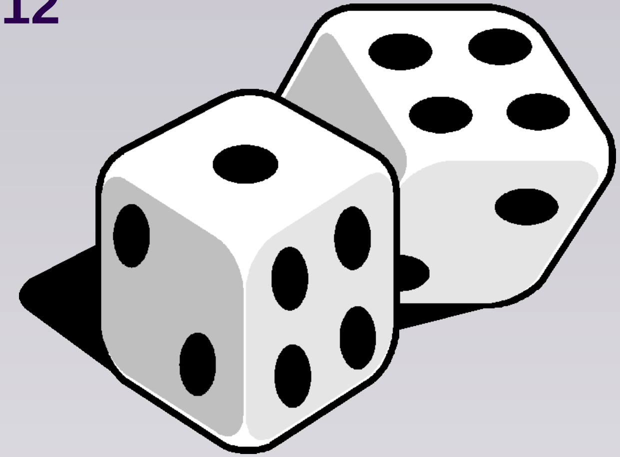
Introduction to Statistics

CERN Summer Student Lecture Program 2012



Helge Voss

… and Machine Learning

(in the last lecture)

Outline

- Why Statistics
- What is Probability :
  - axioms
  - **frequentist / Bayesian interpretation**
- **Lecture 2**
  - **Hypothesis testing**
    - error types and Neyman-Pearson Lemma
    - confidence level α and p-value
    - new particle searches
- Lecture (3-4)
  - Maximum Likelihood fit
  - Neyman Confidence belts
  - Monte Carlo Methods (Random numbers/Integration)
  - Machine Learning / Pattern Recognition

Frequentist vs. Bayesian

Bayes' Theorem

$$P(\mu|n) = P(n|\mu)\frac{P(\mu)}{P(n)}$$

- $P(n|\mu)$: **Likelihood function**
- $P(\mu|n)$:**posterior probability of μ**
- $P(\mu)$: **the "prior"**
- $P(n)$: **just some normalisation**

B.t.w.: Nobody doubts Bayes' Theorem:
discussion starts ONLY if it is used to turn

frequentist statements:

- probability of the observed data given a certain model: $P(Data|Model)$

into Bayesian probability statements:

- probability of a the model begin correct (given data): $P(Model\,|Data)$

- **… there can be heated debates about 'pro' and 'cons' of either….**

P (Data|Theory) ≠ P (Theory|Data)

- Higgs search at LEP: the statement
  - the probability that the data is in agreement with the Standard Model background is less than 1% (i.e. P(data| SMbkg) < 1%)

  went out to the press and got turned round to:

  ~~P(data|SMbkg) = P(SMbkg|data)~~ < 1% → P(Higgs|data) > 99% !

WRONG!

- easy Example: Theory = female (hypothesis) .. male (alternative)

  Data = pregnant or not pregnant

P (pregnant | female) ~ 2-3% but P (female | pregnant) = ?? ☺

→o.k… but what DOES it say?

The correct frequentist interpretation

we know: P (Data|Theory) ≠ P (Theory|Data)

rather: Bayes Theorem: $P\,(\text{Theory}|\text{Data}) = P\,(\text{Data}|\text{Theory})\,\frac{P(\text{Theory})}{P(\text{Data})}$

Frequentists answer ONLY: P (Data|Theory)

… although.. let's be honest, we are all interested in P(Theory…)

We only learn about the "probability" to observe certain data under a given theory. Without knowledge of how likely the theory (or a possible "alternative" theory) is .. that doesn't say anything about how unlikely this makes our current theory !

Later: we'll define "confidence levels" … i.e. if P(data) < 5%, discard theory.

- **→ can accept/discard theory and state how often/likely we will be wrong in doing so. But again: It does not say how "likely" the theory itself (or the alternative) is true**
- **→ note the subtle difference !!**

Frequentist vs. Bayesian

- Certainly: both have their "right-to-exist"
  - Some "probably" reasonable and interesting questions cannot even be ASKED in a frequentist framework :
    - "How much do I trust the simulation"
    - "How likely is it that it will raining tomorrow?"
    - "How likely is it that climate change is going to…
  - after all.. the "Bayesian" answer sounds much more like what you really want to know: i.e.
    "How likely is the "parameter value" to be correct/true ?"
- <u>BUT:</u>
  - NO Bayesian interpretation w/o "prior probability" of the parameter
    - where do we get that from?
    - all the actual measurement can provide is "frequentist"!

Bayesian Prior Probabilties

- "flat" prior $\pi(\theta)$ to state "no previous" knowledge (assumptions) about the theory?
  - **often done, BUT WRONG:**
    - e.g. flat prior in M_{Higgs} → not flat in M_{Higgs}^2
  - **Choose a prior that is invariant under parameter transformations**

  → Jeffrey's Prior → "objective Bayesian":
    - "flat" prior in Fisher's information space → independent of parameterisation!
    - $\pi(\theta) \propto \sqrt{\mathrm{I}(\theta)}$ $\quad (\pi(\vec{\theta}) \propto \sqrt{\det \mathrm{I}(\vec{\theta})}$ if several parameters)

$I(\theta) = -E_x[\frac{\partial^2}{\partial\theta^2} log(f(x;\theta))]$:

- $f(x;\theta)$: Likelihood function of θ, probability to observe x for a give parameter θ
- amount of "information" that data x is 'expected' to contain about the parameter θ

- **personal remark: nice idea, but "WHY" would you want to dot that?**
  - **still use a "arbitrary" prior, only make sure everyone does the same way**
  - **loose all "advantages" of using a "reasonable" prior if you choose already to use a Bayesian interpretation!**

Frequentist or Bayesian?

"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"

Louis Lyons, Academic Lecture at Fermilab, August 17, 2004

- **Traditionally: most scientists are/were "frequentists"**
  - **no NEED to make "decisions" of your degree of believe if the data is not 99.9999% "conclusive"…**
  - **it's ENOUGH to present data, and how likely they are under certain scenarios**
    - **keep doing so and combine measurements**
- **Bayesians are growing**
  - **well, at least now we have the means to do lots of prior comparisons: Computing power/ Markov Chain Monte Carlos**

Hypothesis Testing

- a **hypothesis** H **specifies some process/condition/model which might lie at the origin of the data** x
  - **e.g.** H **a particular event type**
    - signal or background (on event by event basis)
    - NEW PHYSICS or Standard Model (your full data set)
  - **e.g.** H **a particular parameter in a diff. cross section**
    - some mass / coupling strength / $\cancel{CP}$ parameter
- **former: Simple (point) hypothesis**
  - **completely specified, no free parameter**
    - → PDF: $PDF(x) \equiv PDF(x; H)$
- **latter: Composite hypothesis**
  - H **contains unspecified parameters (mass, systematic uncertainties, …)**
    - → a whole band of $PDF(x; H(\theta))$
    - → for given x the $PDF(x; H(\theta))$ can be interpreted as a function of θ → Likelihood
    - → $L(x|H(\theta))$ the probability to observe x in this model H with parameter θ (sometimes also denoted: $L(\theta|x)$ or $L(\theta)$) **! Note: this is not a PDF !**

Why talking about “NULL Hypothesis”

- Statistical tests are most often formulated using a
  - “null”-hypothesis and its
  - “alternative”-hypothesis

- Why?

→ it is much easier to “exclude” something rather than to prove that something is true.

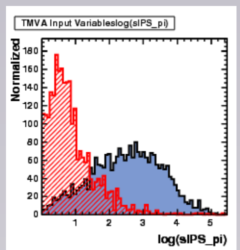
→ excluding: I need only ONE detail that clearly contradicts

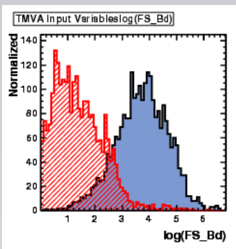
→ assume you search for the “unknown” new physics.

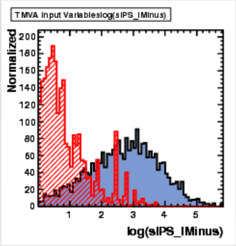
“null”-hypothesis :	Standard Model (background) only
“alternative”:	everything else

Hypothesis Testing

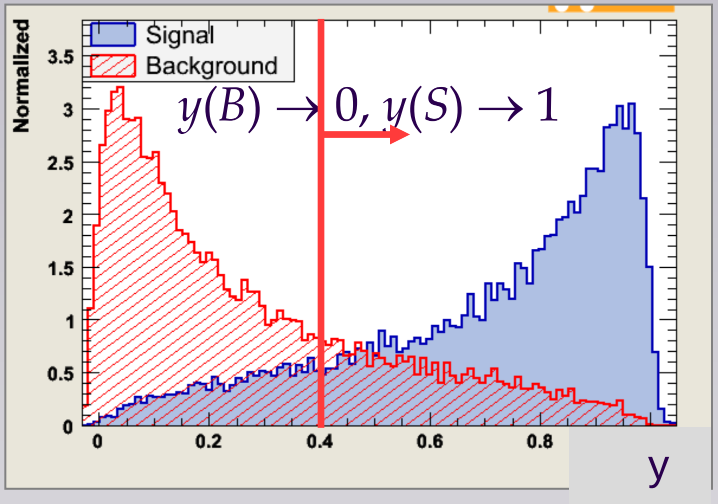
<u>Example:</u> event classification   Signal(H_1) or Background(H_0)







- **Test statistic**
  $y(x_1, x_2, \ldots, x_n): R^n \rightarrow R$
- **PDF$(y|Signal)$ and PDF$(y|Bkg)$**
- **choose cut value: i.e. a region where you "reject" the null- (background-) hypothesis**
  ("size" of the region based on signal purity or efficiency needs)



$$y(x): \begin{cases} > \text{cut: signal} \\ = \text{cut: decision boundary} \\ < \text{cut: background} \end{cases}$$

- **You are bound to making the wrong decision, too…**

Hypothesis Testing

Type-1 error: (false positive)

→ accept as signal (reject backgr. hypothesis) although it IS background

Type-2 error: (false negative)

→ accept background hypothesis although it is signal

Trying to select signal events:
(i.e. try to disprove the null-hypothesis stating it were "only" a background event)

truly is: \ accept as:	Signal	Back-ground
Signal	☺	Type-2 error
Back-ground	Type-1 error	☺

Hypothesis Testing

Type-1 error: (false positive)

reject the null-hypothesis although it would have been the correct one

→ accept alternative hypothesis although it is false

Type-2 error: (false negative)

fail to reject the null-hypothesis/accept null hypothesis although it is false

→ reject alternative hypothesis although it would have been the correct/true one

Try to exclude the null-hypothesis (as being unlikely to be at the basis of the observation):

truly is: \ accept as:	H_1	H_0
H_1	☺	Type-2 error
H_0	Type-1 error	☺

"C": "critical" region: if data fall in there → REJECT the null-hypothesis

Significance α: Type-1 error rate:
(rate of "false discovery")

$$\alpha = \int_C P(x|H_0)\,dx$$

should be small

Size β: Type-2 error rate:
Power: 1– β (sensitivity to the "alternative" theory)

$$\beta = \int_{!C} P(x|H_1)\,dx$$

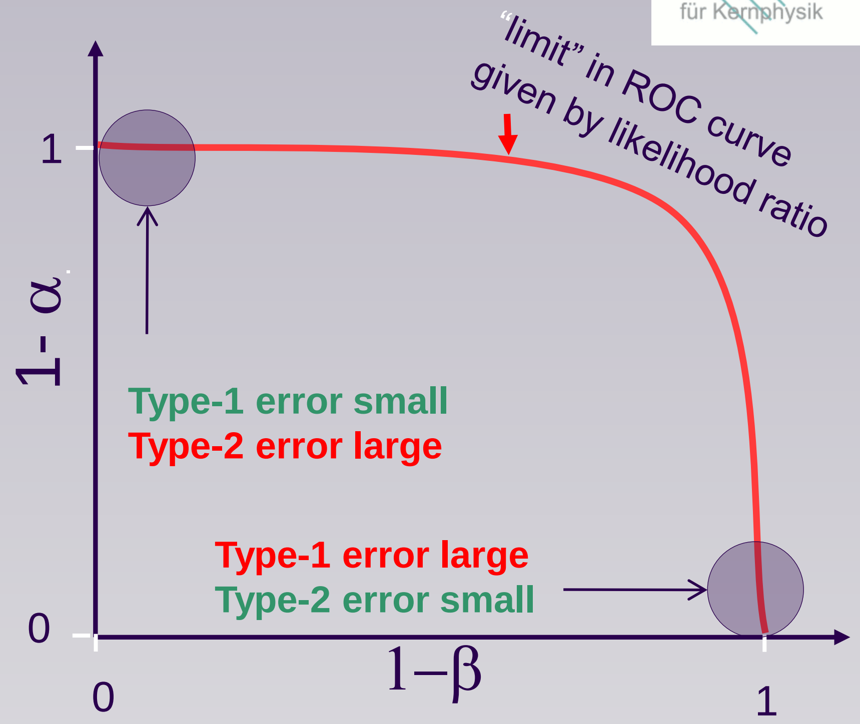
should be small

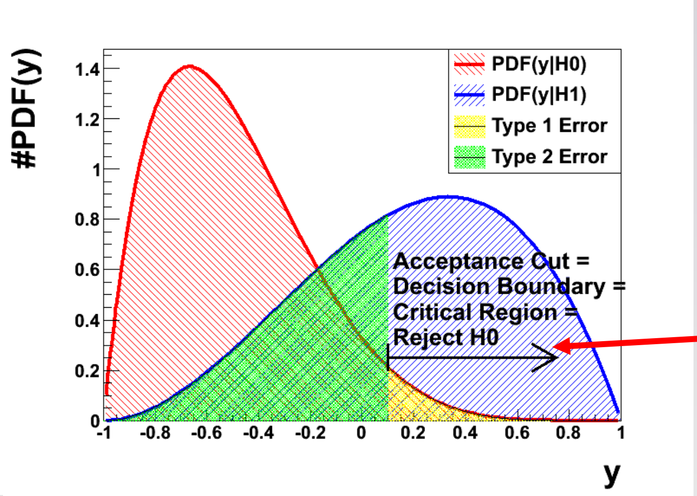
Neyman Pearson Lemma

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

- or any monotonic function thereof, e.g. log(L)

Neyman-Peason:
The Likelihood ratio used as "test statistics" t(x) gives for each significance α the test (critical region) with the largest power $1-\beta$.
i.e. it maximises the area under the "*Receiver Operation Characteristics*" (ROC) curve



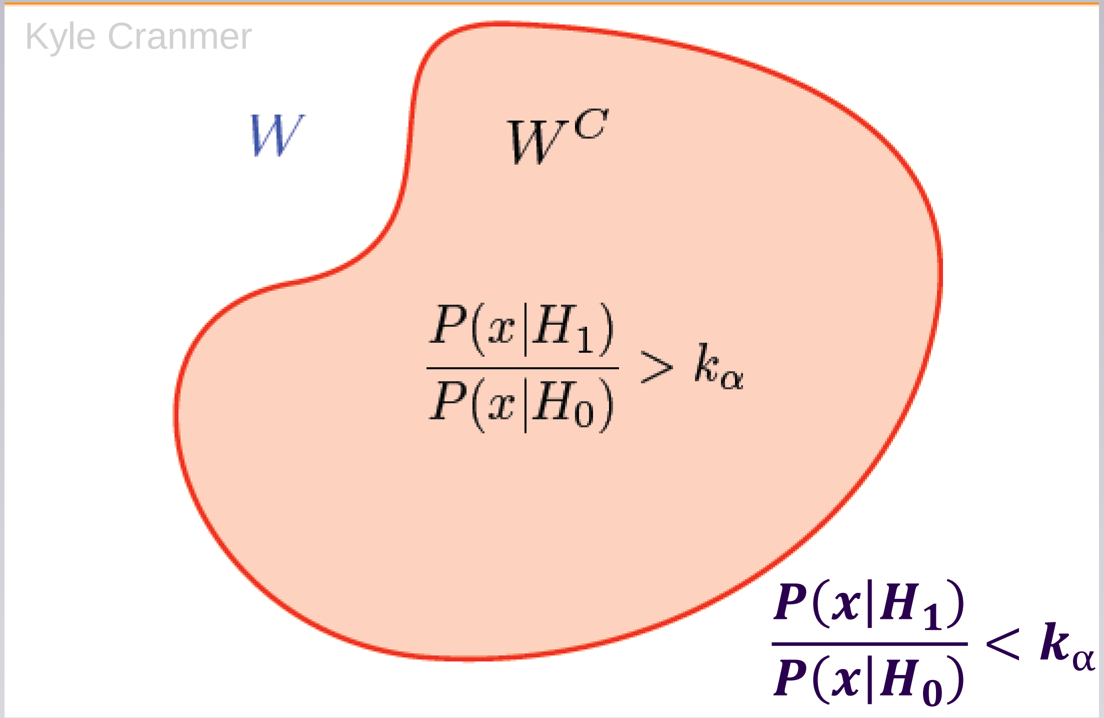


- **measure x --- want to discriminate model H_1 from H_0**
- **H_0 predicts x to be distributed acc. to $P(x|H_0)$**
- **H_1 predicts x to be distributed acc. to $P(x|H_1)$**

→ get distribution of $y(x)$ if H_0 were true: $PDF(y|H_0)$
→ same for H_1 : $PDF(y|H_1)$
→ get ROC curve / critical region

→ calculate test statistics $y(x_{data})$ after measurement and see if you have to reject H_0 or not

Neyman Pearson Lemma



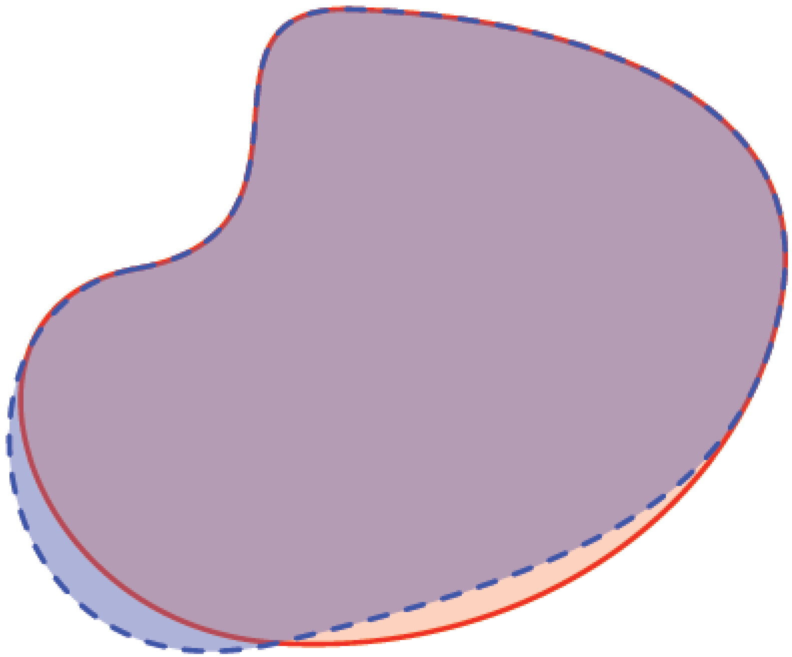
graphical proof of Neyman Pearson's Lemma:
(graphics/idea taken from Kyle Cranmer)

- **the critical region W^C given by the likelihood ratio $\frac{P(x|H_1)}{P(x|H_0)}$**

→ **for each given size α** **(risk of e.g. actually making a false discovery)**
= **the statistical test with the largest power $1-\beta$** **(chances of actually discovering something given it's there)**

Neyman Pearson Lemma

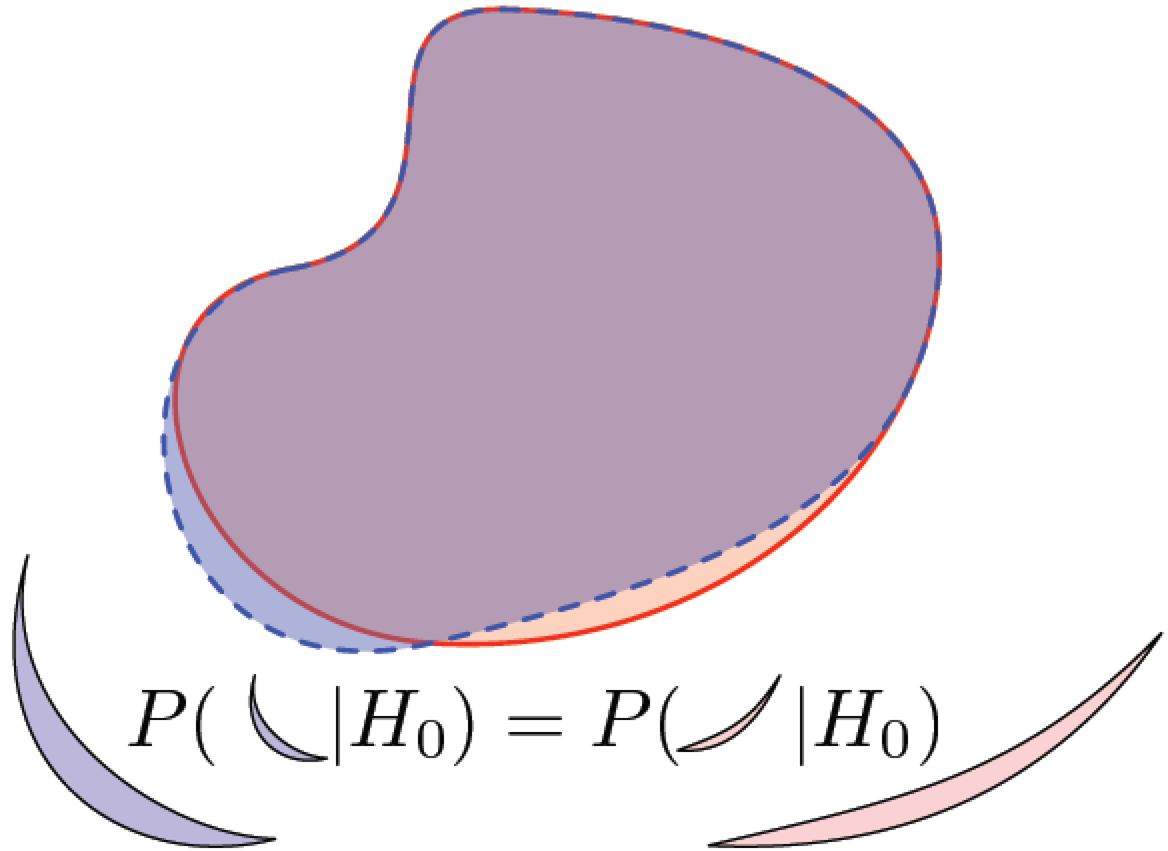
Kyle Cranmer



assume we want to modify/find another "critical" region with same size (α) i.e. same probability under H_0

Neyman Pearson Lemma

Kyle Cranmer



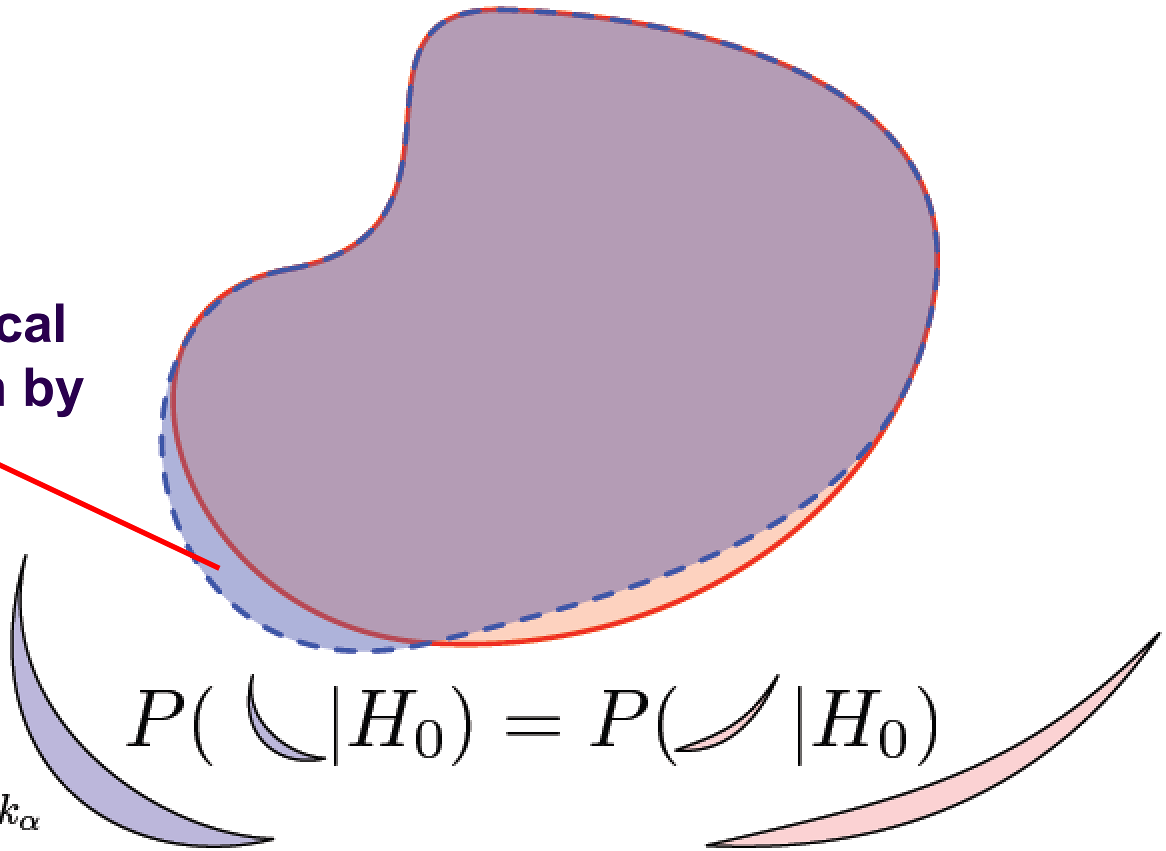
… as size (α) is fixed

$$\alpha = \int_C P(x|H_0)dx$$

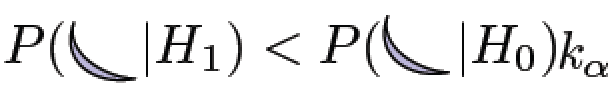
Neyman Pearson Lemma

Kyle Cranmer

outside "critical region" given by LL-ratio



$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha$$

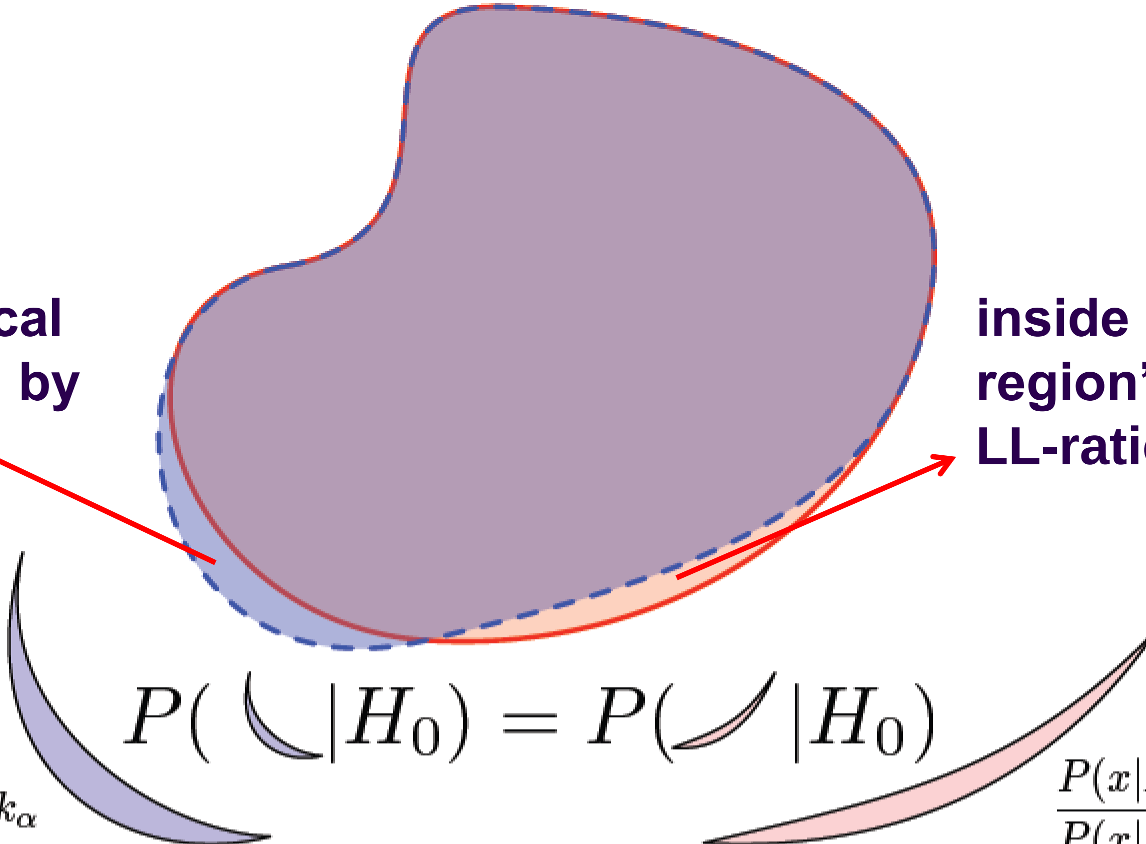


Neyman Pearson Lemma

Kyle Cranmer

outside "critical region" given by LL-ratio

inside "critical region" given by LL-ratio



$$P(\;|H_0) = P(\;|H_0)$$

$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha$$

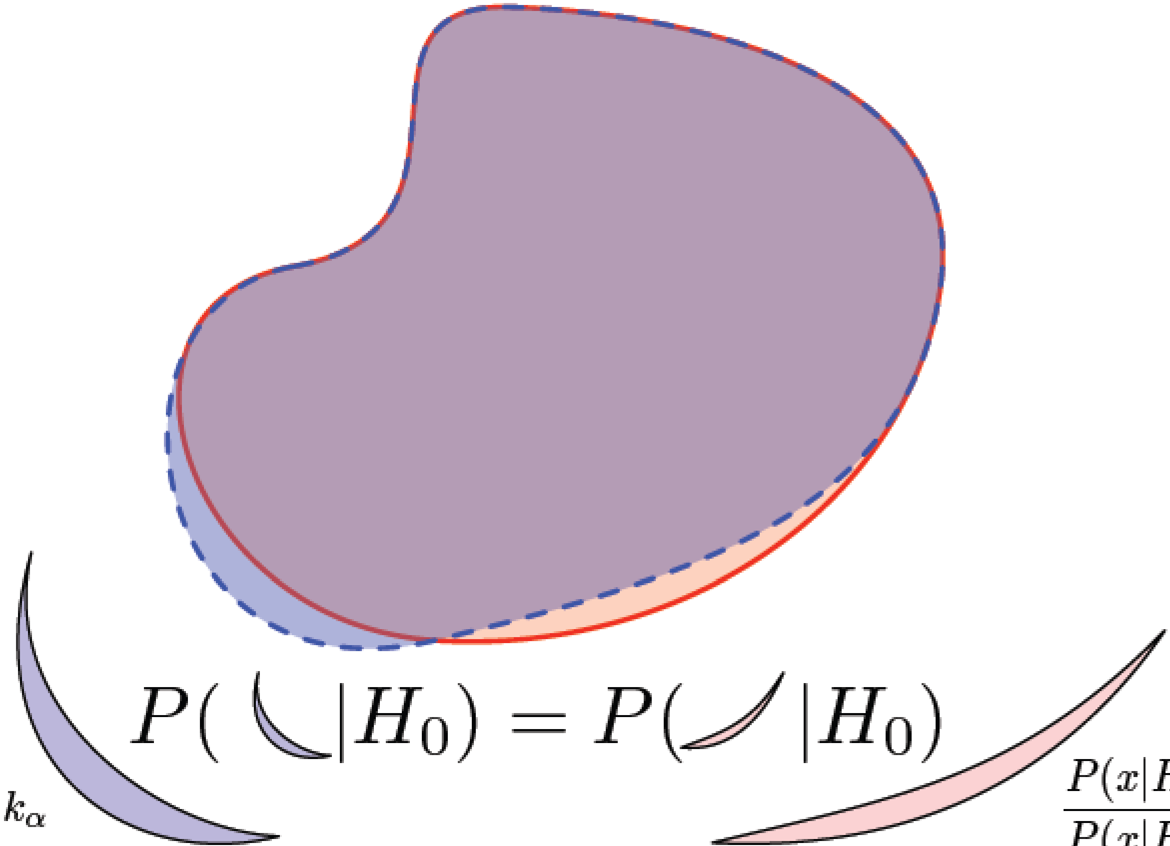
$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$$P(\;|H_1) < P(\;|H_0)k_\alpha$$

$$P(\;|H_1) > P(\;|H_0)k_\alpha$$

Neyman Pearson Lemma

Kyle Cranmer



$$P(\ \ |H_0) = P(\ \ |H_0)$$

$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad\qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$$P(\ \ |H_1) < P(\ \ |H_0)k_\alpha \qquad\qquad P(\ \ |H_1) > P(\ \ |H_0)k_\alpha$$

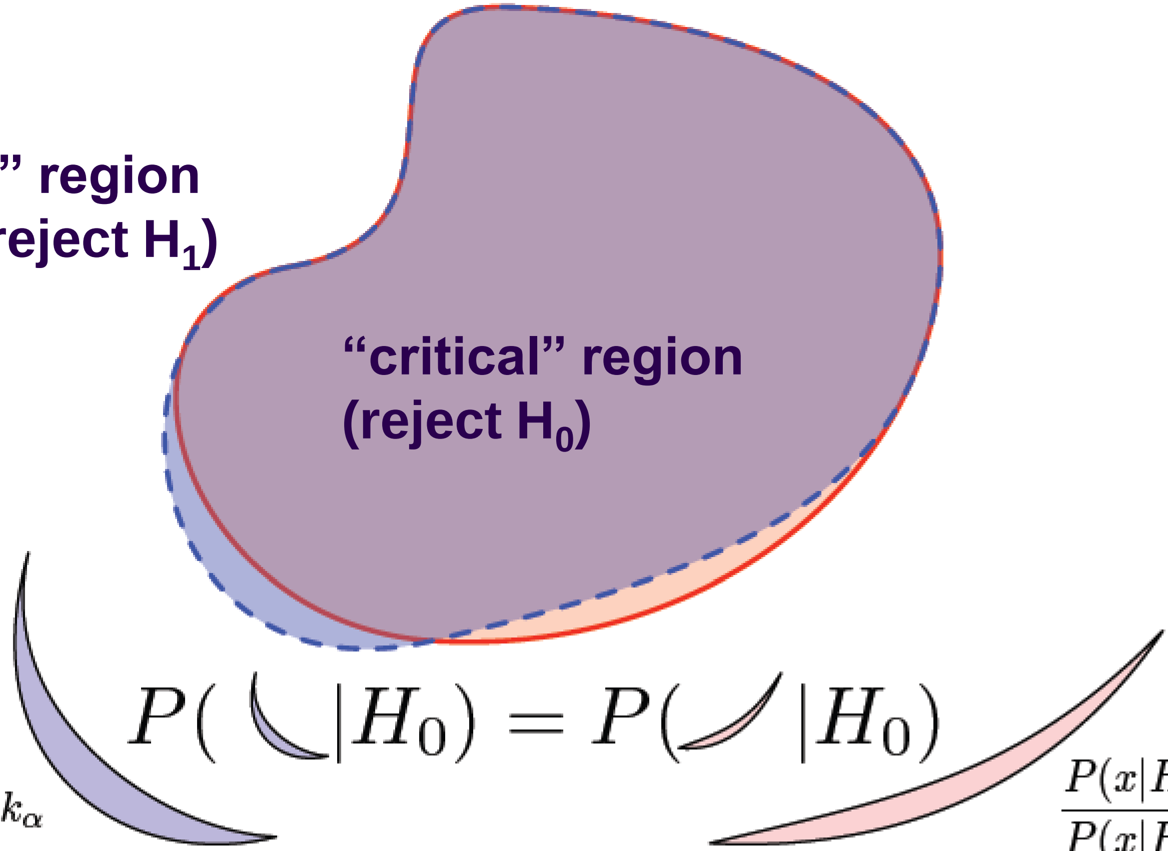
$$\boxed{P(\ \ |H_1) < P(\ \ |H_1)}$$

$$\beta = \int_{!C} P(x|H_1)dx$$

Neyman Pearson Lemma

Kyle Cranmer

"acceptance" region
(accept H_0 (reject H_1)



"critical" region
(reject H_0)

$$P(\ \ |H_0) = P(\ \ |H_0)$$

$$\frac{P(x|H_1)}{P(x|H_0)} < k_\alpha \qquad\qquad \frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

$$P(\ \ |H_1) < P(\ \ |H_0)k_\alpha \qquad\qquad P(\ \ |H_1) > P(\ \ |H_0)k_\alpha$$

The NEW "acceptance" region has less power! (i.e. probability under H_1) q.e.d

Neyman Pearson Lemma

- **Unfortunatly:**
  - **Neyman Pearson's lemma only holds for SIMPLE hypothesis (i.e. w/o free parameters)**
  - **If $H_1 = H_1(\theta)$ i.e. a "composite hypothesis" it is not even sure that there is a so called "Uniformly Most Powerful" test i.e. one that for each given size α is the most powerful (largest 1– β)**
- **Note: even with systematic uncertainties (as free parameters) it is not certain anymore that the Likelihood ratio is optimal**

However: It's probably always your "best guess" ☺

Frequentist Confidence Intervals

- Frequentists CANNOT make statements like: the probability of the theory being true/ parameter having this values is….
  - although we might do so "in our head" at the end anyway (or by "crowd sourcing" like Heuer) → beware!

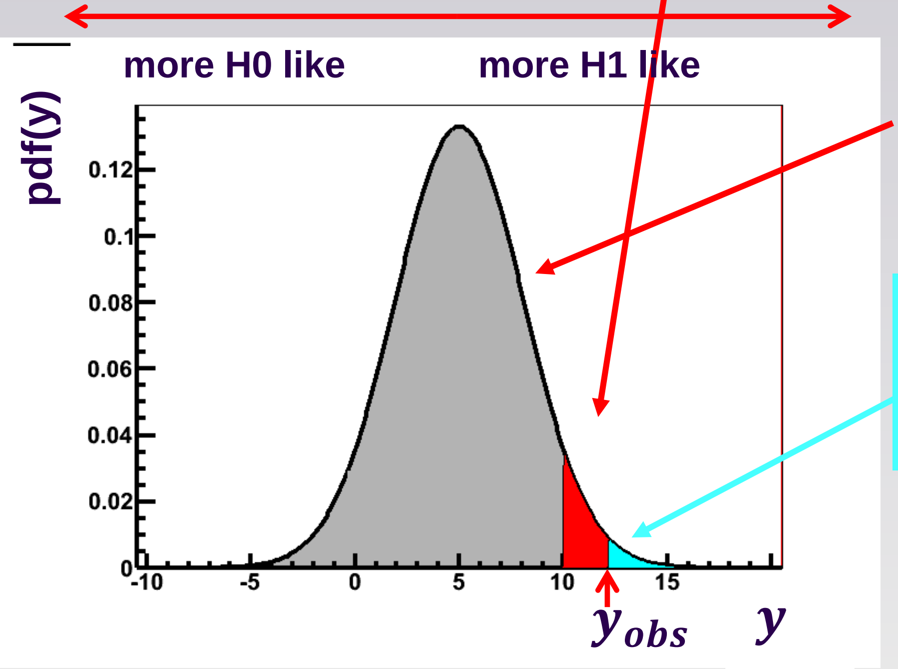
- What do frequentists do?
  - define acceptance (rejection) region of a test (α)
  - measurement/data → just one outcome of a whole set of possible data
  - accept or reject H0 with "confidence level" given by α

  - we might also report how "likely" the particular measurement/observation/data is for given "theories"/"true values" → p-value

→ quote result and it's confidence:

Typical Frequentist Analsysis

- specify your "estimator" (i.e. the Likelihood ratio)
- specify the "significance" α of the test
  - → i.e. how likely you are willing to claim a false discovery
  - → Confidence Level 95% ⇔ $\alpha = 5\%$
- measurement → y_{obs}
- check: result inside or outside the "critical region" ? → decide on H0
  - calculate p-value → how "well" you are within the critical region



PDF of your "estimator" y under H0

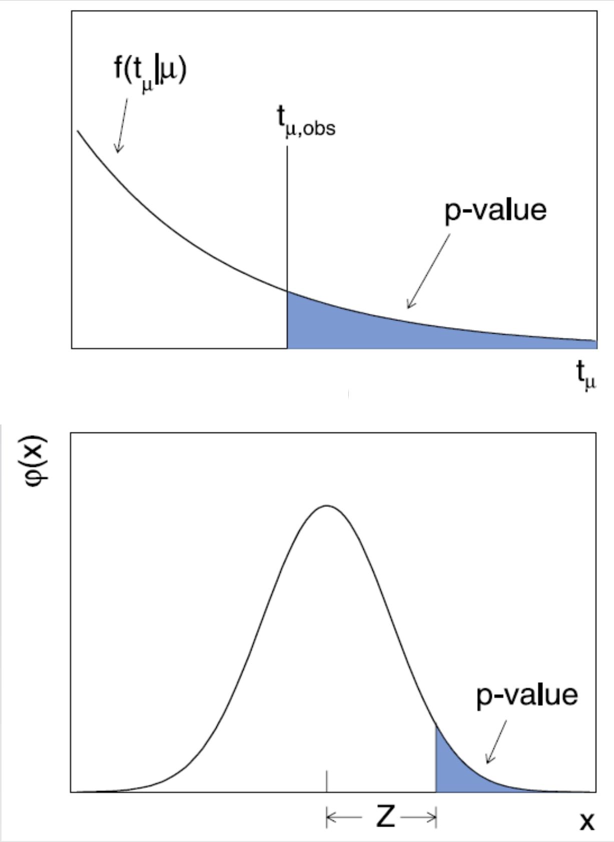
(probability density function)

$$\int_{x_{obs}}^{\infty} p(x')dx' \equiv p - value$$

Size α and P-Value

Note:

- α (significance) is specified "BEFORE" the measurement/test
- p-value is property of the actual measurement
- p-value is NOT a measure of how probably the hypothesis is



Note:

the Confidence Level of your "discovery" (or limit) is given by α, NOT the p-value!!

it's custom to translate p-values to the common "sigma"

→ how many standard deviations "Z" for same p-value on one sided Gaussian

→ 5σ = p-value of $2.87 \cdot 10^{-7}$

Size α and P-Value

- Why α i.e. needs to be specified "BEFORE" the measurement, isn't the p-value afterwards enough to "decide" ?
- see what would happen:
  - measurement
  - determine p-value
  - discard H0 and state how many other measurements would give p-values "worse" than yours

  → you would simply "always" reject H0 (accept H1) and just state in how many other measurements one would measure s.th. more H1 like..

    → but CL : assume H0 were in fact true, of the ensemble of all possible measurement done following this procedure, 1-CL of them would discard H0

    → 1-CL is the probability that "you" falsely discarded H0

    → BUT: this procedure ALWAYS discards H0 !!!

→ for discoveries: $\alpha = 2.87{\cdot}10^{-7}$ by convention, the famous "5σ"

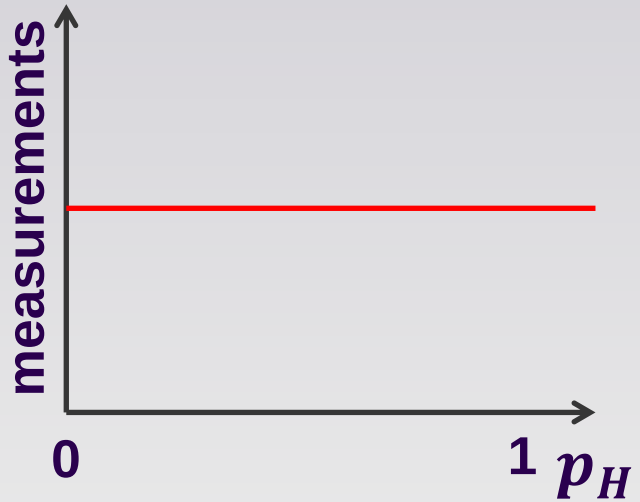
Distribution of P-Values

assume:

- t **: some test statistic** (the thing you measure, i.e. $t = t(x) = t(m, p_t, \ldots)$ or n_{events})
- $p(t|H)$ **: distribution of** t (expected distribution of results that would be obtained if we were to make many independent measurements/experiments)
- **p-value :** $p_H = \int_t^{\infty} p(t'|H)dt'$ (for each hypothetical measurement)

→ **p-values are "random variables" → distribution**

→ **derived from the "cumulative distribution" → FLAT under H**



- **remember: χ^2 and e.g. straight line fit**
  - **χ^2 probability is flat**
  - **value tell you "how unlucky" you were with your "goodness-of-fit" (χ^2 at the best fit)**
  - **up to you to decide if you still trust the model**

Statistical Tests in Particle Searches

exclusion limits

- **upper limit on cross section**
  (↔ or lower limit on mass scale)
- **(**σ < limit as otherwise we would have seen it**)**

➧ **need to estimate probability of downward fluctuation of s+b**

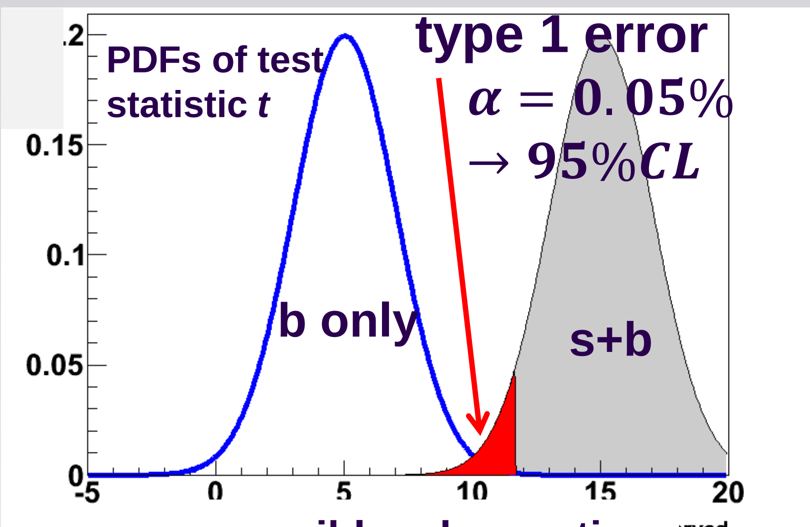
➧ **try to "disprove" H_0 = s+b**

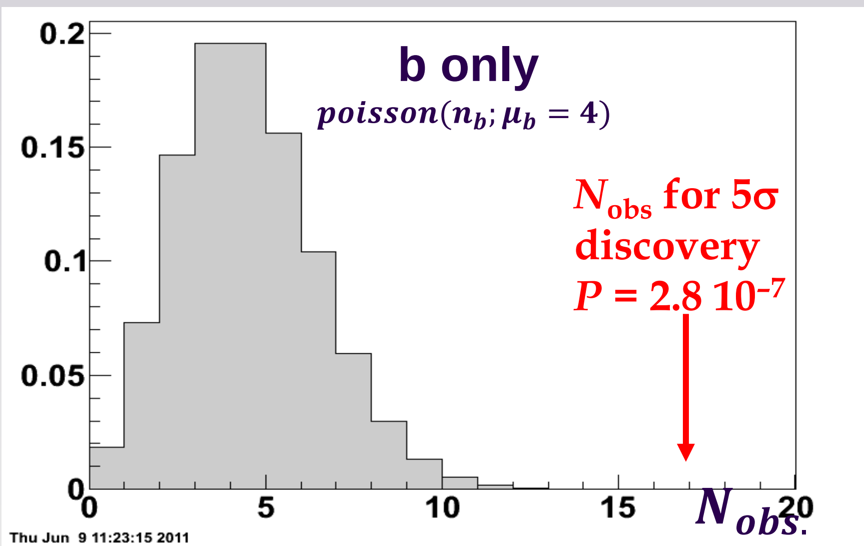
➧ **or: find minimal s, ($\mu=\sigma/\sigma_{SM}$) for which you can still exclude H_0 = s+b at pre-specified Confidence Level**

discoveries

➧ **need to estimate probability of upward fluctuation of b**

➧ **try to disprove H_0 ="background only"**





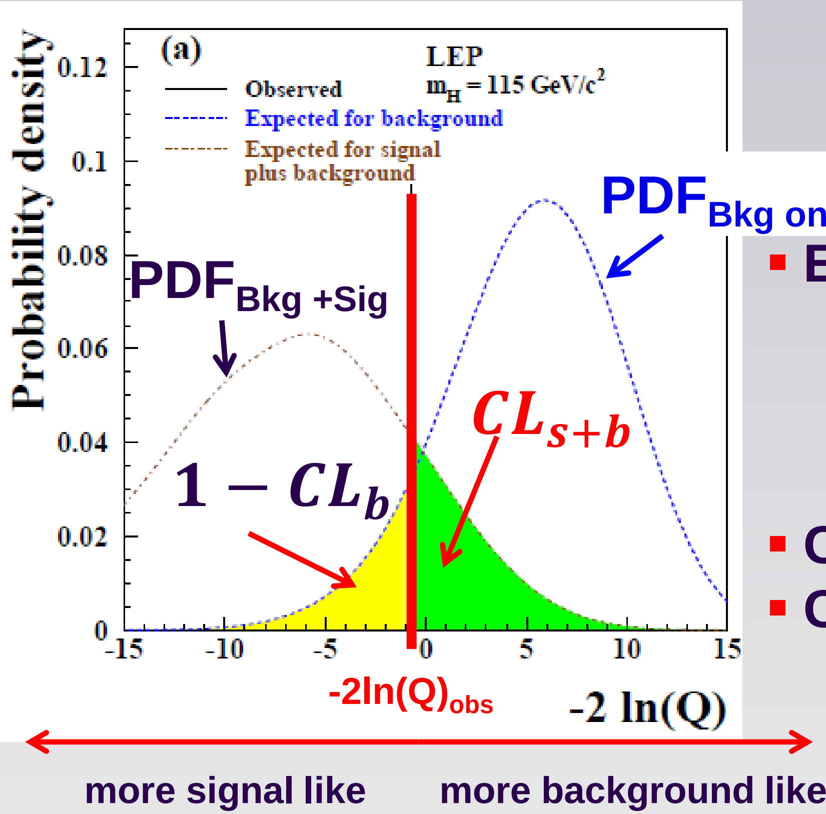
Example :LEP-Higgs search

which test statistic?

$$t(n_{obs}) = \frac{Poisson(n_{obs}; s, b)}{Poisson(n_{obs}; b)}$$

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

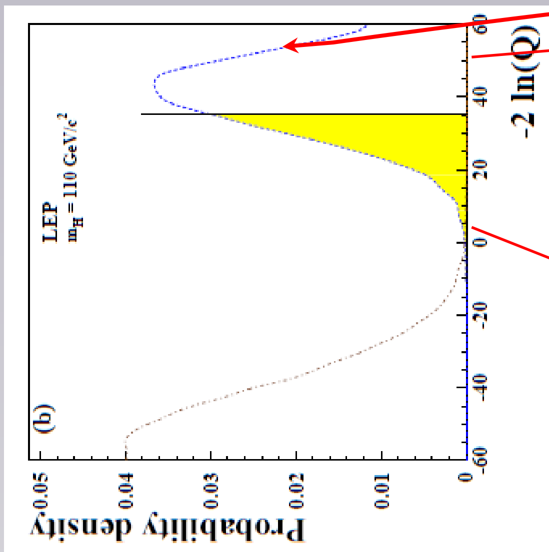
$$q = \ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$

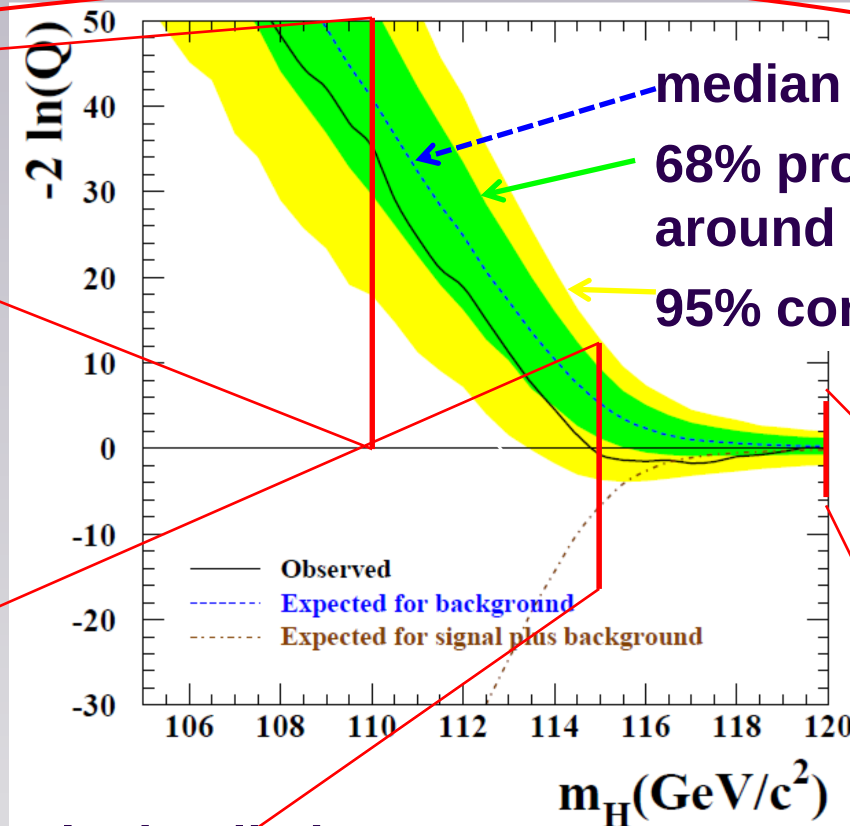


- Evaluate how the -2lnQ is distributed for
  - background only
  - signal (m_H=115GeV/c²) +background
    - (note: needs to be done for all Higgs masses)

- CL_{s+b}: p-value under s+b hypothesis
- CL_b : p-value under bkg only hypothesis

Example: LEP SM Higgs Limit

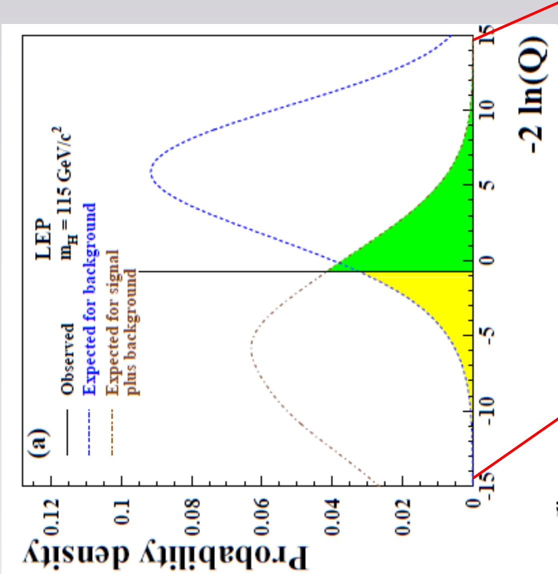




median

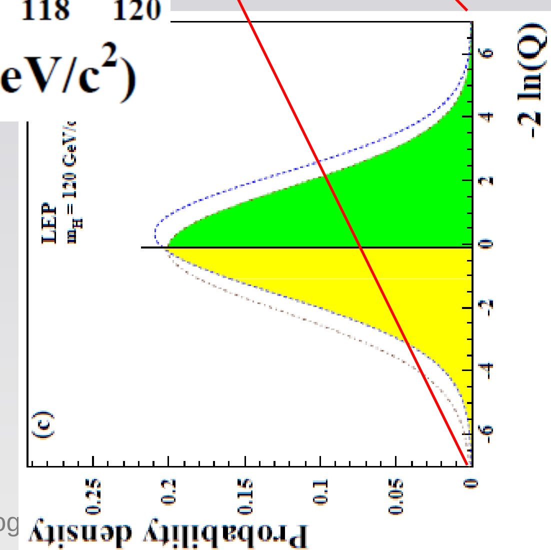
68% prob. content around median

95% content of



Exclusion limit → CL_{s+b}

- draw "bands" around expectation for signal+bkg
- excluded at 95%CL where "observed" lies outside 95% CL band



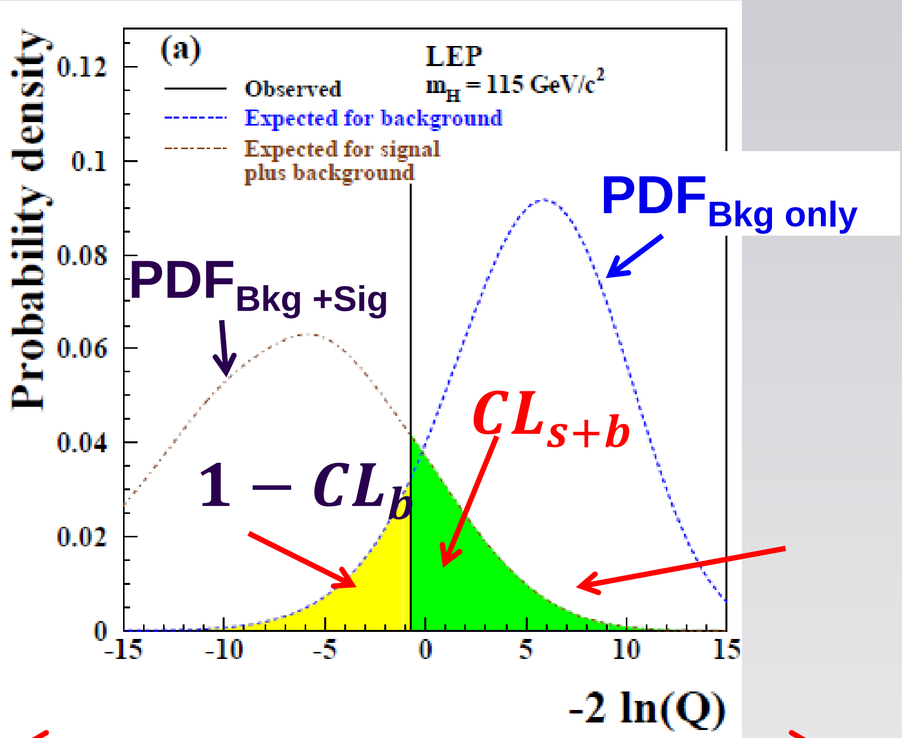
Example LEP Higgs Search

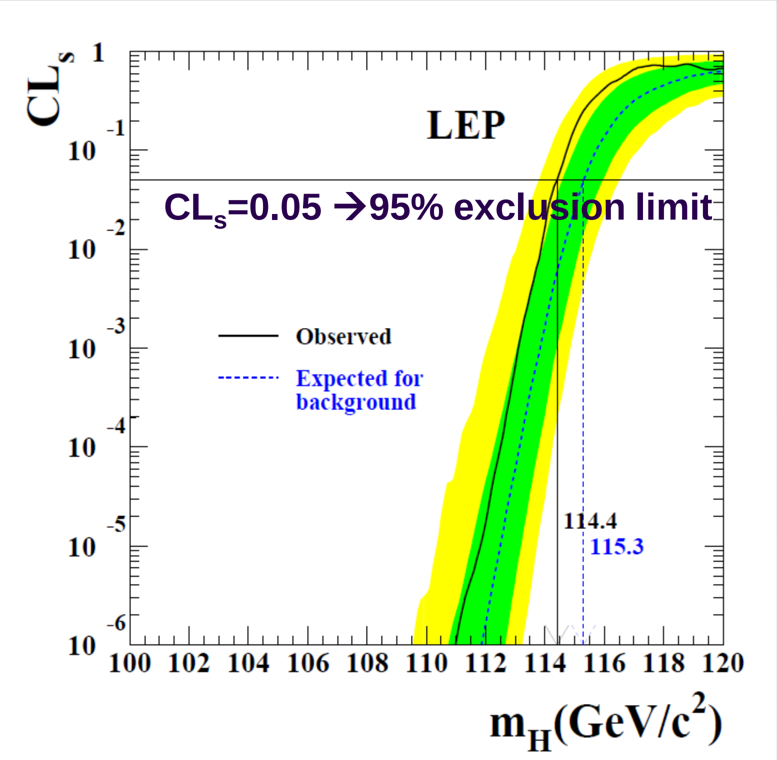
- "avoid" the possible "problem" of Being Lucky when setting the limit
  (we'll come back to that, later…)

→ rather than "quoting" in addition the expected sensitivity

→ weight your CL_{s+b} by it:

$$CL_s = \frac{p_{s+b}}{1-p_b} = \frac{CL_{s+b}}{1-CL_b} = \frac{P(LLR \geq LLR_{obs}|H_1)}{P(LLR \leq LLR_{obs}|H_0)}$$

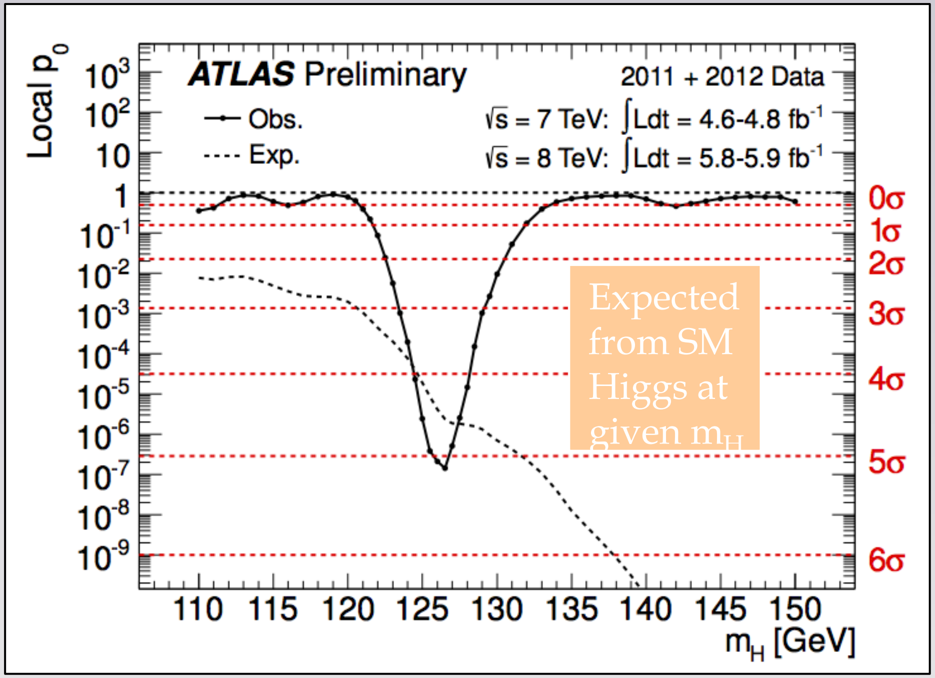




more signal like ↔ more background like

ATLAS/CMS Higgs Search

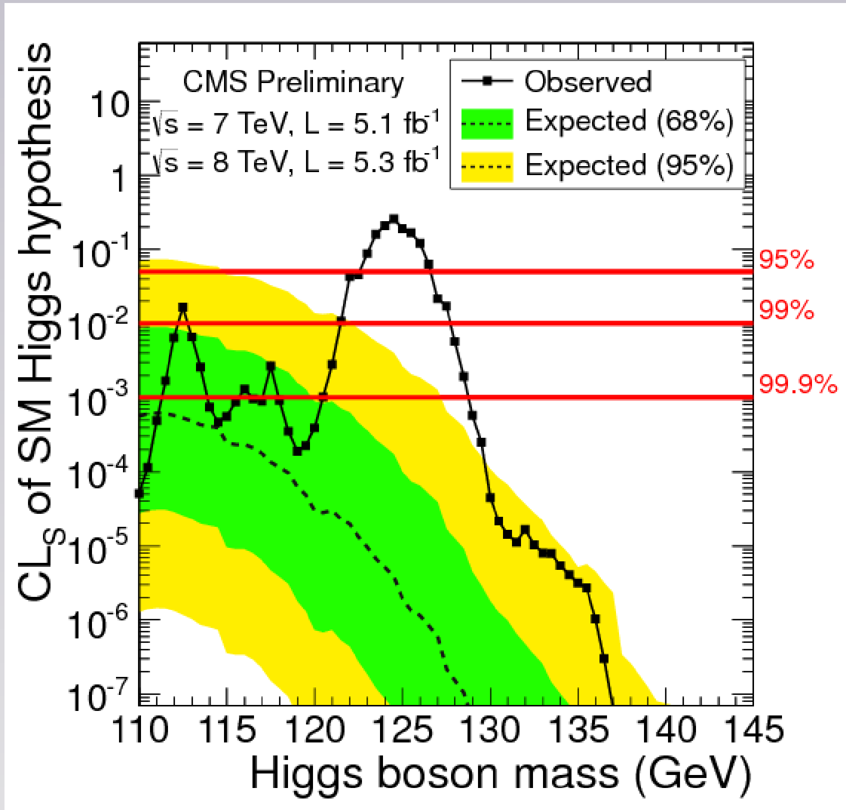
- Aim for DISCOVERY → disprove H_0 = background ONLY
  - somewhat different test statistic: profile Likelihood ratio of Likelihood function $L(\mu,\theta), with\ \ \mu = \frac{\sigma}{\sigma_{SM}}, \theta{:}\, nuisance\ parameters$
  - p-value for discovery: Bkg only hypothesis ($\mu = 0$)
- p-value calculated "locally" every Higgs mass
- Look at any "dip" in p-values over whole mass range
  - think as "binned" in Higgs mass resolution
- Random samples of a distribution, histogram it → 1 out of 20 bins (5%) will deviate 2σ from expectation.. e.t.c.



- LOOK-ELSEWHERE-EFFECT
- not taken into account → local p-value

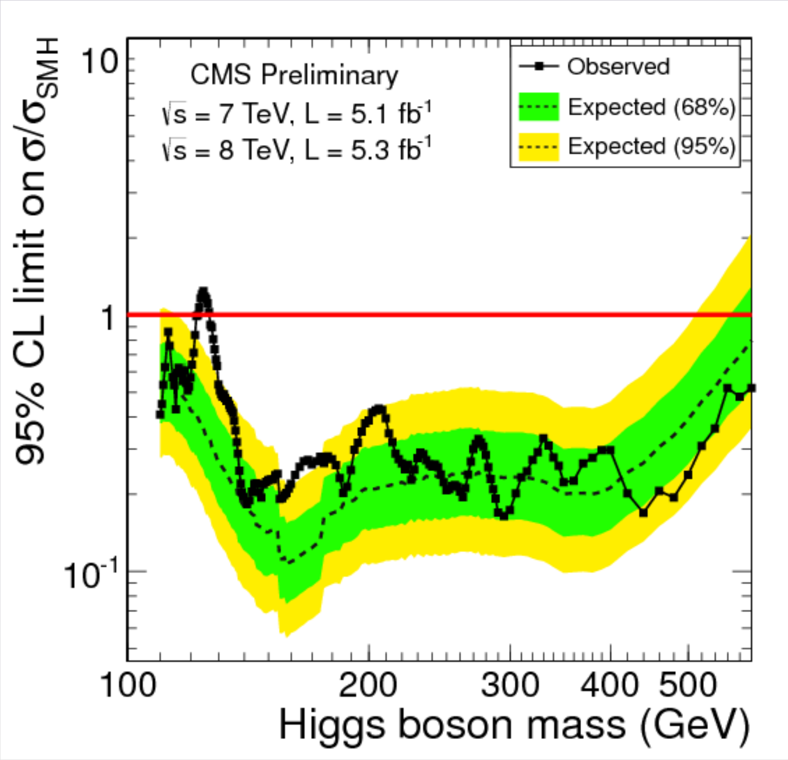
CL_s and Excluded Cross Section

- $CL_s = \frac{p_{s+b}}{1-p_b}$



- adjust $\mu = \frac{\sigma}{\sigma_{SM}}$ such that $CL_s = 95\%$

→ limit on $\mu = \frac{\sigma}{\sigma_{SM}}$



Message:

They can nicely exclude everything at "high Confidence levels" apart from where they see the signal

Summary

- **Reiterated differences of Bayesian ↔ frequentist**
- **Frequentist Hypothesis testing**
  - **Neyman Pearson → Likelihood ratio**
  - **HEP particle searches**
    - limits
    - discoveries

- **Example: LEP: CLs … the HEP limit; ATLAS/CMS Higgs discovery**
  - **CLs … ratio of "p-values" … statisticians don't like that**
  - **new idea: Power Constrained limits**
    - rather than specifying "sensitivity" and "Neyman conf. interval"
    - decide beforehand that you'll "accept" limits only if the where your experiment has sufficient "power" i.e. "sensitivity !

    → lots of "different" ideas floating around how to "set limits"

    → Hey! We don't need that anymore …well at least not for the Higgs.. ☺