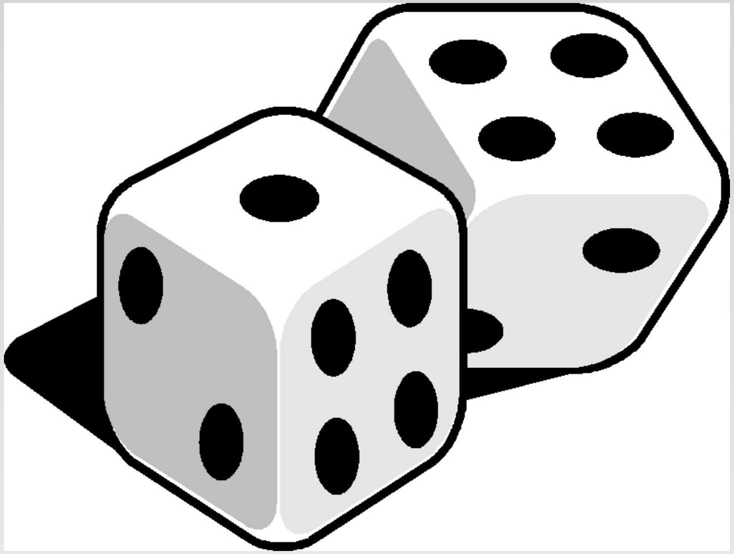
Statistics In HEP

How do we understand/interpret our measurements



Outline

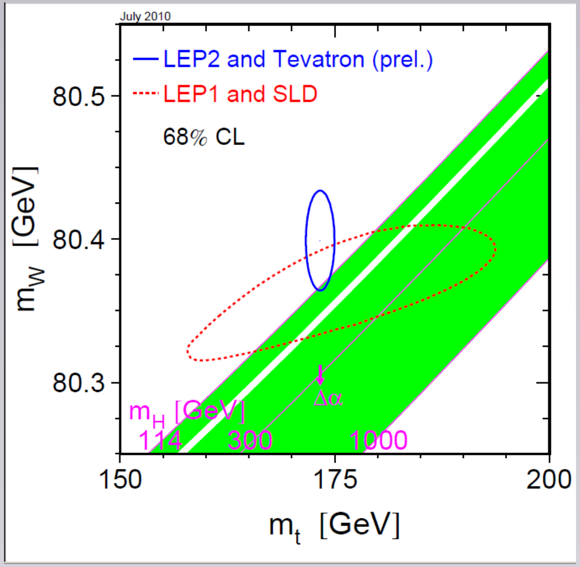
- What is Probability : frequentist / Bayesian
  - review PDFs and some of their properties

- Hypothesis testing
  - test statistic
  - power and size of a test
  - error types
  - Neyman-Pearson → What is the best test statistic
  - concept of confidence level/p-value
- Maximum Likelihood fit
- strict frequentist Neyman – confidence intervals
  - what “bothers” people with them
- Feldmans/Cousins confidence belts/intervals

- Yes, but what about systematic uncertainties?

Interpreting your measurement

- What do we REALLY mean by:
  - m_W = 80.399+/- 0.023 ;
  - M_{Higgs}< 114.4GeV/c^2 @95%CL



- these things are results of:
  - involved measurements
  - many "assumptions"

- correct statistical interpretation:

→ most 'honest' presentation of the result

→ unless: provide all details/assumptions that went into obtaining the results

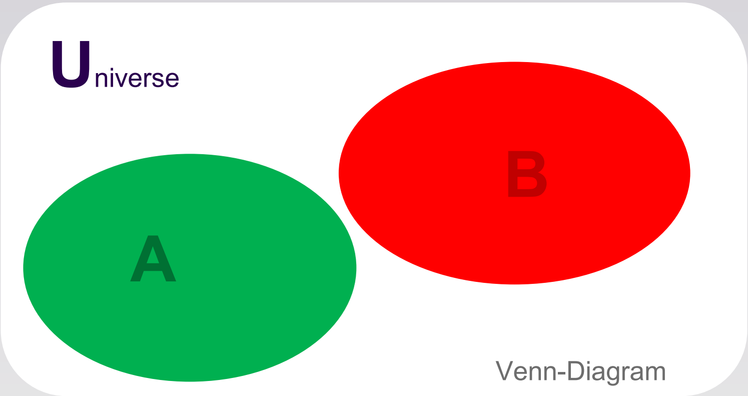
→ needed to correctly "combine" with others (unless we do a fully combined analysis)

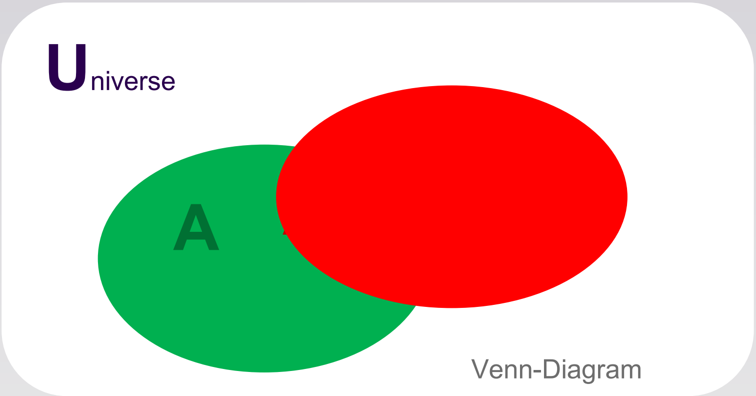
What is Probability

- **Axioms of probability: Kolmogorov (1933)**
  - $P(E) \geq 0$
  - $\int_U P(E)\,dU = 1$
  - if $A \cap B = 0$ (i.e disjoint/independent events) then $P(A \cup B) = P(A) + P(B)$

→ given those we can define e.g.: **conditional probability:**

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





What is Probability

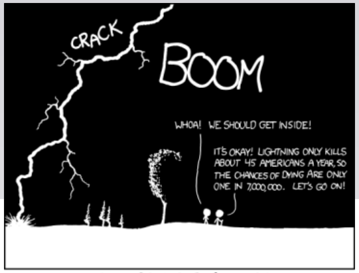
- Axioms of probability: → pure “set-theory”

1) a measure of how likely it is that some event will occur; a number expressing the ratio of favorable – to – all cases

- Frequentist probability

$$P(\text{“Event”}) = \lim_{n\to\infty}\left(\frac{\#\text{outcome is "Event"}}{n-\text{"trials"}}\right)$$

2) the quality of being probable; a probable event or the most probable event (WordNet® Princeton)



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

- **Bayesian probability:**
  - **P(“Event”): degree of believe that “Event” is going to happen**
  - **fraction of possible worlds in which “Event” is going to happen….**

Frequentist vs. Bayesian

Bayes' Theorem $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(B|A)\,\frac{P(B)}{P(A)}$$

- This follows simply from the "conditional probabilities":

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

$$P(A|B)P(B) = P(B|A)P(A)$$

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

Frequentist vs. Bayesian

Bayes' Theorem $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = P(B|A)\,\frac{P(B)}{P(A)}$$

- Some people like to spend hours talking about this…

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: $P(Data|Model)$

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

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 is raining tomorrow?"
    - "How likely is it that the LHC/Tevatron will also be operated next year?"
  - 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"!

Probability Distribution/Density of a Random Variable

random variable x : characteristic quantity of point in sample space

discrete variables

$$P(x_i) = p_i$$

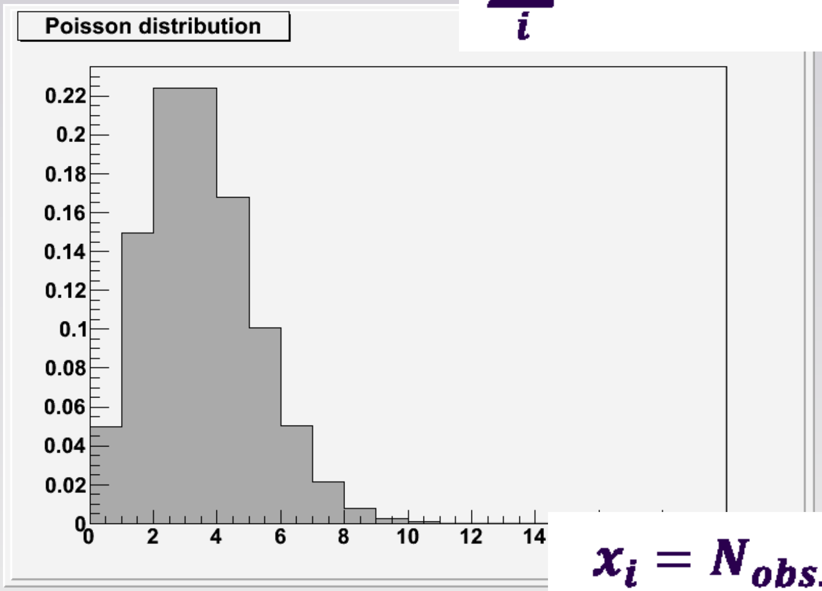
continuous variables

$$P(x \in [x, x+dx]) = f(x)dx$$

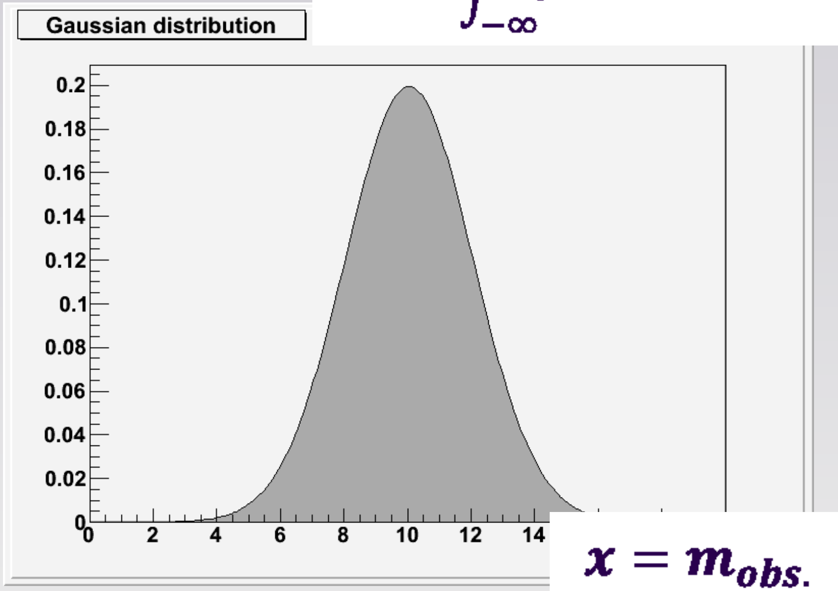
normalisation (It has to be 'somewhere')

$$\sum_i P(x_i) = 1$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

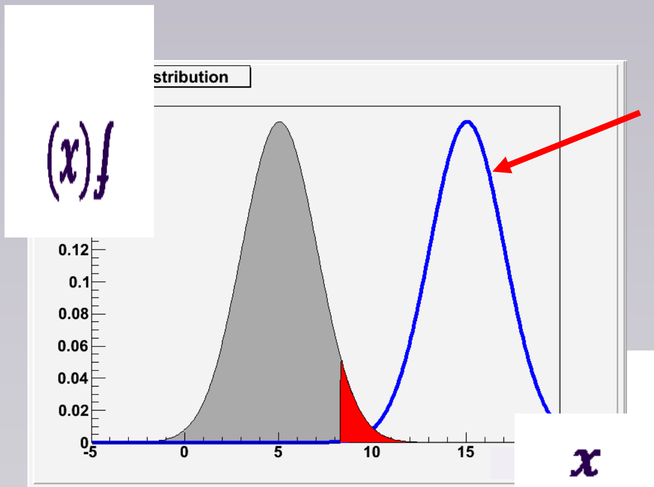


$$x_i = N_{obs.}$$



$$x = m_{obs.}$$

Cumulative Distribution

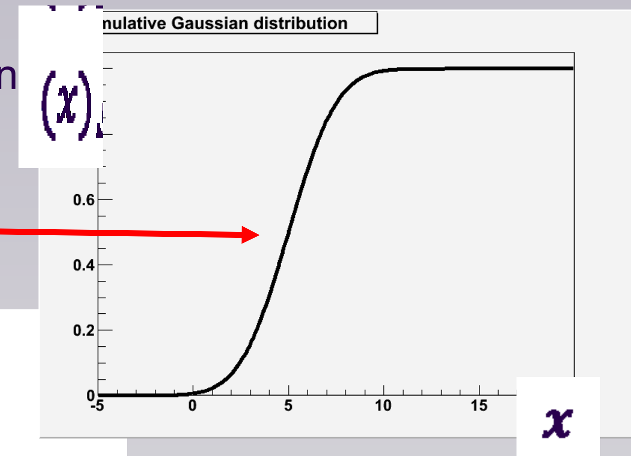


PDF (probability density function

Cumulative PDF distribution:

$$\int_{-\infty}^{x} f(x')dx' \equiv F(x)$$

$$\rightarrow f(x) = dF(x)/dx$$



assume:

- **$f(x)$: probability distribution for some "measurement" x under the assumption of some model (parameter)**
- **$f'(x)$: PDF of some alternative model (parameter)**
- **Imagine you measure x_{obs}**
  - **1-$\int_{-\infty}^{x_{obs}} f(x')dx' \equiv p - value$ for observing something at least as far away from what you expect**
- **red area: the data space where the $p - values$ are < 5%**

we will come back to this..

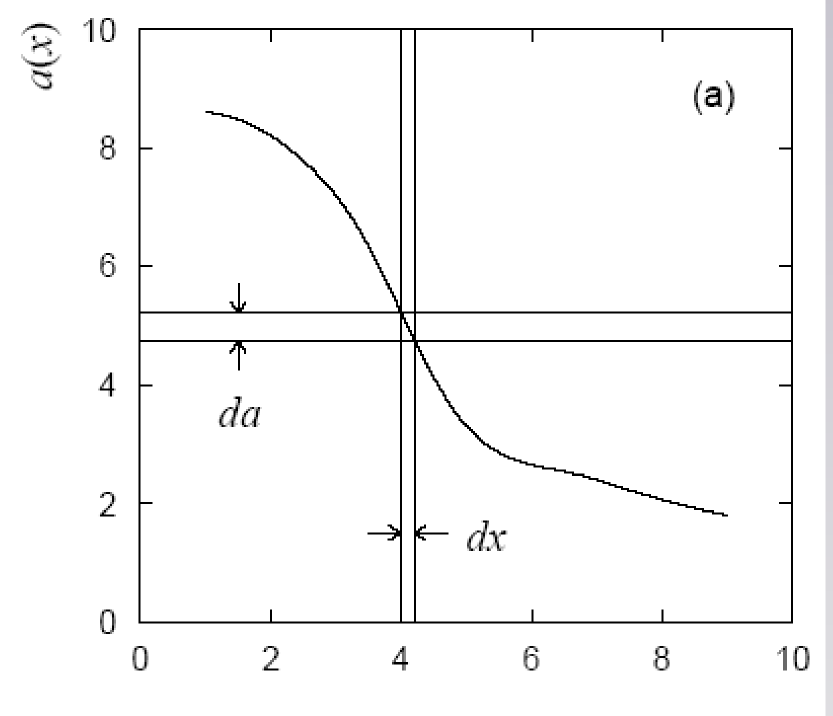
Functions of Random Variables

- A function of a random variable is itself a random variable.
  - x with PDF $f(x)$
  - function a(x)
- PDF $g(a)$?

$$g(a)da = \int_{dS} f(x)dx$$

here: dS =region of x space for which

- a is in [a, a+da].
- For one-variable case with unique
- inverse this is simply:



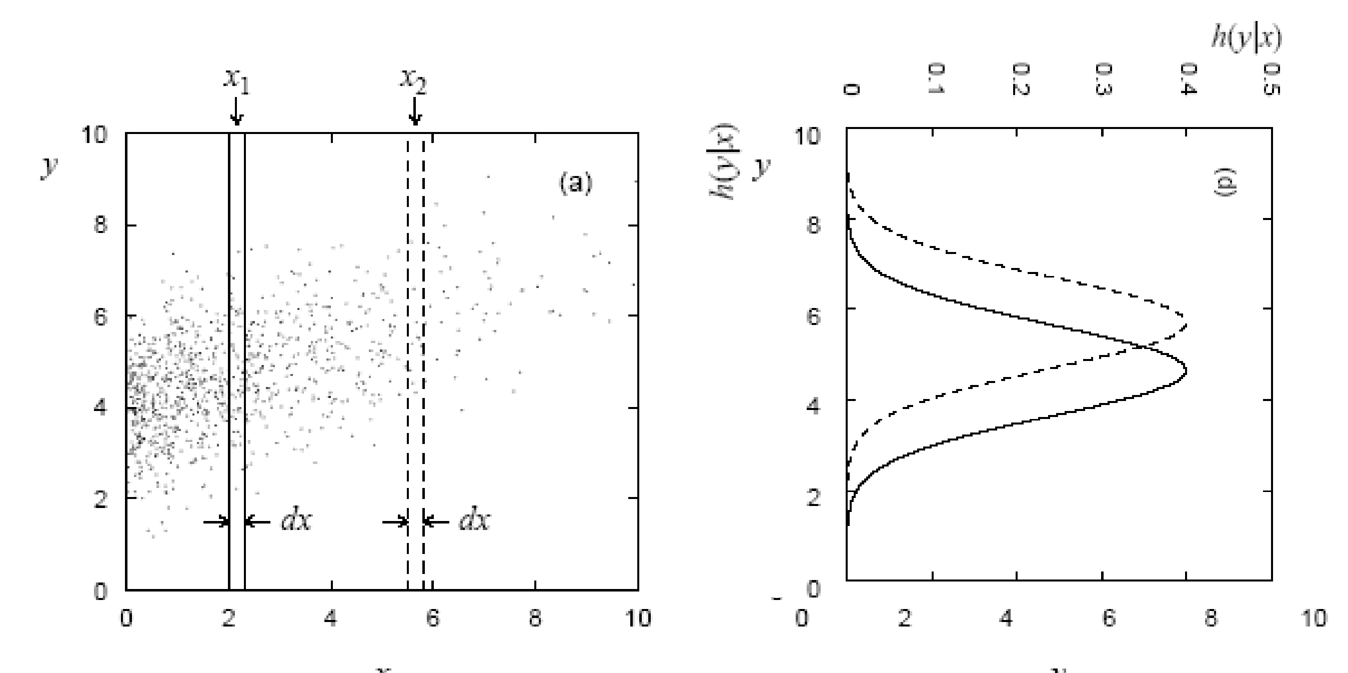
Glen Cowan: Statistical data analysis

$$g(a)da = f(x)dx \dashrightarrow g(a) = f(x(a))\left|\frac{dx}{da}\right|$$

Conditioning and Marginalisation

- **conditional probability:** $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{f(x,y)dxdy}{f_x(x)dx}$

↔ consider some variable in the joint PDF(x,y) as constant (given):



Glen Cowan: Statistical data analysis

- **marginalisation:** If you are not interested in the dependence on "x"
→ project onto "y" (integrate "x out")

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
    - some NEW PHYSICS or Standard Model
  - e.g. H a particular parameter in a diff. cross section
    - some mass / coupling strength / ~~CP~~ parameter

- Simple (point) hypothesis
  - completely specified, no free parameter
    - → PDF: $f(x) \equiv f(x; H)$
- Composite hypothesis
  - H contains unspecified parameters (mass, systematic uncertainties, …)
    - → a whole band of $f(x; H(\theta))$
    - → for given x the $f(x; H(\theta))$ 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

Why talking about "NULL Hypothesis"

- Statistical tests are most often formulated using a
  - "null"-hypothesis and its
  - "alternative"-hypotheisis

- Why?

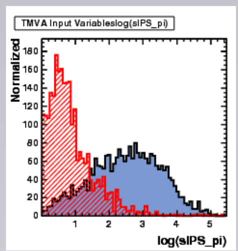
  → it is much easier to "exclude" something rather than to prove that something is true.

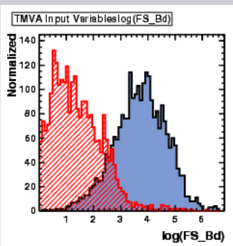
    → excluding: I need only ONE detail that clearly contradicts

  → assume you search for the "unknown" new physics.

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

Example: event classification Signal(H_1) or Background(H_0)

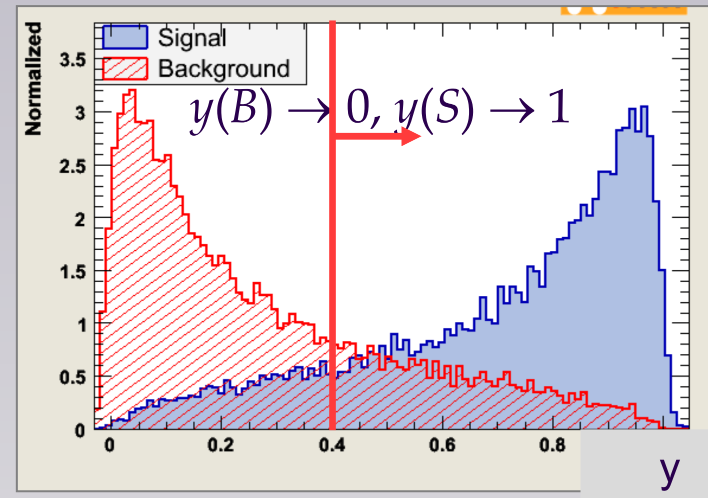




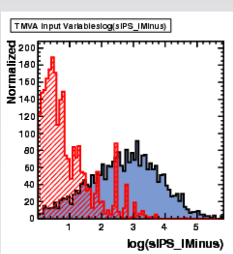
- **Test statistic**
  $y(x_1, x_2, \ldots, x_n): R^n \rightarrow R$

- **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): > cut: signal
= cut: decision boundary
< cut: background



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

Hypothesis Testing

Type-1 error: (false positive)

→ accept as signal although it is background

Type-2 error: (false negative)

→ reject as background 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:

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

should be small

Size β: Type-2 error rate:

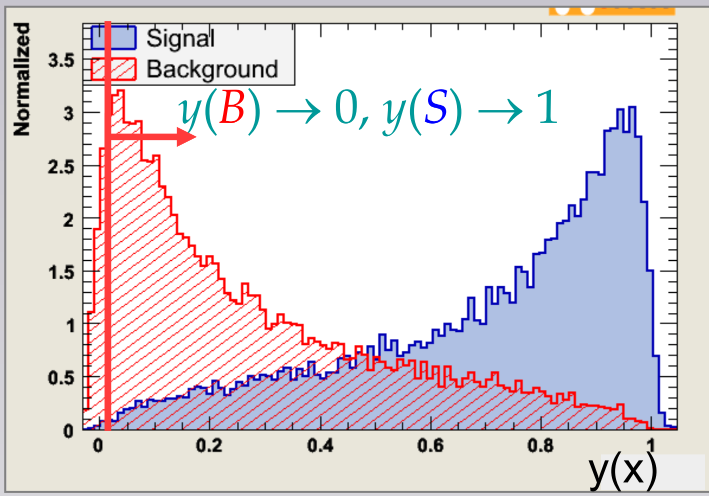
Power: $1-\beta$

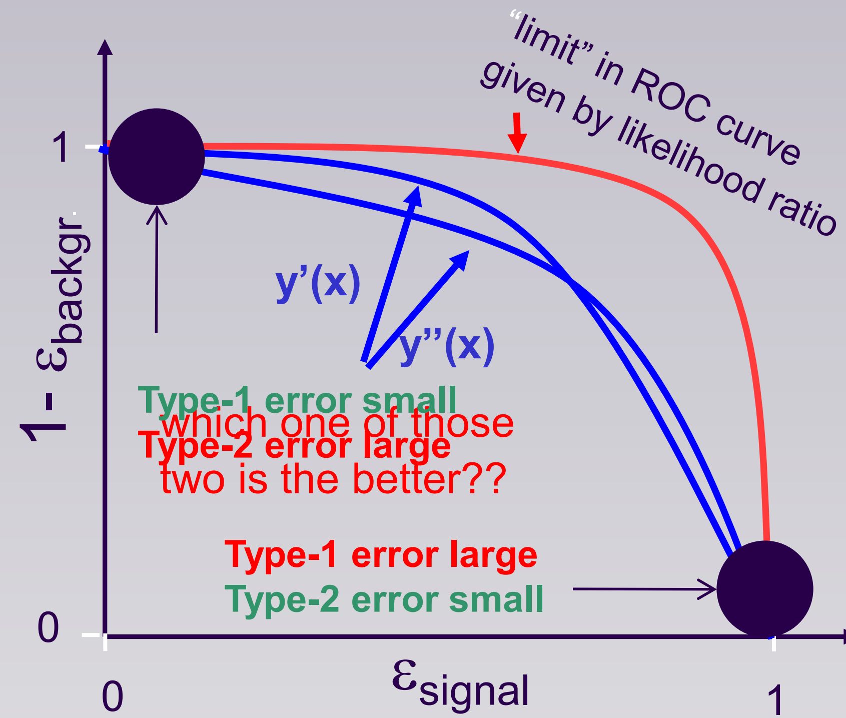
$$\beta = \int_{!C} P(x|H_1)dx$$

should be small

Type 1 versus Type 2 Errors

Signal(H_1) /Background(H_0) discrimination:





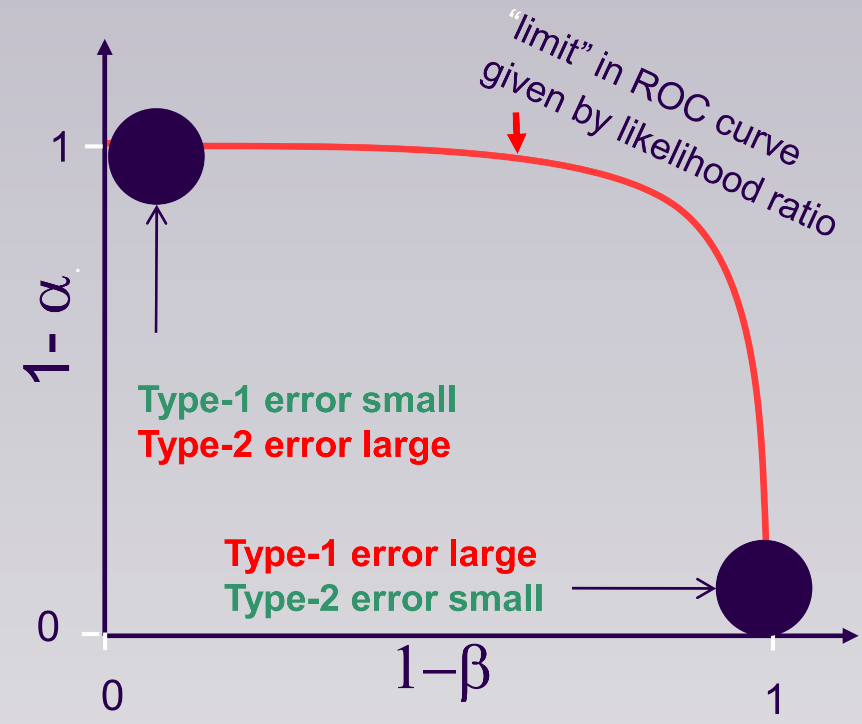
Signal(H_1) /Background(H_0) :

- **Type 1 error: reject H_0 although true → background contamination**
  - Significance α: background sel. efficiency 1– α: background rejection
- **Type 2 error: accept H_0 although false → loss of efficiency**
  - Power: 1– β signal selection efficiency

Neyman Pearson Lemma

Likelihood Ratio: $y(x) = \frac{P(x|S)}{P(x|B)}$

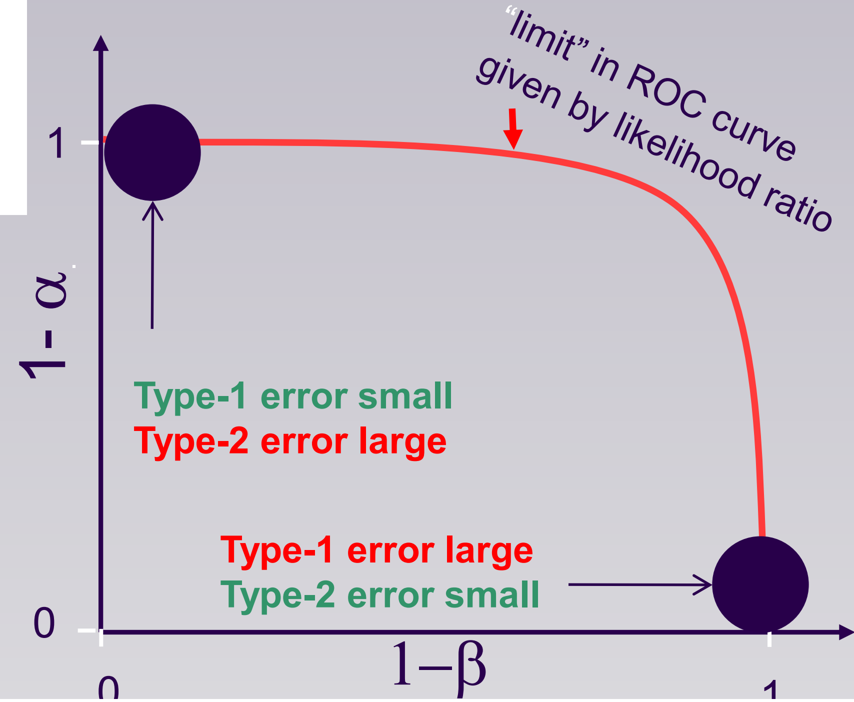
Neyman-Peason:
The Likelihood ratio used as "selection criterium" y(x) gives for each selection efficiency the best possible background rejection.
i.e. it maximises the area under the "*Receiver Operation Characteristics*" (ROC) curve



Neyman Pearson Lemma

Likelihood Ratio: $t(x) = \dfrac{P(x|H_1)}{P(x|H_0)}$

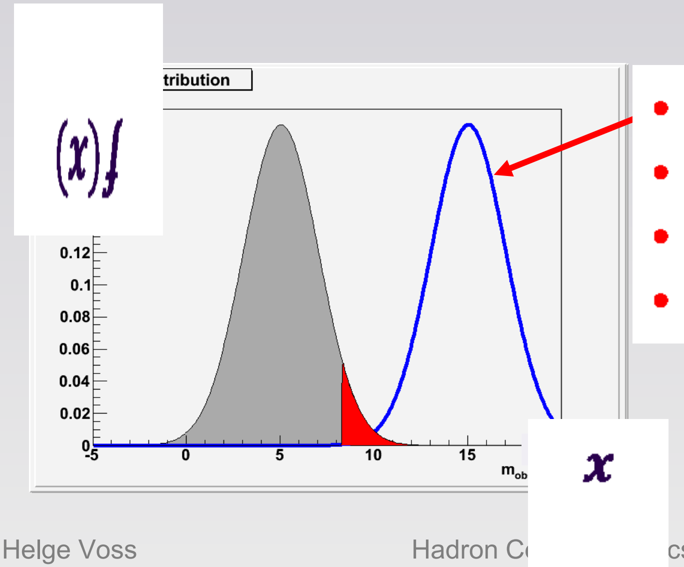
- or any monotonic function thereof, e.g. log(L)



Neyman-Peason:

The Likelihood ratio used as “test statistics” t(x) gives for each size α the test with the largest power $1-\beta$.

i.e. it maximises the area under the “*Receiver Operation Characteristics*” (ROC) curve

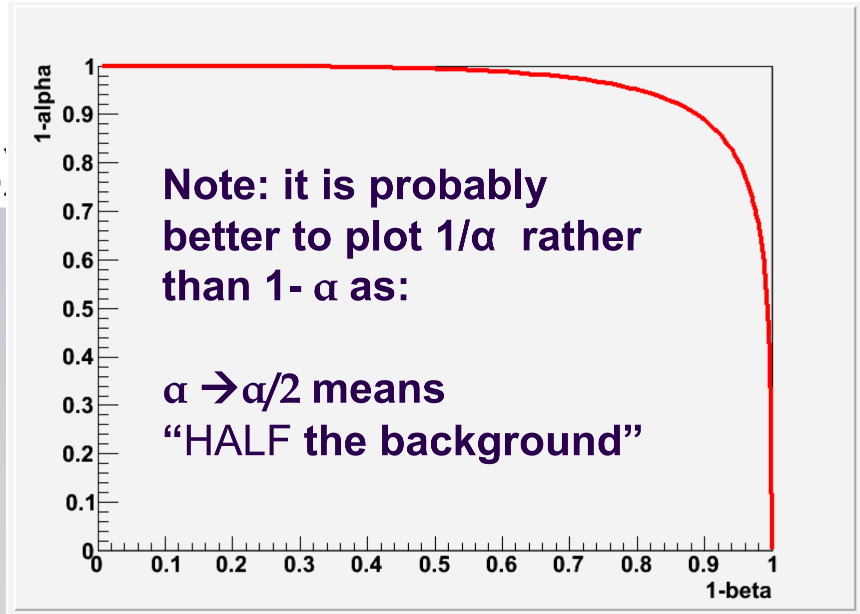


- measure x
- want to discriminate model H_1 from H_0
- H_1 predicts x to be distributed acc. to $P(x|H_1)$
- H_0 predicts x to be distributed acc. to $P(x|H_1)$

Neyman Pearson Lemma

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

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

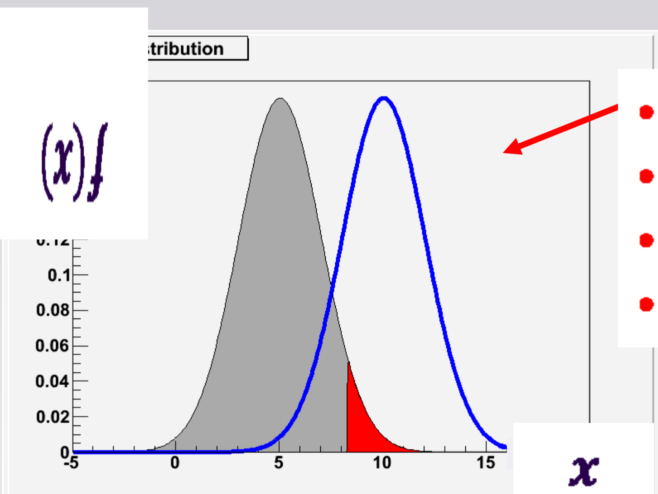


Note: it is probably better to plot 1/α rather than 1- α as:

α → α/2 means
"HALF the background"

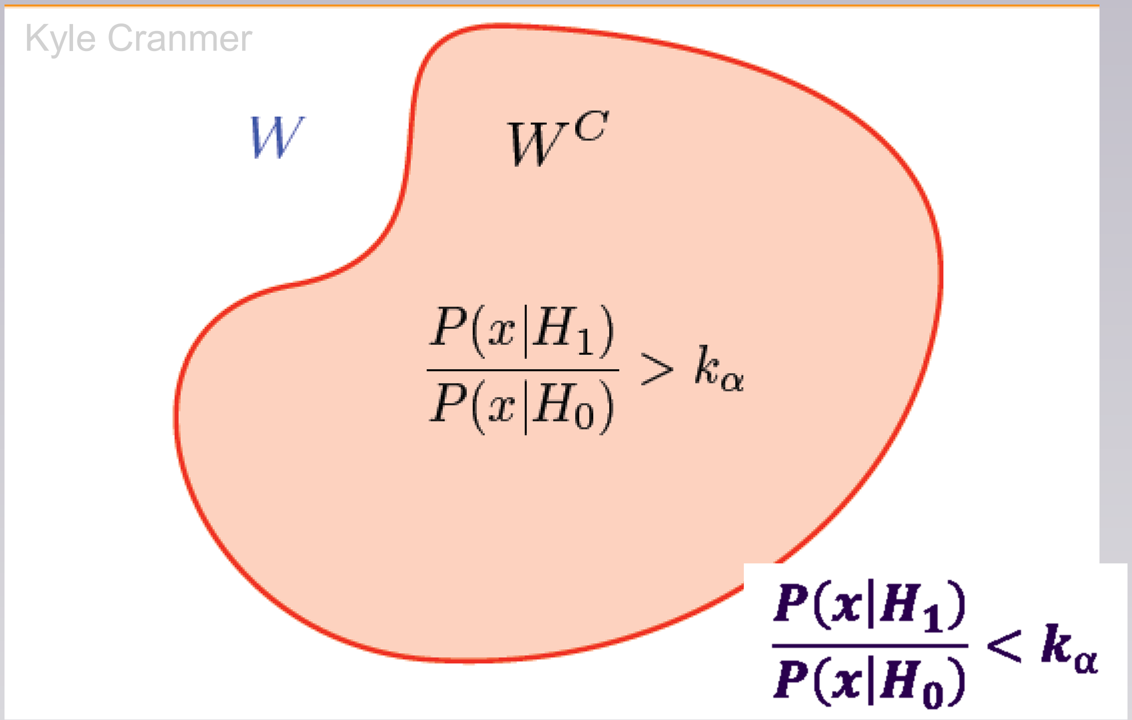
Neyman-Peason:

The Likelihood ratio used as "test statistics" t(x) gives for each size α the test 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_1 predicts x to be distributed acc. to $P(x|H_1)$**
- **H_0 predicts x to be distributed acc. to $P(x|H_1)$**

Neyman Pearson Lemma



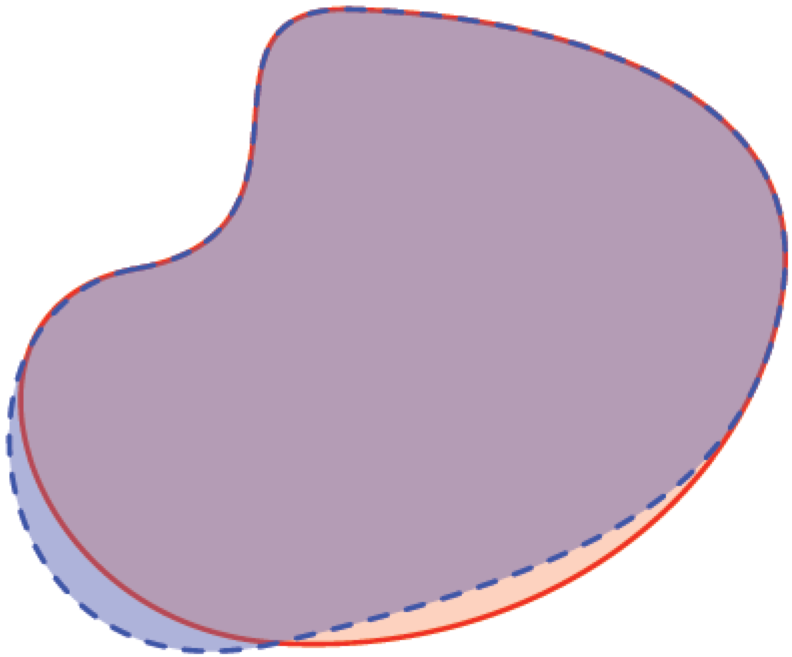
graphical proof of Neyman Pearson's Lemma:
(graphics/idea taken from Kyle Cranmer)

- **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

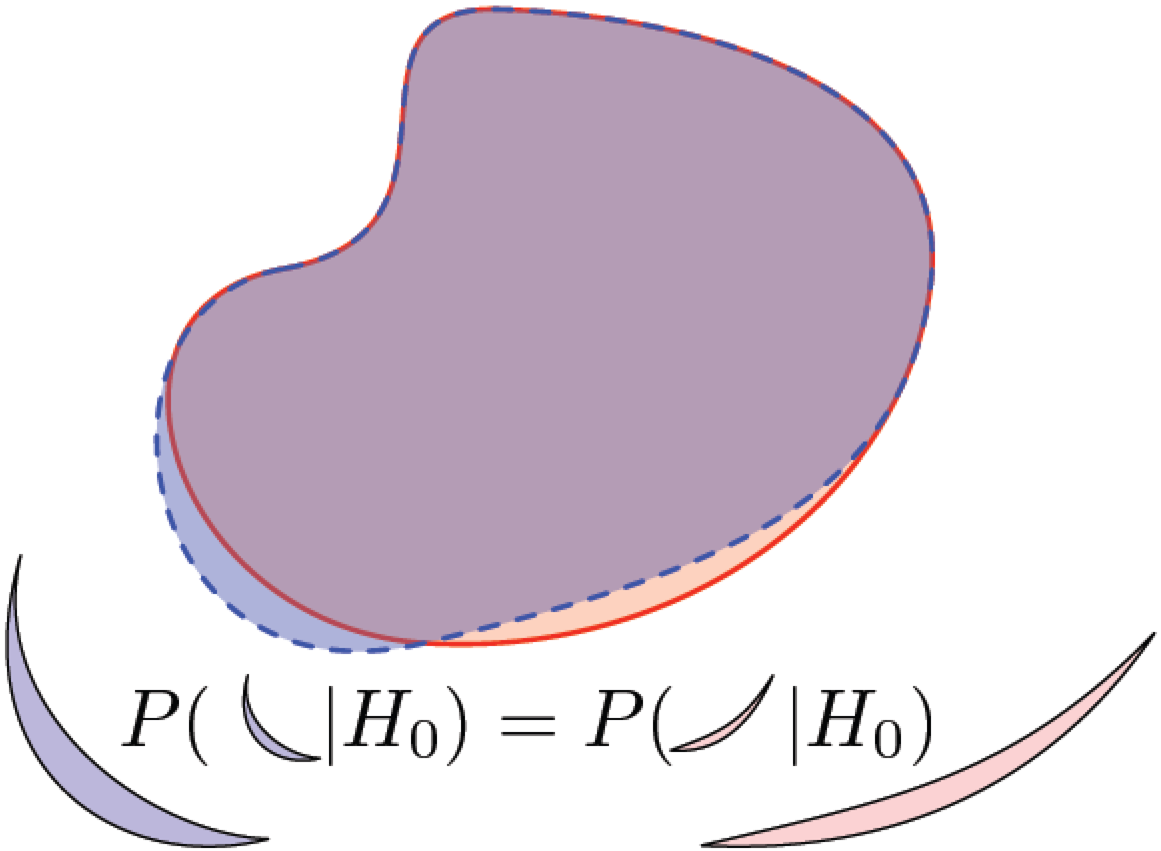
Kyle Cranmer



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



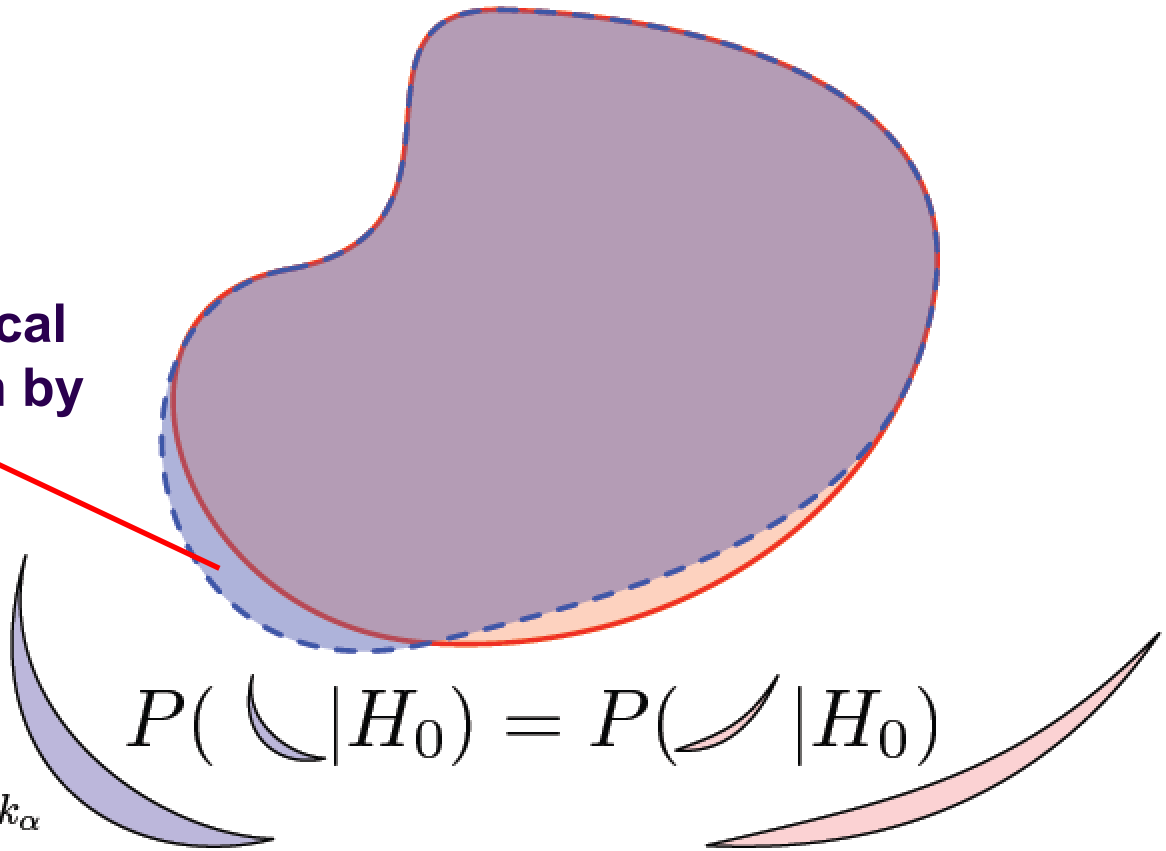
… as size (α) is fixed

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

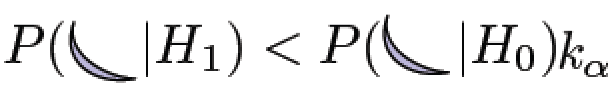
Neyman Pearson Lemma

Kyle Cranmer

outside "critical region" given by LL-ratio



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

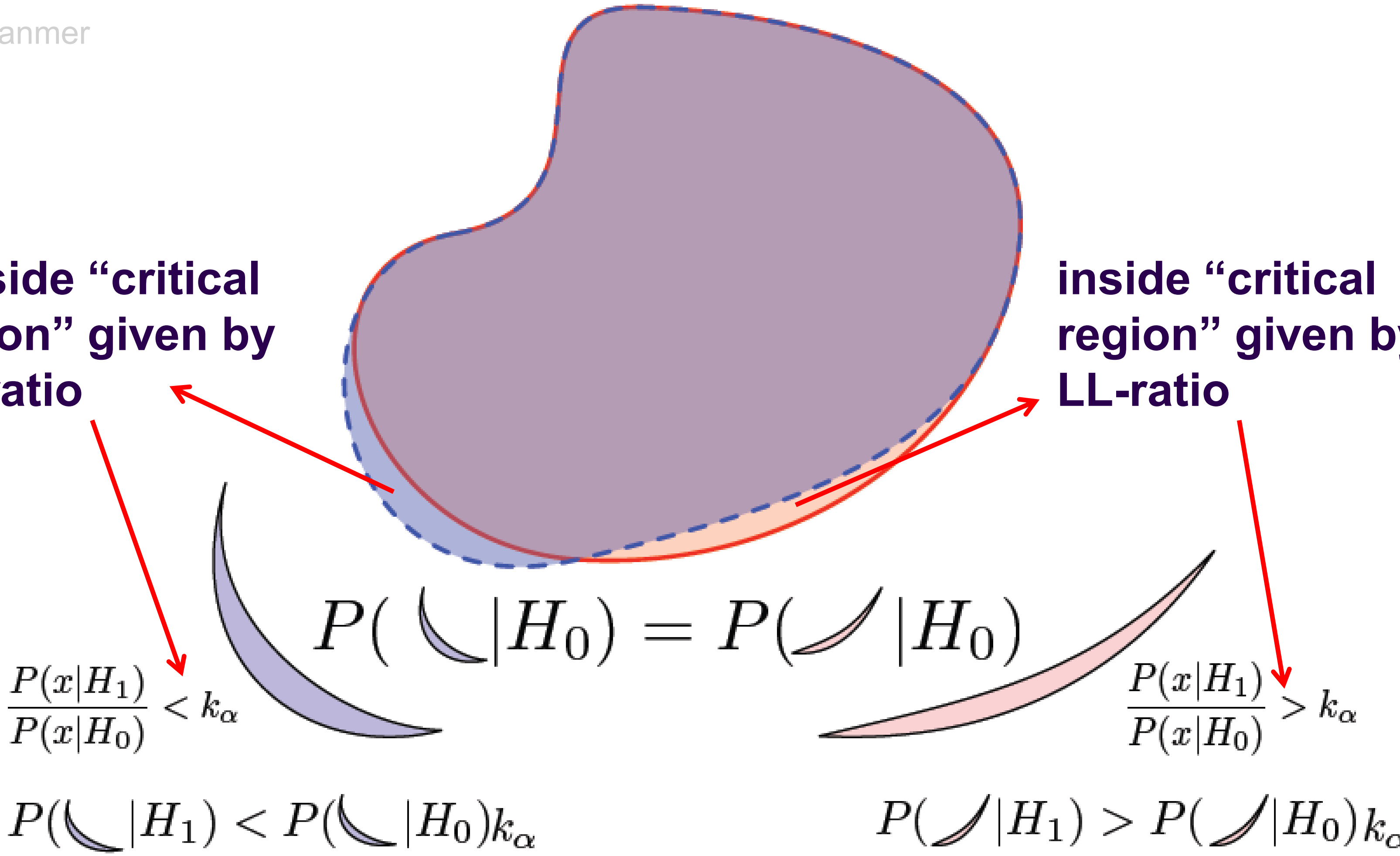


Neyman Pearson Lemma

Kyle Cranmer

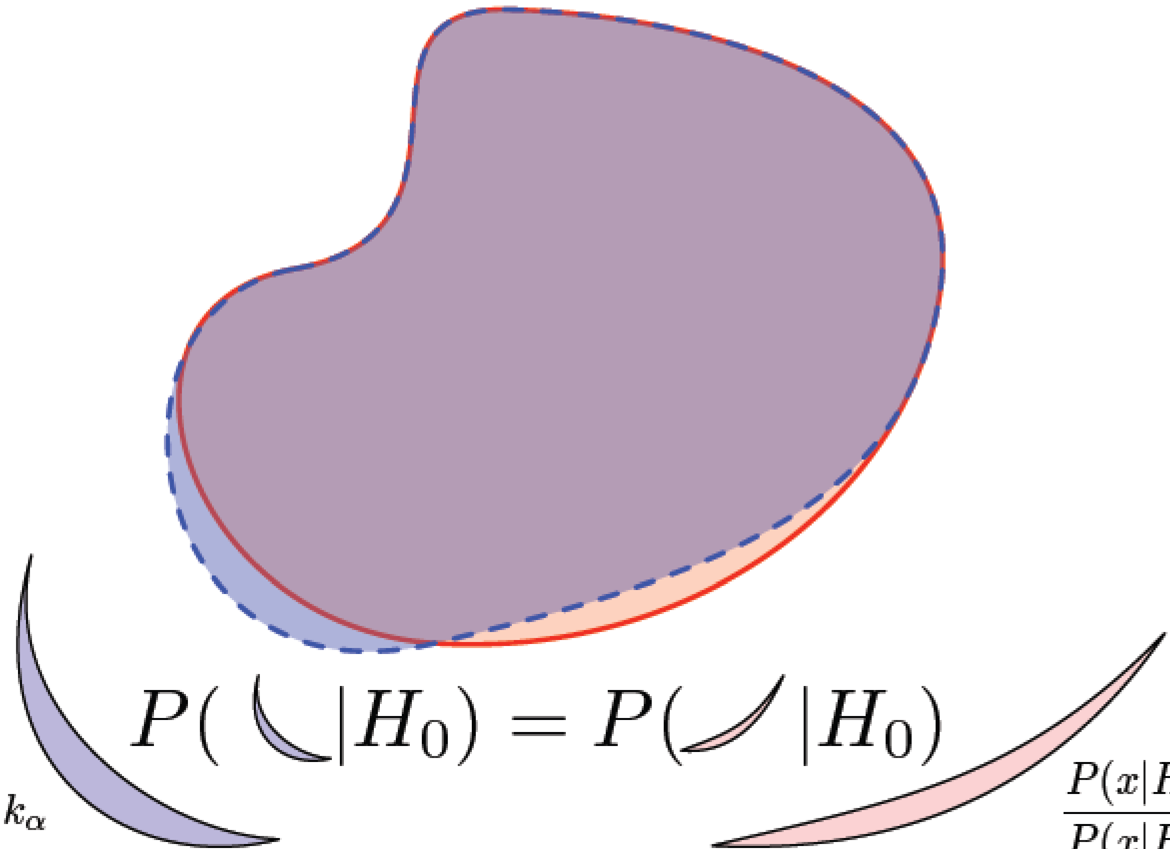
outside "critical region" given by LL-ratio

inside "critical region" given by LL-ratio



Neyman Pearson Lemma

Kyle Cranmer



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

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

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

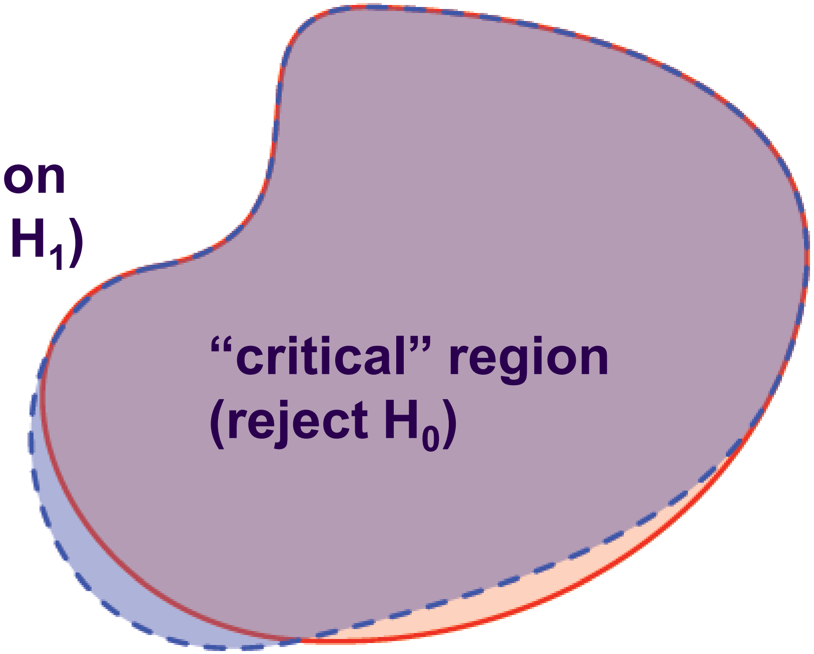
$$\boxed{P(\;\;|H_1) < P(\;\;|H_1)}$$

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

Neyman Pearson Lemma

Kyle Cranmer

"acceptance" region
(accept H_0 (reject H_1)



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

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

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

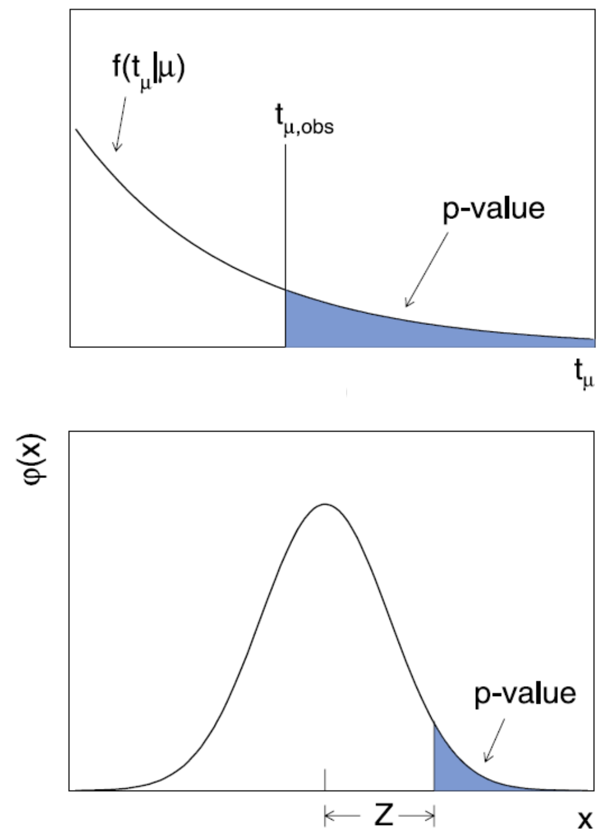
The NEW "acceptance" region has less power! (i.e. probability under H_1) q.e.d

Neyman Pearson Lemma

- **Unfortunatly:**
  - **Neyman Pearsons 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**

P-Value

- typical test setup: specify size α (i.e. the rate at which you tolerate "false discoveries")
- do the measurement an observe t_{obs} in your test statistic
- p-value: Probability to find data (a result) at least as much in disagreement with the hypothesis as the observed one



Note:
- p-value is property of the actual measurement
- p-value is NOT a measure of how probably the hypothesis is

translate to the common "sigma"
→ how many standard deviations "Z" for same p-value on one sided Gaussian

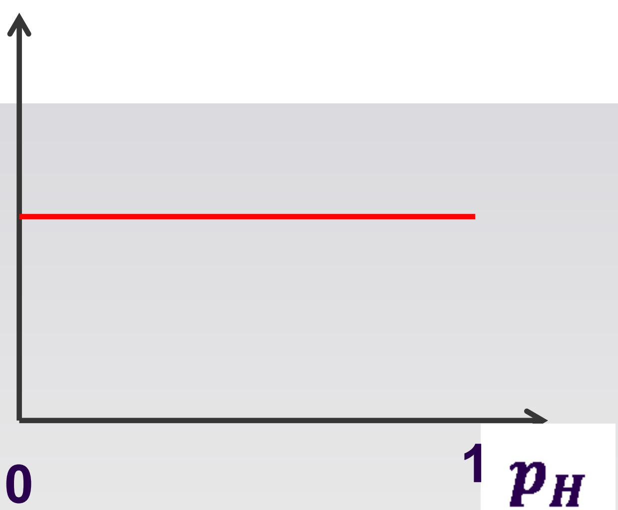
→ 5σ = p-value of $2.87 \cdot 10^{-7}$

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})
- $f(t|H)$: distribution of t (expected distribution of results that would be obtaind if we were to make many independent measuremnts/experiments)
- p-value : $p_H = \int_t^\infty f(t'|H)dt'$ (for each hypothetical measurement)
  - Type equation here.

→ p-values are "random variables" → distribution



- 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

Summary

- **Hypothesis testing**
  - **for simple hypothesis:**
    - Likleihood Ration as teststatistic gives the most powerful test
      - for each given $size = \alpha$ (probability of "false discoveries", accidnetally rejecting H_0 ($background\ only$) hypothesis
      - the largest power power= $1 - \beta$ (chance to actually SEE the effect given it's there)
  - **concept of confidence level / p-value**
    - specify you test to "reject" if outcome falls in "rejection/critical" region (typically some 95% CL or higher for discoveries)
    - p-value (probability to observes s.th. even less in agreement with the (null) hypothesis as my measurement