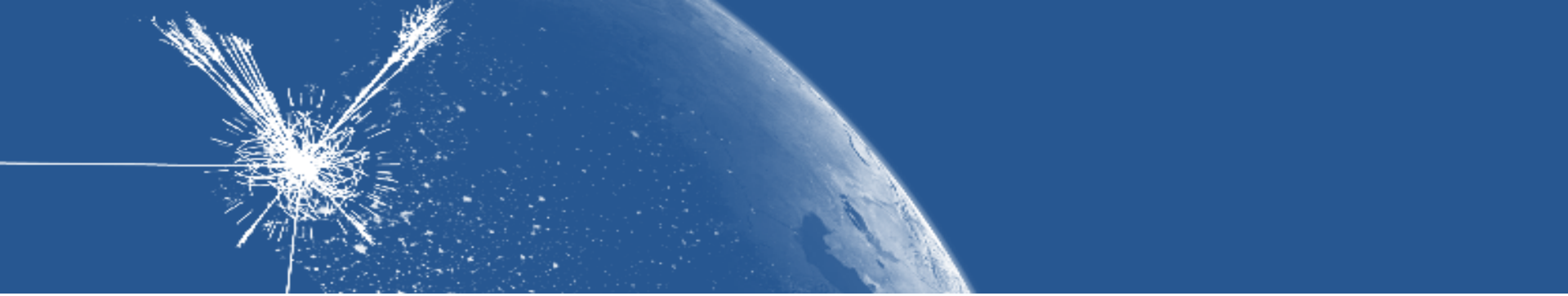
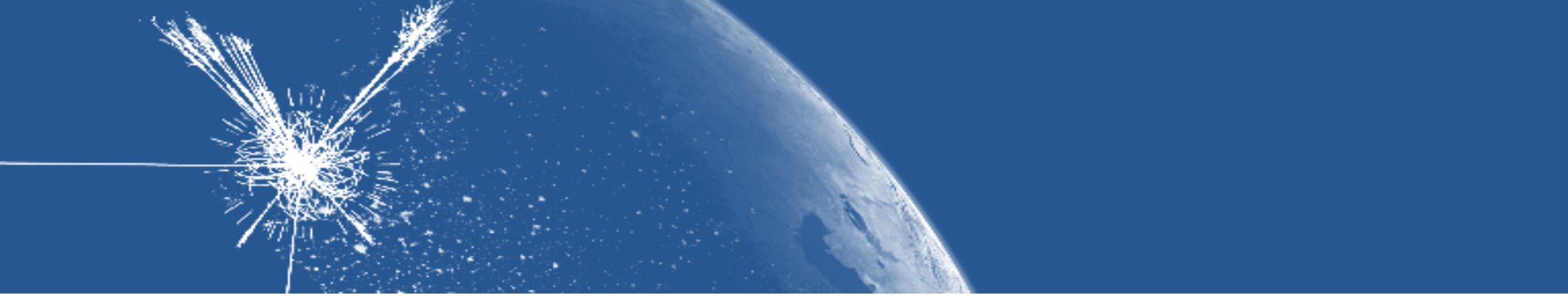


Introduction to probability and statistics (2)

Andreas Hoecker (CERN)

CERN Summer Student Lecture, 17–21 July 2017

If you have questions, please do not hesitate to contact me: **andreas.hoecker@cern.ch**



Outline (4 lectures)

1st lecture:

- Introduction
- Probability (…some catch-up to do)

2nd lecture:

- Probability axioms and hypothesis testing
- Parameter estimation
- Confidence levels

3rd lecture:

- Maximum likelihood fits
- Monte Carlo methods
- Data unfolding

4th lecture:

- Multivariate techniques and machine learning

Catch-up from yesterday

Multidimensional random variables

What if a measurement consists of two variables?

Let:

$\boldsymbol{A}$ = measurement $\boldsymbol{x}$ in $[\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{dx}]$

$\boldsymbol{B}$ = measurement $\boldsymbol{y}$ in $[\boldsymbol{y}, \boldsymbol{y} + \boldsymbol{dy}]$

Joint probability: $P(A \cap B) = p_{xy}(x, y)dxdy$

(where $p_{xy}(x, y)$ is joint PDF)

If the two variables are independent:

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

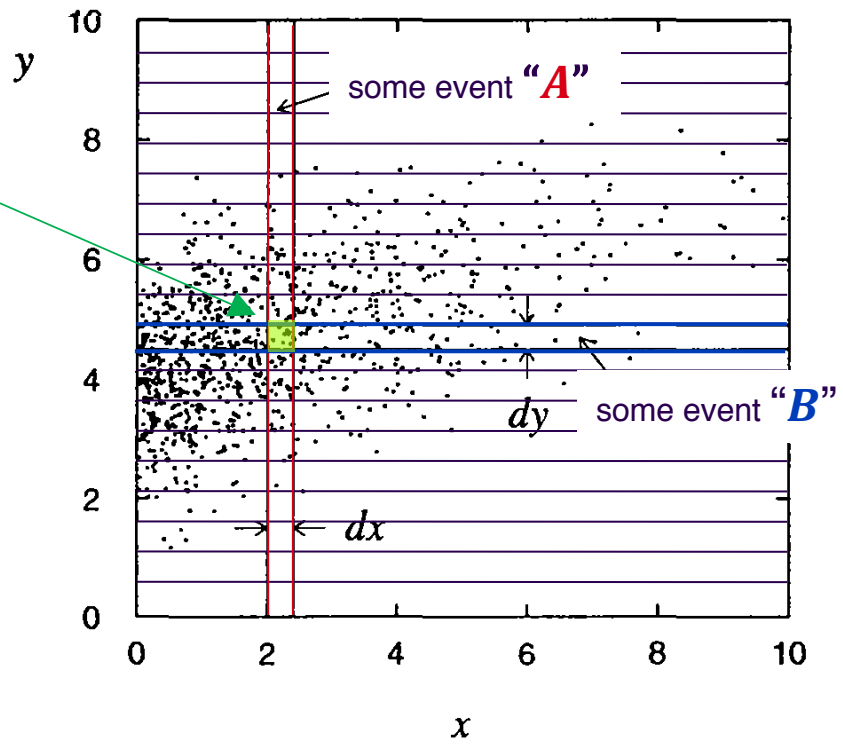
$$p_{xy}(x, y) = p_x(x) \cdot p_y\,(y)$$

Marginal PDF: if one is not interested in dependence on $\boldsymbol{y}$ (or cannot measure it),

→ integrate out ("marginalise") $\boldsymbol{y}$, ie, project onto $\boldsymbol{x}$

→ resulting one-dimensional PDF: $p_x(x) = \int p_{xy}(x, y)dy$

From: Glen Cowan, Statistical data analysis



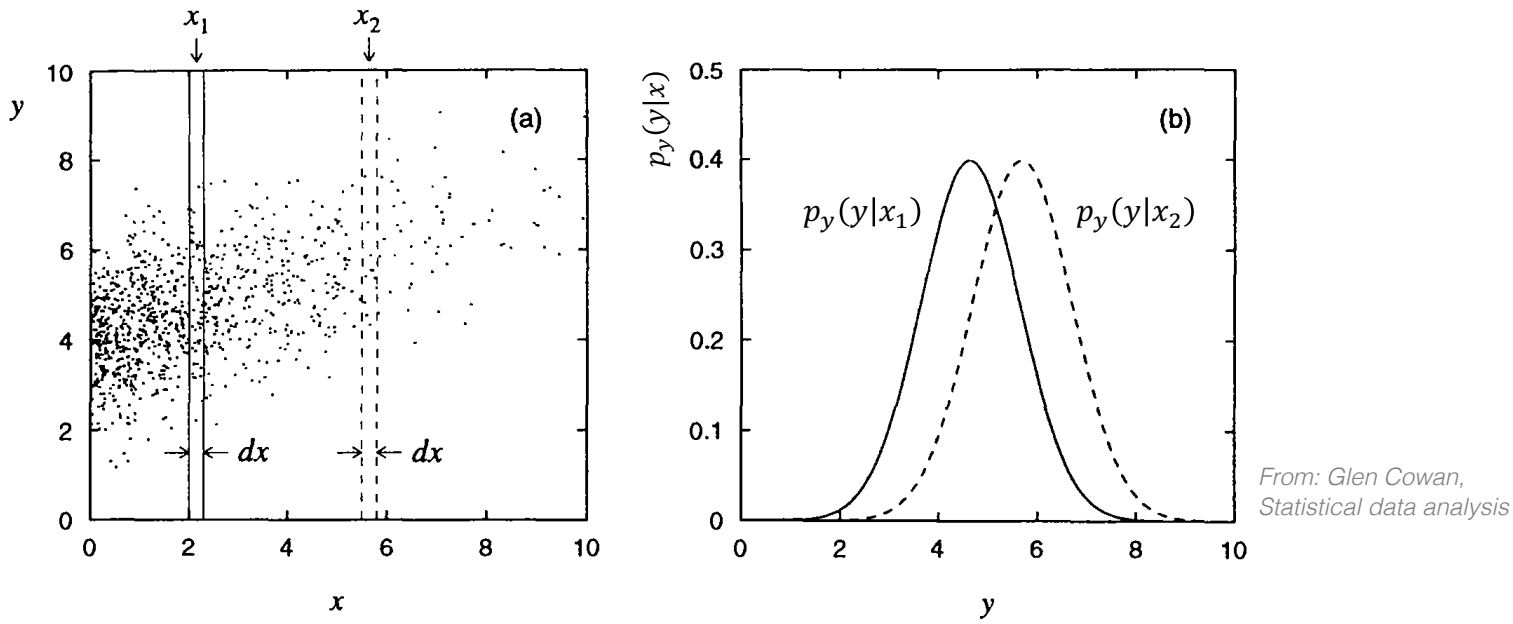
Conditioning versus marginalisation

Conditional probability $\boldsymbol{P(A|B)}$: [read: $P(A|B)$ = "probability of A *given* B"]

$$P(A \cap B) = P(A|B) \cdot P(B) \qquad \Leftrightarrow \qquad \boldsymbol{P(A|B)} = \frac{P(A \cap B)}{P(B)} = \frac{p_{xy}(x,y)dxdy}{p_y(y)dx}$$

Rather than integrating over the whole y region (marginalisation),
look at one-dimensional (1D) slices of the two-dimensional (2D) PDF $p_{xy}(x,y)$:

$$p_y(y|x_1) = p_{xy}(x = \text{const} = x_1, y)$$



From: Glen Cowan, Statistical data analysis

Covariance and correlation

Recall, for 1D PDF $\boldsymbol{p_x(x)}$ we had: $E[x] = \mu_x;\ \ V[x] = \sigma_x^2$

For a 2D PDF $\boldsymbol{p_{xy}(x,y)}$, one correspondingly has: μ_x, μ_y, σ_x, σ_y

How do $\boldsymbol{x}$ and $\boldsymbol{y}$ co-vary ? $\rightarrow$ $\mathrm{C}_{xy} = \mathbf{covariance}_{xy} = E\big[(x-\mu_x)\big(y-\mu_y\big)\big] = E[xy] - \mu_x\mu_y$

From this define the scale / dimension invariant *correlation coefficient*:

$$\rho_{xy} = \frac{\mathrm{C}_{xy}}{\sigma_x\sigma_y}\,,$$ where $\rho_{xy} \subset [-1,+1]$

- If x, y are independent: $\rho_{xy} = 0$, ie, they are *uncorrelated* (or they *factorise*)

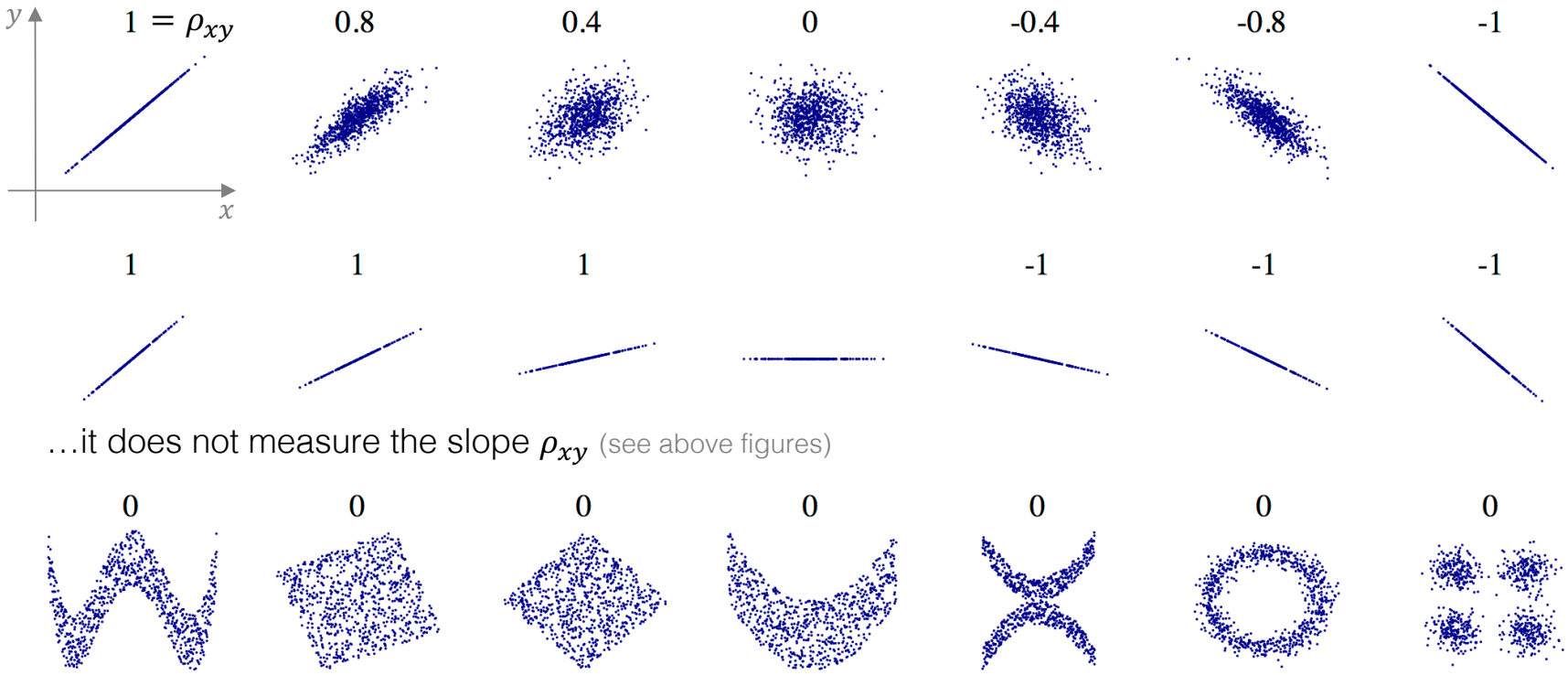
  Proof: $E[xy] = \iint xy \cdot p_{xy}\,(x,y)dxdy = \iint xy \cdot p_x(x)p_y(y)dxdy = \int x \cdot p_x(x)dx \cdot \int y \cdot p_y(y)dy = \mu_x\mu_y$

- Note that the contrary is not always true: non-linear correlations can lead to $\rho_{xy} = 0$,
  $\rightarrow$ see next page

Correlations

Figure from: https://en.wikipedia.org/wiki/Correlation_and_dependence

The correlation coefficient measures the noisiness and direction of a linear relationship:



…it does not measure the slope ρ_{xy} (see above figures)

…and non-linear correlation patterns are not or only approximately captured by ρ_{xy} (see above figures)

Correlations

Non-linear correlation can be captured by the "*mutual information*" quantity $\boldsymbol{I_{xy}}$:

$$I_{xy} = \iint p_{xy}(x,y) \cdot \ln\left(\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}\right) dxdy$$

Measure of mutual dependence between two variables:
"How much information is shared among them"

where I_{xy} =0 only if $\boldsymbol{x, y}$ are fully statistically independent

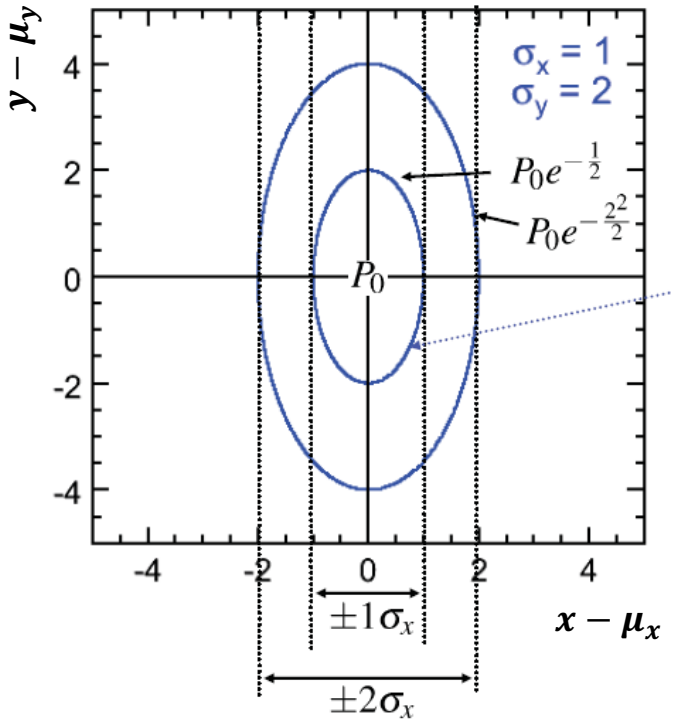
Proof: if independent, then $p_{xy}(x,y) = p_x(x)p_y(y) \Rightarrow \ln(\dots) = 0$

NB: $I_{xy} = H_x - H_x(y) = H_y - H_y(x)$,
where $H_x = -\int p_x(x) \cdot \ln\big(p_x(x)\big)\, dx$ is *entropy*, $H_x(y)$ is *conditional entropy*

2D Gaussian (uncorrelated)

Two variable $\boldsymbol{x}, \boldsymbol{y}$ are independent: [$p_{xy}(x,y) = p_x(x) \cdot p_y(y)$]

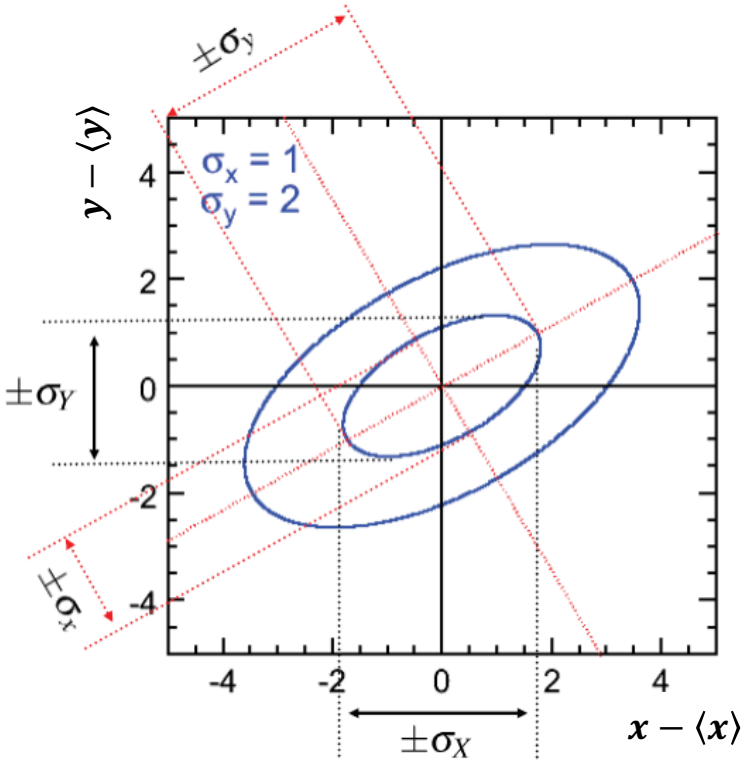
$$p_{xy}(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$$



2D Gaussian (correlated)

Two variable $\boldsymbol{x}, \boldsymbol{y}$ are *not* independent: [$p_{xy}(x,y) \neq p_x(x) \cdot p_y(y)$]

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



where:

$$C = \begin{pmatrix} \langle x^2\rangle - \langle x\rangle^2 & \langle xy\rangle - \langle x\rangle\langle y\rangle \\ \langle xy\rangle - \langle x\rangle\langle y\rangle & \langle y^2\rangle - \langle y\rangle^2 \end{pmatrix}$$

is the (symmetric) *covariance matrix*

Corresponding correlation matrix elements:

$$\rho_{ij} = \rho_{ji} = \frac{\mathrm{C}_{ij}}{\sqrt{\mathrm{C}_{ii} \cdot \mathrm{C}_{jj}}}$$

SQRT decorrelation

Find variable transformation that diagonalises a covariance matrix C

Determine "square-root" C' of C (such that: $C = C' \cdot C'$) by first diagonalising C

$$D = S^T \cdot C \cdot S \quad \Leftrightarrow \quad C' = S \cdot \sqrt{D} \cdot S^T$$

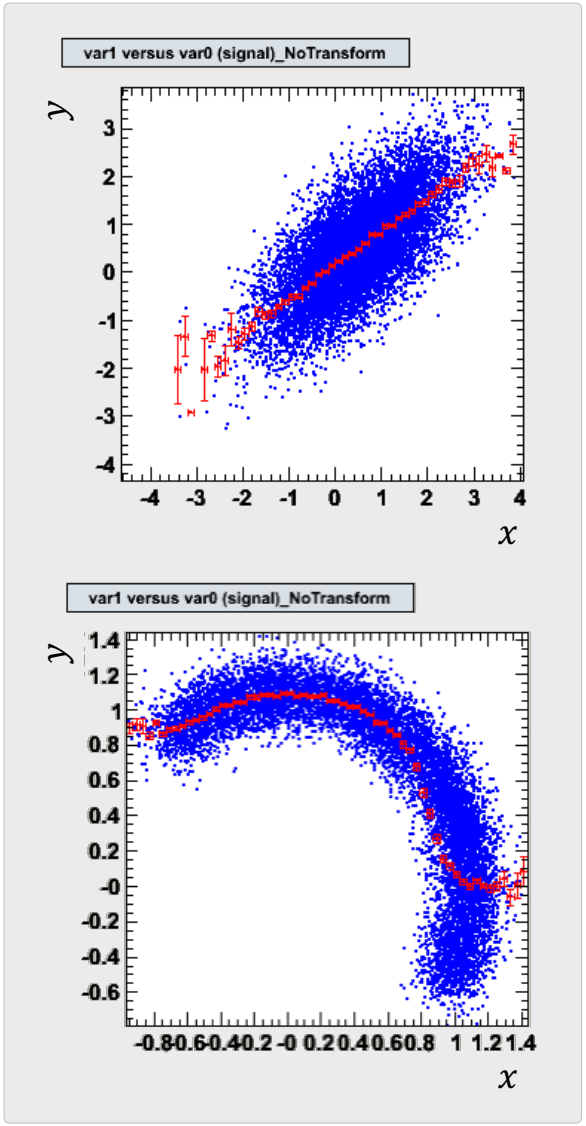
where D is diagonal, $\sqrt{D} = \{\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}\}$, and S an orthogonal matrix

Linear decorrelation of correlated vector $\boldsymbol{x}$ then obtained by

$$\boldsymbol{x}' = (C')^{-1} \cdot \boldsymbol{x}$$

Principle component analysis (PCA) is another convenient method to achieve linear decorrelation

(PCA is linear transformation that rotates a vector such that the maximum variability is visible. It identifies most important gradients)



Example: original correlations

SQRT decorrelation

Find variable transformation that diagonalises a covariance matrix C

Determine "square-root" C' of C (such that: $C = C' \cdot C'$) by first diagonalising C

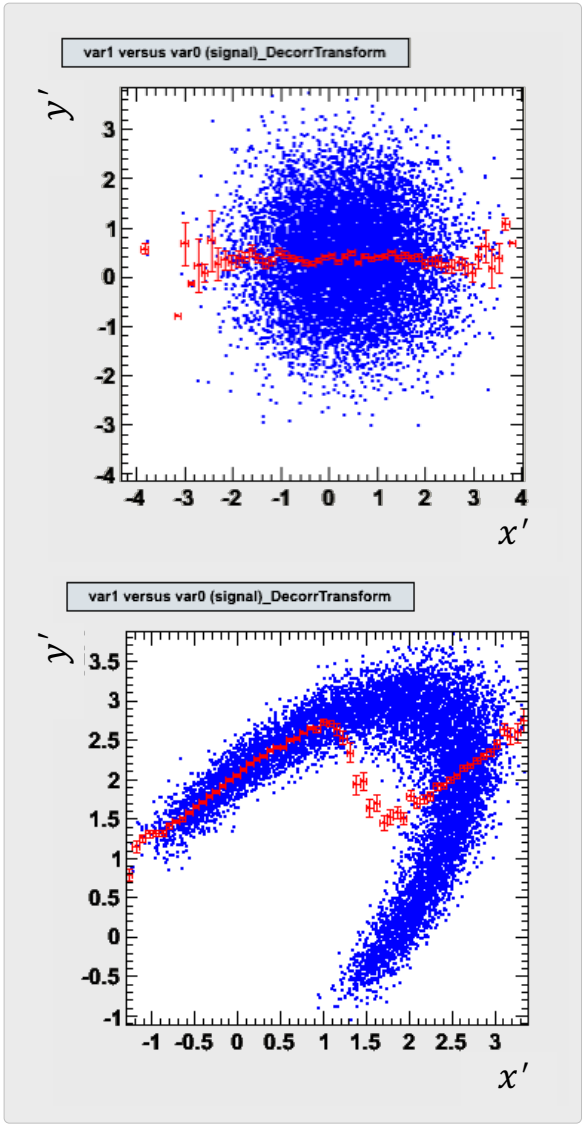
$$D = S^T \cdot C \cdot S \quad \Leftrightarrow \quad C' = S \cdot \sqrt{D} \cdot S^T$$

where D is diagonal, $\sqrt{D} = \{\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}\}$, and S an orthogonal matrix

Linear decorrelation of correlated vector $\boldsymbol{x}$ then obtained by

$$\boldsymbol{x}' = (C')^{-1} \cdot \boldsymbol{x}$$

SQRT decorrelation works only for linear correlations!



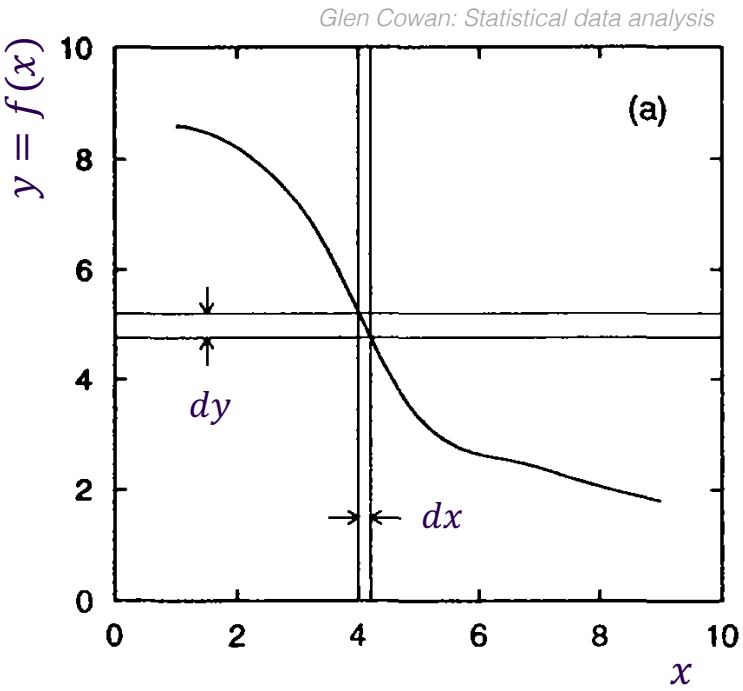
Example: after SQRT decorrelation

Functions of random variables

Any function of a random variable is itself a random variable

E.g., $\boldsymbol{x}$ with PDF $\boldsymbol{p_x(x)}$ becomes: $\boldsymbol{y = f(x)}$

$\boldsymbol{y}$ could be a parameter extracted from a measurement



What is the PDF $\boldsymbol{p_y(y)}$?

- Probability conservation: $p_y(y)|dy| = p_x(x)|dx|$
- For a 1D function $f(x)$ with existing inverse:

$$dy = \frac{df(x)}{dx}dx \quad \Leftrightarrow \quad dx = \frac{df^{-1}(y)}{dy}dy$$

- Hence: $\boldsymbol{p_y(y) = p_x(f^{-1}(y))\left|\frac{dx}{dy}\right|}$

Note: this is **not** the standard error propagation but the full PDF !

Error propagation

Let's assume a measurement $\boldsymbol{x}$ with *unknown* PDF $\boldsymbol{p_x(x)}$, and a transformation $\boldsymbol{y = f(x)}$

- $\bar{x}$ and $\hat{V}$ are estimates of μ and variance σ^2 of $p_x(x)$

What are $E[y]$ and, in particular, $\boldsymbol{\sigma_y^2}$? $\rightarrow$ Taylor-expand $f(x)$ around $\bar{x}$:

- $f(x) = f(\bar{x}) + \left.\frac{df}{dx}\right|_{x=\bar{x}} (x - \bar{x}) + \cdots \quad \Rightarrow \quad E[f(x)] \simeq f(\bar{x})$ (because: $E[x - \bar{x}] = 0$!)

Now define $\bar{y} = f(\bar{x})$, and from the above follows:

$\Leftrightarrow \quad y - \bar{y} \simeq \left.\frac{df}{dx}\right|_{x=\bar{x}} (x - \bar{x})$

$\Leftrightarrow \quad E[(y - \bar{y})^2] = \left(\left.\frac{df}{dx}\right|_{x=\bar{x}}\right)^2 E[(x - \bar{x})^2]$

$\Leftrightarrow \quad \hat{V}_y = \left(\left.\frac{df}{dx}\right|_{x=\bar{x}}\right)^2 \hat{V}_x$

$\Leftrightarrow \quad \sigma_y = \left.\frac{df}{dx}\right|_{x=\bar{x}} \cdot \sigma_x \qquad \rightarrow \qquad$ (approximate) error propagation

Error propagation (continued)

In case of several variables, compute covariance matrix and partial derivatives

- Let $\boldsymbol{f = f(x_1, \dots, x_n)}$ be a function of $\boldsymbol{n}$ randomly distributed variables
- $\left(\frac{df}{dx}\Big|_{x=\bar{x}}\right)^2 \hat{V}_x$ becomes: $\sum_{i,j=1}^{n} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}\Big|_{\bar{x}} \cdot \hat{V}_{i,j}$ (where: $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n)$)
- with the covariance matrix:

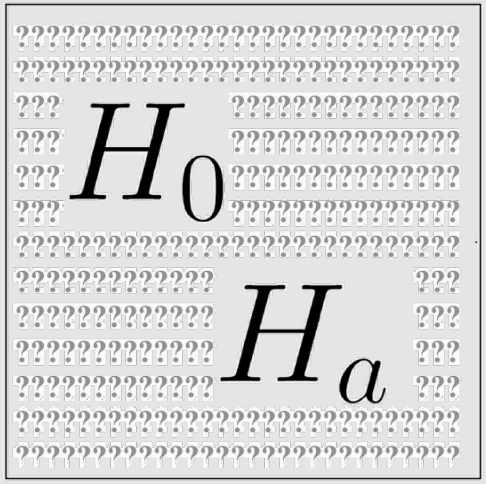
$$\hat{V}_{i,j} = \begin{bmatrix} \sigma_{x_1}^2 & \cdots & \sigma_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ \sigma_{x_n x_1} & \cdots & \sigma_{x_n}^2 \end{bmatrix}$$

→ The resulting "error" (uncertainty) depends on the correlation of the input variables

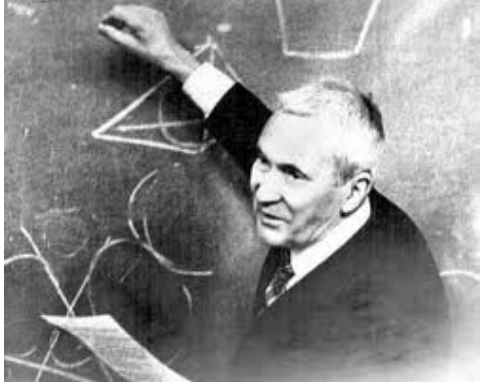
- Typically (not always:) positive correlations lead to an increase of the total error,
- and negative correlations decrease the total error

For very complicated functional dependence $\boldsymbol{f = f(x_1, \dots, x_n)}$, use Monte Carlo techniques ("pseudo MC generation") to propagate uncertainties

Probability (axioms) & Statistics



What is a *Probability*?



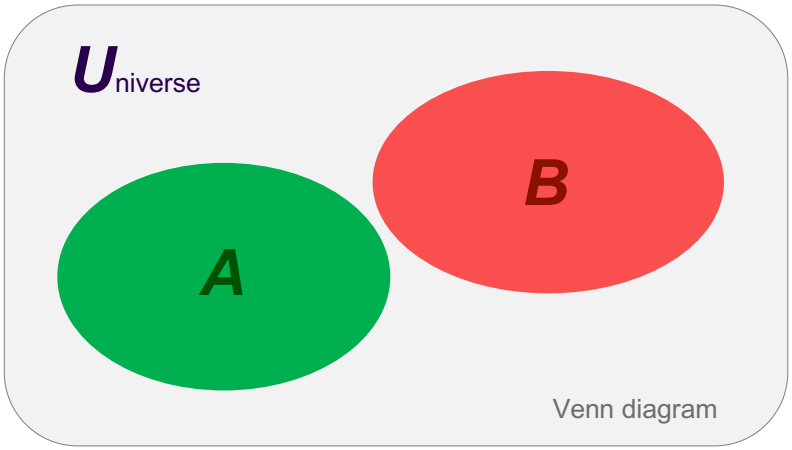
Andrey Nikolaevich Kolmogorov

Axioms of probability (Kolmogorov, 1933)

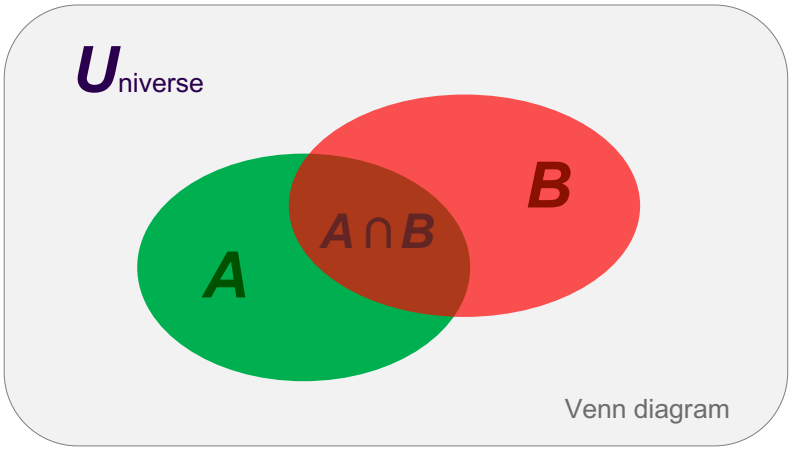
- $\boldsymbol{P(A) \geq 0}$, where $\boldsymbol{A}$ is any subset of sample space (*universe*) $\boldsymbol{U}$
- Unitarity: $\int_{\boldsymbol{U}} \boldsymbol{P(A)dA = 1}$
- If $\boldsymbol{(A \cap B) = 0}$ (read: "A *and* B") $\rightarrow \boldsymbol{P(A \cup B) = P(A) + P(B)}$ (where: $A \cup B =$ "A *or* B")

Recall: conditional probability $P(A|B)$ was defined by $P(A|B) = P(A \cap B)/P(B)$. It is the probability of A in a universe restricted to B

Disjoint/exclusive: $(A \cap B) = 0$



Overlapping



What is a *Probability* ? (continued)

Axioms of probability → *set theory* (a "set" is a collection of things/elements)

1. A measure of how likely an "event" will occur, expressed as the ratio of favourable to all possible cases in repeatable trials

   - **Frequentist** (classical) probability: $P(\text{“event”}) = \lim_{n\to\infty}\left(\frac{\#\text{outcome is “event”}}{n\ \text{“trials”}}\right)$

2. The "degree of belief" that an event is going to happen

   - **Bayesian** probability:

     – $P(\text{“event”})$ is degree of belief that "event" will happen → no need for "repeatable trials"

     – Degree of belief (in view of the data *and* previous knowledge (belief) about the "event") that a parameter has a certain "true" value

   - Bayes' theorem: $P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}$

     The **prior** probability $P(A)$ has been modified by B to become the **posterior** probability $P(A|B)$

     Proof from conditional probability: $P(A|B)P(B) = P(A \cap B) = P(B \cap A) = P(B|A)P(A)$

Frequentist versus Bayesian statistics

*The term "observed data" is sloppy: meant is an observed estimator value, and the probability refers to cumulative estimator values

Frequentist statement:

- *Probability of the "observed data"* to occur given a model (hypothesis)*: $P(\text{data}|\text{model})$

Bayesian statement:

- *Probability of the model given the data*: $P(\text{model}|\text{data})$
- Let's look again at Bayes' theorem written slightly differently (θ = set of parameters fixing the model)

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta) \cdot P(\theta)}{P(\text{data})},$$

here:
- $P(\theta|\text{data})$: posterior probability of θ given the data
- $P(\text{data}|\theta)$: probability of data given θ
- $\boldsymbol{P(\theta)}$: **the "prior" probability for $\boldsymbol{\theta}$**
- $P(\text{data})$: a normalisation

Frequentist statistics is unaware of the "truth", and only allows to exclude unlikely hypotheses (objective statement).

Bayesian statistics speculates about "truth" by injecting arbitrary prior probabilities (subjective statement).

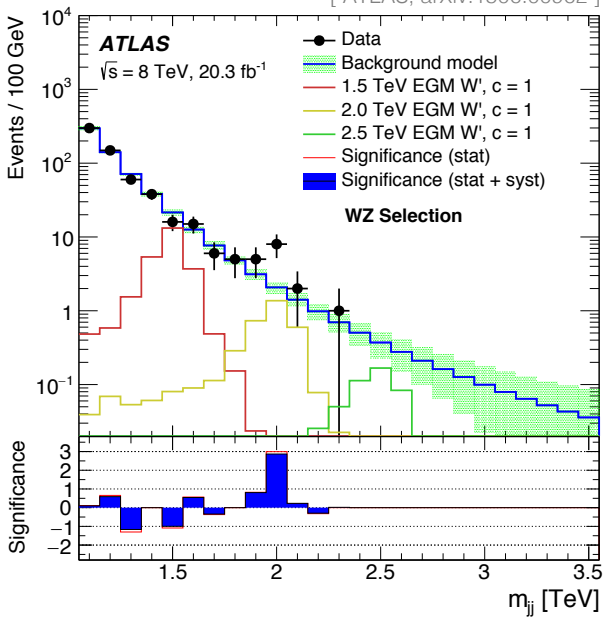
By virtue of the Central Limit Theorem, the prior dependence may be weak in concrete cases.

$P(\text{data} \mid \text{model}) \neq P(\text{model} \mid \text{data})$

Consider a new physics search where a local excess of events has been observed with a (global) significance of 2.5 standard deviations, corresponding to a ~0.6% one-sided probability

- Assuming $P(\mathbf{data} \mid \mathbf{model}) = P(\mathbf{model} \mid \mathbf{data})$, and concluding that for a given well-fitting new physics model $P(\mathbf{new\ physics\ model} \mid \mathbf{data}) \simeq 99.4\%$ is **wrong** (frequently done by the press)

[ATLAS, arXiv:1506.00962]



Frequentist statistics gives the probability to observe certain data under a given hypothesis, but it says nothing about the probability of the hypothesis to be true. Important subtlety!

- Will later define "confidence levels": if $P(\mathbf{data} \mid \mathbf{model}) < 5\%$ $\rightarrow$ discard model

Frequentist versus Bayesian statistics

Both statistical concepts have important applications

- Most LHC results use frequentist statistics as it is objective, and empirical sciences progress by successive exclusion and improvement of the understanding (theory)
  - Nevertheless, there are many Bayesian elements (eg, “decision” on exclusion or discovery given an observed $P(\mathbf{data}|\mathbf{model})$, interpretation of results)
  - It is also possible to define, problem-dependent, “objective priors” in Bayesian statistics
  - A frequentist analysis can become technically very challenging → Bayesian often simpler
- The predictivity of Bayesian statistics is useful when it comes to decision taking, eg:
  - Should I sell, buy or hold certain stocks ?
  - Should I build the LHC ? (Bayesian “no-loose” theorem)
  - Almost any decision in life…

Frequentist versus Bayesian statistics

Both statistical concepts have important applications

- Most LHC results use frequentist statistics as it is objective, and empirical sciences progress by successive exclusion and improvement of the understanding (theory)
  - Nevertheless, there are many Bayesian elements (eg, "decision" on exclusion or discovery given an observed $P(\mathbf{data}|\mathbf{model})$, interpretation of results)
  - It is also possible to define, problem-dependent, "objective priors" in Bayesian statistics
  - A frequentist analysis can become technically very challenging → Bayesian often simpler
- The predictivity of Bayesian statistics is useful when it comes to decision taking, eg:
  - Should I sell, buy or hold certain stocks ?
  - Should I build the LHC ? (Bayesian "no-loose" theorem)
  - Almost any decision in life…

Slightly provocative summary by Louis Lyons (Academic Lecture at Fermilab, August 17, 2004)

Bayesians address the question everyone is interested in, by using assumptions no-one believes

Frequentists use impeccable logic to deal with an issue of no interest to anyone

…to be taken with the grain of salt !

Hypothesis testing

A hypothesis $\boldsymbol{H}$ specifies some model which might lie at the origin of the data $\boldsymbol{x}$

a) *Point hypothesis*: H could be a particular event type (eg, Higgs boson versus background)

b) *Composite hypothesis*: H could be a parameter (eg, Higgs boson mass or coupling strength)

In case a), the PDF is simply $\mathrm{PDF}(x) = \mathrm{PDF}(x; H)$

In case b), H contains unspecified parameters (θ: mass, coupling, systematic uncertainties)

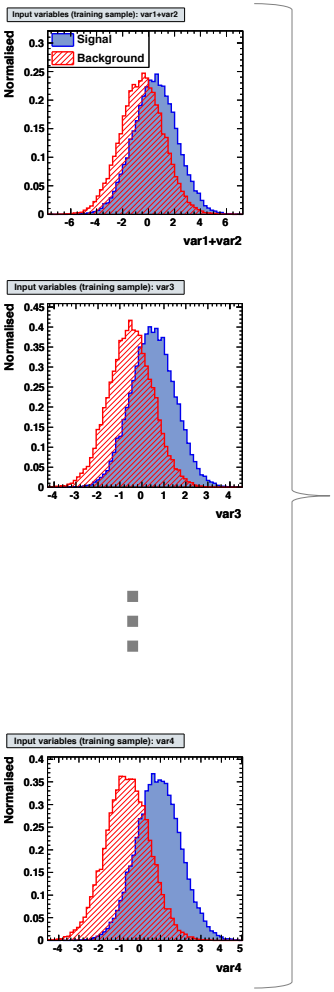
- A whole band of $\mathrm{PDF}(x; H(\theta))$
- For given x, $\mathrm{PDF}(x; H(\theta))$ can be interpreted as a function of $\theta \rightarrow$ *Likelihood function* $\boldsymbol{L}(\theta)$
- $L(\theta) = L(x|H(\theta))$ for fixed θ is the probability density to observe x given the model $H(\theta)$, but note that $L(\theta)$ is *not* the PDF of x versus θ given $H(\theta)$

Statistical tests are often formulated using a

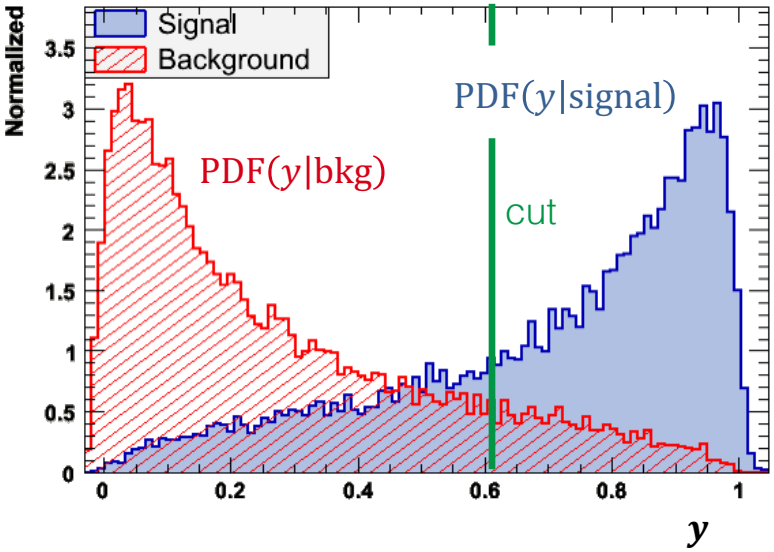
- **Null hypothesis** (eg, Standard Model (SM) background only)
- **Alternative hypothesis** (eg, SM background + new physics)

Hypothesis testing (continued)

Example, a multivariate (→ see last lecture) classification analysis to search for new physics



Take $\boldsymbol{n}$ input variables and combine into single output *discriminant* or ***test statistic*** $\boldsymbol{y}$



$\boldsymbol{y}$:
- > cut: signal region
- = cut: decision boundary
- < cut: background region

Choose *cut* value: i.e. a region where one can "reject" the null-(background-) hypothesis

(optimal cut value depends on signal and background cross-section and purity)

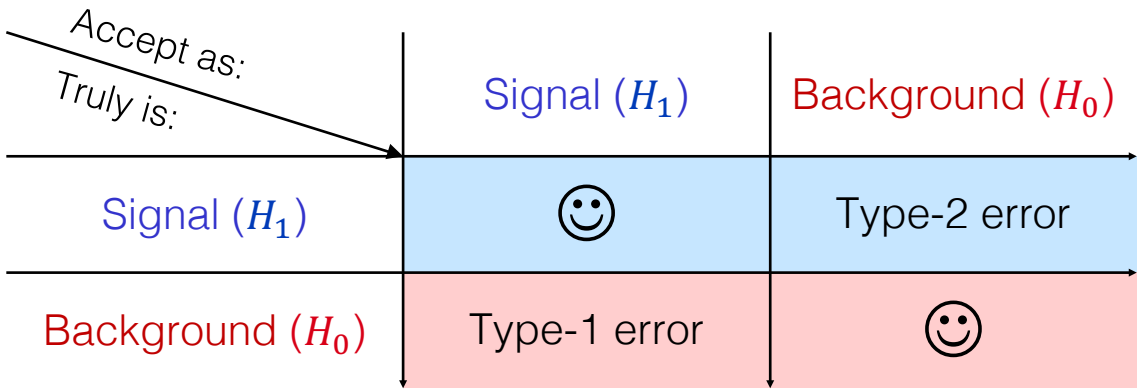
Hypothesis testing (continued)

Example: goal of new physics search: exclude null hypothesis
(as being unlikely the model underlying the observation)

It occurs that one makes mistakes:

Type-1 error: (false positive)
→ reject null hypothesis although it is true (no new physics)

Type-2 error: (false negative)
→ accept null hypothesis although it is not true (there is new physics in the data)



Hypothesis testing (continued)

Example: goal of new physics search: exclude null hypothesis
(as being unlikely the model underlying the observation)

It occurs that one makes mistakes:

Type-1 error: (false positive)
→ reject null hypothesis although it is true (no new physics)

Type-2 error: (false negative)
→ accept null hypothesis although it is not true (there is new physics in the data)

Significance $\boldsymbol{\alpha}$: Type-1 error rate:
Rate ("risk") of "false discovery", background in signal sample

$$\alpha = \int_{y(x) > \text{cut}} P(x|H_0)dx$$

should be small

Size $\boldsymbol{\beta}$: Type-2 error rate:
Power: $1 - \beta$ = sensitivity to the "alternative" theory, signal efficiency

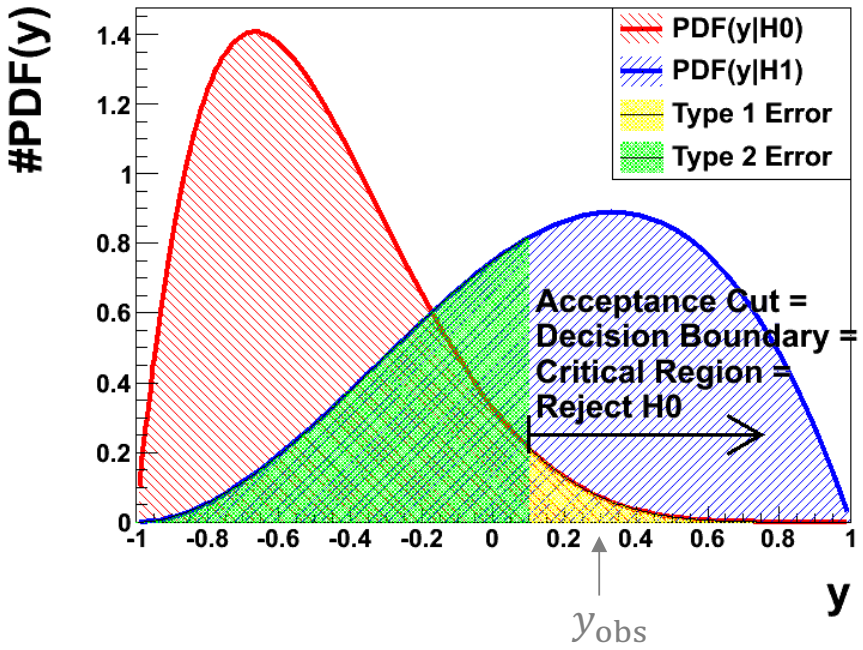
$$\beta = \int_{y(x) < \text{cut}} P(x|H_1)dx$$

should be small

Hypothesis testing (continued)

Define *critical region* $\boldsymbol{C}$ $\rightarrow$ if data (observation) falls there, reject a hypothesis

- Want to discriminate between hypotheses $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$
- Define test statistic $\boldsymbol{y(x)}$ for data $\boldsymbol{x}$
- Compute expected $\boldsymbol{y}$ distributions for two hypotheses: $\mathrm{PDF}(y(x)|H_0)$ and $\mathrm{PDF}(y(x)|H_1)$
- Compute observed test statistic $\boldsymbol{y_{obs}(x)}$ $\rightarrow$ decide on outcome whether or not $\boldsymbol{y_{obs}} \in \boldsymbol{C}$



Neyman-Pearson Lemma

Which test statistic (discriminant) $\boldsymbol{y(x)}$ should one actually choose? What is optimal?

Neyman-Pearson (1933):

> The Likelihood ratio used as test statistic $\boldsymbol{y(x)}$ gives for each significance $\boldsymbol{\alpha}$ the test (critical region) with the largest power $\boldsymbol{1-\beta}$.

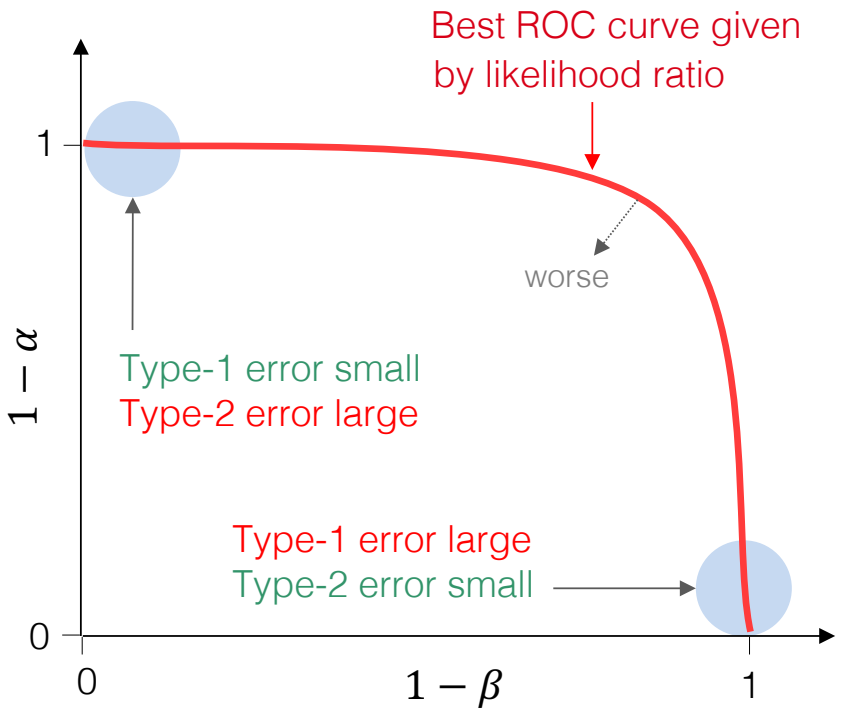
Likelihood Ratio: $\boldsymbol{y(x) = \frac{P(x|H_1)}{P(x|H_0)}}$

or any monotonic function thereof, e.g. $\ln(y(x))$

The likelihood ratio maximises area under "*Receiver Operation Characteristics*" (ROC) curve

The proof of the Neyman-Pearson Lemma is straightforward and almost obvious given the definitions of α and β

See, eg: https://en.wikipedia.org/wiki/Neyman–Pearson_lemma



Neyman-Pearson Lemma

Unfortunately, the Neyman-Pearson Lemma holds strictly only for simple hypotheses without free parameters

If $H_{0/1}$ are "composite hypotheses" $\boldsymbol{H_{0/1}(\theta)}$, it is not even sure that there exists a so-called *uniformly most powerful* test statistic that for *each* given α is the most powerful (largest $1 - \beta$)

Note: already in presence of systematic uncertainties (as varying but constrained "nuisance parameters") it is not certain that the likelihood ratio is the optimal test statistic

However: the likelihood ratio is *probably* close to optimal, it is a very convenient test statistic, and therefore commonly used in experimental particle physics

Frequentist confidence intervals

In frequentist statistics one cannot make a probabilistic statement about the true value of a parameter given the data.

Instead:

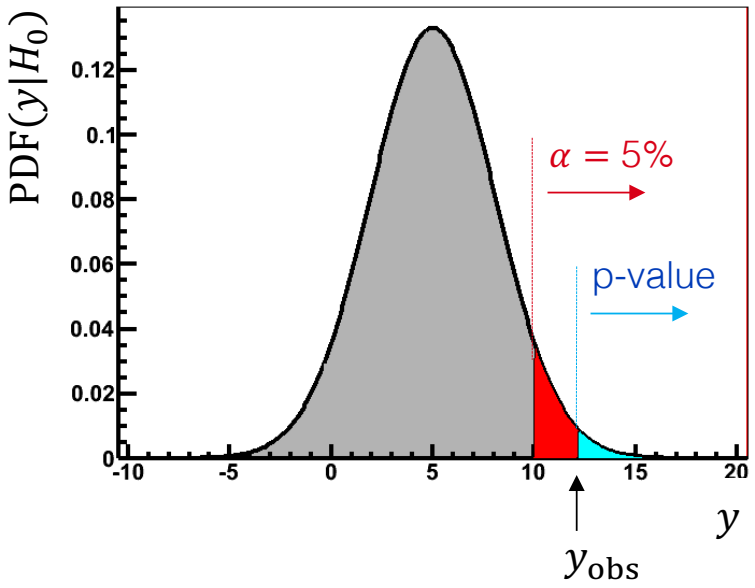
- One defines acceptance / rejection regions of a test statistic ($\boldsymbol{\alpha}$)
- The measurement (data) is one specific outcome of an ensemble of possible data
- One accepts or rejects $\boldsymbol{H_0}$ with ***confidence level*** given by $\boldsymbol{\alpha}$
- It is also possible to state how probable a particular or worse outcome (test statistic measurement) is for a given hypothesis (eg, $\boldsymbol{H_0}$) $\rightarrow$ ***p-value***

One then shows the data and quotes the H_0 outcome given the required confidence level and the hypothesis p-value

A typical (but highly simplified) frequentist analysis

1. Specify a hypothesis H_0 and test statistic or *estimator* ($\rightarrow$ likelihood ratio $\boldsymbol{y}$)
2. Specify the *significance* of the test, ie, how much of a Type-1 error rate to accept: eg, confidence level of 95% $\rightarrow$ $\boldsymbol{\alpha = 5\%}$
3. Take the measurement: $\boldsymbol{y_{obs}}$
4. Check whether $\boldsymbol{y_{obs}}$ lies inside or outside of critical region $\rightarrow$ decide on $\boldsymbol{H_0}$
5. If excluded, compute p-value of $\boldsymbol{H_0}$ to see how deep it lies in the critical region

No composite hypothesis yet (see later). Simple hypothesis test using data



$$= \int_{y_{obs}}^{\infty} \mathrm{PDF}(y|H_0)\, dy$$

Significance and p-values

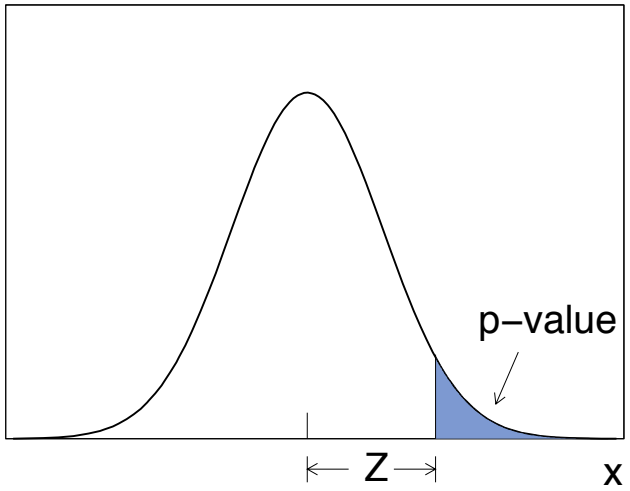
Note: $\boldsymbol{\alpha}$ (significance) must be specified before the hypothesis test is made

The **p-value** is a property of actual measurement (observation)

Again: the p-value is *not* a measure how *probable* the hypothesis is

The confidence level of a hypothesis test (accept / reject) is given by α
not the p-value

arXiv:1007.1727



It is convenient to express observed p-values in terms of Gaussian σ ("sigma"):

- How many standard deviations "Z" for same p-value on one-sided Gaussian
- In ROOT: $\mathbf{TMath::Prob}(Z*Z,1)/2 = p$ (p-value)
  (eg: p-value corresponding to $Z = 5\sigma$ is $2.87 \cdot 10^{-7}$)
- Inverse in ROOT: $\mathbf{sqrt(TMath::ChisquareQuantile}(1-2*p,1)) = Z$

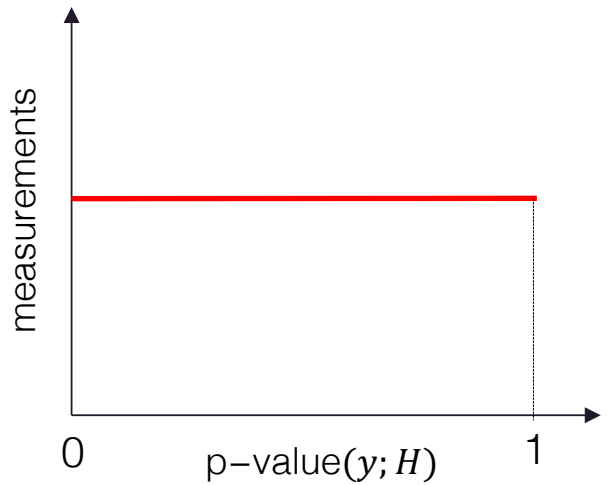
Distribution of p-values

Assume:

- Test statistic: $\boldsymbol{y}$ (function of measured quantities)
- PDF of y for given hypothesis $\boldsymbol{H}$: $\boldsymbol{p_y(y; H)}$
- **p-value**$(\boldsymbol{y}; \boldsymbol{H}) = \int_y^\infty p_y(y'; H)\, dy'$ for each measurement $\boldsymbol{y}$

p-values are random variables → distribution if measurement repeated

Derived from a cumulative distribution → must be uniform for matching hypothesis $\boldsymbol{H}$



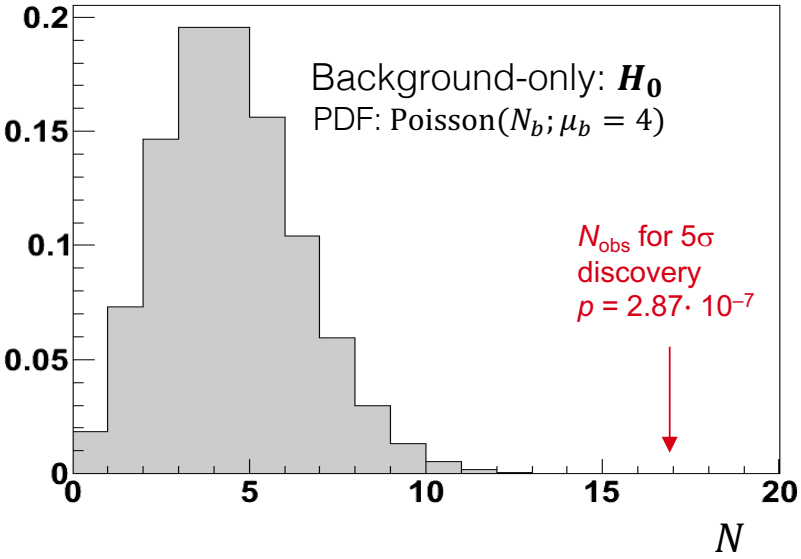
- Hence, in a fraction of times, the p-value of a given measurement may become very small, although $\boldsymbol{H}$ is the correct hypothesis
- If the *true* and *tested* hypotheses are different, the p-value distribution will deviate from uniform (but usually one cannot just repeat a measurement or an experiment to test this)

Statistical tests in new particle/physics searches

Discovery test

- Disprove background-only hypothesis $\boldsymbol{H_0}$
- Estimate probability of “upward” (or “signal-like”) fluctuation of background

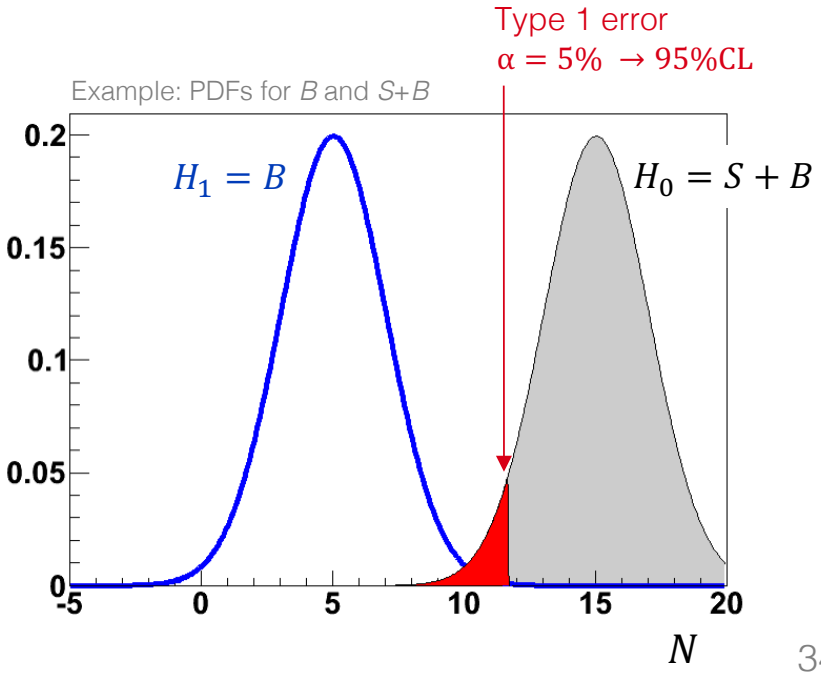
Example: PDF of background-only test statistic



Exclusion limit

- Upper limit on new physics cross section
- Disprove signal + background hypothesis $\boldsymbol{H_0}$
- Estimate probability of downward fluctuation of signal + background: find minimal signal, for which $\boldsymbol{H_0}$ (here: *S*+*B*) can be excluded at specified confidence Level

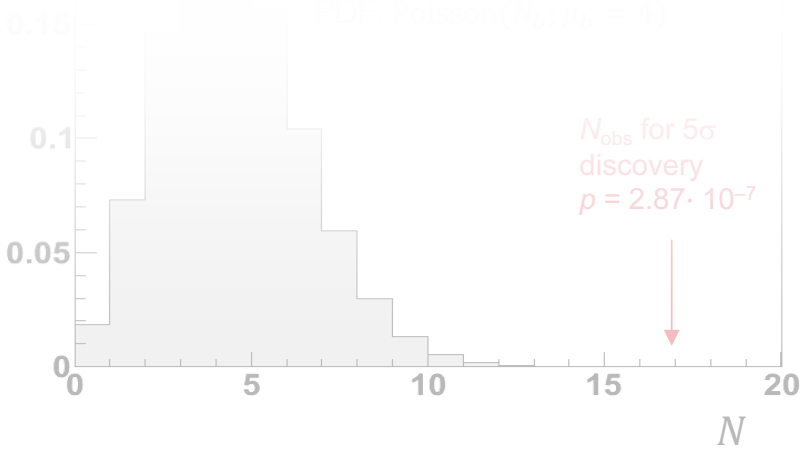
Example: PDFs for *B* and *S*+*B*

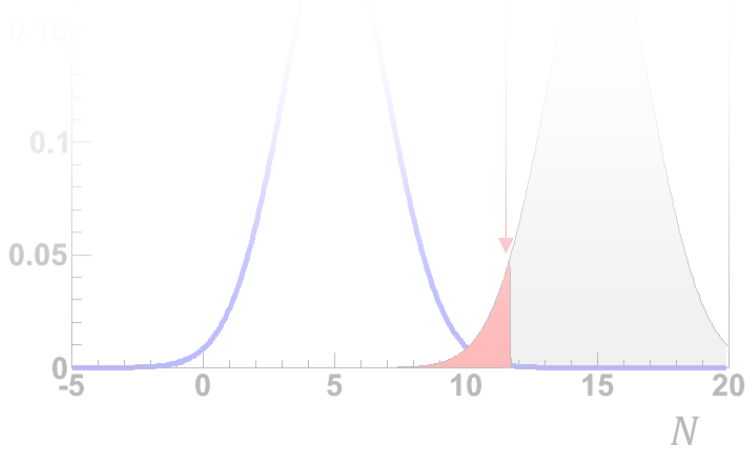


Statistical tests in new particle/physics searches

Realistic discovery and exclusion likelihood tests involve complex fits of several signal and background-normalisation (so-called *control*) regions, signal and background yields, as well as *nuisance parameters* describing systematic uncertainties.

We will come to this, but first need to learn about parameter estimation.

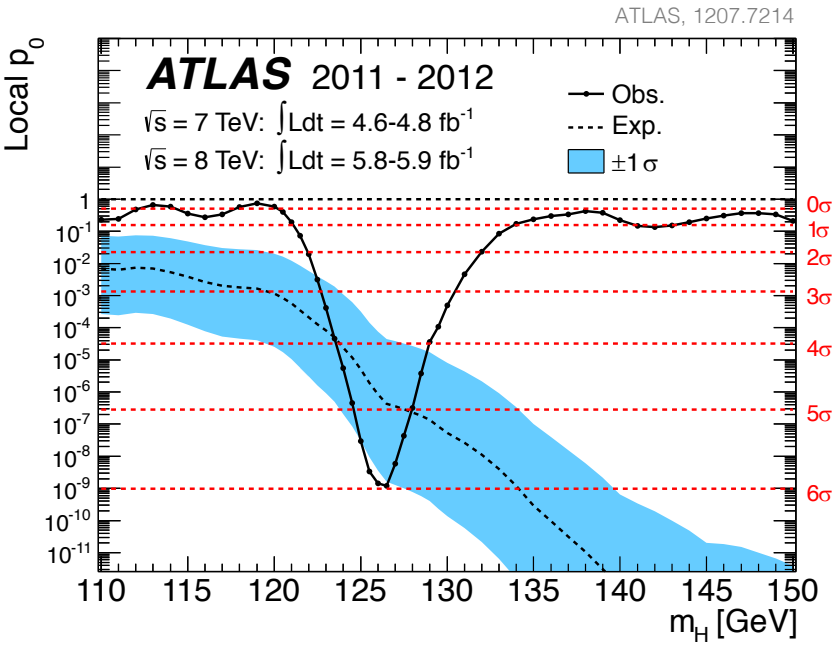




Statistical tests in new particle/physics searches — teaser

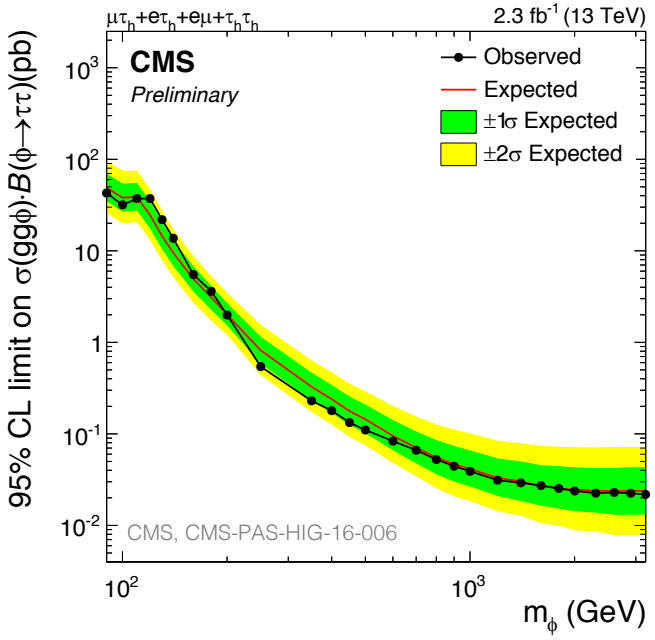
Discovery test — Higgs discovery in 2012

- 5.9σ rejection of background-only hypothesis from statistical combination of dominantly $H \to \gamma\gamma$, ZZ*, WW* decays at m_H = 126 GeV
- No ***trials factor*** (*look-elsewhere-effect*, LEE) taken into account in above number, but would not qualitatively change picture



Exclusion limit

- 13 TeV search for new physics (here: a new heavy Higgs boson) in events with at least two tau leptons
- Figure shows expected and observed 95% confidence level upper limits on cross section times branching fraction



Parameter estimation

An *estimator* is a function of a data sample $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x_1}, \ldots., \boldsymbol{x_N})$ that estimates the characteristic parameter $\boldsymbol{\theta}$ of a parent distribution.

Examples:

- Mean value estimator: $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ (one way to define the mean value, there could be others)
- Variance estimator: $\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- Median estimator …..
- ...but also: CP-asymmetry parameter in *B* meson sample (very complex parameter estimation)

The estimator $\widehat{\boldsymbol{\theta}}$ is a random variable (function of measured data that are random)

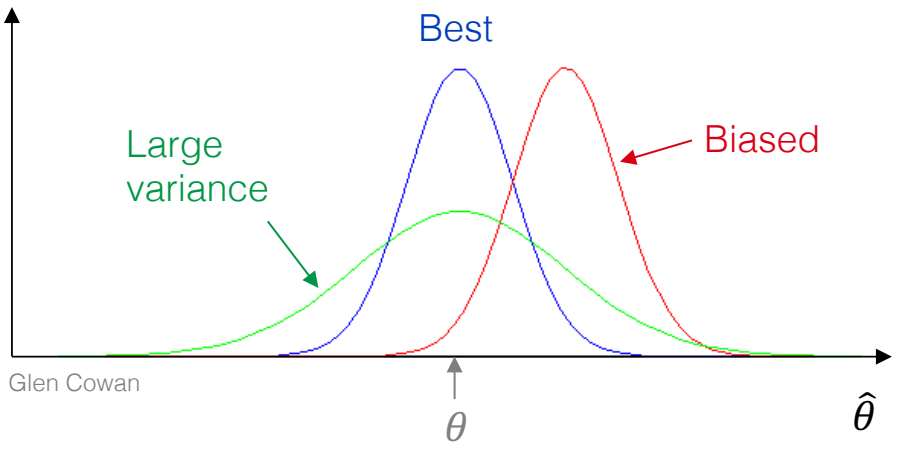
The estimator $\widehat{\boldsymbol{\theta}}$ has itself an expectation value, an expected variance, for given $\boldsymbol{\theta}$:

$\rightarrow \quad E\left[\hat{\theta}(x)\middle|\theta\right] = \int \hat{\theta}\,(x) f(x|\theta) dx$, with $f(x|\theta)$ the distribution (PDF) of the expected data

Parameter estimation

$\hat{\boldsymbol{\theta}}$ is a random variable that follows a PDF. Consider many measurements / experiments:

→ There will be a spread of $\hat{\theta}$ estimates. Different estimators can have different properties:



- Biased or unbiased: if $E\left[\hat{\theta}(x)\middle|\theta\right] = \theta \;\rightarrow$ unbiased
- Small bias and small variance can be "in conflict"
  - asymptotic bias → limit for infinite observations/data samples

Maximum likelihood estimator

Want to estimate (measure !) a parameter $\boldsymbol{\theta}$

Observe $\vec{x}_i = (x_1, \ldots . x_K)_i, i = 1, N$ (ie: K observables per event, and N events)

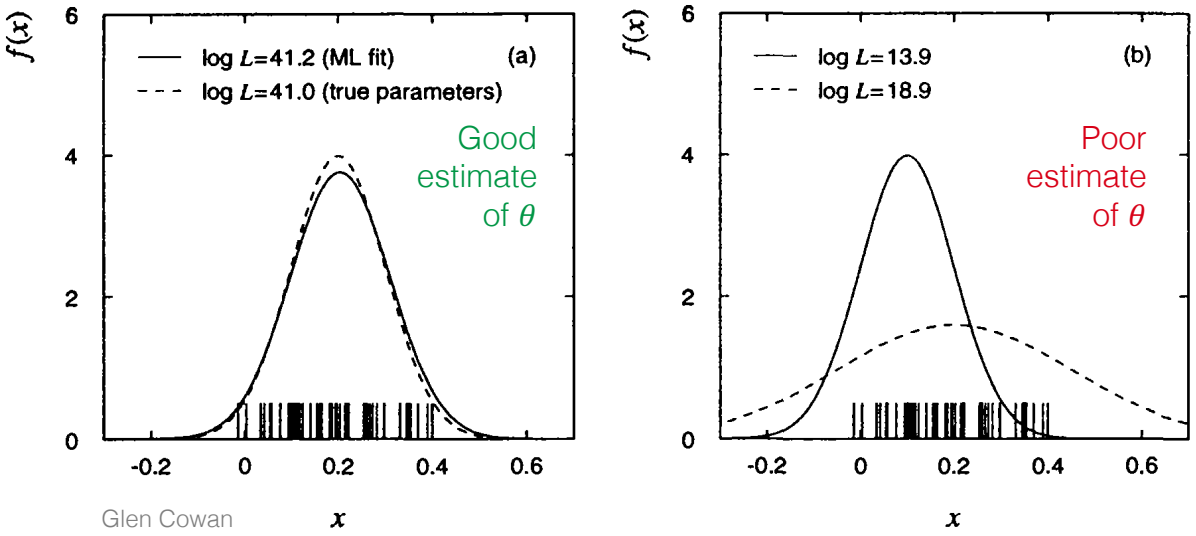
Hypothesis is PDF $p_x(\vec{x}; \theta)$, ie, the distribution of $\vec{x}$ given θ

There are N *independent* events $\rightarrow$ combine their PDFs: $P(\vec{x}_1, .. \vec{x}_N; \theta) = \prod_{i=1}^{N} p_x(\vec{x}_i; \theta)$

For fixed $\vec{x}$ consider $p_x(\vec{x}; \theta)$ as function of θ $\rightarrow$ ***Likelihood*** $\boldsymbol{L(\theta)}$

- $L(\theta)$ is at maximum (if unbiased) for $\hat{\theta} = \theta_{\text{true}}$

50 observations of .Gaussian random variable with mean 0.2 and σ=0.1



Task: maximise $L(\theta)$ to derive best estimate for $\hat{\theta}$

In practice, often minimise $-2 \cdot \ln(L(\theta))$ (see later why)

$\rightarrow$ **Maximum likelihood fit**

Maximum likelihood estimator (continued)

Let's take the Gaussian example from before: $L\,(\mu,\sigma|x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

- Measure N events: $x_1,\ldots,x_N$
- Full likelihood given by: $L(\mu,\sigma|x) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$
- In logarithmic form: $-2\cdot\ln\big(L(\mu,\sigma|x)\big) = \sum_{i=1}^{N}\left(\frac{(x_i-\mu)^2}{\sigma^2}\right) - 2N\cdot\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$

→ In a full maximum likelihood fit one could now determine $\hat{\mu}$ and $\hat{\sigma}$

→ If one is not interested in fitting σ but just μ, one can omit the (then constant) 2nd term:

$$-2\cdot\Delta\ln\big(L(\mu|x)\big) = \sum_{i=1}^{N}\left(\frac{(x_i-\mu)^2}{\sigma^2}\right)$$ → which is the **"least squares" (χ^2)** expression

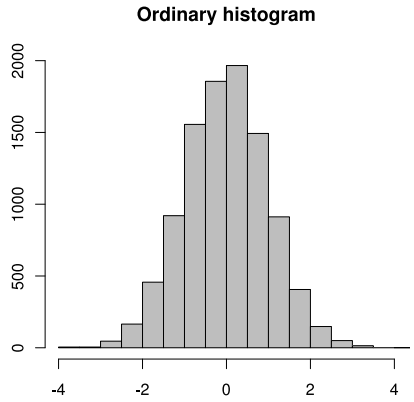
where: $\Delta\ln\big(L(\mu|x)\big) = \ln\big(L(\mu|x)\big) - \text{constant term}$

Maximum likelihood estimator (continued)

So far considered *unbinned* datasets (i.e., likelihood is given by product of PDFs for each event)

One can replace the events by bins of a histogram

- Useful if very large number of events, or PDF has very complex form, or if only broad regions are considered rather than the full shape of a PDF
- Most LHC analyses use binned maximum likelihood fits



Each bin $\boldsymbol{i}$ has $\boldsymbol{N_i}$ events that are Poisson distributed around $\boldsymbol{\mu_i}$

- The prediction of the μ_i can be obtained from Monte Carlo simulation

Likelihood function: $L(\theta) = P\left(N_1, \ldots N_{n_{\text{bins}}}; \theta\right) = \prod_{i=1}^{n_{\text{bins}}} \frac{\mu_i^{N_i}(\theta)}{N_i!} e^{-\mu_i(\theta)}$

…and in log form: $-2 \cdot \ln(L(\theta)) = 2 \sum_{i=1}^{n_{\text{bins}}} \left(\mu_i(\theta) - N_i \ln(\mu_i(\theta)) - \ln(N_i!)\right)$

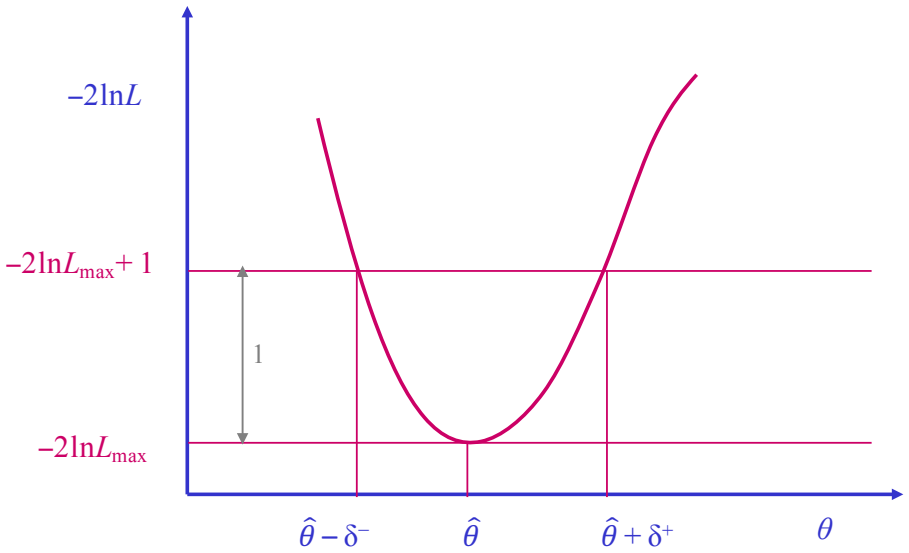
Maximum likelihood estimator (continued)

Maximum likelihood estimator is typically unbiased only in limit $N \to \infty$

If likelihood function is Gaussian (often the case for large N by virtue of central limit theorem):

→ Estimate 1σ confidence interval for θ ("parameter uncertainty") by finding intersections $\mathbf{-2 \cdot \Delta \ln(L) = 1}$ around minimum

→ Resulting uncertainty on θ may be asymmetric



If (very) non-Gaussian:

→ revert typically to (classical) *Neyman confidence intervals* (→ see later)

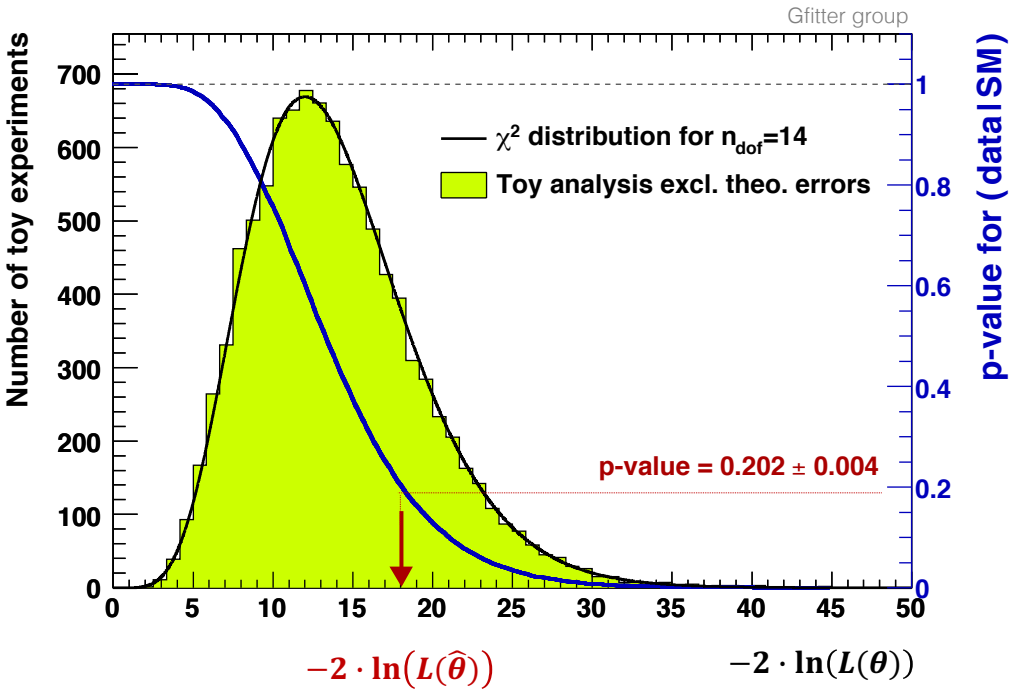
Goodness-of-Fit (GoF)

Maximum likelihood estimator determines the best parameter $\widehat{\boldsymbol{\theta}}$

But: does the model with the best $\hat{\theta}$ fit the data well ?

The value of $-2 \cdot \ln\left(L(\hat{\theta})\right)$ at minimum does not mean much → needs *calibration*

→ Determine the expected distribution of $-2 \cdot \ln\left(L(\hat{\theta})\right)$ using pseudo Monte Carlo events, and compare measured value to expected ones



Goodness-of-Fit (continued)

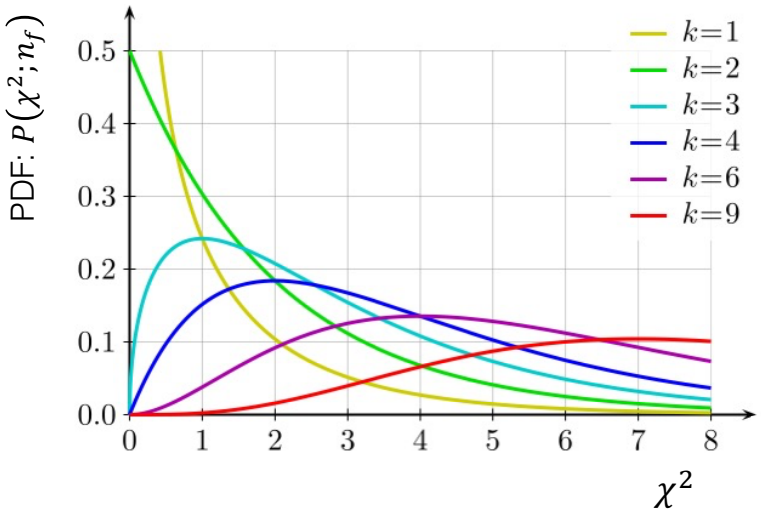
A Goodness-of-fit test is more straightforward with $\boldsymbol{\chi^2}$ estimator

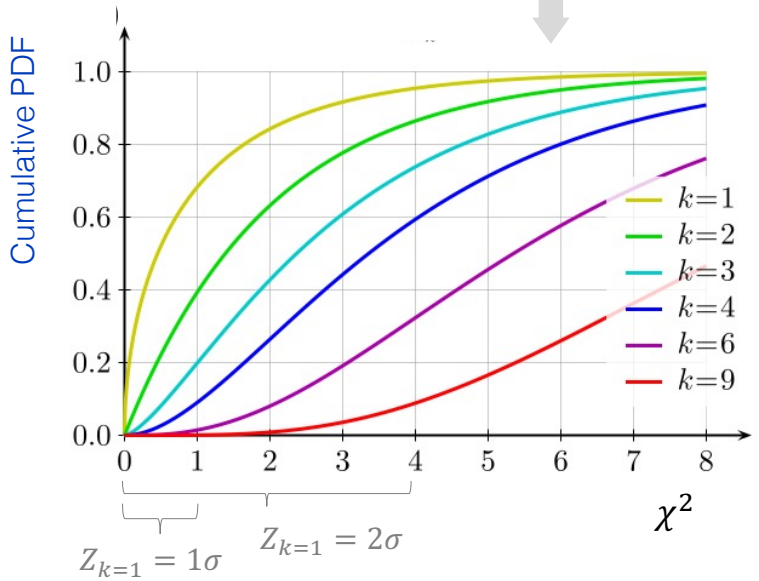
Let's use the binned example again. The task is to minimise versus θ:

$$\chi^2_{\min}(\hat{\theta}) = \min_\theta \left\{ \chi^2(\theta) = \sum_{i=1}^{n_{\text{bins}}} \left(\frac{(N_i - \mu_i(\theta))^2}{\sigma_i^2} \right) \right\}$$

χ^2 has known properties: $E[\chi^2] = n_{\text{d.o.f}} = k$ (= number of degrees of freedom)

Cumulative PDF: probability to find $\chi^2 > \chi^2_{\min}$: $\mathbf{TMath::Prob}(\chi^2_{\min}, k)$



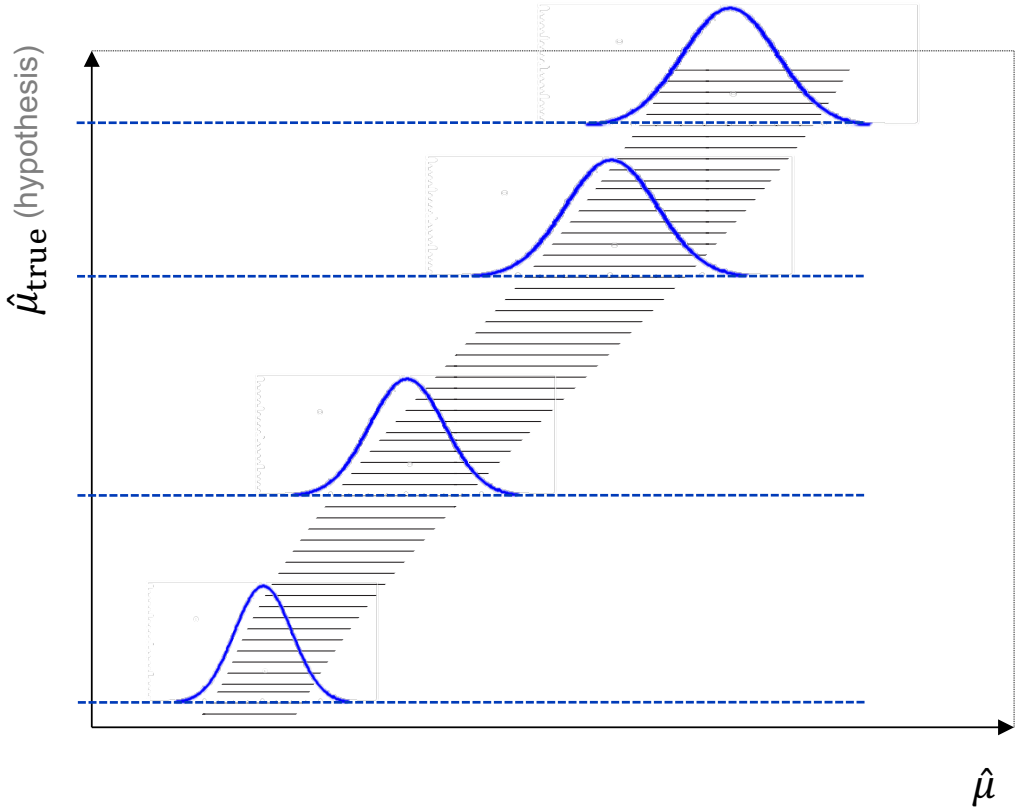


Figures from:
https://en.wikipedia.org/wiki/Chi-squared_distribution

Classical confidence level

Neyman *confidence belt* for confidence level (CL) $\boldsymbol{\alpha}$ (e.g. 95%)

Statement about probability to cover true value $\hat{\boldsymbol{\mu}}_{\mathbf{true}}$ of parameter $\hat{\boldsymbol{\mu}}$ fit to data

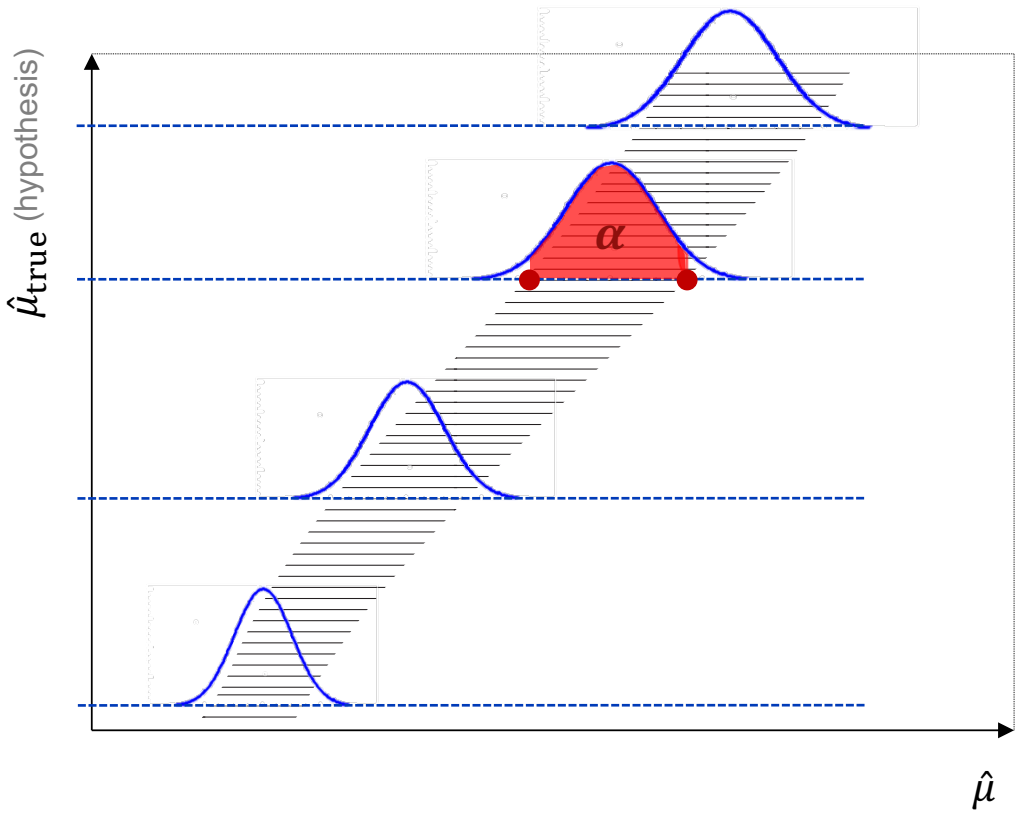


- Each hypothesis $\hat{\mu}_{\mathbf{true}}$ has a PDF of how the measured values $\hat{\mu}_{\mathbf{obs}}$ will be distributed

Classical confidence level

Neyman *confidence belt* for confidence level (CL) $\boldsymbol{\alpha}$ (e.g. 95%)

Statement about probability to cover true value $\hat{\boldsymbol{\mu}}_{\mathbf{true}}$ of parameter $\hat{\boldsymbol{\mu}}$ fit to data

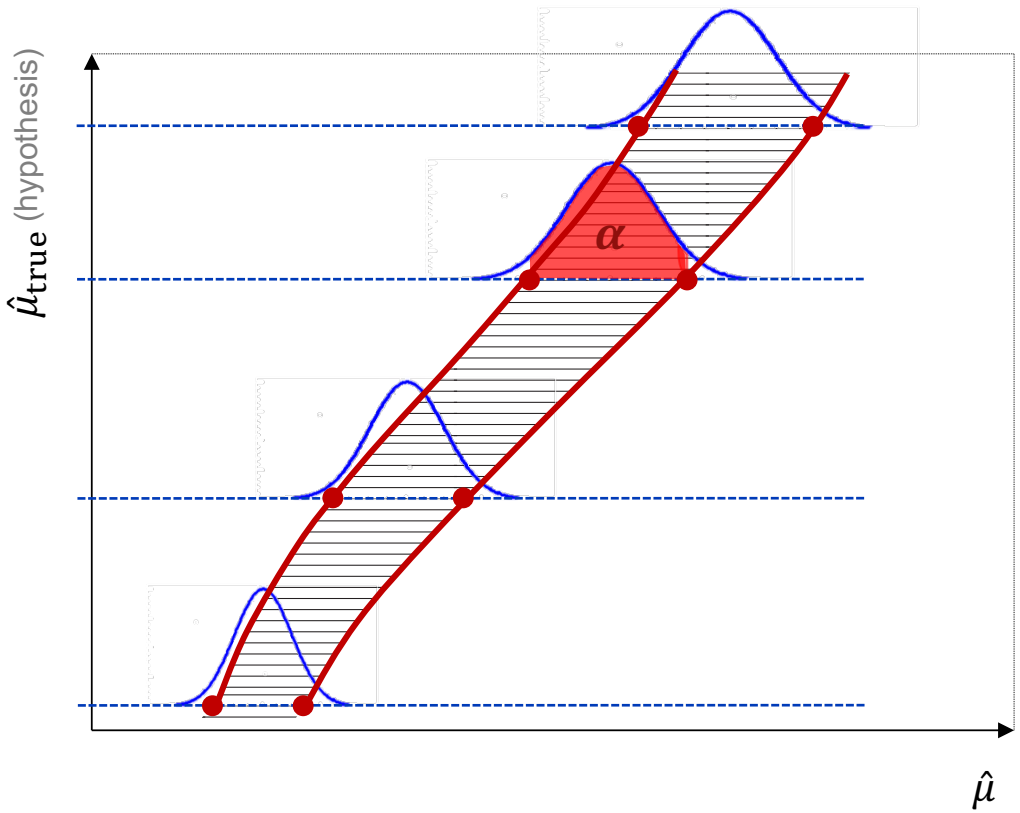


- Each hypothesis $\hat{\mu}_{\mathbf{true}}$ has a PDF of how the measured values $\hat{\mu}_{\mathbf{obs}}$ will be distributed
- Determine the (central) intervals ("acceptance region") in these PDFs such that they contain $\boldsymbol{\alpha}$

Classical confidence level

Neyman *confidence belt* for confidence level (CL) $\boldsymbol{\alpha}$ (e.g. 95%)

Statement about probability to cover true value $\hat{\boldsymbol{\mu}}_{\mathbf{true}}$ of parameter $\hat{\boldsymbol{\mu}}$ fit to data

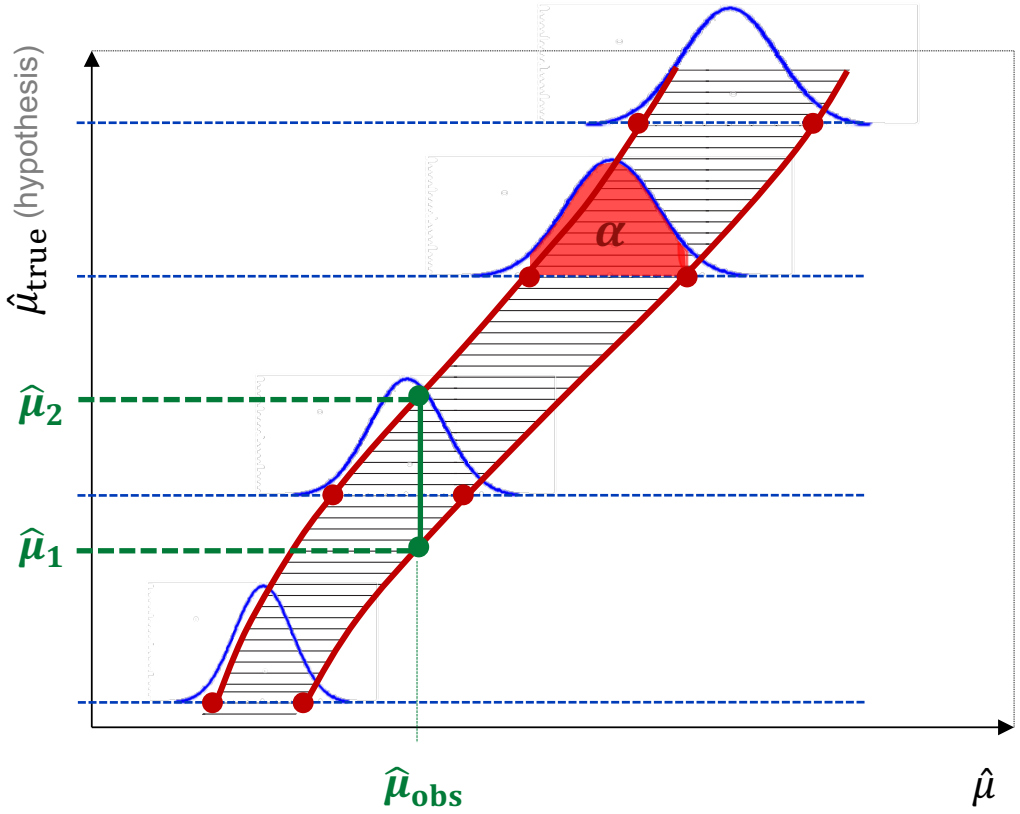


- Each hypothesis $\hat{\mu}_{\mathrm{true}}$ has a PDF of how the measured values $\hat{\mu}_{\mathrm{obs}}$ will be distributed
- Determine the (central) intervals ("acceptance region") in these PDFs such that they contain $\boldsymbol{\alpha}$
- Do this for all $\hat{\mu}_{\mathrm{true}}$ hypotheses
- Connect all the red dots: **confidence belt**

Classical confidence level

Neyman *confidence belt* for confidence level (CL) $\boldsymbol{\alpha}$ (e.g. 95%)

Statement about probability to cover true value $\hat{\boldsymbol{\mu}}_{\mathbf{true}}$ of parameter $\hat{\boldsymbol{\mu}}$ fit to data



- Each hypothesis $\hat{\mu}_{\mathrm{true}}$ has a PDF of how the measured values $\hat{\mu}_{\mathrm{obs}}$ will be distributed
- Determine the (central) intervals ("acceptance region") in these PDFs such that they contain $\boldsymbol{\alpha}$
- Do this for all $\hat{\mu}_{\mathrm{true}}$ hypotheses
- Connect all the red dots: **confidence belt**
- Measure $\hat{\boldsymbol{\mu}}_{\mathbf{obs}}$

→ Confidence interval $[\,\hat{\mu}_1, \hat{\mu}_2\,]$ given by **vertical** line intersecting the belt

→ α = 95% of the intervals $[\,\hat{\mu}_1, \hat{\mu}_2\,]$ contain $\hat{\mu}_{\mathrm{true}}$

Combining confidence intervals

The construction of Neyman intervals may involve large resources if done with pseudo Monte Carlo experiments. In many cases, experiments take "Gaussian" short cut, assuming that the $\mathrm{PDF}(\hat{\mu}_{\mathrm{true}})$ is Gaussian and does not depend on $\hat{\mu}_{\mathrm{true}}$ (see previous slides)

In Gaussian case, measurements can be combined by multiplying their likelihood functions

Otherwise: it is important to combine individual measurements, not the confidence intervals: construct confidence belt of combined measurement

The following "Gaussian shortcut" will be wrong in that case:

SME coefficient determined in [8] and $(CL)_{\bar{\nu}}$ the 99.7% C.L. upper limit determined here. We combine the two limits as

$$1/(CL)^2 = 1/(CL)_{\nu}^2 + 1/(CL)_{\bar{\nu}}^2,$$

where (CL) is the combined 99.7% C.L. upper limit. The most sensitive upper limits we have determined with the MINOS neutrino and antineutrino data are given in Table IV. As discussed, the way we determine the upper lim-

arXiv:1201.2631v2

In a perfectly Gaussian and uncorrelated case, this simple formula is correct

Combining confidence intervals

The construction of Neyman intervals may involve large resources if done with pseudo Monte Carlo experiments. In many cases, experiments take “Gaussian” short cut, assuming that the $\mathrm{PDF}(\hat{\mu}_{\mathrm{true}})$ is Gaussian and does not depend on $\hat{\mu}_{\mathrm{true}}$ (see previous slides)

In Gaussian case, measurements can be combined by multiplying their likelihood functions

Otherwise: it is important to combine individual measurements, not the confidence intervals: construct confidence belt of combined measurement

The following “Gaussian shortcut” will be wrong in that case:

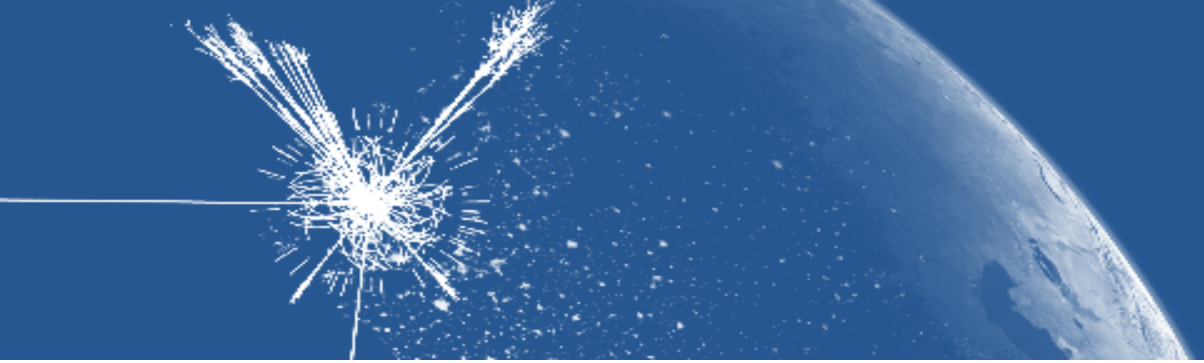
SME coefficient determined in [8] and $(CL)_{\bar{\nu}}$ the 99.7% C.L. upper limit determined here. We combine the two limits as

$$1/(CL)^2 = 1/(CL)_{\nu}^2 + 1/(CL)_{\bar{\nu}}^2,$$

where (CL) is the combined 99.7% C.L. upper limit. The most sensitive upper limits we have determined with the MINOS neutrino and antineutrino data are given in Table IV. As discussed, the way we determine the upper lim-

arXiv:1201.2631v2

In a perfectly Gaussian and uncorrelated case, this simple formula is correct



Summary for today

We have introduced the axioms of probability theory and discussed the difference between frequentist and Bayesian statistics

We have discussed hypothesis testing, introduced Type-1 and 2 errors, and the Neyman-Pearson likelihood-ratio lemma

Test statistics, confidence intervals, significance and p-values were introduced

Parameter estimation with the maximum likelihood technique, goodness-of-fit, and the derivation of a classical Neyman confidence belt were discussed

Next: realistic maximum likelihood fits, Monte Carlo techniques and data unfolding