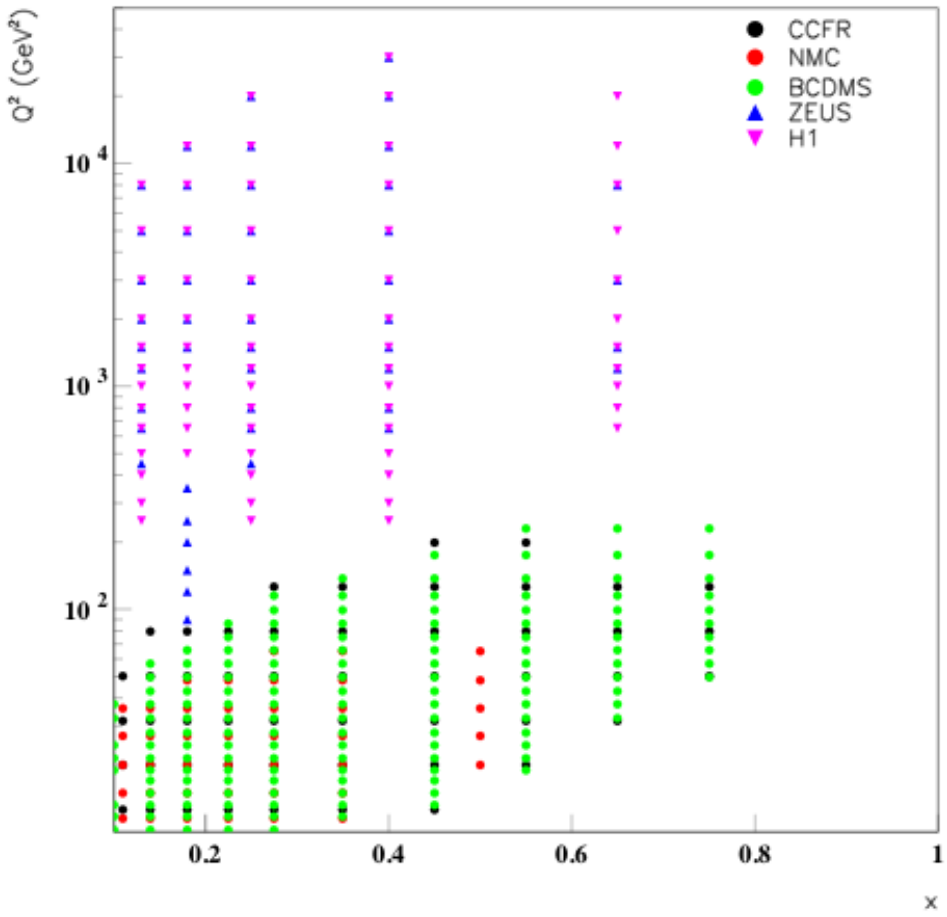
Is there more to learn from the high-x data ?

A. Caldwell

Max Planck Institute for Physics

Motivation

Our information on the very high x behavior of the parton densities is primarily theoretical.



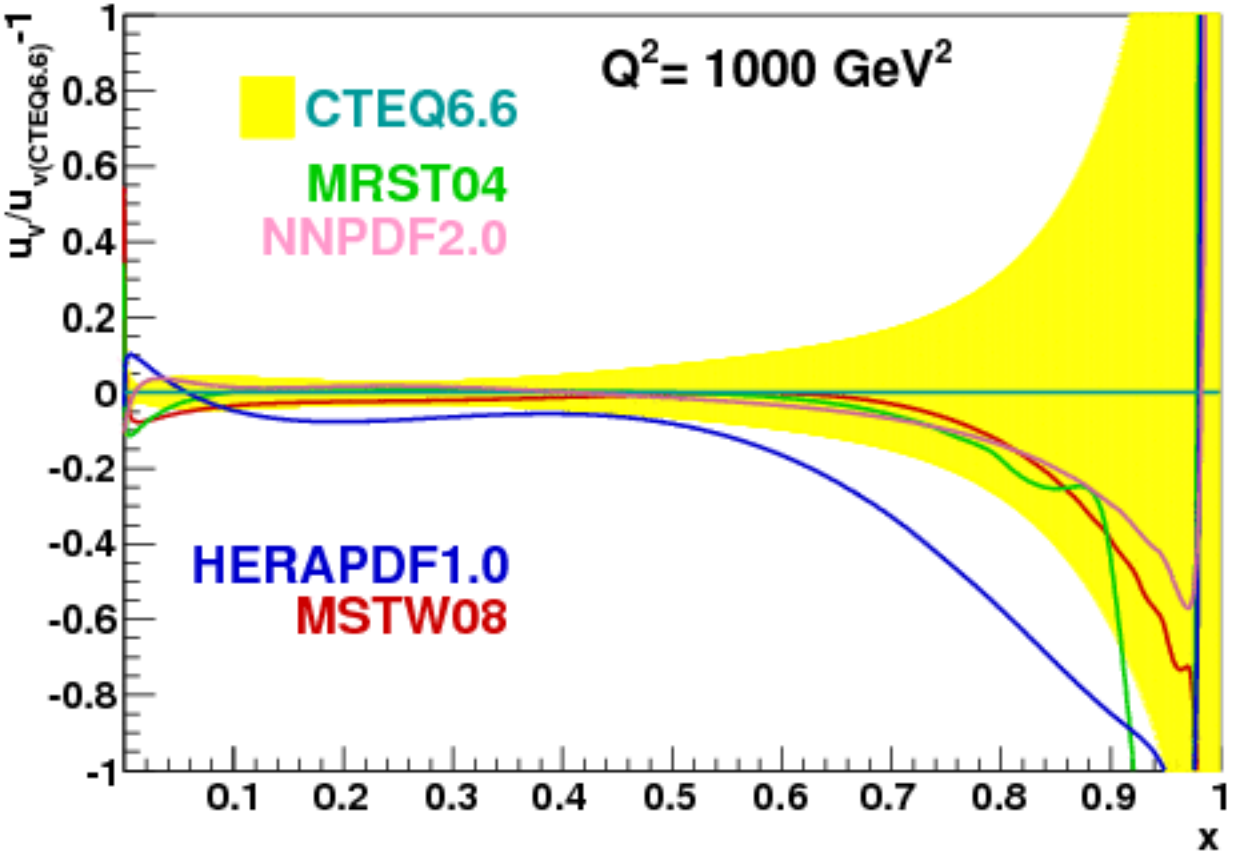
BCDMS has measured F_2 up to x=0.75

H1, ZEUS have measured F_2 up to x=0.65

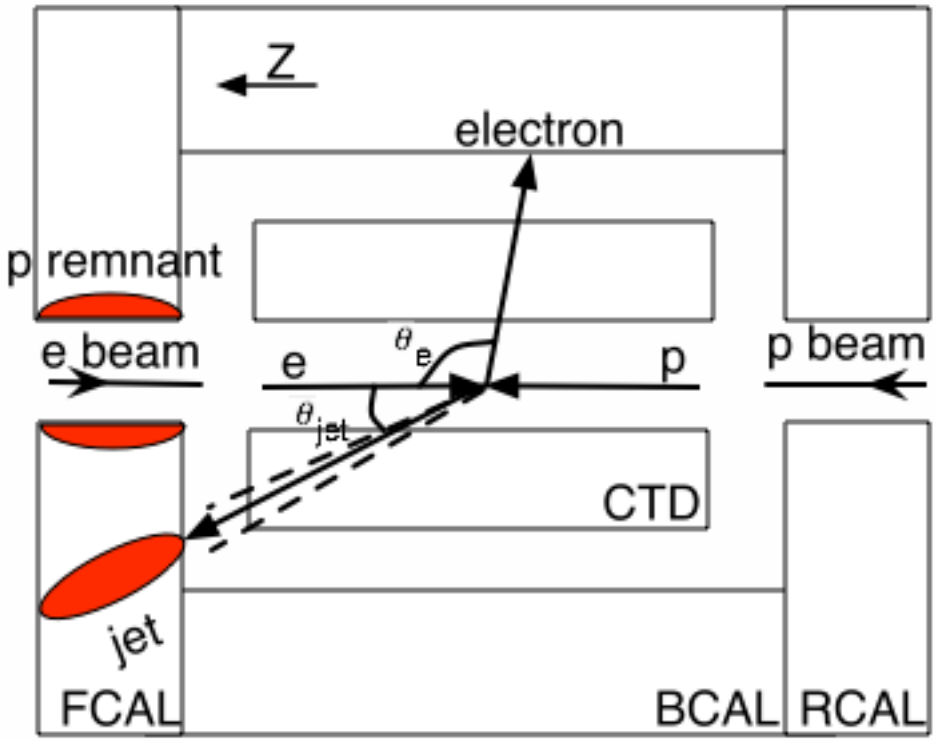
(not including dedicated ZEUS analysis at high-x)

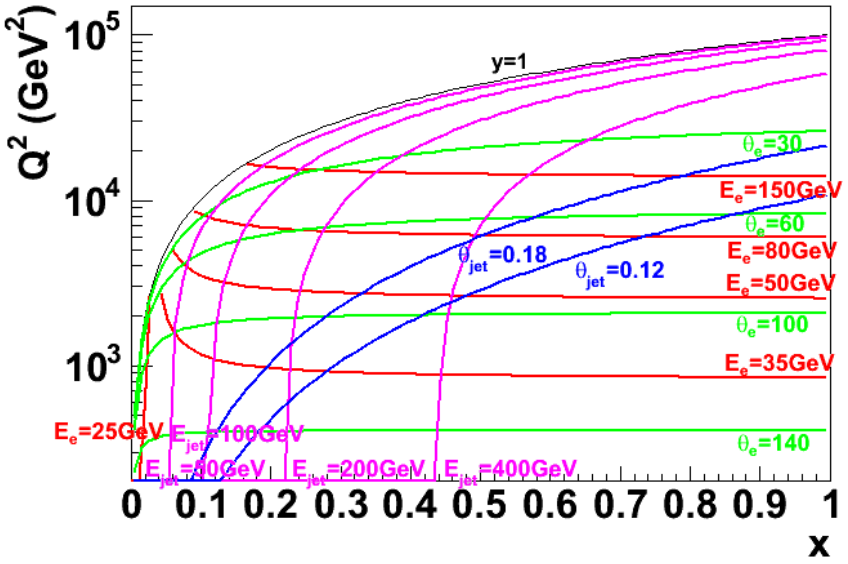
Motivation

The PDF's are poorly determined at high-x. Sizeable differences despite the fact that fits use similar parametrization $xq \propto (1-x)^{\eta}$. Is it possible to improve this situation ?



ZEUS high-x analysis



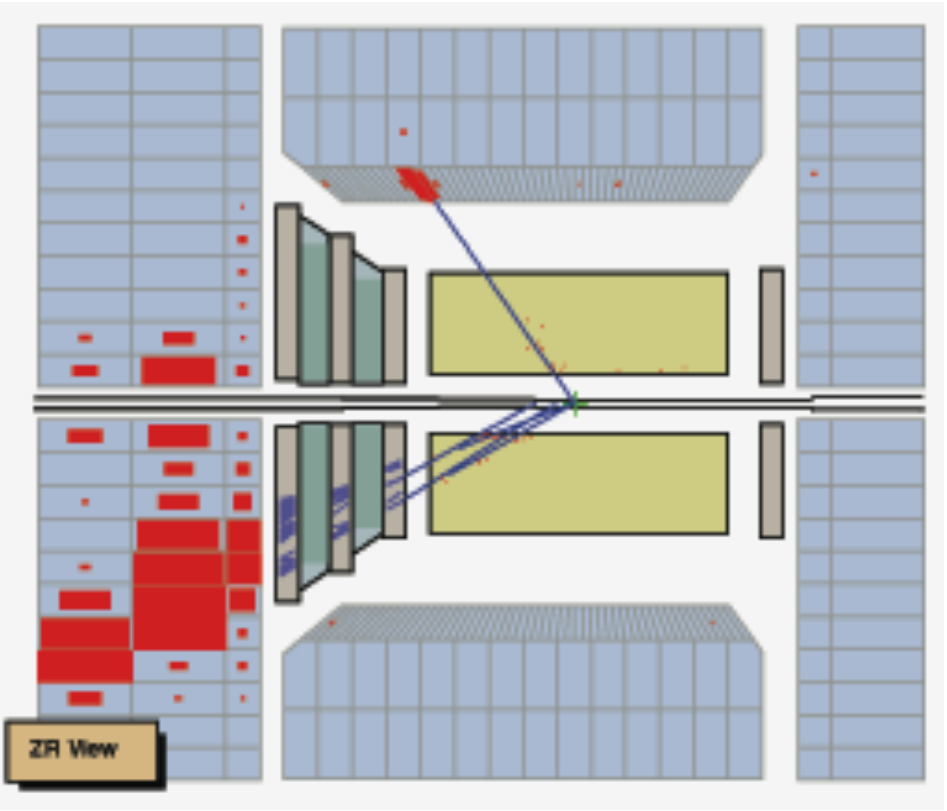


- At high Q^2, scattered electron seen with ≈100% acceptance
- For not too high x, measure x from jet: $\dfrac{d^2\sigma}{dxdQ^2}$
- For $x>x_{Edge}$, measure $\displaystyle\int_{x_{Edge}}^{1} \frac{d^2\sigma}{dxdQ^2}\,dx$

HERA kinematics

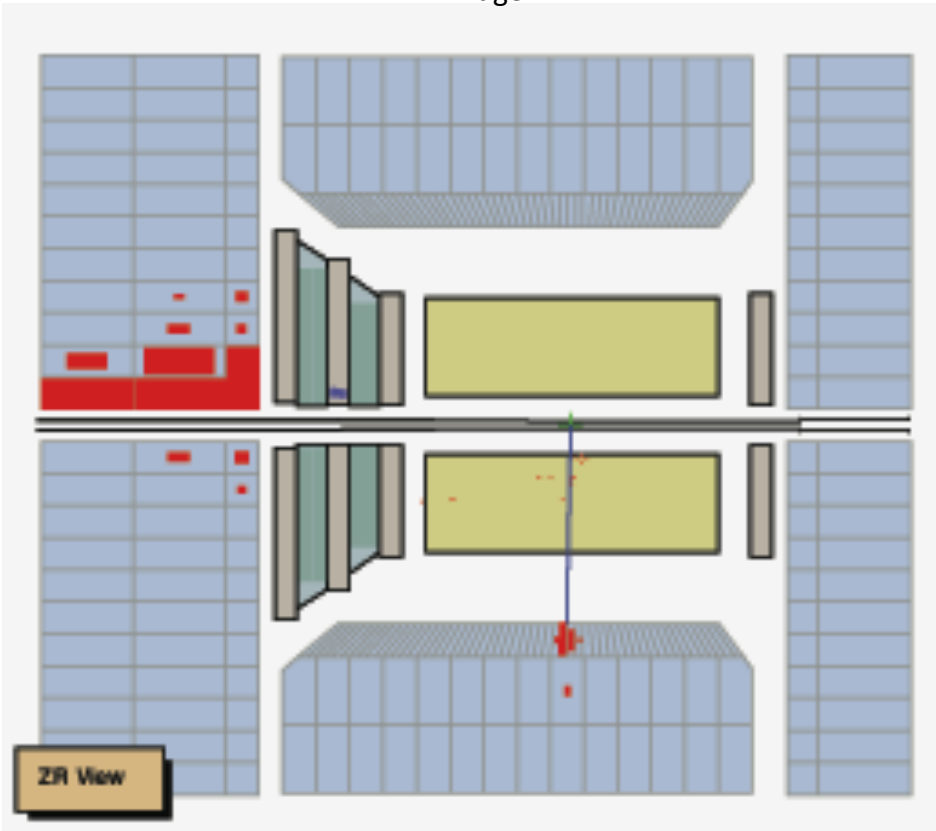
Jet found

$x<x_{Edge}$



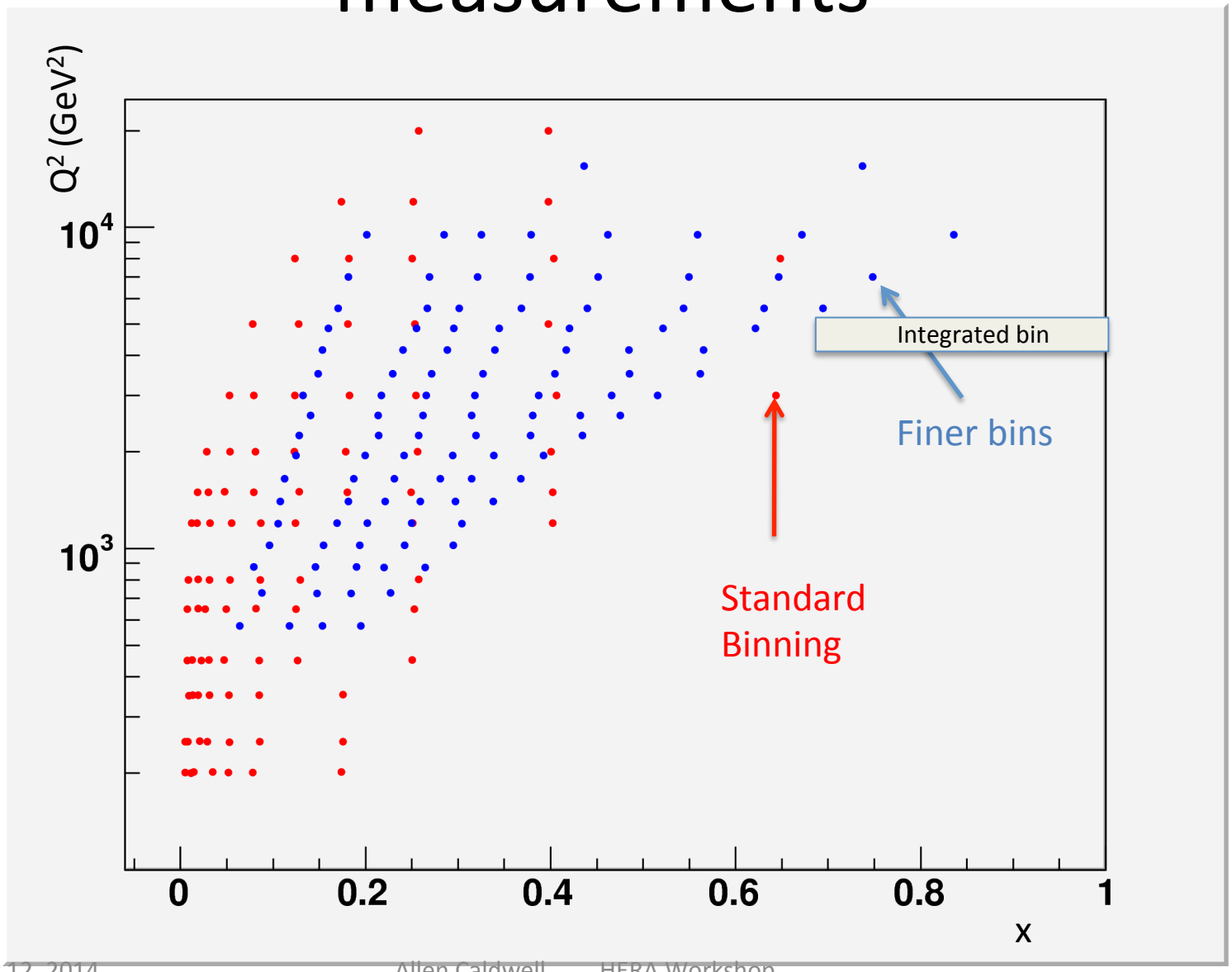
No jet found

$x>x_{Edge}$

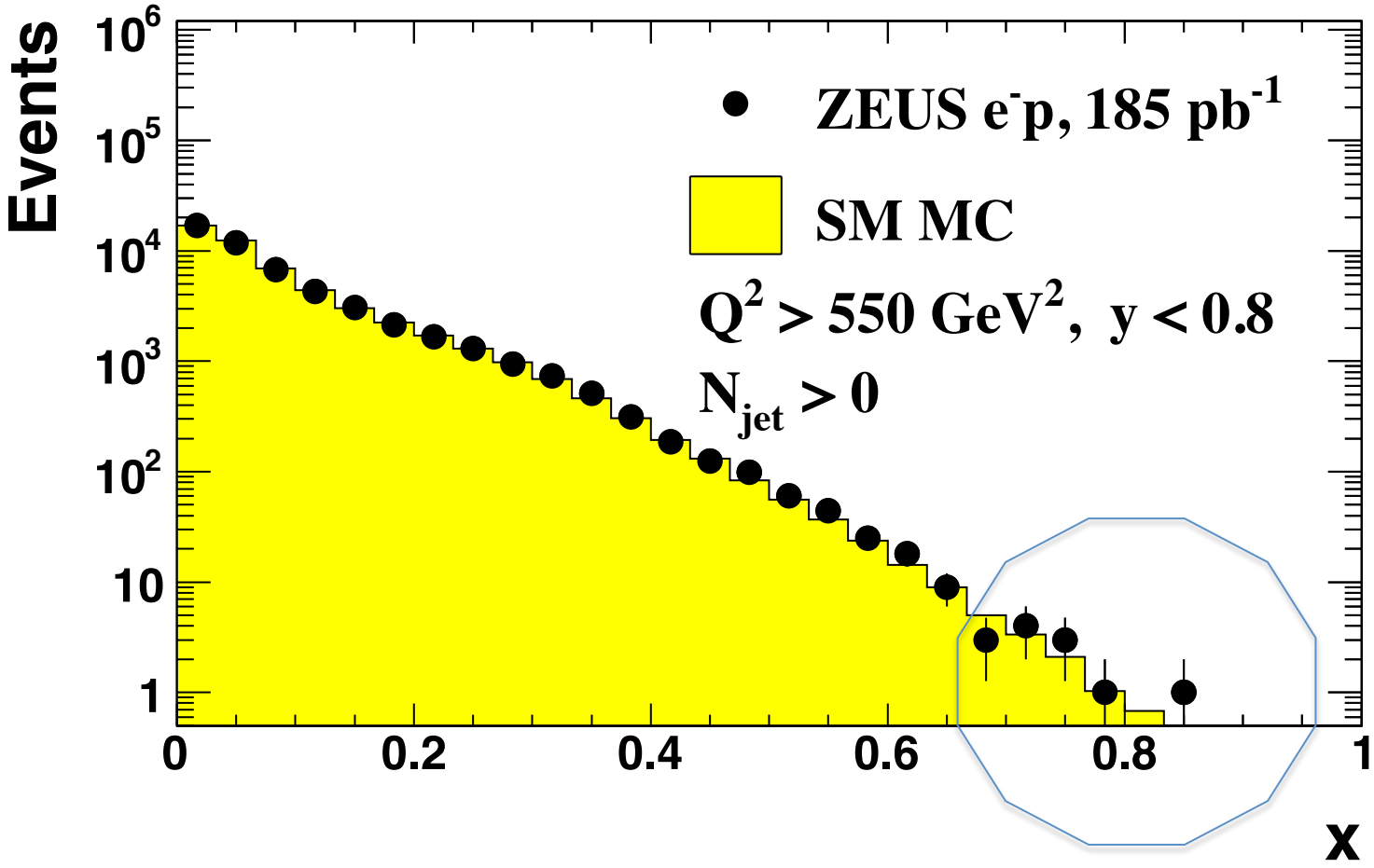


Jet definition: $E_T>10$ GeV, $\theta_{jet}>0.12$ only 0,1 jet events used

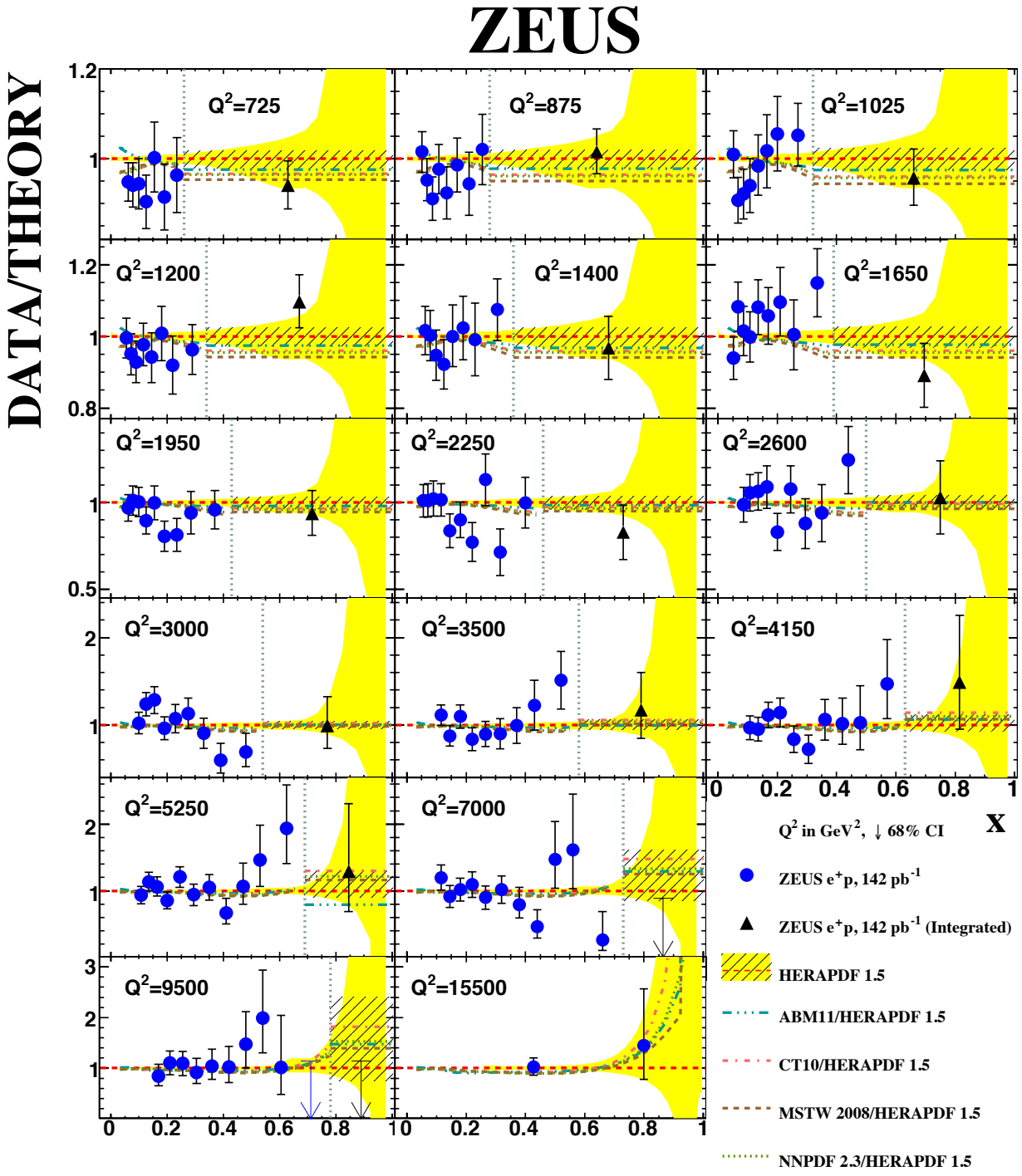
Fine-grained cross section measurements



ZEUS Collaboration; H. Abramowicz et al.
Measurement of Neutral Current ep Cross Sections at High Bjorken x with the ZEUS Detector
Phys. Rev. D 89 (2014) 072007



Small numbers of events – do statistical analysis carefully



Error bars indicate the range of probable values for the underlying cross section given the measured data. How to use this information in a fit ?

Use the observed number of events & calculate the probability to see this number given the model expectation.

5250	0.62	5	$1.76e-04$	$^{+55.2}_{-35.2}$	$^{+11.1}_{-10.5}$	$^{+9.2}_{-9.1}$	$^{-3.2}_{+4.6}$	$^{+0.1}_{+0.1}$	$^{+3.7}_{-3.7}$	$^{+0.6}_{-0.6}$
7000	0.12	93	$1.61e-02$	$^{+10.4}_{-10.4}$	$^{+4.0}_{-5.1}$	$^{+3.3}_{-3.6}$	$^{-1.1}_{+0.8}$	$^{+0.8}_{-0.5}$	$^{-0.7}_{+0.7}$	$^{+0.0}_{-0.0}$
7000	0.14	89	$1.25e-02$	$^{+10.6}_{-10.6}$	$^{+3.7}_{-5.2}$	$^{+3.4}_{-3.5}$	$^{-1.3}_{+1.2}$	$^{-0.5}_{+0.2}$	$^{-0.9}_{+0.9}$	$^{+0.0}_{-0.0}$
7000	0.18	68	$7.02e-03$	$^{+12.1}_{-12.1}$	$^{+3.9}_{-3.6}$	$^{+3.4}_{-3.4}$	$^{-0.6}_{+0.6}$	$^{-0.6}_{+0.4}$	$^{-0.4}_{+0.4}$	$^{+0.0}_{-0.0}$
7000	0.22	56	$5.60e-03$	$^{+13.4}_{-13.4}$	$^{+4.2}_{-4.2}$	$^{+3.9}_{-3.9}$	$^{-1.4}_{+1.1}$	$^{-0.4}_{+1.0}$	$^{-0.4}_{+0.4}$	$^{+0.0}_{-0.0}$
7000	0.26	49	$3.79e-03$	$^{+14.3}_{-14.3}$	$^{+4.6}_{-4.8}$	$^{+3.9}_{-4.0}$	$^{-0.2}_{+2.1}$	$^{+0.2}_{-0.2}$	$^{-0.5}_{+0.5}$	$^{+0.0}_{-0.0}$
7000	0.32	41	$2.70e-03$	$^{+15.6}_{-15.6}$	$^{+5.1}_{-4.7}$	$^{+5.3}_{-4.5}$	$^{-1.4}_{+0.8}$	$^{-0.4}_{+0.4}$	$^{-0.2}_{+0.2}$	$^{+0.0}_{-0.0}$
7000	0.38	23	$1.52e-03$	$^{+20.9}_{-20.9}$	$^{+6.4}_{-6.2}$	$^{+5.5}_{-5.5}$	$^{-1.8}_{+1.7}$	$^{-0.7}_{+0.4}$	$^{+2.0}_{-2.0}$	$^{+0.0}_{-0.0}$
7000	0.44	17	$1.15e-03$	$^{+27.2}_{-21.3}$	$^{+8.4}_{-7.9}$	$^{+7.1}_{-7.1}$	$^{-2.7}_{+2.7}$	$^{-0.0}_{+0.2}$	$^{-2.4}_{+2.4}$	$^{+0.0}_{-0.0}$
7000	0.50	8	$5.38e-04$	$^{+41.8}_{-29.4}$	$^{+9.7}_{-10.3}$	$^{+9.5}_{-9.5}$	$^{-1.6}_{+1.4}$	$^{-0.3}_{+0.4}$	$^{+2.4}_{-2.4}$	$^{+0.1}_{-0.1}$
7000	0.56	4	$2.37e-04$	$^{+63.2}_{-38.2}$	$^{+12.3}_{-11.8}$	$^{+11.3}_{-11.3}$	$^{-3.4}_{+3.4}$	$^{+0.1}_{-0.0}$	$^{+1.2}_{-1.2}$	$^{+0.2}_{-0.2}$
7000	0.66	10	$2.30e-04$	$^{+36.7}_{-26.8}$	$^{+12.6}_{-13.6}$	$^{+12.1}_{-12.3}$	$^{-4.3}_{+2.5}$	$^{-0.3}_{-0.0}$	$^{+2.0}_{-2.0}$	$^{+0.9}_{-0.9}$
9500	0.17	76	$6.77e-03$	$^{+11.5}_{-11.5}$	$^{+5.6}_{-7.7}$	$^{+4.9}_{-4.9}$	$^{-2.0}_{+2.3}$	$^{+0.2}_{-0.2}$	$^{-0.6}_{+0.6}$	$^{+0.0}_{-0.0}$
9500	0.21	53	$3.87e-03$	$^{+13.7}_{-13.7}$	$^{+5.8}_{-5.1}$	$^{+4.3}_{-4.5}$	$^{-1.1}_{+1.8}$	$^{-0.7}_{+0.4}$	$^{-1.1}_{+1.1}$	$^{+0.0}_{-0.0}$
9500	0.25	40	$2.27e-03$	$^{+15.8}_{-15.8}$	$^{+4.8}_{-4.9}$	$^{+4.5}_{-4.5}$	$^{-2.0}_{+1.5}$	$^{+0.2}_{+0.4}$	$^{+0.1}_{-0.1}$	$^{+0.0}_{-0.0}$
9500	0.31	27	$1.50e-03$	$^{+19.2}_{-19.2}$	$^{+5.7}_{-8.1}$	$^{+5.2}_{-5.3}$	$^{-2.6}_{+1.5}$	$^{-0.5}_{+0.2}$	$^{+1.5}_{-1.5}$	$^{+0.0}_{-0.0}$
9500	0.36	19	$8.89e-04$	$^{+25.5}_{-20.3}$	$^{+6.6}_{-6.1}$	$^{+5.9}_{-5.9}$	$^{-1.0}_{+1.9}$	$^{-0.4}_{+0.2}$	$^{-0.3}_{+0.3}$	$^{+0.0}_{-0.0}$
9500	0.42	12	$5.64e-04$	$^{+33.1}_{-24.8}$	$^{+11.3}_{-7.5}$	$^{+13.4}_{-7.3}$	$^{-1.0}_{+2.4}$	$^{-0.8}_{+0.5}$	$^{-0.8}_{+0.8}$	$^{+0.0}_{-0.0}$
9500	0.48	8	$3.63e-04$	$^{+41.7}_{-29.4}$	$^{+10.5}_{-10.4}$	$^{+9.2}_{-9.2}$	$^{-2.6}_{+2.4}$	$^{-0.4}_{+0.6}$	$^{-3.4}_{+3.4}$	$^{+0.0}_{-0.0}$
9500	0.54	5	$2.31e-04$	$^{+55.2}_{-35.2}$	$^{+14.3}_{-13.7}$	$^{+12.5}_{-12.2}$	$^{-1.7}_{+3.6}$	$^{-0.2}_{+0.6}$	$^{+5.7}_{-5.7}$	$^{+0.1}_{-0.1}$
9500	0.61	4	$1.39e-04$	$^{+63.3}_{-38.3}$	$^{+15.5}_{-15.4}$	$^{+14.6}_{-14.8}$	$^{-4.2}_{+4.2}$	$^{+0.0}_{-0.1}$	$^{+0.4}_{-0.4}$	$^{+0.4}_{-0.4}$
9500	0.71	1	$1.50e-05$	$^{+158.0}_{-58.0}$	$^{+21.1}_{-19.8}$	$^{+18.9}_{-18.9}$	$^{-3.3}_{+4.5}$	$^{-0.4}_{+0.3}$	$^{+4.8}_{-4.8}$	$^{+1.3}_{-1.3}$

This uncertainty refers to how well we know the underlying cross section assuming that our only knowledge is the observed number of events. Not the uncertainty that belongs in a fit.

Integrated bins

Q^2 (GeV^2)	$x_{\rm edge}$	N	$I(x)$ (pb/GeV^2)	$\delta_{\rm stat}$ (%)	$\delta_{\rm sys}$ (%)	$\delta_{\rm u}$ (%)	δ_1 (%)	δ_2 (%)	δ_3 (%)	δ_4 (%)
725	0.63	504	$7.71e-02$	$^{+4.5}_{-4.5}$	$^{+2.8}_{-3.2}$	$^{+1.5}_{-1.3}$	$^{+1.3}_{-2.0}$	$^{+0.0}_{-0.1}$	$^{+1.9}_{-1.9}$	$^{+0.3}_{-0.3}$
875	0.64	671	$5.12e-02$	$^{+3.9}_{-3.9}$	$^{+2.3}_{-1.9}$	$^{+1.2}_{-1.2}$	$^{-0.1}_{+1.0}$	$^{+0.0}_{-0.0}$	$^{+1.3}_{-1.3}$	$^{+0.5}_{-0.5}$
1025	0.66	414	$2.75e-02$	$^{+4.9}_{-4.9}$	$^{+3.4}_{-3.6}$	$^{+1.5}_{-1.5}$	$^{-1.6}_{+0.8}$	$^{+0.0}_{-0.0}$	$^{+2.7}_{-2.7}$	$^{+0.6}_{-0.6}$
1200	0.67	368	$1.80e-02$	$^{+5.2}_{-5.2}$	$^{+3.7}_{-2.9}$	$^{+1.7}_{-1.6}$	$^{-1.5}_{+2.5}$	$^{+0.0}_{-0.0}$	$^{+1.4}_{-1.4}$	$^{+0.8}_{-0.8}$
1400	0.68	202	$1.04e-02$	$^{+7.0}_{-7.0}$	$^{+3.5}_{-3.8}$	$^{+2.1}_{-2.0}$	$^{-1.8}_{+1.3}$	$^{+0.0}_{-0.0}$	$^{+2.1}_{-2.1}$	$^{+1.0}_{-1.0}$
1650	0.69	173	$5.91e-03$	$^{+7.6}_{-7.6}$	$^{+4.4}_{-4.1}$	$^{+2.3}_{-2.2}$	$^{-1.7}_{+2.0}$	$^{+0.1}_{-0.1}$	$^{+2.6}_{-2.6}$	$^{+1.2}_{-1.2}$
1950	0.71	74	$2.51e-03$	$^{+11.6}_{-11.6}$	$^{+5.2}_{-5.1}$	$^{+3.1}_{-3.0}$	$^{-1.9}_{+1.9}$	$^{+0.0}_{-0.1}$	$^{+3.1}_{-3.1}$	$^{+1.6}_{-1.6}$
2250	0.73	51	$1.84e-03$	$^{+14.0}_{-14.0}$	$^{+7.1}_{-7.6}$	$^{+4.1}_{-4.1}$	$^{-3.0}_{+2.2}$	$^{+0.0}_{-0.0}$	$^{+4.9}_{-4.9}$	$^{+2.0}_{-2.0}$
2600	0.75	36	$9.65e-04$	$^{+16.7}_{-16.7}$	$^{+6.9}_{-6.6}$	$^{+4.8}_{-4.8}$	$^{-1.9}_{+2.6}$	$^{+0.0}_{-0.0}$	$^{+3.0}_{-3.0}$	$^{+2.5}_{-2.5}$
3000	0.77	19	$4.90e-04$	$^{+25.5}_{-20.3}$	$^{+11.0}_{-10.9}$	$^{+6.6}_{-6.6}$	$^{-3.9}_{+4.1}$	$^{+0.1}_{-0.1}$	$^{+6.8}_{-6.8}$	$^{+3.2}_{-3.2}$
3500	0.79	17	$3.01e-04$	$^{+27.2}_{-21.3}$	$^{+11.6}_{-11.5}$	$^{+8.0}_{-7.8}$	$^{-3.3}_{+3.5}$	$^{+0.1}_{-0.2}$	$^{+6.4}_{-6.4}$	$^{+4.0}_{-4.0}$
4150	0.81	5	$8.19e-05$	$^{+55.2}_{-35.2}$	$^{+14.6}_{-15.0}$	$^{+11.9}_{-11.8}$	$^{-3.7}_{+2.1}$	$^{+0.0}_{-0.3}$	$^{+6.4}_{-6.4}$	$^{+5.1}_{-5.1}$
5250	0.85	3	$1.98e-05$	$^{+75.7}_{-42.3}$	$^{+18.6}_{-18.1}$	$^{+14.3}_{-14.3}$	$^{-2.9}_{+4.7}$	$^{+0.2}_{+0.0}$	$^{+8.1}_{-8.1}$	$^{+6.9}_{-6.9}$
7000	0.87	1	$5.56e-06$	$^{+158.0}_{-58.0}$	$^{+29.0}_{-26.4}$	$^{+24.3}_{-24.1}$	$^{-3.7}_{+10.7}$	$^{-0.4}_{+0.0}$	$^{-1.3}_{+1.3}$	$^{+9.4}_{-9.4}$
9500	0.89	1	$5.60e-06$	$^{+158.0}_{-58.0}$	$^{+58.3}_{-63.4}$	$^{+53.5}_{-53.5}$	$^{-19.0}_{-0.7}$	$^{+0.0}_{+0.0}$	$^{+19.2}_{-19.2}$	$^{+13.0}_{-13.0}$

2. Why use the bin counts

The standard way to estimate the differential cross section at an (x, Q^2) point within a bin is given by

$$\frac{d\sigma^{\rm Data}}{dxdQ^2} = \frac{d\sigma^{\rm theory}}{dxdQ^2}\frac{\sigma^{\rm Data}}{\sigma^{\rm theory}}$$

with

$$\sigma^{\rm Data} = \frac{N}{\mathcal{L}a}$$

where we will choose for concreteness $N = 2$ as the observed number of events in our example below. $\mathcal{L}$ is the integrated luminosity in the data set, and a is the acceptance in the bin:

$$a = \frac{N^{\rm MC}_{\rm reconstructed}}{N^{\rm MC}_{\rm generated}} \quad .$$

(1) The 'old-fashioned-standard-prescription' $\sigma \pm \delta_\sigma$:

$$\delta_\sigma \quad = \quad \frac{\sqrt{N}}{\mathcal{L}a} = \frac{\sigma}{\sqrt{N}} \quad .$$

For $N = 2$, we have a fractional uncertainty of $\delta_\sigma/\sigma = 0.71$. For $N = 0$, this prescription breaks down and it has often been the case that the uncertainty for $N = 1$ is taken.

(2) with confidence levels $\sigma^{+\delta^{\rm up}}_{-\delta^{\rm down}}$ (central interval definition):

$$\delta^{\rm up} \quad \rightarrow \quad \sum_{i=0}^{i=N-1} P(N|\sigma + \delta^{\rm up}) \leq 0.16$$

$$\delta^{\rm down} \quad \rightarrow \quad \sum_{i=N+1}^{i=\infty} P(N|\sigma - \delta^{\rm down}) \leq 0.16$$

(3) with credibility intervals $\sigma^{+\delta^{\rm up}}_{-\delta^{\rm down}}$:

$$\int_{\sigma-\delta^{\rm down}}^{\sigma+\delta^{\rm up}} P(\sigma'|N)d\sigma' \quad = \quad 0.68$$

where $P(\sigma|N)$ is the posterior probability density for σ given the measured number of events. For $N = 2$ and using a flat prior on the expected number of events gives $\delta^{\rm down} = 0.63$ and $\delta^{\rm up} = 2.62$

$2^{+2.62}_{-0.63}$

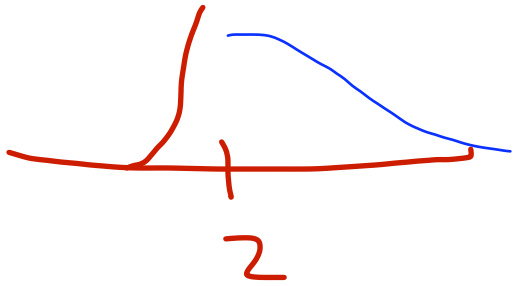
Now imagine that we have two pdf sets, one which predicts $\nu_1 = 0.01$ and a second which predict $\nu_2 = 4.5$ events in the bin. The standard fitting approach is to calculate a χ^2 by comparing the measured differential cross section to the predicted one, using the fractional uncertainty ON THE DATA quoted above. The resulting contributions to χ^2 using the three prescriptions given above are:

(1) 'old-fashioned-standard-prescription' :

$$\begin{aligned} \chi^2_1 &= (N-\nu_1)^2/N = 2.0 \\ \chi^2_2 &= (N-\nu_2)^2/N = 3.1 \end{aligned}$$

We see that the smaller prediction is preferred.

(2) confidence levels (similar to credibility below - not yet done):



(3) Credibility interval:

$$\begin{aligned} \chi^2_1 &= (N-\nu_1)^2/(0.63)^2 = 10.0 \\ \chi^2_2 &= (N-\nu_2)^2/(2.62)^2 = 0.9 \end{aligned}$$

In this case, there is a strong preference for the higher prediction.

Now we see what happens if instead of performing a χ^2 fit using the differential cross section, we instead compare the predicted and observed event counts in the bin. The probabilities are

$$\begin{aligned} P(N=2|\nu_1=0.01) &= \frac{e^{-0.01}0.01^2}{2!} = 5\cdot 10^{-5} \\ P(N=2|\nu_2=4.5) &= \frac{e^{-4.5}4.5^2}{2!} = 0.11 \ . \end{aligned}$$

The preference for the larger prediction is even more pronounced. We see that in this case, there is very strong discriminating power between the different predictions, and very different probabilities would be obtained for $\nu = 0.1, 0.01, 0.001$. This is because we are making use of the probability to observe the given number of events assuming the expectation - the statistical fluctuations are handled correctly. Note also that $N = 0$ poses no problem in the fitting.

3. Information required for event count fitting

In order to make full use of the data as outlined above, more information has to be provided than is normally the case. We need to supply the information which allows the pdf fitters to calculate a predicted number of events in each of the bins in which we report an event count. This prediction has to be calculated for every instance of the pdf parameters by integrating over the full kinematic phase space:

$$\nu(\Delta x, \Delta Q^2) = \int_{\Delta x, \Delta Q^2} \int \frac{d\nu^{pred}(x', Q'^2)}{dx' dQ'^2} P(x, Q^2 | x', Q'^2) dx' dQ'^2 dx dQ^2$$

where $\frac{d\nu^{pred}(x', Q'^2)}{dx' dQ'^2}$ is the probability density to have an event with true kinematics at (x', Q'^2) and $P(x, Q^2|x', Q'^2)$ is the probability density to reconstruct an event at (x, Q^2) given the true kinematics at (x', Q'^2). We approximate this integral with

$$\nu_j = \sum_i a_{ij} \nu_i$$

$$a_{ij} = \frac{\sum_{k=1}^{M_i} \omega_k I(k \in j)}{\sum_{k=1}^{M_i} \omega_k}$$

with M_i the number of events generated in bin i, ω_k the weight given to the $k^{\rm th}$ event, and $I(k \in j) = 1$ if event k is reconstructed in bin j, else $I(k \in j) = 0$.

The matrix a_{ij} should be provided with the bins i covering the full phase space which can give non-negligible contribution to our predictions. I.e., the bins defined in i should go beyond the bins defined in j, and can be made finer to reduce migration uncertainties.

The probability for the model to yield the data is

$$P(D|\vec{\lambda}) = \prod_j \frac{e^{-\nu_j(\vec{\lambda})} \nu_j(\vec{\lambda})^{n_j}}{n_j!}$$

it is being used.

To take account of the systematic uncertainties, we need to further provide the following matrices:

$\delta a^{\rm unc}$ The matrix of uncorrelated systematic uncertainties (given as a difference from the nominal matrix a). E.g., MC statistical uncertainties would enter into this matrix. The MC statistical uncertainty for a_{ij} is

$$\delta a_{ij}^{\rm unc} = \frac{\sqrt{\sum_{k=1}^{M_i} \omega_k^2 I(k \in j)}}{\sum_{k=1}^{M_i} \omega_k}$$

Note that we should have probably at least 25 MC events in all bins and no large spread of weights for this to be reasonably accurate.

$\delta a^{\rm l}$ The correlated systematic uncertainties. For each of our correlated systematic uncertainties, l, we generate a new matrix $a^{\rm l}$ given from a one sigma deviation in the quantity of interest, and then calculate

$$\delta a_{ij}^{\rm l} = a_{ij}^{\rm l} - a_{ij} \quad .$$

The prediction including systematic uncertainties is

$$\nu_j' = \nu_j + \sum_i x_i \delta_{ij}^{\rm unc} \nu_i + \sum_l x_l \sum \delta_{ij}^l \nu_i \quad .$$

The x's are drawn from a Normal distribution of width 1, and a penalty is added to the χ^2 of $\sum_k x_k^2$.

Note that there is additional statistical uncertainty from the data and the systematic

The dream:

• Find someone interested in this project who is also technically savvy with the data

• Discuss the procedure to see if we believe it can be made to work. Should include discussions with the fitting teams.

• Run some simple examples to get a feeling for how much we can learn from 'doing it right'. E.g., translate P(Data|prediction) into chi squared for fits.

• If the gain is significant, then carry this out.

• And hopefully someone in H1 would also be interested …

• worst case – give some prescriptions to fitting teams how better to account for the 'probability of the data given the model' in their analysis.