

Υπολογιστική Φυσική Στοιχειωδών Σωματιδίων

Πρόγραμμα Μεταπτυχιακών Σπουδών Υπολογιστικής Φυσικής
4ο Εξάμηνο 2022-23

Στατιστική Πιθανότητες - Κατανομές

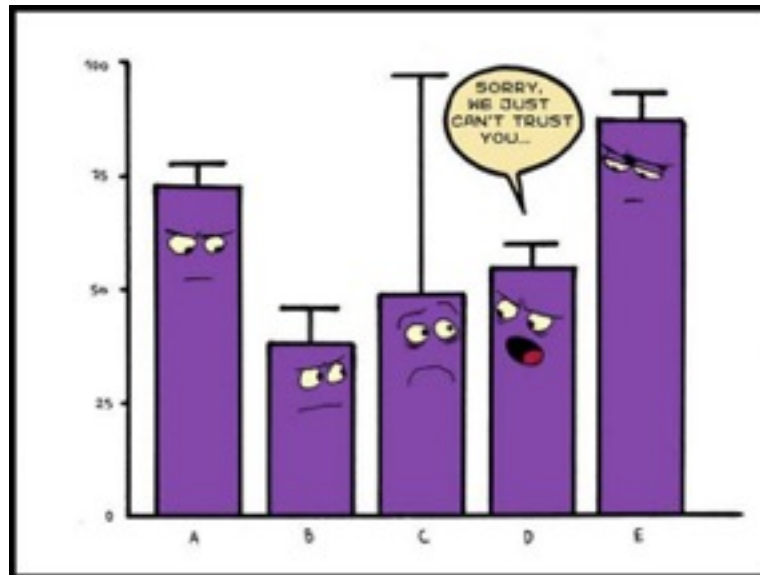
Δήμος Σαμψωνίδης

Γιατί κάνουμε στατιστική ?

- Θέλουμε να προσδιορίσουμε ένα μέγεθος (**parameter estimation**) π.χ. το πλάτος Γ_h ή την μάζα του Higgs: 124.98 ± 0.28 ($\pm 0.19 \pm 0.21$) GeV
- Ελέγχουμε το κατά πόσο τα δεδομένα μας βρίσκονται σε συμφωνία με την θεωρία (**goodness of fit**)
- Ελέγχουμε αν τα δεδομένα έχουν 'προτίμηση' προς την θεωρία 1 ή την
- θεωρία 2 (**hypothesis testing**)
- Σχεδιάζουμε το επόμενο πείραμα (**decision making**)

Γιατί κάνουμε στατιστική ?

- Θέλουμε να προσδιορίσουμε ένα μέγεθος (**parameter estimation**) π.χ. το πλάτος Γ_h ή την μάζα του Higgs: 124.98 ± 0.28 ($\pm 0.19 \pm 0.21$) GeV
- Ελέγχουμε το κατά πόσο τα δεδομένα μας βρίσκονται σε συμφωνία με την θεωρία (**goodness of fit**)
- Ελέγχουμε αν τα δεδομένα έχουν 'προτίμηση' προς την θεωρία 1 ή την θεωρία 2 (**hypothesis testing**)
- Σχεδιάζουμε το επόμενο πείραμα (**decision making**)



Βασική διάκριση: Περιγραφική και Επαγωγική Στατιστική

Descriptive (Περιγραφική) Statistics

- ασχολείται με τη συνοπτική και αποτελεσματική παρουσίαση των δεδομένων
- χρησιμοποιεί αριθμητικούς δείκτες όπως η μέση τιμή, η διάμεσος, η διακύμανση κλπ. (μέτρα θέσης και διασποράς)
- όταν χρησιμοποιείται για την περιγραφή ενός δείγματος, τα μέτρα αφορούν μόνο αυτό και όχι άλλα δείγματα ή ολόκληρο τον πληθυσμό

Βασική διάκριση: Περιγραφική και Επαγωγική Στατιστική

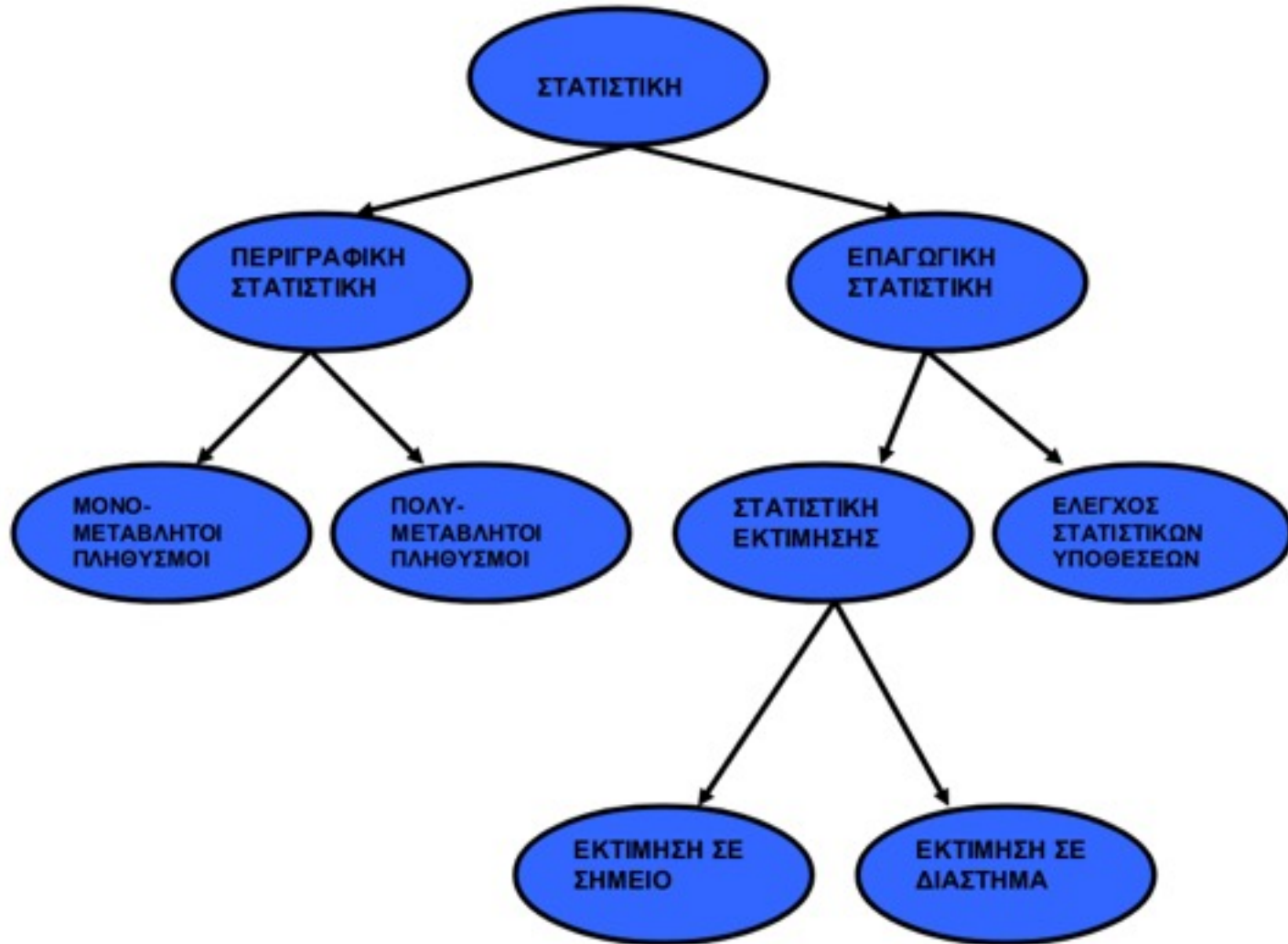
Descriptive (Περιγραφική) Statistics

- ασχολείται με τη συνοπτική και αποτελεσματική παρουσίαση των δεδομένων
- χρησιμοποιεί αριθμητικούς δείκτες όπως η μέση τιμή, η διάμεσος, η διακύμανση κλπ. (μέτρα θέσης και διασποράς)
- όταν χρησιμοποιείται για την περιγραφή ενός δείγματος, τα μέτρα αφορούν μόνο αυτό και όχι άλλα δείγματα ή ολόκληρο τον πληθυσμό

Inferential (Επαγωγική) Statistics

- στοχεύει στην εξαγωγή συμπερασμάτων και την πραγματοποίηση προβλέψεων για τον πληθυσμό, μελετώντας ένα τυχαίο δείγμα του
- πραγματοποιούνται εκτιμήσεις για τα χαρακτηριστικά του πληθυσμού (εκτιμητική) και έλεγχοι στατιστικών υποθέσεων για την ορθότητα μιας υπόθεσης σχετικά με τον πληθυσμό

Βασική διάκριση: Περιγραφική και Επαγωγική Στατιστική



Ο ρόλος της Στατιστικής στα Πειράματα Φυσικής Υψηλών Ενέργειών

Theory $\leftrightarrow$ Statistics $\leftrightarrow$ Experiment

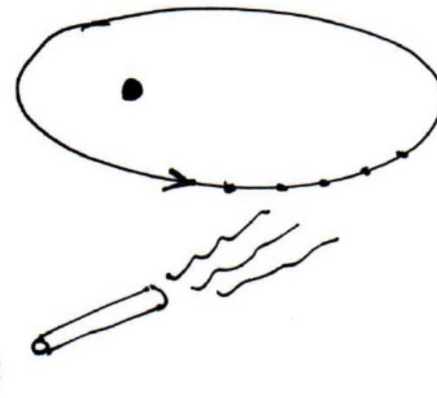
Theory (model, hypothesis):

$$F = -G \frac{m_1 m_2}{r^2}, \dots$$

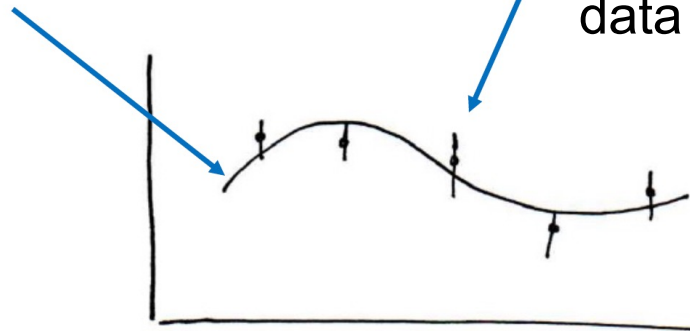
+ simulation of detector and cuts

= model prediction

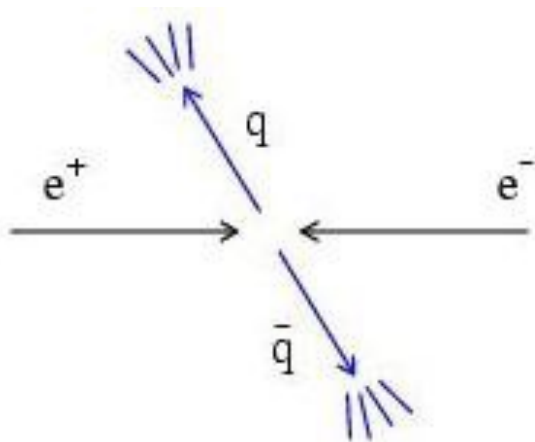
Experiment (observation):



data



Data Analysis in Particle Physics



Παρατήρηση “γεγονότων” συγκεκριμένων τύπων

- Μετράμε τα χαρακτηριστικά κάθε γεγονότος (particle momenta, number of muons, energy of jets,...)
- Οι θεωρίες (π.χ. SM) προβλέπουν κατανομές αυτών των χαρακτηριστικών με ελεύθερες παραμέτρους, π.χ., α , G_F , M_Z , α_s , m_H , ...
- Μερικές εργασίες (tasks) ανάλυσης δεδομένων:
 - ✓ Εκτίμηση (μέτρηση) των παραμέτρων.
 - ✓ Ποσοτικοποίηση της αβεβαιότητας των εκτιμήσεων των παραμέτρων.
 - ✓ Έλεγχος σε ποιο βαθμό στον οποίο οι προβλέψεις μιας θεωρίας συμφωνούν με τα δεδομένα.

Data Analysis in Particle Physics

Η **αβεβαιότητα** μπαίνει σε πολλά επίπεδα

Οι μετρήσεις γενικά δεν είναι απόλυτα επαναλήψιμες (με ακρίβεια)

- *Κβαντικά φαινόμενα*
- *Τυχαία φαινόμενα (ακόμη και χωρίς QM)*

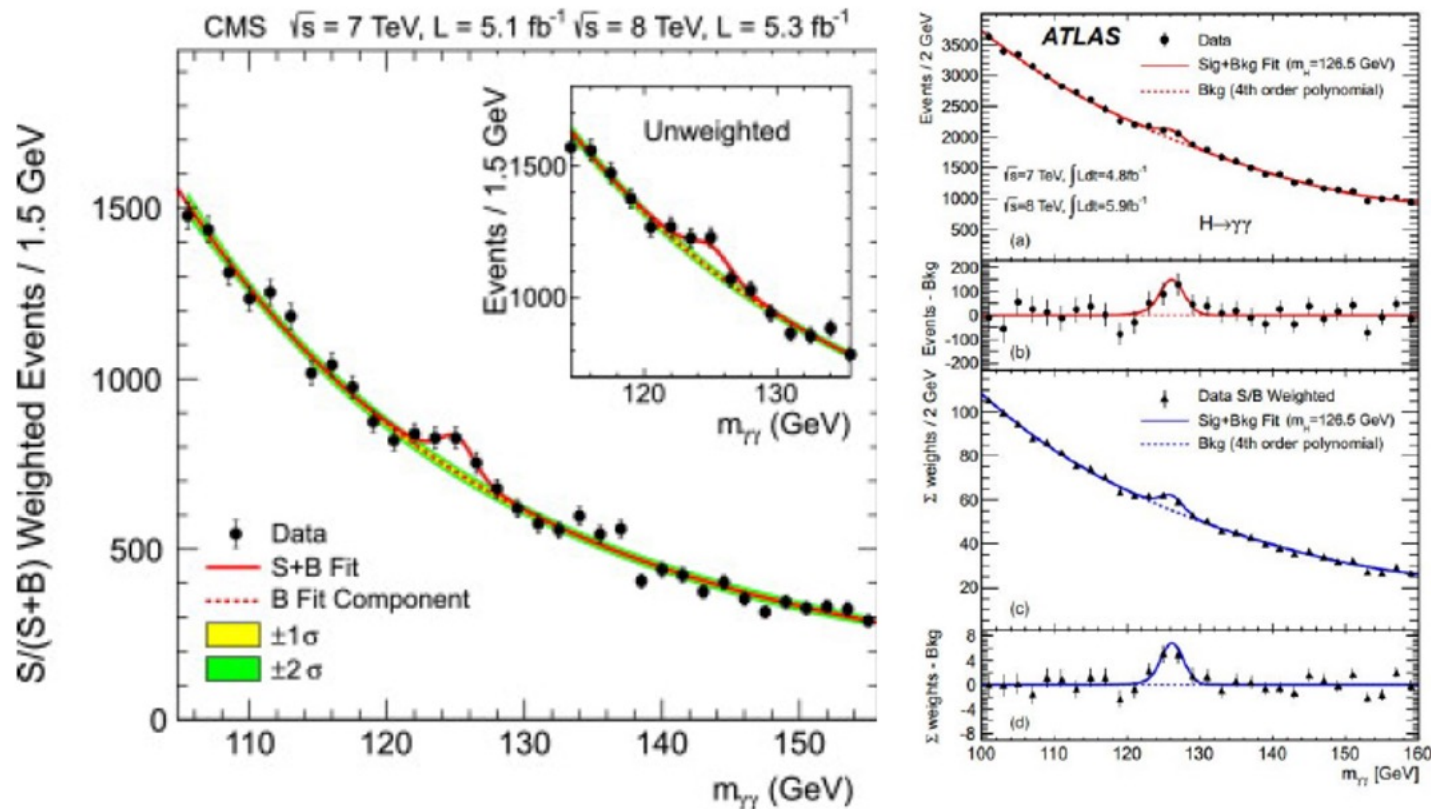
Αβεβαιότητα της πρόβλεψης μοντέλου

- *Προσεγγίσεις που χρησιμοποιούνται για την εξαγωγή θεωρητικής πρόβλεψης*
- *Μοντελοποίηση του ανιχνευτή*

Ποσοτικοποιούμε την αβεβαιότητα χρησιμοποιώντας τις **ΠΙΘΑΝΟΤΗΤΕΣ**



Η ανακάλυψη του Higgs: απόρροια σύνθετης υπολογιστικής διαδικασίας!

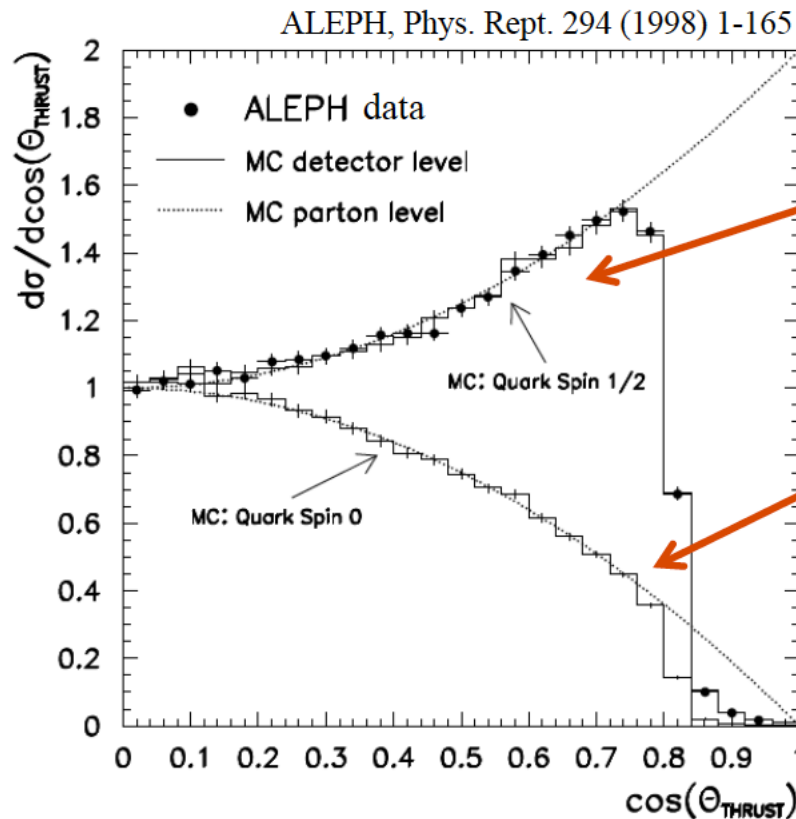


The dots and the error bars represent the observed number of events in a given mass bin. The numbers come from a long chain of data analysis beginning with the trigger event selection, followed by the event reconstruction, the event filtering and analysis for the millions of events recorded by the experiments

The data simulation and analysis is performed on a distributed data grid system called the WLCG (Worldwide LHC Computing Grid). This infrastructure based on tiers computing centres has been installed all around the world to absorb the heavy computing load of the high LHC statistics. The tier 0 is located at the data production centre (CERN) and absorbs 20% of the load, 11 tier-1 connected through direct high bandwidth fibre links (10 Gb/s) are scattered across the main participating countries/regions and ~ 140 tier-2 are providing the specific computing support necessary to research teams. The WLCG provides more than 250,000 processors, and close to 150 PB of disk storage from over 150 sites in 34 countries, producing a massive distributed computing infrastructure that support the needs of more than 8,000 physicists.

Data Analysis in Particle Physics: testing hypotheses

Test the extent to which a given model agrees with the data:

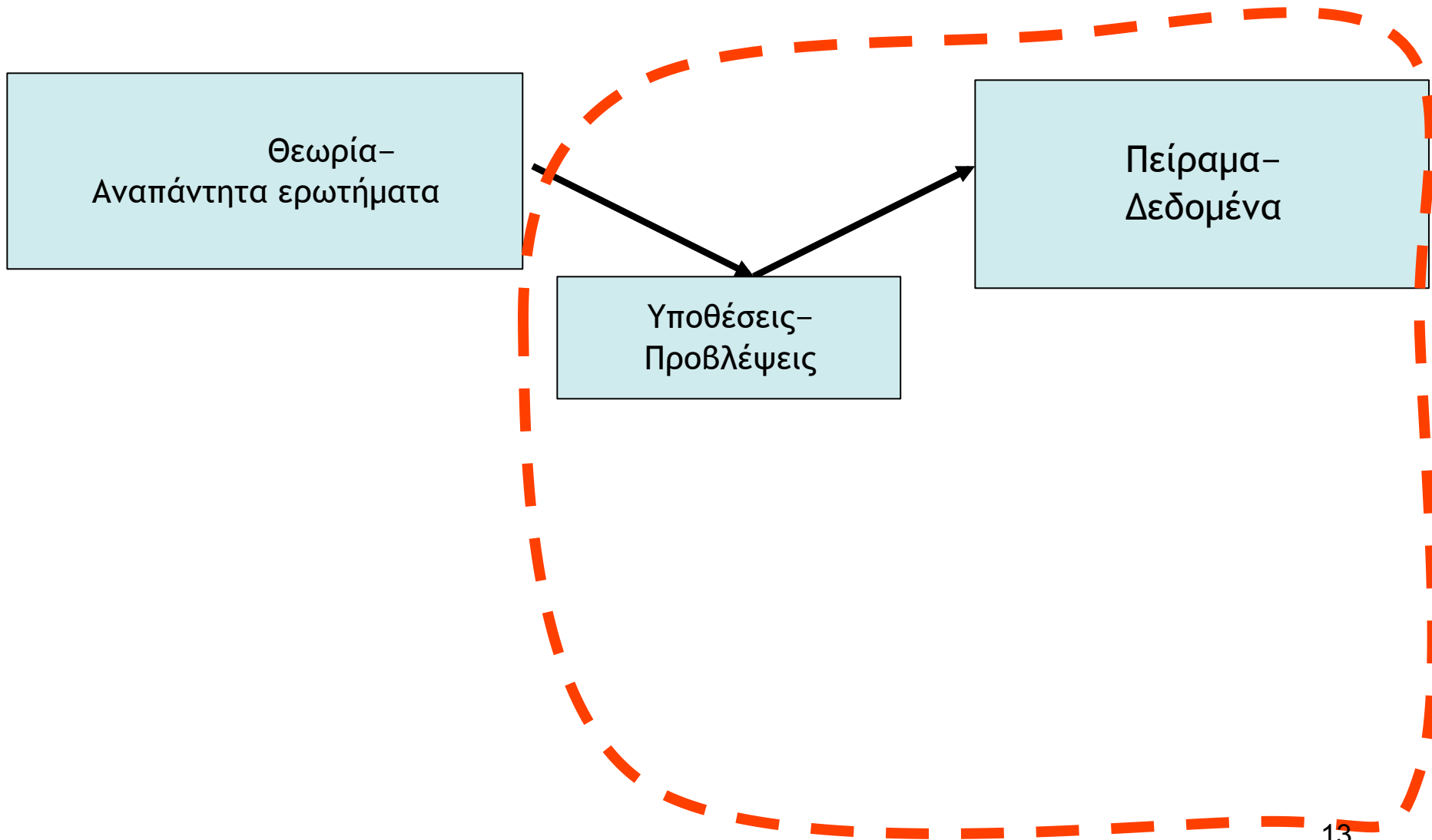


spin-1/2 quark
model “good”

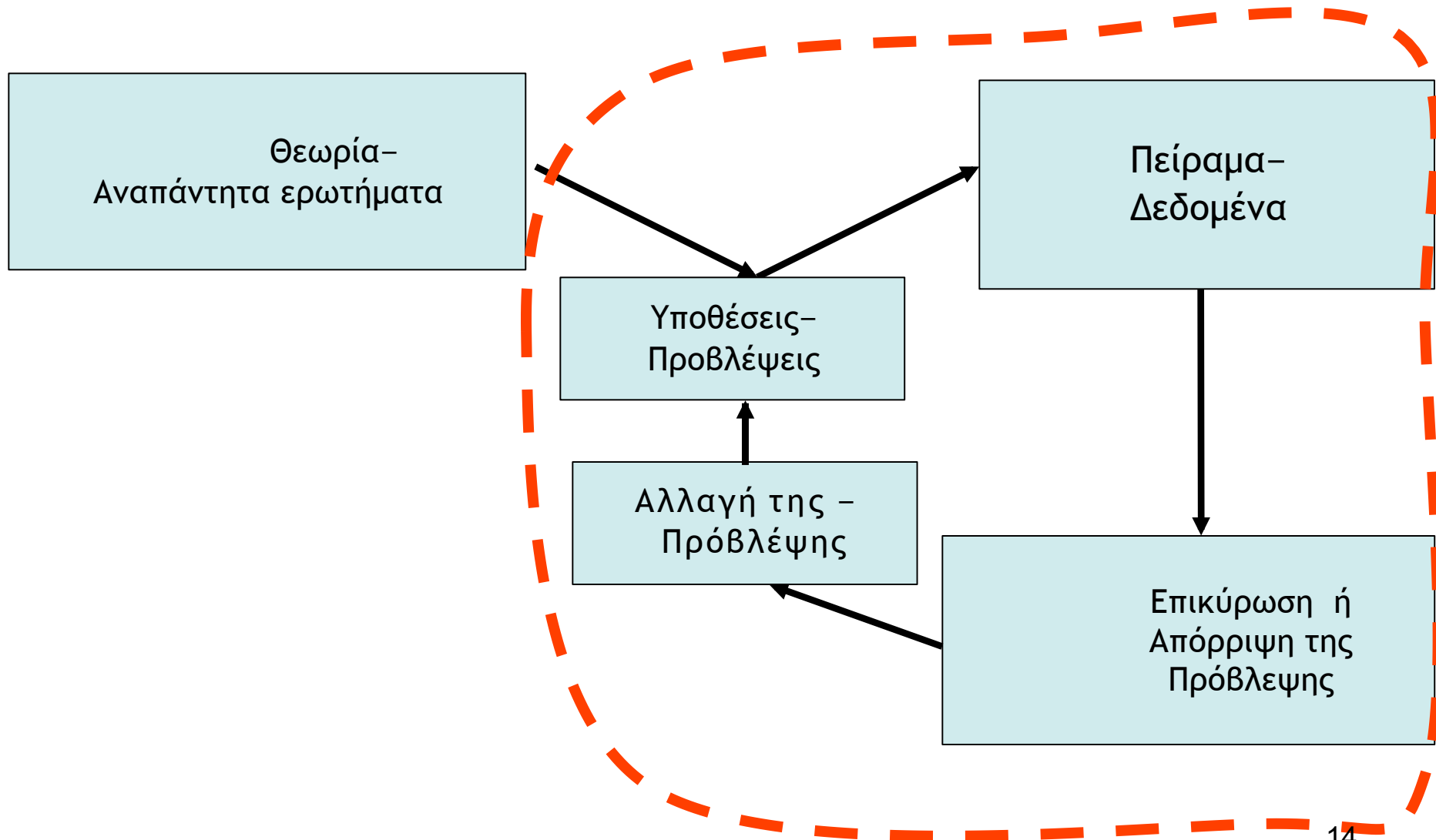
spin-0 quark
model “bad”

In general need tests
with well-defined properties
and quantitative results.

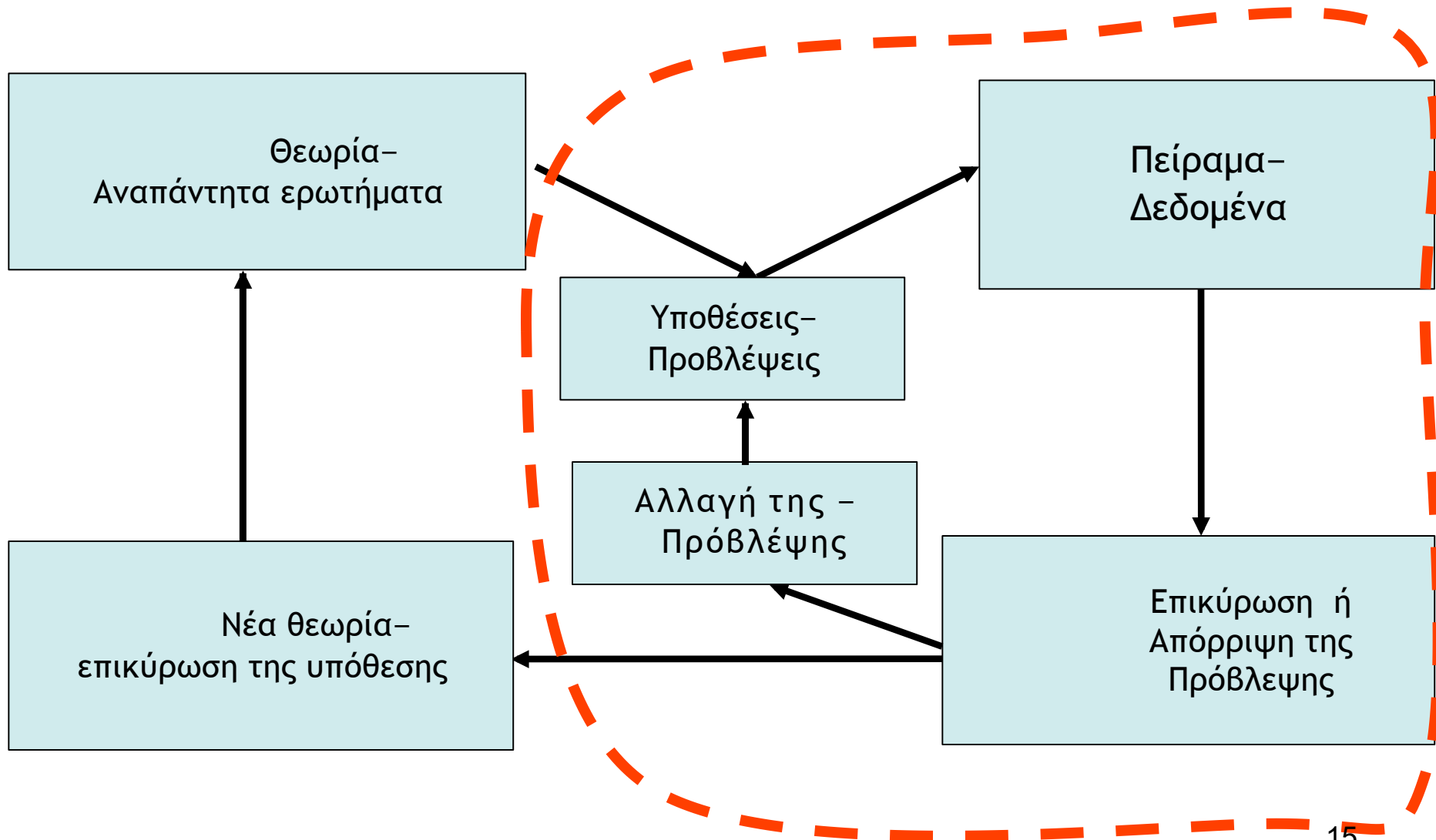
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων I



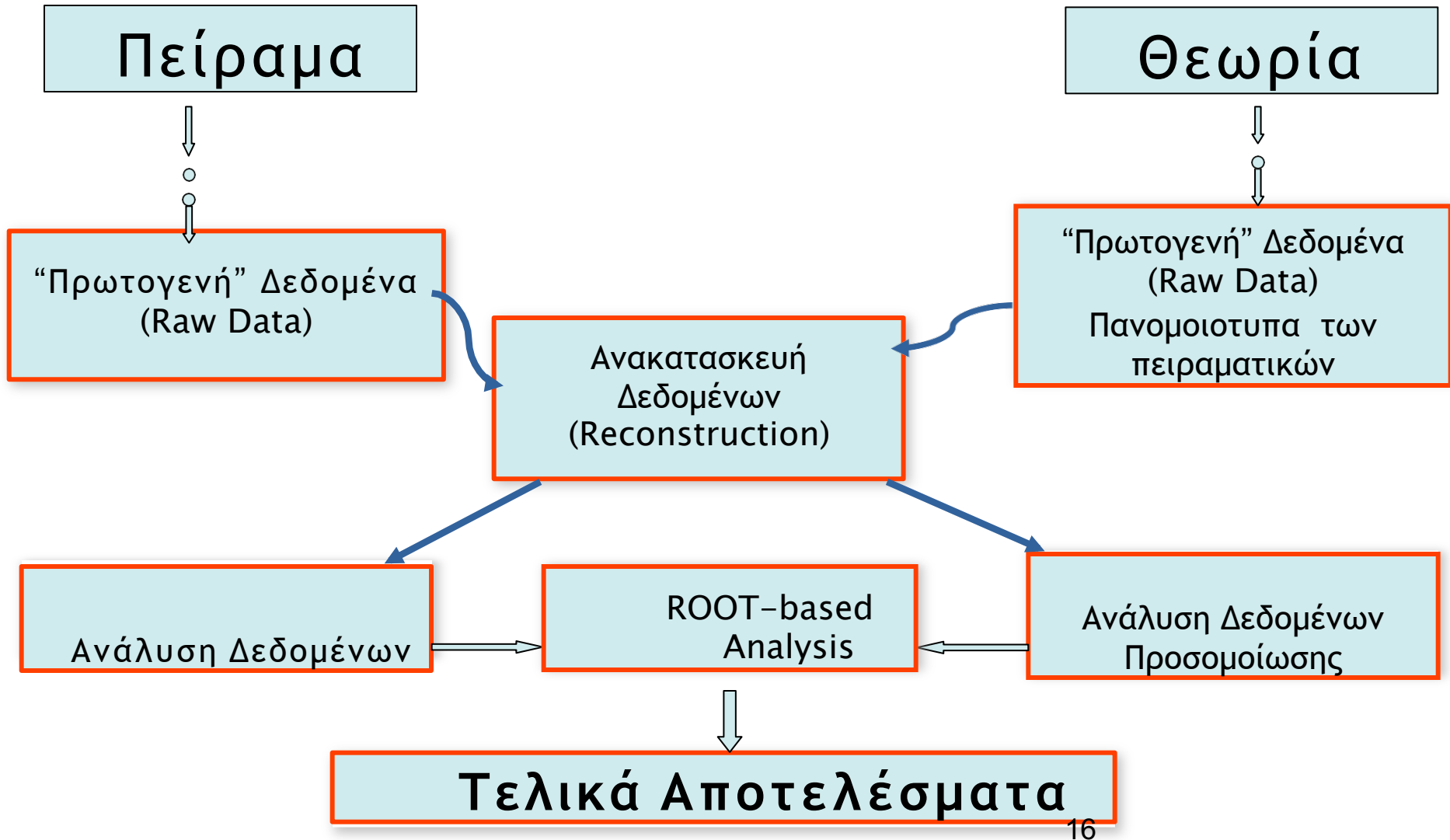
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων I



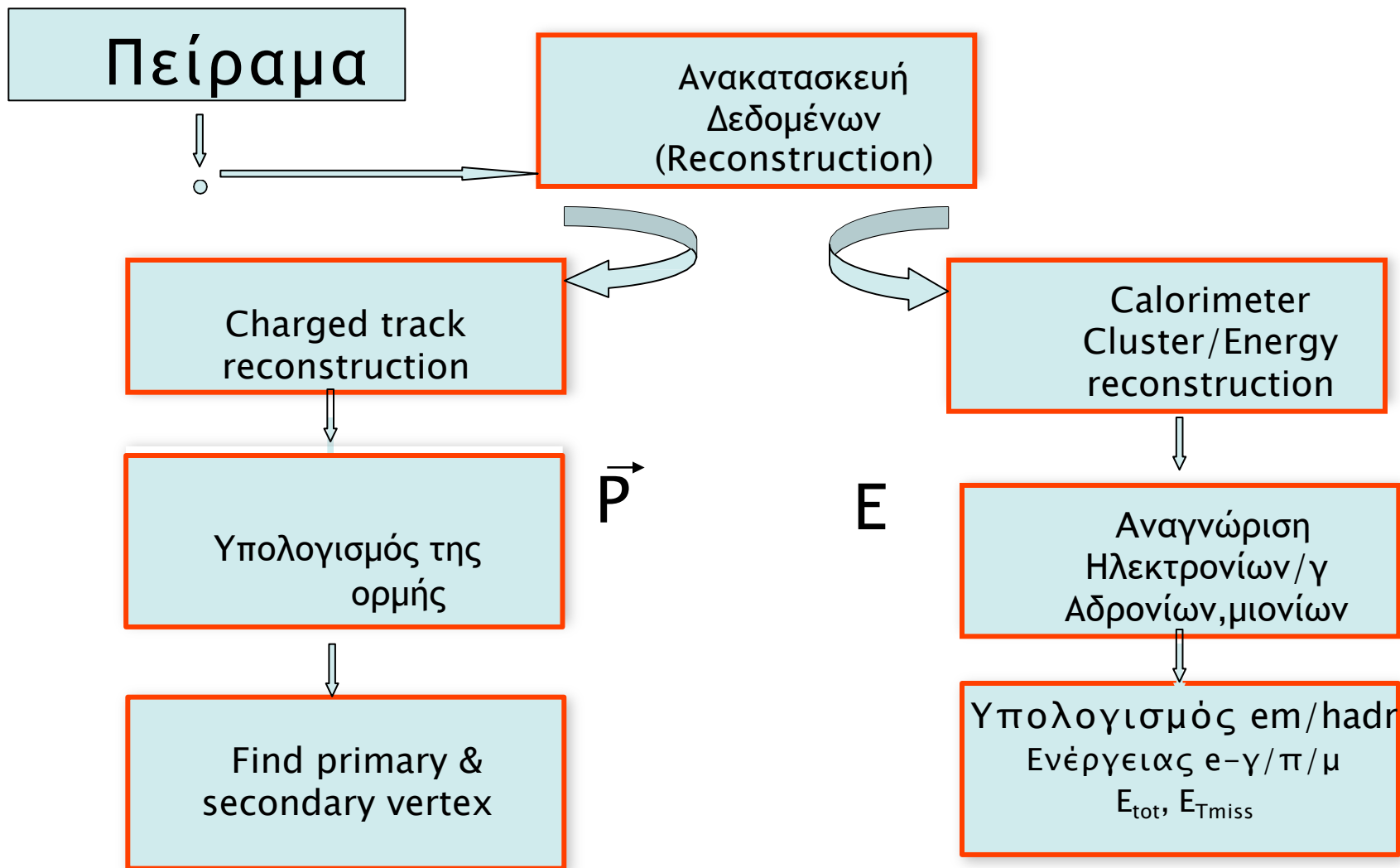
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων I



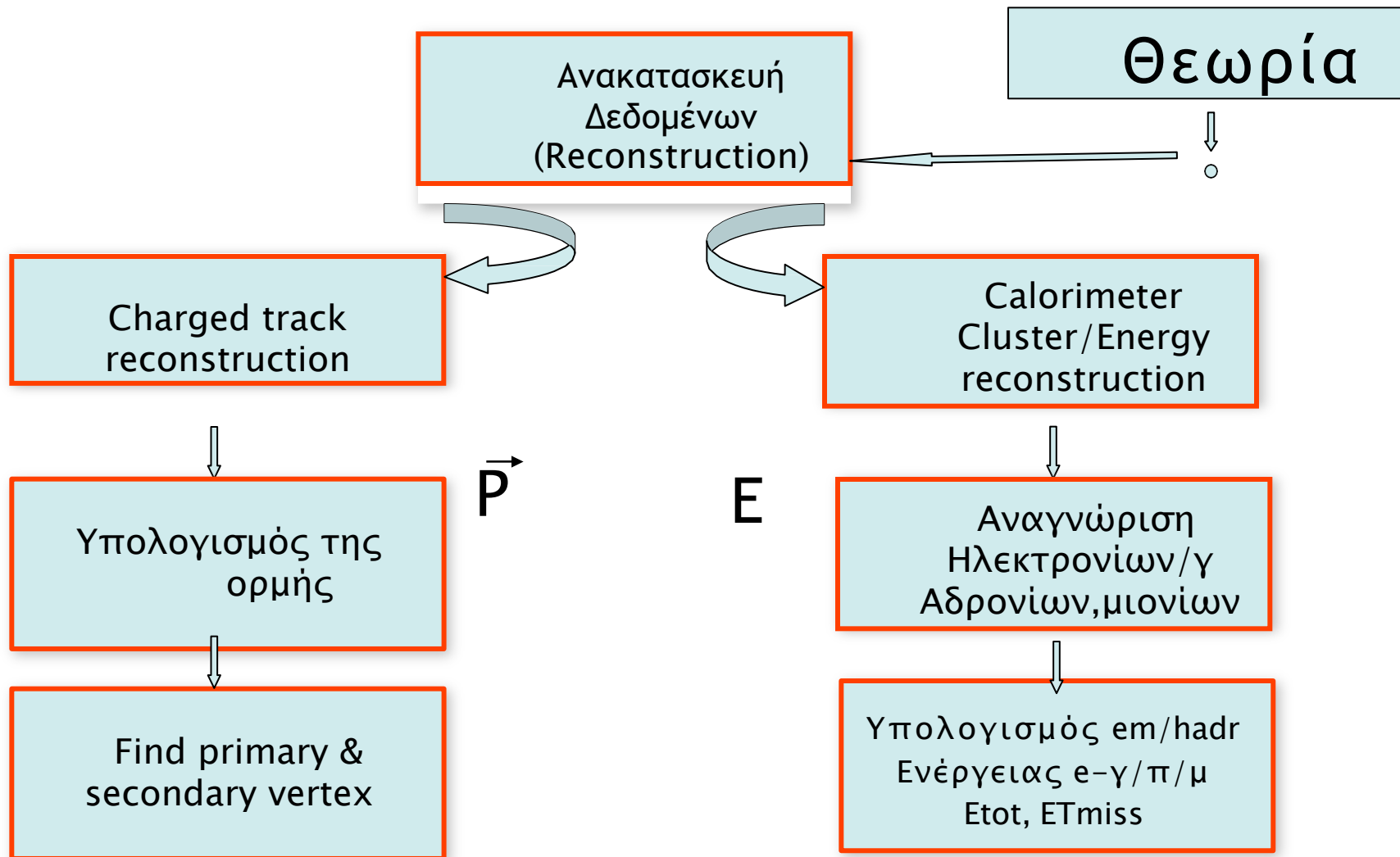
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων II



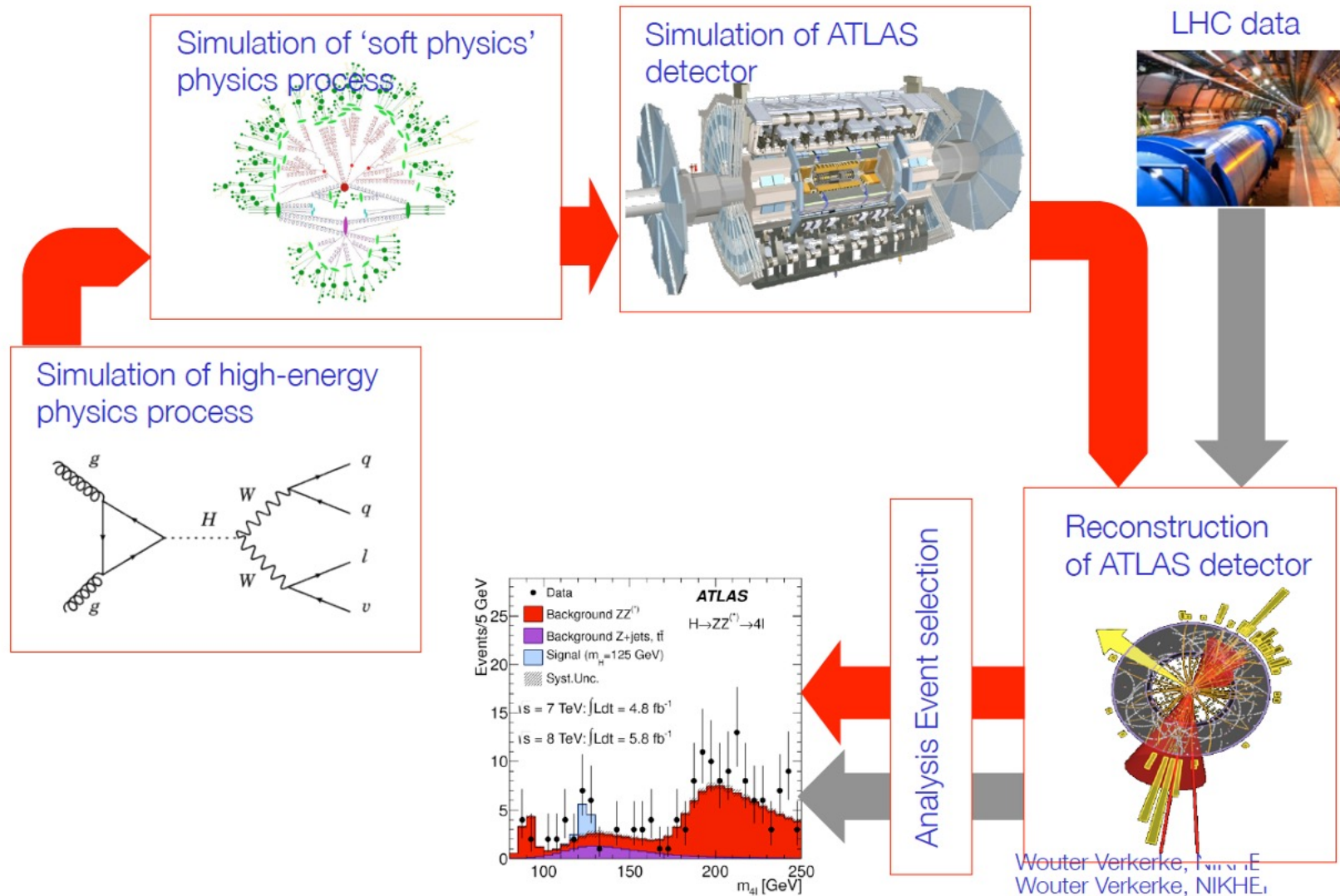
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων III



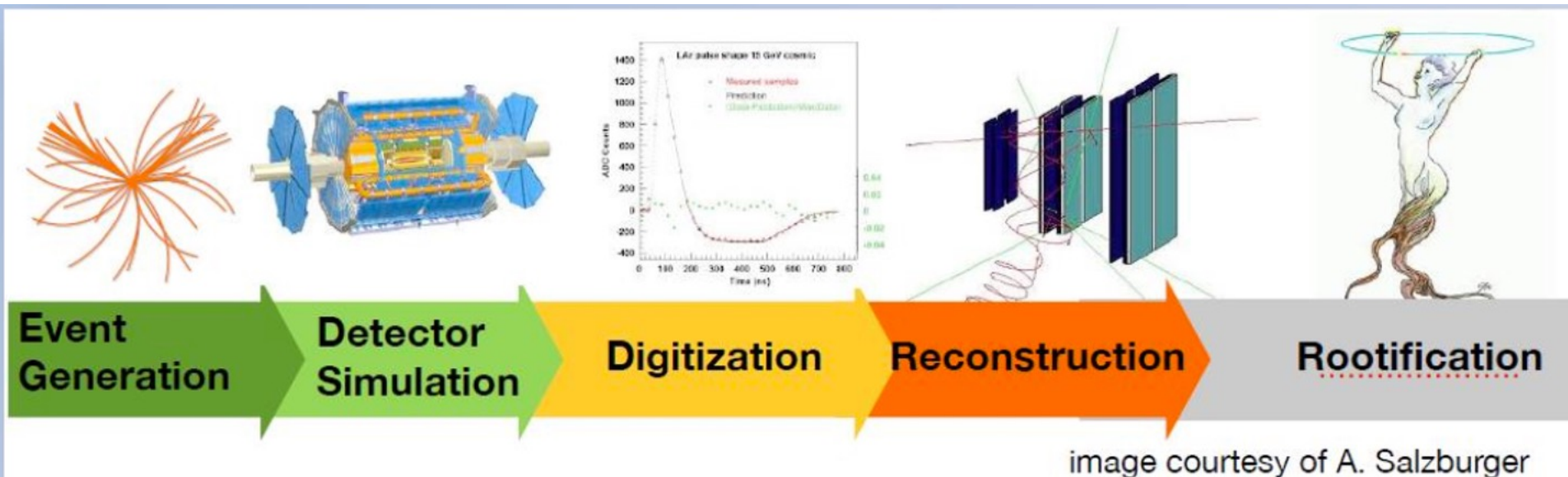
Μεθοδολογία στη Φυσική Στοιχειωδών Σωματιδίων III



The simulation workflow and origin of uncertainties



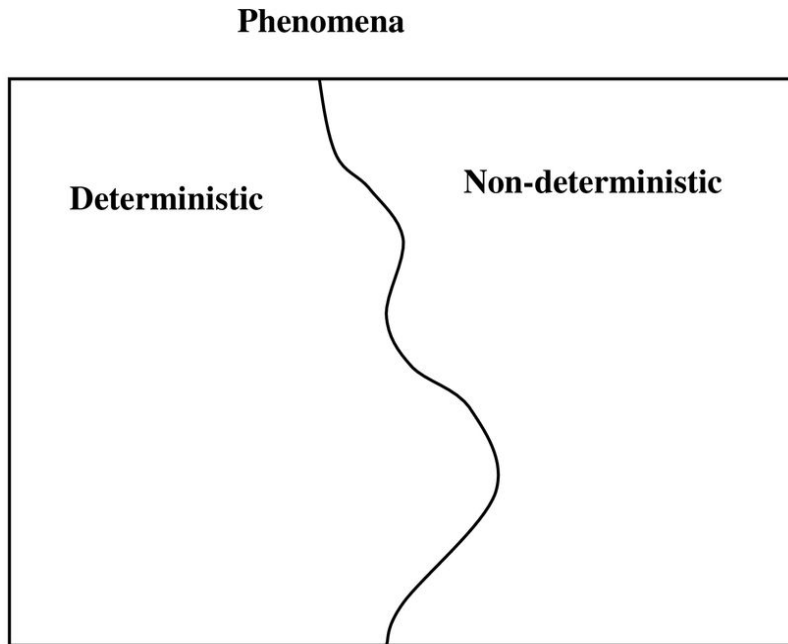
Στάδια της προσομοίωσης μιας αλληλεπίδρασης



Εισαγωγή στην Θεωρία των Πιθανοτήτων



Διαχωρισμός Φαινομένων



Deterministic – Αιτιοκρατικά

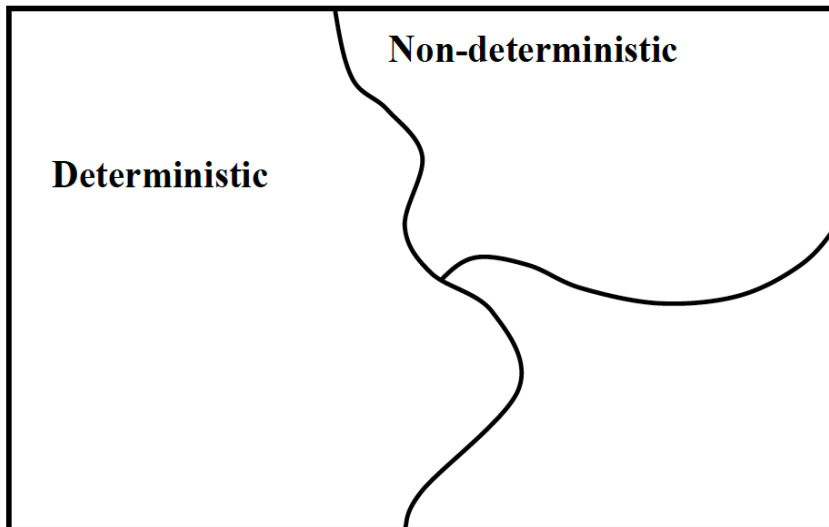
- Όταν υπάρχει ένα μαθηματικό μοντέλο που περιγράφει/προβλέπει 'απόλυτα' το αποτέλεσμα ενός φαινομένου. Η επανάληψη ενός πειράματος (υπό ταυτόσημες συνθήκες) θα φέρει το ίδιο αποτέλεσμα.
- Παραδείγματα από τον χώρο των φυσικών επιστημών.

Non – Deterministic – Στοχαστικά

- Δεν υπάρχει μαθηματικό μοντέλο που να προβλέπει 'απόλυτα' το αποτέλεσμα του φαινομένου. Ακόμη και με ίδιες συνθήκες τα αποτελέσματα διαφέρουν.

Στοχαστικά Φαινόμενα I

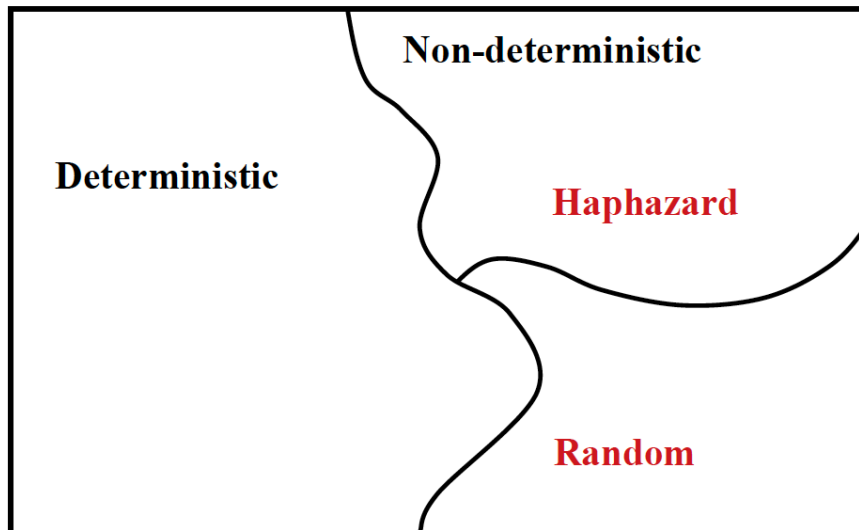
Phenomena



- Επαναλήψεις ενός πειράματος και η παρατήρηση των αποτελεσμάτων αποκαλύπτουν μια 'τυχειότητα' ή 'αβεβαιότητα' στην έκβαση του φαινομένου.
- Ανεξάρτητα από την πληροφορία που κατέχουμε δεν είμαστε σε θέση να καθορίσουμε την εξέλιξη του φαινομένου.

Στοχαστικά Φαινόμενα II

Phenomena



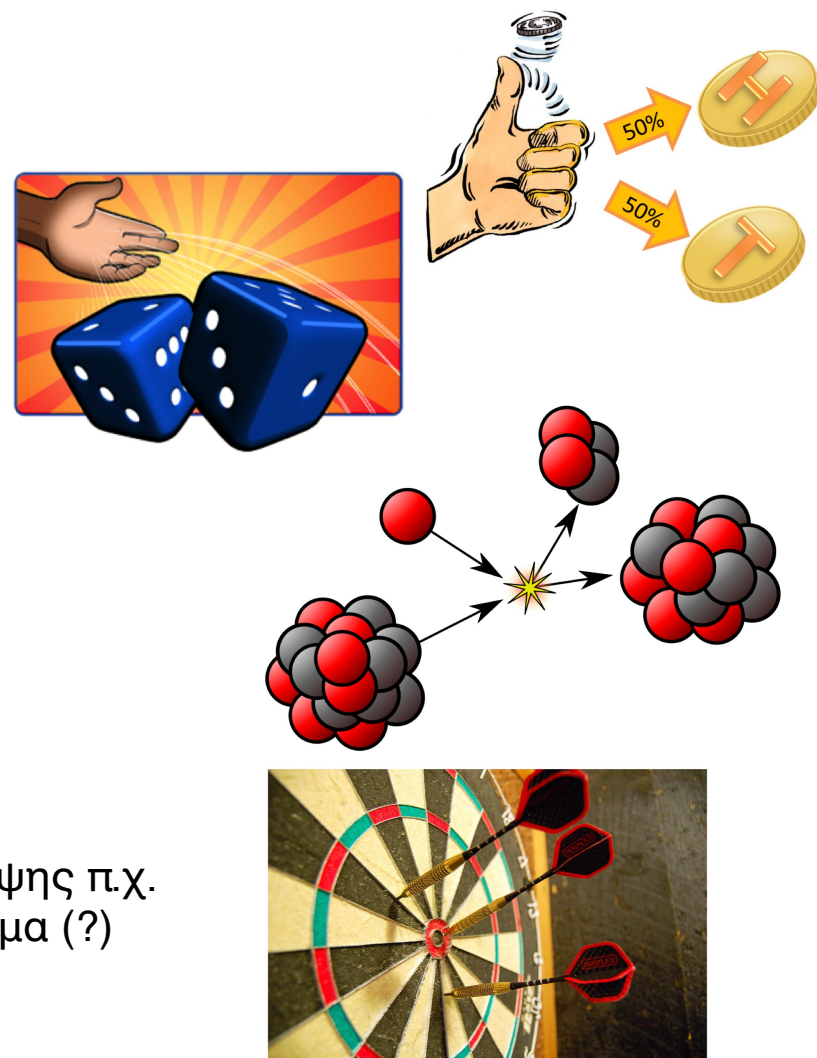
Υποκατηγορίες Στοχαστικών Φαινομένων

1. Τυχαία (Random): Υφίσταται η αδυναμία πρόβλεψης του αποτελέσματος, αλλά σε βάθος χρόνου παρατηρείται ένα 'μοτίβο', μια στατιστική κανονικότητα.

2. 'Ασυνάρτητα' (Haphazard): Ομοίως μη προβλέψιμα αποτελέσματα και δίχως κάποια στατιστική συμπεριφορά σε βάθος Χρόνου.

Στοχαστικά Φαινόμενα: Παραδείγματα

- Ρίψη νομίσματος και παρατήρηση της αλληλουχίας των στοιχειωδών αποτελεσμάτων.
- Ρίψη k ζαριών και παρατήρηση των αποτελεσμάτων.
- Καταγραφή του πλήθους σωματιδίων ραδιενεργού πηγής εντός συγκεκριμένων χρονικών διαστημάτων.
- Καταγραφή απόστασης από το κέντρο του στόχου μετά από ρίψεις βέλους (darts).



Οι επαναλήψεις οδηγούν στην δυνατότητα πρόβλεψης π.χ. 50% εμφάνιση 'γράμματα', για αμερόληπτο πείραμα (?)

Πιθανότητες

- Ορισμός 1 : Κλασικός ορισμός

Η πιθανότητα $P(X)$ είναι η ιδιότητα ενός αντικειμένου που καθορίζει πόσο συχνά συμβαίνει ένα γεγονός A .

Δίνεται από την συμμετρία για εξίσου πιθανά γεγονότα.

Γεγονότα όχι το ίδιο πιθανά τα θεωρούμε εξίσου πιθανά

$$P(X) \equiv \lim_{N \rightarrow \infty} \frac{n_x}{N}$$

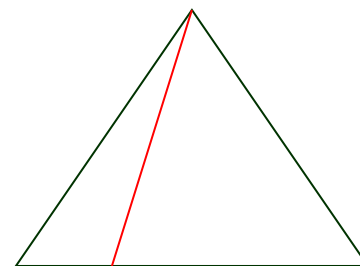
Παράδειγμα : ρίχνουμε ένα νόμισμα : $P(H) = 1/2$

Ρίχνουμε δύο ζάρια : $P(8) = 5/36$? ή $P(8) = 6/36$?

Προβλήματα με τον κλασσικό ορισμό

1. Ποτε πρόκειται για εξισου πιθανές περιπτώσεις? Αν ρίξουμε δύο νομίσματα έχουμε 3 ή 4 ίσες δυνατότητες?
2. Τι κάνουμε σε περιπτώσεις που έχουμε συνεχείς μεταβλητές? Πως να χωρίσουμε ένα τρίγωνο σε δύο τυχαία μέρη?

Έχουμε λύση



Δεν Έχουμε λύση

Πιθανότητες

- **Ορισμός 2 -Μαθηματικός:** $P(A)$ είναι αριθμός που πληρεί τα αξιώματα του Kolmogorov.
- Ας υποτεθεί ότι Ω είναι το σύνολο όλων των δυνατών στοιχειωδών γεγονότων, A_i , που είναι δυνατόν να συμβούν καθώς επίσης και ότι μόνο ένα από αυτά μπορεί να συμβεί κάθε φορά. Τότε η πιθανότητα να συμβεί το γεγονός A_i , $P(A_i)$, ορίζεται από τις ακόλουθες ιδιότητες:

$$\text{Για κάθε } A \subset \Omega, \quad P(A) \geq 0$$

$$P(\Omega) = 1$$

$$\text{Αν } A_1 \cap A_2 = \emptyset, \quad P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

- Ο μαθηματικός ορισμός ΔΕΝ δίνει πληροφορία!

Ιδιότητες των Πιθανοτήτων

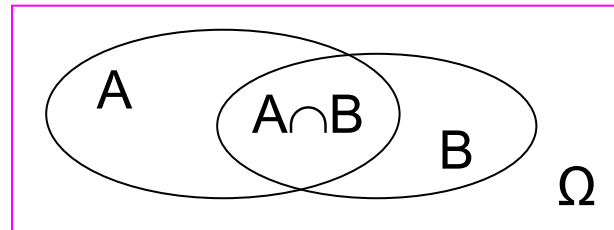
$$P(\bar{A}) = 1 - P(A)$$

$$P(A \cup \bar{A}) = 1$$

$$P(\emptyset) = 0$$

$$\text{if } A \subset B, \text{ then } P(A) \leq P(B)$$

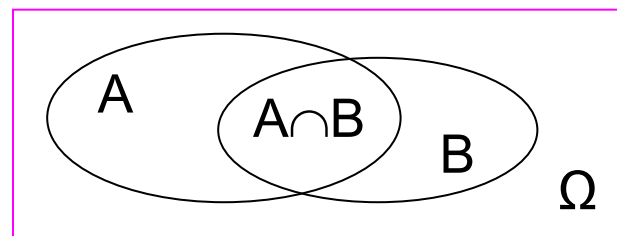
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Πιθανότητα υπό συνθήκη (conditional)

Έστω ότι κάποιο γεγονός ανήκει σίγουρα στο σύνολο γεγονότων B και ίσως ανήκει και στο σύνολο γεγονότων A . Σε τέτοιες περιπτώσεις, συνήθως μας ενδιαφέρει η πιθανότητα ένα γεγονός να ανήκει στο A δοθέντος ότι ανήκει στο B . Αυτή την πιθανότητα την χαρακτηρίζουμε ως “υπό συνθήκη” ή δεσμευμένη πιθανότητα και την συμβολίζουμε ως: $P(A|B)$ (η πιθανότητα να βρεθεί A δοθέντος του B).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Ρίχνοντας 1 ζάρι, ποια η πιθανότητα να βγει άρτιος μικρότερος του 3

$$P(n \leq 3 | n \text{ even}) = \frac{P((n \leq 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

Ερμηνεία της Πιθανότητας

I. Συχνάζουσα πιθανότητα (→ "frequentist statistics")

Τα A, B, είναι τα αποτελέσματα ενός επαναλαμβανόμενου πειράματος π.χ. κβαντομηχανική, σκέδαση σωματιδίων, ραδιενεργή διάσπαση...

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{times outcome is } A}{n}$$

II. Υποκειμενική πιθανότητα (→ "Bayesian statistics")

Οι A, B, ... είναι υποθέσεις (δηλώσεις σωστές ή ψευδείς)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- *Και οι δύο ερμηνείες συνάδουν με τα αξιώματα του Kolmogorov.*
- *Στη σωματιδιακή φυσική η ερμηνεία της σχετικής συχνότητας είναι πιο χρήσιμη, αλλά η υποκειμενική πιθανότητα μπορεί να προσφέρει πιο "φυσική" αντιμετώπιση μη επαναλαμβανόμενων φαινομένων: συστηματικά, αβεβαιότητες, πιθανότητα ύπαρξης μαγνητικών μονοπόλων,...*

Πιθανότητες - Συχνότητα

- **Ορισμός 3 -Frequentist** (‘συχνάζουσα’ πιθανότητα)
Η πιθανότητα $P(A)$ είναι το όριο του λόγου $N(A)/N$ (σε ένα σύνολο N με ιδιότητα A)

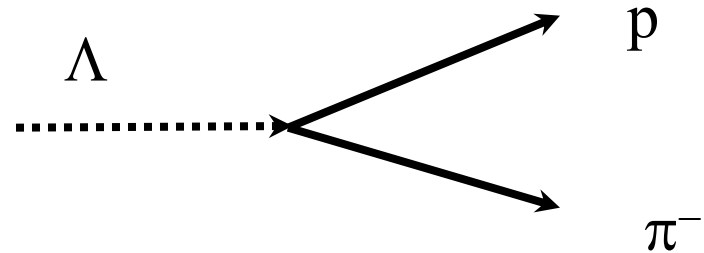
$$P(A) = \lim_{N \rightarrow \infty} \frac{N(A)}{N}$$

Πρόβλημα (περιορισμός) με τον ορισμό της ‘συχνάζουσας’ πιθανότητας

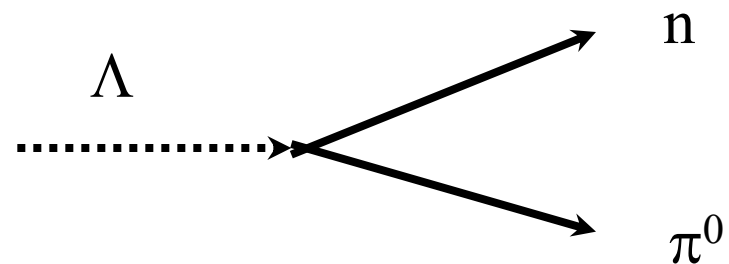
- Η $P(A)$ εξαρτάται και απο το A και απο το σύνολο N . Η πιθανότητα που υπολογίζουμε μπορεί να είναι παραπλανητική αν βασιστούμε στην συχνότητα του γεγονότος :
- ... 10 στους 30 ανθρώπους σε ένα σύνολο είναι ‘πλούσιοι’
 $\Rightarrow P(\text{πλούσιος}) = 1/3 !!!$

Συνέπειες για την Κβαντομηχανική (QM)

- Στην QM υπολογίζουμε πιθανότητες



- Οι πιθανότητες εξαρτώνται από το είδος των γεγονότων που έχουμε (π.χ. Λ^0) και την αλληλεπίδραση



PDG: $P(p\pi^-)=0.639$

$P(n\pi^0)=0.358$

Το πρόβλημα με τον ορισμό της ‘συχνάζουσας’ πιθανότητας

- ΔΕΝ μπορεί να εφαρμοστεί σε μεμονωμένα γεγονότα !
 - ... η έκφραση ‘πιθανόν να βρέξει αύριο’ ΔΕΝ είναι επιστημονική !!!
 - Ενω : ‘η δήλωση “πιθανόν να βρέξει αύριο” είναι σωστή’
 - Η ανακάλυψη ή όχι του Higgs
 - Η ανακάλυψη ή όχι σκοτεινής ύλης

Ορισμός της υποκειμενικής πιθανότητας (Bayesian)

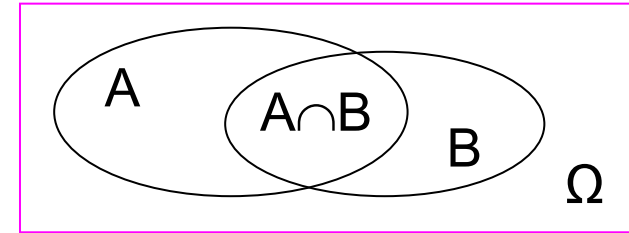
- $P(A)$ είναι η πιθανότητα που **δίνουμε** στην ιδιότητα A να υπάρξει.
- Το A μπορεί να είναι ο,τιδήποτε. Νέα θεωρία, νέο σωματίδιο, διάσπαση σωματιδίου,... βροχή, ... ενα στοίχημα...

Το Θεώρημα του Bayes



Πιθανότητα υπό συνθήκη (conditional probability) :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad P(B|A) = \frac{P(B \cap A)}{P(A)}$$



$P(A|B)$ είναι η πιθανότητα των κοινών σημείων των A και B σε σχέση με το σύνολο B και η $P(A \cap B)$ σε σχέση με το Ω

Αλλά $P(A \cap B) = P(B \cap A)$

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \Rightarrow P(A|B) = \frac{P(B|A)}{P(B)} P(A)$$

Γνωρίζουμε την πιθανότητα να συμβεί B όταν συμβαίνει A και
υπολογίζουμε την αντίστροφη πιθανότητα

Το Θεώρημα του Bayes

Το Θεώρημα Bayes προσφέρει τη δυνατότητα υπολογισμού της πιθανότητας $P(A|B)$ η οποία εκφράζει την εμπιστοσύνη μας στην υπόθεση A μετά την παρατήρηση του αποτελέσματος B (ύστερη γνώση).

Παράδειγμα: Θεώρημα Bayes

Ποιο το κλάσμα των γυναικών μεταξύ των φοιτητών

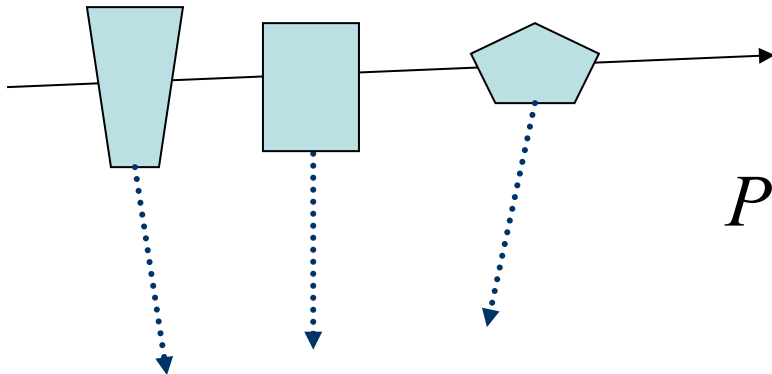
Από την αναλογία φοιτητών και γυναικών στον πληθυσμό και το κλάσμα των φοιτητών μεταξύ των γυναικών υπολογίζουμε το κλάσμα των γυναικών μεταξύ των φοιτητών:

$P(A) = 0,02$	(κλάσμα φοιτητών στον πληθυσμό)
$P(B) = 0,5$	(κλάσμα των γυναικών στον πληθυσμό)
$P(A B) = 0,018$	(κλάσμα φοιτητών μεταξύ γυναικών)
$P(B A) = ??$	(κλάσμα γυναικών μεταξύ των φοιτητών)

$$\begin{aligned} P\{B | A\} &= \frac{P\{A | B\}P\{B\}}{P\{A\}} \\ &= \frac{0.018 \cdot 0.5}{0.02} = 0.45 \end{aligned}$$

Εφαρμογή του Θεωρήματος Bayes για την συχνάζουσα πιθανότητα

- Ταυτοποίηση σωματιδίων
- -Τύποι σωματιδίων :e, π, μ, K, p
- -Σήματα απο ανιχνευτές: DCH, RICH, TOF, TRD



$$P'(e) = P(e | DCH) = \frac{P(DCH | e) P(e)}{P(DCH)}$$

$$P(DCH) = P(DCH | e)P(e) + P(DCH | \mu)P(\mu) + P(DCH | \pi)P(\pi)...$$

- Αντίστοιχα υπολογίζουμε την $P(e|RICH)$ κλπ (δηλ. σήμα στον ανιχνευτή να οφείλεται σε e)

Παράδειγμα χρήσης του θεωρήματος Bayes με την συχνάζουσα πιθανότητα

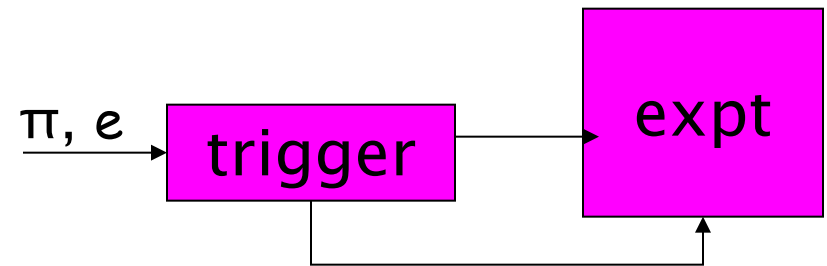
- Δεσμη αποτελείται από π 'ς & e 'ς $N_e/N_\pi=10^{-3}$

- Το σύστημα σκανδαλισμού:

$$\varepsilon_e=0.98, \varepsilon_\pi=0.03$$

$$P(e)=10^{-3} \quad P(\pi)=1-10^{-3}$$

$$P(T|e)=0.98, \quad P(T|\pi)=0.03$$



- $P(e|T)=P(T|e)*P(e)/(P(T|e)*P(e)+P(T|\pi)P(\pi))$ είναι η πιθανότητα το εισερχόμενο σωματίο να είναι ηλεκτρόνιο όταν το σύστημα σκανδαλισμού λέει 'ναι'

Το θεώρημα Bayes και η υποκειμενική πιθανότητα

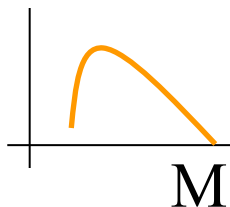
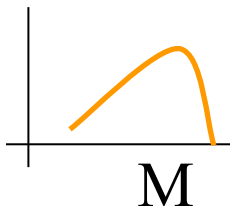
$$P(Theory|Result) = \frac{P(Result|Theory)}{P(Result)} P(Theory)$$

- Η εκ των προτέρων 'πίστη' σε μια θεωρία αλλάζει με βάση τα πειραματικά αποτελέσματα :
- Εάν $(P(Result|Theory) = 0$, (η Θεωρία απαγορεύει το πειραματικό αποτέλεσμα) $\Rightarrow$ η Θεωρία δεν ισχύει
- Εάν $(P(Result|Theory) = \text{Μεγάλο}$, ενισχύεται η πίστη μας στη Θεωρία

Πρόβλημα με την υποκειμενική πιθανότητα

- Η πιθανότητα ‘μου’ $P(A)$ και η πιθανότητα $P(A)$ ενός τρίτου μπορεί να διαφέρουν...
- **Αλλα** οι επιστήμονες θάπρεπε να είναι αντικειμενικοί!
- Δικαιολογείται η άγνοια να δηλώνεται με $P(A)=\text{flat}$?

(αν το A παίρνει διακριτές τιμές το ίδιο πιθανές, ναι. Αν συνεχής μεταβλητή, π.χ. M_{higgs} διαφορετική υπόθεση θα δώσει άλλα αποτελέσματα)



Συνάρτηση Πυκνότητας Πιθανότητας-Κατανομές

- Ορίσαμε τη Πιθανότητα
- Όταν μετρούμε ένα παρατηρήσιμο μέγεθος πολλές φορές, το αποτέλεσμα **x** κατανέμεται σύμφωνα με μιά **κατανομή πιθανότητας (probability distribution)**
- Το αποτέλεσμα έχει μια διακύμανση είτε λόγω καποιων **παραμέτρων** (π.χ. ηλεκτρονικό θόρυβο) είτε λόγω **διακύμανσης** του φυσικού φαινομένου (π.χ. διάσπαση ή σκέδαση σωματιδίων)
- Οι κατανομές πιθανότητας μπορεί να είναι διακριτές ή συνεχείς.

Πιθανότητα vs Στατιστική

- Πιθανότητα: απο την θεωρία στα δεδομένα
 - Υπολογίζουμε όλα τα πιθανά αποτελέσματα (συνέπειες) ενός πειράματος για συγκεκριμένο πρόβλημα
- Στατιστική: απο τα δεδομένα στη θεωρία
 - Λύνουμε το αντίστροφο πρόβλημα: απο τα δεδομένα προσπαθούμε να βρούμε τους κανόνες-νόμους=> **ανάλυση των δεδομένων**
 - => Υπολογισμός παραμέτρων και του σφάλματος
 - => έλεγχος της υπόθεσης: πιστότητα, συμφωνία...

Συναρτήσεις Πυκνότητας Πιθανότητας

- Επανειλημμένη μέτρηση -πειραματικά- μιας συνεχούς μεταβλητής x
- Ορισμός : η πιθανότητα P να μετρήσουμε μια τιμή x στο διάστημα $(x, x+dx)$ δίνεται από την **συνάρτηση πυκνότητας πιθανότητας** : probability density function $f(x)$ (p.d.f.) : $P=f(x) dx$
 - Είναι το μέτρο του πόσο συχνά η τιμή x εμφανίζεται σ'ένα δείγμα

$$Norm = \int_{x \min}^{x \max} f(x') \cdot dx' = 1$$

Κατανομές-ιδιότητες

- Αναμενόμενη τιμή = **Μέση τιμή** (~“centre of gravity”)

$$E[x] \equiv \int_{x \min}^{x \max} x \cdot f(x) \cdot dx = \langle x \rangle = \mu$$

- Διασπορά - **Variance** = σ^2 = το τετράγωνο της απόκλισης σ
Μετράει την διασπορά του x σε σχέση με την μέση τιμή

$$V[x] = E((x - \mu)^2) \equiv \int_{x \min}^{x \max} (x - \mu)^2 f(x) \cdot dx = \sigma^2$$

$$\equiv \langle (x - \mu)^2 \rangle \equiv \langle x^2 \rangle - \mu^2$$

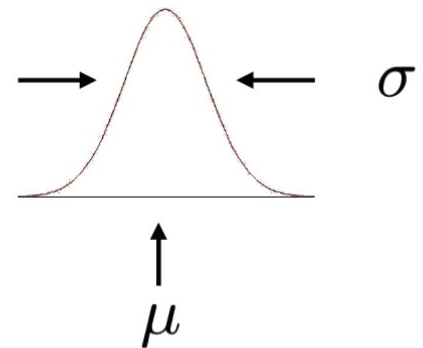
Variance, standard deviation

Variance: $V[x] = E[x^2] - \mu^2 = E[(x - \mu)^2]$

Notation: $V[x] = \sigma^2$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

$\sigma \sim$ width of pdf, same units as x .



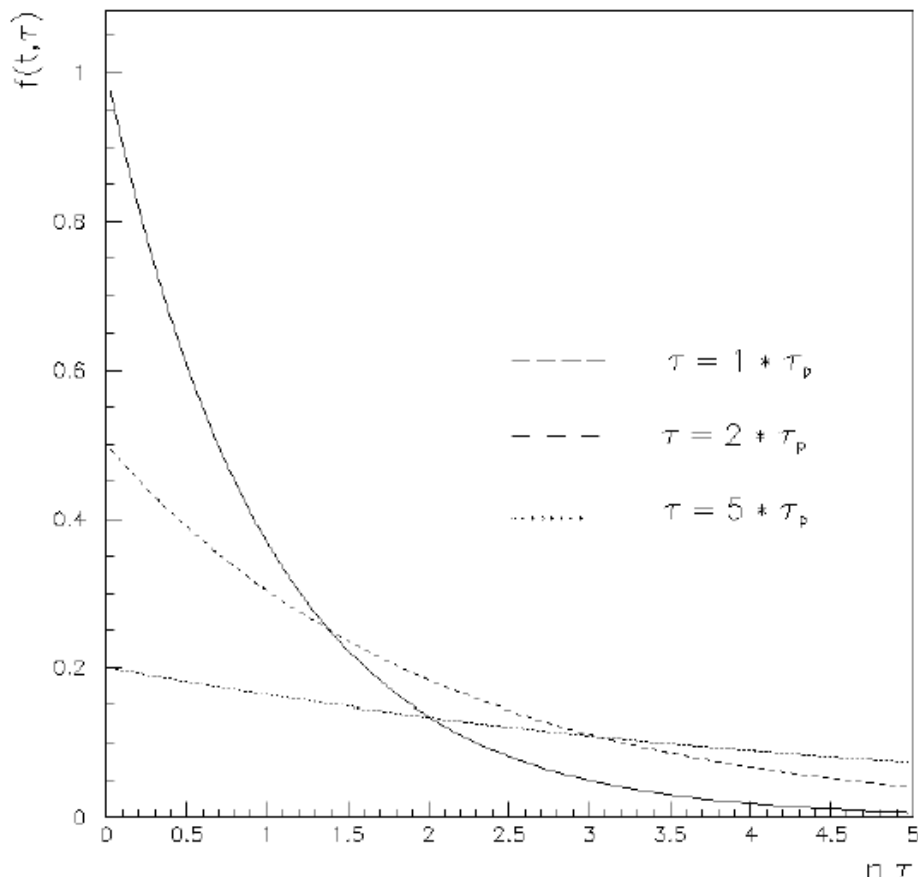
Σχέση μεταξύ σ και άλλων μέτρων πλάτους, π.χ. το πλήρες πλάτος στο μισό μέγιστο (FWHM) εξαρτάται από την pdf, π.χ. $FWHM = 2,35\sigma$ για Gaussian.

Κατανομές Πυκνότητας Πιθανότητας

<u>Distribution/pdf</u>	<u>Example use in Particle Physics</u>
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student's t	Resolution function with adjustable tails

Παραδείγματα

- Ο χρόνος ζωής t σωματιδίου στο σύστημα ηρεμίας του π.χ. Το πiónιο : μέσος χρόνος ζωής $\tau_{\pi} = 2.6 \cdot 10^{-8} \text{ sec}$



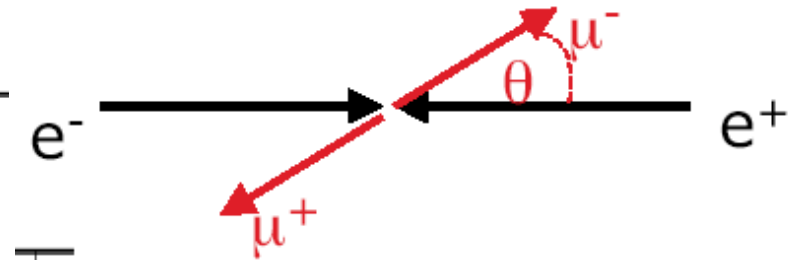
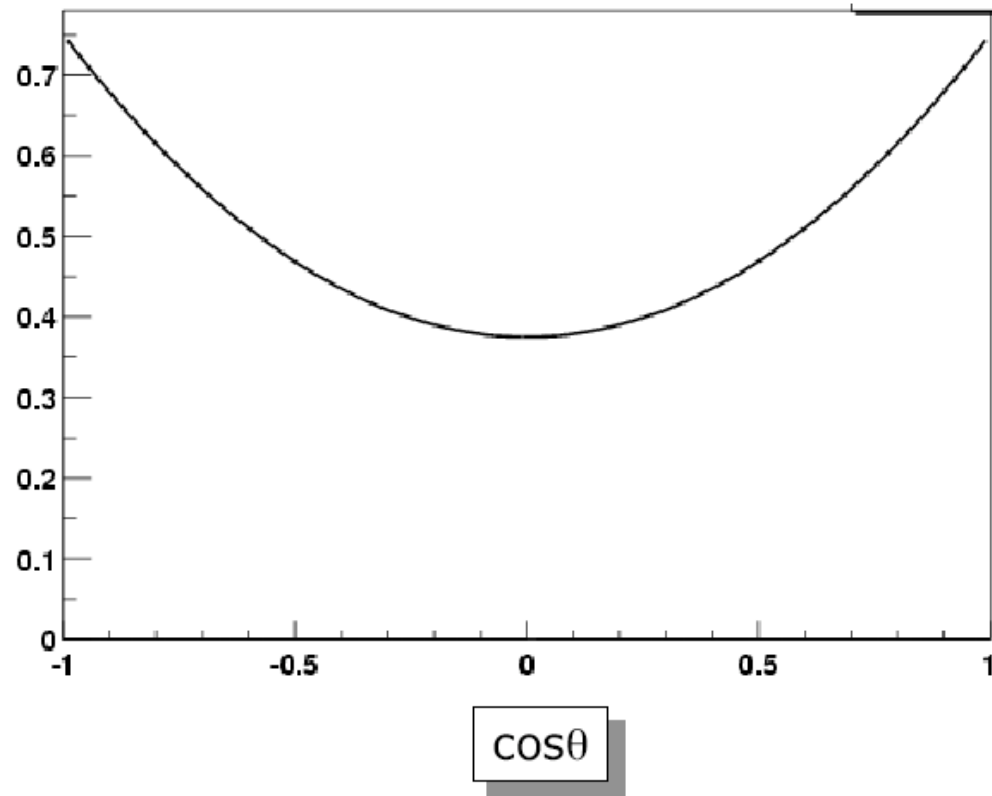
$$f(t; \tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}}$$

$$E[t] = \tau$$

$$V[t] = \tau^2$$

Παραδείγματα

- Κατανομή της πολικής γωνίας θ του μιονίου στην σκέδαση $ee \rightarrow \mu\mu$



$$\frac{1}{\sigma_{tot}} \frac{d\sigma}{d\cos\theta} =$$

$$\equiv f(\cos\theta) = \frac{3}{8} [1 + (\cos\theta)^2]$$

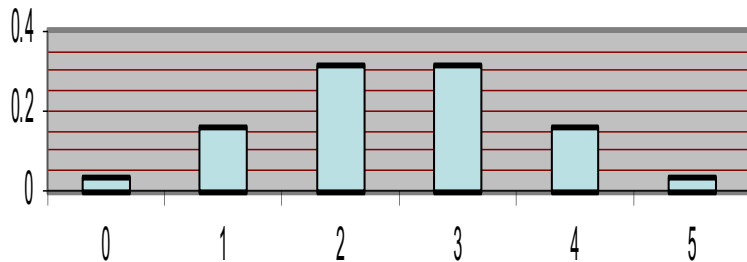
Κατανομές-Διακριτές κατανομές

- Διωνυμη κατανομή: η τυχαία μεταβλητή έχει δύο δυνατότητες

n ανεξάρτητα πειράματα (δοκιμές Bernoulli): το αποτέλεσμα του καθενός είναι «επιτυχία» ή «αποτυχία», η πιθανότητα επιτυχίας σε οποιαδήποτε δεδομένη δοκιμή είναι **p**.

n προσπάθειες και **r** επιτυχίες

p η μεμονωμένη πιθανότητα για επιτυχία



Mean

$$\mu = \langle r \rangle = \sum r P(r) \\ = np$$

$$P(r ; n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

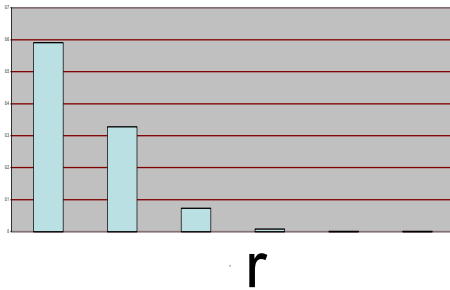
Variance

$$V \equiv \sigma^2 = \langle (r - \mu)^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2 \\ = np(1-p)$$

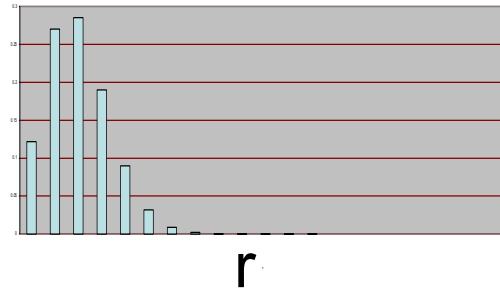
$$1-p \equiv \bar{p} \equiv q$$

Παραδείγματα διώνυμης κατανομής

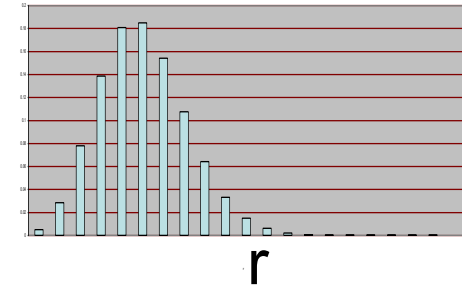
$p=0.1$ $n=5$



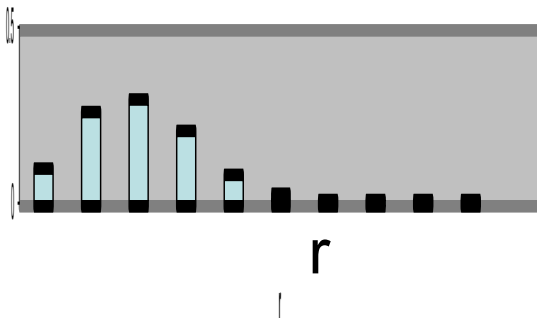
$n=20$



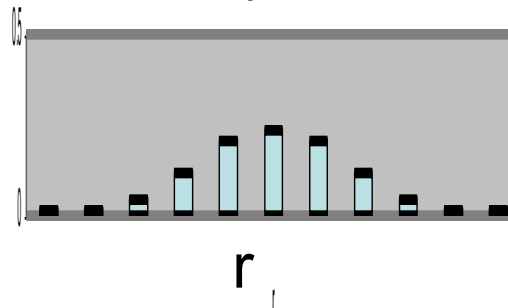
$n=50$



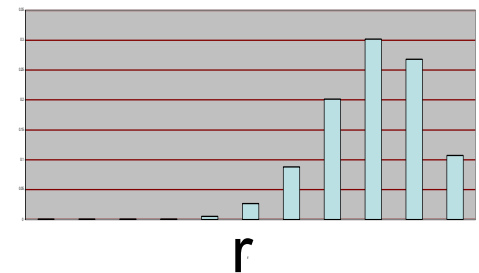
$n=10$ $p=0.2$



$p=0.5$

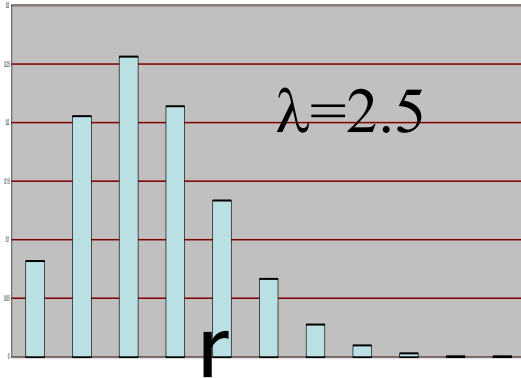


$p=0.8$



Κατανομή Poisson

- Δίνει την πιθανότητα να συμβούν r γεγονότα όταν ο ρυθμός των γεγονότων κατά μέσον όρο είναι λ (είναι το όριο της διωνυμικής κατανομής όταν $n \rightarrow \infty$ $p \rightarrow 0$, $np = \lambda$)



$$P(r; \lambda) = e^{-\lambda} \frac{\lambda^r}{r!}$$

Mean

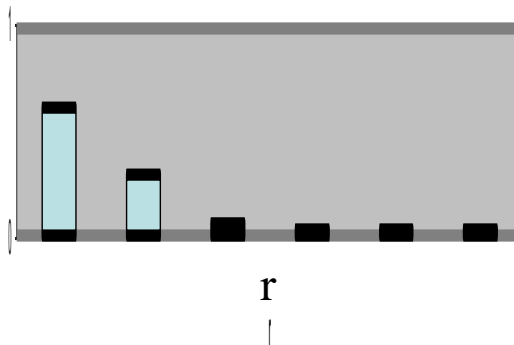
$$\mu = \langle r \rangle = \sum r P(r) \\ = \lambda$$

Variance

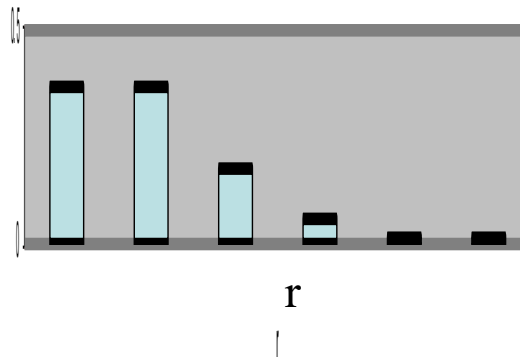
$$V \equiv \sigma^2 = \langle (r - \mu)^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2 \\ = \lambda$$

Παραδείγματα κατανομής Poisson

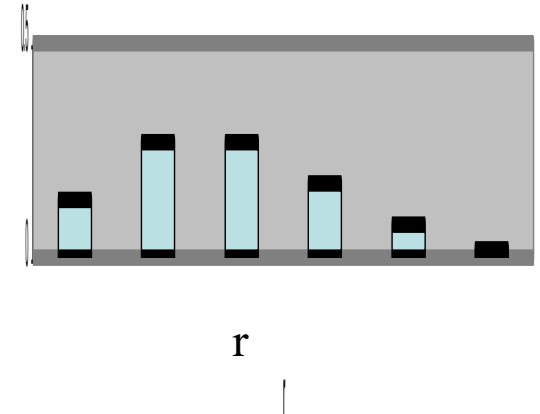
$\lambda=0.5$



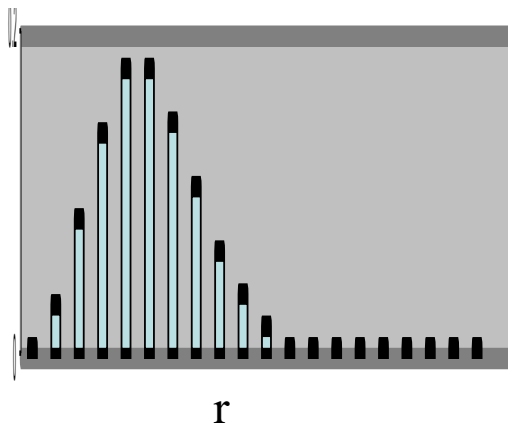
$\lambda=1.0$



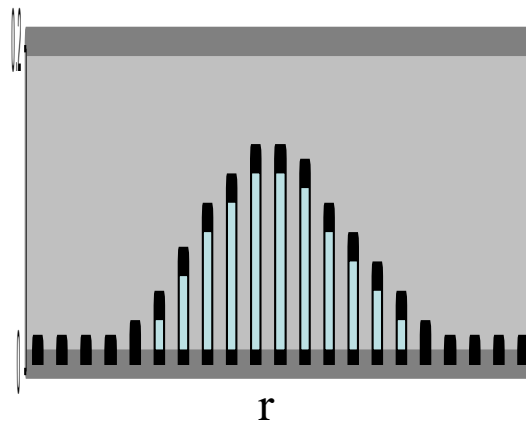
$\lambda=2.0$



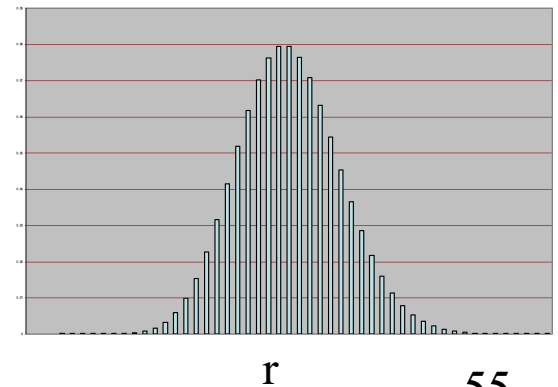
$\lambda=5.0$



$\lambda=10$



$\lambda=25$

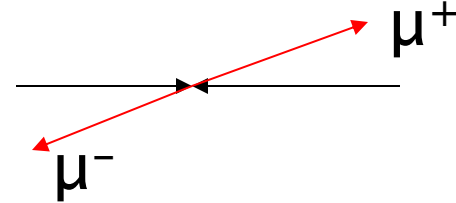
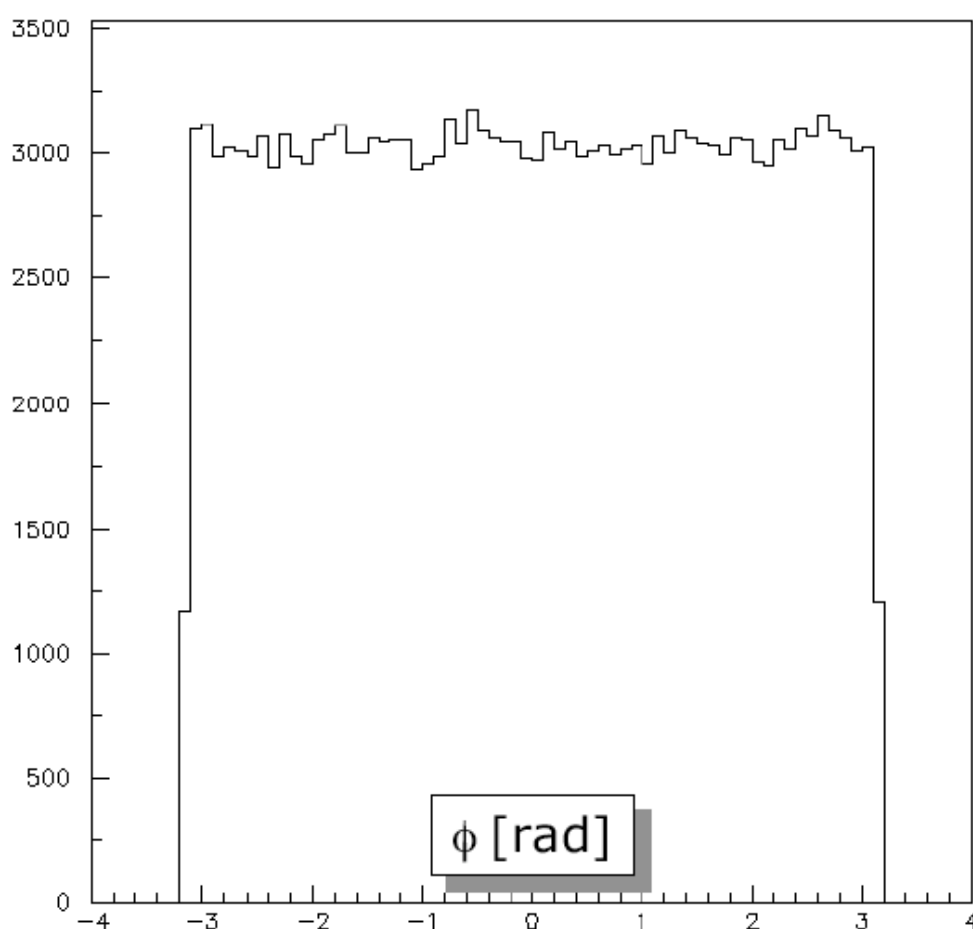


Κατανομές Poisson

- Το πλήθος των ραδιενεργών διασπάσεων σε ορισμένο χρόνο t , **μικρό** σε σχέση με τον χρόνο διάσπασης
- Το πλήθος συγκεκριμένου τύπου σωματιδίων σε σκέδαση σωματιδίου-σωματιδίου όταν ο **συνολικός αριθμός των γεγονότων είναι πολύ μεγάλος** και η συγκεκριμένη διαδικασία **σπάνια**
- Η πιθανότητα παρατήρησης n γεγονότων σε **χρόνο t** , όταν ο μέσος **ρυθμός είναι μ : $\lambda = \mu t$** . Ημέση τιμή **λ -αναμενόμενη τιμή-** και η διακύμανση : **variance** στην κατανομή Poisson είναι **ισες!**
- Απο δώ προκύπτει **ο τύπος : $n \pm \sqrt{n}$** που χρησιμοποιείται στα στατιστικά σφάλματα όταν μετρούμε σε συγκεκριμένο χρονικό διάστημα

Ομοιόμορφες Κατανομές (Uniform Distributions)

- Η πιθανότητα είναι σταθερή σε ένα διάστημα : πχ η κατανομή των μιονίων κατα την αζιμουθιακή γωνία ϕ στη σκέδαση : $ee \rightarrow \mu\mu$



$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

E : αναμενόμενη τιμή

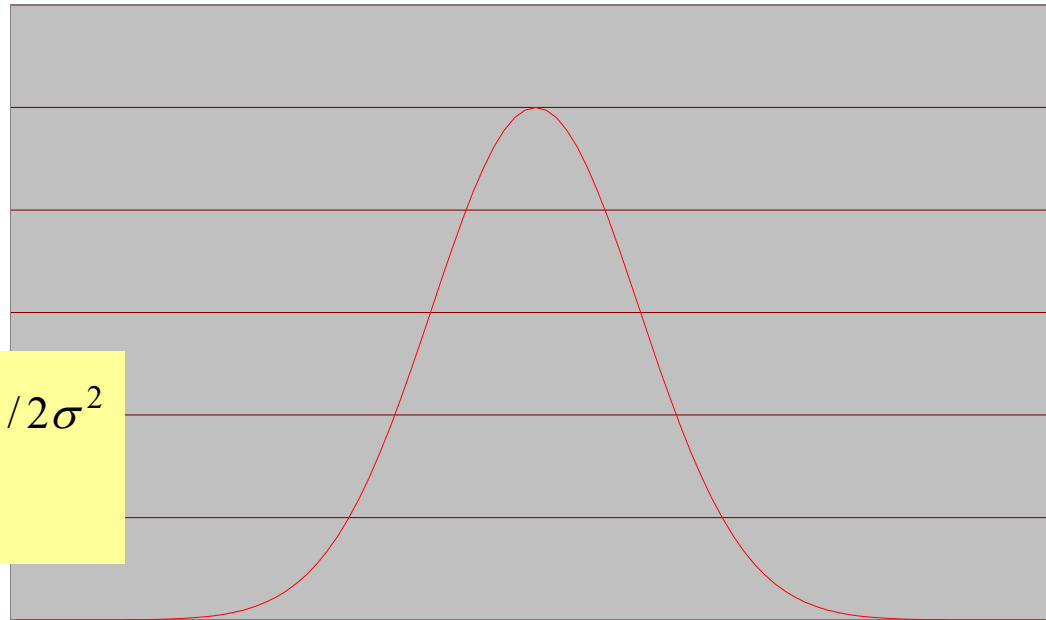
V : variance $E[x] = \frac{1}{2}(\alpha + \beta)$

$$V[x] = \frac{1}{12}(\beta - \alpha)^2$$

Κατανομή Gauss ή Normal

Πυκνότητα πιθανότητας

$$P(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2}$$



Mean

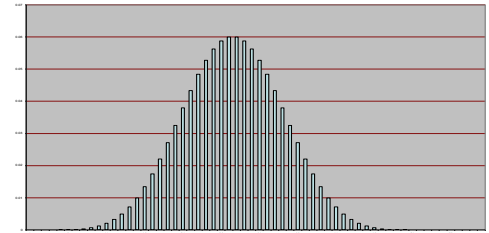
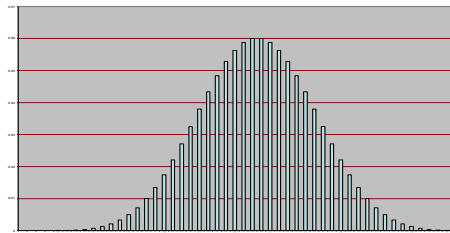
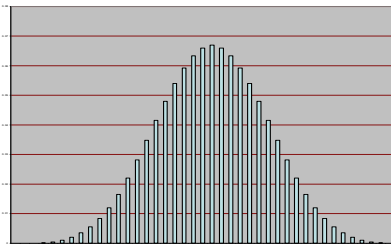
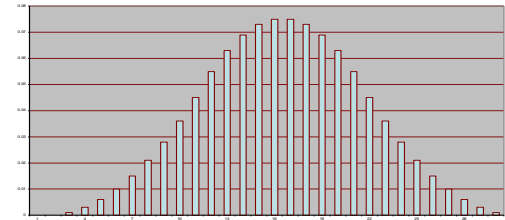
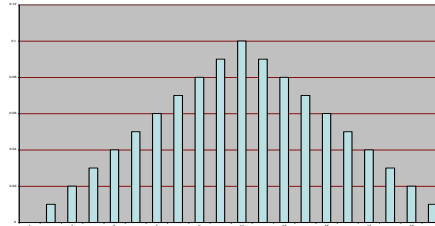
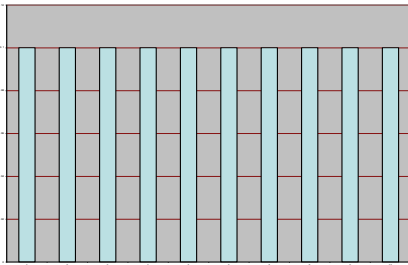
$$\begin{aligned}\mu &= \langle x \rangle = \int x P(x) dx \\ &= \mu\end{aligned}$$

Variance

$$\begin{aligned}V \equiv \sigma^2 &= \langle (x - \mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 \\ &= \sigma^2\end{aligned}$$

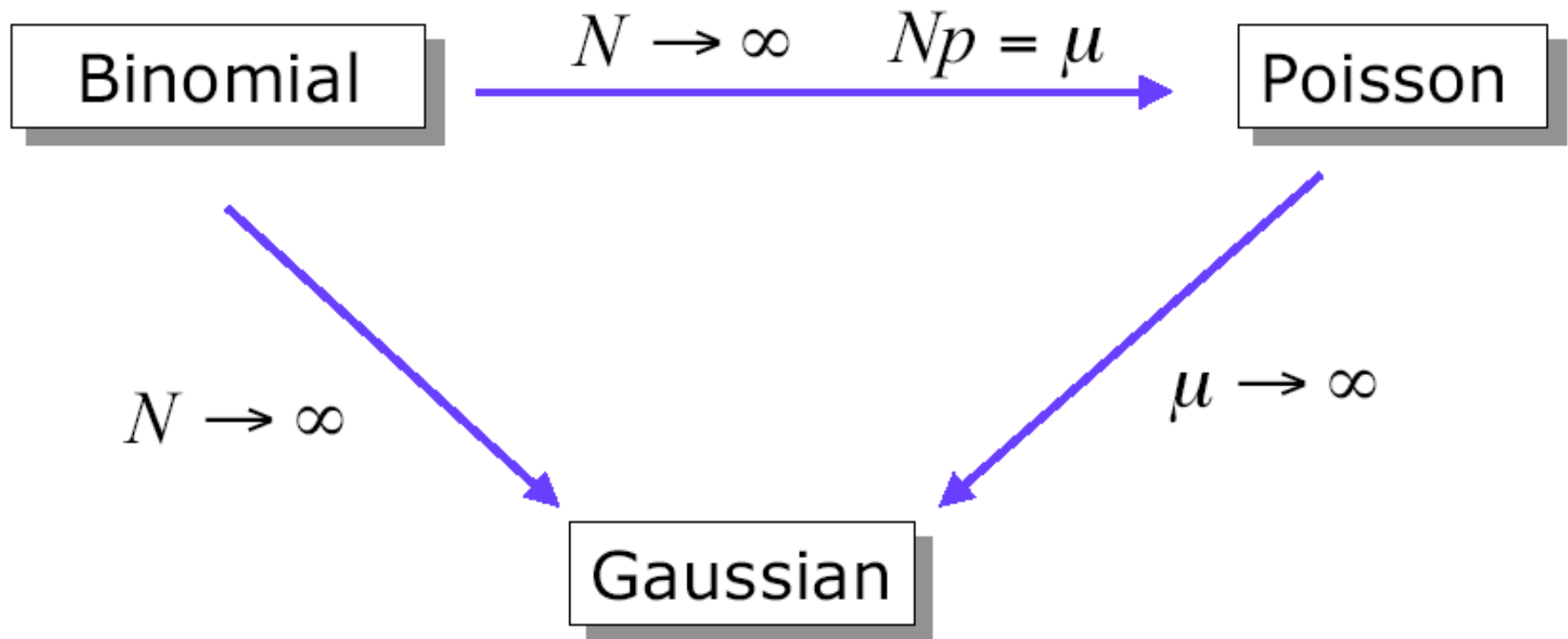
Central Limit Theorem (CLT)

- Το άθροισμα y , n τυχαίων, ανεξάρτητων αριθμών x_i είναι gaussian ($n \rightarrow \infty$) όποια κι αν είναι η κατανομή των x_i
- Convolute uniform distribution με τον εαυτό της



Σχέσεις μεταξύ κατανομών

Διωνυμη : N =το πλήθος των δοκιμών, p =η πιθανότητα επιτυχίας
σε κάθε δοκιμή



Τύποι κατανομών Gauss

Μια μόνο κατανομή υπάρχει!

68.27% within 1σ

95.45% within 2σ

99.73% within 3σ

90% within 1.645σ

95% within 1.960σ

99% within 2.576σ

99.9% within 3.290σ



Normalisation
(if required)

width
scaling
factor

Falls to $1/e$
of peak at
 $x=\mu\pm\sigma$



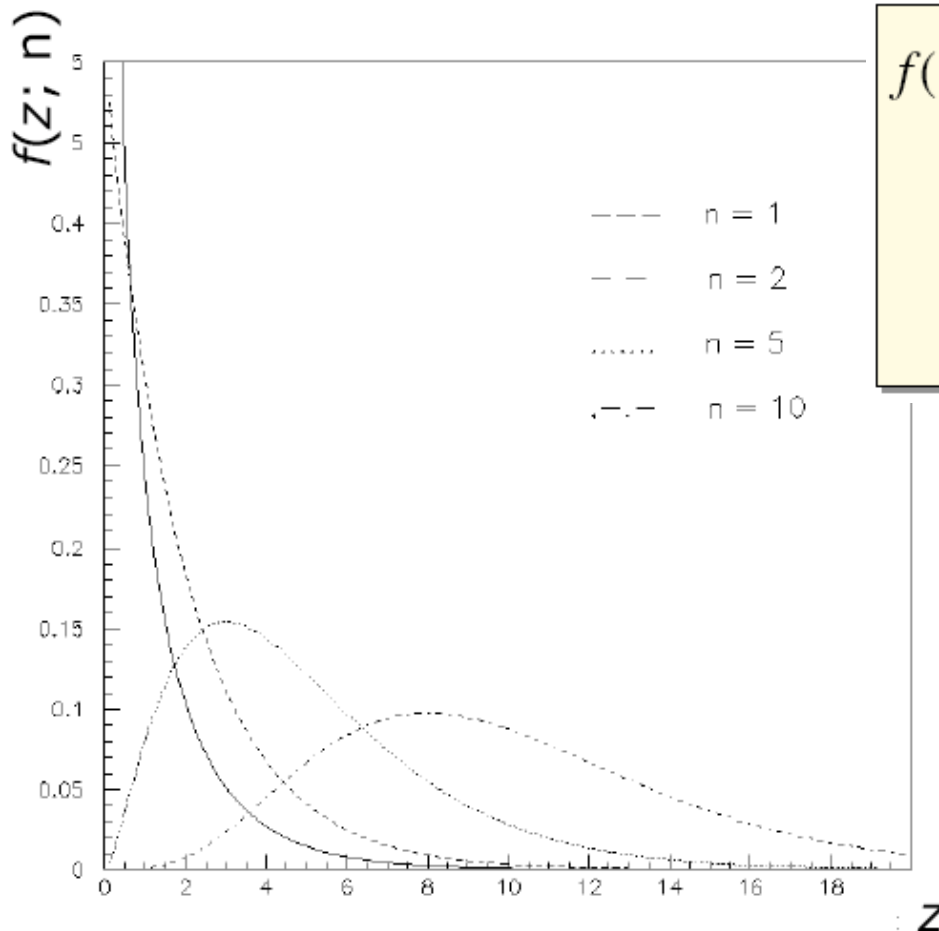
Location
change μ



Κατανομή Chi-Squared (χ^2)

$$z = \chi^2 = \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

x_i : ανεξαρτητες μεταβλητές
με κατανομή Gauss



$$f(z; n) = \frac{1}{2^{n/2} \Gamma(n/2)} z^{n/2-1} e^{-z/2} ; n = 1, 2, \dots$$

$$E[z] = n$$

$$V[z] = 2n$$

Το πλήθος των βαθμών ελευθερίας

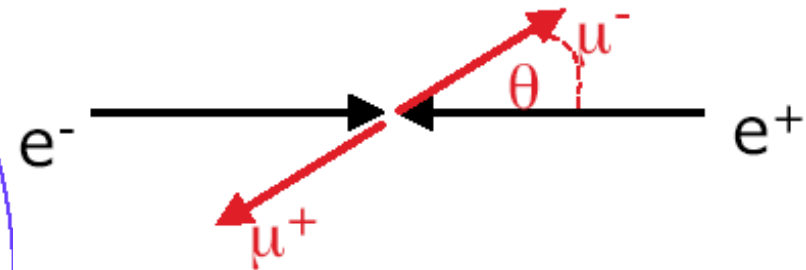
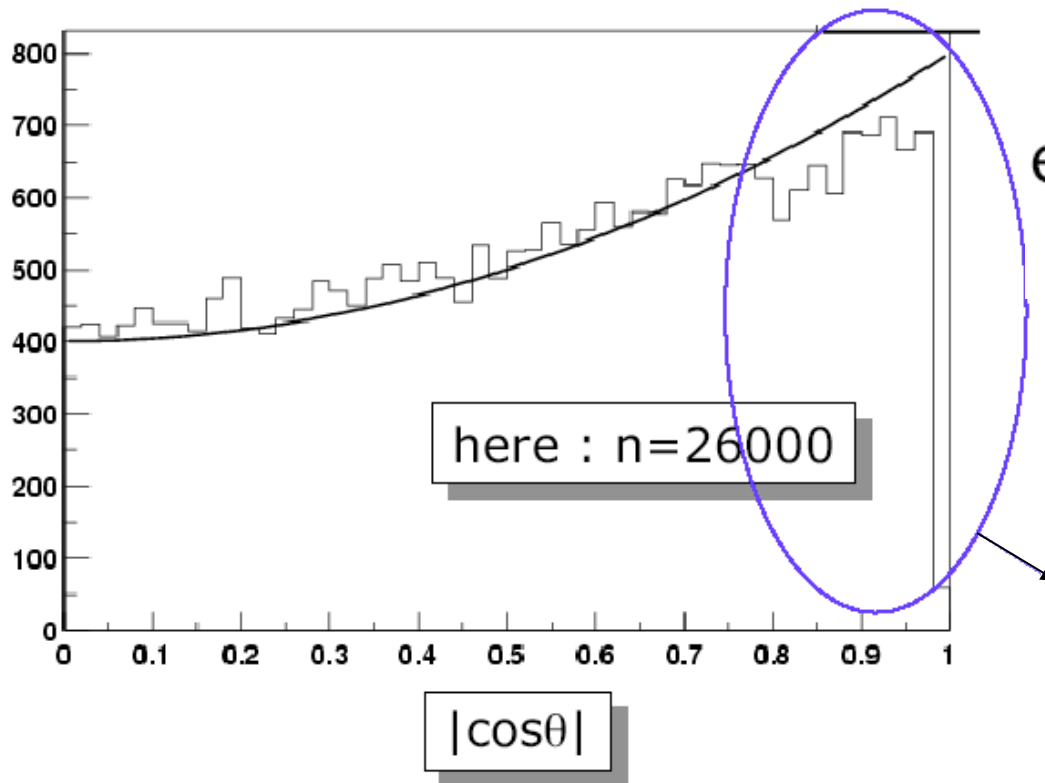
$$\Gamma(x) = e^{-Cx} \frac{1}{x} \prod_{n=1}^{+\infty} \frac{e^{x/n}}{1 + x/n}$$

$$C = \lim_{n \rightarrow \infty} 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \ln(n)$$

Ιστογράμματα (Histograms)

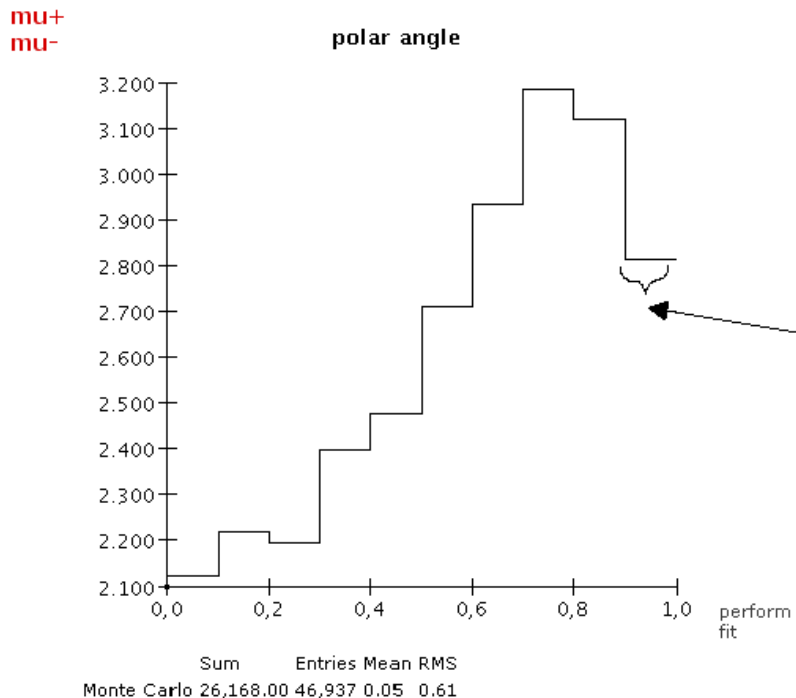
Ιστογράμματα

- Συνήθως οι μεταβλητές που χρησιμοποιούμε είναι διακριτές μεταβλητές $x_1, x_2, \dots, x_n$ που προέρχονται από πειραματικές μετρήσεις
- Παράδειγμα: κατανομή **πολικής γωνίας θ**



Σύγκριση των δεδομένων
με ομαλή κατανομή

Ιστογράμματα



- Entries, normalization,...

- Bin width(s) (στο παράδειγμα ισαπέχοντα bins)

- Πλήθος των bins
- Περιοχή (Range)

Process group: **alephz** h1cc h1nc l3higgs189

Histogram definition:

Title: **polar angl**

Number of Bins: **10**

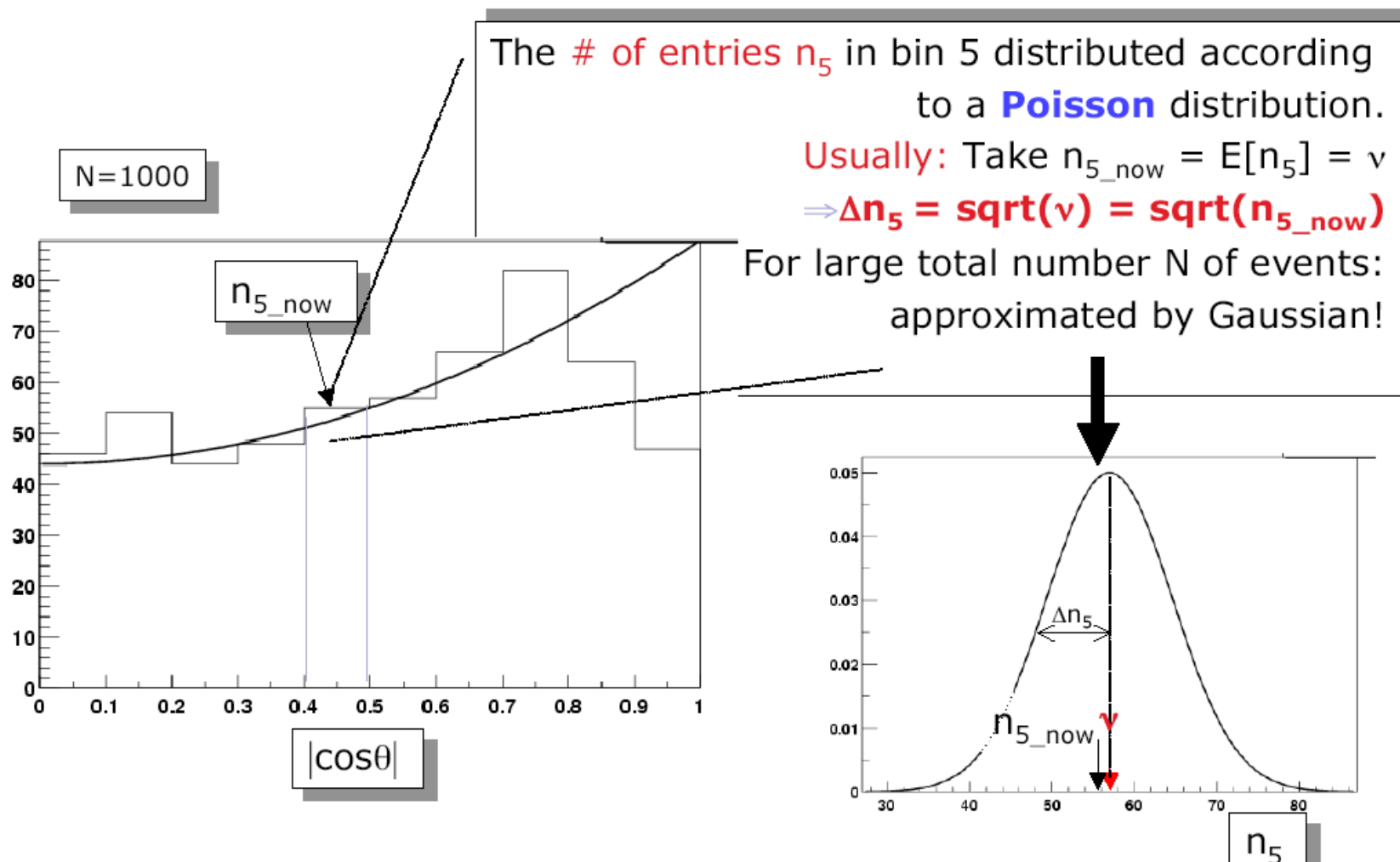
Lower edge: **0**

Upper edge: **1**

Ιστογράμματα

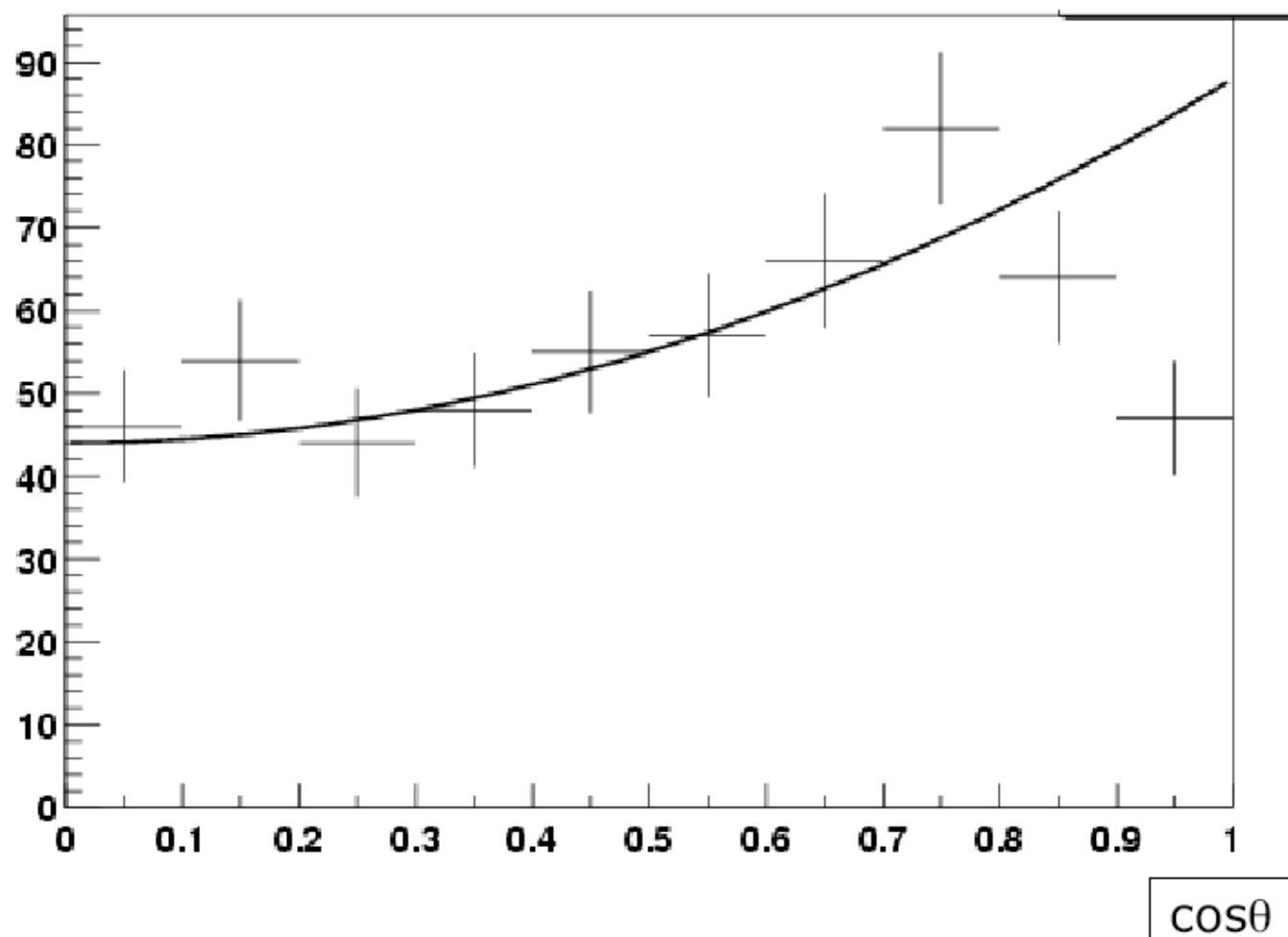
- Εάν N (entries) μεγάλος αριθμός το ιστόγραμμα αντιπροσωπεύει **κατανομή πιθανότητας (pdf)**
- Ενδιαφέρει το πλήθος των γεγονότων σε κάθε bin και το σφάλμα
- Η κανονικοποίηση του ιστογράμματος
- Η μέση τιμή του ιστογράμματος

Σφάλματα στο Ιστόγραμμα



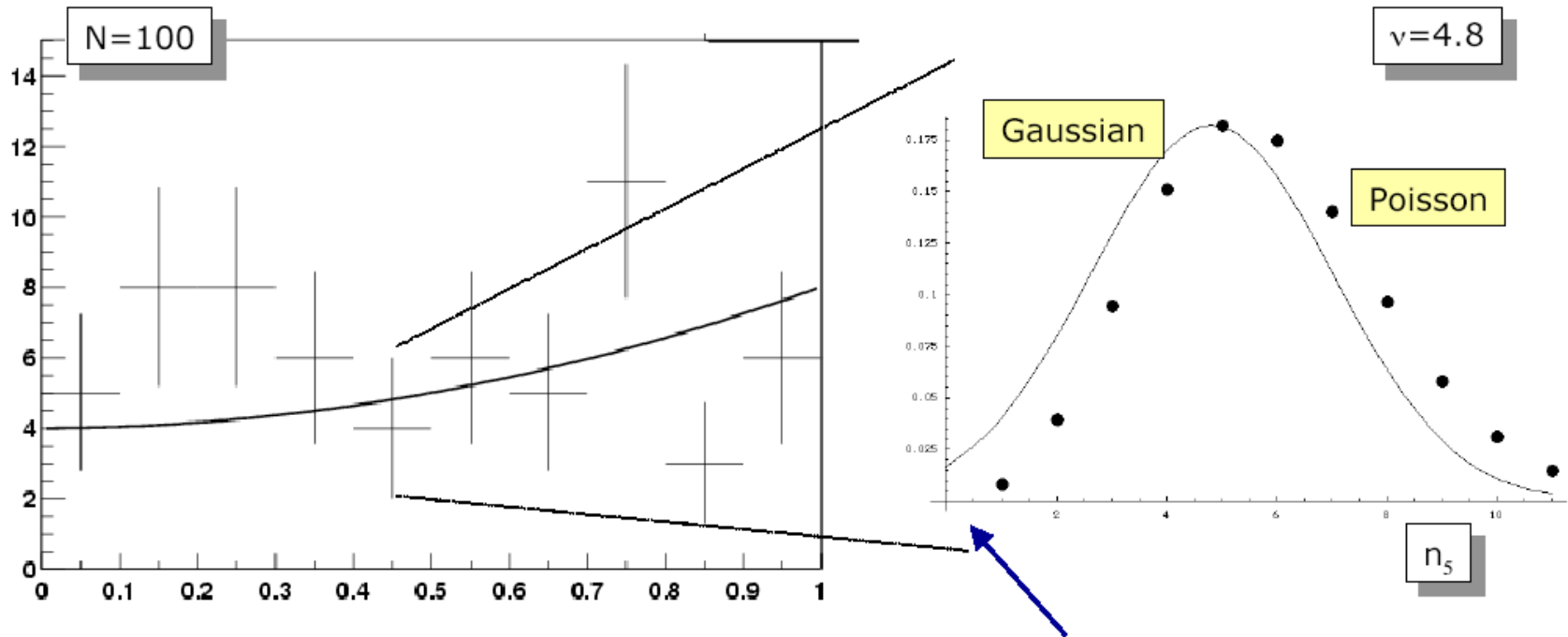
Σφάλματα στο Ιστόγραμμα

- Προσθέτουμε **error bars** στο ιστόγραμμα



Σφάλματα στο Ιστόγραμμα

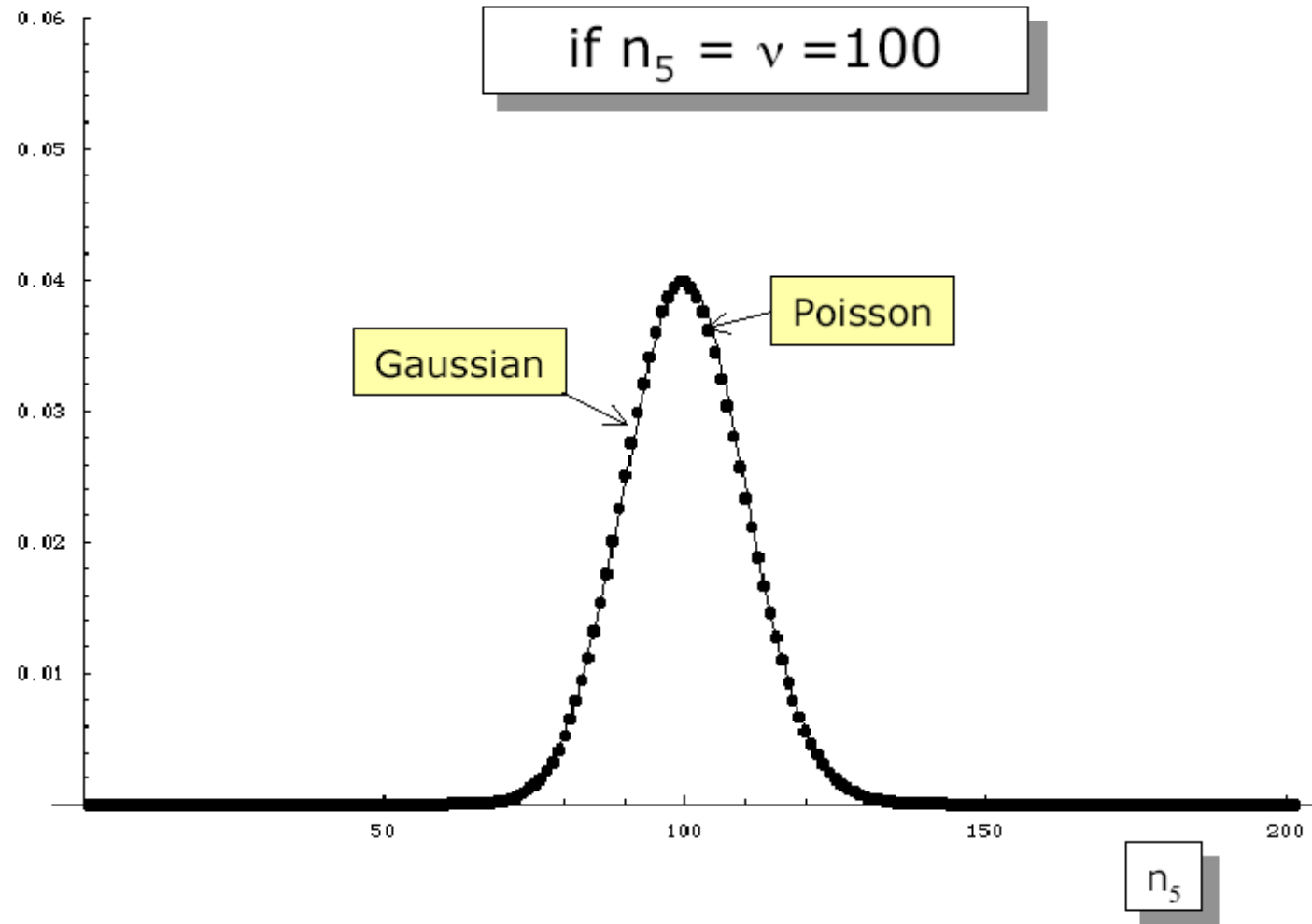
- Προσοχή! Περίπτωση μικρού πλήθους γεγονότων



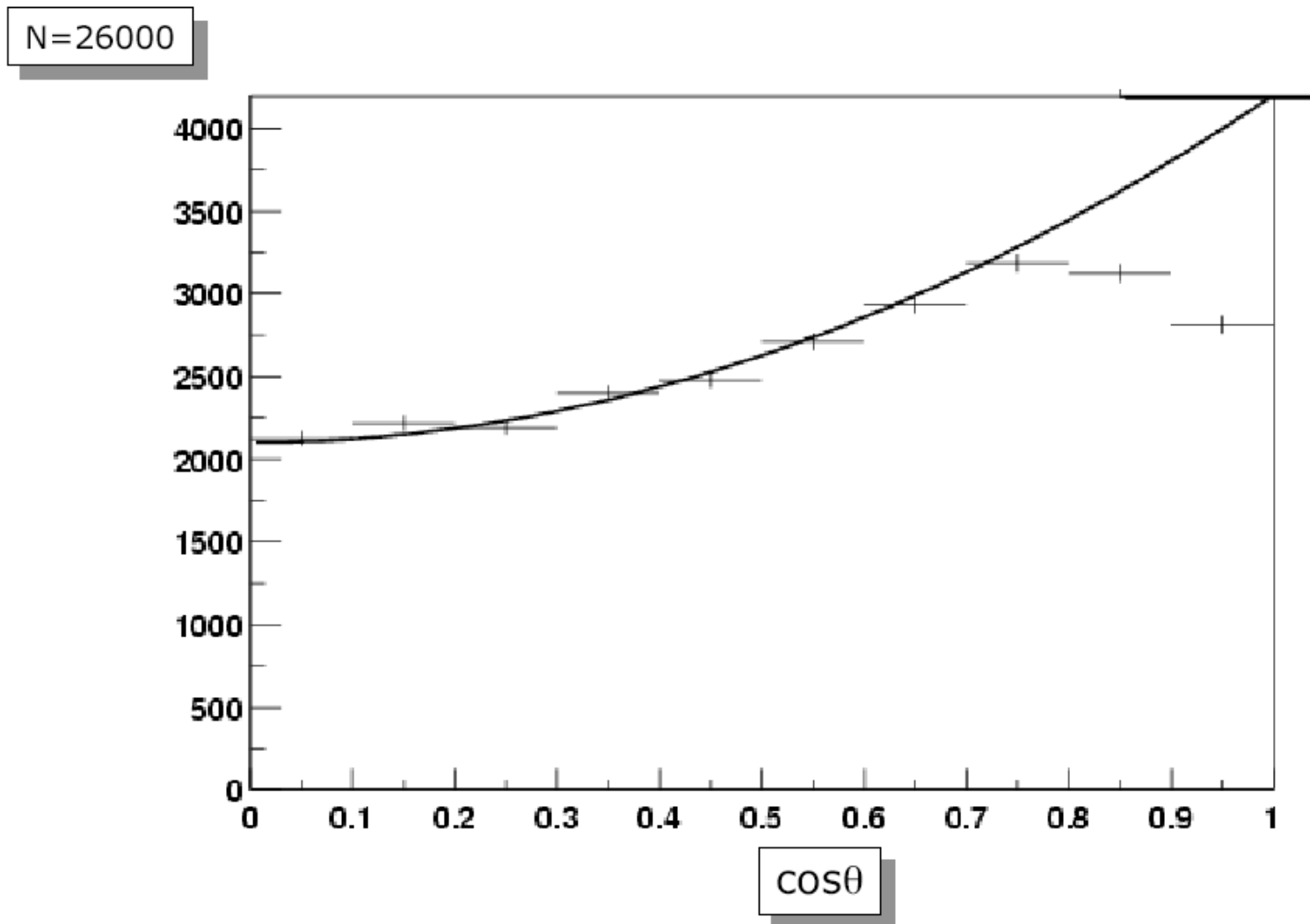
Prob(to see 0) $\neq$ 0 for Gaussian

Prob(to see 0) = 0 for Poisson !!

Poisson vs Gaussian



Σφάλματα στο Ιστόγραμμα



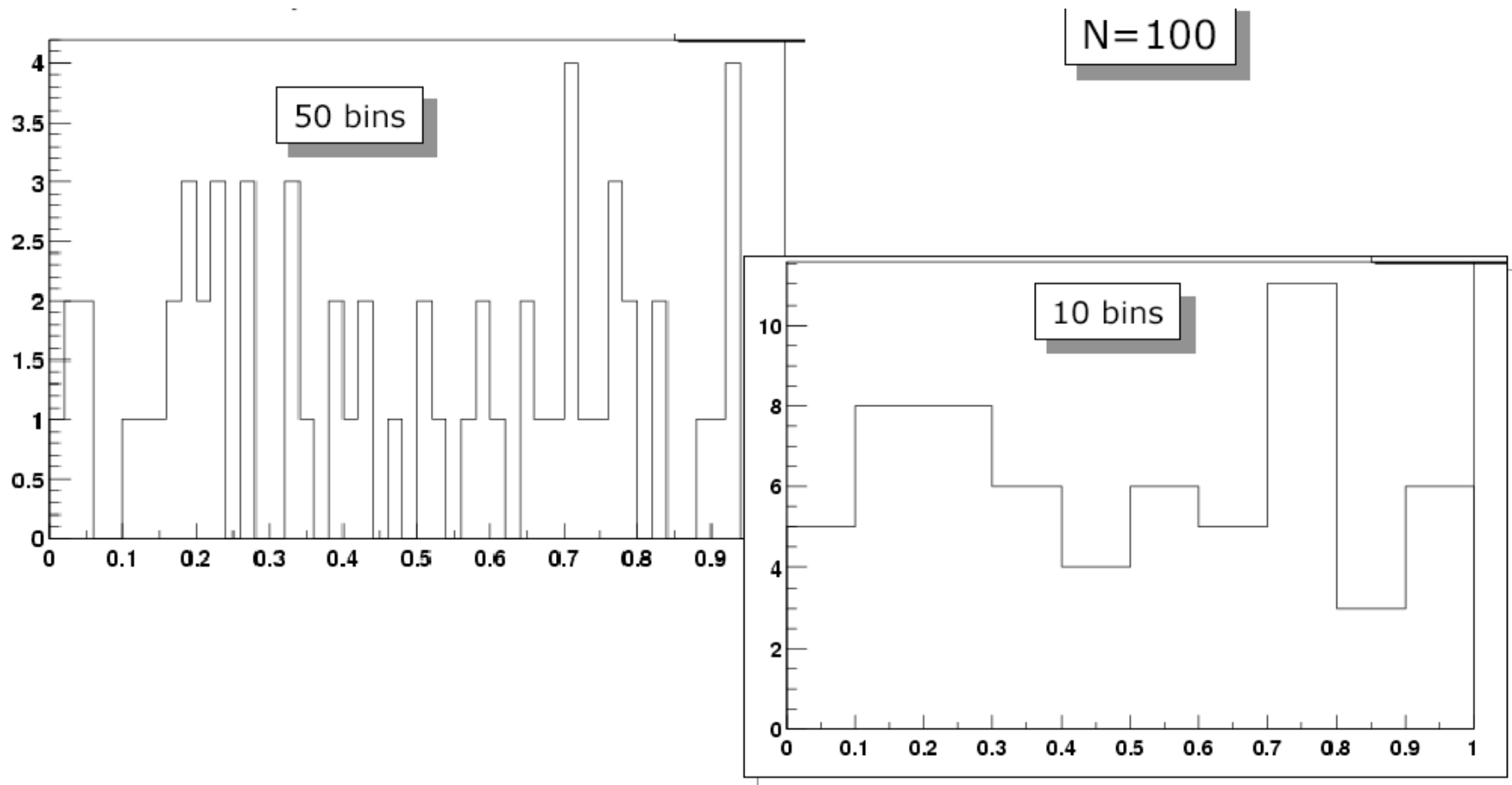
Ιστογράμματα

Τι θα πρέπει να προσέχουμε?

- Το μέγεθος του bin (**bin size**)
- Γεγονότα εκτός των ορίων του ιστογράμματος (**underflows, overflows**)
- Κατανομές με γρήγορες μεταβολές (**steeply falling/fast varying functions**)

Επιλογή του bin width

- Επιλέγουμε την υποδιαίρεση της κλίμακας (bin width) έτσι ώστε το πλήθος των γεγονότων ανά υποδιαίρεση να είναι 'λογικό'

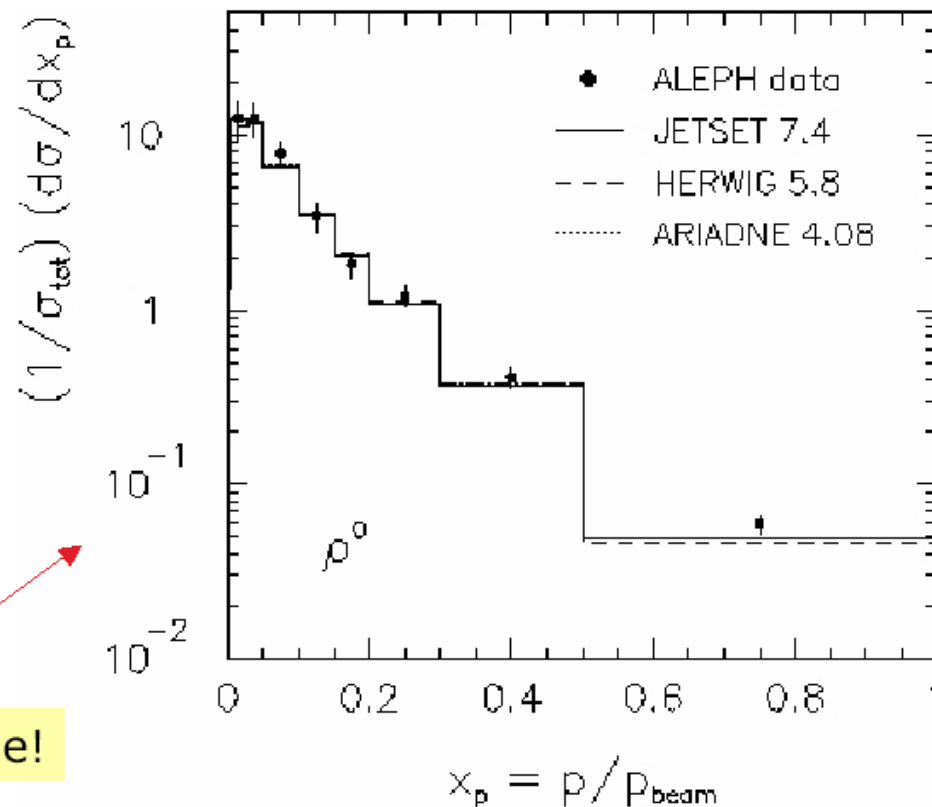


Επιλογή του bin width

- Ο αριθμός των γεγονότων που μετατοπίζονται εντός και εκτός της υποδιαίρεσης πρέπει να είναι παρόμοιος (**bin migration**–σταθερότητα και καθαρότητα του δείγματος)
- Το bin width πρέπει να ταιριάζει με την πειραματική διακριτική ικανότητα στην συγκεκριμένη μεταβλητή (**experimental resolution**)
- Να υπάρχει αρκετή στατιστική για κάθε υποδιαίρεση (**bin statistics**)

Επιλογή του bin width

- Σημαντικό για συναρτήσεις με απότομη κλίση
- Παράδειγμα: η κατανομή της ορμής του σωματιδίου-συντονισμού ρ



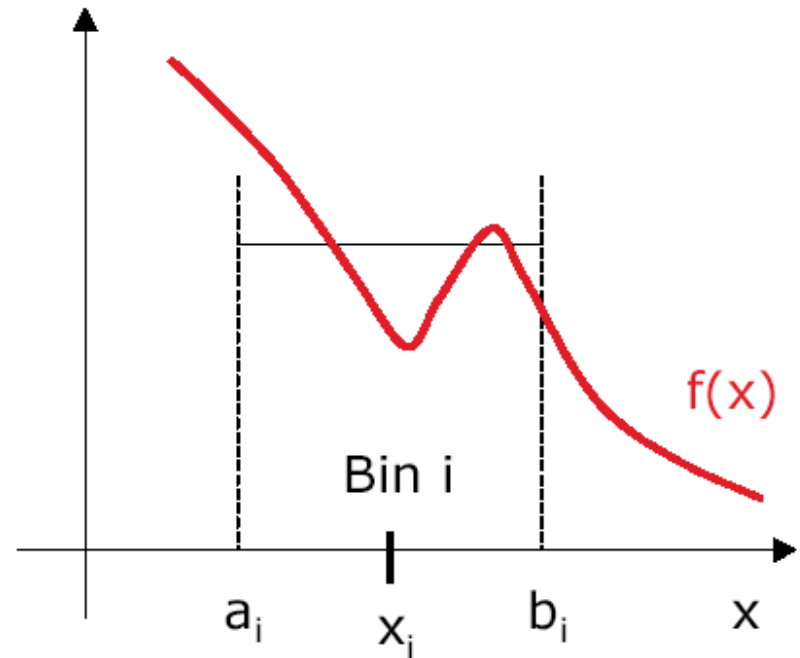
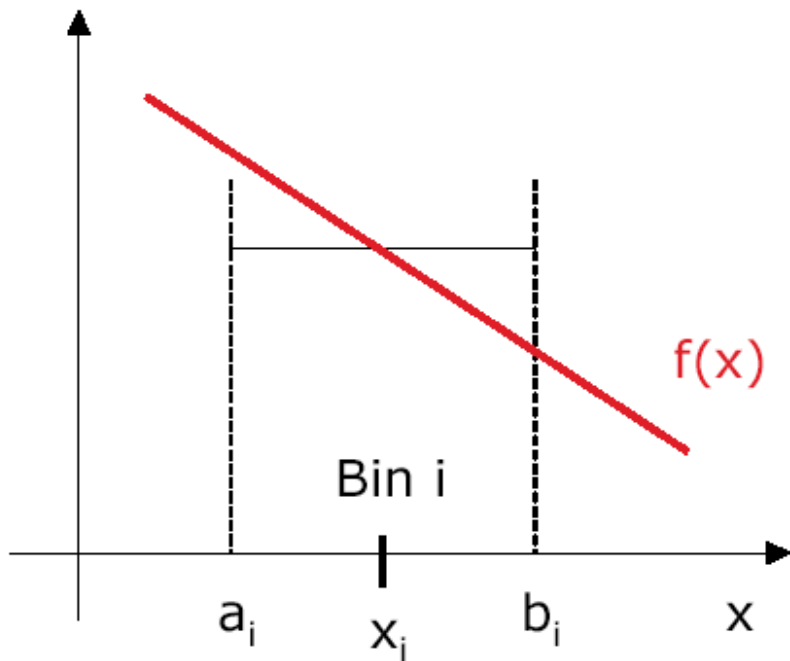
logarithmic scale!

Επιλογή της κλίμακας (Histogram range)

- Κακή επιλογή κλίμακας μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα
- =>Πρέπει να ελέγχουμε να μην υπάρχουν **overflows** και **underflows**
- Η κλίμακα είναι σημαντική και στην περίπτωση κανονικοποίησης

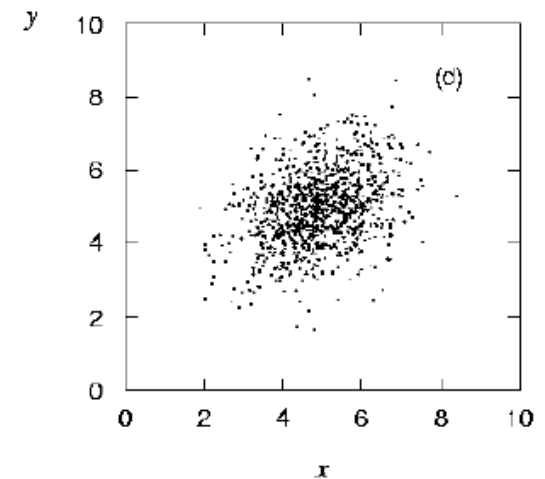
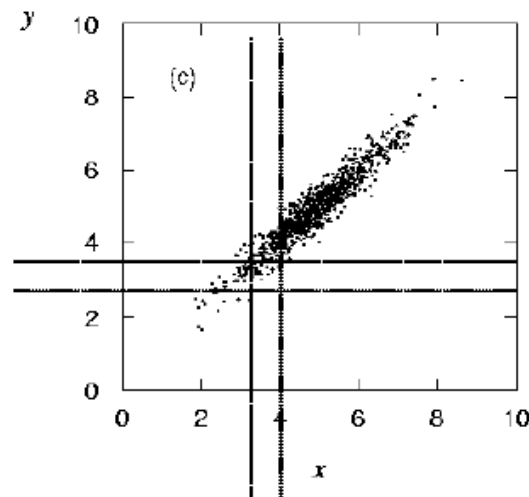
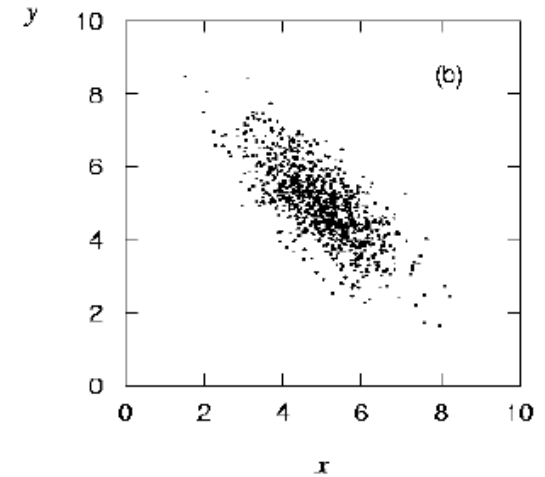
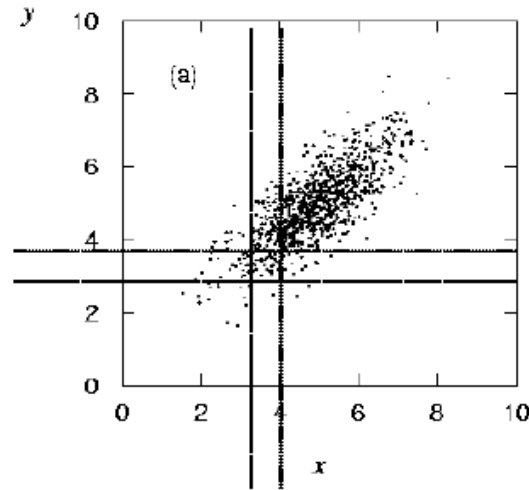
Σύγκριση ιστογράμματος με ομαλή συνάρτηση

- Προσοχή στις συναρτήσεις με απότομες ή γρήγορες μεταβολές



Συσχετίσεις-Συμμεταβολή (correlations-covariance)

- Ιστογράμματα 2-διαστάσεων (**scatterplots**)
- =>Χρήσιμο εργαλείο για να μελετήσουμε συσχετισμούς δύο μεταβλητών (**correlations**)



Ιδιότητες-Συσχετισμοί (correlations)

- Για συναρτήσεις με διάφορες μεταβλητές μπορεί να υπάρχουν συσχετίσεις μεταξύ των μεταβλητών
- Παράδειγμα : για συναρτήσεις δύο τυχαίων μεταβλητών x, y η συμμεταβολή (covariance) $V(x,y)$ (ή $\text{cov}(x,y)$) -error matrix- ορίζεται ως:

$$V[xy] = E((x - \mu_x)(y - \mu_y)) \equiv E[xy] - \mu_x \mu_y$$

$$E[xy] = \int_{y \min}^{y \max} \int_{x \min}^{x \max} x \cdot y \cdot f(x, y) \cdot dx \cdot dy$$

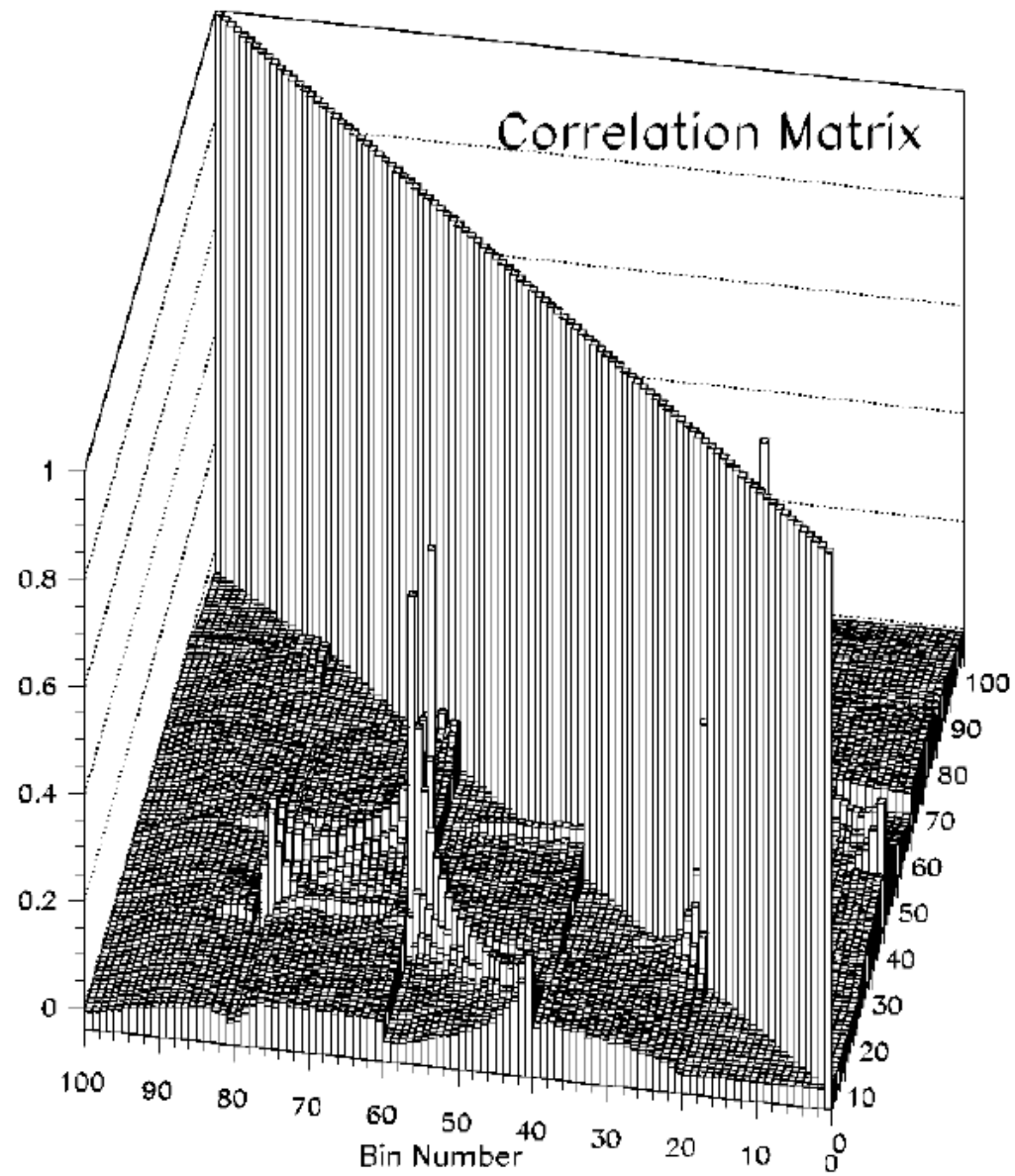
$$V(x, y) = \begin{pmatrix} \sigma_x^2 & V[xy] \\ V[yx] & \sigma_y^2 \end{pmatrix}$$

- Ο παράγοντας συσχέτισης ορίζεται ως:

$$\rho_{xy} = \frac{V[xy]}{\sigma_x \sigma_y}$$

$$-1 \leq \rho_{xy} \leq +1$$

$\rho=1$ για $y=ax+b$



Ιδιότητες-Συσχετισμοί (correlations)

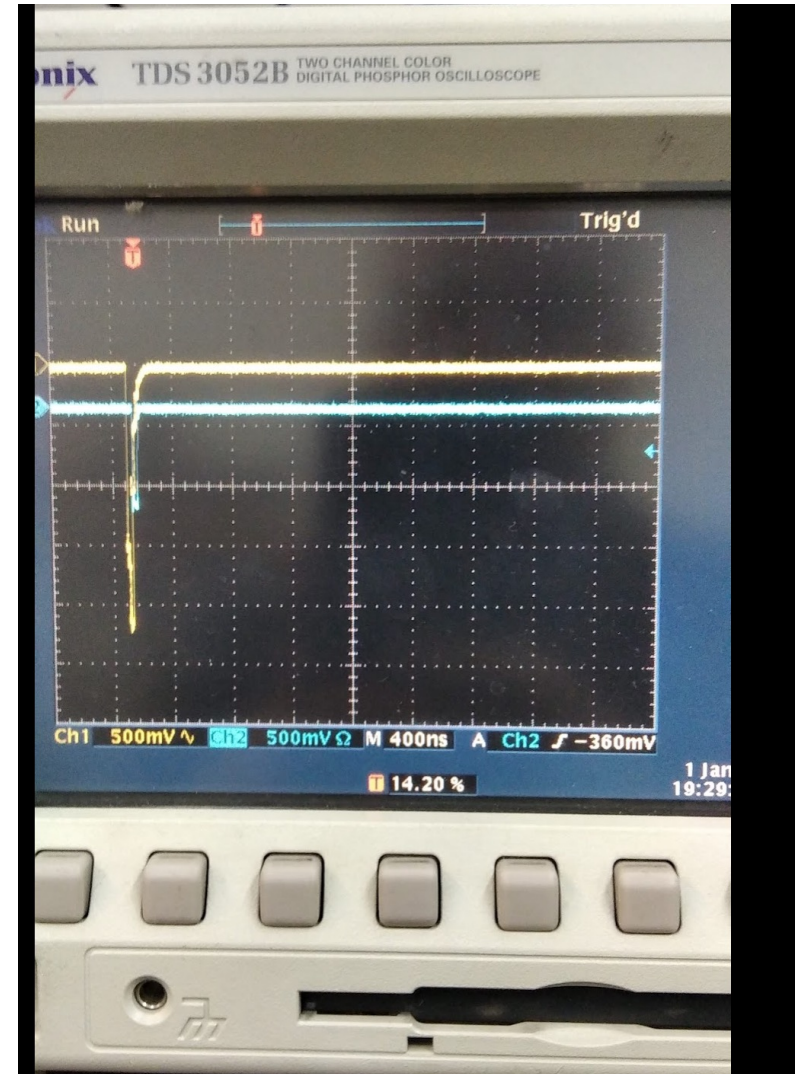
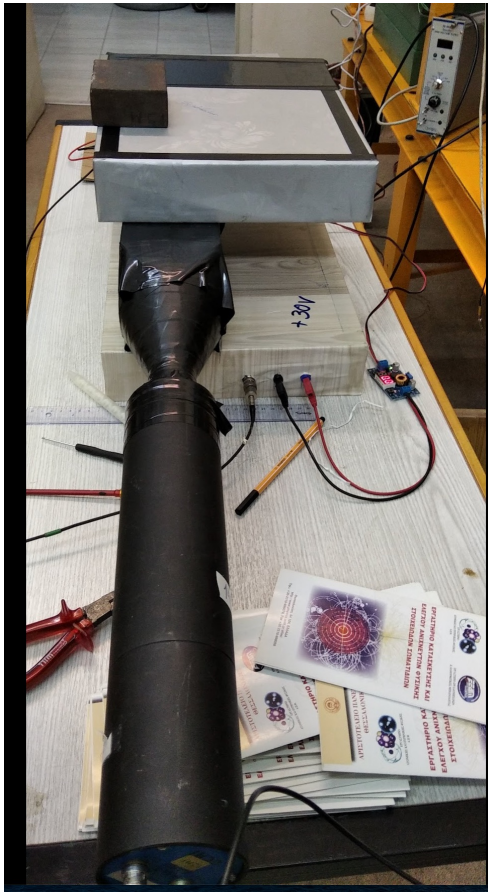
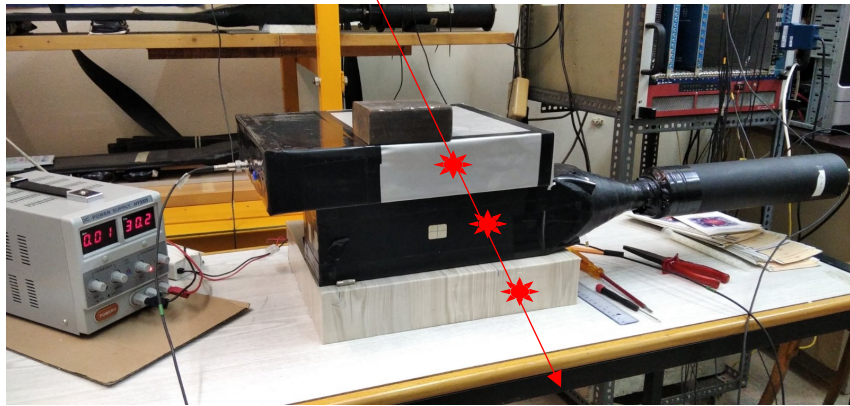
- Υπολογισμός του σφάλματος για συναρτήσεις με συσχετιζόμενες μεταβλητές: **παίρνουμε υπόψη την συσχέτιση**

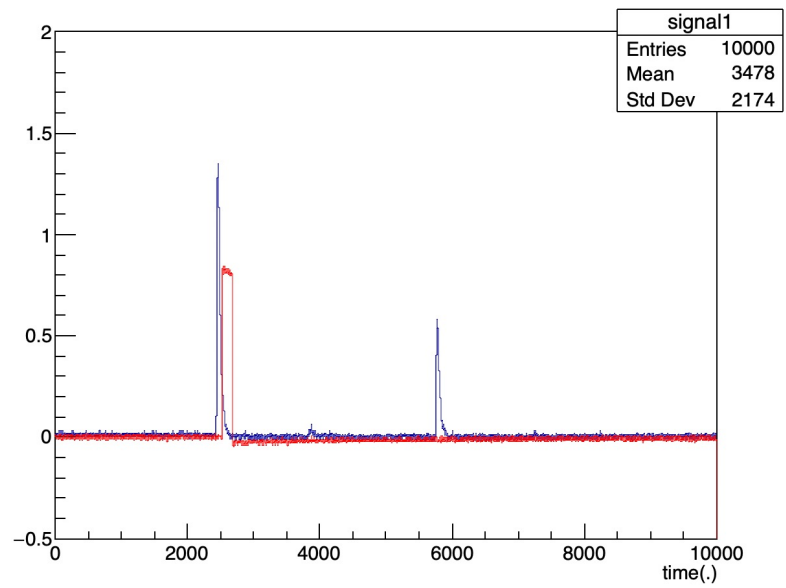
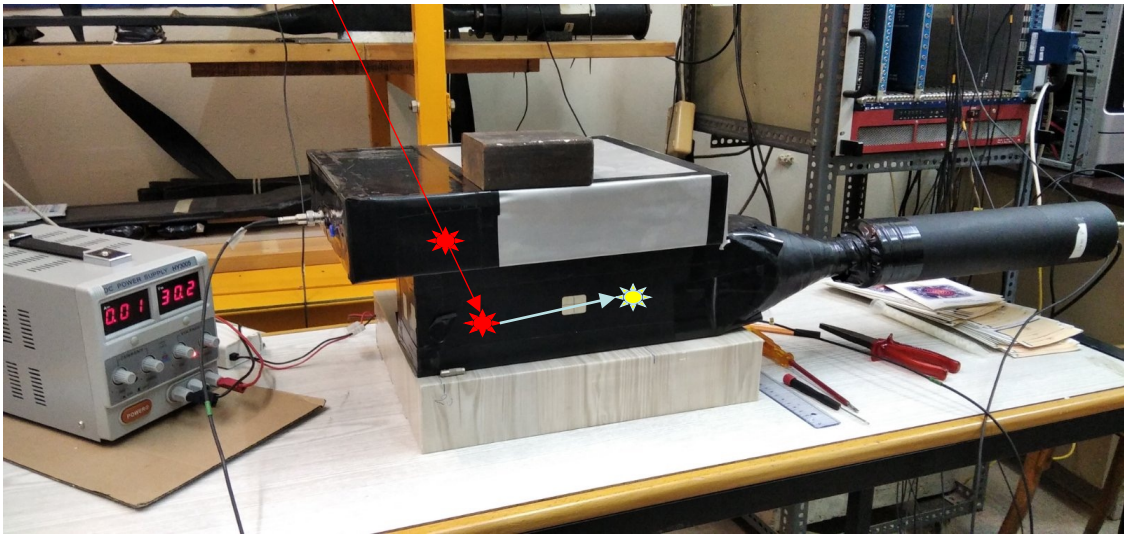
$$\sigma_f^2 = \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \right)^2 \cdot \sigma_{x_i}^2 + \sum_{i=1}^n \sum_{j \neq i}^n \left(\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \right) \cdot \text{cov}(x_i, x_j)$$

Ασκηση

Ένα μιονίο κοσμικής ακτινοβολίας δημιουργείται σε ύψος 30 km στην ατμόσφαιρα, ταξιδεύει στο επίπεδο της θάλασσας και σταματά σε ένα μπλοκ σπινθηριστή, δίνοντας ένα σήμα εκκίνησης (start) τη χρονική στιγμή t_0 . Τη στιγμή t διασπάται σε ηλεκτρόνιο και δύο νετρίνα που δίνει σήμα διάσπασης (stop). Ποια είναι η κατανομή της διαφοράς μεταξύ των χρόνων στάσης και έναρξης, δηλαδή η pdf του $(t - t_0)$ δεδομένου $t > t_0$

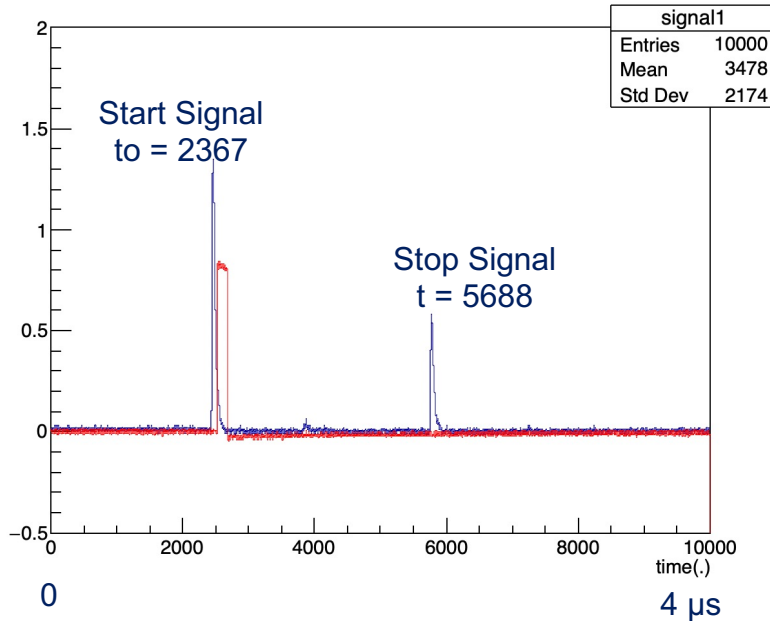
Με τα δεδομένα του πειράματος στο αρχείο allpeaks.txt να βρεθεί ο χρόνος ζωής του μιονίου.





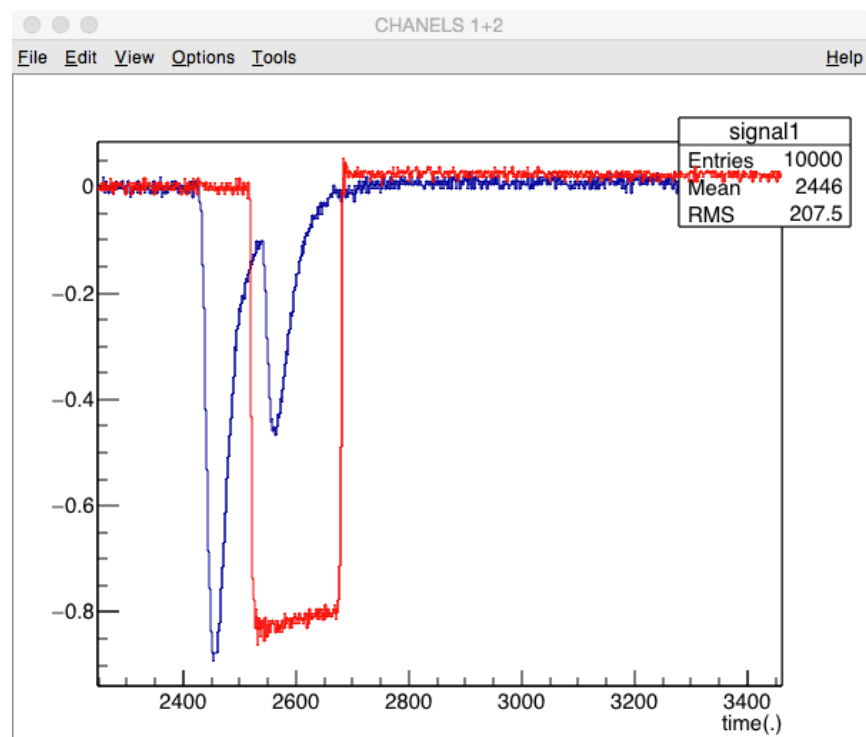
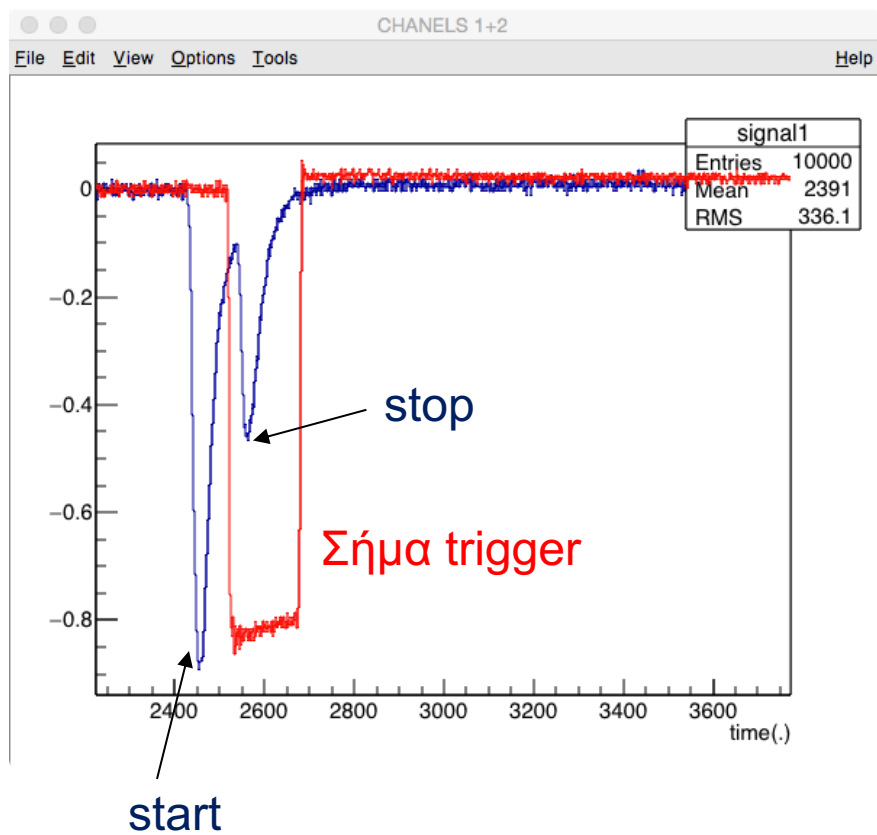
Data

Καταγραφή στον παλμογράφο



event #	# of peaks	t0	Vpeak	t	Vpeak	t2	Vpeak
185	2	1126	3.56	8163	1.28		
336	2	1125	3.56	4771	2.58		
738	2	1129	1.84	3401	0.68		
1090	2	1129	3.56	7069	1.1		
1181	2	1132	3.56	1798	0.88		
1183	2	1132	2.2	3227	1.92		
1337	2	1126	3.56	9556	0.54		
1452	2	1130	2.48	3571	1.78		
1571	2	1134	2.88	2756	1.78		
1578	2	1127	1.46	6843	1.24		
2065	2	1125	3.56	4561	1.62		
2670	2	1131	1.2	1285	0.68		
3227	2	1132	3.56	4248	1.48		
3312	2	1124	3.56	3824	2.46		
3490	2	1128	3.56	1305	1.86		
3510	3	1123	3.56	2595	0.42	3704	0.36
3547	2	1123	3.56	3777	1.9		
3548	2	1129	1.76	10000	0.56		
3706	2	1131	3.26	4281	2.1		

Το παράθυρο του παλμογράφου είναι 4 μs και ψηφιοποιείται με ρυθμό 10 kHz



17527, 17505

