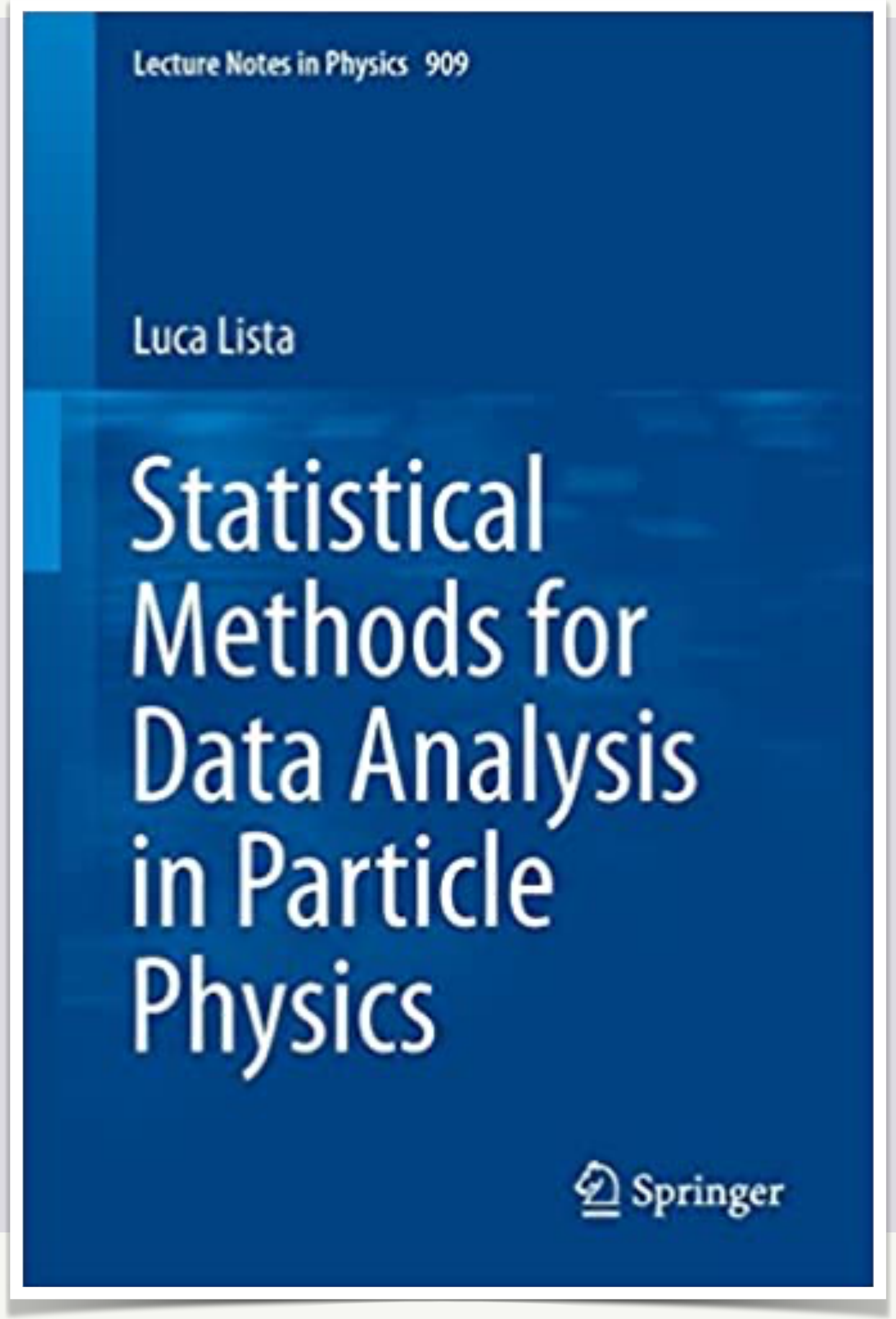
Data Analysis in HEP

Statistics - Part2

Sandro Fonseca de Souza
Eliza Melo Da Costa
Sheila Mara do Amaral

29 Sep 2020

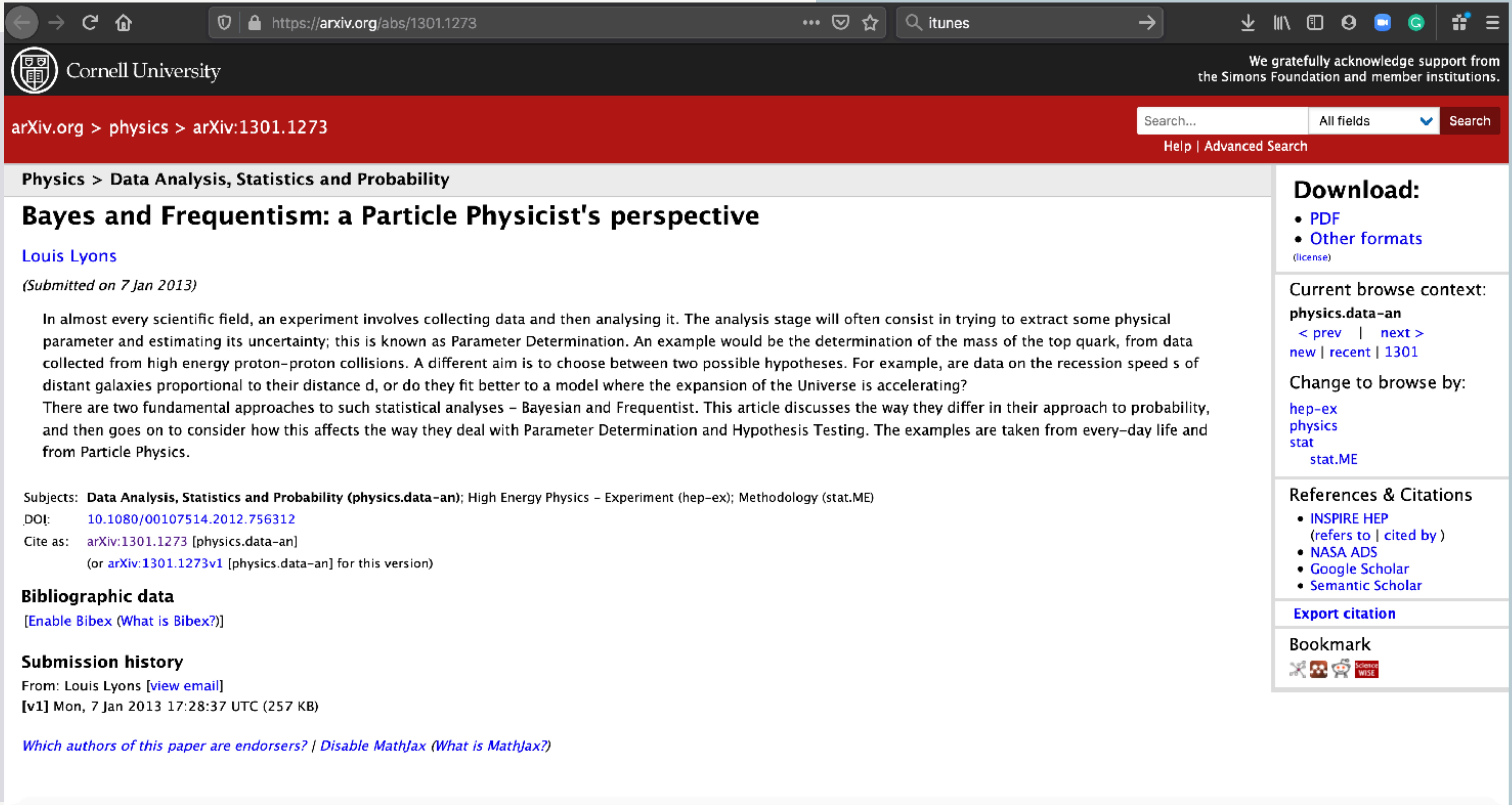
Bibliography



http://dfnae.fis.uerj.br/~oguri/trata.html

Based on material by Stefan Schmitt and Roger Wolf

Paper suggestion



https://arxiv.org/abs/1301.1273 | itunes

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > physics > arXiv:1301.1273

Search... | All fields | Search

Help | Advanced Search

Physics > Data Analysis, Statistics and Probability

Bayes and Frequentism: a Particle Physicist's perspective

Louis Lyons

(Submitted on 7 Jan 2013)

In almost every scientific field, an experiment involves collecting data and then analysing it. The analysis stage will often consist in trying to extract some physical parameter and estimating its uncertainty; this is known as Parameter Determination. An example would be the determination of the mass of the top quark, from data collected from high energy proton-proton collisions. A different aim is to choose between two possible hypotheses. For example, are data on the recession speed s of distant galaxies proportional to their distance d, or do they fit better to a model where the expansion of the Universe is accelerating?
There are two fundamental approaches to such statistical analyses - Bayesian and Frequentist. This article discusses the way they differ in their approach to probability, and then goes on to consider how this affects the way they deal with Parameter Determination and Hypothesis Testing. The examples are taken from every-day life and from Particle Physics.

Subjects: **Data Analysis, Statistics and Probability (physics.data-an)**; High Energy Physics - Experiment (hep-ex); Methodology (stat.ME)
DOI: 10.1080/00107514.2012.756312
Cite as: arXiv:1301.1273 [physics.data-an]
(or arXiv:1301.1273v1 [physics.data-an] for this version)

Bibliographic data

[Enable Bibex (What is Bibex?)]

Submission history

From: Louis Lyons [view email]
[v1] Mon, 7 Jan 2013 17:28:37 UTC (257 KB)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Download:
- PDF
- Other formats

(license)

Current browse context:
physics.data-an
< prev | next >
new | recent | 1301

Change to browse by:
hep-ex
physics
stat
stat.ME

References & Citations
- INSPIRE HEP
(refers to | cited by)
- NASA ADS
- Google Scholar
- Semantic Scholar

Export citation

Bookmark

https://arxiv.org/abs/1301.1273

Outline

1) Some examples distribution useful for Physics

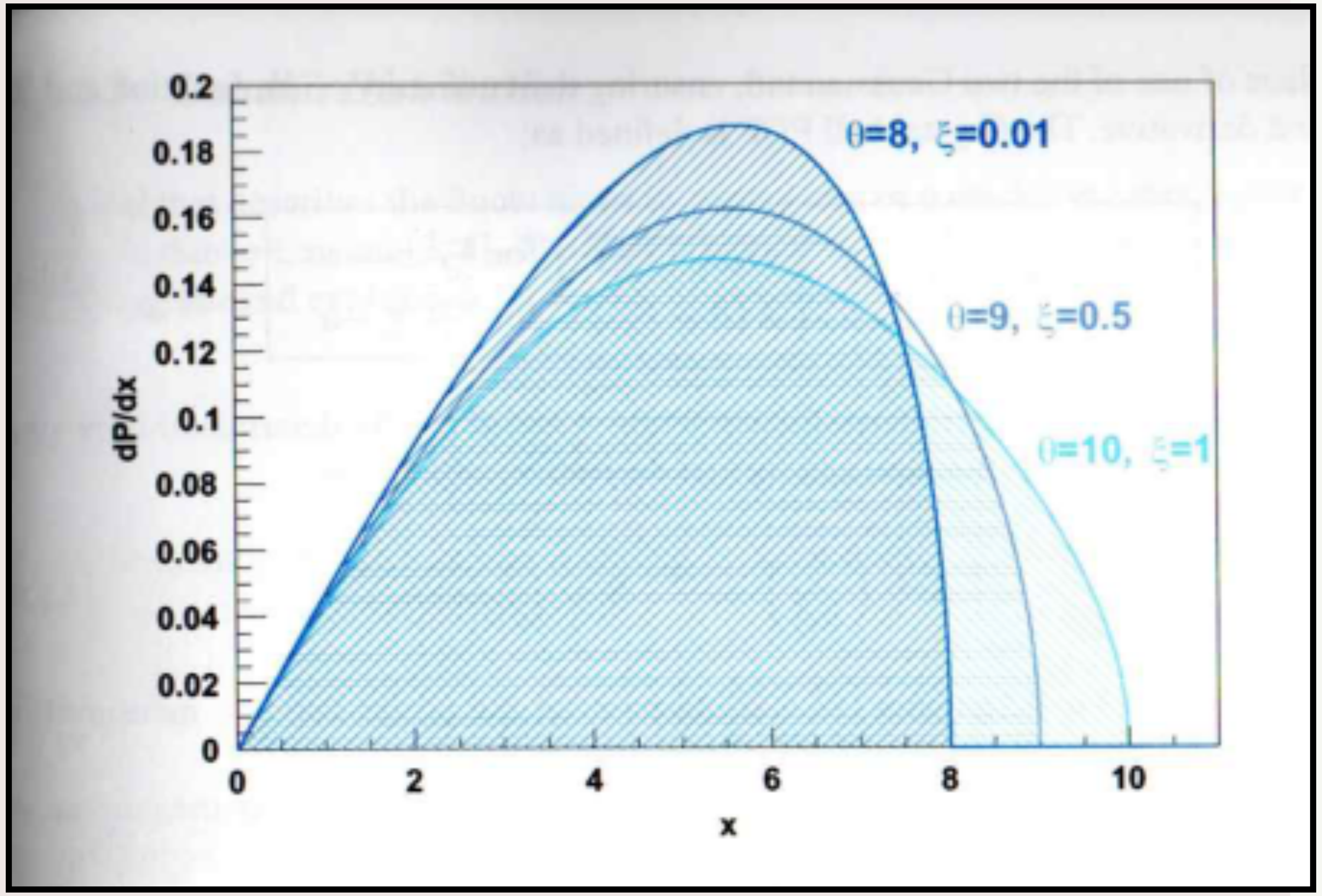
2) Bayesian and Frequentist statistics

Examples of Distributions

- Argus function

$$A(x;\theta;\xi) = Nx\sqrt{1-\left(\frac{x}{\theta}\right)^2}\, e^{-\frac{1}{2}\xi^2\left[1-\left(\frac{x}{\theta}\right)^2\right]},$$

ARGUS (A Russian-German-United States-Swedish Collaboration; later joined by Canada and the former Yugoslavia) was a particle physics experiment that ran at the electron-positron collider ring *DORIS II* at DESY. Its aim was to explore properties of charm and bottom quarks. Its construction started in 1979, the detector was commissioned in 1982 and operated until 1992.[1]



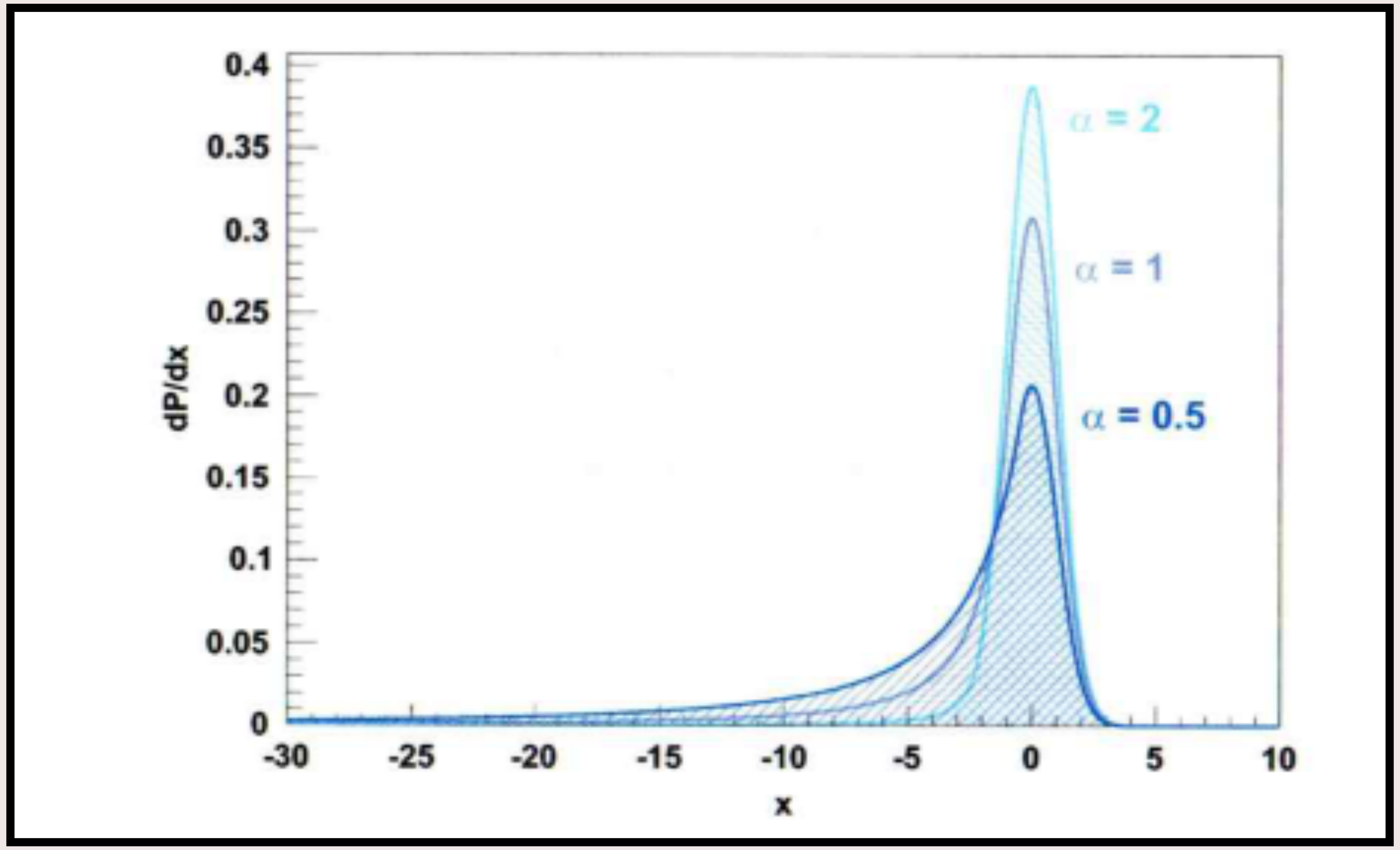
https://en.wikipedia.org/wiki/ARGUS_(experiment)

Examples of Distributions

- Crystall Ball function

$$CB(x;\alpha,n,\bar{x},\sigma) = N \cdot \begin{cases} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right) & \text{for } \frac{x-\bar{x}}{\sigma} > -\alpha \\ A\left(B - \frac{x-\bar{x}}{\sigma}\right)^{-n} & \text{for } \frac{x-\bar{x}}{\sigma} \leq -\alpha \end{cases},$$

The Crystal Ball was a hermetic particle detector used initially with the SPEAR particle accelerator at the Stanford Linear Accelerator Center beginning in 1979. It was designed to detect neutral particles and was used to discover the ηc meson..



https://en.wikipedia.org/wiki/Crystal_Ball_(detector)

Examples of Distributions

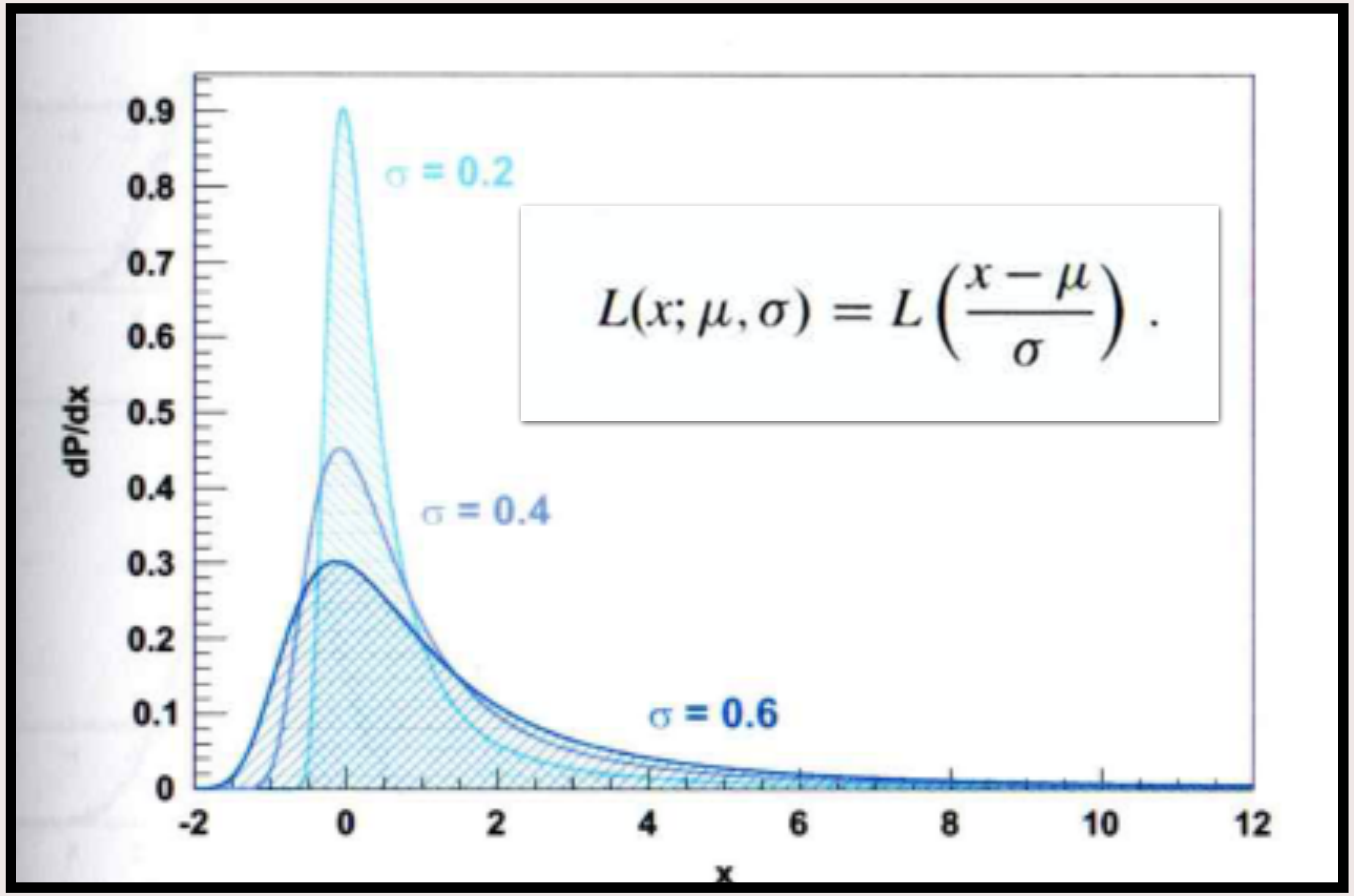
L. Landau, J. Phys. USSR 8 (1944) 201
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. 30 (1980) 253.

- Landau function

Describes energy loss of a charged particle in a thin layer of material

tail with large energy loss due to occasional creation of delta rays

$$L(x) = \frac{1}{\pi} \int_0^\infty e^{-t \log t - xt} \sin(\pi t)\, dt \,.$$



Outline

1) Some examples distribution useful for Physics
2) Bayesian and Frequentist statistics

Axiomatic Probability Definition

Axiomatic Probability Definition

Think of throwing a coin three times in a row:

- $\Omega = \{(\mathrm{Head}, \mathrm{Head}, \mathrm{Head}), (\mathrm{Tail}, \mathrm{Head}, \mathrm{Head}), \ldots\} \rightarrow$ (all possible outcomes) **sample space**.

- Each subset $A \subset \Omega$ is called **event**.

- The set of all possible events (σ-algebra) $\mathfrak{S}(\Omega)$.

- **Probability**:

  $\mathcal{P} : \mathfrak{S}(\Omega) \rightarrow \mathbb{R} \quad ; \quad A \rightarrow \mathcal{P}(A),$

  - Non-negative: $\mathcal{P}(A) \geq 0 \quad \forall A$
  - Linear : $\mathcal{P}(A \cup B) = \mathcal{P}(A) + \mathcal{P}(B) \quad \forall A \cap B = \emptyset$
  - Normalized : $\mathcal{P}(\mathfrak{S}(\Omega)) = 1$

Conditional Probability and Independent Events

Probability of event A given event $B(\neq \emptyset)$:

- $\mathcal{P}_B(A) = \mathcal{P}(A|B) = \frac{\mathcal{P}(A\cap B)}{\mathcal{P}(B)}$ $\rightarrow$ **(conditional probability)**.

Bayes theorem:

- $\mathcal{P}_A(B) \cdot \mathcal{P}(A) = \mathcal{P}_B(A) \cdot \mathcal{P}(B)$ $\rightarrow$ **(Bayes theorem)**.

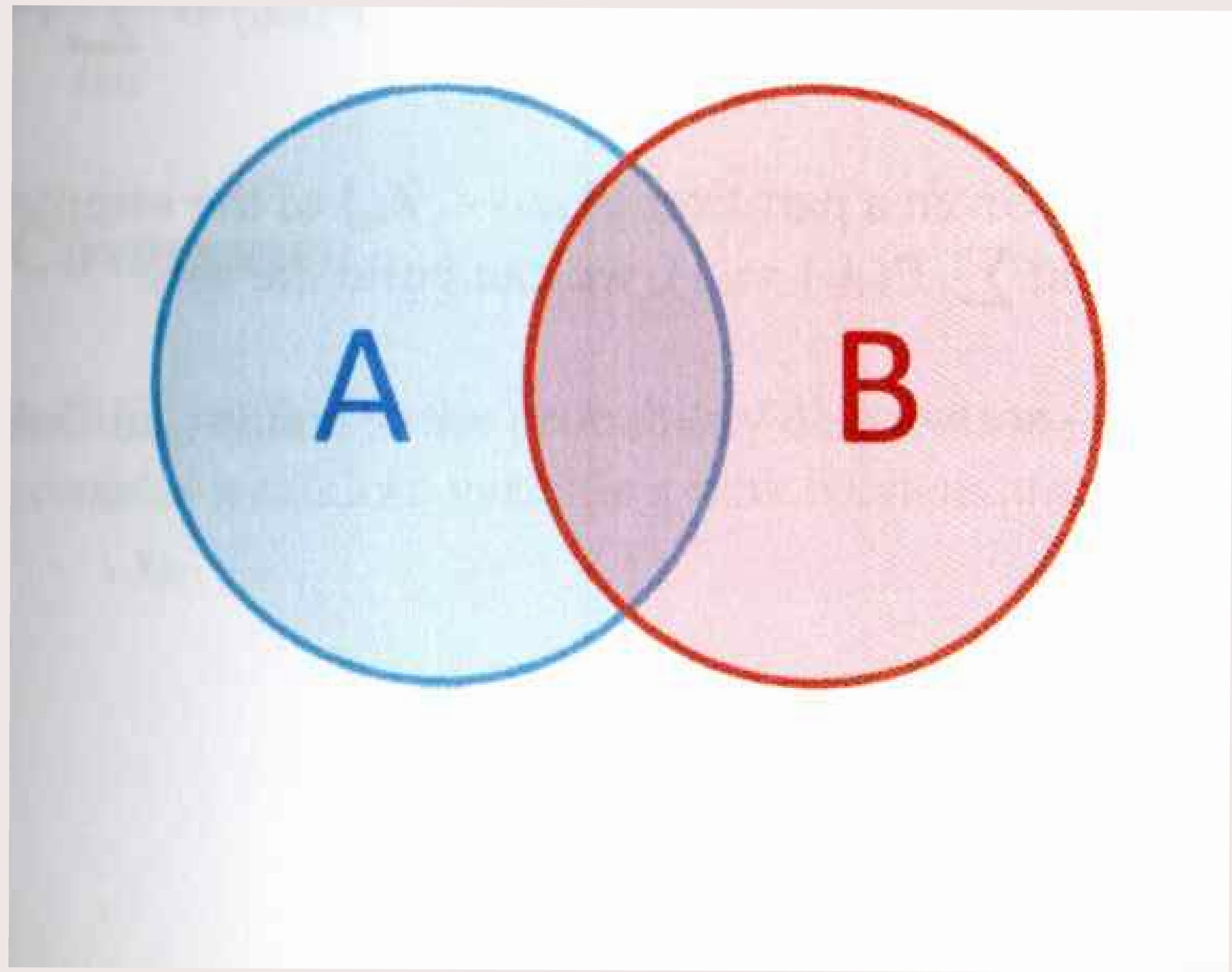
(Statistically) independent events:

- $\mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B)$ $\rightarrow$ **(statistical independent)**.

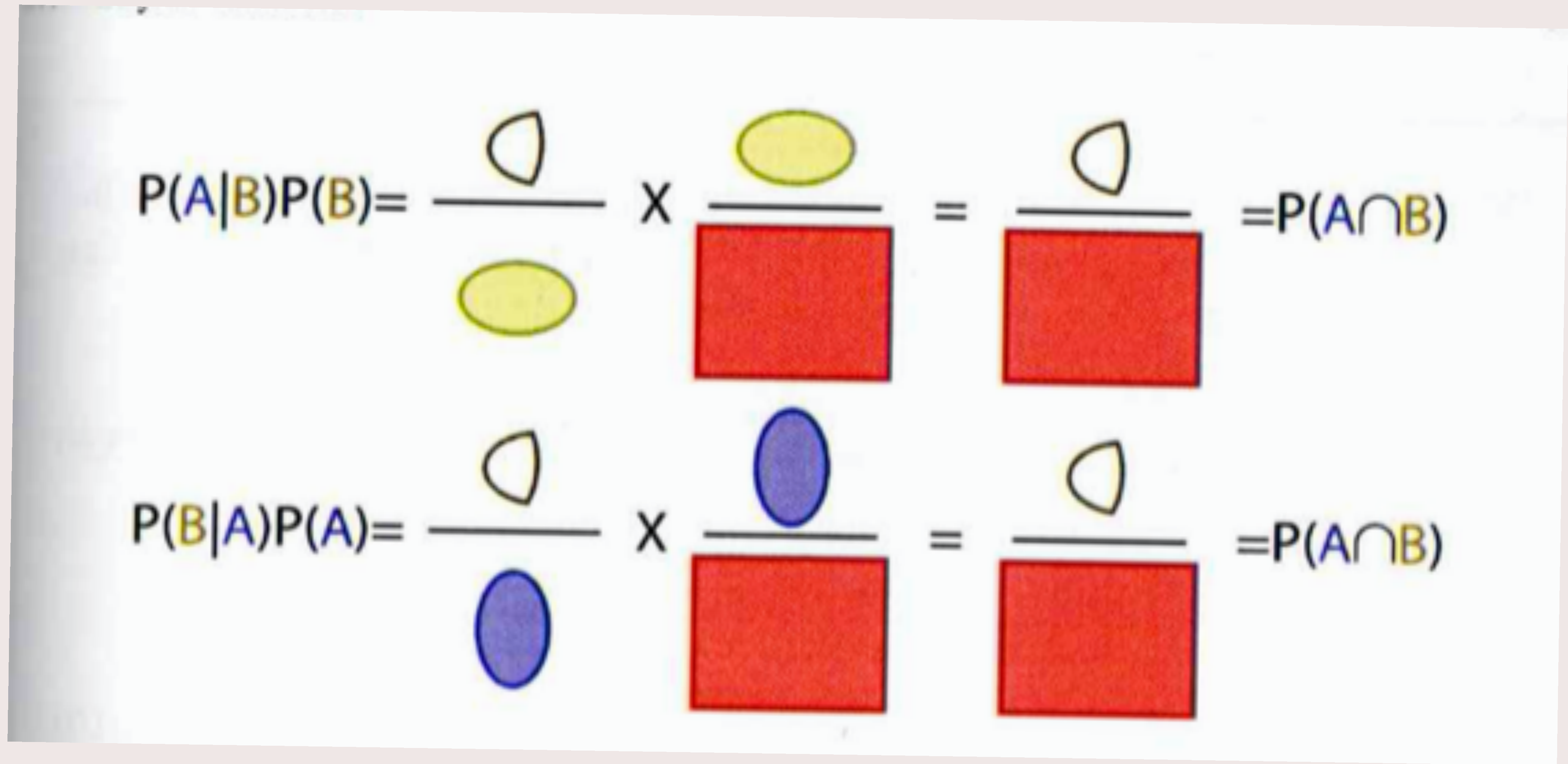
  $\mathcal{P}_A(B) = \mathcal{P}(B)$

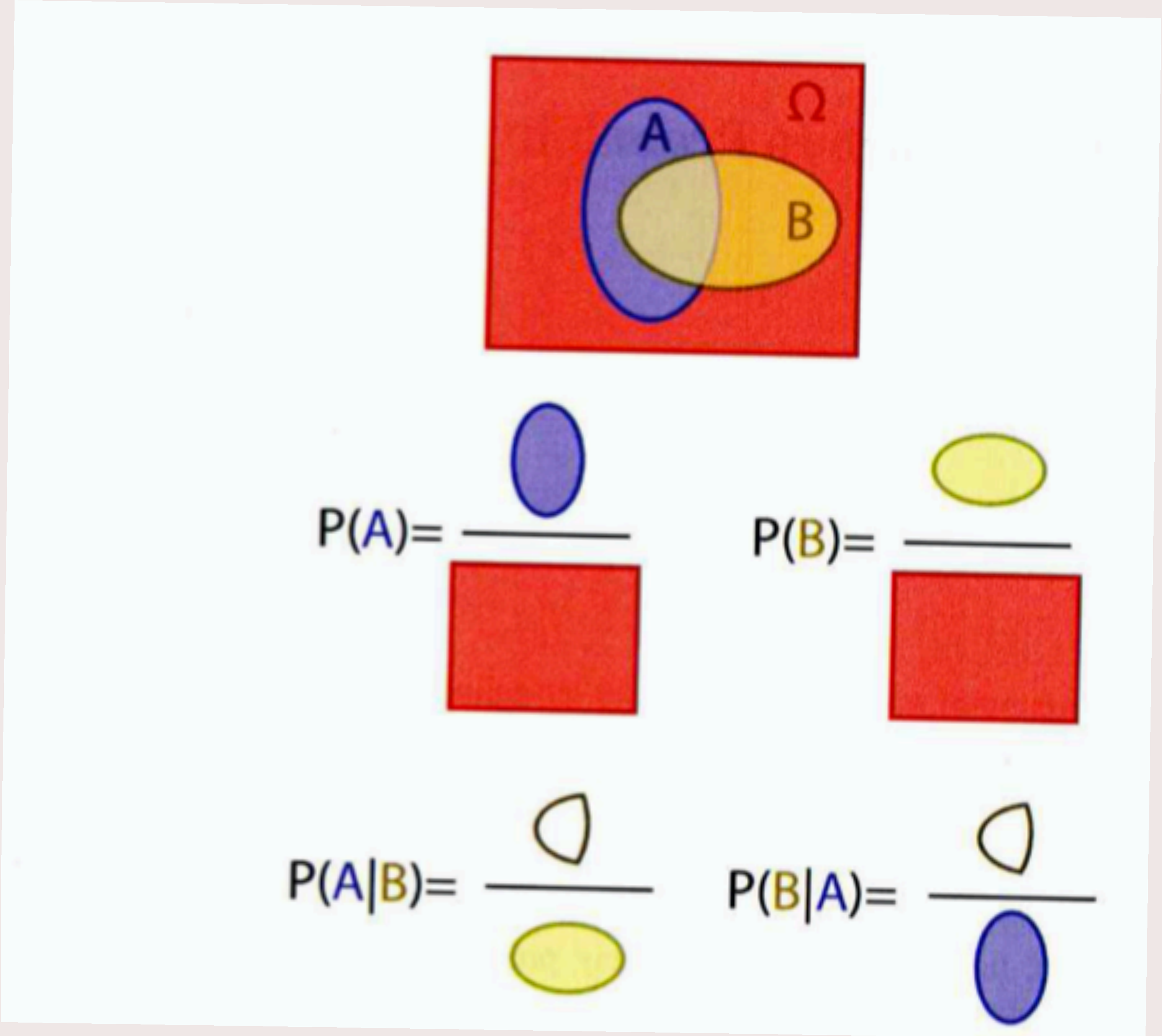
  $\mathcal{P}_B(A) = \mathcal{P}(A)$

- Particle physics is a unique field where statistical independence of event is perfectly fulfilled for an incredibly large sample space.



Bayesian Statistics Definition



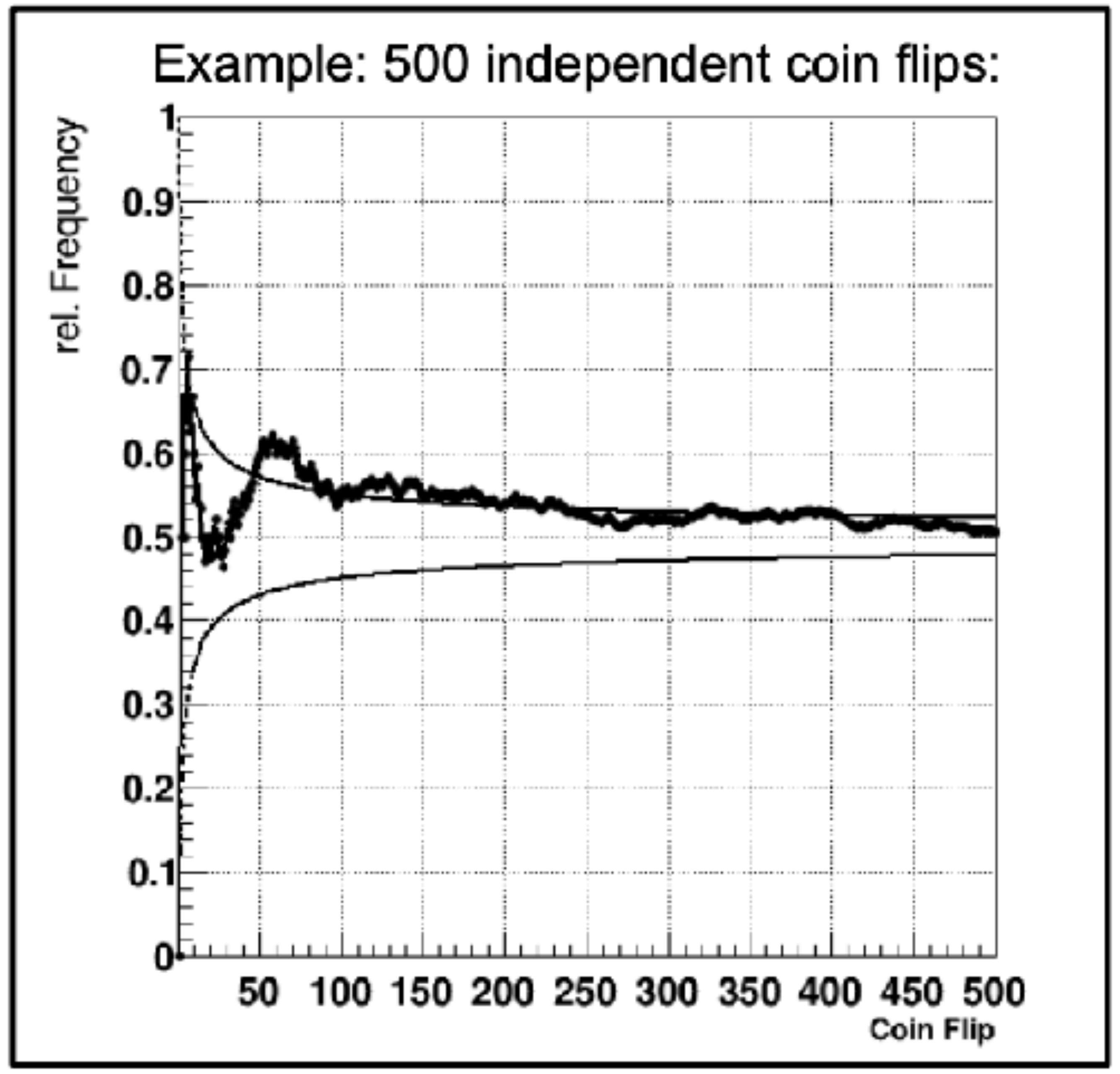


Exercises : you must redo all examples in pages 13- 27 of Vitor's book.

Interpretation Paradigms

- Mathematical results need to be interpreted:

- **Frequentist**:



Relative frequency converges to probability.

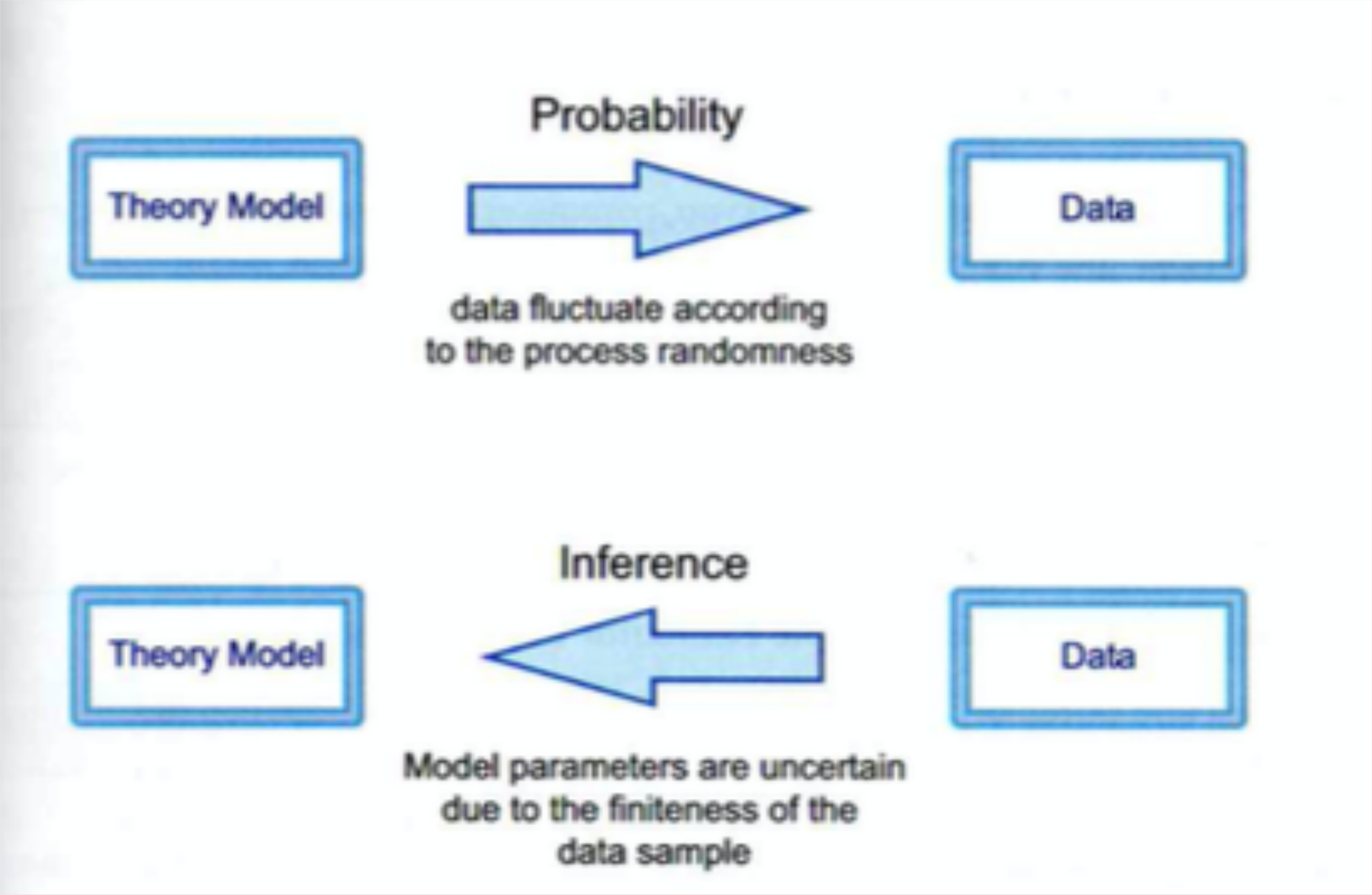
- **Bayesian**:

Quantification of my degree of belief that event A turns out to be true.

+ Makes sense also for "experiments", which can not be repeated.

- Requires reasonable implementation of probability distribution (usually coincides with Frequentist interpretation, where overlaps, but not always).

Relation between Probability and Inference



Bayesian Statistics and prior knowledge

- In Bayesian statistics: "my degree of belief" that event A turns out to be true depends on my prejudice ($\rightarrow$ prior knowledge):

$$\mathcal{P}_{\mathrm{Data}}(\mathrm{Model}) \propto \mathcal{P}_{\mathrm{Model}}(\mathrm{Data}) \cdot \mathcal{P}(\mathrm{Model}) \qquad \rightarrow$$ **(Bayes theorem).**

Posterior Probability

Likelihood

Probability of a given outcome of the experiment for a given model.

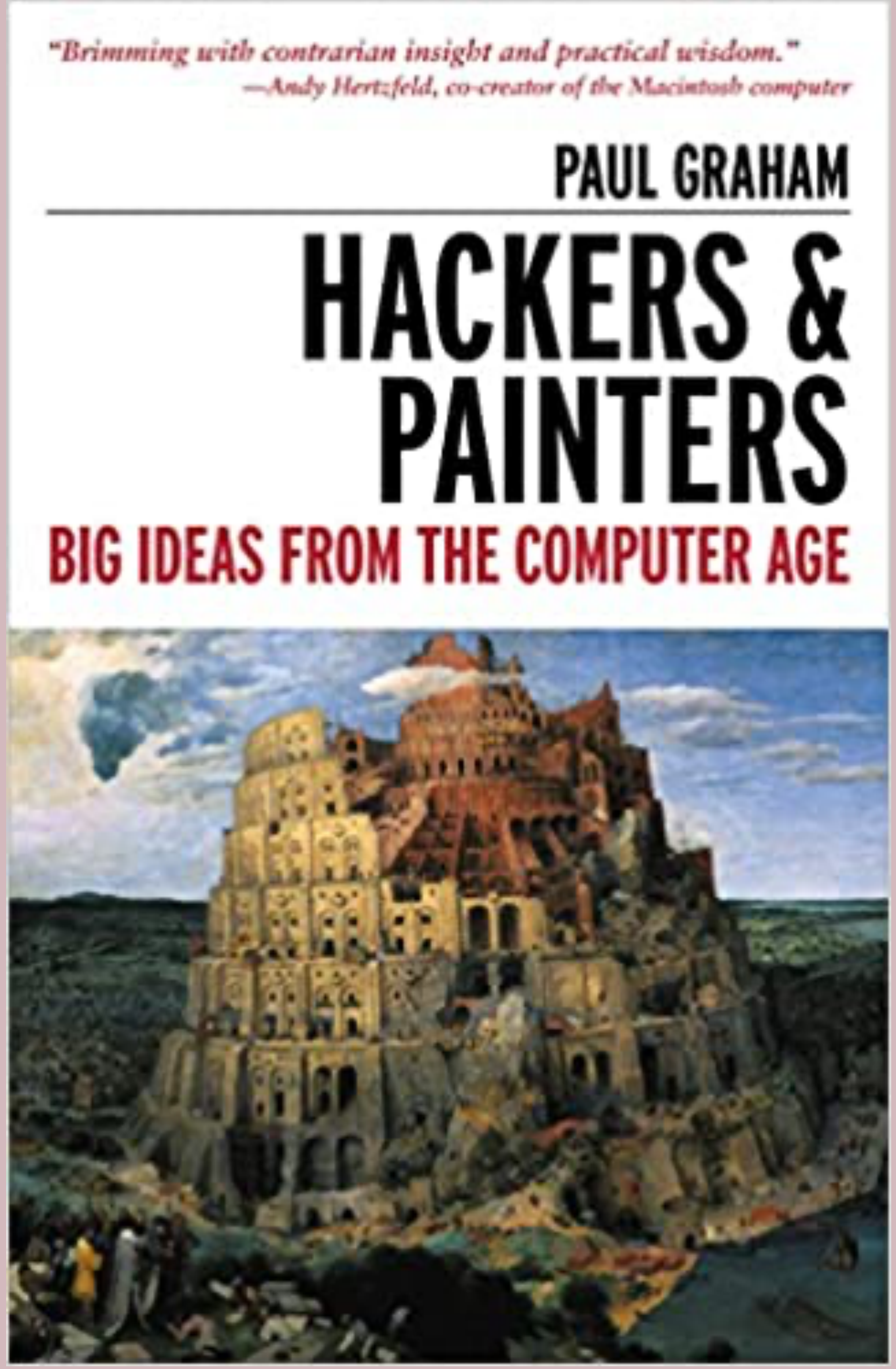
Prior probability

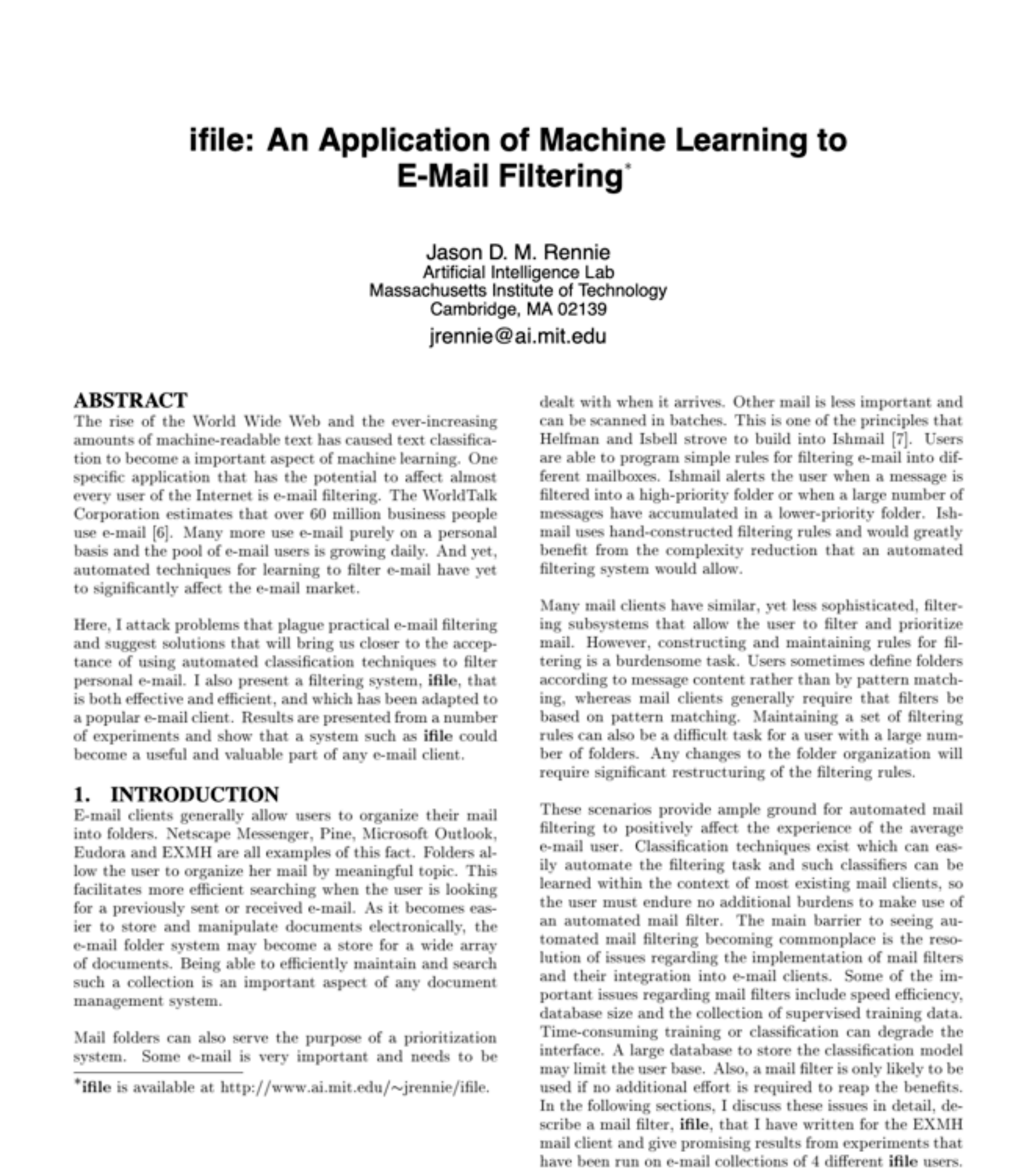
This is usually an open point.

Exercise: you must redo a example in page 26 of Vitor's book.

- Mapped to our use case: does the measurement support my physics model?

Applications outside of the HEP





ifile: An Application of Machine Learning to E-Mail Filtering*

Jason D. M. Rennie
Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139
jrennie@ai.mit.edu

ABSTRACT

The rise of the World Wide Web and the ever-increasing amounts of machine-readable text has caused text classification to become a important aspect of machine learning. One specific application that has the potential to affect almost every user of the Internet is e-mail filtering. The WorldTalk Corporation estimates that over 60 million business people use e-mail [6]. Many more use e-mail purely on a personal basis and the pool of e-mail users is growing daily. And yet, automated techniques for learning to filter e-mail have yet to significantly affect the e-mail market.

Here, I attack problems that plague practical e-mail filtering and suggest solutions that will bring us closer to the acceptance of using automated classification techniques to filter personal e-mail. I also present a filtering system, **ifile**, that is both effective and efficient, and which has been adapted to a popular e-mail client. Results are presented from a number of experiments and show that a system such as **ifile** could become a useful and valuable part of any e-mail client.

1. INTRODUCTION

E-mail clients generally allow users to organize their mail into folders. Netscape Messenger, Pine, Microsoft Outlook, Eudora and EXMH are all examples of this fact. Folders allow the user to organize her mail by meaningful topic. This facilitates more efficient searching when the user is looking for a previously sent or received e-mail. As it becomes easier to store and manipulate documents electronically, the e-mail folder system may become a store for a wide array of documents. Being able to efficiently maintain and search such a collection is an important aspect of any document management system.

Mail folders can also serve the purpose of a prioritization system. Some e-mail is very important and needs to be dealt with when it arrives. Other mail is less important and can be scanned in batches. This is one of the principles that Helfman and Isbell strove to build into Ishmail [7]. Users are able to program simple rules for filtering e-mail into different mailboxes. Ishmail alerts the user when a message is filtered into a high-priority folder or when a large number of messages have accumulated in a lower-priority folder. Ishmail uses hand-constructed filtering rules and would greatly benefit from the complexity reduction that an automated filtering system would allow.

Many mail clients have similar, yet less sophisticated, filtering subsystems that allow the user to filter and prioritize mail. However, constructing and maintaining rules for filtering is a burdensome task. Users sometimes define folders according to message content rather than by pattern matching, whereas mail clients generally require that filters be based on pattern matching. Maintaining a set of filtering rules can also be a difficult task for a user with a large number of folders. Any changes to the folder organization will require significant restructuring of the filtering rules.

These scenarios provide ample ground for automated mail filtering to positively affect the experience of the average e-mail user. Classification techniques exist which can easily automate the filtering task and such classifiers can be learned within the context of most existing mail clients, so the user must endure no additional burdens to make use of an automated mail filter. The main barrier to seeing automated mail filtering becoming commonplace is the resolution of issues regarding the implementation of mail filters and their integration into e-mail clients. Some of the important issues regarding mail filters include speed efficiency, database size and the collection of supervised training data. Time-consuming training or classification can degrade the interface. A large database to store the classification model may limit the user base. Also, a mail filter is only likely to be used if no additional effort is required to reap the benefits. In the following sections, I discuss these issues in detail, describe a mail filter, **ifile**, that I have written for the EXMH mail client and give promising results from experiments that have been run on e-mail collections of 4 different **ifile** users.

*ifile is available at http://www.ai.mit.edu/~jrennie/ifile.

Exercise-1:Pior@work

Assume there is a decease which 0.1% of the population have ($\rightarrow$ this is your prior). Assume there is a test that diagnoses this decease with a probability of 98%, while it gives false positive results with a probability of 3% .

a) Your test is positive. Calculate the probability that you are ill.

Exercise-1 :Pior@work

Assume there is a decease which 0.1% of the population have (→ this is your prior). .
Assume there is a test that diagnoses this decease with a probability of 98%, while it gives false positive results with a probability of 3% .

a) Your test is positive. Calculate the probability that you are ill.

$$\mathcal{P}_+(\text{ill}) = \frac{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})}{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})+\mathcal{P}_{\text{not ill}}(+)\cdot\mathcal{P}(\text{not ill})}$$

$$= \frac{0.98\cdot 0.001}{0.98\cdot 0.001+0.03\cdot 0.999} = 0.032$$

b) How does the result change if you redo the test and it is positive again?

Exercise-1:Pior@work

Assume there is a decease which 0.1% of the population have (→ this is your prior). Assume there is a test that diagnoses this decease with a probability of 98%, while it gives false positive results with a probability of 3% .

a) Your test is positive. Calculate the probability that you are ill.

$$\mathcal{P}_{+}(\text{ill}) = \frac{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})}{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})+\mathcal{P}_{\text{not ill}}(+)\cdot\mathcal{P}(\text{not ill})}$$

$$= \frac{0.98\cdot 0.001}{0.98\cdot 0.001+0.03\cdot 0.999} = 0.032$$

b) How does the result change if you redo the test and it is positive again?

$$\mathcal{P}_{+}(\text{ill}) = \frac{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})}{\mathcal{P}_{\text{ill}}(+)\cdot\mathcal{P}(\text{ill})+\mathcal{P}_{\text{not ill}}(+)\cdot\mathcal{P}(\text{not ill})}$$

$$= \frac{0.98\cdot 0.032}{0.98\cdot 0.032+0.03\cdot 0.968} = 0.519$$

Probability Density Function

- In general we assume the underlying statistics model to follow a given probability density function.

- Most prominent examples:

Poisson:

$$\mathcal{P}(k, \mu) = \frac{\mu^k}{k!} e^{-\mu}$$

e.g. for counting experiments (like for cross sections).

Gaussian:

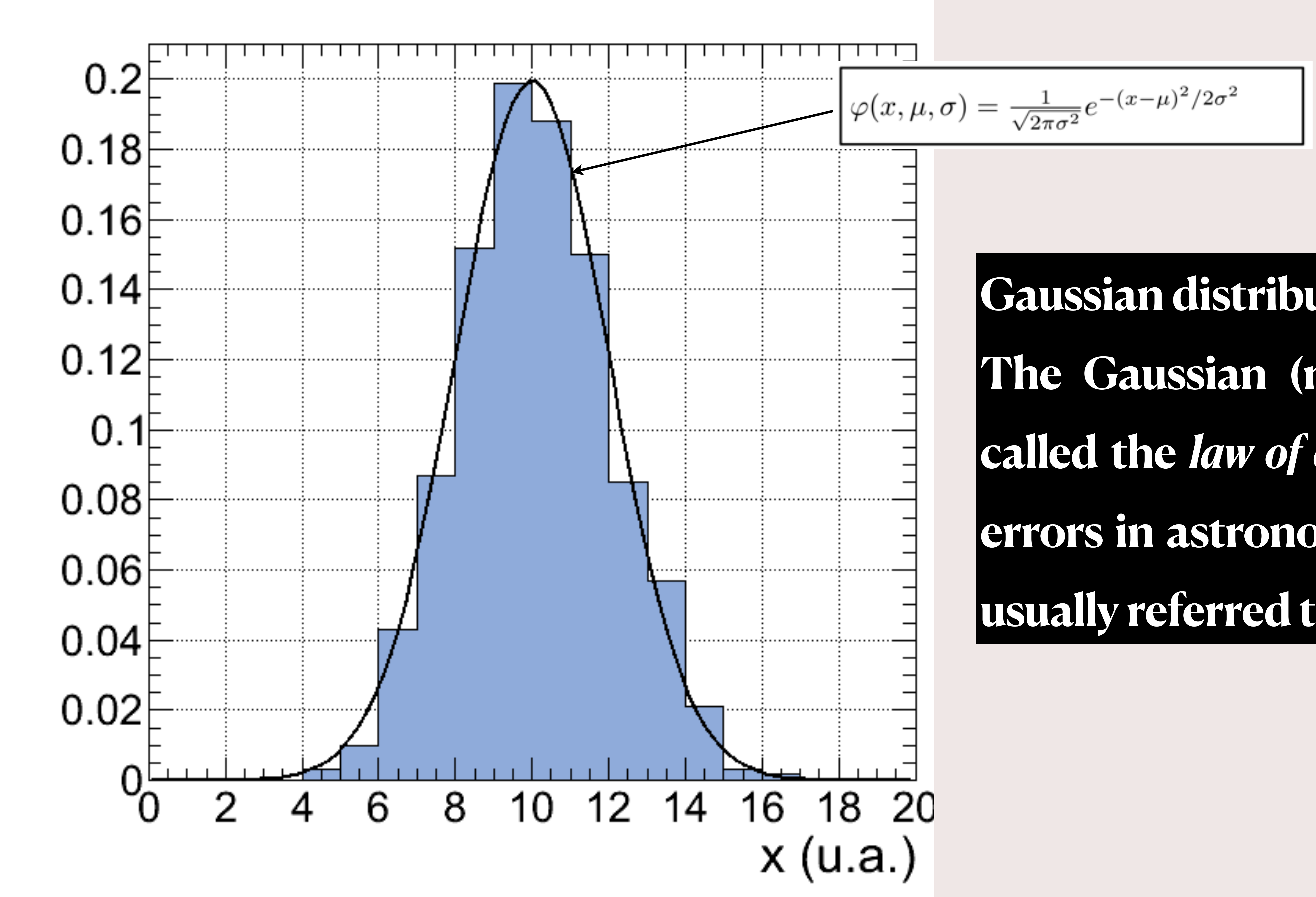
$$\varphi(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

e.g. for parameter estimates (like for mass measurements) .

- Probability density distributions are themselves functions of parameters.

- If it is the target of the measurement we often call it parameter of interest (POI), otherwise we often call it nuisance parameter (often indicated by θ).

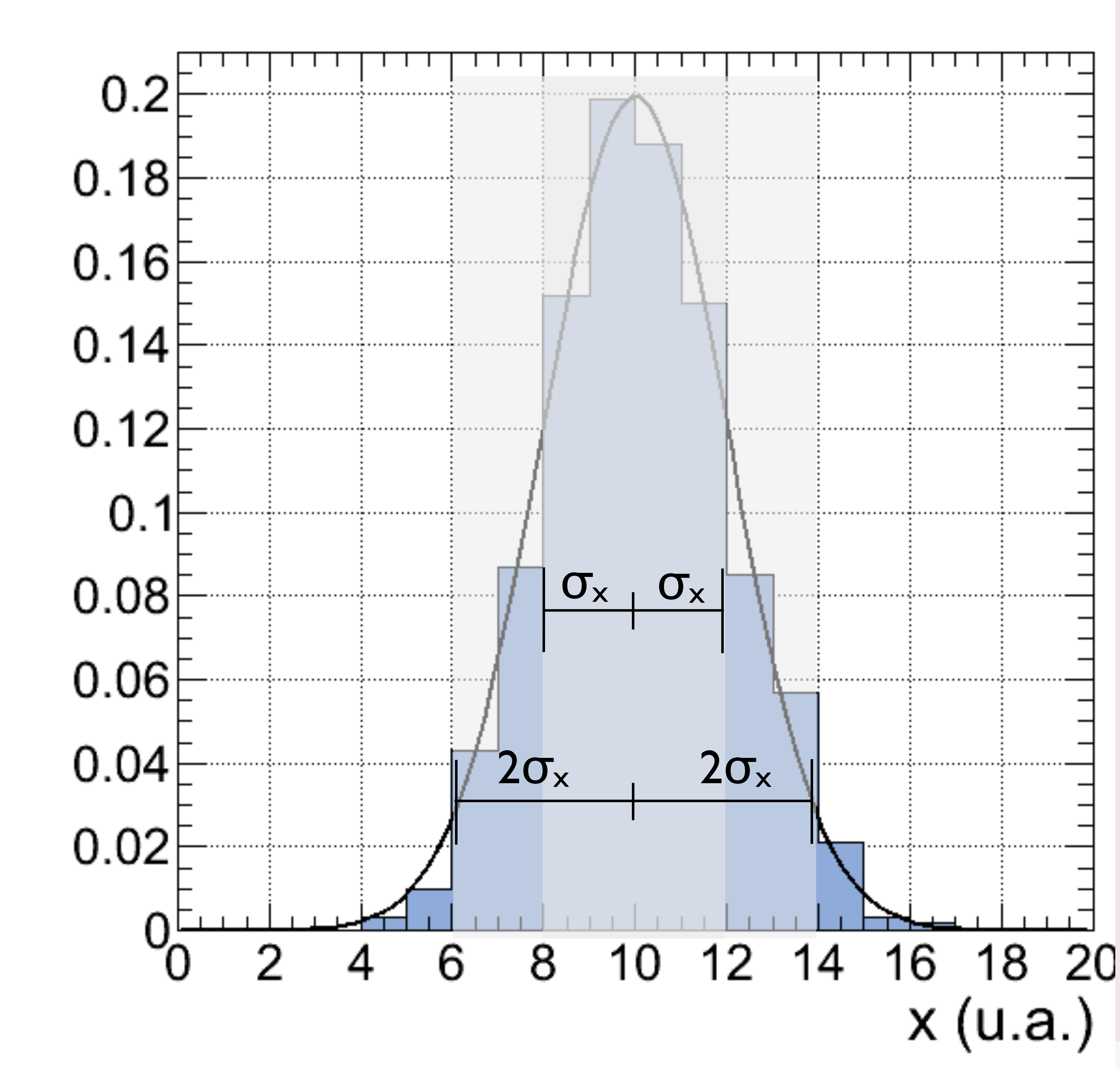
Statistics Uncertainties: Law of Error

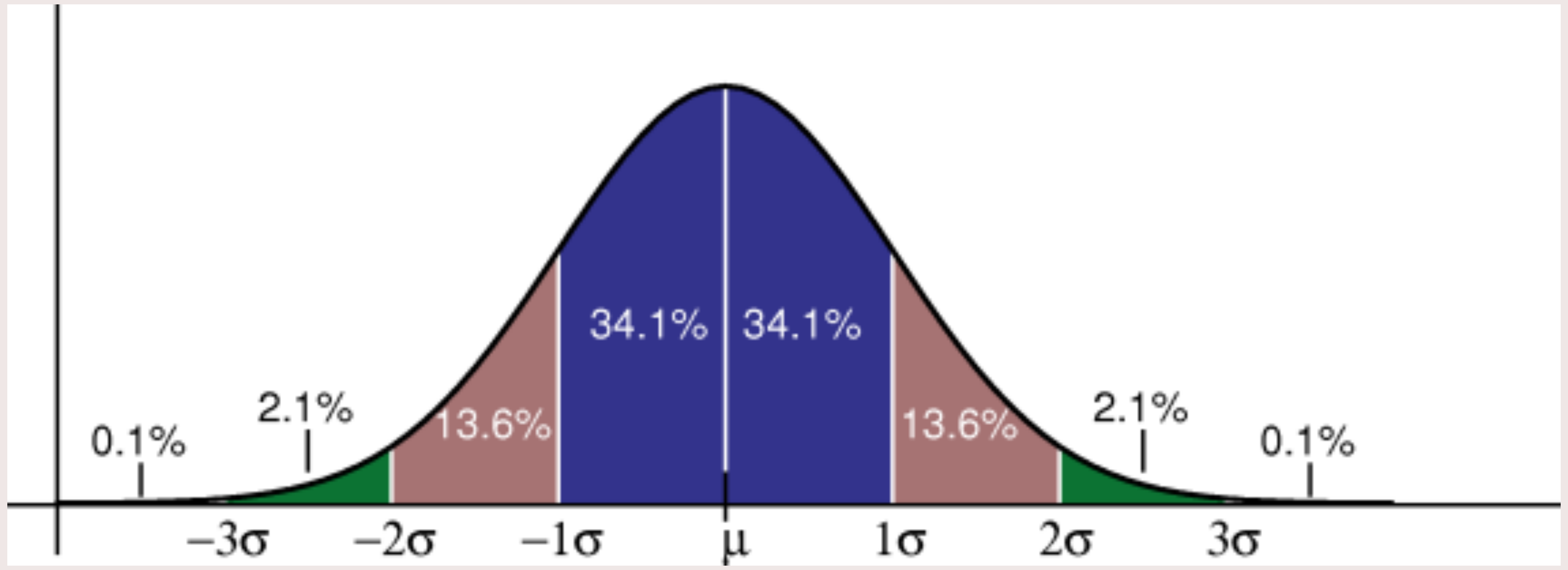


Gaussian distribution.

The Gaussian (normal) distribution was historically called the *law of errors*. It was used by Gauss to model errors in astronomical observations, which is why it is usually referred to as the Gaussian distribution.

Statistics Uncertainties: Confidence Level

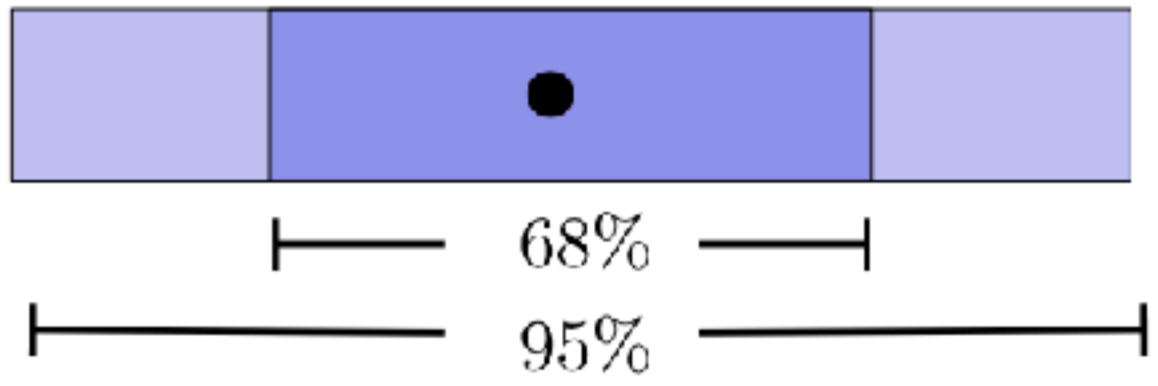




Confidence Intervals

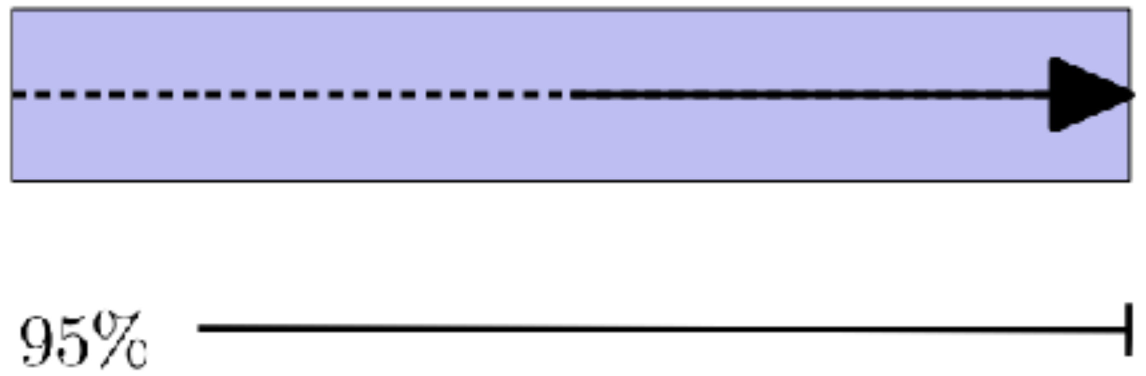
- Confidence intervals allow statements about parameters in models.

Double sided (measurement):



$$m = \mu \pm \sigma_{0.68} \ @ \ 68\% \ CL$$

Single sided (limit):



$$BR \leq \mu_{0.95} \ @ \ 95\% \ CL$$

- Interpretation depending on paradigm:

Frequentist:

Probability to make given observation for a given truth. Esp. no probability for “truth to be true”

Bayesian:

Probability of truth to lie in given interval.

- We will concentrate on single sided confidence intervals (→ used for upper limits).

Upper Limits

- With the upper limit on a model POI μ, for a given observation x_{obs} (or N_{obs}) we search for the largest value of μ for which the probability to make an observation of $x \leq x_{obs}$ (or $N \leq N_{obs}$) is less than a value α .

- We call this value of μ the upper limit on μ at the confidence level (CL) $1 - \alpha$. During the next slides we will indicate this quantity by $\mu_{1-\alpha}$.

- We particle physics we usually use $\alpha = 0.05$ ($\rightarrow$ 95% CL limit).

Meaning:

For $\mu = \mu_{0.95}$ in 95% of all outcomes of the same experiment x (or N) would have been larger then x_{obs} (or N_{obs}) . For $\mu > \mu_{0.95}$ this fraction would be even larger. The observation restricts μ to be not larger than $\mu_{0.95}$ at 95% CL.

Question:

Is $\mu_{0.90} < \mu_{0.95}$ or $\mu_{0.90} > \mu_{0.95}$?

Confidence Intervals Construction -Frequentist

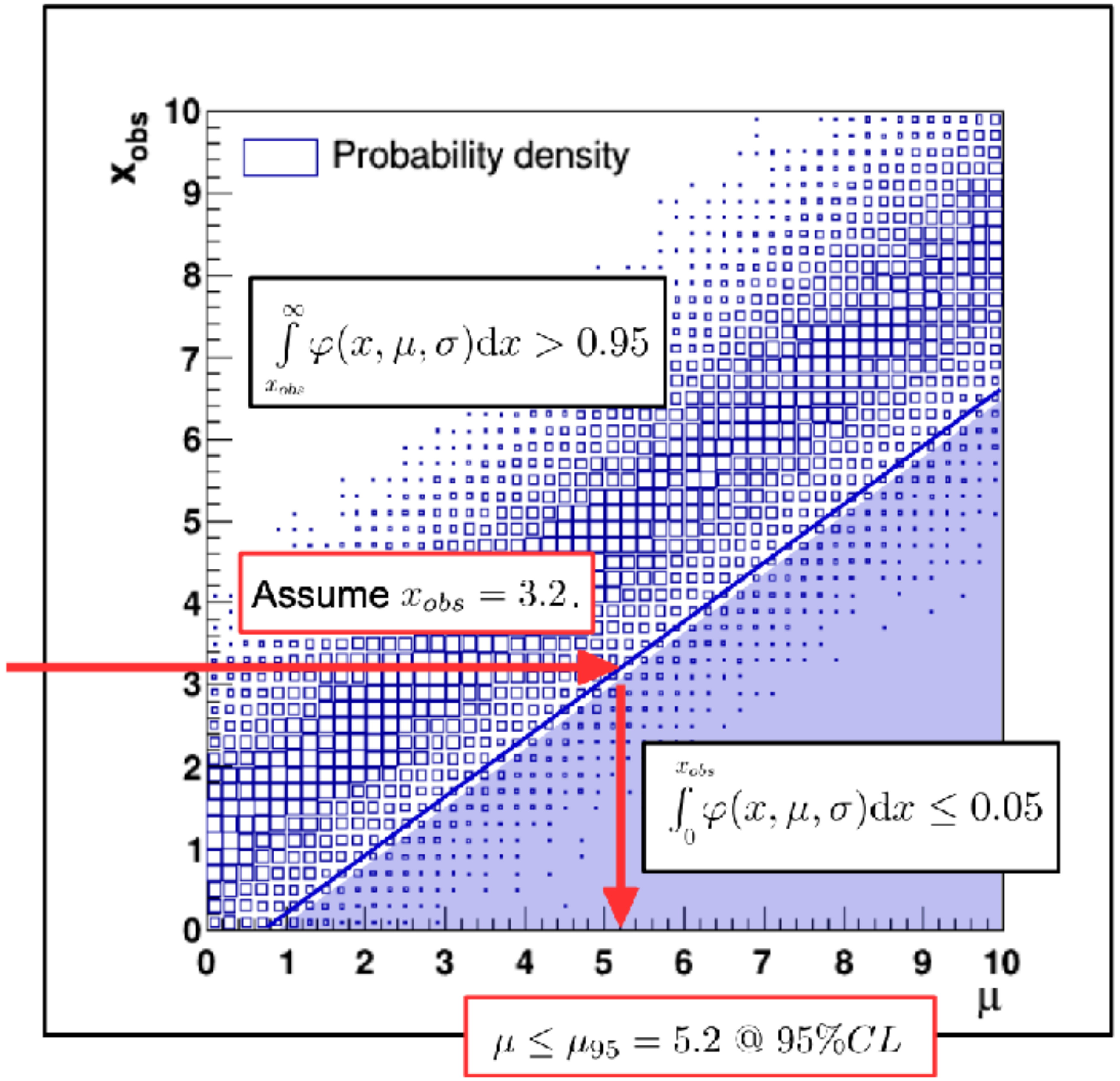
- Here shown for 95% CL upper limit on a parameter μ of a Gaussian distributed random variable x_{obs} :

- **Neyman construction**:

- For each value of μ find single sided confidence interval for given α (e.g. $\alpha = 0.05$).

- Interconnect interval edges.

- For a given observation find the largest value for μ where x_{obs} is still contained in the interval.



Confidence Intervals Construction -Frequentist

- Here shown for 95% CL upper limit on a parameter μ of a Poisson distributed random variable N_{obs}:

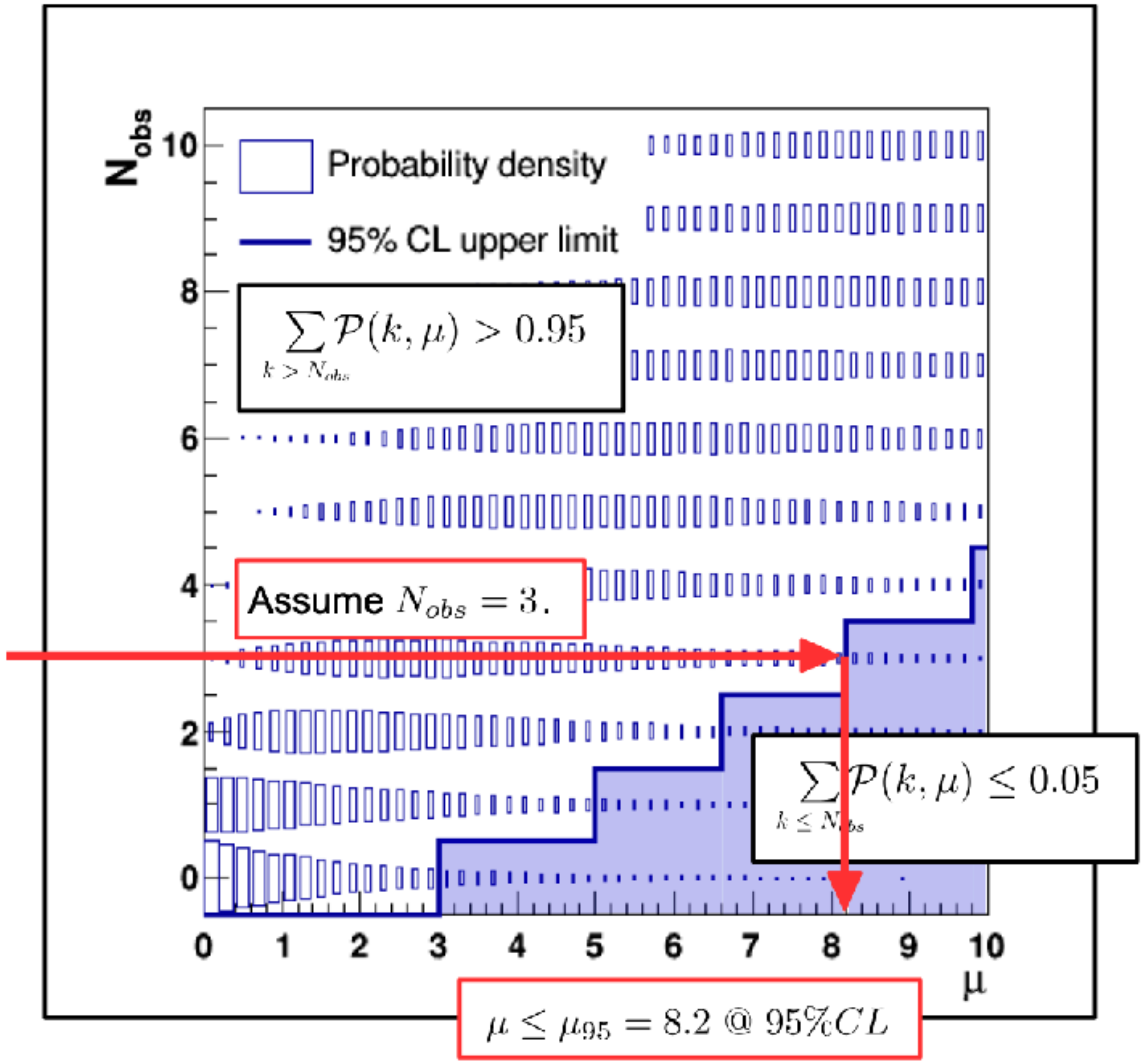
- **Neyman construction**:

- For each value of μ find single sided confidence interval for given α (e.g. $\alpha = 0.05$).

- Interconnect interval edges.

- For a given observation find the largest value for μ where x_{obs} is still contained in the interval.

- Note steps due to discrete nature of Poisson distribution.



Coverage

- For a given limit procedure you can calculate for each value of μ the exact probability to exclude the theory:

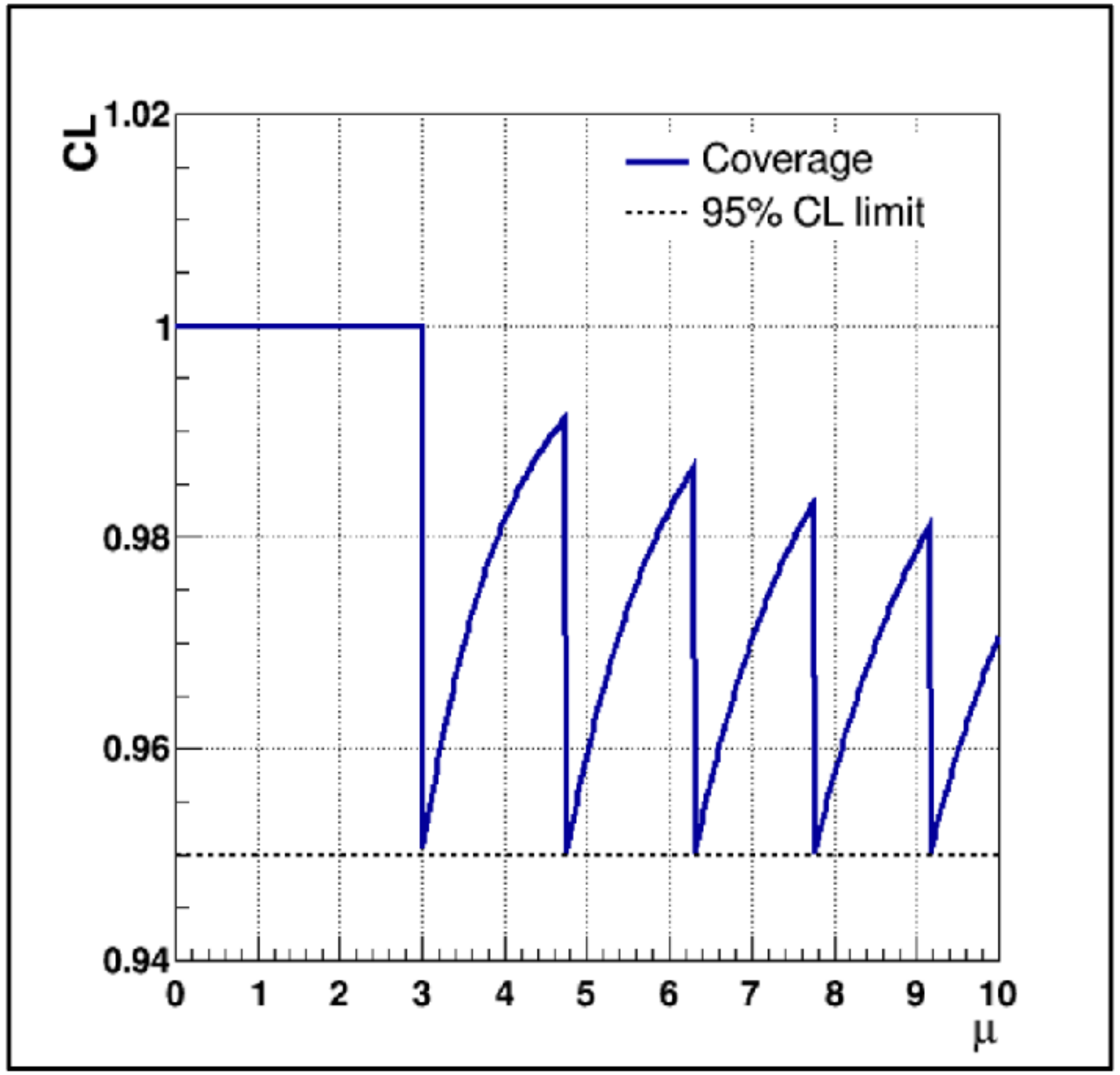
- **Coverage**:

- For our Poisson example:

$$P_{excl}(\mu) = \sum_{N \geq N_{obs}} P_\mu(N) \cdot \theta(\mu \leq \mu_{0.95})$$

$$P_{excl}(\mu) \begin{cases} > 0.95 & \text{over coverage} \\ = 0.95 & \text{exact coverage} \\ < 0.95 & \text{under coverage} \end{cases}$$

- Over coverage ($\rightarrow$ exclusion statement more conservative).



Exercise-2: Frequentist Limit

Assume you have an observation of 1 event, were you expect 0 due to already known processes. You want to quote a 95% CL upper limit on the true value μ of the expected events for a Poisson distributed signal model.

a) Calculate the exclusion probability (i.e. the probability to observe more than 2 events) for the values of μ given in the table on the right:

μ	P_{excl}
2	
3	
4	
5	

$N_{obs} = 1$

NB: in root you can use the function given below for your calculation. Vary μ for fixed $N = N_{obs}$ until $\alpha < 0.05$.

ROOT
Data Analysis Framework

Exercise-2: Frequentist Limit

Assume you have an observation of 1 event, were you expect 0 due to already known processes. You want to quote a 95% CL upper limit on the true value μ of the expected events for a Poisson distributed signal model.

a) Calculate the exclusion probability (i.e. the probability to observe more than 2 events) for the values of μ given in the table on the right:

μ	P_{excl}
2	0.59
3	0.80
4	0.91
5	0.96

$N_{obs} = 1$

b) Calculate the 95% CL limit on μ for the values of N_{obs} given in the table on the right:

N_{obs}	$\mu_{0.95}$
0	
1	
2	
4	

NB: in root you can use the function given below for your calculation. Vary μ for fixed $N = N_{obs}$ until $\alpha < 0.05$.

Exercise-2: Frequentist Limit

Assume you have an observation of 1 event, were you expect 0 due to already known processes. You want to quote a 95% CL upper limit on the true value μ of the expected events for a Poisson distributed signal model.

a) Calculate the exclusion probability (i.e. the probability to observe more than 2 events) for the values of μ given in the table on the right:

μ	P_{excl}
2	0.59
3	0.80
4	0.91
5	0.96

$N_{obs} = 1$

b) Calculate the 95% CL limit on μ for the values of N_{obs} given in the table on the right:

N_{obs}	$\mu_{0.95}$
0	3.00
1	4.74
2	6.30
4	9.15

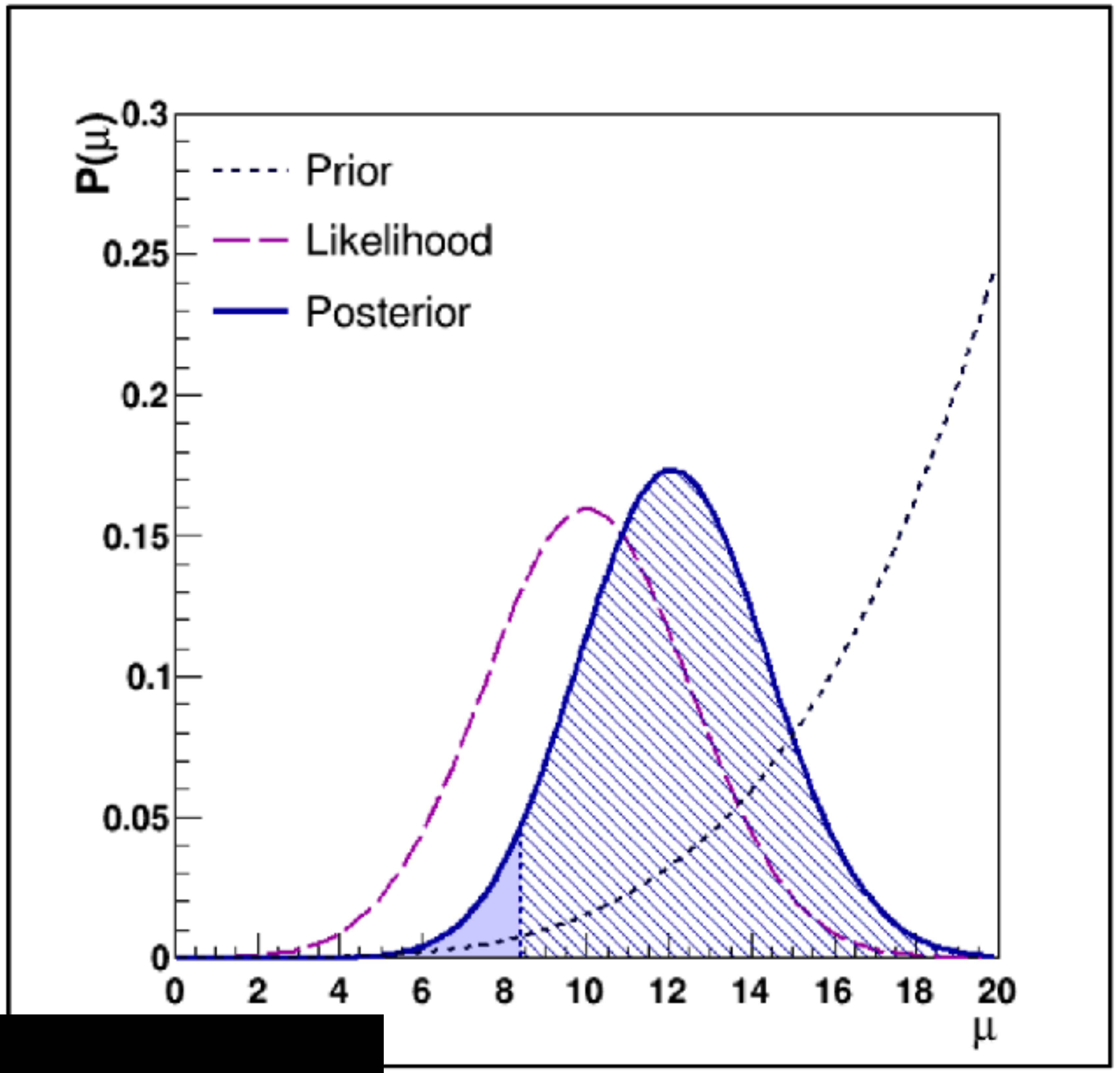
NB: in root you can use the function given below for your calculation. Vary μ for fixed $N = N_{obs}$ until $\alpha < 0.05$.

ROOT
Data Analysis Framework

Confidence Intervals Construction -Bayesian

- Here shown for 95% CL upper limit on a parameter μ of an arbitrarily distributed random variable x_{obs} :
- **Bayesian limit:**
- Assign prior $\mathcal{P}(\mathrm{Model})$ to model to be true and likelihood $\mathcal{P}_{\mathrm{Model}}(x_{obs})$.
- Calculate posterior $\mathcal{P}_{x_{obs}}(\mathrm{Model})$ for known prior, likelihood & x_{obs} .
- Determine $\mu_{0.95}$ such that:

$$\underbrace{\frac{\int_0^{\mu_{0.95}} \mathcal{P}_{x_{obs}}(\mathrm{Model}) \cdot \mathcal{P}(\mathrm{Model}) \mathrm{d}\mu}{\int_0^{\infty} \mathcal{P}_{x_{obs}}(\mathrm{Model}) \cdot \mathcal{P}(\mathrm{Model}) \mathrm{d}\mu}}_{\equiv \alpha(\mu, N)} \leq 0.05$$



Requires numerical integration of posterior

Exercise-3: Bayesian Limit

Assume you have an observation of 1 event, were you expect 0 due to already known processes. You want to quote a 95% CL upper limit on the true value μ of the expected events for a Poisson distributed signal model.

a) Calculate the 95% CL limit on μ for $N_{obs} = 2$ and a flat prior.

b) Do the same calculation for the prior $\mathcal{P}(\text{Model}) \propto \mu$.

N_{obs}	μ_{95}
1	3.00
2	4.74
3	6.30
5	9.15

$\mathcal{P}(\text{Model}) = \text{const}$

N_{obs}	μ_{95}
1	4.75
2	6.27
3	7.76
5	10.51

$\mathcal{P}(\text{Model}) \propto \mu$

NB: the macro below calculates α for you by numerical integration of the posterior. Vary μ for fixed $N = N_{obs}$ until $\alpha < 0.05$.

ROOT
Data Analysis Framework

Summary

- Small Review about probability concepts
- Confidence intervals and limits
- Frequentist and Bayesian construction limit

backup slide