Introduction to data analysis

DESY summer school 2023

Orel Gueta

DESY

August 3 & 4, 2023

HELMHOLTZ
GEMEINSCHAFT

Acknowledgements and further reading

A lot of inspiration taken from the following lectures:

- Louis Lyons - Practical Statistics for Physicists.
- Stephanie Hansmann-Menzemer - Modern Methods of Data Analysis.
- Andreas Hoecker - Foundations of statistics.
- Tommaso Dorigo - Statistics Topics for Data Analysis in Particle Physics: an Introduction.
- Kyle Cranmer - Practical Statistics for Particle Physics.
- Thomas Junk - Data Analysis and Statistical Methods in Experimental Particle Physics.

Books:

- Particle data group statistics review - concise, contains almost everything.
- Glen Cowan - Statistical data analysis.
- Trevor Hastie et al. - The Elements of Statistical Learning.
- Olaf Behnke et al. - Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods.
- J. VanderPlas et al. - Introduction to astroML: Machine learning for astrophysics

Outline

Alternative title could be

“Practical statistics for physicists”

Four lectures

- Lectures 1 & 2 - Introduction to data analysis (orel.gueta@desy.de).
- Lecture 3 & 4 - Dan Parsons - Machine Learning Techniques.

Outline

- Introduction
- Probability and statistical distributions
- Parameter and uncertainty estimation
- Bayesian vs Frequentist
- Hypothesis testing
- Monte Carlo methods

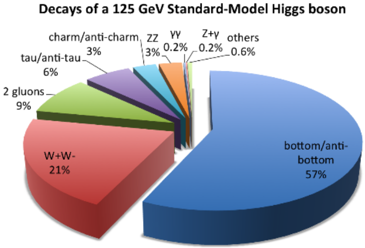
* Slightly biased towards particle physics

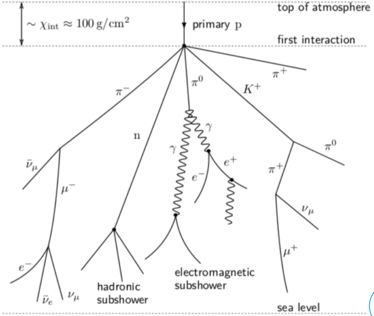
Please feel free to stop me and ask questions!

Introduction

Data analysis in physics involves a lot of probability and statistics.

> Quantum phenomena is probabilistic in nature.
> So is the particle interaction with the detector (e.g., air shower fluctuations).
> Theory only provides probabilities (e.g., Higgs decay channels).
> Analyze large amounts of data and compare to probabilities.
> Utilize Monte Carlo methods to simulate probabilistic phenomena.





Introduction

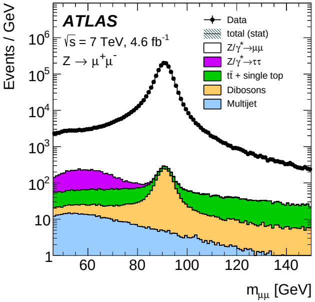
Where is data analysis used?

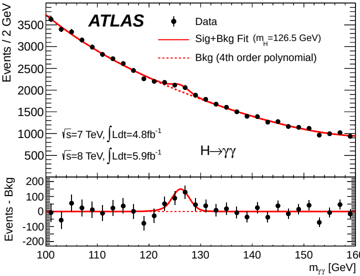
> Measure a known parameter and its uncertainty (mass of the Z boson).

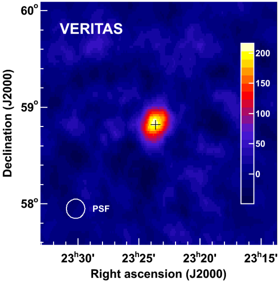
> Discover new phenomena (Higgs, γ-ray/neutrino source).

> Test your theory against the data (hypothesis testing).

$\Rightarrow$ Extract as much as possible from data (experiments are expensive!)







Silly example

Introduction - cheating baker

Simple example of data visualization.

- A restaurant owner orders 30 rolls every day.
- The law in the country states that rolls must weigh $\sim$ 75 grams;
- After changing suppliers, the owner suspects that the new baker sells underweight rolls

$\Rightarrow$ Investigate! Weigh the rolls (1 gram resolution).

```
78, 66, 67, 64, 74, 58, 78, 66, 71, 68, 77, 59, 68, 68, 75, 64, 69, 65, 70, 72
64, 75, 74, 72, 74, 66, 69, 65, 68, 72, 66, 68, 66, 65, 66, 69, 64, 71, 78, 73
69, 65, 66, 78, 70, 66, 70, 80, 70, 73, 71, 68, 64, 68, 68, 72, 74, 74, 71, 74
66, 76, 72, 68, 72, 69, 75, 77, 80, 63, 62, 67, 70, 74, 71, 59, 68, 74, 71, 73
68, 68, 70, 72, 70, 70, 66, 71, 70, 75, 75, 70, 68, 66, 72, 70, 68, 70, 66, 73
67, 76, 72, 72, 64, 70, 73, 65, 68, 70, 63, 71, 74, 65, 71, 63, 69, 61, 75, 72
69, 66, 76, 79, 60, 76, 78, 71, 64, 74, 69, 66, 63, 72, 73, 66, 72, 64, 74, 70
69, 69, 74, 73, 72, 70, 70, 73, 71, 73, 68, 57, 75, 80, 72, 69, 69, 69, 70, 64
67, 65, 71, 68, 72, 69, 74, 71, 80, 60, 66, 74, 82, 68, 70, 68, 76, 68, 73, 68
63, 71, 72, 76, 69, 66, 72, 71, 71, 69, 75, 71, 79, 75, 73, 61, 73, 72, 74, 75
67, 74, 67, 79, 63, 61, 61, 65, 64, 79, 68, 63, 75, 67, 63, 74, 73, 67, 77, 70
78, 72, 77, 70, 63, 82, 67, 69, 66, 68, 71, 71, 73, 78, 69, 63, 64, 66, 61, 74
67, 68, 69, 63, 65, 73, 73, 67, 79, 63, 68, 68, 56, 79, 71, 64, 80, 72, 72, 66
70, 62, 73, 68, 70, 76, 71, 71, 71, 66, 74, 77, 73, 74, 65, 65, 62, 76, 68, 76
66, 67, 70, 74, 70, 71, 70, 70, 64, 70, 69, 69, 72, 66, 69, 68, 72, 73, 65, 72
```

- Raw list of weights is not very useful.

Introduction - cheating baker

A friend suggests to reduce the data,

- combine the measurements, taking into account resolution;
- assume the rolls are produced independently, e.g., neglect changes from week to week.

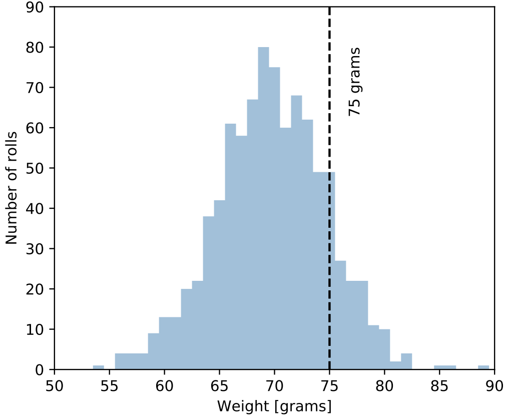
```
Weight[50] = 0  Weight[51] = 0  Weight[52] = 0  Weight[53] = 0  Weight[54] = 1
Weight[55] = 0  Weight[56] = 4  Weight[57] = 4  Weight[58] = 4  Weight[59] = 9
Weight[60] = 13 Weight[61] = 13 Weight[62] = 20 Weight[63] = 22 Weight[64] = 38
Weight[65] = 42 Weight[66] = 61 Weight[67] = 58 Weight[68] = 67 Weight[69] = 80
Weight[70] = 75 Weight[71] = 60 Weight[72] = 68 Weight[73] = 62 Weight[74] = 49
Weight[75] = 49 Weight[76] = 27 Weight[77] = 22 Weight[78] = 22 Weight[79] = 11
Weight[80] = 10 Weight[81] = 2  Weight[82] = 4  Weight[83] = 0  Weight[84] = 0
Weight[85] = 1  Weight[86] = 1  Weight[87] = 0  Weight[88] = 0  Weight[89] = 1
```

- Can see that the majority of rolls weigh less than 75 grams.
- Easier to understand the data this way, but still far from perfect.
- Better idea to visualize the data?

Introduction - cheating baker

Visualize the data with a histogram,

- immediately grasp the distribution of weights;
- mean and standard deviation clearly visible.



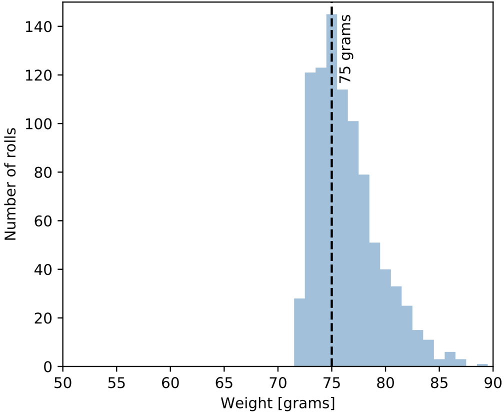
- New baker definitely cheating, rolls are about 5 grams too light.

Introduction - cheating baker

- The owner complains to the baker.
- The baker promises to correct their ways $\rightarrow$ The restaurant owner keeps monitoring.

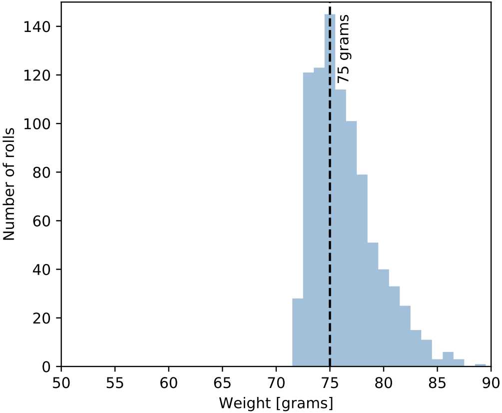
Introduction - cheating baker

- The owner complains to the baker.
- The baker promises to correct their ways → The restaurant owner keeps monitoring



Introduction - cheating baker

- The owner complains to the baker.
- The baker promises to correct their ways → The restaurant owner keeps monitoring



- A month later the owner sees the baker is still cheating, sending the restaurant the heaviest rolls and selling the light ones to others.

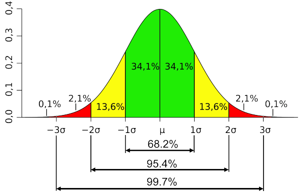
Statistical distributions

Statistical distributions

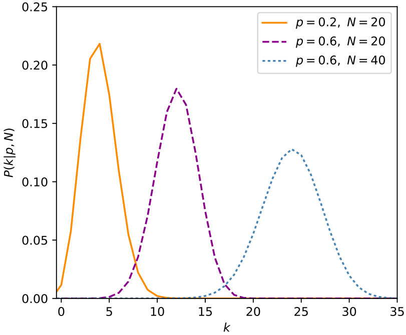
Measurements typically follow a distribution, identifying it could be important

- correct determination of parameters;
- uncertainties estimation;
- for results interpretation (see example later).

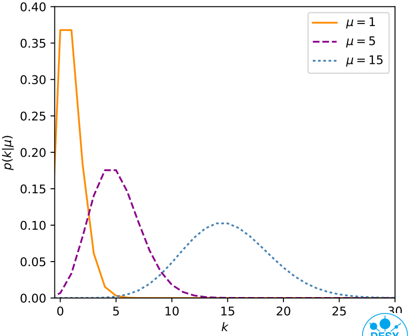
Gaussian



Binomial

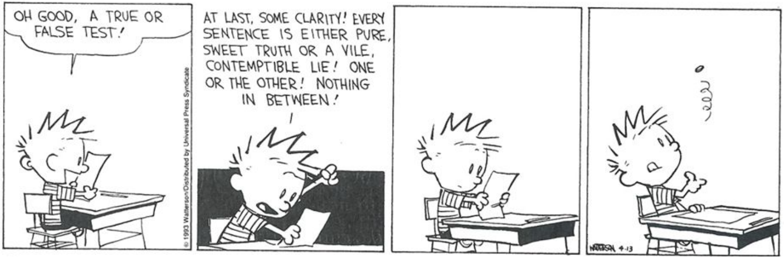


Poisson



Binomial distribution

Experiment has two outcomes,



For N "coins", each with prob. of "success" p,

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

is the prob. of k successes.

- What is the prob. to roll ⚅ 34 times out of 100 throws?
- Selection or reconstruction efficiency (prob. to reconstruct 560 γ's with $p = 0.63$ and $N = 10^3$).

Binomial distribution

Characteristics,

- Expectation value (mean, μ), $E[k] = \sum_k kP(k) = Np$.
- Variance (σ^2), $E[(k - \langle k \rangle)^2] = E[k^2] - (E[k])^2 = Np(1-p)$.

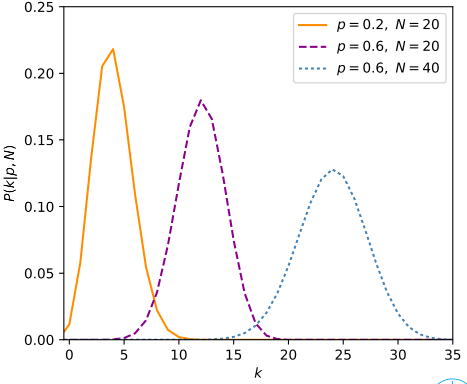
Intuitive, e.g., calculate for 8 coin flips.

Take into account when dealing with efficiencies

- `ROOT` includes various options to use binomial errors for efficiency (e.g., `TEfficiency`). Similar tools exist for `R` and Python.

Limiting cases,

- For $N \to \infty$, $p \to 0$
  $Np =$ const.,
  Binomial $\to$ Poisson.
- For $N \to \infty$, $p =$ const.,
  Binomial $\to$ Gaussian.



Poisson distribution

Prob. of N independent events occurring in time interval Δt with constant rate μ,

$$P(N;\mu) = \frac{\mu^N}{N!}e^{-\mu}$$

> Expectation value, $E[N] = \sum_N NP(N) = \mu$. Variance, $\sigma^2 = \mu$.

Where do we run into this dist.?

Poisson distribution

Prob. of N independent events occurring in time interval Δt with constant rate μ,

$$P(N;\mu) = \frac{\mu^N}{N!}e^{-\mu}$$

- Expectation value, $E[N] = \sum_N NP(N) = \mu$. Variance, $\sigma^2 = \mu$.

Where do we run into this dist.?

- Number of decay events per second from a radioactive source;

Poisson distribution

Prob. of N independent events occurring in time interval Δt with constant rate μ,

$$P(N;\mu) = \frac{\mu^N}{N!} e^{-\mu}$$

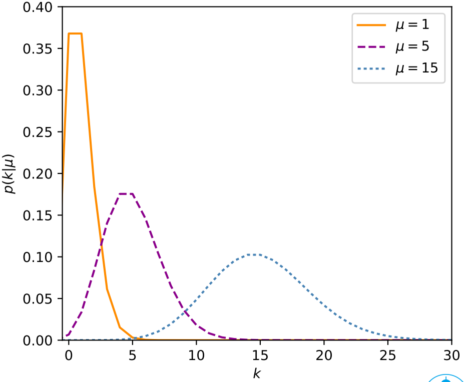
> Expectation value, $E[N] = \sum_N NP(N) = \mu$. Variance, $\sigma^2 = \mu$.

Where do we run into this dist.?

- Number of decay events per second from a radioactive source;
- Number of "rare" interactions occurring per bunch crossing at LHC;
- Number events in a histogram bin.

→ typical, $N \pm \sqrt{N}$ (what about 0 ± 0?).

When $\mu \to \infty$, Poisson → Gaussian.



Importance of distribution identification

Example - evidence of quarks in air showers.

> Researchers observed a track with 110 bubbles (average expected is 229, 55,000 tracks in total).

> They assumed (correctly) bubble formation is a Poisson-distributed quantity.

⇒ Probability of observation $P \sim 10^{-13}$.

⇒ Particles with fractional charge!

In fact,

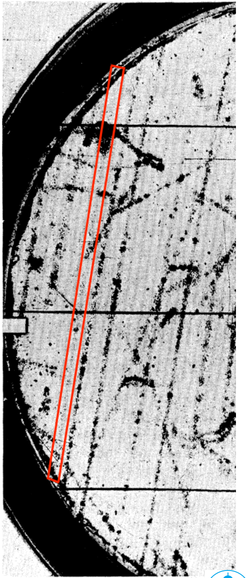
> each scatter of a charged particle off a nucleus produces ∼4 droplets.

⇒ Both particle scattering and bubble formation are Poisson processes.

⇒ Need to use a compound Poisson distribution.

> P to observe one 110 bubble track, $P \approx 5 \cdot 10^{-5}$;

> Observing one such track out of 55,000, $P \sim 92\%$.



EVIDENCE OF QUARKS IN AIR-SHOWER CORES*

C. B. A. McCusker and I. Cairns
Cornell-Sydney University Astronomy Center, Physics Department, The University of Sydney, Sydney, Australia
(Received 3 September 1969)

In a study of air-shower cores using a delayed-expansion cloud chamber, we have ob-

Gaussian distribution

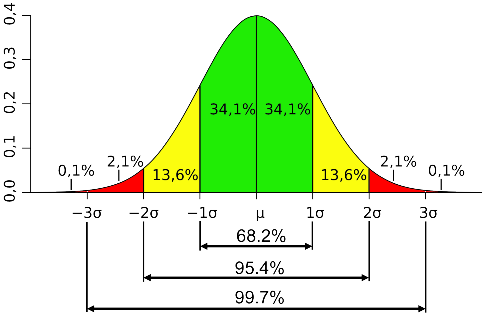
Probably the most common distribution (thanks to Central Limit Theorem),

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

- Expectation value, $E[x] = \mu$.
- Variance, $\sigma^2 = \sigma^2$.
- At $x = \mu \pm \sigma$, $y = y_{\mathrm{max}}/\sqrt{e} \sim 0.606 \times y_{\mathrm{max}}$.

Probability content often used

- $\int_{-\sigma}^{+\sigma} P(x;\mu,\sigma)dx = 68.2\%$;
- $\int_{-2\sigma}^{+2\sigma} P(x;\mu,\sigma)dx = 95.4\%$
- etc.



Central Limit Theorem

Idea:

> pick k random variables from **any** distribution $Q(x)$;

> repeat N times and calculate mean (or sum) between the variables;

⇒ the distribution of the mean values will be Gaussian.

* $Q(x)$ should be well defined.

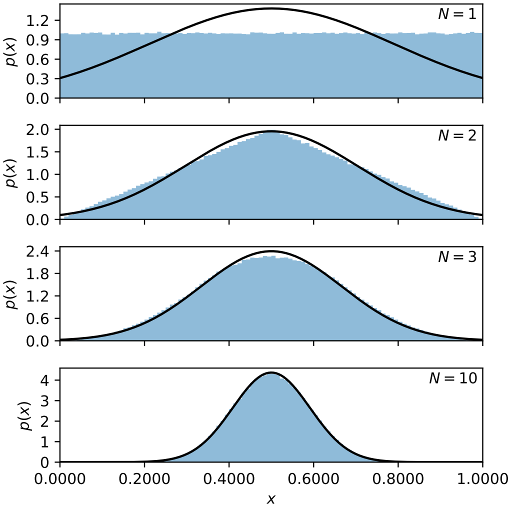
Illustration:

> Uniform $Q(x)$;

> Gaussian is shown for $\mu = 0.5$ and $\sigma = 1/\sqrt{12N}$;

> Already for $N = 10$, Gaussian distribution observed.

> Larger N for non-uniform $Q(x)$.



Central Limit Theorem

Idea:

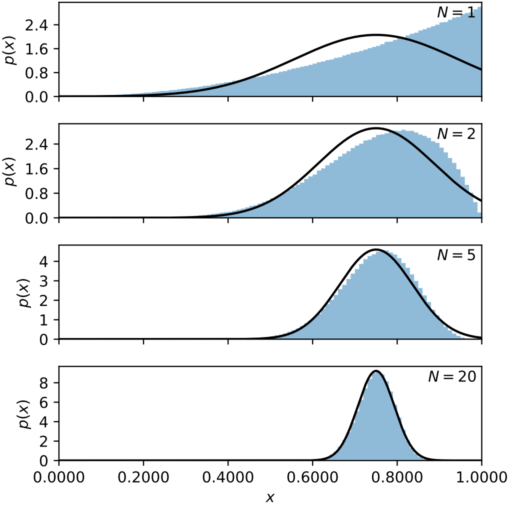
- pick k random variables from **any** distribution $Q(x)$;
- repeat N times and calculate mean (or sum) between the variables;

⇒ the distribution of the mean values will be Gaussian.

- $*$ $Q(x)$ should be well defined.

Illustration:

- Parabolic $Q(x)$;
- Gaussian is shown for $\mu = 0.75$ and $\sigma = \sigma(N)/\sqrt{N}$;
- Requires $N = 20$ to obtain Gaussian distribution.
- Try it yourselves!

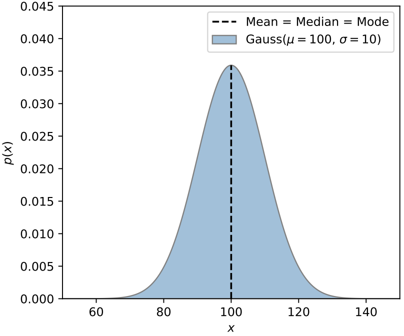


A few extra comments on distributions

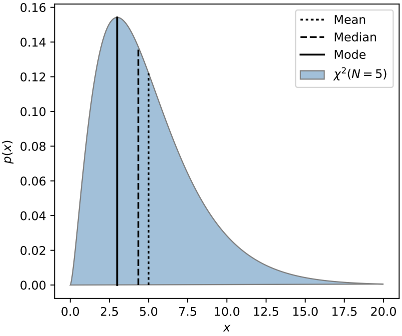
Other characteristics,

- Mode (most-probable value)
- Median (or more generally k-quantiles)

Symmetric



Non-symmetric



Have not mentioned so far,

- continuous or discrete distributions;
- cumulative distributions.

Parameter and uncertainty estimation

Parameter estimation - least square fit

Data: $x_i, y_i \pm \sigma$, Theory: $y = ax + b$.

> Parameter determination.

> Goodness of fit.

Least square fit

$$\chi^2 = \sum_i \left(\frac{(ax_i+b)-y_i}{\sigma}\right)^2$$

* not really χ^2 (convention).

> Linear $\Rightarrow$ minimize analytically
$a = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_i (x_i - \langle x \rangle)^2}$
$b = \langle y \rangle - a\langle x \rangle$

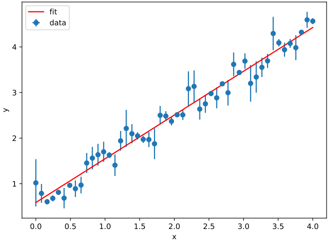
* When $\sigma \to \sigma_i$, perform numerically, assuming normally distributed uncertainties.

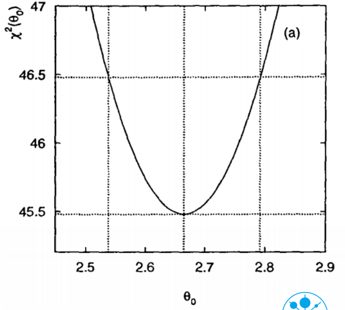
Uncertainties

> with enough data, χ^2 usually parabolic;

> $\sigma_\theta^2 = 2 / \left(d^2\chi^2 / d\theta^2\right)$

> scan parameter space for
$\chi^2(\theta) = \chi^2_{\text{min}}(\theta_{\text{best}}) + 1$;





Uncertainties

Suppose result/theory = 0.970, does the theory describe the data?

Uncertainties

Suppose result/theory = 0.970, does the theory describe the data?

0.970 ± 0.05 $\quad$ 0.970 ± 0.005 $\quad$ 0.970 ± 0.5

Uncertainties

Suppose result/theory = 0.970, does the theory describe the data?

0.970 ± 0.05 $\quad$ 0.970 ± 0.005 $\quad$ 0.970 ± 0.5

Statistical uncertainties

- Random in nature.
- Fluctuates independently per measurement.
- Unavoidable.
- Usually, more data → lower uncertainty ($\propto \sqrt{N}$).
- e.g., counting statistics, electronic noise, etc.

Systematic uncertainties

- Usually originate in the instrument.
- Bias the data by unknown ∼constant offset.
- Hard to detect, correct for, estimate.
- e.g., miscalibration, diff. between data and simulation, simulation statistics, etc.

$$\sigma(\text{tot.}) = \sigma\,(\text{stat.}) \oplus \sigma\,(\text{syst.})$$

- Report uncertainties separately (sometimes diff. syst. contributions).
- Pick your battles.
- Take into account theoretical uncertainty.

Uncertainty propagation

Assume $y = f(x) \Rightarrow \sigma_y = \frac{df(x)}{dx}\Big|_{x=\bar{x}} \cdot \sigma_x$

> Taylor expansion approximation, small uncertainty.

With more variables, $f(x_1, x_2, \ldots, x_N)$, take correlation into account

$$\sigma_y^2 = \sum_{i,j}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j}\Big|_{x=\bar{x}} \cdot V_{x_i,x_j}$$

> V_{x_i,x_j} is the covariance of x_i, x_j (see later).
> Correlated variables lead to increased uncertainty.
> Opposite for anti-correlated.

examples

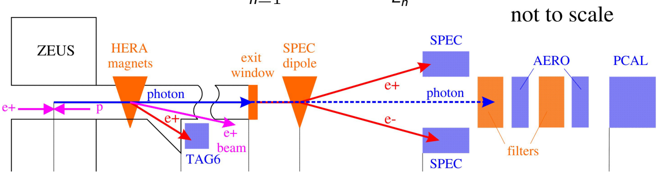
> $y = x_1 - x_2 \Rightarrow \sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 - 2 \cdot V_{x_1,x_2}$.
> $y = x_1^{\alpha} \cdot x_2^{\beta}$, fractional uncertainties are useful (uncorr.)
$\Rightarrow \left(\frac{\sigma_y}{y}\right)^2 = \left(\alpha \frac{\sigma_{x_1}}{x_1}\right)^2 + \left(\beta \frac{\sigma_{x_2}}{x_2}\right)^2$.
> Sometimes easier numerically (uncorr.)
$y_1 = f(x_1 + \sigma_{x_1}, x_2, \ldots, x_N)$, $y_2 = f(x_1, x_2 + \sigma_{x_2}, \ldots, x_N)$, etc.
$\sigma_y^2 = (y - y_1)^2 + (y - y_2)^2 + \ldots + (y - y_N)^2$.

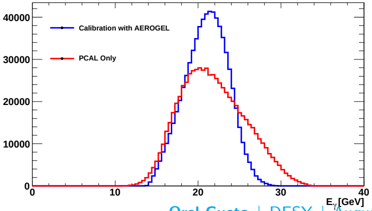
Quick examples - LS

Calibrate detectors

- $E_\gamma = \alpha_1 \cdot E_{\text{AERO}_1} + \alpha_2 \cdot E_{\text{AERO}_2} + \alpha_3 \cdot E_{\text{PCAL}}$
- obtain E_γ from beam energy and other calibrated detector.
- for all data available, minimize

$$\chi^2 = \sum_{n=1}^{N} \frac{\left(\sum_{j=1}^{3} \alpha_j E_{j,n} - E_n^\gamma\right)^2}{\sigma^2_{E_n^\gamma}}$$





Quick examples - LS

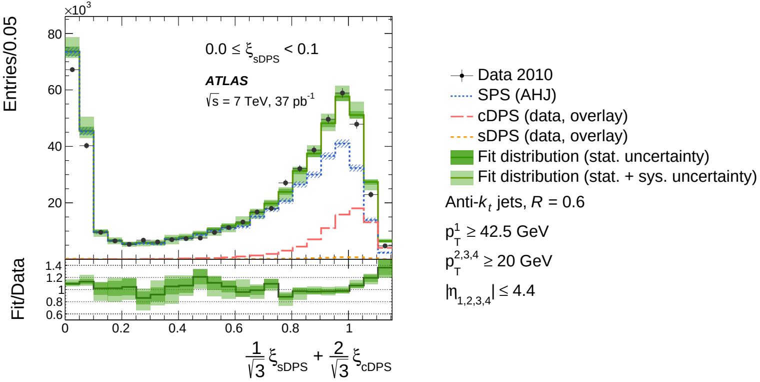
Estimate contributions from signal/background,

> minimize to get optimal relative fractions

$$\mathcal{D} = (1-f)\mathcal{H}_1 + f\mathcal{H}_2$$

> both model and data are binned and have uncertainties

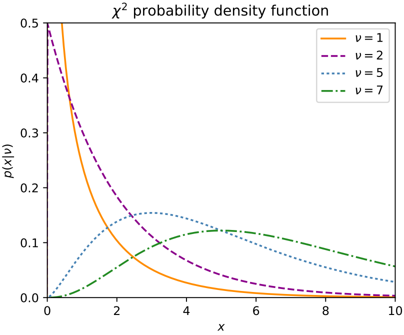
⇒ can only be done numerically (`ROOT`, `Minuit`).

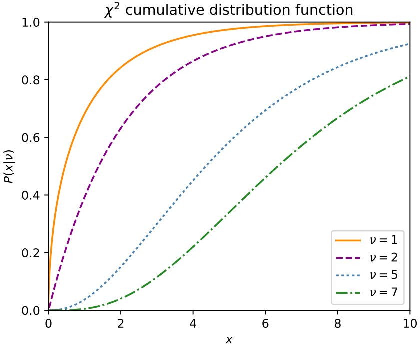


Goodness of fit

In the least squares case, straightforward

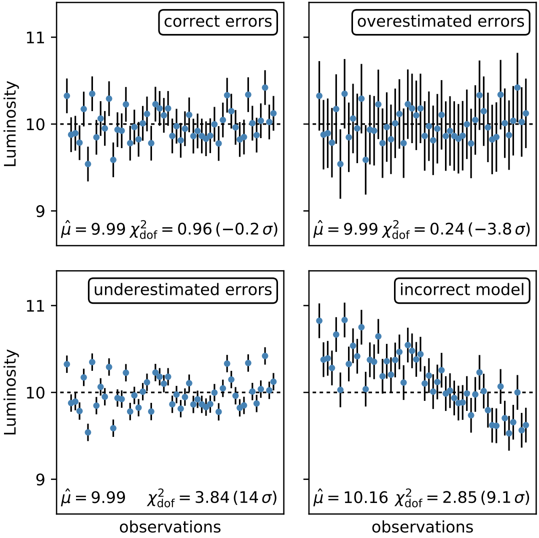
- determine χ^2_{min} and number of degrees of freedom, $\nu = n - p$;
- check probability based on χ^2 distribution (`TMath::Prob(chi2, ndf)`).
  - usually referred to as p-value, prob. to find $\chi^2 > \chi^2_{\text{min}}$ (see later)
- Rule of thumb, $\chi^2_{\text{dof}} = \chi^2/\nu \approx 1$.





Quick example - goodness of fit

Check if brightness of star varies with time



Kolmogorov–Smirnov test

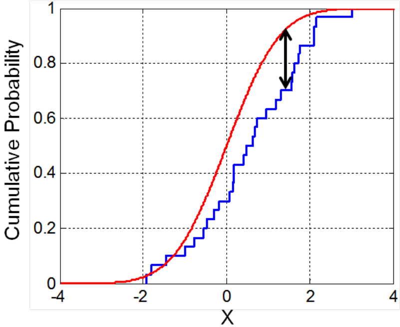
Test if distributions originate from the same underlying PDF.

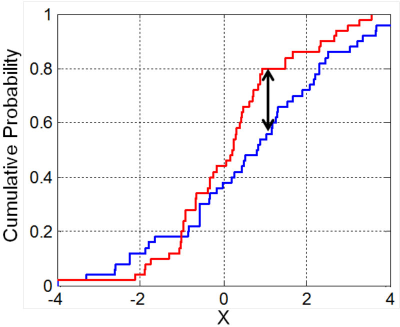
> Search for largest difference between cumulative distributions.

> Useful with small amounts of data (can be used as goodness of fit).

> Fast, non-parametric, sensitive to differences in location and shape of cumulative distributions.

> Example - automatic testing of simulation output distributions.





* χ^2 also available

Probability:
Bayesian vs Frequentist

Brief intro to probability

Axioms (Kolmogorov):

> $P(A) \in \mathbb{R}, \quad P(A) \geq 0, \quad \forall A \in \Omega$ (Ω is the event space).
> $\int_{\Omega} P(A) dA = 1$, i.e., Unitarity, prob. that at least one event will occur is 1.
> if $P(A \cap B) = 0$, then $P(A \cup B) = P(A) + P(B)$.

Conditional probability:

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

Frequentist

> How likely is an event to occur, based on many repeatable trials.
> Not applicable to a single event.
> Objective statement.

$$P(A) = \lim_{n_{\text{trials}} \to \infty} \frac{n_A}{n_{\text{trials}}}$$

Bayesian

> A "degree of belief" that an event will happen.
> Includes previous knowledge in it (prior).

Bayes theorem - $P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$

Bayesian vs Frequentist

Frequentist

⇒ Probability of data given a model, $P(\text{data}|\text{model})$.

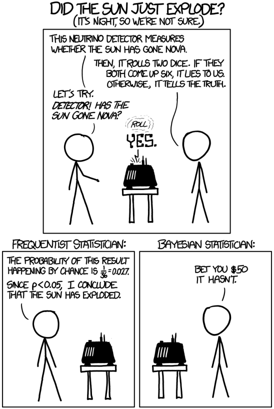
∗ "Frequentist statistics gives the probability to observe data under a given hypothesis, it says nothing about the probability of the hypothesis to be true".

Bayesian

⇒ Probability of model given data, $P(\text{model}|\text{data})$.

> $P(\text{model}|\text{data}) \propto P(\text{data}|\text{model}) \times P(\text{model}) \leftarrow$ prior.

- → could be previous measurements;
- → might be subjective;
- → functional form not always known (necessary?);
- → what if there is no knowledge?



Prior examples

> Physics is "smooth".

> mass squared of neutrino.

> "extraordinary claims require extraordinary evidence".

Covariance and correlation

Covariance and correlation

Consider measurements depending on more than one variable (observable).

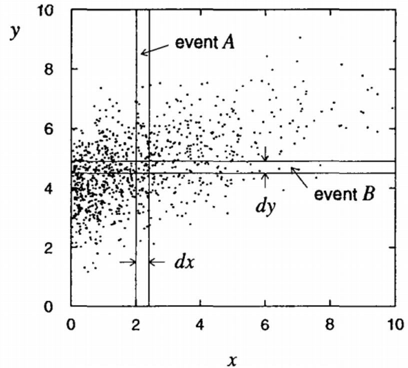
> What is prob. for A **and** B?

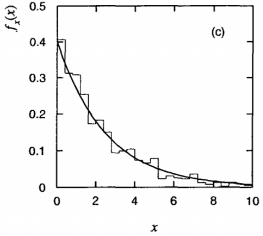
> $P(A) = f(x)dx$, $P(B) = f(y)dy \Rightarrow$ $P(A \cap B) = f(x,y)dxdy$.

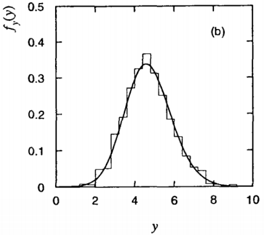
> The joint prob. $f(x,y)$ corresponds to the density of points ($N \to \infty$).

> If not interested in y dependence $\to$ project.

* Profiling (see later)







Covariance and correlation

How correlated are x and y?

$\Rightarrow$ Covariance

> Following the definition of 1D variance,
  $V(x) = \sigma_x^2 = E[(x - \langle x \rangle)^2] = E[x^2] - (E[x])^2$;

> $C(x, y) = V_{x,y} = E[(x - \langle x \rangle)(y - \langle y \rangle)] = E[xy] - E[x]E[y]$.

If x and y uncorrelated,

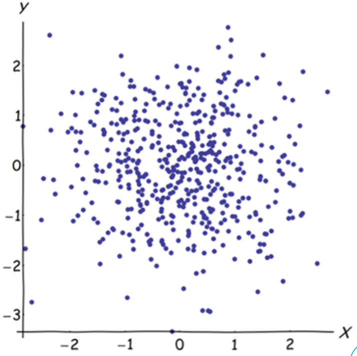
> $P(A \cap B) = P(A) \cdot P(B)$.

> $f(x, y) = f(x) \cdot f(y)$.

Remember uncertainty propagation?
with $y = f(x)$

$$\sigma_y^2 = \sum_{i,j}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \bigg|_{x=\bar{x}} \cdot V_{x_i,x_j}$$

notice $C = V_{x_i,x_j}$ is covariance matrix.



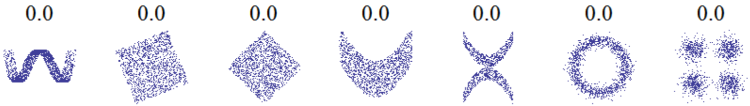
Covariance and correlation

The dimensionless Pearson's correlation coefficient

$$\rho_{x,y} = \frac{C(x,y)}{\sigma_x \sigma_y}$$

* does not measure slope.

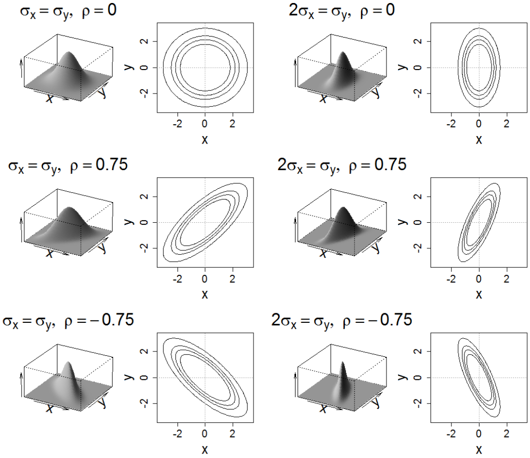
* Test **linear** correlation/anti-correlation. Always plot your data!



Example - 2D Gaussian

$$P(\vec{x}) = \frac{1}{2\pi\sqrt{\det(C)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})\right)$$

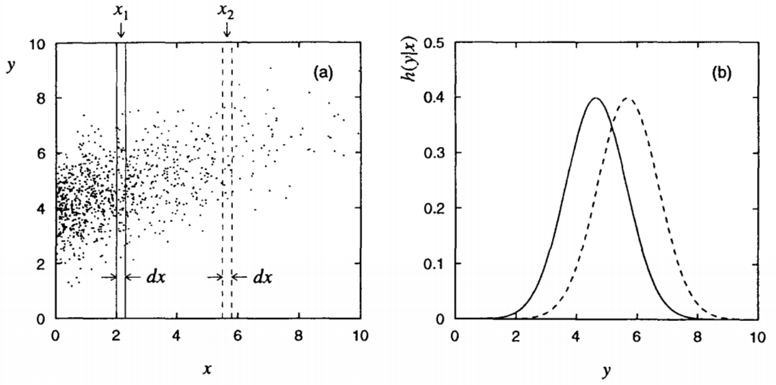
$$\text{for } \vec{x} = (x, y) \Rightarrow C = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & V_{x,y} \\ V_{x,y} & \sigma_y^2 \end{pmatrix}$$



Covariance and correlation

How to deal with correlated variables?

- If one of the variables is not used or cannot measure → project.
- Bin the data (profiling), issues with this method?



- Variable transformation.

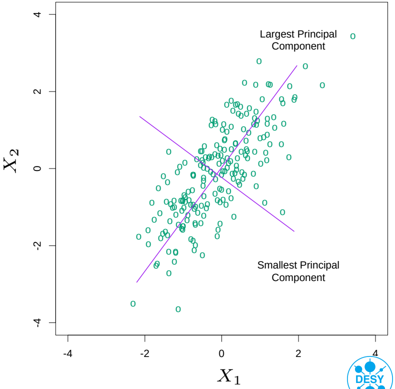
Principal component analysis

Perform orthogonal linear transformation, each component (variable) maximizes variance.
Process (X is data matrix),

- > diagonalize the $X^{\mathsf{T}}X$ matrix, calculate eigenvectors and eigenvalues;
- > the (ordered) eigenvectors are the new observables;
- > the variance "score" is given by the eigenvalues.

Some comments

- > Covariance, $C \propto X^{\mathsf{T}}X$.
- > First n components embody majority of information.
- ⇒ Can be used to reduce dimensionality.
- > Often one of the first steps in multi-variate analysis.
- > Useful only for linearly correlated variables (non-linear options available).
- > Various tools available (`ROOT`, `scikit-learn`, `R`).



Hypothesis testing

Parameter estimation - Maximum likelihood

Maximum likelihood for parameter determination.

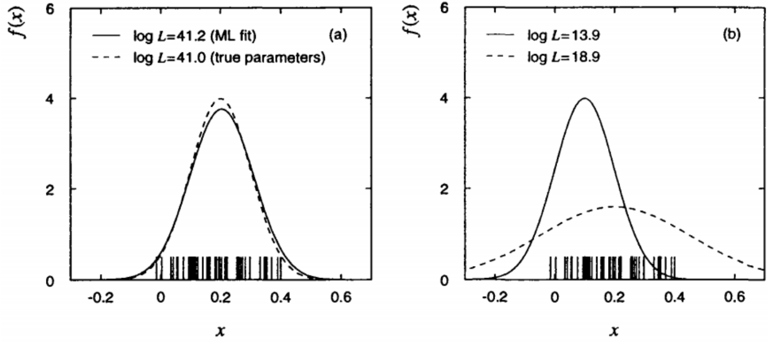
> Assume we observe N **independent** events, y_i.

> The hypothesis to check has a PDF, $p(y, \theta)$, where θ is param.

> Events are independent, combine prob as $\mathcal{L}(\theta) = \prod_i^N p(y_i, \theta)$.

→ calculate $\mathcal{L}(\theta)$ for all θ values (fixed y_i).

> $\mathcal{L}(\theta)$ is at maximum when $\theta = \theta_{\text{true}}$.



Maximum likelihood

Conventional to instead minimise $-2 \cdot \ln \mathcal{L}(\theta)$

- $\ln \mathcal{L}(\theta) = \sum_i^N p(y_i, \theta)$ (numerically easier).

Confidence interval

- $\ln \mathcal{L}(\theta_0 \pm \sigma) = \ln \mathcal{L}(\theta_0) - 1/2$ (also $\frac{d^2 \ln \mathcal{L}(\theta)}{d\theta^2}$).
- For $-2 \cdot \ln \mathcal{L}(\theta) \to -2 \cdot \Delta \ln \mathcal{L}(\theta) = 1$.

When $\mathcal{L}(\theta)$ is ~Gaussian

⇒ confidence interval of ~68% for θ.

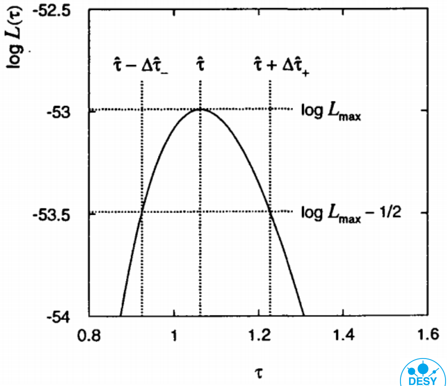
- could be asymmetric.

If $\mathcal{L}(\theta)$ "very" non-Gaussian

⇒ revert to Neyman confidence interval (will not cover).

Goodness of fit

- Not straightforward, at large N $-2 \cdot \Delta \ln \mathcal{L}(\theta) \to \chi^2$ (Wilks' theorem).
- Toy Monte Carlo.



Quick example - MLE

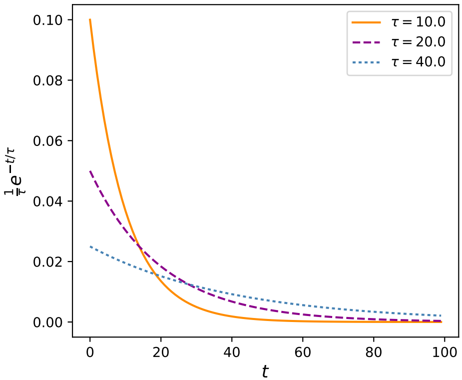
Lifetime determination (L. Lyons)

- Radioactive decay, $\frac{dn}{dt} = \frac{1}{\tau} e^{-\frac{t}{\tau}}$; (normalization, $\frac{1}{\tau}$).
- Observed decays $t_i = t_1, t_2, \ldots, t_N$.
  - neglecting background, time smearing, etc.

Construct likelihood

- $\mathcal{L}(\tau) = \prod_i^N \left(\frac{dn}{dt}\right)_i = \prod_i^N \frac{1}{\tau} e^{-\frac{t_i}{\tau}}$.
- $\ln \mathcal{L}(\tau) = \sum_i^N \left(-\frac{t_i}{\tau} - \ln \tau\right)$.
- $\frac{d \ln \mathcal{L}(\tau)}{d\tau} = \sum_i^N \left(\frac{t_i}{\tau^2} - \frac{1}{\tau}\right) = 0$.

$\Rightarrow$ $\tau = \sum_i^N \frac{t_i}{N} = \langle t_i \rangle$.



Uncertainty estimation

- $\frac{d^2 \ln \mathcal{L}(\tau)}{d\tau^2} = -\sum_i^N \left(\frac{2t_i}{\tau^3} + \frac{1}{\tau^2}\right) = 0$.

$\Rightarrow$ $\sigma_\tau = \frac{\tau}{\sqrt{N}}$ (notice $\frac{1}{\sqrt{N}}$ dependency).

Hypothesis testing

Use likelihood for hypothesis testing, often formulated as

- Null hypothesis, H_0, (e.g., Standard Model only).
- Alternative hypothesis, H_1 (e.g., Standard Model + new physics).

Simple hypothesis

Calculate $\mathcal{L}(H_0(\theta))$

- decide if data is likely for H_0 (p-value).
- If not, claim discovery (of what?)

→ Existence of a particle (Higgs, new particle)

→ A new γ-ray source.

Composite hypothesis

compare $\mathcal{L}(H_0(\theta))$ and $\mathcal{L}(H_1(\theta))$.

- Usually likelihood ratio is used.
- More sensitive to H_1.
- Based on p-values, which H_i is more likely.

→ Particle with certain mass, width, coupling constants.

→ Position and spectra of γ-ray source.

Hypothesis testing - exclude H_0

Types of errors:

> False positive (Type-1 error): wrongly reject H_0 (no new physics).

> False negative (Type-2 error): wrongly accept H_0 (missed new physics in data).

		True State of Nature	
		H_0 is true	H_0 is false
Our Decision	Do not reject H_0	Correct decision	Type II error
	Reject H_0	Type I error	Correct decision

Define the probabilities

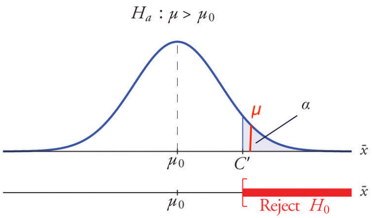
> Type-1 error rate (significance α)
$\alpha = \int_{x \geq x_0} p(x|H_0)dx$
i.e., probability of the data given H_0 (familiar?).

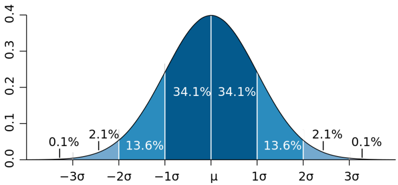
→ relation to p-value in next slides.

Hypothesis testing - exclude H_0

Gaussian example

- Assume PDF is Gaussian distributed around μ_0 and we measure μ.
- The *p*-value is the probability to measure μ or higher.
- α is the probability to measure C' or higher.
- Compare *p*-value to α, decide to accept/exclude H_0.

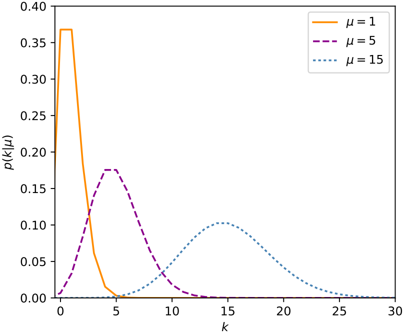


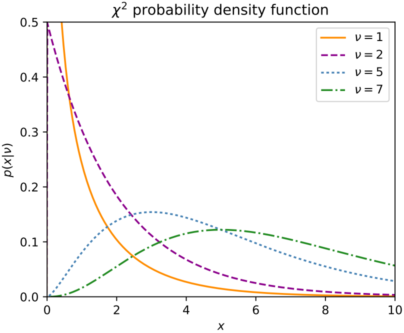


- One-sided *p*-value (or α) at $5\sigma = 3 \cdot 10^{-7}$.
- Sometimes both tails need to be taken into account ($\alpha/2$).
- * See relation to goodness of fit?

p-value

p-values not only for Gaussian distributions





$P(\mu = 5, n \geq 13) = 0.001$ $P(\nu = 5, x \geq 20.5) = 0.001$

Convention

- Convert p-value from any PDF to equivalent one-sided Gaussian σ.
- Does not mean PDF is Gaussian, simply easier to remember.
- p-value is $P(\text{data}|H_0)$, it is **not** $P(H_0|\text{data})$.

Hypothesis testing

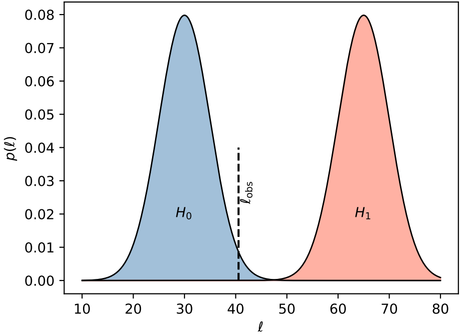
For comparing H_0 & H_1, **Neyman-Pearson Lemma**
$\Rightarrow$ Likelihood ratio test is optimal discriminant (assuming no free parameters).

> Log Likelihood ratio $\ell = -2\ln\left(\frac{\mathcal{L}(\text{data}, H_1)}{\mathcal{L}(\text{data}, H_0)}\right)$.

> If $H_i(\theta)$, use simulation to generate distributions of ℓ for H_i.

> Take measurement and calculate ℓ after maximizing $\mathcal{L}$ for both $H_i(\theta)$.

> Calculate both p-values, decide which H_i to accept.



* More complicated in reality.

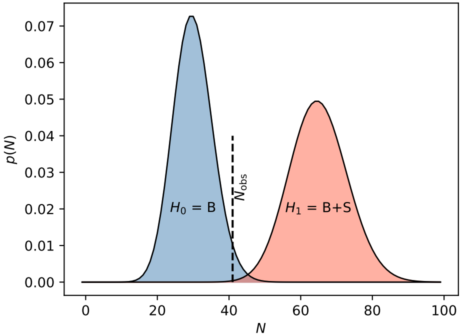
Exclusion limit (simplified)

Assume H_0 = background (SM) and H_1 = background + signal.

- Number of events (cross-section) observed is Poisson distributed.
- From *p*-values, accept H_0.

Set limit

- Find the maximum signal strength for which $p(H_1) < 5\%$.
- Set limit on signal at 95% confidence level (exclusion, 2σ).



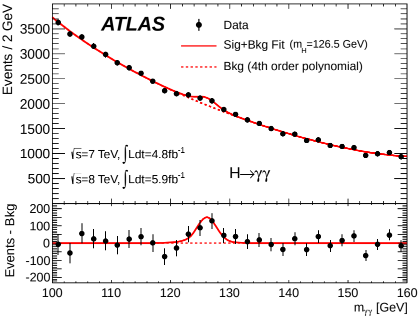
* Usually based on the likelihood ratio test statistics.

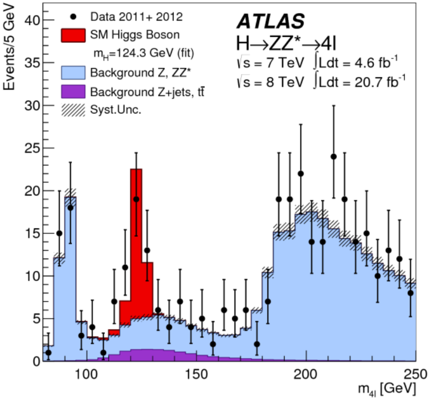
Higgs discovery

> Visible bump in the data.

> For $H \to \gamma\gamma$, background fitted with a smooth distribution.

> Complicated background in $H \to 4\ell$.



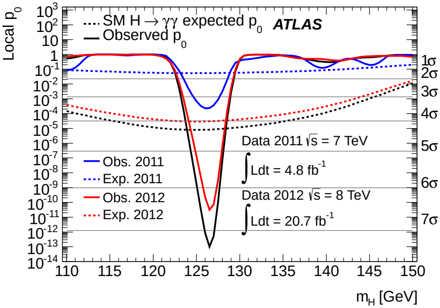


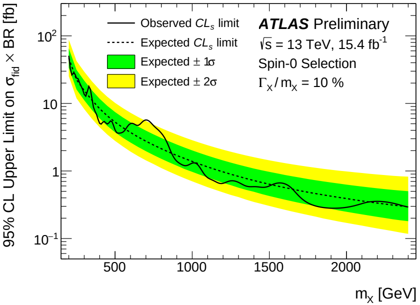
Discovery/Exclusion

> **local** p-value of observed Higgs signal (signal stronger than expected).

> Search for massive scalar decaying to two γ, not found

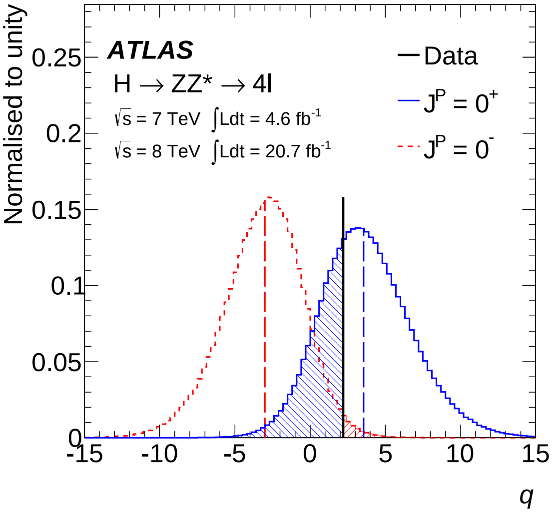
$\Rightarrow$ set an upper limit on the cross-section ($\times$ branching ratio).





Higgs spin

> Use likelihood ratio $q = -2\ln\left(\frac{\mathcal{L}(\text{data},0^+)}{\mathcal{L}(\text{data},0^-)}\right)$ to determine Higgs spin.



Look elsewhere effect

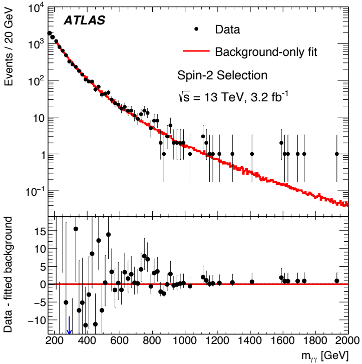
Bump hunting? Peaks can be anywhere!

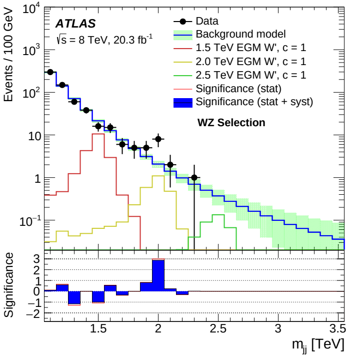
> Increase *p*-value to take into account (quote local and global *p*-value).

> Correction roughly width mass interval divided by width particle.

> Confirmation from other experiment is crucial.

* Consider amount of searches at any given time.





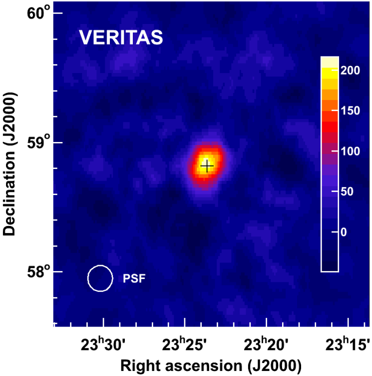
* Remember the track with 110 bubbles?

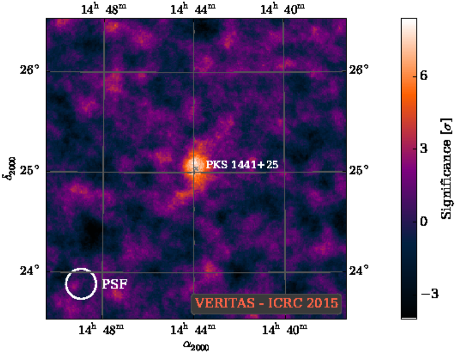
Look elsewhere effect

Also in searches for γ-ray sources

* Usually referred to as trials factor.

> Include also cuts in the correction (not always easy).





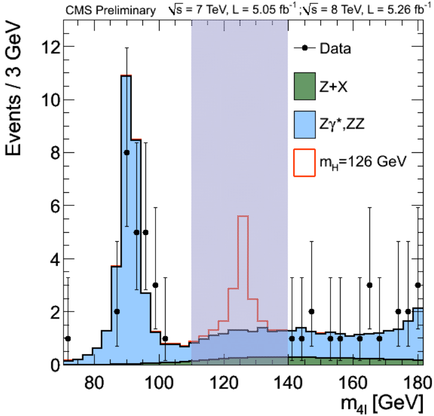
Blind analysis

R. Feynman

The first principle is that you must not fool yourself — and you are the easiest person to fool

Whenever possible, perform blind analysis

- Keep the "signal box" closed.
- Construct and refine analysis on simulation, cannot change after unblinding.
- Use only part of the data available.



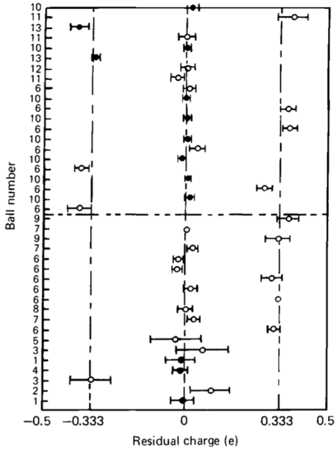
Blind analysis

R. Feynman

The first principle is that you must not fool yourself — and you are the easiest person to fool

Whenever possible, perform blind analysis

- Keep the “signal box” closed.
- Construct and refine analysis on simulation, cannot change after unblinding.
- Use only part of the data available.
- Add random numbers to results.
- Use fake signal to test procedure (done at LIGO).



The 5σ criteria

Probability of fluctuation of 5σ is less than 1 in a million, tiny!

> This was not always the case (and is not in other fields).

> A lot more data these days.

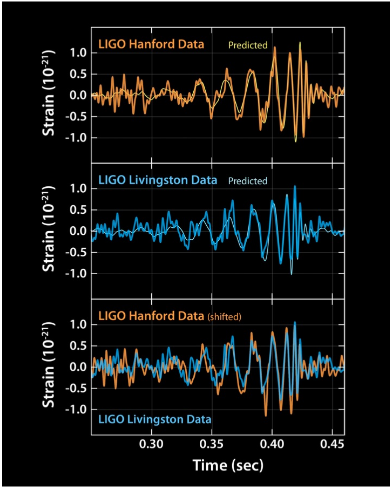
> Sometimes hard to estimate look elsewhere effect.

> Underestimated systematic uncertainties?

> A discovery of new physics will be a game changer, better not take it back.

Bayesian prior

extraordinary claims require extraordinary evidence



Monte Carlo methods

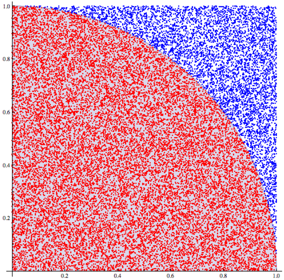
Monte Carlo methods

Wikipedia: "computational algorithms that rely on repeated random sampling to obtain numerical results."

Useful for, e.g.,

- Numerical integration.
- Simulating particle interactions or decay.
- Uncertainty estimation.

Example: estimate π




```python
from random import random
from math import sqrt, pi
inside=0
n=10000000
i_print = [1, 10, 100, 1000, 10000, 100000, 1000000]
for i in range(0,n+1):
    x=random()
    y=random()
    if sqrt(x*x+y*y)<=1:
        inside+=1
    if i in i_print:
        piNow=4*inside/i
        print ('pi(i=%d) = %.4f, error = %.4f' % (i, piNow, abs(piNow - pi)))
```

```
pi(i=1) = 4.0000, error = 0.8584
pi(i=10) = 3.6000, error = 0.4584
pi(i=100) = 3.3600, error = 0.2184
pi(i=1000) = 3.1240, error = 0.0176
pi(i=10000) = 3.1264, error = 0.0152
pi(i=100000) = 3.1433, error = 0.0017
pi(i=1000000) = 3.1402, error = 0.0014
```

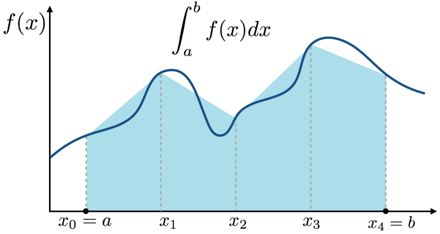
Monte Carlo integration

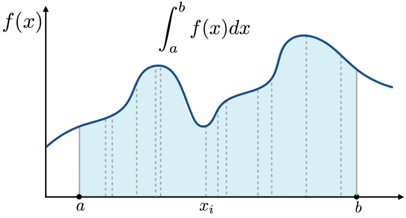
Simple numerical integration

- Divide range to small pieces of known area and sum.
- Suffers from curse of dimensionality, $N_{\text{calc}} = n^d$.

Similar to π estimate example, can sample function at random points.

- Avoids curse of dimensionality of numerical integration, error $\propto 1/\sqrt{N}$.
- Works for any function (including discontinuous ones).
- Faster at large d.
- Used in e.g., phase-space integration of matrix elements.



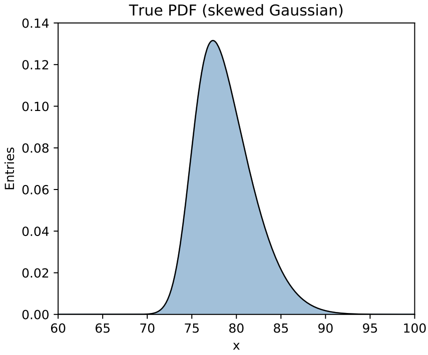


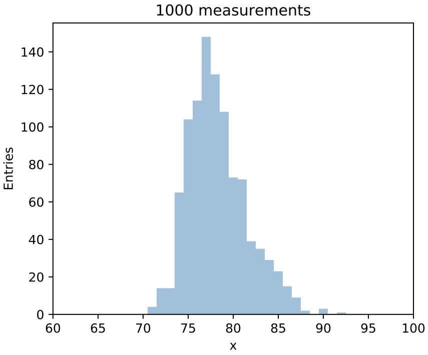
Bootstrap method

Assume N measurements of x, x_i, how to estimate $\mu_x \pm \sigma_{\mu_x}$?
Not easy to estimate σ_{μ_x} without knowing PDF of x.

> Usually there is no access to the “true” PDF.

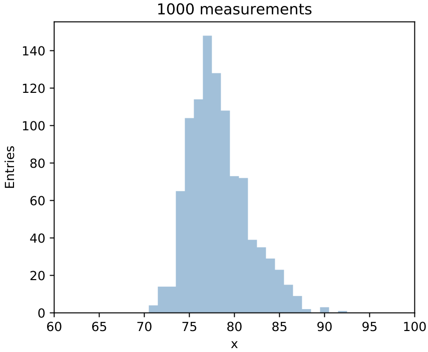
$\Rightarrow$ The distribution of x_i is best approximation of it.

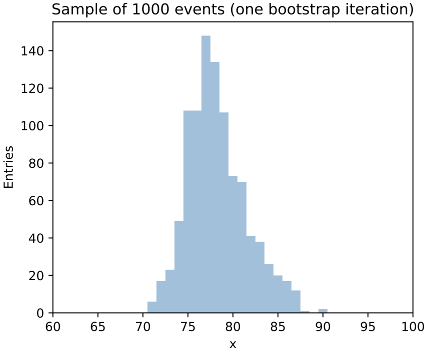




Bootstrap method

- Can generate "new" measurements by sampling from x_i.
- Each iteration sample events from the 1000 measurements, allowing repetition.
- Calculate μ_x for "new" distribution.



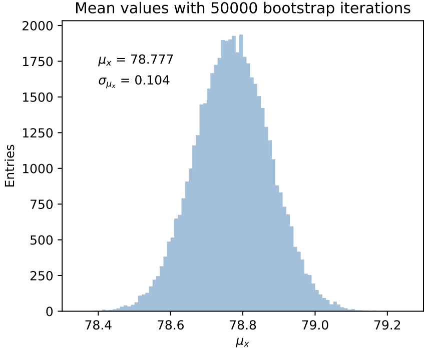


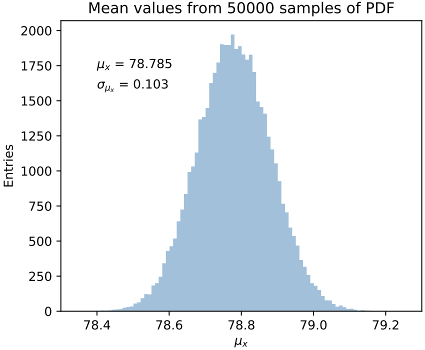
Bootstrap method

> Obtain a distribution of μ_x by repeating 50,000 times.

> For comparison, perform 50,000 experiments using true PDF, each with 1000 events.

⇒ σ_{μ_x} from bootstrap reproduces that from independent experiments.



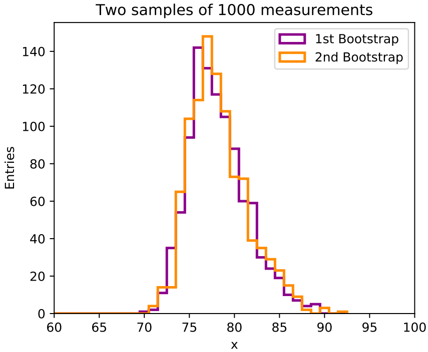


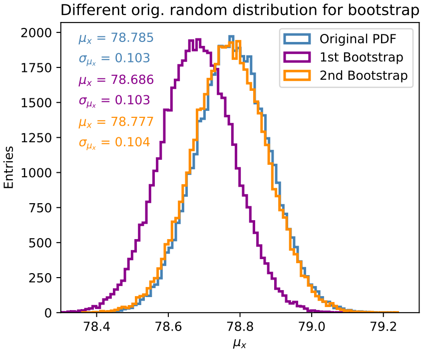
* Toy MC
* Can be used in numerical estimation of uncertainties

Bootstrap method

While making the plots, encountered interesting case,

- In 1st trial, saw a bias between μ_x distributions (not significant, but still).
- Spread was OK $\rightarrow$ test by changing random seed when performing first "measurement" of 1000 events.
- With different random seed, no more bias and same σ_{μ_x}.

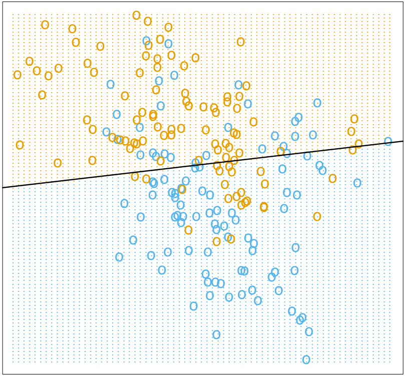


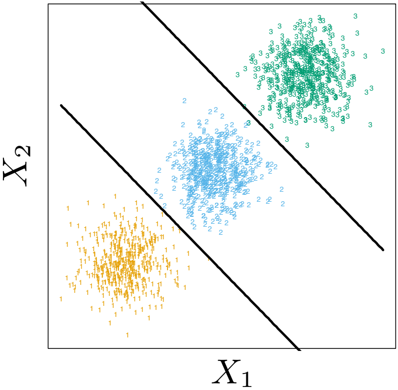


Event classification

Significant part of data analysis, classifying between events.

- Define decision boundaries (cuts).
- Requires prior information (usually from simulation).
- Can it be done "visually"? Usually too many observables/classes.

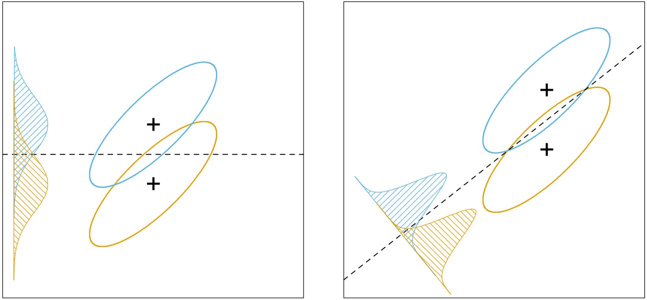




Event classification

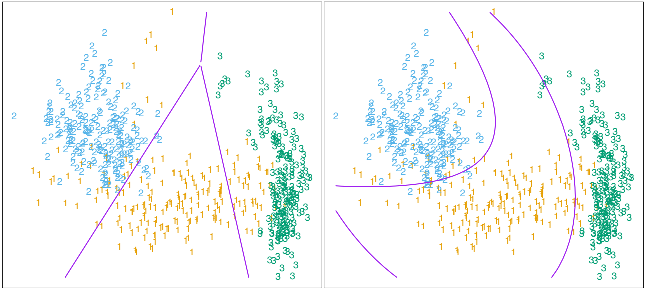
Various ways to deal with it (see 3rd lecture about machine learning)

- > If observables not highly correlated, can define cuts in bins.
- > Many algorithms available to optimize cuts (e.g., linear discriminant analysis, kNN, SVM, BDT, ANN).
- → LDA is similar to Principal Component Analysis, but it maximizes separability between event classes.
- → Maximizes distance between means and minimizes overlap.
- → Can be used to reduce dimensionality and optimize cut hyperplane.



Event classification

- Use data with known labels to train discriminant, apply later to "real" data.
- Can expand space to $5d = X_1, X_2, X_1 \cdot X_2, X_1^2, X_2^2$ to obtain non-linear hyperplanes using linear method.



- Transformed observables useful also for non-linear methods (e.g., $\log E$).
- Take care when using "automatic" classifiers, study results carefully.
- Consider systematic uncertainties when selecting observables.

Miscellaneous

Unfolding

What is unfolding?

The process of correcting the data for detector effects

A measured distribution is affected by

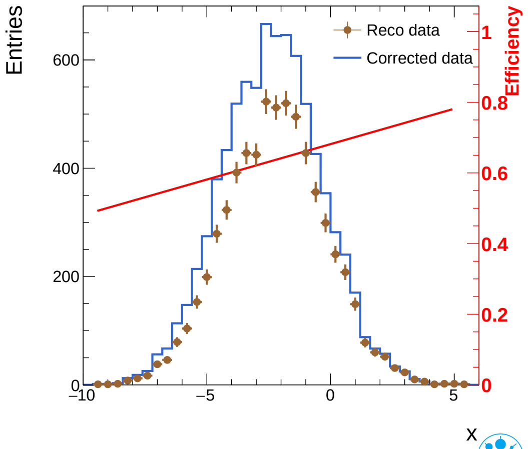
- Inefficiencies in the detector $\rightarrow$ lost events.
- Bias $\rightarrow$ if $\langle x \rangle$ is true mean, measure $\langle x' \rangle = \langle x \rangle + \Delta x$.
- Smearing $\rightarrow$ the detector has finite resolution.

Simple example,

- known efficiency function.
- no bias or smearing.

$\Rightarrow$ correct each bin for fractional loss of events.

* Not really unfolding



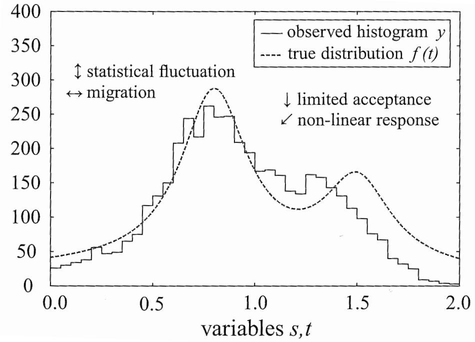
Unfolding

In practice, given a measured histogram y^{data},

- want to obtain “true” distribution x^{data}, where $y^{\text{data}} = R^{\text{data}} \cdot x^{\text{data}}$.
- The matrix R^{data}_{ij} is the response function of the detector.
- Inefficiencies contribute to diagonal elements → per-bin correction;
- bias and smearing to off-diagonal → bin migration.

How to derive R^{data}?

- In simulation we have all necessary information.
- $y^{\text{MC}} = R^{\text{MC}} \cdot x^{\text{MC}}$, where R^{MC} is our detector simulation.
- Assume $R^{\text{data}} = R^{\text{MC}} = R$.
- Notice that in general,
  $y^{\text{data}} \neq y^{\text{MC}}$,
  $x^{\text{data}} \neq x^{\text{MC}}$,
  but should be close.
- Can we then simply use $x^{\text{data}} = R^{-1} \cdot y^{\text{data}}$?



Unfolding

Unfolding is an ill-posed problem
⇒ With finite statistics,
naive unfolding fails.

> Leads to significant statistical fluctuations between bins.

→ Negative correlation coefficients between adjacent bins.

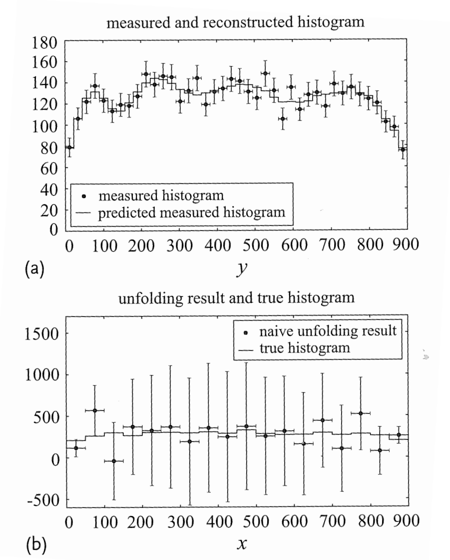
→ Positive coefficients between next-to-nearest neighbours.

How to deal with fluctuations?
Regularization

> Increase weight of “smoother” solutions, damp oscillations.

> Unfold iteratively using Bayes theorem (will not cover).

∗ Various tools available, e.g., `RooUnfold`.



Unfolding

Regularized unfolding

The unfolding problem can be written as a minimization of (simplified)

$$\chi^2(x^{\text{data}}) = (R \cdot x^{\text{data}} - y^{\text{data}})^{\mathsf{T}}(R \cdot x^{\text{data}} - y^{\text{data}}) + \tau(Lx^{\text{data}})^{\mathsf{T}}(Lx^{\text{data}})$$

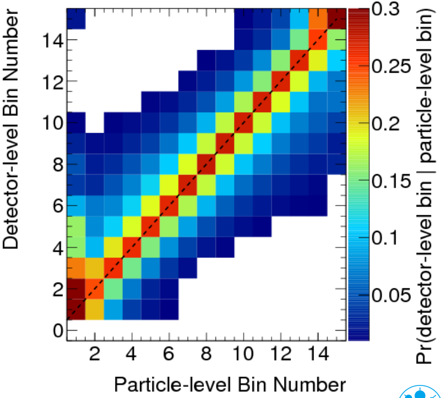
L is regularization matrix (second derivative commonly used).
Second term dampens oscillations.

> τ is regularization parameter,
  * if τ is too small $\rightarrow$ oscillations;
  * if τ is too large $\rightarrow$ x^{data} too smooth and biased towards x^{MC};

> Depends on number of events and binning.

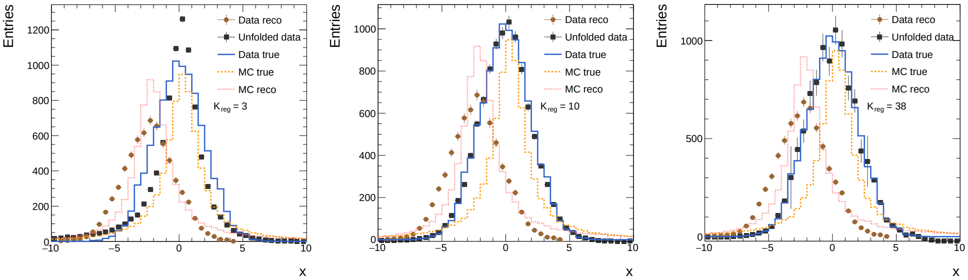
> Some trial & error to choose τ.

> Usually chosen using (independent) MC samples.



Unfolding

Over-regularized **Correctly-regularized** **Under-regularized**



- Notice, $y^{\text{data}} \neq y^{\text{MC}}$, $x^{\text{data}} \neq x^{\text{MC}}$

Why do we bother?

- Allows to compare directly to theoretical models and among experiments.
- "Future proof" the data.

Unfolding vs folding

Folding (or forward folding) is another option

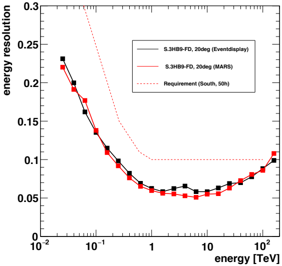
- Instead of correcting data, publish it with corresponding R.
- The problem is then technically simpler,

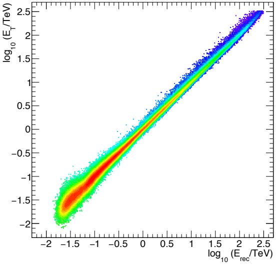
$$\chi^2(x^{\text{theo}}(\theta)) = (R \cdot x^{\text{theo}}(\theta) - y^{\text{data}})^{\mathsf{T}}(R \cdot x^{\text{theo}}(\theta) - y^{\text{data}}).$$

  in the case where $x^{\text{theo}}(\theta)$ is the model one wants to test.
- Avoids unfolding issues (ill-defined problem, converting statistical uncertainties to systematic ones).

Issues with folding

- Does not allow comparison between experiments.
- Harder to test your model against data from various experiments.





What to do?

When possible, unfold.

Extended MLE

The production rate depends on mass of a particle, need to estimate both?

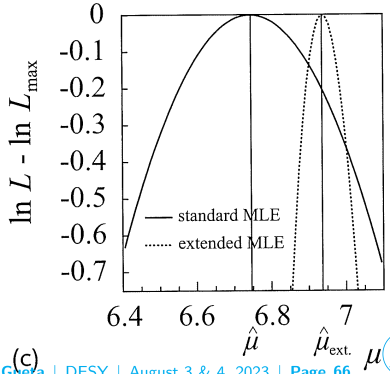
- Extended MLE (Poisson process, PDF $f(x_i;\theta)$)

$$L(x;\nu,\theta) = \frac{\nu^N}{N!}e^{-\nu}\prod_i^N f(x_i;\theta)$$

  maximize with respect to both θ and ν (profile likelihood).
- Improved precision of fitted parameters obtained if θ and ν are correlated (e.g., θ = particle mass).

Nice example in Data Analysis in High Energy Physics book

- PDF, $f(x_i;\mu) = \mathcal{G}(\mu)$.
- Parameter of interest is μ.
- Assume $\nu = 9e^{-4(\mu-7)}$.
- Simulate events with $\mu_{\text{true}} = 7$.
- Perform profile likelihood to obtain more precise $\hat{\mu}$.



(c)

Extended MLE and nuisance parameters

Can be used to include uncertainties in likelihood fit

- Assume signal and background contributions S and B.
- Try to estimate S, include Gaussian uncertainty on background $B \to \theta B$,

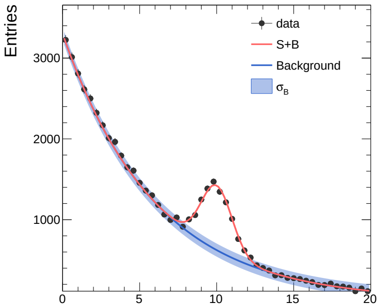
$$\mathcal{L}(N; S, \theta) = \frac{(S+\theta B)^N}{N!} e^{-(S+\theta B)} \mathcal{G}(\theta - 1, \sigma_\theta)$$

- Background is constrained to our best guess ($\theta = 1$), with a σ_θ spread.
- Maximize $\mathcal{L}$ to estimate S while marginalizing θ.

In reality can become complex

- Estimate various parameters of S and B simultaneously (e.g., particle mass).
- S and B affected by various uncertainties (many nuisance parameters).
- Divide data to various regions where different uncertainties contribute.

* Use tools to build models and perform fit, e.g., `RooFit`, `ctools`, `Gammapy`.



Summary

Statistics is everywhere in physics

- Lectures can get a bit abstract $\rightarrow$ learn by doing.
- Likely that your problem was solved already somewhere else, consult books and the web before reinventing the wheel.
- Use software packages as much as possible (`ROOT`, `RooFit`, various Python tools, etc.)

Subjects not covered but worth reading about

- Confidence intervals, coverage and limit setting.
- Dealing with systematic and theory uncertainties.
- Estimating contributions through templates and control regions.
- Combining results.
- Many more.

e-mail: orel.gueta@desy.de