

P-values

Allen Caldwell
Max Planck Institute for Physics
May 8, 2015

Outline:

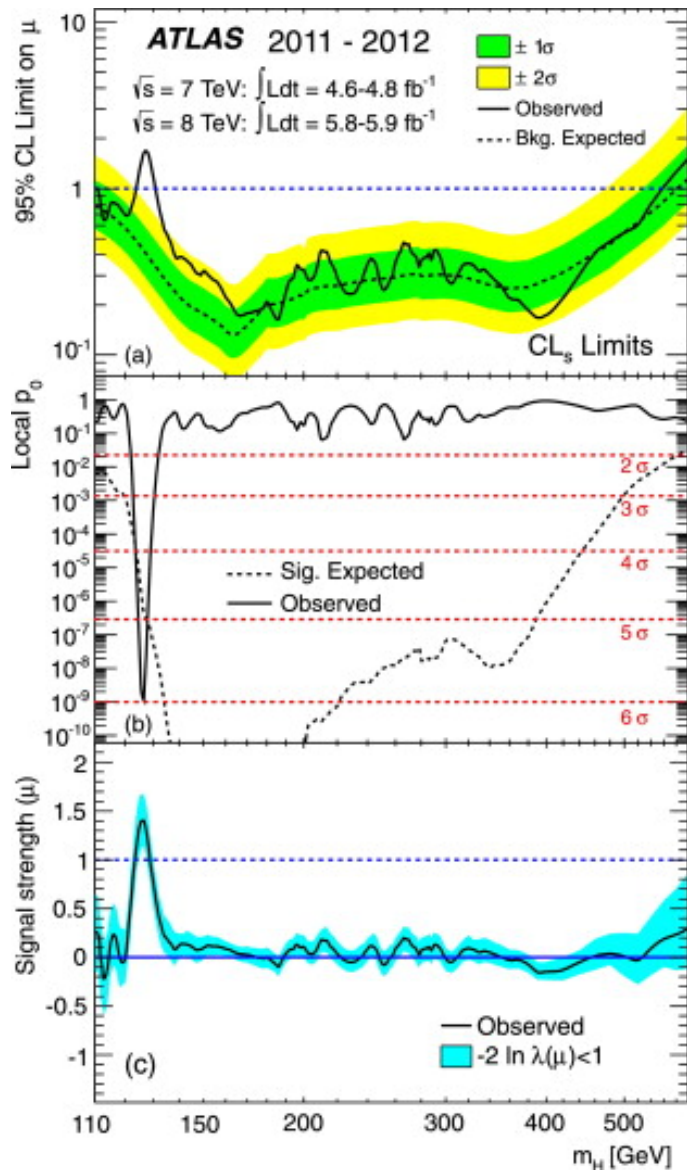
- a) What are p-values ?
- b) Definitions
- c) Model testing
- d) Goodness-of-fit

*Primarily based on
F. Beaujean, A. Caldwell, D. Kollar, K. Kröninger,
'p-values for model evaluation',
Phys.Rev. D83 (2011) 012004*



p-values

Higgs discovery used p-value



Basic and Applied Social Psychology

Editorial

David Trafimow a & Michael Marks

New Mexico State University

Published online: 12 Feb 2015.

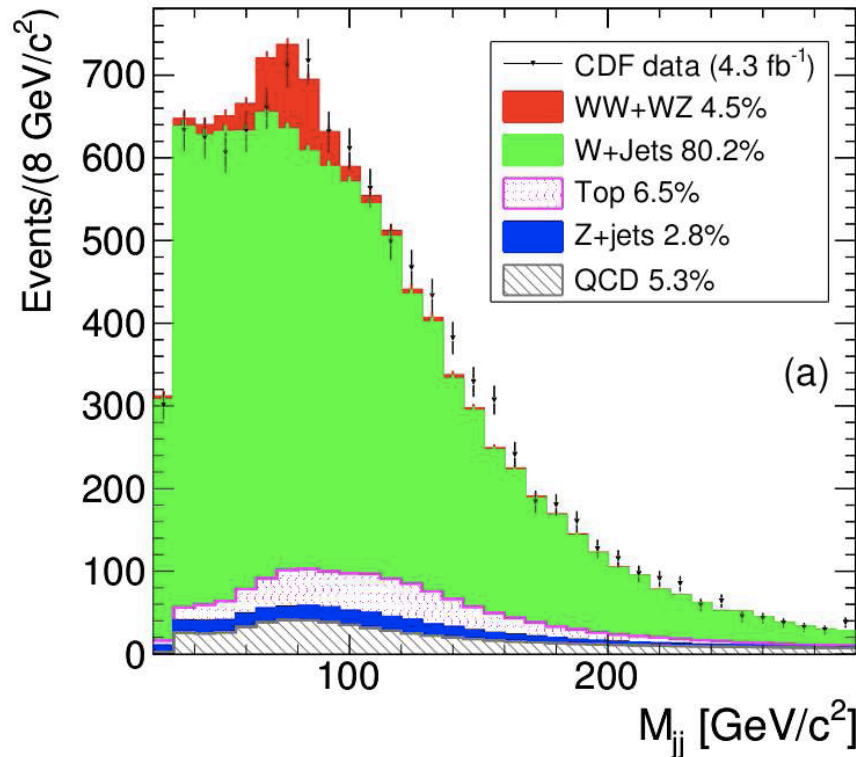
Question 1. Will manuscripts with p-values be desk rejected automatically?

Answer to Question 1. No. If manuscripts pass the preliminary inspection, they will be sent out for review. **But prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about "significant" differences or lack thereof, and so on).**

When/how/if to use ?

Frequent misuse

G. D'Agostini, Probably a discovery: Bad mathematics means rough scientific communication, arXiv:1112.3620v2 [physics.data-an]



Quoting a Discovery article:

It is what is known as a “three-sigma event,” and this refers to the statistical certainty of a given result. In this case, this result has a 99.7 percent chance of being correct (and a 0.3 percent chance of being wrong).

Logic: $1 - P(D|H_0) = P(H_1|D)$

This is logical nonsense – this type of confusion is very widespread !

Intent of p-values

Have a test-statistic and ask ‘**what is the chance that the data could have been more extreme given my model assumptions ?**’ If the chance is small, use this to (make decisions, guide next steps, ...)

Use p-values in

- 1) Significance tests.** Define a hypothesis (null, H_0) and try to reject it. Preset rejection threshold $1-\alpha$. Type I error rate is $1-\alpha$, not the p-value.
- 2) Goodness-of-fit tests.** Evidence in favor of hypothesis. If p-value is large enough ‘fail to reject the null hypothesis’.

The above from ‘Simple Facts about p-values’ by C. Blocker, J. Conway, L. Demortier, J. Heinrich, T. Junk, L. Lyons, G. Punzi (CDF Collaboration internal note)

An Example

First step in using p-value: need a quantity that summarizes the data.

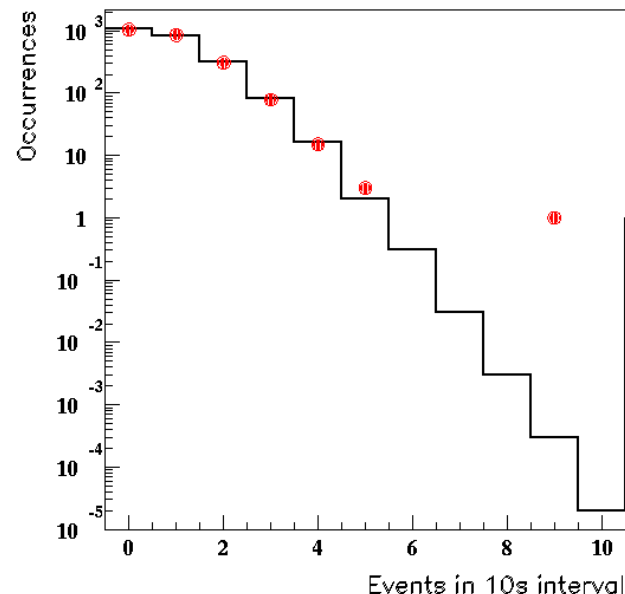
In the simplest case, the data itself (number of observed events)

Example: Observation of Supernovae – IMB experiment

Number of events in 10 sec interval:	0	1	2	3	4	5	6	7	8	9
Frequency	1042	860	307	78	15	3	0	0	0	1
Poisson with mean 0.77	1064	823	318	82	16	2	0.3	0.03	0.003	0.0003

From the data, extract a null hypothesis:

H_0 'The data is from a constant rate background process with 0.077 events/second'



IMB Observations

Test statistic=number of events observed in a ten second interval.

p-value = probability to see this number or greater (tail-area probability)

$$P(n \geq 9|0.77) = \sum_{n=9}^{\infty} \frac{e^{-0.77} 0.77^n}{n!} = 1.3 \cdot 10^{-7} \quad \text{Local p-value}$$

2306 10s intervals analyzed. Apply correction for 'trials factor' or 'look elsewhere effect'

$$\begin{aligned} P(r \geq 9|0.77, 2306 \text{ trials}) &= 1 - P(r < 9|0.77, 2306) \\ &= 1 - P(r < 9|0.77)^{2306} \\ &= 1 - (1 - 1.3 \cdot 10^{-7})^{2306} \\ &\approx 2306 \cdot 1.3 \cdot 10^{-7} \\ &= 3 \cdot 10^{-4} \end{aligned}$$

Numerically small – so we discovered something ?

Bayesian answer – have to consider other options and priors for each option.

Test statistic

In general, we can think of quantities that summarize a 'distance' between the expectation and the observed. E.g., χ^2 is such a quantity. It is a test statistic (scalar function of the data, given the model).

$T(x|M, \lambda)$ Test statistic for possible data x given the model M and parameters λ

Create probability density for this quantity: $P(T|M, \lambda)$

A p-value is a value of the cumulative pdf for the test statistic for some observed value of the data, D (or the complement – depends what is considered extreme).

$$p = F(T(D)) = \int_{T_{\min}}^{T(D)} P(T) dT \quad \text{or} \quad (= 1 - p)$$

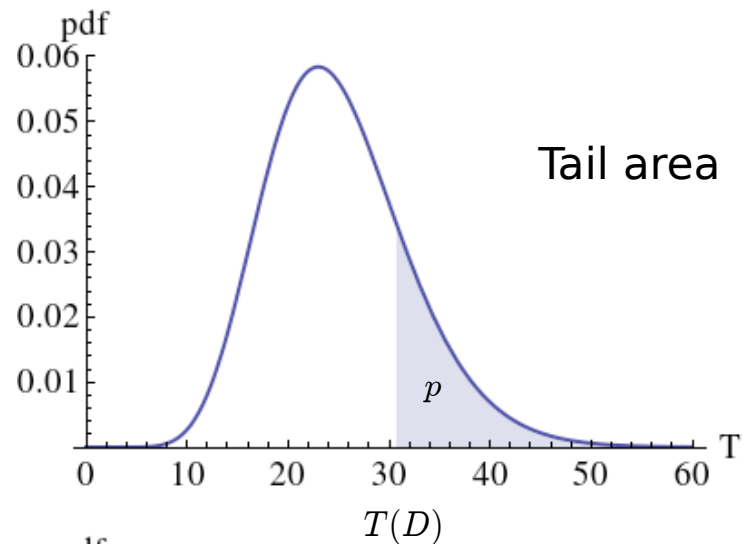
If the model is correct, we expect a flat distribution for p-values between (0,1).

$$P(F) = P(T) \frac{dT}{dF} = \frac{P(T)}{P(T)} = 1$$

p-values and model selection

- Definition:

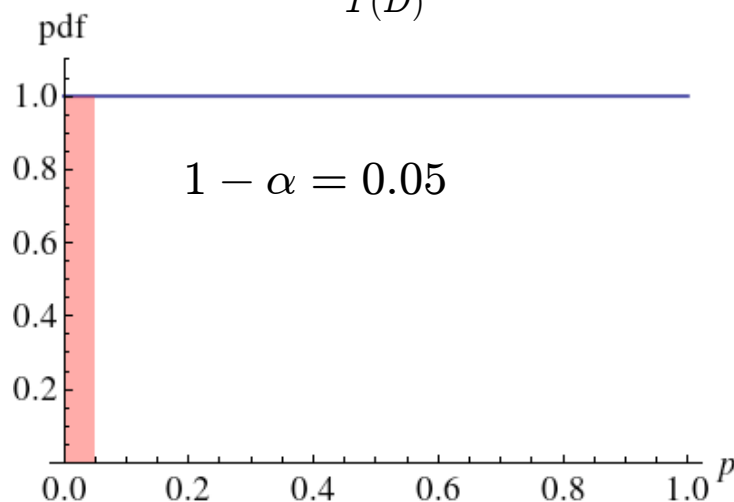
$$p \equiv P(T > T(D)|M)$$



- Assuming M and before data is taken:
 p uniform in $[0,1]$

- Confidence level α :

$$p < 1 - \alpha \Rightarrow \text{reject model}$$



Why do we reject the model for small p-values if all are equally likely ?

p-values are incoherent as measures of support

From M. Schervish, 'P values: What they are and what they are not', The American Statistician, Vol. 50 #3 (1996) 203.

Coherence: “if hypothesis H implies hypothesis H’, then there should be at least as much support for H’ as there is for H.” **p values fail to meet this criterion***.

Argument proven for parameters defined in range [a,b]. Applies to point hypotheses by taking

$$a \rightarrow b$$

and also to one sided intervals by taking

$$a \rightarrow -\infty \text{ or } b \rightarrow \infty$$

*** Bayes factors also fail this criterion.** Only Bayesian posterior probabilities satisfy this requirement.

p-values are incoherent as measures of support

Example: imagine we measure a quantity which we assume comes from a Gauss distribution with unit variance, but unknown mean with

$$\mu \in [a, b]$$

We measure a value x , and calculate a p-value. If we consider a second interval $\mu \in [a', b']$ with

$$a' < a \text{ and } b' > b$$

Coherence: we would expect the p-value in the second case to be at least as large. Not generally true. Concrete example

$$x = 2.18 \quad a = -0.5 \quad b = 0.5 \quad a' = -0.82 \quad b' = 0.52$$

$$p = \frac{2}{b - a} \int_a^b d\mu \int_{|\mu - x|}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 0.036$$

$$p' = \frac{2}{b' - a'} \int_{a'}^{b'} d\mu \int_{|\mu - x|}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 0.0295$$

Note: Schervish used another calculation of p , but with the same conclusion.

Long run: p-value tests will always reject true model

Probability to pass N tests at confidence level α is $1 - \alpha^N$. E.g., if have 10 independent tests at 95% CL, then have 60% chance that at least one of the tests will reject the true model.

Evaluate p values vs amount of data taken - p values perform a random walk. With probability 1, the p-value will cross below any preselected $1-\alpha$ as N becomes infinite.

Example: simple Gaussian with known mean and variance $\mu = 0$ $\sigma = 1$

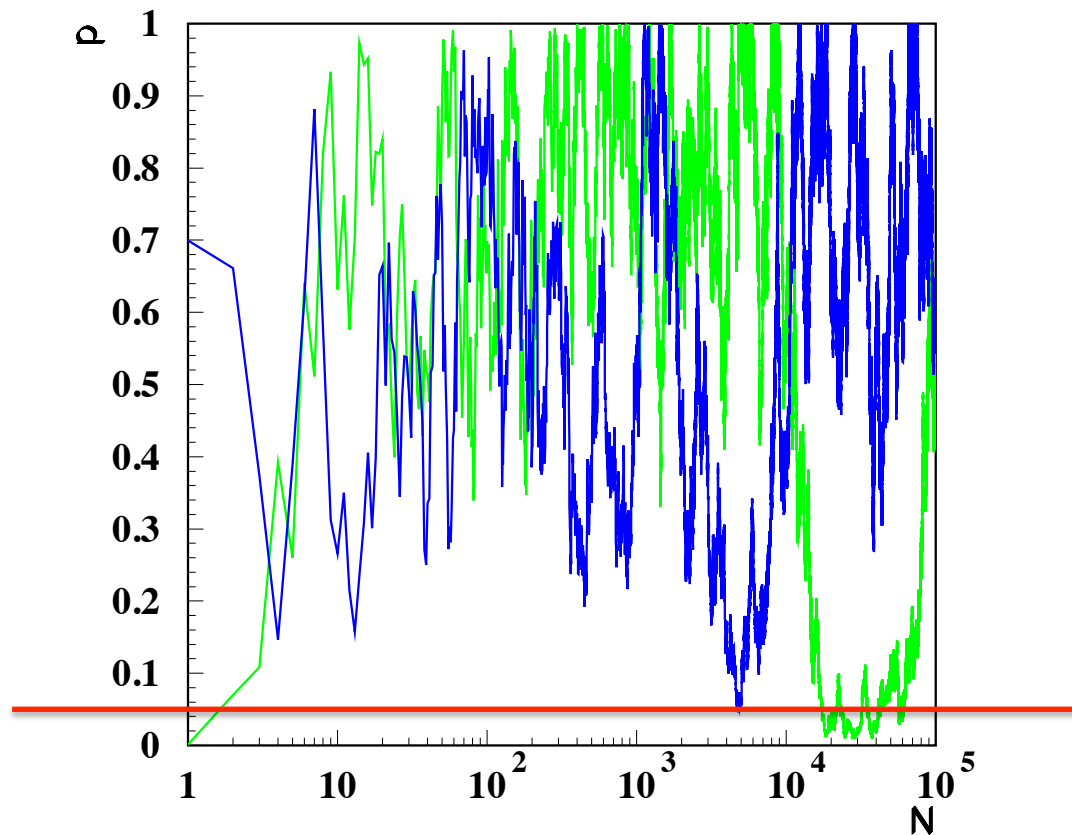
Generate $\{x\}$ according to the correct model and calculate \bar{x} and p-value as

$$p = 2 \int_{|\bar{x}|}^{\infty} \sqrt{\frac{N}{2\pi}} e^{-Nt^2/2} dt \qquad \sigma_{\langle x \rangle} = \frac{\sigma_x}{\sqrt{N}} = \frac{1}{\sqrt{N}}$$

Could also look at one-sided test, same result

p-value tests will always reject true model with enough data

Example – 2 different data sets of 10^5 x values.



A prescription (I. J. Good, "The Bayes/Non-Bayes Compromise: A Brief Review", J. Amer. Statist. Assoc. 87 (1992) 597.) p value cutoff for model selection scaled as $1/\sqrt{N}$.

Better – use Bayes learning rule.

Discussion/Examples

Superluminal neutrinos: p-value for $v \leq c$ very small, yet no discovery claimed or accepted

Bicep2: p-value again small, great excitement but also doubts from the start ...

Higgs: p-value small and Higgs was discovered

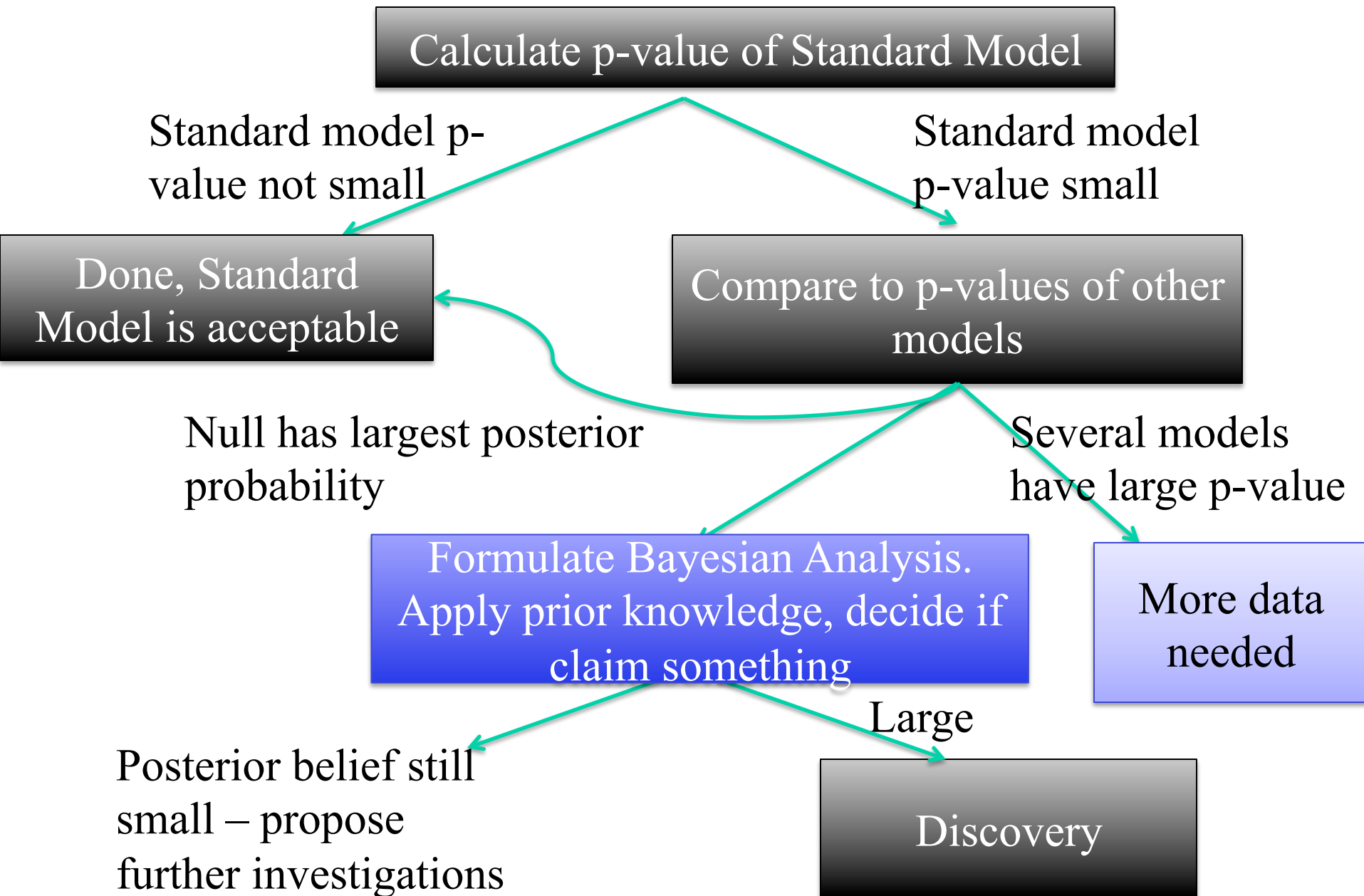
Why OK in Higgs case ?

- The Higgs was expected
- The Higgs was seen in several channels, two different experiments

Shows: If results unexpected or strange, then burden of proof much higher (prior is small).

Also, there is no sound way to account for the 'trials factor'. Rather, use first result to clearly specify the 'region of interest', redo the experiment without the trials factor. The statistics is now understandable.

Incomplete set of models



Goodness of Fit

Example : Use χ^2 as our test statistic. The probability distribution of χ^2 is known analytically. This is one of the main reasons why this test statistic is so popular. **Strictly only applicable in limited cases** (data follow Gaussian distribution from expectation, resolutions are not parameter dependent, if parameters fitted, then function needs to be linear in parameters, ...).

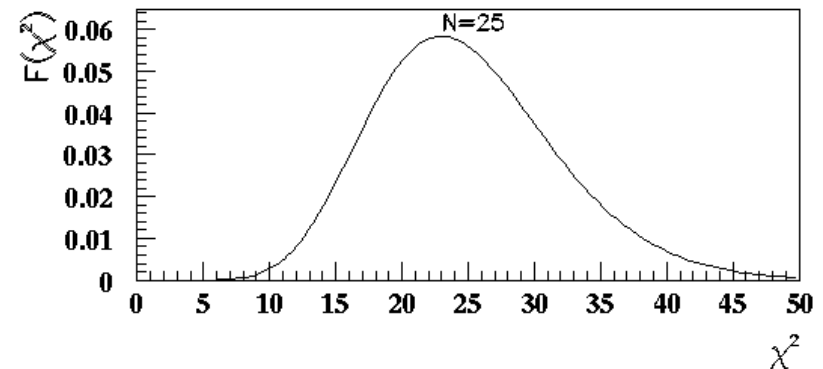
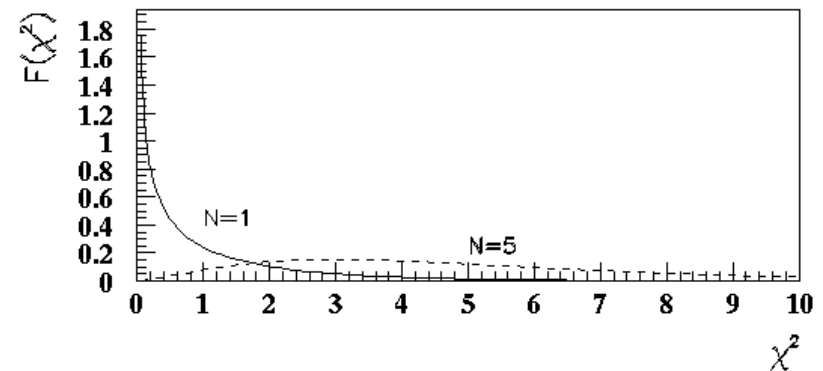
$$P(\chi^2)d\chi^2 = \frac{1}{2^{N/2}\Gamma(N/2)} e^{-\chi^2/2} (\chi^2)^{(N/2)-1} d\chi^2$$

$$p = \int_{\chi_0^2}^{\infty} P(\chi^2)d\chi^2$$

or

$$p = \int_0^{\chi_0^2} P(\chi^2)d\chi^2$$

This is often a good guide (see later)



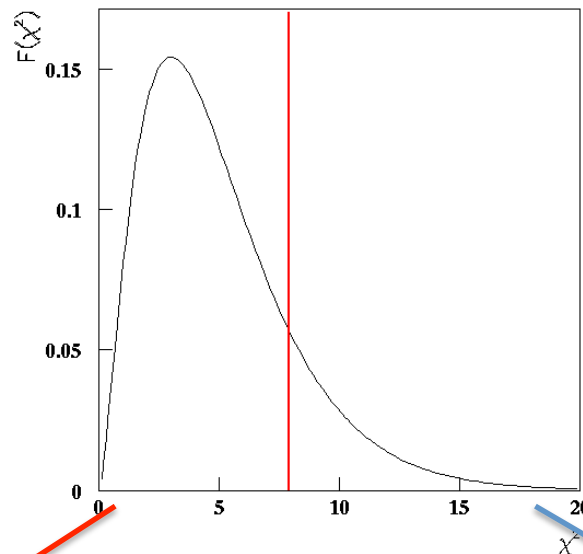
Example -Straight Line Fit

7 data points, 5 DF

$P=0.99$

Missing correlations ?

Errors too big ?

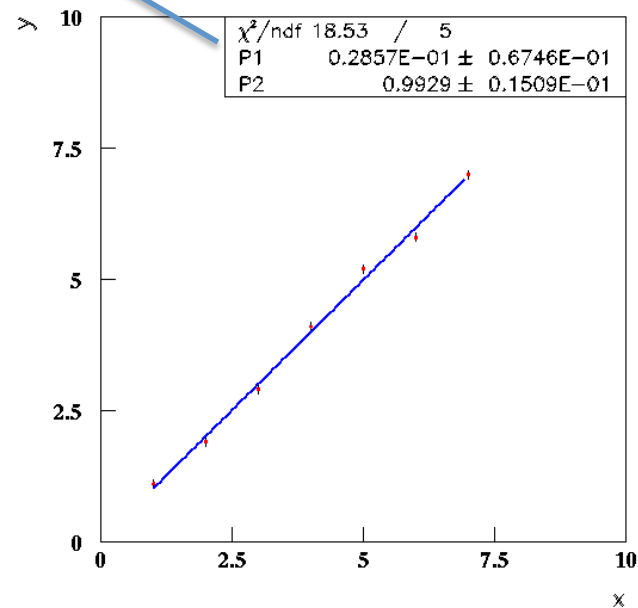
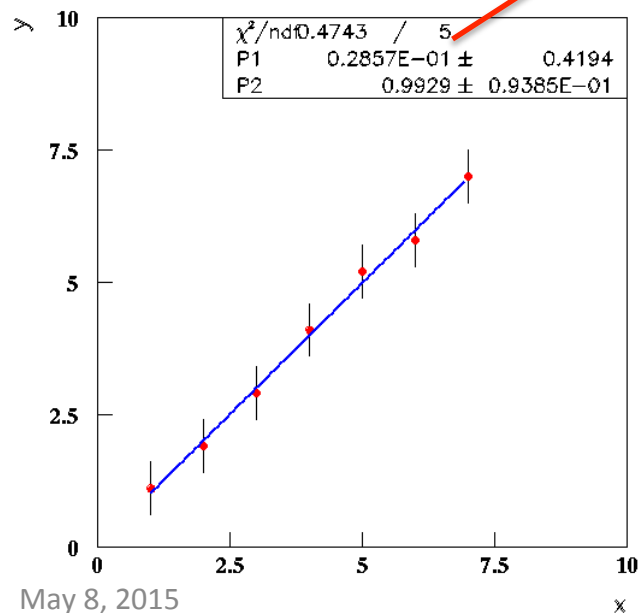


Example of what would usually be considered as a good fit

$P=0.002$,

Wrong model ?

Errors too small ?



Comment p-values for Goodness-of-Fit

Experience shows - χ^2 tests are often useful when trying to find an adequate representation of the data.

Why ?

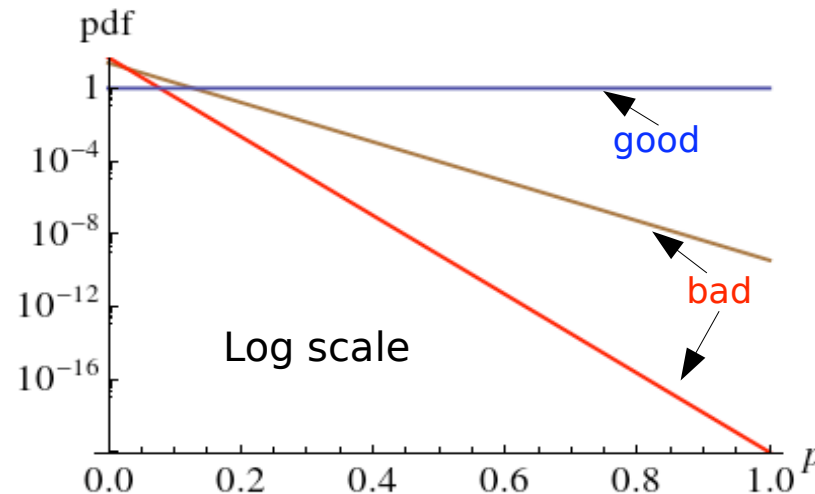
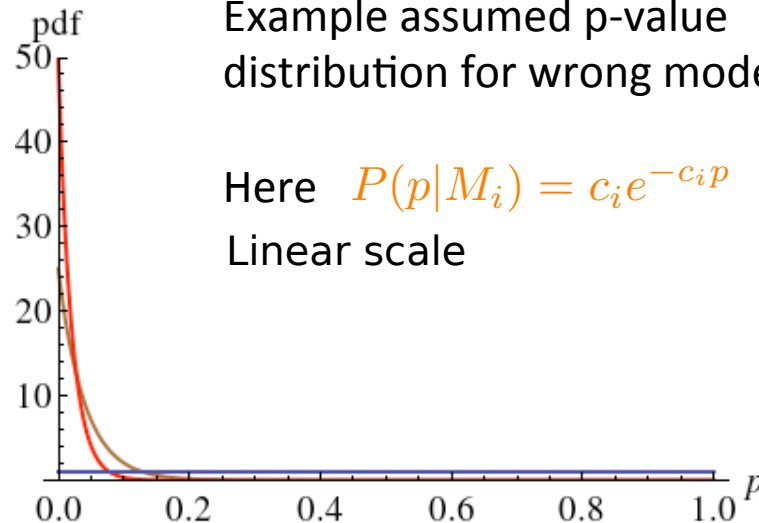
For the correct model, the expected p-value distribution is uniform in $[0,1]$. All p-values have the same probability, so why would we pick any particular values to throw out ?

Because we expect the wrong models to have p-value distributions peaked at zero (although we don't know the p-value distributions for the wrong models).

Accept some rate of rejection of correct model ...

Example assumed p-value distribution for wrong models.

Here $P(p|M_i) = c_i e^{-c_i p}$ $c_i \gg 1$
Linear scale

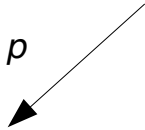


Reasoning behind p-values

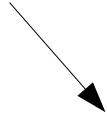
- Similar prior for all models $P(M_i) \approx P(M_j)$

- Bayes Theorem: $P(M_0|p) \approx \frac{P(p|M_0)}{\sum_{i=0}^K P(p|M_i)}$

Small p


$$P(M_0|p \approx 0) \approx \frac{1}{1 + \sum_{i=1}^K c_i} \ll 1$$

Large p


$$P(M_0|p \approx 1) \approx 1$$

Bayes Theorem gives justification to p-values

Pitfalls of p-values

p-values depend critically on how you have chosen the test statistic. The same data set can have hugely varying p-values resulting from different choices of the test statistic.

E.g., consider a model where we assume an exponential decay law. We can define the following probabilities of the data:

Unbinned likelihood

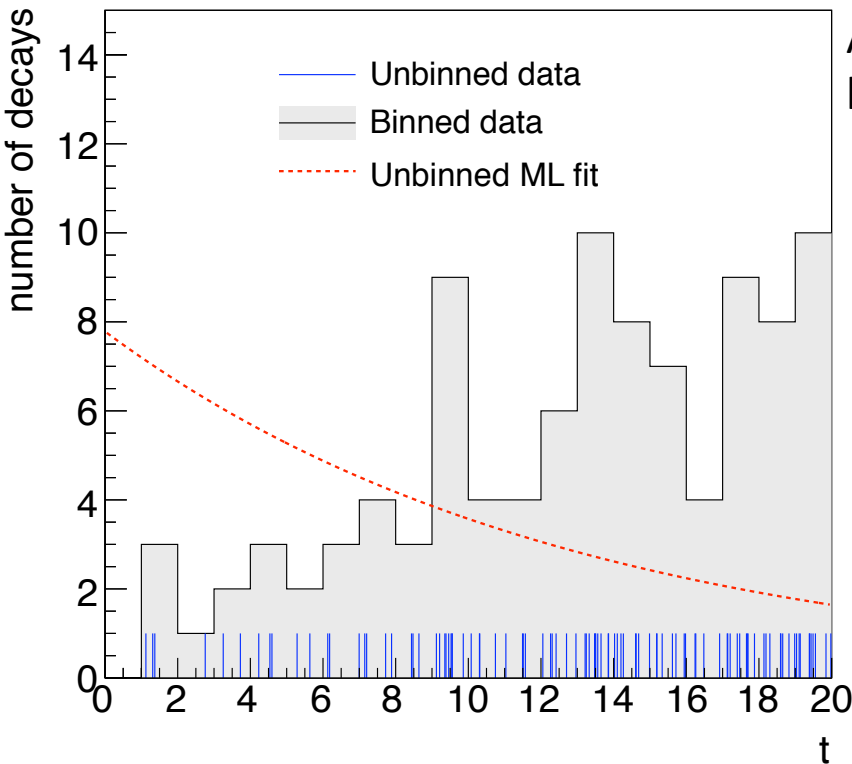
$$P(\vec{t}|\tau) = \prod_{i=1}^N \frac{1}{\tau} e^{-t_i/\tau}$$

Binned Poisson distribution

$$P(\vec{t}|\tau) = \prod_{j=1}^M \frac{e^{-\nu_j} \nu_j^{n_j}}{n_j!}$$

ν_j = expected events in bin j
 n_j = observed events in bin j

pitfalls



Assumed model is exponential. Data actually from linearly increasing function.

Using binned Poisson probabilities, $p \approx 0$

Max likelihood $\tau^* = \frac{1}{N} \sum_{i=1}^N t_i$

$$p = \int_{\sum t'_i > \xi} dt'_1 \int dt'_2 \dots (\tau^*)^{-N} e^{-\sum t'_i / \tau^*} = 1 - P(N, N)$$

$$P(s, x) = \frac{\gamma(s, x)}{\Gamma(s)} = \frac{\int_0^x t^{s-1} e^{-t} dt}{\int_0^\infty t^{s-1} e^{-t} dt}$$

Doesn't depend on the data ! In fact, for large N , $p \approx 0.5$

pitfalls

The p-value from the maximum likelihood is about 0.5 !

The p-value from the binned fit is 0

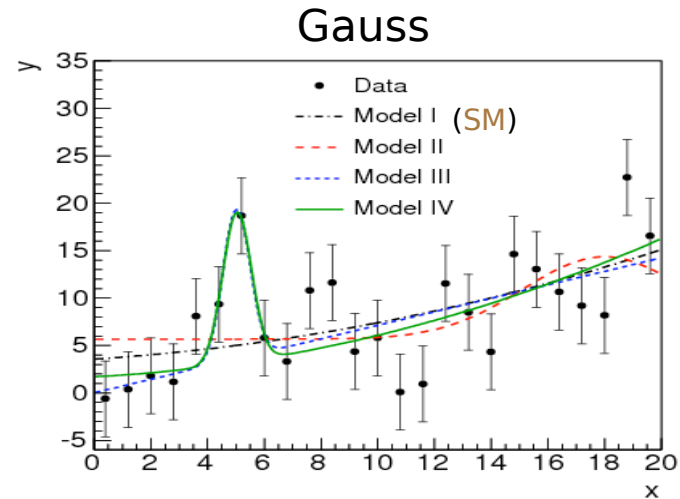
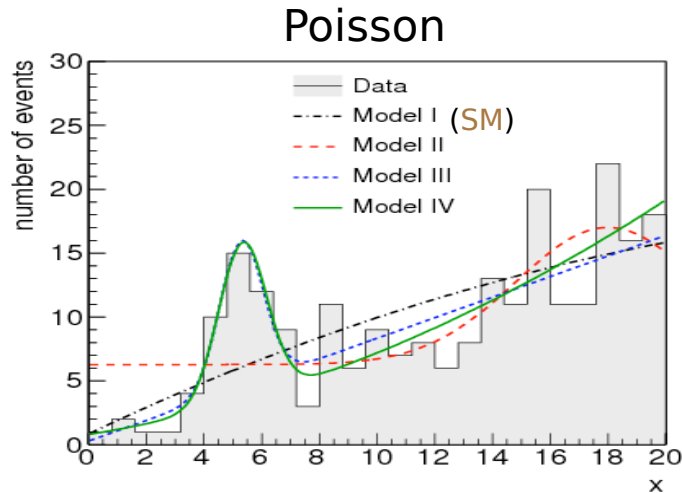
What happened ? The maximum likelihood quantity does not know anything about the distribution of the events, and the result only depends on

$$\tau^* = \frac{1}{N} \sum_{i=1}^N t_i$$

and the p-value only depends on N !

Lesson: make sure your test statistic is sensitive to what you want to test !

p-value numerical study



Fit function

$$f(x|\vec{\lambda}) = \underbrace{A + Bx + Cx^2}_{\text{SM}} + \frac{D}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{\sigma^2}\right)$$

I	: quadratic	}	SM
II	: constant + Gaussian		
III	: linear + Gaussian	}	NP
IV	: quadratic + Gaussian		

More test statistics & examples in

F. Beaujean, A. Caldwell, D. Kollar, K. Kröninger, *Phys.Rev. D83 (2011) 012004*

p-value numerical study

Look at the Poisson case. Focus on χ^2 , but other statistics discussed in paper.

Pearson

$$\chi_P^2 = \sum_i \frac{(n_i - \nu_i)^2}{\nu_i}$$

n_i observed events

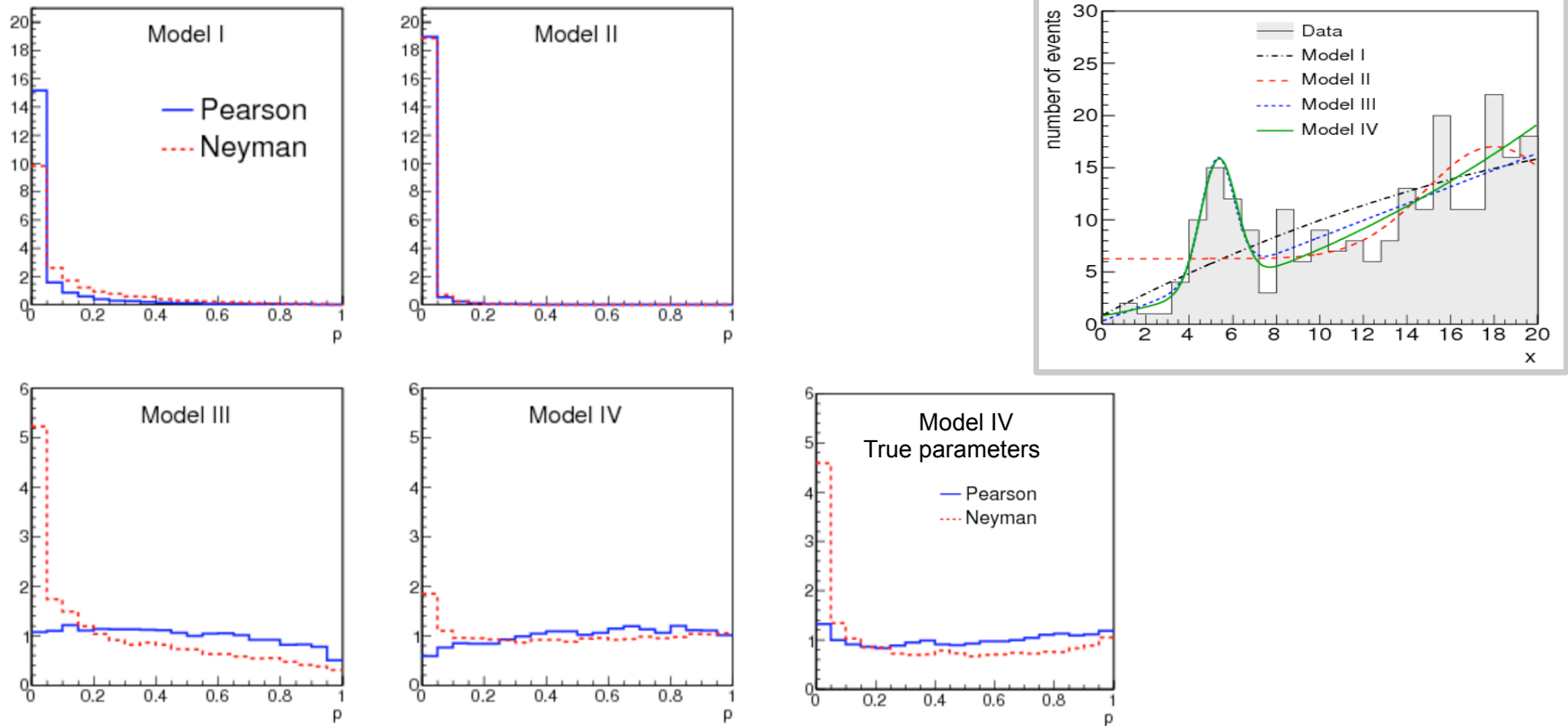
$\nu_i = \nu_i(\vec{\lambda}, M)$ expected events

- Uncertainty if $n_i = 0$? Ignore bin or set uncertainty = 1
- Asymptotically (i.e. infinite data, in **each** bin: $n_i \gg 1$) know distribution of χ_P^2, χ_N^2 .

Neyman

$$\chi_N^2 = \sum_i \frac{(n_i - \nu_i)^2}{n_i}$$

p-value numerical study



- Worrisome peak for Neyman in model III and IV (true)
- Pearson good approximation

p-value distributions are not flat, even for correct model. Fitting important, using correct distribution of test statistic important, finding global mode important...

Summary & Recommendations

p-values are tail probabilities for possible data outcomes. They are intended to signal how rare an outcome is expected to be.

This discussion shows

- **they should not be used directly in model selection**
 - (il)Logical leap
 - Not coherent
 - Further argumentation (priors) necessary
- **In goodness-of-fit tests, care is required**
 - Flat distribution expected ?
 - Choice of test quantity important
 - Numerical issues important

Bayesian argument shows the sense in which p-values can be used to guide reasoning. Since 'all other models' are not specified, this is necessarily a vague procedure.