Subthreshold signals in binned data

Advanced Methods in Applied Statistics, vol. 2021

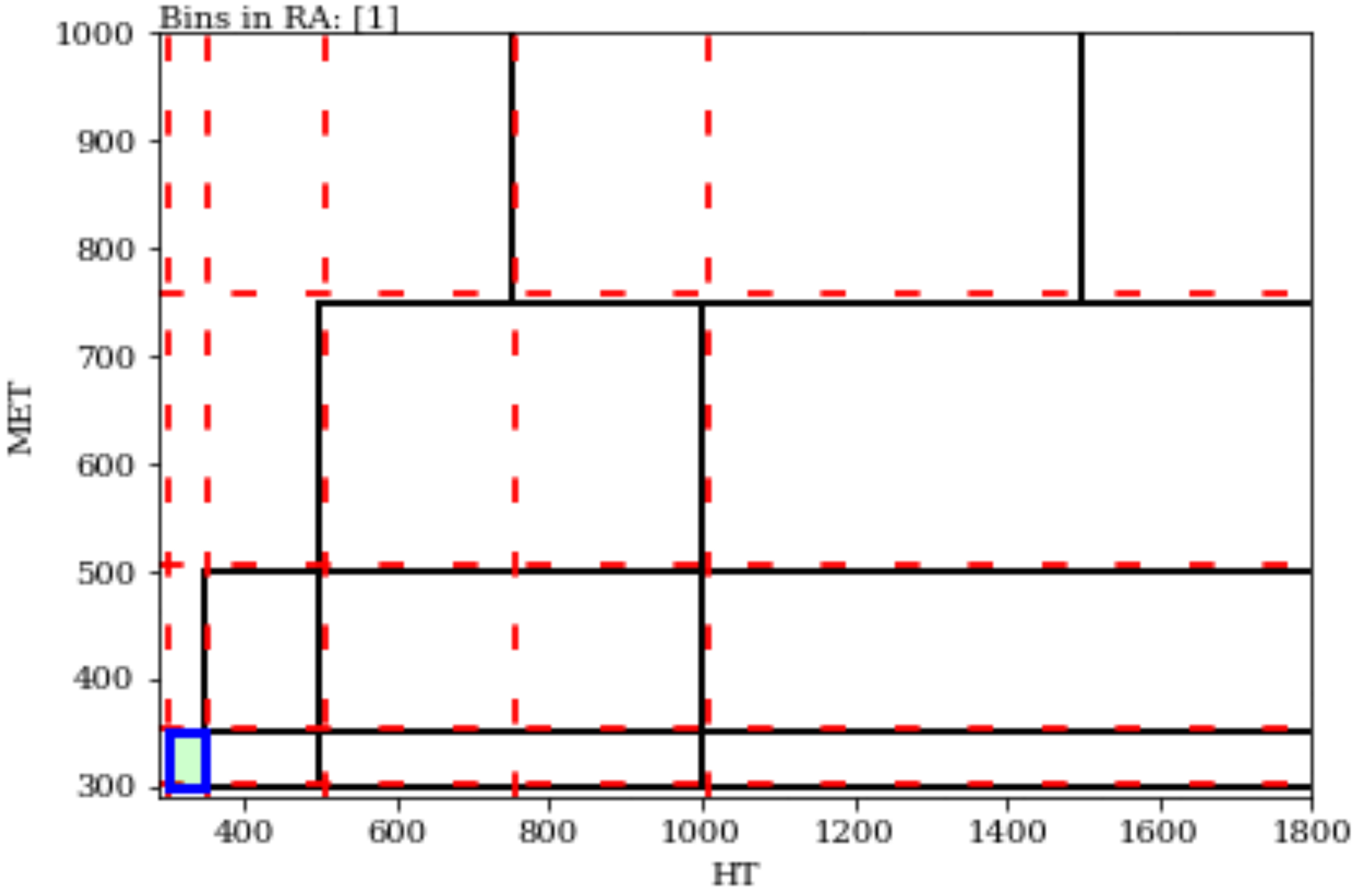
Tania Kozynets
tetiana.kozynets@nbi.ku.dk

March 16, 2021

UNIVERSITY OF COPENHAGEN

How do we know if our binned data is anomalous?

> Jason has hinted at the idea of scanning the binned data using all possible combinations of bin mergers, e.g.:

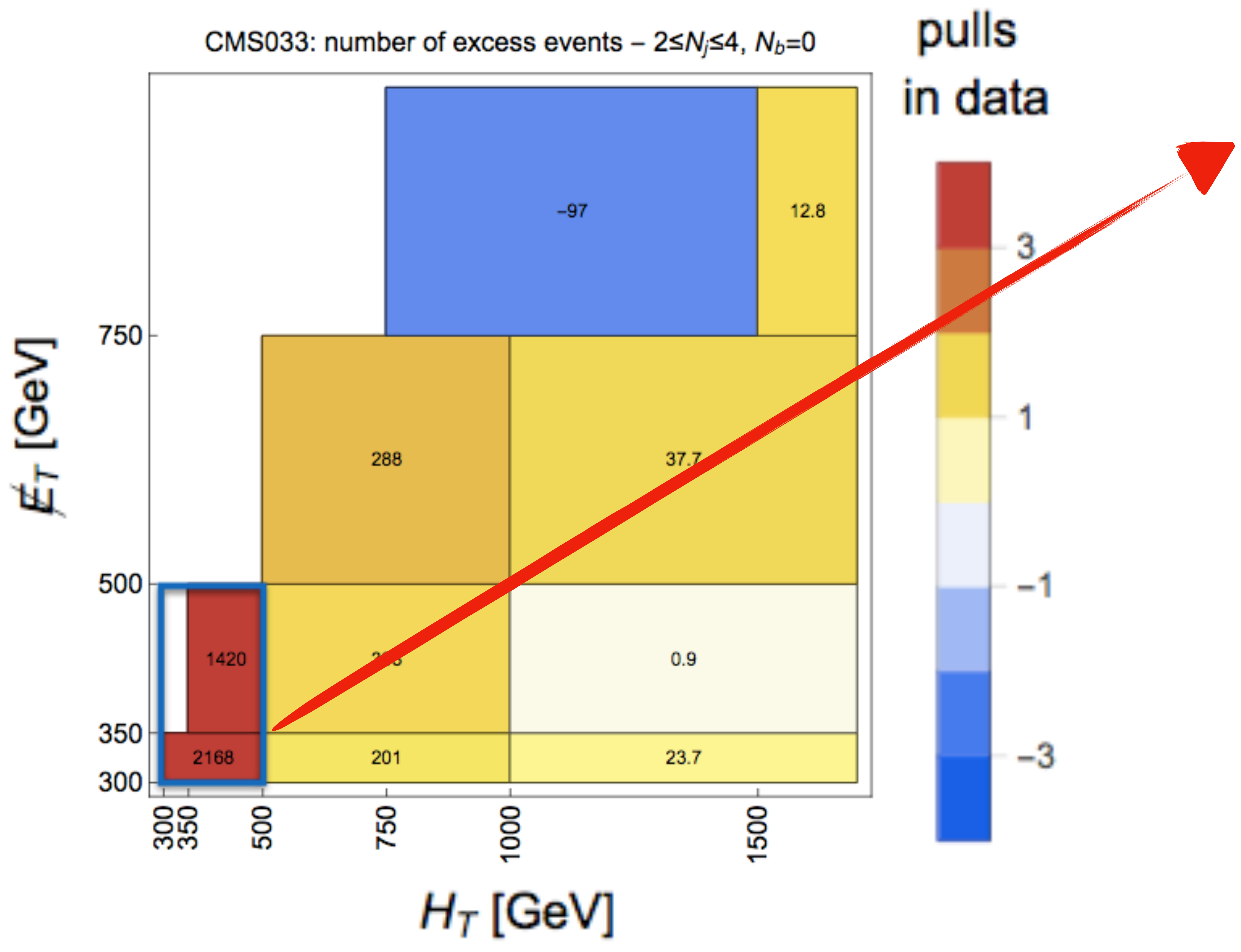


animation by
A.Monteux

*Asadi 2017

What can we get out of the scan?

> After such a scan, one way to proceed is to take just the most anomalous bin (i.e., the one with the highest pull). An example from high-energy physics:



the CMS experiment at CERN found this to be the most anomalous bin merger

the "abnormality" of the whole data grid is then dependent on the "*p*-value" of this particular excess

the rest of the information is lost, plus now we need a model to explain this excess at this particular location…

*Asadi 2017

What do we do to avoid a “single template” search?

> We want to go away from digging into the *p*-values of some very specific signal shapes, which just happen to be “most anomalous”;

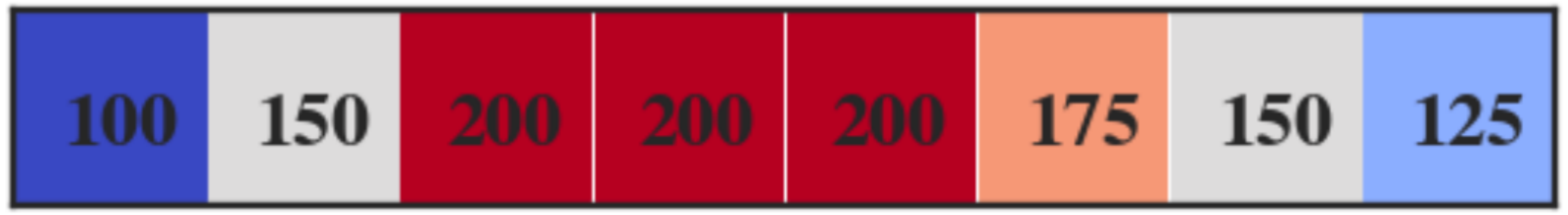
> To do so, we need to incorporate the information on all bin mergers, not just the most anomalous one;

> Exercise 2 will show us the way.

Exercise 2

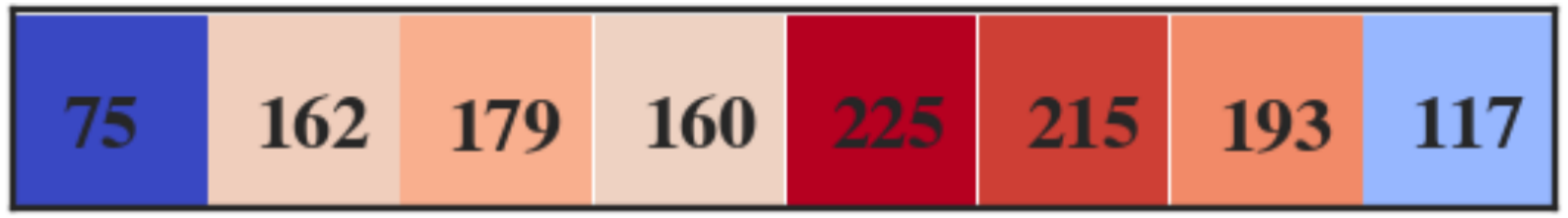
> Here, we will be looking at the following 1D (“1x8”) data:

expectation (*)



100	150	200	200	200	175	150	125

observation (**)

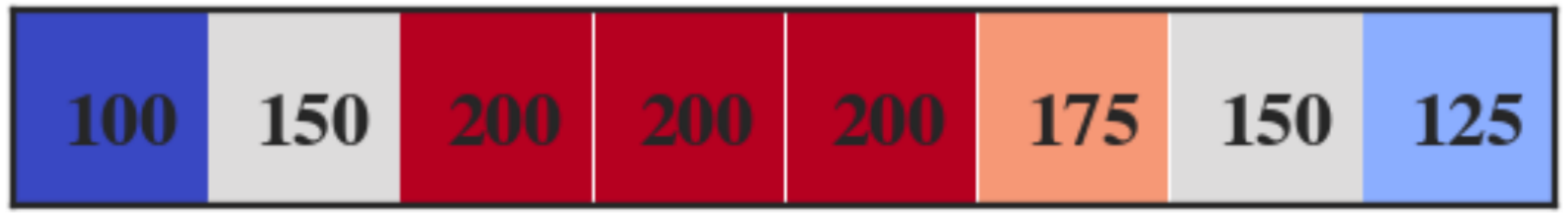


75	162	179	160	225	215	193	117

> The goal: find all the bin mergers and quantify their pulls, for both the “fluctuated expectation” pseudoexperiments and the measured signal.

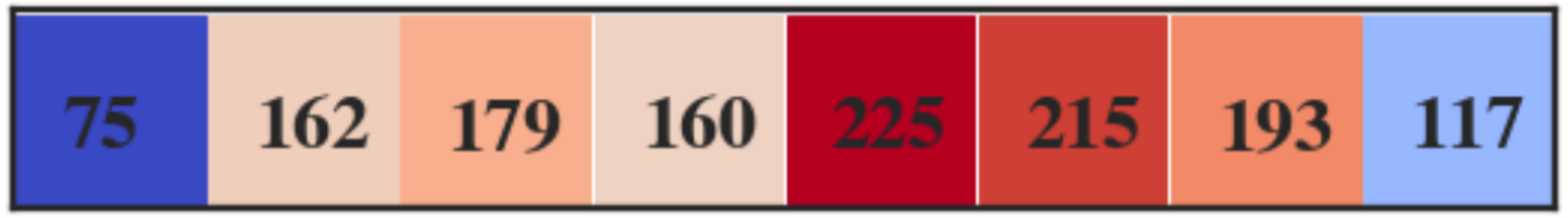
Exercise 2

expectation (*)



100	150	200	200	200	175	150	125

observation (**)



75	162	179	160	225	215	193	117

Tasks:

> For the single observed dataset (**), calculate the pulls for the existing bins as done in Ex.1, given the single expectation (*).

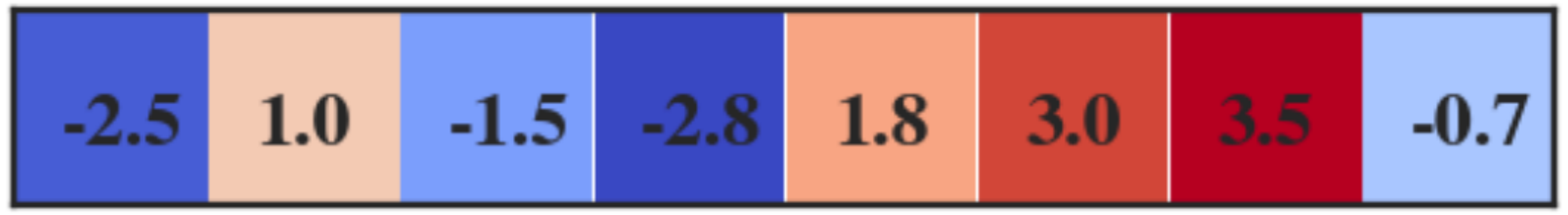
> Then, go through “bin merging” and find how many mergers (***N***$_X$) there are with cumulative pull (significance) above **X=2,3,4 σ**.

> Also, generate 1000 expectation pseudoexperiments from (*) and repeat the above calculation for each pseudoexperiment (assume Poisson fluctuations).

> Histogram ***N***$_X$ outputs for the pseudoexperiments as well as the observation (**). Where does the observation lie with respect to the fluctuated expectations? Is our observation anomalous?

Exercise 2 “solutions”

> I get the following pull values for the observation:



-2.5	1.0	-1.5	-2.8	1.8	3.0	3.5	-0.7

(**)

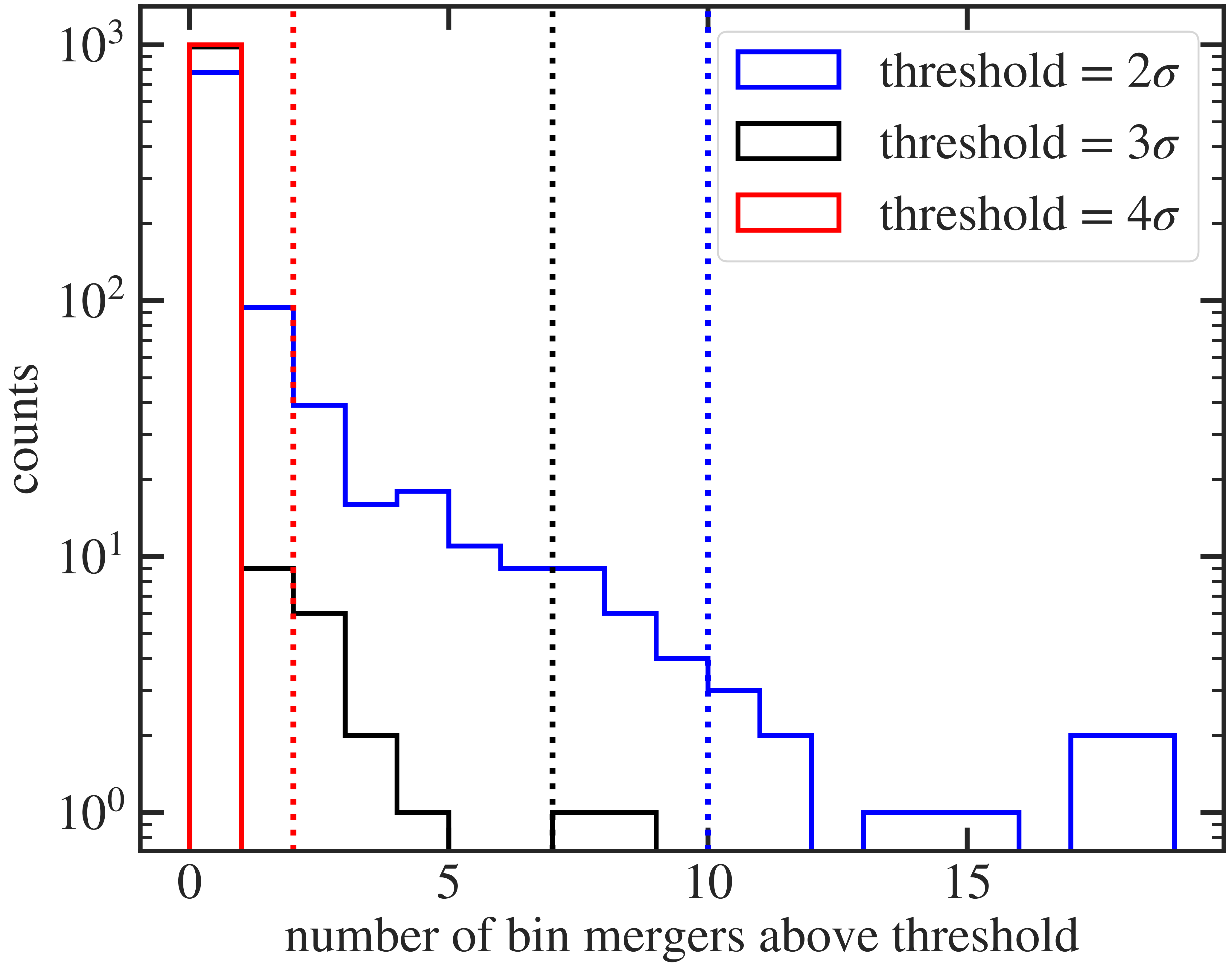
> The number of bin mergers above 2,3,4 σ: ***N*₂** = 10, ***N*₃** = 7, ***N*₄** = 2 (including 1x1 bins).

> Fluctuated background distributions:

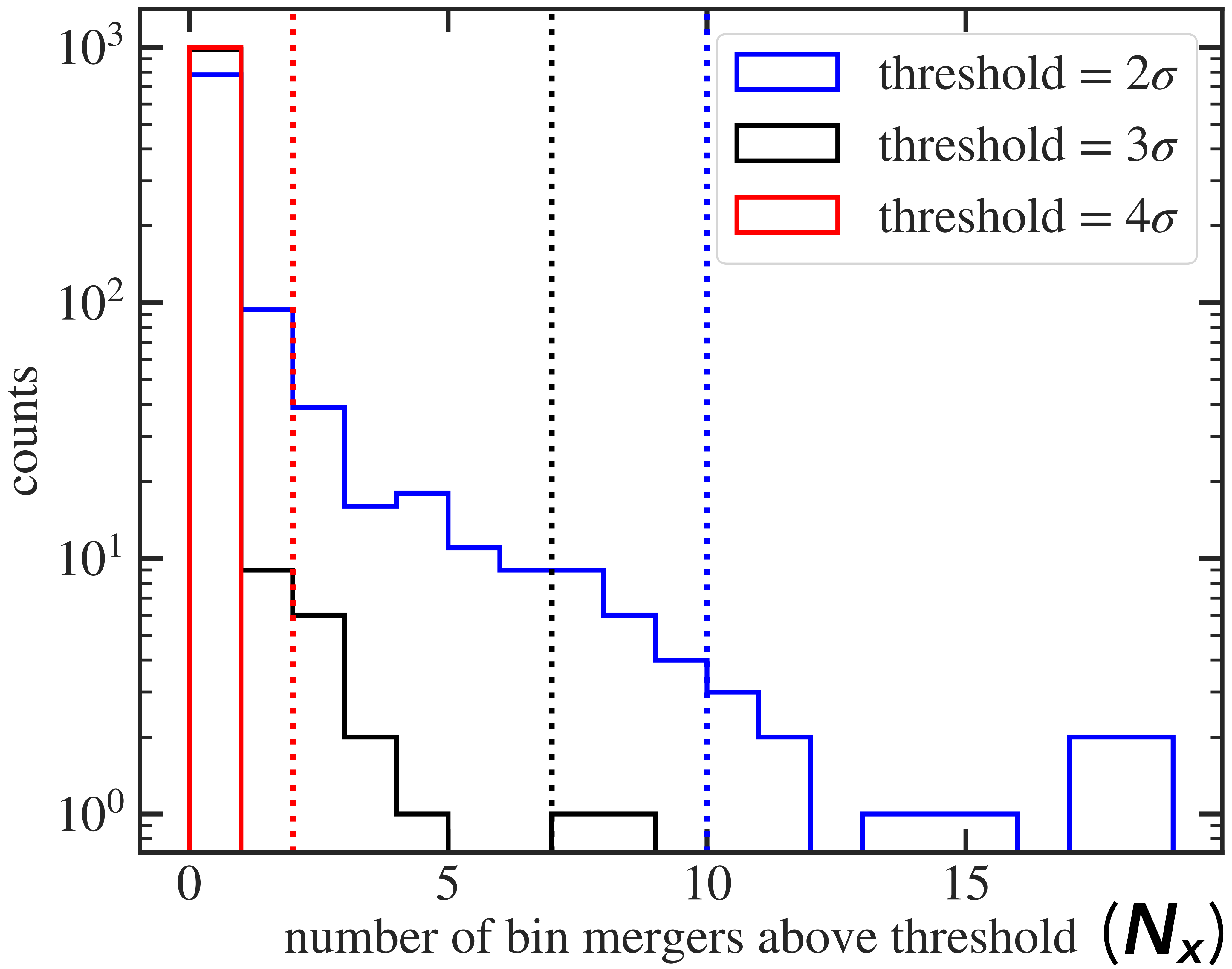
Solid lines = distributions from the 1,000 fluctuated expectation pseudoexperiments;

Dotted lines = results obtained for the observation (**).

So, is our observation anomalous?



Exercise 2 “solutions”



> From my 1,000 pseudoexperiments, there is a…

1.4% chance to get the $\boldsymbol{N_2}$ value as high as the observed or larger;

0.2% chance to get the $\boldsymbol{N_3}$ value as high as the observed or larger;

0% chance to get the $\boldsymbol{N_4}$ value as high as the observed or larger (such large values just never made it into the pseudoexperiments).

So… Multiple *p*-values per observation?

> No. It is not desirable to have many *p*-values like this. We want to have just one, irrespective of the “σ threshold” that we set.

> We see that this problem of “multiple *p*-values” appears already in 1D when we try to characterize the distribution of merged bins with different pulls.

> It won’t get easier as we go to higher-dimensional data… *and we still want a single value for the test statistic!*

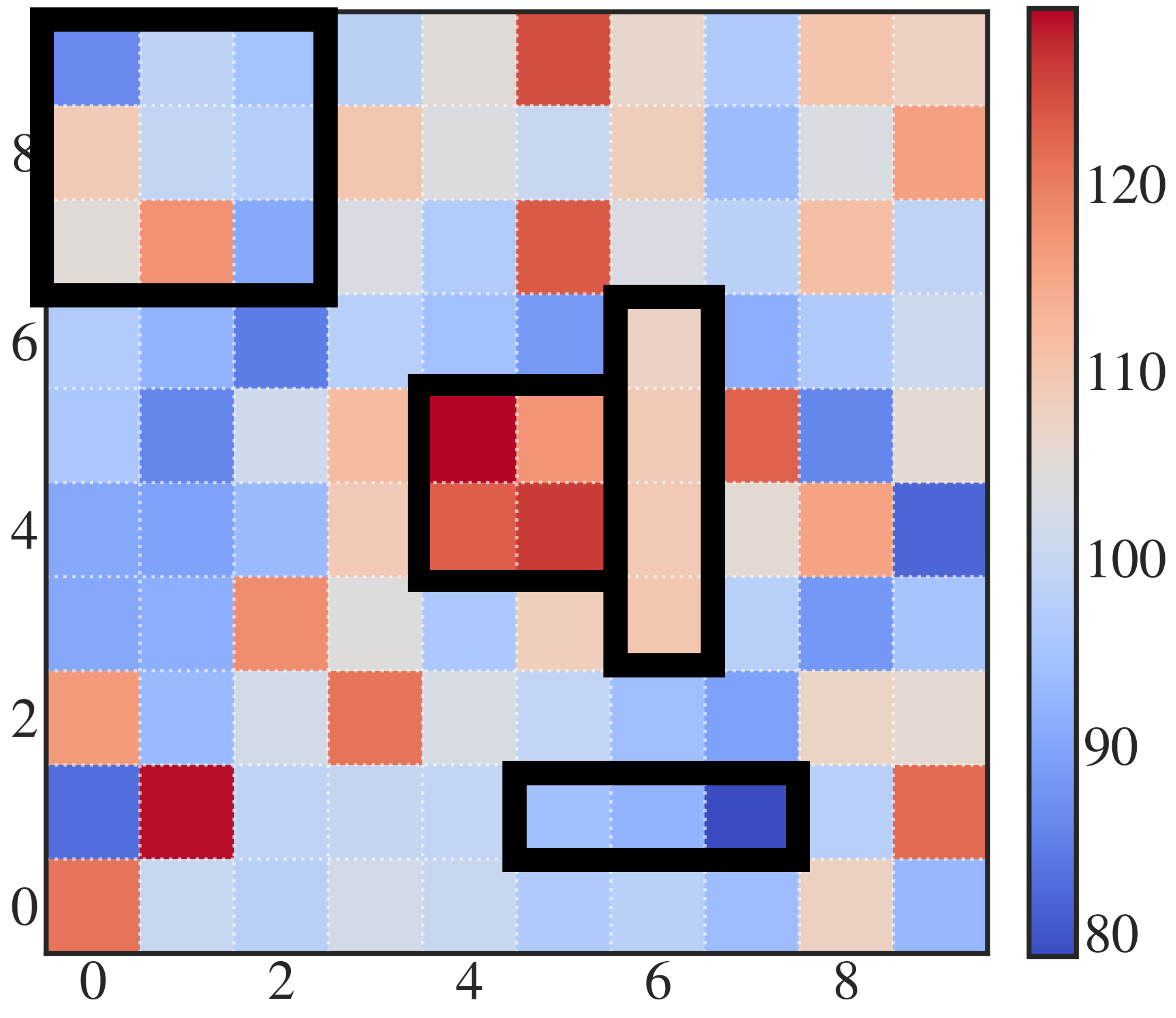
Moving to 2D data…

> The more dimensions/bins in each dimension we have, the more time the merged bin scans will take.

> The bins will also have more “properties” (location in each dimension, extent in each dimension, overall pull,…), and we want to keep the useful ones to define the test statistic.

Our approach:

keep track of the merged bin area in addition to the pull



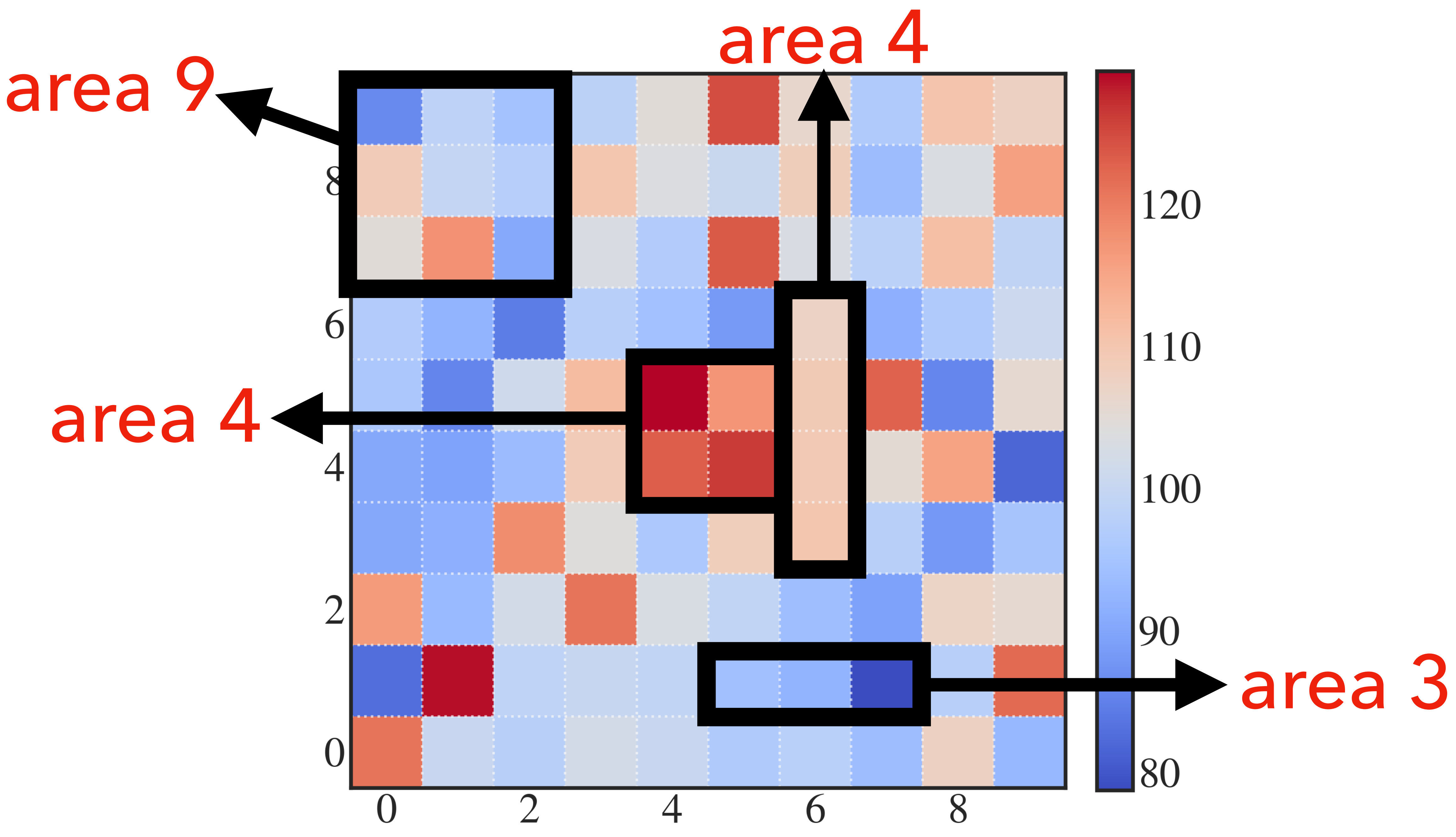
Moving to 2D data…

> The more dimensions/bins in each dimension we have, the more time the merged bin scans will take.

> The bins will also have more "properties" (location in each dimension, extent in each dimension, overall pull,…), and we want to keep the useful ones to define the test statistic.

Our approach:

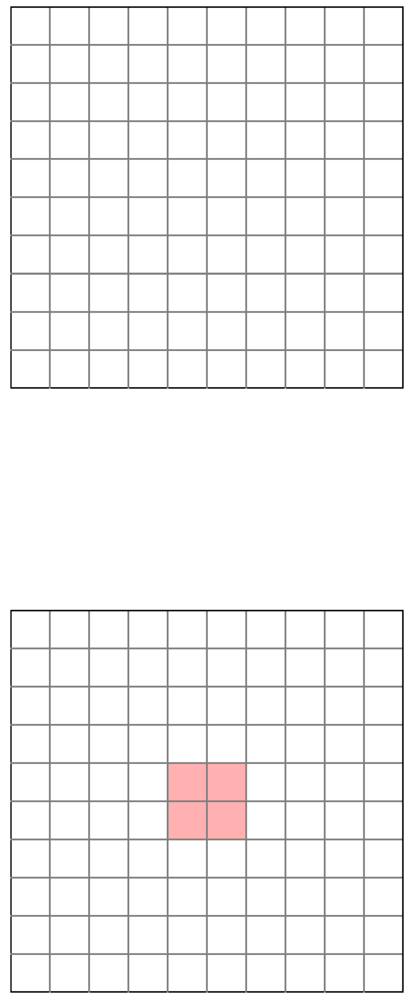
keep track of the merged bin area in addition to the pull

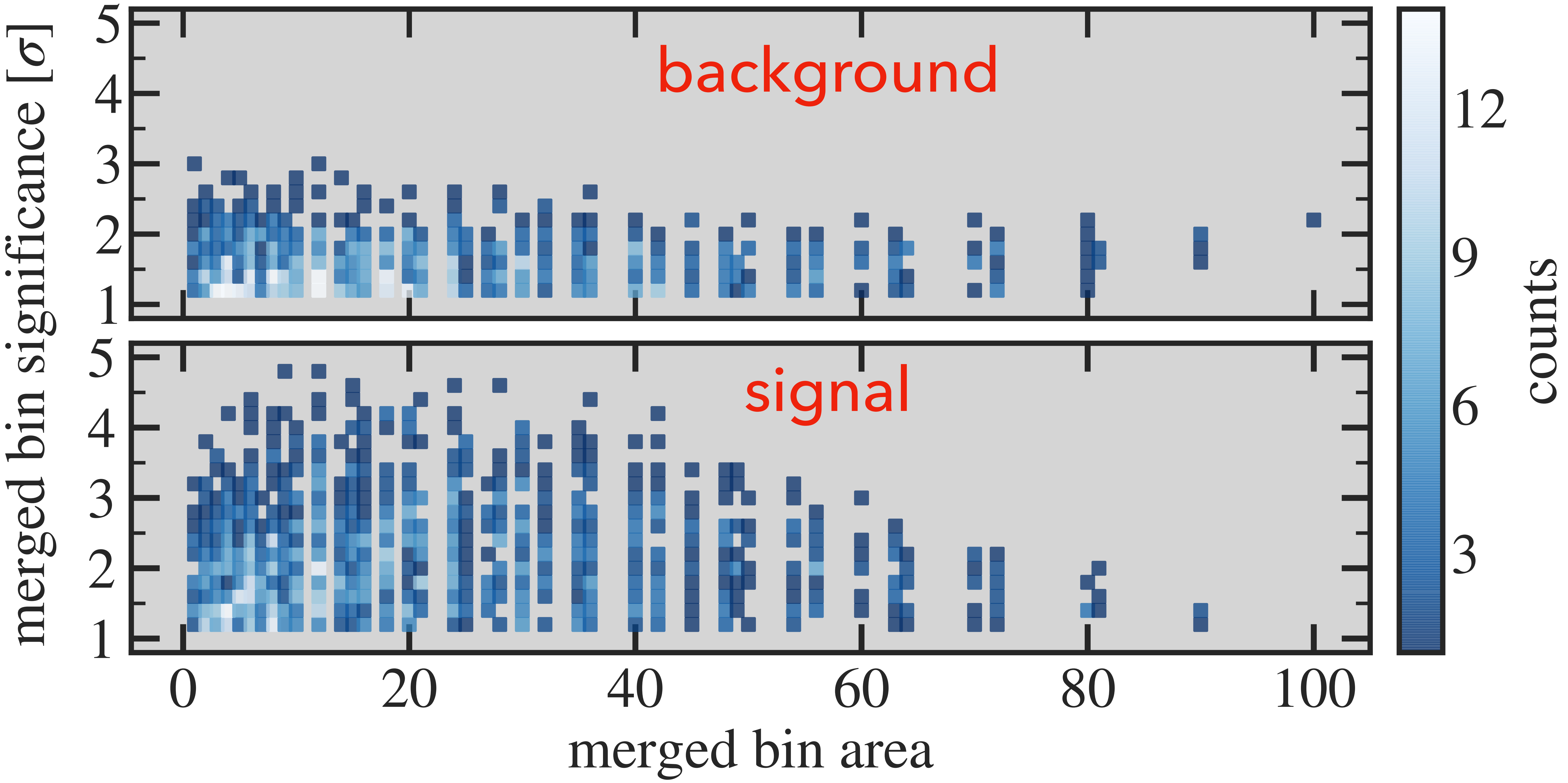


How do we summarize our pull/area findings?

> Once we have measured the areas and the pulls of the merged bins, we can put them in a 2D histogram. For a 10x10 grid, bin areas go from 1 to 100;

> An example of what we would get for a fluctuated background vs signal:

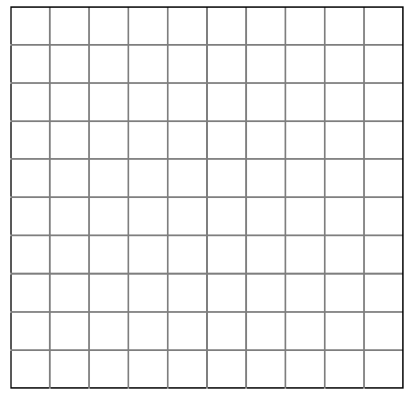


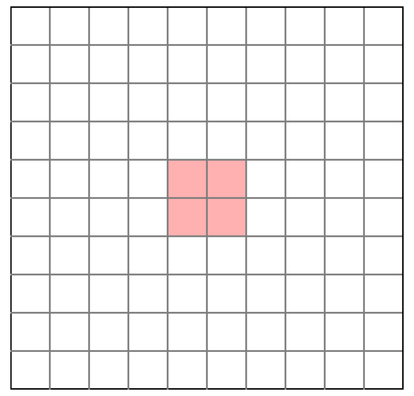


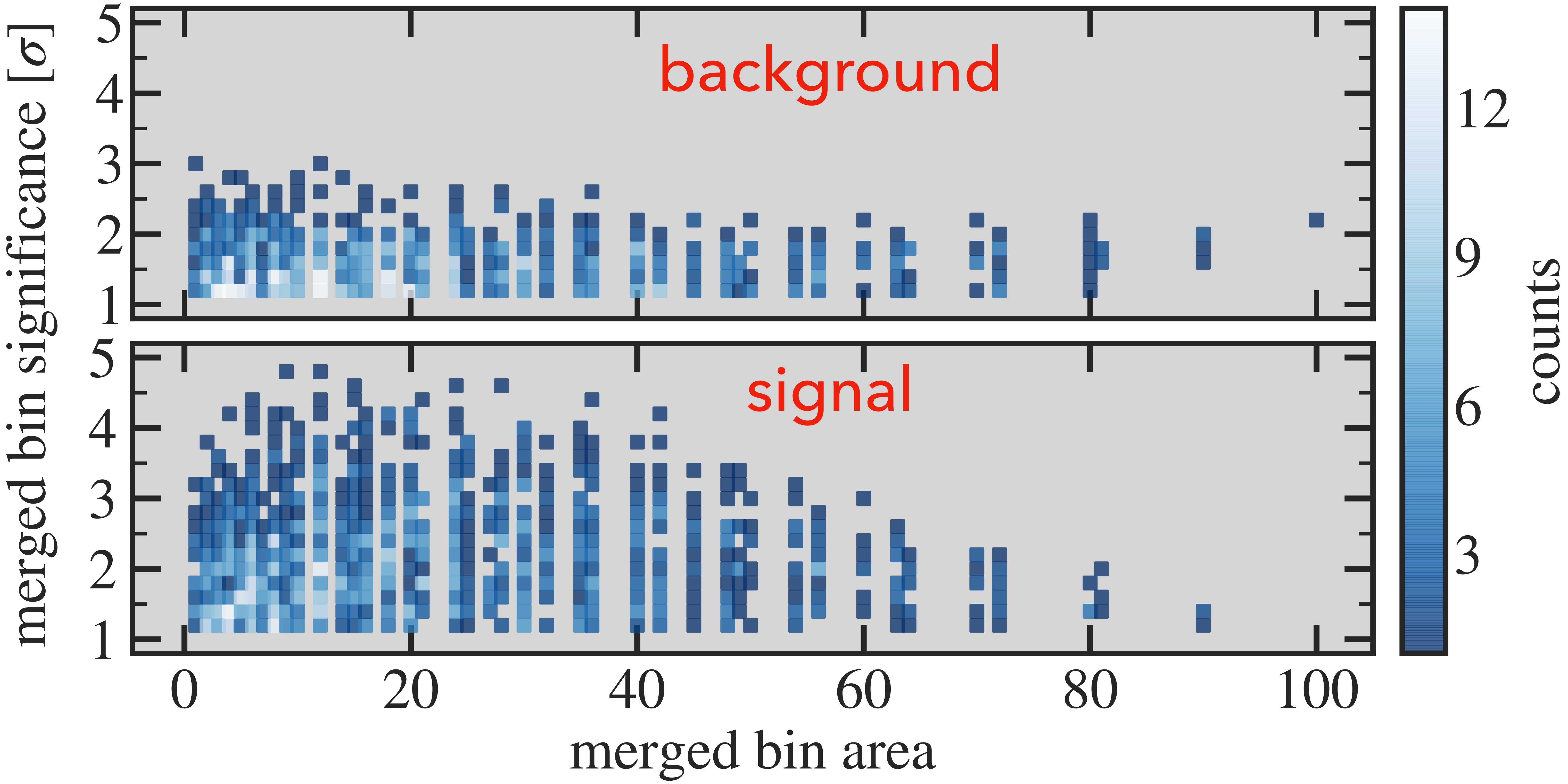
How do we summarize our pull/area findings?

> The “signal” in this case is a 2x2 box (cumulative pull = 4σ) injected in the middle of our 10x10 grid with a flat expectation of a 500 in each bin.

> We can see that this histogram looks *very different* from the background, so our summary of the bin features was useful. How do we produce a single test statistic value though?



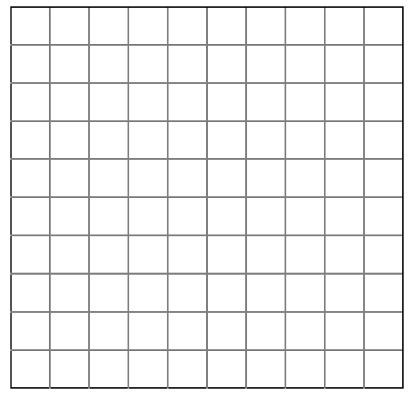


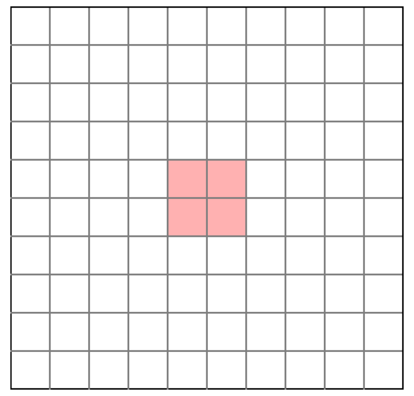


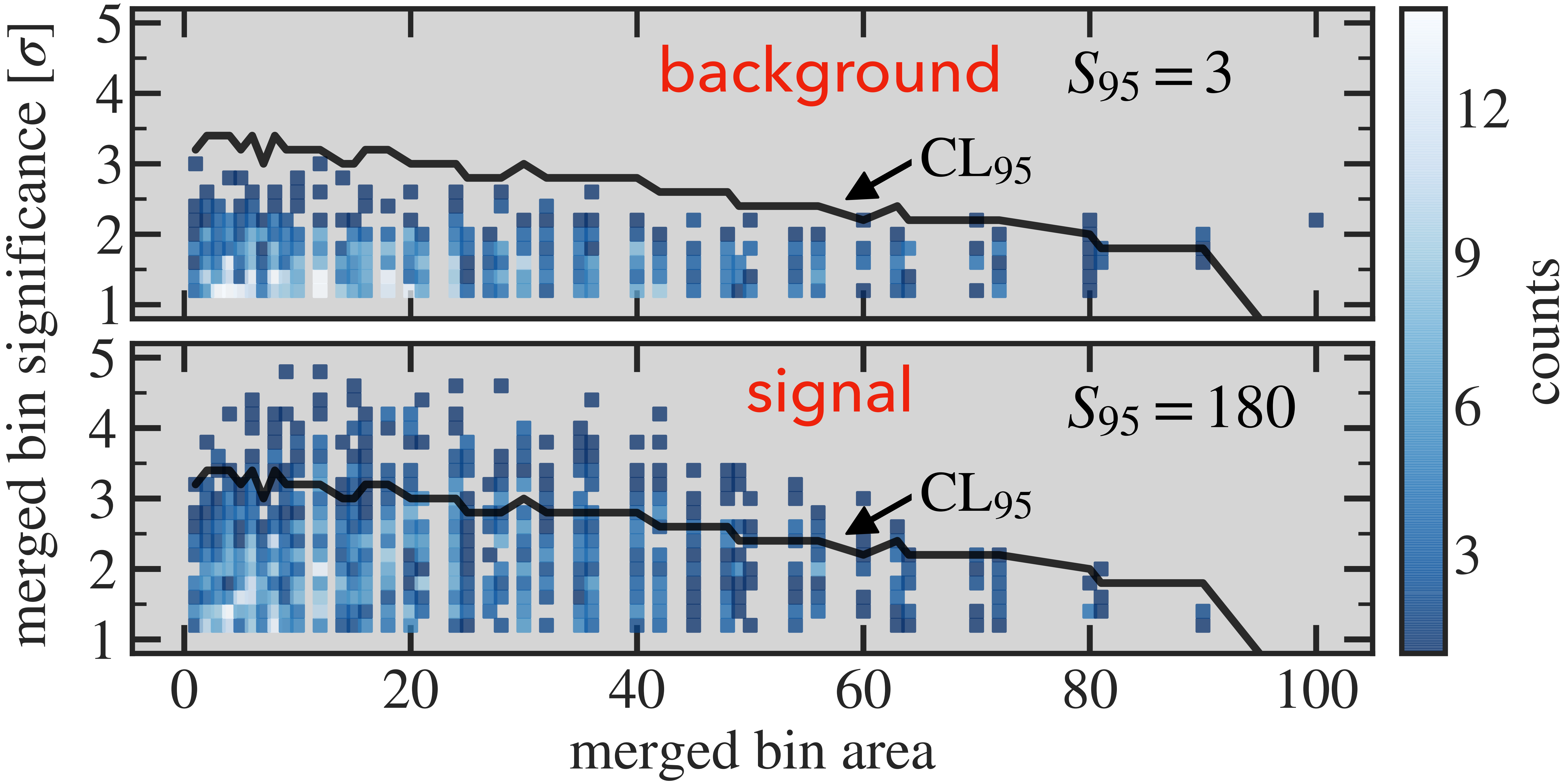
How do we summarize our pull/area findings?

> We will define a score test statistic, S_X, as the number of outliers that lie ***outside of a certain confidence level** X*, which we evaluate from the statistically fluctuated expectation.

> Example for X=95%:



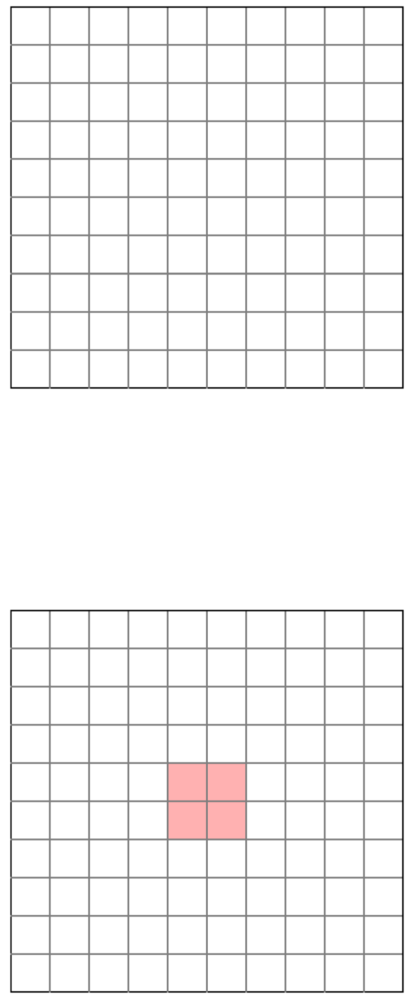


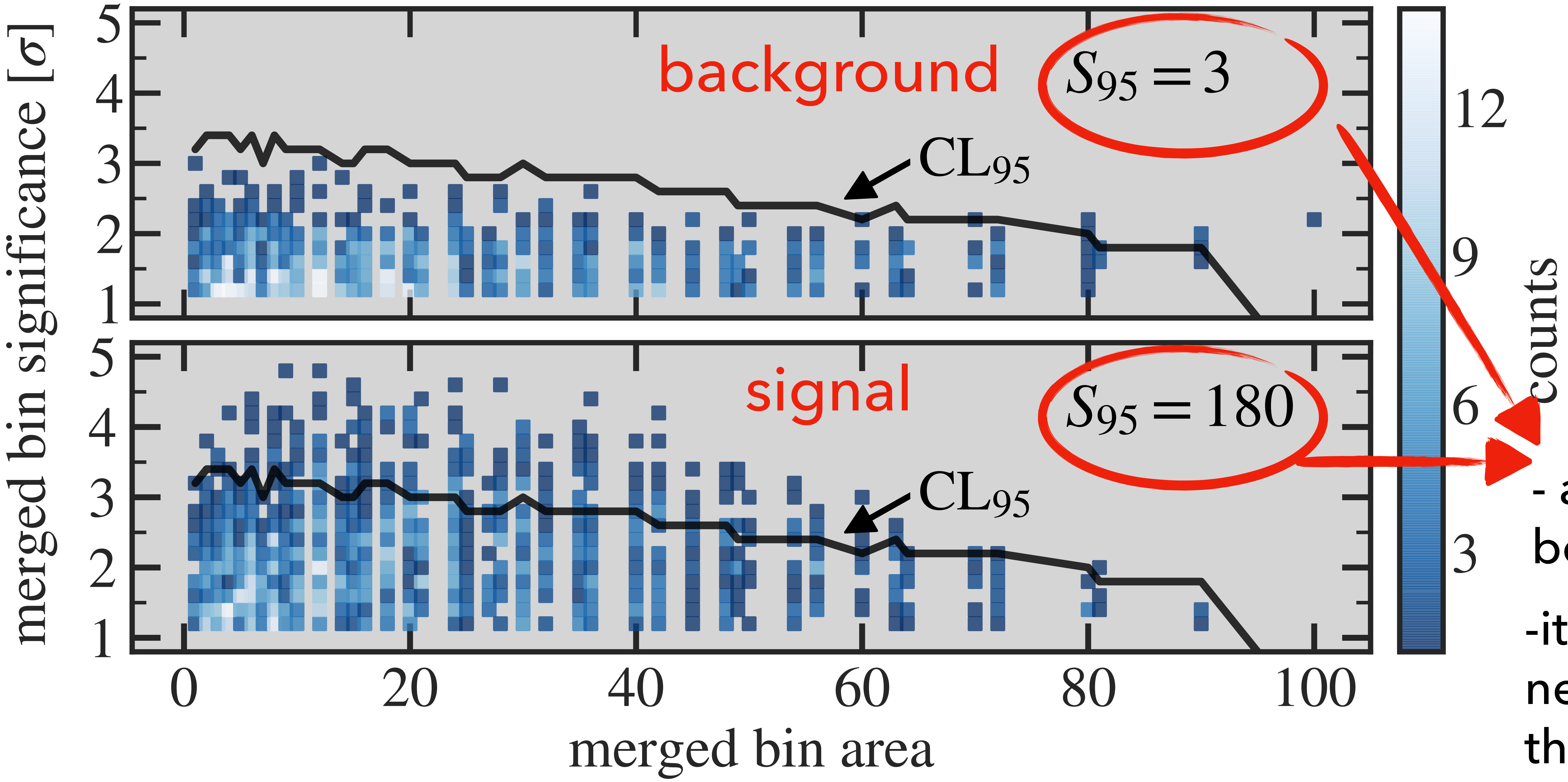


How do we summarize our pull/area findings?

> We will define a score test statistic, S_X, as the number of outliers that lie ***outside of a certain confidence level** X*, which we evaluate from the statistically fluctuated expectation.

> Example for X=95%:





- a single number for both BG and signal!

-it picks up the signal-ness well, at least for this realization of the signal.

Why the 95% level?

> No particular reason; **choosing the 95 in "S_{95}" is completely arbitrary**. We have run tests for a reasonable range of CLs and noticed a negligible difference between the true positive vs false positive rates for the different choices of CL.

> Also, the performance of this test statistic (as any other) will depend on the type of signal. We can't optimize for the CL beforehand because **we don't know (and don't want to know) what we're looking for in the data**.

> The most beautiful thing is… ***we can test it!***

Exercise 3

> We have put online some example code for running the score test statistic.

> The task of the exercise varies depending on how involved you want it to be. The options are:

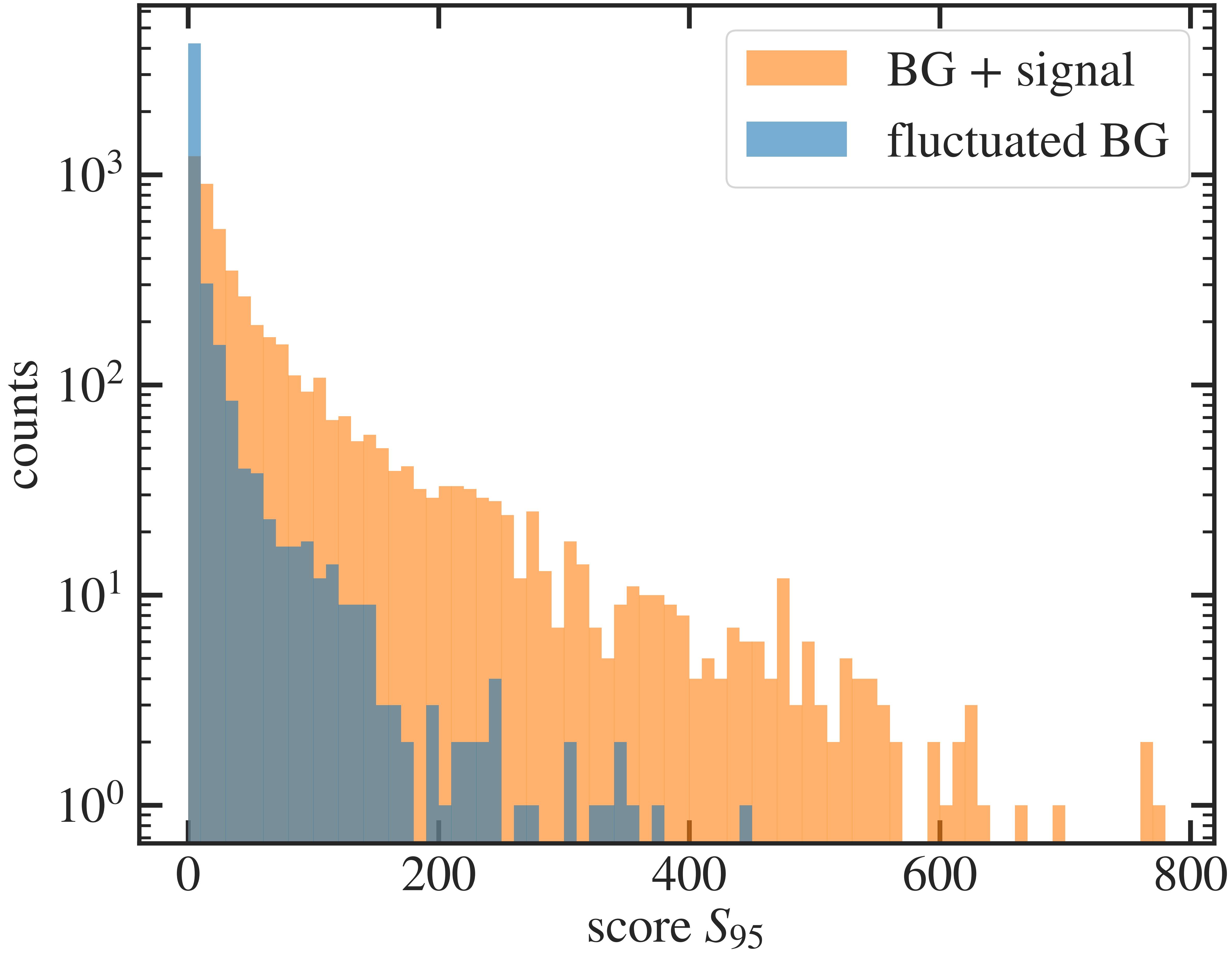
Option A. Download the entire code package, along with the pre-generated data (~200 MB), and run the score test on that pre-generated data as explained in the example notebook, examples/score_test.ipynb.

Option B. Run the score test from `/utils/score/score_TS.py` using the pre-generated data as the background (expectation), and inject your own signal (observation).

Option C. Code up the score test statistic yourself, generate your own expectations and observations, and compare the distributions of S_X for different X.

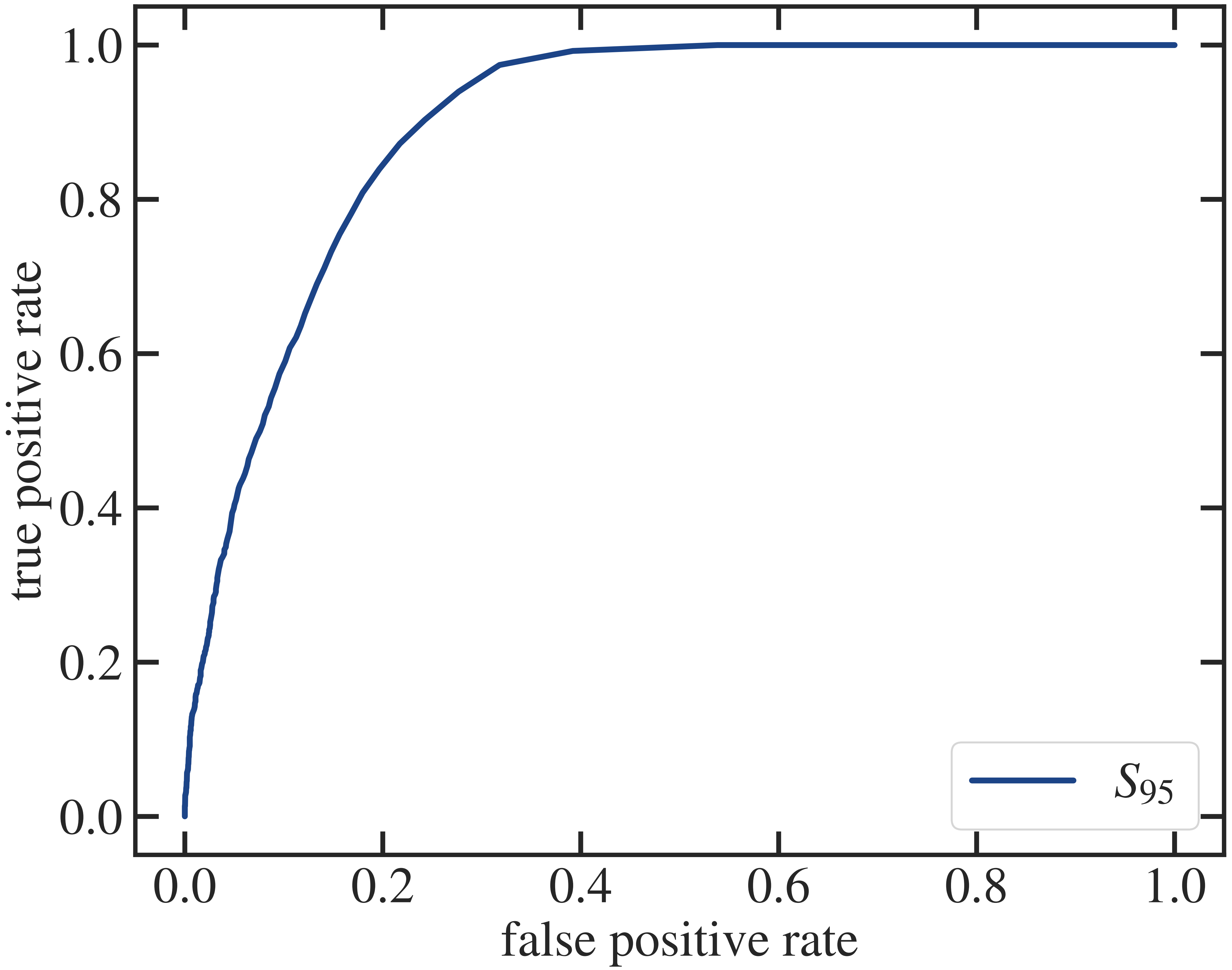
Exercise 3 "solutions"

> The distributions of the scores that I get from fluctuating the flat background map and the 2x2 4σ signal:



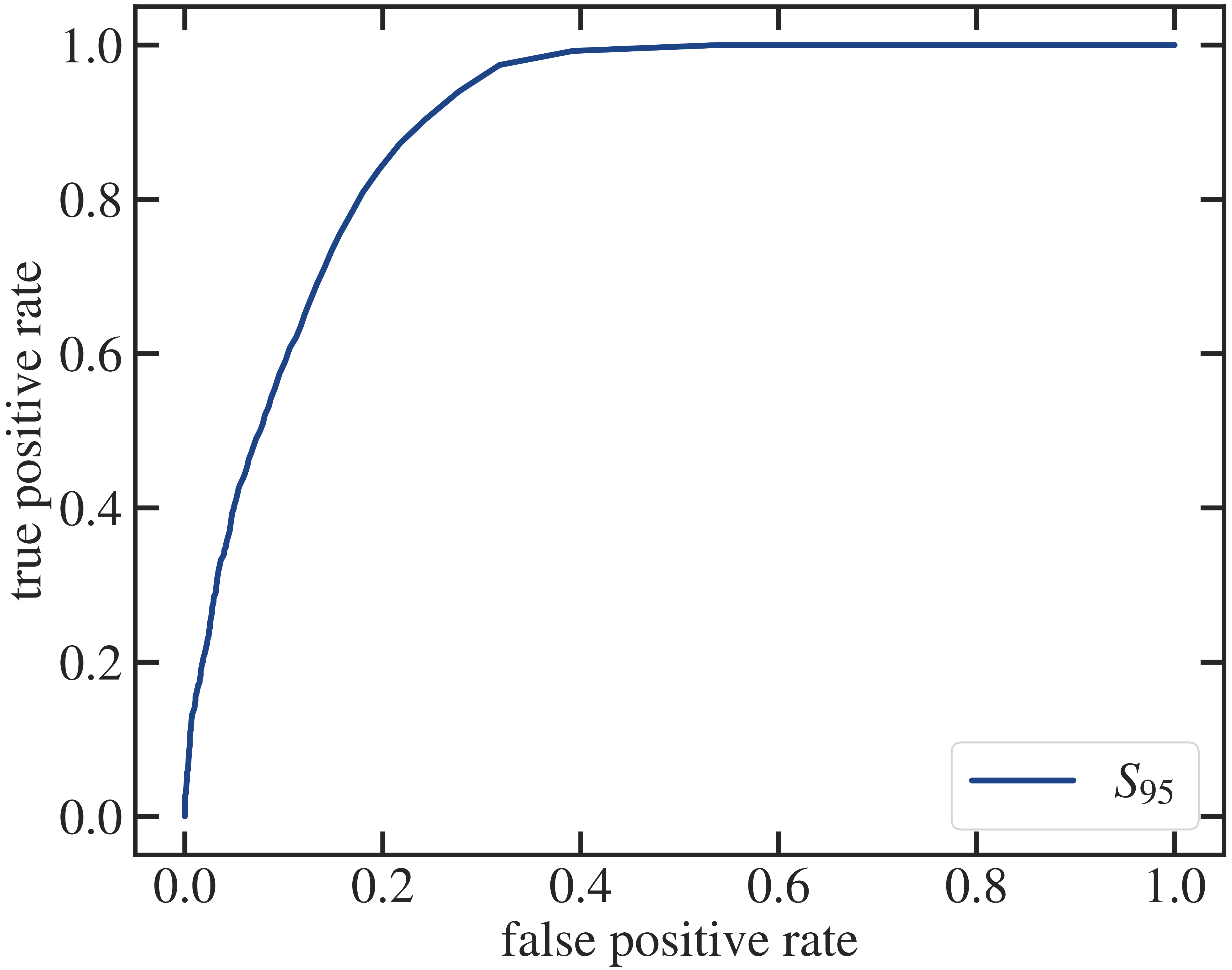
Exercise 3 “solutions”

> Depending on where we “cut” the score (assuming that the “signal” is everything above), we will get different true positive vs false positive rates. This is summarised in the ROC curve:

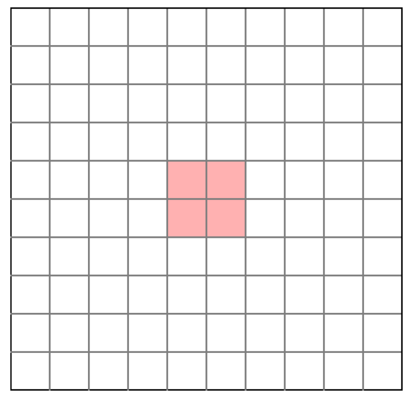


Exercise 3 "solutions"

> Depending on where we "cut" the score (assuming that the "signal" is everything above), we will get different true positive vs false positive rates. This is summarised in the ROC curve:



Remember: this is for our toy signal



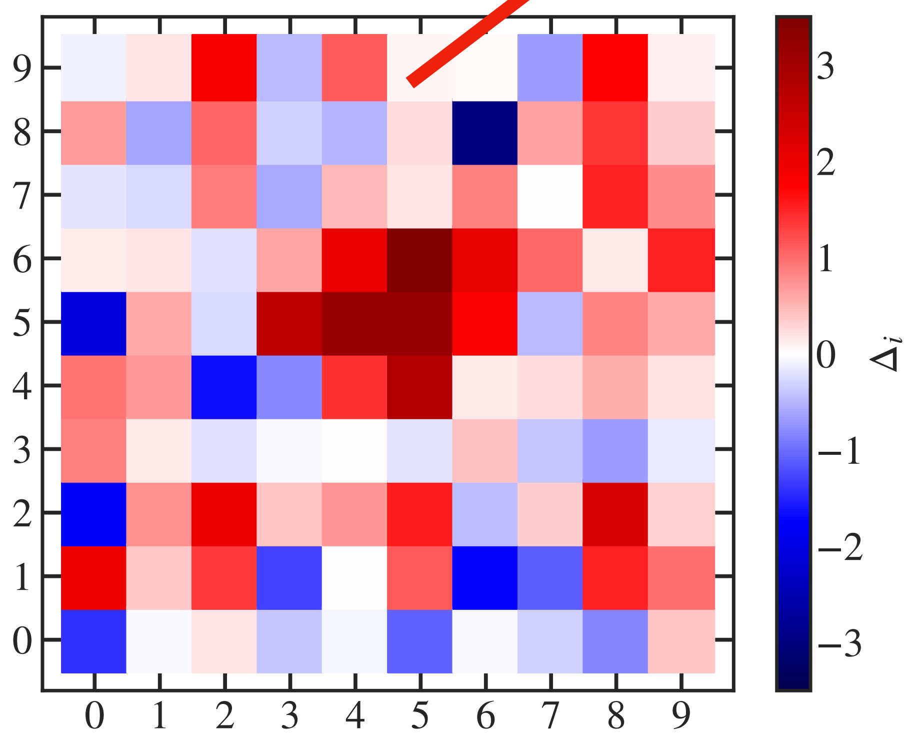
and will look different for other signals.

However, the method itself makes **no assumption about the signal shapes, locations, strengths, etc.**

Extra material

A different view of the binned data: Ising model[†]

s_i = +1(↑) or -1 (↓)



spin-spin interaction range

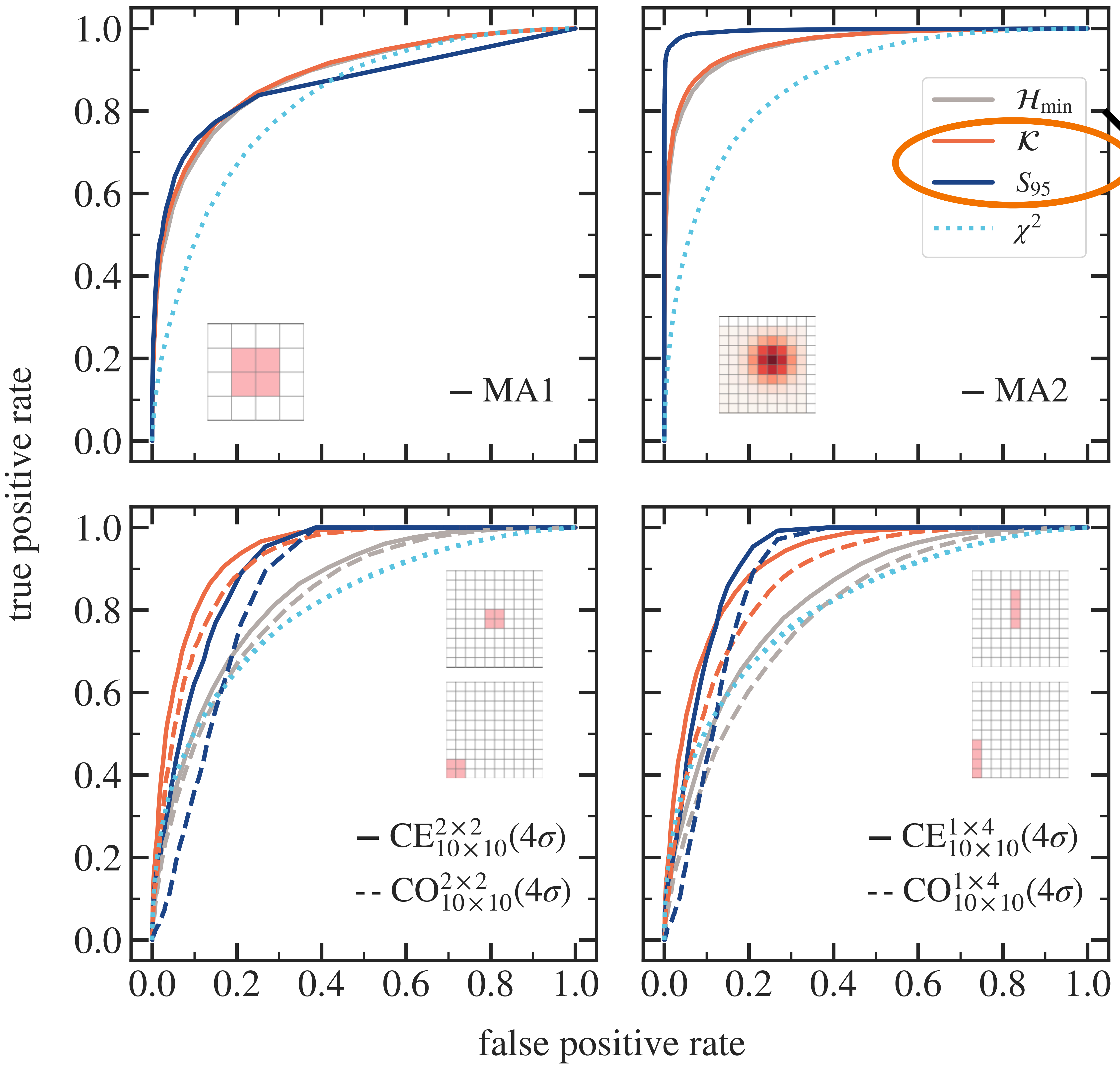
$$\mathcal{H}(\{\Delta_i\},\{s_i\}) = -\sum_{i=1}^{N} \frac{|\Delta_i|\Delta_i}{2}\,\frac{s_i}{2} - \frac{\lambda}{2}\sum_{i,j=1}^{N} w_{ij}\,\frac{(\Delta_i+\Delta_j)^2}{4}\,\frac{1+s_i s_j}{2}$$

deviations from the expectation [σ]

Idea: find the configuration of spins that minimises the Hamiltonian and use $\mathcal{H}_{\text{min}}$ **as a test statistic.**

[†] – Matchev *et al.* (arXiv:2003.02181)

Comparing the different test statistics: ROC curves



the most sensitive method to date;

the test statistics we derived, which beat the most sensitive method to date;

$\mathcal{K}$ is a variation of the Ising model Hamiltonian, which we will not go into.