MAX-PLANCK-GESELLSCHAFT

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

A Test Statistic for Weighted Runs

Frederik Beaujean, Allen Caldwell

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

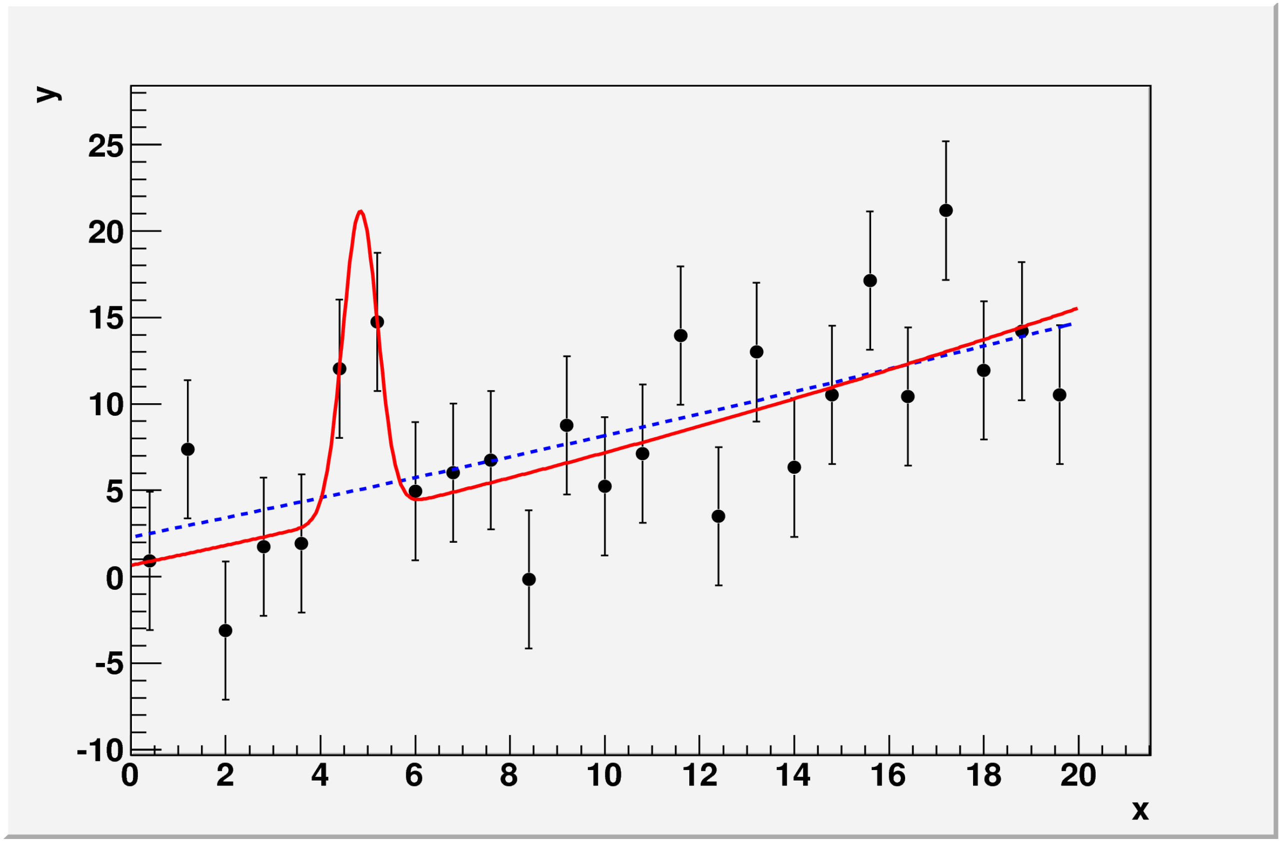
COMPSTAT 2010

Paris, 23.8.2010

Motivating example

Suppose:

- Measurements y_i with Gaussian uncertainty
- Standard Model (SM) background is quadratic
- New physics (NP) predicts signal peak



Goodness of Fit: standard approach

Test statistic:

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

Example:

- Prob. density of the data

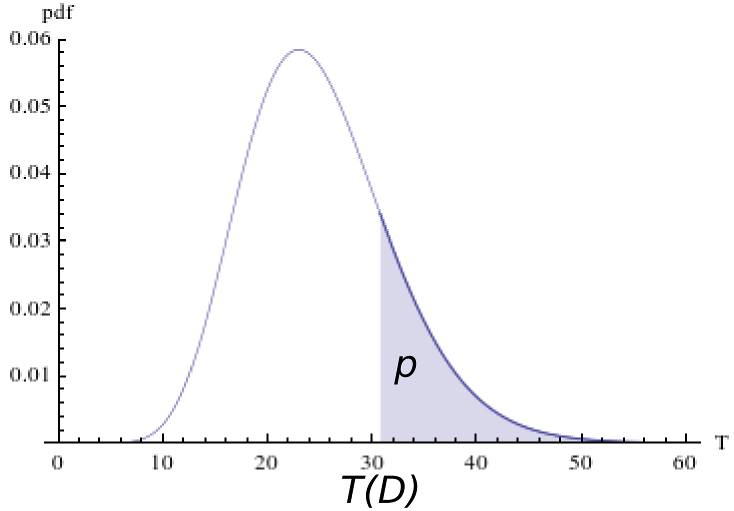
$$P(D|\vec{\lambda}) \propto \prod \exp\left\{ -\frac{\left(y_i - f(x_i|\vec{\lambda})\right)^2}{2\sigma_i^2} \right\} = \exp\left\{ \frac{-\chi^2}{2} \right\}$$

- Familiar choice

$$T(D) \equiv \chi^2(D)$$

p-value

Def: $p \equiv P(T > T(D))$



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

- Critical values: $p < 0.05, 0.01 \Rightarrow \text{reject model}$

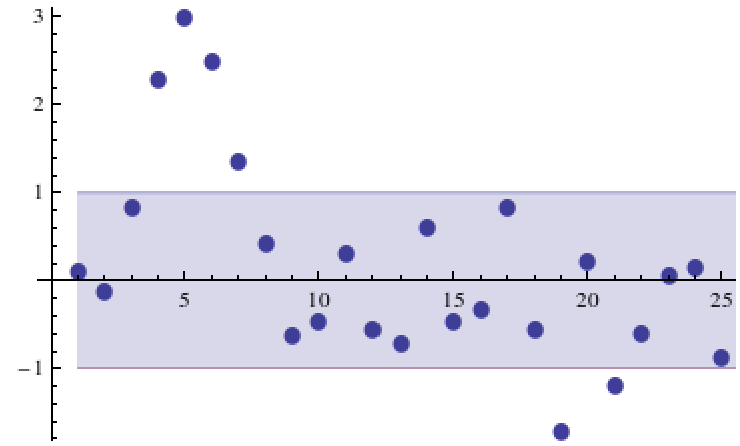
- Warning: p-value *not* the P. that the model is true

Example:

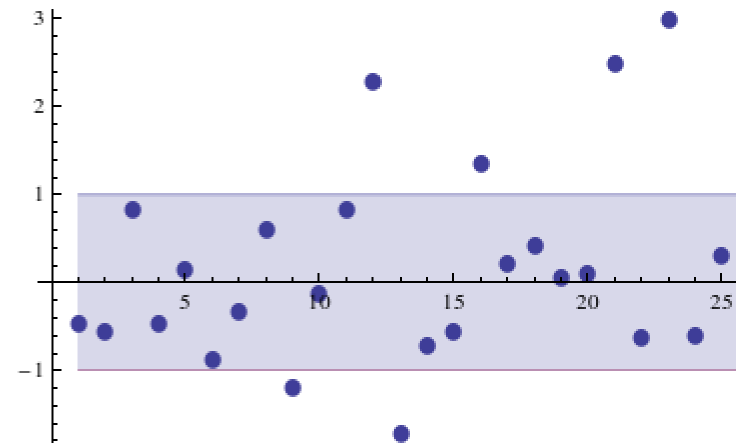
$p_{SM} = 10\%,\ p_{NP} = 37\% \ \Rightarrow$ both OK

Runs

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



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



Example:

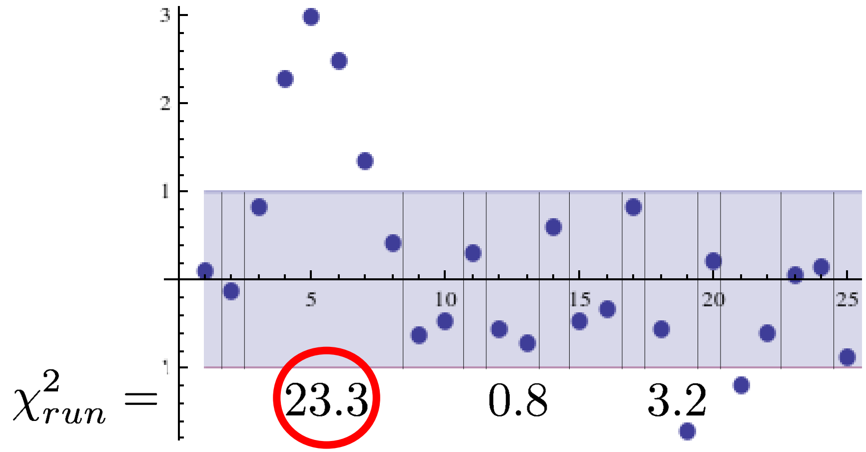
- N=25 datapoints
- Each Gaussian with mean = 0 and variance = 1

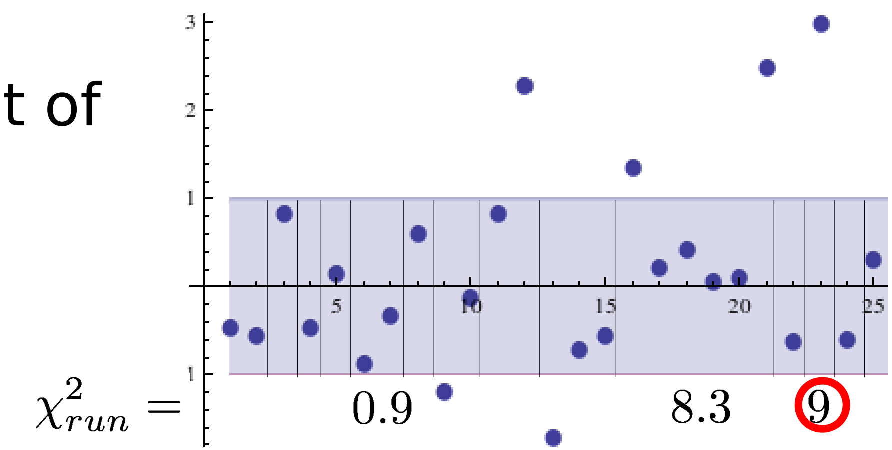
$\Longrightarrow$ Can we combine information about **order** and **magnitude of deviation**?

Runs statistic

Proposal:

- Split data into runs
- Each run has a weight
  Gaussian case: χ^2_{run}





- Test statistic: largest weight of any run

$$T \equiv \max\{\chi^2_{run}\}$$

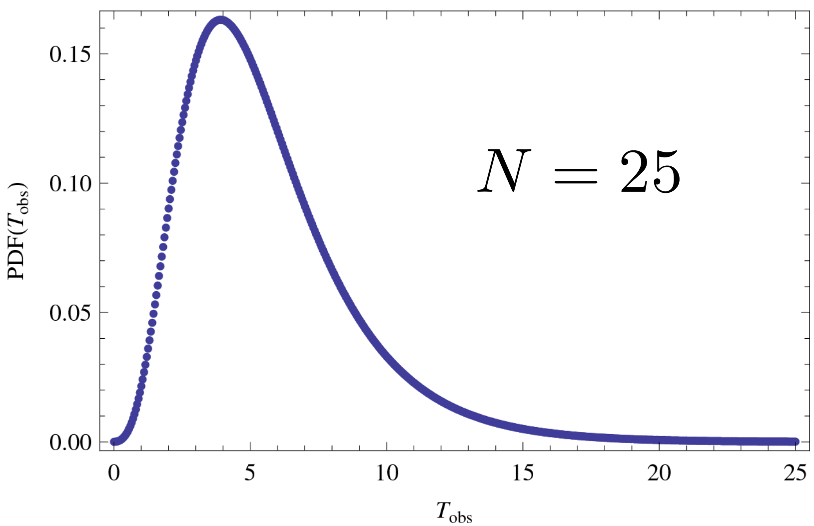
- p-value becomes

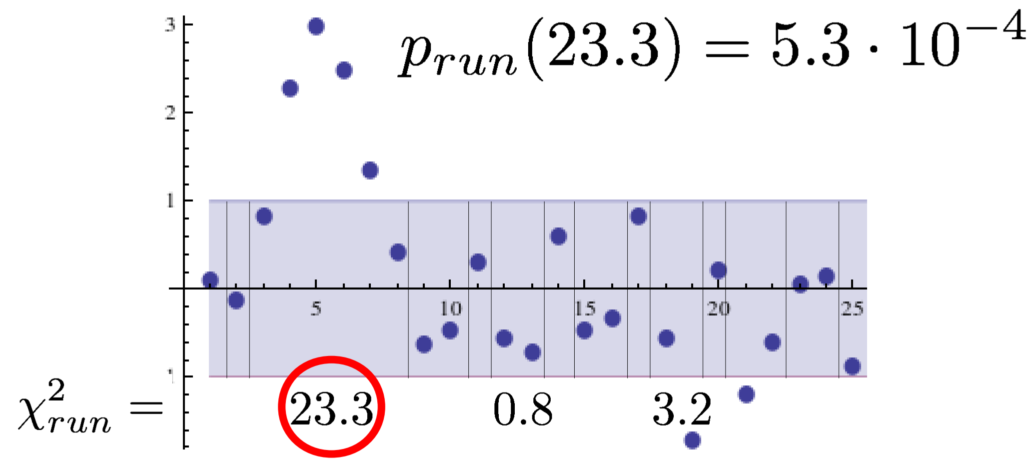
$$p_{run} \equiv P(T > T_{obs})$$

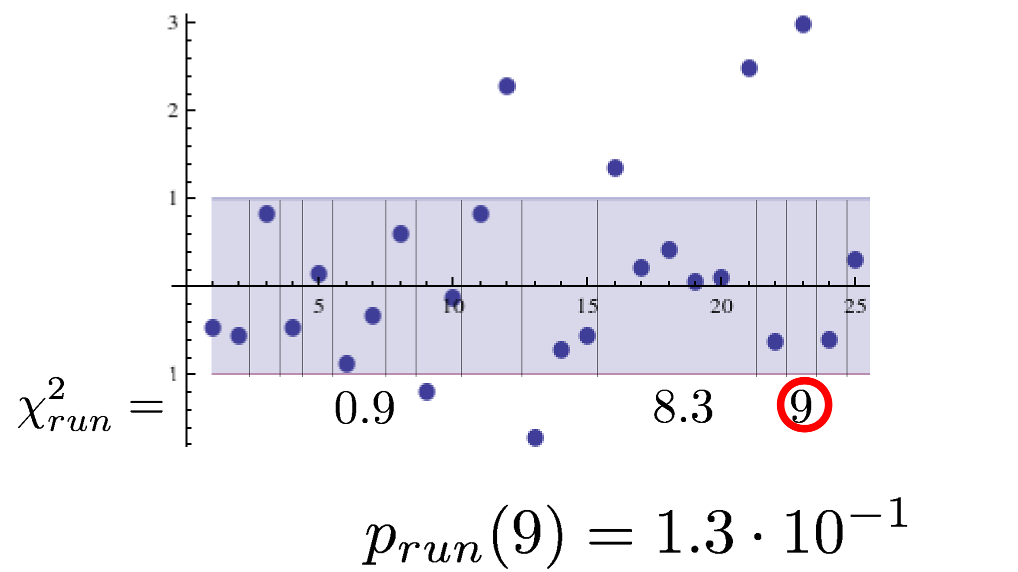
Runs distribution

Gaussian case:

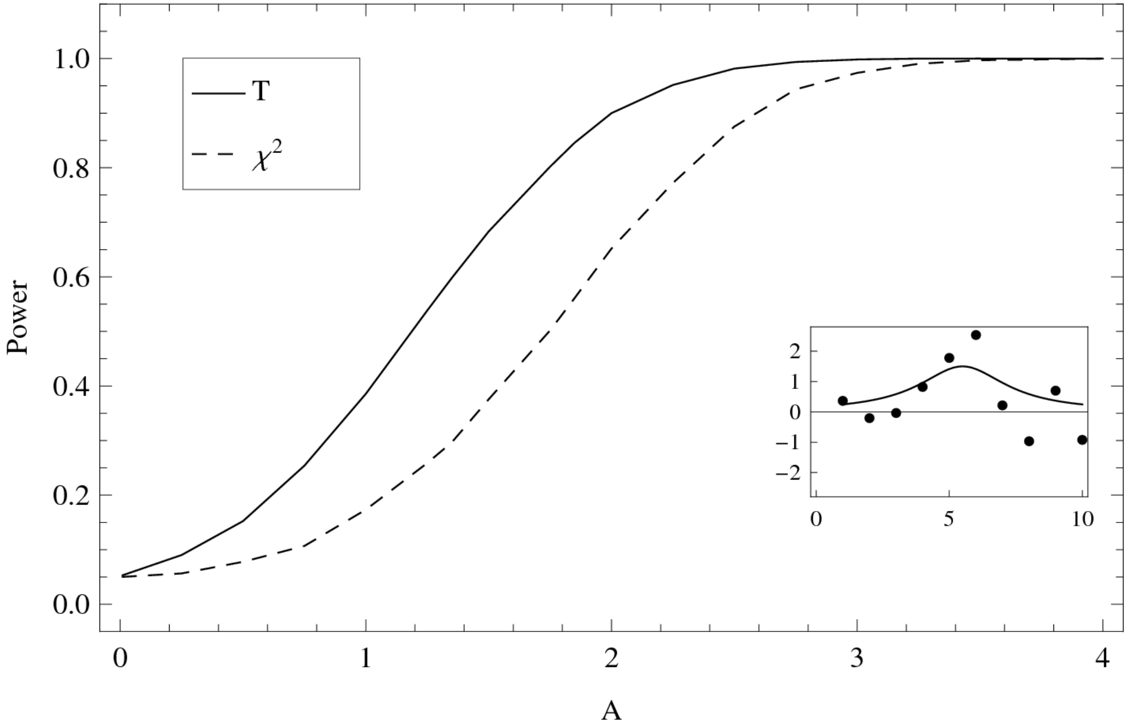
- Distribution of *T* exactly calculated for any *N* (non-parametric)
- Requires sum over integer partitions







Power



5% level

- New physics contribution:

$$y(x) = A \cdot \left(1 + \frac{(x-5.5)^2}{3^2}\right)^{-1}$$

Lorentz peak with amplitude A

- T up to 35% more powerful than classic χ^2 in detecting departures of type *y(x)*

Conclusions

- choose statistic with specific alternative models in mind
- Runs statistic T excellent for “bump hunting”

FINIS

Backup

Exact runs distribution I

$$P(T \geq T_{obs}|N) = 1 - P(T < T_{obs}|N)$$

$$P(T < T_{obs}|N) = \sum_{r=1}^{N} \sum_{M=1}^{\min(r,N-r+1)} P(T < T_{obs}|M,r,N) \cdot P(M,r|N)$$

$$\begin{aligned} M &= \text{number of success runs} \\ r &= \text{number of successes} \\ N &= \text{number of datapoints} \end{aligned}$$

$$P(M,r|N) = \frac{1}{2^N - 1} \cdot R(M|r,N)$$

$$R(M|r,N) = \text{number of permutations with } M \text{ runs given } r, N$$

Exact runs distribution II

$$P(T < T_{obs}|M, r, N) = \sum_{\pi} P(T < T_{obs}|\pi) P(\pi|M, r, N)$$

$$P(\pi|M, r, N) = \frac{W_\pi}{R(M|r, N)}$$

$$\pi = (r_1, \ldots r_l)$$

$$W_\pi = \binom{M}{r_1, \ldots, r_l} \cdot \binom{N - r + 1}{M} = \frac{(N - r + 1)!}{(N - r - M + 1)! \cdot \prod_l r_l!}$$

$$P(T < T_{obs}|\pi) = \prod_l \left[P\left(T < T_{obs}|l\right)^{r_l}\right]$$

Exact runs distribution III

$$P\left(T < T_{obs} | l\right) = \frac{\gamma(l/2, T_{obs}/2)}{\Gamma(l)}$$ CDF of χ^2 distribution

$$\gamma(a, x) = \int_0^x t^{a-1} e^{-t} \mathrm{d}t$$ Lower incomplete gamma function

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \mathrm{d}t$$ Complete gamma function

Computational complexity: Integer partitions

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
1	3	3	2	
1	3	4		
1	4			
1				

= 41 = Part(9)−1

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

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

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