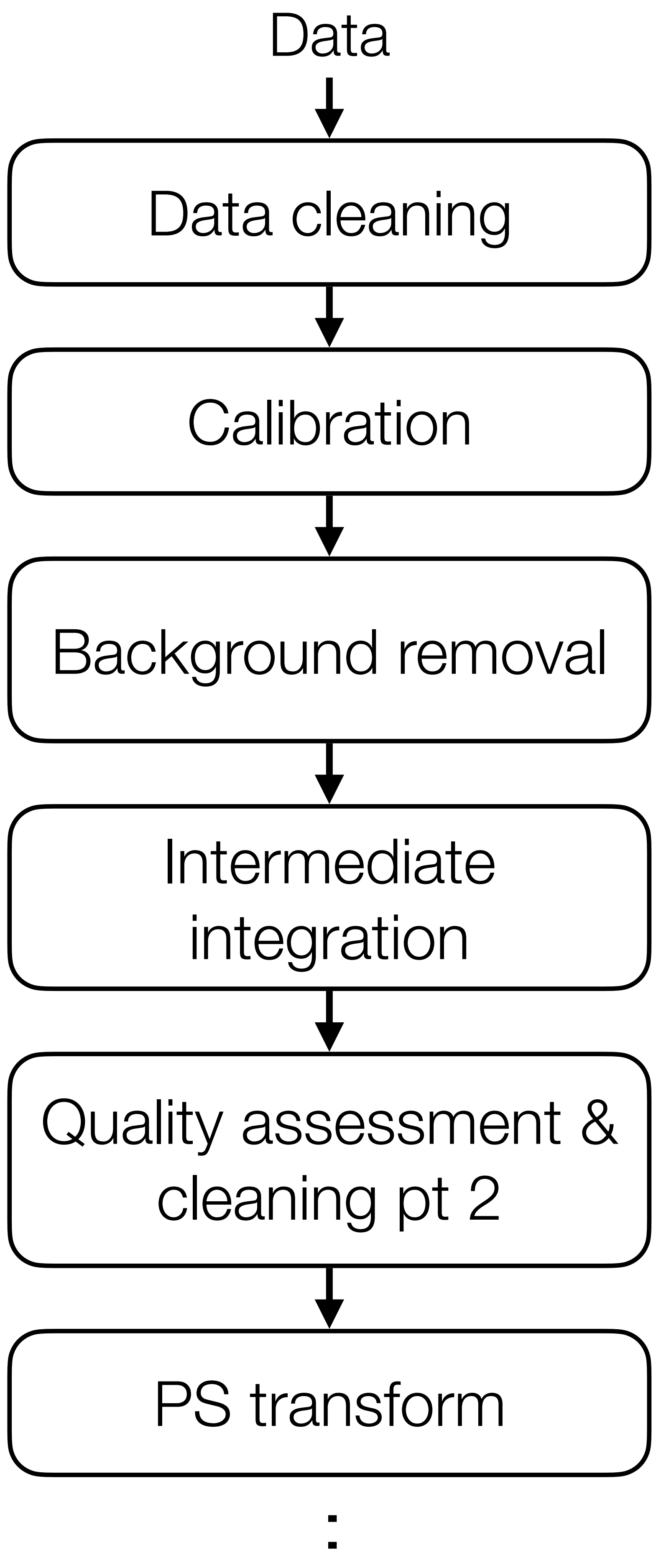
Class 3: Statistical building blocks

Miguel F. Morales
Bryna Hazelton

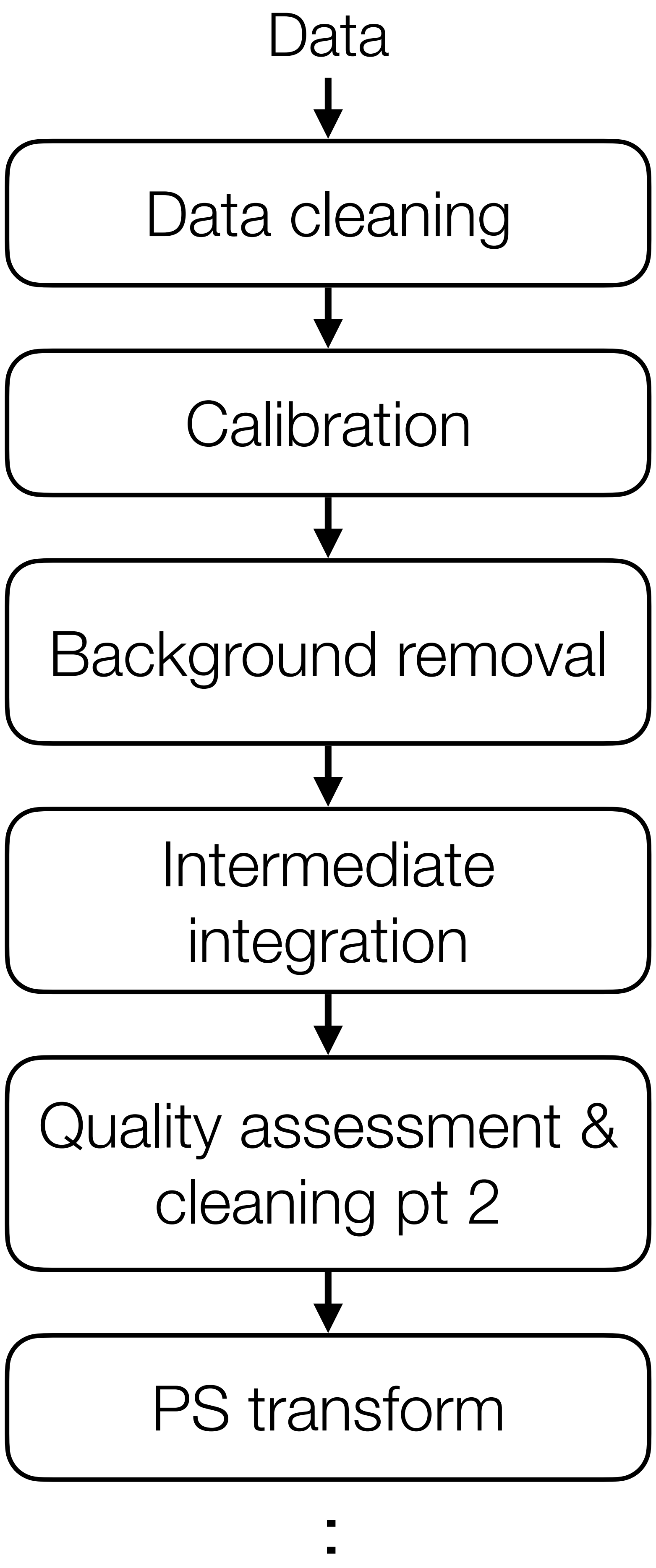
Outline

- Analysis Chains
- Common statistical distributions
- Convolution & central limit theorem
- Finding your background distribution

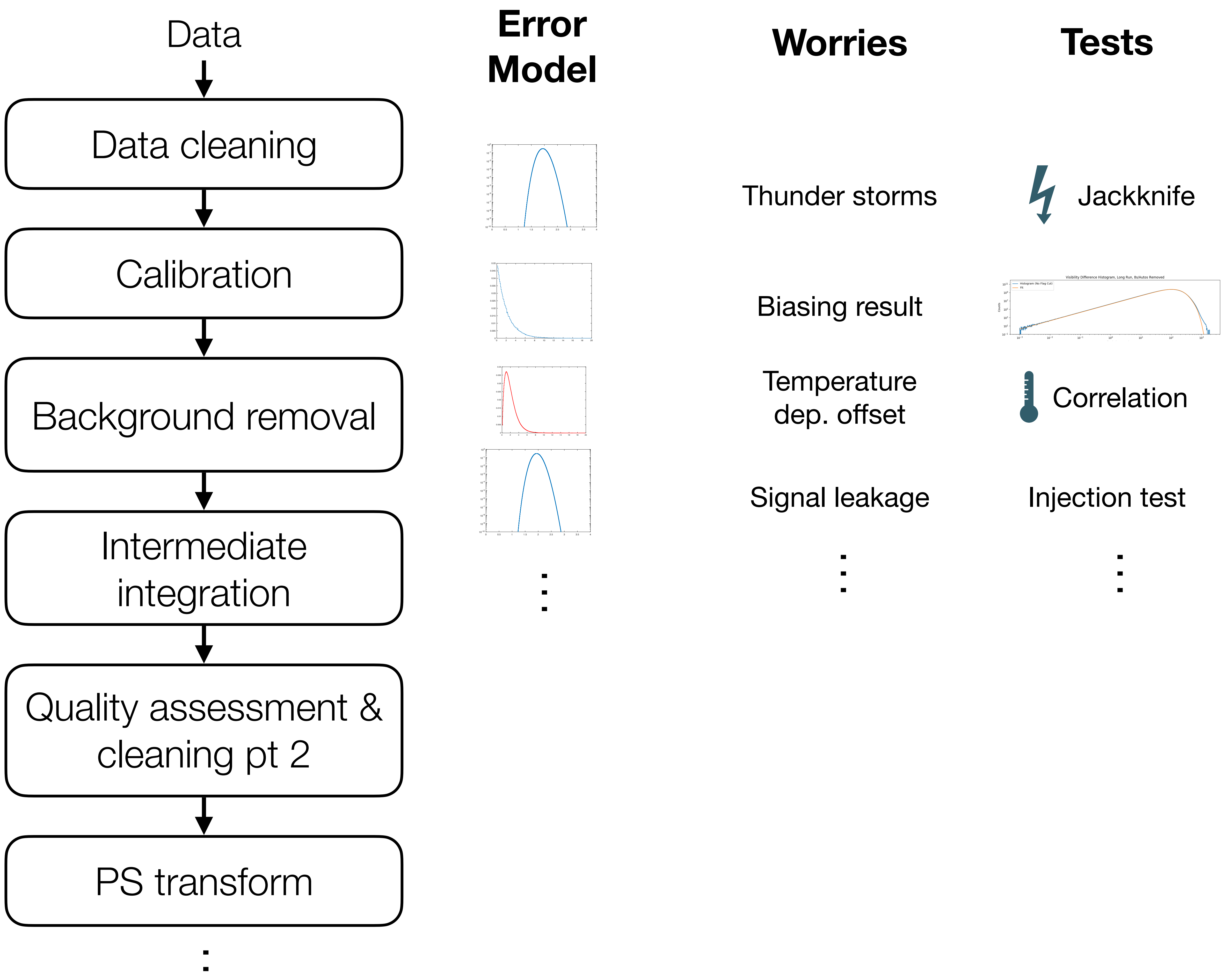
Analysis chains



How do you know the analysis is right?



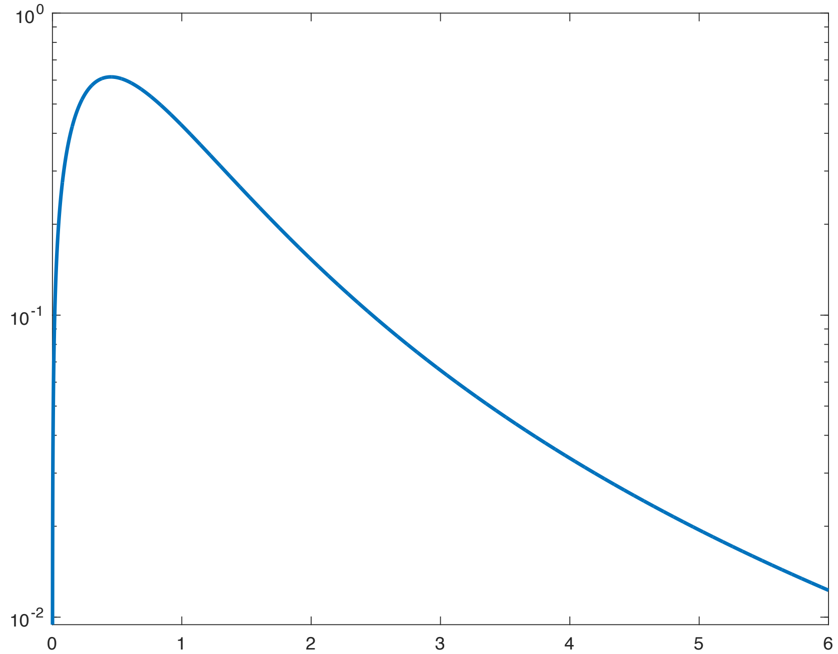
Is it right at step X? (Newton's method of bug finding)

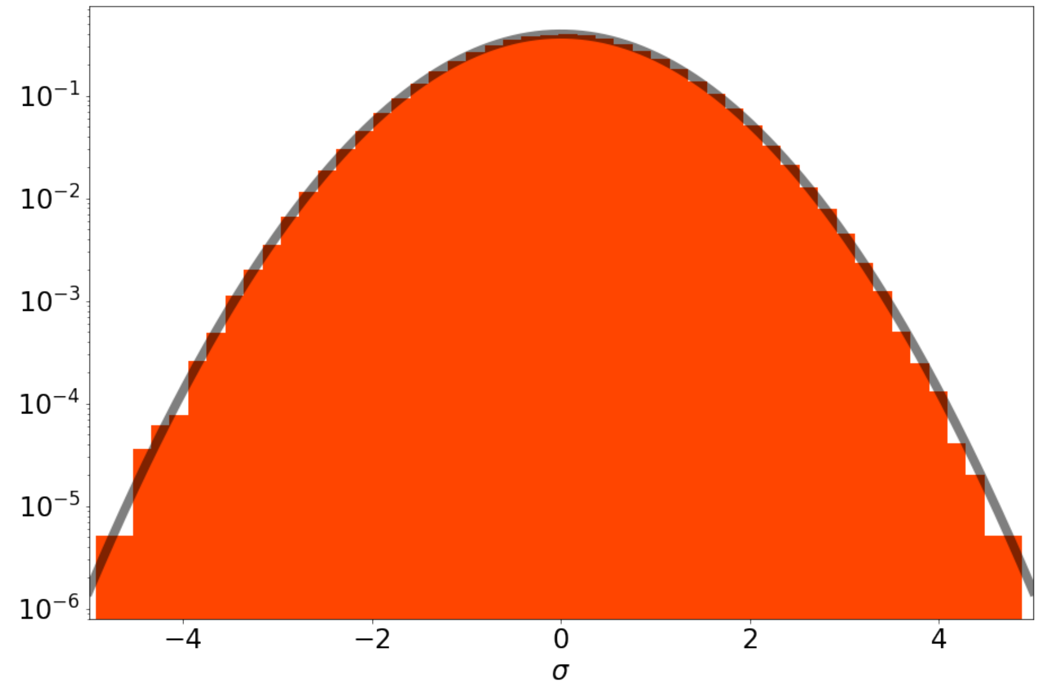

Data
Data cleaning
Calibration
Background removal
Intermediate integration
Quality assessment & cleaning pt 2
PS transform
Error Model
Worries
Tests
Thunder storms
Jackknife
Biasing result
Temperature dep. offset
Correlation
Signal leakage
Injection test

Useful statistical distributions

Key statistical steps

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





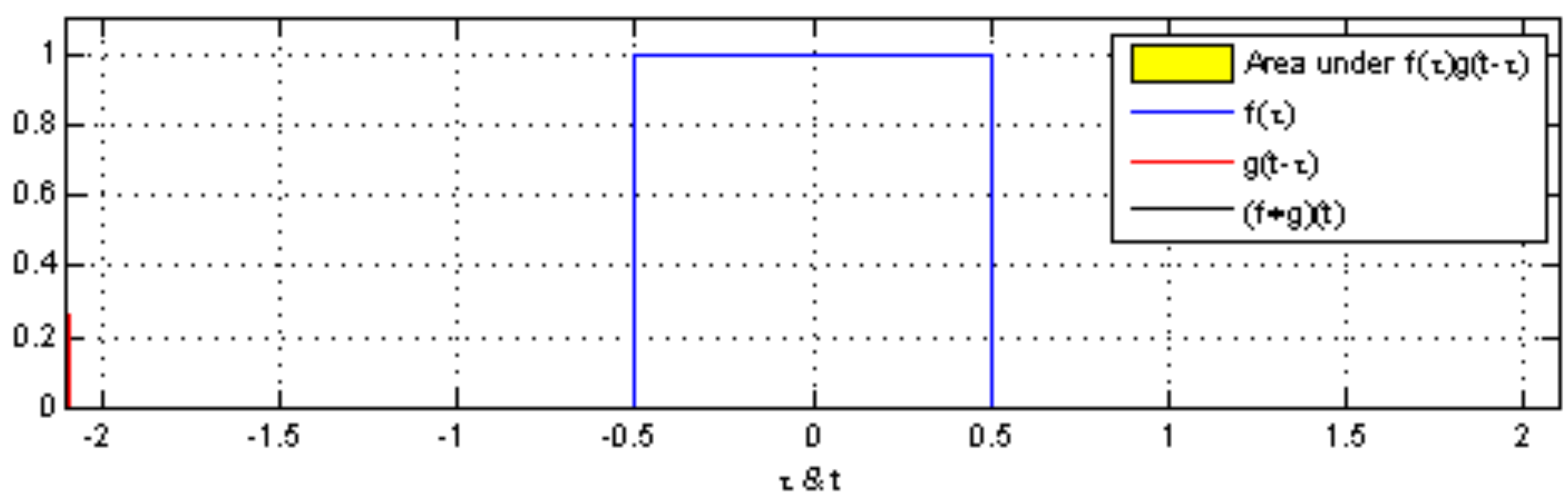
Matlab playing

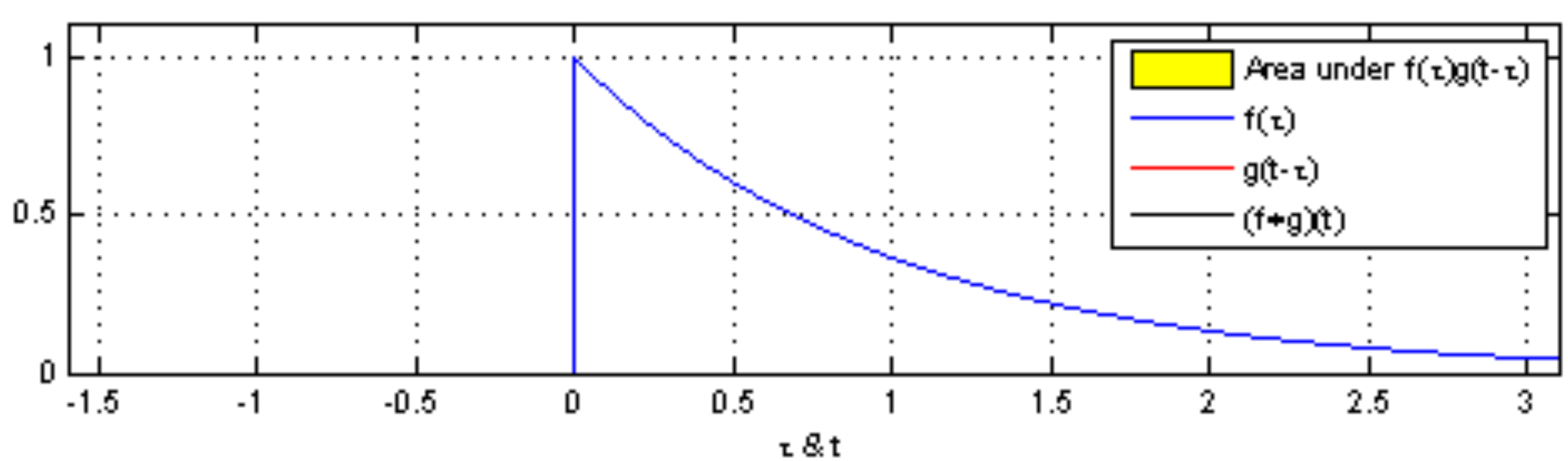
How distributions change

Convolution & the central limit theorem

Convolutions

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





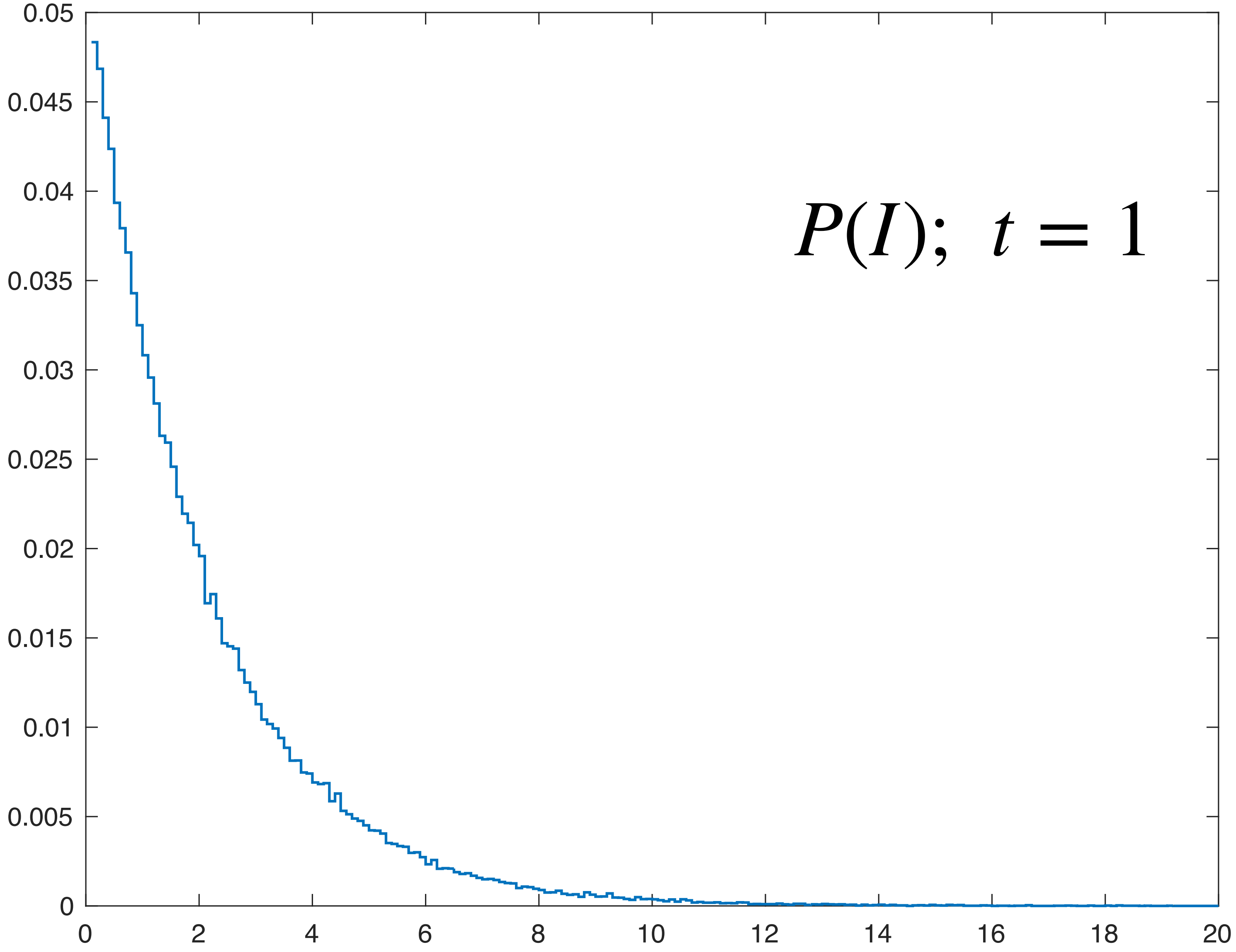
Brian Amberg

Sum & averaging

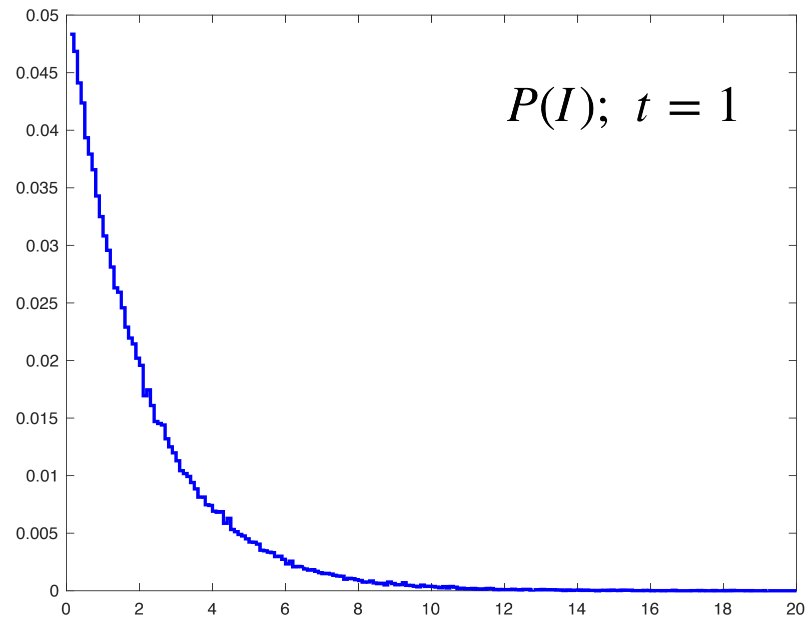
- Sum: pdf of result given by convolution of input pdfs
- Averaging is a sum with a rescaled x axis

Example: power of random electric field

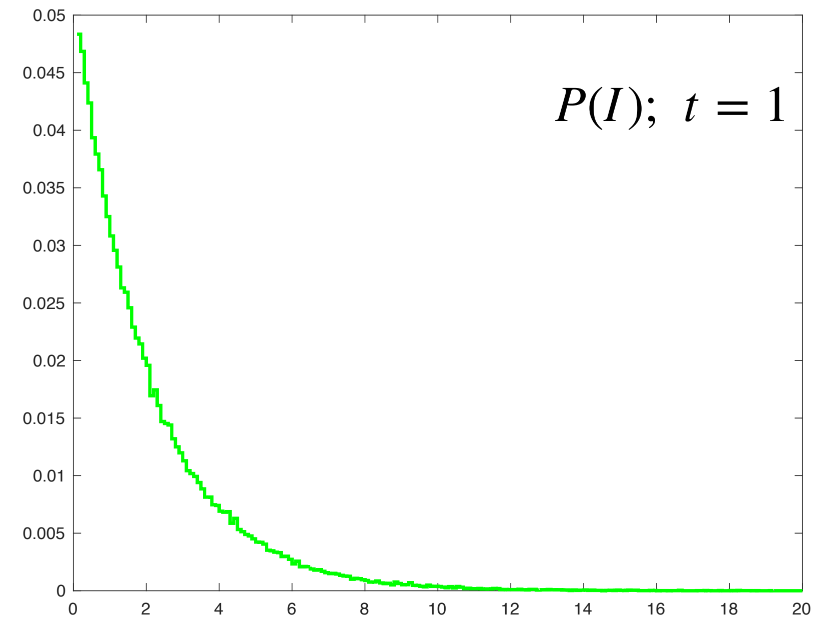
$$I = \langle E^{\dagger} E \rangle_t$$

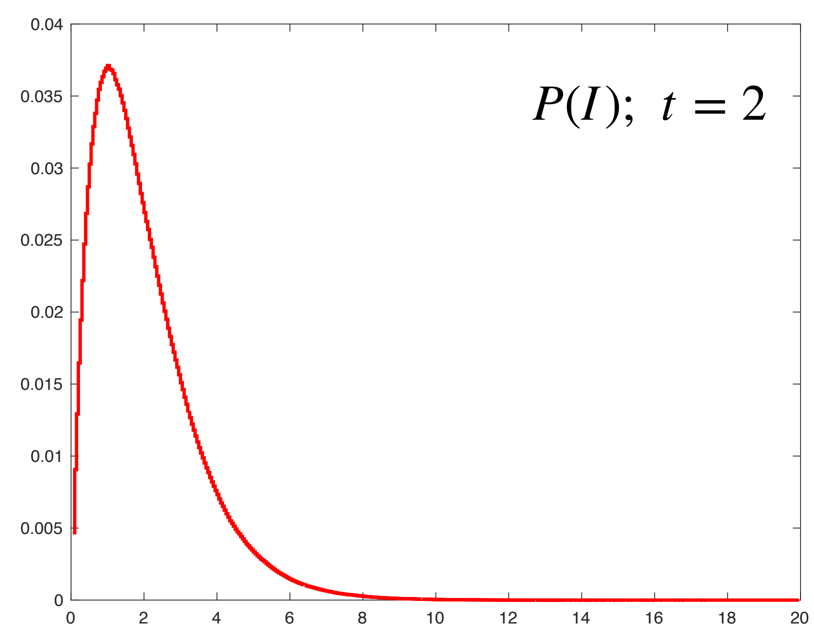


Average (or sum) is convolution of pdfs



*

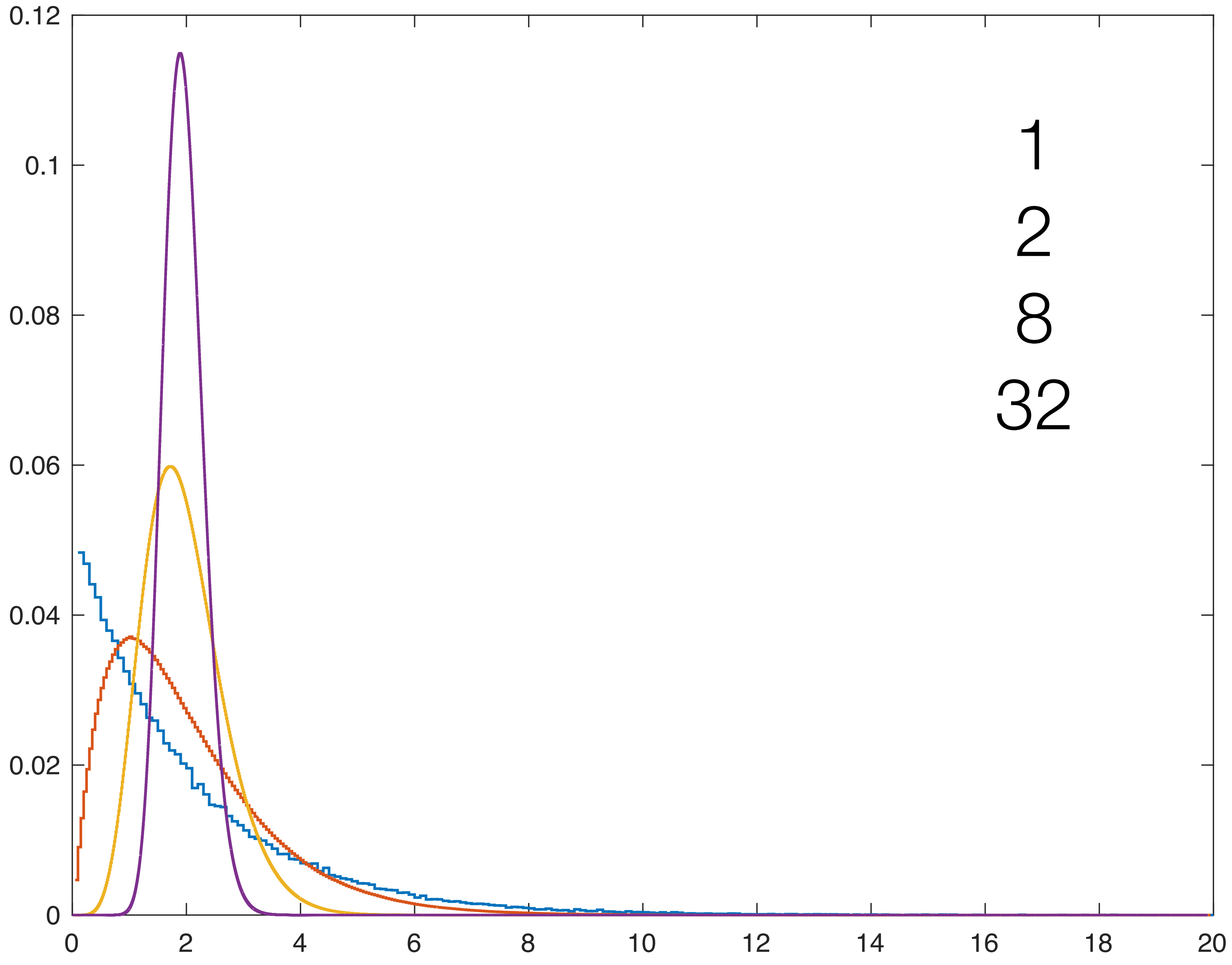




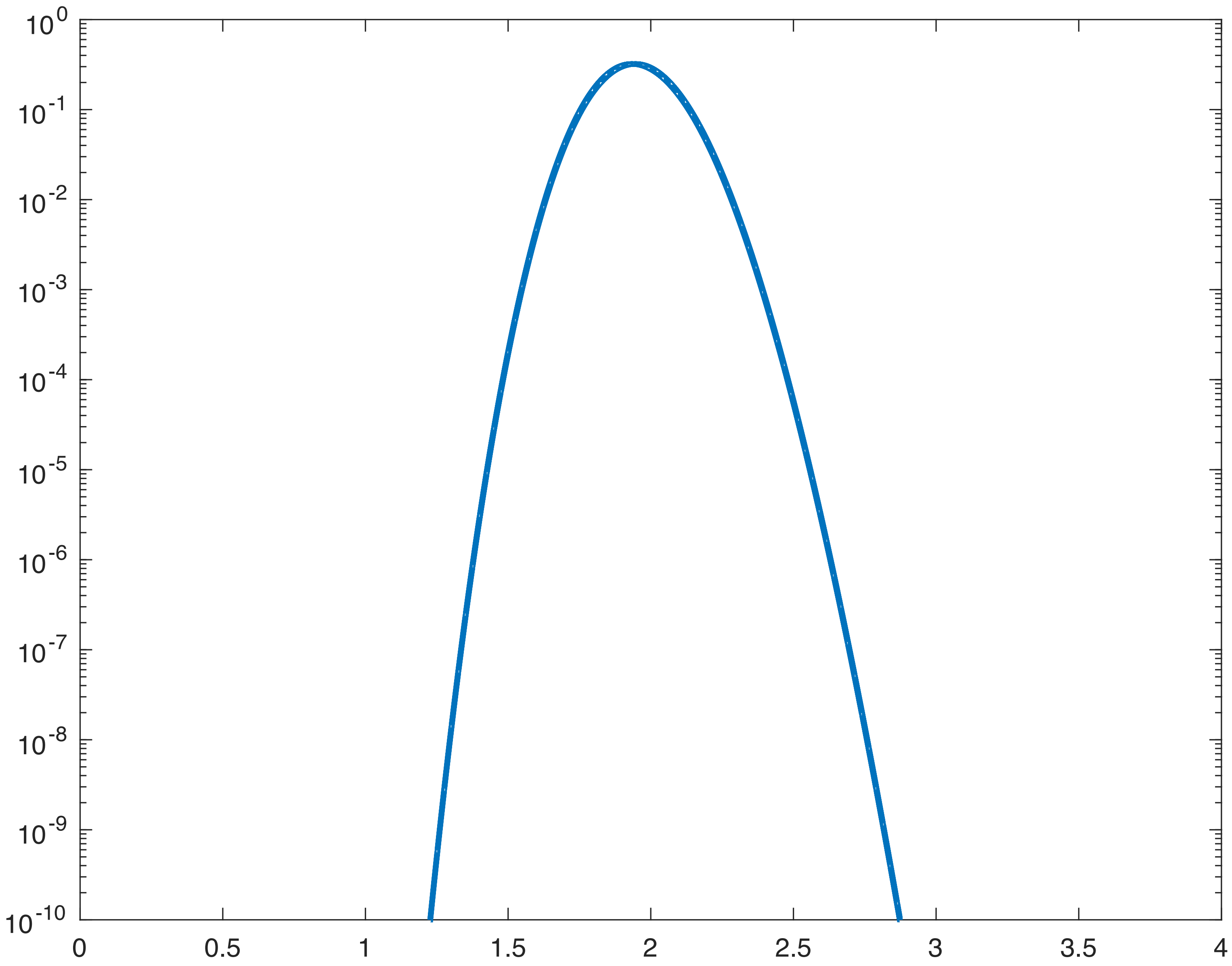
Probability distributions of averages

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

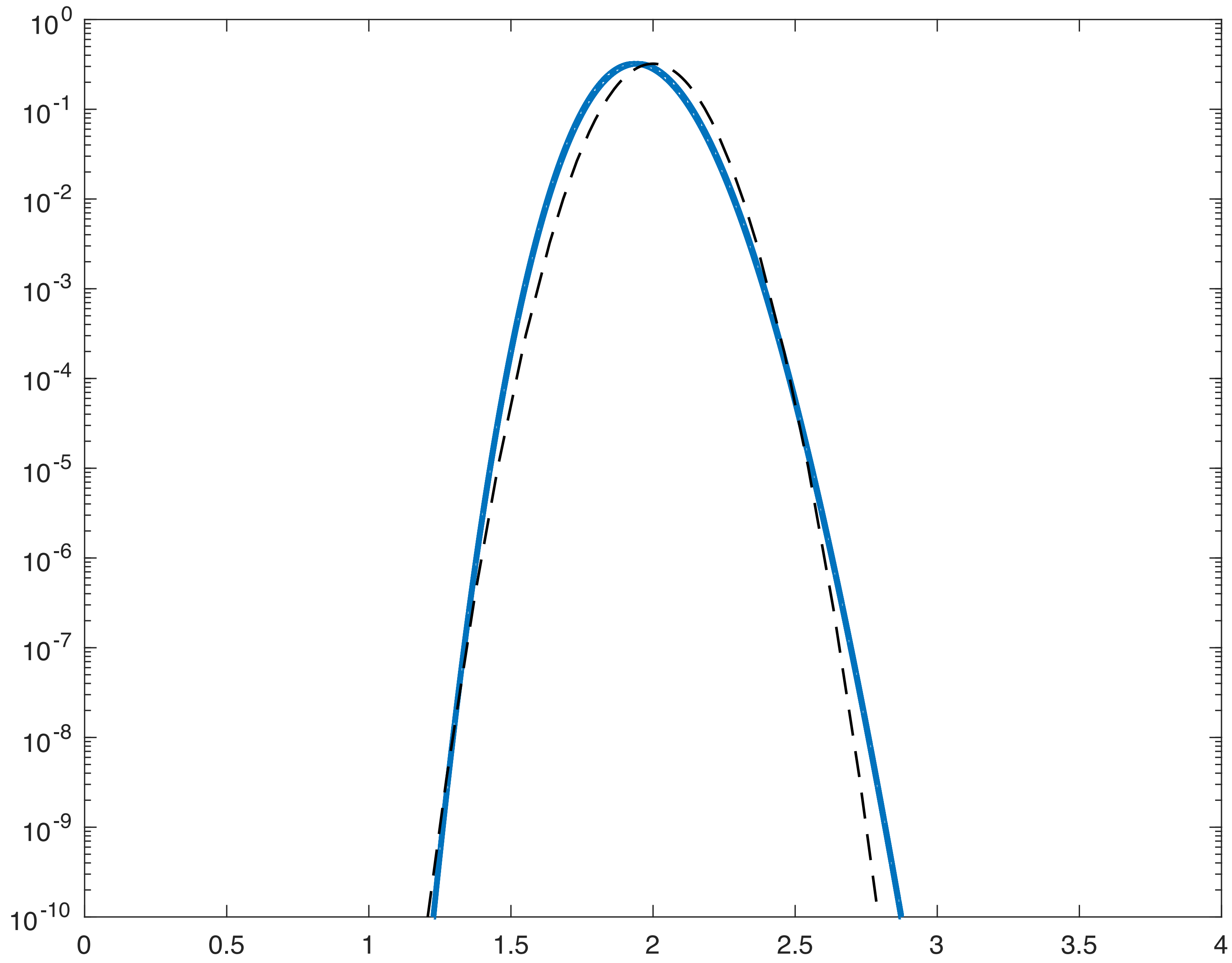
Repeated convolution



256 averages on semilogy

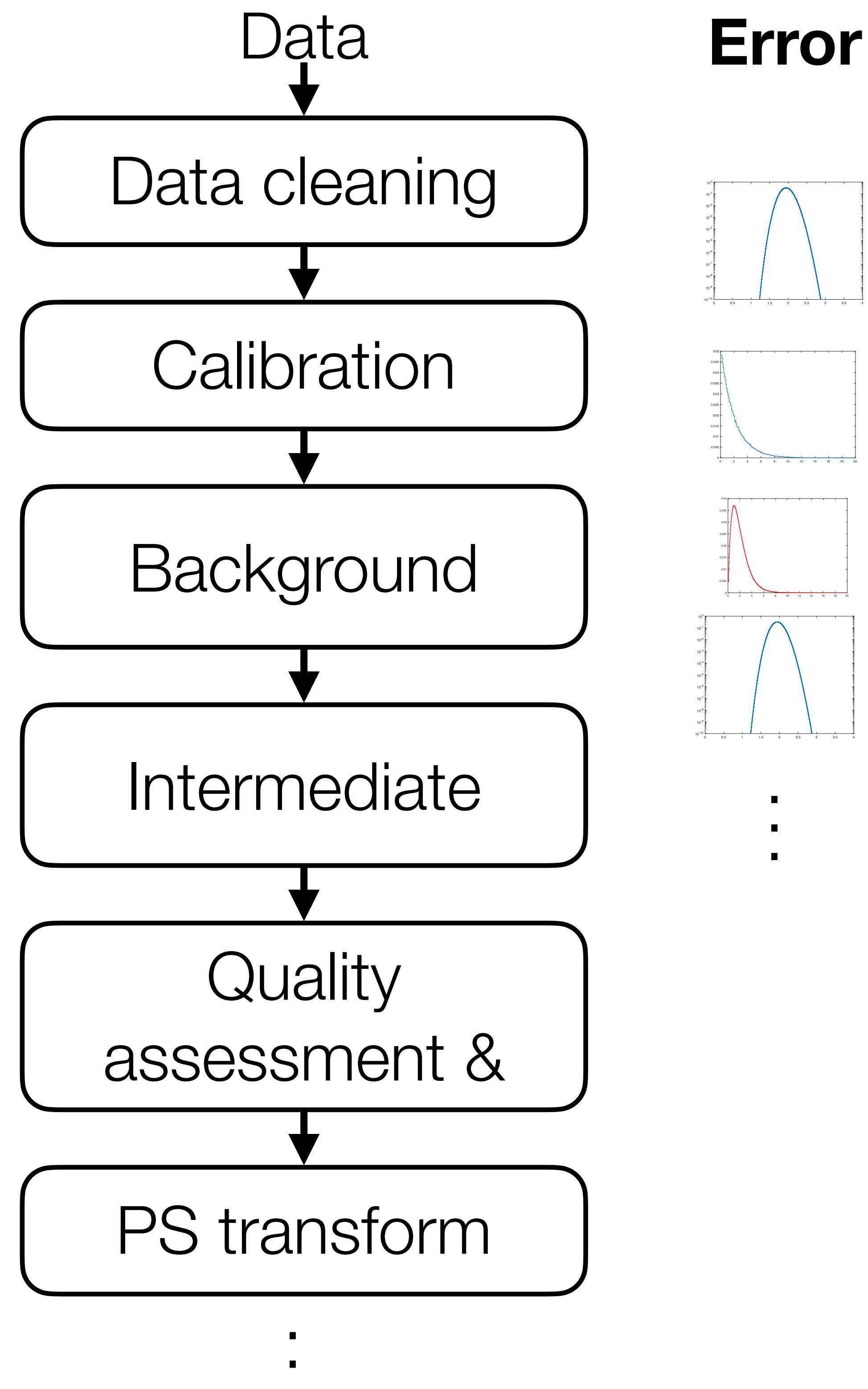


256 averages on semilogy

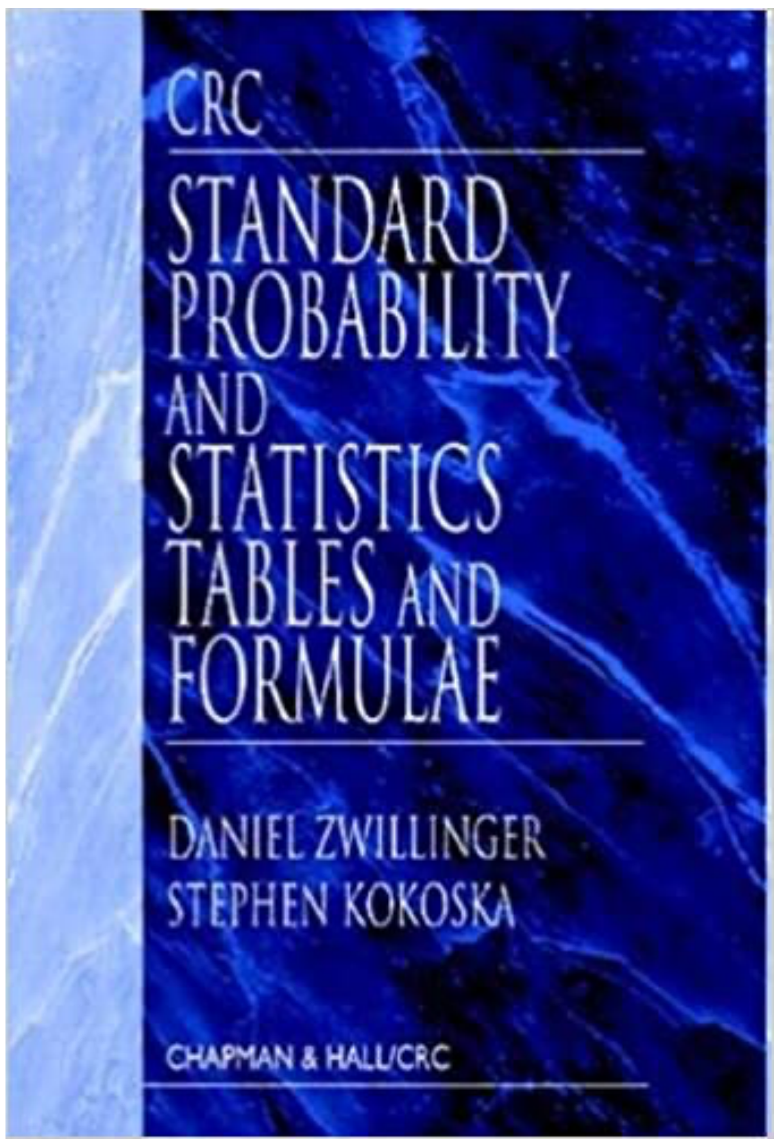


Mathematical relationships between distributions

- Mathematical operations change the noise distributions
- Any analysis step is math…



CRC Standard Probability and Statistics Tables and Formulae



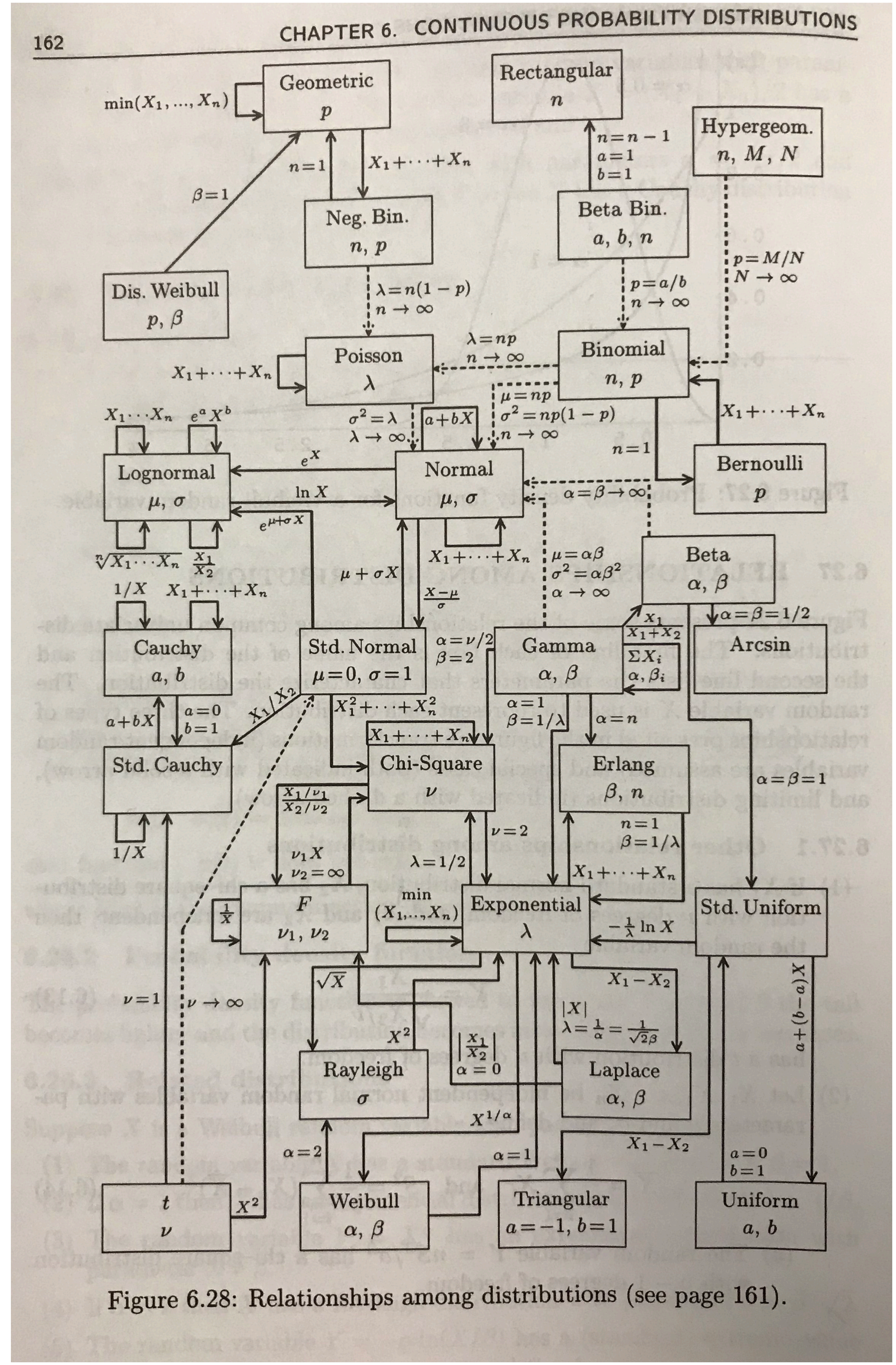
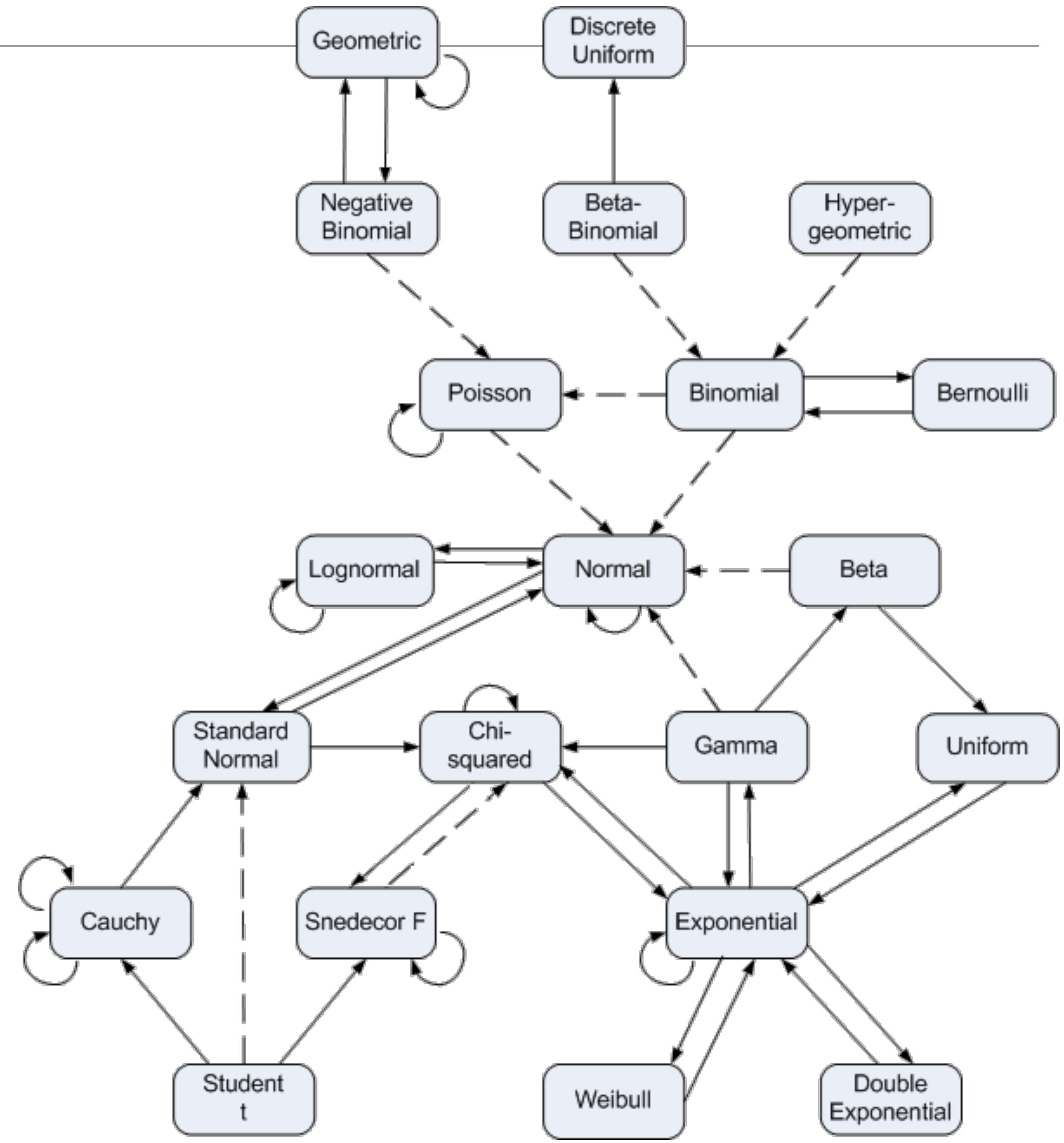

162 CHAPTER 6. CONTINUOUS PROBABILITY DISTRIBUTIONS

Figure 6.28: Relationships among distributions (see page 161).

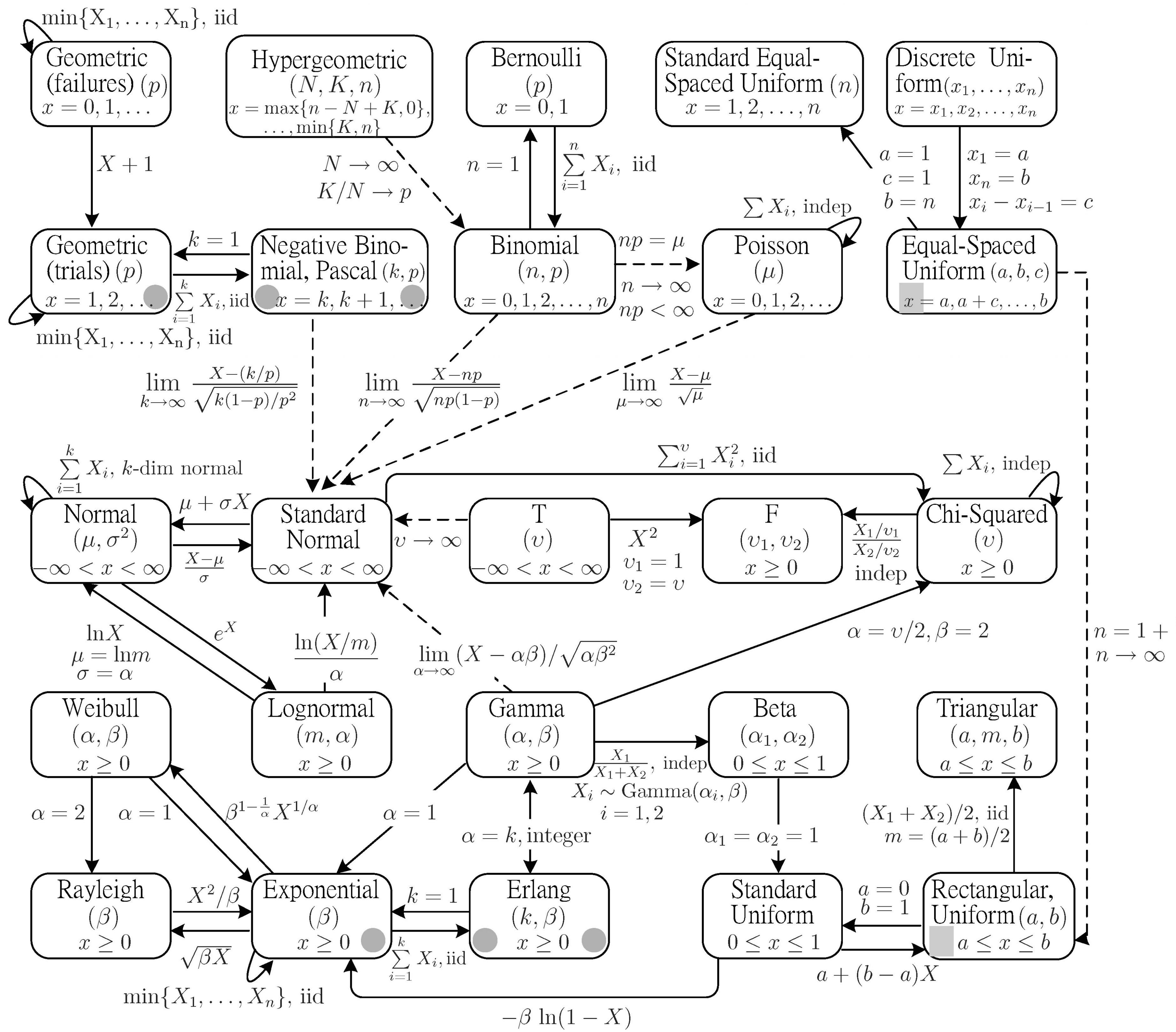
Clickable web site

- https://www.johndcook.com/blog/distribution_chart/#exponential_gamma

- Lives on web so might vanish, selection is OK



Song, IIE Transactions



Wikipedia

- Accurate
- Mathematicians have ruined some entries
- Look for ‘related distributions’

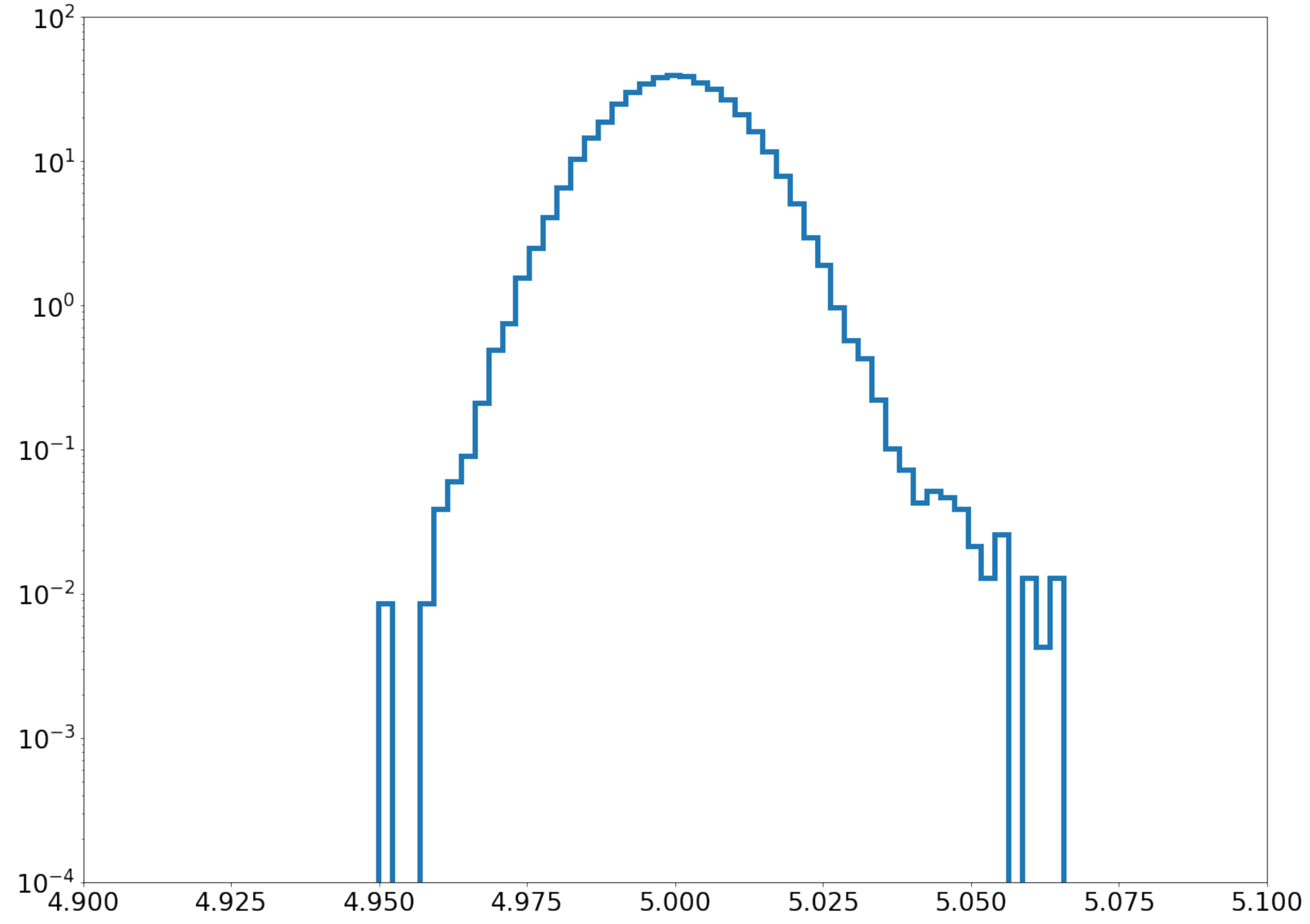
Example (ADMX inspired)

- Electric field is digitized E_t
  - Noise is Gaussian (real)
- Electric field is channelized into narrow frequency channels with a Fourier Transform $E_{f\ \Delta t}$
  - Noise is complex Gaussian
- Electric field channels are squared to create spectrum $|E_f|^2_{\Delta t}$
  - Noise is Exponential
- Spectrum is averaged $\langle |E_f|^2 \rangle_{\Delta t}$
  - Repeated convolution or Erlang distribution
  - Eventually becomes Gaussian through central limit theorem

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

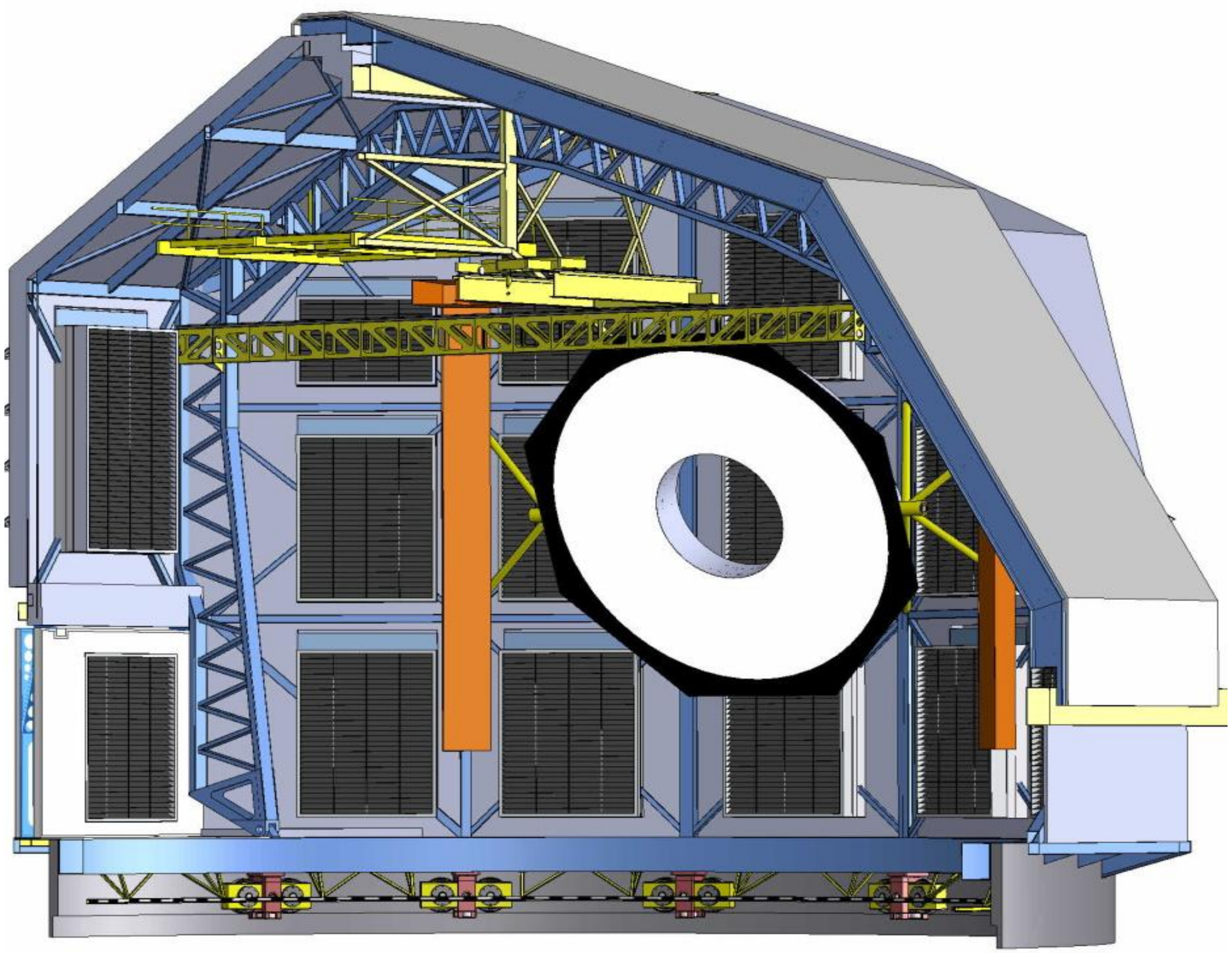


Figure 1: Calibration screen located inside the LSST dome

10 meter screen inside the dome

LSST screen specs:

- Illuminated with white light (UV to IR) and/or a frequency 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:

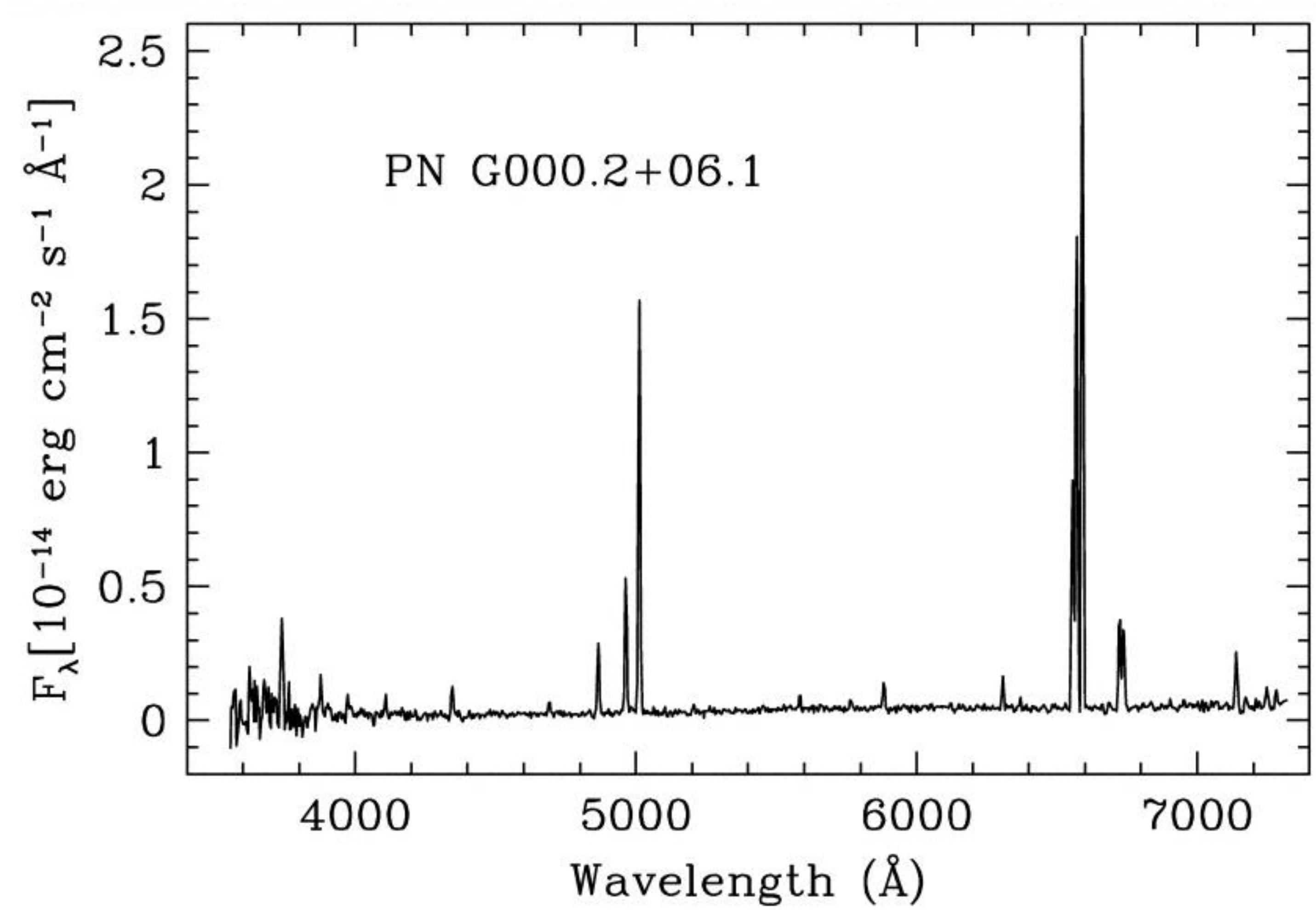
- Randomly shuffling the data so sources disappear
- Subtracting neighboring datasets
- Requires signal be slowly varying

Hope for isolated or rare signals

Isolated signals

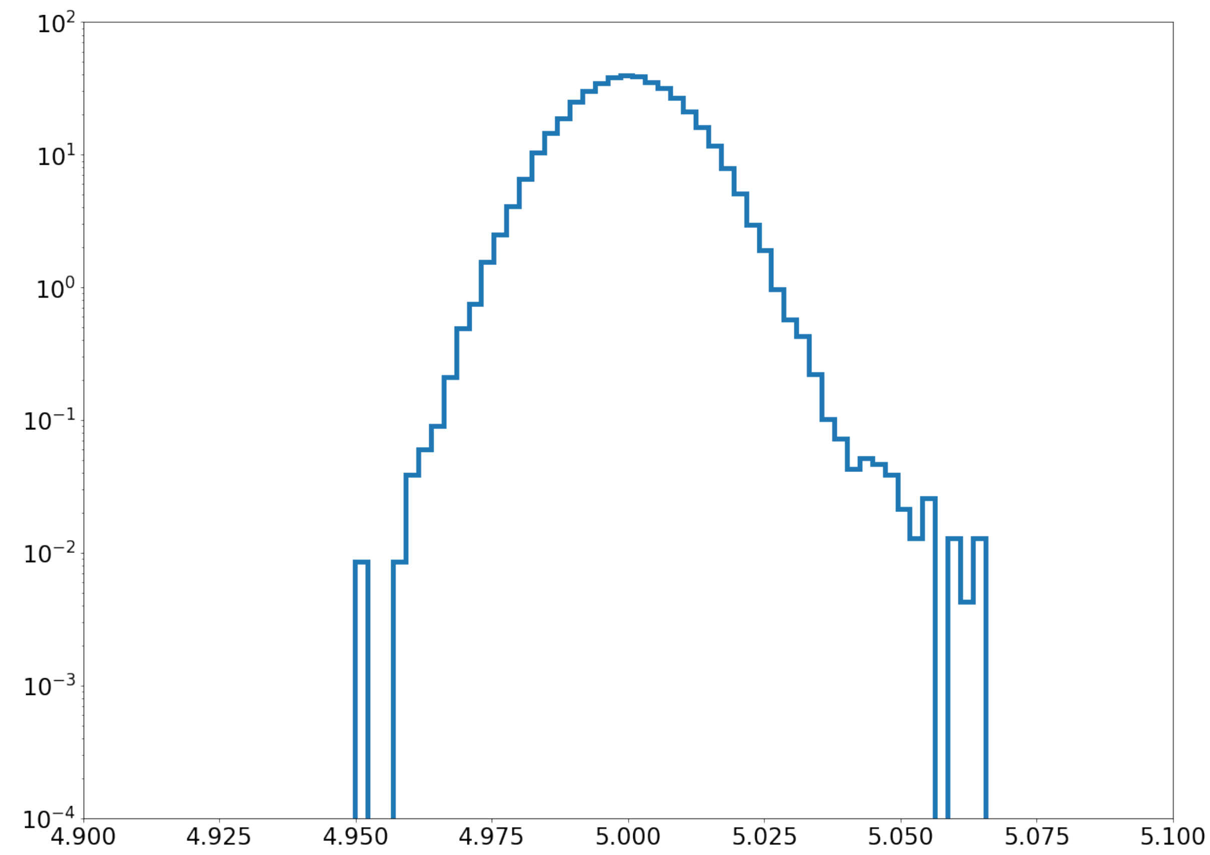
- Identify energies, times, or locations where you don't think there is a signal and extrapolate to the region of interest

- 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)

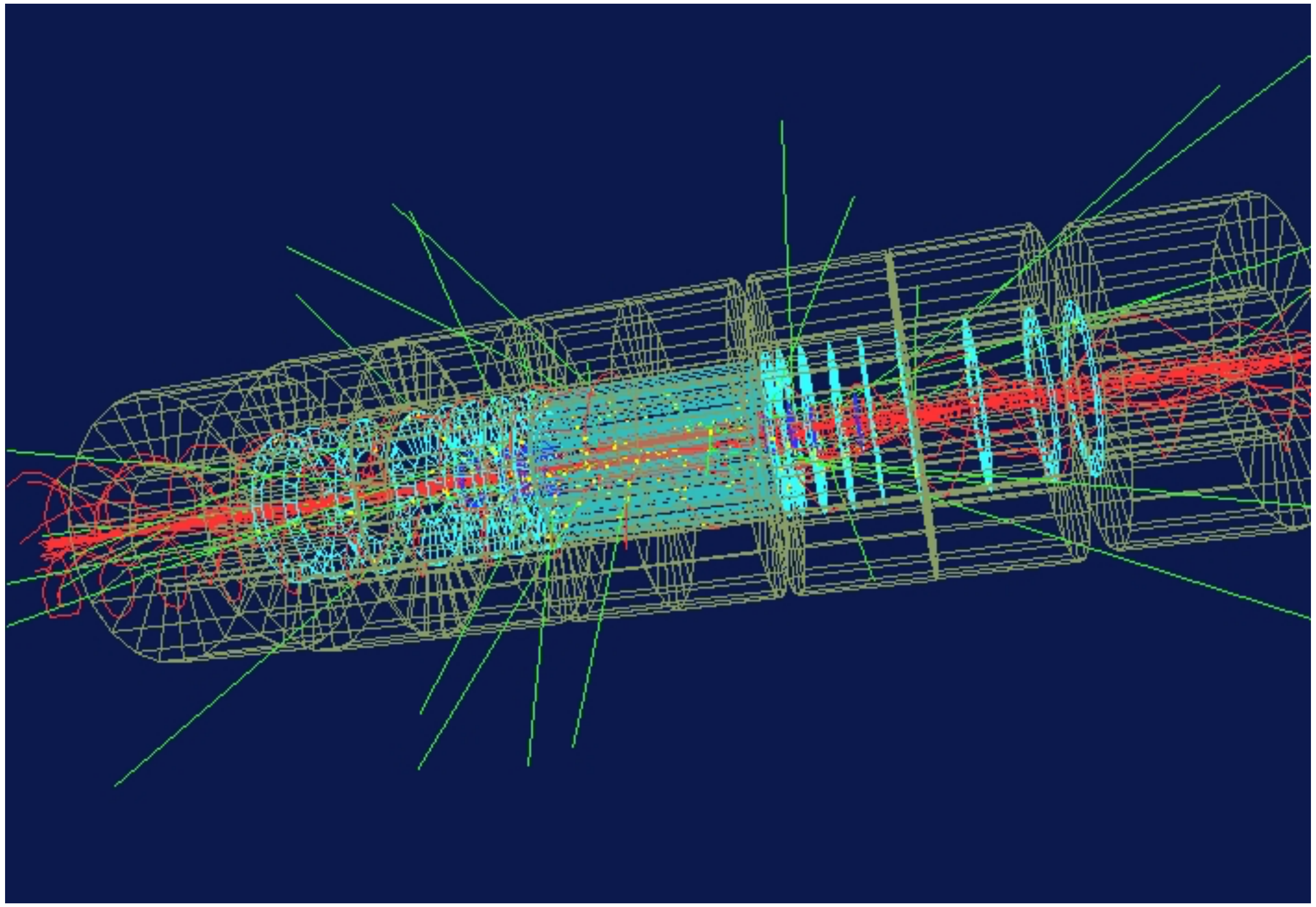
Make fake data without your signal

GEANT

- Purpose built particle physics simulator

- Entire instruments in simulation

- Conferences on how to make it more precise



Simulations fall in two categories

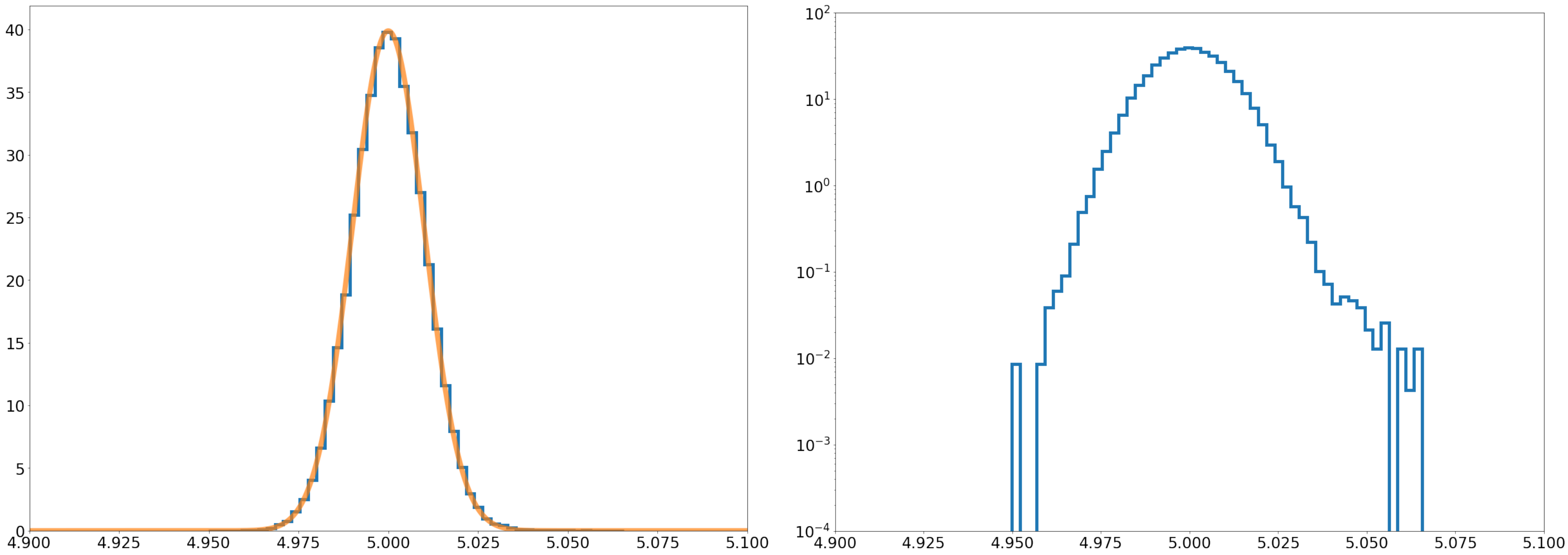
- Helping to understand how the instrument works
- Calculating fake background data (must be much more precise)

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

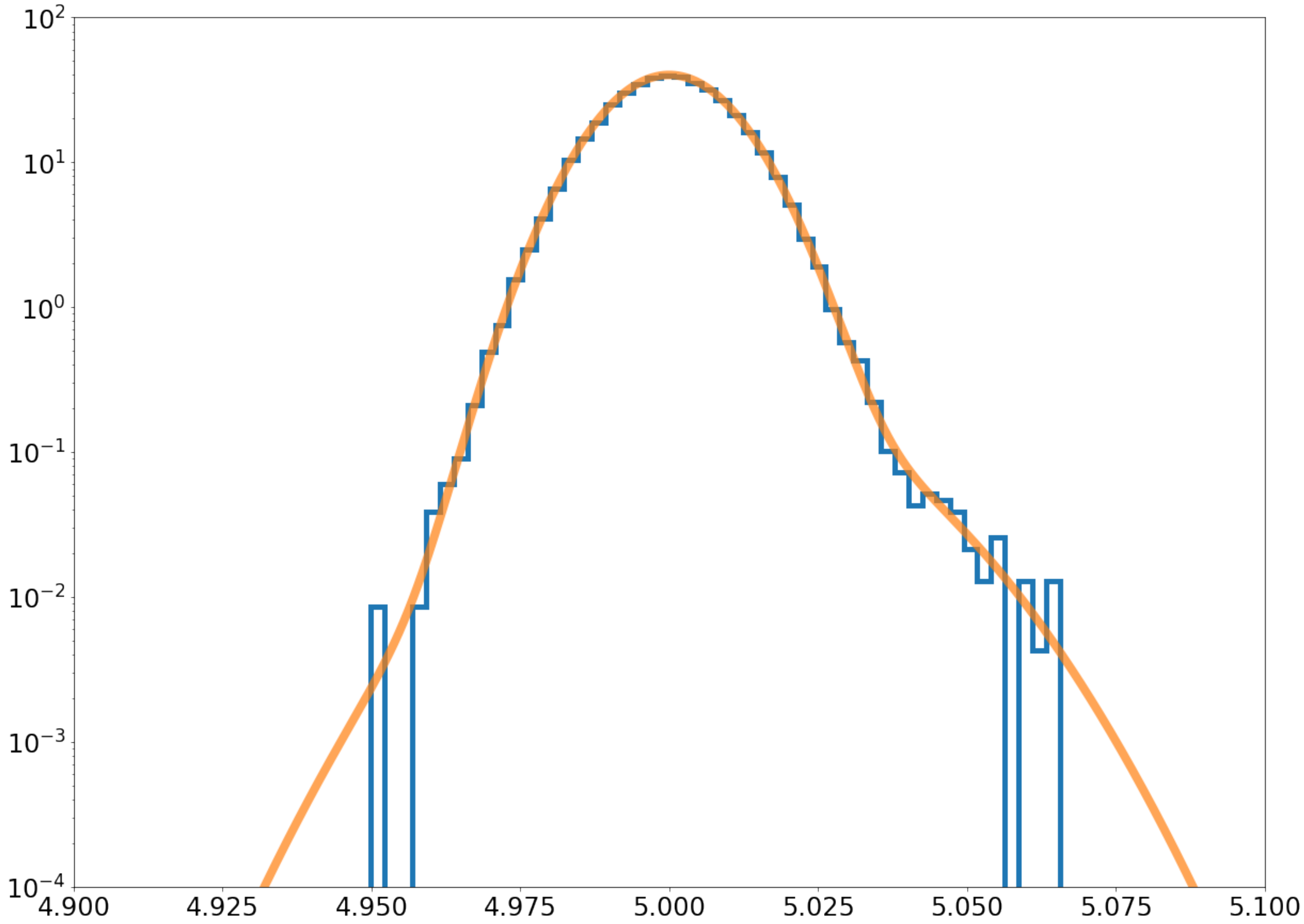


Always use a semilog plot!

3) Think about what kinds of pdfs you might expect

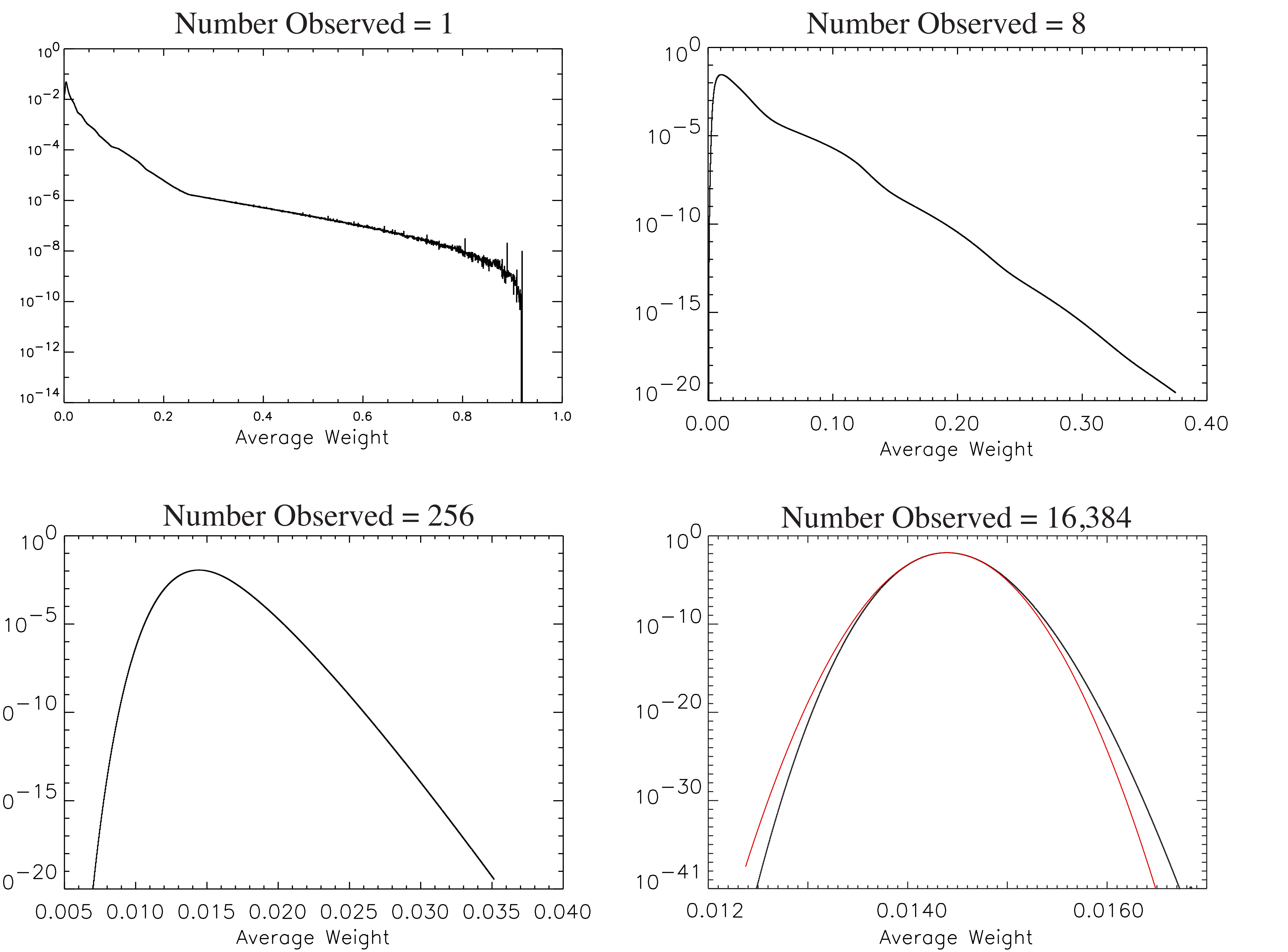
- What creates the background?
  - Thermal noise, radioactive backgrounds, cosmic rays, other uninteresting sources, …
- What systematics should I worry about?
  - Is background stable in time, detector, energy, …

4) Use to build a model of the background



The better the background model, the better the science

My thesis


Number Observed = 1
Average Weight
Number Observed = 8
Average Weight
Number Observed = 256
Average Weight
Number Observed = 16,384
Average Weight