Weights and Measures
HERMES/NIKHEF Group Meeting
18 March 2004

Keith Griffioen
NIKHEF and College of William and Mary

March 18, 2004

Abstract

I discuss the art of error analysis for weighted histograms by mathematical derivation and Monte Carlo simulation. I apply this same approach to the determination of averaged cross sections and asymmetries.

Weights & Measures

HERMES Group Meeting
NIKHEF 18 March 2004

3 questions:

1) How do I calculate errors on a weighted histogram?

2) How do I best calculate a cross section from several data samples?

3) How do I best calculate an aysmmetry from several data samples?

3 answers

1) Depends

2) Not the way you're used to

3) What a can of worms

Weighted Histograms

If $W=\sum_{i=1}^{N} w_i$, then the error σ_W on W is

(a) $\sigma_W = \sqrt{\sum_i w_i^2}$ if N fluctuates with Poisson Statistics

(b) $\sigma_W = \sqrt{\sum_i w_i^2 - \frac{1}{N}\left(\sum_i w_i\right)^2}$ if N is fixed

proof (a) Collect data for a time T
Break this up into T/δ smaller intervals

$$W_T = W_\delta^{(1)} + W_\delta^{(2)} + \ldots + W_\delta^{(T/\delta)}$$

[each W is the sum of weights during that time interval]

★ statistics depend ONLY on how long we count

$$\therefore W_\delta^{(i)} = W_\delta^{(j)} \quad \text{(on average)}$$

★ data in each interval are uncorrelated

$$\therefore \mathrm{Var}(W_T) = \mathrm{Var}\left(W_\delta^{(1)} + W_\delta^{(2)} + \ldots + W_\delta^{(T/\delta)}\right) = \frac{T}{\delta}\mathrm{Var}\,W_\delta$$

★ weights picked from distribution $f(w)$

N_T counts in T $\quad \langle N_T\rangle = RT \quad \delta \to 0$

P_i is probability of i counts in our bin $\quad P_0 \approx 1$, $P_1 \approx R\delta$ etc.

$$\langle W_\delta^2\rangle = \int \left[0^2\cdot P_0 + w_1^2 P_1 + (w_1+w_2)^2 P_2 + \ldots\right] f(w_1) f(w_2)\, dw_1 dw_2$$

$$\therefore \langle W_\delta\rangle = \langle w\rangle R\delta \quad \langle W_\delta^2\rangle = \langle w^2\rangle R\delta \quad \mathrm{Var}(W_\delta) = \langle W_\delta^2\rangle - \langle W_\delta\rangle^2$$
$$= \langle w^2\rangle R\delta + O(R^2\delta^2)$$

$$\therefore \mathrm{Var}(W_T) = \frac{T}{\delta}\langle w^2\rangle R\delta = \langle w^2\rangle\langle N_T\rangle$$

$$\boxed{\Delta W_T = \sqrt{\mathrm{Var}(W_T)} = \sqrt{\frac{1}{N_T}\sum_i w_i^2 N_T} = \sqrt{\sum_i w_i^2}}$$

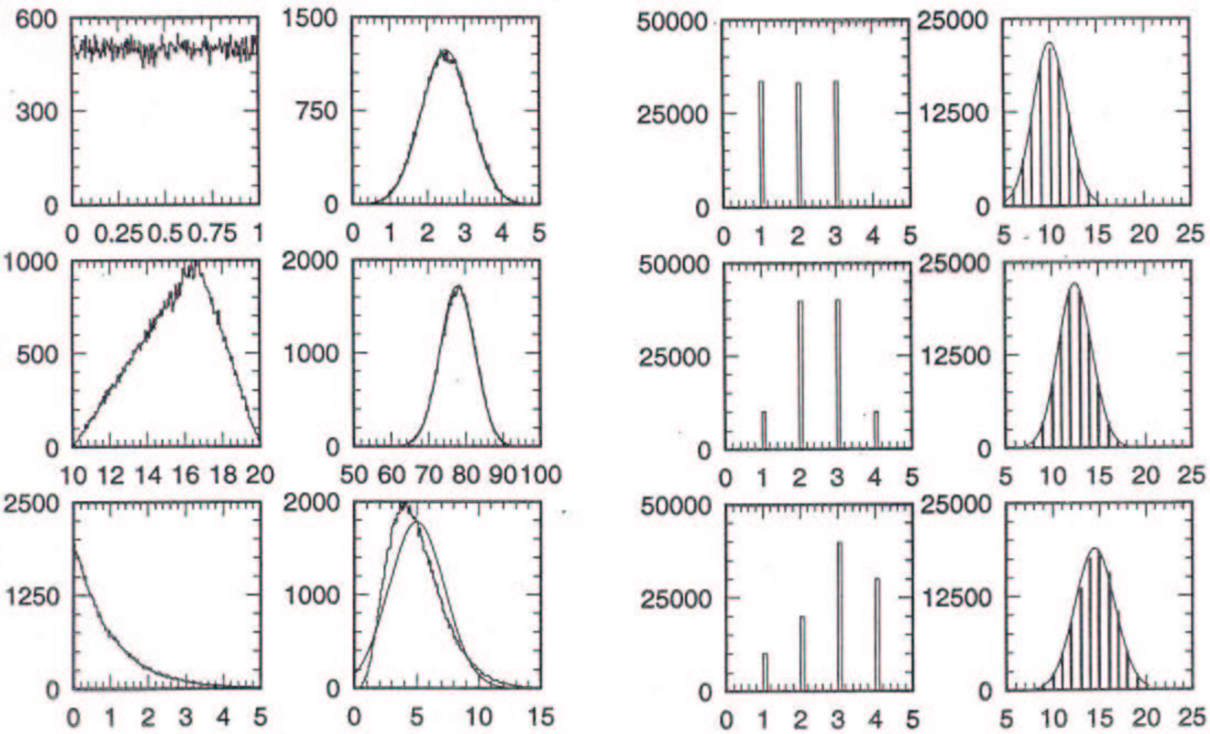
proof (b) Central Limit Theorem

The random variate $X = \frac{1}{N}\sum_i w_i$ is normally distributed with mean $\mu_X = \mu_W$ and variance $\sigma_X^2 = \sigma_W^2/N$ for large N.

$$\sigma_W^2 = \langle w^2 \rangle - \langle w \rangle^2 = \frac{1}{N}\sum_i w_i^2 - \left(\frac{1}{N}\sum_i w_i\right)^2$$

$\therefore W = \sum_i w_i$ has mean $N\mu_W$ and variance $N\sigma_W^2$

How big must N be?

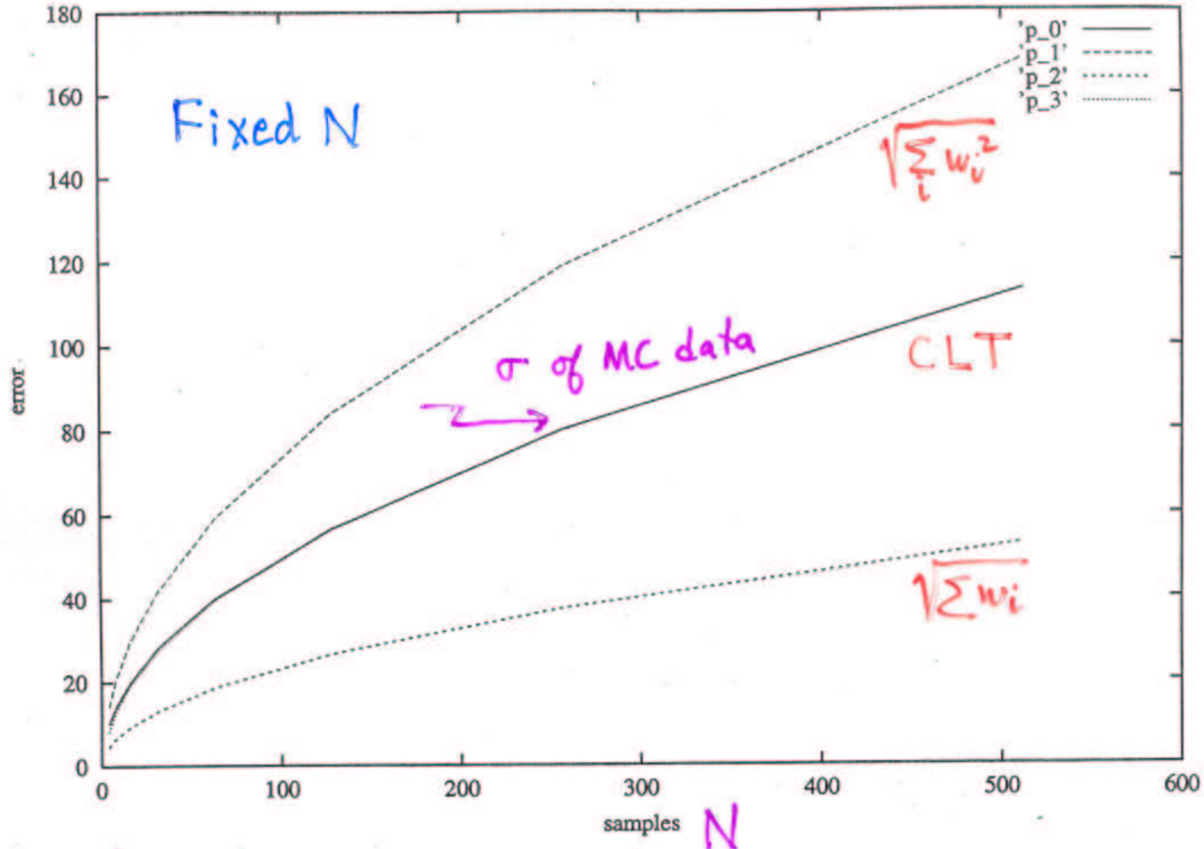


5, unless your distribution $f(w)$ is an exponent.

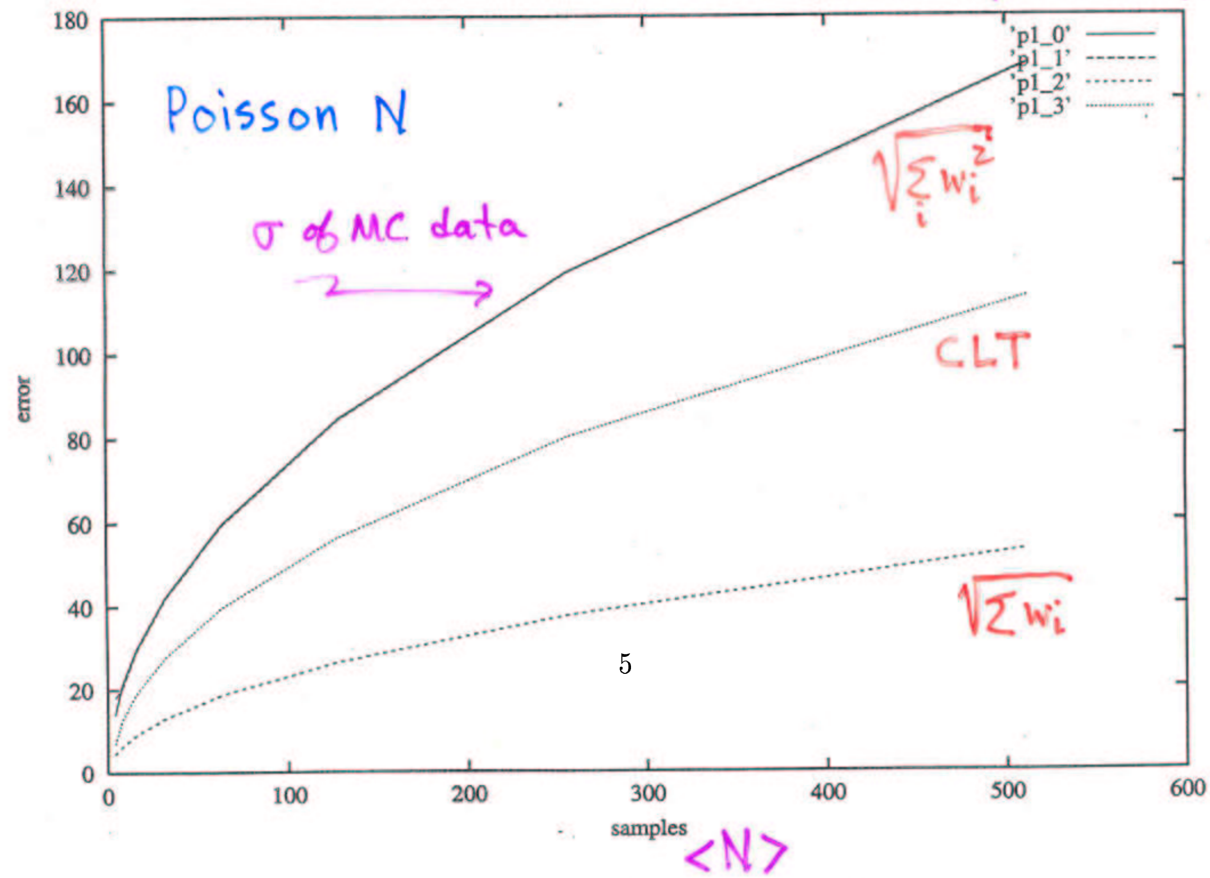
left plots in group: $f(w)$ continuous / discrete

right plots in group: $w_1 + w_2 + w_3 + w_4 + w_5$

Monte Carlo Proof



MC run 100000 times
statistics kept on output.



Conclusions

(a) ☆ If you run for a fixed time such that the number of events in a bin has Poisson fluctuations, then

$$\sigma_W = \sqrt{\sum_i w_i^2}$$

(b) ☆ If you run until you have collected exactly N events in your bin, then

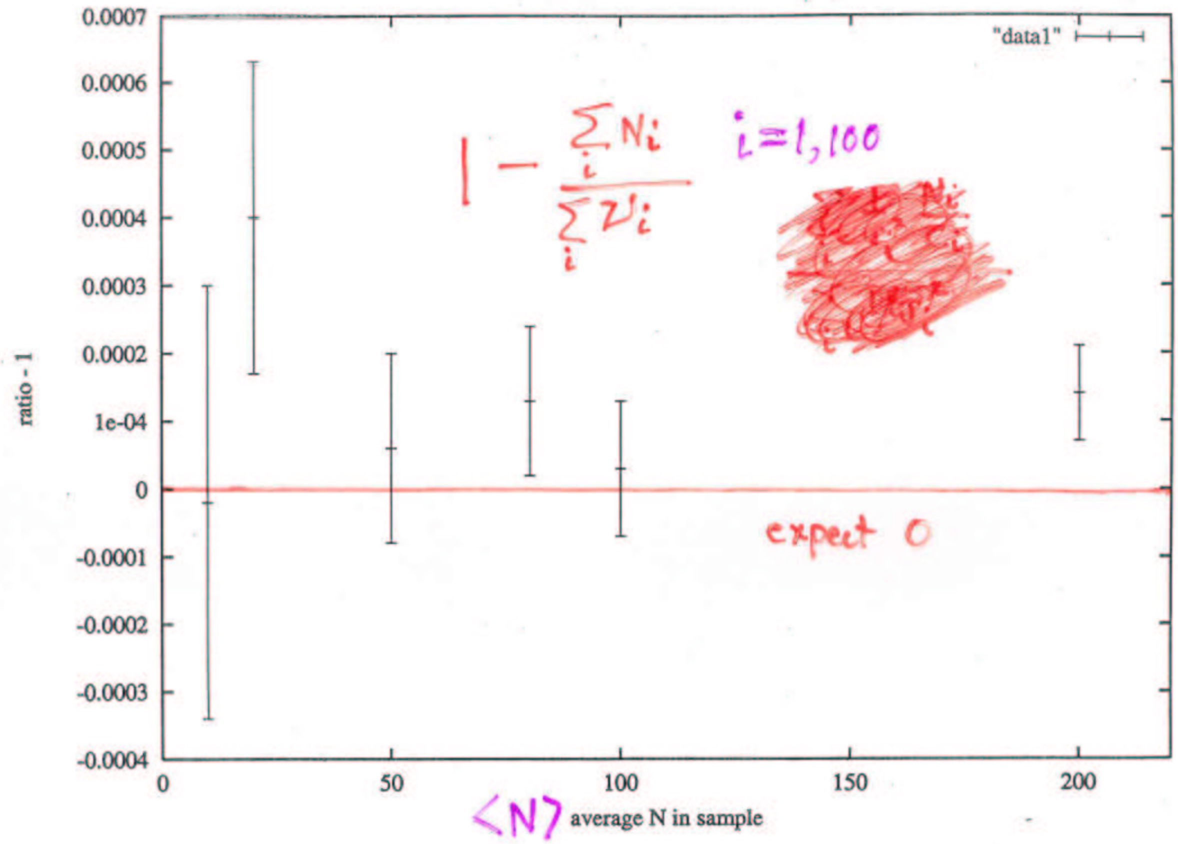
$$\sigma_W = \sigma_{w_i} * N$$

(a) > (b) ∴ (b) contains more information

[for each trial you know N]

→ measures how well you know the centroid of the w_i distribution
(e.g. = 0 if $w_i = 1$ for all i)
(use for $\frac{1}{N}\sum_i \sin\theta_i$, $\frac{1}{N}\sum_i P_3(\cos\theta_i)$, $\frac{1}{N}\sum_i \sin\phi_i$)

→ measures the error on a weighted histogram bin. (e.g. you know the weights w_i but don't know how "full" your bin is going to be)
(use for MC in which each event is weighted)



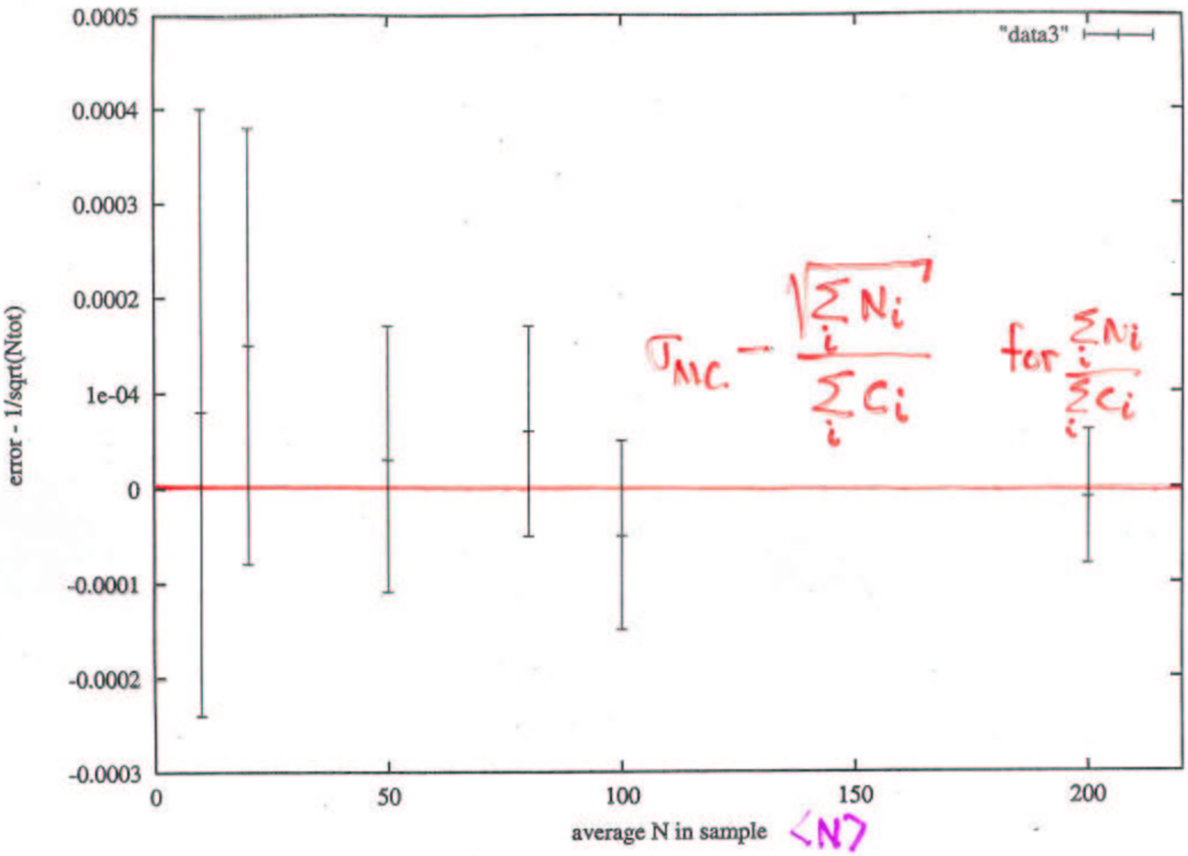
For any experimental quantity $\frac{N}{c}$ in which N is a measured number of counts and c is a normalization factor such that N/c is independent of how long one measures, then

$$\left\langle \frac{N}{c} \right\rangle = \frac{\sum_i N_i}{\sum_i c_i}$$

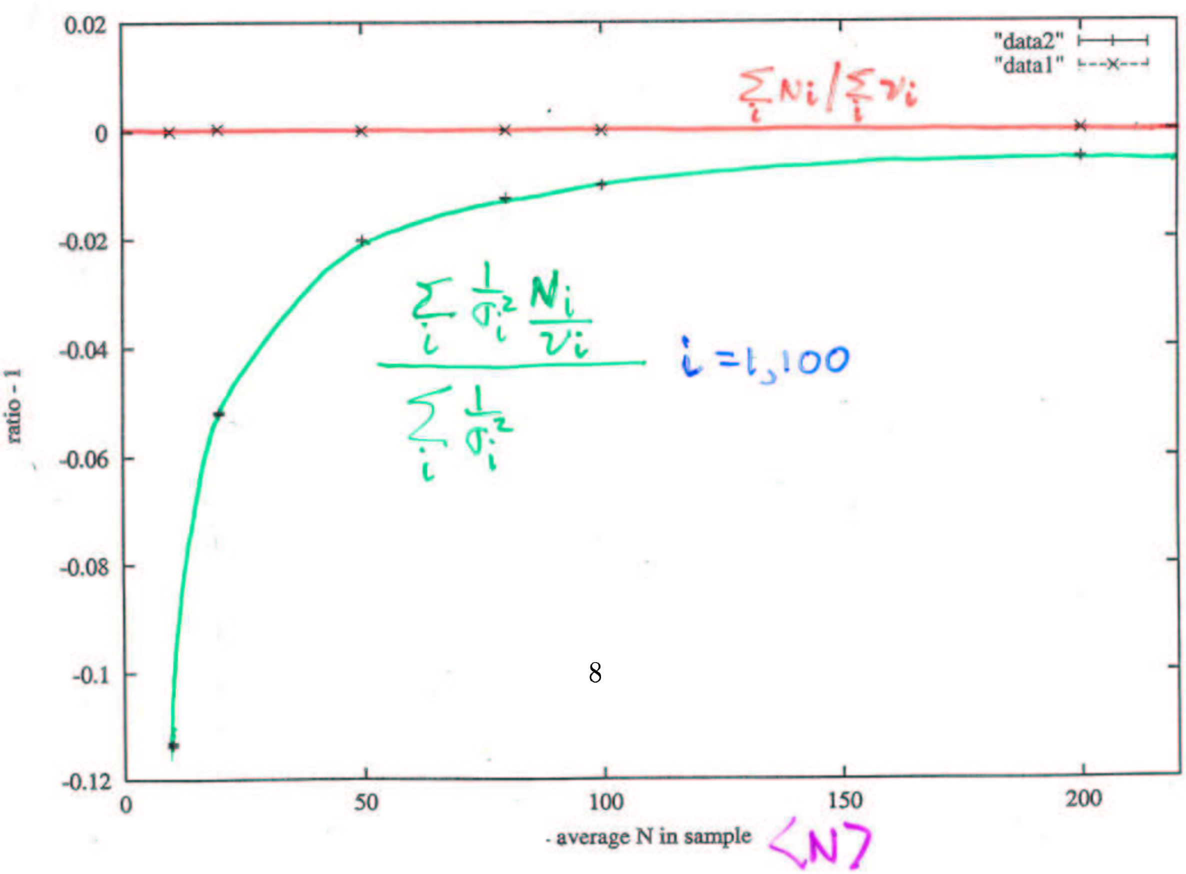
Run MC vs. $\langle N \rangle$ Poisson $P_\nu(n) = \frac{\nu^n}{n!} e^{-\nu}$

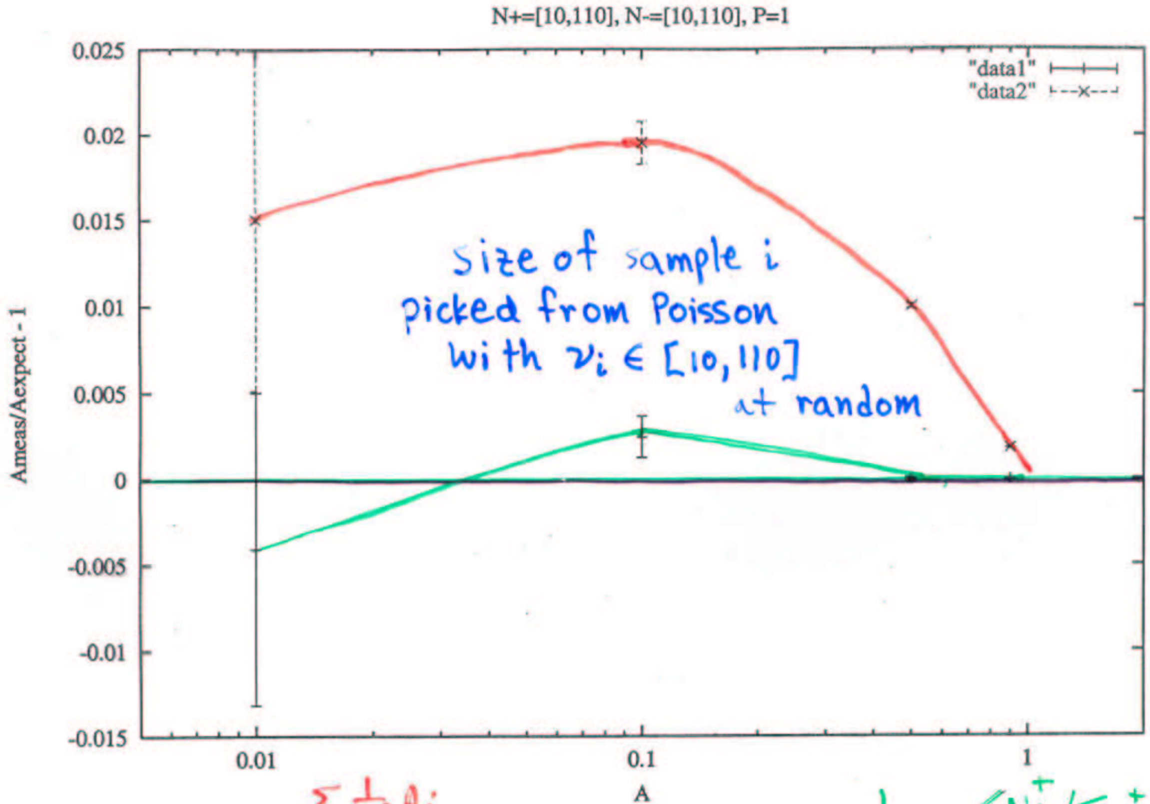
$$c_i \equiv \nu_i \qquad \sigma_i = \sqrt{N_i}/\nu_i$$

- Weighted error method fails!



10000 samples



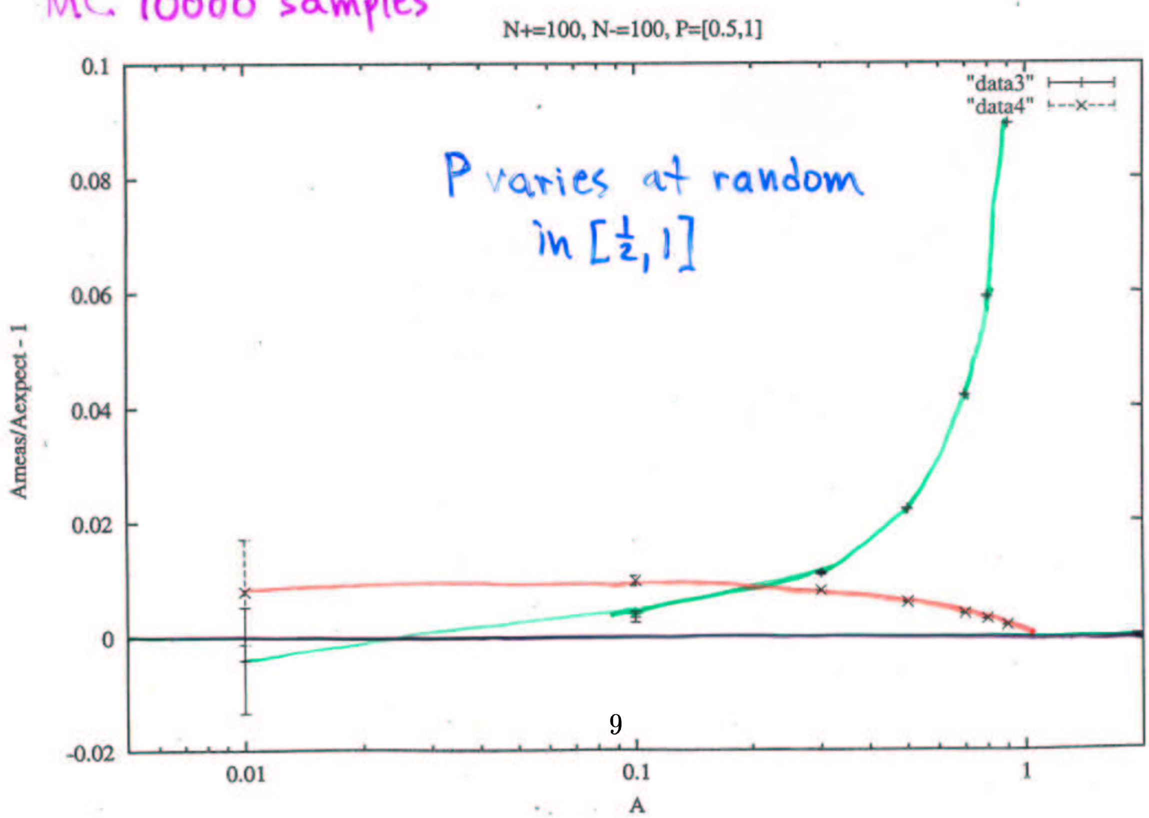


red: $A = \dfrac{\sum_i \frac{1}{\sigma_i^2} A_i}{\sum_i 1/\sigma_i^2}$

green: $A = \dfrac{1}{\langle P \rangle} \dfrac{\sum N_i^+ / \sum \nu_i^+ - \sum N_i^- / \sum \nu_i^-}{\sum N_i^+ / \sum \nu_i^+ + \sum N_i^- / \sum \nu_i^-}$

$i = 1, 100$

MC 10000 samples



Conclusions

- ☆ Uncertainties on the order of $\lesssim 1\%$ arise on A calculated from weighted averages

- ★ For $A < .1$ one asymmetry of sums and averages works better, but it goes bad for varying P at $A > .1$

- ☆ Surest method is probably to break up data into A_i's for roughly constant P, and then calculate weighted averages

- ☆ Use a larger event sample such that errors become gaussian.