Glen Cowan
RHUL Physics
8 February, 2008

Covariance matrix for histogram made using seed events

Consider a histogram created by generating $N_{\rm seed}$ "seed events" according to a given distribution, and then using each seed event to generate a further number $N_{\rm sim}$ events by smearing with a certain resolution function, which results in a final histogram with n_i entries in the ith bin. Because each seed event contributes $N_{\rm sim}$ times to the histogram, there are correlations between bins. This note gives an expression for the covariance $\mathrm{cov}[n_i, n_j]$ between the number of entries in any pair of bins. The result is given in equation (12) below for the case where the number of seed events $N_{\rm seed}$ is treated as a constant and by equation (18) for the case where $N_{\rm seed}$ is a Poisson variable with mean $\nu_{\rm seed}$.

The result for $\mathrm{cov}[n_i, n_j]$ of course also provides the standard deviations

$$\sigma[n_i] = \sqrt{\mathrm{cov}[n_i, n_i]} \ , \tag{1}$$

and the matrix of correlation coefficients,

$$\rho_{ij} = \frac{\mathrm{cov}[n_i, n_j]}{\sigma[n_i]\sigma[n_j]} \ . \tag{2}$$

1 Case of fixed number of seed events

In this section we consider the case where the number of seed events, $N_{\rm seed}$, is taken as a constant; it does not fluctuate upon repetition of the experiment. Suppose the indices i, j, denote a bin of the final histogram (after smearing) and let s be a bin in which the seed event is found. The seed events are labelled with indices $a, b = 1, \ldots, N_{\rm seed}$. Let n_{ia} be the number of events found in bin i from seed event a. The total number of entries found in bin i is obtained by summing over all seed events:

$$n_i = \sum_{a=1}^{N_{\rm seed}} n_{ia} \ . \tag{3}$$

The covariance $\mathrm{cov}[n_i, n_j]$ is given by

$$\mathrm{cov}[n_i, n_j] = E[n_i n_j] - E[n_i]E[n_j] \ . \tag{4}$$

We can find $E[n_i]$ by first considering the case of a single seed event a. This is

$$
\begin{aligned}
E[n_{ia}] &= \sum_{n_{ia}} n_{ia} P(n_{ia}) \\
&= \sum_{s} \sum_{n_{ia}} n_{ia} P(n_{ia}|s) Q_s \\
&= \sum_{s} E[n_{ia}|s] Q_s \\
&= N_{\rm sim} \sum_{s} P_{is} Q_s \equiv N_{\rm sim} \overline{P}_i \ ,
\end{aligned} \tag{5}
$$

where the sums are over all possible values of n_{ia} (0 to $N_{\rm sim}$) and s (over all bins). Here $P(n_{ia})$ is the probability to find n_{ia} entries in bin i from seed event a, $P(n_{ia}|s)$ is the corresponding conditional probability given that the seed event is in bin s, and and Q_s is the probability for the seed event to be in bin s. In the third line of (5), $E[n_{ia}|s] = N_{\rm sim} P_{is}$ is the expected number of entries in bin i given that the seed event is in bin s, and P_{is} is the probability to observe an event in bin i given that the seed event is in bin s. By symmetry this is the same for all seed events and therefore it does not depend on the index a. In the last line of (5) we used the notation

$$
\overline{P}_i \equiv \sum_{s} P_{is} Q_s \ . \tag{6}
$$

Summing over $N_{\rm seed}$ independent seed events gives

$$
E[n_i] = N_{\rm seed} N_{\rm sim} \overline{P}_i \ . \tag{7}
$$

We now need the expectation value $E[n_i n_j]$. This can be written

$$
E[n_i n_j] = E\left[\left(\sum_{a=1}^{N_{\rm seed}} n_{ia} \right) \left(\sum_{b=1}^{N_{\rm seed}} n_{jb} \right) \right] = \sum_{a,b=1}^{N_{\rm seed}} E[n_{ia} n_{jb}] \ . \tag{8}
$$

Of the $N_{\rm seed}^2$ terms in the double sum, $N_{\rm seed}$ have $a = b$. For the $N_{\rm seed}^2 - N_{\rm seed}$ terms with $a \neq b$, the seed events are independent and therefore

$$
E[n_{ai} n_{jb}] = E[n_{ia}] E[n_{jb}] = N_{\rm sim}^2 \overline{P}_i \overline{P}_j \quad (a \neq b) \ . \tag{9}
$$

For the $N_{\rm seed}$ terms with $a = b$, consider again a single seed event a found in a given bin s. For a fixed s, n_{ia} and n_{ja} are multinomially distributed, and therefore the conditional expectation value of $n_{ia} n_{ja}$ for fixed s is

$$
E[n_{ia} n_{ja}|s] = N_{\rm sim}^2 P_{is} P_{js} + N_{\rm sim} P_{is} (\delta_{ij} - P_{js}) \ . \tag{10}
$$

(If one regards $N_{\rm sim}$ as a Poisson variable rather than fixed, then the second term, proportional to $N_{\rm sim}$, in (10) is absent.) We need to average the expectation value (10) over s and multiply by $N_{\rm seed}$ to obtain

$$
\begin{aligned}
E[n_i n_j] &= N_{\text{seed}} \sum_s E[n_{ia} n_{ja} | s] Q_s \\
&= N_{\text{seed}} \left(\sum_s N_{\text{sim}}^2 P_{is} P_{js} Q_s + \sum_s N_{\text{sim}} P_{is} (\delta_{ij} - P_{js}) Q_s \right) \\
&= N_{\text{seed}} N_{\text{sim}}^2 \overline{P_i P_j} + N_{\text{seed}} N_{\text{sim}} (\delta_{ij} \overline{P}_i - \overline{P_i P_j}) \ . \qquad (11)
\end{aligned}
$$

Putting together the ingredients gives the covariance for n_i and n_j,

$$
\text{cov}[n_i, n_j] = N_{\text{seed}} N_{\text{sim}}^2 (\overline{P_i P_j} - \overline{P}_i \overline{P}_j) + N_{\text{seed}} N_{\text{sim}} (\overline{P}_i \delta_{ij} - \overline{P_i P_j}) \ . \qquad (12)
$$

As mentioned above, if one takes N_{sim} to be Poisson distributed rather than fixed, then the second term in (12) proportional to N_{sim} is absent. The required ingredients are thus the matrix of probabilities P_{is} (probability to observe the event in bin i given a seed event in bin s), and the probability to have a seed event in bin s, Q_s, both of which can be estimated, e.g., from Monte Carlo.

2 Case of random number of seed events

In the previous section, the number of seed events N_{seed} was treated as a constant. We can also treat it as a random variable following a Poisson distribution with a mean ν_{seed}. To find the covariance $\text{cov}[n_i, n_j]$ we need the expectation values $E[n_i]$ and $E[n_i n_j]$. For $E[n_i]$ we have

$$
\begin{aligned}
E[n_i] &= \sum_{N_{\text{seed}}=0}^{\infty} P(N_{\text{seed}}; \nu_{\text{seed}}) E[n_i | N_{\text{seed}}] \\
&= \sum_{N_{\text{seed}}=0}^{\infty} P(N_{\text{seed}}; \nu_{\text{seed}}) N_{\text{seed}} N_{\text{sim}} \overline{P}_i \\
&= \nu_{\text{seed}} N_{\text{sim}} \overline{P}_i \ , \qquad (13)
\end{aligned}
$$

where $P(N_{\text{seed}}; \nu_{\text{seed}})$ is the Poisson probability for N_{seed} with mean value ν_{seed}. Equation (7) for $E[n_i]$ is with constant N_{seed} and so this was used for $E[n_i | N_{\text{seed}}]$ to obtain the second line of (13) above.

For $E[n_i n_j]$ we have

$$
\begin{aligned}
E[n_i n_j] &= \sum_{N_{\text{seed}}=0}^{\infty} P(N_{\text{seed}}; \nu_{\text{seed}}) E[n_i n_j | N_{\text{seed}}] \\
&= \sum_{N_{\text{seed}}=0}^{\infty} P(N_{\text{seed}}; \nu_{\text{seed}}) \left(\sum_{a,b=1}^{N_{\text{seed}}} E[n_{ia} n_{jb} | N_{\text{seed}}] \right) \ . \qquad (14)
\end{aligned}
$$

In the sums over a and b in equation (14) there are N_{seed} terms with $a = b$, and for these we can use equation (11). For the remaining $N_{\text{seed}}(N_{\text{seed}} - 1)$ terms with $a \neq b$ we have

$$E[n_{ia}n_{jb}|N_{\text{seed}}] = E[n_{ia}|N_{\text{seed}}]E[n_{jb}|N_{\text{seed}}] = N_{\text{sim}}^2 \overline{P}_i \overline{P}_j \qquad (a \neq b) \; , \tag{15}$$

because two distinct seed events are not correlated. The required expectation value for a given N_{seed} is therefore

$$\begin{aligned} E[n_i n_j|N_{\text{seed}}] &= N_{\text{seed}} E[n_{ia}n_{jb}|N_{\text{seed}}] + N_{\text{seed}}(N_{\text{seed}} - 1)E[n_{ia}|N_{\text{seed}}]E[n_{jb}|N_{\text{seed}}] \qquad (16) \\ &= N_{\text{seed}} \left[N_{\text{sim}}^2 \overline{P_i P_j} + N_{\text{sim}}(\overline{P}_i \delta_{ij} - \overline{P_i P_j}) \right] + N_{\text{seed}}(N_{\text{seed}} - 1) N_{\text{sim}}^2 \overline{P}_i \overline{P}_j \; . \end{aligned}$$

We can then use (16) together with (14). To evaluate the result we need the Poisson expectation value $E[N_{\text{seed}}]$, and we can use the Poisson variance $V[N_{\text{seed}}] = E[N_{\text{seed}}^2] - (E[N_{\text{seed}}])^2 = \nu_{\text{seed}}$ to find

$$E[N_{\text{seed}}^2] = \nu_{\text{seed}}(\nu_{\text{seed}} + 1) \; . \tag{17}$$

Using the resulting value of $E[n_i n_j]$ together with $E[n_i]$ gives the final expression for the covariance,

$$\text{cov}[n_i, n_j] = \nu_{\text{seed}} \left[N_{\text{sim}}^2 \overline{P_i P_j} + N_{\text{sim}}(\overline{P}_i \delta_{ij} - \overline{P_i P_j}) \right] \; . \tag{18}$$