Mitchell's argument for the Schema Theorem mixes probabilities and expected values in an unusual way, and I wanted to give you a more methodical derivation. It turns out that my argument leads to a slightly better bound than the "official" Schema Theorem provides. Everything that follows pertains to a standard genetic algorithm with population size n (an even number), string length L, mutation probability p_m, and one-point crossover probability p_c. The population at time t is $P(t)$, the selection is fitness-proportional, and the (strictly positive) fitness function is f.

In class, we learned that for any individual $x \in P(t)$, the probability that x is selected on any given parental draw is

$$p_{\text{sel}}(x) = \frac{f(x)}{\sum_{y \in P(t)} f(y)} \ .$$

Using this we can prove the following result.

Subset Selection Lemma: If Q is any subset of $P(t)$, then the probability that any given parental draw yields a member of Q is

$$p_{\text{sel}}(Q) = \frac{\sum_{x \in Q} f(x)}{\sum_{y \in P(t)} f(y)} \ . \tag{1}$$

Consequently, the expected number of members of Q selected as parents for $P(t+1)$ is

$$np_{\text{sel}}(Q) = n \left(\frac{\sum_{x \in Q} f(x)}{\sum_{y \in P(t)} f(y)} \right) = |Q| \left(\frac{\frac{1}{|Q|} \sum_{x \in Q} f(x)}{\frac{1}{n} \sum_{y \in P(t)} f(y)} \right) \ , \tag{2}$$

where $|Q|$ is the number of individuals in Q.

Proof: For any $x \in Q$, the probability that x gets selected on any given parental draw is $p_{\text{sel}}(x)$, so the probability that any given parental draw yields some element of Q is simply

$$\sum_{x \in Q} p_{\text{sel}}(x) = \sum_{x \in Q} \left(\frac{f(x)}{\sum_{y \in P(t)} f(y)} \right) \ ,$$

which is the same as the formula for $p_{\text{sel}}(Q)$ in (1). Next, for every i, $1 \leq i \leq n$, define a zero-one random variable Z_i as follows:

$$Z_i = \begin{cases} 1 & \text{if parental draw } i \text{ is in } Q \\ 0 & \text{if parental draw } i \text{ is not in } Q \end{cases}$$

Note that

$$E(Z_i) = 1 \times p_{\text{sel}}(Q) + 0 \times (1 - p_{\text{sel}}(Q)) = p_{\text{sel}}(Q)$$

for all i. The number of Q-members among the parents for $P(t+1)$ is $\sum_{i=1}^{n} Z_i$, so the expected number of Q-members among the parents for $P(t+1)$ is

$$\sum_{i=1}^{n} E(Z_i) = np_{\text{sel}}(Q) \ ,$$

which is what the second part of the theorem statement leading up to (2) asserts. □

Another way of writing the right-hand side of (2) is

$$|Q| \left(\frac{\text{average fitness of individuals in } Q}{\text{average fitness of individuals in } P(t)} \right) .$$

In other words, if the average fitness of individuals in Q exceeds the average fitness of all individuals in $P(t)$, then we "expect" to see at least $|Q|$ representatives of Q among the parents for $P(t+1)$. Now for the Schema Theorem. I will prove it in steps before stating it.

First, what is a schema? I like to think of a schema as a set of strings, but some people like to think of a schema as the defining symbol for such a set. If $L = 8$, for example, the schema with defining symbol $*1******0$ is the set of all bit strings of length 8 with second bit 1 and eighth bit 0. The *order* of a schema H, written $o(H)$, is the number of defined bits (i.e. non-stars) in the defining symbol for H. The example schema above has order 2. The *defining length* of a schema H, written $d(H)$, is the "distance" between the leftmost and rightmost defined bits in the defining symbol for H. The example schema above has defining length 6.

Now let H be a schema and let $Q = H \cap P(t)$. Q is just the subset of all H-representatives in $P(t)$. By the Subset Selection Lemma, the probability that a member of Q is selected on any given parental draw is $p_{\text{sel}}(Q)$ from equation (1). Draw n parents and pair them as follows: 1 mates with 2, 3 mates with 4, ... , $n-1$ mates with n. That yields $n/2$ mating pairs, and we number them using index j, which runs from 1 to $n/2$. For each j, pair j consists of parents $2j-1$ and $2j$. Each pair j gives rise to two offspring. Three things can happen: both offspring are members of H, one of the offspring is a member of H, or neither of the offspring is a member of H. Define W_j for $1 \le j \le n/2$ via

$$W_j = \begin{cases} 2 & \text{if both offspring from pair } j \text{ are in } H \\ 1 & \text{if one offspring from pair } j \text{ is in } H \\ 0 & \text{if neither offspring from pair } j \text{ is in } H \ . \end{cases}$$

Observe that the number of H-representatives in $P(t+1)$ is the random quantity $\sum_{j=1}^{n/2} W_j$.

I'll be calculating a lower bound on the $E(W_j)$, which in turn yields a lower bound on the expected number of H-representatives in $P(t+1)$. First let's figure out a lower bound on $\text{Prob}\{W_j = 2\}$. For now, consider W_1. $W_1 = 2$ precisely when the two offspring of parents 1 and 2 are both in H. This could happen in many ways. For example, any two parents 1 and 2 could give rise to two offspring in H if mutation happens in just the right way, no matter whether the parents cross over or not. While that possibility is unlikely for most pairs of parents, it still exists. A more realistic possibility is for both parents 1 and 2 to be in H, in which case the products of crossing them over with each other will also be in H. If these crossover products don't mutate out of H, we'll have $W_1 = 2$.

To sum up the last observation, one specific way for parents 1 and 2 to give rise to two H-offspring is for the following two things to occur:

- Parents 1 and 2 both lie in H. This occurs with probability $(p_{\text{sel}}(Q))^2$. Whether these parents cross over or not, their crossover products will both lie in H.
- Mutation affects none of the defined bits in either of these crossover products. Since the probability that a bit fails to mutate is $1-p_m$, the probability that mutation affects none of the defined bits on either crossover-product is

$$(1-p_m)^{2o(H)} .$$

Since selection, crossover, and mutation are independent random operations, we can assemble these two criteria into the lower bound

$$\text{Prob}\{W_1 = 2\} \geq (p_{\text{sel}}(Q))^2 (1-p_m)^{2o(H)} . \tag{3}$$

Similar reasoning applies to determining a lower bound on the $\text{Prob}\{W_1 = 1\}$. Here is one way for $W_1 = 1$ to happen.

- Parents 1 and 2 both lie in H. This occurs with probability $(p_{\text{sel}}(Q))^2$. Whether these parents cross over or not, their crossover products will both lie in H.
- Mutation affects none of the defined bits in one of these crossover products but causes the other crossover product to mutate out of H. This happens with probability

$$(1-p_m)^{o(H)} \left(1-(1-p_m)^{o(H)}\right) + \left(1-(1-p_m)^{o(H)}\right)(1-p_m)^{o(H)}$$

which is the same as

$$2(1-p_m)^{o(H)} - 2(1-p_m)^{2o(H)} .$$

The probability that $W_1 = 1$ happens in this way is

$$(p_{\text{sel}}(Q))^2 \left(2(1-p_m)^{o(H)} - 2(1-p_m)^{2o(H)}\right) . \tag{4}$$

Another way for $W_1 = 1$ to occur is as follows.

- Exactly one of parents 1 and 2 lies in H. This occurs with probability

$$p_{\text{sel}}(Q)(1-p_{\text{sel}}(Q)) + (1-p_{\text{sel}}(Q))p_{\text{sel}}(Q) .$$

- Crossover, if it occurs, happens at a point outside the part of the symbol for H that constitutes its defining length. The probability that crossover happens at all is p_c, and the probability that it happens within the defining-length segment is $d(H)/(L-1)$, since $d(H)$ crossover points out of a total of $L-1$ lie within the defining-length segment. Accordingly, the probability that crossover, if it occurs, happens outside the defining-length segment is

$$1 - p_c \frac{d(H)}{L-1} .$$

With at least this probability, exactly one of the crossover-products of parents 1 and 2 will lie in H.
- Mutation affects none of the defined bits in the H-member obtained via crossover. This occurs with probability

$$(1-p_m)^{o(H)} .$$

The probability that $W_1 = 1$ happens in this way is therefore

$$(5) \quad (p_{\rm sel}(Q)(1-p_{\rm sel}(Q)) + (1-p_{\rm sel}(Q))p_{\rm sel}(Q)) \left(1 - p_c \frac{d(H)}{L-1}\right) (1-p_m)^{o(H)} .$$

Summing (4) and (5) gives the lower bound

$$(6) \quad {\rm Prob}\{W_1 = 1\} \geq 2p_{\rm sel}(Q) \left(1 - p_c \frac{d(H)}{L-1}\right) (1-p_m)^{o(H)} + 2(p_{\rm sel}(Q))^2 p_c \frac{d(H)}{L-1}(1-p_m)^{o(H)} - 2(p_{\rm sel}(Q))^2 (1-p_m)^{2o(H)} .$$

Since

$$E(W_1) = {\rm Prob}\{W_1 = 1\} + 2{\rm Prob}\{W_1 = 2\} ,$$

the lower bounds in (3) and (6) imply that

$$(7) \quad E(W_1) \geq 2p_{\rm sel}(Q) \left(1 - p_c \frac{d(H)}{L-1}\right) (1-p_m)^{o(H)} + 2(p_{\rm sel}(Q))^2 p_c \frac{d(H)}{L-1}(1-p_m)^{o(H)} .$$

Since the W_j are identically distributed, right-hand side of (7) bounds all the W_j from below. Accordingly, because

$$E\left(|H \cap P(t+1)|\right) = \sum_{j=1}^{n/2} E(W_j) ,$$

we have

$$(8) \quad E\left(|H \cap P(t+1)|\right) \geq np_{\rm sel}(Q) \left(1 - p_c \frac{d(H)}{L-1}\right) (1-p_m)^{o(H)} + n(p_{\rm sel}(Q))^2 p_c \frac{d(H)}{L-1}(1-p_m)^{o(H)} .$$

The first term on the right-hand side of (8) is the same as the lower bound in Mitchell's version of the Schema Theorem. To see this, note that (1) and (2) imply that

$$np_{\rm sel}(Q) = n\frac{\sum_{x \in Q} f(x)}{\sum_{y \in P(t)} f(y)} = |Q| \left(\frac{\overline{f}(Q)}{\overline{f}(P(t))}\right) ,$$

where $\overline{f}$ means average fitness. The first term on the right-hand side of (8) is therefore

$$|Q|\frac{\overline{f}(Q)}{\overline{f}(P(t))} \left(1 - p_c \frac{d(H)}{L-1}\right) (1-p_m)^{o(H)} .$$

This is identical to the right-hand side of formula (1.2) near the bottom of page 29 of Mitchell's book if you make the appropriate substitutions. Specifically, with Mitchell's notation on the left and mine on the right,

$$m(H,t) = |Q| = |H \cap P(t)| ,$$
$$m(H,t+1) = |H \cap P(t+1)| ,$$
$$\widehat{u}(h,t) = \overline{f}(Q) ,$$

and

$$\overline{f}(t) = \overline{f}(P(t)) .$$

For what it's worth, the right-hand side of (8) contains an additional nonnegative term and thus provides a "tighter" lower bound than the "official" Schema Theorem.