Class 3: Statistical building blocks

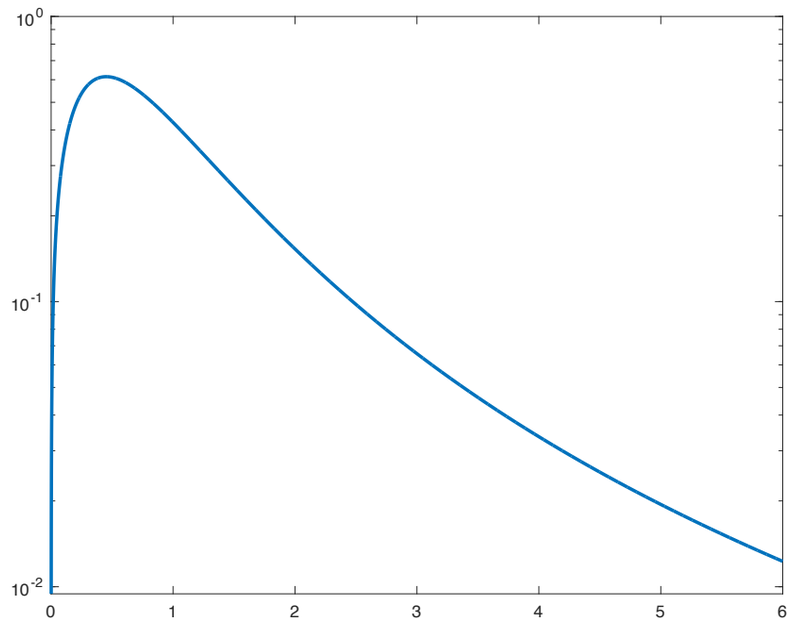
Miguel F. Morales
Bryna Hazelton

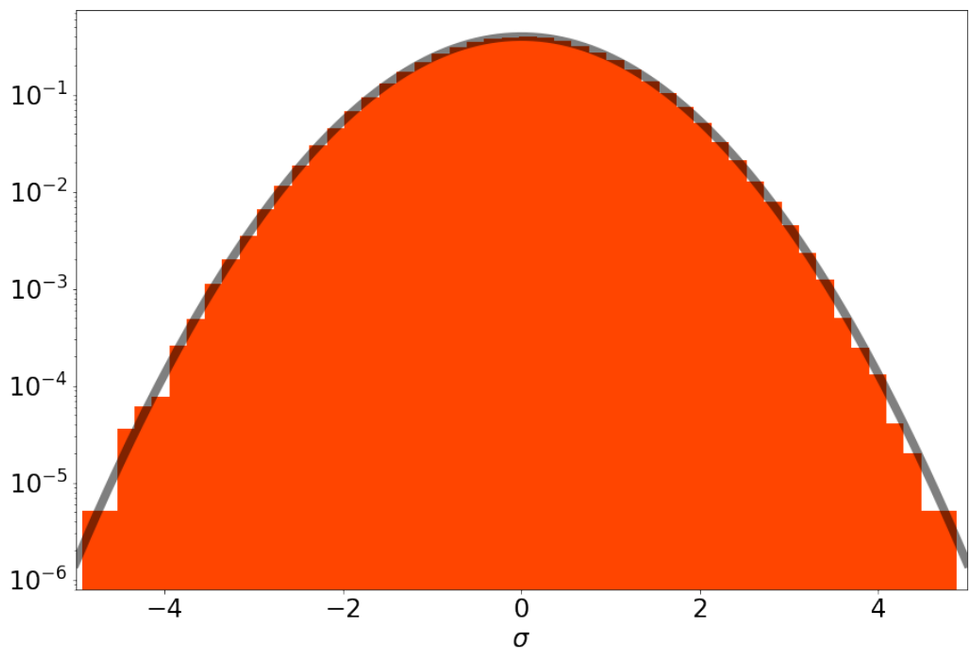
Outline

- Review
- Finding your background distribution
- Common statistical distributions
- Analytic propagation & analysis chains
- Convolution & central limit theorem

Key statistical steps

- Clearly state the question (& turn into math)
- Determine the background distribution
- Integrate background to find probability
- Convert probability into equivalent sigma

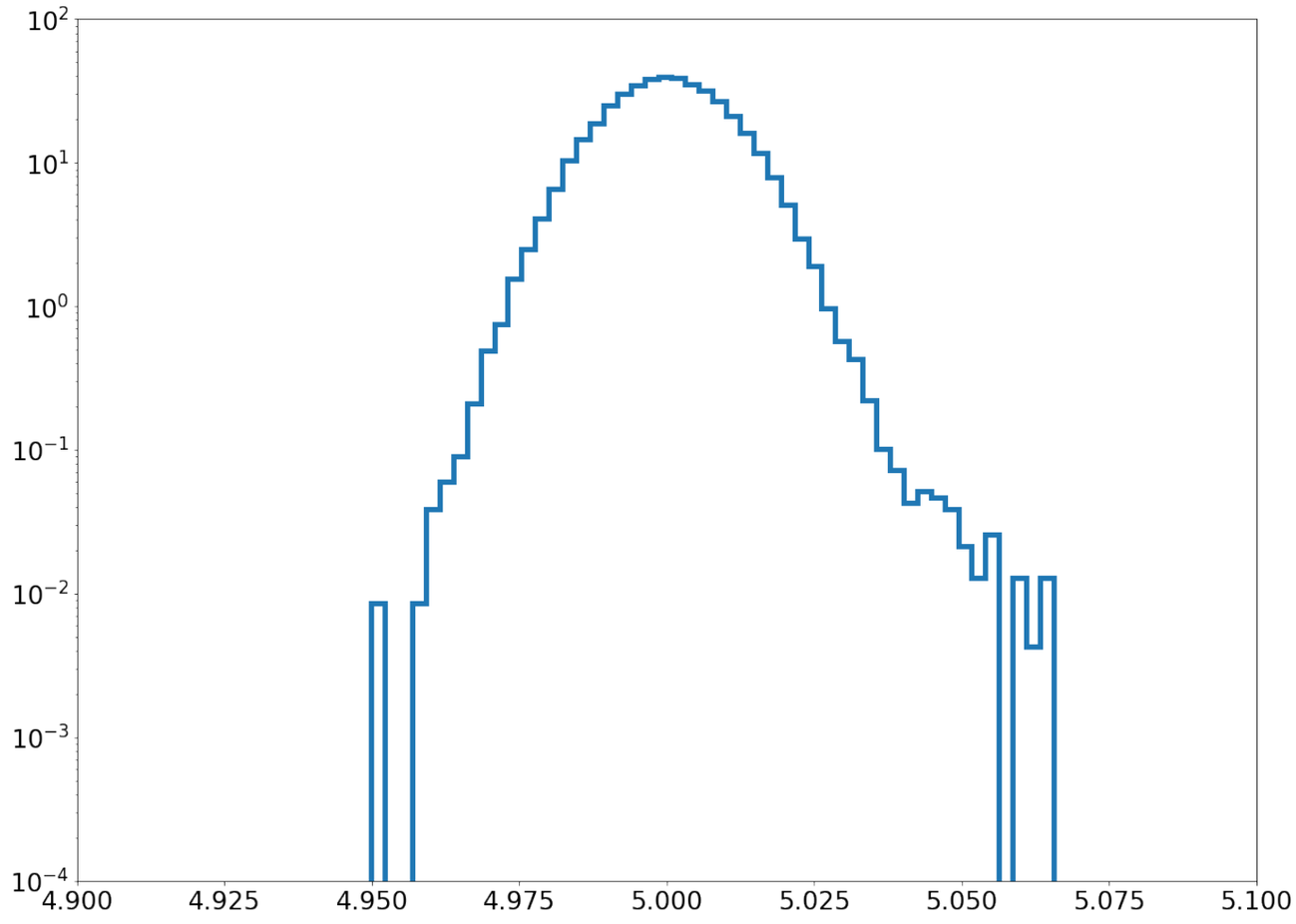




Measuring the background

Two related issues:

- Finding signal-free data
- Finding the ‘shape’ of the background



Measuring the Background

Take signal-free data

LSST calibration

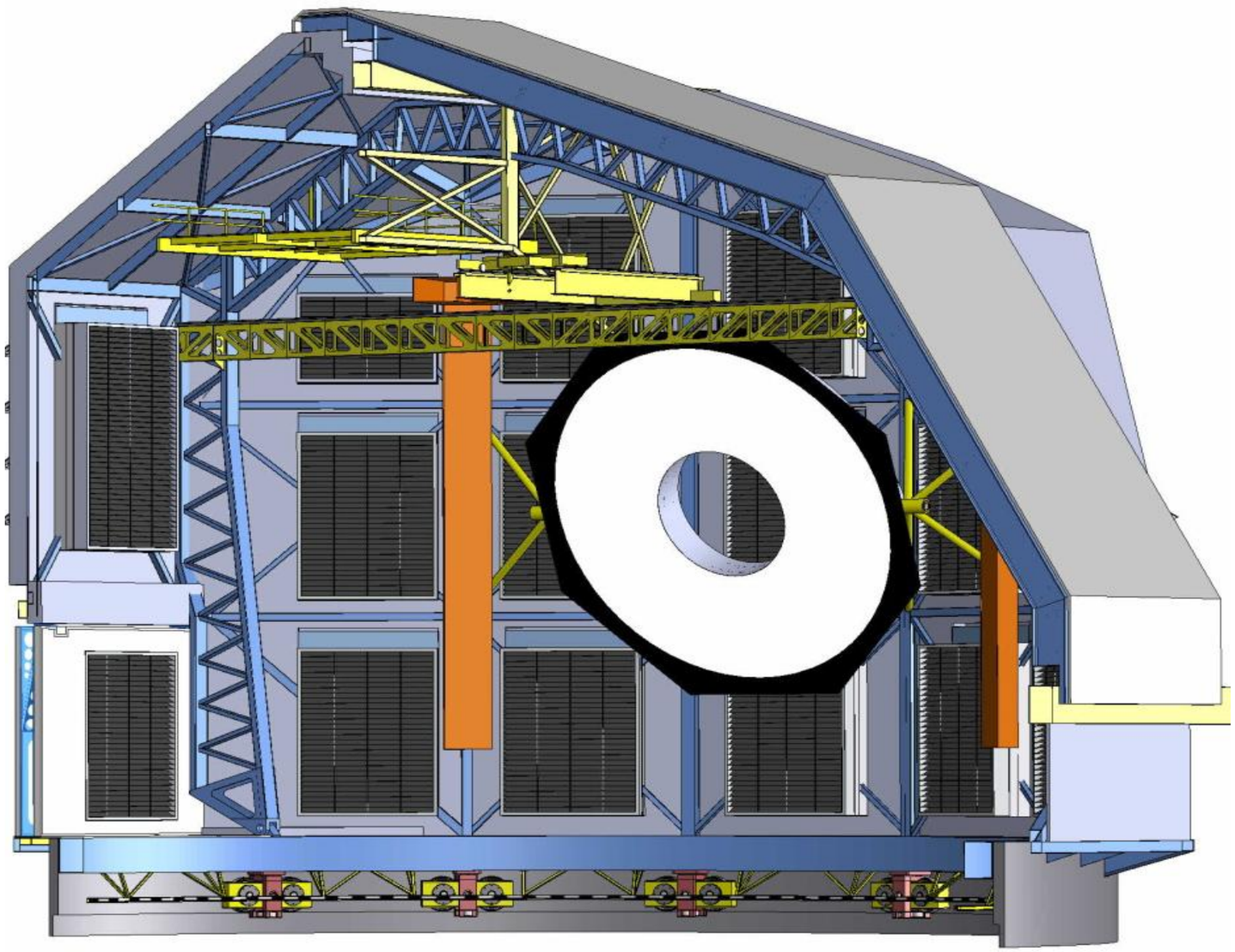


Figure 1: Calibration screen located inside the LSST dome

10 meter screen inside the dome

LSST screen specs:

- Illuminated with white light (UV to IR) and a tunable laser
- Emission must be known to 0.2% across the surface
- Used twice a day (afternoon & morning)
- Must not take more than 4 hours(!)

Manipulate data to make signal disappear

Examples:

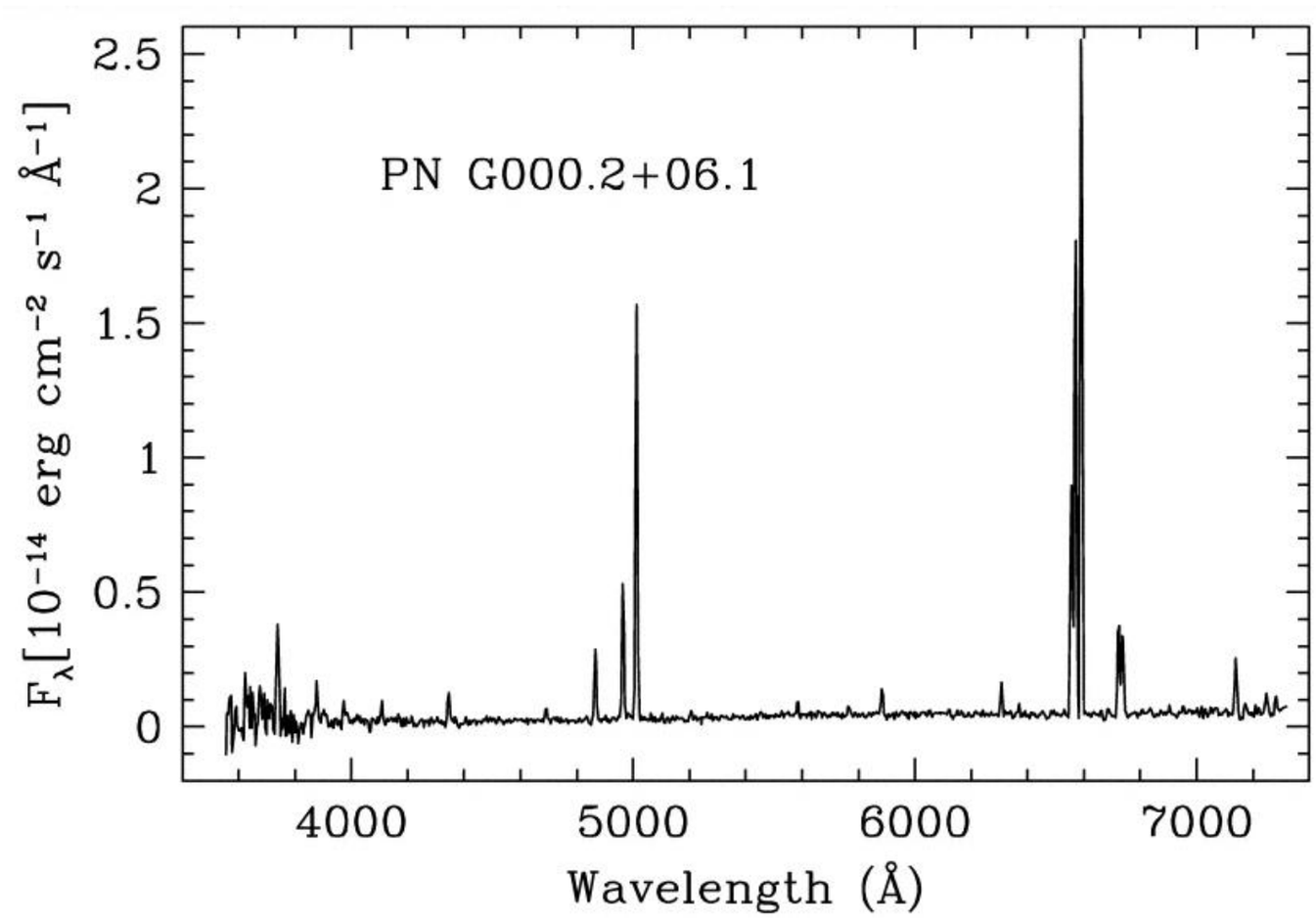
- Randomly shuffling the data so sources disappear
- Subtracting neighboring datasets
- Requires signal be slowly varying

Hope for isolated or rare signals

Isolated signals

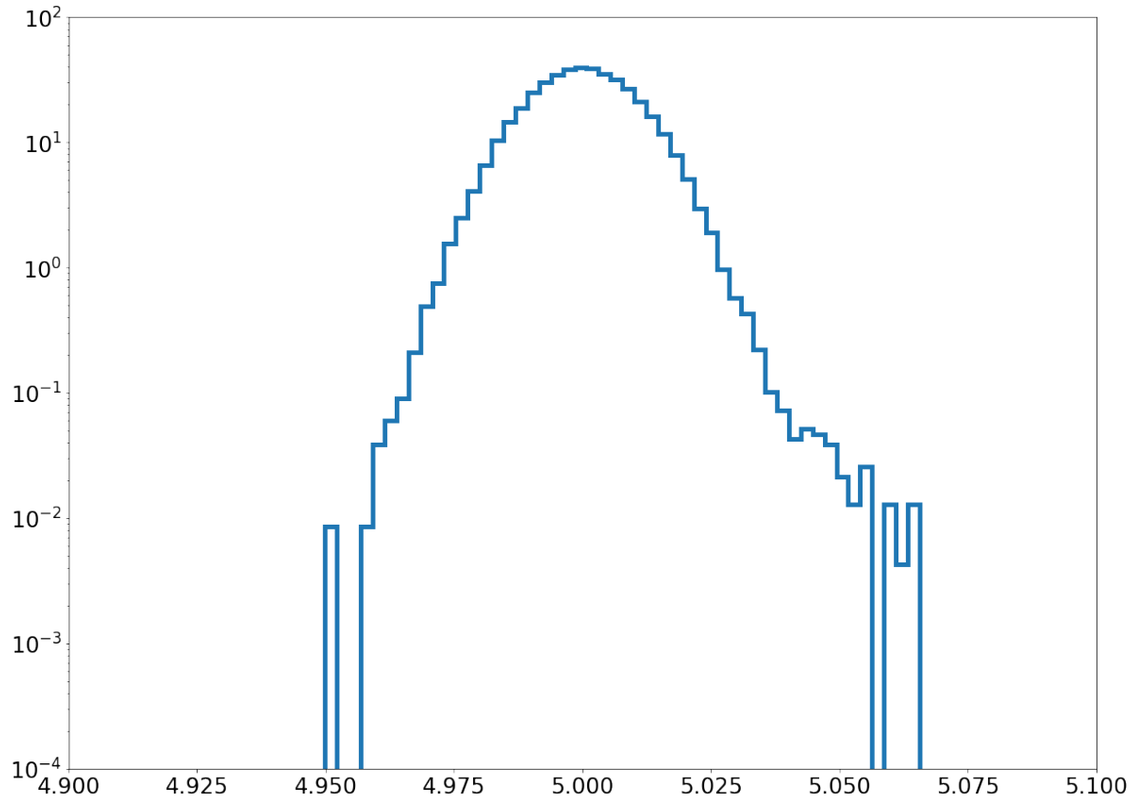
- Identify energies, times, or locations where you don't think there is a signal and extrapolate to the region of interest

- Explicit assumptions about background behavior



Rare signals

- Find a time or area where you are fairly certain there is no signal

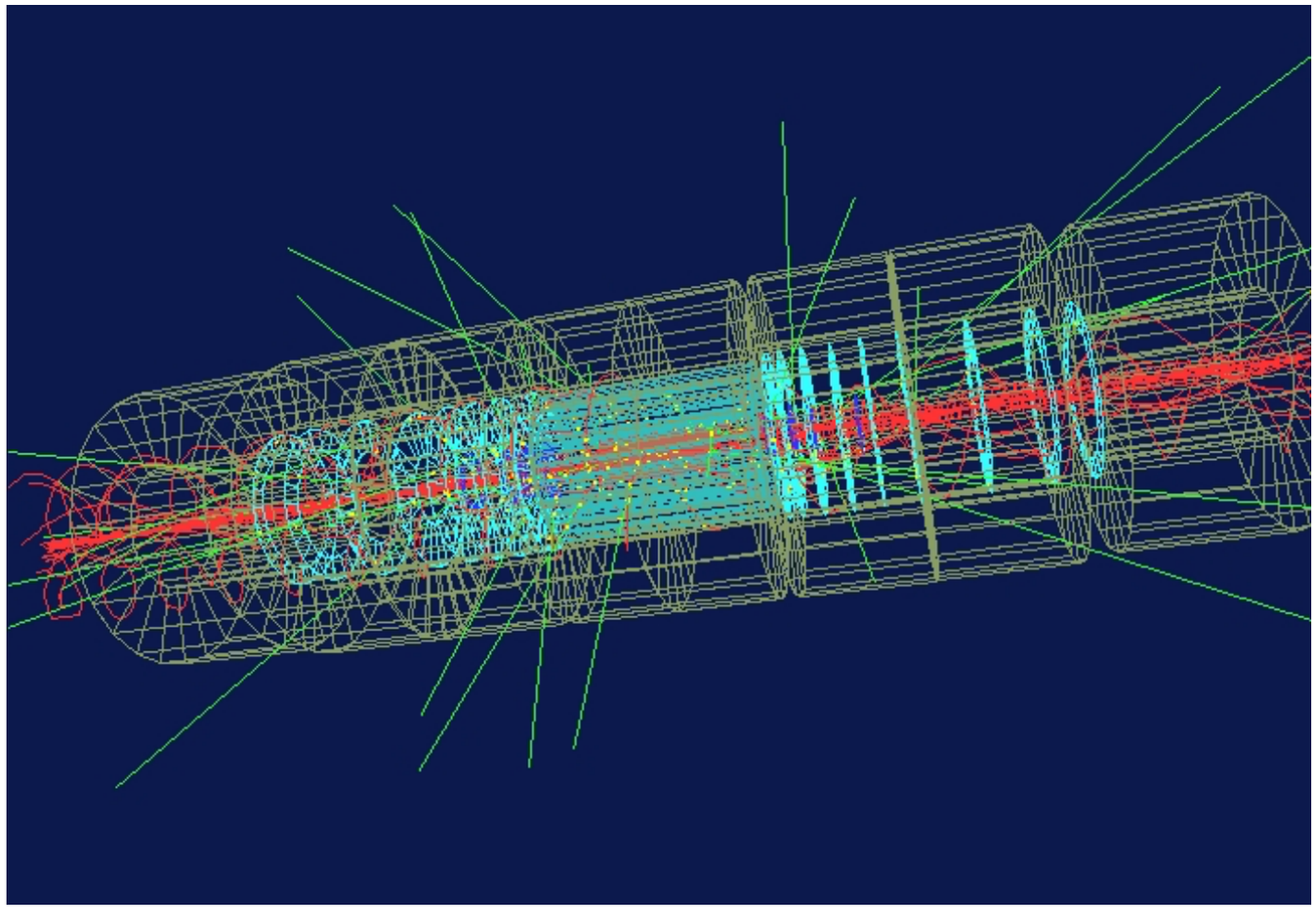


Simulate your background (Monte Carlo)

Make fake data without your signal

GEANT

- Purpose built particle physics simulator
- Entire instruments in simulation
- Conferences on how to make it more precise



Simulations fall in two categories

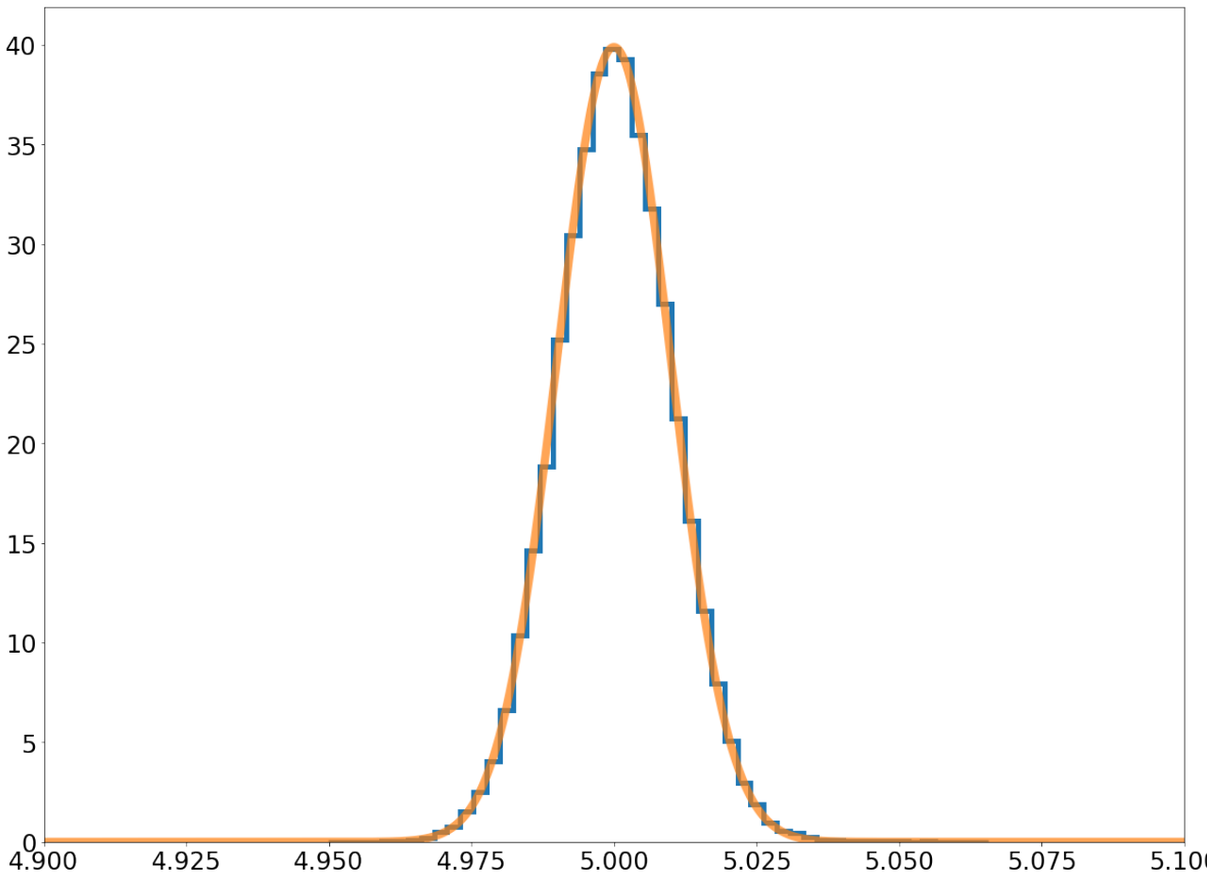
- Helping to understand how the instrument works
- Calculating fake background data (must be much more precise)

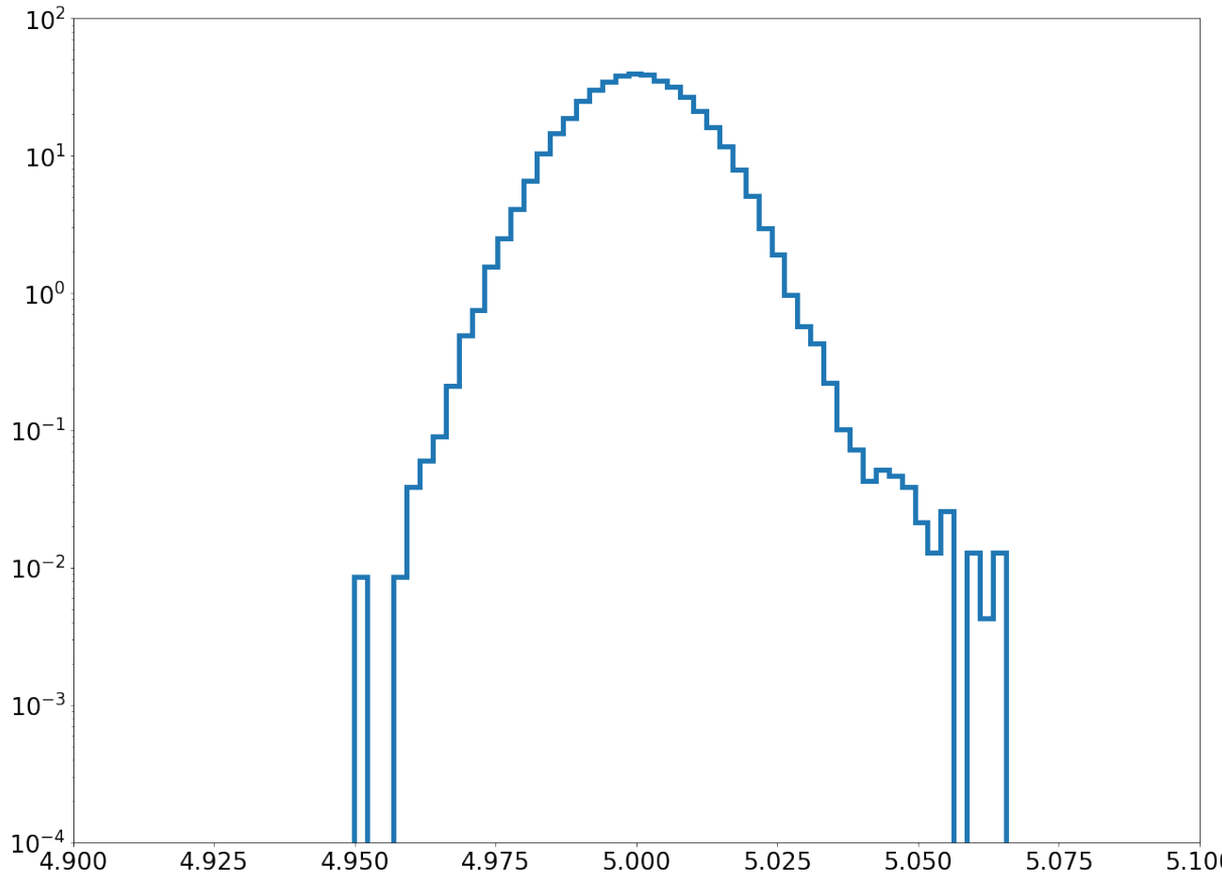
Step-by-step guide to Parametrizing the Background

1) Determine how to measure ‘signal-free’ background

- Take signal-free data
- Manipulate the data to remove the signal
- Hope for isolated or rare signals
- Monte Carlo

2) Make a histogram of background data on log plot



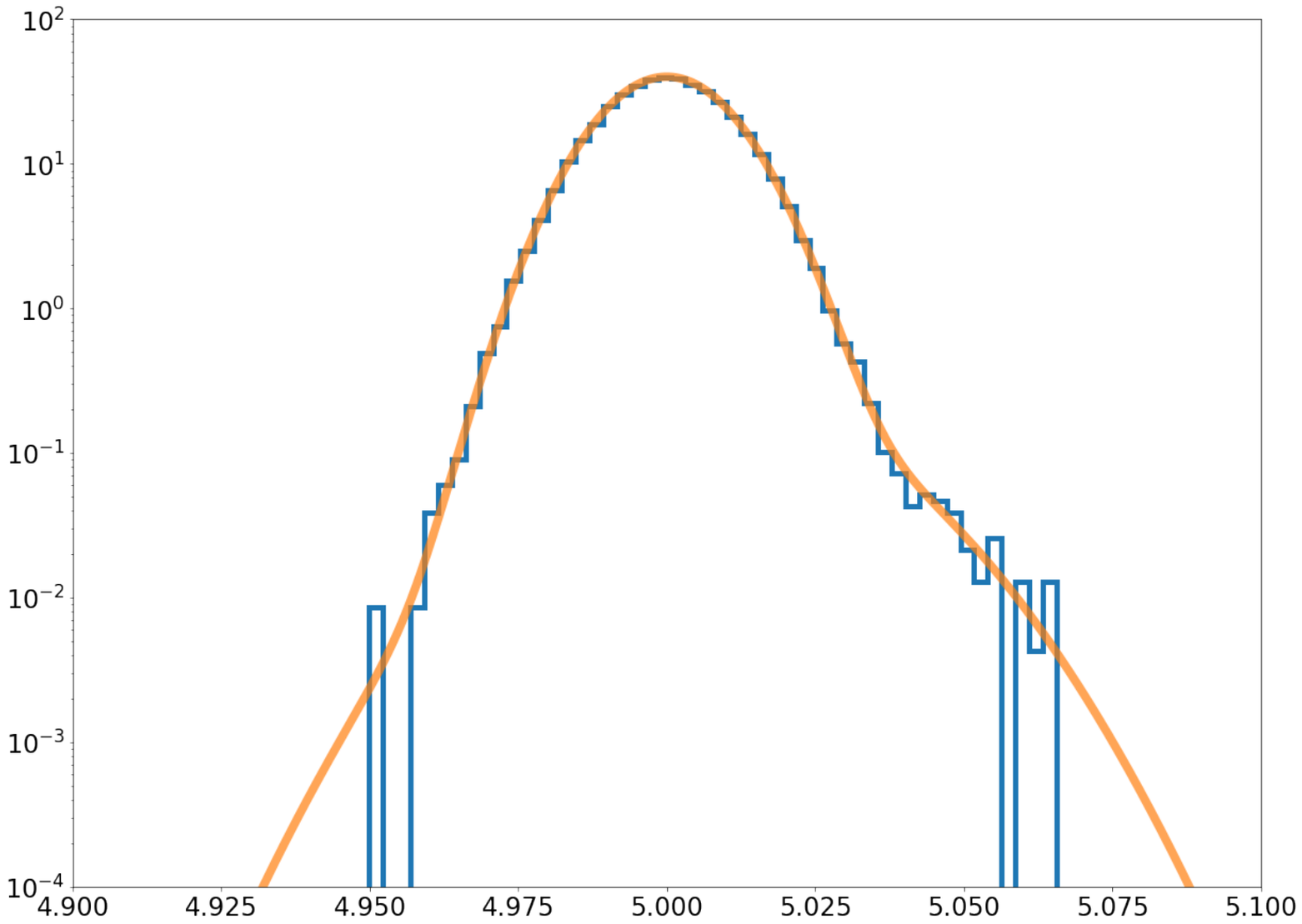


Always use a semilog plot!

3) Think about what kinds of pdfs you might expect

- What creates the background?
  - Thermal noise, radioactive backgrounds, cosmic rays, other uninteresting sources, …
- What systematics should I worry about?
  - Is background stable in time, detector, energy, …

4) Use to build a model of the background



The better the background model, the better the science

Useful statistical distributions

Matlab

Analysis chains & analytic propagation

Data

Data cleaning

Calibration

Background removal

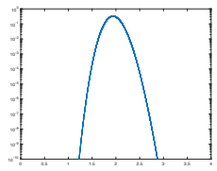
Intermediate integration

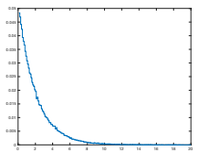
Quality assessment & cleaning pt 2

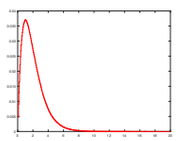
PS transform

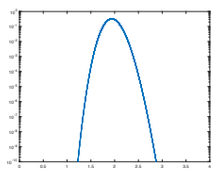
⋮

Error Model









⋮

Worries

Thunder storms

Biasing result

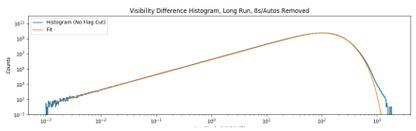
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

⋮

Example

- 2D symmetric Gaussian (normal); amplitude is Rayleigh

Data

Data cleaning

Calibration

Background removal

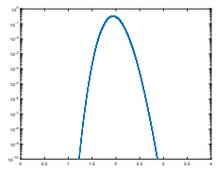
Intermediate integration

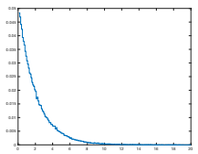
Quality assessment & cleaning pt 2

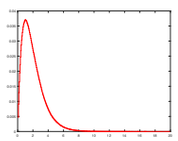
PS transform

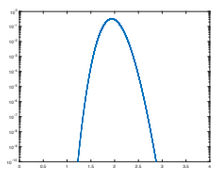
⋮

Error Model









⋮

Worries

Thunder storms

Biasing result

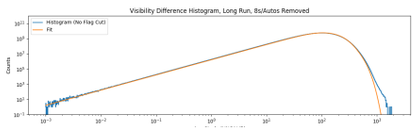
Temperature dep. offset

Signal leakage

⋮

Tests

Jackknife



Correlation

Injection test

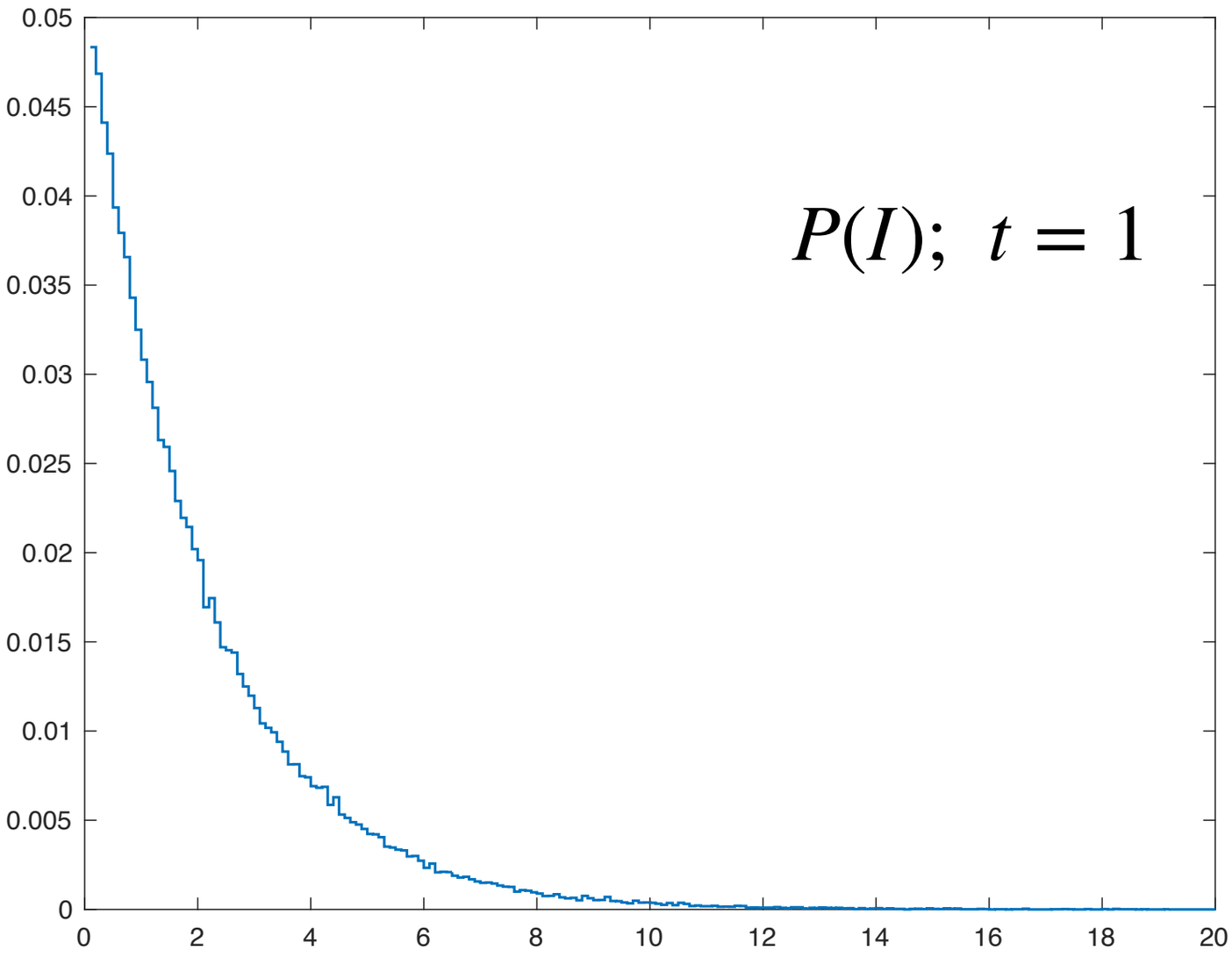
⋮

How distributions change

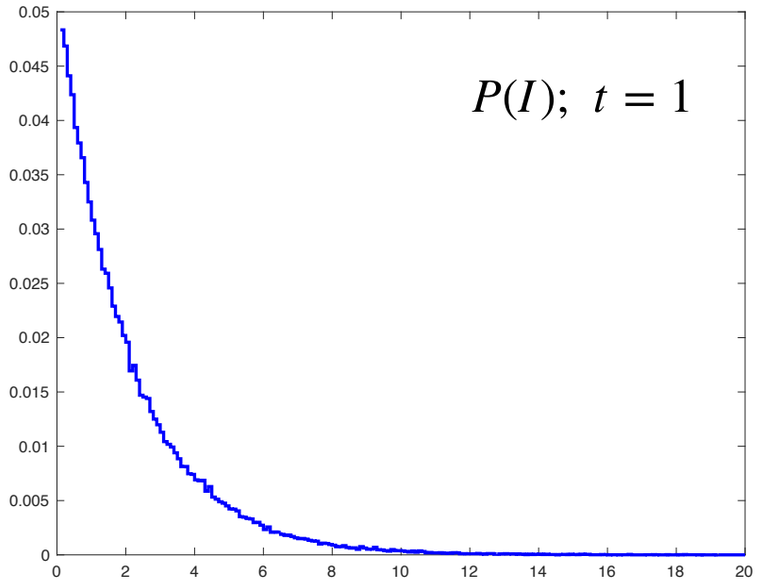
Convolution & the central limit theorem

Example: power of random electric field

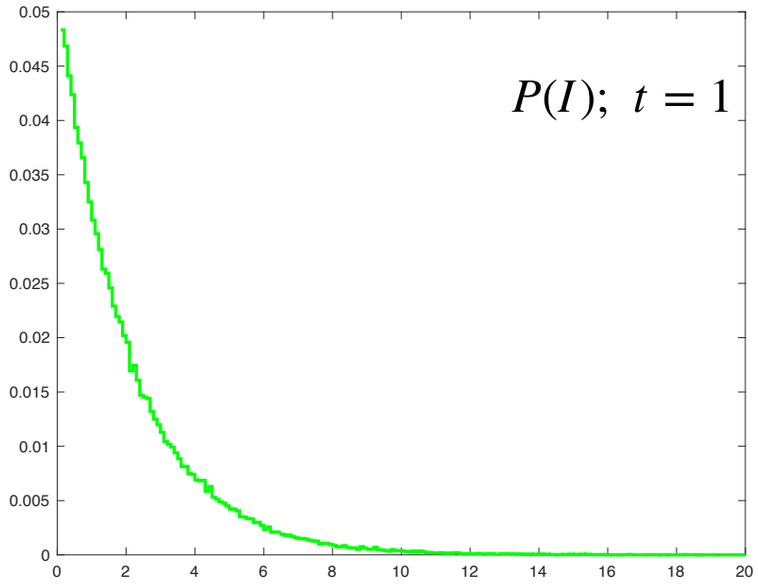
$$I = \langle E^{\dagger} E \rangle_t$$

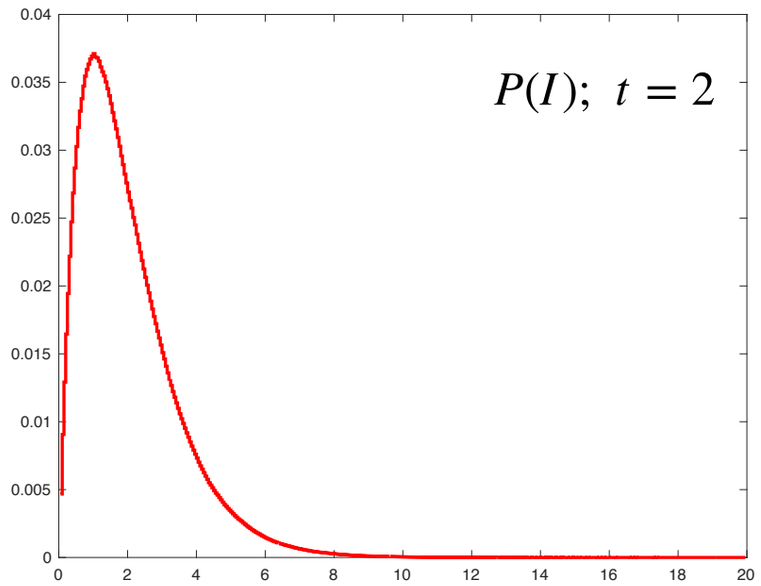


Average (or sum) is convolution of pdfs



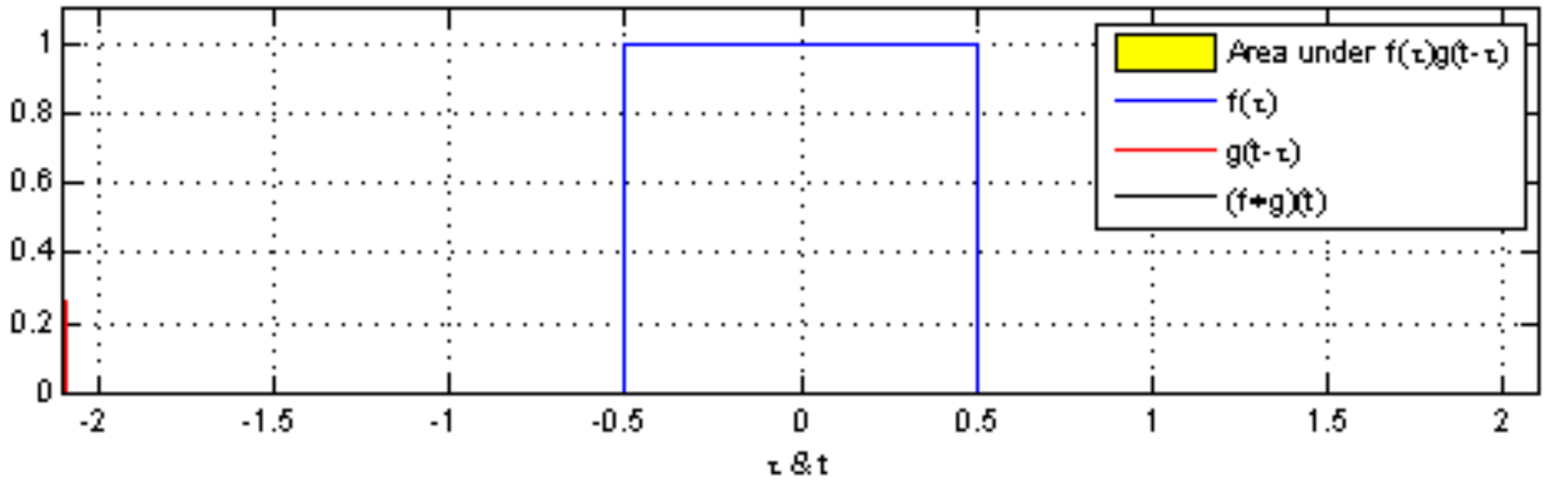
*

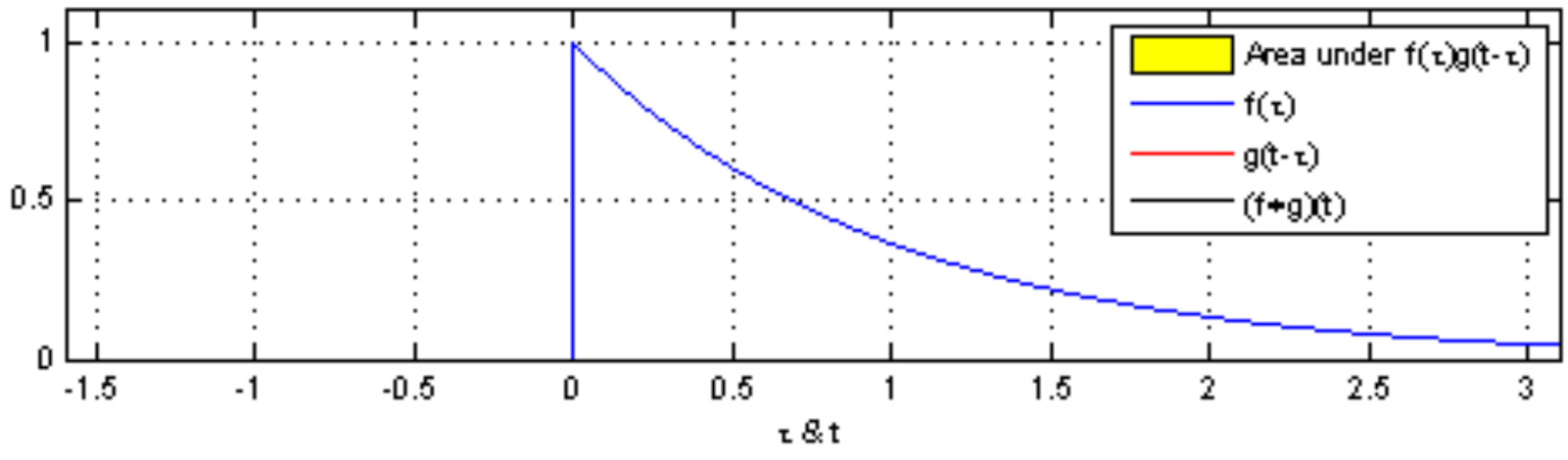




Convolutions

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau$$





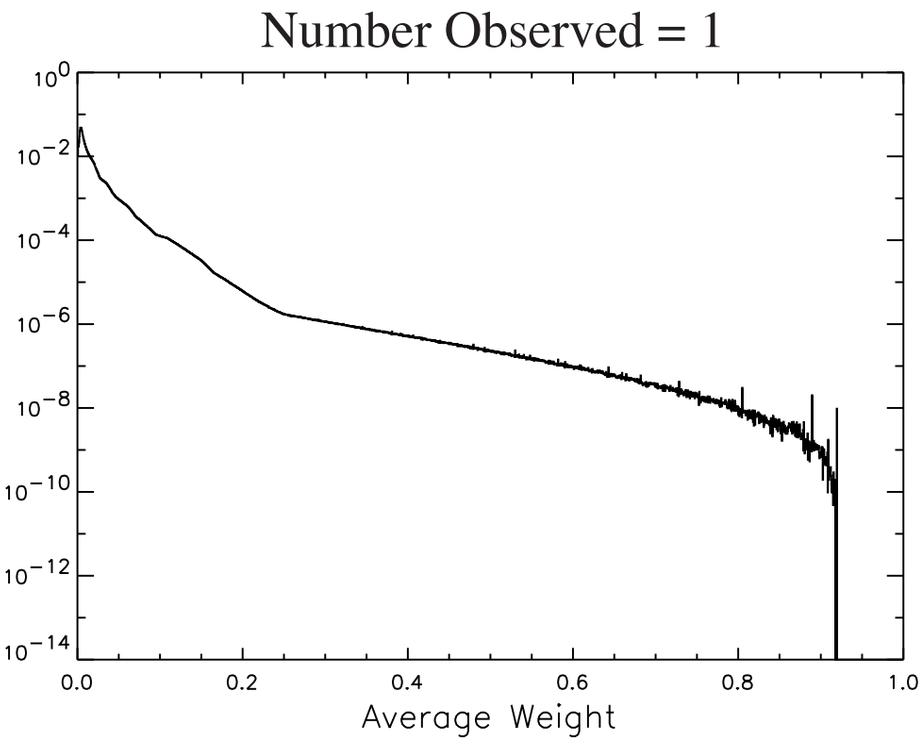
Brian Amberg

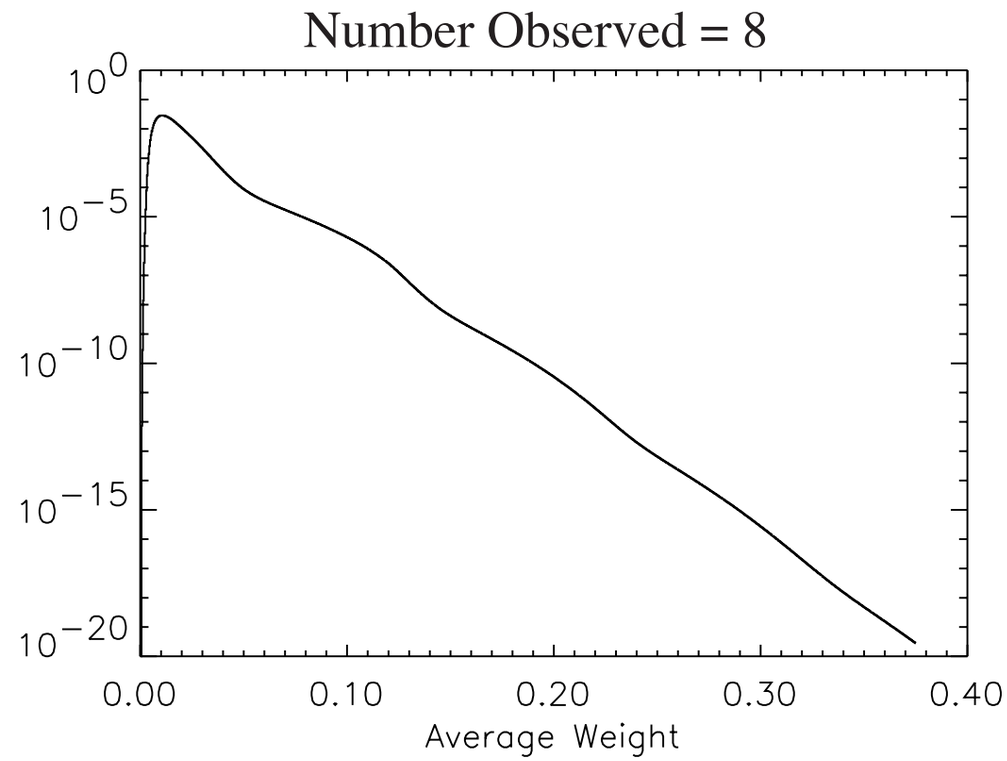
Matlab playing

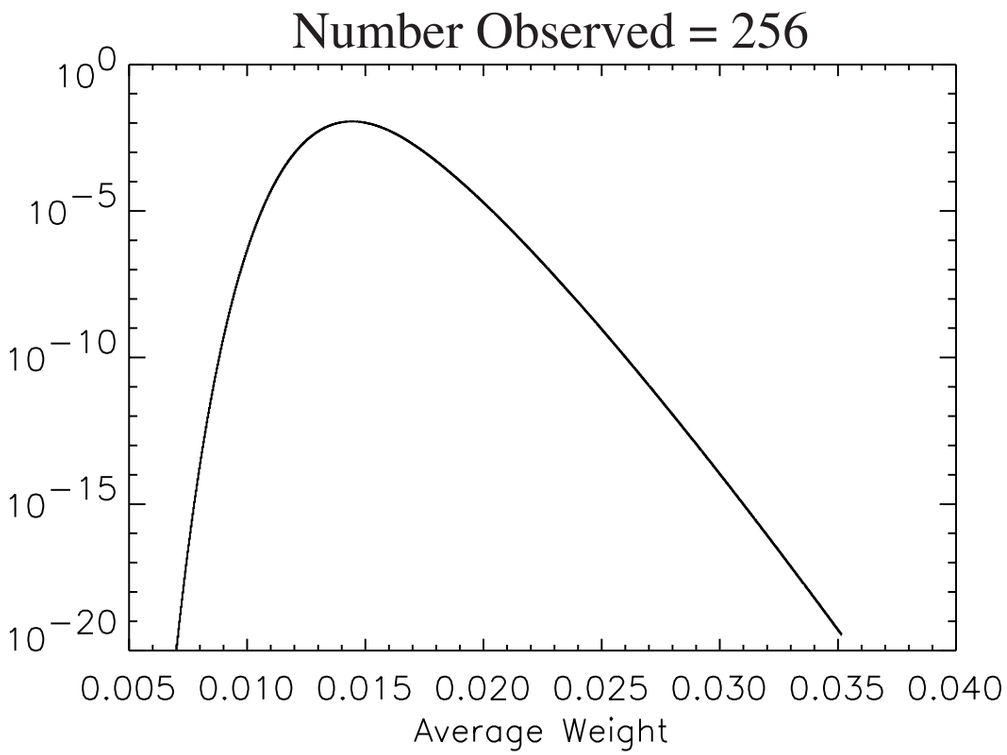
Probability distributions of averages

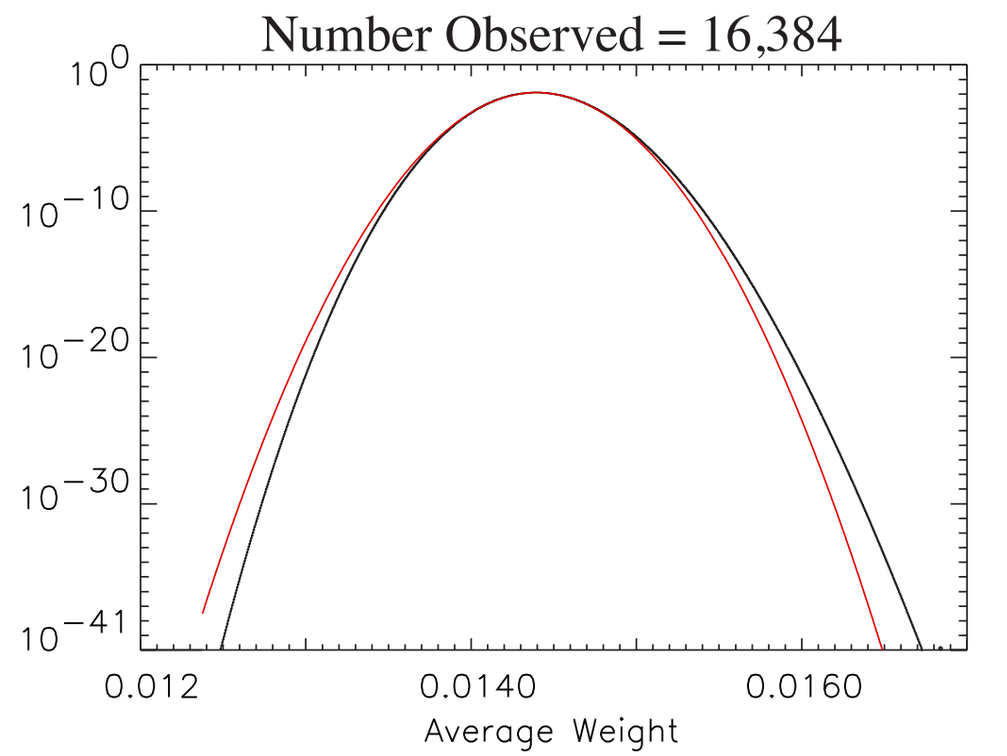
- Averaging involves repeated convolution of pdfs
- Leads to central limit theorem
- *Usually* converges quickly to a Gaussian distribution

My thesis


Number Observed = 1
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
10^{-12}
10^{-14}
0.0
0.2
0.4
0.6
0.8
1.0
Average Weight


Number Observed = 8
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
0.00
0.10
0.20
0.30
0.40
Average Weight


Number Observed = 256
10^{0}
10^{-5}
10^{-10}
10^{-15}
10^{-20}
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Average Weight


Number Observed = 16,384
10^{0}
10^{-10}
10^{-20}
10^{-30}
10^{-41}
0.012
0.0140
0.0160
Average Weight