The effect of animal-to-animal variability on conditional probabilities

September 13, 2011

1 Reorientation sequences

As a worm navigates, it performs a sequence of turns. When turns occur sufficiently close to each other, they are grouped into a reorientation event. These reorientations have several characteristics, e.g. the types of turn of which it is composed, the difference in heading direction before and after, the duration of the run leading into it. We wish to know if the characteristics of one reorientation are independent of the characteristics of previous reorientations. We focus on the types of turn only, and we restrict attention to the two most frequent types: *turn* and *reversal-turn*.

Consider a sequence of two successive reorientations. The characteristics of these reorientations form a pair of outcomes: (X_1, X_2). We are asking whether or not $P(X_2|X_1) = P(X_2)\ \forall (X_1, X_2)$.

Note that this method can confuse animal-to-animal variability with causation. E.g. suppose that one of the worms makes a greater fraction of turns than the others. In this situation, if the first reorientation in a pair is a turn it would make it more likely that the worm in question is the turn-favoring one, making it more likely that the next reorientation is a turn. Thus the probability distributions would not show independence even in the absence of actual causation. We will discuss the effect of animal-to-animal variability quantitatively in §3 and show how to measure the amount of variability in §4 and Appendix A.

2 Measuring probabilities

Each track consists of a series of runs and reorientations. Within each track, all non-overlapping pairs of reorientation such that each reorientation type was either *turn* or *reversal-turn* were found. The number of pairs of type (*turn*,*turn*), (*turn*,*reversal-turn*), (*reversal-turn*,*turn*) and (*reversal-turn*,*reversal-turn*) were counted. The joint probabilities were computed by dividing these counts by their sum. Marginal and conditional probabilities were then computed in the usual way. Standard errors in these probabilities are computed assuming these counts have a binomial distribution.

Note that all probabilities are actually conditional on the reorientations being either of type *turn* or *reversal-turn*.

3 Effect of animal-to-animal variability

In this section we will look at the effect that animal-to-animal variability can have on conditional probabilities in the case of binary output, i.e. the characteristics of each reorientation are reduced to two possible outcomes which we will call $\{+,-\}$. In our specific case these two outcomes will be the two most important reorientation types (*turn* or *reversal-turn*). We will assume that animal-to-animal variability is the only effect and that there is no memory

Label the individual worms with $i = 1 \dots M$. Then, for a single reorientation

$$P(i) = q_i, \qquad P(+|i) = p_i, \qquad P(-|i) = 1 - p_i. \tag{1}$$

The prior probability that a reorientation was from a particular animal, q_i, will be proportional to that animal's reorientation rate, which may not be the same for all animals. We also introduce the notation

$$\bar{p} = \sum_i q_i p_i, \qquad \sigma_p^2 = \sum_i q_i (p_i - \bar{p})^2. \tag{2}$$

So that

$$P(+) = \bar{p}, \qquad P(-) = 1 - \bar{p}. \tag{3}$$

Then, using Bayes' theorem

$$P(i|+) = \frac{P(+|i)P(i)}{P(+)} = \frac{q_i p_i}{\bar{p}}, \qquad P(i|-) = \frac{P(-|i)P(i)}{P(-)} = \frac{q_i(1-p_i)}{1-\bar{p}}. \tag{4}$$

For pairs of successive reorientations, we use the notation

$$P(\text{next} = a|\text{prev} = b) = P(a|b). \tag{5}$$

We can calculate conditional probabilities for next reorientation, given previous reorientation

$$P(a|b) = \sum_i P(a, i|b) = \sum_i P(a|i, b)P(i|b) = \sum_i P(a|i)P(i|b), \tag{6}$$

where we assumed that there is no memory, and that animal-to-animal variability is the only effect, in going from the second expression to the third. This gives

$$\begin{aligned} P(+|+) &= \bar{p} + \frac{\sigma_p^2}{\bar{p}}, & P(+|-) &= \bar{p} - \frac{\sigma_p^2}{1-\bar{p}}, \\ P(-|+) &= 1 - \bar{p} - \frac{\sigma_p^2}{\bar{p}}, & P(-|-) &= 1 - \bar{p} + \frac{\sigma_p^2}{1-\bar{p}}. \end{aligned} \tag{7}$$

4 Estimating animal-to-animal variability

Estimating σ_p in (2) is tricky, as we can't keep track of animal identity during collisions (a problem that does not affect $\bar{p}$, see (9) below). What we will do is to restrict attention to those tracks that start during the first 20 min and are at least 20 min long. This ensures that each animal is sampled at most once and that we have a decent amount of data for each animal. We also only consider the first 20 min of each track (whilst this reduces the amount of data, it makes it possible to estimate bias and standard error, see Appendix A). We will then assume that the estimate for σ_p applies to the whole experiment. Whilst the turning rates do change during the experiment, we only care about the ratio of turning rates between different animals, which we will assume remains constant.

For each track, we count the number of turns in each track, T_i, and the number of reversal-turns, R_i.

$$\hat{n}_i = T_i + R_i, \qquad N = \sum_i \hat{n}_i, \qquad \hat{q}_i = \frac{\hat{n}_i}{N}, \qquad \hat{p}_i = \frac{T_i}{\hat{n}_i}. \tag{8}$$

We can estimate $\bar{p}$ by substituting (8) into (2). However, estimating σ_p^2 is more complicated, as the variance of the sample also receives contributions from the Binomial statistics of each track. In Appendix A, we discuss the construction of approximately unbiased estimators and their variance. The result is

$$\begin{aligned} \hat{\bar{p}} &= \frac{\sum_i T_i}{N}, & \hat{\sigma}_p^2 &= \frac{N+1}{N} \sum_i \left(\hat{q}_i (\hat{p}_i - \hat{\bar{p}})^2 - \frac{\hat{q}_i(1-\hat{q}_i)\hat{p}_i(1-\hat{p}_i)}{\hat{n}_i} \right), \\ \mathrm{Var}(\hat{\bar{p}}) &\approx \frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{N}, & \mathrm{Var}(\hat{\sigma}_p^2) &\approx \sum_i \left(\left(\frac{\partial \hat{\sigma}_p^2}{\partial T_i} \right)^2 T_i + \left(\frac{\partial \hat{\sigma}_p^2}{\partial R_i} \right)^2 R_i \right). \end{aligned} \tag{9}$$

The expressions for the partial derivatives of $\hat{\sigma}_p^2$ are a colossal mess. They can be found in Appendix A, (23).

As the second expression for $\hat{\bar{p}}$ doesn't rely on maintaining animal identity, we can use the full data set to estimate it. As we will use a larger data set for $\hat{\bar{p}}$ than for $\hat{\sigma}_p^2$, we ignore possible correlations in their standard errors when computing standard errors in derived quantities.

Appendix A Bias and standard error

In this appendix, we will remove bias and calculate variance for estimators of the mean probability of a turn and the weighted variance (2). Our approach is similar to that of [1].

We have:

$$P(\hat{p}_i | \hat{n}_i) \sim \frac{\mathrm{Binomial}(\hat{n}_i, p_i)}{\hat{n}_i}, \qquad P(\hat{n}_i) \sim \mathrm{Poisson}(r_i t), \tag{10}$$

where r_i is the reorientation rate of animal i and t is the duration of the tracks. We also define

$$n_i = r_i t, \qquad N_0 = \sum_i n_i, \qquad \epsilon_i = \frac{\hat{n}_i}{n_i} - 1, \qquad q_i = \frac{r_i}{\sum_j r_j} = \frac{n_i}{N_0}. \tag{11}$$

Expanding in ϵ_i leads to

$$\begin{aligned}
\hat{q}_i &= q_i \left[1 + \left(\epsilon_i - \sum_j q_j \epsilon_j\right) - \left(\epsilon_i - \sum_j q_j \epsilon_j\right)\left(\sum_k q_k \epsilon_k\right)\right] + \mathcal{O}\left(\epsilon_i^3\right), \\
\hat{q}_i \hat{q}_j &= q_i q_j \left[1 + \left(\epsilon_i + \epsilon_j - 2\sum_k q_k \epsilon_k\right) + \left(\epsilon_i \epsilon_j - (\sum_k q_k \epsilon_k)^2\right)\right] + \mathcal{O}\left(\epsilon_i^3\right), \\
\frac{1}{N} &= \frac{1}{N_0}\left[1 - \left(\sum_k q_k \epsilon_k\right) + \left(\sum_k q_k \epsilon_k\right)^2\right] + \mathcal{O}\left(\epsilon_i^3\right).
\end{aligned} \tag{12}$$

Using the Poisson distribution, one can show that

$$\mathbb{E}\left[\epsilon_i\right] = 0, \qquad \mathbb{E}\left[\epsilon_i^2\right] = \frac{1}{n_i}, \qquad \mathbb{E}\left[\epsilon_i^{k+3}\right] = \mathcal{O}\left(\frac{1}{n_i^2}\right) \quad k \geq 0, \tag{13}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[\hat{q}_i\right] &= q_i + \mathcal{O}\left(\frac{1}{n_i^2}\right), \\
\mathbb{E}\left[\hat{q}_i \hat{q}_j\right] &= q_i q_j \left[1 + \frac{\delta_{ij}}{n_i} - \frac{1}{N_0}\right] + \mathcal{O}\left(\frac{1}{n_i^2}\right), \\
\mathbb{E}\left[\frac{1}{N}\right] &= \frac{1}{N_0}\left[1 + \frac{1}{N_0}\right] + \mathcal{O}\left(\frac{1}{n_i^2}\right).
\end{aligned} \tag{14}$$

Using the Binomial distribution, one can show that

$$\mathbb{E}\left[\hat{p}_i | \hat{n}_i\right] = p_i, \qquad \mathbb{E}\left[\hat{p}_i^2 | \hat{n}_i\right] = p_i^2 + \frac{p_i(1-p_i)}{n_i}. \tag{15}$$

The obvious estimator for $\bar{p}$ works just fine:

$$\mathbb{E}\left[\sum_i \hat{q}_i \hat{p}_i \middle| \hat{n}_i\right] = \sum_i \hat{q}_i p_i, \qquad \mathbb{E}\left[\sum_i \hat{q}_i p_i\right] = \sum_i q_i p_i + \mathcal{O}\left(\frac{1}{n_i^2}\right) = \bar{p} + \mathcal{O}\left(\frac{1}{n_i^2}\right). \tag{16}$$

However, the obvious estimator for σ_p^2 fails at the first hurdle:

$$\mathbb{E}\left[\sum_i \hat{q}_i (\hat{p}_i - \hat{\bar{p}})^2 \middle| \hat{n}_i\right] = \sum_i \hat{q}_i p_i^2 - \left(\sum_i \hat{q}_i p_i\right)^2 + \sum_i \frac{\hat{q}_i(1-\hat{q}_i) p_i (1-p_i)}{\hat{n}_i}. \tag{17}$$

We can correct this,

$$\mathbb{E}\left[\sum_i \hat{q}_i(\hat{p}_i-\hat{\bar{p}})^2-\sum_i \frac{\hat{q}_i(1-\hat{q}_i)\hat{p}_i(1-\hat{p}_i)}{\hat{n}_i-1}\middle|\hat{n}_i\right]=\sum_i \hat{q}_i p_i^2-\left(\sum_i \hat{q}_i p_i\right)^2, \tag{18}$$

but it fails at the next hurdle:

$$\mathbb{E}\left[\sum_i \hat{q}_i p_i^2-\left(\sum_i \hat{q}_i p_i\right)^2\right]=\frac{N_0-1}{N_0}\left(\sum_i q_i p_i^2-\left(\sum_i q_i p_i\right)^2\right)+\mathcal{O}\left(\frac{1}{n_i^2}\right). \tag{19}$$

Luckily, this can be corrected:

$$\begin{aligned}\mathbb{E}\left[\frac{N+1}{N}\left(\sum_i \hat{q}_i(\hat{p}_i-\hat{\bar{p}})^2-\sum_i \frac{\hat{q}_i(1-\hat{q}_i)\hat{p}_i(1-\hat{p}_i)}{\hat{n}_i-1}\right)\middle|\hat{n}_i\right]&\\ =\frac{N+1}{N}\left(\sum_i \hat{q}_i p_i^2-\left(\sum_i \hat{q}_i p_i\right)^2\right),&\\ \mathbb{E}\left[\frac{N+1}{N}\left(\sum_i \hat{q}_i p_i^2-\left(\sum_i \hat{q}_i p_i\right)^2\right)\right]=\sum_i q_i p_i^2-\left(\sum_i q_i p_i\right)^2+\mathcal{O}\left(\frac{1}{n_i^2}\right)&\\ =\sigma_p^2+\mathcal{O}\left(\frac{1}{n_i^2}\right).&\end{aligned} \tag{20}$$

To estimate the variances, we use the approach

$$\begin{aligned}\mathrm{Var}(\hat{A})&\approx\sum_i\left(\left(\frac{\partial\hat{A}}{\partial T_i}\right)^2\mathrm{Var}(T_i)+\left(\frac{\partial\hat{A}}{\partial R_i}\right)^2\mathrm{Var}(R_i)\right)\\ &=\sum_i\left(\left(\frac{\partial\hat{A}}{\partial T_i}\right)^2 T_i+\left(\frac{\partial\hat{A}}{\partial R_i}\right)^2 R_i\right).\end{aligned} \tag{21}$$

This relatively simple for $\hat{\bar{p}}$:

$$\begin{aligned}\hat{\bar{p}}=\sum_i \hat{q}_i\hat{p}_i=\frac{\sum_i T_i}{\sum_i(T_i+R_i)},\qquad \frac{\partial\hat{\bar{p}}}{\partial T_i}&=\frac{\sum_i R_i}{\left(\sum_i(T_i+R_i)\right)^2},\\ \frac{\partial\hat{\bar{p}}}{\partial R_i}&=-\frac{\sum_i T_i}{\left(\sum_i(T_i+R_i)\right)^2},\\ \mathrm{Var}(\hat{\bar{p}})&\approx\frac{\left(\sum_i T_i\right)\left(\sum_i R_i\right)}{\left(\sum_i(T_i+R_i)\right)^3}=\frac{\hat{\bar{p}}(1-\hat{\bar{p}})}{N}.\end{aligned} \tag{22}$$

but it is much more complicated for $\hat{\sigma}_p^2$:

$$\frac{\partial \hat{\sigma}_p^2}{\partial T_i} = A_i + B_i(1-\hat{p}_i), \qquad \frac{\partial \hat{\sigma}_p^2}{\partial R_i} = A_i + B_i \hat{p}_i, \tag{23}$$

where

$$\begin{aligned} A_i = &- \frac{\hat{\sigma}_p^2}{N(N+1)} + \frac{N+1}{N}\left(\frac{\hat{q}_i(1-\hat{q}_i)\hat{p}_i(1-\hat{p}_i)}{(\hat{n}_i-1)^2}\right) \\ &+ \frac{N+1}{N^2}\left(\hat{p}_i^2 - \left(\sum_j \hat{q}_j \hat{p}_j^2\right) - 2\hat{p}_i\hat{\bar{p}} + 2\hat{\bar{p}}^2 - \frac{(1-2\hat{q}_i)\hat{p}_i(1-\hat{p}_i)}{\hat{n}_i - 1}\right. \\ &\qquad \left. + \left(\sum_j \frac{q_j(1-2\hat{q}_j)\hat{p}_j(1-\hat{p}_j)}{\hat{n}_j - 1}\right)\right), \\ B_i = &\frac{N+1}{N}\left(\frac{2\hat{q}_i\hat{p}_i}{\hat{n}_i} - \frac{2\hat{q}_i\hat{\bar{p}}}{\hat{n}_i} - \frac{\hat{q}_i(1-\hat{q}_i)(1-2\hat{p}_i)}{\hat{n}_i - 1}\right). \end{aligned} \tag{24}$$

The combined expression for $\mathrm{Var}(\hat{\sigma}_p^2)$ is too messy to write out here.

Clearly our estimate of $\mathrm{Var}(\hat{\bar{p}}) \to 0$ as $N \to \infty$. We can show that this is also true for $\hat{\sigma}_p^2$. Let M be the number of animals that make at least one reorientation. Note that $N > M$. We have

$$\begin{aligned} &\hat{q}_i \sim \mathcal{O}(M^{-1}), \qquad \sum_i \sim \mathcal{O}(M), \\ &\hat{p}_i \sim \mathcal{O}(1), \qquad \hat{n}_i, T_i, R_i \sim \mathcal{O}(N/M). \end{aligned} \tag{25}$$

then, we have

$$\begin{aligned} A_i \sim &\mathcal{O}(N^{-2}) + \mathcal{O}(1)\mathcal{O}(M^{-1})\mathcal{O}(M/N) \\ &+ \mathcal{O}(N^{-1})\left(\mathcal{O}(1) + \mathcal{O}(M)\mathcal{O}(M^{-1})\mathcal{O}(M/N)\right) \\ \sim &\mathcal{O}(N^{-1}), \\ B_i \sim &\mathcal{O}(1)\mathcal{O}(M^{-1})\mathcal{O}(M/N) \\ \sim &\mathcal{O}(N^{-1}). \end{aligned} \tag{26}$$

Then

$$\mathrm{Var}(\hat{\sigma}_p^2) \sim \mathcal{O}(M)\mathcal{O}(N^{-2})\mathcal{O}(N/M) \sim \mathcal{O}(N^{-1}). \tag{27}$$

Thus our estimate of $\mathrm{Var}(\hat{\sigma}_p^2) \to 0$ as $N \to \infty$ as well as our estimate of $\mathrm{Var}(\hat{\bar{p}})$. In addition to the biases vanishing as $t \to \infty$, this suggests that our estimators are consistent.

References

[1] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D Nonlinear Phenomena* **125** (Jan., 1999) 285–294.