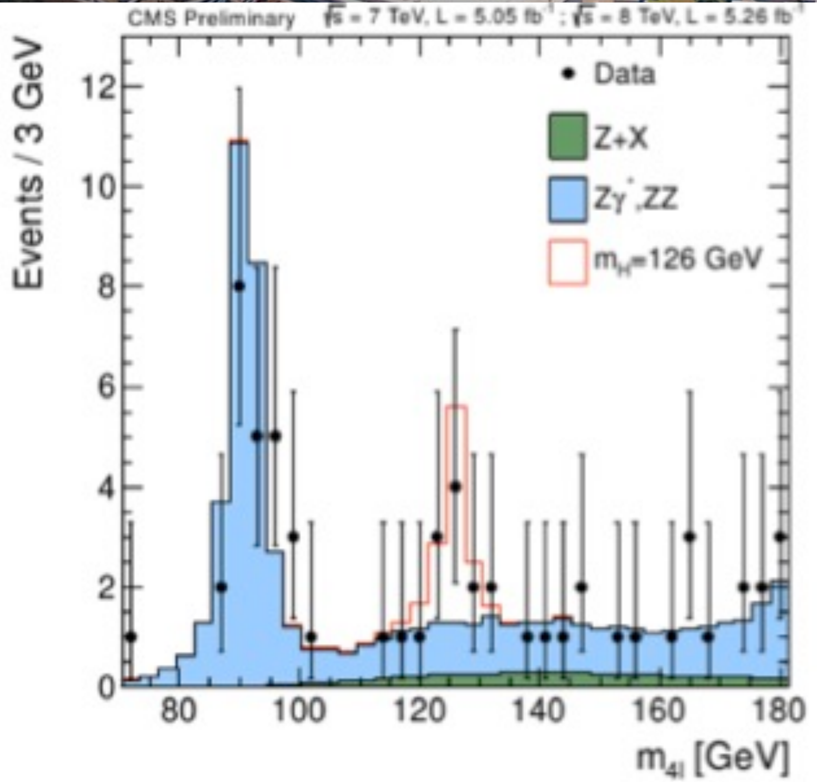
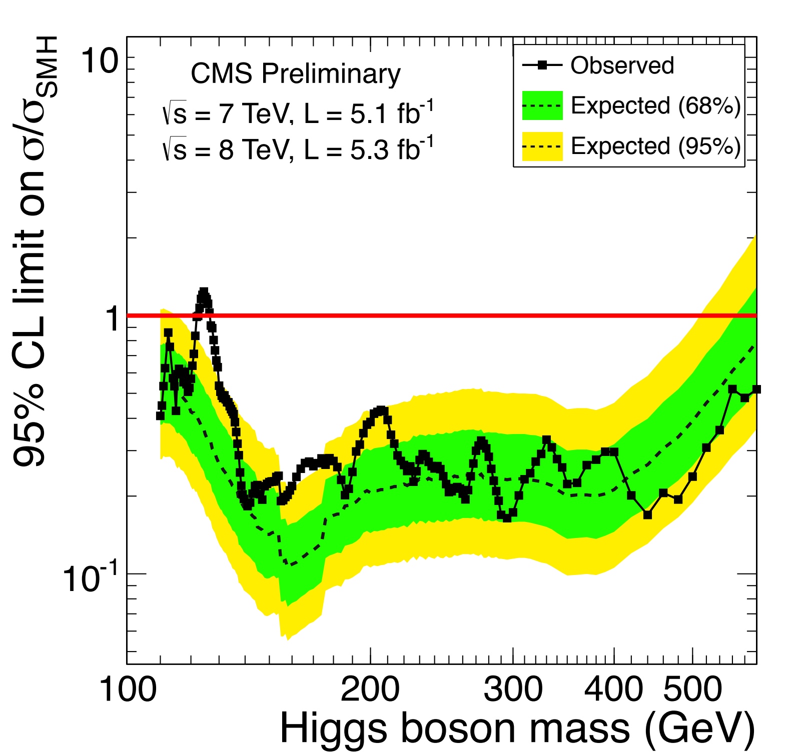
Statistical Methods in Data Analysis

Confidence Intervals





Andreas B. Meyer
DESY

18–22 March 2024

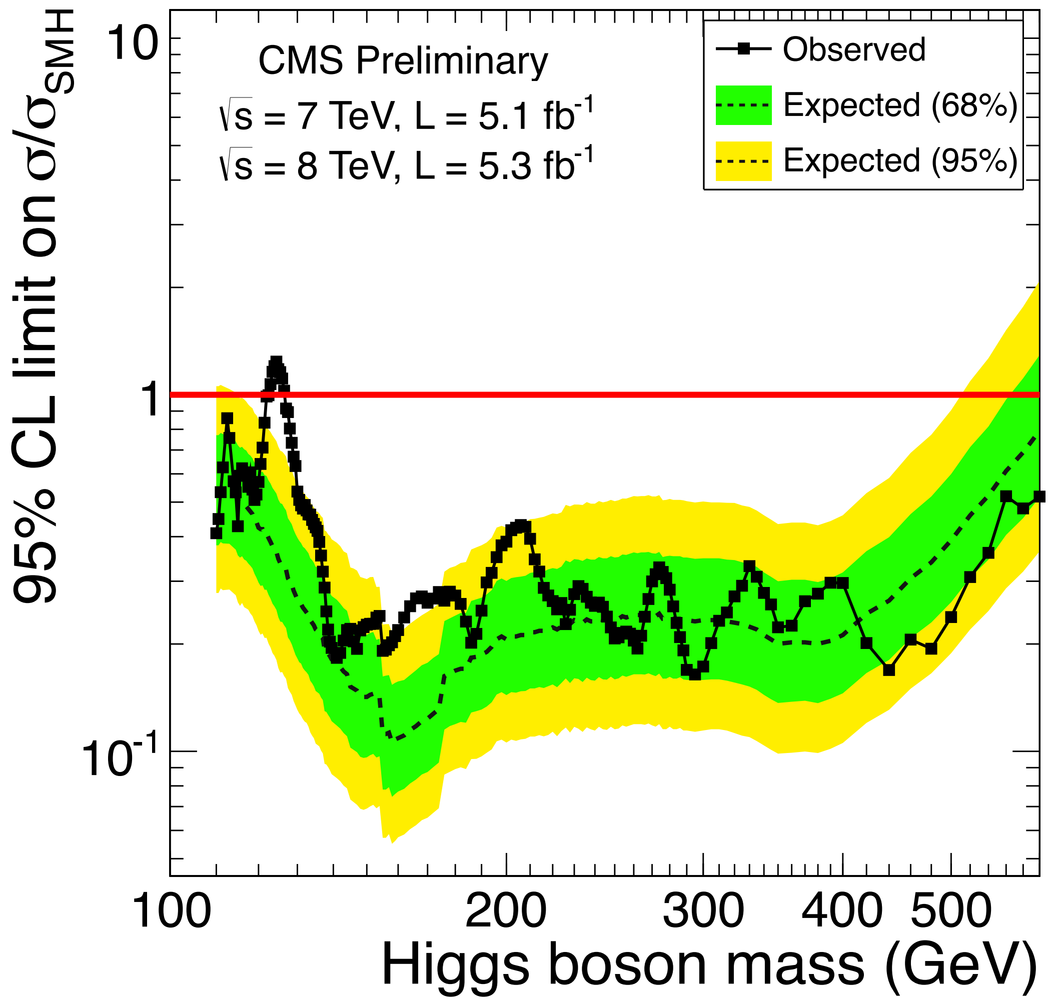
Menu

Confidence Intervals

Tuesday
- Statistical and systematic uncertainties
- Probability
- Parameter estimation

Wednesday
- Hypothesis testing
- Confidence intervals
- Profile likelihood ratio
- Outlook: classification and MVA

Friday
Matthias Komm:
Introduction to machine learning



Higgs discovery: what is shown in this figure ?

Links, Papers and Sources

Statistical Methods in Data Analysis", Terascale, March 2024: https://www.desy.de/~ameyer/da_desy24

Previous lectures:

- Statistical Methods in Data Analysis", Introduction to the Terascale, March 2023: https://indico.desy.de/event/33888 and https://www.desy.de/~ameyer/da_desy23/
- Statistical Methods in Data Analysis", KSETA lecture, Feb 2022: https://www.desy.de/~ameyer/da_kseta_22/
- "Moderne Methoden der Datenanalyse", Course lecture at KIT, SoSe 2017, slides (in German): https://www.desy.de/~ameyer/kit/da_sose17/index.html **Access to slides and material: (user: Students. pw: only)**

Papers and articles:

- Robert Cousins: "Why isn't every physicist a Bayesian ?", Am.J.Phys. 65 (1995).
- Robert Cousins: "Lectures on Statistics in Theory: Prelude to Statistics in Practice" [arXiv]
- G.Cowan, Particle Data Group [pdg] 2020, chapter 40 [pdf] or full PDG book for download (80MB) [pdf]
- G.Cowan, K.Cranmer, E.Gross, O.Vitells: "Asymptotic formulae for likelihood-based tests of new physics" [arXiv]
- ATLAS and CMS Collaborations: "Procedure for the LHC Higgs boson search combination" [CDS]
- T.Junk: "Confidence level computation for combining searches with small statistics", NIM, A 434 (1999) 435-443
- A.Read: "Presentation of search results: the CL_s technique", J.Phys.G: 28 (2002)

Many thanks for discussions, material and help go to:

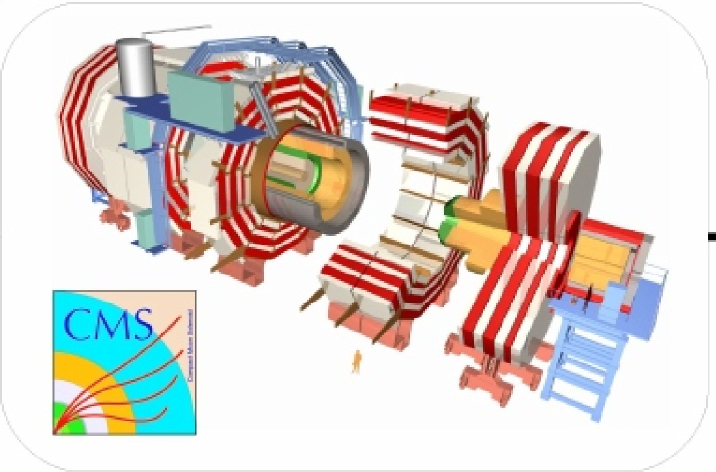
- G. Quast (KIT), R. Wolf (KIT), O. Behnke (DESY), C. Autermann (Aachen)

Recap

The scientific cycle

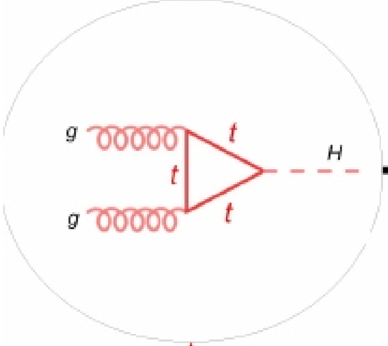
Particle physics

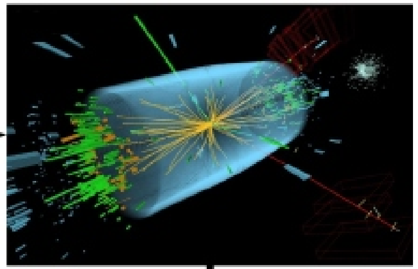
Experiment: measure and test theory predictions

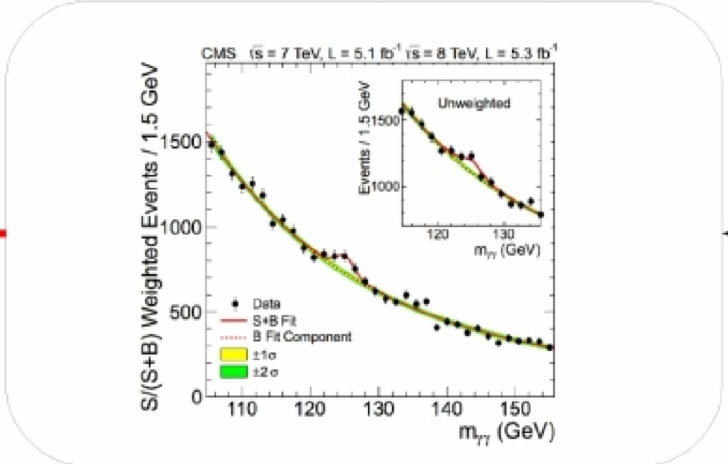


Hypothesis tests

Theory: predict measurement







Experimental input to theory

Statistical analysis and
data interpretation

Parameter estimation

Confidence intervals

Bayes' Theorem

Application in measurements

Thomas Bayes, 1763

"Posterior" "Likelihood" "Prior"

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

"Evidence"

Probability that theory "A" is correct, given data "B" have been measured

Conditional probability to measure data "B" assuming that theory "A" is correct

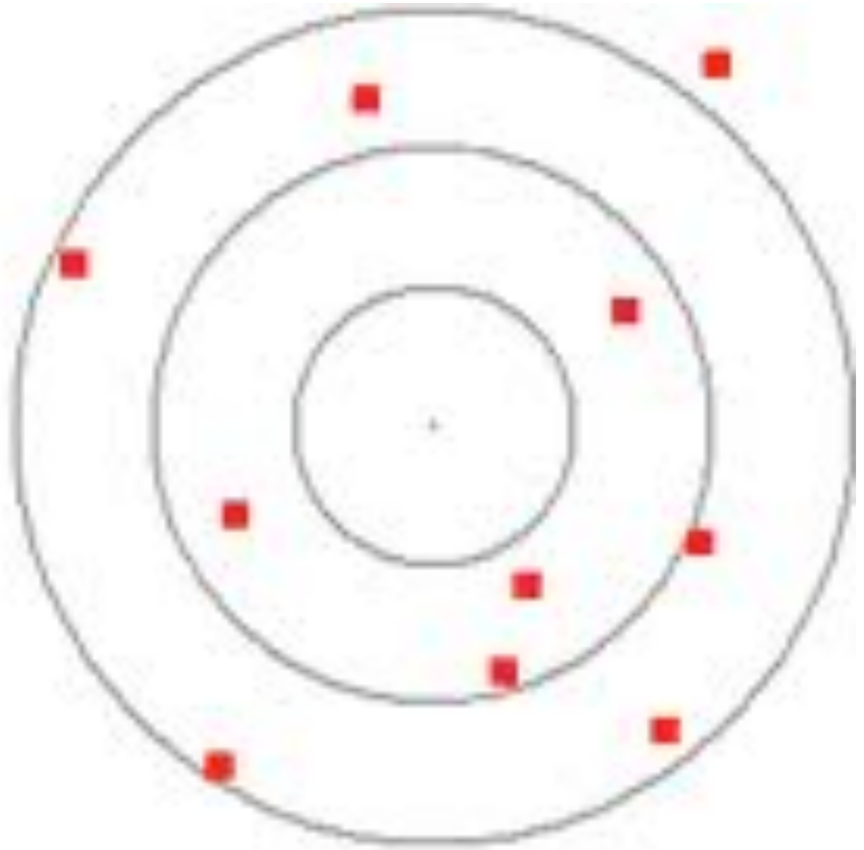
Quantitative relation between: correctness of a theory ↔ observation of actual data

Statistical uncertainties

- Spread of a single measurement for reasons that are practically (e.g. cube) and/or principally (QM) untraceable
  - => Variance: distribution around mean

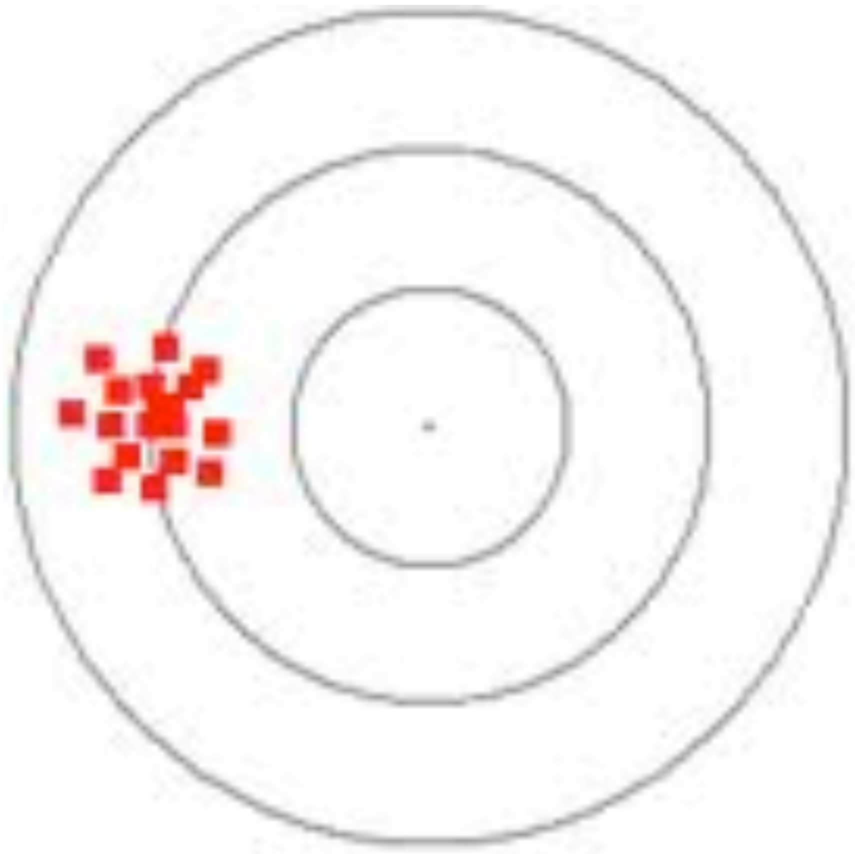
- Repeated measurements are independent (uncorrelated)

Statistical uncertainties are theoretically well understood



Systematic uncertainties

- Bias (distortion) of measurement
- Systematic uncertainties are (in principle) traceable
- Repeated measurements are usually correlated (unless underlying assumptions or analysis approach are changed)
- In practice, no general method for quantification



Estimation requires care and courage

Maximum Likelihood

Maximum likelihood

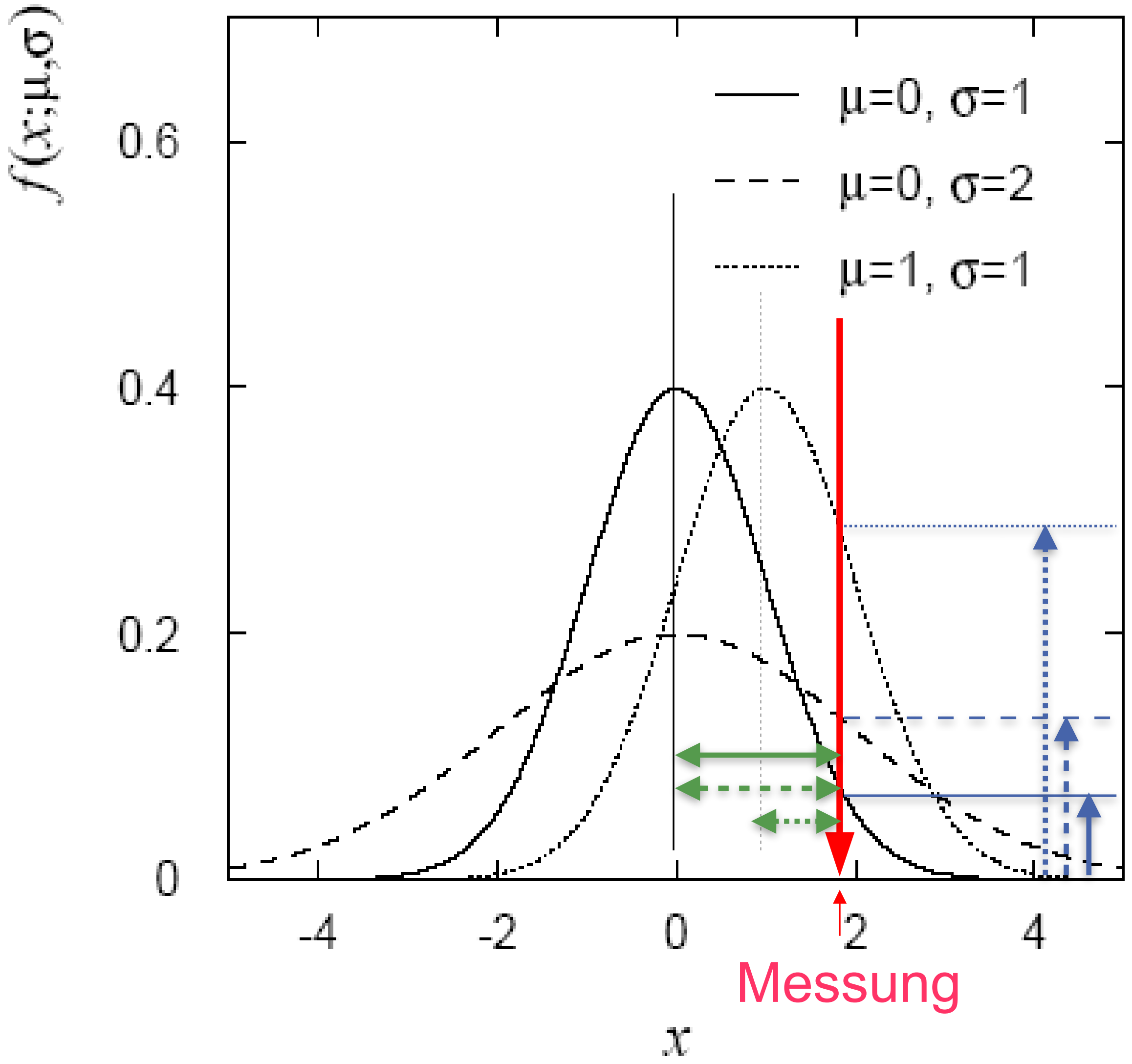
- LS: Least Squares $\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$

  Minimise distance from expectation

- MLE: Maximum Likelihood Estimator

  Maximise PDF value

- Example:
  - Decide between three hypotheses, described by their PDF
  - Measured value: 1.9
  - ➡ MLE and LS both prefer μ=1, σ=1



In general, MLE and LS can lead to different results

Maximum likelihood

- LS: Least Squares $\chi^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$

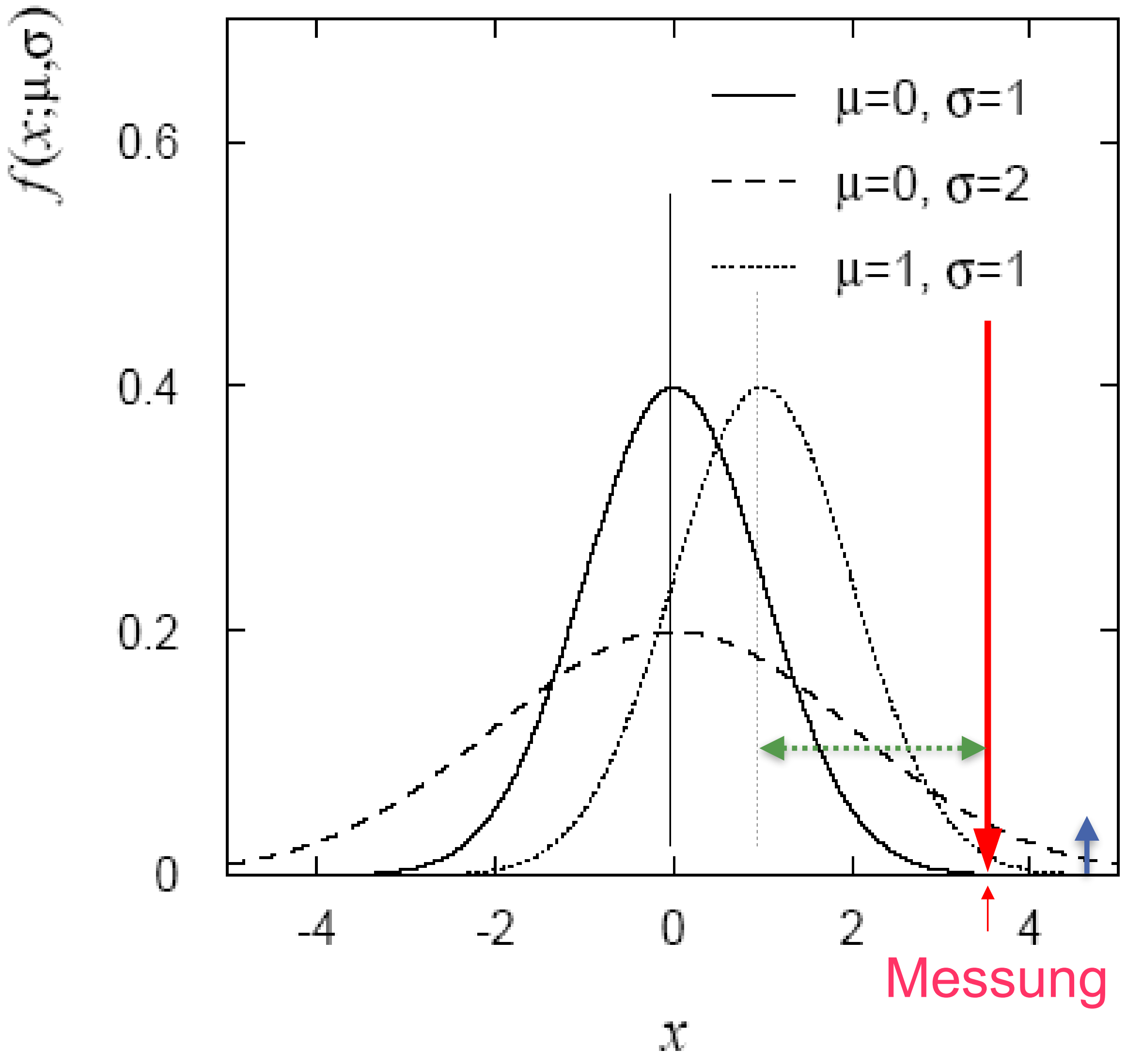
  Minimise distance from expectation

- MLE: Maximum Likelihood Estimator

  Maximise PDF value

- Example:
  - Decide between three hypotheses, described by their PDF
  - Measured value: 3.5

➡ MLE: μ=0, σ=2
➡ LS: μ=1, σ=1



In general, MLE and LS can lead to different results

Maximum likelihood and least squares

- For Gaussian distributions, LS and MLE are equivalent:
  - Likelihood of one x_i given a with a Gaussian-PDF $$f(x_i|a) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left[-\frac{(x_i - a)^2}{2\sigma_i^2}\right]$$
  - Negative logarithm of the likelihood (for all x_i) $$F(a) = -\ln \prod_i f(x_i|a) = -\ln \mathcal{L}(a)$$

$$-\ln \mathcal{L}(a) = \underbrace{\frac{1}{2}\sum_i \frac{(x_i - a)^2}{\sigma_i^2}}_{\chi^2} + \underbrace{\sum_i \ln(\sqrt{2\pi}\sigma_i)}_{\text{const. w.r.t } a \text{ (for fixed } \sigma_i)}$$

- Thus, for the variation: $$\Delta(-\ln \mathcal{L}) = \frac{1}{2}\Delta\chi^2$$

***χ*² is a special case of Maximum Likelihood, for the assumption of a Gaussian PDF**

Comparison MLE and LS

- If MLE is test statistic for a Gaussian PDF:

$$\Delta(-\ln \mathcal{L}) = \frac{1}{2}\Delta\chi^2$$

	Δ(-ln L)	$\Delta\chi^2$
1σ	0.5	1
2σ	2	4
3σ	4.5	9
$n\sigma$	$n^2/2$	n^2

- This is often the case <=> Wilks' theorem

- Things are more difficult if the PDF is not a Gaussian:

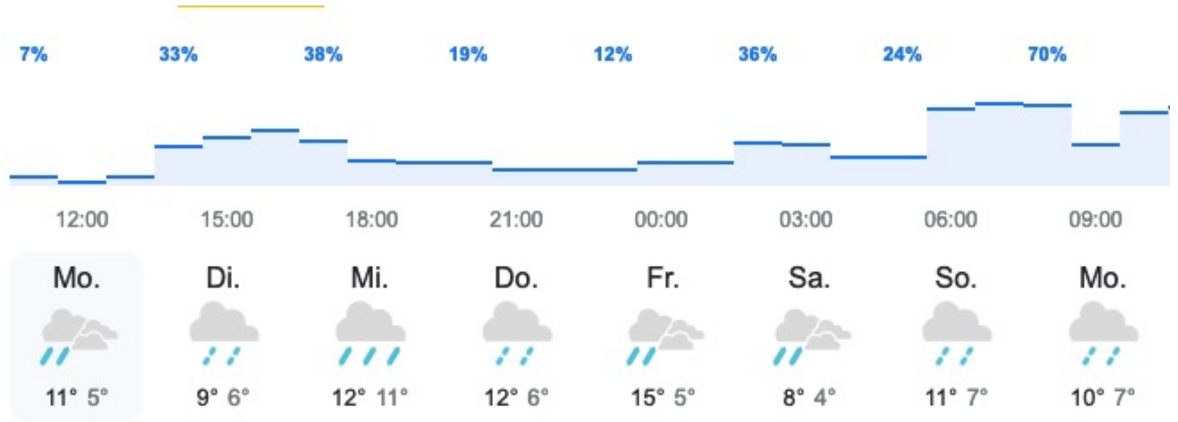
	Maximum Likelihood	**Least Squares (Gaussian)**
Method	PDF value	Distance from mean
Prerequisit	PDF is known	Mean and variance
Efficiency	maximal	maximal in linear problems
Difficulty	difficult	often solvable analytically
Goodness of Fit ?	No	Yes: e.g. χ^2-probability
Robustness	No	No

Hypthesis Tests

Hypothesis tests

Assess plausibility of a hypothesis using data

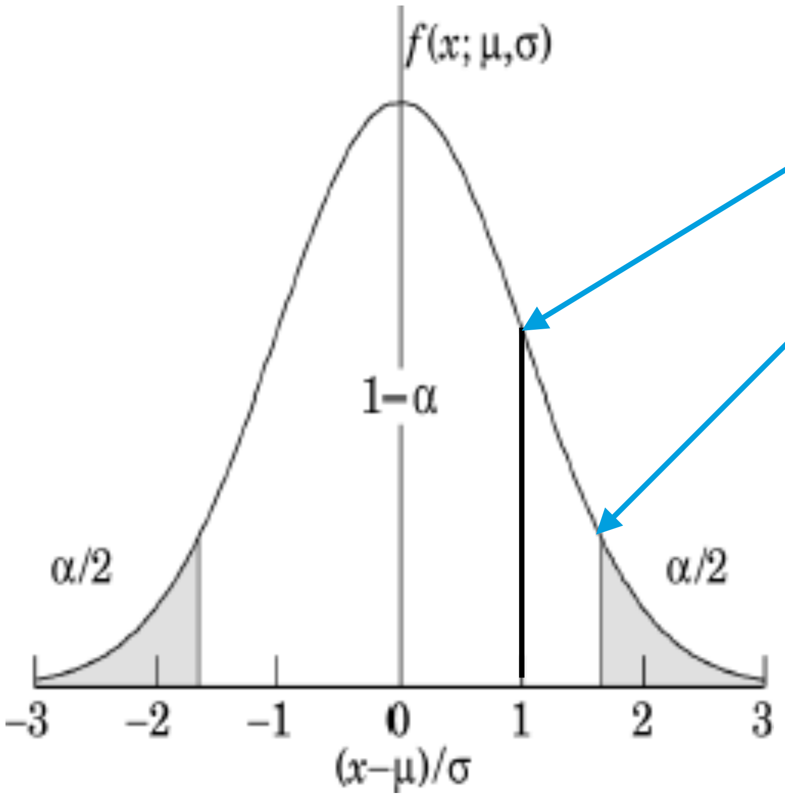
- Should I take an umbrella with me ?
- Is a therapy (medication) effective ?
- Is the discovered signal the Higgs boson predicted by the Standard Model ?



- Hypothesis test: do the data agree, within a pre-defined significance, with the hypothesis (theory) ?
  - Exclusion of hypothetical signals usually at 95% confidence level (p-value = 5%)
  - Discovery of signals requires bigger significance, typically 5σ (p-value ~ $3 \cdot 10^{-7}$)

"Extraordinary claims require extraordinary evidence"

Gaussian quantiles



PDG 2020:
Fig. 40.4

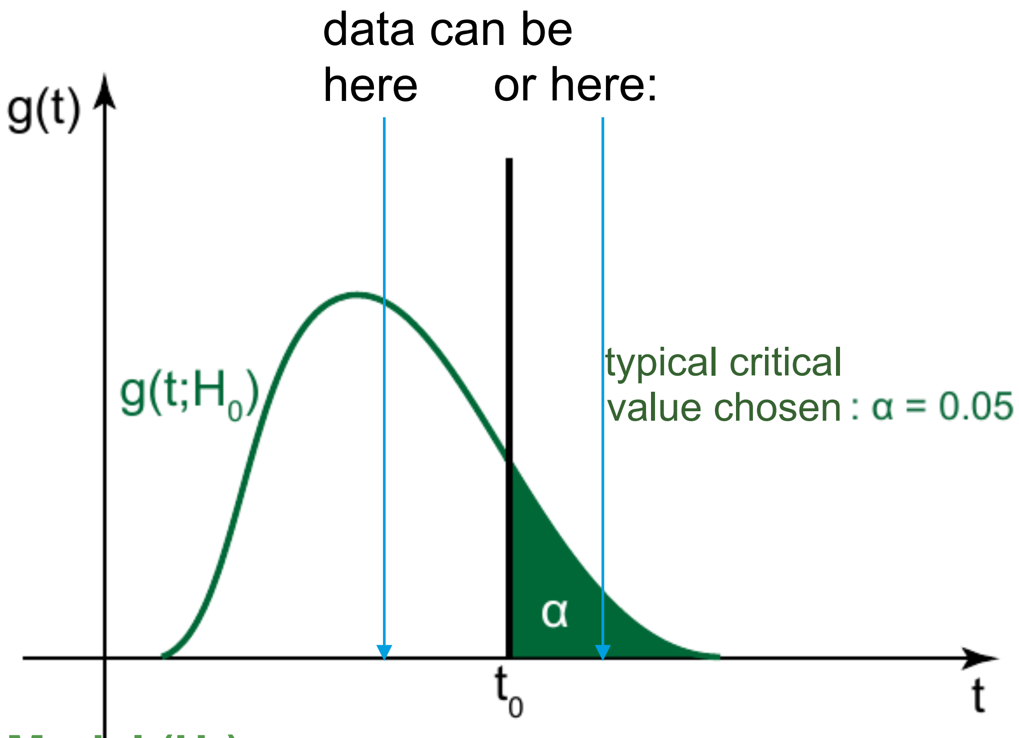
	$1-\alpha$	α	$\alpha/2$
1σ	0.683	0.317	0.158
1.65σ	0.90	0.10	0.05
1.96σ	0.95	0.05	0.025
2σ	0.9545	0.0455	0.0228
3σ	0.9973	0.0027	0.0013
5σ			3×10^{-7}

Measurements: 2-sided interval: *p*-value = α
Exclusion/discovery: 1-sided interval: *p*-value = α/2

- Hypothesis test: do the <u>data</u> agree, within a pre-defined significance, with the hypothesis (<u>theory</u>) ?
  - Exclusion of hypothetical signals usually at 95% confidence level (*p*-value = 5%)
  - Discovery of signals requires bigger significance, typically 5σ (*p*-value ~ $3\cdot10^{-7}$)

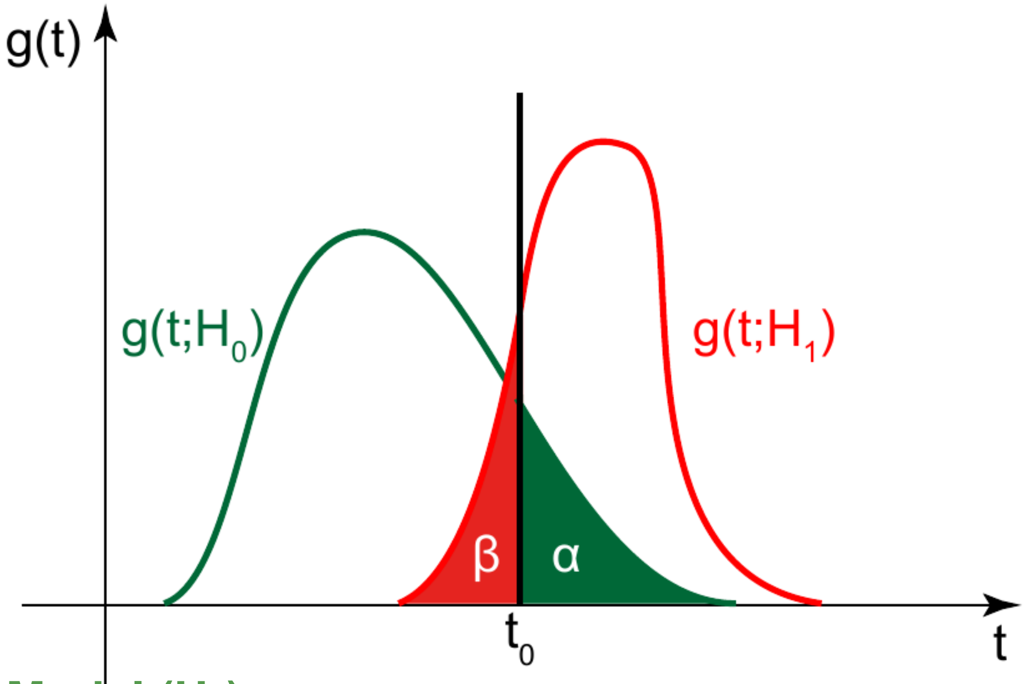
“Extraordinary claims require extraordinary evidence”

Hypothesis tests



- Hypotheses are formulated as PDF of a **test statistic t**
- Comparison of a data sample with one or several hypotheses H_i
- Single hypothesis: <u>null hypothesis H_0</u>
  - Example: test data for consistency with the **Standard Model (H_0)**
  - E.g. using goodness-of-fit tests using χ^2 as test statistic

Hypothesis tests



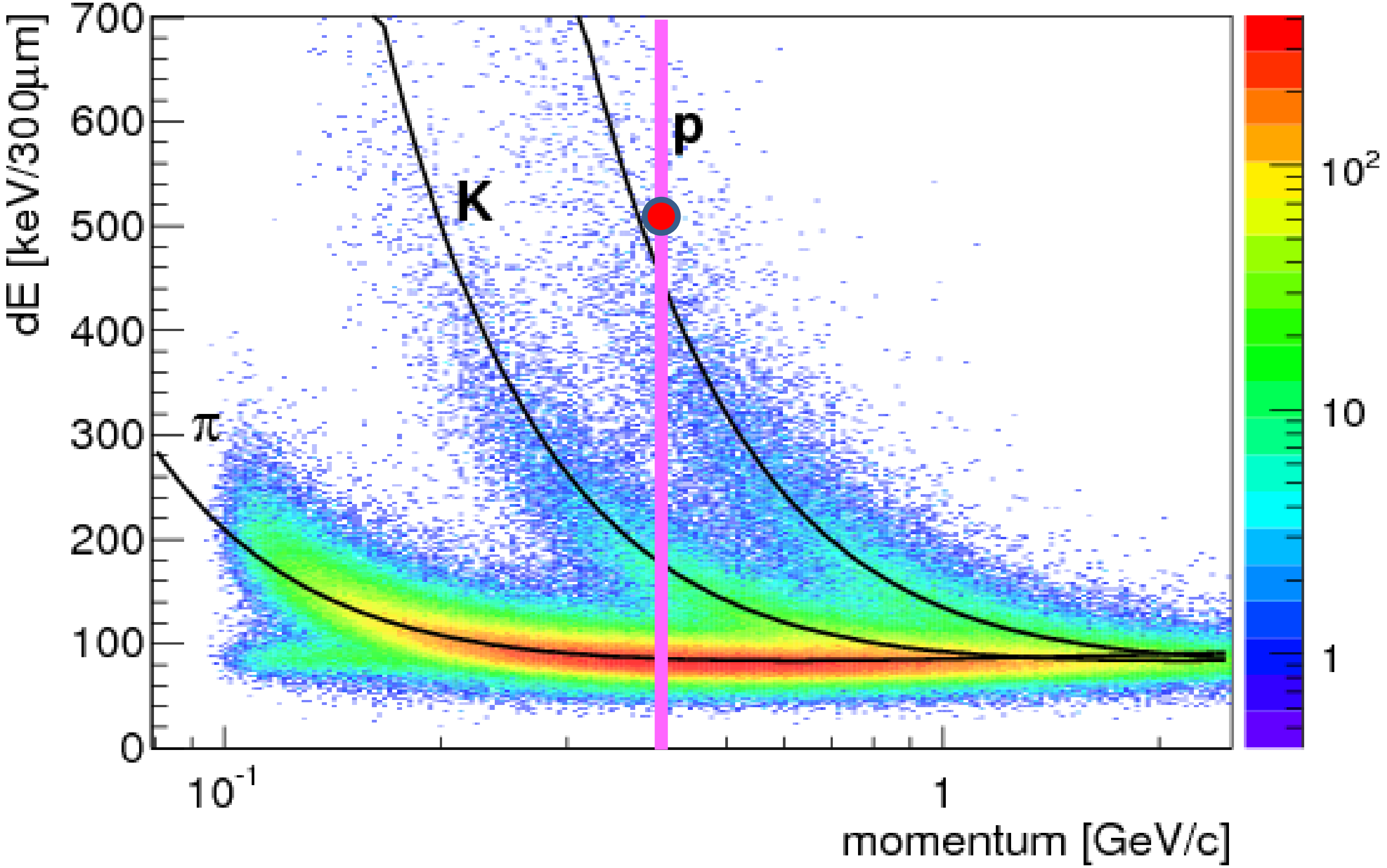
- Hypotheses are formulated as PDF of a **test statistic t**
- Comparison of a data sample with one or several hypotheses H_i
- Single hypothesis: <u>null hypothesis H_0</u>
  - Example: test data for consistency with the **Standard Model (H_0)**
  - E.g. using goodness-of-fit tests using χ^2 as test statistic
- Several hypotheses: H_0 and alternative hypotheses H_i
  - Example: **Standard Model (H_0)** vs **specific New Physics model (H_1)**.

A hypothesis can never be proven, but it can be falsified (usually require a 5σ significance)

Example: particle identification

Energy loss measurement

- Hypotheses H_i:
  - Pion: falsified
  - Kaon: falsified
  - Proton: consistent (but not proven)



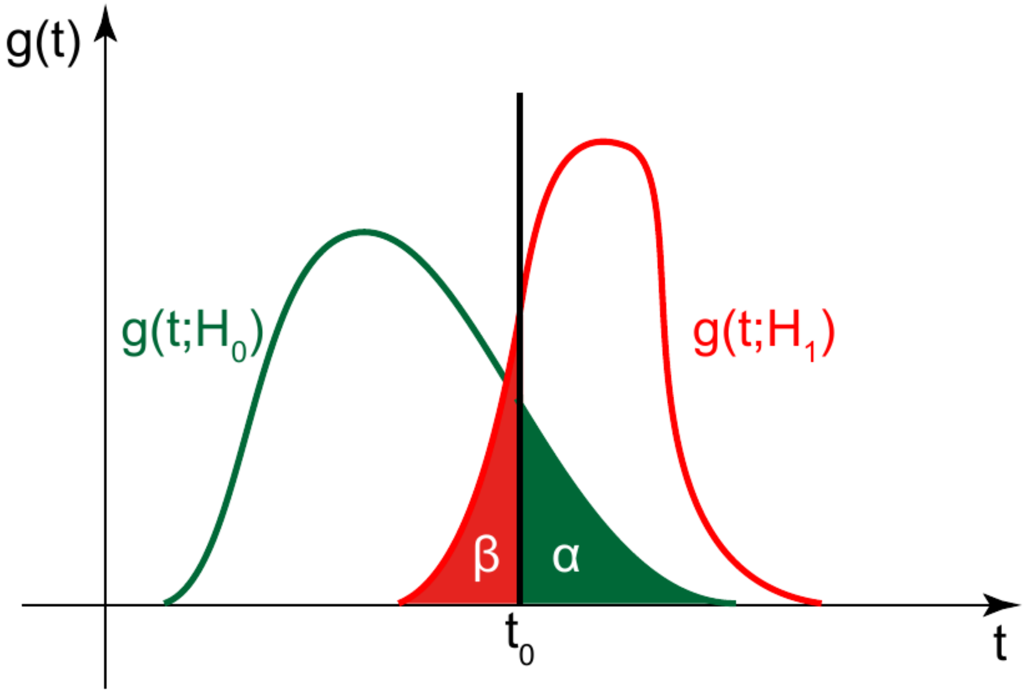
A hypothesis can never be proven, but it can be falsified (usually require a 5σ significance)

Hypothesis tests

Procedure

1. Determine PDF $g(t;H_i)$ for test statistic t
2. Define significance level α (typically 5%)
   - critical value t_0: reject null hypothesis or not
   - in practice, α depends on goal
     - high efficiency ε or high purity p ?

$$\epsilon = 1 - \alpha \qquad\qquad p = \frac{(1-\alpha)N_0}{(1-\alpha)N_0 + \beta N_1}$$



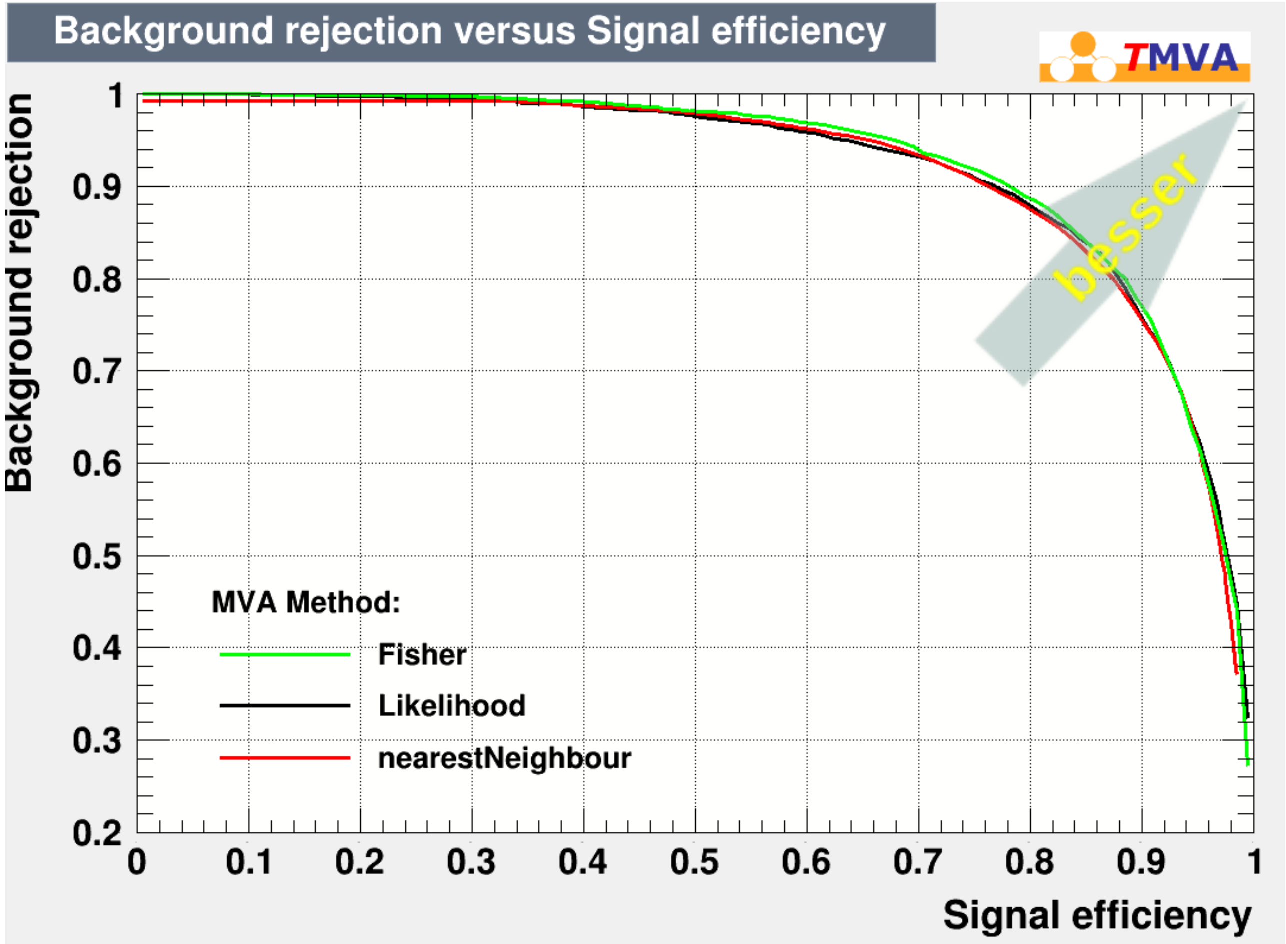
     - separation power: 1-β

Note: trivially, no separation if no separation power
=> large 1-β is fundamentally more important than small α

3. Determine p-value of the measurement

Receiver operating characteristic (ROC)

$$p = \frac{(1-\alpha)N_0}{(1-\alpha)N_0 + \beta N_1}$$



$$\epsilon = 1 - \alpha$$

- Choice of "working point” depends on problem (purity vs. efficiency)
- Area Under Curve (“AUC”) is often used to quantify the separation power

Two hypotheses

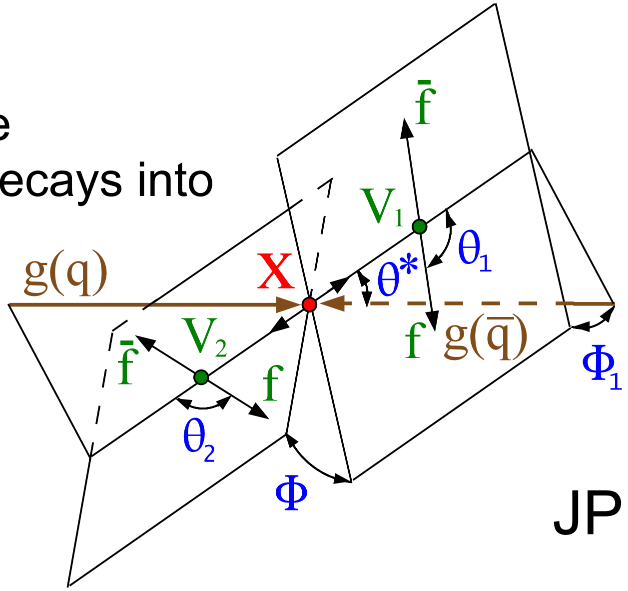
Example: Higgs boson properties

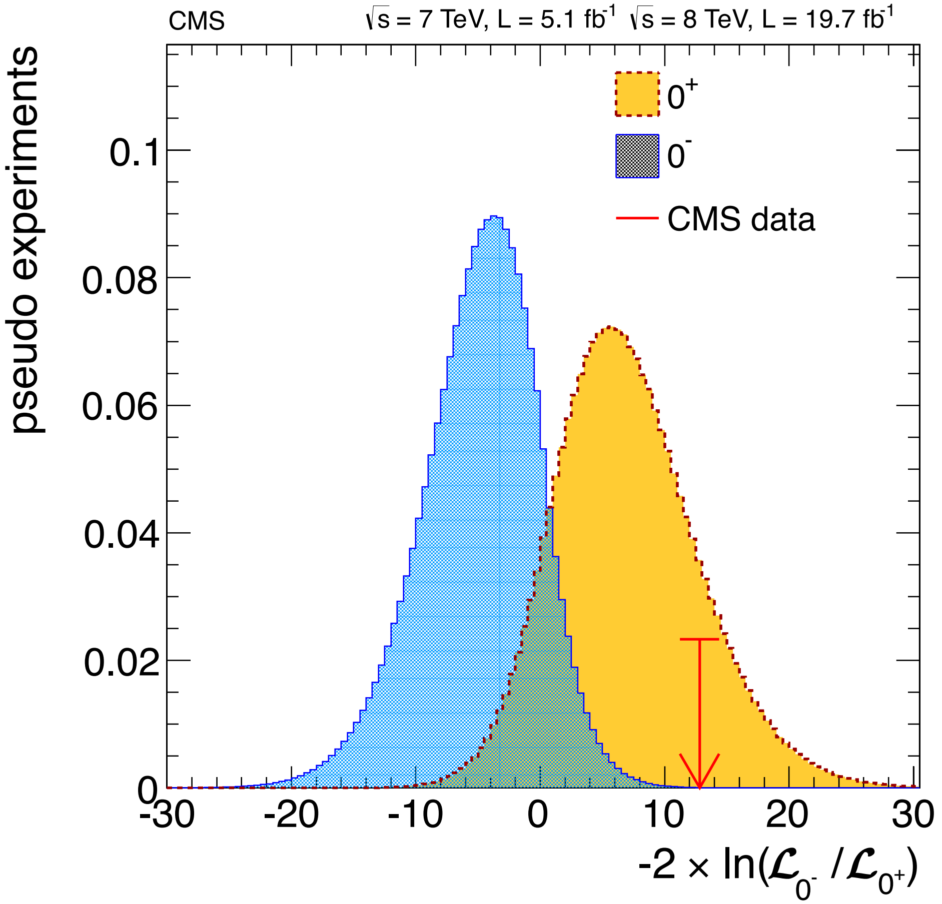
https://arxiv.org/abs/1312.5353

- Is the Higgs boson a scalar particle ?
  - Null hypothesis: $J^P = 0^+$
  - Alternative hypothesis: e.g. $J^P = 0^-$
- Use likelihood ratio as test statistic:

$$q = -2\ln(\mathcal{L}_{0-}/\mathcal{L}_{0+})$$

CP-properties of the Higgs boson from decays into in four leptons





JP = 0^- excluded at 3.8σ observed (2.4σ expected)

Neyman-Pearson lemma

Jerzy Neyman, Egon Pearson, 1933

- For simple hypotheses, i.e. $f(x|H_i)$ are completely known, the likelihood ratio λ(x) provides optimal separation power 1-β (for fixed significance α)

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

- Equivalently: log-likelihood difference:

$$q(x) = -2\ln\lambda(x) = 2(\ln f(x|H_1) - \ln f(x|H_0))$$

- Remarks:
  - Determination of optimal test statistic (signal-to-background separation) is called classification => Friday
  - In practice, MC simulations are used to determine PDF for different hypotheses.
  - The Neyman-Pearson lemma does not generally hold for composite hypotheses, i.e. hypotheses with free parameters, e.g.: $f(x|H(\lambda_i,\mu_i))$ with λ_i known und μ_i free

Wilks' theorem

- For large samples with n data points x_i, $n \to \infty$ (and for a null hypothesis H_0 that determines r=m-m(0) parameters), the distribution of the log-likelihood ratio $q = -2 \ln \lambda$ asymptotically approaches a χ^2 distribution (with r degrees of freedom).
  - r = difference in the number of free parameters for H_1 and H_0

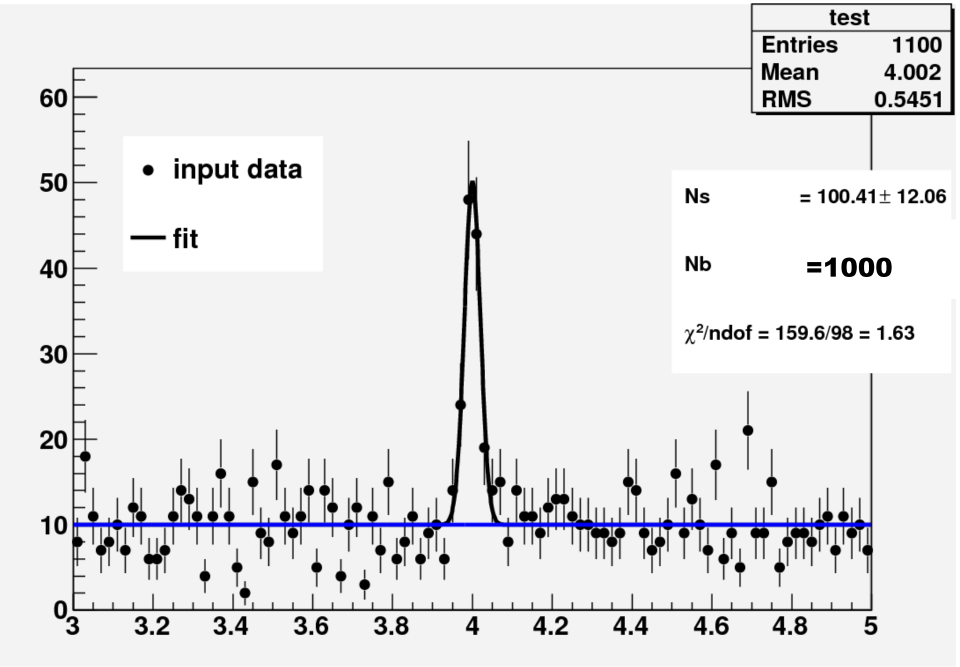
S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. 9, 60–62 (1938)

$$\Delta\chi^2 = -2\ln\lambda = -2\ln\left(\frac{\mathcal{L}(s+b)}{\mathcal{L}(b)}\right) \quad \begin{matrix} H_1 \\ H_0 \end{matrix}$$

Wilks’ theorem

Counting Experiment

- Signal (s) above background (b):
- PDF for each bin in m: n(m) = b(m) + s(m)
  - b: Poisson distributed in each bin -> Gauss for large b
  - s: Number of events in mass peak (fixed mass and width)



- Two hypotheses:
  - H_1 signal-Hypothesis: s≠0 => fit of 2 free parameters b+s → χ^2(b+s)
  - H_0 (background only): s=0 => fit of 1 free parameter b → χ^2(b)

$$\Delta\chi^2 = -2\ln\lambda = -2\ln\left(\frac{\mathcal{L}(s+b)}{\mathcal{L}(b)}\right)$$ = 73 (in this specific case)

Apply Wilks’ theorem:
If H_0 true, then $\Delta\chi^2$ is a χ^2 - distribution with 1 d.o.f: $p(\chi^2 = 73) = 2\text{x}10^{-16}$, corresponds to z = 8.5 σ

In the backup: for small signals and large *n*: $z = \sqrt{\Delta\chi^2} = \sqrt{q} = s/\sqrt{b}$

Cowan, Cranmer, Gross and Vitelis “Asymptotic formulae for likelihood-based tests of new physics” https://arxiv.org/abs/1007.1727

Confidence Intervals

Frequentist and Bayesian approaches

- **Frequentist**: also referred to as "objective" or "classical"
  - Repeatable events, predictions and/or symmetries (e.g. dice, QM)
  - Probability is identified as rate of occurrence (relative frequency) of events
  - For a confidence level CL = *p*%, the confidence interval covers the true value in *p*% of all cases.

    => Neyman construction of the confidence interval

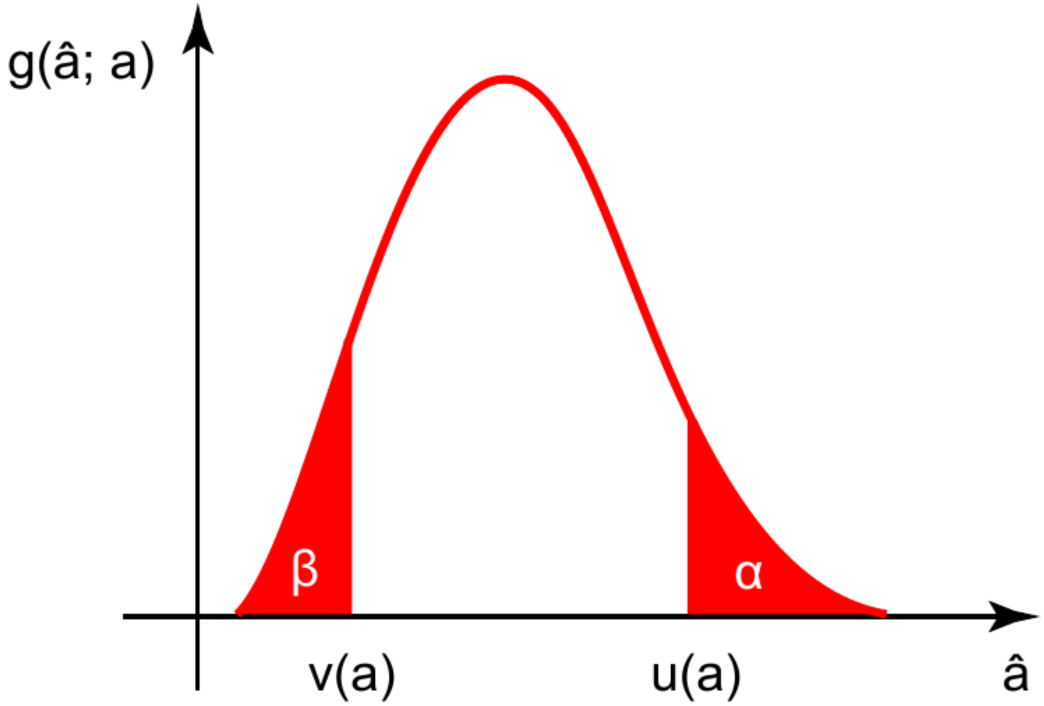
- **Bayesian**: also referred to as "subjective"
  - "Degree of Belief"
  - Applicable also to one-time only events, e.g. probability that it is going to rain tomorrow
  - Priors often consist of non-Frequentist assumptions
  - The posterior density distribution of a, namely f($a|\hat{a}$), is product of the likelihood $\mathcal{L}(\hat{a}|a)$ and the prior $\pi(a)$

$$f(a|\hat{a}) \propto \mathcal{L}(\hat{a}|a) \cdot \pi(a)$$

Frequentist confidence interval

Coverage

- Use measurement of â and uncertainty to determine interval in which the true value a lies for a chosen confidence level (CL)
- Typical CL: 68.3%, 90% or 95%.



Coverage:
probability 1-α-β that true value is contained in the interval

Frequentist confidence interval

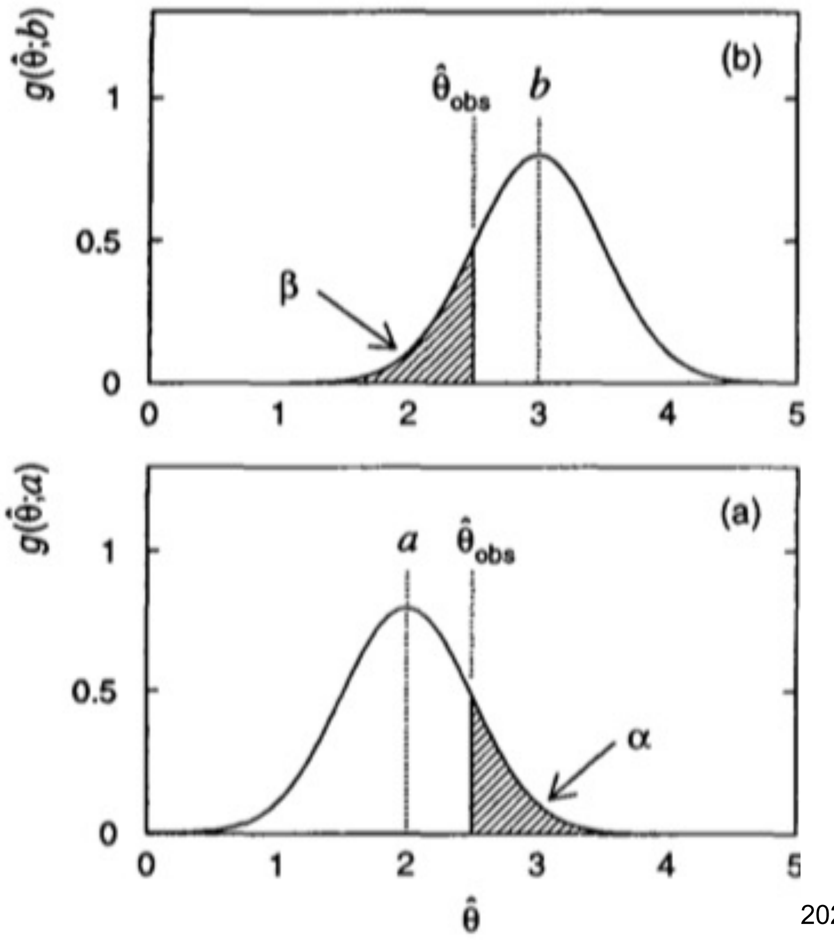
Cowan page 122

Example: Gaussian distribution

- Measurement of a data point $\hat{\theta}_{obs}$ of an observable $\hat{\theta}$ (detector has Gaussian response)
- Construction of a two-sided confidence interval:

Upper limit *b*
For assumed true value *b*, the probability to measure a value $\hat{\theta}_{obs}$ or smaller is β,
e.g. for 1σ: β = (1 – 68%)/2 = 16%

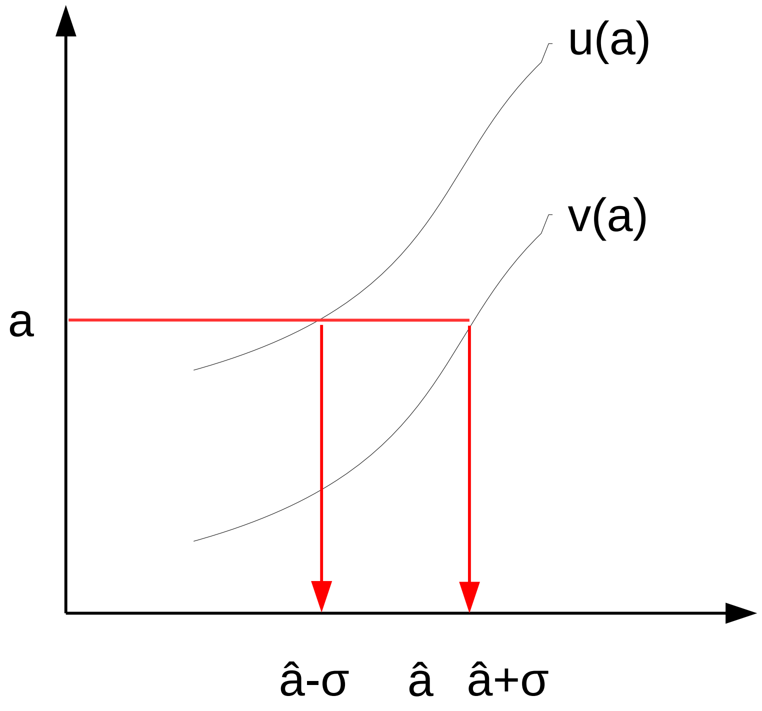
Lower limit *a*
For assumed true value *a*, the probability to measure a value $\hat{\theta}_{obs}$ or bigger is α,
e.g. for 1σ: α = (1 – 68%)/2 = 16%



Neyman construction

Frequentist approach

- For a true value of *a*, there is a measurement *â* with an uncertainty σ.
- *â*-σ und *â*+σ are functions of *a* (here u(*a*) and v(*a*)).
- A confidence belt is constructed for assumed true values of *a*.



In 16% of cases measure < â-σ

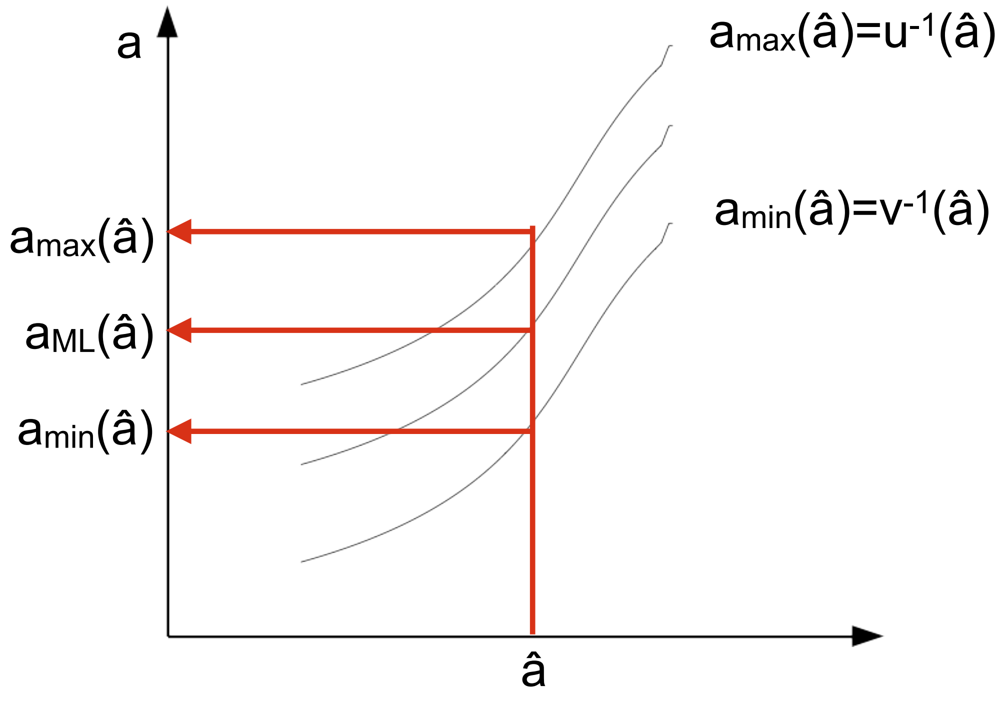
In 16% of cases measure > â+σ

Neyman construction

Frequentist approach

- For a true value of a, there is a measurement $\hat{a}$ with an uncertainty σ.
- $\hat{a}-\sigma$ und $\hat{a}+\sigma$ are functions of a (here $u(a)$ and $v(a)$).
- A confidence belt is constructed for assumed true values of a.
- For a concrete measurement $\hat{a}$, a confidence interval $[a_{min}, a_{max}]$ is determined (vertical axis)

16% of cases:
true value > a_{max}



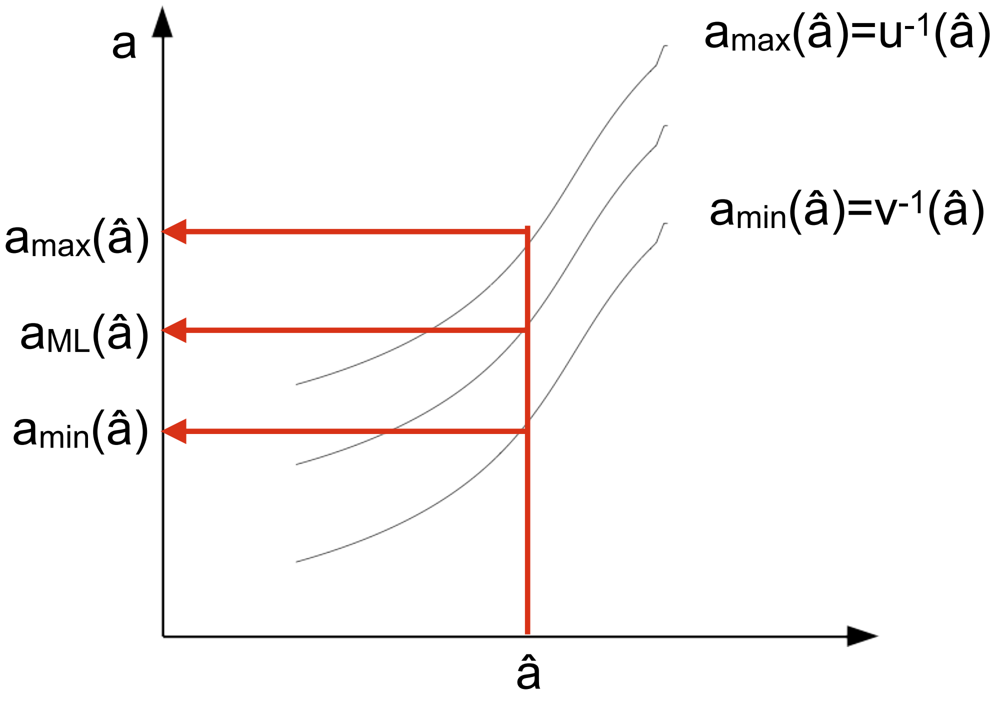
In 16% of cases
true value < a_{min}

Neyman construction

Frequentist approach

- For a true value of *a*, there is a measurement *â* with an uncertainty σ.
- *â*-σ und *â*+σ are functions of *a* (here u(*a*) and v(*a*)).
- A confidence belt is constructed for assumed true values of *a*.
- For a concrete measurement *â*, a confidence interval [a_{min}, a_{max}] is determined (vertical axis)
- Note: the confidence interval is an estimate

16% of cases:
true value > a_{max}



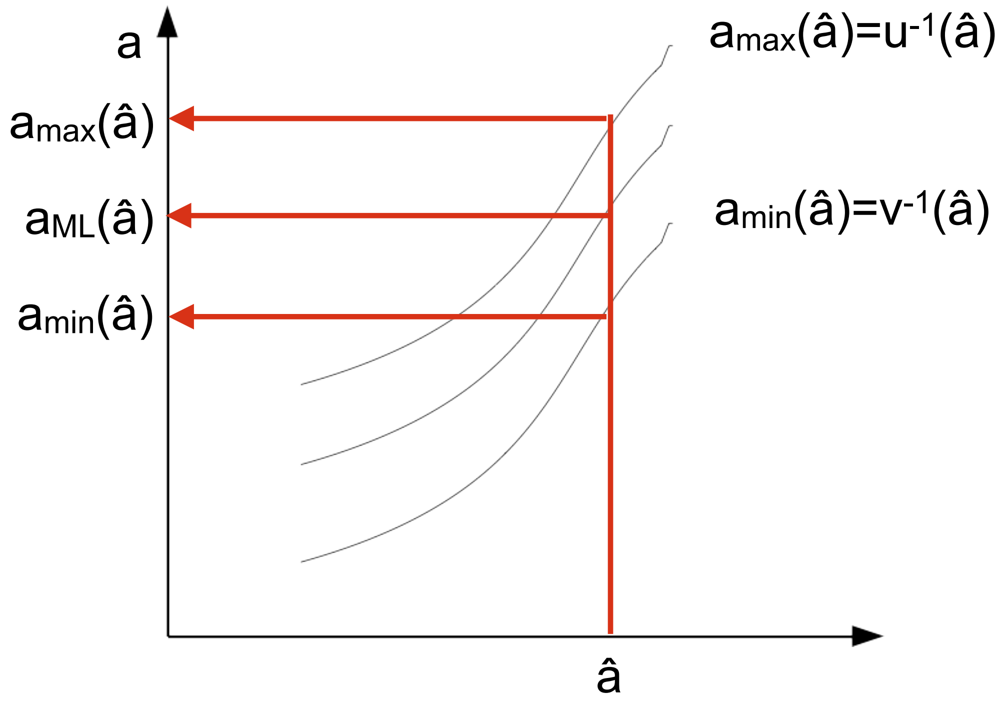
In 16% of cases
true value < a_{min}

Neyman construction

Frequentist approach

- For a true value of a, there is a measurement $\hat{a}$ with an uncertainty σ.
- $\hat{a}$-σ und $\hat{a}$+σ are functions of a (here u(a) and v(a)).
- A confidence belt is constructed for assumed true values of a.
- For a concrete measurement $\hat{a}$, a confidence interval [a_{min}, a_{max}] is determined (vertical axis)
- Note: the confidence interval is an estimate

16% of cases:
true value > a_{max}



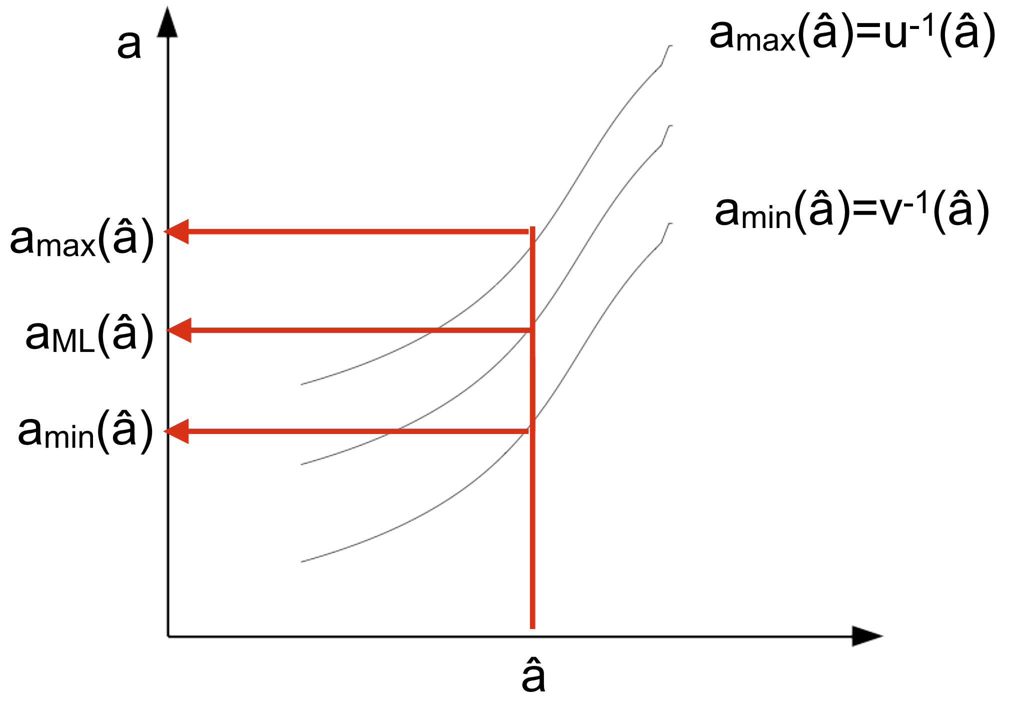
In 16% of cases
true value < a_{min}

Neyman construction

Frequentist approach

- For a true value of a, there is a measurement $\hat{a}$ with an uncertainty σ.
- $\hat{a}$-σ und $\hat{a}$+σ are functions of a (here u(a) and v(a)).
- A confidence belt is constructed for assumed true values of a.
- For a concrete measurement $\hat{a}$, a confidence interval [a_{min}, a_{max}] is determined (vertical axis)
- Note: the confidence interval is an estimate

16% of cases:
true value > a_{max}

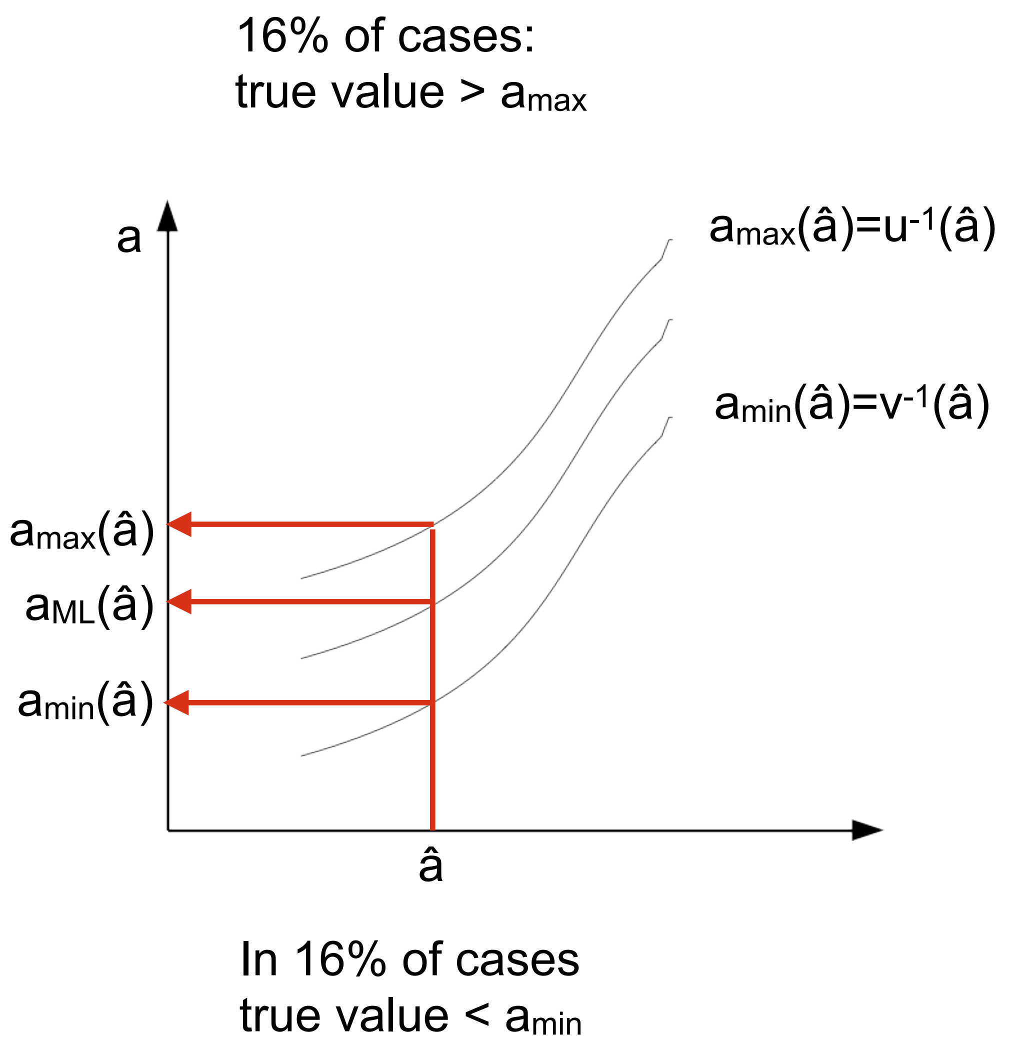


In 16% of cases
true value < a_{min}

Neyman construction

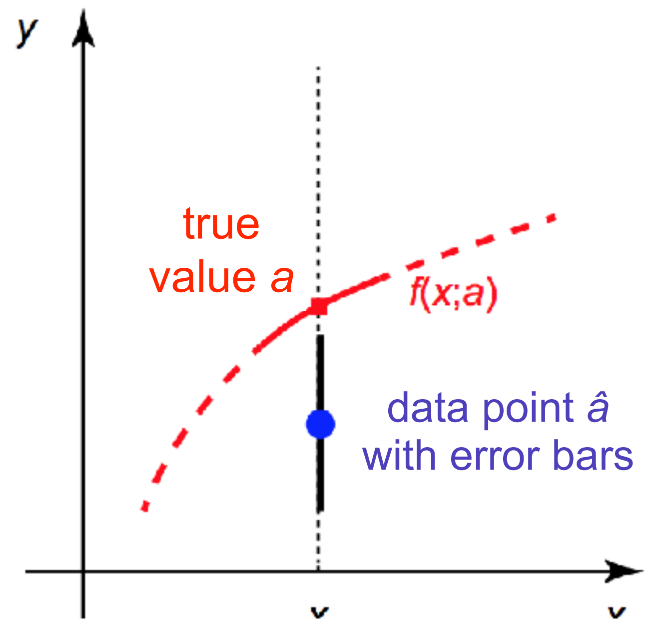
Frequentist approach

- For a true value of *a*, there is a measurement *â* with an uncertainty σ.
- *â*-σ und *â*+σ are functions of *a* (here u(*a*) and v(*a*)).
- A confidence belt is constructed for assumed true values of *a*.
- For a concrete measurement *â*, a confidence interval [a_{min}, a_{max}] is determined (vertical axis)
- Note: the confidence interval is also an estimate

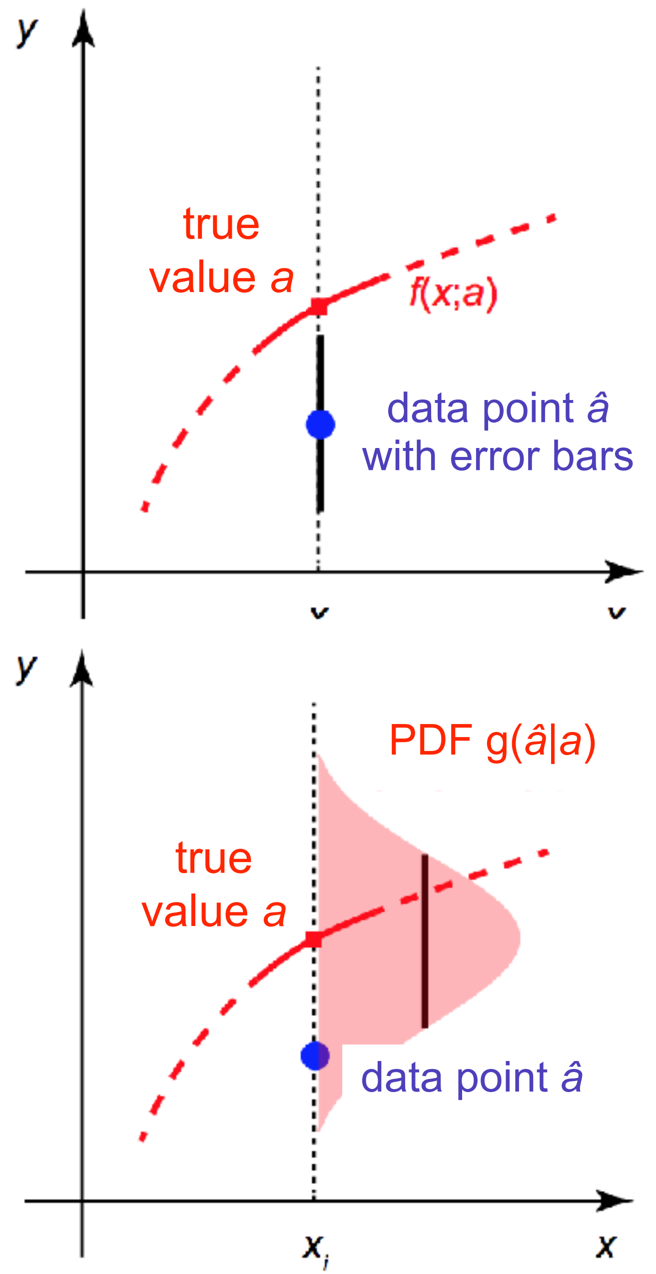


Frequentist confidence interval

- Usual presentation of measurements:
  - Estimator with uncertainty: $\hat{a} \pm \sigma_a$
- Interpretation:
  - The interval $[\hat{a} - \sigma_a, \hat{a} + \sigma_a]$ covers the true value a at 68.3% confidence.



Frequentist confidence interval



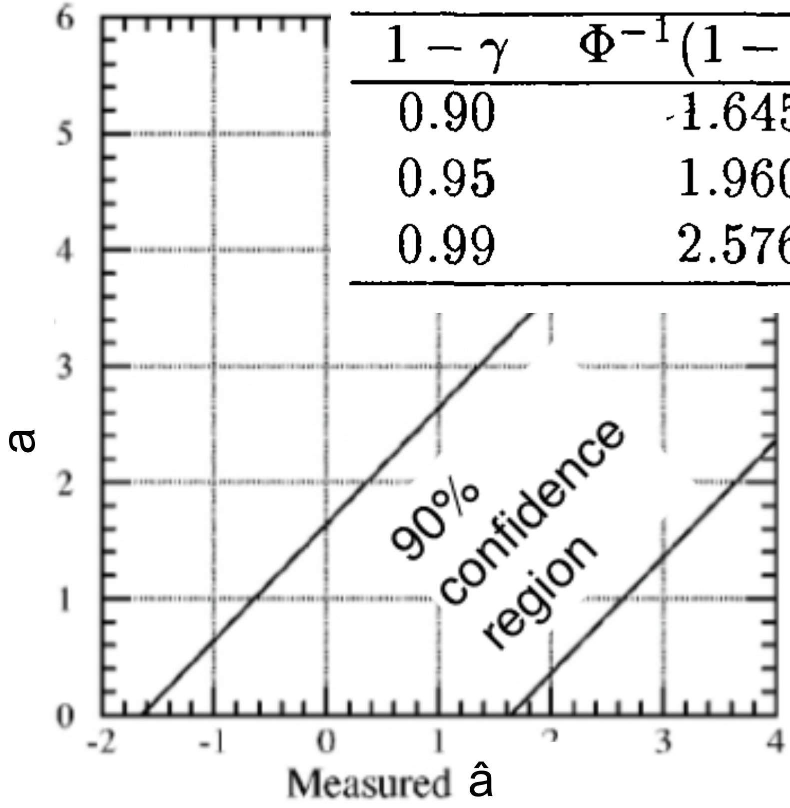
- Usual <u>presentation</u> of measurements:
  - Estimator with uncertainty: $\hat{a} \pm \sigma_a$
- Interpretation:
  - The interval $[\hat{a} - \sigma_a, \hat{a} + \sigma_a]$ covers the true value a at 68.3% confidence.
- Actual meaning:
  - The measured parameter $\hat{a}$ is a random number, given the true value a.
  - PDF $g(\hat{a}|a)$ ist distributed around the true value a.
- Both are equivalent if $g(\hat{a}|a)$ is a Gaussian.
  - This is frequently the case (→ central limit theorem), but not always

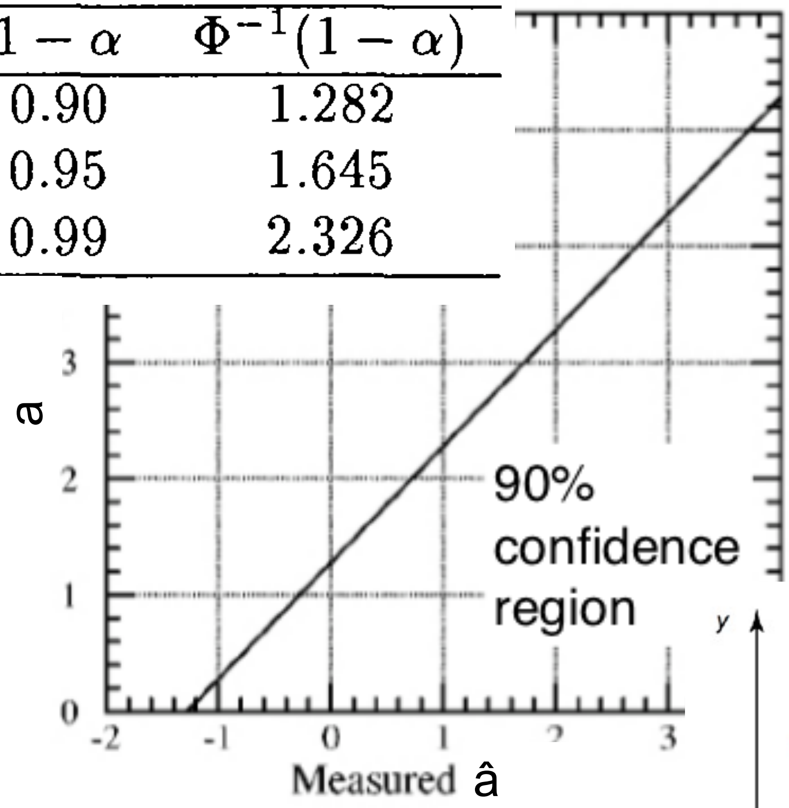
Confidence belt

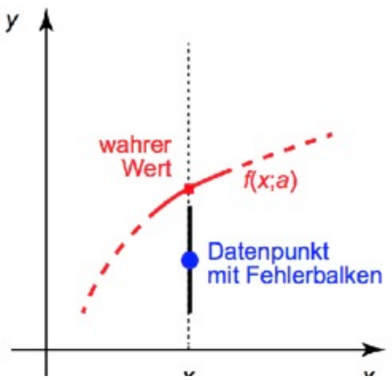
For a Gaussian distribution

Cowan: table 9.2

$1-\gamma$	$\Phi^{-1}(1-\gamma/2)$	$1-\alpha$	$\Phi^{-1}(1-\alpha)$
0.90	1.645	0.90	1.282
0.95	1.960	0.95	1.645
0.99	2.576	0.99	2.326







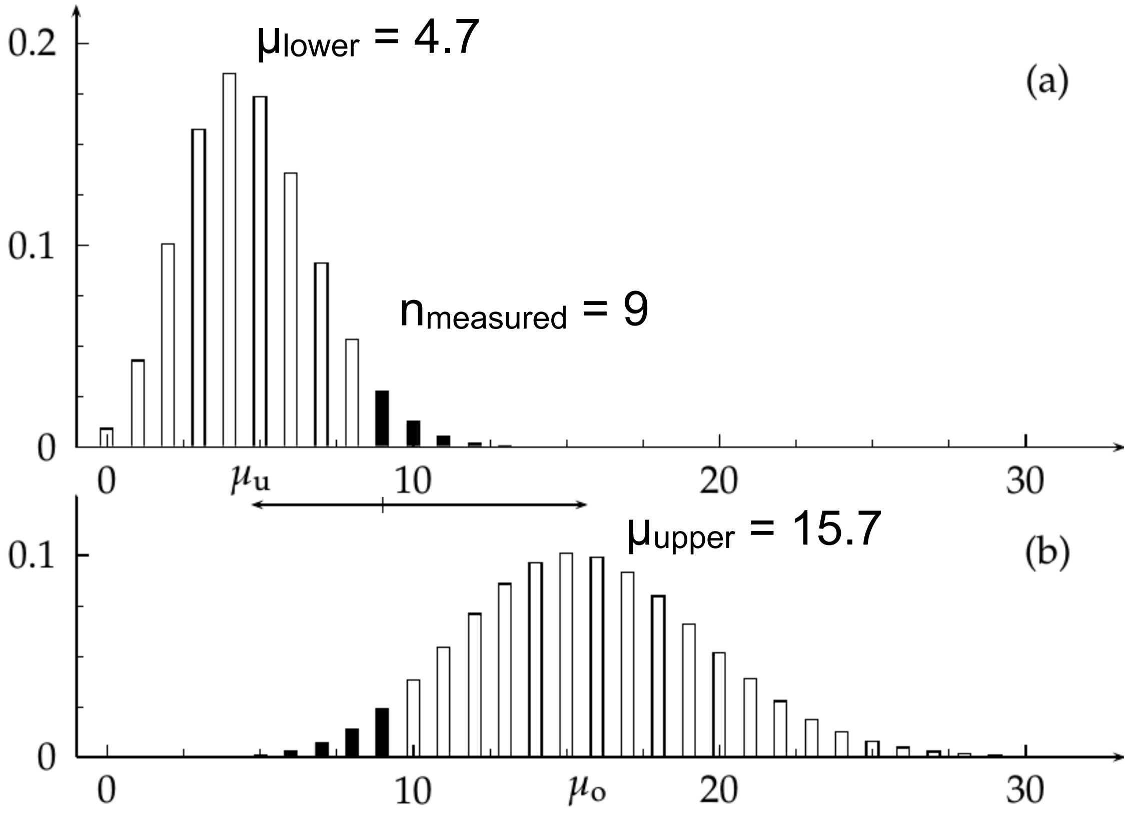
- The confidence belt of a Gaussian PDF is a straight line with a slope of 1.
  - Gauss PDF is symmetric and σ (width) does not depend on μ (mean).
  - This is why it is ok to draw the error bar <u>on</u> the data point, and to interpret it as interval for the true value

Frequentist confidence interval

Blobel/Lohrmann, Section 6.7

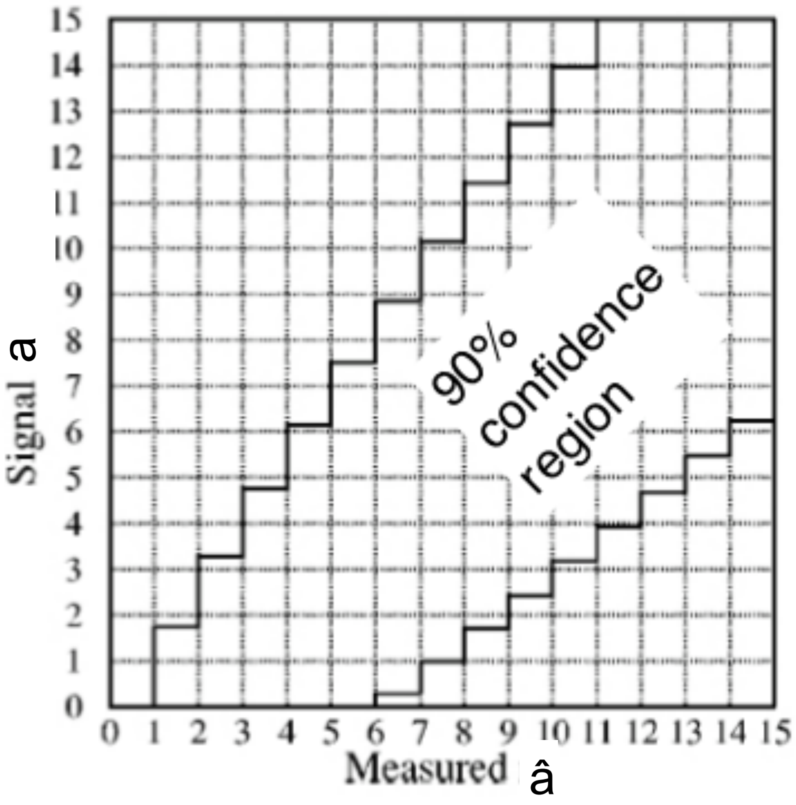
Example: Poisson distribution

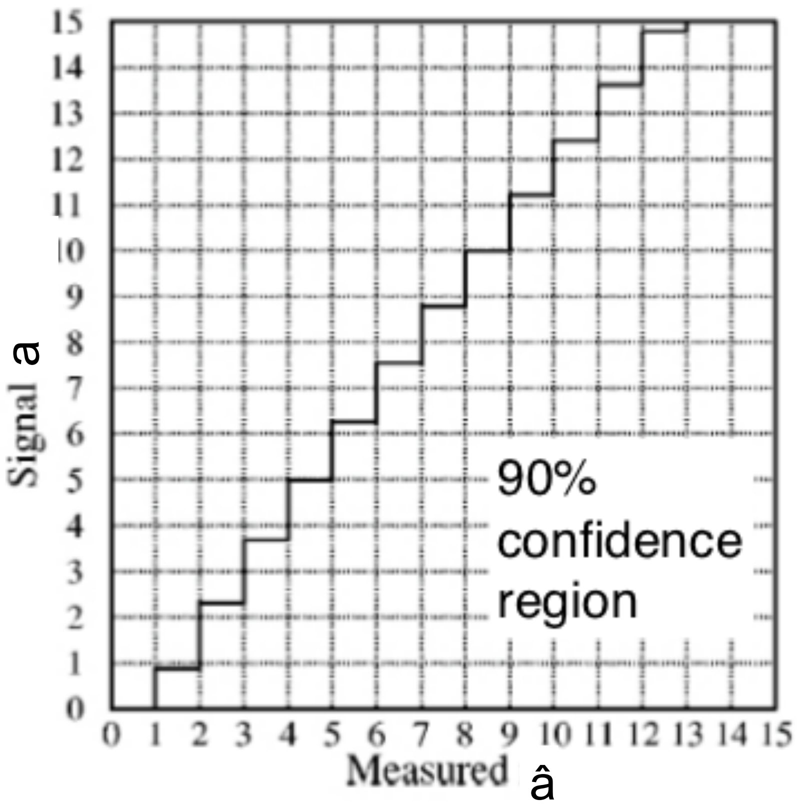
- Determine two-sided 90% confidence interval in a counting experiment with n = 9 observed events
- Poisson probability: $p(n|\mu) = e^{-\mu} \mu^n / n!$
- For a 95% CL, 1-sided interval, the interval border is determined by varying the <u>hypothetical true</u> value μ such that the <u>observed signal</u> is excluded with a p-value of 5%.
- Do this from both sides to obtain the 2-sided 90% CL interval Here: [4.7, 15.7]



Confidence belt

Poisson distribution



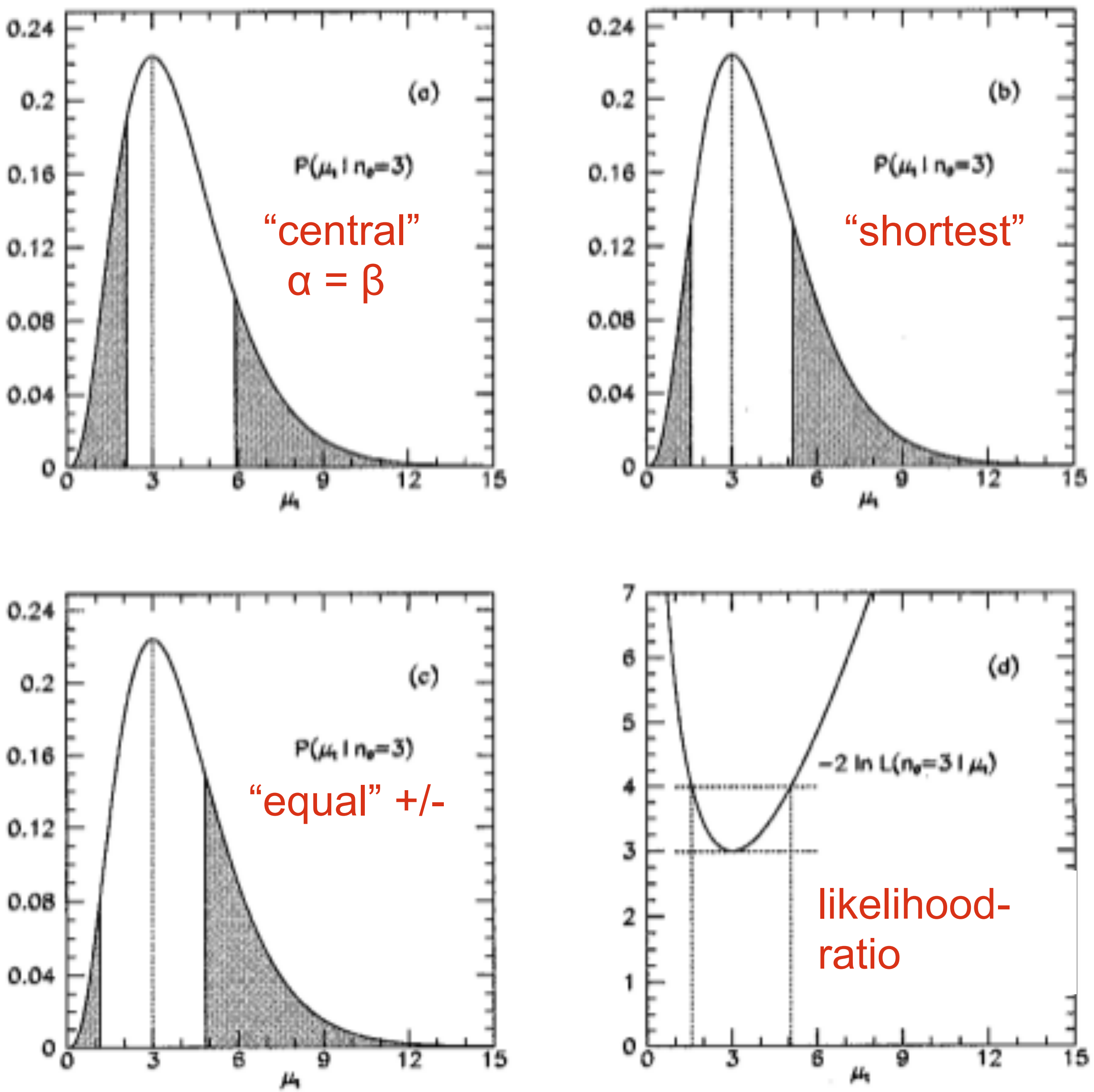


- 90% CL interval for an unknown Poisson-distributed signal with a background of 3 events
- In this case, the band for *â*=0 is empty.

Ordering principle

Example: Intervals for n=3 observed events

- The particular choice of interval borders is determined by an “ordering principle”

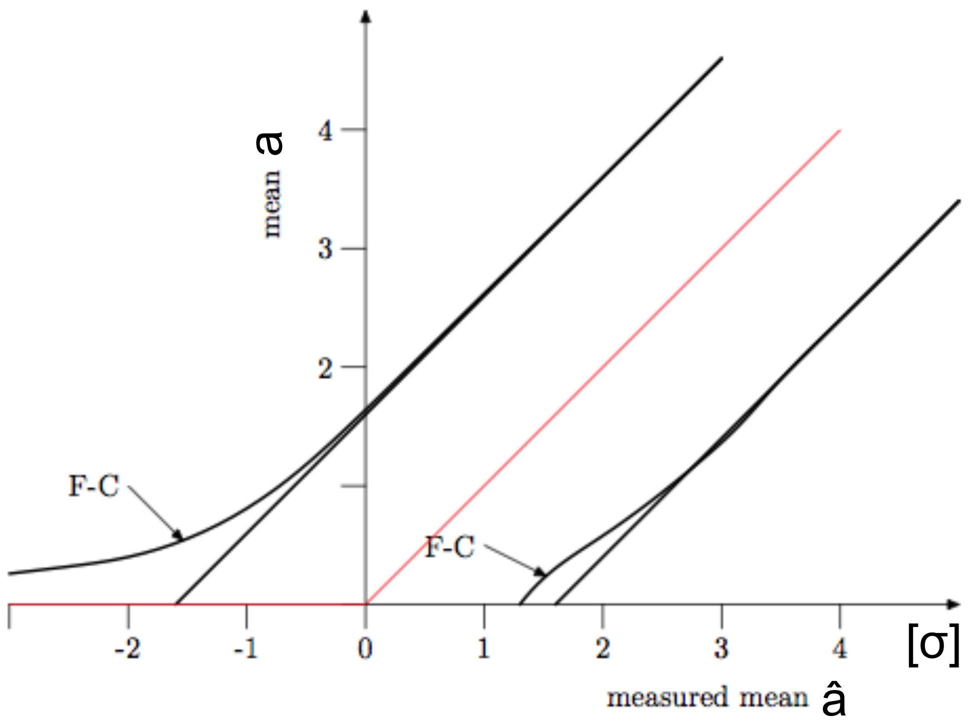


R. Cousins, „Why isn't every Physicist a Bayesian?“

1-sided limits and 2-sided intervals: "Unified Approach"

F.James. Abb. 9.8

- Feldman-Cousins a.k.a. "unified approach": "automatic" decision if measurement or limit



- Construct interval using an ordering principle, based on the likelihood ratio R

$$R(\hat{a}|a) = \frac{g(\hat{a}|a)}{g(\hat{a}|a_{\text{best}})}$$

where $a_{\text{best}} = a$ for which $g(\hat{a}|a)$ is largest

- Recipe:
  - Sum up values of $\hat{a}$ for decreasing values of R until $g(\hat{a}|a)$ reaches the chosen confidence level
  - For $\hat{a} < 0$: add contributions to the left side (no empty interval)

Example in backup

Summary

- **Frequentist**: there is a true value *a*
  - True values are true, they have no uncertainty (!)
  - The interval is a measured (i.e. random) quantity => probability and uncertainty attributed to the interval
  - For a confidence level CL = *p*%, the confidence interval covers the true value in *p*% of all cases
  - Neyman construction to determine interval around true value: coverage by construction

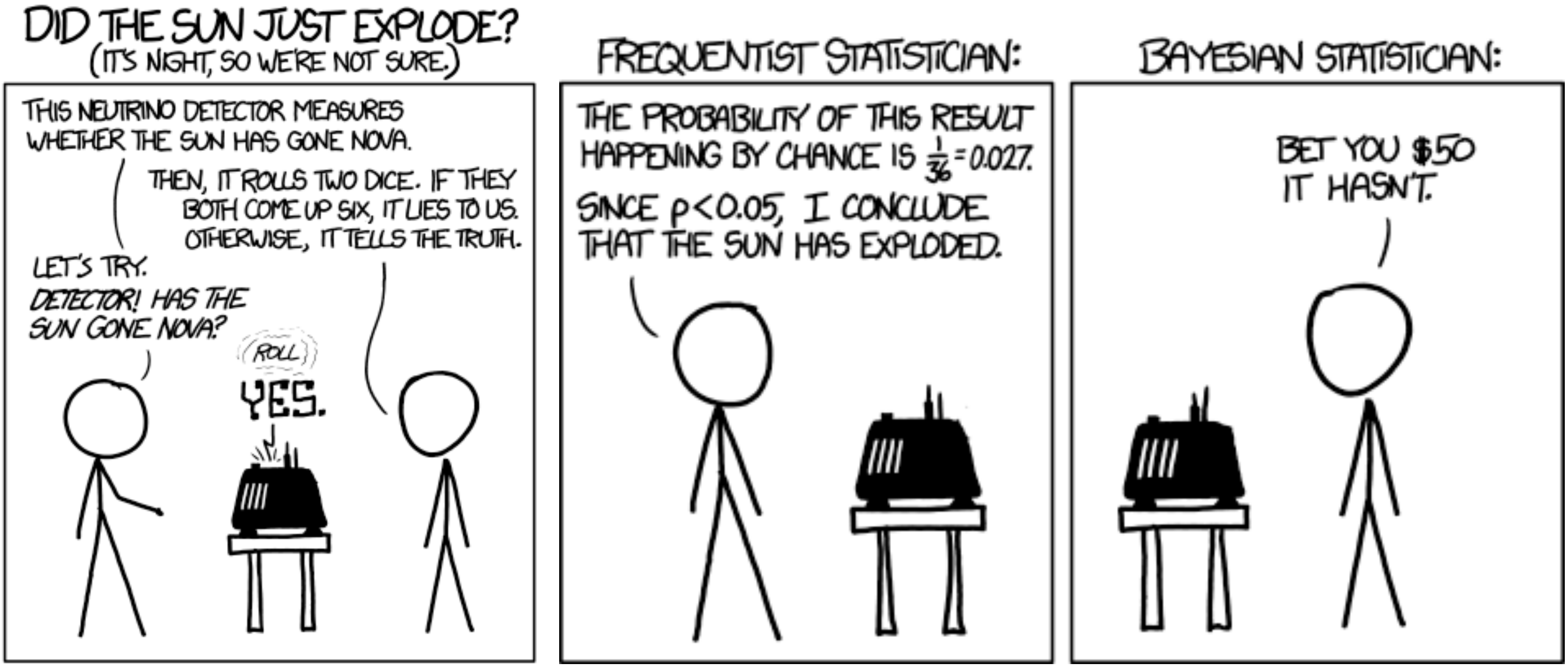
- **Bayesian**: depends on the conditions
  - The “prior” describes the degree of belief that *a* can take certain values
  - The true value has an uncertainty that depends on the measurement
  - The posterior density distribution of *a*, namely $f(a|\hat{a})$, is product of the likelihood $\mathcal{L}(\hat{a}|a)$ and the prior $\pi(a)$

$$f(a|\hat{a}) \propto \mathcal{L}(\hat{a}|a) \cdot \pi(a)$$

  - Coverage must be checked explicitly (e.g. using toy-MC)

Frequentist and Bayesian approaches

https://xkcd.com/1132/



Impact of the prior

Example: Bayesian intervals

$$f(a|\hat{a}) \propto \mathcal{L}(\hat{a}|a) \cdot \pi(a)$$

$\pi(a)$ = const

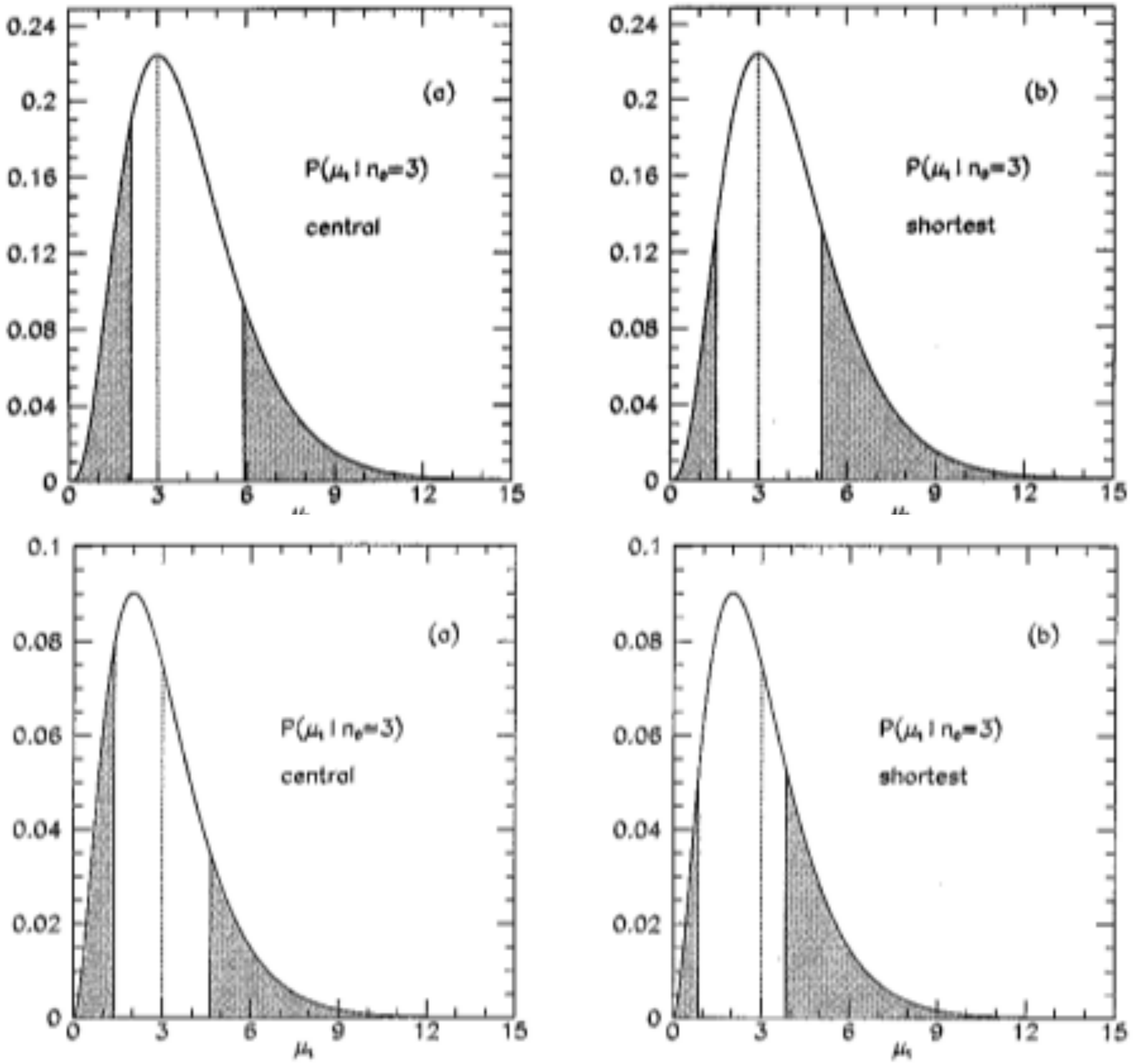
- For $\pi(a)$ = const → Bayesian $f(a|\hat{a})$ = Frequentist $\mathcal{L}(\hat{a}|a)$

$\pi(a) \propto 1/\mu$

Choice of prior is important!

central: α=β

shortest



R. Cousins, „Why isn't every Physicist a Bayesian?“

Poisson signal and background

No prior: “Frequentist”

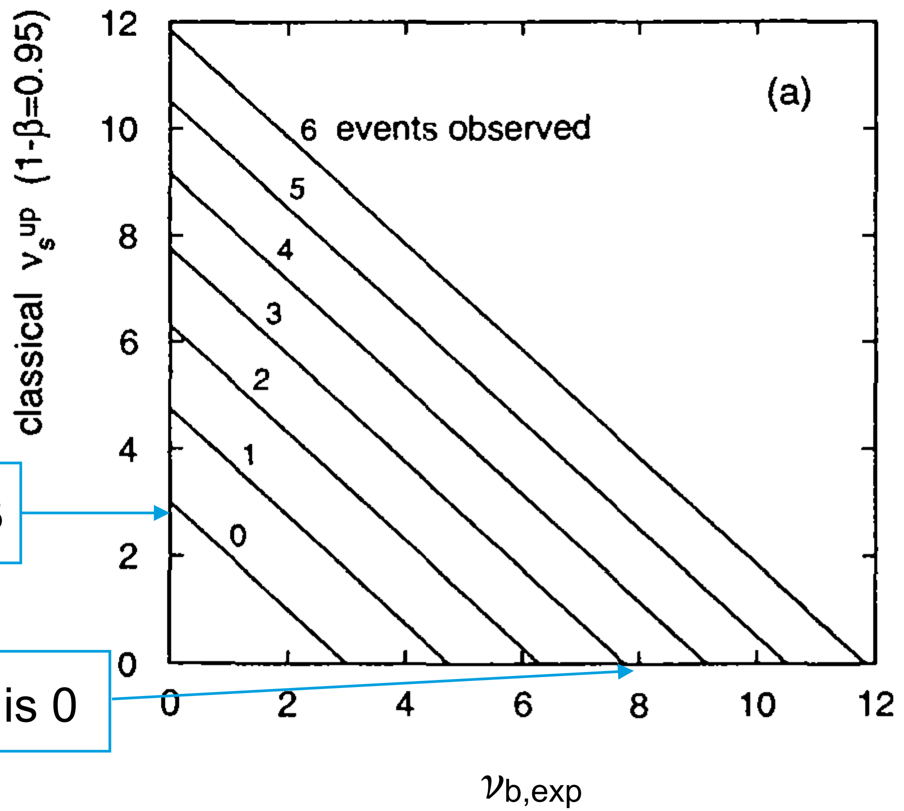
- Typical search analysis, i.e. number of signal events ν_s is small $\hat{\nu}_s = n - \nu_b$
  - Signal + background is Poisson distributed: $p(n|\ \nu_s, \nu_b) = p(n|\nu = \nu_s + \nu_b)$
  - Determine n and subtract $\nu_{b,exp}$ to estimate ν_s
  - Upper limit (95% CL) for ν_s as a function of the expected background $\nu_{b,exp}$, for different n_{obs}

- No positive limit for n_{obs} small against ν_b:

  Experiment with large background could be lucky and measure better limit

$\nu_{b,exp}$ = 0 and 0 observed => upper limit $\nu_{s,up}$ is 3

$\nu_{b,exp}$ = 8 and 3 observed, i.e. ν_s = -5 => upper limit $\nu_{s,up}$ is 0



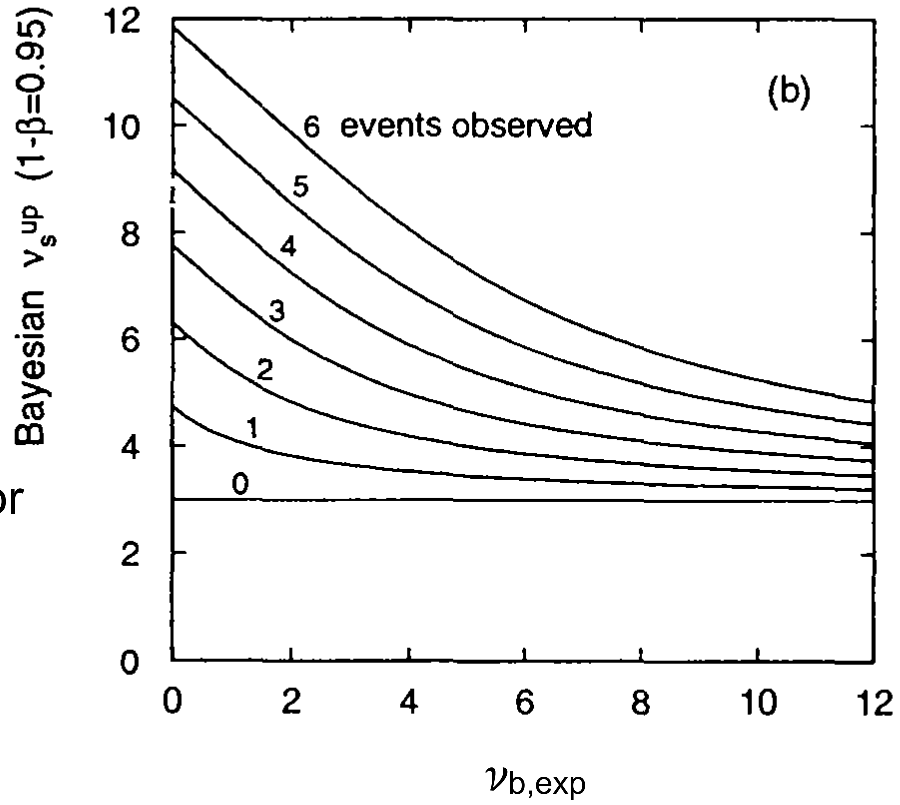
Cowan Fig. 9.9

Poisson signal and background

Prior: "Bayesian"

- Typical search analysis, i.e. number of signal events ν_s is small $\hat{\nu}_s = n - \nu_b$
  - Signal + background is Poisson distributed: $p(n|\nu_s, \nu_b) = p(n|\nu = \nu_s + \nu_b)$
  - Determine n and subtract $\nu_{b,exp}$ to estimate ν_s
  - Upper limit (95% CL) on ν_s as a function of the expected background $\nu_{b,exp}$, for different n_{obs}

- Bayesian prior:
  $\pi(\nu_s < 0) = 0$ and $\pi(\nu_s \geq 0) = \text{const}$
  has good properties:
  - For $\nu_b = 0$: same limit on ν_s
  - For $\nu_b > 0$: higher (i.e. worse) limit on ν_s than flat prior



Cowan Fig. 9.9

"Modified Frequentist approach": CL_S

A Frequentist counter measure

- Consider two hypotheses:
  - H_1: Measured event sample contains both background and signal
    - d = s + b → *p*-value = "CL_{s+b}"
  - H_0: Measured event sample contains just background
    - d = b (i.e. s=0) → *p*-value = "CL_b"

$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{\sum_{k=0}^{d} P(k; s+b)}{\sum_{k=0}^{d} P(k; b)}$$

- Make experiments with different background conditions comparable.
  - CL_s renormalizes measured limit to the background estimate
  - Quantitatively similar effect as Bayesian prior
  - CL_s is always bigger than CL_{s+b} → over-coverage

G.Cowan, PDG, Section 40.4.2.4

Also: T.Junk or A.L.Read

CL_s

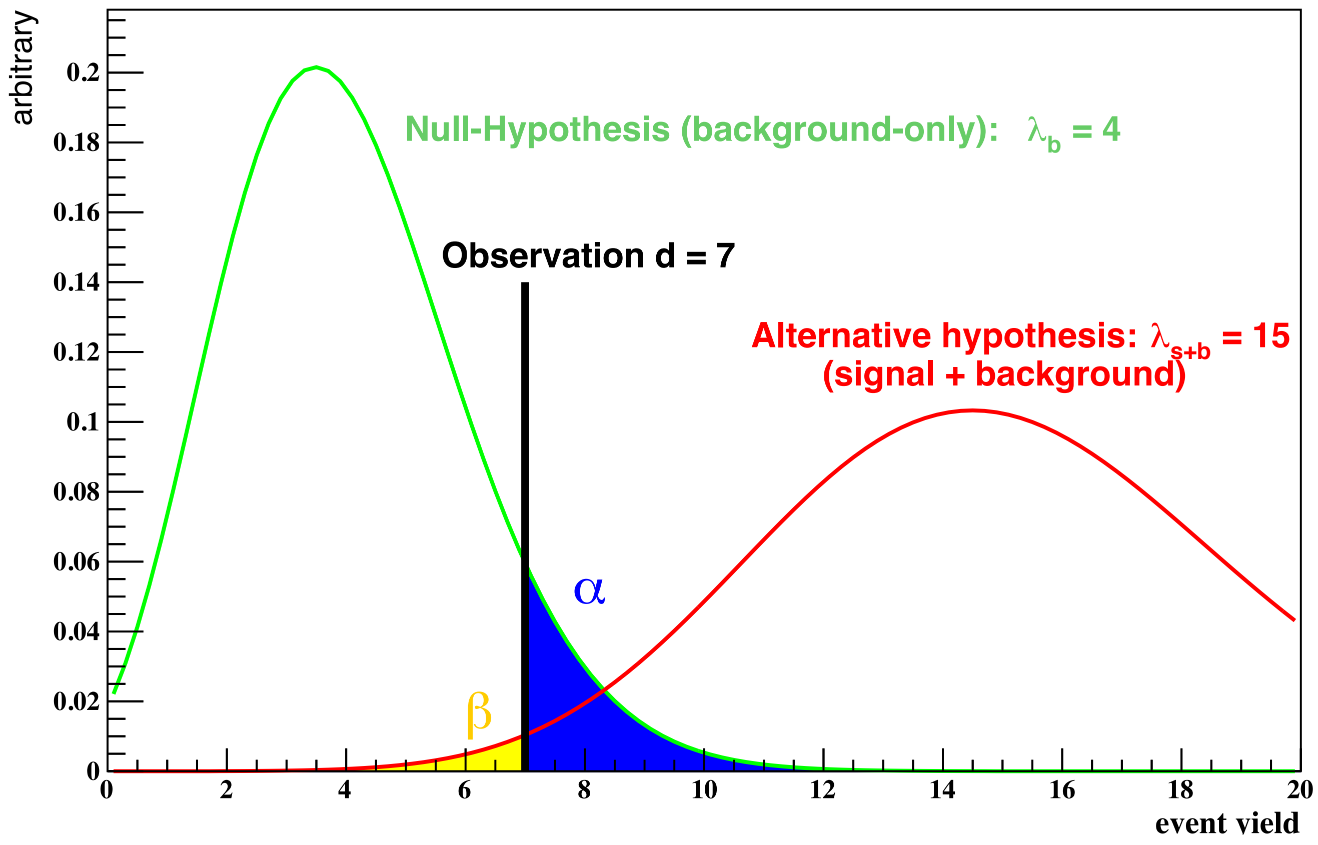
Example: measurement d = 7

- H_0: expected background b=4
- H_1: expected signal s=11 $\longrightarrow$ s+b=15

1-CL_b $$\alpha = 1 - \int_0^d P(x|b)dx$$

CL_{b+s} $$\beta = \int_0^d P(x|b+s)dx$$

- What is the upper limit on s at 95% confidence level for CL_{s+b} and CL_s ? (answer: s_{upper} = 8.5 and 8.7)



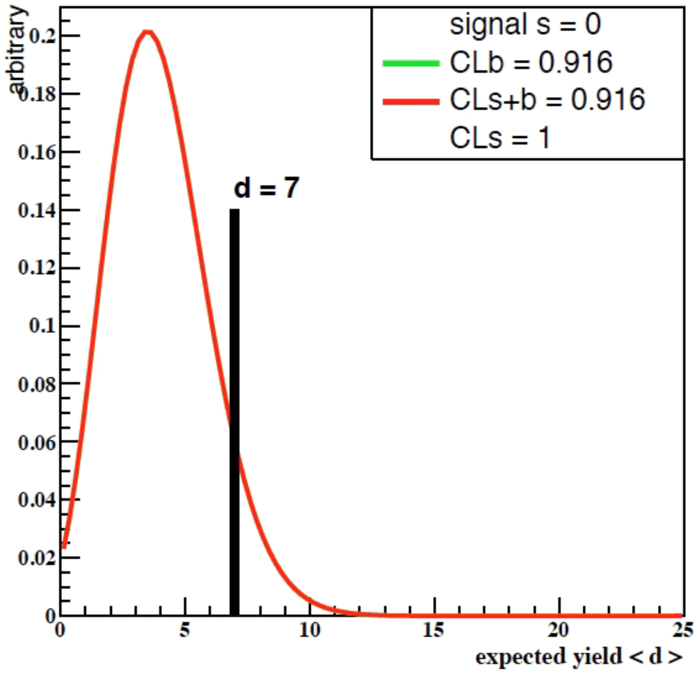
Christian Autermann, RWTH Aachen

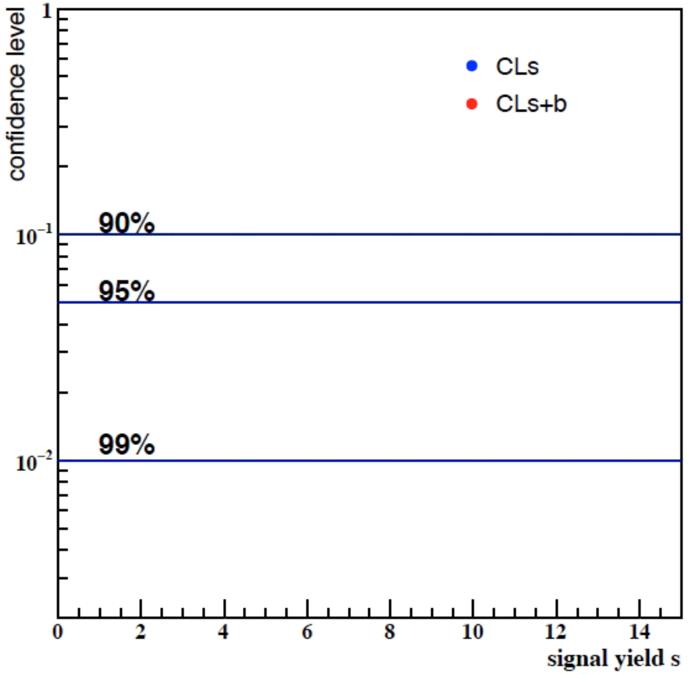
CLs

Example: measurement d = 7

$$\text{CL}_{b+s} = \int_0^d P(x|b+s)dx$$

- Scan for different signal hypotheses and compare with measurement
- For Poisson-distributed b = 4 and d = 7:





- Upper limit on s for CL_{S+B} = 95%: $s_{CLs+b,95\%} = 8.5$
- Upper limit on s for CL_S = 95%: $s_{CLs,95\%} = 8.7$

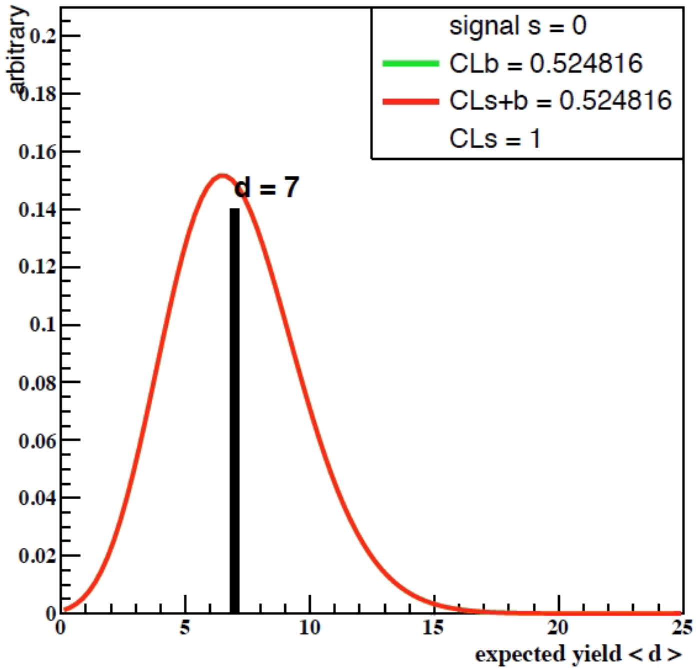
Low background: similar limits on s for CL_{S+B} and CL_S

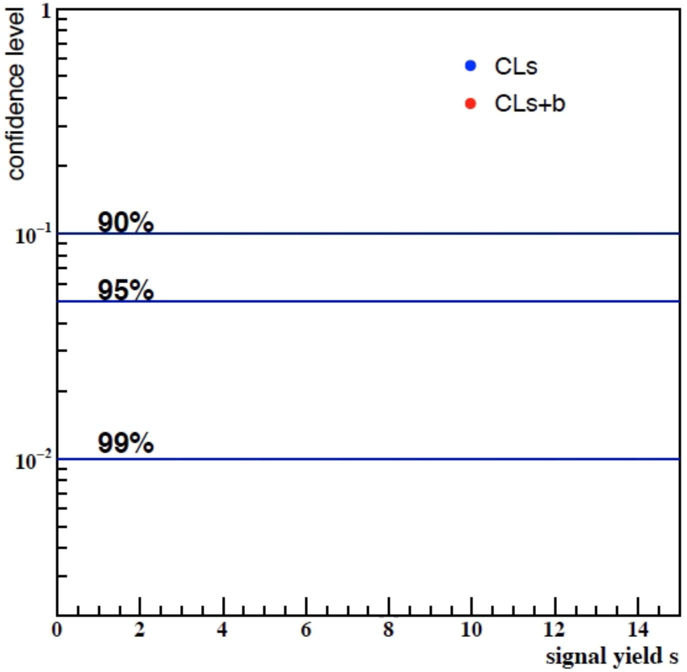
CL_s

$$CL_{b+s} = \int_0^d P(x|b+s)dx$$

Example: measurement d = 7

- Scan for different signal hypotheses and compare with measurement
- For Poisson-distributed b = 7 and d = 7:





- Upper limit on s for CL_{S+B} = 95%: $s_{CLs+b,95\%} = 5.5$
- Upper limit on s for CL_S = 95%: $s_{CLs,95\%} = 6.6$

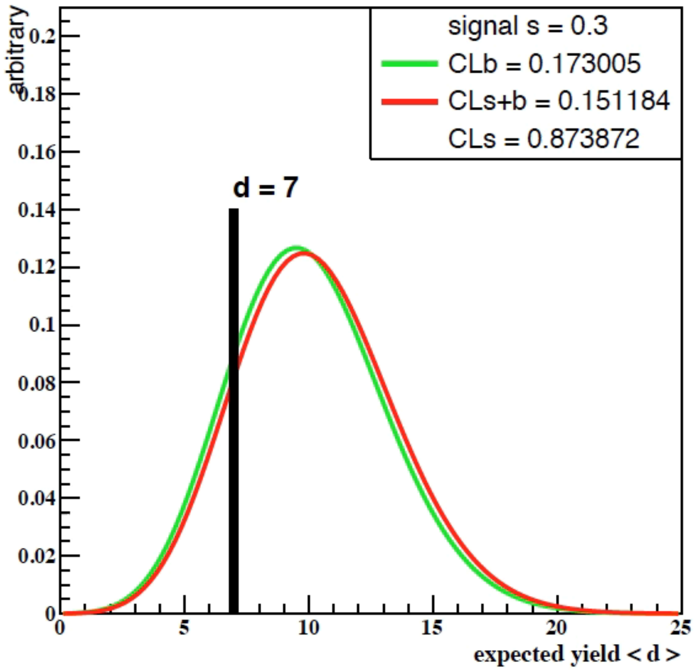
Medium background: CL_S gives worse limit on s than CL_{S+B}

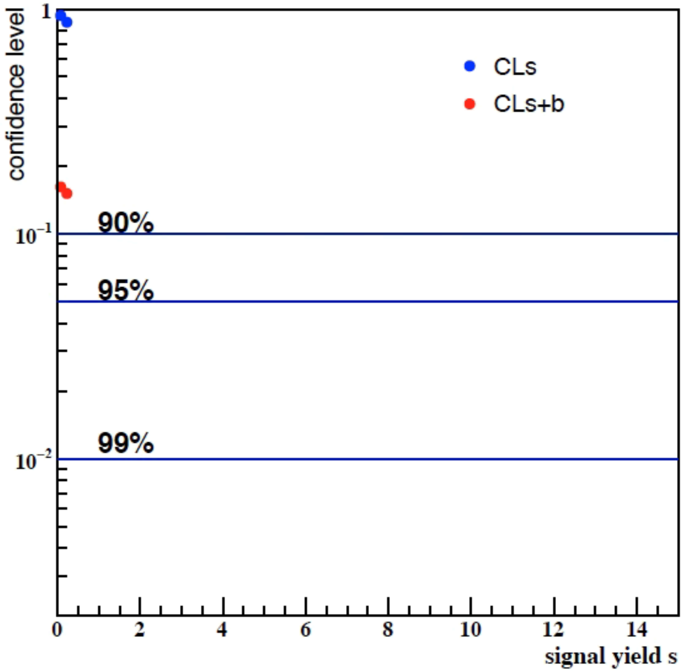
CLs

Example: measurement d = 7

$$\text{CL}_{b+s} = \int_0^d P(x|b+s)dx$$

- Scan for different signal hypotheses and compare with measurement
- For Poisson-distributed b = 10 and d = 7:





- Upper limit on s for CL_{S+B} = 95%: $s_{\text{CLs+b},95\%} = 2.5$
- Upper limit on s for CL_S = 95%: $s_{\text{CLs},95\%} = 5.5$

High background: CL_S gives much worse limit on s than CL_{S+B}

Confidence intervals

Summary

- Interval in which true value lies with pre-defined confidence level.
- Frequentist (or classical) approach:
  - Neyman Construction: correct coverage by construction
  - Unified Frequentist approach: use likelihood ratio as ordering principle to avoid empty intervals.

- Bayesian prior:
  - E.g. to avoid unphysical results.
  - Shape of prior has direct impact on result: possible under-coverage

- Modified frequentist approach CL_S:
  - Robust method to suppress possible effects from downward fluctuations of the background.
  - Price to pay: over-coverage

Profile-Likelihood Ratio

Signal strength μ

- Likelihood in a counting experiment

$$\mathcal{L}(\text{data}|\mu) = \text{Poisson}(\text{data}|\mu \cdot s + b)$$

- Product of Poisson likelihoods to measure n_i events in bin i

$$\text{Poisson}(\text{data}|\mu \cdot s + b) = \prod_i \frac{(\mu \cdot s_i + b_i)^{n_i}}{n_i!} e^{-(\mu \cdot s_i + b_i)}$$

- Signal strength μ: modifies expected signal using data
  - $\mu=0$: H_0 background only
  - $\mu=1$: H_1 expected signal

Signal strength μ

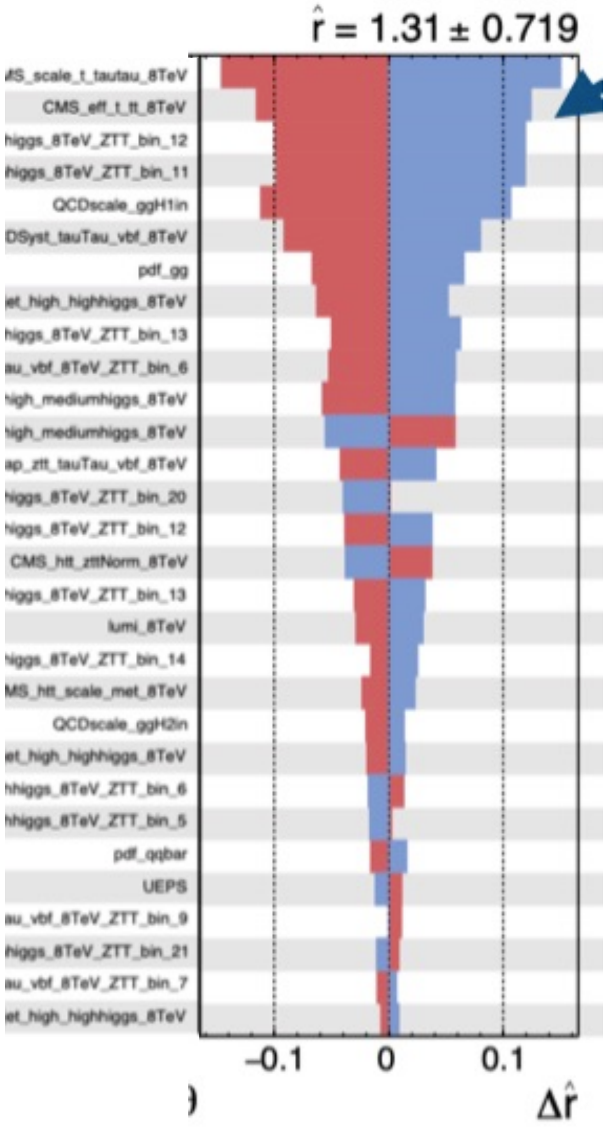
Nuisance parameters θ
impact measurement of s and b

- Likelihood in a counting experiment

$$\mathcal{L}(\text{data}|\mu) = \text{Poisson}(\text{data}|\mu \cdot s(\theta) + b(\theta))$$

- Nuisance parameters θ: parameters that are not of primary interest, but needed for the determination of signal and background, i.e. systematic uncertainties

- Modern particle physics data analyses often use hundreds of nuisance parameters



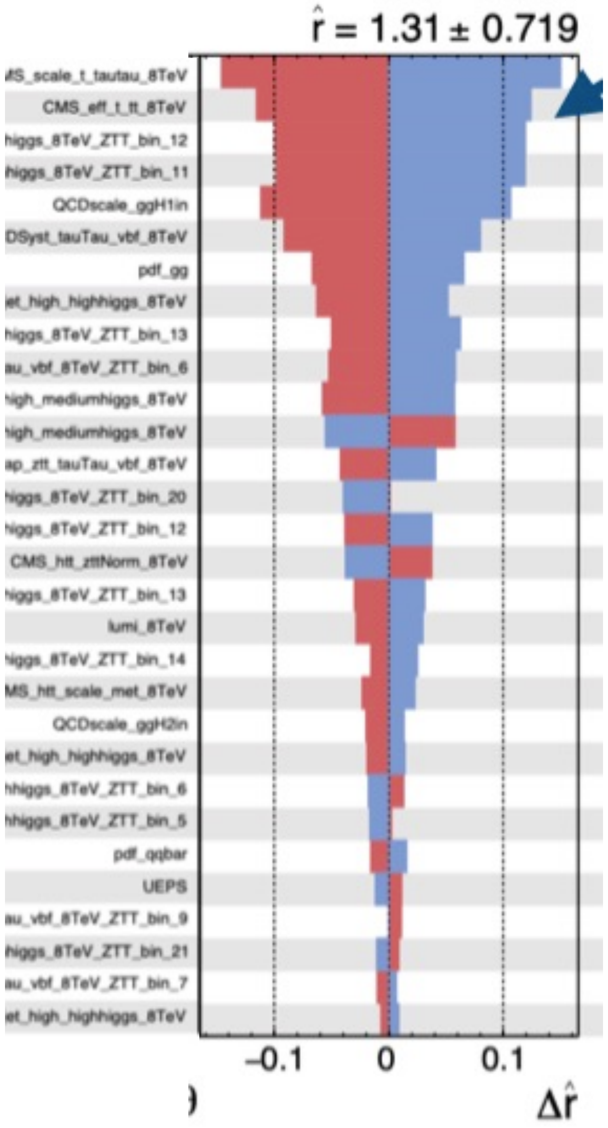
Signal strength μ

- Likelihood in a counting experiment

“Priors”, i.e. PDF determined in other measurements

$$\mathcal{L}(\text{data}|\mu) = \text{Poisson}(\text{data}|\mu \cdot s(\theta) + b(\theta)) \cdot PDF(\theta)$$

- PDF(θ): prior knowledge from ancillary measurements used as constraints for the Frequentist likelihood of the main measurement



Profile-likelihood ratio

- Profile likelihood: determine the interval for the (true) signal strength μ, for optimal nuisance parameters $\hat{\theta}_\mu$, normalized to the global maximum of the likelihood.
- "Profile" = scan, determine q_μ for all μ

Test statistic

Signal strength

Nuisance parameter $\hat{\theta}_\mu$: maximises $\mathcal{L}(\text{data}|\mu,\hat{\theta}_\mu)$ for μ

$$q_\mu = -2\ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\Theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\Theta})}$$

Best-fit values of all parameters

CCGV section 2.5 and
CMS+ATLAS 2.1

Profile-likelihood ratio

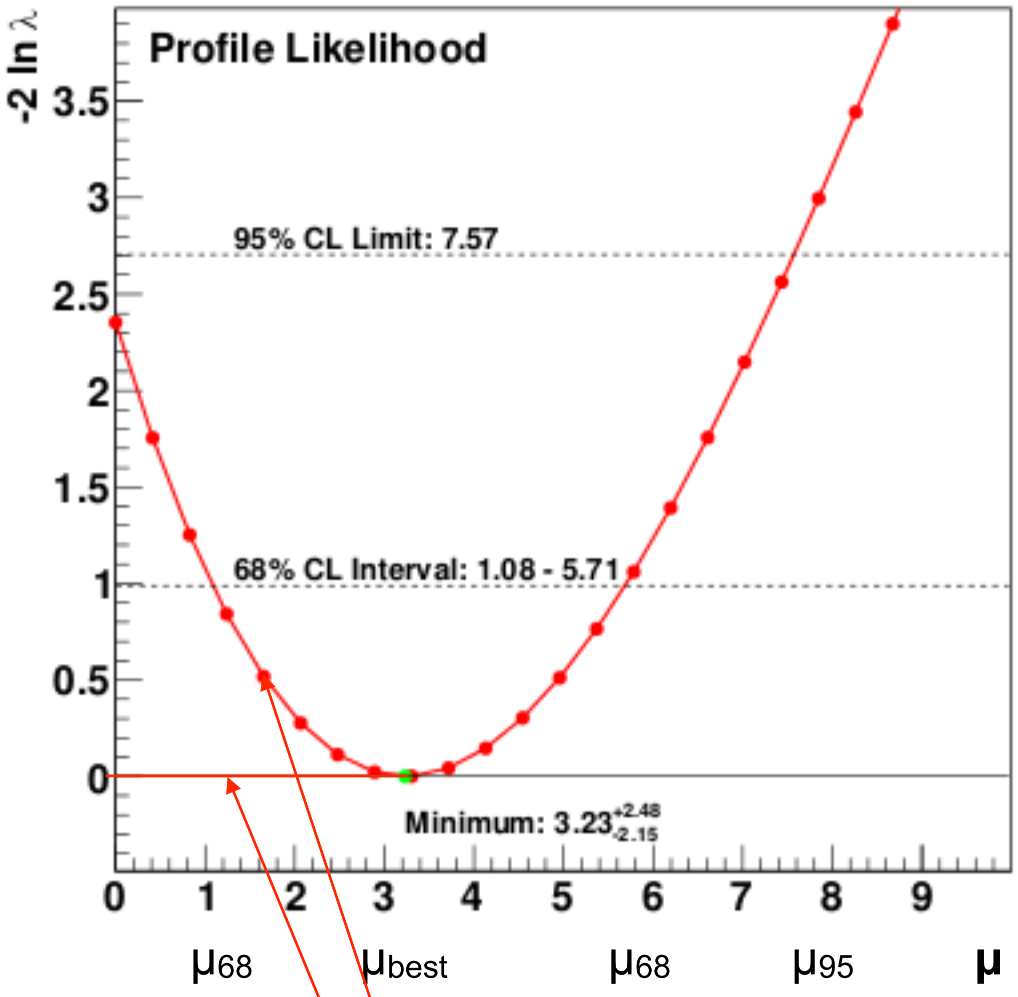
$$q_\mu = -2\Delta \ln \mathcal{L} \approx \frac{(\mu - \hat{\mu})^2}{\sigma^2}$$

- In the limit of high statistics (Wilks), q_μ follows χ^2-distribution (parabola)

- Profile-likelihood distribution has all estimators:
  - Best fit of μ at minimum
  - 2-sided confidence interval: e.g. 68%
  - Exclusion of null-hypothesis:
    - $q(\mu=0) = z^2 = (\text{significance})^2$
    - here: $z \sim \sqrt{2.4} \approx 1.5$
  - Upper limit μ_{95}:

$$-2\Delta \ln \mathcal{L}(\mu_{95}) = 1.645^2 = 2.71$$

Large-Sample Approximation



In general: 2-sided interval

For 1-sided limit set q_μ=0 für $\mu<\hat{\mu}$

CCGV, section 3.7

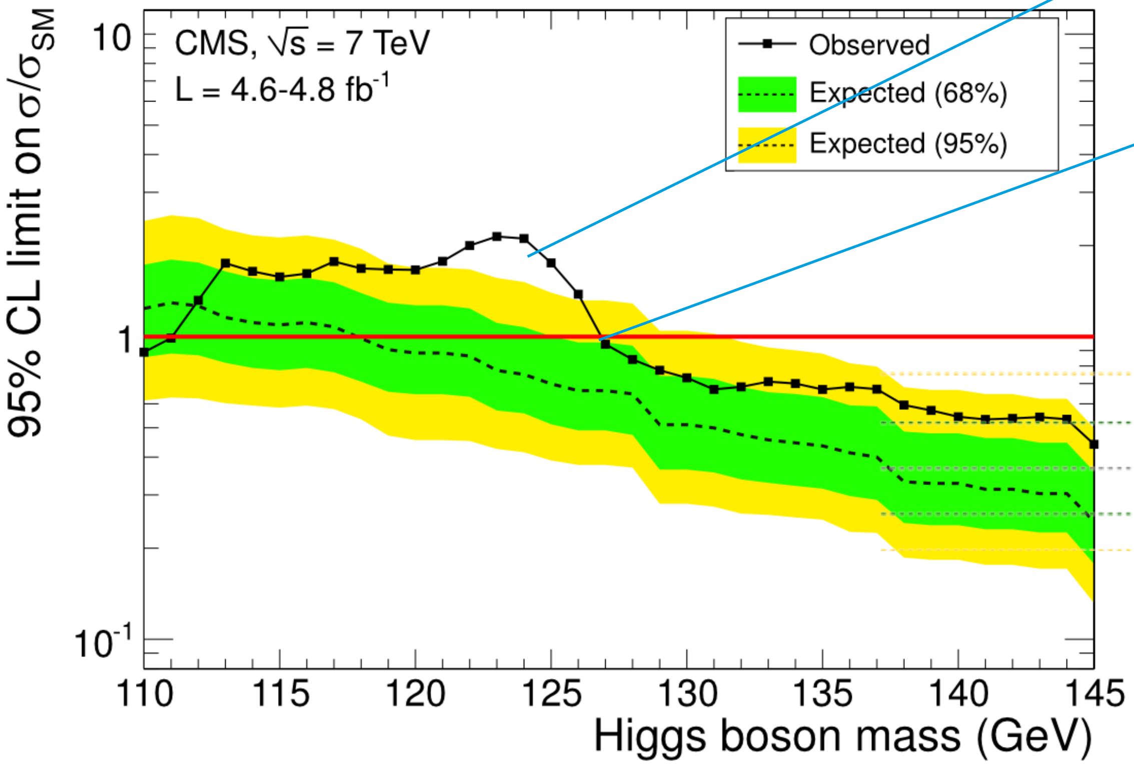
Higgs discovery

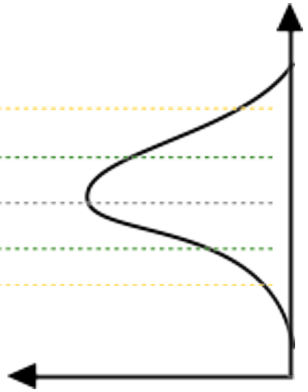
History: status December 2011

- Number of events depends on invariant mass

Excess at a mass of ~124 GeV

Higgs boson mass > 127 GeV
Excluded at 95% confidence level





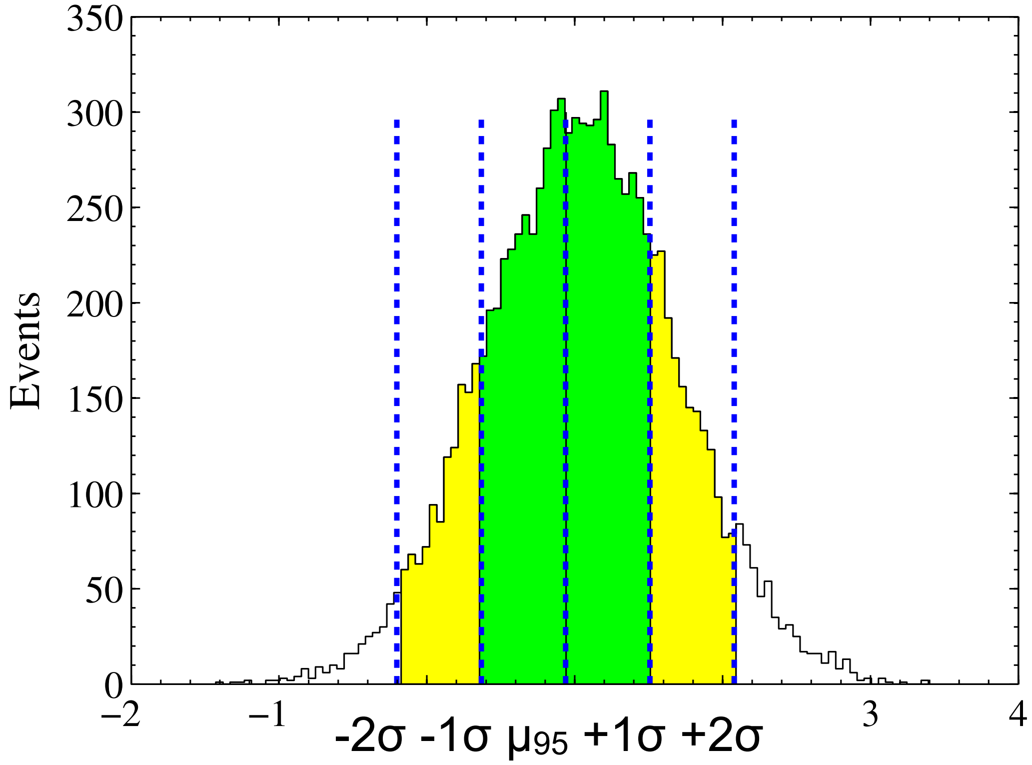
Distribution of pseudo experiments using expected (background-only) limit

- Blind analysis: software and all criteria were all fixed before looking at the data

Higgs discovery

Brazilian-flag figure

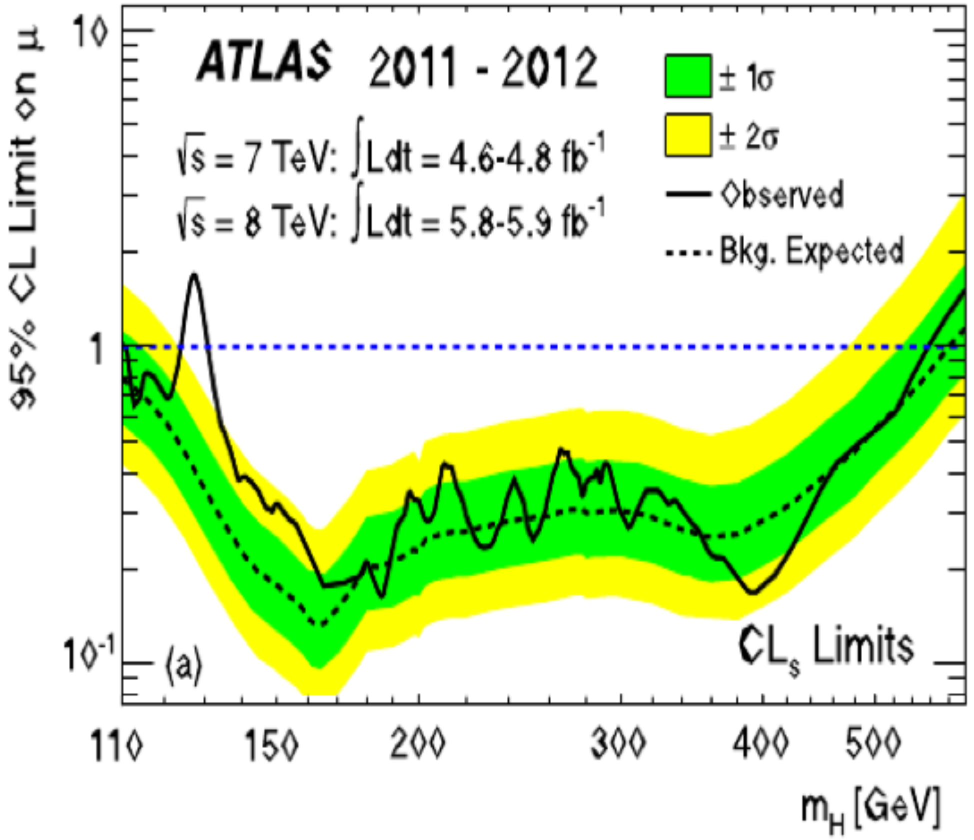
- Determine μ_{95}, i.e. signal strength excluded at 95% CL_s
- Pseudo-experiments to determine the distribution around the 95% limit for the background-only hypothesis, i.e. median and intervals for ±1σ und ±2σ around μ_{95}.

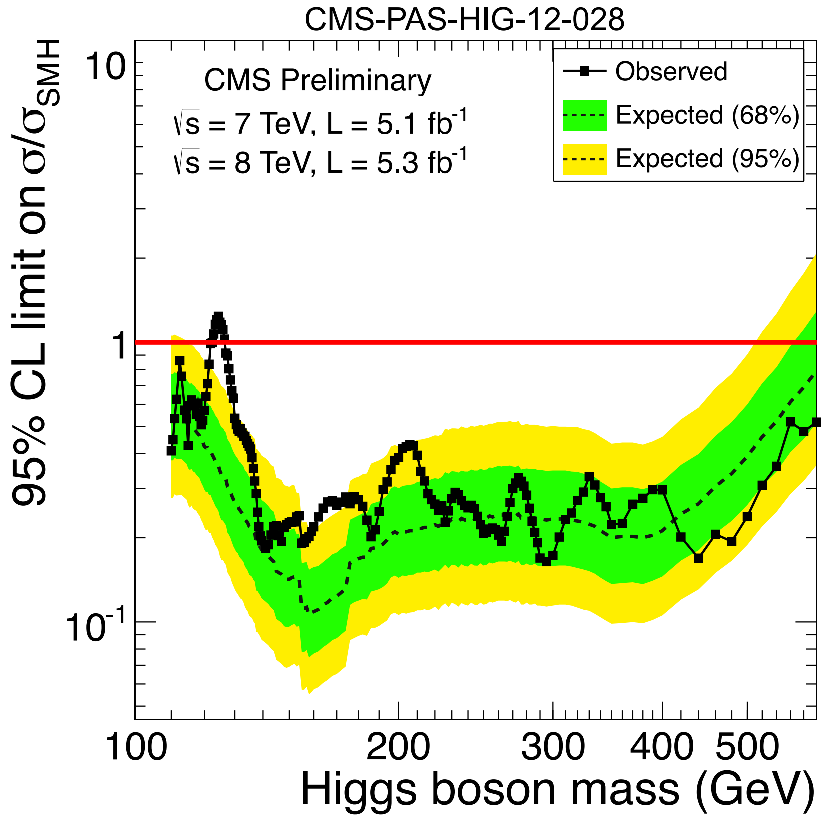


Higgs discovery

4 July 2012

- Public announcement of the discovery at CERN



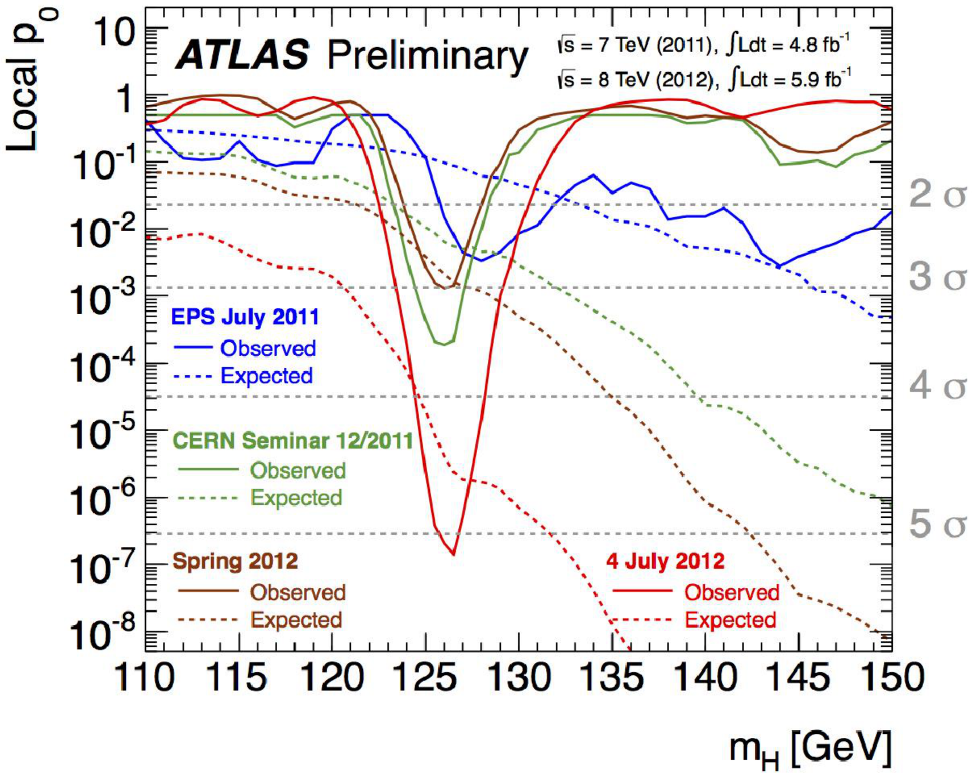


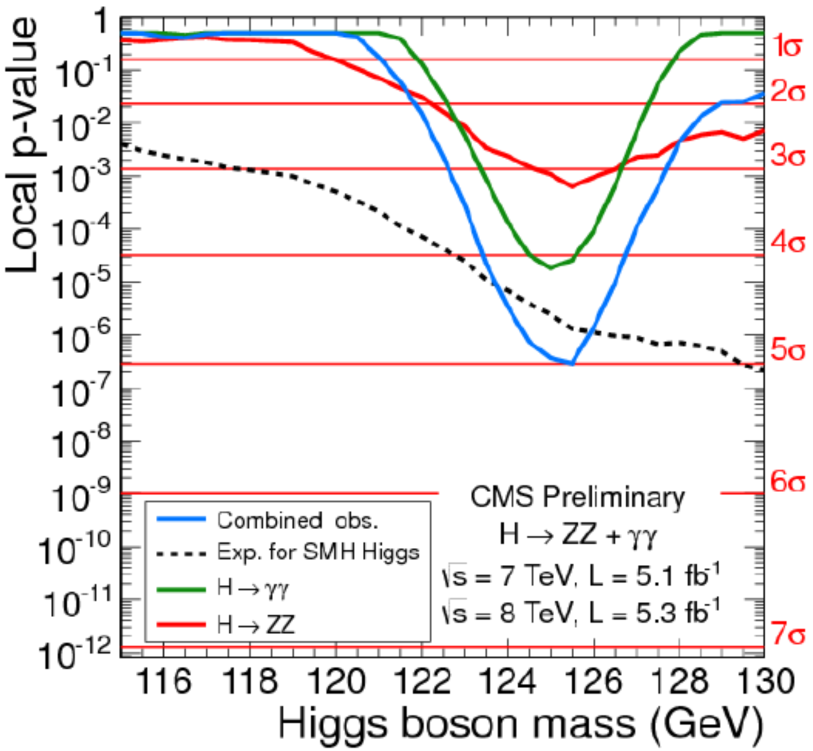
- Exclusion of signals between 131(128) GeV and 523(600) GeV

Higgs discovery

4 July 2012

- Public announcement of the discovery at CERN





- Determine signal significance and local *p*-value by comparison with background hypothesis

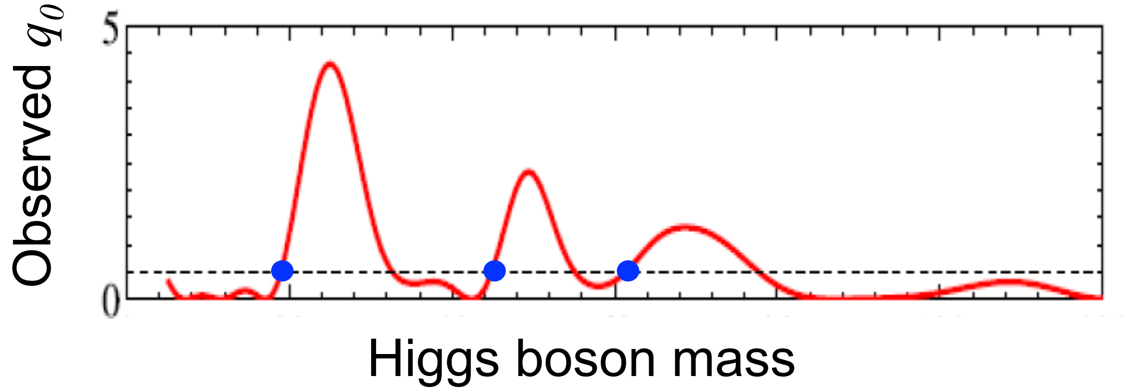
$S_{ATLAS} = 5.9\ \sigma$ $\qquad$ $S_{CMS} = 5.0\ \sigma$

Look-elsewhere effect

- Local p-value: probability that the excess is due to a statistical background fluctuation at a specific value of the Higgs candidate mass (or another observable)
- In global searches (e.g. over the whole mass range) the probability for a fluctuation somewhere increases with the size of the search range → “look-elsewhere effect”

$$\text{global } p = \text{trial factor} \times \text{local } p$$

- The trial factor is generally proportional to the range and inverse proportional to the (mass) resolution
- Determination:
  - Usually by pseudo-experiments: requires a lot of CPU, because fluctuations are rare.
  - Or estimate from frequency of fluctuations in data



CMS+ATLAS, Figure 4

Summary

- Maximum likelihood estimator (MLE)
  - Least-squares method is an important special case of MLE, for the (usually good) assumption of Gaussian behaviour

- Hypothesis testing
  - Neyman-Pearson lemma: likelihood ratio is the best test statistic

- Confidence intervals:
  - Frequentist Neyman construction: coverage by design
  - Wilks' Theorem: asymptotic approach
  - Feldman-Cousins unified approach
  - Bayesian priors
  - Modified Frequentist approach: CL_S method

- Profile-likelihood ratio
  - Posterior likelihoods from scan of signal strength, including systematic uncertainties
  - Higgs discovery figures: "Brazilian-flag" and *p*-value

Backup

1-Sided Limits and 2-Sided Intervals: Unified Approach

Example: Poisson Distribution with μ=4 (95% CL)

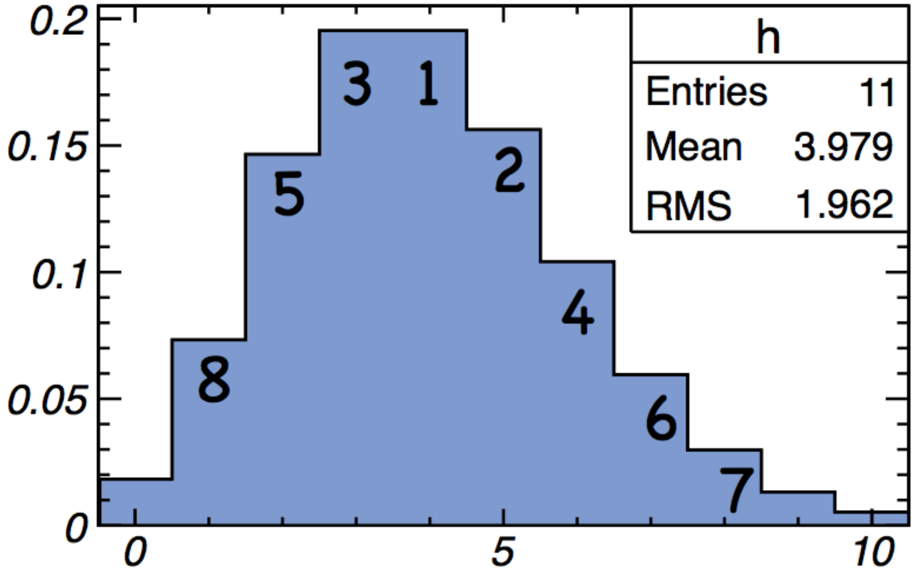
- Construct interval using an ordering principle, based on the likelihood ratio R(n|μ):

$$R(n|\mu) = \frac{g(n|\mu)}{g(n|\mu_{\text{best}})} \quad \text{where} \quad g(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

  and μ_{best} = μ for which g(n|μ) is biggest

  - Calculate R(n|μ=4) for each measurable value of n
  - R defines order of bins
  - Sum up bins until in decreasing order of R until coverage is reached

Poisson distribution for μ=4



- Recipe:
  - Sum up values of â for decreasing values of R until g(â|a) reaches the chosen confidence level
  - For â < 0: add contributions to the left side (no empty interval)

1-Sided Limits and 2-Sided Intervals: Unified Approach

Example: Poisson Distribution with μ=4 (95% CL)

- Construct interval using an ordering principle, based on the likelihood ratio R(n|μ):

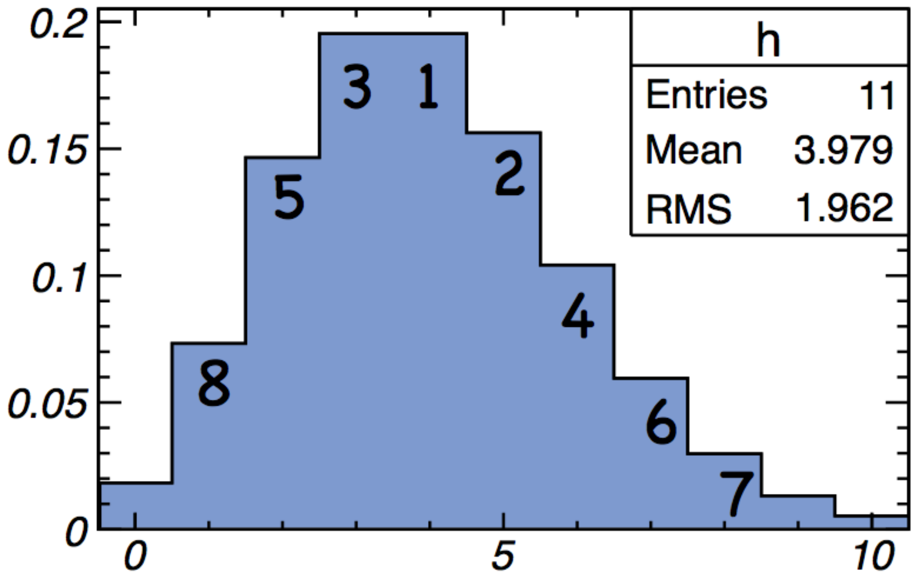
$$R(n|\mu) = \frac{g(n|\mu)}{g(n|\mu_{\text{best}})} \qquad \text{where} \qquad g(n|\mu) = \frac{e^{-\mu}\mu^n}{n!}$$

  and μ_{best} = μ for which g(n|μ) is biggest

  - Calculate R(n|μ=4)
  - R defines <u>order</u> of
  - Sum up bins until i coverage is reache

n	$R(n\|\mu)$	$g(n\|\mu)$	$\sum g$
4	1.000	0.195	0.195
5	0.891	0.156	0.352
3	0.872	0.195	0.547
6	0.649	0.104	0.651
2	0.541	0.147	0.798
7	0.400	0.060	0.857
8	0.213	0.030	0.887
1	0.199	0.073	0.960

Poisson distribution for μ=4



- Result:
  - Confidence interval [1,8] provides coverage of 96%
  - More complex distributions → more computing

Counting Experiment with Known Background

- Observation of *n* events with small signal *s*

$$P_0(n;b) = \frac{1}{n!} b^n e^{-b} \qquad P_1(n;s+b) = \frac{1}{n!}(s+b)^n e^{-(s+b)}$$

$$q = -2\ln\lambda = 2(n\ln(1+\frac{s}{b}) - s)$$

- Background *b* then *n* = *b* + *s*:

$$q = 2(b+s)\ln\left(1+\frac{s}{b}\right) - 2s$$

- For *s* « *b*:

$$\sqrt{q} = s/\sqrt{b} + \mathcal{O}((s/b)^2)$$

- In Wilks' approximation: for a single degree of freedom, the significance of the signal s, expressed by the Gaussian quantile *z* is:

$$z = \sqrt{\Delta\chi^2} = \sqrt{q} = s/\sqrt{b}$$

Cowan, Cranmer, Gross and Vitelis "Asymptotic formulae for likelihood-based tests of new physics" https://arxiv.org/abs/1007.1727