Statistical Methods in Particle Physics

7. Hypothesis Testing and Goodness-of-Fit

Heidelberg University, WS 2020/21

Klaus Reygers (lectures)
Rainer Stamen, Martin Völkl (tutorials)

Hypotheses and tests

Hypothesis test

- 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

Test statistic $t(\vec{x})$:
a (usually scalar) variable which is a function of the data alone that is used to test hypotheses

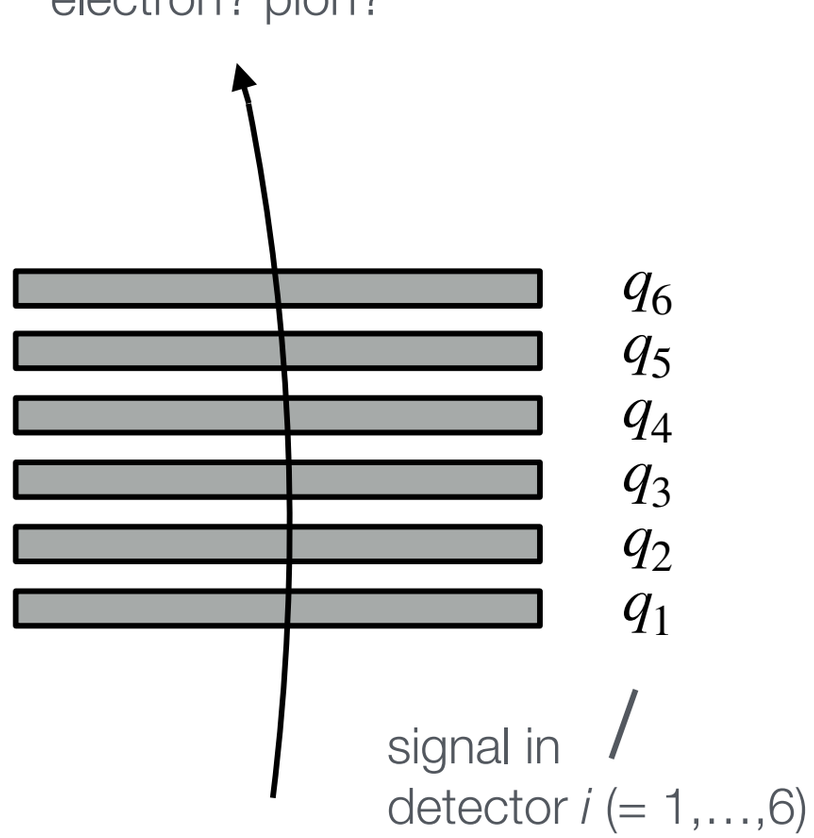
$\vec{x} = (x_1, \ldots, x_n)$: measured features/data

Examples:

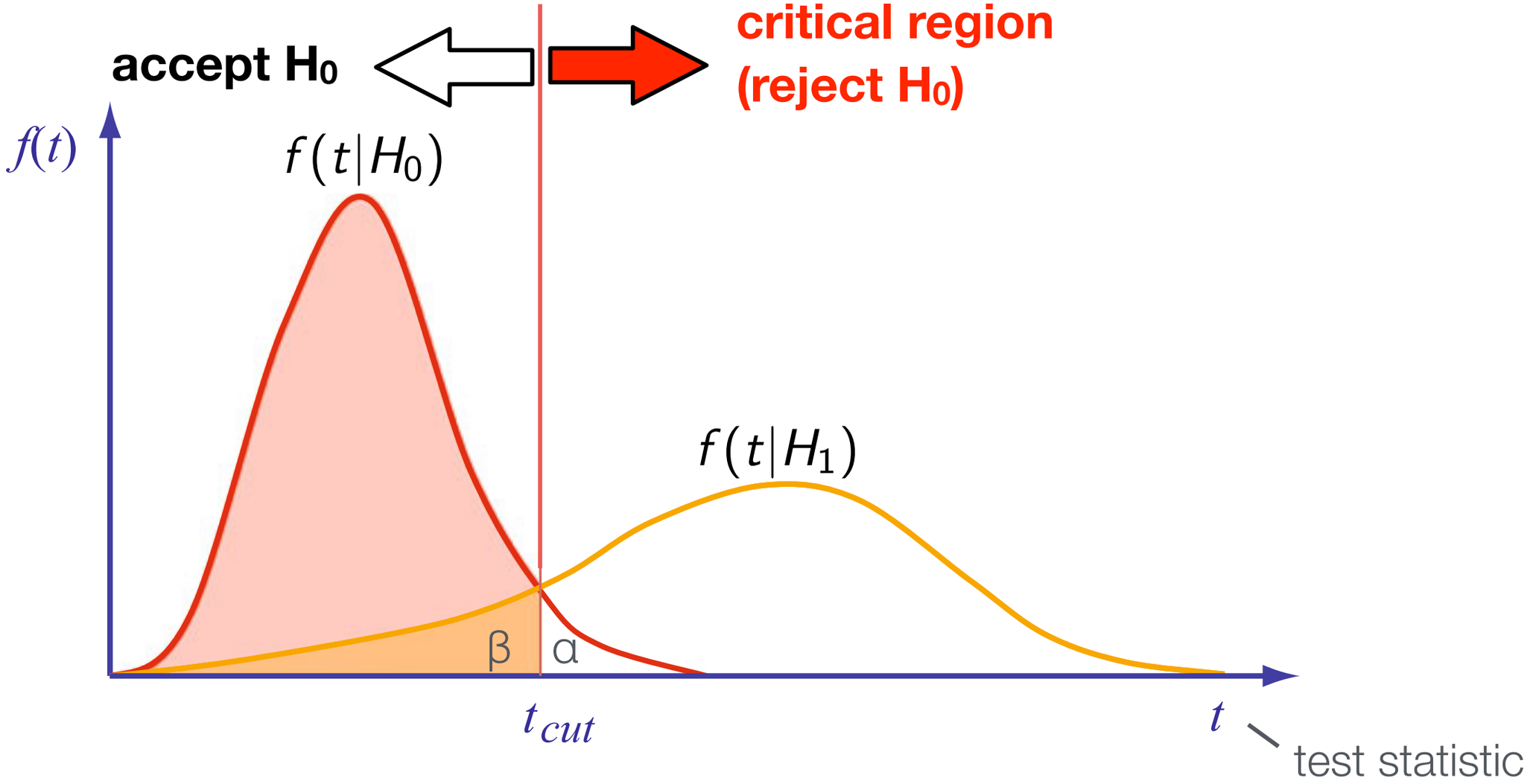
$t = X^2_{\text{min}}$ of a least-squares fit

ALICE TRD: likelihood ratio for electrons and pions: $t = \dfrac{\prod_{i=1}^{6} L_i(q_i \,|\, e)}{\prod_{i=1}^{6} L_i(q_i \,|\, \pi)}$

Output of a boosted decision tree or neural network



Critical region



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

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

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

Probability to reject H_1 even though it is true:

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

1– β:
"power of the test",
prob. to reject H_0 if H_1 is true

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)

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
- …

Test of significance

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 (here 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:
$\nu_b = 1.3$

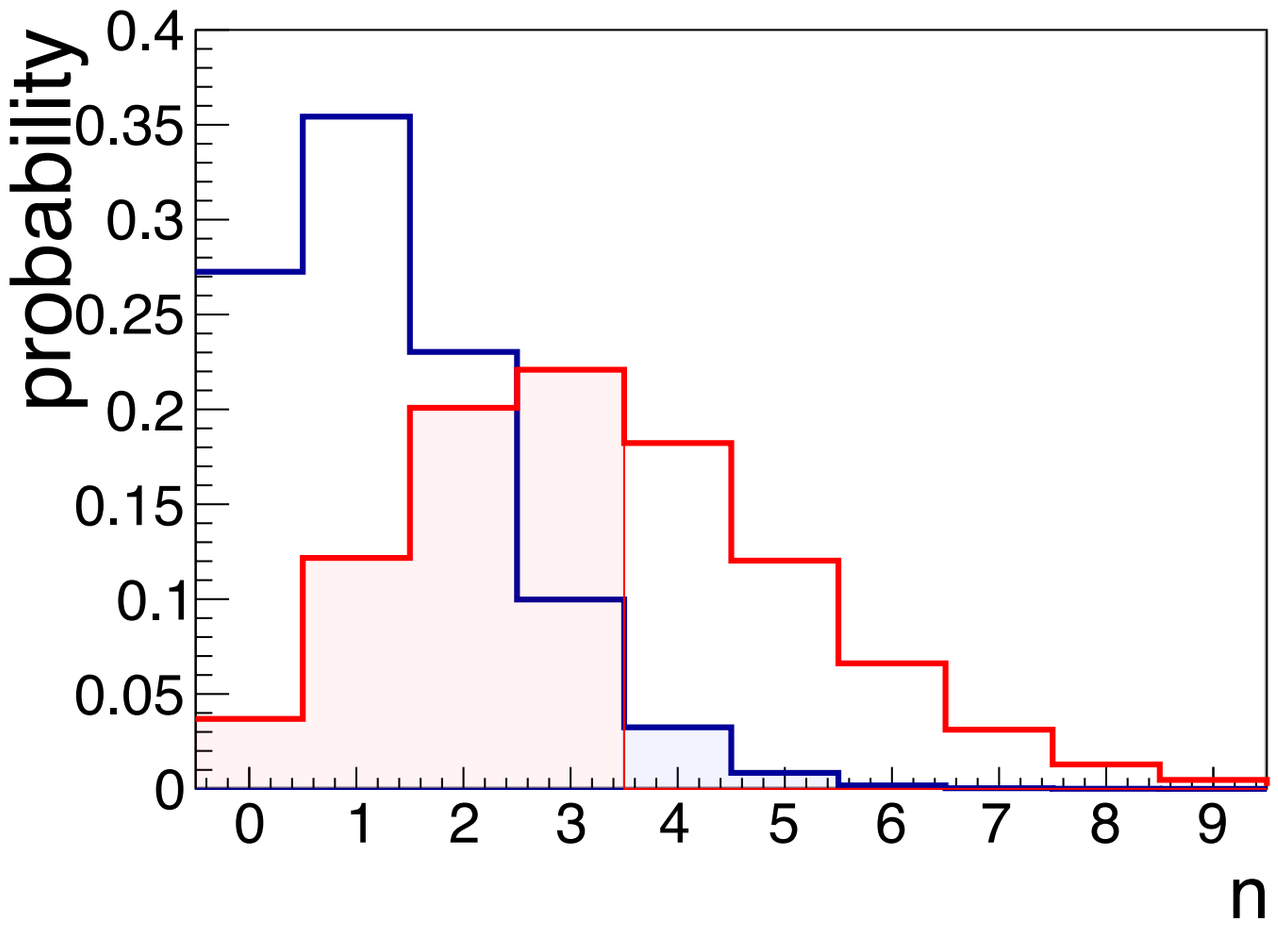
Expected signal events:
$\nu_s = 2$

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**

Goodness-of-fit for least squares fits (1)

The minimum $\chi^2(\hat{\vec{\theta}})$ of a least-squares fit is a measure of the level of agreement between the model and the data:

$$\chi^2_{\text{min}} = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i; \hat{\vec{\theta}})}{\sigma_i} \right)^2$$

Large χ^2_{min}: the model can can be rejected.

If the model is correct, then χ^2_{min} for repeated experiments follows a χ^2 distribution:

$$f(t; n_{\text{df}}) = \frac{1}{2^{n_{\text{df}}/2}\Gamma\left(\frac{n_{\text{df}}}{2}\right)} t^{n_{\text{df}}/2-1} e^{-t/2}, \qquad t = \chi^2_{\text{min}}$$

with $n_{\text{df}} = n - m = \text{number of data points} - \text{number of fit parameters}$

n_{df} = "number of degrees of freedom"

Goodness-of-fit for least squares fits (2)

Expectation value of the χ^2 distribution is n_{df}
→ $\chi^2 \approx n_{\mathrm{df}}$ indicates a good fit

Consistency of a model with the data is quantified with the *p*-value:

$$p\text{-value} = \int_{\chi^2_{\mathrm{min}}}^{\infty} f(t; n_{\mathrm{df}})\,\mathrm{d}t$$

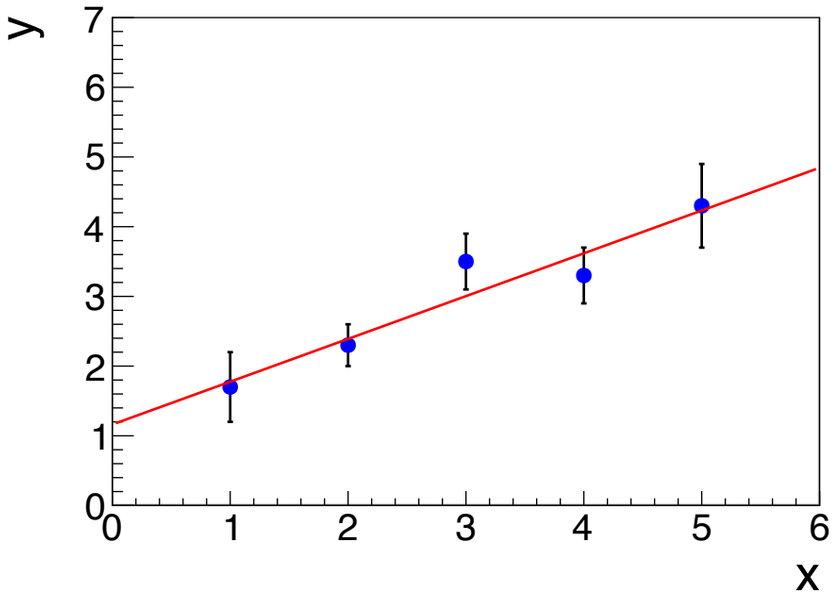
The *p*-value is the probability to get a χ^2_{min} as high as the observed one, or higher, if the model is correct.

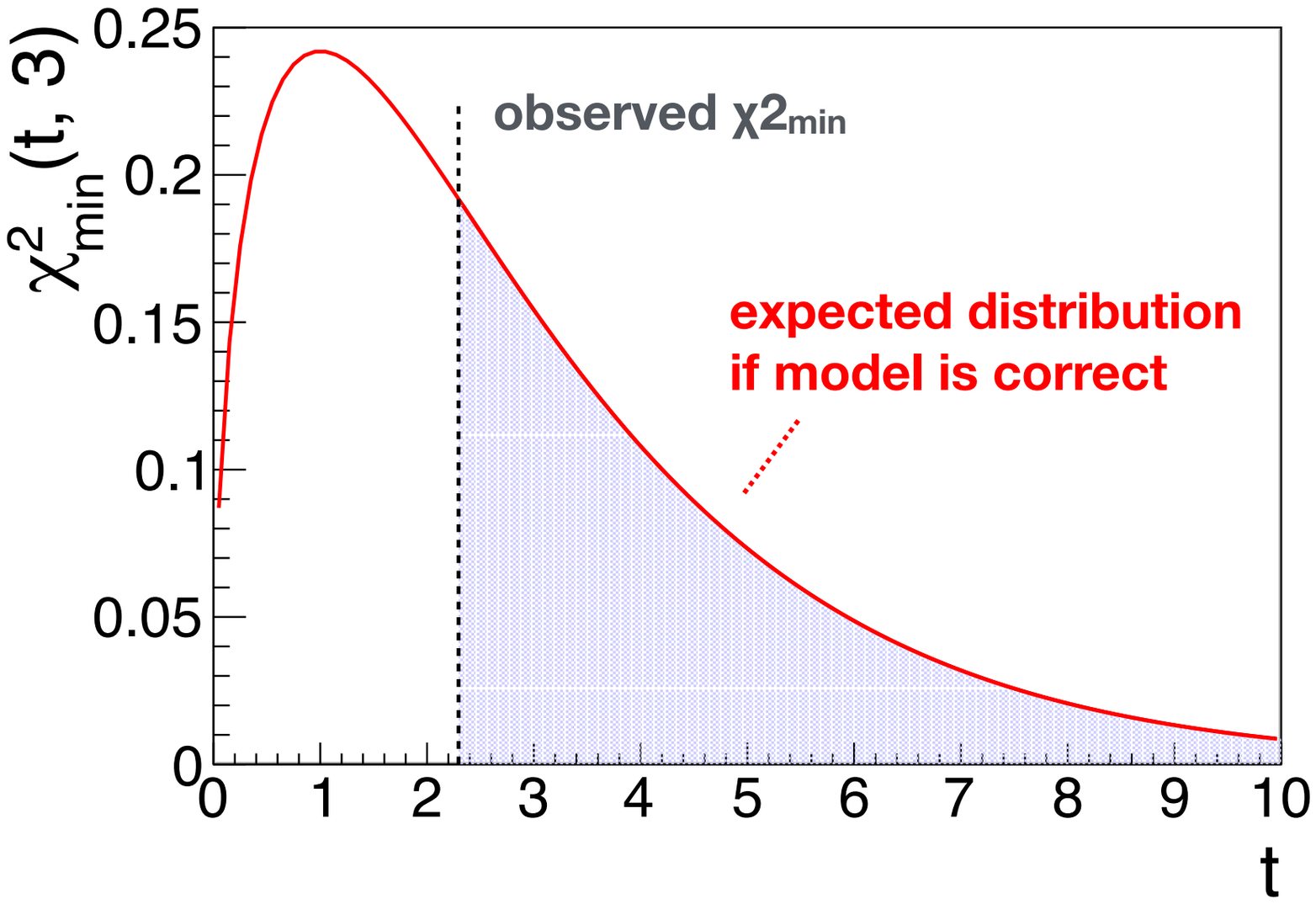
The *p*-value is **not** the probability that the model is correct.

p-value for the straight line fit example

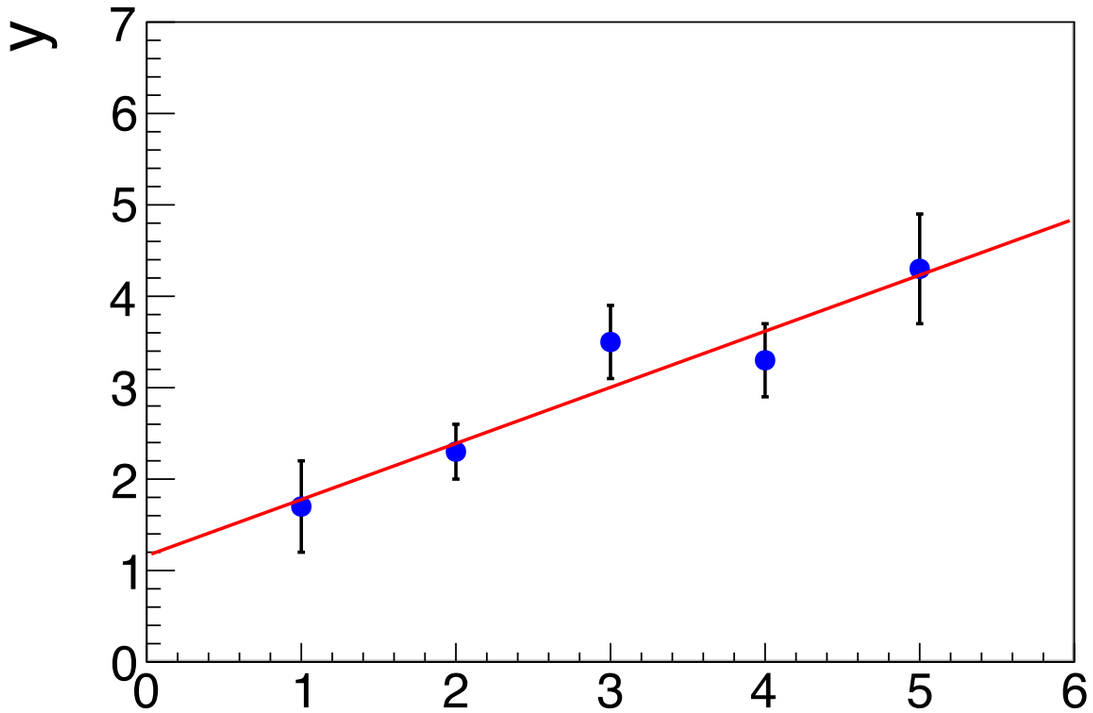
$\chi^2_{min} = 2.29557$, $n_{df} = 3$:

p-value = 0.51337





Constant model ($y = \theta_0$) rejected by small p-value

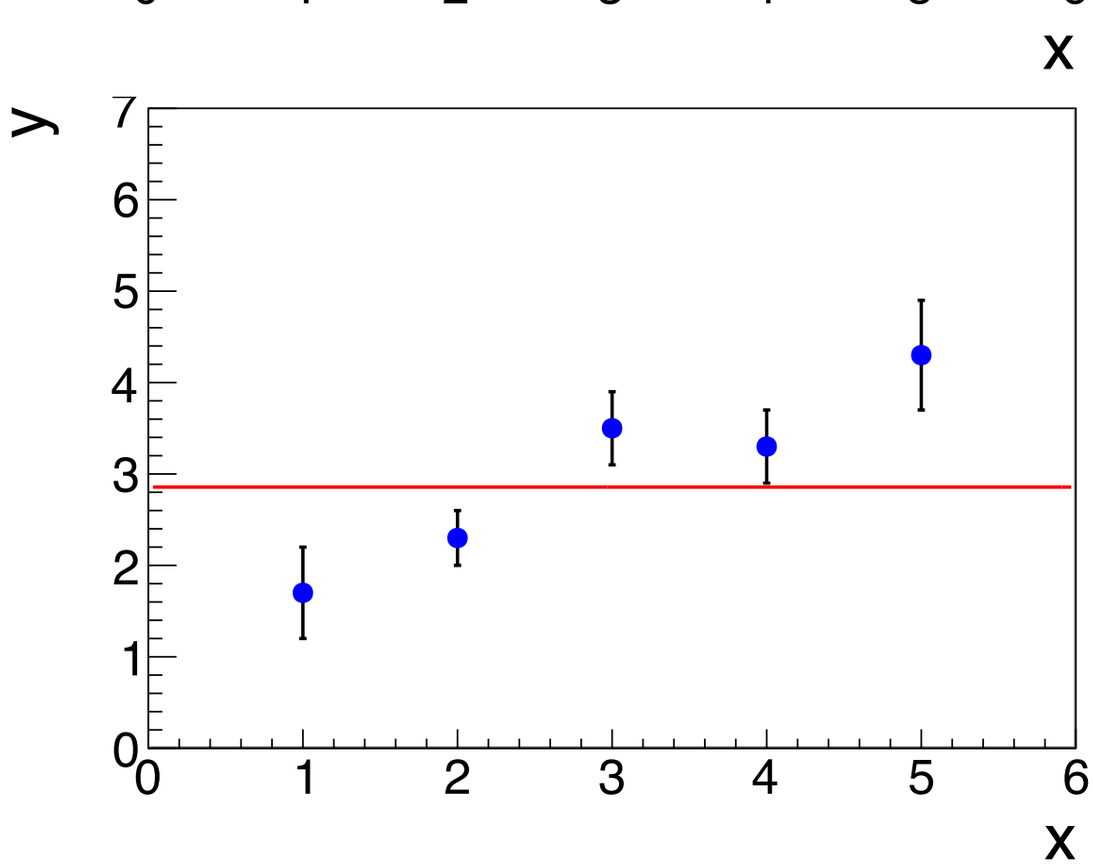


$\chi^2_{min} = 2.29557$, $n_{df} = 3$:

p-value = 0.51337

```
from scipy import stats
pvalue = 1 - stats.chi2.cdf(chi2, n_dof)
```

```
root [1] TMath::Prob(chi2, n_dof)
```



$\chi^2_{min} = 18.3964$, $n_{df} = 4$:

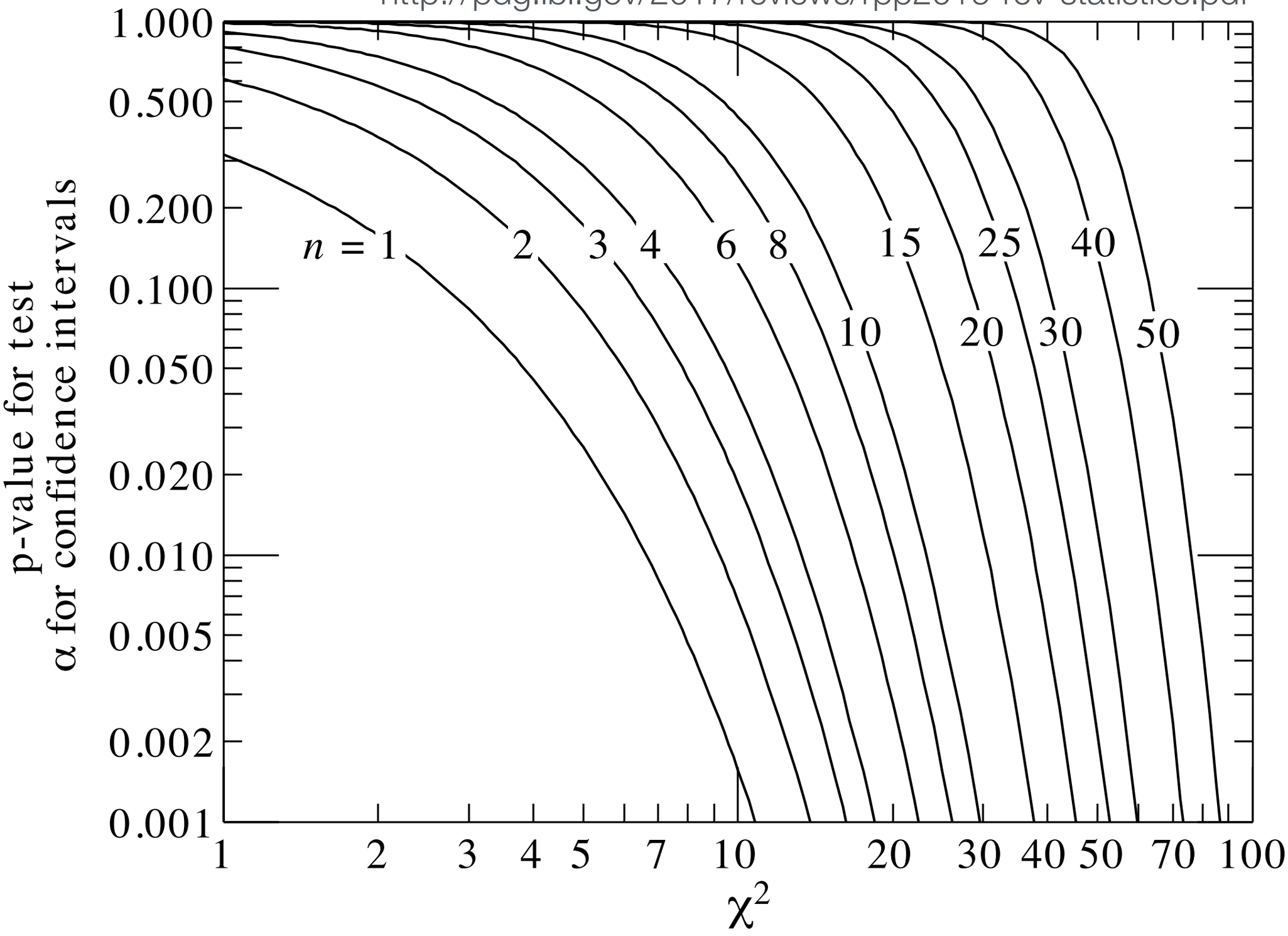
p-value = 0.001032

$\theta_0 = 2.86 \pm 0.18$

Statistical uncertainty of the fit parameter does not tell us whether model is correct!

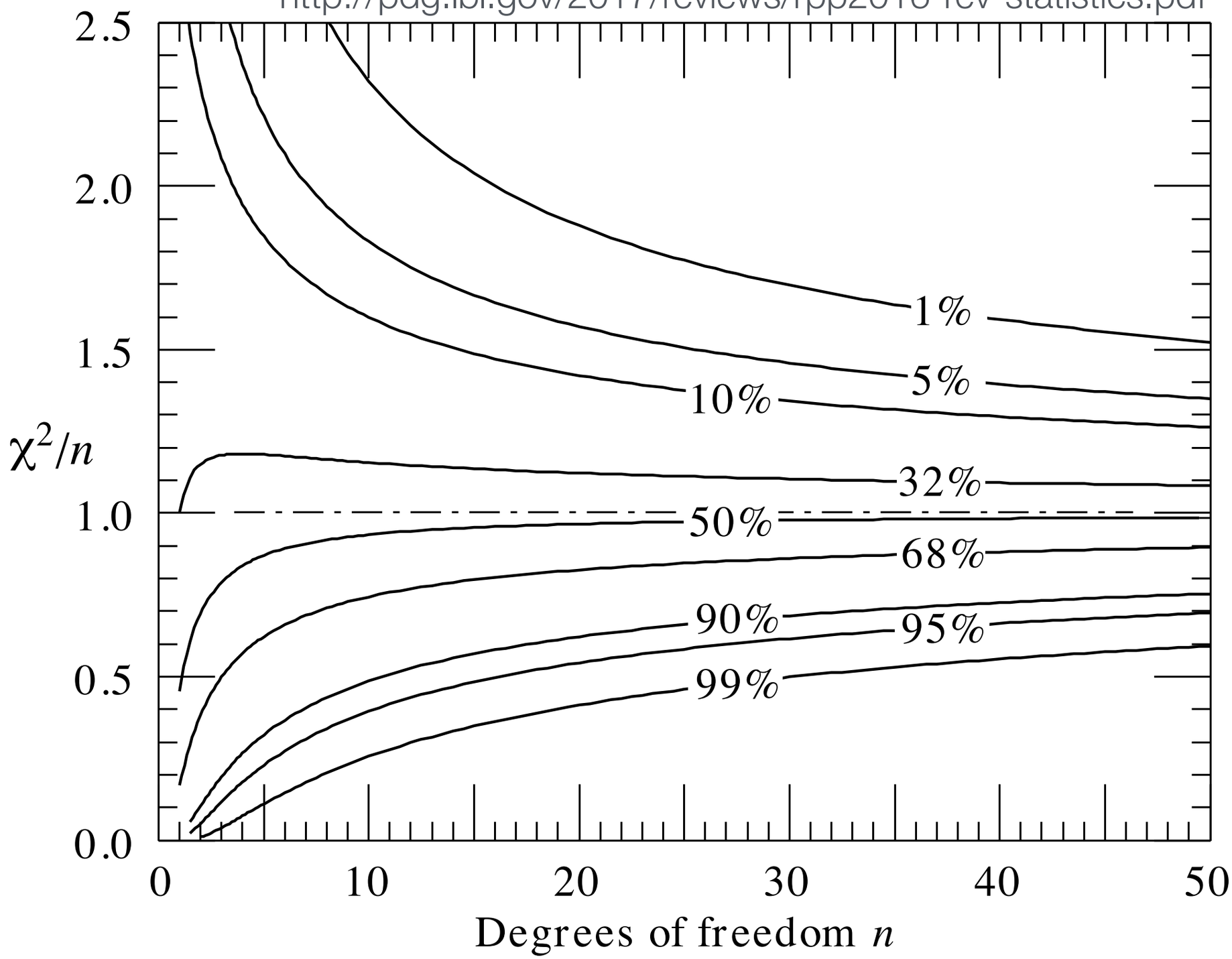
p-value for different χ^2_{min} and n_{df}

http://pdg.lbl.gov/2017/reviews/rpp2016-rev-statistics.pdf



Confidence Intervalls for χ^2_{min} / n_{df} as a fct. of n_{df}

http://pdg.lbl.gov/2017/reviews/rpp2016-rev-statistics.pdf



Goodness-of-fit for unbinned ML fits (1)

In case of an unbinned ML fit one can put data and model prediction into a histogram and perform a χ^2 test.

Consider the ratio

L: likelihood

$$\lambda = \frac{L(\vec{n}|\vec{\nu})}{L(\vec{n}|\vec{n})}, \qquad \vec{\nu} = \vec{\nu}(\vec{\theta}), \quad \vec{\theta} = (\theta_1, ..., \theta_m)$$

For the multinomial ("M", n_{tot} fixed) and Poisson distributed data ("P") one obtains

k: number of bins of the histogram

$$\lambda_{\text{M}} = \prod_{i=1}^{k} \left(\frac{\nu_i}{n_i}\right)^{n_i}, \qquad \lambda_{\text{P}} = e^{n_{\text{tot}} - \nu_{\text{tot}}} \prod_{i=1}^{k} \left(\frac{\nu_i}{n_i}\right)^{n_i}$$

We then consider

$$\chi^2 := -2 \ln \lambda$$

Goodness-of-fit for unbinned ML fits (2)

For multinomially distributed data

$$\chi^2_{\mathrm{M}} := -2 \ln \lambda_{\mathrm{M}} = 2 \sum_{i=1}^{k} n_i \ln \frac{n_i}{\hat{\nu}_i}$$

follows a χ^2 distribution for $k - m - 1$ degrees of freedom in the large sample limit for if the model is correct.

In case of Poisson distributed data

$$\chi^2_{\mathrm{P}} := -2 \ln \lambda_{\mathrm{P}} = 2 \sum_{i=1}^{k} \left(n_i \ln \frac{n_i}{\hat{\nu}_i} + \hat{\nu}_i - n_i \right)$$

follows a χ^2 distribution for $k - m$ degrees of freedom in the large sample limit if the model is correct.

S. Baker, R. D.Cousins
Clarification of the use of CHI-square and likelihood functions in fits to histograms, NIM 221 (1984) 437

Goodness-of-fit ML Test Using L_{max}

For ML fits the value of the likelihood function at the maximum $L_{max}(x|\theta_0) \equiv L_{max,obs}$ is sometimes used as a Goodness-of-Fit test

- Generate pseudo-data based on best-fit parameters
- Repeat fit with pseudo data → L_{max} distribution
- From the L_{max} distribution one can determine how likely it is to find a value $L_{max,obs}$ or smaller

This is briefly discussed in the books by G. Cowan and F. James.

This method is generally discouraged

- Biased and not invariant with respect to change of variables
- From J. Heinrich, PHYSTAT2003, arXiv:physics/0310167
  "The method is fatally flawed in the unbinned case. Don't use it. Complain when you see it used."

Wilks' theorem

Let null hypothesis H_0 be a special case of the hypothesis H_1 ("nested hypotheses")

Example:

$$H_0 : \; f(m) = a_0 + a_1 m$$
$$H_1 : \; f(m) = a_0 + a_1 m + a_2 m^2 + a_3 m^3$$

Define:

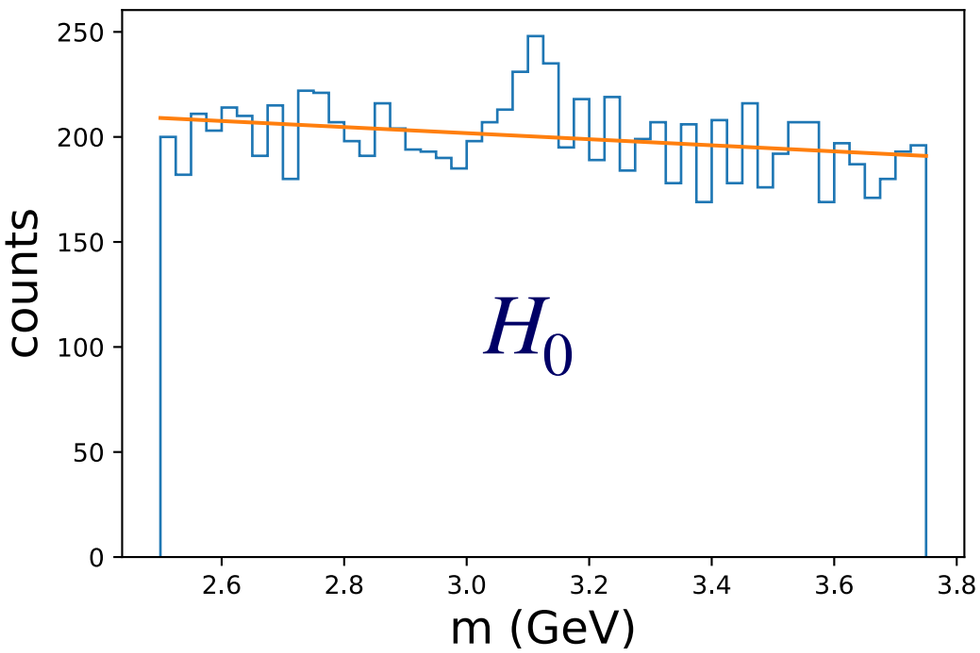
$$\Delta\tilde{\chi}^2 := -2 \ln \left(\frac{L(H_1)}{L(H_0)} \right)$$

Wilks' theorem:
If H_0 is correct then $-\Delta\tilde{\chi}^2$ follows χ^2 distribution with n_{dof} = #added parameters in the large sample limit.

In the above example: $n_{\mathrm{dof}} = 2$

Samuel S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses Ann. Math. Statist., Volume 9, Number 1 (1938), 60-62.

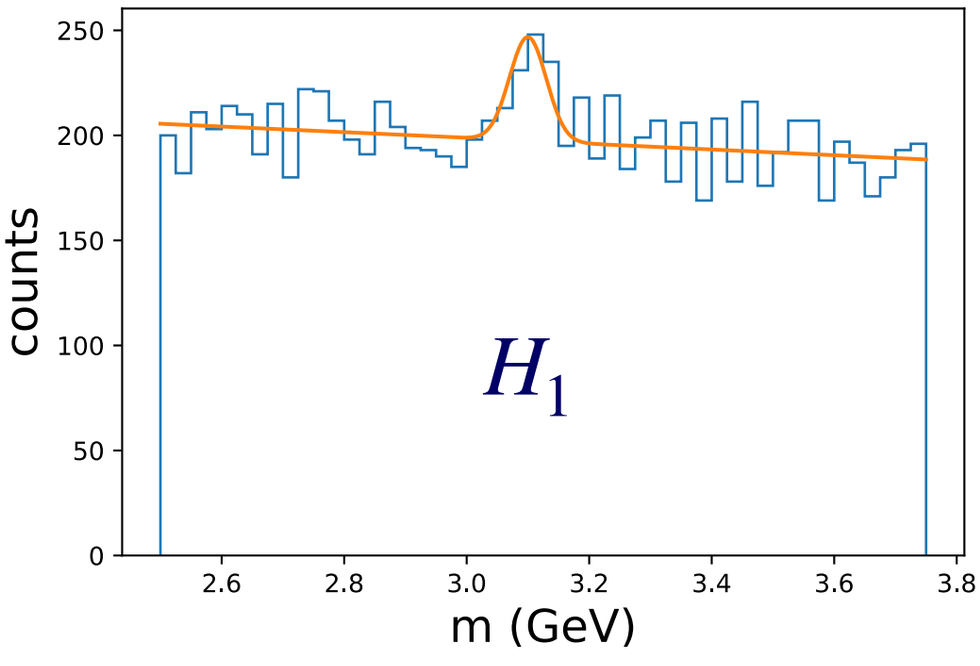
Significance of a peak



$$H_0: f(m) = a_0 + a_1 m$$
$$H_1: f(m) = a_0 + a_1 m + a_2 N(m; \mu, \sigma)$$

$\mu = 3.1, \sigma = 0.03$ fixed in H_1
→ one additional parameter

$$\Delta\tilde{\chi}^2 := -2 \ln \left(\frac{L(H_1)}{L(H_0)} \right) = -22.5$$



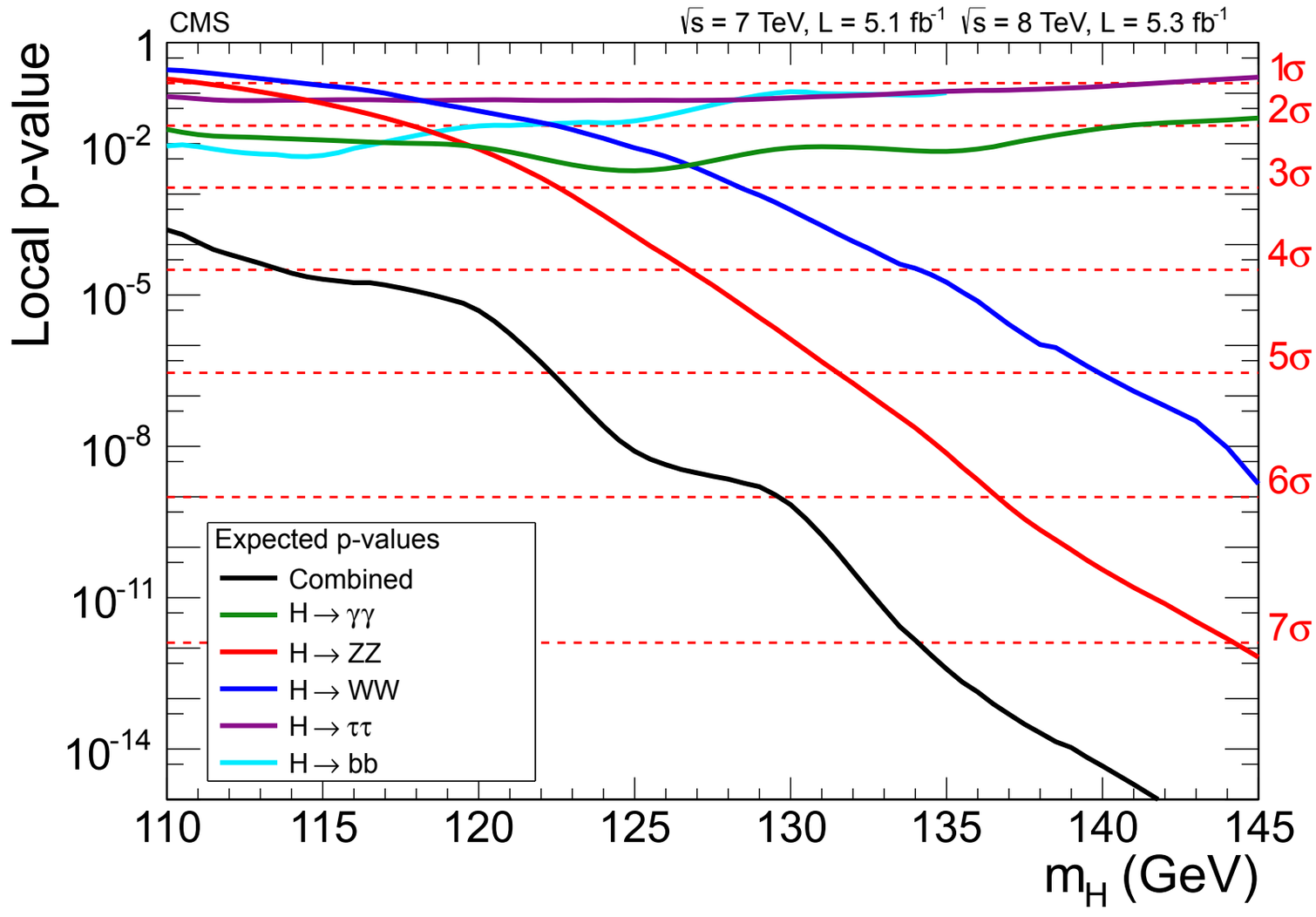
$-\Delta\tilde{\chi}^2$ should follow a χ^2 distribution with $n_{\mathrm{dof}} = 1$ if H_0 ist true

p-value = $2.15 \cdot 10^{-6}$

→ H_0 can be safely rejected

CMS, arXiv:1207.7235

p-values and Higgs measurement: Expected local *p*-values for a Higgs of a given mass

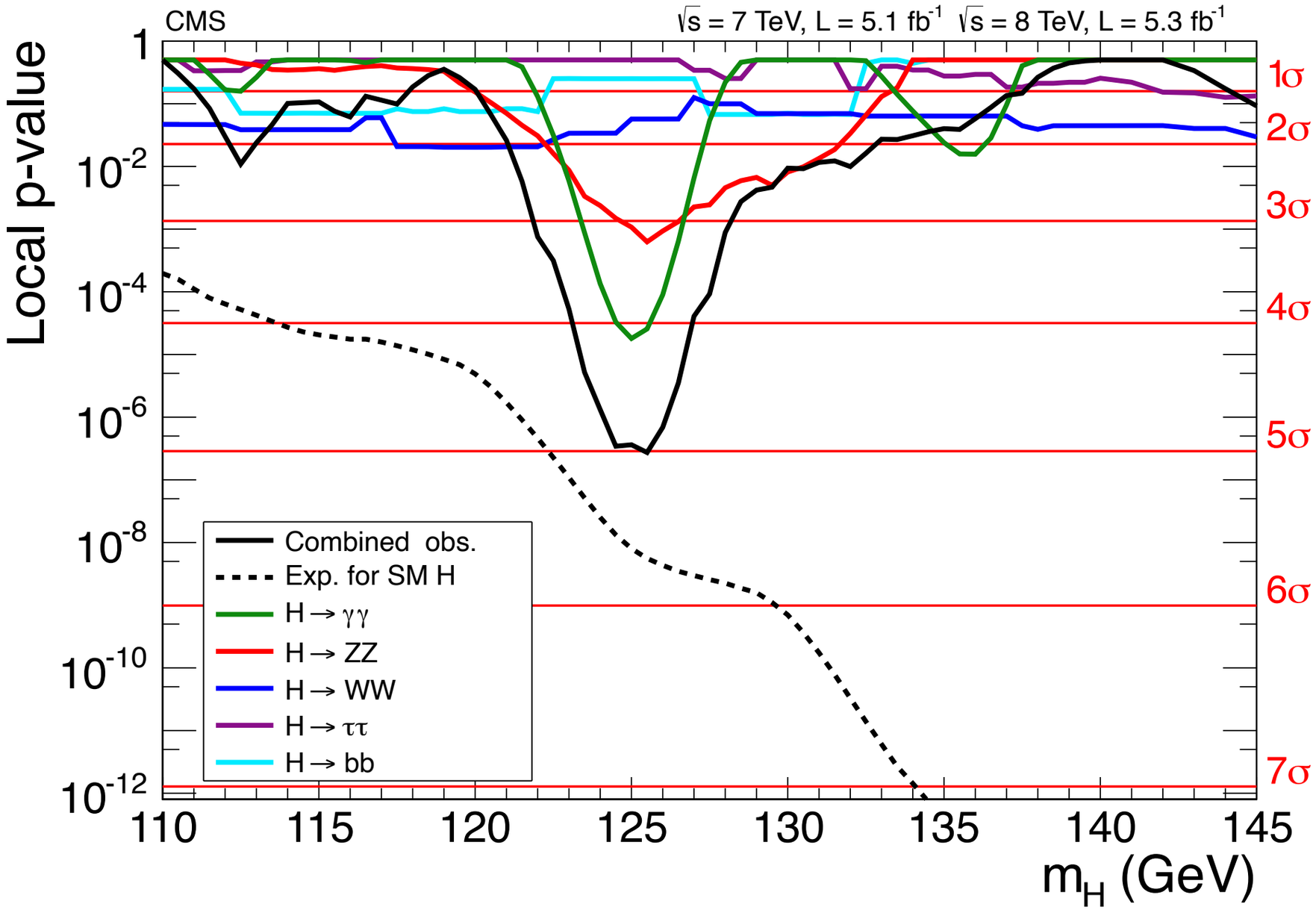


For each assumed Higgs mass (→ local *p*-value)

- Calculate expected signal for Standard Model Higgs boson
- Determine *p*-value for H_0 that only SM background processes contribute
- Pure calculation/simulation, no data involved

CMS, arXiv:1207.7235

p-values and 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, signaling 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

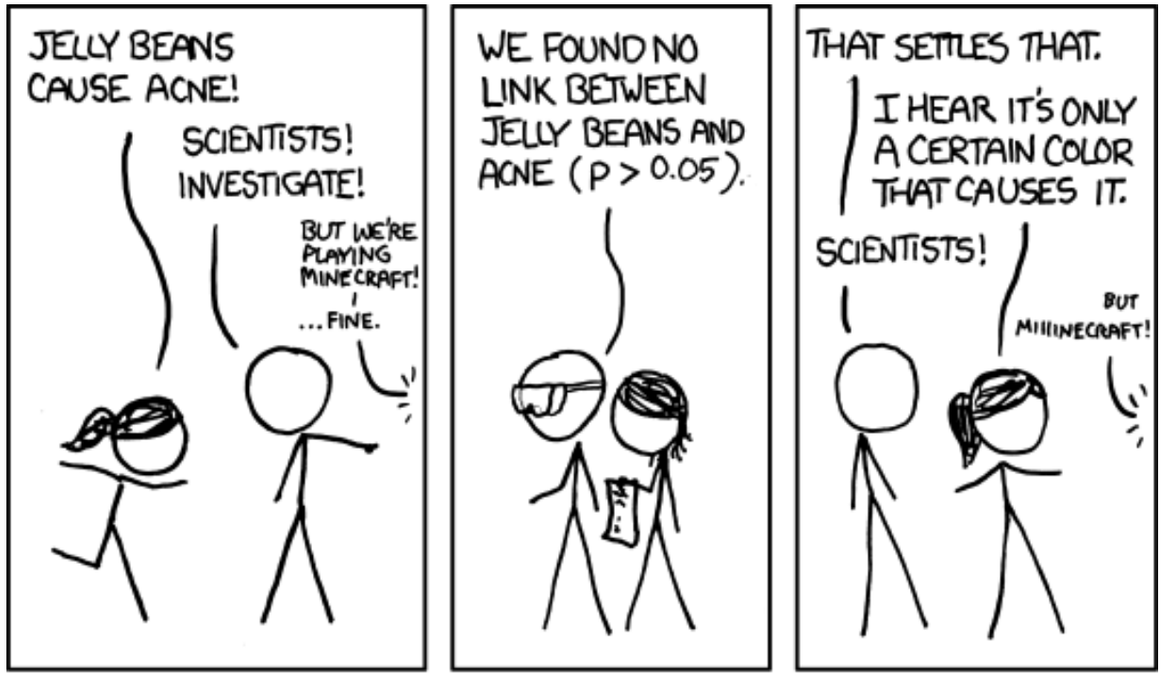
In general:

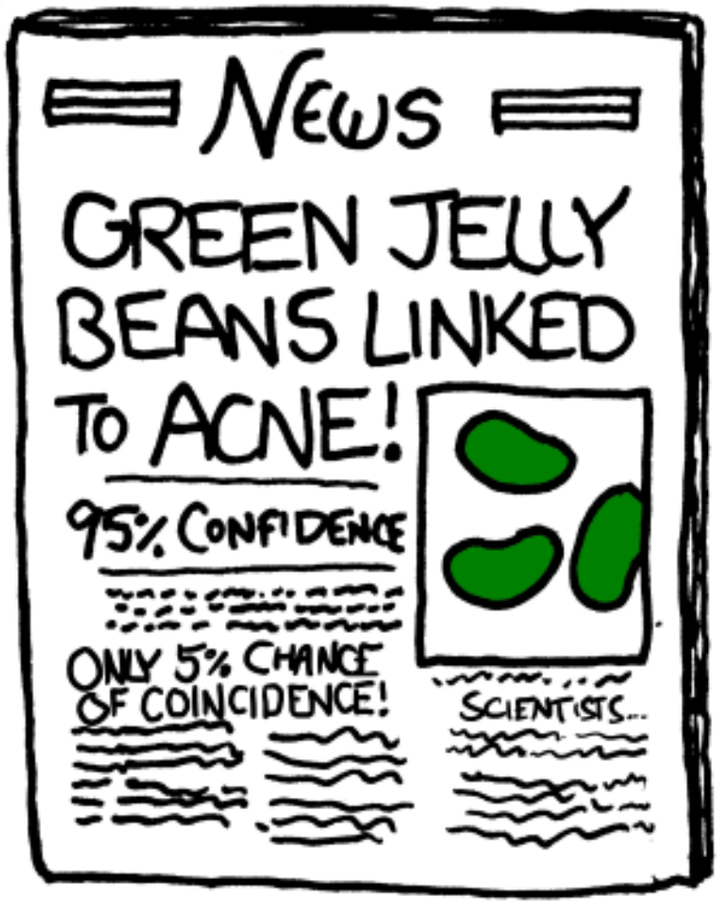
- If one is performing multiple tests then obviously a *p*-value of 1/*n* is likely to occur after *n* tests
- Solution: "trials penalty" or "trials factors", i.e. make threshold a function of *n* (more stringent threshold for larger *n*)

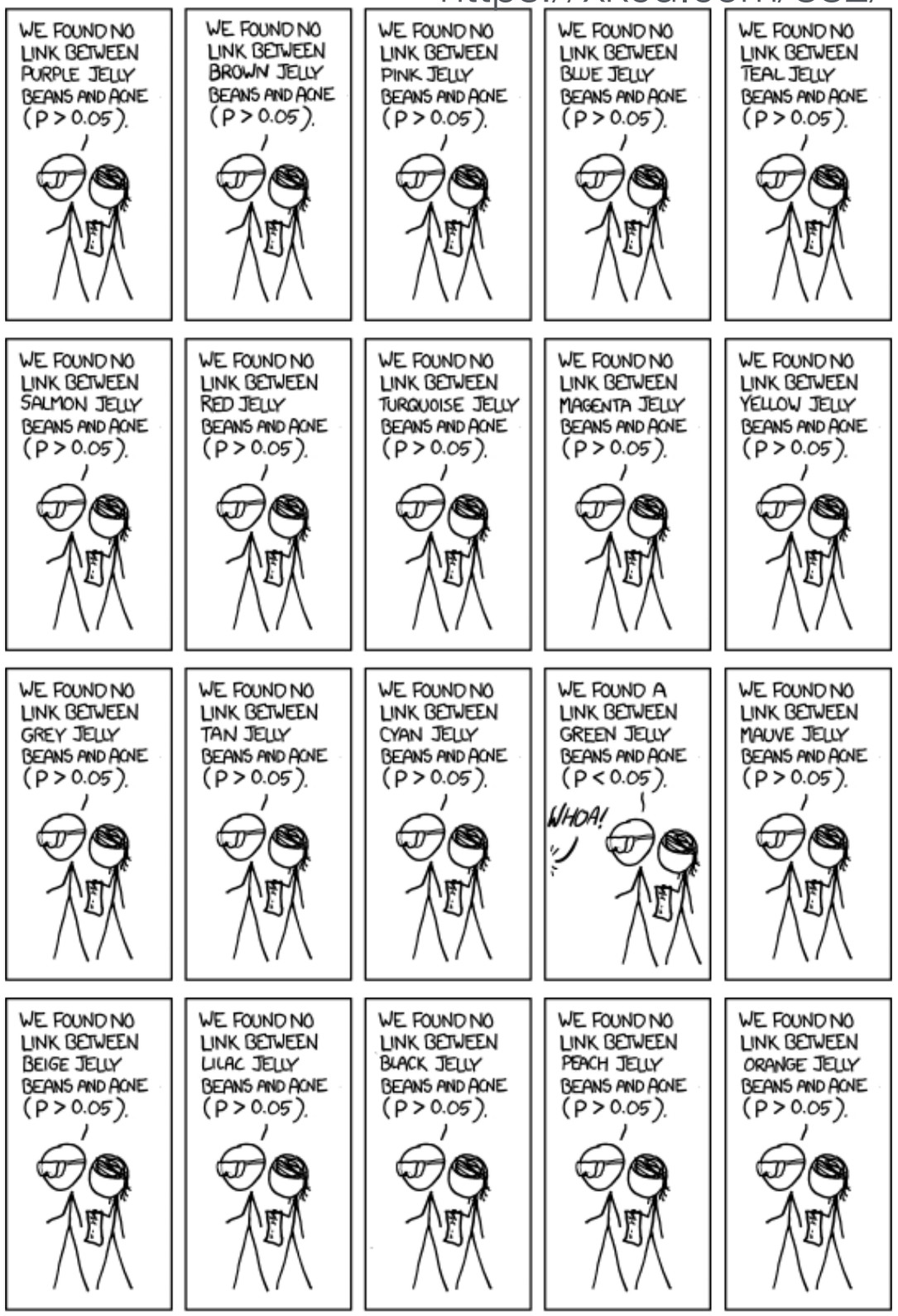
A Swedish study in 1992 tried to determine whether or not power lines caused some kind of poor health effects. The researchers surveyed everyone living within 300 meters of high-voltage power lines over a 25-year period and looked for statistically significant increases in rates of over 800 ailments. 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 result 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

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 (1)

KS test is an unbinned goodness-of-fit test

Q: Do data points come from a given distribution?

Compare cumulative distribution function

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

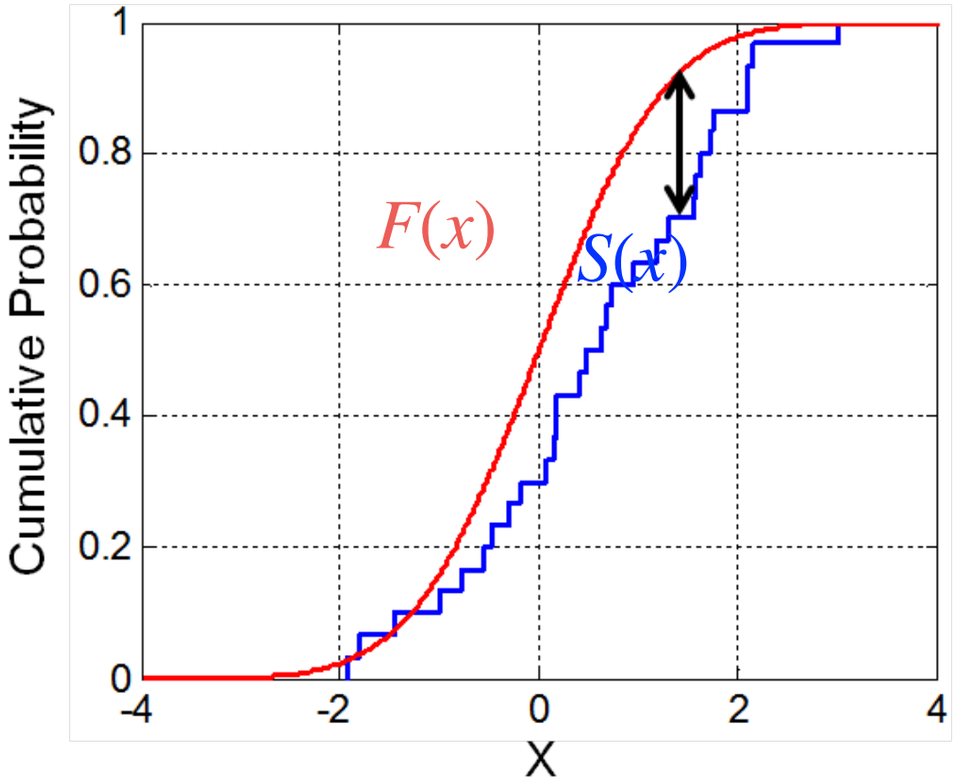
with the so-called Empirical Distribution Function (EDF)

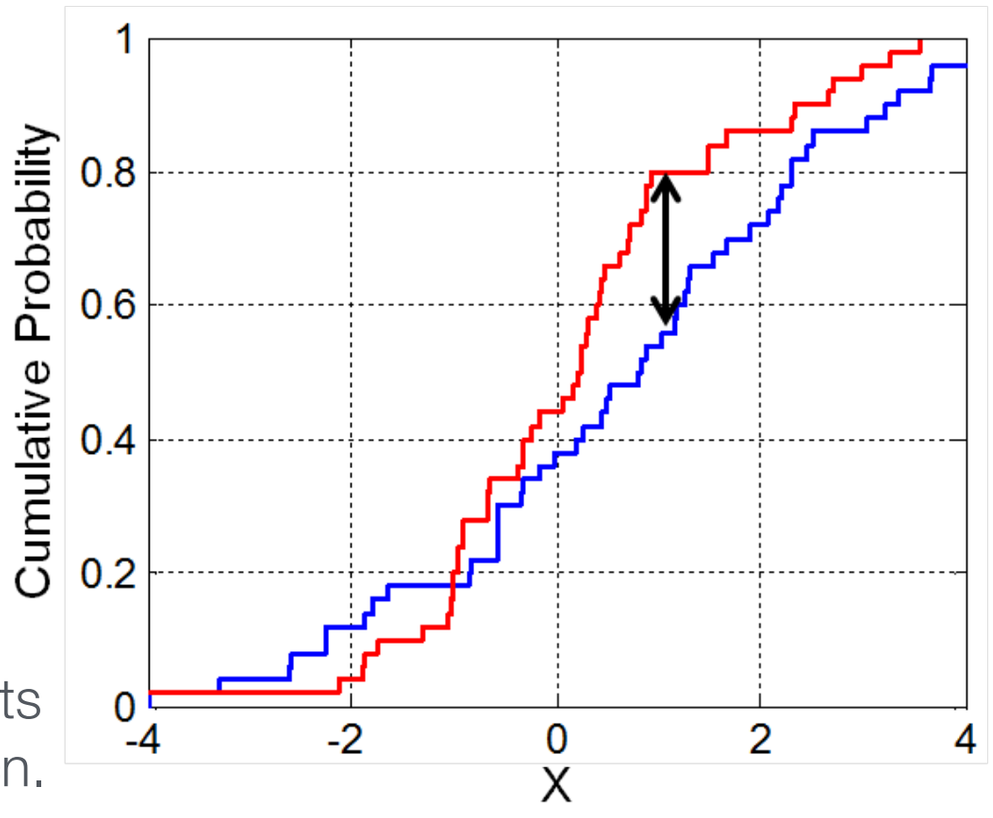
$$S(x) = \frac{\text{number of observations with } x_i < x}{\text{total number of observations}}$$

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.

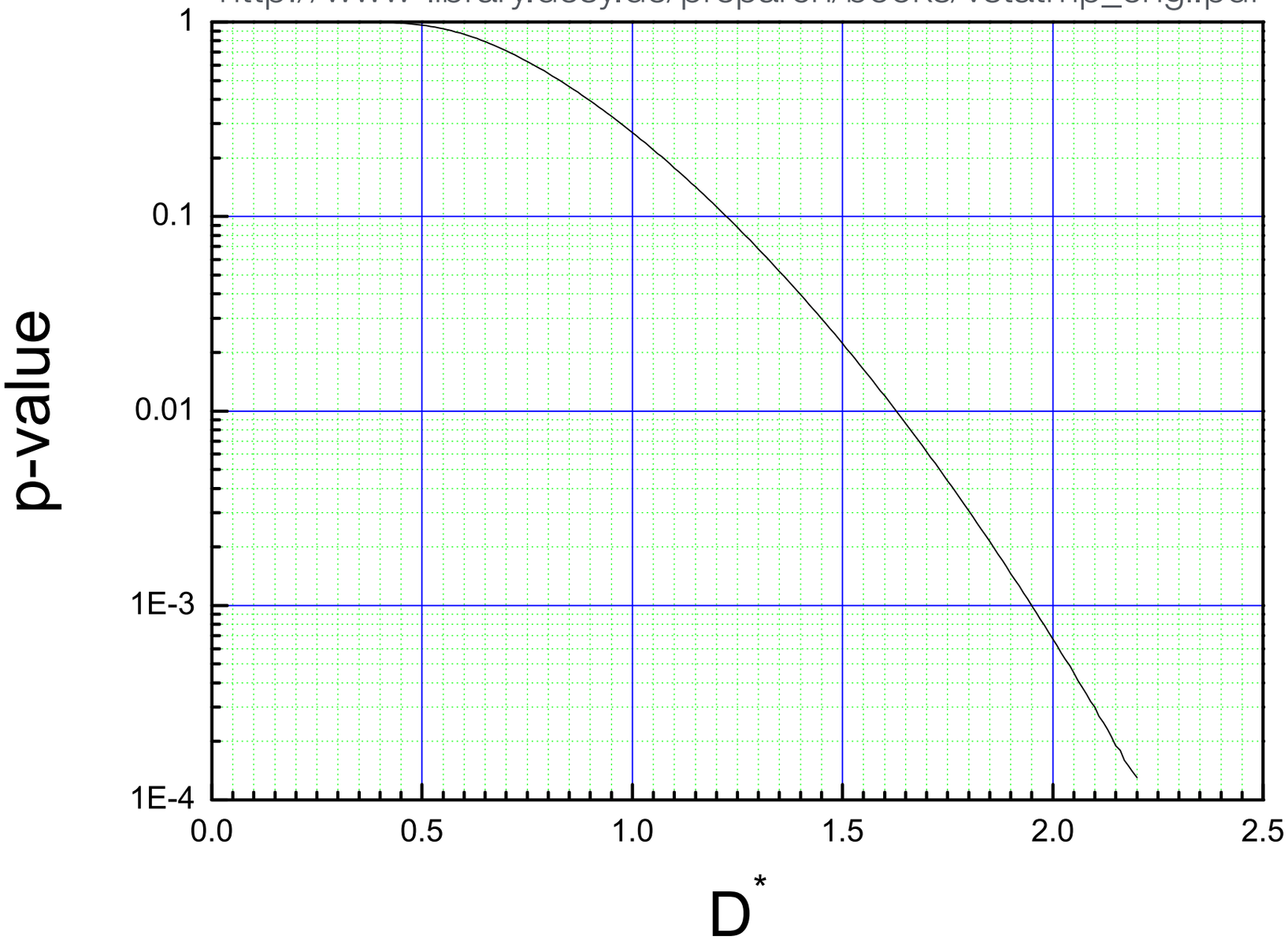




Kolmogorov–Smirnov Test (2)

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

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



$D^* = \sqrt{N} D,$

$N =$ number of data points

Example:
Test whether data x_i come from standard normal distribution N(0,1):

```
from scipy import stats
D, p_value =
stats.kstest(x, stats.norm.cdf)
```

Kolmogorov–Smirnov test: only for 1d data

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

Consider N bins, $N = N_+ + N_-$
N_+: number of positive deviations, N_-: number of negative deviations

run = consecutive bins where the data show deviations in the same direction

++++−−−+++−−+++++−−−− $N = N_+ + N_- = 22$ bins, 6 runs

Mean and variance for the number of runs for the null hypothesis that each element in the sequence is independently drawn from the same distribution (no assumption about prob. for "+" and "–"):

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

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:

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

Profile likelihood ratio as test statistics

Let q be a test statistic and $h(q \mid \theta, \nu)$ its distribution. The p-value depends on the nuisance parameter ν:

$$p_\theta(\nu) = \int_{q_{\text{obs}}}^{\infty} h(q|\theta, \nu)\, dq$$

Independence of the nuisance parameter is achieved approximately by using the *profile likelihood ratio* as test statistic:

$$\lambda_{\text{p}}(\theta) = \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

This is motivated by the fact that $-2 \ln \lambda_{\text{p}}(\theta)$ approaches the χ^2 distribution (with n_{dof} = number of parameters of interest) for a large data sample (→ Wilks' theorem).