Lecture 8:

Hypothesis Testing & Data Presentation

- Hypothesis Selection & Rejection
- Bayesian Information Criterion
- 'Binsmanship' and Dodgy Deviations
- The Meaning of Error Bars (and what to use)
- More Things to Avoid
- Displaying Uncertainties & Multi-Dimensional Data
- Boxes, Whiskers and Violins

Hypothesis Selection

Frequentist:

In general, define the two hypotheses to be compared: H_1 and H_2.
Also define an appropriate statistical test from considerations of:

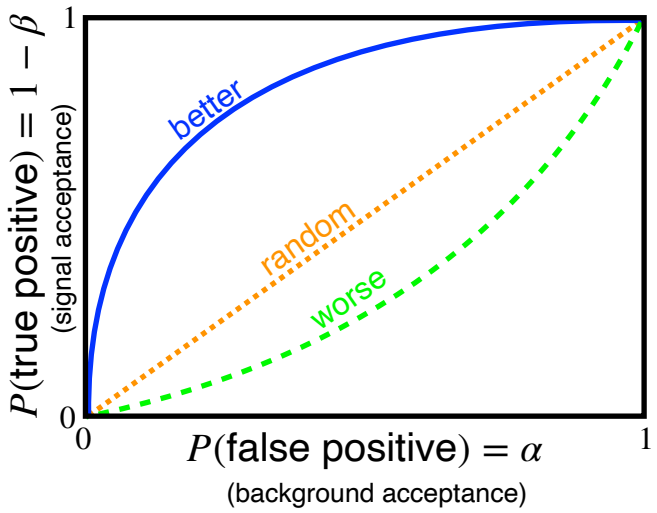
$\alpha \equiv$ probability to incorrectly reject H_1 ("Type I error") → "false positive"
(essentially the p-value)

$\beta \equiv$ probability to incorrectly accept H_2 ("Type II error") → "false negative"
(only makes sense if you know the distribution of an alternative hypothesis)

Define "critical" values for the test statistic in advance, and choose to "accept" or "reject" a particular hypothesis based on the outcome.

If H1 is the background hypothesis and H2 is the signal hypothesis, then the standard Receiver Operating Characteristic (ROC) curve can be drawn as follows:

This can be useful in comparing the performances of test statistics and optimising selection cuts (*e.g.* maximising signal/√background etc.)



Hypothesis Selection

Frequentist:

We've already come across the Neyman-Pearson Lemma, which states that the maximum likelihood ratio provides the uniformly most powerful method to distinguish between simple hypotheses.

But we have the perennial issue with frequentist statistics:

Wikipedia:
"The p-value does not provide the probability that either the null hypothesis or its opposite is correct"

Hypothesis Selection

Bayesian:

Explicitly, the relative belief ("betting odds") in H2 relative to H1 given the data

$$\text{Posterior Odds Ratio} = \frac{P(H_2 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_2)\ P(H_2)}{P(D \mid H_1)\ P(H_1)}$$

Let the indices j & k represent different sets of possible nuisance parameter values within each model. Then, in discreet form:

$$= \frac{\left[\sum_j P(D \mid H_2^{(j)}) P(H_2^{(j)} \mid P(H_2))\right]}{\left[\sum_k P(D \mid H_1^{(k)}) P(H_1^{(k)} \mid P(H_1))\right]} \times \frac{P(H_2)}{P(H_1)}$$

Ratio of marginalised likelihoods, each integrated (in a Bayesian way) over their nuisance parameters (rather than maximised)

Sometimes also called the ratio of "model evidences"

$$\equiv \text{Bayes Factor (BF)} \times \text{Prior Odds Ratio}$$

The heart of Bayesian hypothesis selection is the **posterior odds ratio**, but many statisticians like to separately quote the BF to indicate what the data "prefers" in isolation from the ratio of priors. Quoting both explicitly shows the impact of the priors and, hence, indicates the strength of the data itself.

But caution must be exercised in interpreting the BF on its own, as previously discussed for frequentist statistics: ONLY the posterior probability can give you the betting odds for a particular hypothesis!

Some have advocated using BF as a replacement for p-values* in null hypothesis testing. Amongst the advantages often quoted is that it explicitly weighs the likelihood of alternative hypotheses against that of the null hypothesis, as well as being independent of hypothesis priors. But (as always) there are potential issues here…

*For example, Williams, Baath & Philipp, Research in Human Development, 14:4, 321-337 (2017)

Say we observe some number of events, *n*, from a counting experiment where the expected background is *b* and we wish to test the null hypothesis of zero signal against a possible non-zero signal. We will treat the exact number of signal as a nuisance parameter, allowing it to be anything up to some s_{max}:

density of s values

$$BF = \frac{\int_0^{s_{max}} P(n \mid \mu = s + b)\rho(s)\, ds}{P(n \mid \mu = b)}$$

$$= \frac{\frac{1}{s_{max}} \int_0^{s_{max}} P(n \mid \mu = s + b)\, ds}{P(n \mid \mu = b)}$$

$$\int_0^{s_{max}} \rho(s) ds = 1$$

Let's assume $\rho(s)$ is constant (note: same as a prior for s !)

$$\longrightarrow \rho(s) = \frac{1}{s_{max}}$$

The answer is now highly dependent on the prior for the signal hypothesis!

This is an example of a more general behaviour known as "Lindley's Paradox", which comes about when comparing "point-like" to "diffuse" hypotheses

Reverting to the posterior odds ratio by restoring the ratio of hypothesis priors doesn't necessarily fix things if you assign equal prior probabilities to both the null and non-null hypotheses (PoOR = BF x 1).

On the other hand, if you ascribe equal probabilities to **all** signal hypotheses, including zero, then the term cancels!

$$\text{PoOR} = \text{BF} \times \frac{1}{1/s_{max}}$$

So, as always, it comes down to exactly what you mean when you say that you have "no preference." You are still dependent on this prior choice.

The problem is partly due to the fact that you can alway have an extremely small signal that is non-zero, but indistinguishable from zero for all practical purposes. It therefore becomes essentially impossible to ever rule out a non-zero signal and, thus, prefer the null hypothesis without relying on a prior to focus attention (to a greater or lesser extent) on a region where the signal would be measurable.

An different approach for null hypothesis testing when the alternative is a diffuse set of possibilities is to instead simply concentrate on the region in which H0 is no longer viable. Sounds like p-values again... but we can also consider a Bayesian equivalent too…

Hypothesis Rejection

Frequentist:

In the approach originally proposed by Fisher, the p-value is defined as the chance probability for a fluctuation in an ensemble of possible data sets that is at least as extreme as that observed assuming the null hypothesis. In discreet form, the confidence interval that would exclude the observation is given by:

$$CL_{ex} = \frac{\sum_i \left[P(D_i|H_0) > P(D_{obs}|H_0)\right]}{\sum_i \left[P(D_i|H_0)\right]}$$

summation/integration over the ensemble of **data sets** that are more likely than what has been observed.

and

$$p\text{-value} \equiv 1 - CL_{ex}$$

We have discussed the pit-falls of p-value interpretation before:

Fisher: *"Report the exact level of significance… and do not talk about accepting or rejecting hypotheses." (!!!)*

Neyman-Pearson: *"If the data falls into the rejection region of H1, accept H2; otherwise accept H1. Accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true." (???)*

(Wikipedia)

Hypothesis Rejection

Bayesian:

The parallel Bayesian construction (in discreet form) would then be as follows:

$$CI_{ex} = \frac{\sum_i \left[P(H_i \mid D_{obs}) > P(H_0 \mid D_{obs})\right]}{\sum_i \left[P(H_i \mid D_{obs})\right]}$$

summation/integration over the ensemble of **hypotheses** more likely than H0 given what has been observed.

and

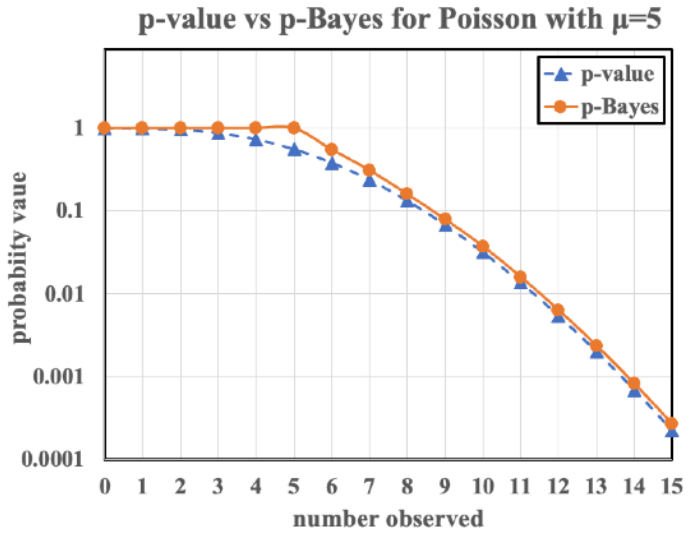
$$p\text{-Bayes} \equiv 1 - CI_{ex}$$

While priors are still in play for this Bayesian parameter, this follows a much better parallel construction to frequentist p-values than the Bayes Factor and avoids many of the previously discussed issues with null-hypothesis testing

Comparisons between p-values and p-Bayes (constant prior) for Poisson and Gaussian excesses:

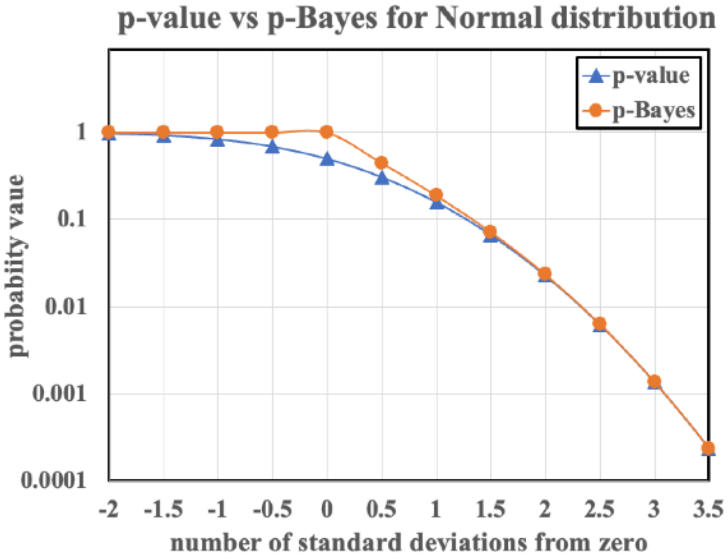
Example python script using a simple adaptive 1-D grid search to find p-Bayes (constant prior) for Poisson process:

```
import math
from scipy.stats import poisson
B = float(input(" Expected Background ="))
n = int(input(" Observed ="))
def pBayes(n,B):
    p0=poisson.pmf(k=n,mu=B)
    i=max(n-B+1,1)
    while(poisson.pmf(k=n,mu=B+i)>p0): i=i+1
    i=i-1
    j=0
    while(poisson.pmf(k=n,mu=B+i+j/10)>p0): j=j+1
    k=0
    while(poisson.pmf(k=n,mu=B+i+j/10+k/100)>p0): k=k+1
    x=i+j/10+k/100
    return poisson.cdf(n,B+x)/poisson.cdf(n,B)
pB=pBayes(n,B)
print("  p-Bayes=",pB)
```



In contrast with p-values, p-Bayes always returns 1 if there is no excess…because you cannot exclude the null hypothesis with **any** credibility in this case!

But, for the constant prior case, the two values quickly converge in the high tail, where judgements are usually made. So the pragmatic and cautious use of p-values is not a bad approximation to p-Bayes for this particular choice of prior



So, after all the fuss and discussion, are p-values ok to use after all??

In fact, it's the same pragmatic conclusion we came to earlier -

Look carefully at context:

Very small p-values, even after careful accounting of trials, confirmed by other observations/crosschecks, which could be explained by self-consistent (plausible) alternatives…

p-values remain one of the central tools for hypothesis testing - it should just be treated carefully!

Effectively a 'loose' Bayesian comparison with other hypotheses!

Reject H0

Can be useful when ambiguities arise in BF and PoOR from how weighting is applied to diffuse hypotheses (*e.g.* is there a non-zero signal *of some kind*?)

It's interesting to see that a Bayesian parallel, based on exclusion rather than acceptance, can also be formulated, which can share similar properties for certain choices of prior!

How do you decide between models that have different numbers of free parameters? Clearly, the more parameters you have, the easier it will be to fit the data… but does this make it a better model??

Let's start by defining something related to the ability of a data set to constrain model parameters (*i.e.* the *information content* of the data) and then consider the overall probability of a hypothesis that is integrated over all possible values of its parameters…

Bayesian Information Criterion (BIC)

$$P(D\,|\,H) = \int \mathscr{L}(D\,|\,\mathbf{q}, H)\; p(\mathbf{q}\,|\,H)\; d\mathbf{q}$$

Recall our expansion around the maximum likelihood point:

$$\log\left[\mathscr{L}(D\,|\,\mathbf{q}, H)\right] \simeq \left[\log \mathscr{L}\right]_{\mathbf{q}=\hat{\mathbf{q}}} + (\mathbf{q}-\hat{\mathbf{q}})\underbrace{\left[\frac{\partial \log \mathscr{L}}{\partial \mathbf{q}}\right]_{\mathbf{q}=\hat{\mathbf{q}}}}_{0} + \frac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})\underbrace{\left[\frac{\partial^2 \log \mathscr{L}}{\partial \mathbf{q}^2}\right]_{\mathbf{q}=\hat{\mathbf{q}}}}_{-\mathscr{I}(\mathbf{q})}(\mathbf{q}-\hat{\mathbf{q}})^T + \ldots$$

$$P(D\,|\,H) \simeq \hat{\mathscr{L}} \int \exp\left(-\frac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})\mathscr{I}(\mathbf{q})(\mathbf{q}-\hat{\mathbf{q}})^T\right) p(\mathbf{q}\,|\,H)\; d\mathbf{q}$$

assume independent identically distributed (IID) data and $\mathscr{I}_i(\mathbf{q}) \sim \mathscr{I}_i(\hat{\mathbf{q}})$ and $p(\mathbf{q}\,|\,H) \sim p(\hat{\mathbf{q}}\,|\,H)$ near the maximum

$$= \hat{\mathscr{L}} \int \exp\left(-\frac{1}{2}(\mathbf{q}-\hat{\mathbf{q}})\left[\sum_{i=1}^{n} \mathscr{I}_i(\mathbf{q})\right](\mathbf{q}-\hat{\mathbf{q}})^T\right)\; p(\mathbf{q}\,|\,H)\; d\mathbf{q}$$

$$\sim \hat{\mathscr{L}} \int \exp\left(-\frac{n}{2}(\mathbf{q}-\hat{\mathbf{q}})\mathscr{I}_x(\hat{\mathbf{q}})(\mathbf{q}-\hat{\mathbf{q}})^T\right) p(\hat{\mathbf{q}}\,|\,H)\; d\mathbf{q} \quad = \hat{\mathscr{L}}\left(\frac{2\pi}{n}\right)^{k/2} \frac{p(\hat{\mathbf{q}}\,|\,H)}{\sqrt{\mathscr{I}_x(\hat{\mathbf{q}})}}$$

k-dimensional Gaussian

$$-2\log P(D\,|\,H) \simeq -2\left[\underbrace{\log \hat{\mathscr{L}} - \frac{k}{2}\log n}_{\text{these terms dominate as } n \to \infty} + \log p(\hat{\mathbf{q}}\,|\,H) - \frac{1}{2}\log \mathscr{I}_x(\hat{\mathbf{q}}) + \frac{k}{2}\log 2\pi\right]$$

$$\sim \; -2\log \hat{\mathscr{L}} + k\log n \equiv BIC$$

(sort of a modified chi-squared to account for the degrees of freedom)

$$BIC \equiv -2\log\hat{\mathscr{L}} + k\log n$$ (to be minimised)

A slightly different criteria derived by Akaike* is based on the "Kullback-Leibler Divergence," which is a measure of separation between probability distributions, with the result:

$$AIC \equiv -2\log\hat{\mathscr{L}} + 2k$$

These penalise models with more free parameters (BIC more than AIC)

This is effectively accounting for the use of additional degrees of freedom (as we do with χ^2), which tends to allow a better fit to the data, but doesn't necessarily indicate a correct model.

This is all in line with Occam's Razor, which does **NOT** say that the more complicated model is incorrect, but merely that simpler models tend to be good starting places. Also, while BIC and AIC can provide a sort of relative goodness of fit between different hypotheses, they do not give an absolute goodness of fit.

So, these criterion should be used with caution, but can provide good guidance on model section.

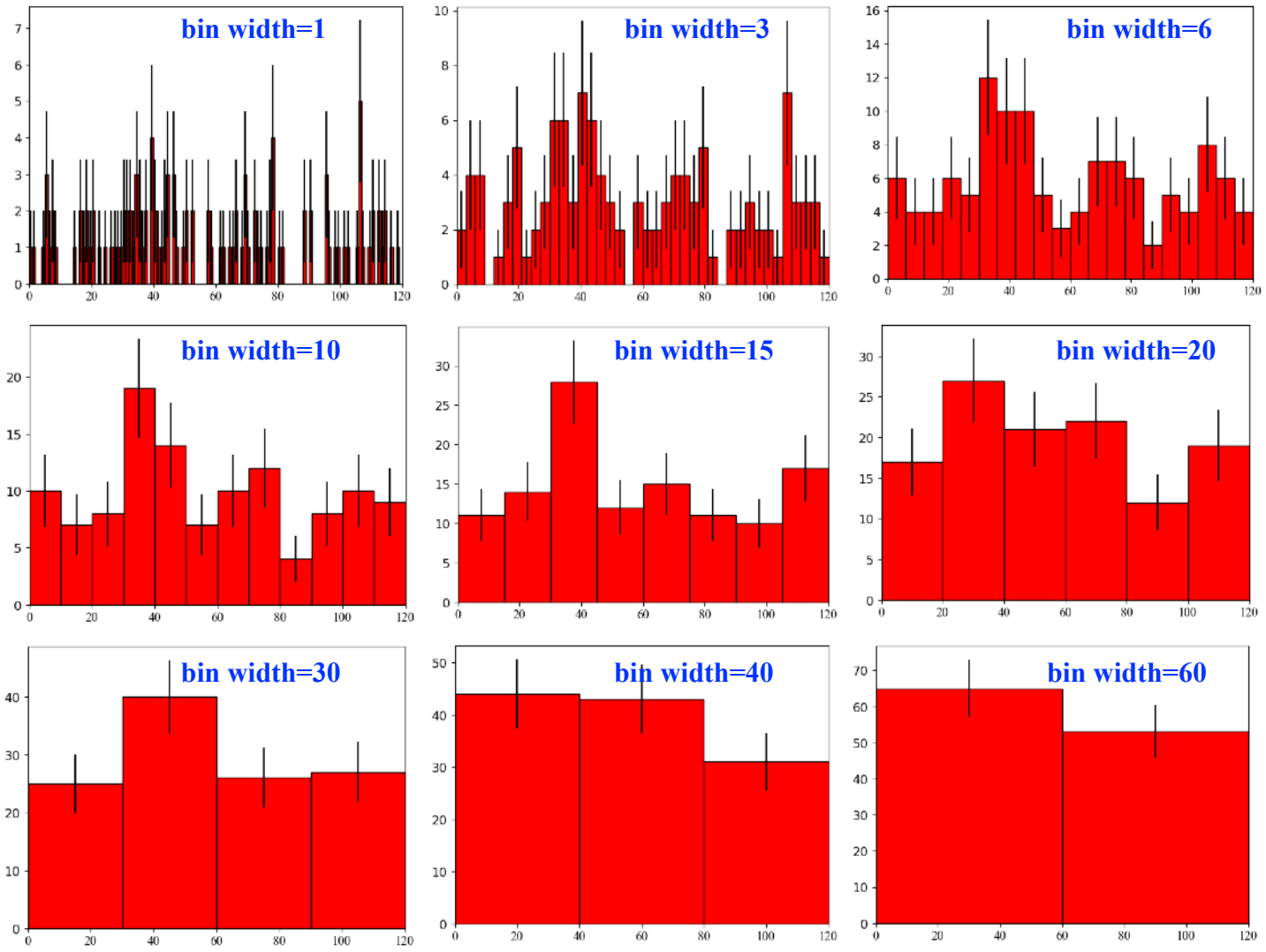
*Akaike, H., "A new look at the statistical model identification", IEEE Transactions on Automatic Control, **19** (6): 716–723 (1974)

WE NEED TO TALK ABOUT ERROR BARS

(and Data Presentation)

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

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



using $\pm\sqrt{n}$ error bars

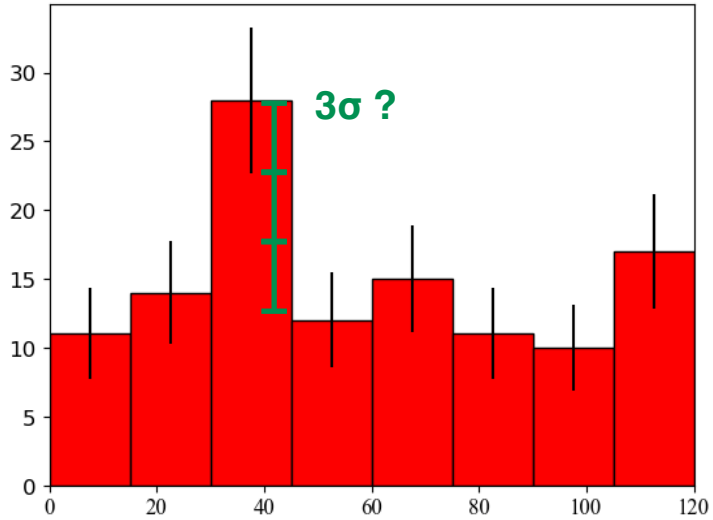
Optimal bin-size for visual inspection is comparable to the resolution and/or scale of the relevant features

Expected in Signal Bin

18 excess compared to average 12.5 background:

$18/\sqrt{12.5} \sim 5\sigma$

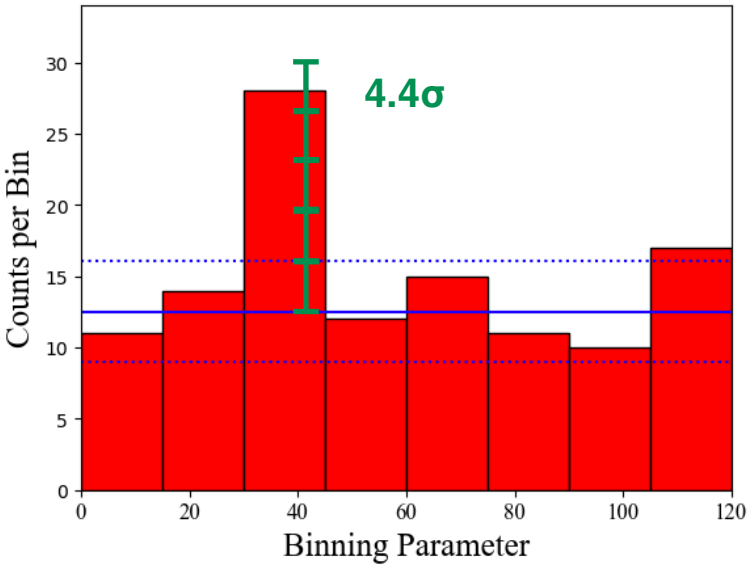
(p-value = 2.9×10^{-7})



Poisson rms 'deviations are based on the **true mean**, not the fluctuated value!

So the common practise of plotting Poisson rms values on individual data points to compare with a model is not formally correct!

consistent with 0.6σ downward fluctuation



But is this even really correct?

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

(~18+12.5)

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

(consistent with 0.4σ downward fluctuation)

What are these error bars supposed to represent?

There is no uncertainty in what was measured:
this IS what was observed!

If you are trying to judge the consistency of the observation with a given model, then you need to look at the observation in light of the **probability distribution predicted for that model**, for which the error bars are an **approximate representation**. So use model-based error bars and do the appropriate statistical test.

What if you want to make some more general statement that relates the observation to the range of possible models? The error bar then represents some bin-by-bin confidence or credibility interval…

Let's start with a Frequentist approach:

Say we're interested in an upper bound. For a given observed number of counts, n, in a given bin, we then want to find the range of possible model means for which the observed number of counts or less would occur with some frequency at least as great as 1-α (Neyman construction).

(so, for a 90% CL, you want values of the mean for which the chance of this occurring is better than 10%)

For an upper bound, the maximum mean value would thus have to satisfy:

$$\sum_{m=0}^{n} \frac{\mu_{max}^{m} e^{-\mu_{max}}}{m!} = 1 - \alpha$$

$$\rightarrow \quad \int_{0}^{\mu_{max}} \frac{\mu^{n} e^{-\mu}}{n!} d\mu = \alpha$$

(can be shown from doing repeated integration by parts)

Which is also identical to a Bayesian upper bound, assuming a prior that is constant with μ !

Recall that, for a χ^2 distribution with k degrees of freedom:

$$P(<\chi^2, k) = \int_{0}^{\frac{\chi^2}{2}} \frac{x^{\frac{k}{2}-1} e^{-x}}{(\frac{k}{2}-1)!} dx$$

(for even k)

So, with a change of variables, we can do the integration using standard tools for χ^2 !

We can then go through a similar process for a lower bound and, thus, we can define a CL (or CI) interval*:

$$\mu_{min} = \frac{1}{2} F_{\chi^2}^{-1}(\alpha_{min}; 2n) \qquad \mu_{max} = \frac{1}{2} F_{\chi^2}^{-1}(1 - \alpha_{max}; 2(n+1))$$

Where $F^{-1}_{\chi^2}$ is the inverse of the cumulative χ^2 distribution and we just need to specify the fractions of the distribution we want above (α_{max}) and below (α_{min}) the observed number. One common protocol** is to define a central confidence interval: $\alpha_{min} = \alpha_{max} = \alpha/2$

HOWEVER, this formalism runs into difficulties in trying to define symmetric, continuous confidence regions for asymmetric, quantised distributions like Poisson!

A trivial example:

The above formalism gives a 1 sigma (68.27% CL) region of 0 - 1.84 for an observation of zero counts. But the Poisson integral for that case is:

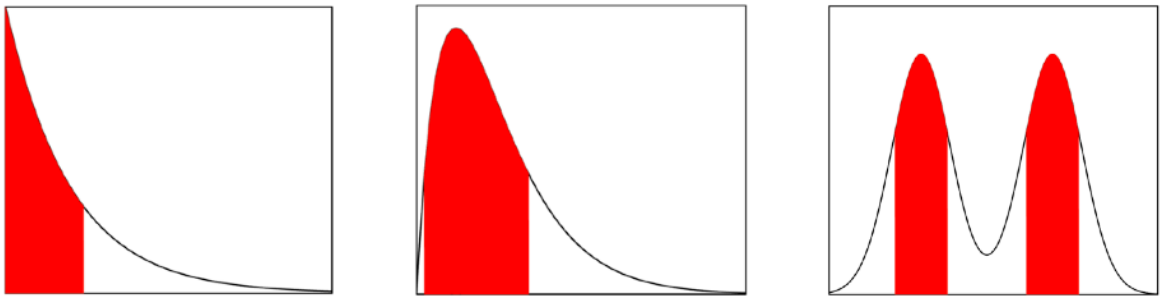
$$1 - e^{-1.84} = 0.84$$

because the 'half' of the interval below zero doesn't exist!

*K. Nakamura et al. (Particle Data Group), J. Phys. G 37, 075021 (2010); pdg.lbl.gov
**Garwood, F. (1936), Biometrika, 28, 437-442; ATLAS Statistics Forum, 15 February, 2011

ROOT error bar option *kPoisson*

An alternative method of interval construction is to use an ordering rule based on the highest probability densities, which more naturally handles asymmetric and multi-peaked distributions:



For single-peaked distributions, such as Poisson, this is equivalent to finding the shortest interval*, which can be numerically determined.

For 1σ bounds on a Poisson distribution, a reasonable parameterisation (mine) is as follows:

$$\mu_{min} \simeq n - \sqrt{n} + 0.33\left[1 - \exp\left(-1.5n^{1/4}\right)\right]$$

$$\mu_{max} \simeq n + \sqrt{n} + 0.34 + 0.81\exp\left(-1.7n^{1/4}\right)$$

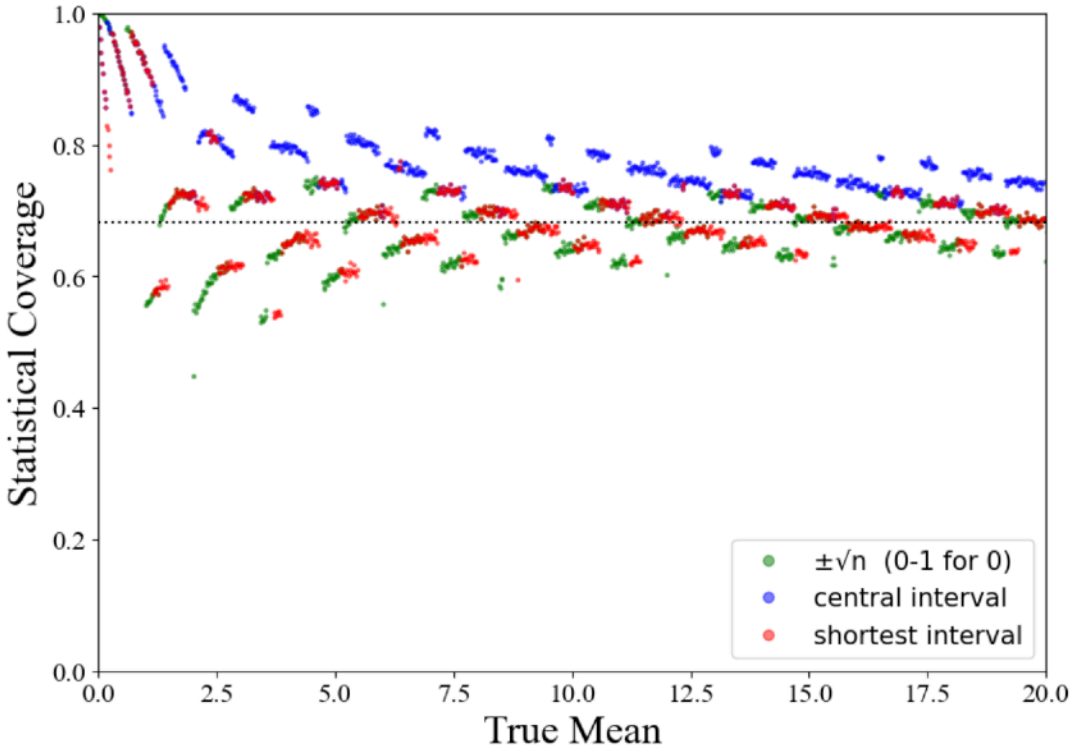
*Casella, G., Robert, C. (1989), The Canadian Journal of Statistics, 17, 45-57; Kabaila, P., Byrne, J. (2000), The Canadian Journal of Statistics, 28, 1-9

The following table compares the "1σ" interval definitions and their (Bayesian) integrals as a function of observed n for central, shortest and also just ±√n, with the latter set to 0-1 for zero counts:

	central interval			shortest interval			±sqrt(n) (with 0–1 for 0)		
n	lower	upper	integral	lower	upper	integral	lower	upper	integral
0	0.000	1.841	0.841	0.000	1.147	0.683	0.000	1.000	0.632
1	0.173	3.300	0.828	0.268	2.501	0.683	0.000	2.000	0.594
2	0.708	4.638	0.806	0.676	3.697	0.683	0.586	3.414	0.641
3	1.367	5.918	0.791	1.479	5.078	0.683	1.268	4.732	0.655
4	2.086	7.163	0.781	2.287	6.400	0.683	2.000	6.000	0.662
5	2.840	8.383	0.773	3.057	7.630	0.683	2.764	7.236	0.666
6	3.620	9.584	0.766	3.847	8.837	0.683	3.551	8.449	0.669
7	4.418	10.770	0.761	4.652	10.029	0.683	4.354	9.646	0.671
8	5.232	11.945	0.757	5.472	11.208	0.683	5.172	10.828	0.673
9	6.056	13.110	0.753	6.302	12.377	0.683	6.000	12.000	0.674
10	6.891	14.267	0.750	7.141	13.536	0.683	6.838	13.162	0.675
11	7.734	15.417	0.748	7.988	14.689	0.683	7.683	14.317	0.675
12	8.585	16.560	0.745	8.842	15.834	0.683	8.536	15.464	0.676
13	9.441	17.698	0.743	9.700	16.974	0.683	9.394	16.606	0.676
14	10.303	18.830	0.741	10.566	18.108	0.683	10.258	17.742	0.677
15	11.171	19.959	0.739	11.435	19.239	0.683	11.127	18.873	0.677
16	12.042	21.083	0.738	12.309	20.364	0.683	12.000	20.000	0.678
17	12.918	22.204	0.736	13.186	21.487	0.683	12.877	21.123	0.678
18	13.797	23.321	0.735	14.068	22.605	0.683	13.757	22.243	0.678
19	14.680	24.435	0.734	14.951	23.720	0.683	14.641	23.359	0.678
20	15.565	25.547	0.732	15.839	24.832	0.683	15.528	24.472	0.679

The integral for the central interval definition is notably too large, the shortest interval is exactly correct (by definition), but the simple ±√n formulation actually isn't bad…

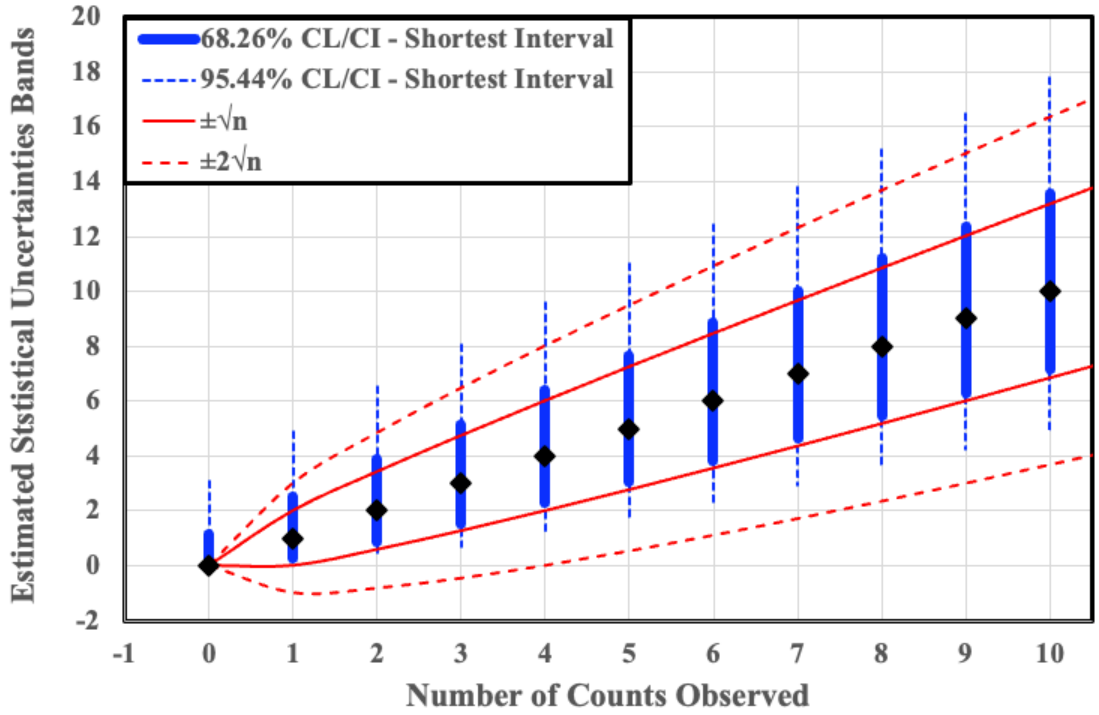
Frequentist coverage of 1σ intervals as a function of true mean:



Note: For quantised observations, such as Poisson, **it is <u>fundamentally</u> impossible to guarantee perfect statistical coverage for all values of μ!**
(because interval boundaries jump discontinuously for different observed counts)

Anyone obsessed with obtaining "exact" coverage for Poisson is barking up the wrong tree!

We see a similar story as with the integration: central intervals notably over-cover, the shortest interval covers at about the right level on average, but the simple ±√n formulation actually isn't bad…



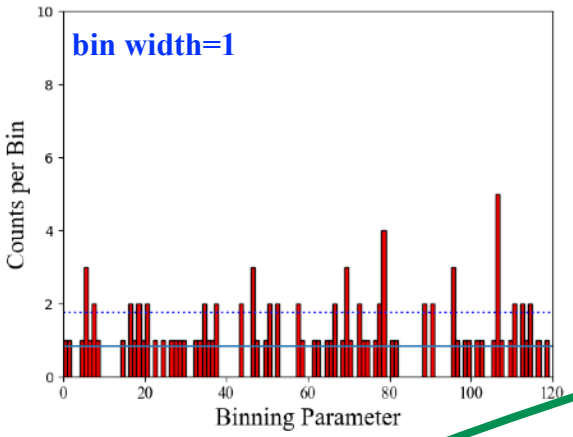
So ±√n isn't terrible as a way to represent approximate '1σ' intervals for an indeterminate model, especially if you ascribe an error bar of 0-1 for zero counts, as is often done.

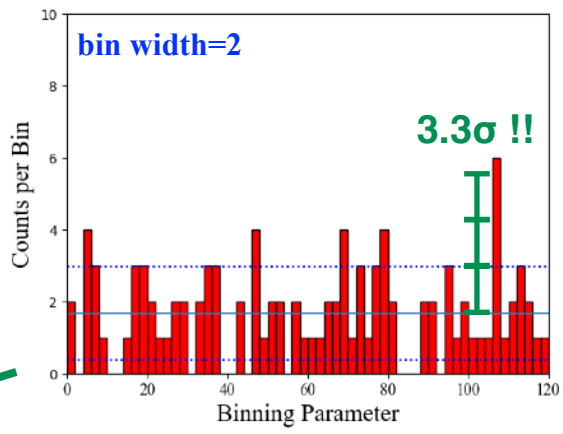
But at ±2σ, problems start to become more obvious, and this will become even worse at higher significance levels.

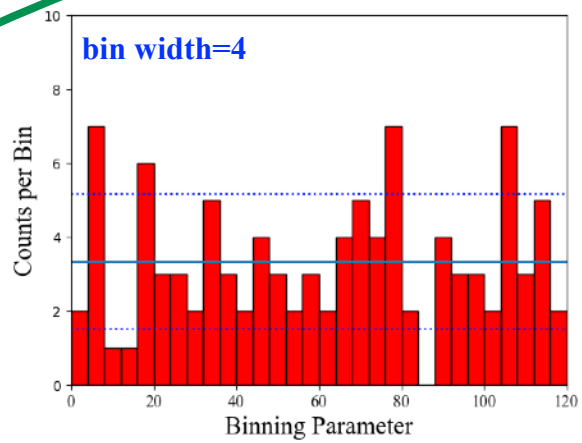
And remember that you shouldn't use these to test a particular model: you need to use model-based probability predictions!!

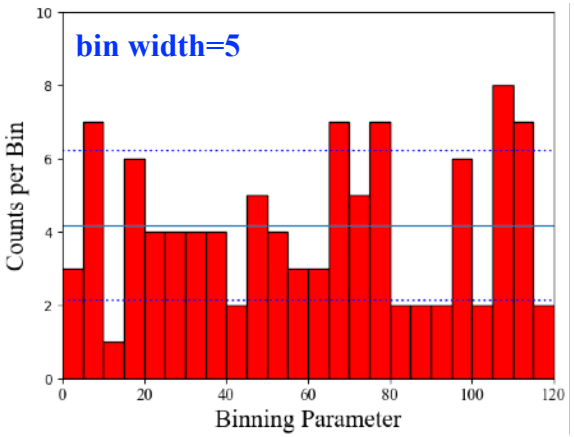
Be careful how you use these!

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)

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

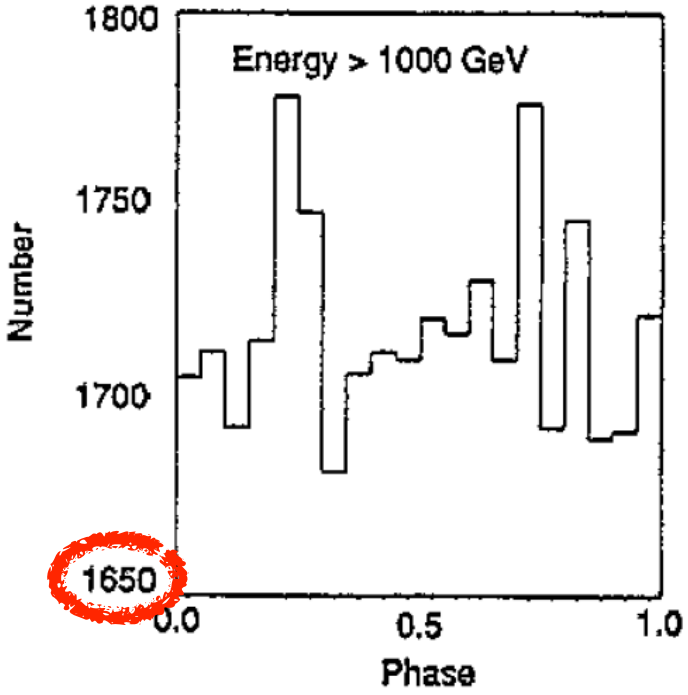
- For visual presentation/inspection of data, choose a binning based on the amount of statistics (to avoid bins with low numbers) and the anticipated scale of possible features.

- 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!

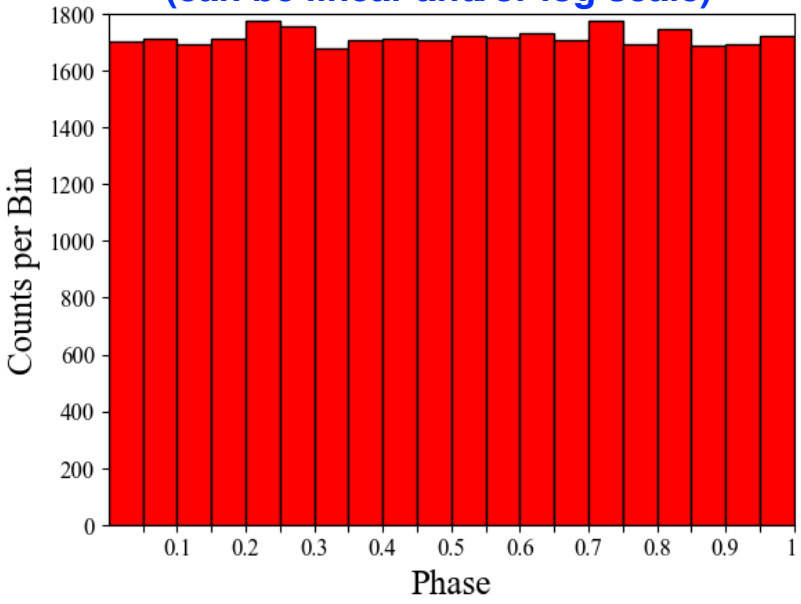
- Fitting and significance tests should be done using the correct probability distributions where appropriate.

- 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.

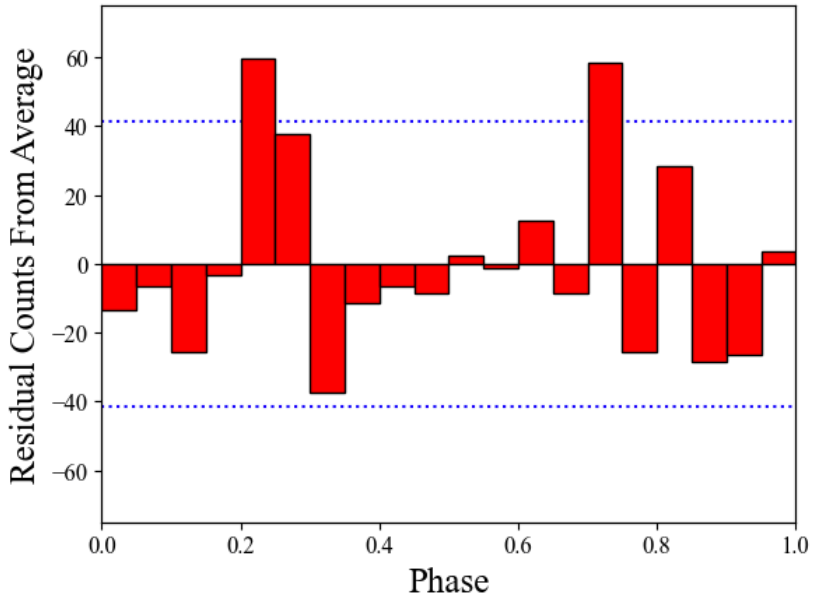
Try to avoid suppressed zeros!



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

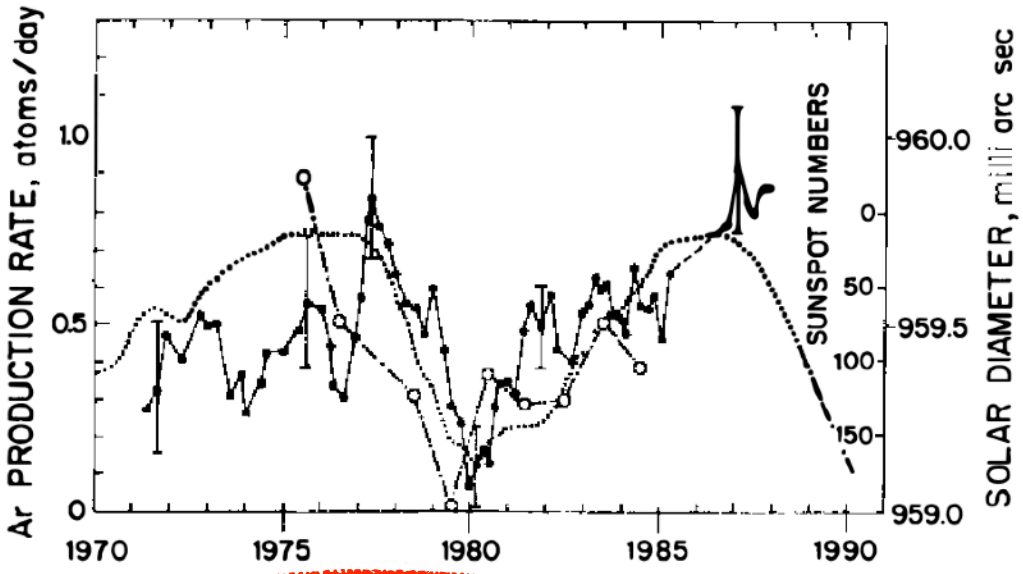


Figure 5 Plots of five-point running average of ^{37}Ar production and smoothed sunspot numbers against time in years (from 130). Solid circles, ^{37}Ar production; dotted curve, sunspot numbers; open circles, solar diameter.

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

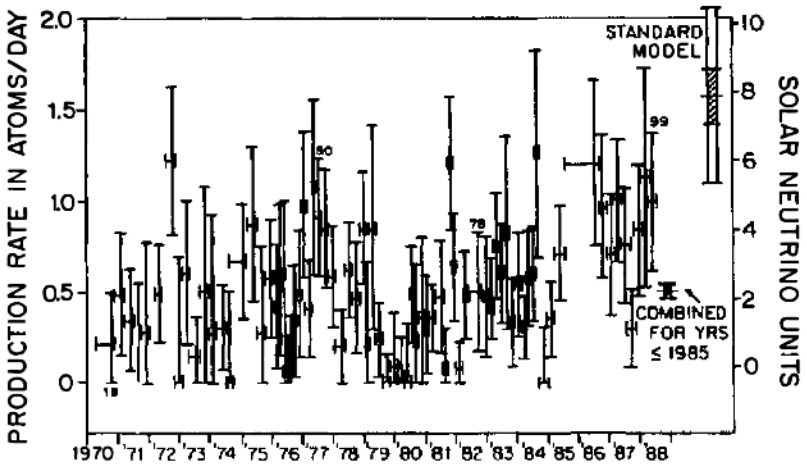
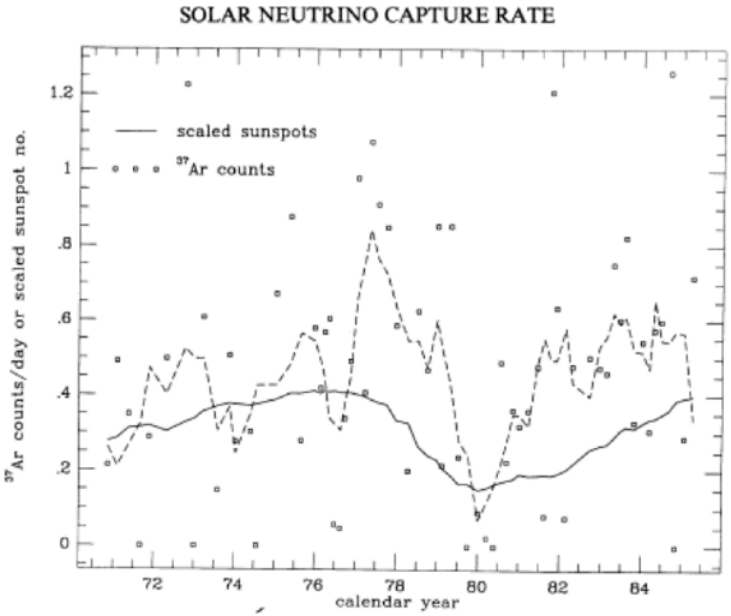


Figure 4 Production rate of ^{37}Ar in the ^{37}Cl detector as a function of time in years (from 66). The average rate in SNU for the period 1970–1985 is shown as 2.1 ± 0.3. Inclusion of the data for 1987–1988 raises the average to 2.3 ± 0.3. The prediction of the SSM from the calculation of Bahcall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



Ways to Display Uncertainties

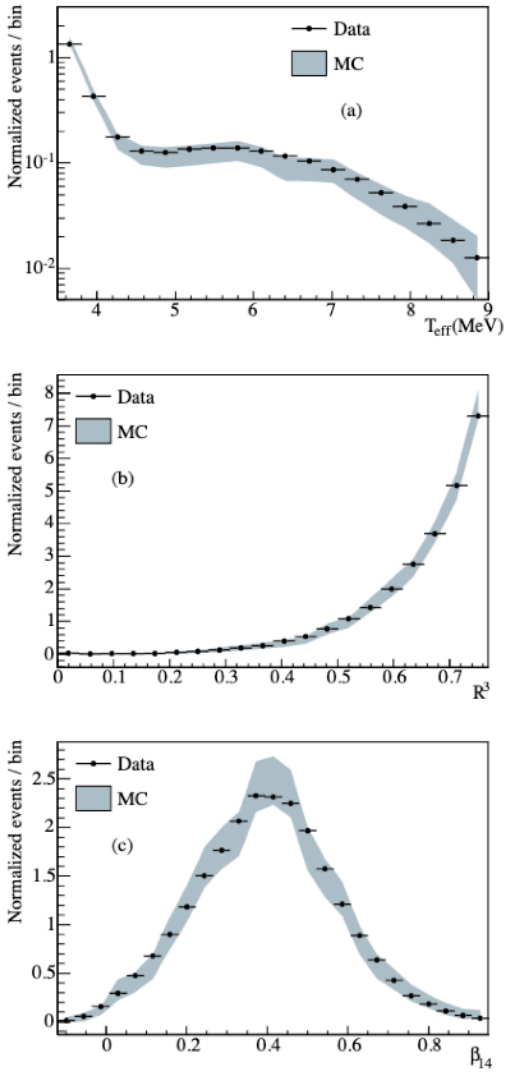


FIG. 24: (Color online) Comparison of data to simulation for ^{232}Th source runs near the AV in Phase II, in (a) $T_{\rm eff}$, (b) R^3, and (c) β_{14}. The band represents the 1σ uncertainty on the Monte Carlo-prediction, taking the quadrature sum of the statistical uncertainties with the effect of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

Phys.Rev.C81:055504,2010

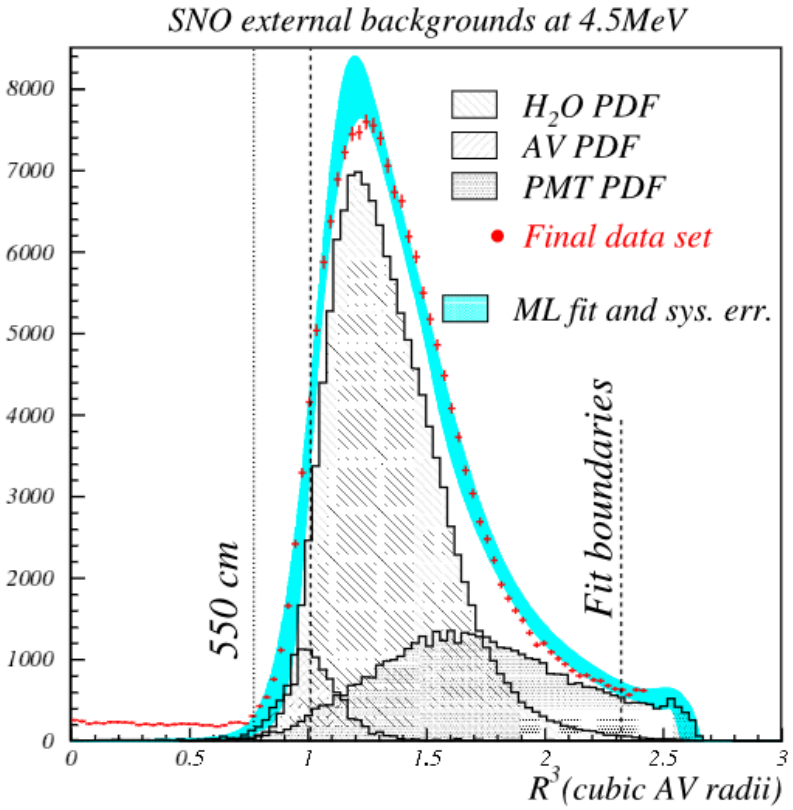


FIG. 34: Fit of of R^3 pdfs created using calibration source data to the neutrino data set, using an energy threshold of $T_{\rm eff} > 4.0$ MeV. The extended maximum likelihood method was used in the fit, and the band represents the systematic uncertainties. The y-axis is in units of Events/0.03 cubic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

Phys.Rev.C75:045502,2007

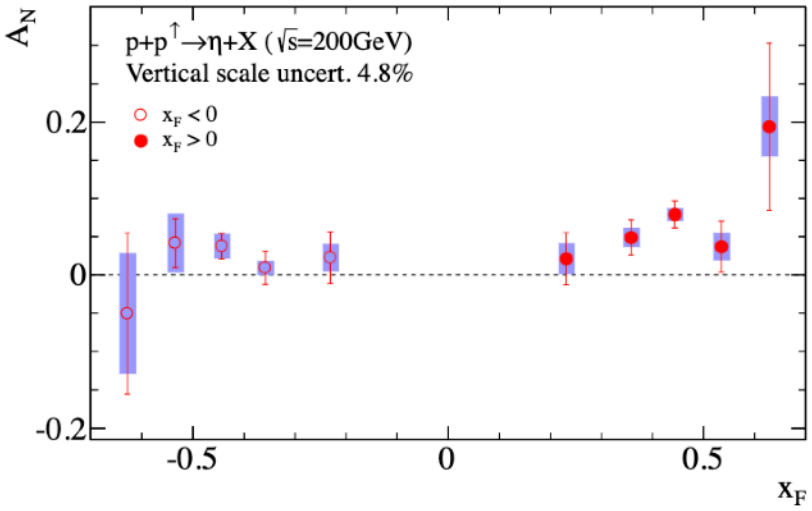


FIG. 9: (Color online) The x_F dependence of A_N. The vertical error bars show the statistical uncertainty, the blue bands represent uncorrelated systematic uncertainties (see text for details). The relative luminosity effect systematic uncertainties are not shown (see text and Table III)

Phys.Rev.D90:072008,2014

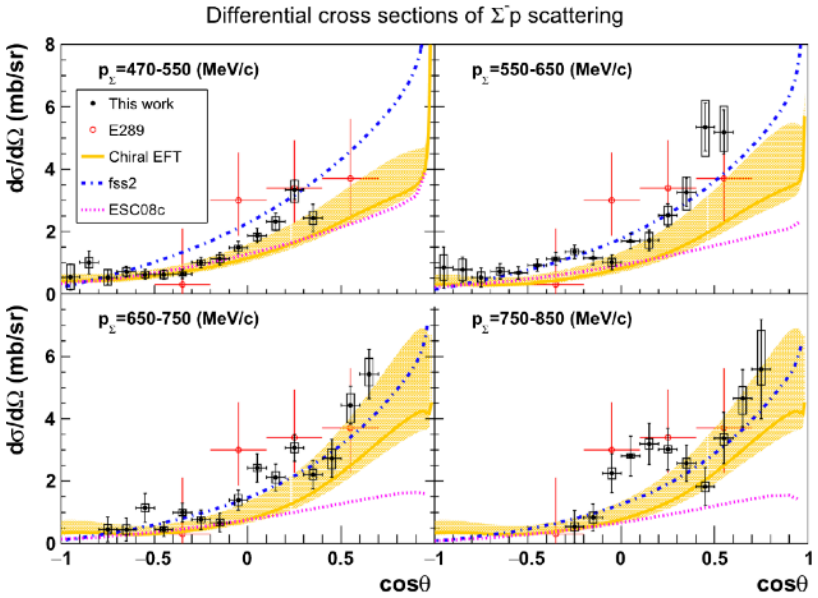
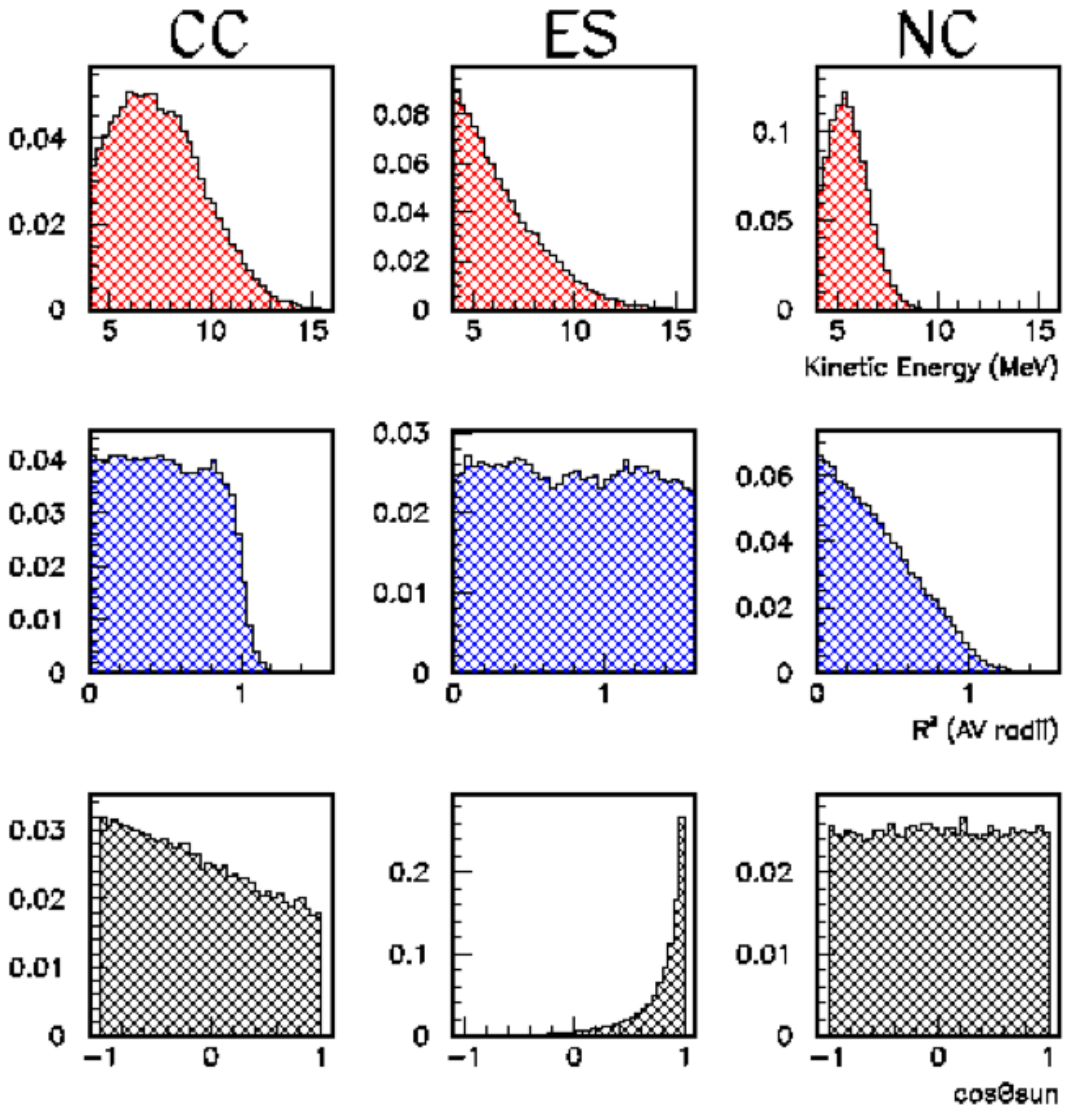


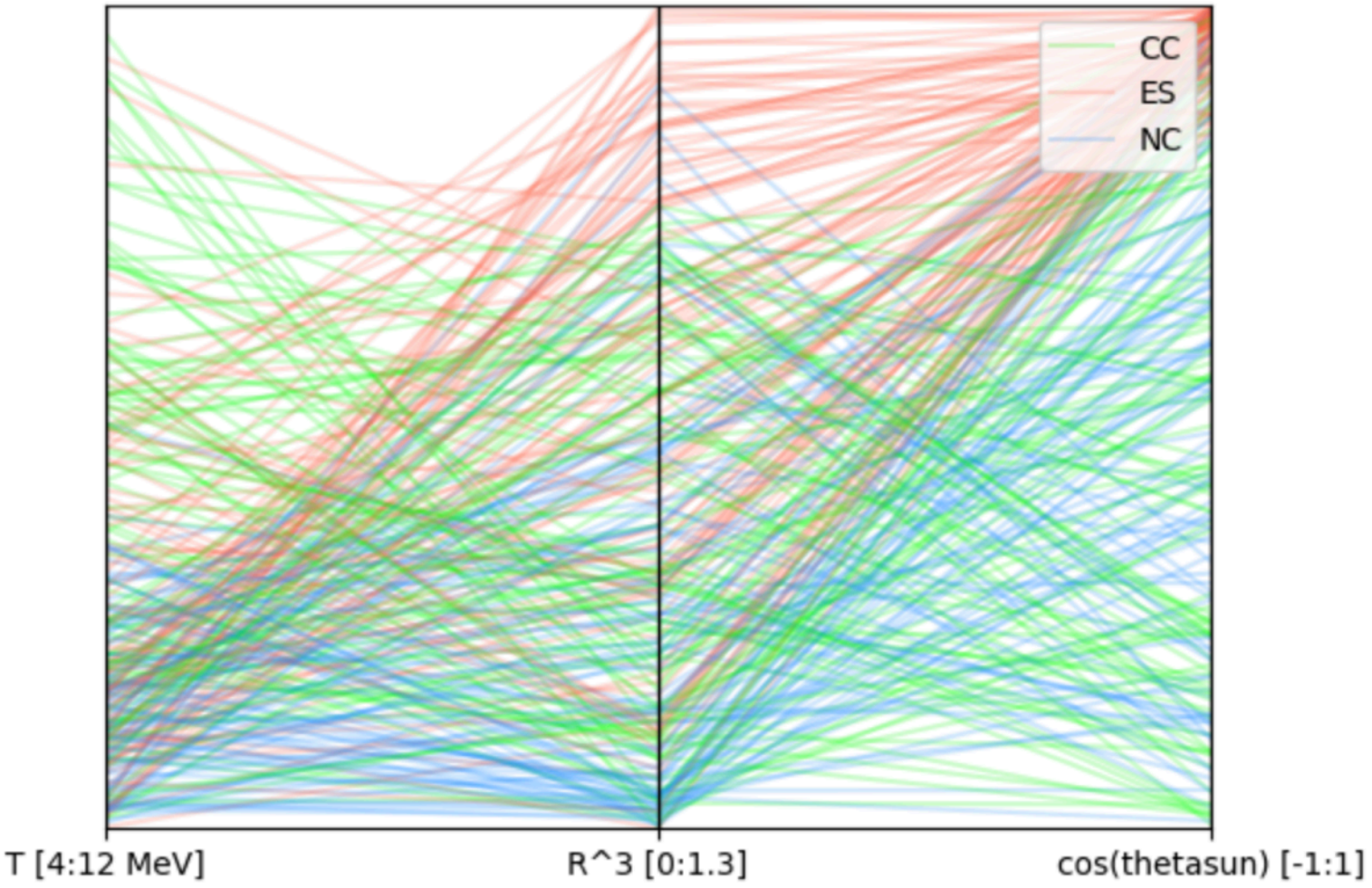
Fig. 5 Derived differential cross sections (black points). The error bars and boxes show the statistical and systematic uncertainties, respectively. Red points are averaged differential cross section of $0.4 < \text{GeV}/c < 0.7$ taken in KEK-PS (the same points are plotted in the four momentum regions). 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



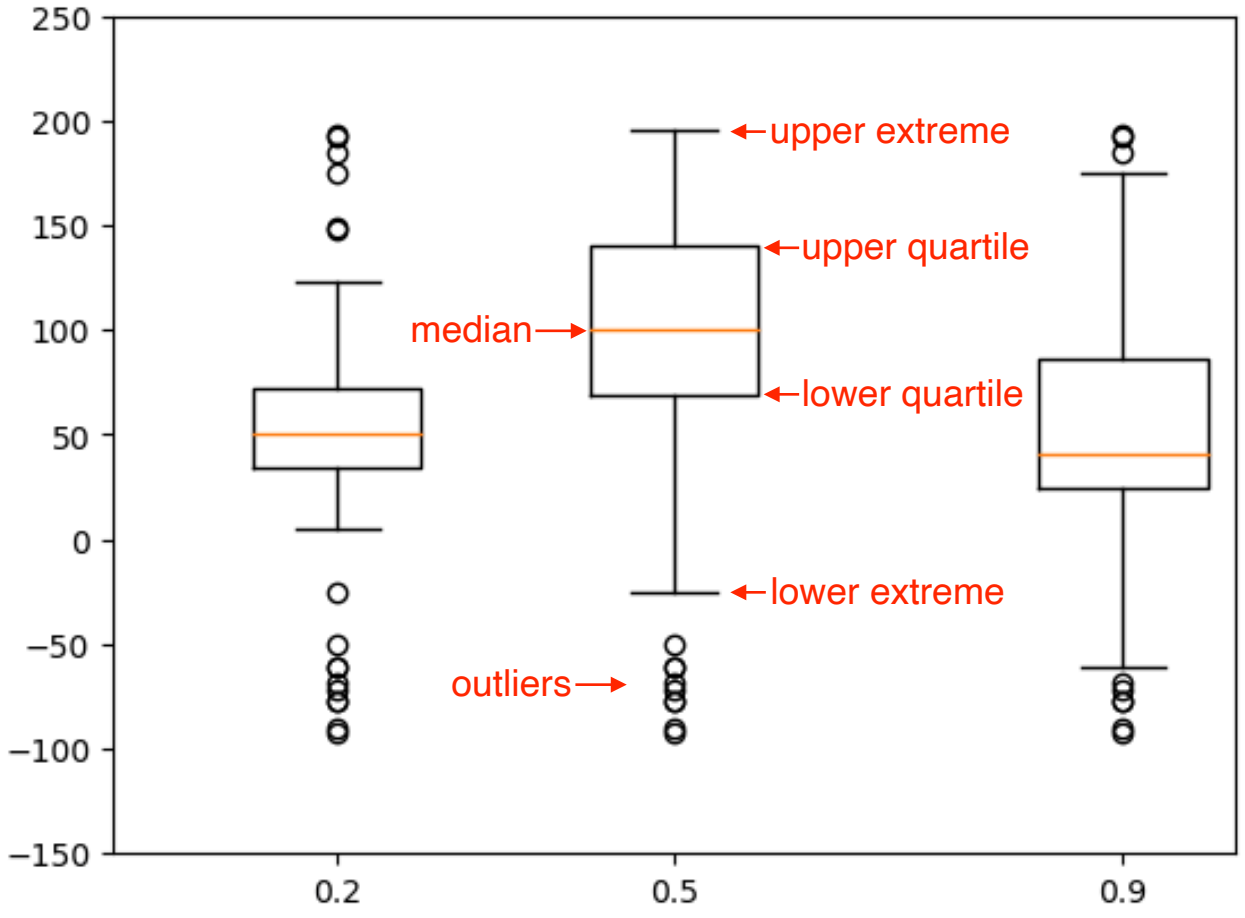
Parallel Coordinate Plot:




```
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:

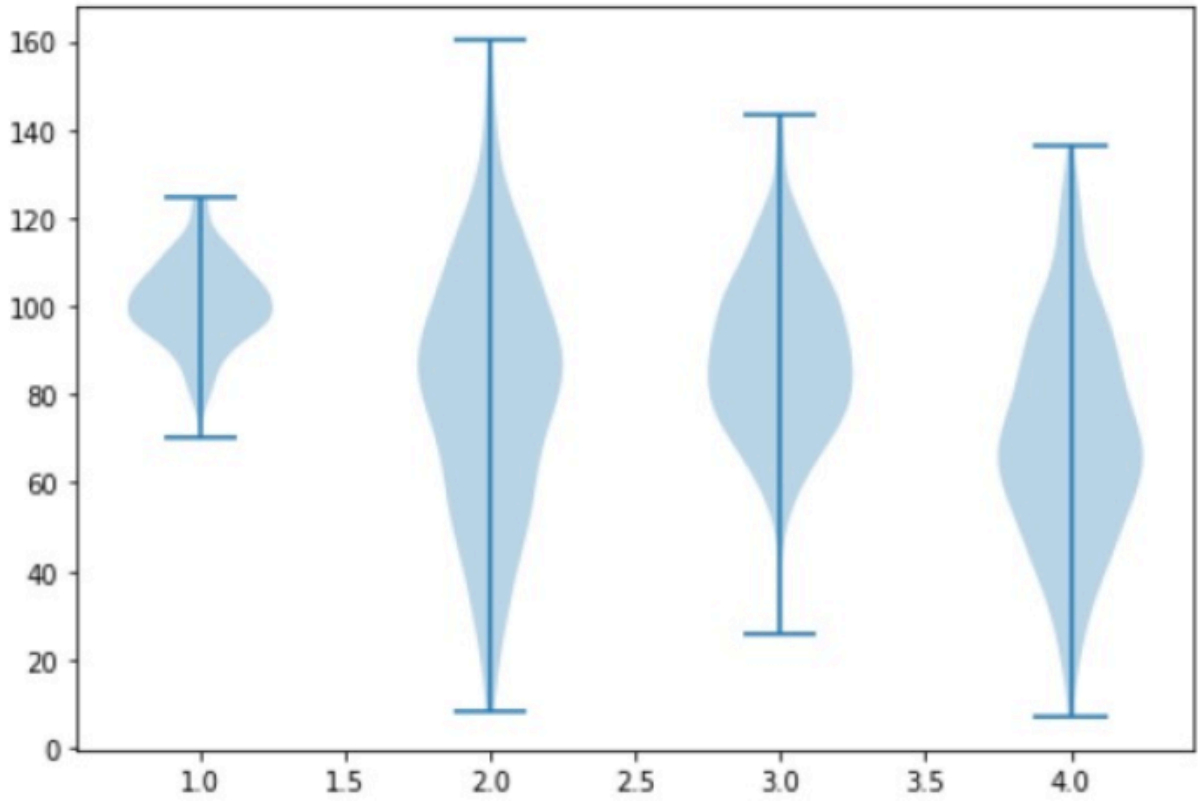
Box and Whisker:




```
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()
```