Lecture 8:

- Loose ends: Real ensembles, Constant priors, Robust data comparisons

- ‘Binsmanship’ and Dodgy Error Bars
- More Things to Avoid
- Ways to Display Uncertainties
- Visualising Multi-Dimensional Data
- Boxes, Whiskers and Violins

As previously stated, frequentist bounds are all about the distribution of the ensemble of hypothetical experiments and not about ascribing meaning to your particular interval.

But what if lots of people do experiments and each defines frequentist bounds, so that you start to have a **real** ensemble. How do you then use this to set bounds on models?

Still Bayesian! Make use of the likelihoods for all these data sets together *(not their frequentist bounds!)* and choose your prior etc. There is no other way! As the ensemble becomes larger and larger, the prior becomes less and less important, and the distinction between frequentist and Bayesian bounds goes away.

Pragmatism: You can use frequentist bounds for models when it gives the same answer as Bayesian bounds.

The Point of Frequentism:

Want to display the results of analyses in a model-independent way that has the most general possible applicability

Absolutely!! Always do this! For example, try to provide sufficient views of the data to allow others to roughly reproduce your results, and show the likelihood distribution, **which gives the full frequentist information content of the data.**

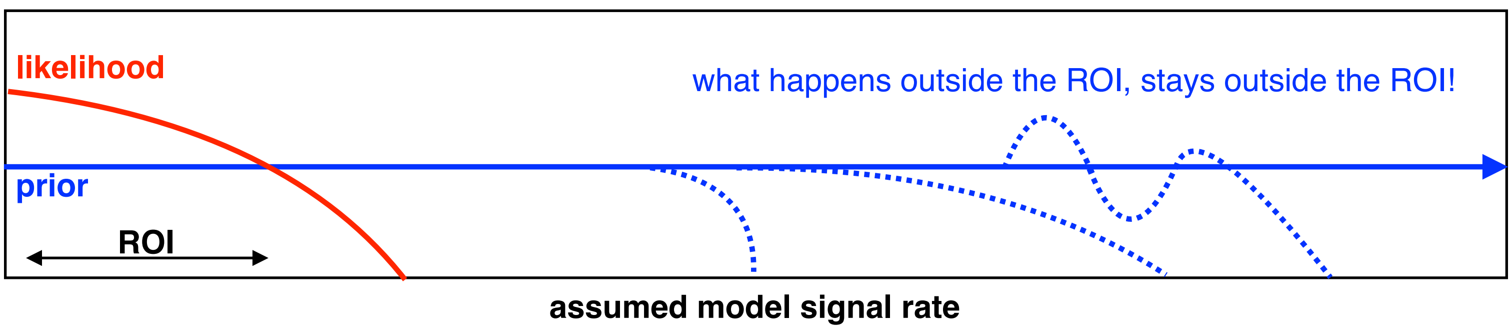
But, if you then want to use this to constrain models, that’s Bayesian!

Both of these are important aspects of data presentation.

The Use of "Uniform" Priors

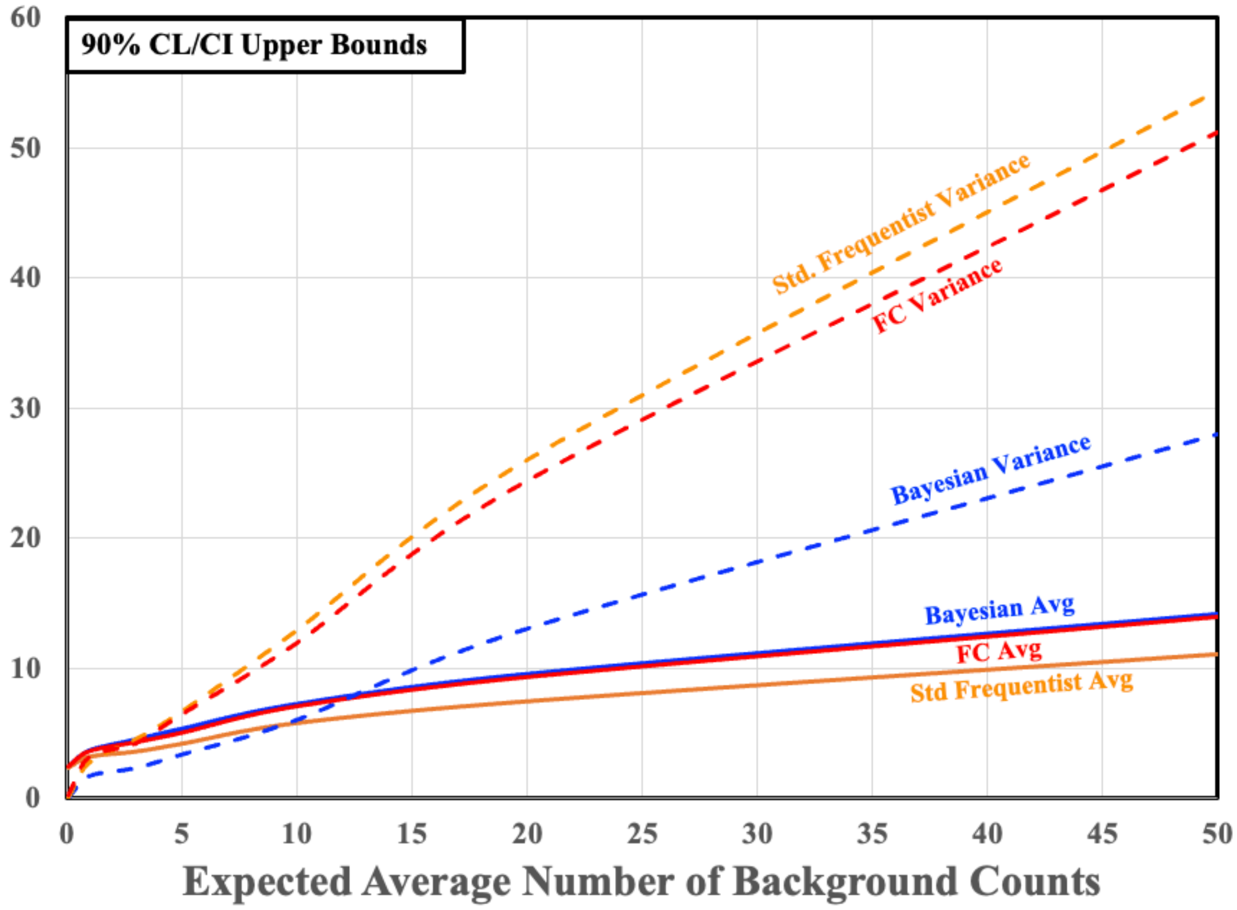
This refers to using a constant prior for a particular model parameter. For example, priors uniform in signal rate mean that you ascribe equal weight to all signal rates.

But is that really realistic? That would allow the possibility of an infinitely large signal and results in a probability distribution that cannot be normalised!!

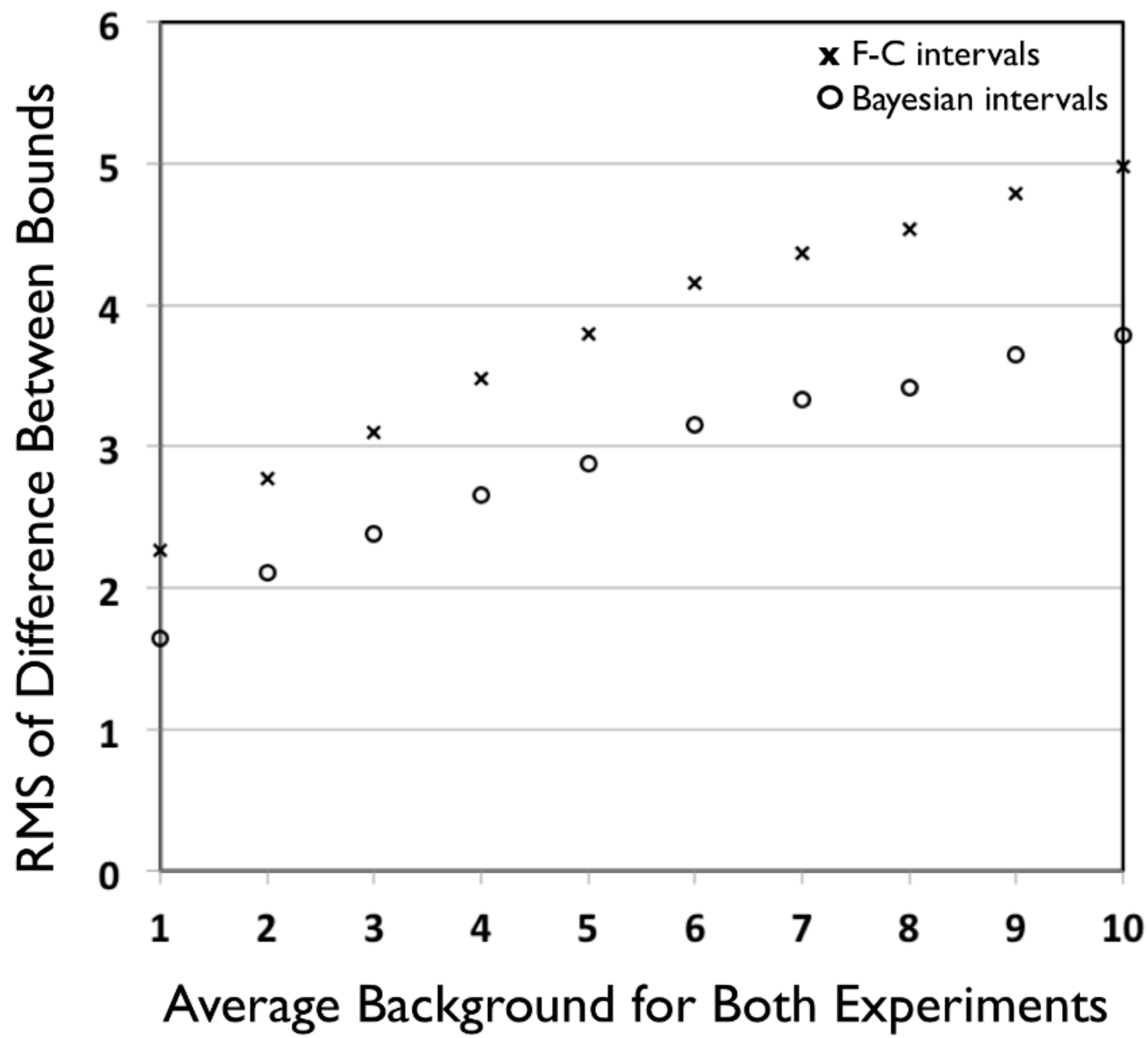


What we actually mean is that the prior is roughly constant in the vicinity of the region of interest, and then tails off in some way that does not need to be specified because the likelihood crushes its impact as soon as you get much outside the ROI

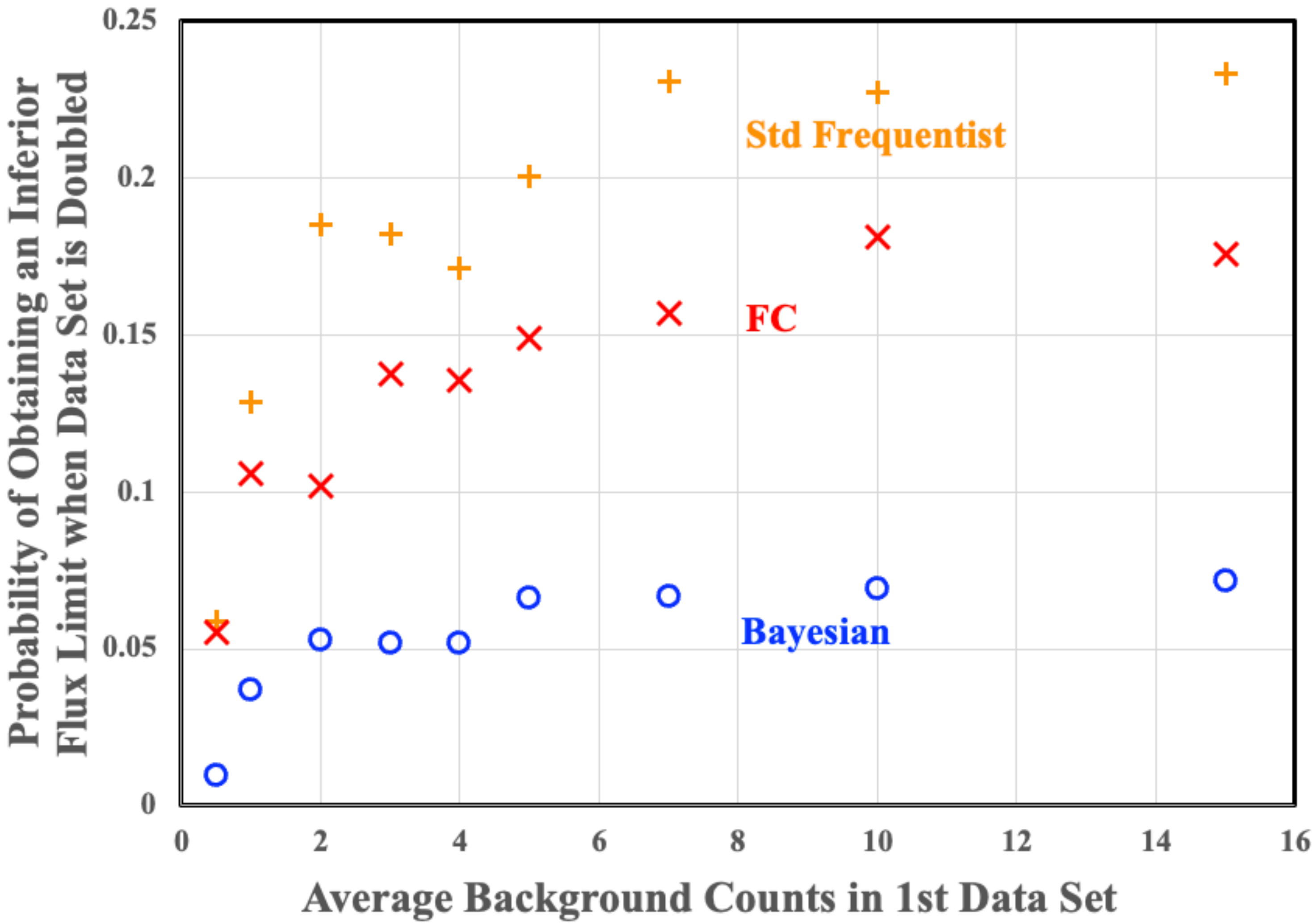
Robustness of Upper Bounds



The numerical values of Bayesian bounds have notably less variance than frequentist methodologies - more robust for comparison of experimental results!

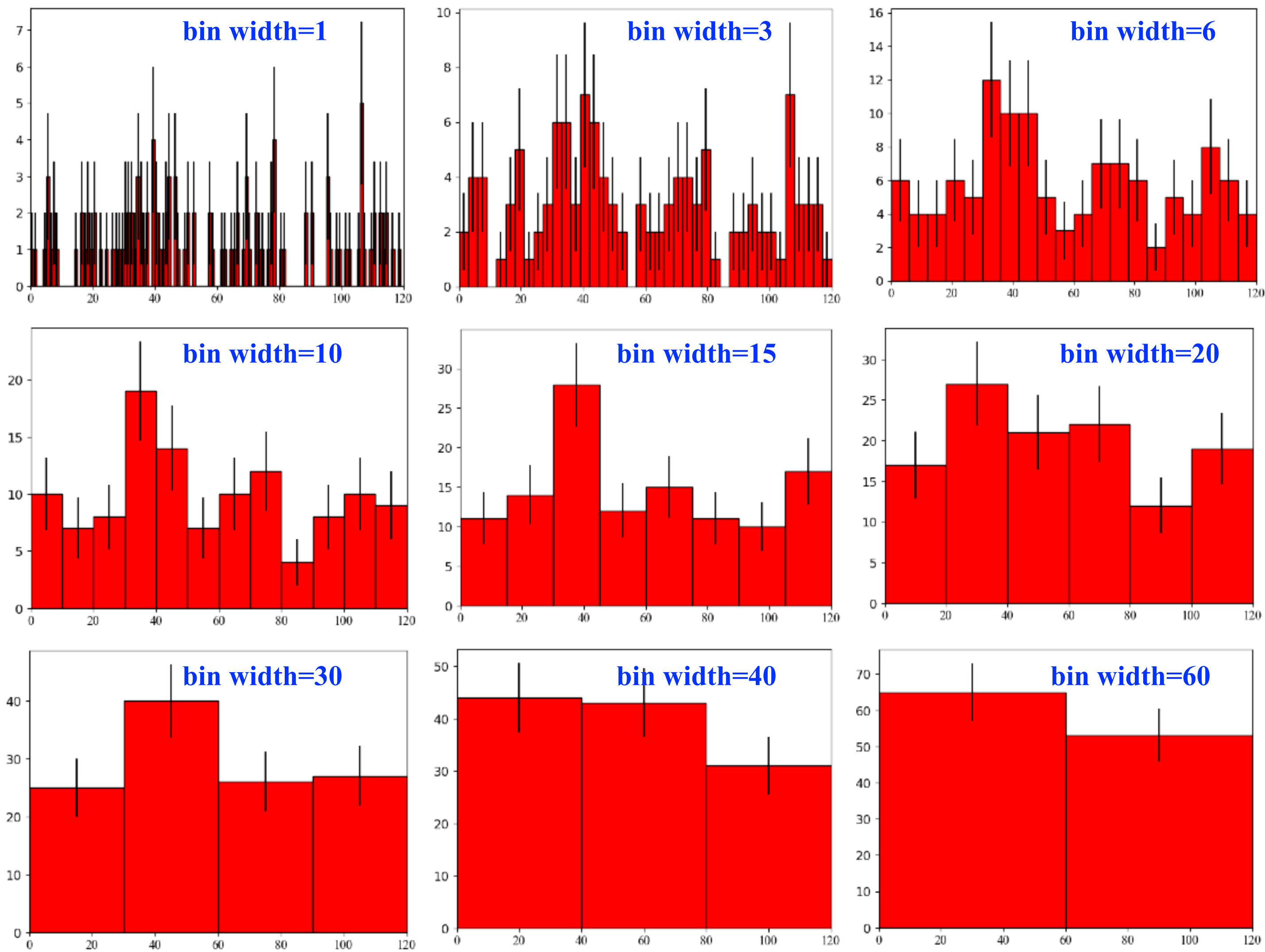


The RMS deviation between 90% CL/CI upper interval bounds from two experiments with the same expected background level. Cases for uniform-prior Bayesian (circles) and Feldman-Cousins (crosses) constructions are shown for the zero signal hypothesis as a function of expected background level. Notably larger fluctuations in the derived bounds are seen for the Feldman-Cousins case.



Probability of Obtaining an Inferior Flux Limit when Data Set is Doubled
0.25
0.2
0.15
0.1
0.05
0
Std Frequentist
FC
Bayesian
0
2
4
6
8
10
12
14
16
Average Background Counts in 1st Data Set

100 uniform ‘background’ events generated with values between 0-120, plus 18 ‘signal’ events with values between 30-45:

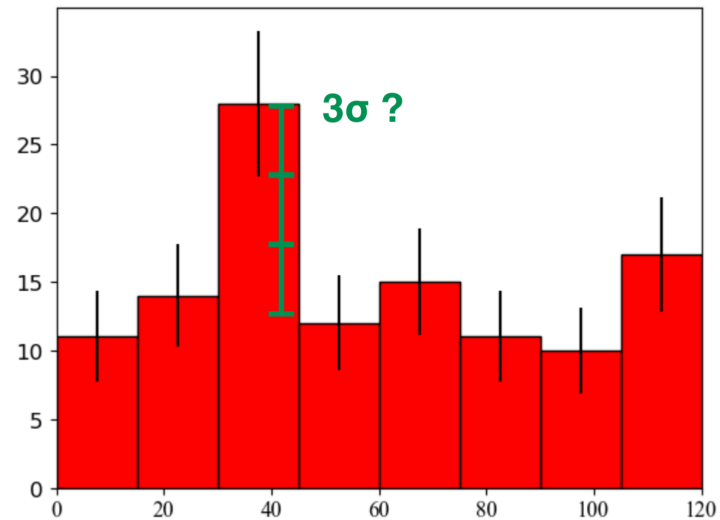


Optimal bin-size for visual inspection is comparable to the resolution and/or scale of the relevant features

Expected in Signal Bin

18 excess compared to average 12.5 background:

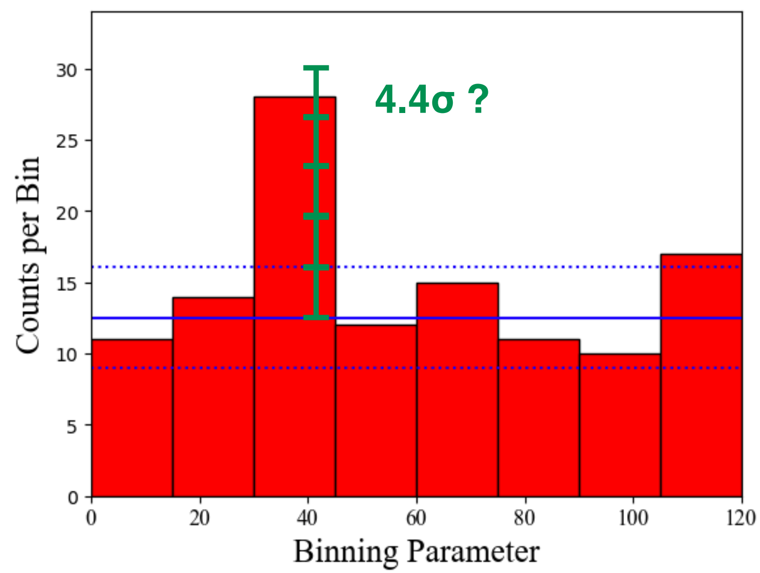
18/√12.5 ~ 5σ



Poisson rms ‘uncertainties’ are based on the **true mean**, not the fluctuated value!

So the common practise of plotting Poisson rms values on individual data points, independent of a model, is formally incorrect!

consistent with 0.6σ downward fluctuation



But is this even really correct?

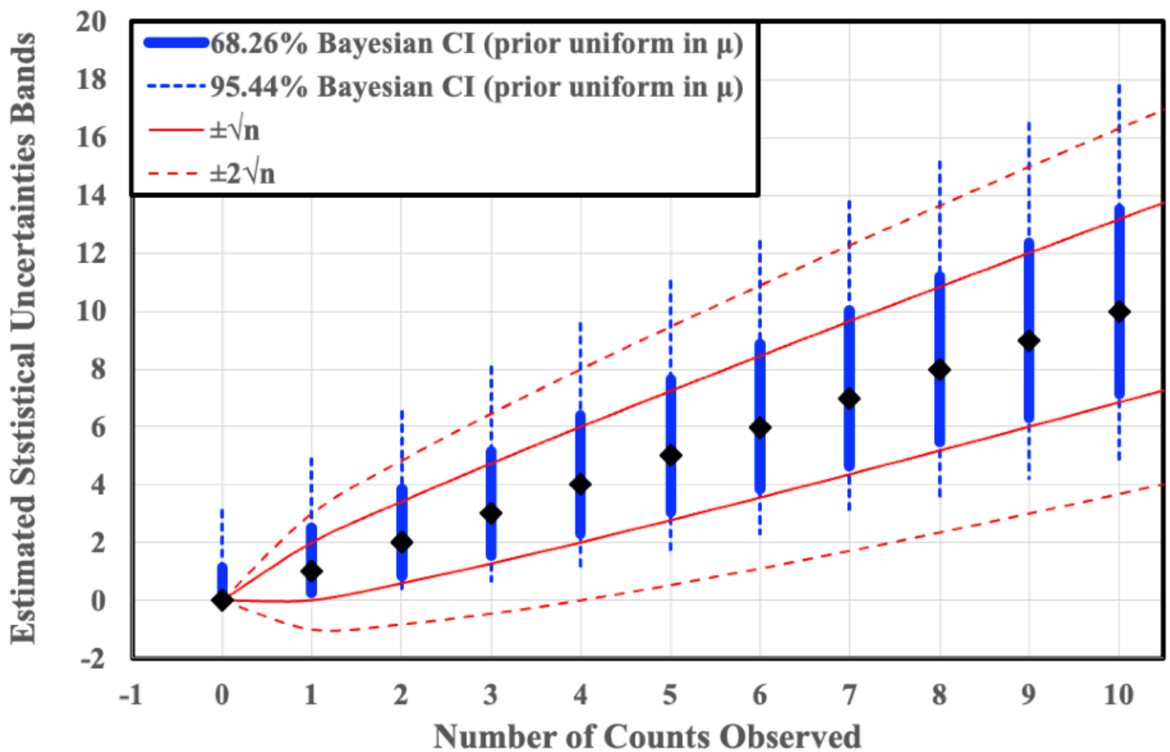
$$P_{Pois}(30\,|\,12.5) = 2 \times 10^{-5} = 4.1\sigma$$

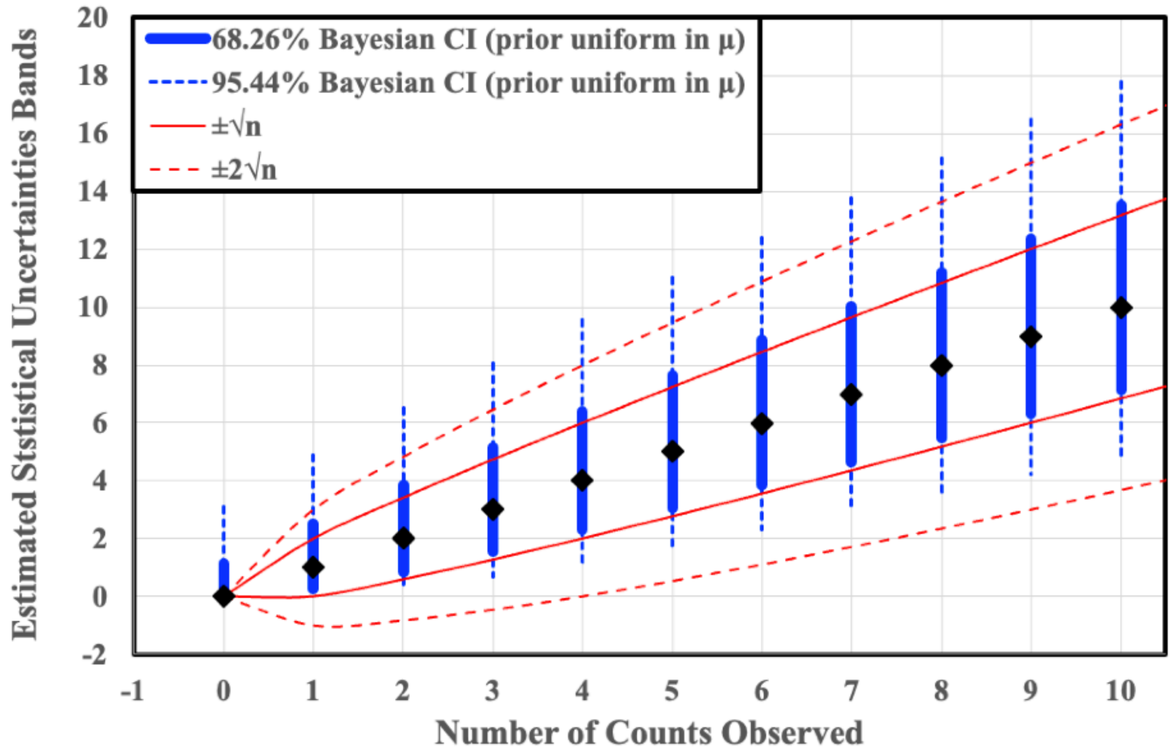
$$P_{Pois}(28\,|\,12.5) = 1.1 \times 10^{-4} = 3.7\sigma$$

(consistent with 0.4σ downward fluctuation)

What about ascribing notional statistical 'uncertainties' to data points independent of a specific model?

Tricky… let's try taking a Bayesian approach, where we assume a uniform (constant) prior for the true rate in the vicinity of the measured number of counts in a particular bin. Then take the region of most probable rates with an integral that spans 68.26% for ±1σ, 95.44% for ±2σ, etc. and compare with the common practise of using √n :



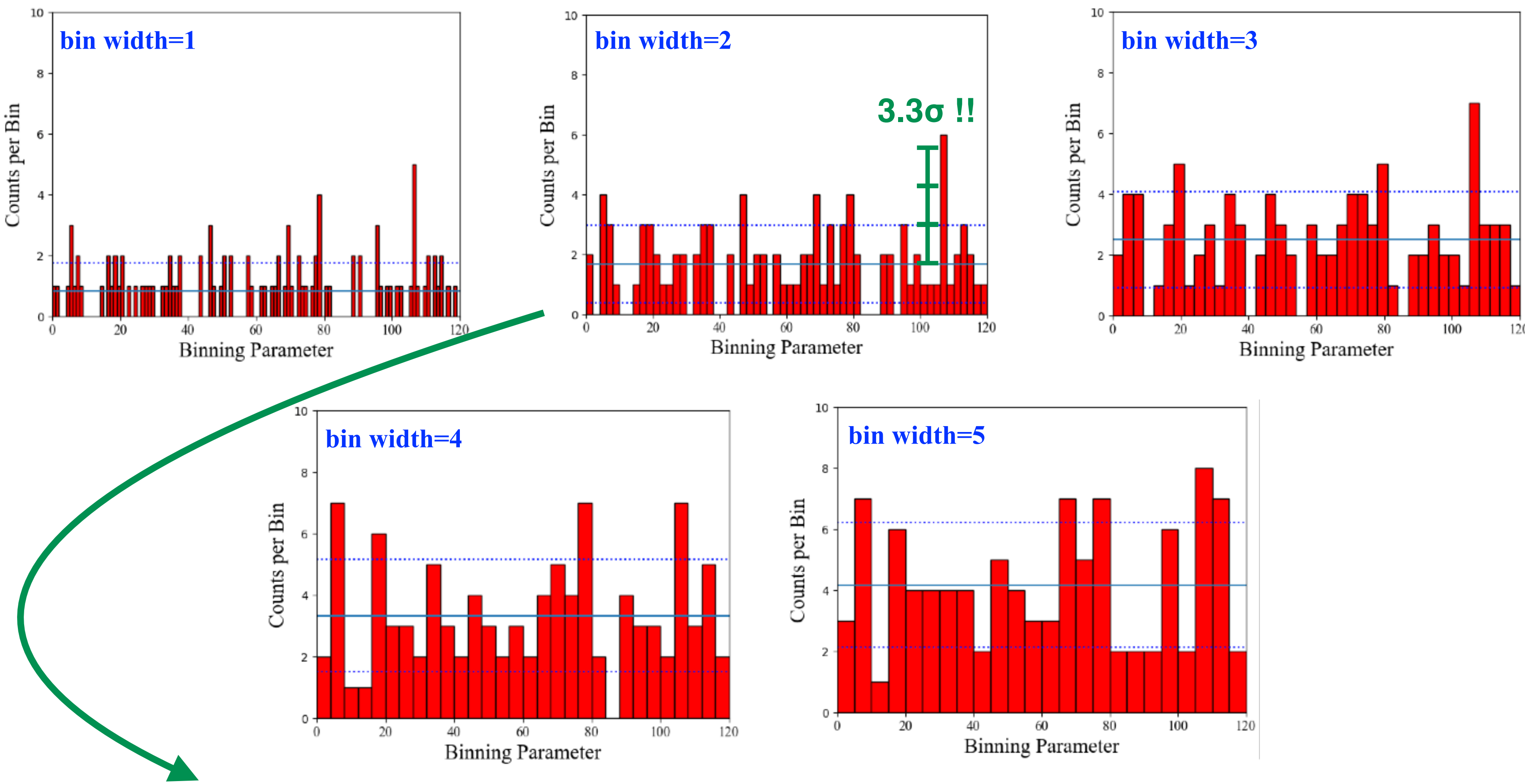


So $\pm\sqrt{n}$ isn't terrible as a way to represent approximate '1σ' Bayesian intervals for an indeterminate model, especially if you ascribe an error bar of 0-1 for zero counts, as is often done.

But at ±2σ, problems start to become more obvious, and this will become worse at higher significance levels. This can lead to biases in fit results and misinterpretations of significance.

Be careful how you use these!

Same data set without any signal (i.e. just uniform ‘background’):



Poisson probability = 0.0073 (2.44σ) rather than 0.0005 (3.3σ)

Trials: taking best of 60 bins and then the best of 5 different binnings

(binnings not entirely independent… assume effective factor of ~2.5)

$$P_{post\ trial} = 1 - (1 - 0.0073)^{(2.5 \times 60)} = 0.67$$

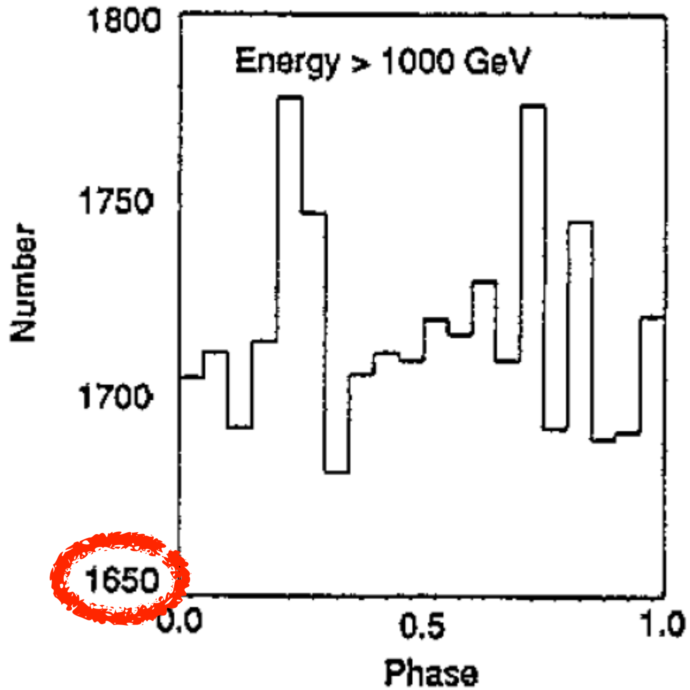
- For visual presentation/inspection of data, choose a binning based on the amount of statistics (to avoid bins with low numbers) and the anticipated scale of possible features.

- Chi-squared tests (and minimisation) using $\sqrt{n}$ errors in bins with a reasonable number of counts are generally ok: it will still get you near to the right minimum and, in the vicinity of the right model, the behaviour is typically dominated by the cumulative effect of small ($\sim 1\sigma$) fluctuations, where the approximation isn't bad. But beware of how you interpret large fluctuations, setting confidence intervals at high significance levels, or generally setting any confidence intervals when the model does not look like a good fit!

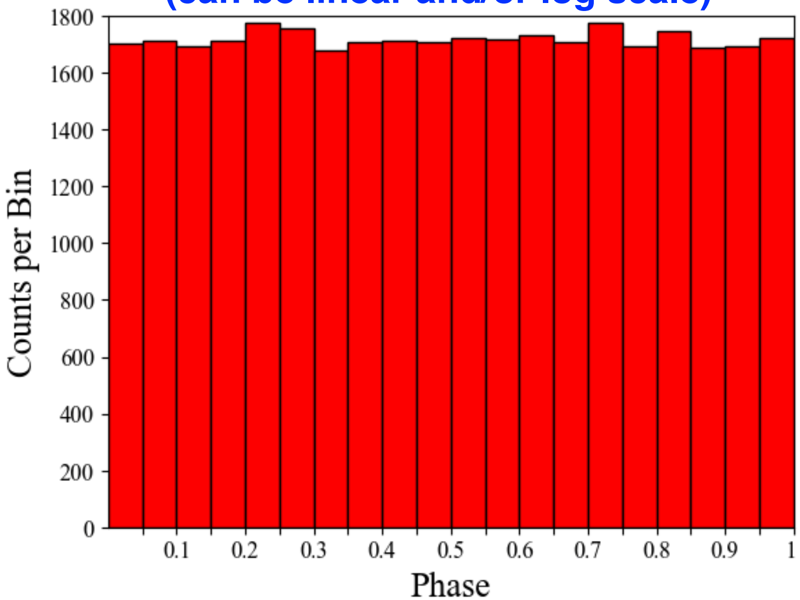
- Fitting and significance tests should be done using the correct probability distributions where appropriate.

- Whenever possible, try to use un-binned tests. Otherwise, it's advisable idea to explicitly check the dependence of your conclusions on the chosen binning.

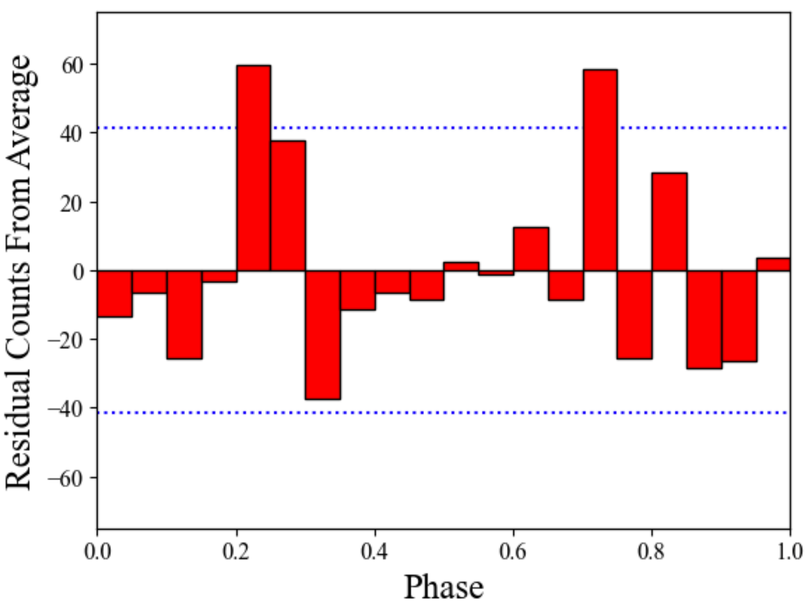
Try to avoid suppressed zeros!



Give full scale whenever possible (can be linear and/or log scale)



Show detail with residual plot



Avoid artificial smoothing (like running averages) where possible, which produce correlated error bars that are hard to interpret and can lead to false conclusions

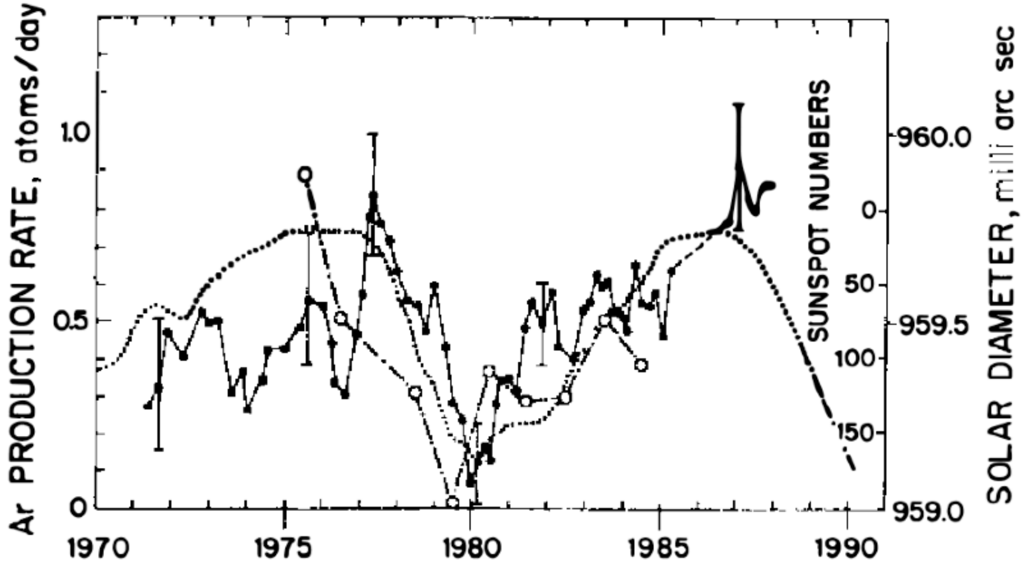


Figure 5 Plots of five-point running average of ^{37}Ar production and smoothed sunspot numbers against time in years (from 130). Solid circles, ^{37}Ar production; dotted curve, sunspot numbers; open circles, solar diameter.

Much better to use appropriate binning to keep data points uncorrelated, and use un-binned tests of significance

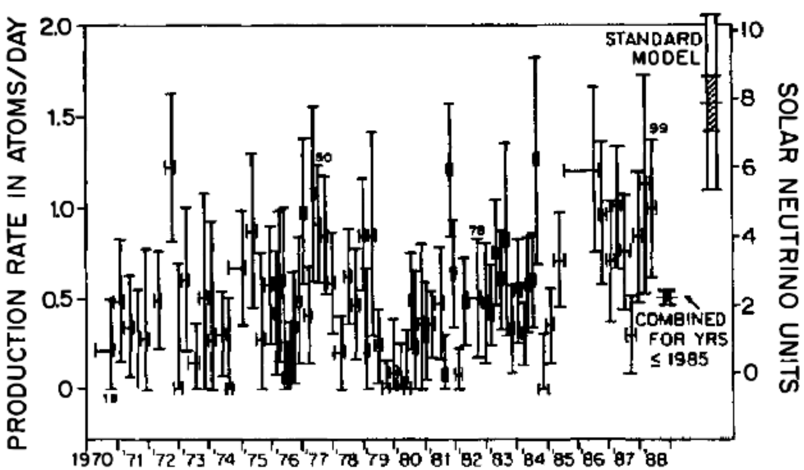
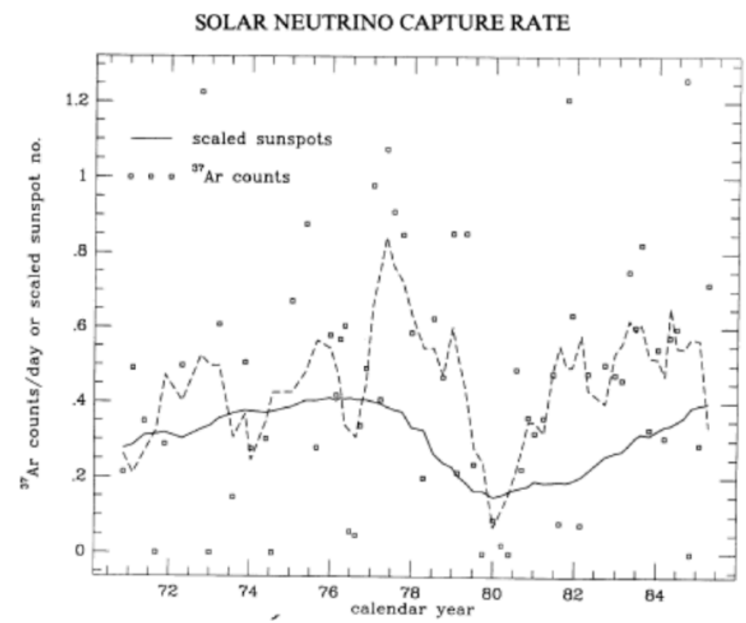


Figure 4 Production rate of ^{37}Ar in the ^{37}Cl detector as a function of time in years (from 66). The average rate in SNU for the period 1970–1985 is shown as 2.1 ± 0.3. Inclusion of the data for 1987–1988 raises the average to 2.3 ± 0.3. The prediction of the SSM from the calculation of Bahcall & Ulrich (1) is also shown, with the 3σ error quoted by them.



Ways to Display Uncertainties

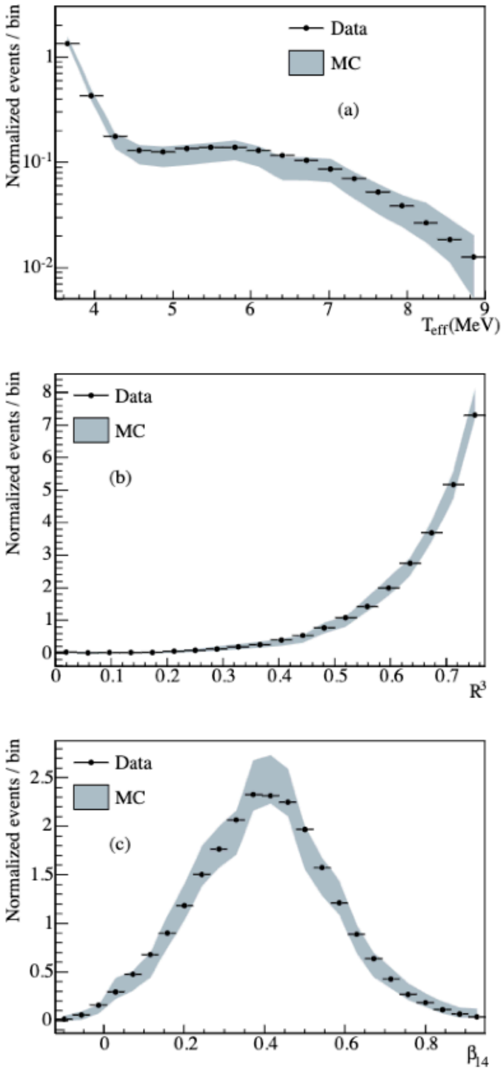


FIG. 24: (Color online) Comparison of data to simulation for ^{232}Th source runs near the AV in Phase II, in (a) $T_{\rm eff}$, (b) R^3, and (c) β_{14}. The band represents the 1σ uncertainty on the Monte Carlo-prediction, taking the quadrature sum of the statistical uncertainties with the effect of applying the dominant systematic uncertainties.

Phys.Rev.C81:055504,2010

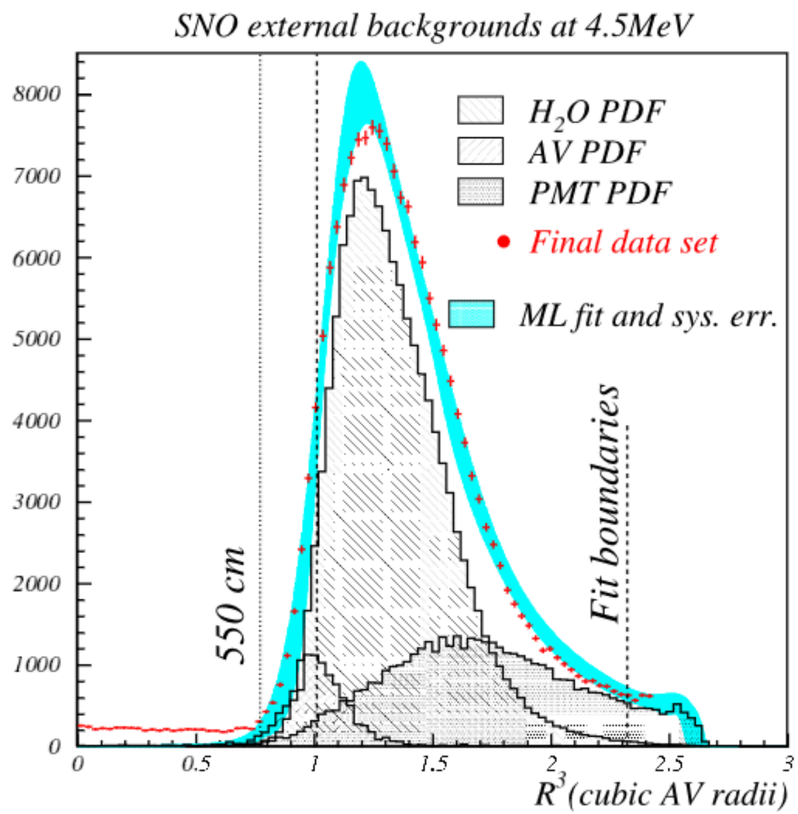


FIG. 34: Fit of of R^3 pdfs created using calibration source data to the neutrino data set, using an energy threshold of $T_{\rm eff} > 4.0$ MeV. The extended maximum likelihood method was used in the fit, and the band represents the systematic uncertainties. The y-axis is in units of Events/0.03 cubic AV radius.

Phys.Rev.C75:045502,2007

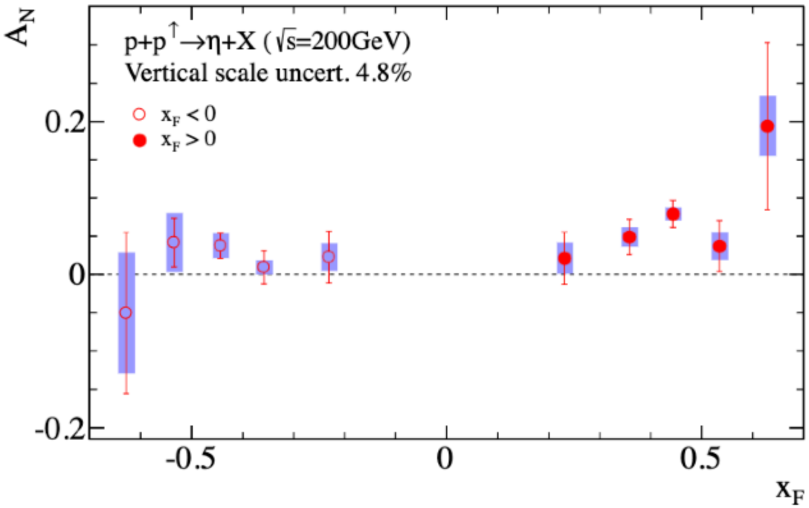


FIG. 9: (Color online) The x_F dependence of A_N. The vertical error bars show the statistical uncertainty, the blue bands represent uncorrelated systematic uncertainties (see text for details). The relative luminosity effect systematic uncertainties are not shown (see text and Table III)

Phys.Rev.D90:072008,2014

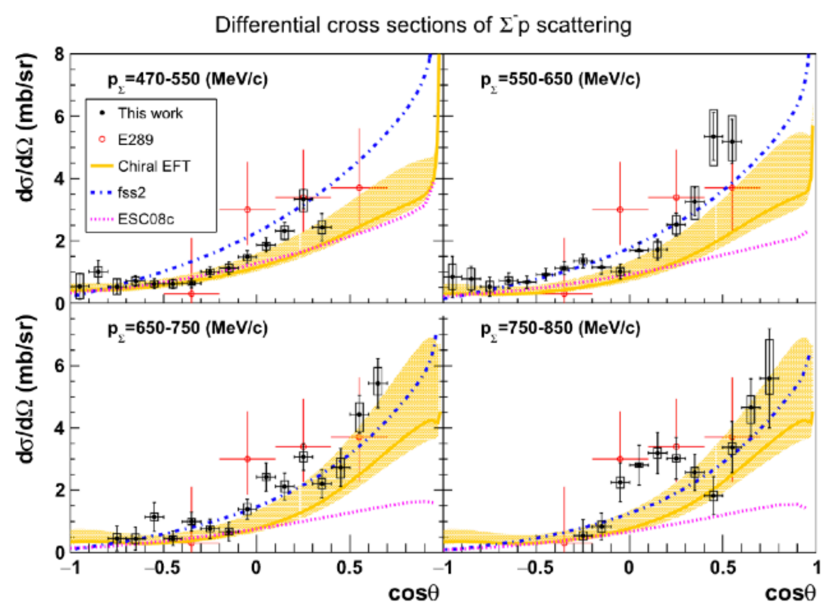
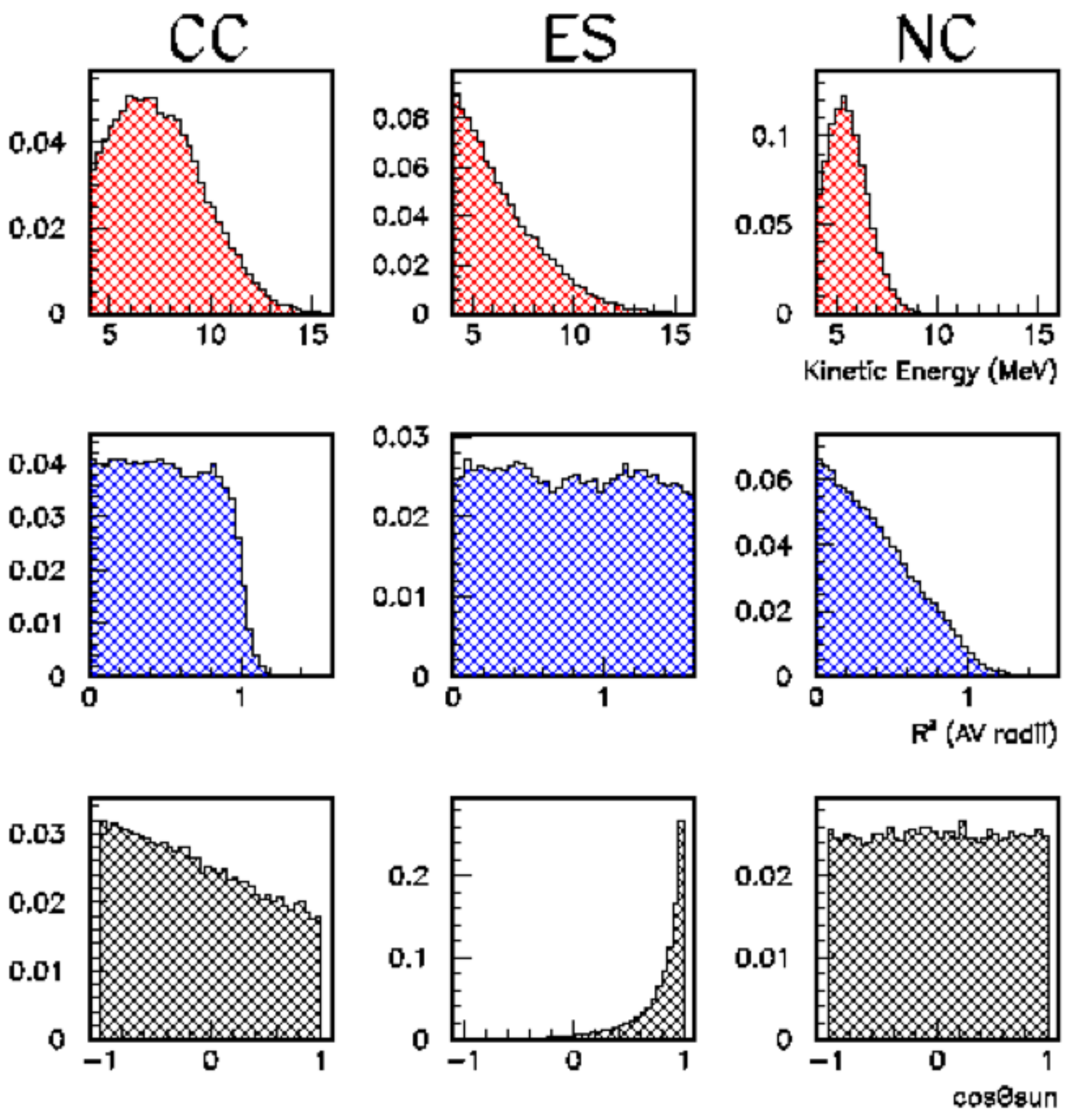


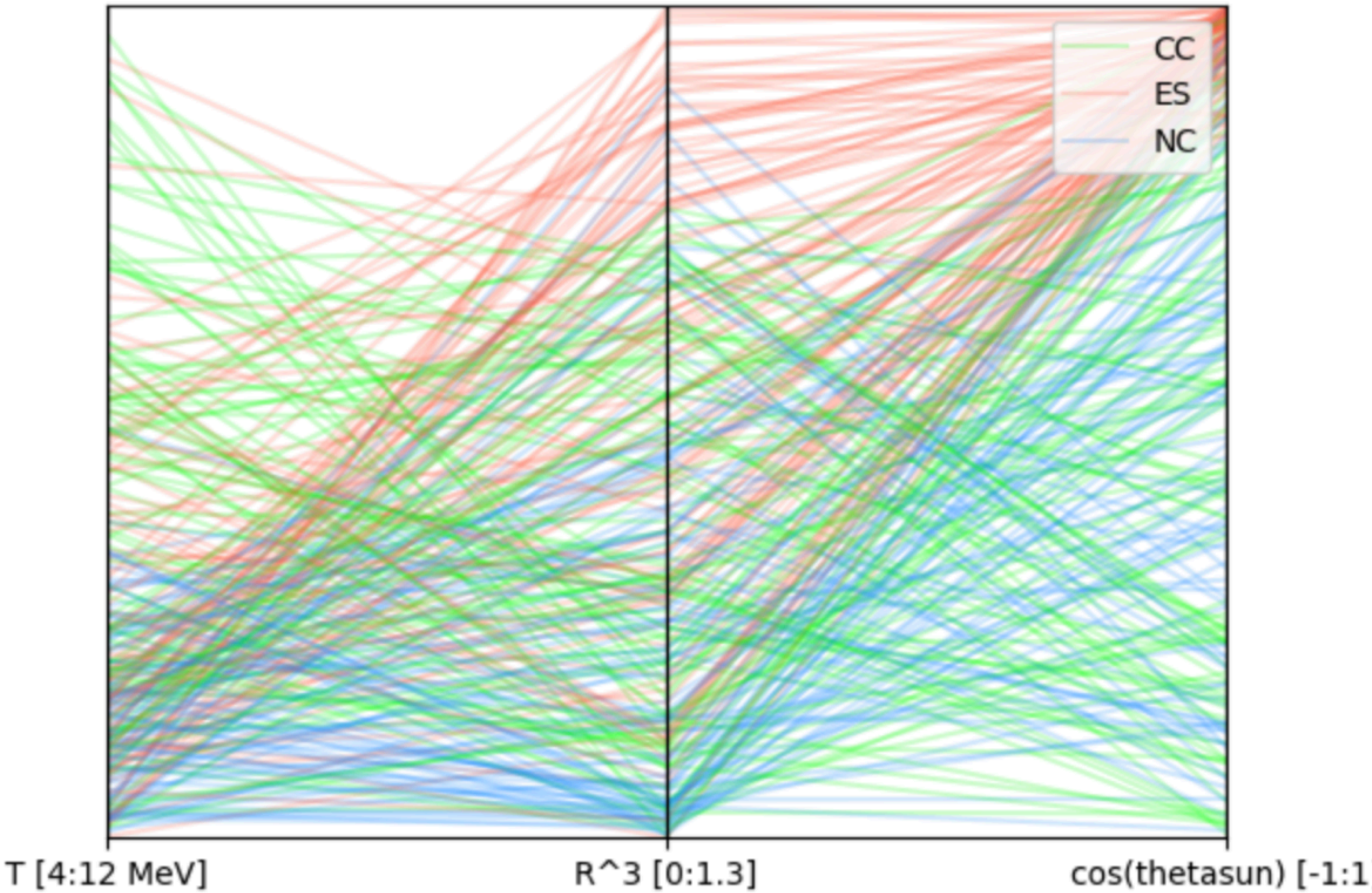
Fig. 5 Derived differential cross sections (black points). The error bars and boxes show the statistical and systematic uncertainties, respectively. Red points are averaged differential cross section of $0.4 < \mathrm{GeV}/c < 0.7$ taken in KEK-PS (the same points are plotted in the four momentum regions). The dotted (magenta), dot-dashed (blue) and solid (yellow) lines represent the Nijmegen ESC08 based on boson-exchange picture, fss2 based on QCM and the extended chiral effective field theory (χEFT), respectively.

Phys. Rev. C 104, 045204 (2021)

Visualising Multi-Dimensional Data



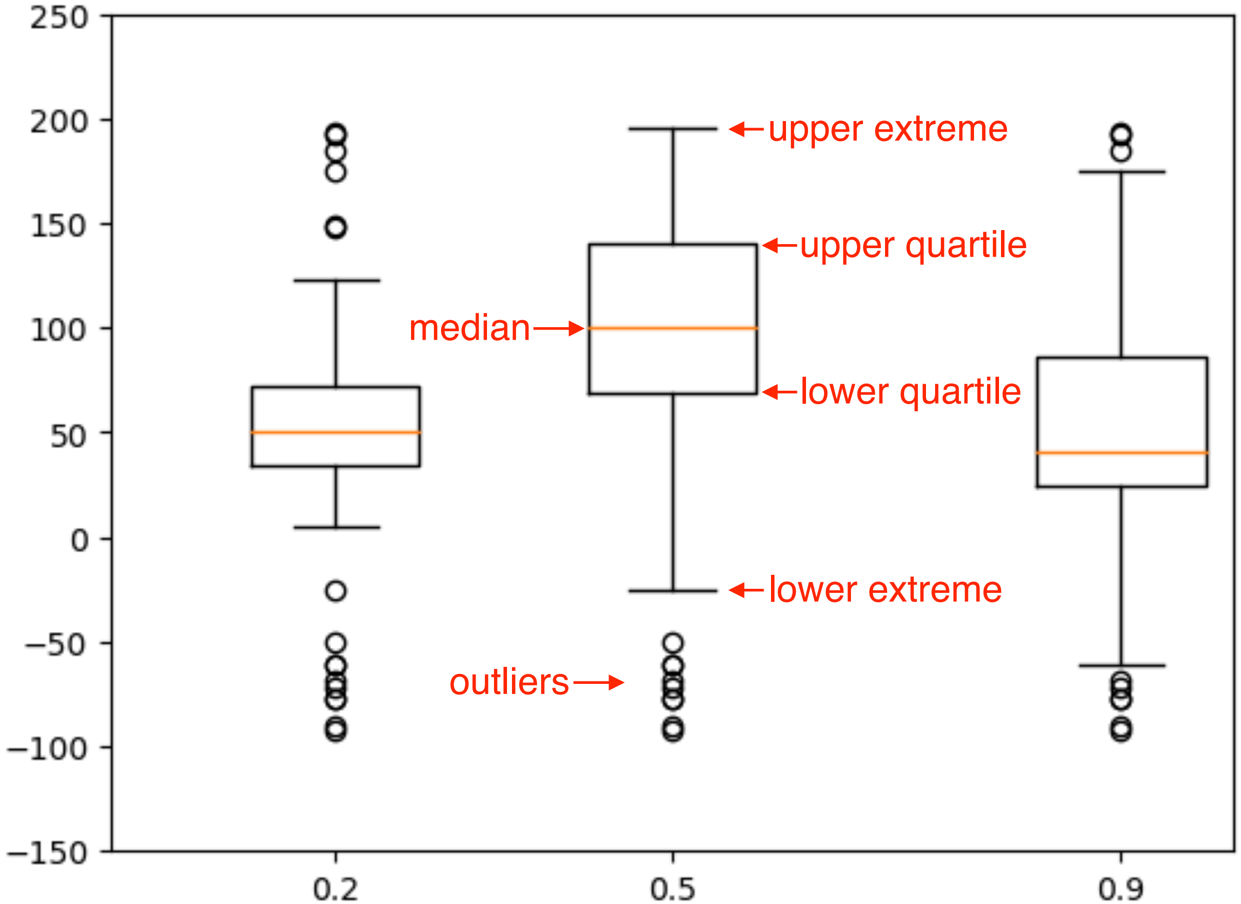
Parallel Coordinate Plot:




```
parallel_coordinates(df_new[['Signal', 'T [4:12 MeV]', 'R^3 [0:1.3]', 'cos(thetasun) [-1:1]']],
                     "Signal", color=["lime", "tomato","dodgerblue"], alpha=0.2)
```

Ways to display information about data point distributions when you're not simply dominated by Poisson statistics:

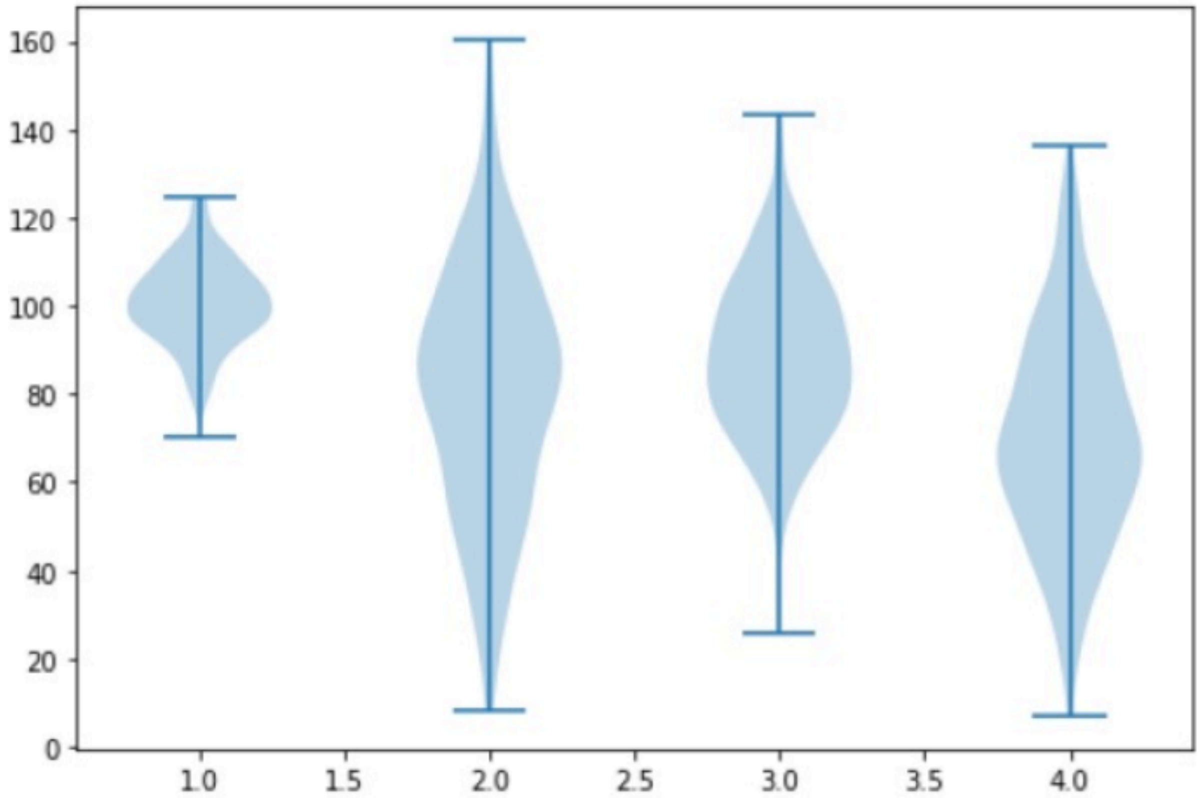
Box and Whisker:




```
data = [data1,data2,data3]
pos = [0.2,0.5,0.9]
plt.xlim([0,1])
plt.ylim([-150,250])
ax1.boxplot(data,positions=pos)
```

Ways to display information about data point distributions when you're not simply dominated by Poisson statistics:

Violin Plot:




```
data_to_plot = [collectn_1, collectn_2, collectn_3, collectn_4]

# Create the boxplot
bp = ax.violinplot(data_to_plot)
plt.show()
```