Run-merging in Rivet

So, just to get my head around this myself, here are my thoughts about merging of runs in Rivet.

Assume we have a number of runs, and that we have access to the cross section for the processes in run i as reported by the generator: σ_i (I will here ignore that there is an uncertainty on the total cross section $\delta\sigma_i$). Also, assume that we have the total sum of weights $S_{w,i}$, and that a given 'raw' histogram bin will have the sum of weights $S_{w,i}^{[b]}$ and the sum of squares $S_{w^2,i}^{[b]}$. When plotted (finalized) the histogram bin will typically have a cross section $\sigma_i^{[b]} = \sigma_i S_{w,i}^{[b]}/S_{w,i}$ with an estimated error of $\delta\sigma_i^{[b]} = \sigma_i\sqrt{S_{w^2,i}^{[b]}}/S_{w,i}$. But we can also imagine other situations where the final plot point is a fraction of the total $r^{[b]} = \sigma_i^{[b]}/\sigma_i$, or if we want to be a bit more general any ratio of cross sections $r^{[b/a]} = \sigma_i^{[b]}/\sigma_i^{[a]}$. We note that for unit weighted events an individual generator run corresponds to an integrated luminosity $\mathcal{L}_i = S_{w,i}/\sigma_i$.

Now if the two runs have exactly **the same** process, the weights in the combined histogram will simply be the added weights,

$$\begin{aligned} S_w &= \sum_i S_{w,i} \\ S_w^{[b]} &= \sum_i S_{w,i}^{[b]} \\ S_{w^2}^{[b]} &= \sum_i S_{w^2,i}^{[b]} \end{aligned}$$

and the cross section for the combined files will be a weighted average

$$\sigma = \frac{1}{S_w}\sum_i S_{w,i}\sigma_i$$

For each bin we will the have the plot value $\sigma^{[b]} = \sigma_i S_i^{[b]}/S_w$ with an estimated error of $\delta\sigma^{[b]} = \sigma_i\sqrt{S_{w^2}^{[b]}}/S_w$ as for the individual histograms.

Now, if the histograms are already normalized to cross section, we can still combine them by doing a weighted average

$$\sigma^{[b]} = \frac{1}{S_w}\sum_i \sigma_i^{[b]} S_{w,i}$$

with the error estimate

$$\delta\sigma^{[b]} = \frac{1}{S_w}\sqrt{\sum_i \left(\delta\sigma_i^{[b]} S_{w,i}\right)^2}$$

which clearly gives the same thing as the unnormalized case, and since all σ_i are the same, it corresponds to weighting with the respective integrated luminosity.

Of course for the ratios of cross sections, the situation is somewhat more complicated. Again the correct thing is to add raw histograms and then do the ratio:

$$r^{[b/a]} = \frac{\sum_i S_{w,i}^{[b]}}{\sum_i S_{w,i}^{[a]}}$$

If we don't have the raw histograms, all we can do is a weighted average

$$r^{[b/a]} = \frac{1}{S_w}\sum_i r_i^{[b/a]} S_{w,i}$$

which is probably reasonable for most cases, but having access to the raw histograms is, of course, better.

Turning now to the case of adding histograms with **different** processes, the case where the histograms are already normalized to cross section is the easiest, since we can then simply add

$$\begin{aligned}
\sigma &= \sum_i \sigma_i \\
\sigma^{[b]} &= \sum_i \sigma_i^{[b]} \\
\delta\sigma^{[b]} &= \sqrt{\sum_i \left(\delta\sigma_i^{[b]}\right)^2}.
\end{aligned}$$

For adding the raw histograms we need to expand out the cross sections in terms of weights

$$\begin{aligned}
\sigma\frac{S_w^{[b]}}{S_w} &= \sum_i \sigma_i \frac{S_{w,i}^{[b]}}{S_{w,i}} \\
\sigma^2\frac{S_{w^2}^{[b]}}{S_w^2} &= \sum_i \frac{\sigma_i^2 S_{w^2,i}^{[b]}}{S_{w,i}^2}.
\end{aligned}$$

In other words, the ratio of the weights to the total is a cross-section-weighted average, and we can write

$$
\begin{aligned}
S_w^{[b]} &= \frac{S_w}{\sigma} \sum_i \sigma_i \frac{S_{w,i}^{[b]}}{S_{w,i}} \\
S_{w^2}^{[b]} &= \frac{S_w^2}{\sigma^2} \sum_i \frac{\sigma_i^2 S_{w^2,i}^{[b]}}{S_{w,i}^2}.
\end{aligned}
$$

However, the S_w is arbitrary (two equations and three unknowns above), and this is related to the fact that the combined histograms no longer necessarily corresponds to a particular integrated luminosity. This, in turn, means that it is not possible to first combine histograms of different processes and then combine these with others of identical combinations.

If the different runs do correspond to the same integrated luminosity, of course the combined run should correspond to the same. One reasonable way of obtaining this could be to let the integrated luminosity for the merged sample be the cross-section-weighted average of the individual samples

$$
\frac{S_w}{\sigma} = \mathcal{L} = \frac{1}{\sigma} \sum_i \sigma_i \mathcal{L}_i = \frac{\sum_i S_{w,i}}{\sigma}
$$

Although it is possible to combine *finalized* histograms of different runs with identical processes, it is only possible to combine finalized histograms which are normalized to cross section in an automated way (although probably doing a cross-section-weighted average would probably be reasonable when dealing with ratios). Instead we want to combine raw histograms, and finalize them after merging. The combination should be done as follows.

For for runs with identical processes:

$$\begin{aligned}
S_w &= \sum_i S_{w,i} \\
\sigma &= \frac{1}{S_w} \sum_i \sigma_i S_{w,i} \\
S_w^{[b]} &= \sum_i S_{w,i}^{[b]} \\
S_{w^2}^{[b]} &= \sum_i S_{w^2,i}^{[b]}
\end{aligned}$$

For runs with different processes:

$$\begin{aligned}
S_w &= \sum_i S_{w,i} \\
\sigma &= \sum_i \sigma_i \\
S_w^{[b]} &= \frac{S_w}{\sigma} \sum_i \sigma_i \frac{S_{w,i}^{[b]}}{S_{w,i}} \\
S_{w^2}^{[b]} &= \frac{S_w^2}{\sigma^2} \sum_i \frac{\sigma_i^2 S_{w^2,i}^{[b]}}{S_{w,i}^2}.
\end{aligned}$$