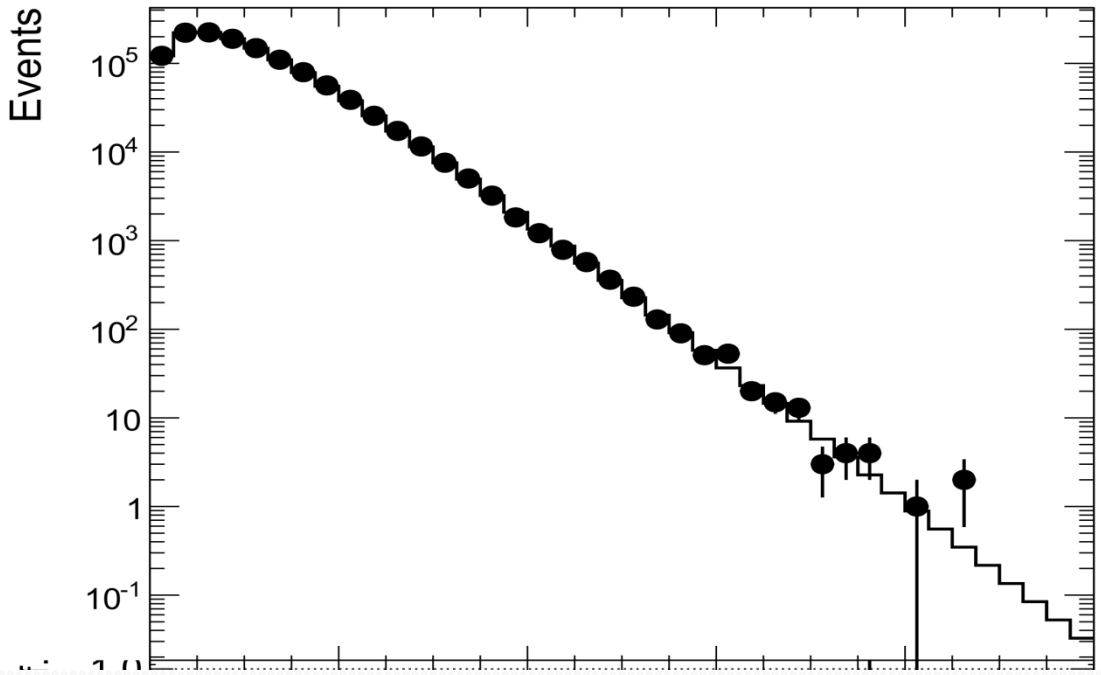
Plotting the Differences between Data and Expectation

Seminar Talk by
Laura Zani (University of Pisa) and Riccardo Di Maria (University of Bologna)
HASCO Summer School 2014

Let's start from a PLOT

Some few questions:

- Can you have a precise and immediate visualization of deviation of data?

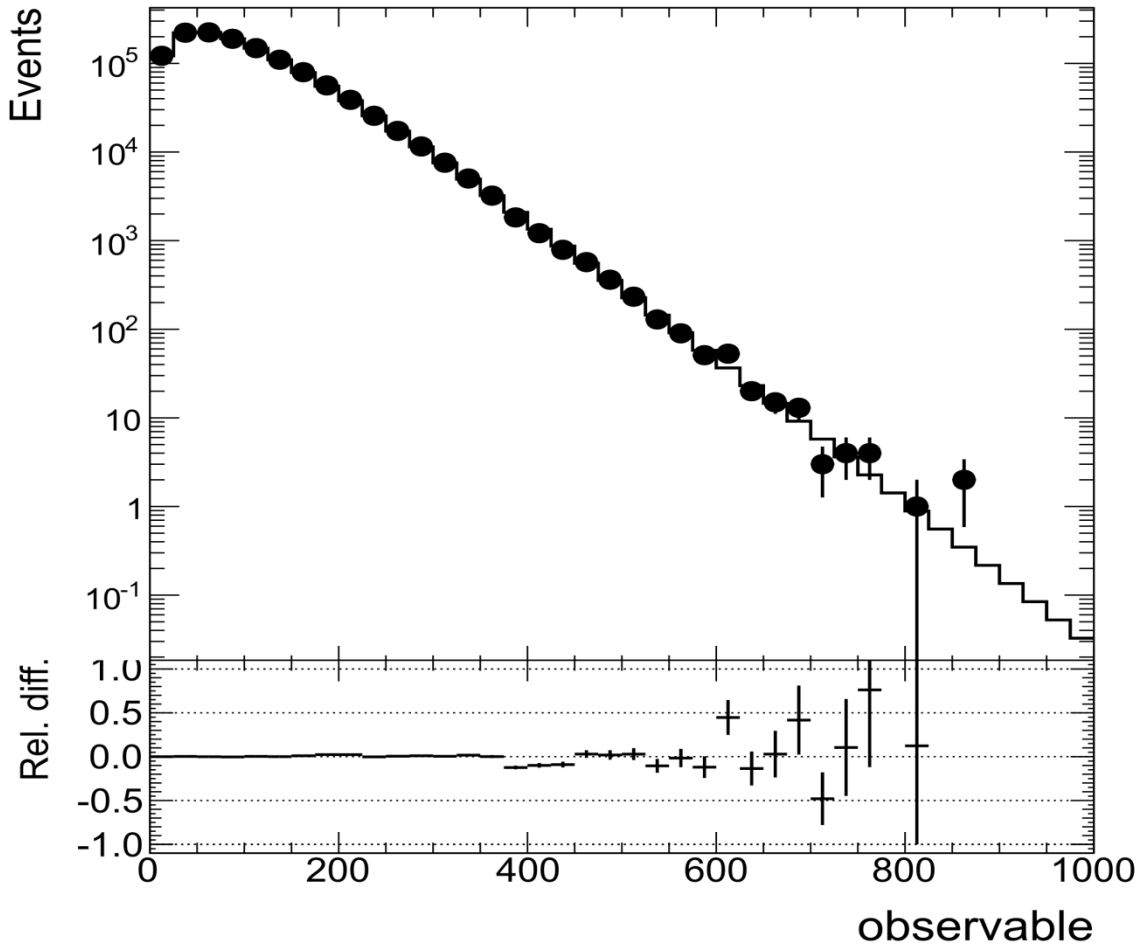


Let's start from a PLOT

Some few questions:

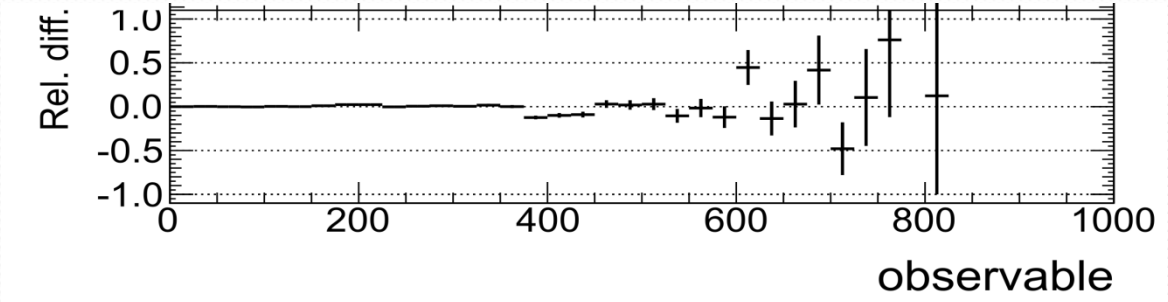
- Can you have a precise and immediate visualization of deviation of data?

Does the INSET PLOT improve the representation of deviations?



Let's start from a PLOT

Is this kind of INSET PLOT the most efficient way to show significance?

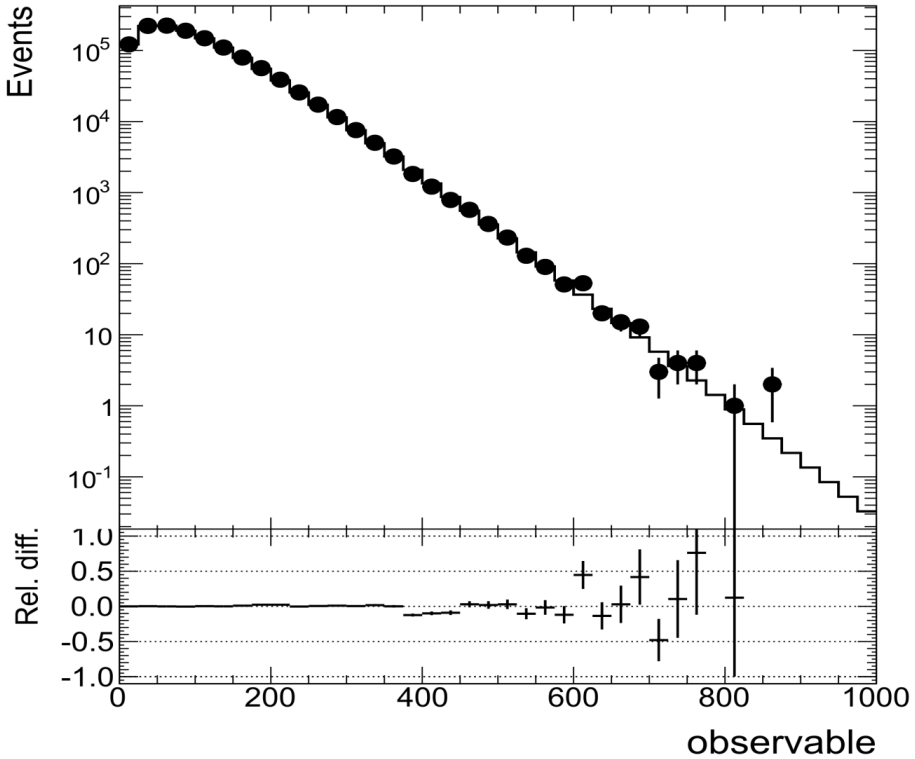


Let's try to focus on different types of

INSET PLOT

Purpose

- The paper focuses on the searching of the most efficient way to present an ***INSET PLOT***
- An inset plot is usually inserted at the bottom of the histogram to explicit the ***SIGNIFICANCE*** of data



Outline

- Motivation
- Statistical significance
- Present deviations in Poisson distributed data
- Considering theoretical uncertainty
- Summary

Motivation

- Histograms in logscale
  - Span lots of orders of magnitude
  - Hide differences → HIDE SIGNIFICANCE!

INSET PLOT

- Intuitive

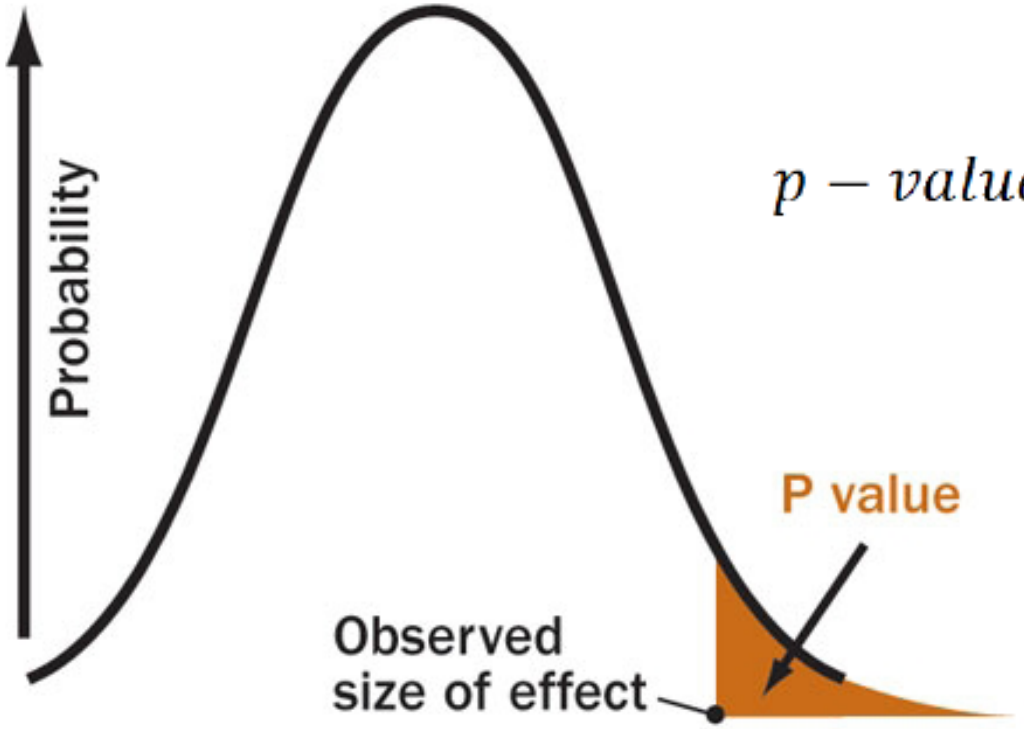
excess/deficit of data must be evident

- Accurate

representing the ACTUAL significance deviation for each bin

Statistical Significance

- DEF: The probability of finding a deviation **AT LEAST** as big as the one observed in data
- *p-value* ⟶ *z-value* (deviation espressed in units of σ)



$$p - value = \int_{z-value}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Statistical Significance

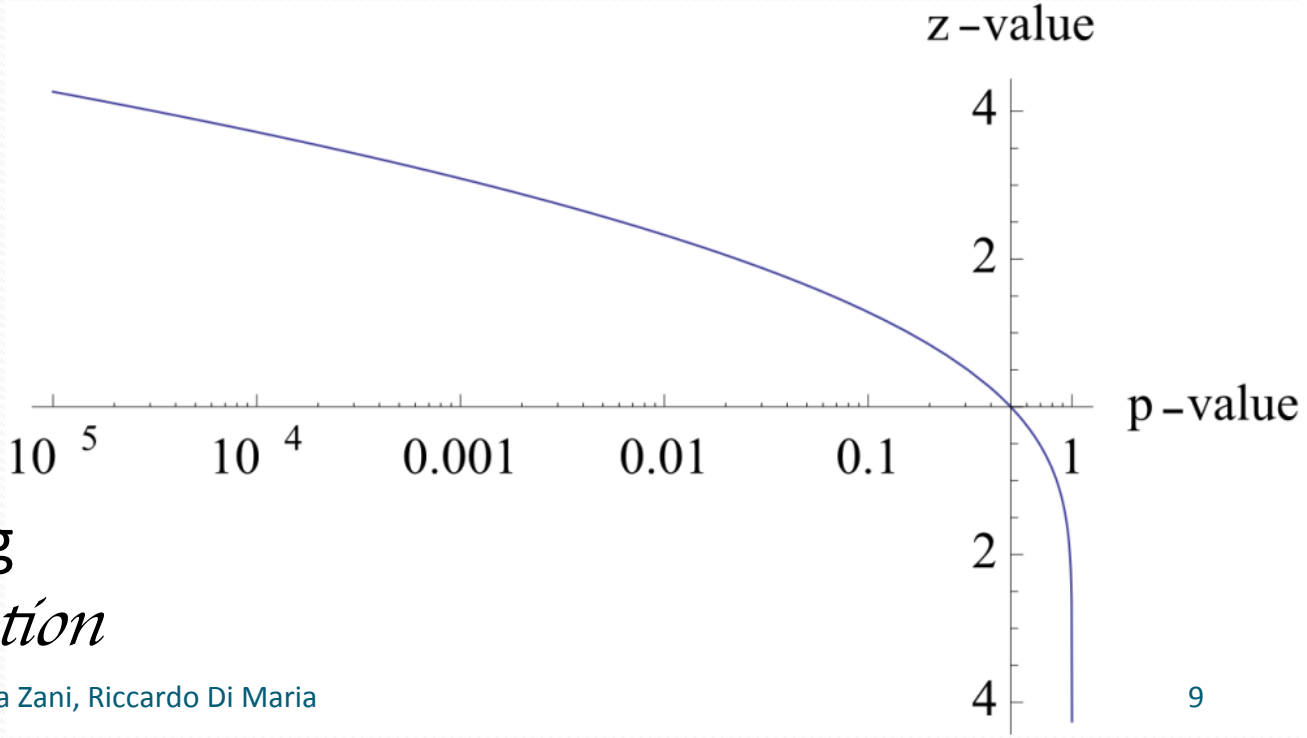
- p-value ≤ 0.5 ⟶ z-value ≥ 0
- p-value > 0.5 ⟶ z-value < 0
  - e.g. $p-value = 2.87 \cdot 10^{-7} \longrightarrow z-value = 5$ 5σ effect

Discovery

$z-value = 1-2$

uninteresting

common fluctuation



How to present DEVIATIONS

- Different types of inset plot
  - The $\frac{D}{B}$ Ratio (relative differences)
  - The $\frac{D-B}{\sqrt{B}}$ Approximation
  - Signed z-values
  - *Final proposal* : plotting signed z-values only if p-value < 0.5

The D/B Ratio

- Assuming Poisson - distributed data for each bin
- $R = \frac{D - B}{B}$ immediately show: EXCESS of data, R > 0
  DEFICIT of data, R < 0
- DISADVANTAGES:
  - several orders of magnitude → significant deviation hidden
  - wrong impression of larger fluctuation for low-population bins

WORST WAY TO COMPARE TWO HISTOGRAMS, even if not "wrong"

It is just not the right "metric" to express the distance *D* - *B*, if you want to show immediately the significance of data

The D/B Ratio

- e.g. Considering two histograms with the same R:

- B = 10
- D = 11

$$\longrightarrow \quad R = \frac{D - B}{B} = 0.1$$

$$\sigma = \sqrt{B} = \sqrt{10}$$

deviation < σ

just fluctuation

- B = 10000
- D = 11000

$$\longrightarrow \quad R = \frac{D - B}{B} = 0.1$$

$$\sigma = \sqrt{B} = 100$$

deviation = 10 σ

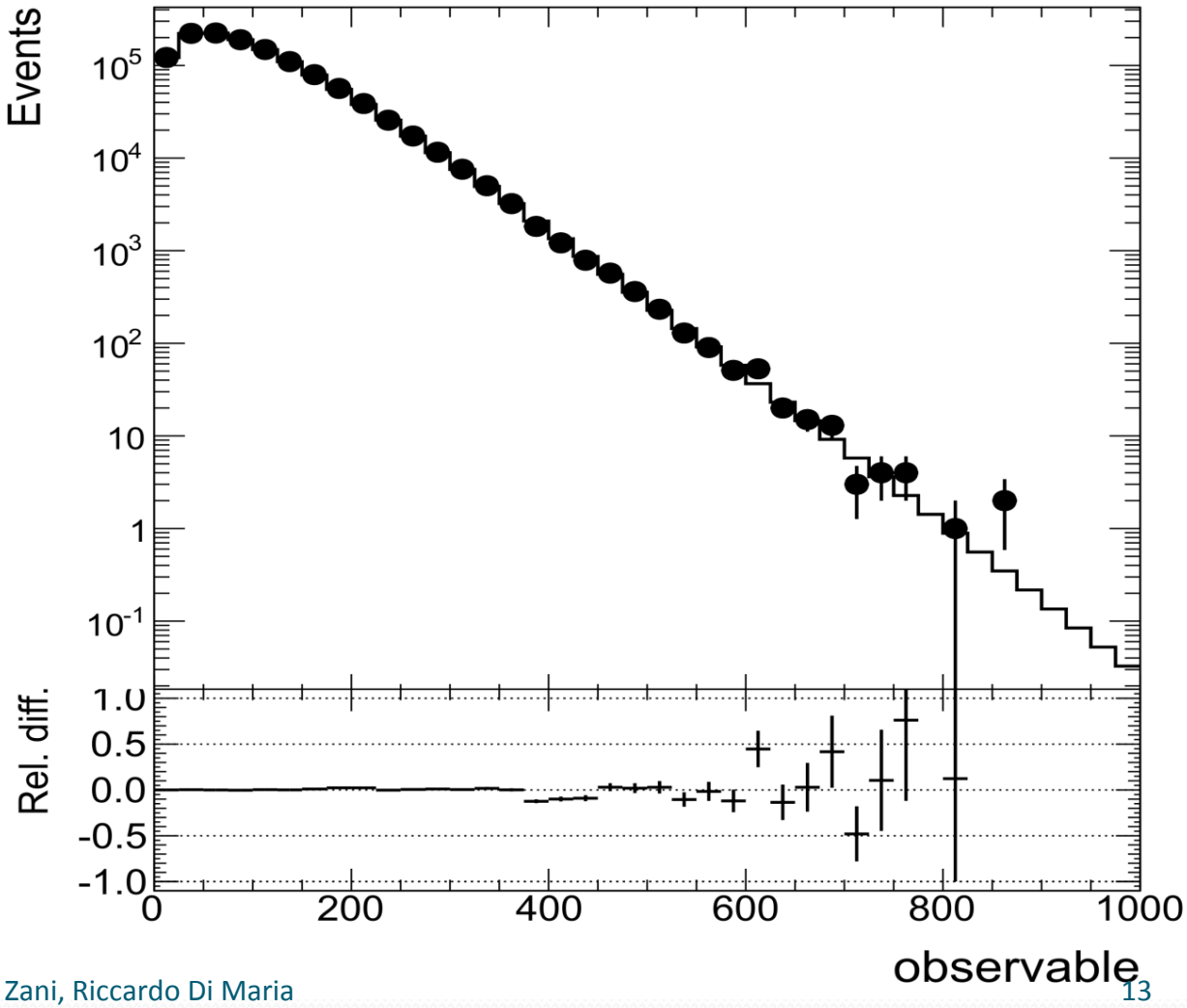
Discovery

The D/B Ratio

Something a bit "*surprising*":

people continue to use this kind of inset plot (also others) as though they know the drawbacks…

So BE CAREFUL!



The $\frac{D-B}{\sqrt{B}}$ Approximation

- Large ***B***: Poisson → Gaussian
  → $z\text{-}value$ can be approximated by $\frac{D-B}{\sqrt{B}}$
  - $mean = B$
  - $deviation = \sqrt{B}$

ADVANTAGE: Differences between data and expectation now expressed in units of σ
→ Significant deviations clearly visible!

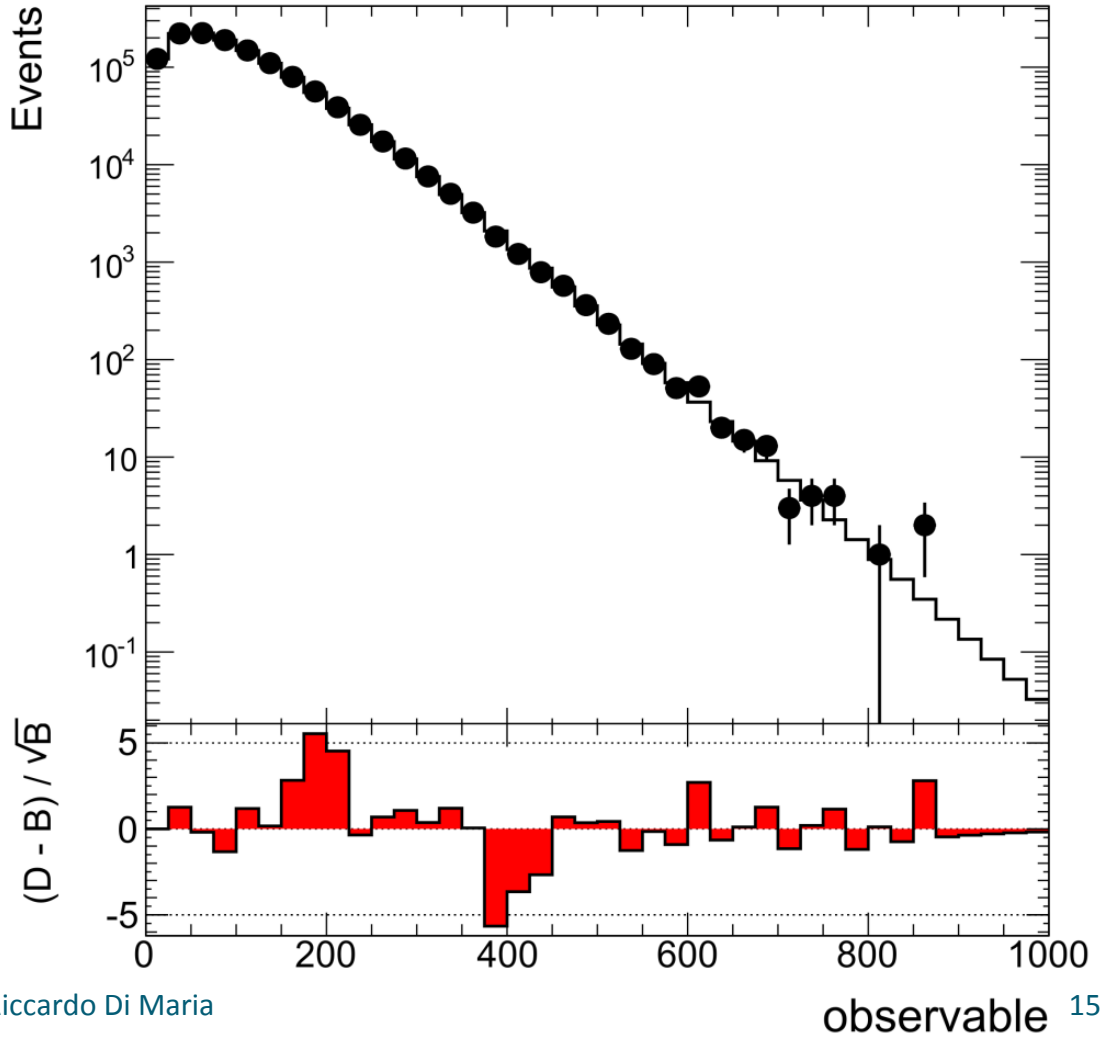
DISADVANTAGE: NOT a good approximation for *low-population bins*

The $\frac{D-B}{\sqrt{B}}$ Approximation

Approximation

FAILS

in the last bins!



→ NOT SIGNIFICANCE
as it appears in the histogram

Plotting signed *z-values*

Plotting the exact z-values:

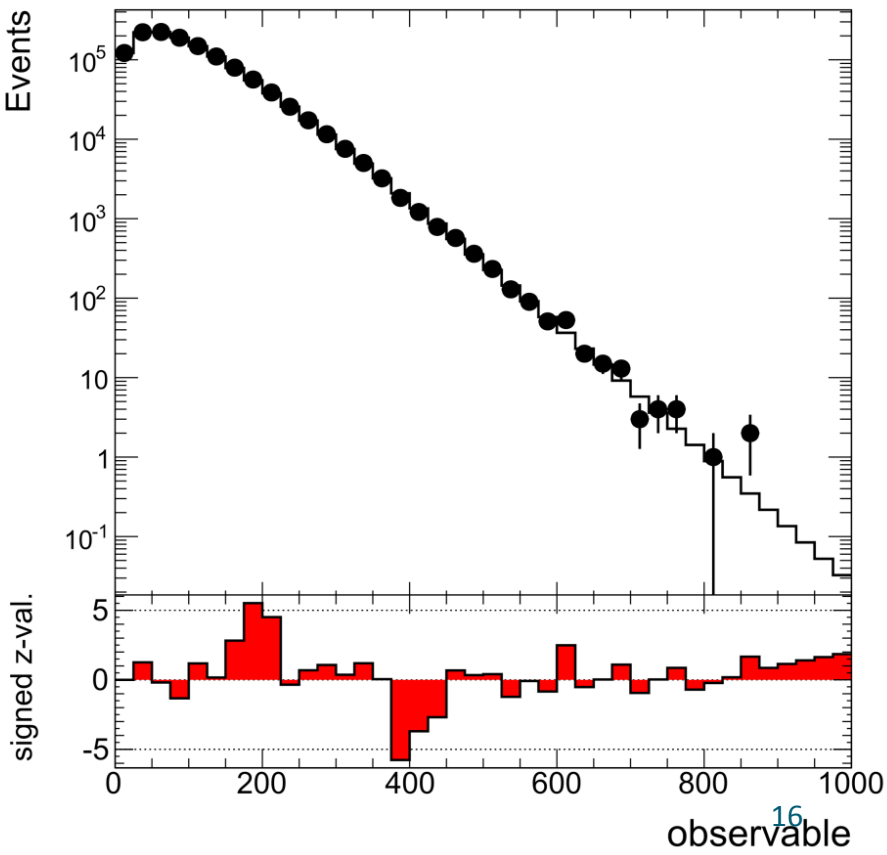
- D > B (EXCESS) ⟶ *+z-value* *(p-value < 0.5)*
- D < B (DEFICIT) ⟶ *- z-value*

DISADVANTAGES

(there is still something confusing)

Problems with *LOW STATISTICS*:
very **insignificant deficits** have *negative z-value* ⟶ with the *sign-flipping* they wrongly appear as an **excess**

(*see last bins of the histogram)*



Final proposal

plot

signed z-values

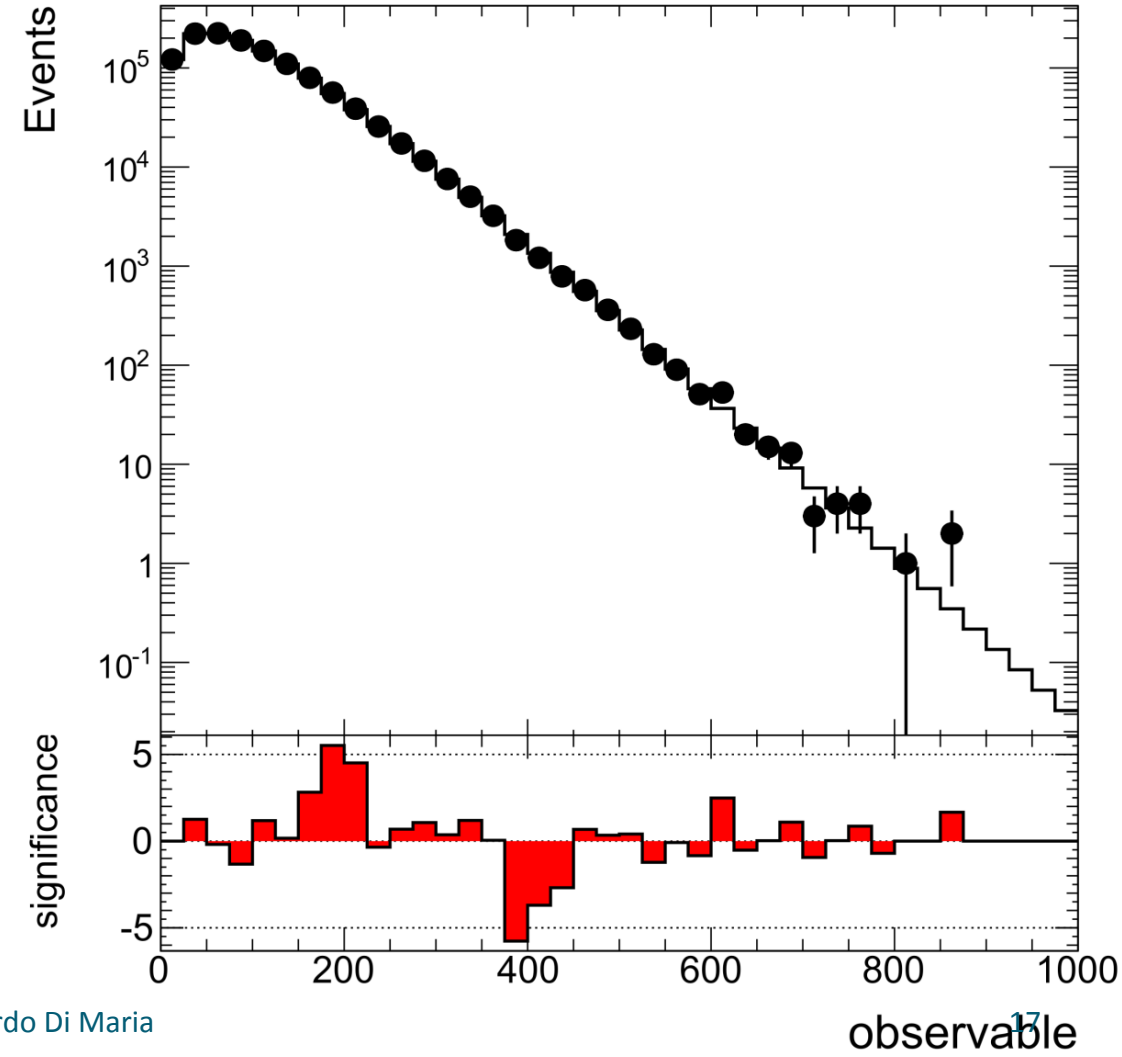
only if

p-value < 0.5

- bins with *p-value* > 0.5 perfectly agree with expectation

  ⟶ uninteresting

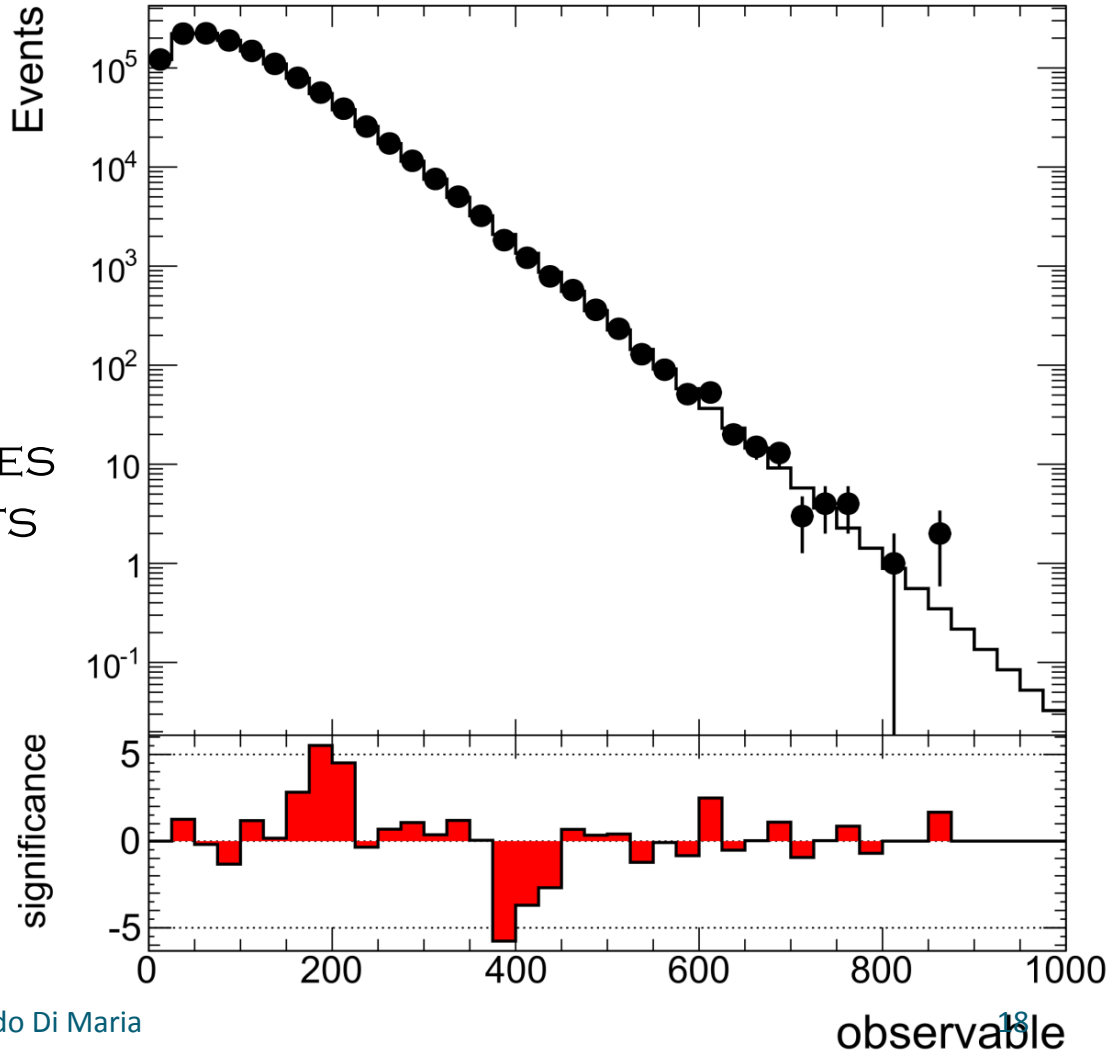
- Even misleading, as just seen before: no need to plot them!



Final proposal

z-value

- *Accurate*
  not an approximation
- *Intuitive*
  positive values represent EXCESSES
  negative values represent DEFICITS
- No significant deviations hidden
  (p-value < 0.5 always shown)
- Same treatment of bins with high and low statistics

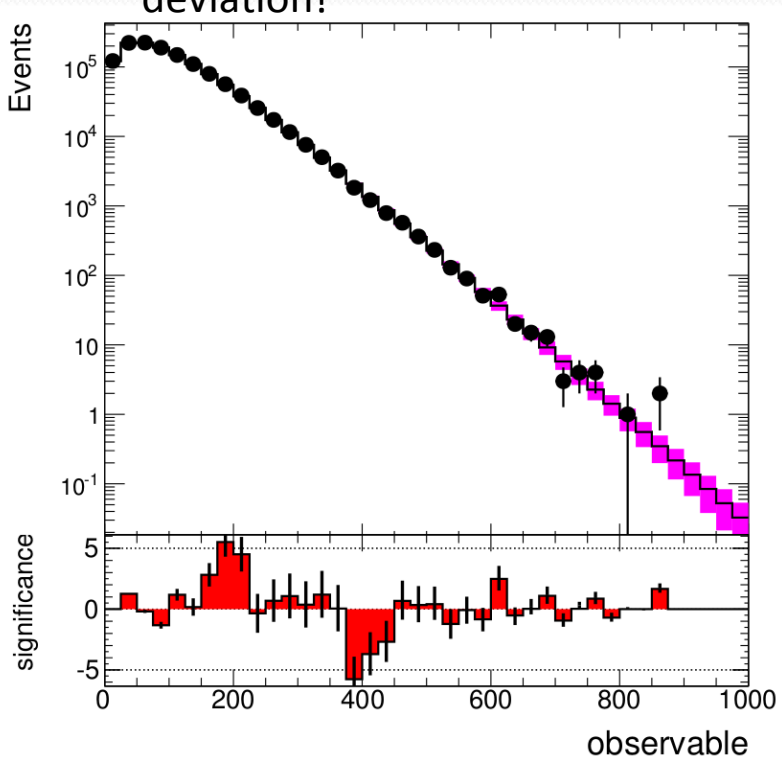


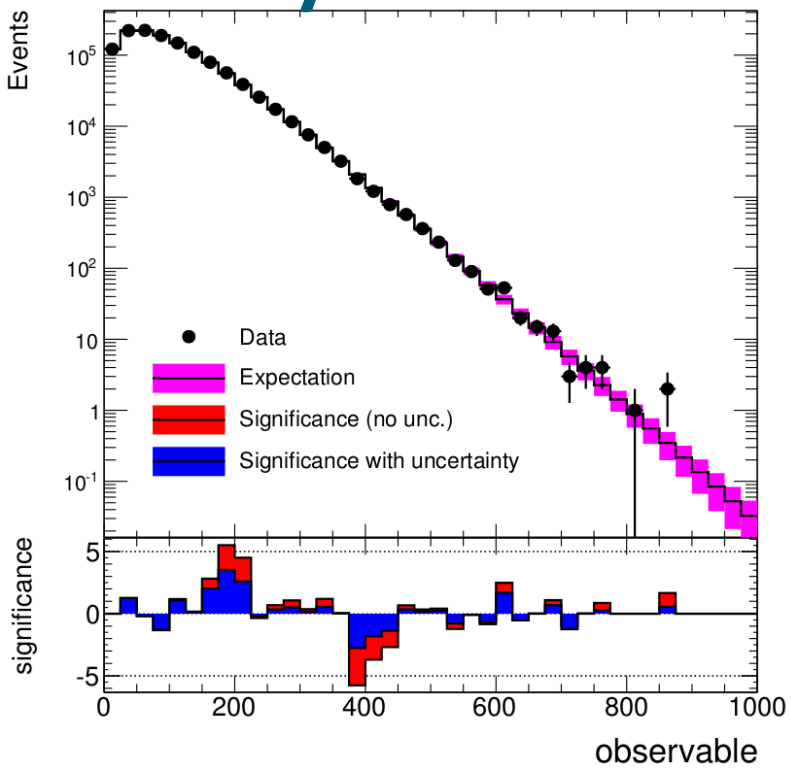
The Uncertainty

- Any theoretical uncertainty in the BACKGROUND will affect the significance of the observation
- Any additional uncertainty will *decrease* the significance of the observed deviation
- Sometimes is not useful to show it:
  - Negligible compared to statistical uncertainty
  - Deviations without theoretical uncertainty already not significant enough

Theoretical Uncertainty

- ERROR BARS: shifting B by $\pm\sigma$ and recomputing significance → try scaling your BKG!
- BUT: any uncertainty will decrease significance of deviation!





Significance computed by neglecting the *theoretical uncertainty* on the expectation (red histogram) and by including it (blue histogram)

Conclusion

It is possible to improve the PLOT of the differences between DATA and EXPECTATION:

- The useful "metric" should show significance in an *accurate* and *intuitive* way
- The exact *p-value* is computed

  → if < 0.5 mapped into *z-value* ≡ deviation in units of Gaussian standard deviations
- The sign (*positive* / *negative* bars) has to show EXCESS/DEFICIT
- Before claiming discovery for important deviations (z-value > 3-4), it is fundamental to check what happens by including the total uncertainty on the expectation!

Thanks For Your Attention!