During the lecture we computed the threshold for ORF length such that if a given ORF has length greater than the threshold, then it will be considered significant with level $\alpha = 0.05$. This computation was based on the probability

$$P(\text{'at least } k \text{ non-stop codons' }) = \left(\frac{61}{64}\right)^k .$$

We present here a bit more detailed derivation for the threshold for considering the ORF significant.

Now the null hypothesis H_0 is that the ORF has been created by chance. The test score is the length l of the ORF. More specifically, here we mean by l the number of non-stop codons between the start and stop codons (the full length of the ORF can be easily computed as $l + 2$). Now the probability that the ORF is exactly of length k is the same as the probability of observing exactly k non-stop codons after the start codon, followed by a stop codon. Thus we can write

$$P(l = k) = \left(\frac{61}{64}\right)^k \left(\frac{3}{64}\right), \quad k = 0, 1, \ldots. \tag{1}$$

Note that Equation (1) does not include the probability of the start codon, because without a start codon there would not be any ORF at all. Distribution in Equation (1) is recognized as the probability mass function of a geometric distribution:

$$P(l = k) = \mathit{Geometric}(k|p = 3/64).$$

Let $F(k|p)$ denote the cumulative distribution function (CDF, kertymäfunktio in Finnish) of a geometric distribution, that is,

$$F(k|p) = P(l \leq k|p).$$

The form of the CDF(see, e.g., Wikipedia) is

$$F(k|p) = 1 - (1 - p)^{k+1}.$$

The threshold k^* for considering the ORF significant with level $\alpha = 0.05$ is obtained by finding the smallest k^* such that $P(l \geq k^*) \leq 0.05$.

$$\begin{aligned} P(l \geq k^*) &= 1 - P(l < k^*) = 1 - P(l \leq k^* - 1) \\ &= 1 - F(k^* - 1|p = 3/64) = 1 - \left[1 - \left(\frac{61}{64}\right)^{k^*}\right] \\ &= \left(\frac{61}{64}\right)^{k^*} . \end{aligned}$$

Therefore, we must find k^* such that $\left(\frac{61}{64}\right)^{k^*}$ is 0.05 or smaller (note that this is the same formula as in the slides). The smallest k^* that satisfies this condition is 63.

To summarize, based on the derivation above, we would consider a single ORF significant (with $\alpha = 0.05$) if the number of non-stop codons between the start and stop codons is 63 or more.