Plotting the Difference Between Data and Expectation

Seminar Talk by
Silvia Biondi, **Tobias Bisanz** and **Lara Bartels**

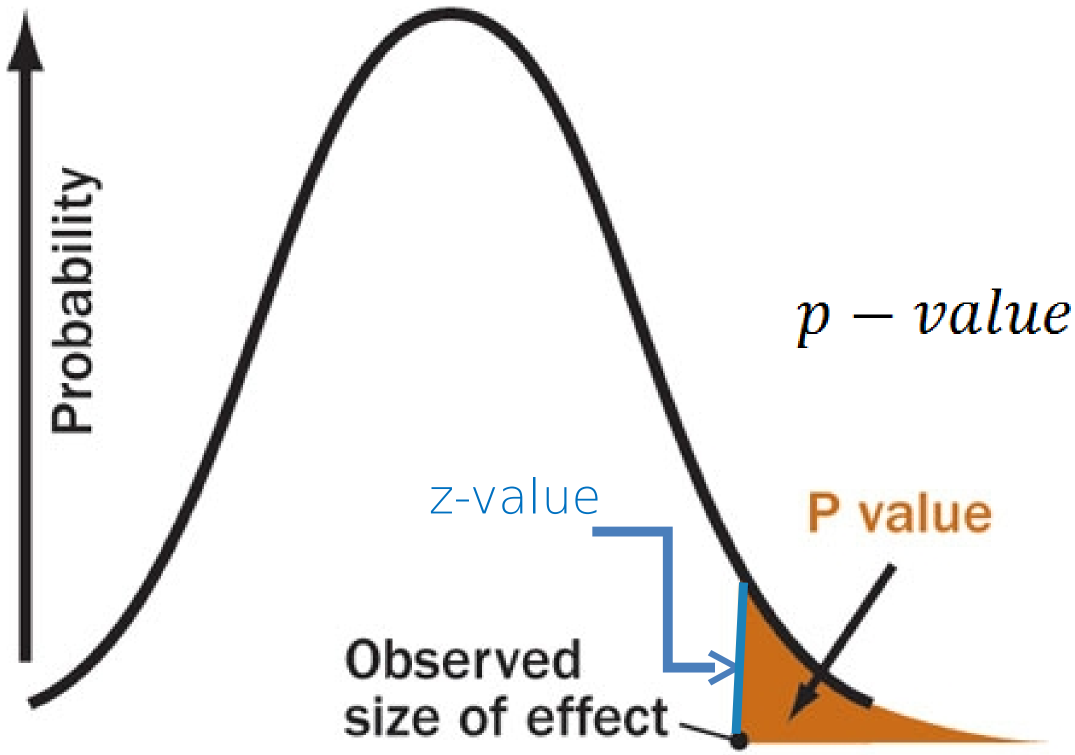
HASCO Summer School 2013

Outline

- Definition of statistical significance
- The Poisson and Binomial model
- Presenting deviations in Poisson distributed data
- Theoretical Uncertainty
- Conclusions

Definition of statistical significance

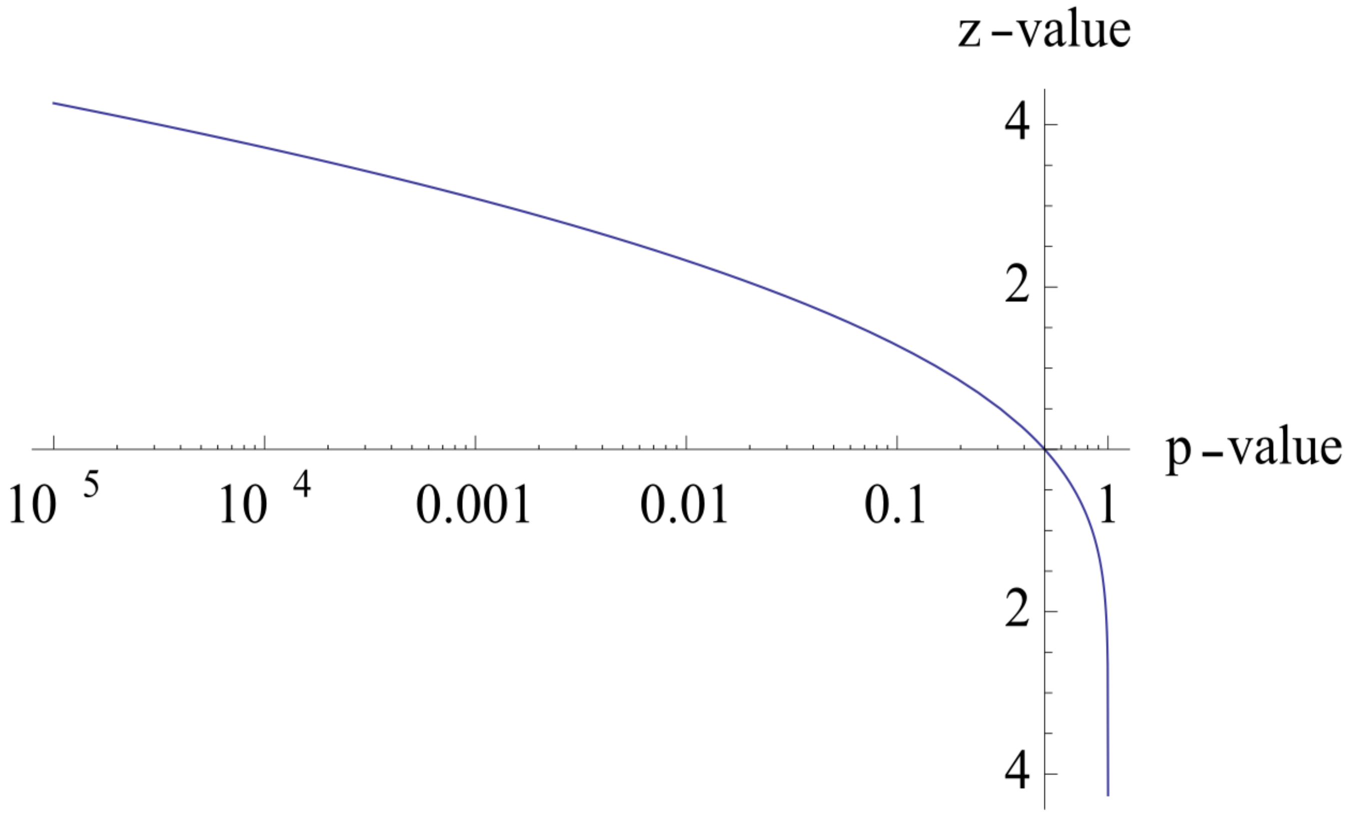
- P-value = Prob(deviation >= observed)



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

- Depends on the probability model

Translating the p-value into a z-value



- Z-value >= 0 p-value <= 0.5
- Z-value < 0 p-value > 0.5

Poisson model

B = number of expected events in a bin
D = number of observed events

$$P(D|B) = \mathrm{Poi}(D|B) = \frac{B^D}{D!}e^{-B}$$

The **poisson p-value** is

$$p\text{-value} = \begin{cases} \sum_{n=D}^{\infty} \frac{B^n}{n!}e^{-B} = 1 - \sum_{n=0}^{D-1} \frac{B^n}{n!}e^{-B} & , D > B \\ \sum_{n=0}^{D} \frac{B^n}{n!}e^{-B} & , D \leq B \end{cases}$$

Thanks to the identity $\sum_{n=0}^{D-1} \frac{B^n}{n!}e^{-B} = \frac{\Gamma(D,B)}{\Gamma(D)}$, we can define the regularized Gamma function

$$Q(s,x) = \frac{\Gamma(s,x)}{\Gamma(s)} = 1 - P(s,x)$$

where P(s,x) is the cumulative distribution function

- In ROOT this function is available as $Q(s,x) = \texttt{ROOT::Math::inc\_gamma\_c(s,x)}$

such that

$$p\text{-value} = \begin{cases} 1 - Q(D,B) = \texttt{ROOT::Math::inc\_gamma\_c(D,B)} & , D > B \\ Q(D+1,B) = \texttt{ROOT::Math::inc\_gamma\_c(D+1,B)} & , D \leq B \end{cases}$$

Binomial model

n = number of initially observed events in a bin
Kobs = observed number of selected events
ε = expected success rate

$$P(k|n,\varepsilon) = \mathrm{Bi}(k|n,\varepsilon) = \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k}$$

The **binomial p-value** is

$$p\text{-value} = \begin{cases} \displaystyle\sum_{n=k_{\mathrm{obs}}}^{n} \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} & , \; k_{\mathrm{obs}} \geq n\varepsilon \\ \displaystyle\sum_{n=0}^{k_{\mathrm{obs}}} \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} & , \; k_{\mathrm{obs}} < n\varepsilon \end{cases}$$

The cumulative distribution function of the binomial model can be represented in terms of the incomplete Beta function: $I_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t / B(a,b)$

So the p-value is:

$$p\text{-value} = P(k \leq k_{\mathrm{obs}}) = I_{1-\varepsilon}(n - k_{\mathrm{obs}}, k_{\mathrm{obs}} + 1)\;, \quad k_{\mathrm{obs}} < n\varepsilon$$

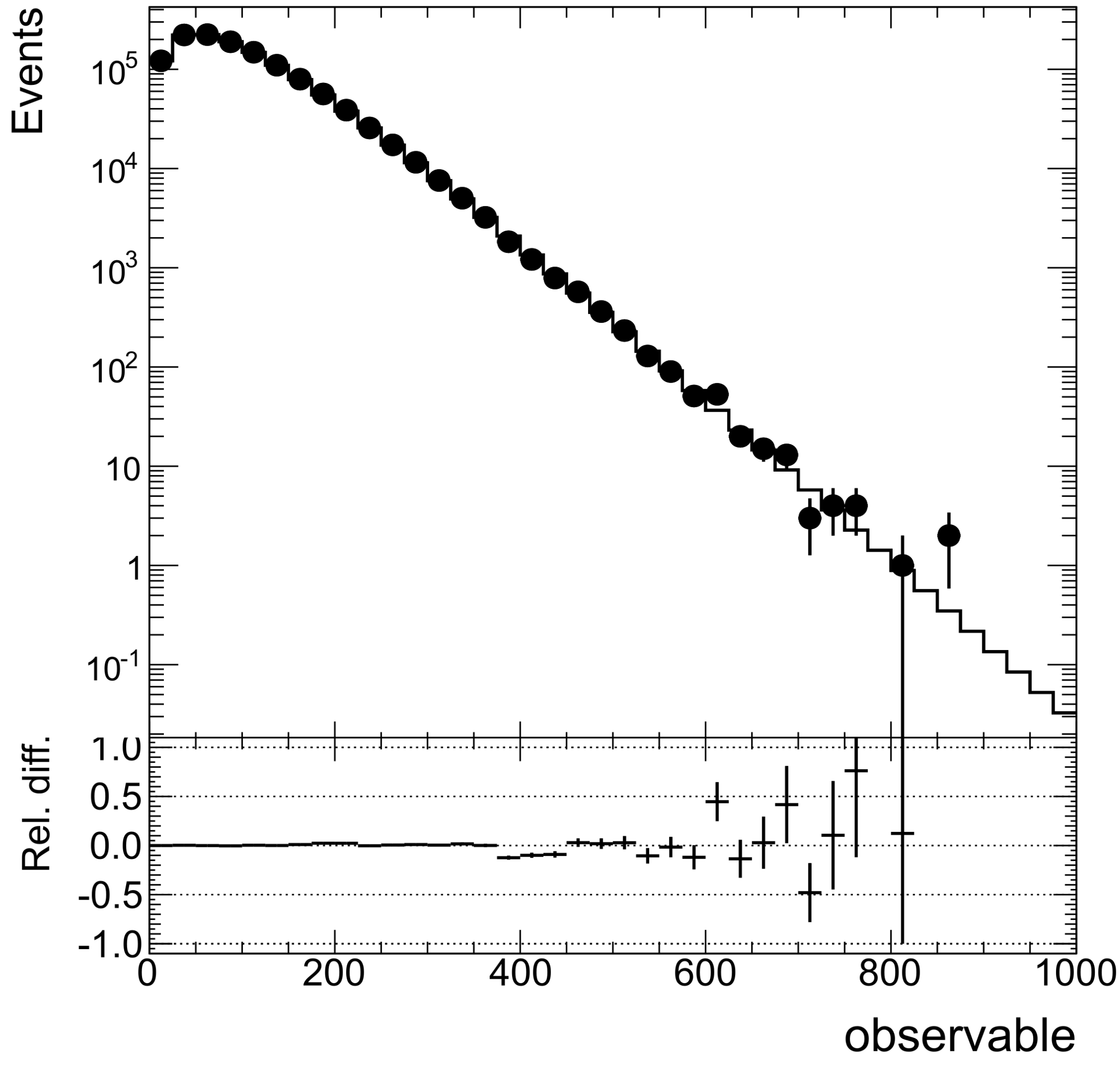
$$p\text{-value} = P(k \geq k_{\mathrm{obs}}) = I_{\varepsilon}(k_{\mathrm{obs}}, n - k_{\mathrm{obs}} - 1)\;, \quad k_{\mathrm{obs}} > n\varepsilon$$

- In ROOT the regularized incomplete Beta function is available as

$$I_x(a,b) = \texttt{ROOT::Math::inc\_beta(x,a,b)}$$

The D/B ratio

- Alternatively: (D-B)/B
- Simple & intuitive
- several orders of magnitude
  ➛ significant deviation hidden
- no statistical significance
  ➛ large fluctuations
  for low stats
- Asymmetry pos./neg.
  deviations

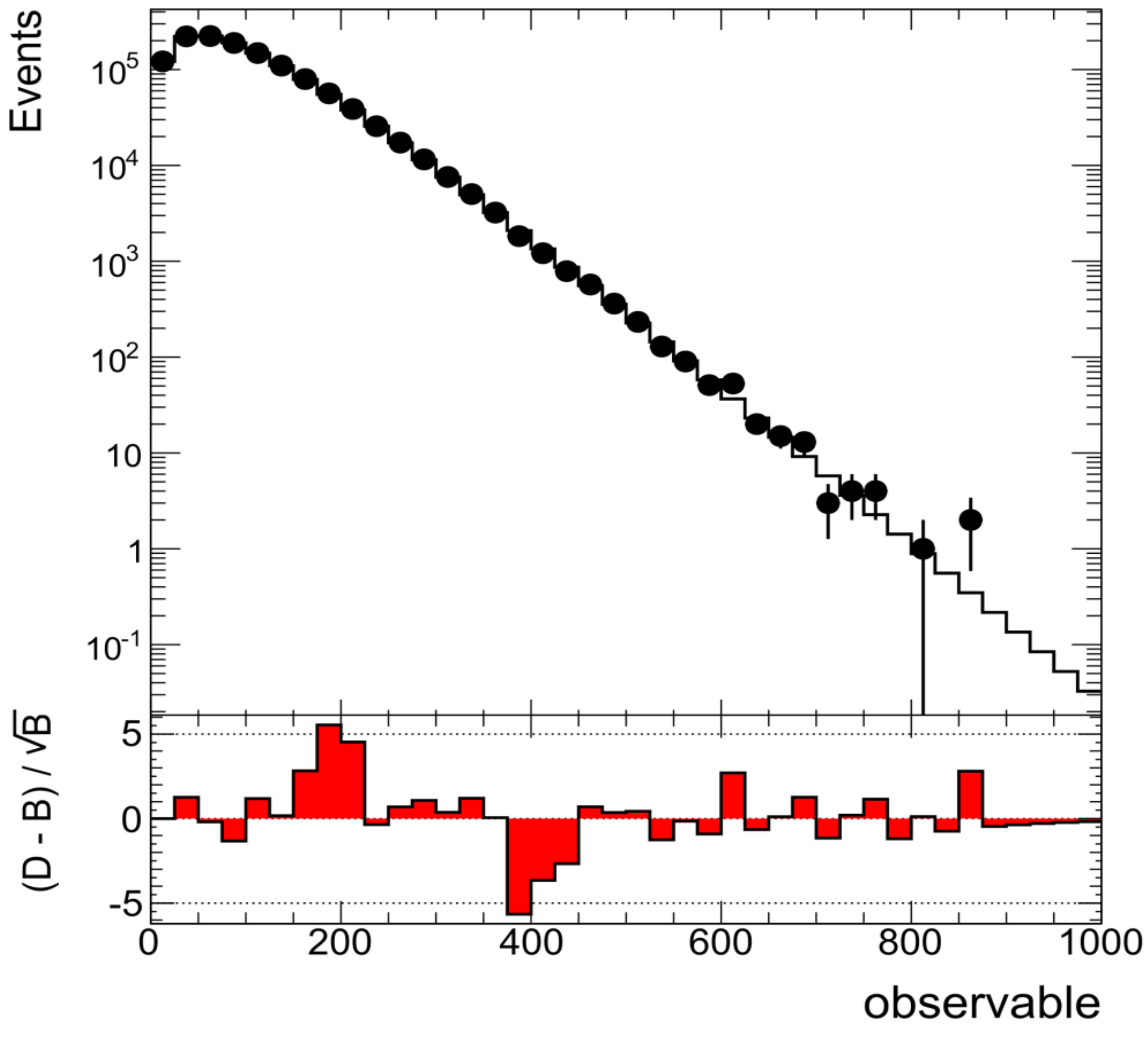


$\left(D-B\right)/\sqrt{B}$ approximation

- Approx: Poisson ⟶ Gaussian
  for large B

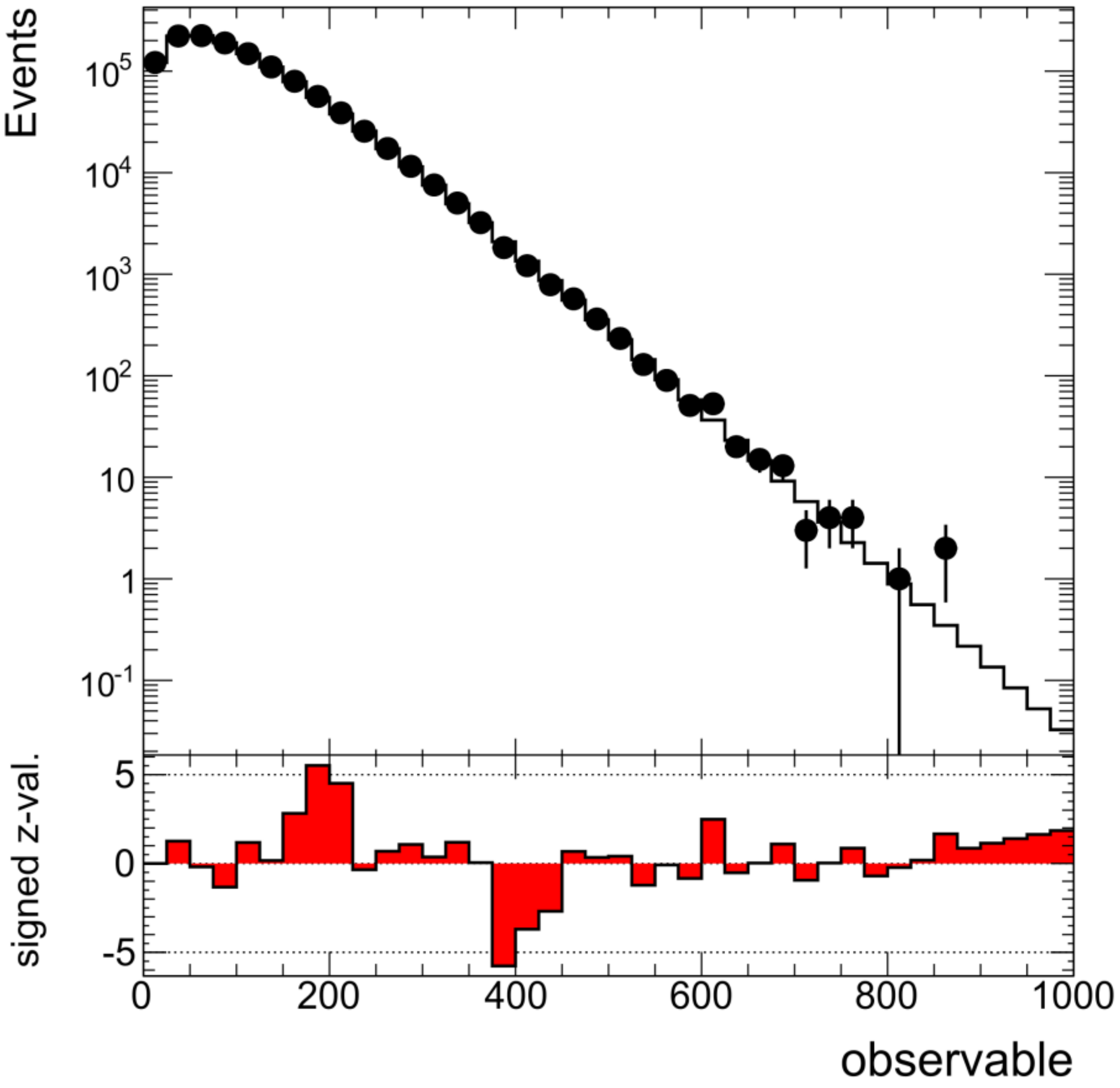
- Significant deviations clearly visible

- Not a good approximation for low-population bins



Plotting signed z-values

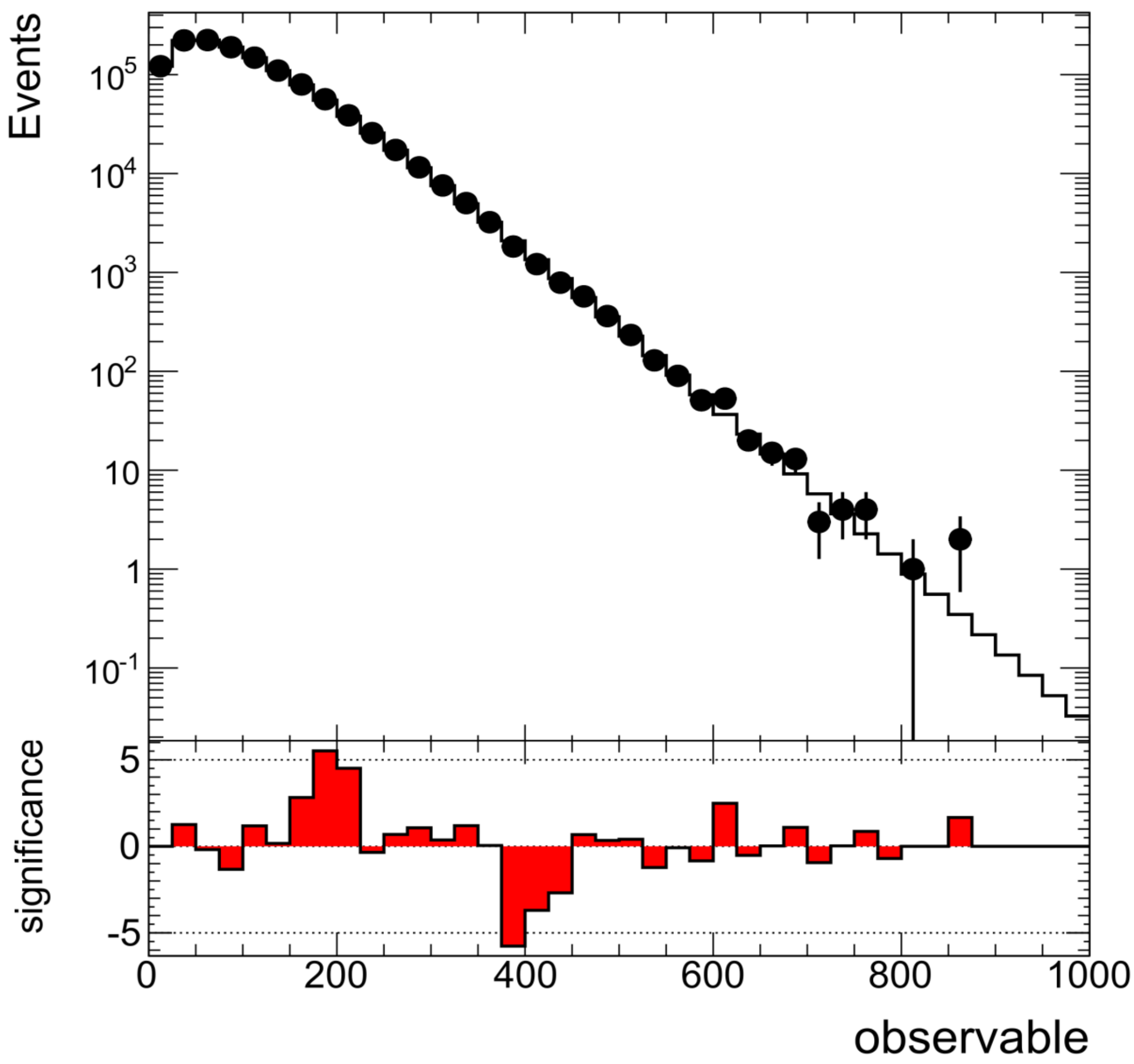
- Plotting the exact z-value:
- D>B (excess):
  + z-value
- D<B (deficit):
  - z-value
- Negative z-values for p-value>0.5
- Problem with low stats: very insignificant deficits
  → negative z-value sign-flipping
  → appear as excess



The final proposal:
Plot signed z-values only if p-value < 0.5

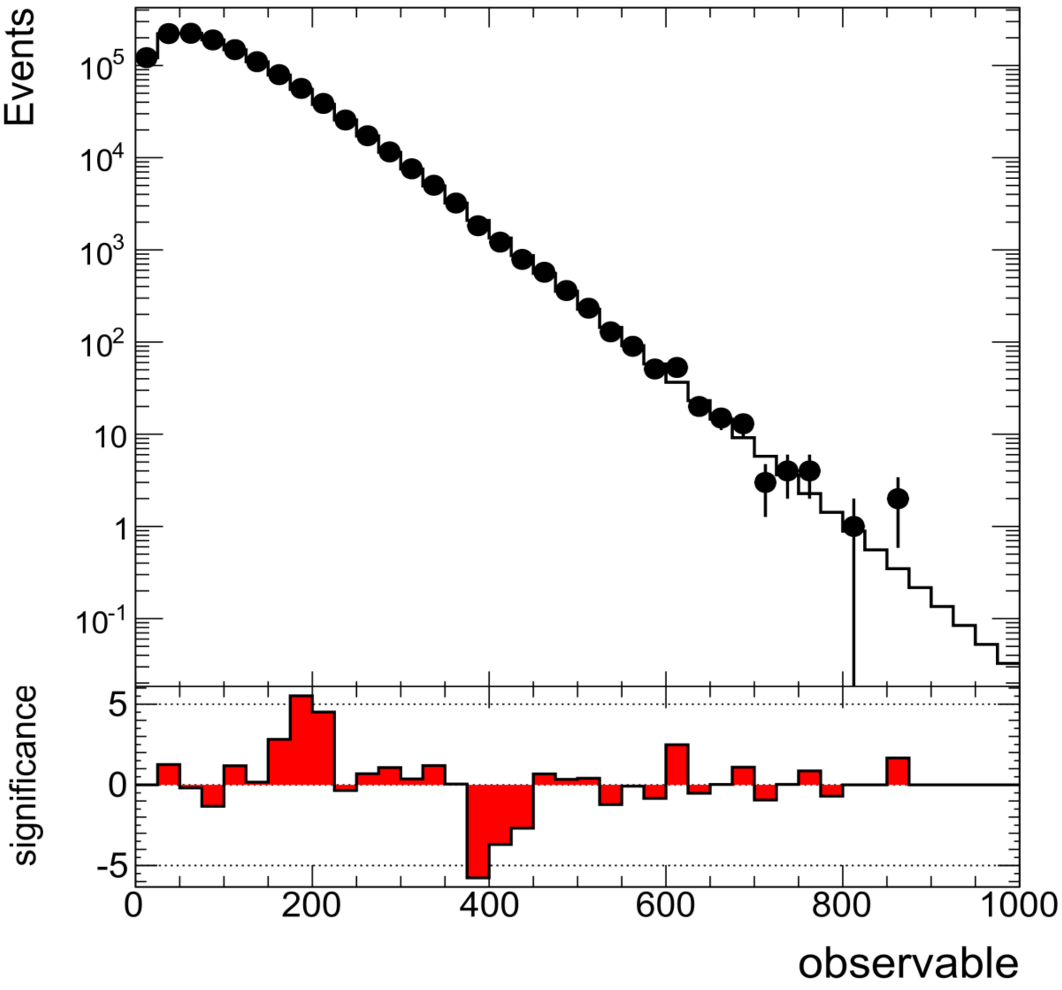
- Plotting z-values as before

- Bins with a corresponding p-value < 0.5 agree perfectly with the expectation
  ➔ As we've just seen, plotting them could be misleading



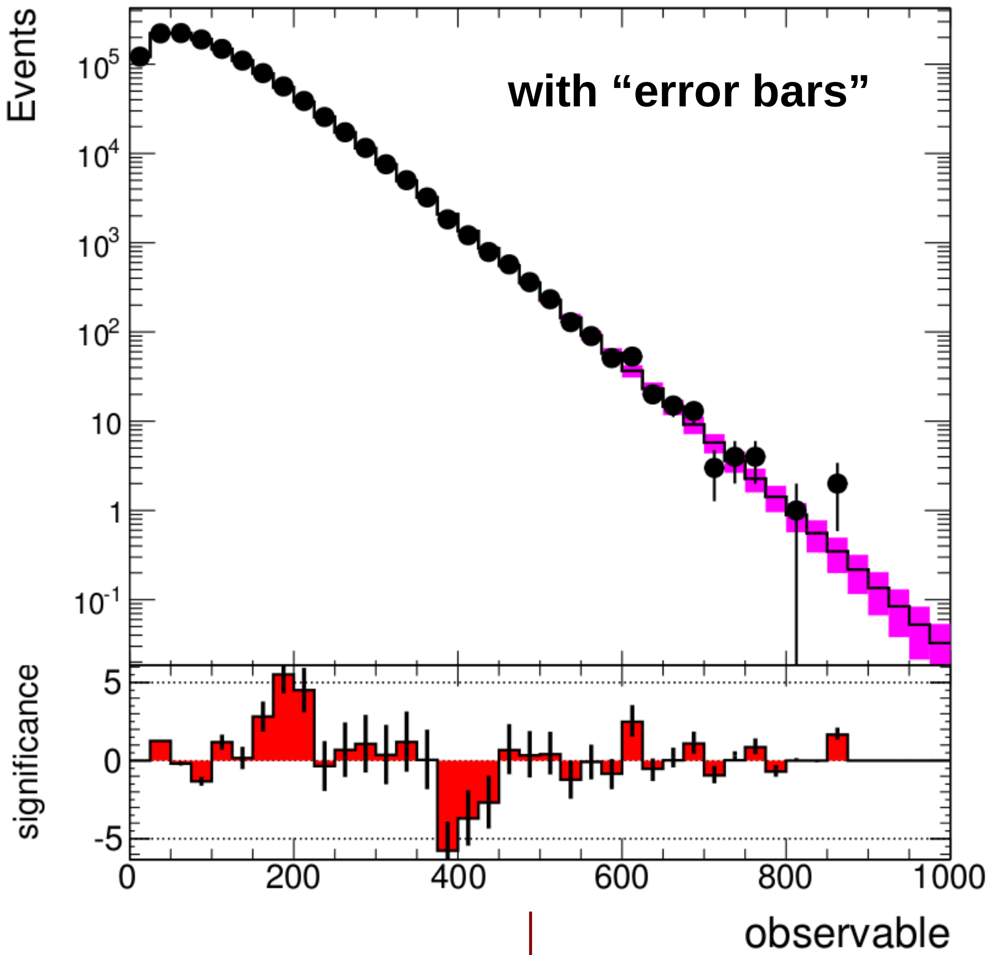
The final proposal:
Plot signed z-values only if p-value < 0.5

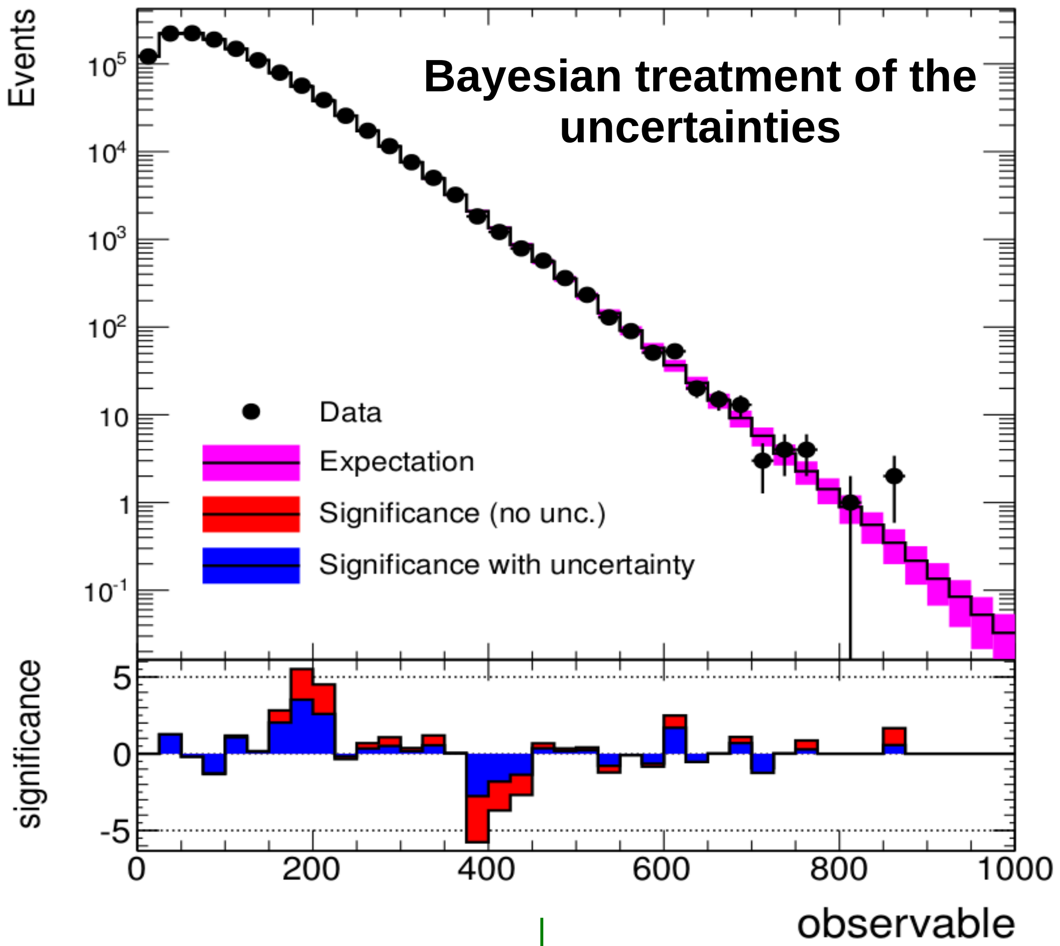
- Z-value is accurate
- Positive values represent excesses of data over expectation
- No significant deviations hidden ($p\text{-value} < 0.5$)
- Same treatment of bins with high and low statistics
- Easy to implement using ROOT



Theoretical uncertainty

Any theoretical uncertainty in the reference value will affect the significance of the observation





It misses the fundamental point:
any additional uncertainty will decrease the significance of the observed deviation

Including additional sources of uncertainties decreases the significance

Conclusions

- shown an improved way of plotting the difference between data and expectation
  - an accurate plot of the statistical significance of the deviation of the bin contents from the expectation
  - an intuitive picture of the relevant deficits and excesses
  - achieved by computing the exact p-value and, when its value is smaller than 50% probability, by mapping it into the z-value
  - the sign of z-values is always positive for excesses and negative for deficits
- fundamental to check what happens by including the total uncertainty on the expectation, before claiming that an excess is really significant
  - always lowers the actual significance
- the focus is only on methods which improve illustrations

Backup

Plot of a cumulative distribution function of the Poisson model

$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$