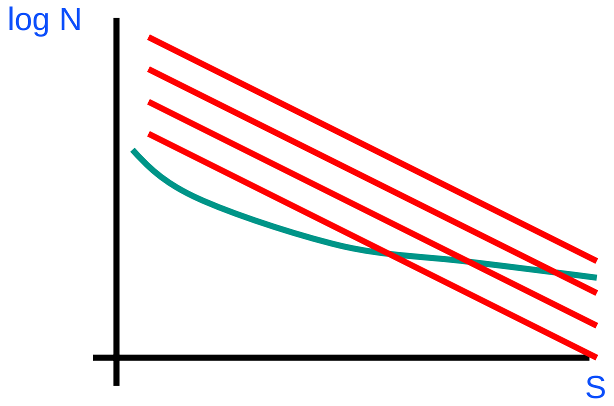
Analyzing Event Data

Lee Samuel Finn
Penn State University

Reference: T030017, T030041

Motivation

- The devil is in the details
  - » Noise obscures, confuses details (waveforms, estimable parameters, etc) in low S/N regime
- "Articulated events" capture principal signal features
  - » E.g., amplitude, duration, time, frequency, bandwidth, etc.
  - » Can be related to physical source characteristics
- Noise event numbers fall with amplitude fast
  - » New populations will emerge from well-defined tails
- More weak signal events than strong ones
  - » *9 of every 10 signal events have* S/N < 2.2 time threshold in isotropic dist; 3.3 times threshold for disk dist



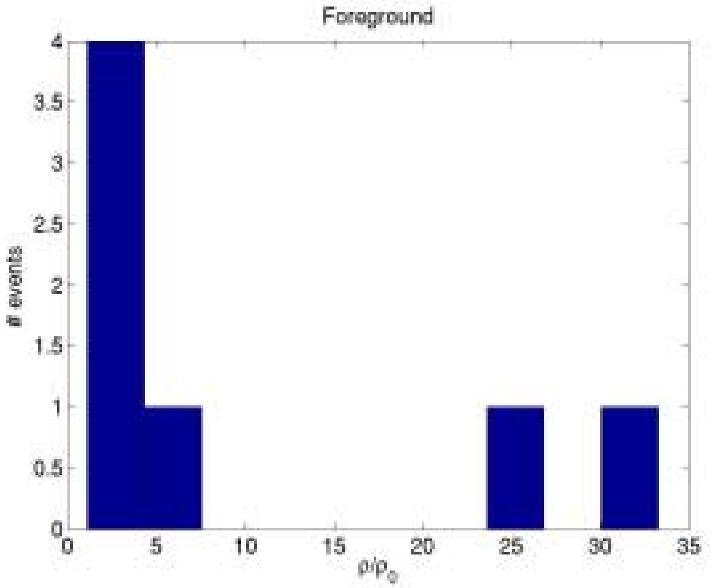
- More events, info/event, better bounds on source properties
- Examples in science
  - » Detection top quark
  - » GRBs are cosmological
  - » Cosmology (distance ladder, Hubble & other parameters, etc.)
  - » COBE & quadrupole anisotropy

From population model to foreground events

- Population model I
  - Sources:
    - Radiation in polarization modes, intrinsic strength, etc.
  - Distribution
    - Spatial, luminosity, other parameters
- Waves at antenna array: “source events”
  - **h**: polarization amplitudes, propagation direction
- Data processing pipeline J leads to “detected events”
  - Pipeline registers only fraction of source events, characterizes events phenomenologically
  - E.g. amplitude, frequency, bandwidth, source location, etc.

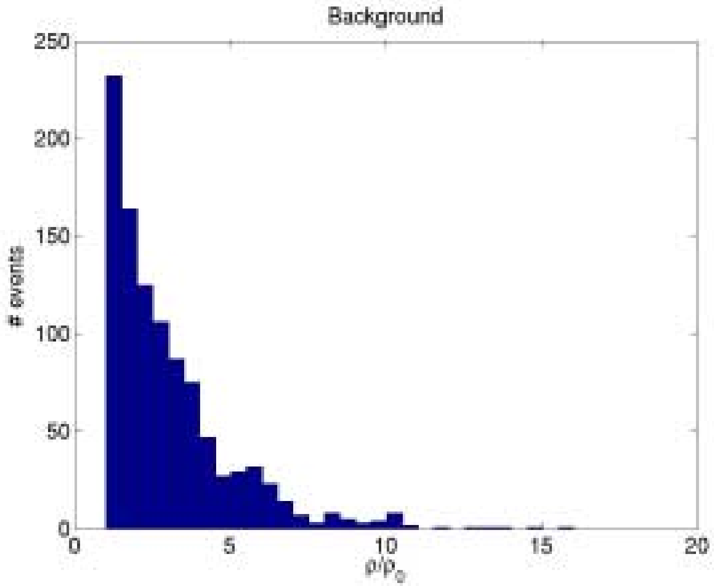
Characterizing detected events

- Detected source events: “foreground events”
  - » $P_F(\mathbf{H}|I\mathcal{J})$: distribution of detected events, owing to sources, in **H**
  - ≈ $\varepsilon(I\mathcal{J})$: fraction of all source events leading to detector events
  - » *Determined by simulation*
- Example: disk distribution
  - » $P(\rho)\sim 1/\rho^2$ for *power* signal-to-noise ρ
- At right:
  - » Draw # events from Poisson (10 expected, 7 actual)
  - » Draw event amplitudes from disk distribution



Background distribution

- Multiple detector correlations among most powerful analysis tools available
  - » Correlation or coincidence
- For event data, estimate distribution, rate from time-delay coincidence
  - » Multiple time delay fit to, e.g., mixture distribution model
  - » “Expectation maximization”
- Example:
  - » Thresholded linear filter output: Exponential distribution in power signal-to-noise
  - » Number drawn from Poisson distribution (1000 expected)



What we observe

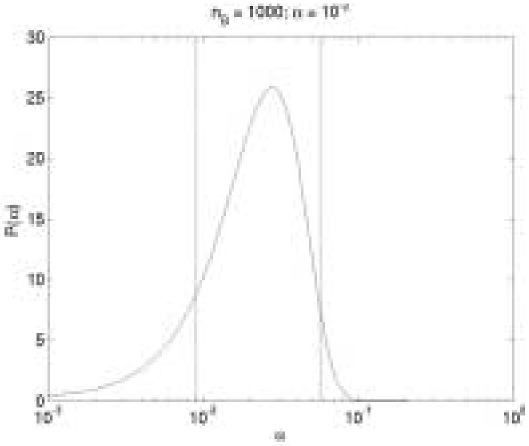
- Observed events are either foreground or background
  - Ratio of foreground number to background number is ratio of foreground rate(unknown) to background rate (known)
- $P(\mathbf{H}|IJn_Bn_S)$: Probability of observing a single event $\mathbf{H}$
  - $P(\mathbf{H}|IJn_Bn_S) = (1-\alpha)P_B(\mathbf{H}|J) + \alpha P_F(\mathbf{H}|IJ)$
  - $\alpha/(1-\alpha) = n_F/n_B$
  - ***Used for Frequentist analysis***
- $P(H|IJn_Bn_ST)$: Probability of observing N events $H = \{H_k : k = 1..N\}$
  - $P(H|IJn_Bn_ST) = P(N|\mu)\,\Pi_k P(\mathbf{H}_k|IJn_Bn_S)$
  - $P(N|\mu)$ is Poisson distribution; $\mu = T[n_B+\epsilon(IJ)n_S]$
  - ***Used for Bayesian analysis***

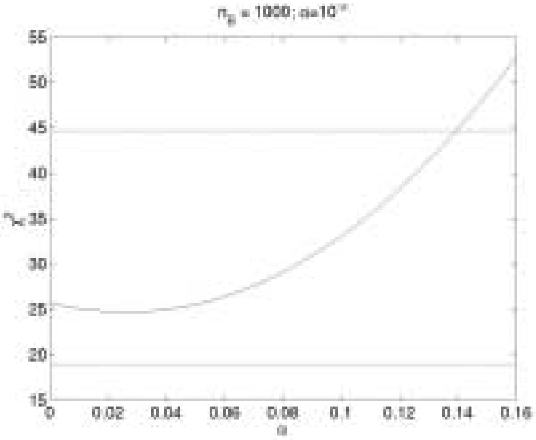
A Frequentist Analysis

- How well does observed distribution fit expected distribution $P(\mathbf{H}|IJn_B n_S)$?
  - N events sample $P(\mathbf{H}|IJn_B n_S)$
  - Evaluate χ^2 test statistic
  - $\chi^2 = \chi^2(H|n_B n_S T I J K)$
- Find interval χ^2 that encloses probability p of χ^2 distribution
  - Choose smallest χ^2 interval
- *For what range of n_S* is χ^2 in probability p interval?
  - Like a CI, but not a CI:
    - CI: range of n_S for which observation is likely with probability p
    - Here: range of n_S for which χ^2 is likely with probability p
- Automatically incorporates “goodness-of-fit” test
  - If observed distribution does not fit well to expected distribution for any n_S, no range of n_S reported

Example

- Disk population, Rayleigh noise
    - » n_S/n_B = 1/100
- Analysis: “See” all events with S/N above threshold
- Expect 1000 background events
    - » Actual number background, foreground Poisson
- Typical result 90% confidence
    - » Bayesian analysis (flat prior) bounds n_F away from zero
    - » Frequentist analysis sets upper limit n_S/n_B<0.14





LIGO-G030065-00-Z

Compare …

- “Excess event” analysis
  - Detection of excess @ 90% confidence requires # observed events greater than ~ $1.5n_B$
  - $n_F/n_B = 1/100$ to $n_F/n_B = 1/2$ requires increase threshold by factor 14
  - After increase, expect 0.7 foreground, 1.4 background!
  - “Detection efficiency” 15%
    - Will have one or more foreground event only 15% of times you look
    - Compare 46% of cases will have Bayesian bound on $n_{F\,\backslash}$ away from 0
- Why is distributional (“log S/log N”) analysis so much better?
  - Populations emerge in the tail
  - Mass of distribution provides context , anchor for measuring, interpreting tail
  - Without the mass of distribution, tail wags dog

Summary & Conclusions

- Source and source population properties are revealed in observed event distribution properties
  - Axi- vs. non-axisymmetry, spatial distribution (disk, sphere), etc., all reflected in observed distribution in amplitude (& frequency, bandwidth, etc.)
- Study event distributions to identify, bound character of sources, source populations
  - Models can be fit to observed event distributions
  - Rate, spatial distribution, luminosity, other properties
  - Bayesian analysis straightforward; Frequentist analysis based on χ^2 statistic
- ***Distributional analyses have greater sensitivity, are more robust against small number statistics***
  - Dig deeper into noise
  - More events make analyses more robust than low-number statistics, single event, low-background analyses

Moral: use coincidence to estimate background & drop thresholds!