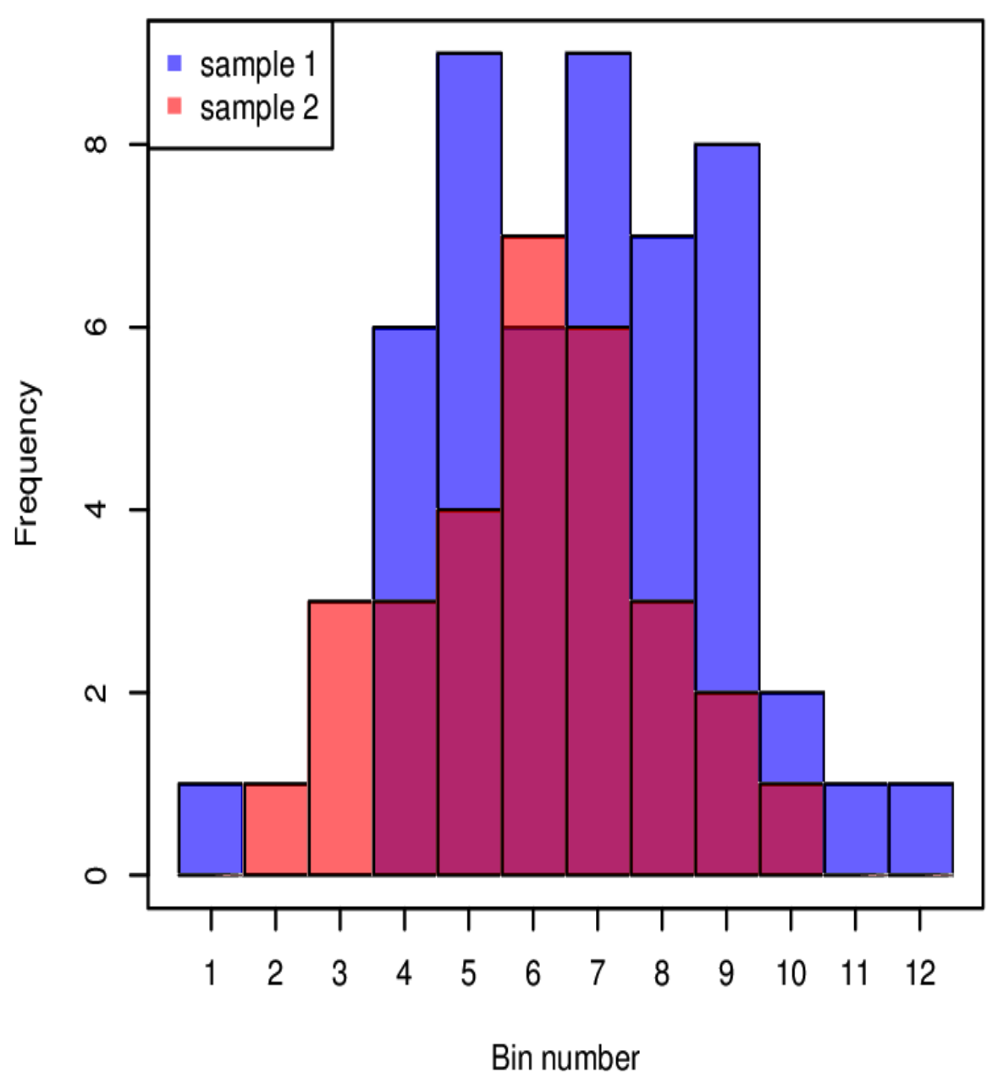
XI REUNIÓN CIENTÍFICA DE LA SOCIEDAD ESPAÑOLA DE ASTRONOMÍA
SEA
TERUEL 8 - 12 SEPTIEMBRE 2014

Do you need to compare two histograms not only by eye?

Nicolás Cardiel
Departamento de Astrofísica y Ciencias de la Atmósfera
Facultad de Ciencias Físicas
Universidad Complutense de Madrid



Avoid histogram comparisons!

The direct comparison of histograms built from continuous data is always a **bad idea** because the actual data have been replaced by the central values of the histogram intervals.

In addition, for a fixed data set, the derived histogram depends strongly on the choice of the bin width and the origin of the intervals.

Instead of comparing two histograms, the comparison should be done using any of the many two-sample comparison tests available in the literature, but over the original un-binned data!

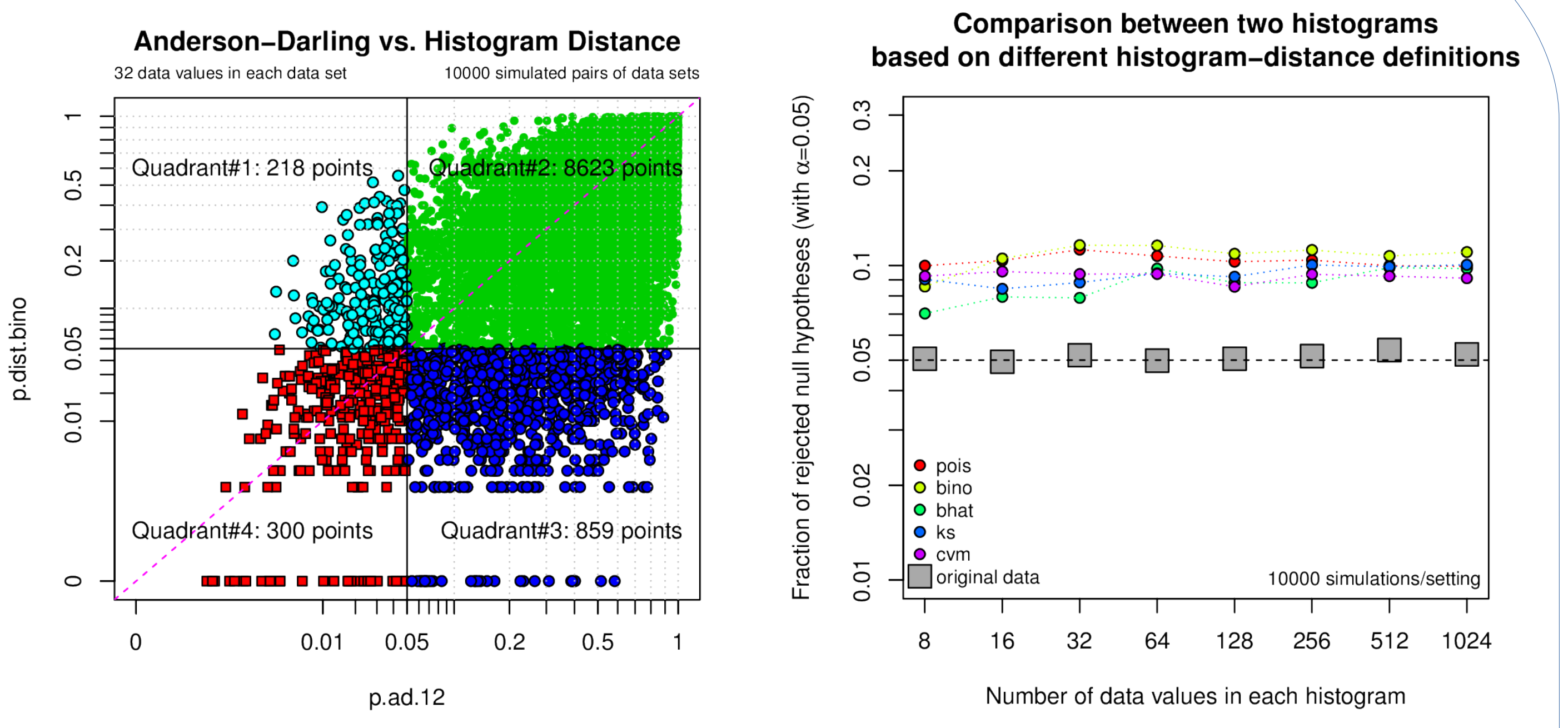
Anyway, histogram representations are frequently employed in scientific communication, particularly in Astrophysics. Sometimes its use is unavoidable when one needs to compare new results with already published data only available in histogram format.

It is not infrequent to find examples in the literature where the similarity between histograms is not statistically quantified but simply justified or discarded "by eye".

Using the "distance" between histograms

Focusing on the comparison of the global shape of histograms (using relative frequencies and ignoring the relative normalization), several methods are discussed in the literature (e.g. Porter 2008, arXiv:0804.0380). One starts by defining a suitable "distance" between the normalized histograms, and then estimates the probability distribution of such "distance" under the null hypothesis of equality of the histogram shapes.

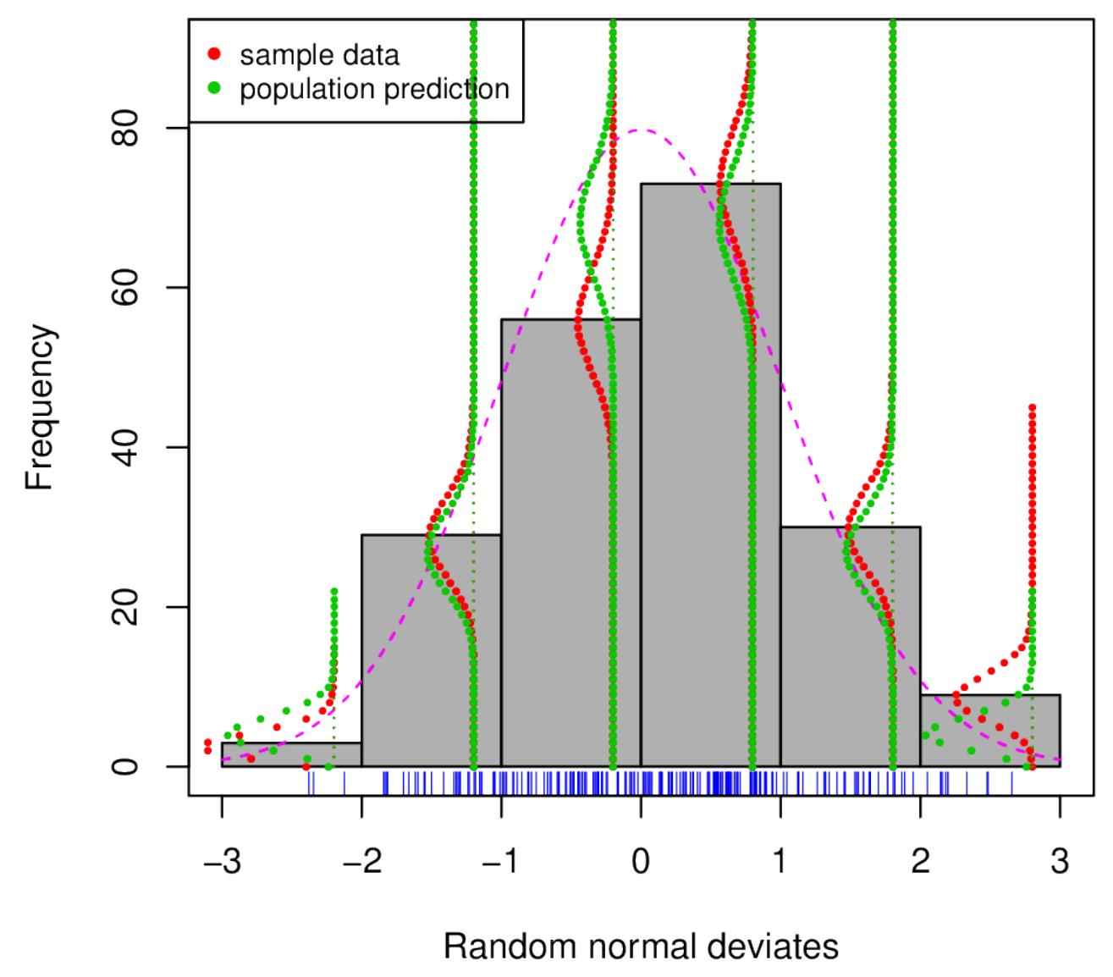
In this work we have explored the application of these methods to pair of histograms built from simulated data following a normal distribution, and compared their results with the application of the well-known Anderson-Darling two sample test for continuos data over the simulated un-binned data. In principle, the null hypothesis should be rejected a fraction of times given by α, the significance level, which is what happens when applying the Anderson-Darling test. However, it is found that, when applied to simulated continuous data following a normal distribution, the methods based on the "distance" between histograms tested in this work typically reject the null hypothesis in excess to the fraction α.



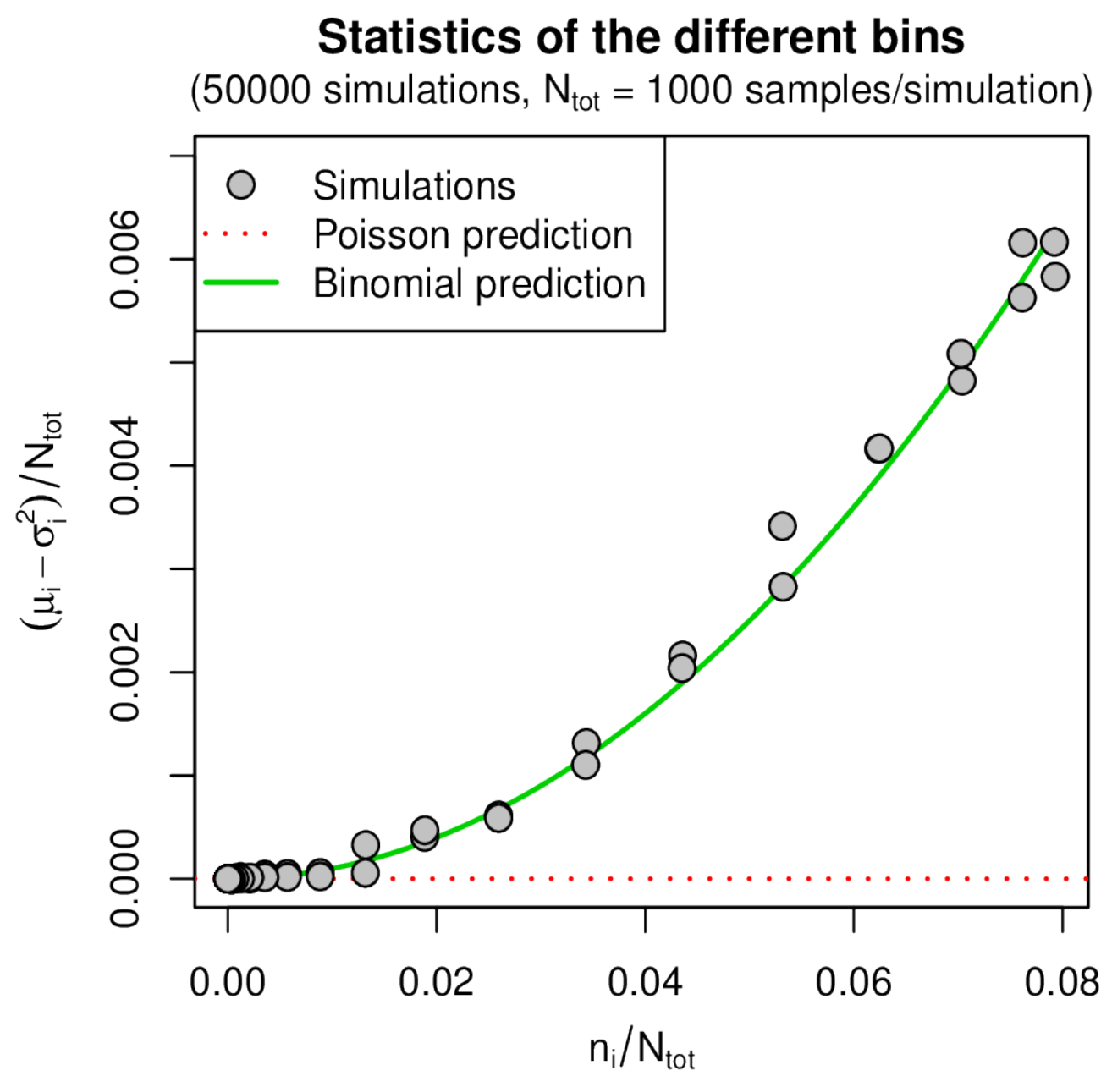
Left: Comparison between the p-values (probability of Type I error) obtained when comparing two simulated histograms using a "distance" based on assuming a binomial distribution of the bin frequencies (p.dist.bino) and when applying the Anderson-Darling test over the un-binned simulated data (p.ad.12). Each point corresponds to one of such comparisons. There is an excess of points in quadrant #3, which reveals that the "distance" method is rejecting the null hypothesis a fraction of times which is larger than the significance level. **Right:** The previous diagram was repeated for several "distance" definitions and different number of data values in each histogram. The figure represents the fraction of times the null hypothesis was rejected in each case by using α=0.05. The methods based on "distance" between histograms tend to reject the null hypothesis in excess to that value of α.

What is going on "inside" a histogram?

If one assumes that a given data set obeys a particular probability distribution, it is not difficult to show that the absolute frequency in any of the bins of a histogram built from this data set follows a binomial distribution (note that many people quote that it follows a Poisson distribution, which is only approximately true for bins with low absolute frequencies).



Left: Simulated histogram built from 200 random deviates following a normal distribution (magenta dashed curve). The green points represent (rotated 90°) the theoretical probability of obtaining any absolute frequency as annotated in the vertical axis. This probability follows a binomial distribution with p (success probability) given by the integral of the normal distribution between the limits of each interval of the histogram. Since the parent population distribution is expected to be unknown, the only estimation that one typically can derive is the probability distribution plotted in red, which corresponds to a binomial distribution where the value of p is assumed to be the relative frequency of each interval of the histogram.



Left: Analysis of the statistical behavior of the frequencies in each interval of a histogram. By simulating 50000 histograms corresponding to 1000 random normal deviates, the mean μ_i and variance σ_i^2 of the absolute frequency in each interval was derived.

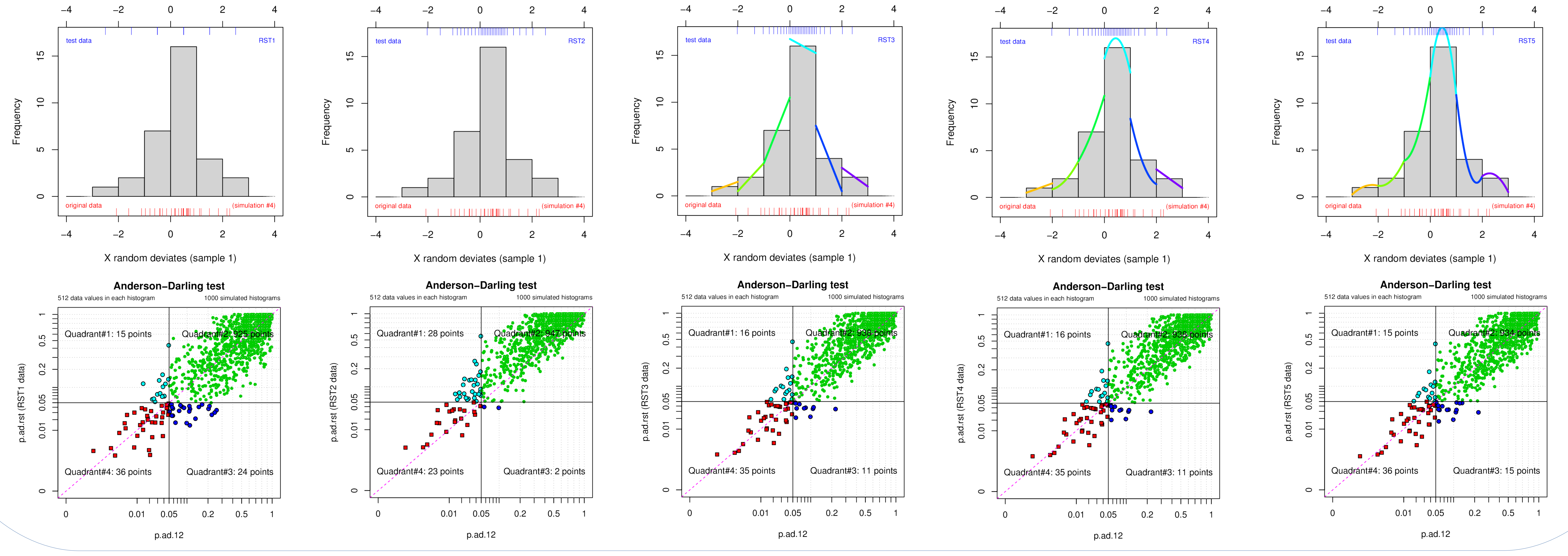
If the frequencies followed a Poisson distribution, one should expect $\mu_i = \sigma_i^2$ (red dotted line), which is not the case. On the contrary, the simulations reveal that the frequencies follow the prediction given by a binomial distribution (green line) with success probability in each interval given by the relative frequency $p_i=n_i/N_{tot}$.

Resampling the binned data within each interval

Even though the information about the exact location of the data within a given interval of a histogram was lost once the histogram was built, one can try to recover part of that information by using the frequencies of the neighbour intervals. If the adjacent intervals are not too noisy (i.e., their frequencies are not excessively small), one should expect that if the original data followed a continuous and not too complex distribution, the location of the data within a particular interval should be more clustered towards the border of the interval which is closer to the neighbour interval with higher frequency. Applying this idea, one can resample the data within each interval, and then perform the comparison between histograms by applying a traditional two-sample test, such as the Anderson-Darling (for those fans of the Kolmogorov-Smirnov test, see *Beware the Kolmogorov-Smirnov test!*, by E. Feigelson & J. Babu, at the ASAIP web site).

This work has explored this approach by using 5 different resampling strategies:
- RST1: the data in each interval is assumed to be exactly in the center of the interval
- RST2: the data is resampled uniformly
- RST3: a straight line is fitted passing through the center of 3 intervals and the data redistributed using the fit as a proxy to the probability density
- RST4: similar to RST3, but using a second order polynomial
- RST4: similar to RST5, but forcing the individual fits of the different intervals to connect at the borders of the intervals.

Note that all these methods do preserve the number of elements within each interval.

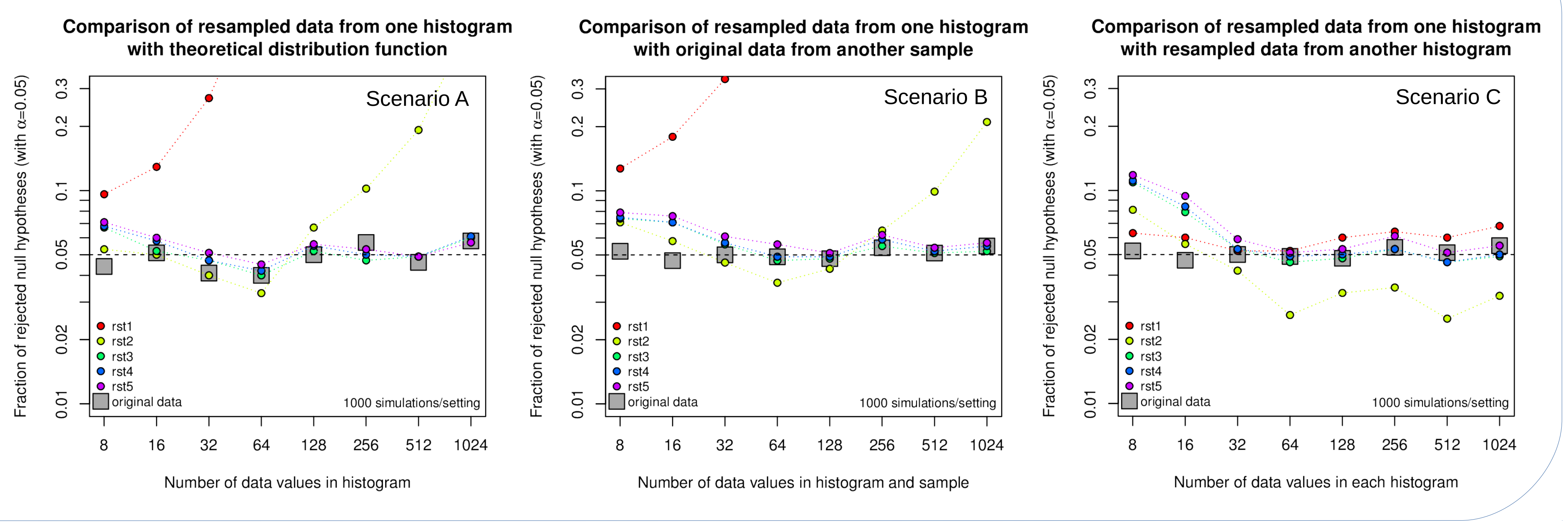


Results

By performing numerical simulations of histograms corresponding to random normal deviates, the five resampling strategies have been analyzed in different scenarios:

- Scenario A: comparison of one histogram data with a theoretical probability distribution.
- Scenario B: comparison of one histogram data with another un-binned sample data.
- Scenario C: comparison of two histogram data.

In all the cases RST3, RST4 and RST5 are in general better choices, specially for scenario A and B. Interestingly, The more complex resampling strategies are not specially good for scenario C when the histograms contains few data. This is not unexpected since in this situation the histogram intervals are too noisy.


Comparison of resampled data from one histogram
with theoretical distribution function
Scenario A
Comparison of resampled data from one histogram
with original data from another sample
Scenario B
Comparison of resampled data from one histogram
with resampled data from another histogram
Scenario C
Fraction of rejected null hypotheses (with α=0.05)
rst1
rst2
rst3
rst4
rst5
original data
10000 simulations/setting
Number of data values in histogram
Number of data values in histogram and sample
Number of data values in each histogram