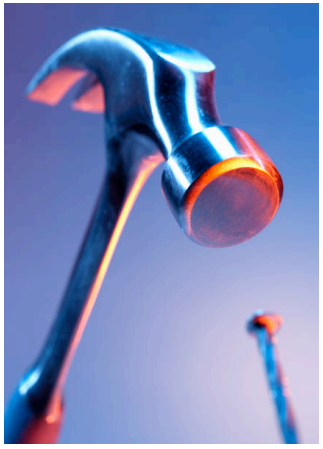
Banff Challenge 2

Tom Junk
Fermilab
BIRS Statistics in HEP Workshop
July 2010



Common Standards of Evidence

Physicists like to talk about how many "sigma" a result corresponds to and generally have less feel for p-values.

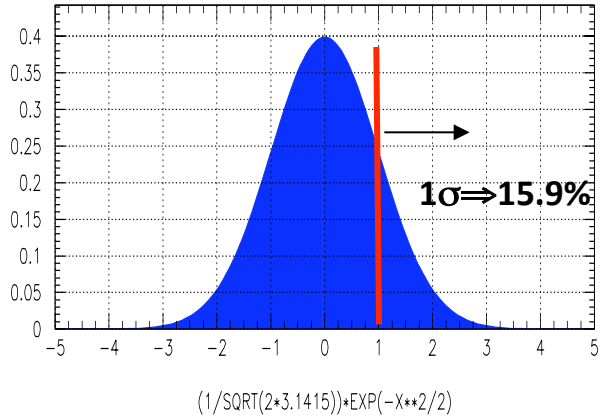
The number of "sigma" is called a "z-value" and is just a translation of a p-value using the integral of one tail of a Gaussian

Double_t zvalue = - TMath::NormQuantile(Double_t pvalue)

Folklore:
95% CL -- good for exclusion
3σ: "evidence"
5σ: "observation"
Some argue for a more subjective scale.

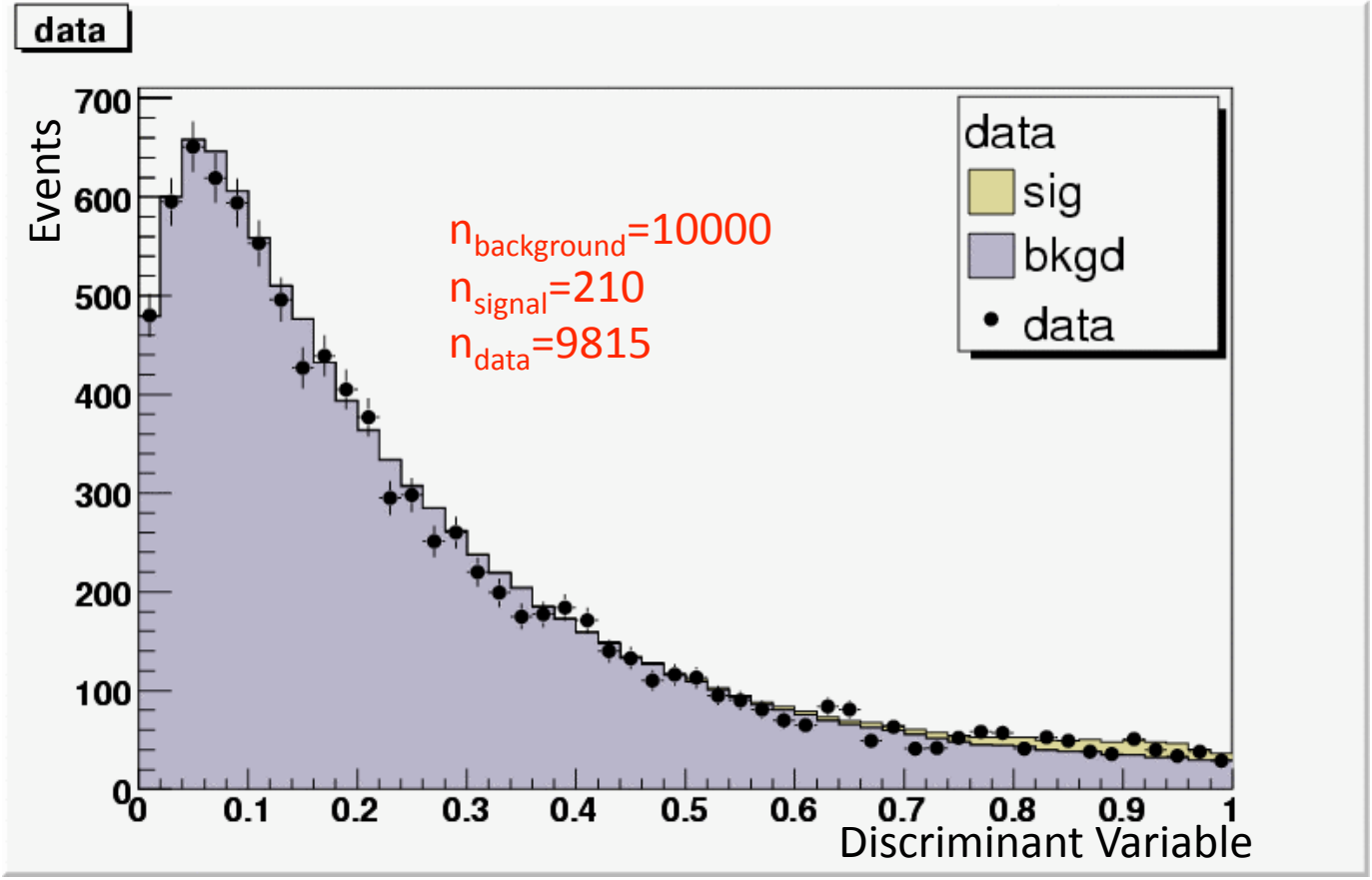
z-value (σ)	p-value
1.0	0.159
2.0	0.0228
3.0	0.00135
4.0	3.17E-5
5.0	2.87E-7

$$pvalue = \frac{\left(1 - erf\left(zvalue/\sqrt{2}\right)\right)}{2}$$



Tip: most physicists talk about p-values now but hardly use the term z-value

Banff Challenge 2 Problem #1 – Stacked plot shown HEP-style



- Observed data shown as points with sqrt(n) error bars (yes, the convention's crazy but that's the way we do it.)
- Signal prediction shown stacked on top of the background prediction.
  Useful because we can compare the the data with H_0 and H_1 with just one plot.

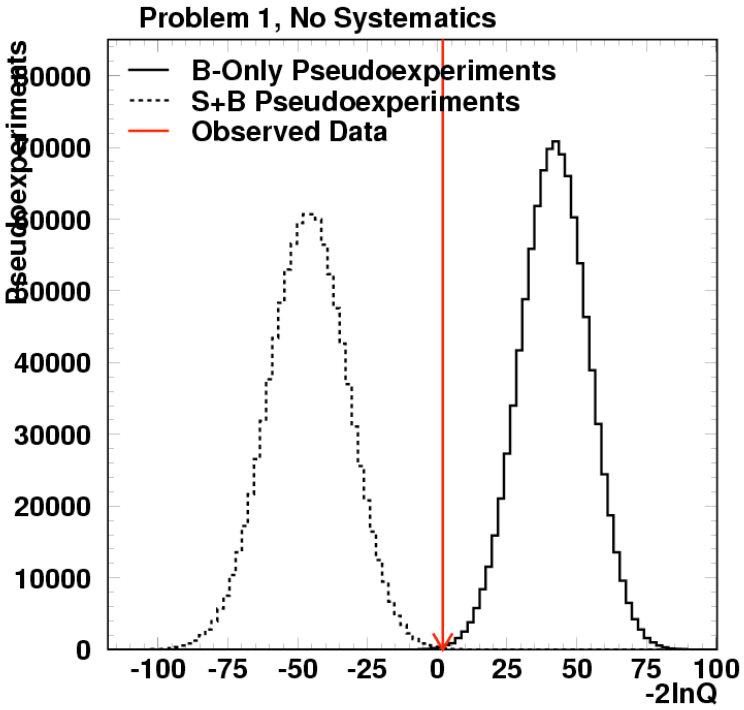
Problem 1, no systematic uncertainty

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

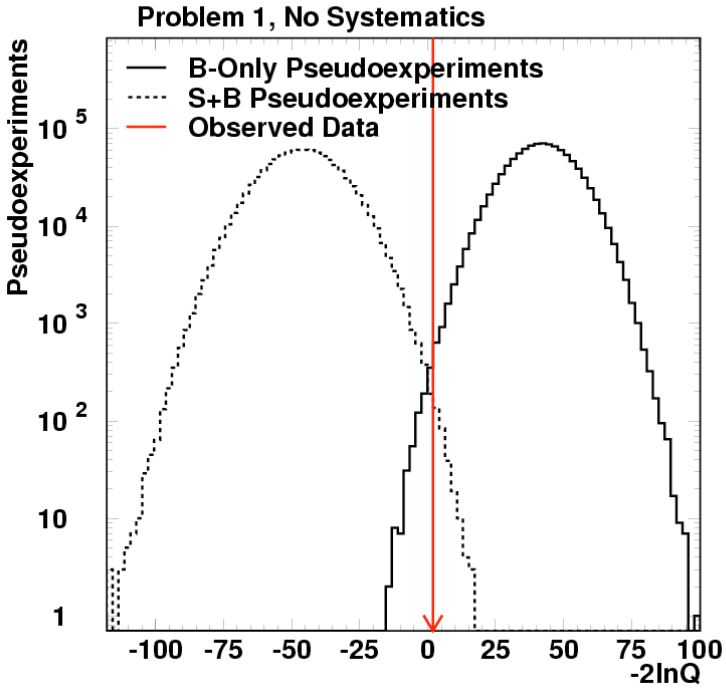
hats don't matter here since there's no fit.

1 Million simulated experiments for H_0 and
1 Million simulated experiments for H_1

Nuisance parameters always at their nominal values



$-2\ln Q_{obs} = 1.98$



p-value = 5.95×10^{-4}
z-value = 3.24

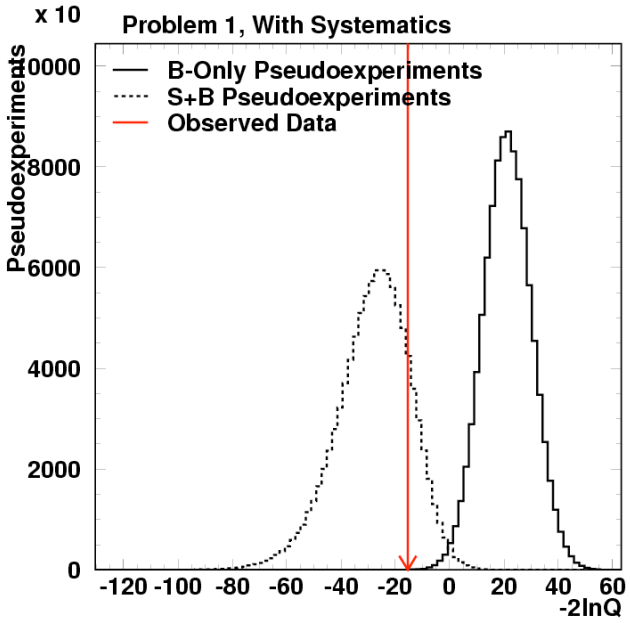
Problem 1, with systematic uncertainty

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

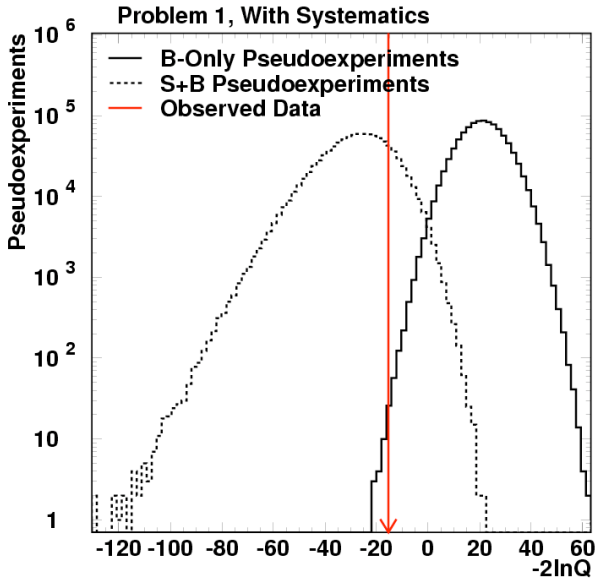
1 Million simulated experiments for H_0 and
1 Million simulated experiments for H_1

now do two fits per simulated experiment
-- fit for all nuisance parameters, rate and shape

Each pseudoexperiment gets randomly fluctuated nuisance parameters ("prior-predictive ensemble")



$-2\ln Q_{obs} = -15.43$

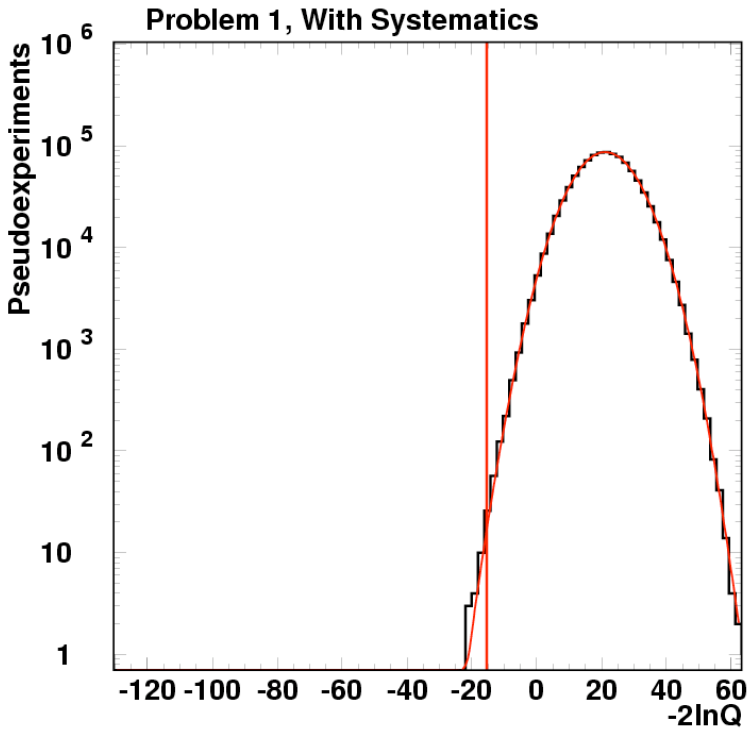


p-value = 1.91×10^{-5}
z-value = 4.11

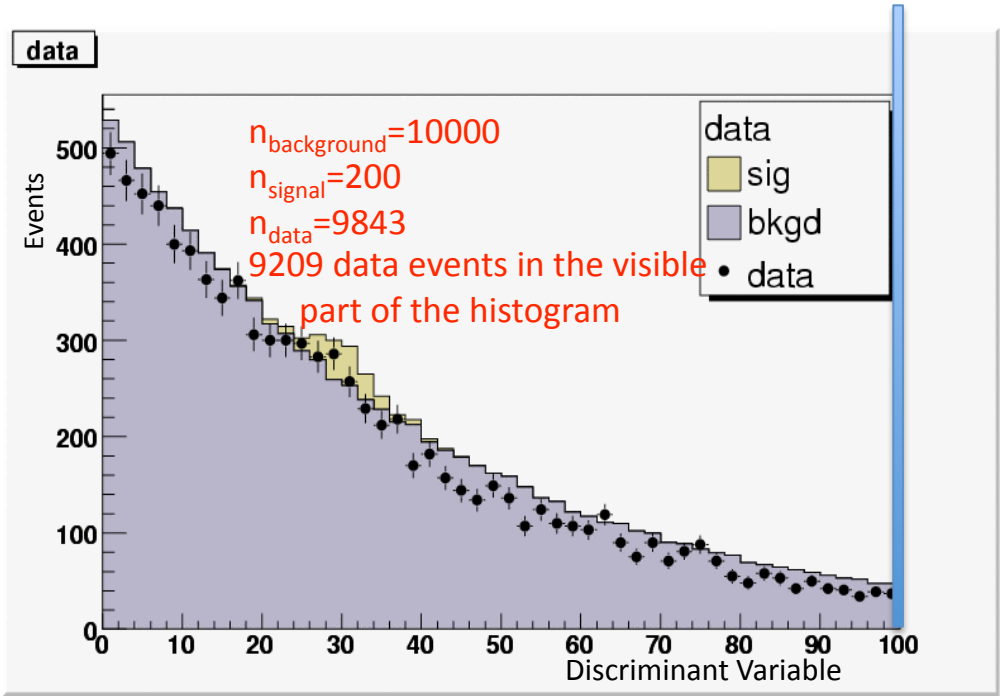
A Trick to Use only 1M Simulated H_0 Experiments

- Fit the distribution of -2lnQ(H_0) to a sum of two Gaussians – can integrate that analytically with erf's.
- Need to check fit quality. A real job would be to estimate the uncertainty (extrapolation uncertainty if need be).
- For a real discovery of a particle, we'd just use the needed CPU. Maybe the fitter gets stuck once in $1x10^7$ experiments – need to know that.

The sum of two Gaussians is a good approximation here
but a poor one if the problem
is more discrete – one bin,
for example, or lots of low s/b
bins and one very high s/b bin
with just a few expected events in it.



A Problem with Problem #2



634 observed events out of 9843 are in the upper overflow bin(!) (~6.44% of them)
Background template models this.

I discourage the use of ROOT's over- and underflow bins for several reasons:

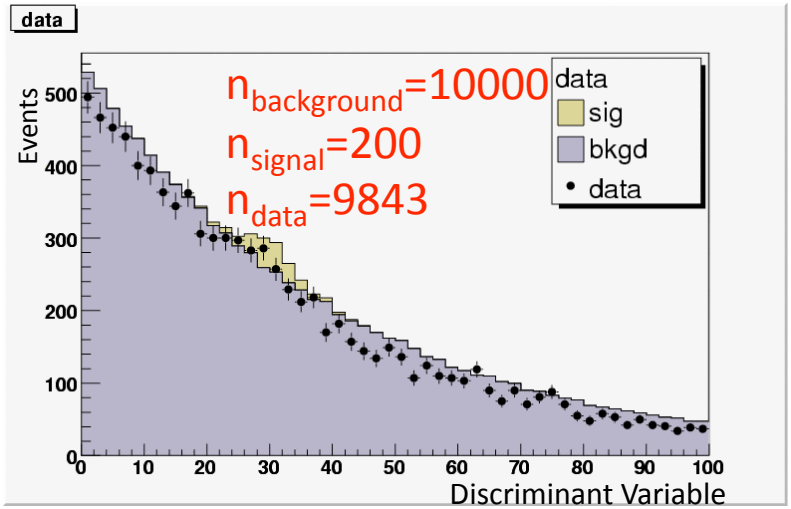
1) They are not (usually) plotted. Hard to validate them if you cannot see them
2) They are not included in TH1::Integral() or in fSumw when dumped. So scaling by dividing by the integral and multiplying by the desired yield won't get it right.

root accumulates entries beyond the histogram edges in underflow and overflow bins, and treats them as special bins (why?)
Suggestion to all students: constrain all selected data to be in visible bins (max and min).

Problems 1 and 3 have no entries in the underflow or overflow bins.

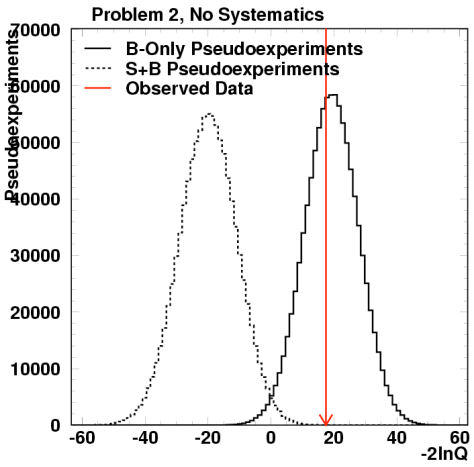
So I solved a problem that is slightly different and possibly more instructive.

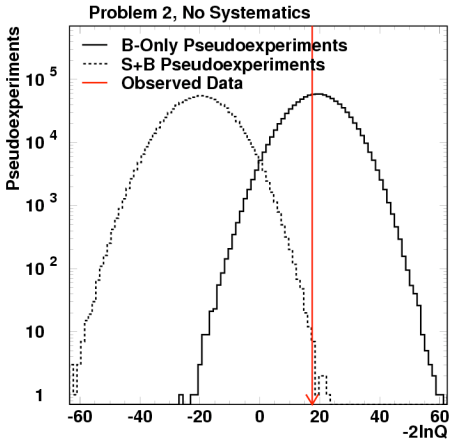
Problem 2



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

GOF not evaluated without systematics – pretty poor though. Shows that the no-systematics interpretation is incorrect.

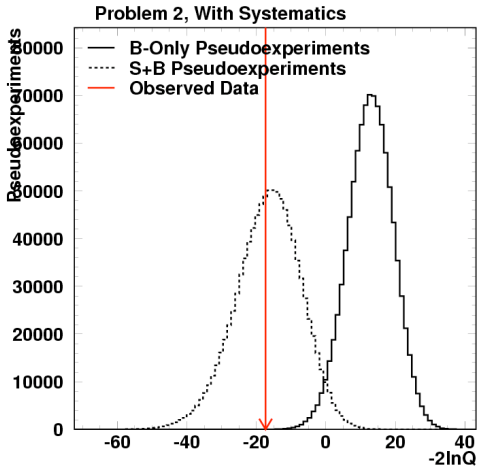


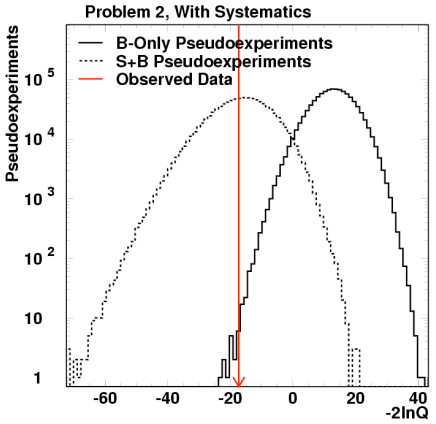


No systematics:

-2lnQ = 17.46 z-value = 0.20

p-value = 0.42





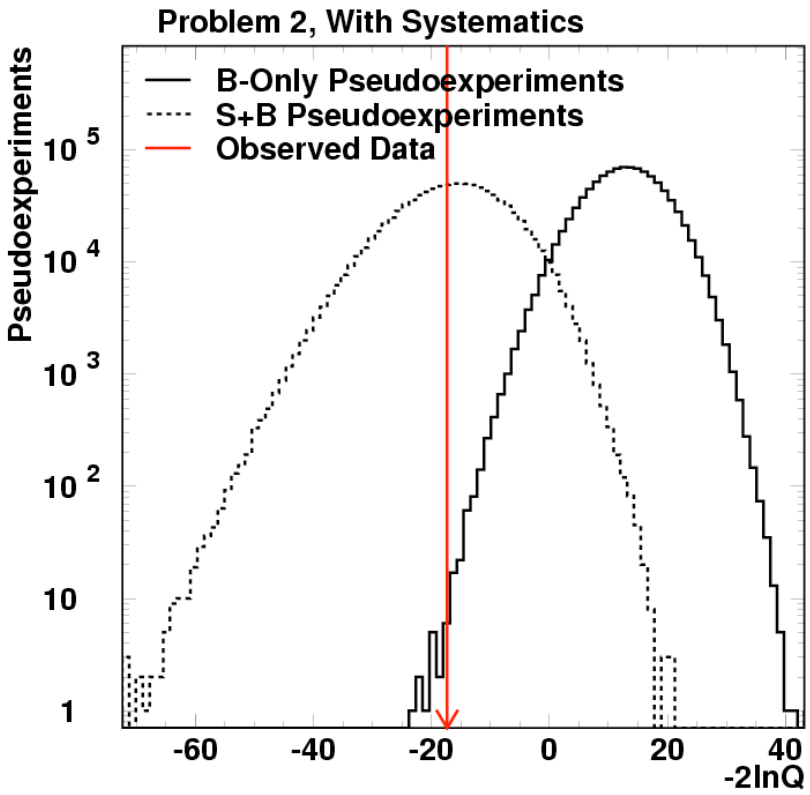
With systematics:

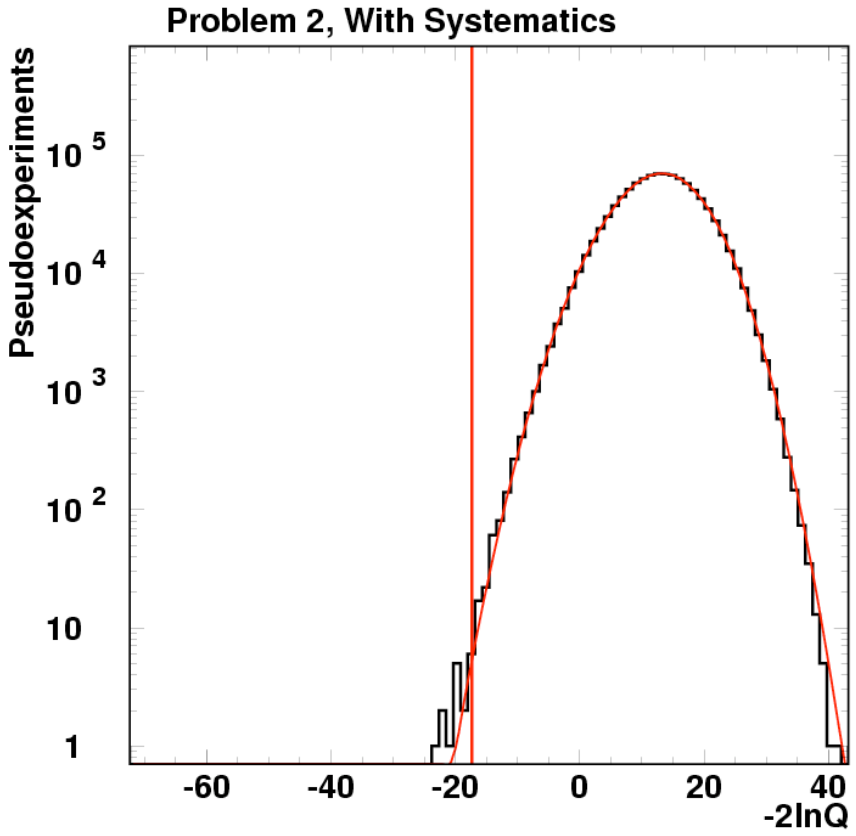
-2lnQ = -17.33 z-value = 0.20

p-value = 2.73×10^{-8}

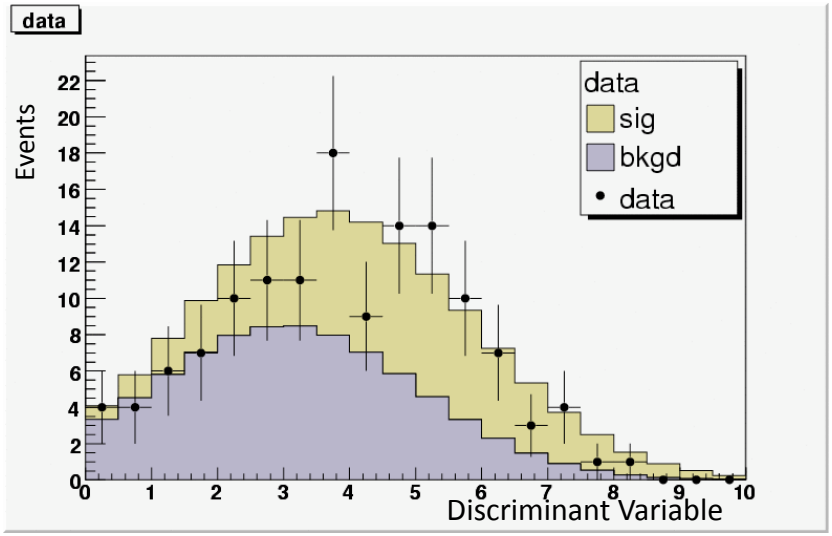
Problem 2's Fit

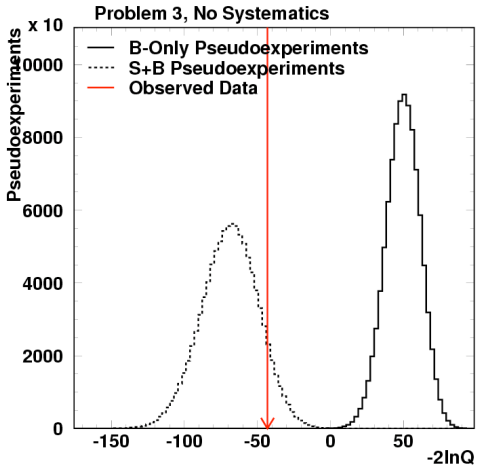
Not perfect on the tail, probably just need to run more pseudoexperiments

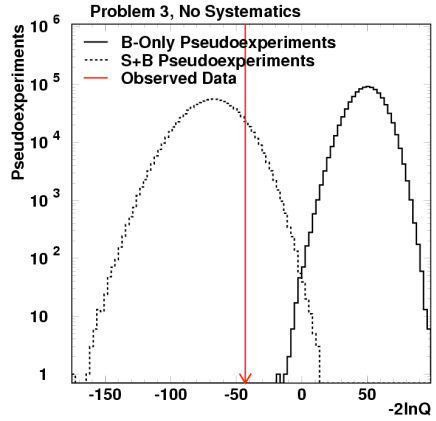




Problem 3







No systematics:

-2lnQ = -43.1 z-value = 7.3

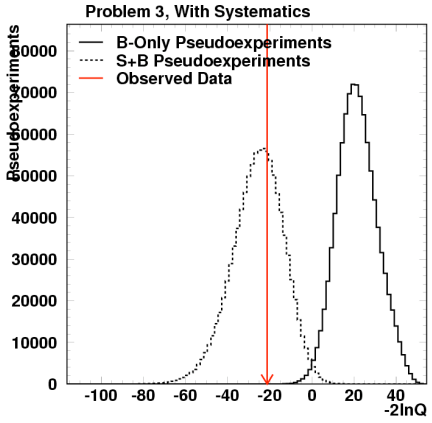
p-value = 1.4×10^{-13}

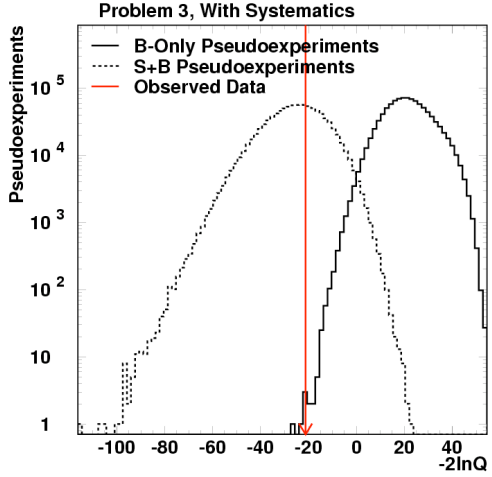
$n_{background}$=80

n_{signal}=72

n_{data}= 134

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$





With systematics:

-2lnQ = -21.2 z-value = 4.44

p-value = 4.5×10^{-6}

Combining Problems 1+2+3

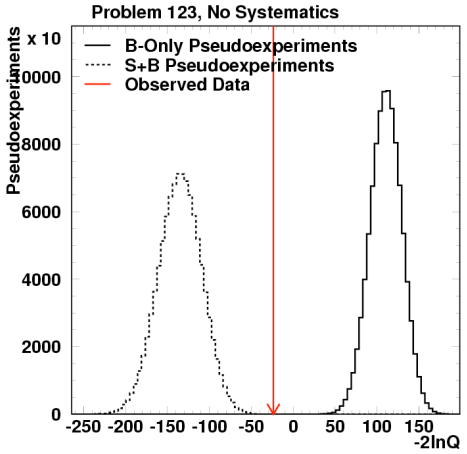
Joint fits – correlated systematic uncertainties in a real problem. We are told to decorrelate the nuisance parameters between channels.

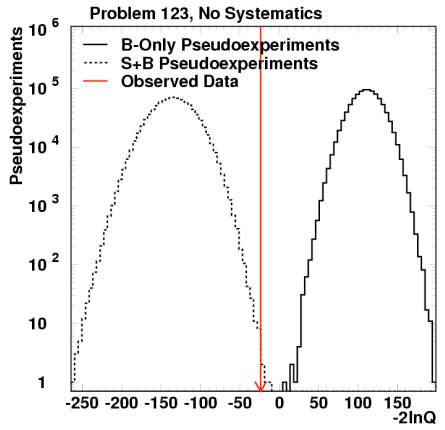
No systematics:

$-2\ln Q_{comb} = -2\ln Q_1 - 2\ln Q_2 - 2\ln Q_3$

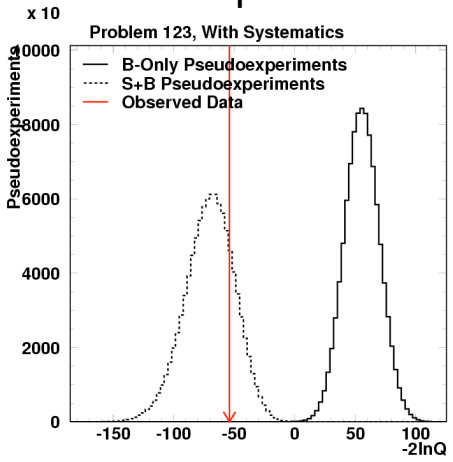
With systematics – spoiled a bit by the different fits, if nuisance parameters are correlated.
In this case the sum rule still works because all data and all nuisance parameters are independent.

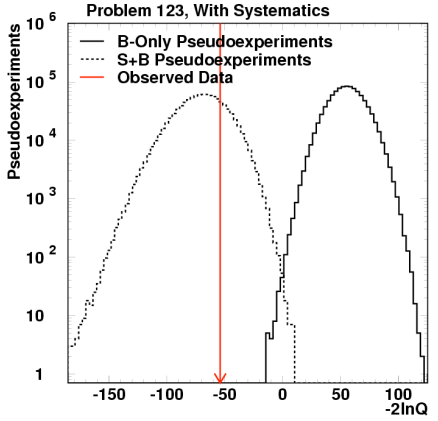
GOF is poor for both hypotheses – see prob. 2. Large sensitivity. No systematics → can rule out both H_0 and H_1.





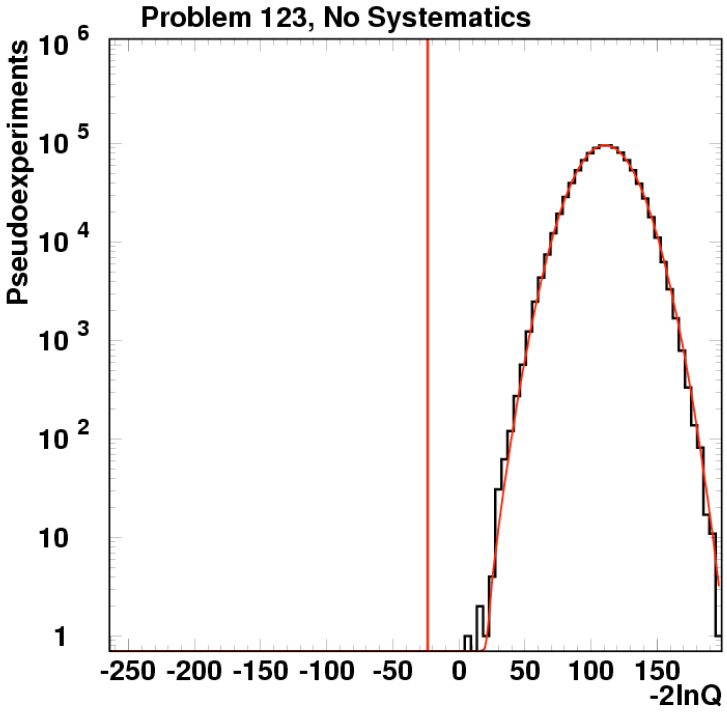
No systematics:
-2lnQ = -23.7 z-value = 7.0
p-value = 1.2×10^{-12}

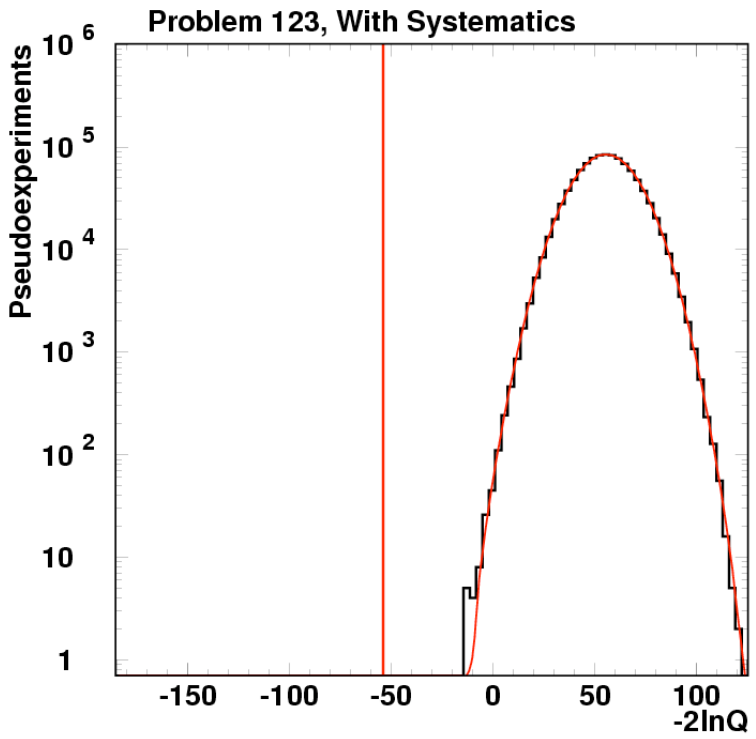




With systematics:
-2lnQ = -53.96 z-value = 7.5
p-value = 3.6×10^{-14}

Combination Significances are a Bit of an Extrapolation with just 1M Simulated Outcomes





Estimates of Sensitivity

- Well, 1 Million simulated experiments isn't enough –
  can get 1 Million more in combination by adding the -2lnQ's from 1, 2, and 3's together.

- Wilks's Theorem probably is a good approximation here too.

- Importance sampling could be used to improve precision in tails

- For discovery, we'd use real CPU as the systeamtics will be correlated and there may be a single bin adding a discrete component to it.

- Our favorite sensitivity estimate: $p_{med,signal}$ is the median expected p-value assuming a signal is present. 1M pseudoexperiments not quite enough.

- A stand-in: the "*o*-value" (named by the CDF Karlsruhe single top team, but we'd used it before.

$$o - \text{value} = \frac{\left(\langle -2\ln Q\rangle_{bkg} - \langle -2\ln Q\rangle_{s+b}\right)}{\sqrt{\sigma^2_{bkg} + \sigma^2_{s+b}}}$$

- Medians can be used instead of means, and the σ's are RMS's of the -2lnQ distribution.

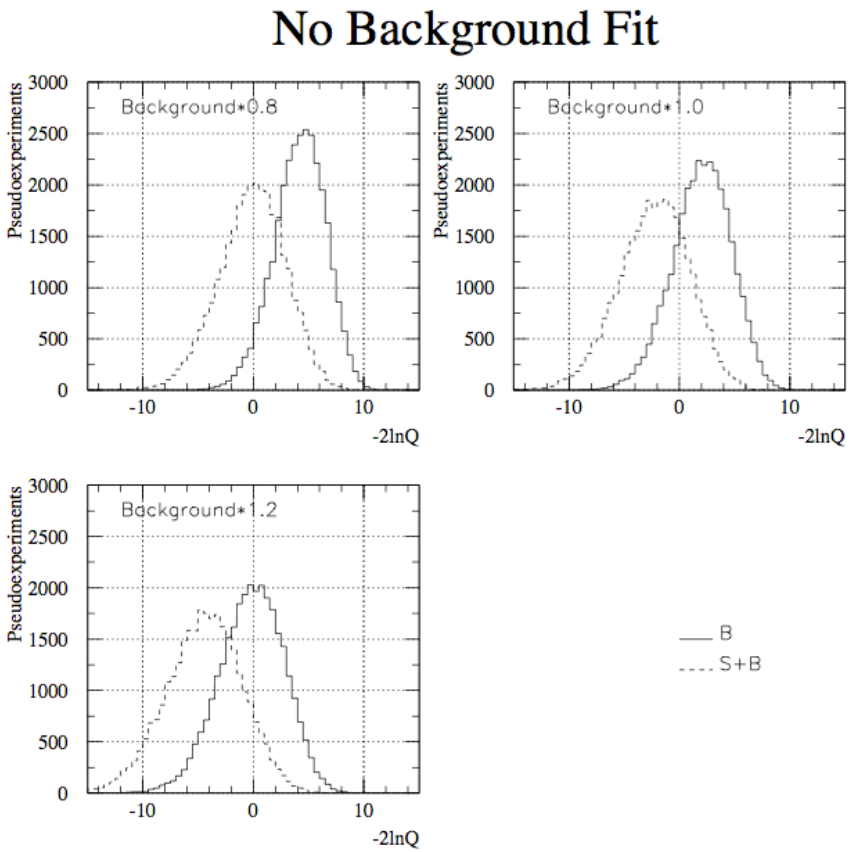
$$o - \text{value} = \frac{\left(\langle -2\ln Q\rangle_{bkg} - \langle -2\ln Q\rangle_{s+b}\right)}{\sqrt{\sigma^2_{bkg} + \sigma^2_{s+b}}}$$

Problem	$<-2\ln Q>_b$	RMS_b	$<-2\ln Q>_{s+b}$	RMS_{s+b}	*o*-value
1 no syst	41.9	12.3	-46.3	14.3	4.7
2 no syst	19.1	8.6	-20.0	9.1	3.1
3 no syst	49.5	11.9	-68.9	19.5	5.2
123 no syst	110.6	19.1	-135.2	25.8	7.6
1 syst	21.2	8.9	-28.1	13.5	3.0
2 syst	12.8	6.6	-16.7	9.3	2.6
3 syst	21.6	9.5	-25.6	12.2	3.0
123 syst	55.5	14.6	-70.3	20.5	5.0

To a good approximation, *o*-values add in quadrature for the combination.
True for this problem, but not true in general.

Backup Material

Fitting Nuisance Parameters to Reduce Sensitivity to Mismodeling



Means of PDF's of -2lnQ very sensitive to background rate estimation.

Still some sensitivity in PDF's residual due to prob. of each outcome varies with bg estimate.