MAX-PLANCK-GESELLSCHAFT

$\Delta p \cdot \Delta q \geq \frac{1}{2}\hbar$

Max-Planck-Institut für Physik
(Werner-Heisenberg-Institut)

How to tell good from bad

Frederik Beaujean

http://arxiv.org/abs/1005.3233v2
http://arxiv.org/abs/1011.1674

Particle Physics Colloquium

Munich, 10.12.2010

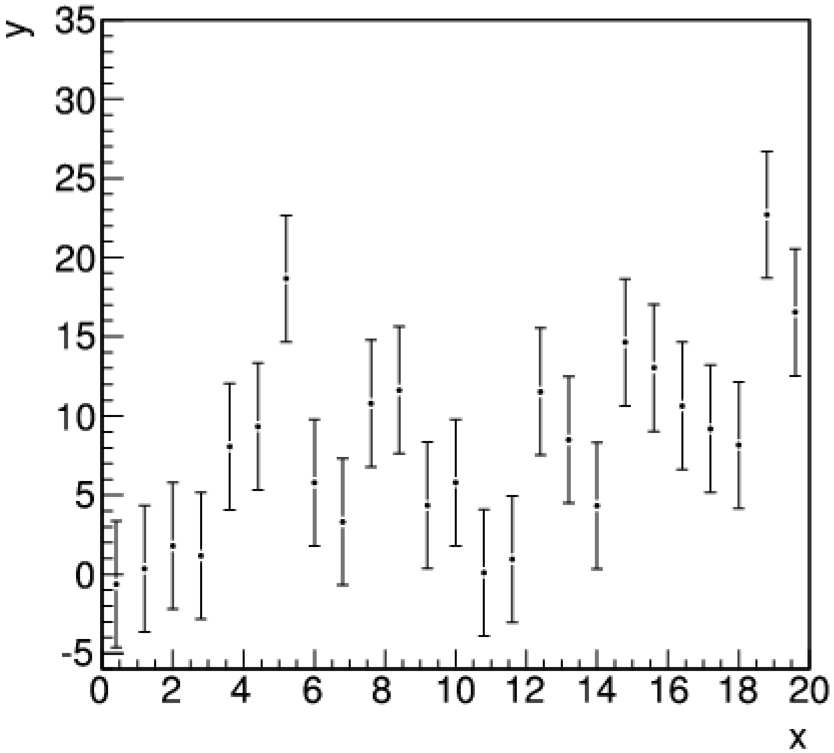
Outline

- Why goodness of fit?
- Statistical approach – p-values
- Common statistics – common pitfalls
- [A bit of number theory: runs and integer partitions]

Motivating Example

Suppose:

- *N* measurements $y_i(x_i)$ with uncertainty
- Standard Model (SM) background is quadratic
- New physics (NP) predicts signal peak (more than one NP model)

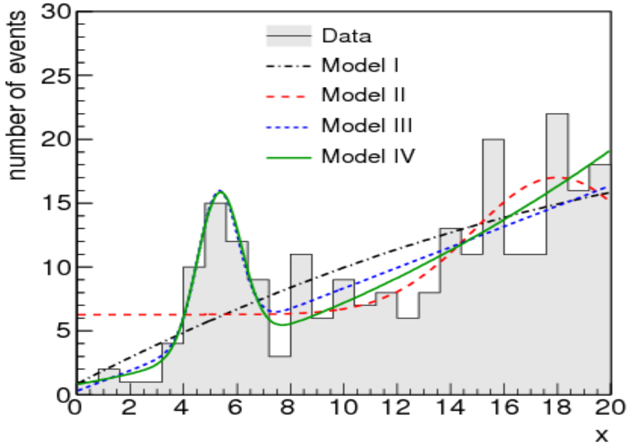


I : quadratic (SM)
II : constant + Gaussian
III : linear + Gaussian
IV : quadratic + Gaussian

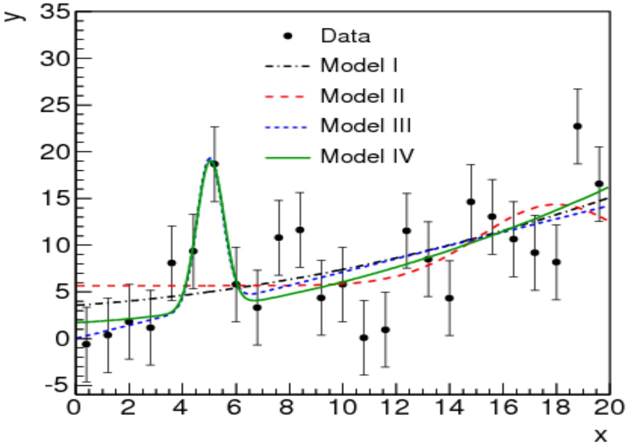
Is Standard model enough to explain data?

Example problem

Poisson



Gauss



Fit function

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

$$\begin{aligned} I : \vec{\lambda} &= (A, B, C) \\ II : \vec{\lambda} &= (A, D, \mu, \sigma) \\ III : \vec{\lambda} &= (A, B, D, \mu, \sigma) \\ IV : \vec{\lambda} &= (A, B, C, D, \mu, \sigma) \end{aligned}$$

Restriction

- Don't compare models directly here. Assume a model is true, then check consistency
- Often no alternative model known
- For model comparison (which one favored by data?) recommend Bayesian approach

Goodness of Fit: standard approach

Requirement:

- Assume a model *M* with parameters *λ*

Test statistic:

- Any scalar function of data, T(D)
- Interpret: large T(D) = poor model

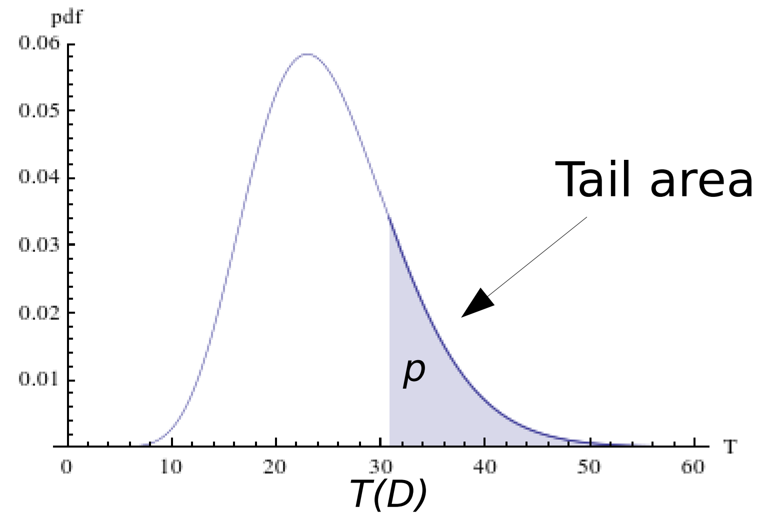
Example:

- Prob. density of the data $$P(D|\vec{\lambda}) \propto \prod \exp\left\{-\frac{\left(y_i - f(x_i|\vec{\lambda})\right)^2}{2\sigma_i^2}\right\} = \exp\left\{\frac{-\chi^2}{2}\right\}$$
- Familiar choice $$T(D) \equiv \chi^2(D)$$
- Extension: discrepancy variable *T(D|λ)*. Fitting procedure important!

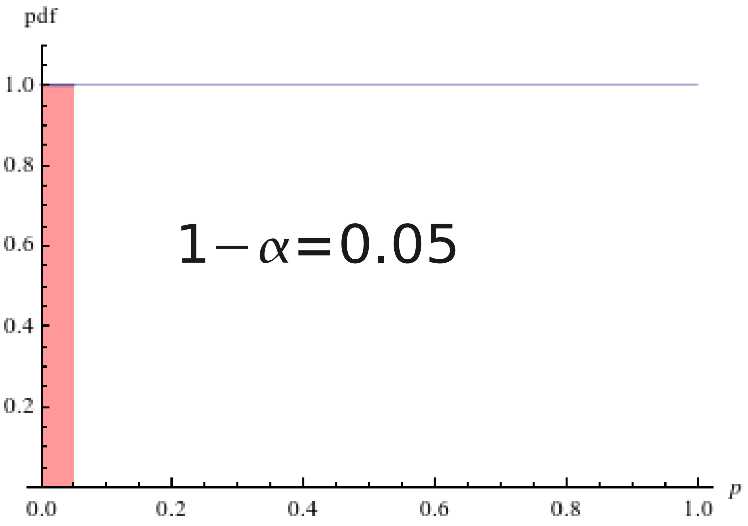
p-value

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

Frequency if experiment repeated



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



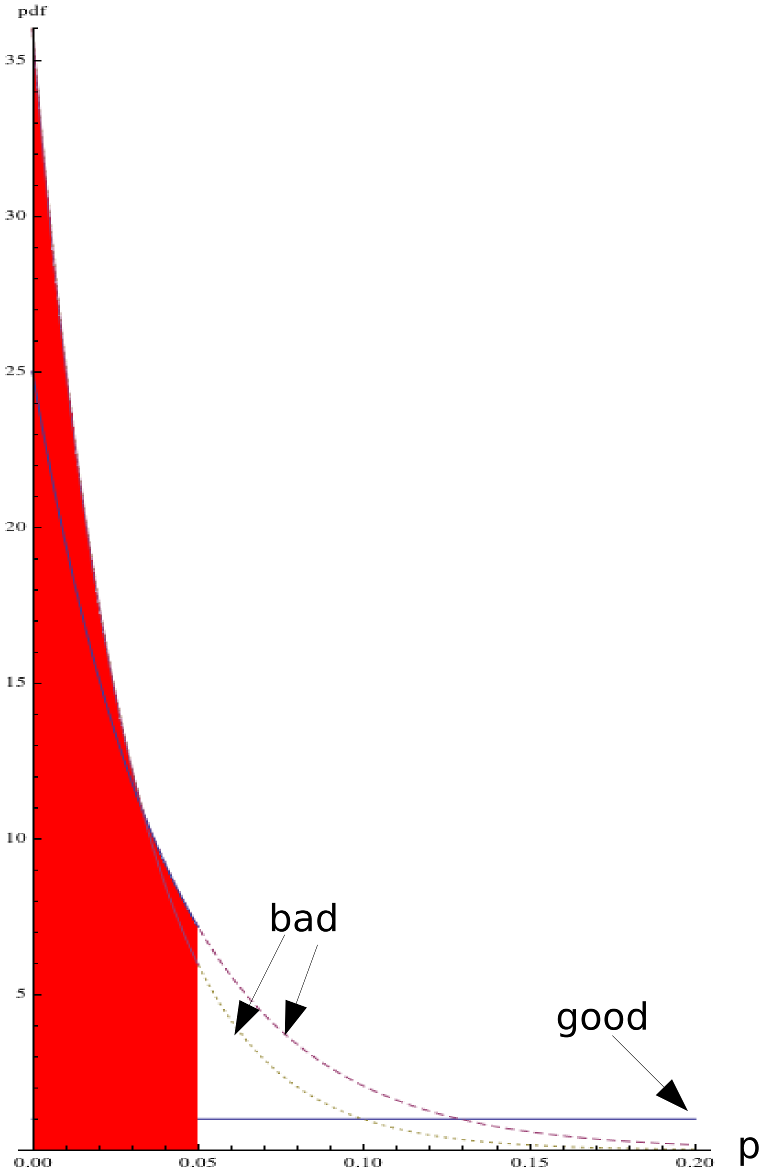
- Confidence level:

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

Warning: p-value *not* the P. that the model is true

Reasoning behind p-values

- Good model: flat p-value
- Bad model: peak at p=0, sharply falling
- Need prior knowledge about alternatives
- vague interpretation



Don't take p-value too seriously!

p-value caveat

Frequent misunderstandings

There are several common misunderstandings about p-values.[3][4]

The p-value is *not* the probability that the null hypothesis is true.
In fact, frequentist statistics does not, and cannot, attach probabilities to hypotheses. Comparison of Bayesian and classical approaches shows that a p-value can be very close to zero while the posterior probability of the null is very close to unity (if there is no alternative hypothesis with a large enough *a priori* probability and which would explain the results more easily). This is the Jeffreys–Lindley paradox.
The p-value is *not* the probability that a finding is "merely a fluke."
As the calculation of a p-value is based on the *assumption* that a finding is the product of chance alone, it patently cannot also be used to gauge the probability of that assumption being true. This is different from the real meaning which is that the p-value is the chance of obtaining such results if the null hypothesis is true.
The p-value is *not* the probability of falsely rejecting the null hypothesis. This error is a version of the so-called prosecutor's fallacy.
The p-value is *not* the probability that a replicating experiment would not yield the same conclusion.
1 − (p-value) is *not* the probability of the alternative hypothesis being true (see (1)).
The significance level of the test is not determined by the p-value.
The significance level of a test is a value that should be decided upon by the agent interpreting the data before the data are viewed, and is compared against the p-value or any other statistic calculated after the test has been performed. (However, reporting a p-value is more useful than simply saying that the results were or were not significant at a given level, and allows the reader to decide for himself whether to consider the results significant.)

The p-value does not indicate the size or importance of the observed effect (compare with effect size). The two do vary together however – the larger the effect, the smaller the p-value will be, other things being equal.

From wikipedia

Choice of statistic

Surprise: T arbitrary

Criteria:

- distribution known
- Easy to compute
- Power to reject alternatives

- Now consider statistics for Gauss, Poisson

Comparison study

- Goal: calculate p-value distribution for common statistics
- 10000 experiments
- Sample *N* data points from Model IV with fixed parameters

- Fit all models with (Markov chain+MINUIT) or MINUIT alone
- Plot the distribution of the p-value for the statistics chosen

Test Statistics: Poisson

Pearson

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

Neyman

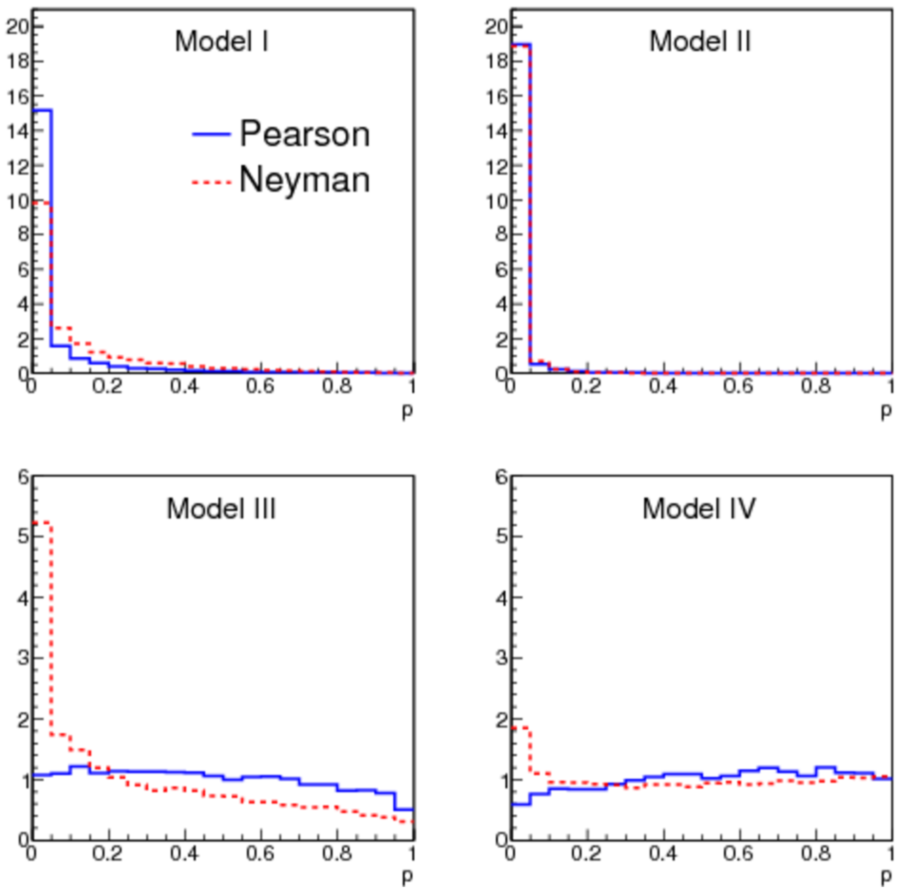
$$\chi_N^2 = \sum_i \frac{(n_i - \nu_i)^2}{n_i}$$

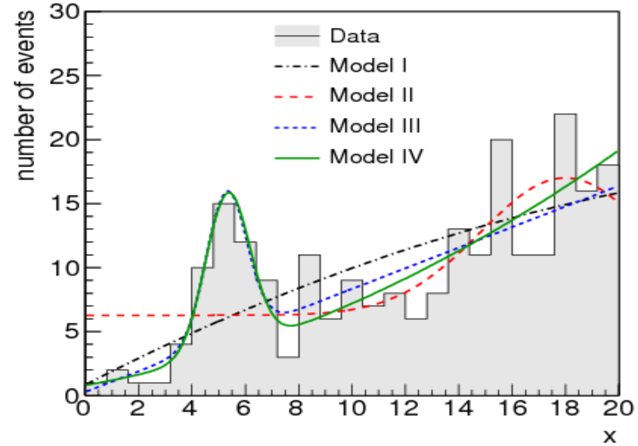
n_i observed events

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

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

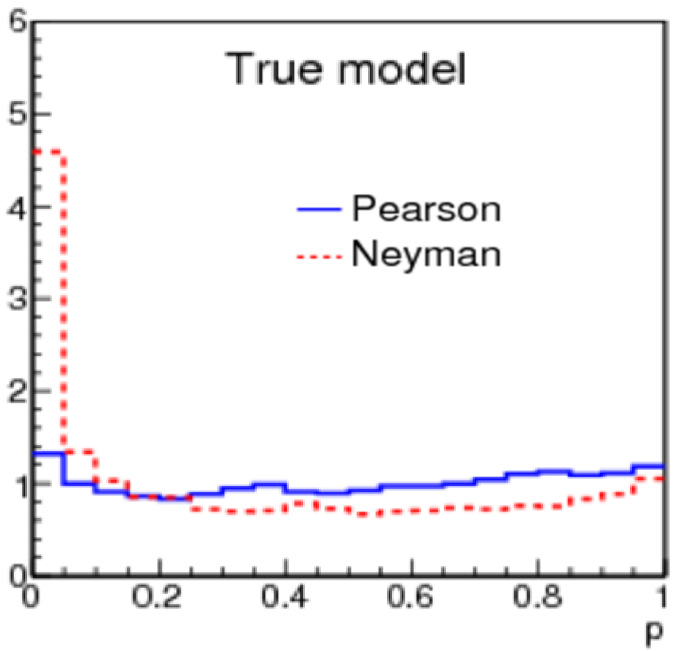
Results

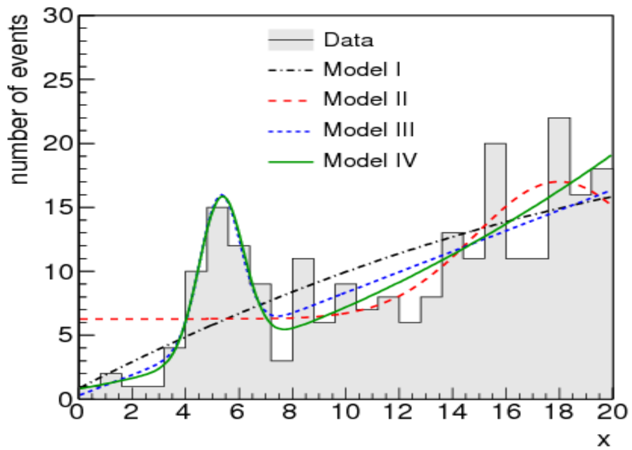




- Worrisome peak for Neyman in model III

Approximations



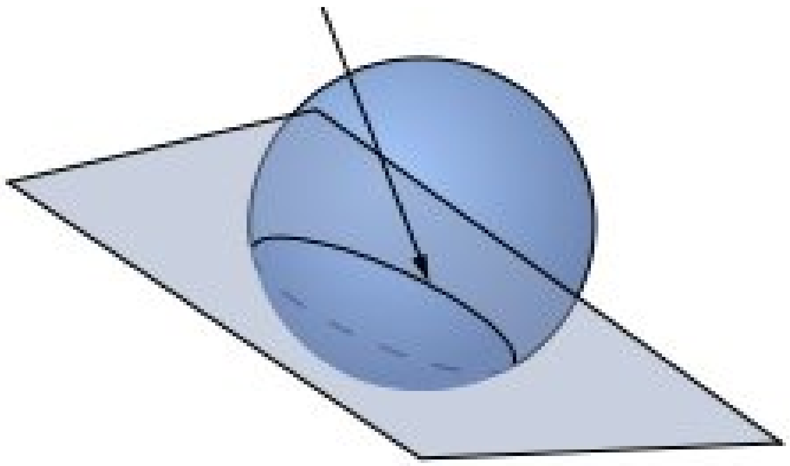


- No parameters fit, just plug in values used to generate the data sets
- Expect flat p-value
- Pearson good approximation
- Neyman bad, gets worse for smaller event numbers

Use Pearson.
No uncertainties without model

Gaussian linear regression

$$\chi^2(\vec{\lambda}) = \sum_i \frac{\left(f(x_i|\vec{\lambda}) - y_i\right)^2}{\sigma_i^2} = \sum_i z_i^2$$



Least squares constraint,
find $\vec{\lambda}^*$ at global minimum:

$$\nabla\chi^2 \equiv \frac{\partial\chi^2}{\partial\lambda_j} = 0 \quad j = 1\ldots n$$

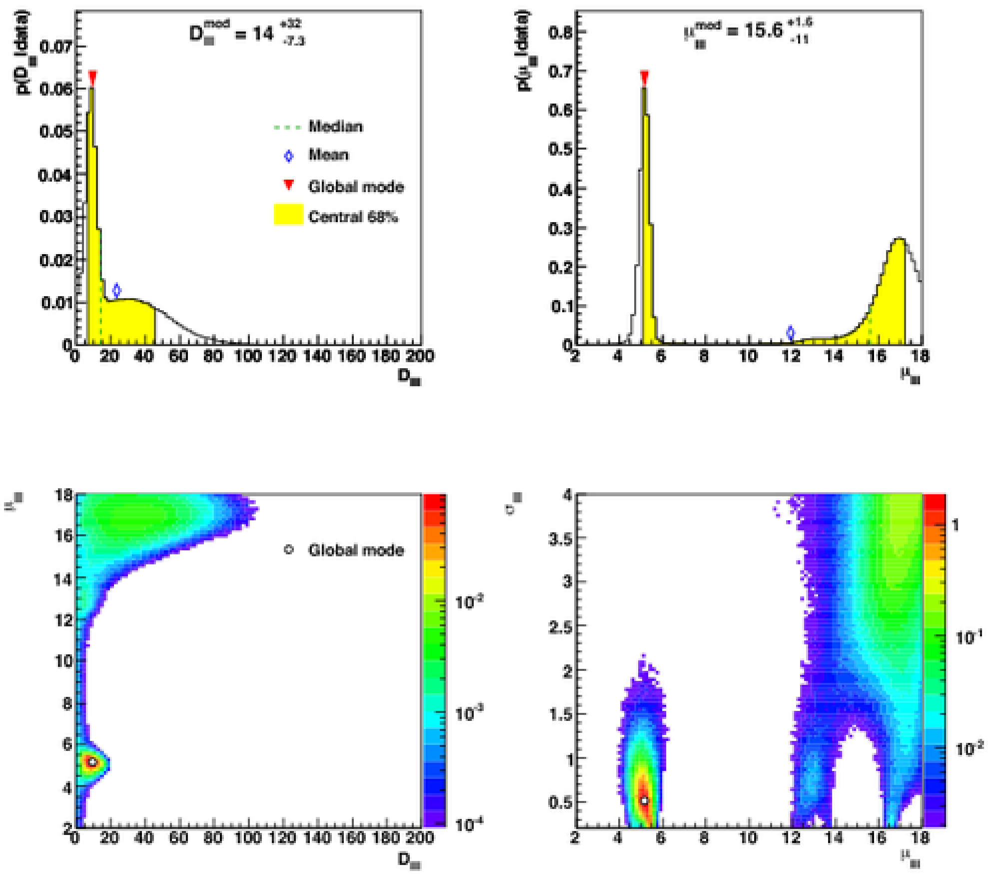
Predictions depend on parameters:

$$f(x_i|\vec{\lambda}) \text{ linear in } \vec{\lambda} \Rightarrow \nabla\chi^2 = 0 \text{ linear in } z_i \Rightarrow (N-n) \text{ DoF}$$

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

In real life, usually don't know exact form of $P(\chi^2|\vec{\lambda}^*, N, n)$

Multimodality



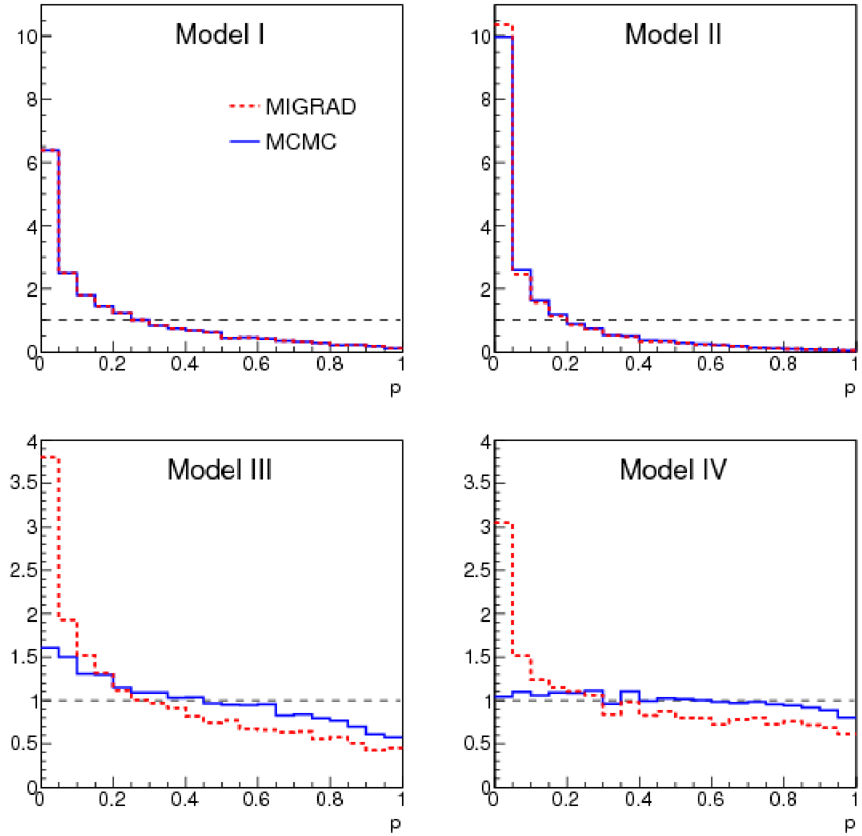
Likelihood of model IV for particular data set and *small* range

- Physics motivates *small* parameter range: e.g. C>0, σ>0.2 ...
- Global minimum may be elsewhere. Compare with *large* range
- Gradient based optimization (MINUIT/MIGRAD): need good starting point
- Clever user guess (difficult) or Markov chain (preferred). [mpp.mpg.de/bat/]

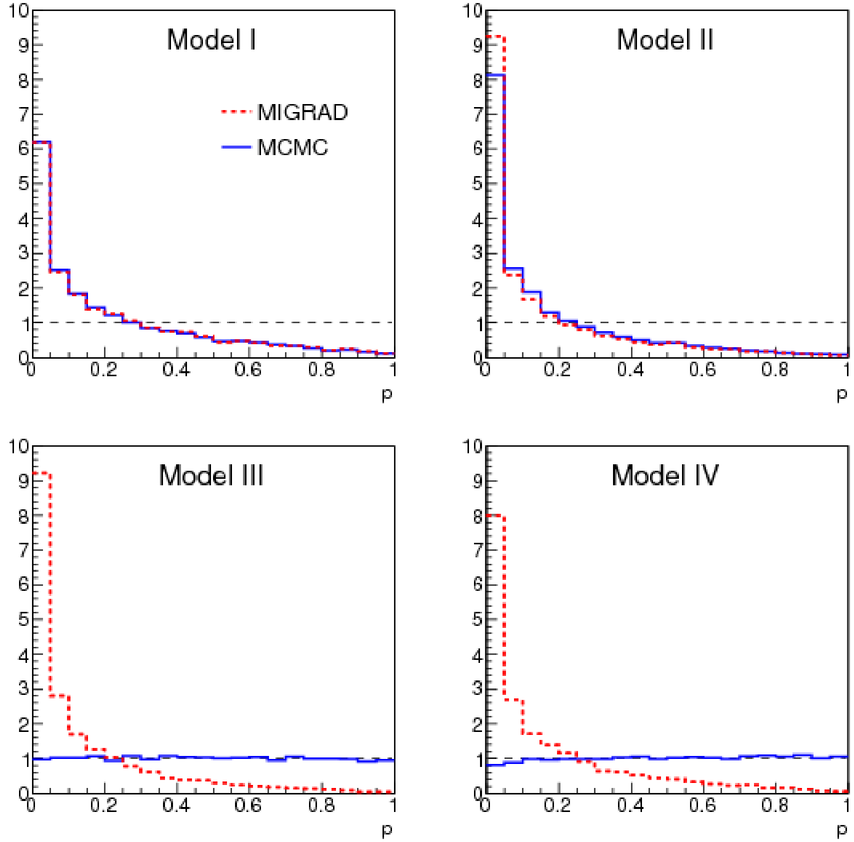
Fitting procedure not trivial!

Results: p-value distribution for χ^2

Small range

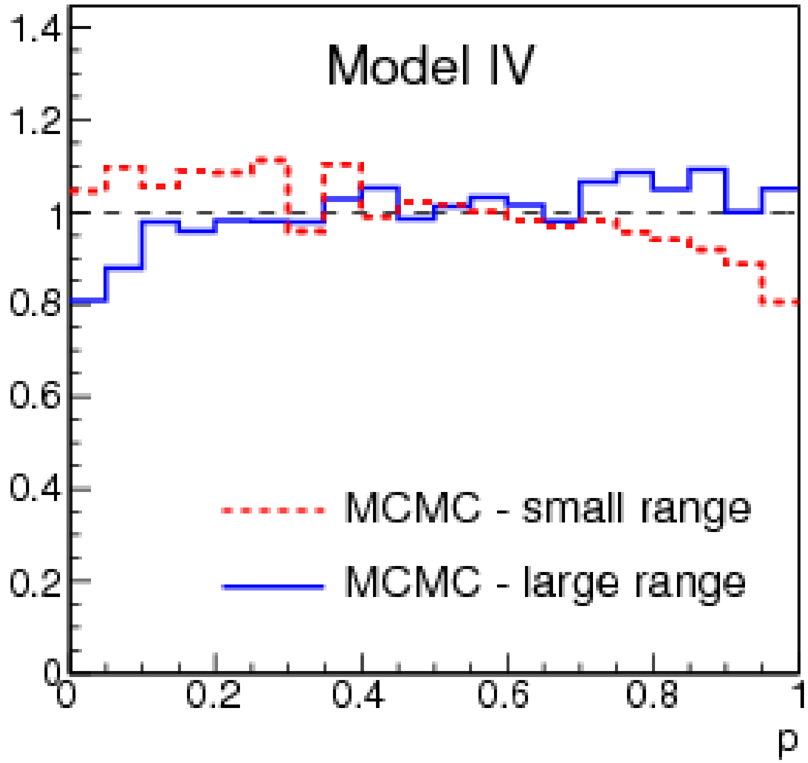


Large range



Fitting procedure and parameter ranges affect distribution of p-value

Local vs Global Minimum



Use χ^2-distribution with $(N-n)$ DoF

True model, global minimum,
but distribution not flat.
→ Nonlinear problem

Constraining parameter range = prior belief

Conclusions

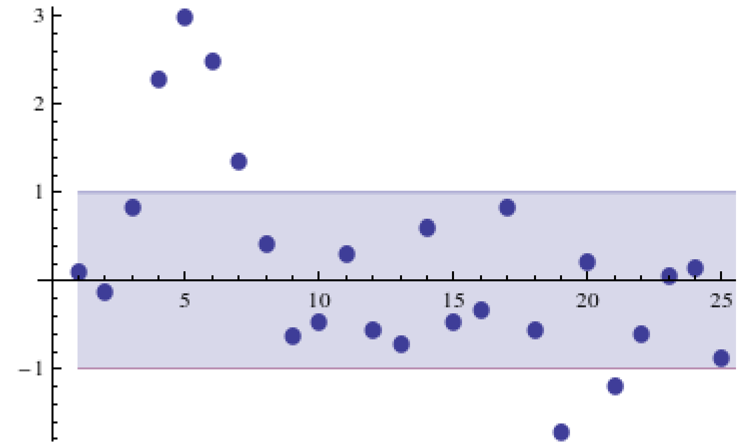
- P-values useful, but need understanding
- Fitting can make big difference
- Choice of statistic crucial
- Beware: distributions usually approximate

FINIS

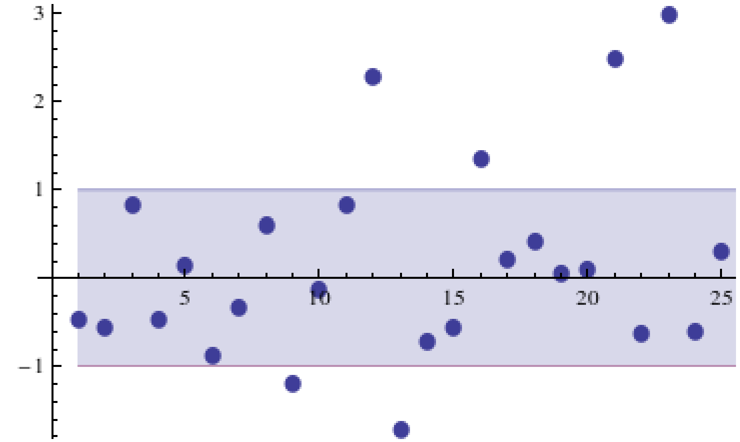
Backup

Runs

- Most statistics disrespect order of data, information wasted
- Human brain good for simple problems



$$\chi^2 = 32.1 \Rightarrow p = 0.16$$



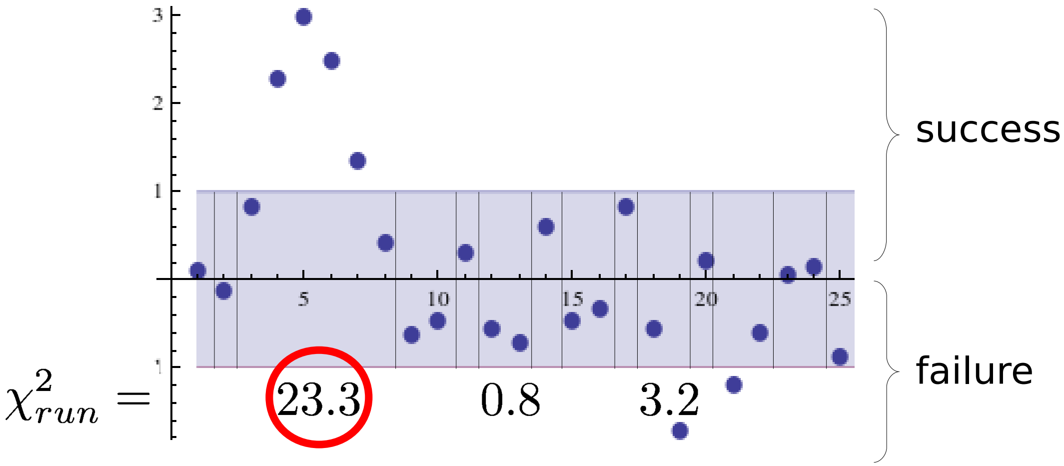
Example:

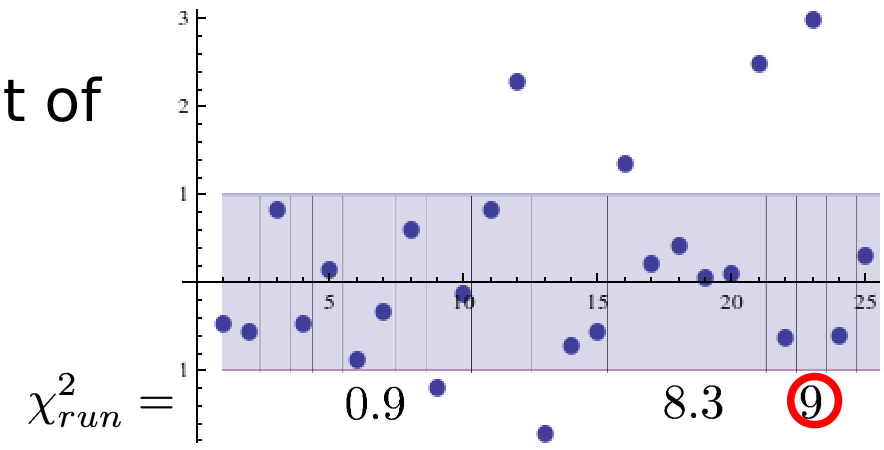
- N=25 datapoints
- Each Gaussian with mean = 0 and variance = 1

$\Rightarrow$ Can we combine information about **order** and **magnitude of deviation**?

Proposal:

- Split data into runs
- Each (success) run has a weight χ^2_{run}
- Test statistic: largest weight of any run $T \equiv \max\{\chi^2_{run}\}$
- p-value becomes $p_{run} \equiv P(T > T_{obs})$

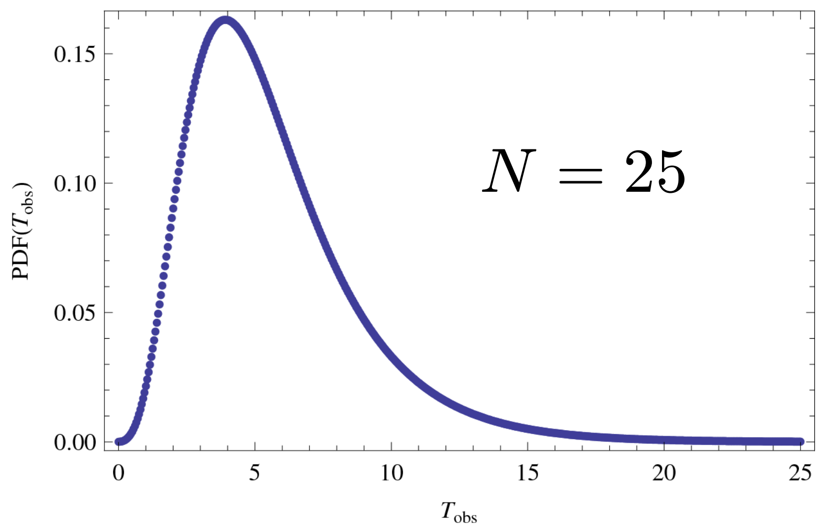




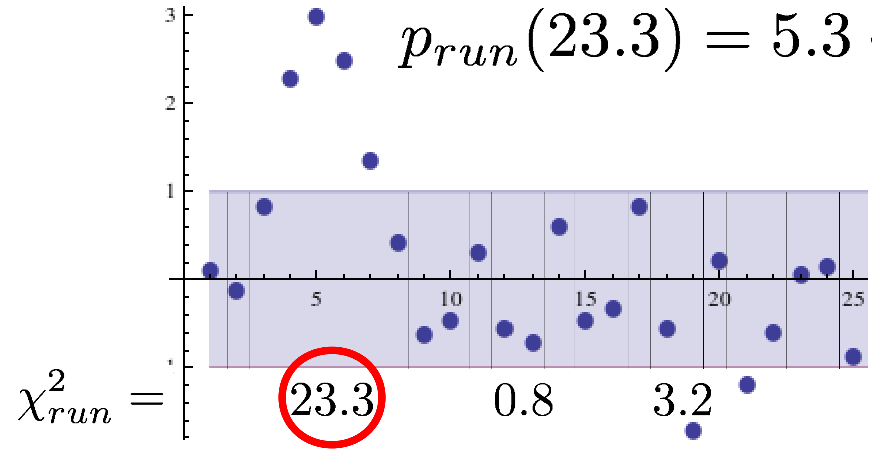
Runs distribution

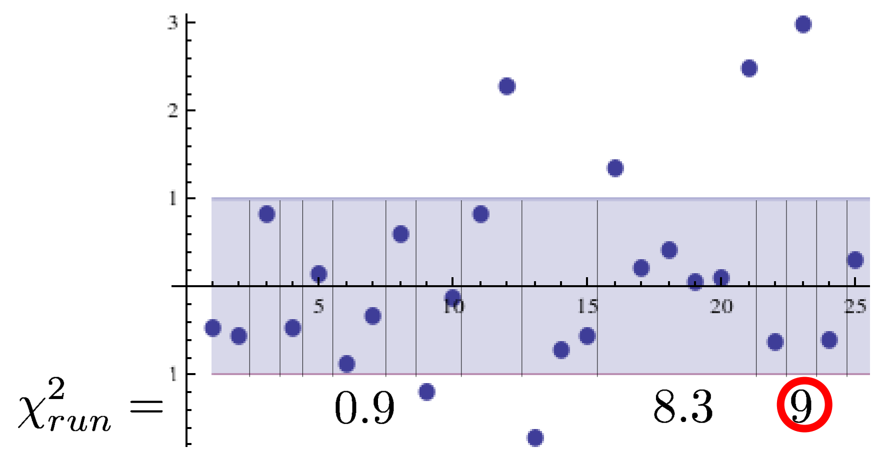
Gaussian case:

- Distribution of *T* exactly calculated for any *N* (non-parametric)
- Requires sum over integer partitions



$$p_{run}(23.3) = 5.3 \cdot 10^{-4}$$





$$p_{run}(9) = 1.3 \cdot 10^{-1}$$

Integer Partitions

- Number of ways to write integer *n* as sum of [*k*] integers
- Example

$$\begin{aligned} 5 &= 5 \\ &= 4+1 \\ &= 3+2 \\ &= 3+1+1 \\ &= 2+2+1 \\ &= 2+1+1+1 \\ &= 1+1+1+1+1 \end{aligned} \quad \Rightarrow \mathrm{Part}\,(5,3) = 2$$

$$\Rightarrow \mathrm{Part}\,(5) = 7$$

- Investigated by famous people (Leibniz, Euler ...)
- Related to strings, solid state, group theory ...

Partition triangle

Part(n,k)

k=1, 2, 3 ...

n=1, 2, 3...

1																								
1	1																							
1	1	1																						
1	2	1	1																					
1	2	2	1	1																				
1	3	3	2	1	1																			
1	3	4	3	2	1	1																		
1	4	5	5	3	2	1	1																	
1	4	7	6	5	3	2	1	1																
1	5	8	9	7	5	3	2	1	1															
1	5	10	11	10	7	5	3	2	1	1														
1	6	12	15	13	11	7	5	3	2	1	1													
1	6	14	18	18	14	11	7	5	3	2	1	1												
1	7	16	23	23	20	15	11	7	5	3	2	1	1											
1	7	19	27	30	26	21	15	11	7	5	3	2	1	1										
1	8	21	34	37	35	28	22	15	11	7	5	3	2	1	1									
1	8	24	39	47	44	38	29	22	15	11	7	5	3	2	1	1								
1	9	27	47	57	58	49	40	30	22	15	11	7	5	3	2	1	1							
1	9	30	54	70	71	65	52	41	30	22	15	11	7	5	3	2	1	1						
1	10	33	64	84	90	82	70	54	42	30	22	15	11	7	5	3	2	1	1					
1	10	37	72	101	110	105	89	73	55	42	30	22	15	11	7	5	3	2	1	1				
1	11	40	84	119	136	131	116	94	75	56	42	30	22	15	11	7	5	3	2	1	1			
1	11	44	94	141	163	164	146	123	97	76	56	42	30	22	15	11	7	5	3	2	1	1		
1	12	48	108	164	199	201	186	157	128	99	77	56	42	30	22	15	11	7	5	3	2	1	1	
1	12	52	120	192	235	248	230	201	164	131	100	77	56	42	30	22	15	11	7	5	3	2	1	1

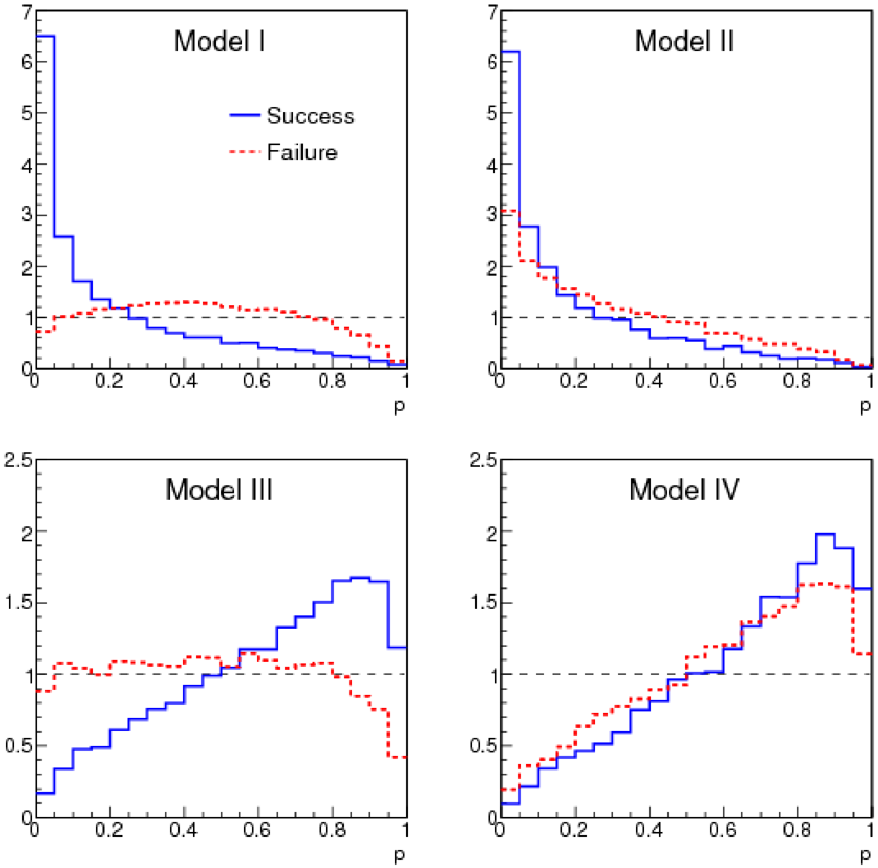
Part(n)

1
1 1
1 1 1
1 2 1 1
1 2 2 1 1 = 41 = Part(9)-1
1 3 3 2
1 3 4
1 4
1

$$\text{Proposition: } Part(n) - 1 = \sum_{r=1}^{N} \sum_{M=1}^{\min(r,N-r+1)} Part(n,k)$$

Runs distribution

Small range, MCMC



- Good model:

a) fitting bias towards p=1

b) Success and failure similar

- Bad model:

a) Success and failure different

b) Bias towards p=0

c) Missed a peak: failures OK

Goodness of Fit: Bayesian approach

Model selection:

- Need explicit alternatives M_1, M_2
- Posterior odds

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)} \times \frac{P(D|M_1)}{P(D|M_2)}$$

Bayes factor:

- (very) sensitive to parameter range

$$P(D|M_1) = \int p(D\,|\,\vec{\lambda})\; p_0(\vec{\lambda})\; d\vec{\lambda}$$

- Occam's razor built in

Example:

- Six (NP) vs three (SM) parameters

$$\frac{P(\mathrm{SM}|D)}{P(\mathrm{NP}|D)} = \frac{P(\mathrm{SM})}{P(\mathrm{NP})} \times 61.7$$