Statistical Methods in Particle Physics

Workshop at the 2020 Retreat of RTG 2149
("Strong and Weak Interactions - from Hadrons to Dark Matter")
9-10 September 2020

Klaus Reygers
Heidelberg University

Contents

Material

https://www.physi.uni-heidelberg.de/~reygers/lectures/2020/smipp-grk-2149/

PI > Startseite

GRK 2149 workshop: Statistical Methods in Particle Physics

Lectures for the Retreat of the Research Training Group (Graduiertenkolleg) 2149 "Strong and Weak Interactions" - 9-10 September 2020

Overview

- Basics
- Maximum likelihood method
- Least Squares
- Hypothesis tests and Goodness-of-fit
- Systematic uncertainties
- Decision trees

We'll use python 3 in the course. Basic knowdegle of the language is useful for this course. We'll work a lot with jupyter notebooks. A nice summary of important python commands is available on the website of the Stanford lecture CS231n.

Here you find the slides of the lecture.

Examples

- basic_chi2_fit_iminuit.ipynb
- error_ellipse.ipynb
- extended_ml_fit_example.ipynb
- ml_fit_example.ipynb
- p-values_and_n-sigma.ipynb
- random_numbers_from_distribution.ipynb

Examples and problems in Python

Problems

1. Construct a Bayesian credible interval (html, notebook) ⧖
2. Linear least squares and error propgation (html, notebook) ⧖⧖
3. Simultaneous least-squares fit to several data sets (blast-wave fit to particle spectra) (html, notebook) ⧖⧖⧖
4. Unbinned maximum likelihood fit (double exponential decay) (html, notebook) ⧖⧖
5. The lighthouse problem: another unbinned maximum likelihood fit (html, notebook) ⧖⧖
6. Unbinned maximum likelihood fit with Gaussian constraint on a parameter (html, notebook) ⧖
7. Kolmogorov-Smirnov test (html, notebook) ⧖
8. Significance of a peak above background (html, notebook) ⧖⧖⧖
9. Least-squares fit with external Gaussian constraint (html, notebook) ⧖
10. Separation of gamma and hadron showers measured with the MAGIC Cherenkov telescope using a boosted decision tree and a random forest (html, notebook) ⧖⧖

⧖ = quick, ⧖⧖ = intermediate, ⧖⧖⧖ = takes a bit longer

Side note:

Python provides a great ecosystem of tools and is "user-friendly", but also slow.

Is Julia the future?
"runs like C, but reads like Python"
(Nature 572, 141-142 (2019))

Hands-on Exercises

Will spend a good fraction of the time on these hands-on exercises.

Problems

1. Construct a Bayesian credible interval (html, notebook) ⧖
2. Unbinned maximum likelihood fit (double exponential decay) (html, notebook) ⧖⧖
3. The lighthouse problem: another unbinned maximum likelihood fit (html, notebook) ⧖⧖
4. Unbinned maximum likelihood fit with Gaussian constraint on a parameter (html, notebook) ⧖
5. Linear least squares and error propgation (html, notebook) ⧖⧖
6. Simultaneous least-squares fit to several data sets (blast-wave fit to particle spectra) (html, notebook) ⧖⧖⧖
7. Kolmogorov-Smirnov test (html, notebook) ⧖
8. Significance of a peak above background (html, notebook) ⧖⧖⧖
9. Least-squares fit with external Gaussian constraint (html, notebook) ⧖
10. Separation of gamma and hadron showers measured with the MAGIC Cherenkov telescope using a boosted decision tree and a random forest (html, notebook) ⧖⧖

⧖ = quick, ⧖⧖ = intermediate, ⧖⧖⧖ = takes a bit longer

Hands-on exercises: Example

Typically you just need to add one or two lines where "your code here" appears

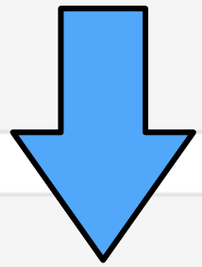
```python
# chi2 for a given particle species
def chi2_particle(particle, Tkin, beta_s, n):

    # blast-wave prediction
    dndpt_bw = dndpt_blastwave(pt_meas[particle], mass[particle], Tkin, beta_s, n)

    # normalization
    A = normalization(particle, Tkin, beta_s, n)

    # ...
    # your code here

    # return np.sum(pulls * pulls)
```



```python
# chi2 for a given particle species
def chi2_particle(particle, Tkin, beta_s, n):

    # blast-wave prediction
    dndpt_bw = dndpt_blastwave(pt_meas[particle], mass[particle], Tkin, beta_s, n)

    # normalization
    A = normalization(particle, Tkin, beta_s, n)

    pulls = (dndpt_meas[particle] - A * dndpt_bw) / dndpt_meas_err[particle]
    return np.sum(pulls * pulls)
```

Working environment (here: macOS)

Assumption: homebrew is installed

install python3 (see https://docs.python-guide.org/starting/install3/osx/)

```
$ brew install python
$ python --version
Python 3.8.5
```

Make sure pip3 is up-to-date (alternative: conda)

```
$ pip3 install --upgrade pip
```

Install needed modules:

```
$ pip3 install --upgrade jupyter matplotlib numpy
pandas scipy scikit-learn xgboost iminuit
```

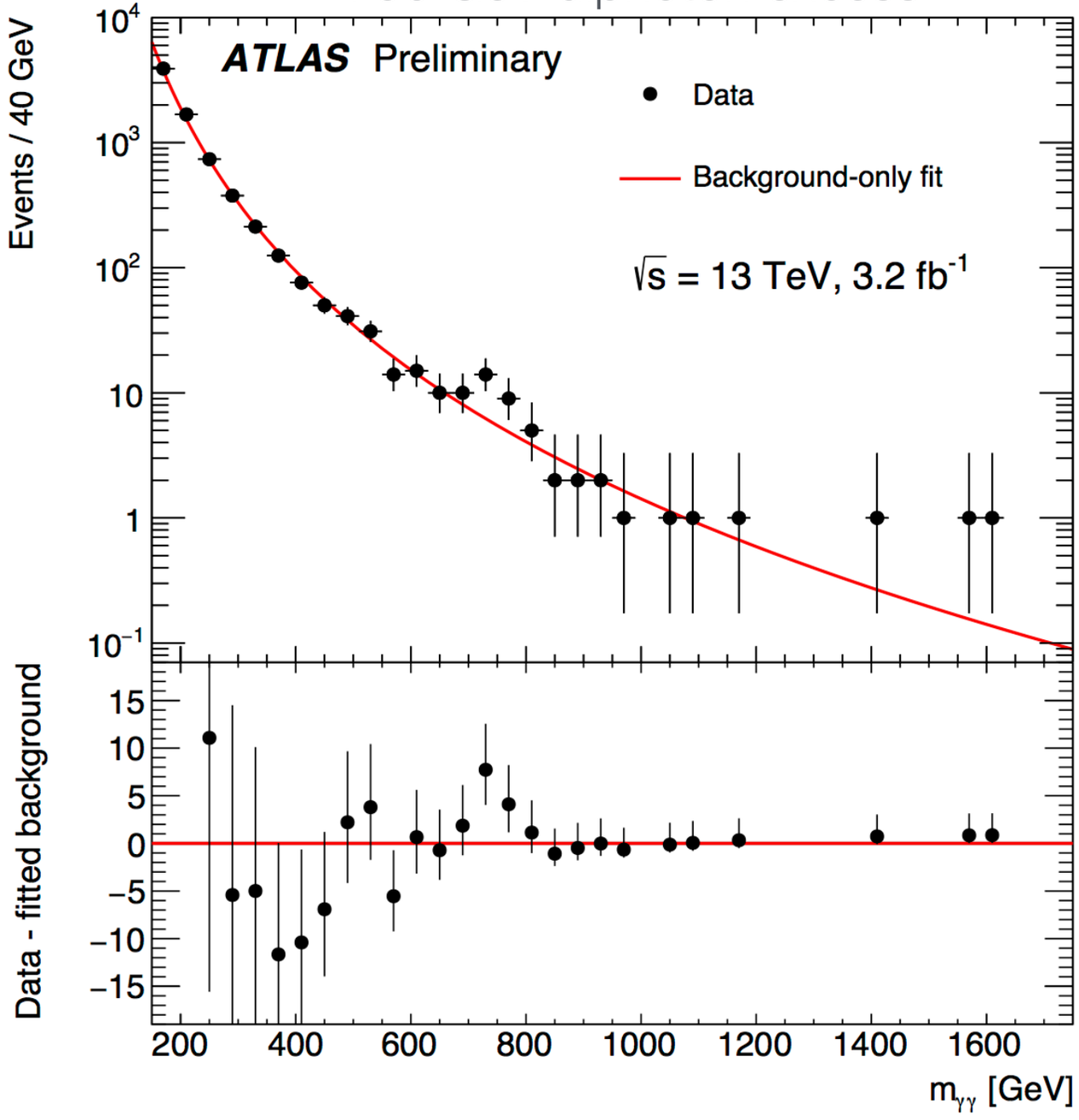
run jupyter:

```
$ jupyter lab      or      $ jupyter notebook
```

Introduction and basic concepts

Why bother with statistical methods?

"750 GeV diphoton excess"



Statistics:
Draw reliable conclusions from data

In case of doubt:
just get more data …

Yes, but not always easy …

A heavy Higgs boson?

Peak disappeared with more data … [link]

Presentations by CMS and ATLAS, December 2015:
https://indico.cern.ch/event/442432/

How Knowledge is Created?

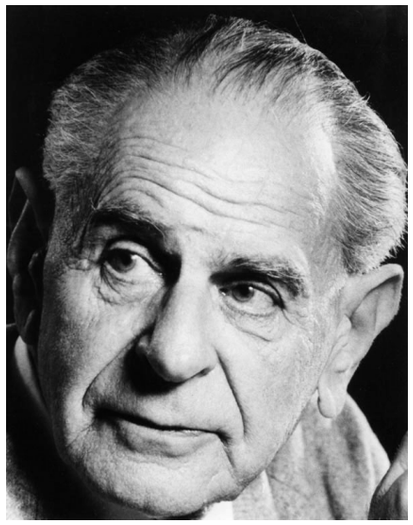
Guess theory/model

- usually mathematical
- self-consistent
- simple explanations, few arbitrary parameters
- testable predictions

Perform experiment

- reject / modify theory in case of disagreement with data
- if theory requires too many adjustments it becomes unattractive

The advance of scientific knowledge is an *evolutionary* process



source: Wikipedia

Karl Popper
(1902–1994)

Statistical methods are an important part of this process

A look at other research fields

"**Why Most Published Research Findings Are False**":
Main thesis: large number, if not the majority, of published medical research papers contain results that cannot be replicated.

Reproducibility crisis:
Affects the social sciences and medicine most severely (in particular psychology)

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviations: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

John Ioannidis
(Stanford School of Medicine)
PLoS Med 2(8): e124., (2005),
doi:10.1371/journal.pmed.0020124

Useful books

- G. Cowan, Statistical Data Analysis
- L. Lista, Statistical Methods for Data Analysis in Particle Physics
- Behnke, Kroeninger, Schott, Schoerner-Sadenius: Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods
- R. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences
- Bohm, Zech, Introduction to Statistics and Data Analysis for Physicist [available online]
- Blobel, Lohrmann: Statistische Methoden der Datenanalyse (in German), [free ebook]
- L. Lyons:
  Statistics for Nuclear and Particle Physicists (Cambridge University Press)
- F. James, Statistical Methods in Experimental physics
- W. Metzger, Statistical Methods in Data Analysis [available online]

Further Material

- Glen Cowan: http://www.pp.rhul.ac.uk/~cowan/stat_course.html
- Scott Oser: http://www.phas.ubc.ca/~oser/p509/
- Terascale Statistics School: https://indico.desy.de/indico/event/25594/other-view?view=standard
- Particle Data Group reviews on Probability and Statistics
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-probability.pdf
  - https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf
- K.R., Statistical Methods in Particle Physics, WS 2017/18: https://uebungen.physik.uni-heidelberg.de/vorlesung/20172/smipp

Interpretations of Probability

- Classical
  - Assign equal probabilities based on symmetry of the problem, e.g., rolling dice: $P(6) = 1/6$
  - difficult to generalize

- Frequentist
  - Let A, B, … be outcomes of an repeatable experiment:

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

- Bayesian (subjective probability)
  - A, B, … are hypotheses (statements that are true or false)

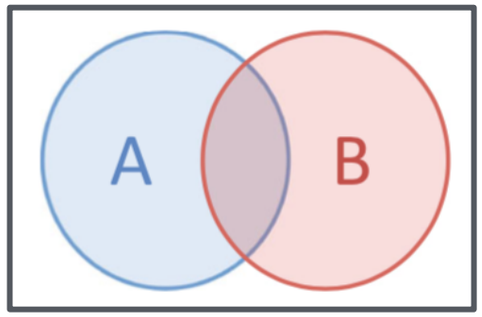
$$P(A) = \text{degree of believe that } A \text{ is true}$$

Criticisms of the Probability Interpretations

- Criticisms of the frequency interpretation
  - $n \to \infty$ can never be achieved in practice. When is n large enough?
  - P is not an intrinsic property of A, it depends on the how the ensemble of possible outcomes was constructed
    - Example: P(person I talk to is a physicist) depends on whether I am in a football stadium or at a physics workshop
  - We want to talk about the probability of events that are not repeatable
    - Example 1: P(it will rain tomorrow), but there is only one tomorrow
    - Example 2: P(Universe started with a Big Bang), but only one universe

- Criticisms of the subjective Bayesian interpretation
  - "Subjective" estimates have no place in science
  - How to quantify the prior state of our knowledge upon which we base our probability estimate?

Bayes' Theorem



Venn diagram

Definition of conditional probability:

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

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

[doubtful whether the portrait actually shows Bayes]

First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

First modern formulation by Pierre-Simon Laplace in 1812

Accepted by everyone if probabilities are not Bayesian probabilities

Bayesian inference: Degree of Believe in a Theory Given a Certain Set of Data (I)

probability of getting the data if theory is true

prior (subjective belief in the theory before seeing the data)

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

posterior probability, i.e., subjective belief in the theory after seeing the data

guarantees normalization:

$$P(\text{data}) = \sum_i P(\text{data}|\text{theory}_i)P(\text{theory}_i)$$

Addresses question: "What should I believe?"

Bayesian inference: Degree of Believe in a Theory Given a Certain Set of Data (II)

For a continuous parameter λ:

$$P_{\text{posterior}}(\lambda|m) = \frac{f(m|\lambda)P_{\text{prior}}(\lambda)}{f_1(m)}$$

λ : true value of a parameter of nature

m : measurement

$$f_1(m) = \int f(m|\lambda')P(\lambda')\,\mathrm{d}\lambda'$$

Problems with Bayesian inference

What functional form to chose for $P_{\text{prior}}(\lambda)$?

Uninformed prior: flat? In which variable, e.g., in $\lambda, \lambda^2, 1/\lambda, \ln\lambda$?

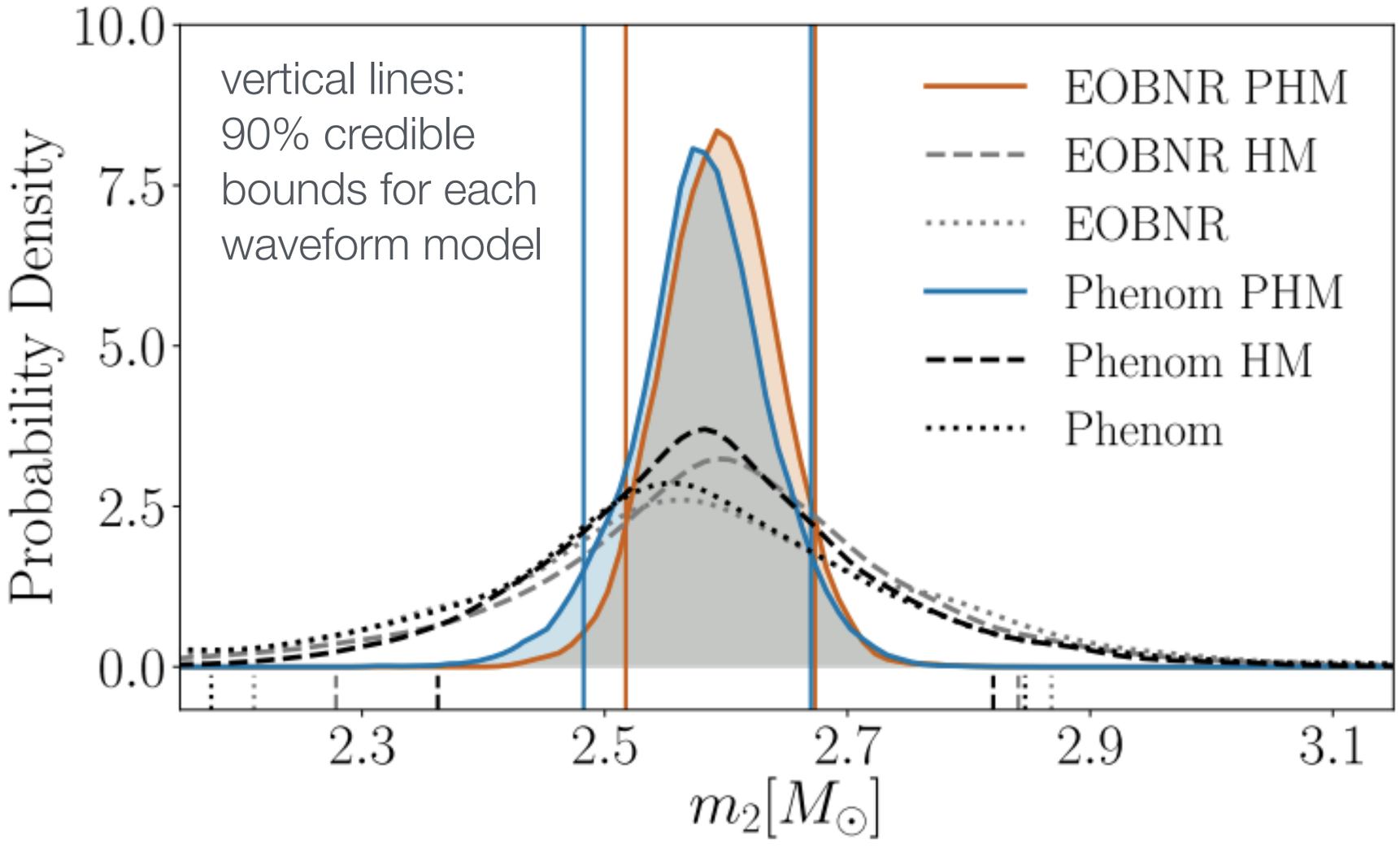
Bayesian reply

Choice of prior usually unimportant after a few experiments

Jaynes' robot: Priors are uniquely determined by your state of knowledge. Thus scientists with the same background knowledge construct the same priors.

Example of a Posterior Distribution

GW190814: Gravitational Waves from the Coalescence of a 23 Solar Mass Black Hole with a 2.6 Solar Mass Compact Object



Posterior Distribution for mass of the lighter objects:

LIGO Scientific Collaboration and Virgo Collaboration:
The Astrophysical Journal Letters, 896:L44 (20pp), 2020 June 20

Note:
Sampling from a multi-parameter posterior distribution typically involves Markov chain Monte Carlo (MCMC)

Are you a Frequentist or a Bayesian?

Suppose mass of a particle is measured with Gaussian resolution σ and the result ist reported as

$$m \pm \sigma$$

Bayesian

$$P(m|m_{\text{true}}) \propto e^{-(m-m_{\text{true}})^2/(2\sigma^2)} \quad \text{flat prior for } m_{\text{true}} \longrightarrow \quad P(m_{\text{true}}|m) \propto e^{-(m-m_{\text{true}})^2/(2\sigma^2)}$$

Frequentist

This is a statement about the interval [m–σ, m+σ]. For a large number of hypothetically repeated experiments the interval would contain the true value in 68% of the cases. In the frequentist approach, a probabilistic statement about the true value is nonsense (the true value is what it is).

"Bayesians address the questions everyone is interested in by using assumptions that no one believes. Frequentist use impeccable logic to deal with an issue that is of no interest to anyone." – Louis Lyons

Bayesian Inference: Jeffreys' Prior

How to model complete ignorance about the value of a parameter θ?

- Uniform distribution in θ, exp θ, ln θ, 1/θ, …?
- Example: Lifetime τ of a particle, uniform distribution in τ or particle's width Γ = 1/τ ?

Jeffreys' prior (non-informative prior) for a model $L(\vec{x}|\vec{\theta})$ of the measurement:

$$\pi(\vec{\theta}) \propto \sqrt{I(\vec{\theta})} \qquad I(\vec{\theta}) = \det\left[\left\langle \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial\theta_i} \frac{\partial \ln L(\vec{x}|\vec{\theta})}{\partial\theta_j} \right\rangle\right]$$

invariant under re-parameterization

determinant of the Fisher information matrix

expectation value evaluated by $\vec{x}$
integrating over all possible results

Examples:

PDF parameter	Jeffreys' prior
Poissonian mean μ	**$p(\mu) \propto 1/\sqrt{\mu}$**
Gaussian mean μ	**$p(\mu) \propto 1$**

Jeffreys' Prior: Example

Exponential distribution: $$L(t \mid \tau) = \frac{1}{\tau} e^{-t/\tau}$$

Jeffreys' prior: $$\pi(\tau) \propto \sqrt{I(\tau)} = \sqrt{\mathrm{E}\left[\left(\frac{d}{d\tau} \ln L(t \mid \tau)\right)^2\right]}$$

$$\frac{d}{d\tau} \ln L(t|\tau) = -\frac{1}{\tau} + \frac{t}{\tau^2}$$

$$E\left[\left(\frac{t}{\tau^2} - \frac{1}{\tau}\right)^2\right] = E\left[\left(\frac{t-\tau}{\tau^2}\right)^2\right] = \frac{1}{\tau^4} V[t] = \frac{\tau^2}{\tau^4} = \frac{1}{\tau^2}$$

$$\rightsquigarrow \quad \pi(\tau) \propto \frac{1}{\tau}$$

Bayesian versus Frequentism

[based on L. Lyons]

	Bayesian	Frequentist
Meaning of probability	**degree of belief**	**frequentist definition**
Probability of parameters	**yes**	**anathema**
Needs prior	**yes**	**no**
Unphysical / empty intervals	**excluded by prior**	**can occur**
Final statement	**posterior probability distribution**	**parameter values, hypothesis test (*p*-value)**
Systematics	**Integrate over nuisance parameter**	**Various methods, e.g., profile likelihood, hard**
Combination of measurements	**can be hard (prior)**	**ok**

Variance and Standard Deviation

Expected value of a random variable x that follows a distribution P(x):

$$\mathrm{E}[x] \equiv \langle x \rangle \equiv \int x\, P(x)\, \mathrm{d}x$$

Variance:

$$\mathrm{V}[x] = E[(x-\mu)^2] = \int \mathrm{d}x\, P(x)(x-\mu)^2$$

$$\begin{aligned}\mathrm{V}[x] &= \mathrm{E}\left[(x-\mathrm{E}[x])^2\right] = \mathrm{E}\left[x^2 - 2x\,\mathrm{E}[x] + \mathrm{E}[x]^2\right] \\ &= \mathrm{E}\left[x^2\right] - 2\,\mathrm{E}[x]\,\mathrm{E}[x] + \mathrm{E}[x]^2 = \mathrm{E}\left[x^2\right] - \mathrm{E}[x]^2 = \langle x^2 \rangle - \langle x \rangle^2\end{aligned}$$

Standard deviation:

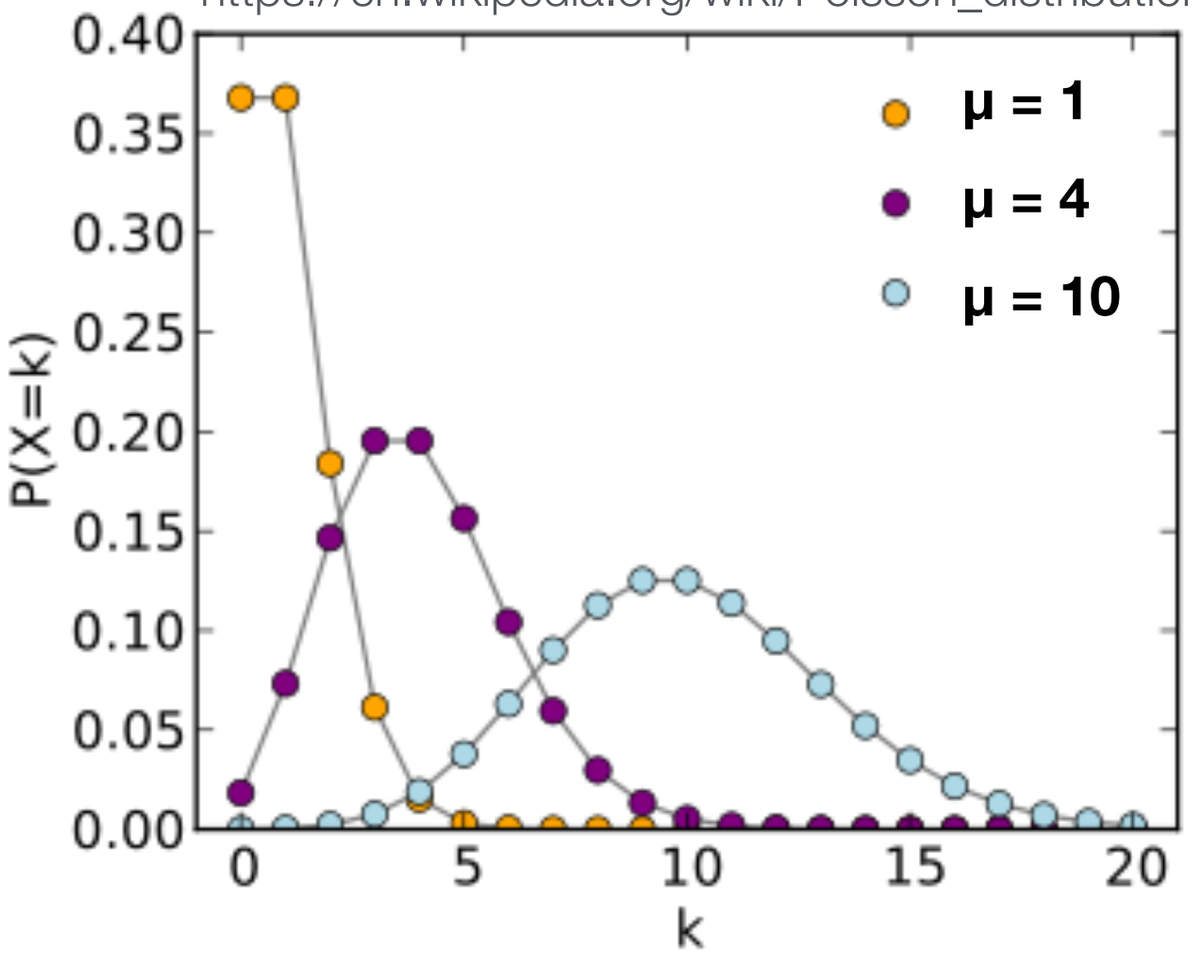
$$\sigma = \sqrt{\mathrm{V}[x]}$$

Poisson Distribution

$$p(k;\mu) = \frac{\mu^k}{k!} e^{-\mu}$$

$$E[k] = \mu, \quad V[k] = \mu$$

https://en.wikipedia.org/wiki/Poisson_distribution



Properties:

- n_1, n_2 follow Poisson distr.
  → n_1+n_2 follows Poisson distr., too
- Can be approximated by a Gaussian for large μ

Counting experiment:

Examples:

- Clicks of a Geiger counter in a given time interval
- Number of Prussian cavalrymen killed by horse-kicks

Number of deaths in 1 corps in 1 year	Actual number of such cases	Poisson prediction
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6

Binomial Distribution

N independent experiments

- Outcome of each is 'success' or 'failure'
- Probability for success is p

$$f(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k} \qquad E[k] = Np \qquad V[k] = Np(1-p)$$

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}$$

binomial coefficient: number of different ways (permutations) to have k successes in N tries

Use binomial distribution to model processes with two outcomes

- Example: Detection efficiency (either we detect particle or not)

For small p, the binomial distribution can be approximated by a Poisson distribution (more exactly, in the limit $N \to \infty$, $p \to 0$, $N \cdot p$ constant)

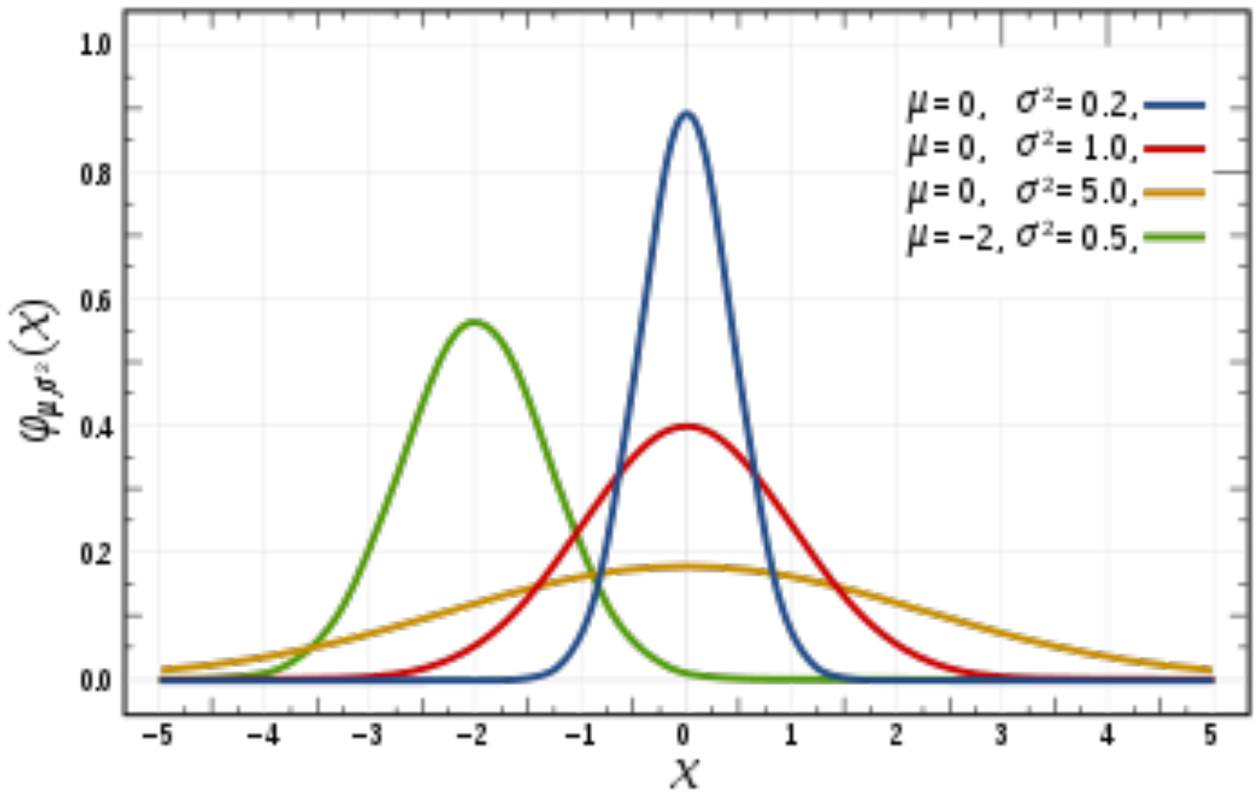
Gaussian Distribution

$$g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \mu$$

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

https://en.wikipedia.org/wiki/Normal_distribution

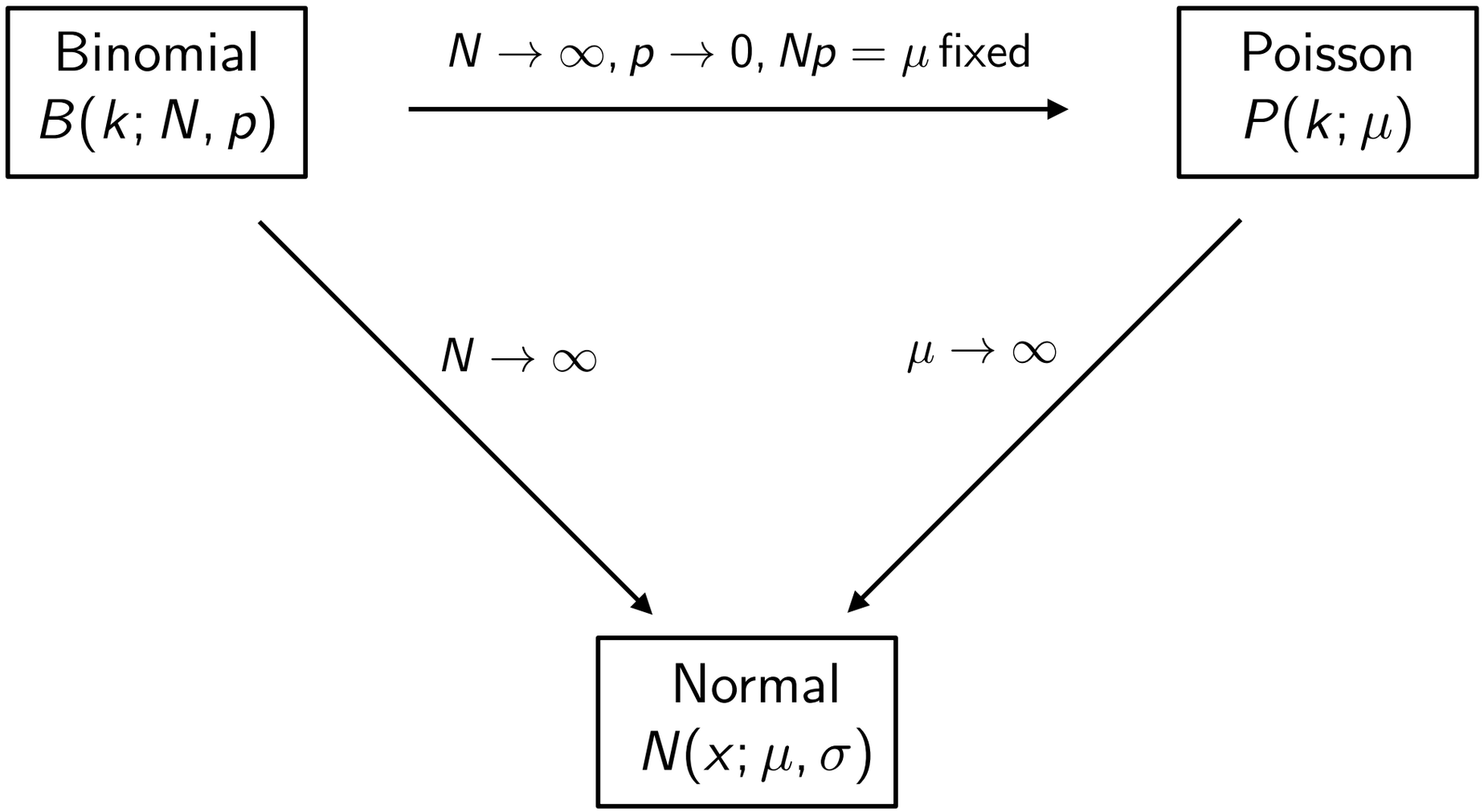


μ = 0, σ = 1 ("standard normal distribution, $N(0,1)$"): $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Cumulative distribution related to error function:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}\,\mathrm{d}z = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)+1\right]$$

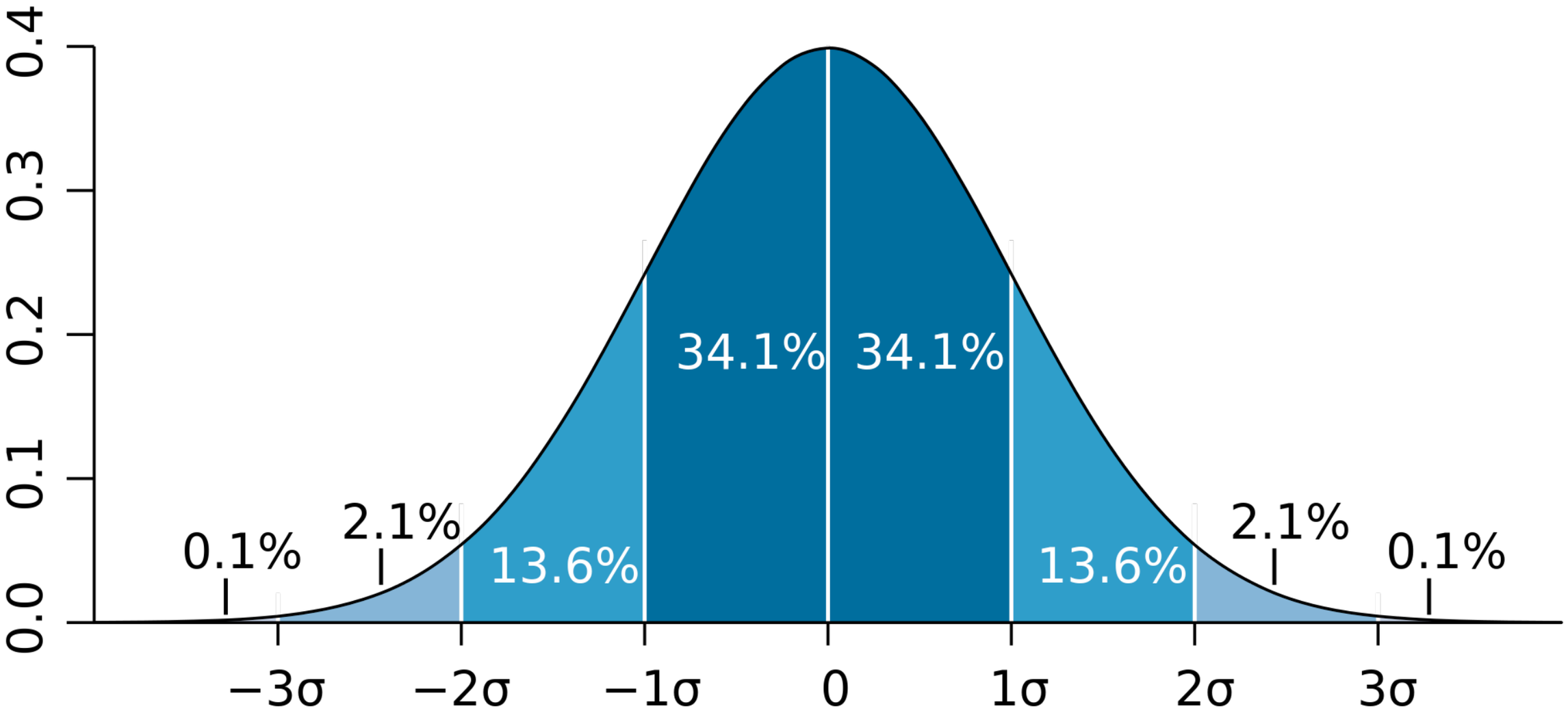
Binomial, Poisson and Normal Distribution



Poisson $P(k;\mu)$: $\quad \dfrac{k-\mu}{\sqrt{\mu}} \to N(0,1) \quad \text{as} \quad \mu \to \infty$

Binomial $B(k;n,p)$: $\quad \dfrac{k-np}{\sqrt{np(1-p)}} \to N(0,1) \quad \text{as} \quad n \to \infty$

Deviation in Units of σ for a Gaussian



$$P(Z\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-Z}^{+Z} e^{-\frac{x^2}{2}} \, \mathrm{d}x$$

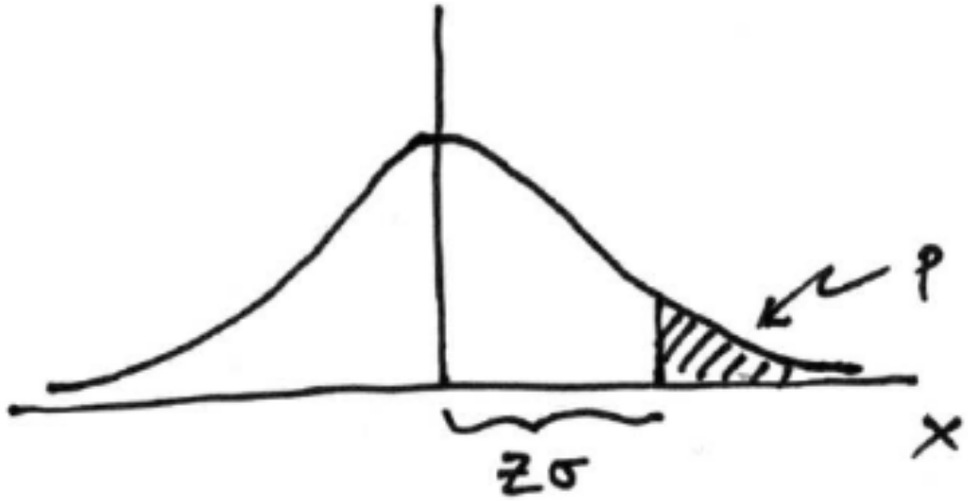
68.27% of area within ±1σ
95.45% of area within ±2σ
99.73% of area within ±3σ

90% of area within ±1.645σ
95% of area within ±1.960σ
99% of area within ±2.576σ

p-value and significance

p-value:
probability that a random process produces a measurement thus far, or further, from the true mean

p-value and significance (one-tailed):



$$p = 1 - \Phi(Z), \quad Z = \Phi^{-1}(1 - p)$$

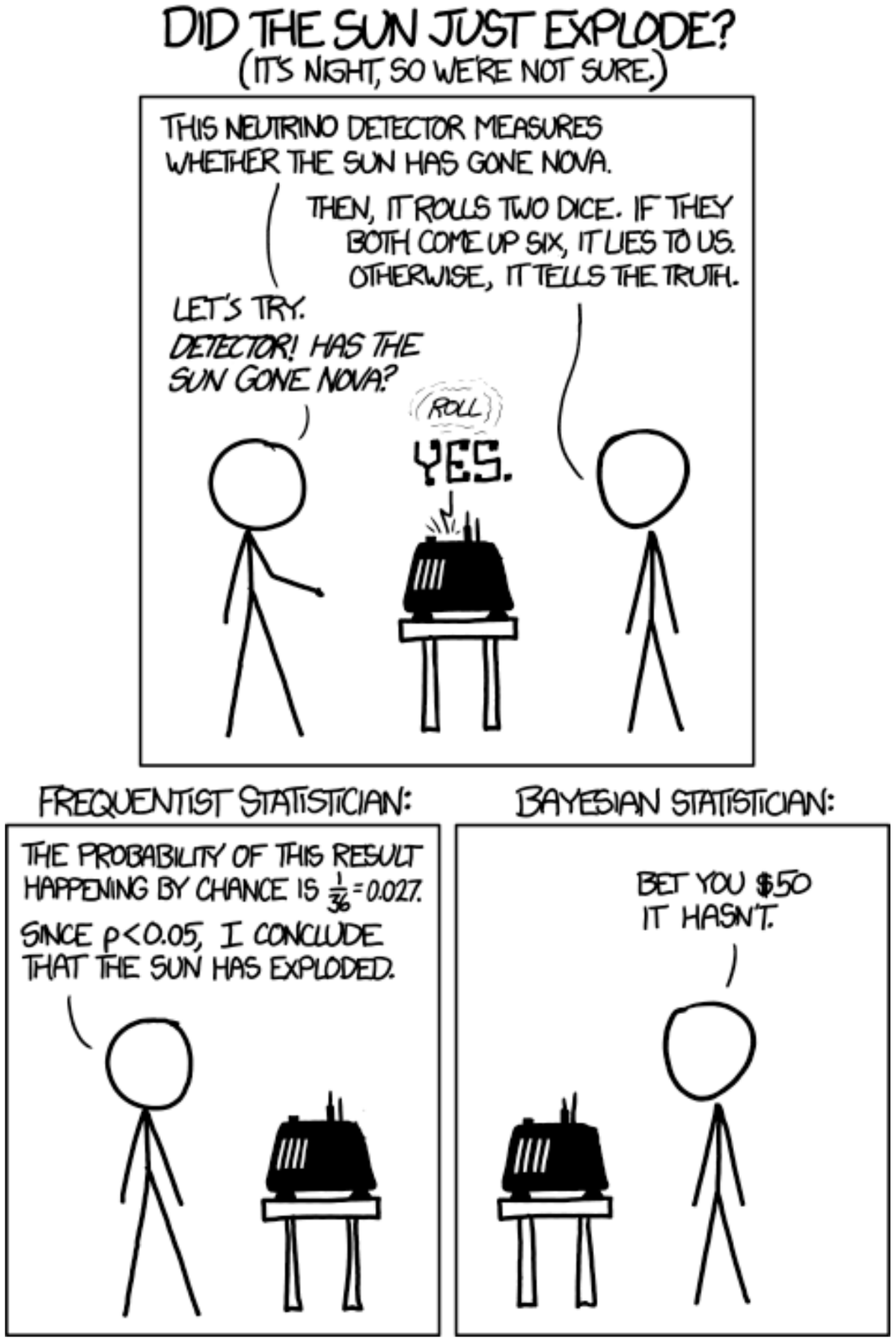
One-tailed Gaussian *p*-values

Deviation	p-value
1 σ	0.16
2 σ	0.023
3 σ	0.0013
4 σ	3.2×10^{-5}
5 σ	2.9×10^{-7}

standard to report a "discovery" →

notebook: p-values_and_n-sigma.ipynb

Frequentist vs. Bayesian Statistics



https://xkcd.com/1132/

Why 5σ for Discovery in Particle Physics?

5σ ⇔ p-value = 2.87×10^{-7} (one-tailed test)

History: There are many cases of 3σ and 4σ effects that have disappeared with more data

The Look-Elsewhere Effect

Systematics:

- Usually more difficult to estimate than statistical uncertainties
- "Safety margin"

Subconscious Bayes factor:

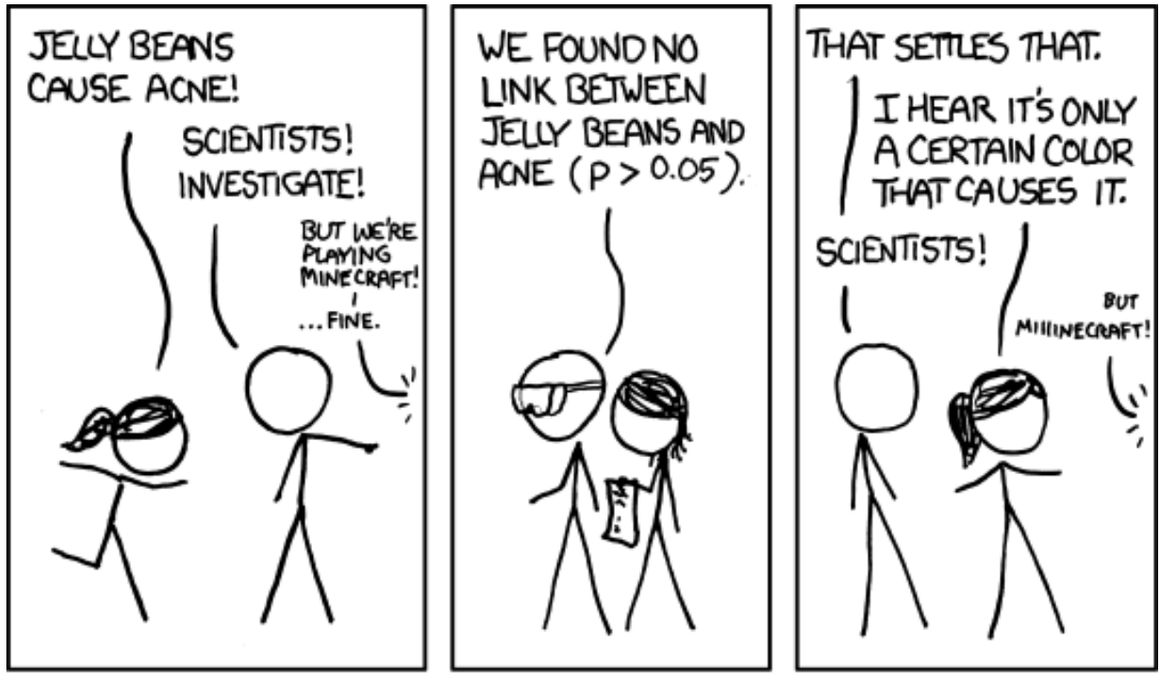
- Physicists subconsciously tend to assess the Bayesian probabilities $p(H_1|\text{data})$ and $p(H_0|\text{data})$
- If H_1 involves something very unexpected (e.g., neutrinos travel faster than the speed of light) then prior probability for null hypothesis H_0 is much larger than for H_1.
- "Extraordinary claims require extraordinary evidence"

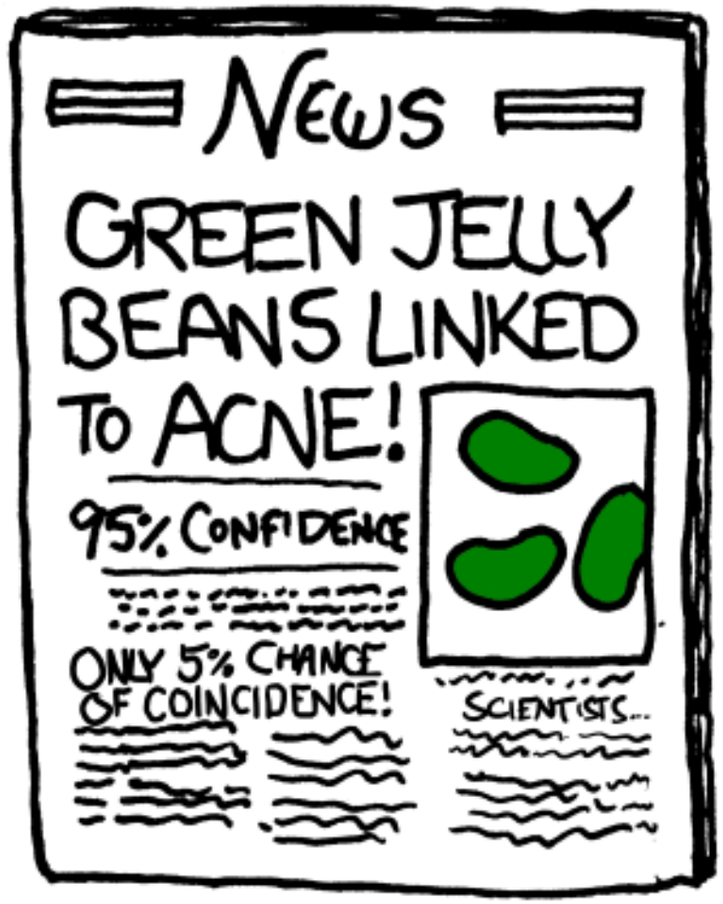
Last point ⇒ unreasonable to have a single criterion (5σ) for all experiments

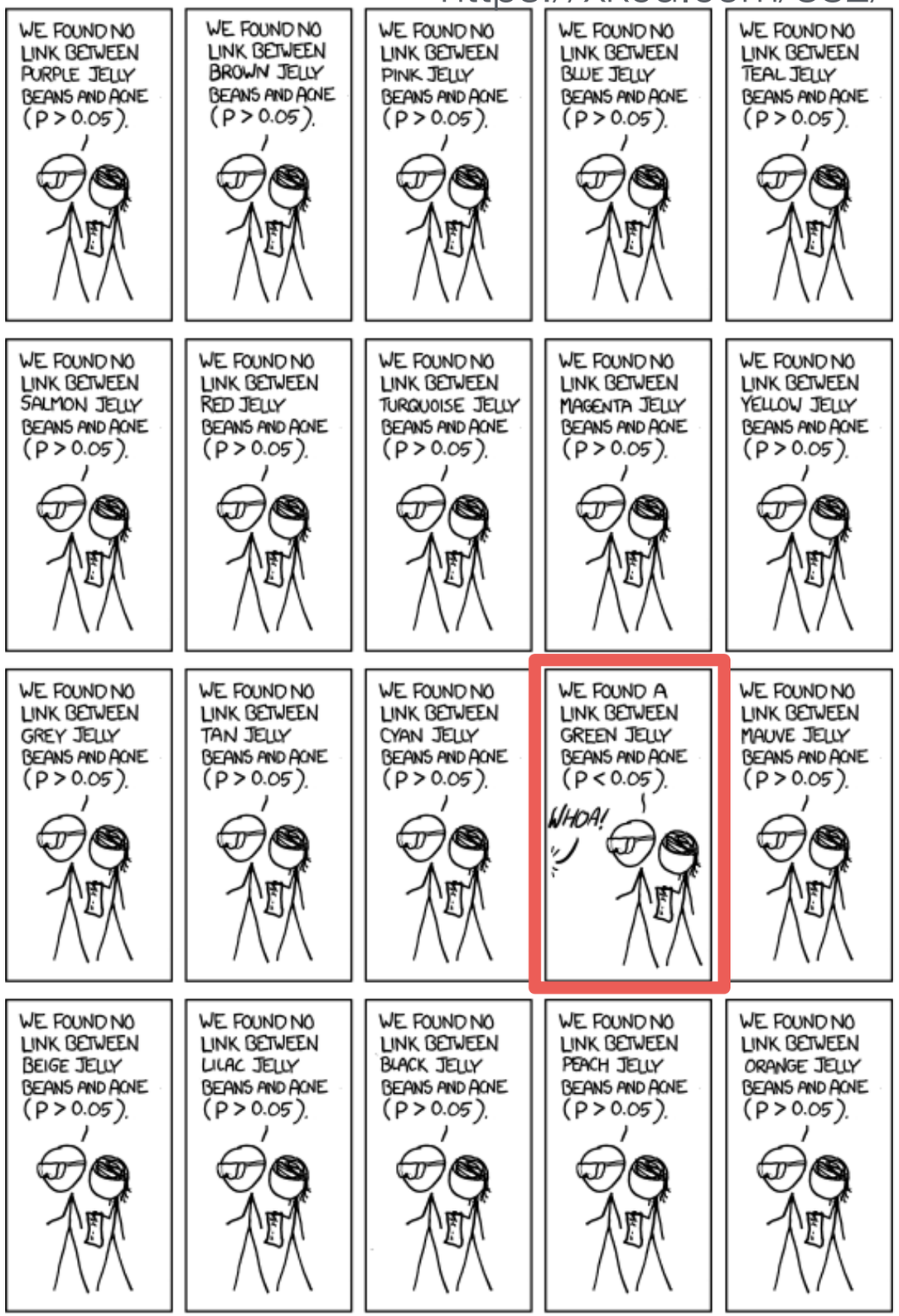
Louis Lyons, Statistical Issues in Searches for New Physics, arXiv:1409.1903

Look-Elsewhere Effect

https://xkcd.com/882/







Covariance and Correlation

Covariance ($\mu_x := \langle x \rangle, \mu_y := \langle y \rangle$):

$$\mathrm{cov}[x, y] = E[(x - \mu_x)(y - \mu_y)]$$

Correlation coefficient (dimensionless):

$$\rho_{xy} = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}$$

x, y independent, i.e., $f(x, y) = f_x(x) \cdot f_y(y)$:

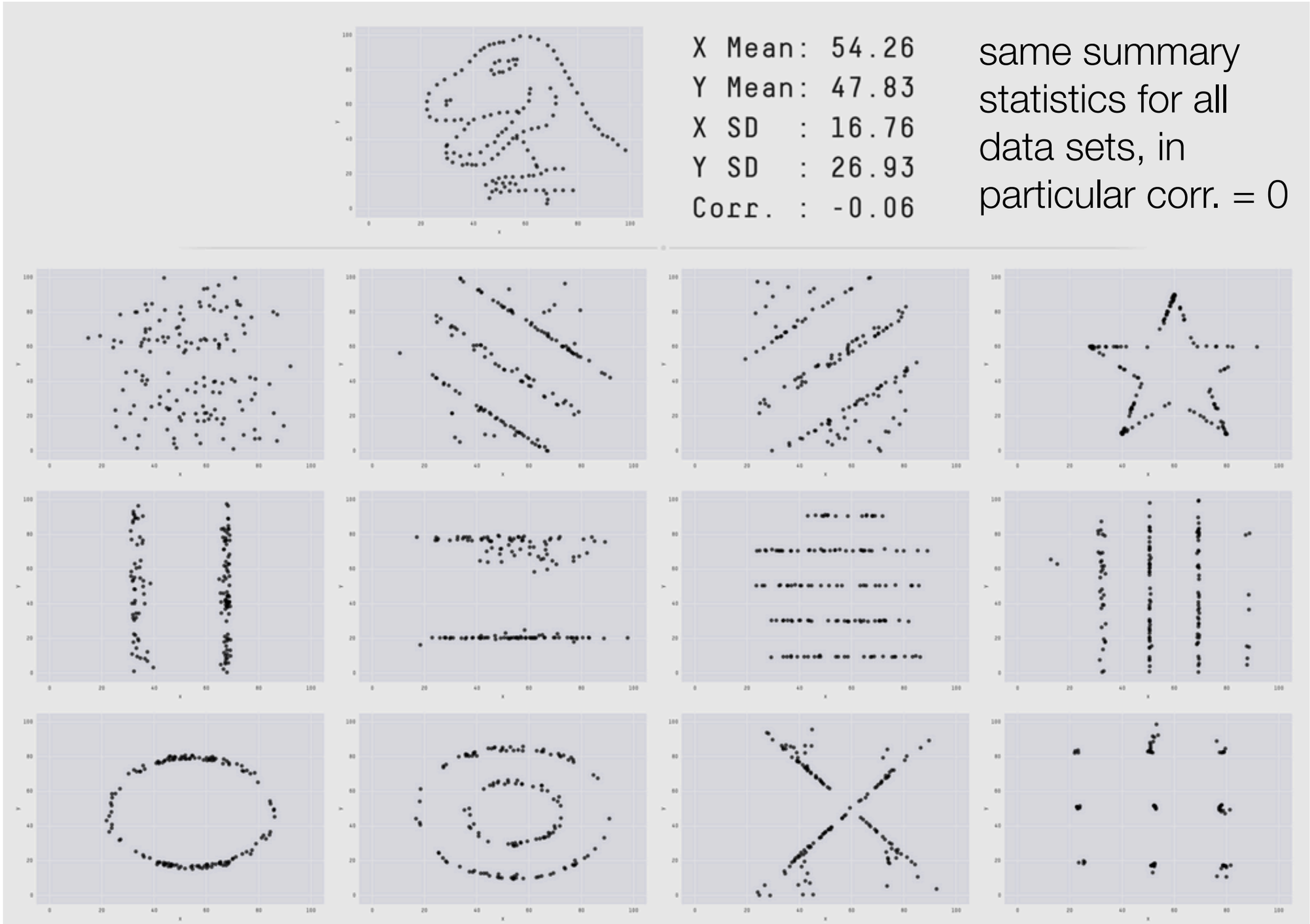
$$E[(x - \mu_x)(y - \mu_y)] = \int (x - \mu_x) f_x(x)\, \mathrm{d}x \int (y - \mu_y) f_y(y)\, \mathrm{d}y = 0$$

$\rightarrow \quad \mathrm{cov}[x, y] = 0$

N.B. converse not always true

Never trust summary statistics alone; always visualize your data

https://www.autodeskresearch.com/publications/samestats



Linear Combinations of Random Variables

Consider two random variables with known covariance cov(x, y):

$$\begin{aligned}
\langle x + y \rangle &= \langle x \rangle + \langle y \rangle \\
\langle ax \rangle &= a\langle x \rangle \\
V[ax] &= a^2 V[x] \\
V[x + y] &= V[x] + V[y] + 2\mathrm{cov}(x, y)
\end{aligned}$$

Calculation:

$$\begin{aligned}
V[x + y] &= E[(x + y - \mu_x - \mu_y)^2] = E[(x - \mu_x + y - \mu_y)^2] \\
&= E[(x - \mu_x)^2 + (y - \mu_y)^2 + 2(x - \mu_x)(y - \mu_y)] \\
&= E[(x - \mu_x)^2] + E[(y - \mu_y)^2] + 2E[(x - \mu_x)(y - \mu_y)] \\
&= V[x] + V[y] + 2\mathrm{cov}(x, y)
\end{aligned}$$

Reduction of the Standard Deviation for Repeated Independent Measurements

Consider the average of n independent observation x_i:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Expectation values and variance of the measurements:

$$E[x_i] = \mu_i \qquad V[x_i] = \sigma^2$$

Standard deviation of the mean:

$$V[\bar{x}] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{1}{n}\sigma^2 \qquad \rightarrow \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of the mean decreases as $1/\sqrt{n}$

Linear Error Propagation

Consider a measurement of values x_i and their covariances:

$$\vec{x} = (x_1, x_2, ..., x_n) \qquad V_{ij} = \mathrm{cov}[x_i, x_j]$$

Let y be a function of the x_i: $\quad y = f(\vec{x})$

What is the variance of y?

Approach: Taylor expansion of y around $\vec{\mu}$ where $\quad \mu_i = E[x_i]$

In practice we estimate μ_i by measured value x_i

$$V[y] \equiv \sigma_y^2 = E[y^2] - E[y]^2$$

Linear Error Propagation Formula

Taylor expansion:
$$y(\vec{x}) \approx y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

$E[y]$ is easy:
$$E[y] \approx y(\vec{\mu}) \quad \text{as} \quad E[x_i - \mu_i] = 0$$

$E[y^2]$:
$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} E[x_i - \mu_i]$$
$$+ E\left[\left(\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)\right)\left(\sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)\right)\right]$$
$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Thus:
$$\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Matrix Notation

Let vector J be given by $\quad \vec{J} = \vec{\nabla} y, \quad$ i.e., $J_j = \left(\frac{\partial y}{\partial x_j} \right)_{\vec{x}=\vec{\mu}}$

Then: $\quad \sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij} \quad = \quad \vec{J}^{\mathsf{T}} V \vec{J}$

Example: $\quad y = \frac{x_1}{x_2}, \quad \vec{J} = \begin{pmatrix} 1/x_2 \\ -x_1/x_2^2 \end{pmatrix}$

$$\sigma_y^2 = \left(\frac{1}{x_2}, -\frac{x_1}{x_2^2} \right) \begin{pmatrix} \sigma_1^2 & \mathrm{cov}[x_1, x_2] \\ \mathrm{cov}[x_1, x_2] & \sigma_2^2 \end{pmatrix} \begin{pmatrix} \frac{1}{x_2} \\ -\frac{x_1}{x_2^2} \end{pmatrix}$$

$$= \left(\frac{1}{x_2}, -\frac{x_1}{x_2^2} \right) \begin{pmatrix} \frac{\sigma_1^2}{x_2} - \frac{x_1}{x_2^2}\mathrm{cov}[x_1, x_2] \\ \frac{1}{x_2}\mathrm{cov}[x_1, x_2] - \frac{x_1}{x_2^2}\sigma_2^2 \end{pmatrix} = \frac{1}{x_2^2}\sigma_1^2 + \frac{x_1^2}{x_2^4}\sigma_2^2 - 2\frac{x_1}{x_2^3}\mathrm{cov}[x_1, x_2]$$

$$\rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2} = \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2\frac{\rho \sigma_1 \sigma_2}{x_1 x_2}$$

Linear Error Proportion: Examples

$$y = ax \quad \rightarrow \quad \sigma_y^2 = a^2\sigma_x^2 \qquad \text{i.e. } \sigma_y = |a|\sigma_x$$

$$y = x^n \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = n^2\frac{\sigma_x^2}{x^2} \qquad \text{i.e. } \frac{\sigma_y}{y} = |n|\frac{\sigma_x}{x}$$

$$y = x_1 + x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\mathrm{cov}[x_1, x_2]$$

$$y = x_1 - x_2 \quad \rightarrow \quad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 - 2\mathrm{cov}[x_1, x_2]$$

$$y = x_1 x_2 \quad \rightarrow \quad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$$

Sanity checks:

Average of fully correlated measurements:

$$y = \frac{1}{2}(x_1 + x_2),\ \sigma_1 = \sigma_2 \equiv \sigma,\ \rho = 1 \quad \rightsquigarrow \quad \sigma_y = \sigma$$

Difference of fully correlated measurements:

$$y = x_1 - x_2,\ \sigma_1 = \sigma_2 \equiv \sigma,\ \rho = 1$$

$$\rightsquigarrow \quad \sigma_y^2 = 2\sigma^2 - 2\sigma^2 = 0$$

Linear Error Propagation for Uncorrelated Measurements

Special case: the x_i are uncorrelated, i.e., $V_{ij} = \delta_{ij}\sigma_i^2$:

$$\sigma_y^2 = \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i}\right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2$$

These formulas are exact only for linear functions.

Approximation breaks down if function is nonlinear over a region comparable in size to the σ_i.

Linear Error Propagation: Generalization from $\mathbb{R}^n \to \mathbb{R}$ to $\mathbb{R}^n \to \mathbb{R}^m$

Generalization: Consider set of m functions:

$$\vec{y}(\vec{x}) = (y_1(\vec{x}), y_2(\vec{x}), \ldots, y_m(\vec{x}))$$

Then:

$$\mathrm{cov}[y_k, y_l] \equiv U_{kl} \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

In matrix notation:

$$U = JVJ^{\mathsf{T}} \qquad J_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

Multivariate Normal distribution

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{N/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^{\mathsf{T}} V^{-1}(\vec{x}-\vec{\mu})\right]$$

transposed (row) vector

column vector

$$\vec{x} = (x_1, ..., x_n), \qquad \vec{\mu} = (\mu_1, ..., \mu_n)$$

Mean: $E[x_i] = \mu_i$ Covariance: $\mathrm{cov}[x_i, x_j] = V_{i,j}$

For $n = 2$:

$$V = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \quad \rightsquigarrow \quad V^{-1} = \frac{1}{(1-\rho^2)} \begin{pmatrix} 1/\sigma_x^2 & -\rho/(\sigma_x\sigma_y) \\ -\rho/(\sigma_x\sigma_y) & 1/\sigma_y^2 \end{pmatrix}$$

ρ = correlation coefficient

2d Gaussian Distribution and Error Ellipse

2d Gaussian distribution:

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1, \sigma, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right)$$

where $\rho = \mathrm{cov}(x_1, x_2)/(\sigma_1\sigma_2)$ is the correlation coefficient.

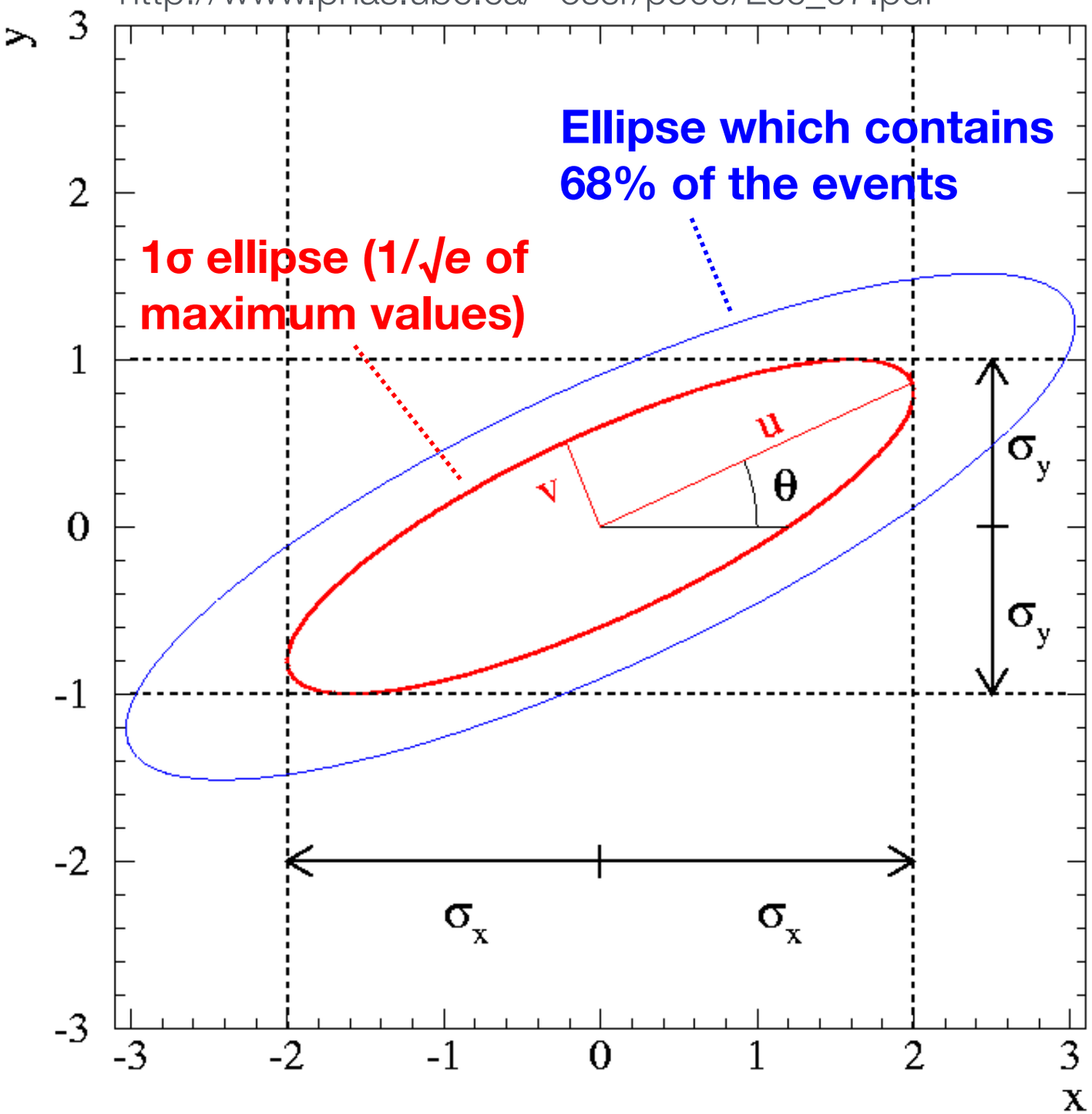
Lines of constant probability correspond to constant argument of exp
→ this defines an ellipse

1σ ellipse: $f(x_1, x_2)$ has dropped to $1/\sqrt{e}$ of its maximum value (argument of exp is –1/2):

$$\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) = 1-\rho^2$$

2d Gaussian: Error Ellipse

http://www.phas.ubc.ca/~oser/p509/Lec_07.pdf



$$f_y(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$

$$= \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right)$$

$$f_x(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)$$

	P_{1D}	P_{2D}
1σ	0.6827	0.3934
2σ	0.9545	0.8647
3σ	0.9973	0.9889
1.515σ		0.6827
2.486σ		0.9545
3.439σ		0.9973

Probability for an event to be within 1σ ellipse: 39.34%

Maximum Likelihood Method

Estimator

Suppose we have a measurement of *n* independent values

$$\vec{x} = (x_1, x_2, ..., x_n)$$

which follow the same underlying distribution f(*x*; θ),
e.g., f(*x*; θ) = 1/θ exp(–*x*/θ).

i.i.d. random variables = independent, identically distributed

An estimator is a function of the data which provides a numerical estimate of the parameter θ:

$$\hat{\theta}(\vec{x})$$

θ often is not only one parameter but a vector of parameters.

Properties of Estimators

Consistency

An estimator is consistent if it converges to the true value

$$\lim_{n\to\infty} \hat{\vec{\theta}} = \vec{\theta}$$

Bias

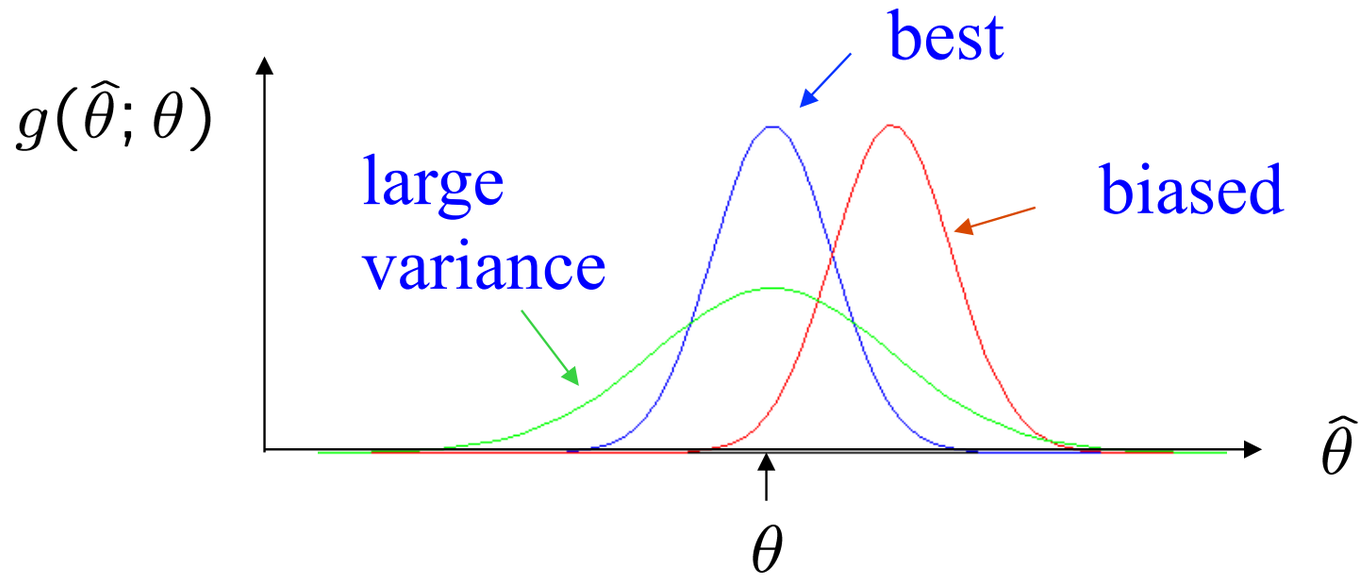
Difference btw. expectation value of estimator and true value

$$\vec{b} := E[\hat{\vec{\theta}}] - \vec{\theta}$$

Efficiency

An estimator is efficient if its variance V(θ) is small

efficient ⇔ Equal-sign in Cramér–Rao inequality holds



http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Example: Estimators for the lifetime of a particle

Estimator	Consistent?	Unbiased?	Efficient?
$\hat{\tau} = \frac{t_1+t_2+\ldots+t_n}{n}$	yes	yes	yes
$\hat{\tau} = \frac{t_1+t_2+\ldots+t_n}{n-1}$	yes	no	no
$\hat{\tau} = t_1$	no	yes	no

http://www.terascale.de/e149980/index_eng.html

Unbiased Estimator for Mean and Variance

Consider n independent and identically distributed measurements x_i drawn from a distribution with mean μ and standard deviation σ:

Estimator for the mean: $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mathrm{E}[\hat{\mu}] = \frac{1}{n}\,\mathrm{E}[\sum_i x_i] = \frac{1}{n}\sum_i \mathrm{E}[x_i] = \mu \quad \rightarrow \text{ estimator is unbiased}$$

$$\mathrm{V}[\hat{\mu}] = \mathrm{V}[\frac{1}{n}\sum_i x_i] = \frac{1}{n^2}\,\mathrm{V}[\sum_i x_i] = \frac{1}{n}\,\mathrm{V}[x] = \frac{\sigma^2}{n}, \text{ i.e., } \sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

Unbiased estimator for the variance:

$$s^2 := \hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Multiplying the sample variance by $n/(n–1)$ is known as Bessel's correction. Note that s is not an unbiased estimator of the standard deviation:

https://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation

Likelihood Function and Maximum Likelihood

Suppose we have a measurement of *n* independent values

$$\vec{x} = (x_1, x_2, ..., x_n)$$

drawn from the distribution

$$f(x; \vec{\theta}), \quad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$$

The joint pdf for the observed values $\vec{x}$ is given by:

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{"likelihood function"}$$

We consider measured values as fixed and the parameters as variables.

Principle of maximum likelihood

The best estimate of the parameters $\vec{\theta}$ is that value which maximizes the likelihood function

Likelihood function is not a probability density function

The integral of $L(\vec{x}, \vec{\theta})$ with respect to the parameter is not necessarily equal to unity ($L(\vec{x}, \vec{\theta})$ might not be integrable at all).

This is why $L(\vec{x}, \vec{\theta})$ is not a probability density function.

Example: exponential decay, one measurement at t = 1h.

$$L(\tau) = \frac{1}{\tau} e^{-t/\tau} \approx \frac{1}{\tau} \quad \text{as} \quad \tau \to \infty, \qquad \int_0^\infty L(\tau)\, d\tau \quad \text{not defined}$$

Note: With Jeffreys' prior 1/τ the posterior $L(\tau)\ \pi(\tau)$ is normalizable.

Maximum Likelihood Example 1: Exponential Decay

Consider exponential pdf: $$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$

Independent measurements drawn from this distribution: $t_1, t_2, \ldots, t_n$

Likelihood function: $$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$$

$L(\tau)$ is maximum when $\ln L(\tau)$ is maximum:

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

Find maximum:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Maximum Likelihood Example 2: Gaussian (I)

Consider $x_1, x_2, \ldots, x_n$ drawn from Gaussian(μ, σ^2)

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

Log-likelihood function:

$$\ln L(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f(x_i;\mu,\sigma^2) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right)$$

Derivatives w.r.t. μ and σ^2:

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2}$$

$$\frac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} \left(\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right)$$

Maximum Likelihood Example 2: Gaussian (II)

Setting the derivatives w.r.t. μ and σ^2 to zero and solving the equations:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

We find that the ML estimator for σ^2 is biased!

Maximum Likelihood Uncertainty

Consider maximum likelihood estimate of a parameter θ. Methods to estimate Uncertainty of θ:

1. $\sigma_{\hat{\theta}}$ **from Monte Carlo**
   Generate pseudo-data by sampling the assumed distribution using the ML estimate $\hat{\theta}$ as parameter

2. **Use minimum variance bound**

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{-\frac{\partial^2}{\partial^2 \theta} \ln L(\theta)}}$$

3. $\Delta \ln L = -1/2$ **method**:

$$\ln L(\hat{\theta} \pm \sigma) = \ln L(\hat{\theta}) - \frac{1}{2}$$

For Gaussian likelihood function all methods agree.

Method 3 usually gives asymmetric uncertainties (which are messy).

Likelihood Function and Minimum Variance Bound

Let's first consider likelihood function with only one parameter:

$$L(\vec{x};\theta) = L(x_1, x_2, ..., x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Let $\hat{\theta}(\vec{x})$ be an unbiased estimator of the parameter θ

It can be shown that the variance (of any unbiased estimator) satisfies:

$$V[\hat{\theta}] \geq \frac{1}{E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right]}$$

For a biased estimator this becomes

$$V[\hat{\theta}] \geq \frac{\left(1+\frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right]}$$

This bound is called Rao-Cramér-Frechet minimum variance bound (MVB)

MVB Example: Exponential Decay

Reminder:

$$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \quad \rightsquigarrow \quad \sum_{i=1}^{n} \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0 \quad \rightsquigarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Minimum variance bound (MVB):

$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n} \left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3} \right) = \frac{n}{\tau^2} - \frac{2}{\tau^3} \sum_{i=1}^{n} t_i = \frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right)$$

$$V[\hat{\tau}] \geq \frac{1}{E\left[-\frac{n}{\tau^2} \left(1 - \frac{2\hat{\tau}}{\tau} \right) \right]} = \frac{1}{-\frac{n}{\tau^2} \left(1 - \frac{2E[\hat{\tau}]}{\tau} \right)} = \frac{\tau^2}{n}$$

Uncertainty of the ML Estimator: Approximating the Minimum Variance Bound

In many cases it is impractical to calculate the MVB analytically. Instead, one uses the following approximation which is good for large n:

$$E\left[-\frac{\partial^2 \ln L}{\partial^2 \theta}\right] \approx -\frac{\partial^2 \ln L}{\partial^2 \theta}\bigg|_{\theta=\hat{\theta}}$$

The variance of the ML estimator is given by:

$$V[\hat{\theta}] = -\frac{1}{\frac{\partial^2 \ln L}{\partial^2 \theta}\Big|_{\theta=\hat{\theta}}}$$

Example: Exponential decay

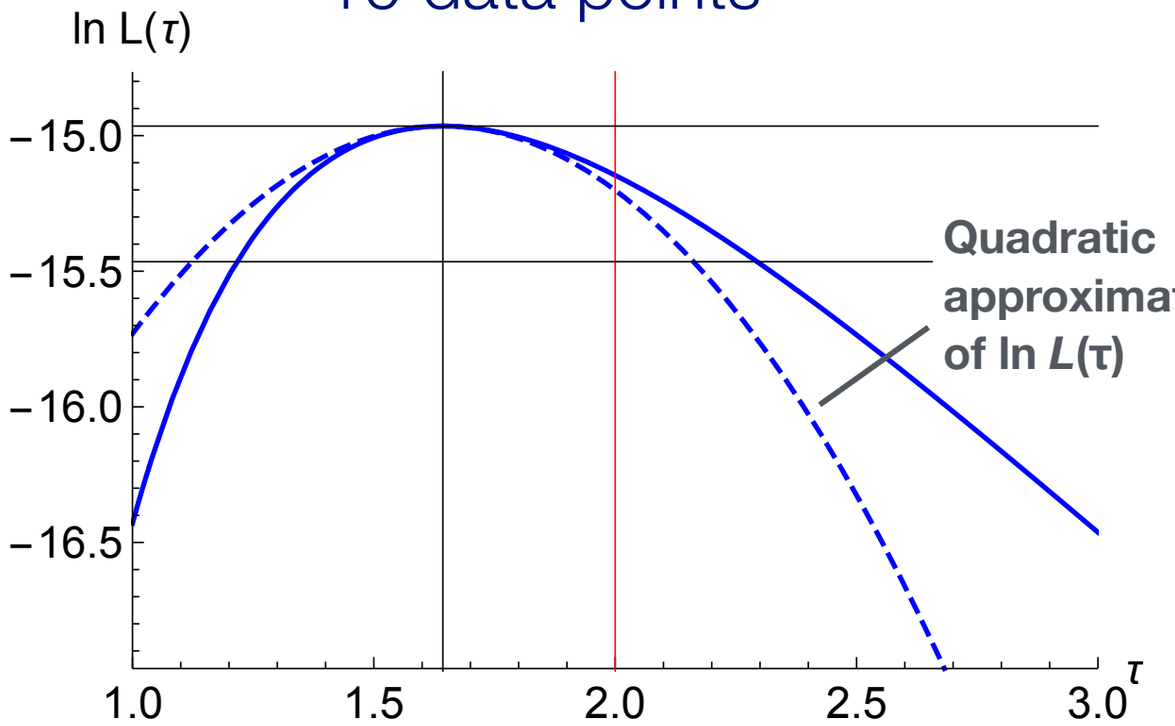
$$\frac{\partial^2 \ln L(\tau)}{\partial^2 \tau} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - 2\frac{t_i}{\tau^3}\right) = \frac{n}{\tau^2} - \frac{2}{\tau^3}\sum_{i=1}^{n} t_i = \frac{n}{\tau^2}\left(1 - \frac{2\hat{\tau}}{\tau}\right)$$

$$V[\hat{\tau}] = -\left(\frac{\partial^2 \ln L}{\partial^2 \theta}\right)^{-1}_{\tau=\hat{\tau}} = \frac{\hat{\tau}^2}{n} \qquad \rightsquigarrow \qquad \hat{\sigma} = \frac{\hat{\tau}}{\sqrt{n}}$$

Asymptotic Normality of the Likelihood function

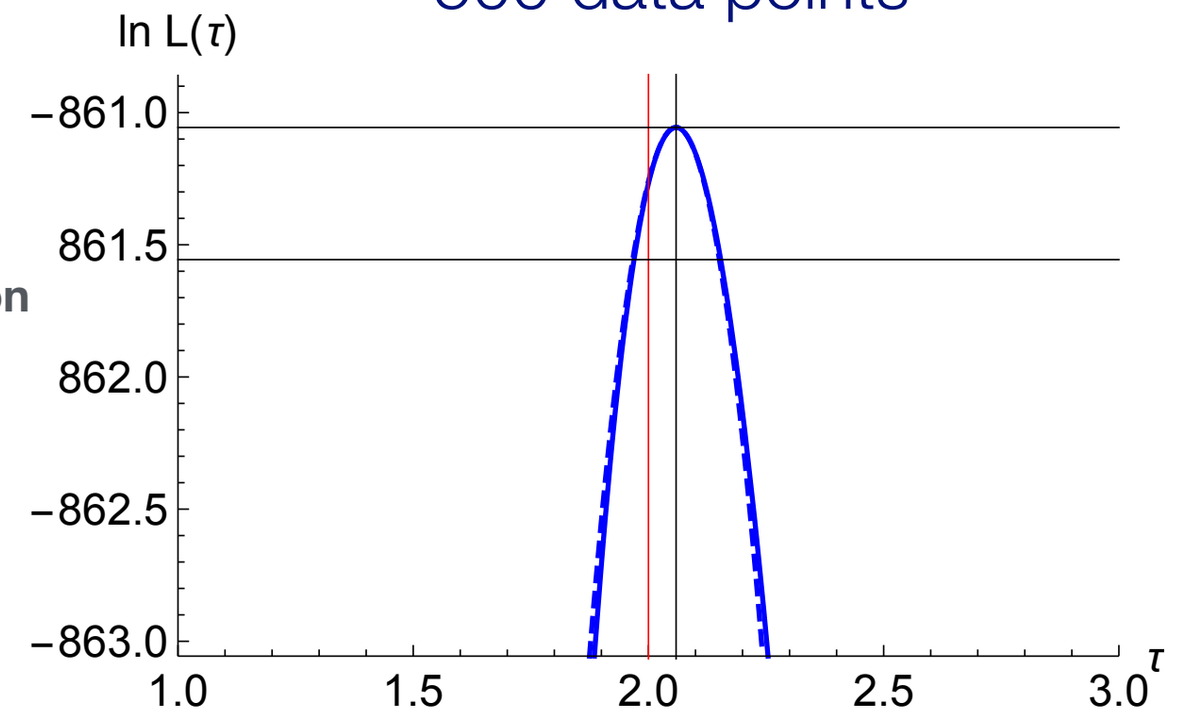
For any probability function $f(x;\theta)$ the likelihood function *L* approaches a Gaussian for large *n*, i.e., for a large number of events, and the variance of the ML estimator reaches the minimum variance bound.

10 data points



quadratic approximation of ln *L*(τ) is not very good

500 data points



quadratic approximation of ln *L*(τ) is excellent

Data points sampled from $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$ with $\tau = 2$

Uncertainty of the ML Estimator:

$\Delta \ln L = -1/2$ method

Taylor expansion of ln L around the maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \underbrace{\left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})}_{=0} + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial^2 \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

$-\frac{1}{\sigma^2}$

[from MVB, or from assuming Gaussian shape]

If $L(\theta)$ is approximately Gaussian ($\ln L(\theta)$ then is a approximately a parabola):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2_{\hat{\theta}}}}$$

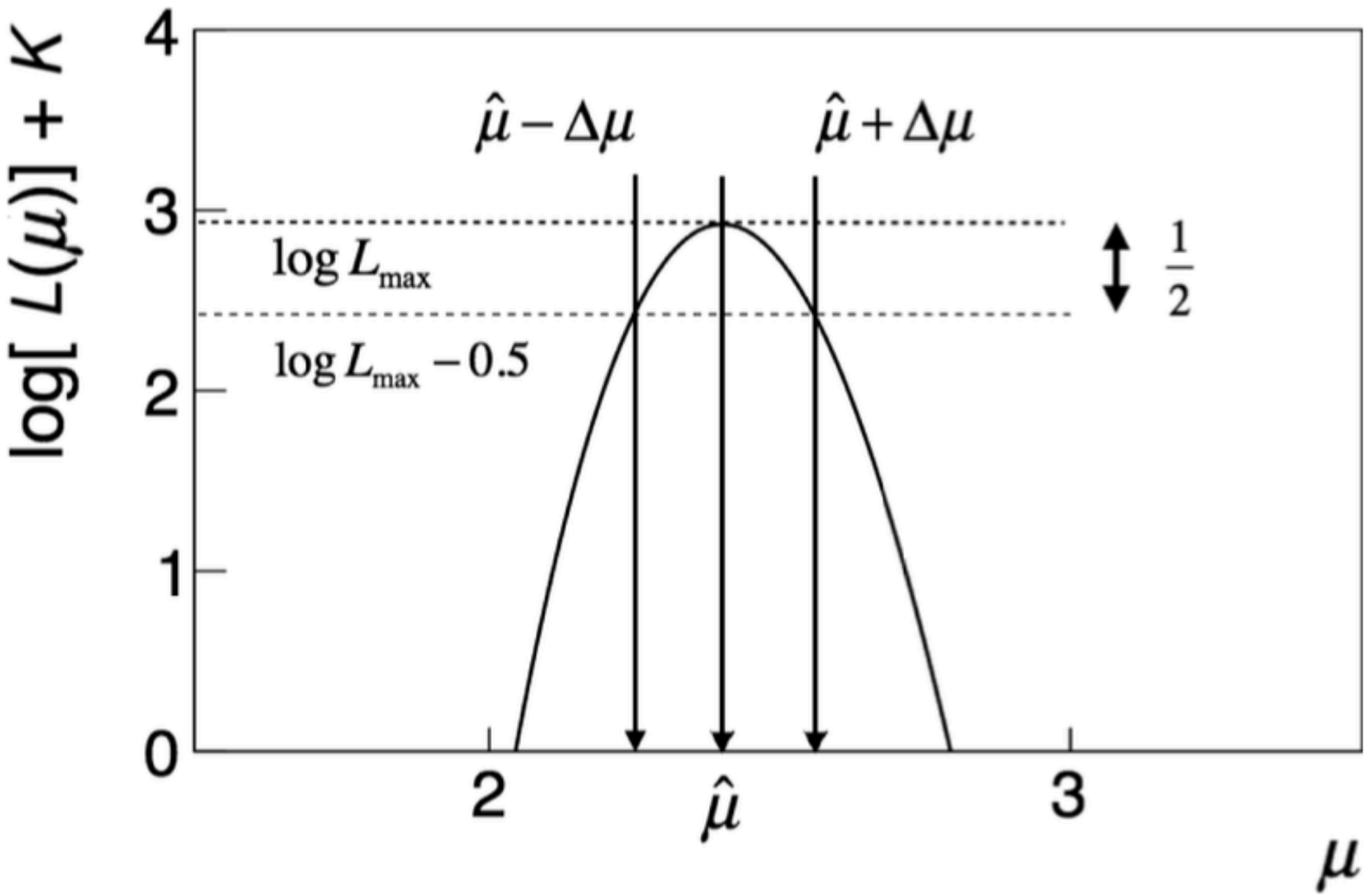
good approximation in the large sample limit

One can then estimate the uncertainties from the points where ln L has dropped by 1/2 from its maximum:

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

Illustration of the $\Delta \ln L = -1/2$ method

L is Gaussian $\longleftrightarrow$ ln *L* is a parabola



Properties of the ML Estimator

The ML estimator is consistent,
i.e., it approaches the true value in the limit of infinite measurements ($n \to \infty$)

ML estimator efficient for large n (you get the smallest possible variance)

For finite n the ML estimator is in general biased

ML efficiency theorem:
the ML estimator will be unbiased and efficient if an unbiased efficient estimator exists

The ML Estimator is invariant under parameter transformation:

$$\psi = g(\theta) \quad \Rightarrow \quad \hat{\psi} = g(\hat{\theta})$$

ML does not provide a goodness-of-fit measure.

Averaging Measurements with Gaussian Uncertainties

pdf for measurement (same mean, different σ):

$$f(x;\mu,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} \qquad \ln L(\mu) = \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}} - \ln \sigma_i - \frac{(x_i-\mu)^2}{2\sigma_i^2} \right)$$

Weighted average = ML estimate

$$\left.\frac{\partial \ln L(\mu)}{\partial \mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n} \frac{x_i - \hat{\mu}}{\sigma_i^2} \stackrel{!}{=} 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}$$

Uncertainty? In this case L is Gaussian and we can write it as

$$L(\mu) \propto e^{-\frac{(\mu-\hat{\mu})^2}{2\sigma_{\hat{\mu}}^2}} \quad \text{with} \quad \sigma_{\hat{\mu}}^2 = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

We obtain the formula for the weighted average:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} \quad \pm \quad \frac{1}{\sqrt{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}}$$

Minimum Variance Bound for *m* Parameters

$$f(x;\vec{\theta}), \quad \vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$$

Fisher information matrix $I(\vec{\theta})$ ($m \times m$ matrix):

$$I_{jk}[\vec{\theta}] = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ln L(x, \vec{\theta})\right]$$

Cramér-Rao-Frechet bound for an unbiased estimator then states that $V - I^{-1}$ is a positive-semidefinite matrix.

In particular one obtained for the variance:

$$V[\hat{\theta}_j] \geq (I(\vec{\theta})^{-1})_{jj}$$

Variance of the ML Estimator for *m* Parameters

For any probability function $f(x;\vec{\theta})$ the likelihood function L approaches a multi-variate Gaussian for large n

$$L(\vec{\theta}) \propto e^{-\frac{1}{2}(\vec{\theta}-\widehat{\vec{\theta}})^{\mathsf{T}}\, V^{-1}[\widehat{\vec{\theta}}]\,(\vec{\theta}-\widehat{\vec{\theta}})}$$

The variance of the ML estimator then reaches the MVB:

$$V[\widehat{\vec{\theta}}] \to I(\vec{\theta})^{-1}$$

Covariance matrix of the estimated parameters:

$$V[\widehat{\vec{\theta}}] \approx \left[-\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial^2 \vec{\theta}} \right|_{\vec{\theta}=\widehat{\vec{\theta}}} \right]^{-1}$$

or equivalently:

$$(V^{-1}[\widehat{\vec{\theta}}])_{ij} = -\left. \frac{\partial^2 \ln L(\vec{x};\vec{\theta})}{\partial\theta_i \partial\theta_j} \right|_{\vec{\theta}=\widehat{\vec{\theta}}}$$

Standard deviation of a single parameters:

$$\hat{\sigma}_{\hat{\theta}_j} = \sqrt{(V[\widehat{\vec{\theta}}])_{jj}}$$

Example: Two-Parameter ML Fit (from Cowan's Book)

Scattering angle distribution, $x = \cos\theta$: $\quad f(x; a, b) = \dfrac{1 + ax + bx^2}{2 + 2b/3}$

Normalization: $\displaystyle\int_{x_{\min}}^{x_{\max}} f(x; a, b)\,\mathrm{d}x = 1$

Example: $\quad a = 0.5,\ b = 0.5;\ x_{\min} = -0.95,\ x_{\max} = 0.95,\ 1000\,\mathrm{MC\ events}$

Numerical minimization with MINUIT:

$$\hat{a} = 0.53 \pm 0.08$$

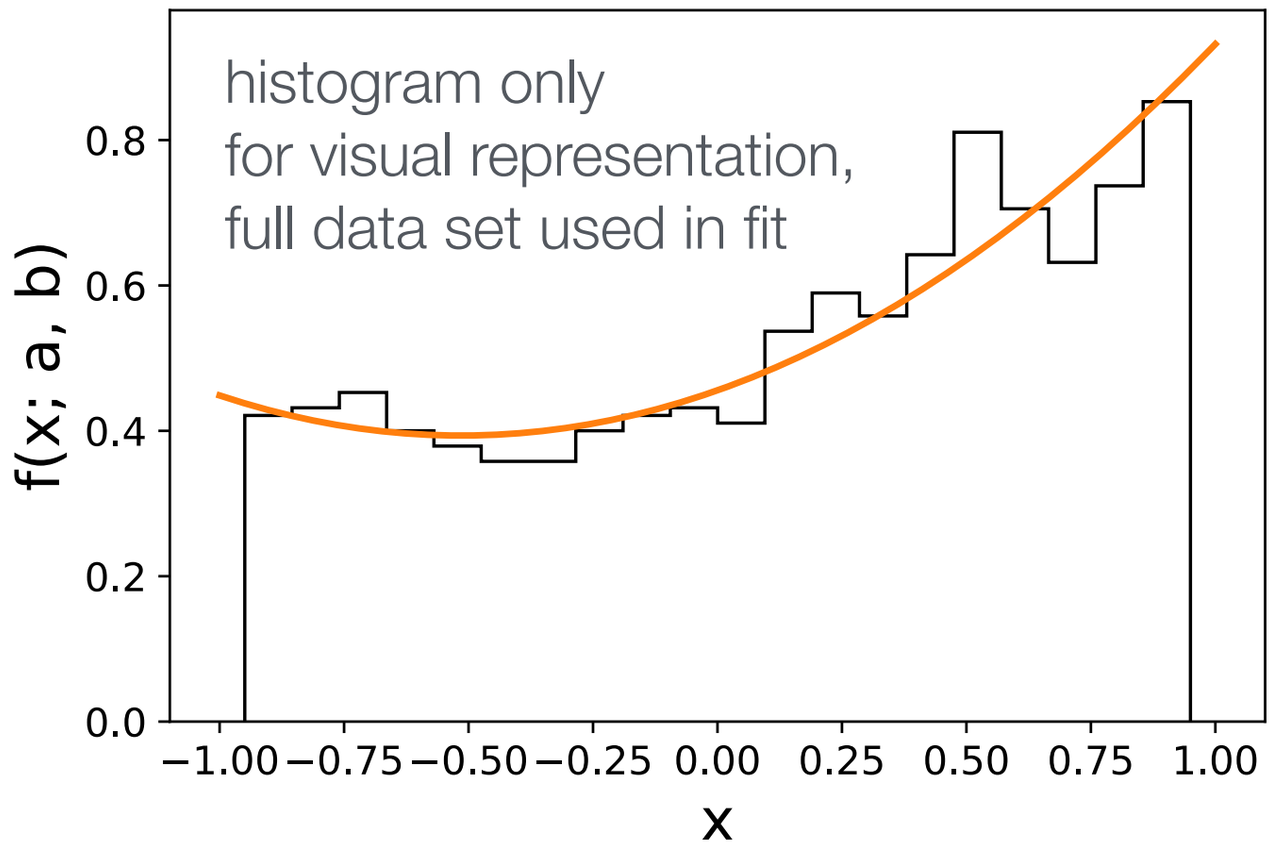
$$\hat{b} = 0.51 \pm 0.16$$

$$\mathrm{cov}[\hat{a}, \hat{b}] = 0.006$$

$$\rho = 0.48$$

Uncertainties and covariance from inverse of Hessian matrix H:

$$\widehat{V} = -H^{-1},\ (H)_{ij} = \left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$$



Example: Two-Parameter ML Fit (iminuit)

```
import numpy as np
import matplotlib.pyplot as plt
from iminuit import Minuit
```

```
x = np.loadtxt("data.txt")
```

```
def f(x, a, b):
    """normalized fit function"""
    xmin = -0.95
    xmax = 0.95
    return (6 * (1 + a * x + b * x * x)) /
           ((xmax - xmin) * (3 * a * (xmax + xmin) +
           2 * (3 + b * (xmax * xmax + xmax * xmin + xmin * xmin))))
```

```
def negative_log_likelihood(a, b):
    p = np.log(f(x, a, b))
    return -np.sum(p)
```

iminuit uses introspection to detect the parameter names of your function

```
m = Minuit(negative_log_likelihood,
           a=1, b=1, error_a=0.01, error_b=0.01, errordef=Minuit.LIKELIHOOD)
```

```
m.migrad()
```

Example: Two-Parameter ML Fit (iminuit)

```
m.migrad()
```

FCN = 606.5			Ncalls = 10 (146 total)	
EDM = 1.33e-10 (Goal: 0.0001)			up = 0.5	
Valid Min.	Valid Param.	Above EDM	Reached call limit	
True	True	False	False	
Hesse failed	Has cov.	Accurate	Pos. def.	Forced
False	True	True	True	False

	Name	Value	Hesse Error	Minos Error-	Minos Error+	Limit-	Limit+	Fixed
0	a	0.53	0.08					
1	b	0.51	0.16					

https://iminuit.readthedocs.io/en/stable/

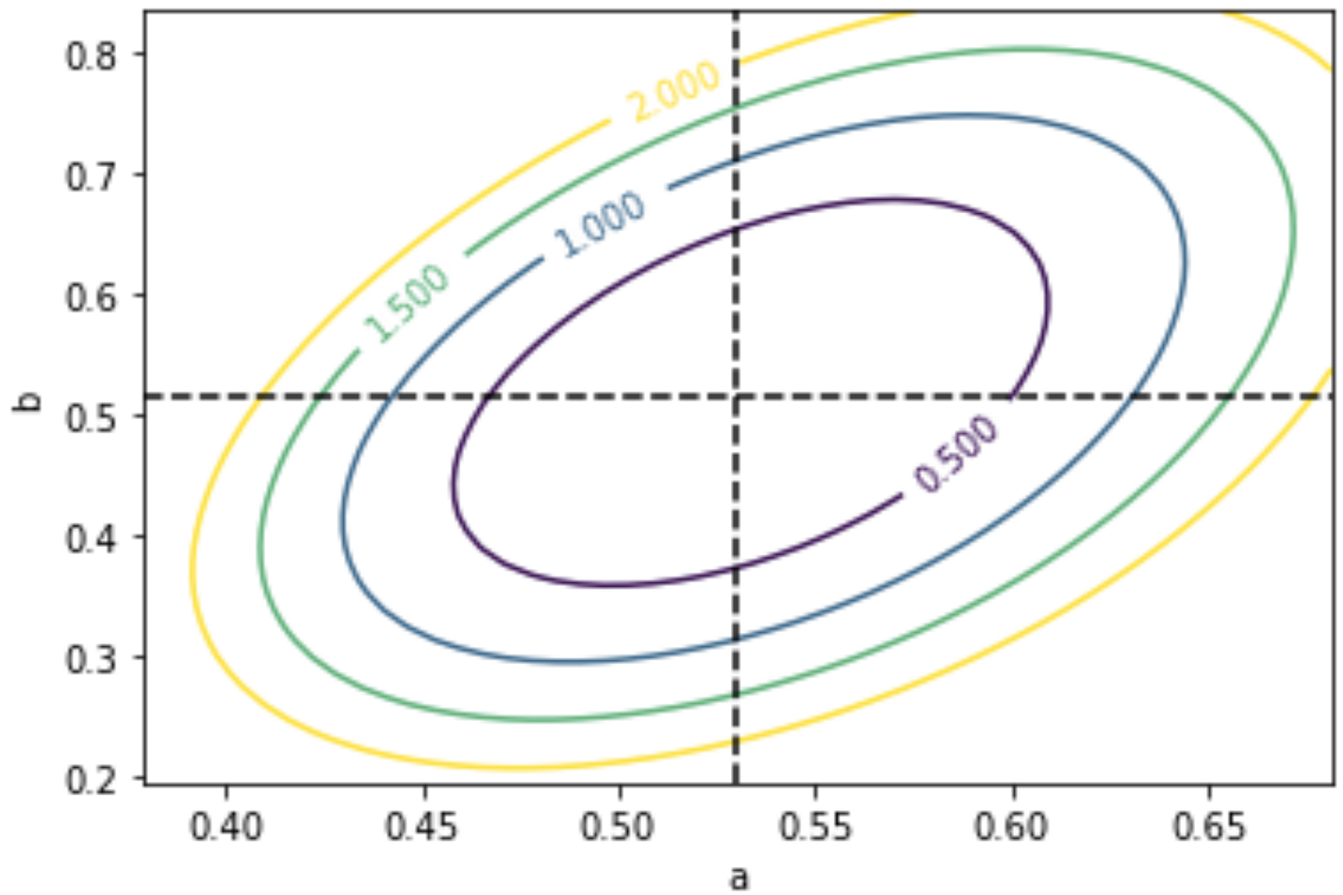
https://nbviewer.jupyter.org/github/scikit-hep/iminuit/blob/master/tutorial/basic_tutorial.ipynb

Example: Two-Parameter ML Fit (iminuit)

```
# covariance matrix
m.matrix()
```

	a	b
a	0.006	0.006
b	0.006	0.026

```
m.draw_contour('a','b');
```



Extended Maximum Likelihood Method (I)

Standard ML fit: information is in the shape of the distribution of the data x_i.

Extended ML fit: normalization becomes a fit parameter

Sometimes the number of observed events contains information about the parameters of interest, e.g., when we measure a rate.

Normal ML method:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Extended ML method:

$$\int q(x, \vec{\theta})\, \mathrm{d}x = \nu(\vec{\theta}) = \text{ predicted number of events}$$

Extended Maximum Likelihood Method (II)

Normalized pdf:

$$\int f(x, \vec{\theta})\, \mathrm{d}x = 1$$

Likelihood function:

$$L(\vec{\theta}) = \frac{\nu^n e^{-\nu}}{n!} \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \text{where} \;\; \nu \equiv \nu(\vec{\theta})$$

Log-Likelihood function:

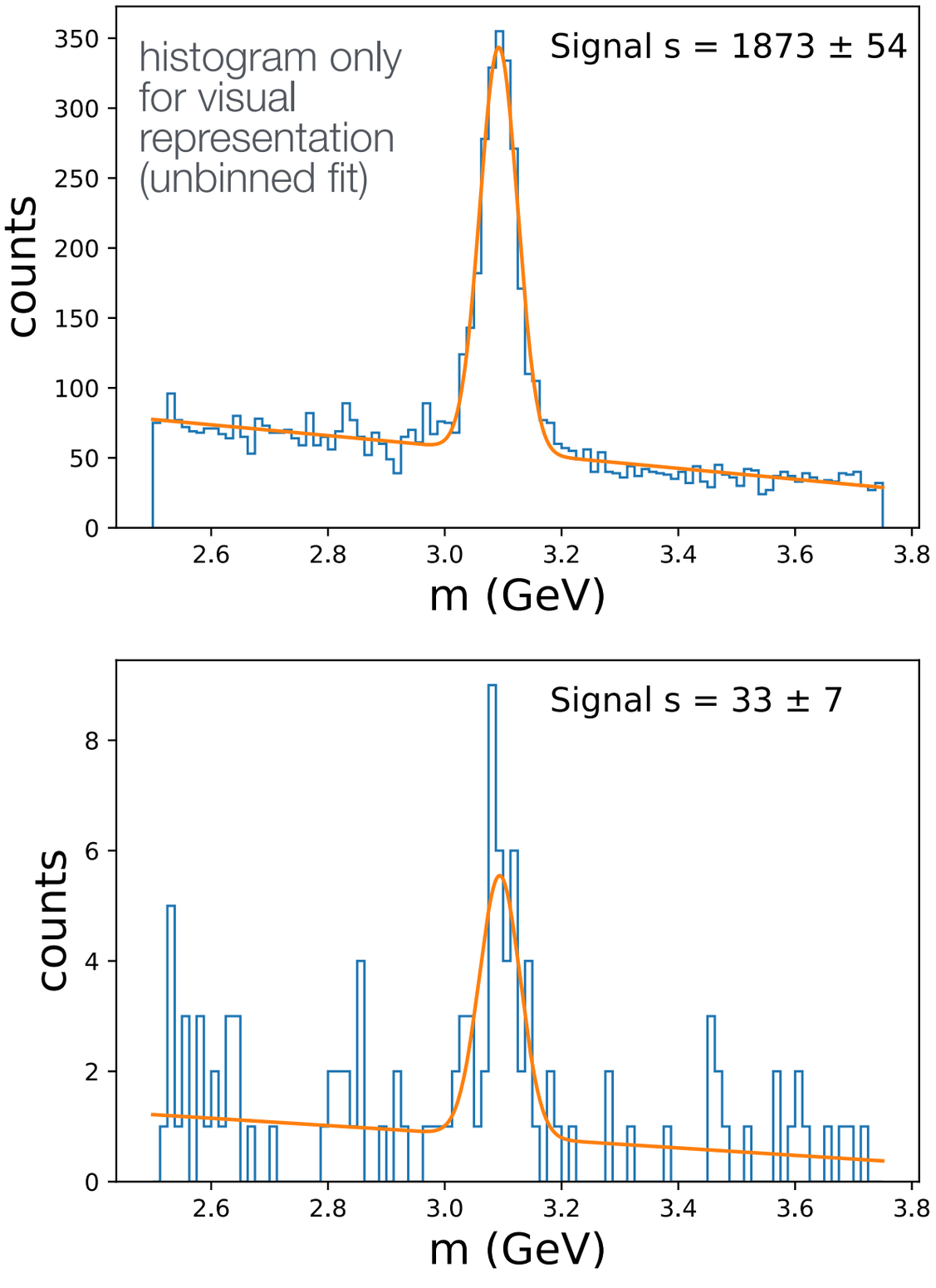
$$\ln L(\vec{\theta}) = -\ln(n!) - \nu(\vec{\theta}) + \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

ln(n!) does not depend on the parameters. So we need to minimize:

$$-\ln \tilde{L}(\vec{\theta}) = \nu(\vec{\theta}) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})\nu(\vec{\theta})]$$

prediction for total number of events

Application of the Extended ML Method: Linear Combination of Signal and Background PDF (I)



Two-component fit (signal + linear background)

Parameters:
- signal counts *s*
- background counts *b*
- linear background (slope, intercept)
- Gaussian peak: μ, σ

Normalized pdf:

$$f(x; r, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

negative log-likelihood:

$$-\ln \tilde{L}(\vec{\theta}) = s + b - n \ln(s + b) - \sum_{i=1}^{n} \ln[f(x_i; \vec{\theta})]$$

$$\nu(s, b) = s + b, \quad r = \frac{s}{s + b}$$

Unbinned ML fit works fine also in case of low statistics

Application of the Extended ML Method: Linear Combination of Signal and Background PDF (II)

Discussion:

We could have just fitted the normalized pdf:

$$f(x; r_s, \vec{\theta}) = r\, f_s(x, \vec{\theta}) + (1 - r)\, f_b(x, \vec{\theta})$$

Good estimate of the number of signal events: $n_{\text{signal}} = r\, n$

However, $\sigma_r\, n$ is not a good estimate of the variation of the number of signal events (ignores fluctuations of n)

[C. Blocker, Maximum Likelihood Primer]

(Trivial) example (L. Lyons):
96 protons and 4 heavy nuclei have been measured in a cosmic ray experiment

	protons	heavy nuclei
ML estimate	96 ± 2	4 ± 2
Extended ML estimate	96 ± 10	4 ± 2

Maximum Likelihood Fits with Binned Data (I)

Common practice: data put into a histogram: $\vec{n} = (n_1, ..., n_k),\ n_{\text{tot}} = \sum_{i=1}^{k} n_i$

Model prediction for the expected counts in bin i for fixed n_{tot}:

$$\nu_i(\vec{\theta}) = n_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta})\, \mathrm{d}x \qquad\qquad \nu_i(\vec{\theta}) = (\nu_1, ..., \nu_k)$$

If n_{tot} is fixed the probability to get a certain $\vec{n}$ is given by the multinomial distribution.

Multinomial distribution (generalization of binomial distribution):

→ k different possible outcomes, probability for outcome i is p_i, $\sum_{i=1}^{k} p_i = 1$

$$f(\vec{n}; n_{\text{tot}}, \vec{p}) = \frac{n_{\text{tot}}!}{n_1! \cdot ... \cdot n_k!}\, p_1^{n_1} \cdot ... \cdot p_k^{n_k} \qquad\qquad \vec{p} = (p_1, ..., p_k)$$

Maximum Likelihood Fits with Binned Data (II)

With $p_i = \nu_i / n_{\text{tot}}$ we write the likelihood of a certain $n_1, \ldots, n_k$ outcome as:

$$L(\vec{\theta}) = \frac{n_{\text{tot}}!}{n_1! \cdot \ldots \cdot n_k!} \left(\frac{\nu_1}{n_{\text{tot}}}\right)^{n_1} \cdot \ldots \cdot \left(\frac{\nu_k}{n_{\text{tot}}}\right)^{n_k} \qquad \nu_i(\vec{\theta}) = (\nu_1, \ldots, \nu_k)$$

Log-likelihood function:

$$\ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i(\vec{\theta}) + C$$

Limit of zero bin width → usual unbinned maximum likelihood method

Treat the n_i as Poisson-distributed (n_{tot} fluctuates,
predicted average $\nu_{\text{tot}} = \nu_1 + \nu_2 + \ldots + \nu_k$ → extended log-likelihood:

$$L(\vec{\theta}) = \prod_{i=1}^{k} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i} \quad \rightarrow \quad \ln L(\vec{\theta}) = \sum_{i=1}^{k} n_i \ln \nu_i - \nu_i = -\nu_{\text{tot}} + \sum_{i=1}^{k} n_i \ln \nu_i$$

Relation to Bayesian Parameter Estimation

Bayesian posterior distribution:

$$p(\vec{\theta};\vec{x}) = \frac{L(\vec{x};\vec{\theta})\pi(\vec{\theta})}{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}}$$

Posterior distribution contains all information about the estimated parameters.

Often the mode (most probable value) of the posterior distribution is reported

→ Coincides with ML estimate for a flat prior distribution

Marginalization in case one is interested in only one parameter of the Bayesian posterior distribution:

$$p(\theta_j;\vec{x}) = \int p(\vec{\theta};\vec{x})\,\mathrm{d}\vec{\theta}_{k\neq j} = \frac{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}_{k\neq j}}{\int L(\vec{x};\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}}$$

The Method of Least Squares

Least Squares from ML (I)

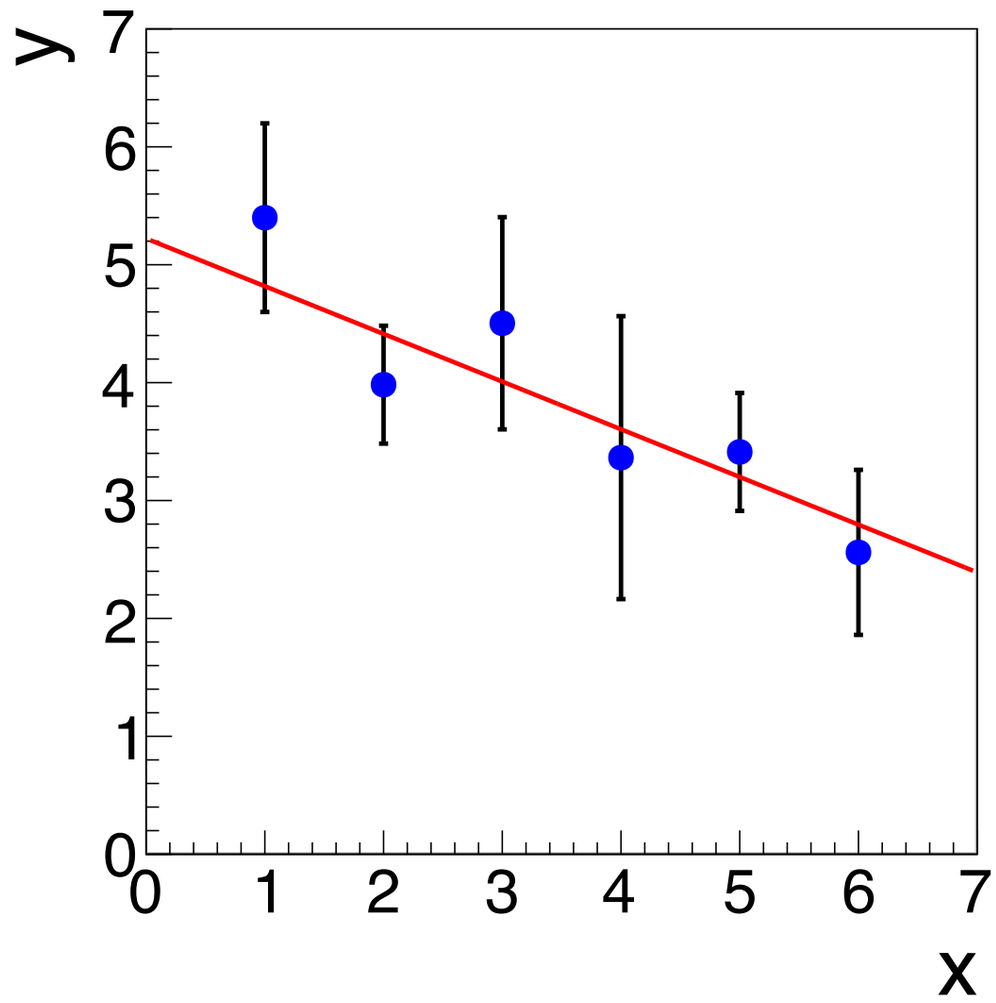
Consider n measured values $y_1(x_1)$, $y_2(x_2)$, … , $y_n(x_n)$ assumed to be independent Gaussian random variables with known variances:

$$V[y_i] = \sigma_i^2$$

Assume we have a function f with

$$E[y_i] = f(x_i; \vec{\theta})$$

We want to estimate $\vec{\theta}$



Likelihood function:

$$L(\vec{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y_i - f(x_i; \vec{\theta})}{\sigma_i}\right)^2\right]$$

Least Squares from ML (II)

Log-likelihood function:

$$\ln L(\vec{\theta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2 + \text{terms not depending on } \vec{\theta}$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{n}\left(\frac{y_i - f(x_i;\vec{\theta})}{\sigma_i}\right)^2$$

Minimizing χ^2 is called the method of least squares, goes back to Gauss and Legendre.

In other words, for Gaussian uncertainties the method of least squares coincides with the maximum likelihood method.

Minimization: $\dfrac{\partial\chi^2}{\partial\theta_j} = 0, \qquad j = 1, \ldots, m$ — Number of parameters

The χ^2 minimization is often done numerically, e.g., using the MINUIT code

https://en.wikipedia.org/wiki/MINUIT

Generalized Least Squares for Correlated y_i

Suppose the y_i have a covariance matrix V and follow a multi-variate Gaussian:

$$g(\vec{y};\vec{\mu},V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{y}-\vec{\mu})^{\mathsf{T}} V^{-1}(\vec{y}-\vec{\mu})\right]$$

The generalized least-squares method then corresponds to minimizing:

$$\chi^2(\vec{\theta}) = (\vec{y}-\vec{f}(\vec{x};\vec{\theta}))^T V^{-1}(\vec{y}-\vec{f}(\vec{x};\vec{\theta}))$$

$$\vec{f}(\vec{x};\vec{\theta}) = (f(x_1;\vec{\theta}),\ldots,f(x_n;\vec{\theta}))$$

We can write this also as

$$\chi^2(\vec{\theta}) = \sum_{i,j}(y_i - f(x_i;\vec{\theta}))^T (V^{-1})_{ij}(y_j - f(x_j;\vec{\theta}))$$

Variance of the Least Squares Estimators

Using

$$\chi^2(\vec{\theta}) = -2 \ln L(\theta) + \text{const.}$$

we can use the result for the variance of the ML estimators and obtain

$$V[\hat{\vec{\theta}}] \approx 2 \left[\left. \frac{\partial^2 \chi^2(\vec{\theta})}{\partial^2 \vec{\theta}} \right|_{\vec{\theta}=\hat{\vec{\theta}}} \right]^{-1} \quad \text{i.e.} \quad (V^{-1}[\hat{\vec{\theta}}])_{ij} = \frac{1}{2} \left. \frac{\partial^2 \chi^2(\vec{x};\vec{\theta})}{\partial\theta_i \partial\theta_j} \right|_{\vec{\theta}=\hat{\vec{\theta}}}$$

Or determine 1σ uncertainties from the contour where

$$\chi^2(\vec{\theta}') = \chi^2_{\text{min}} + 1$$

For z·σ uncertainties the condition is

$$\chi^2(\vec{\theta}') = \chi^2_{\text{min}} + z^2$$

Linear Least Squares

Consider n data points y_i whose uncertainties and correlations are described by a covariance matrix V. The y_i are measured at points x_i.

We would like to fit a linear combination of m functions $a_i(x)$ to the data:

$$f(x;\vec{\theta}) = \sum_{j=1}^{m} \theta_j a_j(x)$$

n data points y_i

m parameters θ_j

examples:

$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2$

$f(x) = \theta_0 + \theta_1 \cos(x)$

The linear least squares problem can be solved in closed form:

Define $n \times m$ matrix A: $A_{i,j} = a_j(x_i)$ "design matrix"

Minimize $\chi^2 = (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta}), \qquad \vec{y} = (y_1, ..., y_n)$

best fit parameters:

$$\hat{\vec{\theta}} = \underbrace{(A^T V^{-1} A)^{-1}}_{\text{symmetric } m \times m \text{ matrix}} A^T V^{-1} \vec{y}$$

covariance matrix of the parameters:

$$U = (A^T V^{-1} A)^{-1}$$

Non-linear Least Squares

Use numerical minimization program like MINUIT if the model is not linear in the parameters.

MINUIT's MIGRAD minimizer: Quasi-Newton Method
https://en.wikipedia.org/wiki/Quasi-Newton_method

See also: Gauss–Newton, Levenberg–Marquardt, …

Choice of initial values of the fit parameters important to converge to the correct solution.

Often numerical minimization program is also used in the linear case for convenience.

"Minuit2 has good performance compared to other minimisers, and it is one of the few codes out there which compute error estimates for your parameters."

https://iminuit.readthedocs.io/en/stable/

Example: Straight Line Fit: $y = \theta_0 + \theta_1 \cdot x$ (I)

The conditions $d\chi^2/d\theta_0$ and $d\chi^2/d\theta_1$ give two linear equations with two variables which is easy to solve.

Here we use the general solution for linear least squares fits:

$$L = (A^T V^{-1} A)^{-1} A^T V^{-1} \qquad \hat{\vec{\theta}} = L\vec{y}$$

$$A^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \qquad \vec{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \qquad V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & & \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ & & & 1/\sigma_n^2 \end{pmatrix}$$

$$A^T V^{-1} = \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \dots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \dots & x_n/\sigma_n^2 \end{pmatrix}$$

$$A^T V^{-1} A = \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \dots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \dots & x_n/\sigma_n^2 \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} \sum_i \frac{1}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{x_i^2}{\sigma_i^2} \end{pmatrix}$$

Example: Straight Line Fit: $y = \theta_0 + \theta_1 \cdot x$ (II)

The 2×2 matrix is easy to invert:

$$(A^T V^{-1} A)^{-1} = \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} [x^2] & -[x] \\ -[x] & [1] \end{pmatrix}$$

shorthand notation for the sum

$$\text{where} \quad [z] := \sum_i \frac{z_i}{\sigma_i^2}$$

This gives:

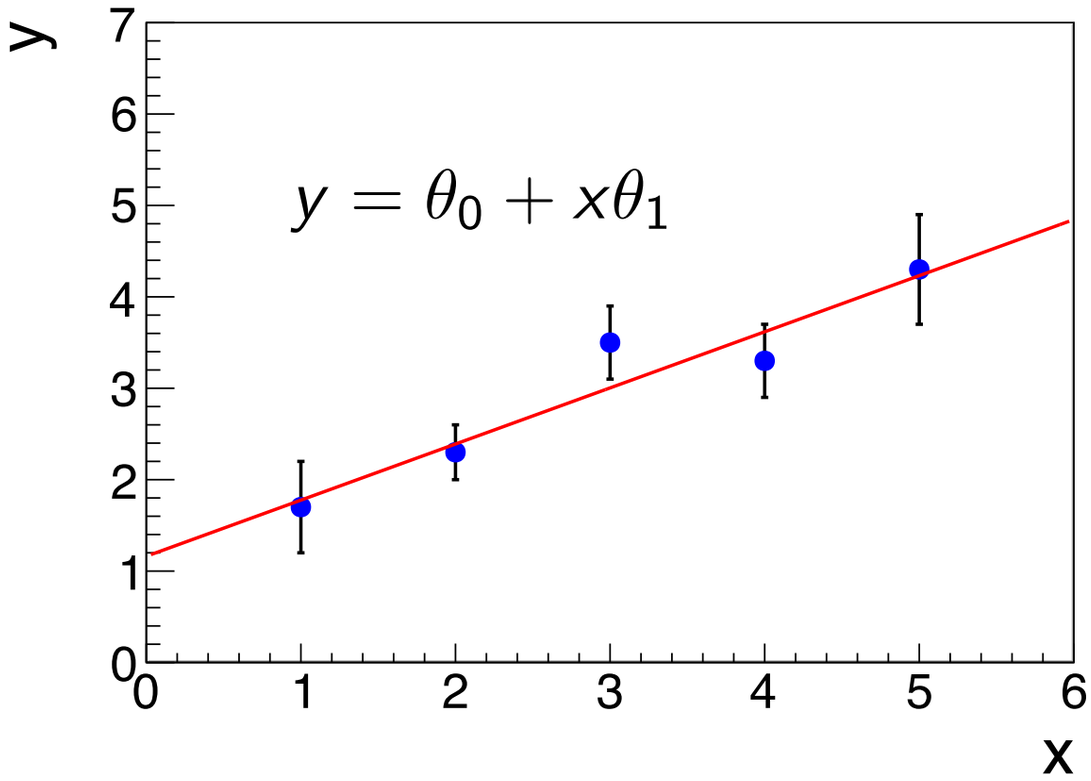
$$\begin{aligned} L &= (A^T V^{-1} A)^{-1} A^T V^{-1} \\ &= \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} [x^2] & -[x] \\ -[x] & [1] \end{pmatrix} \cdot \begin{pmatrix} 1/\sigma_1^2 & 1/\sigma_2^2 & \dots & 1/\sigma_n^2 \\ x_1/\sigma_1^2 & x_2/\sigma_2^2 & \dots & x_n/\sigma_n^2 \end{pmatrix} \\ &= \frac{1}{[1][x^2] - [x][x]} \begin{pmatrix} [x^2]\frac{1}{\sigma_1^2} - [x]\frac{x_1}{\sigma_1^2} & \dots & [x^2]\frac{1}{\sigma_n^2} - [x]\frac{x_n}{\sigma_n^2} \\ -[x]\frac{1}{\sigma_1^2} + [1]\frac{x_1}{\sigma_1^2} & \dots & -[x]\frac{1}{\sigma_n^2} + [1]\frac{x_n}{\sigma_n^2} \end{pmatrix} \end{aligned}$$

We finally obtain:

$$\boxed{\hat{\theta}_0 = \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x][x]} \qquad \hat{\theta}_1 = \frac{-[x][y] + [1][xy]}{[1][x^2] - [x][x]}}$$

$$[xy] := \sum_i \frac{x_i y_i}{\sigma_i^2}$$

Example: Straight Line Fit: $y = \theta_0 + \theta_1 \cdot x$ (III)



x	y	σ_y
1	1.7	0.5
2	2.3	0.3
3	3.5	0.4
4	3.3	0.4
5	4.3	0.6

Fit result:

$$[z] := \sum_i \frac{z}{\sigma_i^2}$$

$$\hat{\theta}_0 = \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x][x]} = 1.16207$$

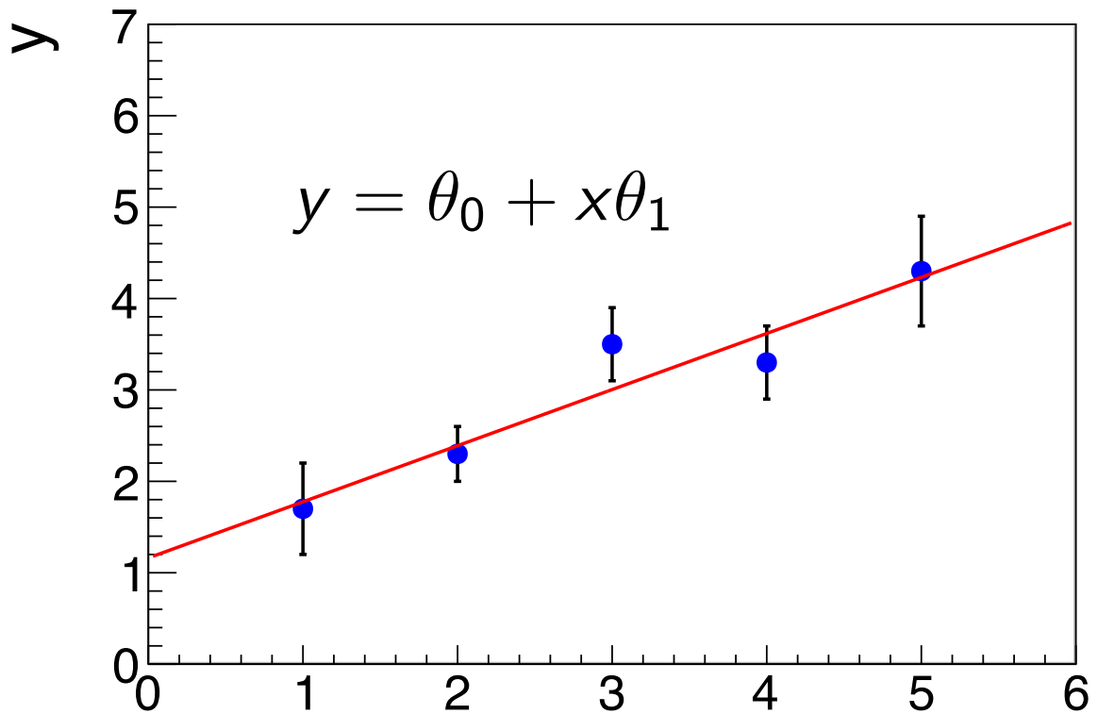
$$\hat{\theta}_1 = \frac{-[x][y] + [1][xy]}{[1][x^2] - [x][x]} = 0.613945$$

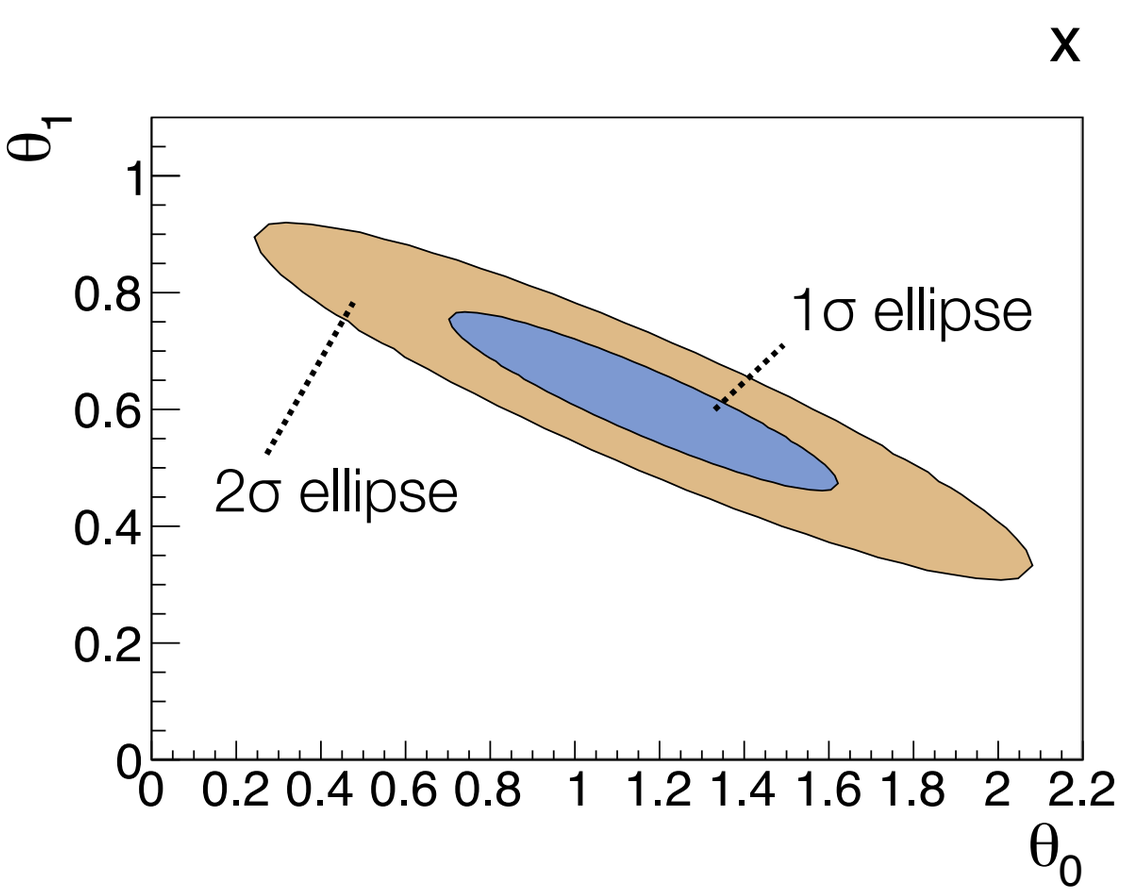
Covariance matrix of (θ_0, θ_1):

$$U = (A^T V^{-1} A)^{-1}$$

$$= \begin{pmatrix} 0.211186 & -0.0646035 \\ -0.0646035 & 0.0234105 \end{pmatrix}$$

Straight Line Fit: Comparison to MINIUT






```
// fit data points with linear function
TF1 *f = new TF1("f", "pol1", 0., 6.);
TFitResultPtr r = g->Fit("f", "F0qS", "", 0., 6.);
r->Print("V");
```

```
Minimizer is Minuit
Chi2                      =      2.29557
NDf                       =            3
Edm                       =  3.23988e-23
NCalls                    =           32
p0                        =      1.16207   +/-   0.45955
p1                        =     0.613945   +/-   0.153005

Covariance Matrix:
                               p0          p1
p0                        0.21119   -0.064603
p1                       -0.064603     0.02341

Correlation Matrix:
                               p0          p1
p0                             1    -0.91879
p1                       -0.91879           1
```

Propagation of Fit Parameter Uncertainties

$$y = \hat{\theta}_0 + \hat{\theta}_1 x \qquad \vec{J} = \begin{pmatrix} \frac{\partial y}{\partial \hat{\theta}_0} \\ \frac{\partial y}{\partial \hat{\theta}_1} \end{pmatrix} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

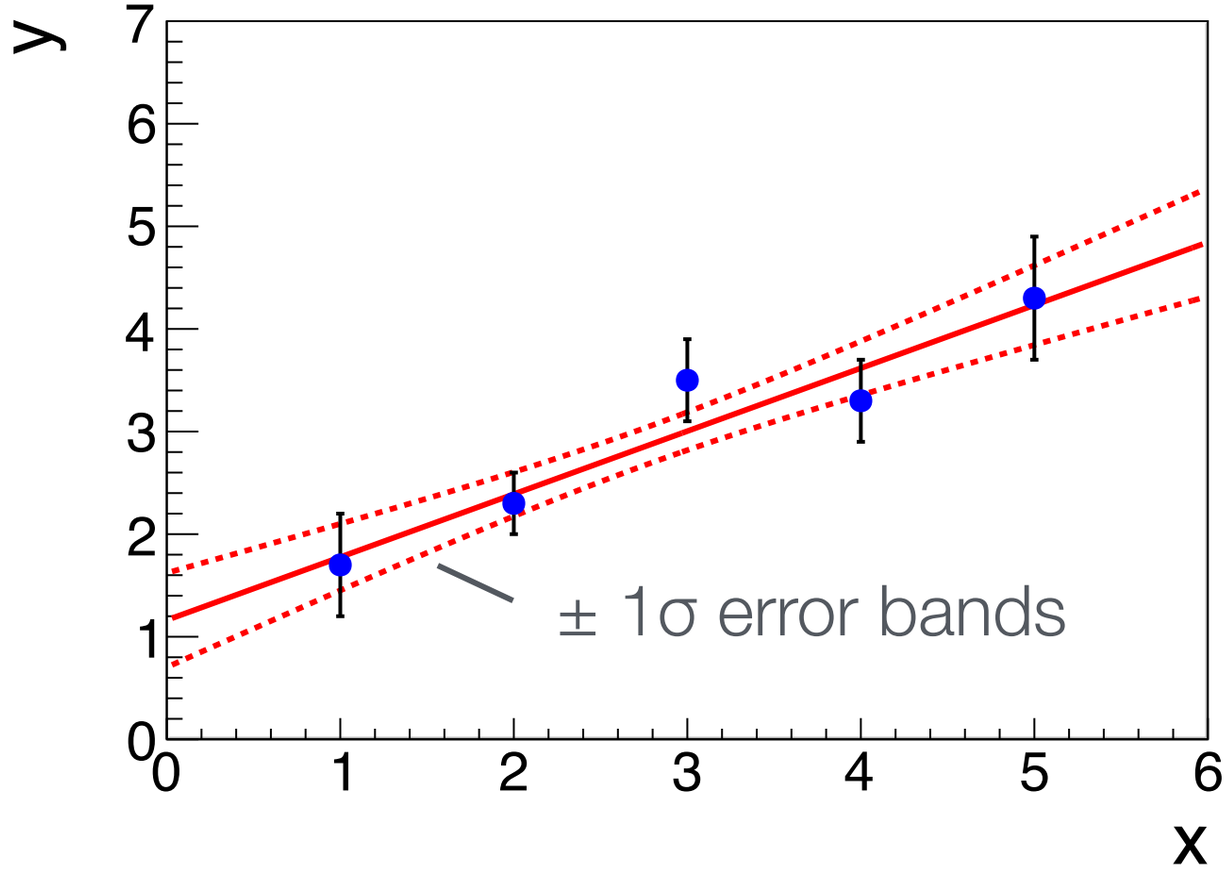
$$\sigma_y^2 = \vec{J}^{\mathsf{T}} U \vec{J} = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \sigma_0^2 & \mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] \\ \mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] & \sigma_1^2 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} \sigma_0^2 + x\,\mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] \\ \mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] + x\sigma_1^2 \end{pmatrix}$$

$$= \sigma_1^2 x^2 + 2\mathrm{cov}[\hat{\theta}_0, \hat{\theta}_1] x + \sigma_0^2$$

Note:
correlation vanishes if you choose

$$y = \theta_0 + \theta_1 (x - \langle x \rangle)$$



Least-Squares Fits to Histograms

Consider histogram with k bins and n_i counts in bin i. If n_i is not too small one can use the Gaussian approximation of the Poisson distribution and apply the least-squares method:

Pearson's χ^2:

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{k} \frac{(n_i - \nu_i(\vec{\theta}))^2}{\nu_i(\vec{\theta})}$$

Neyman's χ^2:

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{k} \frac{(n_i - \nu_i(\vec{\theta}))^2}{n_i}$$

Problems arise in bins with few entries (typically less than 5), in particular in Neyman's χ^2.

Bins with zero entries are problematic, typically omitted from the fit
→ leads to biased fit results

Summary: Maximum Likelihood and χ^2 Method

Maximum likelihood method:

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad \frac{\partial \ln L}{\partial \theta_i} = 0, \quad i = 1, \ldots, m \quad \rightsquigarrow \quad \widehat{\vec{\theta}}$$

$$U[\widehat{\vec{\theta}}] = -H^{-1}, \; h_{ij} = \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\vec{\theta}}}, \qquad H = (h_{ij}), \quad U = (u_{ij}), \quad u_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$$

covariance matrix of the estimated parameters θ_i

Least-squares method:

No correlations btw. the y_i:

$$\chi^2(\vec{\theta}) = -2 \ln L(\vec{\theta}) + \text{constant} = \sum_{i=1}^{n} \frac{(y_i - \mu(x_i; \vec{\theta}))^2}{\sigma_i^2}$$

With correlations btw. the y_i:

$$\chi^2(\vec{\theta}) = (\vec{y} - \vec{\mu}(\theta))^{\mathsf{T}} V^{-1} (\vec{y} - \vec{\mu}(\theta)), \qquad V = (v_{ij}), \quad v_{ij} = \mathrm{cov}[y_i, y_j]$$

covariance matrix of the θ_i

$$\frac{\partial \chi^2}{\partial \theta_i} = 0, \quad i = 1, \ldots, m \quad \rightsquigarrow \quad \widehat{\vec{\theta}} \qquad U[\widehat{\vec{\theta}}] = 2H^{-1}, \; h_{ij} = \left. \frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\vec{\theta}}}$$

Discussion of Fit Methods

[Wouter Verkerke, link]

Unbinned maximum likelihood fit (the best)

+ Don't need to bin data (no loss of information)
+ Works with multi-dimensional data
+ No Gaussian assumption
– No direct goodness of fit estimate
– Can be computationally expensive for large *n*
– Can't plot directly with data

Least-squares fit (the easiest)

+ fast, robust, easy
+ goodness of fit
+ can plot with data
+ works fine at high statistics
– data should be Gaussian
– misses information with feature size < bin size

Binned maximum likelihood fit in between

Hypothesis Tests and Goodness-of-Fit

Hypotheses and Tests

Hypothesis test

- Statement about the validity of a model
- Tells you which of two competing models is more consistent with the data

Simple hypothesis: a hypothesis with no free parameters

- Examples: the detected particle is a pion; data follow Poissonian with mean 5

Composite hypothesis: contains unspecified parameter(s)

- Example: data follow Poissonian with mean > 5

Null hypothesis H_0 and alternative hypothesis H_1

- H_0 often the *background-only hypothesis*
  (e.g. the Standard Model in searches for new physics)
- H_1 often *signal* or *signal + background hypothesis*

Question: Can null hypothesis be rejected by the data?

Tests statistic

Test statistic $t(\vec{x})$:
a (usually scalar) variable which is a function of the data alone that is used to test hypotheses

$\vec{x} = (x_1, \ldots, x_n)$: measured features/data

Examples:

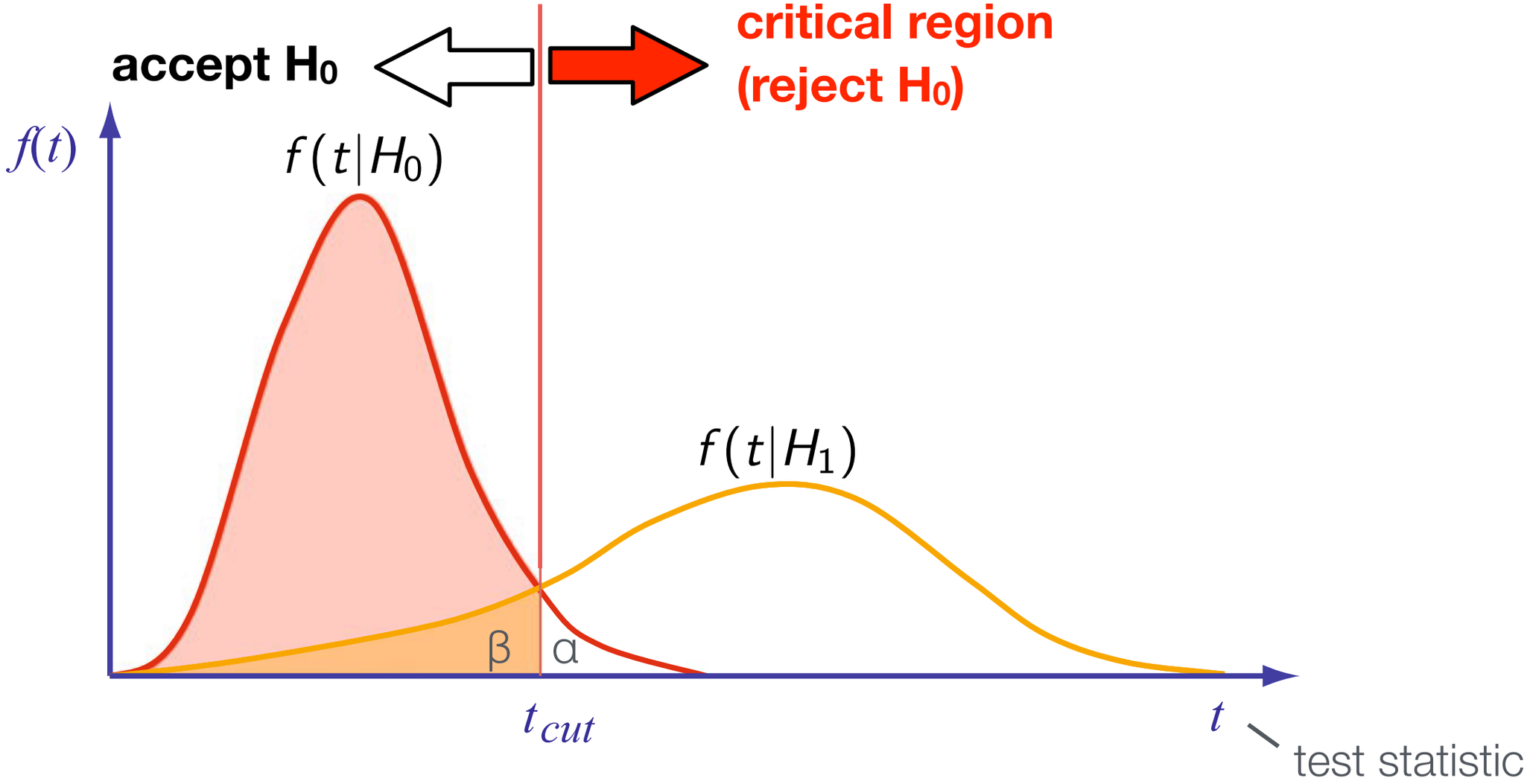
$t = X^2_{\text{min}}$ of a least-squares fit

ALICE TRD: likelihood ratio for electrons and pions: $t = \dfrac{L(q \mid e)}{L(q \mid \pi)}$

ALICE TPC dE/dx: $t = \dfrac{dE/dx - \langle dE/dx \rangle}{\sigma}$

Output of a boosted decision tree or neural network

Critical region



The probability for H_0 to be rejected while H_0 is true:

$$\int_{t_{\text{cut}}}^{\infty} f(t|H_0)\,\mathrm{d}t = \alpha$$

α:
"size" or "significance level" of the test

Probability to reject H_1 even though it is true:

$$\int_{-\infty}^{t_{\text{cut}}} f(t|H_1)\,\mathrm{d}t = \beta$$

1– β:
"power of the test"

Type I and Type II Errors

Type I error:
Null hypothesis is rejected while it is actually true

Type II error:
Test fails to reject null hypothesis while it is actually false

Type I and type II errors and their probabilities:

	H_0 is true	H_0 is false (i.e., H_1 is true)
H_0 is rejected	Type I error (α)	Correct decision ($1-\beta$)
H_0 is not rejected	Correct decision ($1-\alpha$)	Type II error (β)

Neyman–Pearson Lemma

Neyman-Pearson lemma holds for simple hypotheses and states:

To get the highest power (i.e. smallest possible value of β) of a test of H_0 with respect to the alternative H_1 for a given significance level, the critical region W should be chosen such that:

$$t(\vec{x}) := \frac{f(\vec{x}|H_1)}{f(\vec{x}|H_0)} > c \quad \text{inside } W \qquad \text{and} \qquad t(\vec{x}) \leq c \quad \text{outside } W$$

c is a constant chosen to give a test of the desired significance level.

Equivalent formulation: optimal scalar test statistic is the likelihood ratio

$$t(\vec{x}) = \frac{f(\vec{x}|H_1)}{f(\vec{x}|H_0)}$$

Practical Considerations

Problem: often one does not have explicit formulas for $f(x|H_0)$ and $f(x|H_1)$

One rather has Monte Carlo models for signal and background processes which allow one to generate instances of the data

In this case one can use multi-variate classifiers to separate different types of events

- Fisher discriminants
- Neural networks
- Support vector machines
- decision trees
- …

Least Squares Method: Goodness-of-Fit (I)

The minimum value of is a measure of the level of agreement between the model and the data;

$$\chi^2_{\text{min}} = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \hat{\vec{\theta}})}{\sigma_i} \right)^2$$

Large χ^2_{min}: the model can can be rejected.

If the model is correct, then χ^2_{min} for repeated experiments follows a χ^2 distribution:

$$f(t; n_{\text{df}}) = \frac{1}{2^{n_{\text{df}}/2}\Gamma\left(\frac{n_{\text{df}}}{2}\right)} t^{n_{\text{df}}/2-1} e^{-t/2}, \qquad t = \chi^2_{\text{min}}$$

with $n_{\text{df}} = n - m = \text{number of data points} - \text{number of fit parameters}$

n_{df} = "number of degrees of freedom"

Least Squares Method: Goodness-of-Fit (II)

Expectation value of the χ^2 distribution is n_{df}
→ $\chi^2 \approx n_{\text{df}}$ indicates a good fit

Consistency of a model with the data is quantified with the *p*-value:

$$p\text{-value} = \int_{\chi^2_{\text{min}}}^{\infty} f(t; n_{\text{df}})\, \mathrm{d}t$$

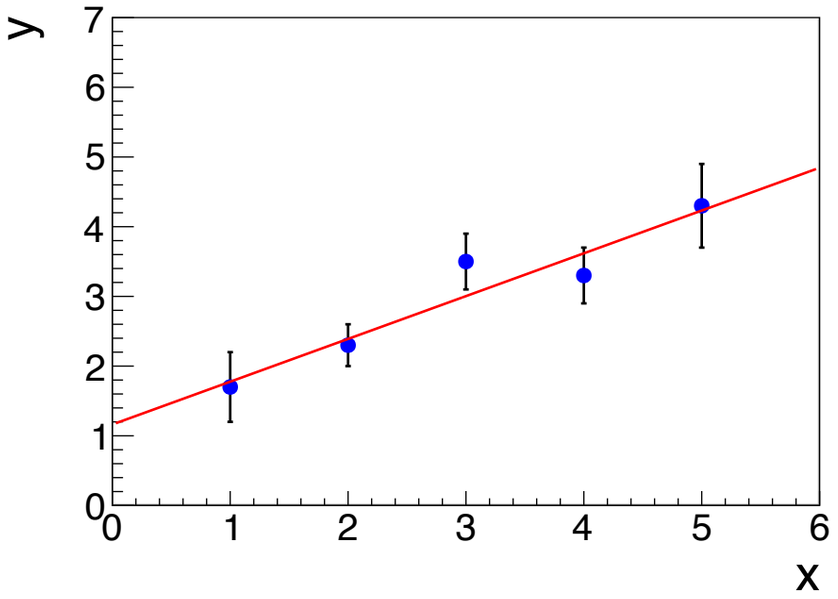
The *p*-value is the probability to get a χ^2_{min} as high as the observed one, or higher, if the model is correct.

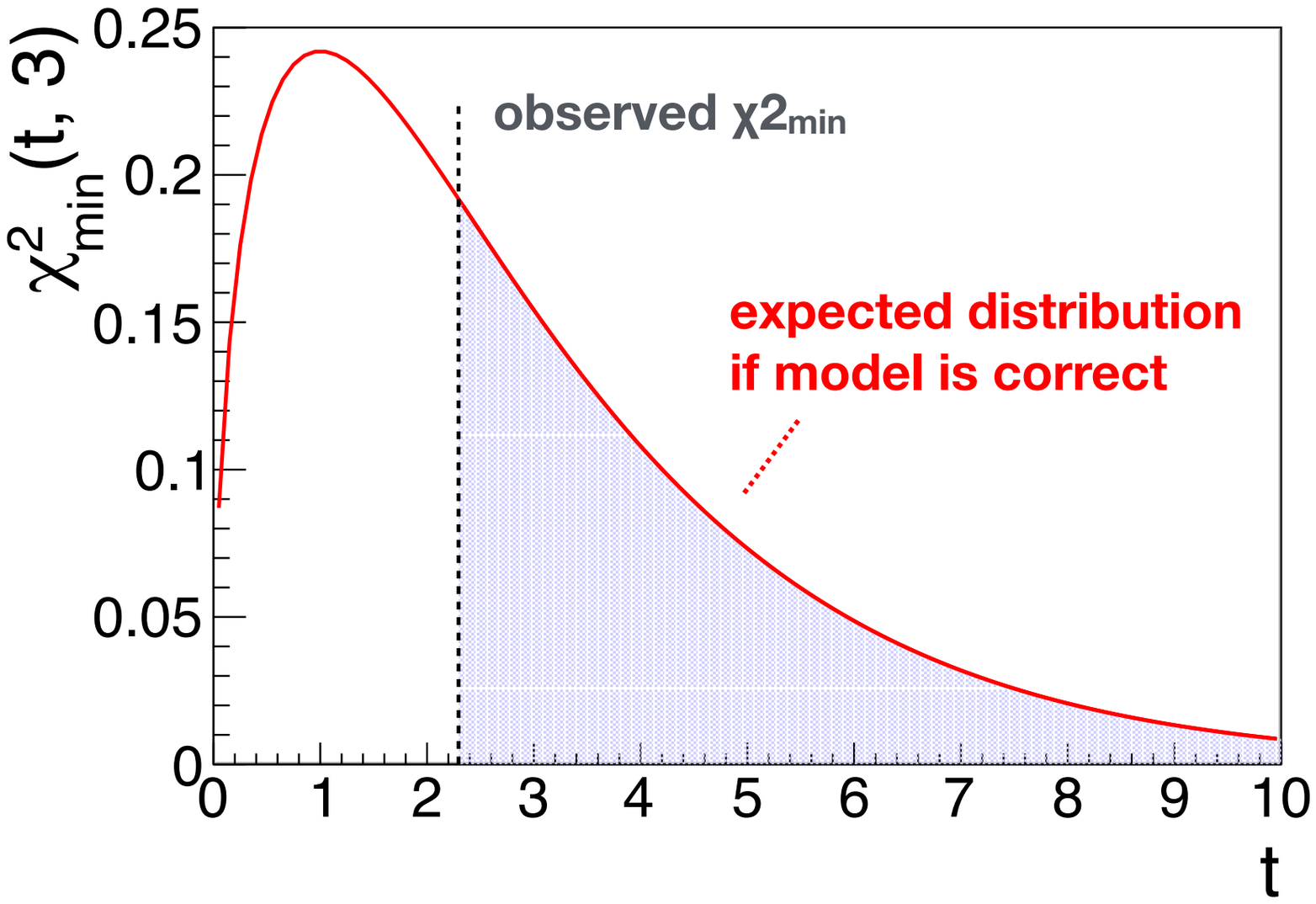
The *p*-value is **not** the probability that the model is correct.

p-value for the Straight Line Fit Example

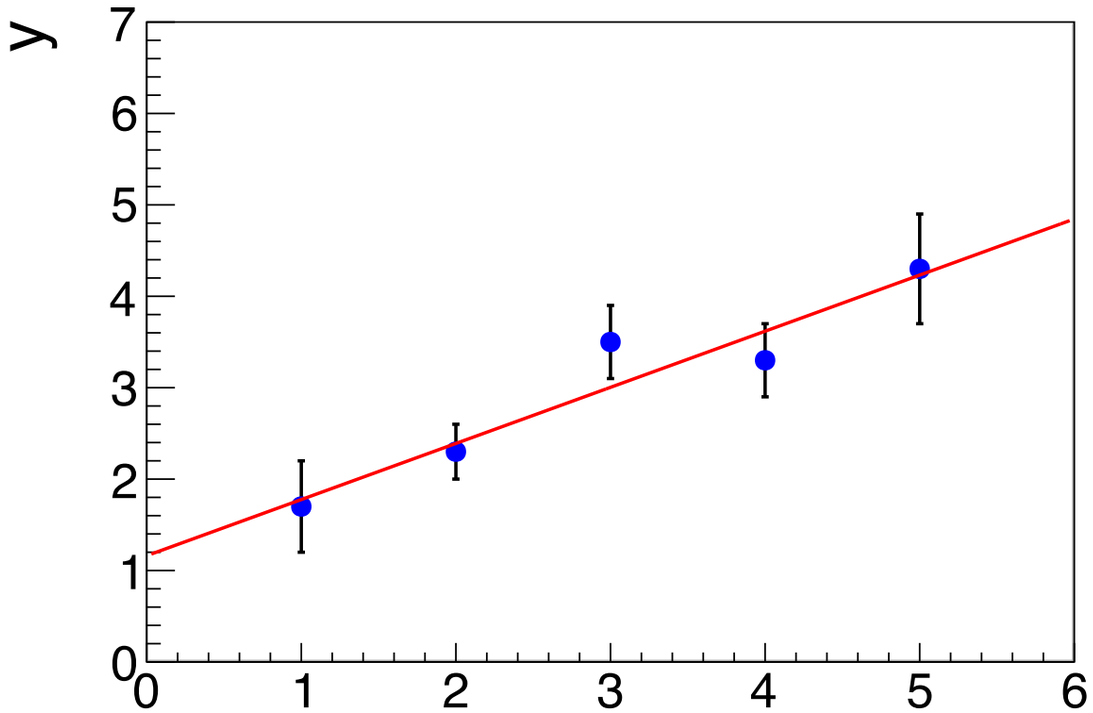
$\chi^2_{min} = 2.29557$, $n_{df} = 3$:

p-value = 0.51337





Constant Model ($y = \theta_0$) Rejected by Small p-value

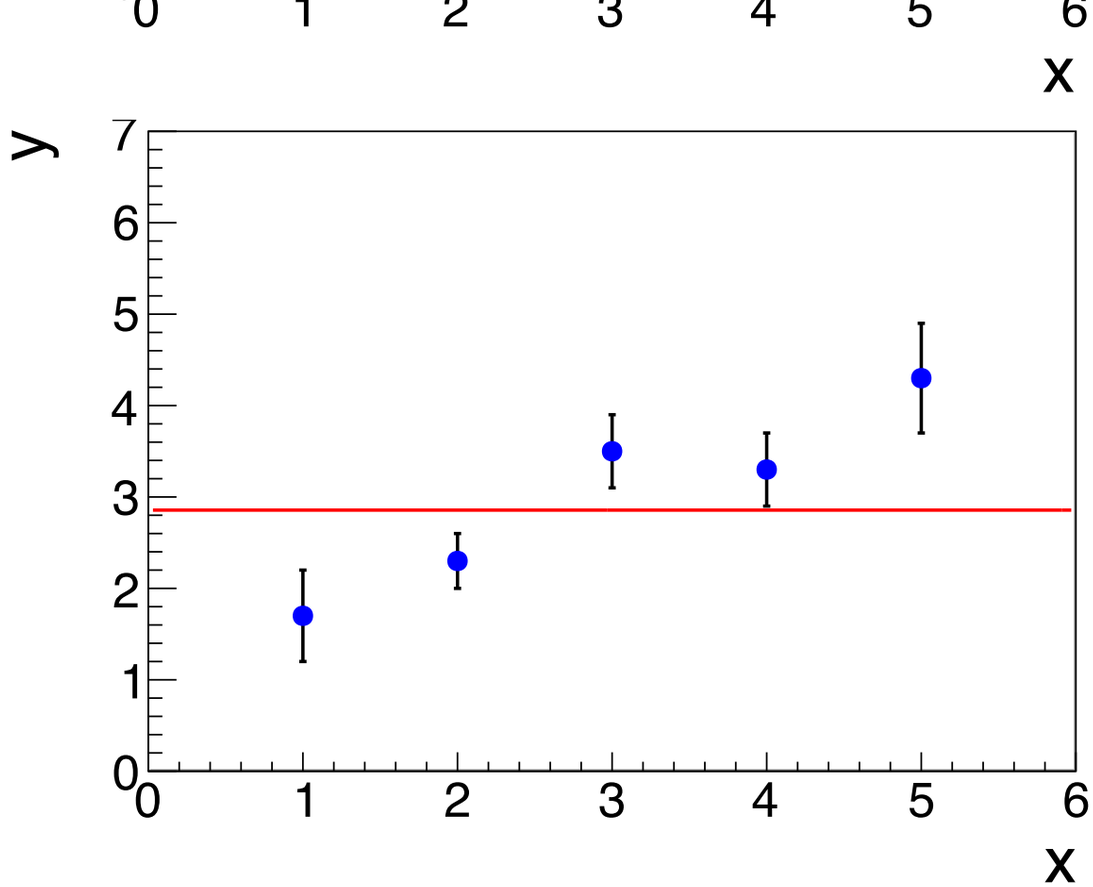


$\chi^2_{min} = 2.29557$, $n_{df} = 3$:

p-value = 0.51337

```
from scipy import stats
pvalue = 1 - stats.chi2.cdf(chi2, n_dof)
```

```
root [1] TMath::Prob(chi2, n_dof)
```



$\chi^2_{min} = 18.3964$, $n_{df} = 4$:

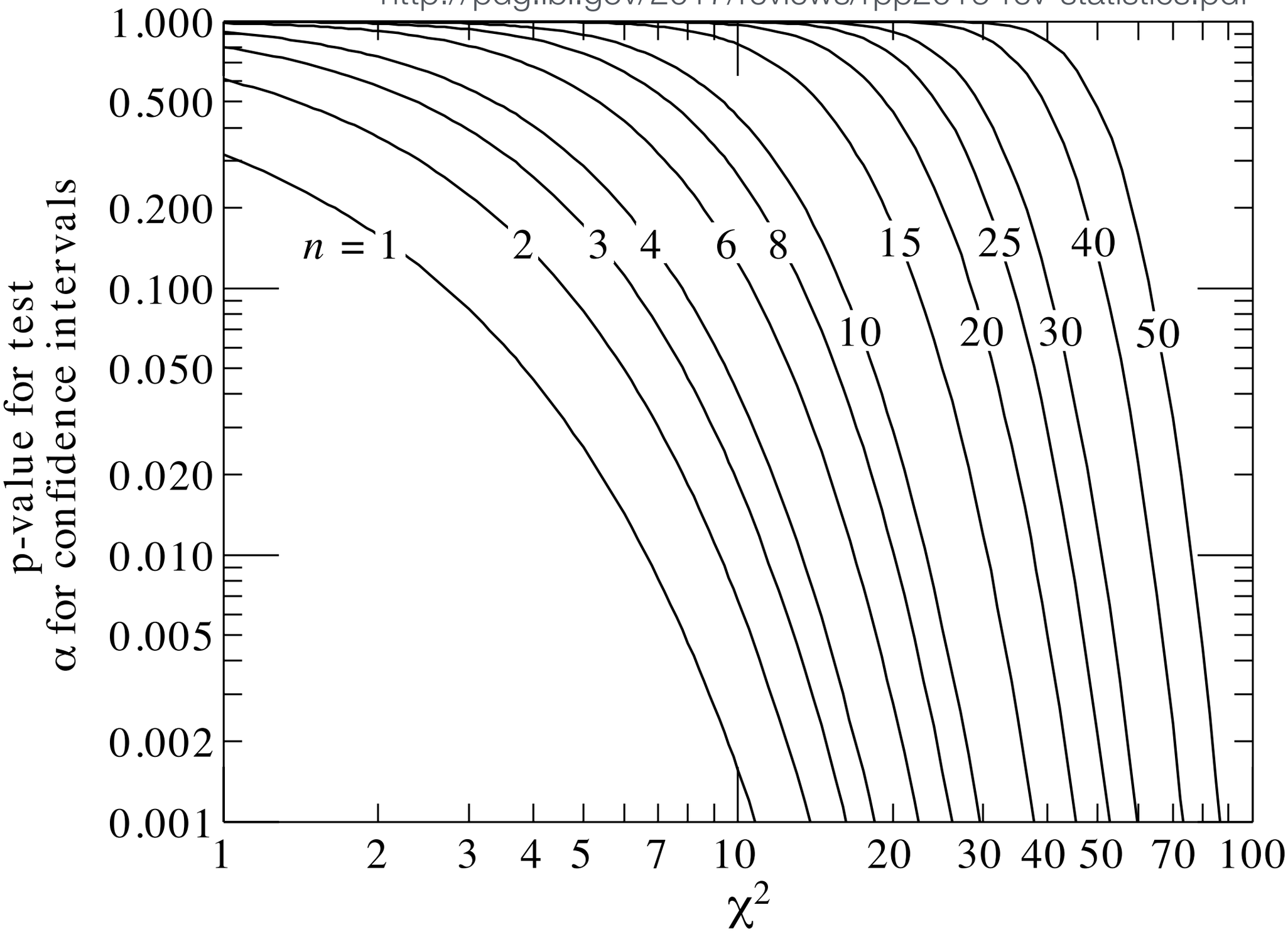
p-value = 0.001032

$\theta_0 = 2.86 \pm 0.18$

Statistical uncertainty of the fit parameter does not tell us whether model is correct!

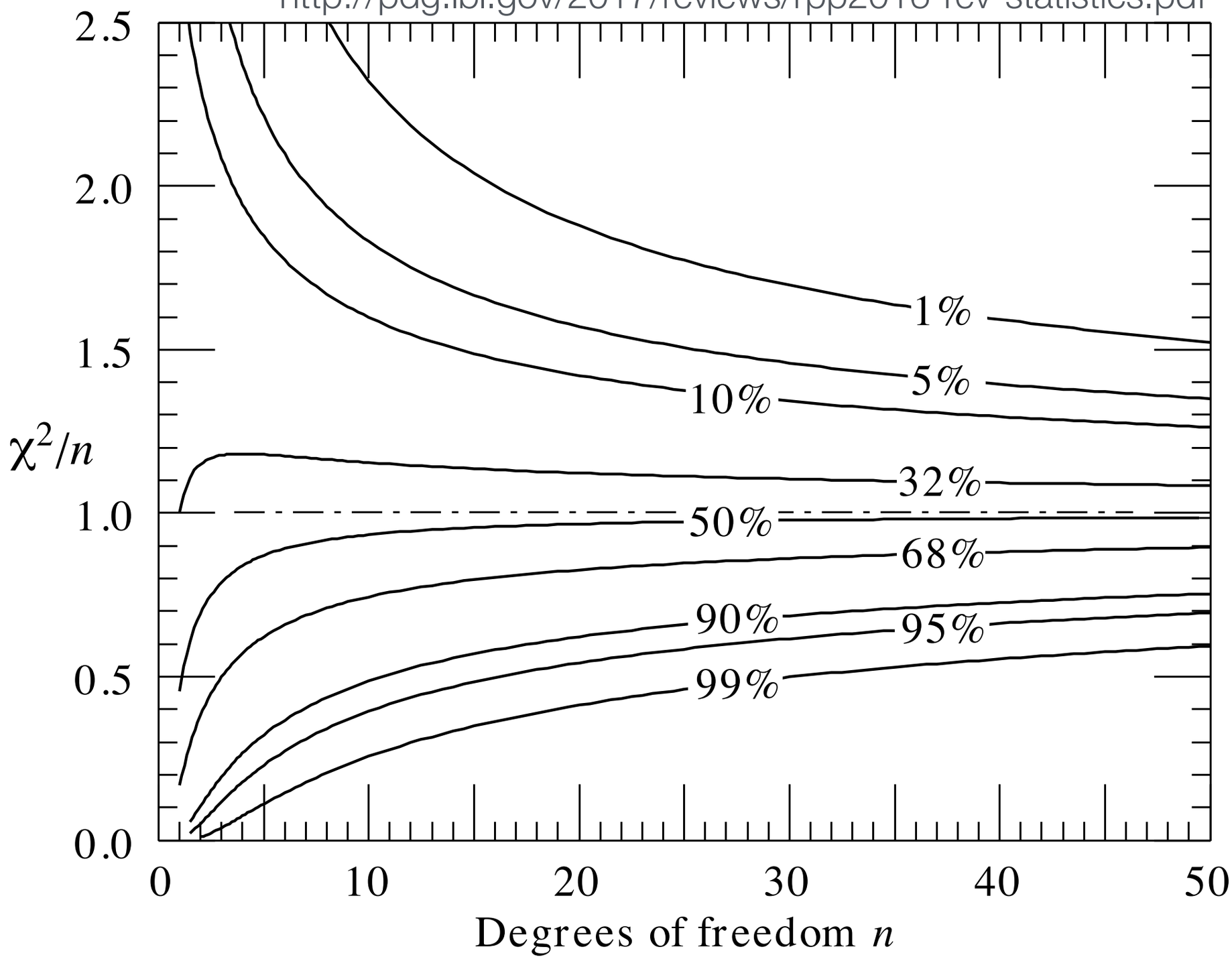
p-value for different χ^2_{min} and n_{df}

http://pdg.lbl.gov/2017/reviews/rpp2016-rev-statistics.pdf



Confidence Intervalls for χ^2_{min} / n_{df} as a fct. of n_{df}

http://pdg.lbl.gov/2017/reviews/rpp2016-rev-statistics.pdf



Goodness-of-Fit for Unbinned ML Fits (I)

In case of an unbinned ML fit one can put data and model prediction into a histogram and perform a χ^2 test.

Consider the ratio

L: likelihood

$$\lambda = \frac{L(\vec{n}|\vec{\nu})}{L(\vec{n}|\vec{n})}, \qquad \vec{\nu} = \vec{\nu}(\vec{\theta}), \quad \vec{\theta} = (\theta_1, ..., \theta_m)$$

For the multinomial ("M", n_{tot} fixed) and Poisson distributed data ("P") one obtains

k: number of bins of the histogram

$$\lambda_{\text{M}} = \prod_{i=1}^{k} \left(\frac{\nu_i}{n_i}\right)^{n_i}, \qquad \lambda_{\text{P}} = e^{n_{\text{tot}} - \nu_{\text{tot}}} \prod_{i=1}^{k} \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

We then consider

$$\chi^2 := -2 \ln \lambda$$

Goodness-of-Fit for Unbinned ML Fits (II)

For multinomially distributed data in the large sample limit

$$\chi^2_{\mathrm{M}} := -2 \ln \lambda_{\mathrm{M}} = 2 \sum_{i=1}^{k} n_i \ln \frac{n_i}{\hat{\nu}_i}$$

follows a χ^2 distribution for $k - m - 1$ degrees of freedom if the model is correct.

In case of Poisson distributed data

$$\chi^2_{\mathrm{P}} := -2 \ln \lambda_{\mathrm{P}} = 2 \sum_{i=1}^{k} \left(n_i \ln \frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i \right)$$

follows a χ^2 distribution for $k - m$ degrees of freedom in the large sample limit if the model is correct.

Wilks' theorem

Let null hypothesis H_0 be a special case of the hypothesis H_1 ("nested hypotheses")

Example:

$$H_0 : \; f(m) = a_0 + a_1 m$$
$$H_1 : \; f(m) = a_0 + a_1 m + a_2 m^2 + a_3 m^3$$

Define:

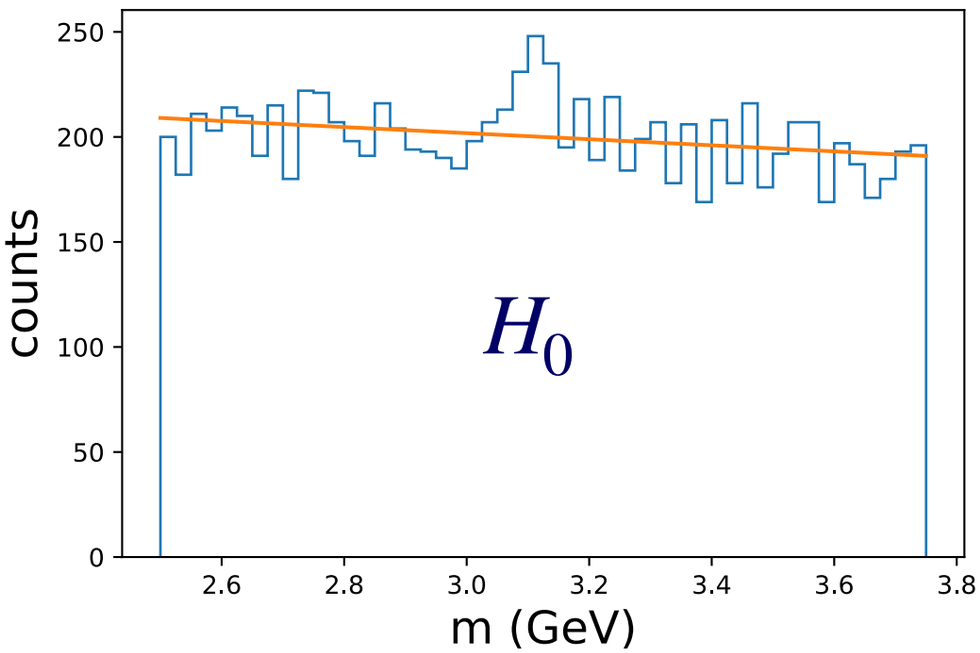
$$\Delta\tilde{\chi}^2 := -2 \ln \left(\frac{L(H_1)}{L(H_0)} \right)$$

Wilks' theorem:
If H_0 is correct then $-\Delta\tilde{\chi}^2$ follows χ^2 distribution with n_{dof} = #added parameters in the large sample limit.

In the above example: $n_{\mathrm{dof}} = 2$

Samuel S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses Ann. Math. Statist., Volume 9, Number 1 (1938), 60-62.

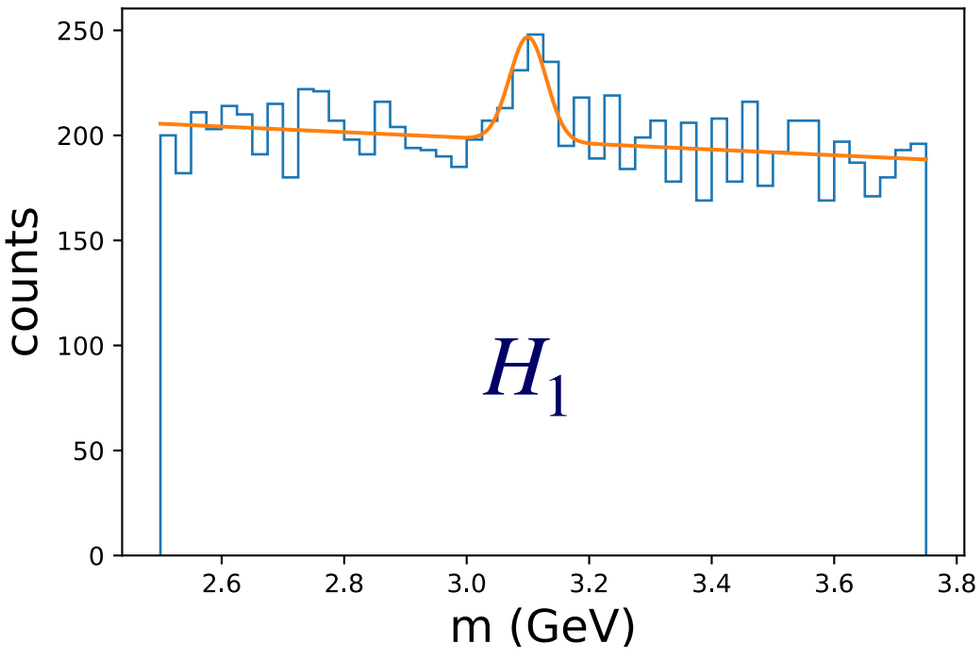
Significance of a peak



$$H_0: \; f(m) = a_0 + a_1 m$$
$$H_1: \; f(m) = a_0 + a_1 m + a_2 N(m; \mu, \sigma)$$

$\mu = 3.1, \sigma = 0.03$ fixed in H_1
→ one additional parameter

$$\Delta\tilde{\chi}^2 := -2 \ln \left(\frac{L(H_1)}{L(H_0)} \right) = -22.5$$



$-\Delta\tilde{\chi}^2$ should follow a χ^2 distribution with $n_{\mathrm{dof}} = 1$ if H_0 ist true

p-value = $2.15 \cdot 10^{-6}$

→ H_0 can be safely rejected

Kolmogorov–Smirnov Test (I)

KS test is an unbinned goodness-of-fit test

Q: Do data points come from a given distribution?

Compare cumulative distribution function

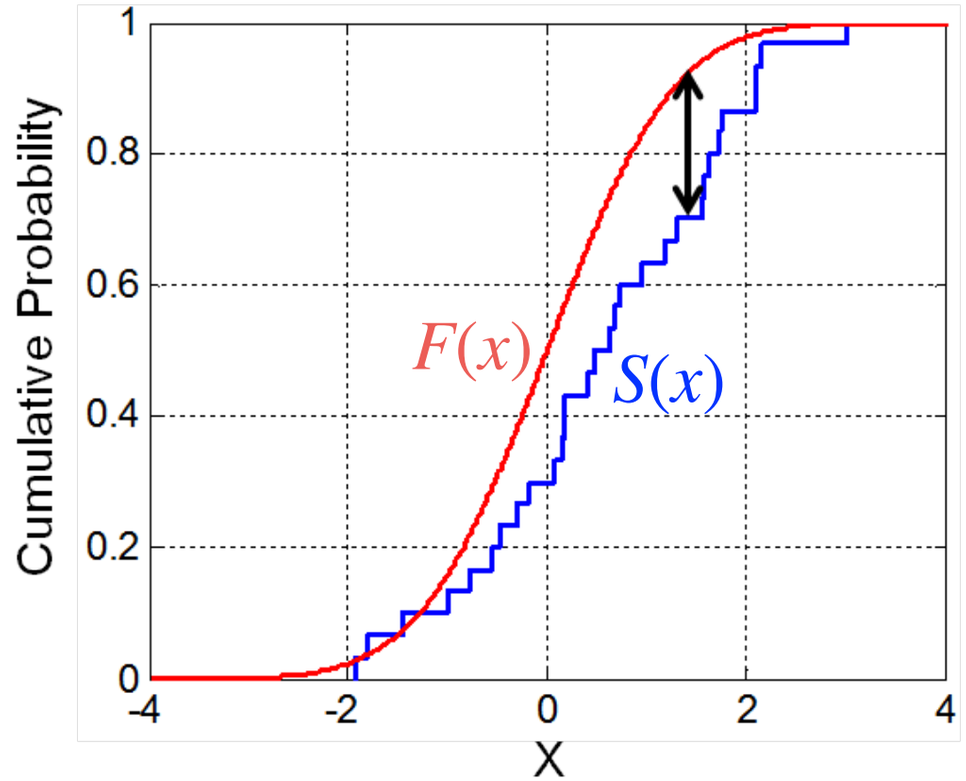
$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'$$

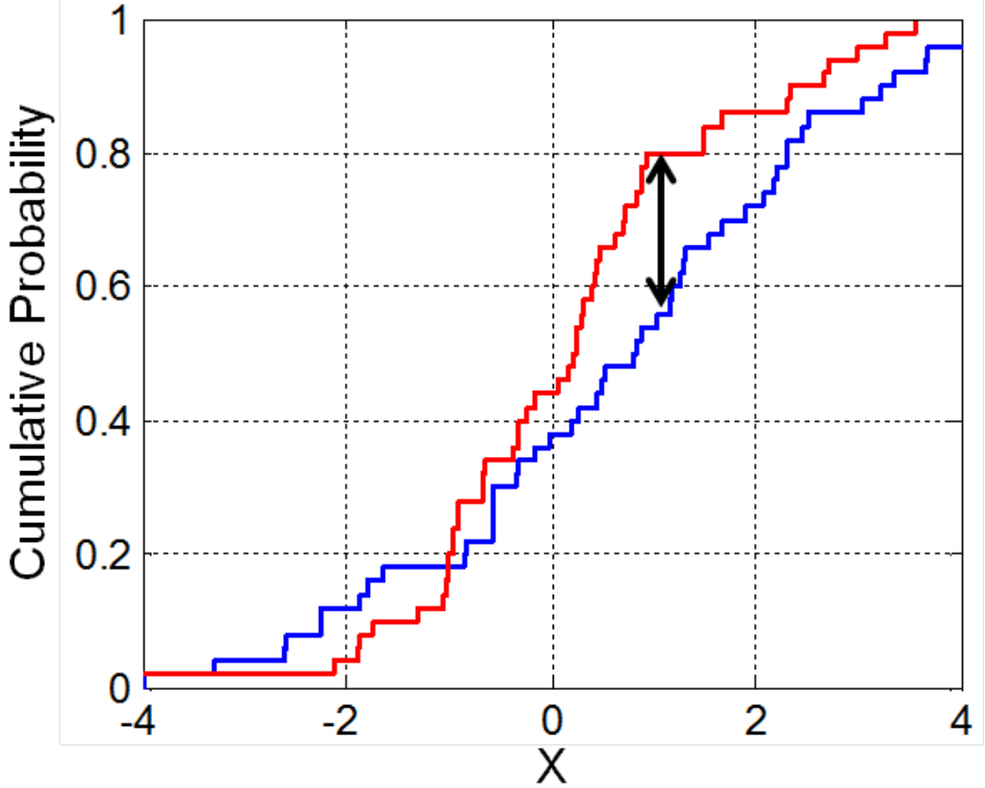
with the so-called Empirical Distribution Function (EDF)

$$S(x) = \frac{\text{number of observations with } x_i < x}{\text{total number of observations}}$$

The test statistic is the maximum difference between the two functions:

$$D = \sup|F(x) - S(x)|$$



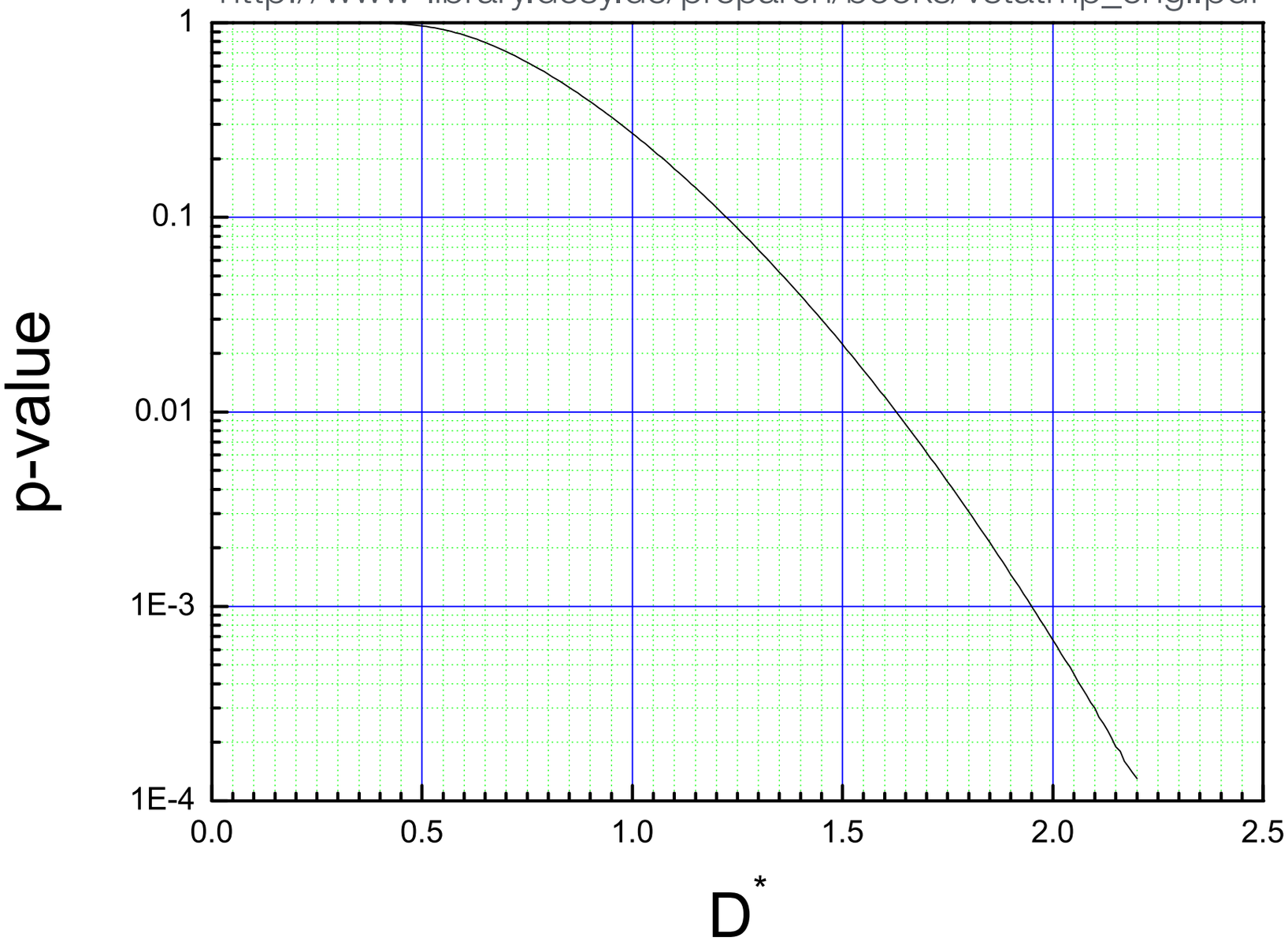


One can also test whether two one-dimensional sets of points are compatible with coming from the same parent distribution.

Kolmogorov–Smirnov Test (II)

Expected distribution of D known for given N → p-value

Bohm, Zech,
http://www-library.desy.de/preparch/books/vstatmp_engl.pdf



$D^* = \sqrt{N} D,$

$N =$ number of data points

Example:
Test whether data x_i come from standard normal distribution N(0,1):

```
from scipy import stats
D, p_value =
stats.kstest(x, stats.norm.cdf)
```

Kolmogorov–Smirnov test: only for 1d data

Bayesian Hypothesis Testing

In Bayesian language, all problems are hypothesis tests!

- Posterior probability for a hypothesis $P(H|\text{data})$ or a parameter $P(\theta|\text{data})$

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

- Parameter estimation amounts to assigning a probability to the proposition that the parameter lies in the interval $[\theta_1, \theta_2]$
  - can reject hypothesis/parameter if posterior prob. is sufficiently small
- Example: LIGO PRL on detection of gravitational waves

In the source frame, the initial black hole masses are $36^{+5}_{-4}M_\odot$ and $29^{+4}_{-4}M_\odot$, and the final black hole mass is $62^{+4}_{-4}M_\odot$, with $3.0^{+0.5}_{-0.5}M_\odot c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

DOI: 10.1103/PhysRevLett.116.061102

- Requires one to explicitly specify alternative hypotheses:

$$P(D) = P(D|H_1) + P(D|H_2) + P(D|H_3) + \ldots$$

Often simply normalization from $\int P(H|D) = 1$

Systematic Uncertainties

Statistical and Systematic Uncertainties

$$x = 2.34 \pm 0.05\,(\text{stat.}) \pm 0.03\,(\text{syst.})$$

quoting stat. and syst. uncertainty separately gives us an idea whether taking more data would be helpful

Statistical or random uncertainties

- Uncertainties that can be reliably estimated by repeating measurements
- They follow a known distribution like a Poisson rate or are determined empirically from the distribution of an unbiased, sufficiently large sample.
- Relative uncertainty reduces as $1/\sqrt{N}$ where N is the sample size

Systematic uncertainties

- Cannot be calculated solely from sampling fluctuations
- In most cases don't reduce as $1/\sqrt{N}$ (but often also become smaller with larger N)
- Difficult to determine, in general less well known than the statistical uncertainty
- Systematic uncertainties ≠ mistakes
  (a bug in your computer code is not a systematic uncertainty)

Systematic Uncertainties: Examples

Calibration uncertainties of the measurement apparatus

- E.g., energy scale uncertainty of a calorimeter

Uncertainty of the detector resolution

Detector acceptance

Limited knowledge about background processes

Uncertainties of auxiliary quantities

- E.g. reference branching ratios uses as input
- Uncertainty of theoretical quantities

...

The uncertainty in the estimation of such a systematic effect is called a systematic uncertainty.

Systematic Uncertainties ≠ Mistakes, but mistakes still happen

R. Barlow
"Systematic Errors, Fact and Fiction," hep-ex/0207026

Look for mistakes be repeating the analysis with changes which *should* make no difference:

Data subsets

Magnet up/down

Different selection cuts

Different histogram bin sizes and fit ranges

Different Event Generator for efficiency calculation

Look for impossibilities

If a check passes the test:
move on and do not add the discrepancy to the systematic uncertainty

If a check fails: try to identify the reason. Only as very last resort, add contribution to total systematic uncertainty. This might underestimate the real uncertainty.

Handling discrete systematic uncertainties

Typical case: choice of model

With 1 preferred model and one other, quote $R_1 \pm |R_1 - R_2|$

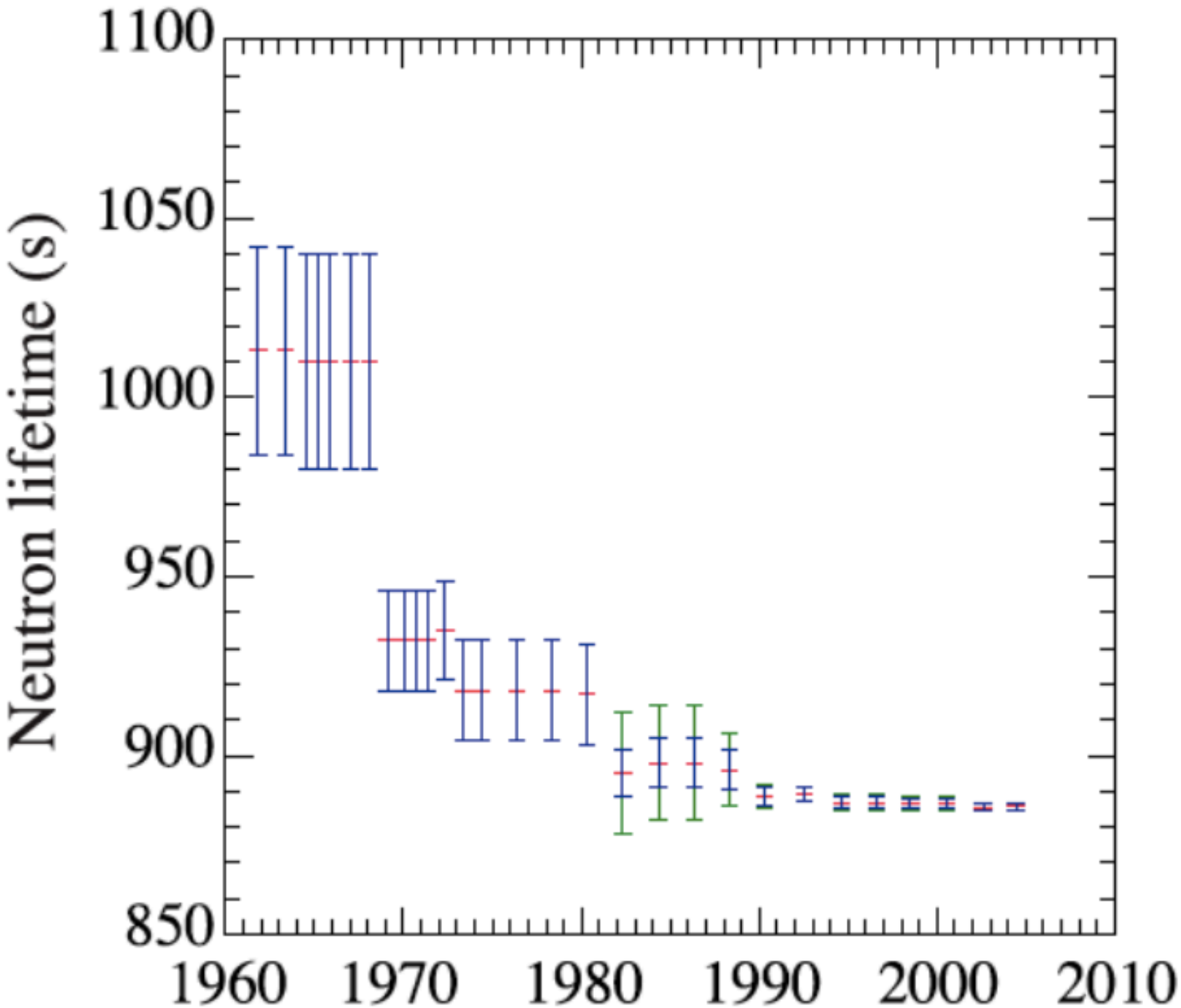
With 2 models of equal status, quote $\dfrac{R_1 + R_2}{2} \pm \dfrac{|R_1 - R_2|}{\sqrt{2}}$

n equal models, quote $\bar{R} \pm \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(R_i - \overline{R})^2} = \sqrt{\dfrac{n}{n-1}(\overline{R^2} - \overline{R}^2)}$

Two extreme model, quote $\dfrac{R_1 + R_2}{2} \pm \dfrac{|R_1 - R_2|}{\sqrt{12}}$

Experimenter's Bias?

Klein JR, Roodman, A. 2005,
Annu. Rev. Nucl. Part. Sci. 55:141–63



Do researchers unconsciously work toward a certain value?

Blind Analyses

Klein JR, Roodman, A. 2005,
Annu. Rev. Nucl. Part. Sci. 55:141–63

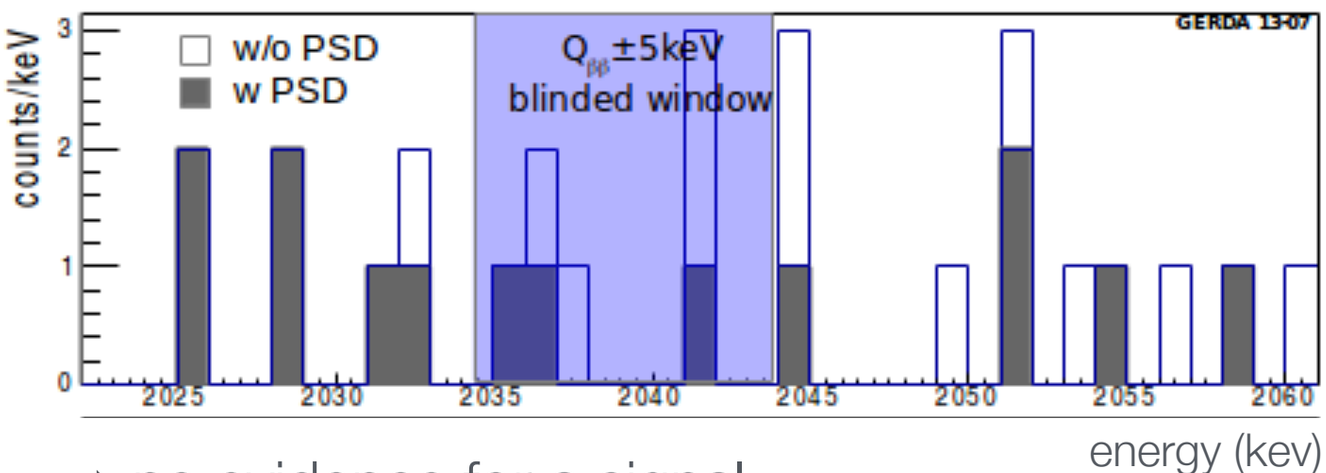
Avoid experimenter's bias by hiding certain aspects of the data.

Things that can be hidden in the analysis:

- The signal events, when the signal occurs in a well-defined region of the experiment's phase space.
- The result, when the numerical answer can be separated from all other aspects of the analysis.
- The number of events in the data set, when the answer relies directly upon their count.
- A fraction of the entire data set.

Example: GERDA experiment

- search for neutrinoless double beta decay
- Signal: sharp peak
- Background model fixed prior to unblinding of signal region



→ no evidence for a signal

Combination of Systematic Uncertainties

In most cases one tries to find independent sources of systematic uncertainties. These independent uncertainties are therefore added in quadrature:

$$\sigma_{\text{tot}}^2 = \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2$$

Often a few source dominate the systematic uncertainty

→ No need to work to hard on correctly estimating the small uncertainties

Systematic Uncertainties: Covariance Matrix Approach (I)

Consider two measurement x_1 and x_2 with with individual random uncertainties $\sigma_{1,\mathrm{r}}$ and $\sigma_{2,\mathrm{r}}$ and a common systematic uncertainty σ_s:

$$x_i = x_\mathrm{true} + \Delta x_{i,\mathrm{r}} + \Delta x_\mathrm{s}$$

$$\langle \Delta x_{i,\mathrm{r}} \rangle = 0, \quad \langle \Delta x_\mathrm{s} \rangle = 0,$$
$$\langle (\Delta x_{i,\mathrm{r}})^2 \rangle = \sigma_{i,\mathrm{r}}^2, \quad \langle (\Delta x_\mathrm{s})^2 \rangle = \sigma_s^2$$

Variance:

$$\begin{aligned} V[x_i] &= \langle x_i^2 \rangle - \langle x_i \rangle^2 \\ &= \langle (x_\mathrm{true} + \Delta x_{i,\mathrm{r}} + \Delta x_\mathrm{s})^2 \rangle - \langle x_\mathrm{true} + \Delta x_{i,\mathrm{r}} + \Delta x_\mathrm{s} \rangle^2 \\ &= \langle (\Delta x_{i,\mathrm{r}} + \Delta x_\mathrm{s})^2 \rangle \\ &= \sigma_{i,r}^2 + \sigma_s^2 \end{aligned}$$

Covariance:

$$\begin{aligned} \mathrm{cov}[x_1, x_2] &= \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle \\ &= \ldots \\ &= \sigma_s^2 \end{aligned}$$

Systematic Uncertainties: Covariance Matrix Approach (II)

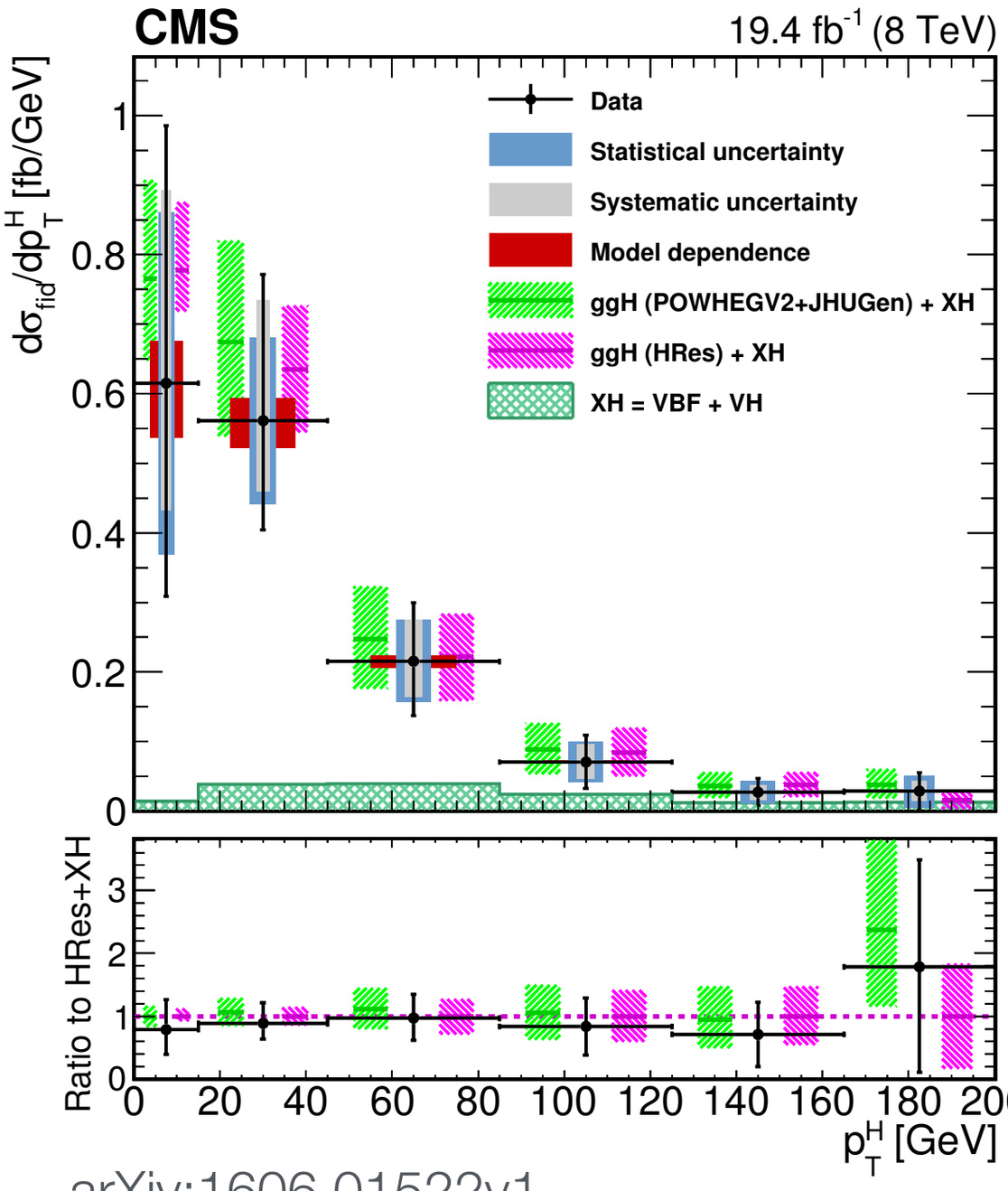
Covariance matrix for x_1 and x_2:

$$V = \begin{pmatrix} \sigma_{1,r}^2 + \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_{2,r}^2 + \sigma_s^2 \end{pmatrix}$$

This also works when the uncertainties are quoted as relative uncertainties:

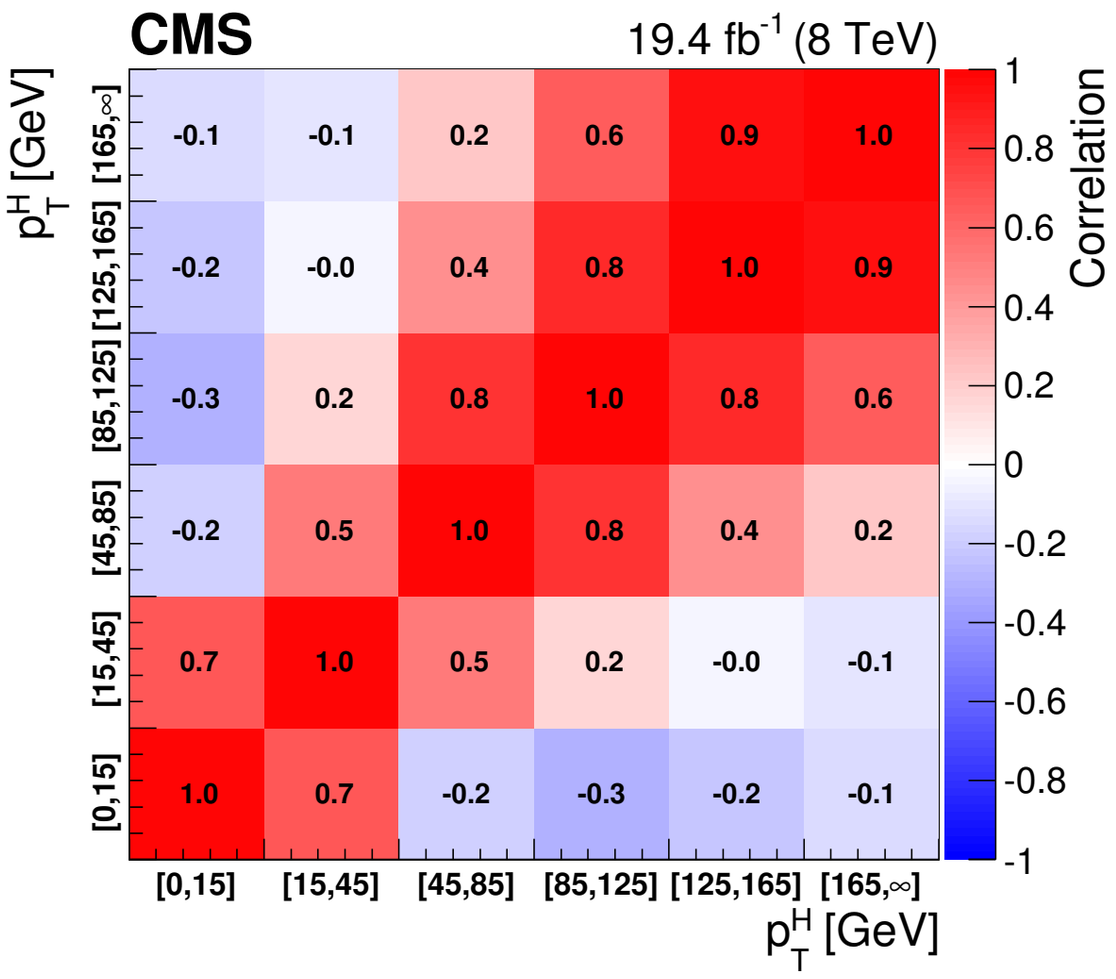
$$\sigma_s = \varepsilon x \qquad \rightsquigarrow \qquad V = \begin{pmatrix} \sigma_{1,r}^2 + \varepsilon^2 x_1^2 & \varepsilon^2 x_1 x_2 \\ \varepsilon^2 x_1 x_2 & \sigma_{2,r}^2 + \varepsilon^2 x_1^2 \end{pmatrix}$$

Example:
Transverse Momentum Spectrum of the Higgs-Boson



arXiv:1606.01522v1

Correlation matrix of the p_T bins:



$$\rho_{i,j} = \frac{V_{i,j}}{\sigma_i \sigma_j}, \qquad V = \text{covariance matrix}$$

Weighted Average of Correlated Data Points

Consider n data points y_i with covariance matrix V: $\vec{y} = (y_1, y_2, \ldots, y_n)$

One can calculate a weighted average λ by minimizing

$$\chi^2(\lambda) = (\vec{y} - \vec{\lambda})^\mathsf{T} V^{-1} (\vec{y} - \vec{\lambda})$$

$\vec{\lambda} := (\lambda, \lambda, \ldots, \lambda)$

One obtains (here without calculation):

$$\hat{\lambda} = \sum_{i=1}^{N} w_i y_i \qquad w_i = \frac{\sum_{j=1}^{n} (V^{-1})_{i,j}}{\sum_{k,l=1}^{n} (V^{-1})_{k,l}}$$

Variance results from error propagation:

$$\sigma^2_{\hat{\lambda}} = \vec{w}^\mathsf{T} V \vec{w} = \sum_{i,j=1}^{n} w_i V_{ij} w_j$$

Minimizing the χ^2 gives the *best linear unbiased estimate* (BLUE) → linear unbiased estimator with the lowest variance

- BLUE combination may be biased if uncertainties not known or are estimated from measured values
- Improvement: iterative approach (rescaling uncertainties based on previous iteration)

Special Case:
Weighted Average of Two Correlated Measurements

Consider two measurements with covariance matrix V (ρ = correlation coeff.):

$$y_1,\ y_2 \qquad V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Applying the formulas from the previous slide:

$$V^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \qquad \hat{\lambda} = wy_1 + (1-w)y_2$$

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \sigma^2 = \frac{(1-\rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

equivalently:

$$\frac{1}{\sigma^2} = \frac{1}{1-\rho^2}\left[\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} - \frac{2\rho}{\sigma_1\sigma_2}\right]$$

Another Approach To Least Squares Fits in Case of Correlated Systematic Uncertainties

Correlated systematic uncertainties can be taken into account with generalized χ^2:

$$\chi^2(\vec{\theta}) = (\vec{y} - \vec{f}(\vec{x};\vec{\theta}))^T V^{-1} (\vec{y} - \vec{f}(\vec{x};\vec{\theta})), \qquad V = \underbrace{V_{\text{stat}}}_{\text{diagonal}} + V_{\text{sys}}$$

Another approach (sometime called 'pull method'):

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i + \varepsilon\sigma_{i,\text{sys}} - f(x_i;\vec{\theta}))^2}{\sigma^2_{i,\text{stat}}} + \varepsilon^2$$

penalty term
("ε = systematic deviation in units of the standard deviation")

The pull method puts nuisance parameters on the same footing as other parameters. The penalty term is none other than a frequentist version of the Bayesian prior on the nuisance parameter.

Bayesian approach to systematic uncertainties

"Bayesians lose no sleep over systematics" (lecture S. Oser)

Quantity of interest: θ, prior knowledge: $\pi(\theta)$

Likelihood depends parameter ν ("nuisance parameter")

We simply treat θ and ν as unknown parameters:

$$P(\theta, \nu|\text{data}) \propto L(\text{data}|\theta, \nu)\pi(\theta, \nu)$$

As we are only interested in θ, we marginalize by integrating over ν:

$$P(\theta) = \int P(\theta, \nu)\, d\nu$$

Prior knowledge on ν often is the result of a calibration measurement.

Example of a Frequentist approach: Profile method

Uncertainty in the probability function for the data described by nuisance parameter ν:

$$L(\theta, \nu) = \prod_i p(x_i|\theta, \nu)$$

If available, can include information on ν from additional measurements y_i:

$$L(\theta, \nu) = \prod_{i,j} p(x_i, y_j|\theta, \nu)$$

Eliminate the nuisance parameter by using the profile likelihood:

$$L_\mathrm{p}(\theta) = L(\theta, \hat{\hat{\nu}}(\theta))$$

$\hat{\hat{\nu}}(\theta)$: value of ν which maximizes $L(\theta, \nu)$ for a given θ

Profile likelihood ratio as test statistics

Let q be a test statistic and $h(q \mid \theta, \nu)$ its distribution. The p-value depends on the nuisance parameter ν:

$$p_\theta(\nu) = \int_{q_{\text{obs}}}^{\infty} h(q|\theta, \nu)\, dq$$

Independence of the nuisance parameter is achieved approximately by using the *profile likelihood ratio* as test statistic:

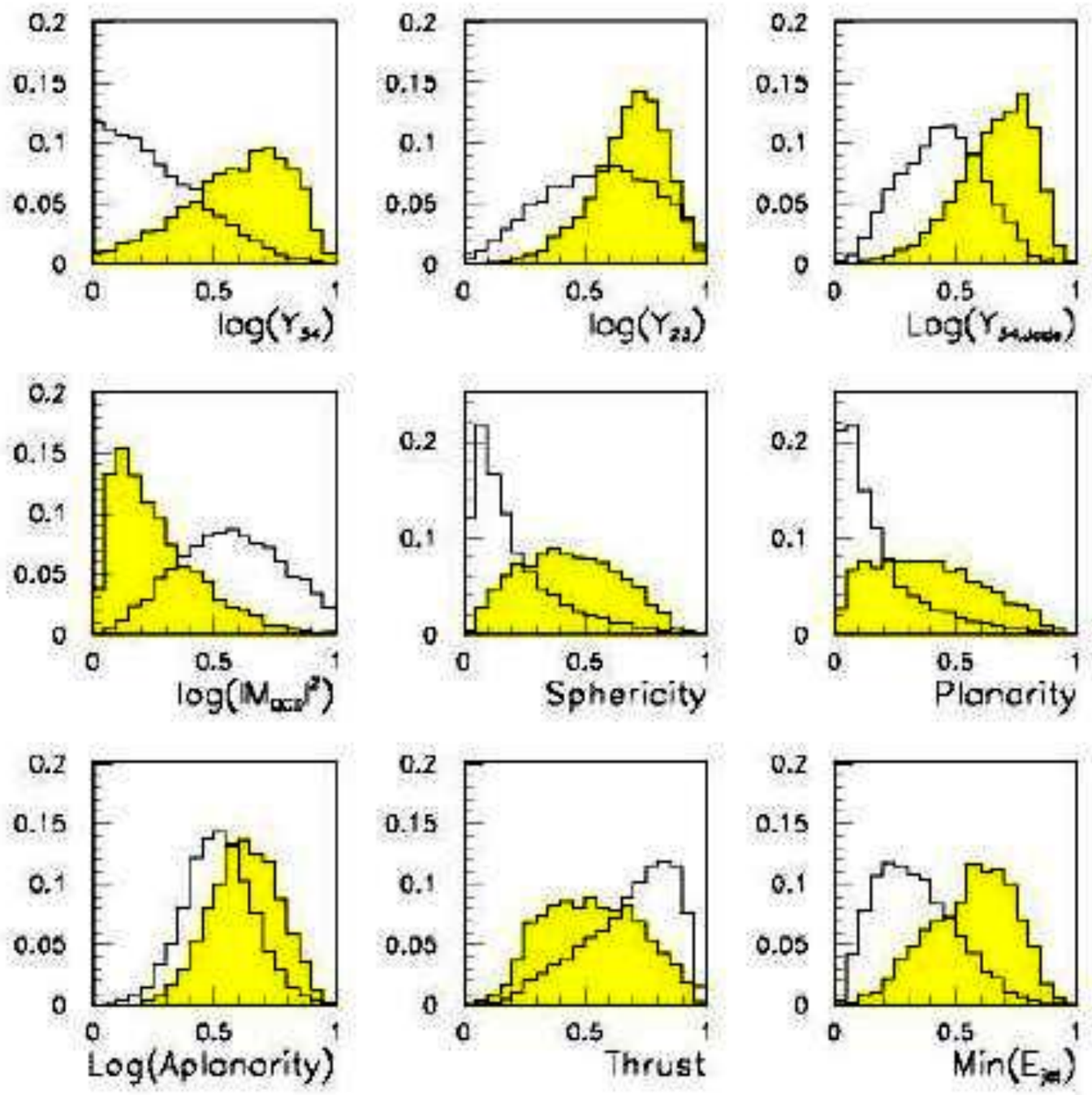
$$\lambda_{\text{p}}(\theta) = \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

This is motivated by the fact that $-2 \ln \lambda_{\text{p}}(\theta)$ approaches the χ^2 distribution (with n_{dof} = number of parameters of interest) for a large data sample (→ Wilks' theorem).

Decision trees

Multivariate Analysis: An Early Example from Particle Physics

G. Cowan, Lecture on Statistical data analysis



(Garrido, Juste and Martinez, ALEPH 96-144)

Signal: $e^+e^- \rightarrow W^+W^-$

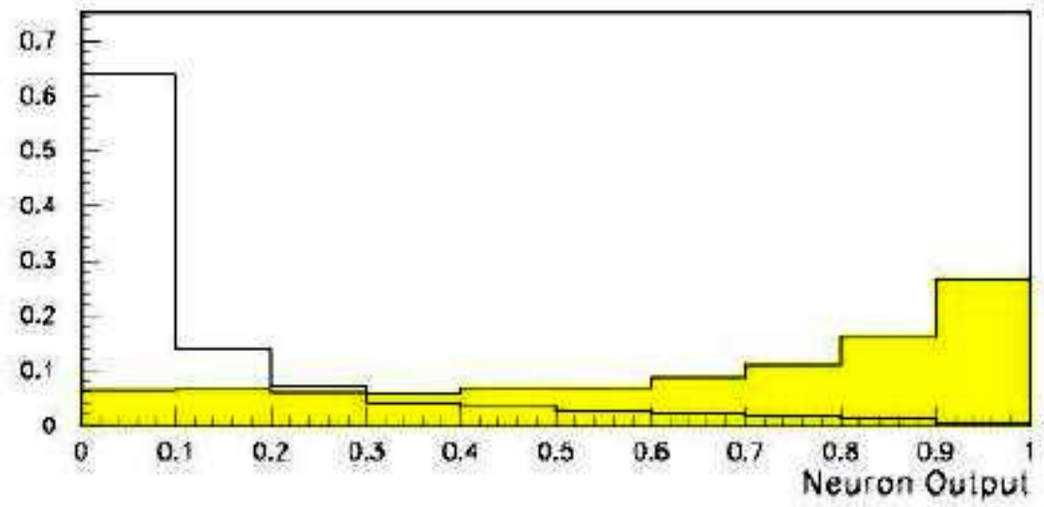
often 4 well separated hadron jets

Background: e+e– → qqgg

4 less well separated hadron jets

← input variables based on jet structure, event shape, ... none by itself gives much separation.

Neural network output:



Multi-Variate Classification

Consider events which can be either signal or background events.

Each event is characterized by *n* observables:

$$\vec{x} = (x_1, ..., x_n)$$ "feature vector"

Goal: classify events as signal or background in an optimal way.

This is usually done by mapping the feature vector to a single variable, i.e., to scalar test statistic:

$$\mathbb{R}^n \to \mathbb{R}: \quad y(\vec{x})$$

A cut $y > c$ to classify events as signal corresponds to selecting a potentially complicated hyper-surface in feature space. In general superior to classical "rectangular" cuts on the x_i.

Classification and Regression

The codomain Y of the function y: $X \rightarrow Y$ can be a set of labels or classes or a continuous domain, e.g., $\mathbb{R}$

Binary classification: $Y = \{0, 1\}$ e.g., signal or background

Multi-class classification $Y = \{c_1, c_2, ..., c_n\}$

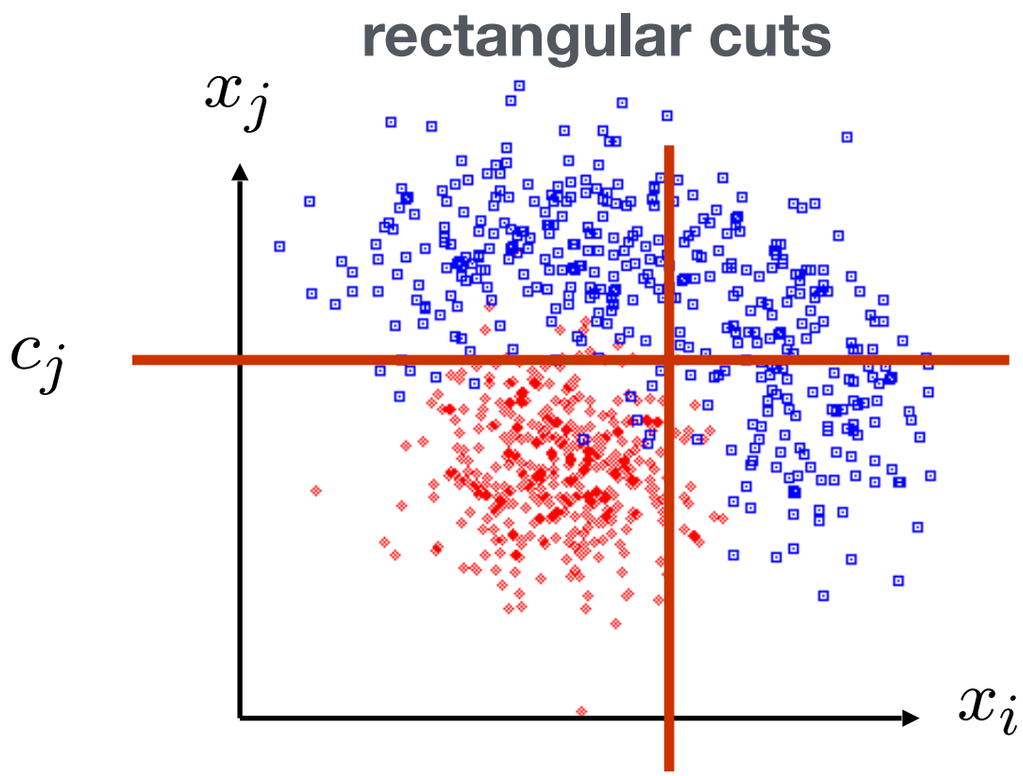
Y = finite set of labels → classification

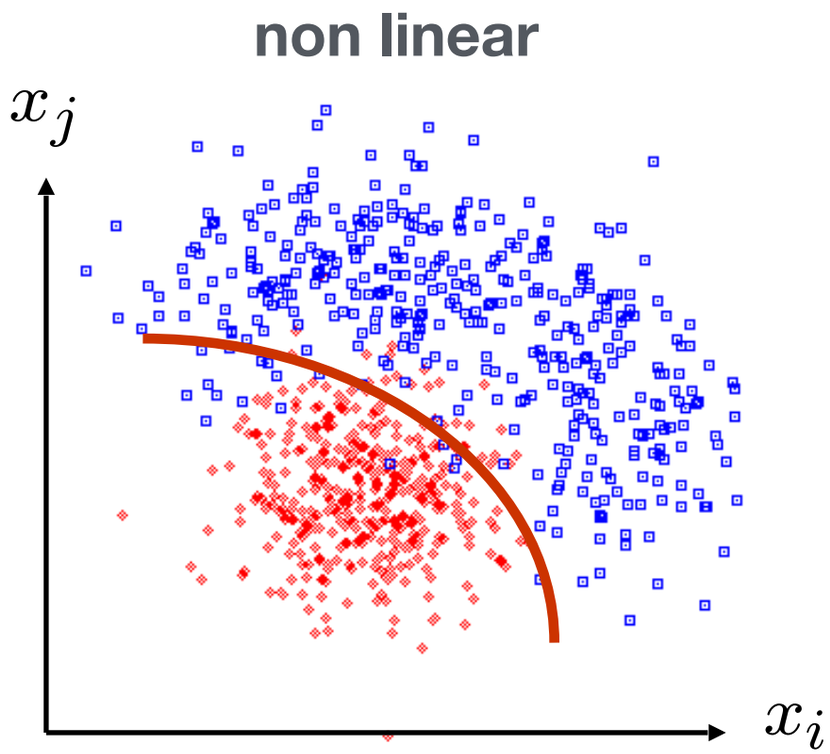
Y = real numbers → regression

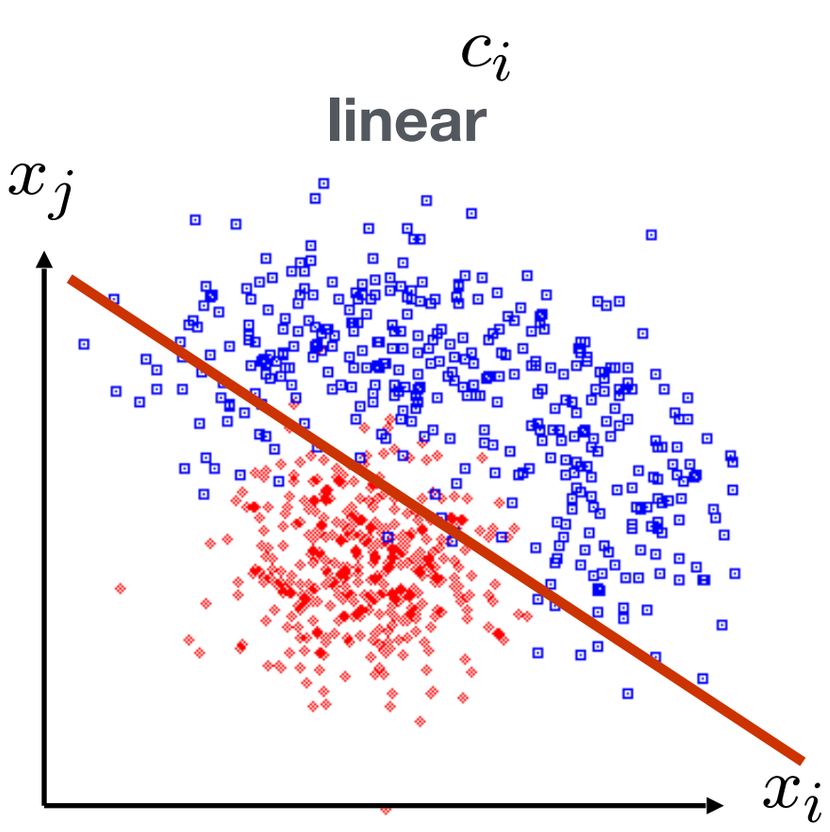
"All the impressive achievements of deep learning amount to just curve fitting"

J. Pearl, Turing Award Winner 2011,
https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

Classification: Different Approaches







k-Nearest-Neighbor,
Boosted Decision Trees,
Random forests
Multi-Layer Perceptrons,
Support Vector Machines
Deep Neural Networks,
…

G. Cowan':
https://www.pp.rhul.ac.uk/~cowan/stat_course.html

Different Approaches to Classification

Neyman-Pearson lemma states that likelihood ratio provides an optimal test statistic for classification:

$$y(\vec{x}) = \frac{p(\vec{x}|S)}{p(\vec{x}|B)}$$

Problem: the underlying pdf's are almost never known explicitly.

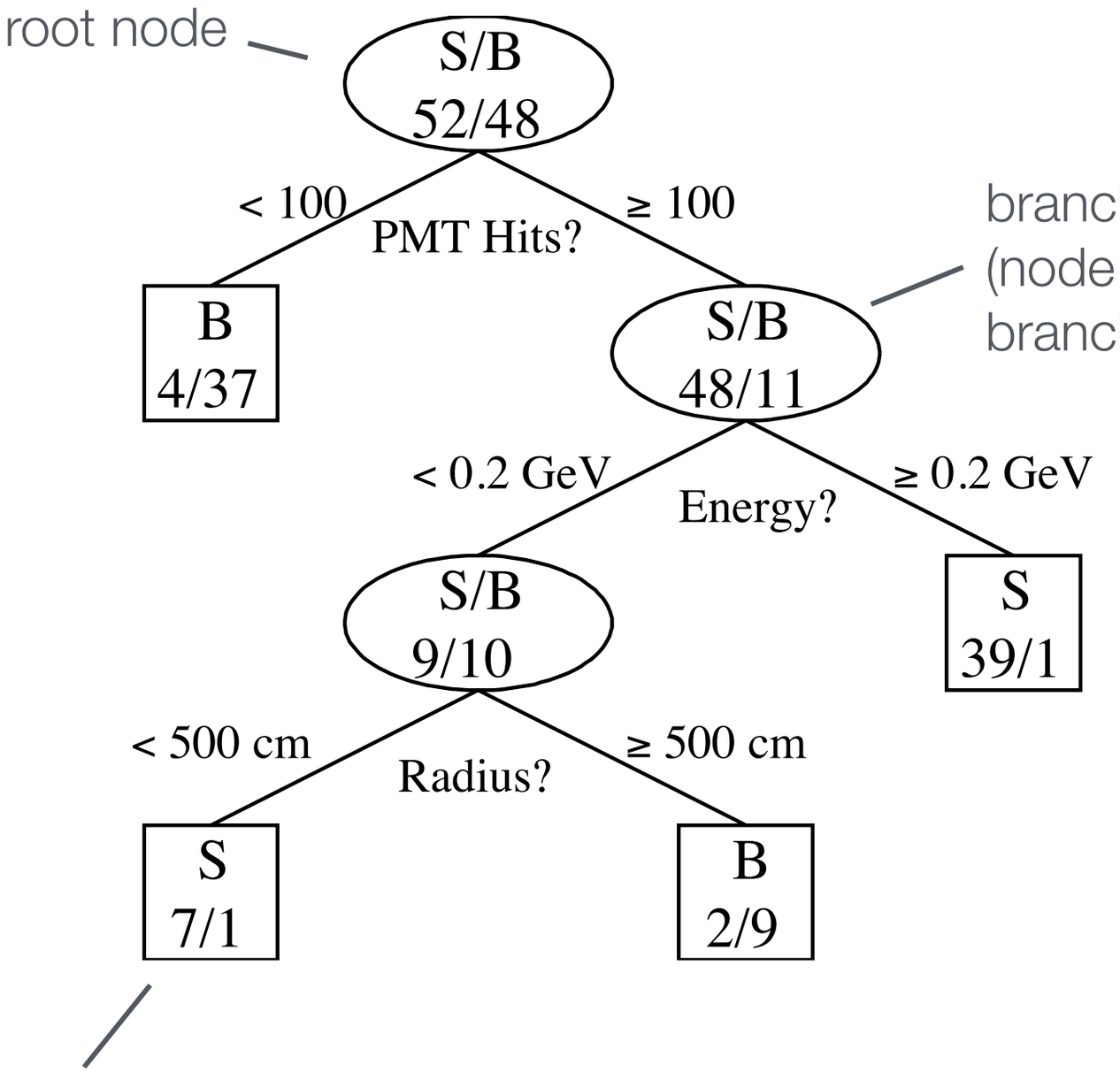
Two basic approaches:

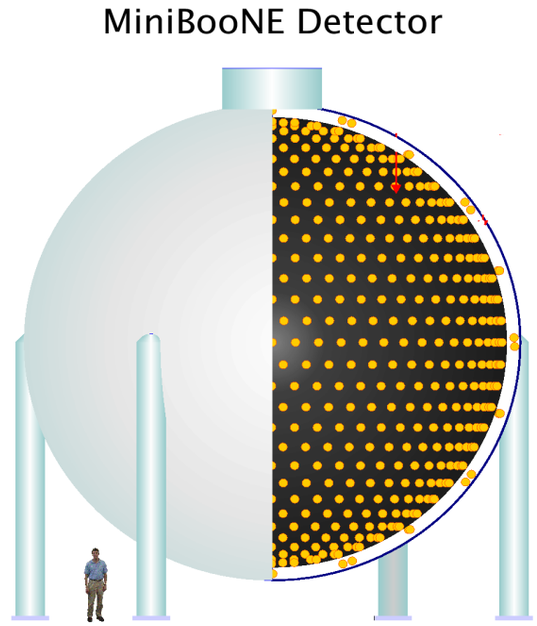
1. Estimate signal and background pdf's and construct test statistic based on Neyman-Pearson lemma, e.g. Naïve Bayes classifier (= Likelihood classifier)

2. Decision boundaries determined directly without approximating the pdf's (linear discriminants, **decision trees**, neural networks, …)

Decision Trees

arXiv:physics/0508045v1





MiniBooNE: 1520 photomultiplier signals, goal: separation of ν_e from ν_μ events

Space of feature vectors split up into rectangular volumes, attributed to either signal or background

Boosted Decision Trees and Random Forests

Drawback of decisions trees:
very sensitive to statistical fluctuations in training sample → boosting

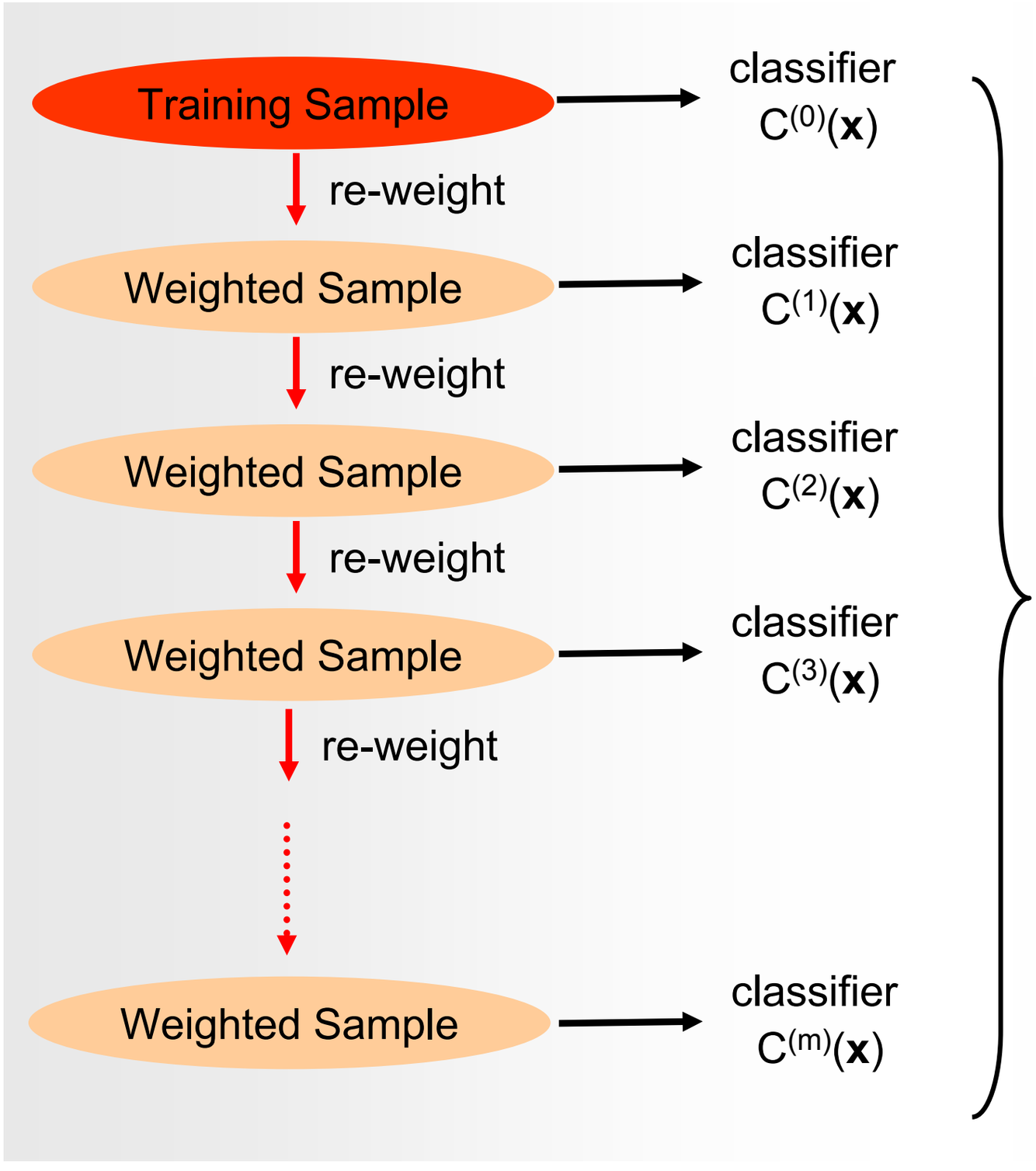
Boosting is a general method of combining a set of classifiers (not necessarily decisions trees) into a new, more stable classifier with smaller error.

Boosted decision trees:
Continuously increase the weight of incorrectly identified events and build new trees; take weighted average

Random forests:
Build many independent trees from random subsets of the training sample; makes the decision more robust to missing data

Both show excellent performance and are easy to train
(not much tuning needed)

Boosted Decision Trees: Idea



Weight is increased if event was misclassified by the previous classifier

→ "Next classifier should pay more attention to misclassified events"

$$y(x) = \sum_{i}^{N_{Classifier}} w_i C^{(i)}(x)$$

Popular example: AdaBoost (Freund, Schapire, 1997)

H. Voss, Lecture: Graduierten-Kolleg, http://tmva.sourceforge.net/talks.shtml

Practical Advice – Which Algorithm to Choose?

M. Kagan, https://indico.cern.ch/event/619370/

From Kaggle competitions:

Structured data: "High level" features that have meaning

- feature engineering + decision trees
- Random forests
- XGBoost

] require little or no preprocessing of the data

Unstructured data: "Low level" features, no individual meaning

- deep neural networks (DNN)
- e.g. image classification: convolutional NN

DNN have some impressive applications, but are they really the future?

Geoffrey Hinton (DNN pioneer), 2017: "My view is throw it all away and start again"