

Banff Challenge 2a Results

Tom Junk
Fermilab

PHYSTAT2011
CERN
18 January 2011

- The Challenge: Signal Significance
- Designing the Problems
- Figures of Merit
- Submissions Received and Performance

A Little History

Banff Challenge 1: Upper limits on a signal in a Poisson counting experiment with a background predicted with the help of another Poisson counting experiment

From the Banff 2006 workshop. Files and results are available at

<http://newton.hep.upenn.edu/~heinrich/birs/>

<http://doc.cern.ch/yellowrep/2008/2008-001/p125.pdf>

The idea: A open forum to test coverage and power using data samples with and without signal present

Blind: signal “answers” hidden from the participants.

Many different strategies used: Frequentist, Bayesian, mixtures....

Quite a success! But...

We'd also like to make discoveries. Only of particles/phenomena that are truly there of course.

Additional issues must be addressed in discovery that aren't relevant for limits.

Original Banff Challenge 2

- Discovery significance with a realistic set of problems
 - Distributions (histograms, or unbinned data) that look like modern multivariate analysis outputs
 - A distribution that has a bump in it (or not)
 - Rate and Shape uncertainties in distributions – some quite large

All of these features are present in most modern searches at hadron colliders.

Often the signal yield is smaller than the uncertainty on the background (as determined from theory or control samples)

Not very many solutions attempted by the time of the workshop (Summer 2010)

Topics discussed at the workshop needed to be addressed in the challenge.

The challenge needed to be advertised to a larger audience

Discovery Issues Addressed by Banff Challenge 2a

- The “Look Elsewhere Effect”, which also goes by these names
 - Trials Factor
 - Multiple Testing / Multiple Comparisons
- Note: setting limits doesn’t really have a LEE although there is a multiple testing effect that makes mass limits weaker
- Dealing with nuisance parameters which are poorly constrained by auxiliary experiments (or theory)
- Large vs. small numbers of events of signal and background – in the same distribution!
- Finite Monte Carlo parameterizations of the expected signal and background predictions
- Measurements – cross sections and masses (“point estimation” in statistics jargon) along with hypothesis testing
- Estimations of power of the test by the analyzer

Discovery Issues Not Addressed by Banff Challenge 2a

- 3σ “evidence” and 5σ “discovery” levels. To keep the computation and storage requirements under control, we asked for a Type-I error rate (= false discovery rate) of no more than 0.01. Far fewer simulated data sets are needed to measure this rate than 3σ and 5σ significances
- Multiple channels – a feature of Banff Challenge 2 but not 2a.
- Shape uncertainties – also a feature of BC2 but not 2a.
- Many components of signal and background – BC2a problem 2 has two kinds of background, while a “real” search can have 10 or more
- Many nuisance parameters

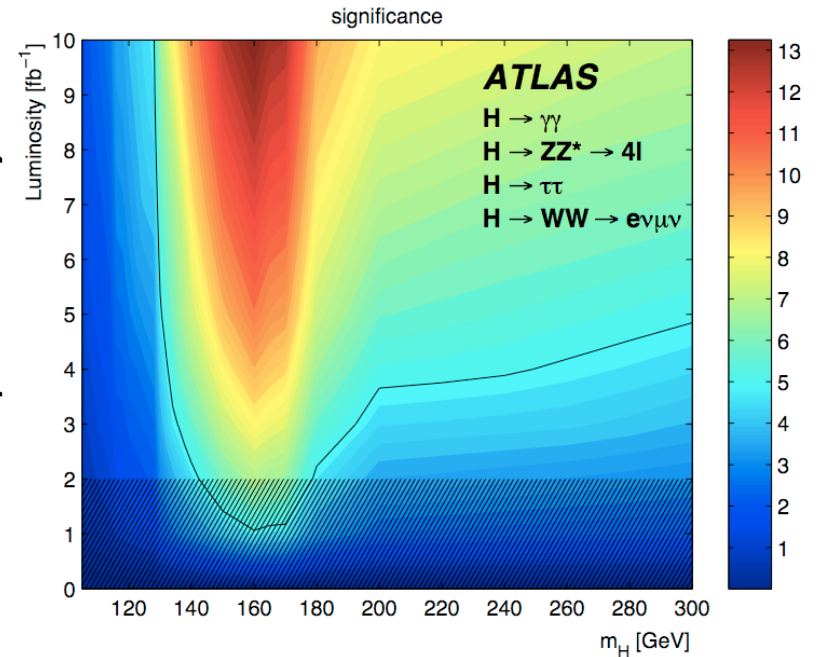
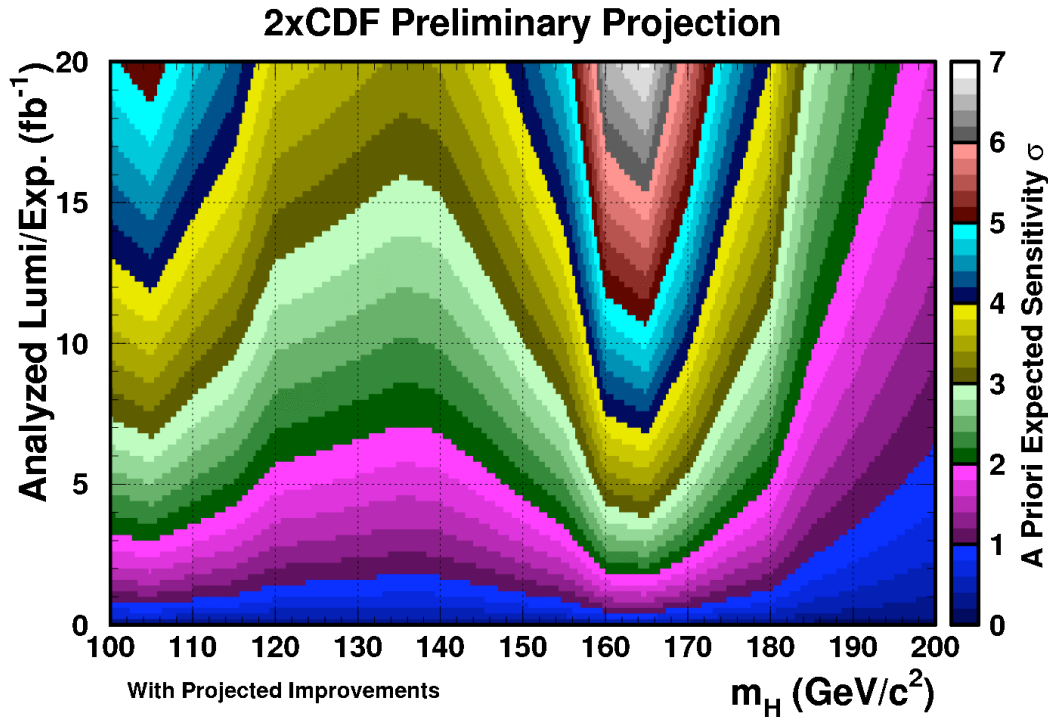
For example, a search for a new particle that decays to W +jets has as backgrounds:

Wbb , Wcc , Wc , $W+LF$, Z +jets (bb , cc , LF), WW , WZ , ZZ , $t\bar{t}$, single top (s , t , tW), non W and these often can be broken into subcategories.

These last four are judgment and bookkeeping exercises – defining an appropriate set of nuisance parameters and applying them properly to all predictions is part of the daily business of a HEP experimentalist. Discoveries often hinge on them.

Importance of Power Estimations

Example: Higgs Sensitivity vs. Integrated Lumi



CERN-OPEN-2008-020, "Expected Performance of the ATLAS Experiment": arXiv:0901.0512v3 [hep-ex]

Plots like these are inputs to decision-making processes!

It is important not to overstate or understate sensitivity (overstating it is worse).

Figures of Merit – how to Pick a Method

I don't speak for everyone, but this is my preference:

- 1) Method should have coverage at the stated level (Type-I error rate $\leq 1\%$ when we use that as a target. 3σ and 5σ are more customary and coverage should hold for those thresholds as well) A “yes-no” issue.
- 2) The quoted sensitivity should not be an overestimate. A “yes-no” issue.
- 3) The most powerful quoted sensitivity then helps us select a method.
- 4) We may use different methods for measuring peak positions and cross sections that have little to do with the method used to make the hypothesis test.
-- Coverage first, then estimated power.

Usual: “check the pulls” and work at it until you get it right. Feldman-Cousins provides a way of calibrating all measurements so coverage is okay – you can always add that as a last step.

Documentation, Challenge Datasets, and Submission Descriptions

The problem specifications and data files with challenge datasets and Monte Carlo signal and background templates:

<http://www-cdf.fnal.gov/~trj/>

Summary note, submitted entry descriptions, the “answer keys” and the programs used to generate the challenge datasets:

<http://www-cdf.fnal.gov/~trj/bc2sub/bc2sub.html>

A Big Thanks to All of the Participants

Participant	Problem 1	Problem 2
Tom Junk	LLR+MINUIT+MC	LLR+MCLIMIT binned
Wolfgang Rolke	LLR+MINUIT+MC	LLR+MINUIT parameterized marks
Valentin Niess	Windowed event counting	KS Test + MC parameterized marks
Stefan Schmitt	Fractional Event Counting aux MC to correct p values for LEE	Fractional Event Counting binned marks
Stanford Challenge Team B. Efron, T. Hastie, O. Muralidharan, B. Narasimhan, J. Scargle, Robert Tibshirani, Ryan Tibshirani	LLR+MC "Lindesy's method" Binned Poisson	LLR+MC Parameterized marks

More complete descriptions are available from the participants notes

"LLR" =
log likelihood
ratio
"-2lnQ"
" $\Delta \log \lambda$ "

Continued
on next page

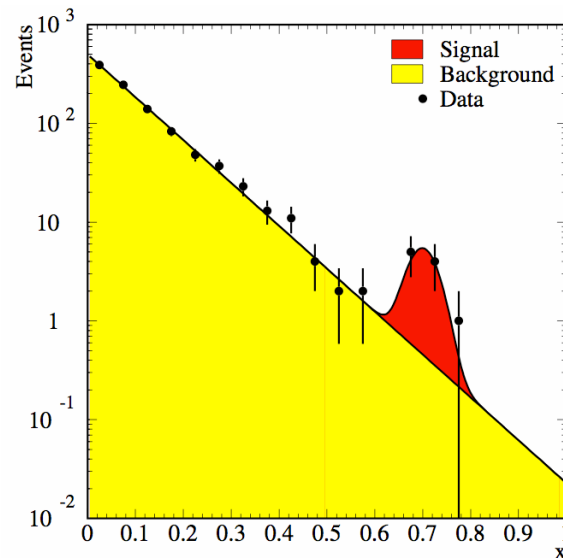
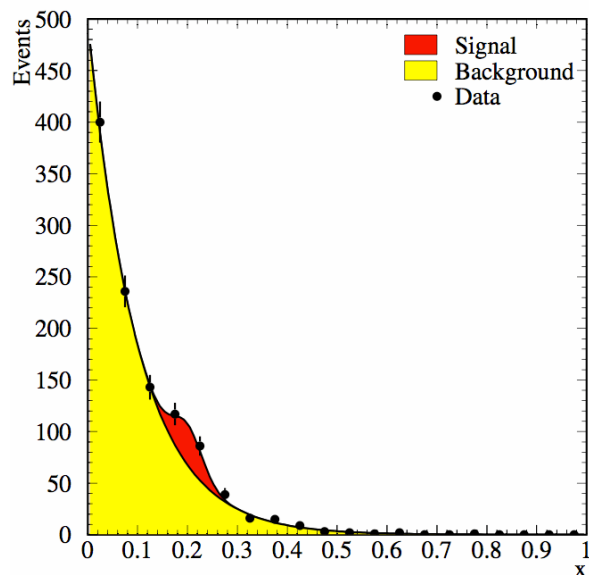
A Big Thanks to All of the Participants

Participant	Problem 1	Problem 2
Ofer Vitells and Eilam Gross	LLR + explicit LEE-corrected p-values from a χ^2 distrib LEE corr calibrated with MC	LLR + χ^2 distrib p values binned marks
BAT Team: F. Beaujean, A. Caldwell, S. Pashapour	Bayesian LEE accounted for in prior p-value first, then further Bayesian selection	
Georgios Choudalakis	BumpHunter	
Mark Allen	LLR+MC – repeated fit with different initial params	
Matt Bellis and Doug Applegate		Bootstrapped nearest-neighbor + MC
Stefano Andreon	Bayesian – arb threshold	

Challenge Parameters – Problem 1

- Would like to measure Type-I error rates of $0.01 \pm 0.001 \rightarrow \mathcal{O}(10000)$ repetitions are needed. We picked 20K total samples, including those with signal.
- Would like a LEE of at least 10 – Signal peak width of 0.03 on a histogram of marks from 0 to 1.

(statistics jargon: A “mark” is a reconstructed quantity, like a measured mass or a neural network output. We have a “marked Poisson process”.)



Discoveries often look like these.

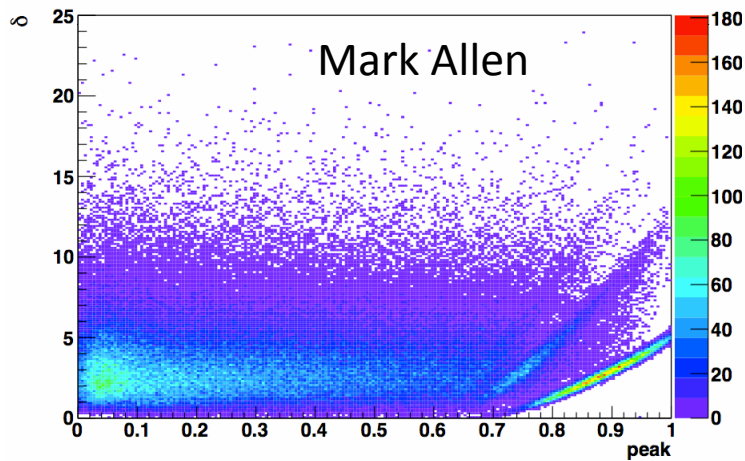
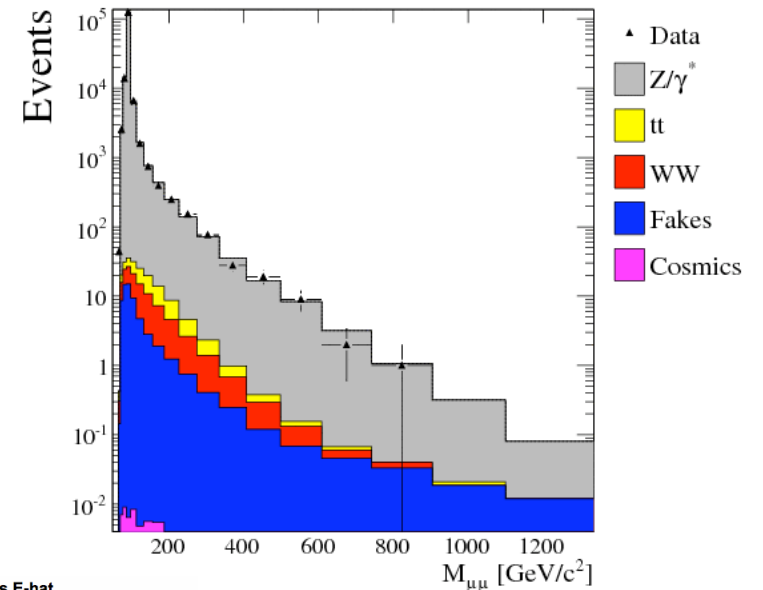
$\mathcal{O}(1000)$ events per simulated experiment to make “high statistics” and “low statistics” outcomes possible (see plots)

Problem 1 Has High-Statistics and Low-Statistics Extremes

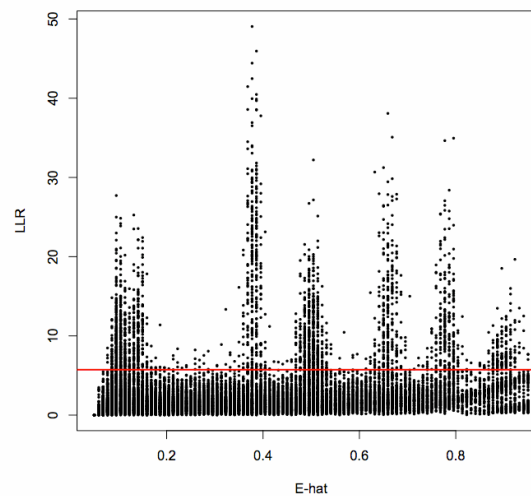
We have to deal with this all the time.

Example: CDF's Z' search in the $\mu^+\mu^-$ channel:
A real search has elements of both Problem 1 and Problem 2

CDF Run II Preliminary 4.6 fb⁻¹



Log Likelihood Ratio vs E-hat



Stanford Challenge Team (B. Efron *et al*)

Challenge Parameters – Problem 1

Background:

$$B(x) = Ae^{-Cx}$$

Signal:

$$S(x) = De^{-(x-E)^2 / 2\sigma^2}$$

$$A = 10000 \pm 1000$$

$$C = 10 \pm 0$$

$$\sigma = 0.03 \pm 0$$

D, E, unknown, but

$$D \geq 0$$

$$0 \leq E \leq 1$$

Category	E_{input}	D_{input}	n_{rep}
1	—	0.00	15400
2	0.50	83.78	200
3	0.38	265.96	200
4	0.10	1010.65	200
5	0.10	478.73	200
6	0.66	66.49	200
7	0.78	39.89	200
8	0.10	744.69	200
9	0.50	136.97	200
10	0.90	15.29	200
11	0.50	190.16	200
12	0.14	664.90	200
13	0.50	163.57	200
14	0.38	531.92	200
15	0.14	1196.83	200
16	0.50	110.37	200
17	0.10	1276.62	200
18	0.90	20.61	200
19	0.66	132.98	200
20	0.90	12.63	200
21	0.90	17.95	200
22	0.90	23.27	200
23	0.78	79.79	200
24	0.10	1542.58	200

Bump Locations and Rates – Problem 1

Wanted to test the entire range
(but didn't generate signals right
at the edges of the physical region)

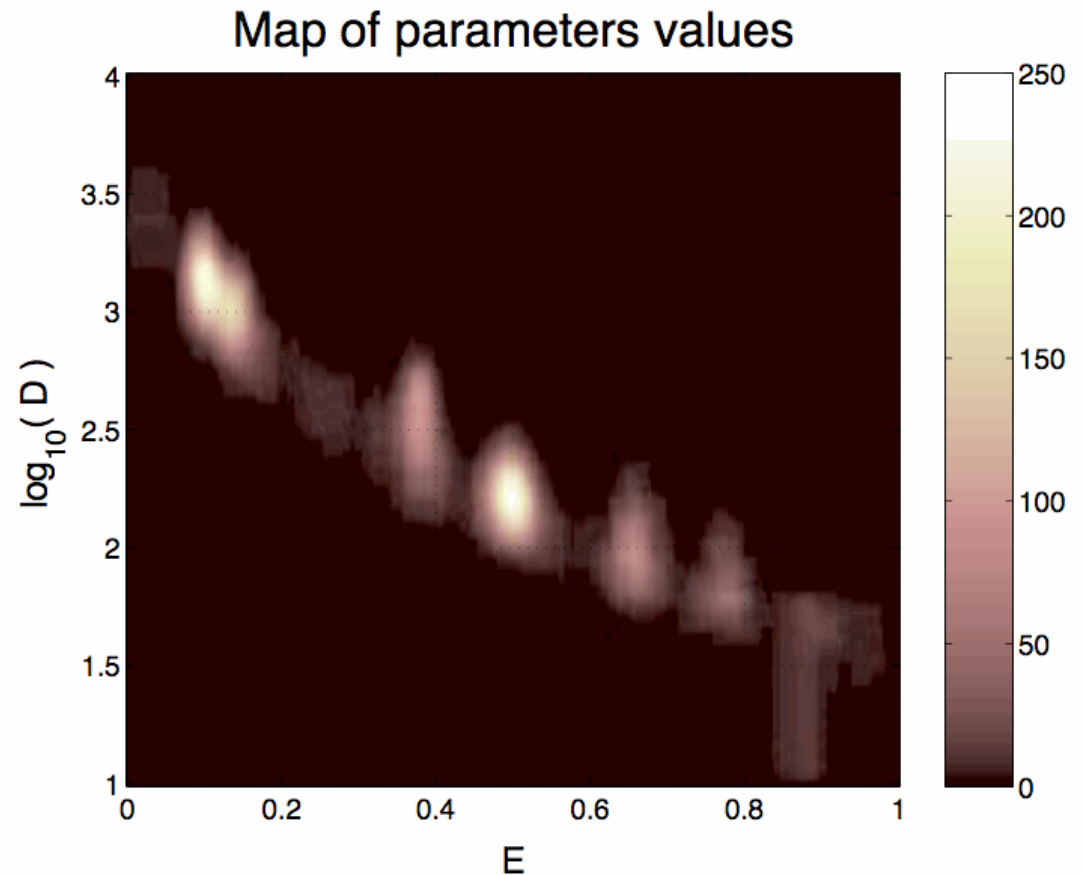
Wanted true discovery rates not
to be all 0% or 100% -- need
rates that are measurable with
the remaining datasets –
200 per signal point.

Three signal models at which
to quote sensitivity

$E=0.1, D=1010$

$E=0.5, D=137$

$E=0.9, D=18$ (the hardest one)



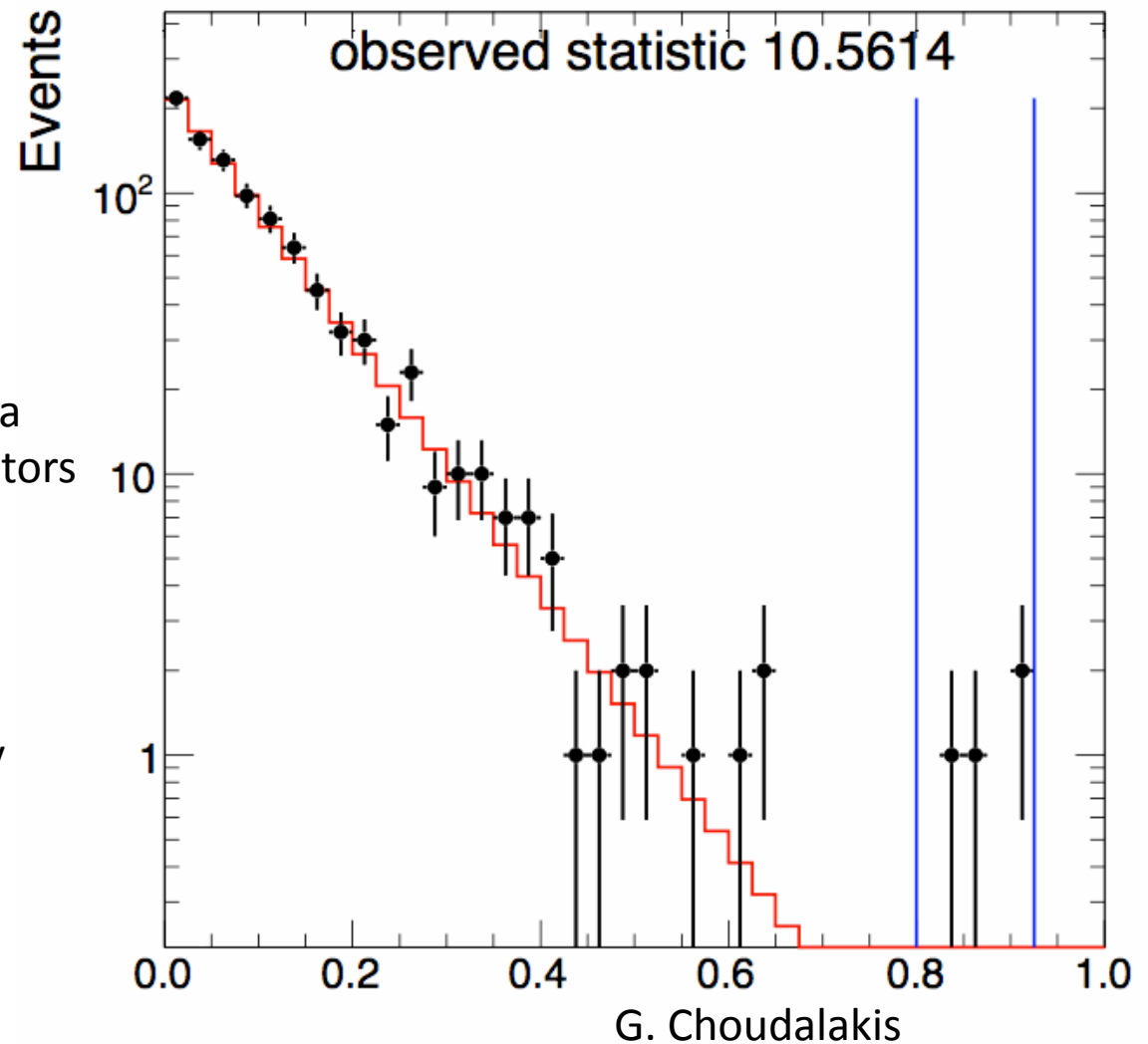
Valentin Niess

Some of the Signals are Small and Way Out on the Tails

We *did* say that the background distribution was exponential, so the few events out on the end of this histogram are evidence of a signal.

But in a real experiment, we'd have a tough time convincing our collaborators it's real. Tails of distributions are notoriously hard to model reliably.

Importance of assigning proper systematic uncertainties, which may be fractionally large on a tail of a distribution



Handling the Nuisance Parameter

D and E are the “parameters of interest” – we ask for intervals on both, and want to test $H_0: D=0$

The background normalization is uncertain at the 10% level however. Fitting the data gives a fractional uncertainty of $1/\sqrt{1000} \sim 0.03$ so the *a priori* (or auxiliary) prediction of $A=10000 \pm 1000$ carries little extra information.

Contributors who fit for A (and all did) thus were immune to the broad distribution of A chosen in the simulated data.

In the early phases of an experiment, or for selections resulting in too few data, the auxiliary predictions can be more powerful than the sideband fits. With more running time, the sidebands overpower the predictions.

The most complicated stage is the crossover – theoretical predictions and sideband measurements have similar uncertainty.

A real problem may have a mixture of these – some backgrounds are better determined with MC, others with the data.

Constructing the Datasets

- Randomly choose an A from a Gaussian distribution centered on 10000 with width 1000 (do not let A be negative)
- Generate a Poisson number from a distribution with mean $A*(0.9999955)$: these will be “background” events n_b
- Generate a number signal events n_s from a Poisson distribution of mean $D\sqrt{2\pi\sigma^2}$
- Generate n_b marks x for the background events from the exponential distribution
- Generate n_s marks x for the signal events from the Gaussian signal
- Shuffle the marks and write to the challenge file.
- Save the injected values of A , D , E , n_b and n_s in an “answer key”
- The signal models were chosen from the list of 23 models.
- The simulated dataset order was shuffled.

The Prior-Predictive Ensemble

- A usual, but not unique choice for handling nuisance parameters in HEP
- Fluctuate all nuisance parameters within their priors for each simulated dataset used to compute p values (and thus coverage)
- Has benefits of consistency of treatment
 - Setting limits or making discovery – a $+2\sigma$ excess shows up as a $+2\sigma$ excess in all plots: $-2\ln Q$, cross section limit, cross section measurement, p values for H_0 and H_{test} .

Alternative: supremum p value – the limit plot could show an excess of data wrt the background while the discovery plot could show an excess, just because we picked the most conservative choice of background for the two interpretations.

Supremum p values don't line up with the cross section measurement – we don't want always the largest cross section or the smallest, but the best with its uncertainty

- Behaves in an intuitive way when systematic uncertainties dominate. E.g., large data counts, counting experiment. Significance of observation depends on where the data fall in the tail of the prior for the background prediction.

Quoting Error Rates

- We use the prior predictive ensemble – drawing A from its prior, D and E are fixed for each model tested.
- The Type-I error rate is just the fraction of $D=0$ simulated datasets on which an analyzer makes a discovery decision
- We want this rate to be $\leq 1\%$ to mimic the procedure used to make a discovery at higher significance levels.
- Analyzers fit for A anyhow. Had the prior information on A been dominant over the fits, we would be discussing how to define the multi- σ tails on the priors for A .
- Similarly for the discovery rates – these are quoted as the fraction of true discoveries in each of the signal categories.

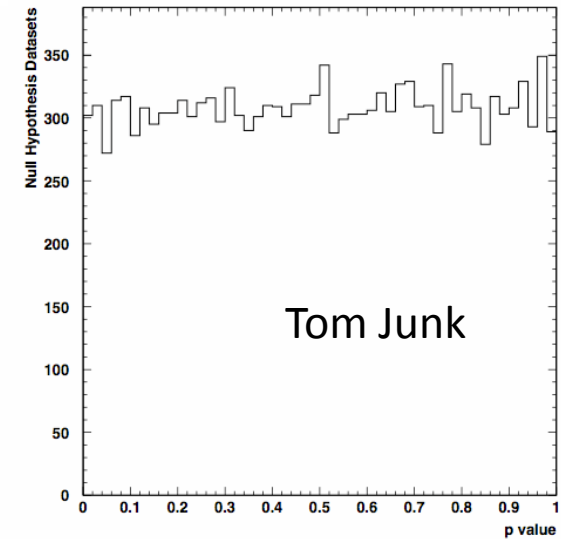
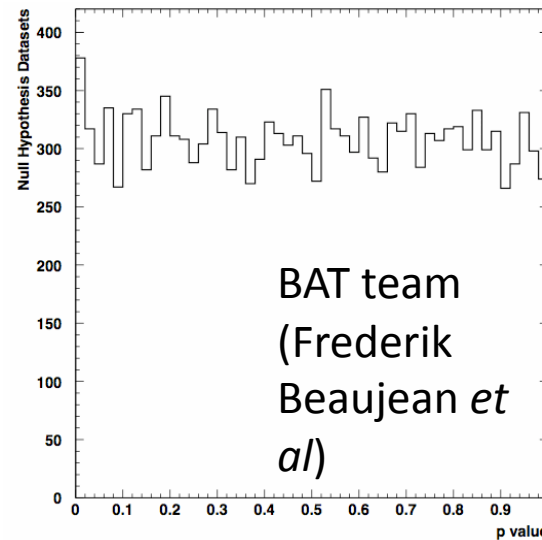
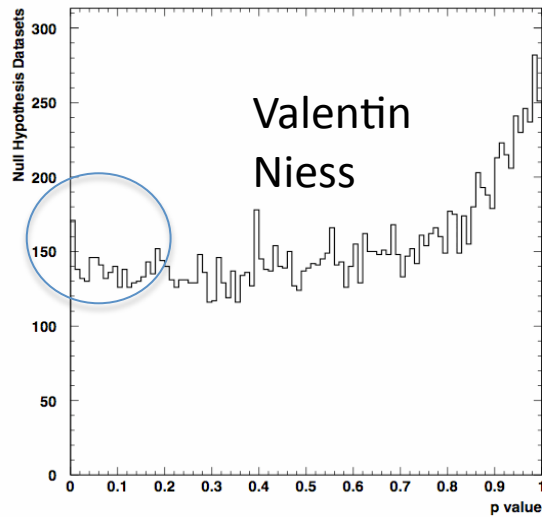
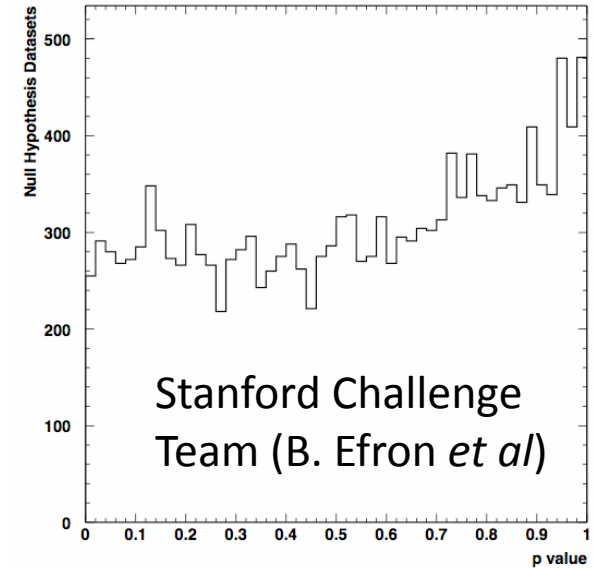
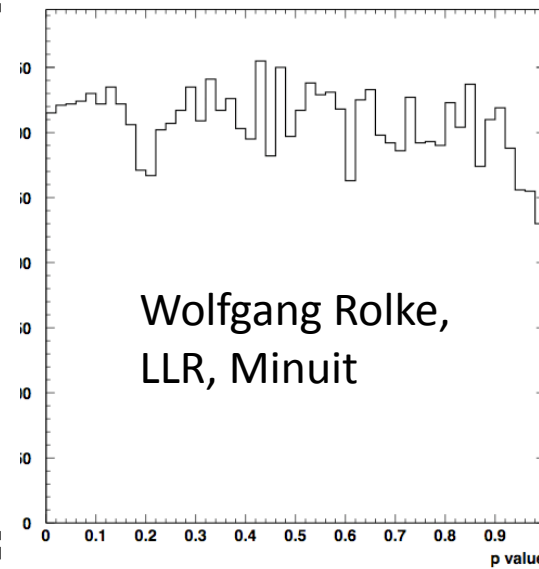
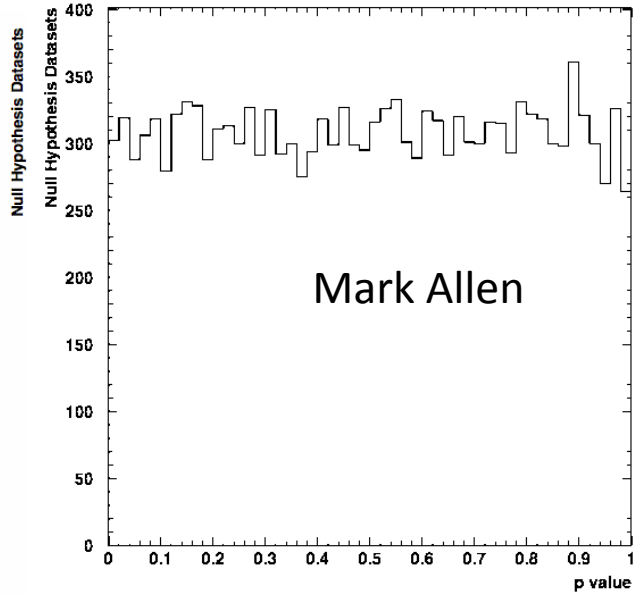
Performance – Problem 1 – Discovery Error Rates

Contributor	Type-I Error Rate Measured	$D = 1010, E = 0.1$		$D = 137, E = 0.5$		$D = 18, E = 0.9$	
		Claimed	Measured	Claimed	Measured	Claimed	Measured
Tom Junk	0.0097 ± 0.0008	0.256	0.3150 ± 0.0328	0.543	0.6100 ± 0.0345	0.108	0.1350 ± 0.0242
Wolfgang Rolke	0.0103 ± 0.0008	0.356	0.3800 ± 0.0343	0.457	0.5250 ± 0.0353	0.184	0.2150 ± 0.0290
Stanford Challenge Team (SCT)	0.0077 ± 0.0007	0.3483	0.3550 ± 0.0338	0.4335	0.5200 ± 0.0353	0.0175	0.2100 ± 0.0288
Eilam Gross & Ofer Vitells	0.0082 ± 0.0007	0.35	0.3600 ± 0.0339	0.46	0.5250 ± 0.0353	0.19	0.2100 ± 0.0288
Valentin Niess	0.0111 ± 0.0008	0.603	0.3250 ± 0.0331	0.87	0.5300 ± 0.0353	0.12	0.1950 ± 0.0280
Georgios Choudalakis	0.0110 ± 0.0008	0.213	0.1600 ± 0.0259	0.290	0.3500 ± 0.0337	0.107	0.1300 ± 0.0238
Mark Allen	0.0106 ± 0.0008	0.385	0.4000 ± 0.0346	0.486	0.5250 ± 0.0353	0.187	0.2100 ± 0.0288
Frederik Beaujean (BAT)	0.0000 ± 0.0000		0.0000 ± 0.0000		0.0300 ± 0.0121		0.0050 ± 0.0050
Stefan Schmitt							
Unbinned	0.0112 ± 0.0009		0.4500 ± 0.0352		0.5450 ± 0.0352		0.1850 ± 0.0275
Binned	0.0110 ± 0.0008	0.37	0.3850 ± 0.0344	0.53	0.5450 ± 0.0352	0.17	0.2200 ± 0.0293
Stefano Andreon							
$p < 3 \times 10^{-3}$	0.0126 ± 0.0013		0.4811 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120
$p < 4 \times 10^{-3}$	0.0191 ± 0.0016		0.5189 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120

Update from V. Niess: powers of 0.34, 0.46, 0.17 quoted after answers released.
(technical issue resolved not related to the answer key)

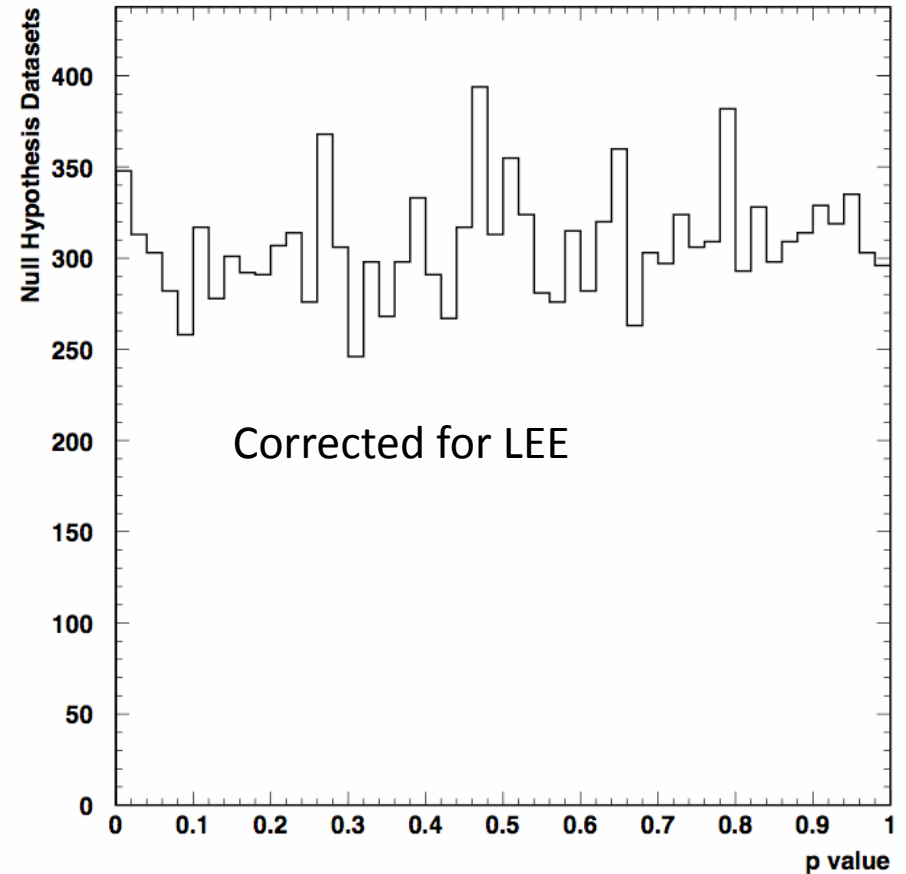
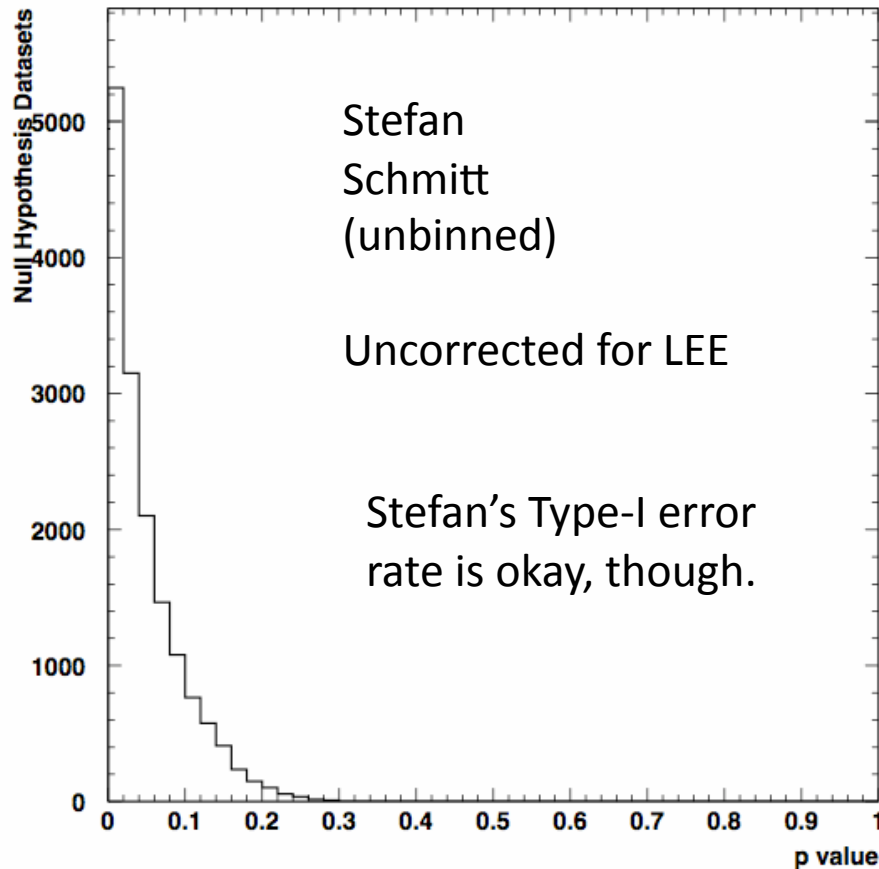
p Value Distributions – Should be uniform on $[0,1]$

Problem 1

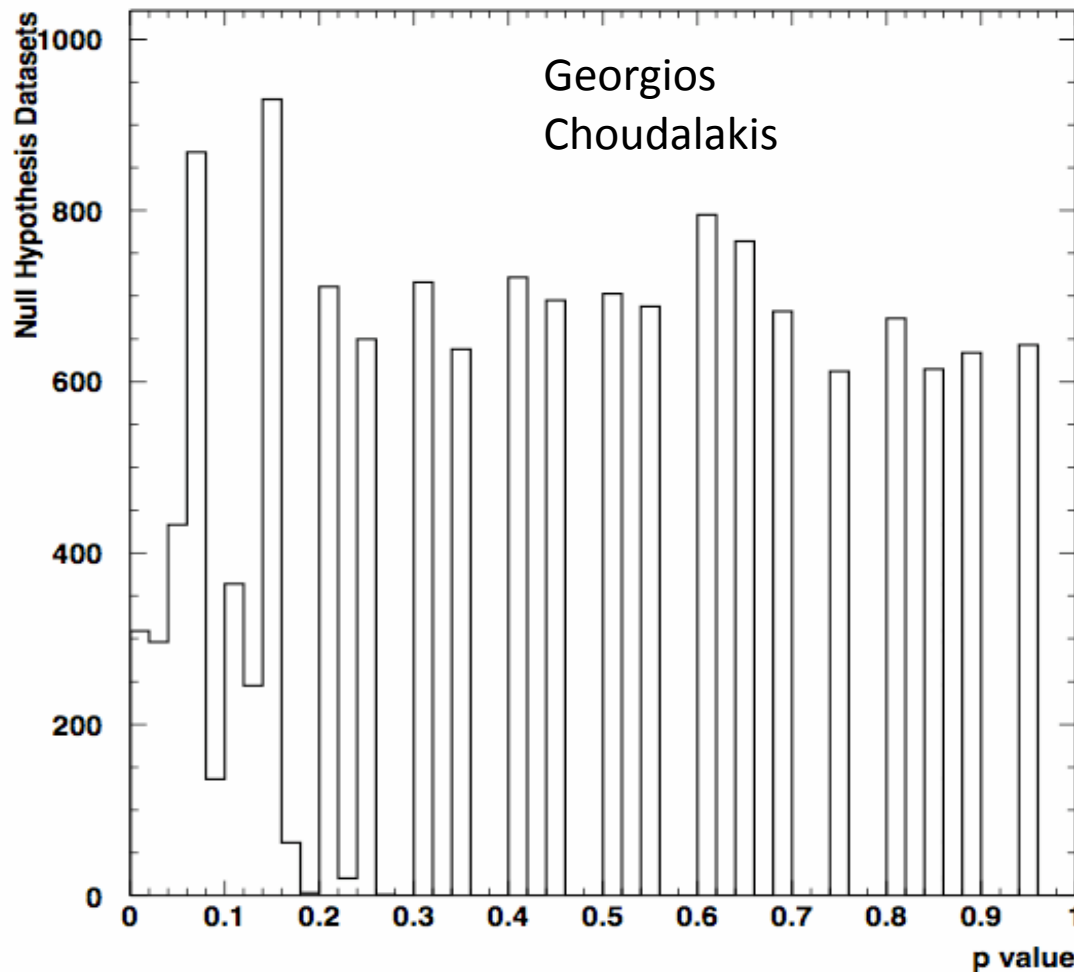


Stefan's p Values – Correcting for Look-Elsewhere

LEE Correction done with Background-Only Monte Carlo simulation



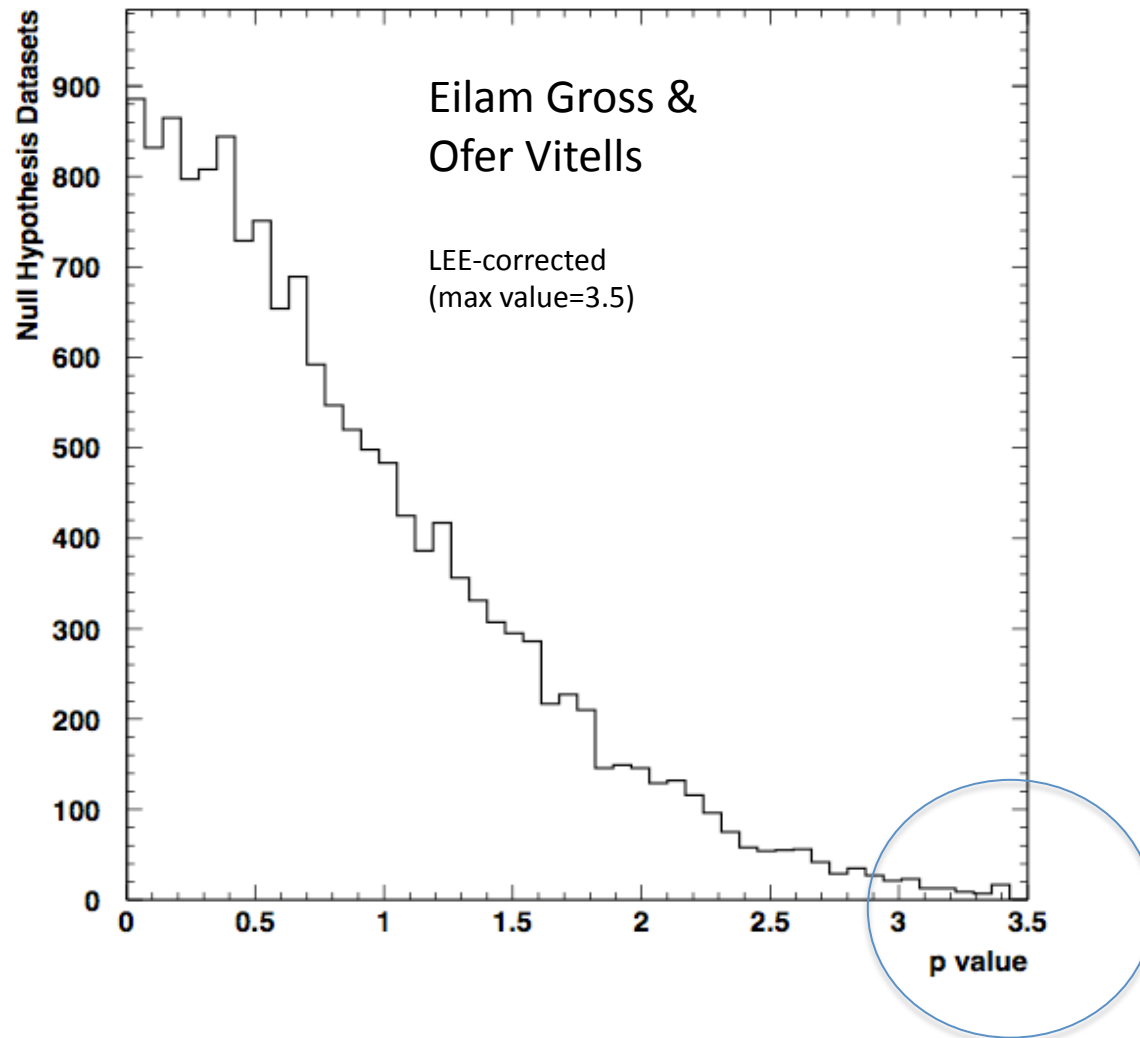
p Value Distributions – Some of the More Interesting Cases for Problem 1



simulation stopped
when very sure $p > 0.01$

Not a problem for
discovery

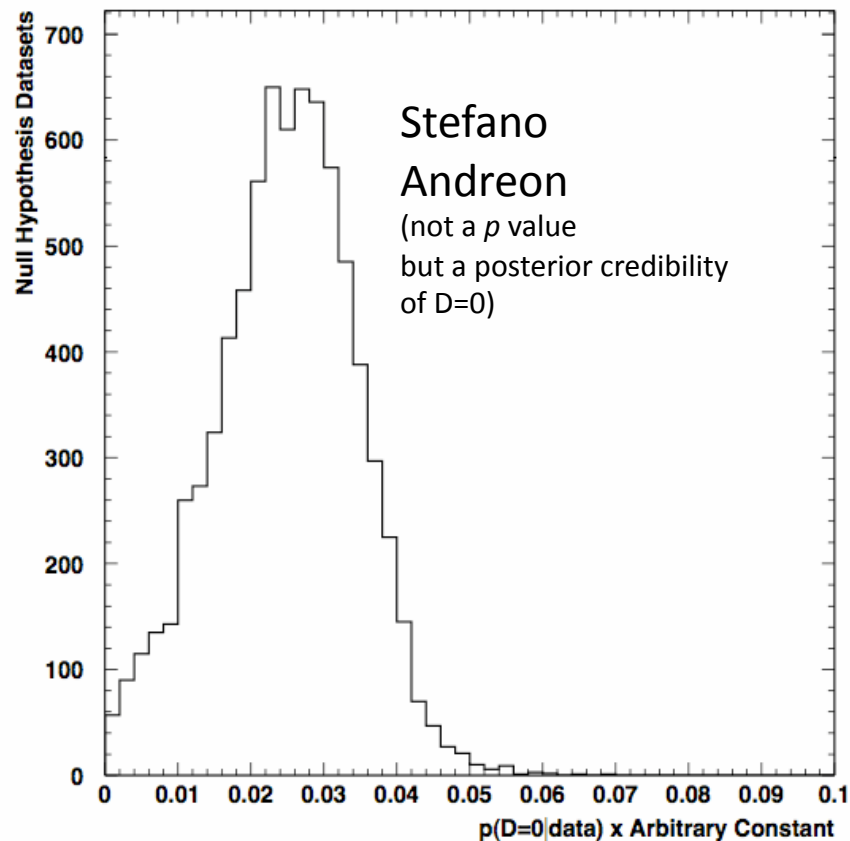
p Value Distributions – Some of the More Interesting Cases for Problem 1



Large maximum p value comes from the LEE correction – See Ofer’s talk tomorrow.

Not a problem for discovery. May want to use this distribution as a test statistic in case there’s a 1σ deficit...

And One Test Statistic Instead of a p Value



$P(D=0 | \text{data})$ – a Bayesian posterior

This is okay too – but what's the critical value?

Stefano chooses 3×10^{-4} and 4×10^{-4} as possibilities

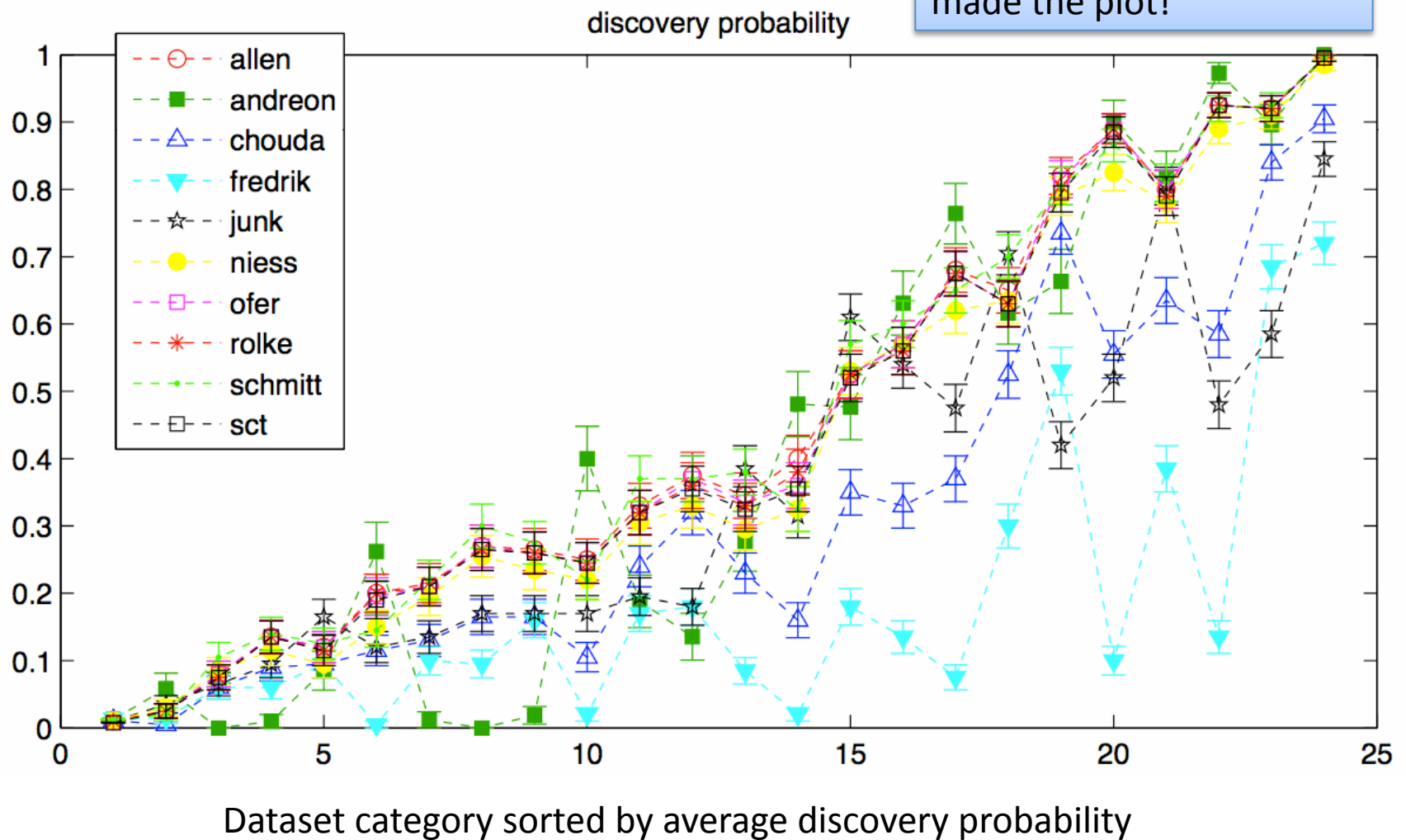
A Typical Error Rate Summary Table

Category	E_{true}	D_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Ecorr}	f_{Ecorr}	n_{Dcorr}	f_{Dcorr}	$\langle E_{\text{wid}} \rangle$	$\langle D_{\text{wid}} \rangle$
1	—	0.00	15400	149	0.0097 ± 0.0008	—	—	6	0.0403	0.0399	309.1451
2	0.50	83.78	200	33	0.1650 ± 0.0262	25	0.7576	26	0.7879	0.0413	163.1197
3	0.38	265.96	200	108	0.5400 ± 0.0352	82	0.7593	91	0.8426	0.0330	270.4944
4	0.10	1010.65	200	63	0.3150 ± 0.0328	57	0.9048	48	0.7619	0.0263	1000.3658
5	0.10	478.73	200	7	0.0350 ± 0.0130	5	0.7143	0	0.0000	0.0320	888.4238
6	0.66	66.49	200	39	0.1950 ± 0.0280	28	0.7179	35	0.8974	0.0406	117.1397
7	0.78	39.89	200	36	0.1800 ± 0.0272	28	0.7778	33	0.9167	0.0445	89.9875
8	0.10	744.69	200	24	0.1200 ± 0.0230	18	0.7500	14	0.5833	0.0278	914.8707
9	0.50	136.97	200	122	0.6100 ± 0.0345	102	0.8361	107	0.8770	0.0350	177.6496
10	0.90	15.29	200	13	0.0650 ± 0.0174	7	0.5385	12	0.9231	0.0524	75.2728
11	0.50	190.16	200	161	0.8050 ± 0.0280	131	0.8137	148	0.9193	0.0317	186.9565
12	0.14	664.90	200	34	0.1700 ± 0.0266	27	0.7941	23	0.6765	0.0289	790.7153
13	0.50	163.57	200	141	0.7050 ± 0.0322	106	0.7518	127	0.9007	0.0334	181.7568
14	0.38	531.92	200	169	0.8450 ± 0.0256	130	0.7692	142	0.8402	0.0214	308.6828
15	0.14	1196.83	200	96	0.4800 ± 0.0353	82	0.8542	89	0.9271	0.0240	807.8491
16	0.50	110.37	200	77	0.3850 ± 0.0344	54	0.7013	61	0.7922	0.0372	173.8191
17	0.10	1276.62	200	95	0.4750 ± 0.0353	80	0.8421	81	0.8526	0.0249	1001.8887
18	0.90	20.61	200	34	0.1700 ± 0.0266	25	0.7353	27	0.7941	0.0492	71.4330
19	0.66	132.98	200	117	0.5850 ± 0.0348	90	0.7692	108	0.9231	0.0319	135.6544
20	0.90	12.63	200	19	0.0950 ± 0.0207	12	0.6316	11	0.5789	0.0474	137.4571
21	0.90	17.95	200	27	0.1350 ± 0.0242	19	0.7037	21	0.7778	0.0491	85.2629
22	0.90	23.27	200	34	0.1700 ± 0.0266	28	0.8235	32	0.9412	0.0500	72.9724
23	0.78	79.79	200	84	0.4200 ± 0.0349	73	0.8690	75	0.8929	0.0379	106.7026
24	0.10	1542.58	200	104	0.5200 ± 0.0353	88	0.8462	91	0.8750	0.0217	1026.1458

From Tom Junk

A summary of the Discovery Probabilities

Thanks to Ofer Vitells who made the plot!



Making Measurements of D and E

signal
rate

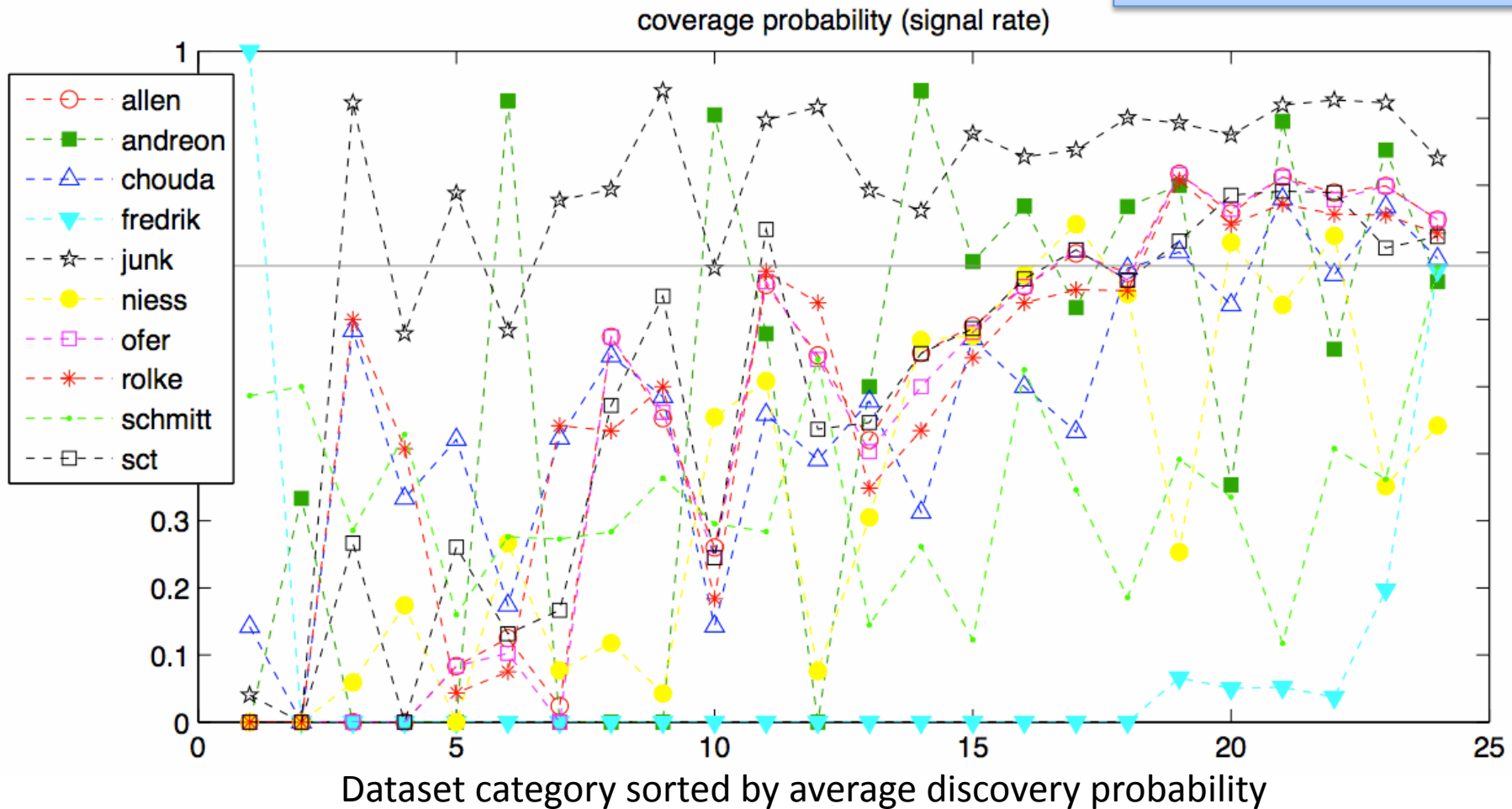
peak
position

- Requested 68% intervals for D and E in the case a decision was made to claim a discovery.
- I regret a little bit not asking for these in case a discovery is not claimed. Reporting only the discovery cases biases the measurements of D upwards.

HEP experiments rarely report measurements of parameters like D and E unless evidence is claimed.

A summary of the Coverage of D Measurements

Thanks to Ofer Vitells who made the plot!

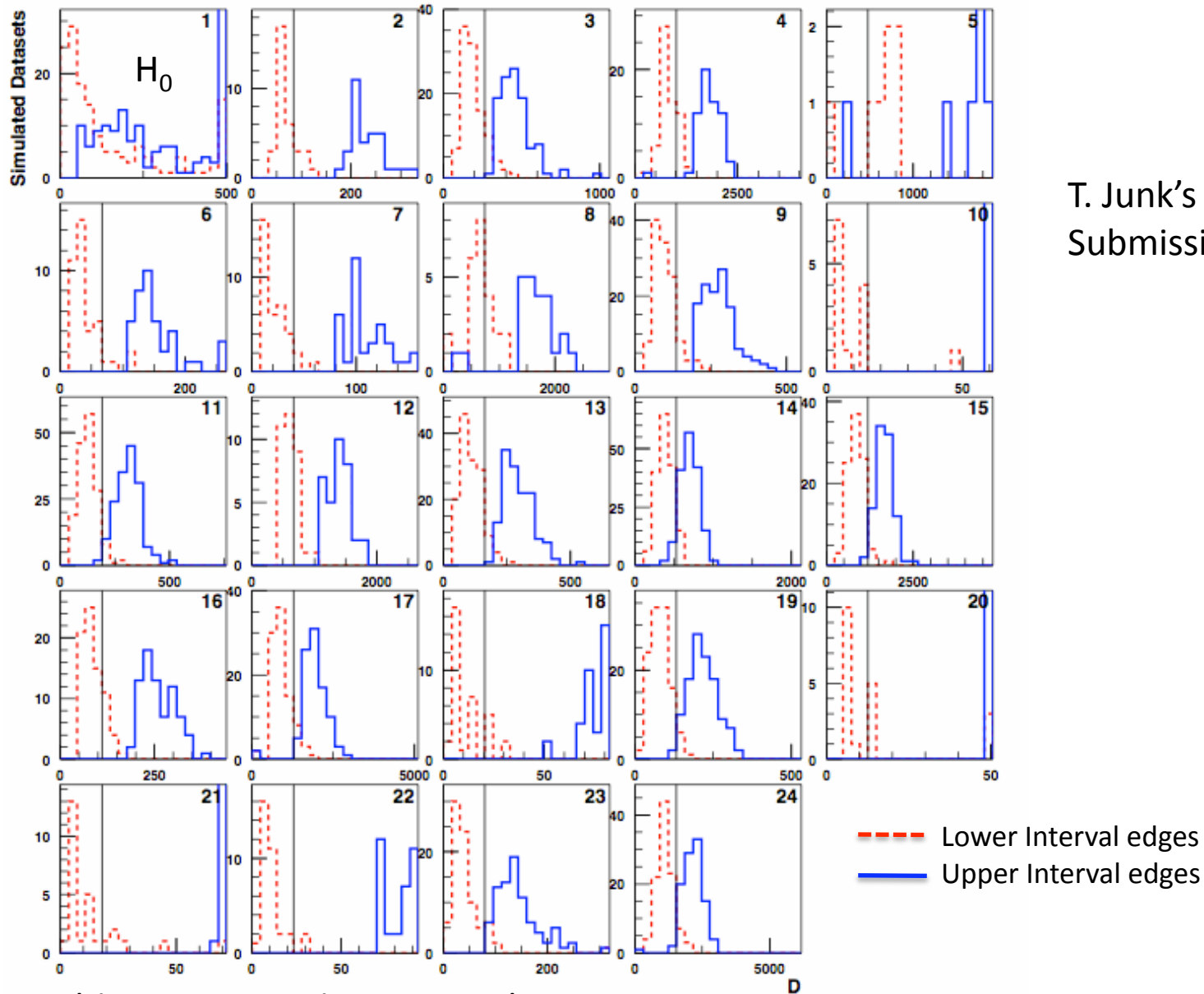


Upper and Lower Interval Edges for D for Each Signal Category

Large signals
OK, small
signals
biased
upwards

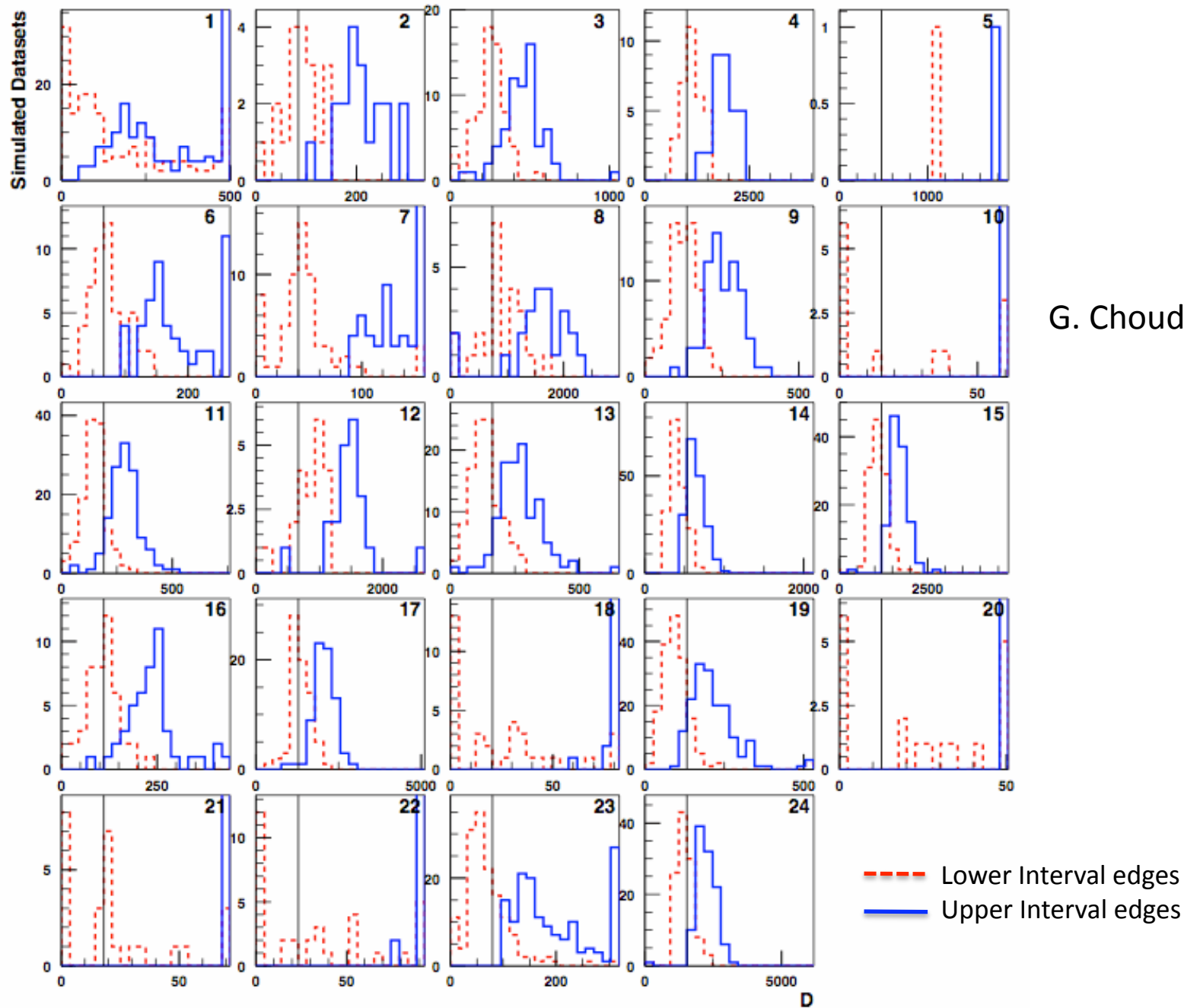
Sum of
fractional
areas of
red > black
and
blue < black
should be < 32%

Tom:
Call mnseek
MINIMIZE
IMPROVE & quote resulting uncertainties on D and E



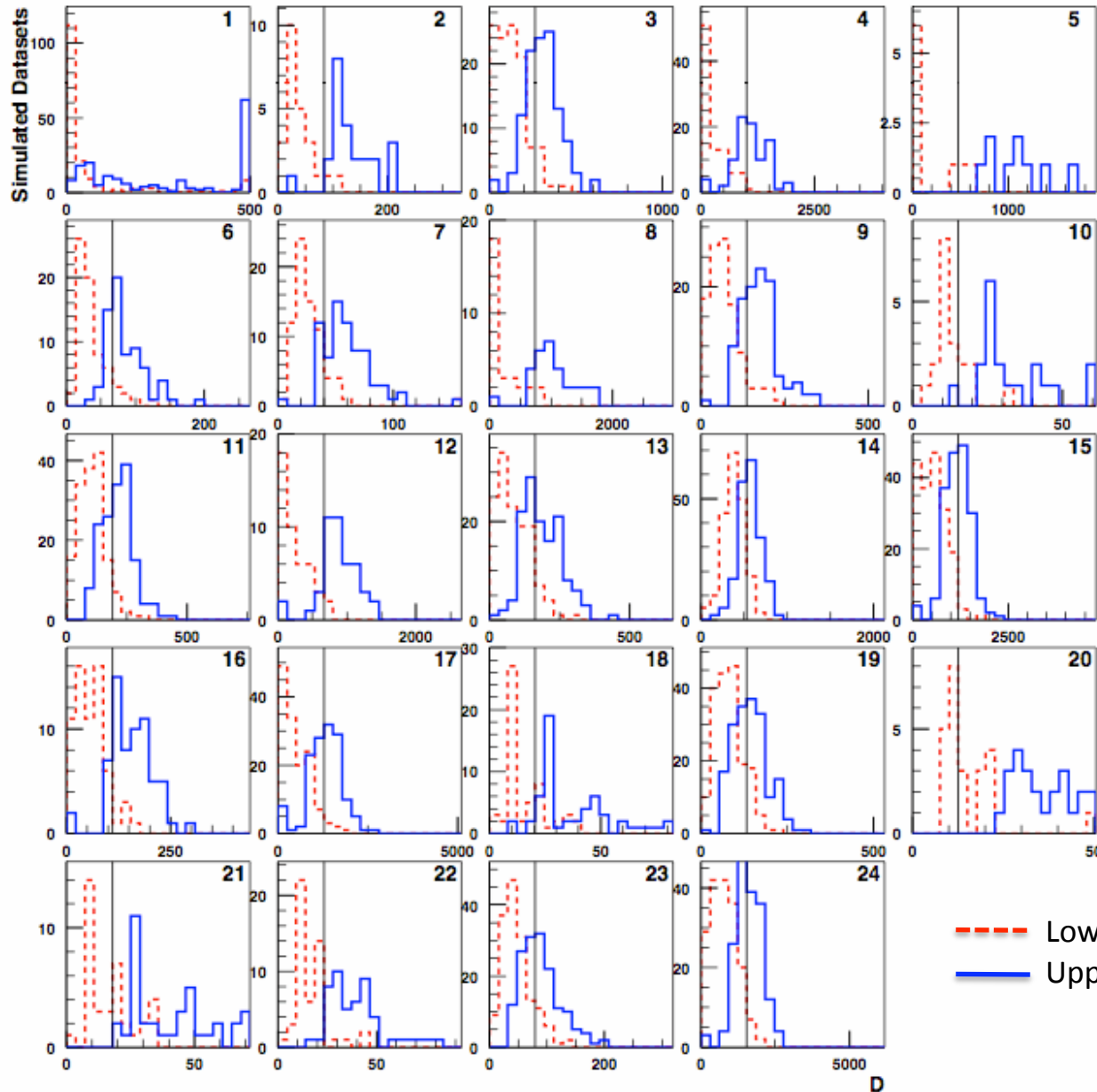
T. Junk's
Submission

Differences in Participants' Intervals – This set tends to give too large D



G. Choudalakis

This set tends to give smaller D

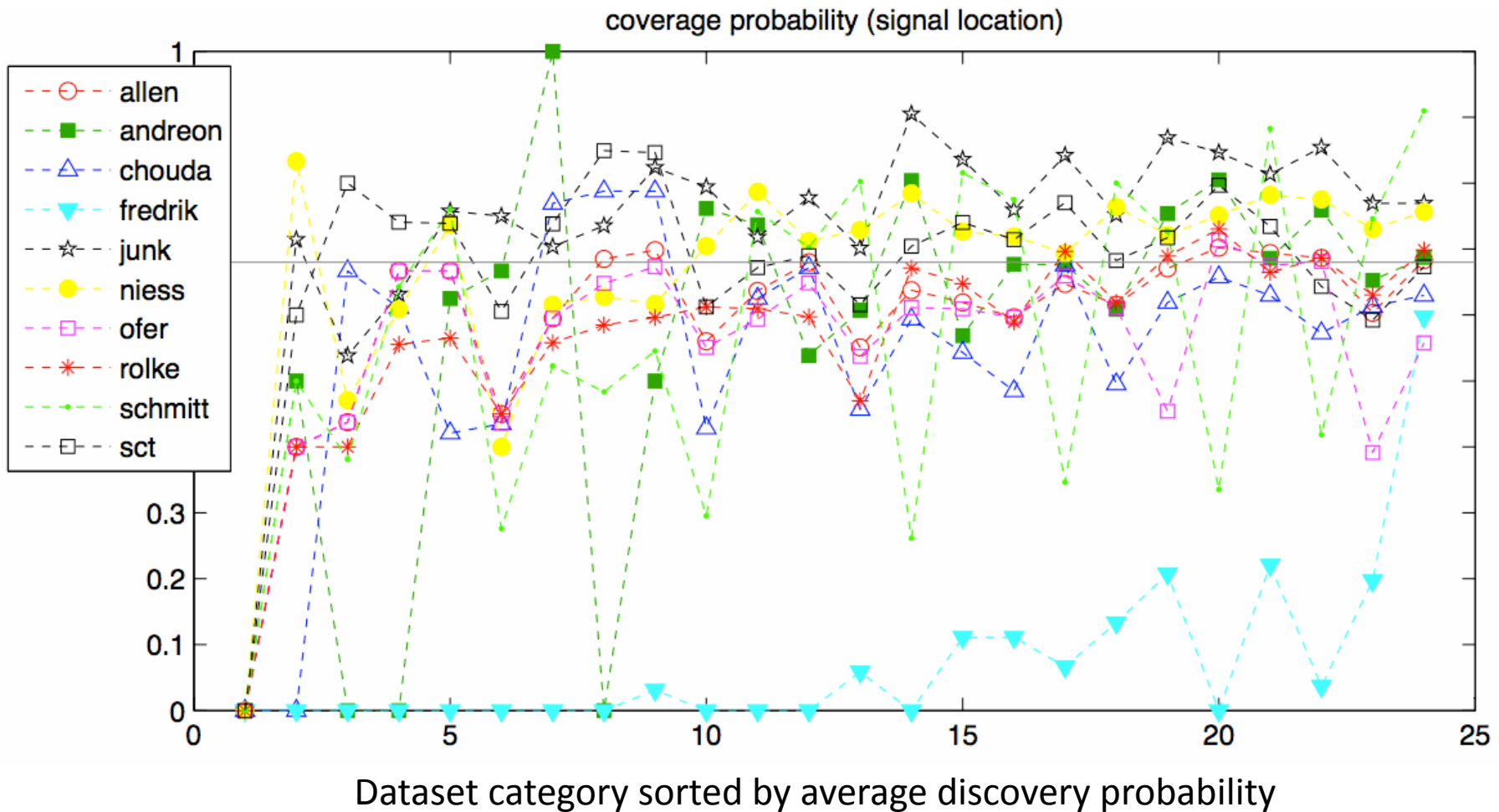


More blue
to the left
of the
black
lines.

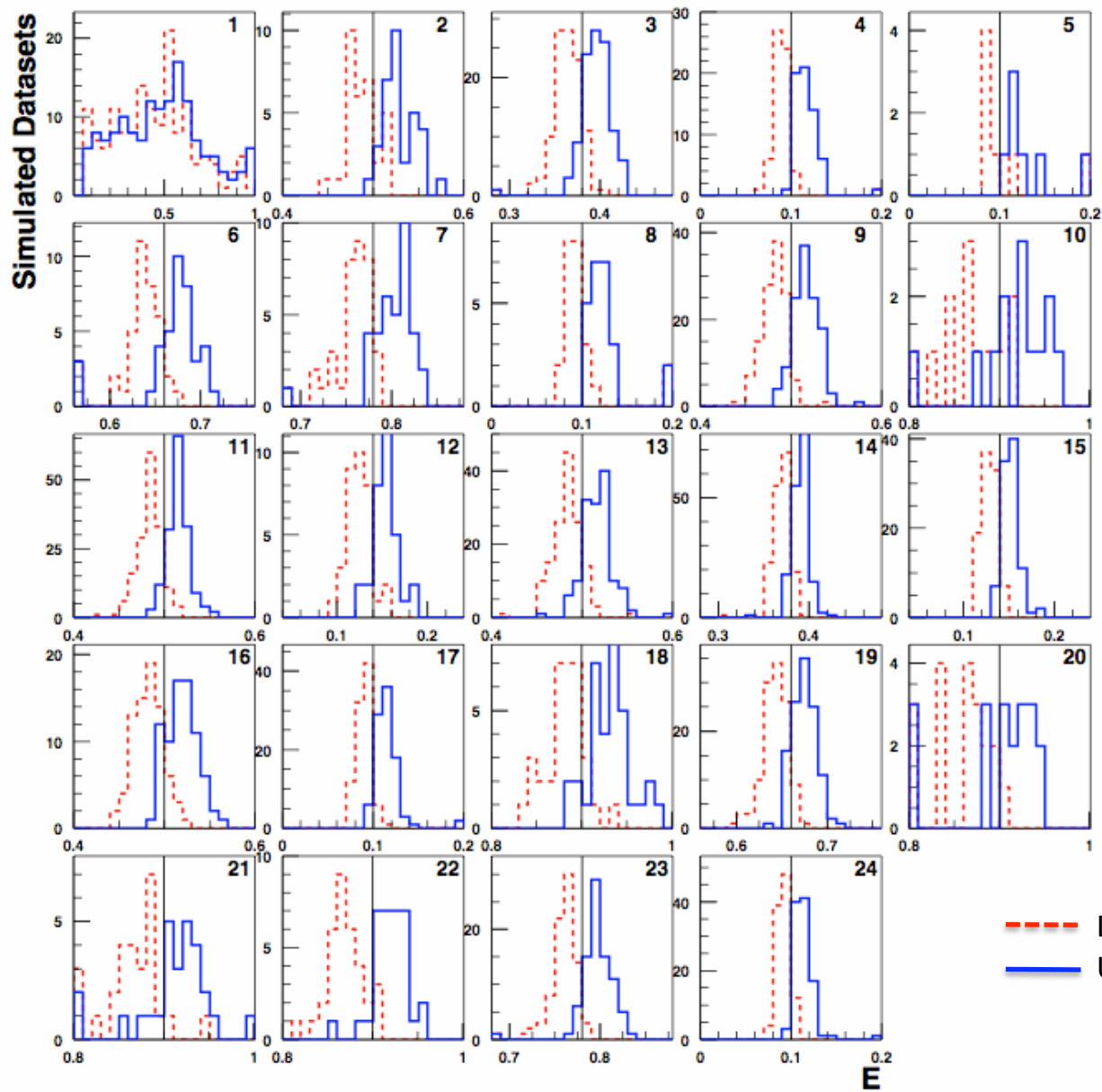
S. Schmitt,
Unbinned

A summary of the Coverage of E Measurements

Thanks to Ofer Vitells who made the plot!



Fraction of Correct Intervals for E



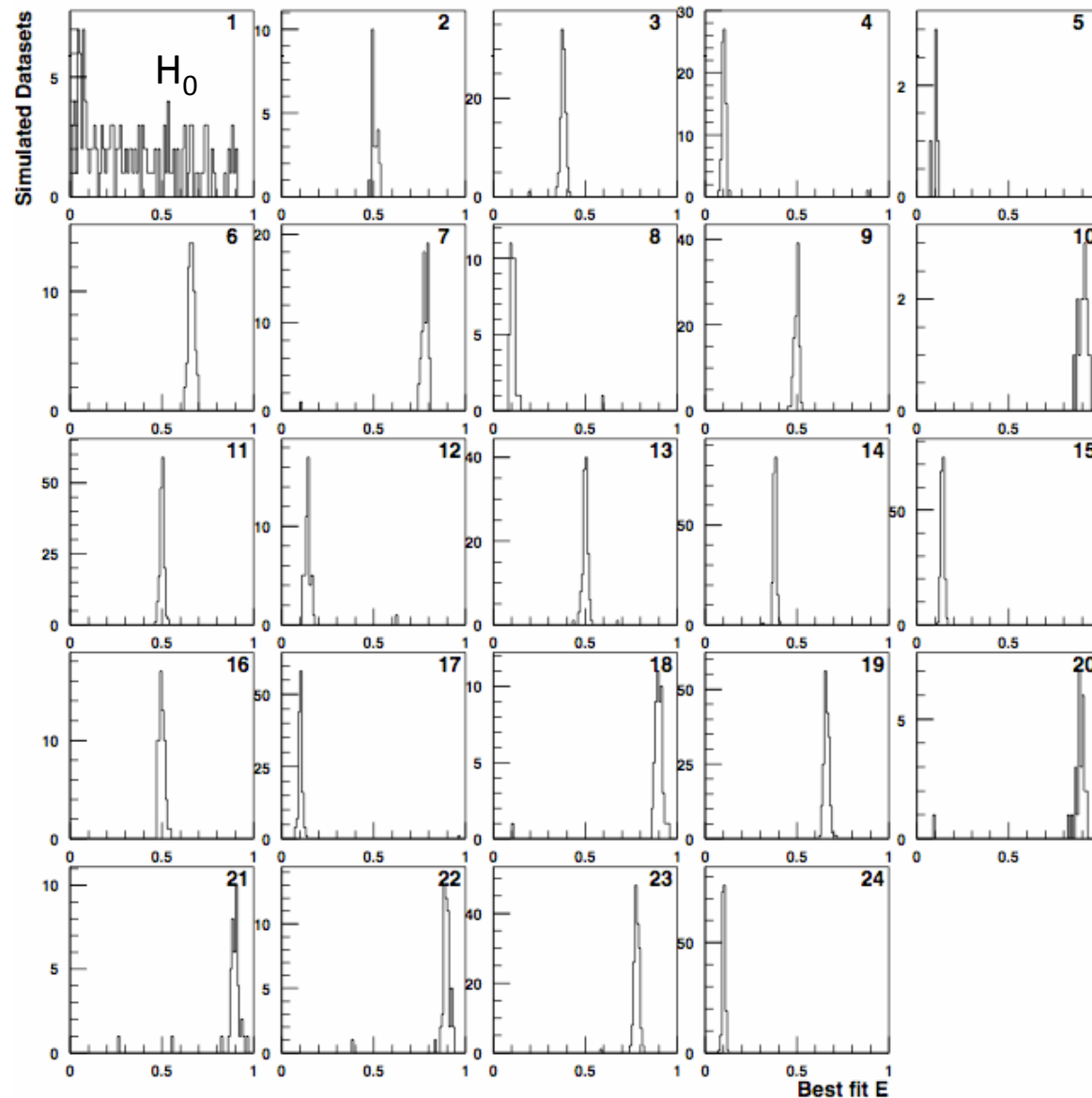
T. Junk's
Submission

Even Intervals that don't contain the true value quite often enough tend to still get the right answer. Intervals just a bit too short.

W. Rolke

No long tails

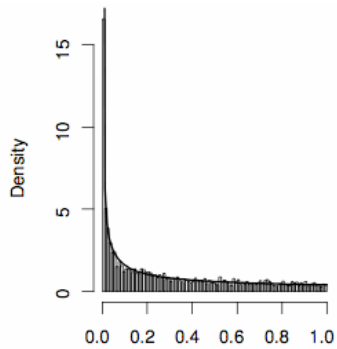
Avg. interval length much shorter than Tom's for example.



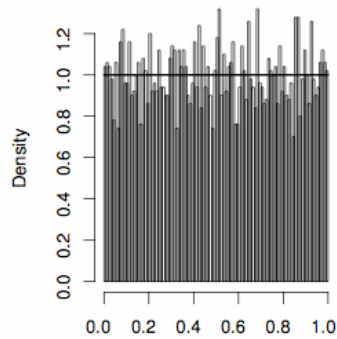
Problem 2 – A Monte-Carlo Parameterized Example

Three processes contribute – Two Backgrounds and one Signal process.

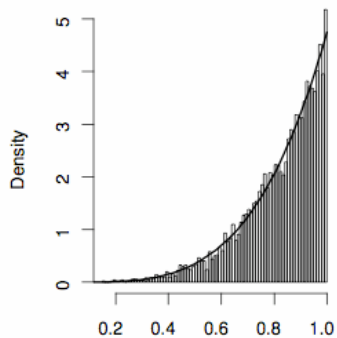
Monte Carlo simulations of each process provided with 5000 events (“marks”) apiece.



Background 1, Beta(0.4,1)

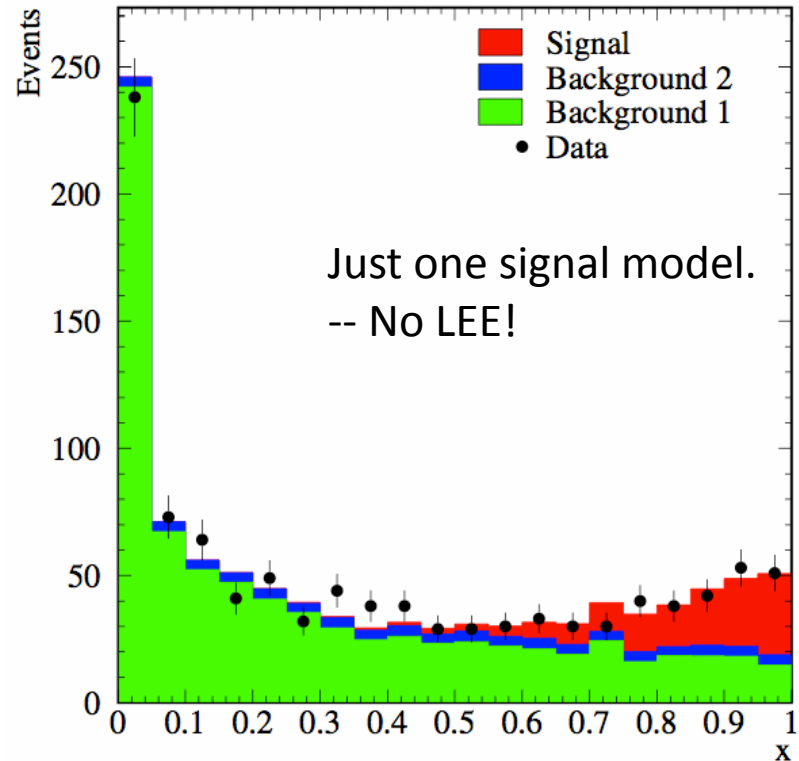


Background 2, Beta(1,1)



Signal, Beta(4.75,1)

W. Rolke

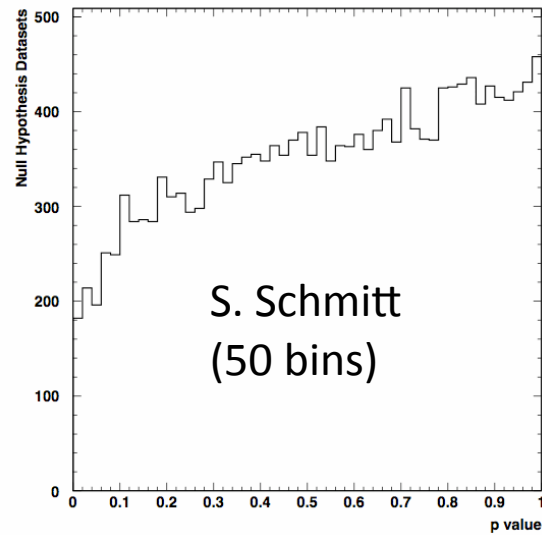
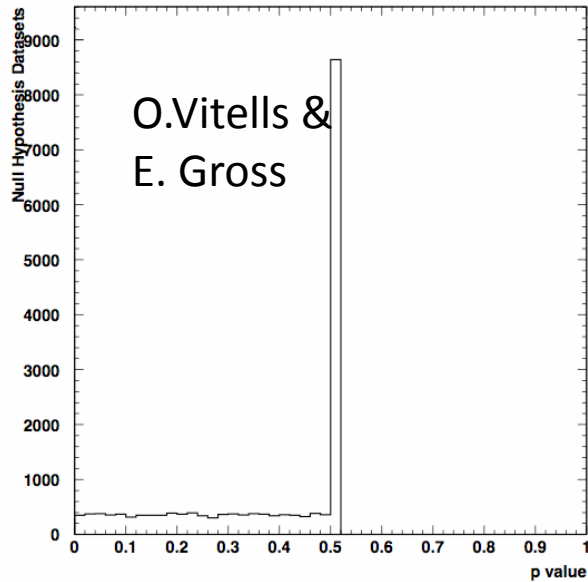
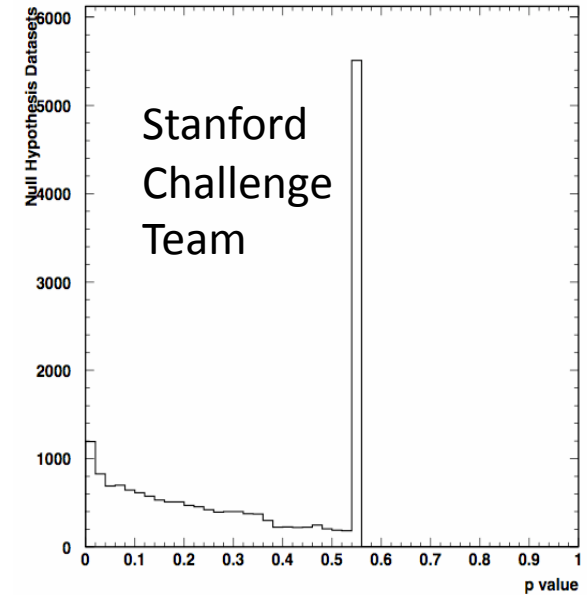
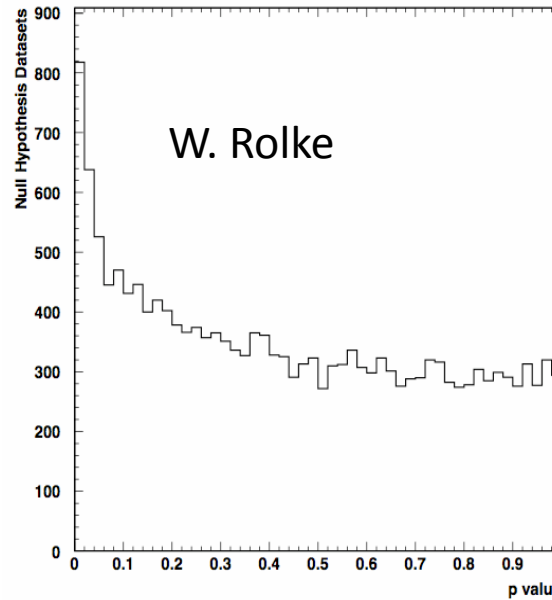
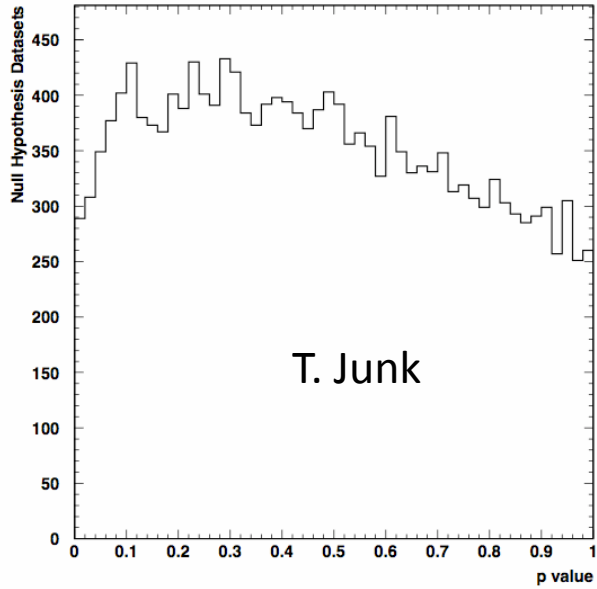


Problem 2 Error Rate Summary

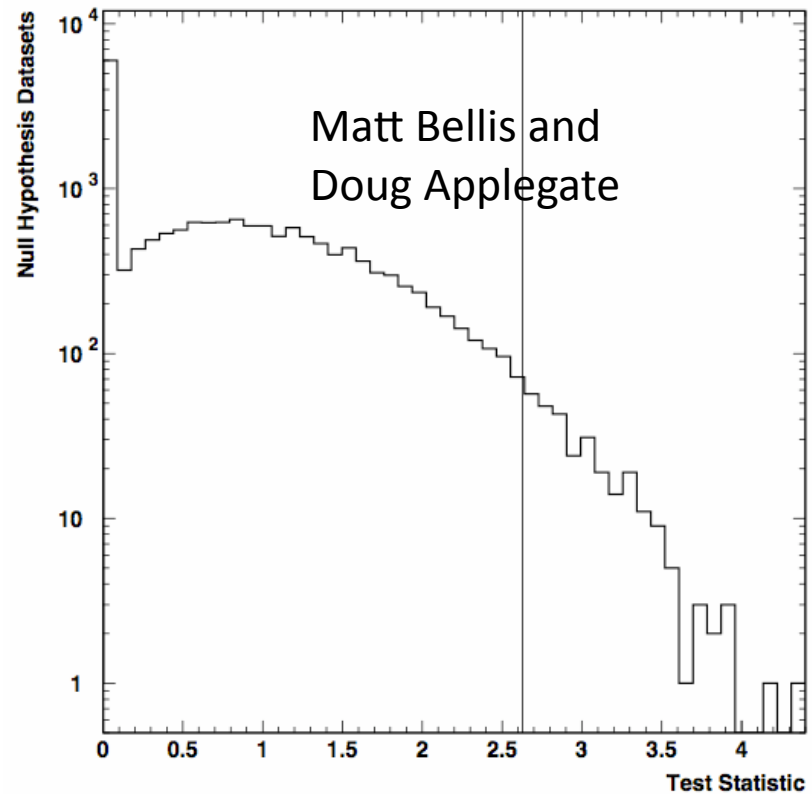
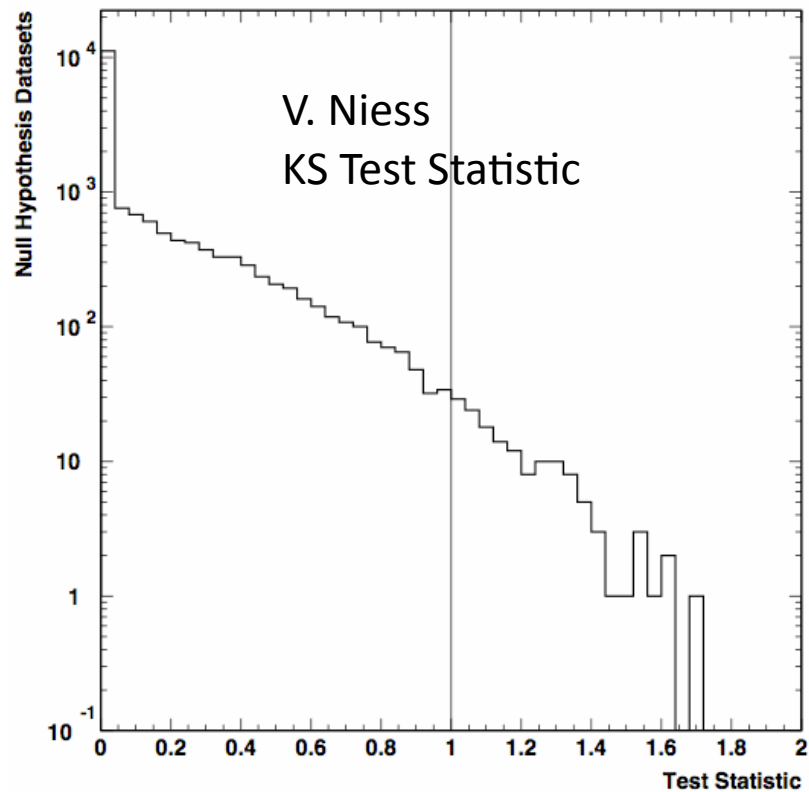
Contributor	Type-I Error Rate Measured	Signal = 75 Events	
		Claimed	Measured
Tom Junk	0.0068 ± 0.0006	0.865	0.870 ± 0.017
Wolfgang Rolke	0.0256 ± 0.0012	0.88	0.8500 ± 0.018
Stanford Challenge Team	0.0389 ± 0.0015	0.84	0.9100 ± 0.0143
Eilam Gross & Ofer Vitells	0.0107 ± 0.0008	0.815	0.7725 ± 0.0210
Valentin Niess	0.0085 ± 0.0007	0.761 ± 0.001	0.7125 ± 0.0226
Stefan Schmitt			
25 Bins	0.0047 ± 0.0005	0.85	0.8200 ± 0.0192
50 Bins	0.0047 ± 0.0005		0.8250 ± 0.0190
Doug Applegate & Matt Bellis	0.0168 ± 0.0010	0.95	0.8950 ± 0.0153

Distributions of the p Value for Problem 2

Should be flat..



And Some Test Statistics Instead for Problem 2



Measuring the Signal Rate in Problem 2

V. Niess
 -- coverage
 is fine for
 most categories.

Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle \text{Swid} \rangle$
1	0.00	17600	150	0.0085 ± 0.0007	0	0.0000	42.4200
2	75.00	400	285	0.7125 ± 0.0226	193	0.6772	52.3930
3	50.00	400	156	0.3900 ± 0.0244	55	0.3526	47.5449
4	25.00	400	47	0.1175 ± 0.0161	0	0.0000	41.5106
5	100.00	400	366	0.9150 ± 0.0139	261	0.7131	54.7268
6	150.00	400	400	1.0000 ± 0.0000	270	0.6750	54.9250
7	125.00	400	391	0.9775 ± 0.0074	269	0.6880	54.7442

For categories

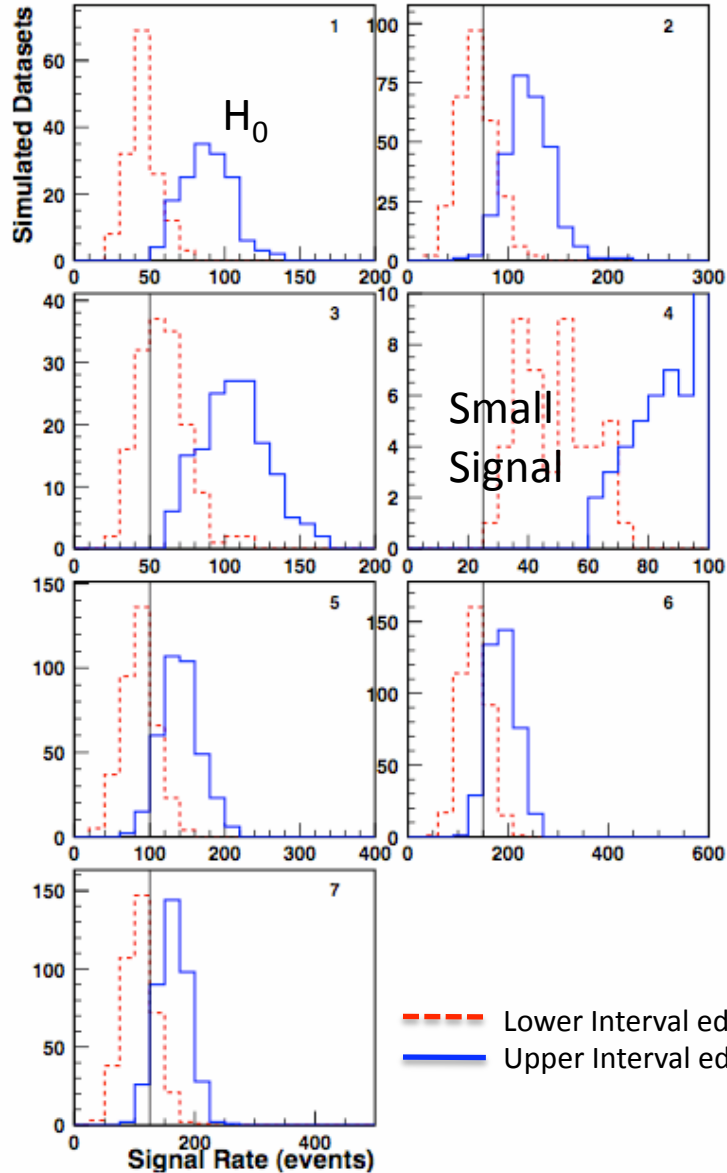
1, 3, and 4, we get the usual bias from small signal rates – quote measurements only for discoveries biases the rates upwards.

Stanford Challenge
 Team: Smaller
 intervals, less
 coverage

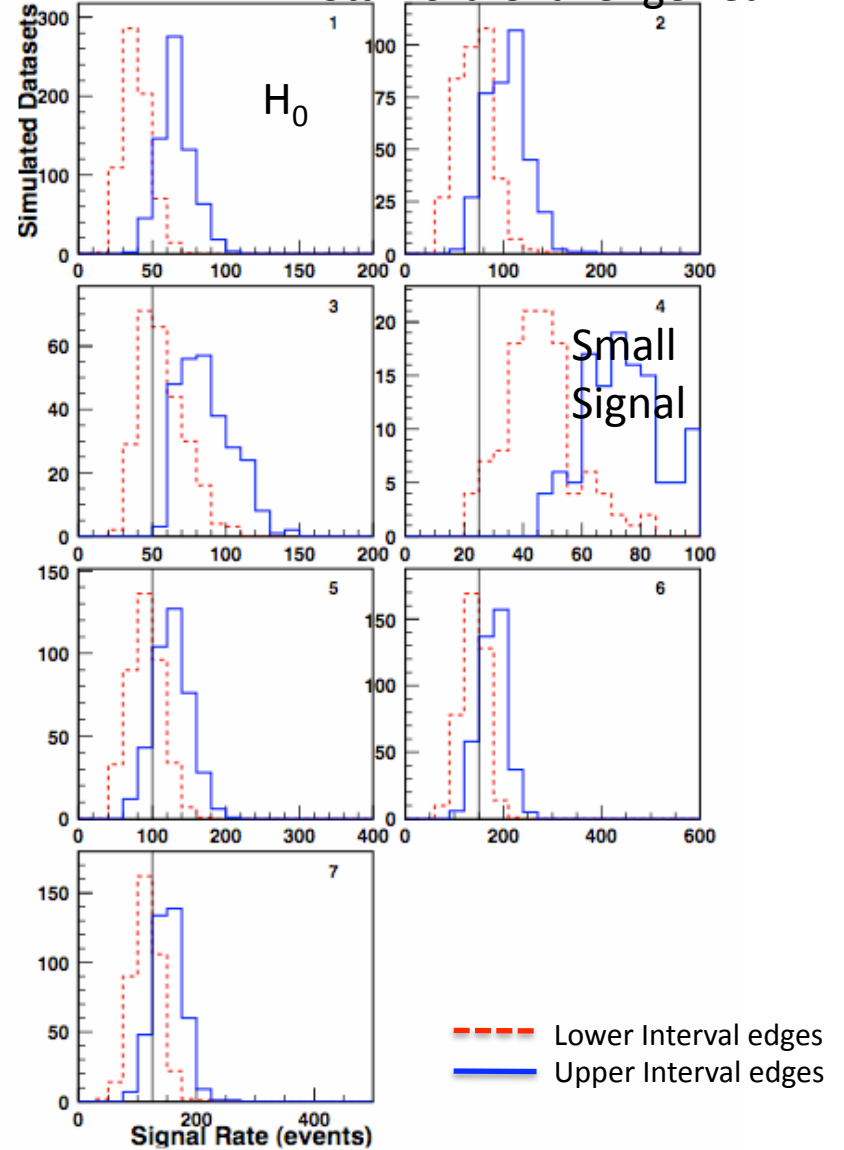
Category	Sig_{true}	n_{rep}	n_{disc}	f_{disc}	n_{Scorr}	f_{Scorr}	$\langle \text{Swid} \rangle$
1	0.00	17600	685	0.0389 ± 0.0015	0	0.0000	27.4298
2	75.00	400	364	0.9100 ± 0.0143	181	0.4973	32.4090
3	50.00	400	265	0.6625 ± 0.0236	102	0.3849	30.4449
4	25.00	400	116	0.2900 ± 0.0227	4	0.0345	28.6076
5	100.00	400	397	0.9925 ± 0.0043	204	0.5139	34.8550
6	150.00	400	400	1.0000 ± 0.0000	193	0.4825	38.8669
7	125.00	400	399	0.9975 ± 0.0025	213	0.5338	36.9672

Signal Rate Intervals for Problem 2

V. Niess




Stanford Challenge Team



Summary

- **Wow!** What a lot of good work went into Banff Challenge 2a. I am very pleased with the variety and quality of the solutions. Congratulations to all the participants!

People must be excited about making discoveries. And have time to solve challenges.

- Some methods perform better than others. Even methods that are nominally identical but differ in the details.
Good to do this with a blind test like BC2a.
- No one was tripped up by the Look-Elsewhere Effect in Problem 1 
- **Any method used to make a discovery within a collaboration must be characterized and approved by the collaboration. Good performance on the challenge problems is *not* a seal of approval! Participants' solutions are not to be considered endorsed by any HEP collaboration.**
- Some technical issues uncovered here are highly problem-specific. A difficulty encountered in Problem 2 may not affect a technique's performance on Problem 1, for example. It's a learning exercise to tune up methods on tractable, fairly realistic problems.

“Winners” – Problem 1

Based on my criteria – coverage first, then estimated power as long as it is not overestimated, I put these four in the top teams for Problem 1:

Mark Allen

Stefan Schmitt

Wolfgang Rolke

Eilam Gross and Ofer Vitells

Stanford Challenge Team (modulo a typo in the sensitivity in case 3)

Difficult to rank them because no one “won” all three operating points.

All winners have to work on the point estimation of D and E though!

Contributor	Type-I Error Rate Measured	$D = 1010, E = 0.1$		$D = 137, E = 0.5$		$D = 18, E = 0.9$	
		Claimed	Measured	Claimed	Measured	Claimed	Measured
Tom Junk	0.0097 ± 0.0008	0.256	0.3150 ± 0.0328	0.543	0.6100 ± 0.0345	0.108	0.1350 ± 0.0242
Wolfgang Rolke	0.0103 ± 0.0008	0.356	0.3800 ± 0.0343	0.457	0.5250 ± 0.0353	0.184	0.2150 ± 0.0290
Stanford Challenge Team (SCT)	0.0077 ± 0.0007	0.3483	0.3550 ± 0.0338	0.4335	0.5200 ± 0.0353	0.0175	0.2100 ± 0.0288
Eilam Gross & Ofer Vitells	0.0082 ± 0.0007	0.35	0.3600 ± 0.0339	0.46	0.5250 ± 0.0353	0.19	0.2100 ± 0.0288
Valentin Niess	0.0111 ± 0.0008	0.603	0.3250 ± 0.0331	0.87	0.5300 ± 0.0353	0.12	0.1950 ± 0.0280
Georgios Choudalakis	0.0110 ± 0.0008	0.213	0.1600 ± 0.0259	0.290	0.3500 ± 0.0337	0.107	0.1300 ± 0.0238
Mark Allen	0.0106 ± 0.0008	0.385	0.4000 ± 0.0346	0.486	0.5250 ± 0.0353	0.187	0.2100 ± 0.0288
Frederik Beaujean	0.0138 ± 0.0009		0.0200 ± 0.0099		0.1800 ± 0.0272		0.1000 ± 0.0212
Stefan Schmitt							
Unbinned	0.0112 ± 0.0009		0.4500 ± 0.0352		0.5450 ± 0.0352		0.1850 ± 0.0275
Binned	0.0110 ± 0.0008	0.37	0.3850 ± 0.0344	0.53	0.5450 ± 0.0352	0.17	0.2200 ± 0.0293
Stefano Andreon							
$p < 3 \times 10^{-3}$	0.0126 ± 0.0013		0.4811 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120
$p < 4 \times 10^{-3}$	0.0191 ± 0.0016		0.5189 ± 0.0485		0.4766 ± 0.0483		0.0120 ± 0.0120

“Winners” – Problem 2

According to the criteria, these solutions were the best:

Tom Junk (but I don't count – I knew the answers!)

Stefan Schmitt

Eilam Gross & Ofer Vitells

Problem 2's priors/aux experiments have more ambiguity. What does $\pm 100\%$ mean on bg2?

Contributor	Type-I Error Rate Measured	Signal = 75 Events	
		Claimed	Measured
Tom Junk	0.0068 ± 0.0006	0.865	0.870 ± 0.017
Wolfgang Rolke	0.0256 ± 0.0012	0.88	0.8500 ± 0.018
Stanford Challenge Team	0.0389 ± 0.0015	0.84	0.9100 ± 0.0143
Eilam Gross & Ofer Vitells	0.0107 ± 0.0008	0.815	0.7725 ± 0.0210
Valentin Niess	0.0085 ± 0.0007	0.761 ± 0.001	0.7125 ± 0.0226
Stefan Schmitt			
25 Bins	0.0047 ± 0.0005	0.85	0.8200 ± 0.0192
50 Bins	0.0047 ± 0.0005		0.8250 ± 0.0190
Doug Applegate & Matt Bellis	0.0168 ± 0.0010	0.95	0.8950 ± 0.0153