MTH notes on JKK's lecture 1

For discrete events, A_i in some sample space of $\mathcal{S}_A$ of possible events, we can make statements about the probability of A_i occurring, denoted $\mathbb{P}(A_i)$.

$$0 \leq \mathbb{P}(A_i) \leq 1 \qquad \forall A_i \in \mathcal{S}_A \tag{1}$$

$$1 = \sum_{A_i \in \mathcal{S}_A} \mathbb{P}(A_i) \tag{2}$$

$\forall$ means "for all".

$\sum$ means a summation over values of a variable. So $\sum_{A_i \in \mathcal{S}_A}$ means we are going to do a summation over all values of A_i that possible in the sample space. We often number the outcomes from 1 to $|\mathcal{S}_A|$ in which case we can write the second eqn as:

$$1 = \sum_{i=1}^{|\mathcal{S}_A|} \mathbb{P}(A_i) \tag{3}$$

where i is indexing the possible events.

Compound event: combinations of simple events

E.g. Let x be an encoding of the result of a random trial into some specified sample space. So, if Trevor's student records the behavior of a lobster we could use the encoding:

Simple vent	code	$\mathbb{P}(x = \text{code})$
Motionless	0	0.5
Random	1	0.1
Walk toward	2	0.1
Walk away	3	0.1
Turn toward	4	0.1
Turn away	5	0.1

What is the probability of "some motion"?

$$\mathbb{P}(A \text{ OR } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A, B) \tag{4}$$

$$\mathbb{P}(c \in [1,2,3,4,5]) = \mathbb{P}(c=1) + \mathbb{P}(c=2) + \mathbb{P}(c=3) + \mathbb{P}(c=4) + \mathbb{P}(c=5) \tag{5}$$

$$= 0.1 + 0.1 + 0.1 + 0.1 + 0.1 = 0.5 \tag{6}$$

Note that each of the simple events are mutually exclusive with the others in our encoding. So the joint probability ($\mathbb{P}(A, B)$) of any of 2 of them occurring in the same random trial is 0. So we don't subtract anything. We could also solve this using eqn 2 above:

$$\mathbb{P}(A \text{ OR not} A) = \mathbb{P}(A) + \mathbb{P}(\text{not } A) = 1 \tag{7}$$

$$\mathbb{P}(\text{not } A) = 1 - \mathbb{P}(A) = 1 - .5 = 0.5 \tag{8}$$

If our dataset had observations for 2 second we could say: $X = [0, 2]$ or $X = [x_0, x_1]$ where $x_0 = 0$ is the encoding of "motionless in the first second" and $x_1 = 2$ meand "Walk toward in second two."

The likelihood is the probability of data identical to what we have observed if the model were true:

$$\mathbb{P}(X) = \mathbb{P}(x_0, x_1) \tag{9}$$

$$= \mathbb{P}(x_0)\mathbb{P}(x_1 \mid x_0) \tag{10}$$

from the general multiplication rule of probabilities: $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B \mid A) = \mathbb{P}(B)\mathbb{P}(A \mid B)$.

To make some progress we could assume that behavior in one second is independent of the behavior in every other seconds. So $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. In this case: $\mathbb{P}(x_1 \mid x_0) = \mathbb{P}(x_1)$

$$\begin{aligned} \mathbb{P}(X) &= \mathbb{P}(x_0)\mathbb{P}(x_1 \mid x_0) && (11) \\ &= .5 \times 0.1 = 0.05 && (12) \end{aligned}$$

What if we have genetic data. $X = [M_g = \text{Aa}, S_g = \text{Aa}]$ where M_g is the genotype of a mother at some locus. and S_g is the genotype of her son. What is the likelihood $\mathbb{P}(X)$?

We need a model. What if we assume that: mating is random, the frequency of A in the population is some unknown parameter q, and there is no selection on the locus or meiotic drive (or other weird stuff)?

$$\begin{aligned} \mathbb{P}(X \mid q) &= \mathbb{P}(M_g = \text{Aa} \mid q)\mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, q) && (13) \\ \mathbb{P}(M_g = \text{Aa} \mid q) &= 2q(1-q) \text{ from Hardy-Weinberg equil.} && (14) \end{aligned}$$

How do we get $\mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, q)$? We can use the law of total probability:

$$\mathbb{P}(A) = \sum_{B \in \mathcal{S}_B} [\mathbb{P}(A \mid B)\mathbb{P}(B)] \quad (15)$$

in this case use this to sum over the possible other event (B) that would be helpful to know: dad's genotype, D_g. Using Hardy-Weinberg again:

event, d	$\mathbb{P}(D_g = d)$	$\mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, D_g = d)$
AA	q^2	0.5 (from Mendel half AA, half Aa)
Aa	$2q(1-q)$	0.5 (from Mendel one quarter AA, half Aa, one quarter aa)
aa	$(1-q)^2$	0.5 (from Mendel half Aa, half aa)

Note that we could condition the third column on q, but the population allele frequencies does not matter when we know both parent's genotype. So we drop that from the notation for the sake of brevity.

$$\begin{aligned} \mathcal{G} &= \{\text{AA, Aa, aa}\} && (16) \\ \mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, q) &= \sum_{d \in \mathcal{G}} [\mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, D_g = d)\mathbb{P}(D_g = d)] \\ &= 0.5 \sum_{d \in \mathcal{G}} [\mathbb{P}(D_g = d)] \\ &= 0.5 && (17) \end{aligned}$$

interestingly this componenet of the likelihood is not a function of q. But the full likelihood is:

$$\begin{aligned} \mathcal{L}(q) = \mathbb{P}(X \mid q) &= \mathbb{P}(M_g = \text{Aa} \mid q)\mathbb{P}(S_g = \text{Aa} \mid M_g = \text{Aa}, q) \\ &= 2q(1-q) \times 0.5 \\ &= q(1-q) && (18) \end{aligned}$$

So, the likelihood is a function of our unknown parameter q. If you play around with this, you will find that the highest likelihood is obtained when $q = 0.5$ and the likelihood is 0.25.