Statistical Methods in Particle Physics

6. Hypothesis Testing

Prof. Dr. Klaus Reygers (lectures)
Dr. Sebastian Neubert (tutorials)

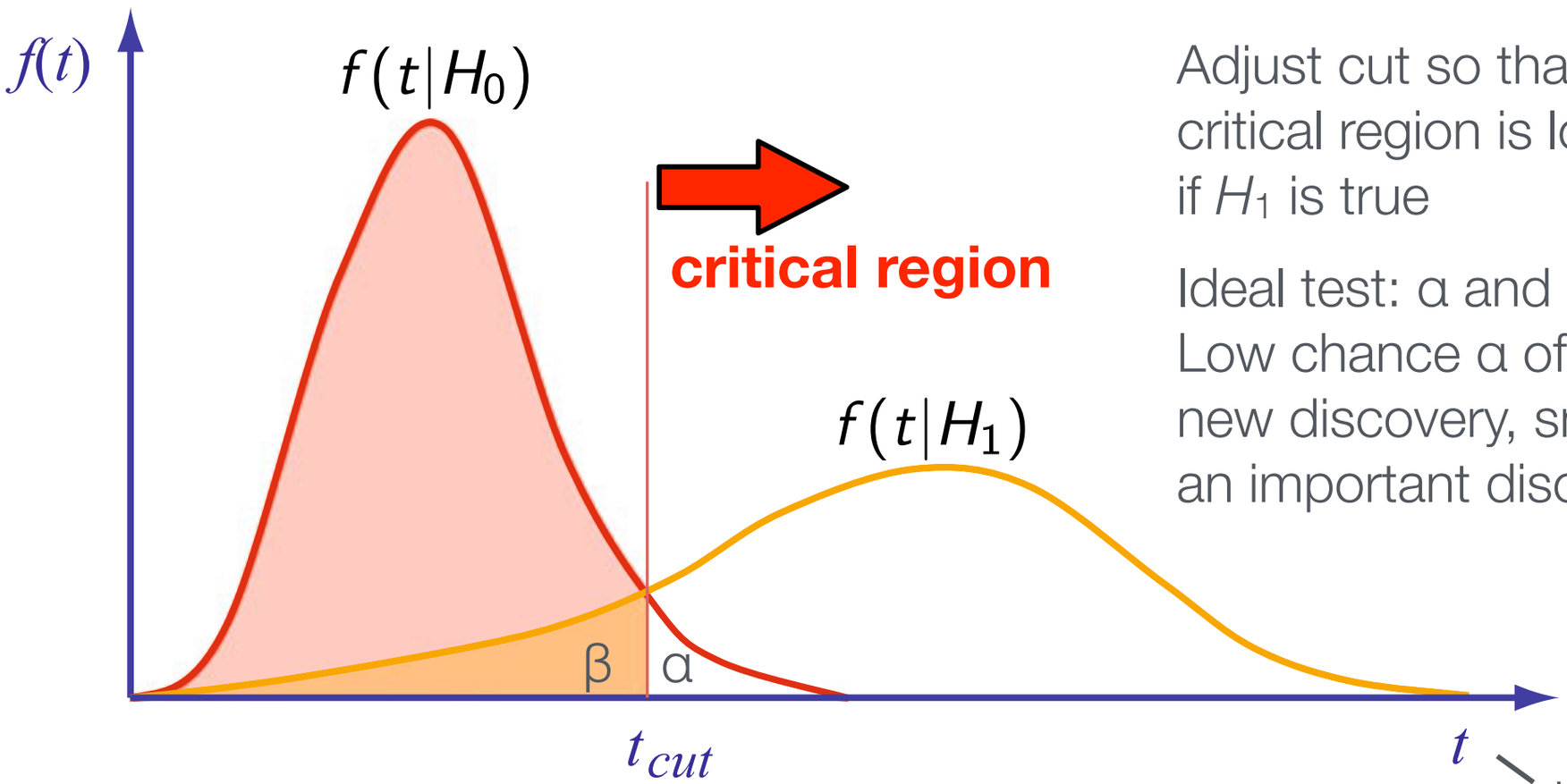
Heidelberg University
WS 2017/18

Hypotheses and Tests

- Hypothesis test
  - Goal: Draw conclusions from the data
  - Statement about the validity of a model
  - Tells you which of two competing models is more consistent with the data
- Simple hypothesis: a hypothesis with no free parameters
  - Examples: the detected particle is a pion; data follow Poissonian with mean 5
- Composite hypothesis: contains unspecified parameter(s)
  - Example: data follow Poissonian with mean > 5
- Null hypothesis H_0 and alternative hypothesis H_1
  - H_0 often the *background-only hypothesis* (e.g. the Standard Model in searches for new physics)
  - H_1 often *signal* or *signal + background hypothesis*
- Question: Can null hypothesis be rejected by the data?
- Test statistic t: a (usually scalar) variable which is a function of the data alone that can be used to test hypotheses
  - Example: $t = X^2_{\text{min}}$ of a least-squares fit

Critical region (I)

Reject null hypothesis if value of t lies in critical region ($t > t_{cut}$)



Adjust cut so that probability to be in critical region is low if H_0 is true and high if H_1 is true

Ideal test: α and β small:
Low chance α of incorrectly claiming a new discovery, small chance β of missing an important discovery

The probability for H_0 to be rejected while H_0 is true:

$$\int_{t_{cut}}^{\infty} f(t|H_0)\,\mathrm{d}t = \alpha$$

α:
"size" or "significance level" of the test

Probability to reject H_1 even though it is true:

$$\int_{-\infty}^{t_{cut}} f(t|H_1)\,\mathrm{d}t = \beta$$

1– β:
"power of the test"

Critical Region (II)

In case of multi-variate data one can also define a critical region *W* directly in data space *S* without using test statistic.

We can then write:

Under H_0, probability to find data in critical region, i.e., to reject H_0:

$$p(\vec{x} \in W | H_0) \leq \alpha$$

"≤" means, that it is not always possible to find critical region that gives exactly significance level α

Under H_1, probability to find data outside critical region, i.e., to reject H_1:

$$p(\vec{x} \notin W | H_1) = \beta$$

In case of H_0 = background, H_1 = signal, think in terms of efficiencies:

$\epsilon_B \equiv \alpha$ "background efficiency", i.e., prob. to misclassify bckg. as signal

$\epsilon_S \equiv 1 - \beta$ "signal efficiency"

Type I and Type II Errors

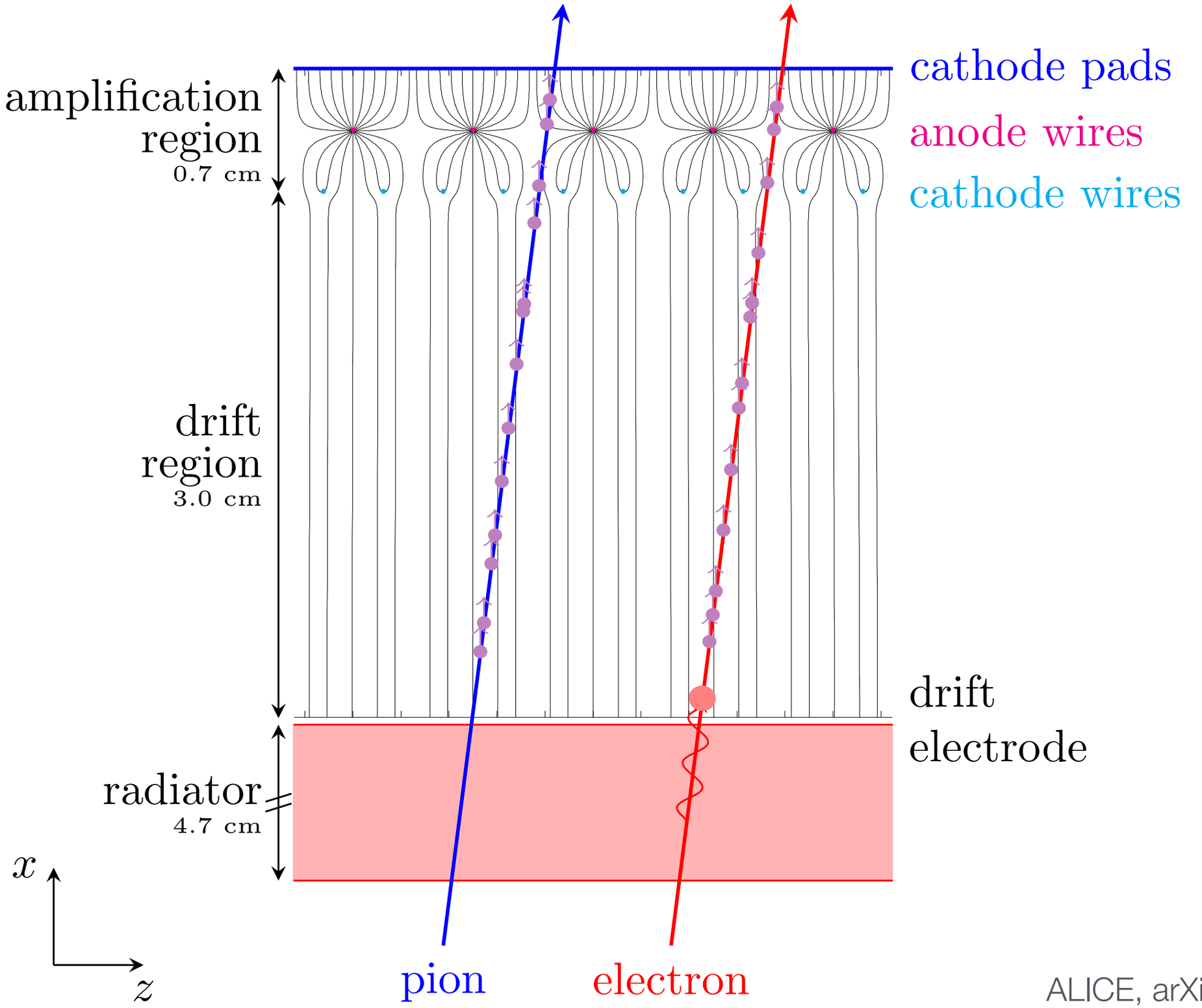
Type I error:
Null hypothesis is rejected while it is actually true

Type II error:
Test fails to reject null hypothesis while it is actually false

Type I and type II errors and their probabilities:

	H_0 is true	H_0 is false (i.e., H_1 is true)
H_0 is rejected	Type I error (α)	Correct decision ($1-\beta$)
H_0 is not rejected	Correct decision ($1-\alpha$)	Type II error (β)

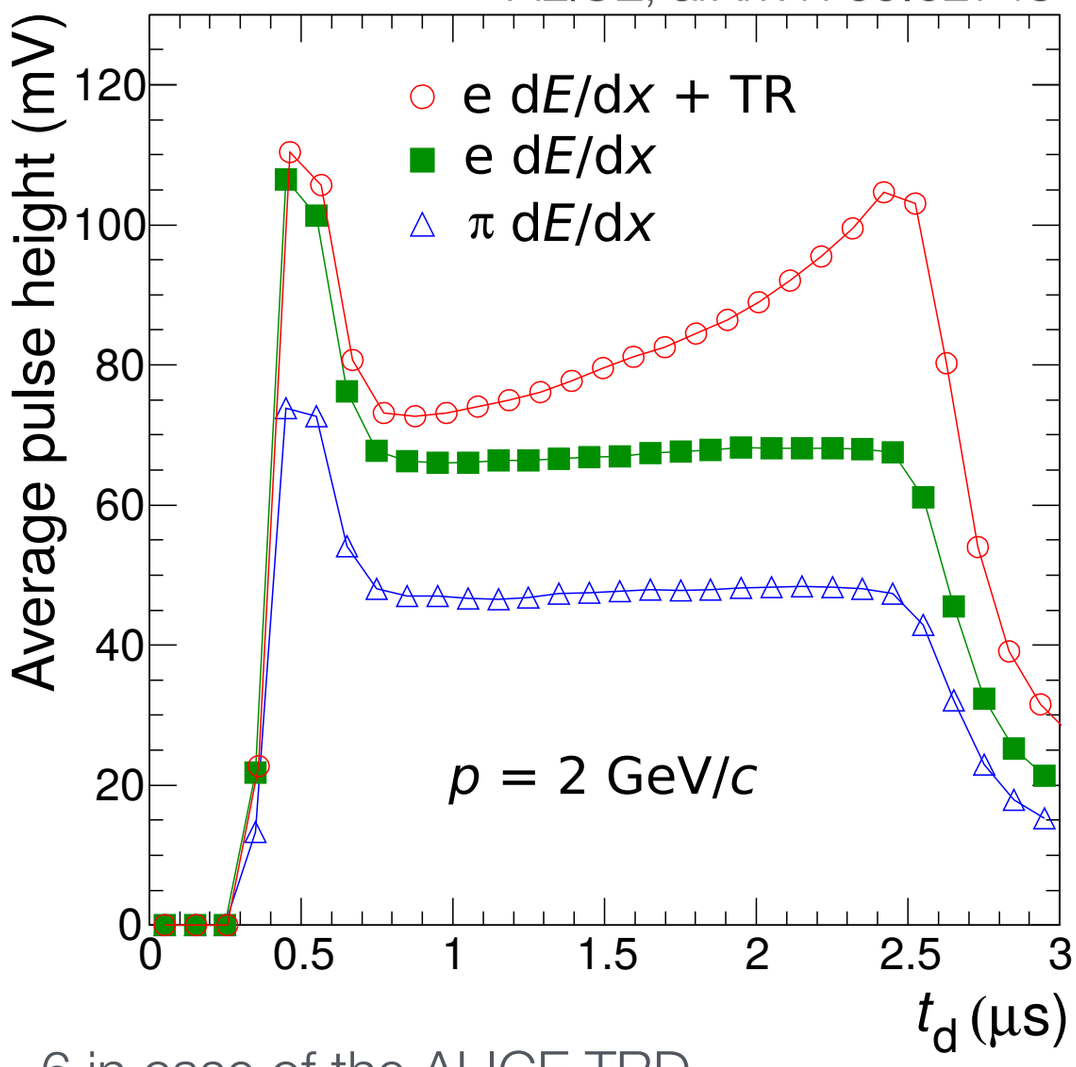
Example: Electron ID with the ALICE TRD (I)



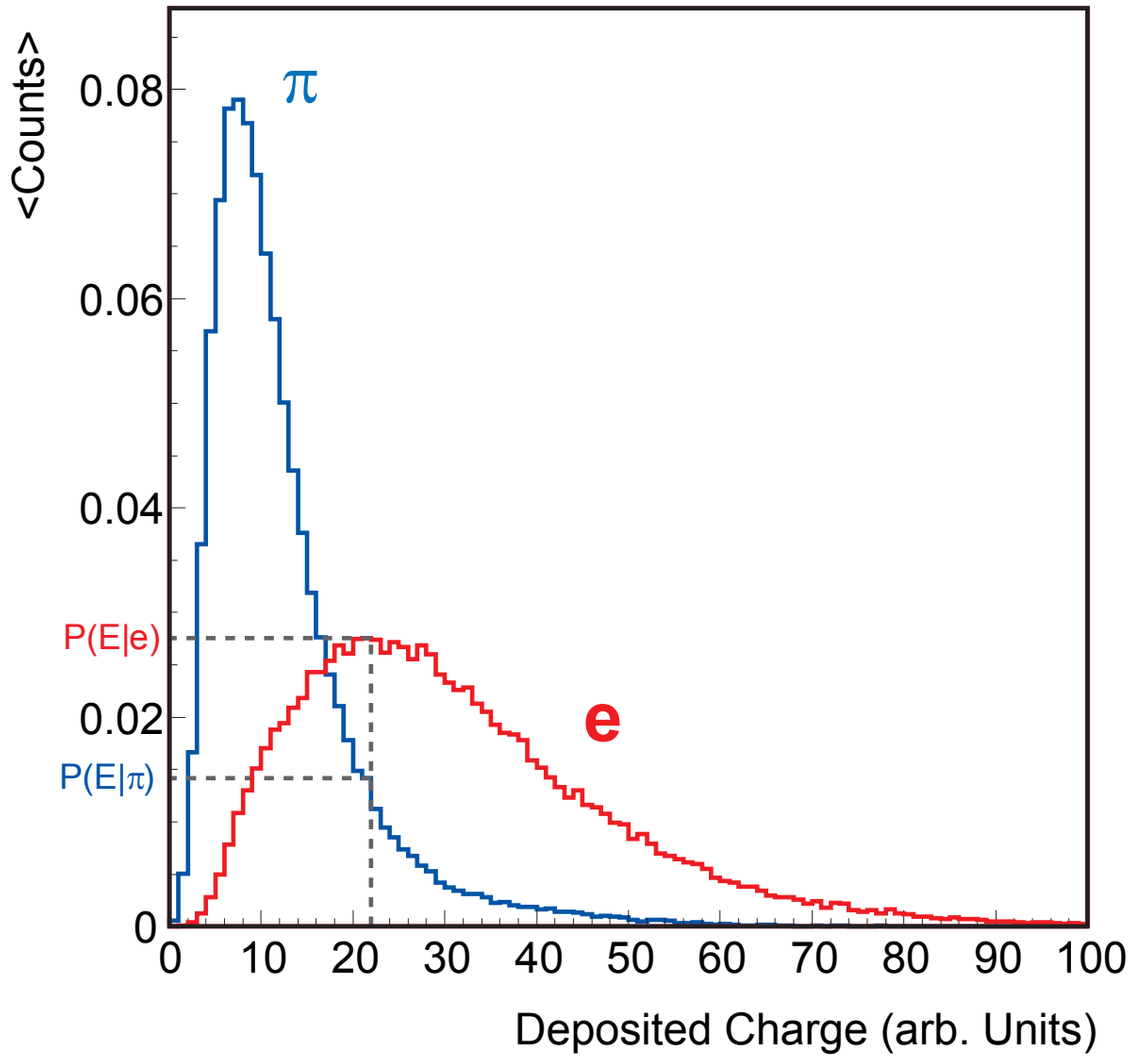
ALICE, arXiv:1709.02743

Example: Electron ID with the ALICE TRD (II)

ALICE, arXiv:1709.02743



Doctoral thesis A. Wilk:
https://inspirehep.net/record/1231193/



6 in case of the ALICE TRD

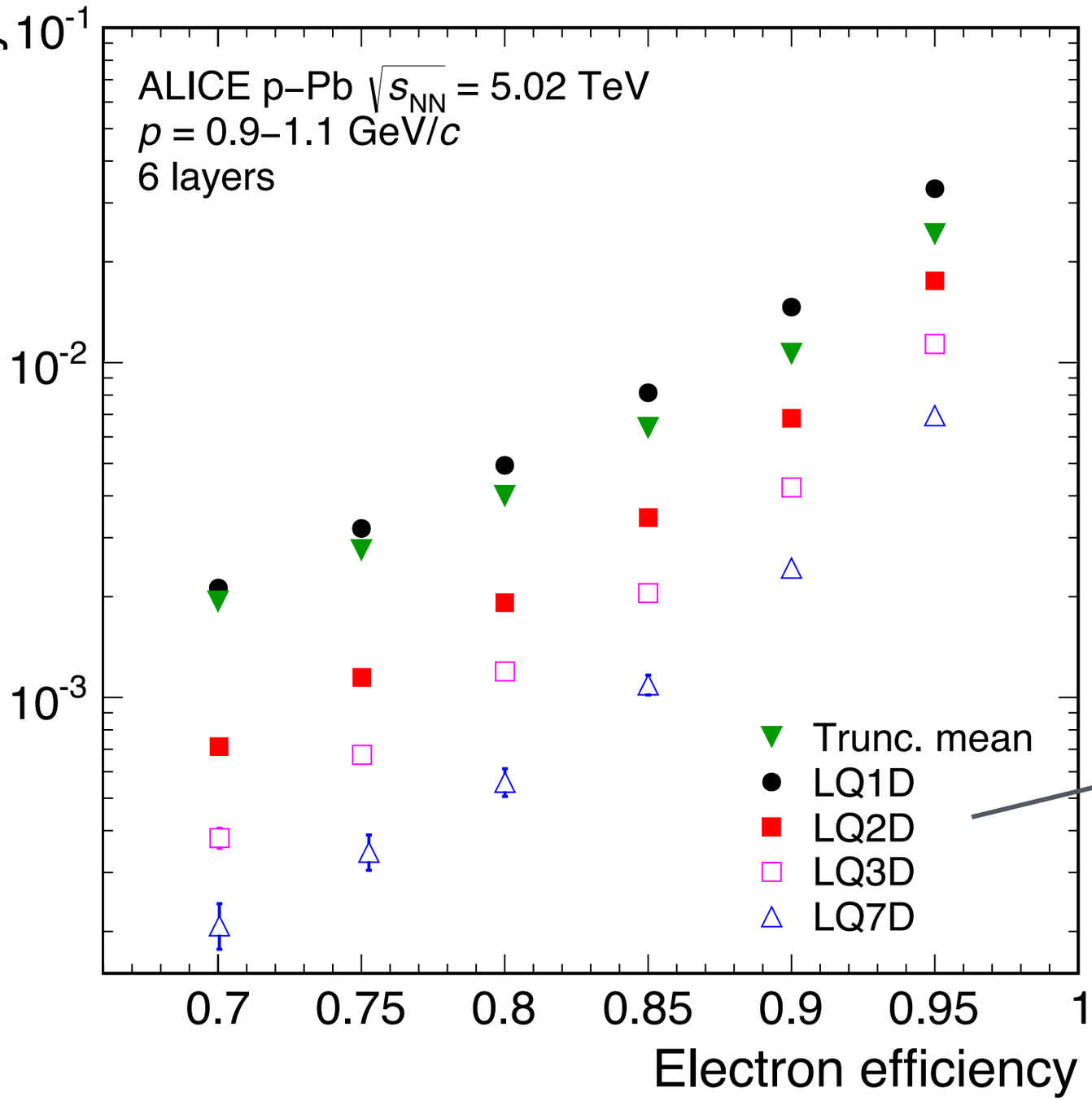
$$P_e = \prod_{i=1}^{n_{\text{chambers}}} P(E_i|e), \quad P_\pi = \prod_{i=1}^{n_{\text{chambers}}} P(E_i|\pi), \quad \text{test statistic } t = \frac{P_e}{P_e + P_\pi}$$

likelihoods can be multiplied here
(independent information)

high values (close to unity)
indicate high prob. for an elentron

Example: Electron ID with the ALICE TRD (III)

ALICE, arXiv:1709.02743



$\epsilon_B \equiv \alpha$

"background efficiency"

$\epsilon_S \equiv 1 - \beta$ "signal efficiency"

Design goal of the ALICE TRD:

pion efficiency < 1% at 90% electron efficiency for 6 chambers

LQ*N*D:
Signal in each chamber divided into *N* time slices

Neyman–Pearson Lemma

Neyman-Pearson lemma holds for simple hypotheses and states:

To get the highest power (i.e. smallest possible value of β) of a test of H_0 with respect to the alternative H_1 for a given significance level, the critical region W should be chosen such that:

$$t(\vec{x}) := \frac{f(\vec{x}|H_1)}{f(\vec{x}|H_0)} > c \quad \text{inside } W \qquad \text{and} \qquad t(\vec{x}) \leq c \quad \text{outside } W$$

c is a constant chosen to give a test of the desired significance level.

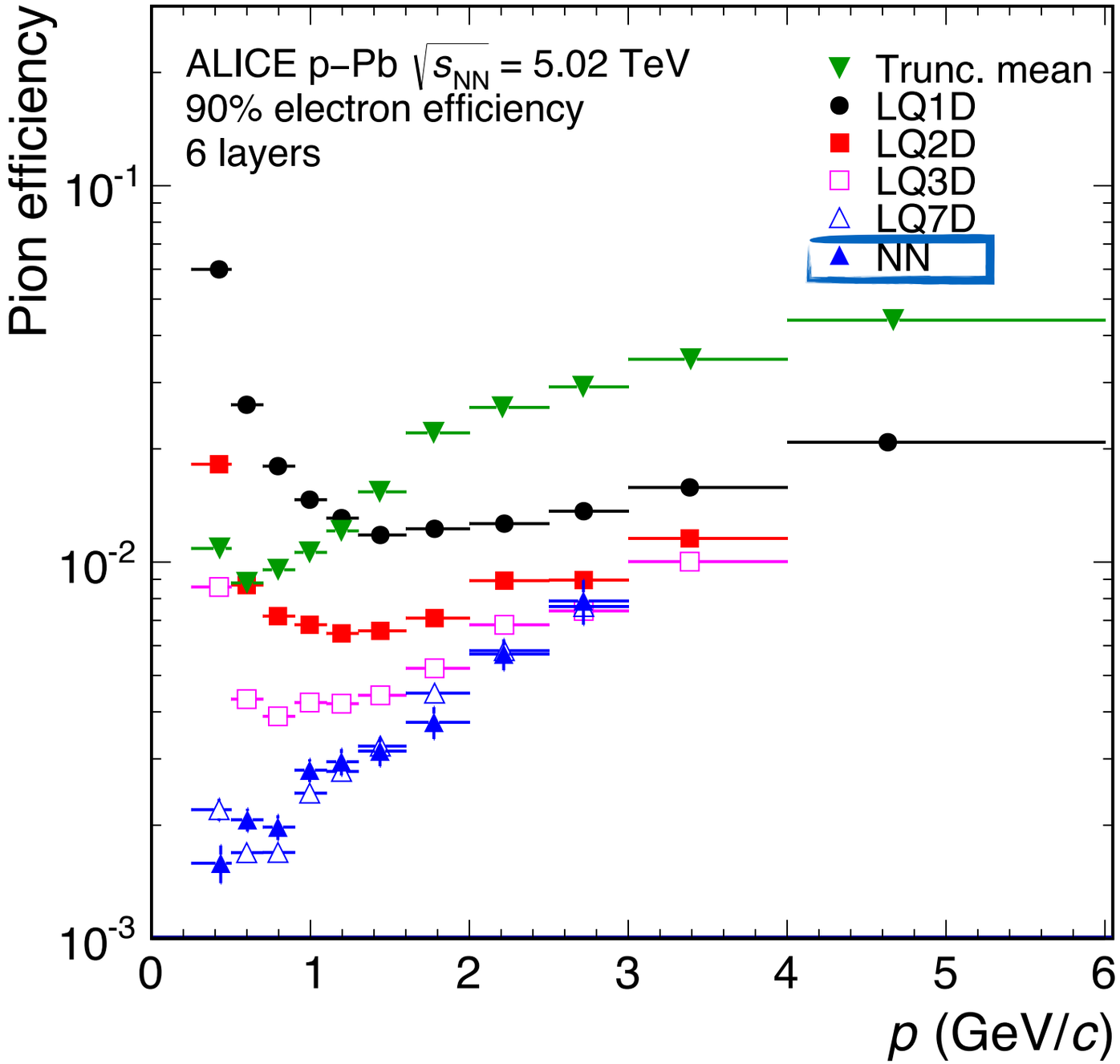
Equivalent formulation: optimal scalar test statistic is the likelihood ratio

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

Practical Considerations

- Problem: often one does not have explicit formulas for $f(x|H_0)$ and $f(x|H_1)$
- One rather has Monte Carlo models for signal and background processes which allow one to generate instances of the data
- In this case one can use multi-variate classifiers to separate different types of events
  - Fisher discriminants
  - Neural networks
  - Support vector machines
  - decision trees
  - …
- Software
  - TMVA (http://tmva.sourceforge.net/#exec_summary, https://root.cern.ch/tmva)
  - …

Example:
Neural Network for e/π Separation with the TRD



Test of Significance (Goodness-of-Fit)

- Often one wants to quantify the level of agreement between the data and a hypothesis without explicit reference to alternative hypotheses
- Define test statistic t that reflects level of agreement with the data
- Determine distribution f($t|H_0$) under hypothesis H_0
- p-value (in case large values of t indicate poor agreement with H_0)

$$p\text{-value} = \int_{t_{\text{obs}}}^{\infty} f(t|H_0)\,\mathrm{d}t$$

- p-value should not be confused with significance level
  - significance level is a pre-specified constant
  - p-value is a function of the data, and is therefore itself a random variable
- p-value is not the probability for the hypothesis; in frequentist statistics, this is not defined

Simple Example:
Counting Experiment (Poisson Statistics)

Expected background events:
$\nu_b = 1.3$

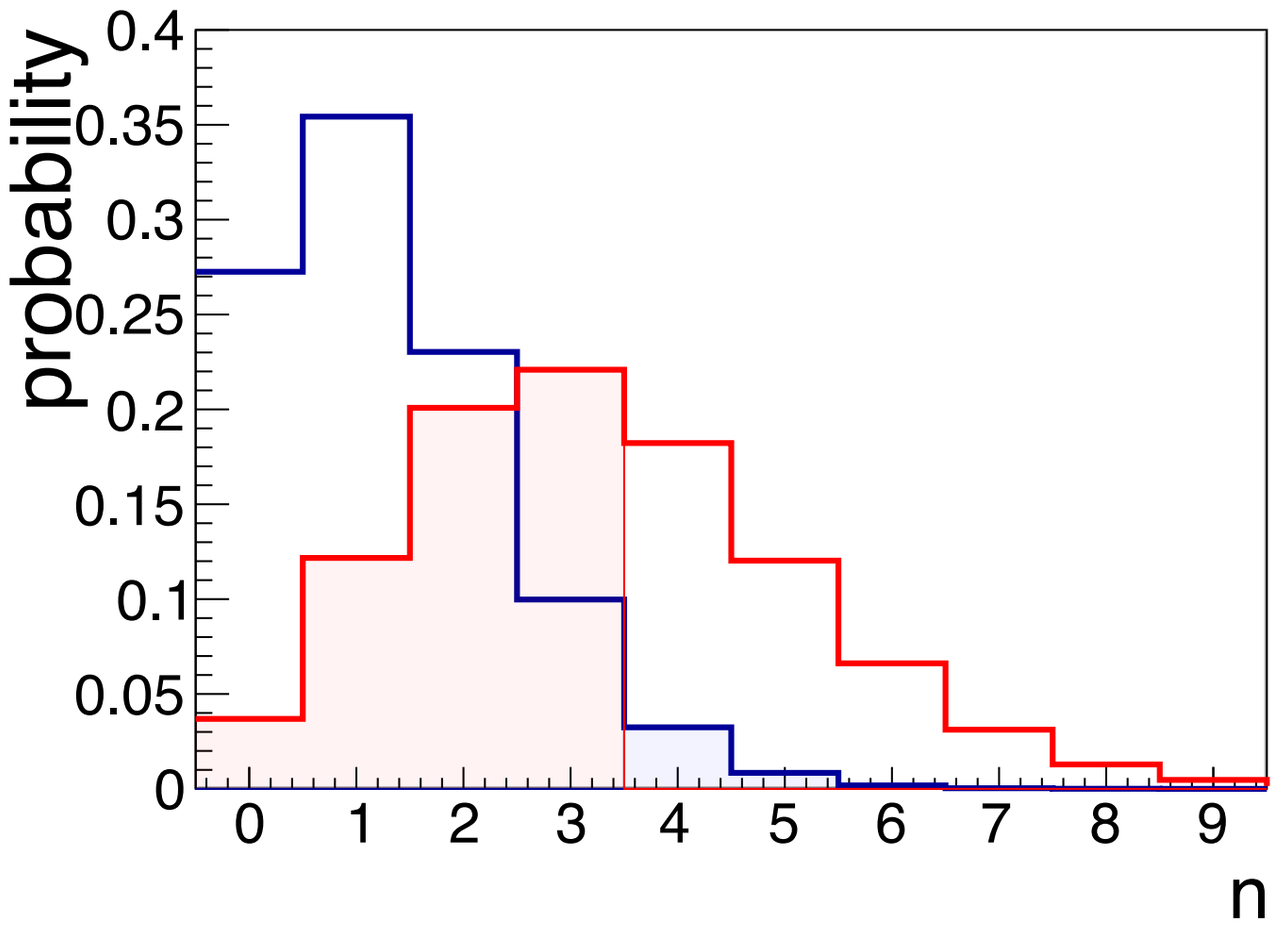
Expected signal events:
$\nu_s = 2$

Expected signal + bckgr. events:
$\nu_{s+b} = 3.3$

Test statistic t =
number of observed events

Critical region $t_c \geqq 4$

- significance of the test $\alpha = 0.043$
- power of the test $1 - \beta = 0.42$

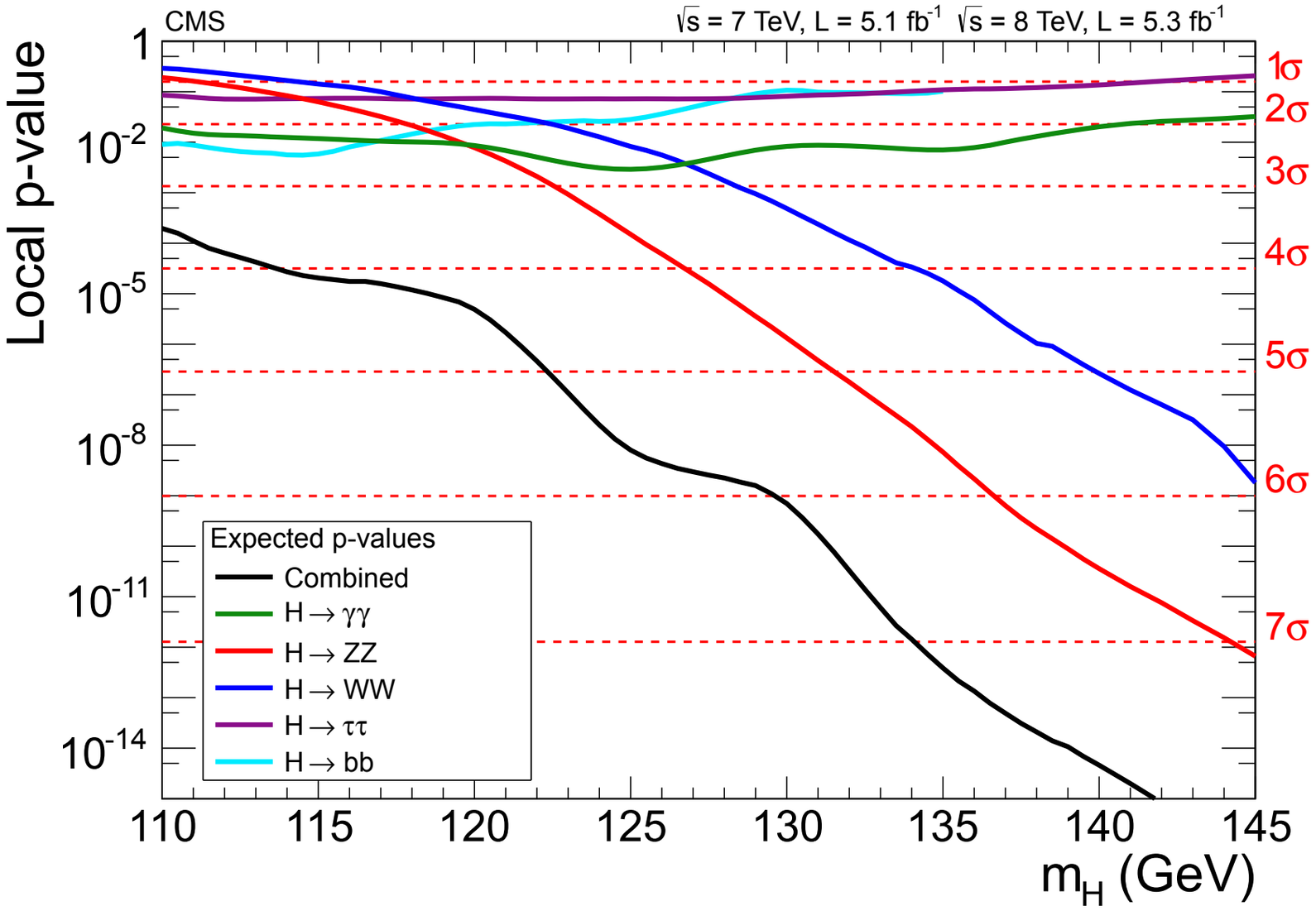


H_0: only background,
H_1: signal + background

Suppose we observe $n = 5$ events

- Under H_0, this correspond to a
  ***p*-value = 0.01**

Example: Higgs Measurement
Expected local *p*-values



p-values translated to number of standard deviations using the one-sided Gaussian tail convention:

$$p\text{-value} = \int_Z^\infty N(x;0,1)\,\mathrm{d}x$$

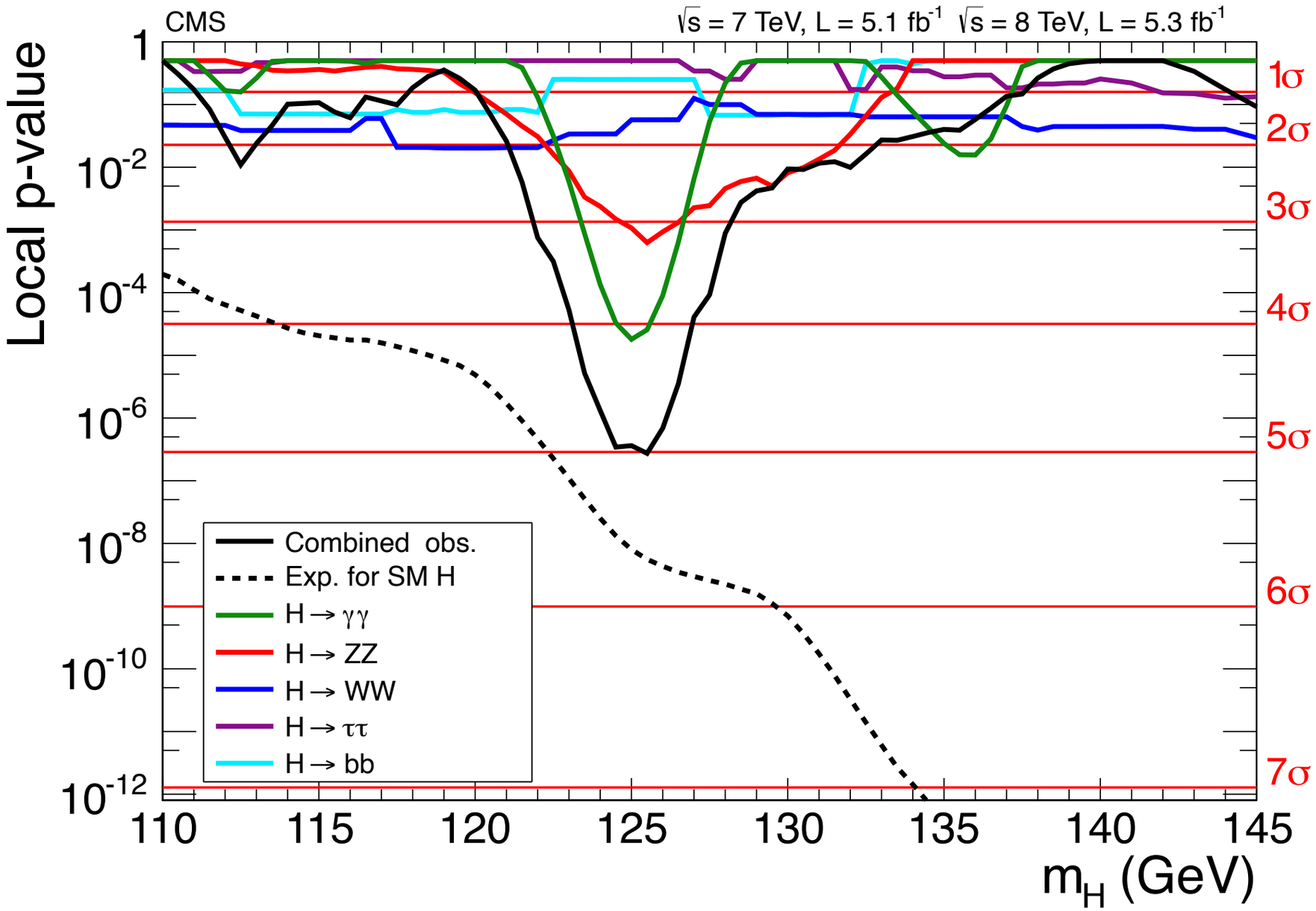
$$\rightsquigarrow Z = \phi^{-1}(1 - p\text{-value})$$

inverse of the CDF of the standard Gaussian distribution

For each assumed Higgs mass (→ local *p*-value)

- Calculate expected signal for Standard Model Higgs boson
- Determine *p*-value for H_0 that only SM background processes contribute

Example: Higgs Measurement
Observed local *p*-values



"An excess of events is observed above the expected background, with a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations."

Look-Elsewhere Effect

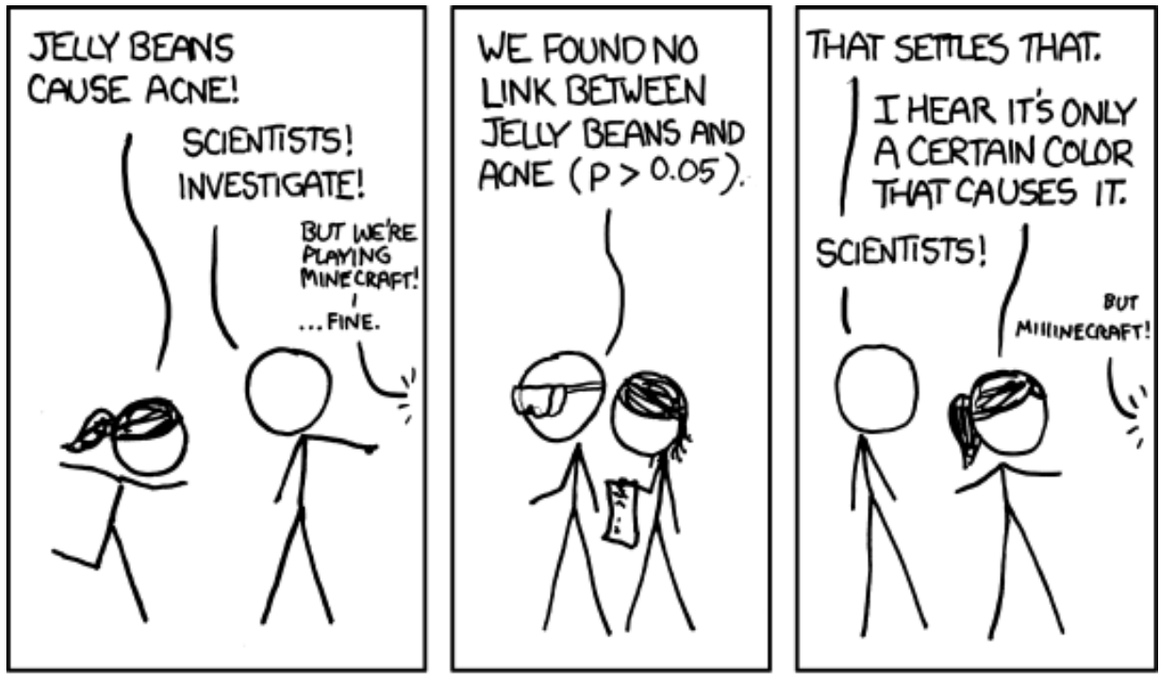
https://en.wikipedia.org/wiki/Look-elsewhere_effect

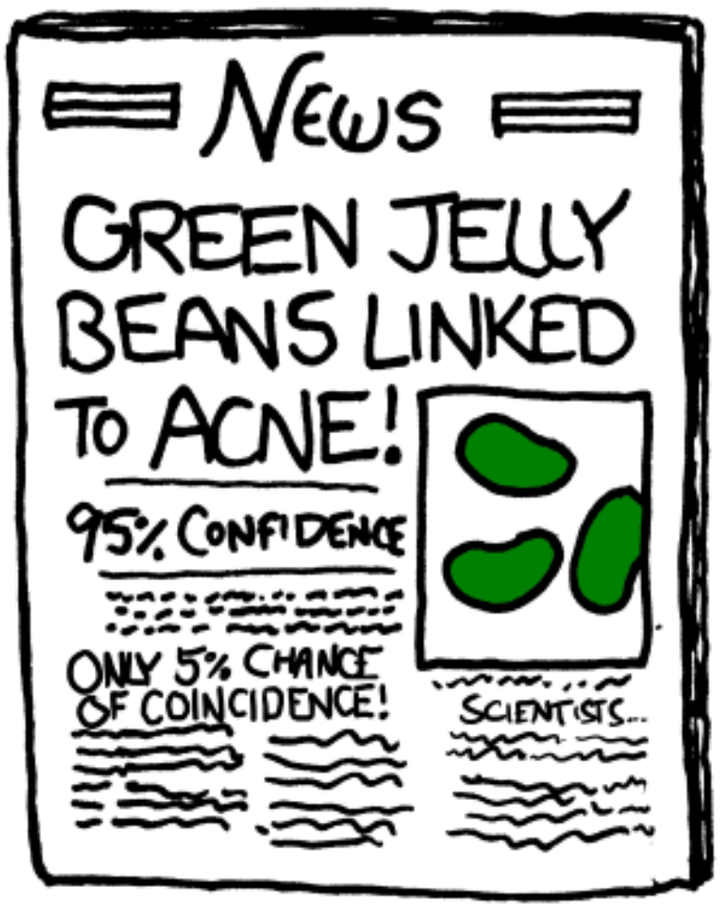
- CMS Higgs paper
  - The probability for a background fluctuation to be at least as large as the observed maximum excess is termed the local *p*-value, and that for an excess anywhere in a specified mass range the global *p*-value.
  - Local *p*-value corresponds to 5σ
  - Global *p*-value for mass range 110–145 GeV corresponds to 4.5σ
- In general:
  - If one is performing multiple tests then obviously a *p*-value of 1/*n* is likely to occur after *n* tests
  - Solution: "trials penalty" or "trials factors", i.e. make threshold a function of *n* (more stringent threshold for larger *n*)

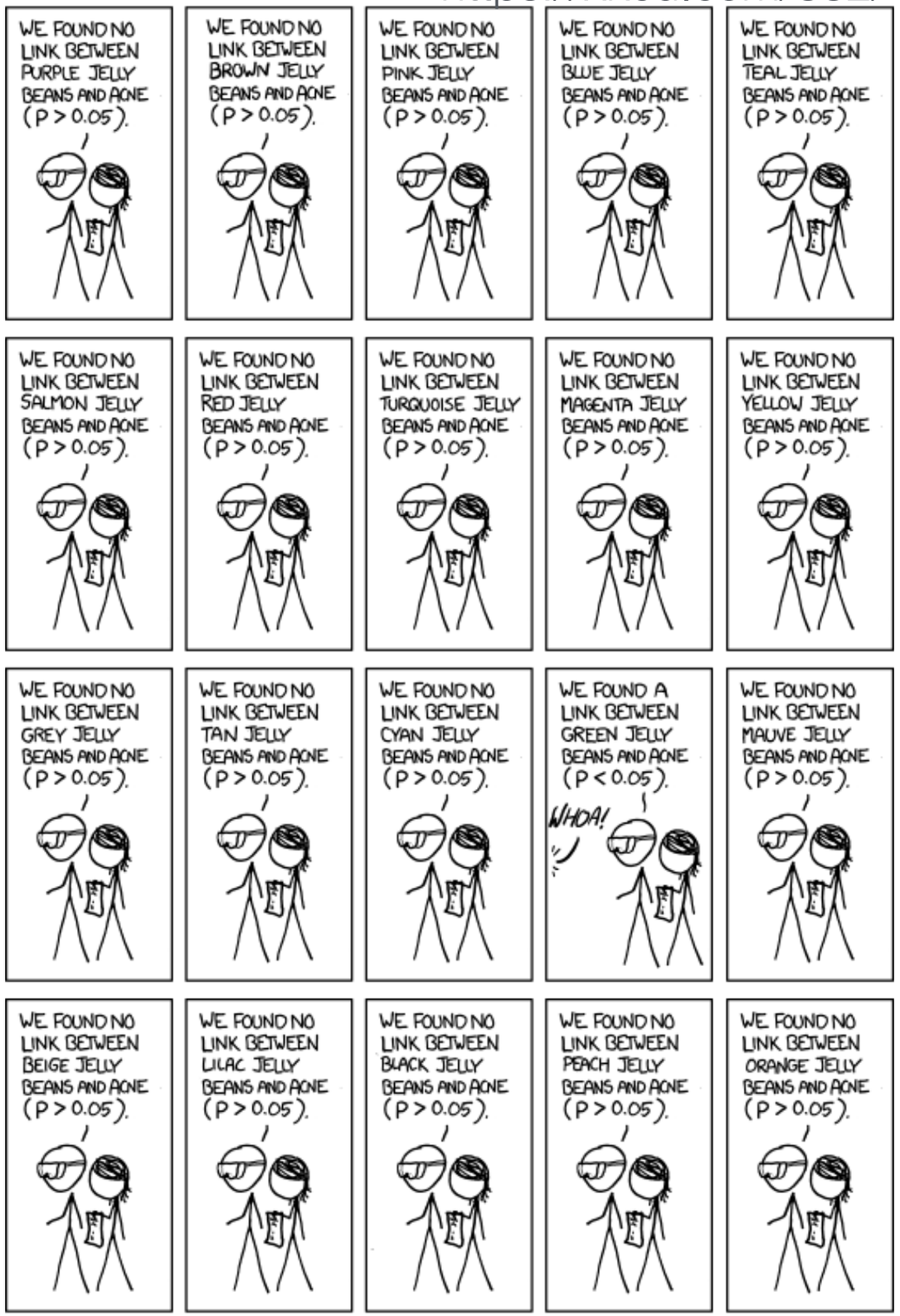
A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 ailments. The study found that the incidence of childhood leukemia was four times higher among those that lived closest to the power lines, and it spurred calls to action by the Swedish government. The problem with the conclusion, however, was that they failed to compensate for the look-elsewhere effect; in any collection of 800 random samples, it is likely that at least one will be at least 3 standard deviations above the expected value, by chance alone. Subsequent studies failed to show any links between power lines and childhood leukemia, neither in causation nor even in correlation.

p-value Hacking

https://xkcd.com/882/







Digression: *p*-value Debate

- Null hypothesis ("no effect") rejected and results deemed statistically significant if p-value < 0.05
- Relatively weak statistical standard, but often not realized as such
- Chance for false positive outcome 1/20
  - Might results in too many false positive results in the literature
  - Social and biomedical sciences in the focus of the discussion
- Problem exacerbated by *p*-value hacking
  - Data gathered by researches without first creating a hypothesis
  - Search for patterns in the data that can be reported as statistically significant
- Probably contributes to reproducibility crisis in science
  - Results of many scientific studies are difficult or impossible to replicate on subsequent investigation
- Proposed solution: lower threshold to p-value < 0.005
  - https://psyarxiv.com/mky9j (published in Nature Human Behavior, https://www.nature.com/articles/s41562-017-0189-z)

https://www.nature.com/news/big-names-in-statistics-want-to-shake-up-much-maligned-p-value-1.22375

Why 5σ for Discovery in Particle Physics?

5σ ⇔ p-value = 2.87×10^{-7} (one-tailed test)

- History: There are many cases of 3σ and 4σ effects that have disappeared with more data
- The Look-Elsewhere Effect
- Systematics:
  - Usually more difficult to estimate than statistical uncertainties
  - "Safety margin"
- Subconscious Bayes factor:
  - Physicists subconsciously tend to assess the Bayesian probabilities $p(H_0|\text{data})$ and $p(H_1|\text{data})$
  - If H_1 involves something very unexpected (e.g., neutrinos travel faster than the speed of light) then prior probability for null hypothesis H_0 is much larger than for H_1.
  - "Extraordinary claims require extraordinary evidence"

Last point ⇒ unreasonable to have a single criterion (5σ) for all experiments

Louis Lyons, Statistical Issues in Searches for New Physics, arXiv:1409.1903

Kolmogorov–Smirnov Test (I)

KS test is an unbinned goodness-of-fit test

Compare cumulative distribution function

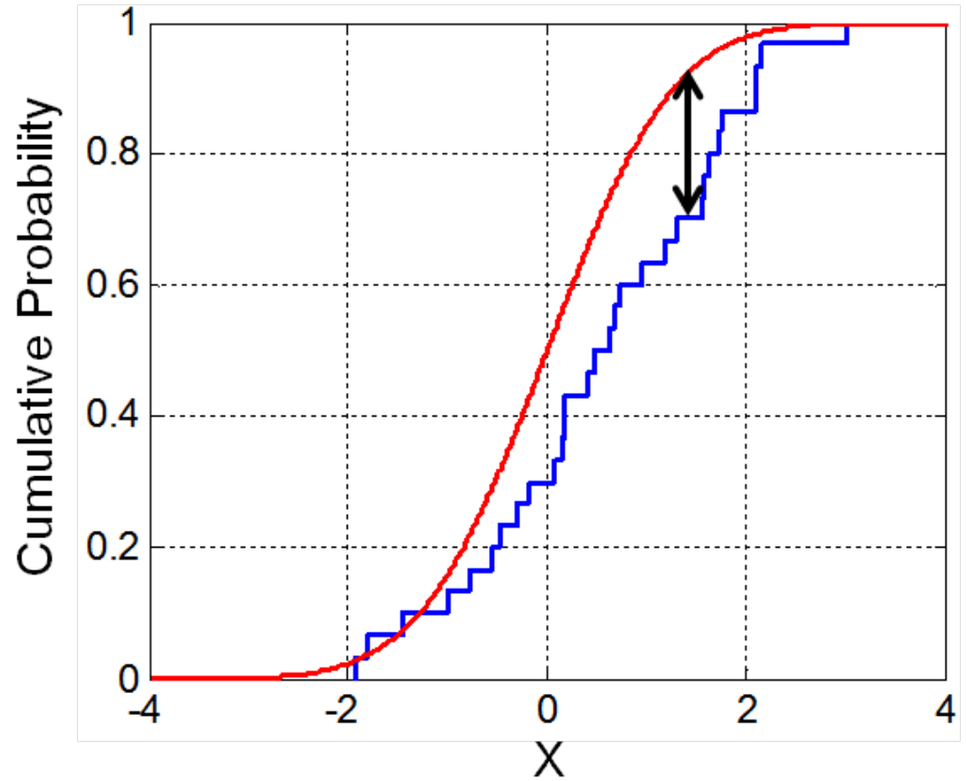
$$F(x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'$$

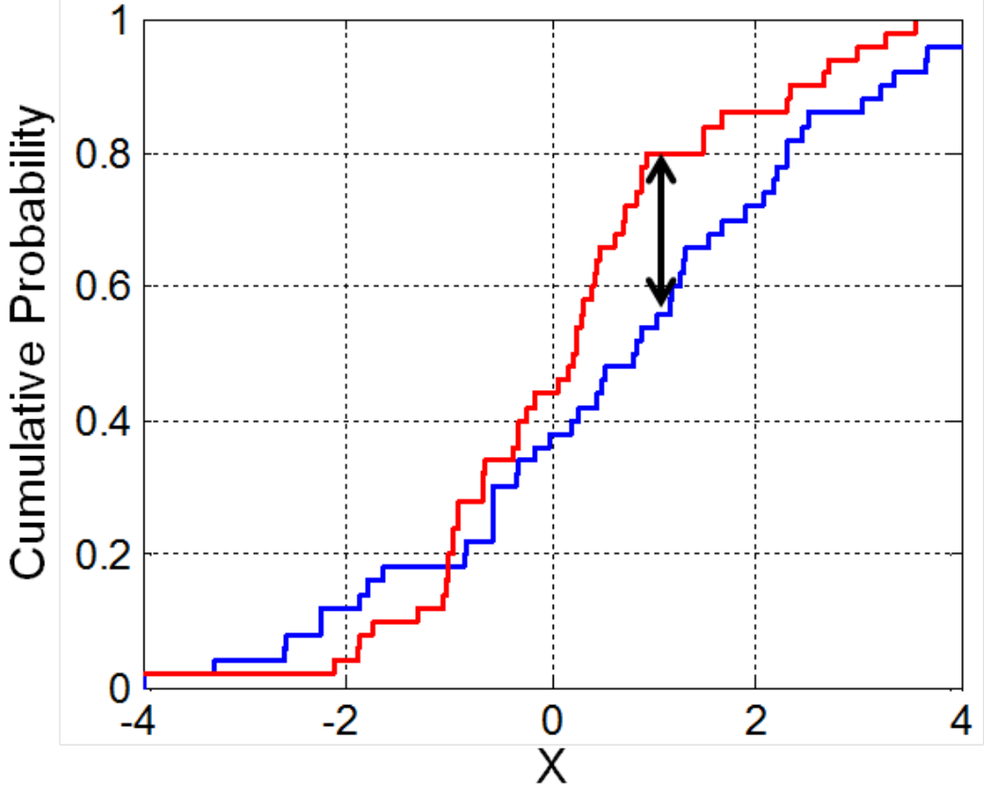
with the so-called Empirical Distribution Function (EDF)

$$S(x) = \frac{\text{number of observations with } x_i < x}{\text{total number of observations}}$$

The test statistic is the maximum difference between the two functions:

$$D = \sup|F(x) - S(x)|$$



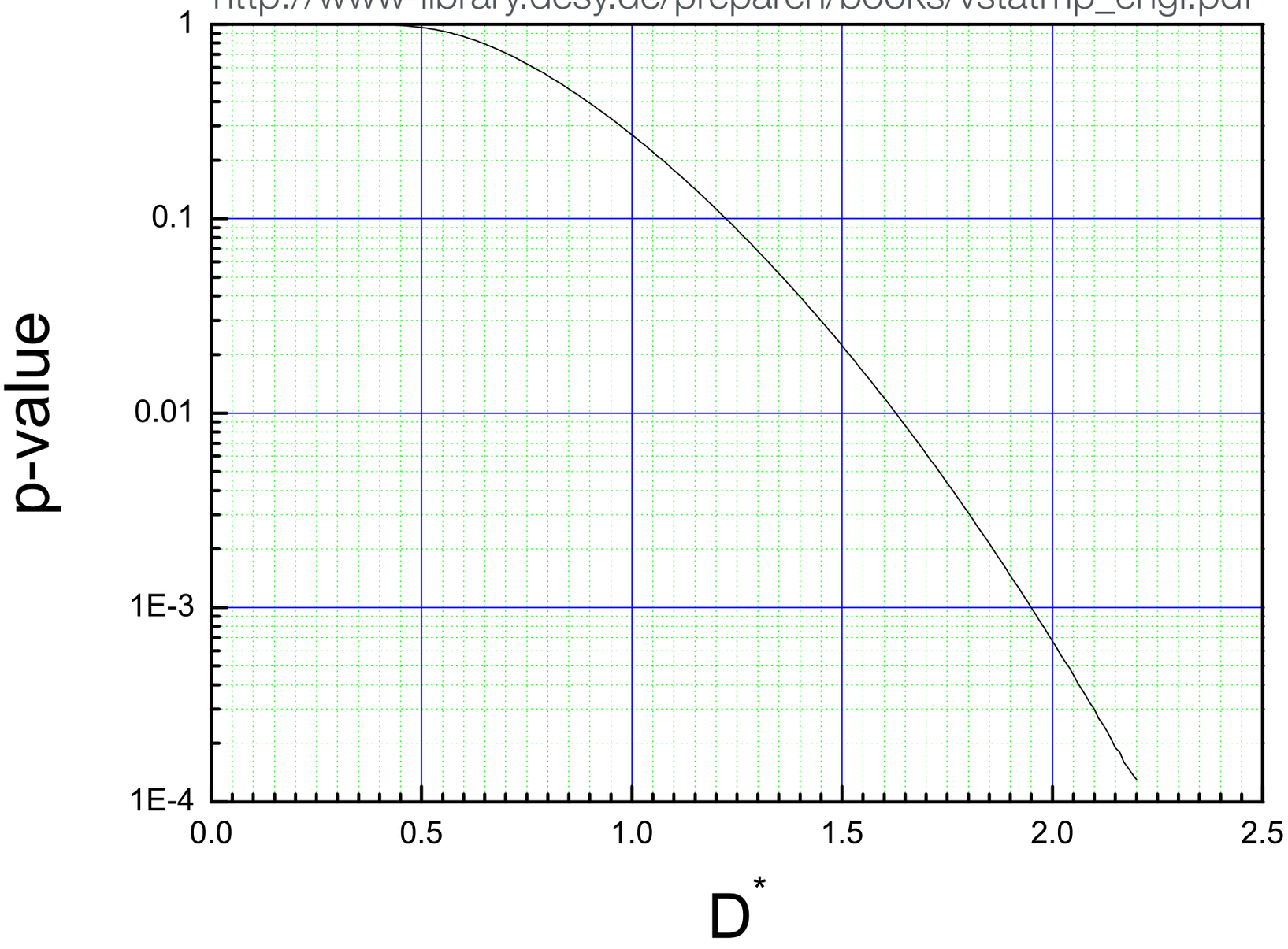


One can also test whether two one-dimensional sets of points are compatible with coming from the same parent distribution: `TMath::KolmogorovTest`

Kolmogorov–Smirnov Test (II)

Expected distribution of D known for given N → p-value

Bohm, Zech,
http://www-library.desy.de/preparch/books/vstatmp_engl.pdf



$$D^* = \sqrt{N} D, \quad N = \text{number of data points}$$

Two-Sample χ^2 Test

Test hypothesis that two binned data sets come from the same underlying distribution.

Two histograms with k bins

Number of entries in bin i: n_i for measurement 1, m_i for measurement 2

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{\sigma_{n_i}^2 + \sigma_{m_i}^2}$$

Run Test (Wald–Wolfowitz Test)

Drawback of the χ^2 test: insensitive to the sign of the deviation

Def. "run":
region of consecutive bins where the data show deviations in the same direction

++++−−−+++−−+++++−−−− $N = N_+ + N_-$ = 22 bins, 3 "+" runs and 3 "–" runs

$$\mu = 1 + \frac{2\ N_+\ N_-}{N}, \qquad \sigma^2 = \frac{2\ N_+\ N_-\ (2\ N_+\ N_- - N)}{N^2\ (N-1)} = \frac{(\mu-1)(\mu-2)}{N-1}$$

μ: expected number of runs in a sequence of N elements (bins).

The run test is based on the null hypothesis that each element in the sequence is independently drawn from the same distribution (no assumption about prob. for "+" and "–")

For more than about 20 bins the Gaussian approximation holds and the significance of the deviation of an observed number r of runs from the expected value in units of the standard deviation is:

$$Z = \frac{r - \mu}{\sigma}$$

Run test is complementary to the χ^2 square test (can be done in addition)

Bayesian Hypothesis Testing

- In Bayesian language, all problems are hypothesis tests!
  - Posterior probability for a hypothesis $P(H|\text{data})$ or a parameter $P(\theta|\text{data})$

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

- Parameter estimation amounts to assigning a probability to the proposition that the parameter lies in the interval $[\theta_1, \theta_2]$
  - can reject hypothesis/parameter if posterior prob. is sufficiently small

- Example: LIGO PRL on detection of gravitational waves

In the source frame, the initial black hole masses are $36^{+5}_{-4} M_\odot$ and $29^{+4}_{-4} M_\odot$, and the final black hole mass is $62^{+4}_{-4} M_\odot$, with $3.0^{+0.5}_{-0.5} M_\odot c^2$ radiated in gravitational waves. All uncertainties define 90% credible intervals. These observations demonstrate the existence of binary stellar-mass black hole systems. This is the first direct detection of gravitational waves and the first observation of a binary black hole merger.

DOI: 10.1103/PhysRevLett.116.061102

- Requires one to explicitly specify alternative hypotheses:

$$P(D) = P(D|H_1) + P(D|H_2) + P(D|H_3) + \ldots$$

Often simply normalization from $\int P(H|D) = 1$

Bayes Factor (I)

Consider two hypotheses with prior probabilities

$$P(H_1), \quad P(H_2) = 1 - P(H_1)$$

We have

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1) \cdot P(H_1)}{P(D|H_2) \cdot P(H_2)}$$

The Bayes factor B_{12} is defined as

$$B_{12} = \frac{P(D|H_1)}{P(D|H_2)}$$

"posterior odds = Bayes factor × prior odds"

If $P(H_1) = P(H_2)$ the Bayes factor is equal to the ratio of posterior probabilities

Example: Search for neutrinoless double beta decay with GERDA

H_0 = no signal, H_1 = previously claimed signal (Heidelberg Moscow exp.)
$P(H_1) / P(H_0) = 0.024$ → previous claim of a 0νββ signal in ^{76}Ge strongly disfavored

arXiv:1307.4720

Bayes Factor (II): Interpretation

$2 \log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

R. Kass, E. Raftery, Bayes factors. J. Am. Stat. Assoc. 90, 773 (1995)
https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf