final analysis

---

**Probability – basic concepts**

- **Definitions of probability**
  - **Mathematical probability**
    - Probability is a basic and an abstract concept
  - **Frequentist probability**
    - Using only measured frequencies
  - **Bayesian probability**
    - Based on a degree of belief
Mathematical probability

Developed in 1933 by Kolmogorov in his “Foundations of the Theory of Probability”

Define $\Omega$ as an exclusive set of all possible elementary events $x_i$.

Exclusive means the occurrence of one of them implies that none of the others occurs.

We define the probability of the occurrence of $x_i$, $P(x_i)$, to obey the **Kolmogorov axioms**:

1. $P(x_i) \geq 0$ for all $i$.
2. $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$.
3. $\sum P(x_i) = 1$.

From these properties more complex probability expressions can be deduced.

- For non-elementary events, i.e., set of elementary events.
- For non-exclusive events, i.e., sets of overlapping events.

Bayesian probability

Based on a concept of “degree of belief”

An operational definition of belief is based on coherent bets by Finetti.

- **What’s amount of money one’s willing to bet based on her/his belief on the future occurrence of the event?**

Bayesian inference uses Bayes’ formula for conditional probability:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

$H$ is a **hypothesis**, and $D$ is the data.

$P(H)$ is the **prior probability** of $H$: the probability that $H$ is correct before the data $D$ was seen.

$P(D \mid H)$ is the **conditional probability** of seeing the data $D$ given that the hypothesis $H$ is true. $P(D)$ is called the **likelihood**.

$P(D)$ is the **marginal probability** of $D$.

- $P(D)$ is the prior probability of witnessing the data $D$ under all possible hypotheses.

$P(H \mid D)$ is the **posterior probability**: the probability that the hypothesis is true, given the data and the previous state of belief about the hypothesis.

Frequentist probability

**Experiment:**
- $N$ events observed
- Out of them $n$ is of type $x$.

**Frequentist probability** that any single event will be of type $x$:

$$P(x) = \lim_{N \to \infty} \frac{n}{N}$$

Important restriction: such a probability can only be applied to repeatable experiments.

- For example one can’t define a probability that it’ll snow tomorrow.
- Although this seems to be a serious problem, a job of scientist is to try to get as close as possible to repeatable experiments and produce reproducible results.

Frequentist statistics is often associated with the names of *Jerzy Neyman and Egon Pearson*.

Example: Who will pay the next round?

You meet an old fried at Gottingen in a pub. He proposes that the next round should be payed by whichever of the two extracts the card of lower value from a pack of cards.

This situation happens many times in the following days. What is the probability that your friend cheats if you end up paying wins consecutive times?²

You assume:

- $P(\text{cheat}) = 5\%$ and $P(\text{honest}) = 95\%$. (Surely an old friend is an unlikely cheat …)
- $P(\text{wins|cheat}) = 1$ and $P(\text{wins|honest}) = 2^{-\text{wins}}$

Bayesian solution:

\[
P(\text{cheat|wins}) = \frac{P(\text{wins|cheat})P(\text{cheat})}{P(\text{wins|cheat})P(\text{cheat}) + P(\text{wins|honest})P(\text{honest})}
\]

\[
P(\text{cheat|0}) = \frac{1P(\text{cheat})}{1P(\text{cheat}) + 2^{-0}P(\text{honest})} = \frac{0.05}{0.05 + 0.95} = 5\%
\]

\[
P(\text{cheat|5}) = \frac{1P(\text{cheat})}{1P(\text{cheat}) + 2^{-5}P(\text{honest})} = \frac{0.05}{0.05 + 0.03} = 63\%
\]

Example: Learning by experience

The process of updating the probability when new experimental data becomes available can be followed easily if we insert

\[ P(\text{cheat}) = P(\text{cheat} | \text{wins} = 1) \]

and

\[ P(\text{honest}) = P(\text{honest} | \text{wins} = 1), \]

where \( \text{wins} = 1 \) indicate the probability assigned after the previous win.

Iterative application of the Bayes formula for \( P(\text{cheat} | \text{wins} = 1) \):

\[
P(\text{cheat} | \text{wins} = 1) = \frac{P(\text{cheat} | \text{wins} = 1) P(\text{win} | \text{cheat}) + \frac{1}{2} P(\text{win} | \text{honest}) P(\text{honest} | \text{wins} = 1)}{P(\text{win} | \text{cheat}) P(\text{cheat} | \text{wins} = 1) + P(\text{win} | \text{honest}) P(\text{honest} | \text{wins} = 1)}
\]

When you learn from the experience, your conclusions no longer depend on the initial assumptions.

Example: Priors and posteriors – expressing degree of belief

Phil is learning from experience:

(From discussion of climate change on Andrew Gelman’s blog.)

Random variables

- **Random event**: event having more than one possible outcome
  - Each outcome may have associated probability
  - Outcome not predictable, only the probabilities known

- Different possible outcomes may take different possible numerical values \( x_1, x_2, ... \) → **random variable** \( x \)
  - The corresponding probabilities \( P(x_1), P(x_2), ... \) form a **probability distribution**

- If observations are **independent** the distribution of each random variable is unaffected by knowledge of any other observation

- When an experiment consists of \( N \) repeated observations of the same random variable \( x \), this can be considered as the single observation of a random vector \( x \), with components \( x_1, ..., x_N \)

Random variables: discrete

- **Rolling a die**:
  - Sample space = \{1,2,3,4,5,6\}
  - Random variable \( x \) is the number rolled
    - 1 if a 1 is rolled
    - 2 if a 2 is rolled
    - 3 if a 3 is rolled
    - 4 if a 4 is rolled
    - 5 if a 5 is rolled
    - 6 if a 6 is rolled

- Discrete probability distribution
  - \( p(x) = \frac{1}{6} \) for \( x = 1, 2, 3, 4, 5, 6 \)
Random variables: continuous

- A spinner
  - Can choose a real number from $[0, 2n]$
  - All values equally likely
  - $X =$ the number spun
  - Probability to select any real number $= 0$
  - Probability to select any range of values $> 0$
    - Probability to choose a number in $[0, n] = \frac{1}{2}$
  - Now we say that probability density $p(x)$ of $x$ is $\frac{1}{2n}$
  - Probability to select a number from any range $D$ is $\frac{D}{2n}$
  - More general

$P(A < x < B) = \int_A^B p(x)dx$

Probability density function

- Let $x$ be a possible outcome of an observation and can take any value from a continuous range
- We write $f(x; \theta)dx$ as the probability that the measurement’s outcome lies between $x$ and $x + dx$
- The function $f(x; \theta)dx$ is called the probability density function (PDF)
  - And may depend on one or more parameters $\theta$
- If $f(x; \theta)$ can take only discrete values then $f(x; \theta)$ is itself a probability
- The p.d.f. is always normalized to unit area (unit sum, if discrete)
- Both $x$ and $\theta$ may have multiple components and then written as vectors
- If $\theta$ is unknown we may wish to estimate its value from a set of measurements of $x$ → Parameter estimation in Lecture 3

Cumulative and marginal distributions

- Cumulative distribution function, CDF
  - For every real number $Y$, the CDF of $Y$ is equal to the probability that the random variable $x$ takes a value less or equal to $Y$
    $$F(Y) = P(x \leq Y) = \int f(x)dx$$
  - If $x$ restricted to $x_{min} < x < x_{max}$ then $F(x_{min}) = 0$, $F(x_{max}) = 1$
  - $F(x)$ is a monotonic function of $x$

- Marginal density function
  - Is the projection of multidimensional density
  - Example: if $f(x, y)$ is two-dimensional PDF the marginal density $g(x)$ is
    $$g(x) = \int_{y_{min}}^{y_{max}} f(x, y)dy$$

Introduction to Monte Carlo method

- Monte Carlo techniques
- Monte Carlo in HEP
I. Puljak: Data Analysis

**Monte Carlo**

**Why simulations?**

- **Design studies**
  - optimize detectors before building them
  - estimation of performances, costs...

- **Development of reconstruction algorithms**
  - exploring different algorithms,
  - tuning parameters
  - optimizing analysis

- Simulation is a good way to save money!

- But, there is even a “deeper” reason!

---

**Example: classical mechanics**

Galilei Brahe observations Kepler

Newton: mechanics and gravitation

\[ F = \frac{GMm}{r^2} \]

From M. Lindf, Experiment Simulation – CSC 2009

---
Monte Carlo method

Monte Carlo methods (MCMs) are a class of computational algorithms that rely on repeated random sampling to compute their results:
- MCMs use random or pseudo-random numbers.
- MCMs tend to be used when it is unfeasible or impossible to compute an exact result with a deterministic algorithm.

The term "Monte Carlo method" was coined in the 1940s by physicists working on nuclear weapon projects in the Los Alamos National Laboratory.

Generally MCMs are used in:
- Studying systems with a large number of coupled (interacting) degrees of freedom, such as fluids, disordered materials, strongly coupled solids, and cellular structures.
- Modeling phenomena with significant uncertainty in inputs, such as the calculation of risk in business.
- Evaluation of definite integrals, particularly multidimensional integrals with complicated boundary conditions.

Example 1: estimating π

1. Define a domain of possible inputs
   - Draw a square on the ground, then inscribe a circle within it.

2. Generate inputs randomly from the domain
   - Uniformly scatter N objects of uniform size throughout the square.

3. Perform a deterministic computation using the inputs
   - Count number of objects in the circle = n

4. Aggregate the results of the individual computations into the final result
   - Estimate π as $\pi \sim 4 \times n / N$

Example 2: integration

Analytical solution

$$f(x) = x$$

$$\int_a^b dx = 1/2 \cdot x^2$$

$$A = 1/2 \cdot (b^2 - a^2)$$

Deterministic algorithm

$$f(x) = x$$

$$n \text{ number of steps}$$

$$\Delta x = (b - a) / n$$

$$A \sim \sum (f(a+i \Delta x) \cdot \Delta x)$$
Monte Carlo solution

Random number generation

- Physical methods
  - "true" random numbers from "unpredictable" process
    - Example: dice, coin flopping, roulette
  - TRUE random numbers from random atomic or subatomic physical phenomena
    - Example: radioactive decay, amplitude of noise in radio

- Computational methods
  - Pseudo-random number generators create long runs (for example, millions of numbers long) with good random properties but eventually the sequence repeats
    - Example: Linear congruential generator \( X_{n+1} = (aX_n + b) \mod m \)

- Generation from a probability distribution \( f(x) \)
  - Generate random numbers distributed according to the \( f(x) \)
  - Method involves transforming an uniform random number in some way
  - Examples: inversion method, acceptance-rejection method

Inversion method

Let \( x \) be a random variable whose distribution can be described by the cumulative distribution function \( F(x) \).

We want to generate values of \( x \) which are distributed according to this distribution.

Method:

1. Generate a random number from the standard uniform distribution; call this \( u \).
2. Compute the value \( x \) such that \( F(x) = u \); call this \( x_{\text{chosen}} \).
3. Take \( x_{\text{chosen}} \) to be the random number drawn from the distribution described by \( F \).

Acceptance-rejection method

It generates sampling values from an arbitrary PDF function \( f(x) \) by using an instrumental distribution \( h(x) \) for which we know how to sample under the only restriction that \( f(x) < Ch(x) \) where \( C > 1 \).

Usually used in cases where the form of \( f(x) \) makes sampling difficult.

Algorithm:

1. Sample \( x \) from \( h(x) \) and \( u \) from \( U(0,1) \)
2. Check whether or not \( u < f(x) / Ch(x) \).
3. Accept \( x \) as a realization of \( f(x) \).
MC methods in engineering

**Wireless network planning**
- Various scenarios depending on: number of users, users' location, services users want to use
- MC used to generate users and their states, so that network performance can be evaluated and optimized

**Computer graphics**
- MC methods efficient in production of photorealistic images of virtual 3D models
- Application in video games, computer generated films, special effects in cinema

**Wind power engineering**
- From measured distributions of wind speeds MC generates single values for wind power system performance evaluation and optimization

Monte Carlo in HEP

**Monte Carlo methods are widely used in High Energy Physics**

**Theoretical calculations**
- Total cross sections, differential cross sections distributions ...

**Event generation**
- Distribute events according to expected probabilities
- Many event generators on the market: e.g. PYTHIA, HERWIG

**Detector simulation**
- Passage of particles through the matter
- GEANT

**Data analysis**
- Background predictions (if not measured from data)
- Signal predictions
- Final results

---

Monte Carlo in HEP

<table>
<thead>
<tr>
<th>Simulation</th>
<th>'Real life'</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event Generation</strong></td>
<td><strong>Collisions</strong></td>
</tr>
<tr>
<td>Tools: MC generators (PYTHIA, ...)</td>
<td>Tools: Accelerator (LHC, Tevatron, ...)</td>
</tr>
<tr>
<td>Output: final state particles</td>
<td>Output: final state particles</td>
</tr>
<tr>
<td><strong>Detector simulation</strong></td>
<td><strong>Data acquisition</strong></td>
</tr>
<tr>
<td>Tools: MC simulators (GEANT)</td>
<td>Tools: Detectors (CMS, ATLAS, ...)</td>
</tr>
<tr>
<td>Output: simulated detector response</td>
<td>Output: detector response</td>
</tr>
<tr>
<td><strong>Event reconstruction</strong></td>
<td></td>
</tr>
<tr>
<td>Tools: Detectors' software packages (custom made; MC used in algorithms)</td>
<td></td>
</tr>
<tr>
<td>Output: reconstructed physical objects (electrons, muons, jets ...)</td>
<td></td>
</tr>
<tr>
<td><strong>Data analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Tools: Statistics (ROOT, ..., MC used in algorithms; e.g. Toy MC)</td>
<td></td>
</tr>
<tr>
<td>Output: new knowledge (parameter/interval estimates, hypothesis tests, article, talks ...)</td>
<td></td>
</tr>
</tbody>
</table>