The Two sample problem

Roger Barlow

11th November 2014

The general problem

Are two samples, A and B, compatible with being drawn from the same parent distribution, or is there a difference?

As we usually meet it in particle physics

- Data is histogrammed (occasionally 2D scatter plots)
- Overall size doesn't matter, only the shape.
- Statistics are high, for at least one of the samples
- No preconceived ideas about what the difference might be
- Very common at early stages of analysis
- No scope for pairing/stratification

Typically general data quality checking:

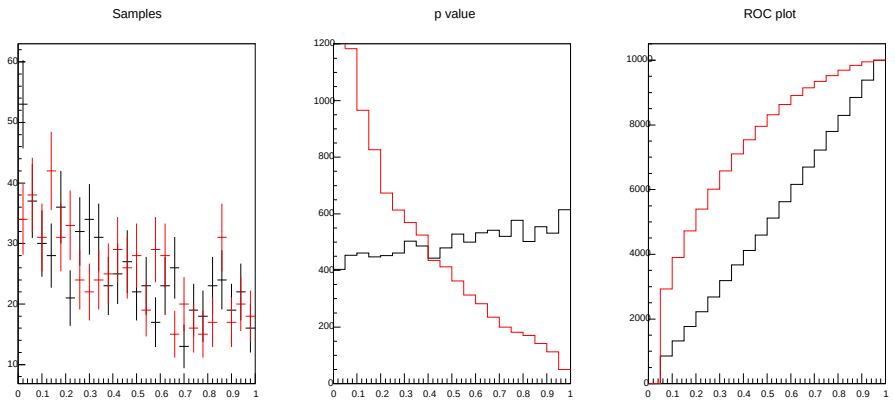
- Data v. Simulation
- Magnetic field up v. down.
- Run 1 v. Run 2

Tests: the general framework

Many available: χ^2, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Run test...

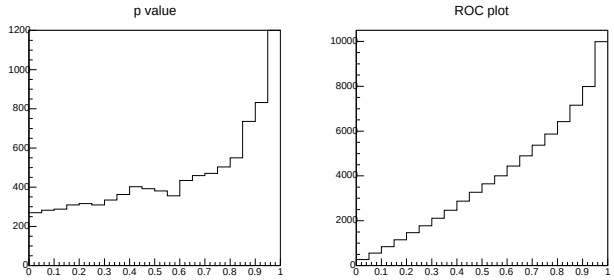
1. Invent a measure of agreement. Call it t
2. From the pdf for t, find $p(t)$ giving probability for agreement between two samples from the same parent being t or worse.
3. A histogram $h_0(p)$ should be flat. Think about this until it is obvious.
4. To test against any alternative hypothesis H_1, histogram $h_1(p)$:
5. Plot the cumulative frequency $\int_0^p h_1(p')dp'$ and get the ROC plot

2 (different) samples: $P_0 \propto e^{-x}$ and $P_1 \propto e^{-1.25x}$. $t \equiv \chi^2$
histograms of $h_0(p)$ (black) and $h_1(p)$ (red); ROC plot



The Kolmogorov-Smirnov test

Easy to use: `h1 -> KolmogorovTest(h2);` returns probability.

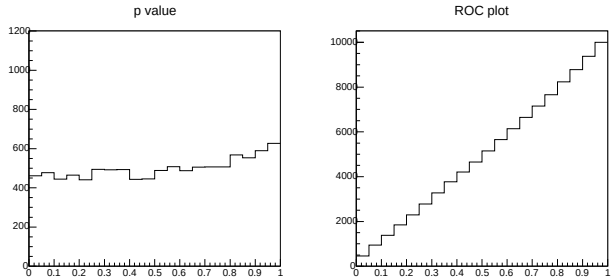


from histos of 1000 events in 25 bins.

Computed histogram of p is not flat!!

Reason: KS test handles ranked data. Not valid for 'tied' data points.

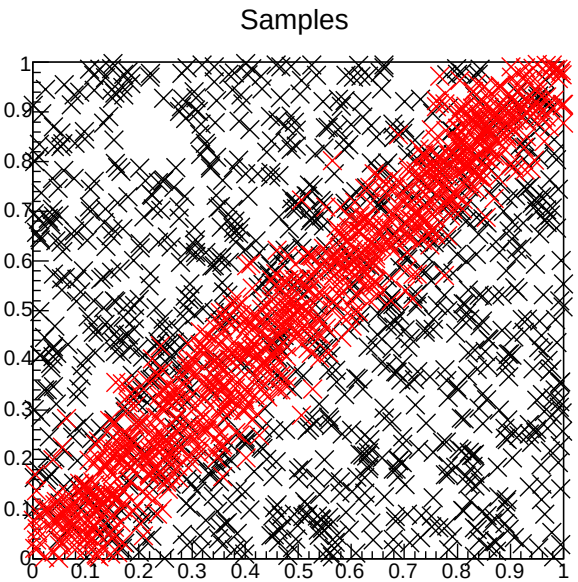
Histogram binning introduces tied events. Results not valid.

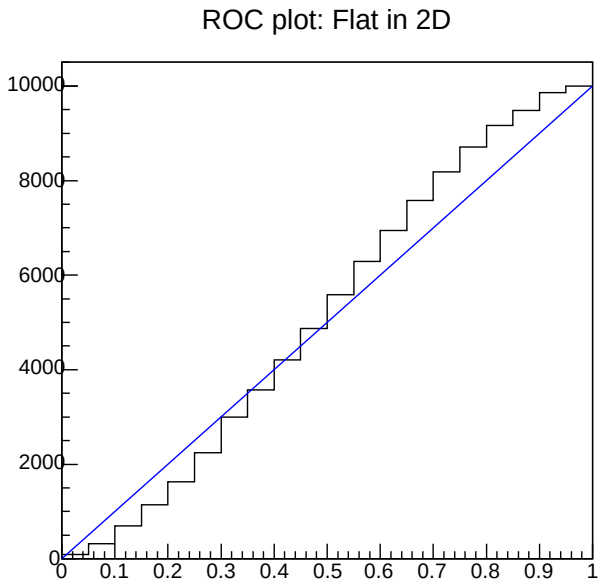


histos of 1000 events in 10000 bins.

KS Test valid iff $N_{bins} >> N_{data}$

The 2D Kolmogorov-Smirnov test

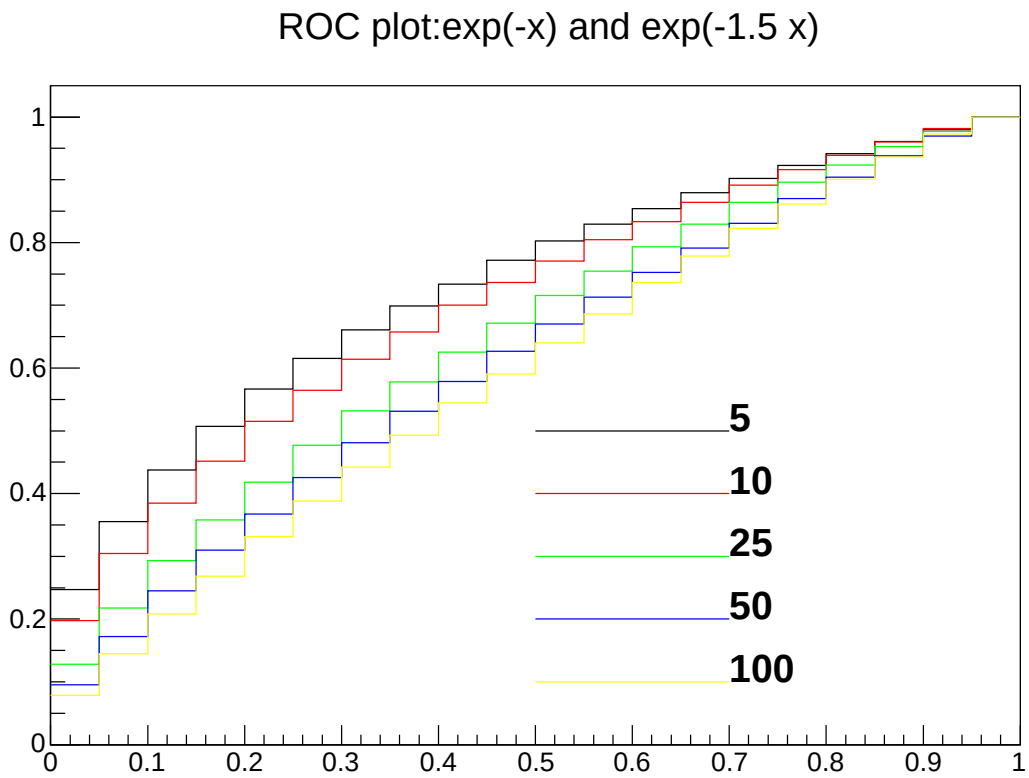




Actually two 1D tests. Looks only at projections.
Datasets on left have probability 14% of agreement
Worse: takes average of max disagreement. Even with flat-flat distributions, ROC is not straight line.

Binning and the χ^2 test

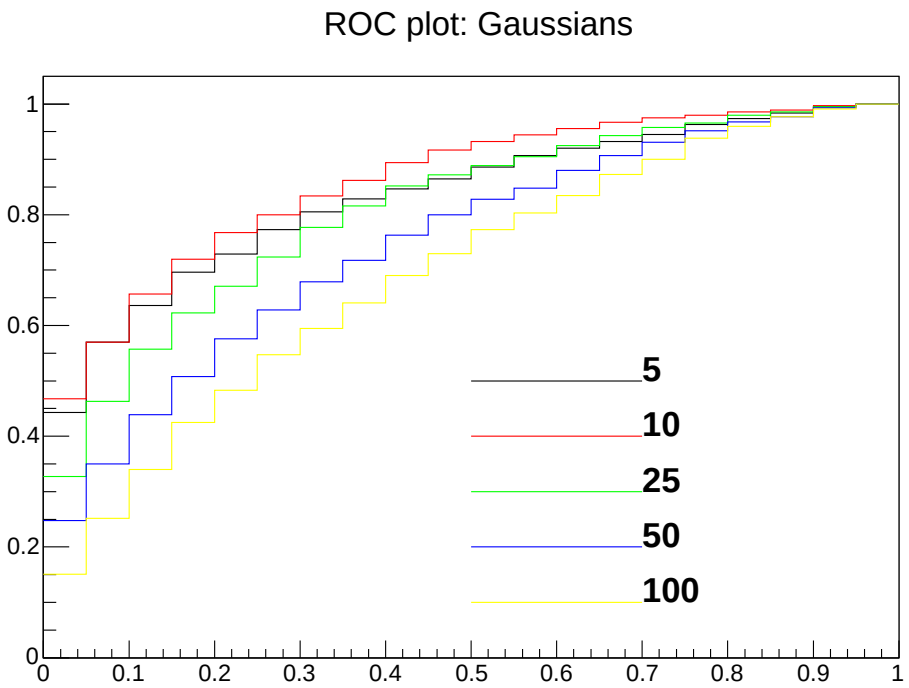
Power of e^{-x} against $e^{-1.5x}$ over $[0, 1]$, for 1000 events in various binnings



The fewer bins, the better...

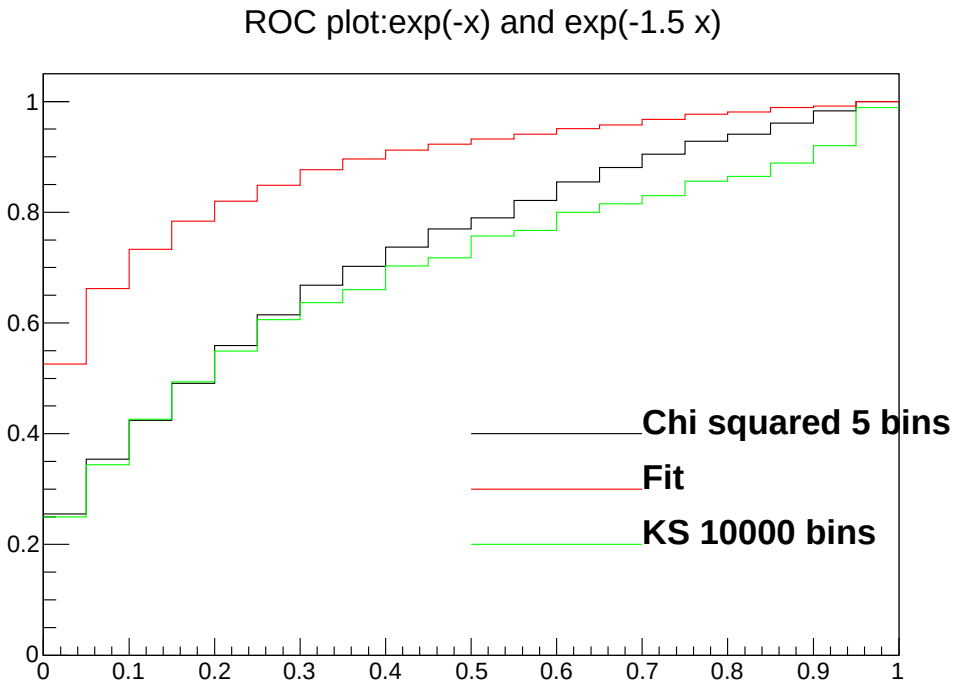
Binning again: χ^2 test

Uniform +Gaussian ($\mu = 0.70, \sigma = 0.1$) against Uniform+Gaussian ($\mu = 0.73, \sigma = 0.1$)



... unless the difference is in the fine detail

Comparison: χ^2 v. KS



Exponential model: best χ^2 and KS about the same.
Compare against fit to $e^{-\alpha x}$ and test α for compatibility: much stronger (but not general)

The Run Test

Forget it.

Some thoughts

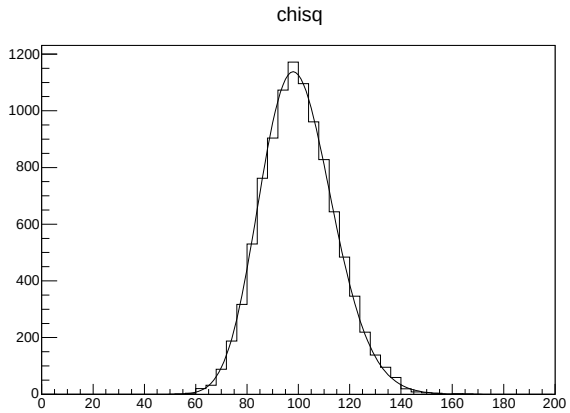
To discuss the 'power' of a test you have to invent an alternative hypothesis. Violates assumption (4). Probably need to consider several different alternatives for intelligent discussion

KS works by (1) ranking the two datasets (2) considering their cumulative distributions (3) finding the largest deviation: that is t. Throws away information from the metric. (AD, CvM similar).

For χ^2, the test does not know which bins are adjacent. High values from a few local bins can be swamped by the rest.

With the two most popular tests clearly lacking something, should be scope for improvement

Bootstrap estimation of p value functions



Comparing χ^2 for 100-bin histograms

Curve is obtained from χ^2 distribution for 100 bins

Histogram from many calls to :

- Generate H1
- H0$->$FillRandom(H1);
- H2$->$FillRandom(H0);

Evaluate $t = \chi^2$ for H1 and H2. Integrate numerically to get p function

Will work for ANY test! No knowledge of parent required.

Note that 2 stages of sampling are necessary.

Summary

1. Only use the 1-D KS test for sparse histograms
2. Never use the 2-D KS test
3. Use χ^2 with very large bins
4. If you have any information about likely differences, use it
5. Devise (and calibrate) your own tests for your own problem

Failure to observe these simple rules will weaken your 2-sample tests.
Significant differences will escape you.
.... maybe that's what you want? Hope not.