Class 5: Data dense plots; trials & more

Miguel F. Morales
Bryna Hazelton

Data density examples

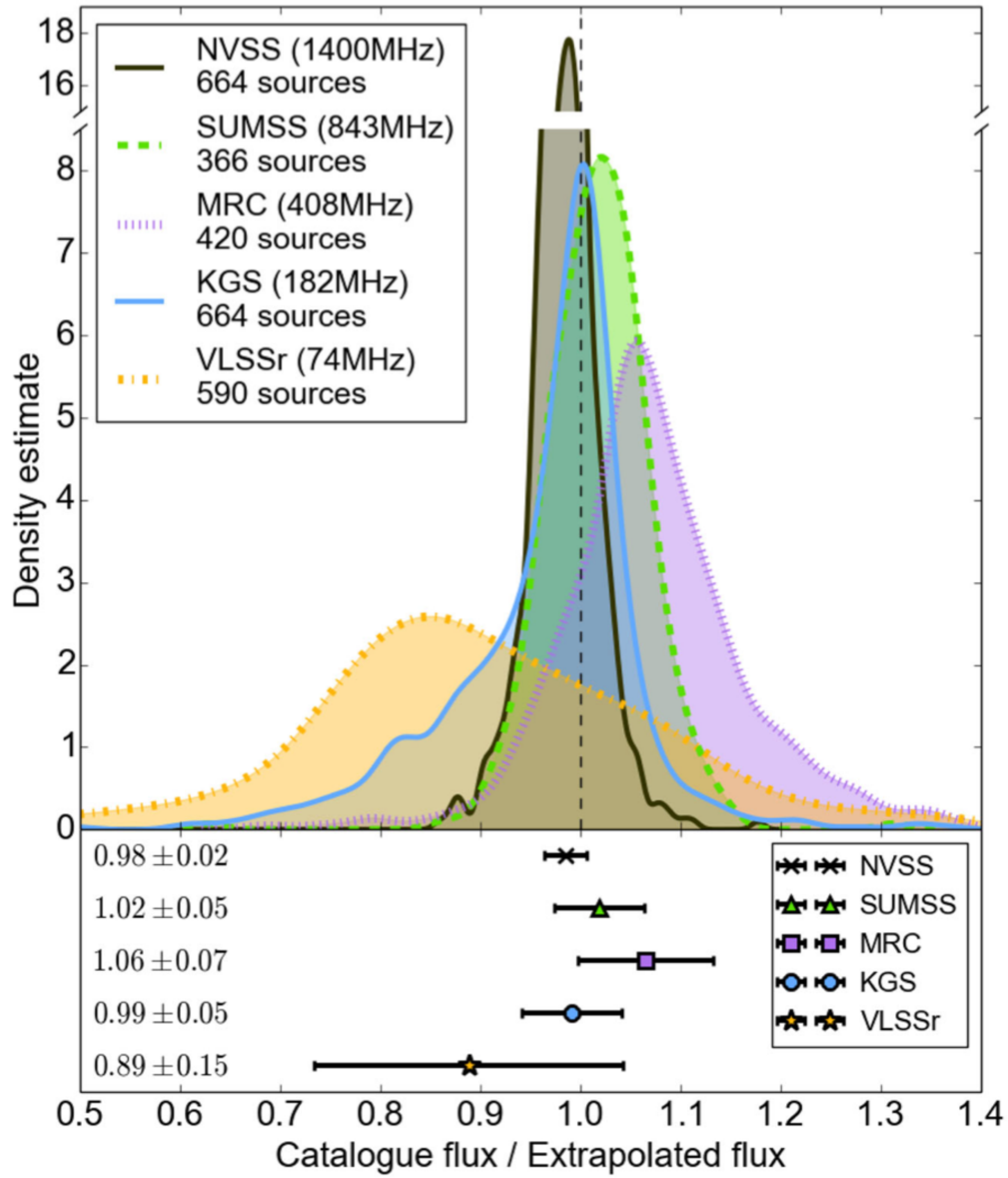
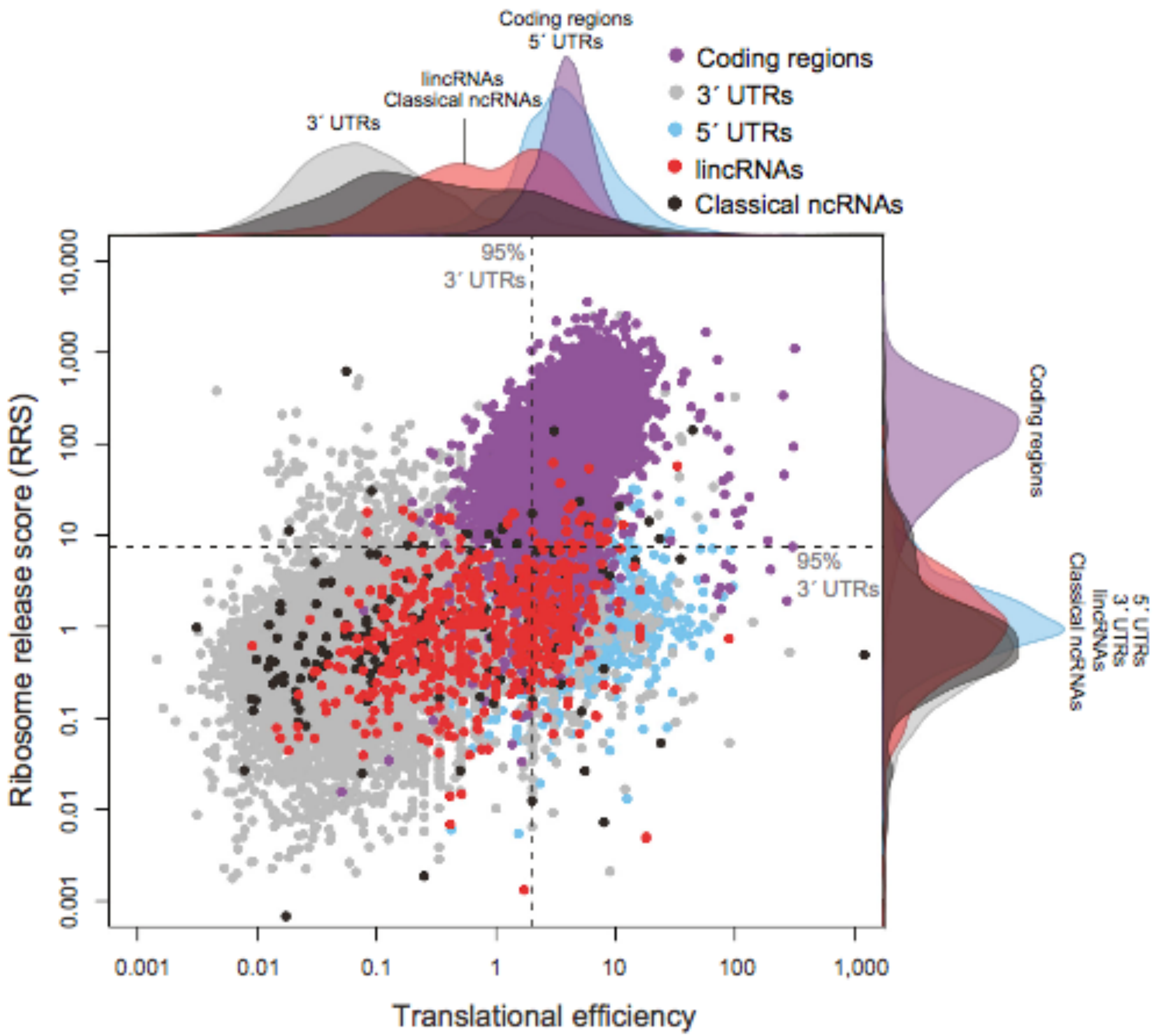
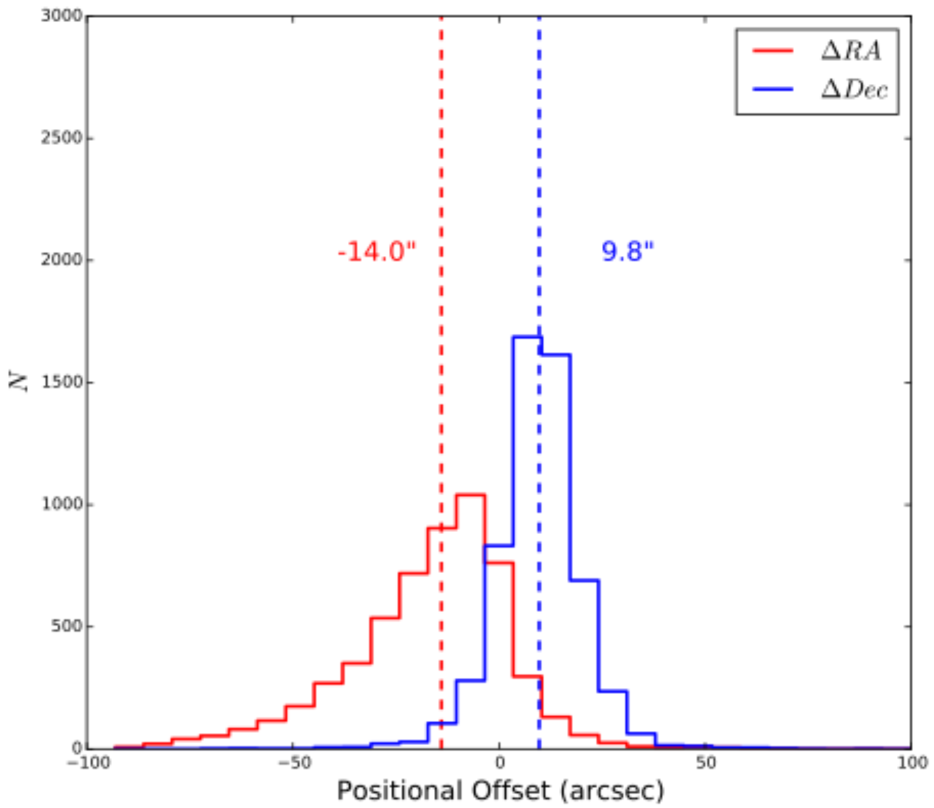


Figure 2. The ratio between observed flux density and extrapolated flux density from a fit to the SED is shown for every time a catalogue appeared in a match with at least two other catalogues for `isolated` sources. The upper panel shows a univariate kernal density estimation of each distribution (note broken y axis due to the sharp peak in the NVSS ratio distribution), while the lower panel shows the median and median absolute deviation of each distribution. The KGS spectral index agrees very well with no indication of flux bias on average.

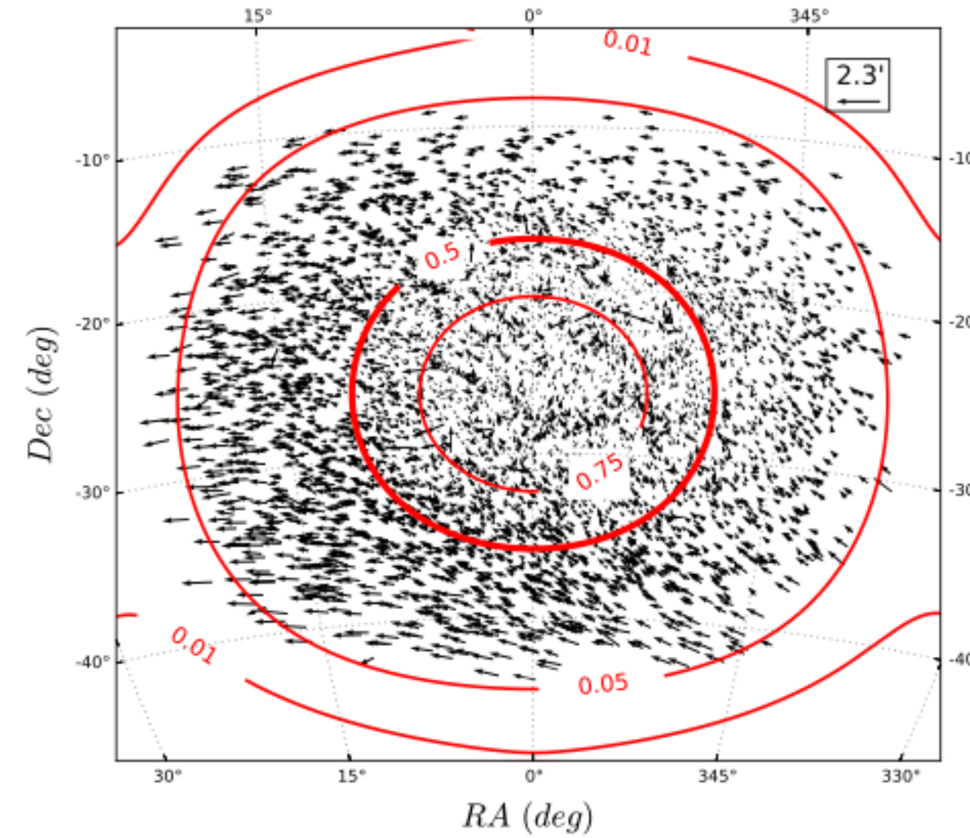


Coding regions
5′ UTRs
lincRNAs
Classical ncRNAs
3′ UTRs
Coding regions
3′ UTRs
5′ UTRs
lincRNAs
Classical ncRNAs
95%
3′ UTRs
95%
3′ UTRs
Ribosome release score (RRS)
10,000
1,000
100
10
1
0.1
0.01
0.001
0.001
0.01
0.1
1
10
100
1,000
Translational efficiency
Coding regions
5′ UTRs
3′ UTRs
lincRNAs
Classical ncRNAs



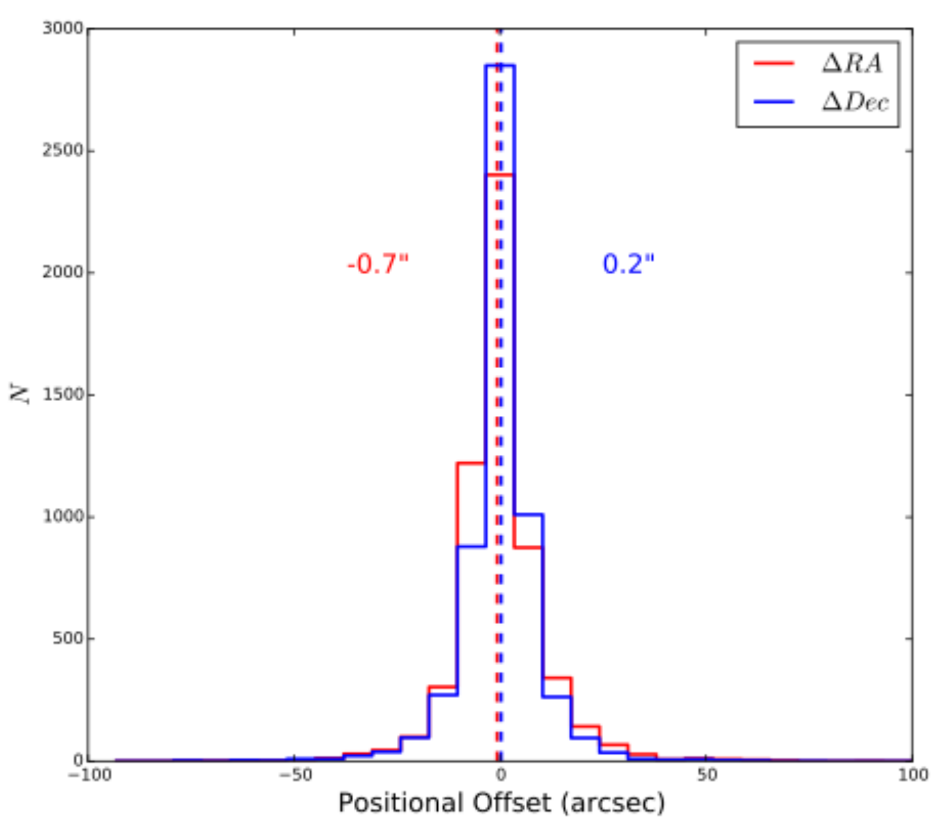
ΔRA
ΔDec
-14.0"
9.8"
N
Positional Offset (arcsec)

(a)



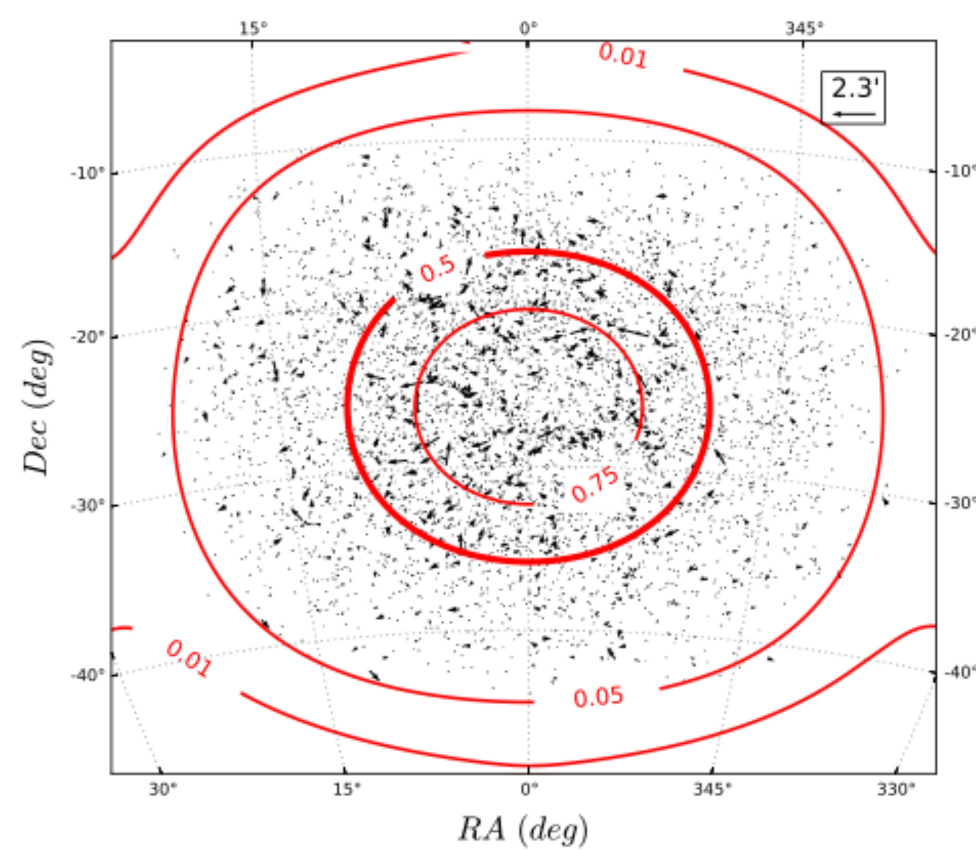
2.3'
0.01
0.5
0.75
0.05
0.01
Dec (deg)
RA (deg)

(b)



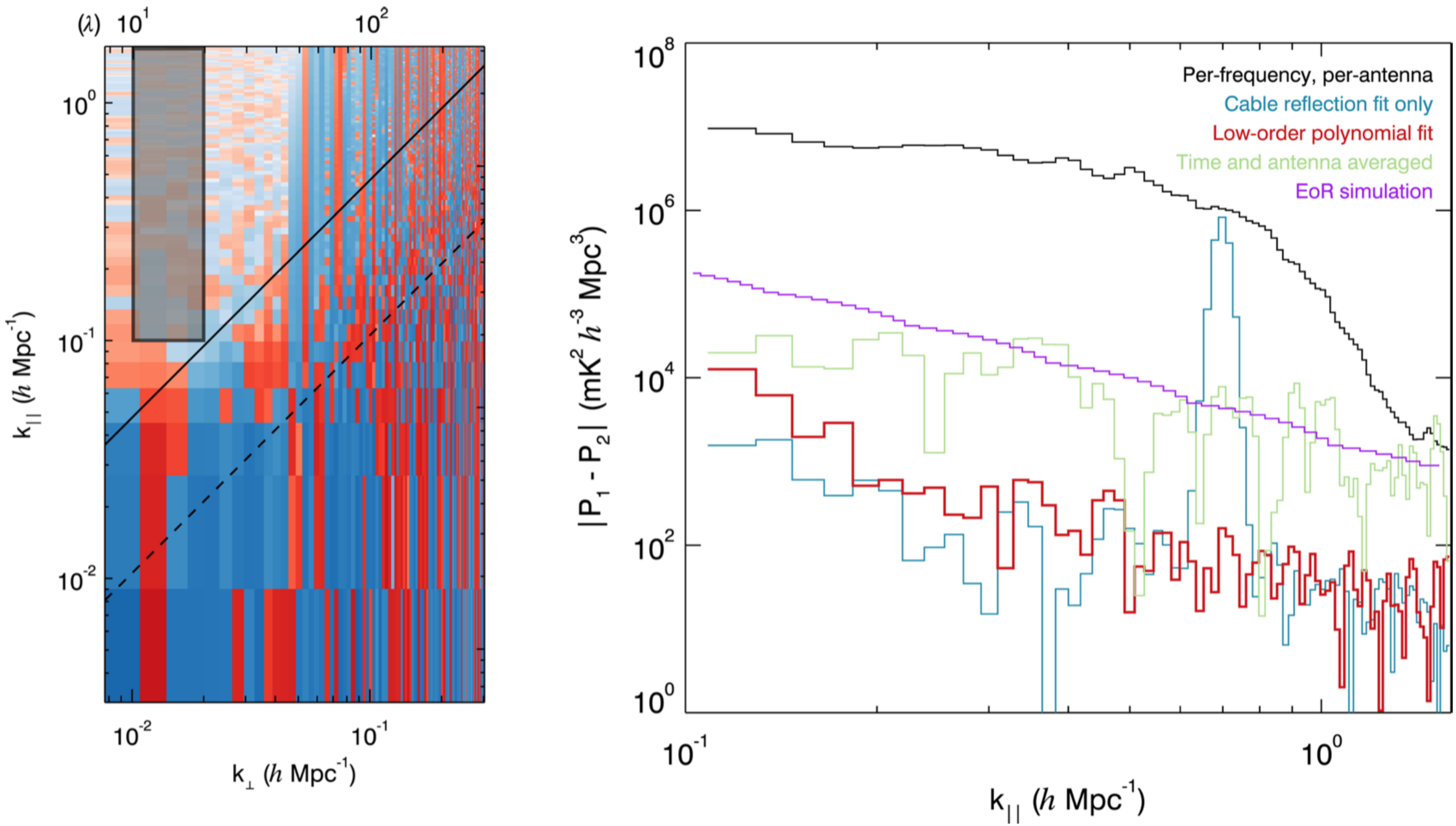
ΔRA
ΔDec
-0.7"
0.2"
N
Positional Offset (arcsec)

(c)



2.3'
0.01
0.5
0.75
0.05
0.01
Dec (deg)
RA (deg)

(d)


(λ) 10^1 10^2
10^0
10^-1
10^-2
k_|| (h Mpc^-1)
10^-2 10^-1
k_⊥ (h Mpc^-1)
10^8
10^6
10^4
10^2
10^0
|P_1 - P_2| (mK^2 h^-3 Mpc^3)
10^-1 10^0
k_|| (h Mpc^-1)
Per-frequency, per-antenna
Cable reflection fit only
Low-order polynomial fit
Time and antenna averaged
EoR simulation

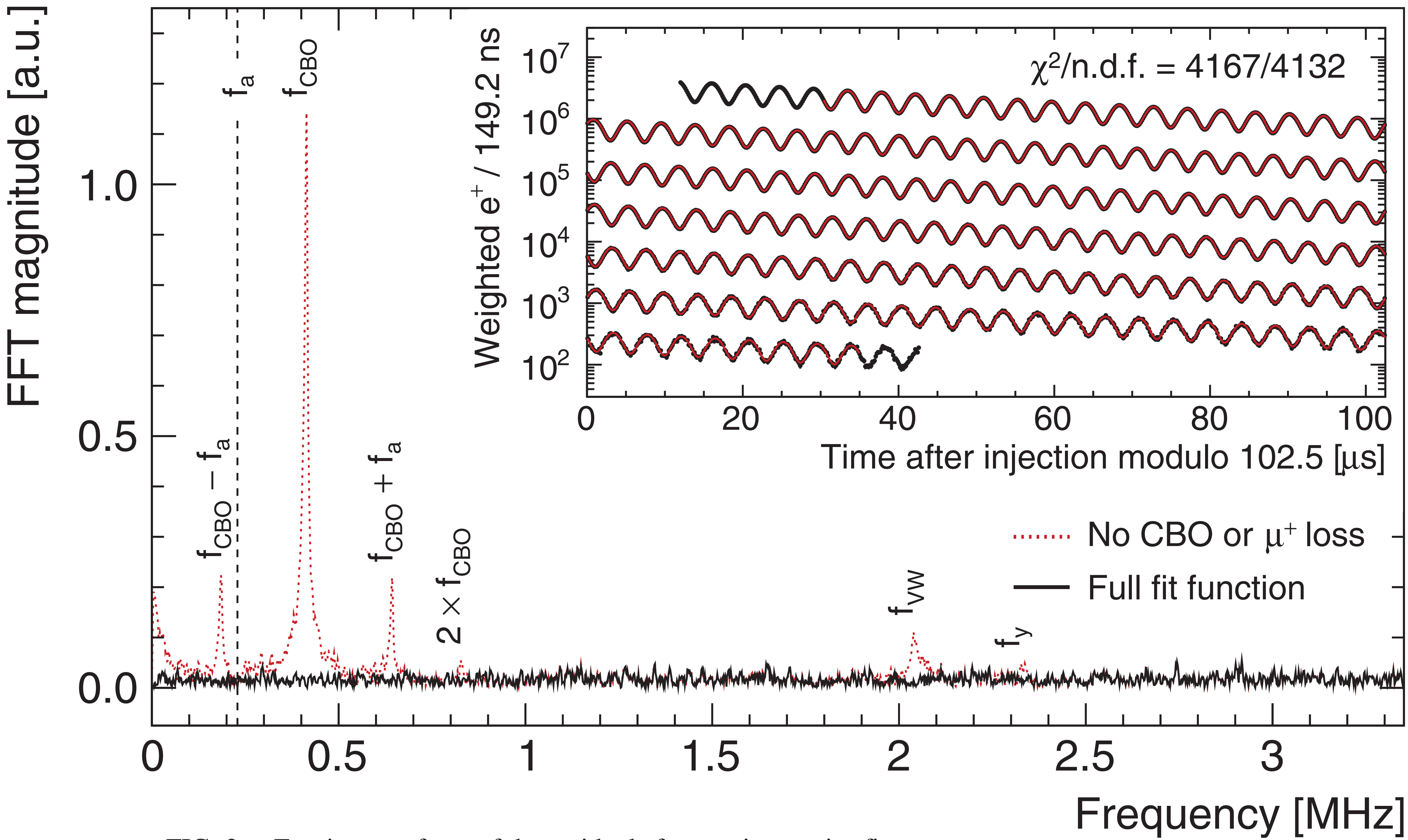
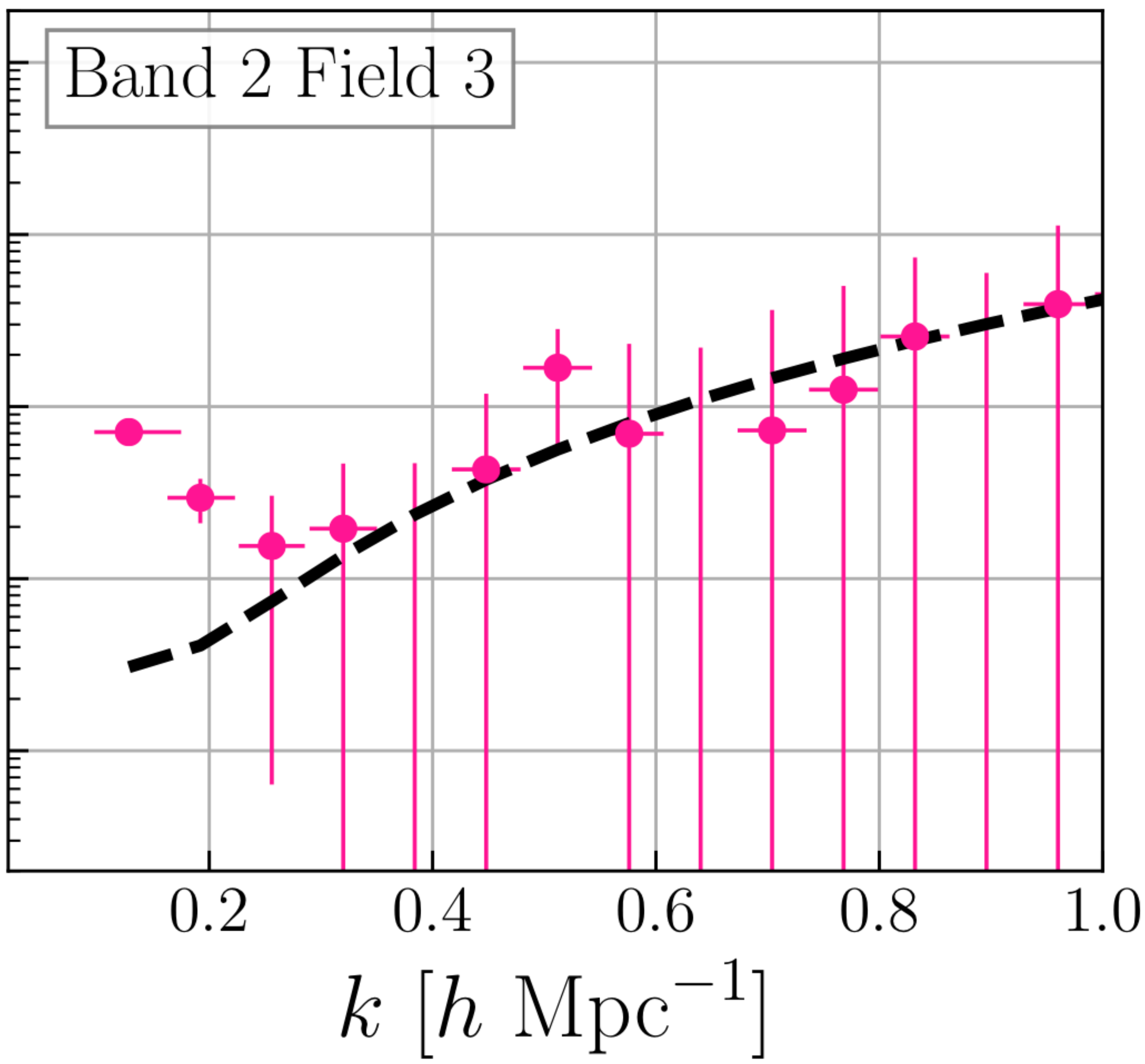


FIG. 2. Fourier transform of the residuals from a time-series fit following Eq. (5) but neglecting betatron motion and muon loss (red dashed), and from the full fit (black). The peaks correspond to the neglected betatron frequencies and muon loss. Inset: asymmetry-weighted e^+ time spectrum (black) from the Run-1c run group fit with the full fit function (red) overlaid.

Emphasis critiques

Upper limit?



Upper limit

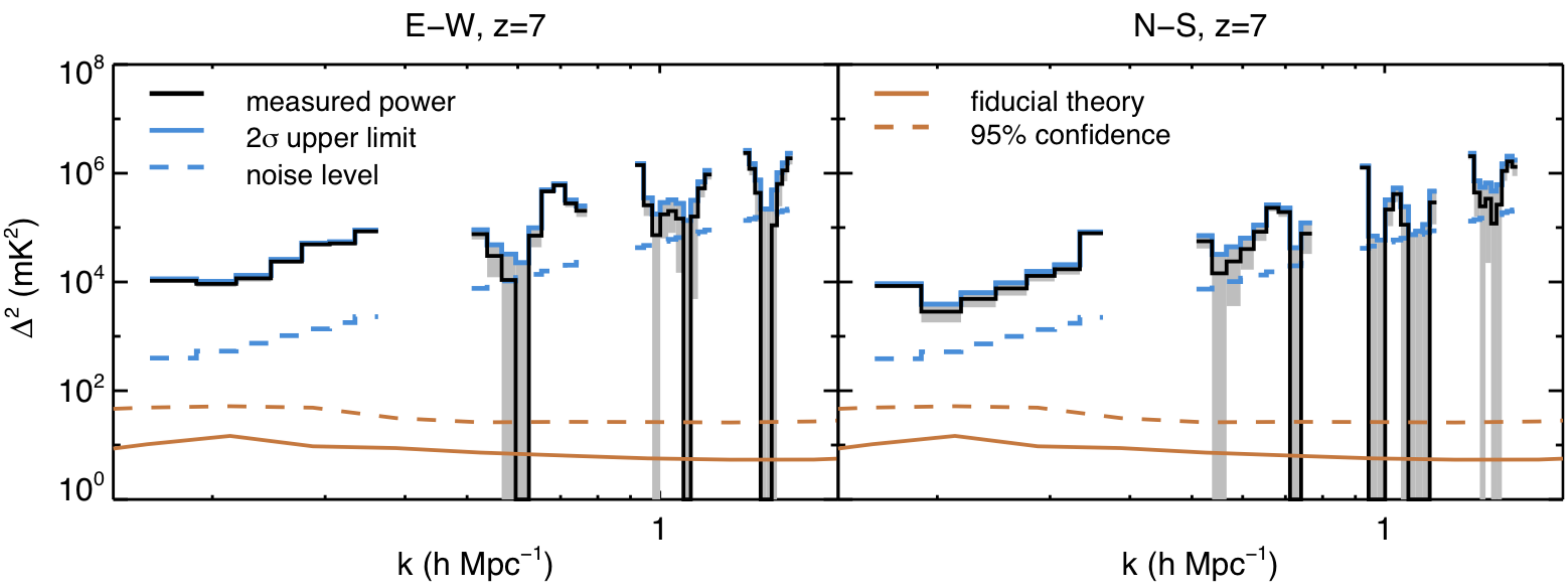
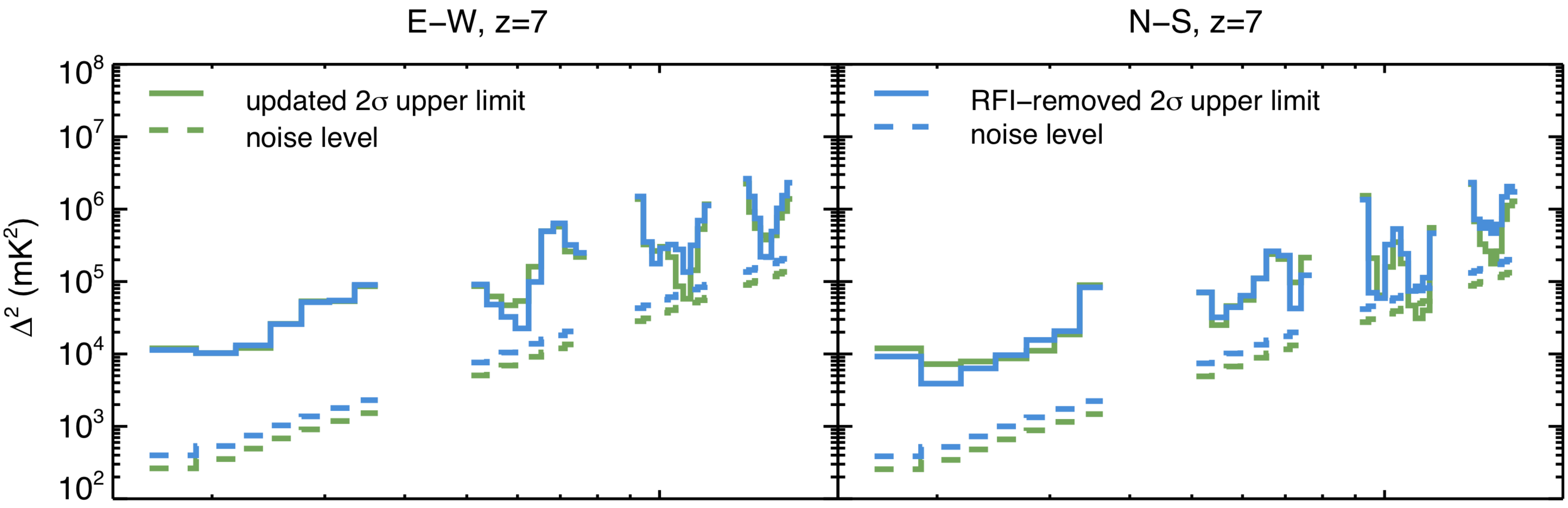


Figure 6. The 1D measured power spectra (black), the 2σ error bars (grey), the 2σ EoR upper limits (solid blue) and the 1σ thermal noise levels (dashed blue) for the E–W and N–S polarizations using 678 observations selected with SSINS. We also present an example fiducial EoR theory power spectrum (solid brown) along with the theoretical 2σ upper limits on the 21 cm power spectrum amplitude (brown dashed) obtained using existing observational constraints (see Appendix A for further details).

These constitute our best EoR upper limits in this work. We are noise dominated for many k-modes, including our lowest EoR upper limit.

Upper limit



E–W, z=7
N–S, z=7
updated 2σ upper limit
noise level
RFI–removed 2σ upper limit
noise level
Δ^2 (mK^2)
10^8
10^7
10^6
10^5
10^4
10^3
10^2

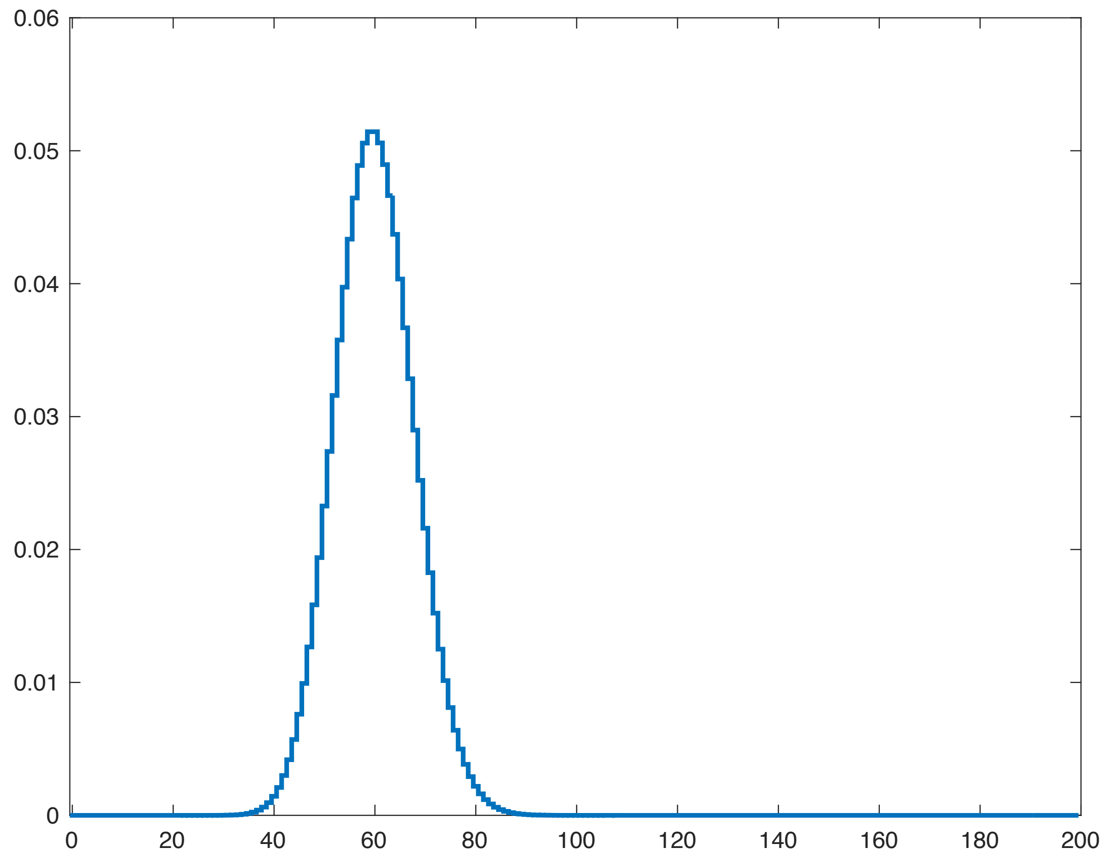
Thursday & HW

- Pick a plot of your data you’d like to improve
- We’ll workshop on Thursday
- HW pt 1 will be to make better version(s) of plot

HW questions

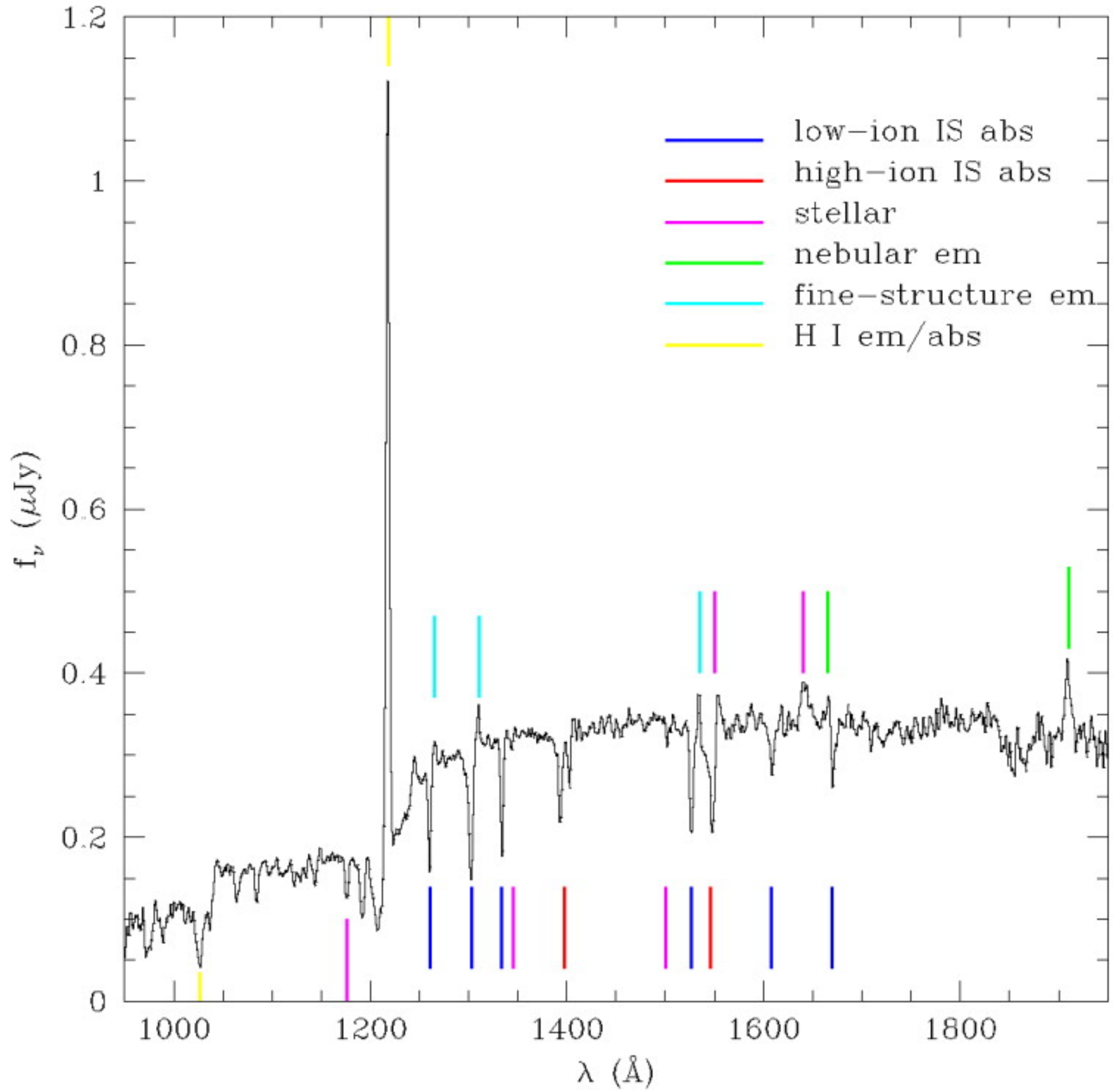
Average of pdf

why is `mean(pdf(x))` wrong?

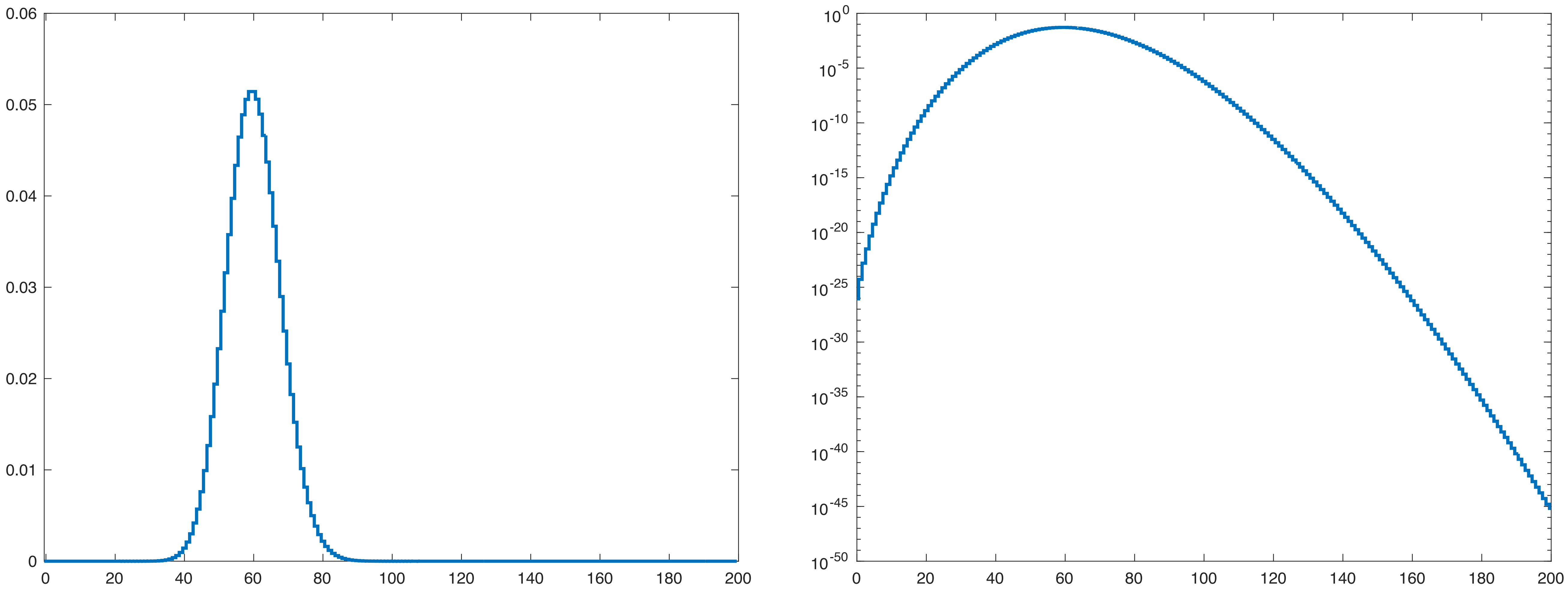


$$\mu = \int x \, \mathrm{pdf}(x) \; dx$$

What is the statistical question?



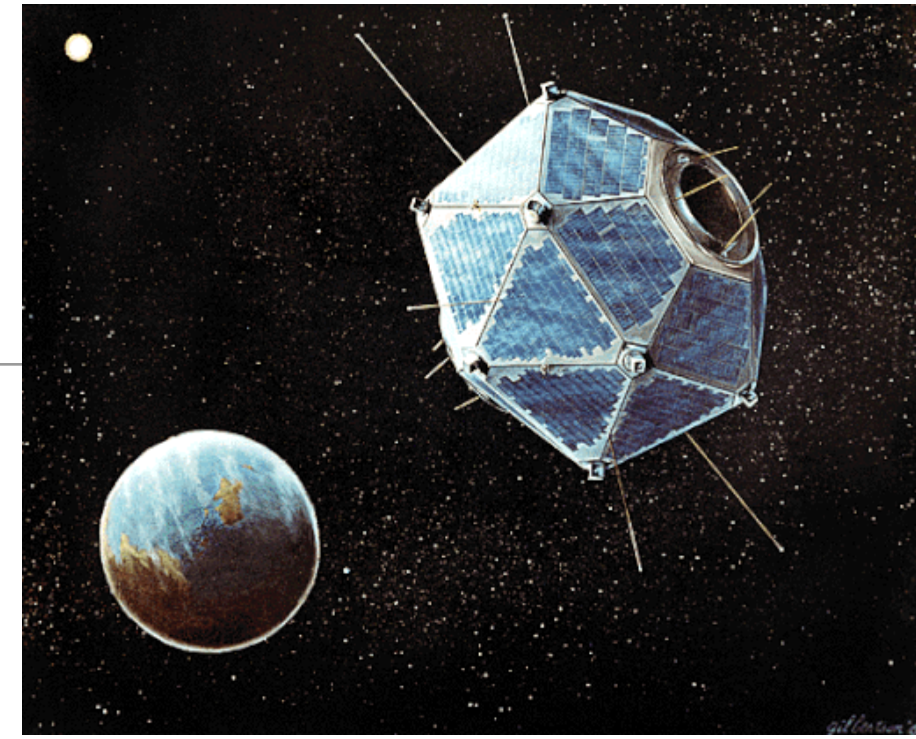
What is the statistical question?



Trials factor

Look elsewhere effect

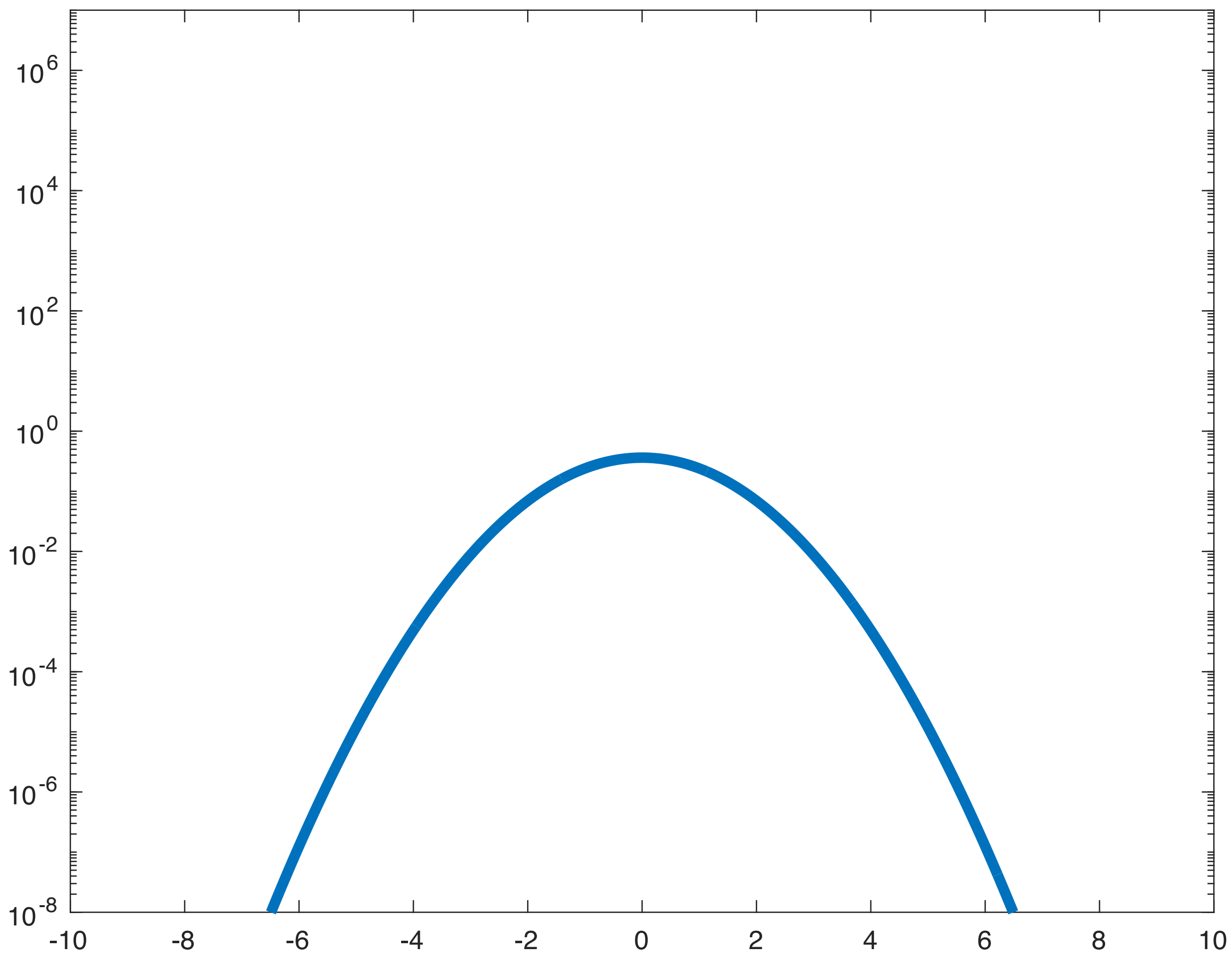
Trials

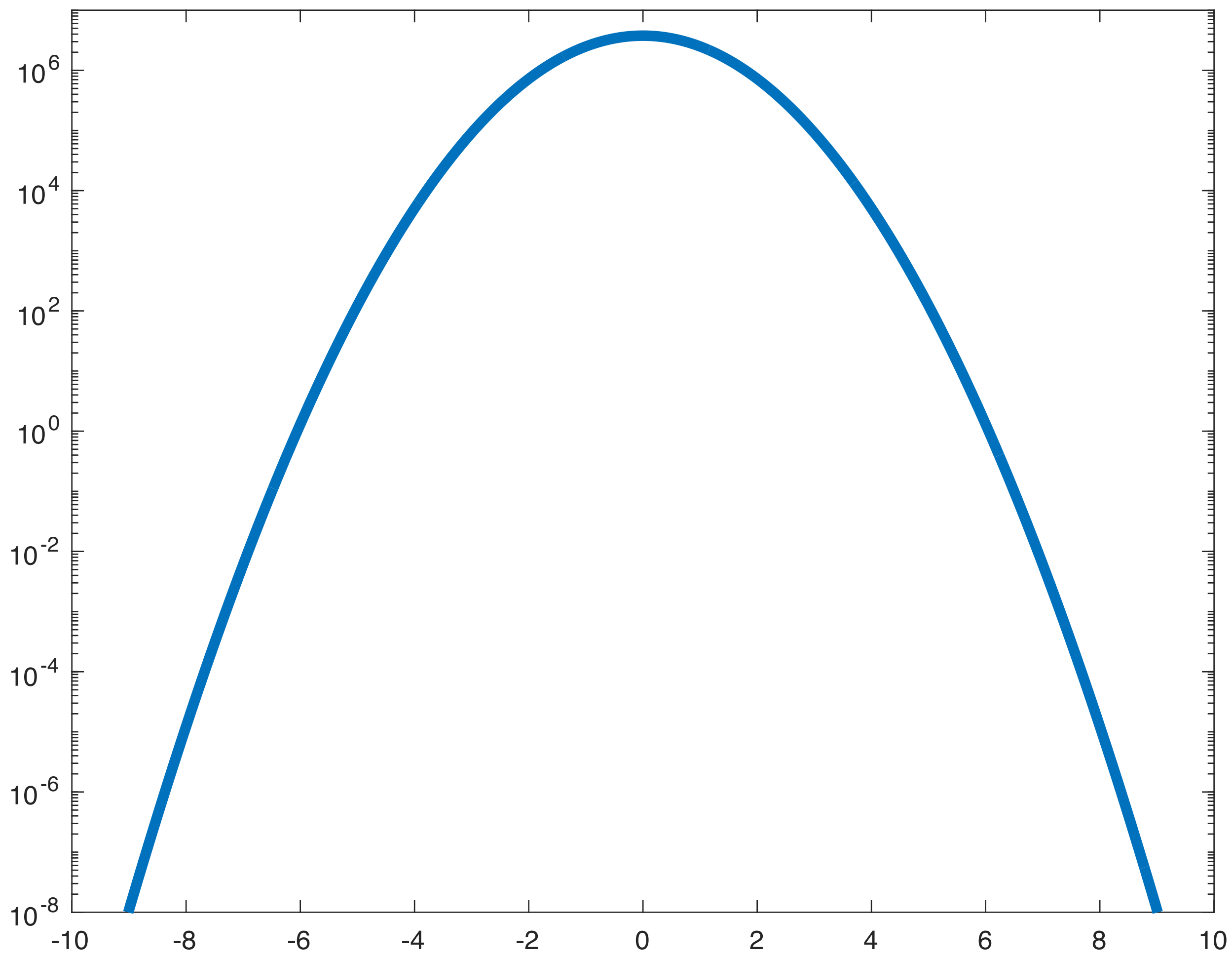


The Vela satellites were launched in the mid 1960’s to look for atmospheric nuclear explosions. Assume the internal gamma-ray detector had a background of 0.85 events per second. After scanning through 120 days of data it has looked at 10.4 million 1 second intervals. How many ‘5 sigma’ 1 second events will it have seen due to the background?

How to deal with trials

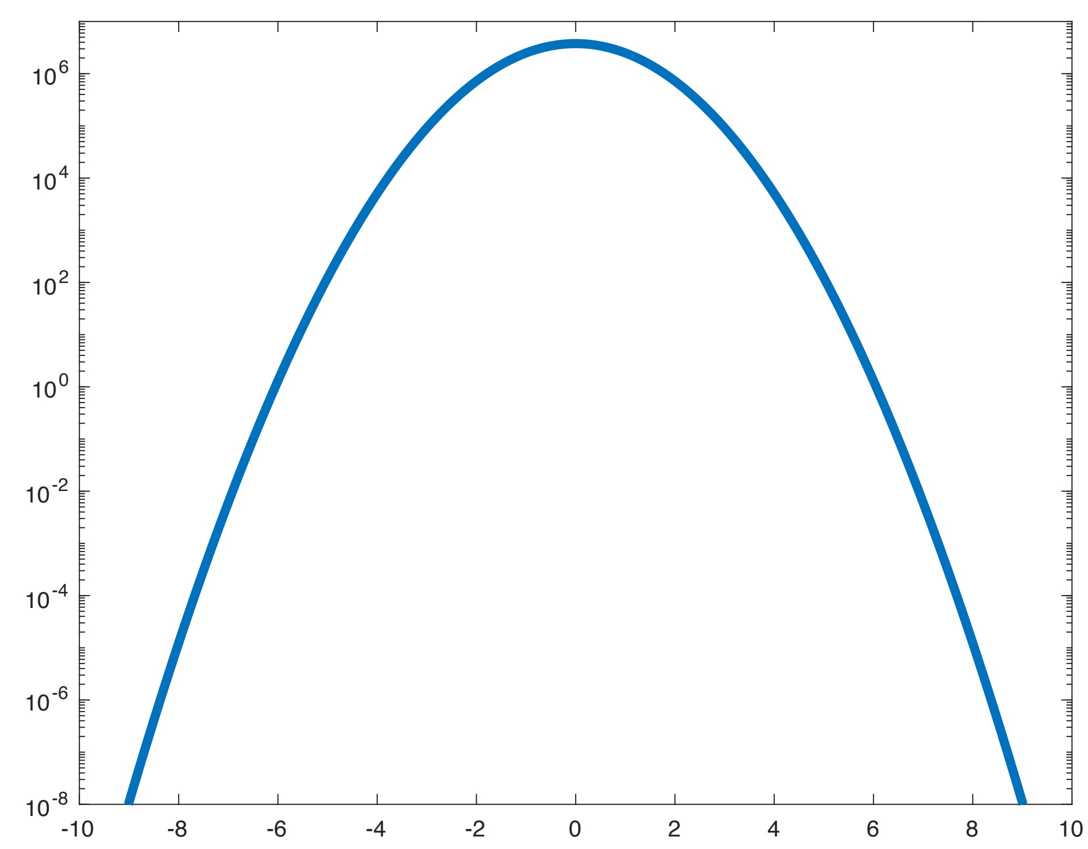
- Calculate the pdf() of the entire run
- Integral of new pdf() is now the expected number of events of a particular brightness





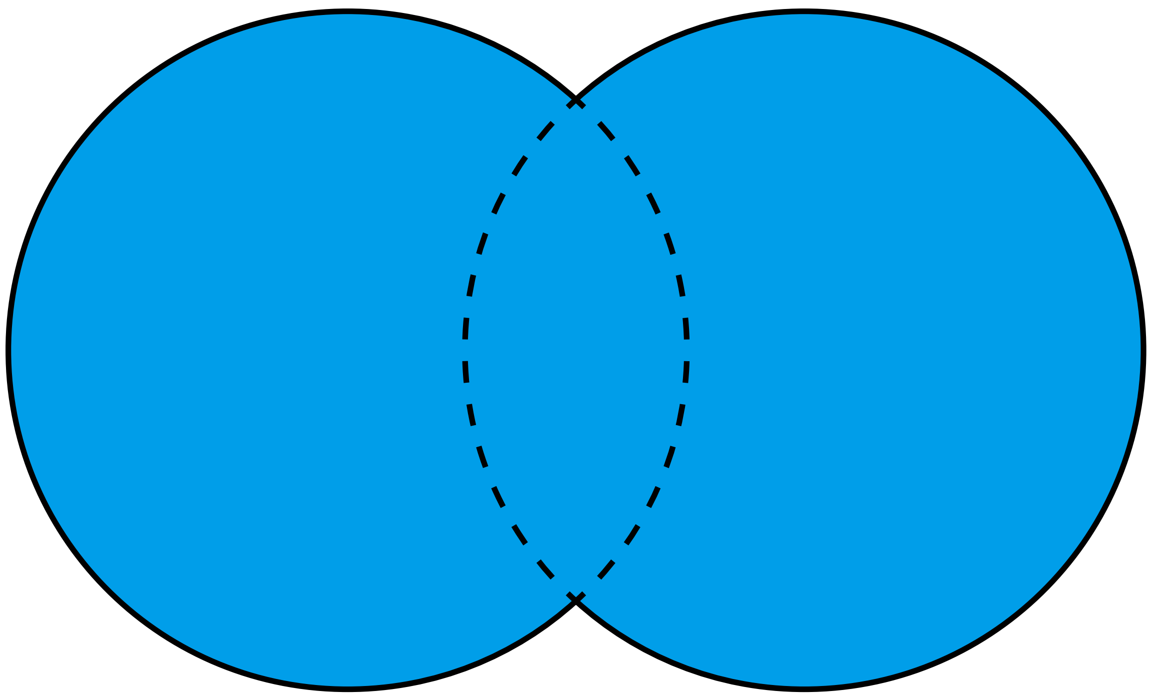
Thinking through carefully

- How many events above X would I expect to see?
- Question we wanted was Probability to see an event stronger than X here <u>OR</u> here <u>OR</u> here…



OR

$$P(A \text{ or } \mathrm{B}) = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A \text{ and } B)}$$

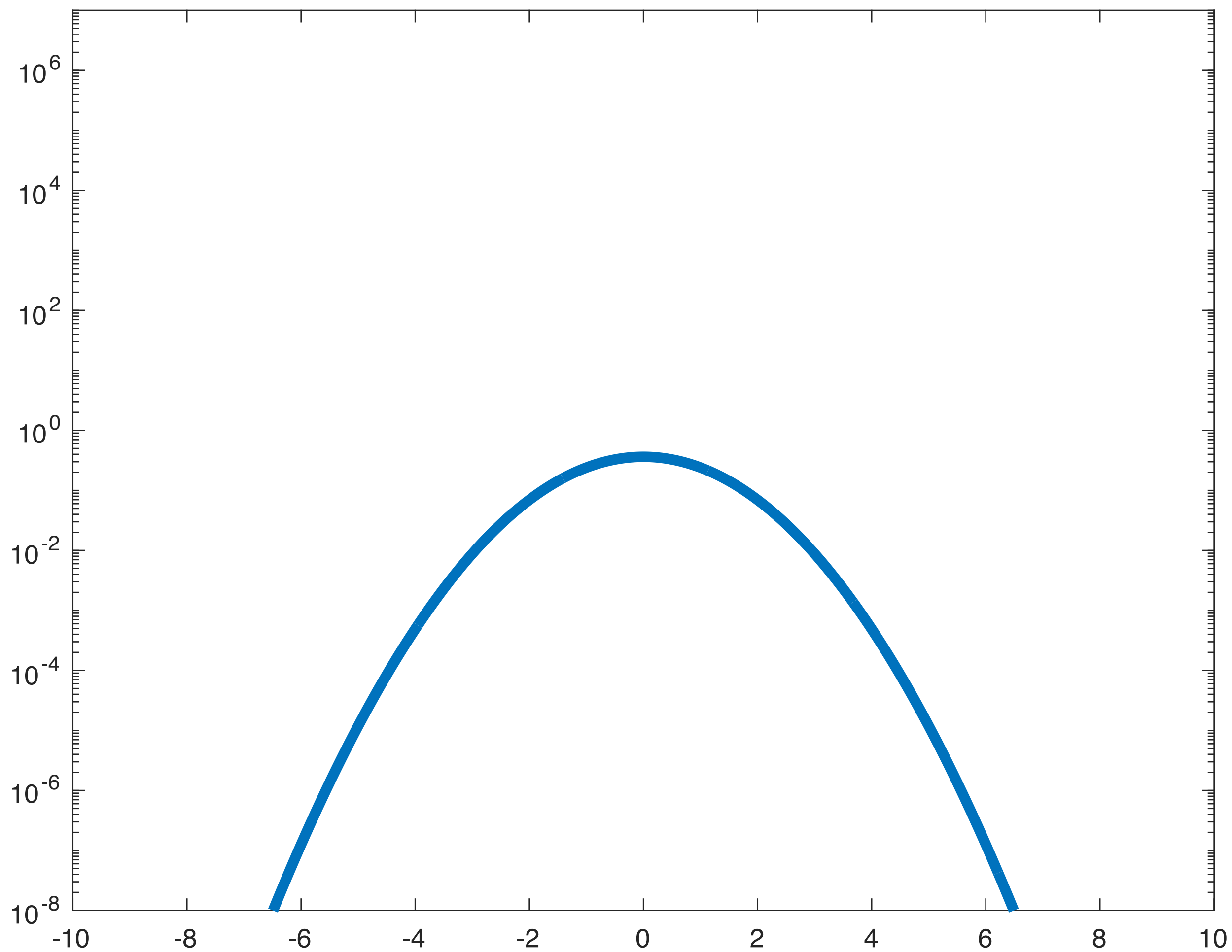


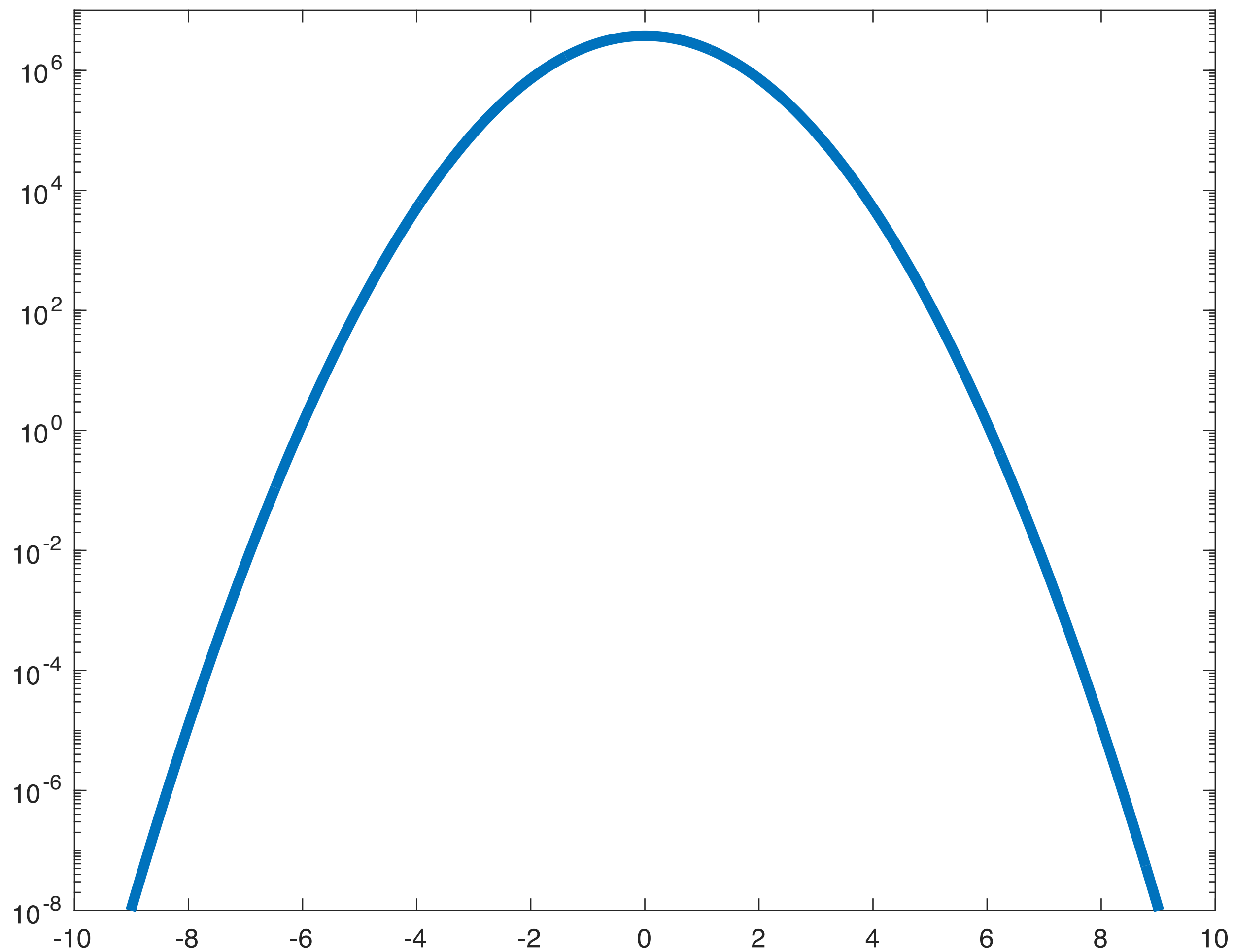
- When small prob

$$P(A \text{ or } \mathrm{B}) \approx \mathrm{P(A)} + \mathrm{P(B)}$$

How to deal with trials

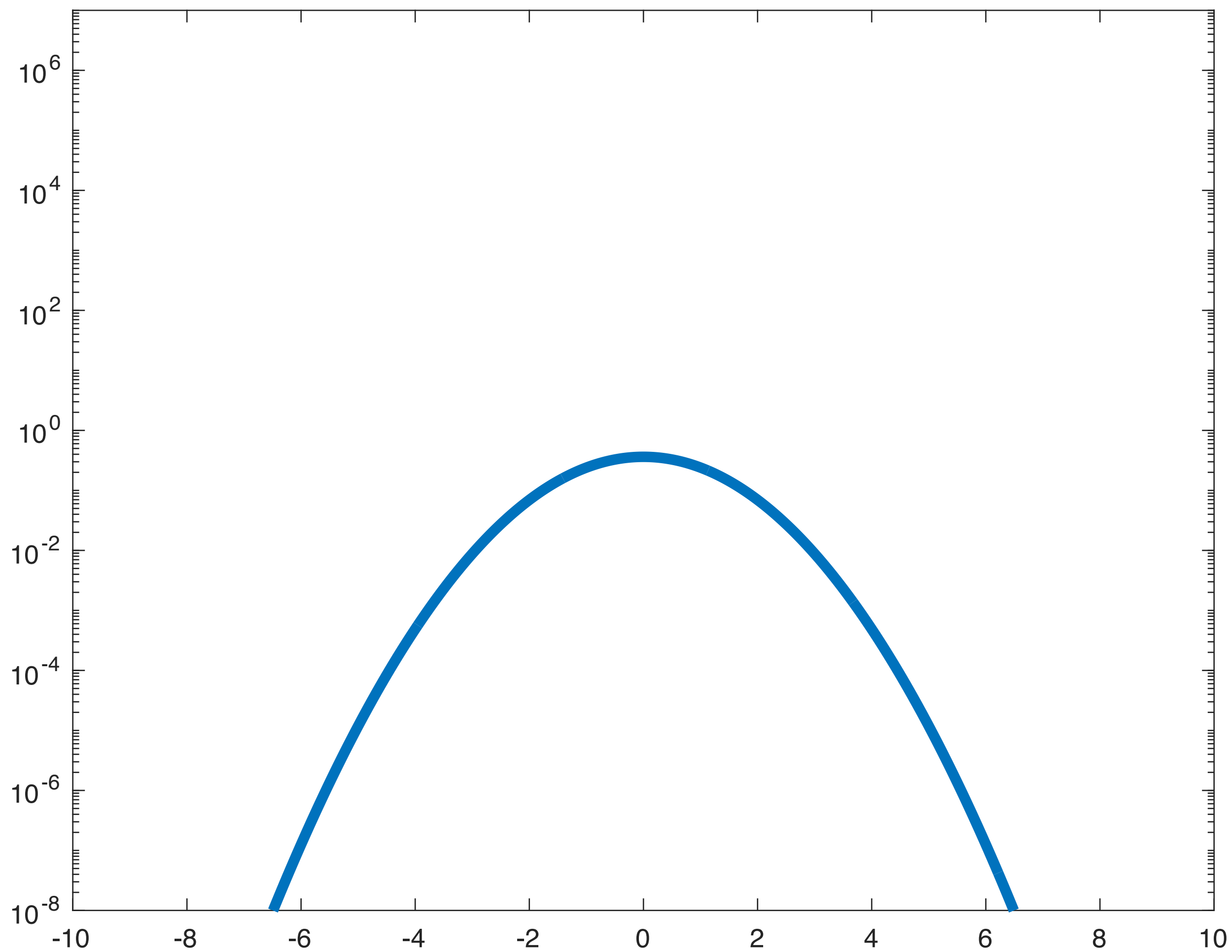
- Multiplying pdf() gives number of an event of strength X over ensemble; integral is number of expected events stronger than X
- When expected number is small; expected number ≈ Prob of 1 event stronger than X

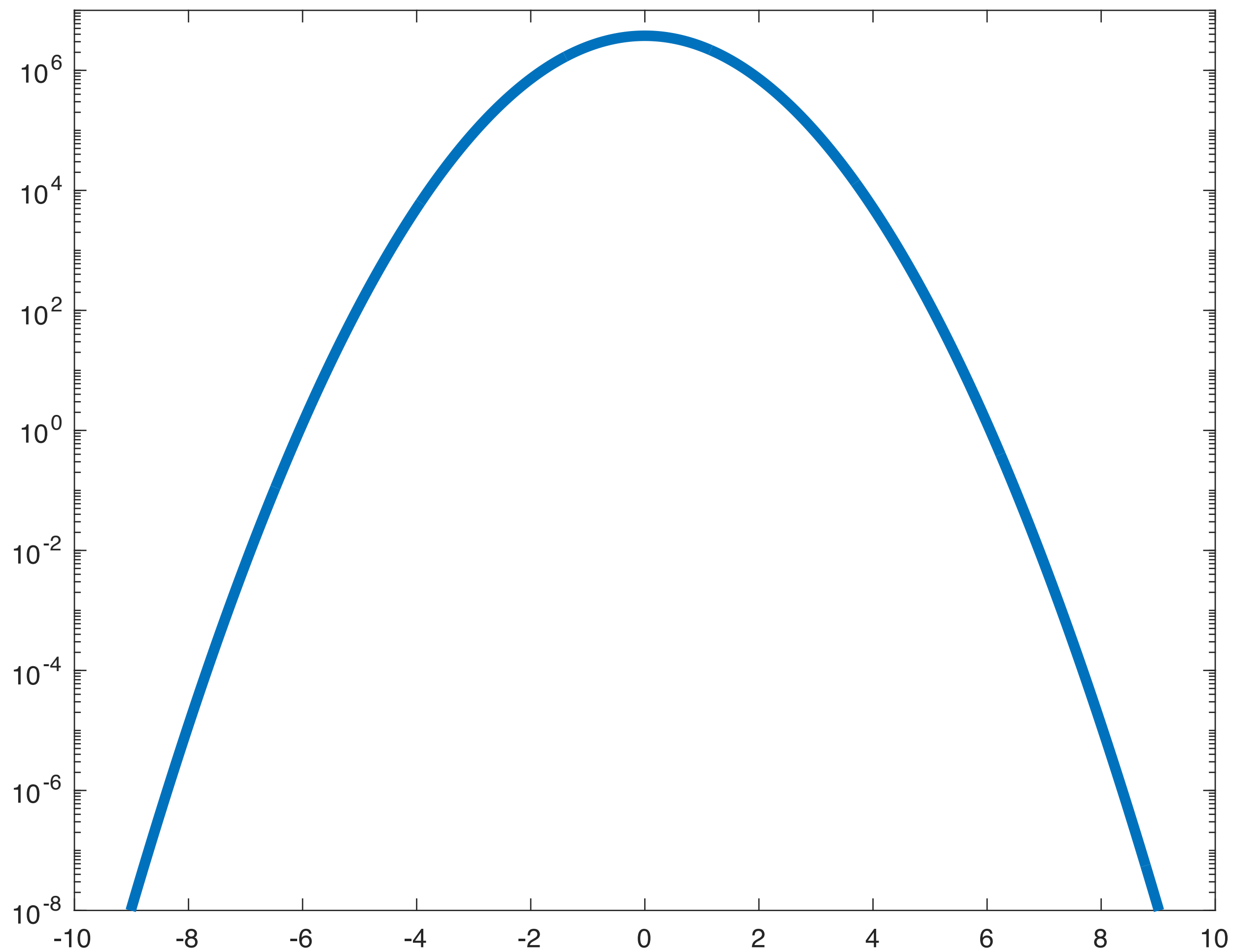




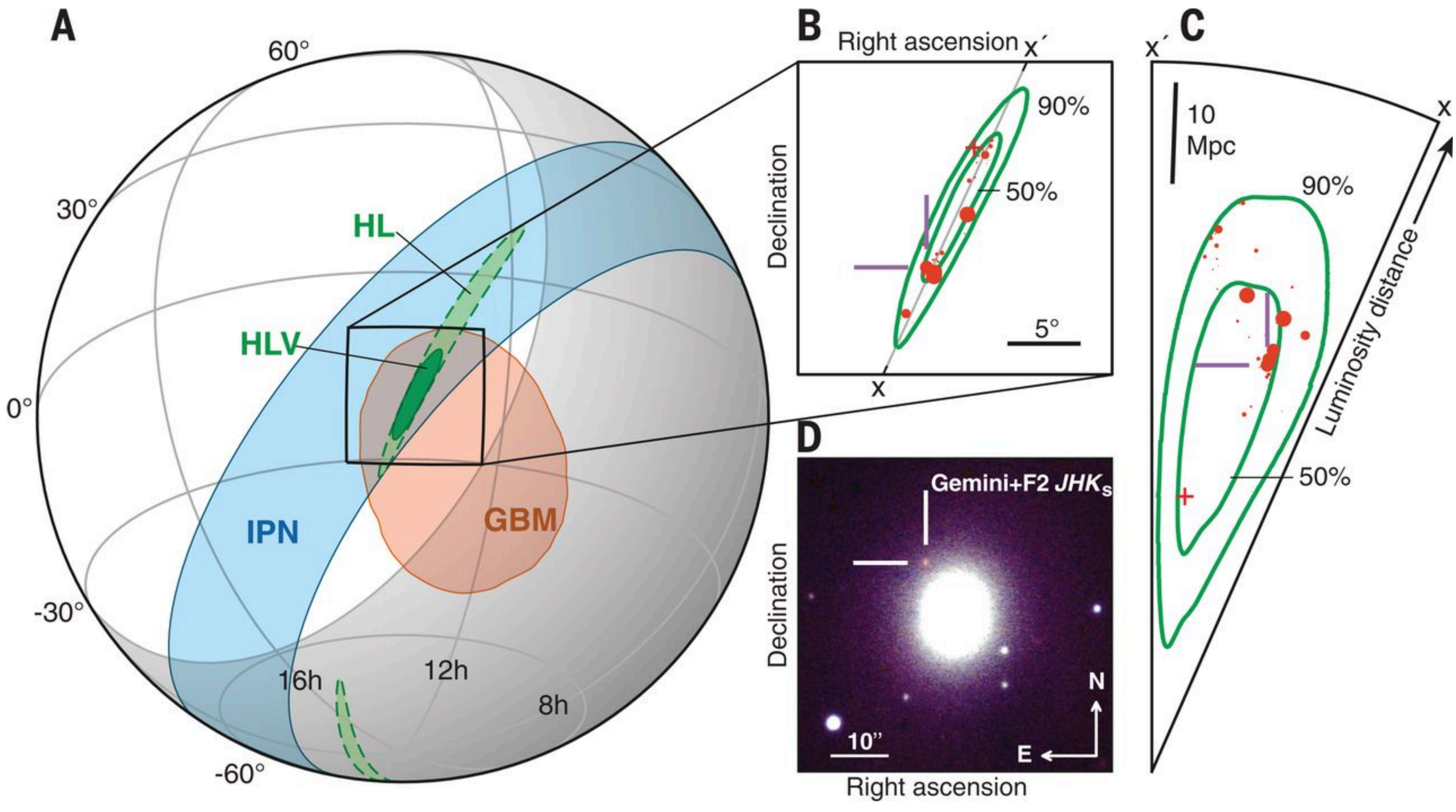
How to deal with trials

- There is a sensitivity penalty
- Really must get tails right
- Sensitivity penalty is small (~3x)





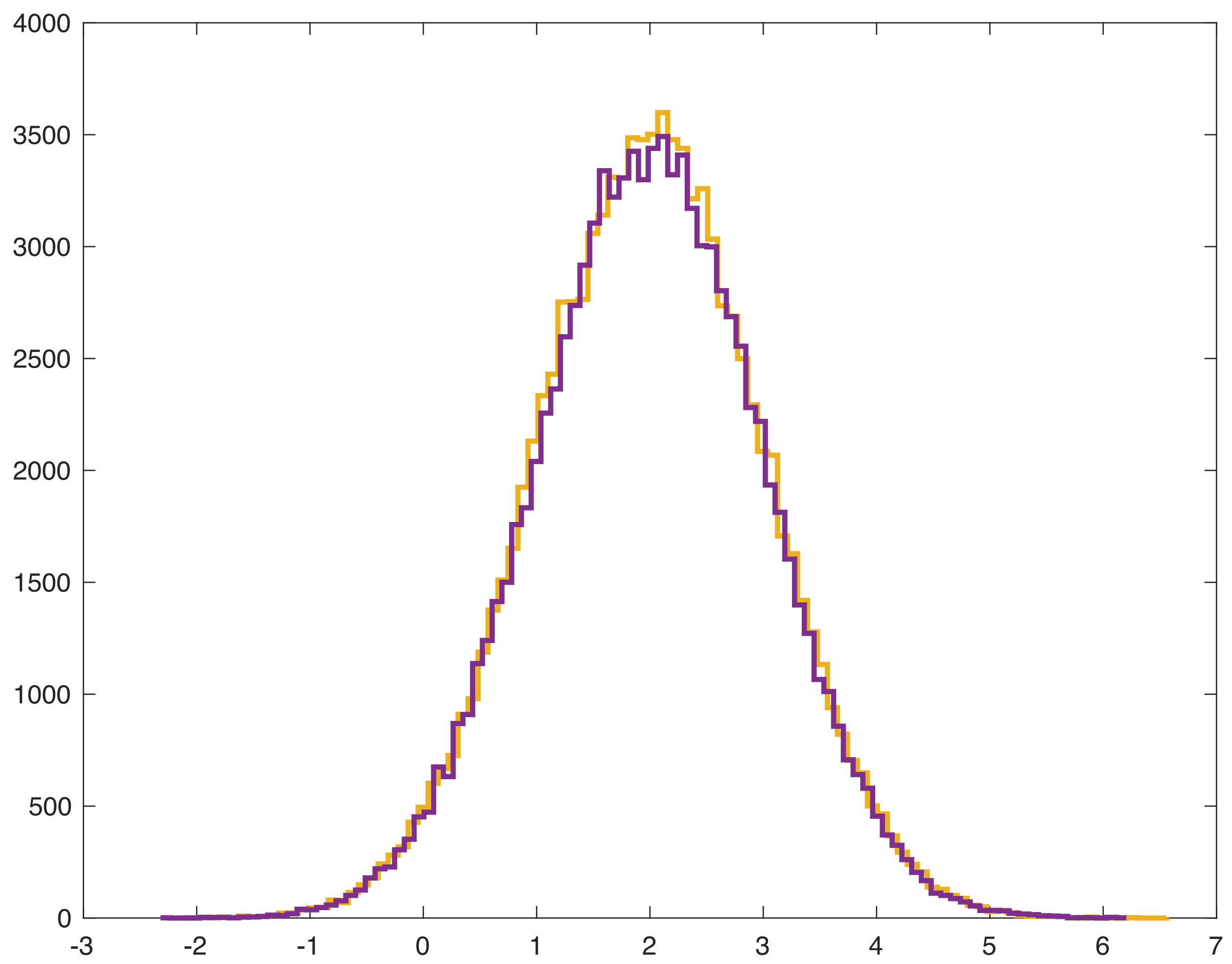
LIGO NS-NS counterpart search



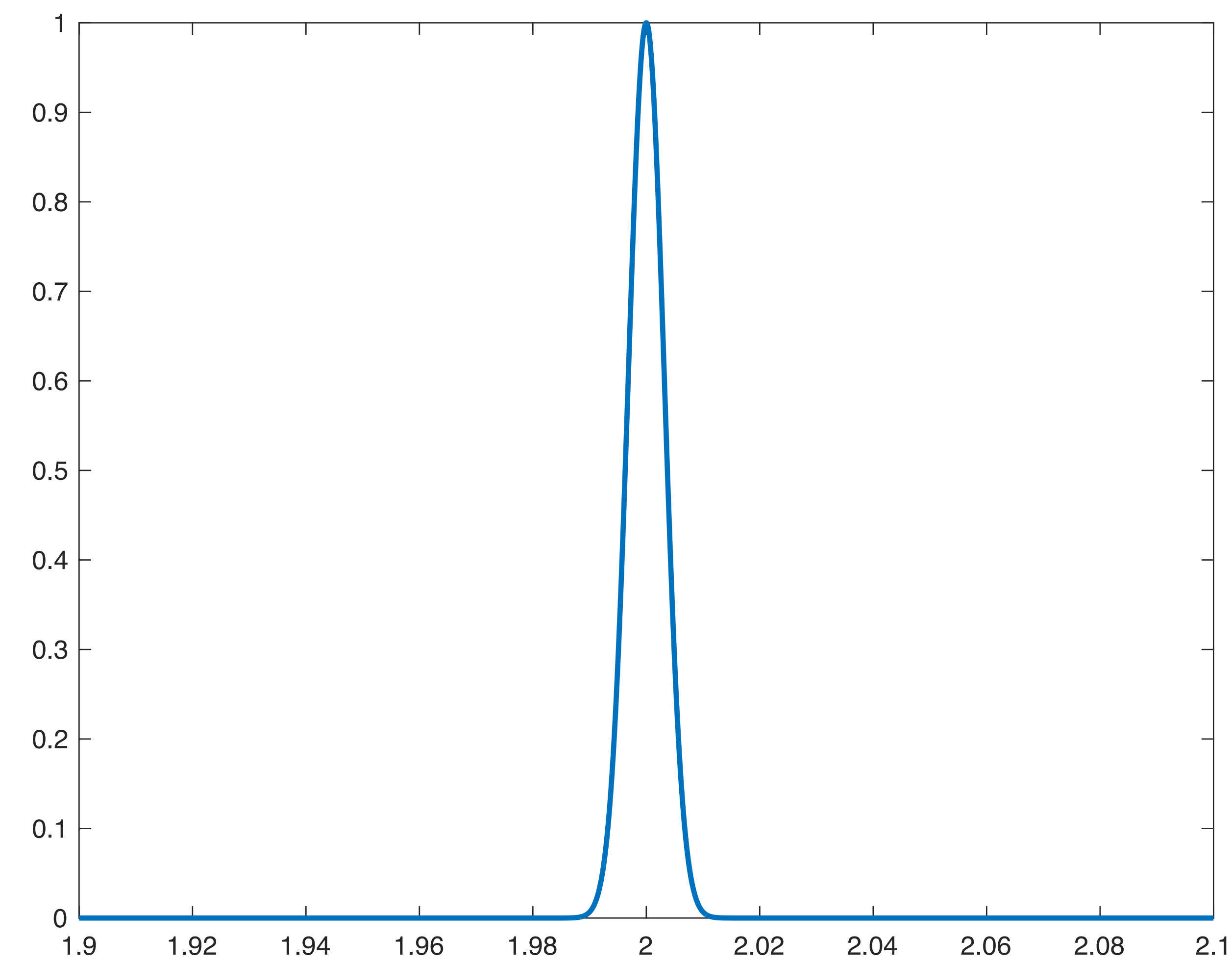
Parameters

Single parameter distribution

Population distribution
(10^5 measurements)



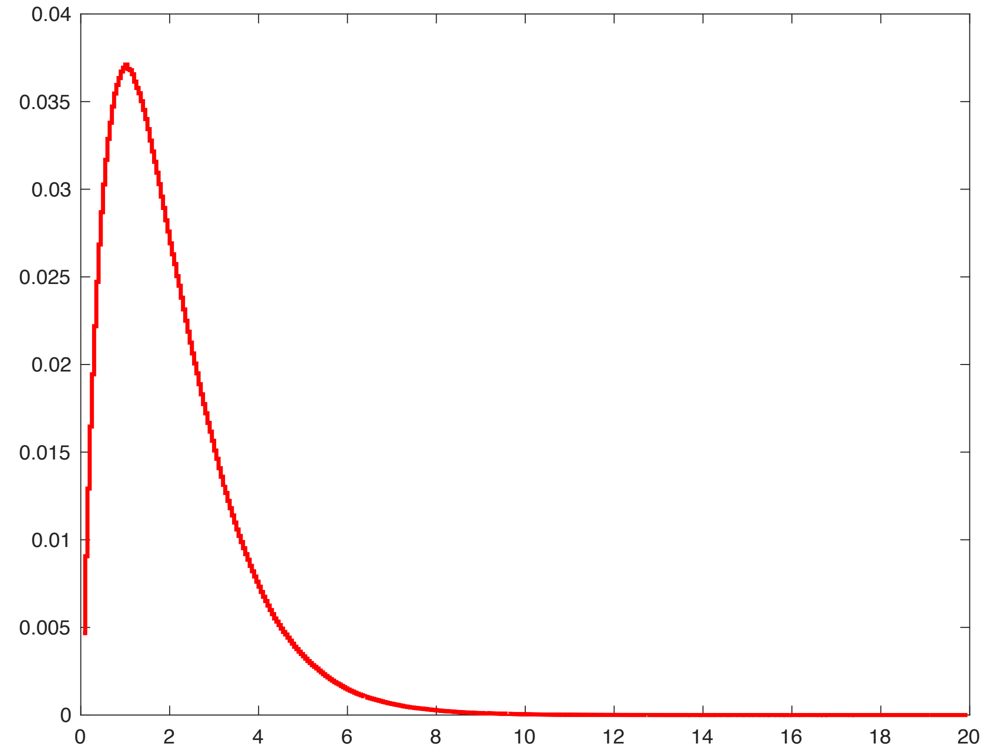
Distribution of the mean



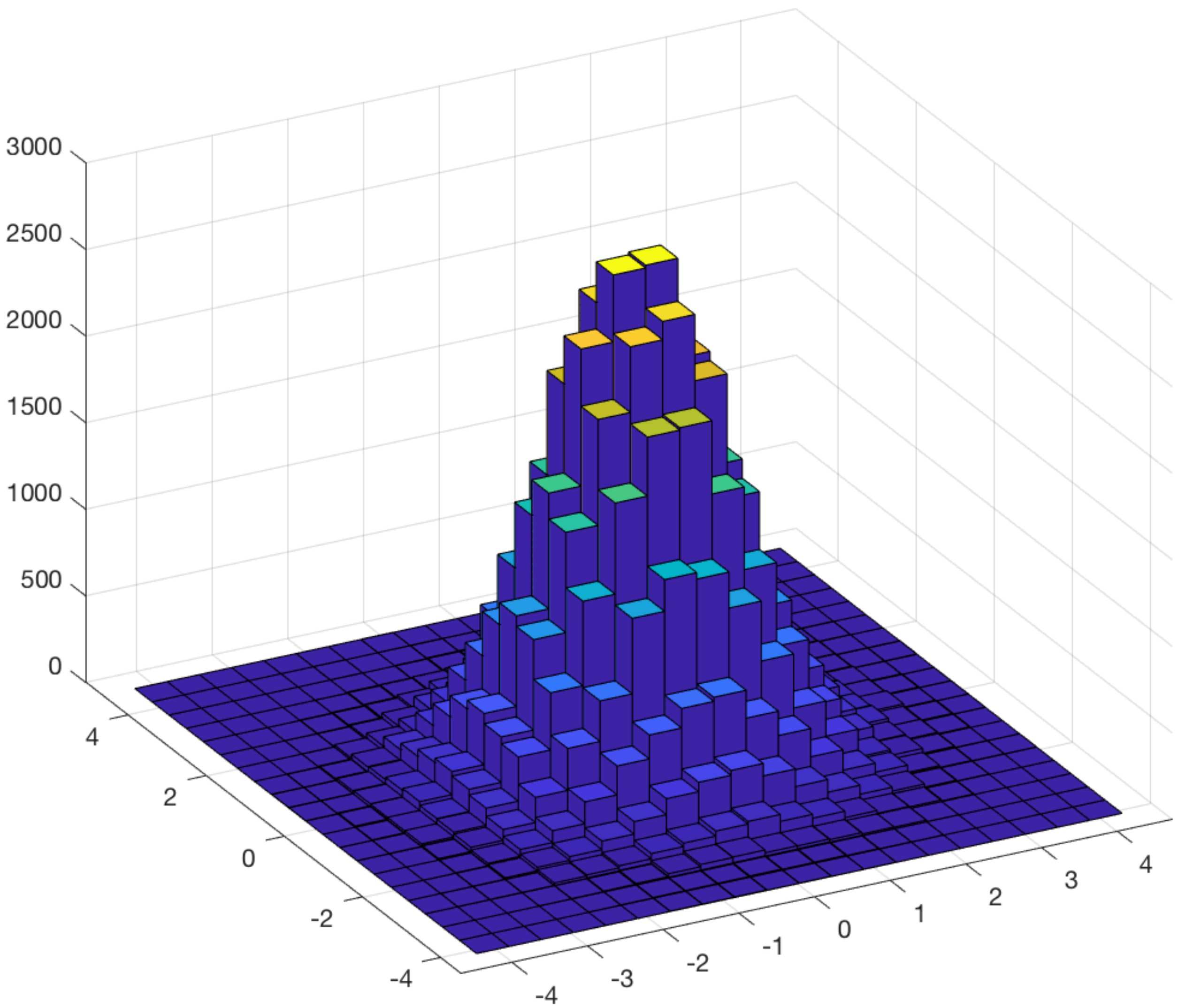
$$\hat{m} = \langle \text{data} \rangle$$

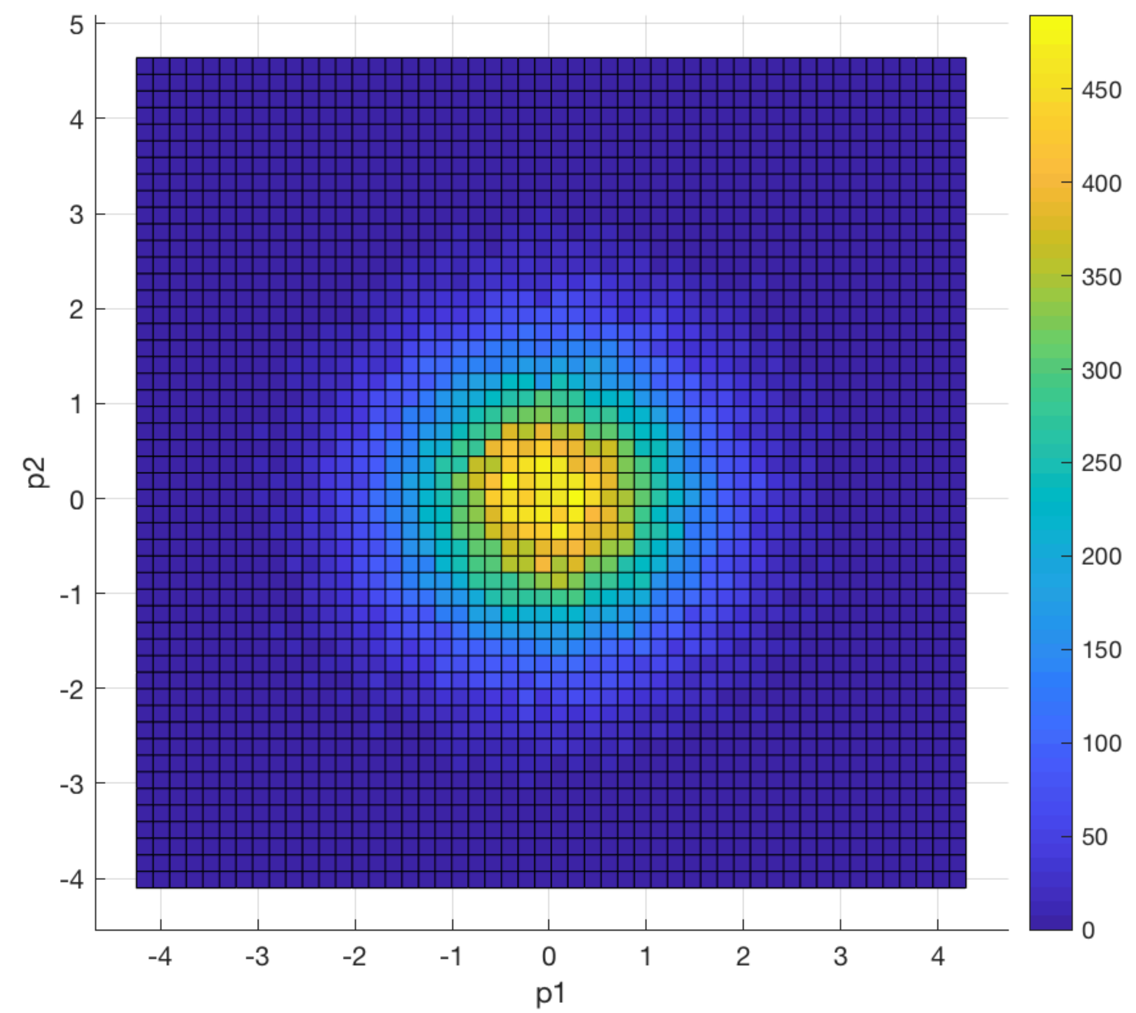
Parameters

- Treat like a measurement—a value with some distribution
- Because the underlying is Gaussian does not mean the parameter is Gaussian, but can often propagate answer

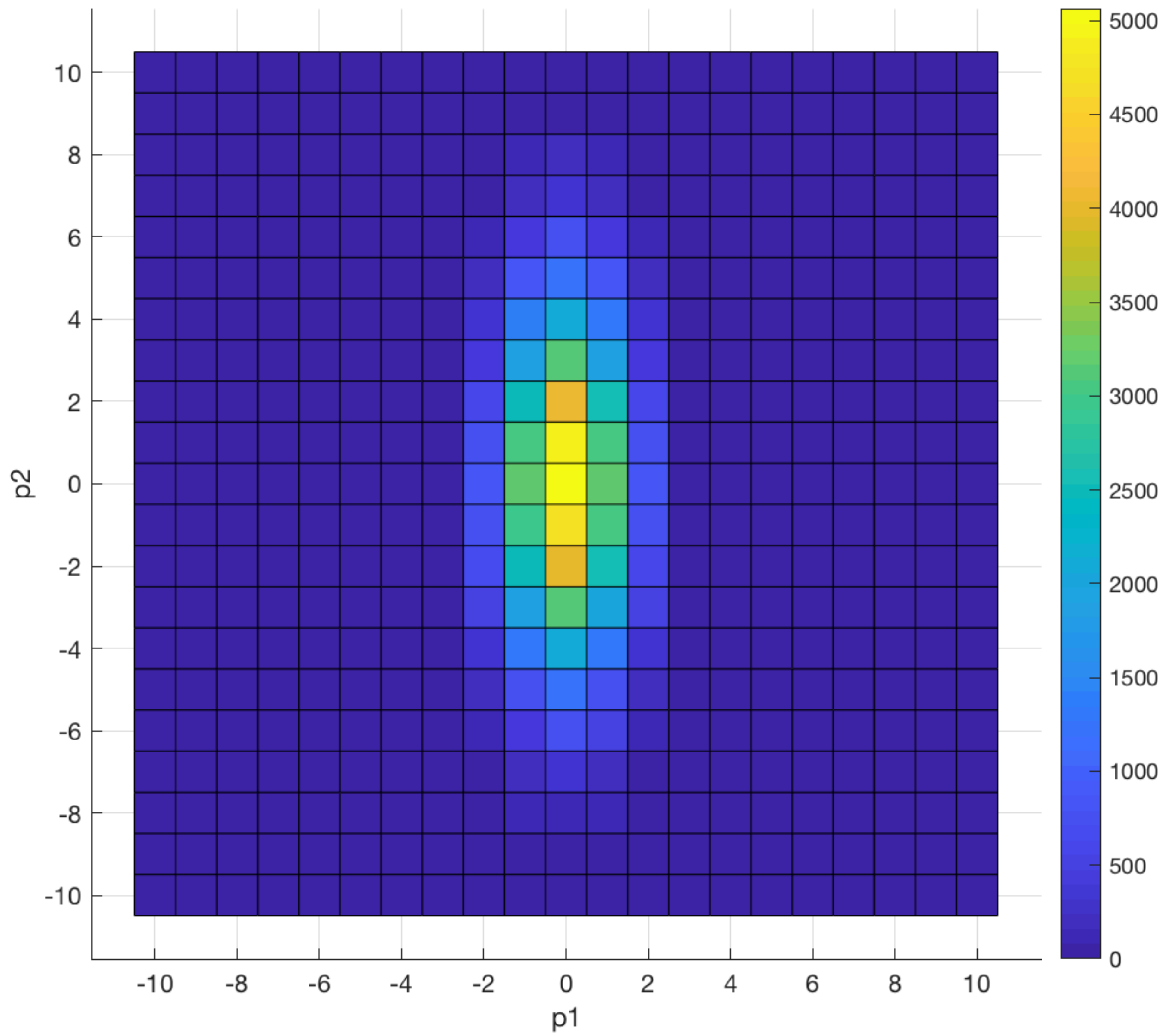


Multiple parameters



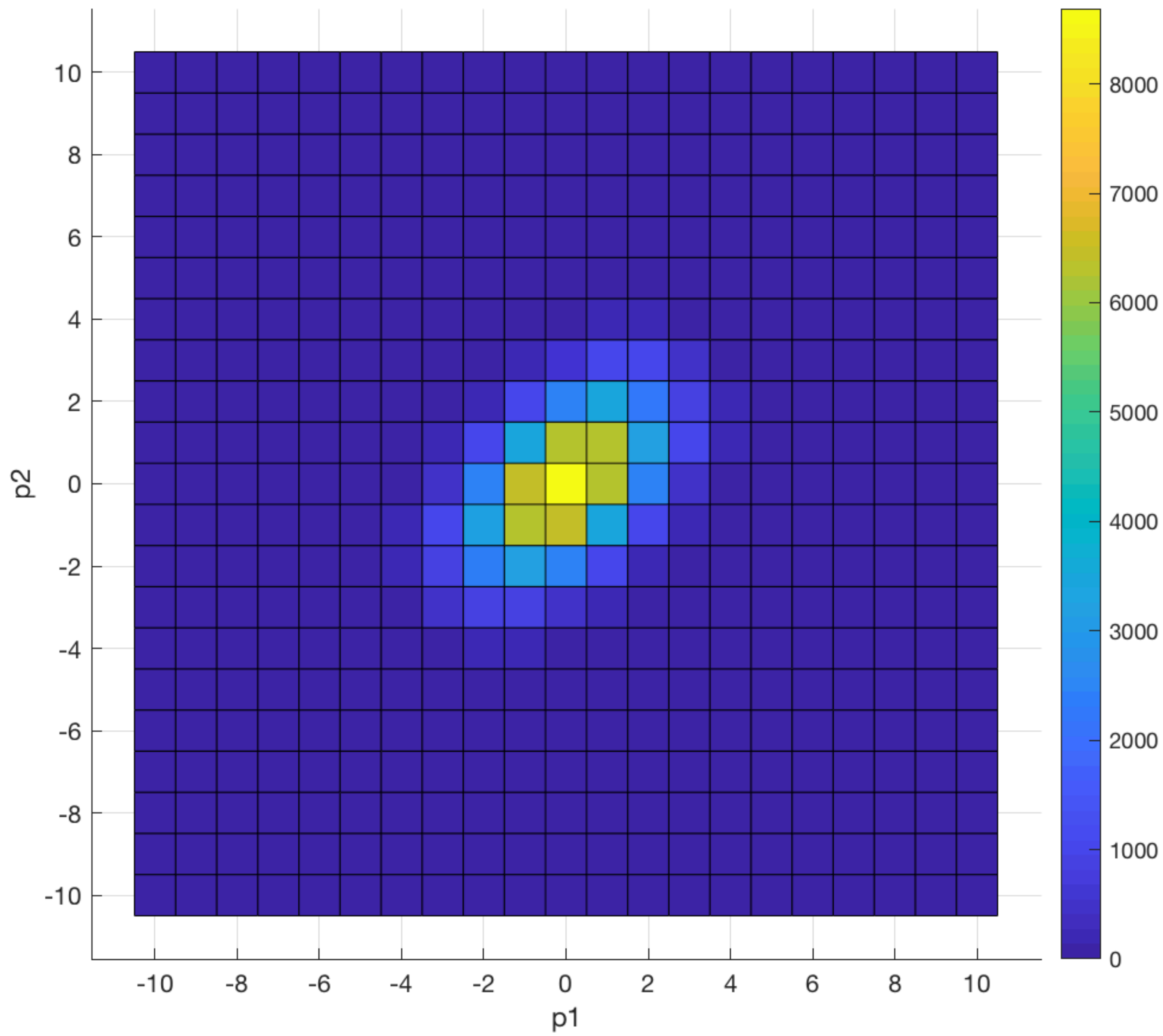


Non-equal variance



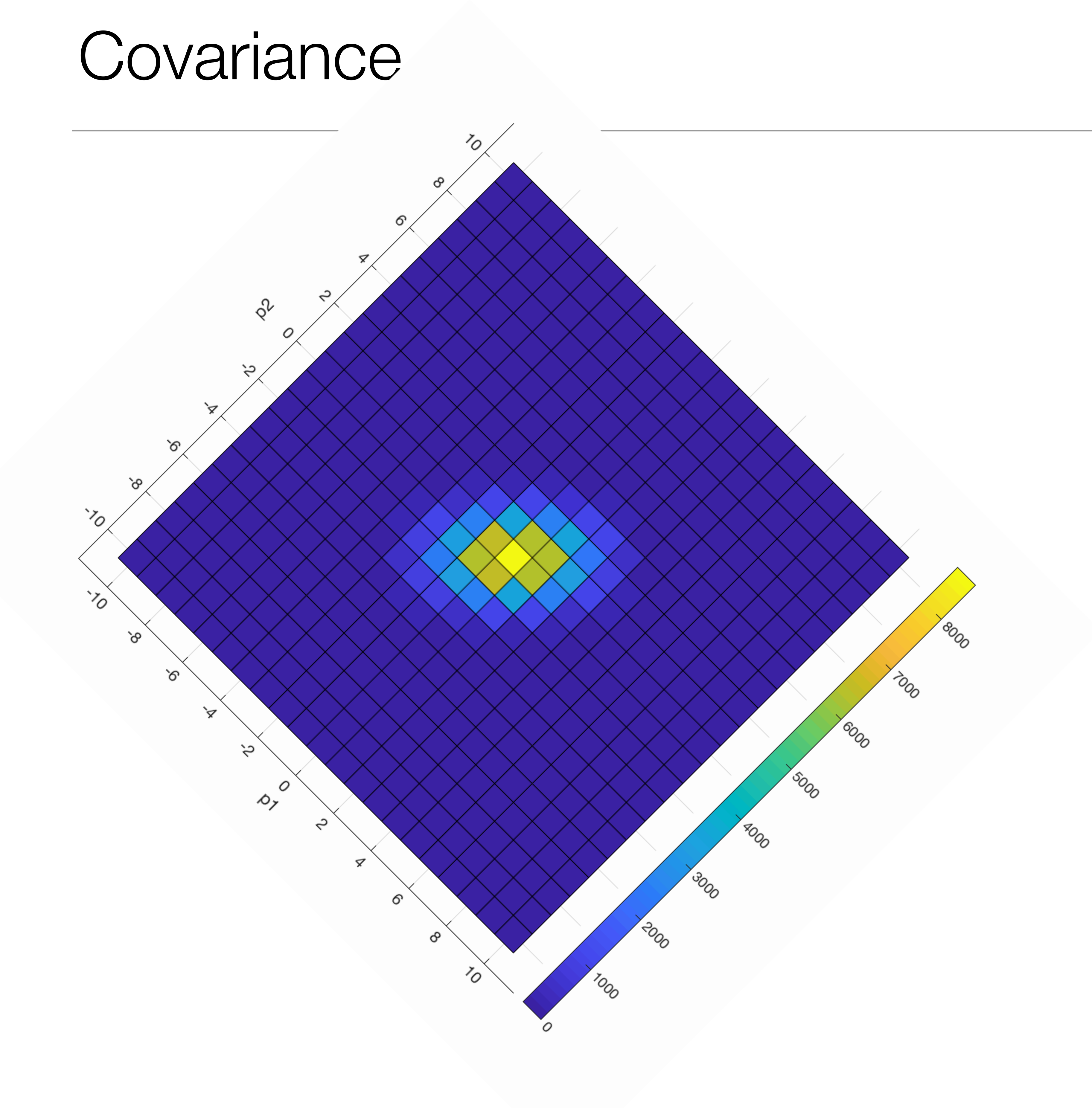
$$\sigma_1^2; \sigma_2^2$$

Covariance



$$C(p_1, p_2) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}$$

Covariance



$$C(p_1', p_2') = \begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix}$$

Triangle plot

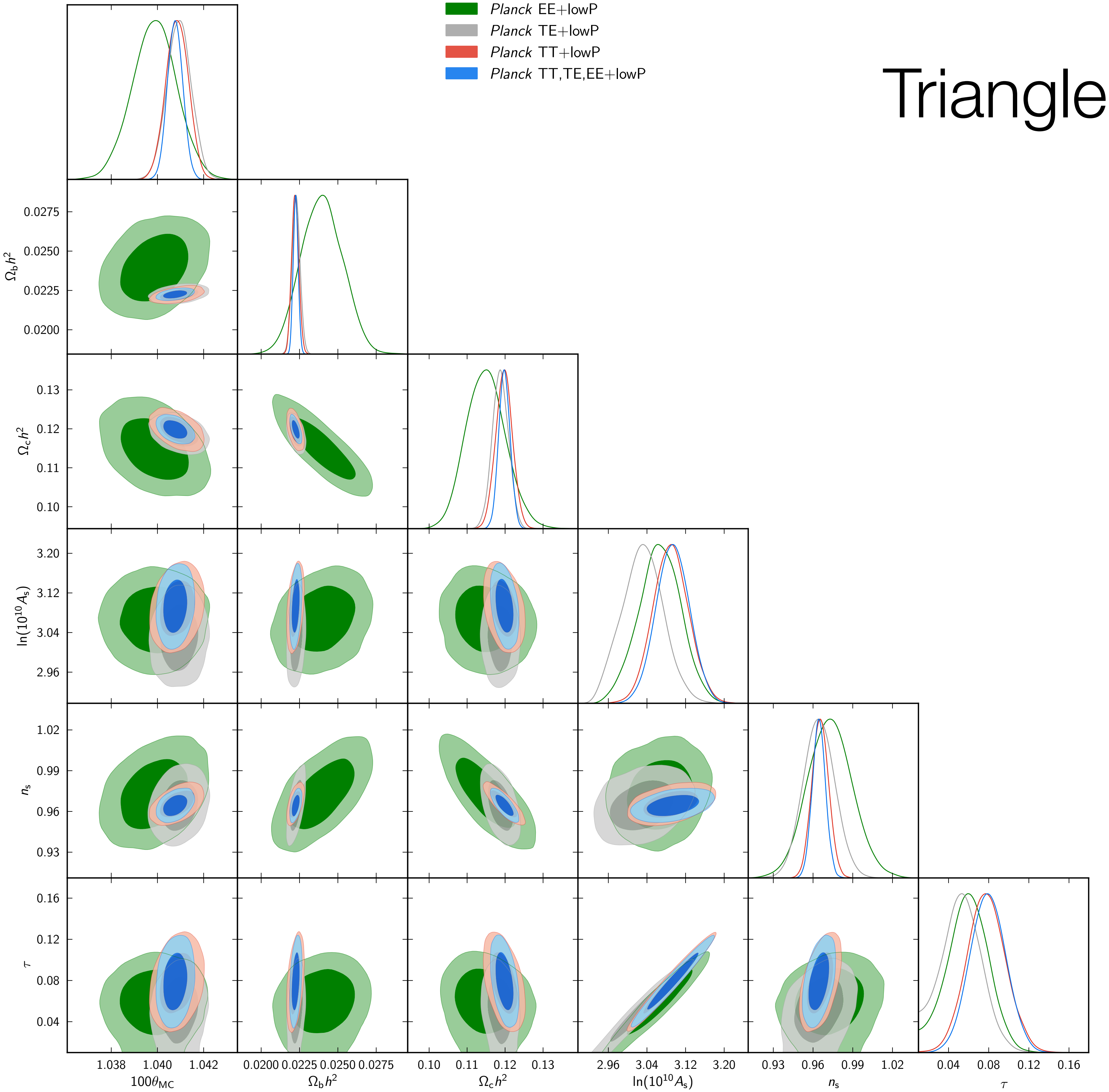
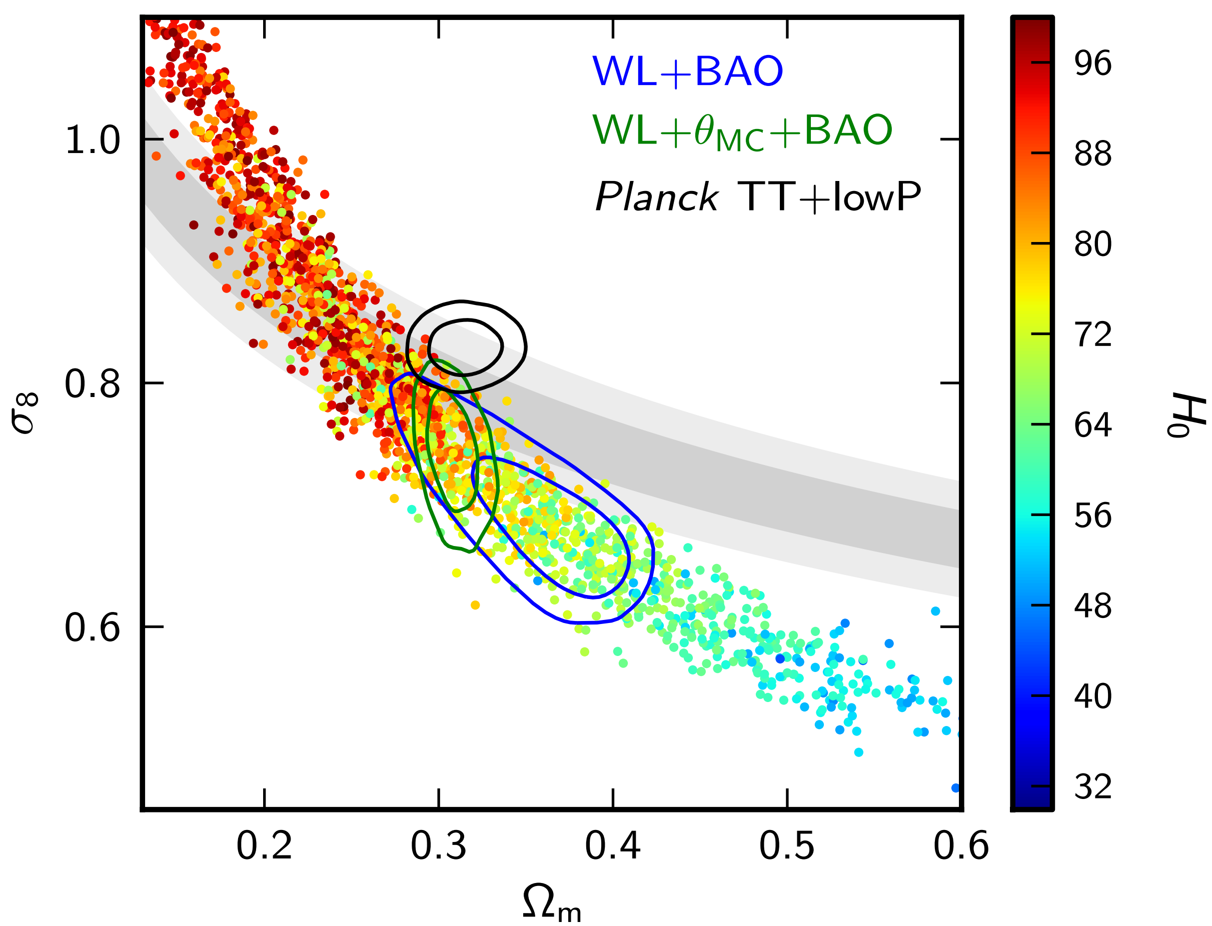
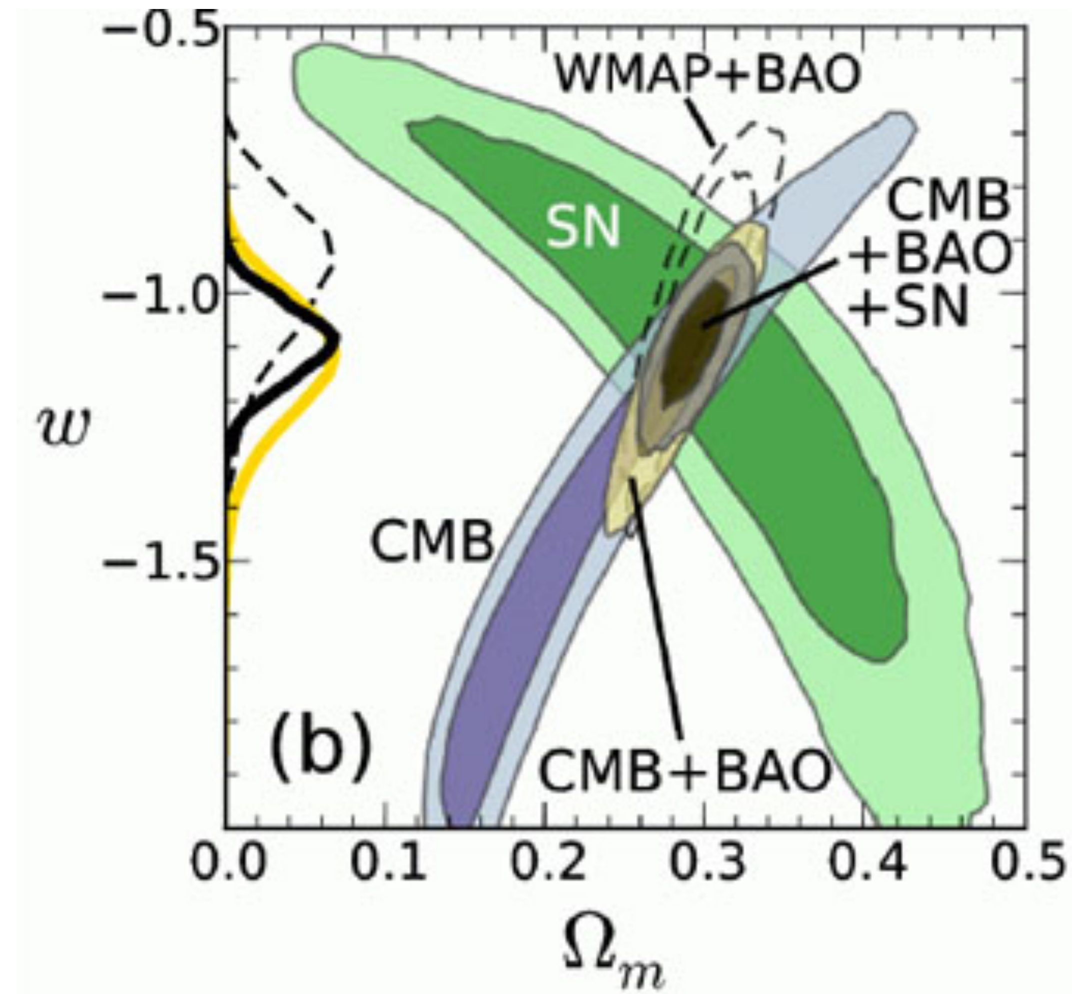


Fig. 6. Comparison of the base ΛCDM model parameter constraints from *Planck* temperature and polarization data.

Cosmology parameters



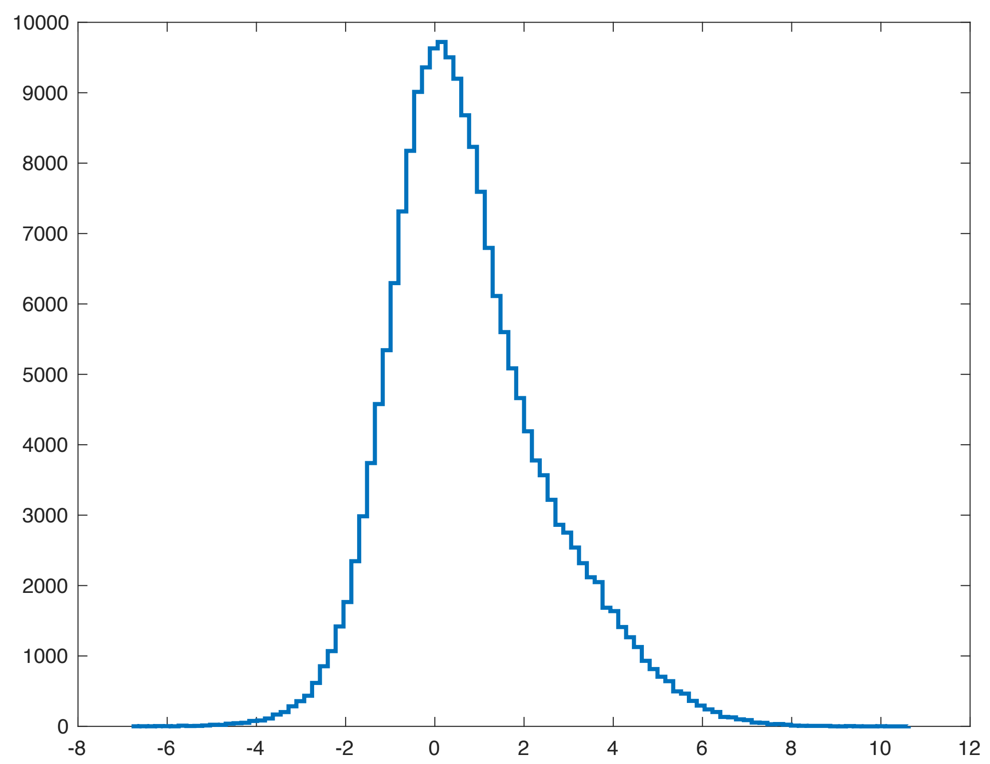


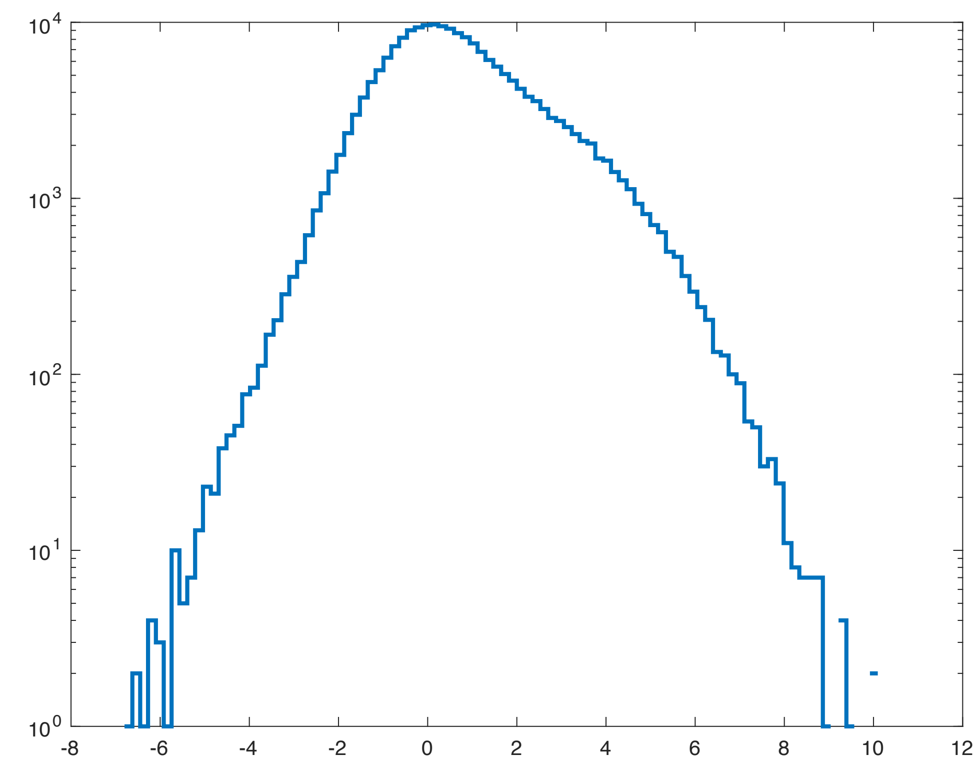
Take home message

- Multiple parameters can depend on each other in complicated ways
- Don’t assume they are independent of one another
- Can be due to Theory, your model, instrument, nature, or interactions

The art of parameterization

So you have a distribution to parametrize…





How do you pick 'good' parameters?

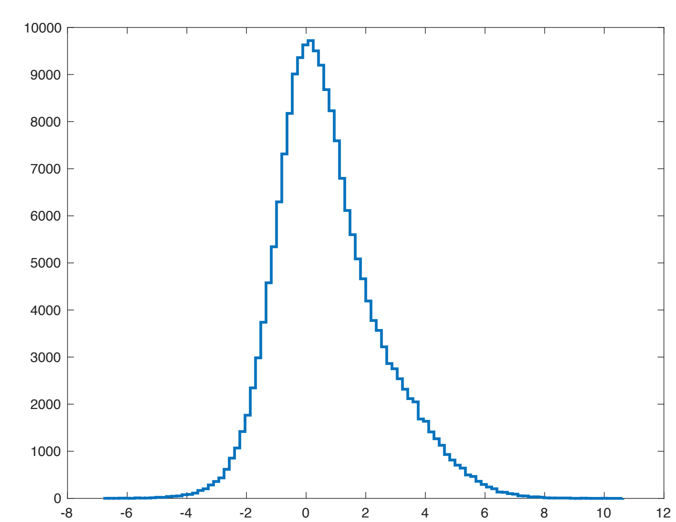
No good answer, topic of the remainder of the course

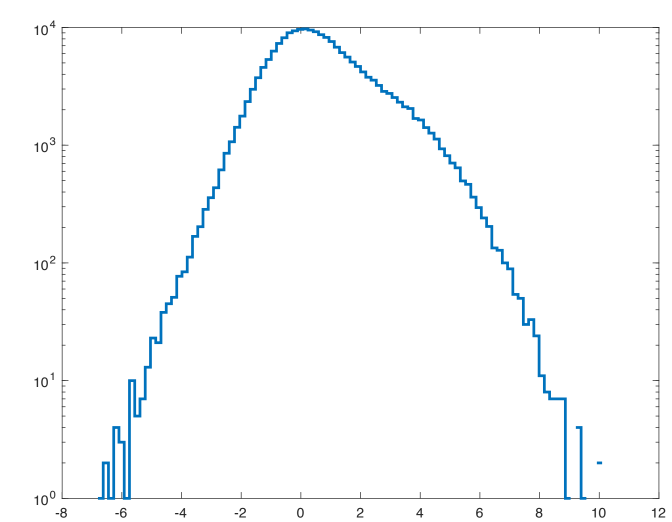
Failure modes

- Too few parameters of the wrong type don’t describe the distribution ⇛ significance calculations wrong

- Too many parameters just ‘fit the elephant’

    poorly constrained and/or highly covariant ⇛ unstable





Maybe it’s not new?

- Sometimes you can propagate the distribution from an earlier one that you understood.

Do you have a physical model?

- Physical models, particularly of an instrument, tend to work better.
- If they don’t work, you often learn something

Maybe there is a systematic?

- Often weirdness is caused by systematics
- Finding systematics is really adding to a physical model

Tests

- What are the covariances of the parameters on subsets of the data?
- Can I take special data to prove/explore a proposed physical model?
- Do parameters respond to my worries? (Does it pass jackknife tests?)

Love your data,
This is the value you add