Statistical Methods in Particle Physics

7. Hypothesis Testing

Heidelberg University, WS 2023/24

Klaus Reygers, Martin Völkl (lectures)
Ulrich Schmidt, (tutorials)

Some difficult topics

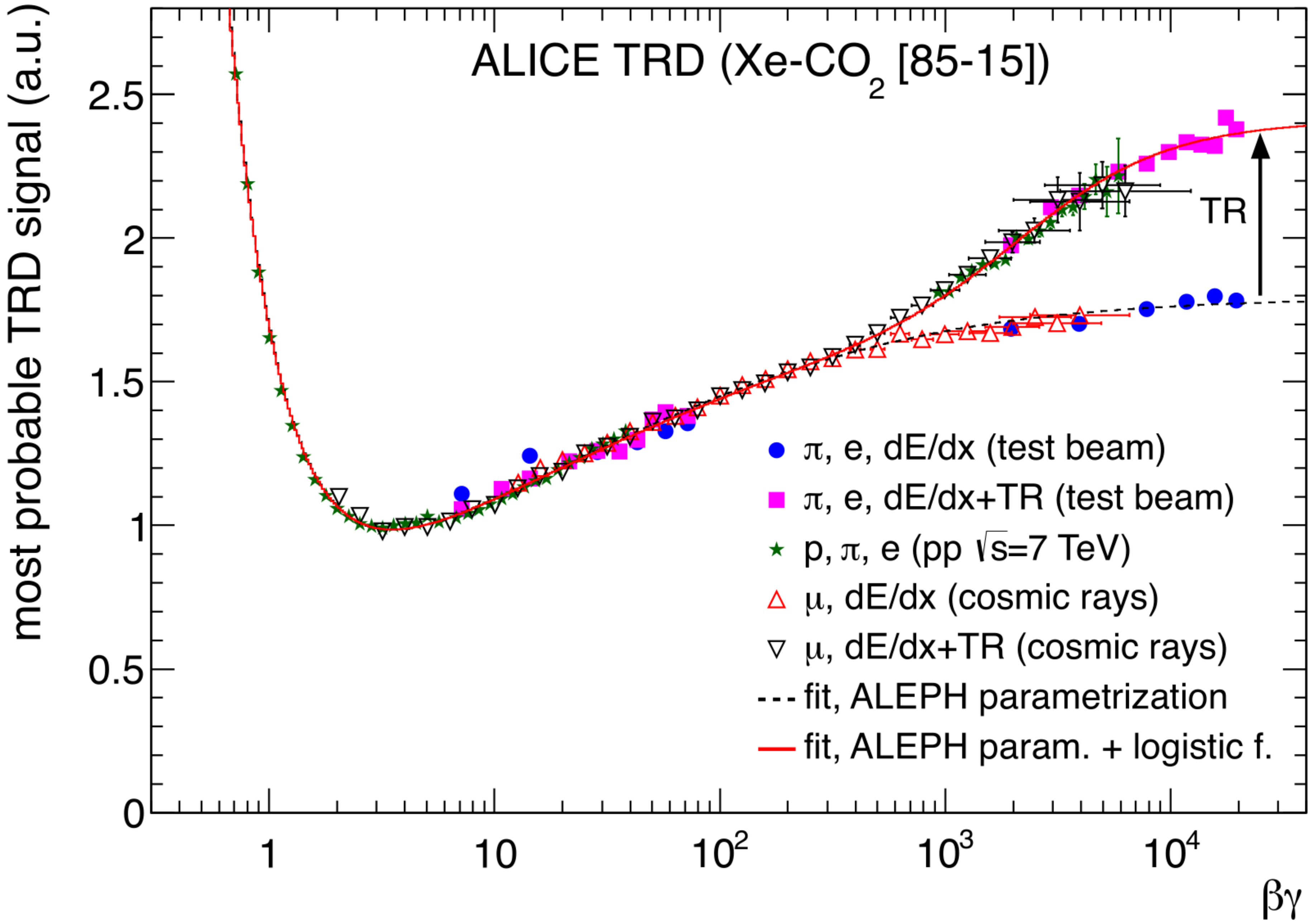
Errors in independent variables

- What happens if we have errors in x *and* y?
- Actually quite complex question
- What happens if we repeat the measurement? Do we:

  A. Draw a new true x-value from some distribution?

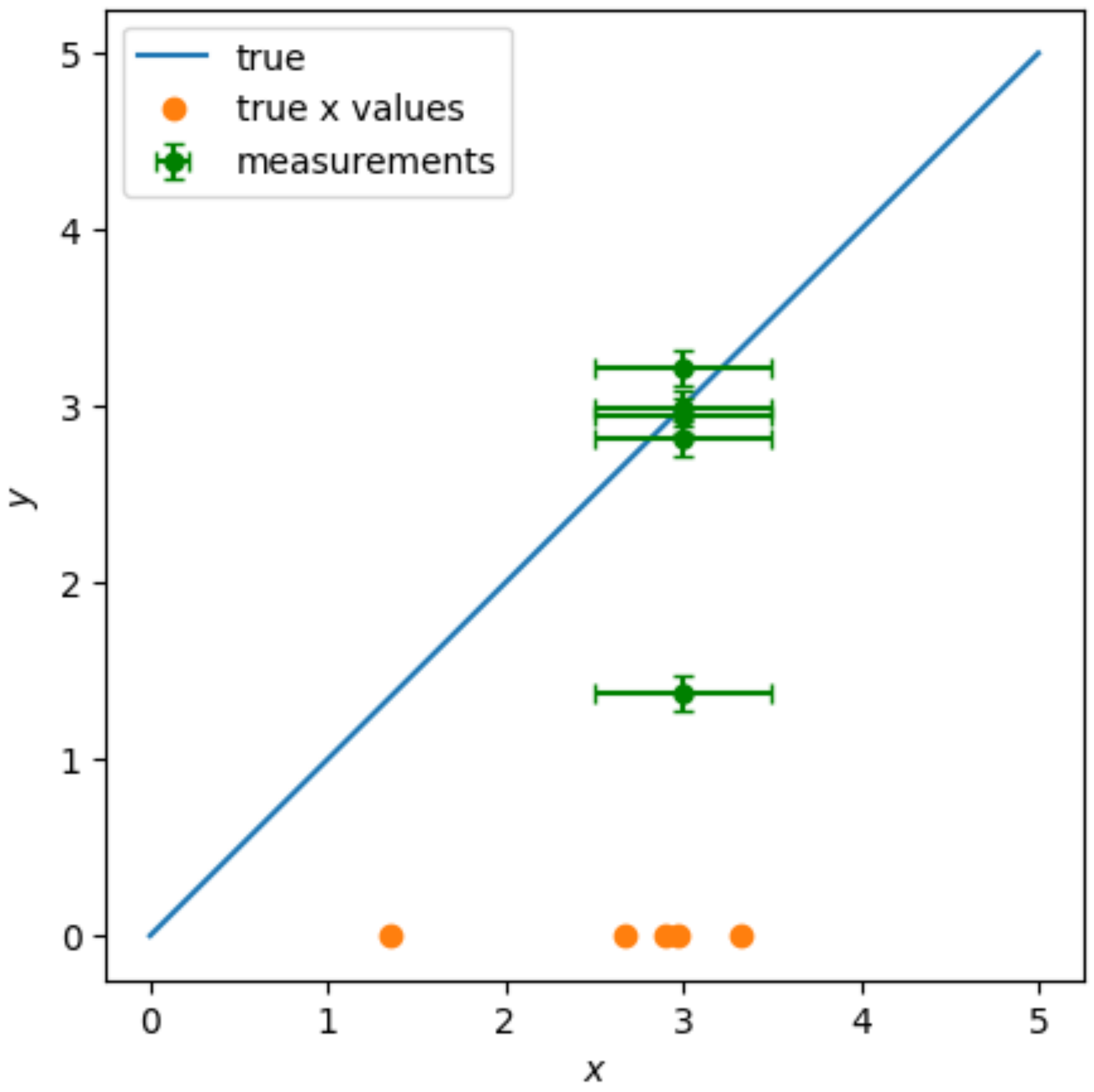
  B. Measure the same unknown x-value again?

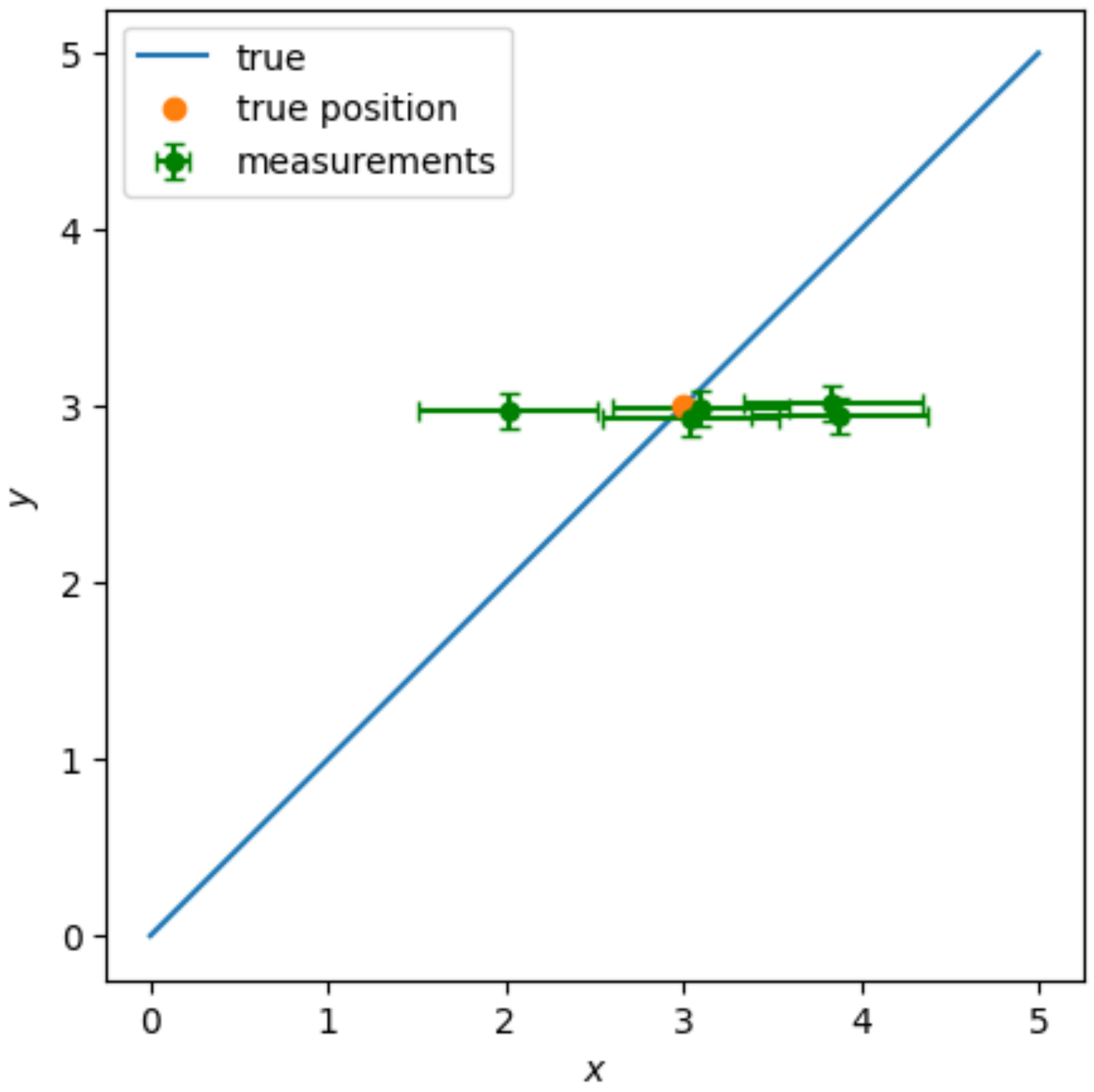
- “A” leads to *structural models*, while “B” leads to *functional models*
- When repetition of measurement is fixed, and the x- and y- variances are known, we can calculate a likelihood
- Many models deal with unknown variances, not usual in physics



Example from ALICE performance report, The measured particle $\beta\gamma$ fluctuates when the experiment is repeated.

Errors in independent variables (2)





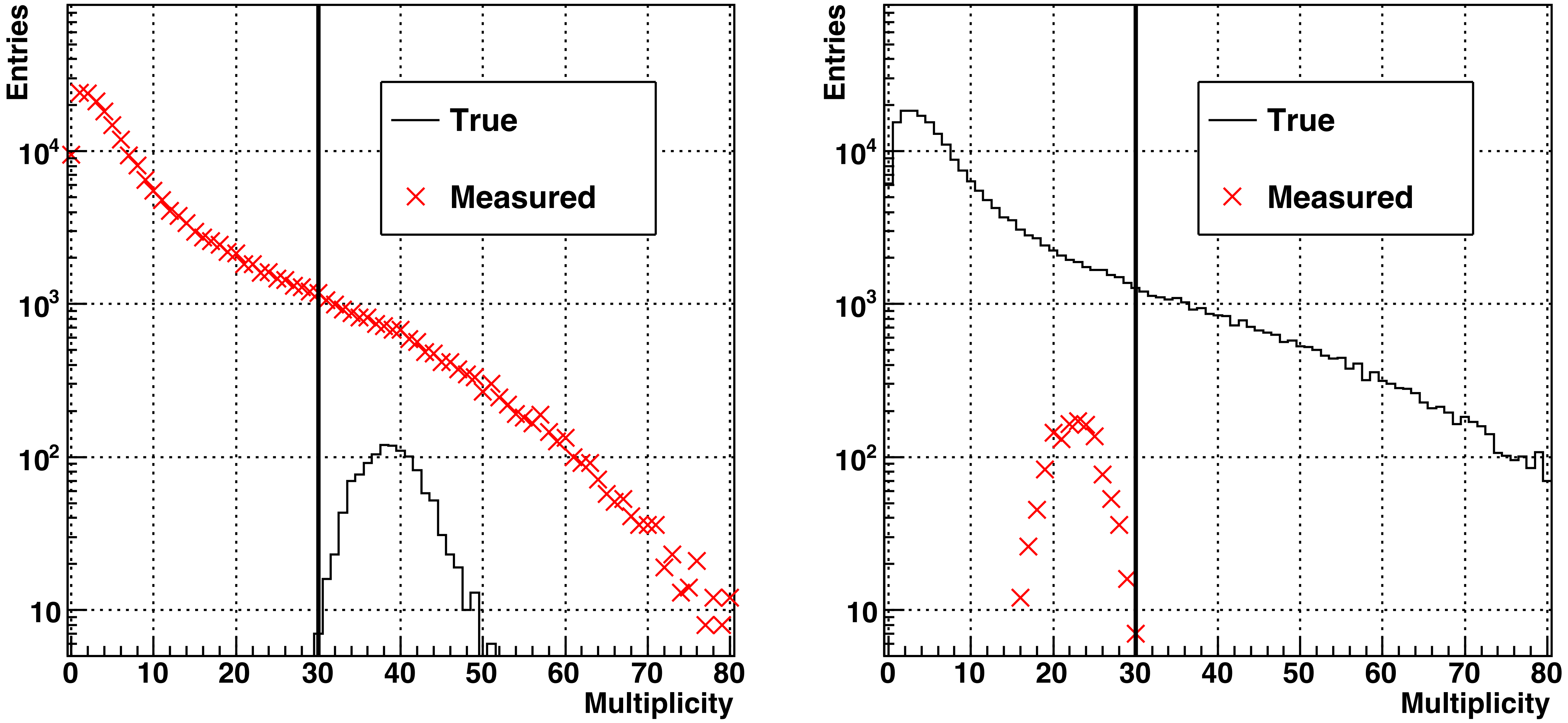
- Example 1: We put the setting on the machine to some x-value (e.g. the voltage), but the true x-value fluctuates around this, giving an uncertainty
- Example 2: There is some true constant x-value, which is unknown. The measurements fluctuate around it.

Since the fluctuations depend on the formulation of the problem, so must the best fit.

Errors in independent variables (3)

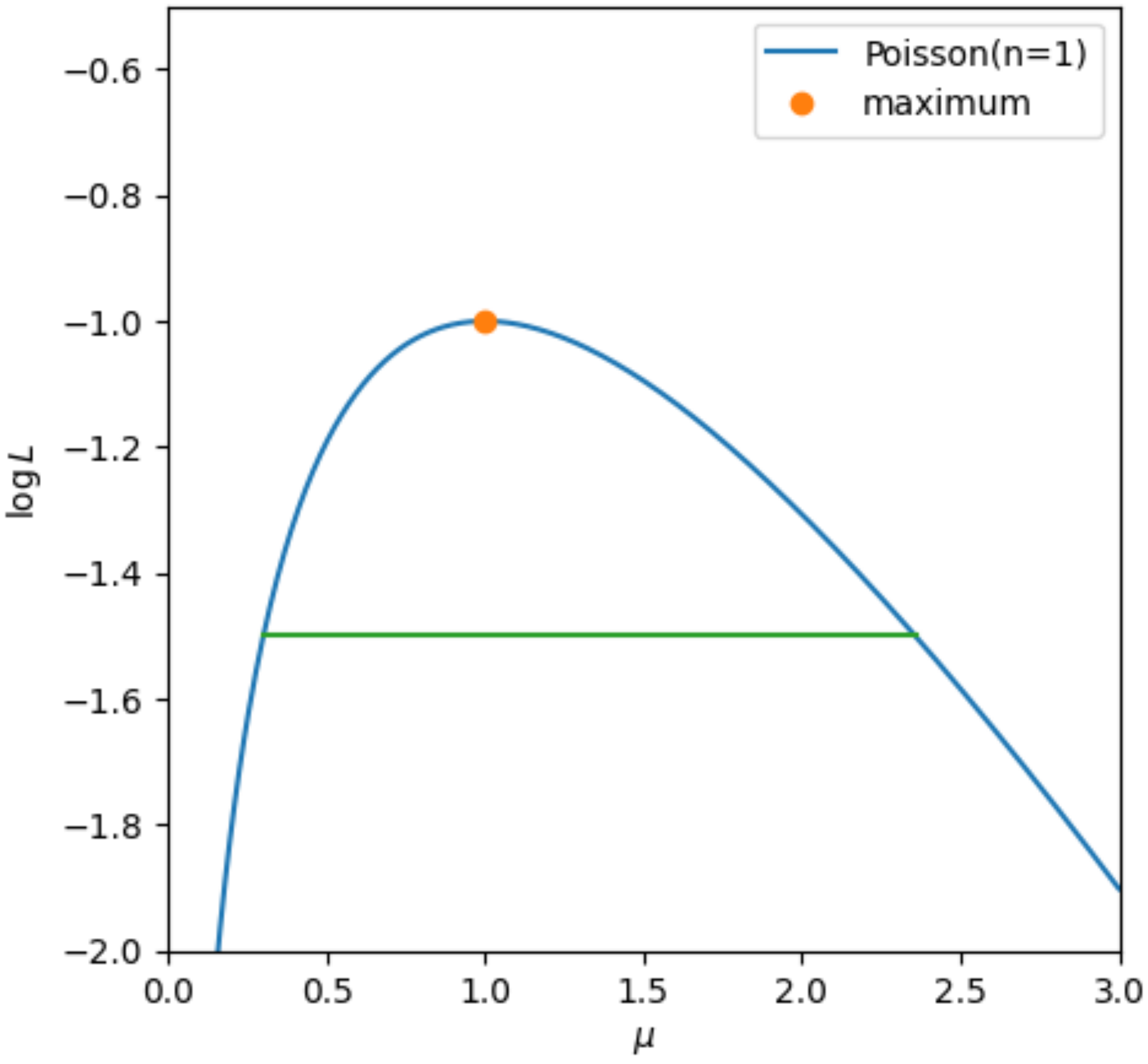
- Typical case in particle physics
  - binned distribution
  - Measurement uncertainty leads to entries landing in the wrong bin
  - Mostly, when uncertainty is larger than the bin width
- Methods for inverting process via “unfolding”
- Discussed later in the lecture

arXiv:0912.0023



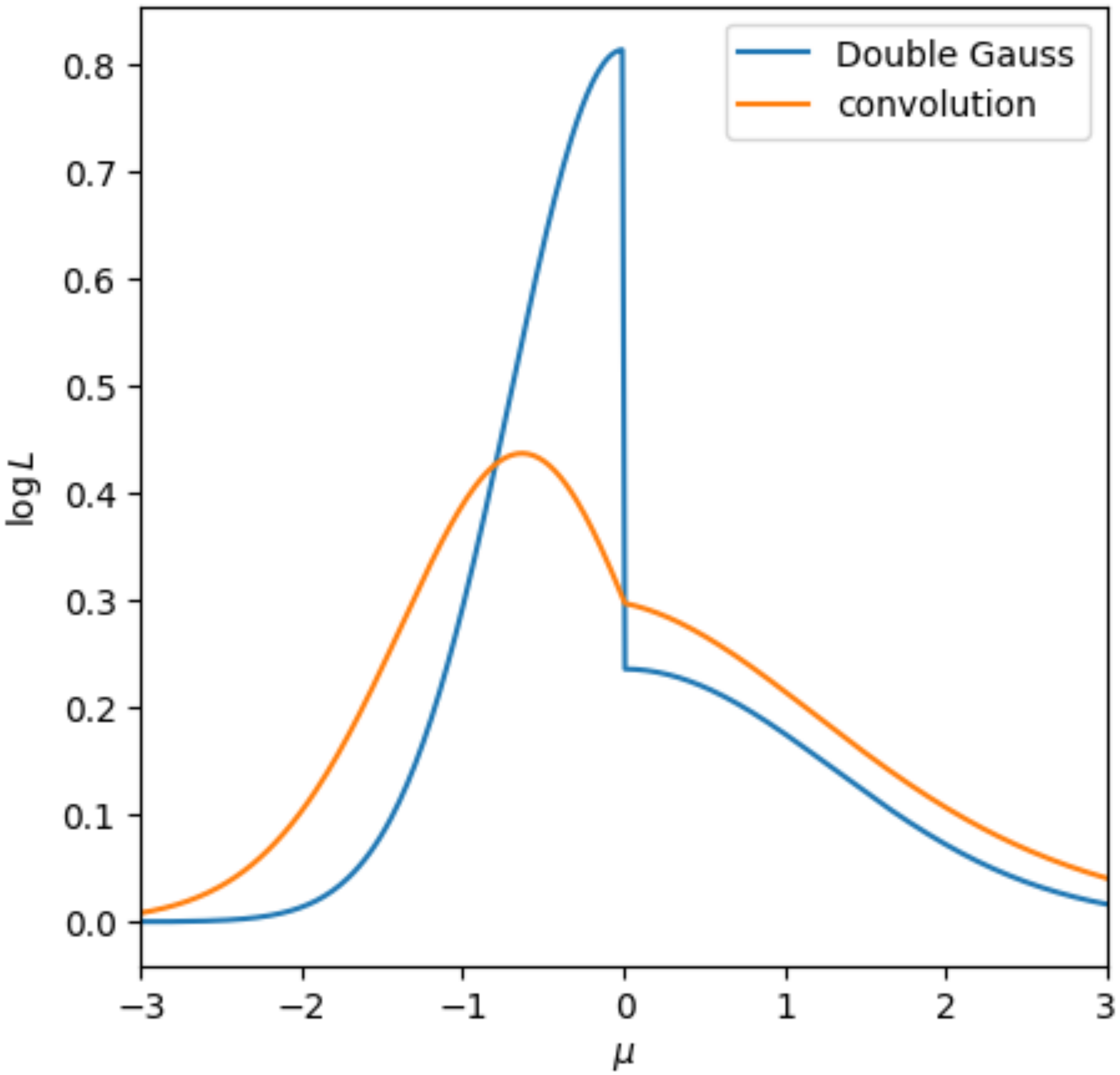
Asymmetric errors

- Sometimes results are given with asymmetric uncertainties $x^{\sigma^+}_{\sigma^-}$
- This is not consistent with the definition of the error as a standard deviation - which does not have a direction
- There are several ways in which an asymmetric error can come about:
  - From the $\log L_{max} - 1/2$ rule of maximum likelihood yielding asymmetric points
  - From using a confidence interval for the edges of the $[x - \sigma^-, x + \sigma^+]$ interval
  - From some other ad hoc rule, which many or may not be explained
- Usually we need some way to combine (average) measurements and to propagate these uncertainties
- The treatment of the results depends on what the asymmetric uncertainties are supposed to represent



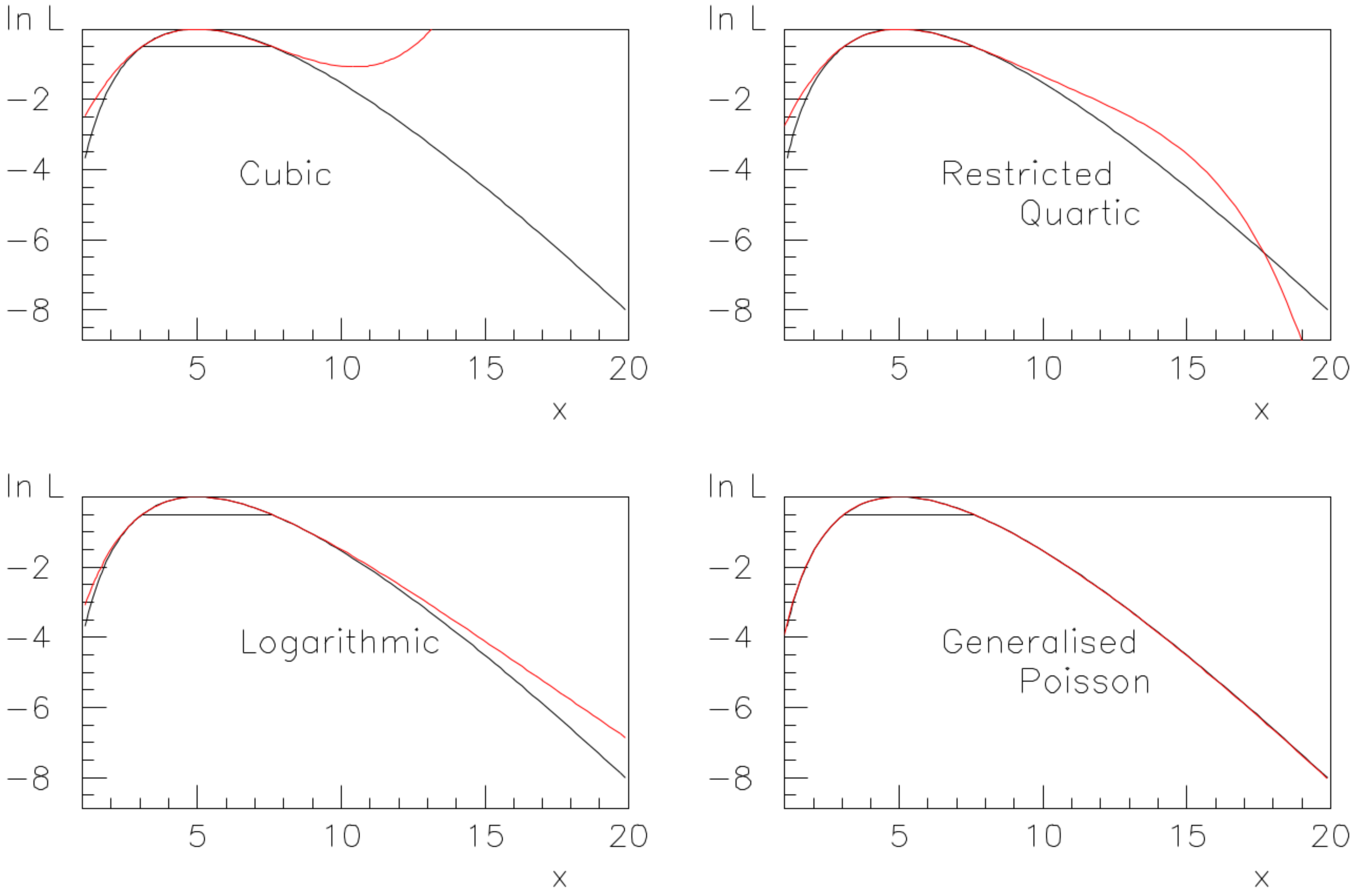
Asymmetric errors (2)

- Sometimes the question arises: Are the errors Gaussian?
- It is unclear what that means
- Assuming: Is the distribution of the estimator Gaussian?
- Then asymmetry could mean: Sum of two half-Gaussian distributions
  $[x > 0] \cdot G_1(x, \sigma^+)/\sigma^+ + [x < 0] \cdot G_2(x, \sigma^-)/\sigma^-$
- Mean at 0 but median below
- Taking the mean of several measurements means convolution
- But convolution must lead closer to symmetric distribution
- Therefore: Whatever the reason for asymmetric error bars, *it is never justified to add the upper and lower errors in quadrature separately!*
- Impossible to calculate χ^2 for goodness-of-fit without specifying what the errors mean
- Similar for weighted mean



Asymmetric errors (3)

- Recommendation for asymmetric uncertainties: Avoid!
- If you have to calculate with other people's errors: Find out precisely what they mean and how they are defined
- R. Barlow looked at two problems:
  - How to combine the $L_{max} - 1/2$ estimates from two measurements
  - How to combine uncertainties with asymmetries from a nonlinear function (arXiv:0306138)
- If no clear definition, then you must use some ad-hoc mechanism, e.g. symmetrising the uncertainties, using the larger of the two … (but don't add them separately)



R. Barlow, PHYSTAT (2005) proceedings
arXiv:0406120

The Guide to Uncertainty in Measurement

- There is a document giving an international standard for evaluating and expressing uncertainty
- Published by the *International Bureau of Weights and Measures, Joint Committee for Guides in Metrology*
- Separates errors into “Type A” and Type B”
  - Type A is estimated via repeated measurements (e.g. from variance of outputs)
  - Type B is everything else
- Essentially adopts a mix of Frequentist and Bayesian methods
  - Type A analysed via unbiased variance estimator
  - "Flat prior” nuisance parameters get $|a-b|/\sqrt{12}$ errors - corresponding to marginal
- It is not clear how much it helps in fundamental physics

BIPM

JCGM 100:2008

GUM 1995 with minor corrections

Evaluation of measurement data — Guide to the expression of uncertainty in measurement

Évaluation des données de mesure — Guide pour l’expression de l’incertitude de mesure

METPΩ XPΩ

GUM, 2008

Hypothesis testing

Reminder: Bayesian and Frequentist diagnosis example

Bayesian:

- Frequency of disease in population is prior
- Probability for *this* patient to have disease is valid concept
    - $p(\mathrm{D}\,|\,+) = 0.032$ is the probability for this patient to have the disease, this encodes the uncertainty

Frequentist:

- Probability for *this* patient to have disease is not a valid concept - there is no random process
- Probability for a patient randomly drawn from the population to have disease is a valid concept
- Two possible statements:
    - "If we randomly select a person from the population, then the people testing positive have a probability of 0.032 of having the disease."
    - "If a patient is healthy, we would get a positive test with a probability of 0.03"
- Neither are probabilities for *this particular person* to have the disease

Can we generalise this approach?

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

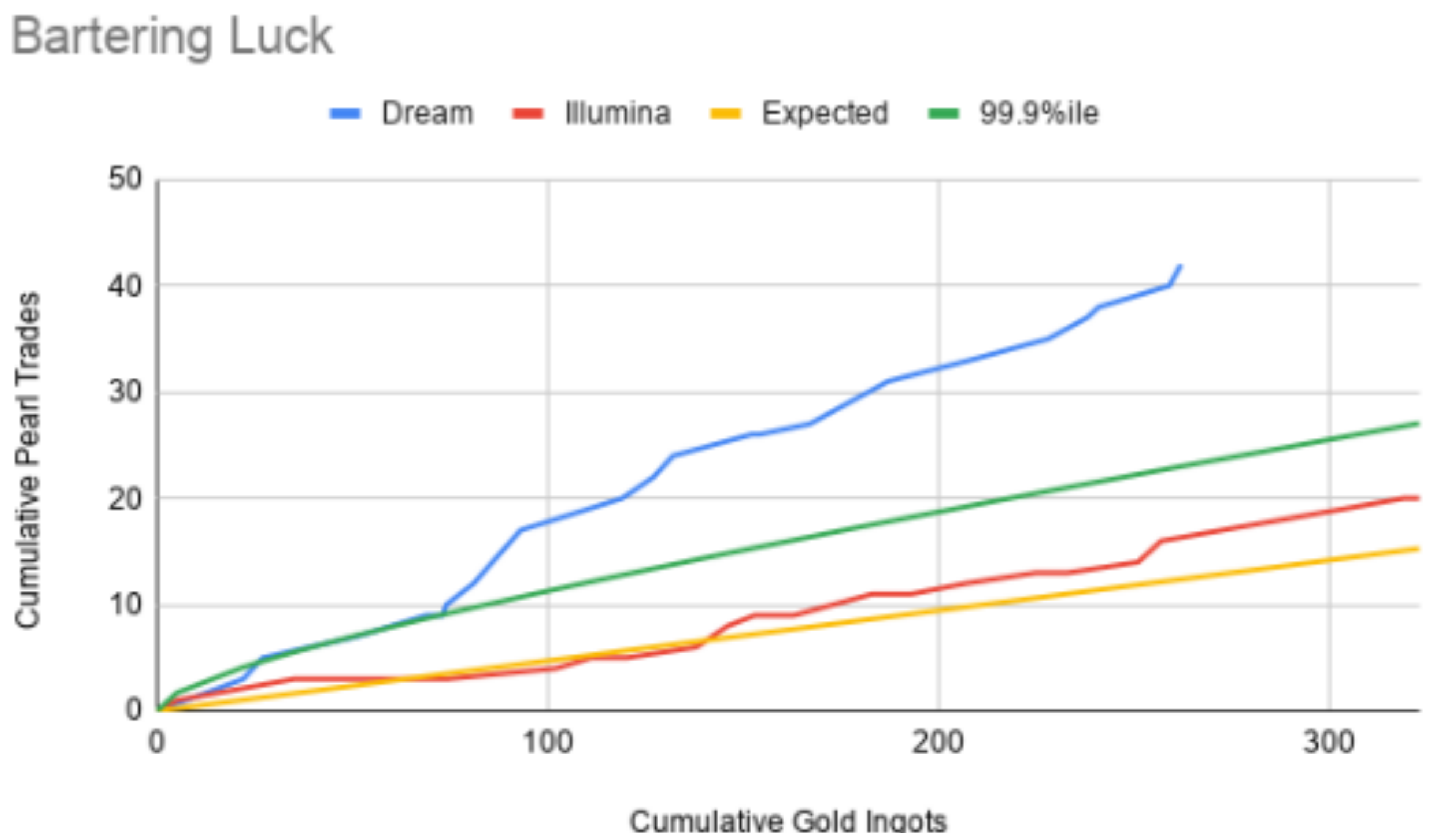
Null hypothesis H_0 and alternative hypothesis H_1

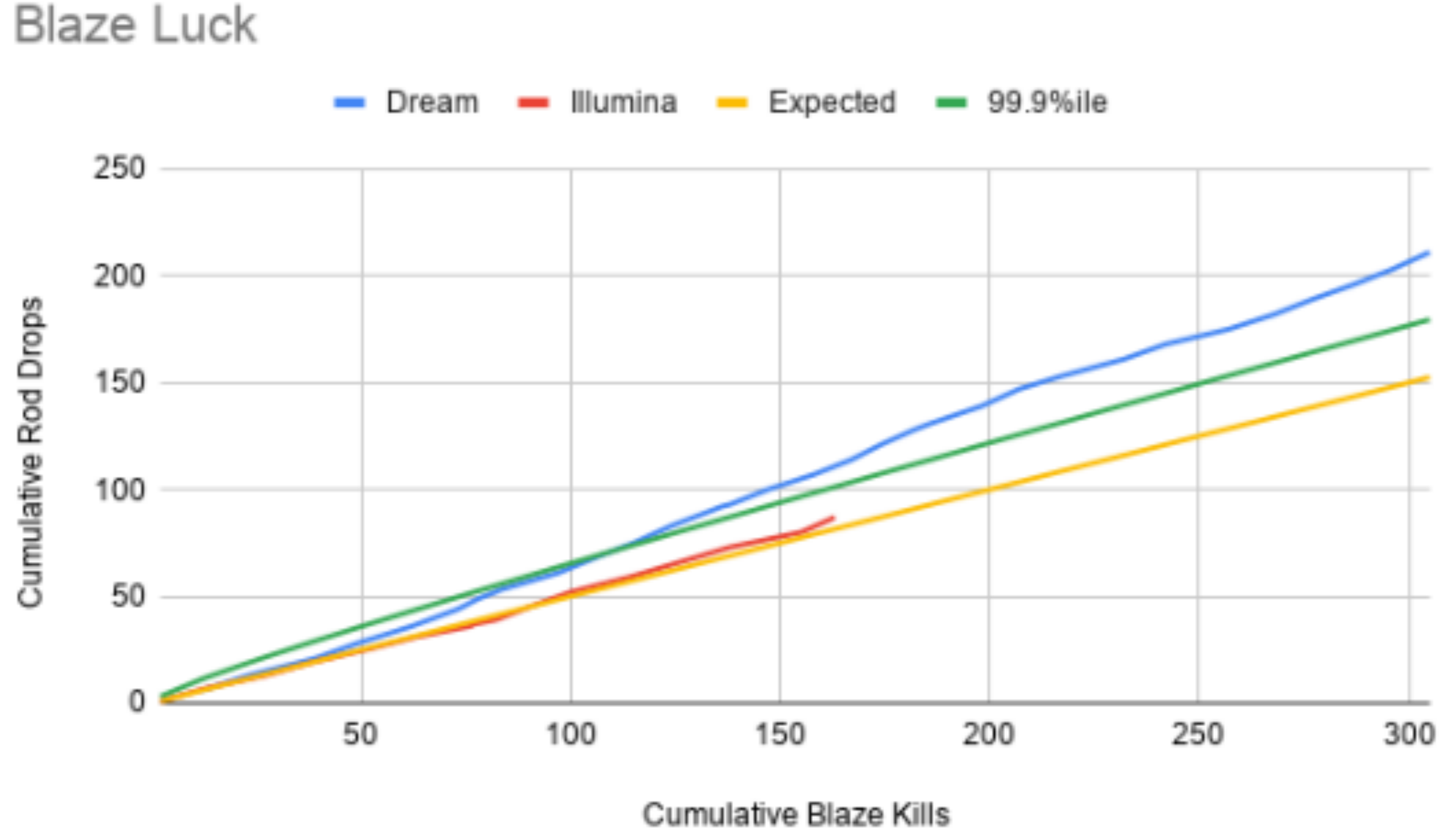
- H_0 often the *background-only hypothesis*
  (e.g. the Standard Model in searches for new physics)
- H_1 often *signal* or *signal + background hypothesis*

Question: Can null hypothesis be rejected by the data?

Example: Dream speedrunning controversy

- In October 2020, Dream reached 5th place in the "glitchless 1.16" Minecraft speedrunning category
- Two main random processes necessary for completing the game: Ender pearls from trading and blaze rods from blazes
- Number of successes should be distributed by a binomial distribution with $\theta_\text{pearls} = 0.05$, $\theta_\text{rods} = 0.5$
- Did he cheat? How to assess from frequentist view?
- Asking: How likely is the result? Does not work
  - $p_\text{binom}(k = 211 \,|\, N = 305, \theta = 0.5) \approx 4.9 \cdot 10^{-12}$ is small
  - But so is the most likely value $p_\text{binom}(k = 152 \,|\, N = 305, \theta = 0.5) \approx 0.046$
- Need another way to quantify

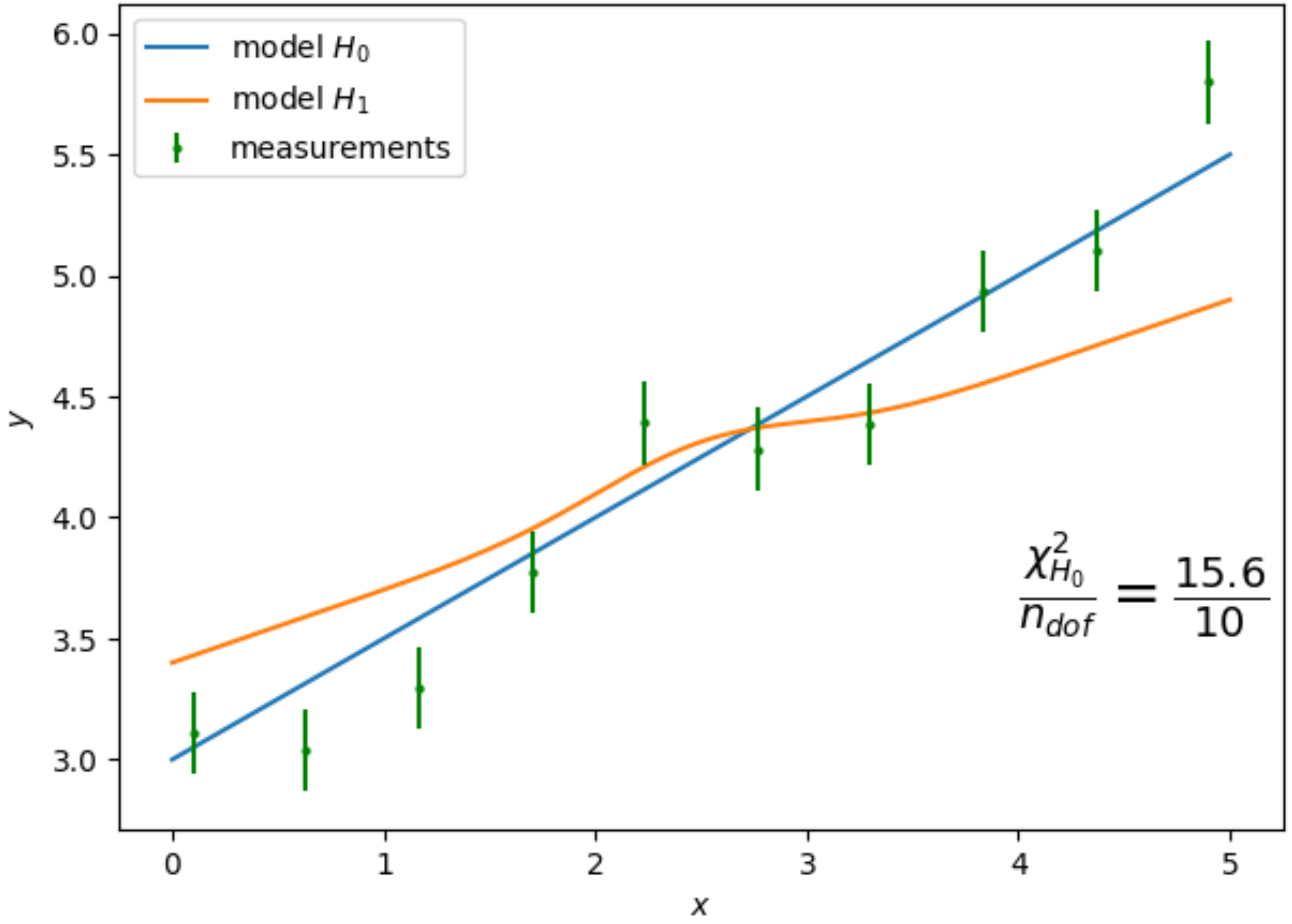




Dream Investigation Results

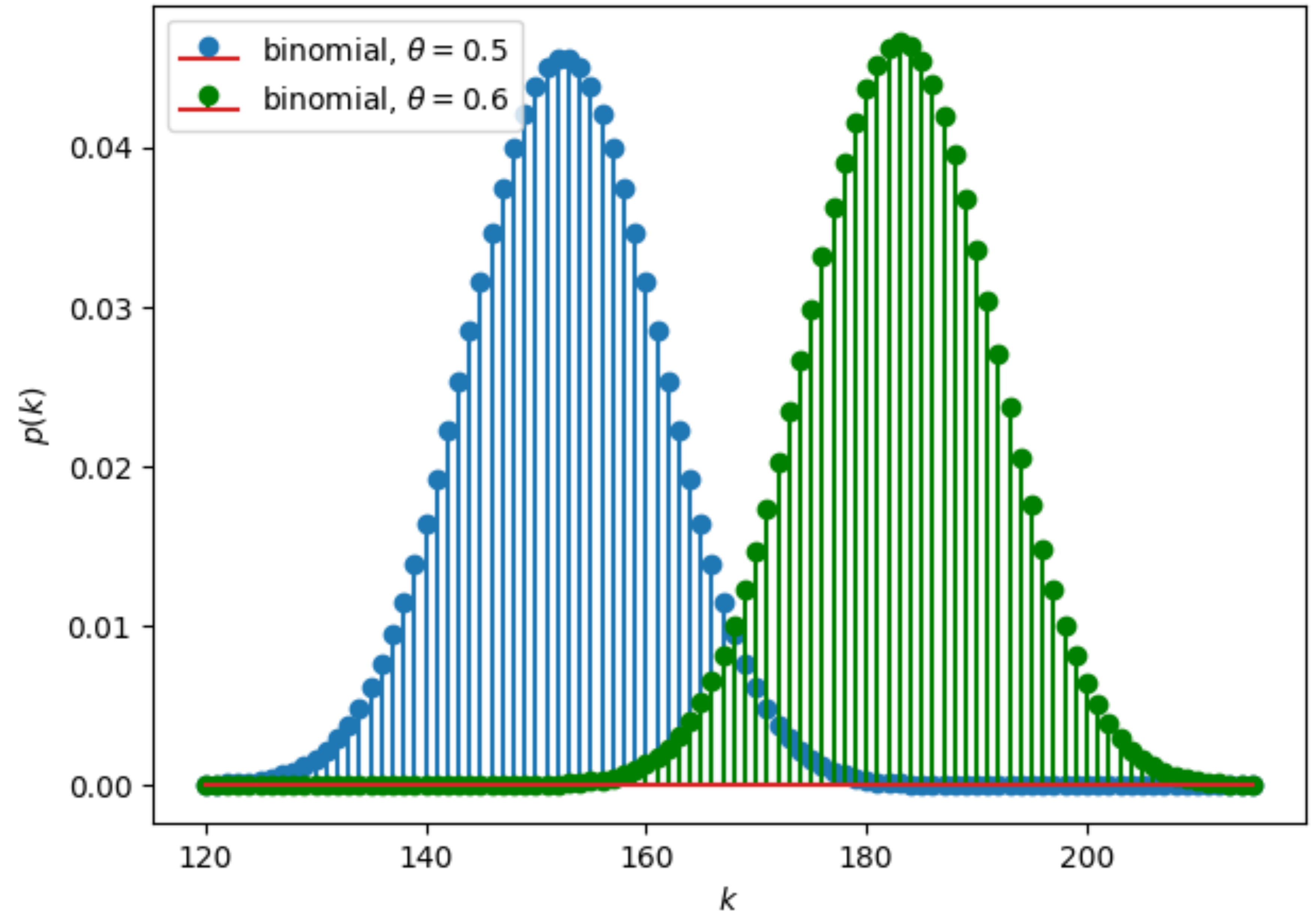
Example 2: Goodness-of-fit

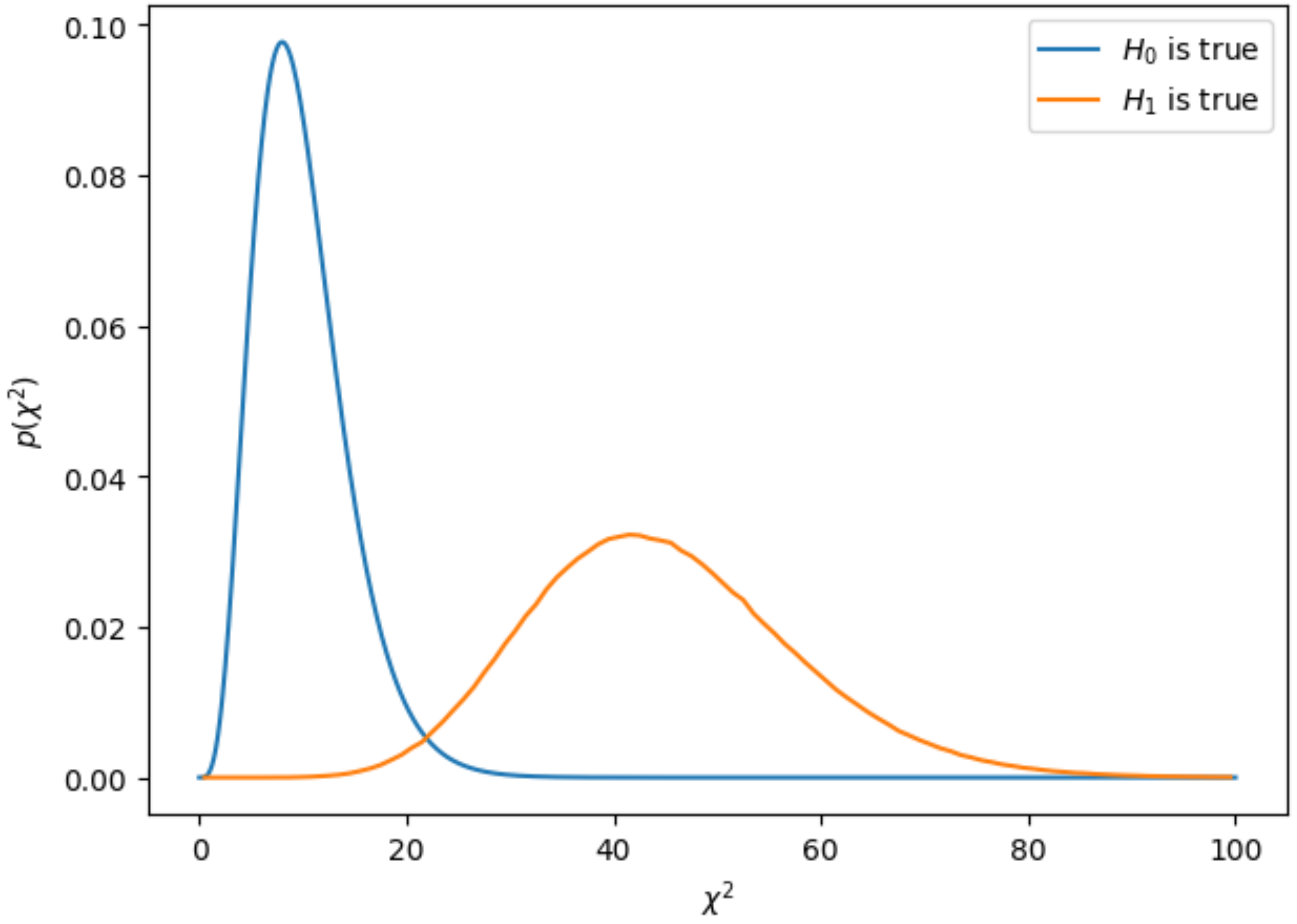
- Want to know if a model (H_0) is consistent with the data
- χ^2 looks "okay", but how to quantify?
- Probability for any particular χ^2 is always 0 (density)
- If e.g. H_1 were the true model, we would likely have a large χ^2 (wrt. the blue line), but this is a qualitative statement



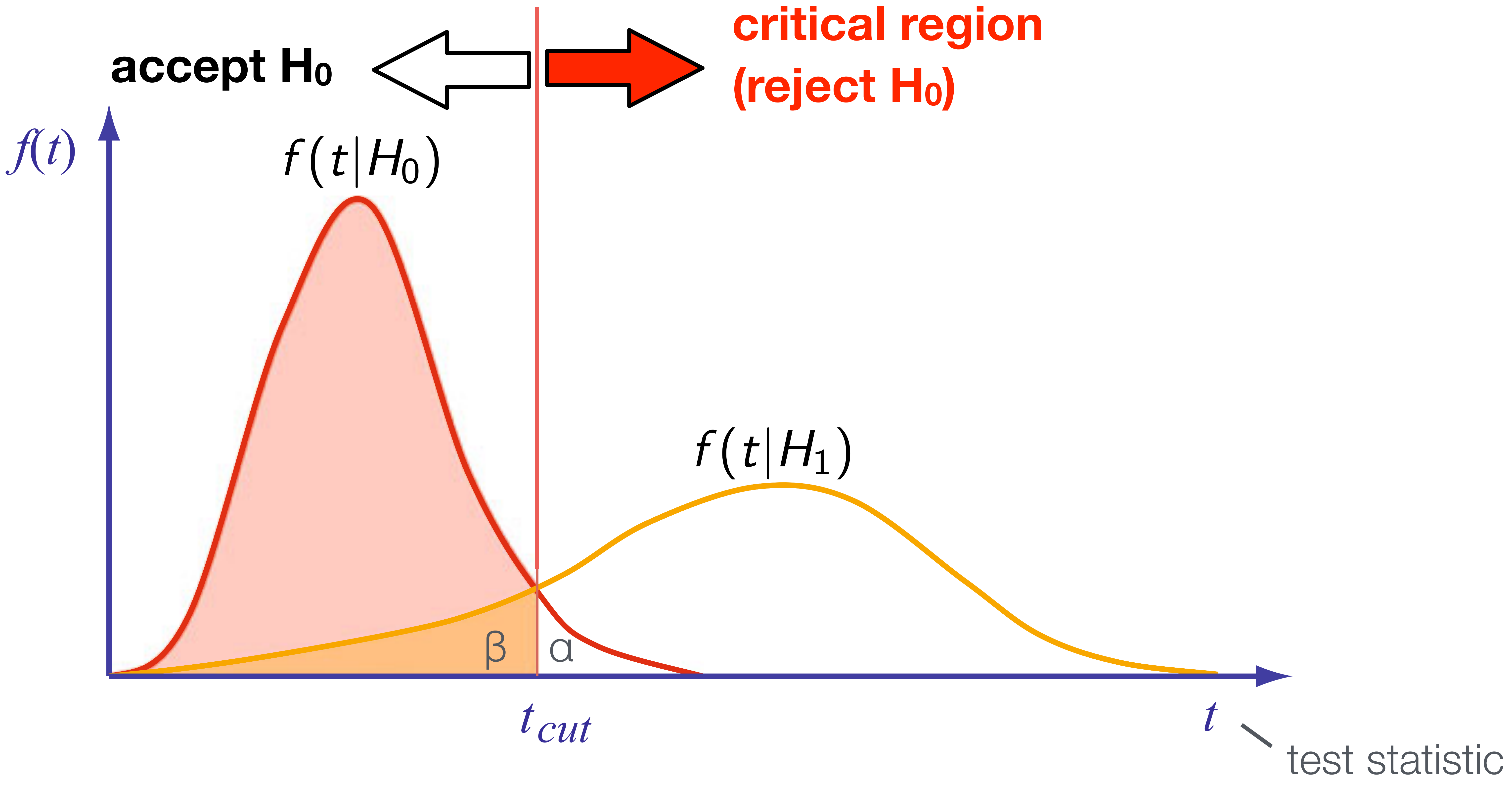
Test statistics

- A *test statistic* $t(\vec{d})$ is a function of the data $\vec{d}$
- It should be chosen, such that for the hypothesis we are testing, H_0, the alternatives H_1 ($H_2, H_3, \dots$) typically have larger or smaller values (we will assume larger in the following)
- For the "Dream" (binomial) case, we can just use the observer number k - a modification to a higher drop probability yields a higher average k
- For the model comparison, we can use the χ^2 as it will be on average higher if another hypothesis is true
- We thus use k and χ^2 as the test statistics here
- The choice of the test statistic determines how well we can distinguish between hypotheses
- We can now decide to reject or accept a hypothesis based on the test statistic





Critical region



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

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

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

Probability to reject H_1 even though it is true:

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

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)

What does such a test mean?

- Remember the two statements from diagnosis:
  - “If we randomly select a person from the population, then the people testing positive have a probability of 0.032 of having the disease.”
  - “If a patient is healthy, we would get a positive test with a probability of 0.03”
- We now define the significance by the second one
- Here, a positive diagnosis would mostly be wrong!
- The Bayesian solution to this would be to use a prior
- The frequentist solution is to require more strict significances, e.g. for particle discoveries $\alpha = 2.9 \cdot 10^{-7}$ also called 5σ.

The p-value - 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
- Larger values should reflect possible alternative hypotheses, but we no not need to specify them
- Determine distribution f(*t*|H_0) under hypothesis H_0

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

The *p-value is the probability of obtaining a test statistic t at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.*

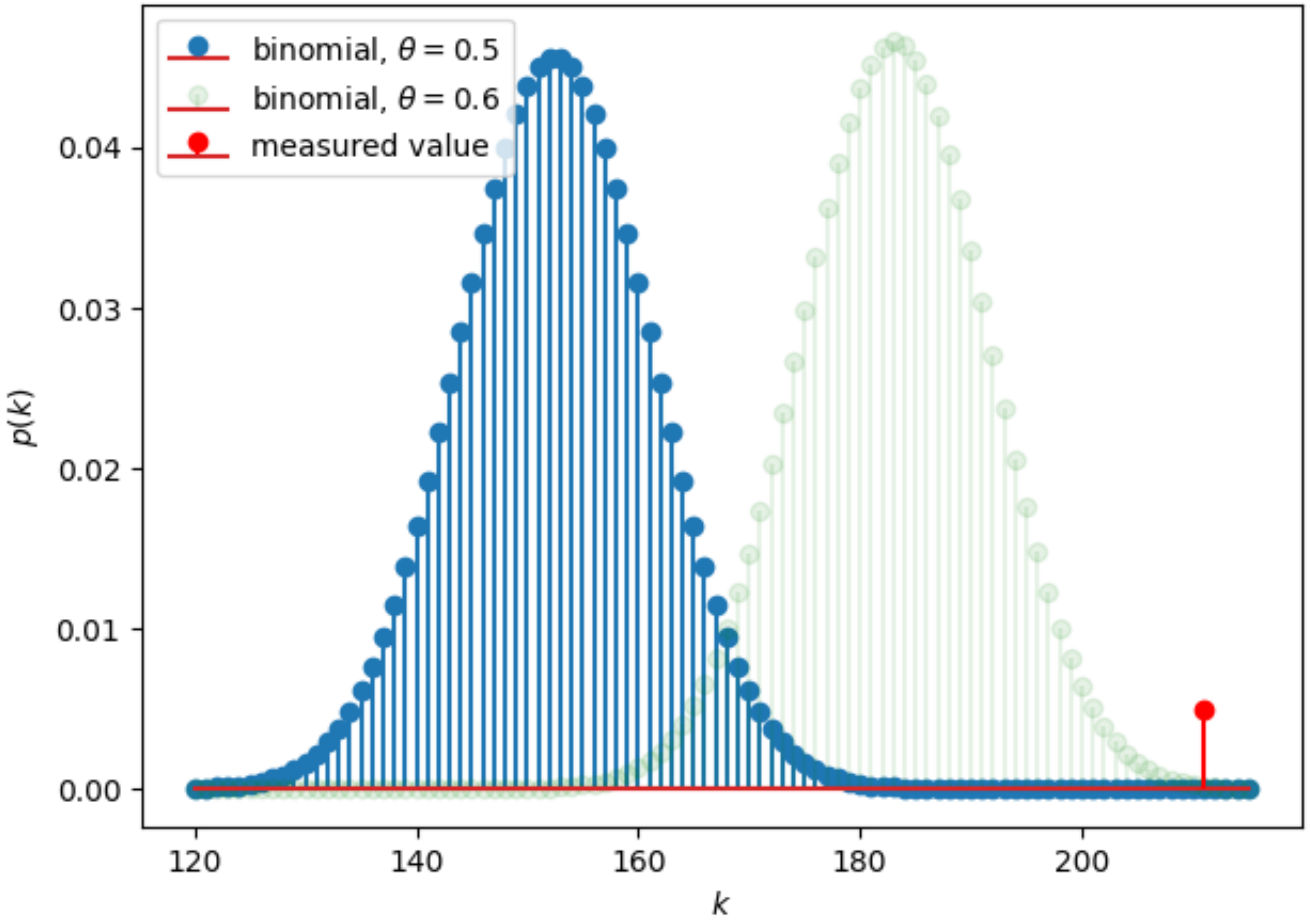
- This means, that the alternative hypothesis is only needed to define what “more extreme” means.

The p-value - example 1

- For the possible cheating: higher numbers of successes k means "more extreme"
- So we need to sum up all cases at least as extreme as the measured one:

$$p\text{-value} = \sum_{k=211}^{305} p_{\text{binom}}(k\,|\,N = 305, \theta = 0.5) = 8.8 \cdot 10^{-12}$$

- Even though this result does not depend on any H_1, we are implicitly comparing to all hypotheses with $\theta > 0.5$



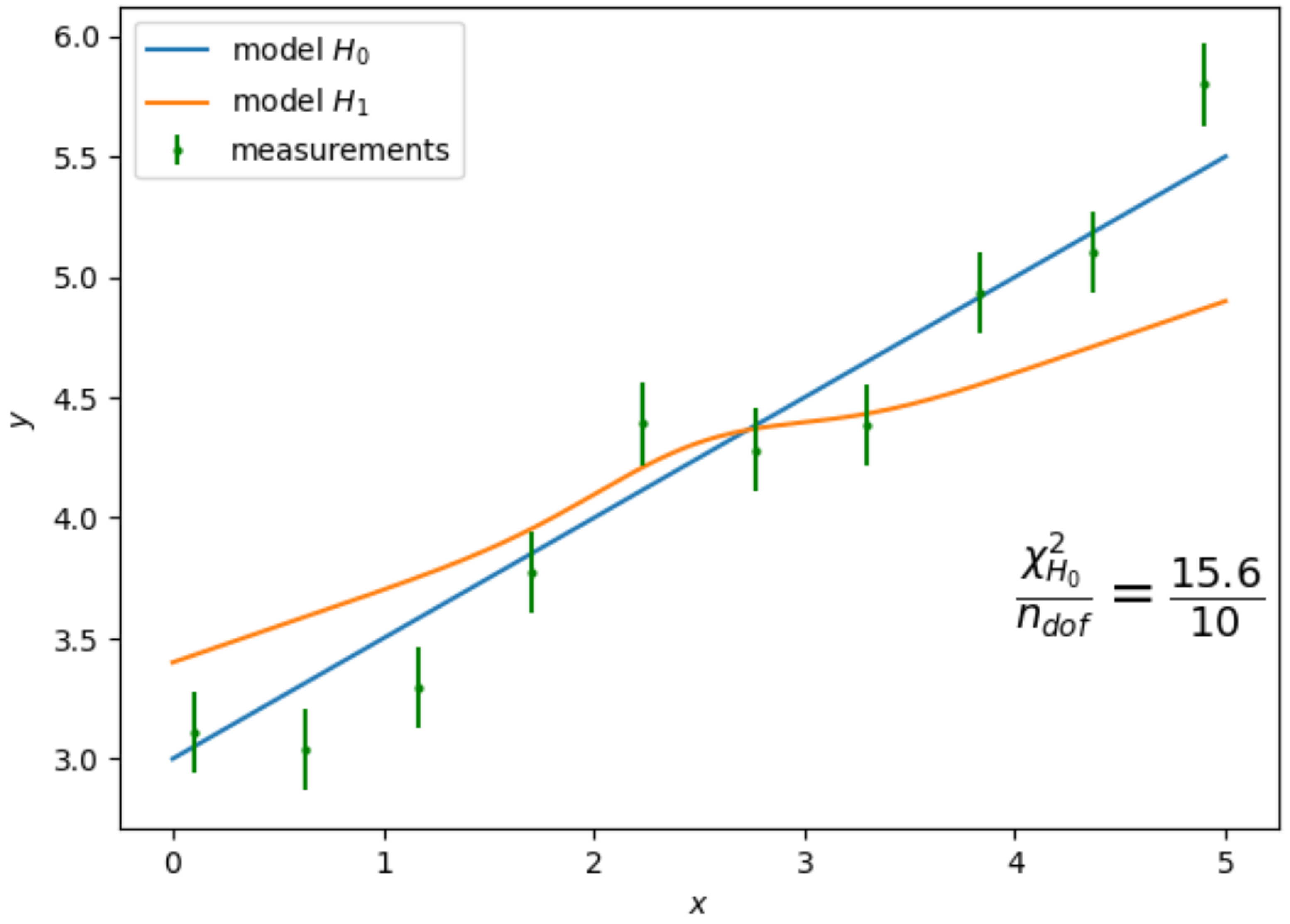
- *p*-value should not be confused with significance level
  - significance level is a pre-specified constant
  - *p*-value is a function of the data, and is therefore itself a random variable
- *p*-value is not the probability for the hypothesis; in frequentist statistics, this is not defined

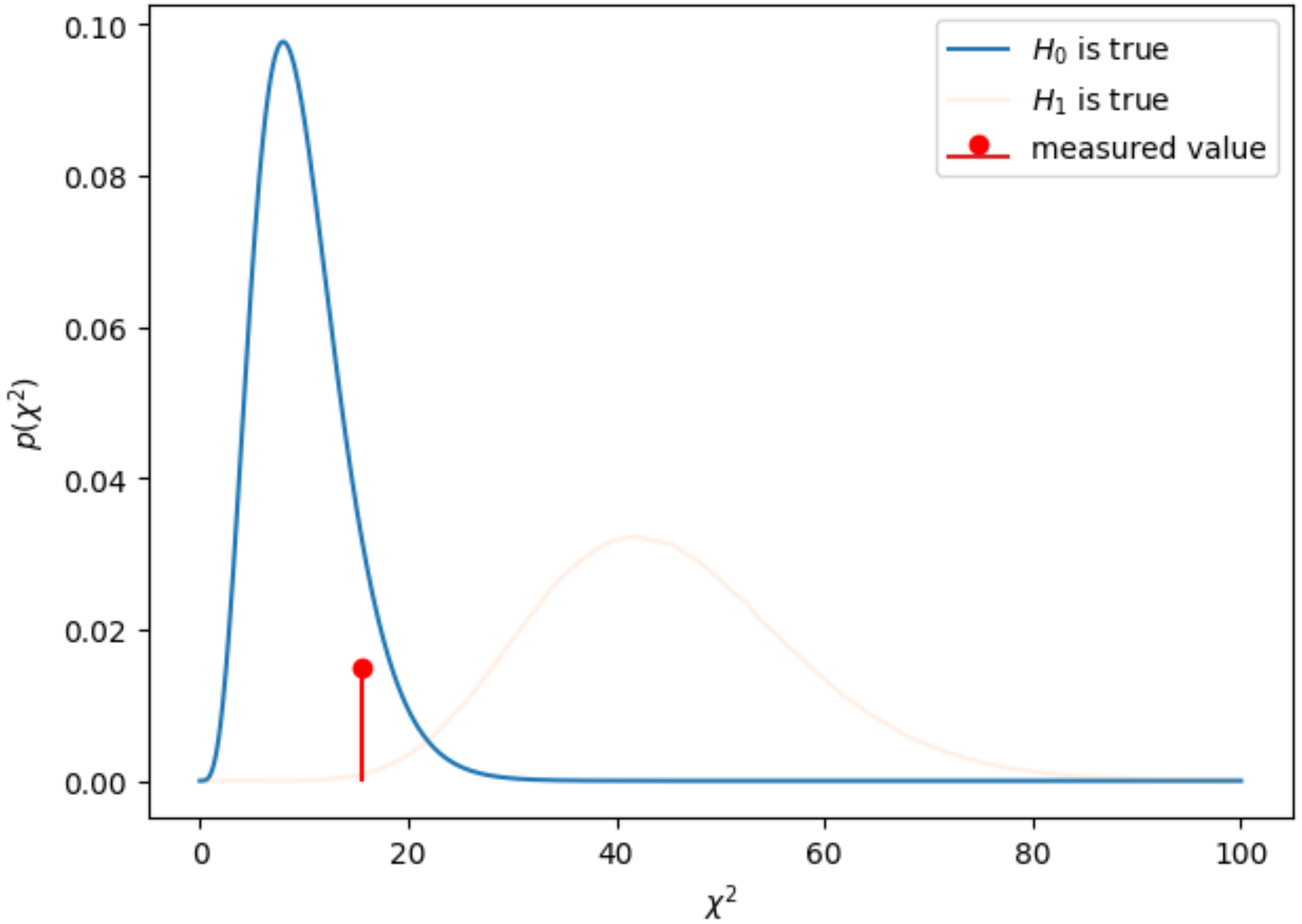
The p-value: example 2

- For the measured $\chi^2 = 15.6$, the more extreme deviation would be towards higher values

$$p\text{-value} = \int_{15.6}^{\infty} p_{\chi^2}(\chi^2 \,|\, N_{dof} = 10) = 0.11$$

- Last point gives large contribution to χ^2
- But it actually disfavours H_1 even more
- This shows, that the χ^2 is not the best possible test statistic here
- When someone quotes a significance, always ask: With respect to which test statistic?





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

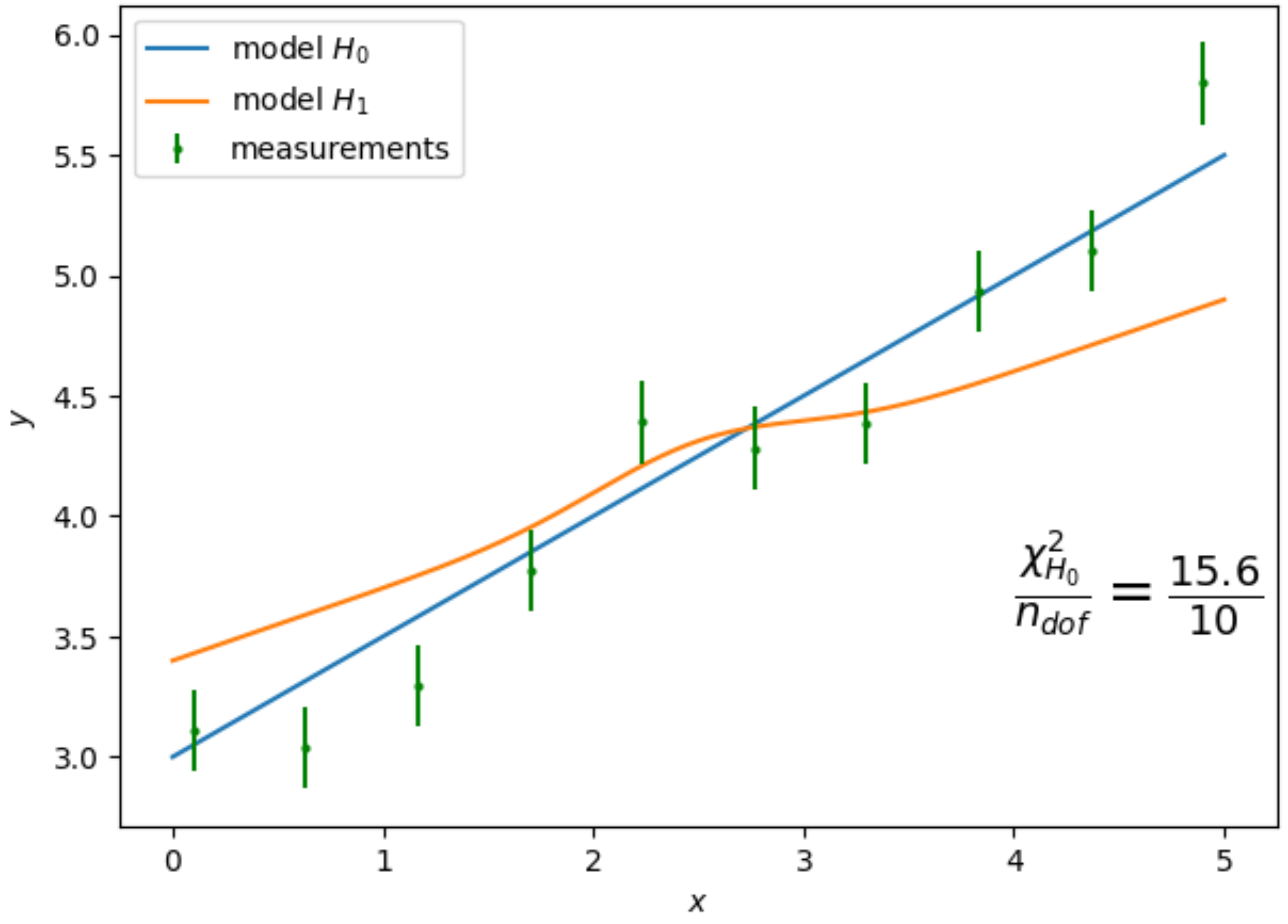
$$t(\vec{x}) = \frac{f(\vec{x}|H_1)}{f(\vec{x}|H_0)}$$

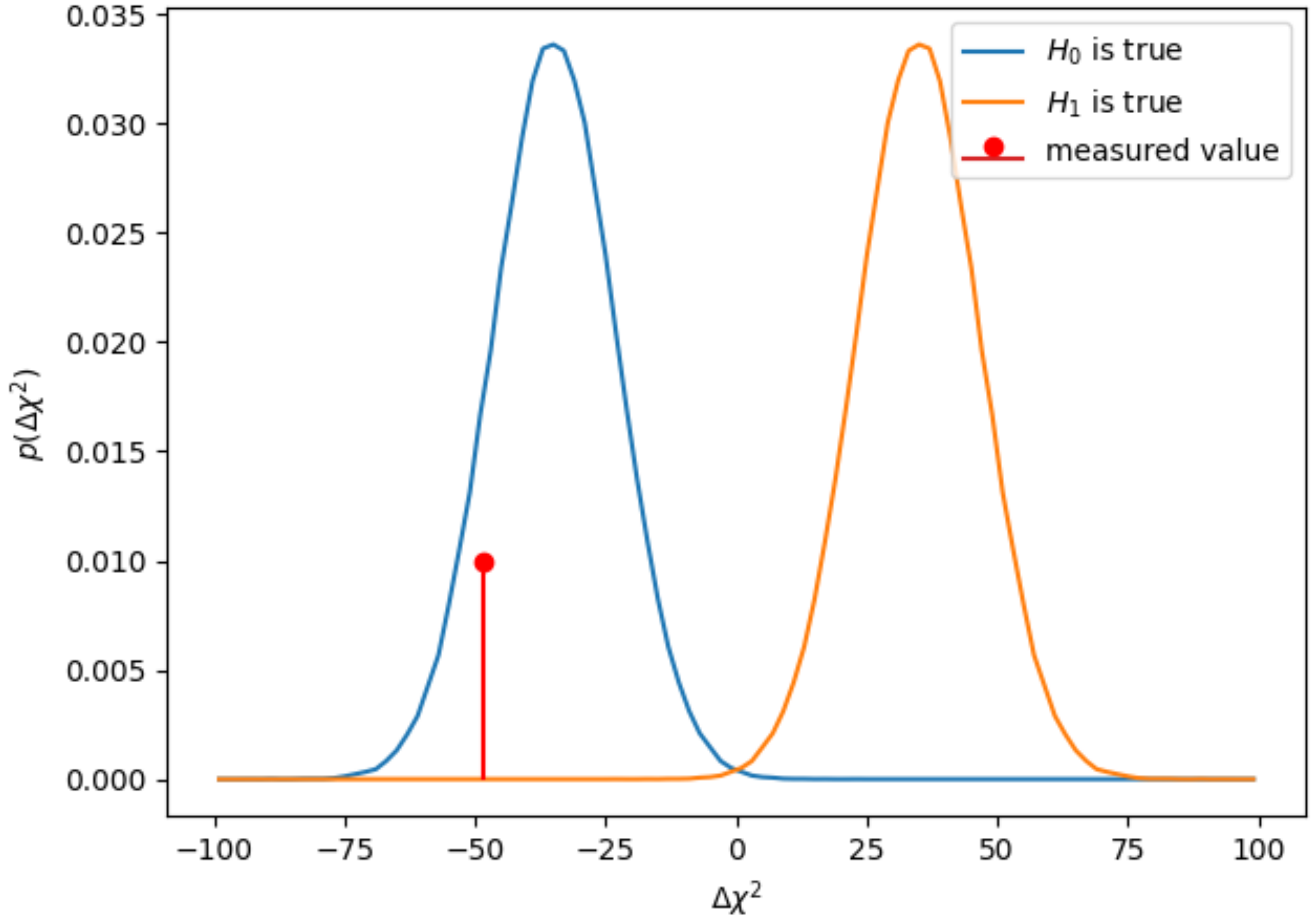
The p-value: example 2

- The χ^2 is the log-L of a multidimensional Gaussian
- Thus, the log of the likelihood ratio leads to a difference in χ^2 for the models with $\Delta\chi^2 = -48.5$

$$p\text{-value} = \int_{-48.5}^{\infty} p(\Delta\chi^2)\,\mathrm{d}\Delta\chi^2 = 0.82$$

- Models are much better separated
- However, we need to specify the alternative model





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
- …

Simple example: Counting experiment (Poisson statistics)

Expected background events:
$\nu_b = 1.3$

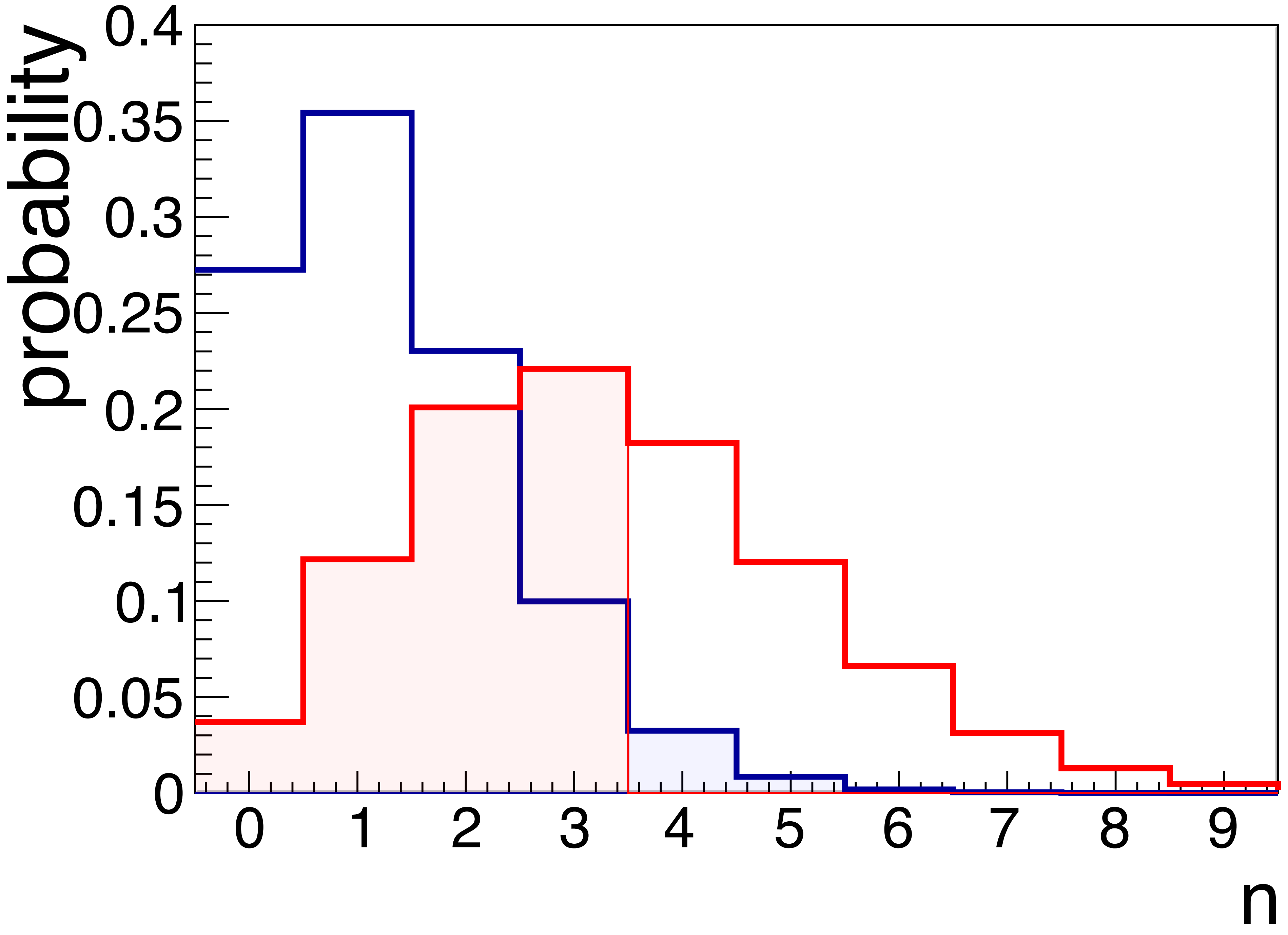
Expected signal events:
$\nu_s = 2$

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**

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

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

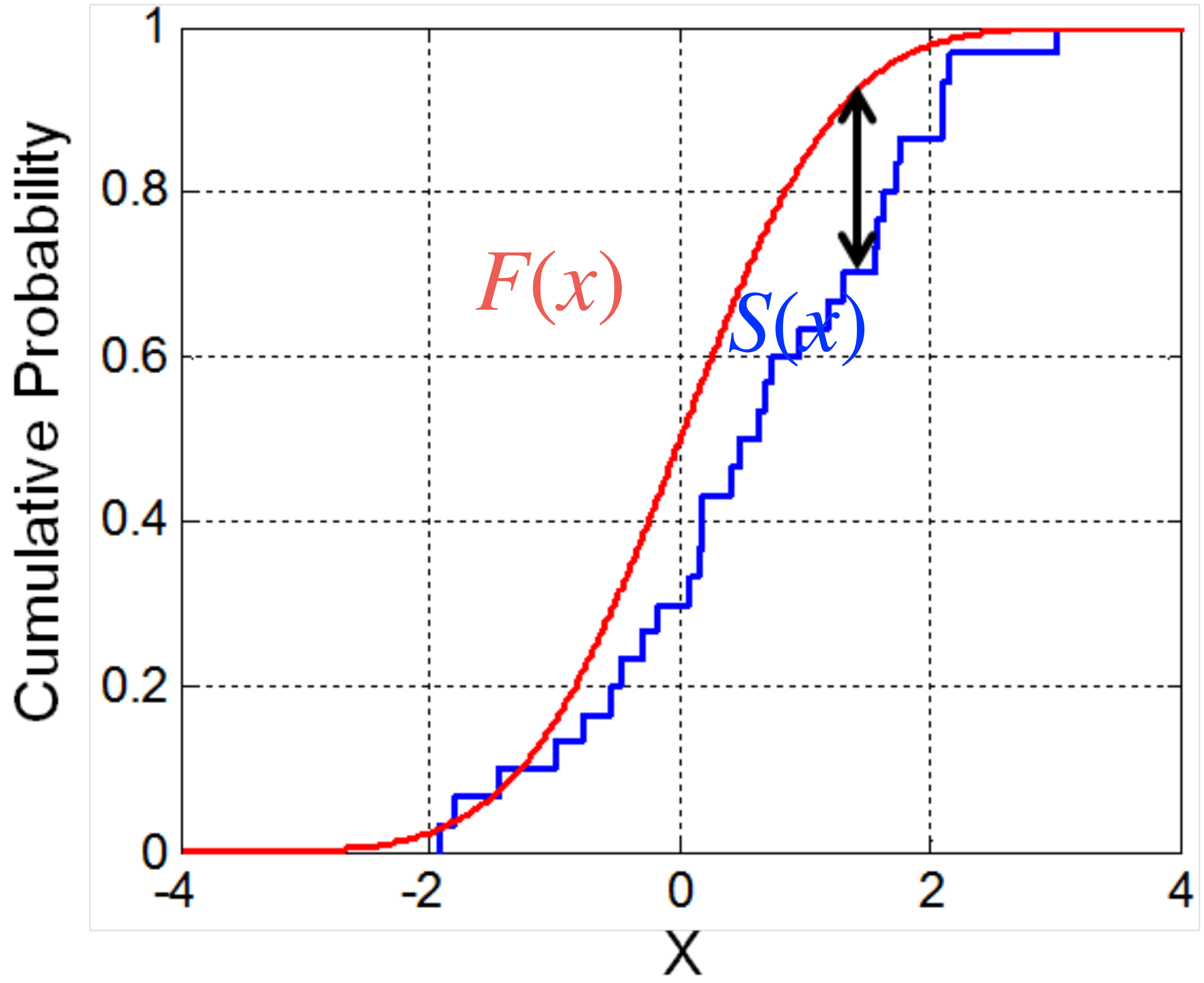
with the so-called Empirical Distribution Function (EDF)

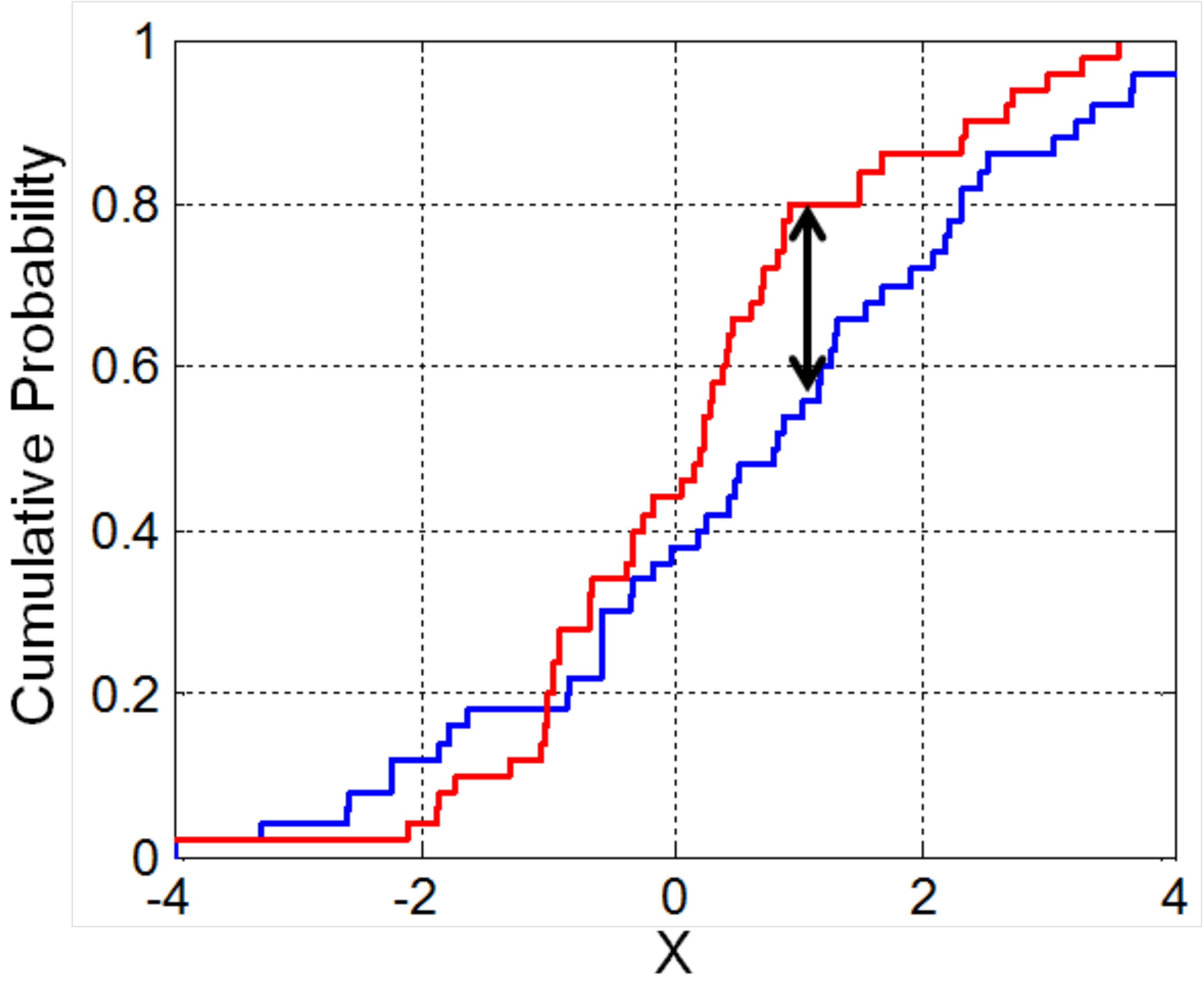
$$S(x) = \frac{\text{number of observations with } x_i < x}{\text{total number of observations}}$$

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.

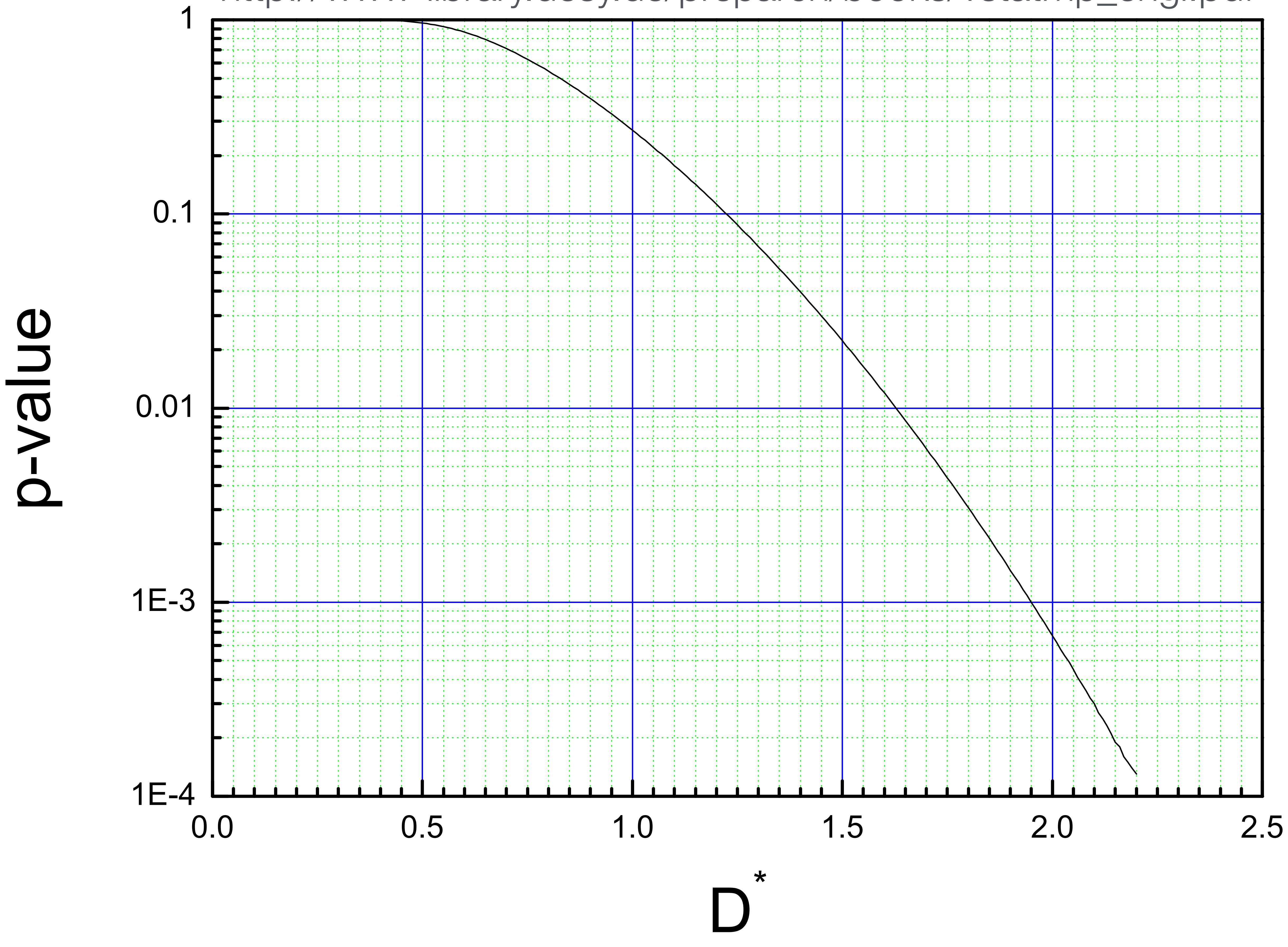




Kolmogorov–Smirnov Test (2)

Expected distribution of *D* known (Kolmogorov distribution) 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:

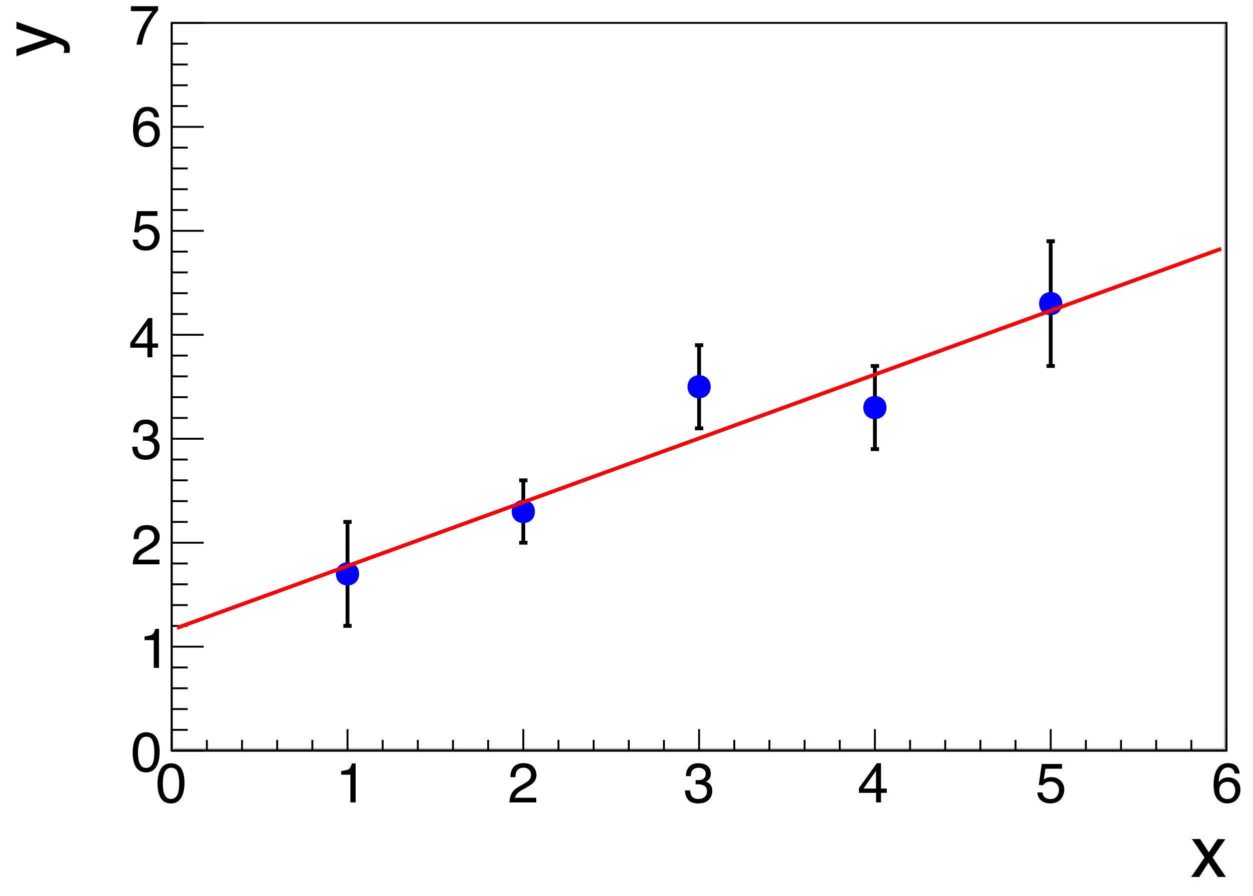
$$Z = \frac{r - \mu}{\sigma}$$

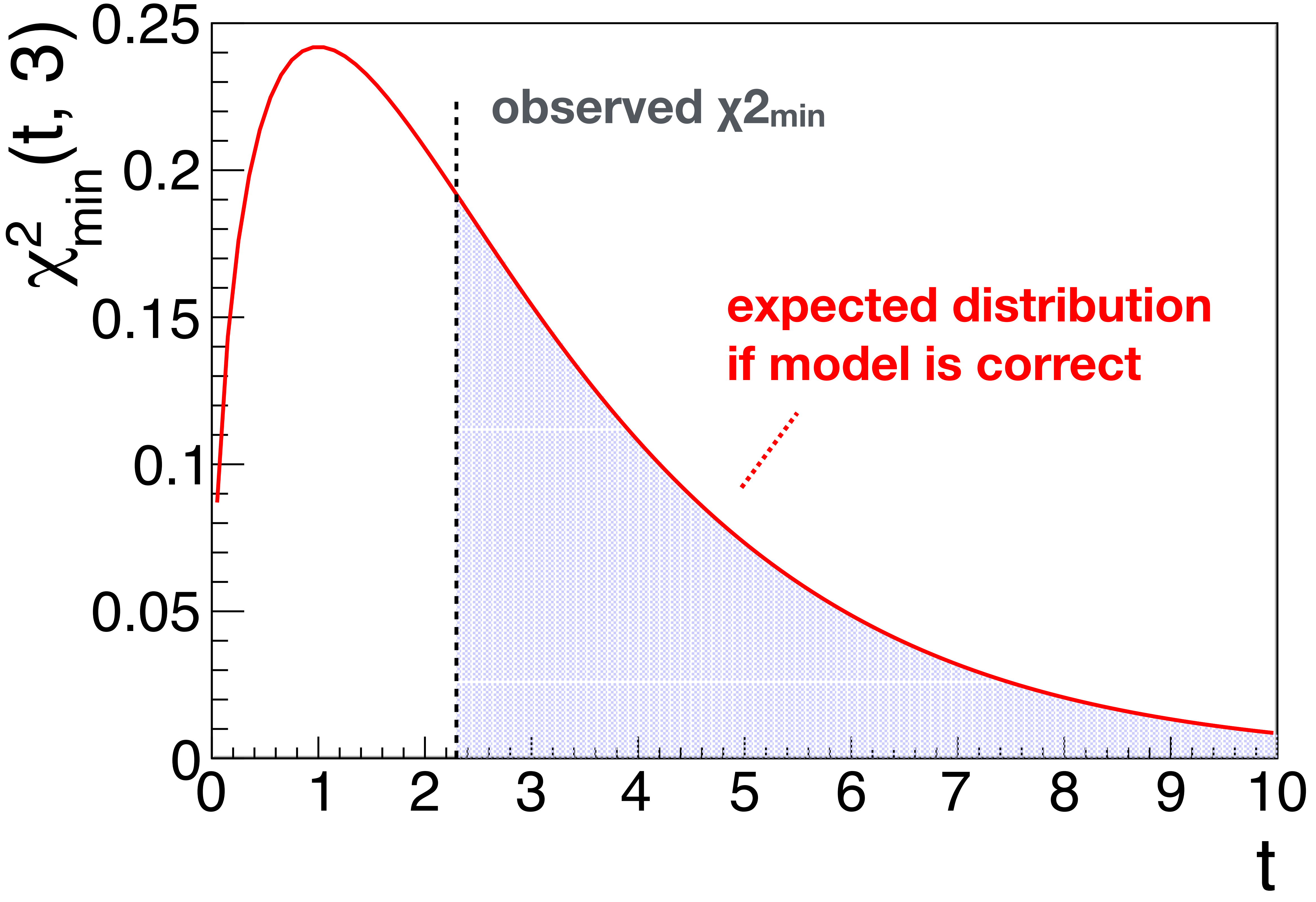
Run test is complementary to the χ^2 square test (can be done in addition)

p-value for a straight line fit

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

p-value = 0.51337

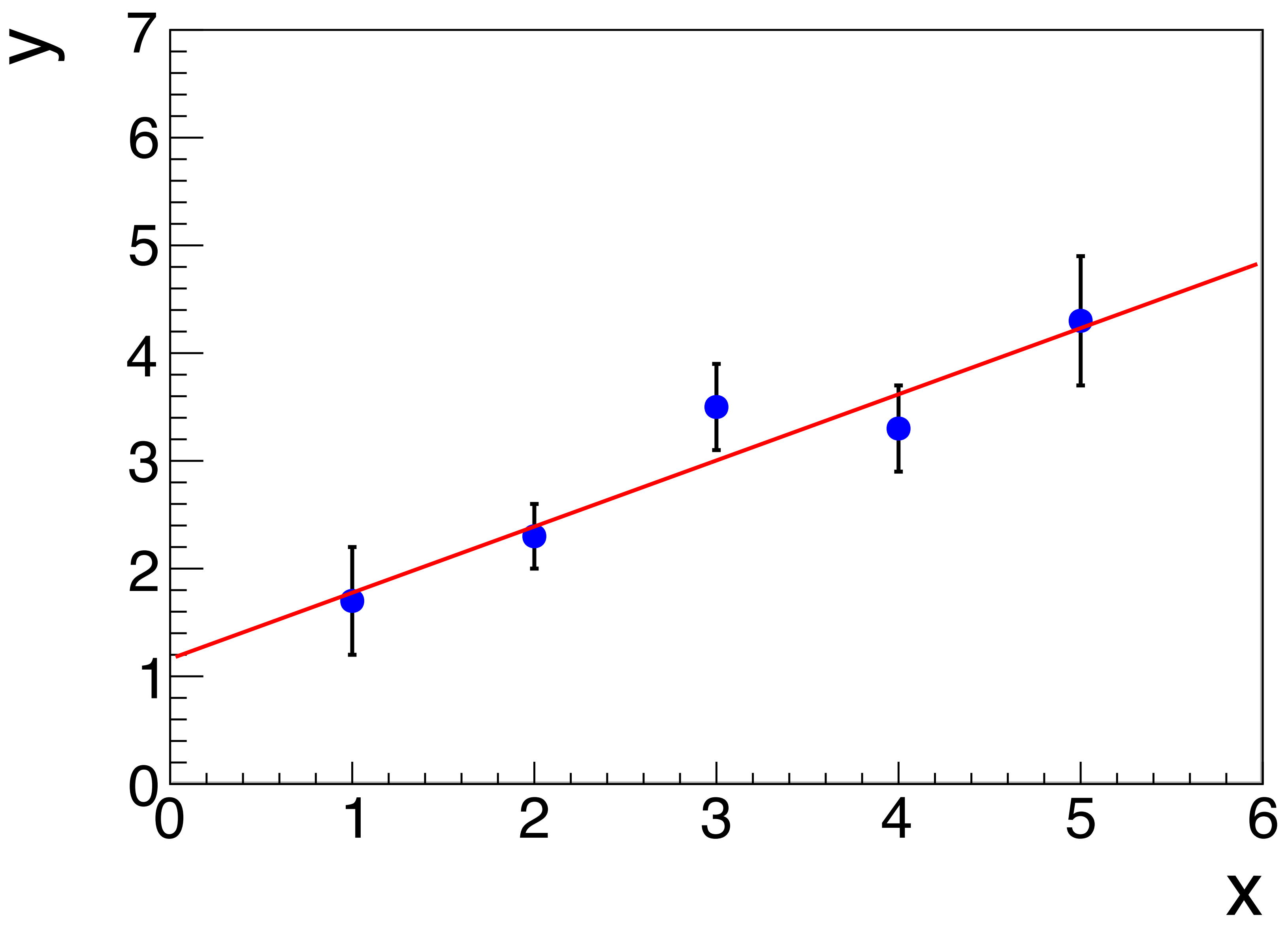




observed $\chi 2_{\text{min}}$

expected distribution if model is correct

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

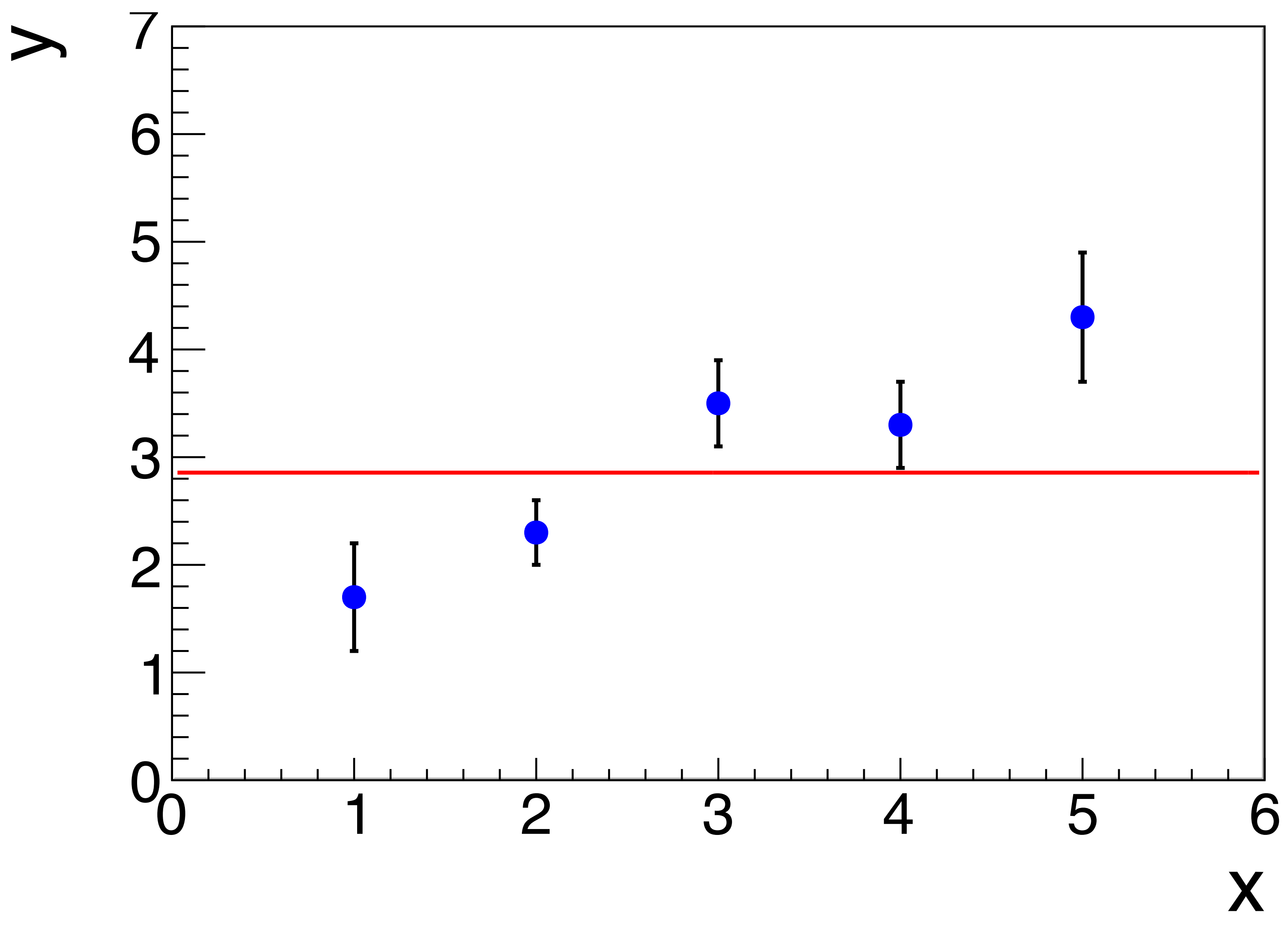


$\chi^2_{\mathrm{min}} = 2.29557$, $n_{\mathrm{df}} = 3$:

p-value = 0.51337

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

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



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

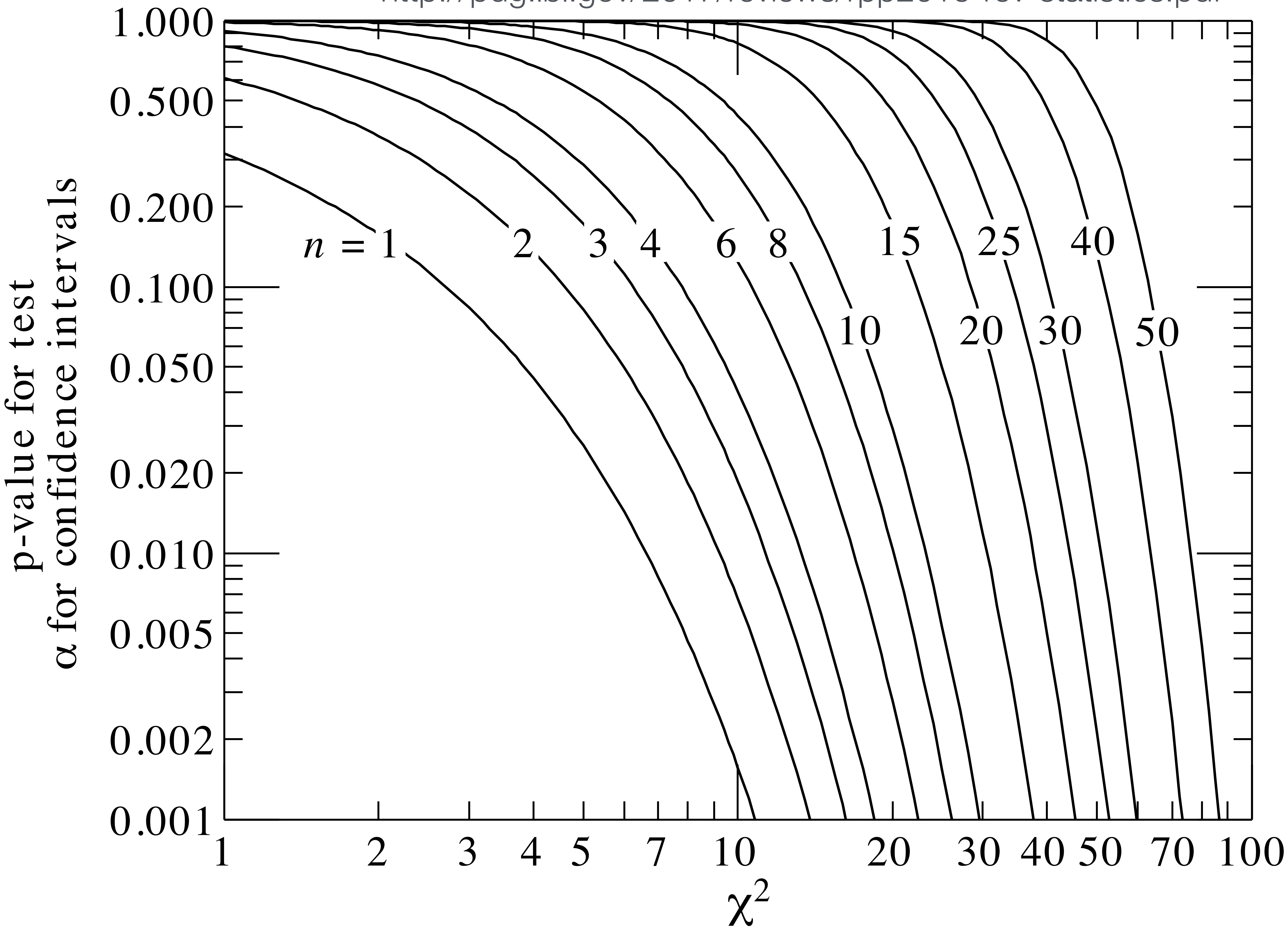
p-value = 0.001032

$\theta_0 = 2.86 \pm 0.18$

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



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:

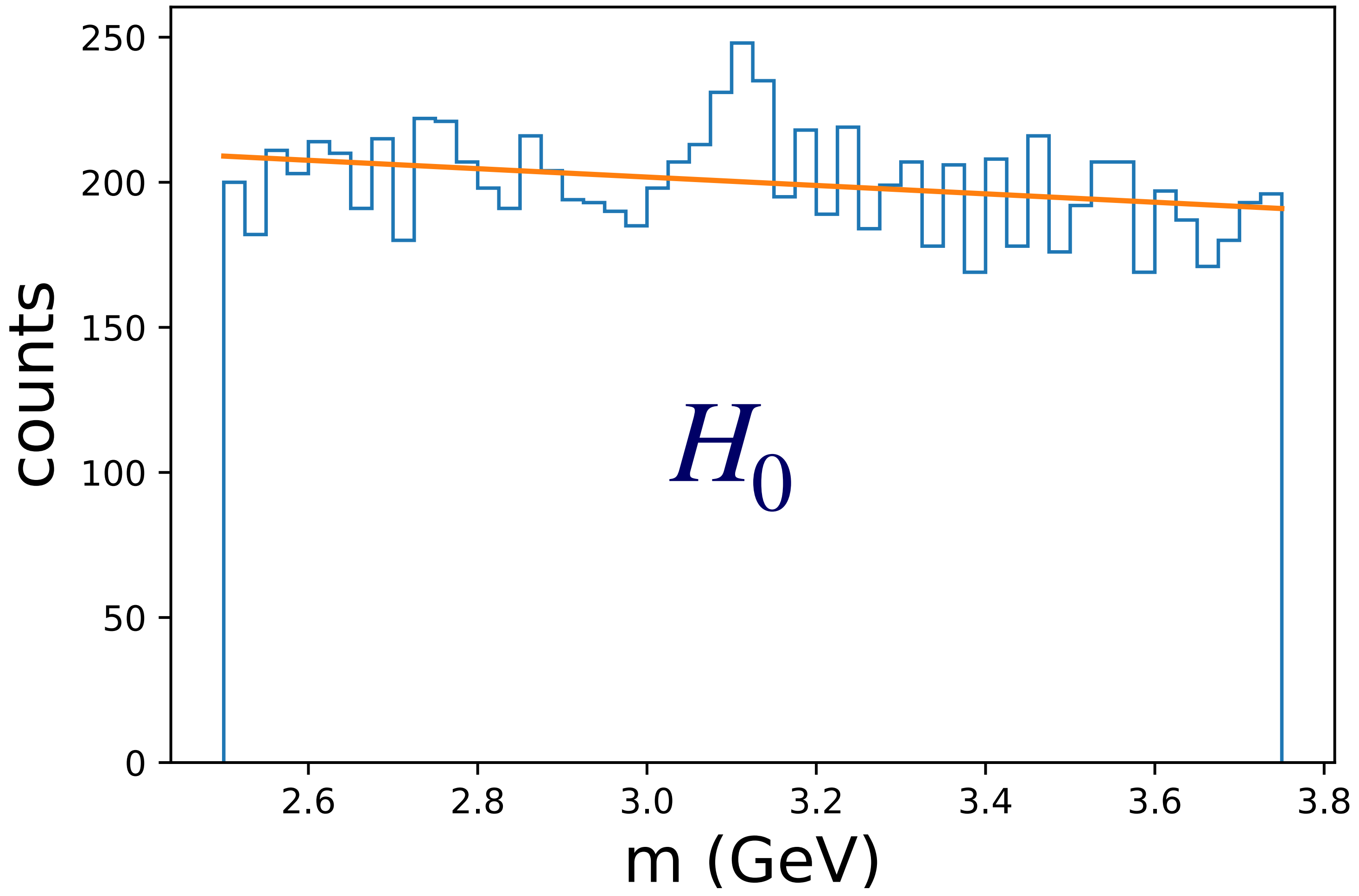
$$\Delta\tilde{\chi}^2 := 2 \ln \left(\frac{L(H_1)}{L(H_0)} \right)$$

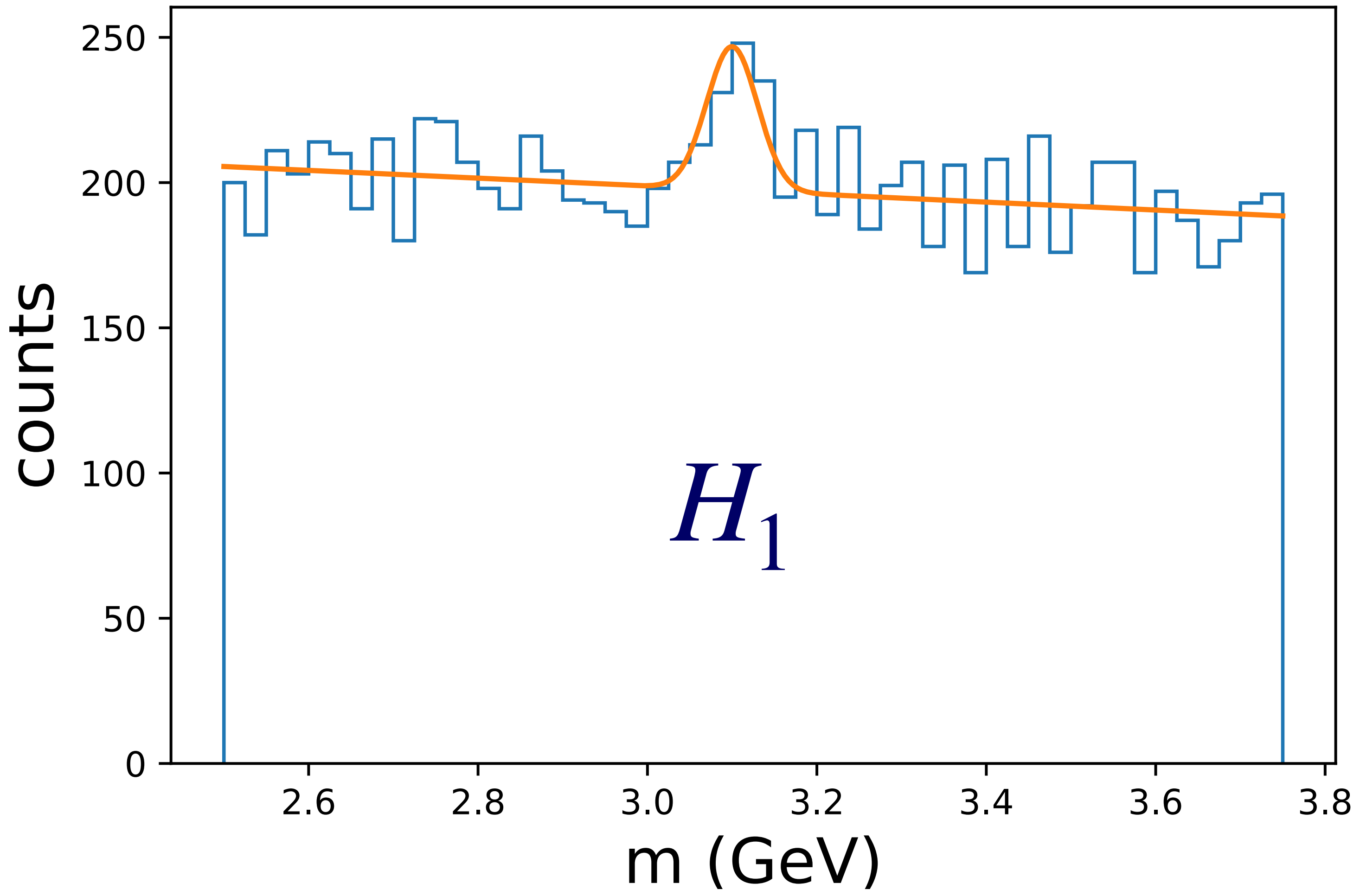
Wilks’ theorem:
If H_0 is correct then $\Delta\tilde{\chi}^2$ follows χ^2 distribution with n_{dof} = #added parameters in the large sample limit.

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

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

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

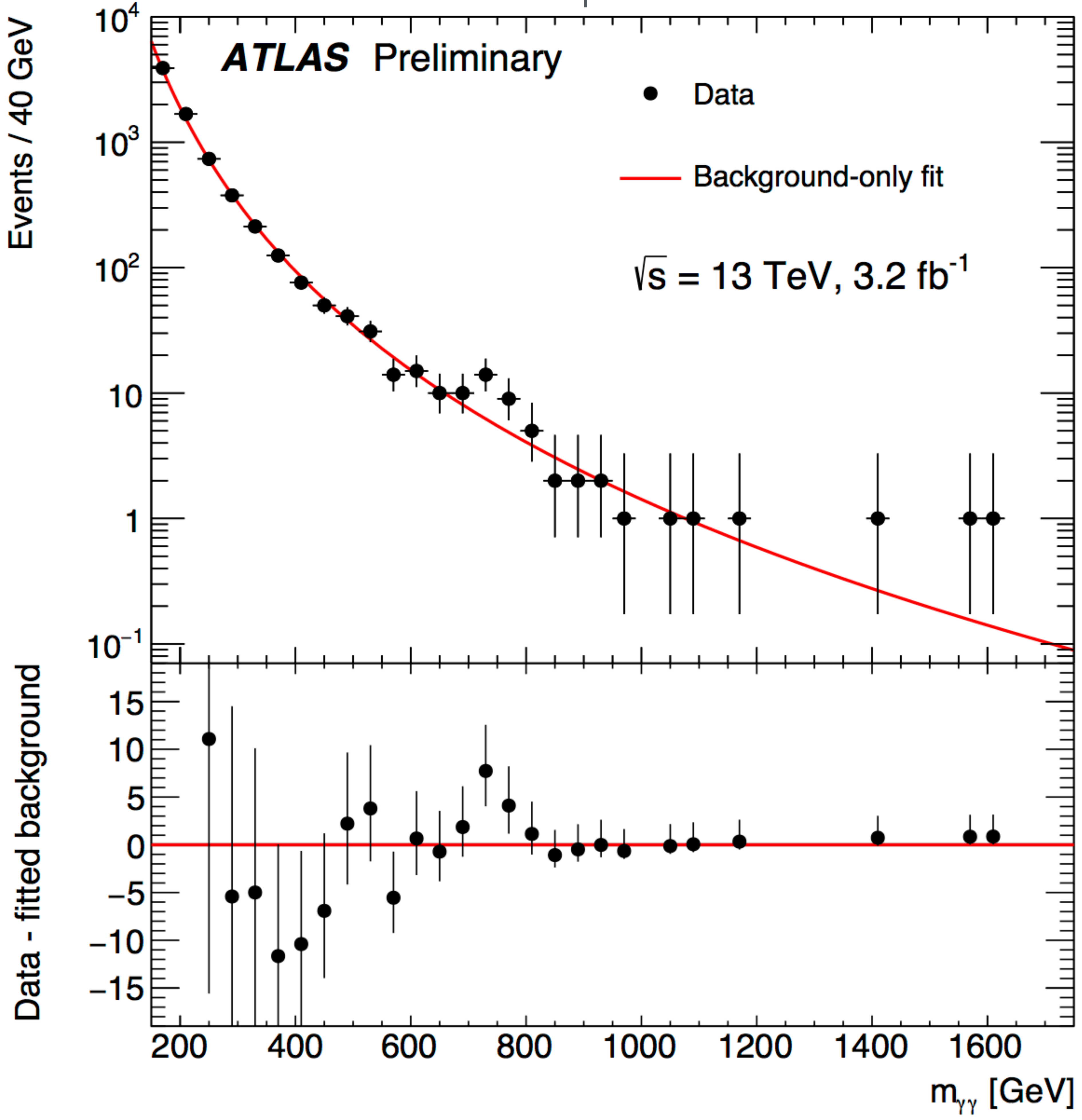
$\Delta \tilde{\chi}^2$ should follow a χ^2 distribution with $n_\mathrm{dof} = 1$ if H_0 ist true

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

→ H_0 can be safely rejected

Why bother with statistical methods?

"750 GeV diphoton excess"



Presentations by CMS and ATLAS, December 2015:
https://indico.cern.ch/event/442432/

Statistics:
Draw reliable conclusions
from data

In case of doubt:
just get more data …

Yes, but not always easy …

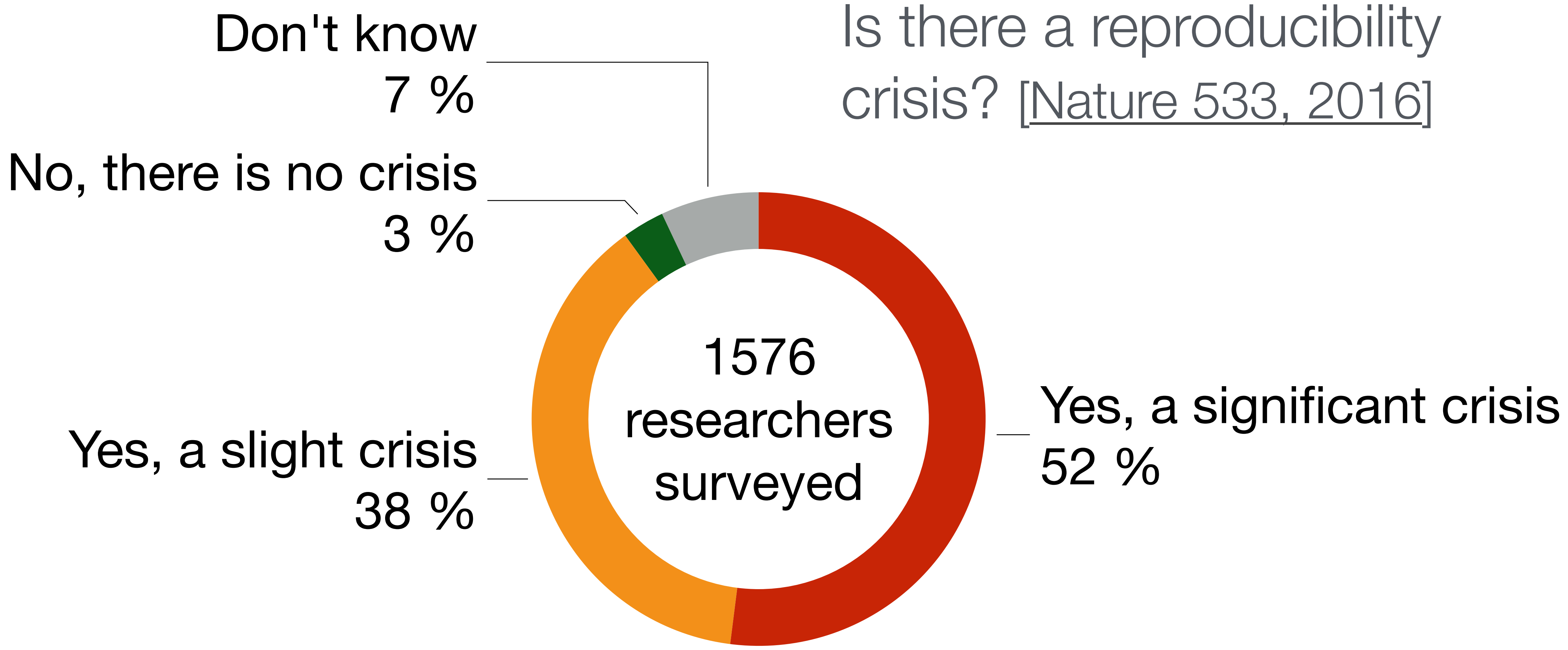
A heavy Higgs boson?

Peak disappeared with more data
… [link]

A look at other research fields

"**Why Most Published Research Findings Are False**":
Main thesis: large number, if not the majority, of published medical research papers contain results that cannot be replicated.

Reproducibility crisis:
Affects the social sciences and life sciences most severely (in particular psychology)



Is there a reproducibility crisis? [Nature 533, 2016]

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

It can be proven that most claimed research findings are false.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2 × 2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2 × 2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

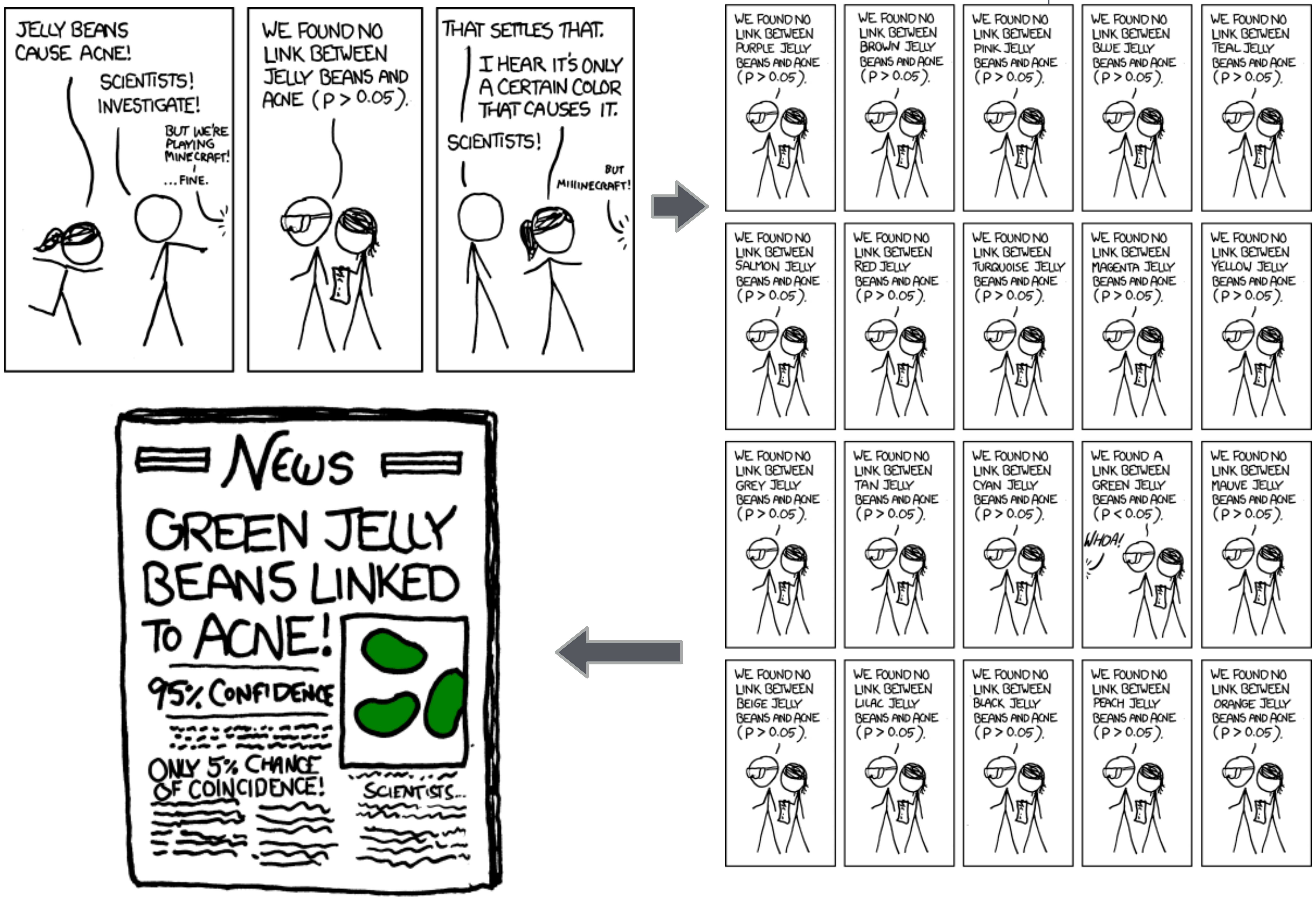
DOI: 10.1371/journal.pmed.0020124

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

John Ioannidis
(Stanford School of Medicine)
PLoS Med 2(8): e124., (2005),
doi:10.1371/journal.pmed.0020124

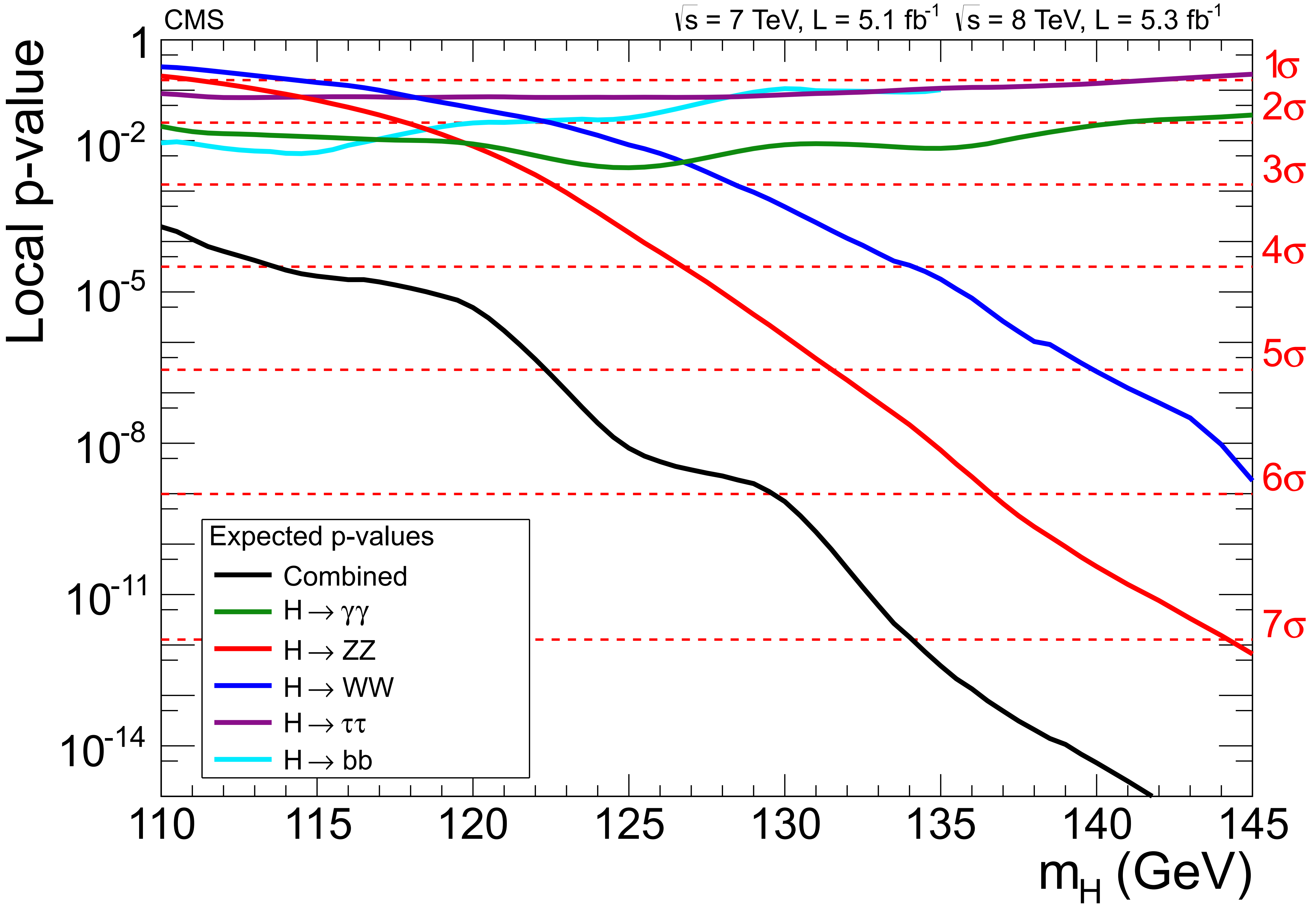
p-value hacking

https://xkcd.com/882/


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

CMS, arXiv:1207.7235

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

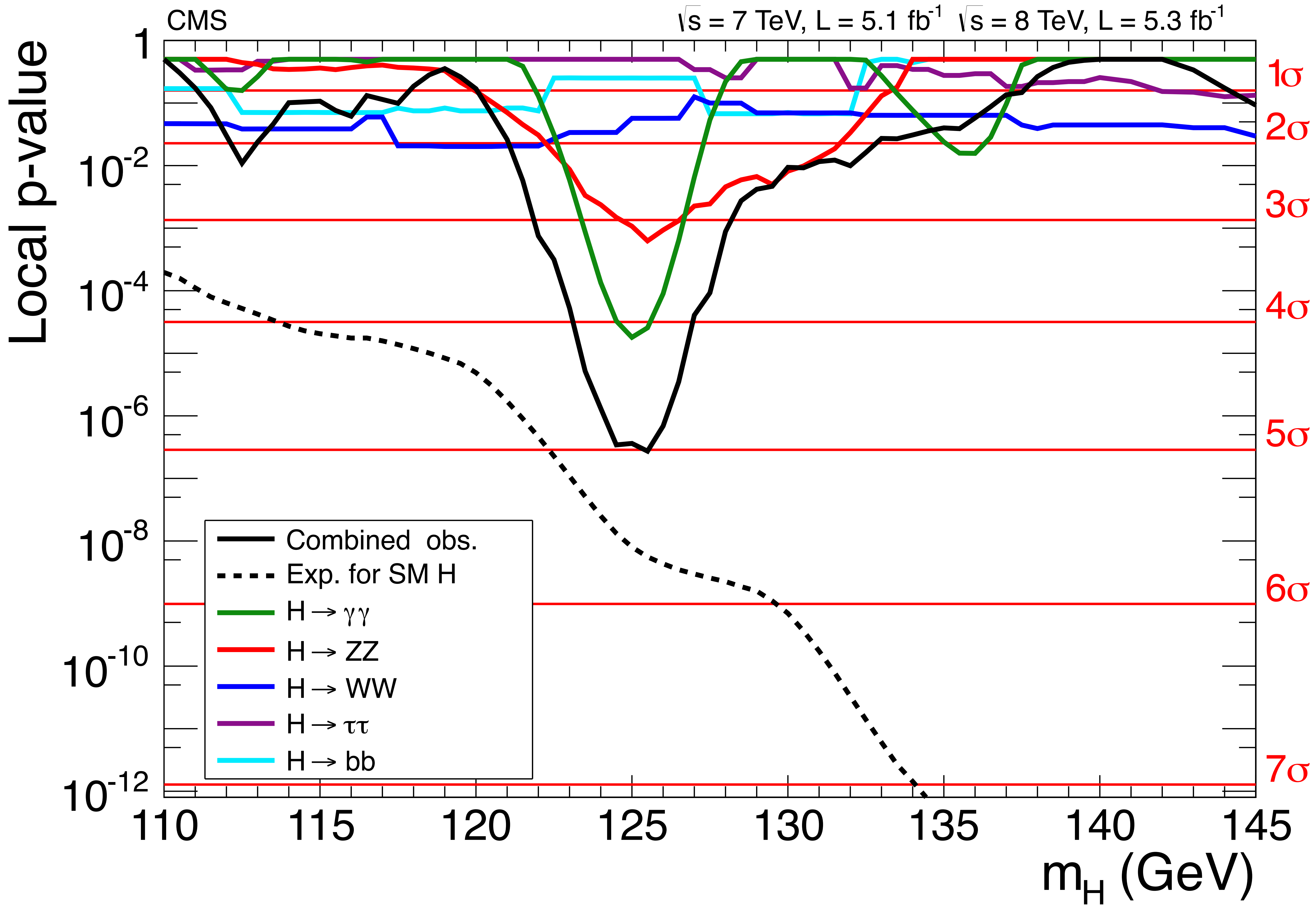


For each assumed Higgs mass (→ local *p*-value)

- 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.

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

Bayesian hypothesis testing

- We can write

$$p(H_0 \,|\, \vec{d}) = \frac{p(\vec{d} \,|\, H_0)\, p(H_0)}{p(\vec{d} \,|\, H_0)\, p(H_0) + p(\vec{d} \,|\, H_1)\, p(H_1)}, \text{ but also}$$

$$p(H_1 \,|\, \vec{d}) = \frac{p(\vec{d} \,|\, H_1)\, p(H_1)}{p(\vec{d} \,|\, H_0)\, p(H_0) + p(\vec{d} \,|\, H_1)\, p(H_1)}$$

thus

$$\frac{p(H_1 \,|\, \vec{d})}{p(H_0 \,|\, \vec{d})} = \frac{p(\vec{d} \,|\, H_1)\, p(H_1)}{p(\vec{d} \,|\, H_0)\, p(H_0)} = \frac{p(\vec{d} \,|\, H_1)}{p(\vec{d} \,|\, H_0)} \, \frac{p(H_1)}{p(H_0)}$$

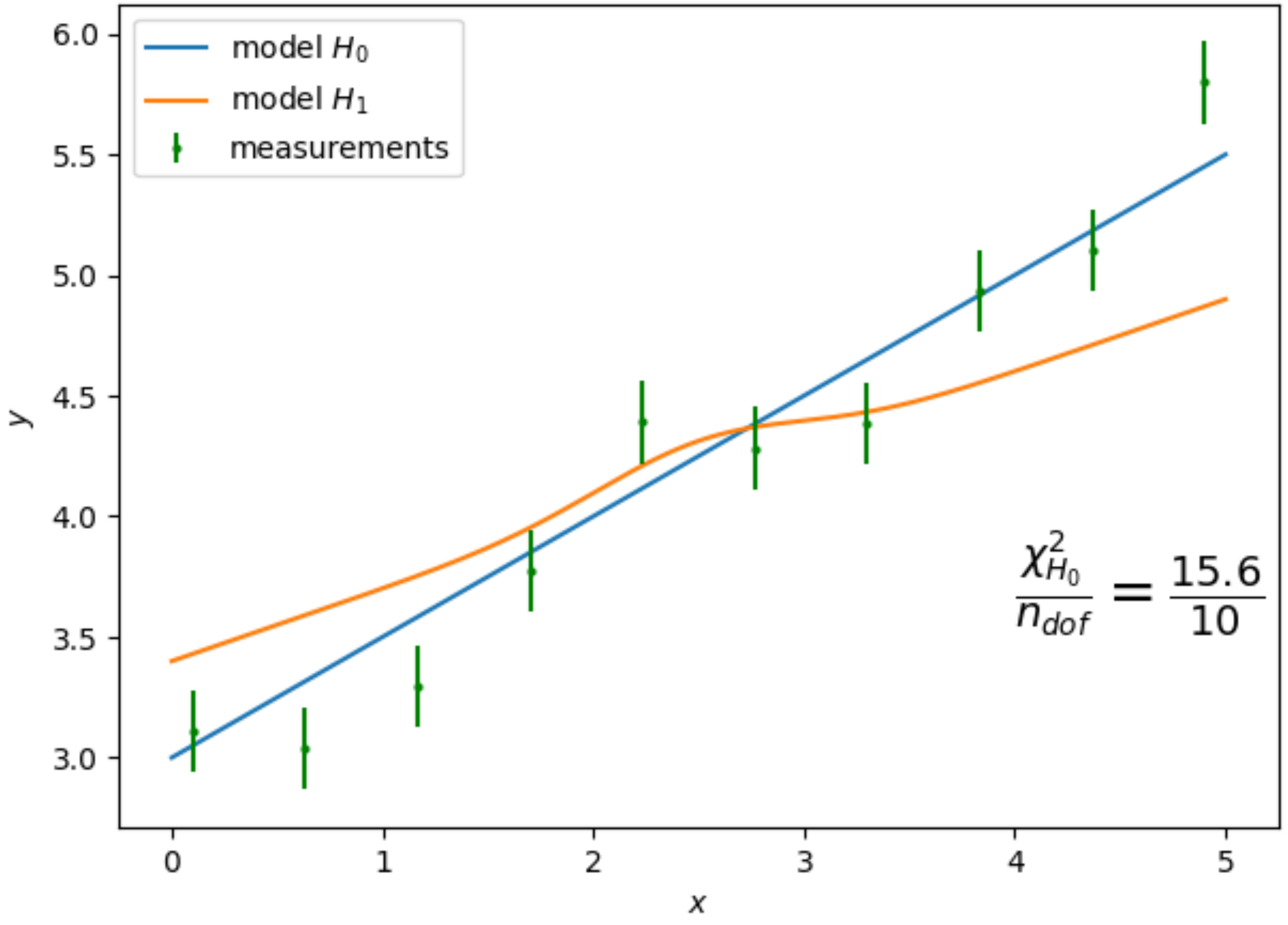
- The factor $\dfrac{p(\vec{d} \,|\, H_1)}{p(\vec{d} \,|\, H_0)}$ is *independent* of the prior! It tells us how the ratio of the probabilities changes. This is called the *Bayes factor*.
- It is also the likelihood ratio
- Provides an objective result from Bayesian analysis (when no free parameters are present)

Bayesian hypothesis testing (2)

- We know the $\Delta\chi^2 = 48.5$
- Thus the Bayes factor ratio is

$$\frac{p(\vec{d}\,|\,H_1)}{p(\vec{d}\,|\,H_0)} = \exp\left(-\frac{48.5}{2}\right) \approx 3 \cdot 10^{-11}$$

- So whatever our priors are, the data pushes the posteriors very much towards H_0
- There is no choice of statistic in Bayesian statistics, just Bayes' theorem
- When the models have free parameters, they need to be marginalised out first
  - Then the Bayes factor depends on the prior of the parameters



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